UNIVERSITY OF BIRMINGHAM

University of Birmingham

A geometric framework for pitch estimation on acoustic musical signals

Goodman, Tom; Van Gemst, Karoline; Tino, Peter

A geometric framework for pitch estimation on acoustic musical signals

Tom Goodman, Karoline van Gemst & Peter Tiňo

Published online: 06 Oct 2021.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

A geometric framework for pitch estimation on acoustic musical signals

Tom Goodman [a*], Karoline van Gemst[b], and Peter Tiňo[a]

[a]*School of Computer Science, University of Birmingham, Birmingham, UK;* [b]*School of Mathematics and Statistics, University of Sheffield, Sheffield, UK*

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 theoretical and experimental perspectives, 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.

Keywords: Pitch estimation; signal processing; geometry; visualization; music information retrieval

1. Introduction

Music information retrieval is increasingly gaining momentum as a cross-disciplinary field of research (Muller et al. 2011), pulling together techniques and researchers from computer science, cognitive science, musicology, and electrical engineering, amongst others. Despite this, many problems that on the surface appear to be trivially solvable from a human perspective have proved intractable for computers, and thus, have remained unsolved.

Pitch estimation is one such problem – the ability to take a musical signal as an input, and at any given position in the signal, be able to ascertain what notes (pitch chroma/pitch height pairs) are present. This is made difficult because of the sheer volume of timbres that could be present, and notes that could be played with various amplitudes, amongst other things. The remarkable variety that exists within music, especially with increasing levels of polyphony, render the problem of multi-pitch estimation (MPE) incredibly challenging indeed.

Many approaches to pitch estimation are restricted to specific instruments (e.g. guitar tuning devices) (Steinberger 1996; Böck and Schedl 2012; Schramm et al. 2017). As a result, researchers are able to exploit certain properties and assumptions (pertaining to the specific context in which

---

*Corresponding author. Email: t.a.goodman@cs.bham.ac.uk

the methods will operate) to increase the accuracy of their approaches. This helps to circumvent the greater difficulty of developing approaches that function from a more generalized perspective.

Relatively recently, there has been a notable influx of approaches to problems in music information retrieval that utilize state-of-the-art machine learning techniques (Wu, Chen, and Su 2018; Kelz, Böck, and Widmer 2019). While the accuracy of such approaches has proved good, the lack of ability to inspect (and further, understand), the inner workings of them has resulted in a lack of deep insight, especially into the underlying mathematical structures that they are approximating. This provides a strong motivation to develop novel algorithmic approaches (Goodman and Batten 2018) in an attempt to discern the intrinsic models, perhaps even those that we, as humans, exploit to perform these tasks.

In this paper, a novel geometric perspective and methodology for both mono- and multi-pitch estimations is presented. Section 2 gives a broad overview of the related work, and a number of inspirations for the paper, following which Section 3 introduces the model, which is then formalized in Section 4. Building on this, Section 5 examines the "edge cases" that arise in more detail. Sections 6 and 7 present an approach to examine the prevalence of edge cases and apply it to data sampled randomly from the total space. Moving on from the more theoretical perspective, Section 8 applies the model to real-world data, with Section 9 more closely analysing the performance of simple algorithms working on the proposed model. Finally, Section 10 proposes future direction for the work and a number of potential applications of the framework.

2.   Related work

Over the past few years, a plethora of approaches have been taken to tackle the challenge of pitch estimation, with methods utilizing the wavelet transform (Kumar and Kumar 2020), two-dimensional spectra (Zhang, Chen, and Yin 2020), and visual information (i.e. by viewing the physical instrument itself) (Koepke, Wiles, and Zisserman 2019), amongst others. Though varied, much of the cutting-edge research is reliant on machine learning (ML) techniques, not necessarily seeking to better understand the underlying structures present, and opting rather to maximize efficacy of the respective approaches (Figure 1).

Elowsson (2020) proposed a method for MPE that relies on "deep layered learning" (Elowsson and Friberg 2014; Elowsson 2018). 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 spectogram) through a spectral summation (their section IV-F), whitening, and logistic regression, which provides a much cleaner basis from which to detect pitch peaks. A neural network is 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 (Miron, Carabias-Orti, and Janer 2014), with subsequent regions then merged (where related), and finally a peak ridge (1D contour) is extracted. This method exhibited state-of-the-art performance on the MAPS (Emiya et al. 2010), Bach10 (Duan and Pardo 2015), TRIOS, and MIREX Woodwind quintet data sets.

Kelz, Böck, and Widmer (2019) derive pitch contours from polyphonic audio specifically from piano. Previous ML approaches tended to use many networks to extract various features, but they instead use a shared representation (Ngiam et al. 2011) to simultaneously predict the ADSR (attack, decay, sustain, release) aspects of each note. By approaching the problem in this way, they are able to increase the potential for generalization to other instruments, as the representations that prove useful to this task can be adapted to others. Further, rather than opting to train another network to learn the ADSR envelopes of the input, they handcraft a hidden Markov model (HMM) with states corresponding to each envelope, as well as an additional
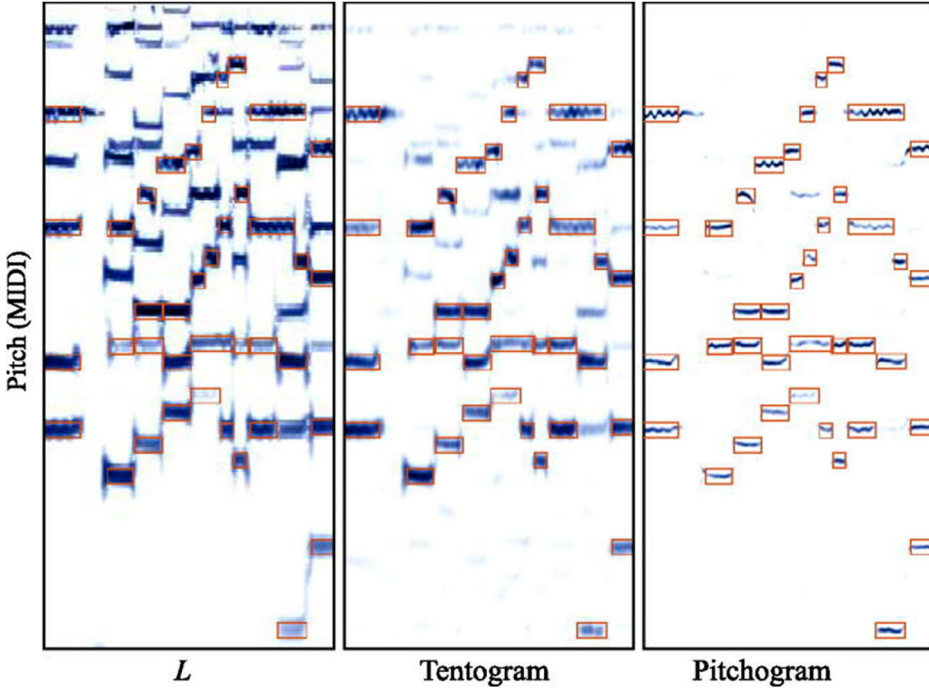
Figure 1. Demonstration of the progression from Spectogram to Tentogram, and then to Pitchogram (Elowsson 2020).

state to represent that a note is not currently sounding. Similarly to this approach, it could be possible to use the proposed framework (Section 4) in a kindred manner (albeit from a different perspective).

The autocorrelation function (ACF) of a signal $\{x(n)\}_{n=0}^{N-1}$ at lag m,

$$r_{xx}(m) = \sum_{n=m}^{N-1} x(n)x(n-m),$$

has been widely used to elicit pitch information from time-domain signals (Rabiner 1977; Amado and Vieira Filho 2008; Kraft and Zölzer 2015). Recently, de Obaldía and Zölzer (2019) conducted a study looking at the efficacy of the ACF on non-stationary sounds (to extract the fundamental period) and presented a number of augmentations that allowed them to improve on the current state-of-the-art approaches for monophonic pitch estimation. They achieve this by utilizing musicological knowledge to construct a heuristic that identifies non-related jumps in the pitch contour and subsequently modifying the signal to compensate for these. They report state-of-the-art results on both speech (including PTDB-TUG Pirker et al. 2011) and musical signals (including Bach10 Duan and Pardo 2015).

First described by Euler (1739), the *Tonnetz* (Figure 2) 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 $(n + 3) \bmod 12$, 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.

Figure 2. A section of Euler's original Tonnetz, taken from his 1774 paper *De Harmoniae Veris Principiis*.

In addition to his work on the generalized *Tonnetz* (Tymoczko 2012), Tymockzo posits a geometrical treatment of music theory in his book *A Geometry of Music* (Tymoczko 2010), driven by underlying musicological knowledge. Wh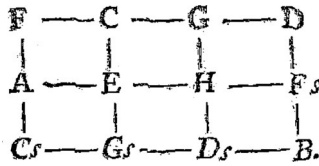ere Lewin (1987) tackled this kind of formalization from a group-theoretical perspective, Tymockzo (while still basing his approach on symmetries) employs a more geometric approach; considering pitch/chord spaces in such a way that proves useful to composers and musicologists alike.

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

- **O**ctave : transposing individual notes by an octave;
- **P**ermutation : reordering (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 (e.g. (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 generalization of *n*-note chord spaces in higher dimensions such as a three-note chord space forming a twisted triangular prism.

