Psychomusicology: Music, Mind & Brain
2011, Vol. 21, No. 1 & No. 2

DOI: 10.1037/h0094008

Automatically extracting performance data from recordings of trained singers

JOHANNA DEVANEY*
McGill University

MICHAEL I. MANDEL*
Université de Montréal

DANIEL P.W. ELLIS
Columbia University

ICHIRO FUJINAGA*
McGill University

**Center for Interdisciplinary Research in Music Media and Technology (CIRMMT)*

ABSTRACT—***Recorded music 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 our technique for automatically extracting information such as note timings, intonation, vibrato rates, and dynamics. An experiment is also presented that focuses on the tuning of semitones in solo soprano performances of Schubert's "Ave Maria" by non-professional and professional singers. We found a small decrease in size of intervals with a leading tone function only in the non-professional group.***

KEYWORDS—***vocal intonation, performance analysis, audio annotation***

This paper describes the challenges that arise when attempting to automatically extract pitch-related performance data from recordings of the singing voice. The first section of the paper provides an overview of the history of analyzing recorded performances. The second section describes an algorithm for automatically extracting performance data from recordings of the singing voice where a score of the performance is available. The algorithm first identifies note onsets and offsets. Once the onsets and offsets have been determined, intonation, vibrato, and dynamic characteristics can be calculated for each note.

The main experiment of the paper, described in the third section, is 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 non-professional and six professional sopranos 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 ascending and descending semitones. We found that the A-B♭ intervals with a leading tone function were on average 8 cents smaller than the non-leading tone A-B♭, but that their average size was approximately the same as the other semitones performed in the piece, regardless of intervallic direction or accompaniment.

PREVIOUS WORK ON THE ANALYSIS OF RECORDED PERFORMANCES

Interest in studying recorded performances dates back almost as far as the birth of recordable media, beginning with Dayton Miller's (1916) work on visualized pitch information in recordings with phonophotographic apparati. The psychologist Carl Seashore and colleagues at the University of Iowa also undertook extensive work in performance

Johanna Devaney, Schulich School of Music, McGill University; Michael I. Mandel, Department of Computer Science and Operations Research, Université de Montréal; Daniel P.W. Ellis, Department of Electrical Engineering, Columbia University; Ichiro Fujinaga, Schulich School of Music, McGill University.

Correspondence concerning this article should be addressed to Johanna Devaney, Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montréal, QC, Canada, H3A 1E3. E-mail: johanna.devaney@mail.mcgill.ca

analysis (Seashore, 1938) employing 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 team also undertook numerous studies of singing. Schoen (1922) studied five performances of Gounod's setting of the "Ave Maria." He found that tuning depended on the direction of the line: notes following a lower tone tended to be flatter whereas notes followed by a higher note tended to be sharper. In general the singers were sharper than either equal temperament or just intonation. Easley's study of vibrato in opera singers found that the rate of the singer's vibrato was faster and the depth broader in songs from opera as compared to concert songs (Easley, 1932). Bartholomew (1934) studied vibrato along with other acoustic features of the singing voice in an attempt to define "good" singing. He observed the vibrato to be sinusoidal in nature and its rate to be approximately 6–7 Hz. H. G. Seashore (1936) also looked at Gounod's setting of the "Ave Maria," as well as Handel's *Messiah*. He studied nine performances and focused on the connections, or glides, between notes. He was able to correlate glide extent with direction, finding that glide extent was larger going up than going down. Miller (1936) provided a large amount of detail through "performance scores," though a lot of the data was not analyzed. His study of vibrato found that the rate of vibrato fell between 5.9–6.7 Hz, and the extent was 44–61 cents (100ths of a semitone), with faster vibrato in shorter tones. This echoed Tiffin's earlier findings that the average rate of vibrato is 6.5 Hz and the average depth is 60 cents (Tiffin, 1932). Similarly, Metfessel found that the range of vibrato rate was 5.5–8.5 Hz, with an average of 7 Hz, and the extent of the vibrato was 10–100 cents, with an average of 50 cents (Metfessel, 1932). Miller's observations about intonation also confirmed earlier findings of deviation from equal temperament or just intonation, and the different characteristics of the gliding transitions between notes. He also detailed dynamics and timing in the performances (Miller, 1936).

Though relatively accurate performance data could be assessed with these methods, they were extremely labor intensive. This limited the number of pieces that could be evaluated. Interest in empirical performance analysis subsequently diminished, in part due to its laboriousness, continuing mainly in the area of ethnomusicology (e.g., Seeger, 1951; Tove, Norman, Isaksson, & Czekajewski, 1966). The resurgence of a more general interest in music performance studies in the late 1970's coincided with both a movement by musicologists away from equating scores with music and an increased interest in music by cognitive psychologists. 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 (Clarke, 1989). In the early 1990s, Repp (1992) performed extensive evaluations of timing in the piano music of Beethoven and Schumann. He found that the degree of *ritardando* in musical phrases 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. A comprehensive survey of research on musical performance up to 2002 can be found in published reviews by Palmer (1997) and Gabrielsson (1999, 2003), and a discussion of the history of performance analysis in musicology is available in Cooper and Sapiro (2006). A discussion of contemporary research on performance analysis of the singing voice is provided below.

Extraction of Performance Data

Historically, the piano has been the primary instrument of performance analysis for several reasons. The large amount of solo repertoire available 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. The piano's percussive nature also 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 Music Instrument Digital Interface (MIDI) technology. MIDI is an industry standard protocol that 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 record the hammer action in either MIDI or a proprietary format. Examples of such pianos are Yamaha's Disklavier and Bösendorfer's Special Edition. For instruments other than the piano, the precision of the mapping between the physical instruments' motions and MIDI is severely limited. The main limitation of this approach is that only performances recorded on specialized instruments can be studied.

The extraction of performance data directly from recordings enables the study of a wider range of instruments and existing performances. The accurate extraction of such data, however, is still an open problem. This is particularly true for the singing voice and instruments with non-percussive onsets and flexible intonation capabilities. Since the mid-1990s there has been an increase in studies on these types of instruments, particularly the violin (Fyk, 1995; Ornoy, 2007) and cello (Hong, 2003). These studies 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 (2007) uses a "manual beat tapping system" for synchronization that is corrected by both a computer-aided system and human intervention.

Studies of the Singing Voice

As noted above, empirical evaluation of the singing voice dates back to the early part of the twentieth century. More recently, a great deal of work has been conducted at the "Speech, Music, and Hearing" group at the KTH Royal Institute of Technology in Stockholm. Hagerman and Sundberg (1980) examined the impact of vowels on intonation accuracy in professional barbershop quartets. They found a high degree of precision in the ensembles studied with limited influence from the type of vowel sung. Sundberg (1982) observed deviations from Pythagorean and pure tuning in singing with vibrato, and he concluded that the presence of vibrato allowed the singers to use a greater range of tunings than singers singing in barbershop style because of the presence of beats. Gramming, Sundberg, Ternström, Leanderson, and Perkins (1987) looked at the relationship between pitch and accuracy in the singing voice in three different populations: professional singers, non-singers, and singers with some form of vocal dysfunction. They did not find any significant differences between the groups. Sundberg also 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 F_0 (fundamental frequency), especially in the presence of vibrato. He also observed some variation in regards to "sharpness" or "flatness" of certain vowels, but general observable trends were limited. Ternström and Sundberg (1988) examined the impact of sound pressure level and spectral properties on choral intonation. They played reference tones for singers individually and found that intonation precision of the singers' response tones was negatively impacted by increased sound pressure levels and by increased amounts of vibrato for simple spectra. Carlsson-Berndtsson and Sundberg (1991) showed that when singers tuned the two lowest formants, in order to project their voices, there was not a discernible impact on vowel perception. Sundberg (1994) also examined the role of vibrato, detailing its acoustics and psychoacoustic features in a thorough review of vocal vibrato research.

Prame (1994, 1997) studied vibrato rate in performances of Schubert's "Ave Maria" by 10 professional sopranos. The fundamental frequency estimates were obtained using a sonogram. The analysis was restricted to the 25 longest notes because only these notes had enough vibrato cycles to measure the vibrato rate accurately. He found that the mean rate was 6 Hz across the 25 notes for each of the

10 singers, and that the rate of the vibrato tended to increase about 15% at the end of the notes. Sundberg, Prame, and Iwarsson (1995) used the same recordings to study both expert listeners' perceptions of whether the 25 tones were in tune and professional singers' ability to replicate pitches in the opening and closing of the first verse. They did not find much agreement among the expert listeners as to which notes were in tune and which ones were not. The singer's ability to replicate the tones was accomplished by comparing the deviation from equal temperament of the mean frequency of each corresponding note in the openings and closings. They found that when the repeated tones were within 7 cents of each other, the expert listeners agreed that they were in tune. Prame (1997) also used these recordings to study vibrato extent and intonation. He found that vibrato excursions ranged from ±34 to ±123 cents and that tones with larger vibrato extent tended to be sharper. The intonation of notes deviated substantially, though not consistently, from equal temperament. Prame also calculated each singer's mean deviation from the accompaniment, and found that the range of these means was 12–20 cents. Jers and Ternström (2005) examined choral intonation in two multi-track recordings of an eight-measure piece by Praetorius, one at a slow tempo and one at faster tempo. They examined the average values across both the whole ensemble and the mean and standard deviation of F_0 for each note produced by the individual singers and found that the amount of scatter was greater at the faster tempo than at the lower tempo. A survey of other research into singing voice performance by the "Speech, Music, and Hearing" group is available in Sundberg (1999) and Ternström (2003).

The past few years have seen an increase in interest in the relationship between singing-voice performance parameters and musical structure. Howard (2007a, 2007b) examined pitch drift and adherence to equal temperament in two *a cappella* SATB quartets. F_0 estimates were calculated by measuring the movement of the larynx with an electroglottograph and SPEAD software by Laryngograph Ltd. Timmers (2007) examined various performance parameters, including tempo, dynamics, and pitch variations, manually with PRAAT software (Boersma & Weenink, n.d.) 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 of performance parameters across performers. She also explored the emotional characteristics of the performances and the ways in which performance style changed throughout the twentieth century, including an increase in vibrato extent and a decrease in vibrato rate. 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 (2007) 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 spectrum 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 José Carreras to determine pitch, duration, and amplitude for each note. They also analyzed the sung lines with Narmour's (1990) implication-realization model and then combined this with a spectral analysis in order to induce classification rules using a decision tree algorithm.

The music information retrieval community has conducted research into the automatic extraction of singing performance data for the purposes of querying databases by singing or humming a tune, an area of research commonly known as "query by humming" (Birmingham et al., 2001). Query by humming research tackles the larger problem of singing transcription, which was divided into six separate sub-tasks by Weihs and 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 and pitch estimation are directly related to the extraction of performance data. There have been several approaches to this: Clarisse et al. (2002) used an energy threshold to determine onsets by measuring the root-mean-square energy as a function of time; Wang, Kan, New, Shenoy, and Yin (2004) used dynamic programming to determine the end points of notes; and Weihs and Ligges combined segmentation with pitch estimation and use pitch differentials to segment the notes. Ryynänen and Klapuri (2004, 2008) developed a more holistic approach, where note events, such as pitch, voicing, phenomenal accents, and metrical accents, are modeled with a hidden Markov model (HMM), and note event transitions are modeled with a musicological model, which performs key estimation and determines the likelihood of two- and three-note sequences. More recent papers addressing this problem include Dannenberg et al. (2007) and Unal, Chew, Georgiou, and Narayanan (2008). 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 an equal-tempered MIDI-like transcription removes the performance data we are interested in examining.

