

An empirical study of the influence of musical context on  
intonation practices in solo singers and SATB ensembles

Johanna Devaney  
Music Technology Area  
Department of Music Research  
Schulich School of Music  
McGill University

Submitted February 2011

A thesis submitted to McGill University in partial fulfilment  
of the requirements of the degree of Doctor of Philosophy

© Johanna Devaney 2011

(This page intentionally left blank)

## Abstract

Intonation in singing is a complex phenomenon that has received only limited empirical attention in the literature to date. Previous studies have observed that singers do not conform to either equal temperament or any other fixed-intonation system. However, none of these studies have explored whether singers' intonation practices are systematically related to musical context. The main objective of this dissertation is to examine the ways in which intonation is related to musical context, in both solo singing and in one voice-per-part SATB ensembles. The intonation-related data was extracted automatically from the recordings using a score-audio alignment algorithm optimized for the singing voice that was developed for this dissertation. Fundamental frequency estimates were made using an existing algorithm and a single perceived pitch for each note was obtained by taking a weighted mean of these estimates. This dissertation also uses the discrete cosine transform to explore novel ways to describe how the fundamental frequency changes over the duration of a note (for example, if singers scoop up into certain notes rather than others).

The dissertation consists of two experiments. The first uses two groups of singers (undergraduate vocal majors and professionals) to study the role of musical context in the intonation of melodic semitones and whole tones. The impact of accompaniment is also explored. Each singer performed the test piece, Schubert's "Ave Maria," three times a cappella and three times with recorded accompaniment. The participants in the second experiment are one SATB ensemble of semi-professional singers and two SATB ensembles of professional singers. This experiment explores the ways in which observable melodic and vertical intonation tendencies are influenced by the organization of musical context. Specifically, the intonation of semitones and whole tones is compared when they occur between different scale degrees and vertical tunings are examined in different harmonic contexts. The repertoire for the second experiment consists of several short composed exercises where the melodic intervals occur over a range of harmonic functions, a chord progression by Benedetti, and Praetorius' "Es ist ein Ros entsprungen."

## Sommaire

L'intonation du chant est un phénomène complexe qui a été peu étudié jusqu'à présent. Les évaluations empiriques qui ont eu lieu indiquent que les chanteurs ne se conforment pas à un tempérament égal, ni à aucun autre système d'intonation fixe, mais ces études n'ont pas cherché à déterminer si les pratiques d'intonation des chanteurs sont reliées de façon systématique au contexte musical. L'objectif principal de cette thèse est d'examiner les façons dont l'intonation est liée à un contexte musical, à la fois pour des chanteurs solistes et pour des choeurs SATB (une seule voix par partie). Les données liées à intonation ont été extraites automatiquement à partir des enregistrements des chanteurs en utilisant un algorithme d'alignement et suivi de partition optimisé pour les voix chantées et développé spécifiquement pour cette thèse. Les estimations de fréquences fondamentales ont été obtenues via un algorithme existant, et une unique hauteur perçue a été calculée, pour chaque note, par moyenne pondérée de ces estimations. Cette thèse utilise également la transformée en cosinus discrète afin d'explorer de nouveaux moyens de décrire la manière dont la fréquence fondamentale varie pendant la durée d'une note ; par exemple, si un chanteur monte plus pendant certaines notes que d'autres.

La thèse se compose de deux expériences. La première fait appel à deux groupes de chanteurs (des étudiants en chant de premier cycle universitaire et des chanteurs professionnels) pour étudier le rôle que joue le contexte musical dans l'intonation des tons et demi-tons mélodiques. L'impact de l'accompagnement musical est aussi étudié. Chaque participant a chanté « Ave Maria » de Schubert trois fois a cappella et trois fois avec un accompagnement enregistré. Les participants à la seconde expérience faisaient partie d'un ensemble SATB semi-professionnel et de deux ensembles SATB professionnels. Cette expérience étudie les manières dont le contexte musical influence les tendances d'intonation observables mélodiques et verticales. Plus précisément, les intonations des demi-tons et des tons sont comparées en différents degrés de la gamme. De plus, les intonations des intervalles verticaux sont examinées dans différents contextes harmoniques. Le répertoire utilisé pour la seconde expérience est constitué de plusieurs courts exercices composés, au cours desquels l'intervalle mélodique apparaît sur une gamme de fonctions harmoniques, d'une progression d'accords de Benedetti, et du « Es ist ein Ros entsprungen » de Praetorius.

## Table of Contents

Chapter 1: Introduction .....	1
1.1 Motivation.....	1
1.2 Prior Research .....	2
1.3 Methodology.....	3
1.4 Contributions.....	5
1.5 Structure .....	7
Chapter 2: Literature Review .....	9
2.1 Tuning Theory.....	10
2.1.1 Overview of Tuning Systems .....	10
2.1.2 History of Tuning and Temperament.....	13
2.2 Performance Analysis.....	17
2.2.1 General Overview.....	17
2.2.2 Extraction of Performance Data .....	18
2.2.3 Studies of Intonation and Vibrato in Instruments .....	19
2.2.3.1 Fyk's work on Melodic Intonation in the Violin .....	21
2.2.4 Physiology and Acoustics of the Singing Voice .....	22
2.2.5 Studies of Intonation and Vibrato in the Singing Voice.....	24
2.2.5.1 Solo Voice.....	24
2.2.5.1.1 Research at KTH .....	25
2.2.5.1.2 Other Research .....	26
2.2.5.1.3 Vocal Ensembles .....	27
2.2.6 Modeling Expressive Performance.....	30
2.2.7 Summary.....	32
2.3 Pitch and Consonance Perception.....	33
2.3.1 Overview of the Mechanisms of the Auditory System.....	33
2.3.2 Pitch Perception.....	36
2.3.2.1 Historical Overview of Pitch Perception .....	37
2.3.2.2 Contemporary Spectral and Temporal Theories of Pitch Perception.....	39
2.3.3 Perception of the Pitch of a Single Tone .....	42
2.3.4 Consonance Perception.....	44
2.3.4.1 Sensory Consonance .....	45
2.3.4.2 Musical Consnance .....	48
2.3.4.3 Consonance and Tuning.....	49
2.3.5 Summary.....	50
2.4 Fundamental Frequency Estimation and Transcription .....	51
2.4.1 Monophonic Estimation .....	51
2.4.2 Polyphonic Estimation .....	53
2.4.3 Transcription of the Singing Voice .....	55
2.4.4 Summary.....	58
2.5 Audio-Score Alignment.....	58
2.5.1 Annotating Note Locations .....	58
2.5.2 Early Score Following.....	59
2.5.2.1 Dannenberg .....	60
2.5.2.2 Vercoe and IRCAM .....	60
2.5.2.3 Second-generation Score Followers .....	61
2.5.2.4 Tracking a Vocal Performer.....	61
2.5.3 Techniques.....	62
2.5.3.1 Dynamic Programming and Time Warping.....	62
2.5.3.1.1 MIDI Data .....	63
2.5.3.1.2 Acoustic Features .....	62
2.5.3.1.3 Spectral Decomposition .....	66
2.5.3.2 Hidden Markov Models and Related Techniques.....	67
2.5.3.2.1. Single-level HMMs .....	68

2.5.3.2.2 Multi-level HMMs .....	68
2.5.3.2.3 Graphical Models .....	69
2.5.3.2.4 Support Vector Machines (SVMs).....	71
2.5.4 Applications .....	71
2.5.4.1 Expressive Performance Studies.....	71
2.5.4.2 Automatic Accompaniment.....	72
2.5.4.3 Query-by-Humming.....	73
2.5.4.4 Digital Music Libraries.....	74
2.5.4.5 Other Applications .....	75
2.5.4.5.1 Karaoke .....	75
2.5.4.5.2 Education .....	75
2.5.4.5.3 Signal Processing .....	75
2.5.5 Evaluation .....	76
2.5.5.1 Ground truth .....	76
2.5.5.2 Formal evaluation .....	77
2.5.5.3 Required Accuracy.....	78
2.5.6 Summary.....	79
Chapter 3: Automatic Extraction of Performance Parameters.....	81
3.1 Annotation of Audio Files with Score-Audio Alignment .....	81
3.1.1 Evaluation of Dynamic Time Warping Approaches to Alignment.....	85
3.1.1.1 Ground Truth Collection .....	85
3.1.1.2 Test Data.....	87
3.1.1.3 Evaluation Framework .....	87
3.1.1.4 Experiment One: Comparison of DTW Approaches .....	88
3.1.1.5 Experiment Two: Evaluation fo Orio and Schwarz Under Different Conditions.....	90
3.1.2 Improving Alignment Accuracy.....	94
3.1.2.1 Acoustical Properties of the Singing Voice .....	94
3.1.2.2 HMM Details .....	96
3.1.2.3 Evaluation.....	99
3.1.2.4 Discussion .....	101
3.1.2.5 Conclusions .....	102
3.1.3 Summary .....	103
3.2 Modeling Perceived Pitch and the Evolution of Fundamental Frequency .....	103
3.2.1 Extracting Fundamental Frequency Information .....	103
3.2.2 Describing the Perceived Pitch .....	105
3.2.3 Evolution of Fundamental Frequency .....	107
3.2.4 Summary .....	122
Chapter 4 Intonation Experiments.....	125
4.1 Intonation in Solo Singing .....	125
4.1.1 Method .....	128
4.1.1.1 Participants.....	128
4.1.1.2 Apparatus .....	128
4.1.1.3 Procedure .....	128
4.1.1.4 Analytical Statistics .....	129
4.1.2 Results.....	132
4.1.2.1 Semitones .....	138
4.1.2.1.1 Interval Size .....	138
4.1.2.1.2 Slope and Curvature.....	145
4.1.2.2 Whole tones .....	164
4.1.2.2.1 Interval Size .....	165
4.1.2.2.2 Slope and Curvature.....	171
4.1.3 Discussion.....	190
4.1.3.1 Semitones .....	190
4.1.3.2 Whole Tones .....	192
4.1.3.3 Slope and Curvature.....	194
4.2 Intonation in SATB Ensemble Singing .....	197
4.2.1 Method .....	205

4.2.1.1 Participants.....	205
4.2.1.2 Apparatus .....	205
4.2.1.3 Procedure .....	206
4.2.2 Results .....	207
4.2.2.1 Part One: Semitone Exercises .....	207
4.2.2.2 Part Two: Whole tone Exercises.....	212
4.2.2.3 Part Three: Benedetti Chord Progression .....	216
4.2.2.4 Part Foul: Praetorius' "Es ist ein Ros' entsprungen" .....	226
4.2.3 Discussion.....	234
4.2.3.1 Semitones .....	234
4.2.3.2 Whole tones .....	235
4.2.3.3 Vertical Intervals .....	236
4.2.3.4 Influence of Syllable .....	239
4.3 Analysis of Both Experiments' Data.....	240
4.3.1 Relationship to Formal Tuning Systems .....	240
4.3.2 Influence of Intervallic Direction on Interval Size.....	241
4.3.3 Role of Musical Context.....	242
4.3.4 Individual Variation Amongst Singers.....	245
4.3.5 Impact of Training and Experience .....	249
4.3.6 R <sup>2</sup> Values in the Regressions .....	250
4.3.7 Conclusions .....	251
Chapter 5 Conclusions.....	253
5.1 Summary of Dissertation .....	253
5.2 Summary of Original Contributions .....	254
5.3 Future Research.....	254
5.3.1 More Controlled Experiments .....	254
5.3.2 Perceptual Questions .....	255
5.3.3 Improving the DCT .....	255
5.3.4 Other Ways of Analyzing the Music .....	256
5.3.5 Intonation and Expression .....	256
5.3.6 Intonation in Non-Western and Popular Music .....	257
5.3.7 Improving the Annotation Tool .....	257
5.3.8 Examination of Existing Recordings .....	258
5.3.9 Modeling Expressive Performance .....	258
References .....	259

## List of Figures

Figure 2.1.1: Some of the interval tunings that can be derived from the first sixteen partials of the overtone series.....	10
Figure 2.1.2: Circle of Fifths .....	12
Figure 2.3.1: Overtone series from the note A1 (55 Hz).....	34
Figure 2.3.2: Schematic of the human ear ( <i>Helix: Structures of the Human Ear</i> . 1997).....	35
Figure 2.3.3: Schematic of the basilar membrane ( <i>Hearing: Basilar Membrane</i> . 1997).....	36
Figure 2.3.4: Pitch processing in the auditory system.....	37
Figure 2.3.5: Plots of consonance as a function of frequency from Plomp and Levelt (1965), as reprinted in Rasch (1999).....	46
Figure 2.3.6: Plots of consonance as a function of frequency as reported by Kameoka and Kuriyagawa (1969a, 1969b) .....	47
Figure 2.5.1: A dynamic time warping similarity matrix.....	63
Figure 3.1.1: Spectrographic representations of 7-second audio clips of solo drum (left) and solo voice (right). .....	82
Figure 3.1.2: Plot results from MIREX 2007 Onset Detection evaluation for solo drum from Downie (2007).....	83
Figure 3.1.3: Plot of the results from MIREX 2007 Onset Detection evaluation for singing voice from Downie (2007). .....	84
Figure 3.1.4: Time-domain representation of audio in Audacity with note onsets and offsets labelled underneath.....	86
Figure 3.1.5: Frequency-domain representation of audio in Audacity with note onsets and offsets labelled underneath.....	86
Figure 3.1.6: Score of the opening of Machaut's <i>Notre Dame Mass</i> .....	87
Figure 3.1.7: Condition 1: Overlay of alignment of a single line aligned to a single voice. ....	92
Figure 3.1.8: Condition 2: Overlay of alignment for all four lines aligned simultaneously to a composite signal. ....	93
Figure 3.1.9: Condition 2: Example of a performance asynchrony for a notated simultaneity. ....	93
Figure 3.1.10: Condition 3: Overlay of alignment for a single line aligned to a composite signal of all the voices. ....	93
Figure 3.1.11: Time domain representation of a sung note's waveform.....	95
Figure 3.1.12: Three-state basic state sequence seed. ....	97
Figure 3.1.13: Basic state sequence seed plus breath. ....	97
Figure 3.1.14: State sequence adapted to sung text. ....	97
Figure 3.1.15: Visualization of the DTW alignment implemented as a prior for the HMM.....	98
Figure 3.1.16: Gaussian distributions for the creation of a prior from the DTW alignment.....	99
Figure 3.1.17: Visualization of the performance of the Algorithm Two-A versus the DTW alignment.....	101
Figure 3.2.1: Example of a spectrographic representation of quasi-monophonic recording from the ensemble experiment in Section 4.2.....	104
Figure 3.2.2: F <sub>0</sub> trace of a single sung note with the robust geometric mean and the weighted mean overlaid. ....	106
Figure 3.2.3: Example of how vertical interval size is calculated. ....	107

Figure 3.2.4: Moving averages for a long note (4.0 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms).....	108
Figure 3.2.5: Moving averages for a medium long note (2.1 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms).....	109
Figure 3.2.6: Moving averages for a medium short note (0.77 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms).....	109
Figure 3.2.7: Moving averages for a short note (0.48 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms).....	110
Figure 3.2.8: Examples of DCT coefficients for simple signals. ....	114
Figure 3.2.9: Comparison of the 1 <sup>st</sup> and 2 <sup>nd</sup> DCT coefficients calculated from the F <sub>0</sub> trace of the 4 s note in Figure 3.2.4 and those calculated from a moving average of the F <sub>0</sub> trace with a window size of 200 ms.....	117
Figure 3.2.10: Discrete cosine transform on a 200 ms moving average of the F <sub>0</sub> trace of a long note (4.0 s)...	119
Figure 3.2.11: Discrete cosine transform on a 200 ms moving average of the F <sub>0</sub> trace of a long note (2.1 s)...	120
Figure 3.2.12: Discrete cosine transform on a 200 ms moving average of the F <sub>0</sub> trace of a medium short note (0.77 s). ..	121
Figure 3.2.13: Discrete cosine transform on a 200 ms moving average of the F <sub>0</sub> trace of a short note (0.48 s). ....	122
Figure 4.1.1: Schubert’s “Ave Maria” with analyzed semitone categories marked.....	127
Figure 4.1.2: Schubert’s “Ave Maria” with analyzed whole tone categories marked.....	127
Figure 4.1.3 Comparison of the opening and closing statements of “Ave Maria” for non-professional singers 1–3. ....	133
Figure 4.1.4: Comparison of the opening and closing statements of “Ave Maria” for non-professional singers 4–6. ....	134
Figure 4.1.5: Comparison of the opening and closing statements of “Ave Maria” for professional singers 1–3. ....	135
Figure 4.1.6: Comparison of the opening and closing statements of “Ave Maria” for professional singers 4–6. ....	136
Figure 4.1.7: Box and whisker plots of semitone interval sizes across all non-professional singers.....	142
Figure 4.1.8: Box and whisker plots of semitone interval sizes across all professional singers. ....	143
Figure 4.1.9: Box and whisker plots of the semitone size in cents for each semitone condition across all non-professional singers. ....	144
Figure 4.1.10: Box and whisker plots of the semitone size in cents for each semitone condition across all professional singers. ....	144
Figure 4.1.11: Box and whisker plots of the 1 <sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the semitones performed by the non-professional group. ....	149
Figure 4.1.12: Box and whisker plots of the 1 <sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the semitones performed by the professional group. ....	150
Figure 4.1.13: Box and whisker plots of the 1 <sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 150 ms of the F <sub>0</sub> trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the semitones performed by the non-professional group. ....	151
Figure 4.1.14: Box and whisker plots of the 1 <sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 150 ms of the F <sub>0</sub> trace smoothed by applying a 200 ms moving average of the first note of all of the semitones performed by the professional group. ....	153

Figure 4.1.15: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of all of the semitones in each condition across all of the non-professional singers. ....	153
Figure 4.1.16: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of all of the semitones in each condition across all of the professional singers. ....	154
Figure 4.1.17: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the semitones performed by the non-professional group. ....	155
Figure 4.1.18: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the semitones performed by the professional group. ....	156
Figure 4.1.19: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 150 ms of the F <sub>0</sub> trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the semitones performed by the non-professional group. ....	157
Figure 4.1.20: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 150 ms of the F <sub>0</sub> trace (smoothed by results of applying a 200 ms moving average of the first note) of all of the semitones performed by the professional group. ....	158
Figure 4.1.21: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of all of the semitones in each condition across all of the non-professional singers. ....	159
Figure 4.1.22: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of all of the semitones in each condition across all of the professional singers. ....	160
Figure 4.1.23: Box and whisker plots of whole tone interval sizes across all non-professional singers. Each subject is represented individually on the x-axis, as well as the combination of all of the subjects. ....	167
Figure 4.1.24: Box and whisker plots of whole tone interval sizes across all professional singers. Each subject is represented individually on the x-axis, as well as the combination of all of the subjects. ....	168
Figure 4.1.25: Box and whisker plots of the whole tone size in cents for each whole tone condition across all non-professional singers. ....	169
Figure 4.1.26: Box and whisker plots of the whole tone size in cents for each whole tone condition across all professional singers. ....	170
Figure 4.1.27: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the whole tones performed by the non-professional group. ....	178
Figure 4.1.28: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the whole tones performed by the professional group. ....	179
Figure 4.1.29: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the last 150 ms of the F <sub>0</sub> trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the whole tones performed by non-professional group. ....	180
Figure 4.1.30: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient, approximating slope, run on the last 150 ms of the F <sub>0</sub> trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the whole tones performed by the professional group. ....	181
Figure 4.1.31: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of each whole tone interval for each condition across all non-professional singers. ....	182

Figure 4.1.32: Box and whisker plots of the 1 <sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of each whole tone interval for each condition across all professional singers.....	183
Figure 4.1.33: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the whole tones performed by the non-professional group. ....	184
Figure 4.1.34: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the F <sub>0</sub> trace of the first note of all of the whole tones performed by the professional group. ....	185
Figure 4.1.35: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient, (approximating curvature) run on the last 150 ms of the F <sub>0</sub> trace (smoothed by applying a 200 ms moving average of the first note) of the first note of all of the whole tones performed by the non-professional group. ....	186
Figure 4.1.36: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature,) run on the last 150 ms of the F <sub>0</sub> trace (smoothed by applying a 200 ms moving average) of the first note of all of the whole tones performed by the professional group. ....	187
Figure 4.1.37: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of each whole tone interval for each condition across all non-professional singers. ....	188
Figure 4.1.38: Box and whisker plots of the 2 <sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of each whole tone interval for each condition across all professional singers. ....	189
Figure 4.2.1: Score for Part One, Progressions 1–18. ....	198
Figure 4.2.2: Score for Part One, Progressions 19–27. ....	199
Figure 4.2.3: Score for Part Two, Progressions 1–18. ....	200
Figure 4.2.4: Score for Part Three, a chord progression by Benedetti. ....	201
Figure 4.2.5: Score for Part Four, Praetorius’ “Es ist ein Ros’ entsprungen,” with semitones marked. ....	202
Figure 4.2.6: Score for Part Four, Praetorius’ “Es ist ein Ros’ entsprungen,” with whole tones marked. ....	203
Figure 4.2.7: Score for Part Four, Praetorius’ “Es ist ein Ros’ entsprungen,” with vertical intervals marked... ....	204
Figure 4.2.8: Box and whisker plots for the semitones interval sizes in Part One across each condition, separated by ensemble.....	209
Figure 4.2.9: Box and whisker plots for the semitone interval sizes in Part One for each singer in each ensemble.....	210
Figure 4.2.10: Box and whisker plot of interval sizes for the vertical intervals in Part One between the bass and the voice singing the melodic semitone interval being evaluated.....	210
Figure 4.2.11: Box and whisker plots for the whole tone interval sizes in Part Two across each condition, separated by ensemble.....	214
Figure 4.2.12: Box and whisker plots for the whole tone interval sizes in Part Two for each singer in each ensemble.....	214
Figure 4.2.13: Box and whisker plot of interval sizes for the vertical intervals between the bass and the voice singing the melodic whole tone interval being evaluated in Part Two. ....	215
Figure 4.2.14: Theoretical tuning for Benedetti progression used in Part Three. ....	217
Figure 4.2.15: Ensembles used in Part Three. ....	218
Figure 4.2.16: Summary of the amount of drift in each ensemble’s renditions of the Benedetti’s chord progression used in Part Three. ....	219

Figure 4.2.17: Box and whisker plots for the whole tones interval sizes in Part Three across all the singers for each ensemble.....	221
Figure 4.2.18: Box and whisker plot of interval sizes for the vertical intervals in Part Three between all of the singers in each ensemble. ....	223
Figure 4.2.19: Box and whisker plots for the sizes of the melodic intervals in Part Four for each ensemble. .	229
Figure 4.2.20: Box and whisker plots for the vertical intervals data in Part Four for each ensemble. ....	230

## List of Tables

Table 2.1.1: Interval sizes in Pythagorean, 5-limit Just Intonation, 1/4-comma Meantone, and equal temperament.....	11
Table 2.1.2 Comparison of the impact of modulation around the circle of fifths in 5-limit.....	13
Table 3.1.1: Mean, minimum, and maximum difference in calculated onset and offset values in the alignment from the ground truth in ms for each of the algorithm evaluated in Experiment One.....	89
Table 3.1.2: Tallies of number of onsets and offsets estimated by the algorithms that are within 50 ms of the ground truth.....	89
Table 3.1.3: Tallies of number of onsets and offsets estimated by the algorithms that are within 100 ms of the ground truth.....	90
Table 3.1.4: The number of onsets and offsets predicted within 100 ms of the ground truth.....	92
Table 3.1.5: Mean and standard deviation in seconds between the onset and offset set alignments and the ground truth.....	92
Table 3.1.6: Results from Algorithms One and Two compared to the original dynamic time warping alignment in milliseconds.....	100
Table 3.2.1: This table shows thirteen simple signals with different curves and slopes—five straight lines, four parabolic curves, and four other curves—and the values for the 1 <sup>st</sup> and 2 <sup>nd</sup> DCT coefficients for each signal.	115
Table 3.2.2: This table shows thirteen simple signals that mirror the signals in Table 3.2.1 with the addition of sinusoids, as well as the values for the 1 <sup>st</sup> and 2 <sup>nd</sup> DCT coefficients for each signal.....	116
Table 4.1.1: Summary of the conditions evaluated in the regression analyses performed in this section.....	131
Table 4.1.2: Summary of the mean interval size and standard deviation in cents for the two subject groups across all of the semitones and whole tones used in this experiment .....	137
Table 4.1.3: Mean and standard deviation of the semitone sizes in cents in the non-professional group.....	139
Table 4.1.4: Mean and standard deviation of the semitone sizes in cents in the professional group. ....	139
Table 4.1.5: Summary of the means and standard deviations of the slope and curvature for the two subject groups across all of the semitones used in this experiments.....	145
Table 4.1.6: Non-professional group's semitone slope values calculated on the original F <sub>0</sub> trace.....	146
Table 4.1.7: Non-professional group's semitone slope values calculated on the F <sub>0</sub> trace with a moving average applied to it. ....	146
Table 4.1.8: Professional group's semitone slope values calculated on the F <sub>0</sub> trace.....	146
Table 4.1.9: Professional group's semitone slope values calculated on the F <sub>0</sub> trace with a moving average applied to it. ....	146
Table 4.1.10: Non-professional group's semitone curvature values calculated on the F <sub>0</sub> trace.....	147
Table 4.1.11: Non-professional group's semitone curvature values calculated on the F <sub>0</sub> trace with a moving average applied to it. ....	147
Table 4.1.12: Professional group's semitone curvature values calculated on the F <sub>0</sub> trace. ....	147
Table 4.1.13: Professional group's semitone curvature values calculated on the F <sub>0</sub> trace with a moving average applied to it. ....	147
Table 4.1.14: Mean and standard deviation of the whole tone sizes (in cents) for each of the tested whole tone conditions for the non-professional singers.....	165
Table 4.1.15: Mean and standard deviation of the whole-tone sizes (in cents) for each of the tested whole tone conditions for the professional singers. ....	165

Table 4.1.16 Summary of the means and standard deviations of the slope and curvature for the two subject groups across all of the whole tones used in this experiment.....	171
Table 4.1.17: Non-professional group's whole tone slope values calculated on the original F <sub>0</sub> trace.....	171
Table 4.1.18: Non-professional group's whole tone slope values calculated on the F <sub>0</sub> trace with a moving average applied. ....	172
Table 4.1.19: Professional group's whole tone slope values calculated on the original F <sub>0</sub> trace. ....	172
Table 4.1.20: Professional group's whole tone slope values calculated on the F <sub>0</sub> trace with a moving average applied.....	172
Table 4.1.21: Non-professional group's whole tone curvature values calculated on the original F <sub>0</sub> trace. ....	172
Table 4.1.22: Non-professional group's whole tone curvature values calculated on the F <sub>0</sub> trace with a moving average. ....	173
Table 4.1.23: Professional group's whole tone curvature values calculated on the original F <sub>0</sub> trace.....	173
Table 4.1.24: Professional group's whole tone curvature values calculated on the F <sub>0</sub> trace with a moving average applied.....	173
Table 4.2.1: Organization of Part One. ....	208
Table 4.2.2: Mean and standard deviation of the melodic semitone sizes for both the Lab and Church ensembles in Part One. ....	208
Table 4.2.3: Mean and standard deviation of the sizes of the vertical intervals in Part One between the first and second notes in the semitone intervals and the bass note for both the Lab and Church ensembles. ....	208
Table 4.2.4: Organization of Part Two. ....	213
Table 4.2.5: Mean and standard deviation of the melodic whole tone sizes in Part Two for both the Lab and Church ensembles. ....	213
Table 4.2.7: Notes in the two-measure seed progression, which is repeated four times to make up the musical material used in Part Three. ....	217
Table 4.2.8: Mean and standard deviation of the ascending and descending melodic whole tone sizes in Part Three for all ensembles, broken down by voice. ....	221
Table 4.2.9: Mean and standard deviation of the sizes of the vertical intervals in Part Three between the three voices across all renditions by each ensemble. ....	221
Table 4.2.10: Results of the <i>t</i> -tests run on the deviations (in cents) from Just Intonation for the grouping of vertical intervals in Part Three into those that share a larger number of harmonics with the fundamental (P8, P5, M3) versus those that share a fewer number of harmonics .....	225
Table 4.2.11: Results of the <i>t</i> -tests run on absolute interval size normalized around zero of the melodic and vertical intervals from Part Three to evaluate if the syllable that the notes were sung to influence interval size. ....	225
Table 4.2.12: Mean and standard deviation of the interval sizes for all of semitones and whole tone sizes in Part Four for each ensemble. ....	227
Table 4.2.13: Mean and standard deviation of the sizes of the vertical intervals in Part Four between the four voices for all of the sonorities with a half note in the bass (as marked in Figure 4.2.7). ....	227
Table 4.2.14: Results of the <i>t</i> -tests run on the deviations (in cents) from Just Intonation tunings for the grouping of vertical intervals in Part Four into the P8, P5, M3 versus the remaining intervals.....	232
Table 4.2.15: Results of the <i>t</i> -tests run on the deviations (in cents) from Just Intonation in vertical intervals in Part Four that occurred in the cadential progression versus those that occurred in non-cadential progressions. ....	232

Table 4.2.16: Results of the <i>t</i> -tests run on the absolute interval size normalized around zero in Part Four and grouped into those takes sung in German and those sung to the syllable “mi” for the Lab and Church ensembles. ....	233
Table 4.2.17: Summary of the means and standard errors for the ascending and descending semitones across each ensemble in Parts One and Four. ....	234
Table 4.2.18: Summary of the means and standard errors for the ascending and descending whole tones across each ensemble in Parts Two, Three, and Four. ....	235
Table 4.2.19: Summary of the means and standard errors (SE) across each ensemble in each experiment part for the vertical intervals where the upper note in the interval had at least 6 harmonics in common with the lower note’s first 32 harmonics. ....	237
Table 4.2.20: Summary of the means and standard errors (SE) across each ensemble in each experiment part for the vertical intervals where the upper note in the interval had less than 6 harmonics in common with the lower note’s first 32 harmonics. ....	238
Table 4.3.1: Summary of the results for intervallic direction from the regressions run on the melodic interval data. ....	242
Table 4.3.2: Summary of the results for different semitone conditions from the regressions run on the melodic semitone data for the solo experiment and Part Four of the ensemble experiment. ....	243
Table 4.3.3: Summary of the results for different semitone conditions from the regressions run on the melodic semitone data for Part One of the ensemble experiment. ....	244
Table 4.3.4: Summary of the results for different whole tone conditions from the regressions run on the melodic whole tone data for the solo experiment. ....	244
Table 4.3.5: Summary of the results for different whole tone conditions from the regressions run on data for Part One of the ensemble experiment. ....	245
Table 4.3.6: Summary of the results for singer identity from the regressions run on the melodic semitone data in the solo experiment. ....	246
Table 4.3.7: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part One of the ensemble experiment. ....	246
Table 4.3.8: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part Four of the ensemble experiment. ....	247
Table 4.3.9: Summary of the results for singer identity from the regressions run on the melodic whole tone data in the solo experiment. ....	247
Table 4.3.10: Summary of the results for singer identity from the regressions run on the melodic whole tone data in Part Two of the ensemble experiment. ....	248
Table 4.3.11: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part Four of the ensemble experiment. ....	248
Table 4.3.12: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part Four of the ensemble experiment. ....	248
Table 4.3.13: Summary of the results for accompaniment from the regressions run on the melodic data in the solo experiment. ....	249
Table 4.3.14: R <sup>2</sup> values for the regressions run on the interval size data in Section 4.1. Only those cells separated by a dashed line share the same regressors, which allows for direct comparison of the R <sup>2</sup> values. ....	250
Table 4.3.15: R <sup>2</sup> values for the regressions run on the interval size data in Section 4.2. Only those cells separated by a dashed line share the same regressors, which allows for direct comparison of the R <sup>2</sup> values ...	251

## Acknowledgements

First, I would like to acknowledge my advisor, Professor Ichiro Fujinaga, who has encouraged me in this research endeavor throughout my time at McGill University. I am grateful for his help in determining which aspects of the massive question of intonation in the singing voice to focus on and how best to test these aspects. Also, his patience and comments while I was writing this dissertation were invaluable. Professors Jonathan Wild and Peter Schubert also provided support and guidance, both in terms of the tuning theory underlying this research and in the selection and design of materials for the experiments. I would also like to acknowledge Professor Peter Schubert's work in finding singers for the experiments, particularly the SATB quartet from his VivaVoce ensemble and for conducting during the ensemble experiment. I would also like to thank Professors Fujinaga, Wild, and Schubert for helping me secure funding to hire the singers for the experiments, though a *Fonds Québécois de la recherche sur la société et la culture* (FQRSC) research creation grant.

I would like to thank Professor Dan Ellis for advice on various signal processing aspects of this research, not only during my time at McGill but also while I was undertaking my MPhil at Columbia University. During my time at Columbia, Professor Fred Lerdahl provided guidance in the early stages of this project. I would also like to acknowledge the contributions of my first academic mentor, Professor Michael Coghlan, who both introduced me to music technology and encouraged me to begin this line of inquiry on intonation in the singing voice.

In addition to hiring singers, the FQRSC research creation grant provided funds to hire a research assistant. In this role, Gabriel Vigliensoni was a godsend. He worked tirelessly, helping with the ensemble recordings, editing the experiment audio, and reviewing/editing the annotations produced by the score-audio alignment algorithm. The experiments were also facilitated by the expertise and efforts of the technical staff at the Center for Interdisciplinary Research in Music Media and Technology (CIRMMT), particularly Harold Kilianski and Yves Méthot. Darryl Cameron also provided invaluable technical support in the Music Technology labs and in acquiring new equipment for the project. I would like to thank Hélène Drouin for guiding me through the administrative red tape related to all aspects of my degree and the submission of this dissertation. I would also like to

acknowledge Lysiane Bouchard, David Haguenauer, and Maïko Paquette-Fujinaga for their assistance in the preparation of the French translation of the abstract.

My past and present labs mates in Distributed Digital Music Archives and Libraries lab, Andrew Hankinson, Cory McKay, Beinan Li, Jason Hockman, Jessica Thompson, John Ashley Burgoyne, and Jordan Smith, have provided me with much moral support and intellectual stimulation throughout my time at McGill. Andrew Hankinson also provided the invaluable service of taking care of the initial submission of the physical copies of my dissertation in my absence. I also owe a debt of gratitude to Caylin Smith for her work copy-editing the text prior to the final submission of this dissertation.

Financial support was provided by the *Social Sciences and Humanities Research Council of Canada* (SSHRC) through a doctoral scholarship, the FQRSC, as noted above, through a research creation grant, and CIRMMT through a student award. Travel funding was also provided through CIRMMT, as well as the McGill Alma Mater Fund and the Schulich School of Music's Graduate Research Enhancement and Travel Awards (GREAT Awards).

I would also like to acknowledge the ongoing love and support of my family. My parents, Lucia and Kevin Devaney, always prioritized my intellectual development and have shown unwavering support for my academic choices. The sacrifices they made for me to attend good schools and be successful in university are the reason I was able to undertake this research and produce this dissertation. I would also like to thank my mother for copy editing the literature review section of this dissertation. My sister, Julie Devaney, has also provided much encouragement, as well as invaluable moral support throughout my various academic and musical endeavours.

Last, but certainly not least, I would like to thank Michael Mandel, my life partner, sometime co-author, and guardian of my sanity. Michael both advised me on signal processing issues and undertook a sizable amount of grunt work, reviewing/editing the annotations produced by the score-audio alignment algorithm and copy editing almost every line of this document. His unwavering support allowed me to get through this dissertation process in a timely manner and with my good humour intact.

(This page intentionally left blank)

# Chapter 1 Introduction

The main objective of this dissertation is to examine whether intonation in the Western tradition is related to musical context, in both solo singing and in SATB (Soprano, Alto, Tenor, Bass) ensembles with one voice per part. Specifically, to determine whether the tuning of melodic semitone and whole tone intervals is influenced by intervallic direction and/or the scale degrees between which the intervals occur. This research also explores whether the tuning of vertical intervals is influenced by the number of shared harmonics between the two simultaneous notes or whether vertical intervals that occur in a cadential non-cadential context. Within this dissertation, intonation refers to the adjustment of the tuning of the musical tones that singers produce, as measured by interval size.

## 1.1 Motivation

Intonation in singing is a complex phenomenon that has received only limited attention in the literature to date. Previous studies have observed that singers, like non-fretted string instruments, do not strictly conform to either equal temperament or any other fixed-intonation system, such as Just Intonation or Pythagorean tuning (Barbour 1953; Backus 1977). None of these earlier studies, however, have explored whether singers' intonation practices in the western tradition are systematically related to musical context.

Intonation is an interesting object of study for a number of reasons. At the most basic level, there is the unanswered question of what singers are doing rather than what they are not doing. If they are not singing in a particular system, what is governing the consistencies that can be observed in repeat performances by a single singer or across singers? There are also issues surrounding how much of the singers' intonation practice is intentional and whether this changes with training or experience. Moreover, studying how intonation practices relate to the musical context can help us understand the link between what is written in the score and the actual music that the audience hears. Findings from this type of research could possibly be used to generate more "natural" sounding digital re-creations and also have potential pedagogical applications for training vocalists.

## 1.2 Prior Research

In the early 20<sup>th</sup> century, psychologist Carl Seashore and his colleagues at the University of Iowa undertook extensive work in performance analysis on singing, examining dynamics, intonation, and vibrato (Seashore 1936, 1938). Their analyses were based on amplitude and frequency information extracted from recordings with phonophotographic apparatus. Seashore and his colleagues manually measured fundamental frequency ( $F_0$ ) and found that the intonation of these estimates deviated from equal temperament. They also analyzed how  $F_0$  changed over the duration of each note and the impact of note duration on the amount and characteristics of this change. They did not, however, consider how variation in the collected data might be understood in terms of musical context: due to the laborious nature of the analysis, they only examined a limited number of recordings.

In recent years, much of the work relating to intonation has been conducted at the Speech, Music, and Hearing department at the Royal Institute of Technology (KTH) in Stockholm. Sundberg (Sundberg 1987) pioneered much of this research, including the examination of variations in intonation between solo and choral performance, as well as the influence of certain vowels on tuning. The work at KTH in the 1980s and 1990s was based on manual annotation of the note onsets and offsets. The  $F_0$  estimates were also done manually by finding an isolated overtone in a spectrogram representation of the audio. Prame used this method in his study of intonation in solo soprano singing, where he found that the intonation of notes deviated substantially, though not consistently, from equal temperament (Prame 1997). More recent work has used a semi-manual approach, where the beginning and ending of notes are annotated by hand before  $F_0$  estimation algorithms are employed. This approach has been used to study vibrato and pitch glides in solo performances of several Schubert songs by Timmers (2007), as well as chromaticism and pitch inflection, with a focus on the chromatic inflections for leading tones, in traditional solo Lithuanian singing by Ambrazevičius and Wiśniewska (2008). In contrast, Marinescu and Ramirez (2008) used a machine learning approach to extract information about pitch, duration, and amplitude for sung notes in several monophonic recordings and developed a model of expressiveness for timing and amplitude.

Vocal ensembles have been studied to a lesser degree than soloists, primarily because of the challenge of extracting information from a polyphonic signal (Ternström 2003). For example, Hagerman and Sundberg (1980) studied intonation in barbershop quartets, a style known for its ‘straight tone’ singing, i.e., singing with minimal vibrato. Jers and Terström (2005) used a semi-automated approach to examine intonation and vibrato in a multi-track recording of an eight-measure piece by Praetorius. They looked at the average values across both the whole ensemble and the individual sections for the mean and standard deviation of  $F_0$  for each note. Howard (2007a; 2007b) examined pitch drift and adherence to either equal temperament or just intonation in an SATB quartet using electroglottalgraphs to obtain  $F_0$  estimates. The results of these experiments are detailed in Section 2.2.

### 1.3 Methodology

A number of technical and perceptual issues needed to be addressed in order to answer the question of whether intonation in Western music is related to musical context. While intonation data can be extracted manually, it is an extremely time consuming procedure that limits the number of recordings that can be evaluated. Automatic extraction of intonation data allows for more data to be collected, which in turn enables more robust generalizations. There are two stages to the automatic extraction method: the first concerns the labelling of note onsets and offsets, and the second involves extracting fundamental frequency ( $F_0$ ) estimates for each frame of audio.

Labelling note onsets and offsets delineates the temporal period in the signal where each note occurs. Since scores were available for the recordings analyzed in this dissertation, score-audio alignment techniques were used to automatically label note onsets and offsets (Scheirer 1998). For the purpose of this study, an existing score-audio alignment method for identifying not only note onset and offsets, but also transient and steady-state sections of each note was developed by building on an existing algorithm (Orio and Schwarz 2001), and was fine-tuned for monophonic recordings of the singing voice. Once the note onsets and offsets were identified,  $F_0$  estimates were made with the YIN algorithm developed by de Cheveigné and Kawahara (2002). YIN is an autocorrelation-based monophonic  $F_0$  estimator, producing estimates at intervals of several milliseconds. To address the issue of pitch extraction in the ensembles, each singer was miked separately. The minimal bleed-through

from the other voices was resolved by using the score to constrain the YIN  $F_0$  estimates to improve their accuracy in this quasi-polyphonic context. Perceived pitch estimates for each note were made by taking the weighted mean across the series of  $F_0$  estimates, a method developed and perceptually tested by Gockel, Moore, and Carlyon (2001). This research also explores an improved way of describing how  $F_0$  changes over the duration of the note. For example, the first two coefficients of the discrete cosine transform are used to examine whether singers adjust the  $F_0$  at the end of the first note in a melodic interval in order to prepare the arrival of the second note.

This dissertation describes two experiments, one with solo singers and one with SATB ensembles. The first experiment looked of two groups of singers (undergraduate soprano vocal majors and professionals) and examined the role of intervallic direction and musical context on the intonation of melodic semitones. Each participant sung three *a cappella* renditions of the first verse of Franz Schubert's "Ave Maria," followed by three renditions with recorded accompaniment, which allowed for the role of accompaniment in melodic intonation tendencies to also be explored. Three SATB ensembles participated in the second experiment. One was a group of semi-professional singers who performed without a conductor and the other two were professional ensembles conducted by Peter Schubert, a music theory professor at McGill University. The experimental material consisted of four parts, which were each sung at least three times: a set of progressions by Jonathan Wild, another music theory professor at McGill, in which semitones occurred in different contexts; a set of progressions by Peter Schubert in which whole tones occurred in different contexts; an exercise by Giambattista Benedetti built on a two-measure chord progression designed to show that singers do not really perform in Just Intonation; and the first verse of Michael Praetorius' "Lo, how a rose e'er blooming," which was included so that, as with "Ave Maria," the participants were also performing a piece with which they were familiar. As with the solo singer experiment, the intonation of melodic semitones and whole tones were studied. This experiment also allowed for the study of vertical intonation tendencies, specifically whether vertical intervals are tuned closer to Just Intonation when there is a greater coincidence of harmonics and if this is influenced by whether the intervals occur in cadential versus non-cadential contexts.

In order to ensure that the renditions that were studied were “in tune,” the solo singers were asked to listen to their recordings and confirm that they considered their intonation to be accurate. During the recording sessions for the professional SATB ensembles, the conductor indicated which renditions were acceptably “in tune.” In the absence of a conductor for the semi-professional ensemble, I appraised the recordings.

#### 1.4 Contributions

This dissertation explores a number of issues that have not been addressed in prior research on intonation and makes four main contributions to the existing body of work on intonation practices.

- Although it has been shown that singers do not conform to a prescribed intonation system, there has not been, with the exception of Ambrazevičius and Wiśniewska’s work on Lithuanian music (2008), a systematic study of whether singers tend to perform musically similar patterns in similar ways. This dissertation analyzes intonation practices for melodic semitones and whole tones and vertical intervals with regards to local musical context in order to explore the consistency of singers’ intonation tendencies when singing similar material.
- In earlier research, only a limited number of performances were analyzed and the issue of reconciling variation between multiple performances of the same piece was not properly addressed. The experiments for this dissertation were designed to collect a large amount of similar data from different singers in order to explore the range of intonation practices that exist within groups of singers with similar levels of ability.
- Most of the prior research only looked at the mean of the  $F_0$  over the duration of the note. With its use of discrete cosine transform coefficients, this dissertation explores a new way of describing the evolution of the  $F_0$ . Specifically, at the end of the first note in melodic intervals.
- The development of a tool for automatically annotating onsets and offsets in the recordings in the singing voice facilitated the analysis of the experimental data in this dissertation and will be made available to other researchers.

Aspects of this dissertation have been published in a number of scholastic venues and presented at several national and international conferences. I performed the majority of the research and writing in all of these publications and presentations, and the specific contributions of my co-authored are detailed below. The early stages of this work were presented at the 2006 International Conference on Music Perception and Cognition (Devaney 2006) and at the 2007 Conference on Interdisciplinary Musicology (Devaney and Ellis 2007). The ideas in these conference papers were later expanded in an article for the *Journal of Interdisciplinary Musicology* (Devaney and Ellis 2008) and in paper presentations at the 2008 Digital Music Research Network Conference (Devaney et al. 2008) and at the Fourth Conference on the Physiology and Acoustics of Singing (Devaney and Wild 2009). The co-authors for these papers, Dan Ellis, Jonathan Wild, and Ichiro Fujinaga, all provided guidance in the development of the methodology. Dan Ellis also provided some programming source code for the *Journal of Interdisciplinary Musicology* article, which was ultimately not implemented in this dissertation. Some of the work presented in these papers also builds on the graduate work I completed at Columbia University with Fred Lerdahl.

The issues related to studying expressive performance using alignment methods were presented at the 2009 International Computer Music Conference (Devaney and Ellis 2009), and the refined method used in the experiments was presented at the 2009 Workshop on Applications of Signal Processing to Audio and Acoustics (Devaney et al. 2009a). Both of these papers were co-authored with Dan Ellis and the latter was also co-authored by Michael Mandel. Dan Ellis provided overall guidance for the digital signal processing aspects of the papers, and Michael Mandel assisted with the implementation of the hidden Markov model used in the refined alignment method. This work was summarized for a psychology audience in a forthcoming publication in *Psychomusicology* (Devaney et al. Forthcoming).

Preliminary results from the two experiments in this dissertation were presented at conferences. Some preliminary results for the solo experiment were presented at the 2009 Indiana University Symposium of Research in Music Theory (Devaney 2009), the 2009 Society for Music Perception and Cognition (Devaney et al. 2009b), and in the forthcoming publication in *Psychomusicology* (Devaney et al. Forthcoming). In the *Psychomusicology* article, Michael Mandel implemented the discrete cosine transform approach, which was the basis of the approach described in Section 3.2. Preliminary results from the ensemble experiment

were presented at the Fifth Conference on the Physiology and Acoustics of Singing (Devaney et al. 2010b) and at the 2010 International Conference on Music Perception and Cognition (Devaney et al. 2010a). The co-authors for these papers, Jonathan Wild, Peter Schubert, and Ichiro Fujinaga, provided assistance in the general formation of the experiments and the choice of the experimental material.

Some of the music theoretical implications of this research were presented at the 2007 meeting of the Society for Music Theory (Devaney 2008b), the 2008 meeting of the Canadian University Music Society (Devaney 2008a), and the 2010 Indiana University Symposium of Research in Music Theory (Devaney et al. 2010c). Some ways in which the tools developed for analyzing the recordings discussed in this dissertation could be applied to other expressive performance studies were presented at the 2010 Meeting of the Society for Music Theory (Devaney and Fujinaga 2010). My co-authors, Jonathan Wild, Peter Schubert, and Ichiro Fujinaga, provided general guidance in the projects presented in these papers.

In all of the papers mentioned above, I was the first author, completing the majority of the research, and except where indicated above, writing all of the text.

## 1.5 Structure

This dissertation has three main chapters, which detail the existing relevant literature (Chapter 2), the methods used to extract the intonation data (Chapter 3), and the two experiments (Chapter 4). Chapter 2 is divided into five sections, each of which surveys scholarly work in a different area related to this thesis: Section 2.1 in the area of tuning theory, Section 2.2 in the area of performance analysis, Section 2.3 in the area of pitch and consonance perception, Section 2.4 in the area of fundamental frequency estimation and transcription, and Section 2.5 in the area of audio-score alignment. Chapter 3 is divided into two sections. Section 3.1 details a number of evaluations that were performed for annotating note onsets and offsets in audio files with score-audio alignment and describes the algorithm that is used. Section 3.2 describes how perceived pitch is modeled for the purposes of this dissertation and how the discrete cosine transform is applied to describe the evolution of fundamental frequency. Chapter Four has three sections. Section 4.1 describes the set up and results of the experiment on intonation in solo singing. Section 4.2 examines the setup and

results of the experiment on intonation in SATB ensemble singing. Lastly, Section 4.3 analyzes and compares the results of the both experiments.

## Chapter 2: Literature Review

This chapter presents a review of existing literature relevant to the research described in Chapters 3 (Automatic Extraction of Performance Parameters) and Chapter 4 (Intonation Experiments). The first and second main sections of this chapter provide the context for the main research question being posed by this dissertation: how do singers tune? Section 2.1 focuses on tuning theory and consists of both the mathematical definitions of the most commonly discussed tuning systems (Pythagorean, Just Intonation, Meantone, and equal temperaments) and a survey of the history of tuning theory from antiquity to the present. Section 2.2 discusses the history of performance analysis. This section includes a comprehensive survey of studies of intonation and vibrato in the singing voice, as well as relevant studies on intonation, vibrato, timing, and dynamics in other instruments. Section 2.2 concludes with a discussion of various attempts to model performance practices.

The third, fourth, and fifth main sections of this chapter provide background on the techniques used to extract and analyze intonation-related data in the experimental recordings used in this dissertation. Section 2.3 surveys the literature on pitch and consonance perception, which is important both for calculating estimates of the perceived pitch with the fundamental frequency estimates extracted from the recordings and for understanding the relationship between vertical consonance and tuning. The literature discussed in Section 2.3 forms the basis for the analysis techniques described in Section 3.2 and provides a context for the experimental data in Chapter 4. Section 2.4 describes the range of approaches for automatically extracting fundamental frequency data from recordings and draws links between these technical approaches and the perceptual findings detailed in Section 2.3. Section 2.5 surveys the current state of the art in determining the location of note onsets and offsets in recordings. Both Sections 2.4 and 2.5 provide a background for the techniques for extracting intonation-related data described in Section 3.1.

## 2.1 Tuning Theory

This section provides an overview of the mathematics of the most widely discussed tuning systems, as well as the historical context in which they developed. Section 2.1.1 presents an overview of the mathematic and acoustical details of Pythagorean tuning, 5-limit Just Intonation tuning, 1/4-comma Meantone temperament, and equal temperament. Section 2.1.2 details the historical context for the development of these and other tuning systems, including well temperaments.

### 2.1.1 Overview of Tuning Systems

Harmonic tones, such as those produced by the singing voice or musical instruments, contain an overtone series of frequencies in whole number ratios to the fundamental frequency. The distance between these harmonics can be used to derive the tunings of intervals, as demonstrated in Figure 2.1.1. The decision about which harmonics are admissible into a system is a defining feature between such systems as Pythagorean and 5-limit Just Intonation system. The historical context for these system is discussed in Section 2.1.2.

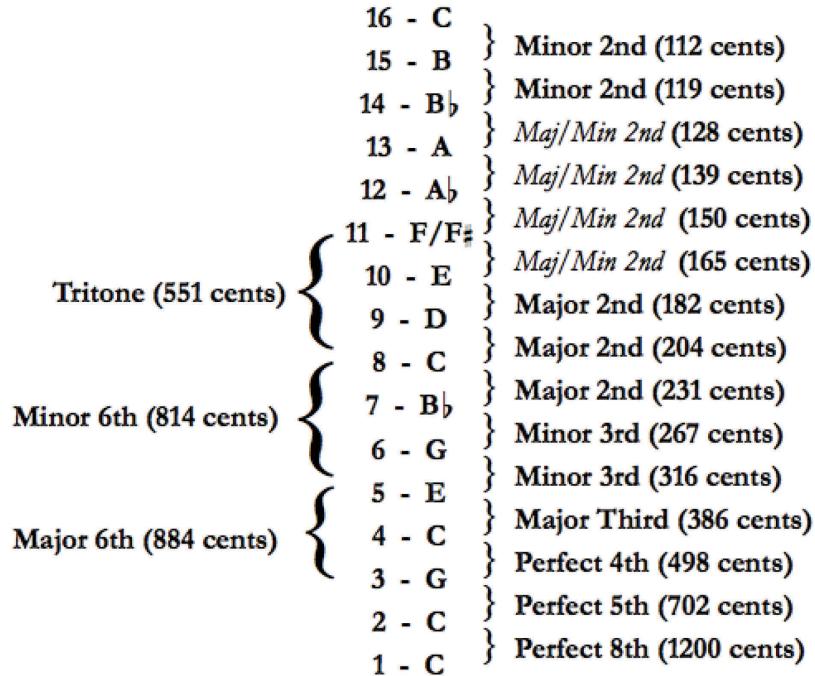


Figure 2.1.1: Some of the interval tunings that can be derived from the first sixteen partials of the overtone series.

	Pythagorean		5-limit Just Intonation		Meantone	Equal Temperament
	Ratio	Cents	Ratio	Cents	Cents	Cents
m2	<b>256:243</b>	90	<b>16:15</b>	112	86	100
M2	<b>9:8</b>	204	<b>9:8</b>	204	193	200
			<b>10:9</b>	182		
<b>m3</b>	<b>32:27</b>	<b>294</b>	<b>6:5</b>	316	<b>312</b>	<b>300</b>
<b>M3</b>	<b>81:64</b>	<b>408</b>	<b>5:4</b>	386	<b>386</b>	<b>400</b>
P4	<b>4:3</b>	498	<b>4:3</b>	498	503	500
TT	<b>1024:729</b>	588	<b>45:32</b>	590	578	600
	<b>729:512</b>	612	<b>64:45</b>	610	620	
P5	<b>3:2</b>	702	<b>3:2</b>	702	697	700
<b>m6</b>	<b>128:81</b>	<b>792</b>	<b>8:5</b>	<b>814</b>	<b>770</b>	<b>800</b>
<b>M6</b>	<b>27:16</b>	<b>905</b>	<b>5:3</b>	<b>884</b>	<b>888</b>	<b>900</b>
m7	<b>16:9</b>	996	<b>9:5</b>	1018	1008	1000
M7	<b>243:128</b>	1110	<b>15:8</b>	1088	1082	1100
P8	<b>2:1</b>	1200	<b>2:1</b>	1200	1200	1200

Table 2.1.1: Interval sizes in Pythagorean, 5-limit Just Intonation, 1/4-comma Meantone, and equal temperament. The perfect consonances in Western harmonic practice (perfect fourth, fifth, and octave) differ by only a few cents. The bolded values indicate where the largest tuning differences occur between the systems, especially for the imperfect consonances (major/minor thirds and sixths).

Pythagorean tuning is essentially a 3-limit system (Partch 1974). In this instance, limit refers to the largest prime number harmonic (or its multiple) from which intervals can be constructed, such that in a 3-limit system only intervals consisting of ratios of 2 and 3 and their multiples are included. In a 3-limit system, only perfect unisons, fourths, fifths, and octaves occur within the first few partials. Other intervals do not occur until much further up the overtone series, leading to the critique that the major and minor thirds and sixths are not “in tune.” For example, the ratio of the first occurring major third in a 3-limit system is 81:64, whereas in a 5-limit system the ratio of the first occurring major third is 5:4. The most common use of Just Intonation is as a 5-limit system, though 7-, 11-, 13-, 17-limits have also been used (Partch 1974; Johnston 2006). Just Intonation is sometimes referred to a “pure” tuning in the literature because of its close adherence to the lower overtones in comparison to not only Pythagorean, but also to Meantone and equal temperament. Table 2.1.1 shows the difference in interval sizes between these two systems.

One practical issue with both Pythagorean and 5-limit Just Intonation is that the systems are only partially in tune for keys within a small number of steps on the circle of fifths (see Figure 2.1.2). When the modulation moves further away on the circle of fifths, intervals in the new key move further and further away from their tuning in the home key. For example, in a simple implementation of a fixed 5-limit Just Intonation tuning, a modulation of one step around the circle, from C to G, would result in intervals in the tonic triad, G-B-D, that are the same size as the original triad, C-E-G: the B would still be 386 cents above the G, and the D would still be 702 cents above the G. However, taking two steps, C to D, would result in a tonic triad with a condensed fifth: the F# would be 386 cents (a 5/4 major third) above D (assuming a 45:32 tritone in relation to C), but the A would be only 680 cents above D. In order to achieve robustness in the face of modulation, it is necessary to temper, or adjust the size, of some of the intervals. This is what is done in both Meantone and equal temperaments.

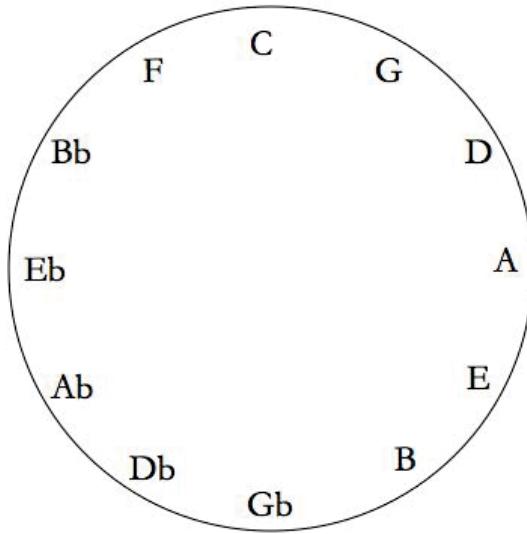


Figure 2.1.2: Circle of Fifths.

Meantone and equal temperaments can both be described as redistributions of tuning commas, the difference between incompatible compound intervals, such as two chains of intervals that end on the same note but not the same tuning. The Pythagorean comma, ~24 cents, is the difference between twelve perfect fifths (8424 cents) and seven perfect octaves (8400 cents). The Syntonic comma, ~22 cents, is the difference between four perfect fifths

(2808 cents) and two perfect octaves plus a 5:4 major third (2786 cents), as well as the difference between the Pythagorean major third (408 cents) and Just major third (386 cents) and between the 9:8 major whole tone (112 cents) and the 10:9 minor whole tone (90 cents).

In Meantone temperaments, the Syntonic comma is split over a specified number of fifths. Therefore, 1/4-comma Meantone tempers all the fifths in the chain by 1/4 of the Syntonic comma, 1/5-comma Meantone tempers the fifths by 1/5 of the Syntonic comma, 1/6-comma Meantone tempers the fifths by a 1/6 of a Syntonic comma, and so on. In 1/4 Meantone, the first modulation described above (C to G) results in a tonic triad with a B that is 386 cents above the G and a D that is 697 cents above the G. The second modulation (C to D) results in a tonic triad with the F# 386 cents above D and the A 697 cents above D. A comparison of modulations in Just-Intonation and Meantone temperament, as well as equal temperament, is shown in Table 2.1.2.

	Original key			1 step on circle of 5ths			2 steps on circle of 5ths		
	C	E	G	G	B	D	D	F#	A
<b>5-limit Just Intonation</b>	0	386	702	0	386	702	0	386	680
<b>1/4 Comma Meantone temperament</b>	0	386	697	0	386	697	0	386	697
<b>Equal Temperament</b>	0	400	700	0	400	700	0	400	700

Table 2.1.2 Comparison of the impact of modulation around the circle of fifths in 5-limit.

Equal Temperament can be understood as either the equal redistribution of 1/12 of the Pythagorean comma across all twelve fifths in the chain or as a system where the size of each semitone is  $\sqrt[12]{2}$ . In equal temperament, for example, the size of the tonic triad remains the same regardless of how many steps around the circle of fifths are taken. The third of the chord is always 400 cents above the tonic and the fifth of the chord is 700 cents above the tonic. The trade off for the increased modulation robustness in these systems is with intervals that more closely align with those in the overtone series. In 1/4-comma Meantone, only the unisons, major thirds and octaves are “in tune” and in equal temperament only the unisons and octaves are “in tune.”

## 2.1.2 History of Tuning and Temperament

Questions of tuning have preoccupied many theorists from antiquity up until the present. For some ancient Greeks, the definitions of good tuning and consonance were numerically based. The Pythagoreans, c. 500 BCE, limited their definition of consonance to intervals that

corresponded to monochord divisions, which employed only super-particular ratios of the numbers 1, 2, 3, and 4. This series of numbers was known as the tetractys and has the property of summing to ten. The tetractys was geometrically represented as a triangle with the top row having one point, the second row having two, the third row having three, and the fourth row having four. It also had important numerological significance for the Pythagoreans: the use of these numbers resulted in a system in which only the octave (with a frequency ratio of 2:1), the fifth (3:2), the fourth (4:3), and the compounds of the octave and the fifth (i.e., the 4:1 perfect fifteenth and the 3:1 perfect twelfth) were considered consonant. The octave compound of the perfect fourth perfect eleventh (8:3) was not considered consonant because the number eight does not occur in the tetractys (Barbera 1984). In contrast, Aristoxenus (c. 335 BCE) argued that the ear, rather than strict mathematics, should be the guide for determining consonance. Ptolemy (c. 120 CE) 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 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. His Syntonic Diatonic tuning system was one of the first complete articulations of 5-limit Just Intonation, which mirrors the sequence of intervals that can be derived from the harmonic series (unlike Didymus, whose earlier system was closer to the arrangement of the monochord) (Barbour 1953).

The early medieval period's main contribution to tuning theory was the transmission of Greek ideas. The major source of Greek music theory in the late medieval and early renaissance periods was Boethius' *De institutione arithmeticā musica libri quinque* (c. 520 CE), which is known in English as *Foundations of Music*. This text laid out an extensive discussion of consonance and dissonance in the Pythagorean tradition. While the Pythagorean doctrine of limited consonance was sufficient for the music of the early and mid medieval eras, the increased use of the third and the sixth as imperfect consonances in the music of the late medieval period created a need for tuning theories and systems in which these intervals sounded agreeable (e.g., 5-limit Just Intonation).

There was an increasingly widespread acceptance of thirds and sixths as consonant intervals, and a number of High Renaissance theorists adjusted the Pythagorean method of monochord division to account for these intervals, e.g., Ramis' *Musica Practica* (1482). In late Renaissance, this method of monochord divisions was expanded to include chromatic and enharmonic pitches in Glarean's *Dodechordon* (1547) and 5-limit just-intonation divisions by Salinas's *De musica libri septem* of 1577. The first explicit discussions concerning intonation arose around the same time and were inspired by the increased interest in keyboard tuning/temperament systems that highlighted the difference between singers' tuning practices and the fixed tuning of keyboard instruments.

The first attempt to systematically address the issue of singers' intonation practices in performance, as well as to develop a keyboard instrument to guide them, was Nicola Vicentino's *L'antica musica ridotta alla moderna prattica* (1555). Vicentino put forth two tuning systems for his 31-tone gamut. The first was explicitly named "Tuning System for the Purposes of Accompanying Vocal Music," which was conceived of as an augmented  $\frac{1}{4}$ -comma Meantone system, with the first 19 fifths tempered by  $\frac{1}{4}$  of a syntonic comma and the remaining 12 fifths tuned pure. The second system was built with a chain of 31 fifths tempered by a  $\frac{1}{4}$ -comma, which Vicentino conceived, somewhat erroneously, as 31-tone equal temperament. This is characteristic of a recurring conflict throughout the late Renaissance and Baroque periods, a time when there was the desire for idealized systems and the need for practicality in keyboard tuning. The conflict was also articulated in Zarlino's *Istitutioni harmoniche* (1558), where both 5-limit Just Intonation and the need to systematically temper intervals when tuning string and keyboard instruments are discussed. See Wild and Schubert (2008) for a discussion of the mathematics underlying theories of vocal tuning in the late Renaissance.

Debates concerning the most appropriate keyboard temperament continued to dominate tuning theory for the next hundred and fifty years. Equal temperament, however, eventually emerged as the preferred system over the various meantone and well-tempered systems that were developed during this time. Salinas's *De musica libri septem*'s also described a number of meantone systems, which were developed by distributing different proportions of the syntonic comma over a chain of fifths. In his 1581 treatise, *Dialogo di Vincentino Galilei... della musica antica, et della moderna*, Vincenzo Galilei provided practical advice on how to derive

equal temperament, suggesting that the 18:17 Just semitone (99 cents) could be used as a basis for calculation. In his unpublished treatise from the 1580s, *De spiegheling der singconst*, Stevin described a method for calculating roots. These roots eventually became the basis of equal tempered intervals as the  $\sqrt[12]{2}$ . In *Harmonie universelle*, Mersenne (1636) applied Stevin's roots to the question of tuning and presented a mathematical derivation for the equal tempered semitone.

In spite of these mathematical advancements, well-temperaments were used for tuning keyboards until the 20<sup>th</sup> century (Jorgensen 1991). Well-temperaments were used because they minimized the wolf-tone found in 1/4-comma Meantone systems and sounded better in the keys with the least numbers of accidentals, which were favoured in the pre-Romantic eras. Well-temperaments distribute the Pythagorean comma amongst a chain of fifths in different ways. In 1691, Werkmeister described a number of cyclic tunings, including those based on 1/4 and 1/3 commas. In 1754, Vallotti, and later Young in 1800, independently developed a well temperament where 1/6 of the Pythagorean comma was distributed over the first six fifths. Later systems distributed the comma more equally, allowing for a greater number of keys to be usable.

With the decline of interest in Just Intonation in the eighteenth and nineteenth centuries, discussions of tuning ratios were limited to music theoretic treatises that considered arguments for consonance and dissonance. The most notable of these treatises were Rameau's theory of harmony (1722) and Helmholtz's theory of *Konsonanz* (1863), which will be discussed in Section 2.3.4. More details on the history of tuning and temperament can be found in Barbour (1953), Burns (1999), Rasch (2002), and Page (2004).

By the early twentieth century, tuning theory had become something of a fringe interest that was restricted primarily to compositional pursuits (Wilkinson 1988). The question of intonation in practice was only rarely considered, as there was no reliable way of assessing exactly what pitches were being performed. One notable exception is Boomsliter and Creel's (1961; 1963) attempt to develop a theory of melody based on their experiments with musicians' preferences for various tuning systems on a monochord-like instrument.

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

on how they should modify their usual intonation practices in order to achieve a more historically accurate performance, e.g., Jackson (2005). In contrast, the research described in this dissertation is descriptive, rather than prescriptive, 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 a small, but growing, number of studies that address questions related to intonational aspects of performance, which are described in Section 2.2.

## 2.2 Performance Analysis

This section surveys research conducted in the area of performance analysis from the early of psychologists in the first half of the twentieth century, as described in Section 2.2.1. Section 2.2.2 discusses different methods of extracting performance data, with a focus on audio recordings. Studies of intonation and vibrato in non-fretted string instruments are examined in Section 2.2.3. The focus turns to singing in Section 2.2.4 with a description of the physiology and acoustics of the singing voice. Studies of intonation and vibrato in the singing voice are subsequently detailed in Section 2.2.5. The question of how to model the collected performance data is addressed in Section 2.2.6.

### 2.2.1 General Overview

Interest in studying recorded performances dates back almost as far as the birth of recordable media. One of the earliest contributions to this field was Dayton Miller's (1916) work on visualized pitch information in recordings with phonophotographic apparatus. The psychologist Carl Seashore and colleagues at the University of Iowa also undertook extensive work in performance analysis (Seashore 1936a, 1938) of pianists, violinists, and singers. The researchers employed a number of methods to study recorded performances, including a stroboscope technique for frequency estimation and an oscillograph for intensity estimation. Piano performances were studied from both piano rolls and films that showed the movement of the hammers during the performance. Vernon (1937) studied asynchrony in the timing of individual chord notes in performances by four different pianists. He found that the degree of asynchrony was performer dependent and ranged from 30–200 ms. He also found that the amount of asynchrony was related to melody and phrasing. Studies on fretless stringed instruments will be described in Section 2.2.3, and studies on singing will be described in Section 2.2.5.

Though relatively accurate performance data could be assessed with these methods, the methods were extremely labour intensive, which limited the number of pieces that could be evaluated. Interest in empirical performance analysis subsequently diminished due, in part, to its laboriousness. It continued mainly in the area of ethnomusicology (Seeger 1951; Tove et al. 1966) and in smaller-scale studies of acoustic features of instruments (Fletcher and Sanders 1967; Beauchamp 1974).

The resurgence of a more general interest in music performance studies in the late 1970s coincided with musicologists moving away from equating scores with music, as well as cognitive psychologists' increased interest in music. Much of this work was on rhythm. Bengtsson and Gabrielsson (1980, 1983) undertook a number of systematic experiments on musical rhythm in performance. Following up on this earlier research. Todd (1985; 1992) studied both rubato and dynamics in piano performance, developing models to account for their individual relationships to musical structure and their interaction. Similarly, Clarke (1989) examined how rhythm in piano performance could be related to both the structural hierarchy of a piece and note-level expressive gestures. In the 1990s, Repp (1990, 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, as well as 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 for various instruments up to 2003 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).

### **2.2.2 Extraction of Performance Data**

Historically, the piano has been the primary instrument of performance analysis for several reasons. One reason is the large amount of solo repertoire available. This allows for the examination of the performer in a context to which he or she is accustomed, in contrast to instruments where it is more typical to play in an ensemble. Another reason is the piano's percussive nature, which makes it possible to study timing with a high degree of precision.

One can also acquire accurate, minimally intrusive performance measurements from a pianist via MIDI or another technology 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 a digital 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. Recently, other approaches have been developed, such as Chen, Wollacott, Pologe, and Moore's (2008) system for extracting pitch information from finger-board positioning on the cello, though unlike comparable approaches for the piano this does not provide information about note onsets and offsets.

This extraction of performance data directly from recordings enables the study of a wider range of instruments and existing performances. Accuracy, however is still an issue. 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 2008) and cello (Hong 2003). The recorded performances in these studies were analyzed using either manual or semi-automatic methods, where the notes onsets and offsets were manually annotated and the F0 estimation was done with an algorithm. Semi-automated systems are also used for analyzing recordings of piano music. For example, the system proposed by Earis uses a "manual beat tapping system" for synchronization that is corrected by both a computer-aided system and human intervention (Earis 2007). The research described in this dissertation uses a fully automated method that is described in Chapter 3.

### **2.2.3 Studies of Intonation and Vibrato in Instruments**

Early work completed at the University of Iowa included studies by Hattwick (1932) on vibrato in wind players and found that wind players typically employed a minimal amount of vibrato. Green (1937) looked at whether solo violinists played in equal temperament, Pythagorean, or Just Intonation. He observed that violinists tended towards Pythagorean tunings. Mason (1960) looked at the same question as Green for intonation in wind quartets. He found that the intonation in the wind quartets he studied did not conform to equal

temperament, Pythagorean, or Just Intonation. Contemporaneously to Mason, Shackford (1961, 1962b, 1962a) studied intonation in string trios and found a wide variety of sizes for all intervals studied, including major seconds, major and minor thirds, fourths, tritones, and fifths. Loosen (1993) built on Green's work, exploring whether solo violinists' intonation is closer to equal temperament, Pythagorean, or Just Intonation. He found that in the performances he studied the intonation was closer to Pythagorean and equal temperament than Just Intonation.

Loosen also studied the relationship between musical experience and intonation, looking specifically at whether instruments influenced musicians' tuning preferences. In a task where the participants adjusted a scale of synthetic tones, he found that violinists' adjustments were closer to Pythagorean tuning, pianists were closer to equal temperament, and non-musicians' adjustments did not converge on a particular tuning system (Loosen 1994). In a listening experiment using synthetic scales tuned to different systems, Loosen found similar results in regards to instrument and preference for Pythagorean and equal temperament. He also found that both groups of instrumentalists judged scales tuned in Just Intonation to be less in tune than either Pythagorean or equal temperament (Loosen 1995). Nordmark and Ternström (1996) performed a pitch adjustment experiment with synthesized major thirds using both undergraduate students and orchestral musicians as subjects. Their subjects' preferred tunings ranged from 388–407 cents, with the average being 397 cents.

The connection between intonation and other factors has also been considered. The relationship between string players' intonation and training has been explored by Salzberg (1980), and the relationship between wind player's intonation and timbre has been studied by Ely (1992). More recently, Chen and colleagues (2008) looked at the role of the physicality of cello playing, specifically bowing, on pitch accuracy. Other studies of intonation are described in Morrison and Fyk (2002), including the role of training in intonation tendencies. Overall, they found that a number of studies support the idea that trained instrumentalists tend to play sharp, particularly for intervals larger than a third, and that listeners prefer the sharper tunings. Morrison and Fyk contrasted this finding with other studies that showed that musical context impacts intonation; however, they did not find much agreement in the studies they considered about exactly how ascending versus descending contexts impact

intonation. Moirrison and Fyk also discuss the pedagogical implications of such intonation studies.

#### 2.2.3.1 Fyk's work on Melodic Intonation in the Violin

The largest of intonation in solo violin performance to date is Fyk's *Melodic Intonation: Psychoacoustics and the Violin* (1995). Fyk proposed a multi-level model of intonation with a degree of independence for interval tuning that corresponds inversely with the structural significance of the interval. Through listening tests, she found that both pitch discrimination and production are learned. Fyk also explored the role of expectation in the perception of intonation, suggesting that some sequences are perceived holistically, and the role of categorical perception of interval size. Fyk termed these tuning ranges "tolerance zones" and argued that their parameters are influenced by musical context.

Fyk also undertook empirical evaluation of recordings of two professional violinists performing both individual intervals, as well as performances of a piece in theme and variations form. From her analysis of these performances, Fyk made a number of generalizations about melodic intonation in the violin. She found that both notes in the middle of phrases and notes with vibrato were less stable. Also, when violinists repeated the same phrase, there was some variation in intonation, particularly at slower tempos, but the general shape, or "colour," of the intonation was maintained. Fyk argues that the glides at the end of the notes are intentional and suggests that note connections may influence pitch perception.

Fyk found that there were not any general tendencies towards any prescribed tuning system. She observed that smaller intervals (smaller than a fifth) tended to be smaller than equal temperament, while larger intervals tended to be larger than equal temperament. She also found that when players made a tuning mistake, they tended to adjust the following interval to compensate. Overall, Fyk argues that melodic intonation is "dynamically charged," which means that it is influenced by tonal function rather than a prescribed system. She also concludes that "correct" intonation is ultimately a combination of acoustic, cultural, and individual factors.

#### **2.2.4 Physiology and Acoustics of the Singing Voice**

The anatomy of the human voice organ includes the vocal folds, the breathing system, and the vocal and nasal tract. In the vocal tract, when the glottis is narrowed, the resulting air stream restriction both causes the vocal folds to vibrate and creates a Bernoulli force, which in turn tries to close the glottis. The sound created by the vibration of the vocal folds is phonation, with the audible frequency equal to the frequency of the vibration of the folds. This frequency is affected by the overpressure from the lungs, subglottis pressure, and laryngeal musculature, which determines the length and tension of the vocal folds. The vibration patterns of the vocal folds vary according to phonation frequency. At lower frequencies, the rate of glottis closure is slower. At higher frequencies, the vocal folds are long, thin, and tense, with no glottal waves.

Lindblom and Sundberg (1970) proposed a theory of how the articulatory system is used in singing based on measurements of the x-rays of subjects pronouncing various long vowels. They determined that the articulation of long vowels is influenced by the jaw; the lip opening, which is primarily dependent on the jaw position, but can also be widened or narrowed horizontally; the thickness of the tongue body; the velum, which controls the amount of air which flows into the nasal cavity; and the larynx, which can be raised or lowered. The physiological basis of vocal vibrato has not been conclusively determined. Recently, Titze, Story, Smith, and Long (2002) proposed that vibrato could be due to the physiological pairing of the cricothyroid muscle with either the thyroarytenoid or the lateral cricoarytenoid muscles as an agonist–antagonist pair.

One of the most important characteristics of the vocal tract is that the formant frequencies are more easily transferred through it than other frequencies. The result is that the frequencies in the voice source that are close to the formant frequencies are more audible. Formants determine the vowel and are sometimes modified by singers to increase their perceptibility in orchestral contexts (known as the singer's formant). The acoustical differences between solo and choir singing have been explored by Rossing, Sundberg, and Ternström (1984). They observed that solo singers tended to emphasize the “singers' formant,” while choir singers tended to emphasize the fundamental frequency. The results of this study were replicated with different singers in Rossing, Sundberg, and Ternström (1985), as well as Rossing, Sundberg, and Ternström (1987). In their later work, they also explored

gender differences in the “singers formant.” The formants are dependent on the length and shape of the vocal tract (generally longer vocal tracts have lower formant frequencies) and are also influenced by the shape of the lips, the jaw opening, the tongue, the velum, and the larynx. The complete anatomy of the human voice and how it functions in singing is described in Sundberg’s *The Science of the Singing Voice* (1987) and in slightly more detail in Titze’s *Principles of Voice Production* (1994).

The physiology of the singing voice has been used to inform singing synthesis. Some of the earliest work in this area was done by Sundberg (1978b), who used models of the glottal voice source and vocal tract resonator. Cook (1993) created a physical model of the singing voice that could be controlled in real time through a text-based software synthesis environment. The physical model was developed with waveguides and re-creates both the vocal and nasal tracts, as well as the acoustic energy that radiates through the throat. Cook (1996) provided an overview of earlier work in speech and singing synthesis, including the use of linear predictive coding in speech synthesis (Atal and Hanauer 1971), sinusoidal modeling in both speech (Mcaulay and Quatieri 1976) and singing synthesis (Serra and Smith 1990), as well as models for controlling formants (Rodet 1984). A more recent survey of models of the singing voice is available in Kim (2009). Kim’s survey included two recent dissertations on the subject: Kob’s physical model (Kob 2002), which can capture different registers and pathologies, and Kim’s own analysis-synthesis framework, which automatically estimates the modeling parameters (Kim 2003). Recently, another approach has been developed by d’Alessandro and collaborators (2008), who describe a refined technique for separating the vocal source from the glottal source in order to better capture the expressive aspects of singing. A review of KTH’s work on singing synthesis is available in Sundberg (2006).

Other work concerning singing synthesis has focused on vocal vibrato. Maher (1990) presented a wavetable approach for tracking formants to determine the rate and depth of vibrato. Herrera and Bonada (1998) described a sinusoidal modeling framework, where vibrato-related peaks in the frequency envelope of the sound could be identified in analysis and modified in resynthesis. Arroabarren and Carlosena (2004) described a source-filter model of vibrato that can be combined with sinusoidal approaches. More recently, Arroabarren and collaborators (2002a; 2002b; 2003) developed a method for measuring

amplitude and frequency modulation in vibrato and explored the relationship between the two types of modulation. In later work, they extended this to include instantaneous frequency and amplitude analysis of the partials in vibrato tones (2006).

Perceptual tests by Howes and colleagues (2004) showed a discrepancy between acoustic analyses of vibrato and the listeners' reported perceptual judgement. Verfaille, Guastavino, and Depalle (2005) used a listening test for evaluating different vibrato models and found that vibrato with modulation of the spectral envelope was preferred over vibrato with just frequency modulation pulses, amplitude modulation pulses, or a combination of frequency and amplitude modulation pulses.

## **2.2.5 Studies of Intonation and Vibrato in the Singing Voice**

### **2.2.5.1 Solo Voice**

As noted above, empirical evaluation of the singing voice dates back to the early part of the twentieth century. Schoen (1922) studied five performances of Gounod's setting of the "Ave Maria" and found that tuning depended on the direction of the line: notes in a descending line tended to be flatter, whereas notes in an ascending line tended to be higher. He found that in general the singers' tunings 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 was broader when they sung opera songs, compared to when they sung 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 (1936b) also looked at singers' performances of Gounod's setting of the "Ave Maria," as well as Handel's Messiah. He studied nine performances, focusing on the connections, or glides, between notes. He was able to correlate glide extent with direction and found that glide extent was larger going up than going down. Miller (1936) also studied vibrato and observed a 5.9–6.7 Hz range of vibrato rate and a 44–61 cent range of depth, with faster vibrato in shorter tones. He also provided a large amount of detail through "performance scores" about tuning, though a lot of the data was not analyzed. Miller's finding on vibrato echoed Tiffin's earlier findings (1932) that the average rate of vibrato is 6.5 Hz and the average depth is 60 cents and Metfessel's findings (1932) that the range of

vibrato rate ranged from 5.5–8.5 Hz (with an average of 7 Hz) and the depth of the vibrato ranged from 10–100 cents (with an average of 50 cents). Miller's observations about intonation also confirmed earlier findings that singers deviate from either equal temperament or Just Intonation. Miller also described different characteristics of the gliding transitions between notes. He also detailed dynamics and timing in the performances. Bjorklund (1961) studied the influence of training on vibrato and timbre in soprano singers. His results showed that with more training, singers had greater control over their vibrato.

#### 2.2.5.1.1 Research at KTH

More recently, there has been a great deal of work done at the “Speech, Music, and Hearing” research group at the Royal Institute of Technology (KTH) in Stockholm, Sweden. Sundberg (1982) observed deviations from Pythagorean and Just Intonation in singing with vibrato and concluded that the presence of vibrato allowed the singers to use greater range of tunings than singers singing in straight-tone barbershop style because of the presence of beats. Gramming, Sundberg, Ternstrom, 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 (1987) also examined variations in intonation between solo and choral performance, as well as the influence of certain vowels on tuning. He found a significant amount of variation in  $F_0$  across choirs, especially when vibrato is present. He also observed some variation in regards to “sharpness” or “flatness” of certain vowels, but general observable trends were limited. Carlsson-Berndtsson and Sundberg (1991) showed that singers tuned the two lowest formants in order to project their voices and that this did not produce a discernible perceptual impact on vowel perception. Sundberg (1994) also examined the role of vibrato in classical singing, detailing its acoustics and psychoacoustic features in a thorough review of vocal vibrato research.

Prame (1994; 1995) studied vibrato rate in ten professional sopranos' performances of Schubert's “Ave Maria.” 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 accurately measure the vibrato rate. He found that the mean rate of vibrato was 6 Hz 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 or not the 25 tones are in tune, as well as the professional singer's ability to replicate pitches in the opening and closing statements of "Ave Maria." They did not find much agreement amongst the expert listeners as to which notes were in tune and which ones were not. The singer's ability to replicate the tones was done by comparing the deviation of mean frequency of each corresponding note in the opening and closing statements in "Ave Maria" from equal temperament. They found that when the corresponding tones were within 7 cents of each other, the expert listeners agreed that they were in tune. Prame (1997) also used the "Ave Maria" recordings to study vibrato extent and intonation. He found that the vibrato extent in these performances ranged from 34–123 cents and that tones with larger vibrato depth tended to be sharper. The intonation of the notes deviated substantially, though not consistently, from equal temperament. Prame also calculated each singer's mean deviation from the accompaniment. Overall, the range of these means was from 12 cents below equal temperament to 20 cents above equal temperament.

A survey of other research into singing voice performance by the "Speech, Music, and Hearing" research group is available in Sundberg (1999). Some more recent work includes Murbe, Pabst, Hofmann, and Sundberg's (2002) study of the role of auditory and kinaesthetic feedback on pitch control under various conditions, including fast vs. slow and legato vs. staccato singing. They found that a reduction in auditory feedback reduced intonation accuracy in staccato and fast singing. Bretos and Sundberg (2003) examined vibrato in sustained notes with vibrato and found that vibrato rate was singer dependent, but that both vibrato extent and F0 was related to sound level. Their results also confirmed Prame's finding of an increase in vibrato rate towards the end of a note.

#### 2.2.5.1.2 Other Research

Other work on vibrato includes a number of articles grouped together as a special topic in a 1987 issue of the *Journal of Voice*. Myers and Michel (1987) looked at vibrato and pitch transitions and observed small perturbations in the vibrato rate and depth that facilitated changes of note. Ramig and Shipp (1987) argued that vocal vibrato in professional singers and vocal tremors in people with vocal pathologies are related. More recently, van Besouw, Brereton, and Howard (2008) used a listening experiment to determine the range of

acceptable tunings for notes sung with vibrato. They found that the range for tones without vibrato was 24 cents and with vibrato was 34 cents. Reviews of other vibrato literature are available in Reinders (1995), Rothman (1987), and Timmers (2000).

The past few years have seen an increased interest in the relationship between singing-voice performance parameters and musical structure. Timmers (2007) examined various performance parameters, including tempo, dynamics, and pitch variations manually with PRAAT (Boersma 1993; 2001) for professional recordings of several Schubert songs whose recording dates spanned the last century. In relating these parameters to the musical structure of the piece, she found consistency across performers. She also explored the emotional characteristics of the performances and the ways in which performance style changed throughout the twentieth century. Ambrazevičius and Wiśniewska (2008) studied chromaticism and pitch inflection in traditional Lithuanian singing. They also used PRAAT for analysis and derived a number of rules to explain chromatic inflections for leading tones, as well as ascending and descending sequences. Rapoport (2008) manually analyzed the spectrograms of recordings of the songs by Berlioz, Schubert, Puccini, and Offenbach, and classified each tone based on the relative strength of the harmonics in its spectrum and vibrato rate and depth. He used this analysis to assess the similarities and differences between individual singers' interpretations of the songs. Marinescu and Ramirez (2008) performed spectral analysis to determine pitch, duration, and amplitude for each note in several monophonic excerpts from several arias performed by Josep Carreras. They also analyzed the sung lines with Narmour's implication-realization model (Narmour 1990) and then combined this with a spectral analysis in order to induce classification rules using a decision tree algorithm.

#### **2.2.5.2 Vocal Ensembles**

Hagerman and Sundberg (1980) examined the impact of vowels on intonation accuracy in professional barbershop quartets. These singers were chosen because of their straight-tone style of singing. They found that there was a high degree of precision in the ensembles' they studied and that there was limited influence on their intonation by the type of vowel sung. Ternström and Sundberg (1982) studied amateurs singing intervals against synthesized tones and found that intonation accuracy was highest when the sung notes had strong common

partials with the stimuli tones. They also found that intonation accuracy was greater with less vibrato.

Sundberg (1987) observed a significant amount of variance in phonation frequencies across choirs, especially when vibrato was present. He also noted that accurate intonation required that the singers must be able to hear one another. Sundberg also discussed the intrinsic pitch associated with vowels, but noted that general observable trends concerning the “sharpness” or “flatness” of certain vowels were limited. Ternström and Sundberg (1988) looked at intonation in choirs, specifically the impact of sound pressure level and spectral properties within a choir on their 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 increased amounts of vibrato for simple spectra.