Inspired in particular by Tymockzo's work, and historic algorithms such as the harmonic pitch spectrum (HPS) (Noll 1970) and the YIN algorithm (De Cheveigné and Kawahara 2002), this paper sets out to look at the problem of pitch estimation from a geometric point of view and construct algorithms from the building blocks laid out herein.

3. Reaching a model

Consider a frequency-sorted (low to high) set of tones, e.g. {(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 3 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 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
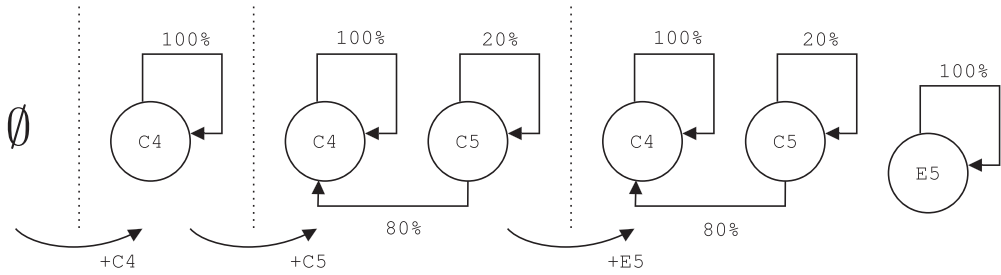
Figure 3. The build-up of a simplistic probabilistic representation.

that B is the ith 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.

Further, one can imagine that trying to represent a large number of vertices (and therefore a likely larger number of edges) renders this representation visually messy and hard to follow or decipher. By placing each vertex into a grid (with the horizontal axis representing the pitch 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, this is alleviated (Figure 4). 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 5).

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
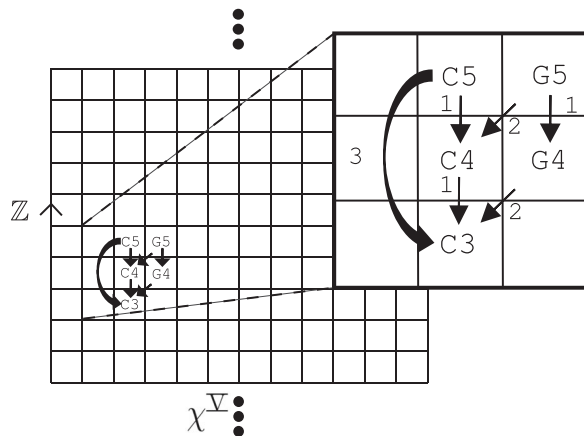
Figure 4. Visualization of the graphical structure overlaid onto the grid.
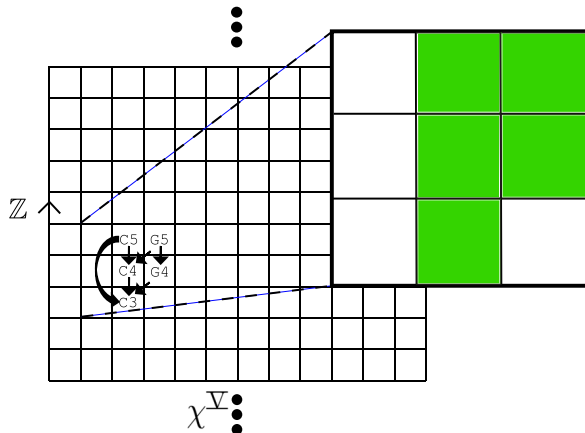
Figure 5.   Visualization of the final grid structure, where shaded cells represent a Boolean value of true.

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 6). By annotating each vertex, v, with its indegree ($\deg^-(v)$) and outdegree ($\deg^+(v)$), 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 3 is always a fundamental with its first three harmonics present. Thus a simple (yet somewhat effective) algorithm is to take each vertex with indegree 3 (for each distinct part of the graph, as it may not be connected), categorize them as fundamentals, and remove all categorized vertices. This can then be iteratively applied to the graph until no sinks with indegree 3 remain (as shown in Figure 6). 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, presenting some algorithms that utilize them.
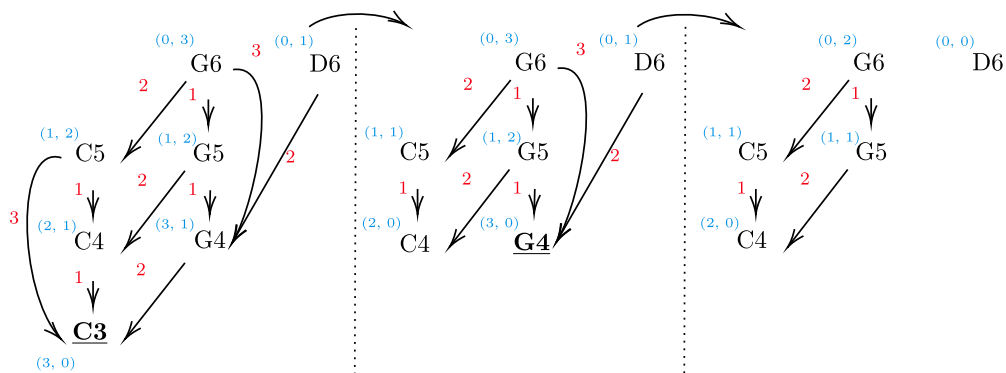
Figure 6.   (*left*) A directed graph depicting C3 and G4 (and each of their first three harmonics sounding). The degrees of each vertex are shown in parentheses. The following steps represent the steps of the simple algorithm described. The bolded/underlined tone at each step is the one selected as a fundamental.
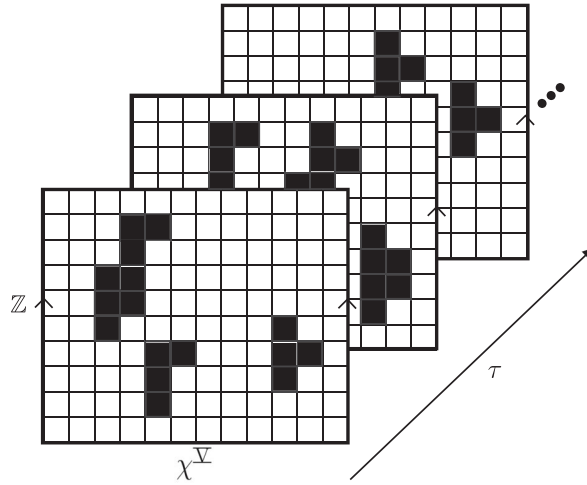
Figure 7. Visualization of a sequence of interpretations (representing temporal slices), indexed by τ.

4. The proposed model

As in Section 3, let $\chi^{\mathbf{V}}$ be the set of pitch chromas, ordered by fifths: $\{C, G, D, A, \ldots, F\}$.[1] Let the grid, $\mathcal{N}^{\mathbf{V}} := \chi^{\mathbf{V}} \times \mathbb{Z}$. That is, an element $v_{i,j} \in \mathcal{N}^{\mathbf{V}}$ is a pair (χ_i, j) representing a tone with pitch chroma χ_i and pitch height j. $\mathcal{N}^{\mathbf{V}}$ 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 realize $\mathcal{N}^{\mathbf{V}}$ as a discretized infinite cylinder. Furthermore, let $\mathcal{N}_\alpha^{\mathbf{V}}$ be the finite subset of $\mathcal{N}^{\mathbf{V}}$, consisting of the 'sub-cylinder' with octaves $[0, 9]$. This represents the human-audible spectrum of sound. Further, define the predicate, $\mathcal{I}_\tau : \mathcal{N}^{\mathbf{V}} \to \mathbb{B}$,

$$\mathcal{I}_\tau(v_{i,j}) = \begin{cases} \top & \text{if } v_{i,j} \text{ is observed at } \tau \\ \bot & \text{if } v_{i,j} \text{ is not observed at } \tau. \end{cases} \tag{1}$$

This is called an interpretation (i.e. of $\mathcal{N}^{\mathbf{V}}$) 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 $(\mathcal{N}^{\mathbf{V}}, \mathcal{I}_\tau)$ (Figure 7).

By viewing the ordered collection of pairs as a whole, one can uniformly stretch each slice and identify the appropriate $\mathcal{N}^{\mathbf{V}}$ faces to create a three-dimensional heatmap, with each tone now represented by a cube as opposed to a square.[2] Thus the interpretations are now indexed by an interval, where previously they had been indexed by an instantaneous point in time, that is,

$$(\mathcal{N}^{\mathbf{V}}, \mathcal{I}_\tau) \longmapsto (\mathcal{N}^{\mathbf{V}}, \mathcal{I}_{[\tau, \tau+1)}). \tag{2}$$

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

---

[1] Note here that the exponent, $\mathbf{V}$, is a label representing that the set is ordered by fifths.
[2] The importance of this becomes apparent in Section 8, in particular when considering the 3D heatmap.

Figure 8. Demonstration of the Γ and $\vdash$ shapes in $\mathcal{N}^{\underline{\mathbf{V}}}$.
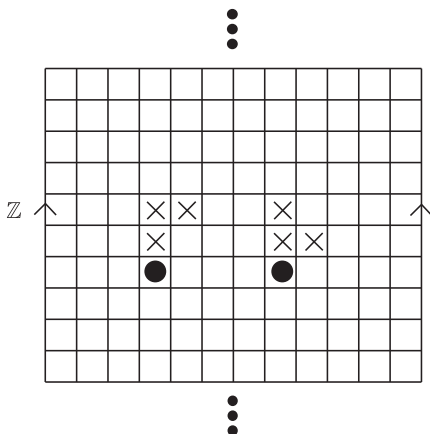
a pitch contour representation of the signal. Further, by viewing the heatmap as a translucent construction, it is possible to consider all aspects simultaneously.

