UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MACHINE LISTENING AS A GENERATIVE MODEL: HAPPY VALLEY BAND

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF MUSICAL ARTS

in

MUSIC COMPOSITION

by

David Andrew Kant

June 2019

The Dissertation of David Andrew Kant
is approved:

Professor Larry Polansky, Chair

Professor David Dunn

Professor Tom Erbe

Lori Kletzer
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Excerpts

List of Tables

Abstract

Machine Listening as a Generative Model:
Happy Valley Band

by

David Andrew Kant

ORGANVM PERCEPTVS is a collection of 11 songs for mixed ensemble written by translating machine listening analysis of pop songs into musical notation. Motivated by the idea that analysis algorithms inherently carry the values of the communities that produce them, I see my compositional process as a way of understanding how musical ideas, beliefs, preferences, and aesthetics are embedded within analysis algorithms. The musical scores are overly-specific, complex, and brimming with the artifacts of the machine listening process, and I formed a dedicated ensemble, the Happy Valley Band, to develop a performance practice unique to the idiosyncratic music. This dissertation essay documents the analysis algorithms used, my compositional process, the performance practice developed, the process of recording and releasing an album of music, and the public reception. I discuss my compositional ideas in the context of a number of twentieth century aesthetic traditions, including plunderphonics and sampling, computer analysis driven composition, and complexity in computer-assisted composition. I see this project as relevant to emerging cultural concerns of algorithmic bias and discrimination, and I relate my experience to contemporary dialogues around digital automation.

Acknowledgements

This project would not be possible without the endless time, energy, support, and dedication of the musicians of the Happy Valley Band. Thank you to Beau Sievers, Mustafa Walker, Andrew Smith, Alexander Dupuis, Pauline Kim, Conrad Harris, and to all who have played with us over the years.

To my advisory Larry Polansky, who played in the very first Happy Valley Band performance and taught me the importance of organizing and performing music, not just writing it. To my committee members David Dunn and Tom Erbe, for your invaluable insight, questions, and patience. To UCSC Professors Amy Beal, Dard Neuman, and Tanya Merchant, who generously offered their time and energy to help shape this project.

To my parents, who lovingly encouraged and supported me through all the stages of higher education, no matter how protracted and despite all better knowledge. To my partner Madison Heying, who has endured more Happy Valley Band than anyone else, for your support, encouragement, feedback, and advice. You have made me a better thinker.

Chapter 1

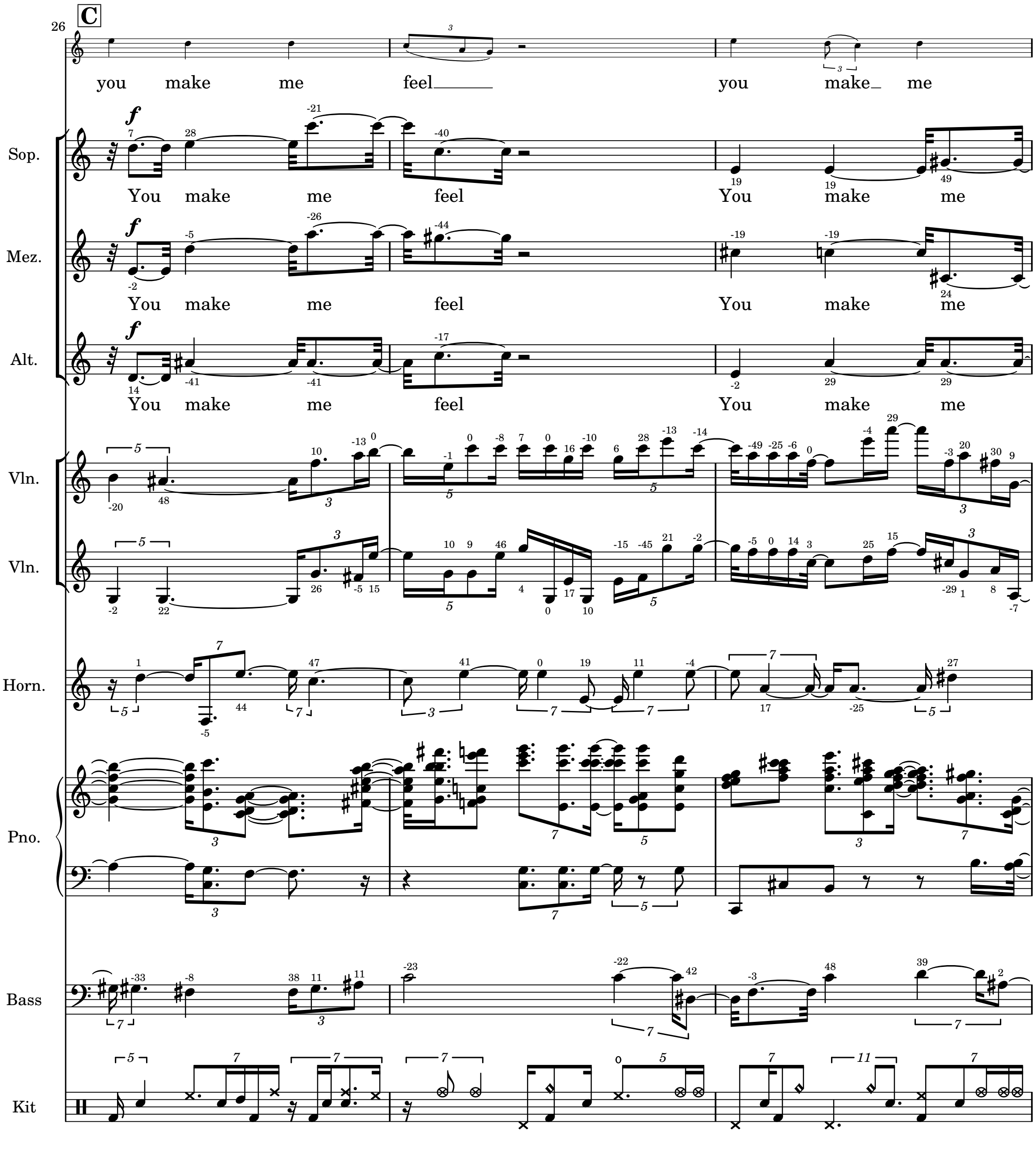
Introduction

As automated algorithms are increasingly woven into the fabric of contemporary life, it occurs to me not only that the individuals, institutions, and communities that produce them wield a tremendous amount of power, but more importantly, public understanding around computation — what the results of computational processes represent and how to interpret them — is sorely in need of a software update. Over the past eight years, I have explored this intuition through the Happy Valley Band (HVB), a project in which I transcribe machine listening analysis of pop songs into music notation for human musicians to perform. The motivations for this project extend back at least nine years to when I first became interested in sound analysis. The prospect that all of the subtle and ineffable qualities of musical experience are somehow encoded within a sequence of numbers and available for quantification enthralled me. However, as I came to understand the communities of thought that produce ideas and technologies in sound analysis, including engineering, industry, and higher education, I grew skeptical of many underlying musical values and assumptions.

As a musician interested in a variety of twentieth century contemporary music practices, including free improvisation, live electronic music, soundscape recording, noise, and avant jazz, I did not always feel that my musical values and interests were represented in the algorithms that I studied and in the communities that produced them. The more I learned about sound analysis, the more I came to understand that assumptions are necessary and inherent to it, assumptions about what one is even analyzing for in the first place — analyzing for pitch or rhythm requires an idea of what pitch or rhythm is. At a certain point analysis is tautologically true: you find what you look for. This struck me as deeply concerning, especially given the ubiquity of automated computation present today, whether musical or not. I became interested in what it means for values to be embedded within an analysis algorithm, the limits of digital sound analysis, what ideas and assumptions in particular motivate algorithms that are widespread today, and most importantly, the implications when analysis algorithms are automated, scaled, and left to operate autonomously.

1.1 The Happy Valley Band Project

This dissertation is a collection of songs written by translating machine listening analysis of pop songs into musical notation. Motivated by the idea that analysis algorithms inherently carry the values of the communities that produce them, I see my compositional process as a way to understand how musical ideas, beliefs, preferences, and aesthetics are embedded within analysis algorithms. HVB transcriptions are microtonal and rhythmically complex; pitch and rhythm are specified to exacting and inhuman degrees of precision, and the scores are rife with additional notes, artifacts of the machine listening process. The differences between the original recordings and HVB transcriptions are a result of my approach to machine listening. I highlight the idiosyncrasies of the machine listening process and the challenges of translating between scales of digital analysis and human perception. I often use well-known songs by artists such as Madonna, James Brown, Patsy Cline, and Phil Collins. Excerpt 1.1 shows a typical HVB score.


C
26
you make me feel you make me
Sop.
You make me feel You make me
Mez.
You make me feel You make me
Alt.
You make me feel You make me
Vln.
Vln.
Horn.
Pno.
Bass
Kit

Excerpt 1.1: *(You Make Me Feel Like) A Natural Woman*, Full Score

Over the past eight years, the HVB project has developed to include a suite of custom machine listening and music transcription tools — tools for spatial filtering, spectrogram decomposition, multiple fundamental frequency estimation, feature extraction, onset detection, and music notation — as well as a specific compositional process and approach to sound transcription. Composition and algorithm design are synonymous in HVB, from the musical ideas that motivate algorithm design down to technical implementation.

Most important, this project has grown to include a dedicated performing ensemble, the Happy Valley Band, who developed a unique performance practice in response to the unusual performance demands of the computer transcriptions. Performers devise their own strategies for navigating the overly complex and detailed scores but adapt in response to one another in live performance, allowing a group feeling to emerge. When I made the first machine listening transcriptions, I never imagined people would play them. In summer 2011, at the suggestion of two composer-performer colleagues, Beau Sievers and Mustafa Walker, we formed a dedicated ensemble to play the computer generated transcriptions. The ensemble has evolved into a collective structure, centered around a core group of musicians with whom I work closely — Alex Dupuis, Andrew Smith, Beau Sievers, and Mustafa Walker. The analysis tools, music notation, performance practice, and my interpretation of the project have emerged in the context of this group, shaped by their ideas and input.

When I first became interested in sound analysis, I wanted to know how machines hear. My advisor at the time gently suggested there is no way of hearing intrinsic to technology, rather machines reflect people's ideas about how hearing works. From the design of machine listening algorithms down to the constructs of boolean logic and material sciences that form the framework of modern computation within which algorithms are expressed, machines reflect human ideas. In many ways, this project has been my reconciling the desire for a way of hearing intrinsic to machines. I realized that as much as this desire seemed a fantasy of science fiction, there is a machine way of hearing, although it does not reside in the technological processes but in an assemblage of forces — the musical values of communities that produce them as well as the history of science, technology, and engineering that has led to the current understandings of sound and signal processing. This project argues the need to consider machine listening algorithms as a social construct if we are to understand how values are embedded within them.

Ultimately this project is motivated by the conviction that we use machine listening technologies to hear more inclusively, less be at risk of amplifying, in a positive feedback loop, the values of the communities that produce them to the exclusion of more diverse forms of music. Eight years ago this conviction was a vague intuition, more of a feeling than a well-formed idea. While I did not have the language necessary to articulate it, I was concerned by the implications for autonomous machine listening systems such as music recommendation or surveillance. Now more than eight years later, algorithmic bias — the idea that algorithms are not neutral but can unfairly discriminate against certain population groups — has emerged as a pressing and widespread social concern. Automated algorithms are involved in many aspects of daily life, from hiring practices to jail sentencing to autonomous surveillance, and news bubbles that came to light in the 2016 presidential election demonstrated the danger of using automation to reify rather than diversify our experiences.

1.2 Reflections

I have come to see this project from a number of different perspectives, which have helped me to understand and explain aspects of the project and draw various conclusions. First, HVB scores are as much a product of the music analyzed as the values and assumptions inherent to the algorithms. I have come to understand the music as *emergent* between the two; it resides neither in one nor the other but in the interaction between them. The concept of emergence provides for me a way of understanding the agency of the machine listening process. I believe that if we are to understand the agency of a distributed system of forces — a technological object such as an algorithm for instance being a network of factors, including technical and social — then the framework of emergent and dynamical systems is necessary.

Second, I find a concept from statistical modeling to be a useful metaphor for explaining my compositional approach, including the decisions that I make and how. I like to think of machine listening algorithms as a *generative model.* In contrast with other kinds of models that are only capable of describing phenomena, a generative model can be used to generate new instances of a phenomena, because they learn such rich representations of the phenomena they model. Extending this concept as a metaphor, I frame machine listening algorithms as generative models, and liken composition to exploring the possibility space of a high-dimensional parametric model. The metaphor of a generative model provides a way for me to see a description of sound embedded within an analysis algorithms and connects to Constructivist ideas of perception, which stress the extent to which our perception is shaped by prior held mental frameworks.

Third, the recent proliferation of concern for algorithmic bias gave language, concepts, and frameworks to my nascent intuitions that algorithms embody the values of the communities that produce them. The concept of a positive feedback loop is readily described in the literature to explain how self-reifying systems perpetuate biases to the exclusion of other ideas. Academic writing and news articles confirmed my suspicion that algorithms are commonly thought to be objective and unbiased — machine learning algorithms in particular — a finding which further amplified my sense that we need to better understand how bias is embedded in algorithms. While the field of algorithmic bias contains many open questions, most pressing among them are how to identify and ameliorate bias in autonomous systems. While HVB may not provide practical solutions that translate to other domains, I have come to think of it as a heuristic analysis of algorithmic bias, a way of coming to access the values inherent in an algorithms through musical interactions with them.

1.3 Related Work

George Lewis' work provided an important context for my ideas, in particular his 1987 piece *Voyager*. Lewis argued that algorithms are not neutral but reflect the values of the communities that produce them and furthermore that interaction with algorithms reveals said values. At the time (and arguable still today) most research and composition in computer music was produced by communities that were almost entirely all male and all white and dominated by trans European musical ideas. In response, Lewis designed *Voyager*, an interactive and spontaneous music system with a decidedly African American provenance, structuring the system around Jeff Donaldson idea of multidominance. In the context of Lewis' idea of interaction, I think of performing HVB music as a form of interaction that provides access to the values embedded within the listening algorithms.

I see HVB as related to a larger trend of work currently undertaken by a generation of young scholars who are exploring questions pertaining to the relationship between technology and culture, including how technologies embody and exhibit agency: Asha Tamirisa's work on how gender norms are encoded in electronic music technologies; Madison Heying's study of Carla Scaletti, the Kyma computer music language, and the community of users; Ezra Teboul's alternative history of electronic music examined at the component level of electrical design; Nick Seaver's ethnographic concept of the algorithm as well as his fieldwork with researchers in computer audition and recommender systems; Ted Gordon's study of Bay Area experimentalism and the fluidity between technological paradigms, composition, and lifestyle; and Josh Hudelson's account of the cultural effects of the concept of the "frequency domain" in signal processing. Through a variety of approaches and methodologies, these studies all address questions pertaining to how values are embedded within technologies and the implications, both cultural and technological.

Ultimately, my means of addressing such questions draws neither from the hard sciences, nor the social sciences, nor engineering, but is enacted through an artistic practice of music composition and performance. I believe that musicality is a sense unto itself, that through performance and listening we can come to know aspects of the world that are not otherwise explicable. By translating the artifacts of machine listening processes into music, my intuition was I would come to understand aspects of such technologies. Translating this experience back into words, however, is another challenge.

1.4 Contents of the Dissertation Essay

The purpose of this essay is to document the HVB project and to give context to the ideas explored. I describe the compositional process, detailing the algorithms used and why, and discuss the evolution of the ensemble as a collective entity. I also discuss the recording and production of HVB's 2016 album *ORGANVM PERCEPTVS*, as well as the public reception. I develop the concept of machine listening as a generative model, describe HVB music as emergent, and in the final chapter, I relate my experience to contemporary discussions of algorithmic bias.

In Chapter 2, I document the machine listening process and algorithms used. My intention in this chapter is to articulate the musical ideas underlying the various algorithms and my reasons for choosing them. I also identify the numerous decisions that need to be made and parameters that need to be set in order to produce a HVB transcription. Chapter 2 is the most technical chapter of the dissertation. I address the details of digital signal processing a well as music theory and music analysis.

In Chapter 3, I discuss the HVB ensemble. I describe the ensemble as a collective social structure and discuss how the group's performance practice evolved in response to the specific challenges of the music. I relate the ensemble's approach to traditions in contemporary music such as Complexity and computer-assisted composition. Despite the affinities, the defining aspect of HVB's performance strategy arises from group dynamics within a collaborative social structure.

Chapter 4 focusses on the ideas that guide my decisions when making a transcription. Having previously identified in Chapter 2 the various parameters that need to be determined, in Chapter 4 I discuss what considerations influence my decisions. I develop the metaphor of machine listening as generative model to explain my compositional approach.

Chapter 5 documents the process of recording and releasing *ORGANVM PERCEPTVS*. I reflect on the public reception of the record, one reviewer in particular who found the record to be “conceptually fascinating” but simply “cannot bear to listen to it.” I reflect on intellectual property and Artificial Intelligence, briefly discussing the history of sampling and new challenges to intellectual property posed by big data, automation, and learning algorithms.

I conclude in Chapter 6 by relating my experience to emerging concerns of algorithmic bias and discrimination. While the field is in need of practical, scalable solutions to identify and ameliorate bias in automated algorithms, it is admittedly difficult to argue what answers art can provide. Drawing on George Lewis’ idea that interacting with technologies reveals aspects of the communities that produce them, I develop the idea of a heuristic analysis of algorithmic discrimination, and reflect on the kinds of bias I find to be present in machine listening systems.

Chapter 2

Algorithms Used

The Happy Valley Band transcription process has four stages: 1) source separation, extracting individual instruments from the original recording; 2) pitch and rhythm analysis, analyzing the separated tracks for pitch and rhythm as well as other features of musical performance such as dynamics or articulation; 3) musical notation, expressing the results of pitch and rhythm analysis in the form of musical notation; and 4) performance, assembling an ensemble to perform the computer generated transcriptions. Figure 2.1 diagrams the analysis stages of a typical HVB transcription. For each step of the process, I developed custom software, sometimes drawing from well-known algorithms and other times designing tools from scratch. I developed my own software tools in part out of curiosity — I wanted to understand each step of the process — but, more importantly, because, in this project, composition and algorithm design are synonymous. The choice of analysis algorithms and parameter selection are the channels that I use to influence the resulting music.

This chapter focusses on my transcription process, the algorithms used, and how I use them. I document and discuss the technologies that I developed for each stage of the process, and I explain my motivations in the context of perception, acoustics, sound analysis, and music theory. This chapter is also an effort to make explicit the musical ideas, values, and preferences implicit within my transcription system.

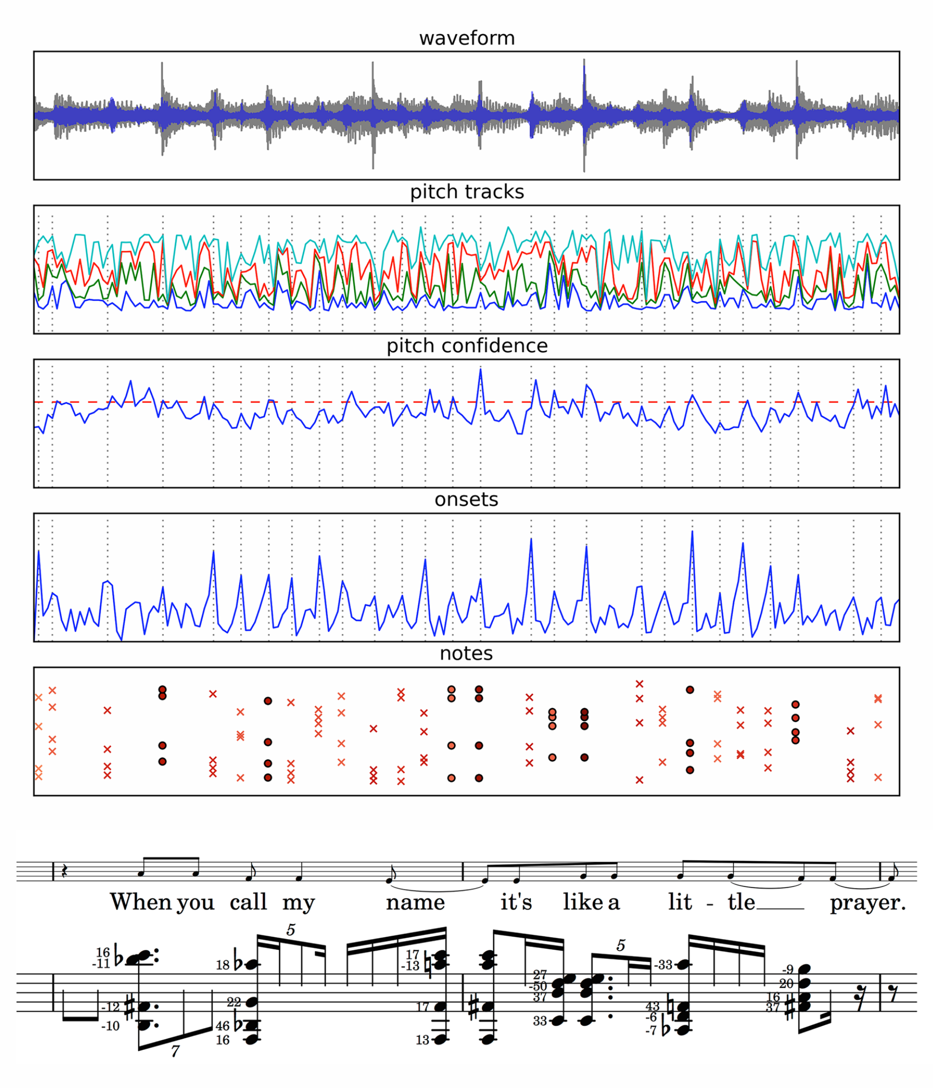


Figure 2.1: Stages of Happy Valley Band analysis.

2.1 Source Separation

Source separation is the task of identifying and extracting individual acoustic sources from a mixture of sources. The human auditory perception exhibits a remarkable capacity to recognize and isolate sound. Humans are able to hear, within a complex sonic mixture, individual elements — instruments, voices, or performers — and shift focus from one sound to another. How does human auditory perception identify and separate sound and and how might a computer algorithm mimic it?

In general, source separation is a very difficult problem. The difficulty is due, in part, to the complexities of the physical properties of acoustic mixing. Due to positive and destructive interference between sound waves, mixing is an information losing process. An ideal source separation technique would know how and when to put back missing or cancelled sound information. Furthermore, the perceptual mechanisms involved in parsing multiple sound sources are also complex, and, in many cases, what we hear is in part an illusion that is not physically present. An instrument may sound drastically different solo than in the mix. This happens when frequencies overlap and mask one another, and is well-known to recording engineers, who intentionally remove frequencies that would otherwise compete. How should we model the complexities of both the physical and perceptual domains, as well as the prior knowledge and expectations of listeners?

There are a variety of computational approaches to source separation. Some model the human auditory system in terms of a set of fundamental perceptual rules that hierarchically organize and group sound energy into sources. These generally fall under the category of Computational Auditory Scene Analysis (Wang and Brown 2006) and are based on Albert Bregman's pioneering work on Auditory Scene Analysis (Bregman 1994), which draws on expert knowledge developed over decades of empirical study. Other approaches, such as a more recent trend in machine-learning based algorithms, rely less on expert knowledge, but rather attempt to extrapolate and generalize patterns from data. As a result, these models need to be trained to produce results that correspond with human auditory perception. In my workflow, I use a number of different tools to separate mixtures, often in combination, depending on the particular production qualities of the acoustic mixture and of the instruments being separated, including spatial filtering and spectrogram decomposition.

2.2 Source Separation by Spatial Filtering

Humans perceive sound along a number of spatial dimensions — horizontal location (azimuth), vertical location (zenith), and depth, as well as acoustic qualities such as reverberation (dry versus wet) and width (wide versus narrow). These features, which function as perceptual cues to the nature of the acoustic environment, are captured in acoustic recording or fabricated by mixing and sound processing effects. *Spatial filtering* refers to a variety of techniques for filtering sound along these spatial dimensions. In particular, the problem of localization — estimating the location of a sound in space and isolating sounds at a given spatial location — has received considerable attention from researchers in sound, music, and acoustics as well as the sciences and engineering more broadly.

There exists a vast body of literature and diversity of approaches to sound localization, with different applications, engineering constraints, and problem formulations. Approaches tend to vary in terms of sound propagation model; audio features and spatial cues (interaural time difference, interaural intensity difference, spectral notches, spectral cues); number of signals/microphones (stereo and binaural versus array-based techniques); number of sources to localize (single versus multiple); and end results (location estimation, source identification, source separation (Rascon and Meza 2017). Different applications require different analysis techniques, but among the most common is beamforming, which uses timing difference between multiple microphones to localize and separate sounds.

In popular music production, however, the predominant localization cue is most often intensity differences, not timing differences. This is due to the nature of studio production techniques which generally rely on amplitude-based panning to place close-miked monaural sources in a stereo mix. Even the multi-microphone stereo recording techniques commonly employed in pop music production, such as coincident and near-coincident pairs or spaced microphones, tend to carry amplitude-based cues due to the use of directional microphones or microphone placement in close proximity to sound sources. Phase-based panning is less common in popular music studio production, and, when present, phase-based location cues, or time differences, often produce a sense of width or reverberation. These are often introduced as signal processing effects rather than captured acoustically via microphones.

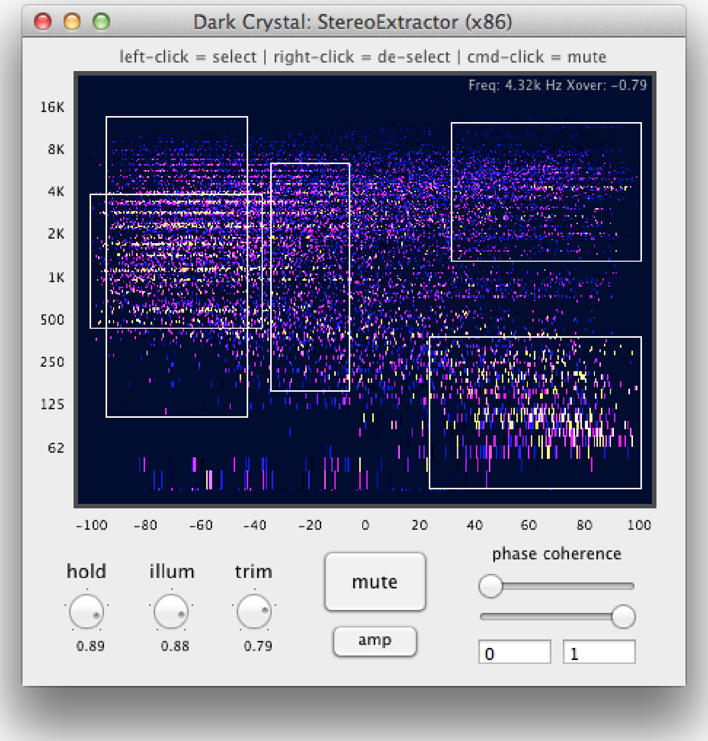
My approach to spatial filtering relies on both interaural intensity differences (IID) and interaural time differences (ITD). It is related to CASA systems that separate mixtures according to sound localization, such as Richard Lyon's binaural separation model (Lyon 1983), which groups audio components by ITD estimates from a cross-correlogram and was later extended by Markus Bodden (Bodden 1993) to include head-related transfer functions (HRTF) as well IID.

2.2.1 xtrk

xtrk is an audio plug-in that I built to implement spatial filtering in real-time.[1] xtrk visualizes the stereo image of an audio signal in a two-dimensional space — frequency is plotted along the vertical axis against estimated stereo location (azimuth) plotted along the horizontal axis — and allows the user to isolate or mute regions of the stereo image by drawing rectangular outlines, as shown in Fig. 2.2. xtrk has two modes, one for IID or amplitude-based estimation and another for ITD or phase-based estimation.

xtrk performs Short-time Fourier Transforms (STFT) of the left and right channels and estimates the perceived stereo location according to differences in phase and amplitude for each STFT time-frequency bin. The user-selected regions provide a set of frequency domain binary masks that are applied during iFFT resynthesis to selectively pass or attenuate time-frequency bins, effectively muting or soloing regions of the stereo image. In the following sections, I discuss the details of the amplitude-based and a phase-based masking.

[1] The UI design of xtrk is based on the Elevayta Extra Boy Pro plug-in by Paul R. Harvey.



Dark Crystal: StereoExtractor (x86)

left-click = select | right-click = de-select | cmd-click = mute

Freq: 4.32k Hz Xover: -0.79

hold 0.89

illum 0.88

trim 0.79

phase coherence 0 1

Figure 2.2: `xtrk` user interface.

2.2.2 Amplitude-based Masking

The perceived stereo location can be estimated from the amplitudes of the left and right signals by solving amplitude-based panning equations. Panning a monarual source is generally implemented by mixing more or less signal to one channel or the other of a stereo mix. A source is panned to the right by attenuating the signal mixed to the left channel and boosting the signal mixed to the right channel. How much to attenuate or boost is determined by a pair of *panning equations*, which represent the amplitudes of the left and right signals when panning a monaural source into a two-channel stereo field.

