Geometric Approaches to the Pitch Estimation of Acoustic Musical Signals

By

Thomas A. Goodman

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
November 2021

ABSTRACT

Multi Pitch Estimation (MPE) is a challenging problem in the field of Music Information Retrieval (MIR). In recent literature in particular, it has been approached with Machine Learning (ML) methods, which are largely opaque, hard to interpret, and often difficult to reproduce from just the information provided in the literature. This Thesis presents a model for pitch detection that reduces the problem of MPE to that of distinguishing between false fundamentals ($\otimes$) and their real counterparts. It further provides an in-depth characterisation of precisely the ways in which these so-called edge cases can occur, looking in particular at the notion of 'basic' edge cases—ones in which the constituent parts are satisfied precisely once. From there, their occurrence is reduced to eight basic edge types (and a ninth type, which is proved to be the only irreducible non-basic type). The results of analysing simulated data on the model are then presented, highlighting the prevalence of the various types with respect to the number of simultaneous fundamentals. In addition, some insight into the use of the model on real data is given, alongside evaluation of a number of simple algorithms utilising the acquired knowledge of edge cases. Finally, this paper presents a range of logical future additions and directions for research, including the possibility of adopting a similar approach for other data—not necessarily musical audio.

DEDICATION

To my family and friends, without whom I would never have reached this point, and in particular to my incredible Nanny Ol for being a lifelong inspiration to me.

ACKNOWLEDGMENTS

I think it wise to start this with a quick acknowledgement to the countless trees that are going to be sacrificed in order to indulge the ridiculous length that this acknowledgements section is going to end up being... There are far too many people to thank for where I am today, and in large part, for finishing this Thesis. I'm sorry to those that I haven't been able to include explicitly, but thank you so much to everyone nonetheless.

First and foremost, I would like to give my utmost thanks to my incredible supervisor, Professor Peter Tino, who has been perhaps the best PhD supervisor I could have asked for. From actually taking me on for a project pretty unrelated to his specific research areas (even though I'd never spoken to him before!), to believing that my poor grade in Achim's maths course was because *"I can't count, but I swear I can do the maths"*, and all of the support, guidance, and fearsome intellect that he has provided these last few years, I'm not sure I can really put into words how much I have appreciated him throughout my PhD.

A tough act to follow, but one that is certainly matched by my partner in mathemagical crime—Karoline van Gemst—without whom, I'm not sure this PhD would have ended up how it has. Her sharp wit, mathematical rigour, immense consumption of coffee, and stubbornness to match only my own has spurred me on through this research, and made me a much better mathematician, researcher, and person.

To my Thesis group, Drs Eike Ritter and Ian Batten—thank you for the guidance, particularly toward the start of this project, and being willing to entertain my vaguely manic ramblings at random points of random days, often without any warning at all.

To all of my housemates and ex-housemates, thank you for putting up with me (to put it frankly), and often dealing with my spontaneous board games, takeaways, cocktail evenings, and 'welsh beer pong' that one time. It has been an absolute pleasure. On that note, a huge thanks to Keith and Cheryl MacFarlane, who have been a wonderful landlord and landlady for the last four and a bit years. In addition, I think a special shoutout to everyone that has played D&D with me in the last six years deserves a special mention—for introducing me to a wonderful game, and providing a plethora of memorable moments.

I want to say a special thank-you to Jake Foster, who has always been an incredible friend to—even at my lowest. I couldn't have asked for a more loyal, trustworthy, and genuinely wonderful friend. Please for the love of god though, don't ever grow that lockdown beard ever again.

To Bede, Peran, Aran, Callum, Jack, and Maciek—You are all a huge part of who I am today, and I'm so glad to be able to keep in touch, even though we're all over the place at the moment. From spontaneous holidays to Geneva, to visiting the Peak District and conquering my fears, to just always being down for board games, it has been an absolute blast. In addition, I'd like to thank Fraser Wade, who has been a fantastic friend over the years, and is always a pleasure to spend time with.

Suzi, I can't put into words how fantastic you have been to me, but let's make sure that we have countless karaoke nights in the years to come.

Next, I want to thank everyone who has been a part of the Computer Science Society (CSS) community over the years. In many ways, I'd like to blame Jack Wearden for inspiring me to get involved with the Society in the first place, but in all honesty, it was one of the best things I ever did. To all of those that sat on the committee with me through my three years as President, and two as Secretary, thank you—you helped to inspire hundreds of people to get involved, feel at home in the Department, and make the most of their university experience.

Though I may no longer be on the committee, I am so proud to see what CSS has become over the years, and I really hope it can do what it did for me all those years ago to as many students as possible in the years to come. Without CSS, I wouldn't have came out of my shell at uni, and I struggle to find many things I can't attribute in one way or another to my time in the Society. I am so proud to see what it has become today, and to have been a part of it thus far. A special thanks in particular has to go to Andreea Gheorghe, the best Secretary I could have asked for during my presidencies, and Jacqui Henes, who was a fantastic International Rep and Vice President, and to this day, one of my closest friends.

HackTheMidlands has been a huge part of my life ever since I founded it with Liam Sorta back in 2016, and I'd like to thank all of those that have helped to make it a success—in particular, all of the organisers past and present that have poured countless hours into making it happen year in year out. Liam, on that note, I need to extend my immeasurable thanks to you. Over the years, you have been a rock to me, a role model, and quite honestly, my best friend. Though you may soon be moving away to Japan, I know we'll always be in touch, and nothing will ever make me lose track of all of the incredible memories we've forged together. You'll be happy to know my coat no longer smells like Jasper has peed on it.

On that note, I'd like to thank AFNOM for having me along (at least to some extent) for the ride. I still remember solving that audio steganography challenge and feeling absolutely chuffed that perhaps my PhD work was maybe giving me some semblance of use... I'd also like to apologise for my horrific vim-based challenge that I made for the first-ever WhatTheCTF!?, but hope that I made up for it with my weird cryptic one.

My next thanks goes to the fantastic Dr Andreea Radu, who has been an absolute rock for me over the last few years (yes, I know, I'm a lemon...). Andreea, you have been so immensely wonderful to me, and I'll always look back fondly on our chats—even the ones in

the cold on those steps round from CS whilst you had a quick smoke.

To everyone 'behind the scenes' in the School of Computer Science that made (and make) it such a wonderful place to work—in particular, Helen Whitby, Kara (sorry for losing your pink stapler!), Caroline, Julie, Sarah, Jason, Mia, Ellie, Kate, Hayley, and Manesha. It has been an absolute pleasure to work, chat, and otherwise interact with you all over the years. From the academic side, it's hard to pinpoint just a few people to thank, but in particular, I'd like to thank Dan Clark, Dr Rich Thomas, Kit Murdock, Ollie Kamperis, Dan Fentham, Marco Canducci, Anna Laura Suarez, Fatma Faruq, Dr Bram Geron, Dr Sandy Gould, Professor Volker Sorge and Professor Iain Styles in particular for being a huge part of my time at the University. In addition, to all of the members of the SRSCC, it has been awesome to be a part of improving the time of PhD students in the School, and I really hope I can still pop along to some socials as we start to move back to normal.

A huge thanks needs to go out to Professor Achim Jung, who was not only willing to put up with (countably) infinitely many mathematical questions from me, but also helped time and time again to steer my research from the sidelines—perhaps even without knowing it. I cannot thank you enough Achim, even if you did laugh at me for weeks after you saw how appalling my final grade was on your maths exam in the 2nd year of my undergrad...

To the other 2019/20 Westmere Scholars—Sarah, Alice, and Beth—you were absolutely fabulous to work with, and though a lot got in our way during our tenure, we sure as hell did accomplish a lot! In addition, a huge thank you to Eren Bilgen, who was a brilliant line manager, and also spurred me on to applying not just for my FHEA, but my SFHEA shortly after, and always believed in me.

Thinking back to societies, I'd like to thank oSTEM for providing a fantastic space for me to learn more about myself, others, and putting so much of your time into making such an inclusive and open space for LGBTQ+ folk. Thanks in particular to Ailsa, Zia, Calum,

Ethan, and the rest of y'all that were a part of my interactions with oSTEM. To Sober Socials, I may have drank more with y'all (outside of official events) than I did otherwise, but I made some damn good friends out of it, and had a fab time.

I couldn't have asked for someone better than Avery Cunningham to spur me on over the years—nor a better role model. It's hard to express quite how much I admire you Avery, but I'm so glad I met you, and hope we can keep in touch as you embark on your own PhD journey now. On that note, I want to thank Beth Soanes and Helena Dodd, both of whom have also been bloody incredible, not just with random spontaneous scheming, but also in supporting me, and one another, even when things have gotten tough. Helena, I hope I can come and stroke Boots soon—I'm sure he misses me!

To everyone at the Guild of Students that I've worked with, thank you so much for giving me experience and insight that I'd never have gotten otherwise, and being fantastic friends to me. On that note, to all of my colleagues at the University of Birmingham—in particular those that I've sat on committees with—thank you for the experience, support, and giving me the confidence to meaningfully contribute to discussions.

Next up, I couldn't go without thanking Grace Surman, who is one of the most incredible people working at the University of Birmingham that I have had the pleasure of knowing. She puts so much energy into all that she does, and helps to empower student societies from across the College to do awesome things. The EPS Awards were fun as hell to organise too, and they wouldn't have existed without the tireless work of herself and Daisy. On that note Grace, we still need to go and grab that GT sometime!

I want to thank Sean Russell for being a fantastic mentor to me this past year and a half. You've spurred me on, given me incredible advice, and helped me to be far more confident in myself than I think I'd ever have managed to be otherwise.

To Dr Emily Rozier, you have been wonderful. We may have only really bumped into eachother a few times until recently, but working with you on the graduations was fantastic, and I can't wait to geek out a bit more about academic dress with you, and maybe play D&D sometime?

Dr Bhalroo, thank you so much for your help over the years, and for always being patient and understanding with me. I also want to thank all of my colleagues in the NHS for being possibly the loveliest possible team I've ever witnessed, and for pulling me through many parts of the pandemic we've found ourselves in.

Craig, thank you so much for always being there and lending a helping hand. I really appreciate it, and I wouldn't have gotten nearly as far as I have without your help and support, and I'm eternally grateful.

Finally, I want to give a huge thanks to my family. Mum, you have always been there for me, and I know we argue sometimes, but I love you more than I can put into words. You're an inspiration to me, and I hope I get to spend some more time with you when I perhaps finally become a 'proper adult'. Dad, we don't keep in touch as much as I'd like to, but I know you're always my biggest advocate. Beckie, I am so proud of who you've become over the years, and I'm so excited to see you flourishing into a fantastic young woman, pharmacist, and person. To Grandma Sheila, Granddad Jim, Annmarie, Auntie Maureen, and Grandma Jean—you're always there to listen to me ramble about my work, watch me get *a little* too merry, and just generally have a laugh with me. Nanny Ol, I've already specifically dedicated this Thesis to you, but I can't imagine a more inspirational, incredible woman. You've always been there for me, and have taught me to be who I am today. To the rest of my family—there are too many of you to write out in full, but I really wouldn't be even a fraction of who I am without you all, and I am so grateful for your love, support, and friendship.

This Thesis wouldn't be complete without a thanks to my pets—Gin, Tonic, Pascal, my fish, Monty, Hazel, and those that went before them—Cresty, Tiger, and Houdini. I guess on that note, my houseplants also deserve a mention. You've all given me companionship in one way or another, helped me to appreciate the little things, and slow down when things are manic or overwhelming.

Of course, I also owe a thanks to SARS-CoV-2, which has possibly had one of the biggest impacts on my Thesis, having been around for more than half of my PhD. Though it has been a nightmarish pandemic—to put it lightly—I hope that I have come out of this more resilient, tenacious, and most importantly, more grateful for what I have in life. It has certainly made me realise what's truly important.

Nanny Ol: "Are you alright pet?"

Me: "I'm not a pet nana, I'm a person."

Contents

List of Figures

List of Tables

Acronyms

12-TET Twelve-tone Equal Temperament. 12

ACF Autocorrelation Function. 48–50, 55

ADSR Attack, Decay, Sustain, Release. 54, 55

AM Acoustic Model. 50, 51

AMT Automatic Music Transcription. 43, 46, 50, 52, 55, 56

CNN Convolutional Neural Network. 50–52, 54

CRNN Convolutional Recurrent Neural Network. 56

CTC Conectionist Temporal Classification. 56

DL Deep Learning. 52, 55

DNN Deep Neural Network. 50, 51

FFN Feed-Forward Network. 50

HMM Hidden Markov Model. 51, 55

HPS Harmonic Product Spectrum. vii, 45, 50

IBM International Business Machine. 41, 42

MIR Music Information Retrieval. i, 1–3, 6, 39–44, 48–50, 52

ML Machine Learning. i, 2, 43, 56

MLM Music Language Model. 50, 51

MPE Multi Pitch Estimation. i, 2, 3, 7, 39, 48, 51, 52, 54, 55, 57, 131, 134

MSE Minimum Squared Error. 52

NADE Neural Autoregressive Distribution Estimator. 51

NN Neural Network. vii, 50, 52, 55, 56

NNMF Non-Negative Matrix Factorisation. 49

PE Pitch Extraction. 43, 45, 48, 49, 55–57, 65

PLCA Probabilistic Latent Component Analysis. 51

RNN Recurrent Neural Network. 50, 51

SDF Square Difference Function. 52

STFT Short-Time Fourier Transform. 54

ZCR Zero-Crossing Rate. 49

Chapter One

Introduction

1.1 Statement of the Problem

Music is complex. Even with a single voice, a plethora of aspects coalesce to sculpt what we, as humans, would consider to be musical. Timbre, dynamics, and articulation - to name but a few - all contribute to this astounding intricacy, and yet over the years, we have become increasingly extravagant in both our creation and appreciation of this art. It has been a core aspect of human culture since time immemorial - with even instruments dating back c. 36,000 years being found [120].

It's no surprise then, that since the dawn of computing, we have striven to utilise new technologies for a range of musical tasks - including analysis, composition, and performance. Around the first two applications especially, computers - unlike humans - are not particularly good at this to say the least. They struggle to comprehend and pick apart the complex signals that constitute music, to make sense of the various intricacies that we have become accustomed to, and they lack creativity and insight. This desire to equip computers with ever more sophisticated methods of tackling problems in this space led to what is now the field of Music Information Retrieval (MIR).

Of course, MIR was preceded by a related, and perhaps initially more pertinent field—that concerning the properties of speech, and analysis and modification thereof. Notably explored in depth by Alexander Graham Bell in his work, and then necessitated by the invention of the telephone, this 'sister field' dates back at least into the mid-19th Century, and arguably much further—with the Greek philosopher Galen (129 CE - c. 210 CE) first describing the Larynx [83].

Recent advances in the field (See Chapter 3 for more detail) have focused particularly around Machine Learning (ML) methods - seeking broadly to imitate the function of the brain with respect to music. Therein lies the motivation for this work. Unlike traditional algorithmic approaches, these ML methods often lack explainability—the ability for us to interpret and adequately describe the process by which such a model has reached its conclusion(s) (in terms of mathematical and musical intuition, and high-level concepts). The aim of the research presented in this Thesis is therefore to provide a fresh perspective and direction to tackling MIR problems (in particular Multi Pitch Estimation (MPE)) in a more explainable and intuitive fashion. Further, the goal is to develop a set of tools, insights, and thoughts that may provide some suitable framework to extend and iterate on this work in the future.

The problem, therefore, is not to develop another approach to MPE in the traditional sense of engineering a way in which to tackle the problem, but rather, to build a model from the ground up that provides a new perspective on MPE, and lays the framework for future research along the same lines.

1.2 Contributions

This Thesis makes a number of novel contributions to the field of Music Information Retrieval, including:

- A geometric framework for modelling acoustic musical signals, which provides a novel way in which to approach related problems such as Multi Pitch Estimation.

- Reduction of the problem of Multi Pitch Estimation to that of distinction between real and false fundamentals in the given framework.

- Characterisation of edge cases – looking in depth at the circumstances in which false fundamentals can occur.

- Definition of basic edge types to unify cases across the three possible configurations ($\Gamma\Gamma$, $\Gamma \vdash$, and $\vdash \Gamma$).

- Definition of edge case to basic edge case reduction, and the use of this tool to perform an analysis of the prevalence of basic edge types by number of simultaneous fundamentals in the signal.

- Proof of the existence of precisely one irreducible non-basic edge case.

- Exploration and presentation of real-world data in an adapted version of the framework (using $\mathbb{R}$ instead of $\mathbb{B}$).

- Evaluation of simple algorithms on simulated data, and discussion around building and testing further algorithms that utilise the framework.

1.3 Publications Arising from This Work

A Geometric Framework for Pitch Estimation on Acoustic Musical Signals

As highlighted in Chapter 4, much of the work in Part II is published in the Journal of Mathematics and Music, albeit in a somewhat condensed format in comparison to the contents of this Thesis. Building broadly off of my work in [67], and exploring new ways of representing musical signals initially resulted in the formulation of $\mathcal{N}^{\mathbb{V}}$—the space depicted in Figure 4.5—and this paper concerns both the formulation of a more sophisticated model, a characterisation of the edge cases that arise, and a brief experimental look at the model in practice.

Goodman, van Gemst, and Tino	
Journal of Mathematics and Music	
Submitted	November 2020
Accepted	September 2021
Published	TBC
DOI	10.1080/17459737.2021.1979116

Abstract: *This paper presents a geometric approach to pitch estimation (PE) - an important problem in Music Information Retrieval (MIR), and a precursor to a variety of other problems in the field. Though there exist a number of highly-accurate methods, both mono-pitch estimation and multi-pitch estimation (particularly with unspecified polyphonic timbre) prove computationally and conceptually challenging. A number of current techniques, while incredibly effective, are not targeted towards eliciting the underlying mathematical structures that underpin the complex musical patterns exhibited by acoustic musical signals. Tackling the approach from both a theoretical and experimental perspective, we*

present a novel framework, a basis for further work in the area, and results that (while not state of the art) demonstrate relative efficacy. The framework presented in this paper opens up a completely new way to tackle PE problems, and may have uses both in traditional analytical approaches, as well as in the emerging machine learning (ML) methods that currently dominate the literature.

Real-Time Polyphonic Pitch Detection on Acoustic Musical Signals

Though based on the work from my Bachelor's Thesis (and therefore no more than tangentially related to this Thesis in content), this paper was adapted into a conference paper during the start of my PhD, and—in an attempt to improve on the achieved results—unwittingly led me to the beginnings of the geometric model that this Thesis presents.

Goodman and Batten
2018 IEEE International Symposium on Signal
Processing and Information Technology (ISSPIT)

Submitted	September 2018
Accepted	November 2018
Conference	6^{th}-8^{th} December 2018
Published	February 2019
DOI	10.1109/ISSPIT.2018.8642626

Abstract: *This paper presents an algorithm for fundamental frequency detection on polyphonic acoustic musical signals, based on a new 'raking' method over the frequency-domain spectra. The algorithm is evaluated as a classifier, and boasts a good accuracy (*83.20%*) compared to other such methods, as well as the ability to function effectively in real-time, with a running-speed below* 140*ms per window evaluated. This proves to be real-*

time for the use-case, as the latency between an auditory stimulus and its perception by a person has been shown to be longer than this. The algorithm itself runs in linear-time, but is thus slowed by the $O(nlog(n))$ Fast Fourier Transform during preprocessing. Though the algorithm fails to account for certain edge-cases with overlapping harmonics as well as certain instruments, future work and improvements are also presented, paving the way for further research.

1.4 Thesis Overview

Part I: Preliminaries

Chapter 2: Background

In this Chapter, we cover much of the underlying material and background knowledge that I found useful, necessary, and in large parts, *interesting* whilst working on this Thesis. By no means feel obliged to read all of the Sections, but I'd certainly encourage dipping in and out to get an idea of the framing in which I conducted my research.

Chapter 3: Related Work

This Chapter provides a literature review, with a specific focus on pitch estimation within MIR, and covers four main periods,

- Pre-MIR, which looks at work related to pitch estimation that was largely developed in the late 19^{th} and early 20^{th} Centuries for speech-related applications;
- Early MIR, presenting the foundational work behind the field of MIR;
- MIR in the 2000s and Early 2010s, at which point MIR (and pitch estimation specifi-

cally) began to more closely align with adjacent work in speech processing; and

- State of the Art MIR—bringing us up to date with the cutting-edge research in Multi Pitch Estimation at the time of writing.

Part II: A Geometric Model for Pitch Estimation

Chapter 4: A Geometric Framework for Pitch Detection

In this Chapter, we present the titular mathematical model that was designed to be used in pitch estimation settings, and characterise the occurrence of edge cases— false fundamentals, $\otimes$—that we anticipate are the prevailing source of error in classification approaches using this model. In essence, our characterisation works towards some distinguishment between real and false fundamentals in a signal that is transformed into our representation.

Chapter 5: Experimental Investigations

Chapter 5 looks to ground the work of this Thesis a little more firmly into the application of the model itself—both by looking at the types of edge cases that occur in (simulated) reality, as well as displaying real signals in a slightly modified model, and evaluating some simple pitch estimation algorithms that work on it.

Part IV: Concluding Remarks

Chapter 6: Future Work and Conclusion

Here, I present what I believe are the logical extensions (or at least a subset thereof) to my work, and a number of avenues that I ultimately wish I had the time or inhuman endurance to explore more thoroughly. It is then rounded off with some concluding remarks.

Chapter Two

Background

> "There is geometry in the humming of the strings. There is music in the spacing of the spheres. Geometry is knowledge of the eternally existent."
>
> — *Pythagoras*

WHILST tackling this research, I oft found myself having to tap into knowledge from a plethora of seemingly unrelated fields. In an effort to guide the reader somewhat, this introductory section contains some hopefully approachable (and generally independent) sections that cover these areas of background knowledge. I would highly encourage the reader to at least dip into the more unfamiliar sections before moving on to the bulk of the Thesis.

2.1 The Mathematics of Music

Perhaps the first recorded musings of the relationship between mathematics and music sit squarely with Pythagoras (c. 570 BC - c. 495 BC) - an ancient Ionian Greek philosopher.

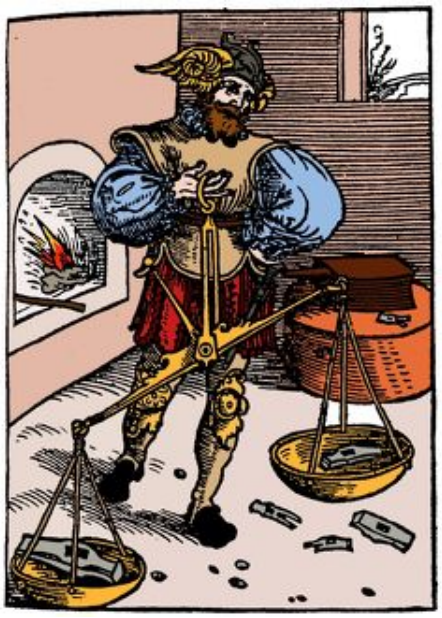


Figure 2.1: Woodcut from Martin Agricola's 'Musica Instrumentalis Deudsch' [80] depicting Pythagoras determining the ratios between the weights of blacksmiths hammers.

Though certainly better-known for his theorem concerning the hypotenuse of a right-angled triangle, Pythagoras was also instrumental in eliciting the foundations of tuning - the ratios between certain related notes. According to the legend of the Pythagorean hammers (Figure 2.1), it is stipulated that Pythagoras discovered these relationships by observing the sounds of four (differently weighted) blacksmiths' hammers as they struck their anvils, and the resulting harmony when they were stuck simultaneously. For reasons noted in Section 2.2 (with respect to standing waves), the story itself is likely to be little more than a tale, but it is still very much the case that Pythagoras did indeed describe the ratios of 2:1 (an octave), 4:3 (a perfect fourth), and 3:2 (a perfect fifth). These observations led to a system of tuning based of ratios (in particular, perfect fifths), known as Pythagorean tuning.

Another example is Fibonacci (c. 1170 - c. 1240-50), an Italian mathematician from the middle ages. Though not himself connected to music as such, his eponymous sequence has certainly had an influence on, or at least often appears in music. The Fibonacci sequence

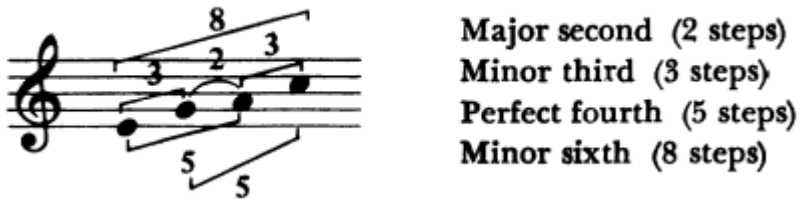


Figure 2.2: Extract from Bartók's 'The Castle of Prince Bluebeard' demonstrating the choice of intervals of sizes (in semitones) that correspond to elements of the Fibonacci sequence. Figure taken from [8].

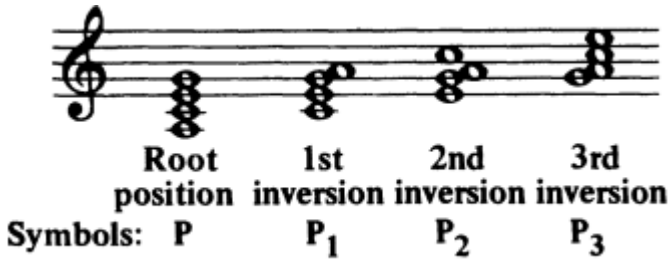


Figure 2.3: The root position and first three inversions of the pentatonic chord. Figure taken from [8].

starts

$$1, 1, 2, 3, 5, 8, 13, ...,$$

with each term equalling the sum of the two preceding it. But how does this relate to music? For one, the ratio of consecutive Fibonacci numbers (e.g. 13:8) converges to φ, the golden ratio, which - as in many disciplines over the years - has been incorporated into many musical works. Take Béla Bartók's 'The Castle of Prince Bluebeard' (Figure 2.2) [8], which utilises the pentatonic chord in its second inversion (P_2) (Figure 2.3). Notably the intervals between ascending notes in P_2 are 2, 3, 5, and 8 semitones respectively - directly corresponding to a section of the Fibonacci sequence. This can be observed 'in the wild' in the aforementioned piece (Figure 2.2). Rather loosely, even the notes in the major triad - arguably the 'purest' consonance - match up to the Fibonacci sequence (Figure 2.4).

Van Gend [66] elicits similar observations in their paper, and presents a short compo-

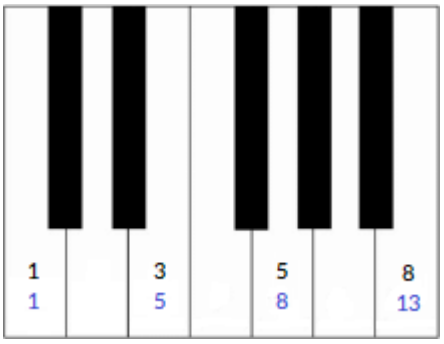


Figure 2.4: One octave of a piano, with the major triad of C (C, E, G), and the octave annotated with their position in the C major scale (black), and position in the chromatic scale starting on C (purple).

sition which utilises the Fibonacci sequence in a number of contexts, including the lengths (in bars) of 'thematic' sections of the piece, and intervals between notes. Whether or not such a composition is palatable (or perhaps even enjoyable) is left to those readers with either the extraordinary talent to easily hum it to themselves, or the time and dedication to this Thesis to go away and learn to play it. Regardless, such uses of the Fibonacci sequence (and φ) are relatively commonplace in music [96], even in the current day - whether intentionally or not. It is hard to believe, however, that it is entirely a coincidence.

Franchinus Gaffurius (1451 - 1522), an Italian music theorist, authored a trilogy of manuscripts (the Trilogia Gaffuria) which offered a broad perspective on music theory [115]. Notably, the second book in the trilogy, authored in 1492, includes some treatment Pythagoras's system of tuning (Figure 2.6). In his diagrams, and accompanying prose, Gaffurius illustrates the tuning of various instruments to Pythagoras's system, including bells, pipes, a monochord, and even a glass harmonica. The most amazing feat of course being the preservation of this knowledge over more than two millennia since Pythagoras had first discovered it.

The modern system of tuning (in Western music at least) is known as Twelve-tone Equal Temperament (12-TET), and was first mathematically 'solved' in 1584 by the Chinese

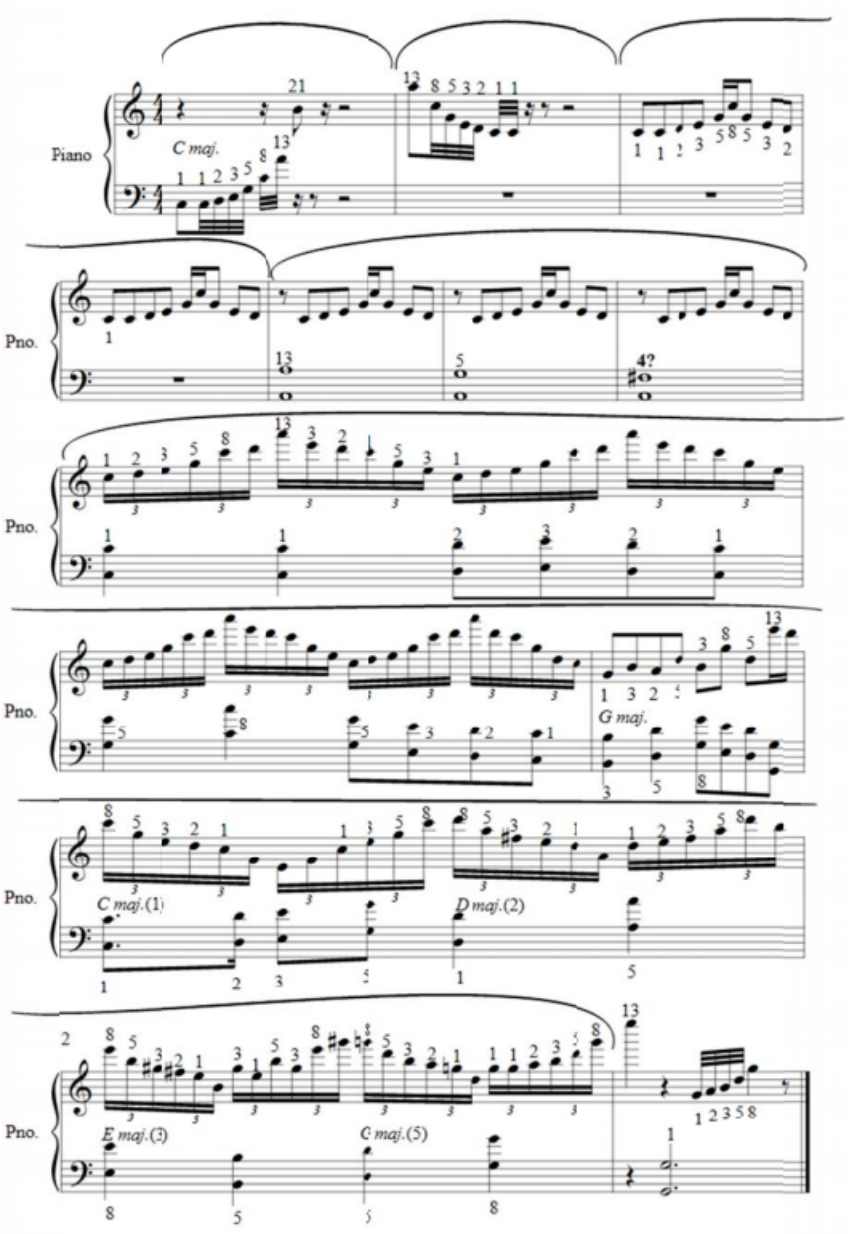


Figure 2.5: 'Fibonacci Composition' from [66] demonstrating the use of numbers from the Fibonacci sequence to architect almost every aspect of the piece. Figure taken from [66].

Figure 2.6: A page taken from Gaffurius's 'Theorica Musicae', depicting a variety of instruments tuned according to Pythagoras's system, as well as what appears to be a depiction of the story of the Pythagorean hammers (top left).

mathematician and musician (amongst other vocations), Zhu Zaiyu (1536 - 1611) [169]. Zaiyu not only became the first to describe a system in which consecutive semitones could be calculated using $\sqrt[12]{2}$ (i.e. by multiplying the frequency of one note by $\sqrt[12]{2}$ to obtain the frequency of the note a semitone up), but also tuned several instruments, including bamboo pipes (Figure 2.7), to the new system.

Moving chronologically onward, another important in mathematics and music was the circle of fifths [28, 106] (Figure 2.8) - an arrangement of the twelve chromatic pitches such that the intervals between each are perfect fifths in one direction. An adjacent concept was first laid out by Heinichen in his 'Musicalischer Circul' (Figure 2.9a), with Nikolay Diletsky being their first to present the circle of fifths itself in his 'Idea Grammatiki Musikiyskoy' (Figure 2.9b) in 1679. In essence, it describes a closeness (from a harmonic perspective) of not just individual notes, but indeed, keys.

The concept has historically been used extensively in composition, predominantly as

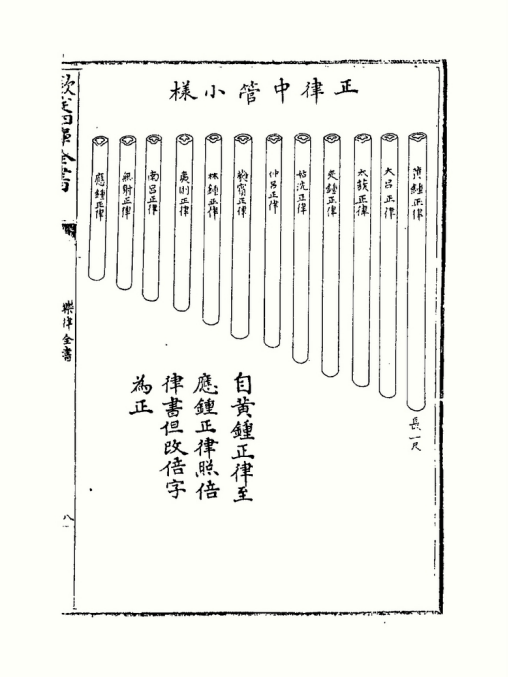


Figure 2.7: Diagram from Zhu Zaiyu's 'Complete Compendium of Music and Pitch' (1584) depicting pipes tuned to 12-TET by precisely choosing their length using $\sqrt[12]{2}$.

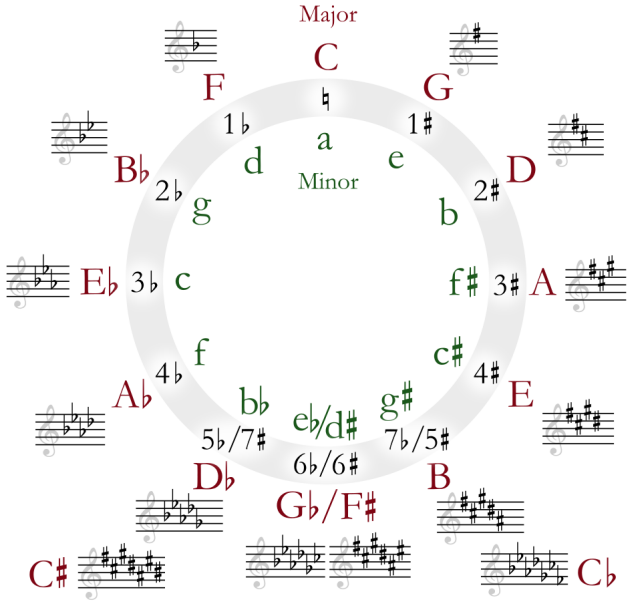
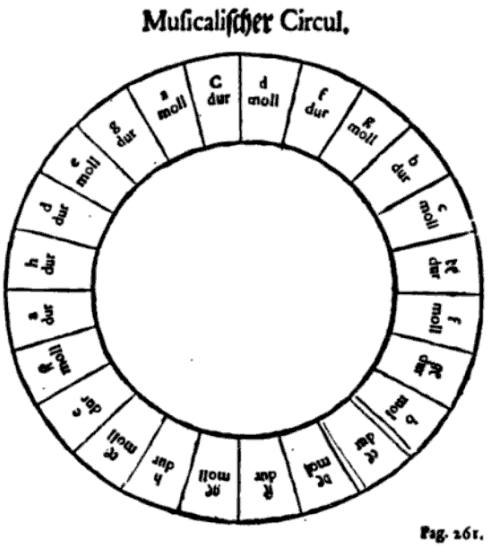
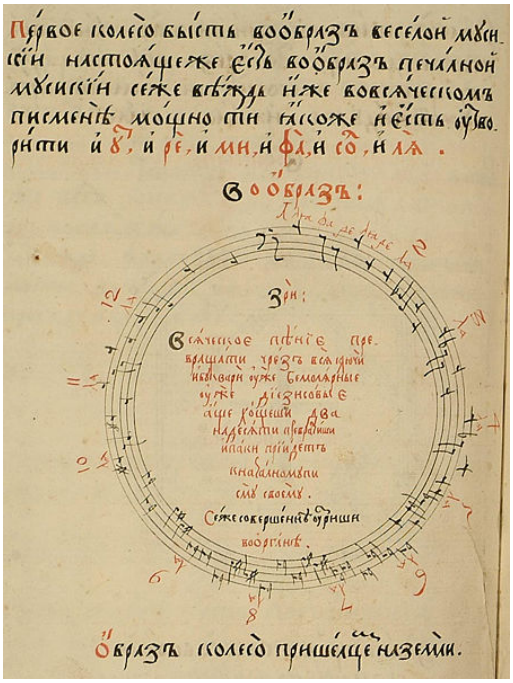


Figure 2.8: The circle of fifths, with the inner ring corresponding to the relative minor keys. Note that moving clockwise round the circle corresponds to a perfect fourth, and moving anticlockwise corresponds to a perfect fifth. Section 2.3 covers the relevant music theory in more depth.



(a) Heinichen's 'Musicalischer Circul' from 'Neu erfundene und gründliche Anweisung' (1711).



(b) Diletsky's Circle of Fifths from 'Idea Grammatiki Musikiyskoy' (1679).

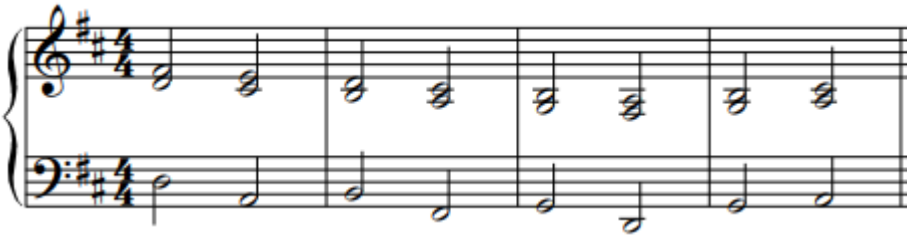
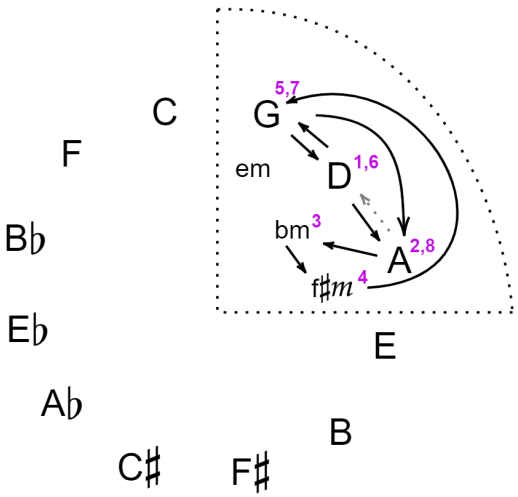


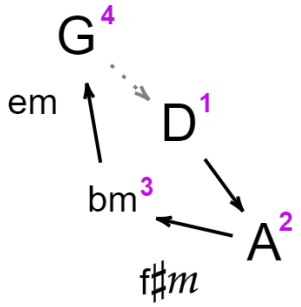
Figure 2.10: The opening four bars of Johann Pachelbel's 'Canon in D Major'.

a method by which to move between keys in a more natural fashion. Consider Pachelbel's canon in D major (Figure 2.10). The piece revolves around a repeated sequence of eight chords (corresponding to each half bar in the Figure) - D, A, bm, f♯m, G, D, G, A - over which there is significant thematic and textural growth as the piece progresses. Considering the movement of the piece through the circle of fifths (Figure 2.11a), it is not only clear that each transition involves two particularly closely-related keys - either moving a perfect fourth (clockwise), perfect fifth (anticlockwise), or by a semitone (f♯m→G) or tone (G→A).

Many composers and artists have sought to use a simplified (notably shorter) chord



(a) The path taken through a section of the circle of fifths by the repeating chords in Pachelbel's 'Canon in D Major'.



(b) Visualisation of the simplified (compared to Pachelbel's 'Canon in D Major') path taken through the circle of fifths by pieces following the I-V-vi-IV progression.

progression of D, A, bm, G (more generally known as the I-V-vi-IV progression). It can be found in pieces throughout history - though used most widely over the last Century or so. From Toto to the Beatles, James Blunt, and Lady Gaga - it appears extensively in popular music, likely due to its simplicity, or a predisposition in us humans that results in it being particularly aesthetically pleasing. Note that each transition here moves at most $\frac{1}{12}$ (i.e. one 'step') around the circle of fifths.

The circle of fifths will crop up throughout this Thesis (and music more generally), so it's well worth ensuring a firm understanding of it before moving on to Chapter 4. It's not the only geometric concept found in music though - far from it. J S Bach's (1685 - 1750) 'Crab Canon' (Figure 2.12) is another particularly interesting example. It consists of one line which essentially forms a musical palindrome - intended to be played by two performers (one from the start, forwards, and one from the end, backwards). To this end, it effectively lies on a Möbius strip. This canon was one of a number of 'riddle canons' presented in Bach's 'Musikalisches Opfer' (Musical Offering), and is largely considered to be the simplest to solve - the others involving augmenting (i.e. lengthening) notes, or modulating the piece.