A number of intonation-related studies in the mid-late 1990s were based on listening tests. Ternström (1993) looked at experienced listeners’ preferences of vocal scatter in choirs; the amount of variation in the singers’ mean  $F_0$ . He found that the listeners preferred scatter of less than 5 cents, but that scatter up to 14 cents was acceptable. Bell (1994) performed intonation preference listening tests with trained musicians on synthesized versions of four-part Bach chorales and solo melodies from the same chorales. The solo melody and/or one voice in the chorale were systematically detuned to create drift of 0.5, 2, or 8 cents per second. He found that tuning variations of 2 or 8 cents were more perceptible in the chorale than in the solo melody. Swann (1999) looked at tuning preferences and pitch discrimination in choir singers. He presented participants with chords tuned to equal temperament, Pythagorean, Just Intonation, and mean-tone systems, as well as chords with randomized mistuning. The chords were presented in root position and in inversion. He found that the subjects preferred just-intonation for root position triads, Pythagorean for first inversion triads, and mean-tone for second inversion triads.

Reviews of choral intonation studies until 2002 are available in Ternström and Karna (2002) and Ternström (2002, 2003). Ternström and Karna provide a general overview on studies of choirs, including vocal production, as well as acoustical and perceptual issues. In their discussion of intonation, they proposed a number of factors that influence tuning: the intrinsic pitch of vowels (Ternström et al. 1988), the impact of amplitude on pitch perception, the ensembles’ ability to maintain enough breath support to achieve accurate

pitch, and the influence of the musical materials. Ternström provided an overview of choir acoustics that dated back to the work done by Lottermoser and Meyer in the 1960s. He also described Lottermoser and Meyer's work on the size of major and minor thirds in three choirs. They found that the major thirds tended to be sharp (416 cents) and that the minor thirds tended to be flat (276 cents). Ternström also reviewed work on relative dynamics, dynamic ranges, choral spacing, and his own work on “self-to-other-ratios” in choral settings.

More recently, Jers and Ternström (2005) examined intonation and vibrato in two *a cappella* multi-track recordings of an eight-measure piece by Praetorius. One was recorded at a slow tempo and the other was recorded at a faster tempo. Pitch estimates were made semi-automatically using a correlogram, which visually represents several  $F_0$  estimates for monophonic vocal signals (Granqvist and Hammarberg 2003). Jers and Terström averaged across both the whole ensemble and the individual singers for each note's mean  $F_0$  and standard deviation. Overall, they found that the amount of scatter was greater at the faster tempo than at the lower tempo. They also found increase in the standard deviations of  $F_0$  at note transitions. In terms of vibrato, Jers and Terström observed a certain degree of synchronization between singers at both tempos.

Vurma and Ross (2006) had thirteen professional singers sing isolated major second, perfect fifths, and tritones. They analyzed the recording using PRAAT and found that the major seconds tended to be smaller than equal temperament and that both the perfect fifths and tritones were larger than equal temperament. However, there were some small differences in the averages for ascending and descending versions of each interval. They also ran a listening test with both the singers in the experiment and seventeen amateur musicians, where the subjects were asked to classify the sung intervals as being either in or out of tune. The details of this experiment are discussed in Section 2.3.1.

Howard (2007b; 2007a) examined pitch drift and adherence to equal temperament in two *a capella* SATB quartets.  $F_0$  estimates were calculated by measuring the movement of the larynx with an electroglottograph and SPEAD software by Laryngograph Ltd. Howard argues that in pieces with modulation, choirs used non-equal temperament and that pitch drift is necessary for choirs to stay in tune (Howard 2007b). Howard replicated the experiment with

a second SATB ensemble and found a tendency toward, although not full compliance with, Just Intonation (Howard 2007a).

Vurma (2010) studied *a cappella* two-part singing in an experiment where 15 professional singers sung the upper part from a two-part vocal exercise against a synthesized lower part. In the first half of the exercises, the voices began in unison with the upper voice moving up a minor second against the lower voice moving up by a perfect fourth before both returned to the unison. In the second half, the voices again began in unison with the upper voice moving up and down by a perfect fifth and the lower voice moving up and down by a major third. The accompaniment was presented either in equal temperament or with the melodic intervals either compressed or expanded by 20 or 40 cents. Vurma found that the singers did not respond to the detuning in the accompaniment and remained relatively consistent in their own melodic interval sizes against all of the different tuning conditions. Overall, the median of major seconds tunings for each singer was 94–103 cents and the median of the perfect fifths was 697–706 cents. Vurma also ran a listening test on the recorded data with seven professional musicians. The details of this experiment are discussed in Section 2.3.1.

## 2.2.6 Modeling Expressive Performance

Analyses of performances have demonstrated that all parameters of human performances (timing, dynamics, intonation, and vibrato) show a deviation from what would be produced in a mechanical performance. This section details various approaches to modeling expressive human performances. Comprehensive reviews of the challenges and approaches not covered in this section are available in De Poli (2004), Widmer and Goebl (2004), Widmer et al. (2007) and Goebl et al. (2008).

Todd (1985) developed a model of expressive timing based on Lerdahl and Jackendoff's generative theory of tonal music (Lerdahl and Jackendoff 1983). Todd compared the results of his model with three performances on a piano that could measure the timing of each note attack via photocells near the hammers and found varying degrees of agreement to his model's hierarchical phrase final lengthening projections. He later extended his hierarchical phrase-based model to account for dynamics, defining the relationship between timing and dynamics as “the faster the louder, the slower the softer” (1992) with an increase in tempo and dynamics of the middle of the phrases and a reduction of both at the end of the phrases. The revised model was shown to correspond with a performance of a Chopin prelude.

Windsor and Clarke (1997) evaluated Todd's model on several performances of a Schubert Impromptu. They found that Todd's model did not account for all of the timing and dynamic activity, but they were able to achieve a good fit by applying weightings that put more emphasis on timing at lower structural levels and on dynamics at higher structural levels.

The Director Musices system for generating expressive performance was developed at the Royal Institute of Technology in Stockholm in the 1990s and 2000s (Bresin et al. 2002). The system was built on an “analysis-by-synthesis” approach, where listening tests determined the parameters for the model, in contrast to the measurement-based approach employed in other approaches by Todd. This “analysis-by-synthesis” approach was first described by Sundberg, Askenfelt, and Frydén (1983), and early work on determining the thresholds for the different parameters was done by Sundberg, Friberg, and Frydén (1991). As with Todd's work, a lot of emphasis has been placed on phrase-related rules (Friberg 1995; Sundberg et al. 2003). The researchers have also developed a number of rules related to intonation, specifically that higher pitches are sharper, that melodic intervals tend towards Pythagorean sizes, and that harmonic intervals tend towards beat-free tunings. They also developed a hybrid form of the melodic and harmonic rules for ensemble performances (Bresin et al. 2002; Friberg et al. 2006).

The Director Musices system was evaluated by Zanon and De Poli (2003a, 2003b), who found that the parameters reported in the KTH publications did not conform to their findings. This led Zanon and De Poli to develop a method of tuning these values. Gabrielsson and Juslin (1996) related the KTH model to the emotional aspects of music appraised through listeners' ratings. This work was later expanded by Juslin, Friberg, and Bresin (2002) into a generative computational model of emotion. Ornoy (2008) studied flute and cello intonation in relation to the KTH rules and found that there was not a strong correspondence between the predictions of the rules and the intonation in the recordings.

In the past decade, machine learning techniques have been used to model expressive performance in piano performance. The piano is popular because of the large amount of data that can be obtained from pianos, such as the Bösendorfer SE, which is capable of recording the hammer attack timing and velocity to a computer. Widmer and his/her colleagues at Vienna have developed a multi-level approach where both note-level and

phrase-level timing and dynamics rules are learned from data collected from a set of Mozart piano sonatas (Widmer 2002, 2003; Widmer et al. 2003; Widmer and Tobudic 2003). Their model makes use of two different algorithms for learning activity at the two different levels: a rule-learning algorithm for the note-level activity and a nearest neighbour algorithm for identifying phrase-level activity. An evaluation of their model showed a good agreement between their predictions and actual performances (Widmer 2002; Widmer and Tobudic 2003).

More recent work has explored the use of other machine learning algorithms. Grindlay and Helmbold (2006) made use of hierarchical HMMs to both model and generate timing changes and dynamics in piano performance. Flossman, Gratchen, and Widmer experimented with using both support vector machines (Flossmann et al. 2008) and discrete-state HMMs (Flossmann et al. 2009) for determining both tempo and local note-level deviations from the overall tempo in piano performances. They found that the use of HMMs allowed them to better incorporate performance context.

### 2.2.7 Summary

This section has surveyed the existing work on extracting and describing performance data. The majority of the work in performance analysis has been in timing and dynamics, but there is a small tradition of literature on vibrato and intonation research for both non-fretted string instruments and the singing voice. Early studies used manual methods to extract pitch and timing information from recorded performances. More recent studies have used automatic methods for extracting pitch information, but the identification of note onset and offset locations is still done manually. Overall, studies of intonation in both solo and ensemble vocal performances have shown that singers do not sing in a fixed tuning system. However, the findings have been variable as to whether singers tend to be sharper or flatter than equal temperament or if singers are closer to Just Intonation than equal temperament. The discrepancies in the findings may be related to the smaller number of performances that were studied. The findings for vibrato rate have been more consistent, with several studies reporting the range to be 5–7 Hz. Vibrato depth is more variable, with reports of up to 100 cents.

As detailed in Chapter 1, this dissertation addresses a number of questions not raised in the existing research on intonation. Specifically, this research provides a systematic study of

whether singers tend to perform musically similar patterns in similar ways by studying a larger number of repeated performances and developing ways of analyzing the variations between multiple performances. Also, this dissertation uses a method for automatically extracting intonation data from recorded performances. The background on this method is described in Sections 2.3–5, and the method itself is detailed in Chapter 3.

### **2.3 Pitch and Consonance Perception**

This section provides an overview of both historical and contemporary theories of pitch and consonance perception. Section 2.3.1 describes the basic mechanisms of the auditory system. Section 2.3.2 details the history of the debate over how pitch processing occurs within the auditory system, specifically the spectral and temporal perspectives of pitch perception. Section 2.3.2 discusses the perception of continuous pitch, surveying theories of how the varying pitches within a note, such as vibrato, are perceived as a single percept. Section 2.3.4 extends the discussion of perception to consonance and dissonance.

#### **2.3.1 Overview of the Mechanisms of the Auditory System**

Sound is conveyed as sound waves, perpetuations, or fluctuations in the air. Sound waves can be described in terms of their frequency, the number of cycles of a waveform occurring in a given time, which are measured in cycles per second as hertz (Hz), their amplitude, the amount of difference in air pressure measured in decibels (dB), and their relative phase to one another. When multiple sound waves in whole number ratios are sounded simultaneously with decaying amplitudes as the frequency rises and synchronized phases, they are generally perceived as a single tone. Typically, the lowest frequency is heard as the fundamental. In musical tones, the pattern of amplitude in these harmonics does not decay monotonically, and the relative strengths of the upper partials contribute to the timbre of the tone. This whole-number relationship is referred to as the overtone, or harmonic, series. Figure 2.3.1 shows the first seven notes close to the overtone series of the note A1. In general, the individual partials can only be heard out by someone practiced in analytical hearing, which is how, with the assistance of resonators, 19<sup>th</sup>-century theorists developed their ideas about the structures of complex tones, as discussed in Section 2.3.2–3.

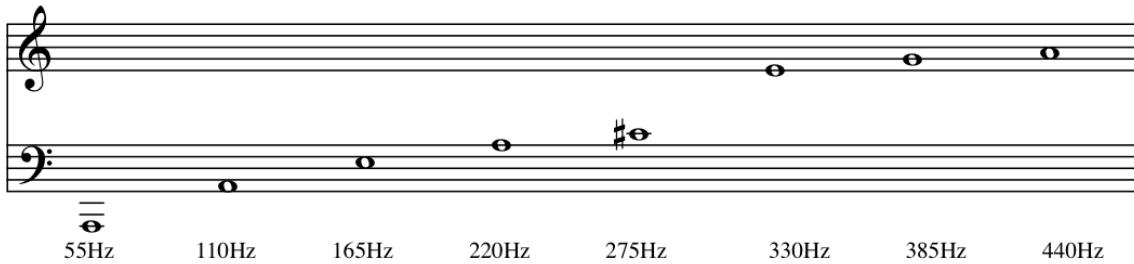


Figure 2.3.1: Overtone series from the note A1 (55 Hz).

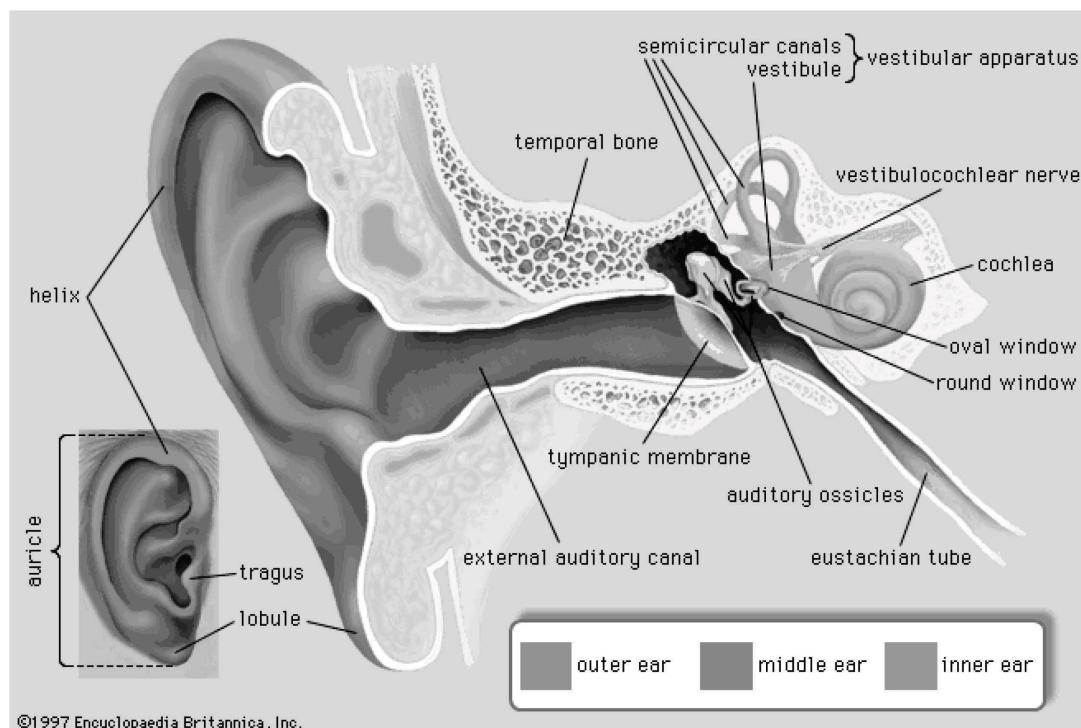
The fact that the intervals between 55 Hz and 110 Hz and between 110 Hz and 220 Hz are perceived as being the same (i.e., one octave) illustrates the logarithmic nature of pitch height perception. Intervals can be represented linearly in cents, where each cent is 1/100 of an equal-tempered semitone. The relationship between cents and frequency is defined in Equation 1, where *note1* is the lower note in the interval, measured in Hz, and *note2* is the upper note, also measured in Hz.

$$cents = 1200 \times \log_2 \left( \frac{note2}{note1} \right) \quad (1)$$

Under ideal conditions, the human auditory system perceives frequencies between 20–20,000 Hz, although the upper limits range decreases with age and exposure to loud sounds. The loss of the upper range mainly affects the perception of timbre, as the range of the fundamental of the human voice for speaking is approximately 85–180 Hz for men and 165–255 Hz for women, extending up to 1100 Hz for the singing voice. A human's ability to distinguish differences in frequency between two notes, referred to as the just noticeable difference (JND) or difference limen (DL), depends on a number of factors, including frequency, amplitude, note duration, timbre (pure vs. complex tones), and listening conditions. Most studies on the JND had subjects either adjust or listen to intervals of synthetic tones (either sinusoidal or complex), many of which are described in Burns (1978). Vurma, in his work with Ross (2006) and alone (Vurma 2010), has studied JND in intervals with sung stimuli. In both experiments, expert listeners were asked to indicate whether a target note in an interval was in tune, sharp, or flat. Vurma and Ross (2006) focused on melodic intervals and found tuning deviations as large as 20–25 cents were considered to be in tune by their subjects. They also found that their subjects tended to classify the perfect fifths and tritones as out of tune more frequently than the thirds. Vurma (2010) used

harmonic major and minor thirds between a synthesized lower part (that deviated either 0, 20, or 40 cents from equal temperament) and a sung upper part as stimuli. He asked his subjects to assess the tuning of both parts in separate trials and found that there was a higher correlation between the measured pitch deviations and the subjects' assessments for the synthesized tones than for the sung tones. Overall Vurma's subjects had only a 34% success rate for tuning deviations of 20 cents, and a 58% success rate for tuning deviations of 40 cents. The difference between the JND in isolated conditions with synthesized and musical contexts can be attributed to categorical perception of intervals (Fyk 1995).

The peripheral human auditory system consists of both the ear and the nerves that connect the ear to the brainstem, as shown in Figure 2.3.2. Sound waves pass through the external auditory canal into the outer ear and are conducted by the tympanic membrane, commonly known as the eardrum, and the three ossicles (the malleus, incus, and stapes) through the middle ear to the oval window of the cochlea, in the inner ear. Both the oval and round windows are membranes that vibrate in opposite phases to one another. The auditory processing in the inner ear takes place in the cochlea.



©1997 Encyclopaedia Britannica, Inc.

Figure 2.3.2: Schematic of the human ear (*Helix: Structures of the Human Ear*. 1997).

Within the cochlea is the organ of Corti, which is divided in two by the basilar membrane. Different parts of the basilar membrane are attuned to different frequencies. On the basilar membrane are the auditory hair cells, and on each hair cell are multiple projections called stereocilia. The frequency response of the basilar membrane is organized from high to low along, a schematic of which is shown in Figure 2.3.3, and function like an overlapping filter bank (Patterson et al. 1992). Studies with notch-noise and rippled-noise have led to the development of a formula for calculating the equivalent rectangular bandwidth (ERB) of each filter (Moore and Glasberg 1983; Glasberg and Moore 1990). The output of the basilar membrane is a neural spike code, which travels, along with balance information, through the Vestibulocochlear nerve to the brain.

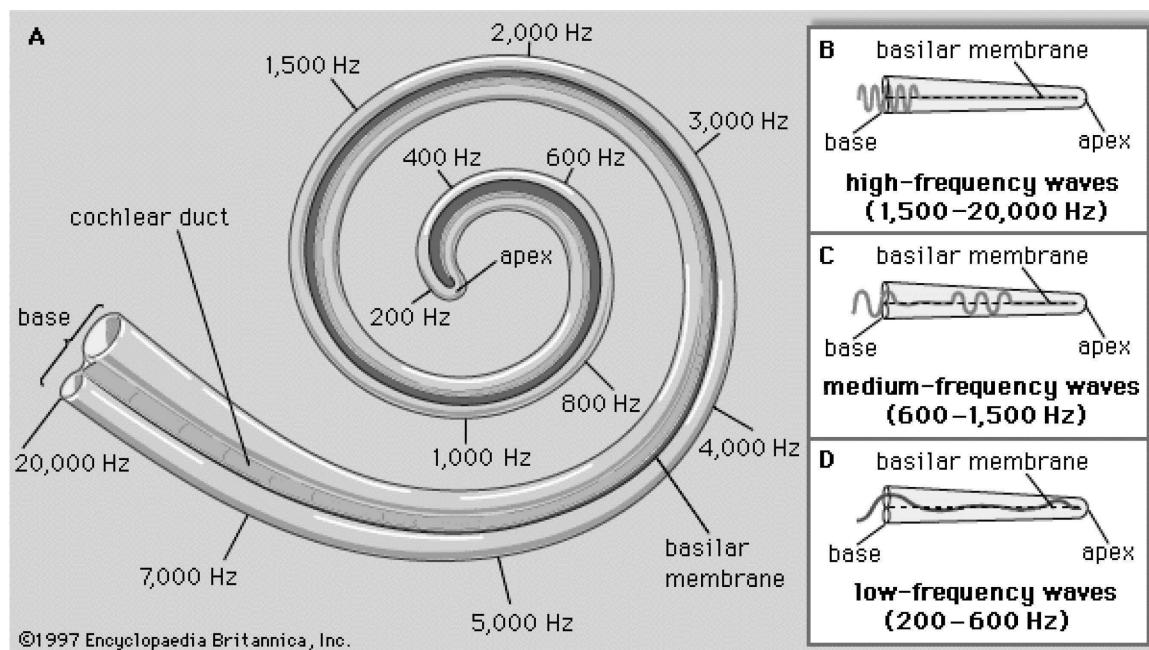


Figure 2.3.3: Schematic of the basilar membrane (*Hearing: Basilar Membrane*. 1997).

### 2.3.2 Pitch Perception

As detailed in Section 2.3.1, initial frequency analysis takes place in the cochlea, when there is a place to frequency correspondence; however, this frequency analysis does not create a pitch percept. Rather, it is likely formed in the cortex or the brainstem, and the frequency analysis information is sent there by neural spike code. Figure 2.3.4 shows a rough model of where different processes occur in the auditory system.

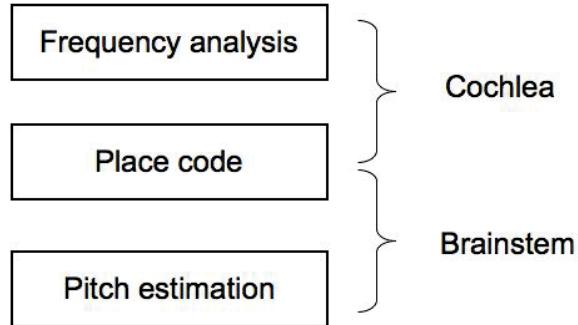


Figure 2.3.4: Pitch processing in the auditory system.

Section 2.3.2.1 provides an overview of the history of developments in pitch perception from ancient Greece to the 1960s, and Section 2.3.2.2 describes the main divide in more recent literature between the spectral and temporal theories. This section sets up the next main section of this chapter, 2.4, and discusses the link between these theories and signal processing approaches to automatic pitch estimation.

### 2.3.2.1 Historical Overview of Pitch Perception

The spectral approach to pitch perception argues that pitch is determined solely by the cochlea's reaction to frequencies. De Cheveigné (2005) argues that the roots of pitch perception theories related to the place of excitation in the cochlea, known as place theories, can be traced back to the ancient Greeks with Pythagoras in the 6<sup>th</sup> century BCE, who performed a number of investigations on a monochord in order to explore the relationship between string length and frequency. Their investigations demonstrated that when a plucked string is divided in half, its resultant pitch is double the pitch of the whole string. De Cheveigné also suggests that temporal theories of pitch perception date back to Nichomachus (c. 100 CE), who observed that sound is made up of a sequence of pulses and that the speed of a sequence of sounds determines its pitch. Boethius later echoed Nichomachus' observations (c. 520 CE) and was the main source for the transmission of Greek thought in Europe (Bower 2002).

In the 16<sup>th</sup> and 17<sup>th</sup> centuries, Mersenne's and Sauveur's explorations of the partials of complex instrumental sounds laid the foundation for modern spectral theories. Mersenne (1636) observed that the fundamental frequency of the string is inversely proportional to the mass. Sauveur (1701) found that a string could vibrate at several frequencies simultaneously,

and coined the words fundamental and harmonics for this phenomenon. Other works from the 16<sup>th</sup> and 17<sup>th</sup> centuries more closely related to temporal pitch perception theories. Both Galileo and Mersenne experimented with measuring sound vibrations, while both Mersenne and Descartes studied the relationship between vibration rate and the speed of propagation (de Cheveigné 2005).

It was only in the 18<sup>th</sup> century that more complete models of how the auditory system perceives pitch began to emerge. Seebeck, Ohm, and Helmholtz all developed their own theories of pitch perception (Turner 1971), much of these theories built on Fourier's observation that any complex signal can be represented as a sum of weighted sinusoids (Fourier 1820). These frequency-based theories argued that pitch is determined by the fundamental. In contrast, periodicity-based theories, for example, by Schouten, argued that pitch is determined by contributions from all of the partials and that the perceived pitch is the period at which they all coincide (i.e., the fundamental). By including partials in his model, Seebeck was able to explain the phenomenon of the "missing fundamental," where the same pitch is perceived when the fundamental is absent. Ohm believed that the perceived pitch was conveyed solely by the fundamental and that the partials contributed only to the timbre of the tone. He dismissed Seebeck's missing fundamental theory as an illusion. Helmholtz (1863) sided with Ohm, arguing that in cases where the fundamental is missing acoustically, it is generated by the interaction of the partials in the middle ear, thus creating the perception of pitch. Helmholtz also put forth a physiological model of pitch perception related to Fourier's theorem. Helmholtz posited that the basilar membrane had a set of taut fibres that are attuned to different frequencies and that resonate when the corresponding frequencies occur in a sound.

As noted above, the "missing fundamental" issue marked a divide between Seebeck's and Ohm's theories. Nearly a century later, Schouten's work on residue pitch demonstrated the validity of Seebeck's theory (Schouten et al. 1962). Schouten showed that even when the fundamental was removed from a complex tone, the perceived pitch remained the same. He termed this pitch percept the residue pitch, though it is also known in the literature as periodicity, virtual, and low pitch. Seebeck's findings were confirmed by Licklider (1954) in an experiment where he showed that the pitch remained unchanged when the fundamental was masked by noise.

Schouten's revisit of Seebeck's "missing fundamental" explanation led to a number of inquiries regarding the dominance region for pitch perception. The significance of the dominance region is that partials that occur within it have more influence on the perceived pitch than partials outside of it. Plomp (1967) and Ritsma (1967) found the overall dominance region for pitch perception to be 500–2000 Hz. Plomp observed that the dominance region for complex tones depends on the fundamental frequency. For example, for tones above 1400 Hz, the fundamental is dominant; for tones 700–1400 Hz, the second harmonic and higher are dominant; for tones 350–700 Hz, the third harmonic and higher are dominant; and for tones up to 350 Hz, the fourth harmonic and higher are dominant. Ritsma demonstrated that the dominance region within the range of typical musical instruments (100–400 Hz) was the third, fourth, and fifth harmonics. Within this range, the third and fourth harmonics dominated in tones between 100 and 200 Hz, and the second harmonic dominated in tones above 400 Hz.

The dominance region for frequency is closely related to another important development for theories of pitch perception: the critical bandwidth for the auditory filters. First described by Fletcher (Fletcher 1940), the critical bands indicate which partials are resolved in the cochlea and which are unresolved. With resolved partials, there is sufficient separation between partials that they excite different parts of the basilar membrane. For unresolved partials, frequency information can only be extracted by the temporal relationship between the neural spikes generated by the basilar membrane. According to Shackleton and Carlyon (1994), partials are resolved when there are fewer than 2 partials in a critical band and unresolved when there are more than 3.25. They considered 2–3.25 partials per band to be a transition region. Resolved and unresolved partials have implications for spectral and temporal theories of pitch perception since spectral theories rely on the place of resolution on the basilar membrane that only occurs with resolved partials. In contrast, temporal theories rely on peaks in the pattern of neural spikes that are produced in the cochlea, often referred to as inter-spike intervals. The next section describes the development and differences between these two approaches in greater detail.

### 2.3.2.2 Contemporary Spectral and Temporal Theories of Pitch Perception

Spectral, or place, approaches to perception fall into the tonotopic tradition that began with Helmholtz, where pitch perception is based on the resolution of partials on the basilar

membrane. Goldstein (1973) developed a model of pitch perception for a complex tone that uses a “harmonic sieve” to match the partials of the tone to an estimated pitch percept. Goldstein’s approach was limited to resolved partials and single tones. Wightman (1973) also used a pattern recognition approach; his two-step process began with a Fourier-based analysis of the power spectrum followed by peak-picking to determine pitch. Terhardt (1974) expanded the idea of the “missing fundamental” in his virtual pitch theory, which builds on Helmholtz’s theory that there are two types of listening: analytic, where each partial is heard individually, and synthetic, where the composite of the partials is heard as a single pitch. Terhardt argues that synthetic listening uses a pattern matching mechanism that is learned from exposure to speech (Terhardt 1979). The implications of Terhardt’s theory on the perception of consonance are discussed in Section 2.3.4.3.

As noted above, problems with spectral approaches arise due to the resolvability of partials. The cochlea can resolve lower partials because they are separated by more than one critical band, so the place theory holds for pitch perception of pure tones or complex tones from low frequency components. However, even higher, unresolvable harmonics can produce the perception of a weak pitch, which indicates that there is another process in the auditory system that can derive pitch from these cues (de Cheveigné 2005). This process is based on the temporal relationships of the neural spike code generated by the cochlea, where frequency information is derived from the spacing of the neural spike code.

Licklider (1951) proposed that pitch perception could be explained by autocorrelation, a type of self-similarity measure, on the intervals between the neural auditory spikes that are generated by the hair cells on the basilar membrane. The activity of the hair cells was later modeled by Meddis and Hewitt (1991) as compression, half-wave rectification, and low-pass filtering. They also proposed the use of a summary autocorrelation approach across the analyzed frequencies to predict pitch, which was later expanded by Meddis and O’Mard (1997). De Cheveigné (1998) described a variation on the autocorrelation approach that uses a difference function to implement a cancellation model of pitch that is, unlike autocorrelation approaches, sensitive to phase and capable of explaining polyphonic pitch perception. More recently, Balaguer-Ballester and colleagues (2008; 2010) have proposed a cascade autocorrelation model, where a cascade of filters was added to the summary autocorrelation function of Meddis and O’Mard.

It has not been determined exactly what the role of neural spike timing is in pitch perception, that is, how much the spikes contribute to the determination of frequency for resolved partials (Oxenham et al. 2009). It cannot be assumed that the timing between these spikes is solely responsible for pitch perception: experiments with dichotic, or binaural, pitches have shown that when the harmonics are spread across the ears, pitch can still be perceived (Bilsen 1977; Raatgever and Bilsen 1986). Such experiments support the theory that there is a pattern matching mechanism for resolved harmonics that works in combination with a temporal mechanism (Bernstein and Oxenham 2003).

The role of unresolved harmonics in pitch perception has been examined recently with experiments using mistuned harmonics. Bernstein and Oxenham (2008) looked at the relationship between  $F_0$  discrimination and resolved versus unresolved individual harmonics. They found that detuning of odd harmonics increased  $F_0$  discrimination in tones with only the tenth harmonic and higher. They argued that  $F_0$  discrimination ability is more impacted by whether the individual harmonics fall into separate spectral filters rather than spectral resolution of the harmonics. They related their findings to their earlier revisions to the autocorrelation model of pitch perception (Bernstein and Oxenham 2005), which they expanded with a summary autocorrelation function across the auditory channels. Moore and Glasberg (2010) revisited their findings and argued that the results of Bernstein and Oxenham (2008) could instead be explained by temporal fine structure in each auditory channel, and that details are lost when a summary autocorrelation is performed. More information about the role of temporal fine structure in pitch perception can be found in Moore (2008), and a further investigation of the role of temporal fine structure in  $F_0$  determination using unresolved harmonics is available in Oxenham, Micheyl, and Keebler (2009).

Some recent work by Pressnitzer and Patterson (2001) looked at the role of combination tones in pitch perception. Combination tones occur when either the sum or the difference of two simultaneously sounding tones is perceptible. Pressnitzer and Patterson recognized that such tones are not necessary for pitch perception, as Helmholtz had initially argued, but which was later disproved by Schouten. Rather, they explored the question of whether combination tones contribute to pitch perception in the absence of partials in the dominance region. Researchers such as Dai (2010) have argued that the existence of combination tones,

along with spectral edges, allow subjects to perceive the pitch of such tones without relying on temporal mechanisms.

There has also been a small body of work on the perception of pitch in musical contexts. McDermott and Oxenham (2008) provided a survey of what is known about auditory mechanisms for pitch perception and the perception of pitch relationships in a musical context. They concluded that while the discernment of individual pitches has been shown to take place in the auditory system, the more structurally related aspects of music listening are taken care of by neural mechanisms that have not yet been fully explored. Marmel, Tillman, and Dowling (2008) looked at the role of tonal expectations on pitch perception, finding that when tonally related notes were mistuned, the pitch processing time increased. Bharucha (2009) explored the parallels between the mapping of frequency to pitch and the mapping of tones to tonal relationships, such as chords and keys. He also looked at the role of hierarchical self-organization in the perception of musical structures. Other open questions about pitch perception are surveyed in Plack (2005), and more thorough reviews of pitch perception are available in de Cheveigné (2005) and Yost (2009).

### **2.3.3 Perception of the Pitch of a Single Tone**

The synthetic stimuli used in pitch perception experiments do not replicate the acoustic variability found in natural tones. Though variability may occur in terms of pitch, loudness, and timbre, this section will focus on how variability in pitch influences pitch perception, specifically for the singing voice and non-fretted string instruments. Outside of the pitch-related instability at the beginnings and endings of notes, these tones generally have a degree of vibrato. Vibrato is a systematic variation of fundamental frequency over the stable portion of the note that is characterized by its depth (the amount of pitch change) and rate (the speed of the pitch change). Vibrato is most prevalent in the singing voice and unfretted string instruments; therefore, the majority of research on the perception of a single pitch in tones with vibrato has focused on these instruments. The physiology of vibrato is described in Hirano (1995), and studies of vibrato in performance practice were surveyed in Section 2.2.

The first studies of the perceived pitch of tones with vibrato were done in the 1930s at the University of Iowa (Tiffin 1931; Metfessel 1932). These early studies used a tone whistle mounted on a crank to produce a synthetic tone of 420 Hz with vibrato ranging from 0–200

cents. Metfessel simply reported that the perceived pitch was the mean of the vibrato, but that there was some variation in his subjects' responses. Tiffin provided more detail, reporting the average perceived pitch across subjects was 1 cent flat of the geometric mean with a standard deviation of 10 cents. In Japan, Hirose (1934) ran experiments using stimuli with frequencies of 900, 1100, and 1500 Hz and vibrato ranging from 12–361 cents. He found that the perceived pitch depended on the width of the vibrato: vibratos with smaller widths were perceived as slightly sharp of the mean, while vibratos with larger widths were slightly flat of the mean. Seashore (1938) later reported that the studies undertaken at Iowa used the simplified assumption that pitch of a tone with vibrato is the mean. This assumption was also used by Bjorklund (1961) in his study of soprano voices.

Sundberg (1972, 1978a) studied the perceived pitch of synthesized sung tones. He looked at the effects of both vibrato and the singer's formant by presenting subjects with different versions of the stimuli tones (straight, with vibrato, with singer's formant, and with both vibrato and singer's formant), and asked them to adjust a reference tone to match the pitch of the stimuli tone. He found that, on average, neither the addition of vibrato with a maximum width of +/- 1.7% of  $F_0$  around the center frequency or the singer's formant changed the perception of the pitch, though there were some differences observed in both the average and standard deviation of individual responses. Shonle and Horan (1980) used a similar adjustment method for square-wave stimuli tones ranging from 220–1500 Hz to test the influence of vibrato on perceived pitch. They found that the average perceived pitch was closer to the geometric mean than the arithmetic mean. Iwamiya, Kosugi, and Kitamura (1983) revisited the findings of Tiffin, Hirose, and Shonle and Horan with experiments on both vibrato tones and trills. They also found that the perceived pitch of both vibrato and symmetrical trills is roughly the center frequency; however, there was a shift either up or down for asymmetrical trills that corresponded to the direction of the asymmetry. Later, Iwamiya and colleagues expanded their study to examine the role of amplitude modulation in the perceived pitch of vibrato (Iwamiya and Fujiwara 1985; Iwamiya et al. 1994).

D'Alessandro and Castellengo (1994, 1995) studied the perceived pitch in short vibrato tones in an attempt to better model the perception of pitch in actual singing practice. They found that the  $F_0$  at the end of the note was more significant for the pitch perception than the beginning of the note. They also argued that taking the mean of the steady-state portion of

the note rather than the mid-point between the maximum and minimum frequencies produces a more robust estimate of the perceived pitch. Brown and Vaughn (1996) studied pitch perception for vibrato in stringed instruments, using real samples of violins rather than synthesized tones. The subjects were presented with a pair of tones and asked to determine whether the second tone was higher or lower than the first. Brown and Vaughn's results confirmed the results of experiments with synthetic tones. They also provided a good survey of the results of earlier research, including the work of Hirose, Tiffin, Seashore, Shonle and Horan, and Iwamiya and colleagues. Yoo, Sullivan, Moore, and Fujinaga (1998) built on Brown and Vaughn's study and looked at differences in response time when reporting whether the second tone was higher or lower. They found that it took subjects longer to make a determination for vibrato tones than for non-vibrato tones.

Gockel, Moore, and Carlyon (2001) revisited d'Alessandro and Castellengo's finding that certain parts of the tone influenced the perceived pitch more than others. In their experiments with sinusoids, they found that the parts of the tone with slower frequency modulation contributed to the perceived pitch more than the parts of the tone with faster frequency modulation. They proposed that the perceived pitch be calculated as a weighted average favouring the slower moving portions of the note, rather than taking an unweighted mean over the duration of the note. Mesz and Eguia (2009) developed a model to better explain how the frequency instability that occurs in natural tones affects the perceived pitch for vibrato, which can also predict Gockel, Moore, and Carlyon's results. They propose a three-part algorithm that begins with an initial instantaneous frequency analysis for each frequency band, or channel. The instantaneous frequency of each channel is used to determine the rate of the  $F_0$  fluctuations, and then the rate calculations are used to determine the weightings for calculating the perceived pitch.

### **2.3.4 Consonance Perception**

Consonance, and by extension dissonance, can be considered as both a sensory phenomenon and as a cultural, or musical, one. Section 2.3.4.1 details the study of sensory consonance since Helmholtz, for both pure and complex tones. A survey of precursors to Helmholtz's work is available in Hoffman-Engl (2010). Section 2.3.4.2 details recent theories of musical consonance.

#### 2.3.4.1 Sensory Consonance

In his theory of consonance and dissonance, Helmholtz (1863) 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. He argued that the degree of consonance between complex tones was dictated by the degree of roughness, or beating between partials. Beats are audible up to  $\sim 20\text{Hz}$  and beyond that a rattling sensation, or roughness is produced. Beating is a tremolo produced by interference between tones of proximate frequency. The rate of the tremolo is determined by the difference in frequency between the two tones. 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, or less dissonant, the resultant sound is. The theory of sensory consonance makes a case for Justly tuned vertical intervals, as there is a greater coincidence of partials between them than with tempered intervals.

Plomp and Levelt (1965) revisited some of Helmholtz's ideas through a series of tests on untrained subjects. They also summarized the relevant work since Helmholtz, including studies on difference tones and fusion. Plomp and Levelt discovered that an interval size, called the critical band, is also a significant factor in the perception of consonance. The size of the critical band is a function of frequency, but it is about a minor third for pitches in the range of typical musical instruments or the voice. Their general results, using sine tones, indicated that their subjects judged intervals less than a minor third and greater than a unison as a sensory dissonance and judged intervals a minor third or greater as sensory consonance. They also demonstrated that the same interval in a lower frequency range was generally perceived as being less consonant than the same interval in a higher frequency range.

Plomp and Levelt also applied their critical band findings to the interactions between the partials of pairs of complex tones using the results of their sine tone experiments. They calculated the dissonance of each complex interval by summing the dissonance of adjacent pairs of the first six partials. This resulted in a different relationship between consonance and frequency than for simple tones (see Figure 2.3.5). Rather than maximum consonance centering around the minor third, the consonance peaks for complex tones occur for intervals with simple ratios (2:1, 3:2, 5:3) where there is a maximum coincidence of partials.

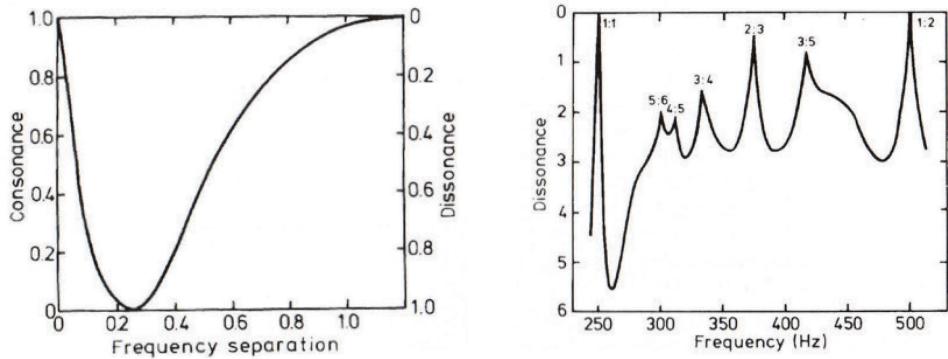


Figure 2.3.5: Plots of consonance as a function of frequency from Plomp and Levelt (1965), as reprinted in Rasch (1999). The plot on the left represents the relationship between consonance (y-axis) and frequency separation (x-axis) for simple tones. The plot on the right represents the relationship between consonance (y-axis) and frequency separation (x-axis) for complex tones with six harmonics.

A study by Kameoka and Kuriyagawa (1969a) showed that the level of consonance associated with two-tone intervals of pure tones forms a V-shape when plotted over an octave. Maximal sensory consonance is perceived at the unison before it declines towards the critical band. The amount of consonance then increases again towards the octave, though it does not reach the same level as for the unison. They observed that consonance was determined by frequency separation and sound pressure level in combination with frequency ratio rather than just being a function of frequency ratio alone. Kameoka and Kuriyagawa (1969b) extended this study to complex tones with 3–12 components, using the dissonance calculation methods that they developed in their earlier work. Their results are similar to those of Plomp and Levelt's for simple complex tones when the sound pressure level is held constant. This is demonstrated in the comparison of the plots in Figure 2.3.5 and Figure 2.3.6.

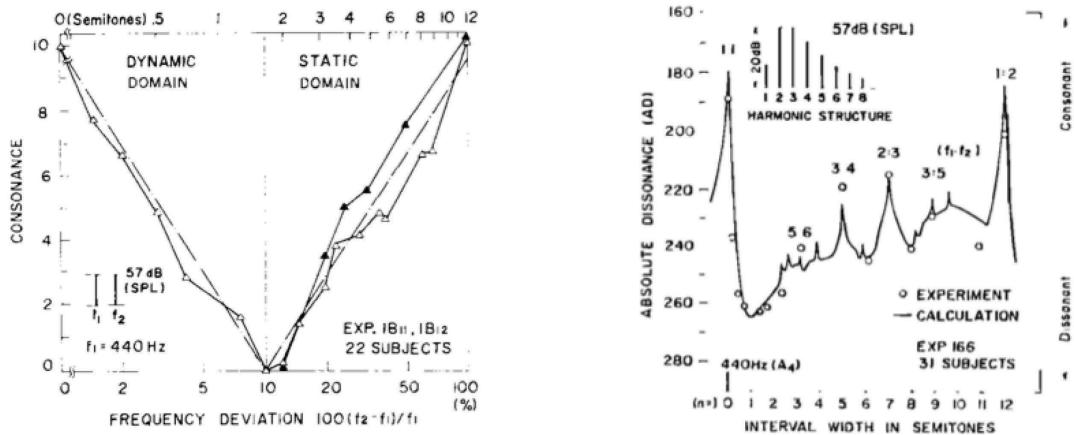


Figure 2.3.6: Plots of consonance as a function of frequency as reported by Kameoka and Kuriyagawa (1969a, 1969b). The plot on the left represents the relationship between consonance (y-axis) and frequency deviation (x-axis) for simple tones. The plot on the right represents the relationship between consonance (y-axis) and intervals for complex tones with eight harmonics.

Hutchinson and Knopoff (1978) developed a formalism, built on Plomp and Levelt's work, for predicting the sensory dissonance of dyads of complex tones, ranging from a minor second to a double octave. Hutchinson and Knopoff found an agreement between the results of their method and the relative consonance ratings reported by Malmberg (1918) from his experiments with piano dyads. They also considered the impact of real instrument timbres on the perception of consonance and dissonance. Sethares (1993) delved deeper into the timbral implications of Plomp and Levelt's work, presenting a computational method for predicting sensory consonance for different natural and synthesized timbres.

Swallowe, Perrin, Sattar, Colley, and Hargreaves (1997) evaluated the role of exposure in the amount of unpleasantness experienced by listeners when they hear acoustic dissonance with a study using four component complex tones. They found that outside of the critical band, cultural rather than acoustic factors influenced a subject's perception. Pressnitzer and McAdams (1999) explored the role of temporal aspects of timbre in the perception of roughness, which is associated with fluctuations in spectral components. To test this, they varied the phase of one of the partials and the spectral envelope of a synthetic complex tone.

They found that such modifications have a significant effect on the perceived roughness. Both of these studies call into question the ability of purely spectral based models, such as Hutchinson and Knopoff's and Sethares' to predict sensory consonance and dissonance. Schön, Regnault, Ystad, and Besson (2005) reinforced the role of musical context in the perception of consonance and dissonance in their study on whether exposure to sensory consonance and dissonance elicited any event-related brain potential (ERP) effects. This view has been challenged by McDermott, Lehr, and Oxenham (2010), who found that harmonicity (i.e., the presence of a harmonic spectra) was the only consistent correlate with listeners' reports of consonance. Their finding that the use of harmonicity as a perceptual cue is not experience-based was later reinforced by Plack (2010).

#### 2.3.4.2 Musical Consonance

Terhardt (1984) expanded the work of Plomp and Levelt with a theory of consonance that reconciles psychoacoustic phenomena of sensory consonance, which he linked to Helmholtz's concept of *Konsonanz*, with musical consonance, which he aligned with both Helmholtz's theory of *Klangverwandtschaft* and his own virtual pitch theory (Terhardt 1974). His theory of consonance prioritizes intervals with low ratios, which can be thought of as those occurring lower in the harmonic series; the unison (1:1); the octave (2:1); the fifth (3:2); and the major third (5:4). There is a discrepancy between this hierarchy and some of the ordering of consonance that emerges from studies of sensory consonance: the unison (1:1), the octave (2:1), the fifth (3:2), the major sixth (5:3), the fourth (4:3), the minor third (6:5), and the major third (5:4) (as demonstrated in Figure 2.3.5).

Terhardt's virtual pitch theory draws on both Schouten's residue theory and Rameau's theory of the fundamental bass (Rameau 1722), and argues that consonance is created by a whole number frequency relationship between the elements of a chord and its fundamental bass. In his 1737 treatise, *Generation Harmonique*, Rameau attempted to reconcile his fundamental bass theory with his understanding of harmonics, which he referred to as the *corps sonore* (Rameau 1737). Rameau's fundamental bass did not coincide with the root of the chord: rather, it was a sub-root for which all of the chord tones coincided with its partials. Rameau was, however, unable to provide an acoustical foundation for all of the aspects of his theory of harmony, such as the minor triad (Christensen 1993). For example, the fundamental bass of a close position major chord on C4 would be C2 (the C two octaves

below) since C4-E4-G4 coincide with the 4<sup>th</sup>, 5<sup>th</sup>, and 6<sup>th</sup> of C2's overtone series. The virtual pitch theory suggests that the perception of harmonic consonance in Western art music is dependent on the mind's acquisition of an acoustical template. Terhardt argues that this template, based on the harmonic series, allows the listener to perceive the pitch of a complex tone 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 a chord corresponds to the virtual fundamental note that is suggested by the template. When the sounded bass note and the virtual fundamental note align, the sonority is perceived as consonant. The learning process associated with the acquisition of this template allows for 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 theory to support a further argument that the basis of harmonic consonance, like sensory consonance, is psychoacoustical rather than cultural. Terhardt's theory of consonance has been used by Parncutt (1989) as the basis for a reinterpretation of western harmonic practices and by Huron (2001) in a study of voice leading practices.

In contrast with Terhardt's explanation of musical consonance, there have been attempts to explain western harmonic practice without reference to any cultural factors. Cook's model of musical consonance (Cook et al. 2004; Cook 2009) incorporates a psychoacoustic model of harmonic instability that extends Plomp and Levelt's model to triads with a tension model of triads and modes. Lots and Stone (2008) made an argument for the primacy of certain intervals in the Western tonal system on account of neural synchrony.

#### **2.3.4.3 Consonance and Tuning**

Vos has addressed the issue of how consonance relates to tuning perception and preferences. In his earlier work, Vos (1982) showed that there was a lower discrimination threshold for Just vs. tempered perfect fifths between complex tones (702 vs. 700 cents) than for Just vs. tempered major thirds (384 vs. 400 cents). Vos looked at three different tone duration (250, 500, and 1000 milliseconds) and found that the discrimination threshold

decreased as the tones got longer. He also showed a correlation between the perceived strength of beats and the discrimination threshold for tempered intervals (Vos 1984).

Vos and van Vianen (1985b) built on Vos' earlier findings and looked at all the intervals that could be formed by frequency ratios of integers up to 8. They evaluated the role of complexity by equating an increase in complexity with an increase in the value of  $p + q$ , where  $p$  and  $q$  are the integers in the ratio ( $p:q$ ). Vos and van Vianen demonstrated that, for synthetic tones, the thresholds for discrimination between pure and tempered intervals increased when the complexity of the interval increased. They also found that the discrimination threshold was not influenced by fundamental frequency (Vos and Van Vianen 1985a).

### 2.3.5 Summary

This section began by briefly surveying the process by which humans perceive pitch, both in terms of the mechanisms of the human auditory systems (Section 2.3.1) and theories about how humans process frequency information (Section 2.3.2). Section 2.3.3 addresses the question of how continuous fluctuations in pitch are perceived as a single percept (e.g., vibrato). The latter part of the section focused on consonance as a sensory phenomenon and how this type of consonance impacts tuning discrimination.

The research described in this section impacts the work done in this dissertation in two ways. First, the work described in Section 2.3.3 provides a guide for describing the  $F_0$  estimates extracted from the experimental recordings. As will be detailed in Section 3.2, these findings are used to describe the perceived pitch of a single tone. Specifically, the Gockel, Moore, and Carlyon method(2001), which takes into account the rate of change of the frame-wise  $F_0$  estimates by weighting each frame's  $F_0$  in the overall pitch estimate, is used. The work on the perception of consonance in Section 2.3.4 is useful for understanding the pitch estimates, and the work on sensory consonance in Section 2.3.4.1 is an important consideration for vertical tunings. Also, the research described in Section 2.3.4.3 done individually by Vos (1982) and collaboratively with von Viannen (1984; 1985a, 1985b) is used in interpreting the results of the ensemble experiments described in Section 4.2.

## 2.4 Fundamental Frequency Estimation and Transcription

This section builds on Section 2.3 by considering how spectral and temporal understandings of pitch perception influence work in the area of automatic  $F_0$  estimation. The following sections describe a range of techniques for both monophonic and polyphonic estimation methods using a categorization proposed by de Cheveigné (2006): spectral, temporal, spectro-temporal, and, for polyphonic estimation, learning-based approaches. For most harmonic signals, monophonic fundamental frequency estimation can be considered a solved problem. Polyphonic fundamental frequency estimation, however, is still an open problem. Section 2.4.1 discusses approaches to monophonic estimation, including the YIN algorithm (de Cheveigné and Kawahara 2002) that is used for the experiments in this dissertation.

### 2.4.1 Monophonic Estimation

Spectral approaches to fundamental frequency estimation can take advantage of the efficiency of the fast Fourier transform (FFT), but are dependent on good frequency resolution. There are a number of different spectral approaches to monophonic fundamental frequency estimation. One makes use of a filter bank, where the difference between peaks in the output is tallied in a histogram and the bin with the largest value is assumed to be the  $F_0$ . This approach can run into problems when the partials of the note are stronger than the fundamental (Martin 1982). Another approach is to take the cepstrum of the signal (the inverse FFT of the log of the FFT), which gives the slow moving (resonant) characteristics of the signal in the lower part of the result and the fast moving characteristics of the signal in the higher part—the first large bin of the higher part can be used as an estimate for the fundamental frequency (Noll 1967). A third approach is to apply an FFT to the signal, which provides phase and amplitude information at each frequency bin for each temporal frame (Schroeder 1968). A mapping of the frequencies in the signal can be constructed by taking the derivative of the phase of the bins with the largest amplitude value. These frequencies, however, are quantized to the mid-point of the bin, so it is advantageous to look at the phase values in two sequential frames to improve frequency estimation accuracy. The  $F_0$  can either be assumed to be the strongest frequency or a harmonic (timbral) mapping can be applied to determine the fundamental of the harmonic complex. The filter-bank and FFT approaches can be extended to account for the presence of multiple  $F_0$ s, as will be discussed in Section 2.4.2. However, the cepstrum method is only useful for single  $F_0$  estimation.

Temporal approaches to fundamental frequency estimation are dependant on good temporal resolution. They can also deteriorate in the presence of high numbers of inharmonic partials because they rely on the detection of self-similarity in the signal. At the same time, temporal approaches are an appealing alternative to spectral approaches because they can be formalized in a way that is close to the mathematical definition of periodicity. Simple temporal approaches include measuring intervals between land marks, counting positive going zero-crossings, performing full wave rectification, or squaring after low-pass filtering. These simple approaches do not work for all signals, though their effectiveness can be increased through pre-processing. A more robust approach is autocorrelation, which was, as detailed in Section 2.3.2.2, first suggested by Licklider (1951). This approach was first implemented into an automated  $F_0$  estimation algorithm by Rabiner (1977) as a running autocorrelation function:

$$r_t(\tau) = \frac{1}{\omega} \sum_{j=t+1}^{t+\omega} x(j)x(j + \tau) \quad (1)$$

$\tau$  is the lag

$\omega$  is the size of the window

$t$  is the time at which the calculation is made

The autocorrelation function compares one frame of the signal to another frame that was extracted from the signal after a certain temporal lag. The lags with the highest scores are retained and typically the longest lag is assumed to be the fundamental frequency. One problem with autocorrelation is that there can be a spike around lag zero, so typically a normalized version of the function that accounts for this is generally used.

Following this approach, de Cheveigné and Kawahara (2002) developed the YIN algorithm, which uses a squared difference function instead of the running autocorrelation algorithm shown in Equation 2.

$$d_t(\tau) = \frac{1}{\omega} \sum_{j=t+1}^{t+\omega} [x(j) - x(j + \tau)]^2 \quad (2)$$

The squared difference function is prone to subharmonic errors. As shown in Equation 3, the YIN algorithm normalizes the results by taking the cumulative mean.

$$d_t'(\tau) = \begin{cases} 1, & \text{if } \tau = 0, \\ d_t(\tau) / \left[ (1/\tau) \sum_{j=1}^{\tau} d_t(j) \right] & \text{otherwise} \end{cases} \quad (3)$$

Evaluation on speech data reports that 99% of estimates are accurate to within 20% of the correct  $F_0$ , 94% to within 5%, and approximately 60% were accurate to within 1% (de Cheveigné & Kawahara, 2002). In the same evaluation, YIN 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 (Boersma 1993). The algorithm was also evaluated on the singing voice by de Cheveigné & Henrich (2002).

Spectro-temporal approaches are motivated by the division of labour in the auditory system: The cochlea performs a spectral representation of the sound and then this representation is coded into temporal neural spikes for further processing. In such approaches, the spectral analysis is typically done by gamma-tone or constant-Q filter banks, depending on which aspect of the auditory system one wants to model, and this output is then subject to some type of temporal analysis, often autocorrelation, in each channel.

#### 2.4.2 Polyphonic Estimation

When approaching polyphonic  $F_0$  estimation, one can use a monophonic algorithm to estimate one voice, suppress it, and then move to the next voice, or one can estimate the voices jointly. One of the main challenges is automatically estimating the number of sources, as there may be some confusion between fundamentals and partials. Thus, in much of the early literature on polyphonic  $F_0$  estimation, the number of sources was specified *a priori*.

In terms of spectral methods, Parsons (1976) described a method for estimating a signal with two periodic components using the histogram technique developed by Schroder (1968) to identify one of the components before removing it from the signal and estimating the second. The process is then iteratively repeated to improve the estimate. Duifhuis, Willems, and Sluyter (1982) developed a polyphonic  $F_0$  estimator based on pattern matching and harmonic sieve model of pitch perception by Goldstein (1973) (described in Section 2.3.2.2). More recently, Yeh, Robel, and Rodet (2005) described a frame-wise method for polyphonic  $F_0$  estimation in the spectral domain that produces a hypothetical partial sequence constrained by three principles: spectral match, spectral smoothness, and spectral continuity.

The hypothetical partial sequence is subject to a score function that picks the best fundamental frequency candidates.

An example of a temporal approach to polyphonic  $F_0$  estimation is de Cheveigné and Baskind's (2003) proposed extension to the YIN algorithm called MMM, which uses a double difference function in order to estimate multiple  $F_0$ s. They also included a method of determining the number of voices, which involved fitting monophonic and polyphonic models and selecting the best fit. The self-similarity approaches can be made more efficient by ordering the lags.

A spectro-temporal approach was employed by Klapuri (2006) in a method that makes use of a gammatone filter bank. Following Meddis and Hewitt's hair cell model, the outputs of the filter bank were subject to compression, half-wave rectification, and low-pass filtering. Similarly, Marolt (2004) uses a gammatone filterbank, but he applies a bank of time-varying oscillators to model its output. Both of these techniques subsequently apply a learning-based model to estimate the individual notes: Klapuri uses a hidden Markov model (HMM) which will be described in greater detail below, while Marolt uses a neural network on the output of the time-varying oscillators to determine the most likely notes which are present in the signal.

Learning-based techniques encompass several approaches, including neural networks, generative approaches, and non-negative matrix factorization. Davy (2006) describes a number of generative approaches that interpret the problem of polyphonic fundamental frequency estimation in terms of priors, transitions, and probabilities. This allows for a great deal of flexibility in terms of handling an unknown number of sources. Non-negative matrix factorization assumes the signal to be the product of two matrices. The resulting representation can be constrained to be sparse, which is helpful for a task like polyphonic  $F_0$  estimation if it can be done in such a way that the harmonics are preserved and stochastic noise in the signal is removed. Cont, Dubnov, and Wessel (2007a) assured sparsity by introducing assumptions about the nature of musical timbre. Abdallah and Plumley (2004) applied independent component analysis to the power spectra of the signal and used a learned dictionary to create a sparse representation. Vincent, Bertin, and Badeau (2007) constrained their non-negative matrix factorization to be harmonic, which they found to improve their results.

### 2.4.3 Transcription of the Singing Voice

Much of the work on  $F_0$  estimation in the singing voice has been in the context of transcription. Weihs and Ligges (2003) divided the task of singing transcription into different steps: note segmentation,  $F_0$  estimation, note estimation, quantization, and transcription. A preliminary question sometimes addressed in the literature on singing transcription is whether there are vocals at all in some sections of the audio. Berenzweig and Ellis (2001) presented a technique based on hidden Markov models for locating vocal segments in polyphonic audio, which is particularly useful for pop songs where there are instrumental bridges or sections that should be ignored. Tsai and Wang (2004) also developed a technique for segmenting the audio in vocal and non-vocal sections before applying a method for modeling each singer's vocal characteristics and tracking them in a polyphonic vocal recording using parametric Gaussian mixture models.

The issue of note segmentation is challenging for the singing voice since, unlike an instrument like the piano, it lacks a clear percussive attack. Sundberg and Bauer-Huppmann (2007) studied the perceived onset time by analyzing the synchronization of pianists accompanying singers in aligned performances. They found that overall the accompanists entries were synchronized with the start of the steady-state portion of the tone rather than the transient. Various approaches have been experimented with for note onset estimation. Transient detection often does not work very well because of the lack of sharp onsets in the sung voice. Clarisse et al. (2002) use an energy threshold to determine onsets by measuring the root mean square energy as a function of time. Wang et al. (2003) use dynamic programming to determine the end point of the notes. Weihs and Leigges (2003) combine segmentation with pitch estimation and use pitch differentials to segment the notes. The method used in this dissertation, based on audio-score alignment, will be described in Section 3.1.

The timbre of the singing voice can also pose some challenges for  $F_0$  estimation, specifically the presence of formants. Formants are resonant frequencies produced by the vocal tract that define vowels and timbre. They are relatively consistent in speech, but can be modified in the singing voice when sopranos sing higher than the first formant (in which case they raise it to match the fundamental) (Sundberg 1999). In the case of the singer's formant, one or more formants are adjusted by bass, tenor, and alto singers in order to assist them in

projecting their voice over an orchestral accompaniment (Bartholomew 1934). The benefit of temporal versus spectral approaches is that they are more robust to the presence of formants. Thus singing transcription methods generally rely on autocorrelation-related techniques, for which de Cheveigné and Kawahara's YIN algorithm (2002) is particularly popular. An alternative approach that is used is to filter the signal in a manner similar to the filtering done by the cochlea (Clarisso et al. 2002). The timbre characteristics of the singing voice can make it more challenging to use learning-based techniques, since the same note sung by the same singer would have a slightly different timbre depending on the syllable being sung.

Note estimation is the assignment of a pitch class or MIDI note to the  $F_0$  estimates over the duration of the note and is complicated by two factors: the presence of vibrato and the fact that the human voice is flexible in its intonation capabilities. The steady-state portion of the note generally contains vibrato (typically up to +/- 71 cents, as reported by Prame (1997)), which means that estimation of a single fundamental frequency, or perceived pitch, requires the application of a heuristic or rule. As discussed in Section 2.3.3, the perceived pitch is often assumed to be the mean of the frequencies over the duration of the note. This mean value can then be assigned to a MIDI note using the "round MIDI" techniques described in the singing transcription literature (Clarisso et al. 2002; Viitaniemi et al. 2003; Wang et al. 2003). These approaches set up a specified range of frequencies that translate to each MIDI note: this is either done automatically, using a fixed tuning reference, or by assessing the relative tuning of the estimated fundamental frequencies.

The quantization of the notes, necessary to represent them in a metre, is often performed by hand, but there is a relatively workable automatic Bayesian framework developed by Cemgil and collaborators (2000). Transcription is the process of generating a MIDI file from the quantized note values. Nienhuys and Nieuwenhuizen's Lilypond software (2003) is currently the de facto tool for this final step in the task of singing transcription.

There are alternative approaches that use a variation of the five steps (note segmentation,  $F_0$  estimation, note estimation, quantization, and transcription) outlined by Weihs and Ligges (2003). These include Wang and collaborators' (2003) method for determining where the vowels are in the signal and building "islands" around them in the representation of signal where the analysis should be focused. They also use a particularly sophisticated stochastic

technique, named Adaptive Round Semitones, for adaptively assigning the estimated frequencies to MIDI notes. The work of Ryynanen and Kalpuri also deviates substantially from the Weihs and Ligges conception of the problem.

Ryynanen (2006) discusses some additional problems associated with transcription of the singing voices, where transcription results in a piano roll representation rather than a musical score. These include variations in tuning and the detection of note-offsets. In order to account for the fact that the voice has flexible intonation capabilities, the transcription system needs to have some rules for assigning fundamental frequency estimates to note names. This can be done by either rounding the frequencies to the nearest note on an equal-tempered scale, which is generally quite problematic, or by taking a more adaptive approach, which may or may not allow for tuning drift. The note-offset problem is quite challenging. In legato singing, the onsets can be located at the terminus point of the previous note, but in detached singing, this is not a viable option. However, in legato singing, onset detection itself is much more challenging.

Ryynanen's work with Klapuri (2004; 2008) on singing voice transcription uses an HMM-based note recognizer that works around the onset, and to a certain extent, offset challenges, by applying a metrical model to the transcription problem. For simple monophonic signals, the note offset can be determined by a decay in amplitude; however, there exists the degenerate situation in monophonic music where reverb may blur the offset and the general situation of polyphonic music. Once note events, such as pitch, voicing, phenomenal accents (as defined in Lerdahl and Jackendoff (1983)), and metrical accents are modeled with a hidden Markov model, note event transitions are modeled with a musicological model, which performs key estimation and determines the likelihood of two- and three-note sequences. Other recent papers addressing this problem include Dannenberg and collaborators (2007), as well as Unal, Chew, Georgiou, and Narayanan (2008).

There are applications of this research in a number of areas. For Query-by-Humming (Kosugi et al. 2000; Birmingham et al. 2001; Clarisse et al. 2002), current monophonic singing transcriptions systems are sufficient for analyzing the queries. However, the problem remains as to whether people can accurately reproduce the song they are thinking of. This research is also helpful for synchronizing lyrics to a sung melody, which is useful for karaoke

(Tsai and Wang 2004). This technology is also useful in systems for training vocalists (Mayor et al. 2006).

#### 2.4.4 Summary

Single-voice  $F_0$  estimation is effectively a solved problem. In particular, de Cheveigné and Kawahara's YIN algorithm has been shown to be robust for the singing voice. However, none of the existing polyphonic  $F_0$  estimation techniques are robust enough for use in this dissertation. The techniques described in Section 2.4.3 make use of quantization to create a MIDI-like transcription. These systems are evaluated in terms of the overall accuracy of their transcription rather than frame-wise accuracy of the individual components. Therefore, they are not useful for extracting the detailed pitch information required for the study of intonation. This research makes use of monophonic recordings, either in the study of a single singer or of an ensemble where each singer is miked individually. Note segmentation is done with the audio-score alignment method described in Section 3.1 and the YIN algorithm for  $F_0$  estimation.

### 2.5 Audio-Score Alignment

This section surveys existing work related to annotating note onsets and offsets in the singing voice. As described in Section 2.5.1, this type of signal is not well suited to blind onset detection algorithm. The subsequent sections describe different types of audio-score alignment algorithms (2.5.2–3) and their utility in obtaining information about onsets and offsets (2.5.4–5).

#### 2.5.1 Annotating Note Locations

Note onsets and offsets are 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 done in the area of note onset detection (Bello et al. 2005), accurate detection of onsets for the singing voice and other instruments without percussive onsets is not a solved problem. Available onset detection algorithms are discussed in Goebl, Dixon, Knees, Pampalk, and Pohle (2008), specifically Dixon's Beatroot system (Dixon 2001; Gouyon and Dixon 2005). However, these algorithms often require a significant amount of manual correction. 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. This degree of accuracy makes this the state of the art, but it still is insufficient for our purposes.

For music where a score is available, audio-score alignment techniques can be used to guide signal-processing algorithms. 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. This research question has been an active area of inquiry for over twenty-five years, and although most of the fundamental issues have been addressed, there remain some open questions for certain applications. In early work, the research question was defined in terms of an online problem for following a soloist while generating a time-sensitive musical accompaniment, which is known as score following. Later work explored the applications of offline implementations, including expressive performance analysis, audio database search, and synchronization for digital libraries. Different applications require different degrees of accuracy, ranging from the note or measure level for digital libraries to the order of milliseconds for expressive performance studies. The first part of this section reviews the early history of score following before describing the various techniques that are used for music alignment and their applications. The second discusses the challenges in evaluating music alignments and the open research questions that remain, particularly for expressive performance applications.

### **2.5.2 Early Score Following**

The published history of score matching began at the 1984 International Computer Music Conference, where Dannenberg (1984) and Vercoe (1984) presented separate papers on the topic of automating computer accompaniment for a live musician. Both Dannenberg and Vercoe broke the overarching problem of automatic accompaniment into a set of sub problems. Dannenberg identified three problems: Continuous and accurate detection of

what the soloist plays, matching this against a score, and using this information to generate an accompaniment. Vercoe took a slightly different approach than Dannenberg and explicitly included learning in his framework in addition to listening and performing. This survey considers only the score-following aspects of his work, not the generation of an accompaniment. Typically, in these early works, either pitch tracking is used to generate MIDI data or MIDI data is transmitted directly from the instrument. The performance data is then aligned to the stored MIDI data, generally using string-matching techniques.

### **2.5.2.1 Dannenberg**

Dannenberg used dynamic programming to find the best match between the incoming pitch tracking information and the note information available in the score. He made the assumption that the order of events in the performance was fixed, which made the system usable for only monophonic performances. Bloch and Dannenberg (1985) extended the matching system to polyphonic keyboard performances. They discussed two approaches: a “static” version, which collapses polyphonic events into a single moment in the score, and a “dynamic grouping” version, which allows for matching between an incoming sequence and notated simultaneities (chords). Dannenberg and Mukaino (1988) further increased the robustness of the system by extending the framework to allow for additional matching algorithms that could accommodate alternative interpretations of the performance. They also demonstrated the utility of grouping notes together to account for ornamentation. The issue of tracking an ensemble performance, rather than a soloist, was addressed by Grubb and Dannenberg (1994). Their algorithm dealt only with MIDI data, taking a weighted average of the pitch and timing information available in the MIDI messages to calculate an estimate of the ensemble’s score position.

### **2.5.2.2 Vercoe and IRCAM**

Vercoe’s work was specifically motivated by a commission for a piece for flute and live electronics by IRCAM, but also more generally by the idea of developing a “synthetic performer” that is able to respond to live performers. Vercoe used optical sensors on the flautist’s fingers to guide the pitch-tracker algorithm. Strategies for improving the score following through learning were further explored by Vercoe and Puckette (1985). They described a technique for reanalyzing the rehearsals offline to improve the accuracy of the

score following during performances. The learning approach was not incorporated into the actual IRCAM system because of the risks involved in relying on these heuristics in a live system. These risks were described by Puckette and Lippe (1992), specifically the way in which they make the system less predictable and more likely to fail catastrophically. The system could fail even with the incorporation of heuristics, which meant that someone needed to supervise the score following, lest it go off track. Puckette (1990) describes a graphical sequence editor for MAX called EXPLODE that allows for easy specification of when the system should expect input from a performer.

#### **2.5.2.3 Second-generation Score Followers**

Baird, Blevins, and Zahler (1990; 1993) presented what they termed a “second generation” score follower. Their work built on Dannenberg’s and Vercoe’s by using segments of the musical score to improve the musicality of score following. The segmentation was achieved by doing a phrase-based analysis of the music and used performance heuristics to improve the system’s expectations of dynamics, note durations, and rests in relation to these segments.

Stammen and Pennycook (1993) pioneered the use of a particular form of dynamic programming called dynamic time warping (DTW) in the context of score following. They extracted pitch and rhythmic contours from incoming MIDI data and matched this to the stored MIDI score via DTW. Vantomme (1995) explored the use of MIDI timing information for score following as an alternative to the pitch-based techniques. The paper discusses techniques for account for timing deviation in performance, particularly asynchrony in the performance of simultaneities.

#### **2.5.2.4 Tracking a Vocal Performer**

Following a vocal performer raises some additional challenges, as separate pitches are not so clearly delineated as in instrumental music. For example, the instantaneous pitch of a vocal onset does not necessarily relate to the pitch of the note and there is more variability depth and rate of vibrato of a sung note, as well as the amplitude envelope. Katayose, Kanamori, Kamei, Nagashima, Sato, Inokuchi, and Simura (1993) developed a method tracking a vocal performer that is tuning to attend to where plosives will occur in the score as a type of acoustic marker. Puckette (1995) presented a system that not only attempts to assign a pitch

as soon as possible to minimize latency, but also allows for *a posteriori* corrections if the system finds that certain notes were not accounted for as they were performed.

Grubb and Dannenberg (1997) introduced a stochastic approach to estimating a score position pointer in a vocal performance by observing only pitch data. They refined the approach the following year by using not only pitch information, but also spectral envelope, and note onset estimates, which decreased the latency in the score follower by ~10% (Grubb and Dannenberg 1998). The score position pointer was estimated using probability density functions that were calculated from hand-labelled data.

### 2.5.3 Techniques

This section describes three different approaches to score matching: dynamic programming, particularly dynamic time warping (DTW), hidden Markov models (HMMs), and support vector machines (SVM). DTW can be considered a constrained form of an HMM, where the state sequence always moves forward and each transition has the same probability for all states. The original DTW formulation does not allow for a meaningful training procedure. Due to its constrained nature, DTW is better suited to offline applications where there is a known correspondence between the performance and the score since it is not as flexible as HMMs in dealing with performance errors in online contexts. HMMs are a type of generative learning algorithm, where the learning can be either supervised or unsupervised. In contrast, support vector machines (SVMs) are a type of discriminative supervised learning algorithm.

#### 2.5.3.1 Dynamic Programming and Time Warping

DTW, a type of dynamic programming, allows for the alignment of similar linear patterns, or sequences, evolving at different rates. Through DTW, the two sequences are aligned by warping them to minimize a cost function that penalizes both local and sequential mismatch. The time warp can be represented visually as a path through a similarity matrix (Rabiner and Juang 1993), as shown in Figure 2.5.1. In the similarity matrix, black indicates maximum similarity and white indicates maximum dissimilarity; shades of grey indicate intermediate steps. The best path through the similarity matrix is a warping from note events in the MIDI to their occurrences in the audio. The black line in Figure 2.5.1 represents the best path, which was calculated using a cost function that considers all possible paths through the

similarity matrix (from the bottom left corner to the top right corner) and which penalizes for both distance and dissimilarity.

Early work with dynamic programming focused on matching polyphonic MIDI performances to the stored MIDI data. Later work applied DTW to audio-score matching, where both the MIDI and audio files are reduced to a set of features. As these features are generally spectral, the MIDI file was first converted to audio, or some type of spectral-like representation, for feature extraction. The question of which features are the most appropriate for this task has been the topic of some debate in the literature. Spectral decomposition methods have also been explored as an alternative to feature-based matching.

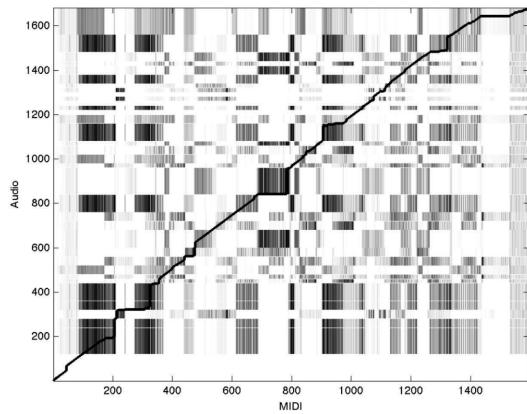


Figure 2.5.1: A dynamic time warping similarity matrix. The black line indicates the optimal path through the similarity matrix, which is used to warp the timing in the audio and MIDI to match each other. The y-axis is the number of audio frames and the x-axis is the number of MIDI frames. Black indicates high similarity and white indicates low similarity.

#### 2.5.3.1.1 MIDI Data

Large (1993) utilized an offline dynamic programming algorithm for coding pitch errors in MIDI piano performances. He scored the performances based on the number of matches, substitutions, additions, and deletions. Large reports that his algorithm was computationally efficient and achieved a greater than 90% accuracy in error detection. Also, in the early 1990s, Honing and Desain began working on an offline matching algorithm for polyphonic MIDI data (Honing 1990; Desain and Honing 1992). This algorithm, along with some later variants, was described in their later articles with Heijink (Desain et al. 1997; Heijink et al.

2000a; Heijink et al. 2000b). These articles identify three main challenges in polyphonic matching: the ordering of events notated as simultaneities in the score, performance errors (including missed notes, extra notes, and wrong notes), and ornaments. They term their earliest algorithm as an “incremental matcher,” where they match notes in the simultaneities in which they occur. They also describe a “non-incremental matcher,” which considers all of the possible orderings of notes in a simultaneity before selecting the most appropriate one, as well as a “structure-based matcher,” which uses both pitch and onset information. The “structure-based matcher” has the flexibility to act like either the “incremental” or “non-incremental” matcher, depending on how many notes it considers at a time. The various types of matchers were evaluated in Heijink, Windsor, and Desain (2000b) on the same dataset of MIDI piano performances used in Hoshishiba, Horiguchi, and Fujinaga (1996). The error rate for the “incremental matcher” was 1.2%, the “non-incremental matcher” is 1.1%, and the “structure-based matcher” was 0.1%.

#### 2.5.3.1.2 Acoustic Features

Orio and Schwarz (2001) experimented with spectral peak structural distance as a feature for DTW-based alignment. They found that the use of the delta peak structural distance, as well as a model for attacks and silence, improved their initial results. Their algorithm was tested on 708 sequences of sample-based synthesized music with different timbres and articulations. The test material included monophony and two- and three-voice polyphony. Two evaluation criteria were used: error rate, which was defined as the number of estimates that were more than 200ms off the ground truth, and average offset, which referred to the average amount that the estimates were off the ground truth in ms. The average error for monophonic performances was 0.42% and polyphonic was 3.6%. The average offset for all of the performances was 31 ms without the attack/sustain model and 18 ms with.

In the same year, Pardo and Birmingham (2001) described a score-following system to match a MIDI performance to a lead sheet using dynamic programming. The system automatically segments the performance based on a metrical reduction analysis and groups the segments into chords that are then aligned to the lead sheet using dynamic programming. The following year, Pardo and Birmingham (2002) described some improvements for a monophonic score follower based on a probabilistic modeling transcription error and timing

information. They evaluated their dynamic programming-based alignment algorithm on solo alto saxophone performance.

Soulez, Rodet, and Schwarz (2003) improved the robustness of the algorithm in Orio and Schwarz (2001) to include a sustain model. They tested their algorithm on real world, rather than synthesized, audio recordings that varied in their musical content from the corresponding MIDI file. The lack of direct correspondence made it difficult to evaluate, so only global alignment was considered. They defined a correct alignment as one where the estimated onset in the performance is closer to its correct match than any other onset in the score. Using this evaluation metric, they had an error rate of 9.7%.

In the same year, Danneberg and Hu (2003) evaluated the usability of chromagrams for alignment. They evaluated their algorithm both in terms of visualizations of the DTW path in a similarity matrix and against hand annotated points in complex polyphonic audio. The visualization showed that the alignment methods were sensitive to systematic adjustments in the MIDI file. The evaluation of the algorithm was made against five manually annotated points in three different pieces, two pieces by Beethoven and one by the Beatles. Overall, the average error ranged from 34–76 ms.

Turetsky and Ellis (2003) presented a DTW-based score-matching algorithm. They explored the use of combinations of the cosine difference of spectral power, first-order difference between channels, and first-order difference in frequency as features. These features were extracted from both the audio and a sonified version of the MIDI. They explored the performance of “greedy” alignment, which finds a smoother path but has the potential to fail, as well as “unconstrained” alignment, which might not find the most optimal path but will always find a path. Turetsky and Ellis used a two-step alignment approach, where an “unconstrained” alignment is refined by a “greedy” one. The algorithm was evaluated aurally, as well as empirically, by calculating the similarity of paired values in the alignment to determine how good the alignment was overall. They did not provide any evaluations on the accuracy of the alignment at the note level.

Izmirli, Seward, and Zahler (2003) expanded the work done in Baird, Blevins, and Zahler (1990; 1993) with the development of a method for automatically analyzing the score for melodic “anchor” points that could be used as an alternative to arbitrarily segmenting the score by length. These anchor points are then used to guide the alignment algorithm. The

system was successfully implemented in a performance evaluation, and visual evaluation was used to compare their system's performance with and without the “anchor” points.

Dixon (2005) presented an online version of DTW using spectral features for audio-to-audio alignment in the context of score following. He tested the system on piano music recorded from a Bosendorfer SE piano, which provided precise information about the timing of each note played. The average note-level alignment error was 59 ms.

#### 2.5.3.1.3 Spectral Decomposition

Arifi, Clausen, Kurth, and Müller (2004) looked at several types of matching: audio to MIDI, music score data to MIDI, music score data to audio, and audio to audio. Since they were dealing with score data, they opted to extract features from the audio that more closely resemble the score data. To this end, they used a combination of sub-band analysis and onset detection to extract note-like features; the technique was later extended by Müller, Kurth, Roder, and Clausen (2004; 2005). The approach in Müller, Kurth, and Roeder (2004), as in Arifi, Clausen, Kurth, and Müller (2003), focuses exclusively on piano music and adds peak picking of local energy maxima to the sub-band decomposition and onset detection techniques used for sparse representation in Arifi, Clausen, Kurth, and Müller (2003). Müller, Kurth, and Clausen (2005) introduced the Chroma Energy distribution Normalized Statistics (CENS), a non-instrument dependant feature based on chroma and short-time statistics. Although the chroma features use the same filtering method as was used in Müller, Kurth, and Roder (2004) for sub-band decomposition, it differs in that the chroma signal is decomposed in 88 bands corresponding to musical notes. The matching, or synchronizing, algorithm in Arifi, Clausen, Kurth, and Müller (2003); Müller, Kurth, and Roder (2004); and Müller, Kurth, and Clausen (2005) all use a less constrained version of dynamic programming than classic DTW in order to allow for matching when there is information in the audio signal that is not present in the score (e.g., ornamentation). In Müller, Kurth, and Roder (2004), the algorithm was evaluated on Romantic piano music and performed well in aural evaluation. In Müller, Kurth, and Clausen (2005), it was evaluated by querying a collection of Romantic music of various genera and found to perform well for clips of 20 seconds or longer.

In 2006, Müller, Mattes, and Kurth (2006) presented a multi-scale DTW-based algorithm which refines an original alignment using the CENS features described in Müller, Kurth, and

Clausen (2005). The multi-scale approach increases the efficiency of the DTW algorithm by first doing a rough initial alignment before doing a more refined alignment based on the initial one. They assess this approach to have an accuracy of less than 100 ms. In 2008, Müller and Ewert (2008) presented a DTW-based algorithm for assessing structural similarities between two pieces of audio by doing a joint structural analysis on the pieces. In addition to revealing structural differences, the joint analysis also allowed for better alignment with pieces that have significant structural variation from the reference recording, or MIDI file.

Niedermayer (2009) presented a matching algorithm in 2009 using non-negative matrix factorization (NMF), which builds on the approach described by Cont (2006). A dictionary of tone models was learned and then used to decompose the signal into constituent pitches. The alignment was performed with DTW. The algorithm was evaluated on 13 Mozart piano sonatas and was found to have comparable results to chroma vectors. Niedermayer also explored the potential of converting the sparse NMF representation to MIDI and doing alignment in the symbolic, rather than in the audio, domain. The benefit of matching in the symbolic domain is that it would be far less computationally intensive; however, the results for matching with the NMF-derived symbolic representation were significantly worse than for NMF in the audio domain. Only 18.9% of the symbolic domain alignments were within 50 ms of the ground truth versus 68.6% for the audio domain alignments.

### 2.5.3.2 Hidden Markov Models and Related Techniques

A hidden Markov model (HMM) is a statistical model of the temporal evolution of a process. The model is based on the assumption that the future can be predicted from current state, since it summarizes the past sequence of events. In order to model the temporal dynamics of a system, each state has a certain probability of transitioning to every other state; the true state path is hidden in the HMM. The observations of information from the HMM are stochastically related to the state, but the state itself is never observed directly (Rabiner 1989). In the case of music alignment, only the acoustic features of the signal can be observed, and it is not known whether a given frame is from either an attack state or a sustain state. In their earliest applications for music alignment, single-level HMMs were used, but they proved to be unreliable at times. More recent works have explored the use of multi-level and graphical models.

#### 2.5.3.2.1 Single-level HMMs

In 1999, several papers described hidden Markov model-based (HMM) approaches to score matching. Cano, Loscos, and Bonada (1999) focused on monophonic music and created left-to-right HMMs to model notes as attack, sustain, and release states, as well as silence as a single state and no-notes states for all non-notes/non-silence. For its observations, the HMM used energy, zero-crossing, and fundamental frequency and their derivatives. The HMM was trained with hand-labelled audio and used a Viterbi decoder to find the optimal path. There was no formal evaluation of the system in the paper.

Loscos, Cano, and Bonda (1999) described an HMM-based algorithm that incorporated phoneme recognition into score matching of vocal music. The motivation for using phonemes was to simplify the alignment problem and reduce the amount of delay in a real-time context. The authors describe the acoustical difference between singing and speaking voice in terms of voiced/unvoiced ratio, dynamics, fundamental frequency, vibrato, and formants, and develop an HMM architecture that used predominantly mel cepstrum and energy values for observations. As in Cano, Loscos, and Bonada (1999), the algorithm used three left-to-right HMMs. The only difference in this implementation was that the non-note (plosive) HMM was modeled with two states. The system was tested on a number of songs and was evaluated visually with the alignment displayed over a time-domain representation of the audio.

Raphael (1999) described an HMM approach for monophonic signals. Each frame of audio is described as a low-dimensional feature vector and unsupervised learning is used to train the system. Raphael contrasts this work with the pitch-based approaches of earlier work, including Dannenberg (1984), Vercoe (1984), Puckette and Lippe (1992), Puckette (1995), and Grubb and Dannenberg (1997). The data model consists of two note states (articulation and pitch) and one rest state (silence). The system uses the Forward-Backward algorithm to find the optimal segmentation of the audio. Evaluation was done visually, by overlaying the segmentation on the spectrogram, and aurally.

#### 2.5.3.2.2 Multi-level HMMs

Orio and Dechelle (2001) presented a new approach for training a multi-level HMM for polyphonic music that consisted of both a note-level and a score-level. As in Orio and

Schwarz (2001), a bandpass filter was used to assess spectral energy. Here, the filtering was done on a database of sounds to obtain note-level observations. At the score level, a variant of the Baum-Welch algorithm was used for training, where the user had to specify the last correct state. The decoding was done using Viterbi. The system was evaluated sonically on several pieces of contemporary music, as well as repertoire, with various ornamentations and articulations.

This approach was extended by Schwarz, Orio, and Schnell (2004), who implemented an HMM for use with polyphonic MIDI piano data, which included the use of the sustain pedal. The incoming MIDI signal was quantized to account for asynchrony in notated simultaneities. Informal evaluation was performed on a contemporary chamber opera with robust results. The following year, Cont, Schwarz, and Schnell (2005) described a Gaussian Mixture Model approach to model the observations in the IRCAM score follower that was based on Orio and Déchelle (2001). In order to train the score-following system, they learned the mapping between the score and the performance through a discriminative approach.

Cont (2006) presented a method for score following based on Non-negative Matrix Factorization (NMF) and hierarchical HMMs. The NMF approach allows for multiple pitch estimation through the unsupervised learning of dictionary specific to the acoustics of a particular instrument. The hierarchical HMM has two levels: the lower level models the notes, chords, and rests themselves, and the upper level models the temporal relationship between the lower-level events and the score. The method was evaluated visually on a piece of contemporary piano music.