Let f_i denote the ith harmonic, with f_0 being the corresponding fundamental. Depending on the pitch chroma, the second harmonic, f_2, may or may not cross the octave boundary (i.e. be in the next octave up from the first 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 $(A, n+1)$, $(E, n+2)$, and $(A, n+2)$. By considering each chroma in $\chi^{\underline{\mathbf{V}}}$, it is clear that the presence of $\{f_0, f_1, f_2, f_3\} \subset \mathcal{N}^{\underline{\mathbf{V}}}$ make up one of two shapes; a turnstile shape, $\vdash$, or a gamma shape, Γ, depending on the position of the fundamental. Denote the set

$$\chi_\vdash = \{C, C\sharp, D, E\flat, E\}, \quad \text{with } \chi_\Gamma = \chi^{\underline{\mathbf{V}}} \setminus \chi_\vdash \text{ as its complement,}[3]$$

and let π_χ, π_y be the projection of $\mathcal{N}^{\underline{\mathbf{V}}}$ 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 in Figure 8, where a fundamental is denoted by • and its harmonics by ×.

This model (particularly the use of $\chi^{\underline{\mathbf{V}}}$ as opposed to a chromatically ordered column set) is chosen such that the pattern exhibited by a fundamental and its first three harmonics (i.e. $\vdash / \Gamma$) appears spatially compact. This serves to make these patterns more easily discernible and is also of 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 $\mathcal{N}^{\underline{\mathbf{V}}}$. Let δ and ω denote the generators of $\mathbb{Z}_{12}$ (the integers modulo 12) and $\mathbb{Z}$, respectively. Then define a group action $\mathbb{Z}_{12} \times \mathbb{Z} \circlearrowleft \mathcal{N}^{\underline{\mathbf{V}}}$ as follows. δ and ω induce maps on $\mathcal{N}^{\underline{\mathbf{V}}}$ by

$$(\delta, \mathbb{1}_\mathbb{Z}) : \mathcal{N}^{\underline{\mathbf{V}}} \to \mathcal{N}^{\underline{\mathbf{V}}}, \quad (\mathbb{1}_{\mathbb{Z}_{12}}, \omega) : \mathcal{N}^{\underline{\mathbf{V}}} \to \mathcal{N}^{\underline{\mathbf{V}}},$$

$$v_{i,j} \mapsto v_{i+1,j} \quad v_{i,j} \mapsto v_{i,j+1}$$

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

---

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

a vertical shift of one cell downwards. Hence a map

$$\mathbb{Z}_{12} \times \mathbb{Z} \times \mathcal{N}^{\underline{\mathsf{V}}} \to \mathcal{N}^{\underline{\mathsf{V}}} : (\delta^k, \omega^l, v_{i,j}) \mapsto v_{i+k,j+l}, \tag{3}$$

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,

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

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

$$\vdash = \{\mathbb{1}, \omega, \omega\delta, \omega^2\}, \quad \Gamma = \{\mathbb{1}, \omega, \omega^2\delta, \omega^2\}, \tag{4}$$

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 $\mathcal{I}$,

$$\forall_{v \in \mathcal{N}^{\underline{\mathsf{V}}}} \big[(\mathcal{F}(v) \wedge \pi_\chi(v) \in \chi_\vdash) \to (\mathcal{I}(\omega(v)) \wedge \mathcal{I}(\omega\delta(v)) \wedge \mathcal{I}(\omega^2(v))) \big], \tag{5}$$

given that f_1, f_2, and f_3 are observed (audible). Here $\mathcal{F}(v)$ is a predicate that returns $\top$ iff v 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 $\vdash \Gamma$. These are used to categorize a number of properties of the model.

Definition 4.1 (Configuration)   A configuration (denoted as $\Gamma\Gamma$, $\Gamma \vdash$, or $\vdash \Gamma$) represents the shapes generated by fundamentals residing in two adjacent columns in $\mathcal{N}^{\underline{\mathsf{V}}}$.

While every fundamental, together with its first three harmonics, exhibit one of the two afore-mentioned shapes (i.e. Γ or $\vdash$), 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 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. $\pi_\chi(\delta^{-1}(\otimes))$.

Similar to how a fundamental and its first three harmonics corresponds to either a $\vdash$ or a Γ, a given harmonic could only have arisen from a fundamental related to it by the *inverse* shape, i.e. either $\vdash$ or L .
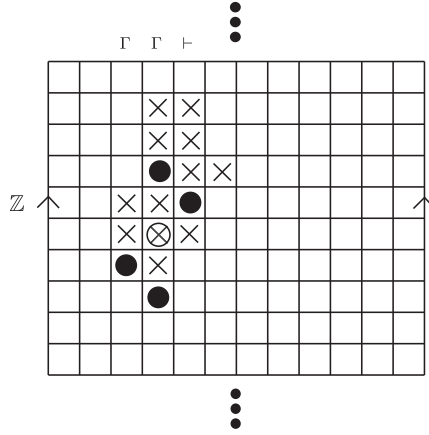
Figure 9.   Counterexample showing that not all Γ shape-exhibiting tones are fundamentals.

The inverse shapes may be written as

$$\dashv = \{\sigma^{-1} \,|\, \sigma \in \vdash\} = \{\mathbb{1}, \omega^{-1}, (\omega\delta)^{-1}, \omega^{-2}\}, \tag{6a}$$

$$\lrcorner = \{\sigma^{-1} \,|\, \sigma \in \Gamma\} = \{\mathbb{1}, \omega^{-1}, (\omega^2\delta)^{-1}, \omega^{-2}\}, \tag{6b}$$

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

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

$$\Psi(\chi_i) = \begin{cases} \vdash & \text{if } \chi_i \in \chi_\vdash \\ \Gamma & \text{otherwise.} \end{cases} \tag{7}$$

Intuitively this function takes a given chroma (i.e. $\pi_\chi(\nu) \in \chi^{\underline{\vee}}$) 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 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 $\vdash \bigcup L$. 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 $\mathcal{N}^{\underline{\vee}}$, 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 $\bigcirc(\otimes)$, with $\bigcirc := \{\sigma'\sigma \,|\, \sigma' \in \dashv \bigcup \lrcorner, \sigma \in \vdash \bigcup \Gamma\}$, as shown in Figure 10. This holds by construction.[4]

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 will naively result in false positives.

Hence, the problem of multi-pitch estimation is reduced to that of distinguishing between fundamentals ($\bullet$) and harmonics masquerading as fundamentals ($\otimes$).

---

[4] Note that the shape traced out by $\bigcirc$ is really the union of three shapes – each obtained from tracing backwards from a false fundamental (and its apparent harmonics) in one of the three configurations.
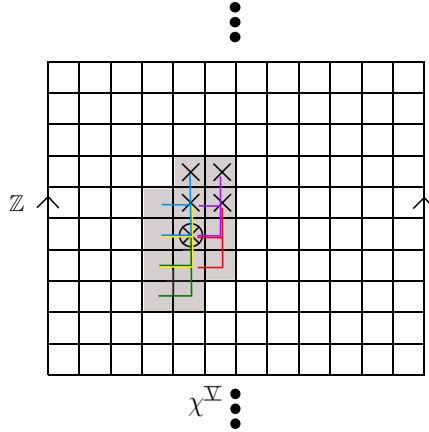
Figure 10. Figure showing where a false fundamental and each of its apparent harmonics could have been generated from, with colours tracing out ⊢ and L for each tone (overlaying both Γ and ⊢ situations).

5. Fantastic edge cases (and where to find them)

This section considers only cases in which a single false fundamental ($\otimes$) is present. In reality, this is expected to be enough of a generalization as long as algorithms consider false fundamentals sequentially, such that $\mathcal{N}_\alpha^{\underline{\vee}}$ is traversed along.

$$\delta^i \omega^j(\nu_{0,0}), \quad \forall_{i \in \{0,\ldots,11\}}, \ \forall_{j \in \{0,\ldots,9\}}, \tag{8}$$

that is, left-to-right, bottom-to-top, where $\nu_{0,0}$ is the bottom-leftmost element of $\mathcal{N}_\alpha^{\underline{\vee}}$.

