AUTOMATICALLY EXTRACTING PERFORMANCE DATA FROM RECORDINGS OF TRAINED SINGERS

Johanna Devaney[1]
Michael I. Mandel[2]
Daniel P.W. Ellis[2]
Ichiro Fujinaga[1]
[1.] McGill University, [2.] Columbia University

Recorded audio offers a wealth of information for studying performance practice. This paper examines the challenges of automatically extracting performance information from audio recordings of the singing voice and discusses a technique the authors have developed for automatically identifying note onsets and offsets. With this information it is possible to study a number of different performance parameters, including note timings, intonation, vibrato rates, and dynamics. In this paper, we focus on the intonation of leading tones in solo soprano performances of Schubert's 'Ave Maria' where we found that the size of leading-tone/tonic intervals were not significantly different than other semitones.

This paper describes the challenges that arise when attempting to automatically extract performance data from unannotated recordings of the singing voice, and presents an overview of the history of performance analysis from recorded performances. The first task in the automatic extraction of performance data is to identify the onsets and offsets of notes, which can be used to calculate intonation, vibrato, and dynamic characteristics for each note. This paper describes an algorithm to automatically identify note onsets and offsets in recordings of the singing voice for which a score of the performance is available.

This paper also presents a study of intonation in solo vocal performance, where both the note onsets and offsets and fundamental frequency were estimated automatically. In the study, six soprano undergraduate vocal majors performed Schubert's 'Ave Maria' three times *a cappella* and three times with a recorded piano accompaniment. Our analysis of these recordings focused on the intonation of leading tones, and we found that the size of the leading-tone/tonic interval was not significantly different than other semitones performed in the piece, regardless of intervallic direction.

Analysis of Recorded Performances

A Brief History

Interest in studying recorded performances dates back almost as far as the birth of recordable media. Some of the earliest, and most extensive, research was done at the University of Iowa in the 1930s by Seashore and his colleagues (Seashore, 1938). Their work dealt primarily with pianists, violinists, and singers, focusing on the areas of timing (including tempo and duration), dynamics, intonation, and vibrato, and the researchers employed a number of techniques to study recorded performances. Piano performances were studied from both piano rolls and films of the movement of the hammers during the performance. The violin and singing analyses were based on amplitude and frequency information from recordings visualized by phonophotographic apparati. Though relatively accurate performance data could be assessed with these methods, they were extremely labour intensive, which limited the number of pieces that could be evaluated. Interest in empirical performance analysis diminished between the Second World War and the 1970's, in part due to its labouriousness. Its resurgence coincided with both a movement by musicologists away from equating scores with music and an increased interest by cognitive psychologists in music. Gabrielsson and Bengtsson undertook a number of systematic experiments on musical rhythm in performance (e.g., Bengtsson & Gabrielsson, 1980; 1983). Following up on this earlier research, Todd studied both rubato and dynamics in piano performance, developing models to account for their individual relationships to musical structure and their interaction (Todd, 1985, 1989). Similarly, Clarke examined how rhythm in piano performance could be related to both the structural hierarchy of a piece and note-level expressive gestures (Clark, 1989). In the early nineties, Repp (1992) performed extensive evaluations of timing in the piano music of Beethoven and Schumann. He found that the degree of *ritardando* in phrases in the performances he studied could be consistently related to the hierarchy of phrases, and observed that the higher the structural level, the more pronounced the *ritardandi*. Repp (1997) also analyzed the collected data for the Schumann performances and performances of a Chopin Etude, and found that the re-created versions of the performances based on the average of the timing variations were pleasing to listeners in perceptual. A comprehensive survey of research on musical performance through 2002 can be found in published reviews by Palme (1997) and Gabrielsson (1999, 2002).

Extraction of Performance Data

Historically, the piano has been the primary instrument of performance analysis for several reasons. One is the large amount of solo repertoire available, which allows for the examination of the performer in a context to which he or she is accustomed—in contrast to instruments where it is more typical to play in an ensemble. Another is the piano's percussive nature, which makes it possible to study timing with a high degree of precision. Also one can acquire accurate, minimally intrusive performance measurements from a pianist via MIDI technology. MIDI is used because it can record information about timing of the note onsets and offsets, as well as the key velocity, which corresponds to dynamics. In typical experiments, regular acoustic pianos are rigged with a system to measure the hammer action and output the measurements in MIDI format, such as Yamaha's Disklavier and Bösendorfer's SE series of pianos, which allows for information to be stored either as MIDI or in their proprietary formats that retain more information about the performance. One problem with piano performance studies is that they are limited to performances done on MIDI-enabled pianos, which often take place in a lab environment. A general, and perhaps more significant, issue for collecting performance data is that the precision of equipment to map physical instruments' motions to MIDI technology is severely limited for instruments other than the piano.

Extraction of performance data directly from recordings allows for the study of other instruments. However, accurate automated extraction of performance data is still an open problem, particularly for studies dealing with instruments that present similar challenges to the singing voice, namely flexible intonation capabilities and non-percussive onsets, such as unfretted string instruments. Since the mid-1990s there has been an increase in studies on these types of instruments particularly the violin (Fyk, 1995; Ornoy, 2008) and cello (Hong, 2003), which used either manual or semi-automatic methods to analyze recorded performances. Semi-automated systems are also used for analyzing recordings of piano music; the system proposed by Earis uses a 'manual beat tapping system' for synchronization that is corrected by both a computer-aided system and a second manual pass (Earis 2007).

Studies of the Singing Voice

As noted above, empirical evaluation of the singing voice dates back to Seashore and his colleagues. In 1936 Seashore measured fundamental frequency (F0) estimates from recordings and compared the intonation of these estimates to equal temperament. He also analyzed how these differences evolved over the duration of the note and the impact of note duration on these evolutions (Seashore, 1936). More recently, there has been a great deal of work done at the "Speech, Music, and Hearing" group at the Royal Institute of Technology in Stockholm. Sundberg examined variations in intonation between solo and choral performance, as well as the influence of certain vowels on tuning (Sundberg, 1987). He found a significant amount of variation in F0 across choirs, especially when vibrato is present. He also observed some variation in regards to 'sharpness' or 'flatness' of certain vowels, but general observable trends were limited. Prame (1997) studied vibrato extent and intonation in soprano singing. He found that vibrato excursions ranged from +/- 34 to +/-123 cents and that the intonation of notes deviated substantially, though not consistently, from equal temperament. A comprehensive survey of research into singing voice performance is available in Sundberg (1999).

The past few years have seen an increase in interest in the relationship between singing-voice performance parameters and musical structure. Howard (2007) examined pitch drift in an *a capella* SATB quartet. F0 estimates were calculated using the SPEAD software by Laryngograph Ltd, which measures the movement of the larynx. Timmers (2007) examined various performance parameters, including tempo dynamics and pitch variations, manually with PRAAT[1] for professional recordings of several Schubert songs whose recording dates spanned the last century. In relating these parameters to the musical structure of the piece, she found consistency across performers. She also explored the emotional characteristics of the performances and the ways in which performance style changed throughout the twentieth century. Ambrazevičius and Wiśniewska (2008) studied chromaticism and pitch inflection in traditional Lithuanian singing. They also used PRAAT for analysis and derived a number of rules to explain chromatic inflections for leading tones, and ascending and descending sequences. Rapoport (2008) manually analyzed the spectrograms of songs by Berlioz, Schubert, Puccini, and Offenbach, and classified each tone based on the relative strength of the harmonics in its timbre and the rate and depth of the vibrato. He then used this analysis to assess the similarities and differences between different singers' interpretations of the songs. Marinescu and Ramirez (2008) used spectral analysis on several monophonic excerpts from several arias performed by Josep Carreras to determine pitch, duration, and amplitude for each note. They also analyzed the sung lines with Narmour's implication-realization model (Narmour, 1990) and then combined this with a spectral analysis in order to induce classification rules using a decision tree algorithm.

