Technology Design Recommendations Informed by Observations of Videos of Popular Musicians Teaching and Learning Songs By Ear

by

Christopher Liscio

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The writing in this thesis is entirely my own, developed through research meetings and collaborations with my supervisor, Prof. Dan Brown.

The work in Chapters 1-3, 6, and 7 is unpublished, and my own.

The work in Chapter 4 is an adaptation of a manuscript in submission, "Watching Popular Musicians Learn by Ear: A Hypothesis-Generating Study of Human-Recording Interactions in YouTube Videos", co-written with Dan Brown. A preprint of this manuscript is available on arXiv: https://arxiv.org/abs/2406.04058. The text in this thesis has been improved by feedback from anonymous peer reviewers of a previous submission.

The work in Chapter 5 is an adaptation of a manuscript in preparation for submission, which will have me and Dan Brown as authors.

Abstract

Instrumentalists who play popular music often learn songs by ear, using recordings in lieu of sheet music or tablature. This practice was made possible by technology that allows musicians to control playback events. Until now, researchers have not studied the human-recording interactions of musicians attempting to learn pop songs by ear. Through a pair of studies analyzing the content of online videos from YouTube, we generate hypotheses and seek a better understanding of by-ear learning from a recording. Combined with results from neuroscience studies of tonal working memory and aural imagery, our findings reveal a model of by-ear learning that highlights note-finding as a core activity. Using what we learned, we discuss opportunities for designers to create a set of novel human-recording interactions, and to provide assistive technology for those who lack the baseline skills to engage in the foundational note-finding activity.

Acknowledgements

First I would like to thank my wife for her love, support, and encouragement to pursue graduate studies. I would also like to thank my two teenage kids, who—along with my wife—helped keep me connected to the real world during this time. To my parents, sister, extended family, and in-laws, I am also grateful for the support and love you've shown me over the years in spite of my *many* peculiarities.

I would also like to thank my friends and colleagues in the Mac and iOS developer communities, some who I have known for more than 20 years. Giving and receiving support—in technical, business, customer, and (at times) emotional matters—has made this solitary career as an *indie* developer feel much less lonely. Additionally, I must express my gratitude to all the customers who have been patient with me, and continued to support me financially during my studies.

I am also grateful for all the researchers who responded to my emails and questions about their papers over the last 14 years, happily clarifying their work for a non-academic like myself: Adam Stark, Taemin Cho, Filip Korzeniowski, and *especially* Brian McFee. Brian not only helped me understand his work, but he also played a role in kicking off my graduate studies—introducing me to Dan Brown and helping with my application to the University of Waterloo.

I certainly do not wish to *thank* the global COVID-19 pandemic—the devastation and loss experienced by me and those around me will be felt for generations. However, I would not have attended the ISMIR 2021 conference if it were not held virtually, and therefore not been encouraged by Brian McFee to meet with Dan Brown.

Finally, I'd like to thank Dan Brown for supervising my studies at the University of Waterloo. To say that I have veered off the well-trodden path is an understatement, but *somehow* you still helped me get there in the end.

Dedication

For Shelley

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DSP digital signal processing 3, 17, 22

FFT Fast Fourier Transform 19

INM implicit note memory 10, 11, 14

INMI involuntary musical imagery 14, 16, 17, 102

MIR music information retrieval 3, 17, 20, 22, 80, 100

Foreword

I've been writing and selling software for musicians that learn by ear[1] since 2009, and have more than 20 years of industry experience—17 of which I've spent working exclusively for myself. I also learn music primarily by ear without the aid of tablature or sheet music, and have done so since I was a child. However, after doing this for more than 30 years I had not quite understood the significance of learning music in this way. But I am a software developer; not much of a musician.

When I started my Master's studies, I appeared—*guns blazing*—ready to embark on building a novel feature for my software. I had a body of literature to follow, and proofs of concept to build. It was not long before I realized the error in my ways. Traditionally, I have built products to satisfy my own needs, and such a feature would absolutely satisfy one of my many wishes as a user of the software. However, I could not explain *how* my idea was going to satisfy what musicians actually need, or want.

While this sounds like a marketing problem, it is far more philosophical. If I design products based on my own needs, and I am a musician—to *some* degree—then why do I find it difficult to explain the reasoning behind my own needs? It turned out that I didn't quite understand what it means to learn music by ear—the prevalence of the practice, the mechanics behind it, etc.

Trying to find answers within the literature, it became clear that there was a big hole—finding information about how popular musicians interact with technology while learning music by ear. My thesis is an attempt to start filling this void with two foundational studies that lead to a collection of design recommendations.

[1] https://capoapp.com

Chapter 1

Introduction

Instrumentalists that play popular music often learn to play melodies, harmonies, complex solos, and entire songs *by ear*. These musicians use recordings in place of sheet music, and they interact with those recordings as they work towards playing the music they hear on their instruments (Bennett 1980). With the growing popularity of music streaming (IFPI 2022), these interactions now typically occur on smartphones and computers using software from music streaming services. Aside from a near-unlimited library of music, these apps offer little beyond the pre-existing set of *human-recording interactions* provided by record, cassette tape, and compact disc (CD) players.

Purpose-built hardware and software can offer an *additional* set of interactions with recordings. These interactions include placing marks at key moments in the recording, repetitive playback between set locations (looping), and controlling the rate and pitch of playback independently. Such features might assist musicians trying to identify and reproduce notes on their instruments. However, we have not discovered any research that confirms this is the case. Rather, it is the underlying technology that is well-documented in the literature (e.g., De Götzen et al. 2000; Duxbury et al. 2002; Röbel 2003a), and it appears that features are built around this technology with little regard to the actual nature of the task.

My aim is to take a step towards improving upon this state of affairs. I want to better understand how musicians learn by ear, observe how they are doing it today, and try to develop novel human-recording interactions based on real-world findings. A great deal of design insight can be derived from observations of people interacting with technology. I argue that this can be achieved—at low cost—through analysis of YouTube videos. Here I present the results of two such studies, and the implications of our findings on the designs of purpose-built features.

In the first study, we analyzed a set of 18 videos depicting real-world examples of skilled popular musicians learning from recordings. In these videos, musicians can be seen interacting with music playback, seeking notes and/or chords on their instrument, and—in some cases—struggling while doing so. Based on our observations, we uncovered a set of hypotheses about the ear-learning task and how it connects to technology, grounded in real-world data. For example, transcribing a recording—to produce some form of notation—does not appear to contribute to the by-ear learning activity. Also, different musicians use different strategies to retain notes in memory. We also see that one's familiarity with a song seems to make by-ear learning go more smoothly.

We then analyzed and coded 29 lesson videos (all but one from YouTube) in the second study to characterize the process of by-ear learning, and identify differences in how it is taught. Using what we learned from the first study, and a survey of neuroscience and psychology literature about short- and long-term memory for music, we also identified where memory is called upon while a musician learns a song by ear. The results from this study helped explain differences we saw in the approaches of experienced musicians—not only are lessons taught differently, but certain techniques for identifying and retaining notes seem to rely less on tonal working memory than others.

Using what we learned, we present a conceptual model of by-ear learning that illustrates how note-finding lies at the foundation of each of its sub-tasks. That is, regardless of whether the musician is seeking the song's key, chords, melody, or solo, they begin by copying one or more individual notes from the recording. This foundational element requires that the player can accurately imitate pitches vocally, or that they can retain a note in pitch working memory for as long as it takes to find the note on their instrument. However, it can be modified such that neither skill is necessary for a musician to copy notes.

Finally, we present a set of recommendations for the designers of purpose-built technology intended for musicians that learn by ear from recordings. Among these recommendations are tools that help a musician develop familiarity with a song, create memories of musical sequences, and playback controls that respect the capacity and fragility of pitch working memory. Each of these recommendations are based upon a collection of real-world observations that came from user-uploaded videos found on YouTube, and all of them can be implemented today—some by using existing techniques from music information retrieval (MIR) and digital signal processing (DSP).

1.1 Overview

Given the breadth of my research, this thesis is organized differently from most others. Instead of a single, all-encompassing Background or Related Work chapter, I have two that serve its purpose.

Learning By Ear (Chapter 2) provides a background of by-ear learning, the kinds of memory that appear related to the task, and the technology that allows it to happen. Opportunities and Limits of Online Video Studies (Chapter 3) shares examples of previous studies that used YouTube as a data source, and explores ethical, methodological, and technical considerations for running a study on YouTube. Following this, A Hypothesis-Generating Study of Musicians Learning by Ear (Chapter 4) describes our YouTube-based study of experienced musicians learning by ear, and the findings we obtained from our observations. Towards an Understanding of the By-Ear Learning Task (Chapter 5) explores lesson videos found on YouTube to identify sub-tasks of by-ear learning and how they relate to one's memory. Finally, Designing Technology Supports for By-Ear Learning (Chapter 6) presents a model of by-ear learning sub-tasks, and a set of design recommendations for purpose-built technology products.

Chapter 2

Learning By Ear

It is common for popular musicians to build their repertoire and improve their skills by mimicking what they hear in recordings. Those who instead rely upon tablature or sheet music to learn popular songs are merely enjoying the hard work of a transcriber who themselves used the recording as source material (Bennett 1980). Still, while "the answers" might be findable online and in printed materials, playing exactly what is written on the page is unlikely to produce a satisfying result, as notation is limited in its ability to communicate subtle details in rhythm and performance that often contribute to what makes a recording memorable.

By-ear learning has been a backbone of western popular music for many decades, and it still plays an important role in the development of pop musicians.

2.1 Popular Musicians and Informal Learning

While one could argue that all genres of music are—at some point in time, or to some group of people—popular, in this thesis we are concerned with musicians that play an instrument in one of the myriad genres of music that (roughly) derive from the blues originating in the southern United States—country, rock, pop, jazz, hip hop, R&B, funk, and so on. H. Stith Bennett, who embedded himself with local musicians in the 1970s to study their progression from amateurs to professionals, defined his use of *rock* to be similarly broad:

> In this respect, my use of the term *rock* is intentionally imprecise. It certainly includes any usage of the term as a combining form: *country-rock, folk-rock,*

> *jazz-rock, progressive-rock,* and even *punk-rock*; and it overlaps a variety of other usages that might typify music, such as *middle-of-the-road, easy listening, old favorites, standards, party music,* and *bar music*. A precise definition is actually unattainable. (Bennett 1980, xvii)

Given the challenging nature of such a definition, and ongoing evolution of music, we propose a more relaxed one: popular music is that which an aspiring musician is unlikely to find among the printed materials offered by their school or private music teachers, yet is popular among their contemporaries. In rare cases where the material may appear in the curriculum, it is unlikely to take its true form, as Campbell (1995) notes. "Rock music that 'makes it' into a school program is thus often antiseptic, a pale imitation of its true colours." Musicians who feel motivated to learn such music have little choice but to become self-sufficient when their needs are unmet at school. We can imagine their exasperation: "Fine! If I can't learn what I want at school or in music lessons, I'll figure it out on my own."

And that's precisely what they do: the by-ear learning of popular music is most often a solitary activity (Bennett 1980; Campbell 1995; Green 2017). Bennett recognized that a musician's initial attempts to learn from a recording happen in private, freeing them from the worry that others may deem their skills inadequate (Bennett 1980). While he characterized this as an incomplete exercise—a precursor to the song getting "worked up" as a group, Groce (1989) and Campbell (1995) later found in interviews that musicians instead learned their parts in full before practice. That way, these musicians could show up prepared to play alongside their bandmates, and ensure practice went smoothly (Groce 1989). Despite these differences in how much players were expected to know before turning up to their group practice sessions, these authors all make the same claim that by-ear learning happens alone, behind closed doors.

Those musicians who spoke with Groce and Bennett engaged in *song getting* to perform covers with their local bands, but Bennett (1980) found that this exercise also helped them develop the musical vocabulary that eventually led to the creation of their own original songs. Lucy Green's interviews with popular musicians during the 1990s built upon this research by focusing more closely on the learning process itself, and discovered that—regardless of their exposure to a formal music education—popular musicians rely upon informal, by-ear methods to develop the skills for the genre they go on to (re)produce professionally (Green 2017).

2.2 Learning By Ear, from Recordings

Lars Lilliestam found that—at least in 1996—there was little research on the practice of playing by ear, partly due to the dominance of Western art music, in written form, and the associated pedagogy in the music literature (Lilliestam 1996). At present, little has changed in this regard—especially when considering the specific needs and practices of popular musicians.

One of Bennett's key insights is that popular musicians use recordings as the formal notation system from which they develop a repertoire, and ultimately their own playing technique (Bennett 1980, 1983). Even when popular musicians turn to a teacher, or refer to notation while they are learning, the teacher will have learned the song by ear, and the notation was transcribed directly from the recording, as Bennett notes.

> [...] rock sheet music is itself derived from recordings in most cases, and although it is transcribed by experts into the conventions of traditional musical notation, the process differs little from the direct song-getting process which I have described. The generally poor repute in which rock sheet music is held among rock musicians is inherent in the limitations of the traditional notation system: Rock musicians tend to play in ways for which conventional notation does not exist. This phenomenon has promoted and will continue to promote experimentation with written notation systems which can more adequately convey unconventional sounds, just as the art music world is now filled with experimental notation systems. In either case the primacy of sound over literature is surfacing. (Bennett 1980, 142-143)

While on-paper notation systems—on a traditional staff, or in tablature form—can serve as a memory aid or a written form of communication with others, the recording stands as the source of truth for popular instrumentalists. The musicians interviewed by Bennett (1980) would learn material almost exclusively from recordings heard on the radio, or during private sessions spent with vinyl records or cassette tapes. Similarly, the musicians interviewed by Groce (1989) would be given cassette tapes from which to learn their individual parts before band practice, and the young rock musicians interviewed by Campbell (1995) interacted with recordings played from either cassettes or CDs.

2.3 Studies of By-Ear Learning Methods

Previous studies of ear learning focused on participants with little to no professional musical experience. Researchers recruited high-school or university students enrolled in music lessons or non-musical novices. Lahav et al. (2005) recruited musically naïve college students in their study of auditory-motor interactions as the students learned melodies by ear on a piano. Varvarigou & Green (2015) categorized the ear learning styles and strategies from in-lesson observations of 75 music students, using the students' initial contacts with isolated recordings of bass guitar melodies to characterize their learning styles, and subsequent interactions to identify their strategies. McPherson et al. (1997) conducted interviews with high school wind instrumentalists after administering ear learning proficiency tests to ask them about their approach to the task. Oswald (2022) studied the methods used by high schoolers to learn melodies by ear using custom-built software that was instrumented to measure the frequency of their interactions with the music. Few studies have focused exclusively on the techniques employed by experienced players. Woody & Lehmann (2010) recruited 24 college musicians with both formal and informal experience to learn melodies by ear, and reported their strategies based on post-activity interviews with these musicians. Johansson (2004) instead studied the by-ear chord learning strategies by observing and interviewing musicians with far more experience—having played an average of approximately 20 years—with even representation among six (reportedly all-male) bass, keyboard, and guitar players trained either informally or formally.

Additionally, researchers who study the methods of those learning by ear tend to focus only on short melodic phrases, and provide participants with audio material that is uncharacteristic of a pop music recording. For example, Lahav et al. (2005) had participants learn the melodies from eight-bar songs using custom-designed software that synthesized MIDI notes on virtual instruments, allowing them to hear the accompaniment alongside the melody they learned to play by ear. Oswald (2022) had students also learn eight-bar melodies, though they were played from solo recordings using custom-designed audio software. Varvarigou & Green (2015) allowed students to listen to a repeating four-bar "pop-funk style" pattern played by a full band, but the students learned the bass melody while listening to a solo recording of it.

In contrast, Johansson (2004) presented participants with full-length band recordings to study the chord learning strategies of experienced rock musicians. However, these musicians were asked to play along with the recording to learn its chords while hearing it for the first time. That is, the participants were not allowed to stop or rewind the recording while identifying chords—their "performance" of the song was recorded in one take, then analyzed.

2.4 Memory and Music

To play songs by ear, musicians rely upon their memories of recordings. On a short time scale, musicians have to retain one or more just-heard notes in *tonal working memory* long enough for them to be found on an instrument. However, musicians can also draw upon longer term memories to sing, or learn to play well-known nursery rhymes such as *Twinkle, Twinkle, Little Star* on their instrument using only their recollection of its melody.

2.4.1 Tonal Working Memory

People can retain—on average—approximately seven digits in working memory (Miller 1956). This allows people to write down a phone number shortly after hearing it, or perform mental math. It is also the case that people can remember and manipulate *auditory memories* in a similar way (Schulze et al. 2018). Here we focus on one specific element of auditory working memory that is closely related to music: *pitch-* or *tonal working memory.*

Capacity

The processes that maintain tonal working memory appear to be separate from verbal memory (Schulze & Koelsch 2012). However, both may involve the same internal *phonological-* or *tonal loop* where words or tones are rehearsed, and it appears that their capacity may be linked (Schulze & Tillmann 2013). That is, those who can remember more digits than others can seemingly also remember more pitches.

Unfortunately, the method often used to evaluate one's tonal memory capacity bears little resemblance to the note seeking task of learning by ear. In many studies, participants hear one or more tones, a silent pause, and another sequence of tones that they are asked to compare with the first—correctly identifying a match or mismatch indicated they were remembered (Akiva-Kabiri et al. 2009; Schulze et al. 2012; Schulze & Tillmann 2013; Ding et al. 2018). By contrast, in ear learning, one must retain pitch(es) in memory, and compare them to attempts made to reproduce it on their instrument.

Despite these differences, these studies suggest that working memory for tones has an upper limit. While testing the hypothesis that the duration of tones would have an impact on memory capacity, Akiva-Kabiri et al. (2009) tested participants using both six- and nine-tone sequences. Schulze et al. (2012) tested the forward recognition task with five- to seven-tone sequences, but used only three- to five-tone-long sequences for their backward recognition task—asking participants to recall if the second tonal sequence was a reversal of the first. Similarly, Schulze and Tillmann (2013) later used five- and six-tone sequences for the forward task, but only three tones for the backward task—compensating for a significant drop in recognition performance they observed during a pilot experiment with longer sequences. Greenspon et al. (2017) opted to test participants with three- and four-note melodies. The selection of sequence lengths for many of these tests was based on earlier work, which includes that of Pembroke (1986) who tested the transcription accuracy of music theory students, or Miller's seminal study (1956) that claims seven digits can be retained in working memory.

Ding et al. (2018) does *directly* measure tonal working memory capacity in terms of note quantity. Here the researchers compared the performance of musicians and non-musicians, and also tested both musical and non-musical (i.e. not aligned to the Western music scale) sequences of tones between 2–24 notes long. Among their results, they report a marked difference between the maximum length of musical sequences for which musicians (16) and non-musicians (8) could demonstrate an above-chance recognition rate. However, both groups perform equally with a maximum of 8-tone sequences of random pitches that they could recognize.

Even though the above experiments look nothing like learning notes or chords from a popular music recording, three things seem clear from this survey of readings on short-term tonal memory: (1) the number of notes that can be held in working memory has an upper limit, (2) that number is likely very small, and (3) it seems to vary based on one's musical training.

Accuracy

Four studies asked participants to reproduce pitches as accurately as possible using an "instrument". In the implicit note memory (INM) task, a sine wave tone with a target pitch is played, and after some delay the participants are asked to make adjustments to a different tone until it matches their memory of the target (Van Hedger et al. 2015). This test provides researchers with a continuous measure of accuracy (i.e. distance from the target pitch) and offers insight beyond the binary result of "remembered" or "forgotten".

While the INM task is more like singing a pitch[1] than locating it on a musical instrument with discrete keys or frets, it more closely resembles what musicians do while learning from a recording.

In the INM task conducted by Van Hedger et al. (2015), participants were played target tones with frequencies aligned to the Western music scale, and after hearing 1000 ms of white noise they were played a different tone 1–7 semitones above or below the target. This second tone is adjustable using on-screen buttons that increment or decrement the pitch by either 33% or 66% of a semitone. For all 17 participants, the average error was within 40 cents (less than half of a semitone) of the target, and this figure decreased with the number of years spent playing an instrument. Their worst performer was within an average of 78 cents of the target—less than a semitone, or a one-fret difference on a guitar.

Wisniewski and Tollefsrud (2023) tested INM by playing a randomly selected tone in the range of 500–1000 Hz for 1s followed by a randomly selected 2, 4, 6, or 8s retention period (i.e. a silent pause). Then, users could apply fine-grained adjustments to a 500 Hz tone using a MIDI touchpad for up to 20 s, or earlier using a confirmation button on the touchpad. Here the researchers found that participants lost precision in their ability to reproduce tones as the retention duration increased.

Tollefsrud et al. (2024) modified the above scheme for testing INM in two separate experiments. For both, the target tone consisted of a fundamental plus a number of harmonics, and when the participants were asked to match the target they did not start at the same 500 Hz, but rather a pitch that corresponded to the first position that their finger was placed on the touchpad. Additionally, while the tones were also random, they were log-spaced within different ranges—134–357 Hz in the first experiment, and a wider 150–642 Hz band in the second. Most significantly, these researchers filled the retention period with either silence, white noise, or a number of tones played at different pitches, and found that the latter appeared to cause the most interference to the participants' memory. However, this interference did not seem to cause them to forget the target tone completely. Rather, their accuracy was lower when those participants tried to match it.

What all of these studies have in common is that participants are asked to match the pitch of a target tone. In all cases, the participants are provided with a tool for producing tones that is conceptually similar to playing an instrument. Here, the participants must rely upon their tonal working memory to maintain the target pitch while comparing it

[1] See also Hutchins & Peretz (2012), who describe a *slider* instrument that was meant to be compared with vocal pitch matching in an INM-like test.

to those they produce mechanically—a process that more closely resembles a musician copying notes from a recording with their instrument.

Duration

In those experiments where participants were asked to recall if a tone sequence matched one played earlier, the time interval between the two sequences was fixed. In Schulze & Tilma's (2013) study, participants were given 3 seconds between the target and comparison sequences. Ding et al. (2018) gave participants only 1 second before they were asked to recall the sequence. Akiva-Kabiri et al. (2009) did not report the interval between sequences.

Those studies that chose to vary the delay period tested single tones. Mathias et al. (2021) specifically recruited musicians for their tests, and asked them to report whether a second tone had a higher or lower pitch than the first after 0.5, 2, 5, or 10 seconds. As discussed earlier, Wisniewski and Tollefsrud (2023) tested participants using randomly selected 2, 4, 6, or 8 s retention periods between the target tone and their attempt to match it.

Given how few studies tried to measure the *lifetime* of a tonal memory, it is difficult to say how long one can expect to retain a note. However, this period of time appears to be very short for single tones—less than 10 s, and presumably shorter as more tones are added.

Vocal Pitch Imitation

There appears to be a link between short-term tonal memory and one's ability to accurately imitate pitches vocally. For example, Greenspon & Pfordresher (2019) reported that those with larger spans of tonal memory and better auditory imagery skills were also more accurate pitch imitators. Halpern & Pfordresher (2022) found in one experiment that those who could accurately reproduce melodies using their voice could more accurately recall notes from memory. In their second experiment, the researchers showed that those with high scores in a test of vocal pitch imitation demonstrated lower error rates when reproducing the starting note of a pop song recording they were familiar with. Additionally, they found years of musical training—0–12 years in this sample, with a mean value of 2—did not correlate with accuracy in this experiment. This connection between one's memory and singing ability may be explained by the *vocal sensorimotor loop* (Berkowska & Dalla Bella 2009; Pfordresher et al. 2015). Specifically, a combination of perception and fine motor planning that may be related to the encoding process for tonal memories.

Unfortunately, research shows that not everyone can reproduce pitches accurately with their voice. Pfordresher & Demorest (2021) found that approximately half the 632 participants they evaluated were able to do so, and the authors estimate the proportion is lower among the general population—that about a third of them can accurately reproduce pitches using their voice. However, those who are unable to reproduce pitches *vocally* can still match pitches using other means. In a study by Hutchins & Peretz (2012), they found that both musicians and non-musicians could more accurately reproduce pitches using a *slider* instrument that produced a voice-like sound than they could with their own voice. Note that in this study, the researchers did not ask participants to match their memory of a tone, but rather one that was playing continuously. Additionally, the continuous slider is unlike most pop music instruments that produce discrete tones, which could suggest that instrumental pitch reproduction is more easily achieved in such an exercise.