AUTOMATIC ESTIMATION OF NOTE ONSETS AND OFFSETS IN THE SINGING VOICE

Identifying note onsets and offsets is an important first stage in the extraction of performance data because they delineate the temporal period in the signal where each note occurs. Note onset information is also useful as timing data. Currently, there are no robust automated methods for estimating note onsets and offsets in the singing voice. Although much work has been conducted 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. Friberg, Schoonderwaldt, and Juslin (2007) developed an onset and offset detection algorithm that was evaluated on electric guitar, piano, flute, violin, and saxophone. On human performances they reported an onset estimation accuracy of 16 ms and an offset estimation accuracy of 146 ms. Toh, Zhang, and Wang (2008) describe a system for automatic onset detection for solo singing voice that accurately predicts 85% of onsets to within 50 ms of the annotated ground truth (i.e., the manually annotated values for the test data). This degree of accuracy makes this the state of the art, but it still is insufficient for our purposes since we need to be able to determine onsets and offsets reliably within at least 30 ms.

For music where a score is available, score-audio alignment techniques can be used to guide signal-processing algorithms for extracting performance data (Dixon, 2003; Goebl et al., 2008; Scheirer, 1998). The challenge in using a musical score to guide the extraction of performance data is that performers do not play or sing with the strict rhythm or pitch 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. Score-audio alignment is considered a nearly solved problem for many application, including score following (Cont, Schwarz, Schnell, & Raphael 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, Müller, Fremerey, Chang, & Clausen, 2007). In existing approaches, HMMs have typically been used for online, or realtime, applications whereas the related, more constrained, technique of dynamic time warping (DTW) has predominantly been used for offline applications; see Appendix A for more details. The online approaches often sacrifice precision for efficiency, low latency, and robustness against incorrect notes (Cont et al., 2007; Downie, 2008). Existing offline approaches are more precise, but they are still not sufficiently precise for detailed analyses of performances. Moreover, alignment of singing voice recordings is particularly challenging for a number of reasons. These include the difficulty of determining note onsets and offsets when notes change under a single syllable (melismas), the differences in onset characteristics between vowels and consonants, and acoustic characteristics that accompany different types of attacks and articulations.

In an earlier work (Devaney & Ellis, 2009), we evaluated the effectiveness of the DTW approach for the alignment of recordings of the singing voice. We used a hand-annotated 40 s 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 identified by the algorithm were within 100 ms of the ground truth for alignment of the individual tracks. The onsets had a mean error of 171 ms (*SD* = 146 ms) while the offsets had a mean error of 147 ms (*SD* = 331 ms). This error rate is insufficient for performance data analysis. For example, timing-focused expressive performance studies require precision on the order of 7–50 ms, if we hope to be able to capture the temporal performance asynchronies between different lines in a music texture (Palmer, 1997). A second issue with existing DTW alignment algorithms is that they do not distinguish between transients (attacks) and steady-state portions of the notes.

An approach for improving MIDI-audio alignment

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 the initial DTW alignment for rough estimates of the locations of the notes. 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 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 (Murphy, 2005). A technical description of its implementation can be found in Appendix B of this paper and more details are available in Devaney, Mandel, and Ellis (2009).

There are certain acoustical properties of the singing voice that we exploited to improve the DTW alignment. Specifically, 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 over 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 (de Cheveigné, 2002; de Cheveigné & Kawahara, 2002). We also use YIN in our intonation investigations (described below) to provide framewise F_0 estimates. YIN is an autocorrelation-based fundamental frequency estimator that was developed originally for speech. Evaluation on speech data showed that 99% of YIN's F_0 estimates were accurate to within 20% of the correct F_0, 94% were accurate to within 5%, and approximately 60% were accurate to within 1% (de Cheveigné & Kawahara, 2002). In the same evaluation it was shown to be robust in terms of minimizing gross error (errors off by more than 20%) than other commonly-used F_0 estimation techniques, including the F_0 estimator in PRAAT. The algorithm was also evaluated on the singing voice by de Cheveigné and Henrich (2002).

Evaluation

Our algorithm was evaluated with the opening three phrases of Schubert's "Ave Maria" by three different singers. The singers exhibited differences in overall timbre, attack time (transient duration), and vibrato rates. Overall, our simple HMM model 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 discussed in Appendix A, the median error was further reduced to 28 ms.

A visual demonstration of the improvement in alignment can be seen in Figure 1. At approximately 400 ms, 800 ms, and 1500 ms (labels 1, 2, and 3, respectively), the DTW alignment estimates the offsets too early and the onsets too late, and at approximately 1800 ms (label 4), the DTW estimates the offset too late. All of these misalignments are

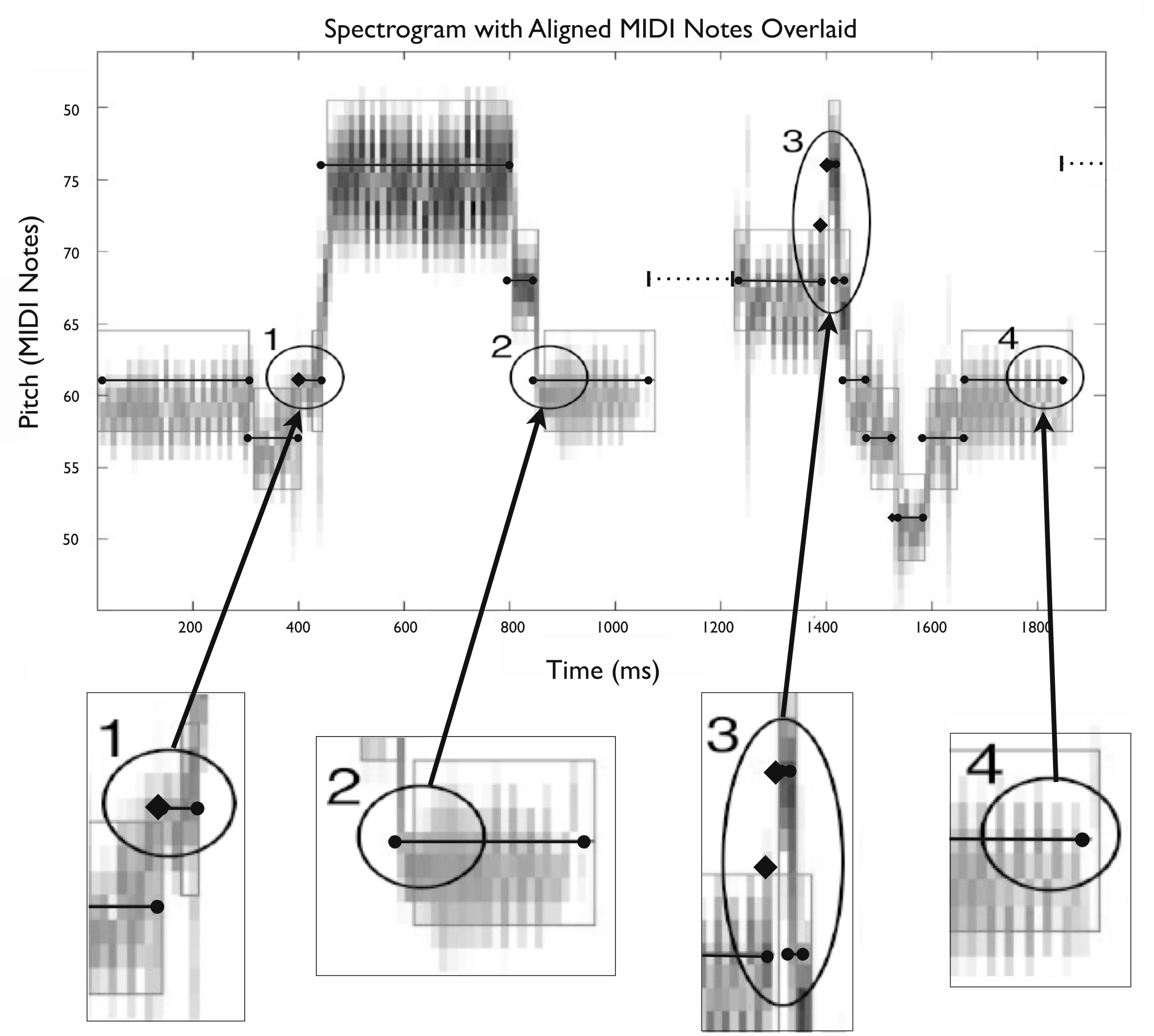


Figure 1. Visual demonstration of the improvements achieved with our HMM-based alignment method over the initial DTW alignment. The opening passage of a recording of the "Ave Maria" is represented as a zoomed-in log-frequency spectrogram. The boxes indicate the note position estimated generated by the initial DTW alignment. The HMM estimates for silence are represented by dotted lines, the estimates for transients are represented by diamond shapes, and the estimates for the steady state portions of the notes are represented by solid lines.

corrected by the HMM. Additionally, at locations 1 and 3, the HMM successfully identifies the presence of the transients at the start of the notes.

We have observed problems that arise with the algorithm in two different conditions: the presence of wide vibrato when annotating neighbor tone sequences and large amounts of reverberation in the recordings. While the algorithm is robust to both various depths of vibratos and normal degrees of compression and expansion of interval sizes, it is sensitive to situations where the singer has a wide vibrato and semitone neighbors that are compressed, as it is difficult to distinguish where the various notes begin and end. The algorithm faces a similar challenge in recordings with excessive reverberation as the note endings, and sometimes the beginnings of the subsequent notes, become ambiguous. In order to indicate to the user how the algorithm is performing, visualizations of the alignment overlaid on a spectrographic representation of the recordings are generated. Manual annotation may be incorporated to correct for any errors made by the algorithm.

Discussion

The system we have described has been developed for use with recordings by trained singers, and it would have difficulty with recordings by less professional performers. This is because the first-stage alignment relies on the sung pitch corresponding reasonably closely to the notated pitches in the reference score. An amateur rendition, however, may include significant relative pitch errors, making this alignment unreliable. There are applications in which automatic alignment of such "naïve" performances would be valuable, for example, in the analysis of children's singing. We are working on extending the algorithm to work in domains where there is no fixed pitch reference. We are working with the assumption that although untrained singers may have unstable pitches and incorrect intervals, they will most likely preserve the basic contour of the melody (e.g., ascending vs. descending pitch changes), as well as the word sequence. We anticipate that this tool will be of use in developmental studies of singing, for example, the work being done on the acquisition of song in development through the Advancing Interdisciplinary Research in Singing (AIRS) project. A MATLAB implementation of the current version of the algorithm is available (Devaney, 2011), and usage instructions are available in Appendix C.

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. 2) because it allows for an exploration of commonalities of intonation tendencies in a well-known piece. Unlike Prame, we did not need to restrict our analysis to the 25 longest notes in the piece, as we were able to obtain reliable onset, offset, and pitch estimates for all of the non-ornamental notes in the piece. However, due to the instability of the pitch in the performance of shorter notes by the singers, only those semitone intervals between notes with a duration greater than a 32nd note were examined. Our assessment of the intonation uses the mean of the frame-wise fundamental frequency estimates for each of these semitones. We also examined the consistency both within each performer's *a cappella* and accompanied renditions and across performers. In this study, we compare the intonation of the semitone interval between the tonic and leading tone, in both directions, with the other semitones in the piece. This allows us to 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. The sharpening of the leading tone leads to a compression of the interval, as discussed in Friberg, Bresin, and Sundberg (2006).

Method

Participants

The first group of participants consisted of six undergraduate soprano vocal majors from McGill University who had completed an average of 2 years (*SD* = 0.6) of full-time course work in the Bachelor of Music degree program. The participants had a mean age of 20.2 years (*SD* = 2.1), and an average of 14.7 years (*SD* = 3.6) of sustained musical activity, with an average of 6 years (*SD* = 2.9) of private voice lessons. They had engaged in daily practice for an average of 5.2 years (*SD* = 3.2), with a current daily practice time average of 1.1 hours (*SD* = .7). The second group consisted of six singers with graduate-level training, who worked professionally in the Montreal area. Their ages ranged from 28 to 58, with a mean of 35.7 years (*SD* = 11.5). They had an average of 26.0 years (*SD* = 8.7) of sustained musical activity, with an average of 10.3 years (*SD* = 6.0) of private voice lessons. They had engaged in daily practice for an average of 16.7 years (*SD* = 11.9), with a current daily practice time average of 1.5 hours (*SD* = 0.5).

