The psychoacoustics and synthesis of singing harmony

Chan, Paul Yaozhu

2020

Chan, P. Y. (2020). The psychoacoustics and synthesis of singing harmony. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/142516>

<https://doi.org/10.32657/10356/142516>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 16:00:19 SGT

THE PSYCHOACOUSTICS AND SYNTHESIS OF SINGING HARMONY

PAUL YAOZHU CHAN

School of Computer Science and Engineering
2020

THE PSYCHOACOUSTICS AND SYNTHESIS OF SINGING HARMONY

PAUL YAOZHU CHAN

A thesis report submitted to
the School of Computer Science and Engineering
in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

June 2020

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

Paul Yaozhu Chan
(21/7/2019)

SUPERVISOR DECLARATION STATEMENT

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

.....
Associate Professor Eng Siong Chng
(21/7/2019)

Authorship Attribution Statement

This thesis contains material from 1 accepted peer-reviewed journal, 3 published conference papers and 3 filed patents, where I was the first and/or corresponding author/inventor.

Chapter 2 has been accepted as Paul Yaozhu Chan, Minghui Dong and Haizhou Li, “The Science of Harmony: A Psychophysical Basis for Perceptual Tensions and Resolutions in Music,” in *Research*, Submitted Aug 2018. The contributions of the co-authors are as follows:

- I proposed the idea, designed the study, wrote the manuscript, and performed the experiments.
- Dr Minghui Dong co-designed the study, and revised the manuscript.
- Prof Haizhou Li co-designed the study, and revised the manuscript.

Part of Chapter 3 has been published as Paul Yaozhu Chan, Minghui Dong, Grace X. H. Ho, and Haizhou Li, “SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-Time Speech and Singing Applications on Mobile Platforms,” in *INTERSPEECH*, pp. 1225-1229, 2016. The contributions of the co-authors are as follows:

- I proposed the idea, designed the study, wrote the manuscript, and performed the experiments.

- Dr Minghui Dong co-designed the study, and revised the manuscript.
- Prof Haizhou Li co-designed the study, and revised the manuscript.

Part of Chapter 3 has been published as Paul Yaozhu Chan, Minghui Dong, Grace X. H. Ho, and Haizhou Li, “SERAPHIM Live! Singing Synthesis for the Performer, the Composer, and the 3D Game Developer,” in *INTERSPEECH*, pp. 1966-1967, 2016. The contributions of the co-authors are as follows:

- I proposed the idea, designed the study, wrote the manuscript, and performed the experiments.
- Dr Minghui Dong co-designed the study, and revised the manuscript.
- Prof Haizhou Li co-designed the study, and revised the manuscript.

Part of Chapter 3 has been patented as Paul Yaozhu Chan, Minghui Dong, Haizhou Li, “A Wavetable Synthesis System with 3D Lip Animation for Real-time Speech and Singing Applications on Mobile Platforms,” *Singapore Patent*, Pub no. SG/P/2016002, Filed 2016. The contributions of the co-authors are as follows:

- I conceptualized the idea, designed the study, wrote the patent, and performed the experiments.
- Dr Minghui Dong co-conceptualized the idea, and revised the drafts.
- Prof Haizhou Li co-conceptualized the idea, and revised the drafts.

Part of Chapter 4 has been published as Paul Yaozhu Chan, Minghui Dong, Siu Wa Lee, Ling Cen, and Haizhou Li, “Solo to A Capella Conversion - Synthesizing Vocal Harmony from Lead Vocals,” in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2011¹. The contributions of the co-authors are as follows:

- I proposed the idea, designed the study, wrote the drafts of the manuscript, and performed the experiments.
- Dr Minghui Dong co-designed the study, and revised the manuscript.

¹This work started during application for candidature, which was before (but extended beyond) the official commencement of candidature.

- Dr Siu Wa Lee co-designed the study, and revised the manuscript.
- Dr Ling Cen co-designed the study, and revised the manuscript.
- Prof Haizhou Li co-designed the study, and revised the manuscript.

Part of Chapter 4 has been patented as Paul Yaozhu Chan, Minghui Dong, Ling Cen and Siu Wa Lee, “Auto-Synchronous Singing Harmonizer,” *US Patent*, Pub no. US20120234158 A1, App no. US 13/418,236, Filed & Published 2012².

The contributions of the co-authors are as follows:

- I conceptualized the idea, designed the study, wrote the drafts of the patent, and performed the experiments.
- Dr Minghui Dong co-conceptualized the idea, and revised the drafts.
- Dr Siu Wa Lee co-conceptualized the idea, and revised the drafts.
- Dr Ling Cen co-conceptualized the idea, and revised the drafts.
- Prof Haizhou Li co-conceptualized the idea, and revised the drafts.

Part of Chapter 4 has been patented as Paul Yaozhu Chan, Minghui Dong, Ling Cen and Siu Wa Lee, “Harmony Synthesizer and Method for Harmonizing Vocal Signals,” *PRC Patent*, Pub no. CN102682762 A, App no. CN 201210068847, Filed & Published 2012².

The contributions of the co-authors are as follows:

- I conceptualized the idea, designed the study, wrote the drafts of the patent, and performed the experiments.
- Dr Minghui Dong co-conceptualized the idea, and revised the drafts.
- Dr Siu Wa Lee co-conceptualized the idea, and revised the drafts.
- Dr Ling Cen co-conceptualized the idea, and revised the drafts.
- Prof Haizhou Li co-conceptualized the idea, and revised the drafts.

Paul Yaozhu Chan
(21/7/2019)

Abstract

The human singing voice is a remarkable instrument that compounds an immense amount of expressivity onto a single dimension. Apart from semantics and melody (pitch, duration and dynamics), accent, age, gender and emotion are all carried in the singing voice. While a single singing voice on its own is aesthetically pleasing to the ear, the addition of concurrent voices of different pitch is commonly known to be capable of producing a pleasing effect far greater than the sum of that produced by each contributing voice. This motivates the use of harmony in singing. Unfortunately, accompaniment voices are difficult to sing, even for professional singers. Thankfully singing synthesis has made it viable for this task to be undertaken by machines. The overall objective of this thesis is to advance today's understanding of singing harmony and ultimately develop novel techniques for its synthetic reproduction. This is broken down into three parts. The first focuses on a psychophysical basis of harmony, the second focuses on the synthesis of the singing voice, while the third combines the first two to focus on the synthesis of harmonized singing.

The first contribution is an attempt to find a psychoacoustic basis of harmony and presented in chapter 2. Apart from stationary harmony (*chords*, or *sonorities*: the aesthetics of a group of concurrent notes at one point of time), this also includes transitional harmony (*chord progression*, or *resolution*: the aesthetics of a similar group of notes progressing to another). In order to explain both stationary and transitional harmony, it introduces a theory of harmony based on the notions of interharmonic and subharmonic modulations. Acoustic measures of stationary and transitional harmony are proposed and the answers to five fundamental questions

of psychoacoustic harmony are presented, both based on this theory. Correlations with existing music theory and perception statistics support this contribution with both stationary and transitional harmony.

The second contribution is in the synthesis of the singing voice and presented in chapter 3. Modern singing synthesis methods are at best capable of word-level runtime synthesis, with only two known ones dedicated to realtime synthesis. This means that they are applicable only towards offline music production. A large part of the art of music and singing, however, is in realtime performance. With both of the existing realtime singing synthesis methods bounded by a phone-coverage to realtime-capability tradeoff, a need for one that overcomes it remains. A novel realtime singing synthesis system, SERAPHIM, is proposed as an answer to this. Apart from overcoming this phone-coverage to realtime-capability trade-off, subjective listening tests also showed that listeners preferred voices synthesized by SERAPHIM as opposed to other realtime systems.

The third contribution is in the synthesis of singing harmony and presented in chapter 4. With this contribution, a novel method for singing harmony synthesis is proposed. Current implementations can be classified into pitch-inaccurate rule-based systems, timing-inaccurate inference-based systems, and hybrid systems that trade off between pitch inaccuracies and timing inaccuracies. This means that existing systems are vulnerable to either pitch errors, timing errors or both in different degrees of compromise. The challenge in the task was to overcome this compromise to develop a robust technique that is simultaneously resilient to both pitch and timing errors while producing harmonious accompaniment. Our strategy was to leverage on the pitch-accurate inference-based method while eliminating timing inaccuracies by use of machine-synchronization. Spectrograms revealed that harmonized voices produced by this method contain the least dissonances amongst existing methods. Subjective listening tests also showed that harmonized voices produced by this method are perceived to be the best sounding, both by vocal experts and by casual listeners.

All in all, the work presented in this thesis contributes to the advancement of the psychoacoustic understanding and machine synthesis of singing harmony across

one journal paper, three conference papers and three patents.

Acknowledgements

I would like to thank my supervisors, A/Prof Eng Siong Chng (Nanyang Technological University), Prof Haizhou Li (National University of Singapore) and Dr Minghui Dong (Institute for Infocomm Research) for giving me the opportunity to undertake this research and nurturing me in my career as a researcher while giving me the much-needed space to learn and grow.

Further to this, I would like to thank my fellow colleagues and students across the Institute for Infocomm Research, Nanyang Technological University and National University of Singapore for the friendship and the camaraderie and for always cheering me on. Especially Ms Aiti Aw for allowing me to work in a parallel field as my research; colleagues at One North Christian Fellowship such as Ms Susan Yap, Dr Yi Yan Yang, Dr Peter Yu Chen and Dr Francois Chin for constantly upholding me in prayer; and the LabRats, the unofficial band of the Agency for Science, Technology and Research, for helping me keep my sanity through the series of gigs and events.

Finally, I would like to thank my parents back home, Ron and Lili; my wife, Jing; and daughter, Dawn, for their love, support, encouragement and prayers.

Contents

Authorship Attribution Statement	iii
Abstract	vi
Acknowledgements	ix
List of Publications	xiv
List of Figures	xix
List of Tables	xx
1 Introduction	1
1.1 Motivation and Scope	2
1.2 Background	2
1.3 Contributions	6
1.3.1 Psychoacoustics of Harmony	6
1.3.2 Singing Synthesis (SERAPHIM)	7
1.3.3 Vocal Harmony Synthesis	8
1.4 Organisation of Thesis	10
2 The Psychoacoustics of Harmony	11
2.1 Background	12
2.1.1 Existing Work	13
2.1.2 Scope	15
2.2 A Psychophysical Basis of Harmony	17
2.3 Modulations in Sinusoidal Summation	18
2.4 Interharmonic Modulations	22
2.4.1 Beating Frequencies and Low-Frequency Modulations	22
2.4.2 Perceptual Responses across the Δf - $\bar{f}$ feature Space	23

2.4.3	Intervals and Second-Order Modulations on the Δf - $\bar{f}$ feature Space	25
2.5	Subharmonic Modulations	29
2.5.1	Subharmonic Modulations in Stationary Harmony	33
2.5.2	Subharmonic Modulations in Transitional Harmony	39
2.6	Experiment and Results	47
2.6.1	Stationary Harmony	47
2.6.2	Transitional Harmony	49
2.7	Addressing the Fundamental Questions of Psychoacoustic Harmony	52
2.8	Conclusion	55
3	Singing Synthesis	56
3.1	Background	56
3.1.1	History of Artificial Voice Production	57
3.1.2	Overview of Singing Synthesis Methods	59
3.1.3	Notable Works	63
3.2	Baseline Synthesizer	66
3.2.1	Mathematical Definition of System Inputs and Outputs . . .	66
3.2.2	System Components	67
3.2.3	Analysis and Synthesis Process of the Vocoder Component .	69
3.3	Proposed System	71
3.3.1	Scope	71
3.3.2	Runtime Algorithm	73
3.3.2.1	Syllable Structure and Phonetic Background	73
3.3.2.2	Phone Model Database	75
3.3.2.3	Syllable Model Database	76
3.3.3	Realtime Algorithm	77
3.4	Experiment and Results	82
3.5	Conclusion	83
4	The Synthesis of Singing Harmony	85
4.1	Background	86
4.1.1	Resynthesis Strategy and Spectral Features	86

4.1.2	Pitch Features with Existing Automatic Harmonization Methods	87
4.1.2.1	The Fourths, Fifths and Octaves Method	88
4.1.2.2	The Thirds and Sixths Method	89
4.1.2.3	The Auxiliary Method	91
4.1.2.4	The Karaoke Method	92
4.1.3	Psychoacoustic Analyses of Existing Methods	93
4.2	The Proposed Method	95
4.2.1	Algorithm	96
4.2.2	Pitch Interpretation	99
4.2.2.1	Fundamental Frequency Estimation	100
4.2.2.2	Octave Correction	101
4.2.2.3	Translation to MIDI Note-Number Scale	101
4.2.2.4	Estimation of Overall Tuning Drift	102
4.2.2.5	Key Determination	104
4.2.2.6	Note Rounding	106
4.2.2.7	Rule-based Transient Segment Correction	107
4.2.3	The Alignment Process	108
4.2.3.1	Dynamic Time Warping	109
4.2.3.2	Realignment	110
4.2.4	Re-Synthesis	111
4.3	Experiment and Evaluation	112
4.3.1	Spectrograms	113
4.3.2	Subjective Listening Tests	115
4.3.2.1	Vocal Experts	115
4.3.2.2	Casual Listeners	116
4.4	Conclusion	117
5	Conclusion and Future Work	118
5.1	Conclusion	118
5.2	Future Work	119
	References	120

Appendices	134
A Music Theory Prerequisites	135
A.1 Vocal Harmony	135
A.2 Special Sets of Notes	135
A.2.1 Keys	136
A.2.2 Chords	138
A.3 Harmonization	140
A.4 Consonances and Dissonances in the Context of Conventional Music Theory	141
B Supplementary Audio, Video and Tables	144

List of Publications

- (i) Paul Yaozhu Chan, Minghui Dong and Haizhou Li, “The Science of Harmony: A Psychophysical Basis for Perceptual Tensions and Resolutions in Music,” in *Research*, vol. 2019, Article ID 2369041, 22 pages, 2019. <https://spj.sciencemag.org/research/2019/2369041/>
- (ii) Paul Yaozhu Chan, Minghui Dong, Haizhou Li, “A Wavetable Synthesis System with 3D Lip Animation for Real-time Speech and Singing Applications on Mobile Platforms,” *Singapore Patent*, Pub no. SG/P/2016002, Filed 2016.
- (iii) Paul Yaozhu Chan, Minghui Dong, Grace X. H. Ho, and Haizhou Li, “SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-Time Speech and Singing Applications on Mobile Platforms,” in *INTERSPEECH*, pp. 1225-1229, 2016.
- (iv) Paul Yaozhu Chan, Minghui Dong, Grace X. H. Ho, and Haizhou Li, “SERAPHIM Live! Singing Synthesis for the Performer, the Composer, and the 3D Game Developer,” in *INTERSPEECH*, pp. 1966-1967, 2016.
- (v) Paul Yaozhu Chan, Minghui Dong, Siu Wa Lee, Ling Cen, and Haizhou Li, “Solo to A Capella Conversion - Synthesizing Vocal Harmony from Lead Vocals,” in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2011².
- (vi) Paul Yaozhu Chan, Minghui Dong, Ling Cen and Siu Wa Lee, “Auto-Synchronous

²This work started during application for candidature, which was before (but extended beyond) the official commencement of candidature.

Singing Harmonizer,” *US Patent*, Pub no. US20120234158 A1, App no. US 13/418,236, Filed & Published 2012².

- (vii) Paul Yaozhu Chan, Minghui Dong, Ling Cen and Siu Wa Lee, “Harmony Synthesizer and Method for Harmonizing Vocal Signals,” *PRC Patent*, Pub no. CN102682762 A, App no. CN 201210068847, Filed & Published 2012².

List of Figures

1.1	Notes of a melody	3
1.2	Notes in harmony	3
1.3	Existing against Ideal Realtime Capability and Phone Coverage of Realtime Singing Synthesis Systems	8
1.4	Example of Perfect Case of Harmonizer Function	9
1.5	Example of Pitch Errors	9
1.6	Example of Timing Errors	10
2.1	Summation of sinusoids of equal (left) versus unequal (right) am- plitudes	19
2.2	$A \cos \omega_1 t + B \cos \omega_2 t$ for various values of B normalized to $A = 1$. .	20
2.3	Identifying the interharmonic modulations across the notes c_3 and e_3^b	22
2.4	Types of Interharmonic Modulation on the scale of Δf . Low- frequency modulation correspond to consonance at small Δf , while beat frequencies [32] correspond to an unpleasant "roughness" termed dissonance at higher Δf . Modulations with near-zero Δf fall below musical significance while those with Δf pass a certain threshold are perceived as distinctly separate.	23

2.5	Even though, as one might expect, emotive responses would be different for every individual, the response for one individual can be plot as an example. The figure shows an example of auditory responses triggered in the mind of the author when exposed to pure-tone frequencies on the horizontal ($\bar{f}$) axis modulated at frequencies on the vertical (Δf) axis. Green, yellow, orange, red and black indicate pleasing, somewhat pleasing, unpleasant, dissonant and beyond beating range, respectively.	24
2.6	Interharmonic plots for perfectly consonant intervals within an octave.	26
2.7	Interharmonic plots for imperfectly consonant intervals in within an octave.	27
2.8	Interharmonic plots for dissonant intervals in within an octave. . . .	28
2.9	Subharmonic wave formation and wave deformation in the C and Cm^7 chords.	30
2.10	Plotting subharmonic wave periods of the C Major chord.	34
2.11	Subharmonic plot of the opening stanza of Pachelbel's Canon in D. Subharmonics are computed from the fundamental wave periods of each note and plot with T_{sub} in milliseconds on the vertical axis and time in bars on the horizontal axis. Subharmonics are plot in the colour matching their corresponding notes on the music score. The subharmonic tensions of each chord, Δt , are marked out on the plot with white arrows. Significant wave periods, along with common subharmonic periods, T_{sub} , are marked against the vertical axis on the right.	35
2.12	An illustration of the effect of harmony, $\varepsilon\{X\}$, on the scale of subharmonic tension, $\Delta\hat{t}$, according to the proposition of $\Delta\hat{t}$ as a measure of tension and dissonance.	38

2.13	Subharmonic plot of the opening line of Beethoven’s Moonlight Sonata with T_{Sub} in the vertical axis and time in bars in the horizontal axis. Subharmonics are coloured to their corresponding notes on the music score. Names of relevant notes are marked out on the left, at T_{Sub} values corresponding to their wave period. The region of each transition is numbered in white. Coloured arrows follow voice leading along the notes across chord changes.	41
2.14	Trajectories of $k_i t_i$ in different states of tension development (states of convergence).	44
3.1	A Timeline of Voice Reproduction and Singing Synthesis	57
3.2	Formant Synthesis	60
3.3	Physical Modeling / Articulatory Synthesis	61
3.4	Parametric Synthesis	61
3.5	Wavenet Synthesis	62
3.6	Wavetable Synthesis	63
3.7	System Definition	66
3.8	A Typical Conventional Singing Synthesizer	67
3.9	The Analysis and Synthesis Processes of a Vocoder	70
3.10	Realtime Capability and Phone Coverage of SERAPHIM against existing Realtime Singing Synthesis Systems	72
3.11	Runtime Synthesis Flowchart	74
3.12	Phone- and Biphone-Model Wavetables	75
3.13	Wave Additive Trajectories across 3 Different Syllable Lengths for the Mandarin Syllable ‘Shuang’	77
3.14	Realtime Synthesis Flowchart	78
3.15	Mandarin Chinese SERAPHIM <i>Live!</i> interface: Initials	78
3.16	Mandarin Chinese SERAPHIM <i>Live!</i> interface: Finals	79
3.17	Mandarin Chinese SERAPHIM <i>Live!</i> interface: Slurring of the word “Shui”. The phone “u” is commonly slurred with certain Mandarin accents	79
3.18	Japanese SERAPHIM <i>Live!</i> interface: Initials	81
3.19	Japanese SERAPHIM <i>Live!</i> interface: Finals	81

3.20	Subjective Listening Tests	83
3.21	Realtime Capability and Phone Coverage of Existing Methods . . .	84
4.1	The Fourths, Fifths and Octaves Method (458)	88
4.2	The Thirds and Sixths Method (458-II)	90
4.3	The Auxiliary Method (AUX)	91
4.4	The Karaoke Method (KTV)	92
4.5	Distortion Tables for 3 rd s above, 3 rd s below, 6 th s above and 6 th s below with the 458-II method	94
4.6	Pitch and Timing Performance Trade-offs of Rule-based, Hybrid and Inference-based Systems against that of the Proposed Strategy	95
4.7	The Proposed Strategy	97
4.8	The Proposed Method, the Solo-to-Acapella Method (S2A)	98
4.9	Pitch Interpretation	99
4.10	Estimation of Overall Tuning Drift, Δf_0	102
4.11	Determination of Key, k	105
4.12	Note Rounding	106
4.13	The Dynamic Time Warping Process	109
4.14	The Realignmmt Process	111
4.15	Spectrogram of the 458, KTV and S2A methods respectively, against that of the human voice	114
A.1	Composition of Singing Harmony	136
A.2	Notes of a Melody	137
A.3	A Special Set of Notes: Key	137
A.4	Simultaneously Sounded Notes	139
A.5	A Special Set of Notes: Chord	139
A.6	Consonances and Dissonances	142

List of Tables

1.1	Existing Methods of Harmony Derivation	9
2.1	Summary of correlations with consonance rankings and historical chord use.	47
2.2	Proposed and Existing Correlates of Stationary Harmony.	49
2.3	Supplementary Table (Tab. S1) Tabulation of $\Delta\Delta\hat{t}$ against Ty- moczko’s chord tendency statistics [26] computed over 11,000 tran- sitions.	51
2.4	Tabulation of correlations between $\widetilde{\Delta\Delta t}$ and Palestrina’s chord use statistics as collated in [26]. Correlations are listed in the top row with corresponding significance in brackets below.	52
3.1	Real-time Singing Synthesis Systems	82
3.2	Subjective Listening Test Results	82
4.1	Current Methods of Harmony Generation	87
4.2	Score across 11 vocal experts on consonance / harmony	115
4.3	Score across 11 vocal experts on smoothness of transition	115
4.4	Score across 12 casual listeners on pleasantness / naturalness	116

Chapter 1

Introduction

The human voice is a remarkable instrument capable of dextrous pitch-bends, phonation in multiple registers, unvoiced sounds and the mimicry of a broad range of timbres. It has the ability to compound an immense amount of expressivity onto a single dimensional acoustic waveform. Apart from semantics in the form of words and melody in the form of pitch, duration and dynamics, components such as accent, age, gender and emotion are all carried in a single singing voice. While a single singing voice on its own is known to be aesthetically pleasing to the ear, the addition of concurrent voices of different pitch is capable of producing a pleasing effect far greater than the sum of that produced by each contributing voice [1,2]. This motivates the use of singing harmony in music. The objective of this thesis is to advance today's scientific understanding of singing harmony and develop novel techniques for its synthetic reproduction.

Harmony is the phenomenon of blending notes in music to produce a pleasing effect greater than that of the sum of its parts [1,2]. Singing harmony is the production of the aforementioned pleasing effect with the blending of two or more singing voices [3-6]. Traditionally, accompaniments are added to a monophonic (single voice) part in the act termed harmonizing. Unfortunately, accompaniment voices are difficult to sing, even for professional singers [7]. Thankfully singing synthesis has made it viable for this task to be undertaken by machines. In this thesis, we

present our work in the psychoacoustics and synthesis of singing harmony.

1.1 Motivation and Scope

This thesis is focused on the creation of singing harmony by the addition of parallel singing voices (accompaniment) to a single singing voice (melody). It is widely used in music, and commonly in singing, for the desirable aesthetic effect that it produces. The difficulty of singing accompaniment parts [7], together with spontaneous advancements in speech synthesis over the recent years [8, 9], provide a general motivation for this task to be undertaken by machines.

The successful generation of aesthetically pleasing singing harmony has two aspects. Apart from a robust method of singing synthesis the choice of notes that make up the accompaniment is just as important. In practice, the latter is usually dictated by music theory. Unfortunately, music theory has widely been regarded to be baseless and unscientific [10, 11]. This further provides a motivation for a better understanding of harmony aesthetics from a psychophysical perspective.

Thus, the scope of this work is in the psychoacoustics and synthesis of singing harmony.

1.2 Background

Harmony, Harmonization and their Role in Music: Harmony, in the context of singing voices, refers to the concurrent use of two or more inter-complementary layers of notes. If melody were depicted as the trajectory of musical pitch across monotonically increasing time, harmonization may be understood to be the addition of one or more complementary trajectories alongside it to produce harmony. The piano rolls of the melody and harmony of Brahms' lullaby are shown in Figures 1.1 and 1.2, respectively. In the vertical axis is pitch and in the horizontal axis is time. In this example, as well as throughout Chapter 4, harmony refers to

Figure 1.1: Notes of a melody

Figure 1.2: Notes in harmony

that for sung parts and is conventionally composed of the melody and one or more accompaniments, as marked *Acc1* and *Acc2* in the figure.

Harmony has a pleasant effect in music, that can be described to be greater than the sum of its components [1, 2]. Apart from its role in the enhancement of a single melody line, it is useful in the expression of emotional valence [12] as well as the development of tensions [1, 13–15] and resolutions [1, 16, 17] in music.

Understanding Harmony with Psychoacoustics: Despite the significance of harmony in music, it still lacks a well-grounded psychoacoustic basis [18]. Advances in deep learning may have had successes at allowing machines to re-create harmony by mimicry but are not aimed at advancing the human understanding of scientific principles of harmony [8, 19–24]. Hence, even today, many aspects of musical harmony still remain disjoint with science and reason [10, 11].

There are three components in our understanding harmony. These are:

1. What: To know which notes sound good together.
2. Why: The explanation of why these notes sound good together.
3. When: Different chords¹ sound better or worse than others when trailing different preceding chords. This aspect should be understood as well.

Both traditional music theory and the field of statistics have had notable success at answering the first and third questions [3, 25]. Traditional music theory does this by the cumulation, justification and clustering of experience [11]. Statistics, on the other hand, does this through the systematic analyses of the historic use of harmony [26]. Both fields approach the problem from the perspective of perception rather than principle [10, 11, 26]. Apart from this, both of fields also approach the problem at a note level rather than an acoustic level [16, 26] .

Psychoacoustics marries the science of sound with the perception of aesthetics to approach this problem from scientific principle. The problem of harmony has proved to be interesting to some of the brightest minds in physic and mathematics throughout history, including Pythagoras, Euclid, Aristotle, Zarlino, Zhu, Galileo, Euler, Helmholtz, Stumpf and Rameau [27–34]. Unfortunately, a thorough psychophysical understanding of harmony has remained elusive [10, 11]. The widely accepted psychoacoustic theories of harmony today are Pythagoras’ rational wave-length theory [35], which is an answer to the aforementioned *what*; and Helmholtz’s beating frequencies [32], which is an incomplete proposition to the aforementioned *why* with several issues that remain unresolved [30, 31, 36–38]. With *why* hanging unresolved, the question of *when* remains largely unapproached in psychoacoustics. In Chapter 2 of this thesis, we will present and validate a theory that bridges Pythagoras with Helmholtz and propose our answer to the *what*-, *why*- and *when*-qualities of harmony.

Harmony in Singing: Harmonized singing in modern Western music draws its roots from *A Capella* singing which dates back to as early as the middle ages [39]. Two to four vocal parts are typically used, usually without further accompaniment from other musical instruments, to greatly enhance the melody.