The results of the above studies suggest there is a connection between one's ability to reproduce pitches vocally and their tonal working memory. However, we have not seen studies that indicate whether one can improve upon their tonal working memory by practicing the ability to reproduce pitches accurately, or vice versa. Fortunately, we see some evidence that indicates those who may be deficient in either skill can still match pitches using other means.

Summary

The above studies do not test whether people can remember or reproduce pitches heard in a recording of pop music. However, they provide us with insight into the most basic mechanisms that may be involved in doing so. Specifically, these studies tell us that people can hold very few notes in short-term memory, they can use their memory of pitches to reproduce them mechanically, and these memories cannot be retained for very long.

2.4.2 The Mind's Ear

Auditory imagery is a form of perceptual long-term memory that allows people to recall—sometimes quite vividly—songs that they have heard in the past (Hubbard 2010). Effectively, these people can seemingly *hear* music with their *mind's ear* (Covington 2005).

Involuntary musical imagery (INMI)—colloquially known as *earworms*—are segments of a recently-heard recording (or performance) that replay in one's mind without any prompting, often during periods when it is left free to wander (Kubit & Janata 2022). This phenomenon is not a unique experience for musicians—it can also occur for those who neither sing nor play an instrument (Beaty et al. 2013; Liikkanen & Jakubowski 2020). That is, memories of recordings can be formed before one learns to sing or play from them.

There is evidence that such long-term memories can be formed intentionally using looping segments of audio. Kubit & Janata (2022) composed a collection of novel instrumental musical loops, and presented them repeatedly to participants before testing their ability to recall them. Here the researchers attempted to induce episodes of INMI, and wanted to see how that might impact the participants' long-term memory for musical sequences. Participants were scored for their ability to correctly recall the loops shortly after hearing them in the first session, and again one week later before completing a survey about the INMI experiences they had since the initial listening session. Kubit & Janata found that those participants who experienced INMI during the week between sessions demonstrated better accuracy when recalling those musical sequences, and suggest that INMI might play a role in the formation of long-term musical memories.

2.4.3 Absolute Pitch

When we say that a person has *absolute pitch*, we generally mean that they can readily name the notes they hear just as one might name colours on sight; but this skill is rare: it is estimated that fewer than 0.01% of the population have this ability (Deutsch 2013).

However, while most people can't name the notes they hear, a significant portion of the population can seemingly *remember* the absolute pitches of music they have heard before. In two experiments performed by Van Hedger et al. (2018), participants listened to segments from well-known recordings of popular music that was either played in the original key, or transposed ±1 semitone, and they could identify whether the song was in the correct key or not more than 60% of the time. Interestingly, those who scored well at this task also scored high in an INM task that was administered before the experiment, suggesting that those who demonstrate good tonal working memory capacity seem to form more accurate memories of songs. A later study by Van Hedger et al. (2023) showed that participants could also identify whether novel versions of popular recordings—either cover performances obtained from YouTube, or melodies played on a synthesized piano—were played in the original key.

Not only can people recognize whether a song is in the correct key, but it appears many can also reproduce notes in the correct key from memory. For example, in an experiment conducted by Halpern & Pfordresher (2022), participants were presented with a curated list of pop songs, and asked to choose the 10 they were most familiar with. For each song, participants attempted to reproduce the first note of the song on a digital (piano) keyboard—without hearing it, or trying to sing or hum the notes first. Out of the 46 undergraduate students they tested, 33% got within ±2 semitones of the true pitch, and 17% were within ±1. In Levitin's (1994) study, participants were asked to sing popular songs they were familiar with—more than half sung the first three notes within a semitone of the correct pitch, transposed to match their vocal range. That is, if the original pitch was C3, it would be considered a perfect match if the participant sung C2 or C4, and -1 semitone away if they sung B2 or B4.

2.4.4 Memory of Melodies

In an experiment conducted by Halpern (1989), participants were asked to sing the first note of familiar songs (a collection that included *Twinkle, Twinkle, Little Star*; *Joy to the World*; and *Somewhere Over The Rainbow*), and while the starting pitches of men and women differed by an average of 11 semitones for the same songs, each person produced four trials that were stable within 1.28 semitones. Unlike the studies we discussed in Section 2.4.3 that asked participants to recall popular music recordings, all but one of the songs—*Somewhere Over the Rainbow*—were found to have notated versions with varying starting notes (and hence, keys) (Halpern 1989, Table 1). That is, these songs are not universally associated with a starting pitch. However, people appeared to maintain an internally consistent stable reference pitch in memory that is reflected in their attempts to reproduce them.

If a person remembers a melody in a specific key, they can still recognize it when heard in a different key—they perceive the two performances as the same piece of music, and their memory likely consists of a more abstract encoding than a set of absolute pitches and durations (Snyder 2014). In two experiments conducted by Plantinga & Trainor (2005), parents familiarized their infant children with an English folk song for seven days, then returned to the lab on the eighth day to measure the infant's preference for one of two melodies. In the first experiment of 32 infants (13 female, 19 male, mean age of 6.02 months), the researchers found that the infants showed preference for transposed versions of the familiar folk songs over ones they were not familiarized with. In the second experiment (32 infants, 15 female, 17 male, mean age of 6.1 months), the infants were presented the same song twice—once in its original key, and once transposed—and the researchers found no significant effect of the transposition on the infant's preference. These experiments provide evidence that—even at a very early age—one's memory for melody is not tied to specific pitches.

Combined with the studies we saw in Section 2.4.3, the literature suggests that the memory of melodies has two major components—a pitch-independent set of contours or intervals, and a pitch reference that anchors the melody to one or more prior listening events.

2.4.5 Long-Term Memory and Musical Structure

Long-term memories are a result of structural changes to the brain, and they can last a lifetime. Such memories are unconsciously held in the mind, and are activated when one's experiences cue associations with existing memories (Snyder 2014). For example, this mechanism is called upon in Halpern's (1989) study where participants were provided with lyrics to help activate their memory of songs.

Additionally, long-term memories can be developed further when they are recalled repeatedly, provided they are semantic and not episodic (or autobiographical) memories; the latter can change upon recall, and the former requires repeated exposure to develop (Snyder 2014). Reflecting on the discussion in Section 2.4.2, it makes sense why INMI helps reinforce one's long-term memory of a song—each experience triggers the recall of the memory formed during (possibly repeated) listening.

The representation of a long-term musical memory is not fully understood, though it appears hierarchical in nature. Segmentation—delineating structural boundaries within a musical piece—is foundational to the encoding of such a long-term memory, and differs at each hierarchical level; for example, significant transitions between segments of a melody at a lower level, and major harmonic changes at a higher level (Snyder 2014). Additionally, one's ability to judge significance and form hierarchies may depend on factors such as their exposure to similar music, or musical training.

2.4.6 Summary

This body of literature provides us with valuable insight into the kind of memory that is called upon while learning songs by ear—from recordings, or even long-held memories. Short-term, tonal working memory allows a musician to recall one or more notes for a brief period of time while attempting to play those same notes on their instrument, and their ability to vocally reproduce pitches may improve their ability to encode such memories. Additionally, musicians can form rich, long-lasting memories of a piece of music that allow them to be replayed in their minds, and their experiences of INMI seem to help in the long-term encoding process. Such memories appear to be somewhat accurate in terms of their connection to the absolute pitch of the original recording, though one's memory of the music is not merely a collection of pitches and durations. Rather, what people remember is more abstract, and allows for recognition and reproduction with respect to changing reference pitches.

2.5 Technological Supports

Here we explore the technology that enables musicians to learn music directly from recordings. From the record players that first allowed them to play music on demand, to DSP and MIR methods that analyze and manipulate digital audio signals, learning popular music by ear was largely a technical revolution.

2.5.1 Purpose-Built Technology

A musician's ability to interact with a recording is limited by the technology used to play it. At one extreme, a radio offers the least control. The musician encounters a song by chance, and would have to wait for another opportunity to hear it again—ideally with their instrument in tow. By obtaining a physical copy of the recording, musicians could use a record player to start and stop playback at will. Having this level of control is essential, as Bennett notes that:

> [...] recorded songs are not gotten through the usual mode of audience exposure to playback events, but by the specifically defined event of copying a recording by playing along with it and using the technical ability to play parts of it over and over again. (Bennett 1980, 138)

If their turntable allowed it, the musician could also slow the playback of a 33rpm record to 16rpm with the press of a button. This interaction allowed musicians to hear quickly-played phrases at half speed. However, some musicians appeared to require more than this, and used their ingenuity. As Jerry Garcia of the Grateful Dead recounts in an interview with Bill Barich, one could manually alter the rate of playback:

> [. . .] I'd picked up the five-string banjo in the Army. I listened to records, slowed them down with a finger, and learned the tunings note by note. (Barich 1993)

Fortunately, finer control did not always require such lateral thinking: some record and cassette players offered additional features that improved the ear-learning musician's quality of life. For example, the Marantz Superscope C-190 was a cassette player sold in the late 1970s that offered variable playback speed with adjustments of ±20%. It also had a resettable counter mechanism and review/cue features that—together—could help musicians locate specific points in a recording. Devices like these that offered useful features for those learning by ear were not necessarily marketed at musicians. For example, one might wish to slow playback while transcribing a spoken interview.

Unfortunately, both vinyl records and cassette tapes had significant limitations. Reducing playback speed would cause the pitch of the recording to change, and the media would degrade with repeated use. As digital audio became more readily available in CDs and later formats, this limitation disappeared; provided the physical media was handled delicately and kept clean. Digital audio also introduced precision to navigating recordings: the beginning of a track was easily found and revisited, and players with LCD displays indicated the track position in minutes and seconds. Such affordances allow the musician to note the time where a verse begins, for example, and return to the same spot during a later session—even when using another CD player.

What we refer to as purpose-built technology in this domain are features—not specific to musicians—that the manufacturer *intends* its customers to use as they interact with recordings. For example, the C-190 cassette player provides a purpose-built continuous speed adjustment control, but a record player leaving enough space for a finger to drag along the turntable is very likely a "happy accident". Long after the introduction of CDs, more intentional designs started to appear once sufficient DSP capabilities became available. For example, the TASCAM CD-GT1[2] is a product that was targeted specifically at musicians (the GT stands for Guitar Trainer). This device could play standard CDs, but most importantly it allowed playback to be slowed without changing the pitch of the original recording.

In addition to hardware, many software packages were introduced with purpose-built features that help musicians learning by ear. For example, the Amazing Slow Downer first appeared on the Mac in 2000, and is still actively maintained.[3] Unlike hardware offer-

[2] https://tascam.com/int/product/cd-gt1/top

[3] https://www.ronimusic.com/download/versioninfo/asd_mac_history/History.txt

ings, software enjoys the benefit of a near-infinite design space in which new, musically-oriented human-recording interactions can be explored further.

2.5.2 Manipulating Playback Speed

The playback speed of a digital audio recording can be modified without affecting its pitch by using techniques such as waveform similarity overlap-add (WSOLA), and the phase vocoder that operates in the frequency domain (Portnoff 1976; Driedger & Müller 2016). The latter is better suited for use with pop song recordings as frequency domain techniques are far more adept at manipulating polyphonic audio. However, care must be taken in the implementation in order to avoid creating undesirable, audible artifacts such as *phasiness*, and *transient smearing* (Laroche & Dolson 1997).

Briefly, the phase vocoder consists two main stages. During the analysis stage, the incoming signal is divided into evenly-spaced overlapping time blocks that are processed using the Fast Fourier Transform (FFT) to generate spectral frames. In the synthesis stage this is reversed: an inverse transform of the spectral frames produce the time domain output signal. This is an identity transformation that can reproduce the input signal unmodified (Portnoff 1976). However, to *slow* audio, new spectral frames are generated by interpolating the analysis frames before synthesizing the output. Audible artifacts are introduced as a result of this interpolation stage.

To combat audible *phasiness* artifacts, it is important that the phase components are interpolated correctly between the spectral frames. Puckette (1995) introduced a phase locking technique that ensures the phase component of adjacent FFT channels move in sync with one another. Laroche & Dolson (1997) extend this core idea by selectively applying phase locking to only peaks in the FFT spectrum. Průša & Holighaus (2017) later expanded the approach by tracking and propagating phase gradients in the time (inter-frame) and frequency (inter-channel) axes, and claim that this resolves *all* phase vocoder artifacts. However, the researchers admit that listeners may actually prefer a combination of their technique with one that deliberately *sharpens* transients.

Eliminating *transient smearing* requires that transients from the original signal are preserved in the synthesized output. Duxbury et al. (2002) split the signal into steady-state and transient components, and they ensure that the original transients appear in the output by briefly resetting the stretch ratio to 1.0 and re-initializing phase components during periods of high transience. To overcome the impact of these temporary speed-ups, the stretch ratio is adjusted adaptively so that the output signal retains the tempo of the original recording. Röbel (Röbel 2003a,b) instead proposes a method that precisely locates transients within the analysis windows, and reinitializes phase components in narrower bands of the spectrum. Both approaches leverage the core idea that the onset of a new sound is not continuing those vibrations that came prior—a new starting point must be established by reinitializing the phase components that belong to the source of the transient.

Using these transient-preserving techniques requires a compromise between frequency and temporal resolution: shorter analysis windows offer more temporal detail, but do so at the expense of coarsely spaced frequency bins. Conversely, longer ones trade temporal detail for the ability to more accurately resolve sinusoidal peaks. Juillerat & Hirsbrunner (2017) attempt to get the best of both worlds by splitting the input signal into three parts—low, medium, and high transience—and each is processed at a different resolution. In this novel approach, phase processing is performed only on the frames with the highest frequency resolution, while a magnitude-only phase vocoder manages the medium- and high-transience signals that enjoy progressively better temporal resolutions.

As we have shown, a time stretching implementation grows more complex as it becomes important to preserve the clarity of the original recording when it is slowed. It is important that—even at extreme stretching ratios— the intelligibility of individual notes is preserved so that musicians can hear them clearly, and with the same relative spacing in time as the original recording.

2.5.3 Computational Understanding of Recorded Music

Today we can extract a great deal of musically meaningful information directly from digital audio recordings. Research in the field of MIR has provided methods that make it seem unnecessary for musicians to learn music by ear. For example, the ability to locate beats in the recording, recognize the song's key, and estimate the chords. These individual tasks each have a rather long history of research, which we will briefly touch upon below.

In general, most tasks in MIR are performed in one of two ways: using heuristic methods, or with the aid of deep learning. Here we are primarily interested in heuristic methods due to the practicality of their implementation. Specifically, it is difficult to obtain a sufficiently large corpus of (often commercial) audio recordings to train with, and it can be time-consuming to align existing labels to a collection of audio data (Bittner et al. 2019). Further, most of the canonical data sets suffer from issues like the under-representation of certain rare chord classes (McFee & Bello 2017). Additionally, expanding a training corpus requires a great deal of expert labour (Burgoyne et al. 2011). In light of these challenges, and my specific research goals, deep learning methods will receive far less attention below.

The locations of every beat in a recording—which may not be spaced equally in a live performance—can be extracted using both heuristic and deep learning methods. Heuristic beat tracking methods generally begin by measuring perceptually significant events called *onsets* to generate a continuous-valued novelty curve, using filter banks in the time domain (Scheirer 1998), analyzing successive frames of spectral data in the frequency domain (Bello et al. 2005; Dixon 2006; Thornburg et al. 2007; Robertson et al. 2013; McFee & Ellis 2014b), or synthesizing a more consistent pulse from those frames (Grosche & Müller 2009, 2011). Such a novelty curve can be turned into a set of beat locations using a dynamic programming algorithm by Ellis (2007) that retains values above a threshold, locates the peaks, then identifies those that most likely correspond to the *tactus*—the pulse that corresponds to the times when someone would tap their foot along with a song (Klapuri et al. 2006). In contrast to heuristic methods, those that rely on deep learning can often provide the beat locations with no need for a novelty curve (Böck & Schedl 2011; Böck et al. 2014; Zhao et al. 2022; Cheng & Goto 2023).

It is also possible to estimate the key of a song from its recording. That is, the specific major or minor scale that defines the *tonal centre* of the piece of music. While we class many of these methods as *heuristic*, early methods rely upon supervised training, or statistical data that was collected from a corpus of notated music (Temperley 1999). For example, the key estimation algorithm by Gòmez (2006) tests a vector that represents the mean harmonic content of a recording against a set of statistically-informed templates for each possible key. Unlike more modern methods, some early key detection algorithms still rely upon the supervised training of Hidden Markov Models (Peeters 2006a,b; Lee 2007; Noland & Sandler 2007), though their architectures are largely informed by heuristics. A recent approach by Korzeniowski & Widmer (2017) instead uses a dense convolutional neural network (CNN) to build a classifier that can estimate a song's key using its spectral representation as input.

Extracting the chords from a recording is of particular interest to popular musicians, and there are very many approaches for doing so. The importance of this task is apparent, as chord recognition has enjoyed more than 20 years of active research (Pauwels et al. 2019). As with key estimation, early methods for chord recognition rely upon supervised training of HMMs (Sheh & Ellis 2003; Papadopoulos & Peeters 2007; Harte 2010; McVicar et al. 2014). Such systems work from *chromagrams*, a modified version of the spectrogram that collapses the full range spectral data to a 12-dimensional vector at each time instant. The choice of method for generating this representation can have a significant impact on the system's performance (Cho et al. 2010; Cho & Bello 2011, 2014). Deep learning methods are also plentiful, with a wide variety of architectures that range from CNNs (Korzeniowski & Widmer 2016), to encoder-decoder RNNs (McFee & Bello 2017), and even the use of transformers (Chen & Su 2019; Park et al. 2019). Most of these systems only consider the simplest chord qualities: major and minor. However, some researchers have proposed systems that can draw from larger vocabularies (Cho & Bello 2014; McFee & Bello 2017; Park et al. 2019).

Tasks such as these can work together to develop a semantic understanding of what is happening in a musical recording. While they are theoretically capable of producing sheet music as output (see, e.g., Donahue et al. 2022), this understanding can also be leveraged for the sake of a musician that is trying to learn a recording by ear. We explore these possibilities further in Chapter 6.

2.5.4 Summary

By-ear learning is made possible by the technology that allows for on-demand playback of recordings. When they are provided with controls to reduce the speed of playback, musicians can gain an advantage by hearing notes that are played too quickly in the original recording at a more comfortable pace. However, it was a set of DSP algorithms that let them do so without a corresponding drop in the pitch of playback. Using techniques from the field of MIR, we have an opportunity to provide novel tools that further improve upon the experience of musicians learning by ear.

2.6 Summary

Popular musicians have been learning music by ear for decades, and they do so by using recordings in place of sheet music. However, we have seen nothing in the literature that attempts to characterize exactly *how* experienced musicians interact with recordings as they learn from them. What we do find in the literature is evidence that these interactions are occurring. We also find a great deal of literature dedicated to the technology that powers the specific interaction of slowing playback, as well as newer technologies that have the potential to drive novel interactions in the future.

Chapter 3

Opportunities and Limits of Online Video Studies

In this chapter we explore the benefits of using online video as research data, discuss how it has been used in various fields, and present some difficulties researchers may face when conducting studies of their own. Underlying all good research is a keen consideration of ethics, so we also discuss how online video studies face unique challenges in this regard.

Much of this chapter is applicable to any online service that hosts video content. However, our focus is almost exclusively directed at research conducted using videos obtained from YouTube—it is prevalent in the literature, and also what we used in our own studies.

3.1 Why Study Online Videos?

A study of videos obtained from an online source such as YouTube is especially attractive for researchers as it requires very little investment to begin. Additionally, studying a skill that requires access to a participant's working environment, or their collection of non-portable tools can incur prohibitive travel and/or shipping costs. While it is conceivable that a researcher could make a convincing case to obtain the necessary funding, it might be wasteful to do so for early-stage studies.

For instance, a video-based study can allow a researcher without sufficient domain expertise to prepare for more conventional in-person studies. By analyzing a collection of videos that depict skilled human activities, a researcher could reinforce their domain-specific knowledge in preparation for an ethnography, or a contextual inquiry. Additionally, the results of such an online video study can generate hypotheses (e.g., Chapter 4) that shape the research questions in the researcher's future work.

User-generated videos posted online can serve as a proxy for in-person observations of what people do "in the wild". This is evidenced by studies in many fields of research, such as healthcare (Sampson et al. 2013) and human-computer interaction (Bartolome & Niu 2023). Additionally, researchers have studied a diverse set of human activities in this way, such as cooking (Paay et al. 2015), the use of touchscreens (Hourcade et al. 2015; Vatavu et al. 2022), gaming (Dao et al. 2021; Wentzel et al. 2022; Gonçalves et al. 2023), administering healthcare to infants (Harrison et al. 2014, 2018), and physical altercations in public places (Weenink et al. 2022a,b).

3.1.1 Unparalleled Access to People and Spaces

By studying user-generated online videos, researchers have the opportunity to observe populations that would be difficult (or impossible) to recruit, and in locations that would otherwise be inaccessible.

Some spaces are challenging to replicate in a lab environment, or impractical to observe in person. For example, Paay et al. (2015) studied YouTube videos of people cooking together in kitchens, and claimed that placing a researcher and/or camera in peoples' homes would be both impractical and detrimental to such a study. Li et al. (2021) *also* studied cooking videos found on YouTube, though they focused on the ways that blind or visually impaired people prepared food in their kitchens. Dao et al. (2021) gathered clips from "VR fails" compilations on YouTube that were captured in private residences. Given the physical contact and injuries to both person and property, a researcher would be unlikely—for ethical reasons, and also self-preservation—to capture such events in an in-person study.

Another opportunity presented by online video content is the potential for gathering information within contexts that may otherwise be "off limits", such as commercial spaces. For example, Chattopadhyay et al. (2021) were able to observe software developers in "day in the life" videos, who often included footage from their workplaces. Researchers may not be so lucky to obtain such candid material if they instead approached companies for on-site access of their employees. Similarly, Harrison et al. (2014; 2018) were able to collect videos of infants receiving health care that—presumably—were all captured in medical facilities that may not have allowed such wide access to conduct their observations.

The quantity and geographic diversity of spaces that are accessible to researchers is also quite wide. For example, Nielsen et al. (2023) were able to observe human-robot interactions in a variety of public spaces that included grocery stores, shopping malls, and airports—the latter of which were located in South Korea, the United States, and Singapore.

Studies have also used YouTube to gather observations from virtual environments, often collecting data of events that researchers would be unlikely to capture during intentional ethnographic embeddings within those spaces. Zheng et al. (2023) studied instances of harassment and other risks to users in social VR environments, using video captures uploaded to YouTube as a way to gain access to spaces that only exist in VR. Gonçalves et al. (2023) similarly used video captures of blind users playing visual-centric video games to discover various strategies they used to navigate those virtual worlds.

In other HCI studies, the use of YouTube videos gave researchers direct access to observe how people with a range of physical disabilities interacted with various technologies like touch screens and game controllers (Anthony et al. 2013; Wentzel et al. 2022; Vatavu et al. 2022; Gonçalves et al. 2023). For example, Wentzel et al. (2022) performed a content analysis of 74 YouTube videos to identify the different ways that people configured multi-modal inputs to control PC and console games for disabled gamers. It would be challenging to execute such studies with in-person subjects, let alone recruit a sufficient number of willing participants that fit the complex needs of those studies.

Many of the above studies also benefit from their potential to gather data across time. For example, if one were to try and replicate the study by Anthony et al. (2013), a researcher could identify whether touchscreen interactions have changed meaningfully over the past 10 years. Additionally, by focusing on specific individuals by way of their YouTube channels a researcher could conduct something that resembles a longitudinal study. For example, in a study like O'Leary's (2020) that focused on popular ukulele channels, a researcher could measure the rate at which one's skills as a musician have progressed over a number of years.

