journal of interdisciplinary music studies
spring/fall 2008, volume 2, issue 1&2, art. #0821209, pp. 141-156

An Empirical Approach to Studying Intonation Tendencies in Polyphonic Vocal Performances

Johanna Devaney[1] and Daniel P.W. Ellis[2]

[1] Schulich School of Music, McGill University
[2] Department of Electrical Engineering, Columbia University

Background in music theory and analysis. Polyphonic vocal intonation practices have been addressed in a number of studies on vocal acoustics. Our research both builds on this work and supplements it with a theoretical paradigm based on work done in the areas of sensory consonance and tonal attraction.
Background in computing. Recent work in the field of music information retrieval has discussed the main obstacles related to tracking pitches in a polyphonic signal and has provided some techniques for working around these problems. Our method for analyzing the pitch content of recorded performances draws extensively on this work and on the knowledge made available to us by the musical scores of the pieces being performed.
Aims. Our research is focused on the study and modeling of polyphonic vocal intonation practices through the intersection of computational and theoretical approaches. We present a methodology that allows for a detailed model of this aspect of polyphonic vocal performance practice to be built from analyses of numerous recordings of real-world performances, while working within a robust theoretical paradigm.
Main contribution. In the computational component of the research a number of *a cappella* polyphonic vocal recordings are analyzed with signal processing techniques to estimate the perceived fundamental frequencies for the sung notes. These observations can be related to the musical context of the score through machine learning techniques to determine likely intonation tendencies for regularly occurring musical patterns. A major issue in developing a theory of intonation practices is the potential conflict between the vertical and horizontal intonational impetuses. To assess this conflict in greater detail we have constructed a theoretical approach where theories of sensory consonance account for vertical tuning tendencies and theories of tonal attraction account for the horizontal tendencies.
Implications. In the field of music cognition, our research relates to work being done in the area of musical expression. If the intonation tendencies inferred from the end results of this research are taken as a norm, then deviations from this norm, when these deviations are musically appropriate, can be viewed as expressive phenomena. Computer software implementing such results will allow composers and musicologists to hear more intonationally accurate digital re-creations and may also function as a training guide for vocalists.

Keywords: Intonation, singing, sensory consonance, melodic attraction, signal processing, machine learning

•*Correspondence:* Johanna Devaney, Schulich School of Music, McGill University, Canada; e-mail: devaney@music.mcgill.ca

Introduction

This paper presents a methodology for studying polyphonic vocal intonation practices. We recognize that the human voice, like non-fretted string instruments, is not locked into a single tuning system or temperament; rather the exact tuning of each pitch may vary each time it is sounded. The central assertion of this work is that a vocal ensemble's tuning cannot be consistently related to a single reference point; rather a combination of horizontal and vertical musical factors form the reference point for the tuning (Backus 1969; Barbour 1953). Assessing the relation between the horizontal and vertical factors is complicated by the strong probability that the weightings of these factors often differ in different musical contexts. We attempt to address these issues by employing a methodology that is built on the interaction of two distinct approaches. The first is theoretical, drawing on the various branches of music theory that address harmonic and voice leading practices, so-called musical forces and musical expectation, acoustics and psychoacoustics, and tuning, temperament, and intonation. The second is computational based on frequency analyses of a number of polyphonic vocal recordings. This computational approach will provide the model with an objective basis, while the theoretical approach will provide a contextual depth that empirical analysis cannot produce. We believe it is a strength of this methodology that the approaches inform one another throughout.

Historical Background

Our methodology builds and expands on centuries of theory about tuning and intonation practices and attempts to address these practices in greater detail. Questions of tuning have preoccupied a significant number of theorists from antiquity through the present. For some ancient Greeks, tuning was predominantly numerically based; the Pythagoreans, c. 500 BC, limited their definition of consonance to intervals corresponding to monochord divisions which employed only super-particular ratios of the numbers 1, 2, 3, and 4 (a series of numbers which, because they sum to ten, was known as the tectractys). This generated a system of pure octaves, fifths, and fourths, in which the thirds were significantly sharper than those occurring in the overtone series. In contrast, Aristoxenus, c. 350 BC, argued that the ear, rather than strict mathematics, should be the guide for determining consonance. Ptolemy, c. 120 AD, took the middle road between the Pythagorean numerically based methodology and the Aristoxenean aurally based methodology, contending that the Pythagorean approach was essentially correct but that it should be informed by aural perception. In the course of his study of harmonics, the aim of which was to address both physical and perceived musical phenomena, Ptolemy defined a seven-note diatonic scale system with a variety of tunings. In particular, his Syntonic Diatonic tuning system was influential on later tuning theory as it was, in effect, the first articulation of 5-limit Just Intonation, a system that is limited to those tunings that can be derived by fifth or larger divisions of a string.