Automatic extraction of singing performance data has been pursued in the music information retrieval community for its application in querying databases by singing or humming a tune, commonly known as 'query by humming' (Birmingham et al., 2001). 'Query by humming' tackles the larger problem of singing transcription, which can be divided into six separate sub-tasks (Weihs & Ligges, 2003): voice separation (for polyphonic recordings), note segmentation, pitch estimation, note estimation, quantization, and transcription (or notation). Of these sub-tasks only note segmentation is directly related to the extraction of performance data, of which there have been several different approaches: Clarisse et al. (2002) use an energy threshold to determine onsets by measuring the root-mean-square energy as a function of time; Wang et al.

[1] http://www.fon.hum.uva.nl/praat

(2003) use dynamic programming to determine the end points of notes; and Weihs and Leigges (2003) combine segmentation with pitch estimation and use pitch differentials to segment the notes. Ryynanen and Klapuri (2004) have developed a more holistic approach, where note events, such as pitch, voicing, phenomenal accents, and metrical accents, are modeled with an HMM (hidden Markov model), and note event transitions are modeled with a musicological model, which performs key estimation and determines the likelihood of two- and three-note sequences. These systems are evaluated in terms of the overall accuracy of their transcription, rather than the accuracy of the individual components. The use of quantization to create a MIDI-like transcription removes the performance data we are interested in examining.

Automatic Estimation of Note Onsets and Offsets in the Singing Voice

Note onsets and offsets are an important first stage in the extraction of performance data because they must delineate the temporal period in the signal where each note occurs. Note onset information is also a means unto itself when considering timing data. Currently, there are no robust automated methods for automatically estimating note onsets and offsets in the singing voice. Although much work has been done in the area of note onset detection (Bello et al., 2005), accurate detection of onsets for the singing voice and other instruments without percussive onsets is not a solved problem. Toh, Zhang, and Wang (2008) describe a system for automatic onset detection for solo singing voice that accurately predicts 85% of onsets to within 50ms of the annotated ground truth. This degree of accuracy makes this the state of the art, but it still is insufficient for our purposes. Also, this and other onset detection approaches do not address the issue of note offset detection. For music where a score is available, score-audio alignment techniques can be used to guide signal-processing algorithms for extracting performance data (Scheirer, 1998; Dixon, 2003; Goebl et al., 2008).

Existing Score-Audio Alignment Techniques

The challenge in using the score of a piece to guide the extraction of performance data is that performers do not play or sing with the strict rhythm of the notation. In order to serve as a reference, the temporal events in the score must be aligned with the temporal events in the audio file, a process for which numerous algorithms exist. This type of alignment is a difficult, but also largely solved problem for many of the applications that require it, these include score following (Cont et al., 2007), generation of ground truth for polyphonic transcription (Turetsky & Ellis, 2003), database search and retrieval (Pardo & Sanghi, 2005), and synchronization of MIDI and audio for digital libraries (Kurth et al., 2007). Hidden Markov models are typically used for online (or realtime) applications whereas the related, more constrained, technique of dynamic time warping has predominantly been used for offline applications. The online approaches often sacrifice precision for efficiency, low latency, and robustness against incorrect notes (Cont et al., 2007; Downie, 2008). While the offline approaches are more precise, they are still not sufficiently precise for the detailed analysis of performance required here. Moreover, alignment of singing voice recordings is particularly challenging because note onsets and offsets are often difficult to determine, particularly when notes change under a single syllable.

Dynamic time warping (DTW) allows for the alignment of similar sequences moving at different rates. It warps two sequences to match each other while minimizing the number of insertions and deletions necessary to align the sequences. When used to align audio with MIDI data, both sequences must be converted to sets of features. A range of features have been used in different DTW-based alignment systems: Orio and Schwartz (2001) used the structure of peaks in the spectrum from the audio against a sinusoidal realization of the MIDI file given by the harmonic sinusoidal partials; Hu, Dannenberg and Tzanetakis (2003) used chromagrams computed directly from the MIDI data; and Turetsky and Ellis (2003) used a short-time spectral analyses of frames in the audio and a sonified version of the MIDI. We performed a comparative evaluation of different features and found peak spectral difference (Orio & Schwarz, 2001) to be the most effective single feature for aligning recordings of the singing voice. Once the features have been calculated, they are compared with one another to generate a similarity matrix, as shown in Fig. 1. In the similarity matrix, black indicates maximum similarity while white indicates maximum dissimilarity, with shades of grey indicating intermediate steps. The best path through the similarity matrix is a warping from note events in the MIDI to their occurrences in the original. The black line in Fig. 1 represents the best path, which was calculated using a cost function that considers all possible paths through the similarity matrix (from the bottom left corner to the top right corner) and which penalizes for both distance and dissimilarity.

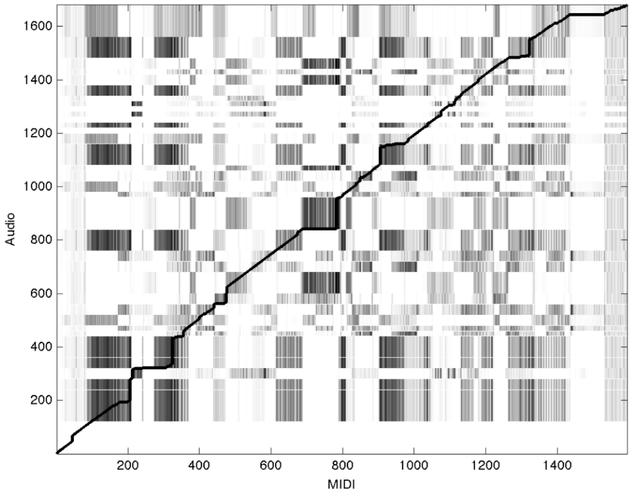

Figure 1: Similarity matrix between a set of feature extracted from an audio file and a MIDI file of the same piece. The y-axis is the number of audio frames and the x-axis is the number of MIDI frames. The black line indicates the optimal path through the similarity matrix, which is used to warp the timing in the audio and MIDI to match each other.

In an earlier work (Devaney and Ellis 2009), we evaluated the effectiveness the DTW approach for alignment of recordings of the singing voice. We used a hand-annotated forty-second excerpt of multi-tracked recordings of the Kyrie from Machaut's *Notre Dame Mass*. We found that only 31 percent of the onsets and 63 percent of the offsets were within 100 ms of the ground truth for alignment of the individual tracks. The onsets had a mean error of 171 ms (with a 146 ms standard deviation) while the offsets had a mean error of 147 ms (with a 331 ms standard deviation). While this error rate is sufficient for some applications, it is not precise enough for performance data analysis. Also, such implementations of DTW do not distinguish between transients and steady-state portions of the notes.

A hidden Markov model (HMM) is a statistical model of the temporal evolution of a process. The model is based on the assumption that the future can be predicted from current *state*, since it summarizes the past sequence of events. As a musical example, a simple HMM to determine whether a note is present or not might have only two states: note and rest. A slightly more involved HMM might have four states: attack, sustain, release, and rest. In order to model the temporal dynamics of a system, each state has a certain probability of transitioning to every other state, known as the *transition probability*. What is hidden in the HMM is the true state path, the *observations* of information from the model are stochastically related to the state, but the state itself is never observed directly. All we can observe is the singer's voice, we do not know, for example, whether the sound is in the attack state or the sustain state. HMMs have been extensively used in speech recognition (Rabiner, 1989) and score following (Cano, Loscos, & Bonada, 1999; Orio & Déchelle, 2001; Peeling, Cemgil, & Godsill, 2007, Raphael, 2004), where a system tracks a live performance in order to synchronize computerized accompaniment in real-time. They have also been used for singing transcription (Shih, Narayanan, & Kuo, 2003; Ryynanen & Klapuri, 2004), but, as noted above, these approaches make use of quantization in both the pitch and time domains to create a MIDI-like transcription, making them unusable for describing features of the performance rather than the score itself.