¹Groups of notes simultaneously presented in harmony. See Appendix A

Unfortunately, humans have a strong tendency of gravitating to the main melody when singing, making vocal harmony extremely difficult to sing, even for professional singers [40]. This is the main motivation behind the synthesis of singing harmony. In order to synthesize singing harmony, we have to take a closer look at singing synthesis.

Singing Voice Synthesis: Singing synthesis is the artificial production of the human singing voice by electronic means. Typically, the lyrics and pitch will be specified by a user.

Since the launch of Yamaha VOCALOID’s Hatsune Miku in 2007, singing synthesis took the world by storm [41]. The virtual singer opened the doors to every indie composer without access to professional singers by allowing them to produce freely for Miku and publishing it directly through the *creative commons license*. Apart from Miku’s influence on popular music culture with her numerous albums, music videos and sell-out concerts across Japan and America, she also reaps revenue from video games, arcade games, merchandise and brand endorsements [41, 42]. These provide a strong motivation for further work in this area.

Synthesizing Singing Harmony: The synthesis of singing harmony typically involves the synthesis of accompaniment parts to harmonize with a melody sung by a human singer. It involves the application of singing synthesis while leveraging on information extracted from the melody sung by the human singer. Existing methods are typically plagued by either timing problems or pitch problems or trade off between the two. An ideal solution should overcome this trade off to resolve both problems.

1.3 Contributions

There are two prerequisites to the synthesis of singing harmony. They are a thorough understanding of the basis of harmony and a robust method of singing synthesis. Hence, the three contributions are presented in this thesis towards the ultimate fulfilment of machine-generated singing harmony. The first of which is towards a better human understanding of harmony. The second is towards the machine generation of singing voice. The third combines the understanding of harmony and the method of singing synthesis into a robust singing harmony system.

1.3.1 Psychoacoustics of Harmony

In the first study, a psychophysical basis for harmony is explored. Harmony studies the phenomenal use of simultaneous notes in music to produce a pleasing effect greater than the sum of its parts [1, 2]. With stationary harmony², acoustic theories explaining consonances and dissonances that have been widely accepted are centred around two schools: rational relationships (commonly credited to Pythagoras) and Helmholtz’s beating frequencies. The first is more of an attribution (*what*) than a psychoacoustic explanation (*why*) while several have raised discrepancies with the second not agreeing with certain acoustic or physiological studies. [30, 31, 36–38, 43, 44]. What transitional harmony³ is a more complex problem that currently remains unaddressed by the science of acoustics. In order to explain both stationary and transitional harmony, we propose the notion of interharmonic and subharmonic modulations. Earlier parts of this contribution bridges the two schools and show how they stem from a single equation. Later parts of the contribution focus on subharmonic modulations in order to explain aspects of harmony that interharmonic modulations cannot. Introducing the concepts of stationary and transitional subharmonic tensions, it is shown to explain the concepts of not only stationary (*what* and *why*) harmony, but also transitional (*what*, *why* and *when*) harmony. By which, the fundamental questions of psychoacoustic harmony such as why the pleasing effect of harmony is greater than that of the sum of its

²sonorities, or notes that are presented together at a single point in time

³progression, or how well a group of notes transits to another group of notes

parts is addressed. Finally, strong correlations with traditional music theory and perception statistics are presented to support this theory with both stationary and transitional harmony.

This novel contribution was submitted to Research ⁴ [45] and will be covered in detail in Chapter 2.

1.3.2 Singing Synthesis (SERAPHIM)

The second objective of this work is to synthesize the human singing voice. State-of-the-art synthesizers are suitable for offline music production purposes and unable to meet realtime demands for live onstage use. A large part of music and singing, however, involves the realtime feedback, influence and interaction amongst the singers, instrumentalists, and audiences. This provides a strong motivation for work in realtime singing synthesis systems.

There are presently two existing methods for realtime singing synthesis: LIMSI's Cantor Digitalis [46] and Yamaha's VOCALOID Keyboard [47–50]. The Cantor Digitalis is a touch-based vowel synthesizer, while VOCALOID Keyboard is a syllable-level keypad/keyboard-based synthesizer.

Figure 1.3 shows the realtime capabilities and phone coverage of the two in blue and green, respectively. As shown in the figure, Cantor Digitalis is realtime to the sub-frame level (frame level or lower) but only covers vowels. VOCALOID Keyboard, on the other hand, covers all phones but is realtime only to the syllable level. The objective, as illustrated in red in the figure, is to achieve:

1. for vowels and voiced consonants: sub-frame capability for better response and control.
2. for unvoiced consonants: phone level capability for segments too fast for human control.

⁴The official journal of CAST. A Science Partner Journal.

Figure 1.3: Existing against Ideal Realtime Capability and Phone Coverage of Realtime Singing Synthesis Systems

Subframe response is desirable for ordinary segments for maximal expressivity. For unvoiced consonants, however, sudden changes in spectrum can be faster non-verbal human response and control, hence, for these, we wish to pass some level of control back to the synthesizer and settle for phone-level capability.

The novel method described here was published to Interspeech 2016 [51, 52]. Patent for parts of this work was filed in Singapore [53]. This is presented in detail in Chapter 3.

1.3.3 Vocal Harmony Synthesis

The third objective of this work is to automatically generate pleasant sounding accompaniment voices given the voice of the sung melody.

Existing methods are tabulated in Table 1.1. Each method is affected by its own

Table 1.1: Existing Methods of Harmony Derivation

Strategy	Method	Auxiliary Input	Applications
Rule-based	458 [7, 54–57]	None	Stage
Hybrid	458-II [7, 54, 58, 59]	Key (User)	Stage, Studio
Hybrid	AUX [7, 54, 55, 58, 59]	Guitar (User)	Stage
Inference-based	KTV [7, 54, 56, 60]	MIDI (User/System)	KTV, Studio

shortcomings which may be attributed to either incorrect pitch or incorrect timing. Figure 1.4 shows an example of an ideal accompaniment (magenta) for a given input (green). Figure 1.5 shows an example of accompaniment with pitch errors (orange), for the same input (green), in which errors are circled, with the correct accompaniment is shown in (magenta). In this example, the accompaniment is incorrect as it does not conform to the key nor chords of the song. Figure 1.6 shows an example of an accompaniment with timing errors (orange), for the same input (green), in which, timing errors are indicated by arrows, with the correct timing indicated in dotted lines. Existing methods are prone to either pitch errors or timing errors or trades off between the two. The third objective of this work aims to develop a robust method that is not bound by this pitch-timing trade-off.

Figure 1.4: Example of Perfect Case of Harmonizer Function

Figure 1.5: Example of Pitch Errors

The novel method described here was published to the IEEE International Conference on Multimedia and Expo in 2011 [7]. Patent for parts of this work was filed

Figure 1.6: Example of Timing Errors

in the United States of America [54] and in the People’s Republic of China [61]. This is presented in detail in Chapter 4.

1.4 Organisation of Thesis

This thesis is organized as follows.

Chapter 2 first provides a psychophysical basis of harmony by presenting original work harmony psychoacoustics. Chapter 3 then presents original work in singing synthesis. With the understanding of harmony and synthesis from Chapter 2 and Chapter 3, Chapter 4 then presents our work in singing harmony synthesis. Finally, Chapter 5 closes with the conclusion and future work.

Chapter 2

The Psychoacoustics of Harmony

This chapter¹ attempts to establish psychophysical basis for harmony that is empirically verifiable. Stationary harmony refers to the harmony within a group of notes that are present concurrently at any single point in time and transitional harmony refers to how well one such group of notes transits to another. While the psychoacoustics of stationary harmony has been broadly researched through history, work in transitional harmony remains largely on the note-level. Both are addressed in this chapter with the notion of interharmonic and subharmonic modulations - which are the terms introduced to refer to different classes of modulations produced by the summations of sinusoidal components produced from different notes. Interharmonic modulations are produced by the summations of neighbouring sinusoidal components and are studied in the frequency domain while subharmonic modulations are produced by the summations of distant sinusoidal components and are studied in the time domain. Stationary and transitional subharmonic tensions are proposed as a measure of these modulations and shown to relate to perceptual tensions [1, 13–15] and resolutions [1, 16] in music. Correlations with traditional music theory and perception statistics affirm the proposed theory. Finally, the work in this chapter is used to address several fundamental questions of

¹The work in this chapter was accepted in [45].

harmony.

The chapter is organized as follows. Section 2.1 presents an introduction to harmony psychoacoustics, a background of existing work and the scope of work covered by this chapter. Section 2.2 gives an overview of the proposed approach, which is covered in detail across the chapter. Section 2.3 zooms in on the mathematical derivations of amplitude modulations as a result of sinusoidal summations by first principle. Section 2.4 focuses on interharmonic modulations in detail, while Section 2.5 then follows with subharmonic modulations. Section 2.6 will move on to present the correlation between tension in these modulations and statistics of music perception and theory. Finally, Section 2.7 will address the five fundamental questions of psychoacoustic harmony and Section 2.8 closes this chapter with the conclusion.

2.1 Background

Even though it is one of the most important components in music, and possibly the most widely studied [3], the definition of harmony differs vastly across time, genre and individuals, reflecting how little is understood about it [28, 62].

There are three aspects to the understanding of the human perception of harmony, which will be, for brevity, simply referred to as its *what*, *why* and *when*. The *what* of harmony refers to an attribution to its defining quality. Its *why* goes further to explain the means by which such a quality ascribes to consonance or dissonance. Finally, it should be recognised that the same harmony perceived as consonant in one context can be perceived as dissonant in another. This takes the *what* and *why* of stationary harmony into the context of transitional harmony – which will be referred to as the *when* of harmony. Unfortunately, existing work in the *when* of harmony has remained largely based on perceptual experiences or statistics on the note level and one of the contributions of this work is to advance this to scientific reasoning grounded on the acoustic level.

2.1.1 Existing Work

Early works effectively attributed the *what* of harmony to rational relationships [3, 63]. This ascribes a chord’s consonance to the ratio amongst its contributing string lengths (and consequently, wave periods and fundamental frequencies,) being fractional with integer numerators and denominators. A fascinating number of esteemed mathematicians, physicists and philosophers have made different contributions in this aspect. The development of the Pythagorean tuning system is commonly credited to Pythagoras in the fourth century BC [27–29]. Euclid wrote the earliest surviving record on the tuning of the monochord [64] and recorded numerous experiments on rational tuning [32]. Aristotle and Plato made various contributions to the development of ancient Grecian (ratio based) music that was later integrated into the diatonic system [32, 65]. Ptolemy developed the syntonic diatonic system as early as the second century [66]. Euler proposed a grading system of chord aesthetics based on the assertion that the notes have a least common multiple (i.e. that they are rational) [67]. Since string lengths correspond to wavelengths, which correspond to wave period, the Pythagorean school effectively attributes harmony to temporal features.

It was not until 1877 that Helmholtz pioneered the psychoacoustic approach [28, 30–32]. He was able to isolate adjacent harmonic sinusoids from different notes using specifically devised acoustic resonators and recorded how amplitude modulation that resulted from their summation grew perceptually unpleasant as their modulation frequency increased towards a certain threshold [32] - thus attributing dissonance to what he called beating frequencies and addressing the questions *what sounds bad and why*. Numerous others [1, 18, 36, 68–75] conducted further studies in this approach, while others raised several questions with Helmholtz’s theory [31, 36, 37]. For example, Plomp and Levelt [30] and Schellenberg and Trehub [38] have separately shown that consonances and dissonances are still perceived in harmonies with pure tones (tones without harmonics). Itoh [43] and Bidelman [44], amongst others, also showed that electrophysiological responses to

pure-tone intervals did not agree with Helmholtz. All in all, the Helmholtz school attributes harmony to frequency features and comprises a large part of what is referred to in this chapter as interharmonic modulations.

In 1898, a notable but short-lived [28] attempt at *what sounds good and why* was seen in Stumpf’s tonal fusion theory [76], which theorized that harmony was the effect the harmonics of its component notes fusing together to sound like a single note with a common fundamental [30, 31, 37, 76].

Because of the non-linear relationship between tonal scale and frequency, scales derived from rational lengths of a string tended to leave certain intervals more rational than others. As Helmholtz’s theory shed light on the dissonances with less rational intervals, Western music eventually adopted 12-tone equal temperament scale. This equally divides the octave on the log-frequency scale [77] such that each semitone interval is a factor of $2^{1/12}$, evenly re-distributing the dissonances to work well with different keys. Despite its late adoption, original development of this scale predates Helmholtz to the 1500s. Vincenzo Galilei (father of Galileo Galilei) made the earliest known estimate of this in the West by approximating $2^{1/12}$ with $\frac{18}{17}$, while Zhu was credited for perfecting it in the East by computing it to accurately to the 25th decimal, both in the 1580s [30]. About the earliest recorded estimate of this in the East was by He in the 5th century, whose estimate was already about as accurate as Galilei’s [33, 34]

In Rameau’s Treatise on Harmony [3], which paved the foundations of harmony in modern music theory, notes of basic chords are derived from the division of the length of a common string [35]. However, this remains disjoint with the rest of the treatise, and modern music theory remains a compilation of rules and deductions from the pattern clustering of perceptual experiences [14, 16, 73, 78–80] – addressing the questions *what sounds good and when* without a scientific reasoning of *why* [81].

More recently, several studies have found high correlations between harmony and

the periodicity of the resultant signal [82, 83]. This novel leap advances the Pythagorean school while presenting a persuasive attribute of *what sounds good and why*.

Several notable studies have been conducted that relate harmony to non-acoustic attributes such as statistics and geometry. An example is Tymoczko’s exploration of how multi-dimensional geometric patterns correlate strongly with patterns that exist in historic harmony use, addressing *what sounds good and when* [25, 26, 84]. Honingh and Bod [85] explored properties of musical scales on the Euler lattice, addressing the *what* of harmony. Numerous others such as Balzano [86, 87] and Carey [88] have worked on other mathematical relationships in harmony, addressing its *what*.

Yet others have looked towards a biological rationale towards our perception of harmony to address *what sounds good and why*. A recent example is Purves’ attribution of the effect of the tonal scale to the familiarity of excited or subdued speech [18, 89–91]. Other examples are the works of Stolzenburg [82, 92] and Langner [93, 94] in the neuronal mechanism of harmony perception.

2.1.2 Scope

The scope of this work is as follows:

- To seek a psychophysical basis for harmony to bridge both acoustic schools.
- To introduce interharmonic modulations, interharmonic tension, subharmonic modulation, stationary subharmonic tension and transitional subharmonic tension (*tension resolution*).
- To show how the empirical tensions introduced relate to perceptual tensions [1, 13–15] and resolutions [1, 16] in music by correlations with subjective consonance rankings [82] and chord use statistics [26].
- To answer the five fundamental questions of psychoacoustic harmony, which

are as follows:

1. The phenomenon that the effect of harmony is greater than that of the sum of its parts [1,2]. Mathematically, this concept may be represented by

$$\varepsilon\{x_1 + x_2 + x_3\} \gg \varepsilon\{x_1\} + \varepsilon\{x_2\} + \varepsilon\{x_3\} \quad (2.1)$$

where ε denotes the conceptual ‘*harmonious effect of*’², x_1 , x_2 and x_3 represents notes of the chord and ‘+’ denotes simultaneous presentation or cumulation.

2. The definition and explanation of stationary harmony. i.e. *what sounds good and why*, or, mathematically, to quantify $\varepsilon\{X_n\}$, where X_n represents chord n .
3. The definition and explanation of transitional harmony. i.e. *what sounds good, why and when*, or, mathematically, to quantify $\varepsilon\{X_1 \rightarrow X_2\}$, where ‘ $\rightarrow$ ’ denotes the transition from one chord to another.
4. The phenomena that:
 - (a) a chord that sounds better than another out of context can sound worse than it in context [16]. Given $\varepsilon\{X_2\} > \varepsilon\{X_3\}$ show that $\varepsilon\{X_1 \rightarrow X_2\} < \varepsilon\{X_1 \rightarrow X_3\}$
 - (b) a chord that sounds better than another in one context can sound worse than it in another context [16]. Given $\varepsilon\{X_4 \rightarrow X_2\} > \varepsilon\{X_4 \rightarrow X_3\}$ show that $\varepsilon\{X_1 \rightarrow X_2\} < \varepsilon\{X_1 \rightarrow X_3\}$
5. The phenomenon that the transition from a chord of lower perceptual tension [14] to one of higher tension can still bring about the effect of

²Consonance and dissonance are loosely defined by how agreeable notes sound together [3]. It does not take into account the fact that certain groups of notes (such as octaves) may sound too agreeable to play a part to harmony. Hence, the need to define ε as the ‘*harmonious effect of*’

tension release. Given $\varepsilon\{X_1\} < \varepsilon\{X_2\}$ show that $\varepsilon\{X_1 \rightarrow X_2\} > 0$

This concludes the scope of this chapter. In the next section is an introduction to the proposed psychophysical basis of harmony.

2.2 A Psychophysical Basis of Harmony

This section, proposes a psychophysical basis of harmony as follows:

The human perception of harmony is composed of auditory events produced by the summation of different clusters of sinusoids that are contributed by each note in the harmony. These may be classified into interharmonic and subharmonic modulations.

First-order interharmonic modulations are those produced by the interplay of adjacent sinusoids from differing notes. These are loosely categorized by the frequency of the resultant amplitude modulation into dissonant beating frequencies [32] and consonant low-frequency modulations, triggering a variety of emotions according to their modulation and carrier frequencies. Second-order interharmonic modulations are produced by the alignment of first-order ones. The consonance types of different intervals may be identified according to patterns cast by first- and second-order interharmonic modulations on their interharmonic plot.

Despite the significance of interharmonic modulations, the effect of consonances and dissonances are still experienced in the absence of harmonics with pure tone harmonies [30, 38]. This, amongst several other discrepancies [31, 36, 37, 43, 44], suggest that interharmonic modulations are not exclusive in our perception of harmony. From this, it may be deduced that subharmonic modulations also play a significant role. These are produced by the interplay of sinusoids much further apart than interharmonic modulations and they comprise of two parts. The first part is subharmonic wave formation, which occurs with the summation of component waveforms from each note to produce a waveform largely periodic to a

common subharmonic frequency. The second is subharmonic wave deformation, which is a slew-like distortion to every successive period of this composite subharmonic waveform due to the imperfect alignment of contributing wave periods. Stationary and transitional tensions may both be computed from subharmonic features which may serve as measures of stationary and transitional harmony.

To explain interharmonic and subharmonic modulations in detail and how they unify the two prevailing schools of harmony will require an approach from first principle by looking at the notes of a chord as the sum of their composite sinusoids. This is presented in the next section.

2.3 Modulations in Sinusoidal Summation

When waveform of two notes, $x_1(t)$ and $x_2(t)$, at amplitudes α and β respectively, are presented together, the result may be expressed as a sum of their composite sinusoids such that

$$\alpha x_1(t) + \beta x_2(t) = \alpha \sum_{n=1}^N q_n \cos(2\pi n f_1 t + \rho_n) + \beta \sum_{m=1}^M r_m \cos(2\pi m f_2 t + \varphi_m) \quad (2.2)$$

where, respectively, n and m represent the individual harmonics from each note, N and M represent the highest harmonics being considered, q_n and r_m represent the amplitude coefficients of each harmonic, $n f_1$ and $m f_2$ represent the frequencies of each harmonic with f_1 and f_2 representing the fundamental frequency of each note, ρ_n and φ_m represent the starting phases of each harmonic and t represents monotonically increasing time.

Isolating a single pair of adjacent sinusoids from differing notes,

$$A h_1(t) + B h_2(t) = A \cos \omega_1 t + B \cos \omega_2 t \quad (2.3)$$

where $h_1(t)$ and $h_2(t)$ are the pair of harmonics from differing notes, $A = \alpha q_n$, $B = \beta r_m$, $\omega_1 = 2\pi n f_1$, $\omega_2 = 2\pi m f_2$ and phase is ignored since the interest is in frequency.

Figure 2.1: Summation of sinusoids of equal (left) versus unequal (right) amplitudes

In the case of $A = B$, the resultant modulation is trivial and, as illustrated in the left portion of Figure 2.1 is given by the sum-to-product rule

$$\cos \omega_1 t + \cos \omega_2 t = 2 \cos \frac{\Delta \omega}{2} t \cos \bar{\omega} t \quad (2.4)$$

where $\Delta \omega / 2$ is the normalized modulating frequency and is given by

$$\frac{\Delta \omega}{2} = \frac{|\omega_1 - \omega_2|}{2} \quad (2.5)$$

$\bar{\omega}$ is the normalized carrier frequency given by

$$\bar{\omega} = \frac{\omega_1 + \omega_2}{2} \quad (2.6)$$

and the values of A and B are normalized to 1.

Figure 2.2: $A \cos \omega_1 t + B \cos \omega_2 t$ for various values of B normalized to $A = 1$.

However, in most cases, $A \neq B$ [4], and the problem becomes non-trivial, because of the change in modulation frequency as the modulating waveform no longer crosses zero. This can be seen in the right portion of Figure 2.1.

The summation of these sinusoids of unequal magnitudes³ may be approximated⁴ by

$$A \cos \omega_1 t + B \cos \omega_2 t \approx B - A + 2A \left\| \cos^{2-\frac{A}{B}} \frac{\Delta\omega}{2} t \right\| \cos \omega_c t \quad (2.7)$$

where ω_c is bounded by ω_1 and ω_2 , and is approximated to be $\bar{\omega}$ (which denormalizes to $\bar{f}$), $\left\| \cos^{2-A/B}(\Delta\omega/2)t \right\|$ denotes the magnitude value of $\cos^{2-A/B}(\Delta\omega/2)t$ signed according to its quadrant, B denotes the larger of the amplitudes and A and B are normalized to $A = 1$. When $A = B$, this simplifies to Equation 2.4. As B increases with respect to A , however, $2 - A/B$ gravitates towards 2, and

$$A \cos \omega_1 t + B \cos \omega_2 t \approx B - A + 2A \left\| \cos^2 \frac{\Delta\omega}{2} t \right\| \cos \omega_c t \approx B + A \cos \Delta\omega t \cos \omega_c t \quad (2.8)$$

For which, the modulating term is $\Delta\omega$.

It can be seen from the plots in Figure 2.2 that this estimation is good for values of B marginally larger than A to much larger than A .

For consistency, in the case of $B = 1$, the effective modulating frequency should be considered by its rectified modulating waveform which is $\Delta\omega$. In music, the interest is in its frequency in hertz. Hence, this is denormalized to

$$\Delta f = |f_1 - f_2| \quad (2.9)$$

In the next two sections, it will be seen how this is applicable not only to the summation of adjacent harmonics in interharmonic modulations but also to distant sinusoids in subharmonic modulations.

³A more precise computation of this may be made in the polar form [95]. This chapter trades off precision to express this summation as a product of two sinusoids.

⁴obtained by generalizing from the sum-to-product rule

2.4 Interharmonic Modulations

Interharmonic modulation⁵ refers to the modulation across an adjacent pair of sinusoids from different notes that fall within a certain threshold, with modulation frequency corresponding to Δf in Equation 2.9.

Figure 2.3 shows a plot of all harmonics of notes⁶ c_3 and e_3^b under 3kHz. All adjacent sinusoids less than 120Hz apart are identified in the figure, with their centre, $\bar{f}$ and modulating, Δf , frequencies labeled accordingly.

Figure 2.3: Identifying the interharmonic modulations across the notes c_3 and e_3^b

2.4.1 Beating Frequencies and Low-Frequency Modulations

Interharmonic modulations with Δf that increase towards a certain threshold are known to become increasingly dissonant, and, as coined by Helmholtz, are known as beating frequencies [32]. On the other hand, interharmonic modulations with small Δf contribute to the harmonious effect perceived in consonance [98]. Figure 2.4 illustrates this.

⁵Interharmonics [45] here, as defined earlier in this thesis, refer to the distance between adjacent harmonics, and are disambiguated from other uses [96] of the word in engineering. *Modulation* here borrows from the same use of the word in *amplitude modulation* from RF electronics [97] and refers to fluctuations in the amplitude of a carrier sinusoid.

⁶ c_3 reads *c in the third octave and plot blue in the figure*. e_3^b reads *b-flat in the third octave and plot red in the figure*

Figure 2.4: Types of Interharmonic Modulation on the scale of Δf . Low-frequency modulation correspond to consonance at small Δf , while beat frequencies [32] correspond to an unpleasant "roughness" termed dissonance at higher Δf . Modulations with near-zero Δf fall below musical significance while those with Δf pass a certain threshold are perceived as distinctly separate.

2.4.2 Perceptual Responses across the Δf - $\bar{f}$ feature Space

It is known that different combinations of notes contribute to different emotive valences [99]. This too may be decomposed into a sum of its harmonics. Hence, further to the consonances and dissonances, emotive responses may also be mapped onto the interharmonic plot. Although, as one might imagine, such responses would be different for every individual, the response for an individual can be plot as an example. Figure 2.5 shows an example of auditory responses triggered in the mind of the author when exposed to frequencies in the horizontal ($\bar{f}$) axis modulated by frequencies in the vertical (Δf) axis. The value of $\bar{f}$ is indicated in the horizontal axis in both Hz and its corresponding note names. The green regions are perceived to be pleasing, yellow as somewhat pleasing, orange as unpleasant, but not to the point of annoying, red as dissonant and black as beyond beating range. The black dots mark the locations of the thoughts or emotions labelled. This shows that interharmonic modulations bring about a large variety of thoughts or emotions. If several of these are triggered simultaneously when just one pair of notes sound simultaneously, one can imagine how ten fingers on a piano or all the instruments in an orchestra could combine several (thoughts or emotions) to paint stories on the interharmonic feature-space over time.

Figure 2.5: Even though, as one might expect, emotive responses would be different for every individual, the response for one individual can be plot as an example. The figure shows an example of auditory responses triggered in the mind of the author when exposed to pure-tone frequencies on the horizontal ($\bar{f}$) axis modulated at frequencies on the vertical (Δf) axis. Green, yellow, orange, red and black indicate pleasing, somewhat pleasing, unpleasant, dissonant and beyond beating range, respectively.

2.4.3 Intervals and Second-Order Modulations on the Δf - $\bar{f}$ feature Space

The interharmonic modulations of each interval within an octave are similarly plot in Figures 2.6, 2.7 and 2.8. However, this time, the plots are in the linear scale. Green, yellow, orange and red represent regions of different degrees of consonance or dissonance according to the same colour scheme as Figure 2.5. However, because this time both horizontal and vertical axes are in the linear scale, consonance-dissonance levels that populate the space on the non-linear plot in Figure 2.5 now populates lower right regions of these linear plots. The remaining upper left regions are then populated with dissonance levels from [30]. These colors provide a simple visual aid in the background as a reference for the dark blue dots, that each represent a modulation at their corresponding Δf and $\bar{f}$ values - which results from the summation of neighboring pairs of sinusoids (at frequencies $\bar{f} + \frac{\delta f}{2}$ and $\bar{f} - \frac{\delta f}{2}$) from the notes specified by the indicated interval. Also for reference, are the two dissonant lines that run across each plot in white, indicating the locations where the values of Δf coincide with a semitone (gentler slope) and a tone (steeper slope) of the corresponding values of $\bar{f}$ (where $\Delta f = (2^{1/12} - 1)\bar{f}$ and $\Delta f = (2^{2/12} - 1)\bar{f}$, respectively). The semitone and the tone are regarded as the most dissonant intervals up to halfway in either direction around the cyclic chroma [30, 72, 91].

The plots of perfect consonances are presented in Figure 2.6. These intervals are described with a bit of a dilemma in classical music theory [100]. They may be described as so consonant that they sound almost like a single note [32]. As such, their use contributes in a limited way to harmony [68]. For example, the use of perfect fifths is forbidden in parallel motion, the use of parallel fourths is likewise restricted, and octaves are regarded as the same note in a different register [16].