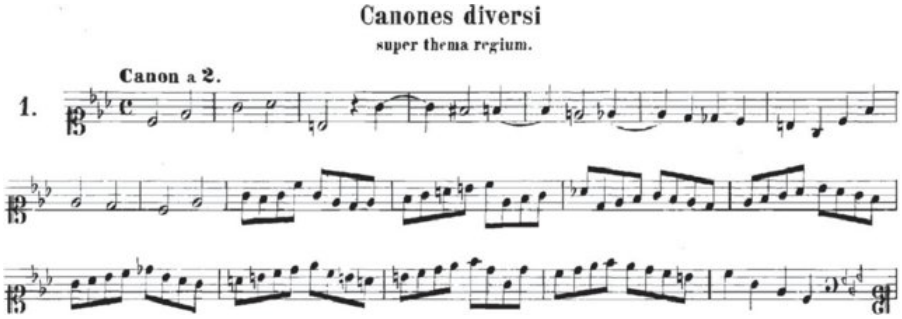


Figure 2.12: The original score for J S Bach's 'Crab Canon'.

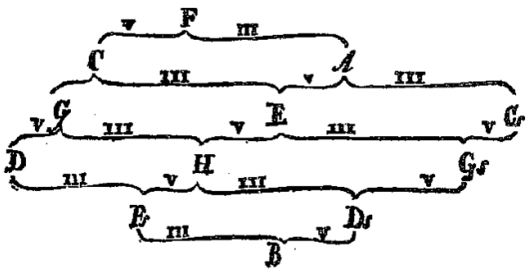
Before moving on to the close of this section, it is worth considering an observation from Claude Palisca's 'Scientific Empiricism in Musical Thought' [129, 61],

> "In any discussion of science in the seventeenth century, among the names that inevitably arise are those of Galileo Galilei, Marin Mersenne, Rene Descartes, Johannes Kepler, and Christian Huyghens. It is no mere coincidence that these . . . were all trained musicians and authors on musical subjects .. . because music until the seventeenth century was a branch of science and held a place among the four mathematical disciplines of the quadrivium beside arithmetic, geometry, and astronomy"

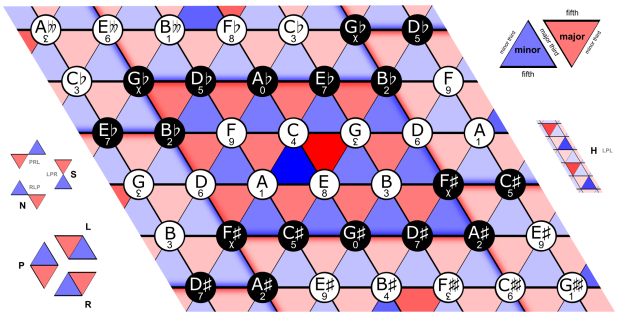
It really should be no surprise then to find so much mathematics (and indeed so many mathematicians) lurking in such vast swathes of music, and influencing musical thought, throughout the years.

We move finally to Leonhard Euler (1707 - 1783) - the legendary Swiss mathematician, whose work spanned a broad spectrum of fields, and occasionally crossed into music theory - albeit with a strong mathematical foundation. Most notably, his Speculum Musicum (Figure 2.13a) [55], a musical lattice, now known as the Tonnetz (Figure 3.8).

The Tonnetz presents a way to spatially demonstrate the relationship between chords. Each row corresponds to the circle of fifths, with each subsequent row corresponding to the



(a) Euler's Speculum Musicum as it appeared in his 'Tentamen novae theoriae musicae ex certissismis harmoniae principiis dilucide expositae'.



(b) A modern rendering of the Tonnetz, with red-filled triangles corresponding to major triads, and blue-filled triangles corresponding to minor ones.

previous one with each element shifted from position n, to position $(n + 3)$ mod 12, and aligned such that the closest two notes diagonally 'upwards' correspond to a major third in each direction. By identifying both vertically and horizontally, it is clear that the Tonnetz is in fact toroidal in nature. Particularly used now in non-Riemmanian theory [29], the Tonnetz encapsulates a number of important musical properties, including the major and minor triads, as well as 'efficient' voice leadings [163], which appear as pairs of triangles sharing an edge. More generally, distance between triangles (i.e. chords) on the Tonnetz corresponds to musical 'distance' between chords.

2.2 The Physics of Sound

The simplest sound (and indeed musical tone) is a pure sine wave (Figure 2.14). A sine wave corresponds to a single frequency, inverse to its wavelength, λ. In reality, sound consists of vibrations carried through a medium, in general with music, through compression and rarefaction of air (Figure 2.15) - a gaseous collection of dynamic, constantly moving particles, between which forces are exerted. Compression and rarefaction pertain in particular to air pressure which, put simply, can be seen as the concentration of molecules in a space of given size. With respect to our sine wave, compression and rarefaction correspond to the

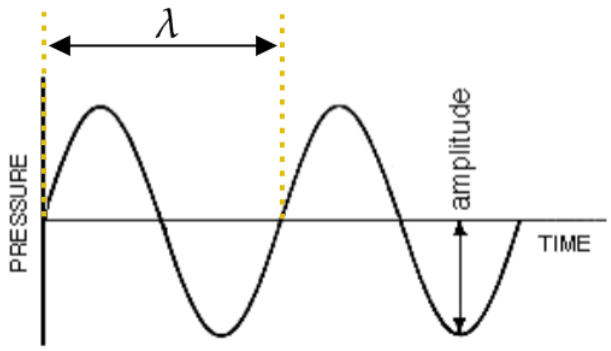


Figure 2.14: An annotated sine wave with wavelength, λ.

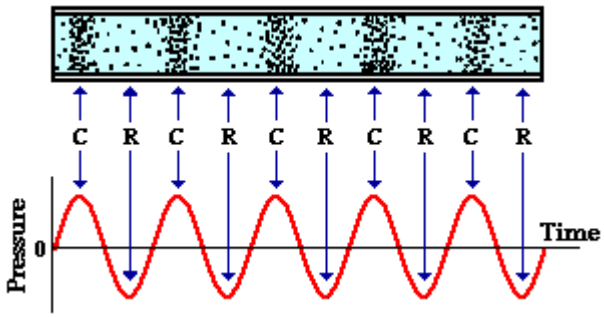
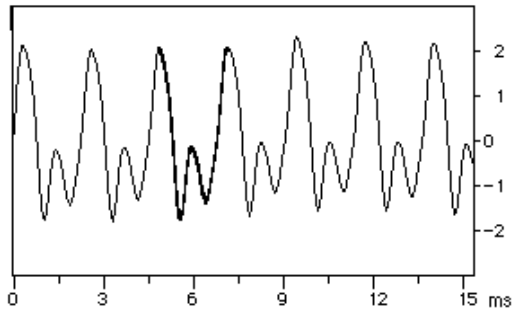


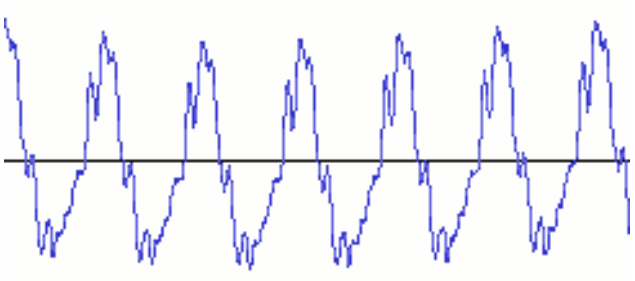
Figure 2.15: Compression (C) and Rarefaction (R) of a sine wave. Taken from [132].

nodes - points of greatest displacement from a relative pressure, generally taken to be the atmospheric pressure. There are musical parallels to each of these physical properties - namely with amplitude corresponding to 'loudness' of a certain pitch, and the frequency corresponding to the musical pitch (in a broad sense).

Figure 2.16a depicts the waveform of a flute sounding the note A4. It should quite quickly be noticed that this is significantly more complex than a pure sine wave, but still exhibits the periodic nature with which we have become accustomed. Consider then also the waveform of a violin playing the same note (Fig. 2.16b - though the period is the same, there are notable differences between the waveforms. These correspond to a property known as the timbre of the sound (broadly, the 'feel'), and account for why instruments sound different to one another.



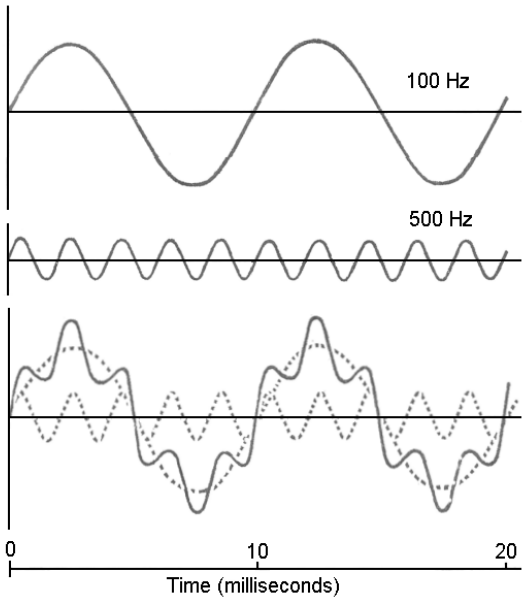
(a) Waveform of a flute playing the note A4 (440Hz). Taken from [60].



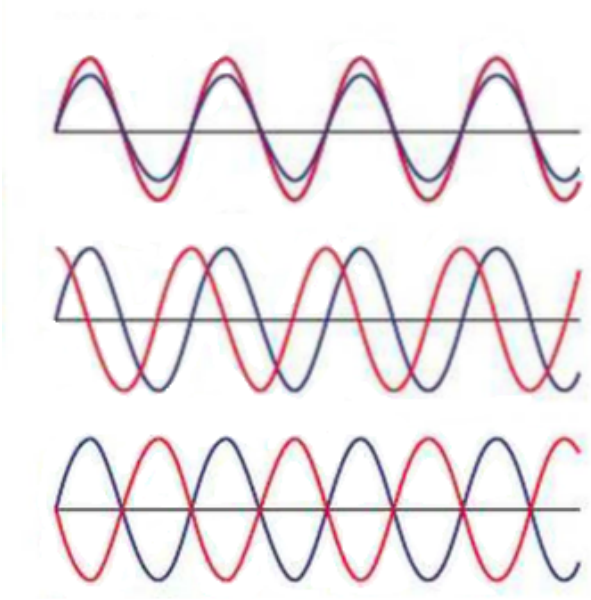
(b) Waveform of a violin playing the note A4 (440Hz).

How then (and why?) are single musical tones from instruments more than just a sine wave? To answer the "how?", we must consider the superposition of waves. Figure 2.17a shows a simple superposition whereby a more intricate wave is formed from the constructive superposition of two sine tones. Note that when considering or describing the position of two waves with respect to one another, it is eminently important to consider their phase (as depicted in Figure 2.17b). If the peaks and troughs of one wave are lined up with those of another (i.e. peaks aligned with peaks, and troughs with troughs), the two waves are said to be in phase. Conversely, if the peaks of one wave align with the troughs of another, they are said to be in antiphase. All other configurations of the two may described as being out of phase. One should now be tempted to ask "but where do these multiple waves come from when we are considering musical instruments?" - the answer to which lies in the concept of standing waves.

Most waves exist in an effectively infinite medium (i.e. the Universe), and thus, on the whole, it is relatively unlikely that any such wave should undergo superposition (or at least, regular superposition). Consider, however, a medium fixed between two stationary points - perhaps a violin string, fixed between the neck (onto which the string is being pressed by a performer) and the bridge, or indeed, the column of air present in woodwind instruments like the flute. The musical note (or fundamental) being played corresponds to the length of this medium, and the fixed points result in standing waves. In this case, superposition is



(a) Superposition of two waves. Taken from [41], adapted from [100].



(b) Two waves in phase, out of phase, and in antiphase respectively.

inevitable - particularly as wave are reflected off of one end (one of the two fixed points) of the medium, and back toward the other. In these cases, our waves undergo superposition as previously explained, and in most cases (i.e. when the two are out of phase) result in little of interest - with the waves essentially undergoing arbitrary superposition. When the two are 180° out of phase (that is, in antiphase), the superposition is completely destructive - resulting in a fixpoint (or node) - a point where the medium is completely stationary. In contrast, when the two waves are in phase, an antinode occurs - where the displacement of the medium is maximal. Figure 2.18 visually depicts the first three standing waves of a fixed medium.

It can be observed that these standing waves occur when their wavelengths correspond to certain factors of the size of the medium (i.e. the length of a string), with the fundamental (corresponding to the musical note being played) having a wavelength $\lambda = 2L$, where L is the aforementioned length. Such frequencies are referred to as resonant frequencies, or harmonics. The harmonics then occur with wavelengths that are integer divisions of this

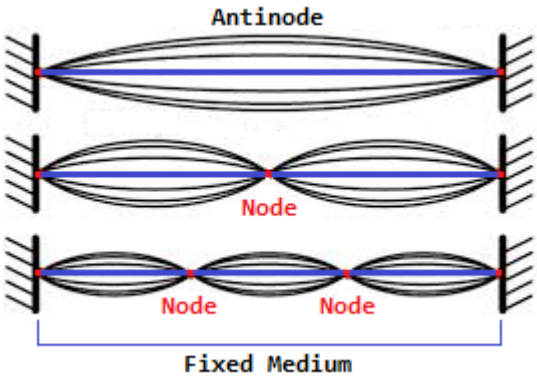


Figure 2.18: The first three standing waves of a fixed medium - namely the fundamental (f_0), first harmonic (f_1), and second harmonic (f_2), from top to bottom.

length, i.e. L, $\frac{2L}{3}$, $\frac{L}{2}$, $\frac{2L}{5}$, and so on. These harmonics, or overtones, and crucially the *amount* of each for a given note determines its timbre (and more broadly, the timbre of an instrument) - effectively the characteristics that set it apart from a pure tone. Timbre can be seen to be the 'sound' or 'feel' of an instrument being played in a particular way (note, for example, the difference between even bowing and plucking a string instrument such as a violin). It is what makes a flute sound different to a trumpet, and (one would hope), my voice sound different to yours. As explained in Section 2.5, however, the human brain is incredibly competent at filtering out ('abstracting') this information. Not only is this of use musically - for example, distinguishing between the melody and accompaniment of a song - but more generally, it is required in everyday life to abstract the timbre of human voices away. Without this, we would be unable to strip heard speech down to a more symbolic (perhaps *syntactic*) representation - that is, the words that make up human language.

Of course, it is largely these harmonics, and the many possible timbres, that render the problem of pitch estimation so difficult. Without the human body's seemingly innate ability to abstract timbral information, computers are left with a complex and hard to decompose (and therefore analyse) signal. Thus, there are two key approaches that could be applied to MPE - namely source separation of the input signal into monophonic signals, or

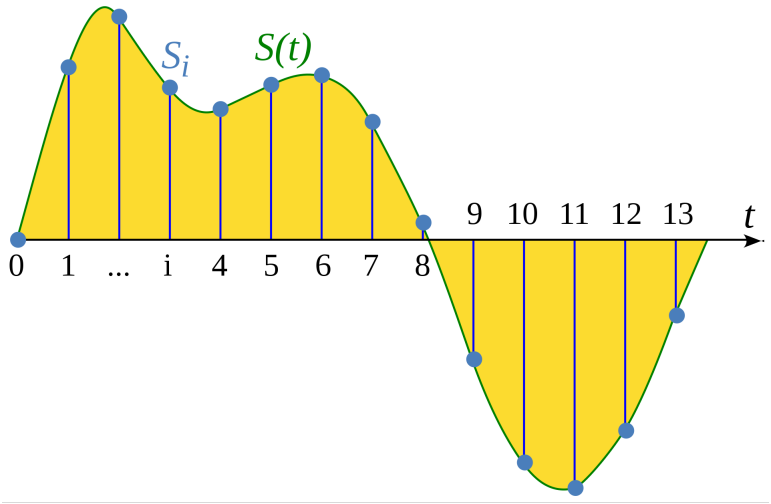


Figure 2.19: A demonstration of discrete sampling of a continuous signal. The green line ($S(t)$) represents the continuous signal, whilst the vertical blue lines, S_i, represent the samples.

some abstraction of the timbral qualities of a polyphonic signal. The latter is the approach taken in this Thesis - seeking to elicit patterns in polyphonic sound that are exhibited by all fundamentals, irrespective of timbre.

To round off this Section, it's important to understand how to take a signal from the 'real world' and computerise it. As sound is continuous, it is infinitely complex - irrespective of how far we may choose to increase the resolution at which we view it (i.e. zoom in), we could always look 'deeper'. How then do we represent sound computationally? We lack an infinite storage medium, and therefore it is necessary to store a *representation* of our continuous signal (i.e. whatever sound we're aiming to record), which sits at a sensible compromise between space (i.e. computational memory), and retaining an accurate (often referred to as 'faithful') likeness to the signal we have observed. Thus, we end up with a discrete signal, as shown in Figure 2.19. How many of these individual points - or rather, their density - is referred to as the sampling rate; the number of samples taken from a signal each second. The reader may indeed be familiar with the sampling rate of 44.1KHz (i.e. 44100 samples per second), which is often taken to be the standard sampling rate for audio.

As first proposed by Whittaker [178], a sample with sample rate x can only represent frequencies up to $\frac{x}{2}$Hz, and thus a sample rate of 44.1KHz allows us to represent frequencies of 22050Hz and below - *roughly* corresponding with the limits of human hearing. Thus this represents the ideal compromise for the vast majority of audio applications - after all, there's likely little reason to diligently store sound that we cannot in fact perceive. This is known as the Nyquist-Shannon sampling theorem, and broadly speaking states the sampling rate to capture *all* of the information from a continuous signal.

The remaining question then, is what will allow us to take the signal we want to record, and convert it into a computational format? The answer lies in a piece of technology that is actually rather ubiquitous in the present day - namely, the microphone.

The 'anatomy' of dynamic microphones (likely the kind you are familiar with in everyday life) is surprisingly simple at heart (Figure 2.20). A lightweight (and therefore quite sensitive to small changes in air pressure) diaphragm sits atop a coil of wire which itself surrounds a magnet. As air compresses and rarefacts near the diaphragm, the coil is moved 'up' and 'down' over the magnet, inducing a current in the wire proportional to the amount of movement of the coil, which then corresponds to a deviation from atmospheric pressure. These currents are then sampled to piece together a (discrete) waveform, as previously shown in this Section.

2.3 Fundamentals of Western Music Theory

Notes in western music are denoted as having two distinct properties (Figure 2.21): their pitch chroma, corresponding to the 'type' or 'sound' of the note; and their pitch height, which describes how high or low the note is. More mathematically, one can see the pitch chromas as elements of the set $\{C, C\sharp, D\flat, D, D\sharp, E\flat, E, F, F\sharp, G\flat, G, ...\}$, and pitch height as

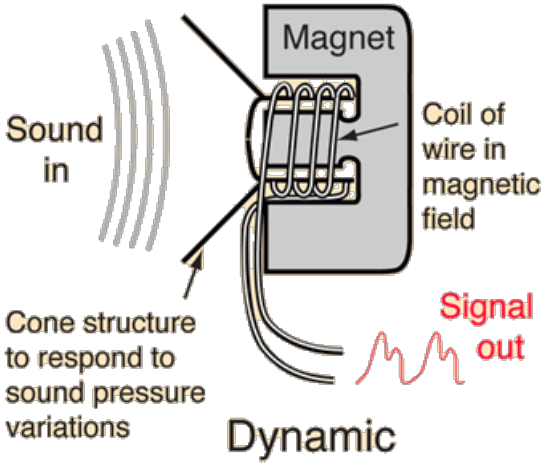


Figure 2.20: 'Anatomy' of a dynamic microphone. Taken from [1]

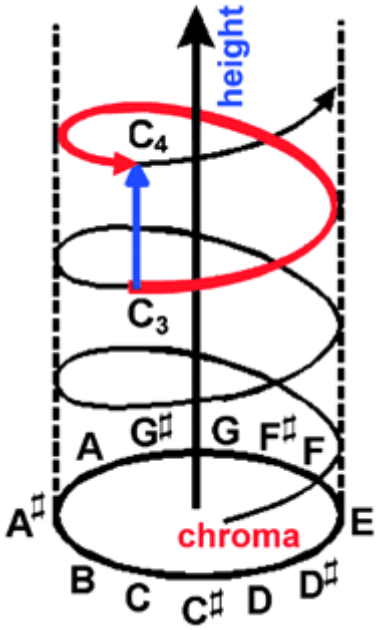


Figure 2.21: Pitch chroma/pitch height in a helical representation. Adapted from [175].

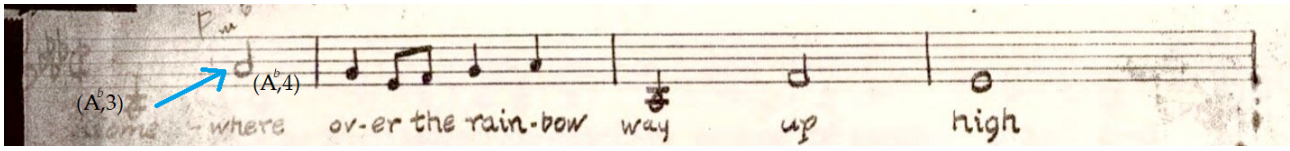


Figure 2.22: The first few bars of "Over the Rainbow" from the Wizard of Oz, with the initial interval highlighted. Adapted from the original score.

some integer $n \in \mathbb{Z}$. A note can then be expressed as a pair, e.g. $(G, 4)$. Considering each pitch chroma in an octave, it is clear that some of the aforementioned chromas correspond to the same tone (e.g. $C\sharp$ and $D\flat$). These are known as enharmonics.

Two notes with the same pitch chroma but different pitch heights generally sound similar, but of different height. Think, for example, of the two notes at the start of Harold Arlen's "Over the Rainbow" from the Wizard of Oz (Figure 2.22). The corresponding pairs in the original score are $(A\flat, 3)$ and $(A\flat, 4)$, and when listening to the piece, these two notes have the same 'sound', but a different height. In particular, the interval of one step (i.e. $3 \to 4$) in pitch height with an invariant pitch chroma is known as an octave.

There is really an ordering to the set of pitch chromas, as denoted in Figure 2.23. One step (e.g. $G \to A\flat$) is a semitone, and as the pattern repeats (i.e. $B \to C$ is itself a semitone), it may be observed that twelve semitones constitute an octave. Thinking back to our physical knowledge of sound (Section 2.2), what then do these steps correspond to? Consider that each note corresponds to a particular frequency. The reason that notes an octave apart sound similar but with different pitch height is due to the fact that a note one octave up from another has precisely double the frequency. For example, $(A, 4)$ has a frequency of $440Hz$, and $(A, 5)$ has one of $880Hz$. Given this relationship, each ascending semitone can therefore be expressed as a multiplication of the frequency by $\sqrt[12]{2}$. For example, from $(A, 4)$ to $(B\flat, 4)$, we reach $440 \times \sqrt[12]{2} \approx 466Hz$.

In order to discern the notes (i.e. pitch chroma/pitch height pairs) from a written piece of music, it is necessary to have a look at the notation commonly used.

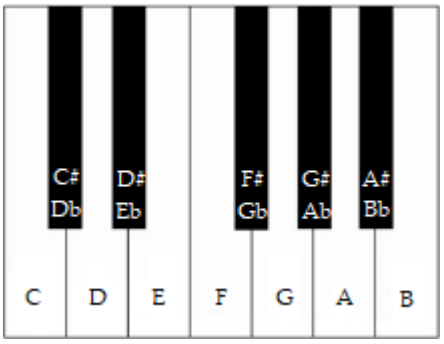


Figure 2.23: The twelve semitones in an octave, demonstrating the ordering to the pitch chromas.

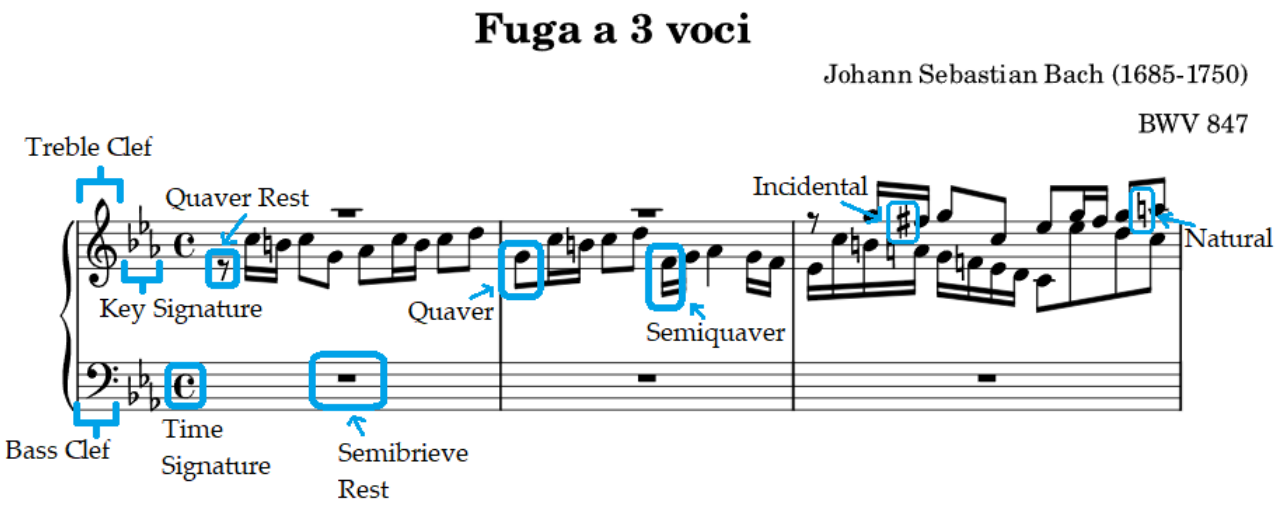


Figure 2.24: Annotated extract from Bach's BWV 847, highlighting the key parts of musical notation present.

Figure 2.24 shows an extract from Bach's BMV 847, with many of the parts of musical notation labelled. Each stave consists of five parallel lines (and four gaps)—on which notes, rests, and other notation sits. Notes falling above or below the staff are depicted using ledger lines (e.g. the first note in Figure 2.22).

The notes that the lines and spaces of the staff correspond to are reliant on the clef. Generally one of three is used—the Treble clef, Bass clef, and (rarely) the Alto clef. the centre of the Treble clef sits on a line corresponding to $(G, 4)$, and the $\bullet$ of the Bass clef sits on a line corresponding to $(F, 3)$.

Each note is represented by a symbol—the position of which corresponds to its pitch chroma and pitch height. The symbol itself corresponds to the note's duration. For example, quavers, ♪, (eighth notes), and semiquavers, ♬, (sixteenth notes), as shown in Figure 2.24. Further, a semibrieve, 𝅝, is a 'whole note', a minim, 𝅗𝅥, a 'half note', and a crotchet, ♩, is a 'quarter note'. In order to obtain all possible note lengths, notes can be dotted (e.g. 𝅗𝅥.), which corresponds to adding another half of their original length—so a dotted minim is a 'three quarter note'. In addition, notes can be tied (by a curved line), which sums their durations together. For each note length, there also exists a corresponding rest that can be used to denote a lack of notes at a given point.

The time signature of a piece predominantly denotes the duration of notes within each bar. For example a time signature of 3/4 indicates that there are 3 quarter notes (i.e. crotchets) in a bar. In the example, 𝐜 stands for 'common time', and corresponds to a time signature of 4/4—i.e. four crotchets in a bar.

Finally, the key signature of a piece shows which of the notes (i.e. lines and spaces of the staff) should be sharpened or flattened. With no key signature, the lines and spaces of the staff correspond to the 'white notes' on a piano. In the example, the key signature has three flats, meaning that all Bs, Es, and As are flattened, which is indicative of either the key of $E\flat$ major, or the relative minor, C minor.

In order to represent notes that fall outside of the key, incidentals—♯, ♭, 𝄪, 𝄫—are used. They are applied on the line or gap corresponding to a note, directly before it, and apply until the end of the current bar. In addition, a natural, ♮, negates the key signature and any incidentals for the given note until the end of the bar.

Figure 2.25: An original Spirograph set. Taken from [123].

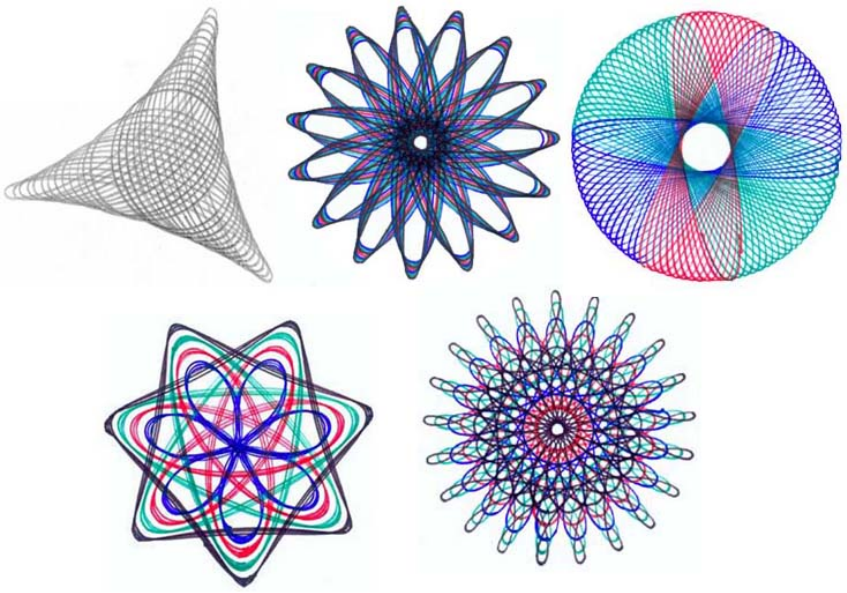


Figure 2.26: Various designs drawn using a Spirograph, demonstrating just some of the possible complexity achievable, even with the toy.

2.4 The Fourier Transform

The Fourier transform, first touched on by Fourier in his in his Théorie analytique de la chaleur [62], allows us to take a time-domain signal, and translate it into the frequency domain. Developed by Denys Fisher, and first sold in 1965, the Spirograph (Figure 2.25) is a children's toy and geometric drawing device, which incidentally ends up providing a rather good analogy for explaining the Fourier transform in an intuitive fashion.

Consider a (very boring) Spirograph with just a single circle. We can represent this as an orbit around the origin in complex space (Figure 2.27) as

$$re^{i\omega t}, \tag{2.1}$$

where r is the radius of the circle around which the point is orbiting, ω is the angular velocity at which it is moving, and t is the point in time at which we're observing the system.

Now imagine adding a gear to our Spirograph, so there is now another circle added to the orbit—your pen now orbits the gear, and the gear orbits the larger circle (Figure 2.28). This can then be expressed as

$$r_0e^{i\omega_0 t} + r_1e^{i\omega_1 t}, \tag{2.2}$$

where each orbit has its own radius and velocity.

Further, consider now that one year, you were a particularly well-behaved and auspicious child, and Santa—in his infinite wisdom and generosity—has gifted you an infinitely-large, infinitely-complex Spirograph, with as many gears as you'd like (*those poor, poor elves...*)! Following the same pattern, we could then represent the orbit as

$$\sum_{n=0}^{\infty} r_n e^{i\omega_n t}. \tag{2.3}$$

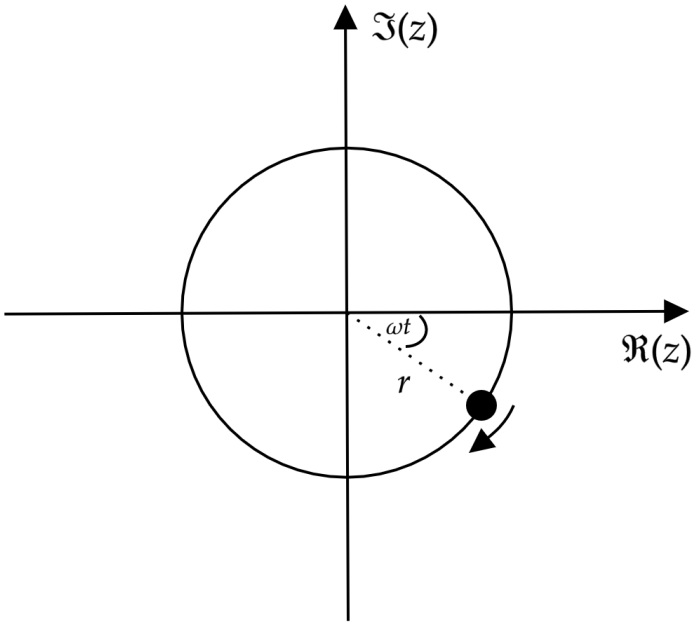


Figure 2.27: The orbit of a point around a single circle in the complex plane.

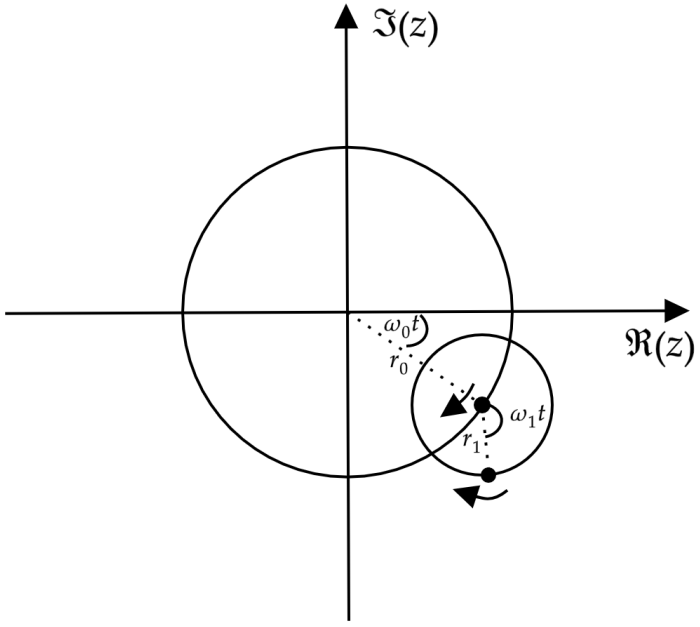


Figure 2.28: The orbit of a point around a bi-cyclic path in the complex plane.

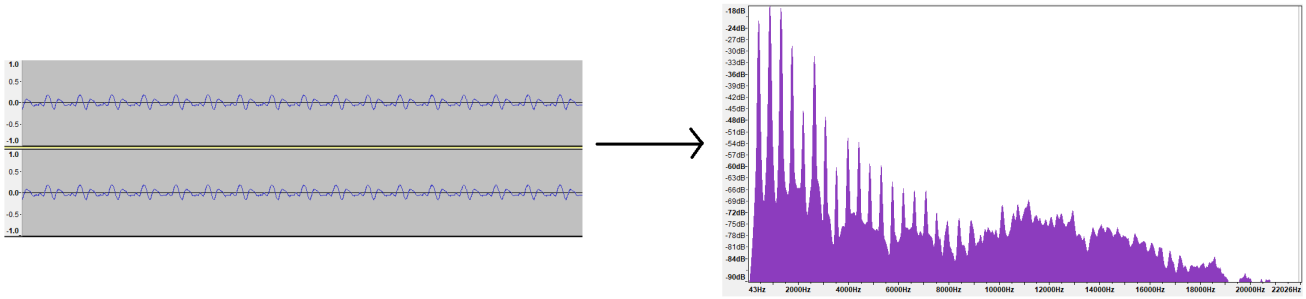


Figure 2.29: Fourier Transform (right) of the waveform of a Flute playing A440 (left).

Thus, as long as you can convince yourself that you can draw *anything* with your snazzy new Spirograph, and by Euler's formula,

$$\sum_{n=0}^{\infty} r_n e^{i\omega_n t} = \sum_{n=0}^{\infty} r_n isin(\omega_n t) + r_n cos(\omega_n t), \tag{2.4}$$

then all (continuous periodic) functions can be expressed as a sum of their sines and cosines.

Thinking back to the idea of superposition (Figure 2.17a), this is equivalent to finding the amplitude of functions with each frequency. In a musical context, this means that we can take a time-domain signal (e.g. a waveform), and work out how much of each tone is present (e.g. Figure 2.29).

Note that we cannot actually record, store, nor process infinitely-large amounts of data. Instead, we have to use a discrete Fourier transform. Because this no longer utilises a continuous signal, it introduces error in the form of spectral leakage [72]. This can be alleviated through the use of windowing, and also by increasing the sampling frequency.

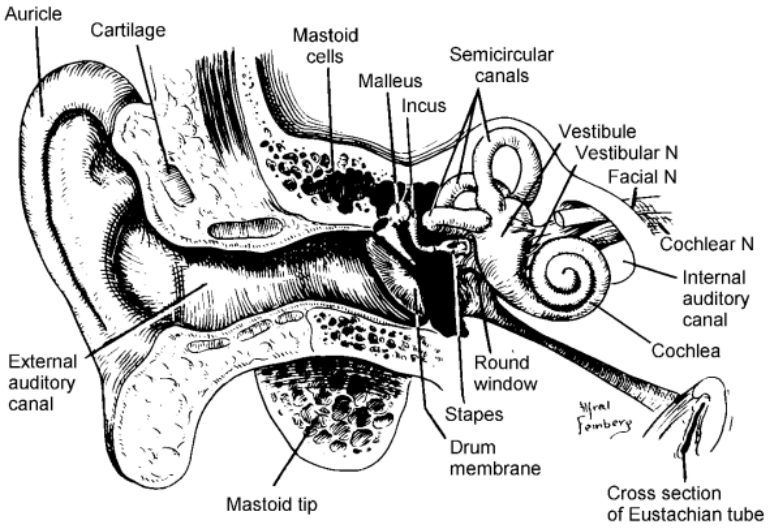


Figure 2.30: Anatomy of the human ear, showing the three main sections. Taken from [2].

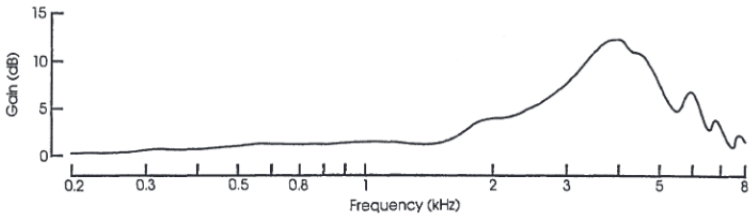


Figure 2.31: Graph showing the variable gain caused by the external auditory meatus across the audible frequency spectrum. Taken from [6].

2.5 Sound, the Ear, and the Brain

The human ear (Figure 2.30) is split into three parts: the outer, middle, and inner ear, respectively [73].

The outer ear is known as the auricle (or pinna) - the funnel-shaped skin and cartilage that is visible protruding from the head. The auricle serves two purposes - both the amplification of, and localisation of, sound [114]. It achieves amplification by acting similarly to a parabolic reflector [59], focusing the sound into the external auditory meatus, or ear canal. The effect of sound localisation is predominantly caused by the shape of the auricle. For low frequency sound, it simply focuses it into the ear canal. For higher frequency sounds, however, much of it bounces around the auricle, whilst some immediately reaches the auditory meatus. This results in a temporal delay between the arrival of different sounds to the inner ear, allowing for some assertion to be made on the direction. This is because the auricle is angled forward - sound from the front is more likely to enter the ear canal directly, whereas sound from the side or behind is more likely to bounce around the auricle, thus taking longer.

Not only does the external auditory meatus direct sound to the tympanic membrane (the boundary between the outer and middle ear), it also provides some amplification to the sound. This is a result of its shape, which perhaps more interestingly induces variable amounts of gain across the frequency spectrum (Figure 2.31) - a property first noted by Wiener and Ross [179].

The middle ear consists of three sections - the tympanic membrane, the eustachian tube, and the ossicles, a set of three tiny bones known as the malleus, incus, and stapes. It is an air-filled cavity, connected to the back of the nose by the eustachian tube (which regulates pressure in the ear). Sound is conducted from the tympanic membrane, and across the malleus and incus to the stapes, the base of which 'plugs' the oval window - the gateway

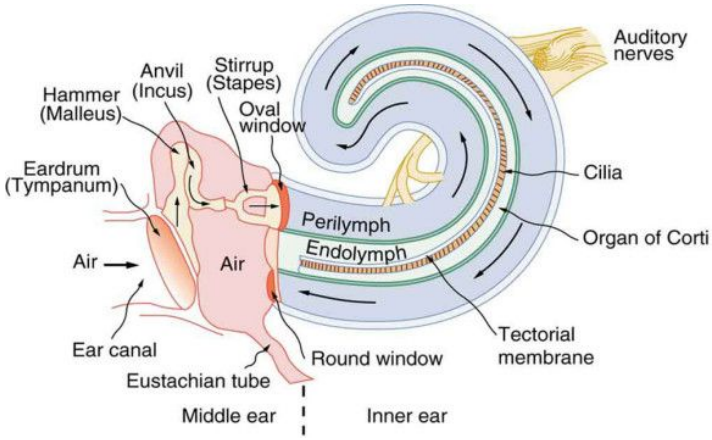


Figure 2.32: Diagram showing the middle and inner ear sections

to the inner ear. Each of the ossicles that make up the ossicular chain are suspended in the cavity by miniscule ligaments, allowing them to freely vibrate - effectively carrying the vibrations through the middle ear.

The inner ear is comprised of two labyrinths - a membranous labyrinth, encapsulated within a bony labyrinth. Part of this bony labyrinth is a shell-shaped tube called the cochlea, which is fundamentally the organ of hearing. The cochlea is filled with a virtually incompressible fluid called perilymph, which has a similar composition to cerebrospinal fluid [23].

The inside of the cochlea is lined with tens of thousands of tiny hairs, which vibrate, and induce impulses along the cochlear nerve - essentially transducing vibrations into nervous impulses corresponding to the frequency of the sound. To this end, the cochlea is essentially performing a Fourier transform on the perceived sound - taking it from the time domain into the frequency domain.

In many ways, the ear functions much the same as a dynamic microphone (Figure 2.20)—sound waves are transduced by the tympanic membrane, and corresponding impulses are passed via the cochlea to the brain (via the eighth cranial nerve [5]). The spiral ganglion neurons that are involved in this transmission of data (including intensity, frequency,

and timing) to the brain, accomplish this task through a complex neural circuit, and also utilise an active feedback loop from brain to ear in order to improve performance. It may well be the case that it is possible to use similar adaptations as the ear in computational approaches to audio analysis.