Improved Alignment Technique

We have developed an algorithm to improve the accuracy of an initial DTW alignment using a hidden Markov model (HMM) that models the acoustical properties of the singing voice. The HMM uses an initial DTW alignment as a prior to inform it of the notes' rough location.. This simplifies the problem that the HMM has to address, as the HMM is only responsible for local adjustments of the alignment. This approach of using an initial alignment to guide a secondary process is similar in this respect to the bootstrapping algorithm for onset detection described in Hu and Dannenberg (2006), where an initial DTW alignment is used to establish note boundaries that are in turn used to train a neural network for onset detection. The HMM was implemented in Matlab with Kevin Murphy's HMM Toolbox[2] and a full technical description its implementation can be found in Devaney, Mandel, and Ellis (2009).

There are certain acoustical properties of the singing voice that can be exploited to improve the DTW alignment. The amplitude envelope and periodic characteristics of a sung note are influenced by the words that are being sung. Transients occur when a consonant starts or ends a syllable, while vowels produce the steady-state portion of the note. The type of consonant affects the characteristics of the transient, as does the particular manner in which the singer attacks or enunciates the consonant. The motivation for identifying transients is to determine where the voiced section of the note begins for estimating a single fundamental frequency of the duration of the note.

The observations for the HMM are the pitch estimate and the square-roots of frame-based periodicity and power estimates from Alain de Cheveigné's YIN algorithm[3] (de Cheveigné and Kawahara, 2002). YIN is an autocorrelation-related fundamental frequency estimator, which we also use in our intonation investigates to provide frame-wise fundamental frequency estimates. The HMM has three basic states: silence, transient, and steady state, which were defined in terms of average periodicity and power estimates for each state. The general characteristics of each state can be observed in Fig. 1; the silence state has high aperiodicity and low power, the transient state has mid to low aperiodicity and low power, and the steady-state state has low aperiodicity and high power. The estimated pitch provides a somewhat noisy cue, especially for the silence and transient states, and the standard deviation used to model it was varied

[2] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

[3] http://audition.ens.fr/adc/sw/yin.zip

accordingly. It does, however, assist the alignment in cases where the note changes under the same vowel.

The probabilities of a state repeating instead of changing were calculated by observing the relative number of frames in each state in several recordings of Schubert's 'Ave Maria' and Machaut's *Notre Dame Mass* that were hand-labeled by the authors using Audacity, which is an audio editing software. The probabilities of a state changing were estimated by examining the corresponding scores for trends in note length and text-underlay. The transition probabilities to transient states reflect the likelihood of syllables beginning and ending with consonants in the Latin text. The transition probabilities to silences were based on the average frequency of rests in the scores. The transition probabilities to the steady-state state were based on the average length of notes.

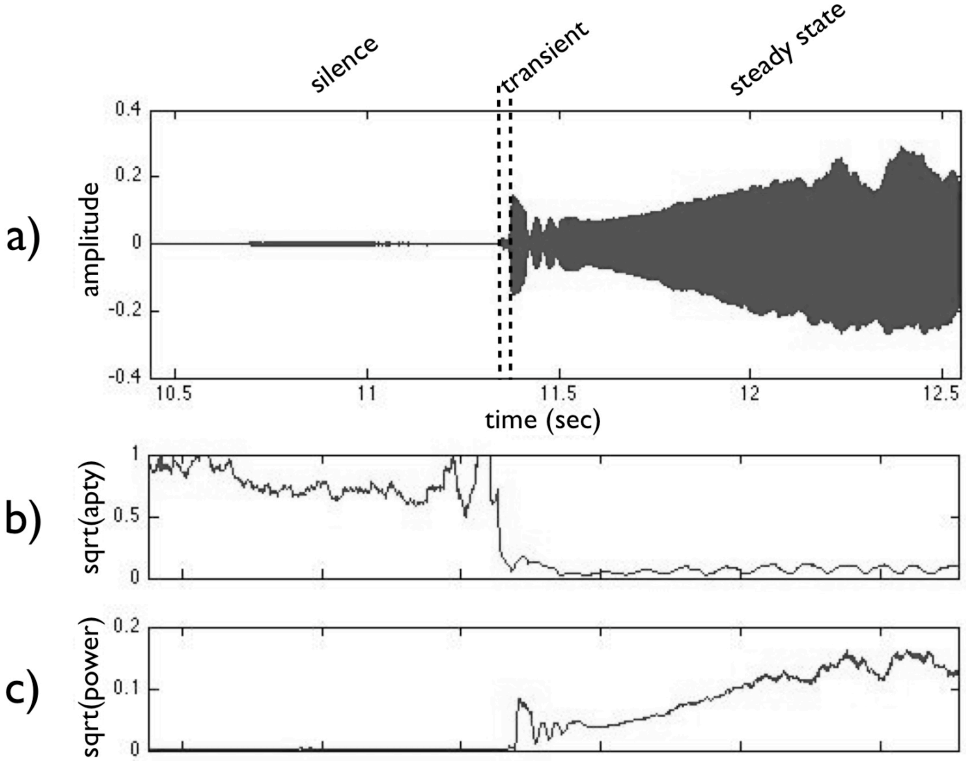

Figure 2: Visualization of the HMM states; (a) is time domain representation of a sung note with the HMM states labeled, (b) is the aperiodicity measure, and (c) is the power measure.

The general state sequence for the HMM representing a single note is shown in Fig. 3a. Here, for the purposes of capturing the proper temporal structure, a distinction is made between beginning- and ending-transients, though the acoustical properties of the two types of transients are modeled identically. In addition to self-looping, a steady-state state can be followed by an ending-transient state, a silence state or a beginning-transient state; an ending-transient state can be followed by a silence state, a beginning-transient state, or a steady-state state; a silence state can be followed by a beginning-transient state or a steady-state state; and a beginning-transient state can be followed only by a steady-state state. We then refined the state sequence to reflect the particular lyrics being sung; transients were only inserted when a consonant began or ended a syllable and silences were inserted only at the end of phrases. The state sequence for the opening phrase of Schubert's 'Ave Maria' is shown in Figure 3b.

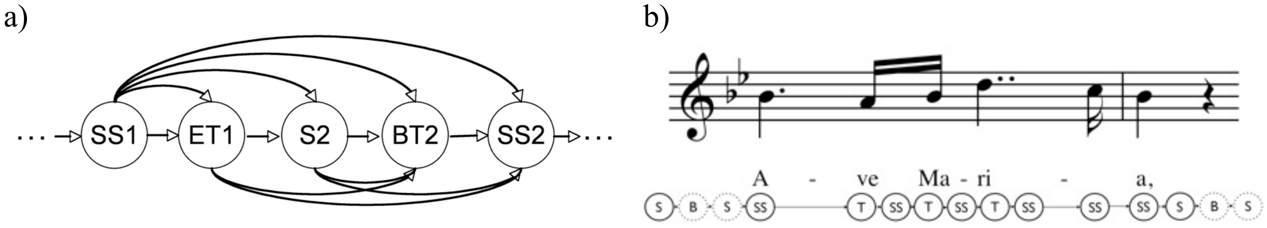

Figure 3: (a) State sequence seed: steady state (SS 1/2), ending transient (ET1), silence (S2), and beginning transient (BT2) and (b) state sequence adapted to sung text.