3.1.2 Observing Human Behaviour and Interactions With Technology

Studying a collection of videos allows researchers to observe more subtle elements of their content. For example, one can study the actions and expressions that accompany what is said in a video; identify non-verbal sounds produced by instruments, animals, or other physical objects that may not be seen; and also the interactions between people and their environment. For example, Harrison et al. (2014; 2018) performed a qualitative content analysis of YouTube videos to review the methods used to soothe infants during immunizations and blood tests. Weenink et al. (Weenink et al. 2022a,b) studied interactions between people during the events leading up to physical altercations in public. Gibson (2022) very closely analyzed video reviews of a single guitar pedal, and characterized details about how the product was presented to viewers, such as reviewers trying to describe its sound verbally.

However, researchers have also used YouTube to gather observations of people interacting with technology in real-life situations. For example, Anthony et al. (2013) studied how people with limited motor abilities interact with touchscreens, and found that some use their nose or feet to activate the touch-sensitive surface. Lovato & Piper (2015) analyzed videos of children interacting with Apple's Siri voice assistant, and found it was used for making phone calls, to express anger, or simply to satisfy the child's curiosity. Hourcade et al. (2015) analyzed videos of toddlers and infants interacting with touchscreen tablets, capturing information such as the position of children in relation to the device, and whether it was held by a parent, laid on a table, propped up, etc. Mauriello et al. (2018b) analyzed video footage of novices using thermal cameras, and identified situations where they are used such as pointing the cameras at everyday objects to see how they appear differently when viewed in infrared. Komkaite et al. (2019) studied videos of people demonstrating the use of insertable devices—those implanted in the body, often placed beneath one's skin—to understand both their interactions with the technology, and why they chose to have it implanted. In the study of VR fails by Dao et al. (2021), interactions with VR technology were observed in a collection of videos depicting negative outcomes for VR users, and identified a collection of *breakdowns* in those interactions when the users hit the limits of the technology—for example, slapping spectators, or hitting walls. Wentzel et al. (2022) observed how a variety of assistive devices are used and configured by users with mobility challenges so that they could play video games on both consoles and gaming PCs. Vatavu et al. (2022) were able to identify a number of problems faced by wheelchair users interacting with public displays, such as input areas that were out of reach. Nielsen et al. (2023) studied how people interacted with robots that were installed in public places, specifically focusing on unguided interactions—those that occurred without training, or help from staff. Gonçalves et al. (2023) studied the video captures of blind and visually impaired users playing video games, and discovered how they interact in novel ways with those games to navigate them, or make it easier for themselves to perform game-specific tasks.

In nearly all of these studies, the researchers' findings could be used to guide the design of novel technologies, or improvements to existing ones. By watching footage of people in their natural environment—at home, or their workplace—researchers can access insights that would otherwise be difficult to obtain.

3.2 How Online Video Studies Are Conducted

3.2.1 Obtaining Videos

YouTube offers a search function that allows its users to find videos matching a given set of keywords. While many researchers use the website (e.g., Rotman et al. 2009; Harrison et al. 2018; Kong et al. 2019; Vatavu et al. 2022), the availability of an official API provides an opportunity for programmatic queries of YouTube's video collection (e.g., Rieder et al. 2018; Wentzel et al. 2022; Zheng et al. 2023). The latter method can offer some insulation from YouTube's recommender system, though researchers who use the web front-end have employed private (or incognito) browsing functionality to achieve the same goal (e.g., Vatavu et al. 2022; Altunisik et al. 2022). Regardless of the method used to access YouTube's search functionality, we focus here on the various *query strategies* that researchers use to obtain a collection of videos.

Quite often, researchers gather videos with searches that specify single keywords. For example, Basch et al. only specified "Ebola" (Basch et al. 2015) or "Zika virus" (Basch et al. 2017) in their studies of the information spread about those infectious diseases, and Blythe & Cairns (2009) searched for "iPhone 3G" to study the public's reception of Apple's device.

As the researcher's concerns broaden, or the topic of study cannot be characterized with a single term, it is necessary for researchers to conduct multiple searches. For example, Kong et al. (2019) specified 11 query strings that included "vape tricks", "e-cig smoke tricks", and "how to do vape tricks", but did not explain how they generated their list. By contrast, Harrison et al. (2014) indicated that "baby injection" and "baby vaccine" were chosen based on results from Google Trends indicating both terms were frequently specified in searches.

A common method found in the HCI literature is to form a cartesian product from two sets of keywords, and perform queries with each possible combination of them (e.g., Anthony et al. 2013; Li et al. 2022; Wentzel et al. 2022). For example, combining

[Apple, Android, . . .] with [tablet, smartphone, . . .] would yield “Apple tablet”, “Apple smartphone”, “Android tablet”, and “Android smartphone”. This approach can yield many queries, and hence a large number of videos to consider. For example, Wentzel et al. (2022) performed 480 queries to obtain a total of 2061 unique candidate videos that ultimately produced the final set of 74 videos that were studied. Unfortunately, it is not made clear exactly how that set was whittled down to such a small size, or how many of the 2061 videos had to be viewed during that process. Mauriello et al. (2018b) expanded upon this idea by automatically generating additional search terms using video metadata gathered during an initial round of queries.

Unfortunately, the increasing complexity in researchers’ methods is a direct consequence of YouTube’s limited query functionality, which lacks the kind of features one might expect to find in an academic database—for example, boolean operators, or the ability to specify that only titles or descriptions are matched. We discuss this issue further in Section 3.3.1. A related problem is that these strategies yield an *overwhelming* number of videos—most of which are irrelevant. We discuss how researchers try to work with these large data sets efficiently in Section 3.3.2.

3.2.2 Evaluating Information in Videos

Quite often, researchers are not concerned with the details of the physical activities of people who appear in videos. Rather, it is the information being conveyed—often verbally—that is most salient.

In public health, there are many studies that focus on the spread of (mis)information about viruses. For example, Basch et al. analyze the content of YouTube videos that discuss Zika (2017) and COVID-19 (2020) viruses, and both Basch et al. (2015) and Pathak et al. (2015) studied videos about Ebola Virus Disease. In the field of health informatics, Madathil et al. (2015) conducted a systematic review of studies that evaluate the overall quality of healthcare information that can be found in YouTube videos.

There are also studies concerned with public safety and well-being that aim to derive meaning from the content in the videos. In these studies, what the researchers seek to understand is not directly tied to what it is they are searching for. For example, Kong et al. (2019) studied YouTube videos that depict 25 different vape tricks to better understand how vaping is promoted to youth online. Kelly-Hedrick et al. (2018) performed a content analysis of videos discussing experiences with infertility to determine what topics or attributes appear to attract more viewers. Hawkins and Filtness (2017) analyzed the content of videos on YouTube to study perceptions of driver sleepiness.

In music education, researchers have used YouTube to study the content of instrumental lesson videos. For example, Kruse & Veblen (2012) analyzed guitar, banjo, fiddle, and mandolin lesson videos to identify the topics covered, teaching methods, and other details such as where the video was filmed, and the perceived age of the teacher. Whitaker et al. (2014) studied a wider range of videos related to music education, and extracted characteristics such as the genre or instrument that was featured, and whether the video contained a lesson, a performance, or if it was technology-focused. O'Leary (2020) studied lesson videos from popular ukulele channels on YouTube, reporting the distribution of video types, and the proportion of views for each type.

3.2.3 Analyzing Audiovisual Content

Video is an attractive medium for research because it is an incredibly rich source of data. Not only can researchers study what is said in a video, but they can also analyze body language, identify objects in the video, or judge its musical content. However, the methods used to study audiovisual media differs greatly depending on the field, the subject matter of the videos, and the researchers' goals. In many cases, these methods have been developed long before online video was used as research data.

In some studies, videos are labeled superficially, often to supplement a more in-depth analysis. For example, Blythe & Cairns (2009) assigned categories such as unboxing, review, or demonstration, to a set of iPhone-related videos on YouTube. They applied labels based on a mixture of their metadata (e.g., the word "unboxing" appears in the title) and what is observed in the video (a product is removed from its packaging). Paay et al. (2012; 2013; 2015) similarly assigned categories to cooking videos, such as home videos, providing cooking advice, or documentary-like. In studies of consumer health videos, researchers consider whether YouTube videos were produced by news organizations, health professionals, or independent sources (e.g., Basch et al. 2015; Pathak et al. 2015; Basch et al. 2017, 2020). O'Leary (2020) studied ukulele video *channels* on YouTube, and videos were assigned to categories such as tutorials, performances, or equipment reviews. These labeling practices are necessary for studies where researchers expect to discover variety in the *kinds* of videos that are considered. That is, these researchers are often not seeking videos that depict the exact same task. For example, if O'Leary (2020) focused only on tutorials, or Blythe & Cairns (2009) on unboxing videos.

When researchers wish to study data that can only be gathered by *watching* the videos, the axes on which they are labeled become more fine-grained. These axes may be based upon existing theoretical frameworks, or developed inductively using a small subset of the data. For example, Anthony et al. (2013) extracted characteristics of the videos they watched such as the kinds of devices used, the context where the video was captured (e.g., at home, at work), and the perceived age of those who appeared in the videos. Wentzel et al. (2022) coded videos that depicted the gaming setups of users with limited mobility, noting the quantity and kind of devices that appeared in videos, as well as the various ways people interacted with these devices. Similarly, Vatavu et al. (2022) gathered information about the touch capabilities and orientation of interactive displays that appeared in videos of wheelchair users interacting with them. When Dao et al. (2021) studied VR “breakdowns” by analyzing clips from a set of YouTube compilation videos, they considered details such as the presence of spectators, and if the video feed from the VR headset was made visible to them. In their study of dashcam footage of motor vehicle collisions and near-misses with moose, Rea et al. (2018) noted whether the driver swerved or slowed their vehicle, or which direction the moose approached from.

Sometimes researchers aim to quantify certain elements of a video based on what they observe while watching it. For example, Harrison et al. (2014; 2018) used the FLACC (Face, Legs, Activity, Cry, Consolability) scale and the Neonatal Facial Coding System (NFCS) to obtain a measure of pain from videos of infants receiving immunizations and blood tests. In their dashcam study, Rea et al. (2018) timed how long the moose was visible before a collision or near-miss, and measured—in lane widths—how much vegetation was cleared from the driver’s view.

Researchers can also extract very rich data from videos that allows them to perform even deeper analyses of their content. For example, Paay et al. (2012; 2013; 2015) generated layout diagrams of the kitchens depicted in the videos, including details such as the camera’s position and where people stood in relation to one another. As an alternative to transcription, researchers can also use established methods from conversation analysis and ethnomethodology (ten Have 2023) to yield rich data from videos. These methods have been used to analyze human behaviour in online videos (Weenink et al. 2022a,b), or study details about how products are reviewed (Gibson 2022), but they also have their place in HCI (Crabtree et al. 2000; Crabtree & Rodden 2004; Suchman 2006; Fischer et al. 2016; Tuncer et al. 2020, 2021). For example, Suchman’s (2006, ch. 9) seminal study in the 1980s analyzed users’ interactions with a photocopy machine to identify flaws in the design of a built-in “expert help system”. More recently, Tuner et al. (2020; 2021) analyzed the recordings of 10 pairs of participants interacting with YouTube as they followed instructional videos to perform everyday tasks such as applying makeup or changing the brakes on a bicycle.

There is clearly a wide range of techniques that researchers have used to extract data from audiovisual content. Additionally, these techniques can be combined to suit the researcher's goals, the kind of videos that are being considered, or the information that is to be obtained.

3.2.4 Grounded Theory on YouTube

Qualitative studies of online videos appear to be well-suited for the application of the grounded theory method (Corbin & Strauss 2008; Charmaz 2014). For example, researchers can expect to encounter variability in the information density within a collection of user-generated videos, and grounded theory allows for a mixture of data sources that contribute different details to the study (Corbin & Strauss 2008). Additionally, the wealth of video data that is available from sources like YouTube allows researchers to engage in purposeful sampling methods (Patton 2014) that can guide the collection of additional data based upon emerging concepts.

In the case of a video-based study, analysis consists of the researcher *coding* videos based on their observations, comparing them with codes applied to other videos, and gradually discovering those things that videos have in common with each other, or where they differ. The codes are collected into *concepts*, and then *categories (*or *themes)*. As new data and insights emerge, existing data may be revisited to refine the findings. During this process, researchers create *memos* that capture the evolution of their thinking, and ultimately these analytical notes are meant to become the results intended for publication. This iterative practice is meant to cease when *theoretical* or *conceptual saturation* is achieved—the point at which nothing new is learned from additional data (Corbin & Strauss 2008).

True to its name, this research practice is meant to produce *theories* that are *grounded* in real-world data. However, this practice rarely allows a theory to develop from only a single study. For practical reasons, studies that follow this method can only offer a step in that direction. That is, a typical grounded theory study yields what Corbin & Strauss (2008) call a "rich, thick description", and an analysis of the concepts that emerged. This discussion in the literature occurs alongside—and remains *grounded* in—the data that was collected. For example, each of the categories will contain elements from the data (usually, quotes) to help illustrate their concepts.

Unfortunately, the seemingly flexible nature of grounded theory leads to an apparent difficulty for researchers to apply it consistently across a variety of studies (Qureshi & Ünlü 2020). This challenge can also be seen in HCI research. For example, Li et al. (2022) use a mixture of open coding and affinity diagramming to generate *themes* under which codes were grouped and refined. By contrast, Rotman & Preece (2010) collected codes and concepts while analyzing the video content, later arranging them hierarchically to reveal *higher-level concepts*. Both studies ultimately appear to produce the same kind of results, but these kinds of variations in their descriptions make it challenging for other researchers to identify a sound practice to follow.

An additional challenge—not unique to video-based studies in HCI—is that what may *appear* to be an application of grounded theory is actually qualitative content analysis (Cho & Lee 2014). For example, Niu et al. (2022) describes a *grounded-theory-based* approach that only yields the code book that is used for a more conventional content analysis. However, they do not appear to use purposeful sampling, collect data in tandem with coding, nor do they claim to have reached *theoretical* or *conceptual saturation*—the point at which nothing new is learned from additional data (Corbin & Strauss 2008). By contrast, we find all of these elements in a study by Rotman et al. (2009), who studied users' feelings towards the YouTube community using both video transcripts and comments from viewers.

3.3 Why Online Video Studies Are Challenging

3.3.1 Systematicity and Exhaustiveness

Many YouTube video studies claim that one of their benefits is systematicity. We find this claim suspect. While not all reviews that claim to be systematic can be conducted like those performed by medical researchers (Clark 2013), reviews of video content from YouTube are often so far removed from such a methodology that it feels disingenuous to label them as systematic.

One characteristic of systematic literature reviews is reproducibility—that one can expect a given search query using academic databases can be repeated at a later date (Sampson et al. 2013), and only those papers published since the initial query would differ. A future researcher that repeats this query must consider that authors may withdraw publications, however such events are uncommon and unlikely to impact the set of literature significantly. Unfortunately, such repetition is impossible when querying YouTube using default parameters, as it returns unstable results—even when executed on the same calendar day. Experimental results that demonstrate this phenomenon are presented in Section 3.3.4.

I am certainly not the first to express discomfort with the idea of systematicity in a YouTube video study. Sampson et al. (2013) performed a systematic review of YouTube-based consumer health studies to design a methodology that informed those of subsequent systematic video review studies (e.g., Harrison et al. 2014, 2018). In this review, the researchers pointed out that sites like YouTube have content that changes daily, and a relevance algorithm that is proprietary and unstable. They propose that one can overcome this challenge by ending the screening process based on pre-defined stopping criteria—e.g., once 20 consecutive videos are found ineligible. Sampson et al. suggest that the researcher must be comfortable knowing that videos will be missed, yet they also claim—without evidence—that “the likelihood of missing a large number is low given the relevance ranking” (Sampson et al. 2013, p.11).

Online video services like YouTube rely upon the uploader to supply metadata with their videos—a title, description, and tags—that serve as both a description of the content for viewers, and something that the search engine can index. While my own experiments (Section 3.3.4) suggest that a video’s transcript, or other aspects of their content, may not be considered in searches via YouTube’s own website, the Google search engine allows content creators to opt in to a more sophisticated indexing that can surface relevant moments within videos[1].

However, there are specific phenomena that can only be analyzed based on what is *seen* and *heard* in the videos. That is, behaviours may not be accompanied by spoken descriptions of what is happening, and certain activities could be deemed too insignificant to describe in either the title or description of a video when it is uploaded. Therefore, what researchers seek in the videos cannot be directly queried, and identifying those that are relevant to the study requires human review.

3.3.2 Attempts to Achieve Efficiency and Quality

As we have seen in Section 3.2.1, researchers compensate for YouTube’s limited query features by executing many queries. Unfortunately, this process yields very many videos that are largely irrelevant. Further, videos do not have abstracts that researchers can rapidly *skim*. When their metadata fails to indicate the videos are ineligible, they must be viewed by researchers.

Watching a large number of videos to determine their eligibility can be slow work, so it is important that researchers adopt a filtering procedure that maximizes their efficiency.

[1] https://developers.google.com/search/docs/appearance/structured-data/video

Using a purposeful sampling strategy (Rotman & Preece 2010; Patton 2014) to guide their YouTube searches, Nielsen et al. (2023) collected and watched a total of 494 videos. After removing duplicates and unavailable videos, the researchers reduced this set to 104 videos after two separate rounds of filtering. First, they coarsely evaluated each video—quickly applying one or more labels to help them eliminate videos from future viewings. Next, the researchers watched the videos more closely to identify more subtle issues, such as whether the video appeared to be staged. Using this two-stage filtering strategy, they minimized how much time was spent watching irrelevant footage.

Of course, it would be ideal if researchers could reduce or eliminate the need to watch videos altogether. Mauriello et al. (2018a) developed a tool that automates both the expansion of the video collection and the identification of those that are most relevant. In a study employing this tool, Mauriello et al. (2018b) began with a collection of keywords that returned 1,092 unique videos from a number of (unreported) initial searches. Using query-expansion techniques, they performed more searches using additional keywords obtained from the titles and descriptions of the initial set, raising their total to 6,790 videos. After manually coding a subset (772) of the videos, the researchers trained a ML model to automatically classify the video's *relevance* and *topic* using a bag-of-words model of their titles and descriptions. This process led to a dataset of 1,686 unique videos from which they randomly sampled 1,000. Finally, they removed videos that were deemed off-topic to obtain a final total of 675 that were used in the study.

While the automated method developed by Mauriello et al. (2018a) can save a great deal of manual labour, it is unlikely to capture relevant videos for studies of human behaviours as it relies upon the video's metadata. In our own studies, the metadata field *rarely* matched the activities we sought to study. For example, some users filled the description with promotional materials such as links to a subscription, or directing viewers to purchase goods using affiliate links. Rarely did we encounter descriptions that characterized what could be observed in the videos.

Ideally, an automated tool would be adapted to take more features into account, perhaps obtained directly from audiovisual content. For example, if the phenomenon is associated with a specific object (e.g., a guitar is in frame), or spoken phrase (e.g., "learn by ear"), researchers could use object recognition and automatic transcriptions to extract additional details that could be queried. However, this approach has its limits—a researcher could surface videos where a guitar or piano appears in the frame, but cannot identify *how* or *if* the guitar is used.

However, even if researchers could overcome these limitations, it would be impractical to build such a tool. Object recognition and transcription would require direct access to the video files for hundreds or thousands of videos found on YouTube. Each of them must be downloaded, which would almost certainly violate YouTube's terms of service.

Setting that issue aside, it seems unlikely that AI-driven tools could answer more subtle questions, such as "Are the workers in this video using safe practices?", "Are they wearing PPE correctly?", or "Is the musician in the video learning music from a recording, or just playing alongside it?" Until science has caught up to provide such capabilities, manual viewing is still necessary. While we are skeptical that AI-driven content analysis can provide researchers with a *more searchable panopticon*, it is plausible that near-term advancements to large language models could power tools that make such queries possible. In the meantime, researchers must resign themselves to the human-powered process of reviewing large swaths of content as efficiently as they can.

3.3.3 My Own Such Attempts

An Application for Video Triage

In March 2023, I developed a prototype application for macOS called YouTubeTranscripts (YTT) to help me maintain a *paper trail* while querying and qualifying videos for study. While it was helpful during the early stages of my research projects, the tool is not suitable for distribution.

YTT is a document-based application. Each document represents a collection of queries and their results. The user adds queries to the document whenever they execute a YouTube search. The date, keywords, and parameters used to conduct each query are retained, which would allow a researcher to see a history of all the queries they have attempted, and how many results each of them returned.

Upon selecting a query, the user can *triage* its results by navigating videos in a master-detail arrangement. That is, selecting a video from the table will reveal additional metadata in a detail view with an embedded YouTube player. Based on the video's content, the user may choose to either *favourite* or *reject* the video, at which point a modal dialog prompts them to enter a reason for doing so. Only those videos marked as favourites are available for further analysis.

Video analysis consists of both watching the video and seeing its transcript at the same time. As the video progresses, the corresponding line of transcript is highlighted. Similarly, selecting a line of transcript will seek to that place in the video. Lines of transcript could then be converted into *annotations*, which retained the timestamp and transcript text and allowed the user to capture notes about those statements.

While this appears to comprise much of what a YouTube video researcher would need to obtain, deem eligible, and ultimately code their collection of videos, it is far from a complete solution in the state it now sits. In my own studies, YTT was only helpful insofar as it helped me evaluate the suitability of certain query terms, and allowed me to rapidly navigate through large lists of videos. The latter was afforded by the fact that YTT is a native macOS application developed using SwiftUI, and accessing YouTube via an embedded player—far more responsive compared to managing browser tabs, facing un-skippable advertisements, etc.

I built YTT with the help of third-party Python libraries that almost certainly violate YouTube's Terms of Service by accessing clandestine API endpoints. I did this out of necessity, because the default quota for the official API is much too restrictive to use during development or actual research. I could certainly integrate the official APIs to eliminate this concern. However, I would have to apply for additional quota as a researcher to get any utility out of the application, and it would also require that every other user does the same in order to use the application. For these reasons, I do not feel comfortable distributing YTT in its current form as an open-source project.

Ultimately, YTT was motivated by a lack of efficiency in the process of manually reviewing a large collection of videos, and the overhead that is added by maintaining a paper trail. Resolving these issues seem to be desiderata for any video-based researcher who works with YouTube, and I hope that one day I or another researcher can make this process require far less labour and organizational discipline than it does now.

Rapid Clip Analysis

In the hypothesis-generating study (Chapter 4), many of the videos in the collection contained a lot of irrelevant footage. For example, discussion about the song featured in the video, or interaction with a virtual audience during a live stream. Even though I took care to capture the timestamps of significant events in the videos, it eventually became too cumbersome to navigate between browser tabs to compare clips with one another. For example, when watching two clips from different videos that demonstrate an instrument being played while the recording is also playing, I could identify that one was for the purpose of seeking notes, and the other player was practicing what they had learned.

Having worked with Final Cut Pro[2] in the past, I had experience resolving a similar problem—isolating only the good takes among hours of raw footage. This was achieved

[2] https://www.apple.com/final-cut-pro/

using Final Cut's *keywords* feature[3] that allows (possibly overlapping) segments of video to be assigned one or more keywords. After a collection of videos are marked up in this way, the interface allows a user to filter and browse only those clips with specific keywords. For example, upon selecting the "playing-while-listening" keyword, I could rapidly review this specific behaviour across the entire set of videos.

While I found this helpful to complete my project, the experience left much to be desired from the perspective of a researcher. For example, entering longer keywords with spaces, and defining keywords that overlap others can exercise one's patience. An application that offers a user interface that takes inspiration from Final Cut Pro, but designed specifically for qualitative research, would certainly be welcome.

This strategy required that I first download each of the videos before study, which required the use of command-line tools that violate YouTube's terms of service. It is up to other researchers to determine whether obtaining the videos in this way is an ethical practice for their particular work. In my own situation, the videos are merely being *cached* in local storage until the study is concluded, and I have no intention to distribute these files.

3.3.4 Observations of YouTube Search Behaviour

On May 2, 2023, I performed an experiment using the results of queries obtained using the YouTube Data Tools[4] website. The test was simple: I ran the same query (`"learn songs by ear"|"learn music by ear"|"learn tunes by ear"`) five times using different orderings: relevance, date, rating, viewCount, and relevance again. For each search, the query string, date range (before Jan 1, 2023), and maximum number of videos (200) was held constant. All queries were performed within a 5 minute time span.