Chapter Three

Related Work

This Chapter of the Thesis focuses on the research and other work in the field of Music Information Retrieval that has been conducted in the past, with a particular focus on pitch estimation - both monophonic, and polyphonic. The literature is approached in a broadly chronological fashion, and split into five rough sections: *Pre-MIR* (Section 3.1), a brief glance at the early work leading to the field as a whole, including the psychoacoustic research of the 1800s and early 1900s, and the research of speech signals necessitated by study of telephony; *Early MIR* (Section 3.2), looking at the emergence of MIR as a standalone field, and the first applications of computers as we know them to such problems; *MIR in the 2000s and Early 2010s* (Section 3.3), a relatively in-depth look at techniques toward Multi Pitch Estimation, amongst other applications, many of which have formed the basis of current work; *State of the Art MIR* (Section 3.4), which presents cutting-edge developments from over the last four years, bringing the reader up to speed with present methods and work; and finally *Other Related Work* (Section 3.5), encapsulating a few works that provide context or motivation for the research of this Thesis.

3.1 Pre-MIR

As put by Crandall [34], the work of German physicist Hermann von Helmholtz really marks the beginning of scientific study of speech. His compiled work on psychoacoustics, "The Sensations of Tone" [76] was approached from both a physical, and from a physiological perspective. In early study, particularly of the production of speech as well as its perception, psychoacousticians were largely split into two camps - the "Harmonic Theory", championed by Helmholtz, purporting that speech (much like harmonic musical instruments) consisted of fundamentals and harmonics, amplified by the oral cavity amongst other resonant aspects of the human anatomy. On the other side was the "Inharmonic Theory" of Willis [180], which claimed that there are not necessarily harmonic relationships between the vowels observed in speech and the fundamental frequency producing them. At least as of Crandall's paper, not far from a century from the initiation of this debate, there was certainly no resolution - with both sides having contributed massively to the field. Perhaps it is no surprise then, that Babbitt [7] refers to auditory perception as the "most refractory of areas". A full treatment of the psychoacoustics underpinning much MIR is beyond the scope of this Thesis, but Kursell [99] provides a fantastic account should the reader be interested.

Crandall's paper goes on to describe the pioneering work of Alexander Graham Bell, who undertook much work in visualising and understanding the vibrations of sound, with a focus on speech. In 1874, rather than using a model ear, Dr. Clarence Blake convinced Bell to use the human ear itself, which allowed him to obtain accurate tracings of the sounds that the contraption had 'heard' (Figure 3.1). Such an understanding, afforded by the ability to visualise signals, was crucial to improving the telephony pioneered by Bell, and over the years, much of the work done in the area has proved somewhat applicable to parallel problems in MIR.

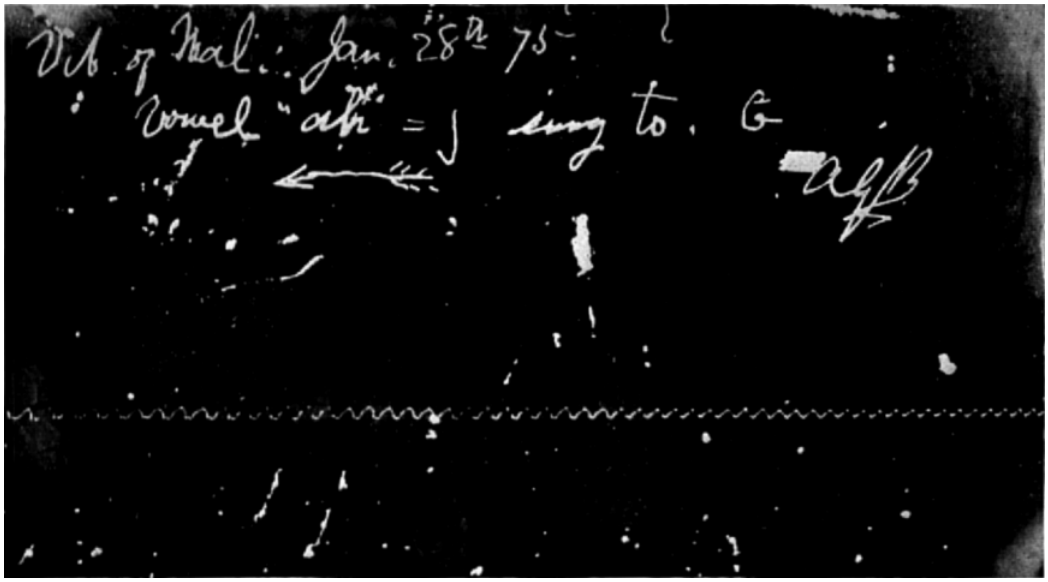


Figure 3.1: A tracing of human speech, made by Bell in 1875. Taken from [34].

3.2 Early MIR

Perhaps the first recorded study of mechanical or computational MIR is in 1949, when Bronson [19] sought to use the International Business Machine (IBM) to facilitate research over a catalogue of folk songs that would be completely intractable for a human to perform. As put in the paper,

> "The machine is not asked to do what a machine should not attempt: it cannot solve aesthetic problems on the basis of figures, but where facts and figures are necessary, it can give factual answers with a startling economy of time and effort, and free the student for questions of a higher order. Successful use of the method to be outlined depends on the thoughtful and sensitive analysis of melody, and follows, not precedes, such analysis."

Even now, this is an astoundingly pertinent point, and largely reinforces the place of computers in musicological research - to perform those tasks that human researchers are

not necessarily required, and allow them to instead investigate deeper and more intellectual considerations. Bronson's approach involved encoding certain properties of the folk songs to be analysed (for example time signature and range) on individual IBM punchcards, and then utilising the machine to sort these based on musicological queries. It was posited that this kind of technique could be used to analyse similarities, and derive facts and statements from the dataset without entrusting the task to humans. Though a seemingly small step towards MIR as we know it today, this work laid the foundations for a range of investigations in the second half of the 20^{th} century.

Babbitt gave a fantastic review and snapshot of the state of the art as of 1965 [7]. He presents mathematically interesting problems and insights around the concept of all-interval twelve-tone sets (amongst other things), and shows how computers can be used to provide clarity. In addition, he postulates on the possibility of having "complexly" searchable works in a computerised fashion, such that they could be used for analysis (akin to Bronson's work), and lays out the role of computers as "information amplifiers". Further, he talks about similarities between concepts or problems of musical style, and those of linguistic style - pointing in particular to parallels with Mosteller and Wallace's analysis of the Federalist Papers (of which the authorship had been a source of much debate) [121], the techniques behind which he believed would be applicable to MIR research too. The paper concludes by musing on the sheer difficulty of many problems in MIR, stating,

> "And, if one wishes to reach the analytical point where all of the dimensions of musical events are considered in their interaction, how many paths must there be through a realistically constrained musical work, when it has been estimated that there are probably some 10^{120} paths through a game of chess?"

Perhaps too, the speedy evolution of computing power thanks to the now-waning Moore's law [101] has left us dull to such considerations in this day and age. Particularly

with the boom of Machine Learning methods, as seen in Section 3.4, this kind of rumination on computational resource, especially as it is so readily available in unfathomable volumes nowadays, seems to be a rarity.

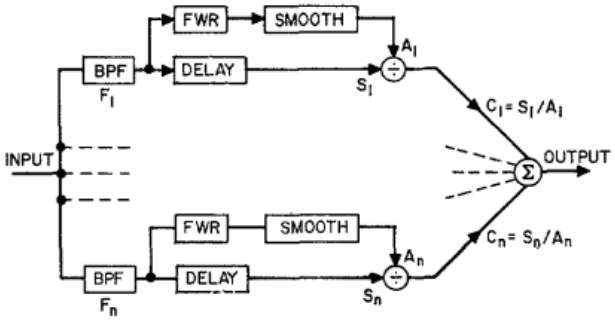
Coming from a completely different angle, Kassler [86] puts forward a 'formal' language for encoding and resolving MIR statements on computerised data. At the time, this was likely quite groundbreaking, and came along with a number of 'query languages' designed to do the same - namely LMT and IML [7]. Concluding his work, Kassler says,

> "Musical information retrieval, as I conceive it, is the task of extracting, from a large quantity of musical data, the portions of that data with respect to which some particular musicological statement is true."

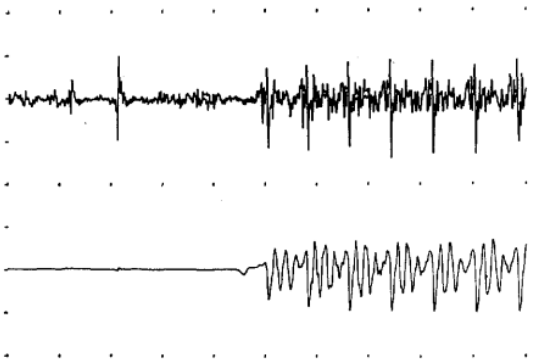
To some extent, this could seem somewhat myopic of Kassler, but one has to remember that this was from the perspective of taking computerised *scores* of music, and eliciting information from them. Through the lens of hindsight, we now know that the field has evolved to include the analysis of raw signals too, but this was likely no more than a distant thought in the context of the time - in the same way that truly complete Automatic Music Transcription (AMT) is now.

In his 1967 paper, Forte [61] divides music and computing as a field into two distinct subfields - namely "sound-generation by computer", and "music research", which encapsulated MIR, as well as the more musicological applications such as style analysis, study of musical systems, and computational representation of music. He draws an interesting distinction between 'numeric' and 'non-numeric' processing, with the former generally dealing with incomplete data (e.g. raw audio), and the latter with complete data (e.g. faithful representations of musical scores in a computerised format).

Looking at Pitch Extraction (PE) from the perspective of speech, Man Mohan Sondhi



(a) A schematic for a spectrum flattener. Taken from [157].



(b) An example of spectrum flattening (top signal) being applied to a speech signal (bottom). Adapted from [157].

- a researcher at Bell laboratories - presented three new methods of extracting fundamental pitch from a raw audio signal containing speech [157]. All three of these were preceded by a technique known as spectrum flattening (Figures 3.2a and 3.2b). This effectively causes the voiced intervals (i.e. those aspects of speech with harmonic components) to display 'phase trains' - a sequence of noticeable and periodic sharp peaks that allow for the determination of a fundamental frequency. Sondhi further showed that the latter two methods - both of which used autocorrelation - were able to perform under surprisingly difficult conditions (i.e. in the presence of noise, or with poor audio quality). What should really strike the reader here is the sheer contrast between this research, and the MIR research present at the time - whilst those working in MIR are still getting to grips with the very basics of computational research, scientists working on speech and telephony are developing novel mathematical and electronic approaches with definite potential for parallel application in MIR. This lag between the fields shortens as MIR grows into maturity, but is certainly stark at this point. Later in 1977, Rabiner also presented his research on the use of autocorrelation analysis for pitch detection, along similar lines to Sondhi [137].

Mendel rather humbly delves into his work with Lockwood on the "Josquin Project" - an attempt to perform analysis on the musical style of composer Josquin Desprez [113]. He talks in particular about using computers to automate menial or tedious tasks over large

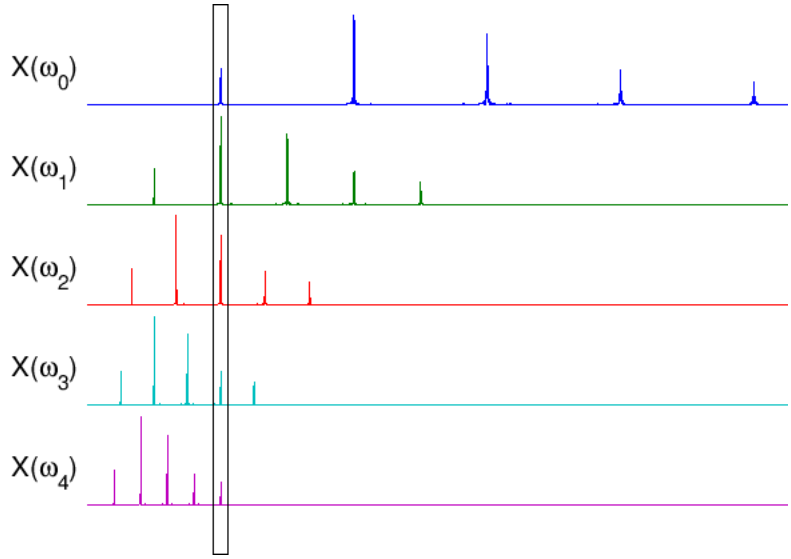


Figure 3.3: Visual representation of Noll's HPS algorithm. Taken from [156].

datasets, and as with Babbitt back in 1965, sought to do so in order to free up human researchers to answer higher-order musicological questions. In doing so, he also highlights one of the major issues of this kind of research - the need to manually encode hundreds or thousands of musical works in a computational format, and the human error that inevitably follows. The sheer time and effort that not only went into creating punchcards, but also proofreading them is astounding, and certainly makes clear that though it may save the *researchers* from menial tasks, it is difficult to abstain from them in their entirety. In fact, this is touched on by Lincoln - twenty years on from Bronson's paper - who asserts a number of limiting factors, noting that *"Programming, keypunching, and computer time are all expensive and time-consuming at present"* [104]. Interestingly for his time, Lincoln also muses on the concept of 'melody querying' (think Shazam, Soundhound, etc.), and the possibility that we might one day be able to search for music with snippets of music - indeed, we can.

The Harmonic Product Spectrum (HPS) proposed by Noll [126] (Figure 3.3) provides another good example of an approach initially designed for Pitch Extraction of speech that has proved quite effective when applied to musical signals too. Working frame by frame, the

algorithm computes

$$Y(\omega) = \prod_{r=1}^{R} |X(\omega r)|$$

,

for a given frame X. From this output spectrum, Y, a maximum value,

$$\hat{Y} = \max_{\omega_i} \{Y(\omega_i)\}$$

,

is computed, which corresponds to the choice of fundamental frequency. That is, for each harmonic being taken into account, the frame is mapped to a corresponding spectrum such that each peak moves to a frequency equal to the product of the frequency multiplied and the inverse of the just intonation interval for the harmonic - i.e. for the first harmonic, the octave, each peak is relocated to the frequency at $\frac{1}{2}$ its current frequency. A peak at 440Hz becoming a peak at 220Hz with the same amplitude, and so on. For the second harmonic, the multiplication is by $(\frac{3}{2})^{-1}$, and so on. Finally Y is formed from the product of all of these spectra. Hence, the peak at the fundamental will remain large in the output, but the peak of every harmonic will be essentially zero. The highest peak in the output spectrum, $\hat{Y}$, is selected as the fundamental.

One of the first pieces of work to attempt Automatic Music Transcription (AMT) in any form was Moorer's doctoral thesis, "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer" [119], which tackled transcription of signals with some notable restrictions,

- Each signal consisted of at most two independent voices;

- There was no vibrato or glissando present in the signals;
- No note was shorter than 80ms in length;
- and the fundamental frequency of a sounding note could not coincide with a harmonic of another simultaneously sounding note.

As to be expected from one of the first forays into the problem, these are quite significantly restrictive assumptions, but clearly allowed Moorer to take this first step within the remit of a doctoral programme of study. The proposed method utilises a directed bank of sharp-cutoff bandpass filters, followed by inference from the set of detected possible fundamentals to select the "best" note hypotheses, which are then passed to a manuscripting program to produce a score. Though the system performed well on the input given, it was noted that the computational cost was extreme (at the time of writing), and therefore the system itself was impractical for widespread use. In addition, Moorer presented a number of useful insights, in particular that,

> "...there is considerably more activity in a piece of music than is perceived by the listener. This is especially common with stringed instruments, because the strings that are not being manipulated invariably resonate and produce sounds independently which are generally not heard due to aural masking."

This phenomenon is well-documented, and proves frustrating at best when dealing with acoustic instruments, particularly stringed instruments as identified.

In a 1977 paper [118], Moorer summarised much of his thesis, and poses an interesting question—*when do a group of harmonics fuse into a single percept (note in our usage) and when do they remain distinct?* The answer to this isn't at all clear, but quite interestingly may well be related with some of the work done to distinguish between fundamentals and 'false fundamentals' in this Thesis.

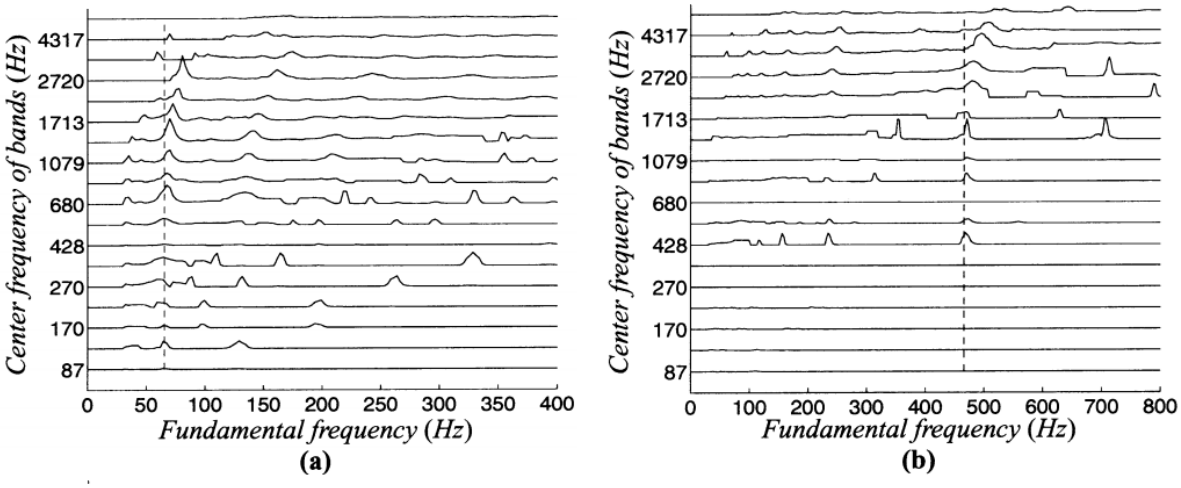


Figure 3.4: Bandwise-calculated weights of two piano tones, (a) f_0 = 65Hz and (b) f_0 = 470Hz. The actual pitches of the harmonics are depicted by the vertical dashed lines. Adapted from [93].

3.3 MIR in the 2000s and Early 2010s

One of the most time-honoured approaches to Pitch Extraction in the literature is YIN—a pitch estimator for both speech and music, developed by Cheveigné and Kawahara [39]. This was based off of the Autocorrelation Function (ACF), and utilised a number of additions to help to distinguish between harmonic and inharmonic components of sound. At the time of publication, YIN was an order of three more effective than other approaches, and to this day is used as a benchmark for MPE systems.

In the same year, Futrelle and Downie of the University of Illinois set out expectations of MIR as an area of research, and emphasised the necessity for researchers to make it easier to compare their work to one another [64] - a notion that Downie reinforced and expanded on in a future paper [44]. Given the appetite in the field almost twenty years on to do precisely what they called for, it seems clear that this call for more structured testing and comparison had a positive impact on the MIR research community - with datasets such as MAPS [52] and Bach10 [36] widely used in the present day.

Early in the 2000s, Klapuri presented three iterative methods for pitch detection based on harmonicity and spectral smoothness [93], human perception [92], and summing harmonic amplitudes [90] respectively. In [93], he visually demonstrates the inharmonicity phenomenon described earlier by Moorer (Figure 3.4). This deviation from the 'ideal' harmonics, in particular when reaching higher harmonics, led partly to the choice to look solely at the first three harmonics (of a given fundamental) in the research of this Thesis.

A number of papers over the next few years provide a brilliant summary of the state of play in MIR heading into the 2010s [165, 4, 122]. Typke, Wiering, and Veltkamp [165] focus in particular on 'content-based' systems, which broadly could be seen as 'higher-order' in comparison to pitch estimation, such as Musipedia (formerly 'Tuneserver') [136] and Shazam [171], whilst Amado and Filho [4] describe two historically-used techniques of interest - namely the Zero-Crossing Rate (ZCR) and ACF, presenting some improvements to these to improve their efficacy specifically for Pitch Extraction of musical signals as opposed to speech.

In [122] a particularly thorough overview is given, pointing at reviews of early approaches to Pitch Extraction of speech [79, 78], and a range of earlier attempts at Pitch Extraction of musical signals, demarcated by tactic—frequency domain approaches [147, 108]; time domain approaches [160, 22]; and mixed domain approaches [112, 131]. Further, they give an overview of current work [172, 91, 25], again grouped by approach—some working in the spectral domain [9, 26, 151, 135, 107, 10]; others working with Non-Negative Matrix Factorisation (NNMF) [166, 167, 154]; some with timbral information [85, 50, 143]; and finally, a couple that utilised knowledge of the mechanisms of auditory perception in humans [89, 182]. Importantly, Müller, Ellis, Klapuri, and Richard note that

> "...to be successful, music audio signal processing techniques must be informed by a deep and thorough insight into the nature of music itself."

This is a thought-provoking and pertinent insight to this day, and should remain at the forefront of work in MIR. Without a deep understanding of the work that we are performing, and further, the data that we work with, it is unreasonable to expect good results. Especially as we look for increasingly sophisticated and effective methods of AMT, we must constantly bring ourselves back to this notion.

Böck and Schedl [17] used a Recurrent Neural Network (RNN) to determine both pitch, and temporal onset and offset of notes in piano music. They opted for this approach over a Feed-Forward Network (FFN) as the hidden layers of RNNs have connections that 'loop back', allowing them to more effectively model the temporal aspects of music.

A 2012 paper by Volk and Honingh [168], much like Futrelle and Downie's earlier paper, looked to set out challenges and opportunities facing MIR as a research community. In it, they discuss mathematical and computational approaches to music in general, and collate the viewpoints of panellists from the "Bridging the Gap" panel at MCM 2011: Alan Marsden, Guerino Mazzola, and Geraint Wiggins, who were prompted around four key areas: benefits, failures, challenges, and interdisciplinary discourse.

Kraft and Zölzer [95] propose a method building off of the ACF, which selects peaks (i.e. local maxima) that are present in both a "spectral peaks" and salience representation. This method of taking the product of multiple spectra is quite reminiscent of Noll's HPS algorithm.

To round out the 'early' 2010s, we consider a paper by Sigtia, Benetos, and Dixon [153], which presents a supervised NN model for piano transcription, which splits the problem into two broad parts - an "Acoustic Model (AM)", and a "Music Language Model (MLM)". The acoustic model is a NN that estimates the probabilities of pitches in each frame, and the MLM is a RNN that models the temporal relationships between pitches in different frames. They dedicate significant time to discussing the benefits of DNNs, RNNs, and CNNs for

auditory tasks (i.e. as the acoustic model). They further put forward three possible candidates for the Music Language Model—a generative RNN, Neural Autoregressive Distribution Estimator (NADE), and a hybrid RNN-NADE approach.

Their proposed model uses a CNN for the AM, and an RNN MLM, which outperformed two methods recent to their work, and they suggest future improvements including enhancing the robustness of their models by augmenting the training data - artificially adding noise to help it to cope in the face of suboptimal conditions. Though their method performs well, it appeared to struggle in particular with faster-played (i.e. shorter) notes, but on the whole, approximated the general shape of the example piece that they presented rather well.

3.4 State of the Art MIR

Thomé and Ahlbäck [161] presented a method that took raw audio data, performed a variation of the constant-Q transform known as SliCQ [81] on it, and passed the transformed data to a DNN that was trained to convert spectrograms to piano roll notation. They claimed to use a series of convolutional and padding layers to remove timbral and inharmonic aspects of the signal before then converting to the note-level notation, but though the model appears quite effective, not enough information about its architecture was presented to allow for faithful reproduction.

As with [153], [146] opts for a dual Acoustic Model/Music Language Model approach, but uses Probabilistic Latent Component Analysis (PLCA) for the AM, and a Hidden Markov Model (HMM) for the MLM. On the whole, the model struggles with harmonic errors in particular—a common issue in MPE—as well as mistaking the variations in pitch that result from use of vibrato as distinct pitches, which is quite a significant error when dealing with singing voices in particular.

Pushing off of a similar basis to YIN [39], another method using the Square Difference Function (SDF) is presented in [98]. They instead used Fourier approximation followed by a calculation of the Minimum Squared Error (MSE), and used global minima to estimate the fundamental period. From there, they considered the strengths of the first sixteen harmonics to choose the most likely candidate. In particular, this more algorithmic approach is quite refreshing amongst so many state of the art Deep Learning (DL) methods, and is quite intuitively linked to underlying properties of musical signals. The evaluation, however, is somewhat lacking, so it is hard to determine how effective the proposed algorithm really is.

Nakamura et al. [124] present a step towards end-to-end polyphonic music transcription. First applying the MPE approach in [13] that utilises spectrogram factorisation, they pass the results to a note-tracking algorithm, leading to the usual "piano roll" (i.e. note-level) data. Though they achieved decent results, more sophisticated MPE methods would further improve the system. As quite well-put by Moorer [118], the "computer" ends up being quite literal-minded, particularly with respect to quantisation and when dealing with musical techniques such as vibrato, which lead to ambiguity.

In an attempt to standardise testing of approaches to automatic polyphonic music transcription, McLeod and Steedman [110] developed and proposed a metric (MV2H), aimed specifically for use on note-based output (as opposed to frame-based), but direct readers to [135] for a frame-based metric.

Choi et al. [24] provide an insightful and well-explained tutorial of the uses of Deep Learning in the context of MIR, as well as a broad overview of relevant preprocessing techniques in particular. Figure 3.5 shows the relationships between NN layers and MIR - in this case, a spectrogram.

Building on the historic use of CNNs for AMT, Cong et al. [31] present a CNN-based approach that produces note-level (i.e. piano roll) output. Based on three models - a

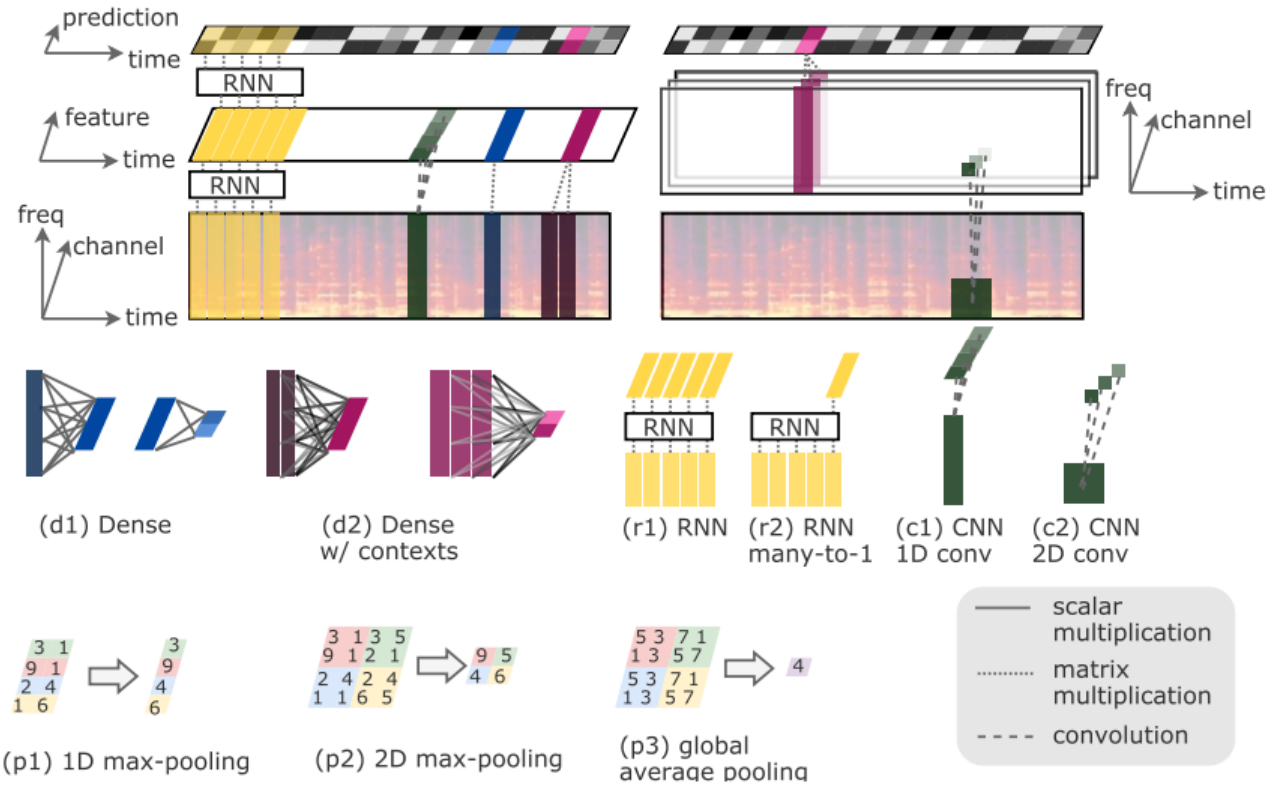


Figure 3.5: Neural Network layers in the context of MIR. Taken from [24].

single-channel CNN "pitch detection model" that performs MPE, a dual-channel "onset/offset model" that determines temporal onset and offset of notes, and finally a rule-based "note-search model" that combines the outputs of the prior two models, their approach improves in particular on the results of [153]. As with many methods, however, they constrain themselves solely to piano music (likely due to the prevalence of the MAPS dataset), which renders the problem somewhat easier - though certainly does not detract from the improvements.

Developing the work in my Bachelor's dissertation, [67] proposes a computationally simple frequency-domain based approach (having applied a STFT to the signal) that produces frame-level results. It works on the assumption that no undertones (i.e. harmonics at frequencies *lower* than f_0) could be present in the signal, and therefore the leftmost peak above some threshold, α, must be a fundamental. From this, the approach 'rakes' left to right through the spectra, iteratively selecting a fundamentals, and using a point to point spline to estimate the amounts of each harmonic generated by the fundamental, subsequently subtracting the estimated fundamental and its harmonics from the spectrum. Though seemingly quite effective, the testing in the paper is certainly somewhat lacking in that it is not based off of any of the large datasets available. Further, though the approach has been shown to work on a variety of instruments, it is not clear that the current approach (requiring successive harmonics to have monotonically decreasing amplitudes) would extend to instruments such as the trumpet or the clarinet - which do not necessarily adhere to that pattern.

Taking a more mathematically-minded tack, Alvarado and Stowell [3] treat signals as linear combinations of sources, which themselves consist of a quasi-periodic component (as touched on by [98]) and an amplitude envelope. They introduced a "Matérn Spectral Mixture" kernel, which they used to model the Attack, Decay, Sustain, Release (ADSR) envelopes of instruments encountered - a key strength of which being the ability to introduce and utilise prior knowledge of the properties of the signal. As with a number of other approaches, the testing presented is quite limited, but exhibited good results on what they

did show.

Also focusing on ADSR envelopes, [88] utilises Deep Learning in tandem with a "hand-crafted" (i.e. with manually-chosen transition probabilities) HMM that models the ADSR envelopes of the input sound. Contrary to other similar approaches [48, 31, 74], they opted for a more compact architecture for their NN, instead using a model which branches off after some shared layers in order to compute (start, end, note) tuples, and achieved state of the art results on the MAPS dataset. The especially unique aspect of this work (i.e. the hand-crafted HMM) is simultaneously an upside and a downside, as it requires a HMM to be assigned transition probabilities for each instrument to be used. Further, it seems likely that this could extend even to variations between instruments (e.g. different pianos), which would impact the efficacy of the method.

In 2019, Hansen, Jensen, and Christensen expanded on their previous work [71] to perform MPE on signals, using directional information (i.e. panned left or right) to simplify the problem [70]. On the whole, they achieved improved results over methods that did not utilise panning data - on both the IOWA and Bach10 datasets. The obvious limitation of this approach is that not all recordings will have stereo channels, but this likely does not present much of an issue on the whole.

Again building on the ACF, [127] propose an extension that removes "unwanted jumps" (i.e. musically unlikely steps) in the time-frequency representation of a signal, and then utilises pitch-tracking to smooth the outputs. This resulted in good results on the Bach10 dataset, as well as speech signals from PTDB-TUG [133]. Importantly however, this approach is restricted solely to Monophonic PE, which is a significantly easier problem than MPE.

Another approach focusing on an 'end-to-end' AMT model, [141] presents a model that aims to demonstrate that the task can be "performed in a single step". They use a

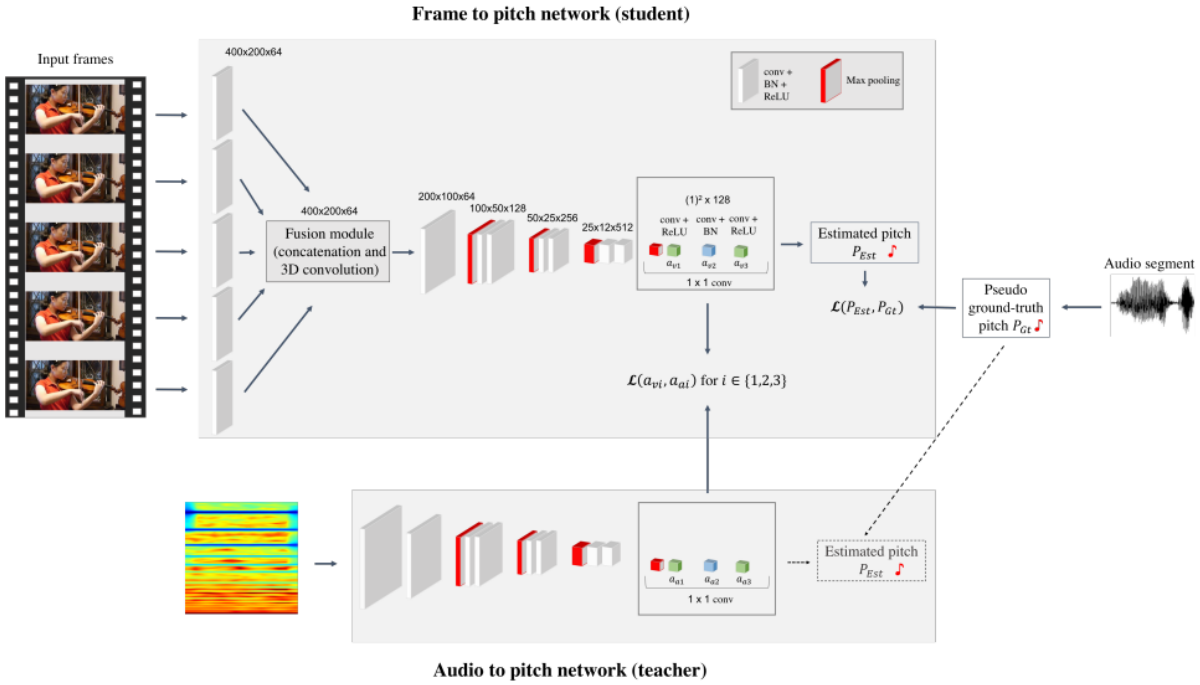


Figure 3.6: An overview of the NN architecture described in [94]. Taken from [94].

Convolutional Recurrent Neural Network (CRNN) with a Conectionist Temporal Classification (CTC) loss function to achieve this. Their claim that the problem can be reduced to a single step seems somewhat disingenuous, however, as there are really many layers to the proposed architecture, which it seems likely may well encapsulate the "intermediate goals" that they describe. Further, from a research perspective, it is not entirely clear what the benefit of such a demonstration would be - not least because the 'stepwise refinement' of a problem results in independent or modular components of a larger system (in this case, AMT), which can be individually developed, pieced together, and swapped out to facilitate improvements to the overarching AMT system. That said, the results garnered from the proposed method—at least that which is showcased in the paper—is nothing short of impressive, and if nothing else, demonstrates the incredible utility of ML methods from an engineering or implementation perspective.

Looking from a completely different perspective, [94] attempts to approach PE using

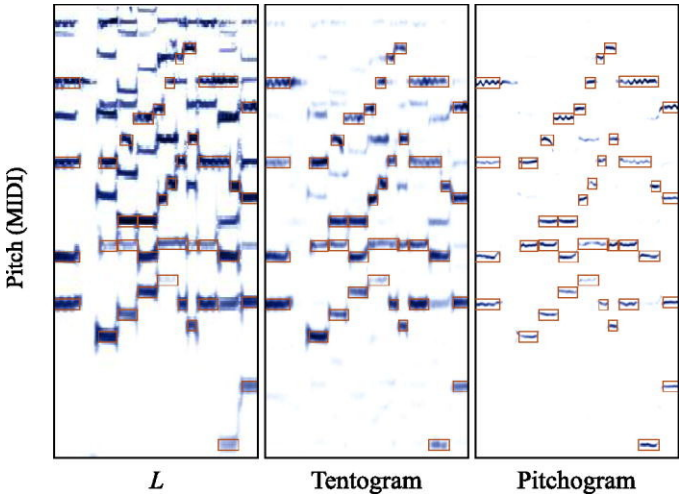


Figure 3.7: Demonstration of the progression from Spectrogram to Tentogram, and then to Pitchogram. Taken from [48].

visual (i.e. video) data as opposed to audio. Specifically, they built a model to estimate monophonic pitch from video recordings of violinists. They further posit that this could be of use in (hybrid) recordings with poor audio quality, or even to aid with polyphonic PE. In order to achieve this, they exploited two tricks - a teacher-student training strategy [74] (Figure 3.6), to teach the model to associate visual cues with corresponding sounds, as well as using multiple frames of video as opposed to a still frame, which helped them to resolve ambiguities that arose from troubles distinguishing which string was vibrating, amongst other things.

Considering the incredibly challenging task on which they embarked, the results achieved were surprisingly good, with the added benefit that the method generalises well, as it relies on "automated" annotation (i.e. through the use of a preexisting mono-pitch estimator).

Elowsson [48] proposed a method for MPE that relies on 'deep layered learning' [49, 47]. It uses a multi-stage system of neural networks and processing steps to elicit pitch contours - i.e. pitch information coupled with an onset and offset for each distinct note. From the MPE side, they opted to create a 'Tentogram' (i.e. a tentative spectrogram)

through a spectral summation (their Section IV-F), whitening, and logistic regression, which provided a much cleaner basis from which to detect pitch peaks. A neural network was then used to convert the Tentogram into a 'Pitchogram', using parabolic interpolation to achieve a 1 cent resolution. From the Pitchogram, 'blobs' are then identified [117], with subsequent regions then merged (where related), and finally a peak ridge (1D contour) was extracted. Figure 3.7 provides a useful visual representation of the approach. This method exhibited state-of-the-art performance on the MAPS [52], Bach10 [45], TRIOS, and MIREX Woodwind quintet datasets.

Taking a different approach to much of the recent literature, Zhang, Chen, and Yin [183] focus on eliciting patterns from a "pseudo-2d spectrum"—with both axes representing frequency—as opposed to solely the frequency domain, or the typical hybrid time-frequency techniques. They show that this results in distinctive 2D shapes corresponding to the fundamentals, and apply pattern-matching techniques framewise to the signal to elicit pitch candidates, further filtering out spurious fundamentals by considering neighbouring frames and pitch contour lengths. This approach achieved improvements over the use of 1D spectra on both the Bach10 and MAPS datasets. They also present their results in comparison to other techniques, in which this method performs somewhat on-par with a range of state of the art parallels. Interestingly, the notion of eliciting spatially-compact 2D representations of fundamentals has a number of parallels to the research presented in this Thesis.

3.5 Other Related Work

In 2012, Wang and Walker presented a short review of studies into the auditory processing of pitch information in the cortices of mammals—looking in particular at macaques, marmosets, cats, and ferrets [174]. The cortices of the mammals studied all comprised of similar structures - with a core region (A1/R/RT) and "belt"/"para-belt" regions surround-

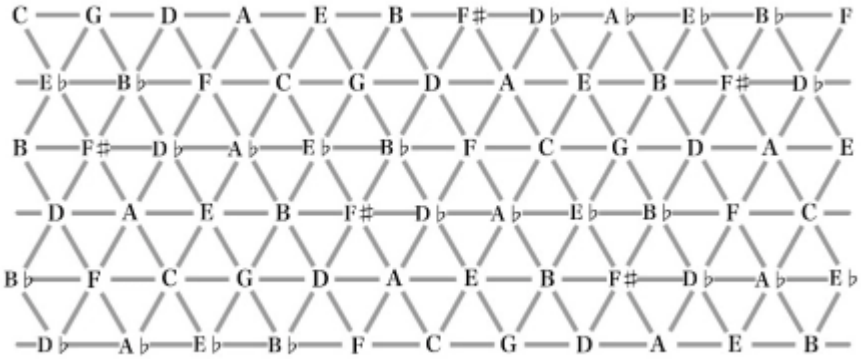


Figure 3.8: A Tonnetz without identified edges - clearly showing its periodic nature in both the vertical and horizontal directions. Taken from [15].

ing it. Notably, it has been found that macaques and cats [162, 75] respectively are able to determine pitch (i.e. fundamental frequency), even when harmonics are missing. Further, Bendor and Wang [12] demonstrated the existence of individual "pitch-selective neurons" in marmosets, which were able to perform pitch extraction. They proceeded to show that these neurons were sensitive in particular to the "temporal regularity" (i.e. fundamental period) of sound (unlike neighbouring zones of the brain) [11]. Wang and Walker state that

> "Given that humans and animals do experience a percept of pitch that generalises across a variety of sounds with the same periodicity (including missing fundamental sounds), it seems reasonable to expect to find neurons at some level of the auditory system that integrate the periodicity cues... to compute a stimulus-invariant pitch representation."

This suggests that there may indeed be some mathematical model of pitch that is invariant of the timbre of the sound being considered. This provides further motivation for attempts to elicit such patterns from acoustic music—such as in this Thesis.

First described by Euler in 1739, the Tonnetz (Figure 3.8) [54] presents a way to spatially demonstrate the relationship between chords. Each row corresponds to the circle

of fifths, with each subsequent row corresponding to the previous one with each element shifted from position n, to position $(n + 3)$ mod 12, and aligned such that the closest two notes diagonally 'upwards' correspond to a major third in one direction, and a minor third in the other. By identifying both vertically and horizontally, it is clear that the Tonnetz is in fact toroidal in nature. It has proven particularly useful in describing voice leadings in music - with distance between triangles (i.e. chords) on the Tonnetz corresponding to musical 'distance' between chords.

In addition to his work on the generalised Tonnetz [164], Tymockzo posits a geometrical treatment of music theory in his book 'A Geometry of Music' [163] - driven by underlying musicological knowledge. Where Lewin tackled this kind of formalisation from a group-theoretical perspective [102], Tymockzo (whilst still basing his approach on symmetries) opted to employ a more geometric approach; considering pitch/chord spaces in such a way that proves useful to composers and musicologists alike, without a necessarily profound understanding of the underpinning mathematics.