Perhaps the simplest form of amplitude panning is linear panning, given by the set of equations

$$\begin{aligned} x_l(t) &= x(t) \cdot \theta \\ x_r(t) &= x(t) \cdot (1-\theta) \end{aligned} \tag{2.1}$$

where $x(t)$ is the monaural source, $x_l(t)$ and $x_r(t)$ are the left and right signals, and $\theta \in [0,1]$ is the pan position expressed between 0 and 1. When estimating the perceived stereo location, the amplitudes of the left and right signals are given, and the location is estimated by solving the system of panning equations for unknowns θ and $x(t)$:

$$\begin{aligned} x(t) &= x_l(t) + x_r(t) \\ \theta &= \frac{x_l(t)}{x_l(t) + x_r(t)} \end{aligned} \tag{2.2}$$

Analyzing in the frequency domain, however, gives a separate panning estimate for each frequency component. Because θ is a scaler, or DC source, the frequency-domain representation of linear panning equations (2.1) simplifies to multiplication by a constant θ (rather than convolution) and is given by the set of equations

$$\begin{aligned} |X_l(\omega)| &= |X(\omega)| \cdot \theta(\omega) \\ |X_r(\omega)| &= |X(\omega)| \cdot (1-\theta(\omega)) \end{aligned} \tag{2.3}$$

where $X(\omega)$ is the Fourier Transform of the monaural source $x(t)$ with frequency denoted ω, $X_l(\omega)$ and $X_r(\omega)$ are the Fourier Transforms of the left and right signals, $|X_l(\omega)|$ denotes the magnitude spectrum, and the scalar θ is the panning position. The panning position θ is parametrized over frequency ω having a separate panning position for each Fourier component ω. The system of equations (2.3) can be solved for θ at each frequency component ω, giving an estimate of the panning position and amplitude of each Fourier component in terms of the amplitudes of the left and right signals:

$$\begin{aligned} |X(\omega)| &= |X_l(\omega)| + |X_r(\omega)| \\ \theta(\omega) &= \frac{|X_l(\omega)|}{|X_l(\omega)| + |X_r(\omega)|} \end{aligned} \tag{2.4}$$

In many audio signals, $X_l(\omega)$ and $X_r(\omega)$ represent a mixture of multiple sources panned in different locations, not a single source panned in one location. Equations (2.4) give only one amplitude and pan estimate for each Fourier component, which, in the case of sources overlapping in frequency, collapses multiple panned sources to one location estimate. Each source, however, contributes to $\theta(\omega)$ at most proportional to its amplitude in that Fourier component, depending on phase alignment. For time-frequency components that are dominated by amplitude of one source, the pan estimate $\theta(\omega)$ will be a good approximation. For time-frequency components that are not dominated by a single source — two or more are relatively balanced sources, for instance — the single pan estimate given by $\theta(\omega)$ will be a poor approximation. However, because the relative amplitudes of multiple sources often vary over time, an STFT frequency component may be dominated by one source at one time moment and another source at another time moment. As a result, with good temporal resolution (an STFT hop of 1024 samples at a sample rate of 44100), enough time-frequency bins are generally left unattenuated to maintain the perceptual integrity of the isolated source while degrading beyond recognition the intelligibility of the background sources. As currently implemented, `xtrk` does not group adjacent bins through time. The same binary mask is applied each frame, unless the user manually adjusts the selection regions. Grouping adjacent bins between hops could mitigate the effects of bins abruptly turning on or off over time, and could be accomplished with smoothing filters, hysteresis, or a Hidden Markov Model.

2.2.3 Phase-based Masking

As mentioned above, many common techniques for source localization and separation rely on the time difference(s) between two or more signals. In popular music, however, time difference is more often a cue of stereo width and reverberation rather than of precise spatial location — the result of stereo miking a single instrumental source or the application of delays and other time-based based effects. With `xtrk`, I use time differences to filter sounds based on the *phase coherence*, a measure of phase consistency across STFT frequency bins and through time, rather than localize and separate sources based on stereo location in the azimuth. The *phase error* between the left and right channels is measured at each STFT time-frequency component, and components that fall outside a threshold of phase coherence are attenuated. The phase error is computed by the difference between phases of the corresponding left and right Fourier components:

$$\Delta phase(\omega) = \angle X_r(\omega) - \angle X_l(\omega) \tag{2.5}$$

where $\angle X_l(\omega)$ and $\angle X_r(\omega)$ are the phases of the left and right Fourier Transforms at frequency component ω.

Components with nonzero phase differences indicate the presence of noise, reverberation, or stereo sources. Sounds with time-based effects, such as stereo delays, have nonzero phase error due to timing differences between the left and right channels. The presence of noise or reverberation in a signal causes phase differences to spread vertically across frequency bins, producing a wide phase portrait across the frequency spectrum. Limiting the phase coherence can effectively separate sounds with wide versus narrow phase portraits, isolating reverberant or stereo sources from a mix.

The rhythm guitar in Madonna's recording of *Like a Prayer*, for instance, has an extremely wide stereo image in the mix. The guitar is heard from both the far left and far right sides of the mix, due to a time delay between the two signals — likely the result of a stereo delay effect, stereo miking, or double tracking. The guitar is perceived as spread across the stereo field, rather than localized to a single point, and can be isolated by using phase coherence to filter narrow from wide phase portraits. Similarly, I use phase coherence in Black Sabbath's *War Pigs* to help isolate the singing voice from the ringing cymbals. The voice, which is a monoaural signal, has a tight phase portrait, where as the drums, recorded using stereo multi-microphone techniques, has wider phase errors, especially in the cymbals.

As with amplitude-based masking, in the presence of multiple sources that overlap in frequency, $\Delta phase(\omega)$ gives only one phase estimate per frequency component, collapsing multiple sources to a single value. Similarly, since sources contribute to phase proportional to their relative amplitudes, frequency components that are dominated by a single source will give good estimates, and often applying the mask through time is sufficient to degrade the background sources beyond recognition while maintaining intelligibility of the intended source.

2.2.4 Related Work

My phase-based filtering approach is related to the use of the Generalized Correlation Coefficient (GCC) (Knapp and Carter 1976) and frequency domain masking approaches in beamforming. One of the difficulties of beamforming is estimating of the Time Difference of Arrival (TDOA) between microphone signals of the intended source. TDOA is usually estimated using a measure of cross correlation, such as the Pearson Correlation Coefficient, although this suffers from distortions due to noise, reverberation, and the presence of multiple sources. The Generalized Correlation Coefficient is a frequency domain extension that mitigates these errors by weighting different frequency regions more or less heavily. The GCC-Phat, in particular, normalizes out amplitude information, leaving only phase information, as given by the equation

$$\text{GCC-PHAT}(\omega) = \frac{X_l(\omega)X_r(\omega)^*}{|X_l(\omega)X_r(\omega)^*|} \tag{2.6}$$

where $X_r(\omega)^*$ denotes the complex conjugate. The denominator normalizes out the magnitude information, leaving just the phase difference for each Fourier component. A number of authors have proposed frequency domain masking (see D. Wang 2005 for a discussion), and Arabi and Shi (Aarabi and Shi 2004) in particular, derive, from the GCC-Phat, a similar phase-based masking technique of punishing STFT time-frequency bins based on the phase error.

2.2.5 Discussion

I use spatial filtering to separate sounds that can be isolated according to stereo location or by spatial features such as width or reverberation. Spatial filtering can be very effective because mix engineers tend to place instruments that have similar frequency profiles in different spatial locations to increase mix clarity, or use spatial effects such as reverb and delay to give instruments unique spatial footprints. Spatial filtering tends to work best on older mixes, dating from the 1960s and 1970s, when hard panning was popular. Most mixes, especially more recently, feature sounds that overlap in both frequency and spatial location, requiring more sophisticated approaches to source separation.

2.3 Source Separation by Probabilistic Latent Component Analysis

For mixtures that are more complex, containing sources overlapping in both frequency and spatial location, I use a spectrogram decomposition technique, Probabilistic Latent Component Analysis (PLCA) (Smaragdis 2007). PLCA decomposes the short-time magnitude spectrum into a set of basis functions and time activations. The basis functions represent the spectra of individual sources in the mixture and the corresponding time activations represent the temporal locations and amplitudes of their occurrences over time. The basis functions recombine according to the temporal locations and amplitudes of the activation weights to reconstruct an approximate of the original spectrogram.

PLCA is a form of matrix factorization. The magnitude spectrum is represented as a matrix **S** and expressed as the product of two lower dimensional matrices

$$\mathbf{S} \approx \mathbf{W} \cdot \mathbf{H} \tag{2.7}$$

where the columns of **W** represent spectral basis functions and the rows of **H** represent the corresponding activations of each basis function over time.

Many algorithms have been used to find the bases **W** and activations **H**, and all produce different perceptual results due to the variety mathematical constraints assumed. Bases found with Principal Component Analysis (PCA), for instance, generally do not exhibit substantial separation because the algorithm does not impose perceptually informed constraints. Independent Component Analysis (ICA), enforces statistical independence between components, producing basis functions that are more perceptually distinct. The assumption of statistical independence, however, is often too strong to model acoustic mixtures, and Sparse Component Analysis (SCA) instead finds good separation by enforcing sparsity — commonly measured using the L0 norm which counts the number of zero or near zero entries in the basis vector. PCA, ICA, and SCA, however, all allow for negative values, which in the case of magnitude spectra, are not perceptually meaningful and lead to noise and other artifacts. Non-negative Matrix Factorization (NMF) improves upon this by restricting basis functions to positive values only, giving a more acoustically and perceptually meaningful separation.

PLCA formulates the spectrogram decomposition problem in a probabilistic framework, treating the magnitude spectra as a distribution or histogram across the dimensions of frequency and time. In addition to enforcing positivity, this probabilistic formulation allows for extensions to learning frameworks, providing for musically, acoustically, and perceptually meaningful constraints and transformations, such as sparsity, and transpositional invariances in both the dimensions of frequency and time.

PLCA is an example of a recent trend of machine-learning approaches to sources separation, which parallels a rise more generally of machine-learning applications in artificial intelligence. Earlier techniques, such as Computational Auditory Scene Analysis (CASA), a general category of approaches to source separation that were previously popular, identify basic perceptual mechanisms of the human auditory system that operate to form high level representations of the world. Grounded in physiology, perception, and empirical study, CASA systems are designed to model the perpetual rules and mechanisms of the human auditory system that govern how we hierarchically organize auditory phenomena. Machine-learning based approaches, however, are motivated by the assumption that the laws, both perceptual and physical, that govern acoustic mixing and perception can be extrapolated from example data. As such, this knowledge does not need to be explicitly programmed into a source separation algorithm but can be learned. Importantly, machine-learning techniques, such as PLCA, need to be trained to produce results that correspond with human perception.

2.3.1 The PLCA Model

PLCA models the magnitude spectrogram as a probability distribution, or histogram. The magnitude spectrogram S is interpreted as a two-dimensional probability distribution $P(f,t)$ and expressed as the product of its two marginal distributions. The marginals, one over frequency $P(f)$ and the other over time $P(t)$, provide the basis functions $\mathbf{W}$ and activations functions $\mathbf{H}$ of the spectrogram decomposition. They are computed by integrating out (or summing over) the other dimension:

$$P(f) = \int P(f,t)\,dt \quad \text{and} \quad P(t) = \int P(f,t)\,df \tag{2.8}$$

For one source, the marginals are equivalent to the power spectrum and amplitude envelope of the original mixture signal.

While a single pair of time and frequency marginals are not a particularly meaningful decomposition for the purposes of source separation, multiple pairs of time and frequency marginals can be extracted by introducing a latent variable z. The latent variable model represents $P(\mathbf{x})$ as the sum of multiple component distributions, indexed by the value of the latent variable z. Each component distribution is similarly expressed as the product of its two marginal distributions, one over time and the other over frequency. This gives a set of frequency marginals, or basis functions, for $\mathbf{W}$ and a set of time marginals, or activation functions, for $\mathbf{H}$. The general form of the latent variable model Probabilistic Latent Component Analysis is

$$P(\mathbf{x}) = \sum_z P(z) \prod_{j=1}^{N} P(x_j|z) \tag{2.9}$$

where $P(\mathbf{x})$ is an N-dimensional distribution of the random variable $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, the latent variable z is a discrete variable taking on integer values $\{1, 2, \ldots, n\}$ up to the number of latent components n, and the marginals $P(x_j|z)\ \forall j \in N$ are one-dimensional distributions across each of x's dimensions.

PLCA is an optimization algorithm. PLCA finds the marginals $P(x_j|z)$ and latent weights $P(z)$ that best approximate $P(\mathbf{x})$ using an Expectation-Maximization algorithm (EM), an iterative method that alternatives between expectation (E) and maximization (M) steps. In the expectation step, the relative contribution of each value of the latent variable z is estimated, normalized by the sum total contribution over all values of z:

$$R(\mathbf{x}, z) = \frac{P(z) \prod_{j=1}^{N} P(x_j|z)}{\sum_{z'} P(z') \prod_{j=1}^{N} P(x_j|z')} \tag{2.10}$$

This gives a collection of time-frequency distributions in which the value at each time-frequency location is the relative amount of observed energy contributed by that component distribution.

In the maximization step, the marginals $P(x_j|z)$ and latent variable distribution $P(z)$ are reestimated to maximize the contribution weightings found in the expectation step. This is done by multiplying the observed distribution $P(\mathbf{x})$ by the relative contributions $R(\mathbf{x}, z)$, giving the amount of observed energy contributed by each value of z. The latent distribution $P(z)$ is reestimated to be the total amount of observed energy contributed, computed by integrating across all dimensions, and the marginals $P(x_j|z)$ are reestimated by integrating over all but the desired dimension:

$$P(z) = \int P(\mathbf{x})R(\mathbf{x}, z)\, d\mathbf{x} \tag{2.11}$$

$$P(x_j|z) = \frac{\int \dots \int P(\mathbf{x})R(\mathbf{x}, z)\, dx_i}{P(z)} \; \forall i \in N, i \neq j \tag{2.12}$$

This has the effect of updating each of the latent marginal $P(x_j|z)$ to account for an amount of $P(\mathbf{x})$ relative to the marginal's prior contribution. The E and M steps are successively repeated over and over until either reaching a convergence thresholds or exhausting a given number of iterations.

Applied to magnitude spectra, $\mathbf{x}$ is a two-dimensional distribution, x_1 across frequency and x_2 across time. The marginals $P(x_1|z)$ and $P(x_2|z)$ are a collection of multiple frequency distributions and time distributions, corresponding to different values of the latent variable z. The multiple frequency and time distributions provide the basis and activation functions of the spectrogram decomposition. Effectively, PLCA finds a set of spectral kernels that reoccur through time to construct $P(\mathbf{x})$.

Because PLCA is a matrix factorization technique, the model can be applied to any positively valued matrix, regardless of what spectral transform the data represents. STFT spectrum, Constant-Q spectrum, and Mel-Frequency spectrum are all commonly used. However, since PLCA operates only on the magnitude spectra, discarding phase and phase delta information, the model does not account for destructive interference in mixing, a limitation which leads to analysis and resynthesis artifacts.

2.3.2 Application to Musical Source Separation

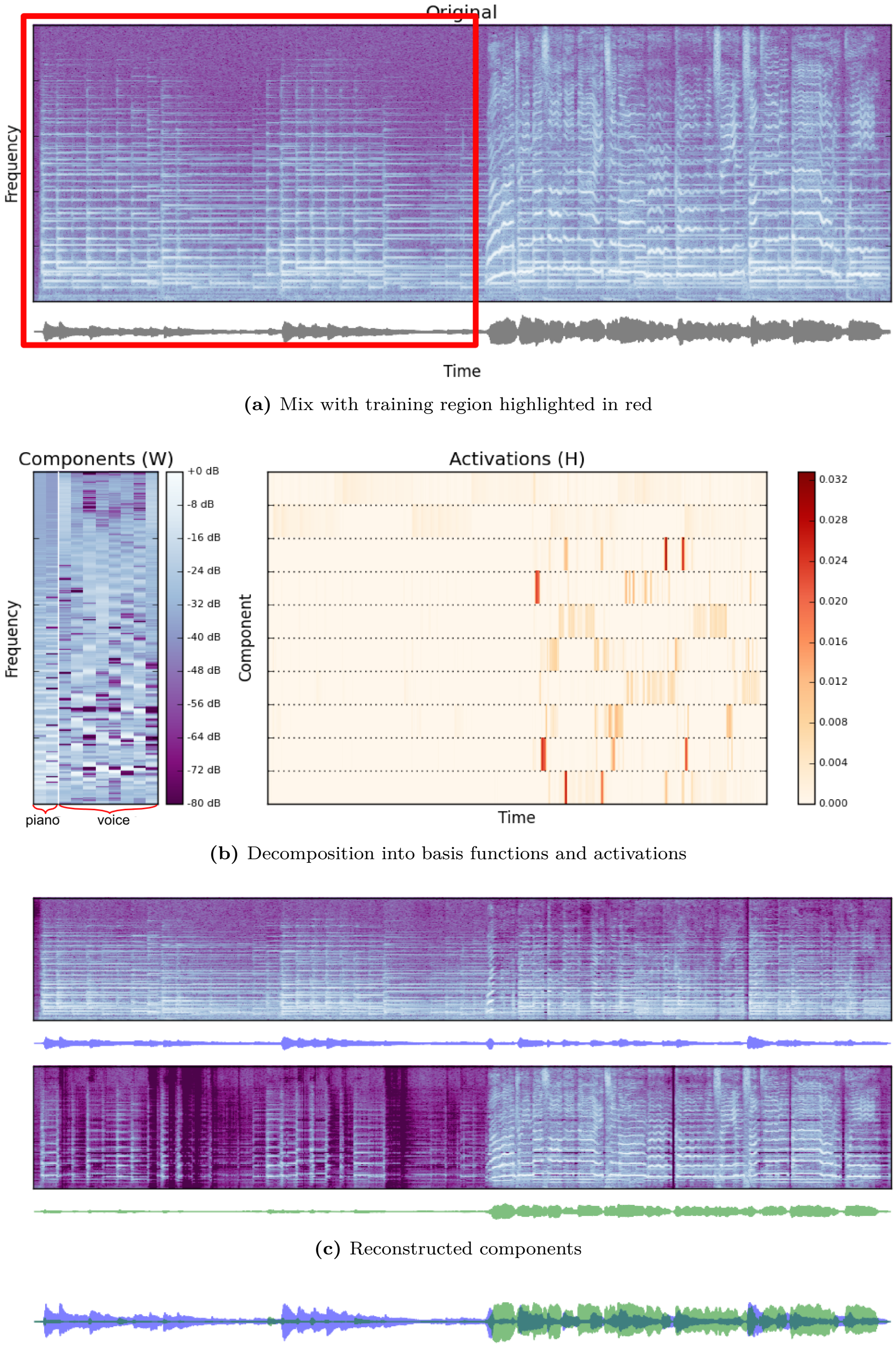
PLCA can be used for musical source separation in a number of different ways, including unsupervised, supervised, and semi-supervised learning. In unsupervised learning, none of the sources are known in advance, and the goal is usually to extract basis functions that match each of the sources. As with NMF, unsupervised separation can be successful, but PLCA parameters, such as the number of components and sparsity, must be chosen carefully to match the sources. In supervised learning, all of the sources are known in advance, and basis functions are derived from tagged examples rather than from the mixture itself. During the training phase, basis functions are learned from isolated audio clips of the known sources. The learned basis functions are then used to decompose, or fit, the mixture, giving new activation functions that reconstruct the mixture according to the learned basis functions. Multiplying the pre-learned basis functions by the new activations resynthesizes each source component.

Generally, I learn multiple basis functions per source, anywhere from $20-80$ components each. Not much audio is necessary for training. A few seconds or even less can be effective, depending on the amount of variation of each source in the mixture and in the training audio. Supervised learning is based on the assumption that learned basis functions can adequately describe new instance of sound from that source and its success depends on the amount of variation and extent to which the model is required to generalize. Too few components and the model will not generalize and account for variation in a source; too many components and the learned basis functions will be too fine, such that components of one source may easily fit another source, causing poor separation.

Most often, I use PLCA in a semi-supervised context, in which one or more sources are known in advance, but others are unknown. The unknown source may represent a single instrument (such as a voice or instrument) or an entire ensemble of sources (such as the entire band minus the voice). The target source to isolate may be the known source, unknown source, or both. I learn multiple sets of basis functions, one set for each known source and another set for the unknown source. The known sets are learned during training phase but kept fixed during fitting. The unknown set is learned during fitting.

Figure 2.3 shows a mixture of two sources, voice and piano, from Neil Young's recording of *After the Gold Rush*. The first half (segment 1) of the clip contains piano solo, and the second half (segment 2) contains a mixture of piano and voice. I first learn basis functions for the piano by training on segment 1 (outlined in red), and then fit the entire clip (segments 1 and 2 together) using the piano basis functions together with new, untrained basis functions. The piano bases are held constant (not updated), since they should do a good job describing the piano, and the untrained bases are updated to describe the remaining source, the voice. The piano is reconstructed using the piano components, and the voice is reconstructed using voice components. Spectrograms are shown in (c) and inverted back to time domain waveforms. The separated components are overlaid in (d), illustrating that PLCA achieves very good separation, about 20dB between the desired source and background in each reconstruction, even though the sources overlap in time, frequency, and stereo location.

I use two components for the piano, and six for the voice. The basis functions represent spectral components of each source. If too few piano components are used, the basis functions will be too constrained to described the variety of piano



(a) Mix with training region highlighted in red

(b) Decomposition into basis functions and activations

(c) Reconstructed components

(d) Reconstructed component waveforms overlay

Figure 2.3: PLCA *After the Gold Rush* voice and piano separation.

sound in the mixture. As a result, some piano sound will be accounted for by the voice components instead, causing piano to bleed into the voice. If too many components are used, the basis functions will be too general, or too fine, and some piano basis functions will fit aspects of the voice, causing voice to bleed into the piano reconstruction as well as producing spectral holes in the isolated voice.

2.3.3 Discussion

To achieve good separation, there are a number of analysis parameters that need to be tuned. STFT parameters — primarily hop size, window size, and FFT size — must be selected appropriately to capture relevant spectral features. There must be sufficient temporal and frequency resolution to represent distinct sources as distinct spectral features, otherwise the spectrogram cannot be decomposed into separate sources. Good resolution is generally preferred, although coarse time resolution can smooth percussive transients and coarse frequency resolution can help eliminate low frequency bleed. With PLCA, the number of components per each source is critical, and the appropriate number depends largely on the sources, their spectral similarity, the training clips available, and amount of sonic variation of each source in the mixture. The numbers of components of each source relative to the others is also critical. Adjusting sources to have greater or fewer sources relative to another effectively pushes or pulls information from one source to another. I parameterize the number of components in terms of two values: the relative number of components per each source multiplied by a resolution scalar.

PLCA can be very effective, especially when clean training data is available. There is always some bleed between sounds, or holes within sounds, but these artifacts are, in part, due to the physical and perceptual complexities of the mixing, as well as the limitations of PLCA. In many ways, these artifacts reflect the complexities of the physical and perceptual processes involved in hearing. As I worked more and more with source separation algorithms, I became interested in the question, is there some way to translate these artifacts back into music? The remaining steps of the HVB process — pitch and rhythm analysis and music notation — are my attempt to express my fascination with source separation into the form of music notation and performance.

2.4 Pitch Analysis

Pitch analysis, often referred to as (multiple) *fundamental frequency estimation*, is the task of estimating, from a waveform, the pitch, or set of multiple pitches, that a human listener would perceive. Although pitch can be explained as the perceptual feature of sound that corresponds with the physical feature of waveform frequency, pitch perception is a complex phenomena that is not well defined. Most waveforms, for instance, do not repeat exactly, yet humans still experience the perception of pitch, to a greater or less extent, and, in the case of complex, aperiodic waveforms, pitch perception can be highly subjective. There are many approaches to pitch estimation, some mathematically motivated, some physiologically motivated, and, more recently, others data-driven using machine learning algorithms.

2.4.1 Maximum Likelihood Pitch Estimation

My approach to multiple fundamental frequency estimation is a maximum likelihood method based on `[fiddle~]` (Puckette, Apel, and Zicarelli 1998), in which the salience of a fundamental frequency is estimated according to the presence of peaks in the frequency spectrum at or near harmonics of that fundamental. The algorithm design is motivated by the perceptual idea that harmonics reinforce the perception of pitch. I chose this algorithm for a few reasons: I am interested in the compositional controls that the parameters of the maximum likelihood estimator afford, but, more importantly, I am interested in the harmonic series as a structuring concept in twentieth century music.

The spectrum is first reduced to a limited set of peak frequencies, after performing an STFT transform with phase vocoder to estimate the instantaneous frequency, which are then used to compute a likelihood function for possible fundamental frequencies. The probability that a given frequency f is a fundamental is computed by the likelihood function

$$L(f) = \sum_{i=0}^{k} A(a_i)\Delta(t_i)H(n_i) \tag{2.13}$$

where f is the frequency candidate, k is the number of peaks in the spectrum, $A(a_i)$ is a function that depends on the amplitude of the i^{th} peak, $\Delta(t_i)$ is a function that depends on the distance of the i^{th} peak to the nearest harmonic of f, and $H(n_i)$ is a function that depends on the order of that nearest harmonic — whether it is a low or high multiple of f. $A(a_i)$ normalizes by the sum total amplitude within an STFT frame, $\Delta(t_i)$ normalizes according to a maximum bandwidth for inclusion, outside of which peaks are discounted or zeroed, and $H(n_i)$ weights harmonics inversely on an power scale $(\frac{1}{n_i})^p$, preferencing lower harmonics. In summary, $L(f)$ is the sum total amplitude of peaks that fall within a bandwidth of inclusion from harmonics of f, and peaks are weighted less strongly the further they are from actual harmonics and the higher they are into the harmonic series.

To find multiple fundamentals, $L(f)$ is computed over set of possible fundamental frequencies, spanning, in quarter tone steps, a given range from minimum to maximum fundamental. The frequencies are ranked according to likelihood score $L(f)$ from highest to lowest, and the top n are considered multiple fundamentals. Finally, because frequency candidates are measured at quarter tone intervals, once a candidate is determined to be a fundamental, a more precise estimate is found using a weighted least squares linear regression

$$WAx \approx Wb \tag{2.14}$$

where b is a matrix of spectrogram peaks, A is a matrix of the orders (integer multipliers) of their nearest harmonics, and W is a matrix of corresponding weights for amplitude and harmonic order as in $L(f)$. Solving for x gives a more precise estimate of the fundamental by minimizing the distance between the found spectrogram peaks and their nearest harmonics. Finally, when analyzing for multiple fundamental frequencies, it is necessary to connect pitch changes over time to prevent false onsets or offsets. Pitches are connected between adjacent analysis frames to form tracks using a nearest neighbor model that allows for the termination and creation of new tracks. Figure 2.4 illustrates the sequence of steps involved in pitch estimation.

2.4.2 Compositional Parameters: Harmonicity

A challenge of this approach, however, is that harmonics of salient fundamental frequency candidates tend to also have high likelihood scores, producing harmonic duplicates and octave jumps from analysis frame to frame. Fundamental frequency candidates that are in harmonic ratio with one another are suppressed, returning only the top candidate (highest likelihood) of the group. Parameters specifying which harmonic ratios are to be suppressed, the bandwidth for inclusion, as well

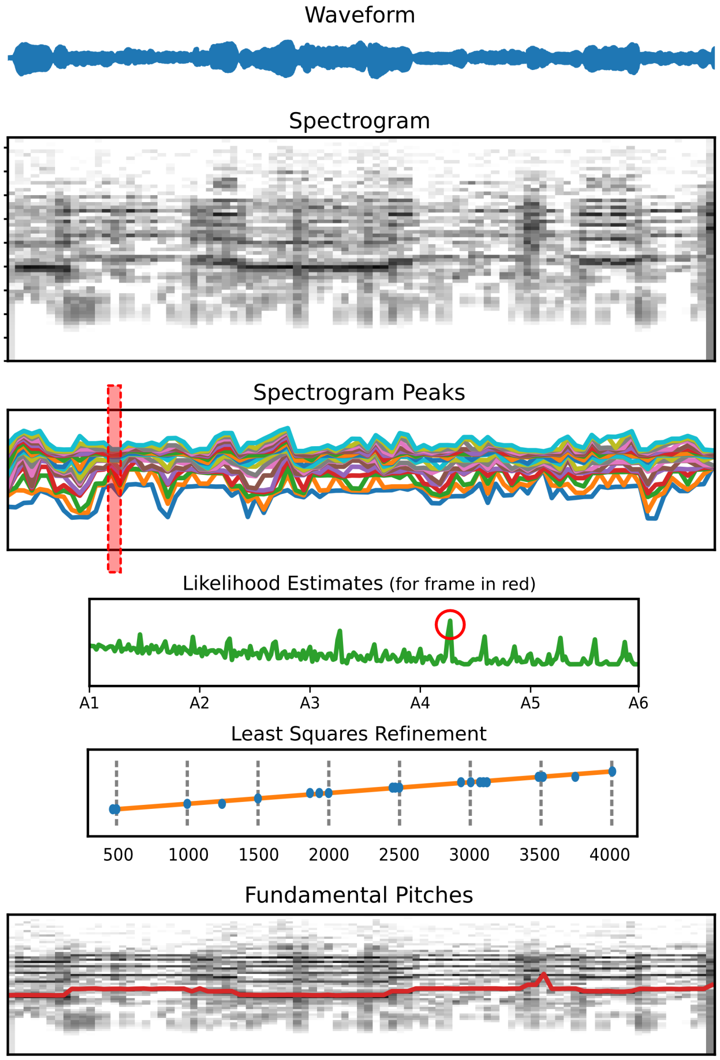
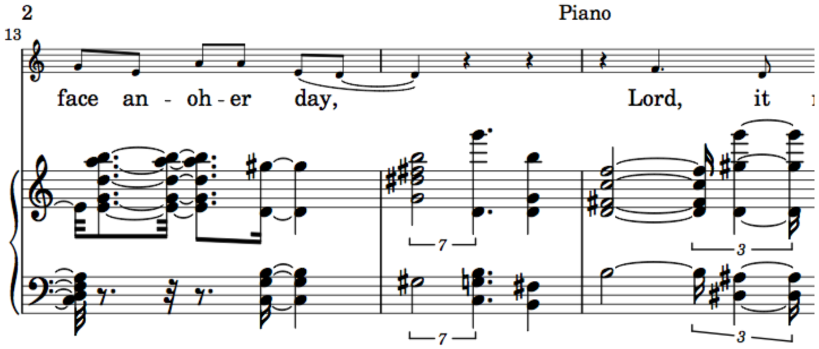


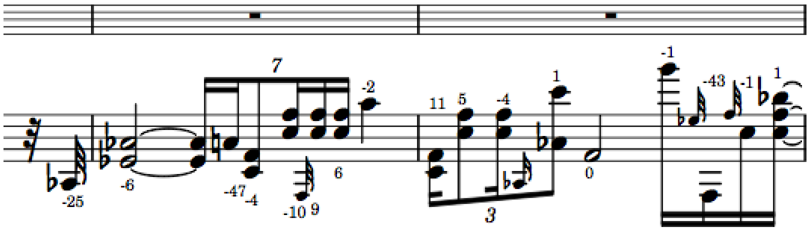
Figure 2.4: Pitch estimation signal flow.

as how far back in time to look for duplicates — within a single frame or over multiple frames — are made variable, controlling of the *harmonicity* of the set of multiple fundamentals. These harmonicity parameters have a substantial impact on the resulting music, and I tend to tune them from song to song, to reflect different musical and sonic aspects of the original songs.