The interharmonic plot reveals the perceived traits of each category of intervals in a way that explains why they sound the way they do, and in a way music theory alone has never been able to. As shown in the figure, the constellations formed by

Figure 2.6: Interharmonic plots for perfectly consonant intervals within an octave.

interharmonic modulations of perfect intervals line up almost horizontally⁷. Since each blue dot on a horizontal has the same Δf , this means that they modulate synchronously and may be perceived collectively as a single modulation. This may be interpreted as fewer modulating micro-events taking place, making them less interesting than regular consonance.

Dissonant intervals are presented in Figure 2.7. As can be seen in the figure these intervals have dots that fall mostly within the central dissonant region and line up along the two dissonant lines. Evenly spaced dots along a line that pass through

⁷These dots are aligned to form up parallel to the horizontal axis.

Figure 2.7: Interharmonic plots for imperfectly consonant intervals in within an octave.

the origin also reveal that their Δf share a harmonic relationship. This has a similar redundant effect, to that of the synchronous modulation which was described with perfect consonances.

Consonances that properly contribute to harmony are called imperfect consonances [100] and are presented in Figure 2.7. As can be seen in the figure, imperfectly consonant intervals have dots well distributed. This may be interpreted as erratic modulations that create a continuous stream of unpredictable events to stimulate aural attention, and thus, interest.

Figure 2.8: Interharmonic plots for dissonant intervals in within an octave.

A lot of work has already been done on interharmonics since Helmholtz [30, 70–72, 75]. While the main focus of this work is not interharmonics, one purpose of this

section is, nevertheless, to provide sufficient background to complete our theory of how the human experience of stationary harmony is based around modulations of both interharmonic and subharmonic nature. From the interharmonic plots in Figures 2.6 through 2.8, a simple predictor of dissonance may be identified to be

$$C(\Delta\hat{f}) = C\left(\frac{\Delta f}{\bar{f}}\right) = \sum_{i=1}^n \left[2^{\frac{r_{lower}}{12}} - 1 < \frac{\Delta f_i}{\bar{f}_i} < 2^{\frac{r_{upper}}{12}} - 1 \right] \quad (2.10)$$

where $\Delta\hat{f}$ represents $\frac{\Delta f}{\bar{f}}$, $C(\Delta\hat{f})$ or $C(\frac{\Delta f}{\bar{f}})$ refers to the number of interharmonic modulations that fall within the central region of dissonance region, $\sum_{i=1}^n[\]$ is the Iversion bracket notation for *number of*, i iterates through all interharmonic modulations on the plot, n is the total number of modulations considered, and Δf_i and $\bar{f}_i$ refer to the pair of Δf and $\bar{f}$ that describe the i^{th} interharmonic modulation respectively, and r_{lower} and r_{upper} define the lower and upper boundaries of the region on the interharmonic plot, respectively.

In this section, it has been seen how interharmonic modulations are significant to our perception of consonances, dissonances, and emotive response in music. When listening to a duet of instruments with no overtones such as a sinewave theremin [101] or a very pure musical saw, it is realized that consonance and dissonance and a variety of emotions remain present even in harmony without harmonics (i.e. across a pair of distant fundamental frequencies alone). This is one amongst the several different ways [5, 30, 31, 36, 43, 102] it can be deduced that interharmonic modulations cannot be the only determinant of the human perception of harmony. This leads to the following hypothesis on the significance of subharmonic modulations [45].

2.5 Subharmonic Modulations

Apart from the modulations that arise from the summation of adjacent harmonic sinusoids across differing notes, it can (as explained above) be deduced that another category of modulations is significant to the human perception of harmony. These are introduced in this section as subharmonic modulations [45]. There are two levels of subharmonic modulations, which will be referred to here as dub subharmonic wave formation and subharmonic wave deformation [45]. In this section, it will be shown how these are significant to the human perception of not only stationary harmony, but also transitional harmony.

Figure 2.9: Subharmonic wave formation and wave deformation in the C and Cm^7 chords.

Figure 2.9⁸ shows the waveforms of a C chord and a Cm^7 chord composed of the fundamental sinusoids of each composite note. Each sinusoid can be shown to start

⁸Supplementary video S1 in Appendix B animates this figure to show how subharmonic wave deformation takes place differently across high and low tension chords over time.

at phase zero since we are only interested in its period. Only the fundament needs to be considered for the same reason. In both cases, the waveform resultant of this summation repeats at a frequency approximately subharmonic to all its composite waveforms. In the figure, its period is marked T_{sub} . This is here referred to as *subharmonic wave formation* and T_{sub} is termed a *common subharmonic* to all its composite waveforms.

Chords are groups of notes that are presented simultaneously as explained in detail in appendix A. In music theory, the C chord is regarded consonant while the Cm^7 contains some dissonances. This can be visualized in the plot. In the case of the C chord, as shown in the figure, each composite sinusoid crosses zero at nearly the same point around $t = T_{sub}$. As marked in the Figure, Δt (which is the difference between the first and the last negative-to-positive zero-crossing around the $t = T_{sub}$ region) is small. However, in the case of the Cm^7 chord, Δt is much larger. One can imagine that each successive period of the resultant waveform looks less and less like the first as it gets more and more deformed by a slew-like distortion. This happens slowly for the C chord because of the small Δt but faster for the Cm^7 because of the large Δt . This is here referred to as *subharmonic wave deformation*. Supplementary Video S1 compares subharmonic wave deformation in a low tension C chord to that in a high tension Cm^7 chord.

Recalling the wave equation, Equation 2.3, $A \cos \omega_1 t + B \cos \omega_2 t$, or $A \cos 2\pi f_1 t + B \cos 2\pi f_2 t$ can be re-written as

$$A \cos 2\pi f_1 t + B \cos 2\pi f_2 t = A \cos 2\pi (k_1 f_{sub} + \Delta f_1) t + B \cos 2\pi (k_2 f_{sub} + \Delta f_2) t \quad (2.11)$$

where frequency, f_{sub} , is an approximate common factor of f_1 and f_2 , k_1 and k_2 are integer multipliers, and Δf_1 and Δf_2 are small values that balance the equation by making up for the discrepancies that arise with finding a common factor.

In the equation, two fundamental frequencies f_1 and f_2 are described as the multiple of a lower subharmonic frequency that is common to them (f_{sub}). This is termed to their *common subharmonic* [45].

Since all harmonics are multiples of their fundamental [63], a subharmonic to any fundamental would inherently be subharmonic to all its harmonics. For this reason, only the fundamental of each note needs to be considered.

Since harmony in music is commonly composed of more than just two notes, this can be generalized to describe fundamentals and common subharmonics from any number of notes to get

$$\begin{aligned} \sum_{i=1}^N A_i \cos 2\pi f_i t = & A_1 \cos 2\pi(k_1 f_{sub} + \Delta f_1)t + A_2 \cos 2\pi(k_2 f_{sub} + \Delta f_2)t + \dots \\ & + A_i \cos 2\pi(k_i f_{sub} + \Delta f_i)t + \dots + A_N \cos 2\pi(k_N f_{sub} + \Delta f_N)t \end{aligned} \quad (2.12)$$

where N is the number of notes in the chord, i cycles through each of them, and A_i is the amplitude coefficient of note i .

Beyond this point, it would be easier to visualize subharmonics in the time domain. With the fundamental frequency of note i given by

$$f_i = k_i f_{sub} + \Delta f_i \quad (2.13)$$

the fundamental period of each note i is then

$$t_i = \frac{1}{f_i} = \frac{1}{k_i f_{sub} + \Delta f_i} \quad (2.14)$$

where t_i is the fundamental period of the note.

Hence, the period of any common subharmonic can be expressed as $k_i t_i$. The non-integral discrepancies may then be compensated in period rather than in frequency. Doing so obtains

$$T_{sub} = \frac{1}{f_{sub}} = k_i t_i + \Delta t_i \quad (2.15)$$

for all i since T_{sub} is common across all notes in the chord⁹, and where T_{sub} is the common subharmonic wave period (simply referred to as the *common subharmonic* in the rest of the chapter) of the chord. What carries over as $k_i t_i$ is essentially just the k^{th} subharmonic of note i which lies in the region of T_{sub} . Since this is true for all pairs of k_i and t_i across all values of i when they are balanced by appropriate Δt_i , the i will be dropped on the left hand side of the equation.

Although the common subharmonic was introduced as the period between primary zero crossings as in Figure 2.9, it shall be, for computational simplicity, re-defined as the mean of $k_i t_i$ across all notes of the chord. Hence,

$$T_{sub} = \overline{k_i t_i} \quad (2.16)$$

Figure 2.10 shows how only the periods of each subharmonic in the C Major chord from Figure 2.9 may be plot. The left column first shows how the period of each subharmonic of c_3 may be plotted in red. The right column then extends this to every remaining note in the chord, with orange, yellow and blue for the notes e_3 , g_3 , and c_4 , respectively. It may be seen in the right column that a subharmonic period from every note in the chord nearly coincides at around 30ms. Hence, it is referred to as its *common* subharmonic, T_{sub} , as defined in equation 2.16.

Having reduced the waveform plot to subharmonic periods in the vertical axis, time spanned by each subharmonic can be represented in the horizontal axis. This will be done in the next section in the subharmonic plot of an actual song.

⁹This will become apparent with the example given in Figure 2.10 in the next paragraph

Figure 2.10: Plotting subharmonic wave periods of the C Major chord.

2.5.1 Subharmonic Modulations in Stationary Harmony

Figure 2.11¹⁰ shows an example of a subharmonic plot. In the horizontal axis is time in bars and in the vertical axis is the subharmonic wave period in milliseconds. Note that the subharmonic axis runs top-down to put shorter wave periods at the top because they correspond to higher frequencies. Larger wave periods, which correspond with lower frequencies sit, conversely, at the bottom. The tails that run horizontally represent the span of time covered by each note. Subharmonics

¹⁰In the interest of visiting all common chords of the key, *Em* is used in the 7th bar instead of *G*, which already occurs in the 5th bar. Considering the fact that this example is not used for transitional harmony, all chords are presented in its root position at the expense of introducing parallel 5ths in the interest of normalization for fairer comparison.

Figure 2.11: Subharmonic plot of the opening stanza of Pachelbel's Canon in D. Subharmonics are computed from the fundamental wave periods of each note and plot with T_{sub} in milliseconds on the vertical axis and time in bars on the horizontal axis. Subharmonics are plot in the colour matching their corresponding notes on the music score. The subharmonic tensions of each chord, Δt , are marked out on the plot with white arrows. Significant wave periods, along with common subharmonic periods, T_{sub} , are marked against the vertical axis on the right.

are colored to match their corresponding notes on the music score. For example, in the first bar, all subharmonics of $f^\#_5$ are marked out in red, followed by d_5 in orange, a_4 in yellow, d_4 in green, a_3 in blue and d_3 in purple. The musical score runs in parallel at the bottom of the plot as reference. Once again, all plots and

computations in our examples assume equal temperament unless stated otherwise. This example shows the opening stanza of Pachelbel’s Canon in D [103] and focuses on stationary harmony, leaving transitional harmony to a later example.

Subharmonics: For every bar, the dashes that flush with the reference point at 0ms mark $0 \times t_0$. Carrying on top down for each bar in accordance to colour, subharmonics are marked at $1 \times t_0$, $2 \times t_0$, $3 \times t_0$, $4 \times t_0$, etc.

Notes and Melody Line: Since the topmost dash of each colour for every bar below the 0ms reference represents $1 \times t_0$, they relate to the fundamental period of each note; of these, the topmost ones of every bar across all colours mark the melody line, $f_5^\sharp - e_5 - d_5 - c_5^\sharp - b_4 - a_4 - b_4 - c_5^\sharp$. (They are red in this particular example.) Hence, it is easy to interpret the melody line in a subharmonic plot. The periods, t_i , of each note of the melody are marked against the vertical axis in milliseconds as well as their common note names.

Chords and Coincidence: Common subharmonics may be visualized in regions with the (approximate) coincidence of dashes of every color. Again, the common subharmonics (T_{sub}) of each chord in the stanza are marked out against the vertical axis in both milliseconds and their respective chord names.

Key: Every note of the key share a common subharmonic. Hence, it is possible to identify the key of a songs by its common subharmonic, assuming all notes used in the songs stays within its key. The common subharmonic associated with the key of this song is marked out much further down the plot.

Stationary Tension: Most of the time, contributing subharmonics from different notes are not precisely coincident. Major chords have better coincidence than minor chords, and triads coincide better than sevenths and extended chords¹¹. With subharmonic modulations, perceptual tension arises with the non-coincidence

¹¹Major and minor chords are fundamental to music harmony. Major chords are known to be the most consonant triads (three-note chords), followed by minor chords. Sevenths (four note chords) are naturally more dissonant than basic major and minor chords.

of common subharmonics. Non-coincidence is measured by an overall Δt as reflected in Figures 2.9 and 2.11. This will be known as its (stationary) subharmonic tension [45].

This Δt is given by the difference between the largest and smallest subharmonics in the chord that coincide around T_{sub} .

$$\Delta t = t_{max} - t_{min} \quad (2.17)$$

where t_{max} and t_{min} denote the largest and the smallest subharmonics in the chord that (nearly) coincide around T_{sub} - i.e. the maximum and minimum values of $k_i t_i$ respectively.

Δt and T_{sub} are the primary features of stationary tension. Δt may also be normalized by expressing it in duty cycle by taking

$$\Delta \hat{t} = \frac{\Delta t}{T_{sub}} \quad (2.18)$$

From Figure 2.4, recall that dissonances increased and decreased with interharmonic modulation frequency while consonances behaved inversely. This happens only within a certain range. In the figure, when interharmonic modulation frequency is approximately zero, it falls below musical significance.

Subharmonic tension behaves similarly. Figure 2.12 describes different types of harmony on the subharmonic tension scale. As can be seen in the figure, our response to subharmonic tension is likewise. Perceived dissonances increase and decrease with subharmonic tension while perceived consonances behave inversely within common range. Mathematically,

$$\varepsilon\{X\} \propto \frac{1}{\Delta \hat{t}_X} \quad (2.19)$$

where $\varepsilon\{X\}$ is the *harmonious effect of chord X* and $\Delta \hat{t}_X$ is its stationary subhar-

monic tension (its Δt).

However, as described in the figure, modulations from subharmonic tension fall below musical significance, the effect of harmony drops to zero as modulations from subharmonic tension fall below musical significance. Hence, where $\Delta \hat{t}_{threshold}$ is the said threshold of musical significance, as $\Delta \hat{t} < \Delta \hat{t}_{threshold}$,

$$\lim_{\Delta \hat{t} \rightarrow 0} \varepsilon \{X\} = 0 \quad (2.20)$$

Figure 2.12: An illustration of the effect of harmony, $\varepsilon\{X\}$, on the scale of subharmonic tension, $\Delta \hat{t}$, according to the proposition of $\Delta \hat{t}$ as a measure of tension and dissonance.

Back to the theory [45], perceptual tensions and consonances are experienced in slew-like modulations of the waveform at common subharmonic locations. (This is the effect of periodically changing phase relationships amongst the contributing waveforms, for which, Δt is a measure.) While there may be several common subharmonics for every chord within reasonable range, it is theorized that the ears identify most with the shortest few [45]. Subharmonic consonances are described by gentler modulations (small Δt) at the shortest common subharmonic locations (short T_{sub}), while subharmonic dissonances are described by more turbulent ones (associated with absence of small Δt at short T_{sub}) [45].

The sensation of a chord can be highly complex, with different tensions and consonances perceived simultaneously -- an experience inadequately represented if by a single term for dissonance. Attempting to rate every chord by its dissonance level alone can be compared to rating every variety of chocolate in a candy store by how sweet or bitter it is. The advantage of Δt , as opposed to existing correlates of harmony [28, 31, 82, 91], is the way it explains abstract notions of perceptual tensions and consonances by ascribing them to regions across the subharmonic spectrum with a strong sense of attribution or identification. While, for purpose of illustration, Figures 2.10 and 2.11 have shown examples where a modal T_{sub} (shortest T_{sub} with smallest Δt) are easiest to identify, it is observed that with complex chords with ambiguous T_{sub} (where it is difficult to identify a modal), the ears can identify with several common subharmonics simultaneously [45]. In other words, indeterminate cases could possibly arise with particularly discordant harmonies without small Δt at short T_{sub} . However, for programmatic analysis of a large number of chords, it is, nevertheless, useful to have a single term to represent the overall dissonance of each chord. Hence, in such cases

$$\widetilde{\Delta t} = n \left(\frac{1}{\sum_{n:m} \frac{1}{T_{sub,j} (\Delta t_j)^c}} \right)^{\frac{1}{c}} \quad (2.21)$$

where a single term, $\widetilde{\Delta t}$, represents the overall subharmonic tension, $T_{sub,j}$ and Δt_j refers to individual candidates of T_{sub} and Δt with j iterating through each candidate pair, c is the pre-emphasis (while $\frac{1}{c}$ serves as “post de-emphasis”), and $\Sigma_{n:m}$ denotes summing over the m smallest values out of a range of n values considered. In our work, m is always chosen to be half of n unless stated otherwise. Note $T_{sub,j}$ here serves as a weighting factor to weight down higher subharmonics, which, as aforementioned in our theory, are less significant. Inverting before (and rectifying after) summation mimics our hearing by allowing smaller values of Δt_j to contribute better towards a smaller $\widetilde{\Delta t}$.

In the next section, how representative $\widetilde{\Delta t}$ is of stationary harmony will be evaluated. However, before that, the next subsection will present how resolutions [1, 16]

in transitional harmony can be visualized in subharmonic modulations.

2.5.2 Subharmonic Modulations in Transitional Harmony

While stationary harmony studies how a chord sounds on its own, transitional harmony deals with how chords transit from one to another [45]. It is remarkable how a low tension (consonant) chord can transit to a high tension (dissonant) one yet still bring about the perceptual effect of tension release (resolution) [1, 16]. From this it may be deduced that transitional harmony stands largely independent of stationary harmony — even though both are considered when assigning harmony in composition. Even though numerous studies have been conducted on stationary harmony from the psychoacoustic approach, work on transitional harmony remain primarily non-psychophysical.

Traditional classical music theory uses the term resolution to describe the perception of tension released when a chord is suitably followed by another chord [1]. With subharmonic modulation, It is proposed that these abstract perceptions of tensions released may be identified and quantified in the perceived trajectories of subharmonics as one chord progresses to the next. Figure 2.13 illustrates this.

Figure 2.13 shows the opening line of Beethoven’s Moonlight Sonata [104]. Before analysis, one should note that unlike in Pachelbel’s Canon, the use of arpeggios (or broken chords) means that notes contributing to the harmony may not necessarily start at the same time, but, when the sustain pedal on the piano is applied, sustain and overlap until the end of each bar. The names of the chords formed by the notes are labelled along the top of the score to aid the reader in this analysis. Another thing to note would be the fact that this piece maintains a strong sense of *voice-leading* [105], which means that each note from a chord has strong progressive associations with a note from the previous and another from the succeeding chord. The subharmonics of all notes that are associated in this way (i.e. of the same voicing) across the song are coded with the same colour to aid the reader in this analysis. For example, all notes in red on the music score represent the bass (lowest) notes throughout the song, and every subharmonic of these notes are

Figure 2.13: Subharmonic plot of the opening line of Beethoven's Moonlight Sonata with T_{sub} in the vertical axis and time in bars in the horizontal axis. Subharmonics are coloured to their corresponding notes on the music score. Names of relevant notes are marked out on the left, at T_{sub} values corresponding to their wave period. The region of each transition is numbered in white. Coloured arrows follow voice leading along the notes across chord changes.

portrayed in red.

It is theorized that in chord transitions, every subharmonic ($k_i t_i$) that (nearly) coincides around the common subharmonic (T_{sub}) of a succeeding chord is perceived to transit from the nearest corresponding (i.e. of the same voicing) subharmonics in the preceding chord. These transitions are marked out by the arrows in Figure 2.13, which are coloured according to the notes they are associated with. Arrows

are usually convergent because the subharmonics of the succeeding chord always identify with the common subharmonic whereas those of the preceding chord usually do not.

The central hypothesis of transitional subharmonic theory is that perceptual tension resolution so often described in traditional music theory but never physically identified in acoustics lies in the degree of convergence seen here.

Assuming transition to be abrupt (since notes do not commonly glide from one pitch to another in music) a Δt may be computed for the preceding subharmonics and a Δt for the succeeding common subharmonic simply measure this degree of convergence as the difference between the two. As such,

$$\Delta\Delta t = \Delta t_p - \Delta t_s \quad (2.22)$$

where Δt_s refers to the Δt of the succeeding chord and Δt_p refers to the Δt of its nearest preceding subharmonics.

This can be normalized by dividing by T_{sub} such that

$$\Delta\Delta\hat{t} = \frac{\Delta t_p - \Delta t_s}{T_{sub}} \quad (2.23)$$

where $\Delta\Delta\hat{t}$ denotes normalized $\Delta\Delta t$. The T_{sub} of the succeeding chord is used in this normalization.

$\Delta\Delta t$ is, thus, a quantification of the tension, Δt , released over the transition at the wave period of the succeeding common subharmonic.

By our theory, tension resolution [16] is perceived in the release of this tension across each transition. Thus, mathematically,

$$\varepsilon \{X_1 \rightarrow X_2\} \propto \Delta\Delta\hat{t}_{X_1 \rightarrow X_2} \quad (2.24)$$

where $\varepsilon \{ \rightarrow \}$ denotes the perceptual resolving effect of tension release, $\Delta\Delta\hat{t}_{X_1 \rightarrow X_2}$ denotes the $\Delta\Delta\hat{t}$ of the transition of chord X_1 to chord X_2 .

Since resolution (tension release) in harmony progression [16] is perceived in the convergence of $\Delta\hat{t}$, complication (build-up of tension or negative resolution) is, thus, seen in its divergence, where $\Delta\Delta\hat{t} < 0$ and $\varepsilon \{X_1 \rightarrow X_2\}$ is negative.

Three possibilities arise when looking at T_{sub} and Δt from this perspective. As illustrated in Figure 2.14, these are:

1. **Resolution**, also called *tension release* [45]: This is the most common occurrence and occurs with the *convergence* of Δt and a positive $\Delta\Delta t$. The larger the $\Delta\Delta t$, the larger the perceptual tension release.
2. **Complication**, also called *tension buildup* [45]: This is the least common occurrence and occurs with the *divergence* of Δt and a negative $\Delta\Delta t$. *Just as negative aesthetics may be used expressively in a painting, it may similarly be used in music [106].* The larger the magnitude of $-\Delta\Delta t$, the larger the perceptual tension buildup. Complications usually only occur when the preceding T_{sub} is equal or nearly equal to the succeeding T_{sub} .¹²
3. **Excursion** [45]: Because of the circular nature of the musical chroma, the preceding T_{sub} and the succeeding T_{sub} may be computed to differ by up to 6 semitones in either direction. When the difference is 1 or 2 semitones, the collective effect of melodic movement (i.e. melody) across each note of the chord can overpower the effect of harmony. In such cases, our ears are persuaded to identify Δt_p with $[k_i t_i]_{max} - [k_i t_i]_{min}$ of the nearest preceding T_{sub} . When this happens, $[k_i t_i]_{max}$ and $[k_i t_i]_{min}$ (and each $[k_i t_i]$ between them) move in the same direction, hence, *neither convergence nor divergence*

¹²Musically speaking, it usually occurs when a simpler chord is followed by a more complex chord of the same root.

is perceived [45]. There are 2 such cases:

- (a) **Escalation** [45]: This occurs when each $[k_i t_i]$ shortens simultaneously, T_{sub} shortens by a factor equivalent to 1 or 2 semitones ($2^{1/12}$ to $2^{2/12}$ times) and f_{sub} rises, producing the uplifting effect of melodies rising by 1 or 2 semitones.
- (b) **Descent** [45]: This occurs when each $[k_i t_i]$ lengthens simultaneously, T_{sub} lengthens by a factor equivalent to 1 or 2 semitones and f_{sub} falls, producing the detrimental effect of melodies falling by 1 or 2 semitones.

Figure 2.14: Trajectories of $k_i t_i$ in different states of tension development (states of convergence).

It is fascinating to note how the perceptual build-up and resolution (i.e. transition) of tension that is so often described in music [1, 14, 15, 107] but never identifiable with a psychophysical attribute may here be visualized in the convergence and divergence of common subharmonics. Figure 2.14 further illustrates how $k_i t_i$ trajectories reflect the development of tension build-up and release. Additionally, trajectories for excursions are illustrated in the same figure.

Returning to Figure 2.13, the transitions between each chord are labeled 1 to 7 in the figure, and correspond to 1 to 7 as follows ¹³:

¹³This portion is intended for readers with prerequisites from Appendix A and music theory [3]. Other readers may skip the rest of this subsection to view Supplementary videos S2 and S3 in Appendix B instead.

1. The song starts off with a $C^\sharp m$ chord. Hence, the common subharmonic is observed around a wave period of $c^\sharp$ wave periods. Our ears adhere especially to the shortest one, which is at $c_2^\sharp$. Large Δt is attributed to the complex tensions within of a minor chord. At the region marked 1, this transits to a $C^\sharp m/B$ chord. The tension built up with the divergence of Δt may be visualized in the divergence of the arrows in the figure. (Of which, the dotted ones across the plot are used to indicate the continuation of subharmonics – i.e. $k_i t_i$, that do not change.) Both perceptually in music [14] and acoustically, as defined above, this translates to a further complication to the minor tension in the music. The tension built up with the divergence of Δt may be visualized in the arrows. Of which, the dotted ones across the plot are used to indicate the continuation of subharmonics, $k_i t_i$, that do not change in length.
2. At region 2, there is a convergence to a momentary (half-bar) low-tension A chord. The uplifting effect of a large tension release, $\Delta \hat{\Delta} t \gg 0$, is counter-balanced by the detrimental effect of a falling melodic sequence (lengthening T_{sub}), adding to the complexity of the song.
3. At region 3, A transits to a $D/F^\sharp$, which is a Neapolitan 2nd. The low $F^\sharp$ bass extends over 2 octaves below the treble notes, putting a strong T_{sub} at a non-root period $f_1^\sharp$ and creating an amount of stationary tension that is unusual for a major chord. (In such cases, there is usually another common subharmonic with lower Δt but at a wave period corresponding to a root at a much larger T_{sub} .)
4. At region 4, the Neapolitan 2nd resolves to the Dominant 7th, marked $G^\sharp 7$ in the figure, with a large perceptual resolution that is signature to $^b I I^6 - V^7$ transitions in music [16]. This large tension release is visualized in the subharmonic plot as indicated by the arrows.
5. Musically, the Dominant 7th typically plays the musical role of building an anticipation for the upcoming return to the Tonic, and Beethoven enhanced this function particularly well with a double suspension in regions 5a to 5c. The subharmonic plot gives tangibility to the perceptual details with

suspension-resolution long theorized about in music that can now be affirmed with visualization.