Tymockzo defines musical objects, which are essentially ordered collections of notes (Eg. (C4, E4, G4)), and five "OPTIC" operations -

- **O**ctave - transposing individual notes by an octave;
- **P**ermutation - reordering the object (changing which voice has which note);
- **T**ransposition - uniformly shifting all notes in an object by a given offset (and direction);
- **I**nversion - essentially reflection about a point in pitch space (i.e. pitches ordered chromatically along a 1D line);
- **C**ardinality change - introducing a new voice that duplicates a note that is already present in the object.

These describe transformations between musical objects. Further, he goes on to define a variety of musical constructs (such as chords and scales) in terms of the set of OPTIC transformations under which each construct remains invariant.

Building on this framework, he defines a two-note chord space, containing progressions between dyads (Eg. (C4, E4)→(C4, E♭4)). By enumerating the whole space, and identifying the edges with a twist (which is necessary as, when enumerated fully, the vertical edges of the two-note chord space are the reverse of one another), the two-note chord space forms a Möbius strip. The use of this space in analysis is then demonstrated practically by applying it to elicit musical insights on pieces (such as Brahms' Op. 116, No. 5), that would otherwise be obscure if viewed in traditional notation. He goes on to provide a generalisation of n-note chord spaces in higher dimensions such as a three-note chord space forming a twisted triangular prism.

Chapter Four

A Geometric Framework for Pitch Detection

MUCH of the core work in this chapter was submitted to the Journal of Mathematics and Music in November 2020, and was accepted for publication in September 2021. We anticipate that it will be published by the journal soon. Notably there was collaboration with Karoline van Gemst, a colleague and friend, without whose mathematical guidance, input, and participation in immeasurably long whiteboard sessions with copious volumes of coffee, the work presented would be far less rigorous, and Peter Tiňo, my fantastic supervisor, who has been nothing short of excellent in his role, and provided immeasurable academic counsel—and appropriate admonishment—over the years.

4.1 Considerations

This section lays out the context in which the model was developed, and aims to present a couple of ideas, alongside the assumptions under which the model is proposed. Though these may seem to be quite restrictive on first pass, they generalise relatively well

to acoustic musical signals in general, and some (such as the lack of undertones) build off of earlier work [67].

4.1.1 Timbre and 'Timbre-Invariance'

Humans have an extraordinary ability to subconsciously abstract the notion of timbre from a signal [173] - a skill that proves useful both in music, and in speech. It would be beyond inconvenient if every time one spoke to another person, it was necessary to jump through some kind of neural hoop along the lines of *'Ah! It's Pete from the shop - I'd better remove the timbre of his voice so I can understand what he's saying'*. Instead, this process occurs automatically, with no conscious thought or cognition about it. Indeed, the same is true of music - it takes little to no effort to pick out and hum the melody to a song, even when the voice is but one of a plethora of instruments playing simultaneously.

Given this skill, it is possible to define an algorithm for pitch detection that most, if not all, humans could consciously execute to a consistently near-perfect standard. It is worth noting that no claims are being made as to how the human brain subconsciously or implicitly performs this task, however. For each monophonic track in a mixture, iterate through each tone, and compare it to a reference point. If the tone is the same as the reference point, then output the total offset taken to reach it, and continue to the next tone in the track. Otherwise, ascertain whether it is higher or lower than the reference point, and move the tone one semitone in the direction of the reference point (keeping track of the total offset). Continue this process until the tone is the same as the reference point, and simply use the overall offset, along with the known reference point, to calculate the original tone (Algorithm 1).

Algorithm 1 Human PD Algorithm

Input: $\mathcal{M}_\tau = \{m_0, m_1, ..., m_n\}$
Output: $\mathcal{P}$

```
P ⟵ ∅
R ⟵ (A, 4)
for m_i ∈ M_τ do
  for tone ∈ m_i do
    offset⟵ 0
    while (tone ◇ offset) != R do
      if tone > R then
        offset ⟵ offset + 1
        tone ⟵ tone ◇ 1
        continue
      end if
      offset ⟵ offset - 1
      tone ⟵ tone ◇ (-1)
    end while
    P ⟵ P ⋃ {tone}
  end for
end for
return P
```

where $\mathcal{M}_\tau$ is a mixture of signals at a given point in time, τ, $\mathcal{P}$ is the set of tones present at τ, and $\diamond$ is some operator that takes a tone, and applies an offset of n semitones.

From both the above example, and the work of Bendor and Wong, it seems likely that there exists some timbre-invariant model of pitch. That is, one that can elicit the fundamental frequency of a signal not just irrespective of the instrument being perceived, but further, irrespective of whether the "system" (whether human or computer) has *ever* perceived such an instrument before. Broadly, this Thesis seeks to define one such model—again, with no connotations that the human brain works in a similar way—such that it may be built upon in the future, and used to develop PE techniques from a wholly different perspective to the current literature.

4.1.2 Considerations for the Model

In building such a model, we took a variety of considerations into account. In particular,

- First and foremost, it was crucial for it to be "easy" to identify fundamentals (or as will be observed later, harmonics masquerading as fundamentals). This required us to identify a property that is true of all fundamentals (though not solely true of fundamentals - i.e. not a bijection) that we could use to this end.
- The model needed to be extensible - by starting with an oversimplified model, with relatively restrictive assumptions, it is then possible to iteratively build upon it, adding complexity to it in the process.
- In addition, it was imperative that it has some grounding in music and music theory. As identified in [122], this insight of the nature of music was paramount to developing successful techniques for music audio signal processing.
- Finally where possible, the model had to be applicable to all harmonic instruments.

4.1.3 Assumptions

In order to help facilitate the goals, a number of assumptions were made:

1. All fundamentals are present with at least their first three harmonics, i.e. given f_0 is present in the signal, so too f_1, f_2, and f_3 will be. Further, only these will be added to the model.
2. No fundamentals below (C, 0) $\approx$ 16.35Hz or above (B, 9) $\approx$ 15804.27Hz will be present in the signal.

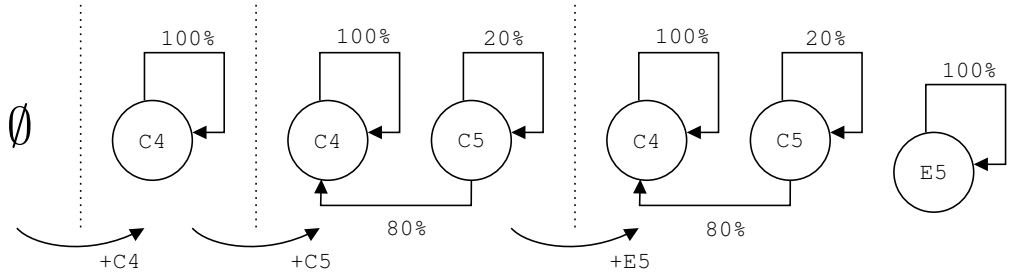


Figure 4.1: The build-up of a simplistic probabilistic representation.

3. All instruments will be harmonic, that is, inharmonic instruments such as drums, bells, etc., with harmonics that do not fall at intervals corresponding to the harmonic series are excluded. Importantly here, the related inharmonicity phenomenon, which regularly occurs in the upper harmonics of fundamentals from harmonic instruments is mitigated by (1).

4. All instruments are acoustic.

5. No undertones are generated by the instruments.

In addition to these assumptions, it would likely be necessary to impose further restrictions on the signal, specific to the application in which the model is being used.

4.2 Reaching a Model

Consider a frequency-sorted (low to high) set of tones[1], for example, {(C, 4), (C, 5), (E, 5)}. One can imagine wanting to build up some representation of where each tone is likely to have originated. Figure 4.1 shows the iterative construction of one such model - which is somewhat adjacent conceptually to Markov chains. This could instead be viewed as a directed graph, with an edge from each vertex to every other vertex that it could potentially

[1]Note the use of 'tone' here as opposed to 'note' – the key distinction being that a tone describes an 'object', whereas a note is a specific instance of a tone.

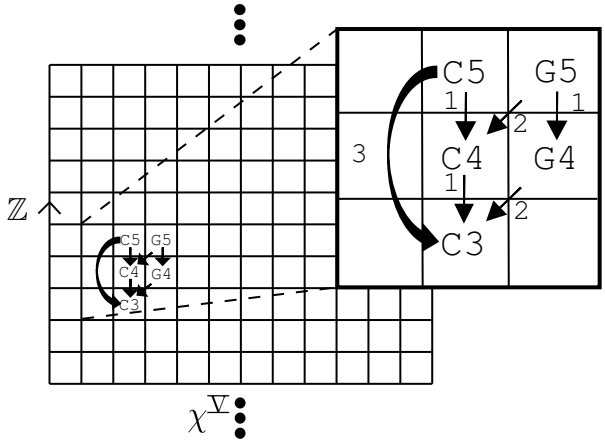


Figure 4.2: Visualisation of the graphical structure overlaid onto the grid.

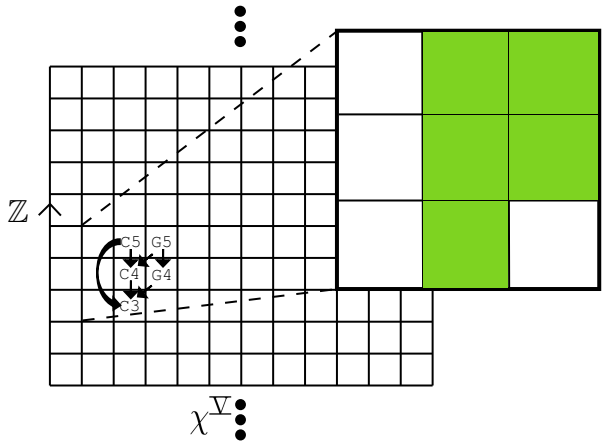


Figure 4.3: Visualisation of the final grid structure, where shaded cells represent a Boolean value of true.

be a harmonic of. Weights are chosen from some vertex, B, to another vertex, A (of which B is potentially a harmonic) to be i, such that B is the i^{th} harmonic of A. Here, each weight corresponds to some measure of the likelihood that a given tone is generated by another. For example, given the presence of C4, the probability that C5 is a fundamental (i.e. generated by itself) may be 20%, whereas there may be an 80% probability that it is instead a harmonic of C4. These are, of course, toy values, and it is more than likely more realistic to readjust all weights following the addition of each tone, but the underlying concept remains the same.

By placing each vertex into a grid (with the horizontal axis representing the pitch

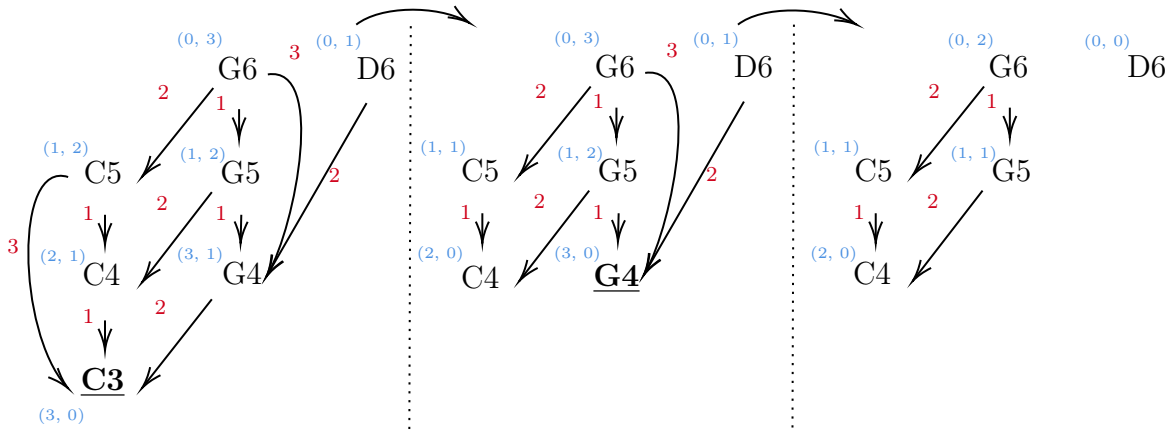


Figure 4.4: (*left*) A directed graph depicting C3 and G4 (and each of their first three harmonics sounding). The degrees of each vertex are shown in parentheses. The following steps represent the steps of the simple algorithm described. The bolded/underlined tone at each step is the one selected as a fundamental.

chromas ordered according to the circle of fifths, and the vertical axis representing the pitch height in octaves), and restricting the edges to only the first three harmonics, it is possible to remove the edges – leaving this information implicit in the new representation – and derive a simpler model, with the added benefit that a large number of vertices and edges do not render the representation visually messy or hard to decipher (Figure 4.2). From here, it becomes clear that it is in fact possible to dispose of the notion of this representation of a graph altogether: by removing the edges (the information from which becomes implicit), and instead assigning a Boolean value to each cell, representing whether the tone is audible or not (Figure 4.3).

The problem of pitch detection is then reduced to the problem of finding the decomposition of the grid (into shapes) that corresponds to the tones played in the input signal. As becomes apparent, this involves discarding a number of false positive cases from the interpretation. From a graphical perspective, this is equivalent to identifying the vertices that correspond to fundamentals, removing them, and repeating the process until no more are present. Note that such methods are impacted by the presence of noise in the signal - the reduction of which is beyond the scope of this paper.

For example, consider the graph in which (C, 3) and (G, 4) are sounding along with their first three harmonics (Figure 4.4). By annotating each vertex, ν, with its indegree ($\deg^-(\nu)$) and outdegree ($\deg^+(\nu)$), some vertices present as sinks (i.e. with degree $(n^-, 0)$ for some $n^- > 0$), and some as sources (i.e. with degree $(0, n^+)$ for some $n^+ > 0$). For the purposes of pitch estimation, and because of the chosen edge direction (from harmonic to fundamental), a sink with indegree three is always a fundamental with its first three harmonics present. Thus, a simple (yet somewhat effective) algorithm is to take each vertex with indegree three (for each distinct part of the graph, as it may not be connected), categorise them as fundamentals, and remove all categorised vertices. This can then be iteratively applied to the graph until no sinks with indegree three remain (as shown in Figure 4.4). Clearly this algorithm is a vast oversimplification of the problem, but it nicely illustrates the benefits of geometric approaches.

The following sections will build upon this grid-based model, proving some useful properties about it, and presenting some ways in which future algorithms may utilise this characterisation.

4.3 The Proposed Model

As in Section 4.2, let $\chi^{\mathbb{V}}$ be the set of pitch chromas, ordered by fifths - $\{C, G, D, A, ..., F\}$[2]. Let the grid, $\mathcal{N}^{\mathbb{V}} \coloneqq \chi^{\mathbb{V}} \times \mathbb{Z}$. That is, an element $\nu_{i,j} \in \mathcal{N}^{\mathbb{V}}$ is a pair (χ_i, j) representing a tone with pitch chroma χ_i, and pitch height j. $\mathcal{N}^{\mathbb{V}}$ forms the backbone of the model. As the circle of fifths exhibits a periodic nature, the left and right edges of the grid may be identified, or *glued*, to realise $\mathcal{N}^{\mathbb{V}}$ as a discretised infinite cylinder (Figure 4.5). Furthermore, Let $\mathcal{N}^{\mathbb{V}}_{\alpha}$ be the finite subset of $\mathcal{N}^{\mathbb{V}}$, consisting of the 'sub-cylinder' with octaves $[0, 9]$. This represents

[2]Note here that the exponent, $\mathbb{V}$, is a label representing that the set is ordered by fifths.

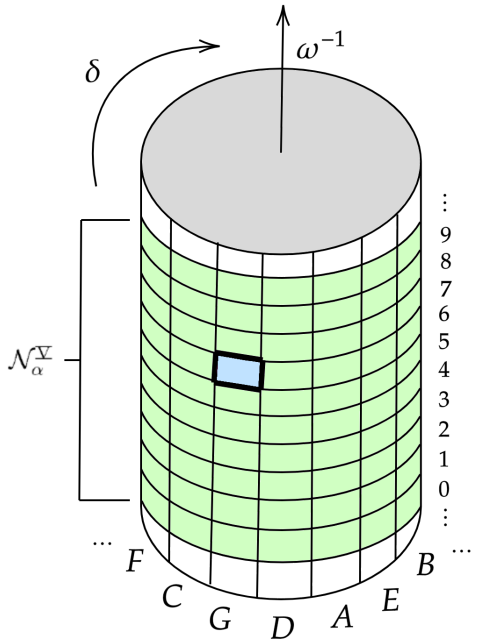


Figure 4.5: The discretised infinite cylinder of $\mathcal{N}^{\mathbb{V}}$ with $\mathcal{N}_\alpha^{\mathbb{V}}$ indicated. Relative to a *fixed* viewpoint (outlined black, filled blue), δ corresponds to a clockwise rotation of the cylinder by one cell, and ω corresponds to a vertical shift of the cylinder one cell downwards.

the human-audible spectrum of sound. Further, define the predicate, $\mathcal{I}_\tau \colon \mathcal{N}^{\mathbb{V}} \to \mathbb{B}$,

$$\mathcal{I}_\tau(\nu_{i,j}) = \begin{cases} \top & \text{if } \nu_{i,j} \text{ is observed at } \tau \\ \bot & \text{if } \nu_{i,j} \text{ is not observed at } \tau. \end{cases} \tag{4.1}$$

This is called an interpretation (i.e. of $\mathcal{N}^{\mathbb{V}}$), and can be seen as a single time slice of a signal, indexed by an instantaneous point in time, τ. By viewing a musical signal as a sequence of temporal slices, we obtain an interpretation of the signal for any given τ by the pair $(\mathcal{N}^{\mathbb{V}}, \mathcal{I}_\tau)$ (Figure 4.6).

By viewing the ordered collection of pairs as a whole, one can uniformly stretch each slice, and identify the appropriate $\mathcal{N}^{\mathbb{V}}$ faces to create a three-dimensional heatmap, with each tone now represented by a cube as opposed to a square[3]. Thus, the interpretations are now indexed by an interval, where previously they had been indexed by an instantaneous

[3]The importance of this becomes apparent in Section 5.2, in particular when considering the 3D heatmap.

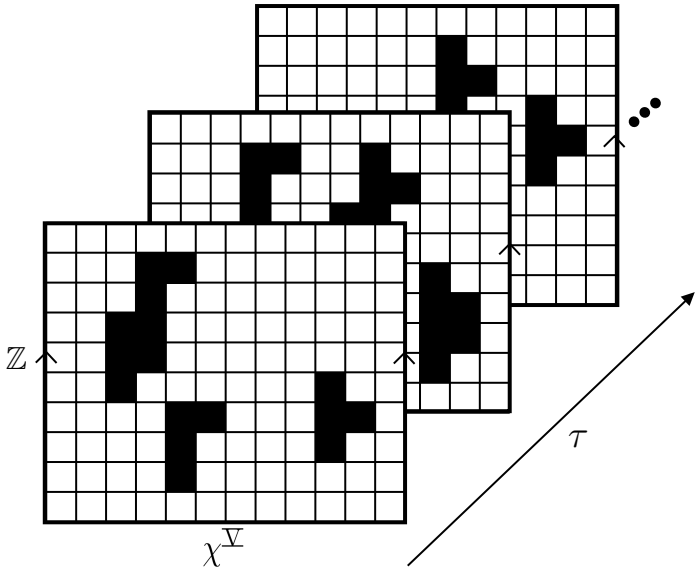


Figure 4.6: Visualisation of a sequence of interpretations (representing temporal slices), indexed by τ.

point in time, that is,

$$(\mathcal{N}^{\mathbb{V}}, \mathcal{I}_\tau) \longmapsto (\mathcal{N}^{\mathbb{V}}, \mathcal{I}_{[\tau,\tau+1)}). \tag{4.2}$$

A real-world example of this is shown later in Figure 5.24.

In general, when referring to *any* interpretation henceforth, $\mathcal{I}_\tau$ may be replaced by $\mathcal{I}$ for simplicity.

By projecting onto hyperplanes parallel to the faces of the cuboid, one can consider the signal from different perspectives - that is, with constant time, constant pitch chroma, or with constant pitch height. For example, considering the projection with constant pitch height, one can elicit a pitch contour representation of the signal. Further, by viewing the heatmap as a translucent construction, it is possible to consider all aspects simultaneously, as if the construction was opaque, only the cubes on each face of the cuboid would be visible.

Let f_i denote the i^{th} harmonic, with f_0 being the corresponding fundamental. Depend-

ing on the pitch chroma, the 2nd harmonic, f_2, may or may not cross the octave boundary (i.e. be in the next octave up from the 1st harmonic). For example, the first three harmonics of $(C\sharp, n)$ are $f_1 : (C\sharp, n+1)$, $f_2 : (G\sharp, n+1)$, and $f_3 : (C\sharp, n+2)$, whereas the first three harmonics of (A, n) are $f_1 : (A, n+1)$, $f_2 : (E, n+2)$, and $f_3 : (A, n+2)$. By considering each chroma in $\chi^{\mathbb{V}}$, it is clear that the presence of $\{f_0, f_1, f_2, f_3\} \subset \mathcal{N}^{\mathbb{V}}$ make up one of two shapes; a turnstile shape, $\vdash$, or a gamma shape, Γ, depending on the position of the fundamental. In particular, when f_2 lies across the octave boundary, then the fundamental and its harmonics present as a Γ shape, and otherwise present as a $\vdash$. Denote the set

$$\chi_{\vdash} = \{C, C\sharp, D, E\flat, E\}, \quad \text{with} \quad \chi_{\Gamma} = \chi^{\mathbb{V}} \setminus \chi_{\vdash} \quad \text{as its complement,} \ [4]$$

and let π_χ, π_y be the projection of $\mathcal{N}^{\mathbb{V}}$ onto the horizontal and vertical axes respectively. Then, when $\pi_\chi(f_0) \in \chi_{\vdash}$ one observes the $\vdash$ shape, and Γ otherwise.

This is shown on Figure 4.7, where a fundamental is denoted by $\bullet$ and its harmonics by $\times$. This notation will be used liberally through the remainder of this Thesis, so be sure to commit it to memory!

This model (particularly the use of $\chi^{\mathbb{V}}$ as opposed to a chromatically-ordered column set) is deliberately chosen such that the pattern exhibited by a fundamental and its first three harmonics (i.e. $\vdash$ /Γ) appears spatially-compact. This serves to make these patterns more easily discernible – both to the human observer, and computationally – and is of particular use when looking to decompose more complicated polyphonic signals into their constituent parts.

The different cells, or tones, on the cylinder can be related to each other by considering a group action on $\mathcal{N}^{\mathbb{V}}$. Let δ and ω denote the generators of $\mathbb{Z}_{12}$ (the integers modulo 12) and $\mathbb{Z}$, respectively. Then define a group action $\mathbb{Z}_{12} \times \mathbb{Z} \circlearrowleft \mathcal{N}^{\mathbb{V}}$ as follows. δ and ω induce

[4] Recall that chromatically, (C, 1) directly follows (B, 0).

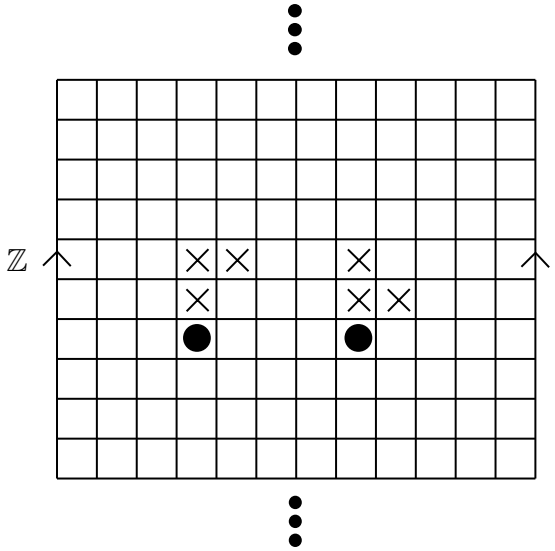


Figure 4.7: Demonstration of the Γ and $\vdash$ shapes in $\mathcal{N}^{\mathbb{V}}$.

maps on $\mathcal{N}^{\mathbb{V}}$ by

$$(\delta, \mathbb{1}_{\mathbb{Z}}) : \mathcal{N}^{\mathbb{V}} \to \mathcal{N}^{\mathbb{V}} \qquad (\mathbb{1}_{\mathbb{Z}_{12}}, \omega) : \mathcal{N}^{\mathbb{V}} \to \mathcal{N}^{\mathbb{V}}$$
$$\nu_{i,j} \mapsto \nu_{i+1,j}, \qquad \nu_{i,j} \mapsto \nu_{i,j+1},$$

where $\mathbb{1}_{\mathbb{Z}}$ and $\mathbb{1}_{\mathbb{Z}_{12}}$ are the identity elements in $\mathbb{Z}$ and $\mathbb{Z}_{12}$, respectively. In other words, $(\delta, \mathbb{1}_{\mathbb{Z}})$ acts on the cylinder by rotating it clockwise by one cell, while applying $(\mathbb{1}_{\mathbb{Z}_{12}}, \omega)$ corresponds to a vertical shift of one cell downwards (It may be worth revisiting Figure 4.5 to cement this intuitively). Hence a map

$$\mathbb{Z}_{12} \times \mathbb{Z} \times \mathcal{N}^{\mathbb{V}} \to \mathcal{N}^{\mathbb{V}} : \ (\delta^k, \omega^l, \nu_{i,j}) \mapsto \nu_{i+k,j+l}, \tag{4.3}$$

is achieved, where $k, l \in \mathbb{Z}$, and δ^{12n} for any integer n is the identity. For notational simplicity $(\delta, \mathbb{1}_{\mathbb{Z}})$ is identified with δ, and similarly for ω. Note that this means that, relative to a reference point, δ translates the tone by a fifth, and ω moves the tone up an octave.

In terms of this action,

$$\omega(f_0) = f_1, \qquad \omega\delta(f_0) = f_2, \qquad \text{and} \qquad \omega^2(f_0) = f_3,$$

for $\pi_\chi(f_0) \in \chi_\vdash$, where $\omega \circ \delta$ is identified with $\omega\delta$. For $\pi_\chi(f_0) \in \chi_\Gamma$ the above holds with the exception of f_2 which in this case is given by

$$\omega^2\delta(f_0) = f_2.$$

Furthermore, using this action, the $\vdash$ and Γ shapes may be written as,

$$\vdash = \{\mathbb{1}, \omega, \omega\delta, \omega^2\}, \qquad \Gamma = \{\mathbb{1}, \omega, \omega^2\delta, \omega^2\}, \tag{4.4}$$

where it is understood that by applying all elements of $\vdash$ to a tone traces out the turnstile shape, and similarly for Γ. In other words, considering the $\vdash$ case, for each fundamental which is mapped to $\top$ by $\mathcal{I}$, there exist harmonics $\omega(f_0)$, $\omega\delta(f_0)$, and $\omega^2(f_0)$ such that each of these are also mapped to $\top$ by $\mathcal{I}$,

$$\forall_{\nu \in \mathcal{N}^{\mathbb{T}}}\big[(\mathcal{F}(\nu) \wedge \pi_\chi(\nu) \in \chi_\vdash) \rightarrow (\mathcal{I}(\omega(\nu)) \wedge \mathcal{I}(\omega\delta(\nu)) \wedge \mathcal{I}(\omega^2(\nu)))\big], \tag{4.5}$$

given that f_1, f_2, and f_3 are observed (audible). Here $\mathcal{F}(\nu)$ is a predicate that returns $\top$ iff ν is a fundamental. Of course, such a construct does not exist in practice, but in essence, the end result of a perfect pitch estimation algorithm is this function, such that it best describes the ground truth of the signal. As before, the Γ case is equivalent under the replacement $\omega\delta \mapsto \omega^2\delta$.

By observation of the corresponding $\vdash$ and Γ shapes over the circle of fifths, it is noted that three two-shape configurations exist - namely $\Gamma\Gamma$, $\Gamma\vdash$, and $\vdash\Gamma$ (Figure 4.8). These are

$\vdash$	Γ	$\vdash$	Γ	$\vdash$	Γ	Γ	$\vdash$	Γ	$\vdash$	Γ	Γ
C	G	D	A	E	B	$F\sharp$	$C\sharp$	$A\flat$	$E\flat$	$B\flat$	F

Figure 4.8: The three two-column configurations, depicted on the circle of fifths. In particular, note the absence of a $\vdash\vdash$ configuration.

used to categorise a number of properties of the model.

Definition 4.3.0.1 (Configuration)**.** *A configuration (denoted as* $\Gamma\Gamma$*,* $\Gamma\vdash$*, or* $\vdash\Gamma$*) represents the shapes generated by fundamentals residing in two adjacent columns in* $\mathcal{N}^{\mathbb{V}}$*.*

Whilst every fundamental, together with its first three harmonics, exhibit one of the two aforementioned shapes (i.e Γ or $\vdash$)[5], the inverse statement is not true. Namely, the presence of a $\vdash$ or Γ shape does not imply that the tone concerned is a fundamental, as shown for the Γ case in Figure 4.9. A similar counterexample for $\vdash$-exhibiting tones can also be constructed. Here, $\otimes$ denotes a harmonic which presents as a fundamental. Such harmonics are called false fundamentals.

Remark. *When referring to a false fundamental,* $\otimes$*, the second column of the configuration always corresponds to that in which* $\otimes$ *lies. Thus, the first column corresponds to the preceding one - i.e.* $\pi_\chi(\delta^{-1}(\otimes))$*. For example, in Figure 4.9, the* $\otimes$ *sits in a* $\Gamma\Gamma$ *configuration and* not *in a* $\Gamma\vdash$ *one.*

Similarly to how a fundamental and its first three harmonics corresponds to either a $\vdash$ or a Γ, as a result of the assumption that any f_0 must present with its first three harmonics – and provided adequate noise removal from the signal – a given harmonic could only have arisen from a fundamental related to it by the *inverse* shape, i.e either $\dashv$ or $\lrcorner$.

[5] As a result of of assumption 1.

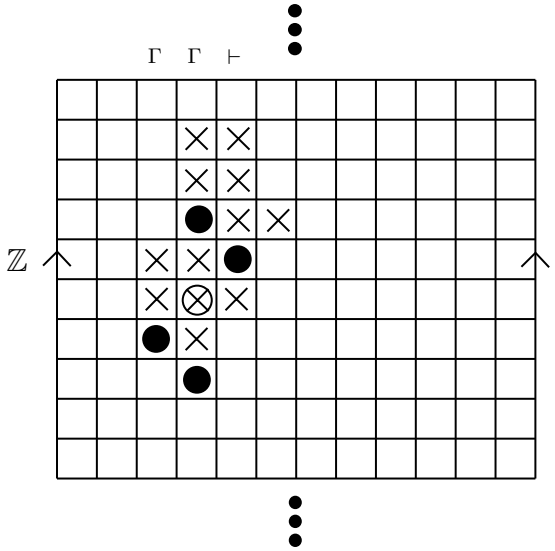


Figure 4.9: Counterexample showing that not all Γ shape-exhibiting tones are fundamentals.

The inverse shapes may be written as,

$$\dashv = \{\sigma^{-1} |\ \sigma \in \vdash\} = \{\mathbb{1}, \omega^{-1}, (\omega\delta)^{-1}, \omega^{-2}\}, \tag{4.6a}$$

$$\lrcorner = \{\sigma^{-1} |\ \sigma \in \Gamma\} = \{\mathbb{1}, \omega^{-1}, (\omega^2\delta)^{-1}, \omega^{-2}\}, \tag{4.6b}$$

where σ^{-1} is the inverse of σ with respect to the group structure.

Additionally, define the function $\Psi(\chi_i)$ for any $\chi_i \in \chi^{\mathbb{V}}$ as,

$$\Psi(\chi_i) = \begin{cases} \vdash & \text{if } \chi_i \in \chi_{\vdash} \\ \Gamma & \text{otherwise.} \end{cases} \tag{4.7}$$

Intuitively this function takes a given chroma (i.e. $\pi_\chi(\nu) \in \chi^{\mathbb{V}}$), and returns the set of group elements that trace out the corresponding shape when applied to a tone with this chroma.

The generator of a tone is defined as the fundamental that deposited the corresponding

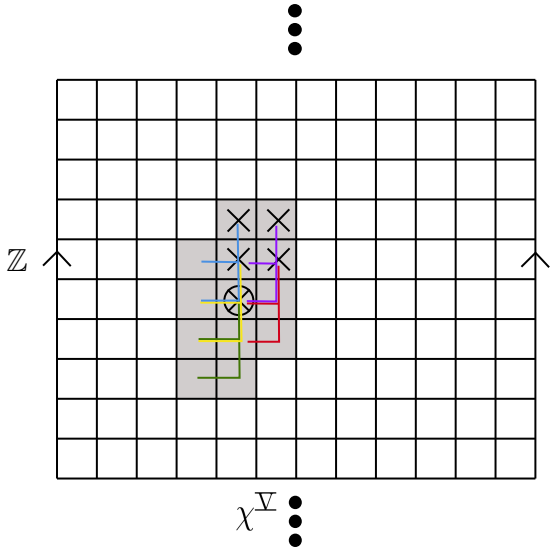


Figure 4.10: Figure showing where a false fundamental and each of its apparent harmonics could have been generated from, with colours tracing out $\dashv$ and $\lrcorner$ for each tone (overlaying both Γ and $\vdash$ situations).

frequency. Note that the generators of a tone may sit both in the same, or proceeding column to itself. This means then when enumerating the possible generators in most cases (i.e. not $\Gamma\Gamma$), it is necessary to consider tones in $\dashv \bigcup \lrcorner$. In many cases there may be multiple generators for a single tone.

Suppose an interpretation is given such that a false fundamental is present. By investigating the possible positions for this tone in $\mathcal{N}^{\mathbb{V}}$, there is a finite region containing the fundamentals that could have created the false fundamental and its first three apparent harmonics. In any of the possible positions for the false fundamental, this region is contained in the region given by $\hexagon(\otimes)$, with $\hexagon \coloneqq \{\sigma'\sigma | \sigma' \in \dashv \bigcup \lrcorner, \sigma \in \vdash \bigcup \Gamma\}$, as shown in Figure 4.10. This holds by construction[6].

Consequently, in order to check whether a tone could be generated by some other tone, it is sufficient to search a 3×5 area centred on the tone. Though this proves sufficient from the perspective of generators, there are a number of cases (i.e. false fundamentals) that

[6]Note that the shape traced out by $\hexagon$ is really the union of three shapes - each obtained from tracing backwards from a false fundamental (and its harmonics) in one of the three configurations.

will naively result in false positives.

Hence, the problem of multi-pitch estimation is reduced to that of distinguishing between fundamentals (•) and harmonics masquerading as fundamentals (⊗). The following sections focus on characterisation of the occurrences of such false fundamentals, seeking to pave a way to suitably distinguish between them and fundamentals.

4.4 Fantastic Edge Cases (and Where to Find Them)

This section deals only with cases in which a single false fundamental (⊗) is considered simultaneously. Note that this does not mean that only one ⊗ is present in each Edge Case, but rather that a sequential traversal is used in order to consider ⊗s one by one. In reality, this is expected to be enough of a generalisation as long as algorithms consider false fundamentals sequentially, such that $\mathcal{N}_\alpha^{\nabla}$ is traversed along

$$\delta^i \omega^j(\nu_{0,0}),\ \forall_{i\in\{0,\cdots,11\}},\ \forall_{j\in\{0,\cdots,9\}}, \tag{4.8}$$

that is, left-to-right, bottom-to-top, where $\nu_{0,0}$ is the bottom-leftmost element of $\mathcal{N}_\alpha^{\nabla}$. It is believed that this traversal will be sufficient to provide information of previously classified tones to aid in the classification of ones further along in the traversal. This is adjacent to the left-to-right traversal (i.e. low to high) of the frequency domain in [67].

Definition 4.4.0.1 (Edge Case)**.** *An edge case is a set of fundamentals and their first three harmonics, in which a tone that presents as a fundamental, is in fact not one. In other words, let*

$$\otimes(\nu) \ = \ \big(\forall_{\sigma\in\Psi(\pi_\chi(\nu))}\big[\mathcal{I}(\sigma\nu)\big]\big) \wedge \neg\mathcal{F}(\nu),$$

be the predicate that returns $\top$ *iff* ν *is a false fundamental. Then a set of tones, S, is an*

edge case when

$$\exists_{\nu \in S}[\otimes(\nu)].$$

It is worth noting that although this Section looks only at cases in which one $\otimes$ is considered at a time, this definition is broad enough to encapsulate the more complex cases with multiple $\otimes$ too.

By considering $\dashv \bigcup \lrcorner$ for each constituent tone [7] (similar to Figure 4.10), it is possible to construct sets of possible generators for any edge case. Through knowledge of the specific configuration (which is always known for a given tone), it is possible to use the appropriate subset of $\dashv \bigcup \lrcorner$. Then the sets of possible generators for a false fundamental and its apparent harmonics are given by

$$f_0 : \{\omega^{-1}, \omega^{-2}, x\}, \quad \text{where } x = \begin{cases} (\omega\delta)^{-1} & \text{if } \Psi(\pi_\chi(\ \delta^{-1}(\otimes\ ))) \ = \vdash \\ \omega^{-2}\delta^{-1} & \text{otherwise,} \end{cases} \tag{4.9a}$$

$$f_1 : \{\omega, \omega^{-1}, x\}, \quad \text{where } x = \begin{cases} \delta^{-1} & \text{if } \Psi(\pi_\chi(\ \delta^{-1}(\otimes\ ))) \ = \vdash \\ (\omega\delta)^{-1} & \text{otherwise,} \end{cases} \tag{4.9b}$$

$$f_2 : \{\omega\delta, \delta, x\}, \quad \text{where } x = \begin{cases} \omega^{-1}\delta & \text{if } \Psi(\pi_\chi(\ \otimes\ )) \ = \vdash \\ \omega^{2}\delta & \text{otherwise,} \end{cases} \tag{4.9c}$$

$$f_3 : \{\omega, \omega^{2}, x\}, \quad \text{where } x = \begin{cases} \omega\delta^{-1} & \text{if } \Psi(\pi_\chi(\ \delta^{-1}(\otimes\ ))) \ = \vdash \\ \delta^{-1} & \text{otherwise.} \end{cases} \tag{4.9d}$$

Note the use of shorthand here - these actions are all relative to (and applied to) a false

[7] Note that "constituent tone" refers to each individual ν in $\otimes(f_0)$ - that is, the fundamental, and its first three apparent harmonics.

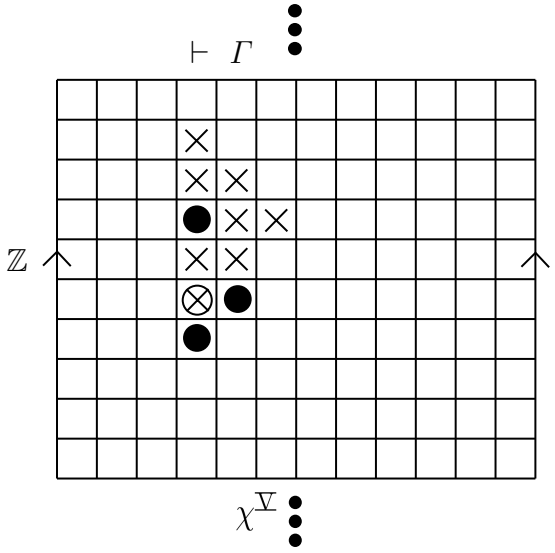


Figure 4.11: A basic edge case in a $\vdash \Gamma$ configuration - note that each of the harmonics associated to the constituent tones of the false fundamental, $\otimes$, have a single generator.

fundamental, $\otimes$.

Further, it is possible to define the notion of a basic edge case;

Definition 4.4.0.2 (Basic Edge Case)**.** *A basic edge case is an edge case such that each constituent tone of a $\otimes$ has precisely one generator.*

In other words, only one element in each of the sets (4.9) is a generator for each f_i in a given $\otimes$. See Figure 4.11 for an example of a basic edge case.

Following from this, it is possible to enumerate every basic edge case for a specific configuration, and ascertain the total number of possible basic edge cases for that configuration. Initially the answer for four choices, each with three options would simply be $3^4 = 81$. Due to overlap in which generators satisfy the constituent, however, the actual result is significantly lower, and can be enumerated with a simple counting method (Table 4.1). Note that f_2 has been omitted as it has no overlap (and therefore, multiplying the end result by 3 is sufficient).

f_0	f_1	f_3
ω^{-1}	-	ω^2
		$\omega\delta^{-1}$
ω^{-2}	ω	-
	δ^{-1}	ω^2
		$\omega\delta^{-1}$
$(\omega\delta)^{-1}$	ω	-
	δ^{-1}	ω^2
		$\omega\delta^{-1}$
	Total	$8 \times 3 = 24$

Table 4.1: Table showing the enumeration of possible basic edge cases for the $\vdash \Gamma$ configuration, with each basic edge case corresponding to a row in the table. Note that a hyphen represents that a choice need not be made as a previous choice already satisfies the harmonic.

The same holds true for the $\Gamma\Gamma$ and $\Gamma \vdash$ configurations, as although the composed actions are different, there are still the same overlaps ($f_0/f_1 : \omega^{-1}$, and $f_1/f_3 : \omega$), and the same number of overall choices.

One might be tempted to claim, therefore, that there are $24 \times 3 = 72$ basic edge cases. Though technically this may be true, we instead define a number of basic edge types - similar to Lent Davis and Maclagan's definition of cap types regarding the card game SET [37]. This allows for comparisons to be made irrespective of configuration. Further, there are a number of invariants that hold across all configurations, for each basic edge type, which can be used to categorise them.

Let

$$g : \mathcal{N}^{\nabla} \to \mathcal{N}^{\nabla} \tag{4.10}$$

be the map sending a tone to its generating fundamental. While this could easily be multi-valued for a generic interpretation, in particular the map gives subsets of (4.9) for edge cases, for basic edge cases the map gives a unique generator, which is the situation in which

this map will be considered. Note that it is assumed here that no inharmonic noise is present, meaning that all tones have a defined generator. Hence the map (4.10) is well-defined. As a demonstration, take the basic edge case shown in Figure 4.11. Applying g to f_1, for example, results in $\omega^{-1}(\otimes)$.