Despite requesting a maximum of 200 results, no list of videos was that large. In the queries that specified an ordering of date, rating, and viewCount, a collection of 68 videos were returned. However, the same query sorted by relevance returned 191 videos, then 187 when it was repeated five minutes later. Such discrepancies between the results of identical queries triggered further investigation.

Each of the 68-video lists contained the same set of video IDs. Both of the larger, relevance-ordered lists returned the same 68 videos before the rest of their results, but

[3] https://support.apple.com/en-ca/guide/final-cut-pro/ver68416335/mac

[4] https://tools.digitalmethods.net/netvizz/youtube/

their rankings were inconsistent between the two. For example, the third-ranked video moved up to the second position, and the three videos ranked 10–12th moved up to positions 8, 10, and 9, respectively. Additionally, 26 videos disappeared from the relevance-ordered results when it was repeated 5 minutes later. Upon reviewing this list of missing videos, they were ranked at position 90 and below, and all but two of them had titles that were relevant to the query.

The 68 videos common to all queries contained either “learn songs by ear” or “learn music by ear” in their title. However, the titles of videos appearing only in the relevance-ordered searches contained substrings such as “learning songs by ear”, “learn any song by ear”, etc. The inclusion of these possibly-relevant videos suggest that relevance-ordered searches on YouTube are capable of matching semantic meanings, and perhaps the date-, viewCount-, and rating-ordered queries apply more simplistic string matching to the titles of videos. However, I discovered that two videos in the relevance-ordered queries contained “learn songs by ear” in their title, yet they did not appear in the other sets. This evidence suggests that those queries returning 68 videos provided incomplete results.

In many studies, researchers perform searches directly on the website (e.g., Altunisik et al. 2022; Basch et al. 2020). To confirm that YouTube Data Tools was not behaving differently from a search performed on the YouTube website, I searched on May 2, 2023 using the same query string, and specified that results should be ordered by date. The search returned near-identical results to my prior tests—68 videos matched the set I had, plus an additional 10 that were uploaded *after* January 1, 2023. This discrepancy was caused by my inability to specify such a date range on the website to match my earlier results.

On May 4th, 2023—two days after my first experiments—I repeated the same queries using YouTube Data Tools. They were executed in the same order as the first experiment, but there were only three minutes separating the first and last relevance queries. The same set of 68 videos were returned in the date-, viewCount-, and rating-ordered queries as before, and two new sets of 197 videos each were returned by the two relevance queries. Again, both relevance-ordered queries ranked the same 68 videos before the rest.

My first video study (Chapter 4) was conducted using the entirety of the above data. That is, I formed the union set of all the above search results for a total of 255 videos that required screening. The screening process was time-consuming, and returned a final set of only 18 videos that were used in the study. Remarkably, all 18 appeared in the smallest set of 68 videos.

Based on the above observations, I formed the following six hypotheses. First, it appears that date-, rating-, and viewCount-ordered queries are quite stable. Second, the results from these queries rank highest in a relevance-ordered search. Third, the titles in the date-, rating-, and viewCount-ordered query results all match the query string. Fourth, those queries do not include all results with a match for the keyword in the title. Fifth, the relevance-ordered queries demonstrate instability between repeated trials. And sixth, the videos that appear only in relevance-ordered queries *are largely irrelevant*.

Recommendations

When conducting a study that is based upon the experience of ordinary people who use the YouTube website, researchers should use the default relevance ordering when searching for videos. This was the appropriate choice for studies like those focusing on the content of videos presented to consumers seeking healthcare information (e.g., Basch et al. 2015; Pathak et al. 2015; Basch et al. 2017, 2020).

If the researcher's aim is to find videos with titles that closely match a query string, a date-, rating-, or viewCount-ordered search appears to provide the most stable results. However, such a strategy is only useful when the subject matter allows for it. For example, our search for by-ear learning videos provided only 68 results, but search terms that return thousands of matches will get truncated. In such cases, the ordering should be selected carefully so that the results are not impacted. For example, a date-ordered search that gets truncated will omit older videos that may have high ratings, and possibly high-quality content.

Considering my observations, it may seem strange for a researcher to consider giving up more than 100 videos if they chose a date-ordered instead of a relevance-ordered query. However, after conducting our hypothesis-generating study (Chapter 4), which used the union set of all the relevance-ordered queries, each of the 18 videos obtained from our filtering process were all contained in the stable set of 68. Considering that we had to manually filter the list of 255 videos, sticking with the date-ordered query would have saved us a considerable amount of time.

3.4 Ethical Considerations

In-person studies require ethics review, which creates barriers between the would-be researcher and their study, and rightly so. However, this can be a major problem during preliminary work, when the methodology is not yet concrete, or researchers need to expand their domain knowledge based on real-world observations. For every change to the method, or as additional rounds of interviews or observations are required, the ethics review board must be consulted before the study can continue. This makes online video studies particularly attractive, as there is no need for the researchers to interact with those they study. However, researchers must understand that studying online video does not completely absolve them from their ethical responsibilities. Legewie & Nassauer (2018) provide a list of five ethical areas that the researchers should consider when assessing the ethics of their video-based studies: *informed consent*, *unique opportunities*, *privacy*, *transparency*, and *minimizing harm*. Here we summarize each of these, focusing specifically on those elements that are relevant to my own research, and those that relate to the previous work discussed above.

On the surface, it may appear that any video a researcher can watch on YouTube is available to use in their research because the uploader chose not to mark it as private. However, nobody that appeared in the video—including the uploader—consented to have their actions analyzed by researchers. Additionally, it is possible to encounter years-old videos that have not been widely viewed. For example, in Chapter 4 we use three videos that were 1–3 years old and had only 22-25 views. Such videos are effectively invisible to the public until a researcher calls attention to them, which is problematic if the uploader was unaware they were visible to the public, or they had intended to take them down. As with many ethical considerations, this lack of consent is not necessarily a dead end for the researcher. When weighed against the assessments made in the following areas, the use of certain videos like these can indeed be ethically acceptable.

Online video data allows researchers to observe events that would otherwise be impossible (or itself unethical) to capture in a lab environment. For example, Dao et al. (2021) cannot ethically ask participants in a VR headset to stand too close to a wall, hoping they walk into it. Similarly, Rea et al. (2018) cannot ask participants to drive dangerously close to moose, and Weenink et al. (2022b; 2022a) should not provoke fights in public. User-generated videos found online also give researchers an opportunity to observe real-world behaviour in a naturalistic setting, without influencing the actions of those performing them—e.g., in their homes, cars, or on public transit. Moreover, the financial cost to execute online video studies is often far less than the travel or remuneration required to do them in person, and thus more likely to be conducted.

Researchers should take extra care to respect the privacy of those they observe in videos, *especially* in the absence of informed consent. While researchers may simply omit the names or faces of those appearing in videos, it is prudent to also prevent those details from being easily obtained. For example, taking care to omit any information—video IDs, URLs, or titles—that can be easily traced back to those videos. A common practice in the literature only refers to videos based on researcher-assigned identifiers (e.g., Komkaite et al. 2019; Dao et al. 2021). In some cases, researchers may wish to exercise additional caution when including quotes from videos—particularly those dealing with sensitive subjects. For example, when Schuman et al. (2019) could not obtain consent to publish certain quotes from war veterans, the researchers took care to ensure that YouTube searches for those quotes did not reveal the videos in the first 20 results.

Naturally, researchers should be prepared to share their data in its raw form with those peers who wish to either verify or replicate their work. This can be particularly challenging in the face of copyright, and the terms of service for sites like YouTube. However, sometimes it is necessary to download videos locally to help facilitate their analysis. To use an example from our first study (Chapter 4), local copies of each video let me review their footage more efficiently, and in the second (Chapter 5), local copies allowed me to obtain higher quality automated transcriptions. In cases like these, video files should be omitted from the data set that is shared with others. Of course, there is a risk that one or more videos have been removed from YouTube, or made unavailable for viewing in certain geographic regions. Researchers should therefore consider whether such a potential scenario could threaten the validity of their work, and take reasonable measures.

It is most important that researchers ensure their work does not inflict any kind of harm upon those people they are studying. Of course, withholding their identity from publication is a good place to start. However, when topics are particularly sensitive (e.g., religious or political beliefs, sexuality, gender identity, trauma, or disability) and people who appear in videos could face any kind of harm (e.g., loss of income, disruption of family life, or physical violence) upon having their identity revealed, then researchers should exercise *additional* caution when sharing video data with other researchers. For example, anonymizing and/or redacting transcript data, or producing a set of video files that are modified to conceal peoples' identities.

As with all ethical considerations, researchers must exercise judgement in all of the above areas before deciding to proceed with a study. However, researchers may not value all concerns equally, and the importance of each varies depending on the subject matter of the research. For example, one may choose to prioritize transparency over privacy when they are studying the behaviour of celebrities, or the videos have tens of millions of views. In contrast, a researcher investigating insurance fraud may have accidentally discovered whistle-blowers, and place more emphasis on obtaining informed consent, and protecting their identity to avoid retaliation.

Whether or not these decisions are clear, researchers should still consider discussing their concerns with members of an ethics review board. While many studies using YouTube videos omit any discussion about ethics approval (Sampson et al. 2013), we encountered those where an IRB deemed approval was unnecessary (e.g., Basch et al. 2017; Kelly-Hedrick et al. 2018), and others where approval was obtained (e.g., Harrison et al. 2014; Borgos-Rodriguez et al. 2019; Schuman et al. 2019).

3.5 Summary

The wide use of online video in research is a testament to its value as a source of data. It provides benefits such as access to a massive library of videos, the opportunity to observe real-world phenomena, and the potential for researchers to study populations that may otherwise be unreachable. However, there are also limits to what can be achieved in a study that relies upon services such as YouTube.

YouTube is certainly not the only place that video is created or watched online; TikTok, Snapchat, Twitch, Facebook, and Twitter all offer significant video features, and researchers have certainly studied the content found on such platforms (see, e.g., Bartolome & Niu 2023). However, YouTube's use in research for analyzing video content appears most prevalent. It offers a mechanism to query and view its content without an account, and also a set of APIs that can do the same.[5] Hence, it is currently the most convenient place for researchers to gather and analyze the content of videos.

For example, YouTube's query mechanism is opaque, does not return consistent results, and appears not to search beyond a video's uploader-provided metadata. These technical challenges can lead to researchers struggling to locate videos that are truly relevant among large collections, unless they take measures to find efficiencies in manual labour earlier in the data collection pipeline.

Despite their convenience, online videos are not an appropriate source of data for all studies. Researchers must not only consider whether their research questions can be answered using a collection of videos obtained online, but also whether it is ethical to do so.

[5] Twitch also offers an API (https://dev.twitch.tv/docs/api/reference/), though it appears to be oriented toward searching for streamers, or specific games that appear in videos.

Chapter 4

A Hypothesis-Generating Study of Musicians Learning by Ear

In this preliminary study, we set out to identify opportunities for the designers of purpose-built technology to improve upon the process of by-ear learning. Our findings are based on a content analysis of 18 videos that were collected from YouTube, each depicting real-world examples of musicians learning from recordings by ear.

We discovered a number of remarkable elements in this set of videos that will drive further inquiry, such as the different strategies used by musicians to retain notes in memory while playing by ear—a topic we explore further in the following chapters. Additionally, we demonstrate that such a study of expert practitioners can be conducted without introducing a financial cost, and without the need for ethics approval. Our methodology provides a model for researchers and designers to gather enough information that would help them decide whether to proceed with larger, more costly studies. Further, it has the potential to shape such future work in ways that can improve the researcher's effectiveness.

4.1 Study Goals and Approach

We wish to understand how modern, experienced popular musicians interact with recordings as they learn music by ear. Specifically, we would like to know more about the kinds of technology they use today, how they control recordings as they learn, their strategies for reproducing the notes and sounds they hear on their instrument, and how they work towards playing the song themselves. Based on these insights, we hope to identify opportunities to help these musicians improve upon their process.

We are interested primarily in musicians with experience learning by ear. They already have a set of strategies that afford them the ability to expand their repertoire in this way. We care less about whether these musicians are *professionals*, because learning music by ear may not be tied to their source of income. For example, popular musicians who make a living from their performances may lack these capabilities, only performing music they wrote, and having little desire to play other people's songs. Similarly, a musician that plays for their own entertainment may have mastered the skill of song acquisition, and ritually learn new songs that they like shortly after release.

Further, we choose to focus on *instrumentalists*, not vocalists. This decision is largely pragmatic. In our filtering methodology, we use the presence of an instrument to help us rapidly select eligible videos. Admitting singers into the study would require in-depth viewing at an early stage that would make the process far less tractable.

Much literature about popular musicians learning by ear comes from before the ubiquitous availability of technology playing an endless supply of music. We argue that to construct a theoretical frame from this material, upon which we would then build and test hypotheses, would be disingenuous. Also, existing studies on popular musicians are weighted heavily towards rock and blues players that learned by ear during the 70s and 80s. Further, many of these studies were based upon interview responses, where researchers sometimes reported difficulty getting musicians to recognize the significance of the activity, and describe their process (Green 2017; Bennett 1980). When studies aimed to observe by-ear learning strategies, they did not choose to study those with experience, or—in the case of Johansson (2004)—when he recruited musicians with experience, they were asked to perform the (rather unusual) task of learning as they heard a recording for the first time.

Thus, we wanted to execute our study with few preconceived ideas about the way musicians learn by ear, and form hypotheses by identifying notable phenomena while observing people doing it. Our study's design was partly influenced by that of Rueben et al. (2021), who similarly lacked a theoretical frame, and conducted a hypothesis-generating study to understand how participants formed mental models of a robot's behaviour. Where our study differs most notably is that we draw upon observations of YouTube videos for which we have no control over content, and no guarantees that relevant examples can even be obtained.

To conduct an in-person study of experienced musicians learning by ear would be highly impractical. Since musicians typically learn by ear in private, we would have to bear the expense of traveling to either meet them where they practice their craft, or have them visit us on-site. We could possibly reduce costs by selecting a small number of cities that are known for having a large and diverse music community. However, we would surely face challenges trying to situate ourselves in the practice spaces of musicians while also adequately recording their activities, similar to what Paay et al. (2015) recognized about most home kitchens. To invite musicians into a lab would require that we ask them to replicate the environment in which they choose to work: bring their instrument(s), and other elements that they need to learn successfully.

We could mitigate these difficulties by asking musicians to participate remotely, but this presents additional challenges. Assuming we find willing participants that are open to sharing this private activity with us, they must also have a certain set of skills and equipment to ensure that we can capture their learning adequately. While cell phone cameras are ubiquitous, and the population has grown more comfortable with videoconferencing software since the COVID-19 pandemic, these sessions are still fraught with technical issues that—if they don't put an end to the session—could negatively impact the participant's demeanour.

Participants who publish videos of themselves learning music by ear typically have attained a baseline level of competency that suits our study, and we can exclude those who do not. This self-selecting nature of our study population is likely to produce the same results as recruiting participants that: (1) have established by-ear learning strategies, (2) can film themselves performing the task, and (3) are able to clearly demonstrate the process while possibly also explaining their actions clearly. A deficiency in any one of these elements could disqualify participants from either of a remote-participant or video study like ours.

4.2 Method

4.2.1 Video Collection

Using the YouTube Data Tools[1] website to perform our queries, we combined the results from 5 searches executed between May 2–5, 2023. For each search, the query string (`"learn songs by ear"|"learn music by ear"|"learn tunes by ear''`), date range (prior to January 1, 2023), result ordering (by relevance) and maximum number of videos (200) was held constant. These query results were obtained from an experiment

[1] https://tools.digitalmethods.net/netvizz/youtube/

that demonstrates YouTube's inconsistent relevance-ordered results (Section 3.3.4), and merged to produce a collection of 255 unique videos.

Conceptually, we treat these videos as a *sample* of the corpus available on YouTube, and do not intend to draw any parallels to a systematic review. We view this collection of videos as analogous to a response to a call for participants—just as recruitment may yield a number of inappropriate or unqualified interviewees, the videos require scrutiny before we decide to include them in the study.

4.2.2 Video Selection

We used a filtering approach inspired by that of Nielsen et al.'s (2023) study of unguided human-robot interactions in public places. Specifically, we selected relevant videos by applying high-level labels to each video after briefly reviewing their content, and retained only the ones depicting genuine instances of learning by ear. We rejected many videos in seconds: if we failed to identify an instrument while scrubbing the timeline and reviewing video thumbnails, the video was eliminated. For example, if the video contained only a talking head or graphical slideshow, but the content still seemed relevant to by-ear learning, the video would be categorized as *describing-not-doing*, and thus rejected. Such efficiencies helped make this video study tractable.

When we encountered videos depicting an instrument in the hands of a musician, they got slightly more scrutiny—we sampled brief segments of those videos to assess whether the player was *legitimately learning* the material in an audio recording, or merely giving a *prepared lesson*. For example, one video contained only hypothetical examples based on nursery rhymes, and the presenters acted out the process of finding notes on their instrument.

168 of the videos in the set were uploaded to the same channel, and largely depicted musical performances or comedic content. The musical performances were given by a solo pianist, but the comedic videos were entirely unrelated: they featured animated musical performances from popular movies and TV shows, with the original soundtrack replaced by over-dubbing the actual notes that would be sounded if the animated character struck the notes that appeared to be played. Fortunately, these two categories of video from this channel used a consistent title scheme that allowed us to apply labels *en masse* based on the video's metadata. One video from this uploader claimed to demonstrate how they learn by ear, however it was a six hour long livestream. While sampling short intervals of this video, we found instances of the player taking requests and performing

for their audience, and the video was excluded.[2]

Three of the videos in the collection were segments from a larger transcription session, and we decided to exclude the second and third parts of the video. This move echoes a strategy we found among the findings from Sampson et al.'s (2013) systematic literature review of YouTube studies: omitting all but the first in a multi-part series.

This filtering process yielded a total of 18 videos for further analysis, which we labelled V1 through V18.

4.2.3 Video Analysis

We watched each of the videos in its entirety as we took notes with varying levels of detail. Our goal was not to transcribe the videos. Rather, we wrote a mixture of high-level summaries, timestamped quotations, and brief descriptions of notable events from each video that could be compared with others, and easily revisited as common themes developed. Videos often contained stretches of irrelevant content, and it was more important to capture the time ranges that depicted the kinds of activities we wanted to analyze. For example, we marked periods where the musician interacted with a recording, and not those where viewers are asked to "like and subscribe".

We met regularly to discuss remarkable findings that emerged from the videos, and those we deemed worthy would trigger further study. Videos were reviewed over the time ranges relevant to the phenomena, paying close attention to different details with each viewing. For example, once it was deemed significant that musicians often sang melodies, we would re-watch those videos, using our notes to direct us to relevant segments. Then, we looked more closely to identify whether they sung alongside the recording, after stopping playback, or while identifying notes on their instrument.

Late in the study, these reviews became more frequent as we continued to refine our findings, and the process became more difficult to maintain. It is at this point that we decided to download all 18 videos to local storage and review their footage in Final Cut Pro (Section 3.3.3).

[2] Later in the study, we discovered this video contained legitimate segments of learning that we missed. Had it passed the initial screening, we would have likely excluded it based on its six-hour duration.

4.2.4 Summary

The above method of querying, filtering, and analysis of the videos' content is largely tailored for use in a preliminary, hypothesis-generating video study. Our query strategy is less comprehensive compared to other HCI studies that combine sets of keywords to generate a large set of search terms in an attempt to maximize the number of relevant videos retrieved (e.g., Anthony et al. 2013; Wentzel et al. 2022; Vatavu et al. 2022). However, breadth was less of a concern for us, because we were not interested in "drinking from the firehose" at this early stage of our research—we did not feel ready to process an overwhelming amount of video material just yet.

4.3 Overview of the Videos

The 18 videos we chose to study had durations that ranged from a minute and 31 seconds to over an hour and 14 minutes. The average duration of the videos was approximately 22 minutes, and half were less than 15 minutes long. In sum, the total viewing time of the videos was more than six hours and 38 minutes. According to the metadata, videos were uploaded to YouTube between November 4, 2017 and November 18, 2022.

Two videos depicted saxophonists, two depicted pianists, and the rest depicted guitarists. All 18 videos in our collection depicted perceptibly male presenters.

Overall, these videos failed to garner a large audience. The most-watched video had 28,923 views, and the median view count was only 541. To put these numbers in perspective, a video from the original set of 255 was viewed more than 8 million times. Despite the low viewership, we were surprised to discover videos that had only a few dozen views, yet carried some of the most valuable footage.

While viewing these videos, it often felt as though we were watching responses that were submitted to us in a participant study. Had we requested our participants to film themselves learning a piece of music by ear while talking us through their process, we would expect to obtain a set of videos like many of those we collected from YouTube. However, not all the videos were as transparent and *raw* as we would like.

For example, two of the videos—V6 and V8—came from the same source in *livestream* format, where the guitarist interacted with his audience via text chat. While this video appeared to be unedited, and the musician in the videos was clearly very adept at learning songs by ear, their behaviour in the video was clearly influenced by the virtual presence of the audience. For example, they learned songs that were requested by viewers, and at one point they got stuck on a chord and exclaimed, “Gosh darn it. This is what I was worried about. Now I’m going to be stumped here.” This may have been an expression of embarrassment and/or discomfort with struggling in front of others, and lends credence to Bennett’s assessment regarding the desire for privacy during this learning activity (Bennett 1980).

We also encountered some videos that were edited heavily, such as V4. It was only 91 seconds in duration, and depicted short segments of the process of ear learning. Despite its brief presentation, the footage that remained in the video demonstrated genuine learning: we observed the player making mistakes as they experimented until the correct notes were found. In other videos, editing was not used to produce succinct content. Rather, it was used to intersperse footage from the presenter’s computer screen recordings with their demonstrations.

4.4 Results

Here, we identify patterns common to many of the videos, and discuss how these give rise to hypotheses for further analysis.

4.4.1 Scope of Learning

In only three of the videos did we observe musicians working to learn the entirety of a song. However, only V2 provides evidence that they did so successfully—the guitarist includes their performance of the whole song at the end of the video. In the rest of the videos, musicians learned only *portions* of songs—solos, riffs, or a subset of the chords. For many popular songs, the repetition in subsequent verses and choruses means that learning one is often sufficient to know how the others are played. The guitarist in V1 helps to explain this before closing their video:

> There’s basically only 3 parts to the entire song. There’s the intro; a little fast riff [*singing*], that leads us right into the verse. The verse always goes twice. Um, chords in the verse: D, G, E minor, A, D, happens again. Then we move into that second part where we play a B major chord into F sharp minor. That happens twice before we hit A and then we hang in on E before hitting the intro again and that leads us back into the verse again, right? That’s how I do it. Rather than thinking of every chord all the time, I think about what chords

> are in the part and then I group them in my mind and like spread it out into different sections.

This repetitive tendency of popular music makes it such that learning the core skill of finding chords, or individual notes by ear, allows any musician—with sufficient time and effort—to learn the entirety of *any* song by learning the first instance of each repeated section. This could explain why so many musicians chose to include such short segments of learning: it may be the case that their interactions within one section of a recording are highly intensive in the beginning, while subsequent repetitions of those sections may be learned more quickly, or skipped altogether. Some musicians may only care to learn segments of songs that carry some degree of novelty. Of course, it is also plausible that these decisions were necessitated by the medium—YouTube limits uploads to 15 minutes in duration for unverified accounts[3], and succinct videos may attract viewers with shrinking attention spans.

This led us to wonder whether there is a not a universal set of recording interactions, but rather ones that are specific to the musician's scope of learning. For instance, a musician learning a solo might reach for tools that allow for manipulating the playback speed, whereas a musician interested in chords or other high-level structural elements might not require such features. However, our videos contain evidence to the contrary. The musician in V16 was learning a guitar solo within a pop song, and the guitarist in V17 was learning to play *fingerstyle* chords on their acoustic guitar to match the recording. However, it was the guitarist in V17 that chose to slow, and loop the playback—the player learning the guitar solo could do so at full speed. This suggests that a musician's need to slow playback is not exclusive to those players learning solos.

We identified two avenues of inquiry that should be considered for future study. First, we would like to further understand this dichotomy between those musicians who strive to learn only segments, and those who work towards playing the entirety of a song. Second, we think that those who wish to learn songs as a whole may benefit from a more structure-oriented graphical representation of a recording—one that exploits the repetition found in many popular songs to aid learning, memory, and recall.