The Pythagorean doctrine of limited consonance served the music of the early and middle centuries of the medieval era, where only the fifth, fourth and the octave were considered consonances. The rise of the third and the sixth as imperfect consonances in the late medieval period led to the use of tuning theories and systems in which these intervals sounded more agreeable, most notably 5-limit Just Intonation. The first explicit discussions concerning intonation arose around the same time; primarily because of the increased interest in keyboard tuning/temperament systems, which highlighted the difference between singers' tuning practices and fixed tuning of keyboard instruments.

The first attempt to systematically address the issue of singers' intonation practices, as well as to develop a keyboard instrument to emulate them, was Nicola Vicentino's *L'antica musica ridotta alla moderna prattica* (1555). Here Vicentino put forth two tuning systems for his 31-tone gamut; one of which explicitly is named "Tuning System for the Purposes of Accompanying Vocal Music" and which he claims, somewhat erroneously, provides pure fifths in every key; in reality a number of the fifths are slightly tempered. The other system produces striking similar results, but is conceived of as an augmented 1/4 comma meantone system. This is characteristic of a recurring conflict throughout late Renaissance and the Baroque, between the desire for idealized systems and the need for practicality in keyboard tuning. The conflict is later articulated in Zarlino's *Le istitutioni harmoniche* (1558), where both an idealized system using a 5/4 ratio for the Major Third and the need to systematically temper intervals when tuning string and keyboard instruments are discussed. Likewise in his *Harmonie universelle* (1636), Mersenne presents a mathematical proof for the size of an equal tempered semitone and argues that consonances are produced through the use of 1, 2, 3, 4, and 5 part divisions of the string, i.e. 5-limit Just Intonation.

Debates concerning the most appropriate keyboard temperament continued to dominate tuning theory for the next hundred and fifty years, until equal temperament eventually won out over the various meantone and well-temperament systems. In the eighteenth and nineteenth centuries, discussions of tuning ratios, and later overtones, were limited to music theoretic treatises considering arguments for consonance and dissonance within the context of equal temperament, most notably in Rameau's theory of harmony (1722, 1737) and Helmholtz's theory of *Konsonanz* (1863).

Contemporary Work

By the early twentieth century tuning theory had become something of a fringe interest. The notions of intonation practices were only rarely considered, as there was no reliable way of assessing exactly what pitches were being sung. Though there was some notable mid-twentieth century interest in intonational practices, or preferences, the relevant studies were in the tradition of the eighteenth and nineteenth century music theoretic approaches cited above. For example, Boomsliter and Creel (1961) attempted to develop a theory of melody based on their experiments with musicians'

preferences for various tuning systems on a monochord-like instrument. Some of the most forward-looking work was done in the 1920s and 30s by Seashore (1938) and his colleagues at the University of Iowa, who studied various aspects of singing performance, including vibrato and intonation.

An increased interest in the accurate performance of early music over the past forty years has prompted deeper investigations into historical tuning and temperament. Most of these studies are, however, a prescriptive endeavor; the musicians/singers are generally instructed of the ways in which they should modify their usual intonation practices in hopes of achieving a more historically accurate performance. Our work, in contrast, is descriptive in its attempt to create a model of common contemporary vocal ensemble intonation practices from actual performances. The descriptive nature of this study falls in line with small, but growing, number of studies that address questions related to intonational aspects of performance, particularly Fyk's large-scale study on violin intonation (1995) where she attributed gravitational attractions at work within the tonal system, and work done at the "Speech, Music, and Hearing" group at the Royal Institute of Technology in Stockholm (Sundberg 1987; Ternström and Sundberg 1988; Nordmark and Ternström 1996; Ternström 2002; Jers and Ternström 2005), as well as studies by Howard (2007), and Vurma and Ross (2007) on the intonation practices of singers, as assessed in laboratory experiments. The use of recordings also contextualizes this study in terms of actual performances, rather than contrived experiments.

Our study of polyphonic vocal intonation practices is aided by a number of these theoretical works, particularly those late Renaissance writings that addressed tuning issues in light of contemporary vocal practices. The large body of *a cappella* repertoire that complemented this theoretical tradition is a fruitful area from which to draw musical examples.

Music Theoretical Considerations

The major issue in theorizing intonation practices is the conflict between vertical and horizontal tendencies at any given point in time. We anticipate that the vertical aspects of the intonation practices will conform to the harmonic series, i.e. that the upper voices will coincide with the partials of the note sung by the bass note, and vice versa. With horizontal aspects there is no acoustical template to refer to; rather we are utilizing recent work on melodic attraction as the basis of our theory of the horizontal intonation tendencies.

Vertical Aspects

Our assumption that the vertical intonation tendencies will conform to the overtone series is rooted in Helmholtz's theory of consonance and dissonance (1863); where he postulated that the coincidence of a significant number of partials between two

pitches produced a consonance whereas the absence of such coincidence produced a dissonance. In his theory, the degree of coincidence can be determined by measuring the number of beats produced when the two tones are played simultaneously. Beating is produced by interference between tones of proximate frequency.