Further consider the triple $g(f_0,\ f_1,\ f_3) = (g(f_0),\ g(f_1),\ g(f_3))$ constructed from applying g to a false fundamental and its first and third apparent harmonics[8]. It is not necessary to consider f_2 as it has no overlap with the other constituent parts (as shown below, with the overlaps bolded for clarity),

$$f_0 : \quad \boldsymbol{\omega^{-1}},\ \omega^{-2}\ ,\ (\omega\delta)^{-1} \tag{4.11a}$$

$$f_1 : \quad \boldsymbol{\omega}\ \ ,\ \boldsymbol{\omega^{-1}},\ \delta^{-1} \tag{4.11b}$$

$$f_2 : \quad \omega^2\delta\ ,\ \omega\delta\ \ ,\ \delta \tag{4.11c}$$

$$f_3 : \quad \boldsymbol{\omega}\ \ ,\ \omega^2\ \ ,\ \omega\delta^{-1}. \tag{4.11d}$$

Any basic edge case can be associated with such a triple, which corresponds to the generating set of the false fundamental and its first and third apparent harmonics.

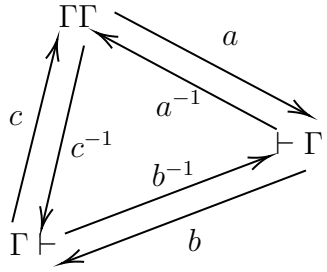
Definition 4.4.0.3. *(Basic Edge Type)*

Two basic edge cases are of the same type iff their two corresponding triples are related by

$$\delta^{-1} \mapsto \omega^k\delta^{-1},\ k \in \{-1, 0, 1\}\ . \tag{4.12}$$

Remark. *Note that if the triple of a basic edge case is obtained from another through the replacement* $\delta^{-1} \mapsto \omega^k\delta^{-1}$ *, then it is possible to move in the opposite direction using the inverse map* $\delta^{-1} \mapsto \omega^{-k}\delta^{-1}$ *.*

[8]Note here that f_0 is assumed to be a false fundamental, so $g(f_0,\ f_1,\ f_3) \neq (g(f_0),\ g(f_0),\ g(f_0))$.



$$a = \delta^{-1} \mapsto \omega\delta^{-1}$$
$$b = \delta^{-1} \mapsto (\omega\delta)^{-1}$$
$$c = \delta^{-1} \mapsto \delta^{-1}$$

Figure 4.12: Diagram showing the relationships between members of the same type between different configurations.

For any given type, there will be precisely *three* members - with precisely one being associated to each of the three configurations. It should be noted that in some cases (I and II in Table 4.2), the member for all configurations is identical, and for all other cases, the members for the ΓΓ and Γ ⊢ configurations are the same. Thus there are in reality one or two *distinct* members, but three configurations. The members of each type (by configuration) are related as according to Figure 4.12.

This becomes clearer following inspection of Table 4.2. Intuitively, the only difference between the ⊢ and Γ shapes is the shifting of f_2. The f_2 in a Γ shape is obtained from the corresponding ⊢ shape by acting ω on its f_2, and its inverse on the contrary.

Lemma 4.4.0.1. *Being of the same type is an equivalence relation.*

Proof. For equivalence, the relation must be reflexive, symmetric, and transitive. The reflexive and symmetric properties may be shown by choosing $k = 0$, and letting $k \mapsto -k$ in Definition 4.4.0.3, respectively. Further, transitivity holds by virtue of the diagram in Figure 4.12. □

Remark. *Note that the types of basic edge cases are independent of tone configuration.*

Figure 4.13 visually represents the eight basic edge types shown in Table 4.2.

As previously mentioned, there are a number of invariants that hold within each type.

These are properties which are the same across each edge type, and provide information about their geometric structure. The invariants considered are; back-δ count ($|\delta^{-1}|$), edge characteristic (ϵ), and generating structure ($G.S$) which are to be defined in the following.

Definition 4.4.0.4 (back-δ count)**.** *The back-δ count, $|\delta^{-1}|$, is a positive integer representing the number of δ^{-1} occurring in a triple $(g(f_0),\ g(f_1),\ g(f_3))$ associated to to a given edge case.*

Example. *The triple $(g(f_0),\ g(f_1),\ g(f_2)) = (\omega^{-1},\ \omega^{-1},\ \omega\delta^{-1})$ has back-δ count $|\delta^{-1}| = 1$.*

Remark. *The back-δ count corresponds to the number of generators in the column to the left of $\otimes$.*

Definition 4.4.0.5 (Edge Characteristic)**.** *The edge characteristic, ϵ, is given by the number of distinct fundamentals that are generators for the given tone. That is, the number of distinct elements in $(g(f_0),\ g(f_1),\ g(f_3))$.*

Example. *The triple $(g(f_0),\ g(f_1),\ g(f_3)) = (\omega^{-1},\ \omega^{-1},\ \omega\delta^{-1})$ has edge characteristic $\epsilon = 2$.*

Remark. *The edge characteristic corresponds to the number of generators for a given $\otimes$, excluding $g(f_2)$.*

In addition to $|\delta^{-1}|$ and ϵ, another invariant is the "Generating Structure" (G.S.), which not only considers the number of generating fundamentals, but also which pairs satisfy overlap (i.e. pairs generated by the same tone). Naively, there are 5 possible generating structures,

$$\text{I. } f_0 = f_1 = f_3,$$

$$\text{II. } f_0 = f_1 \neq f_3,$$

$$\text{III. } f_0 = f_3 \neq f_1,$$

Type	$\lvert\boldsymbol{\delta^{-1}}\rvert$	$\boldsymbol{\epsilon}$	**G.S.**	**Configuration** $\boldsymbol{\Gamma\Gamma/\Gamma\vdash}$	$\boldsymbol{\vdash\Gamma}$
I	0	2	$f_0 = f_1 \neq f_3$	$\{\omega^{-1}, \omega^2\}$	
II			$f_0 \neq f_1 = f_3$	$\{\omega^{-2}, \omega\}$	
III	1	2	$f_0 = f_1 \neq f_3$	$\{\omega^{-1}, \delta^{-1}\}$	$\{\omega^{-1}, \omega\delta^{-1}\}$
IV			$f_0 \neq f_1 = f_3$	$\{\omega^{-2}\delta^{-1}, \omega\}$	$\{(\omega\delta)^{-1}, \omega\}$
V		3	$f_0 \neq f_1 \neq f_3$	$\{\omega^{-2}, (\omega\delta)^{-1}, \omega^2\}$	$\{\omega^{-2}, \delta^{-1}, \omega^2\}$
VI	2	3	$f_0 \neq f_1 \neq f_3$	$\{\omega^{-2}, (\omega\delta)^{-1}, \delta^{-1}\}$	$\{\omega^{-2}, \delta^{-1}, \omega\delta^{-1}\}$
VII				$\{\omega^{-2}\delta^{-1}, (\omega\delta)^{-1}, \omega^2\}$	$\{(\omega\delta)^{-1}, \delta^{-1}, \omega^2\}$
VIII	3	3	$f_0 \neq f_1 \neq f_3$	$\{\omega^{-2}\delta^{-1}, (\omega\delta)^{-1}, \delta^{-1}\}$	$\{(\omega\delta)^{-1}, \delta^{-1}, \omega\delta^{-1}\}$
	Total number of cases			$8 \cdot 3 = 24$	

Table 4.2: Table showing the different types of basic edge case, together with their invariants, and elements (excluding f_2).

IV. $f_0 \neq f_1 = f_3$,

V. $f_0 \neq f_1 \neq f_3$.

Note that each of these really signifies that the *generators* of relevant harmonics are the same - e.g. with I, $g(f_0) = g(f_1) = g(f_3)$. By construction, cases I and III can never occur — III because it is impossible to satisfy with the Γ and $\vdash$ shapes, and I because it would require f_0 to be a generator (and therefore a fundamental). Hence, all basic edge cases exhibit a generating structure of either II, IV, or V.

These invariants help to distinguish between different basic edge cases, and the invariants for each class are enumerated in Table 4.2.

As can be seen in Table 4.2, there is only a single case in which the type is not uniquely determined by $|\delta^{-1}|$, ϵ, and G.S. - namely types 6 and 7. These can be distinguished by considering the position of the generator that sits in the same column as the false fundamental - for type 6, the generator sits below the false fundamental (ω^{-2}), whereas for type 7, the generator sits above the false fundamental (ω^2).

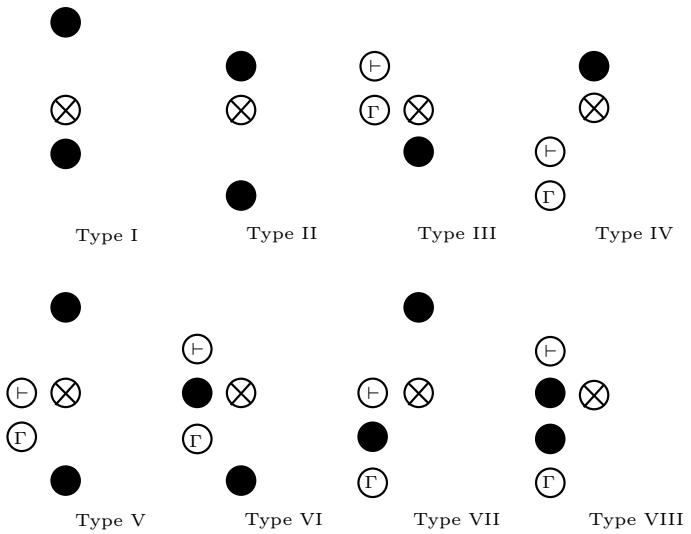


Figure 4.13: The eight basic edge types represented visually. Each • represents a generator, with ⊗ representing the false fundamental, and the unfilled generators containing ⊢ or Γ denoting the shape drawn out in the δ^{-1} column (i.e. corresponding to $\Psi(\pi_\chi(\delta^{-1}(\otimes)))$).

Remark. *Note that, as before, the multiplication by 3 in Table 4.2 is due to there being no restrictions on the choice of generator for the second harmonic.*

Lemma 4.4.0.2. *The second "harmonic" (f_2) of a false fundamental (⊗) must be generated from the column directly to the right of the false fundamental (i.e. $\Psi(\pi_\chi(\delta(\otimes)))$).*

Proof. Assume that there exists some generator $g_i \in \mathcal{N}^{\nabla}$ for f_2 that lies in the same column as the false fundamental[9]. There are two possible cases,

Case 1: $\pi_\chi(g_i) \in \chi_\vdash$.

For g_i to generate the f_2 at $\omega\delta(\otimes)$, it would have to lie at $\mathbb{1}(\otimes)$. Hence, ⊗ would no longer be a false fundamental - resulting in a contradiction.

Case 2: $\pi_\chi(g_i) \in \chi_\Gamma$.

The same argument as Case 1, applying the map $\omega\delta \mapsto \omega^2\delta$.

[9]Note that the only other choice would be the column directly to the right of ⊗, as the ⊢ and Γ shapes trace only once cell to the right, and therefore no harmonic in the column to the right of a ⊗ could be generated by generator in the column to the left of ⊗.

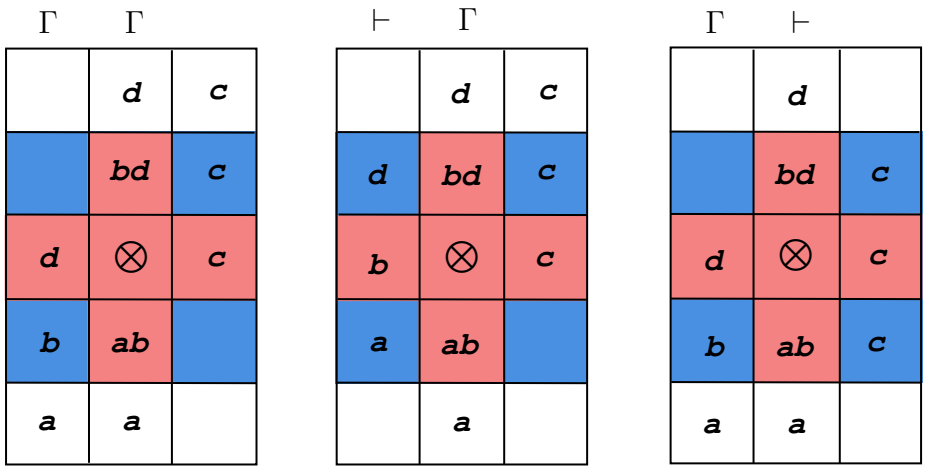


Figure 4.14: The potential generators (a (f_0), b (f_1), c (f_2), d (f_3)) for each configuration, and the von Neumann (red) and Moore (coloured) neighbourhoods.

Thus, as a contradiction is reached in both possible cases, there can be no such generator for f_2. It follows that any generators for f_2 must lie in the column directly to the right of the false fundamental. □

Proposition 4.4.1. *There are 24 basic edge types.*

Proof. This follows from Table 4.2, and Lemmas 6.1.1 and 6.1.2. □

It is also possible to define restrictions on the presence of generators for a false fundamental (with respect to the configuration in which it sits). In order to do this, the minimum number of generators that must fall in certain proximity (such as the von Neumann and Moore neighbourhoods) to a false fundamental can be considered.

In terms of the operators ω, δ, the von Neumann- and Moore-neighbourhoods of some tone ν, are the tones generated by acting the elements of the sets

$$\{\delta^{\pm 1}, \omega^{\pm 1}\} \quad \text{and} \quad \{\delta^{\pm 1}, \omega^{\pm 1}, \omega^{\pm 1}\delta^{\pm 1}\}, \tag{4.13}$$

on ν, respectively.

	$\Gamma\Gamma$	$\vdash\Gamma$	$\Gamma\vdash$
a	ω^{-2} (I)	ω^{-2} (I)	ω^{-2} (I)
b	$(\omega\delta)^{-1}$ (II)	δ^{-1} (III)	$(\omega\delta)^{-1}$ (II)
c	$\omega^2\delta$ (I)	$\omega^2\delta$ (I)	$\omega\delta$ (II)
d	ω^2 (I)	ω^2 (I)	ω^2 (I)
M	1	1	2
v.N.	0	1	0

Table 4.3: Table showing the minimum generators in the von Neumann (v.N.) and Moore neighbourhoods of a false fundamental given its chroma configuration.

The problem of choosing basic edge cases with the least generators in these neighbourhoods can be reduced to the problem of choosing some a, b, c, and d (corresponding to generators for $f_0, ..., f_3$) from Figure 4.14. This can be achieved by choosing generators according to the following order,

I. Outside both neighbourhoods;

II. Inside Moore neighbourhood, outside of von Neumann neighbourhood;

III. Inside von Neumann neighbourhood (and $\therefore$ inside Moore neighbourhood).

If multiple choices are available, it is sufficient to choose any one, without loss of generality, as they have the same effect on the final count as one another. Table 4.3 shows the resulting (minimum) counts of generators in the neighbourhoods for false fundamentals of certain configurations.

Though it may seem that choosing a generator that satisfies multiple parts (i.e. ω^{-1} or ω) may reduce the overall counts, it is always possible in these cases to instead make choices that don't reside in the von Neumann neighbourhood.

By combining the notion of basic edge types (Table 4.2) with the restrictions on neighbourhoods (Table 4.3), it is possible to reach an even more constrained characterisation

Type	Neighbourhood	Configuration		
		$\vdash \Gamma$	$\Gamma \vdash$	$\Gamma\Gamma$
I	v.N.	1	1	1
	M	1	2	1
II	v.N.	1	1	1
	M	1	2	1
III	v.N.	1	2	2
	M	2	3	2
IV	v.N.	1	2	1
	M	2	2	1
V	v.N.	1	0	0
	M	1	2	1
VI	v.N.	1	1	1
	M	2	3	2
VII	v.N.	1	0	0
	M	2	2	1
VIII	v.N.	1	1	1
	M	3	3	2

Table 4.4: Restrictions on the minimum number of generators in the von Neumann (v.N.) and Moore (M) neighbourhoods, by basic edge type and configuration.

of false fundamentals (Table 4.4). For example, for a Type III basic edge case in a $\Gamma \vdash$ configuration, at least two generators must sit in the von Neumann neighbourhood (with a further one in the Moore neighbourhood), whereas considering Table 4.3 there is only a necessity for two generators in the Moore neighbourhood for a case of unspecified basic edge type falling into the same configuration.

Better understanding the occurrence of edge cases is an important step towards identifying them in practice, and gives a deeper understanding of the proposed model itself. Sections 4.5 and 5.1 go on to look at reduction of edge cases to potential basic cases, and the experimental prevalence of basic edge types, and Sections 5.2 and 5.3 investigate the theoretical basis of the model from a more experimental standpoint.

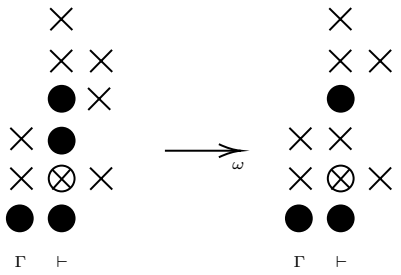


Figure 4.15: A reduction removing a generator in $\vdash \bigcup \Gamma$. Note that $g(f_2)$ is omitted for brevity.

4.5 Reduction and Reducibility of Edge Cases

In order to gain a better understanding of the occurrence of edge cases (and therefore the problem of pitch estimation), it proves useful to be able to classify edge cases by which basic edge types they are related to. In order to achieve this, it is necessary to reduce edge cases (i.e. remove redundancy) by removing potential generators such that the false fundamental in question is still preserved.

Given a set of generators, $\mathcal{G}$, that lie in $\hexagon(\otimes)$ for some false fundamental, they are reducible iff any part, $f_n \in \{f_0,\ f_1,\ f_2,\ f_3\}$ is satisfied more than once (i.e. non-basic) - barring the exception outlined below. Reduction (denoted as $\to_{\mathfrak{g}}$) takes $\mathcal{G}$, and removes some given generator $\mathfrak{g} \in \mathcal{G}$ such that the false fundamental is still satisfied by $\mathcal{G} \setminus \mathfrak{g}$,

$$\mathcal{G} \ \to_{\mathfrak{g}} \ \mathcal{G} \setminus \mathfrak{g}. \tag{4.14}$$

Such a removal of a generator seeks only to remove its 'fundamentalness' - it is entirely possible that it could still be generated elsewhere. Indeed, this must be the case for any reduction via a generator in $\vdash \bigcup \Gamma$, such as in Figure 4.15. Note that reduction is not unique; there may be multiple valid reductions that can be applied to a given set of generators[10].

[10]It is worth further noting that one could reduce 'globally' (i.e. over $\mathcal{N}_\alpha^{\mathbb{V}}$), or 'locally', considering just $\hexagon(\otimes)$. Because false fundamentals generally need only be considered locally (as they can be generated solely by fundamentals within $\hexagon(\otimes)$), it is only necessary to consider at most $3 \times 5 = 15$ possible tones to reduce, which renders the computational complexity significantly lower than might be expected for reduction graphs

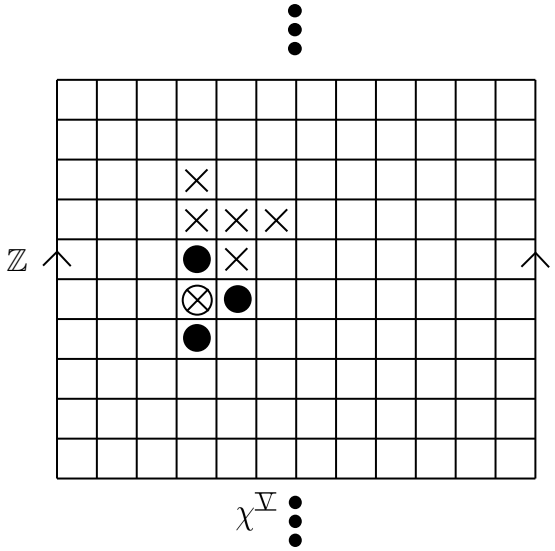


Figure 4.16: An irreducible non-basic edge case.

It would be reasonable to assume that any non-basic edge case can be reduced, therefore, to one of the eight basic edge cases (Table 4.2). On the contrary, however, there exists a case that is both non-basic (i.e. at least one of its parts is satisfied more than once) and irreducible - that is, that no potential generator could be removed whilst preserving the false fundamental (Figure 4.16). Importantly however, this is the only irreducible non-basic edge case.

Proposition 4.5.1. *The only irreducible non-basic edge cases are those containing both* ω^{-1} *and* ω*.*

Proof. Let G_n be the set of generators that generate f_n.

For the proposition to hold, it is sufficient to show that there exists an irreducible non-basic case containing ω^{-1} and ω, and that all other cases (i.e. those with neither generator, or those with precisely one of them) reduce to a basic case. The former statement is shown in Figure 4.16.

Thus, it remains to prove the latter. In addition, the case where both ω^{-1} and ω are

with large numbers of fundamentals.

	B	**C**	**D**
$\mathbf{f_0}$:	$\{\omega^{-1}\} \cup G_0$	G_0	$\{\omega^{-1}\} \cup G_0$
$\mathbf{f_1}$:	$\{\omega^{-1}\} \cup G_1$	$\{\omega\} \cup G_1$	$\{\omega^{-1},\ \omega\} \cup G_1$
$\mathbf{f_2}$:	G_2	G_2	G_2
$\mathbf{f_3}$:	G_3	$\{\omega\} \cup G_3$	$\{\omega\} \cup G_3$

Table 4.5: The most general generators for each case, B (just ω^{-1}), C (just ω), and D (both ω^{-1}, and ω).

present, together with other generators, is considered, and it turns out that such cases are either reducible to a basic edge case, or to the irreducible non-basic case in Figure 4.16.

Note that a lack of overlap, i.e $|G_i| \cap |G_j| = \emptyset$, for all $i \neq j$, implies that a case can always be reduced to a basic edge case. That is, cases where each generator generates precisely one of the constituent parts of the false fundamental can always be reduced to a basic case.

Case A (Neither ω^{-1}, nor ω):

When neither generator is present, there is no overlap in generators, and so one can reduce until $|G_n| = 1$, $\forall\, n$. The resulting basic edge case will be one of type V-VIII.

Case B (ω^{-1}, but not ω):

As shown from Tables 4.5 and 4.6a, there are four cases to consider, related to the number of generators of f_0 and f_1.

1. $|\mathbf{G_0}| = \mathbf{1} \wedge |\mathbf{G_1}| = \mathbf{1}$:

   In this case, both sets are the singleton set $\{\omega^{-1}\}$, and thus the case is irreducible, but basic[11].

2. $|\mathbf{G_0}| > \mathbf{1} \wedge |\mathbf{G_1}| = \mathbf{1}$:

   In this case, it is only possible to reduce by the non-ω^{-1} element(s) of G_0, as a reduction

[11] After arbitrary reduction of f_3. This will be implicit in the remaining cases.

Case	Condition
B-1	$\|G_0\| = 1 \wedge \|G_1\| = 1$
B-2	$\|G_0\| > 1 \wedge \|G_1\| = 1$
B-3	$\|G_0\| = 1 \wedge \|G_1\| > 1$
B-4	$\|G_0\| > 1 \wedge \|G_1\| > 1$

(a) The four cases for case B (just ω^{-1}).

Case	Condition
C-1	$\|G_1\| = 1 \wedge \|G_3\| = 1$
C-2	$\|G_1\| = 1 \wedge \|G_3\| > 1$
C-3	$\|G_1\| > 1 \wedge \|G_3\| = 1$
C-4	$\|G_1\| > 1 \wedge \|G_3\| > 1$

(b) The four cases for case C (just ω).

Case	Condition
D-1	$\|G_0\| = 1 \wedge \|G_1\| = 2 \wedge \|G_3\| = 1$
D-2	$\|G_0\| > 1 \wedge \|G_1\| = 2 \wedge \|G_3\| = 1$
D-3	$\|G_0\| = 1 \wedge \|G_1\| = 2 \wedge \|G_3\| > 1$
D-4	$\|G_0\| > 1 \wedge \|G_1\| = 2 \wedge \|G_3\| > 1$
D-5	$\|G_0\| = 1 \wedge \|G_1\| > 2 \wedge \|G_3\| = 1$
D-6	$\|G_0\| > 1 \wedge \|G_1\| > 2 \wedge \|G_3\| = 1$
D-7	$\|G_0\| = 1 \wedge \|G_1\| > 2 \wedge \|G_3\| > 1$
D-8	$\|G_0\| > 1 \wedge \|G_1\| > 2 \wedge \|G_3\| > 1$

(c) The eight cases for case D (both ω^{-1} and ω).

Table 4.6: The sub-cases examined for cases B through D.

by ω^{-1} would result in f_1 no longer being present. Thus the only reduction is to Case B-1, which is basic.

3. $|\mathbf{G_0}| = \mathbf{1} \wedge |\mathbf{G_1}| > \mathbf{1}$:

   Similar to Case B-2, this can only be reduced by the non-ω^{-1} elements, this time of G_1. Again, this leads to reduction to Case B-1.

4. $|\mathbf{G_0}| > \mathbf{1} \wedge |\mathbf{G_1}| > \mathbf{1}$:

   Finally, with this case there are two possible ways to reduce - either down to a case in which ω^{-1} is the sole generator of f_0 and f_1, corresponding to one of the previous two cases (B-2, or B-3), or reduction by ω^{-1} itself, resulting in Case A. Regardless, neither route results in an irreducible non-basic case, as required.

For **Case C** (just ω) and **Case D** (both ω^{-1} and ω), the approach is almost identical to that laid out in Case B, and both are therefore omitted for brevity. See Tables 4.6b and

4.6c for the sub-cases. Figure 4.17 graphically encapsulates the reductions that relate each case.

Thus, considering Figure 4.17, it is clear that the only terminal nodes are Cases A, B-1, C-1, and D-1. As the first three are basic, this leaves D-1 as the only irreducible non-basic case. Given that this case contains both ω^{-1} and ω, it follows that the only irreducible non-basic edge cases are those containing both ω^{-1} and ω, as required.

□

In fact, the following Corollary holds by Figure 4.17.

Corollary 4.5.1.1. *There is precisely one irreducible non-basic edge case.*

It is worth noting that there are really three such cases, when taking into account the arbitrary choice of the generator for f_2.

Through repeated application of all possible reductions to the vertices (to which reduction is yet to be applied), one may obtain a reduction graph for a given set of generators, $\mathcal{G}$ (Figure 4.18). Given that no reduction could ever produce a set of generators larger than the input, this graph will additionally be acyclic. Further, the terminal vertices (i.e. $\deg^+(n) = 0$, where n is a vertex) of such graphs correspond to irreducible cases[12]. Such a graph can, therefore, be used in order to understand the potential basic edge types that correspond to a given set of generators.

A different approach to reduction would be to look not at proportions of basic edge types – for example $\frac{1}{3}$ Type III, $\frac{1}{3}$ Type $\varnothing$ and $\frac{1}{3}$ Type I in Figure 4.18 – but rather, at the sets of basic edge types that a case reduces to, i.e. $\{III, I, \varnothing\}$ for the same case.

[12]As there may be multiple terminal vertices in a given graph, it would be worthwhile to follow a unified algorithm if traversing in a depth-first manner - for example, by always reducing by the bottom-leftmost option.

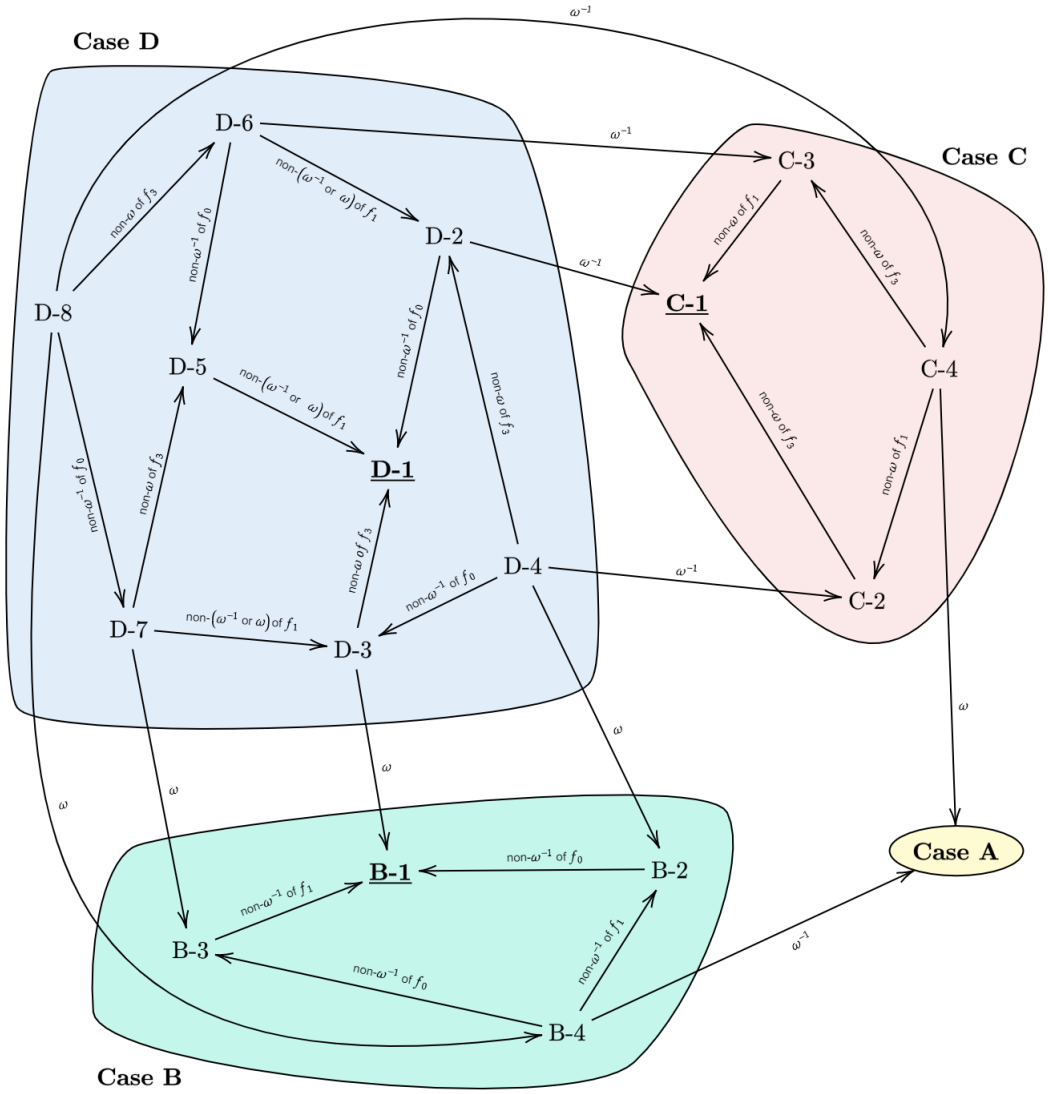


Figure 4.17: A diagram representing the graphical structure relating the various cases laid out in Proposition 6.1. For the terminal vertices (bolded), Cases A, B-1, and C-1 are all basic, and D-1 is the irreducible non-basic case.

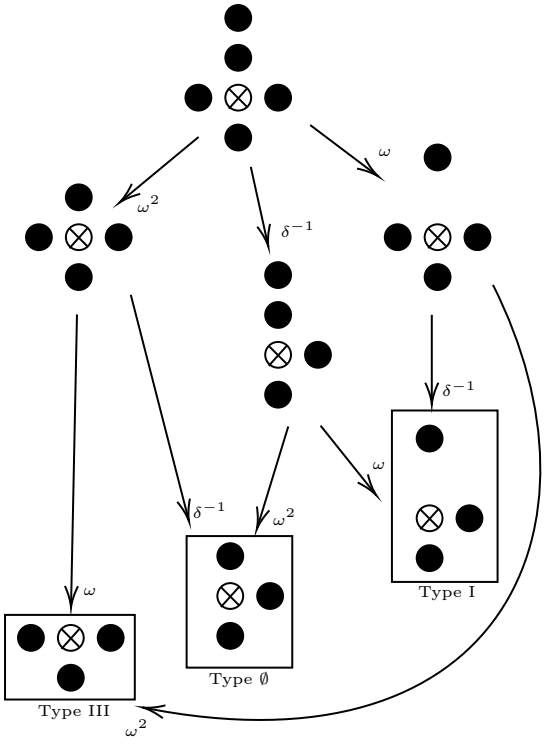


Figure 4.18: An example of a reduction graph, with each step (arrow) showing a reduction in the set of generators. Note that the special case of $\{\omega,\ \omega^{-1},\ g(f_2)\}$ is denoted as 'Type $\varnothing$', and the configuration is $\Gamma\Gamma$ or $\Gamma\vdash$.

Figure 4.19 shows the beginning of a large reduction with 10 generators in $\hexagon(\otimes)$. Though the whole graph would prove too large to contain within this Thesis (whilst respecting the trees), it should be noted that it is absolutely possible to computationally enumerate in its entirety, and a vast overestimate of its total vertices would be $10! = 3628800$. This highlights, however, the need for local reduction of edge cases, as opposed to looking to reduce across the whole of $\mathcal{N}_\alpha^{\nabla}$, as it is relatively clear that we could not computationally handle the reduction of a case with increasingly many vertices.

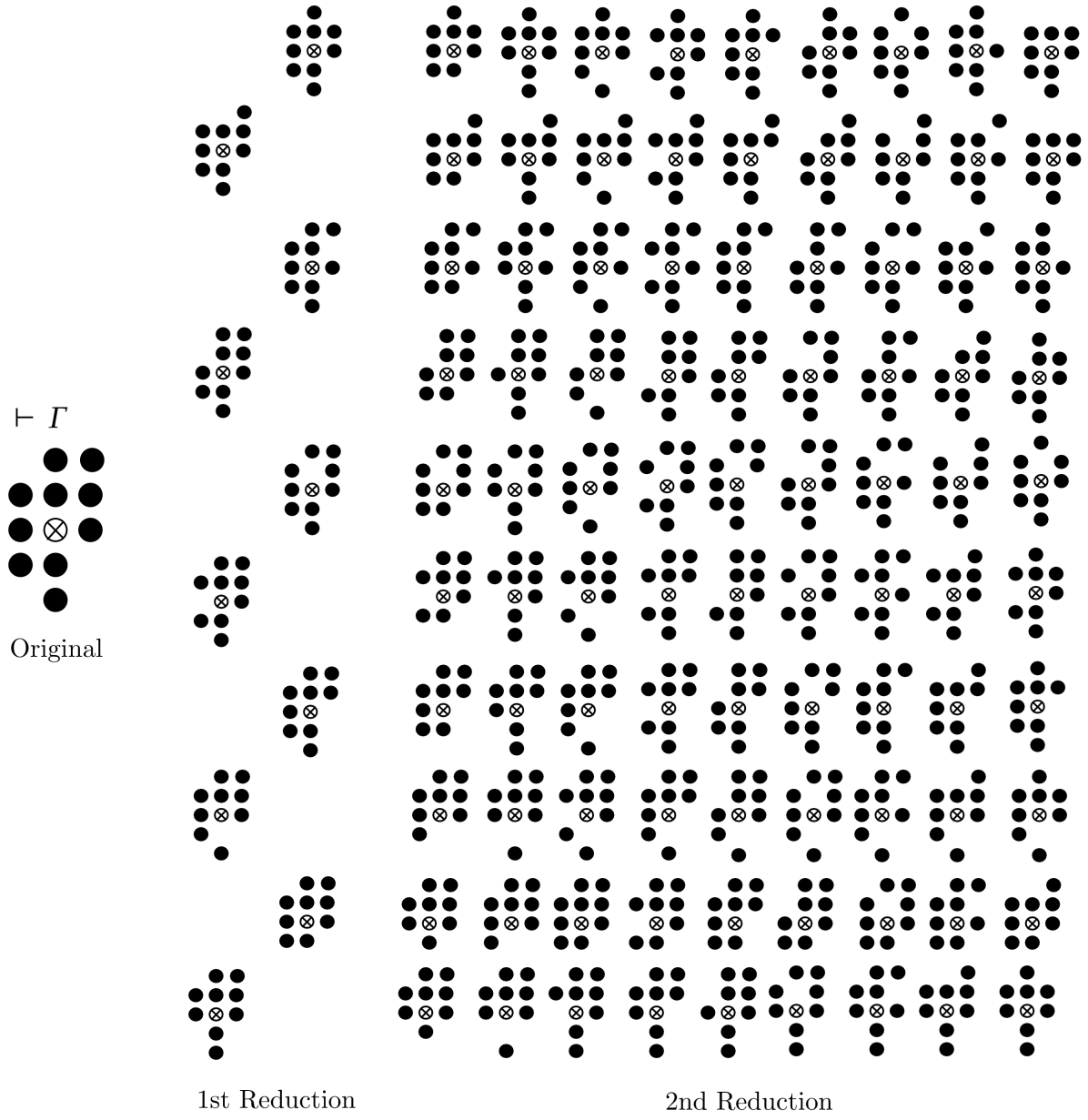


Figure 4.19: Enumeration of the first two reduction steps of a ⬡(⊗) for a ⊢ Γ configuration, given the maximum number of possible generators. Each row shows an edge case reached from a single reduction on the original case, and the corresponding 10 cases following a second reduction.

Chapter Five

Experimental Investigations

In contrast to Chapter 4, in this Chapter we concern ourselves with the application of the model to real-world (and simulated) data. Section 5.1 looks at the theoretical prevalence of different basic edge types by performing repeated reductions (and thus enumerating full reduction graphs) on simulated $\mathcal{N}^{\mathbb{V}}$ data. Section 5.2 presents work on displaying real data in the model by using interpretations $\mathcal{I} : \mathcal{N}^{\mathbb{V}} \to \mathbb{R}$ as opposed to $\mathcal{I} : \mathcal{N}^{\mathbb{V}} \to \mathbb{B}$. Finally, Section 5.3 gives a basic evaluation of simple algorithms working over $\mathcal{N}^{\mathbb{V}}$.

5.1 Prevalence of Basic Edge Types

As previously mentioned, it is important to understand the occurrence of false fundamentals, in order to better differentiate them from genuine ones. One such way is to consider the prevalence of each type in practice. This can be achieved by constructing a reduction graph for sample interpretations (from all combinatorial possibilities) with varying numbers of fundamentals (and their first three harmonics).

Two separate ways of representing the result of such a reduction are used. Subsection 5.1.1 considers each graph to reduce equally to each Type reached (e.g. I, III, and $\varnothing$ in

Figure 4.18). From a different perspective, Subsection 5.1.2 instead considers reduction to a *set* of terminal vertices—i.e. $\{I, III, \varnothing\}$ for the same example. Both approaches lead to interesting results that help to empirically constrain the occurrence of false fundamentals, $\otimes$.

5.1.1 With Reduction to Proportion of Individual Types

In order to sample interpretations from the total sample space, it proves sufficient to construct them by selecting n unique tones, $\nu_0, \nu_1, ..., \nu_n$ from $\mathcal{N}_\alpha^\nabla$, treating all such tones as fundamentals, and thus adding them (and their harmonics) to the interpretation. For the charts in Figures 5.1, 5.2, 5.3 and 5.4, a sample size of 1000 interpretations was taken for each number of simultaneous fundamentals (0, 120], with each generated at random by choosing tones from $\mathcal{N}_\alpha^\nabla$ to act as generators until n unique generators were selected. No effort was made to ensure each interpretation was unique from the next as the chance of this occurring (bar for extraordinarily many or few generators) is statistically improbable. For each of these interpretations, a naive algorithm (simply classifying all $\vdash$ and Γ-exhibiting tones as fundamentals) was applied, and a reduction graph derived from each $\otimes$ - where the difference between the input set and the result of the naive algorithm is the set of false fundamentals. Note that such a naive algorithm does not seek to remove fundamentals in the same way that a more sophisticated algorithm may, and therefore the order of traversal is unimportant. Instead, every tone in $\mathcal{N}_\alpha^\nabla$ that exhibits the expected shape is classified as a fundamental. From each of these reduction graphs, all terminal vertices were classified either as one of the basic edge types, or as the special case, $\varnothing$. In cases where multiple terminal vertices were present, a value was added to each tally such that the sum of all added values was one.

Though 1000 interpretations may at first appear to be a relatively small sample size, it

should be noted that this corresponds to 120,000 interpretations sampled on the whole, with an average of 15.75 (16) false fundamentals per interpretation. These are, as expected, concentrated around the centre of the distribution (of total simultaneous fundamentals), as the number of total possible false fundamentals peaks around the centre. Thus, on average, each set of 1000 interpretations leads to 15750 false fundamentals to classify, but with relatively sparse distribution to the tail-ends (i.e. < 10 simultaneous fundamentals), which resulted in < 100 false fundamentals being classified per 1000 interpretations. In order to ascertain a more reliable picture of the makeup of false fundamentals - particularly with low numbers of simultaneous fundamentals, a significantly larger sample size of 20000 interpretations was used (Figures 5.5, 5.6, and 5.7).

Looking at Figures 5.1, 5.2, 5.3, and 5.4, it is clear that not all basic edge types are equally common. Figure 5.1 gives the overall occurrence of each type, with type III being the most common (along with the other three-fundamental cases), and type V being the least common (along with the other four-fundamental cases). The special case, $\varnothing$ appears to sit between the two, which intuitively coheres with other observations, as it too is a three-fundamental case - albeit not basic. In general, it is hard to draw meaningful insight from this, which incentivises the use of Figure 5.3 - looking at the trends of the prevalences as the number of fundamentals changes.

As Figure 5.3 shows, at low numbers of simultaneous fundamentals, the three-fundamental cases are significantly more dominant than the four-fundamental cases - constituting almost 100% of the cases until around 16 simultaneous fundamentals. Beyond this point, the incidence of four-fundamental cases increases significantly - particularly types 6 and 8 - with the special case $\varnothing$ notably occurring increasingly less often. As before, it is hard to directly relate these results to real-world data (i.e. recorded music), which is much more structured than the random samples that were used, but a number of conclusions can still be drawn,

- With low numbers of simultaneous fundamentals (eg. string quartet), cases 5-8 are incredibly unlikely to occur.

- From Figure 5.4, it is clear that even at large numbers of fundamentals, the accuracy of even the naive algorithm on polyphonic music - with perfect noise removal, recording, playing, etc. - is above around 75%.

Regarding Figure 5.2, graphs appear to have between three and five (of a possible nine) terminal vertices, with the average broadly decreasing as the number of fundamentals grows. The trend appears more turbulent towards the left tail, which is likely due to the low number of samples for these numbers of fundamentals.

By combining this knowledge with a heuristic for the number of fundamentals at a given $\mathcal{I}_\tau$, it may be possible to more easily distinguish between fundamentals and false fundamentals by comparing specific examples to the profile laid out above.