Three annotated recordings of the opening three phrases of Schubert's Ave Maria by three different singers were used to evaluate the system. The singers exhibited differences in overall timbre, attack time (transient length), and vibrato rates. Overall, the HMM was able to improve the results of the standard DTW alignment, decreasing the median alignment error from 52 to 42 ms. When a simple model of the phonetics of the lyrics was taken into consideration, as per Fig. 3b, the median error was further reduced to 28 ms.

A visual demonstration of the improvement in alignment can be seen in Fig. 9. Here the boxes indicate the DTW alignment, the dotted lines are the silences predicted by the model, the diamond shapes are the transients, and the solid lines are the steady-state portion of the notes. At approx. 400ms, 800ms, and 1500ms (labels 1, 2, and 3 respectively) the DTW alignment estimates the offsets too early and the onsets too late and at approx. 1800ms (label 4) the DTW estimates the offset too late. All of these misalignments are corrected by the HMM. Additionally, at location 1 and 3 the HMM successfully identifies the presence of the transients at the start of the notes.

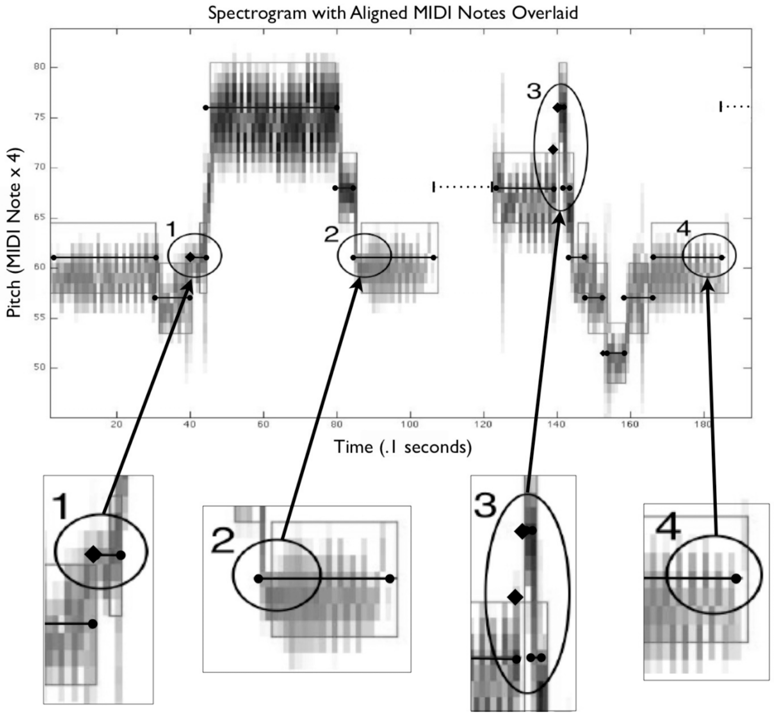

Figure 4: Visual demonstration of the improvements achieved with our HMM-based alignment method over the initial DTW alignment.

Future Work

The system we have described has been developed for use with trained singers, and would have difficulty with data from less professional performers. This is because the first stage alignment relies on the pitch of the recording corresponding reasonably closely to the notated pitches in the reference score. An amateur rendition, however, may not be sung in the correct key, and may include significant relative pitch errors, making this alignment unreliable. There are, however, applications in which automatic alignment of "naïve" performances would be valuable e.g., in the analysis of children's singing. We plan to develop techniques to encompass such domains in which there is no fixed pitch reference: although untrained singers may have unstable pitches and incorrect intervals, they will most likely preserve the basic contour of the melody (e.g., up vs. down pitch changes), as well as the word sequence. We anticipate that this tool will be of use to development psychological studies on singing, particularly the work being done on song acquisition and the developing of singing through the Advancing Interdisciplinary Research in Singing (AIRS) project.

Intonation Experiment on Schubert's 'Ave Maria'

We are interested in exploring whether there is a relationship between melodic interval tuning in solo singing and harmonic function. Following Prame's (1997) work, we used Schubert's 'Ave Maria' (see Fig. 5), as it allows for an exploration of commonalities of intonation tendencies in a well- known piece. Our assessment of the intonation includes both the mean of all of the frame-wise fundamental frequency estimates for each note as well as the evolution of the fundamental frequencies over the duration of the note. We also examined the consistency both within each performer's *a cappella* and accompanied takes and across performers. In this study, we compare the intonation of the semitone interval between the tonic and leading tone, in both directions, against the other semitones in the piece. This allows us to both examine the role of harmonic context in intonation and to assess a commonly held belief, rooted in Pythagorean tuning, that ascending leading-tones are sung sharp relative to descending leading tones or notes in other semitone intervals (Friberg et al., 2000).

Method

In the experiment, each participant performed three *a cappella* renditions of the first verse of the 'Ave Maria', followed by three renditions with recorded accompaniment. The six participants were undergraduate soprano vocal majors from McGill University. The participants had average age of 20.2 years, with a standard deviations of 2.13 years, an average of 14.7 years of sustained musical activity, with a standard deviation of 3.6 years, and an average of 6 years of singing instruction, with a standard deviation of 2.9 years.

The accompaniment was performed on a Bosendorfer SE piano, and subsequently transposed on the instrument to a range of keys when the acoustic recordings were made. This allowed the singers to perform the accompanied version in the key of their choice. The accompaniment was played back to the singers on headphones while they sang so that their singing could be recorded as an isolated monophonic line, which was necessary for signal processing purposes.

We performed all of the data analysis automatically, including the note onset and offset estimation as well as the intonation-related data. Onset and offset estimation was performed with the algorithm described above and frame-wise fundamental frequency estimates were performed using YIN (de Cheveigné & Kawahara, 2003). With the results of these analyses we were able to calculate a single fundamental frequency value for each note by taking the mean across the steady-state frame. We also calculated the evolution of the fundamental frequency from the first and second discrete cosine transform coefficient. The first coefficient approximates its slope and the second its curvature. The slope of the evolution of the fundamental frequency provides information about whether the singers are gliding up into the next note and the curvature indicated the amount that the fundamental frequency deviates from a steady pitch over the course of the note.

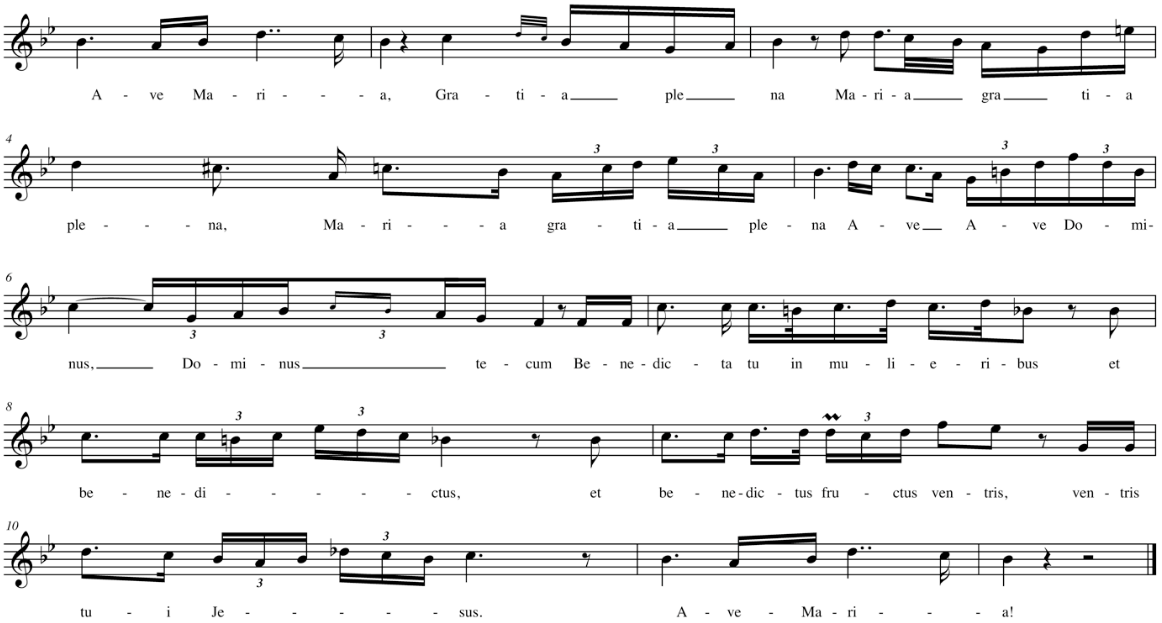
A - ve Ma - ri - - - - a, Gra - ti - a ple na Ma - ri - a gra ti - a
4
ple - - - - na, Ma - ri - - - - a gra - ti - a ple - na A - ve A - ve Do - mi-
6
nus, Do - mi - nus te - cum Be - ne - dic - ta tu in mu - li - e - ri - bus et
8
be - ne - di - - - - - - ctus, et be - ne - dic - tus fru - ctus ven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