Apparatus

The singers were recorded on AKG C 414 B-XLS microphones in a 4.85m x 4.50m x 3.30m lab at the Center for Interdisciplinary Research in Music Media and Technology. The lab had low noise, reflections, and reverberation time (ITU-standard). The microphones were run through a RME Micstasy 8 channel microphone preamplifier and RME Madi Bridge into a Mac Pro computer for recording.

Procedure

In the experiment, each of the six participants performed three *a cappella* renditions of the first verse of the "Ave Maria," followed by three renditions with recorded accompaniment. The performers were asked to produce a neutral performance with minimal vibrato. The accompaniment was performed on a Bösendorfer SE piano, and subsequently transposed on the instrument to a range of keys. This allowed the singers to perform the accompanied version in the key of their choice. The accompaniment was played back to the singers on Sennheiser HD 280 Pro closed headphones while they sang so that their singing could be recorded as an isolated monophonic line, which was necessary for accurate signal processing. The singers only wore the headphones over one ear to allow them to hear themselves.

Results

We performed all of the data analysis automatically, including the note onset and offset estimation as well as the intonation-related data. The time it takes to manually annotate the steady state and (where applicable) transient portions of each note is about 10–12 times real-time (i.e., 10–12 times the length of the audio file). The algorithm requires some amount of manual intervention before it can be run. Specifically, if a MIDI file is not available one must be created or the data must be manually encoded (see Appendix C). This takes approximately 3 times the duration of the audio. The lyrics must also be encoded, which takes about 2 times the duration of the audio. This only has to be done once for each piece; for this experiment a single set-up was needed for all of the 36 recordings evaluated. The alignment algorithm itself currently runs quite slowly, but it does not require any manual intervention while it is running. Using a single core on a Quad-Core Mac Pro with 2.26 GHz processors and 8 GB of RAM, the algorithm runs at about 15 times real-time. Once the algorithm has completed, the user can visually examine the alignment, which runs at about real-time. The amount of time needed for error correction depends on the number of errors present, at the rate of about 10 times the length of each note that needs correcting. Overall the algorithm is not faster in absolute time, but requires far less manual intervention: 5 times real-time for each score plus any necessary error correction compared to 10–12 times real-time to annotate each audio recording manually.

After the onset and offset estimation, frame-wise (frame size = 33.4 ms, hop size = 0.68 ms) F_0 estimates were made with YIN (de Cheveigné & Kawahara, 2002). From these frame-wise F_0 estimates we calculated a single perceived pitch for each note. Following Gockel, Moore, and Carlyon (2001), this research uses a weighted mean based on the F_0's rate of change. This mean is calculated by assigning a higher weighting to the frames where the F_0 has a slower rate of change than those with a faster rate of change. The threshold between fast and slow rates of change is 1.41 octaves per second, based on results reported by Prame (1994, 1997) that professional singers have an average vibrato rate of 6 Hz and a depth of +/- 71 cents.

We first examined the degree of variability between the singers in terms of semitone interval size, using the perceived pitch for each note, for several conditions (see Figure 2): A-B♭ leading tones (36 intervals per group = 2 instances in each rendition x 6 singers x 3 *a cappella* renditions); other A-B♭ intervals (72 intervals per group); B♭-A intervals (72 intervals per group); ascending other semitones (36 intervals per group); descending other semitones (90 intervals per group). There were the same number of conditions per group for the accompanied renditions, resulting in a total of 72 leading tones, 144 A-B♭ intervals, 144 B♭-A intervals, 72 ascending other semitones, and 180 descending

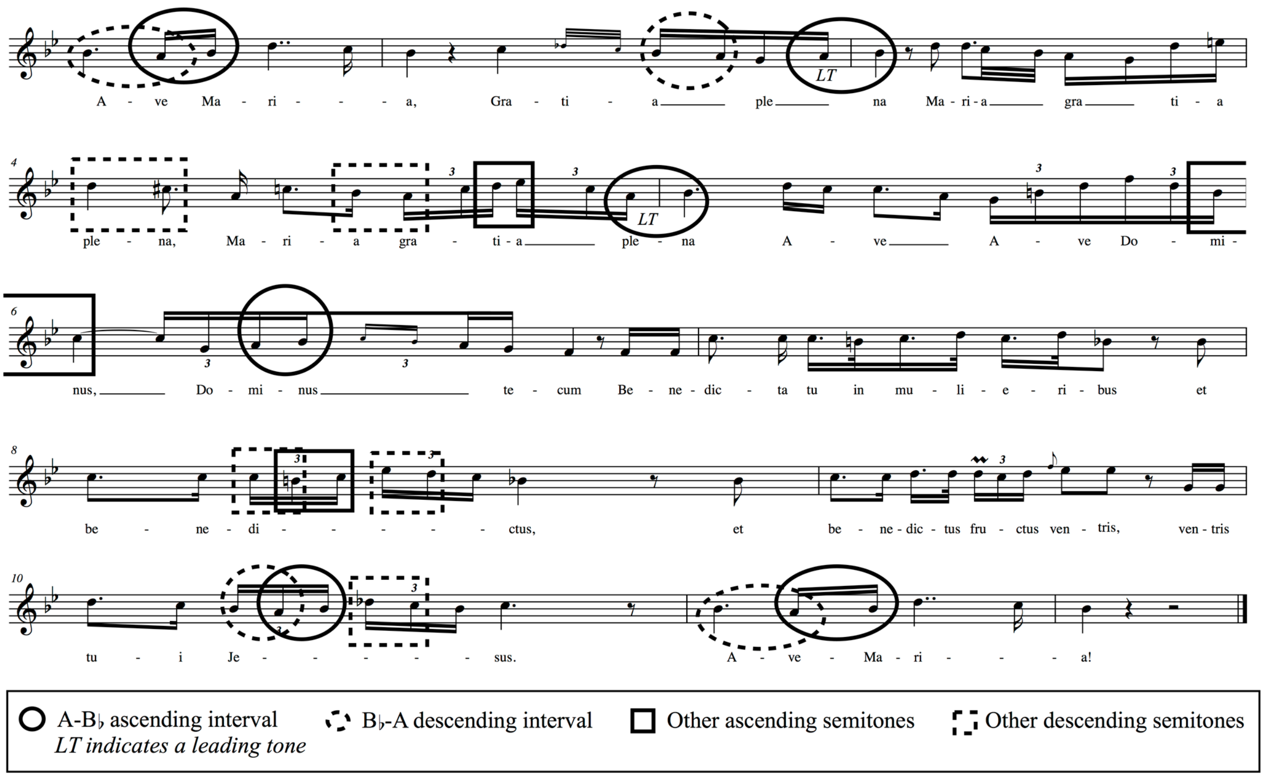

A - ve Ma - ri - - - a, Gra - ti - a ple na Ma - ri - a gra ti - a
LT
4
ple - na, Ma - ri - a gra - ti - a ple - na A - ve A - ve Do - mi -
LT
6
nus, Do - mi - nus te - cum Be - ne - dic - ta tu in mu - li - e - ri - bus et
8
be - ne - di - - - - - - ctus, et be - ne - dic - tus fru - ctus ven - tris,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

Figure 2. Score for Schubert's "Ave Maria" showing all semitone intervals.

other semitones for each group of singers. Overall each group had 144 ascending semitones for each set of *a cappella* and accompanied renditions and 162 descending semitones.

The data were analyzed in a number of ways, including an examination of the means and standard deviations across groupings of interval conditions, a visualization of the data in box and whisker plots, and a linear regression analysis to look for trends.

A 2 (accompaniment) by 2 (direction) by 2 (spelling) by 2 (group) mixed ANOVA with 6 singers per group was used to investigate the effect of musical context on semitone interval size. The between-subjects variable was training group. The rest of the variables were within-subjects. Direction had a significant effect on interval size, F (1, 1352) = 32.56, as did spelling, F= 14.48, and level of training, F = 23.88 (all ps < .001). There were no significant interactions, but this may have been due to the dwindling sample size (and hence power) associated with interaction terms. Accompaniment was only significant, F (1, 611) = 10.11, p < .001, when an ANOVA was run on the non-professional group data.

Overall we found a high degree of variability both between the singers and in terms of the singers' self-consistency. Appendix D demonstrates this variability with data from the opening and closing "Ave Maria" statements. It also reveals an intra-performance consistency in terms of the relative size of the intervals. The mean interval sizes and standard deviations across all of the singers for the various semitone conditions are shown in Table 1 and composite box and whisker plots for the interval size are shown in Figure 3, respectively. Appendix E discusses the interval size data for each singer in greater detail.

To further understand the results, a regression analysis within each group was conducted using only the main effects. There was one regression analysis within each of the professional and non-professional groups. These regressions had five different predictors, or independent variables: 1 binary-coding for accompaniment (present/absent), 1 binary coding for intervallic direction (up down), 1 coding for the leading tone (yes/no) and 1 coding for intervallic spelling, and 5 variables

Table 1
Mean interval sizes in the tested conditions

Conditions	Interval Size (in cents)			
	Non-professionals		Professionals	
	Mean	*SD*	Mean	*SD*
A cappella A-B♭ intervals, leading tones (36)	79.3	15.5	94.8	16.5
A cappella A-B♭ intervals, non-leading tones (72)	95.4	19.6	98.6	16.6
Accompanied A-B♭ intervals, leading tones (36)	90.7	12.7	93.2	14.3
Accompanied A-B♭ intervals, non-leading tones (72)	99.5	13.1	98.7	16.5
A cappella B♭-A intervals (72)	83.4	19.0	87.6	13.4
Accompanied B♭-A intervals (72)	86.3	16.2	89.9	13.9
A cappella other semitones, ascending (36)	89.2	22.3	102.7	19.7
Accompanied other semitones, ascending (36)	90.7	21.4	103.9	17.7
A cappella other semitones, descending (90)	90.4	17.5	97.2	19.4
Accompanied other semitones, descending (90)	91.0	18.6	96.9	17.7

coding for singer identity (using dummy OR contrast coding). Table 2 also details the conditions in each regression that were coded as binary variables. Linear regression analysis was chosen for these analyses because it provides information about the effect direction and size for each of the factors considered, and measures how well the sum of the weighted predictors explains the variance in the data (Cohen, 2002). In the regression analysis, ß values, which are calculated for each predictor, indicate the size and direction of the effect that the predictor has on the variable being predicted. With appropriate encoding, the ß values can also be used to evaluate the difference between the two groups that are defined by the predictor (e.g., *a cappella* or accompanied). The significance of an effect can be determined though ß's 95% confidence interval. If the confidence interval does not include zero then the effect can be considered significant.