Figures 5.5, 5.6, and 5.7 consider specifically the cases for which there are a low (< 10) number of simultaneous fundamentals. In these cases, the total occurrence of four-fundamental basic cases is, on average, 3.6%, with the majority of these weighted towards interpretations with > 6 simultaneous fundamentals (Figures 5.5 & 5.7). Particularly interestingly, the most common case in this subset of interpretations is the special case, $\varnothing$, with 20.4% of the total. Overall, the trend of three-fundamental cases being more common remains, but the ordering within these groupings change - most notably (beyond $\varnothing$'s jump) with type 5 cases being significantly more prevalent than their counterparts compared to the data in Figure 5.3.

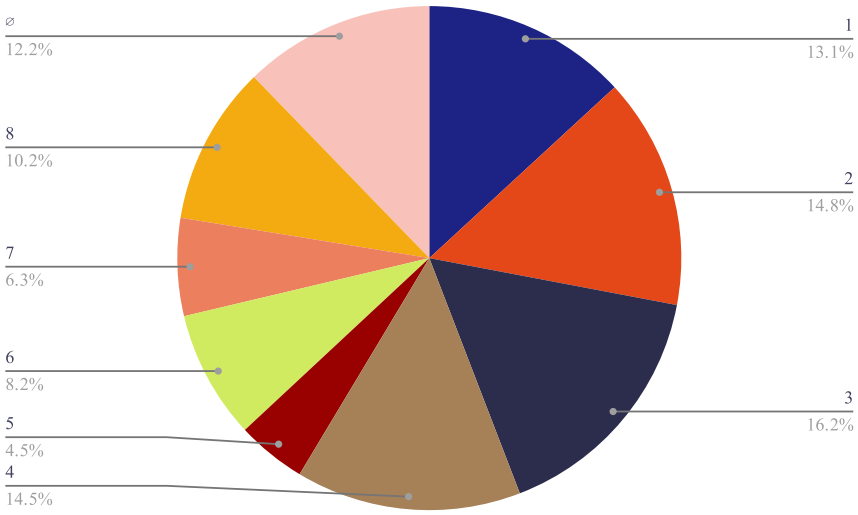


Figure 5.1: A pie chart showing the average (proportional) prevalence of each edge type, and ∅ when considering between 0 and 120 simultaneous fundamentals.

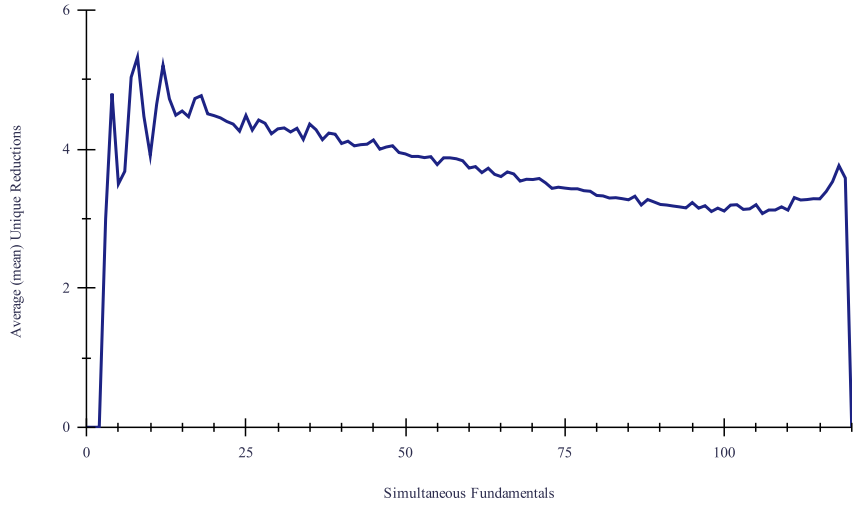


Figure 5.2: A graph showing the estimated average number of terminal vertices for varying numbers of simultaneous unique fundamentals.

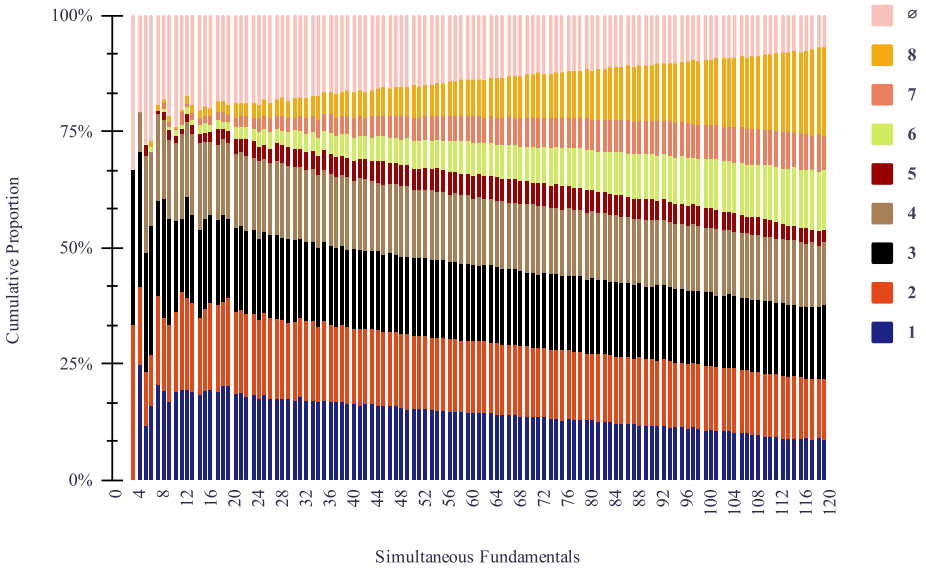


Figure 5.3: A stacked bar chart showing the change in proportional prevalence of basic edge types and ∅ as the number of simultaneous unique fundamentals changes.

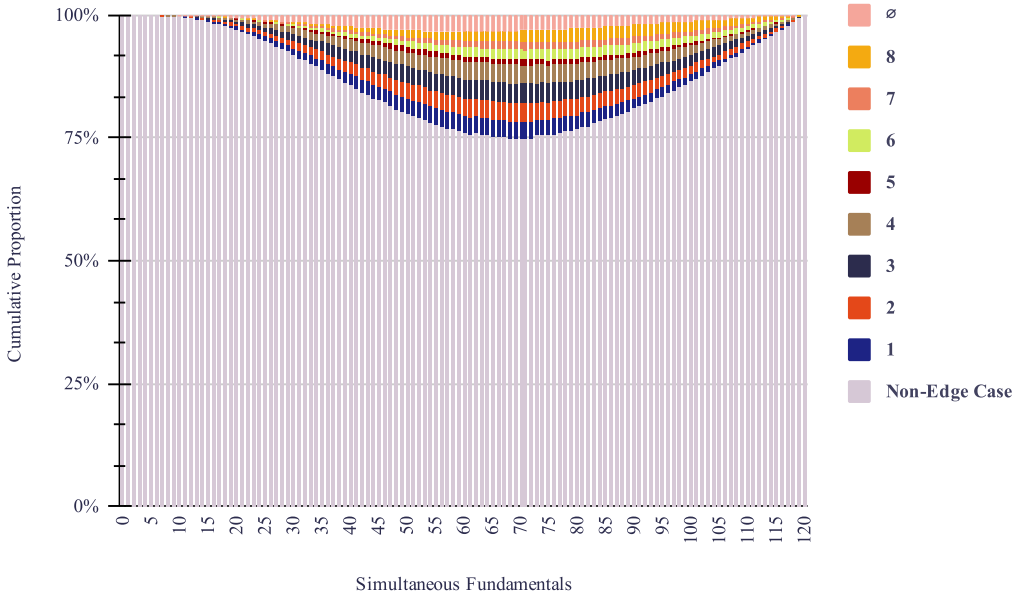


Figure 5.4: Another stacked bar chart, mirroring Figure 5.3, but including the proportion of non-edge cases.

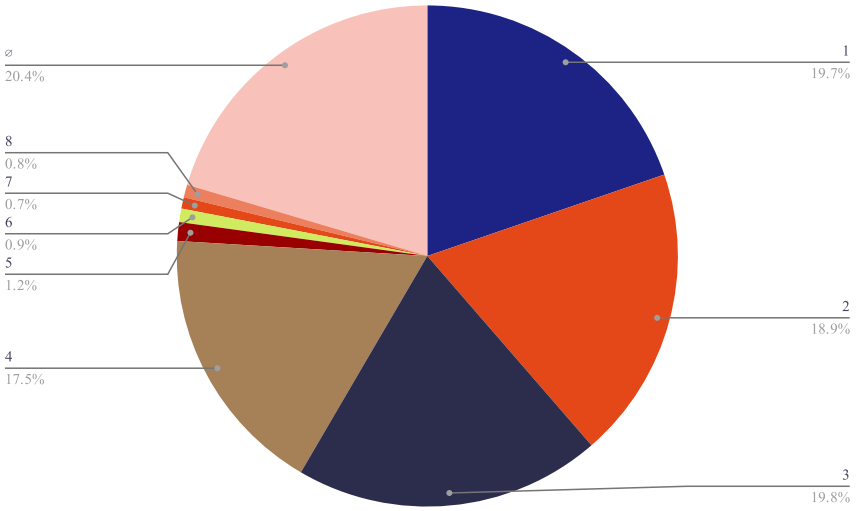


Figure 5.5: A pie chart showing the average (proportional) prevalence of each edge type, and ∅ for total simultaneous fundamentals (0, 10].

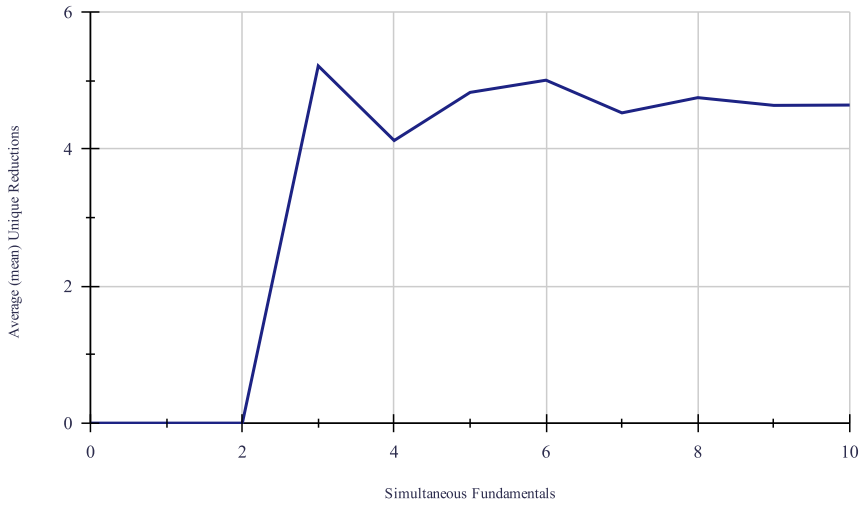


Figure 5.6: A graph showing the estimated average number of terminal vertices for varying numbers of simultaneous unique fundamentals (0, 10] - corresponding to the left hand side of Subfigure 5.2. Note that contrary to Subfigure 5.2, 20000 sample interpretations were taken here, as opposed to 1000.

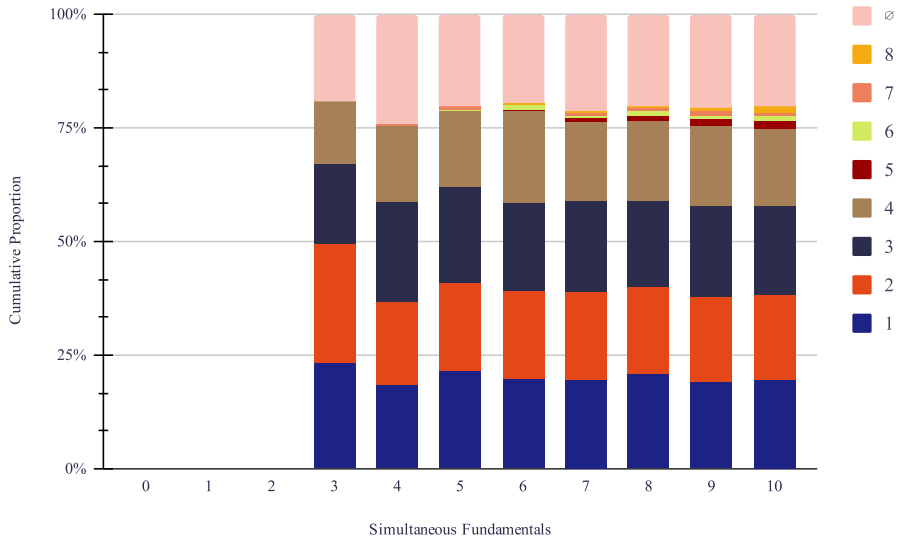


Figure 5.7: A stacked bar chart showing the change in proportional prevalence of basic edge types and ∅ for (0, 10] simultaneous unique fundamentals.

5.1.2 With Reduction to Sets of Terminal Vertices

I would like to start this Subsection with a quick thanks to the reviewer from the Journal of Mathematics and Music that suggested this line of enquiry to me during the review process of the aforementioned paper there [68]. Rather than considering the various terminal vertices to correspond to proportions, here we instead consider the notion of them making up a *terminal vertex set*—that is, a set containing the types of all terminal vertices for a given graph. It may be expected that most, if not all of the $2^9 = 512$ possible subsets of $\{1, 2, 3, ..., \varnothing\}$ would appear in such an analysis, but in fact, only 92 of them actually do in practice. This Subsection will provide an in-depth look at the terminal vertex sets that arise most commonly, as well as a brief explanation of why certain cases cannot ever exist in reality. The simulations of $\mathcal{N}^{\nabla}$ are identical to those used for Subsection 5.1.1, but with a flat sample size of 5000, as opposed to 1000 (and 20000). It is worth noting that in many of the following figures, the notations ∅ and 9 are used interchangeably to represent the irreducible non-basic edge type.

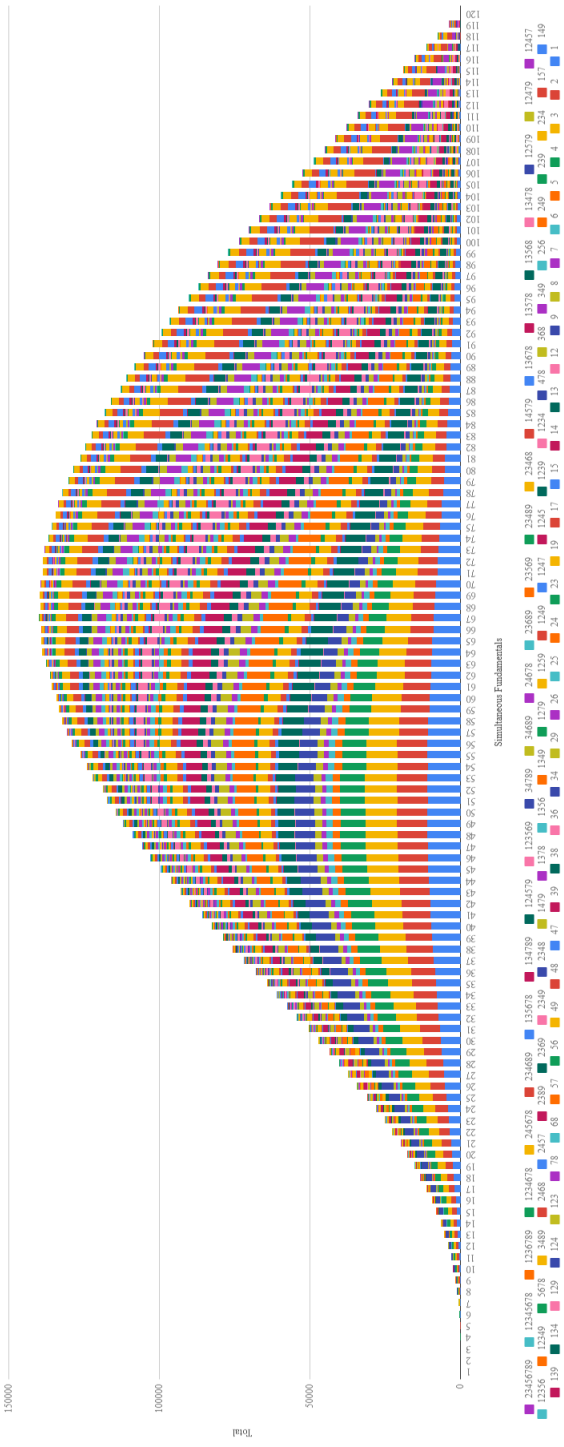


Figure 5.8: Number of terminal vertex sets against the number of simultaneous fundamentals, with a sample size of 5000 simulations. Note that within the stacked bars, the cases count upwards from 1 to 23456789 inclusive.

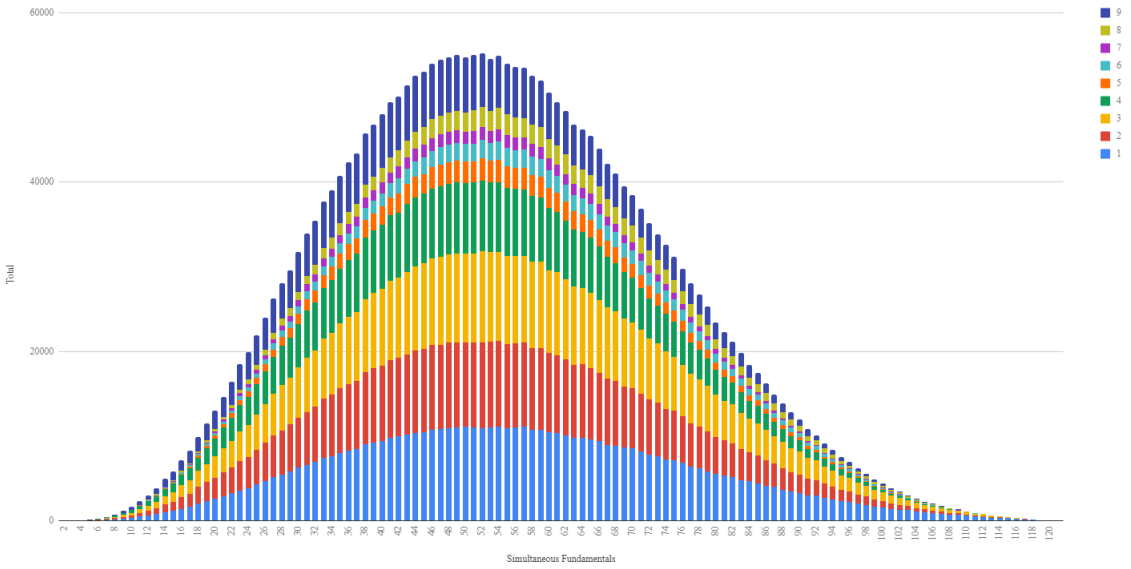


Figure 5.9: Breakdown of terminal vertex sets with cardinality 1 against number of simultaneous fundamentals.

Figure 5.8 shows the prevalence of all of the occurring cases as the number of simultaneous fundamentals changes. This really corresponds to the union of nine roughly normal distributions, corresponding to the sets, S, such that $card(S) = n, n \in \{1, ..., 8\}$, which are depicted in Figures 5.9 through 5.16 respectively. Note that no cases were found with a cardinality of 9 (i.e. all types present).

A number of interesting conclusions can be drawn from looking at how these terminal vertex sets vary with the number of simultaneous fundamentals in $\mathcal{N}^{\nabla}$. Firstly, as the number of simultaneous fundamentals increases, so does the occurrence of higher cardinality sets—and vice versa. Further to this, it can be observed that these higher cardinality sets are significantly rarer than those of a lower cardinality (Figure 5.17).

In addition to this, a closer look at which subsets of $\{1, 2, ..., \varnothing\}$ occur for a given cardinality (and perhaps more pertinently, which *don't* occur), it is possible to highlight constraints on which graphs actually exist. Take the sets, S, with cardinality 2. All such subsets can be enumerated by considering the lexicographical ordering of a 9-bit sequence,

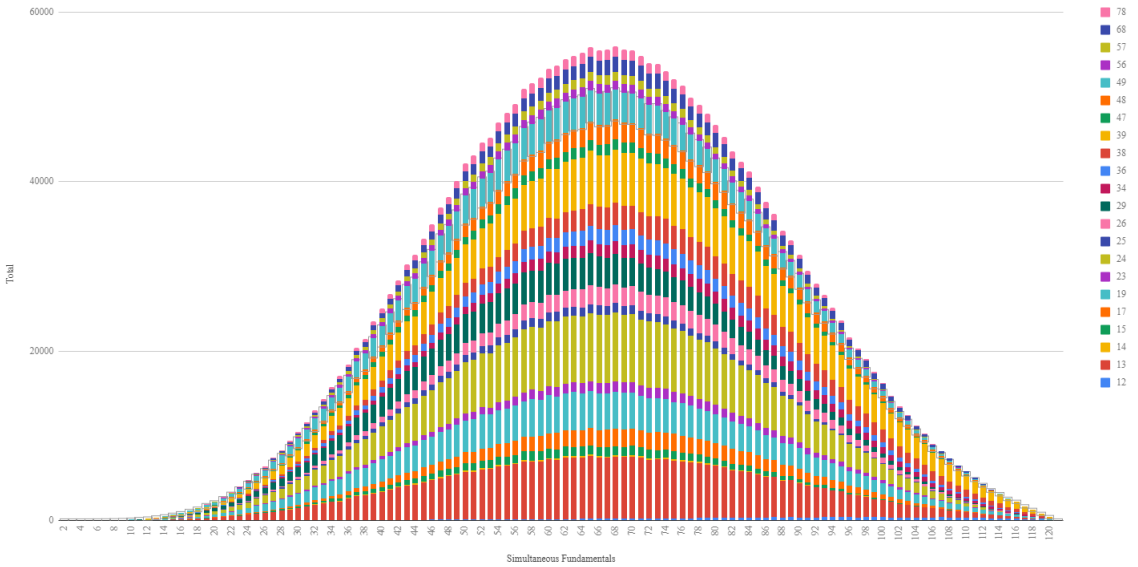


Figure 5.10: Breakdown of terminal vertex sets with cardinality 2 against number of simultaneous fundamentals.

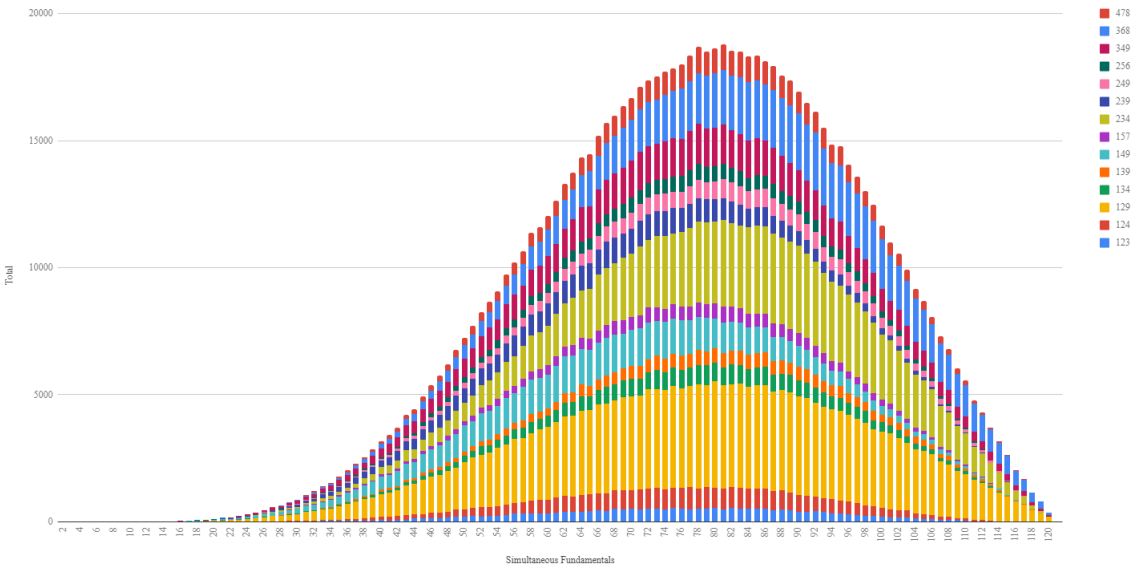


Figure 5.11: Breakdown of terminal vertex sets with cardinality 3 against number of simultaneous fundamentals.

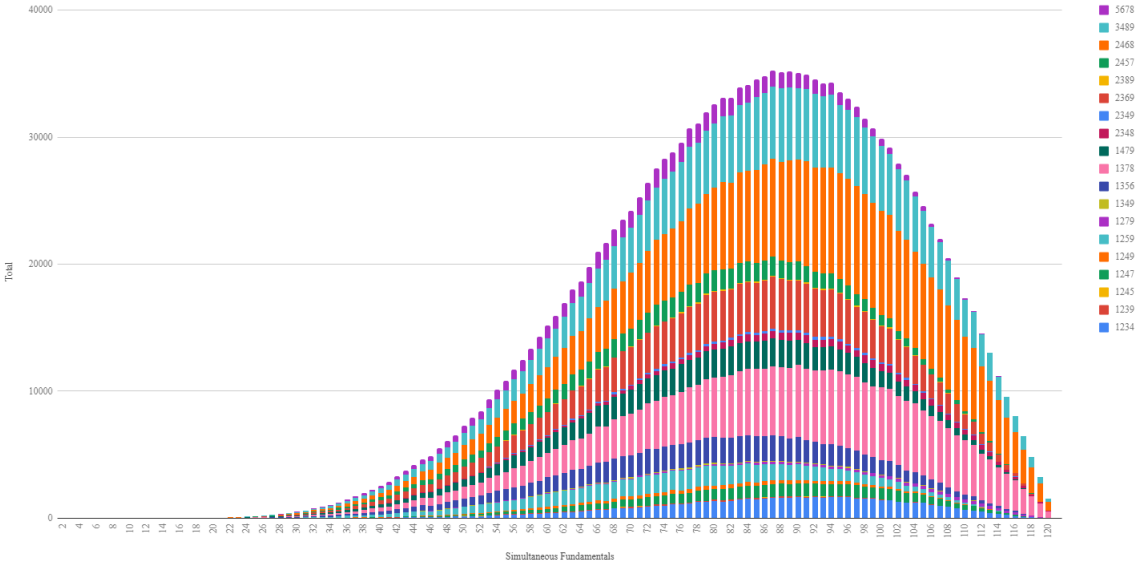


Figure 5.12: Breakdown of terminal vertex sets with cardinality 4 against number of simultaneous fundamentals.

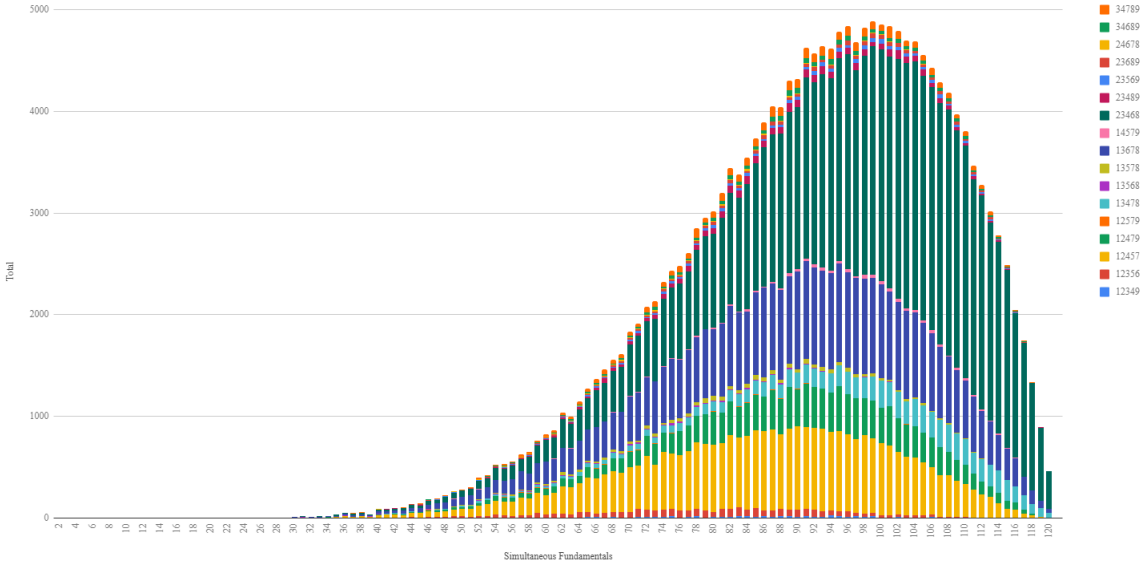


Figure 5.13: Breakdown of terminal vertex sets with cardinality 5 against number of simultaneous fundamentals.

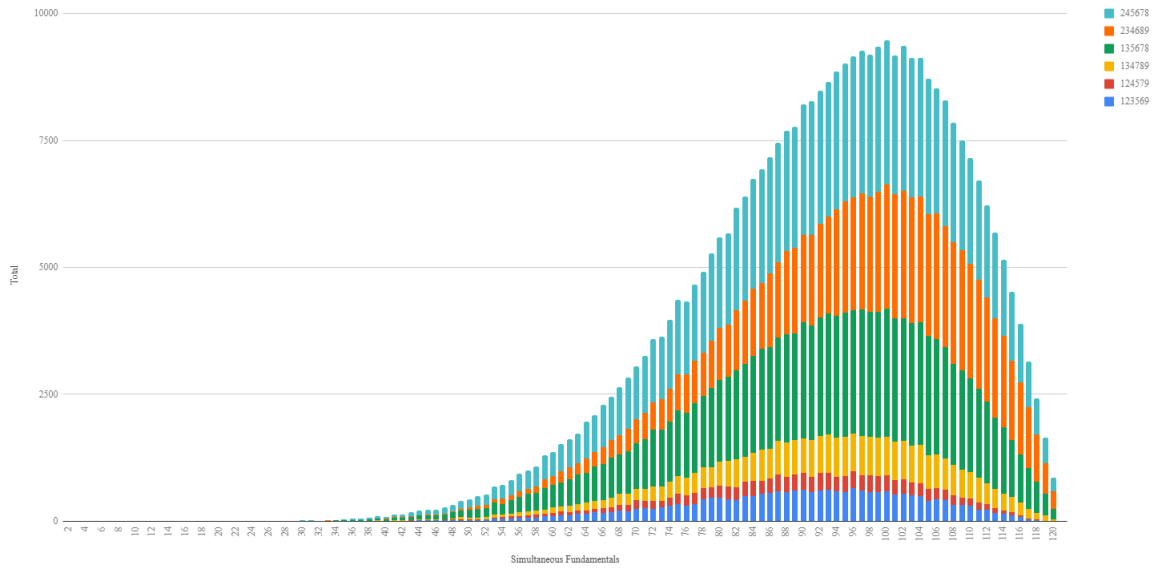


Figure 5.14: Breakdown of terminal vertex sets with cardinality 6 against number of simultaneous fundamentals.

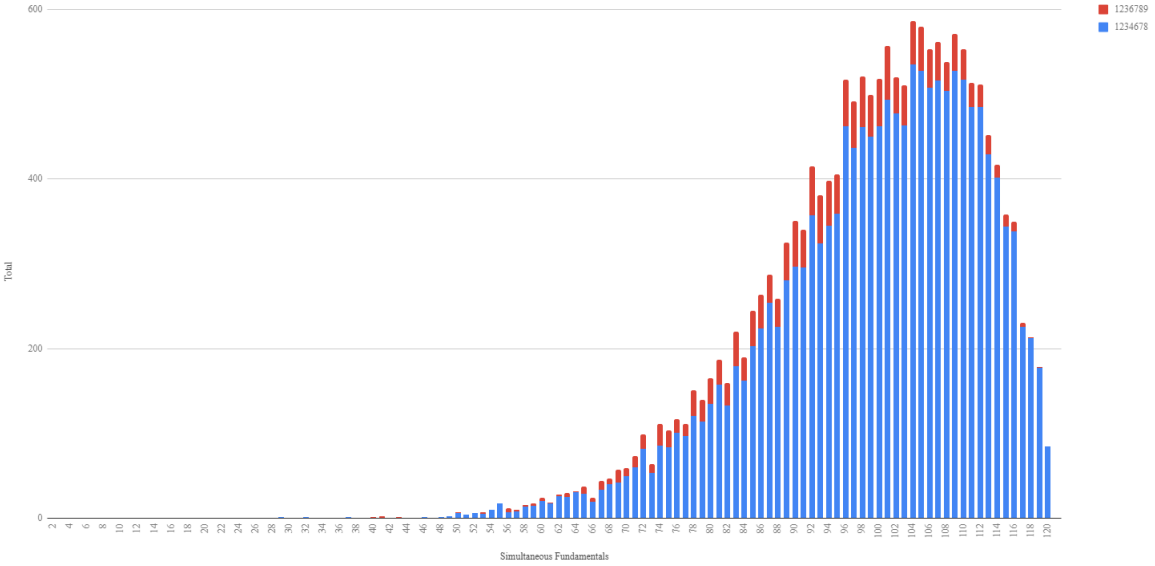


Figure 5.15: Breakdown of terminal vertex sets with cardinality 7 against number of simultaneous fundamentals.

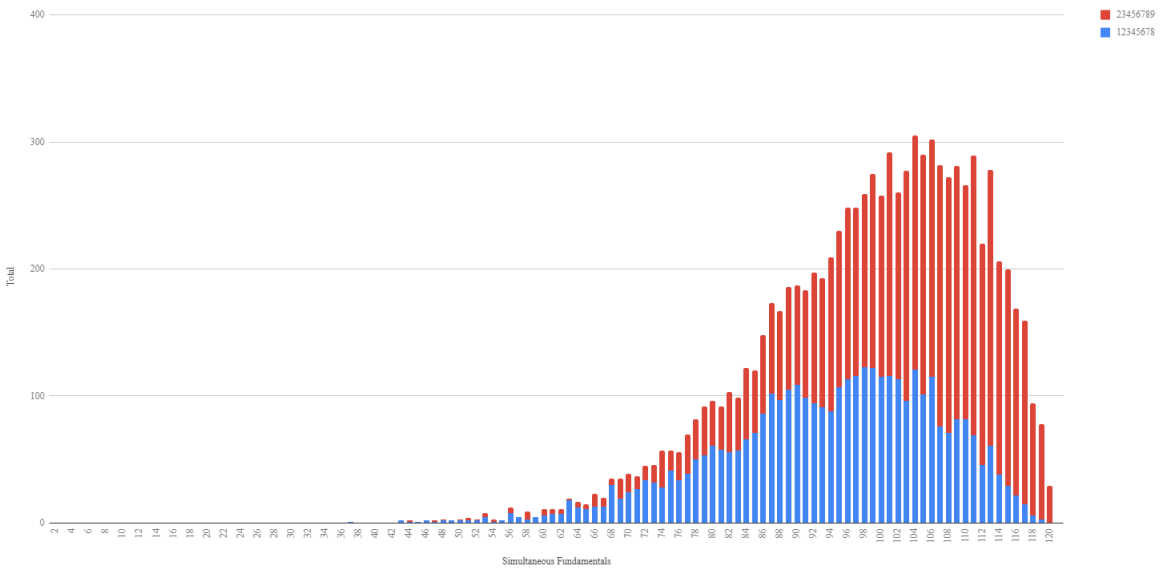


Figure 5.16: Breakdown of terminal vertex sets with cardinality 8 against number of simultaneous fundamentals.

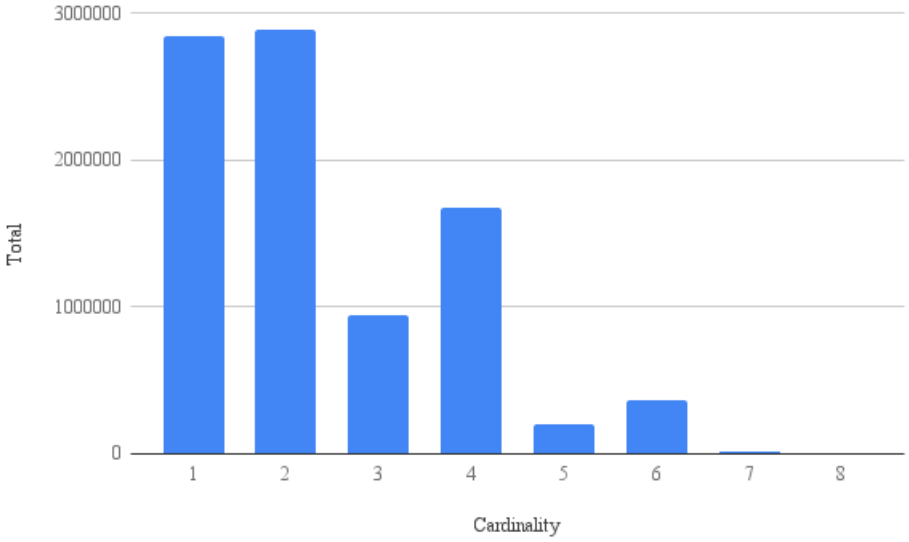


Figure 5.17: Prevalence of graphs with terminal edge set of a given cardinality, averaged across all numbers of simultaneous fundamentals.

with each bit corresponding to the presence (or lack thereof) of a Type in the terminal vertex set. By filtering out all but the cases we care about—for example, those with precisely two terminal vertices—(whilst retaining the order), we reach Figure 5.18.

In trying to unpick at least some of the reasoning behind the lacking cases, it proved useful to build a 'co-occurrence' matrix (Figure 5.19), which helps to elicit patterns behind these gaps. For example, it becomes clear that there are no cardinality 2 graphs with both the irreducible non-basic Type, $\varnothing$, and one of the Types with an edge characteristic (ϵ) of 3. The reasoning behind this may not be intuitively clear, but we can start by assuming the existence of such a terminal edge vertex set—E.g. $\{5, \varnothing\}$. The existence of such a set implies the existence of some vertex in the graph that is (at its simplest) the union of the two types. As shown in Figure 5.20, this necessitates the presence of further basic Types as terminal vertices (beyond 5 and $\varnothing$), which in turn confirms that the aforementioned case can never exist. The same can be shown to be true for Types 6, 7, and 8. A similar look at three terminal vertex sets demonstrates the same phenomenon around $\varnothing$ as Figure 5.19—that is, there are no cardinality three cases containing both $\varnothing$, and one of Types $5 - 8$.

Figure 5.21 shows the proportion that each of the 92 possible cases makes of the total observed during the simulations. Particularly interestingly, the cardinality one cases make up roughly a third of the total, as do the cardinality two cases. To some extent, the former statement means that it may be possible to further-restrict the constraints devised for the occurrence of false fundamentals—for example, by combining this knowledge with the aforementioned, that a lower number of simultaneous fundamentals roughly corresponds to these lower cardinality cases. To this end, and indeed, observing Figure 5.10, it appears that almost all cases dealing with a lower number of simultaneous fundamentals reduce almost exclusively to a single terminal vertex. Thus, it may be the case that in addition to being able to further-restrict our characterisation of false fundamentals, that a parallel of the λ-calculus' Church-Rosser theorem [27] could be applied to these graphs. This has

	Ø	8	7	6	5	4	3	2	1
12	0	0	0	0	0	0	0	1	1
13	0	0	0	0	0	0	1	0	1
23	0	0	0	0	0	0	1	1	0
14	0	0	0	0	0	1	0	0	1
24	0	0	0	0	0	1	0	1	0
34	0	0	0	0	0	1	1	0	0
15	0	0	0	0	1	0	0	0	1
25	0	0	0	0	1	0	0	1	0
35	0	0	0	0	1	0	1	0	0
45	0	0	0	0	1	1	0	0	0
16	0	0	0	1	0	0	0	0	1
26	0	0	0	1	0	0	0	1	0
36	0	0	0	1	0	0	1	0	0
46	0	0	0	1	0	1	0	0	0
56	0	0	0	1	1	0	0	0	0
17	0	0	1	0	0	0	0	0	1
27	0	0	1	0	0	0	0	1	0
37	0		1	0	0	0	1	0	0
47	0	0	1	0	0	1	0	0	0
57	0	0	1	0	1	0	0	0	0
67	0	0	1	1	0	0	0	0	0
18	0	1	0	0	0	0	0	0	1
28	0	1	0	0	0	0	0	1	0
38	0	1	0	0	0	0	1	0	0
48	0	1	0	0	0	1	0	0	0
58	0	1	0	0	1	0	0	0	0
68	0	1	0	1	0	0	0	0	0
78	0	1	1	0	0	0	0	0	0
19	1	0	0	0	0	0	0	0	1
29	1	0	0	0	0	0	0	1	0
39	1	0	0	0	0	0	1	0	0
49	1	0	0	0	0	1	0	0	0
59	1	0	0	0	1	0	0	0	0
69	1	0	0	1	0	0	0	0	0
79	1	0	1	0	0	0	0	0	0
89	1	1	0	0	0	0	0	0	0

Figure 5.18: Lexicographical enumeration of the $\sum_{n=1}^{8} n = 36$ possible two-terminal vertex graphs. Greyed-out rows corresponds to cases that do not occur.

	1	2	3	4	5	6	7	8	Ø
1		1	1	1	1	0	1	0	1
2	1		1	1	1	1	0	0	1
3	1	1		1	0	1	0	1	1
4	1	1	1		0	0	1	1	1
5	1	1	0	0		1	1	0	0
6	0	1	1	0	1		0	1	0
7	1	0	0	1	1	0		1	0
8	0	0	1	1	0	1	1		0
Ø	1	1	1	1	0	0	0	0	

Figure 5.19: 'Co-occurrence' matrix for sets with cardinality 2, showing the valid (and invalid) combinations of Types. Note in particular the lack of cases with both Type ∅, and Types 5, 6, 7, or 8.

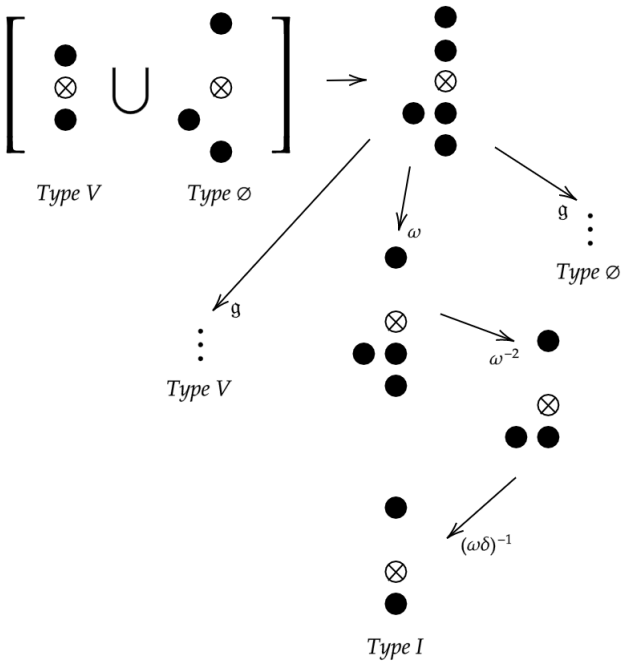


Figure 5.20: Partial reduction graph for the reduction of the union of Types V and ∅, demonstrating the appearance of a further basic edge Type as a terminal vertex. Note that the vertical dots after the reduction by an arbitrary generator, $\mathfrak{g}$, are used to denote that some number of reduction steps lead to the indicated Type.

the potential to speed up processing of reductions, as it would no longer be necessary to enumerate whole graphs to determine the terminal vertex.