Montecchio and Orio (2009) presented an HMM approach that uses the output of a discrete filter bank, rather than a Fast Fourier Transform (FFT), for observations. As in Orio and Déchelle (2001), the HMM is multi-levelled with the acoustics of each musical event modeled at the event level, and these events themselves modeled at the score level. As in earlier works, Viterbi decoding was used. The algorithm was tested on single instrument, chamber, and orchestral music. The results were evaluated both aurally and visually. A comparison with an FFT-based approach was also made using a manually annotated 105-note excerpt. The filter bank approach performed more robustly in terms of event recognition (101 vs. 93).

Cont (2010) described an approach for “anticipatory” score following, where the system makes predictions about the future to inform its current decision. These inferences are made with a hidden hybrid Markov/semi-Markov model that is constructed from information in the score. The observations are based on the FFT of each frame of audio. The system was evaluated using both synthesized audio and the MIREX 2006 dataset (Downie 2006), which is discussed below in Section 2.5.5.3. On the synthesized audio data set the mean onset error ranged from 8.69 to 9.5 ms, and the mean tempo error ranged from 8.13 to 158.78 ms under various tempo manipulations. On the MIREX dataset, the system performed better than any of the systems submitted to either the 2006 or 2008 evaluation with a total precision of 91.49% given a tolerance of 250 ms. The mean offset error on the MIREX dataset ranged from 75.1 to 240.9 ms.

#### 2.5.3.2.3 Graphical Models

Raphael (2004) presented a graphical model method for score matching based on pitch content of the audio. Raphael turned to graphical models in order to address the shortcomings of HMM-based systems when following complex polyphonic audio. He notes that the modeling of note length is often problematic, and in his system, tempo-shifts are modeled in order to use the duration information in the score more effectively. This is achieved with a two-level model: one level models the pitch content in the signal, and the other models the notes and tempo-shifts. In order to make the model tractable and to handle the continuous nature of the tempo-shift variable, a *maximum a priori* estimate is computed to find the most likely collection of paths. The system was evaluated on a 55-minute set of orchestral excerpts that were labelled by tapping and hand correction. The result showed that 95% of the estimate onset times were within 250 ms of the ground truth, and 72% were within 125 ms. Aural evaluation was also used, and the results were made available, as well as the evaluation set. The work was later described in greater detail in a later article (Raphael 2006).

Peeling, Cemgil, and Godsill (2007) presented a new approach for alignment called score position pointer estimation. They described two feature-extraction methods: one using two-element spectrogram-derived vectors for each frame of audio to characterize the signal’s energy and one using a sinusoidal subspace model to extract frequency, amplitude, phase, and damping coefficient. Observation models were learned directly from the data, and the

progression of the score position pointer was calculated with Viterbi. On a test set of sample-based piano performances, the algorithm performed with an average onset resolution of 7.5 ms. The authors also evaluated the algorithm both aurally and visually.

#### 2.5.3.2.4 Support Vector Machines (SVMs)

SVMs find the maximum margin between two classes of data. They are used when the margin cannot be discriminated in a low-dimensional feature space (Christianini and Shawe-Taylor 2000). Shalev-Shwartz, Keshet, and Singer (2004) presented an SVM-approach for offline alignment polyphonic music. They used ten different features, nine of which are FFT-based and one of which is the similarity (or relative tempo) between the performance and the scored MIDI data. The weights for the features are learned from a training set. Once the weights have been applied, dynamic programming is used to determine changes in relative tempo between the recording and the MIDI file over the entire piece. The authors evaluated their approach against a generative HMM model on 12 piano performances for which MIDI files of the performance were available to use as ground truth. The average onset detection error was less than 20 ms compared to 25–78 ms for the variants of generative HMMs that they also evaluated. In a later article with Chazan (Keshet et al. 2007), they expanded this technique to speech-to-phoneme alignment.

### 2.5.4 Applications

There are numerous applications for music alignment techniques. This section surveys different approaches to a range of applications. The most pertinent to this dissertation are the applications for expressive performance studies. Other applications include automatic accompaniment, query-by-humming, and digital libraries.

#### 2.5.4.1 Expressive Performance Studies

As noted above, music alignment can be used for expressive performance studies since it often performs with more precision in identifying note onsets and offsets than blind estimation algorithms. In his master's thesis, Scheirer (1995) described the architecture of a system for transcribing polyphonic piano music using a MIDI score as a guide. He discussed the potential applications for the study of expressive performance both in his thesis and in a book chapter on the topic (Scheirer 1998). His work represents the first attempt to study performance from audio data; however, the study of MIDI recordings remained much more

common until 2003. As discussed above, Large (1993) used dynamic programming to assess errors in piano performance. Hoshishiba, Horiguchi, and Fujinaga (1996) discussed a number of methods for using dynamic programming to find the best alignment between different MIDI performances that contained several errors, as well as a technique for defining a normative performance. Heijink and colleagues (2000a; 2000b) later followed up on the work of Hoshishiba, Horiguchi, and Fujinaga.

Dixon (2003) picked up on the work done by Schierer and described another architecture for a system capable of aligning an audio recording of a piano to a MIDI score. A later paper by Dixon and Widmer (2005) presented a working toolkit for Dixon's algorithm but with a focus on aligning different audio performances and the potential of relating it to MIDI with sonified version of the MIDI (building on Turetsky and Ellis 2003). This work was later summarized in a book chapter by Goebl et al. (2008), which also discussed some techniques for computationally modeling the collected data.

#### **2.5.4.2 Automatic Accompaniment**

Automatic accompaniment motivated the earliest research in score following (Dannenberg 1984; Vercoe 1984) and remains a current topic of interest since improving the response and naturalness of the accompaniment systems is still an open research question. Raphael (2001) described his Music Plus One system, based on the score-following technique described in Raphael (1999). The motivation for Music Plus One was to improve the static performances that were used in existing accompaniment systems, such as Music Minus One, where the accompaniment could not adapt to the expressive characteristics of the soloist's performance. The IRCAM score follower is also implemented in the context of a real-time accompaniment system. Schwarz, Cont, and Schnell (2005) described the implementation of the IRCAM score follower in automatic accompaniment systems for electro-acoustic and vocal music.

Current implementations of both the Music Plus One and the IRCAM score follower work with acoustic input, most optimally from a monophonic source. More recently, Jordanous and Smaill (2009) presented an HMM-based polyphonic MIDI score-following system and undertook surveys with performers to see how they felt about the automatic accompaniment system. They found that the performers felt that the accompaniment system performed well with simple pieces, but that latency became an issue for more complex pieces.

#### 2.5.4.3 Query-by-Humming

Query-by-Humming (QBH) has been an active area of research since the mid-nineties. The requirements for aligning a hummed or sung query to complex audio are a different, but related, problem for score alignment. Recognizing that people may not be completely accurate in their renditions of the melody they are looking for, early researchers used contour rather than absolute pitch to represent the query and matched it against the items in a database. Specifically, they represented the contour in terms of upwards, downwards, and repeating motion. Ghias, Logan, Chamberlin, and Smith (1995) used a fuzzy approach to pattern matching to account for any errors that may occur in the input query or the representation of the items in the database. The following year, McNab, Smith, Witten, Henderson, and Cunningham (1996) described a string-matching approach that utilizes dynamic programming.

Birmingham et al. (2001) presented a framework called MUSART that improved melody-based database queries by generating a thematic index. Searching was done using dynamic programming. The following year, Shlev-Schwartz et al. (2002) presented a probabilistic approach that used spectral and temporal features from the audio. Hu, Dannenberg, and Tzanetakis (2003) applied the method described in Dannenberg and Hu (2003) to database retrieval. They tested chromagrams against mel-frequency cepstrum coefficients, or MFCCs (Logan 2000), and pitch histograms for the task of retrieving pieces from a database of MIDI using audio queries. They found that chromagrams had the best accuracy: 0.95 on their dataset of 10 acoustic Beatles recordings, with pitch estimates coming in slightly below at 0.82 and MFCCs performing poorly at 0.30. The poor performance of the MFCCs is to be expected given that MFCCs highlight timbral information and discard most of the useful pitch information. When comparing MIDI and audio representations, it is likely that there will be significant timbral differences even when the same notes are being played, thus undermining the ability of the algorithm to match the pitch sequences. Adams, Bartsch, Shirfrin, and Wakefield (2004) examined the relative performance of three types of time series representations for queries: sequences of notes, a “smoothed” pitch contour, and sequences of pitch histograms. They used a dynamic programming-based search to evaluate each of these representations. They found that the pitch histograms work almost as well as contour, and both were better than pitch sequences. Pitch histograms had the advantage of

less computational complexity than “smoothed” pitch contours. In the following year, Adams, Marquez, and Wakefield (2005) described a system using pitch histograms with iterative deepening search and dynamic time warping.

Pardo and Sanghi (2005) addressed the issue of queries switching between different polyphonic lines in the targeted recording with an approach that made use of a probabilistic extension to string alignment. Suyoto, Uitdenboger, and Scholer (2007) used a noisy polyphonic transcription with the longest common subspace alignment. The following year, they experimented with relative pitch in the same framework (Suyoto et al. 2008). Overall, they found that they were able to make their DTW-based system more robust for truncated queries by exploring all possible transpositions in the alignment stage.

#### **2.5.4.4 Digital Music Libraries**

Digital music libraries contain both score-based and acoustic-based representations of music. These materials can be synchronized with one another to create a multi-modal browsing experience and music alignment can be used for this synchronization. A paper by Orio (2002) presented an HMM-based approach for synchronizing symbolic and acoustic versions of pieces in a digital library context. Orio discussed the limitations of the MIDI format for representing the information available in the score and, in light of this, designed a framework that can work with other digital representations of the music score. The use of different representations is discussed in more detail in Melucci and Orio (1999). Orio also described how the HMM representations that are used for alignment purposes can also be used for automatic recognition of performances within a library catalogue. This research was extended by Miotto and Orio (2007) to work for orchestra music and by Orio (2010) to work for non-Western music and underspecified scores.

Dunn, Byrd, Notess, Riley, and Scherle (2006) explored the potential for incorporating Raphael’s music alignment method (Raphael 2004) in the Variations3 digital music library project at Indiana University. In the Variations2 phase of the project, the scores and recordings were manually synchronized by someone tapping along while listening to each recording. The synchronization allowed for both fast indexing of the recordings and a multi-modal experience where the score position is shown in real time while the recording plays. This type of large-scale synchronization has been implemented at the Bavarian State Library (Fremerey et al. 2008) by a group of researchers at the University of Bonn, who have made

the technology widely available as a package called SyncPlayer. Kurth, Müller, Damm, Fremerey, Ribbrock, and Clausen (2005) first introduced SyncPlayer, a synchronization and visualization framework for different types of musical documents. The framework uses the alignment technology described in Müller, Kurth, and Roder (2004). Two years later, Kurth, Müller, Fremerey, Chang, and Clausen (2007) reported on the implementation of score alignment data collected through optical music recognition in the SyncPlayer framework.

#### **2.5.4.5 Other Applications**

##### **2.5.4.5.1 Karaoke**

A karaoke-related application of alignment was described by Cano, Loscos, Bonada, de Boer, and Serra (2000). They described a technique for morphing the characteristics of an inputted singing voice to the characteristics of a target singing voice. Alignment was done using the technique described in Cano, Loscos, and Bonada (1999).

##### **2.5.4.5.2 Education**

In their 2006 survey article, Dannenberg and Raphael (2006) described two educational applications of music alignment. The first is a program called Piano Tutor, described more extensively in Dannenberg, Sanchez, Joseph, Joseph, Saul, and Capell (1993), where students' performances are matched in real time to the score that they are playing. The system then makes note of any errors that are made in the performance. The second program is a commercial product called SmartMusic, which provides an accompaniment for students to play along with. AudioZoom, another application, was described by Montecchio and Orio (2008). This tool allows for visual and auditory highlighting of an instrument in a polyphonic recording.

##### **2.5.4.5.3 Signal Processing**

Another application of music alignment is to create a reference for signal processing algorithms. Turetsky and Ellis (2003) described how music alignment could be used to generate ground truth for polyphonic transcription. Woodruff, Pardo, and Dannenberg (2006) used score and spatial information to separate instruments in a polyphonic mixture of instruments. They tested their approach on a performance of sample-based string quartet performances and found that the score-guided separation worked better than blind separation based only on spatial information. The following year, Dannenberg (2007)

described a multi-track music editor that uses music alignment to isolate individual tracks. The alignment estimates are improved with a root mean square (RMS)-based onset detection algorithm. Dynamics for each part were estimated with RMS and fundamental frequency ( $F_0$ ) was estimated using de Cheveigné and Kawahara's YIN algorithm (de Cheveigné and Henrich 2002). Dannenberg also described a method for making adjustments to pitch, timing, and balance within the polyphonic mixture. Recently, Smit and Ellis (2009) described an algorithm for making frame-wise  $F_0$  estimates in four-part vocal music using an aligned MIDI file as a guide. They demonstrated that their probabilistic approach outperformed the YIN algorithm when it was guided by the aligned score.

### 2.5.5 Evaluation

Various evaluation approaches have been employed for assessing music alignment algorithms. When the alignment algorithm is part of a larger system, the quality of the alignment can be evaluated indirectly by measuring the end-to-end performance of the system. Direct evaluations of music alignment accuracy can be done aurally by generating a click track from the aligned onset estimates and then playing it back with the original audio, and/or visually by overlaying a representation of the audio signal in the time or frequency domain with the alignment onset estimates. For more precise evaluation of the timing discrepancies between the alignment algorithm's onsets and offsets estimates and those in the reference recordings, accurate annotations of the reference recordings are needed to act as ground truth.

#### 2.5.5.1 Ground Truth

Obtaining a sufficiently accurate ground truth for the systematic evaluation of audio alignment algorithms is an open problem. MIDI alignment systems, in contrast, are able to use the information in the reference MIDI file that the incoming MIDI file is being aligned to. However, MIDI alignment systems are greatly limited in their utility since they are only applicable to instruments that are capable of outputting MIDI data and are only useable for recordings in MIDI format, which are far less numerous than audio recordings. For audio alignment algorithms, a number of different approaches have been taken, including the use of synthesized audio (Orio and Schwarz 2001), audio recordings from instruments capable of also outputting symbolic data (Raphael 2004), and hand annotation of audio files (Dixon 2005).

The use of synthesized, or sample based, audio solves the problem of having accurate timing information for the notes in the reference recording. The audio signals that such a method creates are not as complex as real-world audio signal. Specifically, the limitation of timbral variability in the instrumental performance and reverberation greatly simplifies the alignment problem compared to its real-world counterpart. Recordings done on instruments capable of outputting symbolic data are generally limited to piano performances on a Bosendorfer SE, a Yamaha Disklavier, and other similar acoustic pianos with MIDI capabilities. Since these acoustic pianos can be recorded in ecologically valid conditions, these recordings provide useful ground truth. However, in the polyphonic context, this approach is only viable for piano music. It is possible to annotate other types of polyphonic recordings manually, but it is a laborious process that can be aided by an initial rough alignment and then making corrections by hand. The problem with this approach is that it is often hard to accurately identify individual onsets when different instruments are playing together. The use of multi-track recordings addresses this problem of asynchrony between the musical lines. However, such recordings are far less numerous than mixed-down recordings, which limits the amount of ground truth that is available in this format.

#### **2.5.5.2 Formal Evaluation**

The first wide-scale formal evaluation of score-following took place as part of the 2006 Music Information Retrieval Evaluation eXchange (MIREX) (Downie 2006). Several metrics were used, including precision rates that were calculated by subtracting the number of “missed notes” from the number of notes in the ground truth both at the level of the audio file and over all of the files. “Missed notes” were defined as those that exist in the ground truth but which are either not reported by the system or which are reported with a discrepancy of more than 2000 ms between the ground truth and the onset estimate. The second definition of “missed notes” is also contained in the “false positive” measurement. The average and mean discrepancies, as well as the discrepancies’ standard deviation and the systems’ average latency, were also calculated.

The systems submitted to the 2006 MIREX task were evaluated on 47.38 minutes of audio, which contained 8957 notes. The audio included a Mozart vocal aria, a Mozart clarinet concerto, a Boulez piece for flute and chamber ensemble, and a Bach violin sonata. The ground truth was obtained by hand correcting an initial offline score alignment. There were

two systems submitted: one by Cont and Schwarz, who obtained a piecewise precision of 90.06%, and one by Puckette, who obtained a piecewise precision of 69.74%. The piecewise average discrepancy error in the Cont and Schwarz system ranged from 91.8–409.7 ms with a standard deviation of 148–905.4 ms. The following year, Cont et al. (2007b) explained the MIREX evaluation system in greater detail and the task was run again in 2008 (Downie 2008) with a dataset submitted by the participants. For this task, Montecchio and Orio obtained an average piecewise precision of 66.50% and Macrae had an average piecewise precision of 22.85%.

### 2.5.5.3 Required Accuracy

Different applications of music alignment require different degrees of accuracy. Work on score-following systems consider estimates within 250 ms (Cont 2010) to 300 ms (Cont et al. 2007b) of the actual note to be correct. The current state of the art for following a range of instruments was presented by Cont (2010), where 91.49% of the onset estimates were within 250 ms of the ground truth and the mean offset. This statistic adds together both the estimation error and the latency of the system, which ranged from 75.1–240.9 ms, with standard deviations of 24.8–253.3 ms, across the pieces in the test set. The values do not take latency into account. Within this 250 ms error rate, there are certain instruments for which score following is effective, but others remain for which it needs to be improved, including the singing voice.

For digital music libraries, where the alignment is used to either visually link the score to the audio during playback or to find a particular section of the piece, the alignment precision could range from note-level, at its most precise, to the bar-level, for certain applications. For a piece in 4/4 that is performed at 120 bpm, this would translate to 500 ms for quarter note-level precision and 2000 ms for measure-level precision. Evaluations of digital libraries are typically made in terms of overall usability or retrieval accuracy rather than onset estimation accuracy.

In contrast, expressive performance studies require much greater precision. For timing-based studies, this is particularly important when evaluating asynchrony in the performance of notated simultaneities either on a polyphonic instrument, such as the piano, or between instrumentalists in an ensemble, which can range from 7 to 50 ms (Palmer 1997). Also, when the alignment is used to guide signal-processing algorithms in complex polyphonic

recordings to estimate pitch or dynamic information, the discrepancy for the onset and offset estimates needs to be as close to 0 ms as possible. Currently, none of the existing approach comes close to this level of precision on non-synthesized audio, and this remains an open area of research. Section 3.1 describes an alignment algorithm optimized for accurate estimation of notes onsets and offsets in monophonic recordings of the singing voice.

### 2.5.6 Summary

This section has surveyed various approaches to audio-score alignment and their applications. The majority of the early work focused on real-time following of monophonic instruments for use in live performance settings. More recently, researchers have also explored the use of audio-score alignment for analyzing the content of musical signals. The next chapter details experiments that demonstrate that none of these existing methods are sufficient for accurately identifying the location of notes onsets and offsets in recordings of the singing voice and presents a new alignment algorithm for addressing this problem.

(This page intentionally left blank)

## Chapter 3: Automatic Extraction of Performance Parameters

This chapter details the methods used to extract and analyse intonation-related data in the audio recordings used in the experiments in Chapter 4. The first main section (3.1) describes the challenge of automatically annotating note onsets and offsets in audio files. This section builds on the literature review in Section 2.5 and focuses on the challenges of determining onset and offset in recordings of the singing voice through score-audio alignment. Section 3.1.1 details a set of experiments on the utility of existing dynamic time warping alignment algorithms on recordings of the singing voice. Section 3.1.2 describes a new approach for score-audio alignment for such signals that bootstraps a hidden Markov model that uses the acoustical properties of the singing voice with an existing dynamic time warping alignment algorithm. The second main section (3.2) discusses the tools used in this dissertation for extracting and describing fundamental frequency information from the audio recordings once the note onsets and offsets have been determined. Section 3.2.1 discusses how de Cheveigné and Kawahara's YIN algorithm (2002) is used to extract  $F_0$  estimates. Section 3.2.2 explains how a perceived pitch for each note is calculated from these estimates. Section 3.2.3 details the use of the discrete cosine transform to model the evolution of  $F_0$  through estimations of the slope and curvature of the  $F_0$  trace for each.

### 3.1 Annotation of Audio Files with Score-Audio Alignment

No robust solutions currently exist for annotating note onsets and offsets in recordings of the singing voice. The current state of the art for score-guided onset and offset estimation is described in Section 2.5. One general theme that emerges from the literature is that blind estimation algorithms are less robust than score-audio alignment algorithms, particularly for audio with non-percussive onsets, such as the singing voice. The singing voice is particularly challenging for several reasons: the difficulty of determining note onsets and offsets when notes change under a single syllable, the differences in onset characteristics between vowels and consonants, and the acoustic characteristics that accompany different types of attacks and articulations. The spectrographic representation in Figure 3.1.1 demonstrates the

acoustic difference between a vocal line and a solo drum line, which can be considered a nearly optimal signal for onset detection. In particular, the broad band energy and higher amplitude at drum hits is easier to track than the pitch-based shifts that are the most salient characteristics of onsets and offsets in solo vocal recordings. Unlike percussive onsets, amplitude is not a useful cue for onset and offset detection in the solo vocal recordings since singers often achieve their highest amplitude level mid-note. The differences between these types of sound clearly impact the onset detection algorithms' performance. Figure 3.1.2 shows the results from the 2007 Music Information Retrieval Evaluation eXchange (MIREX) Onset Detection Task's solo drum examples (Downie 2007). The high values for precision (the number of correct onset estimates divided by the total number of onset estimates) and recall (the number of correct onset estimates divided by the number of onsets in the ground truth) in nearly all of the algorithms demonstrate how good the state of the art is for percussive onsets. In contrast, the results for the singing voice examples in Figure 3.1.3 are much lower and more varied.

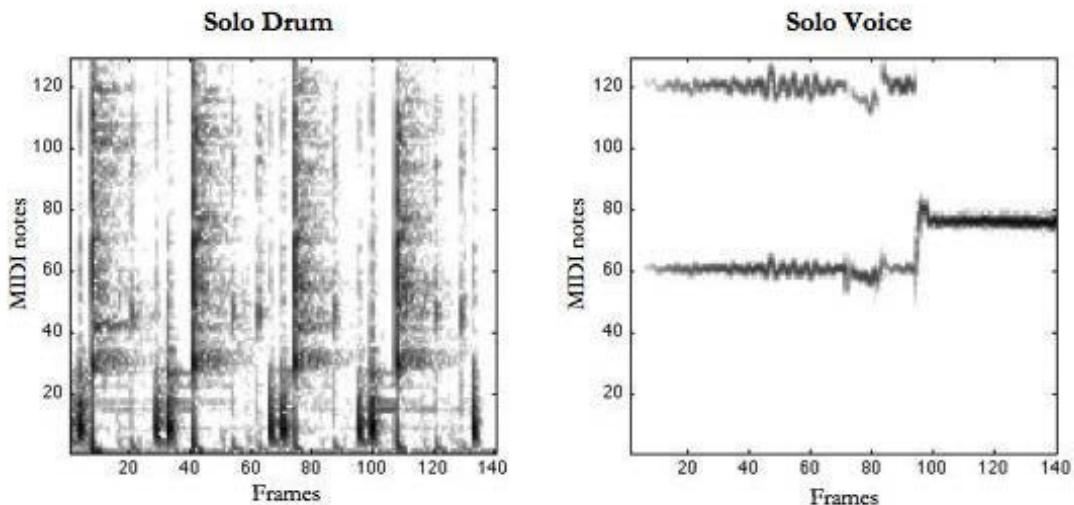


Figure 3.1.1: Spectrographic representations of 7-second audio clips of solo drum (left) and solo voice (right).

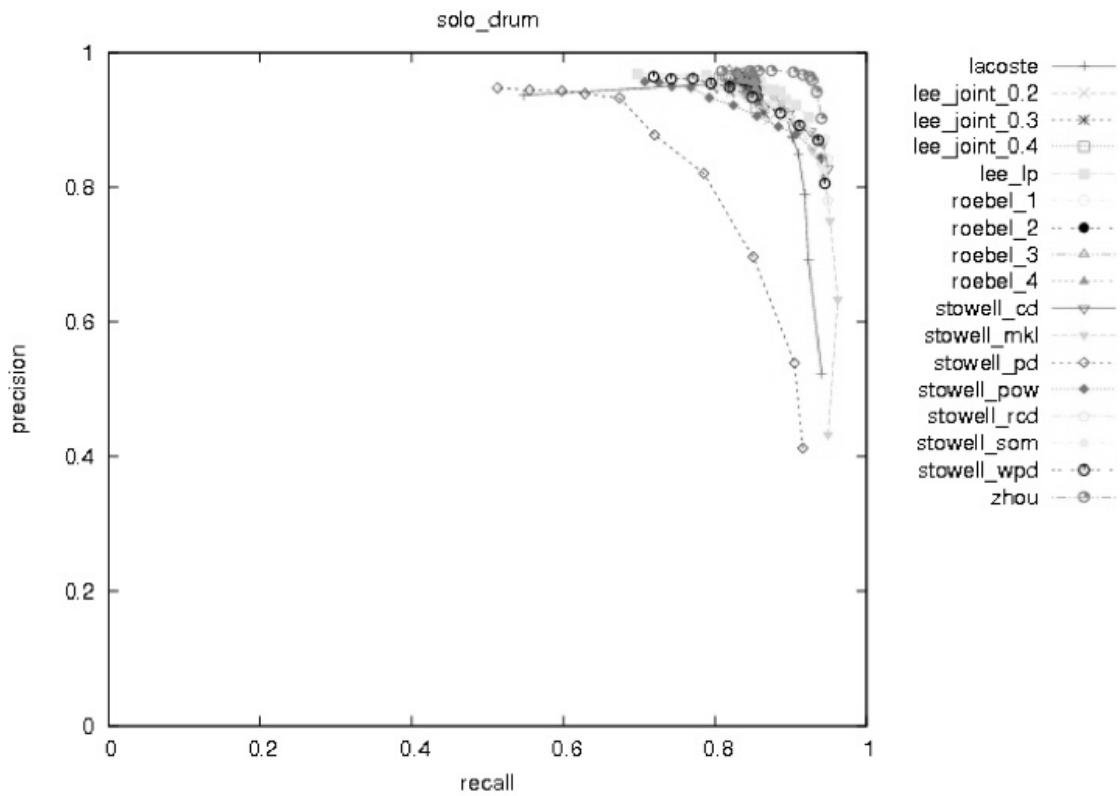


Figure 3.1.2: Plot results from MIREX 2007 Onset Detection evaluation for solo drum from Downie (2007). The y-axis is precision, the number of correct onset estimates divided by the total number of onset estimates, and x-axis is recall, the number of correct onset estimates divided by the number of onsets in the ground truth. The legend on the right indicates which algorithm is which. Details of the algorithms can be found on the MIREX 2007 web page (Downie 2007).

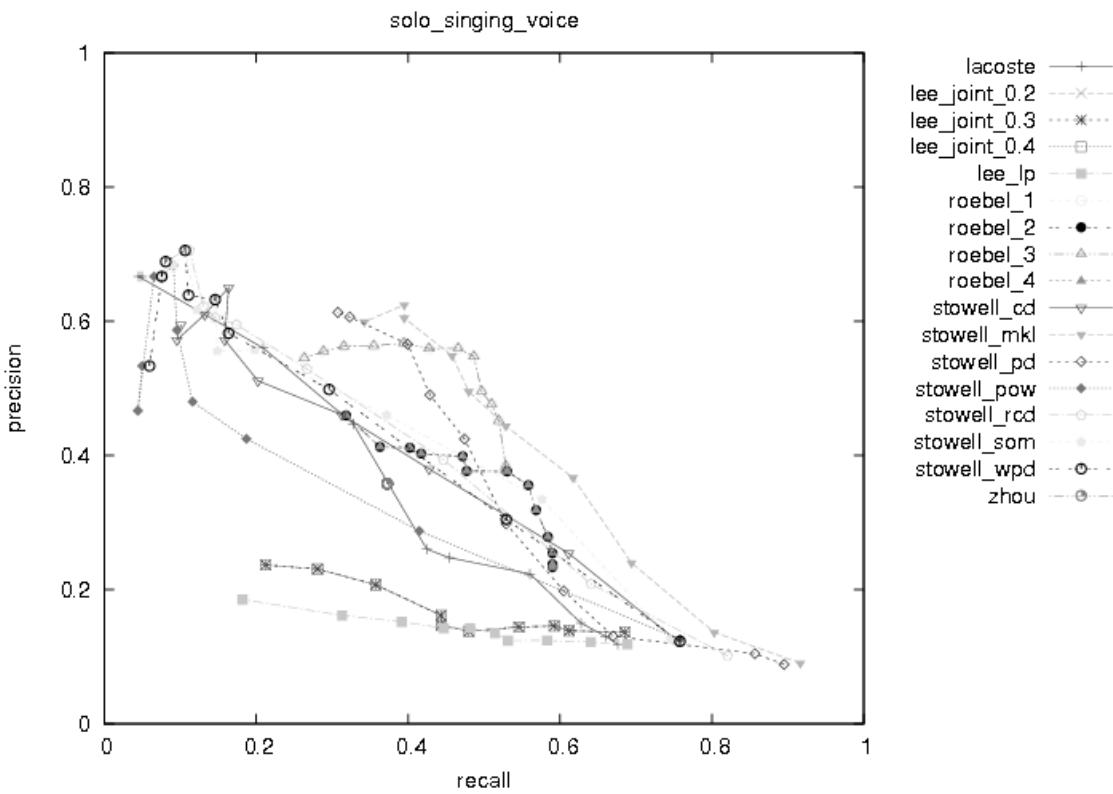


Figure 3.1.3: Plot of the results from MIREX 2007 Onset Detection evaluation for singing voice from Downie (2007). The y-axis is precision, the number of correct onset estimates divided by the total number of onset estimates, and x-axis is recall, the number of correct onset estimates divided by the number of onsets in the ground truth. The legend on the right indicates which algorithm is which (Downie 2007).

For the purposes of this research, the offsets of each note also had to be accurately identified in order for fundamental frequency ( $F_0$ ) estimates to be calculated for each frame of audio in each note. Since scores are available for all of the recordings used in this research, score-audio alignment can be used for annotation. However, none of the existing algorithms have been extensively tested for the singing voice. The emphasis of earlier evaluations was on timing information; therefore, they focused solely on onsets. Also, the evaluations used different data sets, often without any isolated examples of the singing voice.

The first step in the evaluation compared dynamic time warping (DTW)-based algorithms (Orio and Schwarz 2001; Dannenberg and Hu 2003; Turetsky and Ellis 2003) and hidden

Markov model (HMM)-based algorithms (Raphael 2004; Peeling et al. 2007). A qualitative evaluation of these algorithms on recordings of the singing voice showed that offline DTW-based algorithms performed better, likely because they are more constrained than the online HMM-based algorithms. A two-part quantitative evaluation of DTW-based algorithms for onset and offset estimation on recordings is described in Section 3.1.1. These evaluations demonstrate that none of the existing approaches were sufficiently accurate for our purposes, as we require accurate identification of not only the onset, but also the locations of the transients and steady-state portions in the notes. The accuracy of the DTW-based approach was improved with a newly developed two-step approach, where an HMM is bootstrapped with an existing DTW alignment in order to increase its accuracy. The algorithm and its evaluation are detailed in Section 3.1.2.

### **3.1.1 Evaluation of Dynamic Time Warping Approaches to Alignment**

This section demonstrates the limits of existing audio-score alignment approaches for annotating onsets and offsets in recordings of the singing voice. Section 3.1.1.1 details how the ground truth was collected for the evaluations. The following section describes the test data. Section 3.1.1.3 describes the evaluation metric used for the experiments. The first experiment, described in Section 3.1.1.4, is a quantitative evaluation of three different DTW approaches (Orio and Schwarz 2001; Dannenberg and Hu 2003; Turetsky and Ellis 2003). The second experiment, described in Section 3.1.1.5, is a more detailed evaluation of the method by Orio and Schwarz (2001).

#### **3.1.1.1 Ground Truth Collection**

Ground truth was collected by manually labelling the audio in Audacity, an open source audio editor that provides time- and frequency-domain representations of the audio, as well as labelling functionality. The general location of the onset and offset of each note is determined aurally. The estimates were then refined through repeated listening of small segments of audio while alternating between the time-domain (Figure 3.1.4) and frequency-domain (Figure 3.1.5) representations. Each note’s onset and offset estimations were verified several times before the next note was annotated. Manual annotation takes about 10–12 times real-time (i.e., 10–12 times the duration of the audio file). The pieces used for this evaluation were annotated multiple times to ensure consistency.

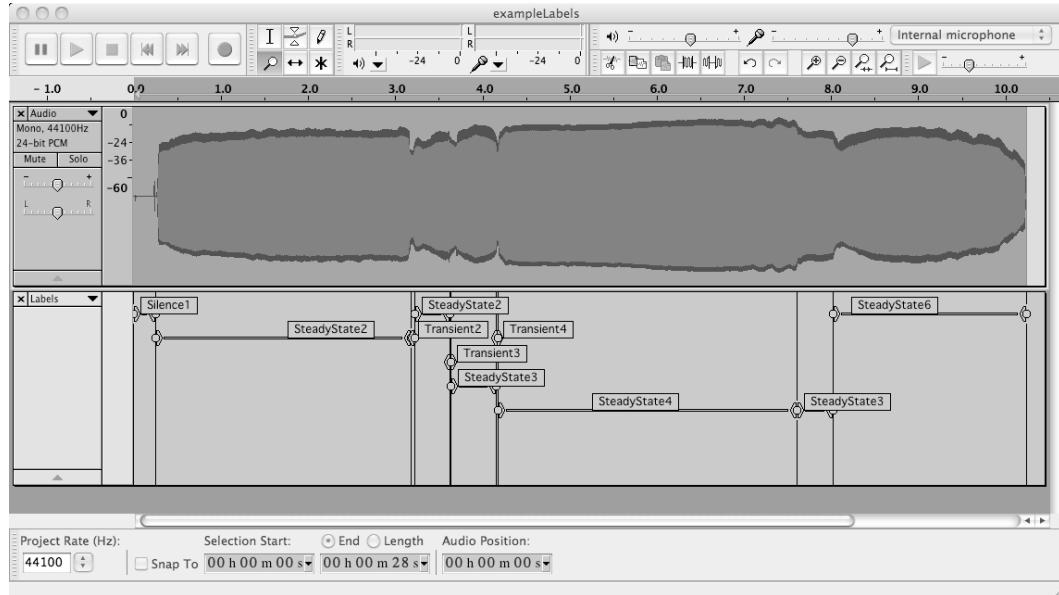


Figure 3.1.4: Time-domain representation of audio in Audacity with note onsets and offsets labelled underneath. In the audio representation in upper window, the x-axis represent time in seconds, and the y-axis represents amplitude in dB. In the lower window, the labels are visualized and labelled as silence, transient, or steady state.

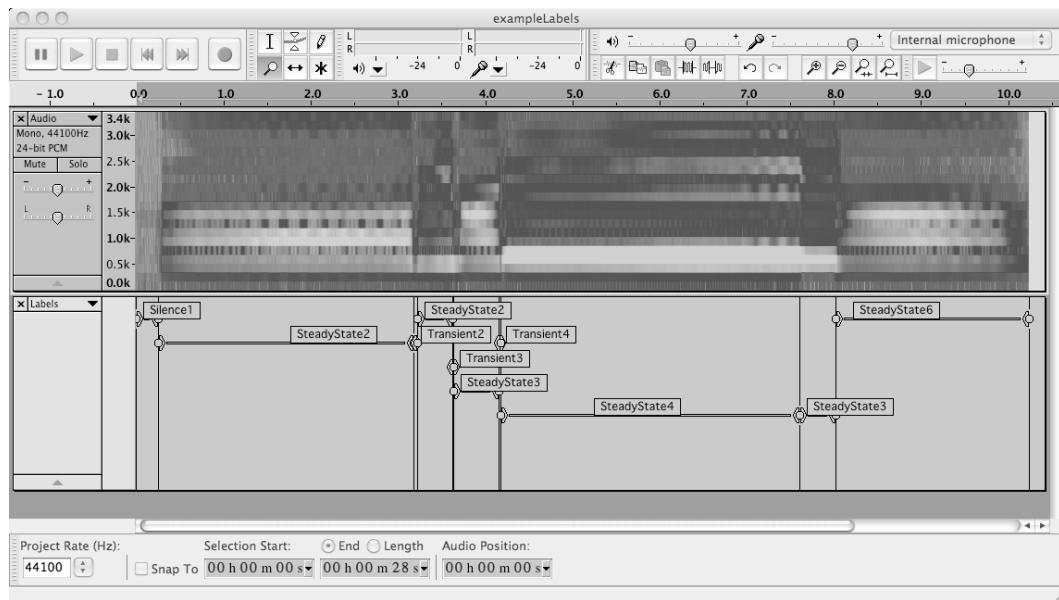


Figure 3.1.5: Frequency-domain representation of audio in Audacity with note onsets and offsets labelled underneath. In the audio representation in upper window, the x-axis represent time in seconds, and the y-axis represents frequency in Hz. In the lower window, the labels are visualized and labelled as silence, transient, or steady state.

### 3.1.1.2 Test Data

The test data for the DTW evaluation consisted of four 40 s multi-tracked audio recordings from the opening of Guillaume de Machaut's *Notre Dame Mass* (see Figure 3.1.6 for score). The recordings were made in the Centre of Research in Music Media and Technology (CIRMMT) labs at McGill University by first recording the singers as an ensemble and then re-recording each singer individually while they listened to the ensemble recording. This allowed for the creation of synchronized yet completely isolated multi-track recordings. Details of the recording process are reported in Wild and Schubert (2008).

A musical score for the opening of Machaut's *Notre Dame Mass*. The score is in common time (indicated by a '3') and consists of four staves. The top staff is labeled 'Triplum' and has a treble clef. The second staff is labeled 'Motetus' and also has a treble clef. The third staff is labeled 'Contratenor' and has a bass clef. The bottom staff is labeled 'Tenor' and has a bass clef. Each staff contains a series of notes and rests, representing the vocal parts of the mass.

Figure 3.1.6: Score of the opening of Machaut's *Notre Dame Mass*.

The multi-track recordings were combined four ways: Triplum (Soprano), Triplum (Soprano) + Motetus (Alto), Triplum (Soprano) + Tenor + Contratenor (Bass), and Triplum (Soprano) + Motetus (Alto) + Tenor + Contratenor (Bass). These combinations were used to represent different combinations of voices and in order to be representative of the different frequency ranges in the piece.

### 3.1.1.3 Evaluation Framework

The evaluation metric looks at the accuracy of note onset and offset alignment by comparing the collected ground truth to the results of the alignment algorithms. As described in Section 2.5.5, this is done by comparing each time value in the ground truth against the list of warped MIDI values generated by the alignment algorithm. If the absolute difference between the aligned value and the ground truth is less than the value defined in the experiment, either 50 or 100 ms, the alignment is considered to be correct. Each correct onset and offset alignment scores one point in their respective counts. This evaluation

method is similar to the metric described in Cont et al. (2007), but differs in that they also discussed some other measures that are not relevant to the offline systems evaluated here: latency in the actual detection, offset in reporting of the detection, missed notes (which are known not to exist in this dataset), and misaligned notes (which are already penalized in our accumulated count).

### 3.1.1.4 Experiment One: Comparison of DTW Approaches

This experiment focused on three different DTW-based alignment approaches: Orio and Schwarz, Dannenberg and Hu, and Turetsky and Ellis. As described in Section 2.5.3, Orio and Schwarz (2001) used peak structural distance as a feature and a model of attacks and silence for matching a note mask representation of a MIDI file to an audio file. Dannenberg and Hu (2003) used chromagograms to match a sonified version of a MIDI file to an audio file. Turetsky & Ellis (2003) used a combination of the cosine difference of spectral power, first order difference between channels, and first order difference in frequency for aligning sonified MIDI to audio files. However, since the recordings in this experiment are monophonic, only the cosine difference of spectral power was used in the evaluation. More details about these algorithms can be found in Section 2.5. The evaluation was done using publicly available MATLAB code (Dannenberg 2003; Ellis 2003, 2008), which was modified in order to account for differences between the offset of one note and the onset of the next by inserting an optional silence between each of the notes.

The performance of each of the alignment algorithms is detailed in Table 3.1.1–Table 3.1.3. Table 3.1.3 shows the mean difference between the ground truth and the alignment algorithms for the onsets and offsets, as well as the minimum and maximum differences. Table 3.1.2 shows a tally of the number of onsets and offsets for each algorithm within 50 ms of the ground truth, and Table 3.1.3 show the tally for the number of onsets and offset within 100 ms. Taken in combination, these tables indicate both the average accuracy (Table 3.1.1) and the robustness (Table 3.1.2 and Table 3.1.3) of the algorithms.

Orio and Schwarz						
	Onsets			Offsets		
	Mean	Min	Max	Mean	Min	Max
<b>Triplum</b>	0.1344	0.0026	0.8373	0.0715	0.0025	0.7655
<b>Triplum+Motetus</b>	0.1301	0.0026	0.8455	0.1110	0.0025	1.0176
<b>Triplum+Tenor+Contratenor</b>	0.2262	0.0029	0.9346	0.1739	0.0026	0.6219
<b>Triplum+Motetus+Tenor+ Contratenor</b>	0.1798	0.0015	0.9346	0.1546	0.0018	1.5641

Dannenberg & Hu						
	Onset			Offset		
	Mean	Min	Max	Mean	Min	Max
<b>Triplum</b>	0.2491	0.0180	0.9708	0.1763	0.0205	0.9345
<b>Triplum+Motetus</b>	0.2057	0.0013	0.4997	0.1749	0.0011	1.0176
<b>Triplum+Tenor+Contratenor</b>	0.3273	0.0039	1.7935	0.3358	0.0205	3.3069
<b>Triplum+Motetus+Tenor+ Contratenor</b>	0.2831	0.0059	2.5435	0.3136	0.0040	4.0569

Turetsky and Ellis						
	Onsets			Offsets		
	Mean	Min	Max	Mean	Min	Max
<b>Triplum</b>	0.1310	0.0026	0.8373	0.0721	0.0025	0.7655
<b>Triplum+Motetus</b>	0.1270	0.0026	0.8455	0.1037	0.0025	0.8176
<b>Triplum+Tenor+Contratenor</b>	0.1536	0.0029	0.6446	0.1324	0.0008	0.6219
<b>Triplum+Motetus+Tenor+ Contratenor</b>	0.1436	0.0029	0.5817	0.1320	0.0008	0.6522

Table 3.1.1: Mean, minimum, and maximum difference in calculated onset and offset values in the alignment from the ground truth in ms for each of the algorithm evaluated in Experiment One.

Filename (# of onsets and offsets)	Orio & Schwarz		Dannenberg & Hu		Turetsky & Ellis	
	Onset	Offset	Onset	Onset	Onset	Offset
Triplum (62)	10	18	1	1	10	10
Triplum+Motetus (122)	15	30	8	9	16	16
Triplum+Tenor+Contratenor (170)	15	20	8	4	18	18
Triplum+Motetus+Tenor+ Contratenor (198)	19	35	15	12	19	19

Table 3.1.2: Tallies of number of onsets and offsets estimated by the algorithms that are within 50 ms of the ground truth.

Filename (# of onsets and offsets)	Orio & Schwarz		Dannenberg & Hu		Turetsky & Ellis	
	Onset	Offset	Onset	Onset	Onset	Offset
Triplum (62)	18	26	1	11	18	18
Triplum+Motetus (122)	33	42	9	25	33	33
Triplum+Tenor+ Contratenor (170)	26	29	8	22	31	31
Triplum+Motetus+Tenor+ Contratenor (198)	37	48	22	42	41	41

Table 3.1.3: Tallies of number of onsets and offsets estimated by the algorithms that are within 100 ms of the ground truth.

Overall, the Turetsky and Ellis and Orio and Schwarz algorithms performed better than Dannenberg and Hu and, as a general rule, the accuracy was better for excerpts with fewer voices. The poorer performance of the Danneberg and Hu algorithm suggests that chroma are not the best features for aligning vocal music, particularly for excerpts with three and four voices. There was no significant difference between the other two algorithms, as both performed fairly consistently for the entire dataset. However, the Orio and Schwarz algorithm has an advantage over the Tuetsky and Ellis algorithm in its implementation. The use of a note mask, instead of a sonified version of the MIDI file, greatly streamlines and expedites the running of the algorithm, particularly in the MATLAB environment.

### 3.1.1.5 Experiment Two: Evaluation of Orio and Schwarz under different conditions

Experiment One evaluated the performance of three different DTW-based alignment algorithms for cases where the same information was available in both the audio and the MIDI. This experiment compares the performance of Orio and Schwarz's algorithm for three conditions: in the first, each line of the monophonic recording of each part was aligned to the corresponding monophonic MIDI data; in the second, all four MIDI parts were aligned to the polyphonic composite of the individual multi-tracks; and in the third the individual MIDI parts were aligned to the polyphonic composite. The first condition allowed the DTW alignment algorithm to perform under the simplest circumstance, where all of the harmonic information in the signal was related to each note in the MIDI file. The second condition presented the algorithm with more material to align. In this condition, simultaneous score events were treated as single events with a single time in the alignment. The third condition evaluated whether aligning each vocal line individually allows for more accurate timing estimates for each line within a polyphonic recording. This third condition is

an important one to evaluate in the context of this thesis since both solo and ensemble performances of the singing voice are studied. Also, as noted above, the second condition cannot accurately account for the asynchronies between simultaneously performed notes because only a single time warp is created. All notes that occur simultaneously in the score are assigned the same onset and offset time; therefore each voice's onsets and offsets cannot be accurately annotated.

The test data for this experiment was the same hand-annotated forty-second excerpt of multi-tracked recordings of the Kyrie from Machaut's *Notre Dame Mass* used in Experiment One. Likewise, the evaluation metric also looks at the note onset and offset alignment estimates against manually annotated ground truth. For this experiment, two measures were considered. The first tallies the number of alignments that are within 100 ms of the ground truth's onsets and offsets (Table 3.1.4) and details the average amount that the alignments in each component of each test were off from the ground truth and their standard deviation (Table 3.1.5).

The results demonstrate that the simultaneous alignment (Condition 2) performs comparably to the individual alignment (Condition 1). At times, the simultaneous alignment outperforms the individual alignment. This was due to the fact that the need to match multiple notes constrains the DTW algorithm and reduces the likelihood of it getting temporarily lost. Figure 3.1.7 and Figure 3.1.8 show that in both conditions, the alignment algorithm is able to consistently find the relevant notes in the audio signal, but that the determination of the exact location of onsets and offsets is not always accurate. Figure 3.1.9 provides a visual example of the asynchrony issue in Condition 2. Around 13.3 s, there is notated simultaneity between the soprano and the bass: the alignment is locked to the onset of the soprano's note, which, in performance, is about 30–40 ms behind the onset of the bass' note. Also, the offset of the tenor note occurs approximately 100 ms before the other voices' offsets. As noted above, one way of addressing the asynchrony is to align the lines one at a time against the composite signal (Condition 3). Figure 3.1.10 shows the main drawback of this approach, which is that the alignment algorithm can easily become lost when aligning a single line in the presence of multiple voices.

Vocal Part (Number of notes)		Test 1 Individual	Test 2 Composite Simultaneous	Test 3 Composite Individual
<b>Soprano (31)</b>	<i>On</i>	7 (22%)	8 (26%)	8 (26%)
	<i>Off</i>	22 (71%)	21 (26%)	18 (58%)
<b>Alto (30)</b>	<i>On</i>	6 (20%)	10 (33%)	7 (23%)
	<i>Off</i>	20 (67%)	14 (70%)	17 (57%)
<b>Tenor (14)</b>	<i>On</i>	4 (29%)	6 (42%)	3 (21%)
	<i>Off</i>	7 (50%)	9 (64%)	2 (14%)
<b>Bass (24)</b>	<i>On</i>	5 (21%)	16 (67%)	8 (33%)
	<i>Off</i>	14 (58%)	15 (62%)	9 (38%)
<b>Totals (99)</b>	<i>On</i>	31 (31%)	40 (40%)	26 (26%)
	<i>Off</i>	63 (64%)	59 (60%)	46 (46%)

Table 3.1.4: The number of onsets and offsets predicted within 100 ms of the ground truth.

Vocal Part (Number of notes)		Test 1 Individual		Test 2 Composite Simultaneous		Test 3 Composite Individual	
		Mean	SD	Mean	SD	Mean	SD
<b>Soprano (31)</b>	<i>On</i>	0.163	0.144	0.146	0.096	0.237	0.254
	<i>Off</i>	0.092	0.063	0.086	0.056	0.185	0.267
<b>Alto (30)</b>	<i>On</i>	0.194	0.146	0.182	0.153	0.229	0.195
	<i>Off</i>	0.154	0.224	0.179	0.174	0.165	0.216
<b>Tenor (14)</b>	<i>On</i>	0.206	0.232	0.124	0.082	1.419	1.598
	<i>Off</i>	0.327	0.082	0.074	0.059	1.815	1.579
<b>Bass (24)</b>	<i>On</i>	0.132	0.065	0.098	0.093	0.228	0.342
	<i>Off</i>	0.108	0.102	0.110	0.119	0.298	0.668
<b>All</b>	<i>On</i>	0.171	0.146	0.142	0.117	0.612	0.836
	<i>Off</i>	0.147	0.331	0.118	0.124	0.693	0.975

Table 3.1.5: Mean and standard deviation in seconds between the onset and offset set alignments and the ground truth.

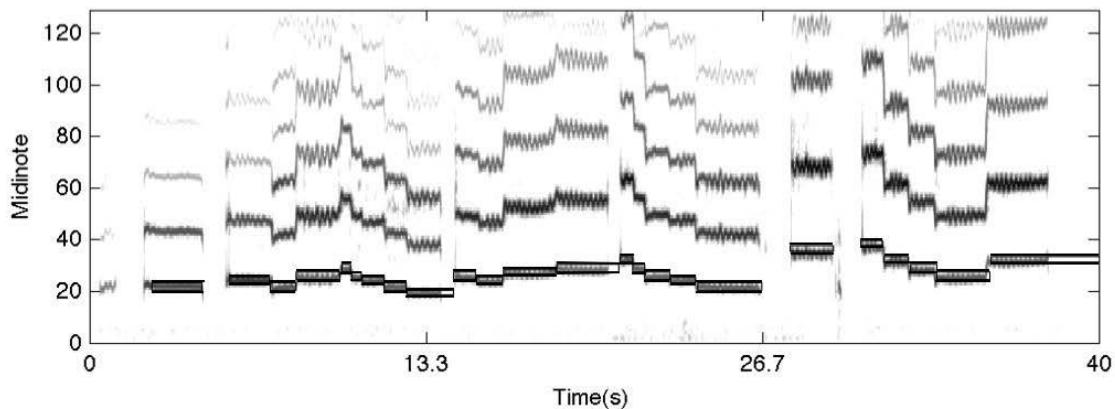


Figure 3.1.7: Condition 1: Overlay of alignment of a single line aligned to a single voice.

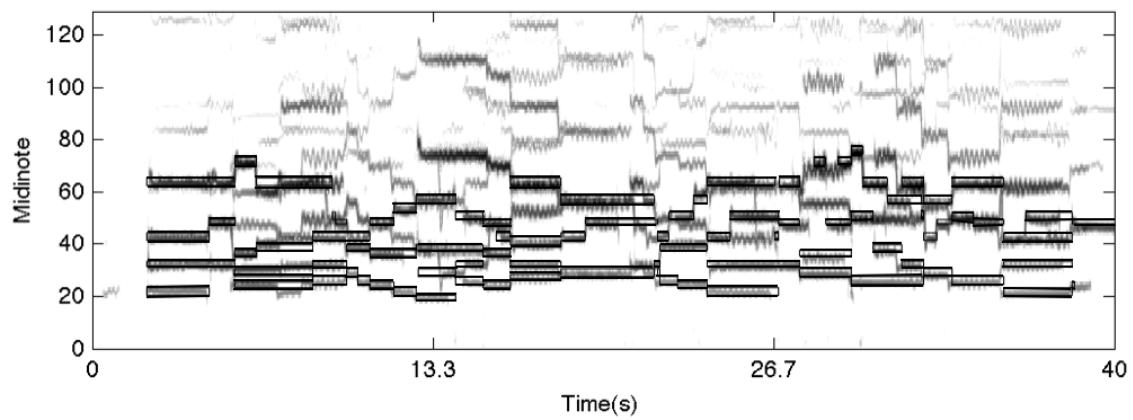


Figure 3.1.8: Condition 2: Overlay of alignment for all four lines aligned simultaneously to a composite signal.

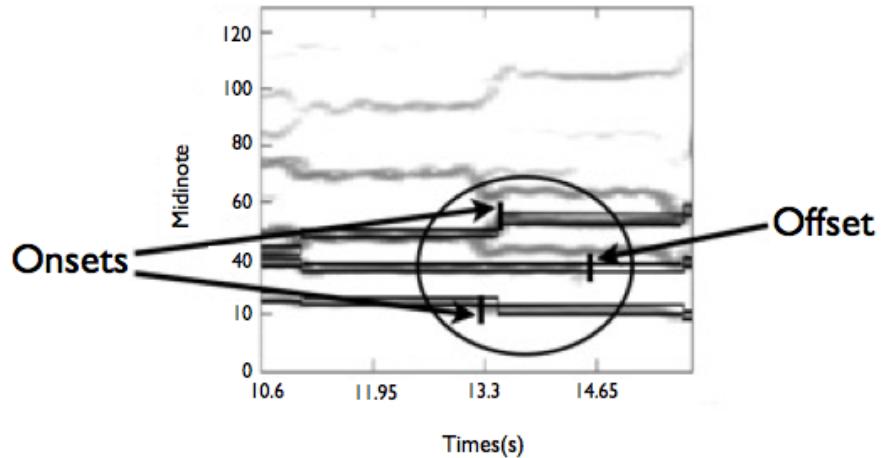


Figure 3.1.9: Condition 2: Example of a performance asynchrony for a notated simultaneity.

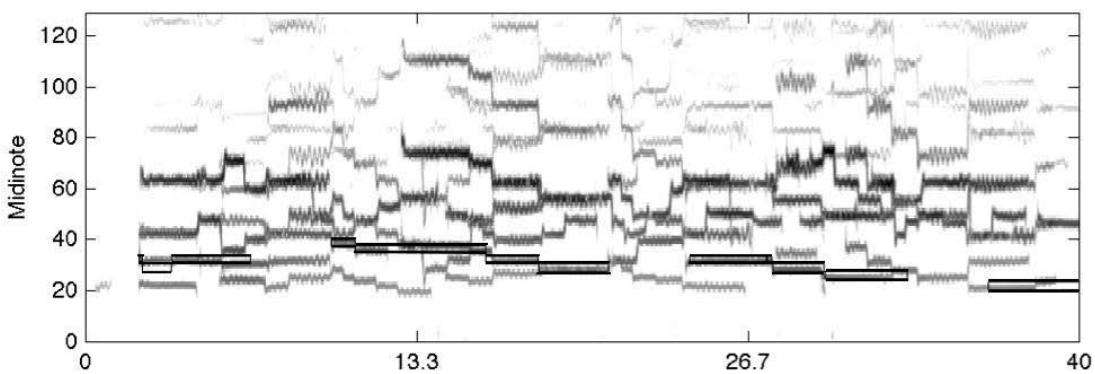


Figure 3.1.10: Condition 3: Overlay of alignment for a single line aligned to a composite signal of all the voices.

### **3.1.2 Improving Alignment Accuracy**

As demonstrated in Section 3.1.1, current score-audio alignment algorithms are not sufficient for accurately estimating note onsets and offsets in recordings of the singing voice. This section describes a technique for improving the accuracy of score-audio alignment by using known acoustical properties of the signal to train a hidden Markov model (HMM) to identify silence, transient, and steady-state portions of each note. The described implementation is for solo singing voice, although the technique could be applied to other instruments by modifying the acoustical features and to polyphonic signals with the use of an algorithm capable of producing the required acoustical descriptions.

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 multi-layer neural network for onset detection. Similarly, HMMs have previously been used for describing signals containing the voice in Shih, Narayanan, and Kuo (2003) and Ryyynanen (2006). In Shih, Narayanan, and Kuo (2003), a three-state HMM was implemented to model the phonemes of hummed notes for a query-by-humming application. Ryyynanen (2006) deals explicitly with transcription of the singing voice and uses a three-state note event HMM and a four-component rest event Gaussian mixture model (GMM) trained on examples of singing and no-singing audio frames, respectively.

Since the HMM performs only local adjustments to the alignment, a relatively accurate initial alignment is important for this technique and achievable with DTW. Following from Section 3.1.1, Orio and Schwartz's algorithm is used for the initial DTW alignment. The HMM was implemented in Matlab with Kevin Murphy's HMM Toolbox (1998) using periodicity and power estimates from Alain de Cheveigné's YIN algorithm (2002).

#### **3.1.2.1 Acoustical Properties of the Singing Voice**

The design of the HMM was based on the acoustical properties of the singing voice. As a result, this implementation is optimized for the singing voice and would require some adjustment to work with other instruments. The amplitude envelope and periodic characteristics of a sung note are influenced by the words that are being sung. The three acoustic events modeled for this system (silence, transient, and sustain/steady state) are

shown in Figure 3.1.11. Transients occur when a consonant starts or ends a syllable, while vowels produce the steady-state portion of the note. The type of consonant, voiced or unvoiced, affects the characteristics of the transient, as does the particular manner in which the singer attacks or enunciates the consonant. The motivation for identifying transients is to determine where the voiced section of the note begins for estimating a single fundamental frequency of the duration of the note.

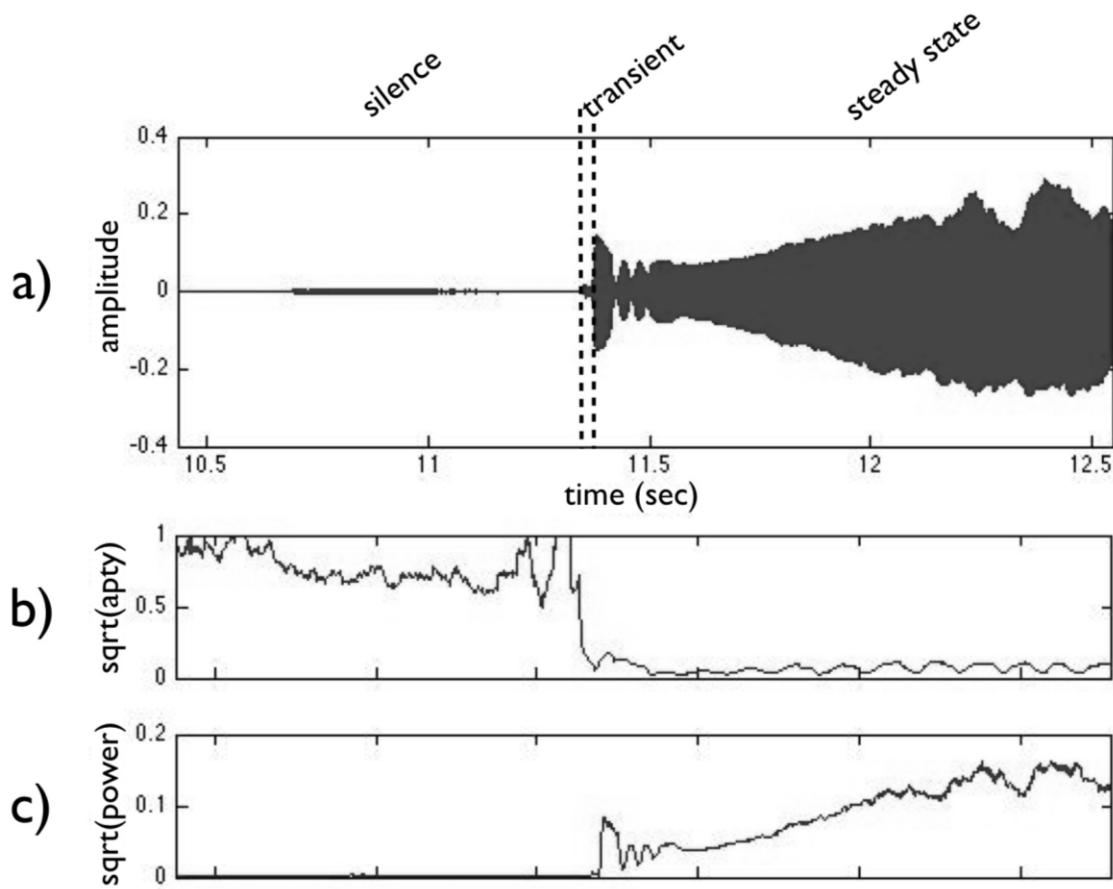


Figure 3.1.11: Time domain representation of a sung note's waveform; (a) is the time domain representation of a sung note with the HMM states labelled, (b) is the aperiodicity measure, and (c) is the power measure. The aperiodicity and power measurements are used as observations for the HMM, whose three states (silence, transient and steady state) are labelled across the top of the figure.

### 3.1.2.2 HMM Details

The basic implementation of this HMM has three states: silence, transient, steady state (see Figure 3.1.12). In the figure, there are two types of transients, beginning and ending, which allow for correct modeling of where the consonants occur in each syllable. An optional fourth state, breath, was introduced experimentally, which improved results in some cases; however, the breath state should not be considered an essential component of the model (see Figure 3.1.13). A second silence after the breath state is added to this state sequence to reflect the common practice among singers of briefly holding the inhaled breath before singing the next note.

The transition probability values were calculated from a superset of the music used in the experiments in Sections 3.1.1.4, 3.1.1.5, and 3.1.2.7, including Schubert’s “Ave Maria” and a Latin mass by Machaut. The silence, breath, transient, and steady-state portions of these pieces were hand-labelled. Self-loop probabilities, the probability of a state to repeat rather than change, were estimated from the average duration of each state in 90 seconds of audio. Non self-loop probabilities were estimated from summary statistics of 318 notes from these scores. Specifically, the transition probabilities to the transient states were set to reflect the likelihood of syllables beginning and ending with consonants in the Latin text. Transition probabilities to the silences were based on the average frequency of rests in the score: it was assumed that in the legato singing style that dominates the singing voice literature, silences would only occur at rest or breath marks.

Two versions of the state sequences were implemented. The first algorithm allows each state to be visited for each note. The second algorithm was determined by the particular lyrics being sung; transients were only inserted when a consonant began or ended a syllable and silences (and breaths for Algorithm Two-B) were inserted only at the end of phrases. The state sequence for the opening phrase of Schubert’s “Ave Maria” is shown in Figure 3.1.14.

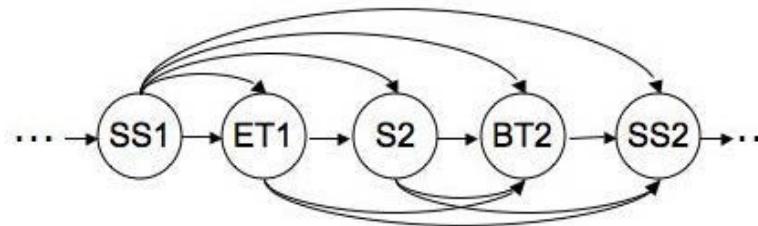


Figure 3.1.12: Three-state basic state sequence seed: steady state (SS), transient (T), silence (S). The ending transient (ET) and the beginning transient (BT) both have the same observation distribution.

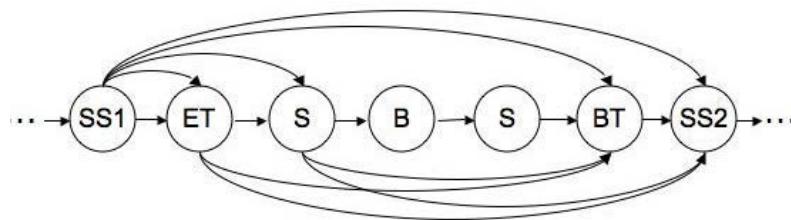


Figure 3.1.13: Basic state sequence seed plus breath (B).

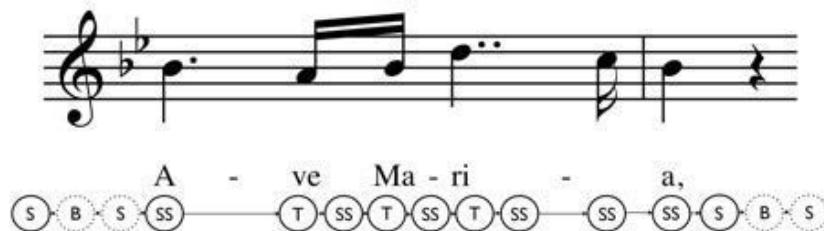


Figure 3.1.14: State sequence adapted to sung text. The circles with solid lines represent those states included in the basic state sequence, while the circles with the dotted lines represent those states that were added when the optional breath state was included.

The observations for the HMM are the square root of periodicity and power estimates provided by the YIN algorithm for each frame. The  $F_0$  estimates from YIN are also used, which provided 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 a single vowel. YIN estimates fundamental frequency by

measuring the self-similarity of a signal over time. While standard autocorrelation uses an inner product to measure similarity, YIN uses the squared difference to measure dissimilarity.

The YIN algorithm was applied to audio sampled at 44,100 samples/s with a frame size of 10 ms and a hop size of 0.7 ms. The mean and variance values for each frame were calculated by isolating representative examples of silence, transient, steady state, and breath from recordings by different singers. In total, 2.25 s of data were used to calculate the means and variances for silence, 13.4 s for steady state, 0.47 s for transients, and 3.83 s for breath.

The initial DTW alignment is used as a prior to guide the HMM (see Figure 3.1.15). The use of the DTW alignment obviates the need to encode information about the score in the HMM. By assuming that the DTW alignment is roughly correct, it is not necessary to encode pitch specific information into the HMM. This drastically simplifies the problem that the HMM has to address, since it simplifies the design of the HMM and allows the same HMM seed to be used for each note. One issue with this approach is that it cannot adjust the initial alignment by more than one note, so the initial alignment has to be relatively accurate.

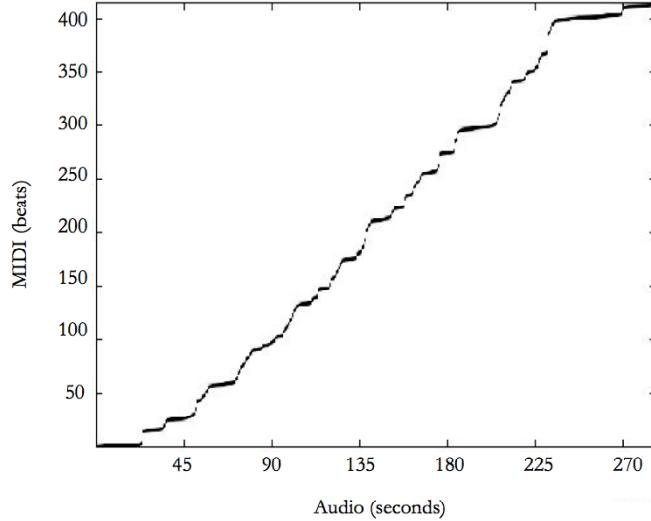


Figure 3.1.15: Visualization of the DTW alignment implemented as a prior for the HMM.

The prior is created by placing a rectangular window with half a Gaussian on each side over the note positions estimated by the DTW alignment. Each state has a different set of rules governing the placement and width of the windows, as well as the half Gaussians. This is detailed in Figure 3.1.16.

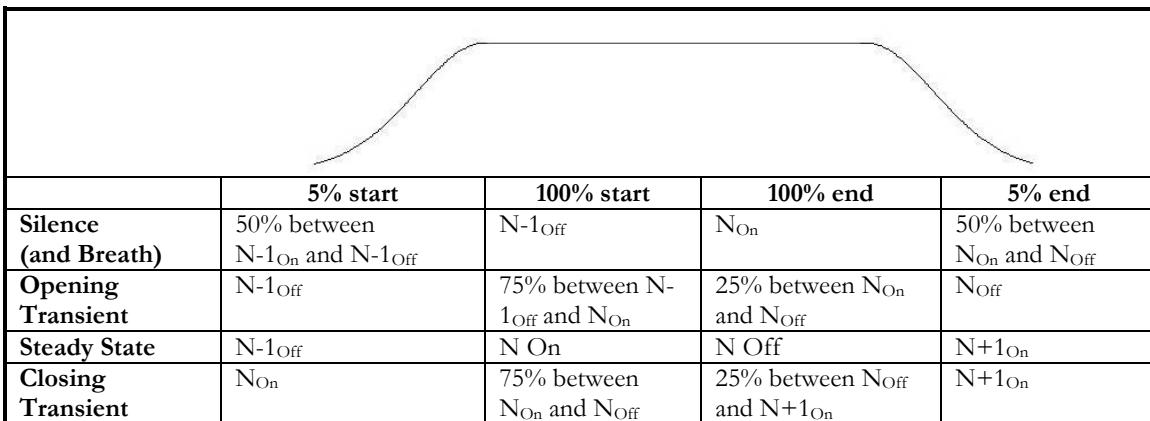


Figure 3.1.16: Gaussian distributions for the creation of a prior from the DTW alignment. N is the current note number.

### 3.1.2.3 Evaluation

Three annotated recordings of the opening of Schubert’s “Ave Maria” by three different singers were used to evaluate the system. The annotations were done manually using Audacity (Mazzoni and Dannenberg 2000), as described in Section 3.1.1.2. All of the singers had soprano voices; one was a professional and the other two were undergraduate vocal majors. The singers exhibited differences in overall timbre, attack time (transient length), and vibrato rates.

In Algorithm One, each note is modeled with a complete set of states. This is the baseline test, to evaluate whether performance is improved when the text is taken into account (Algorithm Two). The first version of this algorithm (One-A) uses the basic three-state HMM model (Figure 3.1.12) and the second (One-B) adds the optional breath state (Figure 3.1.13). In Algorithm Two, the state space is modified based on the presence of consonants in the sung text and phrase endings or rests in the score (Figure 3.1.14). As with Algorithm One, Algorithm Two was run both with the basic three-state HMM (Two-A) and with the optional breath state added (Two-B). The results of the experiments are detailed in Table 3.1.6, which provides the 2.5, 25, 50, 75, and 97.5 percentiles of the absolute difference between the manually annotated ground truth for both the experiments and the original DTW alignment. At the 50<sup>th</sup> percentile, or median, the second version of the algorithm without breath outperforms the DTW alignment with an error rate of 27.8 ms vs. 52.3 ms.

	Percentile				
	2.5	25	50	75	97.5
<b>Dynamic Time Warping</b>	3.2	32.6	52.3	87.9	<b>478.7</b>
<b>One-A: General w/o breath</b>	<b>1.6</b>	<b>13.1</b>	41.8	88.8	564.1
<b>One-B: General w/breath</b>	1.9	13.7	47.4	117.8	923.8
<b>Two-A: Textual w/o breath</b>	<b>1.6</b>	<b>13.1</b>	<b>27.8</b>	<b>78.0</b>	506.0
<b>Two-B: Textual w/breath</b>	2.1	13.7	41.8	91.3	923.1

Table 3.1.6: Results from Algorithms One and Two compared to the original dynamic time warping alignment in milliseconds. The bolding indicates the condition with the lowest error rate in each percentile.

In general, both algorithms provided greater alignment accuracy than the initial DTW alignment. However, the 75<sup>th</sup> and 97.5<sup>th</sup> percentiles for the unmodified state sequence of Algorithms One and Two were less accurate than the DTW. There was also consistent improvement in performance by the modified state sequence used in Algorithm Two over the unmodified sequence in Algorithm One. This was largely to be expected: since in the first algorithm, the HMM had the freedom to select a state that would not have occurred at certain points in the recorded performance. The addition of the breath state did not increase the accuracy of the alignment; rather it led to a small number of quite severe misalignments. Upon inspection, it emerged that these misalignments occurred at the silence-breath-silence states in Algorithm Two and not in the transient and steady-state portions of the notes where accuracy is more important for onset and offset annotation. In terms of the transient and steady-state alignments, the accuracy is comparable to Algorithm One.

A visual demonstration of the improvement in alignment can be seen in Figure 3.1.17. Here the boxes indicate the 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. 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 corrected by the HMM. Moreover, at 1 and 3, the HMM successfully identifies the presence of the transients at the start of the notes.

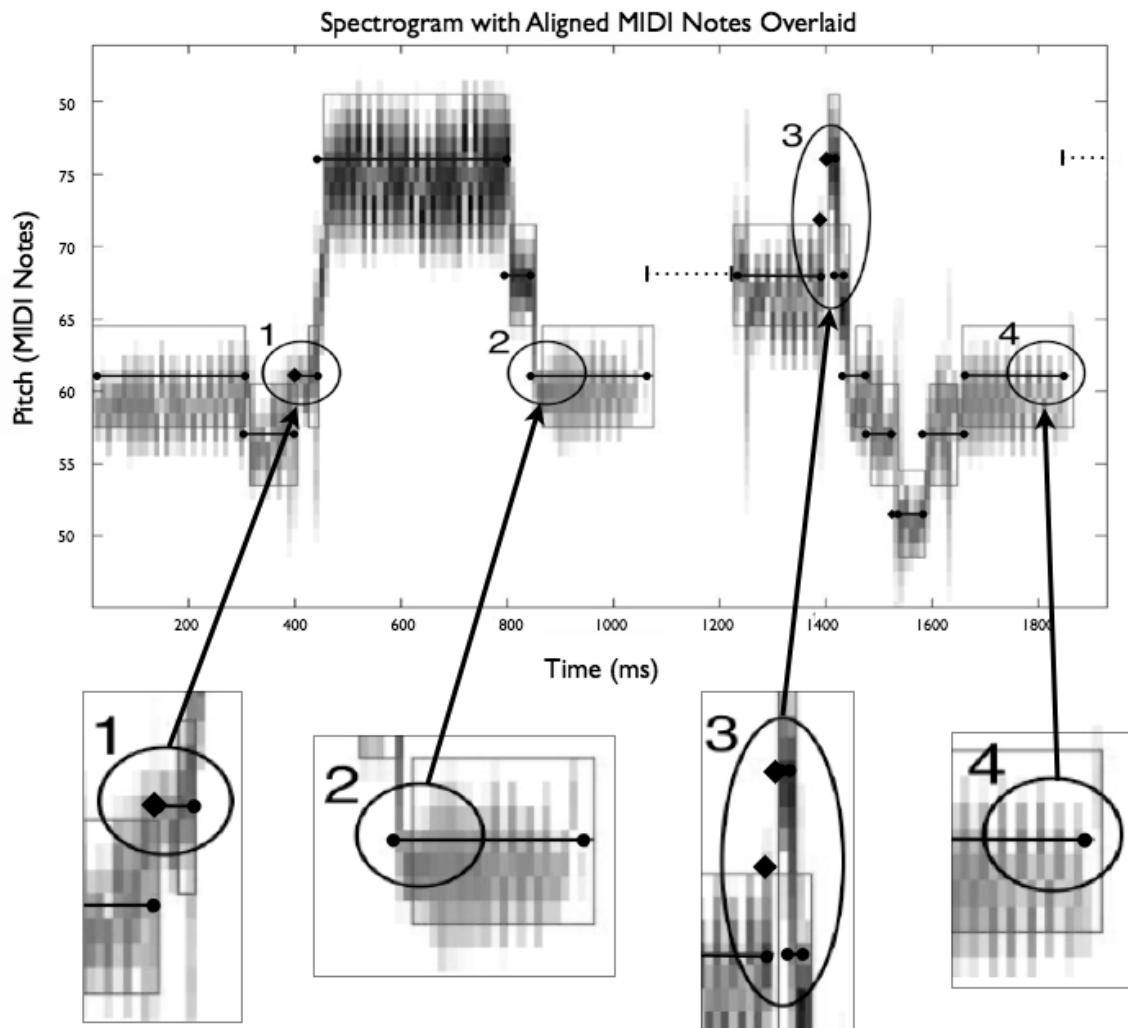


Figure 3.1.17: Visualization of the performance of the Algorithm Two-A versus the DTW alignment. The opening passage of a recording of the “Ave Maria” is represented as a zoomed-in log-frequency spectrogram. The boxes indicate the note positions estimated by the initial DTW alignment. The HMM estimates 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.

### 3.1.2.4 Discussion

Closer examination of where the HMM made incorrect state identifications revealed that some voiced consonants introduced a considerable amount of “noise” in steady-state sections (i.e., when the consonants are rolled). The implementation of the transient state predominantly modeled unvoiced consonants, while allowing for the possibility of several

frames of noise at the start of a voiced consonant. The implementation of the steady-state portion covered both voiced consonants and vowels. The reason for this is that the voiced consonants contribute to the perceived pitch. There is also some ambiguity present in the ground truth. Onsets and offsets in the singing voice are notoriously difficult to identify (Toh et al. 2008), which may affect the accuracy of the ground truth, and thus the results of the experiments, by several tens of milliseconds.

The algorithm does require some amount of manual intervention before it can be run. 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, though this only has to be done once for each piece. 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 run, 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 5 times the length of each note that needs correcting. In contrast, manually annotating the steady state and (where applicable) transient portions of each note takes about 10–12 times real time (i.e., 10–12 times the length of the entire audio file). 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 for each audio recording.

### 3.1.2.5 Conclusions

Overall, the three-state HMM algorithm 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, the median error was further reduced to 28 ms. The HMM algorithm also differentiated between transient and steady-state portions of the note. This differentiation is important when examining pitch-related performance practices, since only the steady-state portion of the note contributes to pitch perception.

### **3.1.3 Summary**

The tests in Section 3.1.1 demonstrate that existing score-alignment algorithms are not sufficiently accurate at identifying note onsets and offsets in recordings of the singing voice for the purposes of this research. Section 3.1.2 describes a new score-alignment algorithm that is informed by the acoustics of the singing voice, the lyrical content of the score, and the audio being aligned. This algorithm not only improves the alignment of notes onset and offsets over existing methods, but also identifies the transient and steady-state portions of the sung notes. The algorithm was used to annotate the recordings used in the experiments in Chapter 4.

## **3.2 Modeling Perceived Pitch and the Evolution of Fundamental Frequency**

Once the onsets and offsets of each note have been determined using the algorithm described in Section 3.1, the pitch-related information can be extracted and modeled. This section details several descriptors that are used in the experiments in Section 4. Section 3.2.1 discusses how fundamental frequency information is extracted from recordings. Section 3.2.2 builds on the perceived pitch literature discussed in Section 2.3.2.4 and details the way in which perceived pitch was calculated. Section 3.2.3 describes ways of modeling the evolution of the estimated fundamental frequency over the duration of each note.

### **3.2.1 Extracting Fundamental Frequency Information**

This research uses the YIN algorithm by de Cheveigné and Kawahara (2002) for fundamental frequency ( $F_0$ ) estimation. The YIN algorithm is an auto-correlation-related  $F_0$  estimator, whose technical details were described in Section 2.4. De Cheveigné's MATLAB implementation of YIN was used for this research (de Cheveigné 2002). This implementation allows for a number of parameters to be specified: minimum and maximum expected  $F_0$ s, window size, buffer size, hop size, and threshold.

The YIN algorithm begins by low-pass filtering the signal with a default setting of one quarter of the sampling rate. The minimum and maximum  $F_0$ s are set according to the note information in the aligned score. For this research, the maximum  $F_0$  is set to two semitones above the corresponding note in the score, and the minimum  $F_0$  is set to two semitones below. This is a very useful feature for recordings that are not strictly monophonic, such as

the scenario in the SATB ensemble experiment that was discussed in Section 4.2, where each singer was closely miked, but where there was a certain amount of bleed-through from the other voices. Figure 3.2.1 shows a spectrographic representation of such a recording. The rectangle represents both the miked singer's vocal line and the settings of the minimum and maximum  $F_0$  parameters. The dotted oval and dotted circle represent the other singers.

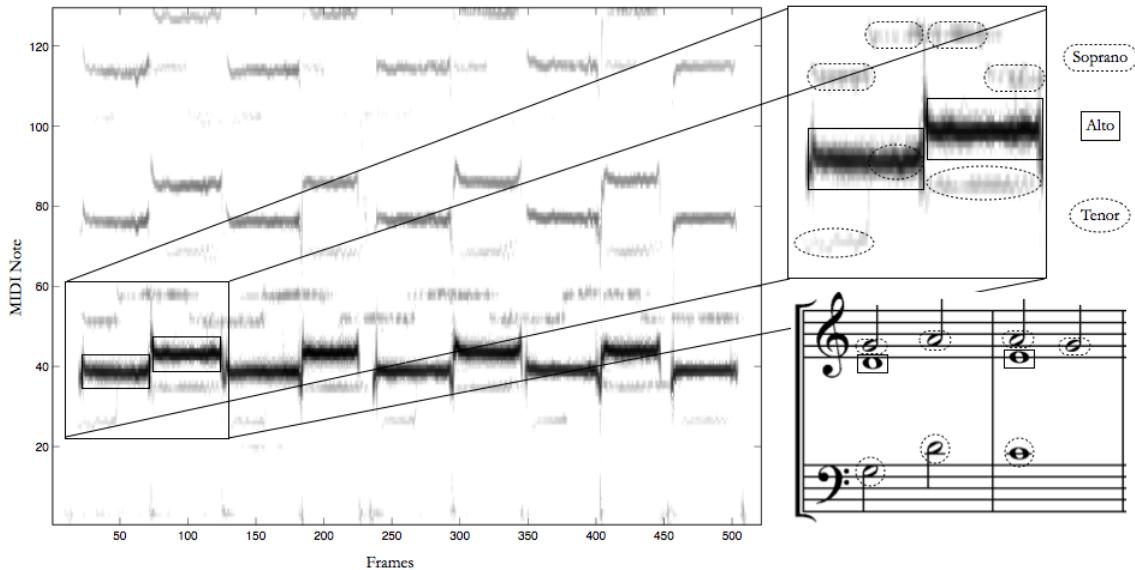


Figure 3.2.1: Example of a spectrographic representation of quasi-monophonic recording from the ensemble experiment in Section 4.2. The solid rectangle indicates the notes in the dominant voice (alto), as well as the minimum and maximum  $F_0$  parameters supplied to YIN. The dotted ovals (soprano) above and dotted circles (tenor) below indicate the locations of the other voices. The score in the bottom right corner represents the musical material that corresponds to the spectrographic representation.

The signal is analyzed in frames, whose size is determined by the window size parameter. For the experiments in this dissertation, the window size is set adaptively as the sampling rate divided by the specified minimum  $F_0$  value. For example, for a soprano singing Schubert's "Ave Maria" in Bb, the minimum  $F_0$  is 329 Hz, which results in a frame size of 3 ms and a hop size of 0.68ms. The frame size is larger for lower voices. YIN returns three

quantities for each frame: an  $F_0$  estimate, an instantaneous power estimate, and an aperiodicity measurement. YIN considers a frame's  $F_0$  estimate as accurate when the aperiodicity measurement is less than two times a specified threshold. The default value for the threshold is 0.1, but for this research we used 0.01 due to the highly periodic nature of the singing voice. The YIN algorithm sets this threshold value adaptively in relation to the minimum of the difference function.

### 3.2.2 Describing the Perceived Pitch

Section 2.3.2.4 described various experiments related to how pitch is perceived, both in general and in the singing voice in particular. The general consensus in the literature is that the perceived pitch is represented by the mean of frame-wise fundamental frequency estimates. However, there is some debate over whether it is the arithmetic or geometric mean. Following from the findings in the experiments described in Section 2.3.3, this research uses the geometric mean, although as noted by Shonle and Horan (1980), the difference is often insignificant. For example, the difference between the arithmetic and geometric mean of the signal in Figure 3.2.2 is only 0.3 cents, with an arithmetic mean of 456.96 Hz and a geometric mean of 456.88 Hz.

The simple arithmetic and geometric mean values are calculated over the entire  $F_0$  trace. A more reliable measure can be achieved with a robust mean, which only uses the central 80% of the sorted frame-wise  $F_0$  estimates. This approach is more robust because it removes outliers that could distort the calculation, though the difference is generally quite small for sung notes. For example, the difference between a regular and a robust geometric mean of the signal in Figure 3.2.2 is 0.75 cents, with a regular geometric mean of 456.89 Hz and a robust geometric mean of 457.08 Hz.

As detailed in Section 2.2, both the simple and robust means have been used in the existing vibrato literature. Following Gockel, Moore, and Carlyon (Gockel et al. 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 set at 1.41 octaves/second. The choice of the value for the threshold was informed by results reported by Prame (1994; 1997) that professional singers have an average vibrato rate of 6 Hz and a

depth of +/- 71 cents. For notes with stable vibrato, this calculation produces value close to the robust geometric mean (see Figure 3.2.2), but it is useful to observe the difference between the two calculations as a measure of the stability in each note. This is particularly true when making generalizations about intonation across different sections of a piece and different performances, as the sizes of intervals with less stable notes are a less reliable basis for generalization than intervals with stable notes.