Helmholtz's work is also the foundation of relevant psychoacoustic theories which corroborate both our ear's ability to discern small variations in frequency and which support the idea that maximal coincidence of partials in a vertical sonority produces maximal consonance. Plomp and Levelt (1965) revisited some of Helmholtz's ideas through a series of tests on untrained subjects. They discovered that interval size, called the critical band, is also a significant factor in the perception of consonance. Their general results, using sine tones, indicated that their subjects judged intervals less than a minor third and greater than a unison as dissonant and intervals a minor third or greater as consonant. Their results also demonstrated that there is some variation based on the frequency range; the same interval in a lower frequency range was generally perceived as being less consonant than in a higher frequency range. Plomp and Levelt's theory also applied their critical band findings to the interactions between the partials of pairs of complex tones.

Terhardt (1984) expanded the work of Plomp and Levelt with a theory of consonance that reconciles psychoacoustic phenomena, which he termed sensory consonance, and tonal significance, or harmony. He aligns sensory consonance with Helmholtz's concept of *Konsonanz*, in this phenomenon a greater degree of consonance corresponds to a lesser amount of beating. According to Terhardt, beats that are slower than 20 Hz are audible; this phenomenon is observable when two instances of the same note are played slightly detuned. If the beating is faster than 20 Hz it is perceived as roughness, an aural sensation akin to rattling. Beating and roughness may occur both between the tones' fundamentals and their partials. The greater the degree of coincidence between the partials of the two tones the less rough, i.e. less dissonant, the resultant sound is. The theory of sensory consonance makes a case for purely tuned vertical intervals, as there is a greater coincidence of partials between them than with tempered intervals.

The second component of Terhardt's theory of consonance is the tonal, or harmonic, context. He aligns this component with both Helmholtz's theory of *Klangverwandtschaft* and his own virtual pitch theory. Both of these theories suggest that the perception of harmonic consonance in Western art music is dependent on the mind's acquisition of an acoustical template. In his virtual pitch theory, Terhardt argues that this template, based on the harmonic series, allows the listener to perceive the pitch of complex tones as being that of the fundamental, whether or not the fundamental is actually present. He expands this to harmonic consonance by arguing that the template acts as a reference point for determining whether or not the bass note of the current sonority corresponds to the virtual fundamental note that is suggested by the template. When the real bass note and the virtual fundamental note align, the sonority is perceived as consonant. The learning process associated with the acquisition of this template allow for the varying degrees of consonance, which

correspond with the different degrees of consonance that are typically assigned to different types of sonorities. Terhardt argues that the majority of this learning comes from exposure to the complex tones found in speech sounds; so although this learning impacts musical perception, its acquisition is predominantly non-musical. He uses this postulation to support a further argument that the basis of harmonic consonance, like sensory consonance, is psycho-acoustical rather than cultural.

This work can be applied to the issue of tuning preferences in vertical sonorities by considering the lower partials in the overtone series as more likely tuning possibilities than those found in the equal tempered system. These tuning tendencies more commonly emerge in sustained notes as the horizontal motion of the voices has less of an impact in such scenarios.

Horizontal Aspects

Theories of melodic attraction, particularly those put forth by Lerdahl (2001) and Larson (2002), offer tools with which to address the horizontal dimension. Lerdahl's model of melodic attraction (2001; Lerdahl and Krumhansl 2007) is a component of his *tonal pitch space* theory. The model formalizes the tendency of a dissonant pitch to resolve to a consonant neighbor (which may be a neighbor at either the chromatic, diatonic, or triadic level of his pitch space model) with a rule which observes both Bharucha's (1996) principles of proximity and stability and proceeds in part on an analogy with Newton's law of gravitation. The attraction of one pitch to another is the anchoring strength of the goal pitch (s_2, derived from a modified version of the model's "basic space") divided by the anchoring strength of the source pitch (s_1) times the inversion of the square of the number of semitones between the two pitches (s_2/s_1 x $1/n^2$). In this context Lerdahl discusses the asymmetries in attraction when moving from unstable pitches to stable ones and from stable pitches to unstable ones. These asymmetries demonstrate how the same interval functions differently in different musical contexts.

Larson (2002) posits a more complex calculation for the phenomenon of melodic attraction in his work on melodic forces, which is more explicitly focused on quantifying how listener's expectations are met or confounded by particular musical patterns. Larson's model correlates the forces of gravity, magnetism, and inertia explicitly within a single equation. Gravity is defined as the tendency of a musical line to go down, and is rooted in Lakoff and Johnson's (1980) notion of embedded metaphors. Magnetism, the tendency of unstable notes to move to stable ones, is, like Lerdahl's attractions, rooted in the psychological principles of proximity and stability. Inertia, the tendency of a musical line to continue rather than vary, is based on the Gestalt principle of good continuation. The culminative force acting on a note in a given context, or pattern, is calculated by summing the results of individual calculations for each force ($F = w_G G + w_M M + w_I I$). Gravity ($G$) is a binary, 1 or 0, as patterns can be assessed as either giving into gravity or not; i.e. if a pattern descends towards a more stable pitch it has a G value of 1, otherwise a G value of 0.