Of the graphs with a cardinality one terminal vertex set, Types $1-4$ make up 74.5% of all cases on average, with Type $\varnothing$ making up a further 11.3%. The remaining Types (all with Edge Characteristic 3) make up a small minority of the remaining graphs. Though these cardinality one cases are ostensibly the most common on average, Figure 5.22 shows that as the number of simultaneous fundamentals increases, they become less common—making up less than half of all cases by 49 simultaneous fundamentals, and tailing off past that point. In fact, graphs with four terminal vertices become the prevailing case past around 87 fundamentals. This likely corresponds to the aforementioned phenomenon around the Types with Edge Characteristic 3, and Type $\varnothing$, as these Types become increasingly common with greater numbers of fundamentals (Figure 5.3), and their co-occurrence with Type $\varnothing$ as a terminal vertex appears to necessitate the existence of at least two further terminal vertices (taking the minimum to four).

Similar conclusions around the most common terminal vertex sets can be drawn from the relevant stacked bar charts, in much the same way as for Figure 5.9. Of particular note, the $\{2, 3, 4, 6, 8\}$ case (Figure 5.13) is by far the most prevalent with cardinality five, which likely because it is the case with the most Types of Edge Characteristic $\epsilon = 3$. This arises from the fact that a cardinality five terminal vertex set of the form $\{1, 2, 3, 4, X\}$ (where X is an arbitrary basic Type) necessitates that $X = \varnothing$, as the overlap of the generators in the unified case (i.e. $1 \cup 2 \cup 3 \cup 4$) always results in a possible reduction to the irreducible non-basic case. It is worth finally noting that the prevalences shown in Figures 5.15 and 5.16 are significantly lower in incidence than all of the other cases—which is especially well-highlighted by Figure 5.22, where each of the two presents as a meagre sliver of the total proportion.

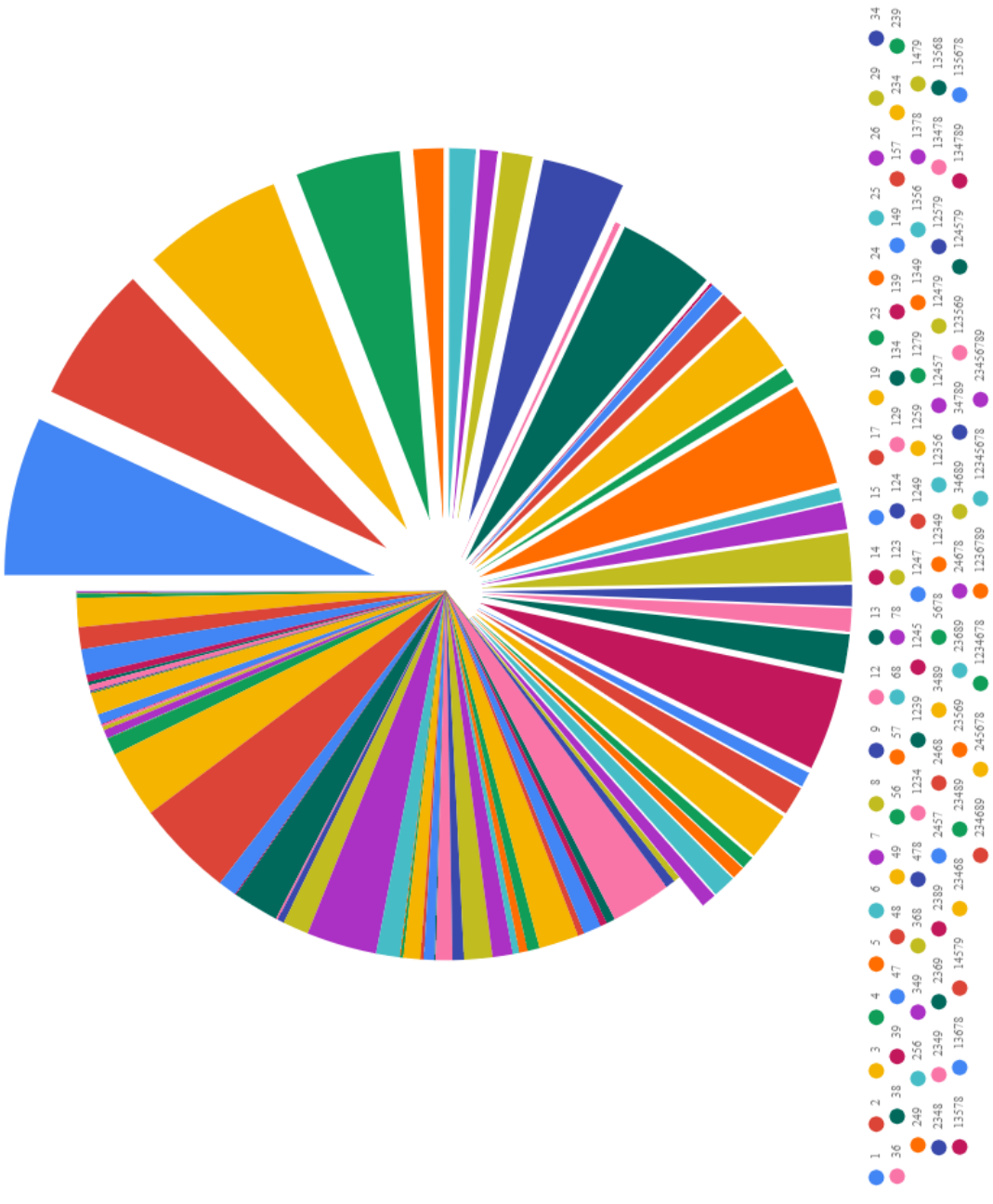


Figure 5.21: Prevalence of all 92 terminal vertex sets, with the cardinality one cases floated out furthest from the centre, and the cardinality two cases floated out half way between these and the rest of the cases. Note that slices are labelled ascending in a clockwise fashion.

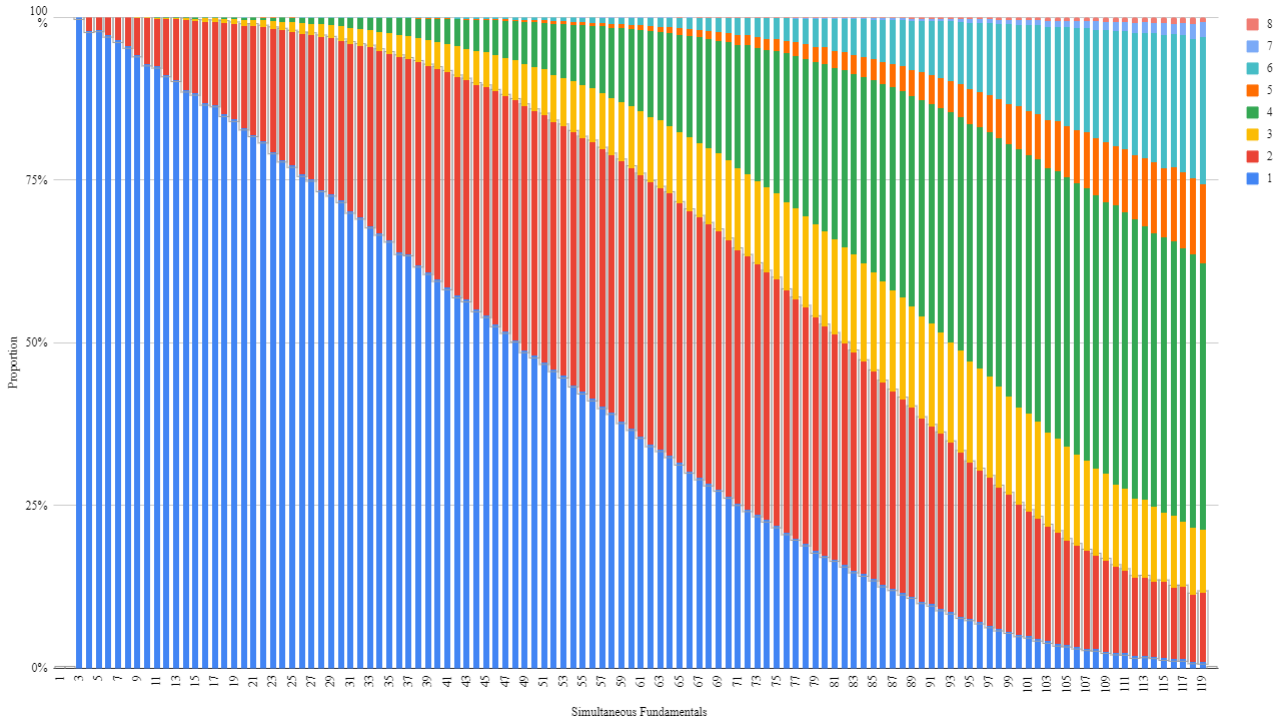


Figure 5.22: Proportion of graphs with terminal vertex sets of varying cardinality as the number of simultaneous fundamentals is varied.

5.2 The Model on Real Data

Though this model is useful theoretically, in practice, real-world applications are rarely so clear-cut or clean - and will remain so unless there exists some perfect approach to noise removal, amongst other preprocessing. Hence, it is prudent to look not at the discrete, but at the continuous in intepretations, $\mathcal{I}$ - i.e. $\mathcal{I} : \mathcal{N}^{\mathbb{V}} \to \mathbb{B}$ becomes $\mathcal{I} : \mathcal{N}^{\mathbb{V}} \to \mathbb{R}$.

Doing so effectively creates a heatmap, in which this additional dimension (perpendicular to $\mathcal{N}^{\mathbb{V}}$) represents an intensity of each tone - for example, their respective amplitudes in the frequency domain [1]. Even in this kind of construction, however, the $\vdash/\Gamma$ shapes are very much still prominent - as demonstrated when applied to some monophonic signals from the University of Iowa (Electronic Music Studios) [63] (Figure 5.23). Here, the intensity is

[1] This construction can be viewed as a real rank 1 trivial vector bundle, with $\mathcal{N}^{\mathbb{V}}$ as a base manifold with trivial topology. In this interpretation, a heatmap is a slice through the topological bundle.

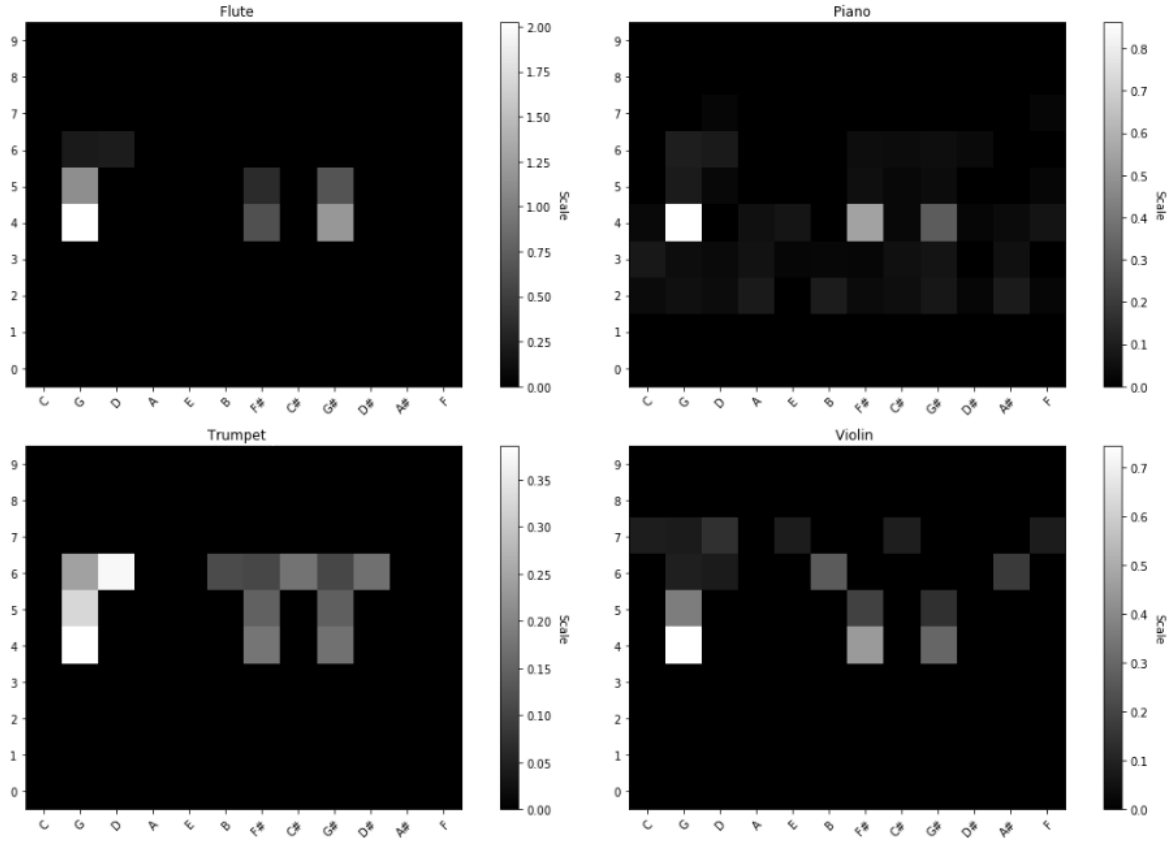


Figure 5.23: The tone G4 being played on a variety of instruments, all exhibiting the Γ shape described in Section 4.3. For flute, trumpet, and violin, the sample was taken from the stationary period, whereas the Piano sample was from part-way through the onset, as this resulted in a clearer image

.

visualised through brightness, with brighter tones representing more prominent frequencies. Though timbrally very different, all of the instruments shown clearly exhibit the Γ shape as anticipated.

Despite this, there are clear differences in the prominence of these shapes between the various instruments. Though flute and trumpet exhibit exceptionally clean examples, the clarity in the piano and violin heatmaps is - whilst still interpretable - somewhat diminished. This is likely a result of multiple media (in the case of piano and violin, strings) vibrating in sympathy to the true fundamental - particularly given that the strings are housed in a shared body. Further, the resonance of this body may also have contributed to the noise.

To build these models, sliding windows were taken from the signal, with a length of 4096 samples, and a hop size of 1024. A constant-Q transform [21] with a Hanning window [134] (using the Librosa implementation [109]) was then applied to achieve a frequency domain representation binned by the 120 semitones of the western musical scale between C0 and B9 inclusive. These values were then normalised across the signal (not just per window), and plotted as a heatmap using matplotlib [82]. Further, each window used here corresponds to a unique interpretation, $\mathcal{I}_\tau$, where τ is the start time index of the window.

It is worth noting that the shapes that appear to mirror the fundamentals and their harmonics in chromatically adjacent columns (i.e. F♯ and G♯ in the case of Figure 5.23) are believed to be a result of spectral leakage, which has not been entirely nullified by the Hanning window. In practice, this could likely be removed, or otherwise accounted for in specific algorithms and approaches.

Looking further, at the three-dimensional heatmap described in Section 4.3 (with each $\mathcal{I}$ indexed as $\mathcal{I}_{[\tau,\tau+1)}$), Figure 5.24 is achieved. A projection onto the $\mathbb{Z}_{12} \times \tau$ plane produces piano-roll notation as expected. Algorithms working in this space may be able to smooth the estimation in the temporal domain by better-exploiting the temporal aspects of music; it is

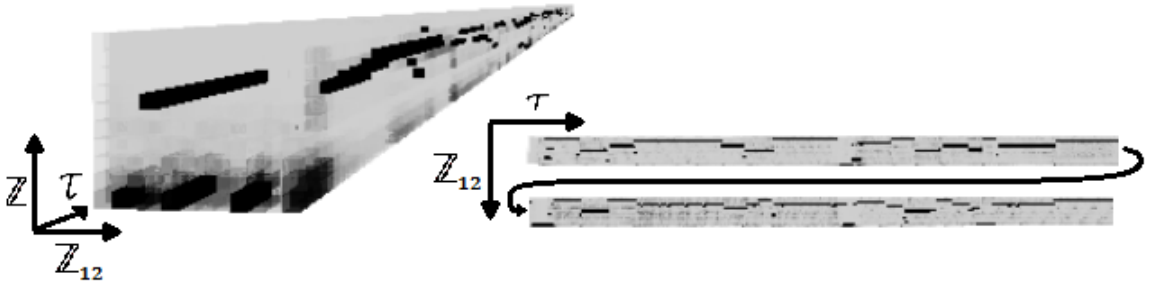


Figure 5.24: Left: Side-on view of a 3D heatmap of the melody of Bach's "Ach Gott und Herr", from the Bach10 data set [45], with darker colours corresponding to greater amplitudes. Right: Projection of the heatmap onto the $\mathbb{Z}_{12} \times \tau$ plane, eliciting piano roll notation of the piece (albeit ordered by the circle of fifths, and not chromatically).

certainly a great oversimplification to treat each window (and therefore each interpretation) as independent of one another.

This was achieved using Python's vpython [144] module, representing each tone as a black (#ffffff) cube, each with opacity proportional to the loudest tone within each specific interpretation[2],

$$Opacity_\nu = \frac{|\nu|}{\mathcal{I}_{MAX}} + 0.05, \tag{5.1}$$

with slight linear scaling to make each individual tone stand out better. Note that $|\nu|$ here refers to the amplitude of a given tone, ν, and $\mathcal{I}_{MAX}$ refers to the largest amplitudes in a given interpretation. It is worth noting that an approach using varying shades of grey as opposed to solid black yielded less optimal results.

Figure 5.25 shows the notable differences between heatmaps constructed from windows from the onset, stationary period, and decay of a single tone. As expected, the shapes are clearest during the stationary period, but in general this raises the more profound issue of choosing an appropriate window, or windows, when given chunks of a signal - such as following onset detection. A simple yet empirically effective heuristic that was found, is to consider both the total number of bins filled above some threshold α (e.g. 3.25μ [67], where

[2]as opposed to in Figure 5.23, where amplitudes are normalised across the whole signal

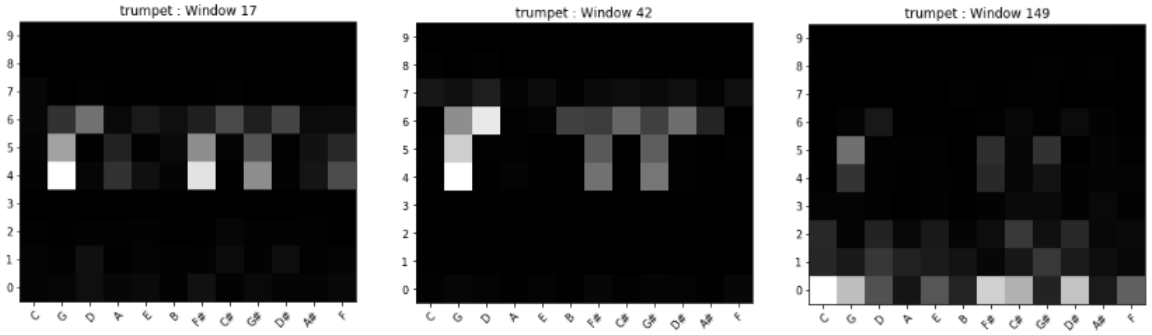


Figure 5.25: A closer look at how heatmaps differ as the tone progresses from onset/attack, to stationary period, to offset/decay (left-to-right).

μ is the mean amplitude of a tone in the window), and the total magnitude of all bins above this threshold in a given window. That is,

$$\frac{\sum_{\nu \in N} \mathcal{I}(\nu)}{|N|}, \; N = \{\nu \mid \mathcal{I}(\nu) \geq \alpha, \; \nu \in \mathcal{N}_{\alpha}^{\Sigma}\}; \tag{5.2}$$

effectively the average amplitude of an audible tone in the window described by $\mathcal{I}$. Doubtless there are more sophisticated approaches, but this serves its purpose if nothing else but as a benchmark. Should time efficiency not be of particular concern, of course, it may be optimal to consider all windows in a chunk (taking their average result) - only discarding a handful of particularly noisy or otherwise useless ones.

Figure 5.26 depicts an edge case built up of the tones D4, A5, and D6, all played on trumpet - exhibiting the special case, $\varnothing$. Though masked somewhat by the spectral leakage on the right-hand side of the image, it is clear to see how such edge cases fool the naive algorithm, even when a threshold is utilised to cut out noise. Of course, the use of Algorithm 2 alleviates this by attempting to remove inharmonic noise, but in doing so, may cause false negatives to arise. Note that when using Algorithm 2 on real-world data, the same switch from $\mathbb{B}$ to $\mathbb{R}$ applies. Appendix A lists the required modifications to the algorithm.

To help the reader better-understand the 'pipeline' at work, Figure 5.27 shows the

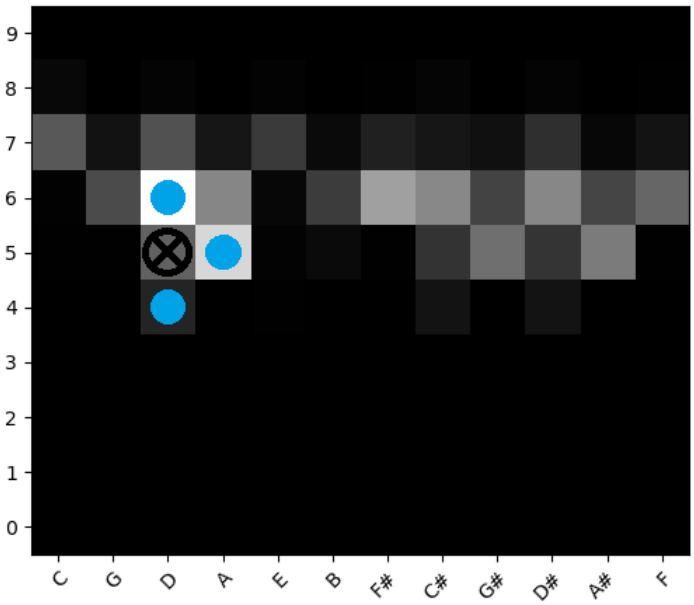


Figure 5.26: An edge case (specifically ∅) being exhibited on real data (trumpet) - with D4, A5, and D6 (dots) as the fundamentals, and D5 being the false fundamental, ⊗.

path from waveform to heatmap visually, with some randomly chosen (and ∴ somewhat noisy) windows. In a further attempt to somewhat ground the work of this Thesis back in the underpinning musical context, Figure 5.28 presents the score for the beginning of the first movement of Bach's Brandenburg Concerto number 2, as well as highlighting the specific presentation of the trumpet's trill of C in the third and fourth bars.

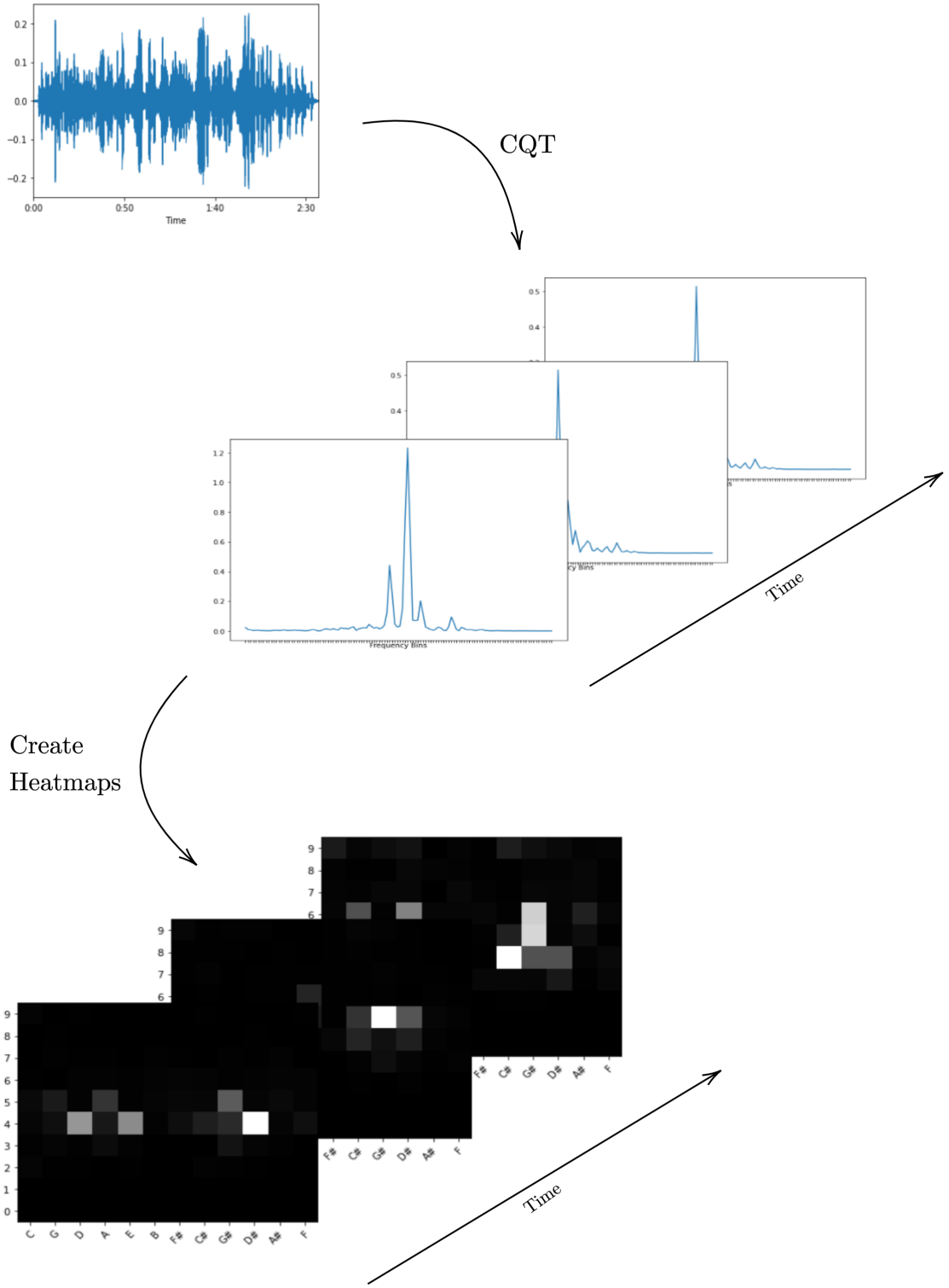


Figure 5.27: Example of the path 'real' data takes from waveform (i.e. time domain signal) to frequency domain signal, and then to heatmap. The data used is from Bach's Prelude and Fugue in C-sharp minor, BWV 849.

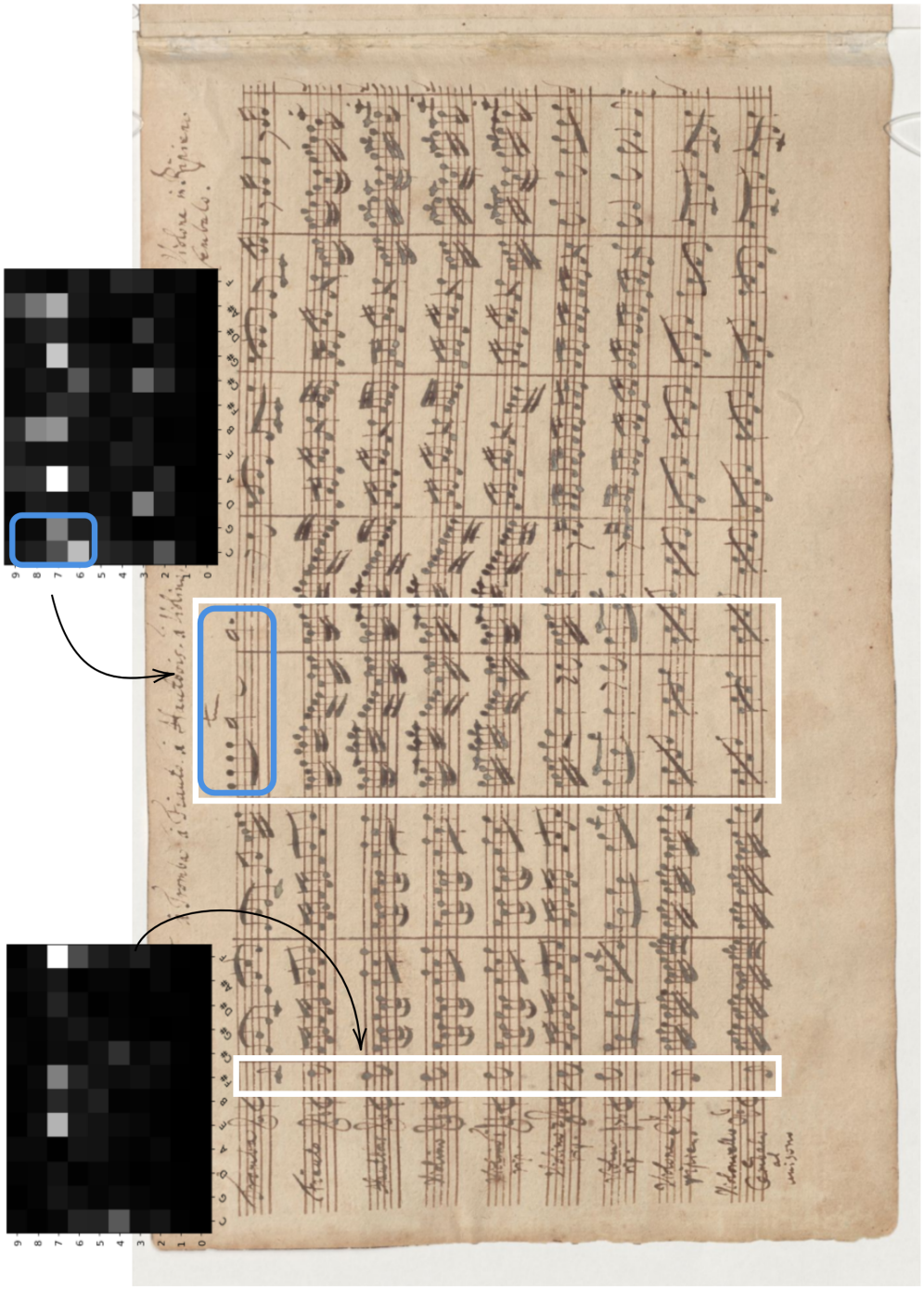


Figure 5.28: Sheet music from the first movement of Bach's Brandenburg Concerto II in F Major, and corresponding heatmaps for some parts. Note the blue circles correspond to the trumpet trill in bars 3 and 4.

5.3 Evaluation of Simple Algorithms

This section presents a brief evaluation of both a naive algorithm on monophonic music, and a more sophisticated (albeit still simplified) algorithm on theoretical polyphonic samples - similar to those utilised in Section 5.1. As noted beforehand, the intention of this Thesis (and investigation on the whole) is not to achieve state of the art results on MPE problems, but rather to lay the foundations for more geometrical approaches to them. Thus, the evaluation is brief, but nonetheless provides insight - particularly surrounding future work.

For monophonic signals, a naive algorithm that treats the relation between fundamentals and $\vdash/\Gamma$-exhibiting tones as an equivalence was used. As shown in Section 4.3, this is untrue due to the presence of edge cases, but nonetheless when only one fundamental is present, such cases can never occur. In response to the spectral leakage, the algorithm was modified slightly to take not only the shape of the potential fundamental and its harmonics into account, but also the corresponding shapes at $\delta^{\pm 5}$.

This was applied to a total of 1395 monophonic samples from the University of Iowa data set, spanning 17 instruments in total (some of which were categorised into vibrato and non-vibrato playing), resulting in a mean accuracy of 73.74%. Removing the outliers (violin/viola/cello/double bass (pizz.), and tuba), this average becomes 88.27%. When considering just whether the pitch chroma is correct (i.e. diregarding octave errors), this increases to 95.08%. Table 5.1 benchmarks this against an implementation of Noll's Harmonic Product Spectrum (HPS) algorithm, also using a Hanning window[3].

A table containing a full breakdown of results, broken down by instrument (and vibrato/non-vibrato playing), can be found in Appendix B.

[3]For more information on the HPS algorithm, or the data used, see Sections **??** and 5.2 respectively. In addition, note that the raw data is converted into interpretations by way of Algorithm 2.

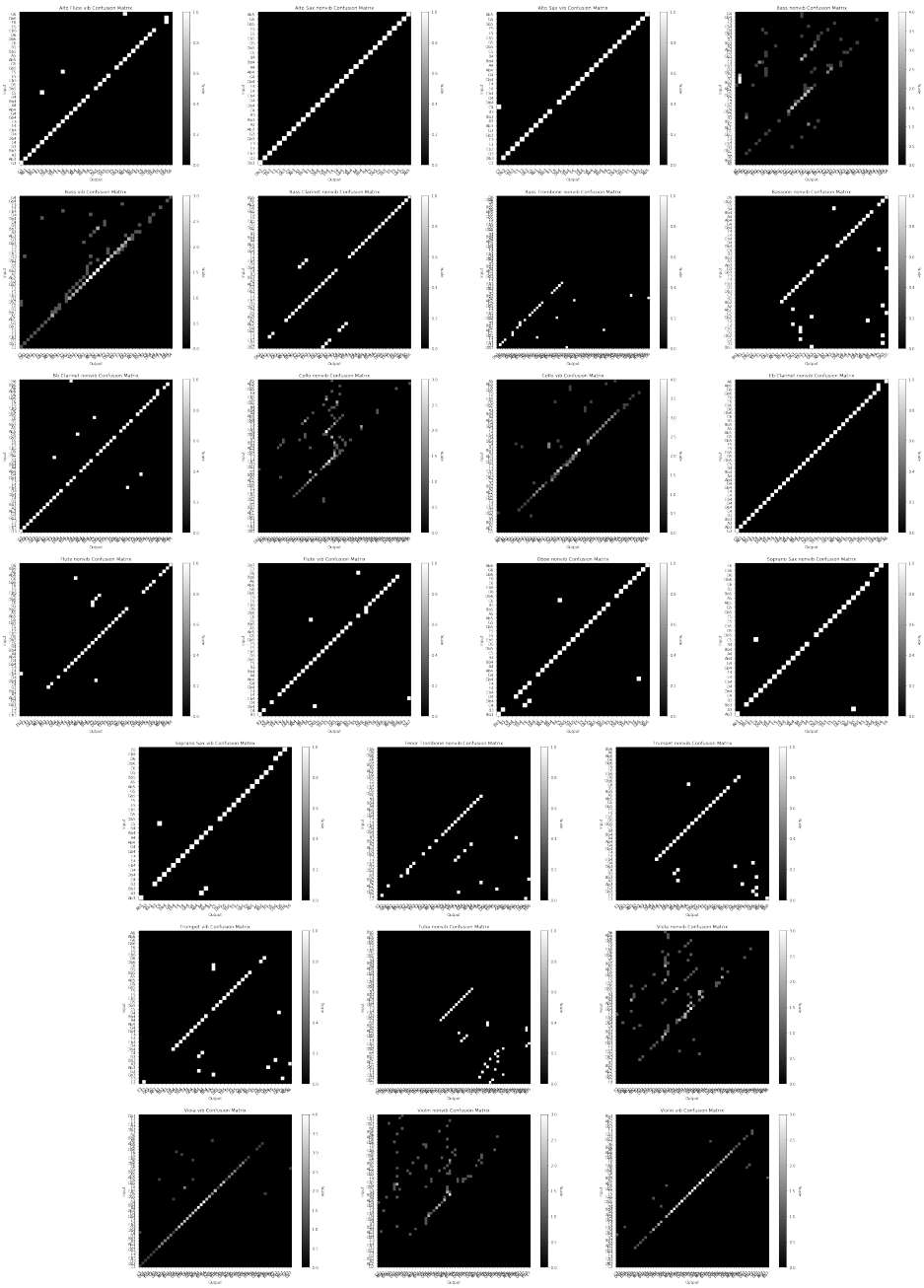


Figure 5.29: Confusion matrices for each set of samples. From top-left to bottom-right: **1**) Alto Flute (vib.); **2**) Alto Sax (non-vib.); **3**) Alto Sax (vib.); **4**) Bass (pizz. non-vib); **5**) Bass (arco, vib); **6**) Bass Clarinet (non-vib.); **7**) Bass Trombone (non-vib.); **8**) Bassoon (non-vib.); **9**) B♭ Clarinet (non-vib.); **10**) Cello (pizz. non-vib); **11**) Cello (arco, vib); **12**) E♭ Clarinet (non-vib.); **13**) Flute (non-vib.); **14**) Flute (vib.); **15**) Oboe (non-vib.); **16**) Soprano Sax (non-vib.); **17**) Soprano Sax (vib.); **18**) Tenor Trombone (non-vib.); **19**) Trumpet (non-vib.); **20**) Trumpet (vib.); **21**) Tuba (non-vib.); **22**) Viola (pizz. non-vib); **23**) Viola (arco, vib); **24**) Violin (pizz. non-vib); **25**) Violin (arco, vib).

	HPS	**Naive**
Overall	58.36%	73.74%
No Outliers	67.73%	88.27%
Chroma Accuracy	77.61%	95.08%

Table 5.1: Table showing the average accuracy of both the naive algorithm and the HPS algorithm as a benchmark when applied to the University of Iowa samples.

Though, as expected, this approach does not reach state of the art results, it still outperforms HPS by a significant margin. Figure 5.29 consists of confusion matrices for each of the instruments (and playing types), showing the algorithm's input (vertical axis) against its output (horizontal axis). Thus, the diagonal is indicative of perfect accuracy, and deviations from this line correspond to errors in the classification. Note that the axes are truncated to match the range of tones tested on each instrument, with the vertical axis running chromatically upwards from bottom to top, and the horizontal axis running chromatically upwards from left to right. Even at first glance, the outliers are relatively clear, and this kind of visualisation has the potential to elicit more profound understanding of how and where an algorithm is failing, and perhaps even (by extrapolation) particular properties of certain instruments that make them more troublesome for pitch detection approaches.

Further, for polyphonic input, a simple extension to the naive monophonic approach, whereby $\mathcal{N}_{\alpha}^{\nabla}$ is traversed left-to-right, bottom-to-top, was utilised[4]. This exploited the assumption that - at least with acoustic music - there will be no undertones. Thus, the bottom-leftmost tone with amplitude above some cutoff will always be a fundamental [67]. The naive algorithm is then applied to subsequent tones, with the following extension: for each potential fundamental, the possible generators for each of its harmonics are enumerated and checked against the current list of perceived fundamentals (i.e. those that have already

[4]It should be noted here, as below, that the polyphonic algorithm was tested on random interpretations, as in Section 5.1. One major drawback of this evaluation is that it likely produces less edge cases than in, for example, consonant music - in which it would be expected that fundamentals would be clustered somewhat closer to one another in $\mathcal{N}^{\nabla}$. In the future, this could be addressed by generating more realistic data, or indeed by utilising polyphonic data sets such as Bach 10.

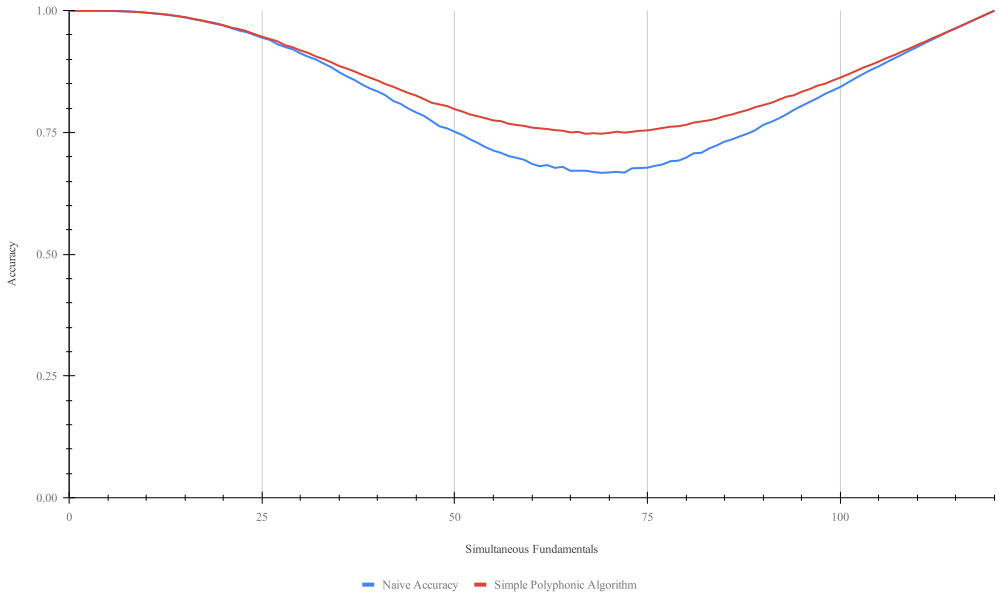


Figure 5.30: Simulated accuracy for a naive approach (blue) (as described in Section 5.1), and simple algorithm (red) when applied to sample polyphonic data.

been classified as such by the algorithm), and if two or more (of a maximum four) of the fundamental and its generators have one or more harmonics that have already been classified as a fundamental, the tone is considered to be a false fundamental, and is discarded. This choice of threshold here may seem somewhat arbitrary, but was chosen as there are a significant number of generators that lie above or to the right of the tone being classified - most notably the harmonics themselves. Thus, whilst they may themselves be fundamentals, it is unclear at this stage of the algorithm. This choice was then tested empirically, with a value of two (from choices [1, 4]) resulting in the best performance.

As in Section 5.1, 1000 sample interpretations were taken for each number of simultaneous unique fundamentals. The accuracy of this simple approach is benchmarked against the accuracy of the naive approach in Figure 5.30. Though there is a clear increase in accuracy, it is anticipated that it will be possible to build on this simplistic approach using the analysis and techniques outlined in Sections 4.4, 4.5, and 5.1, but the implementation of this is beyond the scope of this Thesis.