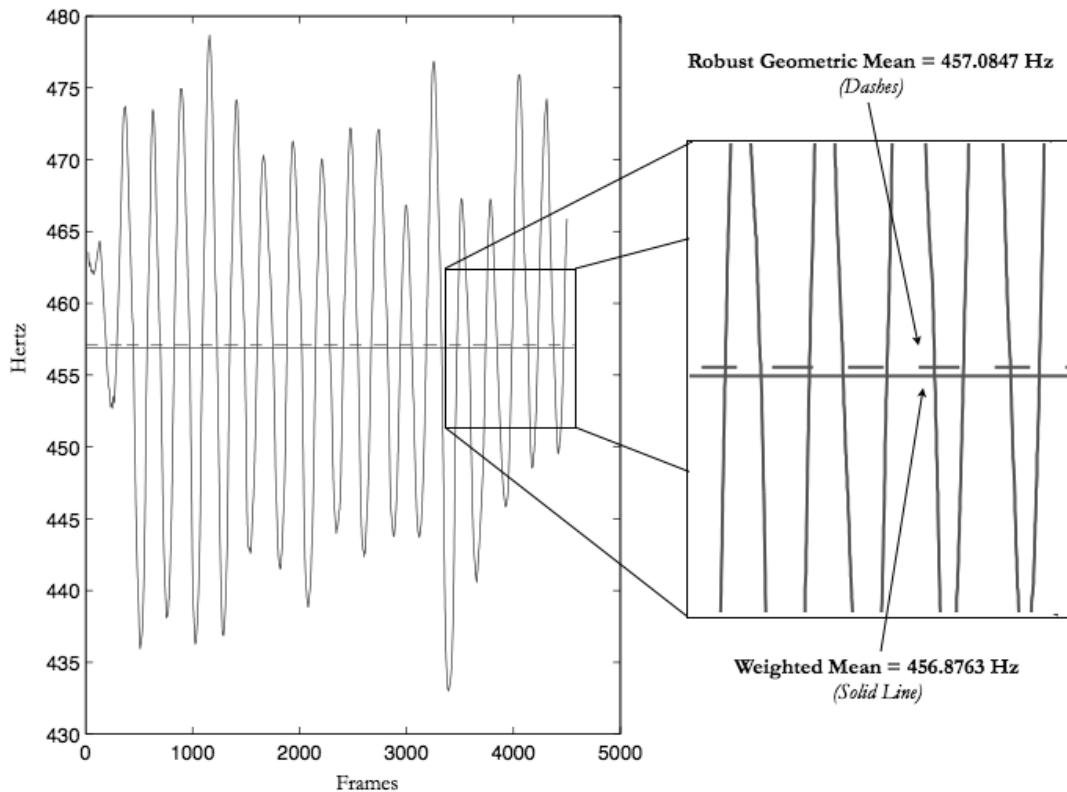


Figure 3.2.2:  $F_0$  trace of a single sung note with the robust geometric mean (dashes) and the weighted mean (lines) overlaid. The zoomed-in image on the right shows that for this note, the robust geometric mean over the duration of the note (457.0847 Hz) is only 1.87 cents different than the weighted mean (456.5913 Hz).

The calculation of a single mean over the duration of each note is most useful for horizontal (melodic) intervals, where it could be argued that a single pitch percept is generated for each note and then related to the subsequent and proceeding notes in the melody. For harmonic intervals, however, this approach does not accurately model the experience of a vertical, or

harmonic, interval between two singers. In this scenario, the singers are tuning continuously to each other, and the analysis of means taken across each note for calculating the size of a vertical interval does not take this into account. Instead, the size of vertical intervals is calculated by measuring the interval size for each frame and then by taking the mean across the series of vertical calculations. Figure 3.2.3 shows the difference in the calculation for an example vertical interval.

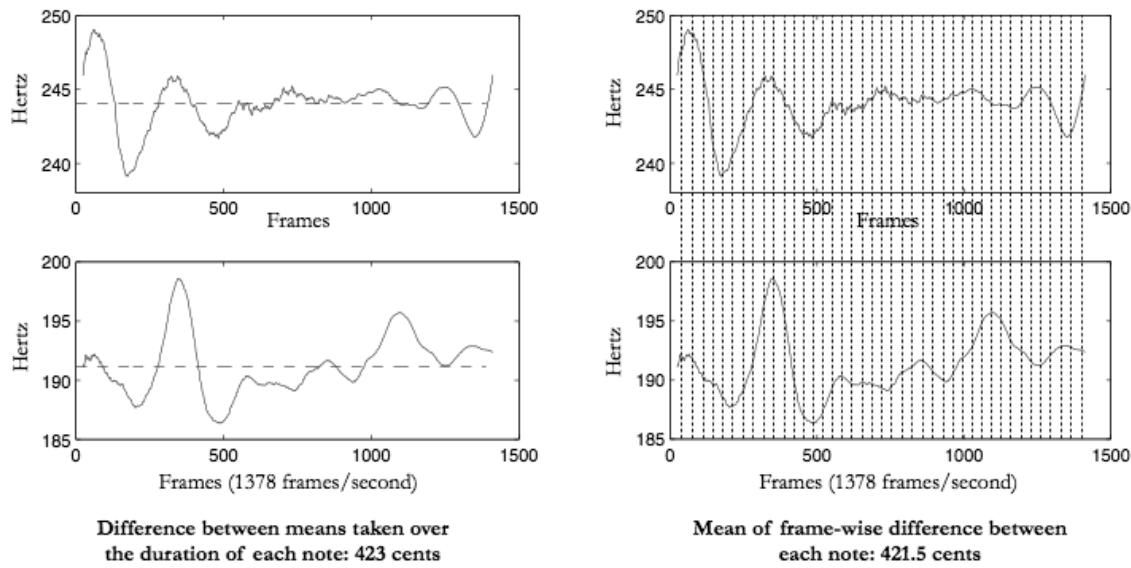


Figure 3.2.3: Example of how vertical interval size is calculated. The plots on the left show the use of a melodic interval approach, where the robust geometric mean across each note's frame-wise  $F_0$  estimates is used to calculate the vertical interval size (423 cents). The plots on the right show the method used for this research, where frame-wise vertical interval size calculations are made, and then the robust geometric mean is taken across these calculations (421.5 cents).

### 3.2.3 Evolution of Fundamental Frequency

The means described in Section 3.2.2 provides only a single value for each note and throw away information about how  $F_0$  changes over the duration of the note. This section discusses how the evolution of  $F_0$  can be described by using discrete cosine transform (DCT) coefficients.

A moving average is obtained by taking the mean of a shifting subset of the  $F_0$  trace and is a type of low-pass filter. The moving average provides information about the slower moving trends in the  $F_0$ , such as whether the frequency is increasing or decreasing, by smoothing out fast moving fluctuations like vibrato. Figure 3.2.4–Figure 3.2.7 show the moving average of four different window sizes (50, 100, 150, and 200 ms) on four notes of different lengths (4.0 s, 2.1 s, 0.77 s, and 0.48 s).

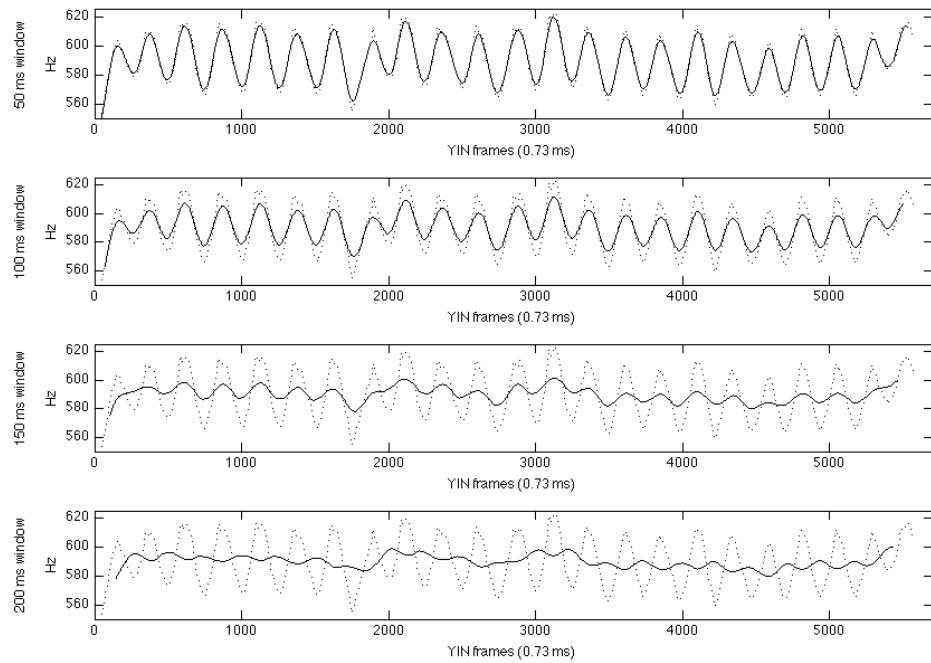


Figure 3.2.4: Moving averages for a long note (4.0 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms). The original signal is represented with a dotted line.

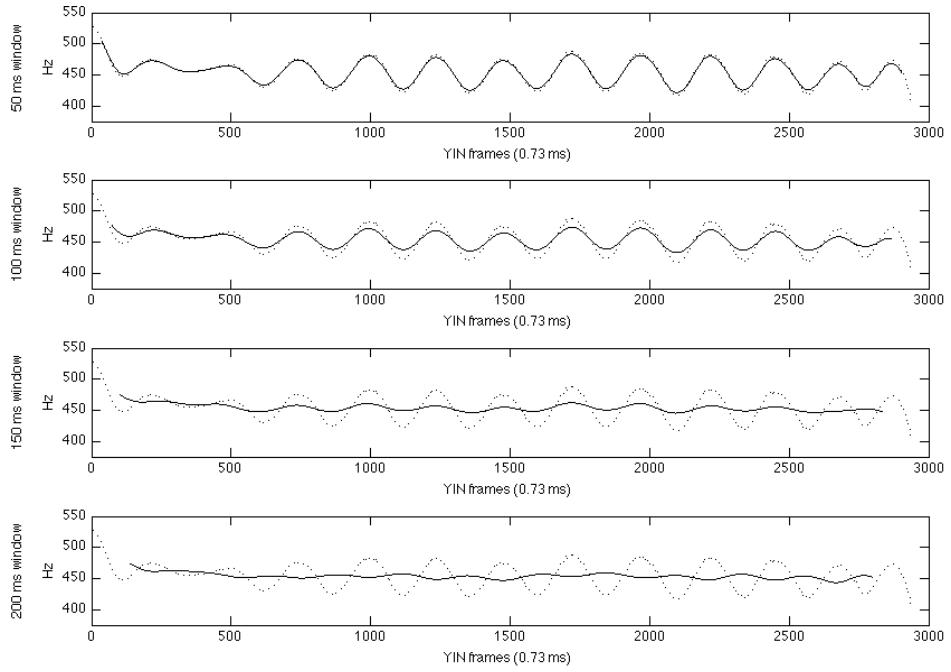


Figure 3.2.5: Moving averages for a medium long note (2.1 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms). The original signal is represented with a dotted line.

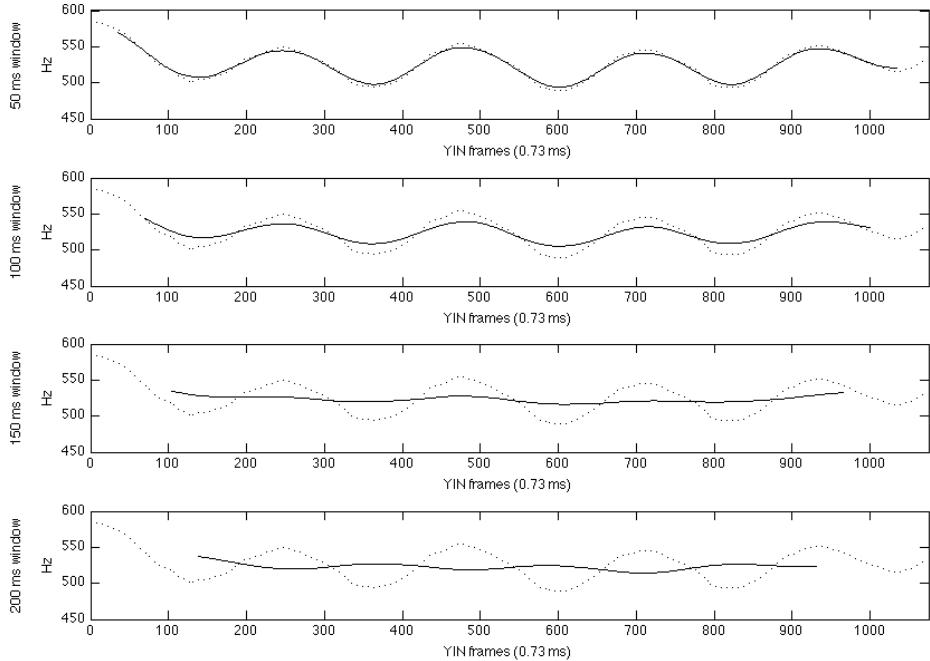


Figure 3.2.6: Moving averages for a medium short note (0.77 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms). The original signal is represented with a dotted line.

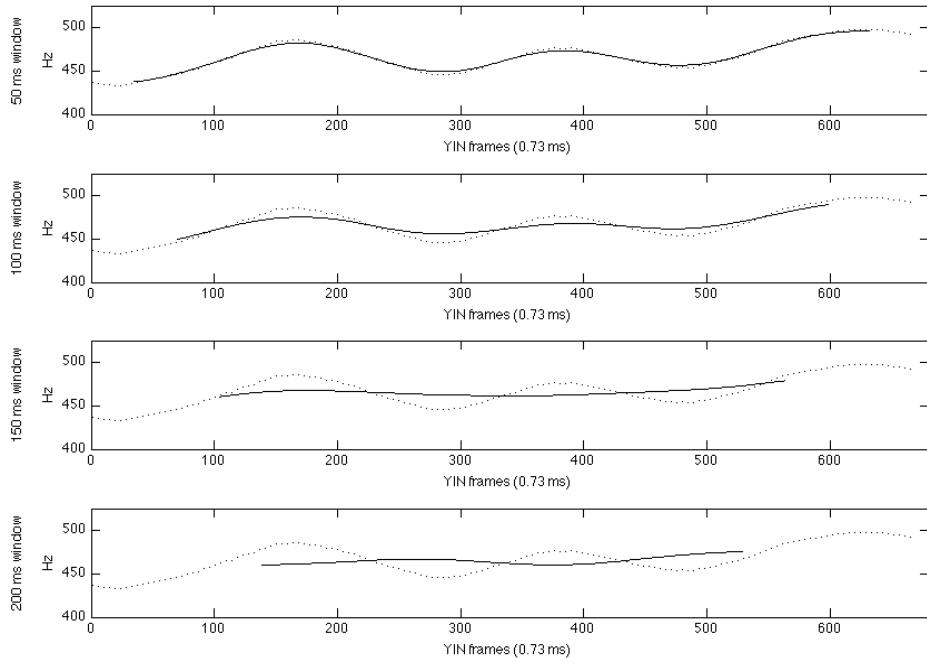


Figure 3.2.7: Moving averages for a short note (0.48 s) with four different window sizes (50 ms, 100 ms, 150 ms, and 200 ms). The original signal is represented with a dotted line.

For the purposes of generalizing the evolution of  $F_0$  over the duration of the note, a larger window size is preferable. With a larger window size, more general trends about whether the  $F_0$  trace is generally moving higher or lower can be observed. If the window size is too small, the vibrato rate will impact subsequent calculations about slope and curvature. In all of the examples, the moving average calculations with 50 and 100 ms window sizes closely follow the vibrato. The moving average signals calculated with 150 and 200 ms window sizes are much smoother, though some artefacts of the vibrato remain in certain cases. While window sizes larger than 200 ms would further smooth the signal, they are not practical for this data since they would start to exceed the size of half of the shortest signal (470 ms) and could smooth more than just the vibrato. This happens to coincide with the findings in the literature that the vibrato range is 5–7Hz.

The moving average generates a string of values, which can be used to visually observe trends. However, additional calculations are needed to generate slope and curvature values that can be compared between different notes. One way of calculating the slope is to take the difference between the first and the last value in the moving average signal and dividing

it by its the length. The problem with this approach is that it is highly sensitive to any noise in the first or last element. A more robust option is to calculate the slope over the duration of the signal rather than from just two points. One way of doing this is to fit a second-order polynomial to the moving average signal. In a first-order polynomial, the constant term is the mean of the signal, and the coefficient on the first order term is the slope. In a second-order polynomial, however, the bases are not orthogonal, so a separation between the slope and curvature is not guaranteed. A more robust approach is the discrete cosine transform (DCT), where the input signal is represented as a sum of scaled cosines and the basis functions are orthogonal (Jain 1989).

The DCT is similar to the discrete Fourier transform (DFT) but differs in that it is real, whereas the DFT is complex. Thus the DCT returns a single coefficient for each frequency with a fixed phase, whereas the DFT returns two coefficients (amplitude and phase) for each frequency. The fixed phase simplifies the comparisons between values. The MATLAB implementation of the DCT used in this research uses Equation 1 for calculating the coefficients ( $y(k)$ ), where  $N$  is the length of the signal  $x$  and  $k$  is the number of coefficients.

$$y(k) = \omega(k) \sum_{n=0}^{N-1} x(n) \cos \frac{k(2n+1)\pi}{2N} \quad k = 0, 1, 2 \quad (1)$$

$$\text{where } \omega(k) = \begin{cases} \frac{1}{\sqrt{N}} & k = 0 \\ \frac{\sqrt{2}}{N} & 1 \leq k \leq 2 \end{cases}$$

As demonstrated in Equation 2, the 0<sup>th</sup> coefficient returned by the DCT is the mean of the signal over the square root of  $N$  (the number of samples in the signal). The 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficients can be used to capture the broad contours of a signal, in this research the F<sub>0</sub> trace, that relate to slope and curvature while ignoring fine details, such as vibrato. In order to transform the DCT coefficients so that they are independent of signal length, the 0<sup>th</sup> coefficient is divided by  $N^{1/2}$ , the 1<sup>st</sup> by  $N^{3/2}$ , and the 2<sup>nd</sup> by  $N^{5/2}$ . After this scaling, the 1<sup>st</sup> coefficient is approximately in units of cents/second, and the 2<sup>nd</sup> coefficient is approximately

in units of cents/second<sup>2</sup>. Also, the 1<sup>st</sup> coefficient represents negative slope, so its sign is reversed so that it describes positive slope.

$$y(0) = \omega(0) \sum_{n=0}^{N-1} n \cos \frac{0(2n+1)\pi}{2N} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} n \quad (2)$$

The 1<sup>st</sup> DCT coefficient approximates the slope of the evolution of F<sub>0</sub>. The slope provides information about whether the singers are gliding up or down or staying relatively stable. The amount and direction of movement (if any) depends on the sign and value of the coefficient. The 2<sup>nd</sup> DCT coefficient approximates the curvature of the evolution of F<sub>0</sub>. The curvature, once the slope has been subtracted, indicates the amount that the F<sub>0</sub> trace is higher or lower in the middle than at the two ends of the time period analyzed. Figure 3.2.8 shows the DCT coefficients for seven simple signals: a straight line, a diagonal line (up and down), a parabolic curve (up and down), and a scooping line (up and down). All of the signals have a mean of 20, so the 0<sup>th</sup> coefficient remains the same. For the flat line, only the 0<sup>th</sup> coefficient has a nonzero value since the signal can be completely described by its mean. For the diagonal line, the 1<sup>st</sup> coefficient has a much larger value than any other (except for the 0<sup>th</sup>), which demonstrates the relationship between the 1<sup>st</sup> coefficient and the slope of the signal. For the parabolic curve, the 2<sup>nd</sup> coefficient has the largest value (again except for the 0<sup>th</sup>), demonstrating that the 2<sup>nd</sup> coefficient can be taken to approximate the curvature of the signal. This relationship is more approximate than between the 1<sup>st</sup> coefficient and the slope, which can be seen in the greater values of the 4<sup>th</sup> and 6<sup>th</sup> DCT coefficients for the parabolic curve. The spread of energy across different coefficients occurs because the signals are generated from second-order polynomials rather than cosines.

Table 3.2.1 and Table 3.2.2 show simple signals with different curves and slopes: five straight lines (a-e), four parabolic curves (f-i), and four other curves (j-k). The signals in Table 3.2.1 are lines with a moving average that closely follows the original signal. In Table 3.2.2, the signals in Table 3.2.1 have been modified with the addition of a sinusoid. For these signals, the moving average follows the midpoint of the sinusoids. For the signals in Table 3.2.1, the 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficients from the original signals and moving average are virtually identical. For the sinusoidal signals in Table 3.2.2, there is a small discrepancy between the 1<sup>st</sup> DCT coefficients calculated on the original signal versus the moving average. The signals in

Table 3.2.2 are derived from those in Table 3.2.1, so the 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficient values should be the same.

The difference between the DCT's coefficients for the original signal and the moving average are likely due to the DCT's sensitivity to the phase of the sinusoidal signal, an issue that is avoided by the moving average's smoothing. With the analysis of more complex real-world signals, such as F<sub>0</sub> traces of a sung note, this issue becomes even more significant, particularly when only a segment of the signal is studied. The DCT coefficients could vary greatly if the starting or ending of the signal moves by some fraction of the vibrato's cycle. This is demonstrated in Figure 3.2.9, which shows the 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficients calculated from the F<sub>0</sub> trace of the 4 s note in Figure 3.2.4 and a moving average of the F<sub>0</sub> trace with a window size of 200 ms. These DCT coefficients were calculated for windows of the same length, shifted one sample at a time. The variance in the DCT values from the F<sub>0</sub> trace demonstrates the sensitivity of the DCT to the phase of the vibrato. The greater stability values from the moving average show that although the moving average is not completely impervious to the phase of the vibrato, its effect is greatly minimized. Using a larger window could further minimize the effect of the vibrato's phase, but, as mentioned above, this runs the risk of smoothing more than just the vibrato. For this reason, the DCT coefficients for the experiments are also calculated on a moving average of the F<sub>0</sub> trace of each note as an alternative method of analysis.

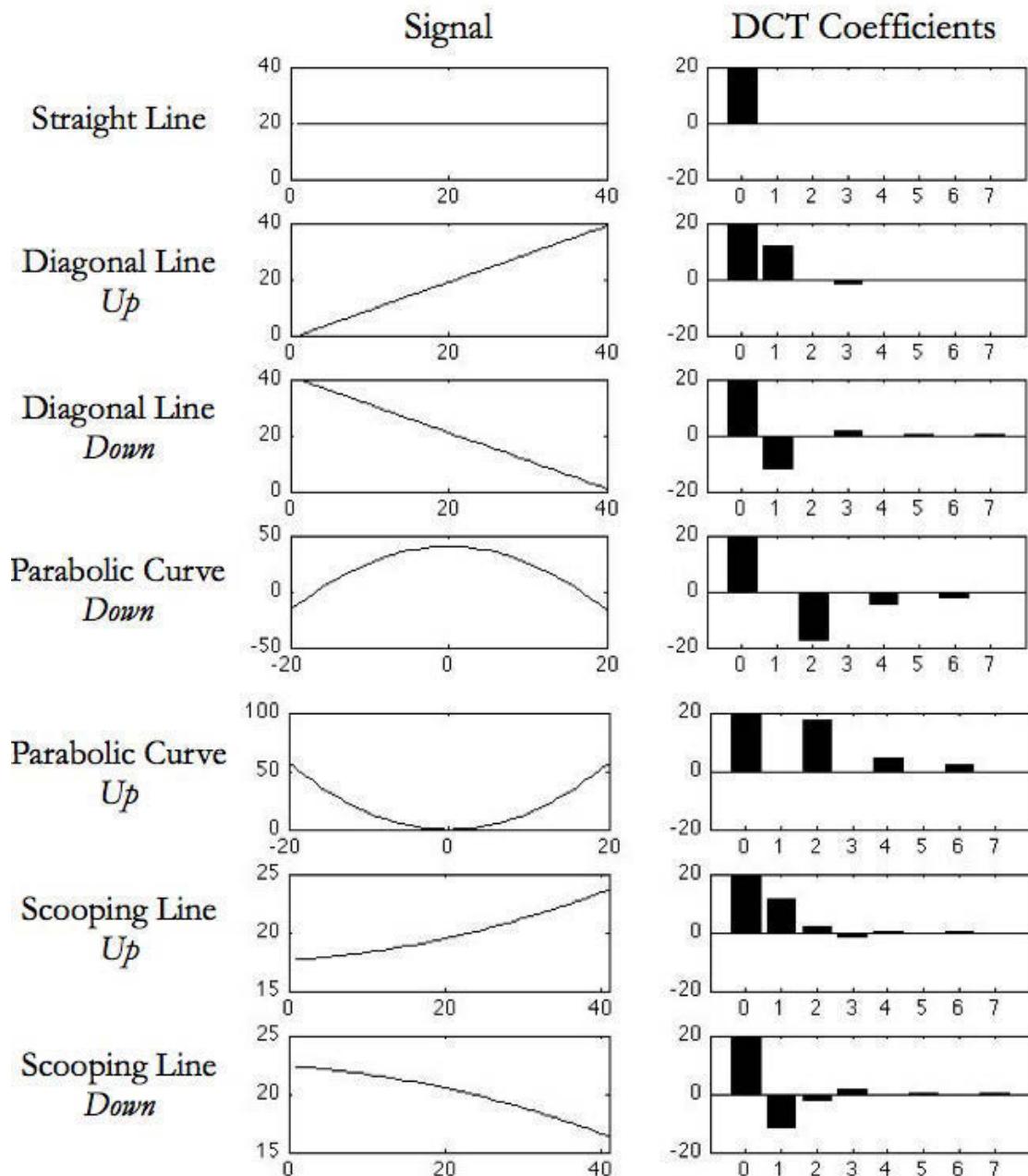


Figure 3.2.8: Examples of DCT coefficients for simple signals. The plots on the left are the original signals (straight line, diagonal line, parabolic curve, and scooping line), and the bar graphs on the right are the values for the DCT coefficients 0–7. The 0<sup>th</sup> coefficient is the mean, the 1<sup>st</sup> coefficient approximates slope, and the 2<sup>nd</sup> coefficient approximates curvature.

Signal	Slope/ Curvature	Signal	Slope/ Curvature
(a)	0 (0)		
(b)	0.0573 (0.057312)	(c)	0.1146 (0.11462)
(d)	-0.0573 (-0.057312)	(e)	-0.1146 (-0.11462)
(f)	0 (0)	(g)	0 (0)
(h)	0.3941 (0.3941)	(i)	-0.7897 (0.7897)
(j)	0 (0)	(k)	0 (0)
(l)	-0.3941 (-0.3941)	(m)	-0.7897 (-0.7897)
	0.057314 (0.057312)		0.11463 (0.11462)
	0.1971 (0.1971)		0.3941 (0.3941)
	-0.057314 (-0.057312)		-0.11463 (-0.11462)
	-0.1971 (-0.1971)		-0.3941 (-0.3941)

Table 3.2.1: This table shows thirteen simple signals with different curves and slopes—five straight lines (a-e), four parabolic curves (f-i), and four other curves (j-k)—and the values for the 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficients for each signal. The 0<sup>th</sup> coefficient (i.e., the mean) for all of the signals is 0. The DCT coefficients were calculated on the original signal and the moving average (indicated in parentheses), which is plotted with a dashed line but not clearly visible because it closely follows the original signal for all of the examples.

Signal	Slope/ Curvature	Signal	Slope/ Curvature
(a)	-0.0017 (0) 0 (0)		
(b)	0.0556 (0.057312) 0 (0)		
(d)	-0.059 (-0.057312) 0 (0)		
(f)	-0.0017 (0) 0.3941 (0.3941)		
(h)	-0.0017 (0) -0.3941 (-0.3941)		
(j)	0.0556 (0.057312) 0.1971 (0.1971)		
(l)	-0.059 (-0.057312) -0.1971 (-0.1971)		
(c)	0.1129 (0.11462) 0 (0)		
(e)	-0.1163 (-0.11462) 0 (0)		
(g)	-0.0017 (0) 0.7897 (0.7897)		
(i)	-0.0017 (0) -0.7897 (-0.7897)		
(k)	0.11292 (0.11462) 0.3941 (0.3941)		
(m)	-0.11634 (-0.11462) -0.3941 (-0.3941)		

Table 3.2.2: This table shows thirteen simple signals that mirror the signals in Table 3.2.1 with the addition of sinusoids, as well as the values for the 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficients for each signal. The 0<sup>th</sup> coefficient (i.e., the mean) for all of the signals is 0. The DCT coefficients were calculated on the original signal and the moving average (indicated in parentheses), which is plotted with a dashed line.

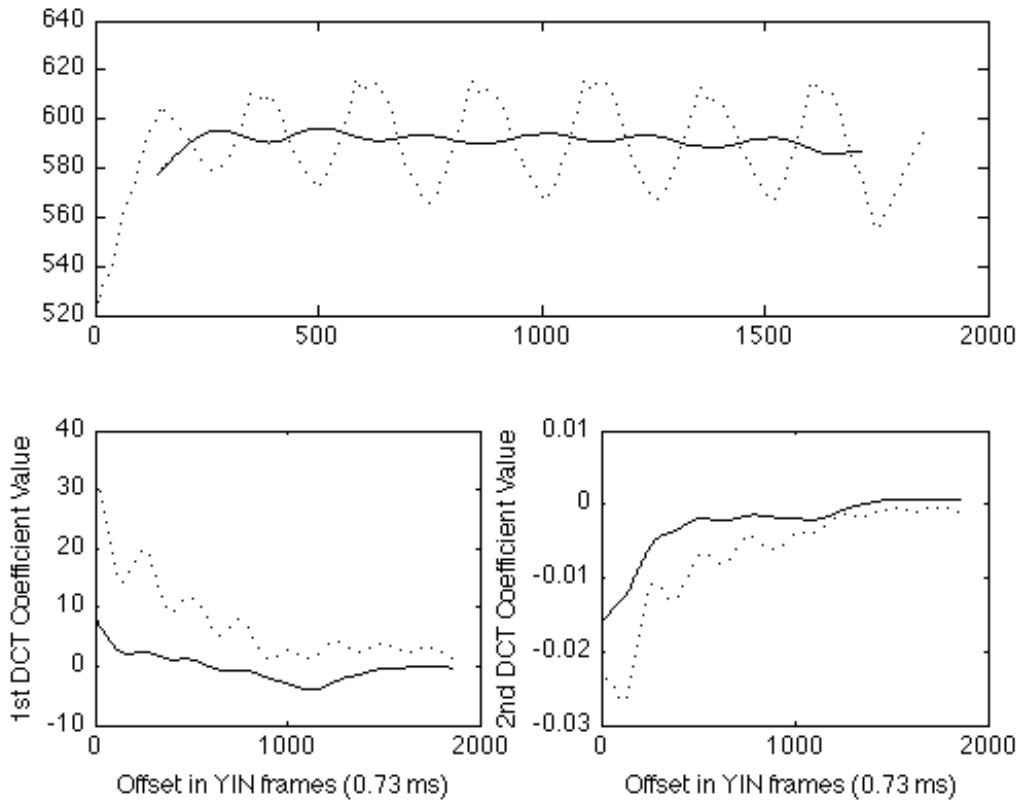


Figure 3.2.9: Comparison of the 1<sup>st</sup> (bottom left) and 2<sup>nd</sup> (bottom right) DCT coefficients calculated from the  $F_0$  trace of the 4 s note in Figure 3.2.4 (dotted line) and those calculated from a moving average of the  $F_0$  trace with a window size of 200 ms (solid line). The 500 DCT coefficient calculations were made on windows of the same length ( $N=500$  or 5479-500, which is 4979 samples), shifted one sample at a time. The variability in the dotted lines demonstrates the sensitivity of the DCT to the phase of the vibrato, which is mitigated by the moving average.

Figure 3.2.10–Figure 3.2.13 show reconstructions of the DCT coefficients over three sections of the four notes in Figure 3.2.4–Figure 3.2.9. The DCT reconstructions were done with the inverse DCT, as defined in equation 3–6, with the DCT coefficients calculated on a 200 ms moving average of the original signal. Equation 3 shows the generalized equation, Equation 4 shows the equation for the 0<sup>th</sup> DCT reconstruction, Equation 5 shows the 0<sup>th</sup>+1<sup>st</sup> DCT coefficients, and Equation 6 shows the 0<sup>th</sup>+1<sup>st</sup>+2<sup>nd</sup> DCT coefficients.

$$x(n) = \sum_{k=0}^N \omega(k) y(k) \cos\left(\frac{\pi(2n+1)k}{2N}\right) \quad n = 1, 2, \dots, N \quad (3)$$

where  $\omega(k) = \begin{cases} 1 & k = 0 \\ \frac{1}{\sqrt{N}} & 1 \leq k \leq 2 \\ \frac{2}{N} & 1 \leq k \leq 2 \end{cases}$

$$\begin{aligned} x_0(n) &= \sum_{k=0}^N y(0) \cos\left(\frac{\pi(2n+1)0}{2N}\right) \quad n = 1, 2, \dots, N \\ x_0(n) &= \frac{y(0)}{\sqrt{N}} \quad n = 1, 2, \dots, N \end{aligned} \quad (4)$$

$$x_1(n) = \frac{y(0)}{\sqrt{N}} + \sqrt{\frac{2}{N}} y(1) \cos\left(\frac{\pi(2n+1)}{2N}\right) \quad n = 1, 2, \dots, N \quad (5)$$

$$x_2(n) = \frac{y(0)}{\sqrt{N}} + \sqrt{\frac{2}{N}} y(1) \cos\left(\frac{\pi(2n+1)}{2N}\right) + \sqrt{\frac{2}{N}} y(2) \cos\left(\frac{\pi(2n+1)}{N}\right) \quad n = 1, 2, \dots, N \quad (6)$$

In Figure 3.2.10–Figure 3.2.13, the top subplots show the original signal with a reconstruction of the signal from the inverse discrete cosine transform using only the 0<sup>th</sup> (mean) coefficient overlaid, the middle plot shows the signal with the 0<sup>th</sup> (mean) + the 1<sup>st</sup> (slope) coefficients overlaid, and the bottom plots show the signal with the 0<sup>th</sup> (mean) + the 1<sup>st</sup> (slope) + the 2<sup>nd</sup> (curvature) coefficients overlaid.

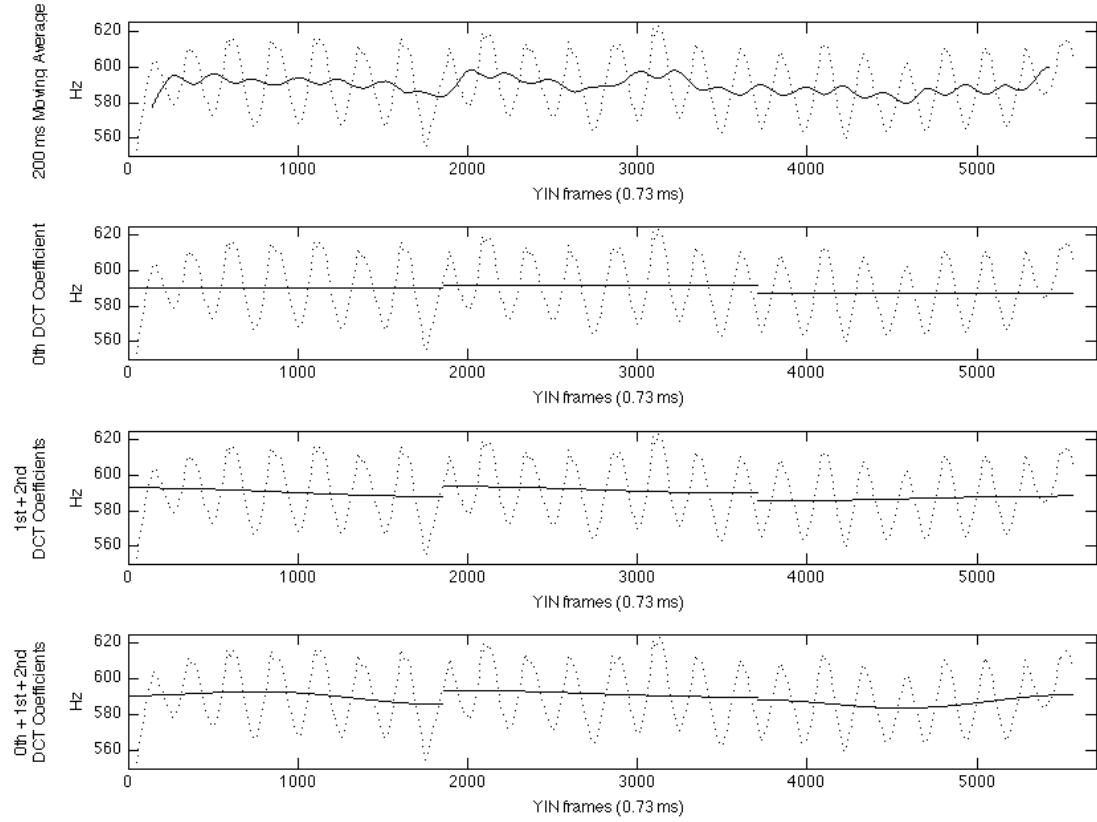


Figure 3.2.10: Discrete cosine transform on a 200 ms moving average of the  $F_0$  trace of a long note (4.0 s). All of the plots represent the original  $F_0$  trace with a dotted line. The top plot shows the 200 ms moving average. The second plot shows the reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) DCT coefficient calculated over each 3<sup>rd</sup> of the note. The third plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) DCT coefficients and the bottom plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) + 2<sup>nd</sup> (curvature) DCT coefficients.

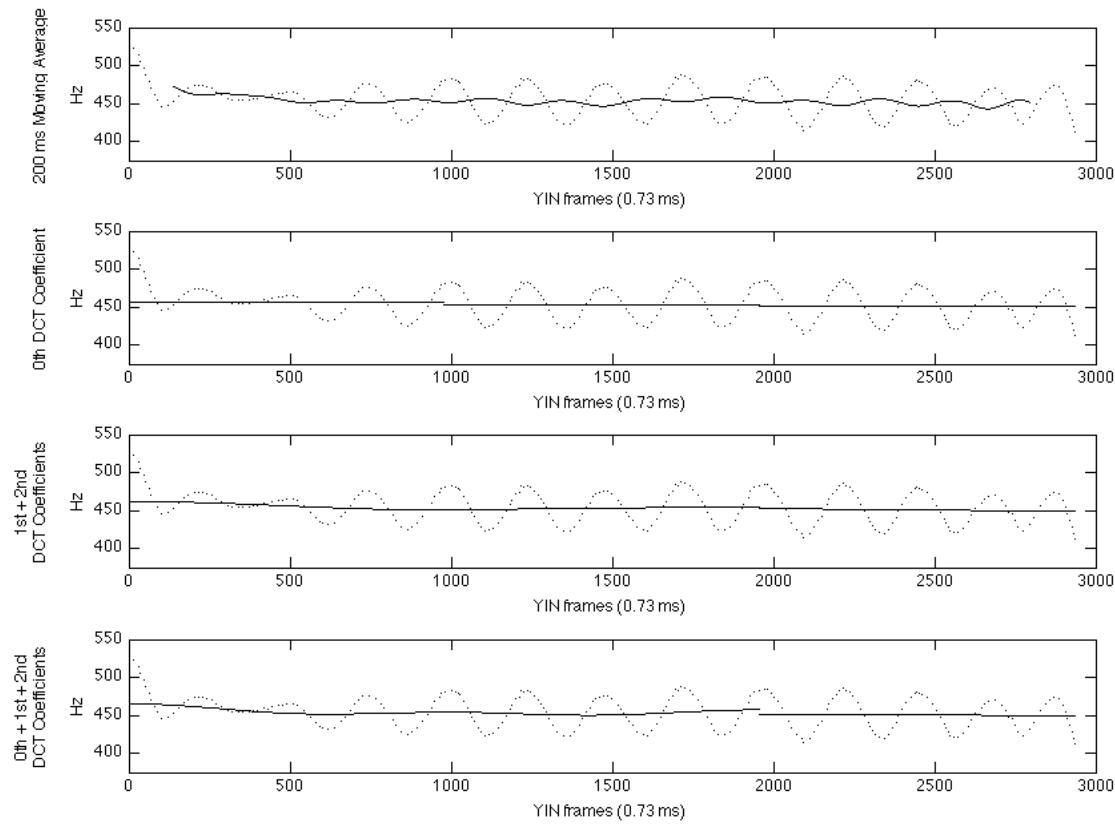


Figure 3.2.11: Discrete cosine transform on a 200 ms moving average of the  $F_0$  trace of a long note (2.1 s). All of the plots represent the original  $F_0$  trace with a dotted line. The top plot shows the 200 ms moving average. The second plot shows the reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) DCT coefficient calculated over each 3<sup>rd</sup> of the note. The third plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) DCT coefficients, and the bottom plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) + 2<sup>nd</sup> (curvature) DCT coefficients.

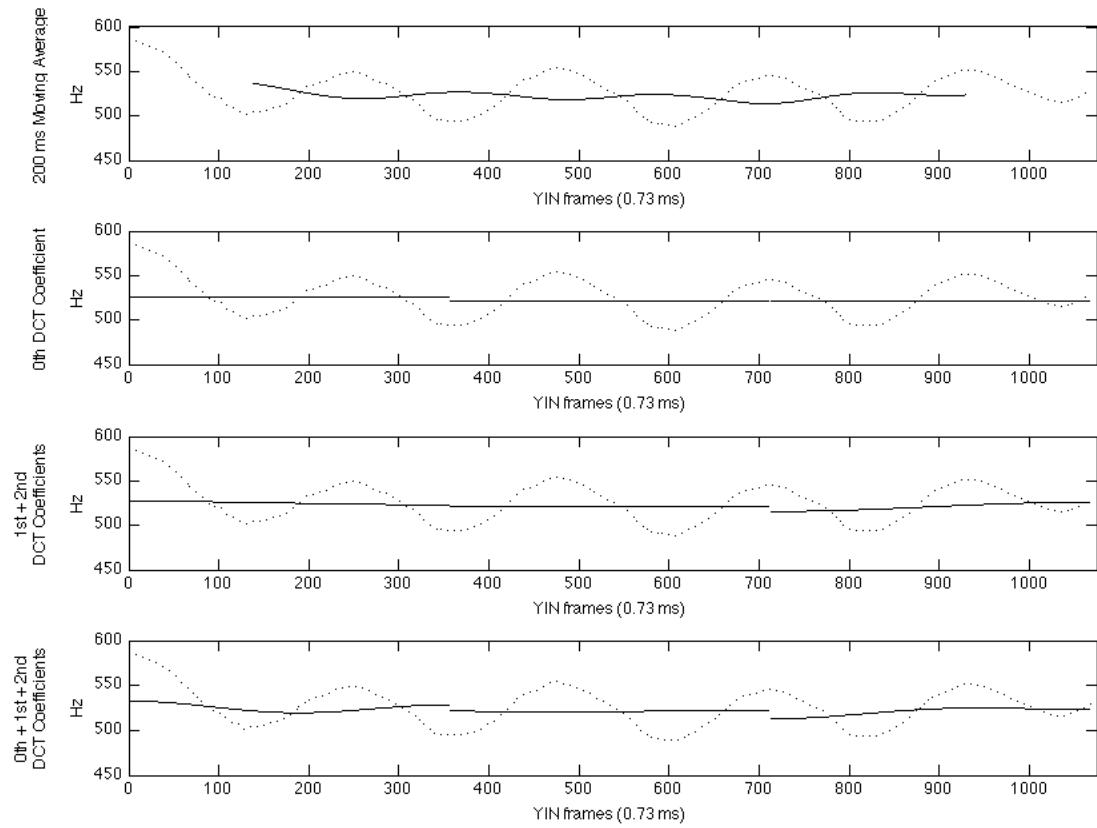


Figure 3.2.12: Discrete cosine transform on a 200 ms moving average of the  $F_0$  trace of a medium short note (0.77 s). All of the plots represent the original  $F_0$  trace with a dotted line. The top plot shows the 200 ms moving average. The second plot shows the reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) DCT coefficient calculated over each 3<sup>rd</sup> of the note. The third plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) DCT coefficients, and the bottom plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) + 2<sup>nd</sup> (curvature) DCT coefficients.

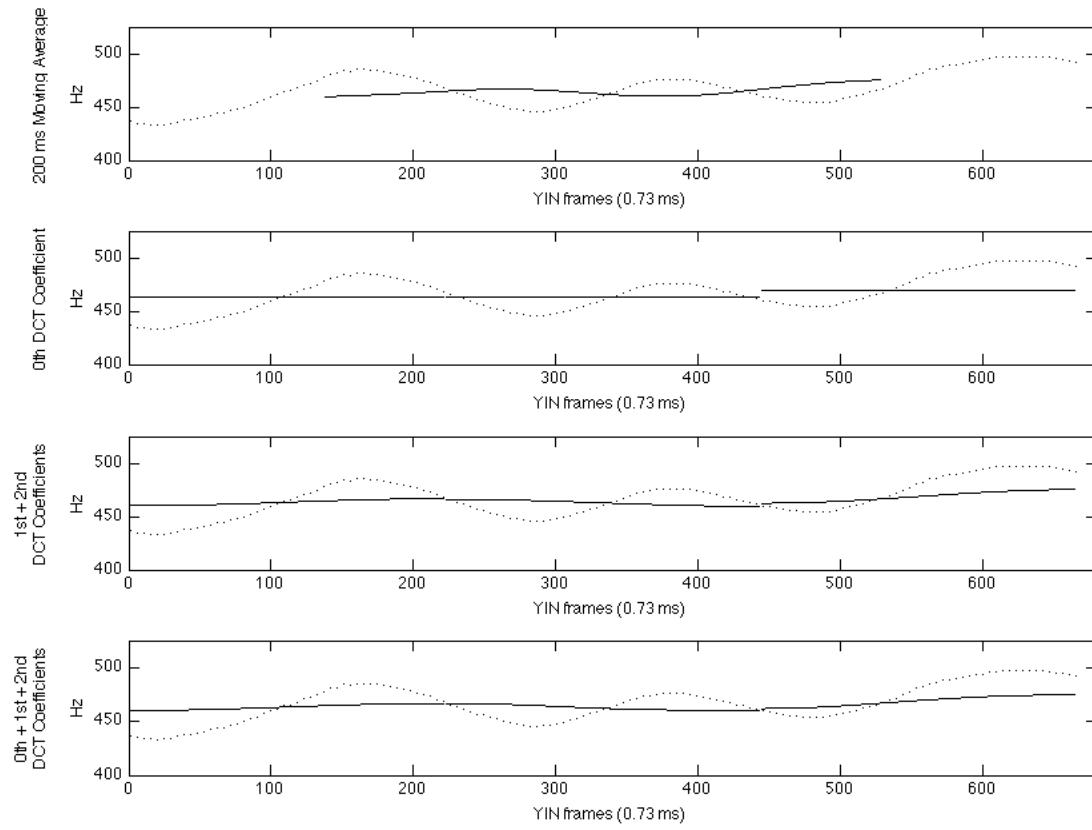


Figure 3.2.13: Discrete cosine transform on a 200 ms moving average of the  $F_0$  trace of a short note (0.48 s). All of the plots represent the original  $F_0$  trace with a dotted line. The top plot shows the 200 ms moving average. The second plot shows the reconstruction of the original  $F_0$  trace with the 0<sup>th</sup> (mean) DCT coefficient calculated over each 3<sup>rd</sup> of the note. The third plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) DCT coefficients, and the bottom plot shows a reconstruction of the  $F_0$  trace with the 0<sup>th</sup> (mean) + 1<sup>st</sup> (slope) + 2<sup>nd</sup> (curvature) DCT coefficients.

### 3.2.4 Summary

Section 3.2.1 explains how the YIN algorithm was used to extract  $F_0$  estimates from both the recordings made for the experiments in Chapter 4. YIN was used on both the monophonic recordings from solo singing experiment (Section 4.1) and the quasi-polyphonic recordings produced by close miking the individual singers in the ensemble singing experiment (Section 4.2). The descriptors detailed in Sections 3.2.2 and 3.2.3 used the  $F_0$  estimates to provide a

good summary of pitch-related characteristics of sung notes. The descriptors in Section 3.2.2 provide an estimation of the perceived pitch of melodic and vertical intervals, and the descriptors in Section 3.2.3 provide information about how a note's  $F_0$  evolves, or changes, over time by calculating its slope and curvature. The descriptors from both sections were used to characterize the intonation-related details of each note in the experiments in Section 4.1, while only the descriptors from Section 3.2.2 is used in Section 4.2. In Section 4.1, the evolution of  $F_0$  is described by calculating the DCT on the end of the first note in each melodic interval on both the original  $F_0$  trace and the  $F_0$  trace with a 200 ms moving average applied to it. The end of the note is defined as the last 250 ms of the  $F_0$  trace or the last 150 ms of the  $F_0$  trace with a 200 ms moving average of the steady-state portion of the note. The points of convergence between the two methods of calculations are considered to be a reliable indication of the evolution of  $F_0$ . Such generalized descriptors are useful because they allow comparisons not only across different notes in the same performance, but also across notes in different performances by the same singer and across different singers.

(This page intentionally left blank)

## Chapter 4 Intonation Experiments

This chapter presents two sets of experiments on intonation practices: one focused on intonation in solo singing and the other on intonation in SATB ensembles with one voice per part. In the experiment in Section 4.1, two groups of singers, one made up of undergraduate vocal majors and the other made up of professional singers, performed Schubert’s “Ave Maria” three times *a cappella* and three times with accompaniment. The melodic semitones and whole tones in the recordings were assessed in regard to interval size, as well as the slope and curvature of  $F_0$  at the end of the first note of each melodic interval. In the experiment described in Section 4.2, three different ensembles were recorded: one semi-professional ensemble that sang without a conductor and two conducted professional ensembles. The two professional ensembles performed two sets of short exercises (Parts One and Two) designed so that melodic semitones and whole tones occurred in different contexts. All of the ensembles performed a repeated chord progression by Giambattista Bendedetti (Part Three) and “Es ist ein Ros entsprungen” by Michael Praetorius (Part Four). Both melodic and vertical intervals were studied: melodic semitones in Parts One and Four; melodic whole tones in Parts Two, Three, and Four; and vertical intervals in all four parts. Section 4.3 draws a connection between the results in both sets of experiments and posits some interpretations about intonation in singing that can be made from the data.

### 4.1 Intonation in Solo Singing

The experiment described in this section explores whether there is a relationship between melodic interval tuning in solo singing and the context in which the interval occurs. Following Prame’s work (1994; 1997), Schubert’s “Ave Maria” was used for the experimental material since it allowed for an exploration of commonalities of intonation tendencies in a well-known piece. There were twelve subjects for the experiment: six undergraduate vocal majors from McGill and six professional singers from the Montreal area. Each singer performed the first verse three times *a cappella* and three times with recorded accompaniment.

Where Prame limited his analysis to the 25 longest notes in the piece, this experiment looks at all of the semitones and whole tones between notes whose durations were greater than a

thirty-second note. This limitation was imposed due to the instability of the pitch in shorter notes. Using the algorithm described in Section 3.1, reliable onset, offset, and pitch estimates could be obtained for the non-ornamental notes in the piece. Accurate annotation of the ornamental notes, however, required manual intervention due to the variability in the way in which the ornaments were performed. As described in Section 3.2, the intonation is described by both the weighted mean across the frame-wise fundamental frequency estimates of each note and the 1<sup>st</sup> and 2<sup>nd</sup> DCT coefficients, which are used to approximate the slope and curvature of the last 250 ms of the first note in each melodic interval.

The data analysis in this section examines the sizes of the semitone and whole tone intervals, specifically whether the context in which the intervals occur influences the intonation tendencies of the singer. The analysis focuses on interval size, rather than absolute pitch, to remove the influence of drift. For both types of intervals, the difference between the ascending and descending intervals is evaluated. The semitone analysis also compares the semitones between the leading tone and the tonic to the other semitones in the piece (Figure 4.1.1). This allows for the assessment of a commonly held belief, rooted in Pythagorean tuning theory, that ascending leading-tones are sung sharp, which would make the leading tone semitone smaller relative to other semitones (Friberg et al. 2006). The whole tone analysis looks at whether the movement towards or away from stable notes influences the intonation tendencies (Figure 4.1.2). The data analysis also considers intonational consistencies both within each performer's *a cappella* and accompanied renditions, as well as across performers.

The musical score consists of six staves of music in G major, 2/4 time. The vocal line is in soprano range. Various semitone categories are marked with specific symbols:

- A-Bb ascending interval**: Circled with a solid line.
- Other ascending semitones**: Circled with a dashed line.
- LT indicates a leading tone**: A label with a circled 'LT' symbol.
- Bb-A descending interval**: Circled with a dashed line.
- Other descending semitones**: Circled with a solid line.

Text below the score:

○ A-Bb ascending interval    □ Other ascending semitones  
 LT indicates a leading tone  
 ○ Bb-A descending interval    □ Other descending semitones

Figure 4.1.1: Schubert’s “Ave Maria” with analyzed semitone categories marked.

The musical score consists of six staves of music in G major, 2/4 time. The vocal line is in soprano range. Various whole tone categories are marked with specific symbols:

- Ascending chord tone to non-chord tone whole tone**: Circled with a solid line.
- Descending chord tone to non-chord tone whole tone**: Circled with a dashed line.
- Ascending non-chord to chord tone whole tone**: Circled with a dashed line.
- Descending non-chord to chord tone whole tone**: Circled with a solid line.
- Ascending chord tone to chord tone whole tone**: Circled with a solid line.
- Descending chord tone to chord tone whole tone**: Circled with a dashed line.

Text below the score:

○ Ascending chord tone to non-chord tone whole tone    □ Descending chord tone to non-chord tone whole tone  
 □ Ascending non-chord to chord tone whole tone    □ Descending non-chord to chord tone whole tone  
 ○ Ascending chord tone to chord tone whole tone    □ Descending chord tone to chord tone whole tone

Figure 4.1.2: Schubert’s “Ave Maria” with analyzed whole tone categories marked.

## **4.1.1 Method**

### **4.1.1.1 Participants**

The first group of participants consisted of six undergraduate soprano vocal majors from McGill University, hereafter referred to as the non-professional group, who had completed an average of 2 years ( $SD = .63$ ) of full-time course work in the Bachelor of Music degree program. The participants had a mean age of 20.2 years ( $SD = 2.13$ ) 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 ( $SD = 0.7$ ).

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 5.2 years ( $SD = 3.2$ ), with a current average daily practice time of 1.5 hours ( $SD = 0.5$ ). None of the singers in either group possessed absolute pitch.

### **4.1.1.2 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 (CIRMMT). The lab had low noise, reflections, and reverberation time (ITU-standard). The microphones were run through an RME Micstasy 8 channel microphone preamplifier and an RME Madi Bridge into a Mac Pro computer for recording.

### **4.1.1.3 Procedure**

In the experiment, each of the singers 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 recorded 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 so that they could hear themselves.

The intonation-related data analysis was extracted using the methods described in Chapter 3 and was checked manually by two trained musicians using Audacity to correct any errors made by the algorithm. Three measurements were extracted for each interval: one for interval size and one for slope and curvature of the end of the first note. The interval size was calculated by taking the difference between the perceived pitch estimates for each note making up the interval. The perceived pitch estimates were made by taking the weighted mean over the  $F_0$  estimates for each note, as described in Section 3.2. Two approaches were used for obtaining slope and curvature estimates. In the first, the discrete cosine transform (DCT) was run on the last 250 ms of the  $F_0$  trace of the first note in the interval. In the second, the DCT was run on the  $F_0$  trace smoothed by results of applying a 200 ms moving average. The moving average was applied to minimize the influence of vibrato on the calculation. For the second approach, the last 150 ms of the smoothed signal was used. If the original  $F_0$  trace was less than 500 ms, or the smooth signal less than 300 ms, then the last half of the signal was used. The 1<sup>st</sup> DCT coefficient was used to approximate slope and the 2<sup>nd</sup> was used to approximate curvature. As discussed in 3.2, the slope provides information about whether, and if so, how much, the singers are gliding up or down, while the curvature indicates the amount that the  $F_0$  is lower in the middle than at the two ends of the analyzed signal once the slope has been subtracted. The slope and curvature of the end of the first note is of interest because it provides an indication about whether the singer is anticipating and preparing for the second note in the interval more so under certain conditions than others. The data were analyzed in a number of ways, including the examination of the means and standard deviations across groupings of interval conditions, visualisation of data in box and whisker plots, and the linear regression analysis to look for significant trends.

#### 4.1.1.4 Analytical Statistics

Linear regression analysis was chosen for some of the analyses over the more commonly used ANOVA technique because it provides information about the effect direction and size for each of the factors considered, without the need for post-hoc hypothesis testing (Platt 1998), and measures how well the sum of the weighted predictors explains the variance in

the data (Cohen 2002). However, the ANOVAs are still useful to explore interactions among independent variables. Linear regression shares a number of assumptions with ANOVA, including those concerning linearity and the normality of the data distribution. The normality of the data's distribution was evaluated. The data distribution did not pass the Kolmogorov-Smirnov test's criteria for normality and a quantile-quantile plot subsequently revealed that the distribution was heavy-tailed. It has been found, however, that this type of departure from normality does not significantly impact the results of ANOVAs, *t*-tests, or regression analyses (Kutner et al. 2005).

In the regression analysis,  $\beta$  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  $\beta$  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 through  $\beta$ 's 95% confidence interval. If the confidence interval does not include zero then the effect can be considered significant.

The linear regression analysis was used to quantify the effect of various potential explanations of the variations in interval size and slope and curvature measurements. The quantity being predicted (the dependent variable) was either interval size in cents, an approximation of cents/second for slope, or an approximation of cents/second<sup>2</sup> for curvature. Six linear regressions were performed, listed in Table 4.1.1, which fall into two broad categories. The first category consists of four regressions: one each for the semitones and whole tones performed by both the non-professional and professional groups. All of the regressions in this category have five different predictors, or independent variables (1 for accompaniment, 1 for intervallic direction, 2 for intervallic condition, and 1 for singer identity). In the second category, the regression was performed over both groups of singers simultaneously for the semitone and whole tone conditions, with singer identity being replaced by group identity. Table 4.1.1 also details the conditions in each regression that were encoded as binary variables. For example, an ascending accompanied leading tone performed by singer four was coded as 101000001 while a descending *a cappella* whole tone between two chord tones performed by singer two was coded as 011101000.

	Accom.	Desc.	Leading tone	Non A-B $\flat$ /B $\flat$ -A	1 <sup>st</sup> Note Chord Tone	2 <sup>nd</sup> Note Chord Tone	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	
Whole tones (Pro)	X	X			X	X	X	X	X	X	X	
Whole tones (Non-pro)	X	X			X	X	X	X	X	X	X	
Semitones (All)	X	X	X	X								X
Whole tones (All)	X	X			X	X						X

Table 4.1.1: Summary of the conditions evaluated in the regression analyses performed in this section. The columns list the various independent variables and the X's indicate which of these were used in each regression.

These regressions were also run on the interval size data with the values rounded to either the nearest five or ten cents in order to assess whether a coarser description of interval size would allow for trends to emerge; however, the results were not significantly different. Logistic regressions were also run to examine whether the singers' intonation was systematically larger or smaller than equal temperament, Pythagorean tuning, or Just Intonation. There was no significant correspondence found between any of the singers and any of the tuning systems.

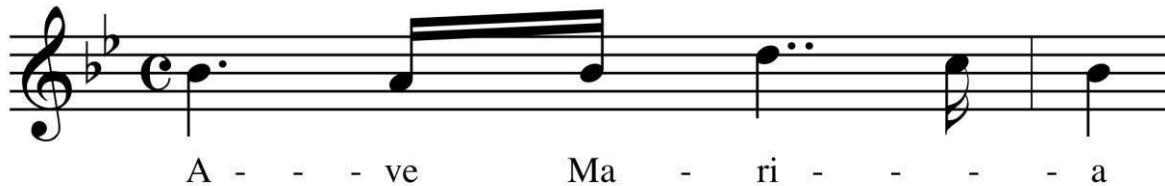
In addition to the linear regression analysis, two ANOVAs were performed to explore interactions between independent variables and to validate the findings from the linear regression analysis. The first ANOVA was performed with semitone size as the dependent variable and the second with whole tone size as the dependent variable. One ANOVA used a 2 (accompaniment) x 2 (direction) x 2 (spelling) x 2 (group) mixed model ANOVA design, with singer nested within group, to investigate the effect of musical context on semitone interval size. The between-subjects variables were singer and group and the within-subject variables for the semitone analysis were accompaniment, direction, and spelling. The other ANOVA used a 2 (accompaniment) x 2 (direction) x 3 (interval type) x 2 (group) mixed model ANOVA design, with singer nested within group, to investigate the effect of musical

context on whole tone interval size. The between-subjects variables were the same as the semitone ANOVA (singer and group), though the within-subject variables differed in that interval type replaced spelling. The spelling factor in the semitone analysis defined whether the semitone was between A-B $\flat$ /B $\flat$ -A or another pair of notes. The interval type factor indicated whether the whole tone occurred between two chord tones, a chord tone and a non-chord tone, or a non-chord tone and a chord tone.

The results of the ANOVA for the individual factors confirmed the results of the linear regression analysis, inasmuch as the various predictors considered in the regression could be assessed in an ANOVA. Specifically, the leading tone function in the semitones and the impact on starting and/or ending on a chord tone versus a non-chord could not be explicitly addressed in the ANOVA since these factors could not be represented independently. There were no significant effects for the interactions of the variables for the semitones. For the whole tones there was only a significant effect for the interaction of direction, interval type, and singer ( $F(20,143) = 3.22, p < 0.01$ ).

#### 4.1.2 Results

Overall, there was a high degree of variability between the singers, though some singers were more self-consistent than others. The following figures (Figure 4.1.3–Figure 4.1.6) not only demonstrate this variability with data from the opening and closing “Ave Maria” statements, but also show intra-performance consistency in the relative size of the intervals.



#### Idealized Interval Sizes

<b>Equal Temperament</b>	-100	100	400	-200	-200
<b>Pythagorean</b>	-90	90	408	-204	-204
<b>5-limit Just Intonation</b>	-112	112	386	-182 (Minor)	-204 (Major)

#### Singer 1

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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)

#### Singer 2

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-63.8 (SD = 3.5)	101.0 (SD = 6.2)	375.83 (SD = 6.1)	-183.2 (SD = 11.3)	-227 (SD = 16.5)

#### Singer 3

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-88.5 (SD = 14.2)	95.4 (SD = 13.4)	376.3 (SD = 7)	-176.7 (SD = 4.6)	-213.6 (SD = 5.6)

Figure 4.1.3 Comparison of the opening and closing statements of “Ave Maria” for non-professional singers 1–3. The Just Intonation tuning calculations were made in relation to the tonic triad.



#### Idealized Interval Sizes

<b>Equal Temperament</b>	-100	100	400	-200	-200
<b>Pythagorean</b>	-90	90	408	-204	-204
<b>5-limit Just Intonation</b>	-112	112	386	-182 (Minor)	-204 (Major)

#### Singer 4

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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)

#### Singer 5

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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)

#### Singer 6

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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)

Figure 4.1.4: Comparison of the opening and closing statements of "Ave Maria" for non-professional singers 4–6. The Just Intonation tuning calculations were made in relation to the tonic triad.



#### 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

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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)

Figure 4.1.5: Comparison of the opening and closing statements of “Ave Maria” for professional singers 1–3. The Just Intonation tuning calculations were made in relation to the tonic triad.



#### 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

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-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)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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

<b>Opening <i>A cappella</i></b>	-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)
<b>Opening Accompanied</b>	-97.0 (SD = 2.8)	103.5 (SD = 1.5)	392.9 (SD = 6.0)	-205 (SD = 9.8)	-199.1 (SD = 8.5)
<b>Closing <i>A cappella</i></b>	-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)
<b>Closing Accompanied</b>	-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)

Figure 4.1.6: Comparison of the opening and closing statements of “Ave Maria” for professional singers 4–6. The Just Intonation tuning calculations were made in relation to the tonic triad.

Amongst the non-professional group, the ascending A-B-flat semitones tended to be larger than the descending B-flat-A semitones, though the absolute semitone size varied amongst the singers. Overall, the descending semitones were generally smaller than Just-Intonation (112

cents) and equal temperament (100 cents), and also smaller than Pythagorean (90 cents) for some singers. For the descending whole tones, there were varying degrees of consistency. Singers 2, 3, 6, and, sometimes, 5 tended to compress the first descending whole tone more than the second. The mean interval sizes for the whole tones were closer to the Pythagorean/Major Just Intonation (204 cents) and equal tempered (200 cents) versions than the minor Just Intonation (182 cents), though some singers sang much larger whole tones than any of these. The variability in the impact of accompaniment seemed to be both singer- and interval-dependent, with some of the singers showing a greater amount of variability for some intervals than others.

Amongst the professional group, there was less consistency in the size of the ascending A-B $\flat$  semitones versus descending B $\flat$ -A semitones across singers. Overall, semitones were generally smaller than Just-Intonation (112 cents) and equal temperament (100 cents), and also smaller than Pythagorean (90 cents) for some singers. As with the non-professional group, there were varying degrees of consistency amongst sizes for the descending whole tone, though mean interval sizes were closer to the Pythagorean/Major Just Intonation (204 cents) and equal tempered (200 cents) versions than the minor Just Intonation (182 cents). In contrast with the non-professional group, there seems to be little impact of accompaniment in this group on the means, though there are some observable effects for standard deviations of some of the singers' whole tones.

Table 4.1.2 provides an overview of the mean interval size and standard deviation for both semitones and whole tones divided into ascending versus descending and *a cappella* versus accompanied. The high standard deviations both show the variability within these conditions and demonstrate that different types of analyses are needed to verify observable trends in the means, such as a tendency for descending semitones to be smaller than ascending ones.

Conditions (Number of Instances)	Non-professional Singers				Professional Singers			
	<i>A cappella</i>		Accompanied		<i>A cappella</i>		Accompanied	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Semitones, ascending (144)	89.8	20.4	95.1	16.1	98.7	17.5	98.6	16.6
Semitones, descending (162)	87.3	18.4	88.9	17.7	93.0	17.6	93.8	16.4
Whole tones, ascending (198)	197.8	22.8	195.0	22.0	199.6	21.1	202.9	21.3
Whole tones, descending (234)	200.5	18.5	200.7	18.7	202.8	18.9	203.4	17.7

Table 4.1.2: Summary of the mean interval size and standard deviation in cents for the two subject groups across all of the semitones and whole tones used in this experiment.

The observations made from Table 4.1.2 and Figure 4.1.3–Figure 4.1.6 are a useful starting point for making generalizations about the intonation tendencies in the experimental recordings. However, the large amount of variability within the broad conditions of ascending versus descending and *a cappella* versus accompanied needs to be explored in more detail since no clear trends emerged within these categories. The following sections explore a more fine-grained categorization within both semitones and whole tones.

#### 4.1.2.1 Semitones

In order to assess how the tuning of a semitone in the piece is influenced by the context in which it occurs, the direction of the interval, and/or the presence of accompaniment the semitone intonation data for the singers in each group, was evaluated for several conditions (see Figure 4.1.1): leading tones (36 intervals per group = 2 instances in each rendition x 6 singers x 3 *a cappella* renditions), other A-B $\flat$  semitones (72 intervals per group), B $\flat$ -A semitones (72 intervals per group), ascending other semitones (36 intervals per group), and 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 $\flat$  intervals, 144 B $\flat$ -A intervals, 72 ascending other semitones, and 180 descending other semitones for each group of singers. Overall, each group had 144 ascending semitones for each set of *a cappella* and accompanied renditions, as well as 162 descending semitones.

##### 4.1.2.1.1 Interval Size

The mean interval sizes and standard deviations across all of the singers for the various semitone conditions are shown for the non-professional group in Figure 4.1.3 and for the professional group in Figure 4.1.4. In the non-professional group, the mean interval size of both the *a cappella* and accompanied leading tones were the smallest: 79.3 cents for the *a cappella* leading tones ( $SD = 15.5$ ) and 90.7 for the accompanied leading tones ( $SD = 12.7$ ). Overall, the standard deviations for the leading tone condition were smaller than other corresponding *a cappella* and accompanied semitone conditions. In the professional group, there was much less variation between the ascending semitone conditions, though amongst the descending semitones, both the mean interval size and standard deviations of the B $\flat$ -A condition are markedly smaller in both the *a cappella* (87.6 cents,  $SD = 13.4$ ) and accompanied conditions (89.9 cent,  $SD = 13.9$ ).

Non-professional Group	<i>A cappella</i>		Accompanied	
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	79.3	15.5	90.7	12.7
A-B $\flat$ semitones, non-leading tones (72)	95.4	19.6	99.5	13.4
B $\flat$ -A semitones (72)	83.4	19.0	86.4	16.2
Other semitones, ascending (36)	89.2	22.3	90.7	21.4
Other semitones, descending (90)	90.4	17.5	91.0	18.6

Table 4.1.3: Mean and standard deviation of the semitone sizes in cents in the non-professional group.

Professional Group	<i>A cappella</i>		Accompanied	
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	94.8	16.5	93.2	14.3
A-B $\flat$ semitones, non-leading tones (72)	98.6	16.6	98.7	16.5
B $\flat$ -A semitones (72)	87.6	13.4	89.9	13.9
Other semitones, ascending (36)	102.7	19.7	103.9	17.7
Other semitones, descending (90)	97.2	19.4	96.9	17.7

Table 4.1.4: Mean and standard deviation of the semitone sizes in cents in the professional group.

The box and whisker plots in Figure 4.1.7 and Figure 4.1.8 show the range of interval sizes for the ascending versus descending and the *a cappella* versus accompanied conditions for each singer for both the non-professional (Figure 4.1.7) and professional (Figure 4.1.8) groups. The plots in Figure 4.1.9 and Figure 4.1.10 show the interval sizes for each semitone condition across all of the singers in each group: non-professionals in Figure 4.1.9 and professionals in Figure 4.1.10.

In the box and whisker plots, the top and bottom of each box represents the 25<sup>th</sup> and 75<sup>th</sup> percentiles, with the solid horizontal line running through the box representing the 50<sup>th</sup> percentile, or median. The short solid horizontal lines at the end of the “whiskers” represent the 5<sup>th</sup> and 95<sup>th</sup> percentiles (the most extreme non-outlier data points), and the plus signs indicate the outliers. One way to interpret these figures is to consider that the smaller the boxes, the more consistent the singer was in the condition. For the non-professionals in Figure 4.1.7, it can be observed that the size of the semitone sung with accompaniment is more consistent than those sung *a cappella* for both the ascending and descending semitone conditions. Also, the median of the interval size was smaller for semitones sung *a cappella* than for the accompanied ones, with the accompanied ones being closer to equal temperament. For the professional singers in Figure 4.1.8, the degree of consistency was

singer dependent, as was the amount that the consistency differed across conditions. Overall, the median interval size was larger for the professionals and more consistent between the *a cappella* and the accompanied conditions.

In terms of the overall behaviour of each group in the more detailed semitone conditions of leading tone, other A-B $\flat$  semitones, B $\flat$ -A semitones, other ascending semitones, and other descending semitones, there are some notable differences between the groups. For the non-professional group (Figure 4.1.9), the interquartile (25<sup>th</sup> to 75<sup>th</sup> percentile) range, shown by the boxes, is quite consistent in the *a cappella* condition and generally smaller in the accompanied one, though there is more variation in the (size/position) of the 95th percentile intervals, which is shown by the whiskers. The median leading tone size is smaller than the other A-B $\flat$  semitones and slightly smaller than the B $\flat$ -A semitones. Similar trends can be observed for the professional group (Figure 4.1.10), though the leading tones' median size is more comparable to the other A-B $\flat$  semitones and larger than the B $\flat$ -A semitones. Also, there is less of a difference between the *a cappella* and accompanied conditions.

The first linear regression analysis was run over intervallic direction, intervallic condition (leading tone versus non-leading tone and A-B $\flat$ /B $\flat$ -A versus other spellings), whether the singer was accompanied, and singer identity analysis for each of the groups. The regression analysis on the semitone data from the non-professional group had a relatively low  $R^2$  value ( $R^2=0.19, p < 0.0001$ ), indicating that only some of the variance in the data was explained by the conditions considered in the regression. The regression did reveal that the leading tone semitones were on average 10 cents smaller than the other semitones (95% confidence interval = [5,14]); however, there was no significant effect for size of the A-B $\flat$ /B $\flat$ -A semitones (including leading tones) compared to the other semitones. The *a cappella* semitones were on average 3 cents (95% confidence interval = [1,6]) smaller than the accompanied ones, and the descending intervals 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, 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. There was not a significant difference in interval size between singers three and six.

For the professional group, the regression analysis'  $R^2$  value was smaller ( $R^2=0.09, p < 0.0001$ ), indicating that less of the variance in the data was explained. The regression analysis did reveal that the A-B $\flat$ /B $\flat$ -A semitones (including leading tones) were on average 7 cents smaller than the other semitones (95% confidence interval = [4,10]); however, there were no significant effects for leading tone function. The descending semitones were on average 8 cents smaller than the ascending ones (95% confidence interval = [4,10]). There were no significant effects for the *a cappella* versus accompanied condition, nor were there any significant effects for singer identity.

The second linear regression analysis, where both groups were combined and singer identity was replaced by group identity, also had a small  $R^2$  value ( $R^2=0.07, p < 0.0001$ ), although it produced significant results for all conditions except for *a cappella* versus accompanied. The descending semitones were on average 8 cents smaller than ascending ones (95% confidence interval = [6,10]). The leading tones were on average 7 cents smaller than the non-leading tone semitones (95% confidence interval = [3,10]), while the average semitone size of the A-B $\flat$ /B $\flat$ -A semitones (including the leading tones) was 4 cents larger than the other semitones (95% confidence interval = [2,6]). The professional group's semitones were on average 6 cents larger than the non-professional group's (95% confidence interval = [4,8]

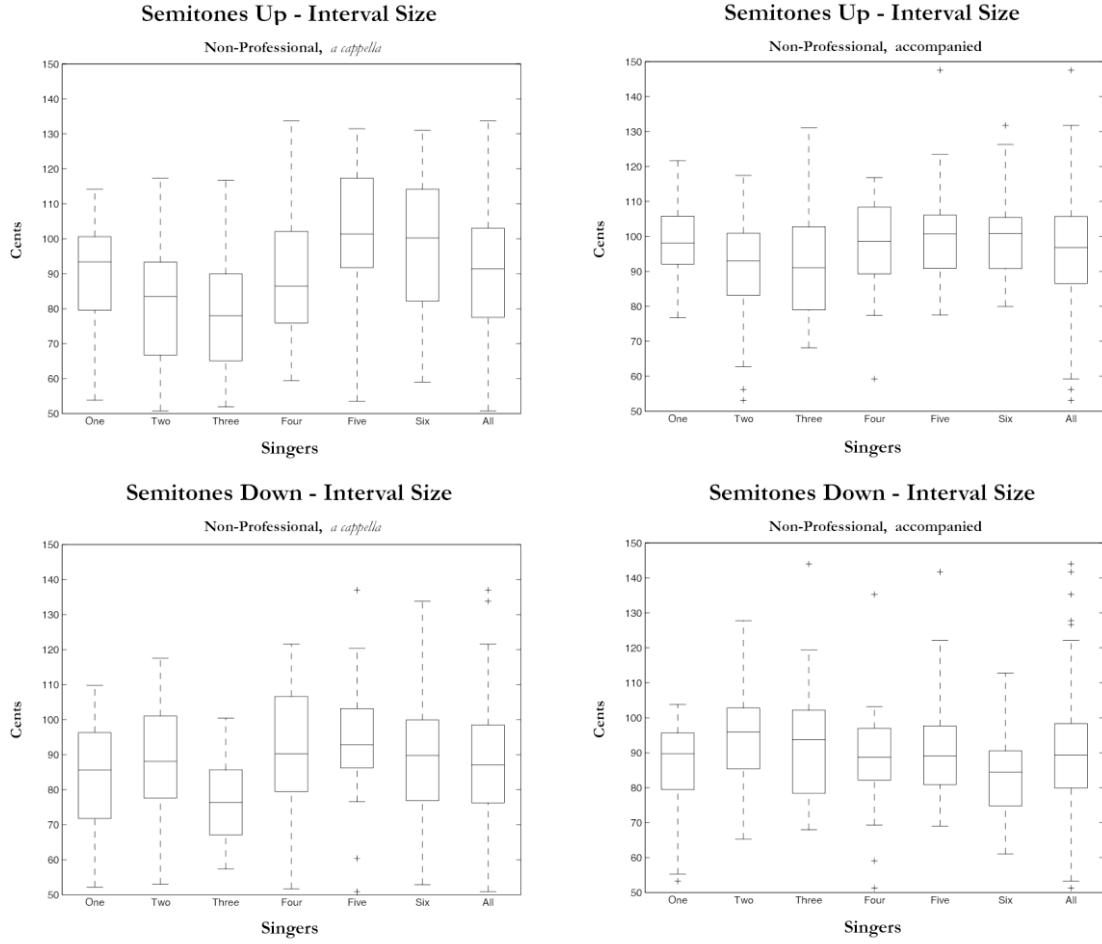


Figure 4.1.7: Box and whisker plots of semitone interval sizes across all non-professional singers. The subjects are represented individually as well as in combination on the x-axis. The y-axis shows the size of the intervals in cents. The plots on the left show the interval sizes for the *a cappella* performances, and the plots on the right show the interval sizes for the performances with accompaniment. The plots on the top show the interval sizes for the ascending semitones, and the plots on the bottom show the interval sizes for the descending semitones.

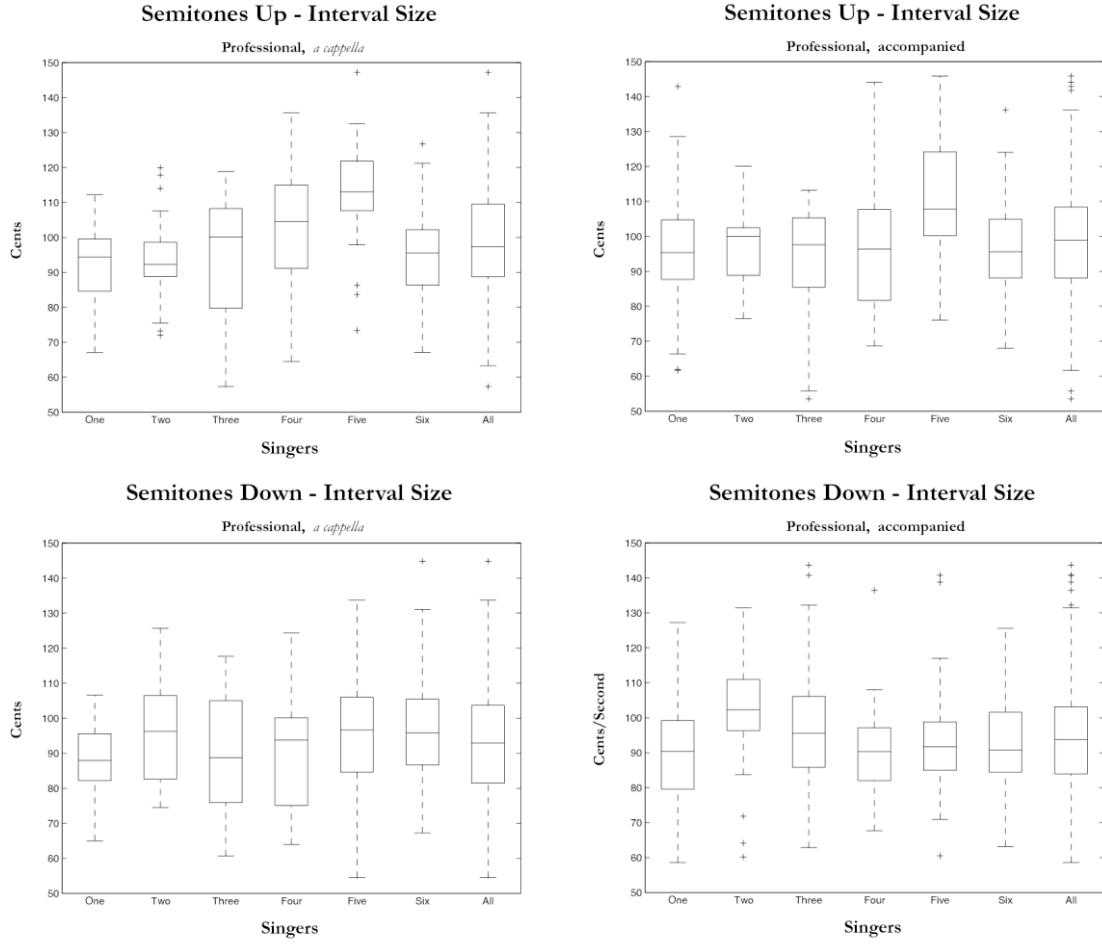


Figure 4.1.8: Box and whisker plots of semitone interval sizes across all professional singers. The subjects are represented individually as well as in combination on the x-axis. The y-axis shows the size of the intervals in cents. The plots on the left show the interval sizes for the *a cappella* performances, and the plots on the right show the interval sizes for the performances with accompaniment. The plots on the top show the interval sizes for the ascending semitones, and the plots on the bottom show the interval sizes for the descending semitones.

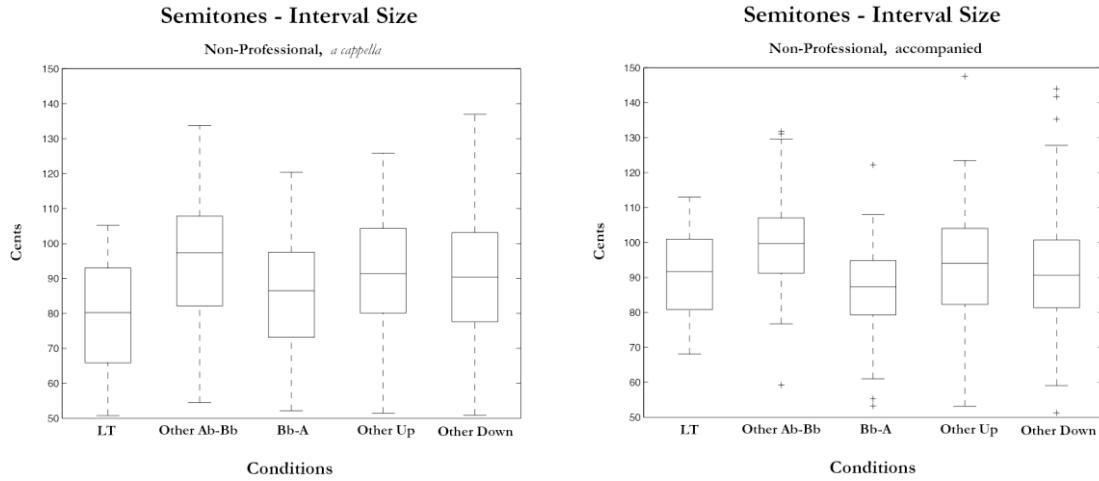


Figure 4.1.9: Box and whisker plots of the semitone size in cents for each semitone condition (leading tones, other A-B $\flat$  semitones, B $\flat$ -A semitones, other ascending semitones, and descending) across all non-professional singers. The plot on the left shows the interval sizes for the *a cappella* performances, and the plot on the right shows the interval sizes for the performances with accompaniment.

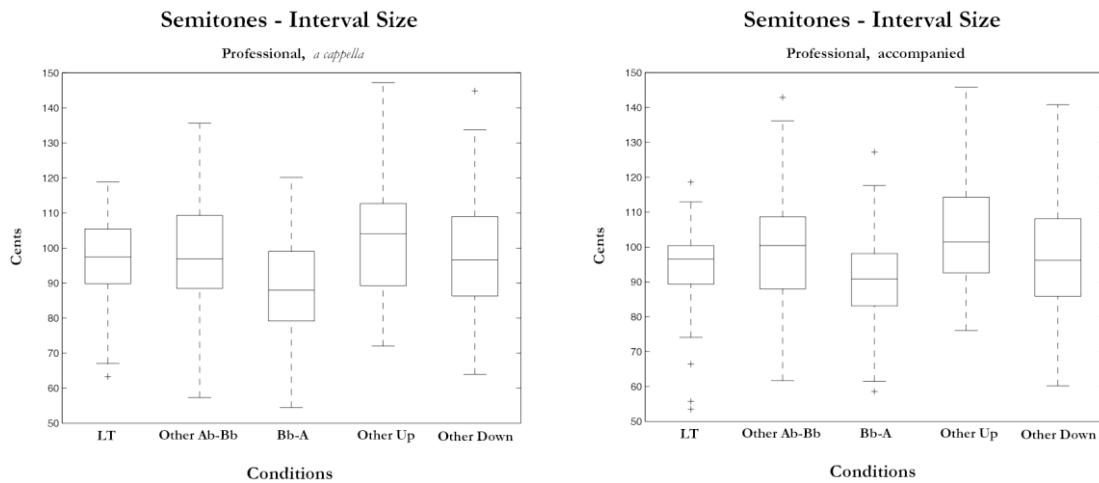


Figure 4.1.10: Box and whisker plots of the semitone size in cents for each semitone condition (leading tones, other A-B $\flat$  semitones, B $\flat$ -A semitones, other ascending semitones, and descending) across all professional singers. The plot on the left shows the interval sizes for the *a cappella* performances, and the plot on the right shows the interval sizes for the performances with accompaniment.

#### 4.1.2.1.2 Slope and Curvature

A summary of the slope and curvature values for the endings of the first note in each semitone interval across the ascending versus descending and *a cappella* versus accompanied conditions are shown for both the non-professional and professional groups in Table 4.1.5. As discussed above, slope is approximated by the 1<sup>st</sup> DCT coefficient and curvature by the 2<sup>nd</sup> DCT coefficient of the last 250 ms of the signal when the DCT is calculated on the original F<sub>0</sub> trace or the last 150 ms of the signal when a moving average with a 200 ms window has been applied to the F<sub>0</sub> trace. The units for slope and curvature are approximations of cents/second and cents/second<sup>2</sup>, respectively. For notes that are less than 500 ms, the last half of the note is evaluated. The data are further broken down into conditions in the subsequent tables. The means and standard deviations for the slope values run on the original F<sub>0</sub> trace are shown in Table 4.1.6 for the non-professional group and Table 4.1.8 for the professional group. The results from the 1<sup>st</sup> DCT run on the F<sub>0</sub> trace with a 200 ms moving average applied to it are shown in Table 4.1.7 for the non-professional group and in Table 4.1.9 for the professional group. Likewise, the means and standard deviations for the curvature values run on the original F<sub>0</sub> trace are shown in Table 4.1.10 for the non-professional and in Table 4.1.12 professional groups. The results from the 2<sup>nd</sup> DCT coefficient run on the F<sub>0</sub> trace with a 200 ms moving average applied to it are shown in Table 4.1.11 for the non-professional group and in Table 4.1.13 for the professional group.