The first regression analysis on the semitone data from the non-professional group had a relatively low but significant R^2 value (R^2 = 0.19, p

Table 2
Summary of the conditions evaluated in the regressions analysis done in this section. The columns list the different conditions tested and the X's indicate which conditions are used in the different regressions

	Accom.	**Desc.**	**Leading tone**	**Non A-B♭/ B♭-A**	**Singers (Baseline: Singer Six)**					**Pro**
					1	2	3	4	5	
Semitones (Pro)	X	X	X	X	X	X	X	X	X	
Semitones (Non-pro)	X	X	X	X	X	X	X	X	X	

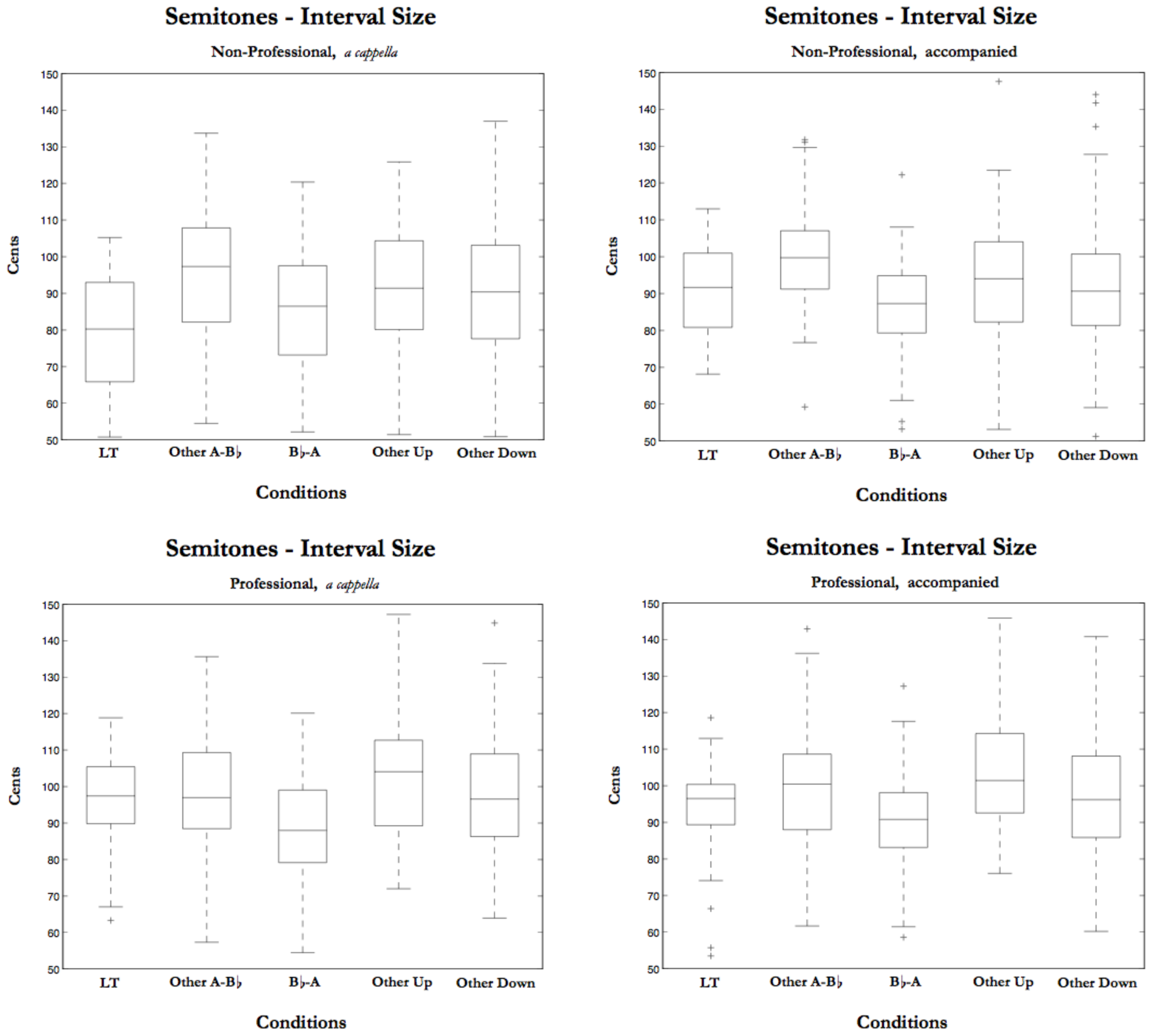


Figure 3. Box and whisker plots of the interval size in semitones for the A-B♭ semitones (leading tone and non-leading tone functions), the B♭-A semitones, and the other semitones (ascending and descending) across all singers. The plot on the left is for the *a cappella* performances and the plot on the right is for performances with accompaniment. The top and bottom of each box represents the 25th and 75th percentiles, with the solid horizontal line running through the box representing the 50th percentile, or median. The short solid horizontal lines at the end of the 'whiskers' represent the most extreme, non-outlier, data points and the plus signs indicate the outliers.

< .001), indicating that some of the variance in the data was explained by the conditions considered in the regression. The regression did reveal that the A-B♭ leading tone semitones, with a mean size of 85 cents, were on average 10 cents smaller than the other semitones (95% confidence interval = [5,14]), however there was no significant effect for the average size of the A-B♭/B♭-A semitones (including leading tones), 90 cents, compared to the other semitones. The *a cappella* semitones, with a mean size of 88 cents, were on average 3 cents (95% confidence interval = [1,6]) smaller than the accompanied ones and the descending intervals, also with a mean size of 88 cents, were on average 7 cents smaller than the ascending ones (95% confidence interval = [4,10]). There were also statistically significant effects for the average interval size of singer one's semitones (8 cents larger,

95% confidence interval = [3,12]), singer two's semitones (7 cents smaller, 95% confidence interval = [2,11]), singer four's semitones (11 cents larger, 95% confidence interval = [6,15]), and singer five's semitones (8 cents smaller, 95% confidence interval = [3,12]) in comparison to singer six, though there were no significant differences in interval size between singers three and six.

The R^2 value for the regression analysis performed on the data from the professional group ($R^2 = 0.09$, $p < .001$) was smaller than that for the non-professional group, indicating less of the variance in the data was explained. This regression analysis revealed no significant effect for the A-B♭ leading tones, with a mean size of 94 cents, compared to the other semitones. There was however, a significant effect for the A-B♭/B♭-A semitones (including leading tones), with a mean interval size of 94 cents, which were on average 7 cents smaller than the other semitones (95% confidence interval = [4,10]). There was no significant difference between the *a cappella* semitones, with a mean size of 95 cents, versus the accompanied semitones. The descending intervals, with a mean size of 93 cents, were on average 8 cents smaller than the ascending ones (95% confidence interval = [4,10]). There were not any significant effects for singer identity.

Discussion

For the singers in the non-professional group, semitones tended to be smaller than the 100 cent equal-tempered semitone, whereas for the professional group, semitones were closer to equal temperament. The 50% confidence intervals for many of the semitones shown in Figure 3 encompass the 90 cent Pythagorean semitone and the 100 cent equal-tempered semitone. Only the ascending non-B♭-A semitones in the professional group encompass the 112 cent major diatonic Just Intonation semitone. Details about the various tuning systems can be found in Barbour (1953) and Rasch (2002).

In terms of the different types of intervals, only the non-professional group showed a significant effect for leading tones versus non-leading tone intervals, with the leading tones being on average 10 cents smaller, although no such effects were found for the professional group. In contrast, the was a significant effect for the professional group's non A-B♭/B♭-A semitones versus the A-B♭/B♭-A semitones, with the non A-B♭/B♭-A semitones being on average 7 cents larger. This suggests that musical function has some influence on the intonation tendencies in the non-professional group's performances, and that note spelling has some influence on the professional group's intonation.

In terms of other influences on intonation, both groups showed a significant effect for intervallic direction on interval size with both the non-professional and the professional singers' descending semitone intervals being smaller than their ascending ones by comparable amounts on average. Only the non-professional group showed an effect for the presence of accompaniment, with their *a cappella* semitones being smaller on average than their accompanied ones. Overall the non-professional group showed more of an effect for singer identity than the professional group for interval size, where only one singer was significantly different from the baseline. There was also a significant group effect for interval size, with the professional group's interval size being 6 cents larger on average than the non-professional group's interval size.

The differences between the non-professional and professional groups suggest different ways in which intonation practices might be influenced by training and/or experience. The absence of an effect in interval size for A-B♭ leading tones in the professional group, in contrast to the non-professional group whose A-B♭ leading tones were on average 10 cents smaller than the non-leading tones, suggests that either with training the singers acquire greater stability in their production of leading tones or that the singers with less training tend to exaggerate them. The existence of a significant effect for accompaniment and greater prevalence of a significant effect for singer identity for the semitones' interval size in the non-professional group suggests that singers become more consistent both between *a cappella* and accompanied versions and with other singers when they acquire more training/experience.

Conclusions

In the opening section of this paper, we presented a brief overview of the history of performance analysis for the singing voice in audio recordings. In the second section, 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 (attack) and steady-state sections of the note. We also discussed our plans for extending this algorithm for application to untrained singers, making it robust enough to accommodate unstable pitches and incorrect intervals. In the third section, we 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," in two groups, non-professional and professional, that showed some differences between the two groups. Most notable, in the non-professional group the A-B♭ intervals with a leading tone function were significantly smaller on average than the other semitones in the piece, though this was not the case for the professional group. The professional group tended to perform the A-B♭/B♭-A semitones smaller than the non-A-B♭/B♭-A semitones. Overall, the means of the non-professionals' semitones tended to be smaller than the equal-tempered semitone while the means of the professionals' semitones were closer to equal temperament. However, the standard deviations were sufficiently large that the differences from equal temperament were negligible.

REFERENCES

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

Barbour, J. M. (1953). *Tuning and temperament: A historical survey*. East Lansing, MI: Michigan State College Press.

Bartholomew, W. T. (1934). A physical definition of "good voice-quality" in the male voice. *Journal of the Acoustical Society of America, 6*, 25–33.

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio, 13*, 1035–1047.

Bengtsson, I., & Gabrielsson, A. (1980). Methods for analyzing performance of musical rhythm. *Scandinavian Journal of Psychology, 21,* 257–268.

Bengtsson, I., & Gabrielsson, A. (1983). Analysis and synthesis of musical rhythm. *Studies of Musical Performance, 39*, 27–60.

Birmingham, W., Dannenberg, R., Wakefield, G., Bartsch, M., Bykowski, D., Mazzoni, D., . . . Rand, W. (2001). MUSART: Music retrieval via aural queries. In S. Downie & D. Bainbridge (Eds.), *International Symposium on Music Information Retrieval 2001* (pp. 73–81). Bloomington, IN: University of Indiana Press.

Boersma, P., & Weenink, D. (n.d.). PRAAT: doing phonetics by computer [Computer software]. Available at http://www.praat.org

Cano, P., Loscos, A., & Bonada, J. (1999). Score-performance matching using HMMs. In A. Horner (Ed.), *International Computer Music Conference 1999* (pp. 441–444). San Francisco, CA: International Computer Music Association.

Carlsson-Berndtsson, G., & Sundberg, J. (1991). Formant frequency tuning in singing. *STL-Quarterly Progress and Status Report, 32*(1), 29–35.

Clarisse, L., Martens, J., Lesaffre, M., Baets, B., Meyer, H., & Leman, M. (2002). An auditory model based transcriber of singing sequences. In M. Fingerhut (Ed.), *International Symposium on Music Information Retrieval 2002* (pp. 116–123). Paris, France: IRCAM.

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

Cohen, C. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge Academic.

Cont, A., Schwarz, D., Schnell, N., & Raphael, C. (2007). Evaluation of real-time audio-to-score alignment. In S. Dixon, D. Bainbridge, & R. Typke (Eds.), *International Symposium on Music Information Retrieval 2007* (pp. 315–316). Vienna, Austria: Österreichische Computer Gesellschaft.

Cooper, D., & Sapiro, I. (2006). Ethnomusicology in the laboratory: From the tonometer to the digital melograph. *Ethnomusicology Forum, 15*, 301–313.

Dannenberg, R., Birmingham, W., Pardo, B., Hu, N., Meek, C., & Tzanetakis, G. (2007). A comparative evaluation of search techniques for query-by-humming using the MUSART testbed. *Journal of the American Society for Information Science and Technology, 58,* 687–701.

de Cheveigné, A. (2002). YIN MATLAB implementation [Computer software]. Available at http://audition.ens.fr/adc/

de Cheveigné, A., & Henrich, N. (2002). Fundamental frequency estimation of musical sounds. *Journal of the Acoustical Society of America, 111*, 2416.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*, 1917–1930.

Devaney, J. (2011). AMPACT: Automatic Music Performance Analysis and Comparison Toolkit [Computer software]. Available at http://www.ampact.org

Devaney, J., & Ellis, D. P. W. (2009). Handling asynchrony in audio-score alignment. In G. Scavone, V. Verfaille, & A. da Silva (Eds.), *International Computer Music Conference 2009* (pp. 29–32). San Francisco, CA: International Computer Music Association.

Devaney, J., Mandel, M. I., & Ellis, D. P. W. (2009). Improving MIDI-audio alignment with acoustic features. In J. Benesty & T. Gaensler (Eds.), *Workshop on Applications of Signal Processing to Audio and Acoustics 2009* (pp. 45–48). Piscataway, NJ: IEEE.