Definition 5.1 (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

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

be the predicate that returns ⊤ iff ν is a false fundamental. Then a set of tones, S, is an edge case when

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

By considering ⊢ ⋃ L for each constituent[5] tone (similar to Figure 10), it is possible to construct logical expressions for the generators of 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 ⊢ ⋃ L. Then the sets of possible generators for a false fundamental and its apparent

---

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

Figure 11. A basic edge case in a ⊢ Γ configuration – note that each of the harmonics associated to the constituent tones of the false fundamental, ⊗, have a single generator.

harmonics are given by

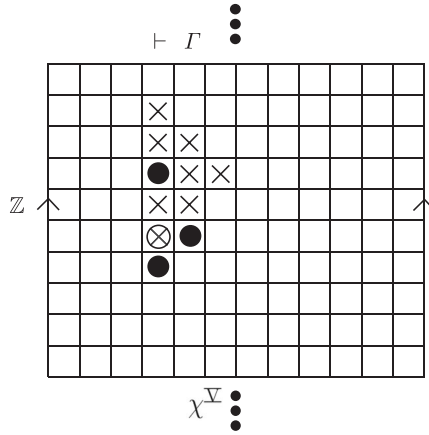$$f_0 : \{\omega^{-1}, \omega^{-2}, x\}, \quad \text{where } x = \begin{cases} (\omega\delta)^{-1} & \text{if } \Psi(\pi_\chi(\delta^{-1}(\otimes))) = \vdash \\ \omega^{-2}\delta^{-1} & \text{otherwise,} \end{cases} \tag{9a}$$

$$f_1 : \{\omega, \omega^{-1}, x\}, \quad \text{where } x = \begin{cases} \delta^{-1} & \text{if } \Psi(\pi_\chi(\delta^{-1}(\otimes))) = \vdash \\ (\omega\delta)^{-1} & \text{otherwise,} \end{cases} \tag{9b}$$

$$f_2 : \{\omega\delta, \delta, x\}, \quad \text{where } x = \begin{cases} \omega^{-1}\delta & \text{if } \Psi(\pi_\chi(\otimes)) = \vdash \\ \omega^2\delta & \text{otherwise,} \end{cases} \tag{9c}$$

$$f_3 : \{\omega, \omega^2, x\}, \quad \text{where } x = \begin{cases} \omega\delta^{-1} & \text{if } \Psi(\pi_\chi(\delta^{-1}(\otimes))) = \vdash \\ \delta^{-1} & \text{otherwise.} \end{cases} \tag{9d}$$

Note the use of shorthand here – these actions are all relative to (and applied to) a false fundamental, ⊗.

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

Definition 5.2 (Basic Edge Case)   A basic edge case is an edge case such that each constituent tone has precisely one generator.

In other words, only one element in each of the sets (9) is a generator. See Figure 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. 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 1). Note that f_2 has been omitted as it has no overlap (and therefore, multiplying the end result by 3 is sufficient).

The same holds true for the ΓΓ and Γ ⊢ 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 the definition of cap types given by Davis and Maclagan (2003) for the card game SET. This

Table 1. The enumeration of possible basic edge cases for the $\vdash \Gamma$ configuration, with each basic edge case corresponding to a row in the table.

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	$8 \times 3 = 24$

Note: A dash represents that a choice need not be made as a previous choice already satisfies the harmonic.

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.

Let

$$g : \mathcal{N}^{\underline{\mathbf{V}}} \to \mathcal{N}^{\underline{\mathbf{V}}} \tag{10}$$

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 (9) for edge cases – for basic edge cases the map gives a unique generator, which is the situation in which we will consider this map. Note that we assume that no inharmonic noise is present, meaning that all tones have a defined generator. Hence the map (10) is well defined. As a demonstration, take the basic edge case shown in Figure 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.[6] 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),

$$f_0 : \quad \boldsymbol{\omega^{-1}}, \omega^{-2}, (\omega\delta)^{-1} \tag{11a}$$

$$f_1 : \quad \boldsymbol{\omega}, \boldsymbol{\omega^{-1}}, \delta^{-1} \tag{11b}$$

$$f_2 : \quad \omega^2\delta, \omega\delta, \delta \tag{11c}$$

$$f_3 : \quad \boldsymbol{\omega}, \omega^2, \omega\delta^{-1}. \tag{11d}$$

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 5.3 (Basic Edge Type)   Two basic edge cases are of the same *type* iff their two corresponding triples are related by

$$\delta^{-1} \mapsto \omega^k\delta^{-1}, \quad k \in \{-1, 0, 1\}. \tag{12}$$

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 $\delta^{-1} \mapsto \omega^{-k}\delta^{-1}$.

---

[6] Note here that f_0 is assumed to be a false fundamental, so $g(f_0, f_1, f_3) \neq (g(f_0), g(f_0), g(f_0))$.
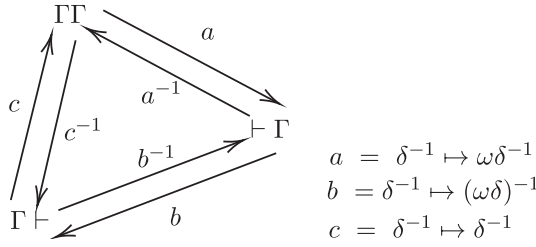
Figure 12. Diagram showing the relationships between members of the same type between different configurations.

Table 2. Different types of basic edge case, together with their invariants, and elements (excluding f_2).

Type	$|\delta^{-1}|$	ϵ	G.S.	Cases	
				$\Gamma\Gamma / \Gamma \vdash$	$\vdash \Gamma$
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		$8 \cdot 3 = 24$	

For any given type, there will be precisely *three* members – with precisely one being associated to each of the three configurations. The members of each type (by configuration) are related as according to Figure 12.

This becomes clearer following inspection of Table 2. Intuitively, the only difference between the $\vdash$ and Γ shapes is the shifting of f_2. The f_2 in a Γ shape is obtained from the corresponding $\vdash$ shape by acting ω on its f_2, and its inverse on the contrary.

LEMMA 5.1   *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 5.3, respectively. Further, transitivity holds by virtue of the diagram in Figure 12. ∎

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

Figure 13 visually represents the eight basic edge types shown in Table 2.

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.).

Definition 5.4 (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 a given edge case.

Example   The triple $(g(f_0), g(f_1), g(f_3)) = (\omega^{-1}, \omega^{-1}, \omega\delta^{-1})$ has back-δ count $|\delta^{-1}| = 1$.
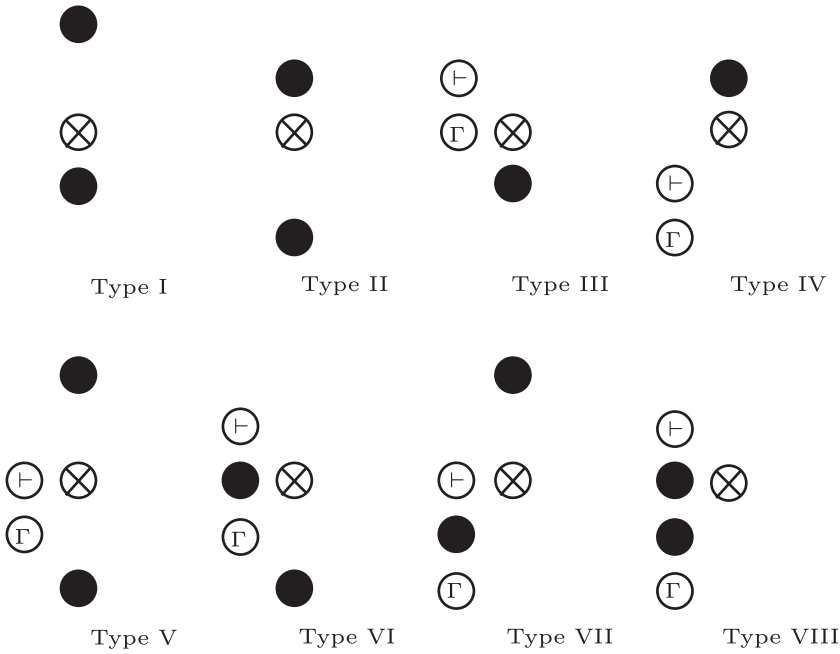
Figure 13. The eight basic edge types represented visually. Each • represents a generator, with ⊗ representing the false fundamental, and the unfilled generators containing ⊢ or Γ denoting the shape drawn out in the δ^{-1} column (i.e. corresponding to $\Psi(\pi_\chi(\delta^{-1}(\otimes)))$).

Definition 5.5 (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 $\epsilon = 2$.

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 five possible generating structures,[7]

$$\text{(I)} \quad f_0 = f_1 = f_3,$$
$$\text{(II)} \quad f_0 = f_1 \neq f_3,$$
$$\text{(III)} \quad f_0 = f_3 \neq f_1,$$
$$\text{(IV)} \quad f_0 \neq f_1 = f_3,$$
$$\text{(V)} \quad f_0 \neq f_1 \neq f_3.$$

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. 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 2.

As can be seen in Table 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

---

[7] Note that here, f_i is shorthand for $g(f_i)$.

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).

Remark   As before, the multiplication by 3 in Table 2 is due to there being no restrictions on the choice of generator for the second harmonic.

LEMMA 5.2   *The second "harmonic" (f_2) of a false fundamental ($\otimes$) 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 $g_i \in \mathcal{N}^{\underline{\mathbb{V}}}$ for f_2 that lies in the same column as the false fundamental.[8] 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, $\otimes$ would no longer be a false fundamental.

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

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

   Thus, as a contradiction is reached in both possible cases, there can be no such generator for f_2. ∎

PROPOSITION 5.1   *There are 24 basic edge types.*

Proof   This follows from Table 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 chroma 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

$$\{\delta^{\pm 1}, \omega^{\pm 1}\} \quad \text{and} \quad \{\delta^{\pm 1}, \omega^{\pm 1}, \omega^{\pm 1}\delta^{\pm 1}\}, \tag{13}$$

on ν, respectively.

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 $f_0, \ldots, f_3$) from Figure 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 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.

---

[8] Note that the only other choice would be the column directly to the right of $\otimes$.
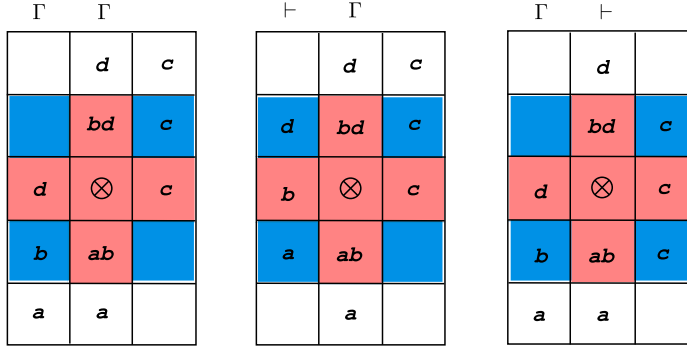
Figure 14. The potential generators (a (f_0), b (f_1), c (f_2), d (f_3)) for each configuration, and the von Neumann (red) and Moore (coloured) neighbourhoods.