Conditions (Number of Instances)	Non-professional Singers				Professional Singers			
	<i>A cappella</i>		Accompanied		<i>A cappella</i>		Accompanied	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Slope, F <sub>0</sub> trace, ascending (144)	32.3	62.9	30.6	55.4	9.3	100.4	-5.1	115.9
Slope , MA, ascending (144)	12.2	32.8	11.3	47.8	15.0	67.5	5.6	43.9
Slope, F <sub>0</sub> trace, descending (162)	-19.7	53.4	-17.4	109.8	-47.8	115.0	-18.7	75.4
Slope, MA, descending (162)	9.4	36.1	2.8	40.5	2.7	34.3	6.8	38.0
Curvature, F <sub>0</sub> trace, ascending (144)	15.2	388.6	45.5	552.6	-141.9	865.5	-179.2	719.4
Curvature, MA, ascending (144)	56.5	204.0	53.7	172.3	-275.6	922.0	-129.5	595.4
Curvature, F <sub>0</sub> trace, descending (162)	-122.7	481.0	-151.6	600.9	-5.5	442.4	13.8	325.2
Curvature, MA, descending (162)	-8.3	124.0	-2.8	190.2	-24.8	226.0	26.2	157.1

Table 4.1.5: Summary of the means and standard deviations of the slope and curvature for the two subject groups across all of the semitones used in this experiments.

Non-professional Group	<i>A cappella</i>	Accompanied		
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	12.3	37.7	15.0	35.3
A-B $\flat$ semitones, non-leading tones (72)	25.5	59.2	24.7	41.5
B $\flat$ -A semitones (72)	-13.1	40.8	-4.1	47.7
Other semitones, ascending (36)	45.3	38.2	33.5	33.9
Other semitones, descending (90)	-26.2	63.0	-28.8	147.4

Table 4.1.6: Non-professional group's semitone slope values calculated on the original F<sub>0</sub> trace.

Non-professional Group	<i>A cappella</i>	Accompanied		
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	10.8	30.8	4.3	34.8
A-B $\flat$ semitones, non-leading tones (72)	14.0	35.1	18.3	35.5
B $\flat$ -A semitones (72)	-0.2	26.3	-2.8	33.2
Other semitones, ascending (36)	19.5	22.6	12.9	27.6
Other semitones, descending (90)	14.6	38.8	9.4	46.6

Table 4.1.7: Non-professional group's semitone slope values calculated on the F<sub>0</sub> trace with a moving average applied to it.

Professional Group	<i>A cappella</i>	Accompanied		
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	8.9	43.1	1.7	35.6
A-B $\flat$ semitones, non-leading tones (72)	5.0	64.4	-3.9	63.3
B $\flat$ -A semitones (72)	-48.1	71.5	-9.8	52
Other semitones, ascending (36)	30.3	83.9	41.6	67.9
Other semitones, descending (90)	-52.7	148.2	-26.3	91.1

Table 4.1.8: Professional group's semitone slope values calculated on the F<sub>0</sub> trace.

Professional Group	<i>A cappella</i>	Accompanied		
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	2.8	23.7	1.5	25.0
A-B $\flat$ semitones, non-leading tones (72)	3.3	31.3	10.4	31.1
B $\flat$ -A semitones (72)	-2.4	36.1	10.4	33.4
Other semitones, ascending (36)	29.8	28.4	17.6	34.1
Other semitones, descending (90)	3.5	30.1	5.5	39.5

Table 4.1.9: Professional group's semitone slope values calculated on the F<sub>0</sub> trace with a moving average applied to it

Non-professional Group	<i>A cappella</i>		Accompanied	
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	-68.9	257.7	103.4	289.4
A-B $\flat$ semitones, non-leading tones (72)	-22.1	334.9	4.1	231.5
B $\flat$ -A semitones (72)	24.8	220.5	0.3	250.8
Other semitones, ascending (36)	84.1	256.3	113.0	186.0
Other semitones, descending (90)	-257.7	609.1	-279.8	796.6

Table 4.1.10: Non-professional group's semitone curvature values calculated on the F<sub>0</sub> trace.

Non-professional Group	<i>A cappella</i>		Accompanied	
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	11.0	71.7	23.4	88.2
A-B $\flat$ semitones, non-leading tones (72)	20.7	140.6	34.5	82.7
B $\flat$ -A semitones (72)	-16.5	110.3	-11.0	108.9
Other semitones, ascending (36)	53.7	79.9	20.7	96.5
Other semitones, descending (90)	6.9	128.2	5.5	239.8

Table 4.1.11: Non-professional group's semitone curvature values calculated on the F<sub>0</sub> trace with a moving average applied to it.

Professional Group	<i>A cappella</i>		Accompanied	
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	-40.0	165.4	35.8	133.7
A-B $\flat$ semitones, non-leading tones (72)	-89.6	520.9	-205.3	431.4
B $\flat$ -A semitones (72)	-110.3	275.6	-81.3	420.3
Other semitones, ascending (36)	73.0	642.2	-59.3	651.9
Other semitones, descending (90)	-456.2	1248.6	-191.6	742.8

Table 4.1.12: Professional group's semitone curvature values calculated on the F<sub>0</sub> trace.

Professional Group	<i>A cappella</i>		Accompanied	
Semitone conditions (Number of Instances)	Mean	SD	Mean	SD
A-B $\flat$ semitones, leading tones (36)	26.2	118.5	15.2	117.1
A-B $\flat$ semitones, non-leading tones (72)	27.6	128.2	48.2	110.3
B $\flat$ -A semitones (72)	-64.8	206.7	22.1	129.5
Other semitones, ascending (36)	35.8	91.0	46.9	122.7
Other semitones, descending (90)	1.4	242.6	44.1	165.4

Table 4.1.13: Professional group's semitone curvature values calculated on the F<sub>0</sub> trace with a moving average applied to it.

The following box and whisker plots show the range of the 1<sup>st</sup> DCT coefficient values, measured in an approximation of cents/second, and the 2<sup>nd</sup> DCT coefficient values, measured in an approximation of cents/second<sup>2</sup>. The plots show the ascending versus descending and *a cappella* versus accompanied conditions for each singer run on both the original F<sub>0</sub> trace for the non-professional group (Figure 4.1.11 for the 1<sup>st</sup> DCT and Figure 4.1.17 for the 2<sup>nd</sup> DCT) and the professional group (Figure 4.1.12 for the 1<sup>st</sup> DCT and Figure 4.1.18 for the 2<sup>nd</sup> DCT) and on results of applying a 200 ms moving average to the original F<sub>0</sub> trace for the non-professional group (Figure 4.1.13 for the 1<sup>st</sup> DCT and Figure 4.1.19 for the 2<sup>nd</sup> DCT) and the professionals group (Figure 4.1.14 for the 1<sup>st</sup> DCT and Figure 4.1.20 for the 2<sup>nd</sup> DCT). The plots in Figure 4.1.15 and Figure 4.1.16 show the 1<sup>st</sup> DCT coefficient for each semitone condition across all of the singers in the non-professional group and professional group, respectively. Similarly, the plots in Figure 4.1.21 and Figure 4.1.22 show the 2<sup>nd</sup> DCT coefficient for each semitone condition across all of the singers in the two groups.

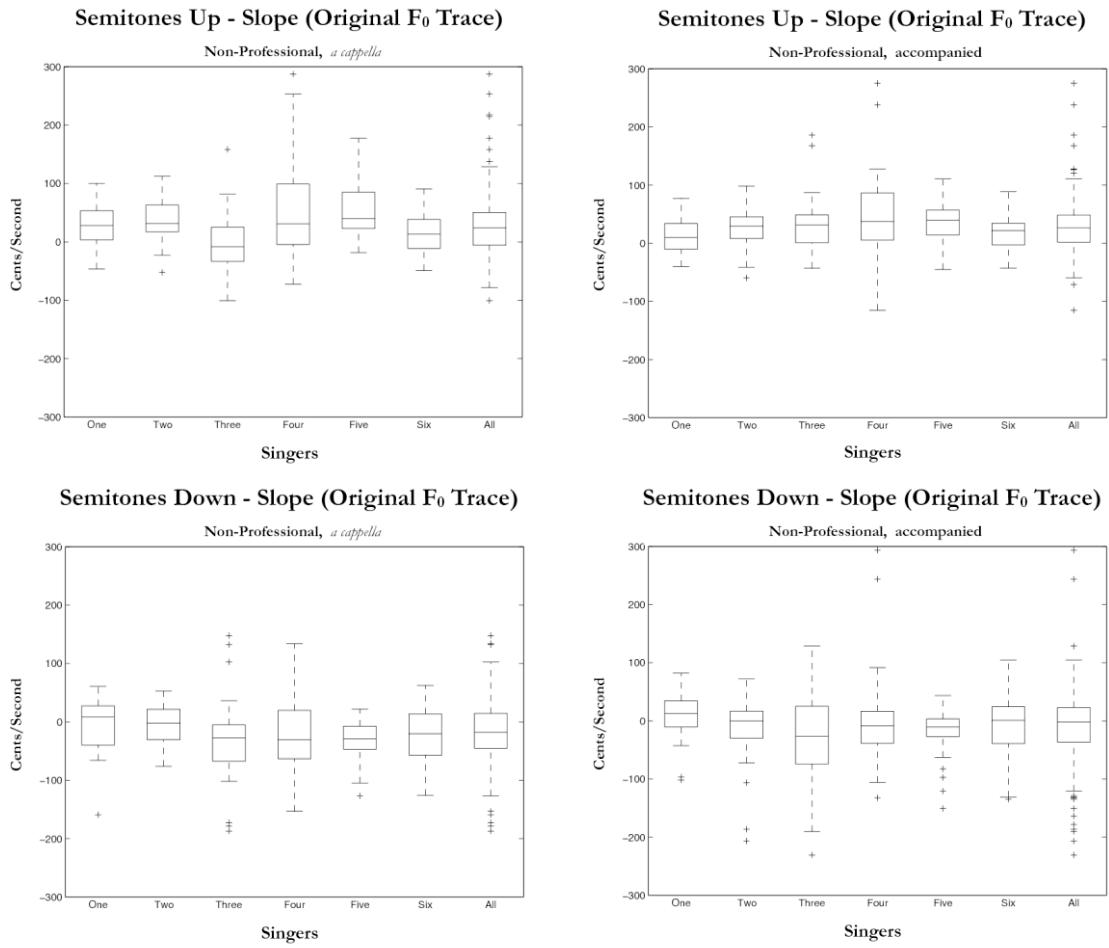


Figure 4.1.11: Box and whisker plots of the 1<sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the  $F_0$  trace of the first note of all of the semitones performed by the non-professional group. Each plot shows the results for the six non-professional singers individually and the mean across all of the singers. The plots on the left show the values of the 1<sup>st</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 1<sup>st</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 1<sup>st</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 1<sup>st</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second.

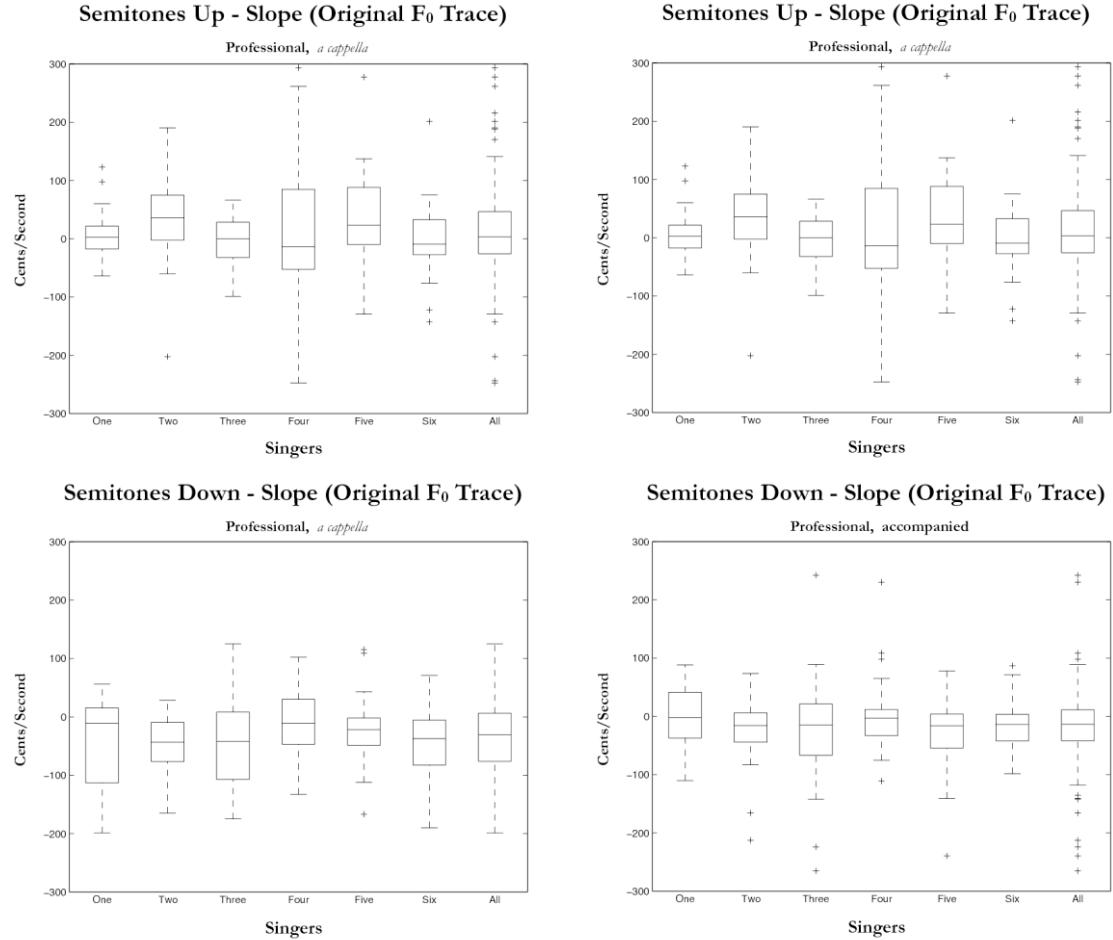


Figure 4.1.12: Box and whisker plots of the 1<sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the  $F_0$  trace of the first note of all of the semitones performed by the professional group. Each plot shows the results for the six professional singers individually and the mean across all of the singers. The plots on the left show the values of the 1<sup>st</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 1<sup>st</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 1<sup>st</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 1<sup>st</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second.

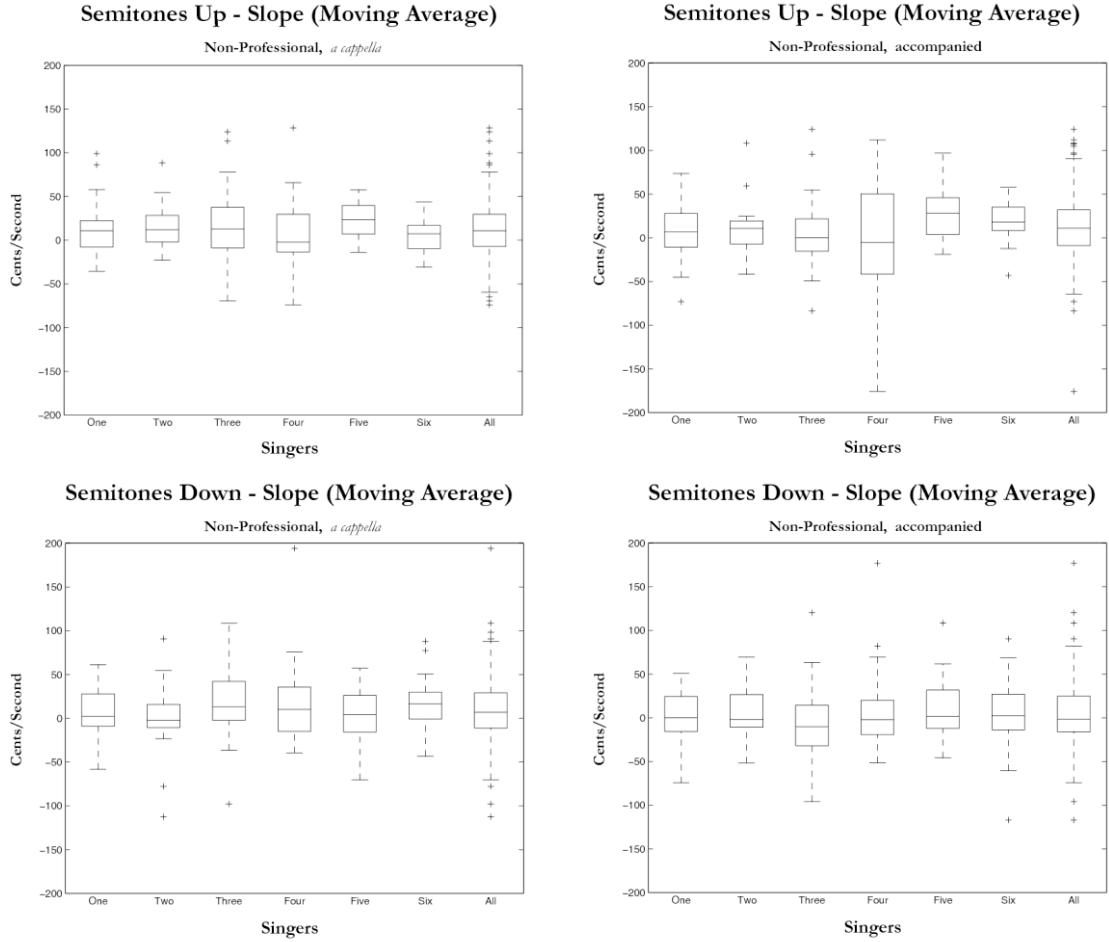


Figure 4.1.13: Box and whisker plots of the 1<sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 150 ms of the  $F_0$  trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the semitones performed by the non-professional group. Each plot shows the values for the six non-professional singers individually and the mean across all of the singers. The plots on the left show the values of the 1<sup>st</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 1<sup>st</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 1<sup>st</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 1<sup>st</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second.

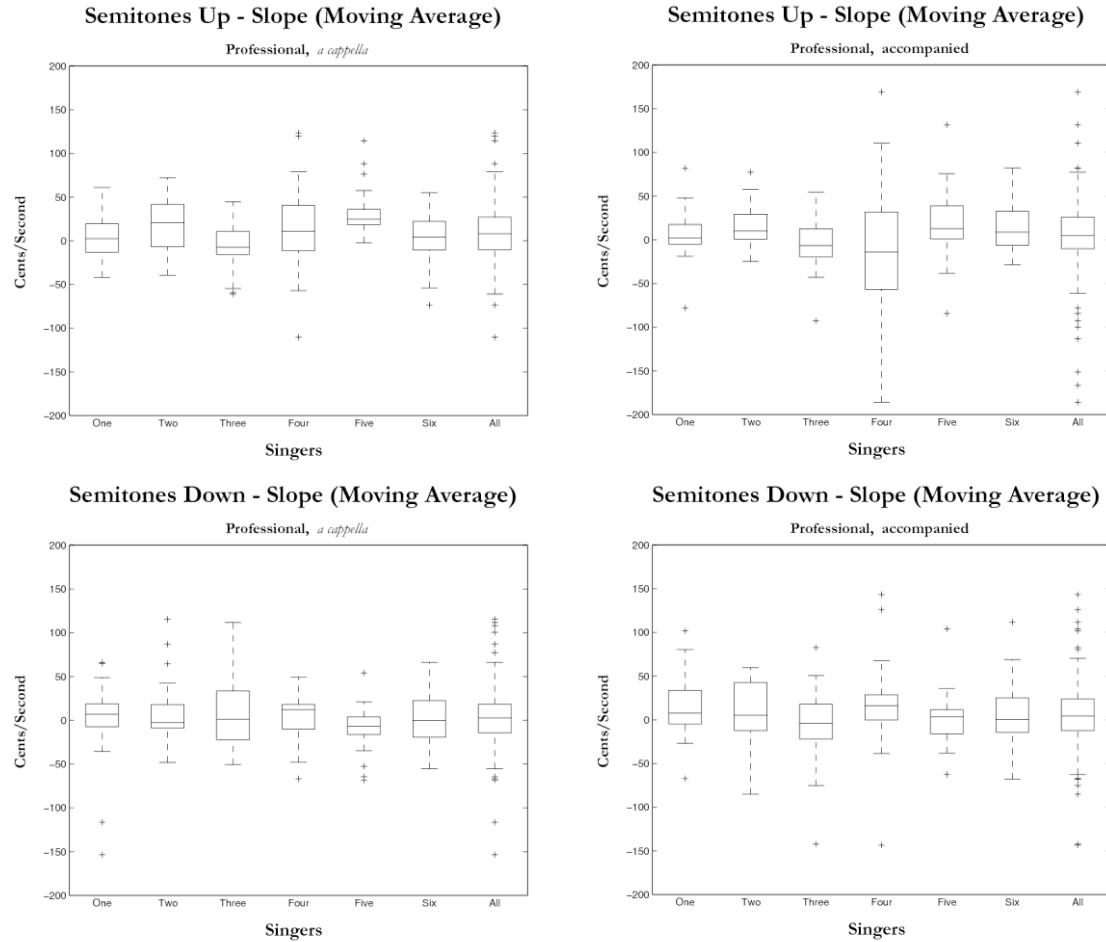


Figure 4.1.14: Box and whisker plots of the 1<sup>st</sup> discrete cosine transform (DCT) coefficient (approximating slope) run on the last 150 ms of the  $F_0$  trace smoothed by applying a 200 ms moving average of the first note of all of the semitones performed by the professional group. Each plot shows the results for the six professional singers individually and the mean across all of the singers. The plots on the left show the values of the 1<sup>st</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 1<sup>st</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 1<sup>st</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 1<sup>st</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second.

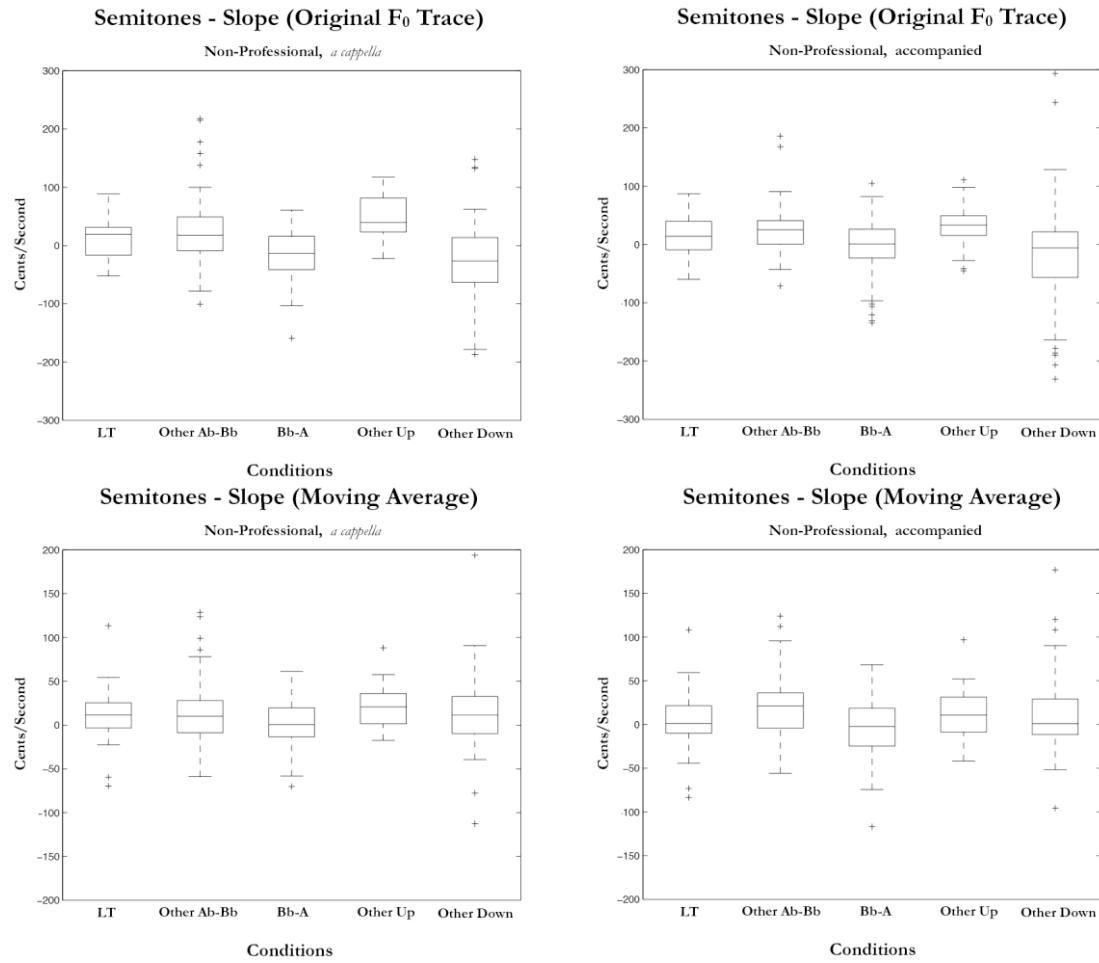


Figure 4.1.15: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of all of the semitones in each condition across all of the non-professional singers. The plots on the left show the values for the 1<sup>st</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values for the 1<sup>st</sup> DCT coefficient for performances with accompaniment. The plots on the top show the values of the 1<sup>st</sup> DCT coefficient run on the original F<sub>0</sub> trace, while the plots on the bottom show the value 1<sup>st</sup> DCT coefficient run on the F<sub>0</sub> trace smoothed by results of applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second.

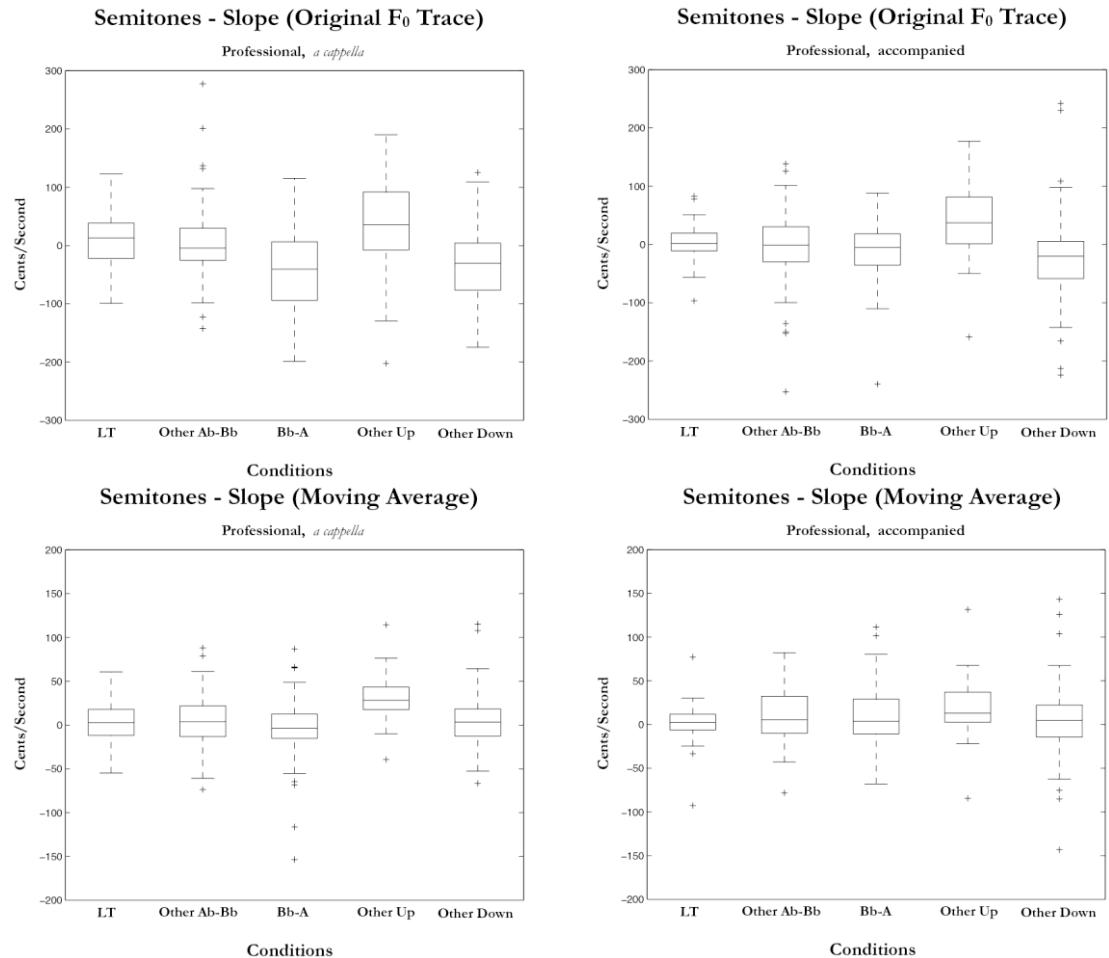


Figure 4.1.16: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of all of the semitones in each condition across all of the professional singers. The plots on the left show the values of the 1<sup>st</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 1<sup>st</sup> DCT coefficient for performances with accompaniment. The plots on the top show the values of the 1<sup>st</sup> DCT coefficient run on the original F<sub>0</sub> trace, while the plots on the bottom show the values of the 1<sup>st</sup> DCT coefficient run on the F<sub>0</sub> trace smoothed by results of applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second.

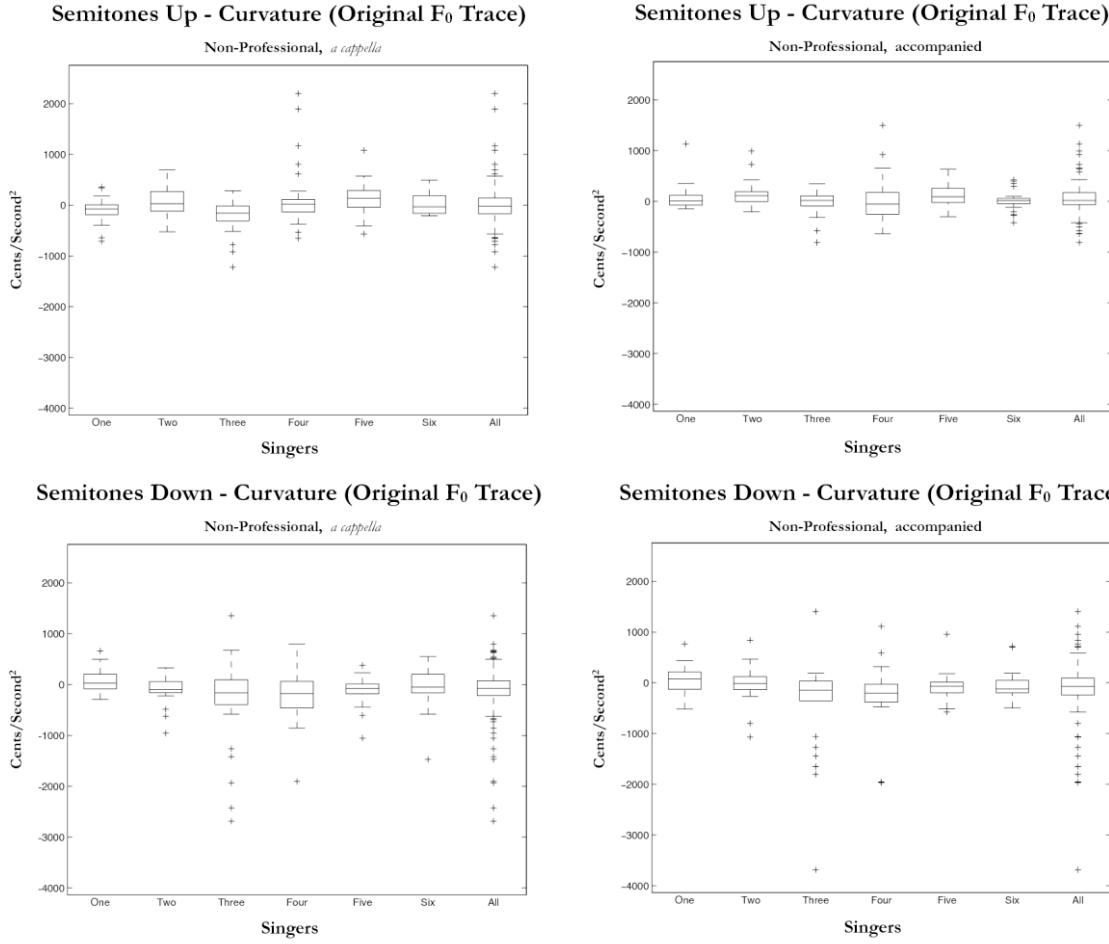


Figure 4.1.17: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the  $F_0$  trace of the first note of all of the semitones performed by the non-professional group. Each plot shows the results for the six non-professional singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

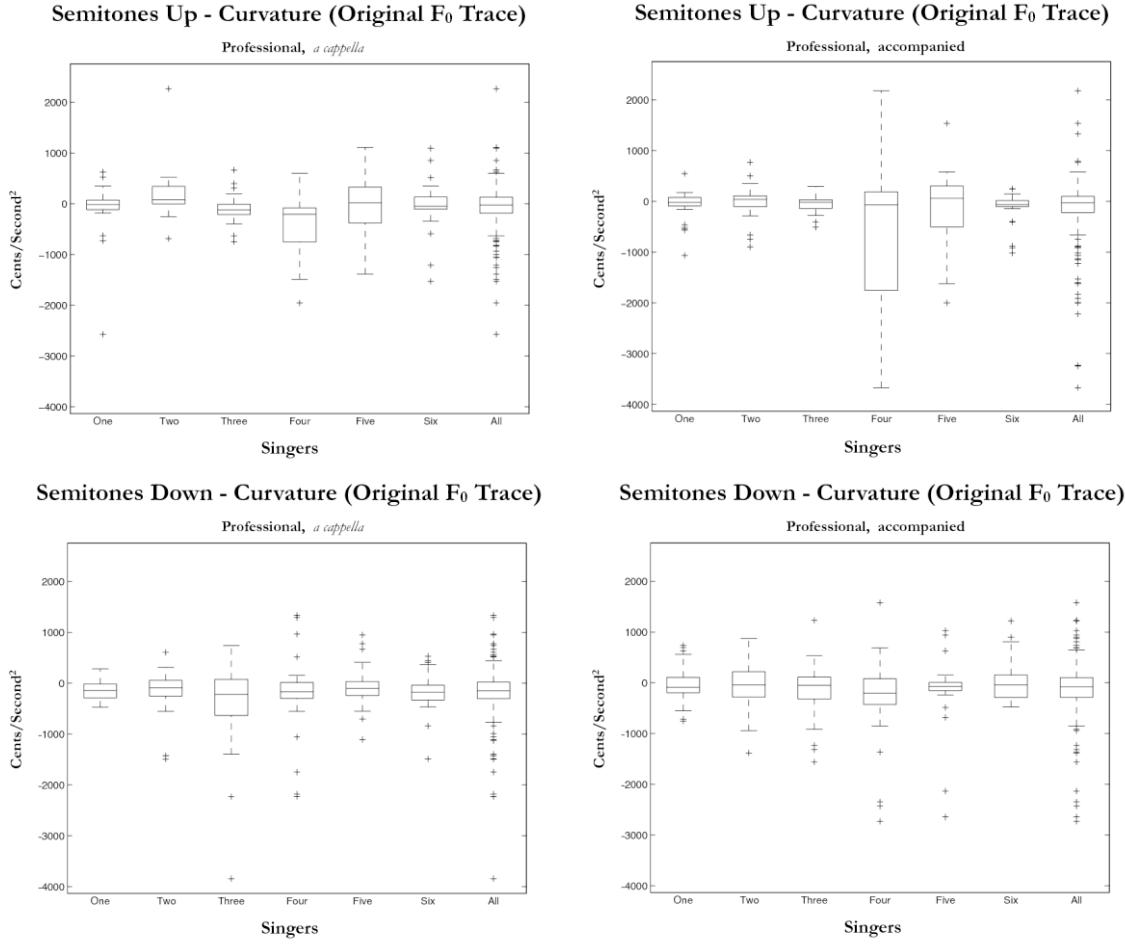


Figure 4.1.18: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the  $F_0$  trace of the first note of all of the semitones performed by the professional group. Each plot shows the results for the six professional singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

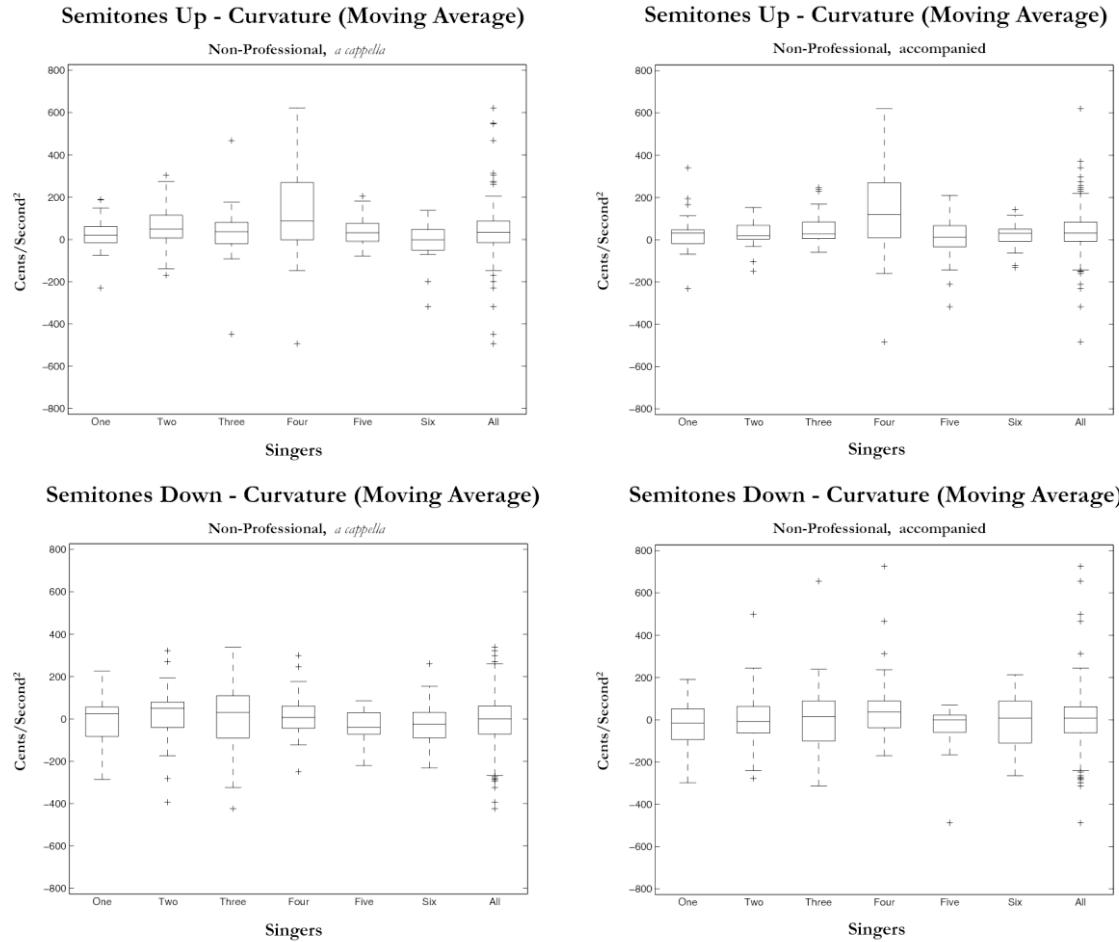


Figure 4.1.19: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 150 ms of the  $F_0$  trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the semitones performed by the non-professional group. Each plot shows the results for the six non-professional singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

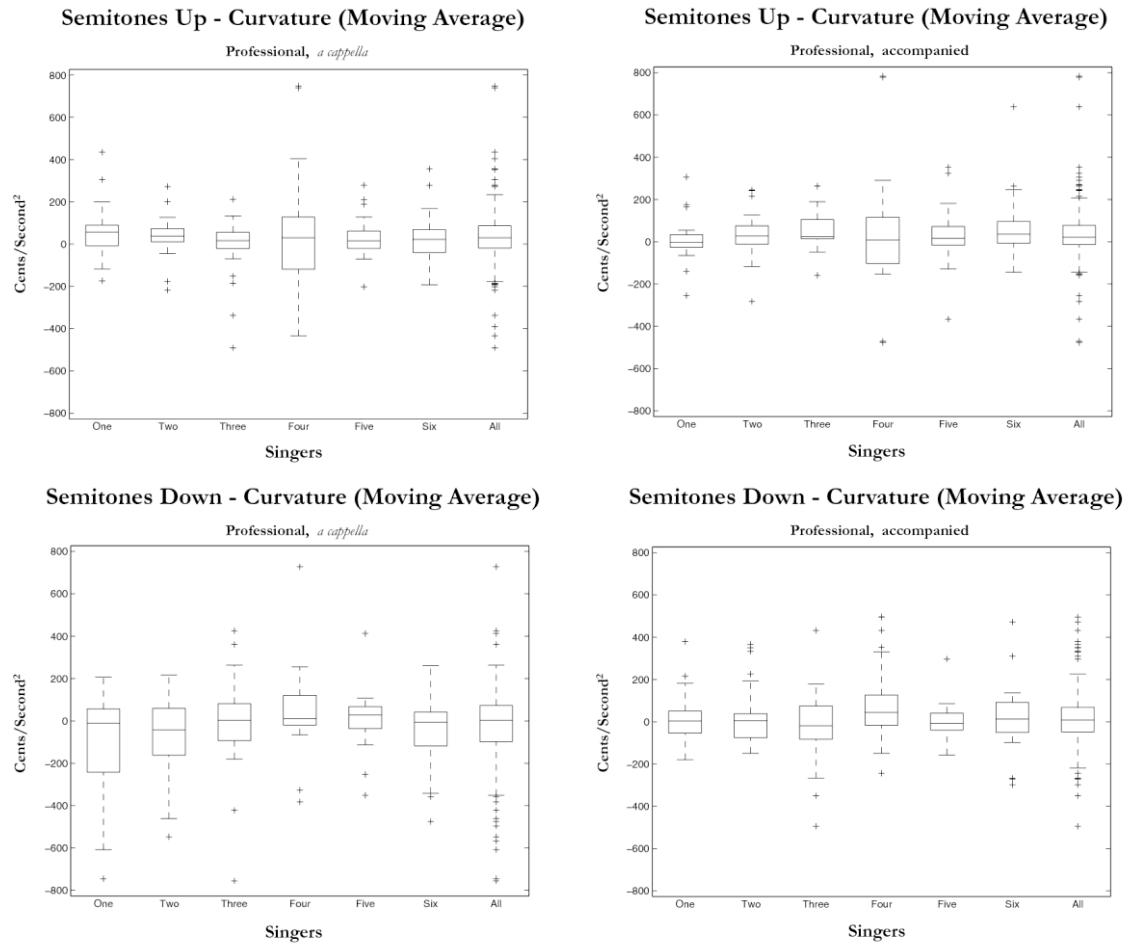


Figure 4.1.20: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 150 ms of the  $F_0$  trace (smoothed by results of applying a 200 ms moving average of the first note) of all of the semitones performed by the professional group. Each plot shows the results for the six professional singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending semitones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending semitones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

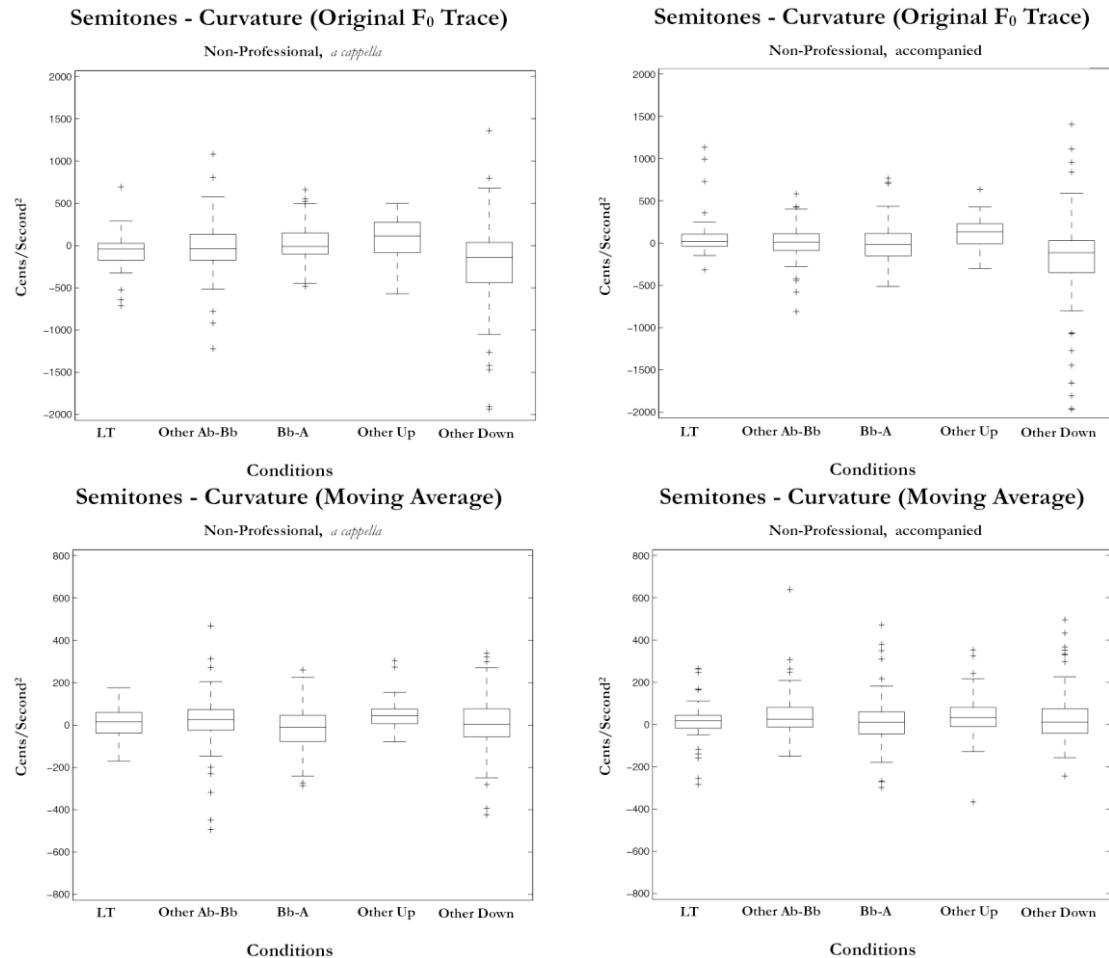


Figure 4.1.21: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of all of the semitones in each condition across all of the non-professional singers. The plots on the left show the 2<sup>nd</sup> DCT coefficient values for the *a cappella* performances, and the plots on the right show the 2<sup>nd</sup> DCT coefficient values for performances with accompaniment. The plots on the top show the values of the 2<sup>nd</sup> DCT coefficient run on the original F<sub>0</sub> trace, while the plots on the bottom show the values of the 2<sup>nd</sup> DCT coefficient run on the F<sub>0</sub> trace smoothed by results of applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

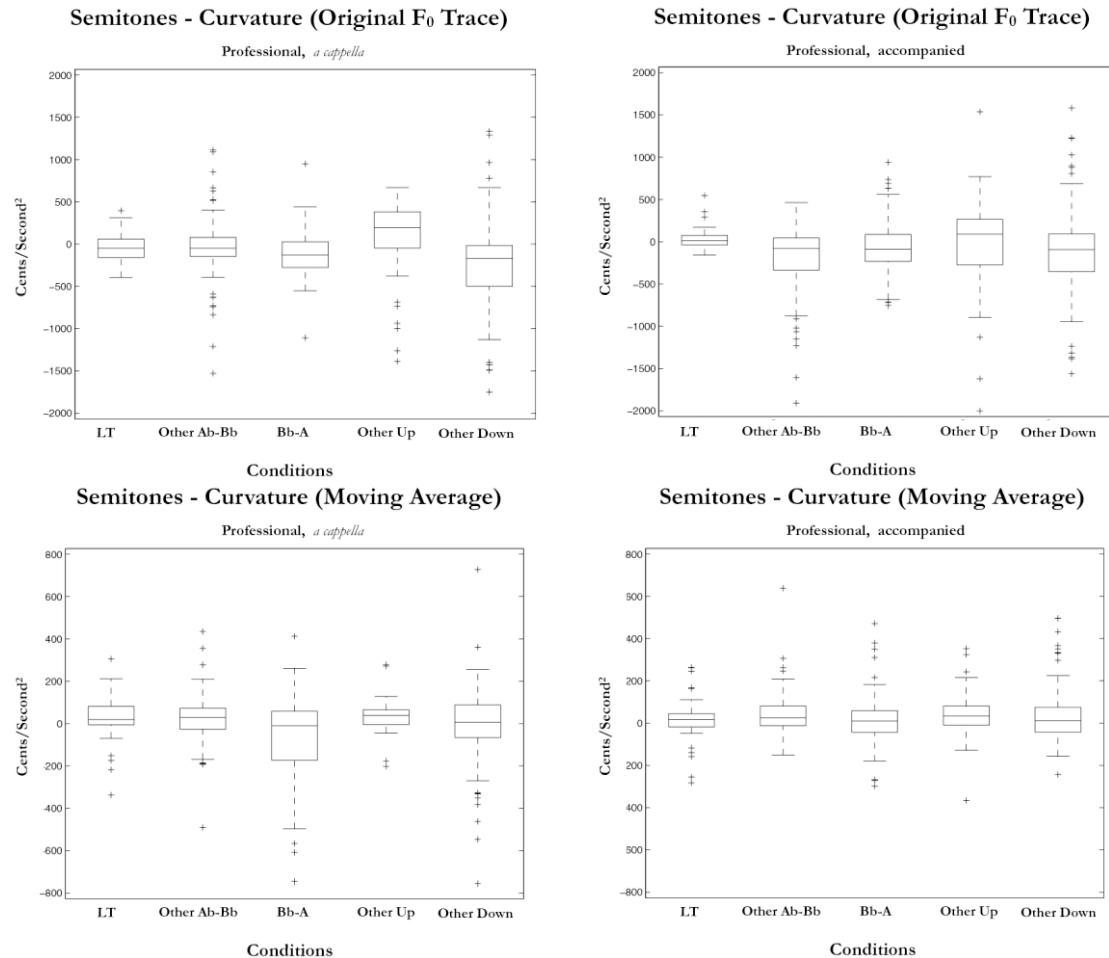


Figure 4.1.22: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of all of the semitones in each condition across all of the professional singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the values of the 2<sup>nd</sup> DCT coefficient run on the original  $F_0$  trace, while the plots on the bottom show the values of the 2<sup>nd</sup> DCT coefficient run on the  $F_0$  trace smoothed by results of applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

In order to analyze these data, the same regressions run on interval size were also run on the slope and curvature data. The first linear regression analysis of the original  $F_0$  trace's 1<sup>st</sup> DCT coefficients in the non-professional group ( $R^2=0.08$ ,  $p < 0.0001$ ) showed significant effects for A-B $\flat$ /B $\flat$ -A semitones versus other semitones, with the other semitones having a slope value 18 cents/second larger than on average (95% confidence interval = [7,28]). There were

also significant effects for singer identity when comparing the singers 1, 2, and 5 against the baseline, singer six. Singer one's slopes were on average smaller by 17 cents/second (95% confidence interval = [1,33]), singer two's were 23 cents/second larger (95% confidence interval = [6,39]), and singer five's were 20 cents/second smaller (95% confidence interval = [4,37]) than singer six's slopes.

For the data from the professional group, the same linear regression analysis ( $R^2=0.14$ ,  $p < 0.0001$ ) showed a significant effect for accompaniment, with the accompanied performances having on average a 16 cents/second smaller slope than the *a cappella* ones (95% confidence interval = [5,27]). As with the non-professional group, there was a significant effect for A-B<sub>b</sub>/B<sub>b</sub>-A semitones versus other semitones, with the other semitones having a 21 cents/second larger slope on average (95% confidence interval = [9,33]). There were also significant singer identity effects for singers one, three, and four. Singer one's slope was on average 23 cents/second larger than singer six's (95% confidence interval = [5,42]), singer three's was on average 56 cents/second larger (95% confidence interval = [37,74]), and singer four's was on average 30 cents/second larger (95% confidence interval = [11,48]). Neither of these regressions had a particularly high  $R^2$  value. Although the professional group (0.14) had a slightly larger  $R^2$  value than the non-professional group (0.08), this still indicates that much of the variation in the data was left unexplained.

The second linear regression analysis, run over the data from both groups together, also had a small  $R^2$  value ( $R^2 = 0.04$ ,  $p < 0.0001$ ). It did, however, show a significant effect again for A-B<sub>b</sub>/B<sub>b</sub>-A semitones versus other semitones, with the other semitones having a 19 cents/second larger slope on average (95% confidence interval = [11,27]). It also showed a significant effect for group identity, with the professional group having on average a 10 cents/second larger slope than the non-professional group (95% confidence interval = [3,18]).

In the first linear regression analysis of the 1<sup>st</sup> DCT coefficients calculated on the result of applying a 250 ms moving average to the  $F_0$  trace ( $R^2=0.09$ ,  $p < 0.0001$ ), there were only significant effects in the non-professional group for singer identity. There were no significant effects for accompaniment, direction, or type of semitone (leading tones versus non-leading tones or A-B<sub>b</sub>/B<sub>b</sub>-A semitones versus other semitones). In terms of singer identity, the all of the singers' slopes except for singer three were significantly different than singer six's slopes.

Singer one's average slope was 12 cents/second smaller than singer six's (95% confidence interval = [5,19]), singer two's average slope was 10 cents/second larger (95% confidence interval = [3,17]), singer four's average slope was 10 cents/second smaller (95% confidence interval = [3,17]), and singer five's average slope was 8 95% confidence interval = [1,15].

In the results of the same linear regression run on the professional singers' data ( $R^2=0.20$ ,  $p < 0.0001$ ), there were significant effects for leading tone function and singer identity. There were no significant effects for accompaniment or direction. The slope of the leading tone semitones was on average 8.3 cents/second smaller than the non-leading tone semitones (95% confidence interval = [2,14]). All of the singers, except for singer five, differed significantly from the baseline. Singer one's average slope was 7 cents/second larger than singer six's (95% confidence interval = [2,13]), singer two's average slope was 6 cents/second larger (95% confidence interval = [0.06,12]), singer three's average slope was 26 cents/second larger (95% confidence interval = [20,31]), and singer four's average slope was 24 cents/second larger (95% confidence interval = [18,30]). The  $R^2$  value was larger for the regression run on the professional group's data (0.2) than on the non-professional group's data (0.09).

In the second linear regression ( $R^2=0.01$ ,  $p < 0.0001$ ), where both groups' average slope data were analyzed, there was only a significant effect for leading tone function. The slopes of the semitones with a leading tone function were on average 6 cents/second smaller than the slopes of those without (95% confidence interval = [0.8,10]).

In the first linear regression analysis of the 2<sup>nd</sup> DCT coefficients calculated on the original  $F_0$  trace showed significant effects for intervallic direction, A-B $\flat$ /B $\flat$ -A semitones versus other semitones, and singer identity for all of the singers except for singer two ( $R^2=0.10$   $p < 0.0001$ ). The descending semitone's curvature was on average 8 second<sup>2</sup> larger (95% confidence interval = [21.4,150.8]) and the non A-B $\flat$ /B $\flat$ -A's curvature was on average 162.0 cents/second<sup>2</sup> larger than the A-B $\flat$ /B $\flat$ -A semitones (95% confidence interval = [98.5,225.6]). In terms of singer identity, all of the singers except for singer two showed a significant effect: singer one's curvature was on average 201.2 cents/second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [100.6,301.8]; singer three's curvature was on average 154.9 cents/second<sup>2</sup> smaller than singer six's curvature (95% confidence interval

= [54.2,255.5]); singer four's curvature was on average 177.1 cents/second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [76.5,277.7]); singer five's curvature was on average 222.8 cents/second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [122.1,323.4]).

For the professional group, there were significant effects for leading tone function, A-B<sub>b</sub>/B<sub>b</sub>-A semitones versus non A-B<sub>b</sub>/B<sub>b</sub>-A semitones, and singer identity for singers one, three, and four ( $R^2=0.19$   $p < 0.0001$ ). The semitones with a leading tone function's curvature were on average 156.2 cents/second<sup>2</sup> smaller than the non-leading tone semitones (95% confidence interval = [8.3,304.1]). The non-A-B<sub>b</sub>/B<sub>b</sub>-A semitones had on average a 248.6 cents/second<sup>2</sup> larger curvature than the A-B<sub>b</sub>/B<sub>b</sub>-A semitones (95% confidence interval = [154.7,342.5]). Singer one's curvature was on average 178                  second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [30.0,327.5]), singer three's curvature was on average 503.9 cents/second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [355.1,652.6]), and singer four's curvature was on average 410.4 cents/second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [261.7,559.2]).

For the second linear regression, across both groups of singers, there were significant effects for A-B<sub>b</sub>/B<sub>b</sub>-A semitones versus other semitones and group identity ( $R^2=0.07$   $p < 0.0001$ ). The non-A-B<sub>b</sub>/B<sub>b</sub>-A semitones' curvature was on average 205.3 cents/second<sup>2</sup> larger than the A-B<sub>b</sub>/B<sub>b</sub>-A semitones (95% confidence interval = [145.4,265.2]). The professional singers' average curvature was on average 102.5 cents/second<sup>2</sup> (95% confidence interval = [47.7,157.3]). As with the regressions run on the slope data, the  $R^2$  values were small overall and larger for the professional group (0.19) than the non-professional group (0.12). Also, like the regressions on the slope data, the  $R^2$  value for the regression run on both groups was quite small (0.07).

In the first linear regression analysis of the 2<sup>nd</sup> DCT coefficients calculated on the result of applying a 200 ms moving average to the F<sub>0</sub> trace, there were only significant effects for singer identity in the non-professional group ( $R^2=0.10$ ,  $p < 0.0001$ ). Singer one's average curvature was 58.5 cents/second<sup>2</sup> smaller than singer six's (95% confidence interval = [30.2,86.7]), singer two's average curvature was 30.6 cents/second<sup>2</sup> larger (95% confidence interval = [2.3,58.8]), singer three's average curvature was 34.5 cents/second<sup>2</sup> smaller (95% confidence interval = [6.3,62.8]), singer four's average curvature was 51.4 cents/second<sup>2</sup>

smaller (95% confidence interval = [23.1,79.7]), and singer five's average curvature was 45.5 cents/second<sup>2</sup> smaller (95% confidence interval = [17.2,73.7]).

For the linear regression analysis on the professional group, there were significant effects for accompaniment, direction, and for singers one, three, and four against the baseline ( $R^2=0.22$ ,  $p < 0.0001$ ). Accompanied semitones' curvature was on average 26.7 cents/second<sup>2</sup> smaller than *a cappella* semitones' average curvature (95% confidence interval = [8.3,45.2]) and descending semitones' curvature was on average 39.3 cents/second<sup>2</sup> larger than ascending semitones' average curvature (95% confidence interval = [18.7,59.8]). Singer one's average curvature was 35.8 cents/second<sup>2</sup> larger than singer six's (95% confidence interval = [3.7,67.8]), singer three's average curvature was 125.8 second<sup>2</sup> larger (95% confidence interval = [93.8,157.8]), and singer four's average curvature was 143.2 cents/second<sup>2</sup> larger (95% confidence interval = [111.2,175.2]). The relative sizes of the regressions' small  $R^2$  values are the same as the other regressions on the curvature and slope data and the  $R^2$  value for the regression run on the professional data (0.22) is larger than that for the non-professional data (0.10).

In the second linear regression analysis run on the combined curvature data for the entire group ( $R^2=0.02$ ,  $p < 0.0001$ ), there were significant effects for direction and group identity. The descending semitones' curvature was on average 28 second<sup>2</sup> larger than ascending semitones' average curvature (95% confidence interval = [13.9,43.5]). The professional group's curvature was on average 15.9 cents/second<sup>2</sup> larger than non-professional group's semitones' average curvature (95% confidence interval = [2.5,29.2]).

#### 4.1.2.2 Whole Tones

In order to assess the degree of variability between the singers in each group in terms of whole tones, interval size was evaluated for several conditions (see Figure 4.1.2): ascending whole tones between chord tones and non-chord tones (72 total intervals per group = 4 instances per rendition x 6 singers x 3 renditions), descending whole tones between chord tones and non-chord tones (54 total intervals per group), ascending whole tones between non-chord tones and chord tones (36 intervals per group), descending whole tones between non-chord tones and chord tones (108 intervals per group), ascending whole tones between chord tones (72 intervals per group), and descending whole tones between chord tones (90 intervals per group). There were the same number of conditions per group for the

accompanied renditions, resulting in a total of 144 ascending whole tones between chord tones and non-chord notes, 108 descending whole tones between chord tones and non-chord tones, 72 ascending whole tones between non-chord tones, 216 descending whole tones between non-chord tones and chord tones, 144 ascending whole tones between chord tones, and 144 descending whole tones between chord tones. Overall each group had 198 ascending and 234 descending whole tones, for each set of *a cappella* and accompanied renditions.

#### 4.1.2.2.1 Interval Size

The mean interval sizes and standard deviations across all of the singers for the various whole tone conditions are shown for the non-professional group in Table 4.1.14 and for the professional group in Table 4.1.15.

Non-professional Group		<i>A Cappella</i>		Accompanied	
Whole tone conditions (Number of Instances)		Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)		207.9	25.4	202.7	23.5
Chord tone – chord tone, descending (54)		-198.1	18.8	-194.5	19.6
Chord tone – non-chord tone, ascending (36)		188.2	15.6	187.7	15.9
Chord tone – non-chord tone, descending (108)		-199.3	18.7	-201.7	16
Non-chord tone – chord tone, ascending (72)		192.6	19.3	191	20.9
Non-chord tone – chord tone, descending (90)		-203.5	18	-203.2	20.5

Table 4.1.14: Mean and standard deviation of the whole tone sizes (in cents) for each of the tested whole tone conditions for the non-professional singers.

Professional Group		<i>A Cappella</i>		Accompanied	
Whole tone conditions (Number of Instances)		Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)		202.5	23.3	209	27.5
Chord tone – chord tone, descending (54)		-203.3	15.6	-200	18.1
Chord tone – non-chord tone, ascending (36)		200.9	19.7	200.3	15.3
Chord tone – non-chord tone, descending (108)		-200.6	20	-201.6	17.8
Non-chord tone – chord tone, ascending (72)		196.2	19.1	198	14.4
Non-chord tone – chord tone, descending (90)		-205.3	19.1	-207.4	16.9

Table 4.1.15: Mean and standard deviation of the whole-tone sizes (in cents) for each of the tested whole tone conditions for the professional singers.

In the non-professional group, ascending chord tone to chord tone whole tones tended to be larger than the corresponding descending whole tones. This was not the case, however, for the chord-tone-to-non-chord-tone whole tones and the non-chord-tone-to-chord-tone whole tones, where the corresponding descending whole tones tended to be larger. The means of the whole tone sizes in the professional group were more consistent across the chord tone to chord tone and chord tone to non-chord tone conditions; except for accompanied chord to chord tone condition, where the ascending intervals tended to be larger. In the non-chord tone to chord tone condition the descending whole tones tended to be larger.

The box and whisker plots in Figure 4.1.23 and Figure 4.1.24 show the range of interval sizes for the ascending versus descending and *a cappella* versus accompanied conditions for each singer for both the non-professional (Figure 4.1.23) and professional (Figure 4.1.24) groups. The plots in Figure 4.1.25 and Figure 4.1.26 show the interval sizes for each whole tone condition across all of the singers in each group, with non-professionals in Figure 4.1.25 and professionals in Figure 4.1.26. As discussed in Section 4.1.2.1, one way to interpret these figures is to consider that the smaller the boxes, the more consistent the singer was in the condition.

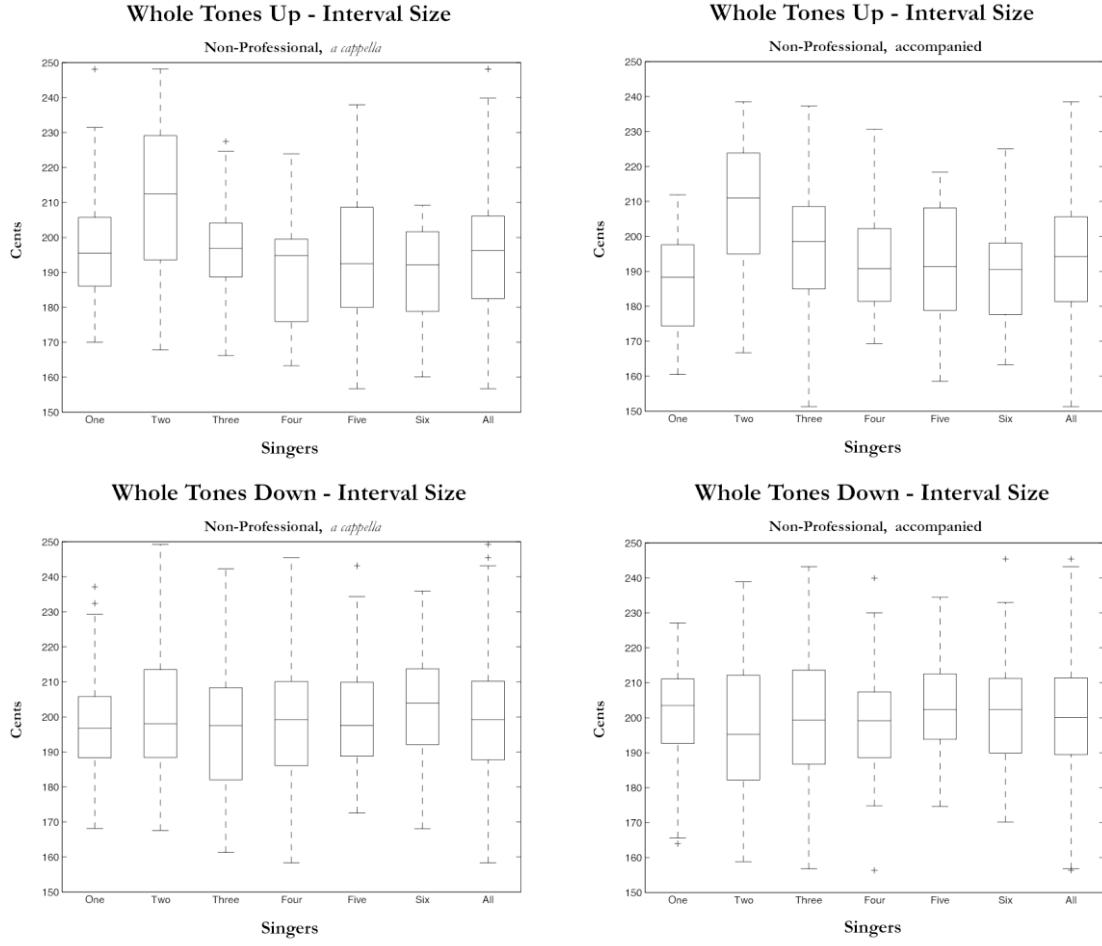


Figure 4.1.23: Box and whisker plots of whole tone interval sizes across all non-professional singers. Each subject is represented individually on the x-axis, as well as the combination of all of the subjects. The y-axis shows the size of the intervals in cents. The plots on the left show the interval sizes for the *a cappella* performances, and the plots on the right show the interval sizes for the performances with accompaniment. The plots on the top show the interval sizes for the ascending whole tones, and the plots on the bottom show the interval sizes for the descending whole tones.

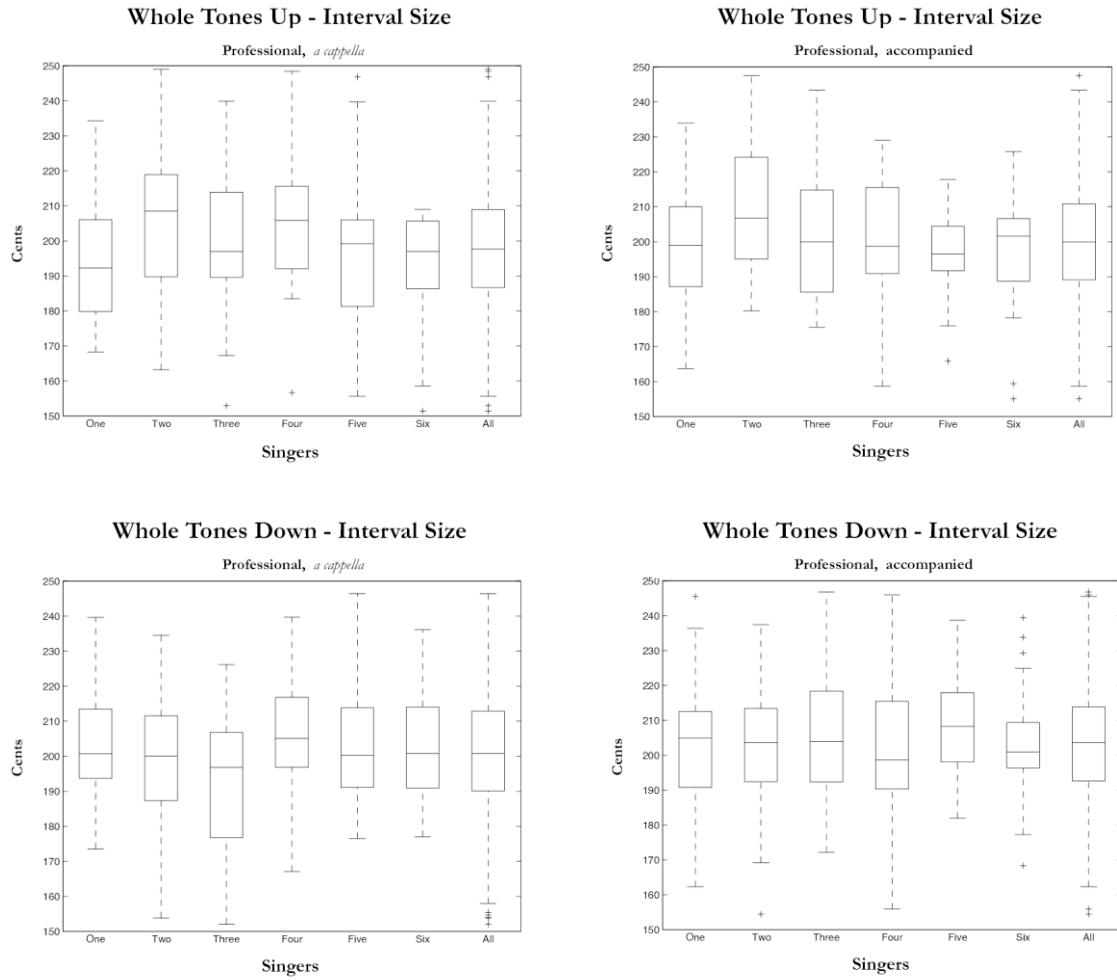


Figure 4.1.24: Box and whisker plots of whole tone interval sizes across all professional singers. Each subject is represented individually on the x-axis, as well as the combination of all of the subjects. The y-axis shows the size of the intervals in cents. The plots on the left show the interval sizes for the *a cappella* performances, and the plots on the right show the interval sizes for the performances with accompaniment. The plots on the top show the interval sizes for the ascending whole tones, and the plots on the bottom show the interval sizes for the descending whole tones.

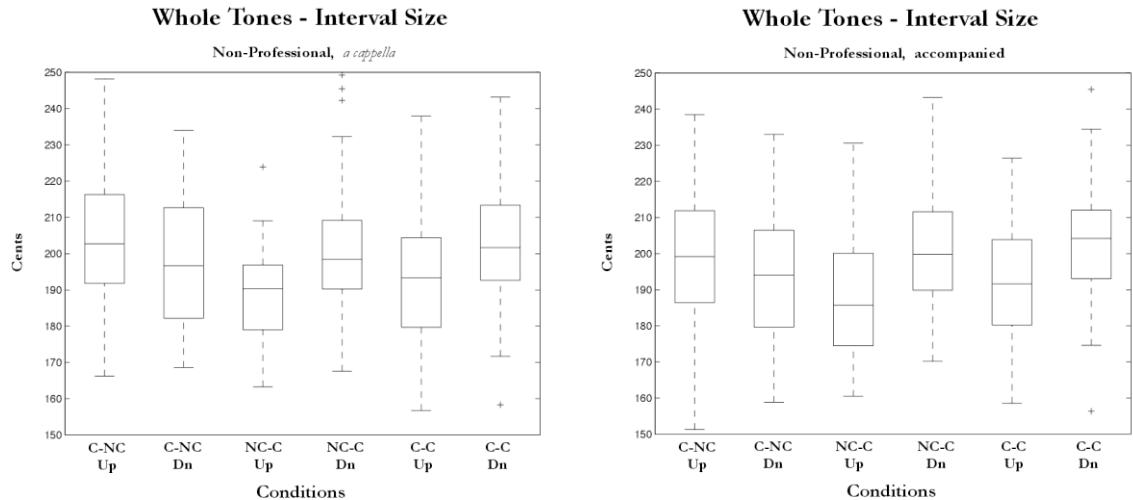


Figure 4.1.25: Box and whisker plots of the whole tone size in cents for each whole tone condition across all non-professional singers. The plot on the left shows the interval sizes for the *a cappella* performances, and the plot on the right shows the performances with accompaniment.

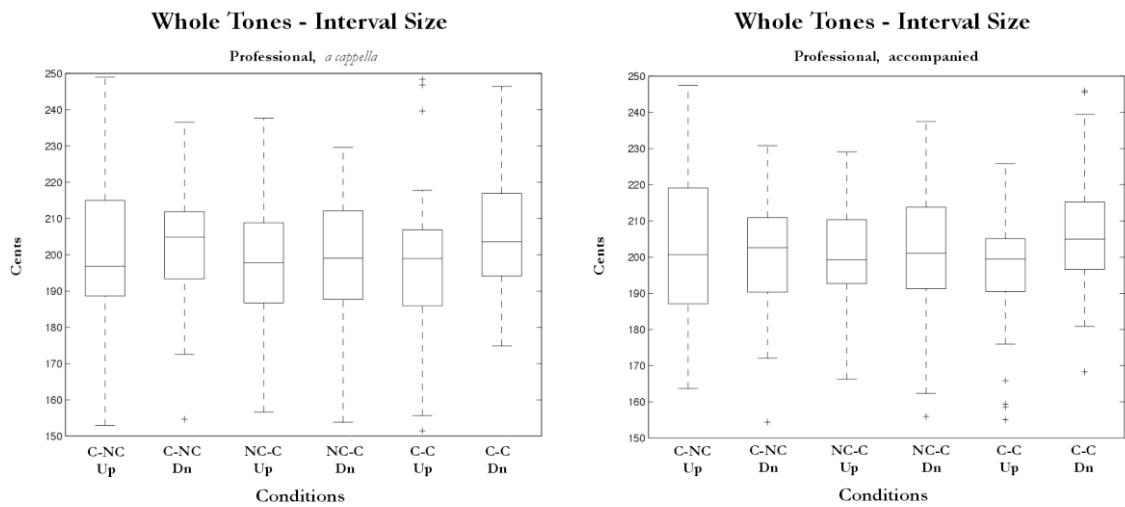


Figure 4.1.26: Box and whisker plots of the whole tone size in cents for each whole tone condition across all professional singers. The plot on the left shows the interval sizes for the *a cappella* performances, and the plot on the right shows the performances with accompaniment.

A linear regression analysis was run over whether the singer was accompanied, intervallic direction, intervallic condition, and singer identity for each of the groups. The results of the regression on the data from the non-professional group ( $R^2=0.09, p < 0.0001$ ) showed that the descending whole tones were on average 5 cents larger than the ascending whole tones (95% confidence interval = [2,8]). The regression also showed that intervals ending with a chord tone were 4 cents smaller on average than those ending with a non-chord tone (95% confidence interval = [1,7]). In terms of singer identity, singers three (10 cents smaller, 95% confidence interval = [5,14]), four (7 cents larger, 95% confidence interval = [3,12]), and five (6 cents smaller, 95% confidence interval = [2,11]) were statistically different than the baseline, singer six. For the professional group, the linear regression ( $R^2=0.04, p < 0.0001$ ) revealed that there was no statistically significant difference between ascending and descending whole tones. There was no statistically significant difference in interval size for the different whole tone conditions regarding chord tones and non-chord tones. There were, however, significant effects for singer identity in the average whole tone size for singers two (14 cents larger, 95% confidence interval = [0.03,9]) and five (95% confidence interval = [3,12]) compared to singer six. Overall, there was no statistically significant effect for the presence of accompaniment in either the non-professional or professional groups.

A second linear regression ( $R^2=0.02, p < 0.0001$ ), where both groups were combined and singer identity was replaced by group identity, produced significant results for all conditions except for *a cappella* versus accompanied. As with the first regression, there was no statistically significant impact of accompaniment on interval size. The descending whole tones were on average 3 cents smaller than ascending one (95% confidence interval = [1,4]). The whole tones ending with a chord tone were on average 3 cents smaller than those ending with a non-chord tone (95% confidence interval = [1,5]). Overall, the professional group's whole tones were on average 3 cents larger than the non-professional group (95% confidence interval = [2,5]).

The results of the ANOVA analysis described in Section 4.1.1.4 revealed a significant effect for the interaction between direction, intervallic conditions, and singer ( $F(20,143) = 3.22, p < 0.01$ ). This suggests that some of the singers sung the combination of ascending versus descending and different intervallic conditions significantly differently than other singers. It should be noted, however, that the related two-way interactions (singer and direction, singer

and intervallic condition, direction and intervallic condition) were not significant, which makes it difficult to directly assess the implications of this three-way interaction.

#### 4.1.2.2.2 Slope and Curvature

A summary of the slope and curvature values for the endings of the first note in each whole tone across the various whole tone contexts is shown for both groups of singers in Table 4.1.16. The means and standard deviations for the slope values, approximately cents/second, from the 1<sup>st</sup> DCT coefficient run on the original F<sub>0</sub> trace for the non-professional group are shown in Table 4.1.17 and for the professional group in Table 4.1.19. The results from the 1<sup>st</sup> DCT run on the F<sub>0</sub> trace with a 200 ms moving average applied to it are shown in Table 4.1.18 for the non-professional group and in Table 4.1.20 for the professional group. The means and standard deviations for the curvature values, approximately cents/second<sup>2</sup>, from the 2<sup>nd</sup> DCT coefficient run on the original F<sub>0</sub> trace are shown in Table 4.1.21 for the non-professional and in Table 4.1.23 for the professional group. The results from the 2<sup>nd</sup> DCT coefficient run on the F<sub>0</sub> trace with a 200 ms moving average applied to it are shown in Table 4.1.22 for the non-professional group and in Table 4.1.24 for the professional group.

Whole Tone Conditions (Number of Instances)	Non-professional Singers				Professional Singers			
	<i>A cappella</i>		Accompanied		<i>A cappella</i>		Accompanied	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Slope, F <sub>0</sub> trace, ascending (198)	46.1	89.6	45.9	93.0	17.5	67.2	16.3	85.7
Slope, MA, ascending (198)	41.2	57.0	36.9	54.4	17.0	54.5	20.9	50.8
Slope, F <sub>0</sub> trace, descending (234)	-28.8	91.8	-13.6	82.4	-36.7	114.9	-23	133.8
Slope, MA, descending (234)	0.1	41.4	-3.3	37.2	1.2	46.8	-5.3	54.8
Curvature, F <sub>0</sub> trace, ascending (198)	-126.8	606.4	-62.0	617.4	-108.9	682.2	-60.6	680.8
Curvature, MA, ascending (198)	85.4	268.7	56.5	308.7	56.5	250.8	34.5	243.9
Curvature, F <sub>0</sub> trace, descending (234)	-246.7	571.9	-136.4	589.8	-419.0	891.6	-351.4	938.5
Curvature, MA, descending (234)	9.6	191.6	27.6	172.3	13.8	272.9	51.0	296.3

Table 4.1.16 Summary of the means and standard deviations of the slope and curvature for the two subject groups across all of the whole tones used in this experiment.

Whole Tone Conditions (Number of Instances)	Non-professional Group		<i>A Cappella</i>		Accompanied	
	Mean	SD	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	71.0	116.4	71.3	101.1		
Chord tone – chord tone, descending (54)	-1.8	102.2	0.3	82.2		
Chord tone – non-chord tone, ascending (36)	49.7	39.5	27	123.7		
Chord tone – non-chord tone, descending (108)	-36.7	96.8	-13.2	77.0		
Non-chord tone – chord tone, ascending (72)	19.5	67.8	30.0	54.4		
Non-chord tone – chord tone, descending (90)	-37.0	71.2	-22.6	71.1		

Table 4.1.17: Non-professional group's whole tone slope values calculated on the original F<sub>0</sub> trace.

Non-professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	53.1	71.8	46.3	65.7
Chord tone – chord tone, descending (54)	7.8	38.1	2.7	32.4
Chord tone – non-chord tone, ascending (36)	28.6	39.3	22.6	42.8
Chord tone – non-chord tone, descending (108)	-1.4	38.8	-8.3	38.2
Non-chord tone – chord tone, ascending (72)	35.5	44.9	34.5	45.3
Non-chord tone – chord tone, descending (90)	-1.2	46.2	-1.1	36

Table 4.1.18: Non-professional group's whole tone slope values calculated on the  $F_0$  trace with a moving average applied.

Professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	35.5	79	38.7	120.7
Chord tone – chord tone, descending (54)	-24	95.4	7.5	104.6
Chord tone – non-chord tone, ascending (36)	35.2	57.7	16.5	42.0
Chord tone – non-chord tone, descending (108)	-37.1	102.7	-36.5	136.0
Non-chord tone – chord tone, ascending (72)	-9.4	47.8	-6.3	45.2
Non-chord tone – chord tone, descending (90)	-47.2	136.7	-31.6	101.2

Table 4.1.19: Professional group's whole tone slope values calculated on the original  $F_0$  trace.

Professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	22	75.9	24.6	68.6
Chord tone – chord tone, descending (54)	-3	53.5	5.2	33.7
Chord tone – non-chord tone, ascending (36)	22.9	38.2	23.2	36.1
Chord tone – non-chord tone, descending (108)	2.7	49.0	0	56.1
Non-chord tone – chord tone, ascending (72)	9.1	30.0	16	33.5
Non-chord tone – chord tone, descending (90)	0.5	33.6	-9.5	51.5

Table 4.1.20: Professional group's whole tone slope values calculated on the  $F_0$  trace with a moving average applied.

Non-professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	-295.5	847.2	-180.1	854.9
Chord tone – chord tone, descending (54)	-28.1	398.9	-23.8	348.3
Chord tone – non-chord tone, ascending (36)	84.2	203.9	34.3	338.8
Chord tone – non-chord tone, descending (108)	-329.6	637.9	-165.3	496.9
Non-chord tone – chord tone, ascending (72)	-62.2	364.8	7.8	385.1
Non-chord tone – chord tone, descending (90)	-284.6	475.9	-181.1	474.0

Table 4.1.21: Non-professional group's whole tone curvature values calculated on the original  $F_0$  trace.

Non-professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	155.5	379.7	121.9	296.9
Chord tone – chord tone, descending (54)	1.1	235	27.3	190.9
Chord tone – non-chord tone, ascending (36)	21.6	121.6	-69.3	498.8
Chord tone – non-chord tone, descending (108)	21.7	192.4	38.1	178.7
Non-chord tone – chord tone, ascending (72)	47	149.1	55.1	131.5
Non-chord tone – chord tone, descending (90)	-16.7	143.9	9	128.5

Table 4.1.22: Non-professional group's whole tone curvature values calculated on the  $F_0$  trace with a moving average.

Professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	-284	958.8	-211.8	968.8
Chord tone – chord tone, descending (54)	-17.1	451.9	-30.7	565.8
Chord tone – non-chord tone, ascending (36)	39	295	135.1	309.6
Chord tone – non-chord tone, descending (108)	-391.1	846.2	-390.8	928.7
Non-chord tone – chord tone, ascending (72)	-9.2	398.4	-7.4	372.1
Non-chord tone – chord tone, descending (90)	-445	722.8	-355.3	752.6

Table 4.1.23: Professional group's whole tone curvature values calculated on the original  $F_0$  trace.

Professional Group	<i>A Cappella</i>	Accompanied		
Whole Tone Conditions (Number of Instances)	Mean	SD	Mean	SD
Chord tone – chord tone, ascending (72)	102.4	345.4	73	323.3
Chord tone – chord tone, descending (54)	-15.1	253.8	60.7	338.6
Chord tone – non-chord tone, ascending (36)	43.8	166.2	-32.5	203.6
Chord tone – non-chord tone, descending (108)	-13.5	231.1	9.9	232.3
Non-chord tone – chord tone, ascending (72)	15.2	148	29.2	148.4
Non-chord tone – chord tone, descending (90)	21.3	277.6	57.8	263.4

Table 4.1.24: Professional group's whole tone curvature values calculated on the  $F_0$  trace with a moving average applied.

The following box and whisker plots show the range of the 1<sup>st</sup> DCT coefficient values, measured in an approximation of cents/second, and the 2<sup>nd</sup> DCT coefficient values, measured in an approximation of cents/second<sup>2</sup>. The plots show the ascending versus descending and *a cappella* versus accompanied conditions for each singer run on both the original  $F_0$  trace for the non-professional (Figure 4.1.27 for the 1<sup>st</sup> DCT and for Figure 4.1.33 the 2<sup>nd</sup> DCT) and professional (Figure 4.1.28 for the 1<sup>st</sup> DCT and Figure 4.1.34 for the 2<sup>nd</sup> DCT) groups and on results of applying a 200 ms moving average to the original  $F_0$  trace for

the non-professionals (Figure 4.1.29 for the 1<sup>st</sup> DCT and Figure 4.1.35 for the 2<sup>nd</sup> DCT) and professionals (Figure 4.1.30 for the 1<sup>st</sup> DCT and Figure 4.1.36 for the 2<sup>nd</sup> DCT). The plots in Figure 4.1.31 and Figure 4.1.32 show the 1<sup>st</sup> DCT coefficient for each whole tone condition across all of the singers in the non-profession and professional group, respectively. The plots in Figure 4.1.37 and Figure 4.1.38 show the 2<sup>nd</sup> DCT coefficient for each whole tone condition across all of the singers in the groups. As with the semitones, linear regressions analyses were also run on the whole tones' slope and curvature data. As with regression analysis on the semitones, the  $R^2$  values for these regressions are quite small.