4.4.2 Transcription and the Role of Notation

In three of the videos, musicians transcribed the notes from the recording to generate sheet music for the songs as they learned them. Unlike the process of transcribing speech,

[3] https://support.google.com/youtube/answer/71673

wherein words are recorded as they are recognized, the musicians in these videos did not enter the notes they heard until they were located on their own instrument.

The musicians recorded notation onto a staff (V18), or as guitar tablature (V2 and V17) using software designed for those purposes. To verify the correctness of their sheet music, the presenters in these videos did not sight-read what they entered. Instead, they used the software's built-in synthesizer to perform their sheet music virtually so they could assess whether the notes were representative of the original recording. This evidence suggests that while by-ear learning is necessary to produce notation, the converse may not be true.

In V17 and V18, the stated goal of the videos was to demonstrate the transcription process. However, it is unclear why they felt motivated to create sheet music. Only the guitarist in V2 states explicitly the role that transcription plays in their own process: "The reason [you generate tablature] is that you can accurately learn even complicated rhythms. Another reason is to remember." That is, this guitarist claims the notation helps them reason about rhythm patterns, and also serves as a memory aid. However, despite this claim about using tablature to learn rhythm accurately, the guitarist contradicts this by learning the song's (quite rhythmically complex) solo later in the video without appearing to enter or use any tablature.

In the 15 remaining videos, musicians do not produce any notation using software, or by placing marks onto paper, but they demonstrate the same set of skills that we see in the videos about transcription. It is apparent that musicians can learn music by ear without writing anything down, and producing notation is not necessary to learn music by ear.

From these observations we identified some ripe opportunities for future work to develop an understanding of the role that notation plays for musicians as they learn by ear. For example, it seems obvious that recording notes as tablature or on a staff—either digitally, or on paper—serves the role of a memory aid or transmission mechanism for what was learned by ear, though it does not seem that this notation facilitates the ear learning process. Further, it would appear that those who rely upon notation for later recall may benefit from technological tools that can store and display notation alongside the recordings they are learning.

4.4.3 Use of Technology

It was not always clear what technology the musicians used to play and interact with recordings, but in eight of the eighteen videos we could observe its use. In some, the creator of the video included a screen capture; the rest merely pointed a camera at the screen. In all the videos depicting technology, none included purpose-built hardware devices—only software running on a smartphone or computer. Among this set of eight videos, musicians used YouTube for music playback in three of them. Three other videos used non-specialized players—Music (on the iPhone), Spotify, and iTunes (on a Mac). In one of the remaining videos, the musician used Digital Audio Workstation (DAW) software. Finally, we discovered a single video that featured purpose-built software—Transcribe![4], running on a Mac—though the musician used it to play audiovisual content (i.e. a video). For this as well as the videos where YouTube was used, we only considered the musician's interactions with the audio component of the media. Because there were so few videos that presented the use of technology visually, we used a more general approach: rather than focusing on specific software or hardware, we instead looked for evidence that suggested the musicians were using specialized features. To do this, we observed how the musicians interacted with the recordings by watching their actions and listening to both the audio playback and their narration as they learned their parts.

Because so few videos placed technology prominently in the frame, we tried to identify purpose-built features using a combination of body language, dialogue, and apparent changes in audio playback. For example, in V12 the saxophonist's shoulder raises slightly before stating YouTube was used to slow playback, we hear the music start, and his shoulder lowers again. This sequence of *turns* in the video allows us to infer slowing was used, and that the saxophone player controlled the playback event (Knoblauch et al. 2014).

Given that the majority of the musicians from the videos we studied do not employ features from purpose-built technology, it would appear that they are no better off than their counterparts were more than 40 years ago. In the 14 videos that did not contain special-purpose technology, these thirteen musicians[5] listened to the recordings at full speed, and they were content to repeat passages with little precision. This group of musicians could be handed a record, cassette, or CD player with the same music they learned in their videos, and—provided they knew how to operate the equipment—the act of learning the music by ear would look very much the same.

With so few occurrences of purpose-built technology use, we are left wondering whether such technology is helpful, or in broad use among experienced players. However, it certainly sparks the need for further inquiry. Additionally, we question whether experienced musicians feel no need for their features once they become proficient, or are simply unaware of them.

[4] https://seventhstring.com

[5] The same musician appears in V6 and V8.

4.4.4 Temporary Note Retention

When learning music by ear, musicians must first recall one or more of the notes from a recording before they can be repeated. Similar to working memory, which lets us carry a phone number or similarly small piece of data before it is utilized, it seems that musicians would have to retain a string of pitches for some duration before they could repeat them on their instrument. We observed three different strategies—often used in combination—that musicians deployed to briefly remember one or more notes.

In eleven of the videos, musicians played their instrument while listening to the recording at the same time. Many of them did this to verify the correctness of the notes they learned using other strategies. However, as a note-finding strategy itself, musicians played atop the recording to hunt for *anchoring notes*—to identify the chords, or key of the song. For example, the saxophonist in V12 plays scales over the recording of the solo he is learning to find the one that suits it best. In those instances where the musician appeared to locate the notes of interest as the recording played, it came after a delay—the player was repeating phrases heard moments before, and was effectively confirming correctness. That is, they could use their *mind's ear* (discussed below), only without stopping the recording so frequently.

Eight of the videos provided us with examples where musicians would sing (or hum) the notes they hear in the recording. Some of the musicians in this group continued to sing these notes as they look for the same pitches on their instrument, but others appear to repeat the notes to simply hold them in memory—just as one might recite a phone number. The musicians that sing while *hunting* for the correct note appeared to create audible *goal tones* they used for comparison as they narrowed the error between their instrument and their voice. In contrast, the other musicians alternated between their singing and playing, implying that these goal tones are "heard" elsewhere.

Ten of our videos—some in common with the aforementioned subset—contain examples of musicians that demonstrate their exclusive use of *the mind's ear* (Covington 2005). That is, these musicians could listen to the recording, retain the notes of interest in their mind, and then locate them on the instrument without supplementary audible feedback. In the case of saxophone players, their breath is used exclusively on the instrument: these musicians cannot produce sounds simultaneously humming and playing notes on the saxophone until they match up. This practice is demonstrated in V12, where a saxophonist repeatedly sings the melody they are trying to learn—both alongside the recording, and after it is stopped—in an attempt to internalize those notes before attempting to play them on their saxophone. Again, this is much like what Covington saw:

> [. . .] performers consistently spoke of mental hearing as preceding what emerged from an instrument. That is, mental hearing is recognized as being more accurate than performing and needs to direct the actual performance. Working out phrasing and fine-tuning one's acuity for pitch need to occur in one's mental ear. The inner ear provides the leadership for performing—mental hearing is the leader for the next note, the dynamic shape of a passage, and intonation. (Covington 2005)

When musicians used humming or singing as a part of their method, their voice plays the role of another instrument on which they can play notes more readily. That is, learning to sing the notes they hear is an exercise in ear learning that comes more naturally to those musicians. It might be the case that more experienced musicians shed this intermediate step once they develop the ability to play notes with their instrument as naturally as they can hum them.

Considering these findings, we wish to understand how memory plays a role while learning by ear. We expect that people with working memory limitations, or limited experience, may need to work differently as they learn by ear. Perhaps they can learn few notes at once, and would benefit from technological supports: the ability to restart playback from a specific note, or work in shorter segments. We should also consider the needs of wind and brass instrumentalists with an under-developed mind's ear. For example, offering repetitive playback of phrases, or continuously sounding individual notes. Finally, those with a highly-developed melodic memory may benefit from technology that exploits this ability—allowing musicians to navigate recordings in musically relevant, bite-sized chunks.

4.4.5 Familiarity with the Music

In eight of the videos we analyzed, the musicians made explicit claims to suggest that their on-camera attempt to learn the recording was their first experience doing so. The guitarist in V11 stated this fact emphatically: "I can promise you I have never heard this song before." Many of these videos were filmed to satisfy requests—from friends, or their online audience—for the musicians to demonstrate their own process for learning songs by ear. It appears that deliberately choosing an unfamiliar song helps the musician convey their intention to portray a *genuine attempt*. However, it is not clear whether this ability to learn a song *cold* (i.e. without hearing it in advance) is a necessary skill for musicians that learn by ear, and why the musicians in these videos chose to approach their demonstration in this way.

Despite making no claims about their unfamiliarity with the recording, the guitarist in V1 provides us with a helpful anecdote that may explain the need for this skill: they were once asked to substitute for a lead guitarist on short notice, and had to learn another band's material to prepare for four performances over two days. Similarly, the guitarist in V15 explained that they often need to learn on the spot when asked by students to teach them songs the guitarist did not know how to play. These anecdotes, combined with the explicit statements in eight of the videos, suggests that learning unfamiliar material is indeed an important skill for some musicians to develop.

While eight of the musicians lend evidence to suggest the need for learning new material *quickly*, it appears that those who are most familiar with the music seem to struggle less while learning it. For example, the guitarist in V14 explains that—despite learning a song that was new to them—they listened repeatedly to the recording before attempting to learn it. That is, they intentionally developed a familiarity with the song first. In doing so, the guitarist appears to have constructed enough of an aural image in their mind that they could demonstrate the song's main riff before starting their concerted effort to learn from the recording. They claimed that they could visualize themselves playing the riff on the fretboard as they were listening, which supports the idea that this guitarist has considerable skill with their instrument.

Even though specific recordings may not have been familiar to the musicians, it was often apparent that most songs in the videos fell within a collection of music they were already well acquainted with. For example, the guitarist in V10 claimed that "I don't even know this one" while listening to one of the songs they attempted to learn in this video, but later suggested there were others by the same artist that they already knew. In the video, this example was one among a set of country songs that he could play himself rather quickly after hearing them. Similarly, in V14 we see a rock guitarist learning a rock song, and in V2 a metal guitarist learning metal.

The guitarist in V7 was clearly familiar with jazz music, though in the video they were trying to identify jazz chords that were originally played on a piano. Unlike the players in V2, V10, and V14, who could readily draw upon guitar-oriented idioms, the guitarist in V7 had to discover the *voicing* on the guitar that best represented what the pianist played in the recording. They worked more slowly because they had not yet developed the required *finger routes*—those shapes and scales that get programmed into the player's brain, and sets the frame of what they can play (Lilliestam 1996). This evidence suggests that familiarity with both the genre and instrument being copied contribute to the ease of learning.

We feel that it would be useful to explore further how one's familiarity with a song impacts the learning experience, and whether the practice of repeated, intentional listenings of a recording beforehand should be prerequisites for the learning process. Additionally, one's familiarity could be exploited in a technological tool that hopes to encourage one's learning by ear—for example, highlighting songs that have a high play count in the user's music library. Although we did not encounter an example of a "fish out of water" learning to play music entirely outside their most familiar genres, performed on different instruments, we wonder whether that is a function of one's limited taste, or perhaps this mismatch adds some level of difficulty to the task that we are unaware of. Future work could certainly explore this further.

4.4.6 Application of Music Theory

Among the set of videos, eight guitarists and one pianist set out to identify the names of chords in the song they were learning. That is, they stated both the root of the chord, and whether it was major or minor (including any extensions or inversions, if applicable.) While naming chords may appear to demonstrate some knowledge of music theory—albeit at a basic level—we instead wish to think about how the musician deploys this knowledge to their advantage.

To give an illustrative example, we compare the chord finding approaches of the guitarists in V6 and V13. In V13, the guitarist starts by identifying the bass notes in the recording, and auditions both major and minor variations of possible chords to identify which one sounds correct, using a trial and error approach. In contrast, the guitarist in V6 first identifies the key of F major, then refers to chords numerically ("the four chord", "the two chord"). By knowing the diatonic chords (i.e. those that occur in the key), the musician can immediately determine whether a chord is major or minor by the position of the bass note in the key. Further, this guitarist can also draw upon a vocabulary of common chord progressions (e.g., vi-ii-V-I, I-IV-V), and anticipate what chords came after those they just identified. The strategies we observed resemble some of those identified by Johansson (2004), though in his study they were divided based on whether they were deployed while listening, or playing.

There appears to be a similar split among the musicians learning individual notes—those who start by determining the key or scale of the recording, and those who instead *hunt* for individual notes on their instrument. The guitarist in V11 describes the latter strategy as follows:

> [. . .] I'll just find a nearby note [*plays note*]. If that note sounds too high, I'll

> go down [*plays note on adjacent fret*]. If it sounds too low—say I hit this note first [*plays note one fret lower than the first*]—then I'll go up.

Interestingly, those musicians who tried to identify the key in the videos also used similar techniques to do so. For example, the saxophone player in V12 suggested they also employed a trial and error approach:

> So I identified by playing through different notes like in church—when I used to play keyboard in church—I would start picking random notes and would start going chromatically up and down until one of those notes stood out to me. And if it did stand out to me, then I would test the scale. So, D stood out to me, so I tested the D minor pentatonic because it sounds bluesy. And, I saw that all the notes worked. So there's a good chance that this song is in D when learning by ear.

Another strategy that we observed for locating the song's key made reference to a song's *home base*—a chord or "note that feels like home", as described by the guitarist in V15. Rather than auditioning the whole scale to hear what fits best, the musicians deploying this strategy would instead look for a single chord or note that sounded like the root of a given scale, then they worked through a limited number of possibilities to locate it.

The key difference between those who found the key before learning individual notes and those who didn't is that once they acquired the key, subsequent experimentation seemed to disappear. For example, the D minor pentatonic scale reduces the musician's *search space* significantly to only 5 out of the 12 chromatic tones in an octave, which undoubtedly helps speed up the process. We observe this benefit in V12 when, after hearing and singing a phrase from the song, the saxophonist could replicate it immediately on the first try.

Generally, those musicians that could apply their knowledge of music theory appeared to struggle less with the acquisition of chords and notes. This certainly deserves further study, especially when we consider the possible implications for the designers of purpose-built technology. However, we wonder whether this apparent grasp of theory is instead a proxy for the proficiency of the player. That is, a saxophone player who has practiced a scale hundreds of times has not only learned its name, but has the muscle memory to play it with ease, and perhaps recognizes its intervals. Similarly, a guitarist who has developed a sizeable repertoire of popular songs has played the most common chord progressions repeatedly, and hence can anticipate—or hear—certain sequences of chords. Therefore, it seems unlikely that developing one's understanding of music theory independently from its application on an instrument would improve their ability to learn how to play it themselves.

4.5 Discussion

4.5.1 Limitations of Our Study

As stated earlier, the majority of musicians only shared footage of themselves learning short portions of songs. This tendency to provide a piecemeal presentation of the process makes it difficult for us to observe the strategies musicians use to work out entire pieces of music. Moreover, we only had one video that contained a performance of the song learned by ear. That means we cannot gauge whether the strategies we observed in the videos were actually effective.

While the self-selecting nature of musicians posting to YouTube allowed us to study those with some baseline level of competency, we risk collecting a set of examples that are largely performative in nature. For example, we have no way to verify that a musician has not heard a recording before filming their video, or that their struggle to locate notes on their instrument is authentic.

The videos we studied all contained perceptibly male musicians, which is especially unfortunate for a modern study. While we are aware of videos on YouTube that feature perceptibly female musicians learning by ear, our queries and filtering strategy failed to capture them for this study. In our follow-up study of lesson videos (Chapter 5), we include three such musicians, though they were still vastly under-represented. We hope to rectify this imbalance by taking concrete steps in future work to increase representation across a more diverse set of gender identities.

We intentionally used a simple query methodology for this study that resulted in a small data set, and acknowledge that it would have taken little effort to multiply its size—using snowball sampling, additional query strings, and other more sophisticated techniques. However, given the nature of our study—generating hypotheses, and not presenting results based on our data—we feel this possible oversight should be forgiven. This was a preliminary study, which required a great deal of effort to categorize the videos, then generate a collection of observation data. Given the wide range of methods used in this aspect of other YouTube video studies, and experimental results that challenge researchers' claims of systematicity, we felt it was necessary to save the design of a more comprehensive strategy for future work.

4.5.2 Future Work

For us to test whether the desired set of interactions changes depending on the musician's scope of learning (i.e. how much of a song to learn), we would have to recruit musicians with experience learning entire songs, melodies, and instrumental solos by ear. Such a study should present all participants with suitable (to their instrument, genre) recorded material to learn from. Ideally, this would occur over two phases: one where participants learn using their preferred tools, and one where researchers supply a tool that has a wide range of human-recording interactions. In the latter configuration, the participants would be provided with sufficient training so that the entirety of the interactions are accessible to them. However, the former is necessary so that we can control for existing habits that may drive participants toward features they have existing familiarity with.

A similar study could test whether musicians learning entire songs would benefit from structure-oriented graphical representations. For example, recordings could be presented with one of three modalities: a simple timeline, a timeline rendered using a waveform representation, and an interface that presents the measures and sections. Each would provide navigation that is appropriate to the interface, such as tapping on a numbered measure to begin playback from that point.

Testing whether transcription plays a role in the by-ear learning process would require either an in-person or remote study where participants are divided into two groups: those who are free to transcribe as they learn a piece by ear, and those who are asked not to record anything onto paper. Because transcriptions vary based on one's chosen instrument, and the detail of the material they are learning, segmentation of these results could be particularly interesting. For example, pianists may be more comfortable sight-reading, and may be more apt to rely upon transcriptions compared to guitarists.

To better understand the proliferation of purpose-built technology products among those who learn by ear, it seems that a survey would provide us with sufficient insight. We not only wish to enumerate the use of such technology among musicians, but also gather additional information about how their usage evolves as their learning skills improve. Ideally, this would be a carried out as a longitudinal study that follows early-intermediate instrumentalists to see how their opinions shift. However, a single survey could gather additional valuable information such as one's frequency of by-ear learning, and the scenarios in which they must employ these skills.

In order to gauge technological supports for musicians that have not yet developed the working memory required to learn longer-running musical phrases, we would need a mixture of participants with varying levels of experience, and also a way to gauge their working memory with respect to one another. What we are hoping to find is that one of a number of provided tools causes those with (measured) limited working memory to learn more efficiently. A difficulty with such a study is accounting for neurodivergent participants—e.g., the limited working memory of those with ADHD (Vassileva et al. 2001).

To explore the impact of one's level of familiarity with a piece of music before learning it by ear, a rather simple experiment could supply participants with a previously-unheard recording of music, and let them hear it an increasing number times before asking them to learn it themselves. Not only would we measure the duration of the learning session—expecting that to negatively correlate with the number of listenings— but also observe how the musician's interactions with the recording change.

Finally, we wish to discover more about the role that a musician's grasp of music theory plays as they learn by ear. It seems unlikely that we would find many experienced musicians who completely lack theoretical knowledge, because they may well develop some baseline vocabulary as they grow more comfortable playing their instrument—especially if they do so with others. Further, there is a good chance that musicians have been exposed to some degree of music theory at a young age—those who showed an early interest in music may have taken piano lessons, or were fortunate to have music classes during elementary school. Therefore, it seems disingenuous to seek out those with experience who eschew such learning. Instead of comparing musicians based upon their breadth of knowledge—seeking evidence that one's grasp of theory contributes to their by-ear learning aptitude—we think observations of by-ear learning sessions are also necessary for such a study. Specifically, we would wish to focus on the written and spoken terminology employed during the learning session in order to discover which parts of music theory are called upon.

4.6 Summary

We set out to clarify our understanding about how musicians learn from recordings by ear, and observed 18 in-the-wild examples of them doing so. From these observations, we formed a set of hypotheses that lead us towards future studies.

This study made it clear that the methods employed by experienced musicians to learn from recordings seemed to play a more significant role than technology did—we observed few interactions with purpose-built technology. However, we saw musicians use different strategies to translate what they heard into a rendition of notes on their instrument.

Remarkably, we witnessed some musicians singing notes while seeking them on the instrument, and others that were unable to do so. However, both had to retain notes in their *mind's ear* while seeking them. We also saw evidence that familiarity with a song or genre of music may provide musicians with some additional benefits—perhaps additional context, or maybe long-held memories that aided with learning. These findings led us to conduct the survey of literature found in Section 2.4, and start to develop an understanding of musical memory.

The variations of the methods used by musicians could be explained by differences in foundational abilities, such as limited working memory, or an inability to sing pitches. However, it could also be the case that musicians are simply *taught* a mixture of strategies for doing so—some of which seem less effective than others. Unfortunately, we have not discovered a body of literature that sheds light on these variations in the methods employed by musicians who learn by ear from recordings. As a result, we conduct our own study of this question in Chapter 5.

Chapter 5

Towards an Understanding of the By-Ear Learning Task

Through an analysis of lesson videos acquired primarily from YouTube, we wish to characterize the task of learning by ear from recordings. Based on findings from the previous study (Chapter 4), we also focus our analytic lens to identify those parts of the lesson where one's memory for music (Section 2.4) is called upon. Specifically, we looked for instances where teachers discuss short-term or working memory, where it is used in the process of learning by ear, and any suggestions that musicians should develop or leverage their familiarity with a piece of music.

What we learn from this study is two-fold. First, we confirmed that musicians teach a wide variety of methods to learn songs by ear from recordings. Second, we find that memory not only plays a significant role in this process, but also that a musician's foundational memory abilities can help explain some of the differences we see between methods.

5.1 Study Goals and Approach

In a preliminary study (Chapter 4), we learned that musicians used varying techniques to learn notes during their interactions with recordings, especially while trying to remember and play the notes heard in a recording. For example, some musicians found it necessary to sing each note while seeking it on their instrument, while others appeared to sing for the sole purpose of remembering the notes. However, we could only hypothesize about the cause of these differences. While a saxophone player is unable to simultaneously sing a note while seeking it on the instrument, this does not explain why a guitarist could find their notes without having to sing them. Perhaps each of these musicians were *taught* to find notes in a certain way, but maybe they happened upon a practice that suited their own memory abilities. In this study, we set out to answer the following research questions.

RQ1: How are musicians taught to learn songs by ear from recordings? We hypothesize that analyzing a collection of lessons should allow us to form an approximate model of the task that a musician is taught to follow when learning a song from a recording.

RQ2: What differences exist between the teachings of by-ear learning's component tasks? Based on our findings in the previous study, we expect to find differences in tasks such as finding notes or chords. For example, some may rely upon their knowledge of theory while others do not.

RQ3: How does the musical memory of a musician play a role in by-ear learning? Again, our previous study suggests that memory plays a role when a musician needs to copy notes from a recording on their instrument, as does our research into musical memory (Section 2.4). Related, *RQ3.1: Do teachers explain how one's memory for music is related to by-ear learning?* An answer to this question could shed light on the importance of one's musical memory, explained from a teacher's point of view.

YouTube contains a wealth of relevant video data, and is an appropriate choice for many of the reasons discussed in Section 3.1. Further, analyzing lesson content on YouTube has already proven useful for other studies in the field of music education (e.g., Kruse & Veblen 2012; Whitaker et al. 2014; O'Leary 2020). We also know that a suitable collection of ear-learning lessons exists on YouTube. Many such videos surfaced during the prior study, but failed to meet our eligibility criteria—they lacked a genuine example of by-ear learning, and instead focused on *explaining* it in a lesson format. In the present study, finding such explanations was precisely our goal.

We also wish to focus on videos that are widely viewed, and those videos that regular YouTube users are going to find when they search for lessons about learning by ear. Videos that rank high in a user's search are likely to have retained viewers' attention, been liked, or commented upon. Presumably, the level of audience engagement should indicate something about the *perceived* value of its content, but this approach could introduce a bias towards entertainment over effectiveness—for example, surfacing poor-quality videos that attract vitriolic comments. To provide contrast, we also analyze the content of a commercially available, multi-instrumental DVD video lesson about learning songs by ear from recordings (Huckabee 2004). This video will provide another source of comparison for both the quality and relevance of content found among the online videos.

Given the apparent lack of a standard pedagogical framework for by-ear learning from recordings, we use an approach inspired by grounded theory (Section 3.2.4). Specifically, the content of videos is coded openly, then grouped into concepts and categories progressively as we gather additional data. Then, we review these data as hypotheses develop, and identify similarities and differences between notable concepts or categories.