Magnetism ($M = 1/d_{to}^2 - 1/d_{from}^2$) is the inverse of the square of the distance in semitones from the initial note to the closest stable pitch ($1/d_{from}^2$) subtracted from the inverse of the square of the distance in semitones from the initial note to the goal note in the current musical context ($1/d_{to}^2$). Inertia (I) has a value of 1, 0, or -1, depending on whether the musical pattern has inertial potential and fulfils it, has no inertial potential, or has inertial potential but goes against it; i.e. if a pattern continues in the direction it started with it has an I value of 1, if it moves in the opposite direction it has an I value of -1, and if it stays on the same pitch it has an I value of 0. Larson uses the technique of multiple regression to find the weightings (w_G, w_M, and w_I) for multiple musical contexts.

Both of these models serve as a starting point for quantizing the musical intuitions that will form the basis of the theoretical model. Quantization of such intuitions is necessary to facilitate the interaction of the theoretical model with the computational one. In turn, the results of this study may be able to provide empirical data to help substantiate further investigations in this area. The melodic attraction models provide a means of exploring intonational tendencies in linear pitch sequences by providing representations of how the pitches within a tonal system exist in relation to one another outside of the harmonic context in which they occur. Lerdahl's model is particularly useful because it is internally consistent, thus it generates a full complement of attractional relations within a musical system. Larson's system has great potential but requires a certain amount of modification because in its current form it has a number of weaknesses: the inability of the model to accommodate a change in the governing tonic part-way through a musical sequence, the inability of the model to calculate attractions from a stable pitch, and the possibility of generating negative values with the equation, which renders comparisons between the different models difficult.

These theories are also a point of intersection with the field of music cognition, specifically the ways in which musical forces shape musical expectation (Narmour 1990; Margulis 2005). In particular, horizontal intonation practices often function as expressive phenomena; thus, they may be related not only to musical expectation but also to musical meaning or emotion, as it relates to performance. In his 1938 text, *Psychology of Music*, Seashore suggested that emotion is conveyed in performance through deviations from a norm. This idea was superseded by Meyer's theory of musical emotion (1956), which argued that emotional responses to music are rooted in the denial of the listener's expectations. Recent work by Gabrielsson (1995), Palmer (1996), and Sloboda (2005) reconsider the issue of musical emotion from the angle of performance through their systematic explorations of expression. This coincides with a rise in studies of intonation practices, in particular Fyk (1995), which have discussed the expressive aspects of intonation. Once our model has been trained with a large number of performances it will be able to provide some quantitative data about the typical deviations in performance. We are confident that our use of probabilistic machine learning techniques to train the model will minimize the influence of intonation 'errors' present in the performances. Deviations would be measured between individual recordings and the normalized results produced by the

model when it is queried with the same musical passage. These data can be correlated with the results of recent psychological experiments on musical emotion, which could serve as the basis of a more nuanced theory of how musical emotion is expressed in performance.

Once the melodic attraction models have been calculated they will suggest horizontal intonation tendencies. While they will not provide exact tuning predictions, the attraction models serve as a more appropriate reference than simply using values from an equal tempered system, particularly in terms of highlighting which notes are more likely to be detuned. The models generate synopses of a melodic line's attractional properties, which provide maps of points of where the melodic attraction rises and falls. The implication that we take from the attraction models is that points where there is more attraction are likely to provoke greater degrees of detuning, either in the sharp or flat direction, depending on the particular musical context. As these attraction patterns are context-dependent, some melodic lines will have more clearly defined tuning implications than others. This makes modelling of the horizontal tuning tendencies far more complex than the vertical ones, and the exact relationship between the attraction models and the intonation tendencies will only emerge through modelling of the data collected in the computational component of the model.

The reconciliation of the vertical and the horizontal is dependent on a number of factors, including the duration and metrical position of a given vertical sonority, its function within the musical context, and the significance of the horizontal lines moving through a vertical sonority, both in terms of the relationship of its pitch material to the current harmonic context and of the overall texture of the music at that point in time. The development of a theory of intonation from these parameters is complicated by the fact that the measurement of such factors may vary greatly under different musical contexts. At points where there are conflicting interpretations, the data-oriented model will provide additional insight into intonation practices. This will require careful querying of the computer-based model with examples that are informed by both the needs of the theoretical model and representative instances in the chosen repertoire. Thus, the range of examples encompasses both composed examples that highlight commonly occurring vertical/horizontal conflicts, such as cadential patterns, and repertoire excerpts that explore more specific tuning issues, such as the asymmetry between a melodic line descending from the tonic to the leading tone versus a melodic line ascending from the leading tone to the tonic.