The first linear regression analysis of the original  $F_0$  trace's 1<sup>st</sup> DCT coefficients in the non-professional group ( $R^2=0.12, p < 0.01$ ) showed significant effects for whether or not the whole tone started or ended with a chord tone, as well as for singer identity for all of the singers except for singer two. The average slope of whole tones starting on a chord tone was 17 cents/second smaller than those starting on a non-chord tone (95% confidence interval = [5,26]), and those ending on a chord tone were 30 cents/second smaller than those ending on a non-chord tone (95% confidence interval = [19,41]). All of the singers' average curvatures were smaller than singer six's average slope: singer one's average slope was 55 cents/second smaller (95% confidence interval = [40,71]), singer two's average slope was 30 cents/second smaller (95% confidence interval = [14,46]), singer four's average slope was 47 cents/second smaller (95% confidence interval = [32,62]), and singer five's average slope was 44 cents/second smaller (95% confidence interval = [28,59]). In the professional group ( $R^2=0.08, p < 0.0001$ ), there were also significant effects for intervallic direction and for singer identity for singer four. Descending whole tones' slopes were on average 25 cents/second larger than ascending whole tones' (95% confidence interval = [13,36]). Singer four's average slope was on average 38 cents/second larger than singer six's (95% confidence interval = [19,56]).

In the second linear regression analysis, across both groups, there were significant effects for intervallic direction and whether or not the whole tone started or ended with a chord note ( $R^2=0.02, p < 0.001$ ). Descending intervals' slopes were on average 9 cents/second larger than ascending intervals (95% confidence interval = [2,17]). The average slope of whole tones starting on a chord tone was 10 cents/second smaller than those starting on a non-chord tone (95% confidence interval = [2,20]), and those ending on a chord tone were 20 cents/second smaller than those ending on a non-chord tone (95% confidence interval = [10,30]).

cents/second smaller than those ending on a non-chord tone (95% confidence interval = [11,29]). Overall there were no significant effects for accompaniment or group identity.

In the first regression analysis of the slope data for the non-professional singers ( $R^2=0.17$ ,  $p < 0.0001$ ) there were significant effects for direction, whether the whole tone ended with a chord tone, and singer identity. There were not, however, significant effects for the presence of accompaniment or whether the whole tone started with a chord tone or non-chord tone. The slope of the descending whole tones was on average 18 cents/second smaller than the slope of the ascending ones (95% confidence interval = [14,23]). Whole tones ending with a chord were 12 cents/second on average smaller than those ending with a non-chord tone (95% confidence interval = [6,18]), In terms of singer identity, the average slopes of singers one (33 cents/second, 95% confidence interval = [24,41]), four (23 cents/second smaller, 95% confidence interval = [15,31]), and five (10 cents/second smaller, 95% confidence interval = [2,18]) were significantly different than singer six's average slope.

For the professional singers ( $R^2=0.17$ ,  $p < 0.0001$ ), there were significant effects for direction: the descending whole tones' slopes were on average 8 the ascending whole tones' (95% confidence interval = [3,13]). There was also a significant effect for whole tones ending on a chord-tone, whose slope was on average 15 cents/second smaller than those whole tones ending on a non-chord tone (95% confidence interval = [9,21]). There were no significant effects for whole tones beginning on chord tones versus non-chord tones or for the presence of accompaniment. There were singer identity effects for singers three (26 cents/second larger, 95% confidence interval = [17,34]) and four (34 cents/second larger, 95% confidence interval = [25,42]) in comparison to the baseline, singer six.

In the second linear regression analysis ( $R^2=0.05$ ,  $p < 0.0001$ ), on the slope data for the both groups, there were significant effects for direction and for whole tones ending on a chord tone versus a non-chord tone. Descending whole tones' slopes were on average 11 cents/second smaller than the ascending whole tones' slopes (95% confidence interval = [8,15]). The slopes of the whole tones ending on a chord tone were on average 13 cents/second smaller than those starting on a non-chord tone (95% confidence interval = [9,18]). There were no significant effects for accompaniment or group identity.

In the linear regression analysis of the 2<sup>nd</sup> DCT coefficients calculated on the original F<sub>0</sub> trace for non-professionals, there was no significant effect for intervallic direction or whether or not a whole tone started on a chord tone. There were, however, significant effects for whether or not the whole tone ended on a chord tone and for singer identity for all of the singers ( $R^2=0.09, p < 0.0001$ ). Whole tones ending on a chord tone's curvature was on average 103 cents/second<sup>2</sup> smaller than whole tones ending on a non-chord tone (95% confidence interval = [31,175]). All of the singers' average curvatures were smaller than singer six's average curvature: singer one's average curvature was 375 cents/second<sup>2</sup> smaller (95% confidence interval = [274,476]), singer two's average curvature was 152 cents/second<sup>2</sup> smaller (95% confidence interval = [51,254]), singer three's average curvature was 114 cents/second<sup>2</sup> smaller (95% confidence interval = [13,215]), singer four's average curvature was 308 cents/second<sup>2</sup> smaller (95% confidence interval = [207,409]), and singer five's average curvature was 254 cents/second<sup>2</sup> smaller (95% confidence interval = [153,579]).

For the professional group, there were significant effects for intervallic direction and for singers two, three, and four ( $R^2=0.13, p < 0.0001$ ). Descending whole's average curvature was 176 cents/second<sup>2</sup> larger than ascending semitones curvature (95% confidence interval = [92,260]). Singer two's curvature was on average 168              second<sup>2</sup> smaller than singer six's curvature (95% confidence interval = [29,306]), whereas singer three's curvature was on average 441 cents/second<sup>2</sup> larger than singer six's curvature (95% confidence interval = [302,58              268 cents/second<sup>2</sup> (95% confidence interval = [129,406]) larger than singer six's curvature.

For the regression run across both groups, there were only significant effects for intervallic direction, the ending notes, and for group identity ( $R^2=0.02, p < 0.0001$ ). Descending whole tones curvature was on average 101 cents/second<sup>2</sup> larger than the ascending whole tones (95% confidence interval = [46,156]). The whole tones ending on a chord tone curvature was on average 79 cents/second<sup>2</sup> smaller than the curvature of those ending on a non-chord tone (95% confidence interval = [13,143]). The profession group's curvature was on average 94 cents/second<sup>2</sup> larger than the non-professional group's curvature (95% confidence interval = [42,147]). There were no significant effects for accompaniment in any of the regressions.

The linear regression analysis of the 2<sup>nd</sup> DCT coefficients calculated on the result of applying a 200 ms moving average to the F<sub>0</sub> trace to the non-professional singers' whole tones ( $R^2=0.11, p < 0.0001$ ) revealed effects for direction and whether the whole tones started or ended on a chord tone. The curvature of the descending whole tones was on average 33 cents/second<sup>2</sup> smaller than ascending whole tones (95% confidence interval = [7,59]). Whole tones starting on a chord tone's curvatures were on average 35 cents/second<sup>2</sup> smaller than those starting on a non-chord tone (95% confidence interval = [5,65]). Similarly, whole tones ending on a chord tone's curvatures were on average 79 cents/second<sup>2</sup> smaller than whole tones ending on a non-chord tone's curvature (95% confidence interval = [49,110]). There were no significant effects for the presence of accompaniment or the starting note of the whole tone. In terms of singer identity, there were significant effects for all of the singers in relation to the baseline, singer six. On average, singer one was 167 cents/second<sup>2</sup> smaller than singer six (95% confidence interval = [123,210]), singer two was 68 cents/second<sup>2</sup> smaller than singer six (95% confidence interval = [25,111]), singer three was 99 cents/second<sup>2</sup> smaller than singer six (95% confidence interval = [56,142]), singer four was 146 cents/second<sup>2</sup> smaller than singer six (95% confidence interval = [102,189]), and singer five was 125 cents/second<sup>2</sup> smaller than singer six (95% confidence interval = [82,169]).

For the professional singers ( $R^2=0.20, p < 0.0001$ ), there were only significant effects for whether the whole tone ended on a chord tone or non-chord tone, as well as for singers three and four. The curvature of the whole tones ending on a chord tone was on average 65 cents/second<sup>2</sup> smaller than those ending on a non-chord tone (95% confidence interval = [34,97]). Singer three's average curvature was 130 cents/second<sup>2</sup> larger than singer six's curvature (95% confidence interval = [87,174]), and singer four's average curvature was 217 cents/second<sup>2</sup> larger than singer six's curvature (95% confidence interval = [173,261]).

In the second regression analysis on the curvature data of both groups combined ( $R^2=0.03, p < 0.0001$ ), there were no significant effects of accompaniment, direction, and the starting note of the whole tone. There were, however, significant effects for the ending notes of the whole tone and group identity. Whole tones ending on a chord tone were on average 72 cents/second<sup>2</sup> smaller than those ending on a non-chord tone (95% confidence interval = [49,96]). The curvatures of professional group's whole tones were on average 19

cents/second<sup>2</sup> larger than the non-professional group's curvatures (95% confidence interval = [0.3,38]).

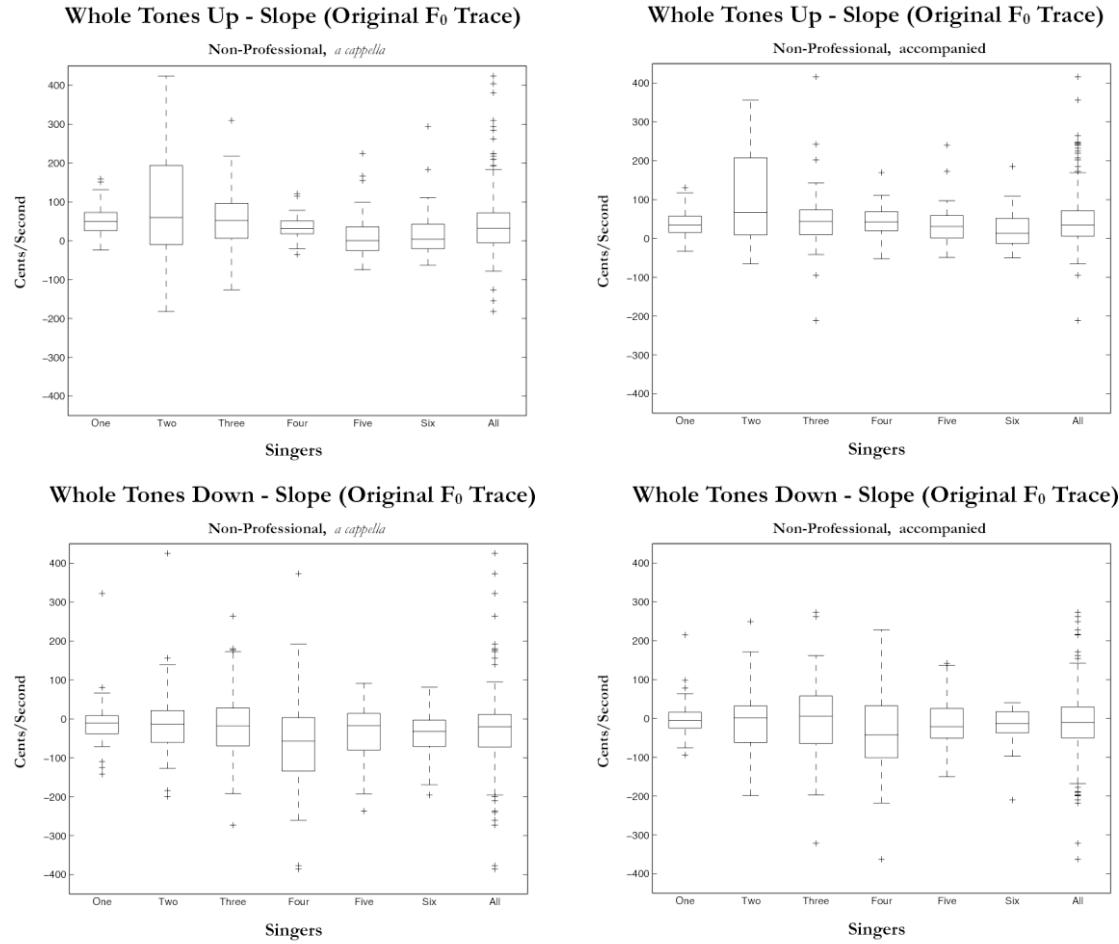


Figure 4.1.27: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the  $F_0$  trace of the first note of all of the whole tones performed by the non-professional group. Each plot shows the results for the six non-professional singers individually and the mean across all of the singers. The plots on the left show the 1st DCT coefficient values from the *a cappella* performances, and the plots on the right show the 1st DCT coefficient values from performances with accompaniment. The plots on the top show the 1st DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 1st DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second.

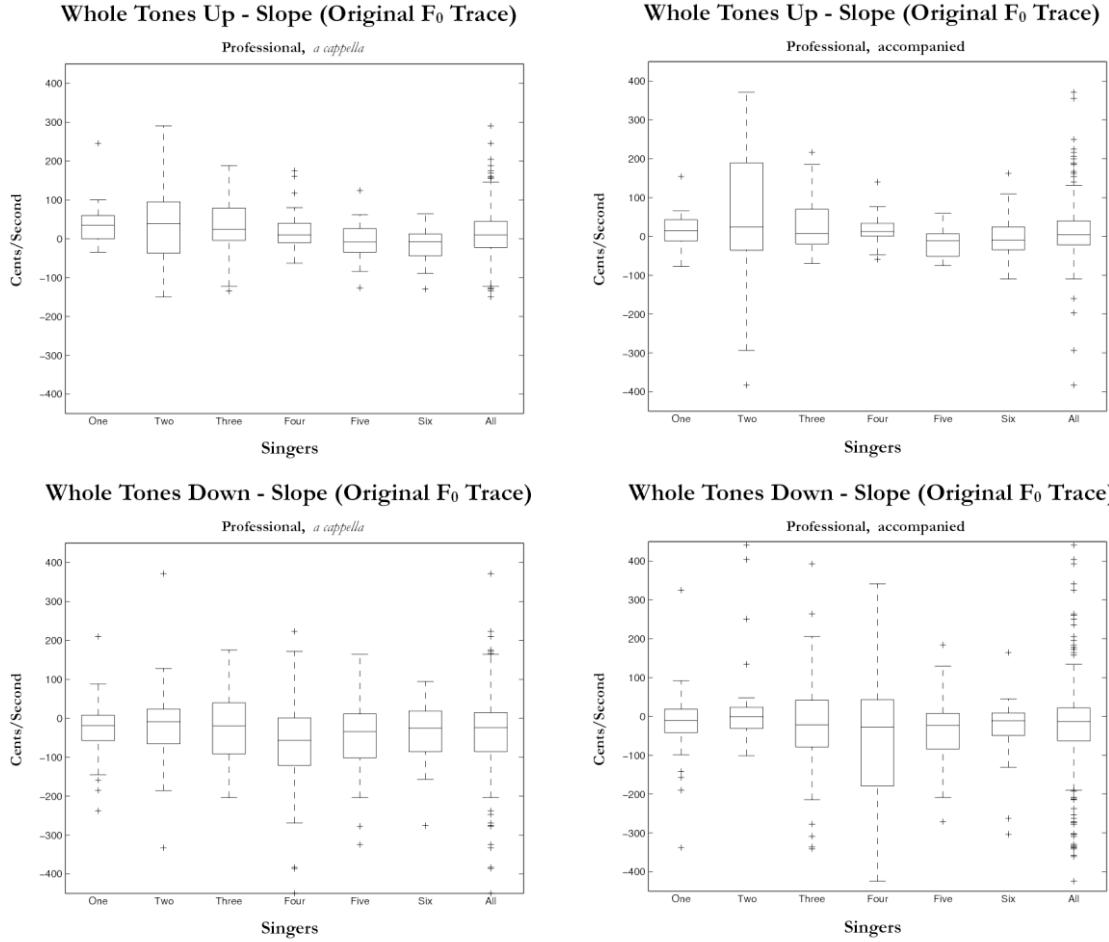


Figure 4.1.28: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the last 250 ms of the  $F_0$  trace of the first note of all of the whole tones performed by the professional group. Each plot shows the results for the six professional singers individually and the mean across all of the singers. The plots on the left show the 1st DCT coefficient values for the *a cappella* performances, and the plots on the right show the 1st DCT coefficient values for performances with accompaniment. The plots on the top show the 1st DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 1st DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second.

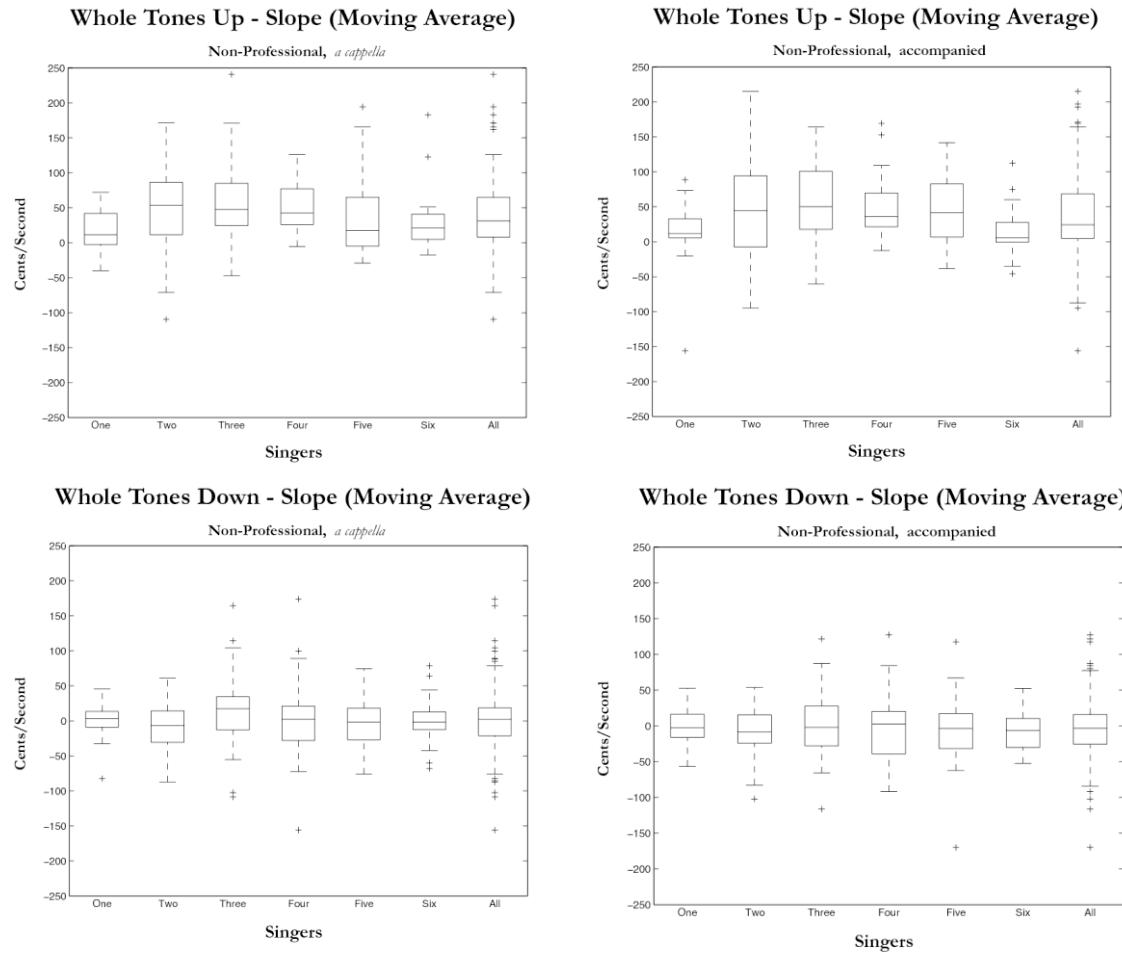


Figure 4.1.29: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the last 150 ms of the  $F_0$  trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the whole tones performed by non-professional group. Each plot shows the results for the six singers individually and the mean across all of the non-professional singers. The plots on the left show the 1st DCT coefficient values for the *a cappella* performances, and the plots on the right show the 1st DCT coefficient values for performances with accompaniment. The plots on the top show the 1st DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 1st DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second.

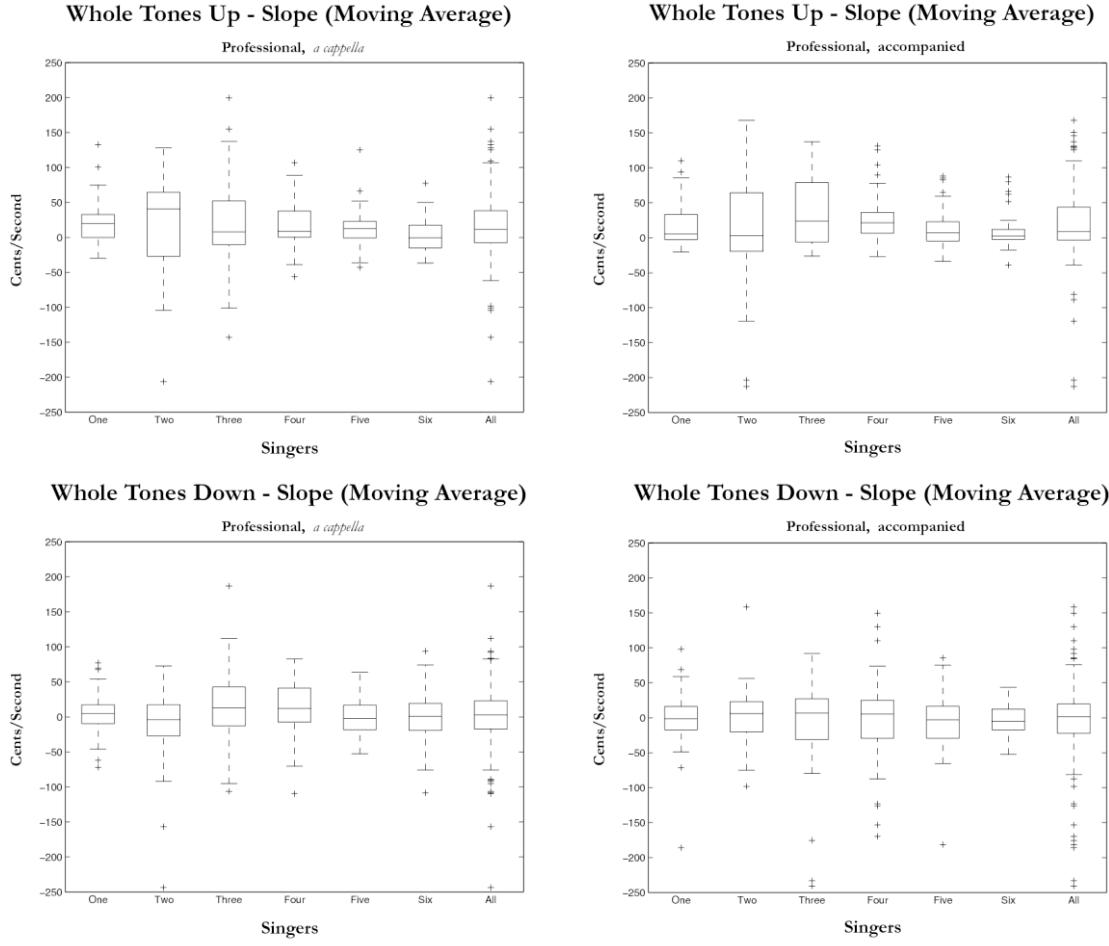


Figure 4.1.30: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient, approximating slope, run on the last 150 ms of the  $F_0$  trace (smoothed by results of applying a 200 ms moving average) of the first note of all of the whole tones performed by the professional group. Each plot shows the results for the six singers individually and the mean across all of the professional singers. The plots on the left show the 1st DCT coefficient values for the *a cappella* performances, and the plots on the right show the 1st DCT coefficient values for performances with accompaniment. The plots on the top show the 1st DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 1st DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second.

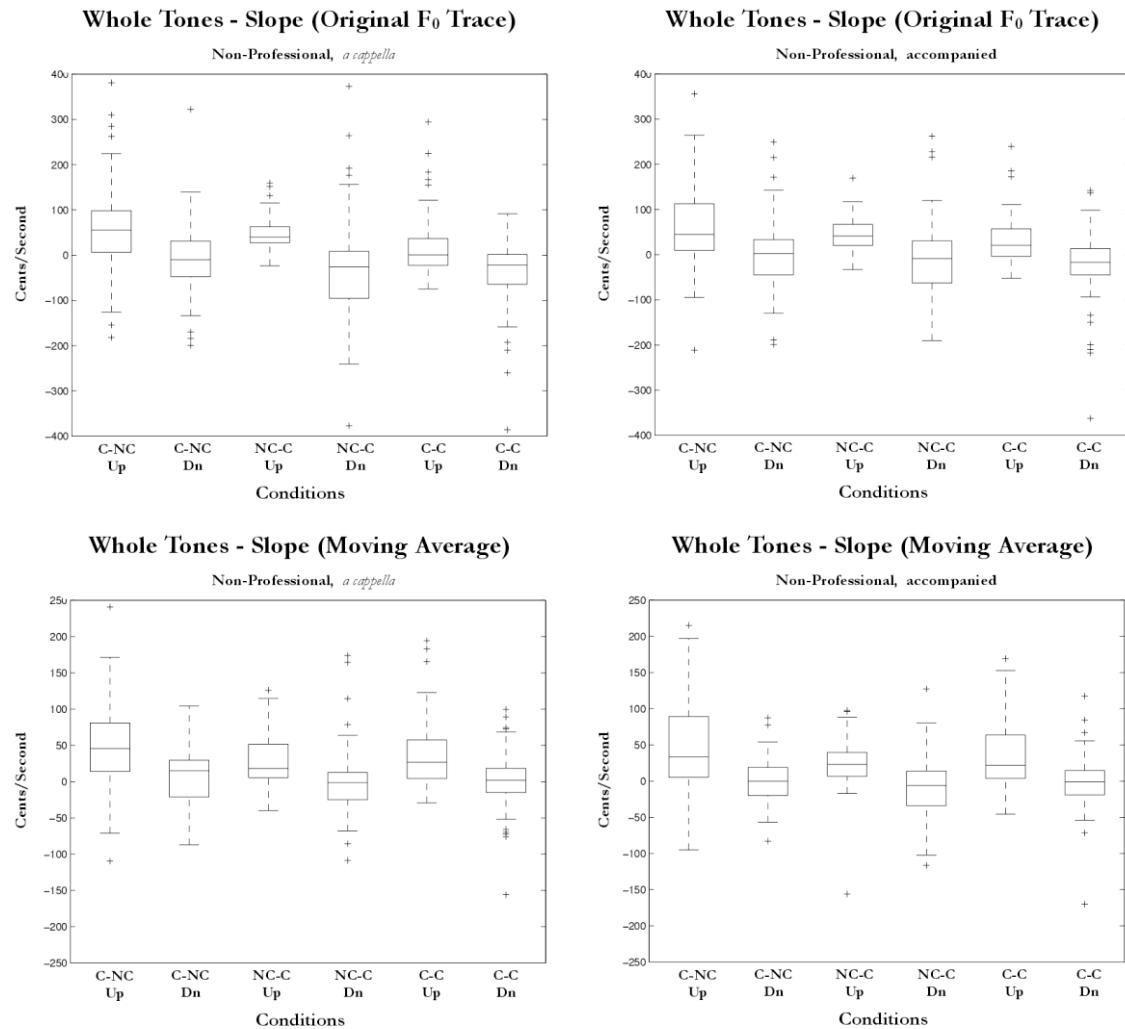


Figure 4.1.31: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of each whole tone interval for each condition across all non-professional singers. The plots on the left show the 1st DCT coefficient values for the *a cappella* performances, and the plots on the right show the 1st DCT coefficient values for performances with accompaniment. The plots on the top show the values of the 1st DCT run on the original  $F_0$  trace, while the plots on the bottom show the values of the 1st DCT coefficient run on the  $F_0$  trace smoothed by applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second.

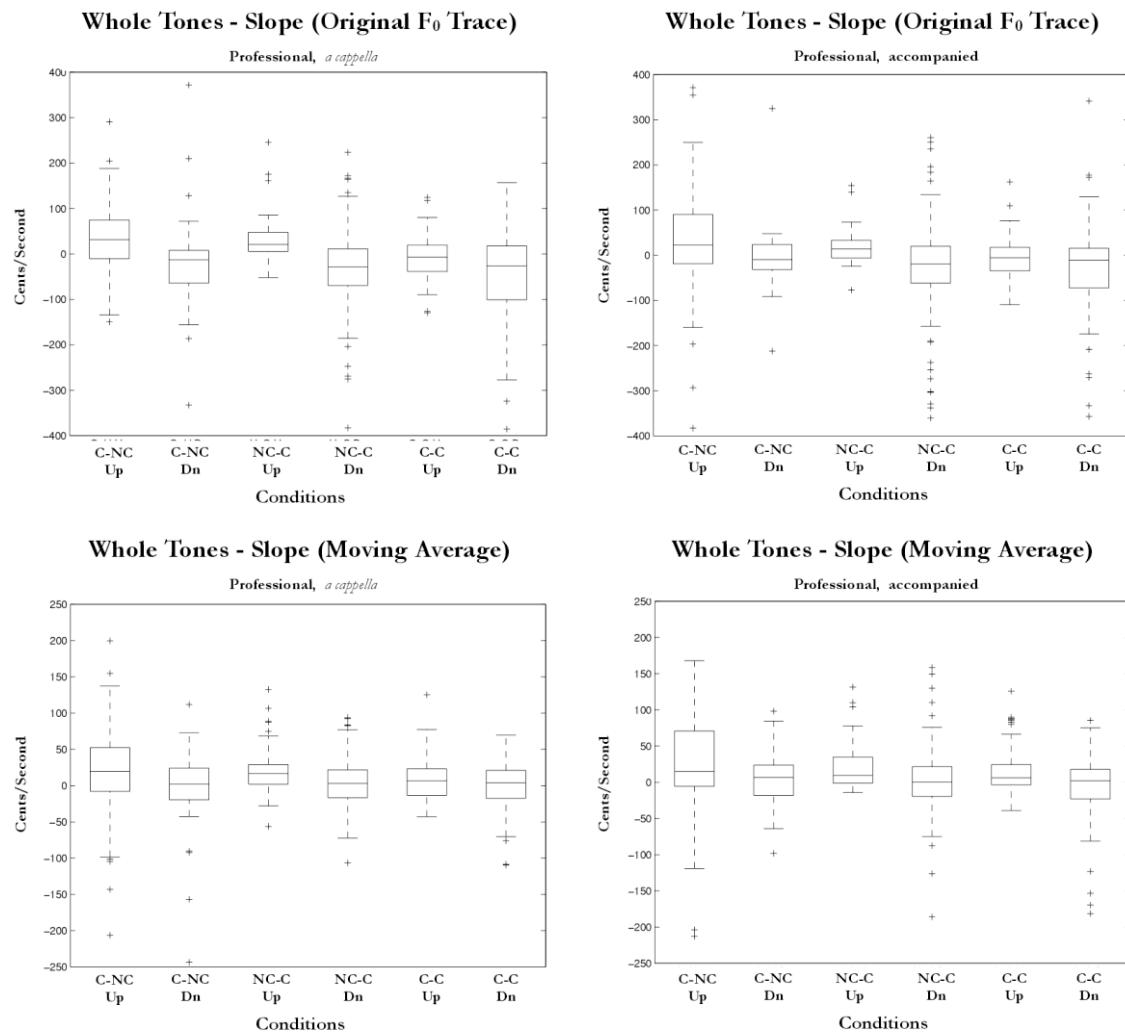
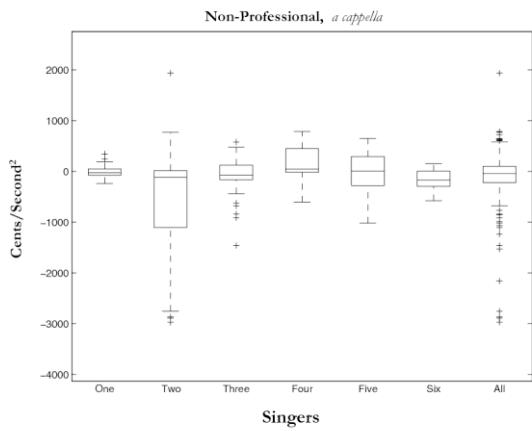
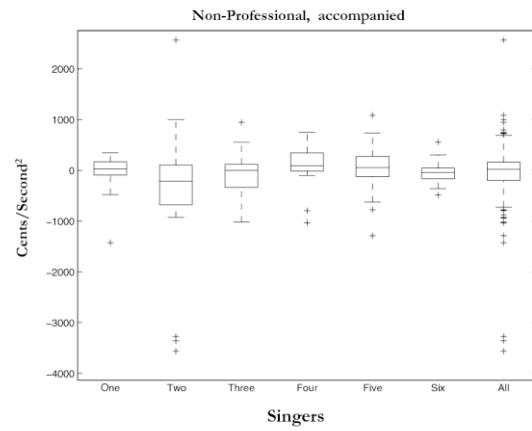


Figure 4.1.32: Box and whisker plots of the 1st discrete cosine transform (DCT) coefficient (approximating slope) run on the end of the first note of each whole tone interval for each condition across all professional singers. The plots on the left show the 1st DCT coefficient values for the *a cappella* performances, and the plots on the right show the 1st DCT coefficient values for performances with accompaniment. The plots on the top show the values of the 1st DCT coefficient run on the original F<sub>0</sub> trace, while the plots on the bottom show the values of 1st DCT coefficient run on the F<sub>0</sub> trace smoothed by applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second.

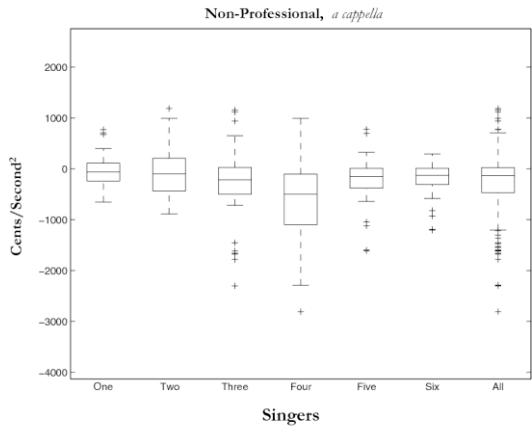
**Whole Tones Up - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones Up - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones Down - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones Down - Curvature (Original F<sub>0</sub> Trace)**

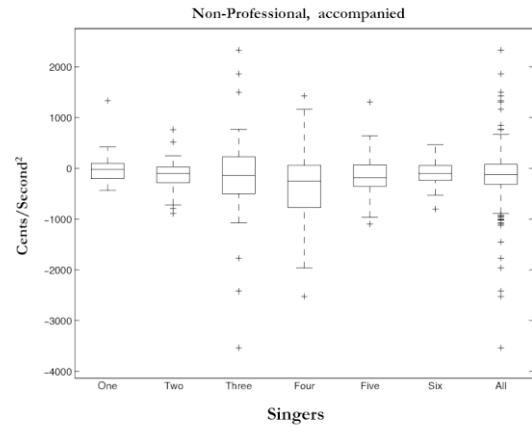


Figure 4.1.33: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the F<sub>0</sub> trace of the first note of all of the whole tones performed by the non-professional group. Each plot shows the results for the non-professional six singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

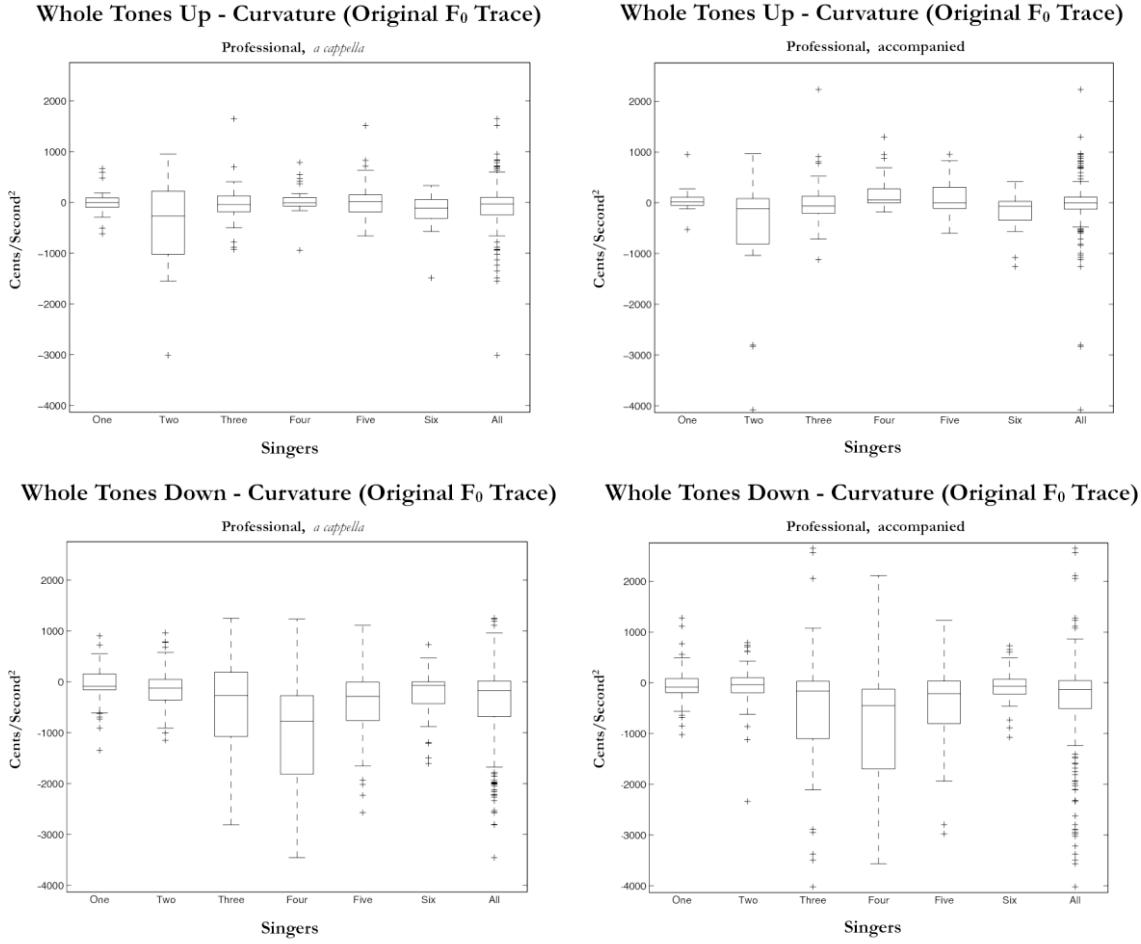


Figure 4.1.34: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 250 ms of the F<sub>0</sub> trace of the first note of all of the whole tones performed by the professional group. Each plot shows the results for the professional six singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

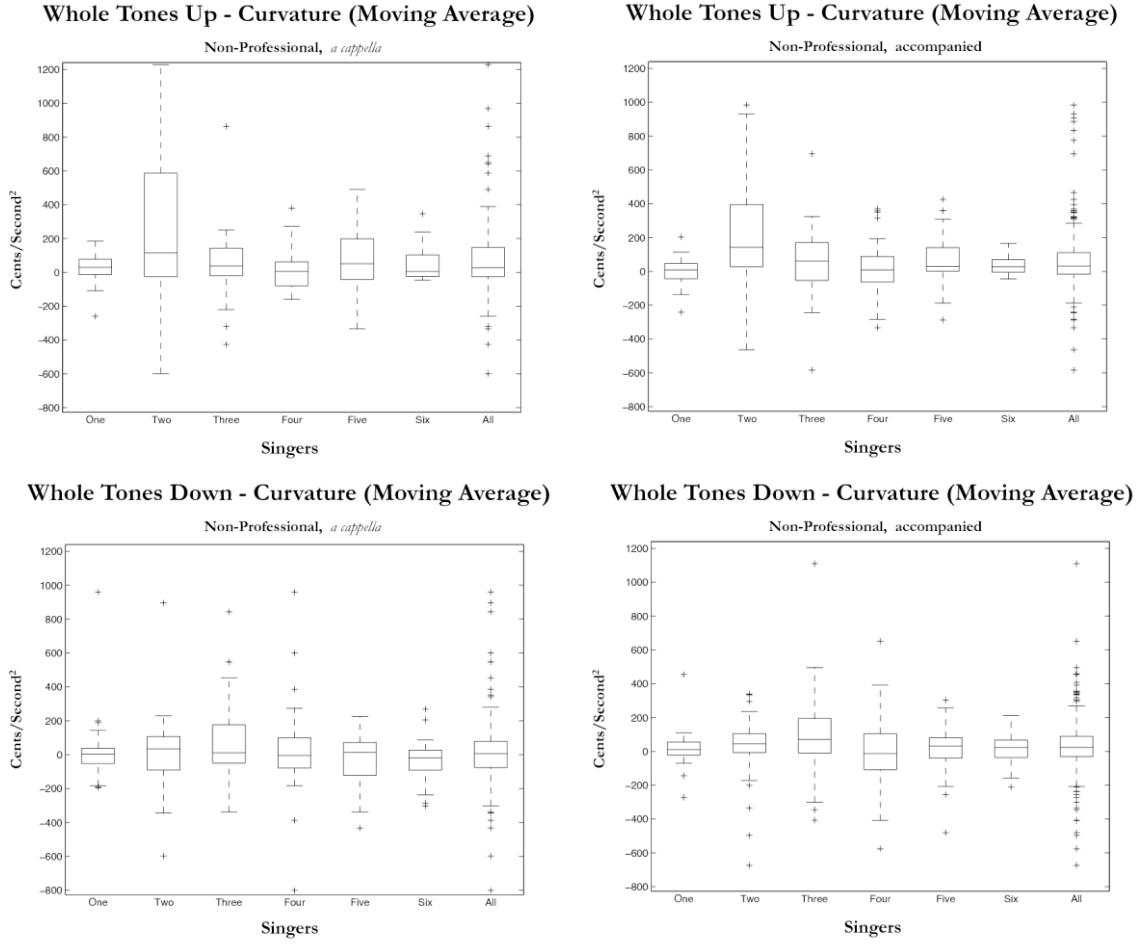


Figure 4.1.35: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient, (approximating curvature) run on the last 150 ms of the  $F_0$  trace (smoothed by applying a 200 ms moving average of the first note) of the first note of all of the whole tones performed by the non-professional group. Each plot shows the results for the six non-professional singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

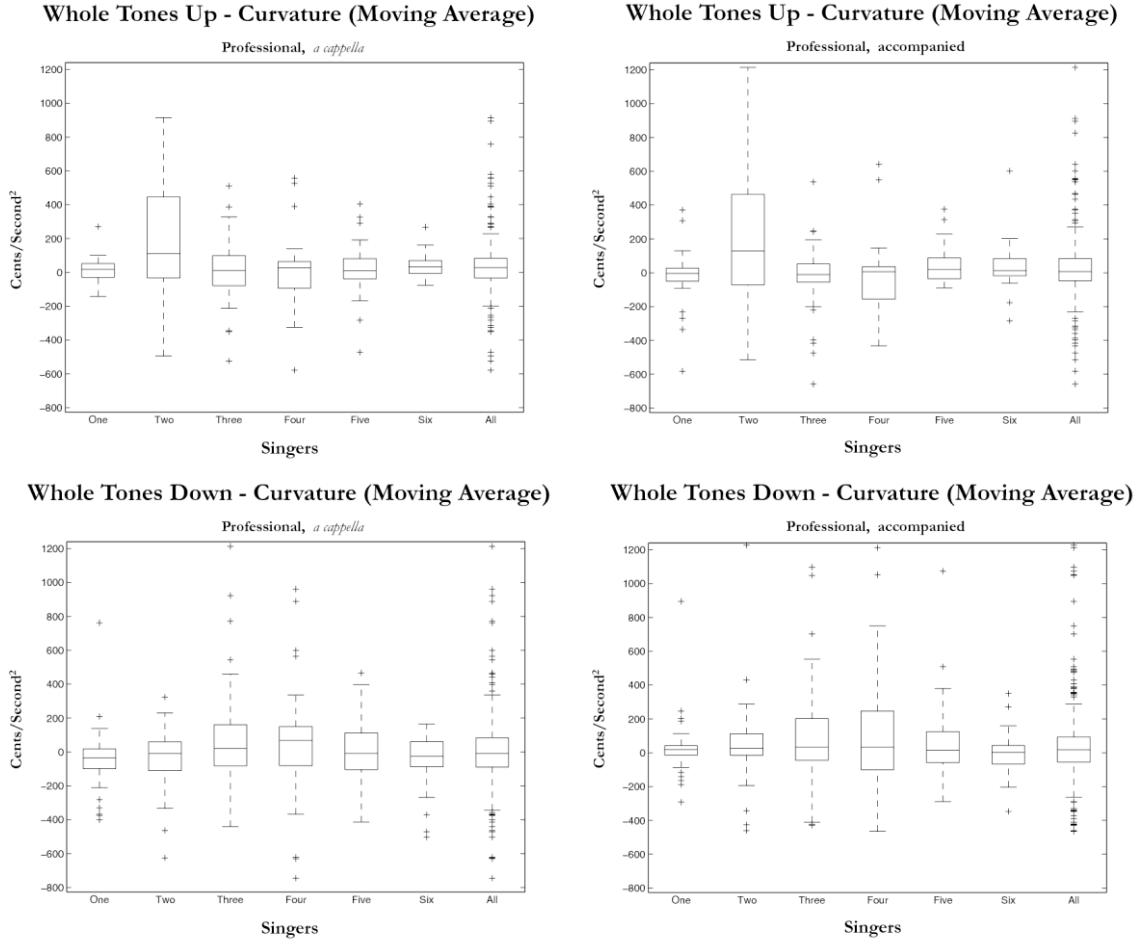
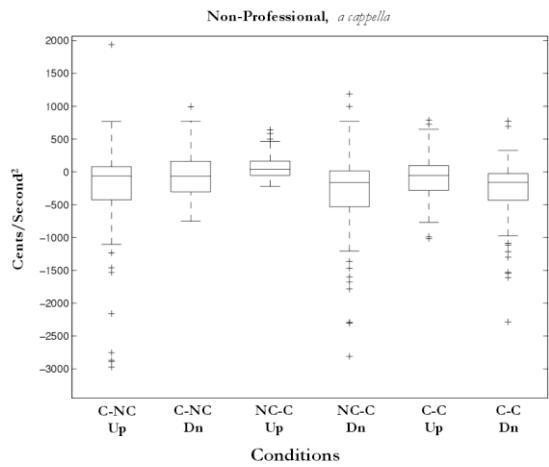
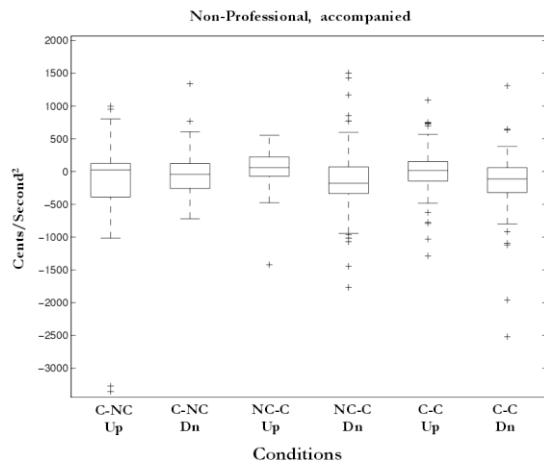


Figure 4.1.36: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the last 150 ms of the  $F_0$  trace (smoothed by applying a 200 ms moving average) of the first note of all of the whole tones performed by the professional group. Each plot shows the results for the six professional singers individually and the mean across all of the singers. The plots on the left show the values of the 2<sup>nd</sup> DCT coefficient for the *a cappella* performances, and the plots on the right show the values of the 2<sup>nd</sup> DCT coefficient for performances with accompaniment. The plots on the top show the 2<sup>nd</sup> DCT coefficient values for the ascending whole tones, and the plots on the bottom show the 2<sup>nd</sup> DCT coefficient values for the descending whole tones. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

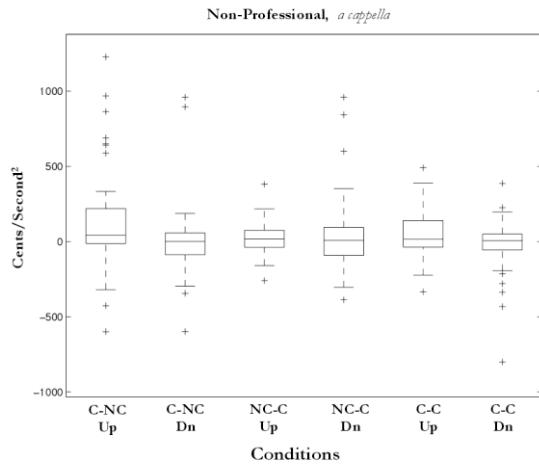
**Whole Tones - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones - Curvature (Moving Average)**



**Whole Tones - Curvature (Moving Average)**

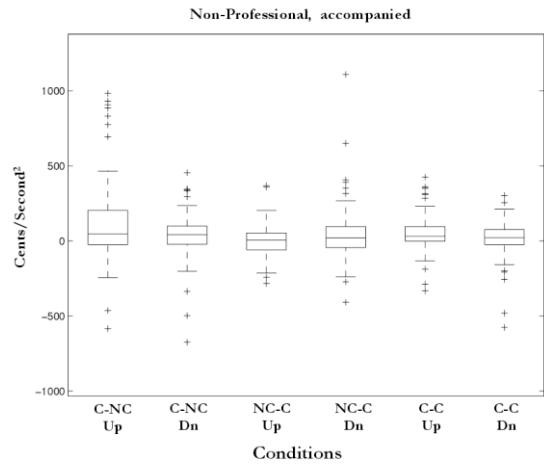
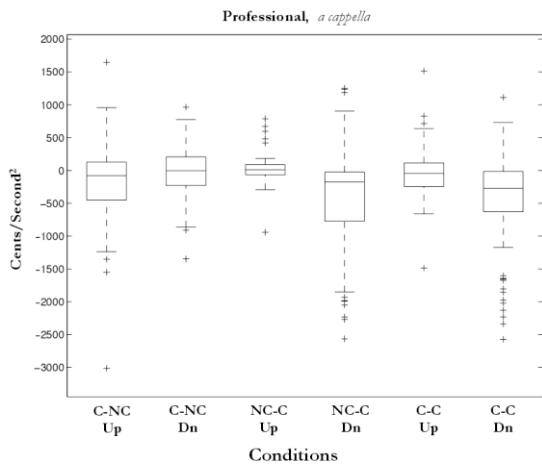
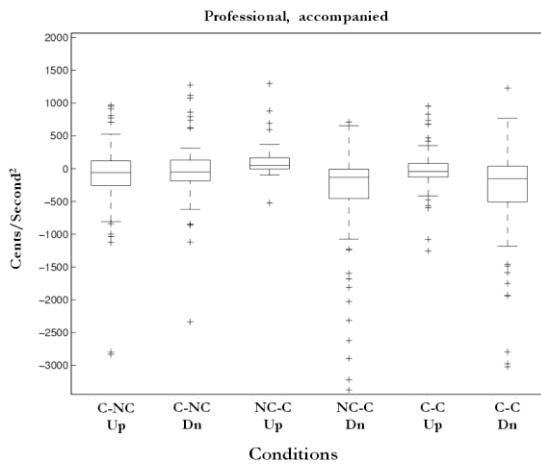


Figure 4.1.37: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of each whole tone interval for each condition across all non-professional singers. The plots on the left show the 2<sup>nd</sup> DCT coefficient values for the *a cappella* performances, and the plots on the right show the 2<sup>nd</sup> DCT coefficient values for performances with accompaniment. The plots on the top show the values of the 2<sup>nd</sup> DCT run on the original F<sub>0</sub> trace, while the plots on the bottom show the values of the 2<sup>nd</sup> DCT coefficient run on F<sub>0</sub> trace smoothed by applying a 200 ms moving average. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

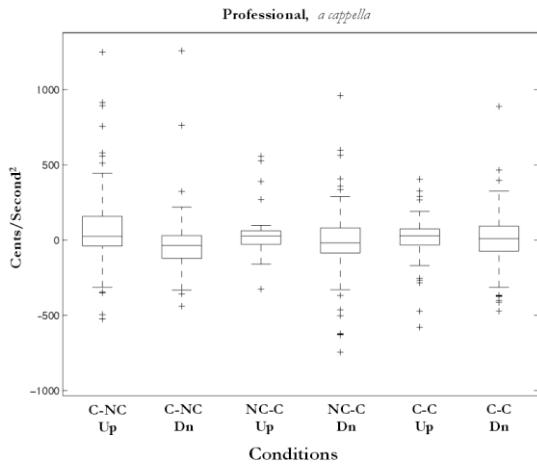
**Whole Tones - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones - Curvature (Original F<sub>0</sub> Trace)**



**Whole Tones - Curvature (Moving Average)**



**Whole Tones - Curvature (Moving Average)**

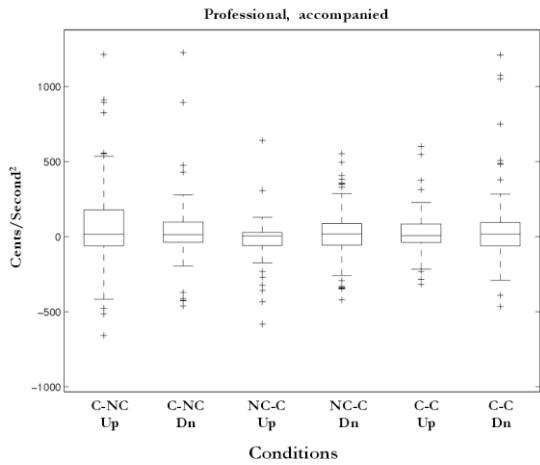


Figure 4.1.38: Box and whisker plots of the 2<sup>nd</sup> discrete cosine transform (DCT) coefficient (approximating curvature) run on the end of the first note of each whole tone interval for each condition across all professional singers. The plots on the left show the 2<sup>nd</sup> DCT coefficient values for the *a cappella* performances, and the plots on the right show the 2<sup>nd</sup> DCT coefficient values for performances with accompaniment. The plots on the top show the values of the 2<sup>nd</sup> DCT run on the original F<sub>0</sub> trace, while the plots on the bottom show the values of the 2<sup>nd</sup> DCT coefficient run on of the F<sub>0</sub> trace smoothed by results a 200 ms moving average. The units on the y-axis are an approximation of cents/second<sup>2</sup>.

### 4.1.3 Discussion

#### 4.1.3.1 Semitones

The mean and median semitone size values across all of the singers in both groups tended to be smaller than the 100 cent equal tempered semitone. The 50% confidence intervals for semitone size, shown in Figure 4.1.8 for the non-professional group and in Figure 4.1.9 for the professional group, encompass the 90 cent Pythagorean semitone, the 100 cent equal tempered semitone, and for some contexts, the 112 cent minor diatonic Just Intonation semitone. Only the descending non-B<sub>b</sub>-A semitones in the progression group encompass the 112 cent major diatonic Just Intonation semitone.

The only interval size measurement for which there was significant effect for the presence of accompaniment was in the non-professional group. The non-professional group's semitones were on average 3 cents smaller for *a cappella* renditions than for those with accompaniment. When the DCT was run on the F<sub>0</sub> trace with the moving average applied, there was a significant effect for accompaniment in the professional group, with the accompanied renditions having a smaller curvature on average than the *a cappella* ones. There was also a significant effect for intervallic direction, with the descending intervals having on average a larger curvature than the ascending ones. There were no specific effects for accompaniment in the non-professional group when the DCT was run on the result of the moving average, or for either group when the DCT was run on the original F<sub>0</sub> trace.

Both groups showed a significant effect for direction on interval size. The linear regression analysis showed that both the non-professional and the professional singers' descending semitones were smaller than their ascending ones. For the slope and curvature measurements, there was only a significant effect for curvature. Specifically, the curvature of descending semitones in the non-professional group, when the DCT was calculated on the original F<sub>0</sub> trace, descending semitones had, on average, a larger curvature than the ascending semitones. Likewise, the descending semitones in the professional group, when the DCT was calculated on the F<sub>0</sub> trace with a moving average applied to it, also had on average a larger curvature measurement than the ascending semitones.

In terms of the different types of semitones, only the non-professional group showed a significant effect for leading tones versus non-leading tone semitones, with the leading tones

being on average 10 cents smaller. In contrast, the professional group had significant effect for non A-B $\flat$ /B $\flat$ -A semitones versus the A-B $\flat$ /B $\flat$ -A semitones, with the non A-B $\flat$ /B $\flat$ -A semitones being on average 7 cents larger. For slope, as measured by the 1<sup>st</sup> DCT coefficient on the end of the first note of the semitone, there were significant effects for A-B $\flat$ /B $\flat$ -A versus other semitones when the DCT was run on the F<sub>0</sub> trace. Both the non-professional and professionals singers had a greater slope on average for the non-A-B $\flat$ /B $\flat$ -A semitones. When the DCT was run on the F<sub>0</sub> trace after a moving average had been applied to it, there was only a significant effect for semitones with a leading tone function in the professional group, which had on average a smaller slope for the semitones with a leading tone function than those without. For curvature measured on the original F<sub>0</sub> trace, there were significant effects for leading tone function and A-B $\flat$ /B $\flat$ -A semitones, with the semitones with a leading tone function having a smaller curvature than the semitones with other functions and the A-B $\flat$ /B $\flat$ -A semitones having a smaller curvature than semitones between different notes. The non-professional group also had a significant effect for the A-B $\flat$ /B $\flat$ -A semitones, which had a smaller curvature than the other semitones.

Overall, the non-professional group showed more of an effect for singer identity than the professional group, particularly for interval size, where only singer five was significantly different from its baseline, professional singer six. In the non-professional group, singers one, two, four, and five were all significantly different from their baseline, non-professional singer six. 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. In terms of slope, both groups had three singers who differed significantly from the baseline when the DCT was calculated on the original F<sub>0</sub> trace (singers one, three, and four in the professional group and singers one, two, and five in the non-professional group), and four singers who differed significantly from the baseline when the DCT was calculated on the F<sub>0</sub> trace after the moving average had been applied to it (singers one, two, three, and four in the professional group and singers one, two, four, and five in the non-professional group). There was only a significant effect for group identity for the slope values calculated on the original F<sub>0</sub> trace. For curvature in the professional group, the same three singers' curvature measurements were significantly different than the baseline (singers one, three, and four) in both calculations. Whereas all of the non-professional singers were significantly

different from the baseline when the DCT was calculated on the result of the moving average, only four differed significantly when it was run on the original  $F_0$  trace (singers one, three, four, and five). There was a group effect for curvature for both calculations, with the professional group having a larger curvature measurement on average in both cases.

Overall, the  $R^2$  values for the regressions were low, the highest value was 0.22 for the regression run on the professional groups' curvature data calculated on the  $F_0$  trace with a moving average applied to it, which indicates that the conditions evaluated only some of the variation in the data. The implications of this will be discussed in Section 4.3.

#### 4.1.3.2 Whole Tones

As with the semitone analysis, the  $R^2$  values for the regressions were low. The highest  $R^2$  value overall was 0.19, for the regression run on the professional group's curvature data calculated on the original  $F_0$  trace. In terms of whole tone size, the mean and median whole tone size values across all of the singers for the various conditions centered around the 200 cent equal tempered whole tone. The 50% confidence intervals for whole tone size, shown in Figure 4.1.25 for the non-professional group and in Figure 4.1.26 for the professional group, encompass the 204 cent Pythagorean/major Just Intonation whole tone and the 200 cent equal tempered whole tone. Only the *a cappella* and accompanied ascending whole tones between non-chord tones and chord tones, *a cappella* and accompanied ascending whole tones between two chord tones, and *a cappella* descending whole tones between two non-chord tones in the non-professional group encompass the 182 cents minor Just Intonation whole tone.

There was no significant effect for accompaniment in either group for any of the measurements: interval size, slope, or curvature. In terms of intervallic direction, only the non-professional group showed an effect for interval size, with their descending whole tones on average 5 cents larger than their ascending whole tones. Both groups showed a significant effect for slope. When the slope was calculated directly on the  $F_0$  trace, the professional group's average slope was larger for descending whole tones than ascending whole tones. The opposite was true for the non-professional group. When the slope was calculated from the  $F_0$  trace with a moving average applied to it, the average slopes of both groups' descending whole tones were smaller than their ascending whole tones. Only the non-profession group, when curvature was calculated from the  $F_0$  trace with a moving average

applied to it, showed a significant effect for direction with the average curvature of the descending whole tones being slightly smaller than the ascending whole tones.

In the non-professional group's whole tone data there was a significant effect for whole tones starting on a chord tone, which were on average 4 cents smaller than whole tones starting on a non-chord tone. There was no effect in the professional group's interval size data. When the slope calculations were made directly on the  $F_0$  trace, there were significant effects for both the professional and non-professional groups. Both non-professional groups' average slopes for whole tones starting or ending on chord tones were, when significant, both smaller than those starting or ending on non-chord tones. When the slope calculations were run on the  $F_0$  trace with the moving average applied, there was only a significant effect for the non-professional group's whole tones ending with a chord tone, which had a smaller slope on average than the whole tones ending on a non-chord tone. For the curvature data calculated directly on the  $F_0$  trace, the non-professional group had a significant effect for whole tones ending on a chord tone, which had a smaller slope on average than the whole tones that ended on a non-chord tone. For the curvature data calculated on the  $F_0$  trace with the moving average applied to it, there was a significant effect for both the professional and non-professional groups for whole tones ending on a chord tone, which had a smaller curvature on average in both groups than the whole tones ending on a non-chord tone. There was also a significant effect in the non-professional group for whole tones starting on a chord tone, which had a smaller curvature on average than the whole tones starting on a chord tone.

In terms of group identity, the professional group's interval size and curvature measurements were on average larger than the non-professional group, although there were no significant group effects for slope. There were comparable effects for singer identity in both groups for the interval size and slope measurements. For interval size, singer identity effects for two professional singers (two and five) and three non-professional singers (three, four and five). For slope, there were significant effects for the four non-professional singers (one, three, four, and five) for the calculations made on the  $F_0$  trace and three non-professional singers (one, four, and five) for calculations made on the  $F_0$  trace after a moving average had been applied to it. For the professional group, there were significant effects for one singer (four) when the slope was calculated directly on the  $F_0$  trace and two singers

(three and four) when the slope was calculated on the  $F_0$  trace after a moving average had been applied to it. For the non-professional singers' curvature, there was a significant effect for four singers (one, three, four, and five) when the curvature was calculated directly on the  $F_0$  trace and for two singers (three and four) when the curvature was calculated on the  $F_0$  trace after a moving average had been applied to it. For the professional group, there were significant effects for three singers (two, three, and four) when the curvature was calculated directly on the  $F_0$  trace and for two singers (three and four) when the curvature was calculated on the  $F_0$  trace after a moving average had been applied to it.

#### 4.1.3.3 Slope and Curvature

As discussed in Section 3.2, slope and curvature data were calculated in two different ways: directly on the  $F_0$  trace and on the  $F_0$  trace after a 200 ms moving average had been applied to it. The calculations were made on the end of the first note in each melodic interval in order to determine if the singer was preparing for the imminent arrival of the upcoming note with changes in  $F_0$ . The moving average was used to minimize the effect of vibrato, although not enough evaluation has been done to determine either if this approach is sufficient to remove the vibrato or if it is also removing some of the information related to the slope and curvature trends in the signal. In light of this uncertainty, more weight is put on those conditions in which both calculations agree.

For slope calculations on the semitone data, the only points of agreement are in singer identity. In contrast, for the slope calculations on the whole tone data, there are observable trends in the non-professional group for the whole tones ending on a chord tone where the singers have a smaller slope on average than when moving towards a non-chord tone. Overall, there is more agreement in the slope data for some singers than others, although the actual slope sizes vary quite substantially, with the calculations run directly on the  $F_0$  trace tending to have larger values and confidence interval ranges.

For curvature calculations on the semitone data, the only point of agreement is in group and singer identity, with the professional group having larger curvature values on average than the non-profession group. This is also the case for the whole tone curvature data. As with the slope data, the curvature data for the whole tones exhibits a correspondence between the whole tones ending on a chord tone for the non-professional group, with the singers having a smaller curvature on average than when moving towards a non-chord tone. Also, similar to

the whole tone slope data, there is close agreement in terms of singer identity for both semitones and whole tones, and the curvature amounts also tend to be larger for the calculations run directly on the  $F_0$  trace.

These findings suggest that singers are more variable in their slope and curvature for semitones than for whole tones, which agrees with what has been observed in this experiment for interval size. In terms of the question of singers “scooping up” (increased slope + increased positive curvature) into leading tones, there does not appear to be any evidence supporting this. Overall, the biggest effect was between intervallic direction and slope in whole tones. This makes intuitive sense given that since there is a larger distance for the singer to traverse in the whole tone than the semitone, the singer might start moving toward it sooner. The finding of smaller curvature values for whole tones ending in a chord tone suggests that the singers might be preparing for the stability of the next note with increased stability in the current one. The larger curvature values in the professional group are harder to interpret, but may suggest that the professionals studied in this experiment were shaping their notes more than the non-professionals. It is interesting to observe that this trend co-occurs with all of the non-professional singers being significantly different from their baseline.

Overall, the use of the 1<sup>st</sup> and 2<sup>nd</sup> Discrete Cosine Transform has provided some interesting data related to way in which the  $F_0$  changes at the end of the first note in the melodic intervals being studied. However, the large amount of variability and the open questions regarding how best to apply the DCT to the signal means that only highly interpretative claims can be made from the data. For this reason, only interval size is examined in the ensemble experiments in the next section, 4.2.

(This page intentionally left blank)

## 4.2 Intonation in SATB Ensemble Singing

This experiment builds on the solo singer experiment in Section 4.1 by looking at the tuning of semitones and whole tones in the context of four-part ensemble singing. As with the previous experiment, both the degree of consistency across performances as well as musical context was considered. Three SATB ensembles were used in the four different parts of the experiment. In Part One (Figure 4.2.1 and Figure 4.2.2), there were 27 short progressions composed by Jonathan Wild, a music theory professor at McGill, where semitones in 9 different contexts occurred in each of the 3 upper voices. In Part Two (Figure 4.2.3), there were 18 short progressions composed by Peter Schubert, also a music theory professor at McGill, where whole tones in 6 different contexts occurred in each of the 3 upper voices. Part Three (Figure 4.2.4) was a short two-measure progression by Giambattista Bendedetti (1530–1590) repeated four times in each of the renditions. Depending on which voice is used as a tuning reference, this repeated progression can, in theory, promote an upward drift in tuning. The fourth part (Figure 4.2.5, Figure 4.2.6, and Figure 4.2.7) was the first verse of Michael Praetorius' (1571–1621) “Es ist ein Ros’ entsprungen” (“Lo, How a Rose E’er Blooming” in English). The first ensemble was used in a pilot study, where only the third and fourth parts were recorded since the first and second parts were designed after the pilot study took place. The second and third ensembles recorded all four parts of the experiment. The intonation data were extracted in the same way as in Section 4.1, and the melodic intervals were calculated in the same manner. The vertical intervals were calculated by measuring the interval size for each frame and then taking the mean across the series of vertical calculations, as described in Section 3.2. As with Section 4.1, this section examines role of musical context on the interval size for semitone and whole tone melodic intervals. Given the weak results in the solo singing experiments in Section 4.1, the slope and curvature were not analyzed for the ensemble experiment. Vertical intervals sizes were also examined, specifically whether the singers tuned the intervals closer to the interval sizes that can be observed between the lower harmonics in the harmonic series.

The figure displays a musical score for four voices: Soprano (S), Alto (A), Tenor (T), and Bass (B). The score is organized into three sections, each containing six numbered progressions (1 through 18). Each progression is composed of two measures. The voices are arranged vertically, with Soprano at the top and Bass at the bottom. Semitones of interest are circled in each measure. In progressions 1–9, the semitones occur in the soprano voice. In progressions 10–18, they occur in the alto voice. In progressions 19–27, they occur in the tenor voice.

Figure 4.2.1: Score for Part One. In its entirety, Part One consists of 27 progressions. Progressions 1–18 are shown in this figure. In each progression, the semitone of interest has been circled. The semitones of interest move between the three upper voices: they occur in the soprano in progressions 1–9, in the alto in 10–18, and in the tenor in 19–27.

Figure 4.2.2: Score for Part One. In its entirety, Part One consists of 27 progressions. Progressions 19–27 are shown in this figure. In each progression, the semitone of interest has been circled. Overall, the semitones of interest move between the three upper voices: they occur in the soprano in progressions 1–9, in the alto in 10–18, and in the tenor in 19–27.

The musical score consists of three staves (Soprano, Alto, Tenor) and 18 numbered progressions (1 through 18). The soprano staff (S) starts with progression 1. The alto staff (A) starts with progression 2. The tenor staff (T) starts with progression 3. The bass staff (B) starts with progression 4. Progressions 1 through 6 are soprano, 7 through 12 are alto, and 13 through 18 are tenor. Boxes highlight specific whole tones in each progression.

Figure 4.2.3: Score for Part Two, which consists of 18 progressions. In each progression, the whole tone of interest has been marked with boxes. The whole tones of interest move between the three upper voices: they occur in the soprano in progressions 1–6, in the alto in progressions 7–12, and the tenor in progressions 13–18.



Figure 4.2.4: Score for Part Three, a chord progression by Benedetti. The seed progression, which is repeated four times, is shown in the box.

The figure consists of three staves of musical notation for four voices: Soprano (S), Alto (A), Tenor (T), and Bass (B). The music is in common time and uses a bass clef for the bass voice.

- Measure 6:** The soprano has a dotted half note followed by eighth notes. The alto has eighth notes. The tenor has a dotted half note followed by eighth notes. The bass has eighth notes. Ascending semitones are circled with solid lines (eighth note in soprano, eighth note in alto, eighth note in tenor). Descending semitones are circled with dashed lines (eighth note in soprano, eighth note in alto, eighth note in tenor).
- Measure 7:** The soprano has eighth notes. The alto has eighth notes. The tenor has eighth notes. The bass has eighth notes. Ascending semitones are circled with solid lines (eighth note in soprano, eighth note in alto, eighth note in tenor). Descending semitones are circled with dashed lines (eighth note in soprano, eighth note in alto, eighth note in tenor).
- Measure 13:** The soprano has eighth notes. The alto has eighth notes. The tenor has eighth notes. The bass has eighth notes. Ascending semitones are circled with solid lines (eighth note in soprano, eighth note in alto, eighth note in tenor). Descending semitones are circled with dashed lines (eighth note in soprano, eighth note in alto, eighth note in tenor).

Figure 4.2.5: Score for Part Four, Praetorius' "Es ist ein Ros' entsprungen." The ascending semitones are marked with circles with solid lines, and descending semitones are marked with circles with dashed lines.

The image displays three staves of musical notation for four voices: Soprano (S), Alto (A), Tenor (T), and Bass (B). The music is in common time, with a key signature of one flat. The notation uses solid black lines for ascending whole tones and dotted black lines for descending whole tones. Boxes are used to highlight specific notes: solid boxes for ascents and dotted boxes for descents. The first staff begins with a solid box over the first note of the soprano line. The second staff begins with a solid box over the first note of the alto line. The third staff begins with a solid box over the first note of the bass line.

Figure 4.2.6: Score for Part Four, Praetorius' "Es ist ein Ros' entsprungen." The ascending whole tones are marked with boxes with solid lines, and descending whole tones are marked with boxes with dotted lines.

The figure consists of three staves of musical notation for four voices: Soprano (S), Alto (A), Tenor (T), and Bass (B). The notation is in common time, with a key signature of one flat. Vertical intervals between voices are highlighted with boxes. Dashed lines indicate cadential contexts, while solid lines indicate non-cadential contexts. Chords are labeled with Roman numerals: V and vi.

- Staff 1 (Measures 1-6):** Shows vertical intervals between voices. Chords labeled V and vi are indicated by dashed boxes. Measures 1-2, 3-4, and 5-6 are grouped by solid boxes.
- Staff 2 (Measures 7-12):** Shows vertical intervals between voices. Chords labeled V and vi are indicated by dashed boxes. Measures 7-8, 9-10, and 11-12 are grouped by solid boxes.
- Staff 3 (Measures 13-18):** Shows vertical intervals between voices. Chords labeled V and vi are indicated by dashed boxes. Measures 13-14, 15-16, and 17-18 are grouped by solid boxes.

Figure 4.2.7: Score for Part Four, Praetorius' "Es ist ein Ros' entsprungen." The vertical intervals studied are marked with boxes. Those in a cadential context are marked with dashed lines, and those in a non-cadential context are marked with solid lines.

## **4.2.1 Method**

### **4.2.1.1 Participants**

Three SATB ensembles participated in this experiment. The first ensemble (pilot) participated in the pilot study without a conductor, with only a subset of the experimental material that was used in the full experiment. This ensemble was semi-professional and was brought together for the experiment. Three of the four singers had previously sung together as section leads at Christ Church Cathedral in Montreal. The ensemble had an average age of 26 years ( $SD = 3.6$ ), with an average of 6.5 years of private voice lessons ( $SD = 4.5$ ), an average of 6.5 years of regular practice ( $SD = 2.5$ ), and an average of 0.75 hours of daily practice ( $SD = 0.84$ ).

The second ensemble (“Lab ensemble”), which performed in the main experiment, regularly sings together as a professional ensemble (“VivaVoce Montréal”) in the Montreal area with the same conductor, Peter Schubert, who also conducted them during the experiment. The third ensemble (“Church ensemble”) was the same as the second ensemble, except for the tenor. The original tenor was not available due to scheduling conflicts. These two ensembles had an average age of 42 years ( $SD = 9$ ), an average of 7.75 years of private voice lessons ( $SD = 0.5$ ), an average of 24 years of regular practice ( $SD = 10$ ), and an average of 1.75 hours of daily practice ( $SD = 1$ ).

### **4.2.1.2 Apparatus**

Both the pilot and professional Lab ensembles were recorded in the same room, a 4.85m x 4.50m x 3.30m lab at the Center for Interdisciplinary Research in Music Media and Technology (CIRMMT). The room had low noise, reflections, and reverberation time (ITU-standard). The singers were miked with cardioid headband mics (DPA 4088-F). The microphones were run through an RME Micstasy 8-channel microphone preamplifier and an RME Madi Bridge into a Mac Pro computer for recording. The professional Church ensemble was recorded on the altar of St. Mathias’ Church, a church in Montreal dating from 1912 with wooden floors, limestone walls, and seating for 350 people. As with the lab environment, the singers were miked with cardioid headband mics (DPA 4088-F), although a portable Zaxcom Deva 16 digital recorder was used for the rest of the recording setup.

#### 4.2.1.3 Procedure

As with the experiment in Section 4.1, the intonation-related data were extracted using the methods described in Chapter 3 and were checked manually by two people to correct any errors made by the alignment algorithm. Overall, there were far fewer errors in the alignments for this data than the recordings from “Ave Maria” due to the lack of ornamentation in the music used in this experiment.

Interval size estimates for the melodic intervals were calculated by taking a weighted mean across the frame-wise  $F_0$  estimates for each note, as used in Section 4.1 and described in Section 3.2. The interval size for vertical intervals was calculated by measuring the interval size between each note’s frame-wise  $F_0$  estimates and then taking the robust mean across this series of vertical calculations. This method for calculating vertical interval size, as described in Section 3.2, was chosen since it best summarizes the moment-to-moment tuning between the two singers. As with the solo singer experiment, the data were analyzed by examining the mean and standard deviation across groupings of intervallic conditions, through visualisation of data in box and whisker plots, and with linear regression analysis to evaluate significance in observable trends.

Linear regression analysis was used to explore the influence of musical context on melodic intervals by evaluating intervallic direction (up or down), intervallic conditions (each type of semitone and whole tone defined below), and singer identity. For Parts One and Two, the intervallic conditions shown in Table 4.2.1 and Table 4.2.4 were evaluated. Due to the repetitive nature and lack of directed harmonic context in Part Three, only intervallic direction and singer identity were evaluated. For Part Four, the same semitone intervallic conditions as “Ave Maria” linear regression were used. However, due to homophonic texture of the music in Part Four, there were not enough instances of the whole tone intervallic conditions from “Ave Maria” (see Table 4.1.1 in Section 4.1.1.3) to analyze, so for whole tones, only direction and singer identity were evaluated.

The main question underlying the analysis of the vertical interval data was whether the singers tend more towards the “pure” tunings found in Just Intonation (i.e., those found in the harmonic series). This hypothesis was tested by dividing intervals into two groups: those intervals where the notes have at least 6 harmonics in common amongst their first 32 harmonics and those intervals where the notes have less than 6 harmonics in common.

Intervals that fall into the first group are the Perfect Octave, the Perfect Fifth, and the Major Third. Intervals in the second group that occur in this piece are the Minor Third, the Tritone, the Minor Sixth, the Major Sixth, and the Augmented Sixth. This division allows for investigation of the historical debate, detailed in Barbour (1953), about whether vocal ensembles tend towards “pure” Perfect Octaves, Perfect Fifths, and Major Thirds.

*T*-tests, with a threshold of 0.05, were run to determine whether singers in the experiment tended to sing the vertical intervals with a greater coincidence of partials (the first group) closer to the tuning found in the harmonic series (Just Intonation) than those with fewer partials in common (the second group). Similarly, *t*-tests were run with the intervals in Part Four divided up into those that occur in cadential progressions and those that occur in non-cadential progressions. *T*-tests were also used to evaluate whether there was any influence of syllable or vowel on intonation in Parts Three and Four amongst the renditions sung to different syllables.

#### 4.2.2 Results

##### 4.2.2.1 Part One: Semitone Exercises

The twenty-seven progressions in Part One were written by Jonathan Wild and were designed to present the singer with melodic semitones between the same two pitches (G-G $\sharp$  and G-A $\flat$  in the soprano and tenor, and D-D $\sharp$  and D-E $\flat$  in the alto) in nine different harmonic contexts. In each context, the harmonic material presented in the other voices was manipulated so that five chromatic semitones occurred as different chord factors in five pairs of different chord types, and four diatonic semitones occur between different scale degrees— $\hat{2}-\hat{3}$ ,  $\hat{3}-\hat{4}$ ,  $\hat{5}-\hat{6}$ , and  $\hat{7}-\hat{8}$ —and also over different chord types, as detailed in Table 4.2.1. Each of the 9 contexts was repeated for each of the top 3 voices (soprano, alto, and tenor) and, each ensemble sang the entire exercise set 3 times, resulting in 9 instances of each condition per ensemble (3 instances per rendition \* 3 renditions). The mean and standard deviations for each semitone condition are shown in Table 4.2.2. Vertical intervals were calculated between the bass and the notes in the melodic semitone intervals being studied, resulting in 18 intervals calculated between each voice and the bass in each rendition. For each ensemble, across all renditions, there were: 27 instances of Minor Thirds, 18 instances of Major Thirds, 45 instances of Perfect Fifths, 36 instances of Major Sixths, 18 instances of Minor Sevenths, and 18 instances of Perfect Octaves. The means and standard

deviations for each type of vertical interval across all of the singers in each ensemble are shown in Table 4.2.3.

	<b>Chromatic semitones</b>	$\hat{1} - \hat{8}$	$\hat{2} - \hat{3}$	$\hat{3} - \hat{4}$	$\hat{5} - \hat{6}$
<b>Soprano</b>	1–5	6	7	8	9
<b>Alto</b>	10–14	15	16	17	18
<b>Tenor</b>	19–23	24	25	26	27

Table 4.2.1: Organization of Part One. The columns indicate the conditions, the scale degrees between which the semitone occurs, and the rows indicate in which progression the conditions occur. These numbers correspond to the progression label in Figure 4.2.1 and Figure 4.2.2.

<b>Semitone conditions (Number of instances)</b>	<b>Lab Ensemble</b>		<b>Church Ensemble</b>	
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
Chromatic semitone 1 (9)	86.9	12.2	111.9	23.3
Chromatic semitone 2 (9)	88.5	9.8	93.0	8.4
Chromatic semitone 3 (9)	93.0	10.8	95.8	11.7
Chromatic semitone 4 (9)	86.2	12.1	97.4	12.4
Chromatic semitone 5 (9)	81.8	15.8	101.1	8.6
$\hat{1} - \hat{8}$ semitone (9)	95.7	10.4	107.2	17.1
$\hat{2} - \hat{3}$ semitone (9)	102.6	13.8	98.6	18.5
$\hat{3} - \hat{4}$ semitone (9)	96.2	8.6	101.3	8.0
$\hat{5} - \hat{6}$ semitone (9)	87.9	7.5	102.5	15.4

Table 4.2.2: Mean and standard deviation of the melodic semitone sizes for both the Lab and Church ensembles in Part One.

<b>Vertical Intervals [JI size] (Number of instances)</b>	<b>Lab Ensemble</b>		<b>Church Ensemble</b>	
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
Minor Third [316 cents] (27)	307.9	14.6	303.0	14.3
Major Third [386 cents] (18)	400.6	15.0	408.3	13.2
Perfect Fifth [702 cents] (45)	713.8	9.8	705.5	13.4
Major Sixth [884 cents] (36)	900.0	14.2	894.4	13.2
Augmented Sixths [977 cents] (18)	989.6	14.6	995.2	24.5
Perfect Octave [1200 cents] (18)	1195.4	15.8	1203.1	13.0

Table 4.2.3: Mean and standard deviation of the sizes of the vertical intervals in Part One between the first and second notes in the semitone intervals and the bass note for both the Lab and Church ensembles.

With the exception of the  $\hat{2} - \hat{3}$  semitone, the means for the Lab ensemble were smaller than the Church ensemble. Overall, the means of the interval sizes in each category for the Lab ensemble's semitones were closer to the Pythagorean semitone (90 cents), whereas the

Church ensemble's were closer to equal temperament (100 cents), with two conditions (Chromatic 1 and  $\hat{7}-\hat{8}$ ) closer to the Just Intonation semitone (112 cents). The groups' mean interval sizes were more consistent with each other for the vertical intervals than for the melodic intervals. Some of the vertical intervals were closer to 5-limit Just Intonation than others; however, in light of the large standard deviations, there was no clear trend even in these intervals to Just Intonation. The standard deviations varied quite substantially between groups and conditions for both melodic and vertical intervals. As with Section 4.1, further analysis is required to see if the musical context has influenced interval size in these performances.

As discussed in Section 4.1, the top and bottom of each box in the box and whisker plots represents the 25<sup>th</sup> and 75<sup>th</sup> percentiles, with the solid horizontal line running through the box representing the 50<sup>th</sup> percentile, or the 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. The box and whisker plot in Figure 4.2.8 shows the sizes of the semitones in each condition in both ensembles, and Figure 4.2.9 shows semitone interval size data organized by singer. The interval size data for each type of vertical interval is shown for each ensemble in Figure 4.2.10.

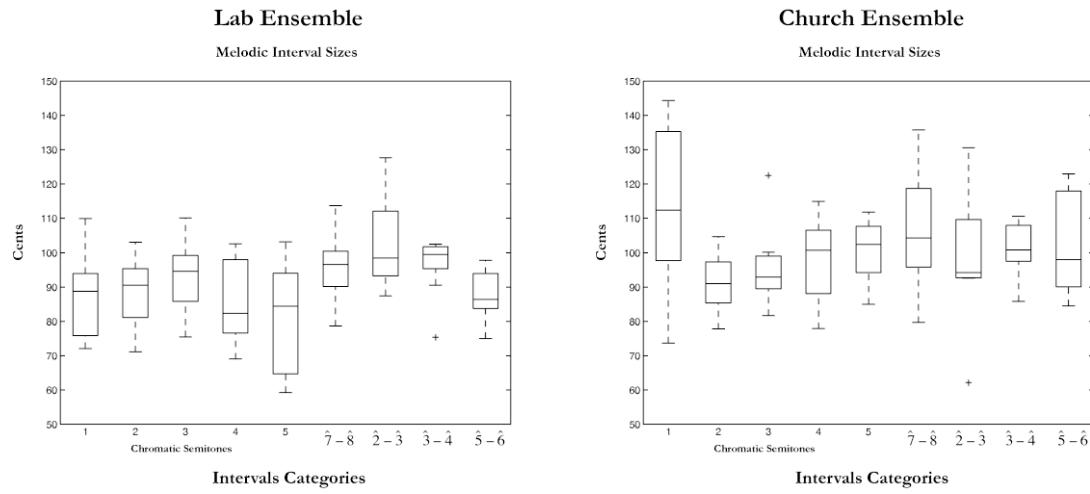


Figure 4.2.8: Box and whisker plots for the semitones interval sizes in Part One across each condition, separated by ensemble.

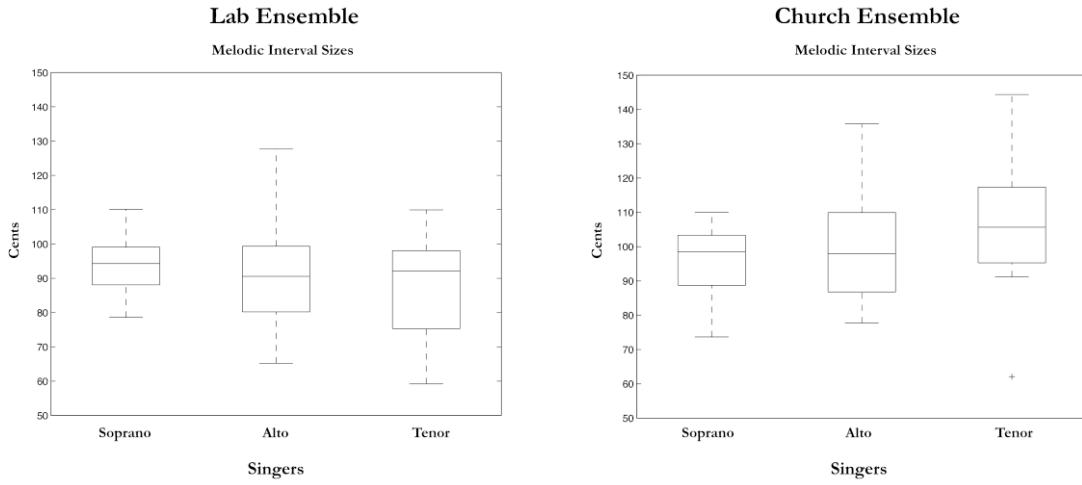


Figure 4.2.9: Box and whisker plots for the semitone interval sizes in Part One for each singer in each ensemble.