Figure 5: Score for Schubert's 'Ave Maria'.

Results

The box and whisker plots in Fig. 6 and 7 show that there is a high degree of variability between the singers in terms of each note's single fundamental frequency for the various conditions: *a cappella* A-Bb intervals; *a cappella* Bb-A intervals; accompanied A-Bb intervals; accompanied Bb-A intervals; *a cappella* ascending other semitones; *a cappella* descending other semitones; accompanied ascending other semitones; and accompanied descending other semitones. There is also a variation amongst the singers in their degrees of self-consistency. The combination of all of the singers (Fig. 8) shows that median and 50% confidence interval for all of the conditions are relatively consistent. A linear regression analysis of the data revealed only weak effects of singer identity and accompaniment on each note's single fundamental frequency ($R^2 = 0.0680$, $p < 0.00001$). Singer 4 tended to be sharp of the mean across the group of singers, while singers 2, 3, and 6 tended to be flat. Our linear regression analyses did not show any effects for the leading tone or the direction of the interval, or a combination of the two (i.e., an ascending leading tone interval, A-Bb, versus a descending leading tone interval, Bb-A).

The slope (Fig. 8) and the curvature (Fig. 9) were consistently centered at 0, and had relatively consistent 50% confidences interval. They showed a high degree of variability in their 95% confidence intervals, however, with a large numbers of outliers for the curvature. The results of a linear regression on the slope values ($R^2 = 0.0332$, $p = 0.0006$) showed weak effects for direction and accompaniment, as well as for singers 4 and 6. The regression on the curvature values ($R^2 = 0.0433$, $p < 0.0000$) only showed effects for singers 3 and 4.

Discussion and Future Work

Overall, there were no observable effects in any of the regressions for the leading-tone intervals versus other semitones. This suggests that intonation tendencies for the semitone, either in terms of mean fundamental frequency or fundamental frequency evolutions, are not influenced by function. The median values across all of the singers for the note's single fundamental frequency in Fig. 8 indicate that there is a general tendency towards semitones that are smaller than the 5-limit Just Intonation semitones (the 16/15 major diatonic at 112 cents and the 17/16 minor diatonic at 105 cents). The range of interval sizes encompassed by the 50% confidence interval includes both the both the equal tempered semitone (100 cents) and the Pythagorean 256/243 diatonic semitone (90.2 cents). In the accompanied performances the semitone sizes were both closer to equal temperament and more consistent, as illustrated by the narrower 50% confidence interval.

This experiment was performed on semi-professional singers with a substantial amount of training. We do however recognize that as students our participants may not have the amount of intonation consistency as more experienced singers. In light of this, we plan to extend this study by recording a similarly sized group of professional singers under the same conditions.

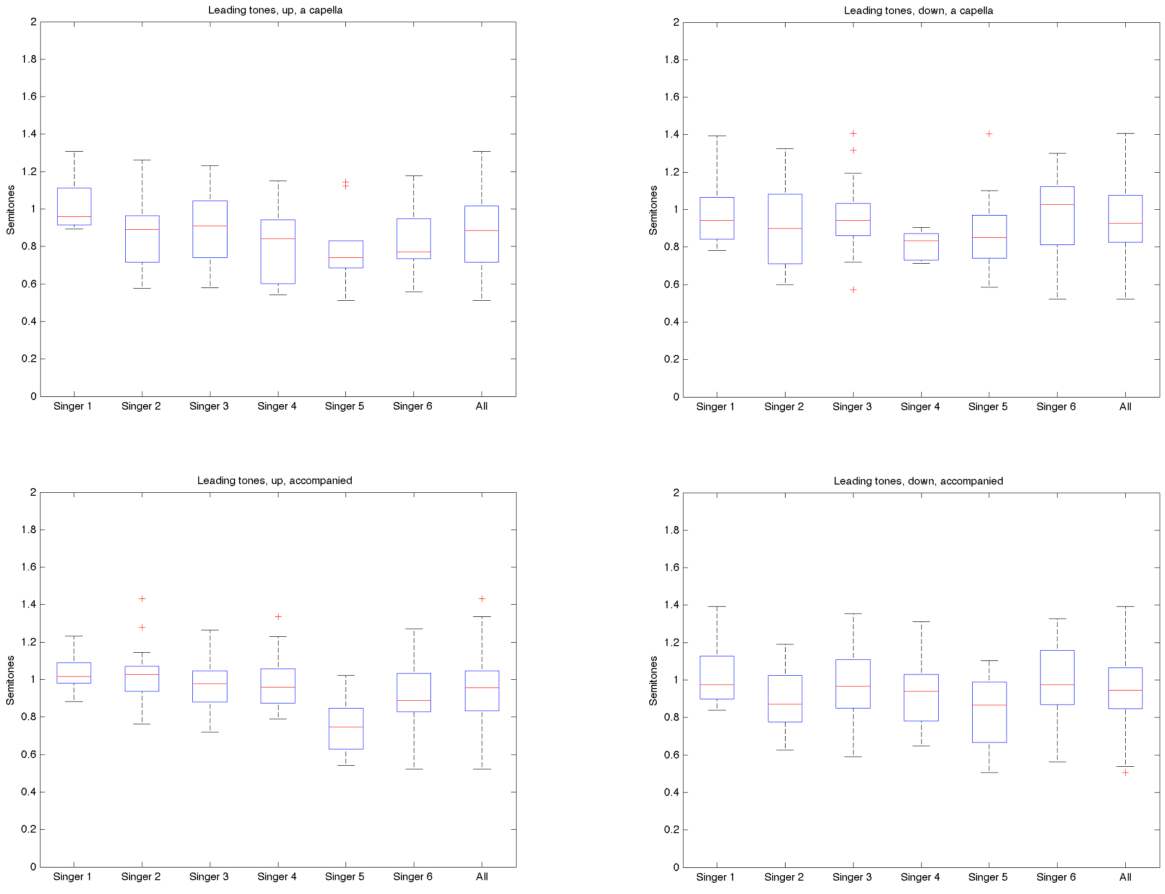

Figure 6: Box and whisker plots of the interval size in semitones for all of the A-Bb occurrences in Schubert's 'Ave Maria'. Each plot shows the results for the six singers individually and the combination of all of the singers.