Dixon, S. (2003). Towards automatic analysis of expressive performance. In R. Kopiez (Ed.), *European Society for the Cognitive Sciences of Music Conference 2003* (pp. 107–110). Hanover, Germany: ESCOM.

Downie, J. (2008). MIREX 2008: Real-time audio to score alignment (a.k.a. score following). Retrieved from http://www.music-ir.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,* 155–182.

Easley, E. (1932). A comparison of the vibrato in concert and opera singing. In C. Seashore (Ed.), *University of Iowa studies in the psychology of music, Vol. I: The vibrato* (pp. 269–275). Iowa City, IA: University of Iowa.

Eerola, T., & Toiviainen, P. (2004). MIDI Toolbox: MATLAB Tools for Music Research [Computer software]. University of Jyväskylä: Kopijyvä, Jyväskylä, Finland. Available at http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/miditoolbox/

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology, 2,* 145–161.

Friberg, A., Schoonderwaldt, E., & Juslin, P. N. (2007). CUEX: An algorithm for extracting expressive tone variables from audio recordings. *Acta Acustica united with Acustica, 93*, 411–420.

Fyk, J. (1995). *Melodic intonation, psychoacoustics, and the violin.* Zielona Góra, Poland: 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*, 221–272.

Gockel, H., Moore, B. C. J., & Carlyon, R. P. (2001). Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *Journal of the Acoustical Society of America, 109*, 701–712.

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.

Gramming, P., Sundberg, J., Ternström, S., Leanderson, R., & Perkins, W. H. (1987). Relationship between changes in voice pitch and loudness. *STL-Quarterly Progress and Status Report, 28*(1), 39–55.

Hagerman, B., & Sundberg, J. (1980). Fundamental frequency adjustment in barbershop singing *STL-Quarterly Progress and Status Report, 21*(1), 28–42.

Hong, J.-L. (2003). Investigating expressive timing and dynamics in recorded cello performances. *Psychology of Music, 31,* 340–352.

Howard, D. M. (2007a). Equal or non-equal temperament in *a cappella* SATB singing. *Logopedics Phoniatrics Vocology, 32*, 87–94.

Howard, D. M. (2007b). Intonation drift in *a capella* soprano, alto, tenor, bass quartet singing with key modulation. *Journal of Voice, 21*, 300–315.

Hu, N., & Dannenberg, R. (2006). Bootstrap learning for accurate onset detection. *Machine Learning, 65,* 457–471.

Hu, N., Dannenberg, R., & Tzanetakis, G. (2003). Polyphonic audio matching and alignment for music retrieval. In G. Wakefield (Ed.), *Workshop on Applications of Signal Processing to Audio and Acoustics 2003* (pp. 185–188). Piscataway, NJ: IEEE.

Jers, H., & Ternström, S. (2005). Intonation analysis of a multi-channel choir recording. *TMH-Quarterly Progress and Status Report, 47*(1), 1–6.

Kurth, F., Müller, M., Fremerey, C., Chang, Y., & Clausen, M. (2007). Automated synchronization of scanned sheet music with audio recordings. In S. Dixon, D. Bainbridge, & R. Typke (Eds.), *International Symposium on Music Information Retrieval 2007* (pp. 261–266). Vienna, Austria: Österreichische Computer Gesellschaft.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). Boston, MA: McGraw-Hill.

Marinescu, M. C., & Ramirez, R. (2008). Expressive performance in the human tenor voice. In M. Supper & S. Weinzierl (Eds.), *Sound and Music Computing Conference 2008*. Berlin, Germany: Technische Universität.

Metfessel, M. (1932). The vibrato in artistic voices. In C. Seashore (Ed.), *University of Iowa studies in the psychology of music, Vol. I: The vibrato* (pp. 14–117). Iowa City, IA: University of Iowa.

Miller, D. C. (1916). *The science of musical sounds*. New York: Macmillan Press.

Miller, R. S. (1936). The pitch of the attack in singing. In C. Seashore (Ed.), *University of Iowa studies in the psychology of music, Vol. IV: Objective analysis of muscial performance* (pp. 158–171). Iowa City, IA: University of Iowa.

Murphy, K. (2005). Hidden Markov Model (HMM) Toolbox for Matlab [Computer software]. Vancouver, BC: University of British Columbia. Available at http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

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. In A. Schloss, R. Dannenberg, & P. Driessen (Eds.), *International Computer Music Conference 2001* (pp. 151–154). San Francisco, CA: International Computer Music Association.

Orio, N., & Schwarz, D. (2001). Alignment of monophonic and polyphonic music to a score. In A. Schloss, R. Dannenberg, & P. Driessen (Eds.), *International Computer Music Conference 2001* (pp. 155–158). San Francisco, CA: International Computer Music Association.

Ornoy, E. (2007). 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,* 115–138.

Pardo, B., & Sanghi, M. (2005). Polyphonic musical sequence alignment for database search. In J. Reiss & G. Wiggins (Eds.), *International Symposium on Music Information Retrieval 2005* (pp. 215–221). London: University of London.

Peeling, P., Cemgil, T., & Godsill, S. (2007). A probabilistic framework for matching music representations. In S. Dixon, D. Bainbridge, & R. Typke. (Eds.), *International Symposium on Music Information Retrieval 2007* (pp. 267–272). Vienna, Austria: Österreichische Computer Gesellschaft.

Platt, R. W. (1998). ANOVA, t tests, and linear regression. *Prevention Injury, 4*, 52–53.

Prame, E. (1994). Measurements of the vibrato rate of ten singers. *Journal of the Acoustical Society of America, 96*, 1979–1984.

Prame, E. (1997). Vibrato extent and intonation in professional western lyric singing. *Journal of the Acoustical Society of America, 102*, 616–621

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77,* 257–286.

Raphael, C. (2004). A hybrid graphical model for aligning polyphonic audio with musical scores. In C. Lomeli Buyoli & R. Loureiro (Eds.) *International Symposium on Music Information Retrieval 2004* (pp. 387–394). Barcelona, Spain: Universitat Pompeu Fabra.

Rapoport, E. (2007). The marvels of the human voice: Poem–melody–vocal performance. *Orbis Musicae, 14,* 7–36.

Rasch, R. (2002). Tuning and temperament. In T. Christensen (Ed.), *The Cambridge history of Western music theory* (pp. 193–222). Cambridge, UK: Cambridge University Press.

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,* 2546–2568.

Repp, B. (1997). The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception, 14*, 419–444.

Ryynänen, M. P., & Klapuri, A. P. (2004). Modelling of note events for singing transcription. In B. Raj, P. Smaragdis, & D. Ellis (Eds.), *Workshop on Statistical and Perceptual Audio Processing 2004* (pp. 319–322). Jeju, South Korea: ICSA Archive.

Ryynänen, M. P., & Klapuri, A. P. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal, 32*(3), 72–86.

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. 361–380). Mahwah, NJ: Lawrence Erlbaum.

Schoen, M. (1922). An experimental study of the pitch factor in artistic singing. *Psychological Monographs, 31,* 230–259.

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

Seashore, H. G. (1936). An objective analysis of artistic singing. In C. Seashore (Ed.), *University of Iowa studies* *in the psychology of music, Vol. IV: Objective analysis of musical performance* (pp. 12–157). Iowa City, IA: University of Iowa.

Seeger, C. (1951). An Instantaneous Music Notator. *Journal of the International Folk Music Council, 3*, 103–106.

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. In A. Rangarajan (Ed.), *Proceedings of the International Conference on Multimedia and Expo, 2003* (pp. 61–64). Baltimore, MD: IEEE.

Sundberg, J. (1982). In tune or not? A study of fundamental frequency in music practise. *STL-Quarterly Progress and Status Report, 23*(1), 49–78.

Sundberg, J. (1987). *The science of the singing voice.* Dekalb, IL: Northern Illinois University Press.

Sundberg, J. (1994). Acoustic and psychoacoustic aspects of vocal vibrato. *STL-Quarterly Progress and Status Report, 35*(2–3), 45–68.

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.

Sundberg, J., Prame, E., & Iwarsson, J. (1995). Replicability and accuracy of pitch patterns in professional singers. *STL-Quarterly Progress and Status Report, 36*(2-3), 51–62.

Ternström S. (2003). Choir acoustics: An overview of scientific research published to date. *International Journal of Research in Choral Singing, 1*, 3–11.

Ternström, S., & Sundberg, J. (1988). Intonation precision of choir singers. *Journal of the Acoustical Society of America, 84*, 59–69.

Tiffin, J. (1932). Phonophotograph apparatus. In C. Seashore (Ed.), *University of Iowa studies in the psychology of music, Vol. I: The vibrato* (pp. 118–133). Iowa City, IA: University of Iowa.

Timmers, R. (2007). Vocal expression in recorded performances of Schubert songs. *Musicae Scientiae, 11*, 237–268.

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

Todd, N. (1989). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91*, 3540–3550.

Toh, C. C., Zhang, B., & Wang, Y. (2008). Multiple-feature fusion based onset detection for solo singing voice. In J. Bello, E. Chew, & D. Turnbull (Eds.), *International Symposium on Music Information Retrieval 2008*, 515–520. Philadelphia, PA: Drexel University.

Tove, P. A., Norman, B., Isaksson, L., & Czekajewski, J. (1966). Direct-recording frequency and amplitude meter for analysis of musical and other sonic waveforms. *Journal of the Acoustical Society of America 39*, 362–371.

Turetsky, R., & Ellis, D. (2003). Ground–truth transcriptions of real music from force–aligned MIDI syntheses. In H. Hoos & D. Bainbridge (Eds.), *International Symposium on Music Information Retrieval 2003* (pp. 135–141). Baltimore, MD: Johns Hopkins University.

Unal, E., Chew, E., Georgiou, P. G., & Narayanan, S. S. (2008). Challenging uncertainty in Query-by-Humming systems: A fingerprinting approach. *IEEE Transactions on Audio, Speech, and Language Processing, 16*, 359–371.

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*, 338–349.

Weihs, C., & Ligges, U. (2003). Automatic transcription of singing performances. *Bulletin of the International Statistical Institute, 60,* 507–510.

APPENDIX A

DYNAMIC TIME WARPING AND HIDDEN MARKOV MODELS

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. A comparative evaluation of different features found peak spectral difference (Orio & Schwarz, 2001) to be the most robust 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 Figure A1. 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 file to their occurrences in the audio. The black line in Figure A1 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.

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 a current state, since the current state summarizes the past sequence of events. A musical example of a simple HMM is one that determines whether a note is present or not using only two states: note and rest. A slightly more complicated HMM for the same problem could 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 any 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. In a singing context, 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), as well as for singing transcription (Ryynänen & Klapuri, 2004; Shih, Narayanan, & Kuo, 2003) 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.

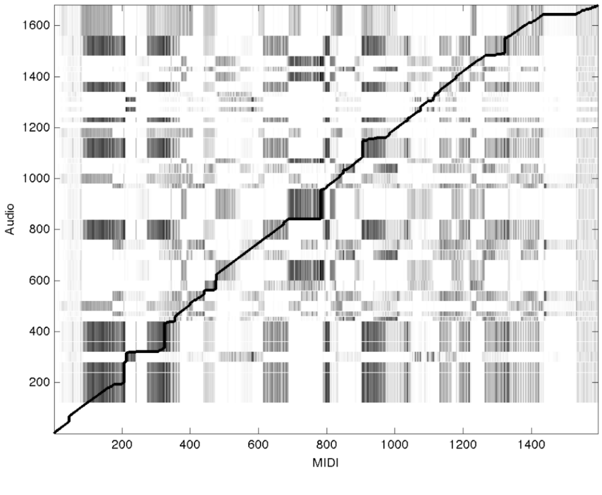


Figure A1. Dynamic Time Warping Similarity Matrix. The black line indicates the optimal path through the similarity matrix, which is used to warp the timing in the audio and MIDI to match each other. The y-axis is the number of audio frames and the x-axis is the number of MIDI frames. Black indicates high similarity and white indicates low similarity.