- (a) At region 5a, the transition from the $G^{\sharp 7}$ progresses to what is labeled $C^{\sharp}m$. However, this $C^{\sharp}m$ is functionally still a $G^{\sharp}$ with a double suspension of the 3rd ($b^{\sharp}$) and 5th ($d^{\sharp}$) to a 4th($c^{\sharp}$) and 6th(e) respectively. The perceptual complication that arises with this transition can be visualized in the subharmonic plot as indicated by the divergence of the green and cyan arrows respectively. The deviation of the suspended notes from the primary triad is visualized as to a deviation of their k_it_i from T_{sub} .
 - (b) At region 5b, the tension resolution with the 6th being resolved back down to the 5th can be visualized in the subharmonic plot by its k_it_i resolving back to T_{sub} as indicated by the convergent cyan arrow. The continuation of the suspended 4th is visualized in the dotted green arrow.
 - (c) At region 5c, the tension resolution with the 4th being resolved back down to the 3rd can be visualized in the subharmonic plot by its k_it_i resolving back to T_{sub} as indicated by the solid green arrow. In preparation for a major resolution back to the upcoming tonic, Beethoven's touch of genius couples this resolution with the simultaneous complication (tension increase) of the 7th at this point. This is visualized in the deviation of its k_it_i away from T_{sub} as indicated by the divergent solid yellow arrow.
6. At region 6, the Dominant 7th is resolved back to the Tonic with a tension release unique to $V_7 - i$ (and $V_7 - I$) cadences that is so immense that it is used as the final resolution in the majority of musical passages. This immense perceptual release of tension, too, is identifiable in the subharmonic plot. From the figure, it may be seen that the common subharmonic, T_{sub} , of $C^{\sharp}m$ (located at the period of $c_1^{\sharp}$), lies right in the middle of two common subharmonics of $G^{\sharp 7}$ (located at the periods $g_1^{\sharp}$ and $g_0^{\sharp}$). This unique subharmonic behaviour allows our ears to identify with both k_it_i for the preceding Δt – making $\Delta \hat{t}_p$ significantly larger than its $\Delta \hat{t}_s$. Its staggering convergence

produces an immense sense of tension resolution with this transition.

7. A final landmark that is interesting to note is at region 7, where the triad in the treble flips from the 1st inversion to the 2nd inversion while the chord remains unchanged. Notice that this brings about no change to both T_{sub} and $\Delta\hat{t}$ while $\Delta\Delta\hat{t} = 0$. This shows how subharmonic analysis agrees with music theory where, despite the change of notes, harmony remains the same at this point.

In this section, it was seen how, even in the context of transitional harmony, perceptual tensions and resolutions in a song may be visualized in its subharmonic modulation. The next section move on to see how well numerical values computed with such modulations verify against listening tests and chord use statistics.

2.6 Experiment and Results

For both stationary and transitional harmony, tensions computed from our models show strong correlations with consonance rankings and historical chord-use statistics. Table 2.1 tabulates a summary of the results of our experiment.

Stationary Harmony		Transitional Harmony		
Dyads/Intervals (2 notes)	Triads (3 notes)	Triads & Tetrads (3 or 4 notes)		
		All Transitions	All Transitions <i>Excl. Comp.</i>	Resolutions
r= 0.922	r=0.907	r= 0.903	r= 0.970	r= 0.996
p=0.0001	p=0.0000	p=0.0000	p=0.0000	p=0.0000

Table 2.1: Summary of correlations with consonance rankings and historical chord use.

Each of these results are explained in detail in the following subsections.

2.6.1 Stationary Harmony

For stationary harmony, the overall tension of a chord is taken to be

$$T_{\Delta f|\Delta t} = w_i T_{\Delta f} + w_s T_{\Delta t} \quad (2.25)$$

where $T_{\Delta f|\Delta t}$ is overall tension, $T_{\Delta f}$ and $T_{\Delta t}$ are taken to represent the tensions contributed by interharmonic and subharmonic modulations respectively (normalized by linearly scaling to fit between 0 and 1) and w_{inter} and w_{sub} are summing coefficients that add up to 1. For which, 0.61 and 0.39 are used, respectively, in the experiment.

A simple estimate of $T_{\Delta f}$ is used, such that

$$T_{\Delta f} = C_1(\Delta \hat{f}) + C_2(\Delta \hat{f}) \quad (2.26)$$

where $C_1(\Delta \hat{f})$ and $C_2(\Delta \hat{f})$ are a tally of interharmonic modulations (given by Equation 2.10). By visual inspection of the interharmonic plot, the two regions of dissonance are defined by r_{lower} values of 0.95 and 1.5 and r_{upper} value of 1.1 and 2.8 respectively.

For $T_{\Delta t}$, $(\widetilde{\Delta t})^2$ is used, where $\widetilde{\Delta t}$ is given by Equation 2.21 pre-emphasized with $c = 2.1$ across a range of $n = 5$. (A preemphasis of just over 2 provided the sufficient discrimination without driving data into saturation. A broad range of n -values are suitable but it was settled on a smaller value of 5 for computational simplicity.)

$$T_{\Delta t} = (\widetilde{\Delta t})^2 \quad (2.27)$$

Numerous previous authors have performed notable work for stationary harmony both within and outside the psychophysical context [1, 4, 30–32, 72, 74, 75, 82, 83, 90, 92, 108] For dyads (intervals, or two-note chords) and triads (three-note chords), the pre-collated information in Tables 2 through 5 from Stolzenburg [82] are used for comparison. Dyads (intervals) are compared against the results of an average across 7 notable studies collated by Schwartz et al [91] on a ranking of 12 chords. Stolzenburg adds the unison to Schwartz’s list, which he reasonably assumes to be

Method	Dyads	Triads
Method (Tuning System) <i>Publication</i>	r (p)	r (p)
$T_{\Delta f \Delta t}$ (Equal Temperament) <i>Proposed</i>	0.922 (0.0000)	0.907 (0.0000)
Log Periodicity (Just) <i>Stolzenburg [82]</i>	0.982 (0.0000)	0.831 (0.0002)
Rel. Periodicity (Just) <i>Stolzenburg [82]</i>	0.982 (0.0000)	0.846 (0.0001)
Log Periodicity (Rational) <i>Stolzenburg [82]</i>	0.936 (0.0000)	0.813 (0.0004)
Rel. Periodicity (Rational) <i>Stolzenburg [82]</i>	0.936 (0.0000)	0.808 (0.0004)
Rel. Periodicity (Pythagorean) <i>Stolzenburg [82]</i>	0.817 (0.0003)	-
Rel. Periodicity (Kirnberger III) <i>Stolzenburg [82]</i>	0.796 (0.0006)	-
Ω measure <i>Stolzenburg [92]</i>	0.886 (0.0000)	-
Consonance Raw Value/Degree [†] <i>Fotlyn*</i>	0.978 (0.0000)	0.826 (0.0016)
Dual Process <i>Johnson-Laird et al* [31]</i>	-	0.791 (0.0006)
Percentage Similarity <i>Gill & Purves* [90]</i>	0.977 (0.0000)	0.802 (0.0005)
Instability <i>Cook & Fujisawa* [1]</i>	-	0.698 (0.0040)
Tension <i>Cook & Fujisawa* [1]</i>	-	0.599 (0.0153)
Sonance Factor <i>Hofmann-Engl* [83]</i>	0.982 (0.0000)	0.434 (0.0692)
Generalized Coincidence <i>Ebeling* [108]</i>	0.841 (0.0002)	-
Consonance Value <i>Brefeld*</i>	0.940 (0.0000)	0.755 (0.0014)
Dissonance Curve <i>Sethares* [72]</i>	0.905 (0.0000)	0.723 (0.0026)
Pure Tonalness <i>Parncutt* [4]</i>	0.938 (0.0000)	0.675 (0.0162)
Complex Tonalness <i>Parncutt* [4]</i>	0.738 (0.0020)	-
Roughness <i>Hutchinson & Knopoff* [74]</i>	0.967 (0.0000)	0.352 (0.1193)
Sensory Dissonance <i>Kameoka & Kuriyagawa* [75]</i>	-	0.607 (0.0139)
Critical Bandwidth <i>Plomp and Levelt* [30]</i>	-	0.570 (0.0210)
Temporal Dissonance <i>Helmholtz* [32]</i>	-	0.503 (0.0399)
Gradus Suavitatis <i>Euler*</i>	0.941 (0.0000)	0.690 (0.0045)

*as cited in Stolzenburg 2015 [82]
[†]Raw Value was used for Dyads and Degree was used for Triads

Table 2.2: Proposed and Existing Correlates of Stationary Harmony.

the most consonant and has been included accordingly. Triads are compared to results from Johnson-Laird’s [31] experiment as cited in Stolzenburg [92]. For consistency with Stolzenburg’s statistics in the comparison, these were first converted to ordinal rankings before computing the correlation as practised in Stolzenburg [82]. Table 2.2 lists our correlations for dyads and triads in stationary harmony against known relevant work as taken from Stolzenburg’s [82]. A detailed tabulation of all available values for each chord is provided in the appendix.

2.6.2 Transitional Harmony

For transitional harmony, $\Delta\Delta t$ from Equation 2.22 is suitable for hand-computation of transitional harmony across individual locations of succeeding common subharmonics, Δt_s . While this is advantageous for visualizing individual complications and resolutions at multiple locations across the tensional soundscape, it requires manual identification of a modal Δt_s for every transition which can be ambiguous for particularly discordant harmonies. For a consistent programmatic approach with larger datasets, the *overall* $\Delta\Delta t$ of a transition is taken to be

$$\widetilde{\Delta\Delta t} = \left(\frac{1}{n} \sum_{j=1, \forall \Delta t_{s,j} < \frac{\Delta T_{sub}}{2}}^N \left(\frac{\Delta\Delta \hat{t}_j}{\Delta t_{s,j} T_{sub,j}} \right)^c \right)^{\frac{1}{c}} \quad (2.28)$$

where $\widetilde{\Delta\Delta t}$ is representative of overall tension resolved, N is the range of nodes considered, j iterates through all $\Delta\Delta t_j$, ΔT_{sub} denotes the distance between two adjacent $T_{sub,j}$, $\sum_{j=1, \forall \Delta t_{s,j} < \frac{T_{sub}}{2}}^N$ denotes summing across all values where $\Delta t_{s,j}$ is less than half the distance between its adjacent $T_{sub,j}$ on either side, n is the number of nodes summed and c is the pre-emphasis as explained with Equation 2.21.

This effectively computes the pre-emphasized, weighted and compensated mean $\Delta\Delta t_j$ across n eligible common subharmonics for a given transition. $T_{sub,j}$ weights down larger subharmonics which are less significant according to the theory.¹⁴ $\Delta t_{s,j}$ compensates for the fact that stationary consonance also affects one's preference for the succeeding chord apart from tension resolution itself. $\Delta t_{s,j} < \frac{1}{2} \Delta T_{sub,j}$ effectively sets the criterion for a node to be considered a common subharmonic. In our experiments, $N = 9$ is used. In consideration of divergent transitions in the dataset, $c = 1$ (i.e. no pre-emphasis) is used. Divergent transitions have negative $\Delta\Delta t, j$ which can be distorted by pre-emphasis.

¹⁴It is a reciprocal as opposed to Equation 2.21 because pleasure is associated with larger the amount of tension released.

With transitional harmony, conducting an accurate listening test is less straightforward. Rather than attempting to acquire a small number of fresh unproven opinions, it is reasonable to use statistics from a large number of well-esteemed pre-made decisions. A simple way to measure how well numerical values of subharmonic transition agree with the music theorists' school is to compare them with statistics of an expert music theorist's chord use. Capturing chord-use statistics from music score is again, however, a labor-intensive process [25, 84, 109]. Details such as melody-harmony discrimination, transition onset and root ambiguity (e.g. Dm7/F versus F6) are often not precisely defined in a song. The largest relevant data readily available (that meets the requirement of precise chord spelling) was found in Tymoczko's *Study on the Origins of Harmonic Tonality* [26]. In this study, Tymoczko interpreted and recorded the statistics of 11,000 chord transitions from Palestrina's [13] corpus. Palestrina was highly regarded for his style of harmony by Helmholtz himself [110] and he is widely considered amongst music theorists to be the pinnacle of contrapuntal harmony [111].

Table 2.3 lists $\widetilde{\Delta\Delta t}$ against frequencies of occurrences for each of the 17 most frequent transitions as read off Tymoczko's [26] chord tendency histogram. C, D, X $\uparrow$ and X $\downarrow$ indicate the convergence type of the progression. Just intonation was used as opposed to equal temperament in this case to be consistent with the database (Palestrina).

	I	V7	iii6	V/V	V2	vi	V6/V	vi7	i	iii	vi6	I6	ii6	viio	V6	ii	IV
Convergence*	C	D	D	C	D	X $\uparrow$	C	X $\uparrow$	C	C	X $\uparrow$	C	C	D	D	C	X $\downarrow$
Tymoczko's [26]	42	7	6	2	2	11	1	1	0	4	1	5	2	0.5	2	2	6
$\widetilde{\Delta\Delta t}$	444	-1.5	1.8	1.9	0.4	3.4	2.3	1.9	6.9	1.2	0.3	39.6	1.9	7.4	4.3	2.5	192

*States of Convergence:

C denotes convergence of $\Delta\hat{t}$

D denotes divergence of $\Delta\hat{t}$

X $\uparrow$ denotes escalating excursion of $\Delta\hat{t}$

X $\downarrow$ denotes descending excursion of $\Delta\hat{t}$

Table 2.3: Supplementary Table (Tab. S1) Tabulation of $\Delta\Delta\hat{t}$ against Tymoczko's chord tendency statistics [26] computed over 11,000 transitions.

Their correlations are listed in Table 2.4. $\widetilde{\Delta\Delta t}$ shows a significantly strong positive correlation of 0.903 with Palestrina's chord tendencies in general. It is close

Resolutions*	Complications [†]	Excursions [‡]		All excl. Comp. [†]	All
		Escalating	Descending [§]		
0.996 (0.0000)	-0.761 (-0.1353)	0.863 (-0.3366)	-	0.970 (0.0000)	0.903 (0.0000)
*Our model is designed to compute tension release in resolution.					
†Complications in music may be interpreted as negative tension resolutions, hence, correlation seen is negative.					
‡Excursions usually encompass tension release, however, apart from resolution alone, perception of succeeding chord is also influenced by the rising or falling of parallel melodies.					
§Apart from the descending excursions leading to IV, insufficient other descending transitions are recorded to compute its correlation.					

Table 2.4: Tabulation of correlations between $\widetilde{\Delta\Delta t}$ and Palestrina’s chord use statistics as collated in [26]. Correlations are listed in the top row with corresponding significance in brackets below.

to perfect at 0.996 for resolutions since the programmatic version of the model was designed with resolutions in mind. Complications may be interpreted as a negative release of tension. Even though a large number of contributing $\Delta\Delta t$ are negative, only one negative $\widetilde{\Delta\Delta t}$ can be seen in the table due to the influence of non-negative candidates. Nevertheless, $\widetilde{\Delta\Delta t}$ shows a strong negative correlation of -0.761 with [26] for complications. As earlier explained, with excursions the perception of a succeeding chord is also influenced by the rising or falling of parallel melodies. Unfortunately, descending excursions were insufficiently popular in Palestrina and only V-IV was being tallied. For escalating excursions, however, there are enough statistics to compute a correlation of 0.863. The correlation across all chords excluding complications was computed (because, as explained, they correlate negatively) to be 0.970.

2.7 Addressing the Fundamental Questions of Psychoacoustic Harmony

At this point, the fundamental questions of psychoacoustic harmony will be addressed in the context of subharmonic modulations. Beginning with question 2,

the first question will be left for the last.

2. The definition and explanation of stationary harmony. i.e. *what sounds good and why*, or, mathematically, to quantify $\varepsilon\{X_n\}$, where ε denotes *the harmonious effect of* and X_n represents chord n .

With large subharmonic tension being perceived as dissonance while small subharmonic modulations are perceived as consonance, the aesthetics of a chord may be visualized in the subharmonic tension acting on its shortest common subharmonics. Mathematically, they are inversely related. As described by Equation 2.20, $\varepsilon\{X\} \propto \frac{1}{\Delta\hat{t}_X}$

3. The definition and explanation of transitional harmony. i.e. *what sounds good, why and when*, or, mathematically, to quantify $\varepsilon\{X_1 \rightarrow X_2\}$, where ‘ $\rightarrow$ ’ denotes transition from one chord to another.

The aesthetics of a chord transition may be visualized in the transition of subharmonic tension acting on its shortest common subharmonics. As explained with Equation 2.16 and indicated by the arrows in Figure 2.13, this refers to the transition to the shortest common subharmonics of the succeeding chord from the nearest subharmonics of the preceding chord. Thus, resolution (tension release) in harmony progression is perceived in the convergence of $\Delta\hat{t}$ (where $\Delta\Delta\hat{t} > 0$) while what has been referred to as complication (build-up of tension or negative resolution) is seen in its divergence (where $\Delta\Delta\hat{t} < 0$) [45]. Mathematically, as described by Equation 2.24, $\varepsilon\{X_1 \rightarrow X_2\} \propto \Delta\Delta\hat{t}_{X_1 \rightarrow X_2}$

4. The phenomena that:
 - (a) a chord that sounds better than another out of context can sound worse than it in context [16]. Given $\varepsilon\{X_2\} > \varepsilon\{X_3\}$ show that $\varepsilon\{X_1 \rightarrow X_2\} < \varepsilon\{X_1 \rightarrow X_3\}$.

The section on subharmonic modulations differentiates between, stationary tension and transitional tension. The tension release brought about by the transition to a chord may be large even for high tension succeeding

chords. An example of such a case is seen with E^7 , G and Am^7 . Taking $E^7 = \{b_3, d_4, e_4, g_4^\sharp\}$, $G = \{g_3, b_3, d_4, g_4\}$ and $Am^7 = \{a_3, c_4, e_4, g_4, a_4\}$, the stationary subharmonic tension for G and Am^7 may be computed by Equation 2.18 to be $\Delta\hat{t}_G = 0.902\%$ and $\Delta\hat{t}_{Am^7} = 6.849\%$, respectively. Thus, $\varepsilon\{G\} > \varepsilon\{Am^7\}$. Whereas, the transitional subharmonic resolution (tension resolution) for $E^7 \rightarrow G$ and $E^7 \rightarrow Am$ may be computed by Equation 2.22 to be $\Delta\Delta\hat{t}_{E^7 \rightarrow G} = 8.783\%$ and $\Delta\Delta\hat{t}_{E^7 \rightarrow Am^7} = 10.748\%$, respectively. Thus, $\varepsilon\{E^7 \rightarrow G\} < \varepsilon\{E^7 \rightarrow Am^7\}$, despite the fact that $\varepsilon\{G\} > \varepsilon\{Am^7\}$.

- (b) a chord that sounds better than another in one context can sound worse than it in another context [16]. Given $\varepsilon\{X_4 \rightarrow X_2\} > \varepsilon\{X_4 \rightarrow X_3\}$ show that $\varepsilon\{X_1 \rightarrow X_2\} < \varepsilon\{X_1 \rightarrow X_3\}$.

With reference to Equation 2.22 and our answer to question 3, since our ears identify the subharmonics of preceding notes that correspond to the succeeding common subharmonic, transitional harmony is contextual. This means that the preceding chord affects the aesthetics of the transition as much as the succeeding chord does. Continuing from our answer to question 4a, take D^7 is taken to be $D^7 = \{c_4, d_4, f_4^\sharp, a_4\}$. The transitional subharmonic resolution (tension resolution) for $D^7 \rightarrow G$ and $D^7 \rightarrow Am^7$ may be computed by Equation 2.22 to be $\Delta\Delta\hat{t}_{D^7 \rightarrow G} = 11.421\%$ and $\Delta\Delta\hat{t}_{D^7 \rightarrow Am^7} = 4.540\%$, respectively. Thus, $\varepsilon\{D^7 \rightarrow G\} > \varepsilon\{D^7 \rightarrow Am^7\}$ despite the fact that $\varepsilon\{E^7 \rightarrow G\} < \varepsilon\{E^7 \rightarrow Am^7\}$.

5. The phenomenon that the transition from a low-tension chord to a high-tension one can still bring about the effect of tension release (resolution). Given $\varepsilon\{X_1\} < \varepsilon\{X_2\}$ show that $\varepsilon\{X_1 \rightarrow X_2\} > 0$.

The answer to this is in the independence of stationary and transitional tension, as established in our answer to Question 4. An example of such a case is seen with $E = \{b_3, e_4, g_4^\sharp\}$ and $Am^7 = \{a_3, c_4, e_4, g_4, a_4\}$, the transitional subharmonic resolution (tension resolution) for $E \rightarrow Am^7$ may be computed by Equation 2.22 to be $\Delta\Delta\hat{t}_{E \rightarrow Am^7} = 4.323\%$. The stationary subharmonic tension for E and Am^7 may be computed by Equation 2.18 to be

$\Delta\hat{t}_E = 0.902\%$ and $\Delta\hat{t}_{Am^7} = 6.849\%$, respectively. Hence, $\varepsilon\{E \rightarrow Am^7\} > 0$ despite the fact that $\varepsilon\{Am^7\} < \varepsilon\{E\}$.

1. Explain the phenomenon that the effect of harmony is greater than that of the sum of its parts [1, 2]. $\varepsilon\{x_1 + x_2 + x_3\} \gg \varepsilon\{x_1\} + \varepsilon\{x_2\} + \varepsilon\{x_3\}$.

With the exception of octaves (which are not usually considered harmony) and extreme examples of precisely tuned chords in rational intonation, the stationary tension of any combination of unique notes is observed to be larger than zero on the subharmonic plot. Hence, by Equation 2.19, $\Delta\hat{t}_{x_1+x_2+x_3} > 0$. Likewise, the stationary tension of each note on its own is observed to be zero on the subharmonic plot. Hence, $\Delta\hat{t}_{x_1} = 0$, $\Delta\hat{t}_{x_2} = 0$ and $\Delta\hat{t}_{x_3} = 0$ for all x_1 , x_2 and x_3 within the musical range. Thus, by Equation 2.19, $\varepsilon\{x_1 + x_2 + x_3\} \gg 0$ Whereas, by Equation 2.20, $\varepsilon\{x_1\} = 0$, $\varepsilon\{x_2\} = 0$ and $\varepsilon\{x_3\} = 0$. Therefore, $\varepsilon\{x_1 + x_2 + x_3\} \gg \varepsilon\{x_1\} + \varepsilon\{x_2\} + \varepsilon\{x_3\}$.

2.8 Conclusion

In this chapter the notion of interharmonic and subharmonic modulations was proposed as a psychophysical basis for stationary and transitional harmony. For stationary harmony, interharmonic and subharmonic tensions were proposed as consonance-dissonance measures with interharmonic and subharmonic modulations respectively. For transitional harmony, transitional subharmonic tensions were proposed as a measure of resolution. Correlations with perceptual [82] and chord-use [26] statistics show their relation to perceptual tensions [1, 14, 15, 107] and resolutions [1, 16] in music. Finally, the work in this chapter is used to address several fundamental questions of harmony.

Chapter 3

Singing Synthesis

Singing synthesis constitutes a significant part of singing harmony synthesis. As described in Chapter 1, this refers to the artificial production of the human singing voice by electronic means. Typically, the lyrics and pitch are specified by a user. Alternative to lyrics, phoneme, formant information, vocal tract shape or spectral shape may, instead, be specified.

This chapter¹, presents our work in singing synthesis, and is organized as follows. Section 3.1 will first give a background on singing synthesis and existing methods. Section 3.2 then describes a baseline system. Section 3.3 follows with our proposed method. Section 3.4 presents our experiment and results. Section 3.5 finally closes the chapter with our conclusion.

3.1 Background

This section will give a background on singing synthesis. Subsection 3.1 will first give a brief history on artificial voice production. Subsection 3.1.2 will then provide an overview of different singing synthesis methods. Subsection 3.1.3 will then list all notable works in the field.

¹Work in this chapter were published in [51] and [52]. This also led to the filing of a patent [53] in singing synthesis.

3.1.1 History of Artificial Voice Production

Attempts at reproducing the human voice dates back to pre-electronic era. Methods evolved as technology progressed, attesting to man’s continued pursuit of artificial voice production. This section covers the background of works in artificial voice production systems leading up to the development of singing synthesis technology. Figure 3.1 plots an overview of significant works across time.

Figure 3.1: A Timeline of Voice Reproduction and Singing Synthesis

Kratzenstein’s Resonators: In 1779, Christian Kratzenstein modeled five common vowel sounds with physical resonators that produced the vowels /a/, /e/, /i/, /o/ and /u/ [112, 113].

The Acoustic-Mechanical Speech Machine and the Euphonia: In 1791, with the dawn of mechanical automation technology, Wolfgang von Kempelen constructed a mechanically actuated acoustic model [112–114]. Lungs were modeled by bellows, the glottis was modeled by a reed, nostrils were modeled by two tiny tubes and the oral cavity was modeled by a manually deformable leather tube. A handle cuts off air through the glottal reed for unvoiced sounds, and sibilances /s/

and /sh/ that are difficult to produce without teeth are produced by individual whistles activated by levers. However, it was only in 1846 that mechanical-actuated acoustic voice reproduction was completed with Joseph Faber’s Euphonia [112], which was complete with a phantom face and sufficient pitch control to enable the generation of singing voices.

Stewart’s Formant Synthesizer: In 1922, following the dawn of analog electronics, Stewart constructed the first formant synthesizer [113]. A glottis was modeled by a single electronic buzzer and two resonant filters modeled the filter effect of the vocal tract. This meant that it could model the first two formants (i.e. voice resonances, or peaks in the voice spectrum) of the human voice. This marked the launch of electronic voice synthesis, but was only capable of producing disjoined vowels.

Dudley’s VOCODER: It was not until 1939, that the first VOCODER [112–114] was unveiled. A vocoder models the human voice production system, decomposing it into features descriptive of voice production in a process called analysis, and recomposing the voice signal using the same set of features in a process called synthesis. This will be covered in detail in Section 3.2.3. Unlike vocoders used in singing synthesis systems, Dudley’s vocoder enabled compression for bandwidth optimization. Hence, the aperiodic part of voiced sounds were not modeled. Nevertheless, it served well, and even enabled encryption for secure communications during the war. More relevant to the subject matter, Dudley’s vocoder paved the way to increasingly better successive vocoders, which ultimately led to technologies that enable singing synthesis.

The Dynamic Analog of the Vocal Tract: In 1958, George Rosen developed the DAVO (Dynamic Analog of the VOcal tract) [115]. This was an articulatory synthesizer (further details in Section 3.1.2) which was able to produce both speech and singing, marking the birth of singing synthesis.

VOCALOID: The first version of the state-of-the-art synthesizer singing synthesizer was first release in about 2004 and the latest release in 2014. It was

jointly developed by Kenmochi Hideki [50], Jordi Bonada [48, 116] and Xavier Serra [117–119] originally at Universitat Pompeu Fabra. This is a concatenative synthesizer and is very popular in Japanese pop culture today [120].

3.1.2 Overview of Singing Synthesis Methods

Singing synthesis systems may be classified into six different methods. The first method is formant synthesis, which generates singing voices by passing a gottal source generator through a multi-resonant filter [121, 122]. The second method is physical modeling synthesis, which aims to model the human voice production system more meticulously based on vocal tract shape [123]. The third method is concatenative synthesis [48–50, 116, 118, 119], which was described in the last section as a baseline, became the de facto when it became more effective to simply concatenate segments of voices from a sample database as data storage technology significantly advanced. The maturity of concatenative synthesis for neutral humanlike singing, and its limitations for expressivity called for the time-modeling of descriptive parameters, giving birth to the fourth method, which is parametric synthesis [124–126]. The notion of dilation in convolution neural networks allowed for the modeling of waveforms in different degrees of resolution which is required for emulating the human voice. This led to the development of the fifth method, which is wavenets [127, 128]. Unfortunately, wavenets remain difficult to condition. The sixth method, wavetable synthesis, is a simplification of concatenative synthesis with realtime suitability.