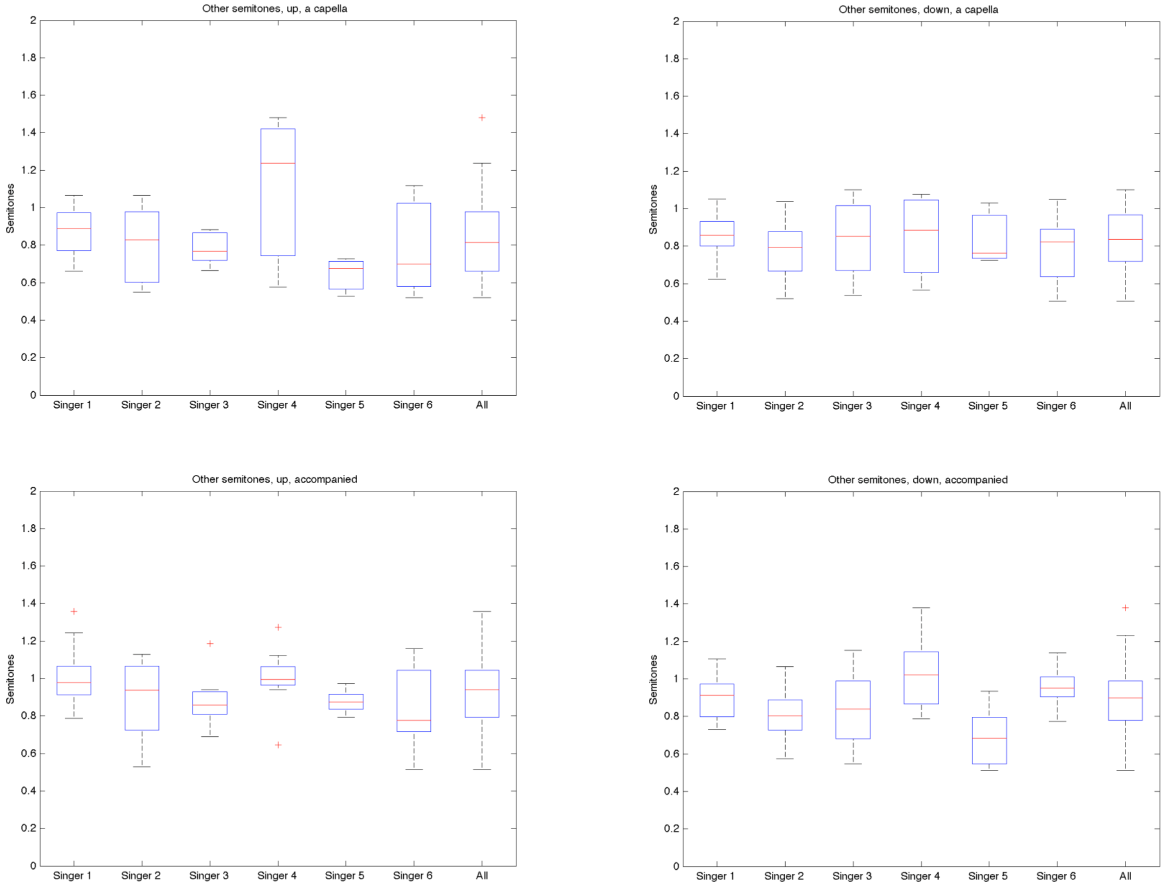

Figure 7: Box and whisker plots of the interval size in semitones for all of the non-Bb-A semitones occurrences in Schubert's 'Ave Maria'. Each plot shows the results for the six singers individually and the mean across all of the singers.

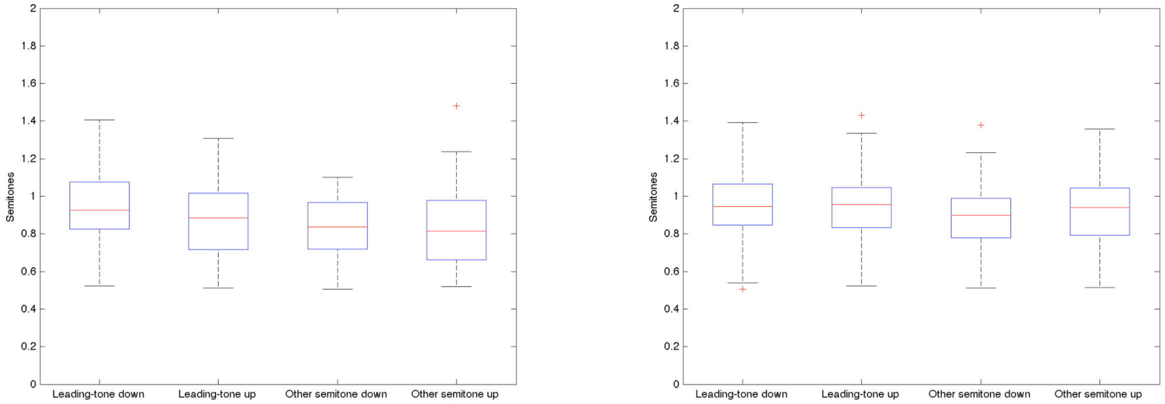

Figure 8: Box and whisker plots of the interval size in semitones for A-Bb and Bb-A intervals and other semitones across all singers. The plot on the left is *a cappella* and the plot on the right is with accompaniment.

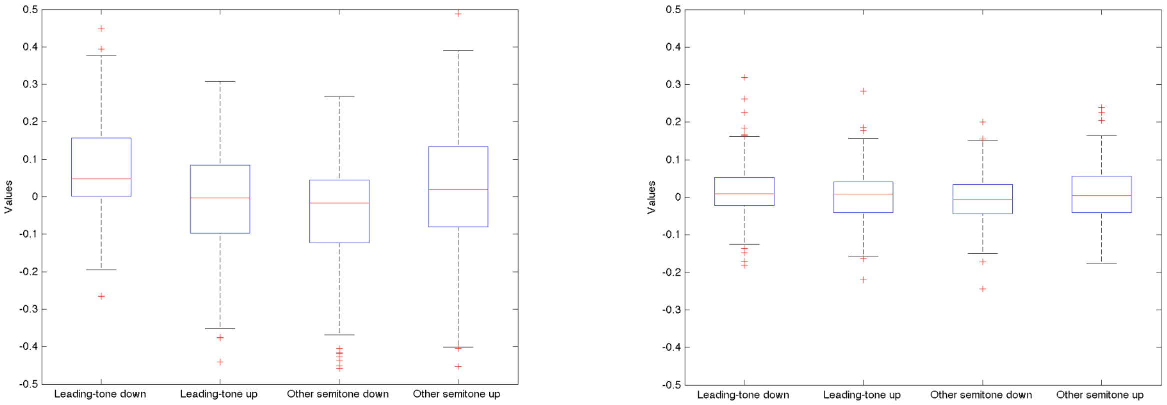

Figure 9: Box and whisker plots for the slope of the fundamental frequencies' evolutions for all of the A-Bb and Bb-A intervals and other semitones across all singers. The plot on the left is a cappella and the plot on the right is with accompaniment.