Computing considerations

The goal of the computational component of the model is to use statistical machine learning techniques to build a model of polyphonic vocal intonation practices from the microtonal pitch variations between recorded performances and twelve-tone equal temperament. Twelve-tone equal temperament is a useful reference because it remains consistent in spite of changes to the tuning references. It is also the standard

system used both by music software and in Western art music practices as a whole, making it the most universal of any available reference points. Statistical machine learning techniques require a large amount of data, so our recent work has focused on developing a system that can automatically collect data from real recordings. Due to the challenges associated with accurately extracting pitch information from a polyphonic signal, the data collection is a two-step process. The first step is the not entirely trivial step of temporally aligning a MIDI score of the work to the audio recording. The second step, and the main obstacle, is developing a method to accurately extract pitch data from the polyphonic vocal recordings.

Data Extraction

In order to build a statistical model of singing intonation, we need to gather accurate pitch information from recordings of real performance. Our task is unlike most previously-studied pitch estimation problems for the following reasons:

a) we need to extract the pitch of multiple voices in polyphonic contexts
b) we can exploit prior knowledge of the intended (notated) pitch sequence
c) we are interested in recovering a single, effective perceived pitch per note, and we want to be able to measure tuning differences in that pitch that may be far smaller than a semitone.

Our solution involves using score information, for instance a MIDI version of the piece, to guide the pitch estimation. The main challenge of using a MIDI file as a frequency template in the pitch-tracking algorithm is that of temporal alignment. Just as the MIDI file provides equal-tempered tunings, it also provides rhythmic renditions according to a static meter. In order to serve as a reference, the temporal events in the MIDI file must be aligned with the temporal events in the audio file. There are a variety of techniques for automated MIDI score-polyphonic audio alignment (Hu, Dannenberg, and Tzanetakis 2003; Soulez, Rodet, and Schwarz 2003; Turetsky and Ellis 2003; Raphael 2004), which deal quite capably with a limited number of simultaneous instruments with well-defined onsets and timbres. There are three main challenges in aligning polyphonic vocal recordings: first, the note onsets are often difficult to determine, particularly when notes change under a single syllable; second, the very nature of a vocal ensemble means that all of the parts have roughly the same timbre, making it more difficult to distinguish between parts in certain musical circumstances; and third, there is often a considerable amount of reverberation present in the recordings. Our dynamic programming approach, derived from Turetsky and Ellis (2003), is in practice accommodating these issues. In particular, our application, in contrast to much of the previous motivation for score alignment, relies on accuracy in pitch, but is less sensitive to exact identification of onset times. We are also exploring algorithmic ways of computationally reducing the amount of reverberation in the recordings (Allen, Berkley, and Blauert 1977). Another potential issue arises when there are several voices singing a single part. Although the signal processing

implications of this are not well understood, we have demonstrated successful pitch extraction for data with multiple voices per part.

Once the MIDI file has been appropriately aligned to the audio file, we can use it to guide our search for relevant frequency information in the signal. The aligned score indicates the approximate time and frequency of each performed note. To obtain the high-resolution estimate of the perceived pitch we use an instantaneous-frequency (IF) based spectral analysis, which calculates a phase derivative (i.e. instantaneous frequency) within each time-frequency cell of a conventional short-time Fourier transform (Abe, Kobayashi, and Imai 1996). This gives us an estimate of the frequency of a sinusoid within each cell whose resolution is not limited to the quantized bin center frequencies of the underlying Fourier transform; the accuracy of this estimate, however, is limited by the amount of other energy (noise or harmonics) that may be present in the bin.

The IF spectrogram thus recovers the estimated energy and frequency of sinusoids at every time-frequency cell. We used 100ms windows with 75% overlap. A criterion that the IF in three spectrally-adjacent cells must differ by less than 25% is used to reject cells that are not dominated by a single sinusoid, since in these cases IF estimates will not be stable. After this, the matching procedure looks for sinusoid components within the time-span indicated by the aligned score, whose frequency is close to the notated pitch; the matching score decreases as the observed frequency moves away from the expected pitch. Only the fundamental harmonic of each note is identified and used for the pitch estimate, which is valid since the spectrum of the voice is purely harmonic. The matching also allows time frames to be associated to 'gaps' between every notated pitch if the best matching score falls below a threshold. Finally, each notated pitch is required to align to a minimum number of time frames; we found that a minimum duration of 4 successive windows (i.e. windows whose center points span 100ms) gave good results. In our examples, these settings appeared successful in identifying the fundamental components associated with the intended voices.

To estimate the single pitch value, we simply used an energy-weighted average of the instantaneous frequencies aligned to each note. This simple averaging is a reasonable approximation to the perceived pitch (Brown and Vaughn 1996), but is a potential weakness in our approach in the presence of vibrato, where pitch variation within each note can be far larger than the tuning nuances we are trying to measure.