In my analysis of (*You Make Me Feel Like) A Natural Woman* (Excerpt 2.1), I suppressed harmonic relationships heavily, because I wanted the harmony to be rich with complex ratios — I also thought it a good opportunity to dig into the inharmonic aspects of the piano spectra. By contrast, in the transcription of Led Zeppelin's *When the Levee Breaks* (Excerpt 2.2), I did not suppress any harmonic relationships, because I wanted the transcription to capture the ringing quality of the droning bass guitar, harmonica, and overdriven lapsteel. The resulting transcription is rich in just-intoned thirds ands sevenths. In many ways I want pitch estimation to reflect not just the perceived fundamentals but also the spectral qualities of sound. I do so because I believe pitch perception and sound spectra are deeply interrelated phenomena and not separable as the conventional "everything else" definition of timbre would suggest.



Excerpt 2.1: *(You Make Me Feel Like) A Natural Woman*, Piano



Excerpt 2.2: *When the Levee Breaks*, Lapsteel

2.4.3 Related Work

Among the many different approaches to pitch estimation are time-domain methods such as zero crossing rate and autocorrelation; autocorrelation based methods such as YIN (Cheveigné and Kawahara 2002) and pYIN (Mauch and Dixon 2014), which include additional processing and heuristics and are generally considered among the state of the art for monophonic pitch estimation; physiologically motivated models such as Klapuri's method (Klapuri 2005), which combines a gammatone filterbank model of the human auditory periphery together with a periodicity analysis stage; and, more recently, corpus driven deep learning algorithms such as CREPE (Wook Kim et al. 2018), a convolutional neural network that operates on time-domain signals. My algorithm belongs to a general class of maximum likelihood estimators, which attempt to find periodicities of peaks in the frequency spectrum. These harmonic analysis based approaches are related to earlier work such as harmonic product spectrum, cepstral pitch determination, and maximum likelihood estimate as described by Noll (Noll 1970), although many more recent authors have developed their own variants, including Puckette's `[fiddle~]`, Klapuri's summation of harmonic amplitudes (Klapuri 2006), Doval and Rodet's maximum likelihood estimator (Doval and Rodet 1991). I chose a harmonic analysis based approach because I am interested in the concept of harmonic structure and its relation to pitch perception.

2.5 Rhythm Analysis

Music is often represented, notated, and heard as a sequence of events over time. These events may be articulated by changes in pitch, spectra, loudness, playing techniques, and other features of sound and musical performance. *Onset/Offset detection* is the process of finding the locations of these events within an audio signal — both the start and end times — as a human listener would perceive them and pertains to how the human auditory perception groups spans of time at a primary level of temporal organization.

2.5.1 Onset and Offset Detection

The perception of onsets and offsets is generally understood to correspond with change over time, and onset detection strategies usually operate by measuring change within an audio signal, or instantaneous difference, from one frame to the next, to find moments of significant change. Onset detection strategies generally involve identifying local peaks in an *onset detection function* (ODF), a time varying audio feature known or expected to correspond with the perceptual onset of musical events. The audio features, however, that serve as perceptual cues for onsets and offsets can be different for different sound sources, because different instruments have different spectral envelopes, especially during the start of a note. As such, a variety of detection functions have been used, including the time-domain amplitude envelope, change in spectra, change in a particular spectral feature, physiologically motivated features (Klapuri 1999, Collins 2005a), probabilistic measures of entropy or surprise, aspects of synthesis models such as spectral modeling synthesis residual, as well as estimated pitch (see Bello et al. 2005 for a comparison of different approaches). In the literature on onset/offset detection, offset detection has received less attention. The end of a musical event may be characterized simply by the beginning of the next event, or by a decrease in amplitude. When analyzing acoustic mixtures or noisy signals, however, these cues can be difficult to identify, and additional features such as pitch confidence can help locate moments when no musical event is present.

For HVB transcription, I choose ODFs to match the character of the music being analyzed — generally using either amplitude envelope, spectral flux, phase deviation, or pitch estimation. I also tend to use multiple ODFs simultaneously, detecting "percussive" onsets through changes in either amplitude or spectra as well as "pitched" onsets through changes in pitch estimation, and balancing between the two. This balance gives me a control that can radically alter the character of the transcription, essentially focussing on different perceptual cues, or features of the sound. I tend to find offsets simply by connecting adjacent onsets, or using an amplitude envelope threshold. When the signal falls below the threshold an offset is triggered. In the following sections, I discuss the various detection functions used and how onsets are interpreted into musical events.

2.5.2 Onset/Offset Detection Functions

A perceptually and computationally simple detection function is the *amplitude envelope*, which corresponds with the perception of loudness, and is measured by the signal amplitude averaged over time. The signal is analyzed in frames, within which it is windowed, rectified, and summed to give the amplitude envelope

$$env(n) = \frac{1}{N} \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} |x(n+m)|w(m) \tag{2.15}$$

where n is the frame index, $x(t)$ is the time-domain signal, $w(m)$ is a window function zero centered, and N is the window size. The amplitude envelope is effective for identifying onsets that are marked by abrupt increases in loudness, such as percussive sounds, or pitched instruments marked by distinct attacks, such strummed bass or guitar. (See Figure 2.5, orange plot).

Frequency-domain detection functions, which measure spectral difference in either the magnitude, phase, or complex domains, are more commonly used and considered state of the art within music information retrieval (MIR) community. *Spectral flux* measures the element-wise difference between magnitude spectra of successive STFT frames. Essentially, STFT frames are interpreted as high-dimensional vectors and any number of vector space distance metrics can be applied. Generally, spectral flux refers to difference in magnitude, not phase. I use the L1 norm of the half-wave rectified difference vector, giving the spectral flux

$$flux(n) = \sum_{k=0}^{\frac{N}{2}-1} H(|X_k(n)| - |X_k(n-1)|) \tag{2.16}$$

where $X_k(n)$ is the STFT of the time-domain signal $x(t)$ at frequency bin index k and frame index n, and $H(x)$ is the half-wave rectify function. The half-wave rectify function passes only positive, or increasing, bins and discounts the negative, or decreasing, bins. Because the spectra are differentiated before they are summed, spectral flux is sensitive to changes in the distribution of spectra, not just changes in the sum total amplitude. Spectral flux can identify changes in timbre or pitch that are not articulated by the amplitude attack envelopes, such as changes in bowed violin notes or onsets that are marked by noise or other spectral change. (See Figure 2.5, green plot).

Additionally, spectral change can be measured in the phase domain. *Phase* *deviation* measures the difference in instantaneous frequency, given by the second difference in phase between STFT frames (instantaneous frequency itself is given by the first difference)

$$pdev(n) = \sum_{k=0}^{\frac{N}{2}-1} |\Delta^2 \varphi_k(n)| \tag{2.17}$$

where $\varphi_k(n)$ is the 2π unwrapped phase of $X_k(n)$, and $\Delta^2 = \varphi_k(n-2) - 2\varphi_k(n-1) + \varphi_k(n)$ is the second difference. Peaks in phase deviation correspond to deviations from steady state spectra, regardless of intensity. These often occur during noisy attack transients and transitions between harmonically related pitched notes. (See Figure 2.5, red plot).

It can be difficult to generalize about the kinds of musical events captured by magnitude- versus phase-based ODFs, because magnitude and phase work together as complex domain Fourier coefficients to express a signal. Changes can be reflected in magnitude, phase, or both features, depending on how the signal lines up in analysis bins, both vertically across frequency and horizontally across time. Spectral flux, for example, will capture changes in pitch distance large enough to register across frequency bins, where as phase deviation will capture change within frequency bins. Because of this, *complex difference* is often used, which combines both spectral flux (magnitude) and phase deviation (phase) into one measure (see Bello et al. 2004 and Dixon 2006 for a discussion of phase and complex-domain onset detection functions).

In addition to amplitude envelope and spectral change, I use pitch as an onset feature, in a manner similar to `[fiddle~]` (Puckette, Apel, and Zicarelli 1998) or Nick Collin's pitch detector method (Collins 2005b), to identify onsets marked by stabilization of estimated pitch, rather than changes in spectra. An onset is found when a change in pitch stabilizes, or remains within a window of its center pitch for a certain amount of time. The center pitch is the average within a running window, from the last change in pitch up through the current candidate, and is measured by the mean log scale pitch distance. While other spectral onset features, such as phase deviation and, to a lesser extent, spectral flux, are also sensitive to pitched change and stabilization, using pitch estimation as an onset features gives a more direct parameterization that is useful for tuning the onset detector to musical sounds. The magnitude and time thresholds determine the onset detector's sensitivity to small or short-lived changes, features which are generally designed to suppress vibrato and reject spurious discontinuities in pitch tracking, and function, in the case of HVB transcription, as controls, tuning the onset detector's sensitivity to minute changes. (See Figure 2.5, purple plot).

Figure 2.5 compares onset detection functions of viola and piano, both playing ascending sequences of four notes. Four discrete note events are easily seen in the spectrogram, as characterized by subsequent vertical shifts in spectra, and are marked by broadband energy at onsets. The viola (left) does not exhibit a substantial change in amplitude envelope (orange), yet onsets correspond with distinct peaks in spectral flux (green) due to the shift in amplitude across bins. Phase deviation (red) is high at note onsets as well as at the decay of each note, making it difficult to distinguish onsets by phase deviation alone — the increase in phase deviation in the latter half of the note reflects the presence of vibrato. By contrast, the piano (right) does exhibit distinct rises in amplitude envelope (orange) at note onsets due to the percussive envelope of the piano. This is also reflected in the spectral flux (green). Phase deviation (red) also exhibits peaks at note onsets, due to the initial noise transient, although the peaks are less distinct, making them difficult to distinguish from false positives. For both sources, changes in pitch estimation are quite clear, changing abruptly at note onsets and remaining relatively stable throughout the duration of the note.

The audio clips in Figure 2.5 were synthesized for the purpose of providing clear examples, free of noise and performance distortions. Figure 2.6 shows ODFs of a human performance, a violin excerpt from Karen Dalton's *Katie Cruel*, after source separated from the mix. As can be seen from the plots, peaks in the ODFs are less distinct, a result of many factors, including noise and artifacts of the separation process as well as aspects of human performance — the violin playing is dynamic and expressive, featuring changes in loudness, vibrato, pitch bends, slurs (notes not articulated by separate bow strokes), and timbral inflections. All of these factors complicate onset detection, many in ways that bleed across ODFs.

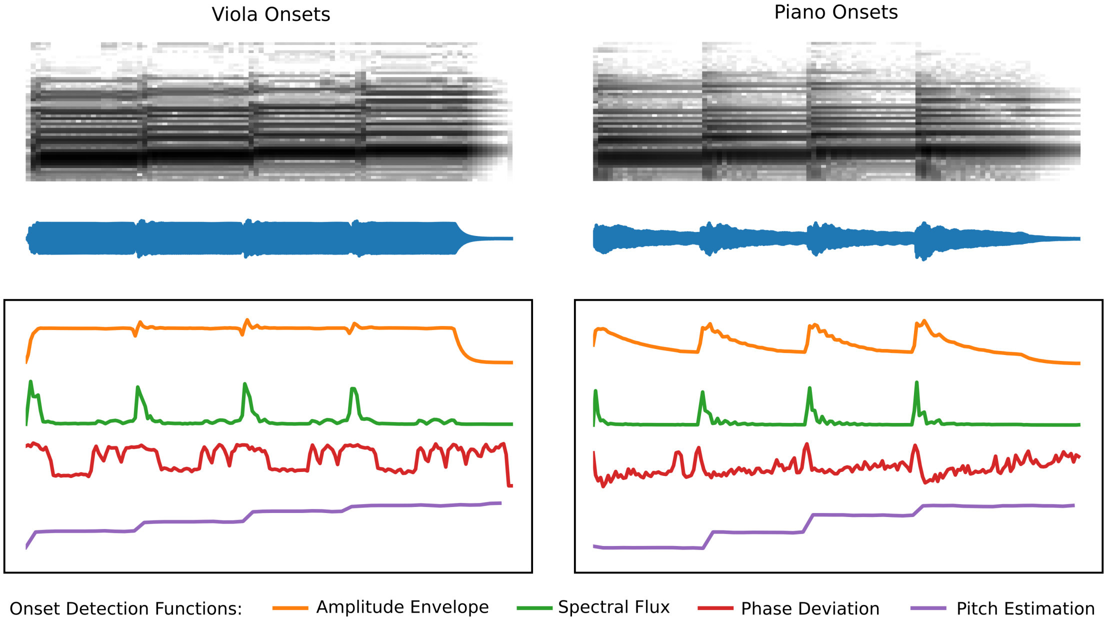


Figure 2.5: Onset Detection Functions for Viola and Piano.

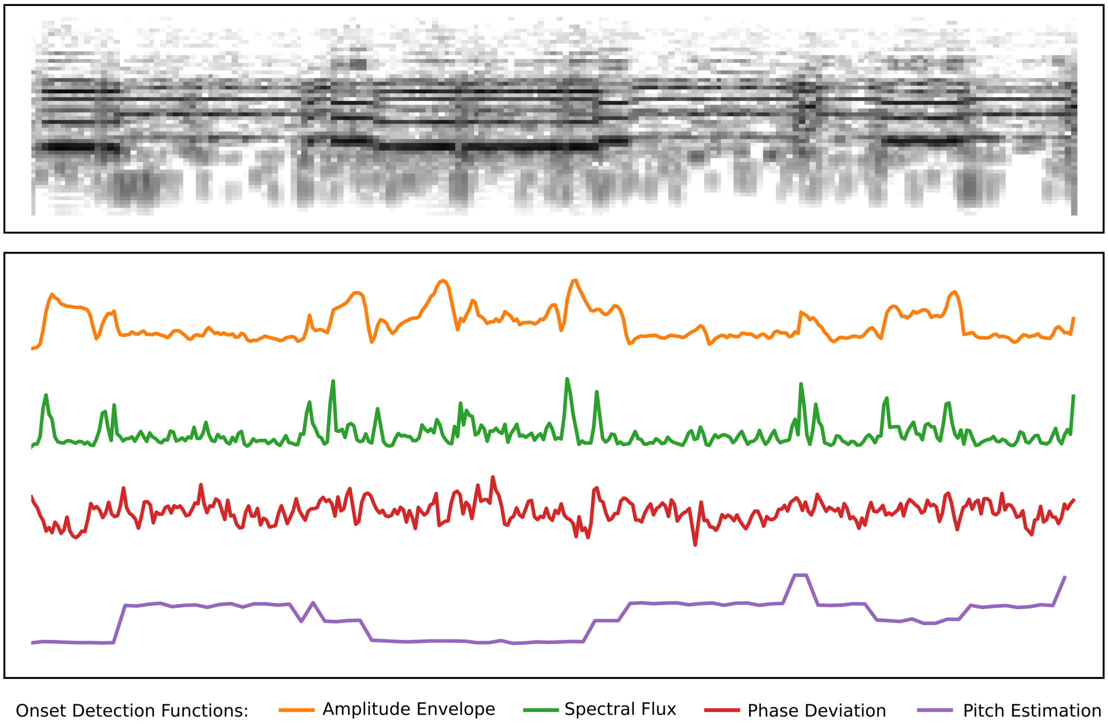


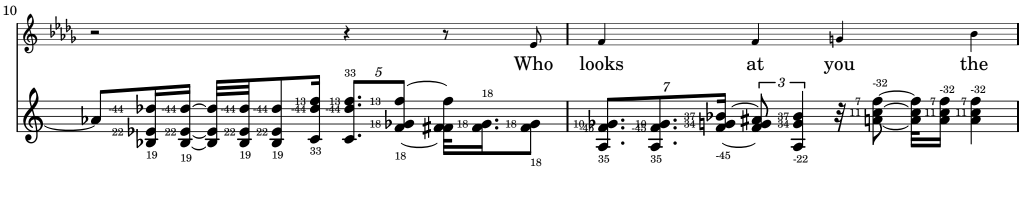
Figure 2.6: Multiple Onset Detection Functions HVB excerpt.

2.5.3 Mapping Analysis Features to Performance Features

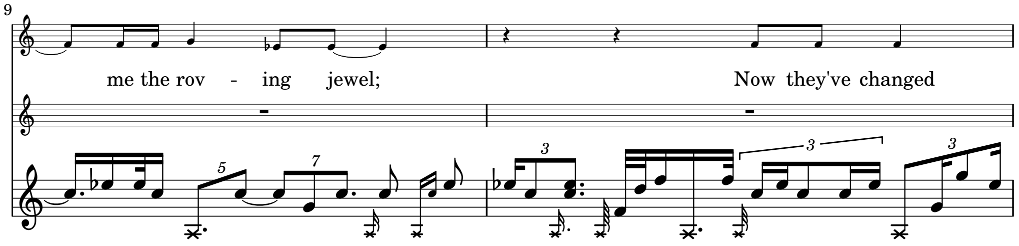
Musical events are generally associated with additional properties beyond time and duration, including pitch value, unpitched sounds, articulation, dynamics, bow position, or additional performance techniques. Once an onset is identified, the properties of the corresponding musical event still need to be determined. I often treat onsets differently depending on which detection function is triggered — "pitched" onsets (onsets that are trigged by pitch estimation ODF) versus "percussive" onsets (onsets that are triggered by amplitude or spectral ODFs). I find that using multiple simultaneous ODFs helps me to produce more detailed, expressive scores, and mimic idiomatic and genre-specific playing styles by mapping different analysis features to different aspects of musical performance.

Usually I use pitched onsets to trigger new pitches, drawing on the estimated pitch value of the corresponding analysis frame, which is filtered and smoothed to debounce changes that are small or short-lived (as described in above). I use percussive onsets to rearticulate the most recent pitch or chord, producing rhythmic figures of repeated notes. The balance of pitched versus percussive onsets controls the rate of new pitch information, and allows me to focus the transcriptions on melodic versus rhythmic figures. This is a flexible control that can be tuned to suit diverse musical figures, from foregrounded melodic lines to backgrounded accompaniment parts, such as rhythm piano or rhythm guitar. Excerpt 2.3 shows a transcription of the rhythm guitar in *This Guy's in Love with You*. I balanced the onset detection more heavily towards percussive onsets slowing the rate of harmonic motion and producing rhythmic figures of repeated chords.

In some transcriptions, I map pitched and percussive onsets to different articulations or aspects of playing technique. Excerpt 2.4 shows the banjo transcription of *Katie Cruel*, in which percussive onsets are interpreted as unpitched muted strokes rather than repeated notes, to mimic the percussive aspects of Karen Dalton's clawhammer banjo style. In other transcriptions, I use additional analysis features to determine playing techniques. In the guitar transcription of *Like a Prayer*, pitch confidence determines whether an onset is a pitched chord or unpitched muted stroke. Pitch confidence is a measure of pitch salience, or how certain the pitch estimator is of the estimated pitch. Onsets below a given confidence threshold are transcribed as muted strokes and above as chords.



Excerpt 2.3: *This Guy's in Love with You*, Rhythm Guitar



Excerpt 2.4: *Katie Cruel*, Banjo

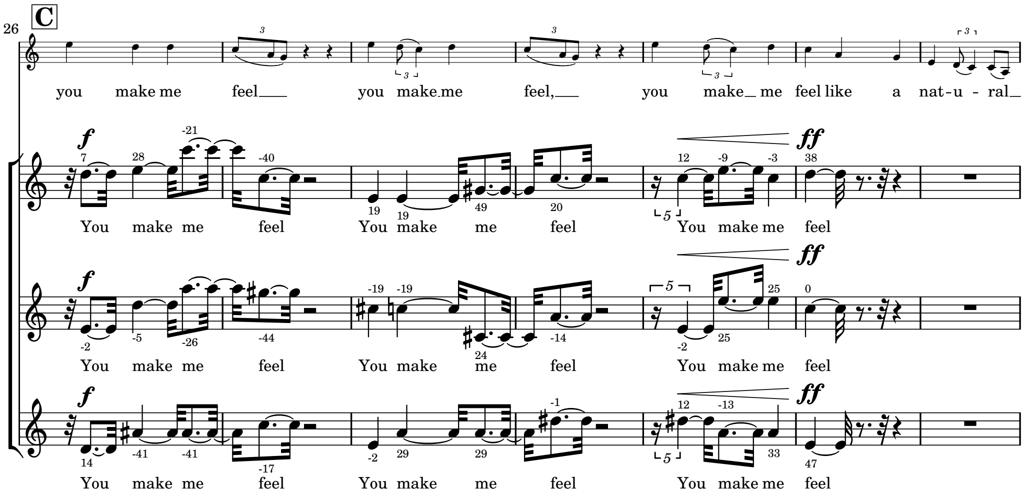
2.5.4 Polyphonic Onsets

When working with polyphonic voices, there are many ways to interpret onsets, from fully coordinated voices to fully independent voices. This pertains equally to individual polyphonic instruments, such as keyboards and guitars, or sections of multiple monophonic instruments, such as a choir or horn section.

An onset (either percussive or pitched) can be used to trigger multiple simultaneous notes, producing rhythmically coordinated voices. This is illustrated by the transcription of the background vocal trio in *(You Make Me Feel Like) A Natural Woman* in Excerpt 2.5. I wanted the voices to be in rhythmic unison, as sung in the original recording, and onsets are triggered simultaneously across each of the three voices. As a result, some voices have rearticulations where others have new pitches, and small discrepancies in rhythm are smoothed into rhythmic unison.

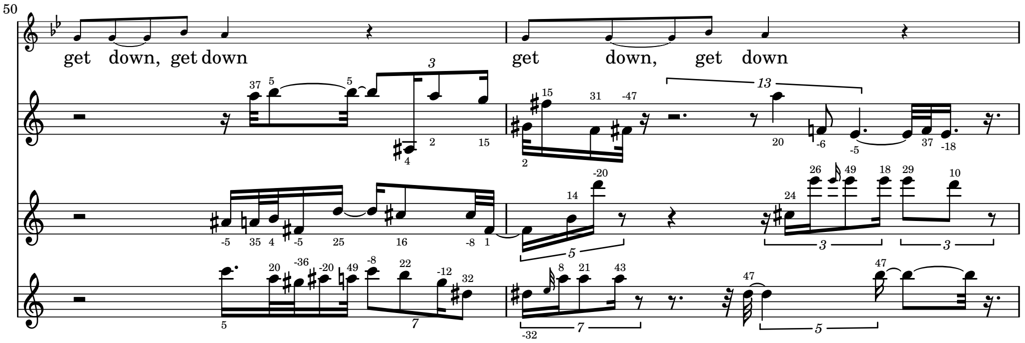
Alternatively, onsets may be detected separately for each voice, producing rhythmically independent polyphonic voices. Excerpt 2.6 shows a transcription of the horn section of *Jungle Boogie*, consisting of trumpet, alto sax, and tenor sax. I wanted the horn section to be uncoordinated, the three instrumental voices unraveling like the effect of a trumpet fall. The three instrumental staves are quantized separately, producing different subdivision grids. The first beat of bar

51, for instance, is subdivisions of eight against five against seven. I often interpolate between these two poles, assigning different functions to different kinds of onsets, which gives a continuum between coordinated and independent voices.



C
26
you make me feel___ you make me feel,___ you make_ me feel like a nat-u - ral_
f
You make me feel You make me feel You make me feel
ff
f
You make me feel You make me feel You make me feel
ff
f
You make me feel You make me feel You make me feel
ff

Excerpt 2.5: *(You Make Me Feel Like) A Natural Woman*, Choir



Excerpt 2.6: *Jungle Boogie*, Horns

2.5.5 Onset Detection for Percussion

Drums transcription requires a different strategy than the onset/offset detection method used for pitched instruments. Drums are a mixture of overlapping discrete unpitched sources. Drum tracks consist of multiple sources — each drum is a separate source — that may occur simultaneously, where as pitched instruments are separated into individual instruments before onset/offset detection. Since drum notes do not exist along a continuous continuum of pitch, but rather are chosen from a discrete set of unpitched sounds, deciding the value of a drum note is not as simple as using the current pitch value, but requires identifying which drum(s) is struck.

For HVB, I developed a drum transcription method using PLCA (the spectrogram decomposition model previously discussed in Section 2.3). The PLCA model is pre-trained on samples of individual drum hits — these are synthesized using the Apple DLS General MIDI Sound Bank. A drum track is transcribed by fitting the pre-trained basis functions to the track, which gives a new set of activation functions, representing the temporal locations and amplitudes of the pre-trained bases within the drum track audio. These activation functions are used as onset detection functions, one for each drum in the trained model, and onsets are found by peak picking, allowing polyphonic drum transcription of multiple overlapping sources. Figure 2.7 shows an example of PLCA activations as used for drum transcription. The audio clip consists of four beats of a simple drum figure. Activations and basis functions are plotted for the four individual drum sources, and it can be seen that abrupt rises in the activations correspond clearly with drum onsets. Importantly, moments when multiple drums are hit simultaneously exhibit energy across multiple activation functions.

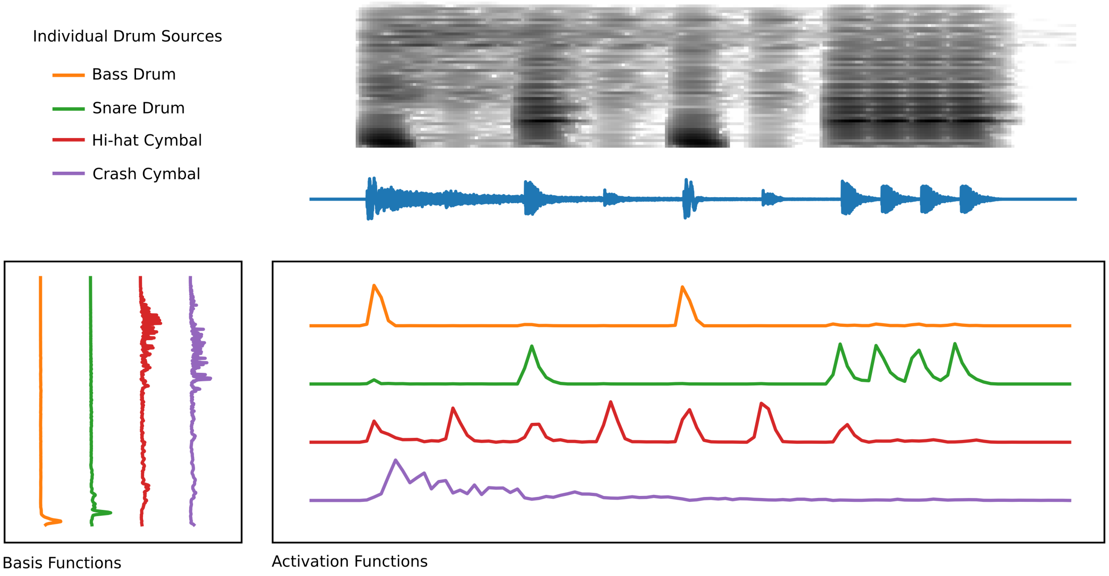


Figure 2.7: Drum transcription using PLCA activations as ODFs.

PLCA can be very effective for drum transcription, and its analysis parameters afford useful controls. Imposing sparsity constraints, for instance, on the activation functions operates as a control of polyphony, otherwise the algorithm finds multiple similar percussion sources in place a single source. The number of components used to represent each training sample influences the model's ability to generalize and account for variability in a single drum source. This number essentially controls the flexibility or "wiggle room" for each source when matching it in the mix, and sources with more flexibility will occur more frequently in the transcription. I also impose weights on the activations to control the likelihood of each source in the transcription. These parameters allow me to control the relative number of occurrences of each drum in the transcription and steer the model towards more or less of a certain drum if that source is overpowering or lacking in the transcription.

2.6 Music Notation

The values produced during pitch and rhythm analysis are found without consideration for any particular musical representations of time or pitch. Timings are expressed in seconds, measured to a degree of resolution afforded by STFT analysis — usually a hop size of 512 samples, or $44100/512 = 1/86^{\text{th}}$ of a second. Pitch is expressed in hertz, measured to floating point numerical precision. *Quantizaton* is the task of fitting these "arbitrary" analysis values (generated without respect for musical values) to a limited set of musical values. This section focusses on rhythmic quantization because my treatment of pitch quantization is less complex; frequency is quantized to the cent ($1/1200^{\text{th}}$ of an octave) and notated as cent deviations from nearest twelve-tone equal temperament pitch. Of course more complex tuning strategies could be used, such as inferring a tuning system — either fixed, variable, or perhaps even paratactical. My treatment of rhythmic quantization, however, is more complex, and involves fitting sequences of time points to musical subdivision grids.