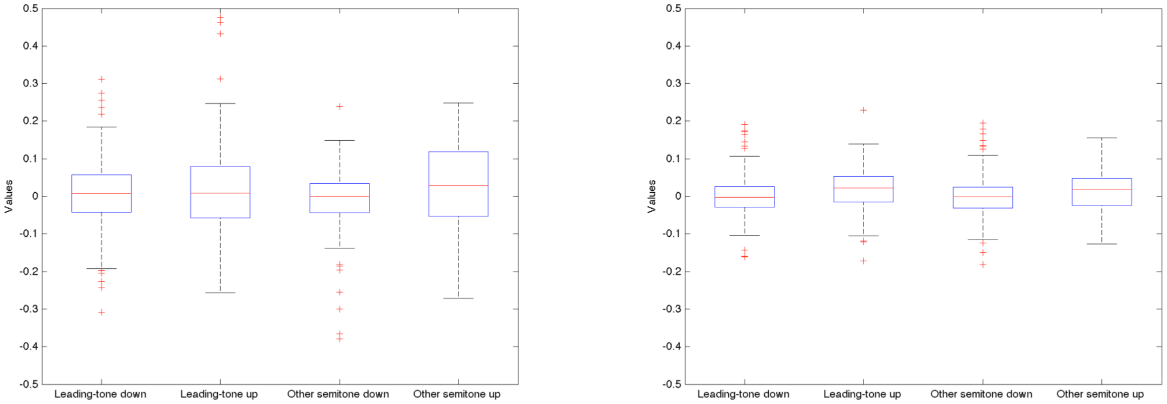

Figure 10: Box and whisker plots for the curvature of the fundamental frequencies' evolutions for all of the A-Bb and Bb-A intervals and other semitones across all singers. The plot on the left is *a cappella* and the plot on the right is with accompaniment.

Conclusions

In this paper, we presented an algorithm that automatically identifies pitch, onsets and offsets for recordings where a symbolic representation of the score is available. It is optimized for trained singing voices such that it can also correctly identify the transient and steady-state sections of the note. We also outlined our plans for extending this algorithm to work for untrained singers, making it robust enough to accommodate unstable pitches and incorrect intervals. We also described some results of a study of intonation that makes use of some of the described techniques for automatically extracting performance data. The study focused on solo soprano performances of Schubert's 'Ave Maria', and found that the size of the leading-tone/tonic interval was not different than other semitones performed in the piece. Specifically, all of the semitones tended to be equal to the size of an equal-tempered semitone or smaller, regardless of intervallic direction (up or down).

Appendix

We are planning to make this algorithm available to other researchers, both in its current form and in the expanded form discussed in the 'Future Works' section. The algorithm is written as a set MATLAB[4] functions and requires the Mathwork's Signal Processing Toolbox[5], as well several freely available toolkits, namely Kevin Murphy's HMM toolkit[2], Alain de Cheveigné's YIN implementation[3], and Tuomas Erola & Petri Toiviainen's MIDI Toolbox[6].

The tool requires:

a) An audio file of the recording

b) A symbolic representation of the score, this can be either in the form of a MIDI file or as a list of MIDI notes and note durations.

c) An annotation of the lyrics in the following format

  i. A list of each isolated syllable and silence

  ii. The number of notes corresponding to each syllable or silence

The algorithm returns the onset and offset times of the transients (where applicable) and steady-state portions of each of the notes defined in the symbolic representation.

Symbolic representation of the score

MIDI notes are in reference to Middle C = 60, and increase/decrease by 1 number per semitone, while rests =130. Note duration are defined in terms of the number of sixty-fourth notes, such that:

1 = 64th note
1.5 = dotted 64th note
2 = 32nd note
3 = dotted 32nd note
4 = 16th note
5 = dotted 16th note
8 = 8th note
12 = dotted 8th note
16 = quarter note
24 = dotted quarter note
32 = half note
48 = dotted half note
64 = whole note
96 = dotted whole note

Triplets are represented as third values, such that each note in a 16th note triplet will have a value of 2.67.

[4] http://www.mathworks.com/products/matlab/

[5] http://www.mathworks.com/products/signal/

[6] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/miditoolbox/

Example of symbolic notation using the 'Ave Maria' score in Fig. 5

Ave Maria

	Bb	*A*	*Bb*	*D*	*C*	*Bb*	*Rest*
MIDI Note:	70	69	70	74	72	70	130
Note Duration:	24	4	4	28	4	16	16

Gratia plena

	C	*D*	*C*	*Bb*	*A*	*G*	*A*	*Bb*	*Rest*
MIDI Note:	72	74	72	70	69	67	69	70	130
Note Duration:	16	1	1	4	4	4	4	16	8

Maria gratia plena

	D	*D*	*C*	*Bb*	*A*	*G*	*D*	*E*	*D*	*C#*
MIDI Note:	74	74	72	70	69	67	74	76	74	73
Note Duration:	8	12	2	2	4	4	4	4	16	12

Maria gratia plena

	A	*C*	*Bb*	*A*	*C*	*D*	*Eb*	*C*	*A*	*Bb*
MIDI Note:	69	72	70	69	72	74	75	72	69	70
Note Duration:	4	12	4	2.67	2.67	2.67	2.67	2.67	2.67	24

Ave, ave Dominus

	D	*C*	*C*	*A*	*G*	*B*	*D*	*F*	*D*	*B*	*C*
MIDI Note:	74	72	72	69	67	71	74	77	74	71	72
Note Duration:	4	4	12	4	2.67	2.67	2.67	2.67	2.67	2.67	18.67

Dominus tecum

	G	*A*	*Bb*	*C*	*Bb*	*A*	*G*	*F*	*Rest*
MIDI Note:	67	69	70	72	70	69	67	65	130
Note Duration:	2.67	2.67	4	1.33	1.33	2.67	2.67	16	8

Benedicta tu in mulieribus

	F	*F*	*C*	*C*	*C*	*B*	*C*	*D*	*C*	*D*	*Bb*	*Rest*
MIDI Note:	65	65	72	72	72	71	72	74	72	74	70	130
Note Duration:	4	4	12	4	6	2	6	2	6	2	8	8

Et benedictus

	Bb	*C*	*C*	*C*	*B*	*C*	*Eb*	*D*	*C*	*Bb*	*Rest*
MIDI Note:	70	72	72	72	71	72	75	74	72	70	130
Note Duration:	8	12	4	2.67	2.67	2.67	2.67	2.67	2.67	16	8

Et benedictus fructus ventris

	Bb	*C*	*C*	*D*	*D*	*D*	*C*	*D*	*F*	*Eb*	*Rest*
MIDI Note:	70	72	72	74	74	74	72	74	77	75	130
Note Duration:	8	12	4	6	2	2.67	2.67	2.67	8	8	8

Ventris tui Jesus

	G	*G*	*D*	*C*	*Bb*	*A*	*Bb*	*Db*	*C*	*B*	*C*	*Rest*
MIDI Note:	67	67	74	72	70	69	70	73	72	70	72	130
Note Duration:	4	4	12	4	2.67	2.67	2.67	2.67	2.67	2.67	24	8

Ave Maria

	Bb	*A*	*Bb*	*D*	*C*	*Bb*	*Rest*
MIDI Note:	70	69	70	74	72	70	130
Note Duration:	24	4	4	28	4	16	16

Example of lyric annotation using the 'Ave Maria' score in Fig. 5

Syllables	A	ve	Ma	ri	a		Gra	ti	a	ple	na	
Notes per syllable	1	1	1	2	1	0	1	2	2	2	1	0

Ma	ri	a	Gra	ti	a	ple	na		Ma	ri	a	Gra	ti	a	ple	na	
1	1	2	2	1	1	1	1	0	1	1	1	2	1	2	1	1	0

A	ve	a	ve	do	mi	nus		Do	mi	nus	te	cum	
2	2	2	1	2	1	1	0	1	1	4	1	1	0

Be	ne	dic	ta	tu	in	mu	li	e	ri	bus		Et	be	ne	dic	tus	
1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	6	1	0

Et	be	ne	dic	tus	fruc	tus	ven	tris		Ven	tris	tu	it	Je	sus	
1	1	1	1	1	2	1	2	1	0	1	1	2	1	6	1	0

A	ve	Ma	ri	a	
1	1	1	2	1	0

References

Ambrazevičius, R., & Wiśniewska, I. (2008). Chromaticisms or performance rules? Evidence from traditional singing. *Journal of Interdisciplinary Music Studies, 2*(1–2), 19–31.