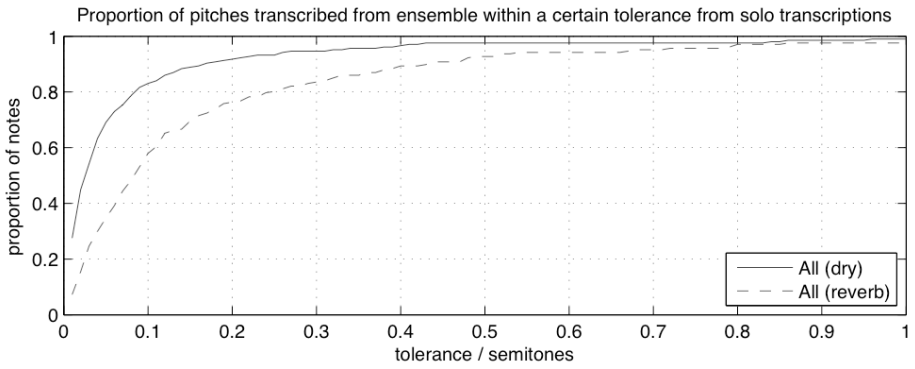

Figure 1. Results of instantaneous-frequency pitch estimation technique evaluation.

In order to evaluate our IF pitch estimation technique we ran an experiment on a multi-track recording of the "Kyrie" from Machaut's four-part *Messe de Notre Dame*, the results of which are shown in Figure 1. For this experiment we used the musical score as guide to identify note onsets and offsets in the recorded signal. Once all of the notes were labeled, we used our IF method to determine the fundamental frequencies in each of the monophonic vocal parts, using the energy-weighted averaging technique discussed above. We then applied the IF to two composite polyphonic versions of the recording, one dry and one with 0.9 second (RT_{60}) artificial reverb added. Figure 1 plots the difference between the pitch estimates of the 207 individual notes in the piece from the composite signal against the estimates from monophonic signal. Though there are some outliers, the overall results demonstrate that we are able to accurately estimate 95% of the notes in the dry composite signal to within 30 cents (less than a sixth of a tone). The presence of reverb complicates the calculation, such that only 58% of the composite notes are within 0.1 semitone of solo estimates versus 83% for the dry signal. However, the results for the composite signal with reverb increase quite sharply as the tolerance increase, up to 84% for 30 cents and over 90% for 50 cents.

Once we had confirmed that we could achieve reasonable results with our IF pitch-tracking technique we applied it, along with our score-audio alignment technique, to a live performance of Allegri's *Miserere*. Figure 2 is scores of the opening of the piece and Figure 3 is the corresponding spectrogram of the performance with the estimated fundamental frequencies overlaid in white. Figure 3 demonstrates effectiveness of our alignment algorithm for both homophonic and florid vocal contrapuntal textures recording in a reverberant environment. This is significant as the success of data analysis part of the computational component of the model is dependent on having a significant amount data available, more data than could reasonably be collected if all the audio files needed to be annotated by hand.

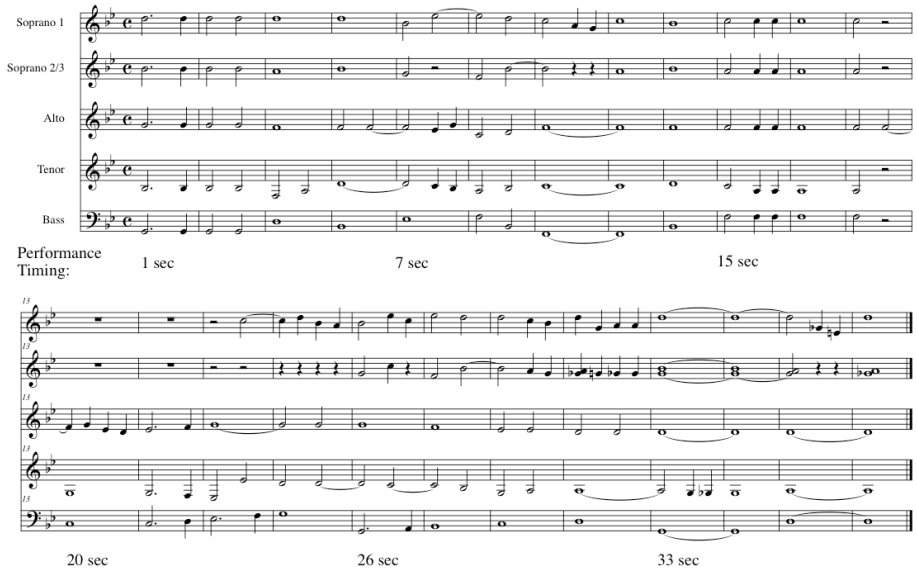

Figure 2. Opening of Allegri's *Miserere*.