2.6.1 Rhythmic Quantization

Conventional music notation represents time according to a metric framework of beats. Musical time is measured in terms of a fundamental unit, the beat, which can be divided into smaller parts, usually of equal size and in small integers ratios — 2, 3, 4, 5, 6. Rhythmic quantization, the task of fitting an arbitrary sequence of time points (generated without respect for musical durations) to the form of musical notation, poses several considerable challenges, which have been addressed in a variety of ways by composers and researchers, particularly those in the domain of computer-assisted composition. Because musical rhythms must fit a specific form of allowable subdivsions, there is often not an exact match, and transcribing an arbitrary sequence of time points, in general, involves finding the closest match, while balancing excessive notational complexity — such as small divisions of the beat, large prime subdivisors, and frequent changes of subdivision grids.

My quantization framework is informed by ideas from Nick Didkovsky (Didkovsky 2004, Didkovsky and Burk 2019) and Paul Nauert (Nauert 1994). I use a quantization technique based on the concept of the *beatspan.* A beatspan is a duration of time (expressed in terms of beats) that may be subdivided, or broken down into smaller parts, generally but not always of equal division — incomplete tuplets, for example are unequal divisions of a beatspan. These parts may be in turn treated as beatspans themselves and further subdivided into smaller and smaller spans, forming a hierarchical structure of nested subdivisions. A rhythmic grid constructed according to this process of subdivisions is called a *quantization grid*, abbreviated *Q-grid*, a term introduced by Nauert. An example Q-grid is shown in Figure 2.8, illustrating the structure through which a beatspan — in this example, four beats — is hierarchically subdivided into finer and finer divisions.

Due to the nested structure of subdivisions, Q-grids can be irregular, having uneven spacing between grid divisions. The flexibility of Q-grid representation is due to the allowance that beatspan durations may be greater than, equal to, or less than a beat, providing for divisions of half the beat or less. Multiple subdivisions may be mixed within a beat, such as dividing the first half into triplet subdivisions and the second half into quintuplet subdivisions. This allows for complex and hierarchical grouping of rhythms. This flexibility to fit arbitrary time points, however, comes at the cost of notational simplicity.

A collection of possible beatspans — called a *beat division scheme list*, a term borrowed from Didkovsky — determines the possible Q-grids, or possible musical durations and that can be used to express timings. Quantization involves finding the best Q-grid to express an arbitrary sequence of time points, and substituting it for the arbitrary sequence. The number of possible Q-grids depends on the possible beatspans and is generally quite large, growing combinatorially with the number of beatspans, as well as the length of the sequence to be quantized. Finding the best Q-grid is a balance between resolution, computation demands, and notational complexity. Large collections of possible beatspans provide finer temporal resolution, but also allow for many more possible Q-grids, which increases the computational difficulty of finding the best fit, as well as the notational complexity of the resulting musical expression — increasing depth of nested subdivision hierarchy and larger subdivisor primes.

An example beat division scheme list is shown in Figure 2.9. A beat divisions scheme is represented as a list of possible beatspans — Nauert uses a tree structure. Each beatspan is defined in terms of its total duration and number of subdivisions. This is encoded by a multiplier and baseline note duration, modeling standard music tuplet notation such as 3:2 (said “three in the span of two”). The total beatspan duration is given by the baseline note duration multiplied by the denominator, and the numerator gives the number of subdivisions. For example, `bds.add_tuplet((5, 4), 1.0/2)` represents a beatspan of five eighth notes in the span of four eighth notes, or an eighth note quintuplet.

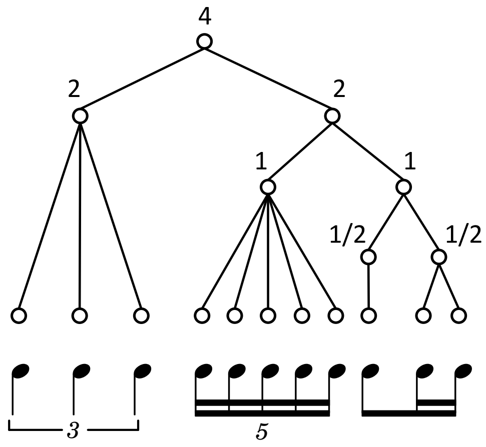
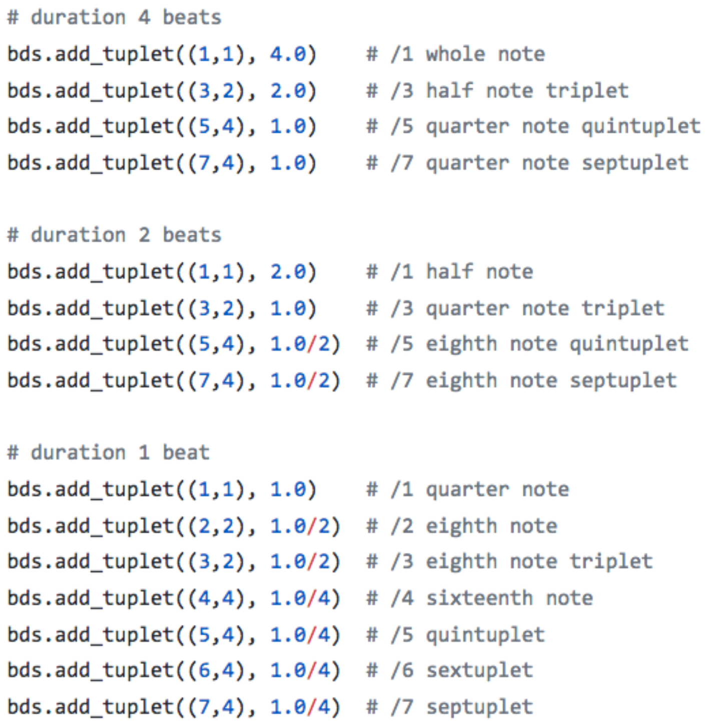


Figure 2.8: An example quantization grid (Q-grid).



```
# duration 4 beats
bds.add_tuplet((1,1), 4.0)    # /1 whole note
bds.add_tuplet((3,2), 2.0)    # /3 half note triplet
bds.add_tuplet((5,4), 1.0)    # /5 quarter note quintuplet
bds.add_tuplet((7,4), 1.0)    # /7 quarter note septuplet

# duration 2 beats
bds.add_tuplet((1,1), 2.0)    # /1 half note
bds.add_tuplet((3,2), 1.0)    # /3 quarter note triplet
bds.add_tuplet((5,4), 1.0/2)  # /5 eighth note quintuplet
bds.add_tuplet((7,4), 1.0/2)  # /7 eighth note septuplet

# duration 1 beat
bds.add_tuplet((1,1), 1.0)    # /1 quarter note
bds.add_tuplet((2,2), 1.0/2)  # /2 eighth note
bds.add_tuplet((3,2), 1.0/2)  # /3 eighth note triplet
bds.add_tuplet((4,4), 1.0/4)  # /4 sixteenth note
bds.add_tuplet((5,4), 1.0/4)  # /5 quintuplet
bds.add_tuplet((6,4), 1.0/4)  # /6 sextuplet
bds.add_tuplet((7,4), 1.0/4)  # /7 septuplet
```

Figure 2.9: Beat division scheme list.

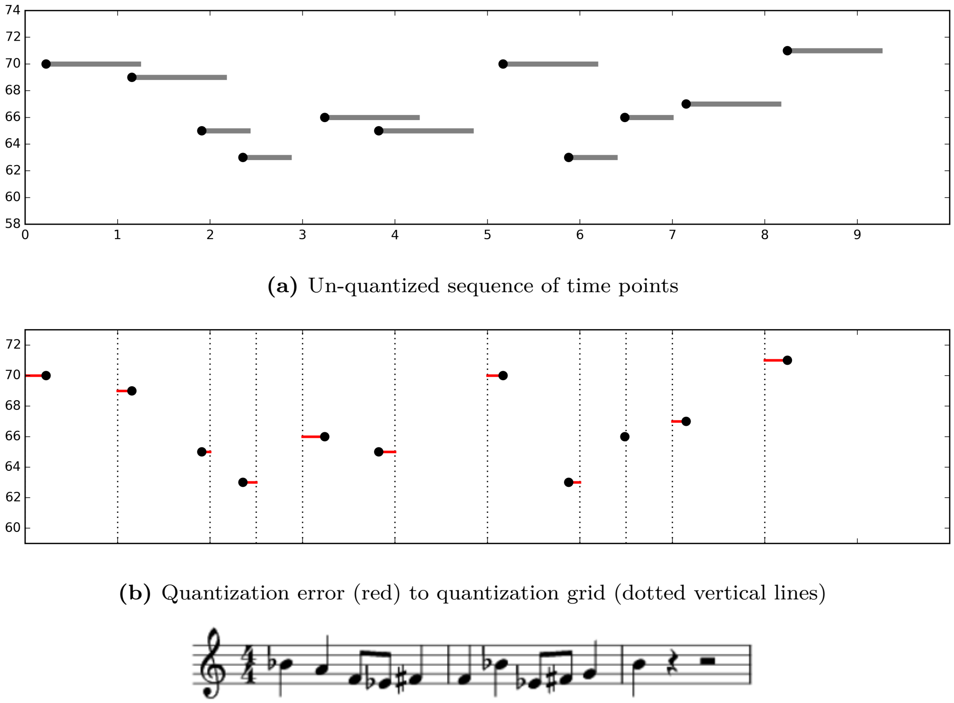
2.6.2 Optimal Q-grids

Finding the best Q-grid to express an arbitrary sequence of time points can be treated as an optimization problem. Given a sequence of arbitrary time points to be quantized $\{x_i\}$, the optimal Q-grid minimizes the quantization error, expressed as the sum of power differences between the actual and quantized points

$$\sum_{i=0}^{N-1} |x_i - y_i|^p \, \frac{1}{n_i} \, w(n_i) \tag{2.18}$$

where $\{y_i\}$ are the corresponding quantized time points, found by taking the closest Q-grid subdivision: $y_i = argmin(|x_i - y|) \, \forall y \in Y$ for each x_i. The term n_i is the number of the subdivisons in the beatspan containing y_i, and the expression $1/n_i$ is a general regularizing term that depends inversely on the magnitude of the number of subdivisions. It is intended to penalize large subdivisors, combatting the problem that smaller divisions of the beat tend to populate quantizations, because they represent finer temporal grains. The term $w(n_i)$ is a manually supplied weight, also depending on the subdivisor n_i of the containing beatspan, for further hand-tuning the regularization. The weights $\{w_{n_i}\}$ afford control of the notational complexity of the quantized rhythms by allowing separate tuning for each beat division. The exponent p scales error along an exponential curve. In summary, the quantization error expresses the best fit Q-grid subject to additional constraints on notational complexity.

The transcriber quantizes a sequence of arbitrary time points by substituting it with the sequence of Q-grid time points that minimizes the quantization error as expressed by Equation 2.18. Figure 2.10 shows the quantization error of an arbitrary sequence of time points. The transcriber finds the subdivision grid, or Q-grid, that minimizes the the sum total amount of error, shown in red. The Q-grid that minimizes the quantization error is found by searching all possible Q-grids. I transcribe one measure at a time to keep the sequence length short, since computation time grows combinatorially with the length of the sequence. Currently, the transcriber implementation enumerates all possible Q-grids and calculates their errors, but this could be made more efficient using AI techniques. Didkovsky, for example, uses a heuristic search with backtracking to reduce computation time.



(a) Un-quantized sequence of time points

(b) Quantization error (red) to quantization grid (dotted vertical lines)

(c) Notated rhythm

Figure 2.10: Quantization error of arbitrary time points.

2.6.3 Constraints on Complexity

Many theorists have devised tools and concepts to measure and constrain rhythmic complexity in music transcription. These frameworks generally consider both the numerical complexity of subdivision ratios as well as the difficulties and cognitive demands of musical performance. Nauert's framework, perhaps one of the most detailed, identifies three categories of factors that contribute to rhythmic complexity: properties of the Q-grid itself, such as large prime divisors and rapid subdivisions of time; changes in the rate of subdivisions, either within or between beatspans; and the number of location of deletions (grid points that are unarticulated or rests) — whether they fall on weak Q-grid subdivisions or on very small divisions of beatspans. Nauert's framework is both musically and numerically motivated. Q-grid properties, such as having large primes, pertain to the numerical complexity of the Q-grid. Deletion properties, on the other hand, how many and where they fall, are motivated by the difficulty of visually parsing and mentally subdividing awkward rhythms.

Didkovsky's JMSL transcriber also includes a useful number of control parameters pertaining to notational complexity. One, in particular, is the minimum number of beatspan divisions that must be articulated to use a given beatspan. Similar to Nauert's deletion principle, requiring larger numbers of articulations guards against sparse Q-grids that are difficult for performers to mentally subdivide.

Barlow's *indispensability* function (Barlow and Lohner 1987), which measures the metric contribution of a given pulse, can be used to measure and constrain performance complexity during quantization. Indispensability identifies the rhythmic locations that are most salient in the perception of meter. Dispensable subdivisions do not contribute substantially to the sense of meter and can be deleted without altering the perception of the type of meter. Indispensability identifies or formalizes the concept of strong and weak, and, if used as a quantization constraint (weighting factor), steers the quantizer away from rhythms that have ambiguous meters and, as such, are difficult to perform. Constraining quantization to relatively indispensable subdivisions will produce clear meters, where as allowing dispensable subdivisions will produce ambiguous meters.

Outside the domain of rhythmic quantization, theorists have given considerable attention to the complexity of ratios, most often in considering *harmonic distance*, which attempts to measure and explain the relative consonance of pitch intervals. Many approaches to harmonic distance, however, can also be applied to rhythm. Barlow's rhythmic indispensability function, for example, is an application to rhythm of his *harmonicity* harmonic distance function and yields similar results in both domains. Harmonic distance functions, such as Tenney's *HD* function (Tenney 2015) and Euler's *gradus suavitatis* (*GS*) function (Polansky 2013), usually consider the number of distinct prime factors as well as the magnitude of those primes and their powers. Large primes and composite numbers, which are more complex numerically, have greater harmonic distance. While harmonic distance functions are usually applied to the simplified relatively prime ratios, when applied to rhythm, it is useful to maintain the unsimplified ratio. Magnitude corresponds to the depth of nested subdivision, where as relative, or simplified ratio, corresponds to the relationship within a given tier. The tuplet 6:4, for example, should be more complex than 3:2, even though the ratios are equivalent when simplified.

In my transcriber, I only consider complexity constraints based on notational complexity, which are implemented in the form of manually supplied subdivisor weights. There are two primary motivations for this decision. First, I am less interested in tempering the music notation according to theoretical ideas of performance complexity as to what is/not playable; I'd rather pass unfiltered notation along to the players. This pertains to my approach to performance (discussed in Chapter 3), which draws on a tradition of twentieth century computer-assisted composition that pushes the limits of musical performance by not simplifying computer composed abstract ideals according to constraints of human performance practice. Second, this part of the project workflow, rhythmic transcription, serves a practical need — to prevent small subdivisions from dominating the transcription — and is intended to achieve this desired result rather than an exploration of a deeper musical question about notation and performance complexity. The weights are supplied manually rather than implemented using a mathematical function or measure of numerical complexity because I want direct control over the distribution and balance of tuplets rather than indirect control through equation coefficients.

2.6.4 Alignment to the Beat

A common problem in rhythm transcription is misalignment between the pulse of the original sequence and the pulse of the transcription. If a sequence is quantized to an arbitrary pulse, rhythms that are perceptually simple may become notationally complex. I transcribe to the beat of the original song, quantizing rhythms to the locations of pulses in the original recording. The transcription remains aligned to the original beat despite rubato, accelerando, ritardando, and minute fluctuations in tempo from beat to beat, and is more semantically meaningful because it is notated with respect to the original beat.

I do so by normalizing fluctuations in tempo to a constant pulse at the average tempo (beats per minute as averaged over the entire song). This requires knowing the location of each beat — knowing the tempo alone is not enough because the pulse may drift in musical performance. I find the location of each beat manually by listening to the song and tapping along on a computer interface, which records the timings of each tap. To ensure the tap tempo is accurate, I listen to the song multiple times in advance to anticipate tempo changes, use punch in/out recording techniques to re-record portions of the track, or manually adjust the beat markers after the fact using a DAW. Given a sequence of time points and list of beat locations, each time point t is normalized to a constant pulse using linear interpolation relative to nearest neighboring lower and upper bounding beats, such that

$$t_{norm} = \frac{t - b_n}{b_{n+1} - b_n} + i \tag{2.19}$$

where t is the time point to be normalized, b_n and b_{n+1} are the beat locations directly before and after, and i is the beat index. Figure 2.11 illustrates a sequence of time points with a fluctuating pulse as normalized to a constant pulse. The actual beat locations and intervening onsets are time warped to a constant pulse. Grid top (a) is the waveform with onsets (dashed lines) and manually annotated beat locations (numbers). Grid bottom (b) shows the onset locations as warped to the constant pulse. Transcription top (c) is the sequence as transcribed to an unaligned pulse. Transcription bottom (d) is the sequence as transcribed with alignment. Not only is the unaligned sequence unnecessarily complex, but the aligned sequence is more semantically meaningful because the rhythm spellings refer to the beats of the original song.

Notating to the pulse of the original songs also allows me to include the vocal line in each instrumental part. Since there is no common subdivision grid shared among instrumental parts, it is difficult for the ensemble to keep time in performance by listening to one another. Rather, the ensemble performs along with the original voice track, which gives a common pulse. Timing fluctuations, which are normalized out during quantization, are reintroduced in performance by slowing down and speeding up in time with the singer.

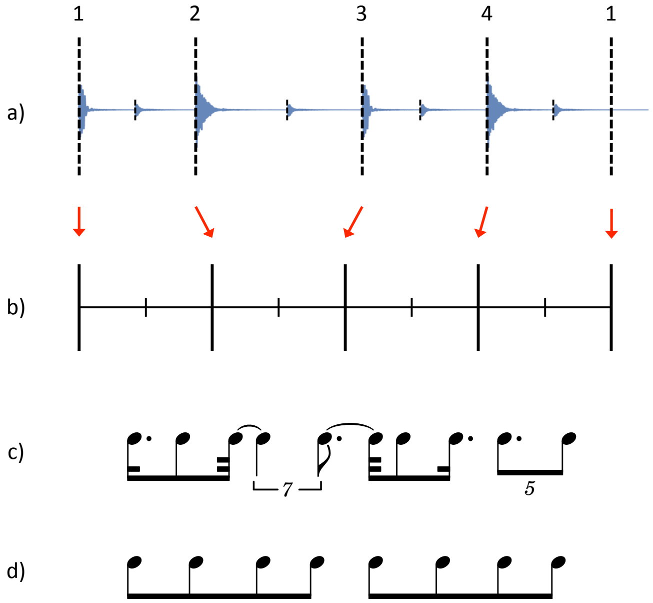


Figure 2.11: Quantization alignment.

Chapter 3

The Ensemble: Performance Practice and Collective Identity

In 2011 I formed an ensemble, the Happy Valley Band, to play the computer generated transcriptions. The ensemble formed at the suggestion of two composer-performer colleagues, Mustafa Walker and Beau Sievers. During an artist talk at Ostrava Days 2011 Festival of New Music, in which I spoke mainly about a recent algorithmically generated orchestra piece, I showed a computer synthesized mockup of the HVB song *Crazy*, an example of a new project and future work direction. Mustafa Walker and Beau Sievers, who were both in the audience, immediately suggested that we form an ensemble to play the machine listening transcriptions. At this point, I had not considered that an ensemble would perform the music. The transcriptions existed as MIDI data and audio renderings, not as notated musical scores. Within a day or two, I produced the first HVB scores — this was done hastily by importing the MIDI data into Finale — and the ensemble had its world premiere days later at a club in Ostrava, Czech Republic, with an ad hoc group consisting of Beau Sievers, Mustafa Walker, Andrew Smith, and Larry

Polansky. Having only had one brief afternoon rehearsal, the performance was a rushed, scrappy interpretation, and the scores looked considerably different than current HVB notation.

Two important things occurred that evening. First, Kurt Gottschalk, a writer and host of the WFMU radio show Miniature Minotaurs, approached me and said, "if you ever do this again you're playing on my radio show." The invitation was sufficient excuse to make the HVB ensemble official and to start expanding the HVB repertoire list. Second, within the span of a week or so at Ostrava Days Festival, I had worked with both a professional orchestra and an ad hoc group of composer-performer colleagues. Working with the ad hoc group of colleagues was a very different experience than working with the professional orchestra. In the orchestra piece, I was careful to temper the difficulty of the music to fit within an hour or two of rehearsal time, and I made sure not to ask too much of any particular instrumentalist. HVB was very much the opposite. I indulged in the difficulty of the notation, and the fun of the music came from its absurd impracticality. Just about everything asked of the performers was impossible, but I found that musicians — or these *particular* musicians — reveled in the challenge. While it was impossible to play the music exactly as written, something emerged in performance that was more than the sum of the parts. A group interpretation or feeling emerged, as the performers together found ways to interpret the impractical notation and excessive pitch and rhythmic resolution of the musical scores. This was the flash of excitement and insight that propelled the project forward. It was not immediately clear how we should approach writing or performing such music, and the HVB ensemble exists to explore these questions.

From its inception, the HVB ensemble has been a collaborative idea and a collective endeavor. It is a core group of seven musicians — Beau Sievers (drums), Mustafa Walker (bass), Andrew Smith (keyboards), Alexander Dupuis (guitar), Pauline Kim (violin), Conrad Harris (violin), and myself (saxophones) — with whom I worked closely over the course of five years, from 2011 to 2016, to develop the project. The ensemble performance practice as well as the music itself grew out of working with these performers, and the music is shaped by their feedback, preferences, and playing styles, not to mention their values, perspectives, and aesthetic preferences. What makes the ensemble unique is both remarkably novel and immensely mundane: it is a collaborative social structure that allowed for the emergence of a shared group identity and the development of a group decision-making process in response to complex musical performance practice challenges — put more colloquially, it is a band. While collective and collaborative musiking is very much the norm in certain cultural contexts, such as a rock band, in the context of contemporary and composed music, it presents challenges to and complications of concepts of authorship, authenticity, and intent.

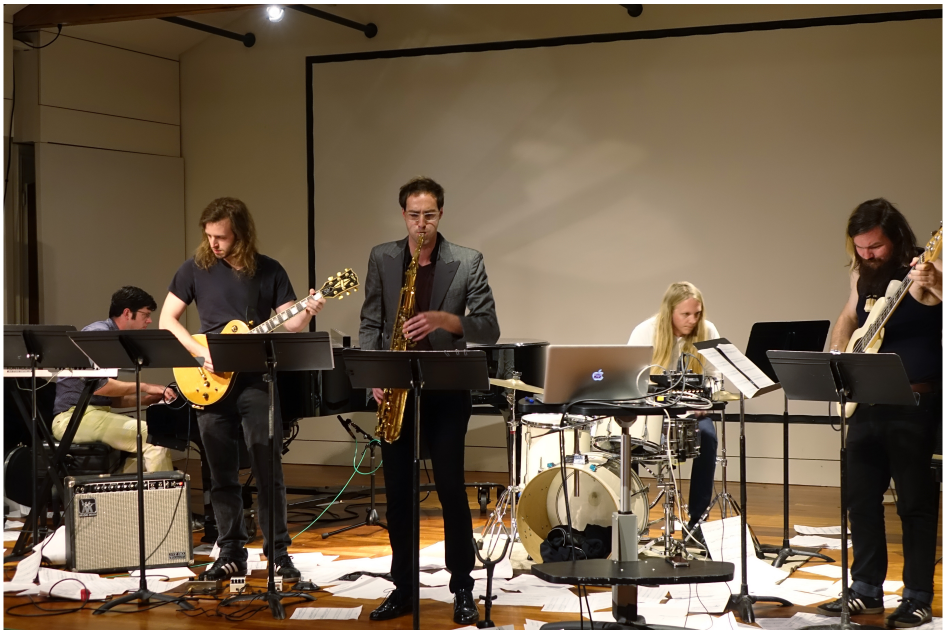


Figure 3.1: Happy Valley Band performing live at CCRMA, May 2017.[1]

3.1 Performance Challenges in HVB

HVB music presents a number of performance practice questions. How should performers approach instrumental writing that is physically awkward, impractical, or non-idiomatic to the instrument, such as difficult leaps across registers and breaks in the instrument, awkward bowing, or uncomfortable fingerings? How should performers approach notation that is impossible, such as pitches above or below the playing range of an instrument, simultaneous notes in monophonic parts, and chords that are wider than the stretch of the hand? Should the transcription algorithm be adjusted to prevent these problems from occurring in the first place by encoding knowledge of human performance limits within the transcription process, should these problems be corrected after the fact as a kind of post-processing manual notation cleanup, or should the performers be tasked with finding ways to deal with them? How should the performers approach notation that is specified beyond the precision of human performance capabilities — pitch is quantized arbitrarily to the nearest cent deviation, not to a limited subset of pitches or specific tuning system, and without respect to a underlying tonal reference, and rhythm is quantized to small subdivisions and complex nested tuplets? Is it even possible to play in the first place? Should performers prioritize one aspect of the music notation or another and how? How should the ensemble keep time in the absence of a discernible beat, or really any coordination at all between parts? Parts are quantized separately; beyond a common pulse, there is no common subdivision grid shared among individual instruments. Should performers attempt to coordinate in time and how? Should the music be conducted? Or should the parts be left to come in and out of alignment?

[1] Photo credit Madison Heying

3.2 A Few Perspectives on Impossibility

Since the latter half of the twentieth century, composers have continually pursued the limits of musical performance. There are precedents to be found for many if not all of these questions within traditions of twentieth century music practice, which, in general, saw both an increase in the complexity of written music as well as an increase in musicians' abilities to play difficult music. At the time that I was wrestling with questions of how to notate and perform HVB music, a handful of perspectives in particular formed the intellectual and aesthetic context in which I approached the project and influenced my to approach to performing HVB music.

In the 1970's, John Cage (1912–1992) composed three sets of etudes that have become iconic of notation that is extremely complex and incredibly difficult to perform — *Etudes Australes* for piano (1974–75), written for Grete Sultan; *Freeman Etudes* for violin (1977–80/1989–90), written for Paul Zukofsky; and *Etudes Borealis* for cello and piano (1978), written for Jack and Jeanne Kirstein. All three sets were composed using star charts and *I Ching* chance procedures. Although the use of star charts is similar to Cage's compositional approach of *Atlas Eclipticalis* (1961–1962), these works are more detailed and determined than Cage's earlier scores. Musical material, including not only pitch and duration but also additional performance instructions, such as articulations and performance techniques, are specified. This wealth of notational instruction is largely responsible for the extreme performance difficulty of the scores.

Perhaps the most challenging of the three sets is *Freeman Etudes*. Beyond rapid figures and difficult fingerings, the music is notated in great detail, specifying articulations, bowing direction, bowing locations, and bowing styles, tremolo and vibrato, detached versus legato, as well as a four types martellato attacks and five degrees of pizzicato. Cage even indicated exactly how many times the bow should bounce when performing ricochet bowing. What makes the music particularly difficult, however, are the abrupt and dramatic changes, from moment to moment, in playing style and technique, a result of using chance procedures to independently determine the many performance parameters of each note. Pitch, dynamics, articulations, and other performance techniques shift rapidly and unpredictably over the entire range of the instrument from one note to the next.

In 1980, after working on the collection for three years, Cage ceased work on the project, having completed only seventeen of the planned thirty-two etudes. While composing the eighteenth etude, passages became so dense that Cage feared the music would be unplayable. While Cage had intended the music to be difficult, this extreme was an unintended result of the composition process — a confluence of the chance procedures used to select the number of notes together with the density of stars in that particular area of the start chart. Cage resumed work on the project, however, nine years later, upon hearing Irvine Arditti play the first sixteen etudes, adding the performance instruction that musicians play "as many as possible" when faced with particularly dense passages. This instruction gives performers artistic license with which to approach the music and, for Cage, proved a solution to the impossible performance demands.

Cage saw meaningful social implications in attempting to play the unplayable, beyond a dedication to a compositional process. When asked about the difficulty, Cage referred to the "practicality of the impossible," expressing a poetic optimism about trying in the face of seemingly difficult or hopeless situations, which connected, for Cage, to serious and seemingly impossible problems in society. Cage explains:

> These [the etudes] are intentionally as difficult as I can make them,

> because I think we're now surrounded by very serious problems in the society, and we tend to think that the situation is hopeless and that it's just impossible to do something that will make everything turn out properly. So I think that this music, which is almost impossible, gives an instance of the practicality of the impossible.[2]

For Cage, as well as many other composers — myself included — there is value, both musical and poetic, in attempting to play music, even when the result will not be a perfectly accurate rendering of what is written, and this interplay between notation and performer becomes a part of the piece. The story of *Freeman Etudes* also illustrates that the designation "impossible" is contingent, both on the abilities of an individual performer as well as, more generally, the historical moment. It is a criteria that can change and can be changed.