Table 3. Table showing the minimum generators in the von Neumann (v.N.) and Moore neighbourhoods of a false fundamental given its chroma configuration.

	$\Gamma\Gamma$	$\vdash\Gamma$	$\Gamma\vdash$
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

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 6 and 7 go on to look at reduction of edge cases to potential basic cases, and the experimental prevalence of basic edge types, and Sections 8 and 9 investigate the theoretical basis of the model from a more experimental standpoint.

6. 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 $\bigcirc(\otimes)$ 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 $\rightarrow_{\mathfrak{g}}$) takes $\mathcal{G}$, and removes some given generator $\mathfrak{g} \in \mathcal{G}$ such that the false fundamental is still satisfied by $\mathcal{G} \setminus \mathfrak{g}$,

$$\mathcal{G} \rightarrow_{\mathfrak{g}} \mathcal{G} \setminus \mathfrak{g}. \tag{14}$$

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
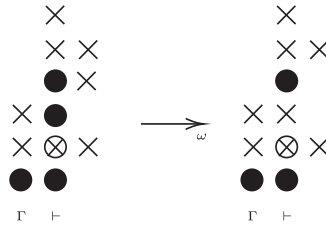
Figure 15. A reduction removing a generator in $\vdash \bigcup \Gamma$. Note that $g(f_2)$ is omitted for brevity.
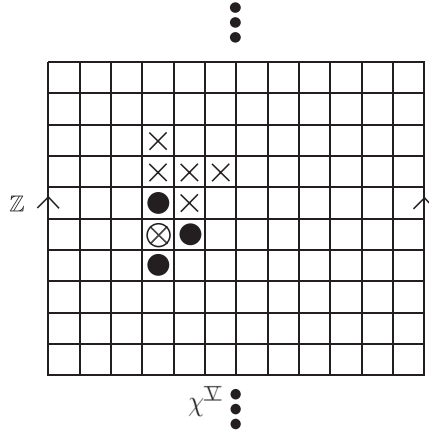

Figure 16. An irreducible non-basic edge case.

generator in $\vdash \bigcup \Gamma$, such as in Figure 15. Note that reduction is not unique; there may be multiple valid reductions that can be applied to a given set of generators.[9]

It would be reasonable therefore to assume that any non-basic edge case can be reduced to one of the eight basic edge cases (Table 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 while preserving the false fundamental (Figure 16). However, this is the only irreducible non-basic edge case.

PROPOSITION 6.1 *The only irreducible non-basic edge cases are those containing both ω^{-1} and ω.*

Proof Let $\mathcal{G}_n$ 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 16.

Thus it remains to prove the latter. In addition, the case where both ω^{-1} and ω are 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 16.

Note that a lack of overlap, i.e. $|\mathcal{G}_i| \cap |\mathcal{G}_j| = \emptyset$, for all $i \neq j$, implies that a case can always be reduced to a basic edge case.

---

[9] It is worth further noting that one could reduce 'globally' (i.e. over $\mathcal{N}_{\overline{\alpha}}^{\underline{\mathbb{V}}}$), or 'locally', considering just $\bigcirc(\otimes)$. Because false fundamentals generally need only be considered locally (as they can be generated solely by fundamentals within $\bigcirc(\otimes)$), 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 with large numbers of fundamentals.
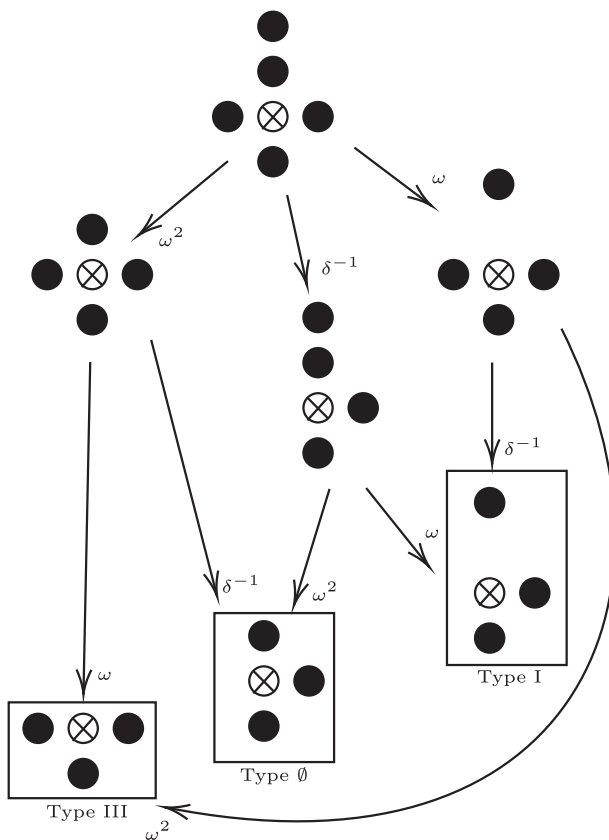
Figure 17. An example of a reduction graph, with each step (arrow) showing a reduction in the set of generators. Note that the special case of $\{\omega, \omega^{-1}, g(f_2)\}$ is denoted as "Type $\varnothing$," and the configuration is $\Gamma\Gamma$ or $\Gamma \vdash$.

For all cases, an enumeration can be performed[10] $\{\omega^{-1}, \omega\}$, and the statement holds as required. ∎

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 17). 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.[11] Such a graph can, therefore, be used in order to understand the potential basic edge types that correspond to a given set of generators.

7. 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

---

[10] By considering $\mathcal{G}$ to contain each possible subset of

[11] 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.

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).

In order to sample interpretations from the total sample space, it proves sufficient to construct them by selecting n unique tones, $\nu_0, \nu_1, \ldots, \nu_n$, from $\mathcal{N}_\alpha^{\underline{\mathrm{V}}}$, treating all such tones as fundamentals, and thus adding them (and their harmonics) to the interpretation. For the charts in Figure 18, 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 $\mathcal{N}_\alpha^{\underline{\mathrm{V}}}$ to act as generators until n unique generators were selected. We made no effort 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, we applied a naive algorithm (simply classifying all $\vdash$ and Γ-exhibiting tones as fundamentals) and derived a reduction graph 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 $\mathcal{N}_\alpha^{\underline{\mathrm{V}}}$ that exhibits the expected shape is classified as a fundamental. From each of these reduction graphs, we classified all terminal vertices as either one of the basic edge types, or as the special case, $\varnothing$. In cases where multiple terminal vertices were present, we added a value to each tally such that the sum of all added values was 1.[12]

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 15,750 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 20,000 interpretations was used (Figure 19).

Looking at Figure 18, it is clear that not all basic edge types are equally common. Figure 18(a) 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 incentivizes the use of Figure 18(c) – looking at the trends of the prevalences as the number of fundamentals changes.

As Figure 18(c) 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 (e.g. string quartet), cases 5–8 are incredibly unlikely to occur.

---

[12] Note that this could easily be reformulated to look at *sets* of types as opposed to single types.

(a) A pie chart showing the average (proportional) prevalence of each edge type, and ∅.

(b) A graph showing the estimated average number of terminal vertices for varying numbers of simultaneous unique fundamentals.

(c) A stacked bar chart showing the change in proportional prevalence of basic edge types and ∅ as the number of simultaneous unique fundamentals changes.

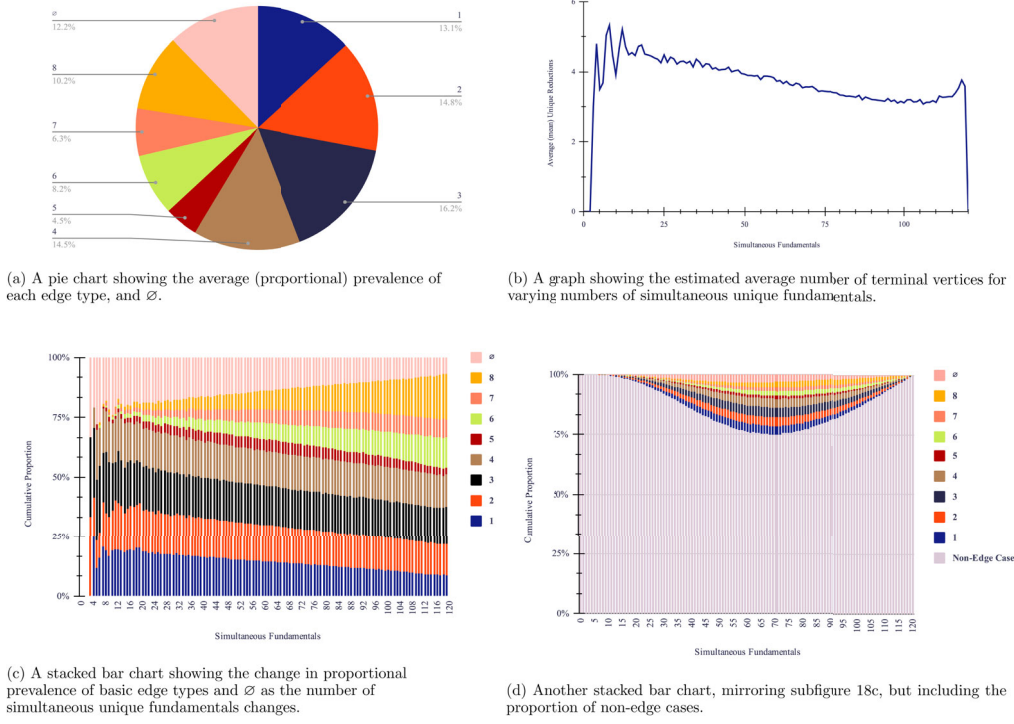(d) Another stacked bar chart, mirroring subfigure 18c, but including the proportion of non-edge cases.
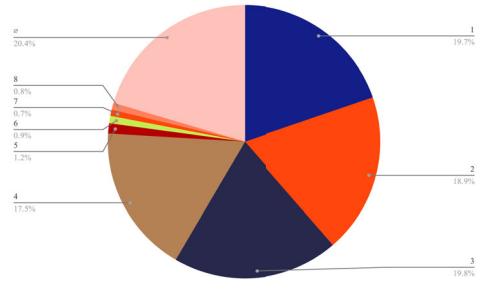
Figure 18. Charts depicting various properties relating to the prevalence of basic edge types with respect to the number of simultaneous unique fundamentals.