Each of the aforementioned methods will be briefly introduced in the following subsections. Their use in existing systems will be covered in the next subsection.

Formant Synthesis Formant synthesis models the human voice with an oscillator and series of resonant filters [121, 122]. As depicted in Figure 3.2, the system accepts two inputs. The first is the target melody, $f_0(t)$ and the second is a time sequence of formant descriptors, $\Lambda(t)$, which contain the formant

intensities $[i_1(t), i_2(t), \dots]$, paired with their respective centre frequency positions $[f_1(t), f_2(t), \dots]$. During synthesis, a glottal-like waveform is generated at fundamental frequency, f_0 , and filtered to the spectral envelope described by $[i_1(t), i_2(t), \dots]$ and $[f_1(t), f_2(t), \dots]$.

Figure 3.2: Formant Synthesis

Physical Modeling / Articulatory Synthesis Physical modeling synthesis, also referred to as articulatory synthesis, is based on a physical model of the vocal tract [123]. As depicted in the example in Figure 3.3, a source generator produces glottal waveform according to the pitch trajectory, $f_0(t)$ specified the input. This is filtered by a physical model of the vocal tract according to parallel physical parameters, $M_\phi(t)$, that describe the shape of the vocal tract.

Unit Selection / Concatenative Synthesis Concatenative, or unit selection synthesis, as depicted in Figure 3.8 and explained in Section 3.2.2, is based on the selection of appropriate phone units from a database and modifying them to meet input specifications before concatenating the modified units to produce the singing voice [48–50, 116, 118, 119].

Figure 3.3: Physical Modeling / Articulatory Synthesis

Figure 3.4: Parametric Synthesis

Parametric Synthesis Parametric systems train models (e.g. Hidden Markov Models) of parameters describing the singing offline as shown in Figure 3.4 [124–126]. At runtime, the models relevant to input melody, $f_0(t)$, and lyrics, $\phi(t)$, are invoked.

This is very similar to physical modeling synthesis in Subsection 3.1.2. Physical

modeling synthesis currently focuses on attempting to parameterize each phone in a way that describes the physical attributes of the human voice production system. There is usually only one articulatory model per phone, hence the database is small. Parametric synthesis, however, focuses on modeling the parametric trajectory over time in a way that describes the sound that is produced. There may be models for different phone transitions across several phones or entire syllables. [124–126].

Figure 3.5: Wavenet Synthesis

Wavenet Synthesis Wavenet synthesis is based on a dilated convolutional neural network [8]. As showed in Figure 3.5, the network is trained by utterances from a singing database offline. With this, the output in the example would produce random singing. It is difficult to condition the network to produce specific melody or lyrics, although, some works have been done with workarounds with varying degrees of success [127, 128].

Wavetable Synthesis Wavetable synthesis may be understood as a simplification of concatenative synthesis based on the modification and summation of a short wavesamples from a database of phonemes as shown in Figure 3.6. In the example shown in the figure, grapheme trajectories, $A_g(t)$, control the weighting

Figure 3.6: Wavetable Synthesis

of the phonemes which are first modified according to the melody, $f_0(t)$, before summation.

Vocoder complexity is reduced in wavetable synthesis as opposed to concatenative synthesis with the trade-off of a painstakingly designed database. Better matched phones allow for the vocoder process to be replaced by resampling, making this method capable of producing realistic sounding singing voices realtime to the sub-frame level.

For this reason our novel Mandarin and Japanese realtime synthesis method, SERAPHIM [51–53] was developed based on wavetable synthesis.

3.1.3 Notable Works

Each synthesis method described in the previous section produces synthesized singing with its own unique musical characteristics and performance. Formant synthesis produces unique robotic characteristics of electronic music of the seventies [129]. Concatenative synthesis currently sounds the most humanlike amongst all functional methods, but lacks expressivity. Parametric synthesis, has the potential of addressing the issue of expressivity but currently still falls behind. Physical

modeling singing synthesizers have not advanced very much because of its complexity. Nevertheless, with today's 3D physical modeling technology [130], this has much potential for highly realistic emotion because of its attention to detail. Wavenet synthesis currently produces the most realistic voices but remains difficult to specify what is being sung [127, 128]. Wavetable synthesis remains highly suitable for subframe realtime synthesis.

Over the last three decades, a number of singing synthesizers have been developed using the methods mentioned in Section 3.1.2. Since Larsson's work on the MUSSE formant synthesizer in 1977 [121], numerous other singing synthesizers have evolved, including Yamaha's Vocaloid [48–50, 131]. The author has also developed and filed a patent [53] for a novel realtime singing synthesis method with 3D lip synchronization [51, 52]. These are described in the paragraphs that follow.

Music and Singing Synthesis Equipment (MUSSE) was the pioneering work in singing synthesis by Larsson [121] of the Royal Institute of Technology (Kungliga Tekniska Högskolan)'s Department for Speech, Music and Hearing based on formant synthesis. This pioneering piece of work was implemented in hardware by passing a source generator and noise through five resonant analog filters. Vibrato may be specified by an auxiliary analog input.

CHANT is an early formant-based singing synthesizer implemented in software [122]. One novel idea with CHANT is the way it exploits the nature of the pulse train to optimize the convolution process in each resonant filter. Since the impulse train is a series of '1's and '0's, the convolution process is reduced to a sequence of summations and hence eliminates the need for any multiplications [122].

CANTOR is a proprietor software synthesizer released in 2004 by VirSyn [132]. It is essentially a formant synthesizer implemented by means of additive synthesis. Unlike MUSSE, in Subsection 3.1.3, CANTOR does not need a filter since individual sinusoid components are generated independently by up to 256 sine wave generators. The sinusoids are generated at harmonic frequencies to form the periodic part of the voice. A noise generator generates the signal for the aperiodic

parts. Filter envelopes describing vowel shapes may be stored as templates and the user can specify how to morph between them.

Singing Physical Articulatory Singing Model, abbreviated SPASM, is an articulatory (physical modeling) singing synthesizer developed by [123]. The vocal tract, nasal tract and mouth are modeled tubes that are decomposed into sections of different diameters. Since each section admits and reflects different amounts of pressure, they are modeled by a filter of the ladder structure.

Lyricos was amongst the first concatenative singing synthesizer and jointly developed by Georgia Tech and Texas Instruments in 1996 [133]. Its structure is simpler than conventional concatenative synthesizers as explained in Subsection 3.1.2 and will be further explained in Section 3.8. However, unlike conventional concatenative synthesizers, no singing database is used at runtime. Instead, a database containing sinusoidal levels similar to the additive method of CANTOR in Subsection 3.1.3 is used. Phone units are chosen from the database are first concatenated and then modified to meet target requirements. Again, this modification is different from conventional concatenative synthesizers as an overlap-and-add method is used instead of a vocoder.

UTAU is a notable concatenative synthesizer that was developed as a shareware response to Yamaha's Vocaloid Section 3.1.3 by Ameya / Ayame in 2008 and garnered a large number of users [134]. Unfortunately, there is an unavailability of publications on its algorithm. Nevertheless, its algorithm is said to be very similar to the one described in Section 3.1.2 and Vocaloid's.

NCKU's Singing Synthesizer [124] and **SinSy** [126, 135] are both based on parametric synthesis as described in Section 3.1.2. Both use HMM [125] for modeling.

Vocaloid is Yamaha's proprietary software singing synthesizer [48–50, 131] and was first released in 2004. The latest version, Vocaloid 4, was released in 2014 and

may be regarded as the state-of-the-art system today.

SERAPHIM was developed by the author in 2016. The system was described [51] and demonstrated [52] in San Francisco at Interspeech 2016 and patent was filed [53] in Singapore in the same year. We will present SERAPHIM in greater detail in the section 3.3.

3.2 Baseline Synthesizer

This section describes a baseline singing synthesis system in three different levels of resolution. Subsection 3.2.1 describes the system by on a general level, defining it by its inputs and outputs. Subsection 3.2.2 further describes the internal blocks of a conventional singing synthesis system. Subsection 3.2.3 examines how the voice is manipulated within the key component of the singing synthesis system, the vocoder.

Figure 3.7: System Definition

3.2.1 Mathematical Definition of System Inputs and Outputs

Conventional singing synthesis systems are defined by inputs of melody and lyrics and outputs of singing voices as shown in Figure 3.7. This is concisely described by the following equation.

$$\Sigma(t) = \xi(f_0(t), \phi(t)) \quad (3.1)$$

where $\Sigma(t)$ is the synthesized singing voice at the output, t represents time, ξ represents the singing synthesis process, $f_0(t)$ represents the melody, which is a sequence of fundamental frequencies of the song, and $\phi(t)$ represents the lyrics in the form of a time sequence of phones.

3.2.2 System Components

Figure 3.8: A Typical Conventional Singing Synthesizer

There are different methods of singing synthesis. The most typical method, as shown in Figure 3.8, uses the method of concatenation. Such a system includes several components, namely, a singing voice database, a parameter generation component, a unit selection component as well as the analysis and synthesis parts of a vocoder. The concatenative singing synthesizer searches its singing voice database for most suitable voice segments, and joins them together, modifying each segment to better match target specifications with a vocoder, where:

- The voice database stores the voices for use in concatenation.
- The parameter generation component sets the target specifications based on the user's input.
- The unit selection component selects the voice segments for concatenation based on the parameter generation component's specifications.
- The vocoder modifies the segments to better match target specifications.

This subsection describes the role of each component in detail.

The singing database stores a large collection of unaccompanied singing $\Sigma_m(t)$, where each $\Sigma_m(t)$ is a recording of a sung song and m is the song index. The recordings are accompanied by transcriptions, that include synchronized information in terms of lyrics in the form of phone sequence, $\phi_m(t)$, pitch sequence, $f_m(t)$, and additional contextual information such as adjacent phones and part of speech, $P_m(t)$. These transcriptions enable the system to select the most appropriate singing segments for concatenation.

The role of the parameter generation component is to interpret the raw inputs from the user (melody $f_0(t)$ and lyrics $\phi(t)$) to generate two sets of outputs, $\{\phi, f_0, P\}$ and $\{f_{Target}\}$. The first set, $\{\phi, f_0, P\}$, describes the phone units to be retrieved by the unit selection component. Since the available phone units in the database often does not completely match the desired specification, they must be transformed prior to concatenation with their adjacent phones. The second output, $\{f_{Target}\}$, specifies the desired pitch of the phone at time t to the synthesis part of the vocoder component to re-synthesize the retrieved phones.

The role of the unit selection component is to search for the best matching phone units, u_{Source} , from the singing database given the desired phone, pitch and contextual information (ϕ , f_0 and P respectively). The selected phone units are then relayed to the analysis part of the vocoder component to be modified by the vocoder.

The vocoder is a key component of the synthesizer. It allows the pitch, voiced spectrum and unvoiced spectrum to be modified independently of one another. Its role is to transform phone units received from the unit selection component, u_{Source} , into phone units to be output by the system, u_{Target} , according to the specifications received from the parameter generation component. It does so in two parts. The first part performs analysis, and describes the input unit, u_{Source} , in terms of a set of descriptive parameters, i.e. f_0 , H and N , where f_0 represents pitch, H represents spectral shape of the periodic part and N represents the noise component. The second part synthesizes the output unit, u_{Target} , according to a similar set of descriptive parameters, f_0 , H and N . In the singing synthesizer in Figure 3.8, the pitch f_0 from the analysis of phone unit u_{Source} is replaced by the pitch f_{Target} from the parameter generation component between the analysis and synthesis parts of the vocoder. Subsection 3.2.3 will describe the analysis and synthesis process in greater detail.

The target singing voice $\Sigma_{Target}(t)$ is finally produced by frame-by-frame concatenation of the output units, u_{Target} .

3.2.3 Analysis and Synthesis Process of the Vocoder Component

A vocoder decomposes a given voice signal into features descriptive of the human voice production system in a process called analysis and recomposing the voice signal based on the same set of features in a process called synthesis.

The analysis process expresses every frame, Σ , of the given voice signal in the form of its glottal source, X , spectral envelope, H , and aperiodic residual, N , components by

$$(X \otimes H) + N = \Sigma \quad (3.2)$$

and the synthesis process conversely recomposes the voice segment, Σ , from the three component features.

Figure 3.9: The Analysis and Synthesis Processes of a Vocoder

Figure 3.9 shows the analysis and synthesis processes in the spectral domain. Traditionally, this is used for bandwidth reduction for transmission. In singing synthesis, however, it allows a voice segment to be modified by changing X , H and N components independently from one another. In the implementation in this section, it is used primarily to modify source segments to the target pitch. Individual vocoders may describe the voice using slightly different components. It

is common for pitch, f_0 , to be used instead of glottal source, X , since it is easy to reconstruct the glottal source given the pitch alone.

3.3 Proposed System

The majority of state-of-the-art synthesizers today are offline systems dedicated for offline music production purposes. Such systems are unable to respond in real-time, as required for live music. A large part of music and singing, however, is in the real-time interaction amongst the singers, instrumentalists, and audiences.

Emotion is expressed in minute variations of pitch, amplitude, timbre and timing [136] of a singer. The emotional expressions in the performance of a singer, for example, influences the emotions of a guitarist in real time - whose expressions of emotions in turn influences the performance of the singer in a feedback loop. Multiple loops are intertwined in a live performance, involving every performer on stage and the audience. Hence, for singing synthesis to be involved in this level of music making, it has to respond in real-time. This provides a strong motivation for work in real-time singing synthesis systems.

In this subsection, we will present SERAPHIM, a novel real-time singing synthesis system [51–53]. Subsection 3.3.1 will first define the scope of this proposed method. Subsection 3.3.2 will then describe the runtime version of the algorithm. Subsection 3.3.3 will finally present the real-time algorithm.

3.3.1 Scope

The scope of this work is real-time singing synthesis to meet onstage performance requirements. State-of-the-art synthesizers are suitable only for offline music production purposes and are unable to meet real-time demands for live onstage use. Singing, being by nature a performing art, gives rise to the desire of a live performance.

Figure 3.10: Realtime Capability and Phone Coverage of SERAPHIM against existing Realtime Singing Synthesis Systems

As mentioned in Chapter 1, there are presently two existing methods for realtime singing synthesis: LIMSI’s Cantor Digitalis [46] and Yamaha’s VOCALOID Keyboard [47–50]. Figure 3.10 shows the realtime capabilities and phone coverage of the two in blue and green, respectively. As shown in the figure, Cantor Digitalis is realtime to the subframe level but only covers vowels. VOCALOID Keyboard, on the other hand, covers all phones but is realtime only to the syllable level.

In general, subframe response is ideal for expressive control. A critical exception is seen in unvoiced consonants, where articulation is often abrupt. These often require drastic changes to the timbre that is often too quick for non-verbal human control. Based on this, an ideal realtime singing synthesis system should have two levels of response. For vowels and voiced consonants, subframe response is desirable. Whereas, for unvoiced consonants that involve sudden changes in spectrum, it is desired to pass some control to the synthesizer and phone-level capability is

desired. The ideal realtime capability and phone coverage is illustrated in Figure 3.10.

3.3.2 Runtime Algorithm

Before presenting the realtime version, we will present a runtime version of SERAPHIM. The runtime version of Seraphim accepts a MIDI sequence of the music and lyrics for studio use. The algorithm for which is illustrated in Figure 3.11. As shown in the figure, the lyrics $\Phi(t)$ at the input is first converted into a sequence of phone labels, $\phi(t)$. Given this, together with the melody, $f_0(t)$, the corresponding phone sample, $\sigma_\phi(f_d, \tau)$ and amplitude trajectory, $a_\phi(t)$ are retrieved from their respective databases. Since the phone sample's pitch may not completely match the specified pitch, $f_0(t)$, it then needs to be shifted to the correct pitch, obtaining $\sigma_\phi(f_0, t)$ before being tapered by coefficient $a_\phi(t)$ to produce a single frame segment of the output singing voice, $\sigma(t) = a_\phi(t)\sigma_\phi(f_0, t)$. (Since the shift is usually very small with adequate samples across the vocal range, it is efficient to do it by interpolation or resampling, and $\tau/t = f_0/f_d$.) These frames are finally concatenated to produce the output singing voice, $\Sigma(t) = [\sigma(1) + \sigma(2) + \sigma(1) + \dots]$.

The following subsections will further explain different components and background of the runtime algorithm. Subsubsection 3.3.2.1 will first explain the syllable structure and phonetic background of Mandarin Chinese and Japanese. Subsubsection 3.3.2.2 will then follow with the phone sample database design strategy. Finally, Subsubsection 3.3.2.3 describes the wave additive trajectory library.

3.3.2.1 Syllable Structure and Phonetic Background

In Mandarin, syllables are made up of an initial consonant phoneme and a final that is compounded of several phones which might include a consonant coda. There are up to about 22 initials and 35 finals, depending on how they are being defined. SERAPHIM classifies them into 4 groups according to their initials for syllable modeling, as follows:

- Nulls - this is a special group of syllables without an initial. They begin with

Figure 3.11: Runtime Synthesis Flowchart

the first phoneme of the final which is a vowel and may be a suprasegmental.

- Voiced Consonants - this consists of all syllables whose initials are voiced.
- Unvoiced Consonants - this loosely groups all syllables whose initials are unvoiced or semi-voiced except those in the following (sibilances) group.
- Sibilances - this loosely groups all syllables whose initials are long, completely unvoiced sibilances and fricatives.

Similarly, in Japanese, syllables are also made up of an initial consonant phoneme and a final, except in Japanese, the final is almost always a single vowel. Suprasegmentals (e.g., 'Y' and 'W') are regarded as initials rather than grouping them as part of the final. SERAPHIM, hence, classifies syllables similarly for Japanese, but the Null group is not applicable.

Since the system is designed with realtime implementation in mind, as much of the computations that may be feasibly pre-performed offline is done so. Hence the two

model databases that are used in SERAPHIM are precomputed as much as possible. These are the syllable model database and the phone model database.

3.3.2.2 Phone Model Database

Figure 3.12: Phone- and Biphone-Model Wavetables

Phone models describe phone spectra at runtime, and since SERAPHIM is a light-weight time-domain system, the phone model is essentially a wavetable sequence of the phone.

Unlike conventional concatenative systems, where diphones are the primary unit for concatenation, in SERAPHIM, phones and biphones are used. Each syllable of the voiced consonants group has its initial consonant modeled by a phone model. Each syllable of the unvoiced consonants group has its initial consonant modeled together with its first vowel by a biphone model. Syllables of the sibilances group

may be modeled like those in the voiced consonants or unvoiced consonants group. All other vowels are modeled by individual phone models, with the exception of certain first vowels of the nulls group - those which are particularly abrupt are modeled like biphones as though they were preceded by unvoiced consonants.

For phone units, one wavetable is used per phone model, with approximately one model for every three semitones (four per octave) across the vocal range. The wavetable is tuned and normalized. Its power envelope is flattened and zero-centred, as shown in part (a) of Figure 3.12. Finally, it is set to start and end at the same phase so they may be set to loop. Biphone units are conditioned in the same way, except two wavetables are used per phone model. The first, as shown in part (b) of Figure 3.12 captures the initial phone of the sequence as well as the transition into the next phone. The second, as shown in part (c) of Figure 3.12 is of the second phone. It is set such that the second wavetable starts at the same phase as the end of both the first wavetable and itself. In this way, the second wavetable may be set to continue after the first wavetable and set to loop.

3.3.2.3 Syllable Model Database

For each class of syllables, for each possible final, a syllable model is built. The syllable model database consists of wave additive trajectories. The syllable models indicate the contribution of each composite phone to the overall timbre at each time frame. Figure 3.13 illustrates the wavetable trajectories across three different syllable lengths for the Mandarin syllable 'Shuang'. Across all three plots, the solid blue line plots the trajectory of the first phone, 'SHU'; the dashed purple line plots the trajectory of the second phone, 'A'; and the dotted red line plots the trajectory of the last phone, 'NG'. At least four lengths per syllable are modeled in SERAPHIM, spanning a semiquaver (quarter beat) to a semibreve (four beats) for a tempo of 120 beats per minute. These are computed using labeled wave recordings of the syllables. The wave recordings to be modeled are first labeled to mark syllable and phone transition boundaries, as well as the centre of each phone. A DTW-like method [137] is then used to traverse a matrix of their spectral similarities. The trajectories are finally superimposed onto the power envelope to produce the wave additive trajectories. Wave additive trajectories are normalized

Figure 3.13: Wave Additive Trajectories across 3 Different Syllable Lengths for the Mandarin Syllable 'Shuang'

in power and time and stacked to form a surface to form the syllable model which may be interpolated to invoke the model.

3.3.3 Realtime Algorithm

The real-time version of our method replaces the syllable models with user input gestures. A flowchart of which is illustrated in Figure 3.14. As shown in the figure, input gestures are first interpreted to acquire phone labels, $\phi(t)$, melody, $f_0(t)$, and amplitude trajectories, $a_\phi(t)$. Given the phone label, $\phi(t)$, the corresponding phone sample, $\sigma_\phi(f_d, \tau)$ is retrieved from the phone model database. This is pitch-shifted to the instantaneous pitch input, $f_0(t)$, to produce $\sigma_\phi(f_d, t)$, which is further tapered by the input amplitude coefficient, $a_\phi(t)$, to produce the current frame of singing voice, $\sigma(t) = a_\phi(t)\sigma_\phi(f_d, t)$.

Figure 3.14: Realtime Synthesis Flowchart

 Figure 3.15: Mandarin Chinese SERAPHIM *Live!* interface: Initials

Table 3.1 lists the features and capabilities of SERAPHIM alongside Yamaha's

Figure 3.16: Mandarin Chinese SERAPHIM *Live!* interface: Finals

Figure 3.17: Mandarin Chinese SERAPHIM *Live!* interface: Slurring of the word "Shui". The phone "u" is commonly slurred with certain Mandarin accents

VOCALOID and LIMSI's Cantor Digitalis. As can be seen from the table, Yamaha's VOCALOID has a static button interface for syllable level control of any syllable in the Japanese language. Similarly, LIMSI's Cantor Digitalis has a static gesture

interface for sub-frame level control of only vowels. In order to provide sub-frame level control of vowels and consonants of both Mandarin and Japanese requires the ability to represent multi-dimensional information on a two-dimensional surface in a simple, intuitive way. SERAPHIM does this by means of a dynamic interface that changes according to input. Based on phone sequence branching restrictions in each language, the interface dynamically anticipates all possible remaining phones based on current and previous inputs. These are presented intuitively to the user with a different bearing for a different type of sound.

Figure 3.15 shows the SERAPHIM *Live!*'s dynamic touch gesture interface prompting the user for the initial phone in Mandarin Chinese.

- **initials** Upon tapping on an initial phone, the onscreen region surrounding the user's right thumb is populated with possible subsequent phones corresponding to the initial phone.
- **subsequent phones** Figure 3.16 shows the interface prompting the user for subsequent phones upon the selection of 'B' as the initial phone.
- **phone transition, tremolos² and slurring³** As the user slides his/her thumb across the phone positions, the sound morphs into vowels or voiced consonants as labeled on screen. The user may transit the phones quickly or slowly according to expression. Back and forth transitions may mimics a tremolo effect. Figure 3.17 shows how slurring is possible by deviating from a particular phone.
- **pitch** Pitch is controlled independently by tapping, bending or gliding across the keyboard on the left. Nudging an up-down movement on the keyboard mimics a vibrato⁴ effect.

Figure 3.18 shows the interface of the Japanese version prompting the user for the initial phone. Like in the Mandarin version, tapping on an initial phone surrounds the user's right thumb with all possible subsequent phones corresponding to the

²amplitude modulation

³bypassing or incomplete phonation of a particular phone

⁴frequency modulation

Figure 3.18: Japanese SERAPHIM *Live!* interface: Initials

Figure 3.19: Japanese SERAPHIM *Live!* interface: Finals

initial phone. An example of which, for the initial phone 'n' is shown in Figure 3.19. The layout of the Japanese version is kept as similar to the layout Japanese text messaging for familiarity.

Table 3.1: Real-time Singing Synthesis Systems

	VOCALOID Keyboard [116]	Cantor Digitalis [121]	SERAPHIM Live! [51–53]
Developers	Yamaha, Universitat Pomeau Fabra	LIMSI	Institute for Infocomm Research
Interface	Static Buttons	Static Gestures	Dynamic Gestures
Real-time Capability	Syllable level	Sub-Frame level: Vowels only	Sub-Frame level: Vowels and Consonants
Languages	Japanese	<i>vowels only</i>	Mandarin Japanese

3.4 Experiment and Results

Table 3.2: Subjective Listening Test Results

Utterance Pair	Systems		Utterance Pair	Systems	
	Cantor Digitalis [46]	SERAPHIM Live!		VOCALOID Keyboard [47]	SERAPHIM Live!
A	20.00%	50.83%	E	60.00%	72.50%
B	25.83%	60.83%	F	53.33%	78.33%
C	34.17%	55.00%	G	64.17%	78.33%
D	36.67%	54.17%	H	58.33%	80.83%
Mean	29.17%	55.21%	Mean	58.96%	77.50%

Figure 3.21 illustrates the realtime capability of our proposed system against that of LIMSI’s Cantor Digitalis and Yamaha’s VOCALOID Keyboard. As shown in the figure, we have managed to achieve phone level realtime capability for unvoiced consonants and subframe level realtime capability for voiced consonants and vowels.

Four segments [46] of singing voices synthesized using LIMSI’s Cantor Digitalis and four [47] using Yamaha’s VOCALOID Keyboard were each mimicked using SERAPHIM *Live!*. The 8 pairs of segments in total were randomized and presented to 12 listeners in a subjective listening test, who were tasked to score the voices on a scale of 0 to 5, with 0 being the worst sounding and 5 being the best sounding. The mean result for each utterance is normalized to 100% and presented

Figure 3.20: Subjective Listening Tests

in Table 3.2. The mean result of each work is plot on the bar chart in Figure 3.20. As shown in the figure, SERAPHIM nearly doubles Cantor Digitalis performance and outperforms VOCALOID Keyboard by 0.927.

3.5 Conclusion

This chapter presents our work in singing synthesis. It covers broad overview of the history of artificial voice production, singing synthesis methods and presented several notable singing synthesis systems. A baseline singing synthesis system was then presented in detail. Following which, a realtime singing synthesizer was proposed and evaluated against two other benchmark realtime singing synthesis systems. The proposed system was evaluated to have the best realtime capability while providing the best phone coverage across existing such systems. It was also the preferred system across listening tests.

Figure 3.21: Realtime Capability and Phone Coverage of Existing Methods

In the next chapter, we will present our work in the synthesis of singing harmony.

Chapter 4

The Synthesis of Singing Harmony

Singing harmony, as defined in Chapter 1, is the blending of two or more voices to produce an effect pleasant to the human ear. The main singing voice is called melody and additional voices are called accompaniment. Together, they form singing harmony, which plays a significant role in classical and contemporary music alike. Accompaniment lines, however, are significantly harder to sing than the melody line because of the human tendency to gravitate towards the melody when singing [138]. This, together with advancements of vocoders and voice synthesis technology, motivates the work covered in this chapter¹ in the automatic generation of vocal harmony.

Section 4.1 first provides the background about singing harmony synthesis including an overview of the resynthesis strategy and existing methods. Section 4.2 then explains our algorithm in detail. Section 4.3 presents our experiment and results. Section 4.4 closes with our conclusion.

¹The work in this chapter was published in [7]. This also led to the filing of patents [54] and [61] in the synthesis of singing harmony.