Brian Ferneyhough (b. 1943) and associated "New Complexity" school of music composition represent another widespread perspective on impossibility in musical performance, which ultimately bears many similarities to Cage's. Although the term New Complexity is contentious and ill-defined — the associated group of composers maintain vastly different compositional approaches and notational strategies, as well as different motivations for and definitions of the term complexity — the term generally refers to an extreme notation specificity and abundance of notes. It is often marked by parametric and other new and invented notation systems. Brian Ferneyhough is a central figure, having worked in this manner dating to the 1960s. In the 1980s, however, Ferneyhough's ideas found renewed interest with a younger generation of composers, among them Michael Finnissy, Chris Dench, Richard Barrett, and Aaron Cassidy, and are now widespread in contemporary music, in part due to Ferneyhough's tenure as a pedagogical fixture at US universities.

[2] Cage quoted in Pritchett 1993

Although the term New Complexity often refers to the surface-level features of extremely dense and detailed notation, the deeper shared musical idea is a perspective on the limits of music notation. Maintaing that no system of notation can "record information encompassing all aspects of sonic phenomena for which it stands," Ferneyhough eschews the concept of accuracy — the exact rendering of musical notation into sound — in favor of a more complex relationships between notation and performance. Similar to Cage, Ferneyhough sees value in attempting the impossible, rather than shying away from it, advocating "the musical effects of as near an approach as possible to this unreachable ideal."[3] Much of his music focusses on devising notation systems to explore this space.

Ferneyhough sees exactitude not as a limitation of performer freedom and interpretation but rather as an opportunity for it, giving the performer license to move within the strata of notational detail. Because the music is so exacting, the performer must devise a strategy for interpreting it, as Ferneyhough explains:

> The goal here, I think, is, therefore, a notation which demands of the performer the formulation of a conscious selection-procedure in respect of the order in which the units of interpretational information contained in the score are surveyed and, as an extension of this choice, a determination of the combination of elements (strata) which are to be assigned preferential status at any given stage of the realization process. The choice made here colors in the most fundamental manner the rehearsal hierarchy of which, in performance, the composition itself is a token.[4]

The performer's decision-making process is a part of the piece, and their strategy, including decisions about which aspects of musical performance to prioritized and how to prepare and rehearse the music, as well as abilities as a performer, forms a filter through which the music is rendered. Difficulty is part of the composition,

[3] Ferneyhough 1995

not necessarily a barrier to overcome. Ferneyhough refers to the score as a “token,” a tangible stand-in for the composition, where as the piece itself takes shape across multiple artifacts — the composer’s sonic idea, its rendering into music notation, a performer’s response to that notation, and a listener’s experience of that performance.

While Cage’s view is couched in a social optimism, many composers associated with New Complexity, Ferneyhough included, connect their music to perception, experience, and the complexity of the world. Composer Erik Ullman offers a definition of New Complexity as “a music that privileges ambiguity and subtlety, nourishing many paths of perception and interpretation.”[5] Composer Kaija Saariaho similarly writes that music should reflect “our personal way of filtering the world,“ rather than “the endless information surrounding us.”[6] Common to many interpretations of New Complexity is an emphasis in the subjectivity of experience in a world that is overwhelming dense and full of information. This complexity is reflected in the notation, the exactitude of which gives performers and listeners autonomy to find their own ways through. Despite composers’ words and intentions, whether or not performers experience the music in this way is another question. In contemporary music communities, Ferneyhough’s music often seems like a conduit for virtuosic displays of technical prowess, a challenge to be mastered, not a celebration of performer freedom. Although stylistically very different, HVB relates to New Complexity in two regards: first, not altogether dissimilar from Ferneyhough’s description, performers develop performance strategies to navigate the notation; and second, HVB explores machine listening as a subjective, rather than objective, metaphor for human perception.

[4] Ibid.

[5] Ullman 1994

[6] Saariaho quoted in Duncan 2010

Beyond the difficulties of exacting and overwhelmingly detailed notation, another kind of impossibility proliferated within twentieth century music practice, this one pertaining to musical ideas that challenge the underlying mechanisms and structures of human music practices. Works that are notationally simple may still be difficult or impossible to perform because musicians are asked to engage with musical performance in ways that are not consistent with how musicians are trained to act, conceive of, or organize music making — some may even be incongruent with the very cognitive mechanisms that function when musiking! A trend in twentieth century art, more generally, was an interrogation of the very medium and material of expression, and twentieth century music composition is rich with works that question, subvert, ignore, and otherwise extend the boundaries, assumptions, and structuring principles of human music practice. In many instances, the challenge to traditional principles causes unusual performance situations and impossibilities.

One composer, in particular, whose work provided a context for my approach to HVB is Larry Polansky (b. 1954). Polansky's work explores how computers can be used to extend and better understand human music practice, often resulting in unique performance challenges. Much of Polansky's work, for instance, explores experimental intonation systems and performers are often asked to play or sing in alternate tunings, and even adjust, adapt, and retune on the fly — in *for jim, ben and lou: Preamble* (1995) the "percussionist" is tasked with retuning the guitar as the guitarist plays it. Other works require musicians to perform difficult cognitive tasks, or even multitask. In *Ensembles of Note* (1998–99) each performer incrementally builds an eight-bar melody in five-four time, adding a few notes with each repetition, much like slowly filling up a tape loop or delay line. While the metaphor is simple, performing the work can be quite difficult, because of how human memory is used.

Proposition (three verbs and a logical operator), the second movement of Polansky's *3 New Hampshire Songs* for 16 part mixed choir (1999), follows the structure of a Rhythmicon, an early electronic instrument designed and built in 1930 by Henry Cowell and Léon Theremin in which pitch and rhythm are analogously structured in harmonic ratios. Performers are asked to sing pitches tuned to harmonic series intervals, up to the seventeenth harmonic of the fundamental D, and in analogous rhythmic proportion, or polyrhythms. The two voices covering the seventeenth and thirteenths harmonics, for example, sing in polyrhythm seventeen against thirteen. While the idea is simple, its execution is quite difficult, because choirs are generally not trained to sing such pitch ratios and polyrhythms, at least not in Western music traditions. At the premiere, the choir sung with click tracks to maintain coordination.

Polansky's ideas, like many if not all composers, are shaped by the contemporaneous cultural moment, including greater societal trends in technology and the sciences. Even when not explicitly using digital technologies and electronic tools for composition, these technologies nevertheless influence our thought, providing metaphors, mental models, and ways of thinking that structure and orient our ideas. In this sense, many works that are incongruent with traditional music practices indicate the influence of technology, even if indirectly — Polansky's tape loop metaphor in *Ensembles of Note* or the Rhythmicon in *3 New Hampshire Songs*. Not only do computers provide new metaphors and structures, but, in the sense that computers tend toward abstraction, they allow composers to work through and realize idealized concepts, removed from the constraints and mechanisms of human performance. As I see it, HVB sits at a nexus between logics that are shaped by technology and logics that are shaped by the traditions of human musical practices. Many of its performance challenges — the overly precise notation and lack of rhythmic coordination, for instance — are artifacts of digital abstraction and machine analysis, discontinuities between traditions of human music practice and affordances of digital technologies. When considering whether or not to edit out such performance challenges, it was seeing HVB in the context of this tradition of computer music that helped me to realize these challenges are artifacts of the translation process, not errors, and are, in fact, the most important aspect of the music.

3.3 Constraints in Computer-Assisted Composition

Of all the performance challenges to be found in HVB music, perhaps the most prevalent is notation that is impossible, awkward, or impractical, not because there is too much information or too many notes, but because the music is transcribed without concern for physical affordances and limitations of the instruments. This includes notation such as chords that necessitate impossible fingerings or stretches wider than a single hand, figures that are clumsy or sit awkwardly on the instrument, and impractical or difficult leaps. Simply put, the transcription process does not take into account basic orchestration knowledge. No consideration is given to the kinds of instrument-specific performance concerns that a skilled composer would likely consider when writing for an instrument.

Generally, in computer-assisted composition, these kinds of orchestration concerns are accounted for — if they are accounted for at all — by ruled-based methods or constraint solving techniques. Implementations often take the form of simple rule checks and sequences of if-then statements, or more sophisticated systems of branching logical conditions and hierarchically prioritized rules sets.[7] Such systems, however, are often tedious to design, requiring complex logic, and can be inflexible, brittle, and overfit, leaving little room for edge cases or the nuances of how a composer might resolve conflicting rule sets. With the HVB transcription process, I chose to not implement orchestration constraints, beyond a "soft" limit on playing range, which is enforced statistically rather than as a "hard" limit — notes generally tend to be in playing range, but out of bounds pitches with high pitch confidence are allowed.

There are a few reasons for this choice. First, it's not entirely clear how an orchestration constraint checker would hook into the music generation process, which is not an infinite generator but rather a sound analysis process. It is not as though there is an endless cue of notes to pull from, check, and reject. Would the algorithm act like a filter, removing conflicting notes without replacement, or would the algorithm apply certain allowable transformations such as octave displacement, reach further into the pitch queue for new pitches that hopefully do not violate constraints, or perhaps even tweak parameters of analysis algorithms to generate new material? While deeply fascinating — a constraint checker, for instance, that could tweak the analysis parameters to find the most playable rendering would be an interesting project — these questions are, in many ways, antithetical to my intentions and ethos with HVB: I am not interested in fitting the results of the analysis process to my judgements of what is or is not playable.

I would rather let the instrumentalists themselves be the filters, each bringing to the music their unique perspectives, preferences, and abilities. My knowledge pales in comparison to that of skilled instrumentalists, who have particular understandings of their instruments and their abilities, and I am continually delighted by the inventive solutions performers find to near impossible performance demands.

Furthermore, performance limits change, not just from performer to performer, but over time. The limits of musical performance are contingent, both historically and culturally. Rather than fix HVB scores in one particular interpretation, my intention is to allow performers to make their own interpretations. Ideally this may even act as a force that challenges, stretches, and extends musical performance in the future.

3.4 Performance Practice in the Happy Valley Band

HVB performance practice is perhaps best summarized as "come up with some sort of strategy and then try to do it." I encourage performers to develop their own ways of interpreting the notation, and consider these interpretations as much a part of the music as the notes on the page. While performers develop their own approaches to the music, performers also adapt their approaches to one another on the fly in live performance, responding to how others are playing, which allows a group feeling to emerge. Performance strategies develop together among the performers as a group in response to one another and to the music notation.

When faced with difficult or impossible passages, performers must find their own solutions, making decisions about which aspects of the music notation to prioritize and which to ignore. In a particularly dense passage, a performer might choose to play as many notes as possible, to prioritize highest and lowest notes of dense chords, or to arpeggiate impossible chord fingers. By presenting too much information and too much musical material, performers must make choices. I find

[7] See Ames 1987 for a general overview of approaches to algorithmic composition including rule-based systems and constraint solving searches. For a specific discussion of a *protocol* or ranking system of tests used to express an order or preference of musical rules, see Ames 1990.

that the exacting and excessive notation gives me, as a performer, agency to shape the music, which is influenced as much by my preparation as well as by the group dynamics of live performance.

Performance strategies vary from not just from instrument to instrument but from song to song as well. Each instrument presents unique performance challenges, and performance strategies are often guided by the physical constraints and affordances of a particular instrument. Larry Polansky's approach to playing guitar on *Crazy* was to prioritize the highest note in each chord and find some hand shape to approximate the notes underneath it. Andrew Smith similarly described his approach to performing the keyboard part on *Born to Run* as "I arpeggiate randomly and make sure the top notes is right." When performing difficult saxophone parts I'll often transpose difficult leaps, drop notes in dense passages, or balance my mental focus between playing exact pitches and following the pitch contour. It is also important to stress that not all HVB music is impossible. Many of the parts can be worked through slowly and learned just like other music.

These strategies are like search heuristics,[8] a kind of "rule of thumb" that is used to quickly prune notes from an overwhelming space of possibilities — when in doubt play the highest note, or look for the stacked fourths, follow the pitch contour, etc. They are cognitive shortcuts and strategies for quickly parsing the excess of notation information. In this sense, I like to think of the performers as an army of heuristic search algorithms simultaneously finding their ways through a space of too many notes. Aware of one another, their criteria are not necessar-

[8] In computer science, a heuristic is a technique used to guide a search algorithm towards a solution more quickly than exhaustively inspecting every possible option. Heuristics rank or eliminate possibilities according to a "rule of thumb" rather than evaluating every possible option. While heuristics can arrive at solutions quickly, they are not guaranteed to find the best solution, trading optimality for speed.

ily fixed but adjust on the fly and in response to one another. These heuristic strategies also give the music a kind distinct character beyond that of a random flailing of limbs because the character is inherent to the heuristic — if the pianist is looking for stacked fourths, the listeners is going to hear stacked fourths.

Such heuristics could be programmed into the transcription process, but why not let the performers develop them for themselves? Giving performers the agency to develop their own strategies allows for flexibility to adapt on the fly and in response to one another, but, more importantly, it gives each performer space to bring their own perspective to the music, rather than being locked into mine. I like to think that this reflects the nature of human auditory perception. The world is too flush with detail to take in everything at the same time; rather, we prioritize, organize, and focus. I try to give performers the raw results of the analysis algorithms and let them carve their own perspective out of it, offering multiple simultaneous perspectives, or "views," into the data.

Perhaps one of the most difficult aspects of performing HVB music is that the notation is unfinished. The notation is also largely devoid of higher level musical indications. Despite the exactitude of pitch and rhythm, other musical details are often left underdetermined, such as phrasing, breath marks, slurs — details pertaining to the hierarchical organization and grouping of notes into higher order temporal units. The analysis algorithms, for the most part, focus on moment to moment measurements, determining musical details one frame at a time, but not further organizing or grouping into higher order musical objects, such as phrases, gestures, sentences, or cadences. The result is a sense of myopia — an overabundance of lower level details but a paucity of higher level grouping indications — which presents difficulties for the performers, since they are not given cues how to shape the music. It is not always clear from the music when to breath, where put emphasis, when to start and stop phrases, how to inflect and articulate gestures, or if to shape the music at all — perhaps the best approach would be to stoically play the notes and let these things coalesce in the listener's perception?

As a performer in the ensemble, I approach my saxophone parts in terms of finding phrases within the overabundance of detail. I work through the music slowly, marking it up according to what I determined by ear to be the beginnings and endings of phrases, appropriate places to slur, breathe, and accent. My strategy usually changes from song to song, depending on the notation challenges as well as the role of my part in the music. I generally work through foreground melodic lines carefully note by note. If the figure is part of a horn section or greater ensemble texture, I may be more inclined to follow the pitch contour, which I draw in by hand, perhaps identifying a few important pitches or key moments. Generally, my interpretation reflects what I perceive to be the key characteristics of the music or what I understand to be the function of my instrumental part. I like to think of this process as a Rorschach test, an amorphous blob of musical information; performers find what they want based on their cognitive priors. Despite my preparation, I find that my interpretation often changes in live performance, depending on the feel of the group, and I often practice multiple different interpretation strategies so that I can adjust in live performance.

Chapter 4

Composition: Sound Analysis as a Generative Model

Composing Happy Valley Band music is primarily about listening. I like to think of the entire compositional process — source separation, pitch analysis, onset detection, and even music notation — as an extended act of listening, aided by computational tools of digital analysis and visualization. Importantly, by rendering the results of digital analysis into music and thus back into sound, the process is way of interacting musically with the analysis algorithms. In this chapter, I explain the considerations that guide my compositional decisions when making HVB transcriptions. The chapter is structured around a number of musical examples that illustrate how I approach decision-making. I begin with a brief description of my workflow, then discuss the musical examples, and finally develop the metaphor of sound analysis as a generative model to explain how I think about the process.

4.1 Workflow

I use the term workflow to refer to the sequence of steps through which a song passes during the transcription process and the tools and methods that I use to navigate and organize it. Most of my tools are consolidated in a collection of Python libraries, and I use the DAW Reaper as a frontend graphical interface, which allows me to quickly hear and see results of the transcription process, as well as control aspects of it.

I usually begin a transcription by separating the vocal from the mix before moving on to the individual instruments. Depending on the recording, I use a combination of PLCA and `xtrk` spatial filtering, and the decisions at this point in the process pertain to parameter settings on each of the separation algorithms, as well as the choice of training segments. Once a song is separated into individual instruments, I render each instrument to a separate mono audio file, which are used as the source audio in the following analysis stage.

The pitch and rhythm detection are performed offline using the Python libraries discussed in Chapter 2. There are a number of parameters to set, including onset and offset thresholds, which onset detection functions to use, parameters of the pitch estimation model such as number of pitches to return, harmonic weighting, and thresholds for harmonic suppression. I set these individually for each instrument and store them in a `parameters.py` file, which contains a list of all possible analysis parameters and their values (see Figure 4.1b).

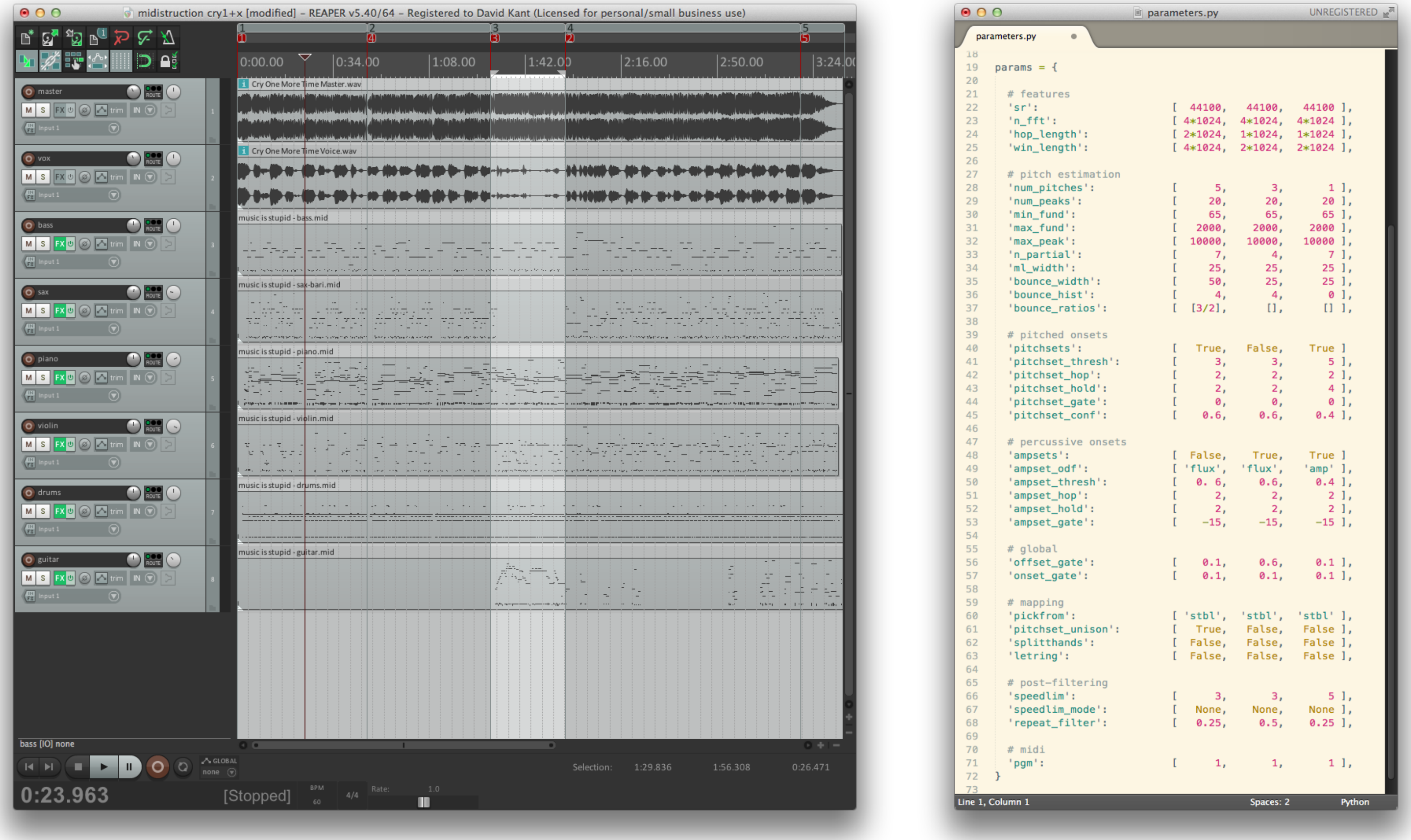
Running an analysis is quick, much faster than realtime — a typical three minute song might take ten to twenty seconds. I use both auditory, visual, and text-based feedback to help me understand the analysis. The results are rendered directly to MIDI files that I use to view and audition the transcription. I monitor a number of basic statistical features, including the number of onsets, the ratio between pitched versus percussive onsets, average note density, time between onsets, and number of notes filtered out. The statistical measures help me understand the impact of changing analysis parameters, because it is not always clear from listening.

It is not uncommon for me to spend a considerable amount of time tweaking analysis parameters and observing the results. Sometimes I plot onset detection functions or other analysis features for visual feedback. Ultimately I rely on my ears as well as sense of musicianship to settle on the analysis parameters, although I will observe the piano roll representation to make sure the transcription is not unreasonably disjoint. With some transcriptions, the initial default parameters work; other times it feels like a protracted battle with the analysis algorithms, chiseling away at shapeless mass of notes until some kind of musical form emerges.

Oftentimes I adjust parameter values throughout a song, between verses and choruses for instance, if the playing style changes radically. It is tempting to focus too narrowly on individual instruments or individual parts of a song, but I try to consider the transcription as a whole and not overfit to individual instruments or short-lived moments of a song. It feels somewhat disingenuous to change the transcription parameters too frequently, in part because it is interesting to hear how different sections of a song drive the transcription algorithms in different ways.

I use the DAW Reaper as a frontend tool (see Figure 4.1a). I insert markers to indicate parameter changes, manage render settings, and to audition playback. Reaper files are saved in a simple file format that is easily parsed, allowing me to automatically extract information such as which section to render and which parameter settings to use. The turnaround in my workflow between annotating parameter settings and rendering the results is quick. While I use Python plotting tools for visual feedback, I often plot analysis data directly in Reaper as waveform data, which maintains alignment between the analysis data and original track while affording the convenience and functionality of navigating audio with a DAW.

Once the analysis parameters are set, I focus on music notation. I have two primary concerns at this point. First, rendering the analysis in music notation helps me identify sections that are excessively difficult or impractical. I do intend the music to be somewhat playable, and the performance concerns by which I judge the music are discussed in Chapter 3. At this point I often notice out of range notes, passages that are excessively dense, or figures that have too many leaps. If the music is too impractical, I return to the analysis stage and adjust parameter settings. Viewing music notation is still a part of the workflow feedback loop, although the music quantizer is relatively slow — it might take 30 seconds to a few minutes to transcribe a typical song, although this largely depends on the size of the beat division scheme list — so I tend to make fewer adjustments at this point. My second concern during the notation stage is quantization, and I adjust the subdivision weights to control the notational complexity of the scores. Since quantization is separate, or downstream, from the analysis data, the notational complexity can be adjusted without changing the music analysis, although it certainly impacts how the performs eventually play it. Often I will go back and forth with musicians at this point, revising parts based on their input about playing techniques and performance limits.


midistruction cry1+x [modified] – REAPER v5.40/64 – Registered to David Kant (Licensed for personal/small business use)
0:00.00
0:34.00
1:08.00
1:42.00
2:16.00
2:50.00
3:24.00
master
vox
bass
sax
piano
violin
drums
guitar
Cry One More Time Master.wav
Cry One More Time Voice.wav
music is stupid - bass.mid
music is stupid - sax-bari.mid
music is stupid - piano.mid
music is stupid - violin.mid
music is stupid - drums.mid
music is stupid - guitar.mid
bass [IO] none
0:23.963
[Stopped]
Selection: 1:29.836 1:56.308 0:26.471
parameters.py
UNREGISTERED
Line 1, Column 1
Spaces: 2
Python

```python

params = {

  # features
  'sr':                 [  44100,   44100,   44100 ],
  'n_fft':              [ 4*1024,  4*1024,  4*1024 ],
  'hop_length':         [ 2*1024,  1*1024,  1*1024 ],
  'win_length':         [ 4*1024,  2*1024,  2*1024 ],

  # pitch estimation
  'num_pitches':        [      5,       3,       1 ],
  'num_peaks':          [     20,      20,      20 ],
  'min_fund':           [     65,      65,      65 ],
  'max_fund':           [   2000,    2000,    2000 ],
  'max_peak':           [  10000,   10000,   10000 ],
  'n_partial':          [      7,       4,       7 ],
  'ml_width':           [     25,      25,      25 ],
  'bounce_width':       [     50,      25,      25 ],
  'bounce_hist':        [      4,       4,       0 ],
  'bounce_ratios':      [  [3/2],      [],      [] ],

  # pitched onsets
  'pitchsets':          [   True,   False,    True ]
  'pitchset_thresh':    [      3,       3,       5 ],
  'pitchset_hop':       [      2,       2,       2 ],
  'pitchset_hold':      [      2,       2,       4 ],
  'pitchset_gate':      [      0,       0,       0 ],
  'pitchset_conf':      [    0.6,     0.6,     0.4 ],

  # percussive onsets
  'ampsets':            [  False,    True,    True ]
  'ampset_odf':         [ 'flux',  'flux',   'amp' ],
  'ampset_thresh':      [   0. 6,     0.6,     0.4 ],
  'ampset_hop':         [      2,       2,       2 ],
  'ampset_hold':        [      2,       2,       2 ],
  'ampset_gate':        [    -15,     -15,     -15 ],

  # global
  'offset_gate':        [    0.1,     0.6,     0.1 ],
  'onset_gate':         [    0.1,     0.1,     0.1 ],

  # mapping
  'pickfrom':           [ 'stbl',  'stbl',  'stbl' ],
  'pitchset_unison':    [   True,   False,   False ],
  'splitthands':        [  False,   False,   False ],
  'letring':            [  False,   False,   False ],

  # post-filtering
  'speedlim':           [      3,       3,       5 ],
  'speedlim_mode':      [   None,    None,    None ],
  'repeat_filter':      [   0.25,     0.5,    0.25 ],

  # midi
  'pgm':                [      1,       1,       1 ],
}

```

(a) Reaper session.

(b) `parameters.py` file.

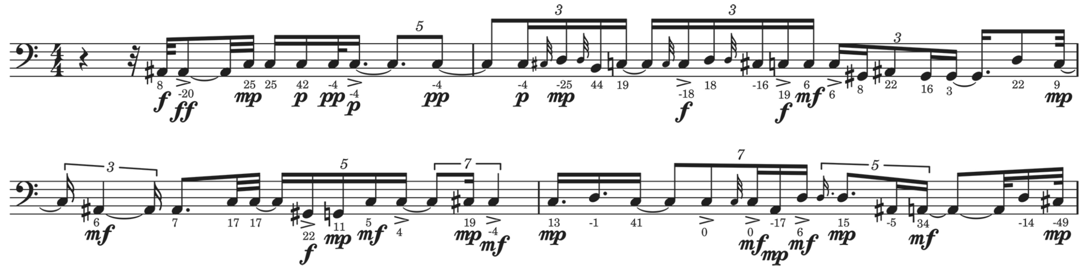
Figure 4.1: Workflow.

4.2 Musical Excerpts

Excerpt 4.1: *It's a Man's Man's Man's World*, Timpani. Excerpt 4.1 is a transcription of the opening timpani roll of James Brown's *It's a Man's Man's Man's World*. While the original recording features a single sustained timpani roll, I wanted the HVB transcription to express the minute fluctuations of pitch and spectra of a single timpani roll that are revealed by spectral analysis. In the HVB transcription, I set the pitch and onset detection thresholds to be very sensitive to minute changes, transforming the simple timpani roll into a complex sequence of irregular rhythms and changes in pitch. The new timpani part is still performed on a single timpani drum, but with constant foot-pedaling adjustments to alter pitch.

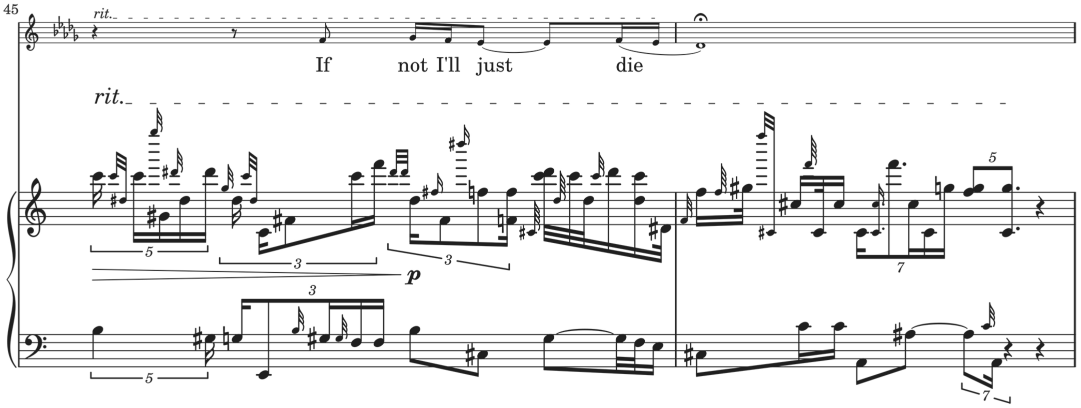
I often use pitch estimation and onset detection to explore the spectral qualities of sound, not just the fundamental frequencies. In this sense, one can think of HVB as Spectralism applied to pop music.[1] While traditions in Western music composition and analysis tend to treat pitch and rhythm as the primary musical features, in much pop music however, timbre is equally if not more important to musical meaning and experience. My intention in transcribing fluctuations in spectra as changes in pitch and rhythm is to articulate the inadequacy of pitch and rhythm to account for what I perceived to be important and semantically meaningful aspects of pop music.