(Appendices continue)

APPENDIX B

TECHNICAL DESCRIPTION OF THE ALIGNMENT ALGORITHM

The HMM in our alignment algorithm (Devaney, Mandel, & Ellis, 2009) has three basic states: silence, transient, and steady state, which were each defined in terms of average estimates of periodicity and power in the audio signal. The general characteristics of each state can be observed in Figure B1; 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 means and covariances for aperiodicty and power were calculated from the YIN estimates of hand-labeled silence, transient, and steady state sections in several recordings of Schubert's "Ave Maria" and Machaut's *Notre Dame Mass*. We also make use of F_0 estimates from YIN, which provides a somewhat noisy cue, especially for the silence and transient states, and the standard deviation used to model it varied accordingly. The F_0 estimates assist alignment when 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 the same labeled audio as was used for the aperiodicity and power ranges. 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.

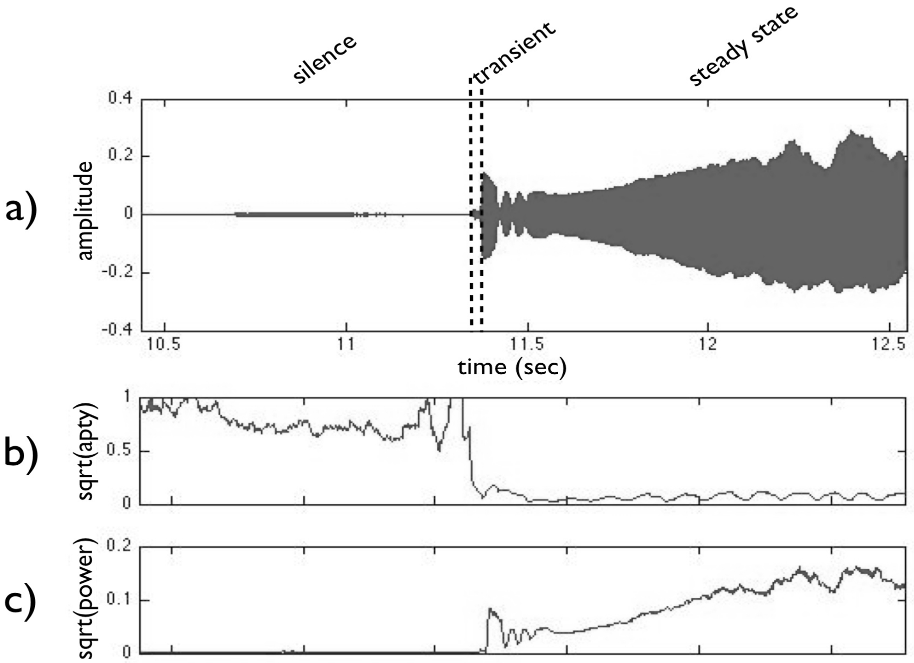


Figure B1. Visualization of the HMM states defined in the alignment algorithm: (a) is the time domain representation of a sung note with the HMM states labeled, (b) is the aperiodicity measure, and (c) is the power measure.

The general state sequence for the HMM representing a single note is shown in Figure B2a. 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 (SS) state can be followed by an ending-transient (ET) state, a silence (S) state or a beginning-transient (BT) 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 B2b.

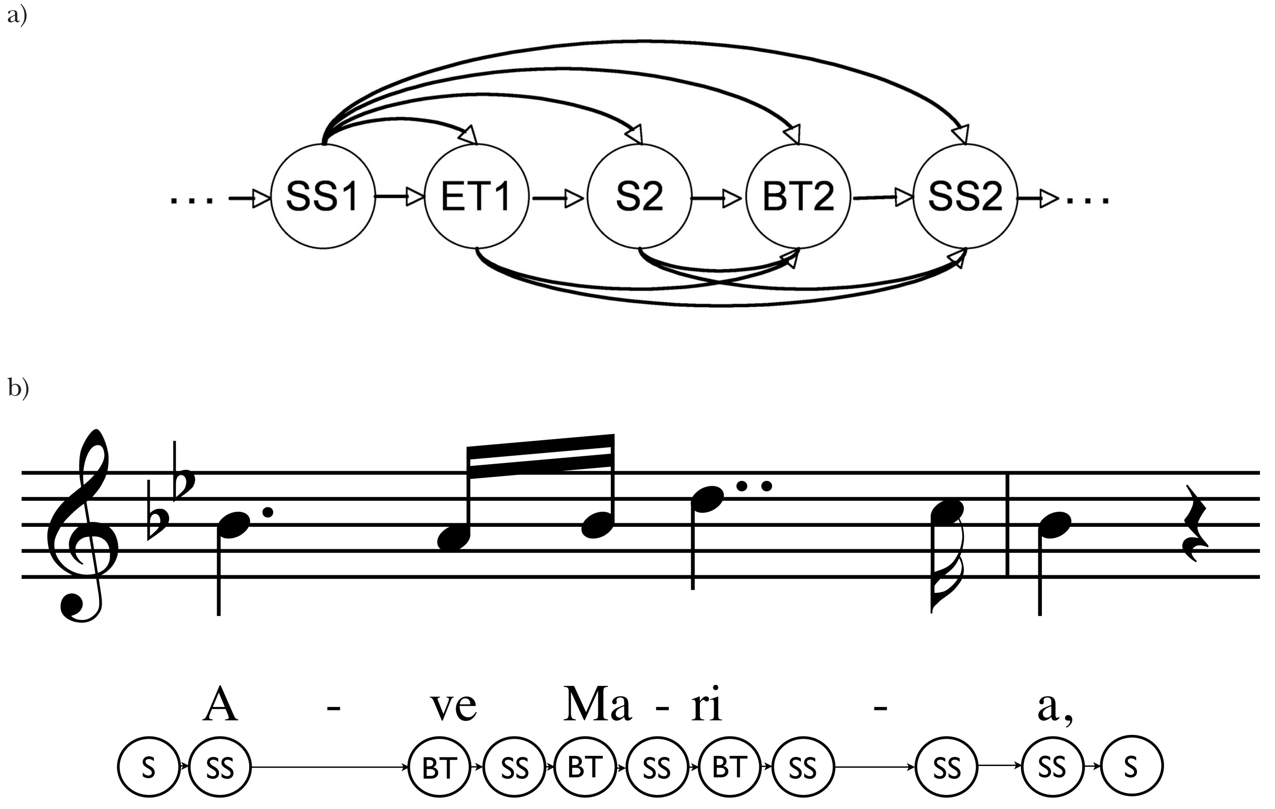


Figure B2. State sequence diagrams for the alignment algorithm's HMM. Subfigure (a) shows the state-sequence seed that is used as the basis for the full HMM state sequence. The four components are steady state (SS1 for the first note and SS2 for the second note), ending transient (ET1), silence (S2), and beginning transient (BT2). In the general version of the state-sequence, the seed is repeated verbatim. Subfigure (b) shows state sequence adapted to sung text, most notably silences are only present when there is a rest in the score and transients are only present when the sung syllable begins or ends with a consonant.

(Appendices continue)

APPENDIX C

ALIGNMENT ALGORITHM USAGE

We are planning to make this algorithm available to other researchers, both in its current form and in an expanded form. The algorithm is written as a set of MATLAB functions and requires the Mathwork's Signal Processing Toolbox, as well several freely available toolkits, namely Kevin Murphy's HMM toolkit (Murphy, 2005), Alain de Cheveigné's YIN implementation (de Cheveigné, 2002), and Tuomas Eerola and Petri Toiviainen's MIDI Toolbox (Eerola & Toiviainen, 2004). 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.

The algorithm requires:

a) An audio file of the recording.

b) A symbolic representation of the score in the form of a MIDI file.

c) An annotation of the lyrics in the following format:

   i. A list of each isolated syllable and silence (rest);

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

Lyric Annotation

The lyric annotation provides information on the relationship between the sung syllables and the notes. The number of notes sung under a syllable is indicated, with rests always assigned a value of 0. See Table C1 for an example of the annotation. Once the lyrics have been annotated the algorithm can calculate the expected characteristics of each note (i.e., whether a beginning or ending transient is present).

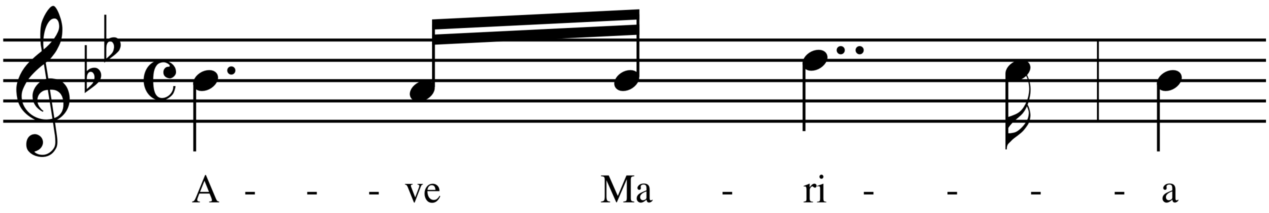
Table C1
Example of lyric annotation using the first two phrases of Schubert's "Ave Maria"

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

APPENDIX D

MEAN INTERVAL SIZE FOR OPENING AND CLOSING "AVE MARIA"S FOR EACH SINGER

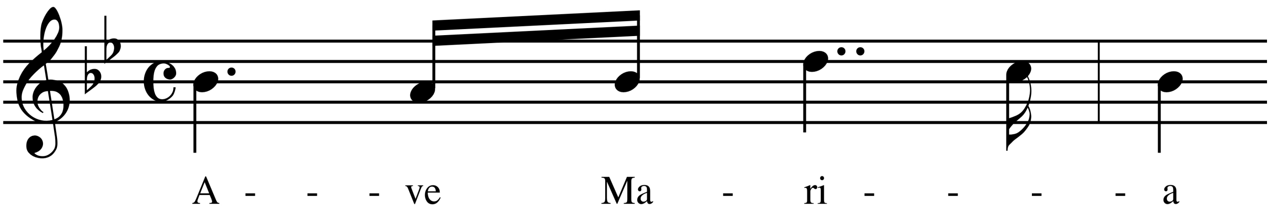
Table D1



Idealized Interval Sizes					
Equal Temperament	-100	100	400	-200	-200
Pythagorean	-90	90	408	-204	-204
5-limit Just Intonation	-112	112	386	-182 (Minor)	-204 (Major)
Non-professional Singer 1					
Opening *A cappella*	-86.6 (*SD* = 14.5)	87.8 (*SD* = 8.3)	397.0 (*SD* = 5.8)	-210.6 (*SD* = 2.0)	-200.2 (*SD* = 3.4)
Opening Accompanied	-95.8 (*SD* = 3.6)	101.1 (*SD* = 5.3)	394.73 (*SD* = 4.9)	-207.6 (*SD* = 8.8)	-197.5 (*SD* = 8.6)
Closing *A cappella*	-87.9 (*SD* = 8.0)	99.1 (*SD* = 3.4)	386.6 (*SD* = 1.6)	-204.0 (*SD* = 4.6)	-204.2 (*SD* = 1.8)
Closing Accompanied	-83.2 (*SD* = 6.1)	96.2 (*SD* = 5.2)	379.6 (*SD* = 5.7)	-197.3 (*SD* = 7.1)	-200.5 (*SD* = 10.0)
Non-professional Singer 2					
Opening *A cappella*	-60.2 (*SD* = 12.1)	101.6 (*SD* = 12.4)	368.1 (*SD* = 14.2)	-196.1 (*SD* = 17.9)	-207.4 (*SD* = 7.2)
Opening Accompanied	-74.8 (*SD* = 16.9)	98.4 (*SD* = 22.5)	368.4 (*SD* = 11.0)	-184.7 (*SD* = 4.4)	-216.5 (*SD* = 9.1)
Closing *A cappella*	-53.4 (*SD* = 0.7)	85.9 (*SD* = 31.4)	370.7 (*SD* = 15.9)	-194.7 (*SD* = 17.9)	-223.6 (*SD* = 9.5)
Closing Accompanied	-63.8 (*SD* = 3.5)	101.0 (*SD* = 6.2)	375.83 (*SD* = 6.1)	-183.2 (*SD* = 11.3)	-227.0 (*SD* = 16.5)
Non-professional Singer 3					
Opening *A cappella*	-91.6 (*SD* = 6.3)	99.5 (*SD* = 5.5)	377.7 (*SD* = 9.6)	-190.0 (*SD* = 4.6)	-194.5 (*SD* = 4.4)
Opening Accompanied	-92.2 (*SD* = 8.4)	97.6 (*SD* = 4.5)	375.5 (*SD* = 10.6)	-192.3 (*SD* = 3.7)	-199.3 (*SD* = 7.9)
Closing *A cappella*	-90.4 (*SD* = 11.8)	108.2 (*SD* = 10.2)	373.0 (*SD* = 15.3)	-185.1 (*SD* = 11.3)	-204.2 (*SD* = 15.3)
Closing Accompanied	-88.5 (*SD* = 14.2)	95.4 (*SD* = 13.4)	376.3 (*SD* = 7)	-176.7 (*SD* = 4.6)	-213.6 (*SD* = 5.6)