4.1 Background

As seen in Chapter 3, synthesis of the singing voice is a problem that may be broken down into two parts. Namely, pitch, or perceived frequency, and lyrics, or semantics in spectra. In singing harmony synthesis, it is sufficient in most cases for the spectral components of the generated voices to remain the same as that of the input voice. Pitch, on the other hand, is derived from various methods usually combining specifications from the user with the pitch of the input voice. Hence, in our strategy for singing harmony synthesis, we will deal with pitch and spectral features separately as opposed to conventional singing synthesis. We call this resynthesis. Subsection 4.1.1 will give a background on how resynthesis leverages on the spectral part of the input while Subsection 4.1.2 will list existing methods of determining the right pitch contours to synthesize in order to harmonize well with the input voice.

4.1.1 Resynthesis Strategy and Spectral Features

Singing harmony synthesis may be realized by leveraging on features extracted from input singing voice. Singing resynthesis refers to the adoption of singing synthesis techniques while leveraging on features extracted from input singing voice. Applications that adopt the resynthesis technique includes pitch shifting [139–141], autotune [56, 139, 142, 143], the vocoder effect [129] and speech-to-singing synthesis [144–146]. In the case of singing harmony synthesis, it uses spectral and modified pitch features from an input singing melody.

Pitch-shifting modifies the pitch of the singing voice while keeping spectral features the same. Autotune rounds off the pitch value of the singing voice to a set of pre-defined values called the scale². The vocoder effect replaces the source signal of the human voice as described in Subsection 3.2.3 with a signal from a musical instrument. Speech-to-Singing modifies the pitch of a spoken voice to mimic that of a singing voice (with modifications to the timing of each syllable). Like the

²See Appendix A.2.1

aforementioned applications, singing harmony synthesis can adopt spectral features of the input voice to synthesize its output voices. The pitch of the input voice may also be used to work out the output pitch. This will be discussed in greater detail in the following subsection.

4.1.2 Pitch Features with Existing Automatic Harmonization Methods

Existing implementations of harmonizers may be classified into 4 classes according to the way they derive harmony information.

Table 4.1: Current Methods of Harmony Generation

Method	Strategy	Synchronization	Dissonances	Other Issues
458 ¹	Fixed Interval	Inherently Synchronized	Type-I & Type-II	Parallel Motion
458-II ²	Interval conforms to key	Inherently Synchronized	Type-II	Requires Key Input
AUX ³	Interval conforms to auxiliary instrument	User Synchronized	Type-II during synch/detection errors	Requires Auxiliary Instrument Requires User to Synchronize
KTV ⁴	Based on score	User Synchronized	Type-II during synch errors	Requires MIDI file Requires User to Synchronize

¹458: The Fourths, Fifths and Octaves Method
²458-II: The Thirds and Sixths Method
³AUX: The Auxiliary Method
⁴KTV: The Karaoke Method

As tabulated in Table 4.1, the Fourths, Fifths and Octaves Method, 458 [7, 54–57] achieves harmony by a fixed interval and thus requires no auxiliary information. However, even when the singer is singing in precise pitch and timing, this strategy is prone to Type-I and Type-II dissonances as well as the negative effect of parallel motion [3]. Statistically, Type-I dissonances may be computed to occur $\frac{1}{7}$ of the time.

The Thirds and Sixths Method, 458-II [7, 54, 58, 59], is an improvement to the 458 system, as it eliminates Type-I errors³ by limiting harmonizing notes to the key pre-specified by the user. Even when the singer is singing in precise pitch and timing, this is still prone to Type-II dissonances³ because chord information is not

available.

The Auxiliary Method, AUX [7, 54, 58, 59], is a further improvement to 458-II, as it reduces Type-II errors in 458-II by deriving chord information, C , from an auxiliary instrument input. This relies on the following:

- 1) accurate polyphonic pitch detection algorithm in the chord derivation process: errors in chord derivation would lead to the Type-II dissonances irregardless of how well the singer sings.
- 2) accurate synchronization between the singer and auxiliary instrument: if the singer sings out of time with the auxiliary instrument, Type-II errors may result.

The Karaoke Method, KTV [56, 60], refers to a MIDI score instead of an auxiliary instrument for harmony information. This method relies on manual synchronization and still produces Type-II errors at junctures with synchronization errors.

Subsections 4.1.2.1 through 4.1.2.4 cover the 458, 458-II, AUX and KTV methods respectively and how they work.

4.1.2.1 The Fourths, Fifths and Octaves Method

Figure 4.1: The Fourths, Fifths and Octaves Method (458)

³See Appendix A.4

This method, referred to as the 458 method in [7, 54], and used in [55–57], uses a predefined fourth, fifth or octave interval throughout the entire song to compute the accompaniment line. As shown in Figure 4.1, the accompaniment, $\Sigma_{Acc}(t)$, is a transposition of the input singing, $\Sigma_{in}(t)$, by means of resynthesis, with its pitch trajectory computed by

$$f_{Acc}(t) = f_{in}(t) \times J_{Hz} \quad (4.1)$$

where $f_{Acc}(t)$ represents the fundamental frequency of the accompaniment, $f_{in}(t)$ represents the fundamental frequency of the input melody and J_{Hz} represents the interval, or relative distance between the fundamental frequencies, expressed as a multiplier.

J_{Hz} of fourths, fifths and octaves correspond to fractions obtained by substituting $J \in \{-12, -7, -5, 5, 7, 12\}$ into $J_{Hz} = 2^{\frac{J}{12}}$. Hence, with this method,

$$J_{Hz} \in \left\{ \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{3}, \frac{3}{2}, \frac{2}{1} \right\} \quad (4.2)$$

An advantage of this method is that the accompaniment is always synchronized with the melody since the interval, J_{Hz} , is a constant. However, the major trade-off is that it produces both Type-I and Type-II dissonances even when the singer is singing in precise pitch and timing. It also suffers from the negative effects of parallel motion [16].

Audio examples of harmony produced by the 458 method are available in Supplementary Audio S5 and S8. These are produced with Supplementary Audio S3 and S4, respectively, at the input. All Supplementary Audio are listed in Appendix B.

4.1.2.2 The Thirds and Sixths Method

This method, known as 458-II in [7, 54], and used in [58, 59], inherits from the 458 method, but introduces the use of thirds and sixths. Unlike fourths, fifths and octaves, where J_{Hz} is constant, thirds ($J = 3$) may be either 3 semitones

Figure 4.2: The Thirds and Sixths Method (458-II)

$(J_{Hz} = 2^{3/12})$ or 4 semitones ($J_{Hz} = 2^{4/12}$) apart, depending on which produces an $f_{Acc}(t)$ that is an element of the key (set K). This is illustrated in Figure 4.2. Likewise, sixths may either be 8 semitones ($J_{Hz} = 2^{8/12}$) or 9 semitones ($J_{Hz} = 2^{9/12}$) depending on the same criteria. This is expressed by,

$$f_{Acc}(t) = K_{Hz}((\argmin_{n \in \mathbb{Z}} |f_{in}(t) - K_{Hz}(n)|) + J_k) \quad (4.3)$$

where $f_{Acc}(t)$ represents the fundamental frequency of the accompaniment; K_{Hz} represents the set of notes that form the key, expressed in hertz; $f_{in}(t)$ represents the fundamental frequency of the input melody; and J_k represents the interval between $f_{in}(t)$ and $f_{Acc}(t)$ expressed as the number of notes between them in key k .

Thirds and sixths represent 2 and 5 notes away in the key, hence,

$$J_k \in \{-5, -2, 2, 5\} \quad (4.4)$$

This method is effective in avoiding Type-I dissonance. However, it makes no attempt to avoid Type-II ones [7].

Figure 4.3: The Auxiliary Method (AUX)

4.1.2.3 The Auxiliary Method

The auxiliary method, known as the AUX in [7, 54], is used in [58, 59], and an improvement from 458-II. It addresses the issue of Type-II dissonance mentioned in Subsection 4.1.2 by referring to an auxiliary input stream of live or pre-recorded backing music (usually, guitar strumming) for the prevailing chord, C_{Hz} . This method imposes the requirement of having a musician to play a backing instrument as well as the ability of the singer to synchronize with the backing instrument. As shown in Figure 4.3, a typical implementation of the auxiliary method computes its accompaniment pitch by

$$f_{Acc}(t) = C_{Hz}((\argmin_{n \in \mathbb{Z}} |f_{in}(t) - K_{Hz}(n)|) \times J_k, t) \quad (4.5)$$

where $f_{Acc}(t)$ represents the fundamental frequency of the accompaniment; C_{Hz} represents the array of notes that form the chord, expressed in hertz; $f_{in}(t)$ represents the fundamental frequency of the input melody; K_{Hz} represents the array of notes that form the key, expressed in hertz; and J represents the interval between $f_0(t)$ and $f_{Acc}(t)$ expressed as the number of notes between them in key k .

This method is prone to Type-II dissonances whenever the singer's timing drifts

from the auxiliary instruments timing and whenever the system interprets the wrong chord from the auxiliary instrument [7].

4.1.2.4 The Karaoke Method

Figure 4.4: The Karaoke Method (KTV)

The karaoke method is known as KTV in [7, 54], and used in [56, 60]. As described in Figure 4.4, it generates harmony voice from the users input voice in karaoke systems based on $f_{Acc}(t)$ stored within the system, typically in the form of a MIDI file.

Its difference from the AUX method is its use of pre-sequenced MIDI information to replace the auxiliary instrument. However, because the MIDI file is synchronized to the backing track, not the user's singing voice, it still relies on the user's ability to synchronize with the backing music. Whenever the user fails to do this, Type-II dissonances are produced [7].

Audio examples of harmony produced by the KTV method are available in Supplementary Audio S6 and S9. These are produced with Supplementary Audio S3 and S4, respectively, at the input, and Supplementary MIDI Files 1 and 2 as reference. All Supplementary Audio and MIDI Files are listed in Appendix B.

4.1.3 Psychoacoustic Analyses of Existing Methods

With the 458 method, stationary subharmonic tension is always

$$\Delta\hat{t} = 0 \quad (4.6)$$

This means that, when using the 458 method, the effect of stationary harmony, according to Equation 2.20, is effectively always 0.

Furthermore, transitional subharmonic tension commonly encounters

$$\Delta\Delta\hat{t} = \Delta\hat{t}_{preceding} - \Delta\hat{t}_{succeeding} = 0 \quad (4.7)$$

Recalling, from Equation 2.24 in Chapter 2, that the sense of gratification from a chord transition is proportional to subharmonic tension released,

$$\varepsilon \{X_1 \rightarrow X_2\} = 0 \quad (4.8)$$

Thus, with the 458 method, there is empirically no gratification contributed by the addition of harmony to the melody in terms of both stationary and transitional harmony.

The 458-II method uses thirds and sixths instead to overcome these problems. This solves the problem with transitional harmony in 458 since $\Delta\Delta\hat{t} = \Delta\hat{t}_{preceding} - \Delta\hat{t}_{succeeding}$ is no longer zero. However, with stationary harmony, even though $\Delta\hat{t}$ is no longer zero, the algorithm allows for errors that produce dissonances. Figure 4.5 shows 4 tables, from which these dissonances may be computed. The top left, top right, bottom left and bottom right ones show the scenarios of when an interval of a 3rd above, a 3rd below, a 6th above and a 6th below is used, respectively. Chords are organised in rows while notes are organised in columns. The number in each cell represents the number of semitones between the harmonizing note produced by the algorithm and the closest note to which that will harmonize well with the melody. Cells representing correct notes are labeled zero and coded in green. Cells with an error of 1 semitone are coded in orange. Cells with an

		+3 rd s		Notes						
				1	2	3	4	5	6	7
Chords	1	0		0			1			
	2		0		0			2		
	3			0		0			2	
	4	1			0			0		
	5		2			0			0	
	6	0		2				0		

		-3 rd s		Notes						
				1	2	3	4	5	6	7
Chords	1	2		0		0				
	2		2		0			0		
	3			1		0			0	
	4	0			2			0		
	5		0			2			0	
	6	0		0				1		

		+6 th s		Notes						
				1	2	3	4	5	6	7
Chords	1	2		0		0				
	2		2		0		0			
	3			1		0			0	
	4	0			2		0			
	5		0			2			0	
	6	0		0				1		

		-6 th s		Notes						
				1	2	3	4	5	6	7
Chords	1	0		0		1				
	2		0		0			2		
	3			0		0			2	
	4	1			0			0		
	5		2			0			0	
	6	0		2				0		

Figure 4.5: Distortion Tables for 3rds above, 3rds below, 6ths above and 6ths below with the 458-II method

error of 2 semitones are coded in red. It may be computed according to Chapter 2, that a semitone interval has a $\Delta\hat{t}$ of 5.775 and 2 semitones have a $\Delta\hat{t}$ of 10.752. From the tables we can also estimate that such dissonances occur about 1/3 of the time. Of which, 1/3 are semitone dissonances and 2/3 are tone dissonances.

With the AUX and KTV methods, errors are less specific and occur at timing discrepancies, producing $\Delta\Delta\hat{t}$ and $\Delta\hat{t}$ with greater variance and is more dependent on the input.

Figure 4.6: Pitch and Timing Performance Trade-offs of Rule-based, Hybrid and Inference-based Systems against that of the Proposed Strategy

4.2 The Proposed Method

Each existing method mentioned in the previous section has its own shortcomings. The relative performance of each existing method may be illustrated qualitatively as in Figure 4.6. The 458 method, being purely a rule-based method, is always synchronized, but, receiving no auxiliary harmony information, commonly generate pitch errors. The KTV method, being a purely inference-based method using auxiliary harmony information, performs well in pitch generation, but poorly in terms of timing because it relies on human synchronization. The 458-II and AUX methods are hybrid systems and trade off the advantages and disadvantages between the two strategies. However, none of these perform well in both pitch and timing.

The proposed method in this report approaches the problem with the strategy of starting with an inference-based system, and further reduces timing problems by implementing automatic synchronization. The motive is to match the timing performance of the 458 method and the pitch performance of the KTV method with the proposed method.

The novel method proposed in this section was published in IEEE International Conference on Multimedia and Expo in 2011 [7]. Patent was subsequently filed and published in the United States of America in 2012 [54] and in the People's Republic of China in 2012 [61].

The rest of this section goes on to explain each of the aforementioned components in detail and is organized as follows. Subsection 4.2.1 first gives a general explanation of the algorithm. Subsection 4.2.2 then elaborates on the pitch interpretation process. Subsection 4.2.3 continues with the alignment and re-alignment process. Finally, Subsection 4.2.4 closes with the re-synthesis process.

4.2.1 Algorithm

Figure 4.7 shows how the proposed system is composed of its major modules. As can be seen in the diagram, the proposed method synthesizes accompaniment singing based on a database of harmony information. It works as follows:

1. *Interpretation*: The refined pitch is extracted from the user's input singing.
2. *Harmony Extraction*: A MIDI score of the target song is retrieved from the MIDI database and its harmony information is extracted.
3. *Alignment*: The extracted harmony information is time-aligned to the refined pitch from the interpretation module.
4. *Harmony Synthesis*: The aligned harmony information is used to synthesize the accompaniment parts.
5. *Mix Down*: The accompaniment parts are summed with the user's input

Figure 4.7: The Proposed Strategy

singing to produce harmonized singing at the output.

This is described in further detail by Figure 4.8. As illustrated in the figure, the system accepts monophonic singing, $\Sigma_{Mono}(t)$, and harmony information in the form of the melody, $f_{Melody}(\tau)$, and one or more harmony lines (two are shown in this example), $f_{Acc1}(\tau)$ and $f_{Acc2}(\tau)$ in the form of a MIDI file. [147]

First the pitch sequence, $f_{Mono}(t)$, of the input singing, $\Sigma_{Mono}(t)$, is extracted and refined through an interpretation process, to produce $f_{Refined}(t)$. Next, the alignment information between $f_{Refined}(t)$ and the melody from the MIDI file, $f_{Melody}(\tau)$, is computed, acquiring $t(\tau)$, where t and τ are the time meters of the user's singing input and the MIDI file respectively. Then, the intervals between the MIDI melody, $f_{Melody}(\tau)$, and the MIDI accompaniments $f_{Acc1}(\tau)$ and $f_{Acc2}(\tau)$ are

Figure 4.8: The Proposed Method, the Solo-to-Acapella Method (S2A)

evaluated by subtraction to obtain $I_{Acc1}(\tau)$ and $I_{Acc2}(\tau)$. They are then realigned to the user's singing input using alignment information, $t(\tau)$, to produce $I_{Acc1}(t(\tau))$ and $I_{Acc2}(t(\tau))$. Subsequently, the realigned intervals are added to the pitch sequence of the input singing, $f_{Mono}(t)$, to acquire the desired accompaniment's fundamental frequencies, $f_{Acc1}(t)$ and $f_{Acc2}(t)$. The accompaniment singing, $\Sigma_{Acc1}(t)$ and $\Sigma_{Acc2}(t)$, are produced by applying the desired pitch sequences, $f_{Acc1}(t)$ and $f_{Acc2}(t)$, to the spectral characteristics, $H(t)$ and $N(t)$, obtained through analysis of the input singing, $\Sigma_{Mono}(t)$, through a re-synthesis process using a vocoder. These are finally summed together with the input singing, $\Sigma_{Mono}(t)$, to produce the output harmonized singing, $\Sigma_{Harmonized}(t)$.

4.2.2 Pitch Interpretation

The pitch interpretation stage accepts the user's singing voice of the melody line and produces a pitch sequence that is pitch-accurate, in the correct key and octave, and linearized to the MIDI scale [147].

It does this by performing pitch estimation, octave correction, scale translation, tuning drift estimation, key determination, note rounding as well as correction of transient segments in sequential order. These are explained in Subsubsections 4.2.2.1 through 4.2.2.7 respectively.

Figure 4.9: Pitch Interpretation

4.2.2.1 Fundamental Frequency Estimation

Fundamental frequency estimation is performed frame by frame using autocorrelation with a step size of 10ms and a window size of 1024. Each sample of the voice signal $x(t)$ is added to $x(t - \tau)$ to obtain $x(t) + x(t - \tau)$ across a range of (integer) values of t , producing

$$R_{xx}(\tau) = \sum_{t=1}^L x(t) + x(t - \tau) \quad (4.9)$$

where L represents the length of each autocorrelation frame; and τ corresponds to a candidate fundamental period.

This is sampled across the range of possible fundamental periods to find the value of τ corresponding to the largest value of $R_{xx}(\tau)$, which is taken to be fundamental period of the voice segment,

$$\tau_0 = \arg \max_{T_1 < \tau < T_2} R_{xx}(\tau) \quad (4.10)$$

where τ_0 represents the fundamental period of the voice segment, and T_1 and T_2 represent the minimum and maximum fundamental periods respectively.

Finally, the fundamental frequency, f_0 , is found by

$$f_0 = \frac{1}{\tau_0} \quad (4.11)$$

Plots A and B in Figure 4.9 shows the input and output of this stage.

Since peaks are easily confused in noise and given the harmonic nature of the human voice, the top three peaks are selected for each frame, and the Viterbi algorithm is further used to determine the most probable f_0 trajectory. However, it is still common for octave errors to occur in noisy segments. Hence, the purpose of the octave correction stage, which will be introduced in the next subsubsection.

4.2.2.2 Octave Correction

As with all methods of fundamental frequency estimation, the estimated fundamental frequency values are commonly confused with their octaves. That is, $f_{Mono}(t)$ is often confused with $2f_{Mono}(t)$. This has to be addressed at this stage of pitch interpretation.

Thankfully, most octave errors occur mostly in the form of short spurious misinterpretations and are easy to identify. In this stage, segments of octave jumps shorter than a certain threshold number of frames are identified and corrected.

Plots B and C in Figure 4.9 shows the input and output of this stage.

4.2.2.3 Translation to MIDI Note-Number Scale

After the correction of spurious octave misinterpretations, the fundamental frequency is translated from hertz to the MIDI note-number scale. This is performed because the melodies from the MIDI file and the singing voice have to be compared during alignment in Subsection 4.2.3.

Musical notes are logarithmically spaced on the frequency scale and linearly spaced on the MIDI scale. It is, thus, easier to perform logical operations in the MIDI scale. Logical operations are required in Subsections 4.2.2.4 through 4.2.2.7 for tuning drift estimation, key determination, the correction of accidentals and the correction of transient segments. Hence, it is advantageous to perform the translation at this point.

In MIDI, notes are represented by integer numbers corresponding to their position on the piano keyboard. Numbers 0 through 127 represent the notes C-1 through G9 and adjacent numbers are a semitone apart. Since note number 69 (A4) cor-

responds to 440Hz, they are translated using the formula

$$n_{MIDI} = 9 + \left(12 \log_2 \left(f_{Hz} \times \frac{32}{440} \right) \right) \quad (4.12)$$

Although the translated pitch sequence map onto the MIDI scale, its floating point values are preserved to exploit the decimal resolution later in Subsections 4.2.2.4 through 4.2.2.6. Plots C and D in Figure 4.9 shows the input and output of the this stage.

4.2.2.4 Estimation of Overall Tuning Drift

Figure 4.10: Estimation of Overall Tuning Drift, Δf_0

The role of the overall tuning drift estimation module is to compute the mean distance of the input singing to the nearest key. Since the system accepts singing voice that does not follow a backing track, the user is not singing to any particular key. The overall tuning drift is thus required for key determination and note rounding. It is performed by computing the unwrapped average across the decimal

part of the sequence of pitch values.

Figure 4.10, shows how tuning drift estimation is performed, explaining the process with a series of note distributions. As shown in the figure, the notes enter this stage from Section 4.2.2.3 represented in the perpetual chromatic scale. [3, 147] They are first mapped to the chroma scale, where the same note in different octaves (e.g. g_3 and g_4) are represented by the same number,

$$D_{Chroma}(n_{Chroma}) = \frac{D_{Chromatic}(n_{MIDI})}{12} - \lfloor \frac{D_{Chromatic}(n_{MIDI})}{12} \rfloor \quad (4.13)$$

where n_{Chroma} and n_{MIDI} represent notes on the chroma and MIDI scale [3] respectively, $D_{Chroma}(n_{Chroma})$ and $D_{Chromatic}(n_{MIDI})$ represent the chromatic and chroma distributions [3] of notes respectively and $\lfloor \rfloor$ denotes the mathematical floor function.

The ends of the chroma scale wrap around (i.e. one note higher than 11 is 0 and notes 12 semitones apart are synonymous) as octave information is discarded. This chroma is later used for key determination in Subsection 4.2.2.5, but here, it is further mapped onto the cents scale by taking the decimal part of each note,

$$D_{Cents}(n_{Cents}) = D_{Chroma}(n_{Chroma}) - \lfloor D_{Chroma}(n_{Chroma}) \rfloor \quad (4.14)$$

where n_{Cents} represents the decimal parts of the notes that are musically called cents and $D_{Cents}(n_{Cents})$ represents the distribution of the cents part of the notes.

Notes in the cents scale may initially wrap from 0 to 1, centering about 0.5, or from -0.5 to 0.5, centering about 0. The circular average is taken by first centering the scale at the peak distribution, and then computing the mean. For example, the circular average of 0.49 and 0.51, is 0.5; but that of 0.99 and 0.01 is 0. This value is further wrapped by subtracting one if it is more than 0.5, so that the absolute distance to 0 remains minimal.

Plots D and E in Figure 4.9 shows the input and output of the tuning drift estimation stage.

4.2.2.5 Key Determination

In order to round all notes to the nearest note in the key, the key the singer is singing in has to be known. Hence it is necessary, to perform key determination at this point.

As shown in Figure 4.11, the notes represented in the chroma scale obtained from the tuning drift estimation stage Section 4.2.2.4 are shifted by Δf_0 to obtain $f_{Centered} = f_0 - \Delta f_0$. After which, they are rounded to the nearest note, effectively tallying the number of notes within the range of each whole note,

$$C(n) = \int_{m=n-0.5}^{n+0.5} D(m) \delta n \quad (4.15)$$

where $D(m)$ is the distribution of chroma distribution of note, and $n + 0.5$ and $n - 0.5$ represents the boundary between two notes. $C(n)$ may be understood as a count of number of times each note appears throughout the song.

Subsequently, the score of each of the 12 possible keys [3] being the key of the input singing is computed by

$$\mathcal{L}(\kappa|C(n)) = \sum_{n=0}^{11} a_n C(n) \quad (4.16)$$

where $\mathcal{L}(\kappa|C(n))$ represents the score of each key, κ , being the key of the input singing, given $C(n)$; and a_n is the weighting coefficient that is determined by how often a particular note, n , appears in each key. If statistics are unavailable, it is sufficient to use 1 for all notes that are part of the key (i.e. $n \in K$), and -1 for all notes that don't (i.e. $n \notin K$).

Having computed this, the key with the highest score is determined to be the key

 Figure 4.11: Determination of Key, k

of the song

$$k = \operatorname{argmax}_{\kappa} \mathcal{L}(\kappa|C(n)) \quad (4.17)$$

Plots E and F in Figure 4.9 shows the input and output of the key determination stage.

4.2.2.6 Note Rounding

Figure 4.12: Note Rounding

This process rounds each note to the nearest note in the key. Figure 4.12 explains how this is done in by looking at the note distributions.

First, having determined the tuning drift, Δf_0 , in Section 4.2.2.4, the output of the tuning drift estimation stage is subtracted from all notes, updating $f_0(t)$ to be

$$f_0(t) = f_{pre-centered}(t) - \Delta f_0 \quad (4.18)$$

Next, having determined the song key, k , in Section 4.2.2.5, the notes of the key are evaluated by

$$K = \{k + [0, 2, 4, 5, 7, 9, 11] + 12i\} \quad (4.19)$$

where

$$i \in \mathbb{Z} \quad (4.20)$$

Finally, they are rounded to the nearest note, effectively tallying the number of notes within the range of each whole note to produce

$$C(K(p)) = \int_{m=K(p-0.5)}^{K(p+0.5)} D(m) \delta K(p) \quad (4.21)$$

Plots F and G in Figure 4.9 shows the input and output of the note rounding stage.

4.2.2.7 Rule-based Transient Segment Correction

The last stage of pitch interpretation corrects any transient segments left in the pitch sequence. Some examples of transient segments include:

Under-pitching It is common for singers to commence phonation at a pitch lower than the target pitch. After which, the singer brings the pitch up to the target pitch. Sometimes this effect is deliberated as part of the song expression.

Over-pitching Conversely to under-pitching, singers can over-pitch as well. However, this effect is seldom deliberated.

Glides When transiting between to notes without an unvoiced phone between them, it is common for singers to smoothly change pitch by gliding through the pitch in between the notes.

Since transient segments are usually not transcribed by MIDI, they have to be identified and corrected at the the pitch interpretation stage in order to better compare the output with the MIDI pitch sequence during alignment. After which, the MIDI-compatible pitch sequence of the user's singing voice is finally produced.

Plots G and H in Figure 4.9 shows the input and output of the note rounding

stage. The MIDI-compatible pitch sequence, $f_{Refined}(t)$, is now ready to be passed to the alignment stage.

4.2.3 The Alignment Process

The alignment process may be broken down into two steps:

- Step one is to find the timing relationship between the input melody and the lead melody from the MIDI file of the target song. These two melodies are unsynchronized and regarded to be running on different time meters.
- Step two is to use the timing relationship found in step one to modify the interval sequence(s) between the lead melody and each accompaniment from the MIDI file of the target song. This effectively synchronizes each interval sequence to the target singing.

With reference to Figure 4.8, the two steps use the timing relationship between $f_{Refined}(t)$ and $f_{Melody}(\tau)$ to time-warp the interval sequences $I_{Acc1}(\tau)$ and $I_{Acc2}(\tau)$ to be in the same meter as $f_{Refined}(t)$. This produces the aligned interval sequences $I_{Acc1}(t)$ and $I_{Acc2}(t)$ respectively.