(a) A pie chart showing the average (proportional) prevalence of each edge type, and ∅ for total simultaneous fundamentals (0, 10].
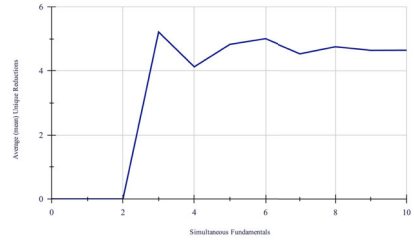
(b) A graph showing the estimated average number of terminal vertices for varying numbers of simultaneous unique fundamentals (0, 10] - corresponding to the left hand side of Subfigure 18b. Note that contrary to Subfigure 18b, 20000 sample interpretations were taken here, as opposed to 1000.
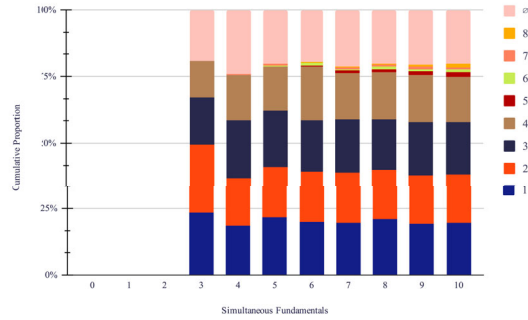
(c) A stacked bar chart showing the change in proportional prevalence of basic edge types and ∅ for (0, 10] simultaneous unique fundamentals.

Figure 19. Charts mirroring those in Figure 18, but considering only interpretations with (0, 10] simultaneous unique fundamentals.

- From Figure 18(d), 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 18(b), 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.

Figure 19 considers 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 19a,b). 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 18.

8. Experimental application

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 interpretations, $\mathcal{I}$. I.e. $\mathcal{I} : \mathcal{N}^{\underline{\mathbf{V}}} \to \mathbb{B}$ becomes $\mathcal{I} : \mathcal{N}^{\underline{\mathbf{V}}} \to \mathbb{R}$.

Doing so effectively creates a heatmap, in which this additional dimension (perpendicular to $\mathcal{N}^{\underline{\mathbf{V}}}$) represents an intensity of each tone – for example, their respective amplitudes in the frequency domain.[13] 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) (Fritts 2012) (Figure 20). Here, the intensity is visualized 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, while 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, we took sliding windows from the signal, with a length of 4096 samples, and a hop size of 1024. A constant-Q transform (Brown 1991) with a Hanning window (Podder et al. 2014) (using the Librosa implementation (McFee et al. 2015)) was then applied to achieve a frequency domain representation binned by the 120 semitones of the western musical scale between C0 and B9 inclusive. We then normalized these values across the signal (not just

---

[13] This construction can be viewed as a real rank 1 trivial vector bundle, with $\mathcal{N}^{\underline{\mathbf{V}}}$ as base manifold with trivial topology. In this interpretation, a heatmap is a slice through the bundle.
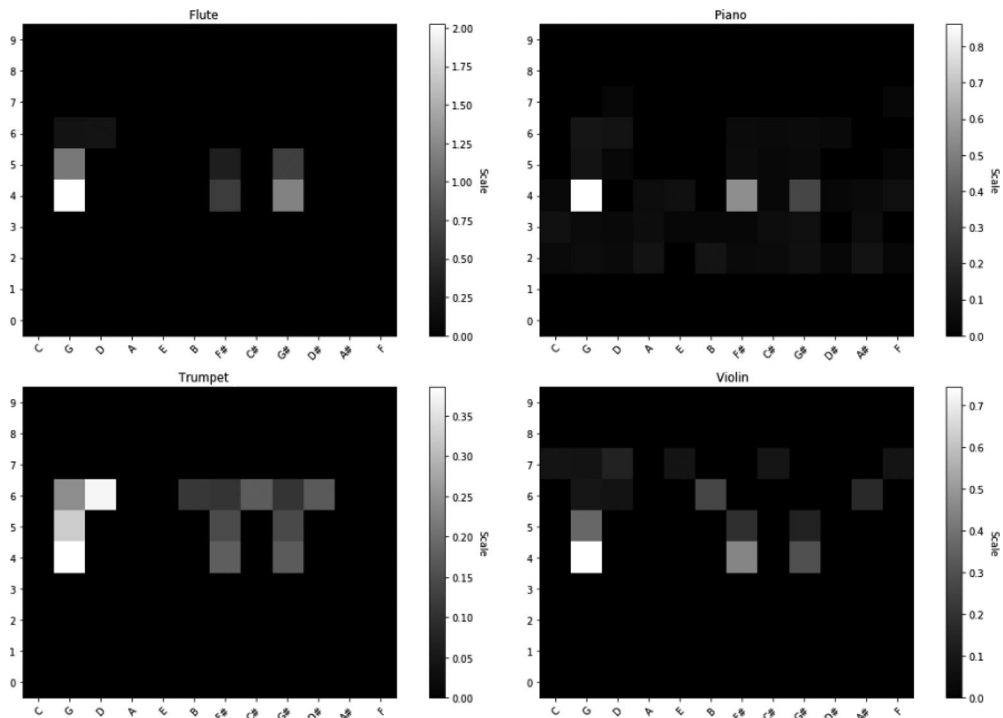
Figure 20.   The tone G4 being played on a variety of instruments, all exhibiting the Γ shape described in Section 4. For flute, trumpet, and violin, the sample was taken from the stationary period, whereas the piano sample was from part-way through the onset, as this resulted in a clearer image.

per window) and plotted them as a heatmap using matplotlib (Hunter 2007). 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 20) are 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.

Figure 21 – created using Python's vpython module (Scherer, Dubois, and Sherwood 2000) – shows the three-dimensional heatmap described in Section 4 (with each $\mathcal{I}$ indexed as $\mathcal{I}_{[\tau,\tau+1)}$). A projection onto the $\mathbb{Z}_{12} \times \tau$ plane produces piano-roll notation. 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 certainly a great oversimplification to treat each window (and therefore each interpretation) as independent of one another.

Figure 22 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 effective heuristic is to consider both the total number of bins filled above some threshold α (e.g. 3.25μ (Goodman and Batten 2018), where 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,

$$\frac{\sum_{\nu \in N} \mathcal{I}(\nu)}{|N|}, \quad N = \{\nu \,|\, \mathcal{I}(\nu) \geq \alpha, \; \nu \in \mathcal{N}_\alpha^{\underline{\nabla}}\}; \tag{15}$$
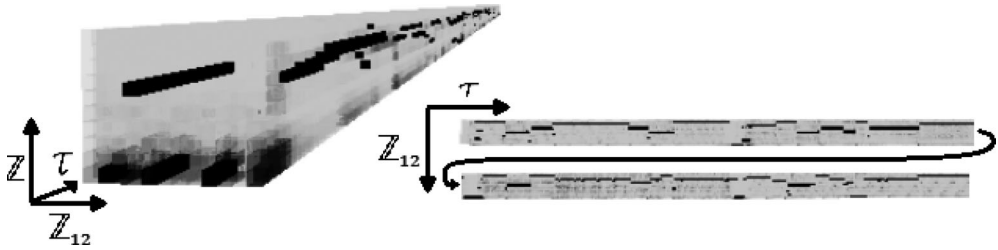
Figure 21. Left: Side-on view of a 3D heatmap of the melody of Bach's "Ach Gott und Herr" from the Bach10 data set (Duan and Pardo 2015), with darker colours corresponding to greater amplitudes. Right: Projection of the heatmap onto the $\mathbb{Z}_{12} \times \tau$ plane, eliciting piano roll notation of the piece (albeit ordered by the circle of fifths, and not chromatically).
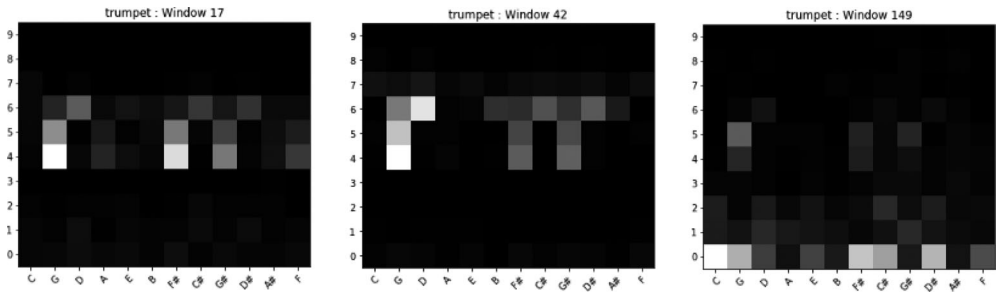

Figure 22. A closer look at how heatmaps differ as the tone progresses from onset/attack, to stationary period, to offset/decay (left-to-right).