5.2 Method

5.2.1 Obtaining Videos and Transcripts

The video collection started with a two-hour-long DVD from my own personal library (Huckabee 2004), and one video from the previous study. We then added the results of a YouTube search performed on February 16th, 2024 with the query string `"how to" "learn songs by ear"|"learn music by ear''`. Results were sorted by view count, and we considered only the top 15 most-seen videos from this list.

The first 16 videos were biased towards guitarists, and the query string did not seem characteristic of a regular YouTube user. Thus, we simplified the query, and added the name of a specific instrument to each. Our instrument selection was based upon the prevalence of instruments in popular music—the piano (/keyboard, 70%), guitar (50%), and bass (30%) were most-credited among the Billboard Hot 100 Songs from 2023[1]. Despite not appearing in this list, we also considered the saxophone because it is inherently monophonic, and at times has been found regularly in popular music (McKinney 2017). Instrument-specific searches used the same prefix string—`learn songs by ear on`—followed by: `piano`, `guitar`, `bass`, or `saxophone`. Each search was sorted by relevance to match the default behaviour on the YouTube website. Searches were performed between February 26th and March 4th, 2024, and we considered only the top 5 results from each.

The above searches were all performed using the YouTube Data Tools[2], a website that accesses YouTube via its official search API. For each online video, we obtained

[1] https://www.billboard.com/charts/year-end/2023/hot-100-songs/. Performer credits were obtained from Apple Music.

[2] https://tools.digitalmethods.net/netvizz/youtube/

IDs	Query	Qty
V0xx	*n/a - seed videos*	2
V1xx	`"how to" "learn songs by ear"|"learn music by ear"`	9
V2xx	`learn songs by ear on piano`	5
V3xx	`learn songs by ear on guitar`	4
V4xx	`learn songs by ear on bass`	4
V5xx	`learn songs by ear on saxophone`	5
	Total	**29**

Table 5.1: The video IDs for each of the six rounds of data collection, including the query terms and the number of eligible videos that were analyzed.

automatically generated transcripts from YouTube. Using Aiko[3], the DVD video lesson was transcribed using a local copy of the *whisper v2* model from OpenAI[4]. The quality of the transcript was superior to those obtained from YouTube, so we downloaded the rest of the videos to local storage and transcribed each of them. For those videos where Aiko failed to generate a usable transcript, we continued with YouTube's version.

The main criteria used to determine video eligibility in this study was duration (up to one hour long), instrument, and language. In the first search, we rejected a drumming video that was longer than one hour, one video in Spanish, and a violinist teaching with Arabic Maqam music. In the instrument-specific searches, one bass video was omitted for being longer than one hour. The two-hour-long DVD was the sole exception to this rule.

We identify videos using a numeric code corresponding to the query they were obtained from. When videos appeared across multiple searches, we assigned them an ID that represents the first instrument-specific search they appeared in. For example, a video in both the 12th position of the general search results and the first position of the saxophone results was assigned V501 rather than V112. Another that appeared 5th in the piano-specific search and 4th in guitar-specific search but was assigned V205. See Table 5.1 for a list of queries, the codes associated with them, and how many videos each contributed to the study.

[3] https://sindresorhus.com/aiko

[4] https://openai.com/research/whisper

5.2.2 Coding Videos

The videos were analyzed iteratively, beginning with the open coding (Corbin & Strauss 2008; Charmaz 2014) of transcripts. We coded fine-grained segments—on the order of 1-3 statements long—in a way that represented what the presenter was saying or doing. Our codes include things such as *Finding one note at a time*, *Pausing playback after first note is heard*, or *Suggesting that one may need to start by learning notes one at a time*. Codes were grouped into concepts, then arranged hierarchically into categories. For example, we grouped the above codes into the concept *One Note at a Time*, and combined it with the *One Chord at a Time* and *Bite-Sized Pieces* concepts to form the *Working in Chunks* sub-category that sat beneath the top-level *Memory for Music* category.

Following an approach inspired by grounded theory (Section 3.2.4), coding was performed in tandem with our data collection. We captured memos as concepts and categories began to form, and met regularly to discuss the emergent findings. Additionally, we continually revisited the data to help refine our thinking about those categories we deemed most remarkable.

We coded a total of 29 videos, and identified 5 top-level categories: *About By-Ear Learning*, *About the Teacher*, *Prescribed Order of Learning*, *Recognition Strategies*, and *Memory for Music*.

5.3 Results

Here we present the most notable categories that we identified, all of which apply to by-ear learning. In these videos, we observed people speaking from a position of experience—*teaching* to the audience. While we can neither assess their credentials or competence, we refer to those people generally as *teachers*, and viewers as *students*.

One category—*About the Teacher*—is entirely superfluous to our study. It captured sub-categories such as *self-promotion*, *openness to audience feedback*, and statements related to *the teacher's own skills*. Here we find (paraphrased) calls to "like and subscribe", "buy my e-book", or "sign up for my online course"; requests to "leave a comment below", or "send me a message if you have questions"; and claims such as "I am still building my music theory knowledge", or demonstrating their skills by learning a set of songs requested by their viewers. We merely gathered these codes to help us shape the coding practices that led to our results.

5.3.1 About By-Ear Learning

Many online teachers shared the opinion that pop music instrumentalists should strive to learn music by ear. Their statements were most often platitudes, merely suggesting that it is the *most important thing to learn* (V001, V503), or that *it makes you a better musician* (V302, V303). Some teachers helped shed light on these ideas with more concrete claims: repeated practice *develops your ear* (V001) over time; also, learning music by-ear *frees you from needing lessons, sheet music, or tablature* (V001, V201, V303) which *generally isn't great* (V001, V303, V504), partly because it doesn't allow you to *learn the nuances of a performance* (V001, V302, V505) such as bending notes on a guitar, or slurring notes on a saxophone. Additionally, teachers claimed that by-ear learning *helps with your ability to improvise* (V114, V501, V503).

Teachers also frequently suggested that—while it can be challenging, or slow at first—by-ear learning gets easier with repeated practice. Sometimes this statement was made with respect to the entire *journey* of learning by ear—that the student should expect the process to take considerable time in the beginning, and that learning more songs by ear will lead to improvements (V001, V109, V203, V204, V303, V305, V402, V403, V501, V503). One teacher characterized this progression as "exponential" (V501), but the rest suggested by contrast that it required consistency, and would develop more slowly over time. Two of the teachers also claimed that *finding notes gets easier over the course of learning a song*, especially once the first note is found (e.g., V001, V303).

Some teachers claimed that learning by ear may prove to be a taxing activity. For example, students may need to set aside certain songs because *sometimes it's just too hard* (V001), or that one may need to take frequent breaks because *you can't do it for long* (V001, V501) as fatigue may set in.

5.3.2 Prescribed Learning Order

When teachers provided students with a set of steps for learning songs by ear, they most often suggested that students *learn the key first* (V101, V115, V202, V203, V204, V301, V302, V305, V502, V505). Whether or not key-finding was the recommended starting point, guitarists were instructed to *identify chords before learning either melodies or solos* (V101, V301). Otherwise, saxophone players learned only melodies or solos, and bass players learned bass lines. Pianists seemed to be mixed in this regard, either *learning the melody before finding chords* (V201), or vice versa—*learning the left-hand harmony before the melody* (V202, V204).

Teachers often expressed that the process of learning songs is a *repeated application of sub-procedures* (V109, V301, V001, V501, V502, V503, V504). That is, learning chords is a repetition of the chord-finding process, melody notes a repetition of note-finding, and this continues until segments are learned, then sections, and finally the whole song. For example, the saxophone player in V501 says, “keep repeating the process and just keep taking it one note at a time.” The guitarist in V301 shares a similar sentiment about chord finding, “it’s more or less a rinse and repeat process to find the rest of your chord progressions.”

In some videos, the overall sequence was difficult to ascertain—the teachers either lacked a lesson plan, diverged from it, or planned to discuss tangential topics. For example, in V110 the content was largely recommending the use of high-quality headphones, an audio interface, and digital audio workstation software—all of which had links (some with affiliate tokens) in the video’s description. In V101 and V205, the videos spent more time discussing music theory than providing concrete suggestions about how to learn from recordings.

In many lessons, teachers assumed students already have certain *prerequisites*, though they were not always stated explicitly. For example, in V401 the bass teacher says, “What you need to know to develop your ear is when you land that note, it’s correct, right?”, whereas most others simply assumed the student could tell when a note was incorrect—for example, when the guitar teacher in V109 says, “hopefully you’ll be able to tell if it’s not the correct note and then you’re going to move in either direction chromatically, so fret by fret so as to make sure you don’t skip over the correct note until you find that note.”

Some teachers mentioned that students should be able to *match pitches vocally* (V401, V503). However, this skill was not to be confused with being able to sing. For example, the bass teacher in V401 says, “You don’t have to have a beautiful singing voice. I certainly don’t have a beautiful singing voice. But what you at least need to do is when you hear a note, you need to try to match the pitch with your voice.”

Other such prerequisite skills were knowing topics from music theory such as *scale degrees* (V305), and certain elements of instrumental proficiency like *knowing the note names on the fretboard* (V305), or *knowing how to play scales* (V305, V502, V503, V504). By contrast, some teachers suggested that *anyone can learn by ear* (V204, V303, V501), or *you can start learning immediately* (V001, V302)—claiming that even those early in their instrumental journey can start to learn how to play music by-ear.

5.3.3 Recognition Strategies

Extracting Salient Notes

While learning from a recording, it is important that the *pitches of interest can be accurately reproduced* by the loudspeakers or headphones with which the musician is listening to the recording (V001, V110, V402, V501). Alternatively, they suggested making adjustments so that *salient notes sound more prominent* (V001) by using an equalizer or other such tone knobs. While this is certainly a concern for bass players, listening for the bass notes is also the most commonly suggested method for finding chords, which is discussed separately below.

Regardless, the musician needs to *listen carefully for the salient notes* (V001, V002, V115, V301, V402, V505), which can require a great deal of focus and concentration. Salient notes are sometimes *difficult to hear when there is a difference in timbre* as the teacher in V001 explains: "In other words, even if your voice and a saxophone and a piano are all playing the same note, they all have a little different texture, and this textural difference can confuse your sense of pitch, at least until you get a little experience." He also describes how recordings can have unique characteristics that may further contribute to this challenge.

> Understand that every recording is different. Each producer that releases a recording treats the mix a little differently. Some like for the solo instrument to be way above the level of the other instruments so it'll really stand out, while others like it just slightly above the others so it'll be well supported. Some like it drowned in reverb, while others like it relatively dry. Some like the stereo panning drastically spread out, while others like it close to centre so that when a radio station plays it and a channel gets lost, the recording will still have all its parts. So each recording you try to figure out will be different as to how vividly you can hear your hero. (V001, 1:01:15)

The bass teacher in V402 describes an exercise for students to practice their listening skills, and demonstrates how hearing can be focused on individual instruments in a song. He explains that—much like our vision—other instruments begin to "blur" as one's focus shifts. In a similar vein, students may need to ignore the song's lyrics so they can focus on listening for only the notes in a melody (V201).

Musicians can also employ *singing to identify and extract salient notes*, such as the individual voices in a chord (V202), or the bass notes (V401). However, singing appears to be more significant as a memory aid (Section 5.3.4). Because not all students can do this, the teacher in V401 suggests practicing "when you're driving in your car, you're on your way to work, you're at home, you're chilling, you're listening to music, wherever you are listening to music, please sing those bass lines."

Finding and Naming Notes

To locate a note on the instrument, it is often taught as a *random search* (V001, V002, V115, V301, V303, V401, V404, V501). That is, the musician plays some arbitrary note, then makes adjustments to match what they hear—if the pitch is too high, they play lower notes, and vice versa. To ease the pain of such a strategy, teachers recommended that students *limit the search to the song's key* (V115, V203, V302, V404) so that they have fewer notes to test.

Once the note is found, it must be named before it can be written down, or used to determine which chord to play on a guitar. For pianists, bassists, and guitarists this is often a matter of looking down at the position of their fingers, and recalling which note is produced at that position (V101, V109, V301, V303, V305). For those who are proficient with scales on a guitar, piano, or saxophone, or they can sing *solfege*, the note's name can be assigned based on their *interval from the tonic note* (V201, V503).

We wish to point out that teachers rarely spoke about *perfect-* or *absolute pitch* (Section 2.4.3)*,* which would make the aforementioned note-finding strategies unnecessary. When teachers mentioned perfect pitch (V302, V501), they seemed to assuage a student's concerns of unattainability. For example, during the opening of V501 the saxophone teacher says, "I don't have perfect pitch and I don't think you need to have perfect pitch in order to be able to learn stuff by ear."

Identifying the Song's Key

Teachers described two methods to identify the *tonic* note—the first note in the key that shares its name (e.g., F is first note of both F major or F minor). Some teachers suggested this note can be found in *common locations of the tonic*. For example, it might be the first or last note (V204, V501) in the song, or the first or last in a verse or chorus. An alternative method has the musician play a single note while also listening to the recording, then evaluating how well it fits (V404, V505).

Once the tonic note was known, an additional step is necessary to determine whether the key is major or minor. To achieve this, the musician can experiment by *auditioning* *scales alongside the recording* (V101, V202, V305, V404, V505) to choose which sounds most correct.

This second step can be skipped if the musician instead searches for chords. That is, if the tonic chord is major, or minor, then so is the key. They do this using similar methods as above, either by looking for them in common locations (V001, V204), or by listening for the chord that sounds most like “home base” (V001, V101, V202, V401).

Using one’s knowledge of music theory, key identification can be done using a different two-stage process that relies on a test of *set membership*. Here the musician starts by identifying the melody notes (V203, V301), or chords (V109, V115, V205). Then, they look for the key that contains all of them. This strategy is notable because it runs counter to the advice to learn the key first, which is supposed to help make note- and chord-finding go more quickly.

Finding Chords

The most-recommended method for finding chords is similar to the two-stage method for identifying the song’s key. That is, first the root of the chord is identified by *finding the bass note* (V001, V002, V107, V115, V202, V205, V301), and then the *quality* of the chord must be identified—whether it is major, minor, diminished, etc.

Quality identification is sometimes taught as a process of *trial and error*. Here, the teacher recommends that the student *auditions major and minor chord candidates* (V001, V109, V204) alongside the recording to hear which sounds correct choice. However, a more efficient method involves some music theory knowledge. That is, by knowing the song’s key, the quality of the chords can be deduced immediately (V109, V202, V204, V301).

Other trial and error methods can yield both the root and chord together. For example, if the key of the song is already known then chords can be drawn at random from the set of seven *diatonic* chords that belong to the key (V202, V305), and by using a process of elimination (V204) the list of choices shrinks as new chords are learned.

A modified version of this method relies upon one’s knowledge of commonly used *chord progressions* (V203, V205), reducing the initial set to only three or four. If the melody is already known, then the musician can choose from those chords having a note in common with the melody (V001), or one that simply sounds correct alongside it (V201, V204). In V204, the teacher sings the melody and listening for those chords that match, and in V201 they listen for the chords that sound correct against the melody of *Twinkle, Twinkle, Little Star*.

What to Expect in Pop Music

Some teachers discussed how the chord progressions in popular music are rather simple in practice—consisting of few chords, often in predictable sequences (V203, V205, V301, V305). Teachers also discussed other commonalities such as the prevalence of 4/4 time (V109), or that pop songs usually contain only major and minor chord qualities (V205). Additionally, teachers would discuss things that are less likely to occur, such as changing keys (V305), the use of non-diatonic harmony (V205), or diminished chords (V301, V305). In V301, the teacher even suggests that hearing a seventh degree bass note more likely indicates the first inversion of the dominant (V) chord than the diminished seventh (vii°). Knowing about these common practices in popular music helps to simplify the task of identifying chords, and very little about music theory is required to achieve these efficiencies.

Music Theory Fundamentals

While knowledge of music theory is not required to identify the individual notes or chords, it certainly appears to help make these tasks go more quickly. In some of the lessons, teachers chose to include brief discussions about theory. In most of these instances, teachers talked about the concept of *diatonic chords* (V201, V202, V205, V301, V305). Some taught this concept in relation to the *circle of fifths* (V101, V205), and one explained it in the context of the *Nashville Number System* (de Clercq 2019) (V301). Presumably, these teachers chose to include such discussions because they felt this was the baseline level of theory that is necessary to learn songs by ear, or at least makes it go more smoothly.

Leveraging Purpose-Built Technology

Many teachers suggested that students try *slowing down playback* while learning by ear (V001, V109, V115, V501, V504, V505) as a way to help them listen for fast-moving notes or chords. Most often, this feature was demonstrated using the slowing functionality built into the YouTube web player. The teacher in V001 explains exactly why this feature is beneficial while learning by ear.

> But the main thing to remember is that slowing down a fast passage can be a real equalizer. It can reduce an advanced solo to being just as easy as London Bridge or Old MacDonald Had a Farm. (V001, 54:54)

The use of *looping playback* was also discussed by teachers (V001, V109, V504). However, the teacher in V001 has specific recommendations for when students should use this feature.

> You may be tempted to repeat something for hundreds of repetitions in hopes that the passage will eventually soak into your brain through osmosis. I've never seen it work on any of my students. Sure, let it repeat four or five times to become acquainted, if you feel like it'll help, but eventually you're going to have to stop the tape, try to hum what you've just listened to, and then try to search for each note one at a time. [. . .] The only time you should ever be playing along with the recording is after you've already learned it or if you're just using the recording for something to jam along with to practice your improvising. (V001, 1:05:17)

Identifying Things Directly By Ear

Some teachers recommended that students try to develop the skill of *relative pitch*—by-ear recognition of *intervals*, or the relative distance between notes (V002, V115, V201, V203, V205, V401, V403, V501). Teachers suggested that this could be improved with the help of *solfege* singing (V002, V201, V302, V401). The most basic form of this technique was demonstrated in V201, where the piano teacher sung the major scale from the tonic note of the song's key, up to the note they were seeking. That revealed its interval from the tonic, and then the note's name.

Sometimes teachers spoke about a simpler, more intuitive sort of recognition. For example, being able to recognize whether a key or chord is major or minor based on whether it sounds "happy" or "sad", respectively (V109, V301, V404). Or, simply trying to make rough judgements about the distances between notes while learning them (V001, V201, V303) so that the trial and error process isn't a series of *completely* random draws.

5.3.4 Memory for Music

Playing Familiar Melodies From Memory

In five lesson videos, the teachers' first recommendation for musicians who want to start learning by ear was roughly the same: begin practicing by learning songs that can readily be sung from memory. Most often, these were children's nursery rhymes.

Among the nursery rhymes, *Twinkle, Twinkle, Little Star* appeared most frequently (V002, V201, V303, V501). In V501, the teacher explains this choice: “So the first thing you’re going to want to do is pick a song that you want to try and learn. This will likely be something that you already are familiar with and can maybe even sing and have it in your head.” However, their lesson demonstrated this process using a recording of the song they found on YouTube. In all other cases, the exercise was meant to be performed from memory.

In V001, the teacher used an entirely different set of nursery rhymes from the rest. He begins his demonstration of note-finding using *My Country ’Tis of Thee, Old McDonald Had a Farm*, and *London Bridge*. Among these demonstrations, he states “If you’re just starting out, I recommend that you pick out as many of these easy children’s songs as possible because it will help to sharpen your ear and hone your skills.”

There were two videos in which nursery rhymes were not called upon, but a core component of the lesson consisted of learning a popular song entirely from memory. In V203, the teacher reconstructs a pop song beginning with his memory of its vocal melody. After identifying a set of melody notes, they deduce the song’s key before proceeding to determine the harmony (Section 5.3.3) that accompanies the melody. A similar practice can be seen in V201 and V204.

Developing Familiarity With Songs

Some teachers made it explicit that students should listen to, and familiarize themselves with a recording before they start trying to learn from it (V105, V109, V402, V404, V504). In V105 the teachers refer to this activity as *active listening*: “when you put everything else away and only concentrate on listening to the song, paying really close attention to the actual song and intently listening to it.” Their rationale for intensively listening to the song beforehand is that it makes the process go more smoothly. “[. . .] the more committed you are to doing this, the easier it’s gonna be to learn the song later on when you get into these other steps, because you’re gonna be familiar with it already and you’re gonna kind of know what’s coming.” In V402, the bass teacher says “Listen and pay attention. Listen to the words, the sections, and as you listen over and over, you’ll be able to anticipate what comes next, you know, as the song moves along. So you’ll be learning the form of the song without even trying [. . .]”

This active listening exercise is meant to happen outside of the musician’s attempts to learn the song on their instrument, as one teacher in V105 describes, “[. . .] give it some really good active listening time where you’re not doing anything, not even walking. Just sit on a couch or in a chair, put your headphones on, listen, no phones, no distractions." We also see this in Chapter 4, where one guitarist claimed to be learning a song that was new to them, and they intentionally listened to it repeatedly beforehand to develop familiarity. That guitarist was doing precisely what the teacher in V109 says: "[...] you're going to want to have listened to that piece of music a lot and be super familiar with it."

Intentional Short-Term Storage

In the lesson videos, it was most often suggested that students try to *sing or hum notes to retain them* (V001, V002, V106, V107, V108, V109, V115, V203, V204, V301, V303, V401, V403, V404, V501, V502, V503, V504, V505). That is, upon recognizing those notes that are salient the musician must then recall and imitate those pitches vocally. Sometimes, bass notes must be sung one or more octaves higher to match the vocal range of the singer (V002, V107, V401, V505).

The guitar teacher in V301 suggests that singing is meant to help students remember notes: "[...] one of the best tips when searching for your note is to hum the notes yourself. This way you can keep it in your head as opposed to having to listen to it over and over." However, this may not be possible right away, as the saxophone teacher in V502 suggests that students need to have "heard it enough times that [they] can sing it or hum it without needing any help."

Some musicians vocalized pitches as a way to retain notes as they seek them on the instrument. For example, in V203 the piano teacher demonstrates that a note is sung continuously while walking chromatically up the piano keys to match its pitch. However, a saxophone player cannot do both things at once due to the nature of the instrument—the teacher in V505 only sings the phrase to remember it, but we hear only his saxophone while he seeks the notes.

Another way that both pitches can be heard at once is to try and seek notes on the instrument while the recording is playing. For example, in V115 we see a guitar teacher that advises playing over the recording while trying to locate the roots of chords. The bass teacher in V401 also demonstrates simultaneous playing over the recording, though it is a reconstruction of his first attempt as a thirteen year old to learn the song—he later recommends that students *sing notes* to remember them.

Short-Term Memory Delicacy, and Chunking

The teacher in V001 expressed that one's working memory for notes can be fragile, and disrupted when incorrect notes are played: "When it sounds wrong, you lose the tune in your head, you go back and play the tape again." This may explain why he also suggests that humming or singing should be done only after playback is stopped: "Figuring out licks, as you've seen in this program, is always done with the recorder off. You play the recording, you stop the recording, and then you search for the note on your instrument." We also also see this practice in other videos, such as the saxophone teacher in V505 who sings phrases to remember them while the recording is stopped, and the teacher in V501 who stops the recording after only a single note is played.

> All I want you to focus on is just how the note sounds. The easiest way to do this is to just pause the song as soon as you hear the first note. [*demonstration*] Okay so there's the first note. Now you just want to try to match it with either your instrument or your voice. Sometimes it can be easier with your voice because you have it in your head then and you can basically listen to that note over and over again until you can match it on your instrument. (V501, 1:26)

Continuing playback may cause a disruption to tonal working memory, but this could also be an issue of capacity. Teachers recommended that notes (V001, V301, V501) or chords (V109) should (initially) be learned one at a time. They may recognize that a beginner has not yet developed the tonal working memory capacity to learn more than that in each *chunk*. However, it seems that teachers believe this should improve with practice (V001, V501). For example, the teacher in V501 says "Once you can consistently and quickly get one note at a time try to get two. Eventually you'll be able to get entire phrases and you won't even need to slow down the song."

It appears that teachers also try to work within the limits of their full tonal working memory. Teachers encourage students to take this into account while they learn, as the saxophone player in V505 describes using the following analogy: "When you have a big hamburger, you don't try to eat the whole thing at one time. You take little bites." Failing to do so leads to the situation encountered by the saxophone player in V501, who—while trying to learn a song that was requested by his audience—says "[...] I always struggle with memorizing longer phrases. I got the first few notes but by the time I got to the end of the phrase I didn't really remember what it sounded like."