In step one, the alignment relationship, $t(\tau)$, between the MIDI melody, $f_{Melody}(\tau)$ and MIDI-compatible pitch sequence, $f_{Refined}(t)$ is found by dynamically time warping. $f_{Melody}(\tau)$ and $f_{Refined}(t)$ both contain sequentially similar fundamental frequency events but are in the different meters, τ and t , respectively.

In step two, the alignment information, $t(\tau)$, is used to realign the interval sequences, $I_{Acc1}(\tau)$ and $I_{Acc2}(\tau)$, to match the meter of $f_{Refined}(t)$ to produce

$$I_{Acc1}(t) = I_{Acc1}(t(\tau)) \quad (4.22)$$

and

$$I_{Acc2}(t) = I_{Acc2}(t(\tau)) \quad (4.23)$$

This will allow it to be added to $f_{Mono}(t)$ to produce $f_{Acc1}(t)$ and $f_{Acc2}(t)$.

The following parts of this section is organized as follows. Section 4.2.3.1 explains how step one is performed using a process called dynamic time warping. Section 4.2.3.2 explains how step two realigns the interval sequences to produce $I_{Acc1}(t)$ and $I_{Acc2}(t)$.

4.2.3.1 Dynamic Time Warping

Figure 4.13: The Dynamic Time Warping Process

Figure 4.13 demonstrates the process of dynamic time warping,

$$t(\tau) = DTW(f_{Refined}(t), f_{Melody}(\tau)) \quad (4.24)$$

The green plot on the left represents MIDI melody, $f_{Melody}(\tau)$, transposed to match

the key of the input singing, with time, τ , flowing from bottom to top, and higher pitch represented by points on the right and lower ones by on the left. Similarly, the red plot at the bottom represents MIDI-compatible pitch sequence of the input singing voice, $f_{Refined}(t)$. With time, t , flowing from left to right, and higher pitch represented by points at the top and lower ones by points at the bottom.

The pitch at each point along the MIDI melody, $f_{Melody}(\tau)$, is compared with the pitch at each point along that of normalized input singing voice, $f_{Refined}(t)$, and their difference forms a matrix, where every point on the matrix is given by

$$dtw(t, \tau) = f_{Refined}(t) - f_{Melody}(\tau) \quad (4.25)$$

To visualize this matrix, Figure 4.13 uses shades of green to represent points where the pitch on the MIDI melody, $f_{Melody}(\tau)$, is higher than that of the singer's $f_{Refined}(t)$, with darker shades of green to represent a greater mismatch. Similarly, it uses shades of red to represent points where the pitch on the MIDI melody, $f_{Melody}(\tau)$, is lower than that of the singer's $f_{Refined}(t)$. Again, darker shades of red represent a greater mismatch. Cyan represents where points on both pitch sequences match perfectly.

Starting from the bottom left and working towards the top right, the matrix is traversed with the objective of traversing an adjacent point with the best match $\text{argmin}_{t, \tau}(|dtw(t, \tau)|)$ at each step.

Finally, the collective indices of points traversed is obtained. This describes which time frame, τ , corresponds to which time frame, t , thus arriving at the alignment information, $t(\tau)$.

4.2.3.2 Realignment

With the alignment information, $t(\tau)$, the interval sequences, $I_{Acc1}(\tau)$ and $I_{Acc1}(\tau)$, may be warped to follow the meter of the input singing voice, $f_{Refined}(t)$. This is carried out by inserting fill-in frames to and removing redundant frames from the

Figure 4.14: The Realignment Process

original time-sequence of intervals as shown in Figure 4.14.

The figure shows examples of frames $I_{Acc}(\tau)$ being omitted when

$$t(\tau - 1) = t(\tau + 1) + 1 \quad (4.26)$$

and frames $I_{Acc}(t(\tau))$ being inserted when

$$t(\tau) = t(\tau + 1) + 2 \quad (4.27)$$

After alignment, the warped sequence of intervals, $I_{Acc1}(t(\tau))$ and $I_{Acc2}(t(\tau))$, are added to the scaled pitch sequence, $f_{Mono}(t)$, to obtain accompaniment frequencies, $f_{Acc1}(t)$ and $f_{Acc2}(t)$. These are used for the synthesis of the output in Section 4.2.4.

4.2.4 Re-Synthesis

As shown in Figure 4.8, the re-synthesis stage generates the output accompaniment voices, $\Sigma_{Acc1}(t)$ and $\Sigma_{Acc2}(t)$, given the input singing voice, $\Sigma_{Mono}(t)$, and

the accompaniment pitch sequences, $f_{Acc1}(t)$ and $f_{Acc2}(t)$. The vocoder used for analysis and synthesis is Tandem-STRAIGHT [148]. As described in Figure 4.8, re-synthesis is performed as follows.

1. Vocoder analysis is performed on the input, $\Sigma_{Mono}(t)$, to obtain a set of descriptive parameters, $P(t)$.
2. The $f_0(t)$ parameter amongst the descriptive parameters are substituted by the accompaniment pitch sequences, $f_{Acc1}(t)$ and $f_{Acc2}(t)$, to produce $P_{Acc1}(t)$ and $P_{Acc2}(t)$.
3. The accompaniment voices, $\Sigma_{Acc1}(t)$ and $\Sigma_{Acc2}(t)$ are re-synthesized according to these new features.

At the end of the re-synthesis process, the accompaniment voices are finally summed with the input singing voice, $\Sigma_{Mono}(t)$, to produce the target harmonized singing voices, $\Sigma_{Harmonized}(t)$.

4.3 Experiment and Evaluation

Our proposed method, Solo to A Capella, or S2A [7, 54], was evaluated against the 458 method of Section 4.1.2.1 (using fifths) as well as the KTV method of Section 4.1.2.4. These two methods are chosen for comparison because the 458 method is known to perform the best amongst existing methods in terms of continuity and the KTV method is known to perform the best amongst existing methods in terms of consonance. Fifths are chosen for the 458 mode because that offers better trade-offs than fourths and octaves. The results are evaluated over a series of spectrogram plots in Subsection 4.3.1 and subjective listening tests in Subsection 4.3.2.

Audio examples of harmony produced by the proposed method are available in Supplementary Audio S7 and S10. These are produced with Supplementary Audio S3 and S4, respectively, at the input, and Supplementary MIDI Files 1 and 2 as reference. All Supplementary Audio and MIDI Files are listed in Appendix B.

4.3.1 Spectrograms

Figure 4.15 compares the spectrograms for the harmonization of the song “Twinkle Twinkle Little Star” using the three methods against that of the human voice. The segment compared is the stanza “How I wonder what you are”.

Spectrograms A, B, C and D in Figure 4.15 are those of the 458 method, the KTV method, the proposed method (S2A) and the human voice respectively. Landmarks have been identified and marked out by hand. In the spectrograms, marker ‘A’ identifies the fundamental frequency of the melody and marker ‘B’ identifies the fundamental frequency of an accompaniment. Marker ‘C’ cites an example of the undesirable parallel motion described in Subsection 4.1.2.1 and [16]. Marker ‘D1’ identifies regions of potential Type-I and Type-II dissonances due to key or chord ignorance. Marker ‘D2’ identifies regions of dissonance due to timing inaccuracies. Marker ‘E’ indicates harmonic discontinuities due to misalignment. The green and yellow ‘+’s, and red ‘-’s compliment ‘D1’ and ‘D2’ by indicating regions of consonance, coincidental consonance, and dissonance respectively. (Coincidental consonance refers to uncommon situations where the alignment is completely off but the harmony just happens to be correct.) The red ‘-’s indicate dissonance.

It is observable from Spectrogram A that an extended part of the phrase synthesized by the 458 method is plagued by both Type-I and II dissonances (marked by D1). The effects of parallel motion is also observable in Spectrogram A, where the third harmonic of the fundement marked A and the second harmonic of the fundement marked ‘B’ completely align at ‘C’. It is observable from Spectrogram B, dissonances due to timing inaccuracies (marked by D2), as well as discontinuities in the harmonics (marked by E) at multiple locations across the phrase produced using the KTV method. Spectrogram C shows the spectrogram of the utterance from the S2A method, which has only a very short segment of Type-II dissonance (marked by D2) and one location with discontinuity (marked by E). It is also the one that looks most identical to Spectrogram D, which, being the spectrogram of the actual human voice, is the target output.

Figure 4.15: Spectrogram of the 458, KTV and S2A methods respectively, against that of the human voice

Table 4.2: Score across 11 vocal experts on consonance / harmony

	458		KTV		S2A	
	mean	var	mean	var	mean	var
Brahms’ Cradlesong	5.64	1.96	2.91	1.38	8.73	1.01
Twinkle Twinkle Little Star	7.64	2.34	3.64	2.16	9.27	1.01
<i>Overall mean as a percentage</i>	66.36%		32.73%		90.00%	

Table 4.3: Score across 11 vocal experts on smoothness of transition

	458		KTV		S2A	
	mean	var	mean	var	mean	var
Brahms’ Cradlesong	4.73	2.24	2.73	1.35	5.64	1.75
Twinkle Twinkle Little Star	5.09	1.87	3.64	2.50	6.73	2.41
<i>Overall mean as a percentage</i>	49.09%		31.82%		61.82%	

Thus, it is observable by spectrogram analysis that S2A is the best performing method. However, music is aural and subjective, hence the need to verify and quantify this with subjective listening tests.

4.3.2 Subjective Listening Tests

Two songs, namely, “Brahms’ Cradlesong” and “Twinkle Twinkle Little Star”, were synthesized using the 458, KTV and S2A methods. Across the 3 methods, the first song was synthesized using two harmony lines while the second song was synthesized with just one. A series of subjective listening tests were carried out separately on a group of vocal experts and a group of casual listeners.

4.3.2.1 Vocal Experts

In the first test, 11 vocal experts were tasked to listen to the 6 songs and evaluate them in terms of consonance and smoothness of continuity (i.e. smoothness of transition). These two characteristics were explicitly specified to evaluate how well the pitch-timing resilience trade-offs mentioned in 4.6 have been overcome: The 458 method, deriving vocal harmony voice by effectively transposing lead vocals by a fixed interval throughout the song, is the best existing method for timing

Table 4.4: Score across 12 casual listeners on pleasantness / naturalness

	458		KTV		S2A	
	mean	var	mean	var	mean	var
Brahms’ Cradlesong	6.67	1.78	6.00	2.17	8.25	1.76
Twinkle Twinkle Little Star	5.50	2.11	4.67	2.27	5.75	2.05
<i>Overall mean as a percentage</i>	60.83%		53.33%		70.00%	

performance; The KTV method on the other hand, deriving its accompaniment from MIDI whilst relying on manual synchronization, is the best existing method for pitch performance.

Table 4.2 and Table 4.3 show their average scores for each of the songs on a scale of 1 to 5, with 1 being worst and 5 being best. The scores are normalized to 10 in the table.

In terms of continuity, as anticipated, the results show that the 458 method is perceived to sound better than the KTV method. Nevertheless, S2A scores significantly better than both.

In terms of consonance, it was unexpected for the KTV method to perform worse than the 458 method. This may be due to 2 reasons. First, the distortion produced by the continuity problems in the KTV method is also perceived to be dissonant, contributing to lowering its consonance score. Second, the unusual absence of a backing track causes Type-I dissonances with 458 to sound more tolerable because it has less sounds to be dissonant with. In any case, S2A is still perceived to sound significantly better than both 458 and KTV in terms of consonance.

4.3.2.2 Casual Listeners

In the second test, 12 casual listeners were tasked to listen to the 6 songs. Because non-experts are not expected to be as attentive to aural detail, we tasked them to give a single score to each song on a scale of 1 to 10 on how pleasant and natural

they thought each song sounded. Table 4.4 lists their ratings on a scale of 1 to 10, with 1 being worst and 10 being best.

Since casual listeners are less attentive to aural detail, the scores are less distinct. Nevertheless, the casual test still shows that the harmonized voices by the S2A method produces the best results.

4.4 Conclusion

This chapter covered the synthesis of singing harmony. Four existing methods of singing harmony synthesis were presented, namely, the 458, 458-II, Auxiliary and Karaoke methods. A method was proposed to overcome the trade-offs across pitch and timing errors in existing methods. Spectrograms revealed that harmonized voices produced by this method contain the least dissonances. Subjective listening tests also showed that harmonized voices produced by this method are perceived to be the best sounding, both by vocal experts and by casual listeners.

Chapter 5

Conclusion and Future Work

Section 5.1 highlights the contributions of this thesis while section 5.2 describes potential future work.

5.1 Conclusion

This thesis presents our work in the psychoacoustics and synthesis of singing harmony.

In Chapter 2 the notion of interharmonic and subharmonic modulations was proposed as a psychophysical basis for both stationary and transitional harmony. Interharmonic and subharmonic tensions were proposed as consonance-dissonance [32] measures with stationary harmony while transitional subharmonic tensions were proposed as a measure of resolution [1, 16] with transitional harmony. Correlations with perceptual [82] and chord-use [26] statistics show their relation to perceptual tensions [1, 14, 15, 107] and resolutions [1, 16] in music. Work in this chapter was used to address several fundamental questions of harmony.

Chapter 3 focused on the synthesis of human singing. The chapter covered conventional singing synthesis systems and presented an overview of different methods of synthesis such as formant synthesis, physical modeling (articulatory synthesis), unit selection (concatenative synthesis), parametric synthesis, wavenets and

wavetable synthesis was made. Apart from singing synthesis, the chapter also covered an overview of singing resynthesis systems such as pitch-shifters, autotune, the vocoder effect, speech-to-singing systems and harmonizers. After which, a real-time wavetable synthesis system, SERAPHIM was proposed. An evaluation was made which concluded SERAPHIM to be the best amongst known systems in terms of overcoming phone coverage and realtime capability trade-offs. Subjective listening tests also showed that SERAPHIM outperformed LIMSI's and Yamaha's system by almost double and over 30% respectively.

In Chapter 4, four existing methods of harmony synthesis were briefly described and a novel method was proposed. This was a pitch-accurate inference-based system, which overcame traditional pitch-timing error trade-offs by improving existing pitch accurate methods with machine-synchronization. Harmony information from a MIDI database was aligned to the user's input singing voice by dynamic time warping. This information was then used to modify the input to create accompaniment voices, which were added back to the input to create harmonized singing. The proposed system was evaluated against two other methods by means of subjective listening tests. Experiment results showed that both vocal experts and casual listeners preferred harmony synthesized by the proposed system.

5.2 Future Work

Having understood how psychophysical attributes of music relates to aesthetic appeal in humans from the work in Chapter 2, it would be ideal if there could be a way to apply it directly to synthesize the audio waveform of harmonized singing in Chapter 4. Unfortunately, the aesthetic component of music is not so straightforward. Often, for example, dissonance is sometimes desirable to develop a plot in a musical storyline. Nevertheless, in future work, means to overcome this could be explored. This could encompass fields such as stochastic composition, music cognition, and end-to-end singing synthesis.

References

- [1] N. D. Cook and T. X. Fujisawa, “The Psychophysics of Harmony Perception: Harmony is a Three-Tone Phenomenon,” 2006.
- [2] E. Terhardt, G. Stoll, and M. Seewann, “Algorithm for Extraction of Pitch and Pitch Saliency from Complex Tonal Signals,” *The Journal of the Acoustical Society of America*, vol. 71, no. 3, pp. 679–688, 1982.
- [3] J.-P. Rameau, *Treatise on Harmony*. Courier Corporation, 1722.
- [4] R. Parncutt, *Harmony: A Psychoacoustical Approach*. Springer-Verlag, 1989.
- [5] D. Harrison, *Harmonic Function in chromatic Music: A Renewed Dualist Theory and an Account of its Precedents*. University of Chicago Press, 1994.
- [6] R. Fink, *The Origin of Music: A Theory of the Universal Development of Music*. The Greenwich Meridian Co., 1981.
- [7] P. Y. Chan, M. Dong, S. W. Lee, L. Cen, and H. Li, “Solo to A Capella Conversion - Synthesizing Vocal Harmony from Lead Vocals,” in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2011.
- [8] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,” in *SSW*, p. 125, 2016.
- [9] Z. Wu, O. Watts, and S. King, “Merlin: An Open Source Neural Network Speech Synthesis System,” in *SSW*, pp. 202–207, 2016.

REFERENCES

- [10] P. F. Broman and N. A. Engebretsen, “What Kind of Theory is Music Theory?: Epistemological Exercises in Music Theory and Analysis,” *Acta Universitatis Stockholmiensis*, 2008.
- [11] M. Brown and D. J. Dempster, “The Scientific Image of Music Theory,” *Journal of Music Theory*, vol. 33, no. 1, pp. 65–106, 1989.
- [12] K. J. Pallesen, E. Brattico, C. Bailey, A. Korvenoja, J. Koivisto, A. Gjedde, and S. Carlson, “Emotion processing of major, minor, and dissonant chords: a functional magnetic resonance imaging study,” *Annals of the New York Academy of Sciences*, vol. 1060, no. 1, pp. 450–453, 2005.
- [13] M. Farbood and B. Schöner, “Analysis and Synthesis of Palestrina-Style Counterpoint using Markov Chains,” in *ICMC*, 2001.
- [14] E. Bigand, R. Parncutt, and F. Lerdahl, “Perception of Musical Tension in Short Chord Sequences: The Influence of Harmonic Function, Sensory Dissonance, Horizontal Motion, and Musical Training,” *Perception & Psychophysics*, vol. 58, no. 1, pp. 125–141, 1996.
- [15] C. K. Madsen and W. E. Fredrickson, “The Experience of Musical Tension: A Replication of Nielsen’s Research using the Continuous Response Digital Interface,” *Journal of Music Therapy*, vol. 30, no. 1, pp. 46–63, 1993.
- [16] P. I. Tchaikovsky, *Guide to the Practical Study of Harmony*. Courier Corporation, 1872/2005.
- [17] F. E. Maus, “Music as Narrative,” *Indiana theory review*, vol. 12, pp. 1–34, 1991.
- [18] D. L. Bowling and D. Purves, “A Biological Rationale for Musical Consonance,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 36, pp. 11155–11160, 2015.
- [19] A. Roberts, J. Engel, C. Hawthorne, I. Simon, E. Waite, S. Oore, N. Jaques, C. Resnick, and D. Eck, “Interactive Musical Improvisation with Magenta,” in *Proc. Neural Information Processing Systems*, 2016.

REFERENCES

- [20] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, “This time with feeling: learning expressive musical performance,” *Neural Computing and Applications*, pp. 1–13, 2018.
- [21] J.-S. Kim, “Deep Learning driven Jazz Generation using Keras & Theano,” 2017.
- [22] F. Liang, *Bachbot: Automatic Composition in the Style of Bach Chorales*. PhD thesis, Masters thesis, University of Cambridge, 2016.
- [23] F. Pachet, P. Roy, and F. Ghedini, “Creativity through Style Manipulation: The Flow Machines Project,” in *2013 Marconi Institute for Creativity Conference, Proc.(Bologna, Italy, vol. 80*, 2013.
- [24] A. Nayebi and M. Vitelli, “Gruv: Algorithmic Music Generation using Recurrent Neural Networks,” *Course CS224D: Deep Learning for Natural Language Processing (Stanford)*, 2015.
- [25] D. Tymoczko, *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. Oxford University Press, 2010.
- [26] D. Tymoczko, “A Study on the Origins of Harmonic Tonality,” *Paper delivered to the national meeting of the Society for Music Theory, Indianapolis*, 2014.
- [27] K. J. Hsü and A. J. Hsü, “Fractal Geometry of Music,” *Proceedings of the National Academy of Sciences*, vol. 87, no. 3, pp. 938–941, 1990.
- [28] N. McLachlan, D. Marco, M. Light, and S. Wilson, “Consonance and Pitch,” *Journal of Experimental Psychology: General*, vol. 142, no. 4, p. 1142, 2013.
- [29] B. V. Rivera, “Theory Ruled by Practice: Zarlino’s Reversal of the Classical System of Proportions,” *Indiana theory review*, vol. 16, pp. 145–170, 1995.
- [30] R. Plomp and W. J. M. Levelt, “Tonal Consonance and Critical Bandwidth,” *The journal of the Acoustical Society of America*, vol. 38, no. 4, pp. 548–560, 1965.

REFERENCES

- [31] P. N. Johnson-Laird, O. E. Kang, and Y. C. Leong, “On Musical Dissonance,” *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 1, pp. 19–35, 2012.
- [32] H. Von Helmholtz and A. J. Ellis, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. London: Longmans, Green and Company, 1875.
- [33] H. L. Goodman and Y. E. Lien, “A Third Century AD Chinese System of Di-Flute Temperament: Matching Ancient Pitch-Standards and Confronting Modal Practice,” *The Galpin Society Journal*, pp. 3–24, 2009.
- [34] G. J. Cho, *The Discovery of Musical Equal Temperament in China and Europe in the Sixteenth Century*, vol. 93. Edwin Mellen Press, 2003.
- [35] T. Christensen and J.-P. Rameau, “Eighteenth-Century Science and the “Corps Sonore:” The Scientific Background to Rameau’s “Principle of Harmony”,” *Journal of Music Theory*, vol. 31, no. 1, pp. 23–50, 1987.
- [36] I. Shapira Lots and L. Stone, “Perception of Musical Consonance and Dissonance: An Outcome of Neural Synchronization,” *Journal of The Royal Society Interface*, vol. 5, no. 29, pp. 1429–1434, 2008.
- [37] P. Lalitte, “The Theories of Helmholtz in the Work of Varese,” *Contemporary Music Review*, vol. 30, no. 5, pp. 327–342, 2011.
- [38] E. G. Schellenberg and S. E. Trehub, “Frequency Ratios and the Perception of Tone Patterns,” *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 191–201, 1994.
- [39] H. Le, “A Brief History of Medieval Music,” *The Histories*, vol. 3, no. 1, p. 11, 2016.
- [40] L. Guelker Cone, “The Unaccompanied Choral Rehearsal: Consistent rehearsing without accompaniment can improve a choir’s sight-singing, intonation, sense of ensemble, and ability to respond to conducting gestures,” *Music educators journal*, vol. 85, no. 2, pp. 17–22, 1998.

- [41] H. Kenmochi, “VOCALOID and Hatsune Miku phenomenon in Japan,” in *Interdisciplinary Workshop on Singing Voice*, 2010.
- [42] L. K. Le, “Examining the Rise of Hatsune Miku: the First International Virtual Idol,” *The UCI Undergraduate Research Journal*, pp. 1–12, 2013.
- [43] K. Itoh, S. Suwazono, and T. Nakada, “Cortical Processing of Musical Consonance: An Evoked Potential Study,” *Neuroreport*, vol. 14, no. 18, pp. 2303–2306, 2003.
- [44] G. M. Bidelman, “The Role of the Auditory Brainstem in Processing Musically Relevant Pitch,” *Frontiers in psychology*, vol. 4, p. 264, 2013.
- [45] P. Y. Chan, M. Dong, H. Li, *et al.*, “The Science of Harmony: A Psychophysical Basis for Perceptual Tensions and Resolutions in Music,” *Research*, vol. 2019, pp. Article ID 2369041, 22 pages, 2019. <https://spj.sciencemag.org/research/aip/2369041/>.
- [46] Audio Acoustique LIMSI CNRS, “Cantor Digitalis - JDEV 2013, Palaiseau - Cantate 2.0,” 2013.
- [47] DigInfo TV, “Yamaha Vocaloid Keyboard - Play Miku Songs Live! #DigInfo,” 2012.
- [48] J. Bonada, *Voice Processing and synthesis by Performance Sampling and Spectral Models*. PhD thesis, La Universitat Pompeu Fabra, 2008.
- [49] H. Kenmochi, “Singing Synthesis as a New Musical Instrument,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5385–5388, 2012.
- [50] H. Kenmochi and H. Ohshita, “VOCALOID - Commercial Singing Synthesizer based on Sample Concatenation.,” in *Interspeech*, no. August, pp. 3–4, 2007.
- [51] P. Y. Chan, M. Dong, G. X. H. Ho, and H. Li, “SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-Time Speech and Singing Applications on Mobile Platforms,” in *INTERSPEECH*, pp. 1225–1229, 2016.

REFERENCES

- [52] P. Y. Chan, M. Dong, G. X. H. Ho, and H. Li, “SERAPHIM Live!-Singing Synthesis for the Performer, the Composer, and the 3D Game Developer,” in *INTERSPEECH*, pp. 1966–1967, 2016.
- [53] P. Y. Chan, M. Dong, and H. Li, “SERAPHIM: A Wavetable Synthesis System with 3D Lip Animation for Real-time Speech and Singing Applications on Mobile Platforms,” SG/P/2016002, 2016.
- [54] P. Y. Chan, M. Dong, L. Cen, and S. W. Lee, “Auto-Synchronous Singing Harmonizer,” US20120234158, 2012.
- [55] TC Helicon, *Harmony Singer*. <http://www.tc-helicon.com/en/products/harmony-singer/>, Date Accessed: 21 Mar 2016.
- [56] Antares Audio Technologies, *Harmony Engine Evo*. http://www.antarestech.com/products/detail.php?product=Harmony_Engine_Evo_4, Date Accessed: 10 Feb 2015.
- [57] Roland Corporation, *VE-20 Vocal Performer*. <http://www.bossus.com/products/ve-20/>, Date Accessed: 10 Feb 2015, 2015.
- [58] Electro-Harmonix, *Voice Box*. <http://www.ehx.com/products/voice-box>, Date Accessed: 02 Feb 2015, 2004.
- [59] Harman International Industries, *DigiTech Live Harmony*. <http://digitech.com/en/products/live-harmony>, Date Accessed: 10 Feb 2015, 2014.
- [60] Kageyama, M. Yasuo, and Hiroshi, “Karaoke Apparatus Using Frequency Of Actual Singing Voice to Synthesize Harmony Voice From Stored Voice Information,” 1996.
- [61] P. Y. Chan, M. Dong, L. Cen, and S. W. Lee, “Harmony Synthesizer and Method for Harmonizing Vocal Signals,” PRC Patent CN102682762, Filed: 15 Mar 2012.
- [62] P. Hindemith, A. Mendel, and O. Ortmann, *The Craft of Musical Composition*, vol. 2. Schott, 1941.

REFERENCES

- [63] J. Sauveur, *Principes D'Acoustique et de Musique: Ou, Système général des Intervalles des Sons*. Editions Minkoff, 1701.
- [64] G. Pont, "Philosophy and Science of Music in Ancient Greece," *Nexus Network Journal*, vol. 6, no. 1, pp. 17–29, 2004.
- [65] E. Wellesz and J. A. Westrup, *Ancient and Oriental Music*, vol. 1. Oxford University Press, 1957.
- [66] J. M. Barbour, *Tuning and Temperament: A Historical Survey*. Courier Corporation, 2004.
- [67] A. Gräf, "On Musical Scale Rationalization," in *ICMC*, 2006.
- [68] H. E. White and D. H. White, *Physics and Music: the Science of Musical Sound*. Courier Corporation, 2014.
- [69] G. Dillon, "Calculating the Dissonance of a Chord according to Helmholtz Theory," *The European Physical Journal Plus*, vol. 128, no. 8, p. 90, 2013.
- [70] Y. I. Fishman, I. O. Volkov, M. D. Noh, P. C. Garell, H. Bakken, J. C. Arezzo, M. A. Howard, and M. Steinschneider, "Consonance and Dissonance of Musical Chords: Neural Correlates in Auditory Cortex of Monkeys and Humans," *Journal of Neurophysiology*, vol. 86, no. 6, pp. 2761–2788, 2001.
- [71] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, vol. 22. Springer Science & Business Media, 2013.
- [72] W. A. Sethares, "Local Consonance and the Relationship between Timbre and Scale," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1218–1228, 1993.
- [73] R. Parncutt, "Revision of Terhardt's Psychoacoustical Model of the Root(s) of a Musical Chord," *Music Perception: An Interdisciplinary Journal*, vol. 6, no. 1, pp. 65–93, 1988.
- [74] W. Hutchinson and L. Knopoff, "The Acoustic Component of Western Consonance," *Journal of New Music Research*, vol. 7, no. 1, pp. 1–29, 1978.