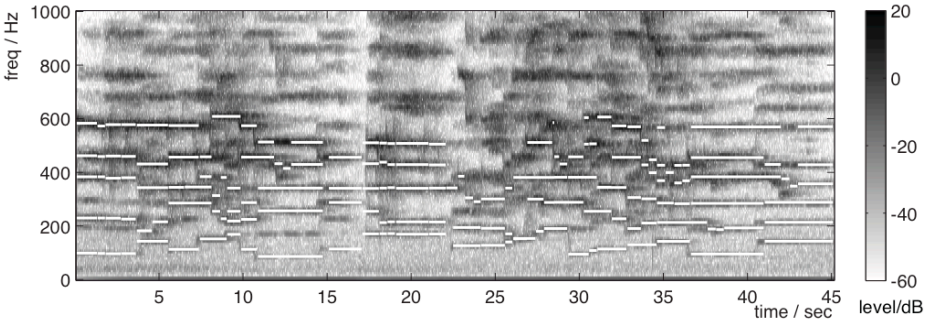

Figure 3. Spectrogram of the opening of Allegri's *Miserere*, white lines indicate estimated fundamental frequencies.

Data Analysis

Apart from its consistency in the challenging pitch estimation task, the appeal of automatic pitch extraction is that it holds the promise of obtaining virtually unlimited data with little manual involvement beyond providing paired audio and score input files. As the scale of the training data increases, the accuracy and complexity of the models that can be learned automatically from the data improves in correspondence.
By using automatically-inferred machine learning models, we will be able to test and quantify the influence and interaction of different candidate factors. One approach we

will investigate is the use of decision trees (Duda, Hart, and Stork 2000), where data items descend a tree from its root to reach a leaf node that determines the class (e.g. the predicted tuning) of that item. Each node of the tree is associated with a test on some aspect of the data item (for instance, questions about the constituents of the vertical or horizontal context). The algorithms for choosing the tests and building the tree typically consider all possible tests at each node, selecting the one that maximizes the information gained on the data used for training. The appeal of this approach is that the resulting structure can be examined to reveal which factors were discovered to be the most important, and how they interact.

Such a machine learning model, trained to emulate some of the intonation practices of a real-world vocal ensemble, can provide predictions against which various recordings of the same piece could be compared, as well as providing some empirical data for relevant music theories. Once such a model has been sufficiently trained to produce reliable results, i.e. results which fall within the range of possible tuning deviations for a given musical context observed in the data, it can be queried with specific musical circumstances. The results of these queries, designed in reference to the theoretical model described above, may either verify or encourage reconsideration of the theoretical model. The data analysis component of this project is conceived as an iterative process, initially to discover which machine learning technique is most appropriate to address this problem and later to refine both the specific application of the chosen technique and the components of the theoretical model.

Other Applications of the Model

Once this model has been implemented into a usable algorithm it will have numerous applications. In the academic world, it will allow composers and musicologists to hear accurate temperament re-creations and it has pedagogical potential as a training guide for vocalists. It also has applications in the area of digital music production as a means of rendering both MIDI and audio recordings more accurate in intonation. In terms of MIDI recordings, the model would be applied through the pitch-bend controller available in the MIDI protocol. This type of MIDI manipulation has a precedent in the 'groove-mapping' algorithms introduced in the mid-nineties, where the timing idiosyncrasies of real drummers were measured, modeled, and then applied in a type of quantization algorithm. It could also be used for intonational corrections in audio signals, e.g. autotuners. Though in this application the model would need to piggyback on one of the numerous methods that are able to achieve pitch-shifting without any time stretching or compression, and would initially be limited to single-voices or very simple polyphonic textures where the individual voices could easily be parsed.

Conclusion

This article has outlined the methodology for an ongoing study of polyphonic vocal intonation practices based on the interaction of music theoretical and computational approaches. The music theoretical approach draws on work done in the areas of sensory consonance and melodic attraction to assess tuning tendencies in the musical scores of recorded performances being studied, while the computational approach uses signal processing and machine learning techniques to analyze the recordings in terms of the actual performance tunings. Overall, the goal of this methodology is to produce a generalized model of some of the most common intonation tendencies in contemporary Western art music vocal ensembles. We have shown through two examples the feasibility of our method for extracting sufficiently accurate tuning information from polyphonic vocal recordings to address this complex performance practice. These computational innovations are particularly significant as there are no existing techniques that can reliably extract such information from polyphonic vocal recordings.

Acknowledgements

The authors would like to thank Fred Lerdahl for his input during the early stages of this project, Jonathan Wild and Peter Schubert for providing us with multi-tracked recordings of the Machaut's *Notre Dame Mass*, and Michael Mandel for his thoughtful suggestions and comments.

References

Abe, T., T. Kobayashi, and, S. Imai. 1996. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In *Proceedings of the International Conference on Spoken Language*. 1277-80.

Allen, J. B., D. A. Berkley, and J. Blauert. 1977. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *Journal of the Acoustical Society of America*. 62:4, 912-5.