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 23 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 utilized to cut out noise. Of course, the use of Algorithm 1 alleviates this by attempting to remove inharmonic noise, but in doing so, may cause false negatives to arise. Note that when using Algorithm 1 on real-world data, the same switch from $\mathbb{B}$ to $\mathbb{R}$ applies. Appendix 1 lists the required modifications.

9. Evaluation

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 utilized in Section 7. As noted beforehand, the intention of this paper (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 regarding future work.

For monophonic signals, we used a naive algorithm that treats the relation between fundamentals and $\vdash$/Γ-exhibiting tones as an equivalence. As shown in Section 4, this is untrue due to the presence of edge cases, but nonetheless when only one fundamental is present, such cases can
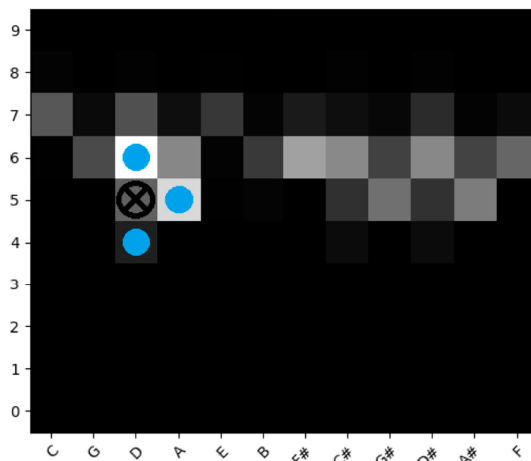
Figure 23. An edge case (specifically ∅) being exhibited on real data (trumpet) – with D4, A5, and D6 (dots) as the fundamentals, and D5 being the false fundamental, ⊗.

Table 4. 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.

	HPS	Naive
Overall	58.36%	73.74%
No outliers	67.73%	88.27%
Chroma accuracy	77.61%	95.08%

never occur. In response to spectral leakage, we slightly modified the algorithm 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 categorized 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. disregarding octave errors), this increases to 95.08%. Table 4 benchmarks this against an implementation of Noll's harmonic product spectrum (HPS) algorithm, also using a Hanning window.[14]

Appendix 2 contains a table with a full breakdown of results, broken down by instrument (and vibrato/non-vibrato playing).

While, as expected, this approach does not reach state of the art results, it still outperforms HPS by a significant margin. Figure 24 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 visualization has the potential to elicit more profound

---

[14] For more information on the HPS algorithm, or the data used, see Sections 2 and 8 respectively. In addition, note that the raw data is converted into interpretations by way of Algorithm 1.
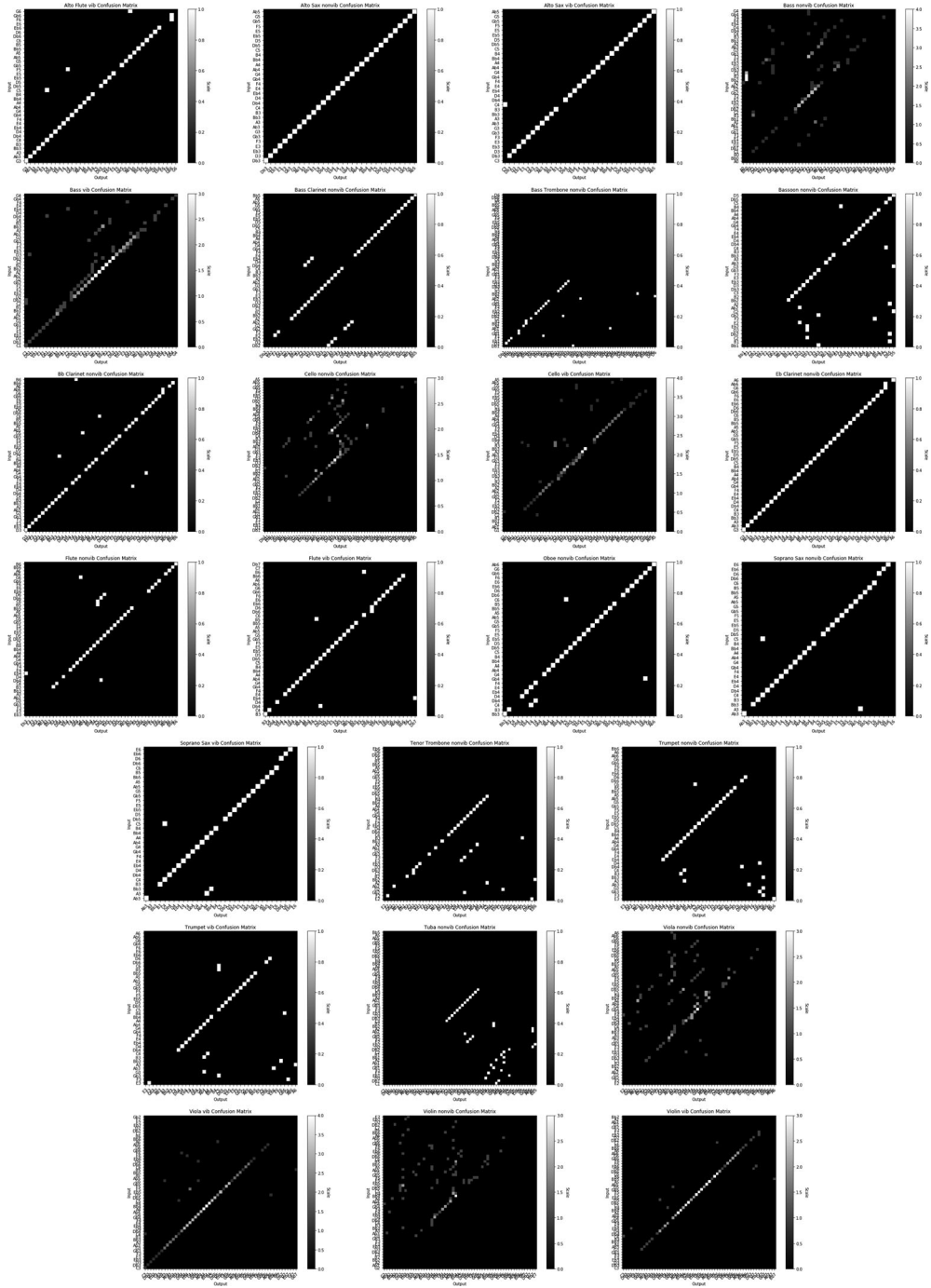
Figure 24. Confusion matrices for each set of samples. From top-left to bottom-right: (**1**) Alto Flute (vib.); (**2**) Alto Sax (non-vib.); (**3**) Alto Sax (vib.); (**4**) Bass (pizz. non-vib); (**5**) Bass (arco, vib); (**6**) Bass Clarinet (non-vib.); (**7**) Bass Trombone (non-vib.); (**8**) Bassoon (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).

Figure 25. Simulated accuracy for a naive approach (blue) (as described in Section 7), and simple algorithm (red) when applied to sample polyphonic data.

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.
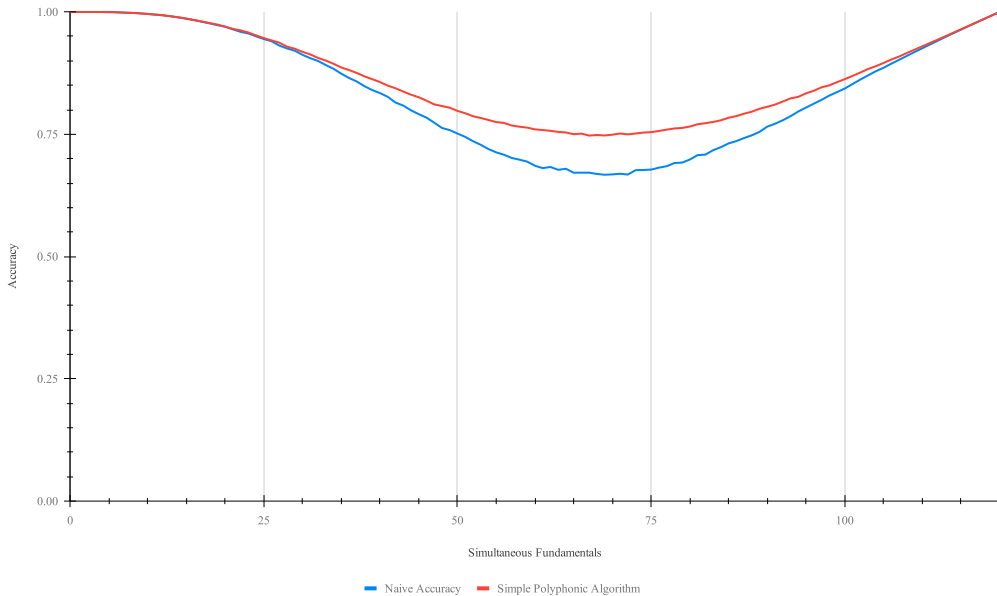
In addition, for polyphonic input, we used a simple extension to the naive monophonic approach, whereby $\mathcal{N}_{\alpha}^{\underline{\vee}}$ is traversed left-to-right, bottom-to-top.[15] This exploited the assumption that, at least with acoustic music, there will be no undertones. Thus the bottom-leftmost tone with amplitude above some cutoff will always be a fundamental (Goodman and Batten 2018). 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 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 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, while 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 7, 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 25. 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

---

[15] It should be noted here, as below, that the polyphonic algorithm was tested on random interpretations, as in Section 7. 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 $\mathcal{N}^{\underline{\vee}}$. In the future, this could be addressed by generating more realistic data, or indeed by utilizing polyphonic data sets such as Bach 10.

outlined in Sections 5, 6, and 7, but the implementation of this is beyond the scope of this paper.