Off of the back of this short evaluation of algorithms on $\mathcal{N}^{\mathbb{V}}$, it seems sensible to highlight two broadly useful points that will hopefully aid in the development of future work. Firstly, much in the way that we used simulations of $\mathcal{N}^{\mathbb{V}}$ to provide a heuristic for the performance of new algorithms (in our case, taking into account how many already-classified generators *could* be responsible for the 'current' note), it should be possible to do the same for *any* new algorithm utilising the model. This could speed up development of such approaches by facilitating the type of comparison made in Figure 5.30, and additionally removing the need to wrangle large and often hard to manipulate datasets in the preliminary stages of development. In addition, we found that confusion matrices such as those in Figure 5.29 quite nicely demonstrated the *kind* of errors that approaches were facing. For example, looking at the panel for B♭ clarinet, the errors fall along lines corresponding to ± 1 octave. Similarly for trumpet, almost all errors occur below a certain frequency—i.e. D4. Perhaps this corresponds to a change in the timbre or quality of the notes played by the trumpet below this point, which the simple algorithm was struggling to handle.

Chapter Six

Future Work and Conclusion

6.1 Future Work

Building on the work presented, there are a number of relatively natural extensions, additions, and broad applications of the framework that I believe it would be beneficial to consider in the future.

Perhaps the most pertinent of these would be the development of a method to convert real signals into the required Boolean representation effectively, such that the properties of the framework could be used to perform the decomposition-based reformulation of MPE on real-world signals. It may well be the case that what is really necessitated to this end is a 'meta-method' for building such methods in specific domains (e.g. string quartet in a room with known acoustics etc.). As alluded to previously, developing such a technique is likely a significant undertaking, but represents the opportunity to much more concretely ground the more abstract aspects of the work of this Thesis in practical application.

In addition to this, it would also be beneficial to develop algorithms that work on $\mathcal{N}^{\mathbb{V}}$ (with Boolean interpretations), but utilise in particular the characterisation of false

fundamentals given in Chapters 4 and 5 to discard them from the final classification. At a base level, such algorithms could have simple rule-based decisions akin to the extension of the naive algorithm given in Section 5.3, but there are no doubt much more sophisticated ways in which to incorporate the additional knowledge.

In order to facilitate faster and potentially more *general* development of such algorithms, it may be worthwhile to further investigate the testing of them using simulated data as in Section 5.3. There are two points for improvement in particular – firstly, an evaluation of how well the results of this kind of approach actually correlate with performance on real signals, and secondly, the creation and use of artificial data that better-emulates musical signals.

Looking more closely at the latter of the two improvements to testing, it should be noted that the completely random approach currently used to generate artificial test data likely results in inflated performance (of the algorithms being testing) than would be expected on real data, or even on dummy data that better-mimics acoustic musical signals. This is largely due to the fact that fundamentals randomly placed are much more likely to be sparse over $\mathcal{N}^{\mathbb{V}}$. Though regarding octaves, it may make sense for there to be a range within one window of a piece of music (e.g. points with a high melody line and bass line simultaneously), but when considering sparsity round the circle of fifths, there is somewhat of an implication (though not in totality) that notes are unrelated or potentially dissonant. For example, the semitones either side of a tone, ν, sit at $\delta^{\pm 5}$ – almost as far away from the tone itself as is possible, yet these are also some of the most dissonant when sounding simultaneously with ν (particularly in the same octave). Even a very light model which necessitated some more structured movement from note to note (and interpretation to interpretation) could greatly improve how realistic the data is, and it is assumed that this would further make the testing more representative of the algorithms' efficacy on real-world data.

From the perspective of building algorithms to identify candidate fundamentals (i.e. potential fundamentals which need to be analysed and classified) in $\mathcal{N}^{\forall}$ with Real interpretations (and therefore useful if the aforementioned method to translate from Real to Boolean representations is not defined), two further extensions come to mind. One avenue of exploration would be to look to identify $\vdash$ and Γ shaped prisms from $\mathcal{N}^{\forall} \times \tau$, as opposed to $\vdash$ and Γ shapes from $\mathcal{N}^{\forall}$. In addition to utilising the temporal aspects of the signal in this way, it may also be possible to otherwise use time-related data to model higher-order musical structures in the signal, with the intention of using these to identify candidate fundamentals, or to help distinguish between real and false fundamentals – for example, by identifying the melodic line, and placing reasonable musical constraints, such as a maximum 'jump' from the current position, on it, and using these constraints to discount potential candidates that don't conform to the musical intuition encoded by them.

One useful application that it may be possible to construct from the work presented in this Thesis (notably that of Sections 5.1.1 and 5.1.2) would be to build some heuristic to estimate the number of fundamentals in a given signal. I believe that by analysing the types of edge cases present in reduction graphs created from real-world signals (given that at least a rough $\mathbb{R} \to \mathbb{B}$ method can be utilised), it would be possible to compare this to the simulated results presented here, and use this as an estimation of the number of fundamentals at a given point. This could then be used to refine the results of pitch estimation algorithms, by discarding the least likely classified fundamentals where the output set is greater than the heuristic.

Finally, it seems entirely plausible that similar models could be applied to other similar data—for example, PCR sequencing data. Though the specific model used in this Thesis applies to music, the notion of constructing spatially-compact geometric shapes, and remodelling the problem to a classification of edge cases is one that hasn't been widely applied as of yet (to the author's knowledge) to a range of problems.

6.2 Conclusion

As has been reiterated through this Thesis, the aim of this work was not to provide a foolproof solution to the problem of MPE, or indeed yet another algorithm for this purpose, but rather to put forward a general (and largely oversimplified) model for acoustic musical signals, on (and from) which novel approaches can be devised, and iterative improvements and refinements can be made. Further, efforts were made to characterise the occurrence of false fundamentals, $\otimes$, which represent false positives from approaches on $\mathcal{N}^{\forall}$.

What this Thesis does present, however, is a novel approach to pitch estimation that I sincerely hope provides an exciting new viewpoint on the problem, and paves the way for practical applications of the work within. In it, I have provided a model for musical tones that groups a fundamental and its first three harmonics into spatially compact shapes, and reduced the problem of multi-pitch estimation to the problem of distinguishing between edge cases—i.e. false fundamentals, $\otimes$—and real fundamentals—that is, the tones being played. This can alternatively be seen as the decomposition of the 'filled' portions of the grid into potentially overlapping $\vdash$ and Γ shapes.

In order to aid in this distinction, I presented an in-depth theoretical characterisation of precisely when such cases can occur, as well as an analysis of which cases predominantly occur for varying numbers of simultaneous fundamentals. Further, I have provided some insight into the presentation of real-world data in this model, performed basic evaluation on monophonic and polyphonic algorithms using it, and put forward a range of potential avenues for future investigation and research.

I sincerely look forward to seeing where this research may lead in the future.

Appendix One

Creation of an Interpretation

A number of changes must be made for this to work for the reals, $\mathbb{R}$ as opposed to booleans, $\mathbb{B}$:

- An additional parameter, α, is required to represent the minimum amplitude for a note to be considered 'audible';
- Lines 13 and 27 should be replaced by $\mathcal{M}[i,\ j]\ =\ |\nu_{i,j}|$ - setting the amplitude of $\nu_{i,j}$, as opposed to a truth value;
- Finally, lines 12 and 26 should instead be the disjunction or conjunction respectively comparing whether the given harmonics are above the threshold, α, for example, "**if** $|f_1| \geq \alpha \vee |f_2| \geq \alpha \vee |f_3| \geq \alpha$ **then**".

Algorithm 2 Creation of an Interpretation, $\mathcal{I}$, from a Sorted Set of Notes

Input: Φ, a chromatically sorted set of notes
Output: $\mathcal{M}$, a (matrix) interpretation of the notes in Φ

$\mathcal{M} \longleftarrow zeroes(10,\ 12)$
for $\nu_{i,j} \in \Phi$ **do**
  // *Is* $\nu_{i,j}$ *a harmonic of another note?*
  $f_1 \longleftarrow \mathcal{M}[i,\ j-1]$
  $f_2 \longleftarrow 0$
  if $\Psi(\chi_{i-1}) = \vdash$ **then**
    $f_2 \longleftarrow \mathcal{M}[i-1,\ j-1]$
  else
    $f_2 \longleftarrow \mathcal{M}[i-1,\ j-2]$
  end if
  $f_3 \longleftarrow \mathcal{M}[i,\ j-2]$
  if $f_1 \vee f_2 \vee f_3$ **then**
    $\mathcal{M}[i,\ j] = 1$
    continue
  end if

  // *Is* $\nu_{i,j}$ *a potential fundamental?*
  $\phi_1 \longleftarrow \nu_{i,j} \in \Phi$
  $\phi_2 \longleftarrow \bot$
  if $\Psi(\chi_{i-1}) = \vdash$ **then**
    $\phi_2 \longleftarrow \nu_{i+1,\ j+1}$
  else
    $\phi_2 \longleftarrow \nu_{i+1,\ j+2}$
  end if
  $\phi_3 \longleftarrow \nu_{i,\ j+2} \in \Phi$
  if $\phi_1 \wedge \phi_2 \wedge \phi_3$ **then**
    $\mathcal{M}[i,\ j] = 1$
    continue
  end if

  $\Phi \longleftarrow \Phi \setminus \nu_{i,j}$
end for
return $\mathcal{M}$

Appendix Two

Full Results - Naive Algorithm

Instrument	Type	HPS			Naive		
		1	2	3	1	2	3
Alto Flute	Vib	88.89%	88.89%	88.89%	97.22%	97.22%	97.22%
Alto Sax	Nonvib	75.00%	75.00%	81.25%	100.00%	100.00%	100.00%
	Vib	68.75%	68.75%	75.00%	100.00%	100.00%	100.00%
Bass	Pizz Nonvib	20.19%	-	-	53.85%	-	-
	Arco Vib	36.54%	36.54%	39.42%	71.15%	53.85%	72.12%
Bass Clarinet	Nonvib	63.04%	63.04%	65.22%	100.00%	100.00%	100.00%
Bass Trombone	Nonvib	0.00%	0.00%	29.63%	44.44%	44.44%	62.96%
Bassoon	Nonvib	45.00%	45.00%	62.50%	75.00%	75.00%	95.00%
B♭ Clarinet	Nonvib	84.78%	84.78%	84.78%	97.83%	97.83%	97.83%
Cello	Pizz Nonvib	18.00%	-	-	46.00%	-	-
	Arco Vib	65.26%	65.26%	68.42%	88.42%	88.42%	88.42%
E♭ Clarinet	Nonvib	82.05%	82.05%	82.05%	94.87%	94.87%	94.87%
Flute	Nonvib	94.59%	94.59%	94.59%	100.00%	100.00%	100.00%
	Vib	94.59%	94.59%	94.59%	100.00%	100.00%	100.00%
Oboe	Nonvib	77.14%	77.14%	97.14%	100.00%	100.00%	100.00%
Soprano Sax	Nonvib	84.38%	84.38%	87.50%	90.63%	90.63%	90.63%
	Vib	78.13%	78.13%	81.25%	90.63%	90.63%	90.63%
Tenor Trombone	Nonvib	33.33%	33.33%	66.67%	78.79%	78.79%	100.00%
Trumpet	Nonvib	51.43%	51.43%	82.86%	74.29%	74.29%	97.14%
	Vib	51.43%	51.43%	82.86%	74.29%	74.29%	100.00%
Tuba	Nonvib	18.92%	-	-	48.65%	-	-
Viola	Pizz Nonvib	22.00%	-	-	28.00%	-	-
	Arco Vib	88.00%	88.00%	91.00%	100.00%	100.00%	100.00%
Violin	Pizz Nonvib	25.27%	-	-	38.46%	-	-
	Arco Vib	92.22%	92.22%	96.67%	97.78%	97.78%	100.00%

Table B.1: Table showing the performance of the naive algorithm on monophonic samples from the University of Iowa Electronic Music Studios data set, benchmarked against Noll's HPS algorithm. 1, 2, and 3 correspond to the whole data set, sans outliers, and chroma accuracy respectively.

References

[1] 2016. URL: http://hyperphysics.phy-astr.gsu.edu/hbase/Audio/mic.html.

[2] Peter W Alberti. "The anatomy and physiology of the ear and hearing". In: (2001). URL: pdfs.semanticscholar.org/6a5f/0832a948dde736208de5ca02ada86ec6593d.pdf.

[3] Pablo A Alvarado and Dan Stowell. "Efficient learning of harmonic priors for pitch detection in polyphonic music". In: *arXiv preprint arXiv:1705.07104* (2017).

[4] Rafael George Amado and Jozue Vieira Filho. "Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes". In: *2008 International Conference on Audio, Language and Image Processing*. IEEE. 2008, pp. 449–454.

[5] Jessica M Appler and Lisa V Goodrich. "Connecting the ear to the brain: Molecular mechanisms of auditory circuit assembly". In: *Progress in neurobiology* 93.4 (2011), pp. 488–508.

[6] "Applications to Hearing". In: (2015). URL: www.phon.ucl.ac.uk/courses/spsci/AUDL4007/12.pdf.

[7] Milton Babbitt. "The use of computers in musicological research". In: *Perspectives of New Music* (1965), pp. 74–83.

[8] TIBOR BACHMANN and PETER J. BACHMANN. "An Analysis of Béla Bartók's Music Through Fibonaccian Numbers and The Golden Mean". In: *The Musical Quarterly* LXV.1 (Jan. 1979), pp. 72–82. ISSN: 0027-4631. DOI: 10.1093/mq/LXV.1.72.

eprint: https://academic.oup.com/mq/article-pdf/LXV/1/72/9906438/72.pdf. URL: https://doi.org/10.1093/mq/LXV.1.72.

[9] Roland Badeau, Valentin Emiya, and Bertrand David. "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2009, pp. 3073–3076.

[10] Juan Pablo Bello, Laurent Daudet, and Mark B Sandler. "Automatic piano transcription using frequency and time-domain information". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.6 (2006), pp. 2242–2251.

[11] Daniel Bendor and Xiaoqin Wang. "Neural coding of periodicity in marmoset auditory cortex". In: *Journal of neurophysiology* 103.4 (2010), pp. 1809–1822.

[12] Daniel Bendor and Xiaoqin Wang. "The neuronal representation of pitch in primate auditory cortex". In: *Nature* 436.7054 (2005), pp. 1161–1165.

[13] Emmanouil Benetos, Tillman Weyde, et al. "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription". In: (2015).

[14] T. Benjamin. *The Craft of Tonal Counterpoint*. Taylor & Francis, 2004. ISBN: 9781135946623. URL: https://books.google.co.uk/books?id=7Mnfsg2MMXwC.

[15] Tony Bergstrom, Karrie Karahalios, and John Hart. "Isochords: visualizing structure in music." In: Jan. 2007, pp. 297–304. DOI: 10.1145/1268517.1268565.

[16] Rachel M Bittner et al. "Deep salience representations for f_0 estimation in polyphonic music". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference, Suzhou, China*. 2017, pp. 23–27.

[17] Sebastian Böck and Markus Schedl. "Polyphonic piano note transcription with recurrent neural networks". In: *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2012, pp. 121–124.

[18] Albert Bregman. "Hearing musical streams". In: *Computer music journal* 3.4 (1979), pp. 26–43.

[19] Bertrand H Bronson. "Mechanical help in the study of folk song". In: *The Journal of American Folklore* 62.244 (1949), pp. 81–86.

[20] Judith C Brown and Miller S Puckette. "An efficient algorithm for the calculation of a constant Q transform". In: *The Journal of the Acoustical Society of America* 92.5 (1992), pp. 2698–2701.

[21] Judith C. Brown. "Calculation of a constant Q spectral transform". In: *The Journal of the Acoustical Society of America* 89.1 (1991), pp. 425–434. DOI: 10.1121/1.400476. eprint: https://doi.org/10.1121/1.400476. URL: https://doi.org/10.1121/1.400476.

[22] Arturo Camacho and John G Harris. "A sawtooth waveform inspired pitch estimator for speech and music". In: *The Journal of the Acoustical Society of America* 124.3 (2008), pp. 1638–1652.

[23] "Cerebrospinal fluid". In: (2018). URL: https://www.britannica.com/science/cerebrospinal-fluid.

[24] Keunwoo Choi et al. "A tutorial on deep learning for music information retrieval". In: *arXiv preprint arXiv:1709.04396* (2017).

[25] Mads Græsbøll Christensen and Andreas Jakobsson. "Multi-pitch estimation". In: *Synthesis Lectures on Speech & Audio Processing* 5.1 (2009), pp. 1–160.

[26] Y Chunghsin. "Multiple fundamental frequency estimation of polyphonic recordings". PhD thesis. Ph. D. dissertation, University Paris 6, 2008.

[27] Alonzo Church and J Barkley Rosser. "Some properties of conversion". In: *Transactions of the American Mathematical Society* 39.3 (1936), pp. 472–482.

[28] John Clough and Gerald Myerson. "Musical Scales and the Generalized Circle of Fifths". In: *The American Mathematical Monthly* 93.9 (1986), pp. 695–701. DOI: 10.1080/00029890.1986.11971924. eprint: https://doi.org/10.1080/00029890.1986.11971924. URL: https://doi.org/10.1080/00029890.1986.11971924.

[29] Richard Cohn. "Introduction to Neo-Riemannian Theory: A Survey and a Historical Perspective". In: *Journal of Music Theory* 42.2 (1998), pp. 167–180. ISSN: 00222909. URL: http://www.jstor.org/stable/843871.

[30] Richard Cohn. "Neo-Riemannian operations, parsimonious trichords, and their Tonnetz representations". In: *Journal of Music Theory* 41.1 (1997), pp. 1–66.

[31] F Cong et al. "A parallel fusion approach to piano music transcription based on convolutional neural network". In: (2018).

[32] Keith Conrad. *Why is group theory important?* URL: https://kconrad.math.uconn.edu/math216/whygroups.html.

[33] James W. Cooley and John W. Tukey. "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90 (1965), pp. 297–301. ISSN: 00255718, 10886842. URL: http://www.jstor.org/stable/2003354.

[34] Irving B Crandall. "The sounds of speech". In: *The bell system technical journal* 4.4 (1925), pp. 586–639.

[35] Mike Crundell and Geoff Goodwin. *Cambridge International AS and A Level Physics*. 2014. ISBN: 978-1471-80921-7.

[36] A Versatile Polyphonic Music Dataset. "Bach10 Dataset". In: ().

[37] Benjamin Lent Davis and Diane Maclagan. "The card game SET". In: *The Mathematical Intelligencer* 25.3 (2003), pp. 33–40.

[38] Alain De Cheveigné and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1917–1930.

[39] Alain De Cheveigné and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1917–1930.

[40] Patricio De La Cuadra, Aaron S Master, and Craig Sapp. "Efficient Pitch Detection Techniques for Interactive Music." In: *ICMC*. 2001.

[41] Macquarie University Linguistics Department. *Adding Waveforms and Phase*. 2020. URL: https://www.mq.edu.au/about/about-the-university/faculties-and-departments/medicine-and-health-sciences/departments-and-centres/department-of-linguistics/our-research/phonetics-and-phonology/speech/acoustics/speech-waveforms/adding-waveforms-and-phase.

[42] Riyanti Djalante. "Key assessments from the IPCC special report on global warming of 1.5°C and the implications for the Sendai framework for disaster risk reduction". In: *Elsevier Progress in Disaster Science* 1 (2019). DOI: https://doi.org/10.1016/j.pdisas.2019.100001.

[43] J. Dongarra and F. Sullivan. "Guest Editors Introduction to the top 10 algorithms". In: *Computing in Science Engineering* 2.1 (Jan. 2000), pp. 22–23. ISSN: 1521-9615. DOI: 10.1109/MCISE.2000.814652.

[44] J Stephen Downie. "The scientific evaluation of music information retrieval systems: Foundations and future". In: *Computer Music Journal* 28.2 (2004), pp. 12–23.

[45] Z Duan and B Pardo. *Bach10 dataset*. 2015.

[46] Mark Eichenlaub. 2017. URL: https://math.stackexchange.com/q/72479.

[47] Anders Elowsson. "Deep layered learning in MIR". In: *arXiv preprint arXiv:1804.07297* (2018).

[48] Anders Elowsson. "Polyphonic pitch tracking with deep layered learning". In: *The Journal of the Acoustical Society of America* 148.1 (2020), pp. 446–468.

[49] Anders Elowsson and Anders Friberg. "Polyphonic transcription with deep layered learning". In: *MIREX. Retrieved from http://www. music-ir. org/mirex/abstracts/2014/EF1. pdf* (2014).

[50] Valentin Emiya, Roland Badeau, and Bertrand David. "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6 (2009), pp. 1643–1654.

[51] Valentin Emiya, Roland Badeau, and Bertrand David. "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6 (2010), pp. 1643–1654.

[52] Valentin Emiya et al. "MAPS-A piano database for multipitch estimation and automatic transcription of music". In: (2010).

[53] A. Cayley Esq. "VII. On the theory of groups, as depending on the symbolic equation n =1". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 7.42 (1854), pp. 40–47. DOI: 10.1080/14786445408647421. URL: https://doi.org/10.1080/14786445408647421.

[54] Leonhard Euler. "1739. Tentamen novae theoriae musicae ex certissismis harmoniae principiis dilucide expositae". In: *St. Petersburg, also in Opera omnia, ser* 3.1 (1739), pp. 197–427.

[55] Leonhard Euler. *Tentamen novae theoriae musicae ex certissimis harmoniae principiis dilucide expositae*. Ex typographia Academiae scientiarum, 1739.

[56] Brian. Everitt. *The Cambridge dictionary of statistics / B.S. Everitt*. English. 2nd ed. Cambridge University Press Cambridge, U.K. ; New York, 2002, ix, 410 p. : ISBN: 052181099. URL: http://www.loc.gov/catdir/toc/cam031/2002514499.html.

[57] Frank A. Farris. "Wheels on Wheels on Wheels-Surprising Symmetry". In: *Mathematics Magazine* 69.3 (1996), pp. 185–189. ISSN: 0025570X, 19300980. URL: http://www.jstor.org/stable/2691465.

[58] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[59] Richard Fitzpatrick. *Spherical Mirrors*. 2007. URL: farside.ph.utexas.edu/teaching/316/lectures/node136.

[60] *Flute Waveform*. 1997. URL: http://hyperphysics.phy-astr.gsu.edu/hbase/Music/flutew.html.

[61] Allen Forte. "Music and computing: The present situation". In: *Proceedings of the November 14-16, 1967, fall joint computer conference*. 1967, pp. 327–329.

[62] Jean-Baptiste-Joseph Fourier. "Théorie analytique de la chaleur. (French) [The analytical theory of heat]". In: (1822).

[63] Lawrence Fritts. *Musical Instrument Samples*. 2012. URL: http://theremin.music.uiowa.edu/MIS.html.

[64] Joe Futrelle and J Stephen Downie. "Interdisciplinary communities and research issues in Music Information Retrieval." In: *ISMIR*. Vol. 2. 2002, pp. 215–221.

[65] Mateusz Gawlik and Wieslaw Wszolek. "Modern pitch detection methods in singing voices analyzes". In: *Proceedings of Euronoise*. 2018, pp. 247–253.

[66] Robert van Gend. "The Fibonacci sequence and the golden ratio in music". In: ().

[67] Thomas A Goodman and Ian Batten. "Real-Time Polyphonic Pitch Detection on Acoustic Musical Signals". In: *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE. 2018, pp. 1–6.

[68] Tom Goodman, Karoline van Gemst, and Peter Tino. "A Geometric Framework for Pitch Estimation on Acoustic Musical Signals". In: *Journal of Mathematics and Music* (2021). DOI: 10.1080/17459737.2021.1979116.

[69] Richard W. Hamming. *The Art of Probability: For Scientists and Engineers*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1991, p. 64. ISBN: 0-201-51058-8.

[70] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Græsbøll Christensen. "Estimation of fundamental frequencies in stereophonic music mixtures". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.2 (2018), pp. 296–310.

[71] Martin Weiss Hansen, Jesper Rindom Jensen, and Mads Græsbøll Christensen. "Estimation of multiple pitches in stereophonic mixtures using a codebook-based approach". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 186–190.

[72] Fredric J Harris. "On the use of windows for harmonic analysis with the discrete Fourier transform". In: *Proceedings of the IEEE* 66.1 (1978), pp. 51–83.

[73] Joseph E. Hawkins. *Anatomy of the Human Ear*. Oct. 2018. URL: www.britannica.com/science/ear/Anatomy-of-the-human-ear.

[74] Curtis Hawthorne et al. "Onsets and frames: Dual-objective piano transcription". In: *arXiv preprint arXiv:1710.11153* (2017).

[75] Henry Heffner and Ian C Whitfield. "Perception of the missing fundamental by cats". In: *The Journal of the Acoustical Society of America* 59.4 (1976), pp. 915–919.

[76] Hermann LF Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. 1863.

[77] W. Hess. *Pitch Determination of Speech Signals*. 3. Springer-Verlag, 1983, p. 355. ISBN: 978-3-642-81928-5. URL: https://www.springer.com/gb/book/9783642819285.

[78] Wolfgang Hess. *Pitch determination of speech signals: algorithms and devices*. Vol. 3. Springer Science & Business Media, 2012.

[79] Wolfgang J Hess. "Pitch and voicing determination of speech with an extension toward music signals". In: *Springer handbook of speech processing* (2008), pp. 181–212.

[80] William E Hettrick, William E Hettrick, John Butt, et al. *The'Musica instrumentalis deudsch'of Martin Agricola: A treatise on musical instruments, 1529 and 1545*. CUP Archive, 1994.

[81] Nicki Holighaus et al. "A Framework for Invertible, Real-Time Constant-Q Transforms". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.4 (2013), pp. 775–785. DOI: 10.1109/TASL.2012.2234114.

[82] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[83] Jonas Hydman. *Recurrent laryngeal nerve injury*. Institutionen för klinisk neurovetenskap/Department of Clinical Neuroscience, 2008.

[84] Steven G. Johnson and Matteo Frigo. *Implementing FFTs in Practice*. 2012. URL: http://cnx.org/contents/ba55ed41-b37b-45bd-89c9-e6b67725401e@15.

[85] Hirokazu Kameoka, Takuya Nishimoto, and Shigeki Sagayama. "A multipitch analyzer based on harmonic temporal structured clustering". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), pp. 982–994.

[86] Michael Kassler. "Toward musical information retrieval". In: *Perspectives of New Music* (1966), pp. 59–67.

[87] R. Kelz, S. Böck, and G. Widmer. "Deep Polyphonic ADSR Piano Note Transcription". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2019, pp. 246–250. DOI: 10.1109/ICASSP.2019.8683582.

[88] Rainer Kelz, Sebastian Böck, and Gerhard Widmer. "Deep polyphonic adsr piano note transcription". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 246–250.

[89] Anssi Klapuri. "Multipitch analysis of polyphonic music and speech signals using an auditory model". In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.2 (2008), pp. 255–266.

[90] Anssi Klapuri. "Multiple fundamental frequency estimation by summing harmonic amplitudes." In: *ISMIR*. 2006, pp. 216–221.

[91] Anssi Klapuri and Manuel Davy. "Signal processing methods for music transcription". In: (2007).

[92] Anssi P Klapuri. "A perceptually motivated multiple-f0 estimation method". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* IEEE. 2005, pp. 291–294.

[93] Anssi P Klapuri. "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness". In: *IEEE Transactions on speech and audio processing* 11.6 (2003), pp. 804–816.

[94] A Sophia Koepke, Olivia Wiles, and Andrew Zisserman. "Visual pitch estimation". In: (2019).

[95] Sebastian Kraft and Udo Zölzer. "Polyphonic pitch detection by matching spectral and autocorrelation peaks". In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE. 2015, pp. 1301–1305.

[96] Jonathan Kramer. "The Fibonacci series in twentieth-century music". In: *Journal of Music Theory* 17.1 (1973), pp. 110–148.

[97] Neeraj Kumar and Raubin Kumar. "Wavelet transform-based multipitch estimation in polyphonic music". In: *Heliyon* 6.1 (2020), e03243.

[98] Balachandra Kumaraswamy and PG Poonacha. "Modified square difference function using fourier series approximation for pitch estimation". In: *2017 international conference on algorithms, methodology, models and applications in emerging technologies (ICAMMAET)*. IEEE. 2017, pp. 1–8.

[99] Julia Kursell. "Sounds of Science–Schall im Labor (1800–1930)". In: (2008).

[100] P. Ladefoged. *Elements of Acoustic Phonetics*. Elements of Acoustic Phonetics. University of Chicago Press, 1962. ISBN: 9780226467849. URL: https://books.google.co.uk/books?id=M11AAQAAIAAJ.

[101] Charles E Leiserson et al. "There's plenty of room at the Top: What will drive computer performance after Moore's law?" In: *Science* 368.6495 (2020).

[102] David Lewin. *Generalized musical intervals and transformations*. Oxford University Press, USA, 1987.

[103] Y. Li and D. Wang. "Pitch Detection in Polyphonic Music using Instrument Tone Models". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 2. Apr. 2007, pp. II-481-II–484. DOI: 10.1109/ICASSP.2007.366277.

[104] Harry B Lincoln. "The computer and music research: Prospects and problems". In: *Bulletin of the Council for Research in Music Education* (1969), pp. 1–9.

[105] Mathias Lohne. *The Computational Complexity of the Fast Fourier Transform*. 2017. URL: https://folk.uio.no/mathialo/texts/fftcomplexity.pdf.

[106] Christopher Mark. "Britten and the Circle of Fifths". In: *Journal of the Royal Musical Association* 119.2 (1994), pp. 268–297. DOI: 10.1093/jrma/119.2.268.

[107] Matija Marolt. "A connectionist approach to automatic transcription of polyphonic piano music". In: *IEEE Transactions on Multimedia* 6.3 (2004), pp. 439–449.

[108] Philippe Martin. "Comparison of pitch detection by cepstrum and spectral comb analysis". In: *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 7. IEEE. 1982, pp. 180–183.

[109] Brian McFee et al. "librosa: Audio and music signal analysis in python". In: *Proceedings of the 14th python in science conference*. Vol. 8. 2015.

[110] Andrew McLeod and Mark Steedman. "Evaluating Automatic Polyphonic Music Transcription." In: *ISMIR*. 2018, pp. 42–49.

[111] Philip McLeod. "Fast, accurate pitch detection tools for music analysis". In: *Academisch proefschrift, University of Otago. Department of Computer Science* (2009).

[112] Ray Meddis and Michael J Hewitt. "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification". In: *The Journal of the Acoustical Society of America* 89.6 (1991), pp. 2866–2882.

[113] Arthur Mendel. "Some preliminary attempts at computer-assisted style analysis in music". In: *Computers and the Humanities* (1969), pp. 41–52.

[114] John C. Middlebrooks and David M. Green. "Sound Localization by Human Listeners". In: *Annual Review of Psychology* 42.1 (1991). PMID: 2018391, pp. 135–159. DOI: 10.1146/annurev.ps.42.020191.001031.

[115] Clement A Miller. "Gaffurius's" Practica Musicae": Origin and Contents". In: *Musica disciplina* 22 (1968), pp. 105–128.

[116] James S. Milne. *Group Theory (v3.13)*. Available at www.jmilne.org/math/. 2013.

[117] Marius Miron, Julio José Carabias-Orti, and Jordi Janer. "Audio-to-score Alignment at the Note Level for Orchestral Recordings." In: *ISMIR*. 2014, pp. 125–130.

[118] James A Moorer. "On the transcription of musical sound by computer". In: *Computer Music Journal* (1977), pp. 32–38.

[119] James Anderson Moorer. *On the segmentation and analysis of continuous musical sound by digital computer*. Stanford University, 1975.

[120] Iain Morley. "An Investigation into the Prehistory of Human Musical Capacities and Behaviours, Using Archaeological, Anthropological, Cognitive and Behavioural Evidence". PhD thesis. 2003.

[121] Frederick Mosteller and David L Wallace. "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers". In: *Journal of the American Statistical Association* 58.302 (1963), pp. 275–309.

[122] Meinard Muller et al. "Signal processing for music analysis". In: *IEEE Journal of Selected Topics in Signal Processing* 5.6 (2011), pp. 1088–1110.

[123] Multicherry. *Spirogram Set*. Aug. 2014. URL: https://commons.wikimedia.org/wiki/File:Spirograph_set_(UK_Palitoy_early_1980s)_(perspective_fixed).jpg.

[124] Eita Nakamura et al. "Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 101–105.

[125] Jiquan Ngiam et al. "Multimodal deep learning". In: *ICML*. 2011.

[126] Michael A Noll. "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate". In:

Symposium on Computer Processing in Communication, ed. Vol. 19. University of Brooklyn Press, New York. 1969, pp. 779–797.

[127] de Obaldıa and Zölzer. "Improving Monophonic Pitch Detection Using the ACF and Simple Heuristics". In: (2019).

[128] A.V. Oppenheim, R.W. Schafer, and J.R. Buck. *Discrete-time Signal Processing.* Prentice Hall international editions. Prentice Hall, 1999. Chap. 8. ISBN: 9780137549207. URL: https://books.google.co.uk/books?id=Bv1SAAAAMAAJ.

[129] Claude V Palisca. *Scientific empiricism in musical thought.* Princeton University Press, 2015.

[130] Karl Pearson. "Note on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58 (1895), pp. 240–242.

[131] Geoffroy Peeters. "Music pitch representation by periodicity measures based on combined temporal and spectral representations". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings.* Vol. 5. IEEE. 2006, pp. V–V.

[132] *Physics Tutorial: Sound Waves as Pressure Waves.* URL: https://www.physicsclassroom.com/class/sound/u11l1c.cfm.

[133] Gregor Pirker et al. "A pitch tracking corpus with evaluation on multipitch tracking scenario". In: *Twelfth Annual Conference of the International Speech Communication Association.* 2011.

[134] Prajoy Podder et al. "Comparative performance analysis of hamming, hanning and blackman window". In: *International Journal of Computer Applications* 96.18 (2014).

[135] Graham E Poliner and Daniel PW Ellis. "A discriminative model for polyphonic piano transcription". In: *EURASIP Journal on Advances in Signal Processing* 2007 (2006), pp. 1–9.

[136] Lutz Prechelt and Rainer Typke. "An interface for melody input". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 8.2 (2001), pp. 133–149.

[137] Lawrence Rabiner. "On the use of autocorrelation analysis for pitch detection". In: *IEEE transactions on acoustics, speech, and signal processing* 25.1 (1977), pp. 24–33.

[138] Lawrence Rabiner. "On the use of autocorrelation analysis for pitch detection". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.1 (1977), pp. 24–33.

[139] Don Michael Randel. *The Harvard Concise Dictionary of Music and Musicians*. 1999, pp. 105–106. ISBN: 0-674-00084-6.

[140] Xavier Rodet and Boris Doval. "Fundamental frequency estimation using a new harmonic matching method". In: *Proceedings of the International Computer Music Conference*. INTERNATIONAL COMPUTER MUSIC ACCOCIATION. 1991, pp. 555–555.

[141] Miguel A Román, Antonio Pertusa, and Jorge Calvo-Zaragoza. "A holistic approach to polyphonic music transcription with neural networks". In: *arXiv preprint arXiv:1910.12086* (2019).

[142] Stuart Rosen. *Signals and Systems for Speech and Hearing*. 2nd ed. BRILL, 2011, p. 163. ISBN: 978-9004252431.

[143] Matti P Ryynänen and Anssi P Klapuri. "Automatic transcription of melody, bass line, and chords in polyphonic music". In: *Computer Music Journal* 32.3 (2008), pp. 72–86.

[144] David Scherer, Paul Dubois, and Bruce Sherwood. "VPython: 3D interactive scientific graphics for students". In: *Computing in Science & Engineering* 2.5 (2000), pp. 56–62.

[145] James Schloss. *What is a Fast Fourier Transform (FFT)? The Cooley-Tukey Algorithm*. 2017. URL: https://www.youtube.com/watch?v=XtypWS8HZco.

[146] Rodrigo Schramm et al. "Multi-pitch detection and voice assignment for a cappella recordings of multiple singers". In: ISMIR. 2017.

[147] Manfred R Schroeder. "Period histogram and product spectrum: New methods for fundamental-frequency measurement". In: *The Journal of the Acoustical Society of America* 43.4 (1968), pp. 829–834.

[148] Claude E. Shannon. "Communication in the Presence of Noise". In: *Proceedings of the IRE* 37.1 (1949). DOI: 10.1109/JRPROC.1949.232969.

[149] Hagit Shatkay. "The Fourier transform-A primer". In: (1995).

[150] Jose Shelton and Gideon Praveen Kumar. "Comparison between auditory and visual simple reaction times". In: *Neuroscience & Medicine* 1.1 (2010), pp. 30–32.

[151] Francois Signol, Claude Barras, and Jean-Sylvain Liénard. "Evaluation of the pitch estimation algorithms in the monopitch and multipich cases". In: *Acoustics' 08*. 2008.

[152] S. Sigtia, E. Benetos, and S. Dixon. "An End-to-End Neural Network for Polyphonic Piano Music Transcription". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.5 (May 2016), pp. 927–939. ISSN: 2329-9290. DOI: 10.1109/TASLP.2016.2533858.

[153] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. "An end-to-end neural network for polyphonic piano music transcription". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.5 (2016), pp. 927–939.

[154] Paris Smaragdis and Judith C Brown. "Non-negative matrix factorization for polyphonic music transcription". In: *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*. IEEE. 2003, pp. 177–180.

[155] Julius O. Smith. *Spectral Audio Signal Processing*. online book, 2011 edition. 2011. URL: https://ccrma.stanford.edu/$%5Csim$jos/sasp/Hamming%5C_Window.html.

[156] Tamara Smyth. *Harmonic Product Spectrum (HPS)*. 2019. URL: http://musicweb.ucsd.edu/~trsmyth/analysis/Harmonic_Product_Spectrum.html.

[157] Mohan Sondhi. "New methods of pitch extraction". In: *IEEE Transactions on audio and electroacoustics* 16.2 (1968), pp. 262–266.

[158] Ned Steinberger. *Chromatic tuner display providing guitar note and precision tuning information*. US Patent 5,549,028. Aug. 1996.

[159] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *arXiv preprint arXiv:1906.02243* (2019).

[160] David Talkin and W Bastiaan Kleijn. "A robust algorithm for pitch tracking (RAPT)". In: *Speech coding and synthesis* 495 (1995), p. 518.

[161] Carl Thomé and Sven Ahlbäck. "Polyphonic pitch detection with convolutional recurrent neural networks". In: *Music Information Retrieval Evaluation eXchange (MIREX)* (2017).

[162] RW Ward Tomlinson and Dietrich WF Schwarz. "Perception of the missing fundamental in nonhuman primates". In: *The Journal of the Acoustical Society of America* 84.2 (1988), pp. 560–565.

[163] Dmitri Tymoczko. *A geometry of music: Harmony and counterpoint in the extended common practice*. Oxford University Press, 2010.

[164] Dmitri Tymoczko. "The generalized tonnetz". In: *Journal of Music Theory* (2012), pp. 1–52.

[165] Rainer Typke, Frans Wiering, and Remco C Veltkamp. "A survey of music information retrieval systems". In: *Proc. 6th international conference on music information retrieval*. Queen Mary, University of London. 2005, pp. 153–160.

[166] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. "Adaptive harmonic spectral decomposition for multiple pitch estimation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2009), pp. 528–537.

[167] Tuomas Virtanen. "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria". In: *IEEE transactions on audio, speech, and language processing* 15.3 (2007), pp. 1066–1074.

[168] Anja Volk and Aline Honingh. "Mathematical and computational approaches to music: challenges in an interdisciplinary enterprise". In: *Journal of Mathematics and Music* 6.2 (2012), pp. 73–81.

[169] Hermann Von Helmholtz. *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*. Vieweg, 1863.

[170] John Von Neumann, Arthur W Burks, et al. "Theory of self-reproducing automata". In: *IEEE Transactions on Neural Networks* 5.1 (1966), pp. 3–14.

[171] Avery Wang et al. "An industrial strength audio search algorithm." In: *Ismir*. Vol. 2003. Citeseer. 2003, pp. 7–13.

[172] DeLiang Wang and Guy J Brown. "Multiple F0 estimation". In: (2006).

[173] Xiaoqin Wang and Kerry M. M. Walker. "Neural Mechanisms for the Abstraction and Use of Pitch Information in Auditory Cortex". In: *Journal of Neuroscience* 32.39 (2012), pp. 13339–13342. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3814-12.2012. eprint: https://www.jneurosci.org/content/32/39/13339.full.pdf. URL: https://www.jneurosci.org/content/32/39/13339.

[174] Xiaoqin Wang and Kerry MM Walker. "Neural mechanisms for the abstraction and use of pitch information in auditory cortex". In: *Journal of Neuroscience* 32.39 (2012), pp. 13339–13342.

[175] J. D. Warren et al. "Separating pitch chroma and pitch height in the human brain". In: *Proceedings of the National Academy of Sciences* 100.17 (2003), pp. 10038–10042. ISSN: 0027-8424. DOI: 10.1073/pnas.1730682100. eprint: http://www.pnas.org/content/100/17/10038.full.pdf. URL: http://www.pnas.org/content/100/17/10038.

[176] H.J. Weaver. *Theory of discrete and continuous Fourier analysis*. A Wiley Interscience publication. Wiley, 1989. URL: https://books.google.co.uk/books?id=1%5C_pQAAAAMAAJ.

[177] Jake Wellens. *A friendly introduction to group theory*.

[178] Edmund Taylor Whittaker. "XVIII.—On the functions which are represented by the expansions of the interpolation-theory". In: *Proceedings of the Royal Society of Edinburgh* 35 (1915), pp. 181–194.

[179] Francis M Wiener and Douglas A Ross. "The pressure distribution in the auditory canal in a progressive sound field". In: *The Journal of the Acoustical Society of America* 18.2 (1946), pp. 401–408.

[180] Robert Willis. "On vowel sounds, and on reed organ pipes". In: *Transactions of the Cambridge philosophical Society* (1830), pp. 231–276.

[181] Yu-Te Wu, Berlin Chen, and Li Su. "Automatic music transcription leveraging generalized cepstral features and deep learning". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 401–405.

[182] Chunghsin Yeh and Axel Roebel. "The expected amplitude of overlapping partials of harmonic sounds". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2009, pp. 3169–3172.

[183] Weiwei Zhang, Zhe Chen, and Fuliang Yin. "Multi-Pitch Estimation of Polyphonic Music Based on Pseudo Two-Dimensional Spectrum". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2095–2108.