Backus, J. 1969. *Acoustical foundations of music*. New York: W.W. Norton & Company Inc.

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

Bharucha, J. J. 1996. Melodic anchoring. *Music Perception*. 13:3, 383-400.

Boomsliter, P., and W. Creel. 1961. The long pattern hypothesis in harmony and hearing. *Journal of Music Theory*. 5:2, 2-30.

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

Duda, R., P. Hart, and D. Stork. 2000. *Pattern classification*. 2nd Edition. New York: Wiley Interscience.

Fyk, J. 1995. *Melodic intonation, psychoacoustics, and the violin*. Zielona Gura: Organon Publishing House.

Gabrielsson, A. 1995. Expressive intention and performance. In R. Steinberg, ed. *Music and the mind machine*. Berlin: Springer-Verlag. 35–47.
Helmholtz, H. 1863. *On the sensation of tone as a physicological basis for the theory of music*. Trans. by A.J. Ellis. 1954. New York: Dover Publications.
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, and G. Tzanetakis. 2003. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 185-8.
Jers, H., and S. Ternstrom. 2005. Intonation analysis of a multi-channel choir recording. *TMH-QPSR Speech, Music and Hearing: Quarterly Progress and Status Report*. 47, 1-6.
Lakoff, G., and Johnson, M. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.
Larson, S. 2002. Musical forces and melodic expectations: Comparing computer models with experimental results. *Music Perception*, 21:4, 457-98.
Lerdahl, F. 2001. *Tonal pitch space*. Oxford: Oxford University Press.
Lerdahl, F., and C. Krumhansl. 2007. Modeling tonal tension. *Music Perception*. 24:4, 329-66.
Margulis, E. 2005. A model of melodic expectation. *Music Perception*, 22:4, 663-714.
Meyer, L. 1956. *Emotion and meaning in music.* Chicago: University of Chicago Press.
Narmour, E. 1990. *The analysis and cognition of basic melodic structures.* Chicago: University of Chicago Press.
Mersenne, M. 1636. *Harmonie universelle*. Trans. Roger E. Chapman. 1957. The Hague, M. Nijhoff.
Nordmark, J., and S. Ternstrom. 1996. Intonation preferences for major thirds with non-beating ensemble sounds. *TMH-QPSR Speech, Music and Hearing: Quarterly Progress and Status Report*. 1, 57-61.
Plomp, R., and W. J. M. Levelt. 1965. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*. 38:4, 548-60.
Palmer, C. 1996. Anatomy of a performance: Sources of musical expression. *Music Perception*. 13, 433-53.
Rameau, J.-P. 1722. *Traité de l'harmonie* [Treatise on Harmony]. Trans. P. Gossett. 1971. New York: Dover.
Rameau, J.-P. 1737. *Generation harmonique.* Trans. D. Hayes. 1968. In *Rameau's theory of harmonic generation: An annotated translation and commentary of Génération harmonique*. Standford University: Ph.D. dissertation.
Raphael, C. 2004. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the Fifth International Conferernce on Music Information Retrieval*. 387-94.
Soulez, F., X. Rodet, and D. Schwarz. 2003. Improving polyphonic and poly-instrumental music to score alignment. In *Proceedings of the Fourth International Conference on Music Information Retrieval*. 143-8.
Seashore, C. E. 1938. *Psychology of music*. New York: Dover Publications.
Sloboda, J. 2005. *Exploring the musical mind: Cognition, emotion, ability, function.* New York: Oxford University Press.
Sundberg, J. 1987. *The science of the singing voice*. Dekalb, IL: Northern Illinois University Press.
Terhardt, E. 1984. The concept of musical consonance: A link between music and psychoacoustics. *Music Perception*. 1:3, 276-295.
Ternstrom, S., and Sundberg, J. 1988. Intonation precision of choir singers. *Journal of the Acoustical Society of America*. 84:1, 59-69.

Ternstrom, S. 2002. Choir acoustics – an overview of scientific research published to date. *TMH-QPSR Speech, Music and Hearing: Quarterly Progress and Status Report*. 43, 1-8.

Turetsky, R., and D. P. W. Ellis. 2003. Ground-truth transcriptions of real music from force-aligned MIDI syntheses. In *Proceedings of the Fourth International Conference on Music Information Retrieval*. 135-41.

Vurma, A., and J. Ross. 2007. Interval tuning in two-part singing. Poster presented at the *3rd Conference on Interdisciplinary Musicology*, Tallinn, Estonia, August 15-19, 2007.

Vicentino, N. 1555. *Antica musica ridotta alla modern prattica* [Ancient music adapted to modern practice]. Trans. by M.R. Maniates. 1996. New Haven: Yale University Press.

Zarlino, G. 1558. *Institutione harmoniche*, 4a pt. [On the modes]. Trans. by V. Cohen. 1983. New Haven: Yale University Press.