[1] Spectralism is a twentieth century composition technique concerned with the computer analysis of timbre. Spectralism is often associated with the French research center Institut de Recherche et Coordination Acoustique/Musique (IRCAM) and composers such as Tristan Murail, Georg Friedrich Haas, or Gérard Grisey, whose piece *Partiels* (1975), is based on a spectrogram analysis of a single, sustained trombone tone. In many ways, HVB is Spectralism applied to pop music, although this phrase is more a way of explaining and locating the work within a larger tradition of Spectral Music rather citing my motivations. While I do not think of my work as a direct consideration of and response to the Spectral tradition, my awareness of Spectral composer such as Grisey and Murail no doubt provided conceptual and aesthetic contexts for my explorations.


5
3
3
3
8 f -20 ff 25 mp 25 42 p -4 pp -4 p -4 pp -4 p -25 mp 44 19 -18 f 18 -16 19 f 6 mf 6 8 22 16 3 22 9 mp
3
5
7
7
5
6 mf 7 17 17 22 f 11 mp 5 mf 4 19 mp -4 mf 13 mp -1 41 0 0 mf mp -17 6 mf 15 mp -5 34 mf -14 -49 mp

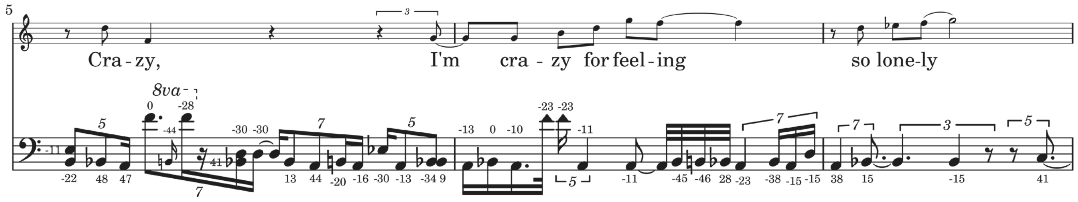
Excerpt 4.1: *It's a Man's Man's Man's World*, Timpani

Excerpt 4.2: *This Guy's in Love with You*, Piano. Excerpt 4.2 shows two bars of music from the piano feature in the HVB transcription of *This Guy's in Love with You*. The music is full of multi-octave leaps and notes that are very short in duration, representing a more extreme example of the kind of complexity to be found in HVB music. The difficulty is intentional; I wanted the piano transcription to reflect the dense, active playing of the original recording, in particular because the song is a piano feature, and because the HVB version was written for virtuosic pianist Joe Kubera. Rather than increase the sensitivity of pitch and onset detection thresholds as in the previous musical excerpt, for this song I increased the number of voices returned by the pitch estimation algorithm. As discussed in Chapter 2, the pitch estimation algorithm ranks a set of possible pitches according to likelihood of being perceived. Increasing the number of voices causes the algorithm to reach further down the list of possible pitches, returning pitches with less and less likelihood. This is more or less equivalent to lowering the threshold of pitch confidence below which a pitch is suppressed. The short-lived notes and abrupt leaps in register are largely a result of collapsing multiple pitch tracks to a single musical voice.



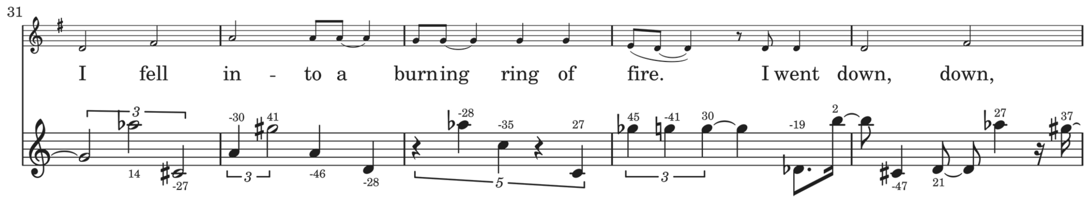
Excerpt 4.2: *This Guy's in Love with You*, Piano

Excerpt 4.3: *Crazy*, Upright Bass. Occasionally the pitch estimation algorithm returns an extremely high and spurious note, far beyond the playing range of the instrument. This is often due to the presence of high-frequency noise in the signal and the absence of low frequency energy, which causes the peak-picking algorithm to concentrate many peaks towards the upper registers. While notes beyond the range of the instrument could be removed or octave reduced to fit the range of the instrument, I like to maintain some indication of their occurrence. In the upright bass part to *Crazy*, these pitch estimation artifacts are notated with the indication "as high as possible," denoted by the triangle notehead, a notational idea that I borrow from composer Christian Wolff. Asking the performer to play as high a note as possible preserves the contour of the pitch estimation but also translates the artifact into a semantically meaningful notation. It is less important which pitch class was reported, and more important to reflect that an artifact occurred.



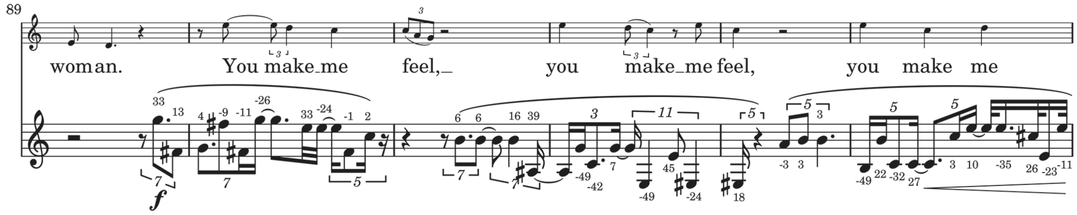
Excerpt 4.3: *Crazy*, Upright Bass

Excerpt 4.4: *Ring of Fire*, Electric Guitar. In the original recording of Johnny Cash's *Ring of Fire*, the electric guitar drives the music almost more like a percussion instrument than like a pitched instrument, playing palm-muted strokes at an even quarter note rhythm. Wanting to reflect this in my transcription, I weighted the onset detection thresholds more heavily in favor of percussive onsets than pitches onsets. The transcribed pitches, however, deviate from what one might expect. This is caused by the envelope of plucked string notes. The attack of a plucked guitar strings is generally marked by a brief noise transient before settling in to a stable pitch. During the noise transient, the perception of pitch is ambiguous and difficult to measure, causing the pitch estimation to give erratic results. A common solution to this problem is to use a slightly delayed pitch estimate, allowing the pitch tracker time to settle into a more confident estimation. Fascinated by the problem, I use the pitch estimation more or less concurrent with the percussive onset.



Excerpt 4.4: *Ring of Fire*, Electric Guitar

Excerpt 4.5: *(You Make Me Feel Like) A Natural Woman*, French Horn. I wish I could say the collection of HVB songs represents a particularly meaningful and considered statement about the history of popular music, but the truth is song selection is driven by performers. I, or more accurately *we* — the entire band usually weighs in on song selection — choose songs with particular musicians in mind. I try to maintain a diversity of style, genre, time period, and production techniques, but ultimately performs and instrumentation is the guiding factor. *When the Levee Breaks* (2018) was written for Sam Friedman, a harmonica player with whom I first worked on the recording of *ORGANVM PERCEPTVS*. Sam recorded four bars of harmonica music remotely, and I made a mental note to work with him again if the opportunity arose. *This Guy's in Love with You* (2012) was written for the composer, performer, and sound artist Gordon Monahan. Gordon witnessed the first ever performance of HVB in Ostrava, Czech Republic, and immediately invited us perform at his *Electric Eclectics Festival* in Meaford, Ontario. The piano part was later revised for pianist Joseph Kubera for the New York recording session. *After the Gold Rush* (2013) was written for Thomas Verchot, a New York City based trumpeter. *(You Make Me Feel Like) A Natural Woman* (2015) was written for French hornist Daniel Costello. After playing a particularly demanding French horn part in an orchestra piece of mine, Daniel and I stayed in touch. Excerpt 4.5 shows a few measure of the French horn part. In addition to the microtonal indications and demanding rhythms, the part features difficult slurred leaps across breaks in the instrument and requires the performer to play demanding passages at the lower and higher extremes of the instrument's range. These kinds of challenges are typical of HVB music.

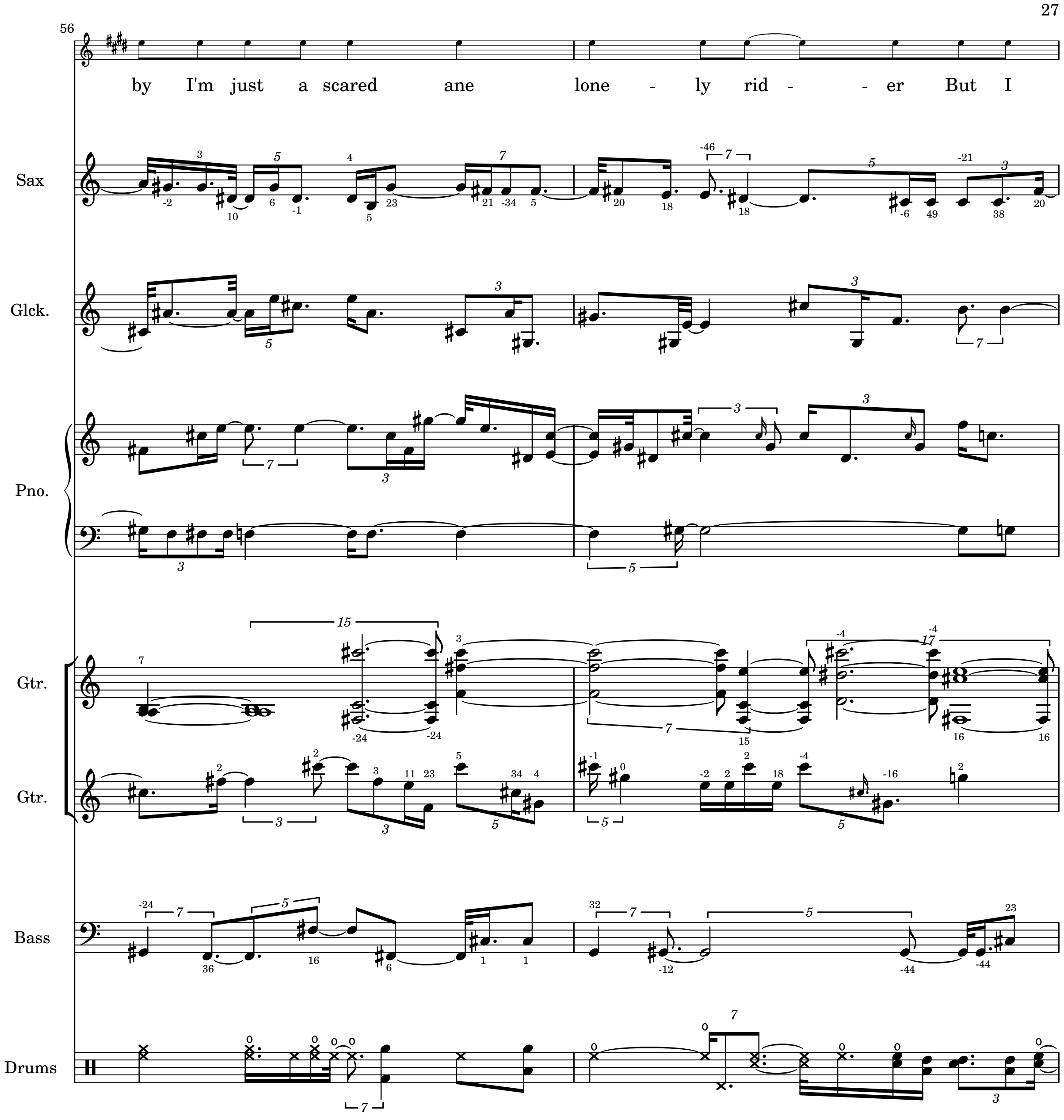


Excerpt 4.5: *(You Make Me Feel Like) A Natural Woman*, French Horn

Excerpt 4.6: *Born to Run*, Full Score. Production qualities of the original recordings have a tremendous impact on HVB transcriptions. Aspects of production and mixing — panning, spatial effects, reverb, compression, equalization, and balance between instruments — all affect HVB music, down to the notes and rhythms transcribed. Instruments that are difficult to isolate produce noisy signals that complicate the downstream pitch and onset analysis stages, causing spurious notes and other pitch and onset detection artifacts. Because of this, I consider separation to be part of my compositional process. When setting analysis and separation parameters I am mindful of how separation artifacts might affect pitch and rhythm analysis. Rarely do I fiddle with separation parameters in effort to change notes and rhythms, but I do bear in mind my observations from one transcription to the next.

Due to the mixing and production techniques of the original recording, Bruce Springsteen's *Born to Run* is perhaps the most difficult mix I tried to separate. The recording is of a loud distorted rock band, the mix featuring layered guitar tracks, heavy use of compression, and distortion effects on both guitars and bass. Many of the instruments in the mix — guitars, bass, saxophones, organs, and other keyboards — overlap in frequency, particularly in the high mid range, masking one another. Overlapping spatial positions and stereo effects further blend the instruments together. In short, the production technique is a wall-of-sound style approach, all instruments woven together in a distorted, ringing mass of sound. I had difficulty differentiating one instrument from another by ear or telling exactly which instruments were in the mix. The separated tracks are full of noise and crosstalk, and the transcribed music reflects it. A kind of musical noise floor permeates the entire transcription — pitch and rhythm artifacts of the analysis process — and musical figures are shared between instruments, sometimes in unison and other times spread across many instruments like a hocket.

Ultimately, some songs work and some songs do not. I can't really explain how or why. Sometimes the process results in music that is remarkable, fascinating, surprising, or astounding, and sometimes it simply does not. The composition process — training source separation models, setting pitch estimation parameters, mapping onset features to musical features, setting detection thresholds, and weighting notation schemes — is ultimately guided by intuition and by my ears. For me, sound analysis is way to explore a signal, and I try to bear in mind both the perceptual implications as well as mathematical features measured. I use these tools to search for aspects of sounds that I find interesting or unexpected and I try to focus the analysis on the features that I understand to be most salient and substantial to the music.



Excerpt 4.6: *Born to Run*, Full Score

4.3 A Generative Model of Machine Perception

A metaphor that I find useful for describing my compositional approach in HVB is that of a generative model, an idea borrowed from statistical modeling. I've come to think of HVB as a generative model of sound analysis. In statistical modeling, models generally fall into one of two categories: *generative* or *discriminative*. While both types can perform similar tasks, such as classification, they differ in terms of internal structure and operation. A generative model learns a fuller description of the phenomena it models, allowing the model to generate new samples of that phenomena, where as a discriminative model cannot generate samples of the observed phenomena. Formally, a generative model learns the joint probability $P(X, Y)$ of the input variable X and label Y, where as a discriminative model learns the conditional probability $P(Y|X)$ of the label Y given the input variable X. The distinction is between a model that learns to *describe* phenomena and one that learns to *produce* new instances of that phenomena.

A simple example would be that of an image classifier, a model that learns from a dataset to classify input images as belonging to one of a given set of output categories, such as cat, dog, or mouse. While the discriminative model will learn to classify images, the generative model will learn not just to classify images but to generate new images of a given category as well — a new image of a cat, dog, or mouse. The ability to generate new images requires the discriminative model to learn a representation of the input data based on the features of the images themselves, such as what visual features constitute a cat, dog, or mouse, where as the discriminative model might simply classify images based on the relative similarity of one image to all the others. Generating new instances of a phenomena requires more information, or a fuller representation, of the phenomena modeled.

Dispensing with precise statistical definitions and employing the distinction rather as a metaphor, it occurs to me that machine listening algorithms are often understood to *describe* sound. Instead, I like to think the HVB machine listening algorithms as a kind of generative model, a means of *producing* new sounds and new music. Rather than passive devices that translate sound into an appropriate musical representation, I think of machine listening algorithms as active devices of sound production. This frame describes my mental state when composing HVB music. I think of the parameters of the machine listening algorithms more like parameters of a complex generative model. The compositional process starts with a consideration of the content of the original recordings, but at some point becomes more like exploring the possibility space of a high-dimensional parametric model.

The theory of *constructivism* provides a framework for considering how machine listening models function as active devices. Often associated with Jean Piaget (1896–1980), a Swiss psychologist who studied cognitive development in children, constructivism is a way of thinking that extends to many domains of thought, from philosophy of mind, to learning theory, social theory, mathematics, and perception. Piaget argued that learning is an active process in which individuals assimilate new information from the environment into a prior held mental frameworks. Key is the role that the prior held mental frameworks plays in constructing a subjective representation of the world. Interpreted in the domain of perception, be it visual or auditory, constructivist theories of perception stress the role the sensory and cognitive mechanisms play in constructing a subjective experience of the world. This is antithetical to theories of direct perception, in which the sense are believed to provide direct access to reality.

I cannot help but see machine listening from a constructivist perspective. The algorithms that I use — as well as those used by others — contain a tremendous amount of information about the phenomena they purport to describe. A pitch estimation model that looks for harmonically related chords, for instance, will find harmonically related chords. The results are determined as much by the internal structure and operation of a model as by the signal that is analyzed. My choices, including which parameters and how to set them, which analysis features to use, and how to structure signal flow within the transcription system, all significantly shape the resulting music. Put simply, how the model is wired affects the music that comes out, and changing the model changes the results.

Importantly, the metaphor provides, for me at least, a way of seeing HVB music as emergent between the input signal and internal structure of the analysis models, residing not entirely in one or the other but in the interaction between the two. The compositional decisions I make are less about reflecting a direct relationship with the original songs. Rather I think of the transcription process as a system in which the original recordings perturb a model of perception. It strikes me that when considering discourses about the agency of technologies, the language of emergence is useful for articulating how the values and beliefs embedded within tools assert agency on users. The relationship does not manifest as a directly observable transfer of value from technologist to user, but as a more complex interaction between the two. The study of how tacit sources of power and agency embedded within technologies assert influence could benefit from dynamical systems theory tools used to identify and measure mutual influence in coupled systems.

Chapter 5

Recording and Release

In March 2017, I released *ORGANVM PERCEPTVS*, an album of eleven songs of the Happy Valley Band. The collection spans a variety of pop music genres, artists, time periods, and production styles, including country, funk, pop, and rock (Table 5.1 lists the songs included on the release). The music was recorded in fall 2015 at the Bunker Studio in Brooklyn, NY, with additional small ensemble work and overdubs continuing into winter 2016. A few parts were recorded remotely, by musicians in Los Angeles, Providence, and Germany, and then layered into the mix. The project was mixed in winter 2016 by Joseph Branciforte of Greyfade Studio in Brooklyn, NY and mastered in early spring of the same year by Cookie Marenco at OTR Studio in Belmont, CA. In addition to the core ensemble, the recording features an expanded personnel of twenty musicians, many of whom had not played with HVB prior. The album was released on vinyl LP in March 2017 by Santa Cruz based record label Indexical and includes an accompanying 5,000 word liner notes essay "The Long Answer" in which I explain the project, my motivations, and compositional process, as well as reflect on broader cultural implications of automation technologies such as machine learning. The vinyl LP, liner notes, and album packaging are pictured in Figure 5.1. See Appendix D for a complete list of album personnel and recording credits.

5.1 Recording *ORGANVM PERCEPTVS*

The recording project presented a number of new challenges, among them, how to incorporate new performers into the ensemble. Up to this point, HVB consisted of a core group of seven musicians with whom I worked closely over the course of five years, from 2011 to 2016, to develop the ensemble performance practice. *ORGANVM PERCEPTVS* required additional musicians to meet the instrumentation of all of the songs. How would I integrate new performers into this dynamic group who together had developed a shared understanding of the project? Would the performance practice be comprehensible to new musicians? Would the introduction of new performers alter the group dynamics? While many of the new performers were already familiar with the project, some were not, and explaining the performance practice to a freelance musician over the phone was difficult. Additionally, the expanded instrumentation presented new playing challenges. Two songs in particular, *Ring of Fire* and *(You Make Me Feel Like) A Natural Woman*, include a trio of vocalists, singing backup (not lead) vocal parts. How should vocalists approach performing HVB music? The erratic jumps in pitch present different performance challenges for vocalists than for instrumentalists.

The recording project also raised interesting questions about recording, production, and mixing. I felt as though every assumption and convention of recording and production was up for reconsideration. The most complicating factor by far was that our lead singer is fixed on a tape track lifted from another recording, sung to another performance. Should we attempt to mimic the production styles and recording techniques of the original recordings? How? And to what extent? Should we use similar instruments, amplifiers, and microphones? Should we mimic the panning, effects, reverbs, and spatial profiles of the original mixes? In pop music, production can be as important to the character and identify of a song as the notes and lyrics, not to mention that recording techniques also bear on the success of the transcription process, affecting the notes and rhythms of the music that we play. Initially I intended to mimic the originals. We matched the original instrumentation and used similar instruments when tracking — we rented a clavinet for one song, used an organ and Wurlitzer piano when appropriate, switched between acoustic and electric guitars and basses, re-miked drums sets to fit rock and jazz styles, matched auxiliary percussion, brought in flugelhorn and French horn players, and even called in a harmonica overdub for just four bars of music. In a few cases, we exaggerated the instrumentation. The B section of *Like a Prayer* is marked by auxiliary percussion, but I had difficulty identifying the percussion instruments in the original recordings, so I scored the HVB percussion part for a battery of skins, woods, and metals, leaving it to the discretion of the performer. We also took a few liberties, such as substituting saxophone in *In the Air Tonight* for a synthesizer patch that sounds like horns.

Although I had intended to mimic the mix styles of the original recordings, this reached a limit during mixing. The original recordings were mixed to suit different performances, different bands, and different music; we needed to mix to fit our music, not constrained to match. As a result, some of the mixes are similar to the originals and others are quite different. In many of the mixes we exaggerated one or two characteristic, identifying, or otherwise interesting production techniques, much like a caricature portrait of the original song. We exaggerate the gated snare reverb on *Like a Prayer*; the final HVB mix of *Born to*

Run is saturated with compression in a nod to the E Street Band's wall-of-sound style guitar tracks; and in *In the Air Tonight* we manually mimic a tape delay effect by staggering multiple violinists one beat apart. In the original version of *Like a Prayer*, bass guitarist Guy Pratt doubled the bass guitar part in unison on Minimoog synthesizer, a common production technique in pop music at the time. We mimic this production technique in the HVB version, using a pitch tracker and software Minimoog emulation.

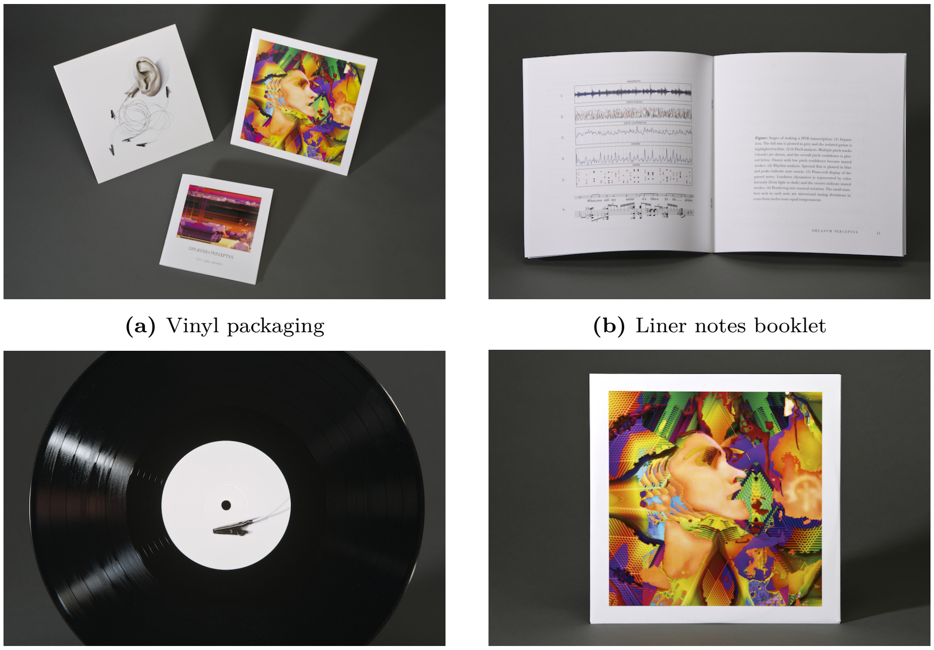
This project has also been an unexpected opportunity for me to engage with some of the history of these recordings. I stumbled upon a number of fascinating stories, often when deciding which version of a song to transcribe, such as the unusual fade out in Elvis Presley's *Suspicious Minds.* Towards the end of the track there is a long fade out, as though the song is ending, but then the music fades back in and a section of the verse loops for another 30 seconds before fading out a second time, this time for good. According to Chips Moman, who produced the recording, the fade is the result of disagreement between Moman and Elvis' longtime producer Felton Jarvis. Jarvis was not happy about the session, and, before releasing the track, added the exaggerated fade, perhaps to mimic the manner in which Elvis performed live at the time.[1] I find it fascinating that a recording can be an artifact of and index to these personal stories, and this is something that I like to consider when approaching a transcription. With *Suspicious Minds* I chose to transcribe the fade itself — rather than recovering the music and applying a new fade. As the music fades out, I allowed the tracks to gradually fall below the thresholds of the pitch and onset detection algorithms, causing the transcriptions to drop out erratically as the fade crosses the edge of

[1] Marc Myers. 2012. "Caught in a Trap: Elvis's last No. 1 Hit." *Wall Street Journal (Online).*

audibility and then enter back in.

	Song Title	**Artist**	**Year**
A1	*Like a Prayer*	Madonna	1991
A2	*Ring of Fire*	Johnny Cash	1963
A3	*Jungle Boogie*	Kool and the Gang	1973
A4	*(You Make Me Feel Like) A Natural Woman*	Aretha Franklin	1968
A5	*Crazy*	Patsy Cline	1961
A6	*Suspicious Minds*	Elvis Presley	1969
B1	*It's a Man's Man's Man's World*	James Brown	1966
B2	*After the Gold Rush*	Neil Young	1970
B3	*In the Air Tonight*	Phil Collins	1981
B4	*This Guy's in Love with You*	Herb Alpert	1968
B5	*Born to Run*	Bruce Springsteen	1975

Table 5.1: *ORGANVM PERCEPTVS* track list.



(a) Vinyl packaging

(b) Liner notes booklet

(c) Vinyl record

(d) Inner sleeve

Figure 5.1: *ORGANVM PERCEPTVS* vinyl packaging.

5.2 Release and Public Response

I released *ORGANVM PERCEPTVS* in March 2017 to polarized public response. Some reviewers were enthusiastic. *The Wire*, an international magazine for avant-garde and experimental music, quickly picked up the project, publishing an interview and exclusive pre-release album stream. Interviewer Emily Bick described the music as "refracted and amplified through the software in often mystifying ways, resulting in warped interpretations that are unexpected, to say the least."[2] The growing online music distribution platform and media outlet *Bandcamp*, listed the release second in their article "Meet the Artists Using Coding AI and Machine Language to Make Music," describing the album as a "skewed, jittery cacophony. . . equal parts bewildering and inspiring, highlighting how AI can help humanity see the familiar from a fresh perspective."[3] The project was featured by numerous other new music publications including *I Care if You Listen*, *Sequenza 21*, *Experimental Music Yearbook*, and *Tiny Mixtapes*, who called the project "pop music's post-human future. . . delirious and discordant."[4] Perhaps my favorite description came from *New Classic LA* writer Elizabeth Hambleton who wrote "I am convinced [it] is an actual recording of my high school's pep band at a snowy football game when every brass instrument detuned after five minutes out of their

[2] Emily Bick. 2017. "Album stream and interview: David Kant of Happy Valley Band talks about their 'machine listening' album." *The Wire.* Accessed April 27, 2019. `https://www.thewire.co.uk/in-writing/interviews/listen-to-the-happy-valley-band-s-new-album-and-read-an-interview-with-its-founder`.

[3] Simon Chandler. 2018. "Meet the Artists Using Coding, AI, and Machine Language to Make Music." *Bandcamp Daily.* Accessed April 27, 2019. `https://daily.bandcamp.com/2018/01/25/music-ai-coding-algorithms/`.

[4] Colin Fitzgerald. 2017. "Happy Valley Band deconstruct pop classics via machine-learning algorithm on debut album ORGANVM PERCEPTVS." *Tiny Mix Tapes.* Accessed April 27, 2019. `https://www.tinymixtapes.com/news/happy-valley-band-deconstruct-pop-classics-machine-learning-algorithm-debut-organvm-perceptvs`.

cases. This album is fresh, deceptive, and insanely fun to listen to."[5] See Appendix B for a list of press and review publications.

Common to many reviews is the sense that the project is both confounding and difficult to listen to — reviewers described the music as "mystifying," "bewildering," "unexpected," as well as "cacophony," "warped," "discordant." These reviews capture a common response to HVB: many people are unsure what to make of it. This is due, I think, to the nature of the project; it is a sprawling and complex idea, a mix of motivations, both technical, aesthetic, and cultural. It is not easy to explain in a succinct manner what HVB is or what is involved in making the transcriptions, why the are "wrong," or what exactly "wrong" might mean, let alone why and how musicians play the music — Is this a big practical joke? Are they making it up? Are they improvising? Why would anyone even do this? The project is simultaneously humorous, facetious, and endlessly serious and involved. Ultimately, most of the reviews resolve in a positive and optimistic valence, describing the project as "fun," "fresh," "inspiring," and even "valuable to humanity."

One reviewer, however, was more conflicted. Having agreed relatively early on to review the project, I emailed to follow up after a few weeks had gone by without communication. Over email the reviewer explained, while they had agreed to review the project with every expectation of enjoying the music, upon listening to the record, they simply did not know what to do. The reviewer wrote, almost apologetically, "I find it a pretty conceptually fascinating record, but I simply cannot bear to listen to it"[6] In their review, "You, With the Violin!