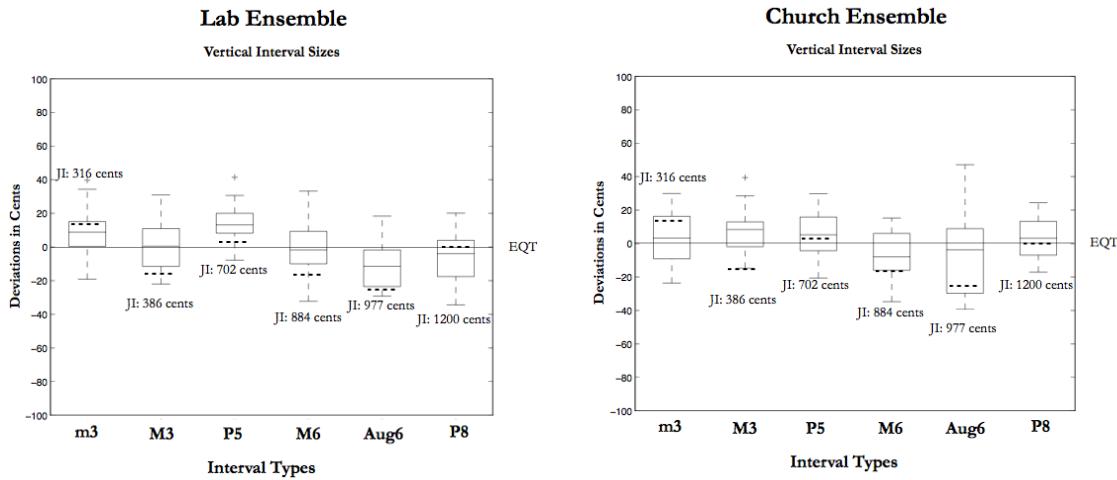


Figure 4.2.10: Box and whisker plot of interval sizes for the vertical intervals in Part One between the bass and the voice singing the melodic semitone interval being evaluated. The vertical interval sizes were calculated for both notes in the melodic semitone interval. The dotted lines represent the position of the idealized 5-limit Just Intonation tunings for each interval, and the solid line represents equal temperament (EQT).

As visualized in the box and whisker plots for the Lab ensemble's different semitone conditions (Figure 4.2.8), there were, with the exception of the  $\hat{2} - \hat{3}$  semitone, smaller 25<sup>th</sup>-75<sup>th</sup> ranges for diatonic semitones than the chromatic semitones. A *t*-test, with a threshold of

0.05, showed that there was a significant difference between the overall means for the chromatic (87.3 cents) and diatonic (95.6 cents) semitones. In contrast, the Church ensemble had much larger 25<sup>th</sup>–75<sup>th</sup> percentile ranges for the diatonic intervals and slightly smaller ranges for the chromatic semitone. There was, however, much more variability within both groups of semitone categories, which meant that the difference between the means of the groups was not statistically significant. However, no such trends were observed for the Church ensemble. Overall, the mean of the Lab ensemble's semitones (90 cents) was significantly smaller than the mean of the Church ensemble's semitones (101 cents), with a threshold of 0.05.

A linear regression analysis was run on the semitone data to see if there were any significant effects on semitone size for the different categories or the singers within each ensemble. For this analysis, the  $\hat{7} - \hat{8}$  semitone was used as a baseline for comparing interval categories, and the soprano was used as a baseline for comparing the singers. For the Lab ensemble ( $R^2 = 0.28$ ,  $p < 0.01$ ), there was only a significant difference between the  $\hat{7} - \hat{8}$  semitones and the Chromatic 4 semitones, which were on average 10 cents smaller than the  $\hat{7} - \hat{8}$  semitones (95% confidence interval=[3,16]). For the Church ensemble ( $R^2 = 0.25$ ,  $p < 0.02$ ), there was a significant effect for the Chromatic 2 semitones, which were on average 16 cents smaller than the  $\hat{7} - \hat{8}$  semitones (95% confidence interval=[3,29]). There were also significant effects for singer identity: the tenor's semitones were on average 12 cents smaller than the soprano's (95% confidence interval=[4,19]), and the alto's were on average 8 cents smaller (95% confidence interval=[1,16]).

Figure 4.2.9 shows the same data as Figure 4.2.8, divided up by singer instead of semitone category. Only the soprano (same singer) was consistent between ensembles (mean interval size of 93.6 cents in the Lab ensemble versus 96.0 in the Church ensemble), while both the alto (same singer) and the tenor (different singer) had significantly larger interval sizes in the Church ensemble. The alto's average semitone size increased from 91.6 to 99 cents, whereas the two tenors' average semitone sizes were 88 cents and 105 cents between the Lab and Church ensembles.

The plots in Figure 4.2.10 show the vertical interval data for each ensemble. In this exercise, the purely tuned Perfect Octaves (2:1), Perfect Fifths (3:2), and Major Thirds (5:4) were put into the first group of vertical intervals and the Minor Thirds (6:5), Major Sixths (5:3), and

Augmented Sixth (255:128) were put in the other. The box and whisker plots show that in general the vertical intervals did not converge around the interval size predicted by Just Intonation for either category listed above, although there was, as with the melodic intervals, a high degree of variability in interval size. In order to evaluate this question, a two-tailed *t*-test was run on the deviations from Just Intonation for each interval in each group. For the Lab ensemble, the absolute distance of the intervals in the first group from Just Intonation tuning was on average 19.7 cents, whereas the intervals in the other group were on average 12.3 cents away. The difference between the groups was not statistically significant. For the Church ensemble, the difference between the groups was significant: the first group was on average 17.4 cents away from Just Intonation tuning, and the second group was on average 38.1 cents away.

#### **4.2.2.2 Part Two: Whole tone Exercises**

The eighteen progressions in Part Two were written by Peter Schubert and were designed to present the singers with melodic whole tones between the same two pitches (A-B in the soprano and tenor, and D-E in the alto) in six different contexts. In each context, the harmonic material presented in the other voices was manipulated so that the melodic whole tones were between different scale degrees:  $\hat{1}-\hat{2}$ ,  $\hat{2}-\hat{3}$ ,  $\hat{3}-\hat{4}$ ,  $\hat{4}-\hat{5}$ ,  $\hat{5}-\hat{6}$ , and  $\hat{6}-\hat{7}$ , as detailed in Table 4.2.4. Each of the six contexts was repeated for each of the top three voices (soprano, alto, and tenor), and each ensemble sang the entire exercise set three times, resulting in 9 instances of each condition per ensemble (3 instances per rendition x 3 renditions). The mean and standard deviations for each condition are shown in Table 4.2.5. Vertical intervals were calculated between the bass and the notes in the melodic whole tone intervals being studied. There were 12 intervals calculated between each voice and the bass in each rendition. Across all renditions there were, for each ensemble: 30 instances of Minor Thirds, 27 instances of Major Thirds, 9 instances of Tritones, 15 instances of Perfect Fifths, 9 instances of Minor Sixths, and 18 instances of Perfect Octaves. The means and standard deviations for each type of vertical interval across all of the singers in each ensemble are shown in Table 4.2.6.

	$\hat{2}-\hat{3}$	$\hat{5}-\hat{6}$	$\hat{4}-\hat{5}$	$\hat{3}-\hat{4}$	$\hat{1}-\hat{2}$	$\hat{6}-\hat{7}$
Soprano	1	2	3	4	5	6
Alto	7	8	9	10	11	12
Tenor	13	14	15	16	17	18

Table 4.2.4: Organization of Part Two. The columns indicate the conditions, the scale degrees between which the whole tone occurs, and the row indicates the progression in which the conditions occur. These numbers correspond to the progression label in Figure 4.2.3.

Whole tone conditions (Number of instances)	Lab Ensemble		Church Ensemble	
	Mean	SD	Mean	SD
$\hat{2}-\hat{3}$ whole tone (9)	199.6	10.9	200.6	28.7
$\hat{5}-\hat{6}$ whole tone (9)	193.0	8.8	193.9	6.5
$\hat{4}-\hat{5}$ whole tone (9)	211.7	12.8	203.5	13.9
$\hat{3}-\hat{4}$ whole tone (9)	204.6	12.9	202.4	10.4
$\hat{1}-\hat{2}$ whole tone (9)	200.1	11.7	198.8	21.2
$\hat{6}-\hat{7}$ whole tone (9)	206.8	7.1	205.6	11.4

Table 4.2.5: Mean and standard deviation of the melodic whole tone sizes in Part Two for both the Lab and Church ensembles.

Vertical Intervals [JI size] (Number of instances)	Lab Ensemble		Church Ensemble	
	Mean	SD	Mean	SD
Minor Third [316 cents] (30)	302.0	17.1	294.7	25.0
Major Third [386 cents] (27)	403.1	19.4	394.1	18.2
Tritone [590 cents] (9)	604.1	35.3	603.8	13.4
Perfect Fifth [702 cents] (15)	715.4	11.7	708.3	10.5
Minor Sixth [814 cents] (9)	797.6	11.7	809.2	14.1
Perfect Octave [1200 cents] (21)	1210.2	14.1	1210.9	12.7

Table 4.2.6: Mean and standard deviation of the sizes of the vertical intervals in Part Two between the first and second notes in the whole tone intervals and the bass note for both the Lab and Church ensembles.

Overall, the means of the melodic interval sizes were comparable between both groups, as were the standard deviations with the exception of the  $\hat{2}-\hat{3}$  and  $\hat{1}-\hat{2}$  whole tones. Overall, the mean values were much closer to the equal tempered (200 cents) and the 9:8 Pythagorean/Major Just-Intonation (204) whole tones than the 10:9 Minor Just Intonation whole tone (182 cents). There was more variation in the groups' mean interval sizes for the vertical intervals than for the melodic intervals, although the standard deviations were similar except for the Tritone in the Lab ensemble and the Minor Third in the Church ensemble.

The same analysis performed in 4.2.2.1 was repeated to see if the musical context has influenced interval sizes in these performances. The box and whisker plots in Figure 4.2.11 correspond to the semitone plots in Figure 4.2.8 and shows the size of the whole tones for each condition in both ensembles. Likewise, Figure 4.2.12 corresponds to Figure 4.2.9 and shows whole interval size data organized by singer. Figure 4.2.13 corresponds to Figure 4.2.10, and shows the data for vertical interval size.

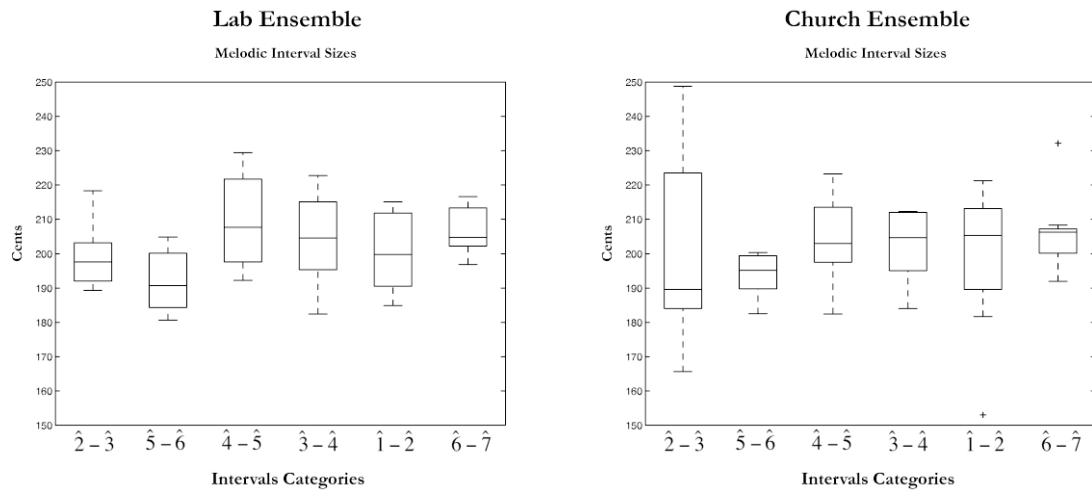


Figure 4.2.11: Box and whisker plots for the whole tone interval sizes in Part Two across each condition, separated by ensemble.

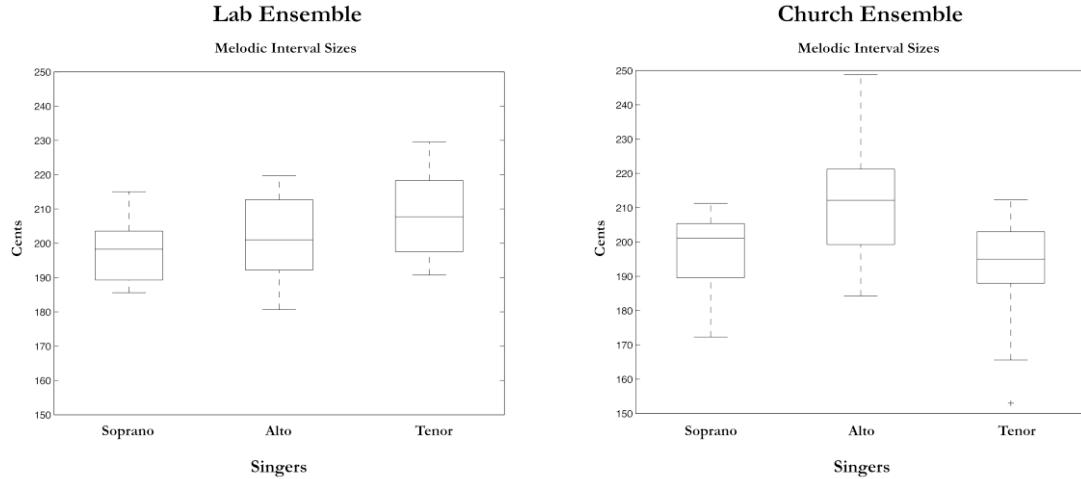


Figure 4.2.12: Box and whisker plots for the whole tone interval sizes in Part Two for each singer in each ensemble.

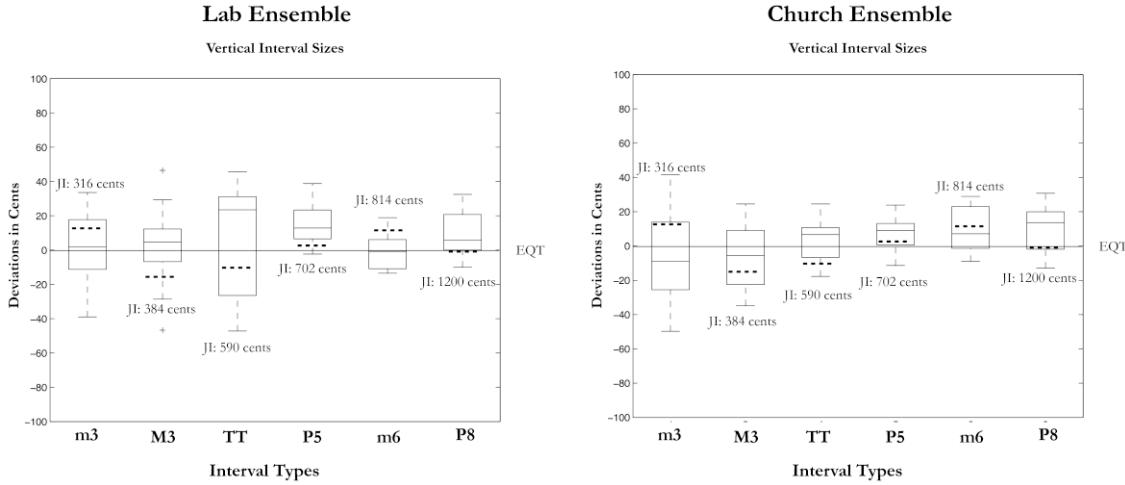


Figure 4.2.13: Box and whisker plot of interval sizes for the vertical intervals between the bass and the voice singing the melodic whole tone interval being evaluated in Part Two. The vertical interval sizes were calculated for both notes in the whole tone interval. The dashed lines represent the position of the idealized 5-limit Just Intonation tunings for each interval.

In the box and whisker plots for the different whole tone conditions (Figure 4.2.11), the relative size of 25<sup>th</sup>–75<sup>th</sup> position between the two ensembles followed similar trends except for the  $\hat{2} - \hat{3}$  whole tones, which had a large amount of variability. In both groups, the median position was lower in the  $\hat{2} - \hat{3}$  and  $\hat{5} - \hat{6}$  whole tones than in the other ones. In the Lab ensemble, the remaining medians showed some variability, whereas in the Church ensemble, the remaining medians were closer. Overall, there was no significant difference between the means of all of the whole tones in the lab (202.6 cents) and the church (200.8 cents) ensembles, using a *t*-test with a threshold of 0.05.

The linear regression analysis was run on the whole tone data using the  $\hat{6} - \hat{7}$  whole tone and tenors as a baseline ( $R^2 = 0.42$ ,  $p < 0.001$  for the Lab ensemble and  $R^2 = 0.32$ ,  $p < 0.01$  for the Church ensemble). Amongst the whole tone conditions, there was only a significant effect for  $\hat{5} - \hat{6}$  in the Lab ensemble, which was on average 14.4 cents smaller than the baseline (95% confidence interval=[5,24]). Both the tenor and the alto were significantly different than the soprano in the Lab ensemble, with the tenor being 12.3 cents smaller (95% confidence interval=[6,19]) and the alto being 9.2 cents smaller than the soprano (95% confidence interval=[3,16]). For the Church ensemble, there was only a significant effect for

the alto, which was on average 20.3 cents larger than the soprano (95% confidence interval=[10,30]).

Figure 4.2.12 shows the same data as Figure 4.2.11, divided up by singer instead of semitone category. As with the Part One, the soprano (same singer) was consistent between ensembles (mean interval size of 198.1 cents in the Lab ensemble versus 197.6 in the Church ensemble), and the alto (same singer) had significantly larger interval sizes in the Church ensemble. The alto's average semitone size increased from 200.6 to 212.6 cents. The difference between the two different tenors in each ensemble was opposite of Part One, with the second tenor having a smaller interval size for whole tones than the first one (204.0 cents versus 196.3 cents); however, this difference was not statistically significant with a threshold of 0.05.

The plots in Figure 4.2.13 show the vertical interval data for each ensemble. The vertical intervals in this set of exercises were slightly different than those in Part One: Tritones and Minor Sixths occurred in place of Major Sixths and Minor Sevenths in the second group of intervals, where the upper note had less than 6 harmonics in common with the first 32 harmonics of the lower note in the interval. As in Part One, the vertical intervals did not converge around the interval size predicted by Just Intonation. For the Lab ensemble, the first group, consisting of the Perfect Octave, the Perfect Fifth, and the Major Third, had an absolute average distance from Just Intonation tuning of 19.7 cents, and the second group had an absolute distance that was 17.0 cents. This difference was not statistically significant, with a threshold of 0.05; however, the difference between the two groups for the Church ensemble was. For the Church ensemble, the first group's average distance from Just Intonation was 17.4 cents, and the second group was 29.4 cents away.

#### 4.2.2.3 Part Three: Benedetti Chord Progression

The experimental material in Part Three was a three-part chord progression written by Giambattista Benedetti (1530–1590) that was designed to show that singers do not sing in Just Intonation since strict adherence to Just Intonation would result in a significant pitch drift that is not observable in performances of the progression (Benedetti 1585; Palisca 1994). The progression, as shown in Figure 4.2.4, is built from a seed two-measure progression that is repeated four times. In the two-measure seed, a note in one of the voices is either sustained or repeated between each chord, as shown in Table 4.2.7.

D	E	E	D
A	(A)	B	(B)
D	A	G	(G)

Table 4.2.7: Notes in the two-measure seed progression, which is repeated four times to make up the musical material used in Part Three. The bolded notes are either sustained or repeated from the previous sonority.

If the singers were to tune in Just Intonation to the sustained note, rather than the bass note, the ensemble would drift up a syntonic comma (21.5 cents) by the end of each seed, resulting in a total upwards drift of 86 cents by the end of the four repetitions. In contrast, if the singers were to tune to the bass in each vertical sonority, with D, A, or G in the bass, there should be no drift. The calculations for both tuning scenarios are shown in Figure 4.2.14.

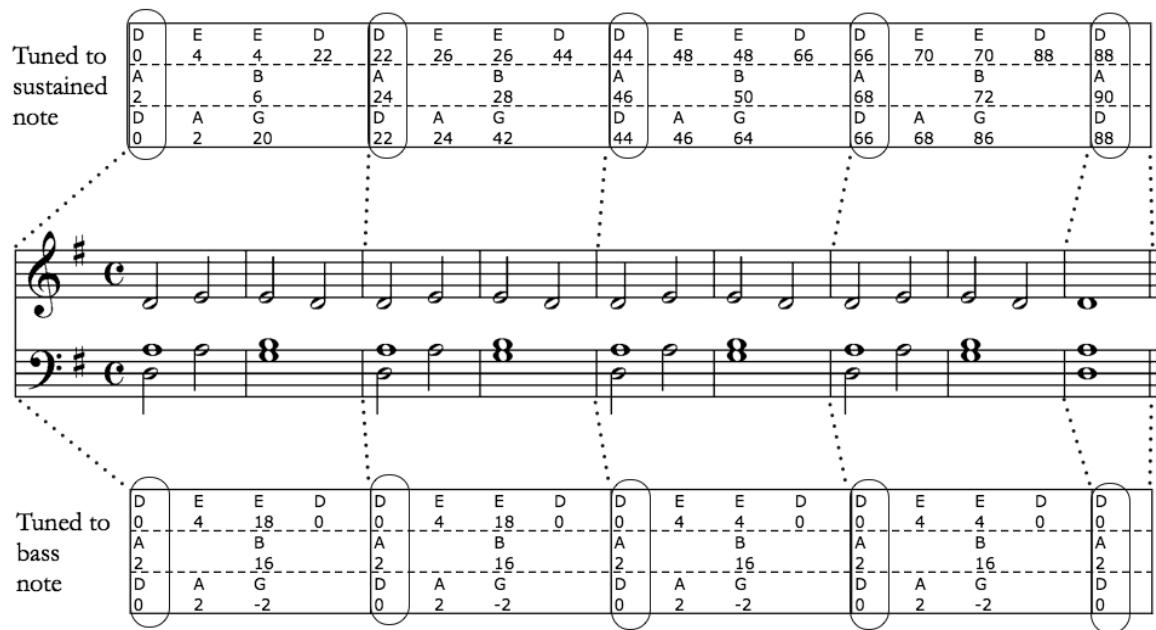


Figure 4.2.14: Theoretical tuning for Benedetti progression used in Part Three. The numbers in the tables at the top and bottom of the figure indicate the difference in cents between the projected tuning and equal temperament.

For this part of the experiment, four different ensembles were recorded. As detailed in Figure 4.2.15, Ensemble 1 consisted of singers from the pilot experiment, Ensemble 2 consisted of singers from the professional Lab ensemble, and Ensembles 3 and 4 consisted of the singers from the professional Church ensemble. Ensemble 1 sung without a conductor, whereas Ensembles 2–4 were conducted by Peter Schubert. Ensemble 1 performed the exercise three times, Ensembles 2 and 3 performed the exercise four times, and Ensemble 4 performed the exercise five times.

Ensemble 1 – Semi-professional singers\* (ATB, pilot)

Ensemble 2 – Professional singers\*\* (ATB, lab)

Ensemble 3 – Professional singers\*\* (SAT, church)

Ensemble 4 – Professional singers\*\* (ATB, church)

\*no conductor

\*\*conducted by Peter Schubert

Figure 4.2.15: Ensembles used in Part Three.

In order to assess whether the ensembles were drifting in the way predicted by Benedetti, the perceived pitch estimates for the D in bass at the start of each seed progression were obtained and plotted for each of the ensembles' renditions in Figure 4.2.16. Ensemble 1 sung all of their renditions on the syllable “du,” Ensemble 2 sung 2 rendition on the syllable “mi” and 2 on the syllable “ma,” Ensemble 3 sung 2 renditions on the syllable “mi” and 2 on the syllable “ma,” and Ensemble 4 sung 3 renditions on the syllable “mi” and 2 on the syllable “ma.”

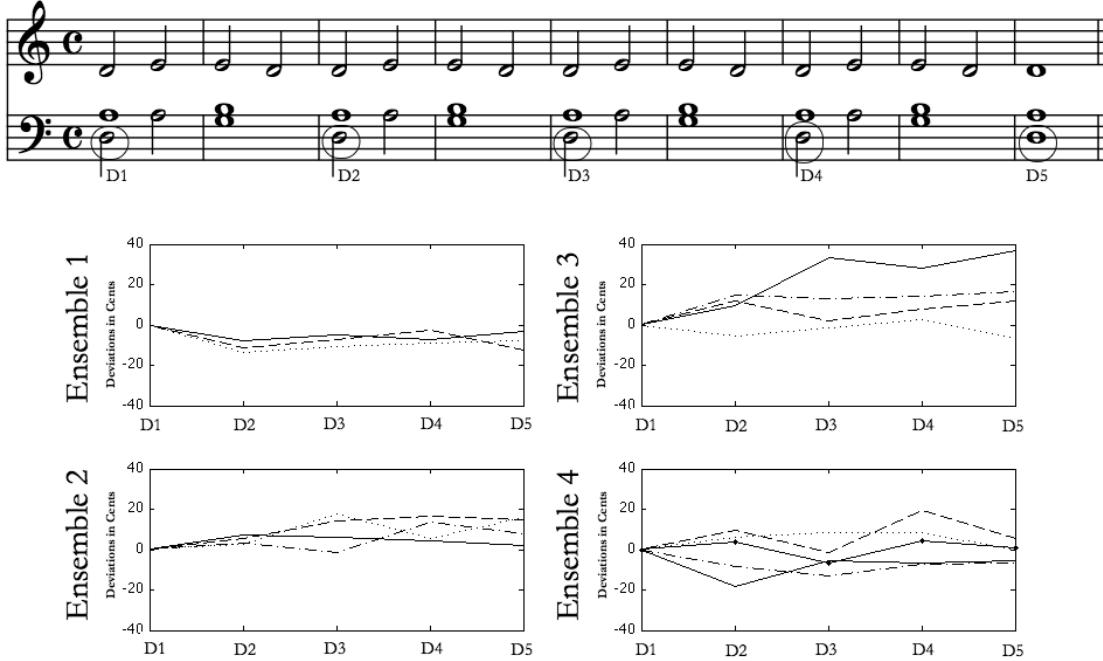


Figure 4.2.16: Summary of the amount of drift in each ensemble's renditions of the Benedetti's chord progression used in Part Three. The lines in the each plot link the perceived pitch estimates for the notes D1–D5 in each rendition.

The plots in Figure 4.2.16 show that none of the ensembles drifted as much as predicted by Benedetti. This is not surprising as such a rapid drift, 88 cents over eight measures, is highly unlikely since it implies that the singers were not retaining their starting pitch as a reference only a few tens of seconds after it was sung. Ensemble 1 was the most consistent with itself across performances, exhibiting only a small amount of drift. Ensembles 2 and 3 both tended to drift upwards with Ensemble 3 showing a greater amount of variability in the amount of drift. Ensemble 4 had little drift overall but showed a large amount of variation within each performance.

Exercise Three also provided the opportunity to examine ascending and descending whole tone melodic intervals. In each rendition, there were 4 ascending and 4 descending whole tones in the upper voice, 4 ascending and 4 descending whole tones in the middle voice, and 4 descending whole tones in the bottom voice for a total of 8 ascending whole tones and 12 descending whole tone intervals in each rendition. Due to the repetitive nature of the musical material in this exercise, there were no contextual whole tone conditions to consider,

so the mean and standard deviation values for only the conditions of ascending and descending are shown in Table 4.2.8. Vertical intervals were calculated between all of the voices: lowest voice to middle voice, lowest voice to upper voice, and middle voice to upper voice. Overall, there were 51 vertical intervals in each rendition: 4 Minor Thirds, 8 Major Thirds, 9 Perfect Fourths, 17 Perfect Fifths, 4 Major Sixths, and 9 Perfect Octaves. The means and standard deviations for each type of vertical interval across all of the singers in each ensemble are shown in Table 4.2.9.

The means of the whole tone interval sizes in Table 4.2.8 were generally smaller than either the equal tempered (200 cent) or the Pythagorean/Major Just Intonation (204 cents) semitones. The most notable variance occurred in the middle voice, where the mean intervals sizes for the ascending whole tones was 185 cents in Ensemble 1 and 207 cents in Ensemble 2 and for the descending whole tones was 183 cents in Ensemble 1 and 210 cents in Ensemble 2. The standard deviations were comparable across the ensembles and voices.

For the vertical interval sizes in Table 4.2.9, there was a wide range in the mean values for both the vertical Minor and Major Thirds, ranging from 300–322 cents for the Minor Third and 375–413 cents for the Major Third. The Minor Third ranges encompassed the equal tempered (300 cents) and Just Intonation (316 cents) tunings, whereas the Major Third range encompassed the Just Intonation (386 cents), the equal tempered (400 cents), and the Pythagorean (408 cents) tunings. When the standard deviations were taken into account, Ensemble 2 also encompassed the Pythagorean tuning (294 cent) for the Minor Thirds. The range of the means for the Major Sixths encompassed only the equal tempered tuning (900 cents) since the means were all larger than the Just Intonation tuning (884 cents) and marginally smaller than the Pythagorean one (905 cents). The tunings for the Perfect Fourth (498 cents), Perfect Fifth (702 cents), and Perfect Octave (1200 cents) were common to both the Pythagorean and Just Intonation systems and were close to the values for equal temperament (500, 700, and 1200 cents, respectively); the ranges for these intervals encompassed all of these tunings. The box and whisker plots in Figure 4.2.17 show the ascending and descending whole tones interval data for each ensemble across the singers. The plots in Figure 4.2.18 show the data for the vertical intervals for each ensemble.

	<b>Ensemble 1</b>	<b>Ensemble 2</b>	<b>Ensemble 3</b>	<b>Ensemble 4</b>
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
<b>Whole tone conditions [Number of instances]</b>				
Top voice, ascending (12/16/16/20)	199.3	5.5	191.7	5.9
Top voice, descending (12/16/16/20)	195.0	4.3	190.8	16.2
Middle voice, ascending (12/16/16/20)	184.6	5.8	207	11.5
Middle voice, descending (12/16/16/20)	182.9	9.9	210.0	12.6
Bottom voice, descending (12/16/16/20)	191.0	6.8	189.1	5.5

Table 4.2.8: Mean and standard deviation of the ascending and descending melodic whole tone sizes in Part Three for all ensembles, broken down by voice.

	<b>Ensemble 1</b>	<b>Ensemble 2</b>	<b>Ensemble 3</b>	<b>Ensemble 4</b>
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
<b>Vertical Intervals [JI values] [Number of instances]</b>				
Minor Thirds [316 cents] (12/16/16/20)	322.4	7.4	299.5	12.1
Major Thirds [386 cents] (24/32/32/40)	375.6	9.1	413.3	11.3
Perfect Fourths [498 cents] (27/36/36/45)	508.7	10.1	497.0	17.2
Perfect Fifths [702 cents] (51/68/68/85)	700.7	6.2	704.9	13.9
Major Sixths [884 cents] (12/16/16/20)	893.0	6.4	903.1	14.5
Perfect Octaves [1200 cents] (27/36/36/45)	1201.3	7.3	1205.9	11.6

Table 4.2.9: Mean and standard deviation of the sizes of the vertical intervals in Part Three between the three voices across all renditions by each ensemble.

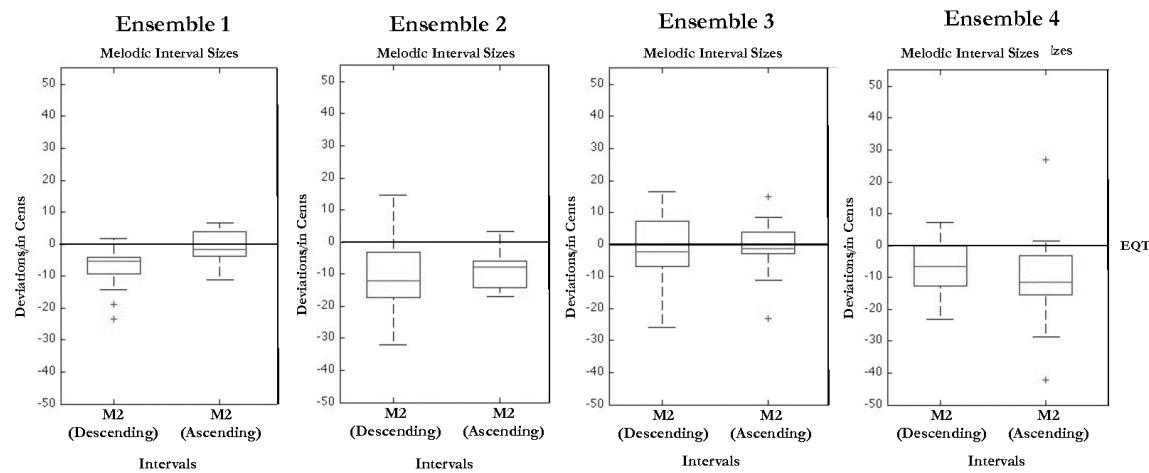


Figure 4.2.17: Box and whisker plots for the whole tones interval sizes in Part Three across all the singers for each ensemble.

In the box and whisker plots in Figure 4.2.17, there are observable similarities between Ensembles 2 and 3 in terms of the relative positioning of the medians and the 5<sup>th</sup>, 25<sup>th</sup>, 75<sup>th</sup>, and 95<sup>th</sup> percentiles of the descending and ascending whole tones. Overall, the medians were slightly higher for the ascending intervals, and the percentile ranges were larger for the descending intervals. In contrast, Ensemble 1 exhibited the same trend for medians, but the percentile ranges were much smaller for the descending whole tones than for Ensembles 2 and 3. Ensemble 4 was anomalous in that the ascending whole tones' medians were lower, and the percentiles ranges were comparable between the ascending and descending semitones.

A linear regression analysis was run on the melodic interval data with intervallic direction and singer identity for conditions. For Ensemble 1 ( $R^2 = 0.48$ ,  $p < 0.0001$ ), there was no effect for direction or the upper voice; the only significant effect was that the middle voice was on average 13.4 cents smaller than the lower voice (95% confidence interval = [9.5, 17.3]). Likewise, for Ensemble 2 ( $R^2 = 0.40$ ,  $p < 0.0001$ ), there were no significant effects for direction or the upper voice, whereas the middle voice was on average 17.3 cents larger than the lower voice (95% confidence interval = [11.7, 22.9]). The regressions for Ensemble 3 ( $R^2 = 0.01$ ,  $p = 0.85$ ) and Ensemble 4 ( $R^2 = 0.04$ ,  $p = 0.28$ ) were not themselves significant.

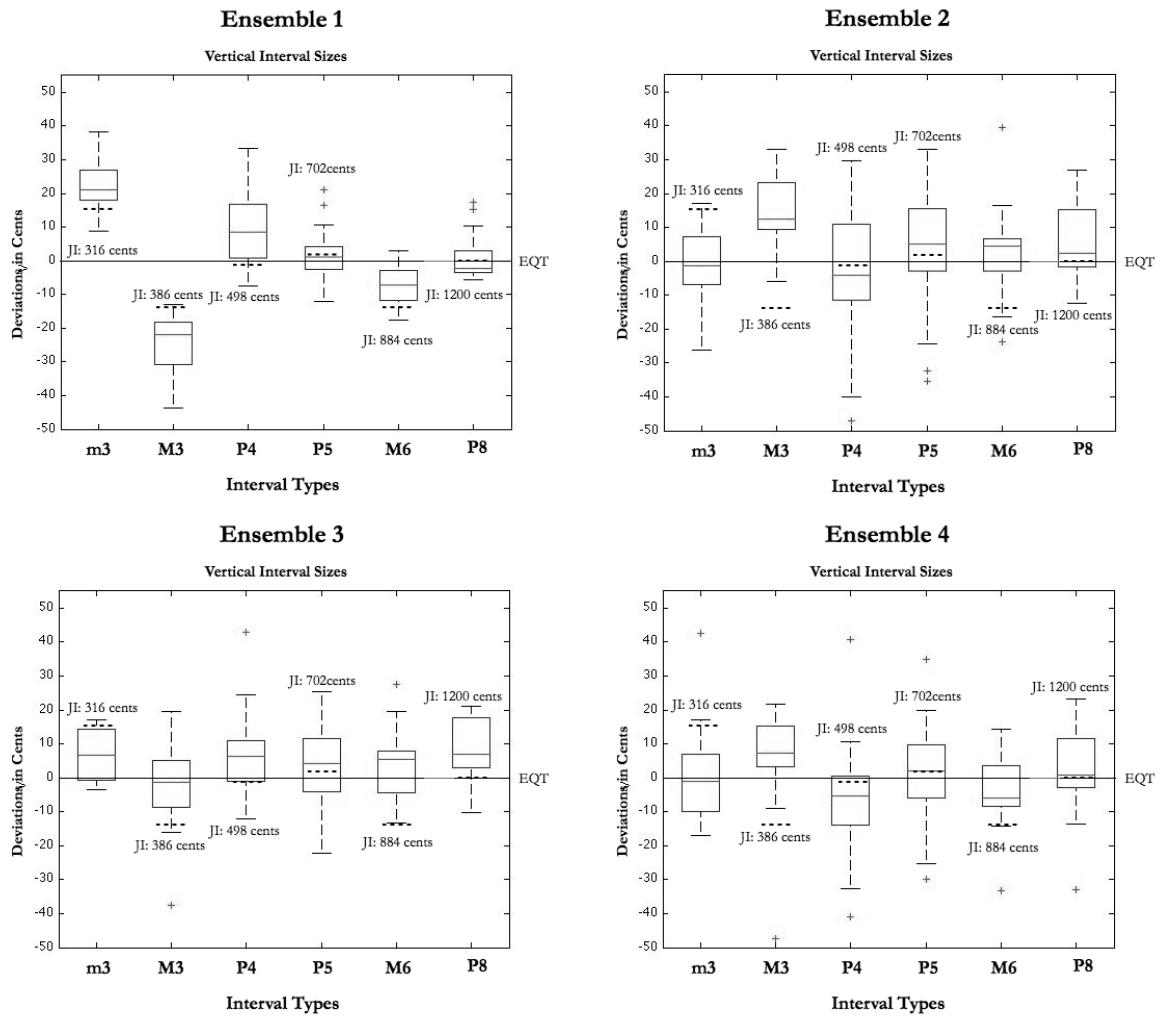


Figure 4.2.18: Box and whisker plot of interval sizes for the vertical intervals in Part Three between all of the singers in each ensemble. The dotted lines represent the position of the idealized 5-limit Just Intonation tunings for each interval.

Figure 4.2.18 shows a high degree variance between Ensemble 1 and the other ensembles for vertical interval size. Specifically, the median values for Ensemble 1 were much closer to Just Intonation tuning than the other ensembles. Ensemble 2 showed the largest ranges for the 5<sup>th</sup>-95<sup>th</sup> percentiles, although there were fairly substantial deviations in all of the ensembles. There were not any observable trends in the interval size data to explain the drift trends for each ensemble, which remains an open research question that will be addressed in the future works section in Chapter 5.

As with Parts One and Two, the vertical intervals were divided into groups for the purpose of statistical analysis. As with the other ensembles, the first group of vertical intervals was made up of the Major Thirds, Perfect Fifths, and Perfect Octaves. The second group was made up of all other vertical intervals: Minor Thirds, Perfect Fourths, and Major Sixths. Three different *t*-tests were run on the vertical interval data: the first for all of the intervals, the second for the intervals between the lower voice and the middle and upper voices, and the third for the intervals between the two upper voices. The division of the intervals for the second and third *t*-test was done to explore if the singers were more likely to tune closer to Just Intonation when the interval occurred between an upper voice and the bass rather than between the upper voices, which could suggest that the singers were tuning to the bass. There were significant effects in Ensemble 1 for the first and second tests, in Ensemble 2 for the first test, and in Ensemble 3 for the first and third tests. There were no significant results for Ensemble 4. Detailed results for the *t*-test are available in Table 4.2.10. These results show that the effects were ensemble-dependant. In Ensembles 1–3, the group of intervals where the upper note had at least 6 harmonics in common with the first 32 harmonics of the interval's lower note were almost always (if not significantly) smaller than the group of intervals with fewer partials in common, although the amount of the difference varied both amongst ensembles and between the voices in which the intervals occurred. In Ensemble 4, in contrast, the distance from Just Intonation tunings were generally the same between the groups of intervals.

	All intervals	Intervals between the middle and upper voices	Intervals between the lower voice and the upper voices
<b>Ensemble 1 – Intervals with at least 6 harmonics in common</b>	6.3865	4.8362	6.625
<b>Ensemble 1 – Other intervals</b>	<b>10.194</b>	<b>10.493</b>	9.2227
<b>Ensemble 2 – Intervals with at least 6 harmonics in common</b>	15.485	11.522	15.921
<b>Ensemble 2 – Other intervals</b>	15.918	14.68	20.152
<b>Ensemble 3 – Intervals with at least 6 harmonics in common</b>	10.178	9.7295	10.117
<b>Ensemble 3 – Other intervals</b>	<b>12.559</b>	8.9419	<b>20.309</b>
<b>Ensemble 4 – Intervals with at least 6 harmonics in common</b>	13.143	13.704	13.056
<b>Ensemble 4 – Other intervals</b>	12.991	12.728	13.846

Table 4.2.10: Results of the *t*-tests run on the deviations (in cents) from Just Intonation for the grouping of vertical intervals in Part Three into those that share a larger number of harmonics with the fundamental (P8, P5, M3) versus those that share a fewer number of harmonics. The bolded items indicate that the mean of the group of intervals was significantly larger than the group it was tested against.

*T*-tests were also used to determine if there was a difference in the interval when Ensembles 2, 3, and 4 sung on “mi” versus “ma” (see Figure 4.2.12). The *t*-tests, with a threshold of 0.05, showed that the mean of the melodic intervals interval size in the renditions sung to “mi” in each ensemble were not significantly different than the mean of those sung to “ma” for any of the ensembles. For the vertical interval sizes, there was no significant difference in the means for Ensemble 2 between the “mi” and “ma” renditions. For Ensemble 3, the “mi” intervals’ mean was 4 cents higher than the “ma” interval’s mean. This was reversed in Ensemble 4, where the “ma” intervals’ mean was 5 cents higher than the “mi” intervals.

	Melodic (“mi”)	Melodic (“ma”)	Vertical (“mi”)	Vertical (“ma”)
<b>Ensemble 2</b>	1.7	-0.1	4.3	8.6
<b>Ensemble 3</b>	2.3	-0.3	<b>5.3</b>	1.0
<b>Ensemble 4</b>	-1.8	0.2	-0.02	<b>4.8</b>

Table 4.2.11: Results of the *t*-tests run on absolute interval size normalized around zero of the melodic and vertical intervals from Part Three to evaluate if the syllable that the notes were sung to influence interval size. The bolded items indicate that the intervals sung to that syllable were significantly larger.

#### 4.2.2.4 Part Four: Praetorius' "Es ist ein Ros' entsprungen"

The experimental material in Part Four was the first verse of a setting of "Es ist ein Ros' entsprungen" by Michael Praetorius (1571–1621), a four-part vocal piece with a predominantly homophonic texture. Three ensembles recorded the piece: Pilot, Lab, and Church. The Pilot ensemble performed all 3 of their renditions of the piece in English. The Lab ensemble performed 7 renditions, 4 in German and 3 to the syllable "mi." The Church ensemble performed 8 renditions of the piece, 4 in German and 4 to the syllable "mi."

Table 4.2.12 shows the means and standard deviations for all of the melodic semitones and whole tones in the piece. The first six rows in the table show the various semitone conditions considered in this experiment: E-F leading tones (2 instances per rendition), non-leading tones E-F semitones (3 instances per rendition), F-E semitones (7 instances per rendition), chromatic semitones (1 instance per rendition), other ascending semitones (6 instances per rendition), and other descending semitones (5 instances per rendition). The bottom two rows show the two whole tone conditions: ascending (18 instances per rendition) and descending (33 instances per rendition). Due to the lack of instances in this piece, the conditions of chord tone versus non-chord tone for the whole tone intervals starting and ending notes, which were used in analysing "Ave Maria," were not applicable for this piece. As with the other parts of the experiment, box and whisker plots were used to visualize the variance of the data in each intervallic condition, and regressions were used to evaluate the significance of differences between the conditions. As with "Ave Maria," the semitone regression tested for intervallic direction, leading tone function, whether the semitone occurred between E-F/F-E or another pair of notes, and singer identity. The additional condition of chromatic versus diatonic was also examined.

Table 4.2.13 shows the means and standard deviations for the vertical intervals in the 21 sonorities marked in Figure 4.2.7: Minor Thirds (16), Major Thirds (20), Perfect Fourths (12), Perfect Fifths (27), Minor Sixths (12), and Major Sevenths (13). Only sonorities with a half note in the bass were considered. This was based on the hypothesis that the singers' vertical tunings are more consistent for longer tones and less influenced by melodic interval tuning. Vertical intervals were calculated between all of the voices: Bass-Tenor, Bass-Alto, Bass-Soprano, Tenor-Alto, Tenor-Soprano, and Alto-Soprano. As with the Parts One-Three, the amount of deviation from Just Intonation was compared between two groups of

intervals. The first group, where there was a greater degree of coincidence between the harmonics, consisted of the Major Thirds, Perfect Fifths, and Perfect Octaves. The second group, with a lesser degree of coincidence between the harmonics, consisted of the Minor Thirds, Perfect Fourths, Minor Sixths, and Minor Sevenths. The influence of cadence on deviation from Just Intonation was also evaluated.

<b>Melodic Intervals Types (Number of instances)</b>	<b>Pilot</b>		<b>Lab</b>		<b>Church</b>	
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
Leading tone semitones (2/14/16)	107.0	9.7	105.9	7.4	100.2	12.5
Non-leading tone E-F semitones (9/21/24)	105.2	19.8	105.3	15.3	107.5	20.1
F-E semitones (21/49/56)	101.3	14.7	97.2	15.0	90.5	18.4
Chromatic semitones (3/7/9)	112.3	10.0	99.9	18.9	96.6	23.0
Other semitones ascending (18/42/48)	99.8	23.1	102.7	17.8	100.4	17.8
Other semitones descending (15/35/40)	109.8	19.6	109.2	15.8	100.9	18.2
Whole tones ascending (54/126/144)	197.3	19.7	195.7	17.8	199.3	19.8
Whole tones descending (99/231/297)	203.0	20.1	201.2	16.6	204.0	19.1

Table 4.2.12: Mean and standard deviation of the interval sizes for all of semitones and whole tone sizes in Part Four for each ensemble.

<b>Vertical Intervals (JI values) (Number of instances)</b>	<b>Ensemble 1</b>		<b>Ensemble 2</b>		<b>Ensemble 3</b>	
	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
Minor Thirds [316 cents] (48/112/128)	303.2	20.5	308.8	18.4	304.5	19.4
Major Thirds [386 cents] (60/140/160)	401.4	16.0	387.3	19.3	397.8	19.6
Perfect Fourths [498 cents] (36/84/96)	504.1	18.8	506.6	19.1	494.5	19.9
Perfect Fifths [702 cents] (81/189/216)	697.3	14.8	696.4	17.8	704.6	18.3
Minor Sixths [814 cents] (36/84/96)	802.2	16.2	807.1	19.2	799.6	16.5
Major Sixths [884 cents] (39/91/104)	900.0	19.4	890.3	17.8	893.2	16.2
Minor Sevenths [1018] (3/7/8)	1017.1	44.8	1011	30.7	999.7	23.4
Perfect Octaves [1200 cents] (66/154/176)	1198	15.5	1195.9	17.4	1202.2	17.0

Table 4.2.13: Mean and standard deviation of the sizes of the vertical intervals in Part Four between the four voices for all of the sonorities with a half note in the bass (as marked in Figure 4.2.7).

Overall, the means of the semitone interval sizes in Table 4.2.12 varied both across conditions and amongst the ensembles. The whole tones' mean interval sizes had less variability, and for each ensemble, the mean size of the ascending whole tone was smaller than the descending one. The vertical interval data in Table 4.2.13 shows greater consistency amongst the ensembles for the Minor Third, Perfect Fourth, Perfect Fifth, Minor Sixth,

Major Sixth, and Perfect Octave intervals. The Pilot and Church ensembles had comparable means for the Major Third, which were all close to the equal tempered tuning (400 cents), whereas the Lab ensemble's mean for the Major Third was much closer to Just Intonation (386 cents). All three of the ensembles' standard deviations for the Major Thirds were comparable. There was also a wide range of means for the Minor Sevenths, with the Church ensemble's mean being closest to equal temperament (1000 cents), the Pilot ensemble's mean being closest to Just Intonation (1018 cents), and the Lab ensemble's mean falling somewhere in between. The box and whisker plots in Figure 4.2.19 and Figure 4.2.20 visualise the melodic and vertical intonation data, respectively.

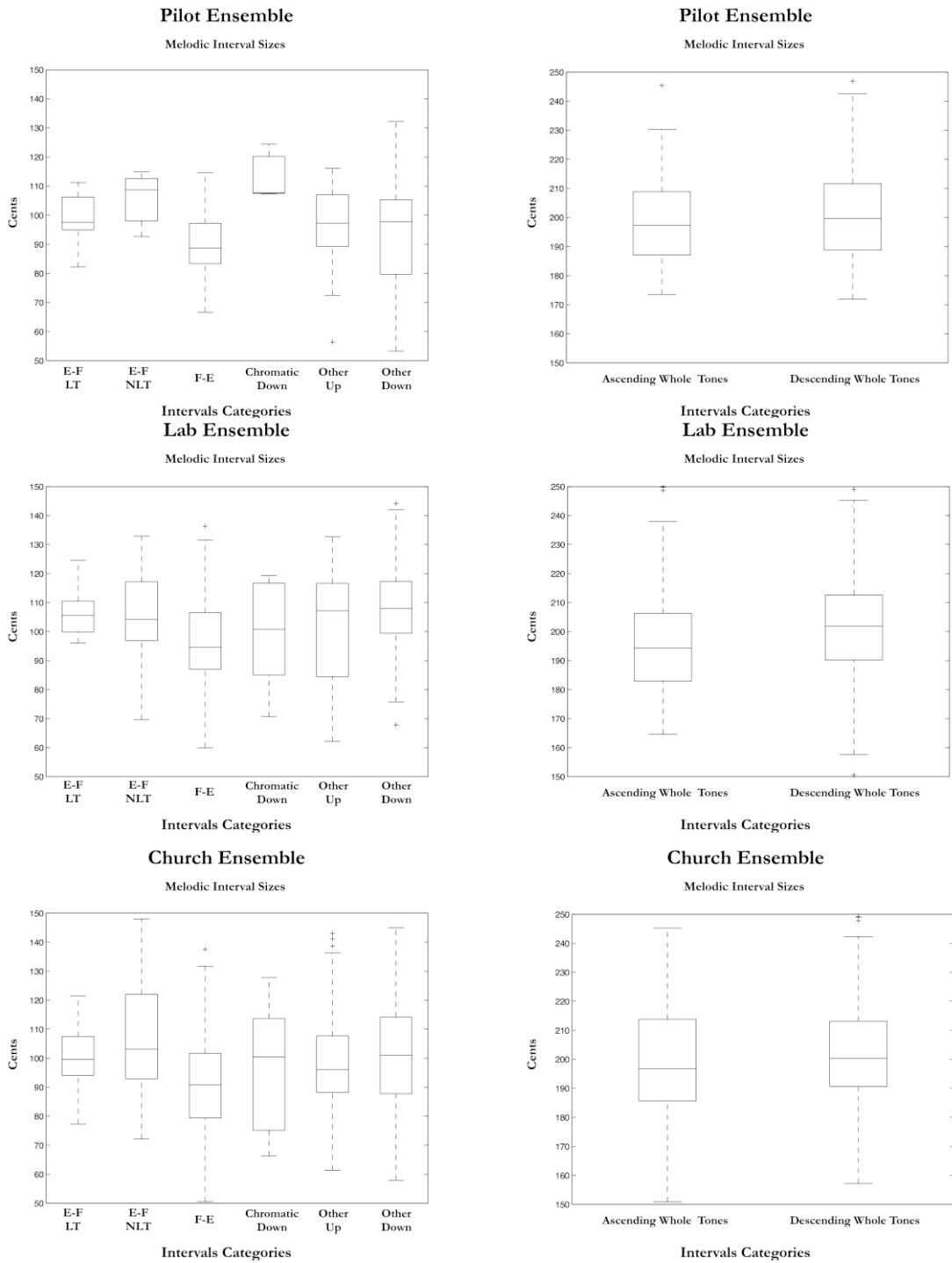


Figure 4.2.19: Box and whisker plots for the sizes of the melodic intervals in Part Four for each ensemble. The plots on the left show the data for the semitone conditions, while the plots on the right show the data for the whole tone conditions.

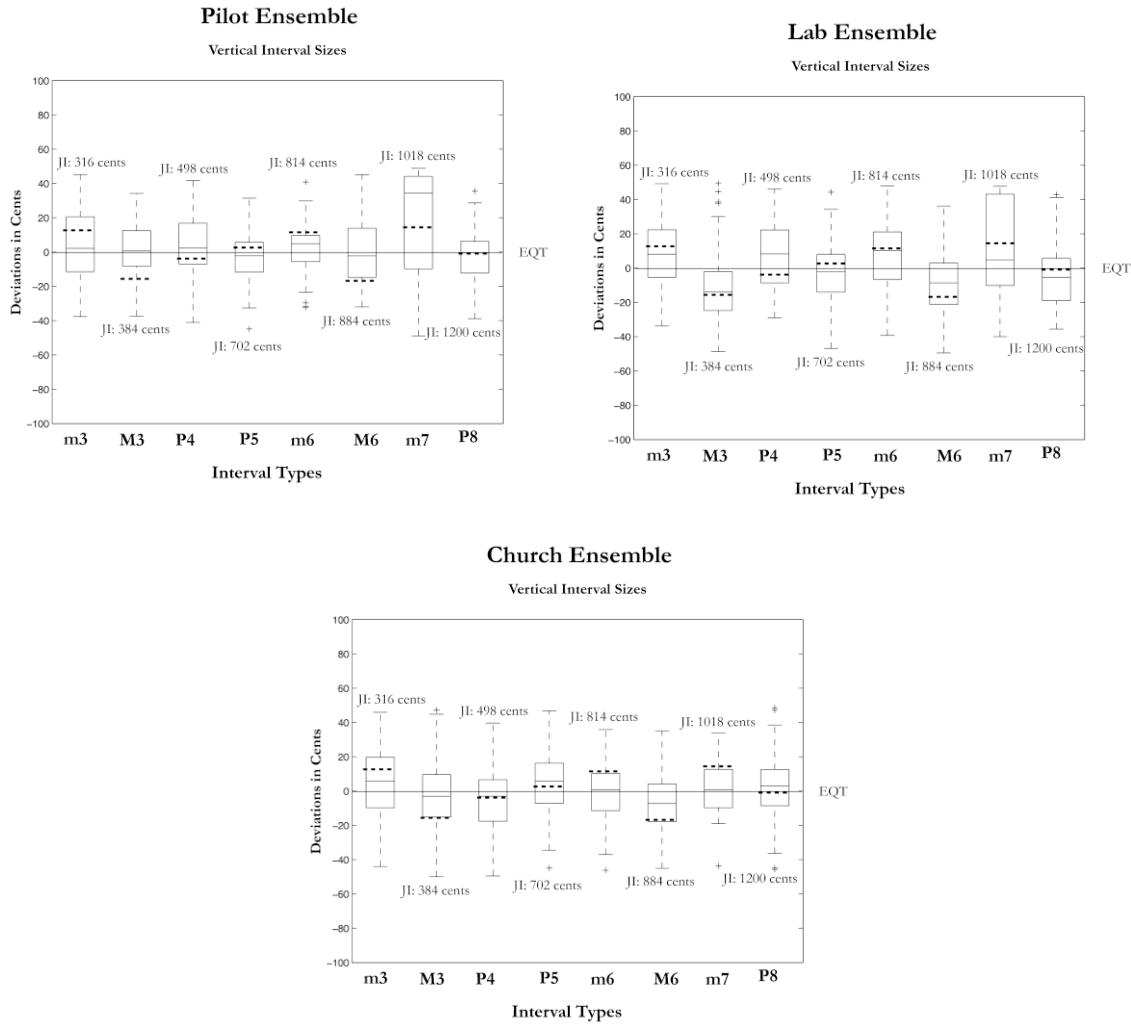


Figure 4.2.20: Box and whisker plots for the vertical intervals data in Part Four for each ensemble. The dotted lines represent the position of the idealized 5-limit Just Intonation tunings for each interval.

The box and whisker plots in Figure 4.2.19 shows that there was a lot of variation in the melodic interval size, as observable in the wide ranges for both the 25<sup>th</sup>–75<sup>th</sup> and 5<sup>th</sup>–95<sup>th</sup> percentiles. The exception to this is the E-F leading tone, although its smaller percentile ranges were likely due to the smaller number of instances for this condition. It is interesting, however, to observe the contrast between the percentile ranges for the E-F leading tones and the chromatic semitones, which only had the number of instances of the leading tones. The much larger percentile ranges for the chromatic semitones suggest that directed

harmonic activity, as occurs with leading tones, may influence singers' tuning, and that in its absence, the singers are less consistent in interval size.

All of the ensembles showed a similar pattern for the relative position of the medians of the E-F leading tone, non-leading tones E-F semitones, and the F-E semitones, with both categories of the E-F semitones being higher than the F-E semitones. The ascending versus descending whole tones exhibited similar relative median positions and percentile ranges across the ensembles. Overall, the vertical intervals in Figure 4.2.20 had comparable 5<sup>th</sup>–95<sup>th</sup> and 25<sup>th</sup>–75<sup>th</sup> intervals. The main exception to this was the wide percentile ranges for the Minor Sevenths in the Pilot and Lab ensembles. Also, with the exception of the Minor Seventh for the Pilot ensemble, the medians for both the Pilot and Church were closer to equal temperament than the Lab ensemble, whose medians tended slightly towards the values predicted by Just Intonation. It is also interesting to observe the large number of outliers for the Major Third in the Lab ensemble, which indicate that while the ensemble had a general trend towards the Just Intonation tuning (386 cents), there were occasions where the thirds were tuned much wider.

A linear regression analysis of the semitone interval size data for the Pilot ensemble was not significant ( $p = 0.13$ ). The analysis was significant for the Lab ensemble, although the  $R^2$  value was small ( $R^2 = 0.10$ ,  $p < 0.01$ ). The Lab Ensemble regression only showed a significant effect for the soprano, which was 13 cents larger than the bass (95% confidence interval = [5,21]). The regression on the Church ensemble explained more of the variance in the ensemble's data ( $R^2 = 0.14$ ,  $p < 0.0001$ ) and showed significant effects for direction, with the descending semitones being on average 8 cents larger than the ascending ones (95% confidence interval = [3,13]), and singer identity. In relation to the soprano, the bass' semitones were on average 14 cents larger (95% confidence interval = [5,22]), the tenor's semitones were on average 3 cents smaller (95% confidence interval = [1,14]), and the alto's were on average 11 cents larger (95% confidence interval = [4,18]). None of the ensembles' regressions showed any effects for leading tone function (i.e., whether the semitone occurred between E-F/F-E or if the spelling was chromatic versus diatonic).

The regression analysis for the whole tone data was also not significant for the Pilot ensemble ( $p = 0.24$ ), and the  $R^2$  values for both the Lab and Church ensembles were even smaller than the corresponding semitone regressions. The regression analysis for the Lab

ensemble's data ( $R^2 = 0.033$ ,  $p < 0.03$ ) showed a significant effect for intervallic direction, with the descending semitones being on average 4 cents larger than the ascending ones (95% confidence interval = [0.3,9]). The regression analysis of the Church ensemble's data ( $R^2 = 0.028$ ,  $p < 0.03$ ) showed significant effects for both interval direction and singer identity, with the descending whole tones being on average 5 cents smaller than the ascending ones (95% confidence interval = [1,9]) and the tenor's whole tones being on average 6 cents smaller than the soprano's (95% confidence interval = [1,11]).

For the vertical interval data, *t*-tests, with the threshold set to 0.05, were run to evaluate two different conditions. The first was whether intervals that share a greater number of harmonics were tuned closer to the interval sizes predicted by 5-limit Just Intonation, and the second was whether the intervals were tuned close to Just Intonation at cadences than at other points in the piece. Table 4.2.14 shows the results of the *t*-tests for the first condition, and Table 4.2.15 shows the results for the second.

	All intervals	Intervals between the middle and upper voices	Intervals between the lower voice and the upper voices
Pilot – Intervals with at least 6 harmonics in common	12.776	13.127	12.675
Pilot – Other intervals	<b>18.197</b>	<b>17.203</b>	<b>24.075</b>
Lab – Intervals with at least 6 harmonics in common	16.768	18.450	16.208
Lab – Other intervals	16.463	15.977	18.740
Church – Intervals with at least 6 harmonics in common	15.273	15.430	15.228
Church – Other intervals	16.971	17.254	15.296

Table 4.2.14: Results of the *t*-tests run on the deviations (in cents) from Just Intonation tunings for the grouping of vertical intervals in Part Four into the P8, P5, M3 versus the remaining intervals. The bolded items indicate that the mean of the group of intervals were significantly larger than the other group it was tested against.

	Pilot	Lab	Church
Cadence	14.356	14.108	14.261
Other	<b>17.027</b>	<b>16.631</b>	<b>17.254</b>

Table 4.2.15: Results of the *t*-tests run on the deviations (in cents) from Just Intonation in vertical intervals in Part Four that occurred in the cadential progression versus those that occurred in non-cadential progressions.

Only the Pilot ensemble showed a significant effect for the different intervals groups in the first set of *t*-tests. For this ensemble, the second group of intervals was significantly larger than the first group, which shared a larger number of harmonics with the fundamental. This was the case regardless of whether the vertical intervals occurred between the bass and an upper voice or within the upper voices. For the second set of *t*-tests, however, there were significant effects for all of the ensembles. In these tests, the ensembles' deviation from Just Intonation was shown to be significantly smaller for those intervals that occurred in a cadence than in other sonorities.

As with Part Three, the influence of syllable was also evaluated with a set of *t*-tests, with a threshold of 0.05, and the interval data normalized around zero. Unlike Part Three, where the only variation between the syllabic condition was a single vowel (i.e., “mi” versus “ma”) here the question being evaluated was whether there was a difference when the piece was sung with its original German lyrics as opposed to the syllable “mi.” Since the Pilot ensemble sang all of their renditions in English, only the data from the Lab and Church ensembles were used in this evaluation. As detailed in Table 4.2.16, the impact was consistent when a significant difference could be observed. The mean of both the Lab ensemble’s and Church ensemble’s melodic intervals were significantly larger when sung to the German text rather than to “mi.” Only the Lab ensemble’s vertical intervals showed a significant effect for which text was sung, with the intervals sung in German being significantly larger.

	Melodic (German)	Melodic ("mi")	Vertical (German)	Vertical ("mi")
Lab	<b>3.180</b>	0.128	<b>0.4284</b>	-4.385
Church	<b>1.971</b>	-1.702	-0.205	1.089

Table 4.2.16: Results of the *t*-tests run on the absolute interval size normalized around zero in Part Four and grouped into those takes sung in German and those sung to the syllable “mi” for the Lab and Church ensembles. The bolded items indicate the groups of intervals that were significantly larger.

### 4.2.3 Discussion

#### 4.2.3.1 Semitones

Semitone melodic intervals occurred in Parts One and Four, although there were only ascending semitones in Part One but both ascending and descending semitones in Part Four. The means and standard errors of the ascending and, where applicable, descending semitones across each ensemble in Parts One and Four are shown in Table 4.2.17. With the exception of the Lab ensemble in Part One, the means of the ascending intervals were closer to equal temperament (100 cents) than either Pythagorean tuning (90 cents) or Just Intonation (112 cents). In contrast, the mean for the Lab ensemble's ascending semitones was much closer to the Pythagorean tuning. The means for the descending intervals for the Pilot and Church ensembles in Part Four were both significantly smaller than the ascending semitones, whereas the mean for Lab ensemble's descending semitones was comparable to the mean for its ascending semitones.

	Ascending Semitone		Descending Semitones	
	Mean	Standard Error	Mean	Standard Error
<b>Part One – Lab</b>	91.0	1.4	-	-
<b>Part One – Church</b>	100.8	1.7	-	-
<b>Part Four – Pilot</b>	99.1	2.3	94.1	2.2
<b>Part Four – Lab</b>	104.1	1.8	104.5	1.4
<b>Part Four – Church</b>	102.3	1.9	97.0	1.5

Table 4.2.17: Summary of the means and standard errors for the ascending and descending semitones across each ensemble in Parts One and Four.

Although Part One contained both chromatic and diatonic semitones, there was no significant effect in interval size between these two groups. Following from this, there was only a minimal effect for the different interval types within the groups, with only one significant effect emerging in the linear regression analysis of the Lab ensemble's and Church ensemble's data. In each case, one of the types of chromatic semitone' average size was significantly smaller than the  $\hat{7}-\hat{8}$  semitones. In Part Four, there were fewer semitone conditions to consider (only leading tone versus non-leading tone and E-F/F-E semitones versus non E-F/F-E semitones) and none of them showed significant effects in the linear regression analysis. In both parts, however, there were significant effects for singer identity.

#### 4.2.3.2 Whole tones

Whole tone melodic intervals occurred in Parts Two, Three, and Four, with only ascending whole tones in Part Two and both ascending and descending intervals in Parts Three and Four. The means and standard errors of each ensemble's whole tones in each part are detailed in Table 4.2.18. Overall, the majority of the means for the ensembles' ascending and descending whole tones were closest to equal temperament (200 cents), with the descending semitones in the Lab and Church ensembles in Part Four sitting between the equal tempered value and the Pythagorean/Major Just Intonation whole tone tuning (204 cents). Ensemble One in Part Three had the smallest means: 191.9 cents for ascending intervals and 189.6 cents for descending intervals, which were the only values that were closer to the Just Intonation Minor whole tone tuning (182 cents) than equal temperament.

	Ascending Whole Tones		Descending Whole Tones	
	Mean	Standard Error	Mean	Standard Error
<b>Part Two – Lab</b>	195.1	5.5	-	-
<b>Part Two – Church</b>	200.8	2.3	-	-
<b>Part Three – Ensemble 1</b>	191.9	1.9	189.6	1.5
<b>Part Three – Ensemble 2</b>	199.6	2.1	196.6	2.2
<b>Part Three – Ensemble 3</b>	199.2	1.5	197.7	1.4
<b>Part Three – Ensemble 4</b>	192.9	2.0	195	1.4
<b>Part Four – Pilot</b>	199.5	2.2	199.9	1.6
<b>Part Four – Lab</b>	195.7	1.6	201.2	1.1
<b>Part Four – Church</b>	198.6	1.6	202.1	1.1

Table 4.2.18: Summary of the means and standard errors for the ascending and descending whole tones across each ensemble in Parts Two, Three, and Four.

In the linear regression analysis, there were no significant effects for direction in Part Three. In Part Four, there was a significant effect for direction in the Lab and Church ensembles, with the average size of the ascending intervals in both ensembles being significantly smaller than the average size of the descending intervals. As with the corresponding semitone exercises in Part One, there was only a minimal effect for the different whole tone types in Part Two. The only statistically significant effect was in the Lab ensemble's data, where the  $\hat{5} - \hat{6}$  whole tones were smaller on average than the  $\hat{6} - \hat{7}$  whole tones. Due to the lack of directed harmonic activity in Part Three and the lack of instances of the non-chord tone conditions in Part 4, different types of whole tones were not considered for these parts.

There were significant effects for singer identity when a regression was run using the upper voice in each ensemble as a baseline. In Part One, there were significant effects for the Church (tenor and alto) ensemble. In Part Two, there were significant effects for both the Lab (tenor and alto) and the Church ensembles (alto). In Part Three, there were significant effects for Ensemble 1 (ATB, pilot; tenor) and Ensemble 2 (ATB, lab; tenor). In Part Four, there were significant effects for both the Lab (bass) and Church (bass, alto, and tenor) ensembles for the semitone data, and for the Church (tenor) ensemble for the whole tone data.

#### 4.2.3.3 Vertical Intervals

Vertical interval tuning was evaluated for Parts 1–4. The main question being evaluated with the vertical intervals was whether intervals with a greater number of harmonics between the upper and lower notes were tuned closer to Just Intonation since the coincidence of harmonics could encourage a tuning that is derived from the lower harmonics of the harmonic series. The intervals were divided into two categories: those where the upper note in the interval have at least 6 harmonics in common with the lower note's first 32 harmonics and those where the upper note has a fewer number of harmonics in common. Of the vertical intervals that occurred in Parts 1–4, the Perfect Octave, Perfect Fifth, and Major Third fell into the first category, whereas the Minor Seventh, Minor third, Perfect Fourth, Minor Sixth, and Major Sixth fell into the second category. The means and standard errors of the intervals in the first category across each ensemble for each part are shown in Table 4.2.19. The means and standard errors for the second category for each part are shown in Table 4.2.20. The means and standard errors in Table 4.2.19 and Table 4.2.20 convey information about how close the different ensembles were on average to the tunings predicted by Just Intonation. Those ensembles with means that were within two standard errors of the Just Intonation tuning can be considered to encompass the Just Intonation tuning. The intervals for which the largest proportion of the ensembles' average tunings encompassed the relevant Just Intonation tuning included, as anticipated, those with the largest number of harmonics in common: Perfect Octave (6/11) and Perfect Fifth (7/11). This was not the case, however, for the Major Third (1/11), which suggests another factor may have influenced the tuning. Specifically, that the Just Tunings for the Perfect Octave and Perfect Fifth are closer to the equal tempered tuning, whereas the Just Tuning for the Major Third is 14 cents smaller than equal temperament.

In the second group of intervals, those with less than 6 harmonics in common between the first 32 harmonics of the two notes in the interval, the close relationship between the Just Intonation and equal temperament tunings for the Perfect Fourth (3/7) and the Tritone (2/2) could also explain why the means of these intervals were relatively close to Just Intonation tuning compared to the other intervals in the group. The Just Intonation tunings for the Minor Third (0/11), the Minor Sixth (1/5) and the Major Sixth (0/9) are 14–18 cents away from equal temperament and only a small proportion of ensembles' means encompassed the Just Intonation tuning. The Minor Seventh was the exception in this regard since the difference between the Just Intonation and equal tempered tuning is 18 cents, yet 2 out of the 5 means encompassed the Just Tuning. This was not, however, because the Minor Seventh was on average closer to equal temperament, but rather because the standard errors were large enough to encompass both the Just Intonation and the equal tempered tuning. These findings suggest that, overall, the singers were tuning their intervals closer to equal temperament than Just Intonation.

	Perfect Octave (JI: 1200)		Perfect Fifth (JI: 702)		Major Third (JI: 386)	
	Mean	Std. Error	Mean	Std. Error	Mean	Std. Error
<b>Part One, Lab (P1 1)</b>	1195.4	3.7	<b>713.8</b>	1.5	<b>400.6</b>	3.5
<b>Part One, Church (P1 2)</b>	1203.1	3.1	705.5	2.0	<b>408.3</b>	3.1
<b>Part Two, Lab (P2 1)</b>	<b>1209.3</b>	2.9	<b>715.5</b>	3.1	<b>403.1</b>	3.7
<b>Part Two, Church (P2 2)</b>	<b>1210.9</b>	3.0	<b>708.3</b>	2.7	<b>394.1</b>	3.5
<b>Part Three, Ensemble 1 (P3 1)</b>	1201.3	1.9	700.7	0.9	<b>375.6</b>	1.9
<b>Part Three, Ensemble 2 (P3 2)</b>	<b>1205.9</b>	2.6	704.9	1.7	<b>413.3</b>	1.9
<b>Part Three, Ensemble 3 (P3 3)</b>	<b>1208.5</b>	2.0	704.0	1.3	<b>397.4</b>	1.9
<b>Part Three, Ensemble 4 (P3 4)</b>	1202.3	2.4	701.6	1.3	<b>405.8</b>	2.4
<b>Part Four, Pilot (P4 1)</b>	1198	1.9	697.3	1.7	<b>401.4</b>	2.1
<b>Part Four, Lab (P4 2)</b>	<b>1195.9</b>	1.4	<b>696.4</b>	1.3	387.3	1.6
<b>Part Four, Church (P4 3)</b>	1202.2	1.3	704.6	1.2	<b>397.8</b>	1.5

Table 4.2.19: Summary of the means and standard errors (SE) across each ensemble in each experiment part for the vertical intervals where the upper note in the interval had at least 6 harmonics in common with the lower note's first 32 harmonics. Bolded values indicate those means that are greater than twice the standard error away from the Just Intonation tuning.

	Minor Seventh (JI = 1018)		Augmented Sixth (JI=977)		Minor Third (JI = 316)		Perfect Fourth (JI = 498)		Minor Sixth (JI = 814)		Major Sixth (JI = 884)		Tritone (JI = 590/ 610)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
P1 1	-	-	989.6	3.5	307.9	2.8	-	-	-	-	900.0	2.4	-	-
P1 2	-	-	995.3	4.9	303.0	2.7	-	-	-	-	894.4	2.4	-	-
P2 1	-	-	-	-	302.0	3.3	-	-	799.3	4.0	-	-	604.1	11.8
P2 2	-	-	-	-	294.7	4.6	-	-	809.3	4.7	-	-	603.8	4.5
P3 1	-	-	-	-	322.4	2.1	508.7	1.9	-	-	893.0	1.8	-	-
P3 2	-	-	-	-	299.5	3.0	497.0	2.8	-	-	903.1	3.6	-	-
P3 3	-	-	-	-	307.0	2.0	507.2	2.0	-	-	904.3	2.8	-	-
P3 4	-	-	-	-	300.8	3.1	492.9	2.0	-	-	896.1	2.3	-	-
P4 1	1017.1	22.4	-	-	303.2	3.0	504.1	3.2	802.2	2.7	900.0	3.1	-	-
P4 2	1011.0	9.7	-	-	308.8	1.8	506.6	2.1	807.1	2.1	890.3	1.9	-	-
P4 3	999.7	8.3	-	-	304.5	1.8	494.5	2.0	799.6	1.7	893.2	1.6	-	-

Table 4.2.20: Summary of the means and standard errors (SE) across each ensemble in each experiment part for the vertical intervals where the upper note in the interval had less than 6 harmonics in common with the lower note's first 32 harmonics. Bolded values indicate those means that are greater than twice the standard error away from the Just Intonation tuning. The abbreviations on the left correspond to the full names for the ensembles listed in Table 4.2.19.

T-tests were used to explore the effect of the coincidence of harmonics on vertical tuning more generally with the groupings shown in Table 4.2.19 and Table 4.2.20. In Parts One and Two, the distance of each group's average interval sizes from Just Intonation were not statistically significant for the Lab ensemble, but were for the Church ensemble. In Parts Three and Four, a distinction was made between whether the intervals occurred between the bass and an upper voice or amongst the upper voices to see if intervals sung in relation to the bass tended to be closer to Just Intonation. This was the case in Ensemble 3 (SAT, church) in Part Three, where intervals between the upper voices and the lowest voice showed a significant difference between the two groups, whereas the intervals between the upper voices did not. In Ensemble 1 (ATB, pilot), the opposite was true. For this ensemble, there was a significant difference between the two groups for intervals occurring in the upper voices, but not for those between the upper and lowest voice. There were no significant differences for Ensemble 2 (ATB, lab) and 4 (ATB, church). In Part Four, there were only significant differences for the Pilot ensemble, where the intervals that had less partials in common with both the group of intervals between upper voices and the bass and

the group of intervals between the upper voices were significantly further away from Just Intonation tunings. Part Four also allowed for the possibility of dividing up the vertical intervals between those that occurred in a cadential context and those that did not. For all of the ensembles, the vertical intervals were on average closer to the Just Intonation tunings than those that occurred in a non-cadential context.

#### 4.2.3.4 Influence of Syllable

The influence of syllables on tuning was also briefly considered, although this is a much larger question that will require additional studies to address properly. Parts Three and Four offered the opportunity to explore this question since some of the ensembles sung the music on different syllables. In Part Three, Ensembles 2–4 sang roughly half of their takes on “mi” and the rest of “ma.” In Part Four, the Lab and Church ensembles sang some of their takes in the original German and some to “mi.” The recordings in Part Three were more controlled since the difference between the two sets is a single syllable. Overall, there were no significant effects for melodic interval size. For vertical interval size, only Ensembles 3 and 4 showed a significant effect. In Ensemble 3, the average size of the vertical intervals sung to “mi” was significantly larger than the average size of the vertical interval sung to “ma.” In contrast, in Ensemble 4, the average interval size of the vertical intervals sung to “ma” was significantly larger than those sung to “mi.” Part Four was more complex since what was being evaluated was the tuning difference between when the piece was sung with its original lyrics versus a single syllable. There were significant effects in the Lab ensemble for both melodic and vertical intervals, with the average size of the melodic intervals being smaller for those sung in German than those sung to “mi” and the average size of the vertical intervals being larger for those sung in German than those sung to “mi.” For the Church ensemble, there was only a significant effect for melodic intervals: with the intervals sung in German also being larger on average than those sung to “mi.”

### **4.3 Analysis of Both Experiments' Data**

This section considers the commonalities and differences in the interval size data across the solo experiment in Section 4.1 and the SATB ensemble experiment in Section 4.2. Specifically, this section summarises the ways in which the interval size data conforms to and deviates from formal tuning systems, the influence of direction on melodic intervals, the impact of musical context on interval size, the amount of individual variation across singers, the influence of training/experience, and some possible explanations for the size of the  $R^2$  values in the regressions that were run on the interval size data. Overall, there was a large amount of variation in interval size both within individual singers' takes and across singers and groups/ensembles, much more than was anticipated before the experiments were undertaken.

#### **4.3.1 Relationship to Formal Tuning Systems**

Overall, the mean interval sizes were closer to equal temperament than either Pythagorean tuning or 5-limit Just Intonation, with a few notable exceptions. For the melodic intervals, this was overwhelmingly the case for both ascending and descending whole tones, but for the semitones, there was a marked difference between ascending and descending, particularly for the solo singers. In the solo experiment, the means of the professional singers' ascending semitones were close to equal temperament (100 cents), while the means of the non-professional group were smaller overall. In both groups of solo singers, however, the means of the descending semitones were smaller than the ascending ones. The means of both the ascending and descending semitones for the non-professional group and the descending semitones for the professional group were closer to the Pythagorean/Major Just Intonation semitones (90 cents) than equal temperament. This was also true for the Lab ensemble's ascending semitones in Part One of the ensemble experiment and the Pilot ensemble's descending semitones in Part Four.

As discussed in Section 4.2.3, the means of the vertical intervals were generally closer to equal temperament than 5-limit Just Intonation, except for those intervals where the sizes of the equal tempered and Just Intonation intervals were close to one another: the Perfect Octave (1200 cents for both), the Perfect Fifth (700 for equal temperament versus 702 for Just Intonation), and the Perfect Fourth (500 for equal temperament versus 498 for Just Intonation). Overall, the interval sizes were also closer to Pythagorean than Just Intonation,

particularly when the Pythagorean tunings were closer to Equal Temperament than Just Intonation. It should, however, be noted that the large standard deviations for both the melodic and vertical intervals indicate that while the means are centered around the values predicted by equal temperament, the singers were not singing in equal temperament.

The motivation for focusing on Just Intonation, rather than Pythagorean tuning, in the statistical analysis of the vertical interval sizes was that the Just Intonation tunings are derived from lower harmonics in the overtone series than Pythagorean tuning. The question underlying this analysis was whether singers tend more towards Just Intonation tunings when there is a greater coincidence of harmonics between the interval's notes. Overall, there were mixed results when the intervals were divided between those that had at least 6 of the first 32 harmonics in common (P8, P5, M3) and the remaining intervals. This is due in part to the wide range of tunings for the Major Third, where the ensembles' means ranged from 376–408 cents.

#### **4.3.2 Influence of Intervallic Direction on Interval Size**

For both the semitones and whole tones, there were significant effects for direction observed in some of the groups/ensembles studied, as detailed in Table 4.3.1, which shows the relationship of the interval size of the descending intervals to the ascending intervals. For both the semitones and whole tones, there were some groups/ensembles for which the difference between ascending and descending was not significant. When the difference was significant, however, two general tendencies emerged: the descending semitones tended to be smaller than the ascending ones, and the descending whole tones tended to be larger than the ascending ones. As described above in Section 4.3.2, the ascending semitones' means were closer to equal temperament (100 cents), while and descending semitones were closer to the Pythagorean tuning (90 cents). This was not the case for the whole tones, as in those ensembles/groups where there was a significant difference between the ascending and descending whole tones, the descending whole tones' means tended to be slightly larger than equal temperament, whereas the ascending whole tones means' tended to be slightly smaller than equal temperament.

Group/Ensemble	Semitones	Whole Tones
Non Pro	8 Cents Smaller	5 Cents Larger
Pro	7 Cents Smaller	NS
P3 Pilot	-	NS
P3 Lab	-	NS
<i>P3 Church 1</i>	-	-
<i>P3 Church 2</i>	-	-
<i>P4 Pilot</i>	-	-
P4 Lab	NS	5 Cents Larger
P4 Church	8 Cents Smaller	5 Cents Larger

Table 4.3.1: Summary of the results for intervallic direction from the regressions run on the melodic interval data. The “Larger/Smaller” labels refer to the relationship of the descending intervals to the ascending intervals. Italics are used to indicate those ensembles for which the regression was not significant ( $p > 0.05$ ), and NS indicates that an individual condition was not significant.

Overall, the differences between the ascending and descending semitones were in the 7–8 cents range, and the differences between ascending and descending whole tones were around 5 cents. It is hard to know if these values, or the differences between them, are significant, considering the wide range of values reported for the Just Noticeable Difference of pitch stimuli (see Section 2.3.1). It is, however, interesting to note that where the differences between the ascending and descending intervals are significant, the trends for the semitones and whole tones are opposite of one another. One possible explanation is that there is a compensatory effect happening to avoid excessive drift, where the ascending semitones are generally larger in order to correct for generally smaller ascending whole tones and vice versa. It is interesting to note that these trends emerge in two different pieces, one for the solo singers and one for the SATB ensembles, which suggests that the trends are not restricted to a particular arrangement of musical materials.

### 4.3.3 Role of Musical Context

The regression analyses of the melodic interval data explored the role of musical context on interval size. Overall, there was a minimal effect for the musical context in the solo experiment and virtually no effect for the ensemble experiment. For both “Ave Maria” in the solo experiment and “Es ist ein Ros entsprungen” (Part Four) in the ensemble experiment, two semitone conditions were evaluated. The first was whether there was a

significant difference between the mean interval sizes of leading tones and non-leading tones. The second was whether there was a difference between the mean interval sizes of semitones between the same pitch classes as the leading tone (whether or not they had a leading tone function) and semitones between other pitch classes. Table 4.3.2 shows which regressions and conditions showed significant effects. For the “Ave Maria,” there was a significant effect for the non-professional group in the first condition, with the non-leading tones being larger. In contrast, the professional group had a significant effect in the second condition, with the semitones between other pitch classes being larger. For “Es ist ein Ros entsprungen,” there were no significant effects found for either condition in any of the ensembles.

Group/Ensemble	LT	Other
Non Pro	7 Cents Smaller	NS
Pro	NS	7 Cents Larger
<i>P4 Pilot</i>	-	-
P4 Lab	NS	NS
P4 Church	NS	NS

Table 4.3.2: Summary of the results for different semitone conditions from the regressions run on the melodic semitone data for the solo experiment and Part Four of the ensemble experiment. In the first column, the “Larger/Smaller” labels refer to the relationship of the semitones with a leading tone function to those without. In the second, it refers to the relationship of semitones between A and B<sub>j</sub> to the semitones between other pitch classes. Italicization is used to indicate those ensembles for which the regression was not significant ( $p > 0.05$ ), and NS indicates that an individual condition was not significant.

Part One of the ensemble experiment offered a more controlled setting in which the influence of musical context on semitone size could be evaluated. In the progressions in Part One, there were five chromatic and six diatonic semitones. The semitones in each group were between the same pitch classes in different harmonic contexts and occurred sequentially in each of the upper three voices. As detailed in Table 4.3.3, the only significant effect was that in each of the ensembles, one of the chromatic semitone types was on average significantly larger in size than the diatonic semitone with a leading tone function.

Group/Ensemble	Chromatic Semitones					$\hat{2}-\hat{3}$	$\hat{3}-\hat{4}$	$\hat{5}-\hat{6}$
	1	2	3	4	5			
P1 Lab	NS	NS	NS	NS	14 Cents Larger	NS	NS	NS
P1 Church	NS	16 Cents Larger	NS	NS	NS	NS	NS	NS

Table 4.3.3: Summary of the results for different semitone conditions from the regressions run on the melodic semitone data for Part One of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the semitones indicated in the column header to the  $\hat{7}-\hat{8}$  semitones, which were used as a baseline. NS indicates that an individual condition was not significant.

The results for the whole tones were similar. In the “Ave Maria,” there was a significant effect in the non-professional group for whether a whole tone started or ended on a chord tone, shown in Table 4.3.4. The whole tones that ended on a chord tone were smaller on average than those that ended on a non-chord tone. There were no significant effects for the non-professional group.

Group/Ensemble	Starting Note	Ending Note
Non Pro	NS	4 Cents Smaller
Pro	NS	NS

Table 4.3.4: Summary of the results for different whole tone conditions from the regressions run on the melodic whole tone data for the solo experiment. The “Larger/Smaller” labels refer to the relationship of the size of the whole tones that either started or ended on a chord tone to those that either started or ended on a non-chord tone. NS indicates that an individual condition was not significant.

Part Two was, like Part One, designed to present whole tones in different musical contexts. However, as with Part One, the significant effects for the average interval size across the different conditions were minimal, as detailed in Table 4.3.5. The only significant effect was that the whole tones between the fifth and sixth scale degrees were smaller on average than those between the sixth and seventh scale degrees, which were used as the baseline.