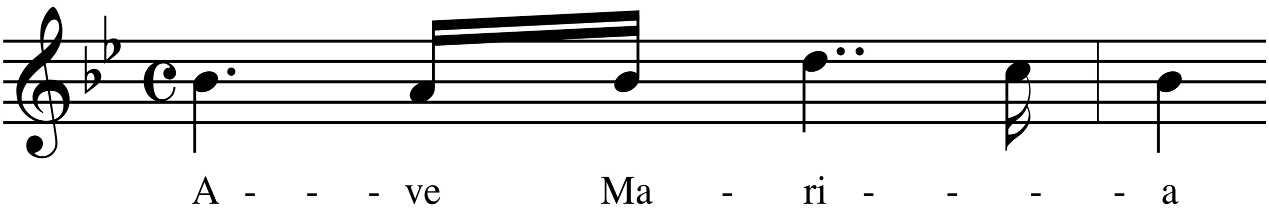
(Appendices continue)

Table D1 *(continued)*



Idealized Interval Sizes					
Equal Temperament	-100	100	400	-200	-200
Pythagorean	-90	90	408	-204	-204
5-limit **Just Intonation**	-112	112	386	-182 (Minor)	-204 (Major)
Non-professional Singer 4					
Opening ***A cappella***	-86.9 (*SD* = 9.0)	88.9 (*SD* = 17.5)	394.1 (*SD* = 10.6)	-217.6 (*SD* = 10.8)	-197.7 (*SD* = 5.3)
Opening **Accompanied**	-89.5 (*SD* = 8.6)	102.4 (*SD* = 12.6)	387.2 (*SD* = 1.6)	-200.3 (*SD* = 9.6)	-200.6 (*SD* = 9.6)
Closing ***A cappella***	-88.9 (*SD* = 8.6)	82.1 (*SD* = 8.8)	416.8 (*SD* = 11.1)	-211.5 (*SD* = 3.9)	-208.1 (*SD* = 7.4)
Closing **Accompanied**	-87.5 (*SD* = 3.0)	96.1 (*SD* = 4.9)	397.3 (*SD* = 6.3)	-205.8 (*SD* = 11.0)	-213.0 (*SD* = 3.9)
Non-professional Singer 5					
Opening ***A cappella***	-55.3 (*SD* = 5.3)	98.0 (*SD* = 5.5)	356.1 (*SD* = 3.2)	-180.9 (*SD* = 11.9)	-223.5 (*SD* = 19.6)
Opening **Accompanied**	-58.7 (*SD* = 7.7)	90.7 (*SD* = 6.7)	378.9 (*SD* = 5.5)	-200.2 (*SD* = 6.5)	-201.8 (*SD* = 9.9)
Closing ***A cappella***	-60.9 (*SD* = 2.9)	100.8 (*SD* = 0.9)	352.8 (*SD* = 6.6)	-195.5 (*SD* = 11.4)	-227.37 (*SD* = 8.7)
Closing **Accompanied**	-71.2 (*SD* = 9.4)	99.1 (*SD* = 15.9)	362.7 (*SD* = 9.8)	-195.5 (*SD* = 6.8)	-191.0 (*SD* = 18.4)
Non-professional Singer 6					
Opening ***A cappella***	-58.8 (*SD* = 13.1)	88.4 (*SD* = 20.4)	394.9 (*SD* = 6.0)	-192.6 (*SD* = 14.4)	-202.5 (*SD* = 2.5)
Opening **Accompanied**	-73.3 (*SD* = 5.2)	95.5 (*SD* = 3.5)	385.3 (*SD* = 4.5)	-207.3 (*SD* = 4.4)	-215.8 (*SD* = 6.8)
Closing ***A cappella***	-80.2 (*SD* = 16.7)	89.5 (*SD* = 10.0)	398.3 (*SD* = 3.4)	-198.7 (*SD* = 9.6)	-218.6 (*SD* = 3.7)
Closing **Accompanied**	-75.1 (*SD* = 6.9)	88.5 (*SD* = 2.2)	381.0 (*SD* = 10.9)	-199.8 (*SD* = 8.7)	-198.7 (*SD* = 13.9)

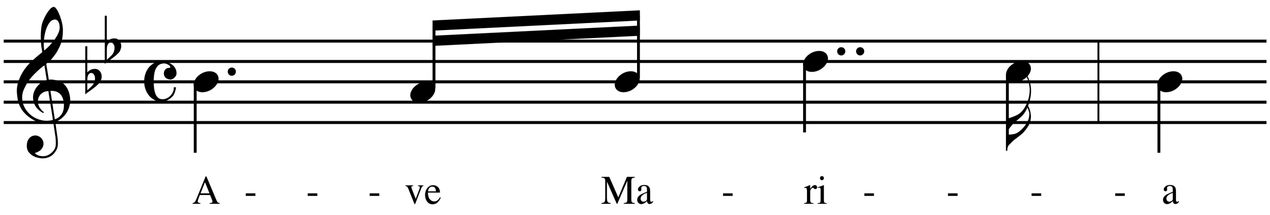
Table D1 *(continued)*



			Idealized Interval Sizes		
Equal Temperament	-100	100	400	-200	-200
Pythagorean	-90	90	408	-204	-204
5-limit **Just Intonation**	-112	112	386	-182 (Minor)	-204 (Major)
			Professional Singer 1		
Opening ***A cappella***	-87.4 (*SD* = 3.4)	96.8 (*SD* = 7.6)	398.7 (*SD* = 4.2)	-224.7 (*SD* = 4.5)	-197.4 (*SD* = 3.9)
Opening **Accompanied**	-97.0 (*SD* = 2.8)	103.5 (*SD* = 1.5)	392.9 (*SD* = 6.0)	-205.7 (*SD* = 9.8)	-199.11 (*SD* = 8.5)
Closing ***A cappella***	-97.1 (*SD* = 10.8)	86.4 (*SD* = 16.7)	420.6 (*SD* = 7.2)	-220.6 (*SD* = 5.3)	-204.3 (*SD* = 9.8)
Closing **Accompanied**	-83.1 (*SD* = 7.3)	92.3 (*SD* = 4.0)	383.7 (*SD* = 5.8)	-197.6 (*SD* = 6.7)	-200.4 (*SD* = 9.6)
			Professional Singer 2		
Opening ***A cappella***	-87.5 (*SD* = 10.2)	90.9 (*SD* = 15.7)	392.4 (*SD* = 10.9)	-216.6 (*SD* = 10.1)	-198.3 (*SD* = 4.7)
Opening **Accompanied**	-89.7 (*SD* = 9.5)	103.4 (*SD* = 13.0)	387.2 (*SD* = 2.0)	-203.9 (*SD* = 13.5)	-198.2 (*SD* = 13.4)
Closing ***A cappella***	-89.1 (*SD* = 3.1)	81.9 (*SD* = 8.0)	417.46 (*SD* = 398.4)	-212.1 (*SD* = 4.8)	-208.4 (*SD* = 7.6)
Closing **Accompanied**	-86.6 (*SD* = 2.3)	94.6 (*SD* = 6.4)	398.4 (*SD* = 7.1)	-206.1 (*SD* = 11.6)	-213.4 (*SD* = 4.5)
			Professional Singer 3		
Opening ***A cappella***	-68.1 (*SD* = 4.4)	81.0 (*SD* = 11.8)	402.8 (*SD* = 7.2)	-188.1 (*SD* = 12.3)	-237.1 (*SD* = 3.6)
Opening **Accompanied**	-76.4 (*SD* = 19.3)	87.1 (*SD* = 22.4)	402.3 (*SD* = 21.9)	-202.5 (*SD* = 9.9)	-224.22 (*SD* = 11.5)
Closing ***A cappella***	-77.0 (*SD* = 10.3)	95.3 (*SD* = 2.4)	398.6 (*SD* = 11.8)	-201.2 (*SD* = 7.4)	-227.0 (*SD* = 9.2)
Closing **Accompanied**	-79.1 (*SD* = 9.8)	99.2 (*SD* = 28.5)	395.4 (*SD* = 10.8)	-204.4 (*SD* = 22.6)	-212.7 (*SD* = 20.6)

(Appendices continue)

Table D1 *(continued)*



		Idealized Interval Sizes			
Equal Temperament	-100	100	400	-200	-200
Pythagorean	-90	90	408	-204	-204
5-limit **Just Intonation**	-112	112	386	-182 (Minor)	-204 (Major)
		Professional Singer 4			
Opening ***A cappella***	-88.9 (*SD* = 19.5)	90.2 (*SD* = 5.6)	417.2 (*SD* = 7.6)	-198.9 (*SD* = 14.1)	-207.03 (*SD* = 16.1)
Opening **Accompanied**	-87.5 (*SD* = 34.4)	91.6 (*SD* = 44.6)	402.82 (*SD* = 14.04)	-187.15 (*SD* = 22.7)	-223.54 (*SD* = 19.1)
Closing ***A cappella***	-82.3 (*SD* = 13.4)	82.3 (*SD* = 7.9)	399.1 (*SD* = 6.3)	-201.2 (*SD* = 18.1)	-204.3 (*SD* = 12.6)
Closing **Accompanied**	-93.7 (*SD* = 9.4)	91.2 (*SD* = 2.9)	407.0 (*SD* = 9.6)	-203.8 (*SD* = 28.4)	-199.6 (*SD* = 27.6)
		Professional Singer 5			
Opening ***A cappella***	-98.0 (*SD* = 6.7)	93.4 (*SD* = 5.7)	399.24 (*SD* = 5.0)	-208.6 (*SD* = 7.2)	-194.14 (*SD* = 5.2)
Opening **Accompanied**	-98.4 (*SD* = 9.5)	110.75 (*SD* = 19.1)	391.4 (*SD* = 9.5)	-191.2 (*SD* = 13.2)	-209.4 (*SD* = 13.6)
Closing ***A cappella***	-102.8 (*SD* = 1.2)	96.4 (*SD* = 4.4)	400.6 (*SD* = 6.8)	-191.2 (*SD* = 1.8)	-208.5 (*SD* = 2.1)
Closing **Accompanied**	-93.7 (*SD* = 6.9)	92.87 (*SD* = 9.1)	397.2 (*SD* = 5.6)	-194.6 (*SD* = 3.8)	-201.6 (*SD* = 4.4)
		Professional Singer 6			
Opening ***A cappella***	-87.4 (*SD* = 3.4)	96.8 (*SD* = 7.6)	398.7 (*SD* = 4.2)	-224.7 (*SD* = 4.5)	-197.4 (*SD* = 3.8)
Opening **Accompanied**	-97.0 (*SD* = 2.8)	103.5 (*SD* = 1.5)	392.9 (*SD* = 6.0)	-205.0 (*SD* = 9.8)	-199.1 (*SD* = 8.5)
Closing ***A cappella***	-97.1 (*SD* = 10.8)	86.4 (*SD* = 16.7)	420.6 (*SD* = 7.2)	-220.6 (*SD* = 5.3)	-204.2 (*SD* = 1.8)
Closing **Accompanied**	-83.13 (*SD* = 7.3)	92.3 (*SD* = 4.0)	383.7 (*SD* = 5.8)	-197.6 (*SD* = 6.7)	-200.5 (*SD* = 10.0)