Bengtsson, I., & Gabrielsson, A. (1980). Methods for analyzing performance of musical rhythm. *Scandinavian Journal of Psychology, 21,* 257–68.

Bengtsson, I., & Gabrielsson, A. (1983). Analysis and synthesis of musical rhythm. *Studies of Musical Performance*, 27–60.

Birmingham, W., Dannenberg, R., Wakefield, G., Bartisch, M., Bykowski, D., Mazzoni, D., et al. (2001). MUSART: Music retrieval via aural queries. *Proceedings of the International Conference on Music Information Retrieval,* 73–81.

Cano, P., Loscos, A., & Bonada, J. (1999). Score-performance matching using HMMs. *Proceedings of the International Computer Music Conference*, 441–4.

Clarisse, L., Martens, J., Lesaffre, M., Baets, B., Meyer, H., & Leman, M. (2002). An auditory model based transcriber of singing sequences. *Proceedings of the International Conference on Music Information Retrieval,* 116–23.

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

Cont, A., Schwarz, D., Schnell, N., & Raphael, C. (2007). Evaluation of real-time audio-to-score alignment. *Proceedings of the International Conference on Music Information Retrieval,* 315–6.

Dannenberg, R. & Raphael, C. (2006). Music score alignment and computer accompaniment. *Communications of the ACM, 49*(8), 38–43.

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

Devaney, J., & Ellis, D.P.W. (2009). Handling asynchrony in audio-score alignment. *Proceedings of the International Computer Music Conference,* 29–32.

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

Dixon, S. (2003). Score-based analysis of expressive performance. *Proceedings of the European Society for the Cognition Sciences for Music Conference,* 107–10.

Downie, J. (2008). MIREX 2008: Real-time audio to score alignment (a.k.a. score following). Retrieved from http://www.musicir.org/mirex/2008/index.php/Realtime_Audio_to_Score_Alignment_(a.k.a_Score_Following)

Earis, A. (2007). An algorithm to extract expressive timing and dynamics from piano recordings, *Musicae Scientiae, 11*(2)*,* 155–82.

Fyk, J. 1995. *Melodic Intonation, Psychoacoustics, and the Violin*. Zielona Góra: Organon.

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

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

Goebl, W., Dixon, S., De Poli, G., Friberg, A., Bresin, R., & Widmer, G. (2008). Sense in expressive music performance: Data acquisition, computational studies, and models. In P. Polotti & D. Rocchesso (Eds.) *Sound to Sense – Sense to Sound: A State of the Art in Sound and Music Computing* (pp. 195–242). Berlin: Logos Verlag.

Hong, J.-L. (2003). Investigating expressive timing and dynamics in recorded cello performances. *Psychology of Music, 31,* 340–52.

Howard, D. M. (2007). Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation. *Journal of Voice, 21*(3), 300–15.

Hu, N., R. Dannenberg, & G. Tzanetakis. (2003). Polyphonic audio matching and alignment for music retrieval. *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics,* 185–8.

Hu, N. & R. Dannenberg. (2006). Bootstrap leaning for accurate onset detection. *Machine Learning*, *65,* 457–71.

Kurth, F., Muller, M., Fremerey, C., Chang, Y., & Clausen, M. (2007). Automated synchronization of scanned sheet music with audio recordings. *Proceedings of the International Conference on Music Information Retrieval,* 261–6.

Marinescu, M.-C., & Ramirez, R. (2008). Expressive performance in the human tenor voice. *Proceedings of the Sound and Music Computing Conference.*

Narmour, E. (1990). *The Analysis and Cognition of Basic Musical Structures*. Chicago, IL: University of Chicago Press.

Orio, N., & Déchelle, F. (2001). Score following using spectral analysis and hidden Markov models. *Proceedings of the International Computer Music Conference*, 151–4.

Orio, N., & D. Schwarz. (2001). Alignment of monophonic and polyphonic music to a score. *Proceedings of the International Computer Music Conference,* 155–8.

Ornoy, E. (2008). An empirical study of intonation in performances of J.S. Bach's Sarabandes: Temperament, 'melodic charge' and 'melodic intonation'. *Orbis Musicae, 14,* 37–76.

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

Pardo, B., & Sanghi, M. (2005). Polyphonic musical sequence alignment for database search. *Proceedings of the International Conference on Music Information Retrieval,* 215–21.

Peeling, P., Cemgil, T., & Godsill, S. (2007). A probabilistic framework for matching music representations. *Proceedings of the International Computer Music Conference*, 267–72.

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

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77,* 257–289.

Raphael, C. (2004). A hybrid graphical model for aligning polyphonic audio with musical scores. *Proceedings of the International Conference on Music Information Retrieval,* 387–94.

Rapoport, E. (2008). The marvels of the human voice: Poem–melody–vocal performance. *Orbis Musicae, 14,* 7–36.

Repp, B. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America, 92(5),* 2546–68.

Repp, B. (1997). The aesthetic quality of a quantitatively average music performance: two preliminary experiments. *Music Perception, 14*(4), 419–44.

Ryynanen, M., & Klapuri, A. (2004). Modelling of note events for singing transcription. *Proceedings of IEEE Workshop on Statistical and Perceptual Audio Processing.*

Scheirer, E. (1998). Using musical knowledge to extract expressive performance information from audio recordings. In H. Okuno & D. Rosenthal (Eds.) *Readings in Computational Auditory Scene Analysis* (pp. 361–80). Mahwah, NJ: Lawrence Erlbaum.

Seashore, C. (1936). *Objective Analysis of Musical Performance*. Iowa City, IA: University of Iowa Press.

Seashore, C. (1938). *Psychology of Music.* New York, NY: Dover Publications.

Shih, H., Narayanan, S.S., & Kuo, C.-C.J. (2003). A statistical multidimensional humming transcription using phone level hidden Markov models for query by humming systems. *Proceedings of the IEEE International Conference on Multimedia, 1,* 61–4.

Sundberg, J. (1987). *The Science of the Singing Voice*. Dekalb, IL: Northern Illinois University Press.

Sundberg, J. (1999). The perception of singing. In D. Deutsch (ed.) *The Psychology of Music* (2nd ed.) (pp. 171–214). San Diego, CA: Academic Press.

Timmers, R. (2007). Vocal expression in recorded performances of Schubert songs. *Musica Scientiae, 11*(2), 237–68.

Todd, N. (1985). A model of expressive timing in tonal music. *Music Perception, 3*(1), 33–58.

Todd, N. (1989). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91*(6), 3540–50.

Toh, C.C., Zhang, B., & Wang, Y. (2008). Multiple-feature fusion based on onset detection for solo singing voice. *Proceedings of the International Conference on Music Information Retreival*, 515–20.

Turetsky, R., & Ellis, D. (2003). Ground–truth transcriptions of real music from force–aligned MIDI syntheses. *Proceedings of the International Conference on Music Information Retrieval,* 135–41.

Wang, Y., Kan, M.-Y., New, T.L., Shenoy, A. & Yin J. (2004). LyricAlly: Automatic synchronization of acoustic musical signals and textual lyrics. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(2), 338-49.

Weihs, C., & U. Ligges. (2003). Automatic transcription of singing performances. *Bulletin of the International Statistical Institute, 60,* 507–10.

Authors' Note

This work was supported by the Center for Research in Music Media and Technology (CIRMMT), the Fonds de recherche sur la société et la culture (FQRSC), the Social Sciences and Humanities Research Council of Canada (SSHRC), and the National Science Foundation under grant IIS-0713334. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.