Musicians may also reach their limits as they begin to accumulate what they learn. While learning a mandolin solo in V001, the teacher says "[...] there was a lot to that, and a lot I'm trying to keep in my head at the same time. In a case like this, you might write all this down." Here they are simultaneously retaining the just-learned notes in memory while also ingesting new ones from the recording and learning how to play them as well. Another recommendation from teachers was for students to write the individual notes (V301) or chords (V109) down as they learned, or producing something like sheet music (V109, V501). In V109, the teacher walks the student through the process of capturing the bars in the song, the chords within, and also suggests that students try to capture strumming patterns.

The Mind-Instrument Connection

One common selling point of by-ear learning is that it improves the student's musicianship, and fluency with their instrument. However, it seems that this fluency can reach that with which we can speak or sing after considerable practice, as described by the teacher in V001: "Eventually, your fingers go automatically [on your instrument] to everything you hear," which appears to come after "search[ing] for hundreds of little tunes with your fingers." Other teachers appear to imply the same outcome from repeated practice—that it eventually *takes less time to locate notes* on the instrument with practice (V001, V201, V303). Just as some teachers suggested that students learn to sing the major scale in solfege (Section 5.3.3), others said that *knowing more scales makes the process easier* (V302, V504). That is, being able to play the major scale (and others) naturally in all 12 keys would effectively provide an instrumentalist with the same benefit as learning to sing those scales—having some degree of familiarity with the production and recognition of their intervals.

The teacher in V001 likens vocal pitch reproduction to playing an instrument that the student already has experience with: "You had hours and hours of practice, and you finally got to where your vocal cords became an extension of your ear and your brain." In a similar vein, the guitar teacher in V302 asks students to repeat a spoken statement and claims, "If you were able to repeat that, you can play by ear."

In those instances when students were not instructed to write down the notes and chords they learned, some teachers recommended that notes or phrases are progressively *fused* together (V502, V504) as a way to memorize them. That is, gradually building up to playing an entire sequence of notes by repeatedly playing them from the start as successive notes or phrases are learned—playing the first note, then the first two, and so on until the entire song or section of interest can be played.

5.4 Discussion

Analyzing lesson videos helped us understand the process of by-ear learning from the perspective of those who teach other musicians to do it. As we have seen in Section 5.3.1, teachers claim it is important for musicians to learn by ear. Admittedly, one expects to find such statements from those who choose to engage in (and especially teach) this method of learning. However, the common use of by-ear learning in popular music (Section 2.1), and the inability of notation to capture what can be heard in recordings (Section 2.2) help reinforce this notion of importance.

5.4.1 RQ1: How are musicians taught to learn songs by ear from recordings?

Based on our analysis, learning a song by ear requires that musicians execute one or more repetitions of *sub-tasks*, which we roughly characterize below. Because we did not find a consensus among teachers with respect to their order (Section 5.3.2), we present the sub-tasks with a weak ordering.

Playing Melodies From Memory—when it is taught—is presented as a preparatory exercise that precedes one's attempts to learn songs from recordings. As we discussed in Section 5.3.4, this is most often demonstrated using well-known nursery rhymes such as *Twinkle, Twinkle, Little Star*. The common appearance of this introductory exercise among the lesson videos suggests that teachers consider it to be a foundational (or prerequisite) skill that is necessary for by-ear learning.

Learning the Key is commonly taught as the entry point to learning a song. Once identified, the song's key make subsequent sub-tasks easier to execute because there are fewer notes—seven, and not twelve—to consider while learning. However, the student must possess a baseline knowledge of music theory to enjoy such efficiencies, which is why we see teachers often discussing music theory in their lessons (Section 5.3.3).

Learning the Chords is often the next (and possibly last) step for guitarists, and usually only one component of what a pianist is asked to learn—we do not see bass and saxophone teachers explaining this to students. This sub-task is taught in a way that is meant to be executed repeatedly throughout the song. Upon learning one or more repeating sequences of chords (*chord progressions*), the player has a complete representation of the song that they can perform with a band, sing alongside, or possibly combine with the melody and/or solo.

Learning the Melody may precede chord-finding—especially when its notes are used to identify the key. While it is an optional sub-task for guitar and piano players, it was often all that a saxophone player was taught to learn from a pop song. *Learning the Bass Line* is largely the same activity, but is specific to bass players. Similar to learning chords, teachers often taught note-finding as a method that is to be applied repetitively, to learn one or more notes at a time.

Learning the Solo was not frequently taught in the videos, though it is the primary goal of learning for some players (e.g., the saxophone player in V505). This differs only slightly from learning the notes in a melody. Specifically, solos often demonstrate advanced instrumental skills, and include specific performance elements that can make them more challenging to mimic—for example, identifying that notes were bent on a guitar, or notes slurred on a saxophone.

Teachers effectively claim that—by applying the above tasks in (roughly) this order—the student can "learn a song" by ear. However, what they are *actually* teaching students to do is create their own *derivative performance* of the original recording. When a musician copies only the portion of a recording they can play on their instrument, then performs their part—without singing, or accompaniment from other band members—they produce an incomplete rendition that is unlikely to be recognized by a listener. To attain a more complete rendition, musicians create *instrumental covers* of songs that can stand on their own; we saw this demonstrated by piano teachers in V202, V203, and V204. Presumably, it is the nature of the piano that explains why only these videos contained this practice: ten distinct notes can be played by ten fingers, allowing the harmony and melody to be performed together.

Given that the goals or abilities of the musician are likely to differ, we expect that the application of the above steps is piecemeal in practice. It also comes as no surprise that we found such a great deal of variation between the lessons, and how each task was taught.

5.4.2 RQ2: What differences exist between the teachings of by-ear learning's component tasks?

For most of the sub-tasks we listed above, there appears to be a dichotomy between techniques that rely upon music theory, and those that do not. For example, we saw in Section 5.3.3 that the key of a song can be found by testing a list of melody notes for set membership—requiring that the musician knows which key contains them all; by contrast, they can simply evaluate whether a song sounds *happy* (major) or *sad* (minor) once the tonic note is identified by ear. Similarly, in Section 5.3.3 we learned that once the key is found, the musician can determine a chord's quality immediately based on their knowledge of diatonic chords; or, they can just play both major and minor variations of a chord to listen for that which sounds correct. To find the notes in the melody, solo, or bass line, we saw in Section 5.3.3 that the musician can limit their search to those in the song's key; or, they can simply test every single one using the trial-and-error method. When teaching these sub-tasks, teachers often framed the theory-based versions as being more efficient. However, we saw lessons teaching *both* theory-based and theory-free methods, as well as lessons teaching one or the other.

Regardless of whether theory was involved—we noticed additional variations between the techniques used to execute the sub-tasks. For example, we saw in Section 5.3.3 that key finding can be done testing the membership of either notes, or chords. The latter method is especially strange, given that the student is identifying the key *after* it would have been beneficial to know that information. For example, the guitarist in V115 locates the chords (and its bass notes) using the theory-free method of trial and error before deducing the song's key using their knowledge of diatonic chords.

Each of the key- and chord-finding methods described above could be made easier with the help of MIR techniques. At one extreme, this exercise could be made redundant by estimating the key or chords using one of the approaches we discussed in Section 2.5.3. However, one could imagine a method that doesn't automate the process completely—allowing the musician to continue participating in the by-ear recognition process. For example, a musician could be provided with theory-based tools that narrow their key or chord choices based on a set of notes they already identified.

We also found significant differences in the way notes are identified by ear. For example, in Section 5.3.4, we discussed how some teachers describe note-finding in a way that advocates singing a note while it is sought on the instrument; or, they suggest that students learn to identify intervals by ear (Section 5.3.3); and we also observed teachers like the one in V505 who sings to remember, but can seemingly compare notes in his mind.

There are indeed differences in the way that by-ear learning is taught to musicians. This confirms our hypothesis, and certainly helps to explain the source of differences we observed between the methods used by musicians to locate notes on the instrument in Chapter 4.

5.4.3 RQ3: How does the musical memory of a musician play a role in by-ear learning?

The student's memory for music is called upon in a number of places in the by-ear learning task, given our results in Section 5.3.4. Also, much of what we observed can be linked to the literature we discussed earlier in Section 2.4.

When suggesting that students try playing nursery rhymes from memory on their instrument (Section 5.3.4), teachers are asking their students to draw upon their long-term memory of that melody. We know from Section 2.4.4 that such memories are not merely a collection of absolute pitches, which explains why the student is often told they can start singing or playing from an arbitrary note. To reproduce the melody on their instrument, the students call upon sensorimotor memories to generate candidate notes, and their tonal working memory (Section 2.4.1) allows them to compare those pitches they hear with the targets in their mind.

To match a pitch from the recording—in the melody, the root note for a chord, or the tonic of the key—students are effectively being asked to encode the salient pitch in tonal working memory. Students were often taught to sing or hum the salient pitch(es) that they heard, which may help the student encode them in memory (Section 2.4.1). Here, the *absolute* pitch must be retained for as long as it takes the student to find the correct note on their instrument. Again, the student's tonal working memory facilitates this comparison. When the memory of a pitch is lost, the student must start over by listening to the recording again (Section 5.3.4).

We saw that students were sometimes taught to continue singing the pitch from the recording while seeking it on the instrument (Section 5.3.4), which gives them more time to do so. However, this method does not appear to rely upon tonal working memory, as the comparison is happening *live* between two audible tones—similar to how one might tune a guitar aurally, relative to another pitch. Unfortunately, this method is inaccessible to saxophone players, as we have also discussed in Section 5.3.4. Therefore, instrumentalists who cannot sing and produce notes at the same time must rely upon their tonal working memory to retain pitches for long enough that they can be found.

Finally, we also found evidence of a different sort of long-term memory—one that is formed when students develop familiarity with a recording intentionally (Section 5.3.4). Here, teachers are not asking students to remember the entire tonal content of the song. Rather, it appears that this exercise helps students develop *context* that will help frame their learning. These listening sessions are helpful because they can either facilitate the creation of long-term memories, or reinforce existing ones by cueing their recall

(Section 2.4.5).

RQ3.1: Do teachers explain how one's memory for music is related to by-ear learning?

In the discussion above, we used our understanding of the literature to connect the by-ear learning sub-tasks to memory. Sometimes, teachers made this connection explicitly, though they did so in general terms. For example, in Section 5.3.4 we saw claims that singing helps "keep it in your head" (V301), and that it's necessary to have listened "enough times [to] sing it or hum it without needing any help" (V502). We did not expect teachers to provide in-depth explanations about short-term memory, but statements such as these helped explain *why* the teacher thinks it is important for students to sing or hum notes, or listen to a segment repeatedly.

In one case, we saw a teacher explaining that excessive attempts to match a pitch may cause the student to "lose the tune in [their] head" (V001). This mirrors findings from the literature (Section 2.4.1) that suggest tonal working memory can be disrupted by additional pitches that are heard after the target that is to be matched. This connection could help explain why some teachers suggested students continue vocalizing pitches while seeking them on the instruments.

Our results also showed that teachers sometimes acknowledged the limited capacity of short-term memory. For example, we reported in Section 5.3.4 that students are taught to work in short segments, and that teachers themselves encountered limits to their working memory. Additionally, students are told that they may need to write notes down as they begin to accumulate, and their short-term memory becomes full.

We also observed that teachers expressed how a student's capacity for notes would continue to develop over time (Section 5.3.4). This suggests that the student's tonal working memory may get trained as they continue to learn from recordings.

5.4.4 Connections to Prior Results

Our findings in this study helped explain some of the behaviours that we observed in Chapter 4. In Section 4.4.1 we hypothesized that musicians learned only portions of whole songs because of the repetitive nature of popular music. While some teachers acknowledged this musical simplicity (Section 5.3.3), in Section 5.3.2 we also see that the ear learning process itself is repetitive. In Section 4.4.2, it appeared that musicians created notation as a long-term memory aid, and teachers recommended doing so for the same reason in Section 5.3.4. We observed very few experienced musicians using purpose-built technology features in Section 4.4.3, but in Section 5.3.3 we found teachers frequently recommending that students slow down playback to help them learn fast-moving passages—supporting our hypothesis that musicians might outgrow the need for certain assistive features as they gain experience.

Many of the observations we discussed in Section 4.4.4 had matching results in the present study. When some of the musicians played alongside the recording, they demonstrated the trial-and-error key- (Section 5.3.3) and chord-finding (Section 5.3.3) methods taught in the lesson videos. Those musicians who sung what they heard—sometimes while seeking notes on the instrument—were doing so in order to help them remember those notes (Section 5.3.4). In Section 4.4.4 we hypothesized that one's ability to remember notes may influence how they proceed to learn songs, which lines up with teachers recommending that students start out by learning one note (or chord) at a time (Section 5.3.4).

We also found evidence in Section 4.4.5 that developing familiarity with a recording may help musicians learn them more quickly, which matches recommendations from teachers that students do the same (Section 5.3.4). In Section 4.4.6, we saw musicians employing strategies we found in lesson videos for finding chords (Section 5.3.3), notes (Section 5.3.3), and also the key (Section 5.3.3) of songs. We hypothesized in Section 4.4.6 that knowledge of theory and common patterns in pop music may speed up the learning process, which is supported by what we observed teachers claiming in lesson videos (Section 5.3.3, Section 5.3.3).

5.4.5 Limitations

A significant challenge of using online videos is that there is a lack of uniformity in the video data. Unlike a (semi-)structured interview, the person in the video cannot be directed to remain on topic, or asked to discuss the task in full. For example, the teacher in V402 said nothing about whole-song learning—they did not discuss finding the song's key, or locating notes on the instrument. However, that same lesson revealed valuable information about focusing one's hearing on salient notes in the recording, and helped shed light on what was said in other videos.

Our understanding of the task is built from a mixture of descriptions and demonstrations that were provided by people teaching others how to learn by ear. This certainly colours the results such that our characterization of the by-ear learning process may be skewed towards beginners. However, our findings mirrored our observations of experienced musicians that demonstrated this skill in Chapter 4. That is, there is evidence suggesting our results are generalizable to non-beginner ear-learners.

A single DVD video (V001) covered the widest range of topics, and explained them most clearly. This could have been a function of its duration—allowing time for exposition—though much of that time was filled with descriptions of technology and repeated demonstrations. However, because this video was produced to be sold, it is conceivable that the lesson content was more deliberately planned; ensuring a high-quality product. It is therefore likely that studying better-prepared materials could have yielded additional data, and *newer* materials may have contained more extensive use of purpose-built technology. Therefore, a future study of lessons may be helped by considering other commercially available materials such as online courses.

Gender imbalance was evident in this study, though we do not believe this biased the results. For example, none of the women appearing in the videos introduced concepts that contradicted what was seen elsewhere. Only three perceptibly female musicians appeared across six of the videos—one taught alongside a male guitarist in four of them, another taught with a guitar on her own, and one taught a solo lesson on a piano. Because we consider the content of lessons that are most likely to be found by musicians seeking online instruction, we do not believe this element can easily be corrected for. However, it may be the case that musicians would continue looking further down the list of results to select teachers that they identify with, just as one might skip over lessons from guitarists who perform different genres. To confirm this, a future study may wish to crowdsource the selection of videos from a diverse set of musicians to see how that impacts the representation reflected by the lessons.

Guitar players were over-represented in this collection of videos, introducing certain biases into the data. For example, playing melodies is not prevalent in their practice except as a preparatory learning activity, whereas melodies are necessary for non-singing solo pianists that must integrate them into their performances. It would be easy to correct this bias in our study—discarding the V0xx and V1xx videos, and ensuring that we capture an equal number of eligible videos from each instrument. However, this bias towards guitarists does not appear to impact our results; none of the categories referred only to guitar (V0xx, V1xx, V3xx) videos.

5.5 Summary

From an analysis of 29 lesson videos, we further developed our understanding of the process musicians follow to learn songs by ear from recordings. Specifically, we mapped out the by-ear learning task as a set of sub-tasks that allow musicians to learn songs from recordings when they are followed in (a weakly ordered) sequence. We also discussed how these sub-tasks differed between lessons—some requiring the use of music theory, for example.

Using what we learned from the neuroscience and psychology research in Section 2.4, we could identify where memory plays a role during the by-ear learning process—e.g., tonal working memory allows musicians to compare a notes on the instrument with a remembered target pitch. Additionally, we found that teachers sometimes acknowledged the connection between memory and learning by ear. For example, some teachers advised working in short segments to account for a beginner's capacity—one note at a time, until the student can remember more than that.

Our findings helped explain differences in by-ear learning that we observed in Chapter 4. For example, some musicians in the prior study continued singing pitches from the recording while seeking the matching note on their instrument. In the present study, we saw teachers recommending this approach as well as those where a note is not sung again once it is remembered. This technique does not appear to engage tonal working memory, and could be a useful one for those musicians who have yet to develop theirs.

We now have a much clearer picture of the by-ear learning task, a better understanding of the variations between its sub-tasks, and we have identified where memory plays a role in the process. Our findings reveal clear opportunities for designers to create or modify existing technology so that musicians can learn by ear more effectively. In Chapter 4 we will use what we learned to present a conceptual model of the by-ear learning task, and provide recommendations for the designers of purpose-built technology for musicians that learn from pop music recordings.

Chapter 6

Designing Technology Supports for By-Ear Learning

We have seen in Section 2.5.3 that researchers can extract a great deal of musical information directly from a digital audio recording. For example, we can locate points in time when notes are played, extract their pitches, estimate the key and harmony—we seem to have all the tools necessary to automatically transcribe a recording. However, we discussed in Chapter 2 how popular music has a long history of informal learning practices that include by-ear learning. Also, not every popular musician can read music. Even if they could, sheet music fails to capture many elements of a performance that musicians seek to copy from a recording (Section 5.3.1). Thus, it makes little sense for designers of by-ear learning tools to consider an approach that strives to produce sheet music.

As we saw in Section 5.3.1, by-ear learning is described as a path to *musicianship*. There, teachers claimed learning by ear is an important skill to develop, and leads to concrete benefits such as developing the ability to improvise. While improvisation may not necessarily be a goal for all popular musicians, it is a skill that demonstrates mastery of the instrument—playing notes that the musician conjures in their mind's ear (Section 2.4.2). Therefore, it is important that by-ear learning tools continue encouraging this practice rather than trying to automate it away.

Based on what we learned in Chapter 4 and Chapter 5, we claim there are opportunities for designers to improve upon the ear-learning musician's experience with novel human-recording interactions. In this chapter, we describe each of these alongside recommendations for technology designers to realize them. We provide Table 6.1 for an overview of our design recommendations and their basis in literature and in our studies.

§	Recommendations	Literature	Video Studies	Key Technologies
6.2	Surface most familiar songs to learn		§ 4.4.5, § 5.3.4	Estimated song familiarity (§ 6.2.2)
6.2	Facilitate active listening sessions	§ 2.4.5	§ 4.4.5, § 5.3.4	Focused playback interface (§ 6.2.2)
6.3	Restrict playback regions	§ 2.4.1	§ 4.4.4, § 5.3.4	Predetermined stopping point (§ 6.3.3)
6.3	Align regions to musical content	§ 2.4.1	§ 5.3.4	Onset and beat tracking (§ 2.5.3)
6.4	Instil memories of musical sequences	§ 2.4.2, § 2.4.3, § 2.4.4	§ 5.3.4	Time stretching (§ 2.5.2) Repetitive playback outside learning context (§ 6.4.3)
6.5	Extend note playback indefinitely	§ 2.4.1	§ 4.4.4, § 5.3.4	Granular or harmonic re-synthesis (§ 6.5.3)
6.6	Isolate salient notes		§ 5.3.3	Music source separation (§ 6.6.2)
6.6	Audition user-entered notes		§ 4.4.2	Synchronized playback of user recording or proxy note entries (§ 6.6.3)

Table 6.1: A summary of our design recommendations. Each is provided with links to their basis in the literature, and our video studies.

However, we will first synthesize what was learned in the two studies to present a conceptual model of the by-ear learning task. In doing so, we highlight the core activity that lies at the foundation of each sub-task that a musician carries out while learning a song by ear.

6.1 Modelling the By-Ear Learning Task

As we discussed in Section 5.4.1, learning a song by ear involves a mixture of discrete sub-tasks: identifying the song's key, its chords, and the individual notes in a melody, bass line, or solo. We also saw how these sub-tasks can be executed in different ways: some methods call upon music theory, and others do not; some require that the musician can sing and remember notes, and others do not. However, when we attempt to model these tasks, we discover a common thread that runs through them.

We begin by thinking about the sub-tasks as a set of *procedures* executed by musicians, wherein they consume input, and produce output. When viewed in this manner, each of the sub-tasks consume the same input (musical audio), but produce different outputs. For example, key-finding results in the song's key, and chord-finding produces chords played on an instrument; both begin with playback of the song recording.

6.1.1 Playing Melodies From Memory

We begin our modelling with one sub-task that is an exception to the above definition—playing a melody from memory, which—as we have seen in Section 5.3.4—is most often taught using well-known nursery rhymes. This procedure differs in that it technically has no input.[1]

As illustrated in Figure 6.1a, melody playing begins with the musician's memory of pitches. Then, they produce a set of audible notes on their instrument, and—if they choose—they can assign names to the notes they played. However, because this procedure lacks external input, the target pitches are *arbitrarily* chosen by the musician—they may begin with a random starting note on the instrument, or attempting to match imagined pitches based upon their ability to sing them.

[1] No input *at the time of performance*, that is. The underlying memory was formed years prior, with input that may have taken the form of a lesson from an early childhood educator or a family activity.

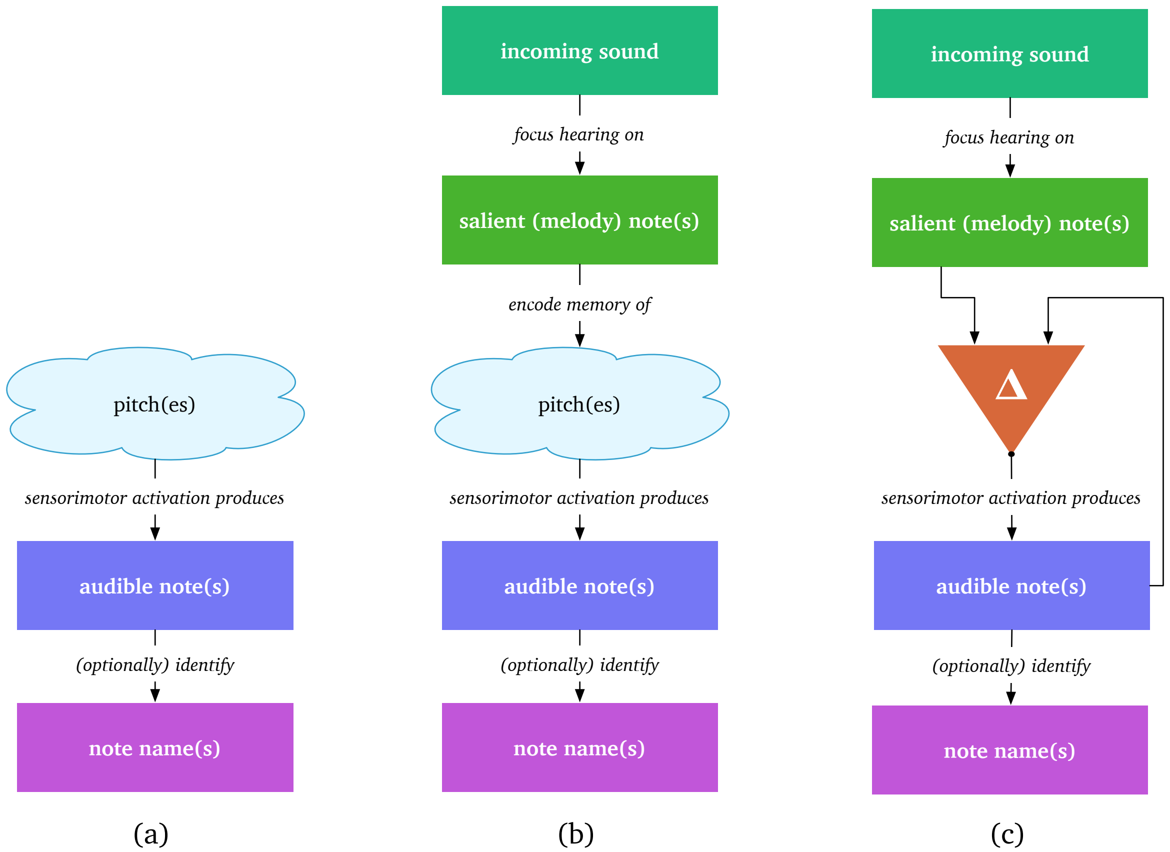


Figure 6.1: A conceptual model of (a) playing melodies from memory, and (b) copying notes from a recording. These diagrams make it evident why learning nursery rhymes from memory (Section 5.3.4) is a foundational exercise. In (c), we modify the note-copying model to not rely on tonal working memory (Section 6.1.6).