10. Summary and future work

Moving forwards, there are a number of improvements and implementations that can be created building upon this framework.

First, the simple polyphonic approach can be refined and extended using the characterizations of edge cases outlined in Section 5. It may be possible to further improve by combining the approaches in Sections 6 and 7 to perform a "backward pass" over $\mathcal{N}_\alpha^{\underline{\mathrm{V}}}$, working right to left, top to bottom, and utilizing reduction to reconsider the likelihood of tones presenting as false fundamentals. This could also utilize the working count of currently perceived generators as a heuristic for the number of distinct simultaneous fundamentals, and similar forward–backward approaches are a fundamental concept in probabilistic latent variable models – a notion which itself aligns well with the problem of pitch estimation as set out in this paper.

In addition, it may prove useful to reformulate the problem as a decomposition of the total heatmap into its constituent (albeit constructively overlapping) $\vdash$ and Γ shapes, potentially using spectrogram subtraction to represent this decomposition. This pivots the work towards combinatorics rather than a necessarily more algorithmic perspective.

Furthermore, one could build on the three-dimensional model to create approaches that effectively utilize the temporal aspects of music in their predictions. For example, the extraction of $\vdash$ and Γ-shaped prisms from the extended heatmap – representing the tone, or tones, sounding for some period of time.

In conclusion, this paper presents a different perspective on approaching pitch estimation problems, doing so from a more geometric viewpoint. To this end, we introduced an idealized model of fundamentals and their harmonics (and later adapted it to real-world scenarios). Importantly, from a geometrical perspective in particular, this model results in spatially close shapes, namely $\vdash$ and Γ. Further, it provides a framework on which to approach pitch estimation problems in this way, along with a thorough investigation into the edge cases that occur in this model.

Though the simple algorithms outlined in Section 9 do not provide state-of-the-art results, the intention is instead to provide a solid foundation on which to construct more sophisticated algorithms – particularly utilizing the characterization of, and insight into, edge cases. This is certainly a step towards more intuitive and innovative geometrical solutions in the field of pitch estimation, and in music information retrieval on the whole.

Disclosure statement

Funding

ORCID

Tom Goodman http://orcid.org/0000-0001-9283-5010

References

Amado, Rafael George, and Jozue Vieira Filho. 2008. "Pitch Detection Algorithms Based on Zero-Cross Rate and Autocorrelation Function for Musical Notes." In *2008 International Conference on Audio, Language and Image Processing*, 449–454. IEEE.

Böck, Sebastian, and Markus Schedl. 2012. "Polyphonic Piano Note Transcription with Recurrent Neural Networks." In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 121–124. IEEE.

Brown, Judith C. 1991. "Calculation of a Constant Q Spectral Transform." *The Journal of the Acoustical Society of America* 89 (1): 425–434. https://doi.org/10.1121/1.400476.

Davis, Benjamin Lent, and Diane Maclagan. 2003. "The Card Game SET." *The Mathematical Intelligencer* 25 (3): 33–40.

De Cheveigné, Alain, and Hideki Kawahara. 2002. "YIN, a Fundamental Frequency Estimator for Speech and Music." *The Journal of the Acoustical Society of America* 111 (4): 1917–1930.

de Obaldía, C., and U. Zölzer. 2019. "Improving Monophonic Pitch Detection Using the ACF and Simple Heuristics."

Duan, Z., and B. Pardo. 2015. "Bach10 Dataset."

Elowsson, Anders. 2018. "Deep Layered Learning in MIR." Preprint, arXiv:1804.07297.

Elowsson, Anders. 2020. "Polyphonic Pitch Tracking with Deep Layered Learning." *The Journal of the Acoustical Society of America* 148 (1): 446–468.

Elowsson, Anders, and Anders Friberg. 2014. "Polyphonic Transcription with Deep Layered Learning." *MIREX*. http://www.music-ir.org/mirex/abstracts/2014/EF1.pdf.

Emiya, Valentin, Nancy Bertin, Bertrand David, and Roland Badeau. 2010. "MAPS-A Piano Database for Multipitch Estimation and Automatic Transcription of Music."

Euler, Leonhard. 1739. "1739. Tentamen Novae Theoriae Musicae Ex Certissismis Harmoniae Principiis Dilucide Expositae." *St. Petersburg, Also in Opera Omnia, Ser* 3 (1): 197–427.

Fritts, Lawrence. 2012. "Musical Instrument Samples." http://theremin.music.uiowa.edu/MIS.html.

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

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.

Kelz, Rainer, Sebastian Böck, and Gerhard Widmer. May 2019. "Deep Polyphonic ADSR Piano Note Transcription." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 246–250. IEEE.

Koepke, A. Sophia, Olivia Wiles, and Andrew Zisserman. 2019. "Visual Pitch Estimation."

Kraft, Sebastian, and Udo Zölzer. 2015. "Polyphonic Pitch Detection by Matching Spectral and Autocorrelation Peaks." In *2015 23rd European Signal Processing Conference (EUSIPCO)*, 1301–1305. IEEE.

Kumar, Neeraj, and Raubin Kumar. 2020. "Wavelet Transform-Based Multipitch Estimation in Polyphonic Music." *Heliyon* 6 (1): Article ID e03243.

Lewin, David. 1987. *Generalized Musical Intervals and Transformations*. Oxford University Press.

McFee, Brian, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. "librosa: Audio and Music Signal Analysis in Python." In *Proceedings of the 14th Python in Science Conference*, Vol. 8.

Miron, Marius, Julio José Carabias-Orti, and Jordi Janer. 2014. "Audio-to-Score Alignment at the Note Level for Orchestral Recordings." In *15th Conference of the International Society for Music Information Retrieval*, 125–130.

Muller, Meinard, Daniel P. W. Ellis, Anssi Klapuri, and Gaël Richard. 2011. "Signal Processing for Music Analysis." *IEEE Journal of Selected Topics in Signal Processing* 5 (6): 1088–1110.

Ngiam, Jiquan, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. "Multimodal Deep Learning." In *28th International Conference on Machine Learning*.

Noll, A. Michael. 1970. "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Surn Spectrum, and a Maximum Likelihood Estimate." In *Symposium on Computer Processing in Communication, ed.*, Vol. 19, 779–797. New York: University of Broodlyn Press.

Pirker, Gregor, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf. 2011. "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario." In *Twelfth Annual Conference of the International Speech Communication Association*.

Podder, Prajoy, Tanvir Zaman Khan, Mamdudul Haque Khan, and M. Muktadir Rahman. 2014. "Comparative Performance Analysis of Hamming, Hanning and Blackman Window." *International Journal of Computer Applications* 96 (18).

Rabiner, Lawrence. 1977. "On the Use of Autocorrelation Analysis for Pitch Detection." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25 (1): 24–33.

Scherer, David, Paul Dubois, and Bruce Sherwood. 2000. "VPython: 3D Interactive Scientific Graphics for Students." *Computing in Science & Engineering* 2 (5): 56–62.

Schramm, Rodrigo, Andrew McLeod, Mark Steedman, and Emmanouil Benetos. 2017. "Multi-Pitch Detection and Voice Assignment for a Cappella Recordings of Multiple Singers." In *12th Annual Conference of the International Speech Communicati 18th Conference of the International Society for Music Information Retrieval*.

Steinberger, Ned. 1996. "Chromatic Tuner Display Providing Guitar Note and Precision Tuning Information." US Patent 5549028.

Tymoczko, Dmitri. 2010. *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. Oxford University Press.

Tymoczko, Dmitri. 2012. "The Generalized Tonnetz." *Journal of Music Theory* 1–52.

Wu, Yu-Te, Berlin Chen, and Li Su. 2018. "Automatic Music Transcription Leveraging Generalized Cepstral Features and Deep Learning." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 401–405. IEEE.

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

Appendices

Appendix 1. Creation of an interpretation, $\mathcal{I}$

Algorithm 1 Creation of an Interpretation, $\mathcal{I}$, from a Sorted Set of Tones

---

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

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

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

$\quad\quad \Phi \longleftarrow \Phi \setminus v_{i,j}$
$\quad$ **end for**
$\quad$ **return** $\mathcal{M}$

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 tone to be considered "audible"
- Lines 13 and 27 should be replaced by $\mathcal{M}[i, j] = |v_{i,j}|$, setting the amplitude of $v_{i,j}$, 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**."

Appendix 2. Full results – naive algorithm (monophonic)

Table A1. 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.

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%

Note: 1, 2, and 3 correspond to the whole data set, sans outliers, and chroma accuracy respectively.