APPENDIX E

COMPARISON OF SEMITONE PERFORMANCES ACROSS SINGERS

In the box and whisker plots in Figures E1 and E2, the top and bottom of boxes represent the 25th and 75th percentiles, with the solid horizontal line running through the box representing the 50th percentile, or median. The short solid horizontal lines at the end of the 'whiskers' represent the most extreme, non-outlier, data points and the plus signs indicate the outliers. Each plot shows the results for the six singers individually and the combination of all of the singers on the right. Figures E1 and E2 compare the sizes of the ascending and descending semitones in the *a cappella* and accompanied contexts for the non-professional and professional singers, respectively. The plots show the high degree of variability both across singers and in terms of

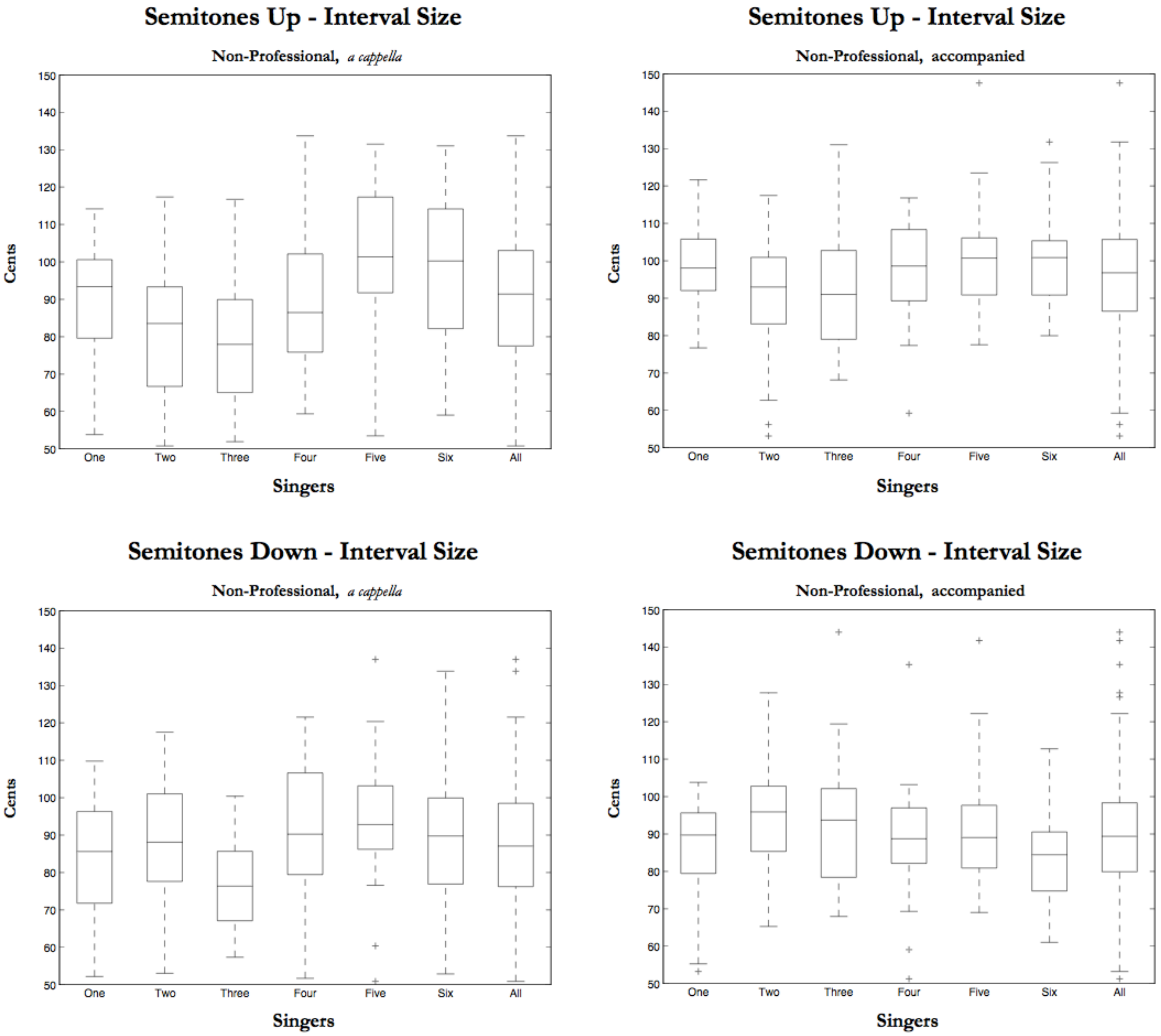


Figure E1. This figure shows box and whisker plots comparing the sizes of all of the ascending versus descending semitones across the non-professional subjects. Each subject is represented individually on the x-axis as well as the combination of all of the subjects. The y-axis shows the size of the intervals in semitones.

(Appendices continue)

each singer's self-consistency. The smaller the boxes, the more consistent the singer was in that condition. Some points of interest include the effect of accompaniment and directionality on variability. With accompaniment, some tend more towards equal temperament while others were more consistent with their interval sizes in *a cappella* version. In terms of directionality, some singers exhibit less variability for the ascending intervals than the descending ones.

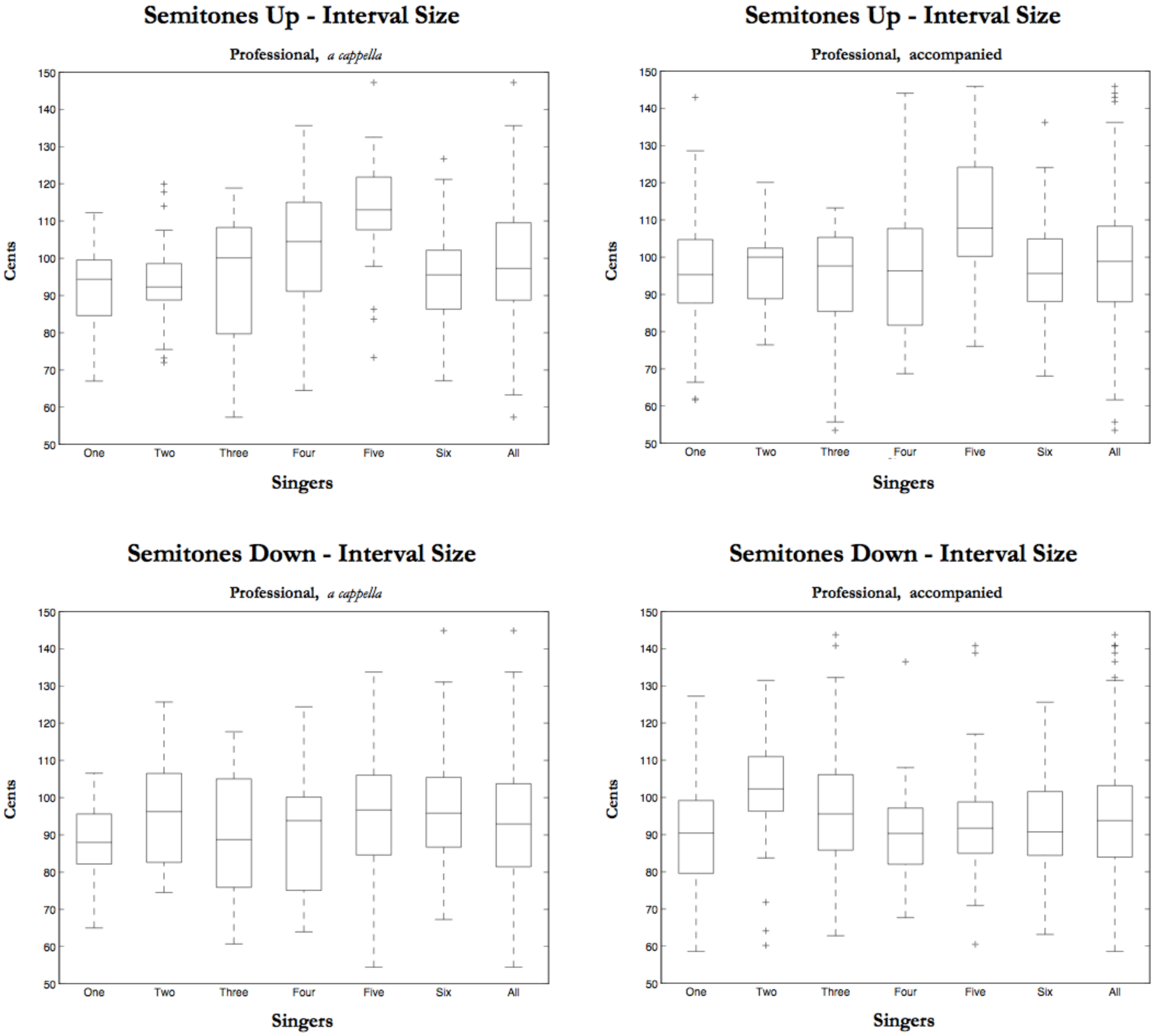


Figure E2. This figure shows box and whisker plots comparing the sizes of all of the ascending versus descending semitone occurrences across the professional subjects. Each plot shows the results for the six singers individually and the mean across all of the singers.

AUTHOR NOTES

This work was supported by the Center for Interdisciplinary 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. The authors acknowledge the encouragement of AIRS in this project.

BIOGRAPHIES

Johanna Devaney

Johanna Devaney completed her PhD in the Music Technology Area of the Department of Music Research at the Schulich School of Music of McGill University, under the supervision of Ichiro Fujinaga. While at McGill, Johanna also collaborated with Jonathan Wild and Peter Schubert in the Music Theory Area and was a student member of both the Centre for Research in Music Media and Technology (CIRMMT) and the International Laboratory for Brain, Music and Sound Research (BRAMS). Johanna holds an MPhil degree in Music Theory from Columbia University, where she worked with Fred Lerdahl and Daniel P. W. Ellis, as well a BFA in Music and History and an MA in Composition from York University in Toronto, where she worked with Michael Coghlan and David Lidov. She also taught at York for several years in the areas of Digital Music and Music Theory. Johanna's research is focused on studying and modeling performance practice. She has published her work in the *Journal of Interdisciplinary Music Studies* and *Ex Tempore* and presented at numerous international and national conferences including the International Conference on Music Perception and Cognition (ICMPC), the International Computer Music Conference (ICMC), the Conference on Interdisciplinary Musicology (CIM), and the annual meeting of the Society of Music Theory (SMT). As of July 2011, Johanna is a postdoctoral scholar at the Center for New Music and Audio Technologies (CNMAT) at the University of California at Berkeley, working with David Wessel.

Michael I. Mandel

Michael I. Mandel received the B.S. degree in Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, in 2004 and the M.S. and Ph.D. degrees in Electrical Engineering from Columbia University, New York, NY, in 2006 and 2010 in Daniel P. W. Ellis' Laboratory for the Recognition and Organization of Speech and Audio. He was a postdoctoral researcher in the Machine Learning laboratory at the Université de Montréal in Montréal, Québec with Yoshua Bengio and Douglas Eck. He has published on sound source separation, music similarity, and music recommendation. His research uses machine learning to model sound perception and understanding.

Daniel P.W. Ellis

Daniel P. W. Ellis received the Ph.D. degree in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, where he was a Research Assistant in the Machine Listening Group of the Media Lab. He spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA. Currently, he is an Associate Professor with the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing.

Ichiro Fujinaga

Ichiro Fujinaga is an Associate Professor and the Chair of the Music Technology Area at the Schulich School of Music at McGill University. He has Bachelor's degrees in Music/Percussion and Mathematics from the University of Alberta, and a Master's degree in Music Theory, and a Ph.D. in Music Technology from McGill University. In 2003-4, he was the Acting Director of the Center for Interdisciplinary Research in Music Media and Technology (CIRMMT). His research interests include music theory, machine learning, music perception, digital signal processing, genetic algorithms, and music information acquisition, preservation, and retrieval.