Ensembles	$\hat{2}-\hat{3}$	$\hat{5}-\hat{6}$	$\hat{4}-\hat{5}$	$\hat{3}-\hat{4}$	$\hat{1}-\hat{2}$
P2 Lab	NS	14 Cents Higher	NS	NS	NS
P2 Church	NS	NS	NS	NS	NS

Table 4.3.5: Summary of the results for different whole tone conditions from the regressions run on data for Part One of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the semitones indicated in the column header to the  $\hat{6}-\hat{7}$  semitones, which were used as a baseline. NS indicates that an individual condition was not significant.

For the vertical intervals, musical context was considered in Part Four of the ensemble experiment with *t*-tests that evaluated whether the ensembles tuned closer to Just Intonation in cadential contexts than in non-cadential contexts. The *t*-tests showed a significant difference for all of the ensembles between the two contexts, with the intervals in the cadential context being significantly closer on average to Just Intonation tuning than the intervals that were not.

The effects for the musical context amongst the solo singers were group-dependant and thus mostly likely due to the singers’ amount of training/experience. The influence of training/experience is discussed below in Section 4.3.5. In the ensembles’ melodic interval size data, the lack of significant effects for the different melodic contexts in the ensemble data indicate that there may be other factors influencing the intonation than the ones considered. It is also possible that the exercises in Part One and Two were too short to create sufficient musical context and that they were too unfamiliar for the singers to be consistent in their performances since they were effectively sight-reading. Ways of improving this experiment will be discussed in the future works section in Chapter 5.

#### 4.3.4 Individual Variation Amongst Singers

The regression analyses also considered singer identity. Overall, there were more significant effects for singer identity than for musical context, although it does not explain all of the variation in the data. Also, the amount of the effect varied across the experiments. The results of the regression analysis for semitone data from the solo experiment are shown in Table 4.3.6. Overall, the non-professionals showed more of an effect than the professional group for singer identity, with four out of the five singers’ mean semitone size differing

significantly from the baseline compared to only one singer in the professional group. There was also a significant effect for group identity, with the non-professional group's semitones being smaller on average than the professional group's semitones.

<b>Group</b>	<b>Singer 1</b>	<b>Singer 2</b>	<b>Singer 3</b>	<b>Singer 4</b>	<b>Singer 5</b>
Non-Professional	8 Cents Larger	7 Cents Smaller	NS	11 Cents Larger	8 Cents Smaller
Professional	NS	NS	NS	NS	2 Cents Larger

Table 4.3.6: Summary of the results for singer identity from the regressions run on the melodic semitone data in the solo experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to singer 6, who was used as a baseline. NS indicates that an individual condition was not significant.

The semitone size data in Parts One and Four in the ensemble experiment also showed significant effects for singer identity in some of the ensembles. In Part One, only the Church ensemble showed a significant effect, with both the tenor and alto singers' semitones being larger on average than the soprano's (Table 4.3.7). In Part Four, the Church ensemble also showed a strong effect, with the bass and alto's semitones being smaller on average than the soprano's and the tenor's being larger. The Lab ensemble only had a significant effect for one singer, the bass, whose semitones were on average smaller than the soprano's in that ensemble (Table 4.3.8).

<b>Ensemble</b>	<b>Tenor</b>	<b>Alto</b>
P1 Lab	NS	NS
P1 Church	12 Cents Larger	8 Cents Larger

Table 4.3.7: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part One of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to the soprano, who was used as a baseline. NS indicates that an individual condition was not significant.

Ensemble	Bass	Tenor	Alto
<i>P4 Pilot</i>	-	-	-
P4 Lab	13 Cents Smaller	NS	NS
P4 Church	14 Cents Smaller	7 Cents Larger	11 Cents Smaller

Table 4.3.8: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part Four of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to the soprano, who was used as a baseline. Italicization is used to indicate those ensembles for which the regression was not significant ( $p > 0.05$ ), and NS indicates that an individual condition was not significant.

The singer and group effects in the solo experiment’s whole tone data were similar to their effects in the semitone data. For the non-professional group, 3 singers’ means were significantly different than their baseline singer’s mean, while 2 of the professional group’s singers were different from their baseline (Table 4.3.9). There was also, as with the semitone data, a significant effect for group identity, with the professional singers having larger whole tones on average than the non-professional singers.

Group	Singer 1	Singer 2	Singer 3	Singer 4	Singer 5
Non Professional	NS	NS	10 Cents Smaller	7 Cents Larger	6 Cents Smaller
Professional	NS	5 Cents Larger	NS	NS	8 Cents Smaller

Table 4.3.9: Summary of the results for singer identity from the regressions run on the melodic whole tone data in the solo experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to singer 6, who was used as a baseline. NS indicates that an individual condition was not significant.

In the ensemble experiment, the regression analysis on the whole tone data from Parts Two, Three, and Four showed similar effects for singer identity to the statistical analysis of the semitone data in Parts One and Four. In Part One, there were significant effects for the Lab ensemble, where the tenor’s and alto’s whole tones were both larger on average than the soprano’s (Table 4.3.10). In Part Three, the middle voices in both the Pilot and Lab experiments differed from the upper voice (Table 4.3.11). In Part Four, there were fewer

significant effects for singer identity, with only the alto in the Church ensemble's whole tone's size differing significantly from the soprano's. (Table 4.3.12).

Ensemble	Tenor	Alto
P2 Lab	12 Cents Larger	9 Cents Larger
P2 Church	NS	Smaller

Table 4.3.10: Summary of the results for singer identity from the regressions run on the melodic whole tone data in Part Two of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to the soprano, who was used as a baseline. NS indicates that an individual condition was not significant.

Ensemble	Lowest Voice	Middle Voice
P3 Pilot	NS	14 Cents Larger
P3 Lab	NS	17 Cents Smaller
<i>P3 Church 1</i>	-	-
<i>P3 Church 2</i>	-	-

Table 4.3.11: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part Four of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to the upper voice, which was used as a baseline. Italicization is used to indicate those ensembles for which the regression was not significant ( $p > 0.05$ ), and NS indicates that an individual condition was not significant.

Ensemble	Bass	Tenor	Alto
<i>P4 Pilot WT</i>	-	-	-
P4 Lab WT	NS	NS	NS
P4 Church WT	NS	NS	6 Cents Larger

Table 4.3.12: Summary of the results for singer identity from the regressions run on the melodic semitone data in Part Four of the ensemble experiment. The “Larger/Smaller” labels refer to the relationship of the singer indicated in the column header to the soprano, who was used as a baseline. Italicization is used to indicate those ensembles for which the regression was not significant ( $p > 0.05$ ), and NS indicates that an individual condition was not significant.

#### 4.3.5 Impact of Training and Experience

In the ensemble experiment, there were no clear differences between the semi-professional ensemble and the two professional ensembles. There were, however, significant differences between non-professional and professional groups of singers in the solo experiment. For semitones, the significant differences for conditions, accompaniment, and singer identity allowed for several interpretations about how intonation practices might be influenced by training and/or experience. Likewise an additional interpretation can be made from the whole tones' significant differences for intervallic direction and whole tone conditions.

The absence of a significant effect in interval size for the leading tones in the professional group (Table 4.3.2), in contrast to the non-professional group which showed a significant effect, allows for two different interpretations. One is that with training the singers acquire greater stability in their production of leading tones or that the singers with less training tend to exaggerate them. Another is that the existence of a significant effect for accompaniment (Table 4.3.13) and greater prevalence of a significant effect for singer identity (Table 4.3.6) 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.

Groups	ST	WT
Non Pro	3 Cents Larger	NS
Pro	NS	NS

Table 4.3.13: Summary of the results for accompaniment from the regressions run on the melodic data in the solo experiment. The “Larger/Smaller” labels refer to the relationship of the accompanied intervals to the *a cappella* intervals for the interval types listed in the columns. NS indicates that an individual condition was not significant.

In terms of whole tone interval size, the main distinction between the non-professional and professional groups was of the significant effect for direction and the ending not in the non-professional groups (Table 4.3.1) versus the lack of significant effects in the professional group (Table 4.3.4). Unlike the semitone intervals, there were no significant effects for the presence of accompaniment. These observations suggest that training/experience reduces

the influence of direction on whole tone interval size and increases the influence of the musical material, such as the stability of the starting and ending notes of the intervals.

#### 4.3.6 $R^2$ Values in the Regressions

The  $R^2$  values, or coefficients of determination, for the regressions run on the data from the two experiments ranged from 0.02–0.42. The  $R^2$  values for the solo experiment's regressions are shown in Table 4.3.14, and the  $R^2$  values for the ensemble experiment's regressions are shown in Table 4.3.15. In Table 4.3.14 and Table 4.3.15, regressions with the same predictors are separated by dashed lines. Direct comparison between all of the  $R^2$  is not possible, since not all of the regressions shared the same predictors; however, general trends emerge when the regressors are taken into account. Overall, the  $R^2$  values in the solo experiment were less than 0.10, with the exception being the regression on the non-professional group's semitone data. For this regression, the significant effects for leading tones and singer identity compared to the professional ensemble's data explained a greater amount of the variation in the data. For the ensemble experiment, the higher  $R^2$  values in Parts One–Three were due to a greater number of significant effects for singer identity. The  $R^2$  values were closer between the solo and ensemble experiment when the musical material was more comparable, as can be observed with the comparable  $R^2$  values for the regressions run on the “Ave Maria” data in the solo experiment and “Es ist ein Ros entsprungen” (Part Four) data in the ensemble experiment. Overall, these  $R^2$  values indicate that there is unexplained variation in the interval size, which is due to some factors not considered, randomness, or both.

	Semitones	Whole Tones
Professional Group	0.09	0.06
Non-Professional Group	0.19	0.08
All Singers	0.07	0.02

Table 4.3.14:  $R^2$  values for the regressions run on the interval size data in Section 4.1. Only those cells separated by a dashed line share the same regressors, which allows for direct comparison of the  $R^2$  values.

	Semitones	Whole Tones
<b>Part One, Lab</b>	0.28	-
<b>Part One, Church</b>	0.25	-
<b>Part Two, Lab</b>	-	0.42
<b>Part Two, Church</b>	-	0.32
<b>Part Three, Ensemble 1</b>	-	0.48
<b>Part Three, Ensemble 2</b>	-	0.40
<b>Part Three, Ensemble 3</b>	-	NS
<b>Part Three, Ensemble 4</b>	-	NS
<b>Part Four, Pilot</b>	NS	0.24
<b>Part Four, Lab</b>	0.10	0.03
<b>Part Four, Church</b>	0.14	0.03

Table 4.3.15:  $R^2$  values for the regressions run on the interval size data in Section 4.2. Only those cells separated by a dashed line share the same regressors, which allows for direct comparison of the  $R^2$  values.

### 4.3.7 Conclusions

Overall, there was much more variation in the data than was expected at the outset of the experiments. It is not possible to determine if this variation is characteristic of the singers that were used or of singers in general, and more experiments are required to determine this. The future works section in Chapter 5 will discuss some possible modifications to the experiments and ways in which perceptual testing could be used to assess both how much variation in cents is significant and whether particular renditions are “in tune.” In the experiments detailed in this chapter, the question of whether a particular rendition was “in tune” was addressed in various ways. In the solo experiment, the solo singers, none of whom had absolute pitch, listened to their recordings and indicated that they considered their intonation to be accurate. For the Lab and Church ensembles, the conductor indicated the acceptability of the intonation for each rendition during the recording sessions. For the Pilot ensemble, which was not conducted, all of the renditions were included.

(This page intentionally left blank)

## Chapter 5 Conclusions

### 5.1 Summary of Dissertation

This dissertation examined the relationship of intonation to musical context through two experiments, one with solo singers and the other with SATB ensembles with one voice per part. Overall, the experiments show that the singers tended towards equal temperament; however, there was a wider range in interval size for both the melodic and vertical intervals than was anticipated at the start of the project. In the solo experiment, as was found in Prame's study (1997), the singers did not conform to equal temperament when performing the "Ave Maria." There were some significant effects found for intervallic direction for some of the singers in the melodic intervals, with the descending semitones tending to be smaller than the ascending semitones and the ascending whole tones tending to be smaller than the descending ones. There were also some significant effects found for the musical context in the solo intonation experiment, but they were group dependent. The non-professional group's leading tones tended to be smaller on average than their non-leading tone semitones. The non-professional group also tended to sing whole tones that ended on chord tones smaller than those that ended on non-chord tones. In the ensemble experiment, however, there were no such tendencies for musical context in the melodic intervals. The vertical intervals were analyzed to see if the amount that they deviated from Just Intonation was related to the amount of coincidence of harmonics between the two notes in the interval or whether the interval occurred in a cadential or non-cadential context. The vertical intervals were divided into two groups: intervals where the notes shared at least 6 harmonics in common amongst the first 32 harmonics and intervals where the notes had less than 6 harmonics in common. Overall, the differences between the two groups of intervals' deviations from Just Intonation were variable and group-dependent. However, when the intervals were divided into those that occurred in cadential versus non-cadential contexts, a general trend emerged that showed that overall the intervals that occurred in a cadential context were significantly closer to Just Intonation than those that occurred in a non-cadential context. These results differed from the findings by Howard (2007a; 2007b), who observed a much closer, although not strict, adherence to Just Intonation in the singers

and exercises he studied, which had far more modulations than those used in this experiment.

## 5.2 Summary of Original Contributions

As discussed in the Introduction, this dissertation explores a number of issues that have not been addressed in earlier research on intonation and presents both results and tools that contribute to the existing body of work on vocal intonation studies. The experiments showed that while overall there was a large amount of variability amongst the singers studied, both within singers' takes and across singers, there were some statistically significant consistencies, particularly for the vertical intervals but also for intervallic direction and musical context for some groups for the melodic intervals. The breadth of the experiment also provided information regarding the effect of training on the intonation practices of solo singers. In terms of tools, this dissertation presented a new algorithm for automatically estimating note onsets and offsets in monophonic recordings of the singing, using a hidden Markov model-based approach that was trained on the acoustics of the singing voice and bootstrapped with an existing Dynamic Time Warping score-audio alignment algorithm. This algorithm allowed for the intonation from the large number of recordings considered in this dissertation to be analyzed more efficiently than manual annotation. This dissertation also introduced a new approach to describing changes in the fundamental frequency over the duration of a note using the discrete cosine transform.

## 5.3 Future Research

### 5.3.1 More Controlled Experiments

The variation in the singers' intonation practices detailed in Chapter 4 may be due to unaccounted-for degrees of variation in the musical material, as well as the behaviour of the other singers in the ensemble experiment. Parts One and Two of the ensemble experiment were designed to minimize some of this variability. Similar exercises could be developed to highlight particular musical features for both solo and ensemble experiments. With Parts One and Two, however, there remained the issue concerning the behaviour of the different singers and the fact that the singers were essentially sight reading this material. One way to minimize the effect of other singers is to have them perform individually against recorded and pitch-corrected versions of the other parts. Since the other voices would be held steady across each rendition, it would be

possible to identify when a singer is simply varying across performances rather than reacting to what another singer is doing. Another benefit of this approach is that alternate versions of the accompanying recordings could be created in different tuning systems.

### 5.3.2 Perceptual Questions

There are a number of perceptual issues that arose during this dissertation for which educated assumptions were made. Given more time, it would be productive to explore these questions in more detail. The first question relates to the perception of pitch for sung notes with vibrato. The existing work on pitch perception in notes with vibrato has been performed with either synthetic or violin tones. One can assume that the results would be the same for sung notes, though this is an assumption that can be tested. Secondly, while existing work has looked at variable rate and depth of vibrato in synthetic tones, the question remains of how pitch is perceived when the average pitch of the vibrato changes over the duration of the note. This relates to the question of how slope and curvature in  $F_0$  are perceived and if they influence the perceived pitch.

Two additional perceptual issues arise when considering the issue of intonation explicitly. The first is the question of which tuning variations are perceived as “sharp,” which are perceived as “flat,” and which are perceived as changes in timbre. These perceptions have not been explored for the singing voice. The second is whether vowels have an inherent pitch. For example, if two notes have the same  $F_0$  trace but are sung with different vowels, are they perceived as being the same pitch?

### 5.3.3 Improving the DCT

In the solo experiments in Chapter 4, the use of the 1<sup>st</sup> and 2<sup>nd</sup> Discrete Cosine Transform has provided some interesting information about the ways in which the  $F_0$  changes at the end of the first note of the melodic intervals studied. However, there remain some open questions regarding how to best apply the DCT to the signal. There are the perceptual issues discussed in 5.3.2, as well as some implementation issues. The first issue relates to minimizing the impact of vibrato on the calculations, which was roughly done in Section 4.1 through the use of a 200 ms moving average. A more sophisticated approach that attempts to estimate the rate and depth of the vibrato for each cycle and cancel it out could be developed.

### **5.3.4 Other Ways of Analyzing the Music**

Theories of melodic attraction, particularly those put forth by Lerdahl (2001) and Larson (2004), offer alternative ways of quantifying the relationship between melodic tones that can supplement the approaches used in Chapter 4. Lerdahl's approach is a component of his Tonal Pitch Space theory; in this method, he formalizes the tendency of a dissonant pitch to resolve to a consonant neighbour (which may be a neighbour at either the chromatic, diatonic, or triad level of his pitch space model) with a rule which observes both Bharucha's principles of proximity and asymmetry (1996) and Newton's law of gravitation. Lerdahl also discusses the asymmetries in attraction when moving from unstable to stable pitches and vice-versa. These asymmetries demonstrate how the same interval functions differently in different musical contexts.

Larson posits a more complex calculation for the same phenomenon, though more explicitly focused on quantifying how listeners' expectations are either met or confounded by particular musical activities (Larson 2004). Larson's model correlates the forces of gravity (the tendency of a musical line to go down), magnetism (the tendency of unstable notes to move to stable ones), and inertia (the tendency of a musical line to continue rather than vary) explicitly in a single equation that is rooted in the gestalt psychological principles of proximity and stability. The cumulative forces acting on a note in a given context or pattern is calculated by summing the results of individual calculations for each force. Both Lerdahl's and Larson's theories have been subject to empirical perceptual tests (Vega 2003; Larson and Vanhandel 2005; Lerdahl and Krumhansl 2007), which produced results that generally support their theories, adding some strength to the notion that these theories could provide a cognitively sound way of quantifying the relative stability of notes in a musical passage.

### **5.3.5 Intonation and Expression**

In the *Psychology of Music*, C. E. Seashore argued that pitch is the “fundamental character of a tone” and that “it determines in large part what emotional reaction we shall have for this tone” (Seashore 1938). Seashore’s idea that emotion is conveyed in performance through deviations from the norm was later subsumed by Meyer’s theory of musical emotion (Meyer 1956), in which Meyer argued that emotional responses to music are rooted in the fulfillment or denial of the listener’s larger-scale expectations. More recently, Palmer (1996, 1997),

Gabrielsson (1999), and Sloboda (2005) have reconsidered the issue of musical emotion from the perspective of performance. Various studies of intonation practices, such as those carried out by Fyk (1995), have also discussed the expressive aspects of intonation. One future application of intonation studies is to provide quantitative data about the typical deviations in singing performance, which could be correlated with the results of psychological experiments on musical emotion.

### **5.3.6 Intonation in Non-Western and Popular Music**

This dissertation focused on singing intonation in Western art music. Issues regarding the intonation practices of non-Western and popular traditions must also be considered. A potential area of inquiry involves the early twentieth-century American folk recordings included in the Smithsonian Folkways collection. This collection is appealing for a number of reasons, including the large amount of monophonic recordings, the existence of multiple versions of songs by different singers, and the availability of transcriptions made by Alan Lomax and other ethnographers. In non-Western classical music, Byzantine chant, and North Indian and Middle Eastern vocal musics are potentially fruitful areas of inquiry for intonation studies.

### **5.3.7 Improving the Annotation Tool**

The alignment algorithm used in this dissertation was developed for use with recordings by singers who are following the score exactly, and it would have difficulty providing accurate annotations for performances that deviate from the score, either in error or by design. This response is due to the fact that the first-stage alignment relies on the sung pitches that closely correspond to the notated pitches in the reference score. An amateur rendition, for example, may include significant relative pitch errors, making this alignment unreliable. There are applications in which automatic alignment of such performances would be valuable, such as in an analysis of children's singing practices. I plan to apply the algorithm to such recordings by relying on contour and word sequence rather than exact pitch matches. This type of algorithm would be useful for developmental psychologists, including those working on the acquisition of song in development through the SSHRC-MCRI funded Advancing Interdisciplinary Research in Singing (AIRS) project.

### **5.3.8 Examination of Existing Recordings**

Another source of data for intonation studies is the wealth of existing recordings of singing performances. Once signal processing tools are developed for accurately analyzing pitch information in polyphonic recordings, it will be possible to do longitudinal studies of intonation practices, both across time and geographic distance. With the currently available tools, these types of studies can be undertaken for existing monophonic recordings.

### **5.3.9 Modeling Expressive Performance**

Once more data is collected, models of singers' intonation practices can be developed. As discussed in Section 2.2.4, the only existing expressive performance model that addresses intonation is the Director Musices system (Bresin et al. 2002), an analysis-by-synthesis model whose intonation rules have not shown a strong correspondence to intonation in practice (Ornoy 2008). Expressive performance models are useful for generating "natural" sounding digital re-creations, and also have potential pedagogical applications for training vocalists.

## References

- Abdallah, S. A., and M. D. Plumley. 2004. Polyphonic transcription by non-negative sparse coding of power spectra. In *Proceedings of the International Conference on Music Information Retrieval*, 318–25.
- Adams, N., M. Bartsch, J. Shifrin, and G. Wakefield. 2004. Time series alignment for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval*, 303–10.
- Adams, N., D. Marquez, and G. Wakefield. 2005. Iterative deepening for melody alignment and retrieval. In *Proceedings of the International Conference on Music Information Retrieval*, 199–206.
- Ambrazevičius, R., and I. Wiśniewska. 2008. Chromaticisms or performance rules? Evidence from traditional singing. *Journal of Interdisciplinary Music Studies* 2 (1–2): 19–31.
- Arifi, V., M. Clausen, F. Kurth, and M. Müller. 2003. Automatic synchronization of music data in score-, MIDI- and PCM-format. In *Proceedings of the International Conference on Music Information Retrieval*, 219–20.
- Arifi, V., M. Clausen, F. Kurth, and M. Müller. 2004. Automatic synchronization of musical data: A mathematical approach. *Computing in Musicology* 13: 9–33.
- Aristoxenus. c. 335 BCE. Elementa harmonica. In *Greek Musical Writings, Vol. 2. Harmonic and Acoustic Theory*, ed. A. Barker, 119–89. Cambridge, UK: Cambridge University Press. Original edition, 1989.
- Arroabarren, I., and A. Carolensa. 2004. Vibrato in singing voice: The link between source-filter and sinusoidal models. *EURASIP Journal on Applied Signal Processing* 7: 1007–20.
- Arroabarren, I., X. Rodet, and A. Carlosena. 2006. On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato. *IEEE Transactions on Audio, Speech, and Language Processing* 14 (4): 1413–21.

- Arroabarren, I., M. Zivanovic, J. Bretos, A. Ezcurra, and A. Carlosena. 2002a. Measurement of vibrato in lyric singers. *IEEE Transactions on Instrumentation and Measurement* 51 (4): 660–5.
- Arroabarren, I., M. Zivanovic, and A. Carlosena. 2002b. Analysis and synthesis of vibrato in lyric singers. In *Proceedings of the European Signal Processing Conference*.
- Arrouburren, I., M. Zivunovic, X. Rodet, and A. Curlosena. 2003. Instantaneous frequency and amplitude of vibrato in singing voice. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 537–40.
- Atal, B. S., and S. L. Hanauer. 1971. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America* 47 (65): 637–55.
- Backus, J. 1977. *The Acoustical Foundations of Music*. 2nd ed. New York, NY: W. W. Norton & Company. Original edition, 1969.
- Baird, B., D. Blevins, and N. Zahler. 1990. The artificially intelligent computer performer: The second generation. *Interface / Journal of New Music Research* 19 (2): 197–204.
- Baird, B., D. Blevins, and N. Zahler. 1993. Artificial intelligence and music: Implementing an interactive computer performer. *Computer Music Journal* 17 (2): 73–9.
- Balaguer-Ballester, E., N. R. Clark, M. Coath, K. Krumbholz, and S. L. Denham. 2010. Understanding pitch perception as a hierarchical process with top-down modulation. In *The Neurophysiological Bases of Auditory Perception*, ed. E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis. New York, NY: Springer Science+Business Media.
- Balaguer-Ballester, E., S. L. Denham, and R. Meddis. 2008. A cascade autocorrelation model of pitch perception. *Journal of the Acoustical Society of America* 124 (4): 2186–95.
- Barbera, A. 1984. The consonant eleventh and the expansion of the musical tetractys: A study in ancient Pythagoreanism. *Journal of Music Theory* 28: 191–223.
- 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.

- Beauchamp, J. W. 1974. Time-variant spectra of violin tones. *Journal of the Acoustical Society of America* 56 (3): 995–1004.
- Bell, M. S. 1994. *The Effects of Rate of Deviation Musical Context on Intonation Perception in Homophonic Four-Part Chorales*. PhD diss., School of Music, Florida State University.
- Bello, J. P., L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler. 2005. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio* 13 (5): 1035–47.
- Benedetti, G. 1585. *Diversarum speculationum liber*. Turin: Sucessors of Nicola Bevilqua.
- Bengtsson, I., and A. Gabrielsson. 1980. Methods for analyzing performance of musical rhythm. *Scandinavian Journal of Psychology* 21: 257–68.
- Bengtsson, I., and A. Gabrielsson. 1983. Analysis and synthesis of musical rhythm. *Studies of Musical Performance* 39: 27–60.
- Berenzweig, A. L., and D. P. W. Ellis. 2001. Locating singing voice segments within musical signals. In *Proceedings of the Workshop on the Applications of Signal Processing to Audio and Acoustics*, 119–22.
- Bernstein, J. G., and A. J. Oxenham. 2008. Harmonic segregation through mistuning can improve fundamental frequency discrimination. *Journal of the Acoustical Society of America* 124: 1653–67.
- Bernstein, J. G. W., and A. J. Oxenham. 2003. Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? *Journal of the Acoustical Society of America* 113 (6): 3323–34.
- Bernstein, J. G. W., and A. J. Oxenham. 2005. An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination. *Journal of the Acoustical Society of America* 117 (6): 3816–31.
- Bharucha, J. J. 1996. Melodic anchoring. *Music Perception* 13 (3): 383–400.
- Bharucha, J. J. 2009. From frequency to pitch, and from pitch class to musical key: Shared principles of learning and perception. *Connection Science* 21 (2): 177–92.

- Bilsen, F. A. 1977. Pitch of noise signals: Evidence for a “central spectrum.” *Journal of the Acoustical Society of America* 61 (1): 150–61.
- Birmingham, W. P., R. Dannenberg, G. Wakefield, M. Bartisch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand. 2001. MUSART: Music retrieval via aural queries. In *Proceedings of the International Conference on Music Information Retrieval*, 73–81.
- Bjorklund, A. 1961. Analysis of soprano voices. *Journal of the Acoustical Society of America* 33 (5): 575–82.
- Bloch, J. J., and R. Dannenberg. 1985. Real-time computer accompaniment of keyboard performances. In *Proceedings of the International Computer Music Conference*, 279–89.
- Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, 97–110.
- Boersma, P. 2001. PRAAT, a system for doing phonetics by computer. *Glot International* 5 (9/10): 341–5.
- Boethius. c. 520 CE. Fundamentals of music. In *Strunk’s Source Reading in Music History*, ed. O. Strunk and L. Treitler, 137–42. New York, NY: W. W. Norton & Company.
- Boomsliter, P., and W. Creel. 1961. The long pattern hypothesis in harmony and hearing. *Journal of Music Theory* 5 (2): 2–30.
- Boomsliter, P., and W. Creel. 1963. An unrecognized dynamic in melody. *Journal of Music Theory* 7 (1): 2–22.
- Bower, C. M. 2002. The transmission of ancient music theory into the Middle Ages. In *The Cambridge History of Western Music Theory*, ed. T. Christensen, 136–67. Cambridge, UK: Cambridge University Press.
- Bresin, R., A. Friberg, and J. Sundberg. 2002. Director Musices: The KTH performance rules system. In *Proceedings of the SIGMUS*, 43–8.
- Bretos, J., and J. Sundberg. 2003. Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos. *Journal of Voice* 17 (3): 343–52.

- 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.
- Burns, E. M. 1999. Intervals, scales, and tuning. In *The Psychology of Music*, ed. D. Deutsch, 215–64. San Diego, CA: Academic Press.
- Burns, E. M., and W. D. Ward. 1978. Categorical perception-phenomenon or epiphenomenon. *Journal of the Acoustical Society of America* 63: 456–68.
- Cano, P., A. Loscos, and J. Bonada. 1999. Score-performance matching using HMMs. In *Proceedings of the International Computer Music Conference*, 441–4.
- Cano, P., A. Loscos, J. Bonada, M. de Boer, and X. Serra. 2000. Voice morphing system for impersonating in karaoke applications. In *Proceedings of the International Computer Music Conference*, 109–12.
- Carlsson-Berndtsson, G., and J. Sundberg. 1991. Formant frequency tuning in singing. *STL-Quarterly Progress and Status Report* 32 (1): 29–35.
- Cemgil, T., P. Desain, and B. Kappen. 2000. Rhythm quantization for transcription. *Computer Music Journal* 24 (2): 60–76.
- Chen, J., M. H. Woollacott, S. Pologe, and G. P. Moore. 2008. Pitch and space maps of skilled cellists: accuracy, variability, and error correction. *Experimental Brain Research* 188: 493–503.
- Christensen, T. 1993. *Rameau and Musical Thought in the Enlightenment*. Cambridge, UK: Cambridge University Press.
- Christianini, N., and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, UK: Cambridge University Press.
- Clarisso, L., J. Martens, M. Lesaffre, B. Baets, H. Meyer, and M. Leman. 2002. An auditory model based transcriber of singing sequences. In *Proceedings of the International Conference on Music Information Retrieval*, 116–23.
- Clarke, E. 1989. The perception of expressive timing in music. *Psychological Research* 51: 2–9.
- Clarke, E. F., and W. L. Windsor. 2000. Real and simulated expression: A listening study. *Music Perception* 17: 277–313.

- Cohen, J., P. Cohen, S. G. West, and L. S. Aiken. 2002. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. New York, NY: Routledge Academic.
- Cont, A. 2006. Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 185–8.
- Cont, A. 2010. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (6): 974–87.
- Cont, A., S. Dubnov, and D. Wessel. 2007a. Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. In *Proceedings of the International Conference on Digital Audio Effects*, 85–92.
- Cont, A., D. Schwarz, and N. Schnell. 2005. Training IRCAM's score follower. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing*, 253–6.
- Cont, A., D. Schwarz, N. Schnell, and C. Raphael. 2007b. Evaluation of real-time audio-to-score alignment. In *Proceedings of the International Conference on Music Information Retrieval*, 315–6.
- Cook, N. D. 2009. Harmony perception: Harmoniousness is more than the sum of interval consonance. *Music Perception* 27 (1): 25–41.
- Cook, N. D., T. Fujisawa, and K. Takami. 2004. A psychophysical model of harmony perception. In *Proceedings of the International Conference on Music Perception and Cognition*.
- Cook, P. R. 1993. SPASM, a real-time vocal tract physical model controller; and Singer, the companion software synthesis system. *Computer Music Journal* 17 (1): 30–44.
- Cook, P. R. 1996. Singing voice synthesis: History, current work, and future directions. *Computer Music Journal* 20 (3): 38–46.
- Cooper, D., and I. Sapiro. 2006. Ethnomusicology in the laboratory: From the Tonometer to the Digital Melograph. *Ethnomusicology Forum* 15 (2): 301–13.
- d'Alessandro, C., and M. Castellengo. 1991. Etude, par la synthese de la perception du vibrato vocal dans les transitions de notes. *International Voice Conference*.

- d'Alessandro, C., and M. Castellengo. 1994. The pitch of short-duration vibrato tones. *Journal of the Acoustical Society of America* 95 (3): 1617–30.
- d'Alessandro, C., and M. Castellengo. 1995. The pitch of short-duration vibrato tones: Experimental data and numerical model. In *Vibrato*, ed. P. H. Dejonckere, M. Hirano, and J. Sundberg, 83–92. San Diego, CA: Singular Publishing Group.
- d'Alessandro, N., O. Babacan, B. Bozkurt, T. Dubuisson, A. Holzapfel, L. Kessous, A. Moinet, and M. Vlieghe. 2008. RAMCESS 2.X framework—expressive voice analysis for realtime and accurate synthesis of singing. *Journal of Multimodal User Interfaces* 2: 133–44.
- Dai, H. 2010. Harmonic pitch: Dependence on resolved partials, spectral edges, and combination tones. *Hearing Research* 270 (1–2): 143–50.
- Dannenberg, R. 1984. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference*, 193–8.
- Dannenberg, R. 2003. *Audio Alignment Software in Matlab*. Available from <http://www.cs.cmu.edu/~music/music.software.html> (accessed 27 February 2010).
- Dannenberg, R. 2007. An intelligent multi-track audio editor. In *Proceedings of the International Computer Music Conference*, 89–94.
- Dannenberg, R., and N. Hu. 2003. Polyphonic audio matching for score following and intelligent audio editors. In *Proceedings of the International Computer Music Conference*, 27–34.
- Dannenberg, R., and H. Mukaino. 1988. New techniques for enhanced quality of computer accompaniment. In *Proceedings of the International Computer Music Conference*, 243–9.
- Dannenberg, R., and C. Raphael. 2006. Music score alignment and computer accompaniment. . *Communications of the ACM* 49 (8): 38–43.
- Dannenberg, R., M. Sanchez, A. Joseph, R. Joseph, R. Saul, and P. Capell. 1993. Results from the piano tutor project. In *Proceedings of the Fourth Biennial Arts and Technology Symposium*, 143–50.

- Dannenberg, R. B., W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis. 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 (3): 687–701.
- Davy, M. 2006. Multiple fundamental frequency estimation based on generative models. In *Signal Processing Methods for Music Transcription*, ed. A. Klapuri and M. Davy, 203–27. New York, NY: Springer.
- de Cheveigné, A. 1998. Cancellation model of pitch perception. *Journal of the Acoustical Society of America* 103 (3): 1261–71.
- de Cheveigné, A. 2002. YIN MATLAB implementation Available from <http://audition.ens.fr/adc/sw/yin.zip> (accessed 30 September 2010).
- de Cheveigné, A. 2005. Pitch perception models. In *Pitch: Neural Coding and Perception*, ed. C. J. Plack, A. J. Oxenham, R. R. Fay, and N. A. Popper, 169–233. New York, NY: Springer.
- de Cheveigné, A. 2006. Multiple F0 estimation. In *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, ed. D. Wang and G. J. Brown, 45–80. Hoboken, NJ: Wiley-IEEE Press.
- de Cheveigné, A., and A. Baskind. 2003. F0 estimation of one or several voices. In *Proceedings of the European Conference on Speech Communication and Technology*, 833–6.
- de Cheveigné, A., and N. Henrich. 2002. Fundamental frequency estimation of musical sounds. *Journal of the Acoustical Society of America* 111: 2416.
- de Cheveigné, A., and H. Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America* 111 (4): 1917–30.
- De Poli, G. 2004. Methodologies for expressiveness modelling of and for music performance. *Journal of New Music Research* 33 (3): 189–202.
- Desain, P., and H. Honing. 1992. *Music, Mind, and Machine: Studies in Computer Music, Music Cognition, and Artificial Intelligence*. Amsterdam, NL: Thesis Publishers.

- Desain, P., H. Honing, and H. Heijink. 1997. Robust score-performance matching: Taking advantage of structural information. In *Proceedings of the International Computer Music Conference*, 337–40.
- Devaney, J. 2006. A methodology for the study and modeling of choral intonation practices. In *Proceedings of the International Conference on Music Perception and Cognition*.
- Devaney, J. 2008a. The potential of recordings in testing quantitative aspects of music theories. Paper presented at the *Canadian University Music Society*.
- Devaney, J. 2008b. “Tonality's gravitational pull”: Intonation as an empirical measure of melodic attraction. Paper presented at the *Society of Music Theory*.
- Devaney, J. 2009. Intonation tendencies in solo a cappella performances. Paper presented at the *Indiana University Symposium of Research in Music Theory: Special Symposium on Performance and Analysis*.
- Devaney, J., and D. P. W. Ellis. 2007. An empirical approach to studying intonation tendencies in choral performances. In *Proceedings of the Conference on Interdisciplinary Musicology*.
- Devaney, J., and D. P. W. Ellis. 2008. An empirical approach to studying intonation tendencies in polyphonic vocal performances. *Journal of Interdisciplinary Music Studies* 2 (1–2): 141–56.
- Devaney, J., and D. P. W. Ellis. 2009. Handing asynchrony in audio-score alignment. In *Proceedings of the International Computer Music Conference*, 29–32.
- Devaney, J., and I. Fujinaga. 2010. AMPACT: Automatic Music Performance Analysis Toolkit. Poster presented at the *Society of Music Theory*.
- Devaney, J., I. Fujinaga, and D. P. W. Ellis. 2008. Intonation tendencies in polyphonic vocal ensembles. Paper presented at the *Digital Music Research Network Workshop*.
- Devaney, J., M. Mandel, I., D. P. W. Ellis, and I. Fujinaga. Forthcoming. Automatically extracting performance data from recordings of trained singers. *Pyschomusicology*.
- Devaney, J., M. I. Mandel, and D. P. W. Ellis. 2009a. Improving MIDI-audio alignment with acoustic features. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, 45–8.

- Devaney, J., and J. Wild. 2009. Empirical, historical and speculative approaches to intonation. Paper presented at the *Physiology and Acoustics of Singing*.
- Devaney, J., J. Wild, and I. Fujinaga. 2009b. Intonation tendencies in solo a cappella performances. Poster presented at the *Society for Music Perception and Cognition*.
- Devaney, J., J. Wild, P. Schubert, and I. Fujinaga. 2010a. Exploring the relationship between voice leading, harmony, and intonation in a cappella SATB vocal ensembles. Paper presented at the *International Conference on Music Perception and Cognition*.
- Devaney, J., J. Wild, P. Schubert, and I. Fujinaga. 2010b. Horizontal and vertical intonation tendencies in SATB ensembles. Paper presented at the *Physiology and Acoustics of Singing*.
- Devaney, J., J. Wild, P. Schubert, and I. Fujinaga. 2010c. What can expressive performance studies tell us about the organization of musical materials? Paper presented at the *Indiana University Symposium of Research in Music Theory: "This is your brain on music."*
- Dixon, S. 2001. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research* 30 (1): 39–58.
- Dixon, S. 2003. Towards automatic analysis of expressive performance. In *Proceedings of the European Society for the Cognition Sciences for Music Conference*, 107–10.
- Dixon, S. 2005. Live tracking of musical performances using on-line time warping. In *Proceedings of the International Conference on Digital Audio Effects*, 92–7.
- Dixon, S., and G. Widmer. 2005. MATCH: A music alignment toolkit. In *Proceedings of the International Conference on Music Information Retrieval*, 492–7.
- Downie, J. S. 2006. Score following. *MIREX 2006*, [http://www.music-ir.org/mirex/2006/index.php/Score\\_Following\\_Proposal](http://www.music-ir.org/mirex/2006/index.php/Score_Following_Proposal).
- Downie, J. S. 2007. *MIREX audio onset detection*. Available from [http://www.music-ir.org/mirex/wiki/2007:Audio\\_Onset\\_Detection](http://www.music-ir.org/mirex/wiki/2007:Audio_Onset_Detection) (accessed 3 November 2009).
- Downie, J. S. 2008. Real-time audio to score alignment. *MIREX 2008*, [http://www.music-ir.org/mirex/2008/index.php/Real-time\\_Audio\\_to\\_Score\\_Alignment\\_\(a.k.a\\_Score\\_Following\)](http://www.music-ir.org/mirex/2008/index.php/Real-time_Audio_to_Score_Alignment_(a.k.a_Score_Following)).

- Duifhuis, H., L. F. Willems, and R. J. Sluyter. 1982. Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *Journal of the Acoustical Society of America* 71 (6): 1568–80.
- Dunn, J. W., D. Byrd, M. Notess, J. Riley, and R. Scherle. 2006. Variations2: Retrieving and using music in an academic setting. *Communications of the ACM* 49 (8): 53–8.
- Earis, A. 2007. An algorithm to extract expressive timing and dynamics from piano recordings. *Musicae Scientiae* 11 (2): 155–82.
- Easley, E. 1932. A comparison of the vibrato in concert and opera singing. In *University of Iowa Studies in the Psychology of Music. Vol. I: The Vibrato*, ed. C. Seashore, 269–75. Iowa City, IA: University of Iowa.
- Ellis, D. P. W. 2003. *Dynamic Time Warp (DTW) in Matlab*. Available from <http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/> (accessed 4 September 2010).
- Ellis, D. P. W. 2008. *Aligning MIDI scores to music audio*. Available from <http://www.ee.columbia.edu/~dpwe/resources/matlab/alignmidiwav/> (accessed 4 September 2010).
- Ely, M. C. 1992. Effects of timbre on college woodwind players' intonational performance and perception. *Journal of Research in Music Education* 40 (2): 158–67.
- Fletcher, H. 1940. Auditory patterns. *Reviews of Modern Physics* 12 (1): 47–65.
- Fletcher, H., and L. C. Sanders. 1967. Quality of violin vibrato tones. *Journal of the Acoustical Society of America* 41 (6): 1534–44.
- Flossmann, S., M. Grachten, and G. Widmer. 2008. Experimentally investigating the use of score features for computational models of expressive timing. In *Proceedings of the International Conference on Music Perception and Cognition*.
- Flossmann, S., M. Grachten, and G. Widmer. 2009. Expressive performance rendering: Introducing performance context. In *Proceedings of the Sound and Music Computing Conference*, 155–60.
- Fourier, J. B. J. 1820. *Traité analytique de la chaleur*. Paris, FR: Didot.

- Fremerey, C., M. Müller, F. Kurth, and M. Clausen. 2008. Automatic mapping of scanned sheet music to audio recordings. In *Proceedings of the International Conference on Music Information Retrieval*, 413–8.
- Friberg, A. 1995. Matching the rule parameters of Phrase Arch to performances of “Träumerei:” A preliminary study. In *Proceedings of the KTH Symposium on Grammars for Music Performance*.
- Friberg, A., R. Bresin, and J. Sundberg. 2006. Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology* 2 (2–3): 145–61.
- Friberg, A., E. Schoonderwaldt, and P. N. Juslin. 2007. CUEX: An algorithm for extracting expressive tone variables from audio recordings. *Acta Acustica united with Acustica* 93: 411–20.
- Fyk, J. 1995. *Melodic Intonation: Psychoacoustics, and the Violin*. Zielona Góra, PL: Organon.
- Gabrielsson, A. 1999. The performance of music. In *The Psychology of Music*, ed. D. Deutsch, 501–602. Original edition, San Diego, CA: Academic Press.
- Gabrielsson, A. 2003. Music performance research at the millennium. *Psychology of Music* 31 (3): 221–72.
- Gabrielsson, A., and P. N. Juslin. 1996. Emotional expression in music performance: Between the performer’s intention and the listeners experience. *Psychology of Music* 24 (1): 68–91.
- Ghias, A., J. Logan, D. Chamberlin, and B. C. Smith. 1995. Query-by-Humming: Musical information retrieval in an audio database. In *Proceedings of the ACM Digital Libraries*, 231–6.
- Glarean, H. 1547. Dodecachordon. In *Strunk’s Source Reading in Music History*, ed. O. Strunk and L. Treitler, 429–35. New York, NY: W. W. Norton & Company.
- Glasberg, B. R., and B. C. J. Moore. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47: 103–38.
- Gockel, H., B. C. J. Moore, and R. P. Carlyon. 2001. Influence of rate of change of frequency on the overall pitch of frequency-modulated tones. *Journal of the Acoustical Society of America* 109 (2): 701–12.

- Goebel, W., S. Dixon, G. De Poli, A. Friberg, R. Bresin, and G. Widmer. 2008. Sense in expressive music performance: Data acquisition, computational studies, and models. In *Sound to Sense – Sense to Sound: A State of the Art in Sound and Music Computing*, ed. P. Polotti and D. Rocchesso, 195–242. Berlin, DE: Logos Verlag.
- Goldstein, J. L. 1973. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America* 54 (6): 1496–516.
- Gouyon, F., and S. Dixon. 2005. A review of automatic rhythm description systems. *Computer Music Journal* 29 (1): 34–54.
- Gramming, P., J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins. 1987. Relationship between changes in voice pitch and loudness. *STL-Quarterly Progress and Status Report* 28 (1): 39–55.
- Granqvist, S., and B. Hammarberg. 2003. The correlogram: A visual display of periodicity. *Journal of the Acoustical Society of America* 114 (5): 2934–45.
- Green, P. C. 1937. Violin performance with reference to tempered, natural, and Pythagorean intonation. In *University of Iowa Studies in the Psychology of Music. Vol. IV. Objective Analysis of Musical Performance*, ed. C. Seashore, 232–51. Iowa City, IA: University of Iowa.
- Grindlay, G., and D. Helmbold. 2006. Modeling, analyzing, and synthesizing expressive piano performance with graphical models. *Machine Learning* 65: 361–87.
- Grubb, L., and R. Dannenberg. 1994. Automating ensemble performance. In *Proceedings of the International Computer Music Conference*, 63–9.
- Grubb, L., and R. Dannenberg. 1997. A stochastic method of tracking a vocal performer. In *Proceedings of the International Computer Music Conference*, 301–8.
- Grubb, L., and R. Dannenberg. 1998. Enhanced vocal performance tracking using multiple information sources. In *Proceedings of the International Computer Music Conference*, 37–44.
- Hagerman, B., and J. Sundberg. 1980. Fundamental frequency adjustment in barbershop singing. *Journal of Research in Singing* 4: 3–17.
- Hattwick, M. 1932. The vibrato in wind instruments. In *University of Iowa Studies in the Psychology of Music. Vol. I: The Vibrato*, 276–80. Iowa City, IA: University of Iowa.

*Hearing: Basilar Membrane.* 1997. Available from <http://www.britannica.com/EBchecked/topic-art/123552/18100/Model-showing-the-distribution-of-frequencies-along-the-basilar-membrane> (accessed 4 October 2010).

Heijink, H., P. Desain, H. Honing, and L. Windsor. 2000a. Make me a match: An evaluation of different approaches to score-performance matching. *Computer Music Journal* 24 (1): 43–56.

Heijink, H., L. Windsor, and P. Desain. 2000b. Data processing in music performance research: Using structural information to improve score-performance matching. *Behavior Research Methods* 32 (4): 546–54.

*Helix: Structures of the Human Ear.* 1997. Available from <http://www.britannica.com/EBchecked/topic-art/123552/111239/The-structures-of-the-outer-middle-and-inner-ear> (accessed 4 October 2010).

Helmholtz, H. 1863. *On the Sensation of Tone as a Physiological Basis for the Theory of Music.* Translated by A. J. Ellis. New York, NY: Dover Publications.

Herrera, P., and J. Bonada. 1998. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proceedings of the Digital Audio Effects Workshop (DAFX98)*.

Hirano, M., S. Hibi, and S. Hagino. 1995. Physiological aspects of vibrato. In *Vibrato*, ed. P. H. Dejonckere, M. Hirano, and J. Sundberg, 9–33. San Diego, CA: Singular Publishing Group.

Hirose, K. 1934. An experimental study on the principal pitch in the vibrato. *Japanese Journal of Psychology* 9: 793–845.

Hofmann-Engl, L. 2010. Consonance/dissonance: A historical perspective. In *Proceedings of the International Conference on Music Perception and Cognition*, 852–6.

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

Honing, H. 1990. POCO: An environment for analysing, modifying, and generating expression in music. In *Proceedings of the International Computer Music Conference*, 364–8.

- Hoshishiba, T., S. Horiguchi, and I. Fujinaga. 1996. Study of expression and individuality in music performance using normative data derived from MIDI recordings of piano music. In *Proceedings of the International Conference on Music Perception and Cognition*, 465–70.
- 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 (3): 300–15.
- Howes, P., J. Callaghan, P. Davis, D. Kenny, and W. Thorpe. 2004. The Relationship Between Measured Vibrato Characteristics and Perception in Western Operatic Singing. *Journal of Voice* 18 (2): 216–30.
- Hu, N., and R. Dannenberg. 2006. Bootstrap learning for accurate onset detection. *Machine Learning* 65 (2): 457–71.
- Hu, N., R. Dannenberg, and G. Tzanetakis. 2003. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, 185–8.
- Huron, D. 2001. Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception* 19 (1): 1–64.
- Hutchinson, W., and L. Knopoff. 1978. The acoustic component of western consonance. *Journal of New Music Research* 7: 1–29.
- Iwamiya, S.-I., and K. Fujiwara. 1985. Perceived principal pitch of FM-AM tones as a function of the phase difference between frequency modulation and amplitude modulation. *Journal of the Acoustical Society of Japan* 6 (3): 193–202.
- Iwamiya, S.-I., K. Kosygi, and O. Kitamura. 1983. Perceived pitch of vibrato tones. *Journal of the Acoustical Society of Japan* 4: 73–82.
- Iwamiya, S.-I., T. Miyakura, N. Satoh, and Y. Hayashi. 1994. Perceived pitch of complex FM-AM tones: Pitch determination process of vibrato sounds. *Annals of Physiological Anthropology* 13 (5): 303–8.
- Izmirli, Ö., R. Seward, and N. Zahler. 2003. Melodic pattern anchoring for score following using score analysis. In *Proceedings of the International Computer Music Conference*, 411–4.

- Jackson, R. 2005. *Performance Practice*. New York, NY: Taylor & Francis Group.
- Jain, A. K. 1989. *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall.
- Jers, H., and S. Ternström. 2005. Intonation analysis of a multi-channel choir recording. *TMH-Quarterly Progress and Status Report* 47 (1): 1–6.
- Johnston, B. 2006. “*Maximum Clarity*” and Other Writings on Music. Edited by B. Gilmore. Urbana, IL: University of Illinois Press.
- Jordanous, A., and A. Smaill. 2009. Investigating the role of score following in automatic musical accompaniment. *Journal of New Music Research* 38 (2): 197–203.
- Jorgensen, O. H. 1991. *Tuning, Containing The Perfection of Eighteenth Century Temperament, The Lost Art of Nineteenth Century Temperament, and The Science of Equal Temperament*. East Lansing, MI: Michigan State University Press.
- Juslin, P. N., A. Friberg, and R. Bresin. 2002. Toward a computational model of expression in performance: The GERM model. *Musicae Scientiae*: 63–122.
- Kameoka, A., and M. Kuriyagawa. 1969a. Consonance theory part I: Consonance of dyads. *Journal of the Acoustical Society of America* 45 (6): 1451–9.
- Kameoka, A., and M. Kuriyagawa. 1969b. Consonance theory part II: Consonance of complex tones and Its calculation method. *Journal of the Acoustical Society of America* 45 (6): 1460–9.
- Katayose, H., T. Kanamori, K. Kamei, Y. Nagashima, K. Sato, S. Inokuchi, and S. Simura. 1993. Virtual performer. In *Proceedings of the International Computer Music Conference*, 138–45.
- Keshet, J., S. Shalev-Shwartz, Y. Singer, and D. Chazan. 2007. A large margin algorithm for speech-to-phoneme and music-to-score alignment. *IEEE Transactions on Audio, Speech and Language Processing* 15 (8): 2373–82.
- Kim, Y. 2003. *Singing Voice Analysis/Synthesis*. PhD diss., School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge, MA.

- Kim, Y. E. 2009. Singing voice analysis, synthesis, and modeling. In *Handbook of Signal Processing in Acoustics*, ed. D. Havelock, S. Kuwano and M. Vorlander, 359–74. New York, NY: Springer Science+Business Media, LLC.
- Klapuri, A. 2006. Auditory model-based methods for multiple fundamental frequency estimation. In *Signal Processing Methods for Music Transcription*, ed. A. Klapuri and M. Davy, 229–65. New York, NY: Springer.
- Kob, M. 2002. *Physical Modeling of the Singing Voice*, RWTH Aachen Hochschulbibliothek.
- Kosugi, N., Y. Nishihara, T. Sakata, M. Yamamuro, and K. Kushima. 2000. A practical Query-by-Humming system for large a music database. In *Proceedings of the 8th ACM International Conference on Multimedia*, 333–42.
- Kurth, F., M. Müller, D. Damm, C. Fremerey, A. Ribbrock, and M. Clausen. 2005. Syncplayer: An advanced system for multimodal music access. In *Proceedings of the International Conference of Music Information Retrieval*, 381–8.
- Kurth, F., M. Müller, C. Fremerey, Y. Chang, and M. Clausen. 2007. Automated synchronization of scanned sheet music with audio recordings. In *Proceedings of the International Conference on Music Information Retrieval*, 261–6.
- Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models*. 5th ed. Boston, MA: McGraw-Hill.
- Large, E. W. 1993. Dynamic programming for the analysis of serial behaviors. *Behavior Research Methods, Instruments, and Computers* 25 (2): 238–41.
- Larson, S. 2004. Musical forces and melodic expectations: Comparing computer models with experimental results. *Music Perception* 21 (4): 457–98.
- Larson, S., and L. VanHandel. 2005. Measuring musical forces. *Music Perception* 23 (2): 119–36.
- Lerdahl, F. 2001. *Tonal Pitch Space*. Oxford, UK: Oxford University Press.
- Lerdahl, F., and R. Jackendoff. 1983. *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.
- Lerdahl, F., and C. Krumhansl. 2007. Modeling tonal tension. *Music Perception* 24 (4): 329–66.

- Licklider, J. C. R. 1951. A duplex theory of pitch perception. *Cellular and Molecular Life Sciences* 7 (4): 128–34.
- Licklider, J. C. R. 1954. “Periodicity” pitch and “place” pitch. *Journal of the Acoustical Society of America* 26 (5): 945.
- Lindblom, B., and J. Sundberg. 1970. Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America* 48 (1): 120.
- Logan, B. 2000. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval*, 23–5.
- Loosen, F. 1993. Intonation of solo violin performance with reference to equally tempered, Pythagorean, and just intonations. *Journal of the Acoustical Society of America* 93 (1): 525–39.
- Loosen, F. 1994. Tuning of diatonic scales by violinists, pianists, and nonmusicians. *Perception & Psychoacoustics* 56 (2): 221–6.
- Loosen, F. 1995. The effect of musical experience on the conception of accurate tuning. *Music Perception* 12 (3): 291–306.
- Loscos, A., P. Cano, and J. Bonda. 1999. Low-delay singing voice alignment to text. In *Proceedings of the International Computer Music Conference*, 437–40.
- Lots, I. S., and L. Stone. 2008. Perception of musical consonance and dissonance: an outcome of neural synchronization. *Journal of the Royal Society: Interface* 5: 1429–34.
- Maher, R., and J. W. Beauchamp. 1990. An investigation of vocal vibrato synthesis. *Applied Acoustics* 30: 219–45.
- Malmberg, C. F. 1918. The perception of consonance and dissonance. *Psychological Monographs* 25: 93–133.
- Marinescu, M.-C., and R. Ramirez. 2008. Expressive performance in the human tenor voice. In *Proceedings of the Sound and Music Computing Conference*.
- Marmel, F., B. Tillman, and W. J. Dowling. 2008. Tonal expectations influence pitch perception. *Perception & Psychoacoustics* 70 (5): 841–52.

- Marolt, M. 2004. Networks of adaptive oscillators for partial tracking and transcription of music recordings. *Journal of New Music Research* 33 (1): 49–59.
- Martin, P. 1982. Comparison of pitch detection by cepstrum and spectral comb analysis. In *Proceedings of the IEEE, International Conference on Acoustics, Speech and Signal Processing*, 180–3.
- Mason, J. A. 1960. Comparison of solo and ensemble performances with reference to Pythagorean, Just, and Equal-Tempered intonations. *Journal of Research in Music Education* 8 (1): 31–8.
- Mayor, O., J. Bonada, and A. Loscos. 2006. The Singing Tutor: Expression categorization and segmentation of the singing voice. In *Proceedings of the Audio Engineering Society Convention*.
- Mazzoni, D., and R. Dannenberg. 2000. *Audacity*. Available from <http://audacity.sourceforge.net/> (accessed June 10 2010).
- McAulay, R., and T. Quatieri. 1976. Speech snalysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics Speech and Signal Processing* 34 (44): 744–54.
- McDermott, J. H., A. J. Lehr, and A. J. Oxenham. 2010. Individual differences reveal the basis of consonance. *Current Biology* 20: 1035–41.
- McDermott, J. H., and A. J. Oxenham. 2008. Music perception, pitch, and the auditory system. *Current Opinion in Neurobiology* 18: 452–63.
- McNab, R. J., L. A. Smith, I. H. Witten, C. L. Henderson, and S. J. Cunningham. 1996. Towards the digital music library: Tune retrieval from acoustic input. In *Proceedings of the ACM International Conference on Digital Libraries*, 11–8.
- Meddis, R., and M. J. Hewitt. 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America* 89 (6): 2866–82.
- Meddis, R., and L. O'Mard. 1997. A unitary model of pitch perception. *Journal of the Acoustical Society of America* 102 (3): 1181–20.

- Melucci, M., and N. Orio. 1999. Musical information retrieval using melodic surface. In *Proceedings of the ACM Digital Libraries Conference*, 152–60.
- Mersenne, M. 1636. *Harmonie Universelle*. Translated by R. E. Chapman. The Hague, NL: Martinus Nijhoff.
- Mesz, B. A., and M. C. Eguia. 2009. The pitch of vibrato tones: A model based on instantaneous frequency decomposition. In *The Neurosciences and Music III—Disorders and Plasticity*, ed. S. Dalla Balla, N. Kraus, K. Overy, C. Pantev, J. S. Snyder, M. Tervaniemi, B. Tillman, and G. Schlaug, 126–30. Boston, MA: Blackwell Publishing.
- Metfessel, M. 1932. The vibrato in artistic voices. In *University of Iowa Studies in the Psychology of Music. Vol. I: The Vibrato*, ed. C. Seashore, 14–117. Iowa City, IA: University of Iowa.
- Miller, D. C. 1916. *The Science of Musical Sounds*. New York, NY: Macmillan Press.
- Miller, R. S. 1936. The pitch of the attack in singing. In *University of Iowa Studies in the Psychology of Music. Vol. IV. Objective Analysis of Musical Performance*, ed. C. Seashore, 158–71. Iowa City, IA: University of Iowa.
- Miotto, R., and N. Orio. 2007. Automatic identification of music works through audio matching. In *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries*, 124–35.
- Montecchio, N., and N. Orio. 2008. Automatic alignment of music performances with scores aimed at educational applications. In *Proceedings of the Automated solutions for Cross Media Content and Multi-channel Distribution*, 17–24.
- Montecchio, N., and N. Orio. 2009. A discrete filter bank approach to audio to score matching for polyphonic music. In *Proceedings of the International Conference on Music Information Retrieval*, 495–500.
- Moore, B. C. J. 2008. The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology* 9: 399–406.
- Moore, B. C. J., and B. R. Glasberg. 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America* 74 (3): 751–3.

- Moore, B. C. J., and B. R. Glasberg. 2010. The role of temporal fine structure in harmonic segregation through mistuning. *Journal of the Acoustical Society of America* 127 (1): 5–8.
- Morrison, S. J., and J. Fyk. 2002. Intonation. In *The Science and Psychology of Music Performance Creative Strategies for Teaching and Learning*, ed. R. Parncutt and G. Mcpherson, 183–98. Oxford, UK: Oxford University Press.
- Müller, M., and S. Ewert. 2008. Joint structure analysis with applications to music annotation and synchronization. In *Proceedings of the International Conference on Music Information Retrieval*, 389–94.
- Müller, M., F. Kurth, and M. Clausen. 2005. Audio matching via chroma-based statistical features. In *Proceedings of the International Conference on Music Information Retrieval*, 288–95.
- Müller, M., F. Kurth, and T. Roder. 2004. Towards an efficient algorithm for automatic score-to-audio synchronization. In *Proceedings of the International Conference on Music Information Retrieval*, 365–72.
- Müller, M., H. Mattes, and F. Kurth. 2006. An efficient multiscale approach to audio synchronization. In *Proceedings of the International Conference on Music information Retrieval*, 192–7.
- Murbe, D., F. Pabst, G. Hofmann, and J. Sundberg. 2002. Effects of a professional solo singer education on auditory and kinesthetic feedback - a longitudinal study of singers pitch control. *TMH-QPSR* 43 (1): 81–7.
- Murphy, K. 1998. *Hidden Markov Model (HMM) Toolbox for Matlab*. Available from <http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html> (accessed April 20 2009).
- Myers, D., and J. F. Michel. 1987. Vibrato and pitch transitions. *Journal of Voice* 1 (2): 157–61.
- Narmour, E. 1990. *The Analysis and Cognition of Basic Musical Structures*. Chicago, IL: University of Chicago Press.
- Nichomachus. c. 100 CE. Enchiridion. In *Greek Musical Writings, Vol. 2. Harmonic and Acoustic Theory*, ed. A. Barker, 245–69. Cambridge, UK: Cambridge University Press.
- Niedermayer, B. 2009. Towards audio to score alignment in the symbolic domain. In *Proceedings of the 6th Sound and Music Computing Conference*, 77–82.

- Nienhuys, H.-W., and J. Nieuwenhuizen. 2003. Lilypond, a system for automatic music engraving. In *Proceedings of the Colloquium on Musical Informatics*, 167–72.
- Noll, A. M. 1967. Cepstrum pitch determination. *Journal of the Acoustical Society of America* 41: 293–309.
- Nordmark, J., and S. Ternström. 1996. Intonation preferences for major thirds with non-beating ensemble sounds. *TMH-Quarterly Progress and Status Report* 1: 57–61.
- Orio, N. 2002. Alignment of performances with scores aimed at content-based music access and retrieval. In *Lecture notes in computer science. Research and Advanced Technology for Digital Libraries* ed. M. Agosti and C. Thanos, 173–84. Berlin, DE: Springer.
- Orio, N. 2010. Automatic identification of audio recordings based on statistical modeling. *Signal Processing* 90 (4): 1064–76.
- Orio, N., and F. Déchelle. 2001. Score following using spectral analysis and hidden Markov models. In *Proceedings of the International Computer Music Conference*, 151–4.
- Orio, N., and D. Schwarz. 2001. Alignment of monophonic and polyphonic music to a score. In *Proceedings of the International Computer Music Conference*, 155–8.
- Ornøy, E. 2008. An empirical study of intonation in performances of J.S. Bach's Sarabandes: Temperament, “melodic charge” and “melodic intonation.” *Orbis Musicae* 14: 37–76.
- Oxenham, A., J. C. Micheyl, and M. V. Keebler. 2009. Can temporal fine structure represent the fundamental frequency of unresolved harmonics? *Journal of the Acoustical Society of America* 125 (4): 2189–9.
- Page, M. F. 2004. Perfect harmony: A mathematical analysis of four historical tunings. *Journal of the Acoustical Society of America* 116 (4): 2416–26.
- Palisca, C. V. 1994. *Studies in the History of Italian Music and Music Theory*. Oxford, UK: Oxford University Press.
- Palmer, C. 1996. Anatomy of a performance: Sources of musical expression. *Music Perception* 13: 433–54.
- Palmer, C. 1997. Music performance. *Annual Review of Psychology* 48: 115–38.

- Pardo, B., and W. P. Birmingham. 2001. Following a musical performance from a partially specified score. In *Proceedings of the Multimedia Technologies and Applications Conference*, 202–7.
- Pardo, B., and W. P. Birmingham. 2002. Improved score following for acoustic performances. In *Proceedings of the International Computer Music Conference*, 262–5.
- Pardo, B., and M. Sanghi. 2005. Polyphonic musical sequence alignment for database search. In *Proceedings of the International Conference on Information Retrieval*, 215–22.
- Parncutt, R. 1989. *Harmony: A Psychoacoustical Approach*. New York, NY: Springer-Verlag.
- Parsons, T. W. 1976. Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America* 60 (4): 911–8.
- Partch, H. 1974. *Genesis of a Music*. New York, NY: Da Capo Press.
- Patterson, R. D., K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. 1992. Complex sounds and auditory images. In *Auditory physiology and perception, Proceedings of the 9th International Symposium on Hearing*, ed. Y. Cazals, L. Demany, and K. Horner, 429–46. Oxford, UK: Pergamon.
- Peeling, P., T. Cemgil, and S. Godsill. 2007. A probabilistic framework for matching music representations. In *Proceedings of the International Conference on Music Information Retrieval*, 267–72.
- Plack, C. J. 2010. Musical consonance: The importance of harmonicity. *Current Biology* 20 (11): 476–8.
- Plack, C. J., and A. J. Oxenham. 2005. The psychophysics of pitch. In *Pitch: Neural Coding and Perception*, ed. C. J. Plack, A. J. Oxenham, R. R. Fay, and A. N. Popper, 7–55. New York, NY: Springer.
- Platt, R., W. 1998. ANOVA, t tests, and linear regression *Injury Prevention* 4: 52–3.
- Plomp, R. 1967. Pitch of complex tones *Journal of the Acoustical Society of America* 41 (6): 1526–33.
- Plomp, R., and W. J. M. Levelt. 1965. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America* 38 (4): 548–60.

- Prame, E. 1994. Measurements of the vibrato rate of ten singers. *Journal of the Acoustical Society of America* 96 (4): 1979–84.
- Prame, E. 1995. Measurement of the vibrato rate of ten singers. In *Vibrato*, ed. P. H. Dejonckere, M. Hirano and J. Sundberg. San Diego, CA: Singular Publishing Group.
- Prame, E. 1997. Vibrato extent and intonation in professional western lyric singing. *Journal of the Acoustical Society of America* 102 (1): 616–21.
- Pressnitzer, D., and S. McAdams. 1999. Two phase effects in roughness perception. *Journal of the Acoustical Society of America* 105 (5): 2773–82.
- Pressnitzer, D., and R. D. Patterson. 2001. Distortion products and the perceived pitch of harmonic complex tones. In *Physiological and psychophysical bases of auditory function*, ed. D. Breebaart, A. Houtsma, A. Kohlrausch, V. Prijs, and R. Schoonhovem. Maastrichtu, NL: Shaker.
- Ptolemy. c. 120 CE. Harmonika. In *Greek Musical Writings, Vol. 2. Harmonic and Acoustic Theory*, ed. A. Barker, 270–91. Cambridge, UK: Cambridge University Press. Original edition, 1989.
- Puckette, M. 1990. EXPLODE: A user interface for sequencing and score following. In *Proceedings of the International Computer Music Conference*, 259–61
- Puckette, M. 1995. Score following using the sung voice. In *Proceedings of the International Computer Music Conference*, 175–8.
- Puckette, M., and C. Lippe. 1992. Score following in practice. In *Proceedings of the International Computer Music Conference*, 185–2.
- Raatgever, J., and F. A. Bilsen. 1986. A central spectrum theory of binaural processing: Evidence from dichotic pitch. *Journal of the Acoustical Society of America* 80 (2): 429–41.
- Rabiner, L. R. 1977. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25 (1): 24–33.
- Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77 (2): 257–89.

- Rabiner, L. R., and B. Juang. 1993. *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.
- Rameau, J.-P. 1722. *Treatise on Harmony*. Translated by P. Gossett. Toronto, ON: Dover.
- Rameau, J.-P. 1737. *Génération Harmonique*. Paris, FR: Prault.
- Ramig, L. A., and T. Shipp. 1987. Comparative measures of vocal tremor and vocal vibrato. *Journal of Voice* 1 (2): 162–7.
- Ramis De Pareia, B. 1482. Musica practica. In *Strunk's Source Reading in Music History*, ed. O. Strunk and L. Treitler, 408–14. New York, NY: W. W. Norton & Company. Original edition, 1950.
- Raphael, C. 1999. Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21 (4): 360–70.
- Raphael, C. 2001. Music Plus One: A system for expressive and flexible musical accompaniment. In *Proceedings of the International Computer Music Conference*, 159–62.
- Raphael, C. 2004. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of the International Conference on Music Information Retrieval*, 387–94.
- Raphael, C. 2006. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning* 65 (2–3): 389–409.
- Rapoport, E. 2008. The marvels of the human voice: Poem–melody–vocal performance. *Orbis Musicae* 14: 7–36.
- Rasch, R. 2002. Tuning and temperament. In *The Cambridge History of Western Music Theory*, ed. T. Christensen, 193–222. Cambridge, UK: Cambridge University Press.
- Rasch, R., and R. Plomp. 1999. The perception of musical tones. In *The Psychology of Music*, ed. D. Deutsch, 89–112. San Diego, CA: Academic Press.
- Reinders, A. 1995. The history of vibrato in the singing voice. In *Vibrato*, ed. P. H. Dejonckere, M. Hirano and J. Sundberg. San Diego, CA: Singular Publishing Group.
- Repp, B. 1990. Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America* 88 (2): 622–41.

- Repp, B. 1992. Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei." *Journal of the Acoustical Society of America* 92 (5): 2546–68.
- Repp, B. 1997. The aesthetic quality of a quantitatively average music performance: Two preliminary experiments. *Music Perception* 14 (4): 419–44.
- Ritsma, R. J. 1967. Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America* 42: 191–8.
- Rodet, X. 1984. Time-domain formant wave-function synthesis. *Computer Music Journal* 8 (3): 9–14.
- Rossing, T., J. Sundberg, and S. Ternström. 1984. Acoustic comparison of voice use in solo and choir singing. *STL-Quarterly Progress and Status Report* 25 (1): 30–43.
- Rossing, T., J. Sundberg, and S. Ternström. 1985. Acoustic comparison of soprano solo and choir singing. *STL-Quarterly Progress and Status Report* 26 (4): 43–58.
- Rossing, T., J. Sundberg, and S. Ternström. 1987. Acoustic comparison of soprano solo and choir singing. *Journal of the Acoustical Society of America* 82 (3): 830–6.
- Rothman, H. B., and A. A. Arroyo. 1987. Acoustic variability in vibrato and its perceptual significance. *Journal of Voice* 1 (2): 123–41.
- Ryynanen, M. 2006. Singing transcription. In *Signal Processing Methods for Music Transcription*, ed. A. Klapuri and M. Davy, 361–90. New York, NY: Springer.
- Ryynanen, M., and A. Klapuri. 2004. Modelling of note events for singing transcription. In *Proceedings of the IEEE Workshop on Statistical and Perceptual Audio Processing*, 319–22.
- Ryynanen, M., and A. Klapuri. 2008. Query-by-Humming of MIDI and audio using locality sensitive hashing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2249–52.
- Salzberg, R. S. 1980. The effects of visual stimulus and instruction on intonation accuracy of string instrumentalists. *Psychology of Music* 8: 42–9.

- Sauveur, J. 1701. Système générale des intervalles du son. In *Acoustics: historical and philosophical development.*, ed. R. B. Lindsay, 88–94. Stroudsburg, PA: Dowden, Hutchinson and Ross. Original edition, 1973.
- Scheirer, E. 1995. *Extracting Expressive Performance Information from Recorded Music*. Master's Thesis, Massachusetts Institute of Technology, Media Laboratory.
- Scheirer, E. D. 1998. Using musical knowledge to extract expressive performance information from audio recordings. In *Computational Auditory Scene Analysis*, ed. H. Okuno and D. Rosenthal, 361–80. Mahwah, NJ: Lawrence Erlbaum.
- Schoen, M. 1922. An experimental study of the pitch factor in artistic singing. *Psychological Monographs* 31 (1): 230–59.
- Schön, D., P. Regnault, S. Ystad, and M. Besson. 2005. Sensory consonance: An ERP study. *Music Perception* 23 (2): 105–17.
- Schouten, J. F., R. J. Ritsma, and B. Lopes Cardozo. 1962. Pitch of the residue. *Journal of the Acoustical Society of America* 34 (8): 1418–24.
- Schroeder, M. R. 1968. Period histogram and product spectrum: new methods for fundamental frequency measurement. *Journal of the Acoustical Society of America* 43: 829–34.
- Schwarz, D., A. Cont, and N. Schnell. 2005. From Boulez to ballads: Training IRCAM's score follower. In *Proceedings of the International Computer Music Conference*, 547–50.
- Schwarz, D., N. Orio, and N. Schnell. 2004. Robust polyphonic MIDI score following with Hidden Markov Models. In *Proceedings of the International Computer Music Conference*, 442–5.
- Seashore, C. 1936a. *Objective Analysis of Musical Performance*. Iowa City, IA: University of Iowa Press.
- Seashore, C. 1938. *Psychology of Music*. Iowa City, IA: University of Iowa Press. Original edition, New York, NY: Dover Publications.
- Seashore, H. G. 1936b. An objective analysis of artistic singing. In *University of Iowa Studies in the Psychology of Music. Vol. IV: Objective Analysis of Musical Performance*, ed. C. Seashore, 12–157. Iowa City, IA: University of Iowa.

- Seeger, C. 1951. An instantaneous music notator. *Journal of the International Folk Music Council* 3: 103–6.
- Serra, X., and J. Smith. 1990. Spectral Modelling Synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal* 14 (4): 12–24.
- Sethares, W. A. 1993. Local consonance and the relationship between timbre and scale. *Journal of the Acoustical Society of America* 94 (3): 1218–28.
- Shackford, C. 1961. Aspects of perception. I: Sizes of harmonic intervals in performance. *Journal of Music Theory* 5 (2): 162–202.
- Shackford, C. 1962a. Aspects of perception III: Addenda. *Journal of Music Theory* 6 (2): 295–303.
- Shackford, C. 1962b. Aspects of perception. II: Interval sizes and tonal dynamics in performance. *Journal of Music Theory* 6 (1): 66–90.
- Shackleton, T. M., and R. P. Carlyon. 1994. The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. *Journal of the Acoustical Society of America* 95 (6): 3529–40.
- Shalev-Shwartz, S., S. Dubnov, N. Friedman, and Y. Singer. 2002. Robust temporal and spectral modeling for Query-by-Melody In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 331–8.
- Shalev-Shwartz, S., J. Keshet, and Y. Singer. 2004. Learning to align polyphonic music. In *Proceedings of the International Conference on Music Information Retrieval*, 381–6.
- Shih, H., S. S. Narayanan, and C.-C. J. Kuo. 2003. A statistical multidimensional humming transcription using phone level hidden Markov models for Query-by-Humming systems. In *Proceedings of the IEEE International Conference on Multimedia*, 61–4.
- Shonle, J. I., and K. E. Horan. 1980. The pitch of vibrato tones. *Journal of the Acoustical Society of America* 67 (1): 246–52.
- Smit, C., and D. Ellis. 2009. Guided harmonic sinusoid estimation in a multi-pitch environment. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, 41–4.

- Soulez, F., X. Rodet, and D. Schwarz. 2003. Improving polyphonic and poly-instrumental music to score alignment. In *Proceedings of the International Conference on Music Information Retrieval*, 143–8.
- Stammen, D., and B. Pennycook. 1993. Real-time recognition of melodic fragments using the dynamic timewarp algorithm. In *Proceedings of the International Computer Music Conference*, 232–5.
- Sundberg, J. 1972. Pitch of synthetic sung vowels. *STL-Quarterly Progress and Status Report* 13 (1): 34–44.
- Sundberg, J. 1978a. Effects of the vibrato and the singing formant on pitch. *Musicologica Slovaca* 6: 51–69.
- Sundberg, J. 1978b. Synthesis of Singing. *Swedish Journal of Musicology* 1: 107–12.
- 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 *The Psychology of Music*, ed. D. Deutsch, 171–214. San Diego, CA: Academic Press.
- Sundberg, J. 2006. The KTH synthesis of singing. *Advances in Cognitive Psychology* 2 (2–3): 131–43.
- Sundberg, J., A. Askenfelt, and L. Frydén. 1983. A synthesis-by-rule approach. *Computer Music Journal* 7 (1): 37–43.
- Sundberg, J., and J. Bauer-Huppmann. 2007. When does a sung tone start? *Journal of Voice* 21 (3): 285–93.
- Sundberg, J., A. Friberg, and R. Bresin. 2003. Attempts to reproduce a pianist's expressive timing with Director Musices performance rules. *Journal of New Music Research* 32: 317–25.

- Sundberg, J., A. Friberg, and L. Frydén. 1991. Threshold and preference quantities of rules for music performance. *Music Perception* 9: 71–92.
- Sundberg, J., E. Prame, and J. Iwarsson. 1995. Replicability and accuracy of pitch patterns in professional singers. *STL-Quarterly Progress and Status Report* 36 (2-3): 51-62.
- Suyoto, I. S. H., A. L. Uitdenbogerd, and F. Scholer. 2007. Effective retrieval of polyphonic audio with polyphonic symbolic queries. In *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, 105–14.
- Suyoto, I. S. H., A. L. Uitdenbogerd, and F. Scholer. 2008. Searching musical audio using symbolic queries. *IEEE Transactions on Audio Speech and Language Processing* 16 (2): 372–81.
- Swallowe, G. M., R. Perrin, G. Sattar, A. M. Colley, and D. J. Hargreaves. 1997. On consonance: Pleasantness and interestingness of four component complex tones. *Acustica* 83: 897–902.
- Swann, R. A. 1999. *An Investigation into the Harmonic Intonation Discrimination and Tuning Preferences of Choral Musicians*. PhD diss., School of Music, Florida State University.
- Terhardt, E. 1979. Calculating virtual pitch. *Hearing Research* 1: 155–82.
- Terhardt, E. 1984. The concept of musical consonance: A link between music and psychoacoustics. *Music Perception* 1 (3): 276–95.
- Terhardt, P. 1974. Pitch, consonance, and harmony. *Journal of the Acoustical Society of America* 55: 1061–9.
- Ternström, S. 1993. Perceptual evaluations of voice scatter in unison choir sounds. *Journal of Voice* 7 (2): 129–35.
- Ternström, S. 2002. Choir acoustics: an overview of scientific research published to date. *TMH-Quarterly Progress and Status Report* 43: 1–8.
- Ternström, S. 2003. Choir acoustics: An overview of research published to date. *International Journal of Research in Choral Singing* 1 (1): 3–12.

- Ternström, S., and D. R. Karna. 2002. Choir. In *The Science and Psychology of Music Performance Creative Strategies for Teaching and Learning*, ed. R. Parncutt and G. Mcpherson, 269–84. Oxford, UK: Oxford University Press.
- Ternström, S., and J. Sundberg. 1982. Acoustical factors related to pitch precision in choir singing. *STL-Quarterly Progress and Status Report* 23 (2–3): 76–90.
- Ternström, S., and J. Sundberg. 1988. Intonation precision of choir singers. *Journal of the Acoustical Society of America* 84: 59–69.
- Ternström, S., J. Sundberg, and A. Colldén. 1988. Articulatory F0 Perturbations and Auditory Feedback. *Journal of Speech and Hearing Research* 31: 187–92.
- Tiffin, J. 1931. Some aspects of the psychophysics of the vibrato. *Psychology Monographs* 41: 1153–200.
- Tiffin, J. 1932. The role of pitch and intensity in the vocal vibrato of students and artists. In *University of Iowa Studies in the Psychology of Music. Vol. I: The Vibrato*, ed. C. Seashore, 134–65. Iowa City, IA: University of Iowa.
- Timmers, R. 2007. Vocal expression in recorded performances of Schubert songs. *Musica Scientiae* 11 (2): 237–68.
- Timmers, R., and P. Desain. 2000. Vibrato: Questions and answers from musicians and science. In *Proceedings of the International Conference on Music Perception and Cognition*.
- Titze, I. R. 1994. *Principles of voice production*. Englewood Cliffs, NJ: Prentice-Hall.
- Titze, I. R., B. Story, M. Smith, and R. Long. 2002. A reflex resonance model of vocal vibrato. *Journal of the Acoustical Society of America* 111 (5): 2272–82.
- Todd, N. 1985. A model of expressive timing in tonal music. *Music Perception* 31 (1): 33–58.
- Todd, N. 1992. The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America* 91 (6): 3540–50.
- Toh, C. C., B. Zhang, and Y. Wang. 2008. Multiple-feature fusion based on onset detection for solo singing voice. In *Proceedings of the International Conference on Music Information Retrieval*, 515–20.

- Tove, P. A., B. Norman, L. Isaksson, and J. Czekajewski. 1966. Direct-recording frequency and amplitude meter for analysis of musical and other sonic waveforms. *Journal of the Acoustical Society of America* 39 (2): 362–71.
- Tsai, W.-H., and H.-M. Wang. 2004. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 221–4.
- Turetsky, R., and D. P. W. Ellis. 2003. Ground-truth transcriptions of real music from forced-aligned MIDI synthesis. In *Proceedings of the International Conference on Music Information Retrieval*, 135–42.
- Turner, R. S. 1971. The Ohm-Seebeck dispute, Hermann von Helmholtz, and the origins of physiological acoustics. *The British Journal for the History of Science* 10 (1): 1–24.
- Unal, E., E. Chew, P. G. Georgiou, and S. S. Narayanan. 2008. Challenging uncertainty in Query-by-Humming systems: A fingerprinting approach. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 16 (2): 359–71.
- van Besouw, R. M., J. S. Brereton, and D. M. Howard. 2008. Range of tuning for tones with and without vibrato. *Music Perception* 26 (2): 145–56.
- Vantomme, J. D. 1995. Score following by temporal pattern. *Computer Music Journal* 19 (3): 50–9.
- Vega, D. 2003. A perceptual experiment on harmonic tension and melodic attraction in Lerdahl's Tonal Pitch Space. *Musicae Scientiae* 7: 35–54.
- Vercoe, B. 1984. The synthetic performer in the context of live performance. In *Proceedings of the International Computer Muisc Conference*, 199–200.
- Vercoe, B., and M. Puckette. 1985. Synthetic rehearsal: Training the synthetic performer. In *Proceedings of the International Computer Muisc Conference*, 275–8.
- Vertfaille, V., C. Guastavino, and P. Depalle. 2005. Perceptual evaluation of vibrato models. In *Proceedings of the Conference on Interdisciplinary Musicology*.
- Vicentino, N. 1555. *Antica musica ridotta alla moderna prattica*. Translated by M. R. Maniates. New Haven, CT: Yale University Press.

- Viitaniemi, T., A. Klapuri, and A. J. Eronen. 2003. A probabilistic model for the transcription of single-voice melodies. In *Proceedings of the Finnish Signal Processing Symposium*, 59–63.
- Vincent, E., N. Bertin, and R. Badeau. 2007. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 109–12.
- Vos, J. 1982. The perception of mistuned fifths and major thirds. *Perception & Psychophysics* 32 (4): 297–313.
- Vos, J. 1984. Spectral effects in the perception of pure and mistuned intervals. *Perception and Psychophysics* 35 (2): 173–85.
- Vos, J., and B. G. van Vianen. 1985a. The effect of fundamental frequency on the discriminability between pure and tempered fifths and major thirds. *Perception & Psychophysics* 37: 507–14.
- Vos, J., and B. G. van Vianen. 1985b. Thresholds for discrimination between pure and tempered intervals: The relevance of nearly coinciding harmonics. *Journal of the Acoustical Society of America* 77 (1): 176–87.
- Vurma, A. 2010. Mistuning in two-part singing. *Logopedics Phoniatrics Vocology* 35: 24–33.
- Vurma, A., and J. Ross. 2006. Production and perception of musical intervals. *Music Perception* 23 (4): 331–44.
- Wang, C.-K., R.-Y. Lyu, and Y.-C. Chiang. 2003. A robust singing melody tracker using adaptive round semitones (ARS). In *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, 18–20.
- Weihs, C., and U. Ligges. 2003. Automatic transcription of singing performances. *Bulletin of the International Statistical Institute* 60: 507–10.
- Widmer, G. 2002. Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research* 31: 37–50.
- Widmer, G. 2003. Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence* 146 (2): 129–48.

- Widmer, G., S. E. Dixon, W. Goebel, E. Pampalk, and A. Tobudic. 2003. In search of the Horowitz factor. *AI Magazine* 24: 111–30.
- Widmer, G., and W. Goebel. 2004. Computational models of expressive music performance: The state of the art. *Journal of New Music Research* 33 (3): 206–16.
- Widmer, G., D. Rocchesso, V. Välimäki, C. Erkut, F. Gouyon, D. Pressnitzer, H. Penttinen, P. Polotti, and G. Volpe. 2007. Sound and music computing: Research trends and some key issues. *Journal of New Music Research* 36 (3): 169–84.
- Widmer, G., and A. Tobudic. 2003. Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research* 32: 259–68.
- Wightman, F. L. 1973. The pattern-transformation model of pitch. *Journal of the Acoustical Society of America* 54 (2): 407–16.
- Wild, J., and P. Schubert. 2008. Historically informed retuning of polyphonic vocal performance. *Journal of Interdisciplinary Music Studies* 2 (1–2): 121–39.
- Wilkinson, S. R. 1988. *Tuning In: Microtonality in Electronic Music*. Milwaukee, WI: Hal Leonard Books.
- Windsor, W. L., and E. F. Clarke. 1997. Expressive timing and dynamics in real and artificial musical performances: Using an algorithm as an analytical tool. *Music Perception* 15: 127–52.
- Woodruff, J., B. Pardo, and R. Dannenberg. 2006. Remixing stereo music with score-informed source separation. In *Proceedings of the International Conference on Music Information Retrieval*, 314–9.
- Yeh, C. H., A. Robel, and X. Rodet. 2005. Multiple fundamental frequency estimation of polyphonic music signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 225–8.
- Yoo, L., D. S. Sullivan Jr., S. Moore, and I. Fujinaga. 1998. The effect of vibrato on response time in determining the pitch relationship of violin tones. In *Proceedings of the International Conference on Music Perception and Cognition*, 477–81.
- Yost, W. A. 2009. Pitch perception. *Attention, Perception, & Psychophysics* 71 (8): 1701–15.

- Zanon, P., and G. De Poli. 2003a. Estimation of parameters in rule systems for expressive rendering in musical performance. *Computer Music Journal* 27: 29–46.
- Zanon, P., and G. De Poli. 2003b. Time-varying estimation of parameters in rule systems for music performance. *Journal of New Music Research* 32: 295–315.
- Zarlino, G. 1558. *Institutione harmoniche*. Translated by V. Cohen. Vol. 4, *On the modes*. New Haven, CT: Yale University Press.