REFERENCES

- [75] A. Kameoka and M. Kuriyagawa, “Consonance Theory Part I: Consonance of Dyads,” *The Journal of the Acoustical Society of America*, vol. 45, no. 6, pp. 1451–1459, 1969.
- [76] C. Stumpf, “Konsonanz und Dissonanz, Beiträge zur Akustik und Musikwissenschaft 1,” *Leipzig: Johann Ambrosius Barth*, 1898.
- [77] W. S. B. Woolhouse, *Essay on Musical Intervals, Harmonics, and the Temperament of the Musical Scale*. 1835.
- [78] E. Bigand and B. Poulin-Charronmat, “Are we “experienced listeners”? A Review of the Musical Capacities that Do Not Depend on Formal Musical Training,” *Cognition*, vol. 100, no. 1, pp. 100–130, 2006.
- [79] T. M. Fiore, “Music and Mathematics,” *Recuperado de: <http://www-personal.umd.umich.edu/tmfiore/1/musictotal.pdf>*, 2007.
- [80] R. Scruton, *The Aesthetics of Music*. Oxford University Press, 1999.
- [81] P. F. Broman, “Music Theory Art, Science, or What?,” *What Kind of Theory Is Music Theory?*, p. 17, 2007.
- [82] F. Stolzenburg, “Harmony Perception by Periodicity Detection,” *Journal of Mathematics and Music*, vol. 9, no. 3, pp. 215–238, 2015.
- [83] L. J. Hofmann-Engl, “Virtual Pitch and the Classification of Chords in Minor and Major Keys,” 2008.
- [84] D. Tymoczko, “Scale Theory, Serial Theory and Voice Leading,” *Music Analysis*, vol. 27, no. 1, pp. 1–49, 2008.
- [85] A. Honingh and R. Bod, “In Search of Universal Properties of Musical Scales,” *Journal of New Music Research*, vol. 40, no. 1, pp. 81–89, 2011.
- [86] G. J. Balzano, “What are Musical Pitch and Timbre?,” *Music Perception: An Interdisciplinary Journal*, vol. 3, no. 3, pp. 297–314, 1986.
- [87] G. J. Balzano, “The pitch set as a level of description for studying musical pitch perception,” in *Music, mind, and brain*, pp. 321–351, Springer, 1982.

- [88] N. Carey and D. Clampitt, “Aspects of Well-formed Scales,” *Music Theory Spectrum*, vol. 11, no. 2, pp. 187–206, 1989.
- [89] D. Purves, *Music as Biology*. Harvard University Press, 2017.
- [90] K. Z. Gill and D. Purves, “A Biological Rationale for Musical Scales,” *PLoS One*, vol. 4, no. 12, p. e8144, 2009.
- [91] D. A. Schwartz, C. Q. Howe, and D. Purves, “The Statistical Structure of Human Speech Sounds Predicts Musical Universals,” *Journal of Neuroscience*, vol. 23, no. 18, pp. 7160–7168, 2003.
- [92] F. Stolzenburg, “Harmony Perception by Periodicity and Granularity Detection,” *Cambouropolos et al.(2012)*, pp. 958–959, 2012.
- [93] G. Langner, M. Sams, P. Heil, and H. Schulze, “Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: Evidence from Magnetoencephalography,” *Journal of comparative Physiology A*, vol. 181, no. 6, pp. 665–676, 1997.
- [94] G. Langner and C. E. Schreiner, “Periodicity Coding in the Inferior Colliculus of the Cat. I. Neuronal Mechanisms,” *Journal of neurophysiology*, vol. 60, no. 6, pp. 1799–1822, 1988.
- [95] R. John, “How to Calculate the Perceived Frequency of Two Sinusoidal Waves Added Together.” Mathematics Stack Exchange. <https://math.stackexchange.com/q/164386>.
- [96] L. Hsiung-Cheng, “Sources, Effects, and Modelling of Interharmonics,” *Mathematical Problems in Engineering*, 2014.
- [97] B. Razavi and R. Behzad, *RF microelectronics*, vol. 2. Prentice Hall New Jersey, 1998.
- [98] S. Ternström, “Physical and Acoustic Factors that Interact with the Singer to Produce the Choral Sound,” *Journal of voice*, vol. 5, no. 2, pp. 128–143, 1991.

REFERENCES

- [99] D. Temperley and D. Tan, “Emotional Connotations of Diatonic Modes,” *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 237–257, 2013.
- [100] E. C. Bairstow, *Counterpoint and Harmony*. Read Books Ltd, 2013.
- [101] T. L. Ssergejewitsch, “Method of and Apparatus for the Generation of Sounds,” Feb. 28 1928. US Patent 1,661,058.
- [102] M. J. Tramo, “Music of the Hemispheres,” *Science*, vol. 291, no. 5501, pp. 54–56, 2001.
- [103] J. Pachelbel, “Canon and Gigue for 3 Violins and Basso Continuo,” 1680–1706.
- [104] L. v. Beethoven, “Piano Sonata No. 14 in C minor, “Quasi una fantasia”, Op. 27, No. 2,” 1801.
- [105] E. Aldwell and A. Cadwallader, *Harmony and Voice Leading*. Cengage Learning, 2018.
- [106] M. Guernsey, “The Role of Consonance and Dissonance in Music,” *The American Journal of Psychology*, pp. 173–204, 1928.
- [107] M. M. Farbood, “A Parametric, Temporal Model of Musical Tension,” *Music Perception: An Interdisciplinary Journal*, vol. 29, no. 4, pp. 387–428, 2012.
- [108] M. Ebeling, “Neuronal Periodicity Detection as a basis for the Perception of Consonance: A Mathematical Model of Tonal Fusion,” *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2320–2329, 2008.
- [109] D. Tymoczko, “The Geometry of Musical Chords,” *Science*, vol. 313, no. 5783, pp. 72–74, 2006.
- [110] J. Kursell, “A Third Note: Helmholtz, Palestrina, and the Early History of Musicology,” *Isis*, vol. 106, no. 2, pp. 353–366, 2015.
- [111] C. Marvin, *Giovanni Pierluigi da Palestrina: A Research Guide*. Routledge, 2013.

REFERENCES

- [112] J. O. i. Font, *Musical and Phonetic Controls in a Singing Voice Synthesizer*. PhD thesis, Polytechnics University of Valencia, 2001.
- [113] S. Lemmetty, “Review of Speech Synthesis Technology,” *Helsinki University of Technology*, vol. 320, pp. 79–90, 1999.
- [114] A. Loscos, *Spectral Processing of the Singing Voice*. PhD thesis, Pompeu Fabra University, 2007.
- [115] A. Rugchatjaroen, *Articulatory-based English Consonant Synthesis in 2-D Digital Waveguide Mesh*. PhD thesis, University of York, 2014.
- [116] J. Bonada, O. Celma, A. Loscos, J. Ortola, and X. Serra, “Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models,” in *Proceedings of International Computer Music Conference*, 2001.
- [117] X. Serra, “Spectral modeling synthesis: Past and present,” *Proceedings of DAFX London*, 1993.
- [118] X. Serra and J. O. Smith, “Spectral Modeling Synthesis,” *Computer Music Journal*, pp. 281–284, 1990.
- [119] X. Serra and J. Smith, “Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition,” *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.
- [120] L. K. Le, “Examining the Rise of Hatsune Miku: The First International Virtual Idol,” *The UCI Undergraduate Research Journal*, 2014.
- [121] B. Larsson, “Music and Singing Synthesis Equipment (MUSSE),” *Dept. for Speech, Music and Hearing, Quarterly Progress and Status Report (STL-QPSR)*, vol. 18, no. 1, pp. 38–40, 1977.
- [122] X. Rodet and X. Rodet, “Synthesis and Processing of the Singing Voice,” in *Proc.1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, (Leuven, Belgium), pp. 99–108, 2002.

REFERENCES

- [123] P. R. Cook, *Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing*. PhD thesis, Stanford University, 1991.
- [124] J.-y. Cheng, Y.-c. Huang, and C.-h. Wu, “HMM-based Mandarin Singing Voice Synthesis Using Tailored Synthesis Units and Question Sets,” *Computational Linguistics and Chinese Language Processing*, vol. 18, no. 4, pp. 63–80, 2013.
- [125] M. Airaksinen, *Analysis/Synthesis Comparison of Vocoders Utilized in Statistical Parametric Speech Synthesis*. PhD thesis, 2012.
- [126] K. Oura, A. Mase, Y. Tomohiko, S. Muto, Y. Nankaku, and K. Tokuda, “Recent Development of the HMM-based Singing Voice Synthesis System - Sinsy,” in *7th ISCA Workshop on Speech Synthesis*, pp. 211–216, 2010.
- [127] M. Blaauw and J. Bonada, “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs,” *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.
- [128] E. Gómez, M. Blaauw, J. Bonada, P. Chandna, and H. Cuesta, “Deep Learning for Singing Processing: Achievements, Challenges and Impact on Singers and Listeners,” *arXiv preprint arXiv:1807.03046*, 2018.
- [129] P. Bussy, *Kraftwerk: Man, Machine and Music*. SAF Publishing Ltd, 2nd revised ed., 2001.
- [130] COMSOL, *Acoustics Module Software for Acoustics and Vibration Analysis*. <https://www.comsol.com/acoustics-module>, Date Accessed: 10 Feb 2015.
- [131] M. Umbert, J. Bonada, and M. Blaauw, “Generating Singing Voice Expression Contours based on Unit Selection,” in *Proc. SMAC*, 2013.
- [132] Virsyn, *CANTOR 2 - The Vocal Machine*. http://www.virsyn.de/en/E_Products/E_CANTOR/e_cantor.html, Date Accessed: 10 Feb 2015.
- [133] M. W. Macom, L. Jensen-Link, J. Oliverio, and M. A. Clements, “A Singing Voice Synthesis System based on Sinusoidal Modeling,” *Acoustics, Speech,*

REFERENCES

- and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, pp. 435–438, 1997.
- [134] Ameya Purojekuto. Ameya/Shobu, “UTAU-Synth.” <http://utau-synth.com>, 2008.
- [135] K. Oura, A. Mase, Y. Nankaku, and K. Tokuda, “Pitch Adaptive Training for HMM-based Singing Voice Synthesis,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5377–5380, 2012.
- [136] G. Chua, Q. C. Chang, Y. W. Park, P. Y. Chan, M. Dong, and H. Li, “The expression of singing emotion - contradicting the constraints of song,” in *Asian Language Processing (IALP), 2015 International Conference on*, pp. 98–102, IEEE, 2015.
- [137] “Using Dynamic Time Warping to find Patterns in Time Series, author=Berndt, Donald J and Clifford, James,” in *KDD workshop*, vol. 10, pp. 359–370, Seattle, WA, 1994.
- [138] J. H. Arnold, “Plainsong Accompaniment,” *The Musical Times*, vol. 68, no. 1015, pp. 816–822, 1927.
- [139] S. Langford, *Digital Audio Editing: Correcting and Enhancing Audio in Pro Tools, Logic Pro, Cubase, and Studio One*. CRC Press, 2013.
- [140] W. S. Marlens, “Duration and Frequency Alteration,” *Journal of the Audio Engineering Society*, vol. 14, no. 2, pp. 132–139, 1966.
- [141] Infotronics Systems Inc, “Eltro Mark II Information Rate Changer,” 1967.
- [142] M. Lech and B. Kostek, “A System for Automatic Detection and Correction of Detuned Singing,” *The Journal of the Acoustical Society of America*, vol. 123, p. 3177, 2008.
- [143] F. Holm, “NAMM 2009,” *Computer Music Journal*, vol. 33, no. 3, pp. 61–64, 2009.
- [144] L. Cen, M. Dong, and P. Y. Chan, “Template-based Personalized Singing Voice Synthesis,” pp. 4509–4512, 2012.

REFERENCES

- [145] M. Dong, S. W. Lee, H. Li, P. Chan, X. Peng, J. W. Ehnes, and D. Huang, “T²R Speech2Singing Perfects Everyone’s Singing,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [146] S. W. Lee, L. Cen, H. Li, Y. P. Chan, and M. Dong, “Method and System for Template-based Personalized Singing Synthesis,” Jan. 22 2015. US Patent App. 14/383,341.
- [147] MIDI Manufacturer’s Association, “The Complete MIDI 1.0 Detailed Specification, Version 96.1 Second Edition,” 2001.
- [148] H. Kawahara, “STRAIGHT, Exploitation of the Other Aspect of Vocoder: Perceptually Isomorphic Decomposition of Speech Sounds,” *Acoustical Science and Technology*, vol. 27, pp. 349–353, 2006.

Appendices

Appendix A

Music Theory Prerequisites

This appendix explains some prerequisites to the thesis from the field of music theory. Part [A.1](#) explains vocal harmony. Part [A.2](#) explains chords and keys. Part [A.3](#) explains harmonization. Part [A.4](#) explains types of consonances and dissonances in the context of conventional music theory.

A.1 Vocal Harmony

As discussed in Chapter [1](#), harmony is defined to be the simultaneous occurrence of two or more voices based on a relationship (i.e. interval) that is pleasant to the human ear (i.e. consonant). The main singing voice is called melody, whilst voices added to it to form the harmony are called accompaniments. Figure [A.1](#) illustrates an example with harmony being composed by melody (green) and two accompaniments (magenta and orange) on a piano roll.

A.2 Special Sets of Notes

In music, the entire audible range of fundamental frequencies is divided into one octave per frequency doubling and each octave is further divided into 12 discrete frequency locations called notes [\[3, 6\]](#). This means, for example, that there are 12 notes between 440Hz and 880Hz and another 12 between 880Hz and 1760Hz,

Figure A.2: Notes of a Melody

Figure A.3: A Special Set of Notes: Key

by the dots in Figure A.3. Since notes that are octaves apart may be regarded as synonymical, we only need to consider the notes within the octave in orange. For purpose of reference and ease of mathematical derivation, notes are numbered as shown in Figure A.3.

Given the above, the set of notes that form a key may be obtained by

$$K_k = \{k + [0, 2, 4, 5, 7, 9, 11] + 12i\} \quad (\text{A.1})$$

where

$$i \in \mathbb{Z} \quad (\text{A.2})$$

and where k is the numeric identifier of the key; K_k is the set of notes of key k ;

and $\mathbb{Z}$ is the set of all integers. It can be seen from the equation that the set of notes are fixed intervals from one another [3,6]. Adding $12i$ extends this across all octaves.

In the example, Figure A.3 may be compared to Equation A.1, to derive $k = 7$ as follows.

$$K_k = \{k + [0, 2, 4, 5, 7, 9, 11] + 12i\} \quad (\text{A.3})$$

$$K_7 = \{[7, 9, 11, 12, 14, 16, 18] + 12i\} \quad (\text{A.4})$$

$$K_7 = \{[7, 9, 11, 0, 2, 4, 6] + 12i\} \quad (\text{A.5})$$

$$k = 7 \quad (\text{A.6})$$

A.2.2 Chords

For each key, there are 6 chords, each comprising of 3 notes. Song harmonies are composed of notes of a chord.

An example of notes of a chord being used in harmony may be seen in Figure A.4, which shows the harmony extracted from the example song in Figure A.1. All the notes in the harmony between $t = 1$ and $t = 3$ of this song may be represented by the notes marked by the dots in Figure A.5. Since notes that are octaves apart may be regarded as synonymical, we only need to consider the notes within the octave in magenta. For purpose of reference and ease of mathematical derivation, notes are numbered as shown in Figure A.5.

Figure A.4: Simultaneously Sounded Notes

Figure A.5: A Special Set of Notes: Chord

Given the above, the set of notes that form a chord may be obtained by

$$C_{c,k} = \{K_k (c + [0, 2, 4] - 1) + 12i\} \quad (\text{A.7})$$

where

$$i \in \mathbb{Z} \quad (\text{A.8})$$

and where c is the numeric identifier of the chord bounded by integer values 1 and

6; C_c is the set of notes of chord c ; and $K_k(n)$ returns the n^{th} member of the set K_k .

A.3 Harmonization

The last section, explained how to derive the notes in each key and chord in music. Given a song's melody, a musician is able to compose accompaniments to harmonize with it by working out the underlying key and chords [3,4]. This section gives an example of how to do this.

There are no fixed rules on how accompaniments may be composed. This section describes a typical method by which an accompaniment might be derived based on theories compiled in [3]. The implementations covered in Chapter 4 simplify from different parts of this general method.

The method is as follows:

1. Determine the song's key from the notes in the melody based on Equation A.1.
2. Classify the notes in the melody into stressed and unstressed notes based on the onset and length of each note. Longer notes that are onset on more important beats are considered stressed and vice versa.
3. Derive chords for each time-segment of the song. This is based on the following two factors:
 - The likelihood of each chord based on the number of notes common with the stressed notes within the section.
 - The likelihood of each chord based on the chords before and after it (progression).

The chords are then assigned to each section based on the two likelihoods through a viterbi-like process.

4. Assign accompaniment notes as follows:
 - For every stressed note, assign another note within the chord assigned in Step 3 for accompaniment.
 - For every unstressed note, assign another note in the key derived from Step 1 for accompaniment.
5. Defining intervals by $J = m_{melody} - m_{acc} + 1$, where m_{melody} and m_{acc} are the numeric identifiers of the melody and accompaniment notes within K respectively, some other considerations include:
 - Intervals of $J = 2$ and $J = 7$ should be avoided at all costs.
 - Consecutive intervals of $J = 4$, $J = 5$ and $J = 8$ should be avoided where possible.
 - Intervals are almost always kept within an octave.
 - Motion trajectory should be considered. It is common to keep to direct motion to the melody (i.e. accompaniment's pitch increases or decreases together with the melody's), although occasional contrary motion (i.e. accompaniment's pitch increases or decreases inversely with the melody's) is welcome.

One simple way to fulfill the above is to use intervals $J = 3$ or $J = 6$ whenever possible, switching to $J = 4$ or $J = 5$ respectively only when the former does not fulfill the criteria.

A.4 Consonances and Dissonances in the Context of Conventional Music Theory

This section defines certain types of consonances and dissonances that are common in harmonizers in the context of classical music theory.

For the example melody in Figure A.6a, where set K is denoted by the orange dots and set C is denoted by the magenta dots:

a) Melody

b) Harmony

c) Type-I Dissonances

d) Type-II Dissonances

e) Parallel Motion

Figure A.6: Consonances and Dissonances

- Consonance may be defined as all the rules listed in Subsection A.3 being followed. An example of consonance is illustrated in Figure A.6b, where, according to Subsection A.3, all stressed notes are selected from the magenta dots, $n_{acc,stressed} \in C$, and all unstressed notes are selected from the orange dots, $n_{acc,unstressed} \in K$.
- *Type-I Dissonances* are defined as dissonances where some accompanying notes are selected from outside the orange dots, $n_{acc} \notin K$, as illustrated in Figure A.6c. These are perceived to sound more severe than Type-II dissonances.
- *Type-II Dissonances* are defined as dissonances where some stressed accompanying notes are selected from outside the orange dots, $n_{acc} \notin K$, as illustrated in Figure A.6d. These are perceived to sound more tolerable than Type-I dissonances.
- *Parallel Motion* is an unusual case of consonance that adds little enhancing effect to the melody. It occurs when $J(t) \in 4, 5, 8$ and $J(t-1) = J(t) = J(t+1)$, and is illustrated in Figure A.6e.

This concludes this appendix on traditional harmonization.

Appendix B

Supplementary Audio, Video and Tables

- **Supplementary Audio S1** Audio example of consonant low-frequency modulation with $\bar{f} = 440Hz$ and $\Delta f = 0.5Hz$:
https://www.dropbox.com/s/e593jtsorjq2u21/S1_deltaF_half.wav?dl=0
- **Supplementary Audio S2** Audio example of dissonant beating frequency with $\bar{f} = 440Hz$ and $\Delta f = 70Hz$:
https://www.dropbox.com/s/c578bq0wn4x6i6z/S2_deltaF_70.wav?dl=0
- **Supplementary Audio S3** Input audio sample of Brahms' Cradlesong before harmonization:
https://www.dropbox.com/s/lv23m9sjwdvgbz8/S3_Cradlesong_Input.mp3?dl=0
- **Supplementary Audio S4** Input audio sample of Twinkle Twinkle Little Star before harmonization:
https://www.dropbox.com/s/kx98a1ibfx09w3q/S4_Twinkle_Input.mp3?dl=0
- **Supplementary Audio S5** Harmonized audio sample of Brahms' Cradlesong using the 458 method, given Supplementary Audio S3 at the input:

https://www.dropbox.com/s/txtaotx84ou769y/S5_Cradlesong_458.mp3?dl=0

- **Supplementary Audio S6** Harmonized audio sample of Brahms' Cradlesong using the KTV method, given Supplementary Audio S3 at the input and Supplementary MIDI File S1 as reference:

https://www.dropbox.com/s/2e45kf8ymj5cpji/S6_Cradlesong_KTV.mp3?dl=0

- **Supplementary Audio S7** Harmonized audio sample of Brahms' Cradlesong using the method proposed in the thesis, given Supplementary Audio S3 at the input and Supplementary MIDI File S1 as reference:

https://www.dropbox.com/s/2f5d7wikwgvjwap/S7_Cradlesong_S2A.mp3?dl=0

- **Supplementary Audio S8** Harmonized audio sample of Twinkle Twinkle Little Star using the 458 method, given Supplementary Audio S4 at the input:

https://www.dropbox.com/s/7i0e7nofjcbjxm4/S8_Twinkle_458.mp3?dl=0

- **Supplementary Audio S9** Harmonized audio sample of Twinkle Twinkle Little Star using the KTV method, given Supplementary Audio S4 at the input and Supplementary MIDI File S2 as reference:

https://www.dropbox.com/s/6oihq03dod4gjy7/S9_Twinkle_KTV.mp3?dl=0

- **Supplementary Audio S10** Harmonized audio sample of Twinkle Twinkle Little Star using the method proposed in the thesis, given Supplementary Audio S4 at the input and Supplementary MIDI File S2 as reference:

https://www.dropbox.com/s/ut132k02v5ajtsh/S10_Twinkle_S2A.mp3?dl=0

- **Supplementary Audio S11** Utterance mimicking Supplementary Audio S12 synthesized using SERAPHIM controlled by realtime touch-screen gestures:

https://www.dropbox.com/s/8h9jrb80aah3vqk/S11_SERAPHIM_1.wav?dl=0

- **Supplementary Audio S12** Utterance synthesized using LIMSI’s Cantor Digitalis:

https://www.dropbox.com/s/5b3ytikqk4o0qyv/S12_LIMSI_1.wav?dl=0

- **Supplementary Audio S13** Utterance mimicking Supplementary Audio S14 synthesized using SERAPHIM controlled by realtime touch-screen gestures:

https://www.dropbox.com/s/tnnrhaz0cju77sf/S13_SERAPHIM_2.wav?dl=0

- **Supplementary Audio S14** Utterance synthesized using LIMSI’s Cantor Digitalis:

https://www.dropbox.com/s/ukisne376cswz03/S14_LIMSI_2.wav?dl=0

- **Supplementary Audio S15** Utterance mimicking Supplementary Audio S16 synthesized using SERAPHIM controlled by realtime touch-screen gestures:

https://www.dropbox.com/s/x5n4zspbk3eroha/S15_SERAPHIM_3.wav?dl=0

- **Supplementary Audio S16** Utterance synthesized using LIMSI’s Cantor Digitalis:

https://www.dropbox.com/s/yk7khde6cvqx0ar/S16_LIMSI_3.wav?dl=0

- **Supplementary Audio S17** Utterance mimicking Supplementary Audio S18 synthesized using SERAPHIM controlled by realtime touch-screen gestures:

https://www.dropbox.com/s/ewjf81d7ygr6pdn/S17_SERAPHIM_4.wav?dl=0

- **Supplementary Audio S18** Utterance synthesized using LIMSI’s Cantor Digitalis:

https://www.dropbox.com/s/u66apzibnn7a6go/S18_LIMSI_4.wav?dl=0

- **Supplementary Audio S19** Utterance mimicking Supplementary Audio S20 synthesized using SERAPHIM controlled by realtime touch-screen gestures:
https://www.dropbox.com/s/nlqa48f4gayrkp8/S19_SERAPHIM_A.wav?dl=0
- **Supplementary Audio S20** Utterance synthesized using Yamaha’s VOCALOID Keyboard:
https://www.dropbox.com/s/682d9o3c5mddg3u/S20_YAMAHA_A.wav?dl=0
- **Supplementary Audio S21** Utterance mimicking Supplementary Audio S22 synthesized using SERAPHIM controlled by realtime touch-screen gestures:
https://www.dropbox.com/s/e3etxsbhir49tyh/S21_SERAPHIM_B.wav?dl=0
- **Supplementary Audio S22** Utterance synthesized using Yamaha’s VOCALOID Keyboard:
https://www.dropbox.com/s/123v2hgrv5h69od/S22_YAMAHA_B.wav?dl=0
- **Supplementary Audio S23** Utterance mimicking Supplementary Audio S24 synthesized using SERAPHIM controlled by realtime touch-screen gestures:
https://www.dropbox.com/s/kundptyj43tk1a6/S23_SERAPHIM_C.wav?dl=0
- **Supplementary Audio S24** Utterance synthesized using Yamaha’s VOCALOID Keyboard:
https://www.dropbox.com/s/lqhx2z07ycl6jzl/S24_YAMAHA_C.wav?dl=0
- **Supplementary Audio S25** Utterance mimicking Supplementary Audio S26 synthesized using SERAPHIM controlled by realtime touch-screen gestures:
https://www.dropbox.com/s/0czwoiud2mudjku/S25_SERAPHIM_D.wav?dl=0
- **Supplementary Audio S26** Utterance synthesized often Yamaha’s VOCALOID

Keyboard:

https://www.dropbox.com/s/erfpemep48zh8f1/S26_YAMAHA_D.wav?dl=0

- **Supplementary MIDI File S1** MIDI reference for Brahms' Cradlesong:
https://www.dropbox.com/s/frofde5egwo1o75/S1_Cradlesong.mid?dl=0
- **Supplementary MIDI File S1** MIDI reference for Twinkle Twinkle Little Star:
https://www.dropbox.com/s/ta1gea6qnozkk4y/S2_Twinkle.mid?dl=0
- **Supplementary Video S1** Video example comparing subharmonic wave deformation in low tension C chord against high tension Cm^7 chord. https://www.dropbox.com/s/9vsw7a3yvrehgy/S1_SubharmonicWaveDeformation.mp4?dl=0
- **Supplementary Video S2** Visualizing stationary harmony with subharmonic tension in Pachelbel's Canon [103]. https://www.dropbox.com/s/kgf1jgl8exyz0rl/S2_StationaryHarmonySubharmonicPlot.mp4?dl=0
- **Supplementary Video S3** Visualizing transitional harmony with subharmonic tension in Beethoven's Moonlight Sonata [104]. https://www.dropbox.com/s/r1t0xjdc12c3s7t/S3_TransitionalHarmonySubharmonicPlot.mp4?dl=0