[5] Elizabeth Hambleton. 2017. "Happy Valley Band's debut album ORGANVM PERCEPTVS." *New Classic LA*. accessed April 27, 2019. `http://newclassic.la/2017/04/27/review-happy-valley-bands-debut-album-organvm-perceptvs/`.

[6] Chris Zaldua. Email message to David Kant. April 4, 2017.

Sight-Read These Computer Algorithms!" they describe the music "like nails on a chalkboard," indicting the project for being, simply put, "bad. Quite bad." At the end of the review, the author raises a question that I find deeply fascinating: "Actually, nails on a chalkboard sounds much better to me; what does it mean that I prefer pure cacophony to off-time, off-kilter pop music?"[7] This caused me to wonder, what is it about this project that some people find so viscerally off-putting?

5.3 Implications for Music Theory and Perception

5.3.1 The Uncanny Valley

Although it was not my intention at the outset of this project, the concept Uncanny Valley provides a useful and difficult-to-ignore metaphor for understanding responses to HVB — the term is to some extent a trend or buzz word in digital culture at the present moment as well as in electronic and digital art. The term, proposed by Japanese roboticist Masahiro Mori in a 1970 article "The Uncanny Valley," refers to an unsettled feeling that humans experience when artificial representation closely resemble human beings.[8] Generally, the affinity that humans feel for an artificial representation increases with the likeness of that representation, except for a critical point, a characteristic dip or valley, as the representation nears resemblance to a human being, and the feeling of affinity gives way to an

[7] Chris Zaldua. 2017. "You, With the Violin! Sight-Read These Computer Algorithms!" *KQED*. accessed April 27, 2019. `https://www.kqed.org/arts/13038360/you-with-the-violin-sight-read-these-computer-algorithms`.

[8] Masahiro Mori. 2012. "The Uncanny Valley." Translated by Karl F. MacDorman and Norri Kageki. *IEEE Robotics and Automation Magazine* 19, no. 2 (June): 98–100.

unsettling sense of strange familiarity, or uncanny. This relationship is plotted in Figure 5.2. The implication is, if automation and AI technology are designed to model the likeness and modes of human operation, then there is a point at which users will exhibit an overwhelmingly negative response. Over the past five to ten years, there has been renewed focus on the uncanny, in the fine arts, design, and technology, as well as a proliferation of scientific studies that attempt to verify, disprove, or explain its cause.

Perhaps the Uncanny Valley pertains to the unsettled feelings of bewildering cacophony that listeners experience of HVB. Perhaps HVB music sits at this critical point of close but not close enough renderings of familiar songs and familiar forms, inducing distress and disorientation. I often receive comments about the Uncanny Valley in response to HVB, and I like to think that if the Uncanny Valley is relevant, then explanations of the Uncanny Valley would shed light on HVB, and maybe even vice versa. Ultimately, I cannot help but wonder why the Uncanny Valley might exist in the first place. Numerous explanations have been proposed, from biological motivations to perceptual and cognitive confusion such as conflicting cues, but I like to think that it is biological warning, a line not to be crossed, an evolutionary imperative that technology be used for its own logic rather than to replicate that of humanity. The Uncanny Valley challenges the assumption that technology should be like us, an assumption that I believe motivates much development in technology, especially in the commercial and public spheres.

Interestingly, the application of Uncanny Valley rests on a key difference between HVB music and that of Ferneyhough, Cage, and much other contemporary and new music. There is something against which to compare HVB transcriptions: the original song. With Ferneyhough, for instance, the musical object itself is abstract; the listener has no access to or prior familiarity with the sound object as Ferneyhough imagines and notates it, other than through a performer's realization. A listener does not know how accurately a performer plays Ferneyhough. This is not the case with HVB. Many listeners are familiar with the original songs that we perform, or at least familiar with the form of pop songs in general, perhaps even a particular genre or artist. There is a sense of what the transcriptions *should* sound like, were they to be "correct."

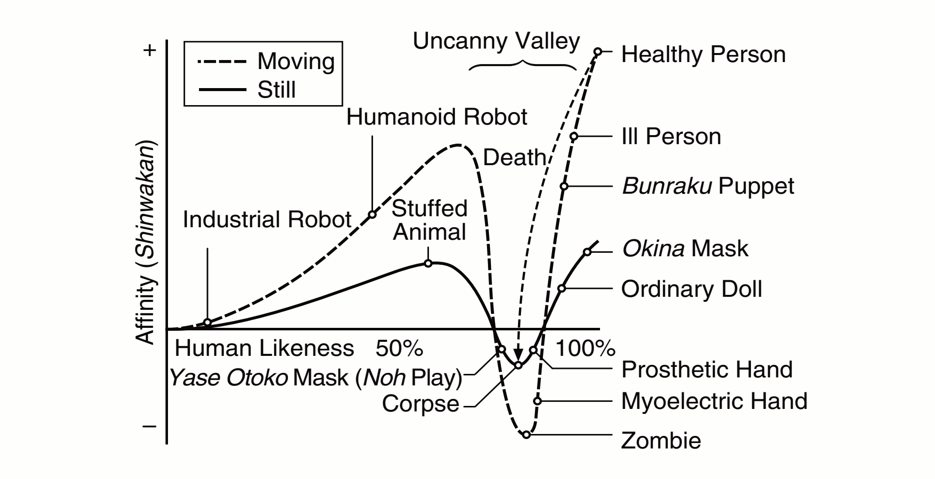


Figure 5.2: Plot showing the Uncanny Valley. The human affinity for a simulation is plotted versus the human likeness of that simulation.[9]

5.3.2 *Meta+Hodos*

Another framework that I cannot help but see in the project is one proposed by James Tenney. Tenney's 1961 Master's thesis *Meta+Hodos* (published in 1984) is an application of Gestalt theory of psychology to music. Premised on the increased complexity of twentieth century composition, in particular the observation that any sound could now serve as the fundamental building block of music, Tenney looked to Gestalt theory to explain how humans organize complex auditory phe-

[9] Reproduced from Mori, *Uncanny*, 99.

nomena. Tenney argued that the fundamental unit of music had changed, citing in particular the music of composers such as Ives, Webern, Bartók, Ruggles, and Schoenberg. Individual "tones" were no longer the only building blocks, but complex sounds and "sound-configurations," even those with considerable variation in time, now had "equal potentiality for use as the elemental building-materials in music."[10] Early twentieth century Gestalt theorists, such as Kurt Koffka or Max Wertheimer, similarly attempted to explain, in the visual domain, the laws by which human visual perception hierarchically groups together elements into wholes, or "gestalts." In Gestalt theory, the whole is something new, taking on an identity distinct or other from the sum of its parts, and the role elements play are contingent on their context and arrangement, not on intrinsic properties. Acknowledging the insufficiency of established ideas in music theory to account for the growing complexity of music, Tenney applied principles of Gestalt psychology to develop an expanded conceptual framework for twentieth century music

Tenney identifies a number of factors that contribute to the impression of discontinuity when listening to a piece of music, including *focus* and *scale*, and I cannot help but wonder what role they might play in experiences of HVB. Tenney's first factor, *focus*, refers to which aspects of sound a listener attends to. If a listener directs their attention towards the less essential parts of a complex sound texture or less essential features[11] of a musical figure, they may, understandably, miss important structural aspects of the musical idea and experience discontinuity. Tenney's use of focus is motivated by the observation that in twentieth century music, features other than pitch articulate musical ideas more so than in earlier musics. Tenney uses the second term, *scale*, not in its ordinary musical context

[10] James Tenney. 1986. *Meta+Hodos: A Phenomenology of 20th-Century Musical Materials and an Approach to the Study of Form.* Edited by Larry Polansky. Frog Peak Music, 10.

but rather in a sense more akin to visual perception, offering the analogy of viewing an image from too close or too far. Simply put, scale refers to being too far zoomed in or out to capture structuring ideas. An incongruity between the scale of detail at which a listener is oriented towards distinguishing elements from larger configurations and the scale at which the piece is organized will produce an experience of discontinuity.

Tenney's concept of discontinuity offers a possible explanation for the discord, cacophony, and confusion that reviewers expressed in response to HVB. Both scale and focus are treated in unusual ways. The listener is not listening to the original song directly, but rather to an interpretation of that song, the results of a perceptual model. The focus and scale of that analysis are fixed, and arguable at a scale not consistent with human perception. The choice of onset detection functions (Chapter 2.5), for example, determines the parameters of focus — ultimately, my best justification of how I choose them rests on what I perceive to be interesting. Scale is similarly fixed and cast at low level details, perhaps zoomed further in than a human listener would. Single notes are transcribed into short melodic figures, and fluctuations in pitch that would ordinarily be considered embellishment are transcribed as though they are separate notes.

Fundamental to Tenney's theory is the formation of perceptual units, called *clangs*, which are further grouped hierarchically into larger units called *sequences*, eventually reaching the level of *form*. Importantly, anything may function as an element of a larger configuration if it is perceived as an element. Tenney identifies a number of factors responsible for the formation of perceptual of units and for

[11] I use the word "feature" synonymously with Tenney's use of the word "parameter" to maintain consistency with my earlier discussion of sound analysis features in Section 2.5. In my terminology, I distinguish between feature and parameter: features is a perceived aspect of sound, where as parameter is numerical quantity of a system or model.

the separation of one unit from another. There are two *primary* factors (*proximity* and *similarity*) and four *secondary* factors (*intensity*, *repetition*, *objective set*, and *subjective set*). My observation is none of these factors operate strongly in HVB music, and perhaps their absence is the reason the music is bewildering.

The first primary factor is *proximity*: sounds that are simultaneous, contiguous, or close together in time will tend to form groups, while separation in time will produce segregations. Figure 5.3a shows a simple example using alphanumeric characters. The zeros cluster together because they are relatively close together in spacing. Figure 5.4a shows a similar musical example. Similarly, three separate groups form; the groups are internally contiguous yet separated from one another by relatively large distances. The second primary factor is *similarity*: sounds that are similar to one another will tend to form groups, while dissimilarity will produce segregations. Again, Figure 5.3b shows a simple alphanumeric example. Despite the absence of blank space, characters form visual groups according to their similarity. Figure 5.4b shows a musical example in which three separate groups cluster due to relative similarity in range, even though there are no breaks between groups.

```
00   000     00      0000    00   000
```

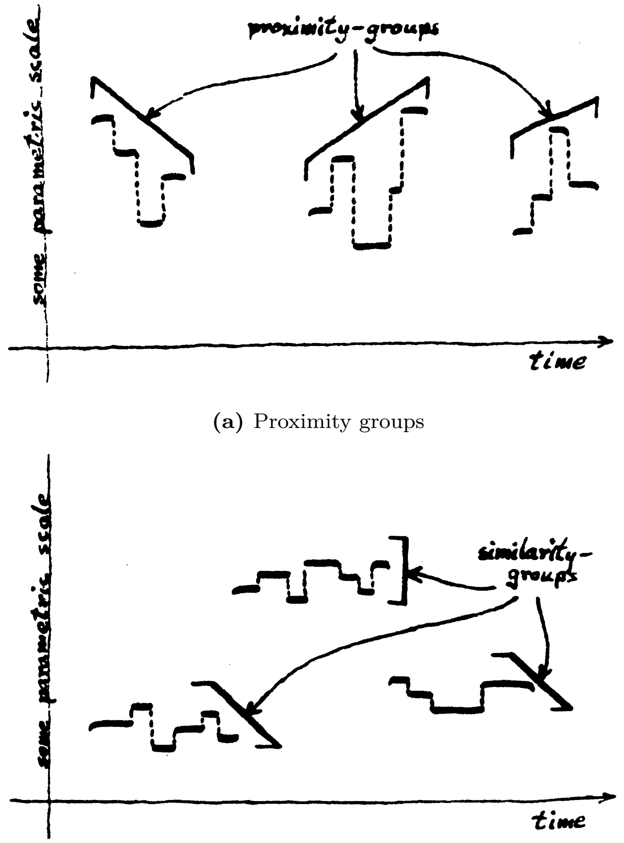
(a) Proximity groups

```
0 0 0 # # # 0 0 0 0 # #
```

(b) Similarity groups

Figure 5.3: Illustrations of Tenney's proximity and similarity factors.[12]

[12] Reproduced from Tenney, *Meta+Hodos*, 28 and Tenney, *Meta+Hodos*, 29.



(a) Proximity groups

(b) Similarity groups

Figure 5.4: Musical examples of Tenney's proximity and similarity factors. Plots represents the intensity of some musical parameter through time.[13]

Many passages in HVB music are undifferentiated sequences of notes that resist the parsing described by Tenney's factors of proximity and similarity. Notes jump erratically between registers and are strung together without interruption. Often there is little coordination between parts, and factors of similarity or proximity would do little to find organization within and between the multiplicity of voices. As discussed in Section 3.3, performers are left to shape the music themselves in the absence of phrasing and articulation indications. This often leads

[13] Reproduced from Tenney, *Meta+Hodos*, 34.

to perceptual grouping cues that are inconsistent between performers or between various features of sound.

Perhaps the most interesting of Tenney's secondary factors, at least in the context of HVB, is the concept of a *set*. Tenney uses the term to mean a prior psychological attitude or perspective which determines or alters one's perception. The *objective set* refers to expectations or anticipations arising *during* a musical experience, which are produced by previous events occurring within the same piece; where as the *subjective set* refers to expectations or anticipations that are a result of *previous* experience, separate from the musical work. Objective set factors that contribute to feelings of expectation and anticipation include internal coherence, structure, inertia, the establishment of specific referential norms, thematic reference, recurrence, or recall. Most of these factors are not present in HVB music. HVB music in general does not provide many cues about what will come next, what to expect, or what to anticipate — beyond, of course, broad brush strokes pertaining to the form and structure of pop music.

Tenney's theory provides a framework for understanding how and why HVB might be considered cacophony. By Tenney's factors, much of the music is, functionally considered, noise — it does not congeal into structures of clang, sequence, form; it does not establish internally coherent patterns and expectations. Despite its seeming lack of coherence, however, the music is not entirely disordered either, and Tenney's theory provides insight about how and why. I like to think of HVB more like random deviation from a form, akin to low amplitude high frequency noise in the presence of a signal. The general form is still present as long as the noise and signal are balanced, but as the noise amplitude increases, the form dissolves.

5.4 Automated Copyright Detection

On February 26, 2017 Indexical received an email notification of possible copyright violation. The email came from DistroKid, a digital music distribution service that enables artists to sell and stream music through large online retailers such as iTunes, Spotify, and Tidal. Many streaming music platforms require a third party service such as DistroKid for artists who are not represented by a major label, although this is beginning to change, with platforms such as Spotify for Artists. In response to HVB's application for distribution, DistroKid responded:

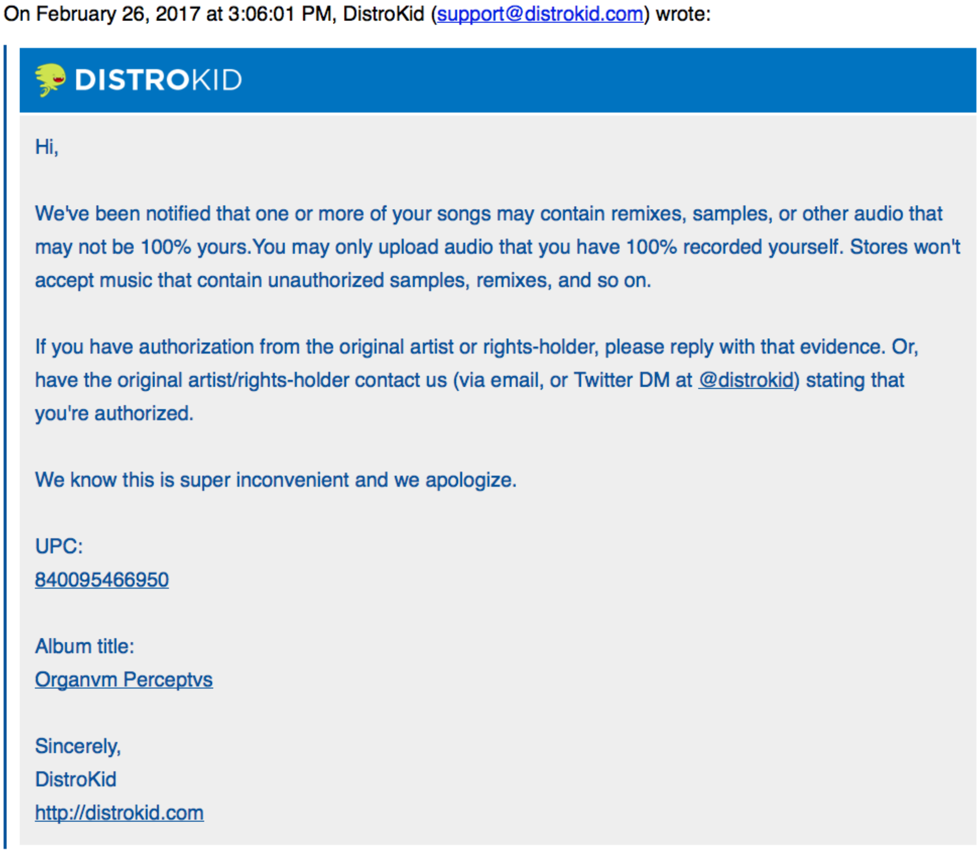
> We've been notified that one or more of your songs may contain remixes, samples, or other audio that may not be 100% yours. You may only upload audio that you have 100% recorded yourself. Stores won't accept music that contain unauthorized samples, remixes, and so on.[14]

The full email is shown in Figure 5.5. DistroKid's sample detection process is automated as Philip Kaplan, founder and CEO of DistroKid, explains:

> When an album is submitted to DistroKid, our automated system processes the music files and artwork, scans for copyright and compliance issues, and then immediately sends everything to the stores. The entire process takes less than a minute per song. It runs 24/7 and requires no human interaction.[15]

[14] DistroKid. Email message to Indexical. February 26, 2017.

[15] Philip Kaplan. 2019. "Open letter to Robb McDaniels, Founder and CEO of InGrooves." Accessed April 27. https://medium.com/@pud/open-letter-to-robb-mcdaniels-founder-and-ceo-of-ingrooves-11b2bc746c2f.



On February 26, 2017 at 3:06:01 PM, DistroKid (support@distrokid.com) wrote:

Hi,

We've been notified that one or more of your songs may contain remixes, samples, or other audio that may not be 100% yours.You may only upload audio that you have 100% recorded yourself. Stores won't accept music that contain unauthorized samples, remixes, and so on.

If you have authorization from the original artist or rights-holder, please reply with that evidence. Or, have the original artist/rights-holder contact us (via email, or Twitter DM at @distrokid) stating that you're authorized.

We know this is super inconvenient and we apologize.

UPC:
840095466950

Album title:
Organvm Perceptvs

Sincerely,
DistroKid
http://distrokid.com

Figure 5.5: DistroKid email message to Indexical.

5.5 Two Thoughts on Musical Borrowing

My reaction to DistroKid's email is two-tiered. First of all, DistroKid's sample detection process is automated. I can't help but find it ironic and a little bit humorous that HVB was taken down, for lack of a better phrase, by the same kinds of automated sound analysis tools and techniques that I use to make the music in the first place. What could be a better illustration of my concern that these tools be used carefully and cautiously otherwise suffer unintended consequences. In many cases those who bear the unintended consequences will be small, independent communities already marginalized to the fringes of mainstream cultural values and largely powerless to fight back. I felt as though I had happened upon a secret gatekeeper to the platforms of streaming music, an internet age version of the kind of dystopian bureaucracy Terry Gilliam depicts in the 1985 film *Brazil.* I found my interaction with DitsroKid to be a disarming mix of candor and charm that belied the gravity of the situation, specifically problems facing fair and equitable access in electronic and streaming media, and diminished the complex history of musical sampling.

Critics have spoken out against large streaming music services in recent years, criticizing a concentration of power that monopolizes the marketplace, funnels access through a centralized and streamlined format, and leaves artists with little control over how their music is distributed and accessed, not to mention the financial challenges of distributing royalties in an equitable way. Mat Dryhurst in particular is an outspoken and prolific advocate of the challenges faced by independent artists in the streaming marketplace.[16] Interestingly, concerns about the centralization of streaming music parallel concerns more generally about the centralization of the internet. What was once a decentralized open platform that anyone could access equally is now disproportionately concentrated in the hands of a few dominant social networks, proprietary developer platforms, and operating systems that exert influence on development and distribution. These companies control access, flow of information, and more importantly, mediate the very information that we encounter online.

Providing users fair and equitable tools with which to navigate the tremendous amount of media available on electronic platforms is an interesting new challenge faced by streaming music services. Automated discovery and recom-

[16] Mat Dryhurst. 2019. "SoundCrowd: Tokenizing and Collectivizing Soundcloud." Accessed April 27. https://medium.com/blockchannel/soundcrowd-tokenizing-collectivizing-soundcloud-5c4f60ed4961.

mendation systems are now widely deployed, with some of the largest streaming platforms, including Spotify, Apple Music, and Pandora, providing personalized music recommendation. Recommender systems, for the most part, operate on either a user's past activity on the platform — what a user plays, skips, and likes — or features of the media itself, which involves automated sound analysis. While there may be a sense that personalization is preferable or perhaps even synonymous with curation, truth be told, these discovery services are built to prioritize and de-prioritize content, and wield too much power over which music is heard. I cannot help but wonder what kinds of latent value judgements are implicit in these systems and as such are reified by their mass deployment.

Second, sampling should be understood as part of a long history of musical borrowing, a tradition that reaches back centuries in Western music practice. This history raises many complex questions pertaining to concerns both artistic and aesthetic as well as ethical and legal. In the history of notated Western music, musical quotation, in which melodic, harmonic, or rhythmic fragments of a composer's work are incorporated into another, is a precursor for electronic sampling. Medieval chants often borrowed melodic patterns from other chants. Paraphrase masses of the Renaissance were based on well-known plainsong chants or secular songs. Throughout the Romantic period, musical quotation was common and composers would incorporate and rework music of their peers. With political revolutions of the nineteenth century, it was also common for composers to quote folk and protest songs. These traditions continued into and throughout the twentieth century — composer Charles Ives in particular is well known for his extensive use of borrowed material.[17]

At the turn of the twentieth century, trends in the visual and literary arts

[17] Based on conversation with musicologist Madison Heying, April 24, 2019.

prefigured attitudes towards sampling, a more recent innovation afforded by twentieth century technologies in which a portion of a sound recording is incorporated into a another. Marcel Duchamp's concept of the readymade, Dada cut-up writing techniques, and even collage as pioneered in painting and visual art by George Braque and Pablo Picasso all set artistic and aesthetic precedent that I argue is more akin to musical sampling than to prior forms of musical borrowing such as quotation. Although some composers worked with phonograph and turntable techniques in the earlier part of the century, the genres of electroacoustic and tape music were more firmly established around mid century when tape recorders became accessible. Composers such as Pierre Schaeffer (1910–1995) and John Cage (1912–1992) established the early norms and techniques of working with recorded sound. Pierre Schaeffer assembled his *Cinq études de bruits* (1948) from samples of train sounds that Schaeffer either recorded himself or found in the libraries of various radio stations. John Cage's *Williams Mix* (1951–53) consists of hundreds of sound clips, recorded by Louis and Bebe Barron, that were edited and spliced together on magnetic tape.[18]

Sampling, as it is colloquially known today, is deeply connected to hip hop, an African American subculture, art movement, and music practice that developed in the 1970s in the United States, particularly in New York City. Sampling is among the defining features of hip hop music. DJs, such as DJ Kool Herc (b. 1955), Afrika Bambaataa (b. 1957), and Grandmaster Flash (b. 1958), pioneered techniques of electronic music using turntables to manipulate borrowed sound — to slow down, speed up, loop, skip, play backwards. These artists represent a contribution to electronic and computer music that is not always mentioned in the traditional

[18] For a history of musical sampling see Jon Leidecker's radio series Jon Leidecker. 2011-2018. "Variations." *Ràdio Web Macba.* Accessed April 27, 2019. `https://rwm.macba.cat/en/variations_tag`.

histories, but is perhaps more widespread culturally than most others.[19] In many ways composer John Oswald (b. 1953) was engaging with this popular culture movement when he responded to new ways people were repurposing recorded media. Oswald developed the idea of "plunderphonics," a work that is a clearly derivative manipulation of another recording. In the opening paragraph of his 1985 talk "Plunderphonics, or Audio Piracy as a Compositional Prerogative" Oswald called attention to hip hop and to the new creative potential of electronic sound manipulation:

> A phonograph in the hands of a hip hop/scratch artist who plays a record like an electronic washboard with a phonographic needle as a plectrum, produces sounds which are unique and not reproduced - the record player becomes a musical instrument. A sampler, in essence a recording, transforming instrument, is simultaneously a documenting device and a creative device, in effect reducing a distinction manifested by copyright.[20]

More or less concurrent with the emergence of hip hop music, the US government revised copyright law, passing the Copyright Act of 1976. Prior to this Act, US copyright law had not been revised since 1910. The Copyright Act of 1976 extended protection of the law to "original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device,"[21] including sound recordings. The Act also codified the terms of "fair use," under which the use of a copyrighted work is not copyright infringement, a category that many artists claim. It is important to note the law is complex; while this Act remains the foundations of US copyright law

[19] For a history of Hip Hop see Tricia Rose. 2008. *The Hip Hop Wars: What We Talk About When We Talk About Hip Hop - and Why It Matters.* New York: Civitas Book.

[20] John Oswald. 1985. "Plunderphonics, or Audio Piracy as a Compositional Prerogative." Accessed April 27, 2019. `http://www.plunderphonics.com/xhtml/xplunder.html`.

[21] 17 U.S.C. Sec. 102.

today, the law is also defined by legal precedent and court case rulings. Copyright law was not generally enforced with respect to music sampling until a number of landmark cases in the late 1980s and early 1990s, notably *Grand Upright Music, Ltd v. Warner Bros. Records Inc.* In this case, Gilbert O'Sullivan, a popular English songwriter challenged DJ and rapper Biz Markie, an extremely influential hip hop artist, for sampling without permission. The court ruled in O'Sullivan's favor.[22]

Sampling continues to be a popular means of artistic expression today. Digital and electronic appropriation is widespread culturally, in part due to the proliferation of digital content and the availability of extremely powerful consumer tools for manipulating digital media, both audio and visual.[23] As I see it, HVB is part of this history, embroiled in the same legal and ethical conundrums, and motivated by similar aesthetic and artistic drives. I also see in the project something new. It points towards new legal and ethical questions surrounding machine learning and big data.

Learning algorithms extrapolate patterns and trends across many thousands, if not millions, of data points. What does it mean for a work to be based on patterns extrapolated from many millions of works, not just a single work? Who holds the rights then — the authors each of the individual works, the algorithm designer, is there no human author at all? On the first day of my class *Music and Artificial Intelligence* this past winter term 2018, a student asked, how often does a machine learning algorithm scrape SoundCloud, an online platform where users upload and share their music with one another, and generate new works based on the works of an entire community? I didn't know, and still don't know, the

[22] Music Copyright Infringment Resource. 2019. "Grand Upright v. Warner." Accessed April 27. `https://blogs.law.gwu.edu/mcir/case/grand-upright-v-warner/`.

answer. It occurred to me, however, that this questions cuts to the core of public concern around digital and online media and the uncertainty and unease around data privacy. It's not clear who should benefit from big data and how, nor what role intellectual privacy and copyright should play. It's not clear what constitutes a copy anymore nor what should fall under the legal category "original works of authorship." This, I believe, will form the basis of intellectual property debates in the coming years.

Chapter 6

Conclusion

When I started the Happy Valley Band project over eight years ago, the terms Machine Learning and Artificial Intelligence did not carry the same cultural connotations they do today. Machine learning technology was nascent, deep learning had not yet boomed, and social issues such as algorithmic bias and discrimination had yet to be taken up by cultural theorists and the Digital Humanities. Within the past ten years the term AI has proliferated. AI has become deeply integrated into the daily life and arguably shapes our understanding of the world. Concurrent with the rise of AI is a growing awareness and public conversation about concerns of algorithmic bias and discrimination. I believe that artists can and should play a critical role in thinking through the social and ethical issues of artificial intelligence and automation. In this chapter, I relate my experience of HVB to these contemporary conversations.

6.1 Overview of Algorithmic Bias

Although algorithmic bias is difficult to define, there is a general sense that autonomous algorithms are in danger of reflecting, reinforcing, and amplifying cultural biases, stereotypes, and historical inequalities. Search algorithms, social networks, and other forms of automated computation can unfairly privilege or discriminate against arbitrary population groups, often along the lines of race, gender, ethnicity, and socioeconomic status. Since automated models are used in many aspects of daily life — policing, determining who is eligible for a loan, hiring practices, and education — it is concerning that such algorithms may encode discrimination and bias. In many cases, the groups against which algorithms are biased are already underrepresented or subject to discrimination, thus reinforcing and perpetuating inequality.

Many alarming examples of algorithmic bias have recently been reported. A 2016 analysis published by investigate journalism organization ProPublica claimed that COMPAS, an algorithm used by judges to help determine bail sentencing decisions, is biased against black defendants. COMPAS predicts the likelihood that a defendant will commit a violent crime, assigning a risk score between 1 and 10 based on a number of factors including age, criminal history, charge degree, and gender. The study found that black defendants were substantially more likely to be incorrectly classified as high risk, while white defendants were more likely to be incorrectly classified as low risk (Angwin et al. 2016). In a 2018 news article, Jeffrey Dastin reported that Amazon's hiring software was biased against women. The software was used to screen resumes of job applicants, scoring candidates on a scale from 1 to 5 stars, and was trained on a collection of prior resumes submitted to Amazon. The collection of prior resumes was largely dominated by male candidates, and the model learned to penalize female candidates, identifying gender not just by keywords such as "women's" but also by indicators that implicitly suggest gender, such as having attended an all-women's college (Dastin 2018). In a 2018 study by Buolamwini, three commercial machine vision systems were shown to be biased on the task of gender classification, exhibiting substantial disparity in the accuracy of classifying darker-skinned females versus lighter-skinned males. The misclassification rate for darker-skinned females being as high as 34.7%, where as the maximum error rate for lighter-skinned males is 0.8%. Since machine vision tools are often used in a pipeline for higher risk tasks such as law enforcement and public surveillance, bias could lead to wrongfully accusations based on confident misidentifications (Buolamwini and Gebru 2018).[1]

In many examples, algorithmic bias is introduced by the dataset on which the algorithm is trained. Buolamwini explains that machine vision models are often trained using fewer images of women and people with dark skin, which leads to poorer performance in such cases; Amazon's hiring mode was trained on a dataset dominated by male applications causing it to preference male candidates; and predictive policing models are often trained on "dirty data" reflecting past histories of flawed, unlawful, and racially biased police practices (Richardson, Schultz, and Crawford 2019). These biases are reflected by the training dataset, captured by the model, and perpetuated in a pernicious feedback loop that reinforces and amplifies historical and cultural biases.

A recent example to illustrate how algorithms learn to encode bias is word embeddings. A word embedding, such as the project word2vev (Mikolov et al. 2013), maps English language words to a high-dimensional vector space — 300 is a common number of dimensions — assigning to each word a 300 dimensional

[1] For a discussion of how mathematical models increase inequality see O'Neil 2016.

numerical value, which is interpreted as a location in that high-dimensional vector space. The mapping is determined by English language usage, as found by training on large bodies of text, and locations of words reflect their meanings. Words with similar meanings are generally close together in the vector space, and semantic relationships between words are encoded as geometric relationships in the vector space.

Word embeddings have been shown to encode biases, such as gender and race, that are implicit in the texts on which they are trained. The relationship between the words “he” and “she,” for example, represents a vector in the embedding space onto which other words can be projected to reveal implicit gender associations (shown in Figure 6.1). Gender neutral words such as “homemaker” and “sewing” are more strongly associated with “female,” where as “genius” and “tactical” are more strongly associate with “male.” In a 2017 report, Caliskan et al. applied Implicit Association Test of subconscious social biases to popular word embeddings, including word2vec, and found significant social biases (Caliskan, Bryson, and Narayanan 2017). The test revealed strong association between male names and “career” while females names were associated with “homemaking” and “family.” While it is not surprising that gender bias lurks in English language writing, it is alarming that machine learning systems capture and encode it. Word embedding is commonly used as a preprocessing feature for Natural Language Processing systems. What makes it powerful also makes it dangerous. Sophisticated semantic relationships are captured, but cultural bias are also integrated into the machine learning system, which can reinforce and perpetuate bias.

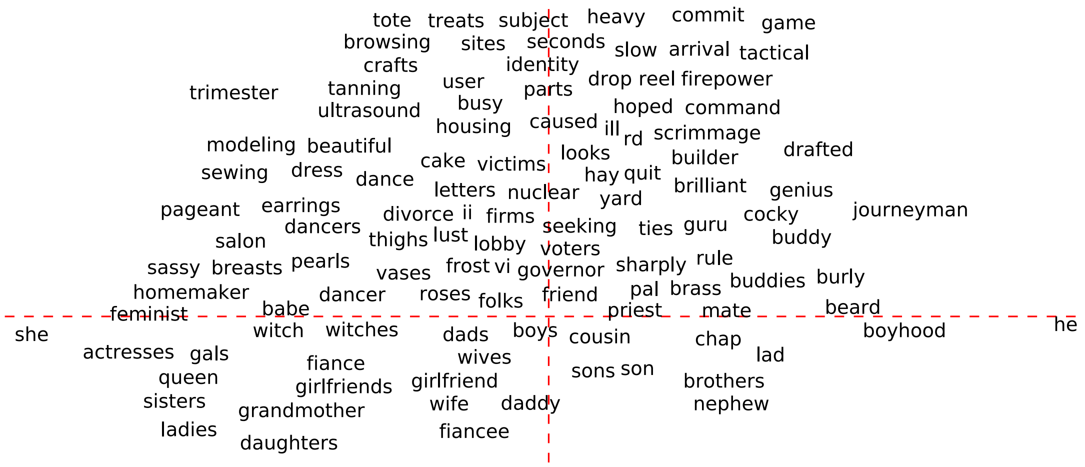


Figure 6.1: Word embedding. Reproduced from Bolukbasi et al. 2016.

In the past few years, many new conferences and institutes have formed with the goal of making automated algorithms more fair, such as the conference on Fairness Accountability and Transparency in Machine Learning (FATML) and the AI Now Institute at New York University. Even within industry, leading companies have formed boards for Artificial Intelligence and ethics — despite dubious motivations, their presence nevertheless is a symbol of the widespread problem. While it is clear that algorithmic bias is a pressing social issue, what to do about it is less clear. The academic literature is vast and scattered across multiple disciplines and applications. It is full of convincing examples and calls to action, but light on pragmatic solutions and generalized frameworks. Many solutions have been explored, such as mathematically removing bias using subspace methods and vector projection (Bolukbasi et al. 2016), developing diverse and robust testing benchmarks (Buolamwini and Gebru 2018), and simply requiring humans to intervene at some point in a decision pipeline — the EU's recent General Data Protection Regulation bans fully automated decision-making in significant situations. Unfortunately, solutions are often fit to specific cases and models are often difficult to implement in mass deployment (Springer, Garcia-Gathright, and Cramer 2018). Practical ways and workflows for identifying, preventing, and fixing algorithmic bias are sorely needed.

Fairness is difficult to identify and define let alone codify in computer code. Many different statistical interpretations of fairness exist, which leads to mutually exclusive metrics that cannot be satisfied simultaneously. The very definition of algorithmic bias itself is not clear. From a computational perspective, bias is necessary; bias is what distinguishes an algorithm that is successful from an algorithm that operates at chance. Algorithmic bias, however, generally refers to bias that is unwanted or unintended. These two forms of bias — the necessary and the unwanted — can be difficult to disentangle. Bolukbasi explains that completing removing any gender variation from word embeddings is not always a viable solution, because in many cases gender is important to meaning. While it is important to remove gender stereotypes, such as the association between the words "receptionist" and "female," we may want to maintain desired gender associations, such as between the words "queen" and "female." The very nature of bias, when to remove it, and how are open questions that need to be addressed in context, requiring a larger guiding framework of understanding and evaluation.

At the root of the crisis of algorithmic bias is public understanding of algorithms, mathematics, and computation. Algorithms are often seen as objective. In fact, in many instances of algorithmic bias, automation is intended to curb the possibility of human bias — a judge, for instance, who is tired after hearing many court cases or allows emotional factors to influence their bail sentencing. At middle and high school education, mathematics tends to be taught as an instrument of scientific objectivity. For many students, it is not until college level education that they learn to appreciate the subjectivity and art in mathematics and logic.

I believe there is a larger cultural association of number with objectivity which has led to a false sense of objectivity that is now being challenged by the crisis of algorithmic bias. In this sense, my intention with HVB is to illustrate the subjectivity of mathematical analysis and to point towards considering algorithms as socially and culturally determined constructs.

6.2 Music as a Heuristic Analysis of Algorithmic Bias

Given that HVB was directly motivated by the question what does it mean for values to be embedded in sound analysis algorithms, it only seems appropriate to try to answer it. My understanding is informed as much by analyzing and coding machine listening algorithms as by interacting with algorithms through composition, performance, and listening.

While much of the current discussion around algorithmic bias focusses on bias introduced by large datasets, it is clear to me, from my work with HVB, that individuals, communities, and cultures that produce technologies also introduce bias, independent of the data ultimately used. Algorithms are cultural products, the results of social, political, and economic forces, not just computational objects, and it is necessary to consider bias as emanating from and residing within all of the various facets of this assemblage. Research on algorithmic bias extends back more than twenty years, and earlier writings provide frameworks for considering types of bias not introduced by data. Friedman and Nissenbaum's framework, in particular, is well-suited for describing how I understand values to be embedded in machine listening algorithms (Friedman and Nissenbaum 1996). Friedman and Nissenbaum identify three categories of algorithmic bias: *preexisting* bias pertains to the values, practices, and attitudes of the individuals or institutions that produce technologies; *technical* bias refers from the limitations or constraints of technical design; and *emergent* bias arises when a technology is used in a context different than intended by its designers.

As addressed by Friedman and Nissenbaum's concept of preexisting bias, the musical ideas of the communities that produce machine listening technologies bias said technologies towards certain musical systems.[2] Given that there is no universal definition of pitch, for example, any algorithm for detecting pitch is value-laden. A harmonic model of pitch perception will find different results than a physiological model of pitch perception, especially when driven to extremes and edge cases, such as in HVB. Deciding to smooth or suppress short-lived and spurious changes in pitch, which is common in many detection algorithms, labels such artifacts as errors, indicating they are neither valid nor correct ways of hearing. The very idea to transcribe music in terms of pitch and rhythm rather than other features of sound, or to quantize to twelve-tone equal temperament, reflects a trans European musical value system that is better suited to describe certain forms of music than others. Such values may be encoded consciously or unconsciously as individuals and communities are influenced by larger cultural forces and scientific and engineering paradigms that lead to ways of describing the world.

As expressed by Friedman and Nissenbaum's concept of technical bias, the limits and constraints of the computational tools used introduce bias, both hardware and software. Developers of machine listening technologies must necessarily find ways to express musical ideas in mathematically computable forms, and the

[2] With the HVB project, I have focussed on the question of how values are embedded in algorithms. For an ethnographic study of the communities that produce machine listening technologies and their value systems, see the work of anthropologist Nick Seaver on music recommender systems (Seaver 2018; Seaver 2017).

available mathematics and computational paradigms not only impose limits and constraints but shape how musical ideas are codified. How problems are framed, the software tools and paradigms used, the underlying structure of binary computation and its hardware implementations, and available mathematical ideas both suggest and limit possibilities. One such pervasive construct in machine listening is Fourier analysis, which has had a profound impact on how we conceive of sound in the twentieth century. Fourier analysis represents sound in terms of the amplitude and phase of harmonically spaced sinusoids, distributing spectral information between the dimensions of amplitude and phase, and this structure is clearly reflected in a number of the onset detection functions discussed in Chapter 2. Spectral flux, phase deviation, and complex difference measure frame to frame differences in the amplitude, phase, and combined domains respectively. While it is tempting to interpret these spectral features as corresponding to independent acoustic features, in reality the relationship is more complex. Furthermore, Fourier analysis does not represent all sounds equally well. Sounds with harmonic structure better fit the model than inharmonic sounds, which suffer from chatter and leakage across analysis bins, and many artifacts in HVB music are themselves results of FFT analysis and its computational implementation.

How can algorithmic bias be effectively identified, measured, and ultimately ameliorated? These are among the most pressing questions facing the field machine learning today. In a recent article “Why Do We Want Our Computers to Improvise,” Georg Lewis extends his idea of improvisation to include interactions with technologies generally, not limited to music technologies (Lewis 2018). Lewis explains that interactions with technologies reveal aspects of the people and environments that produce them. I see HVB in this context, as a set of interactions with machine listening algorithms that reveal aspects of the values that motivate such technologies. Rather than quantified mathematical analysis, it is a heuristic process,[3] a way of coming to know the biases present through individual, hands-on experience with the products of the algorithms. I believe that through experience and though our sense of musicality we can uniquely come to understand the biases implicit in algorithms, although I am admittedly unsure how to translate this into an industry ready workflow.

Through my experience writing, performing, and listening to HVB music, I have come to understand the nature of the machine listening algorithms. I believe that others have as well, including the performers in the ensemble and listeners. To return briefly to the reviewer of *ORGANVM PERCEPTVS* discussed in Chapter 5 who described the music as "conceptually fascinating" but simply "cannot bear to listen to it," upon reading the review my partner Madison Heying responded, "They didn't listen to the record enough." I include this quote not to be flippant and dismissive but in all earnest; composing this music, playing this music, and listening to this music has changed the way I hear. This project has put me in intimate contact with the idiosyncrasies of machine listening algorithms in way that I never planned for. I have spent so much time fitting my brain into these algorithms — listening to their results, tweaking parameters, anticipating new results, then listening back and comparing — that is difficult for me to hear these songs any other way, or really any other song any other way. Ultimately, I have come to sense the logic in the transcription system.

Of all the responses to HVB that I have received, one in particular gives me hope that my experiences will translates to others. On the internet forum `ilxor.com` Milton Parker wrote:

[3] My use of the term heuristic is informed by David Dunn's recent collection of pieces titled *Heuristic Automata (Book 1)* (2012–2016).

> at first it sounds as if it's filled with timing errors and wrong notes, but all the "wrongness" is actually over-precise analysis of some of the original arrangements, and as you listen over time and the pieces somehow hold together and get tighter and tighter, you realize that they kind of aren't strictly errors, there's some kind of turbo-charged high level thinking going on.

I like to think of Parker's "turbo-charged high level thinking" as a utopian future in which technologies extend our value systems to more inclusive ways of hearing rather than reifying the value systems of the communities that produce technologies.

Appendices

Appendix A

Related Publications

Writings that I have authored about Happy Valley Band:

Bick, Emily. 2017. "Album stream and interview: David Kant of Happy Valley Band talks about their 'machine listening' album." *The Wire*. Accessed April 27, 2019. `https://www.thewire.co.uk/in-writing/interviews/listen-to-the-happy-valley-band-s-new-album-and-read-an-interview-with-its-founder`.

Kant, David. 2016a. "The Happy Valley Band: Creative (Mis)Transcription." *Leonardo Music Journal (LMJ)* 26:76–78.

———. 2016b. "The Long Answer." *Experimental Music Yearbook*. Accessed May 4, 2019. `http://www.experimentalmusicyearbook.com/Happy-Valley-Band`.

———. 2017a. "Making music through machine ears." *Humanising Algorithmic Listening*. Accessed May 4, 2019. `http://www.algorithmiclistening.org/introductions/HVB/`.

———. 2017b. *ORGANVM PERCEPTVS*. Happy Valley Band. Indexical Index-2, Vinyl LP + Print Booklet + Digital Download.

Appendix B

Selected Press

Selected press and reviews about Happy Valley Band:

Bick, Emily. 2017. “Album stream and interview: David Kant of Happy Valley Band talks about their ‘machine listening’ album.” *The Wire*. Accessed April 27, 2019. https://www.thewire.co.uk/in-writing/interviews/listen-to-the-happy-valley-band-s-new-album-and-read-an-interview-with-its-founder.

Chandler, Simon. 2018. “Meet the Artists Using Coding, AI, and Machine Language to Make Music.” *Bandcamp Daily*. Accessed April 27, 2019. https://daily.bandcamp.com/2018/01/25/music-ai-coding-algorithms/.

Fitzgerald, Colin. 2017. “Happy Valley Band deconstruct pop classics via machine-learning algorithm on debut album ORGANVM PERCEPTVS.” *Tiny Mix Tapes*. Accessed April 27, 2019. https://www.tinymixtapes.com/news/happy-valley-band-deconstruct-pop-classics-machine-learning-algorithm-debut-organvm-perceptvs.

Hambleton, Elizabeth. 2017. "Happy Valley Band's debut album ORGANVM PERCEPTVS." *New Classic LA*. Accessed April 27, 2019. http://newclassic.la/2017/04/27/review-happy-valley-bands-debut-album-organvm-perceptvs/.

Margasak, Peter. 2017. "Best of Bandcamp Contemporary Classical: March 2017." *Bandcamp Daily*. Accessed May 4, 2019. https://daily.bandcamp.com/2017/03/28/best-of-bandcamp-contemporary-classical-march-2017/.

Zaldua, Chris. 2017. "You, With the Violin! Sight-Read These Computer Algorithms!" *KQED*. Accessed April 27, 2019. https://www.kqed.org/arts/13038360/you-with-the-violin-sight-read-these-computer-algorithms.

Appendix C

Complete List of Songs

List of all Happy Valley Band songs that I have transcribed:

Cry One More Time (2019), Gram Parsons

Stony End (2018), Laura Nyro

Katie Cruel (2017), Karen Dalton

War Pigs (2017), Black Sabbath

When the Levee Break (2017), Led Zepplin

Darling Nikki (2016), Prince

Like a Prayer (2015), Madonna

Jungle Boogie (2015), Kool and the Gang

Born to Run (2015), Bruce Springsteen

(You Make Me Feel Like) A Natural Woman (2015), Aretha Franklin

Ring of Fire (2013), Johnny Cash

After the Gold Rush (2013), Neil Young

Suspicious Minds (2012), Elvis Presley

It's a Man's Man's Man's World (2012), James Brown

In the Air Tonight (2012), Phil Collins

This Guy's in Love with You (2012), Herb Alperb

Wooden Heart (2012), Elvis Presley

Midnight Mover (2012), Wilson Pickett

Crazy (2011), Patsy Cline

Good Luck Charm (2011), Elvis Presley

Appendix D

ORGANVM PERCEPTVS Personnel

List of musicians who performed on *ORGANVM PERCEPTVS*:

Alexander Dupuis, Guitars

Mustafa Walker, Electric and Upright Bass

Beau Sievers, Drum Kit

Andrew Smith, Piano and Keyboards

David Kant, Saxophones

Pauline Kim Harris, Violin

Conrad Harris, Violin

Chris Nappi, Percussion

Joseph Kubera, Piano

Daniel Costello, French Horn

Thomas Verchot, Trumpet and Flugelhorn

Joe Moffett, Trumpet

Christoper Scanlon, Trumpet

Nathaniel Morgan, Alto Saxophone

Sam Friedman, Harmonica

Charlotte Mundy, Voice

Jane Sheldon, Voice

Eve Gigliotti, Voice

John Welsh, Guitar

Larry Polansky, Guitar

Appendix E

Performance History

2019	High Desert Soundings, Joshua Tree, CA
2018	CODAME Art + Tech Festival, San Francisco, CA
2018	Pro Arts Gallery, Oakland, CA
2018	Idea Fab Labs, Santa Cruz, CA
2017	Experimental Music Yearbook, Human Resources, Los Angeles, CA
2017	Center for New Music, San Francisco, CA
2016	CCRMA, Stanford University, Palo Alto, CA
2016	Don Quixote's Musical Hall, Felton, CA
2015	Dog Star Orchestra 11, the wulf., Los Angeles, CA
2015	sfSound series, Center for New Music, San Francisco, CA
2013	Littlefield, Brooklyn, NY
2012	Electric Eclectics 7, Meaford, Ontario, Canada
2012	Miniature Minotaurs, WFMU Radio, Jersey City, NJ
2012	Transient Series I.2, Brooklyn, NY
2011	Ostrava Days Festival 2011, Ostrava, Czech Republic

Bibliography

Aarabi, Parham, and Guangji Shi. 2004. “Phase-Based Dual-Microphone Robust Speech Enhancement.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34 (4): 1763–1773.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine bias: Tere’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica.* Accessed May 1, 2019. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Bača, Trevor, and Josiah Wolf Oberholtzer. 2018. “Abjad API.” Accessed February 16, 2019. `http://abjad.mbrsi.org/api/index.html`.

Barlow, Clarence, and Henning Lohner. 1987. “Two Essays on Theory.” *Computer Music Journal* 11 (1): 44–60.

Bello, Juan Pablo, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. 2005. “A Tutorial on Onset Detection in Music Signals.” In *IEEE Transactions on Speech and Audio Processing,* 13:1035–1047. 5.

Bello, Juan Pablo, Chris Duxbury, Mike Davies, and Mark Sandler. 2004. “On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain.” *IEEE Signal Processing Letters* 11 (6): 553–556.

Bick, Emily. 2017. “Album stream and interview: David Kant of Happy Valley Band talks about their ‘machine listening’ album.” *The Wire*. Accessed April 27, 2019. `https://www.thewire.co.uk/in-writing/interviews/listen-to-the-happy-valley-band-s-new-album-and-read-an-interview-with-its-founder`.

Bodden, Markus. 1993. “Modeling Human Sound-Source Localization and the Cocktail Party Effect.” In *Acta Acustica,* 1:43–55.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.” In *Proceedings of Neural Information Processing Systems (NIPS),* 4349–4357.

Bregman, Albert S. 1994. *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge: The MIT Press.

Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Conference on Fairness, Accountability, and Transparency,* 81:1–15.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Science* 356:183–186.

Casey, Michael, Parag Mital, Andy Sarroff, Tom Stoll, Jessica Thompson, Spencer Topel, Ben Fields, and Christophe Rhodes. 2016. “Bregman Toolkit: Audio and Music Analysis and Synthesis in Python.” Accessed February 17, 2019. `https://github.com/bregmanstudio/BregmanToolkit`.

Chandler, Simon. 2018. “Meet the Artists Using Coding, AI, and Machine Language to Make Music.” *Bandcamp Daily.* Accessed April 27, 2019. https://daily.bandcamp.com/2018/01/25/music-ai-coding-algorithms/.

Cheveigné, Alain de, 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.

Collins, Nick. 2005a. “A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions.” In *118th Convention of the Audio Engineering Society,* 6363–6375.

———. 2005b. “Using a Pitch Detector for Onset Detection.” In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR),* 100–106.

Dastin, Jeffrey. 2018. “Amazon scraps secret AI recruiting tool that showed bias against women.” *Reuters.* Accessed May 14, 2019. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Didkovsky, Nick. 2004. “Java Music Specification Language, v103 update.” In *Proceedings of the International Computer Music Conference.*

Didkovsky, Nick, and Phil Burk. 2019. “JSML Docs.” Accessed February 16, 2019. http://www.algomusic.com/jmsl/index.html.

Dixon, Simon. 2006. “Onset Detection Revisited.” In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06),* 133–137.

Doval, Boris, and Xavier Rodet. 1991. "Estimation of fundamental frequency of musical sound signals." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 3657–3660.

Dryhurst, Mat. 2019. "SoundCrowd: Tokenizing and Collectivizing Soundcloud." Accessed April 27. `https://medium.com/blockchannel/soundcrowd-tokenizing-collectivizing-soundcloud-5c4f60ed4961`.

Fitzgerald, Colin. 2017. "Happy Valley Band deconstruct pop classics via machine-learning algorithm on debut album ORGANVM PERCEPTVS." *Tiny Mix Tapes.* Accessed April 27, 2019. `https://www.tinymixtapes.com/news/happy-valley-band-deconstruct-pop-classics-machine-learning-algorithm-debut-organvm-perceptvs`.

Friedman, Batya, and Helen Nissenbaum. 1996. "Bias in Computer Systems." *ACM Transactions on Information Systems* 14 (3): 330–347.

Hambleton, Elizabeth. 2017. "Happy Valley Band's debut album ORGANVM PERCEPTVS." *New Classic LA*. Accessed April 27, 2019. `http://newclassic.la/2017/04/27/review-happy-valley-bands-debut-album-organvm-perceptvs/`.

Kant, David. 2016a. "The Happy Valley Band: Creative (Mis)Transcription." *Leonardo Music Journal (LMJ)* 26:76–78.

———. 2016b. "The Long Answer." *Experimental Music Yearbook.* Accessed May 4, 2019. `http://www.experimentalmusicyearbook.com/Happy-Valley-Band`.

———. 2017a. "Making music through machine ears." *Humanising Algorithmic Listening.* Accessed May 4, 2019. `http://www.algorithmiclistening.org/introductions/HVB/`.

Kant, David. 2017b. *ORGANVM PERCEPTVS.* Happy Valley Band. Indexical Index-2, Vinyl LP + Print Booklet + Digital Download.

Kaplan, Philip. 2019. "Open letter to Robb McDaniels, Founder and CEO of InGrooves." Accessed April 27. `https://medium.com/@pud/open-letter-to-robb-mcdaniels-founder-and-ceo-of-ingrooves-11b2bc746c2f`.

Klapuri, Anssi P. 1999. "Sound Onset Detection by Applying Psychoacoustic Knowledge." In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 6:3089–3092.

———. 2005. "A Perceptually Motivated Multiple-F0 Estimation Method." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* 291–294.

———. 2006. "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes." In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR),* 216–221.

Knapp, Charles, and G. Clifford Carter. 1976. "The Generalized Correlation Method for Estimation of Time Delay." In *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP),* 24:320–327. 4.

Leidecker, Jon. 2011-2018. "Variations." *Ràdio Web Macba.* Accessed April 27, 2019. `https://rwm.macba.cat/en/variations_tag`.

Lewis, George. 2018. "The Oxford handbook of algorithmic music The Oxford Handbook of Algorithmic Music." Chap. Why Do We Want Our Computers to Improvise?, edited by Alex McLean and Roger T. Dean, 123–130. Oxford University Press.

Lyon, Richard F. 1983. "A Computational Model of Binaural Localization and Separation." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 8:1148–1151.

Margasak, Peter. 2017. "Best of Bandcamp Contemporary Classical: March 2017." *Bandcamp Daily.* Accessed May 4, 2019. `https://daily.bandcamp.com/2017/03/28/best-of-bandcamp-contemporary-classical-march-2017/`.

Mauch, Matthias, and Simon Dixon. 2014. "PYIN: A fundamental frequency estimator using probabilistic threshold distributions." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 659–663.

McFee, Brian, Colin Raffel, Dawn Liang, Daniel PW 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 (SciPy 2015),* edited by Kathryn Huff and James Bergstra, 18–25.

Mikolov, Tomas, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *Proceedings of the International Conference on Learning Representations (ICLR).*

Mori, Masahiro. 2012. "The Uncanny Valley." Translated by Karl F. MacDorman and Norri Kageki. *IEEE Robotics and Automation Magazine* 19 (2): 98–100.

Myers, Marc. 2012. "Caught in a Trap: Elvis's last No. 1 Hit." *Wall Street Journal (Online).*

Nauert, Paul. 1994. "A Theory of Complexity to Constrain the Approximation of Arbitrary Sequences of Timepoints." *Perspectives of New Music* 32 (2): 226–263.

Noll, A. Michael. 1970. "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate." In *Proceedings of the Symposium on Computer Processing in Communications,* XIX:779–797. Brooklyn, New York: Polytechnic Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Broadway Books.

Oswald, John. 1985. "Plunderphonics, or Audio Piracy as a Compositional Prerogative." Accessed April 27, 2019. `http://www.plunderphonics.com/xhtml/xplunder.html`.

Polansky, Larry. 2013. "Pitch, Harmony and Experimental Intonation: A Primer." Class Notes for seminars taught at Dartmouth College (1997) and UC Santa Cruz (2012).

Puckette, Miller S., Theodore Apel, and David Zicarelli. 1998. "Real-time audio analysis tools for Pd and MSP." In *Proceedings of the International Computer Music Conference,* 109–112.

Rascon, Caleb, and Ivan Meza. 2017. "Localization of Sound Sources in Robotics: A review." *Robotics and Autonomous Systems* 96:184–210.

Resource, Music Copyright Infringment. 2019. "Grand Upright v. Warner." Accessed April 27. `https://blogs.law.gwu.edu/mcir/case/grand-upright-v-warner/`.

Richardson, Rashida, Jason Schultz, and Kate Crawford. 2019. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice." *New York University Law Review* 94:192–233.

Rose, Tricia. 2008. *The Hip Hop Wars: What We Talk About When We Talk About Hip Hop - and Why It Matters.* New York: Civitas Book.

Seaver, Nick. 2017. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems." *Big Data & Society* 4 (2).

———. 2018. "Captivating algorithms: Recommender systems as traps." *Journal of Material Culture.*

Smaragdis, Paris. 2007. "Probabilistic Decompositions of Spectra for Sound Separation." Chap. 13 in *Blind Speech Separation,* edited by Shoji Makino, Te-Won Lee, and Hiroshi Sawada, 365–386. The Netherlands: Springer.

Springer, Aaron, Jean Garcia-Gathright, and Henriette Cramer. 2018. "Assessing and Addressing Algorithmic Bias — But Before We Get There." In *Association for the Advancement of Artificial Intelligence Spring Symposium Series.*

Tenney, James. 1986. *Meta+Hodos: A Phenomenology of 20th-Century Musical Materials and an Approach to the Study of Form.* Edited by Larry Polansky. Frog Peak Music.

———. 2015. "John Cage and the Theory of Harmony." Chap. 12 in *From Scratch: Writings in Music Thoery,* edited by Larry Polansky, Laure Pratt, Robert Wannamaker, and Michael Winter, 280–304. Original work published 1983. University of Illinois Press.

Vos, Joos, and Rudolf Rasch. 1981. "The perceptual onset of musical tones." *Perception & Psychophysics* 29 (4): 323–335.

Wang, DeLiang, and Guy J. Brown. 2006. "Fundamentals of Computational Auditory Scene Analysis." Chap. 1 in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications,* edited by DeLiang Wang and Guy J. Brown, 1–44. Hoboken: John Wiley & Sons, Inc.

Wang, DLiang. 2005. "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis." Chap. 12 in *Speech Separation by Humans and Machines,* edited by Pierre Divenyi, 181–197. New York: Springer US.

Wang, Eric. May. "What does it really mean for an algorithm to be biased?" *The Gradient.* Accessed May 1, 2019. https://thegradient.pub/ai-bias/.

Wook Kim, Jong, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. "CREPE: A Convolutional Representation for Pitch Estimation." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 161–165.

Yu, Xianchuan, Dan Hu, and Jindong Xu. 2014. *Blind Source Separation: Theory and Applications.* Beijing: Wiley.

Zaldua, Chris. 2017. "You, With the Violin! Sight-Read These Computer Algorithms!" *KQED.* Accessed April 27, 2019. https://www.kqed.org/arts/13038360/you-with-the-violin-sight-read-these-computer-algorithms.