Starting from this set of imagined target notes, the musician calls upon sensorimotor activations to produce notes on their instrument. A musician with highly developed sensorimotor memories can produce the correct notes almost immediately—just as naturally as one might *sing* them, as we discuss in Section 5.3.4. By contrast, a musician who lacks such a skill will still rely upon the methods discussed in Section 5.3.3 to play each note.

6.1.2 Copying Notes From a Recording

Whether the musician is learning a vocal melody, a bass line, or a solo from a recording, our model of the procedure is identical (Figure 6.1b). It differs from a memorized nursery rhyme in that the musician is now listening for salient notes among what they hear in the recording, and encoding them as short-term tonal memories.

Isolating the salient notes from a whole-band recording is not always a trivial exercise, depending on the notes the musician is trying to copy. As we have seen in Section 5.3.3, musicians may have to practice focused listening, and ensure that their equipment can reproduce (or be adjusted to make prominent) the frequencies of the notes they wish to copy—a potential challenge when seeking low-frequency bass notes. When learning a solo, the musician may struggle to hear notes when they are played too quickly. Additionally they must be sensitive to the performance details such as vibrato, slurs, or the subtle differences between bending and sliding notes on a guitar that may be difficult to ascertain by ear. Musicians can use video footage of a performance to observe these techniques in use, provided it exists; when it does, the camera is more likely to be trained on the singer than any particular instrument.[2]

Once the salient notes have been identified, the musician now encodes them into tonal working memory (Section 2.4.1) so they can be reproduced in the same way that the musician did to play a nursery rhyme. Where this differs from the above procedure is that these memories have a limited capacity, short lifespan, and are easily disrupted by other pitches as we discussed earlier in Section 5.3.4 and Section 2.4.1. This is why the modelled procedure may have to be repeated by the musician for individual notes before they develop the ability to learn two or more at a time.

6.1.3 Identifying Chords

All of the methods that teachers described for finding chords began with the identification of bass notes, as discussed in Section 5.3.3. However, identifying these notes is no different from the procedure we described for copying them from a recording. In Figure 6.2, we see that this procedure is carried over in its entirety, except for one important

[2] As an illustrative example, see this live performance by Stevie Ray Vaughan at 0:43: https://www.youtube.com/watch?v=kfjXp4KTTY8#t=43s. He sings “She’s my pride and joy”, then responds with a guitar riff that is played entirely off-camera. By contrast, we have a full view of his picking and fretting hands during the solo at 1:44.

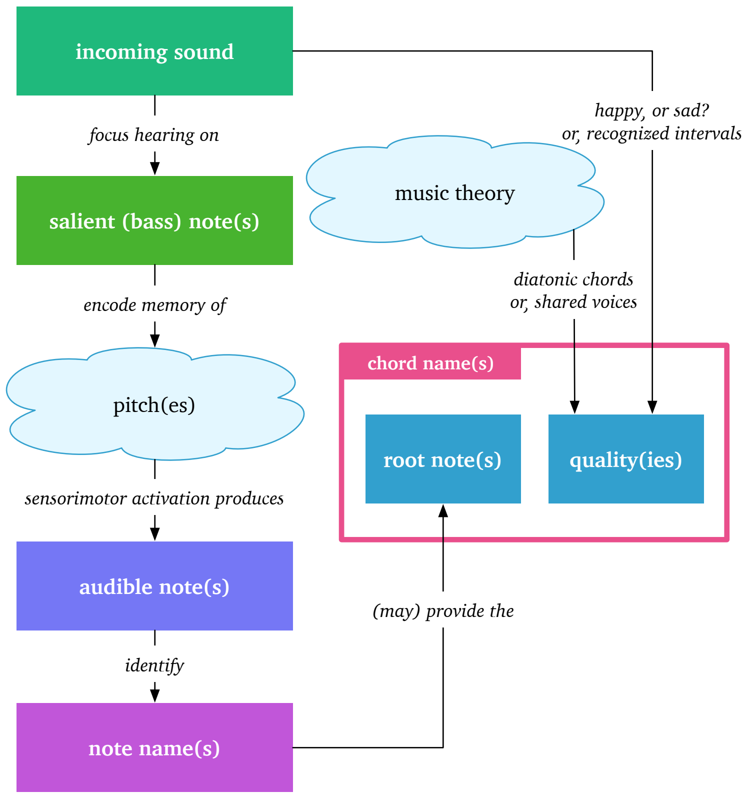


Figure 6.2: A conceptual model of finding chords from a recording, which is built upon the note-finding procedure shown in Figure 6.1b.

change: the musician *must* name the root note in order to identify the chord. All that remains is the identification of the chord's quality.

We saw teachers explaining both theory- and listening-oriented methods for doing this in Section 5.3.3, all of which are captured in Figure 6.2. The theory-oriented methods use the song's key (identification is discussed below), or knowledge of the individual notes (voices) in chords that are common with melody notes that were identified earlier. By contrast, the methods that don't rely upon music theory require that the musician consults the incoming sound (or perhaps a rich aural image stored in memory) to determine the chord's quality—either listening for a *happy*- or *sad-sounding* chord, recognizing the intervals, or (not pictured) playing candidate chords with the same root note—e.g., major and minor variations—to choose the one that sounds correct.

6.1.4 Identifying the Song's Key

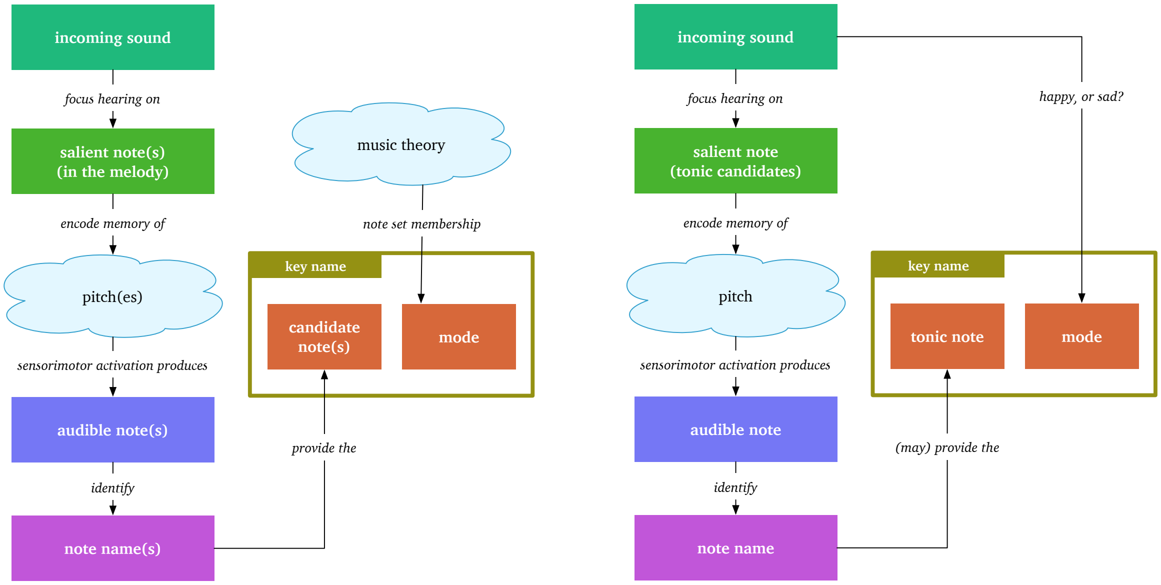
We have seen key-finding methods in Section 5.3.3 that—like chords—are varied in their approach. In Figure 6.3 we see models of two such methods: one that relies upon music theory, and one that does not. Again, at the left of both diagrams we see the same note-finding procedure we discussed above. In the theory-based method (Figure 6.3a), the musician is seeking a collection of melody notes, which must all be named. From this list of names, and the musician's knowledge of the notes belonging to each key, they can identify the key of the song. By contrast, the *intuitive* method (Figure 6.3b) requires that the musician listens for the note that sounds like the tonic of the key, and names it. Then, they consult the recording (or their memory of it) to gauge whether the song has a *happy* or *sad* sound to it to determine the key's *mode*—whether it is major or minor. Alternatively, they can test both major and minor key variations by playing each scale alongside the recording to test which sounds better (not pictured).

Note that we have also chosen not to picture two methods that are nearly identical to what we describe above, except the musician begins with one chord that sounds like the tonic, or a set of chords that they use to test key membership. In the method that begins with the tonic, it has the benefit that the quality of the chord also hints to the mode of the key—e.g., an F major tonic chord indicates the key of F major. In both cases, chord-finding works exactly the same as we discussed above: starting with the same note-finding procedure that yields bass (root) notes.

6.1.5 Isn't It All Just Learning Melodies?

It should be clear by now that each of the sub-tasks for learning a song from a recording begin with the same activity: copying individual notes. We cannot claim that one can learn chords as long as they can learn a melody from a recording. However, we would feel confident arguing the converse—that one has little hope to learn chords, or the song's key using the above methods if they are unable to learn melody notes from a recording.

We also described how learning notes from a recording is itself built upon the foundational task of playing a melody from memory. That is, an in-memory representation of notes is transformed into an attempt to play them on the instrument. However, this does not lead to the same kind of argument as above. That is, while playing notes from memory does not guarantee the ability to learn from a recording, we have seen that one *can indeed* learn melodies from a recording without having to play notes from memory.



(a) (b)

Figure 6.3: A conceptual model of learning the song's key (a) using a method that relies upon music theory, and (b) using an *intuitive* method that seeks both the tonic and quality by ear.

6.1.6 Learning Without Tonal Working Memory

We know from Section 2.4.1 that not everyone is capable of remembering or reproducing pitches accurately, so it stands to reason that some musicians may not possess this ability. We also see in Section 5.3.4 that some of the taught methods—singing while the note is located, or locating the note while the recording plays—do not appear to call upon tonal working memory. Additionally, we have observed experienced musicians using such strategies in Section 4.4.4. However, we do not know whether these musicians are compensating for a tonal working memory deficiency, or they are simply following the procedure they were taught. Regardless, they all demonstrate by-ear learning is still possible without engaging with their tonal working memory.

In the models discussed above, we have assumed that musicians have the tonal working memory skills to remember one or more pitches, and compare them to notes played on the instrument. However, we can eliminate this requirement by replacing the note-finding procedure in all the models with the one pictured in Figure 6.1c. In this diagram, the triangle represents a *difference operation* that comprises the comparison between the salient notes that are to be learned, and those notes the musician plays on their instrument. Effectively, the musician's goal is to eliminate the difference they hear between those two pitches, though the musician must *hear* the salient notes for long enough that they can close this gap. We further discuss this challenge in Section 6.5, when we present technological supports that are designed to aid musicians who lack the requisite skills, or prefer to work this way.

While we can see that it is *possible* to learn songs by ear without tonal working memory, musicians using this method are limited to learning only one note at a time. Those without the requisite skills may become frustrated working at this pace, and unlikely to learn by ear. To improve the experience for these musicians, we discuss tools for measuring and developing tonal working memory in Section 6.7.

6.1.7 Summary

Here we have presented a collection of simplified conceptual models that help characterize the by-ear learning task as we observed it being demonstrated and taught across two studies. Additionally, we illustrated how the sub-tasks relate to one another, and concluded that finding individual notes in recordings lies at the root of all the sub-tasks of by-ear learning. Further, we discussed how musicians can still execute all the sub-tasks without relying upon tonal working memory: by substituting a modified version of the note-finding sub-task.

We have revealed a clear opportunity for technology designers who can exploit these insights to create novel human-recording interactions. Specifically, designers should aim to allow users to effectively *normalize* each of the above recording-driven sub-tasks to one of learning melodies. That is, we should strive to make each sub-task equally challenging in terms of the extraction of salient notes, storing them in working memory, and offering aids for those who are unable to do so.

Designers may also strive to take a further step, and try to usher users towards a level of *familiarization* with recordings that leads to the formation of long-term memories like those that allow them to play *Twinkle, Twinkle, Little Star*. Specifically, aiming for musicians to remember the salient notes, eliminating the need for musicians to engage so intensely with recordings while trying to learn notes one at a time—ultimately making a pop song as easy to learn as a nursery rhyme.

6.2 Technology for Cultivating Familiarity

When musicians need to learn a song—regardless of whether it is new to them, or a longtime favourite—they should first dedicate some time to develop familiarity with it. By doing so, it provides a sort of *orientation* to the song's content. At the time of learning, musicians can recognize the general structure of the song, and anticipate which parts come next.

We suggest that the designers of by-ear learning technology should consider how they might leverage existing familiarity in their users, or allow users to cultivate their familiarity with a song intentionally.

6.2.1 The Importance of Familiarity

Musicians develop familiarity naturally through the processes of *hearing*, *distracted listening*, or *attentive listening* that occur during the musician's enculturation (Green 2017, p. 24). That is, becoming familiar with songs while hearing them in the background at the grocery store, noticing them on the radio, or listening to them for enjoyment.

However, in Chapter 4 we observed experienced musicians making a point to demonstrate by-ear learning with songs they had not heard before. By contrast, none of the teachers we observed in Chapter 5 characterized learning by ear in this way. Rather, we saw teachers suggesting that students develop familiarity with a recording intentionally. This *active listening* exercise we saw in Section 5.3.4 differs from what Lucy Green (2017, p. 61) calls *purposive listening*. The latter characterizes the listening activity that comprises the ear-learning session, often accompanied by an instrument; the former lies somewhere between Green's attentive and purposive listening. Specifically, active listening is meant to develop the musician's *contextual understanding* of a song, which—according to the bass player in V402—allows them to "anticipate what comes next". This may be related to the structural understanding that develops in their memory with repeated listenings (Section 2.4.5).

From our observations in Section 4.4.5, experienced musicians might occasionally challenge themselves to learn songs without first developing familiarity with them. Alternatively, these musicians might encounter real-world scenarios that do not allow them the opportunity to do so. However, we have also seen evidence in Section 4.4.5 that—despite appearing to possess sufficient skill to learn by ear without doing so—experienced musicians might choose to develop familiarity intentionally to gain an advantage while learning.

In addition to the development of a contextual understanding, we hypothesize that intentional, repeated listening sessions with a song improves the musician's ability to learn it because of the associative nature of long-term memory (Section 2.4.5). That is, active listening sessions help the musician strengthen their existing memory of the song by cueing its recollection. When the musician later attempts to learn the song by ear on their instrument, hearing short segments of the recording cues the activation of their long-term memory; this increases their their short-term memory capacity by providing the necessary association for chunking to occur (Snyder 2014).

6.2.2 Measuring and Developing Familiarity

Familiarity may be a neglected element of by-ear learning in technology products, likely because it is easy to assume familiarity has already been established. Therefore, we see a ripe opportunity for software designers to provide a unique advantage for their users who learn by ear. Technological supports can take many forms, but here we discuss two possible directions.

First, designers can try to estimate a user's existing familiarity with songs. For example, using the number of times the user has played a specific song, software can gauge their familiarity in relation to other songs. A list of recommended songs could be provided to users based on estimated familiarity scores, or users could be discouraged from learning songs with low scores. Unfortunately, this technique is only possible if the available interface to a user's music library surfaces this information; for example, Spotify's web API[3] only provides an estimate of *global* popularity, but Apple's API offers access to the number of times a track has been played.[4] However, even if it were easy to access this data, counting a user's exposure from the play count is likely to capture many *distracted listening* (Green 2017, p. 24) events that may not contribute to one's familiarity with a song.

The second approach instead encourages, or facilitates, the *active listening* sessions we saw described by teachers in Section 5.3.4. At the most basic level, designers can simply communicate the importance of active listening sessions in their user interface. However, designers might instead choose an extreme approach: enforcing that songs are

[3] https://developer.spotify.com/documentation/web-api/reference/get-track

[4] https://developer.apple.com/documentation/mediaplayer/mpmediaitem/1621694-playcount. However, this is limited to their iOS platform, and it appears to increment only when the user completes playback of a song in its entirety.

heard in their entirety before users can access song-learning features. Naturally, the latter approach risks irritating (and losing) users, but a well-designed feature can strike a healthy balance between the two. What we recommend is that designers simply aim to make active listening *accessible* in the moment when a user's intention is directed towards learning the song. The design goal would be to motivate users to participate—especially those who might otherwise skip this step—and make the experience enjoyable. For example, the user interface could provide coaching (e.g., recommending the instrument is set aside; a distraction-free space is sought), calming visuals, and present the user with statistics (e.g., time spent listening; number of sessions with this song) at the end of each session. Across repeated attempts to learn the song, users should be encouraged to revisit these active listening sessions so that their long-term memory of the song is reinforced (Section 2.4.5).

Both approaches would be improved by a mechanism to directly assess the user's familiarity with the song they wish to learn. For example, if a user could be *quizzed* about the content of the song they are to learn, it would help establish their level of familiarity and provide metrics that show improvements with additional listening sessions. Designers may be able to achieve this result by asking users to restore a rearranged version of the song to its correct order, or asking users to complete a sequence using one of multiple choices of audio clips. However, future study is required in order to determine the effectiveness of either method for evaluating one's familiarity with a song.

6.2.3 Summary

Whether they have fond memories of songs that drove them to learn to play their instrument, or a catchy song heard on the radio, it seems reasonable to assume that most songs are already familiar to musicians who wish to learn them by ear. However, we have showed that musicians believe this is not enough, and it is beneficial to develop familiarity with songs *intentionally* before learning them, and obtain context through active listening sessions.

Our recommendation to designers is that purpose-built software should not only attempt to estimate a user's existing familiarity with a song, but proactively *foster* the user's active listening practices. By making it easy for users to deeply engage with songs before (and during) their attempts to learn them, designers can potentially improve the user's ability to learn songs by ear.

6.3 Technology That Interfaces With Working Memory

Regardless of their by-ear learning proficiency, musicians must work within the limits of their memory while they learn from a recording. Specifically, they need to consider the capacity, duration, and delicate nature of their tonal working memory while they execute the foundational task of copying notes from a recording.

Here we recommend that technological tools offer a set of playback controls that are designed in a way that matches our understanding of the by-ear learning task, and tonal working memory. That is, we want to offer musicians a set of tools that provide them with the best chance of remembering the notes they heard while seeking the notes on their instrument.

6.3.1 Working in Bite-Sized Pieces

As we have discussed in Section 2.4.1, tonal working memory has limited capacity. We also found that teachers recognize this limitation (Section 5.4.3)—for example, by suggesting students learn only one note at a time in the beginning, or that they take “little bites” (V505) of the recording.

With time, teachers expect that students will experience an increase in their capacity, aligning with research from Ding et al. (2018) (discussed further in Section 2.4.1) that showed musicians could remember longer sequences of notes than non-musicians. However, we also saw in Section 5.3.4 that teachers themselves encountered limits to their memory in demonstrations, so we can see clearly that this progression has a clear upper bound.

Based on what we have seen, musicians can only *ingest* few notes from a recording at a time. Additionally, the number of notes is variable—not all musicians need (or want) to progress through a recording note by note.

6.3.2 Tonal Working Memory Disruption

We know notes do not last very long in short-term memory (Section 2.4.1). However, it appears this is *especially* true when musicians try to find them on the instrument, as we saw in Section 5.3.4. Specifically, as the musician hears the pitches of their incorrect attempts to play it, their memory of the target must be refreshed. This problem is likely a consequence of short-term pitch memories getting disrupted by the sound of new pitches, as we discussed in Section 2.4.1.

However, it is not only the pitches that emit from the musician's instrument that cause disruption, but also the pitches of additional notes heard in the playback of the recording. That is, audible pitches that continue to play after the target note(s) may negatively impact the musician's tonal working memory. In Section 5.3.4, we saw teachers either stating or demonstrating that note-finding should occur exclusively while playback is stopped—ideally before superfluous notes can be heard.

Musicians cannot predict how many attempts it will take to find the correct note on their instrument, and certain notes may prove more challenging than others to copy; such disruptions are inevitable. However, if they can listen only to the note(s) they wish to find, musicians can avoid disrupting their short-term tonal memory prematurely.

6.3.3 Memory-Friendly Playback Controls

Because tonal working memory has a limited capacity, and is easily disrupted, playback technology must be adapted so that users can control the flow of tonal information. That is, designers should ensure that users can avoid becoming overwhelmed while they are learning—playback must not start before, or proceed beyond those notes the user wishes to hear.

Fortunately, technology already allows precise starting and ending points to be defined, as users can mark such locations in purpose-built applications like the Amazing Slow Downer and Transcribe!. By placing these marks, playback only allows the region of audio between them to be heard.

Unfortunately, this affordance—as it is most often implemented—has two significant flaws. First, the user is required to identify the precise locations where playback is to be started and stopped, and—once the note(s) are learned—the user must re-define these locations for each subsequent (group of) note(s) they wish to learn. This procedure is rather *fiddly*, in practice. The second flaw is that these locations are represented as points in time (e.g., 1:09.53), which fails to represent the musical content captured by the recording.

Ideally, designers should offer musicians a mechanism to navigate recordings in a way that is musically significant. That is, in addition to time-oriented movement specified in minutes and seconds, designers should also allow musicians to place cursors or set regions using units that more closely match their mental model of a song—thinking in terms of measures and beats, sections, or individual notes, for example.

To provide these facilities requires the kind of musical understanding that we discuss in Section 2.5.3. For example, beat tracking provides the location of each beat in the recording, and would allow designers to offer navigation that is aligned to this unevenly-spaced *grid* of points. At a finer scale, software designers may wish to also offer access to an even finer grid of points that are aligned to a subdivision of the beats, or individual note onsets.

Armed with these capabilities, users would be able to define regions of playback in a way that is musically relevant—for example, starting at the second beat of bar 12, and ending three notes later. Similarly, they would be able to shift these playback regions such that the start and end always align to note onsets, effectively maintaining the number of notes contained in each playback region.

6.3.4 Summary

While copying notes from a recording, musicians need to work in segments that are sized to correspond with their tonal working memory capacity. Additionally, it is important that musicians hear only the note(s) they wish to learn so that this short-term memory is not disrupted by excessive pitch information.

Designers can help users by allowing them to learn within the limits of their tonal working memory, specifying precisely how much of the recording is heard when the user initiates playback. To improve upon existing methods, designers can leverage techniques from MIR to provide users with a musically meaningful way to define these segments. Doing so reduces the burden on users—it eliminates the need to make fine adjustments in the software, and allows the user to remain focused on the foundational task of copying notes from the recording.

6.4 Technology to Facilitate Playing From Memory

Teachers only seem to ask students to play melodies entirely from memory before learning to play songs from recordings (see Section 5.3.4), but it is never suggested to do so afterwards. However, some musicians can learn songs from a long-lived memory of a pop song recording, and it appears that those who do so struggle less with the limits of tonal working memory.

Here we discuss suggestions for the design of tools that facilitate this method of learning. Specifically, we explain how designers can turn segments of pop songs into long-term memories using techniques inspired by the neuroscience and psychology literature found in Section 2.4.2.

6.4.1 The Nursery Rhyme Exercise

In Section 5.3.4, we saw five teachers describing an exercise that asked students to play nursery rhymes from memory. One teacher noticed a connection between humming nursery rhymes and learning to play solos, which helps explain why this exercise is foundational for learning by ear.

> So what's the difference in humming a tune and figuring out licks off of CDs? Absolutely nothing, really. Humming London Bridge is humming a tune that you've already learned in the past, and a hot solo on a CD is a tune that you're fixing to learn in the future. You learned London Bridge on your vocal cords, and you're learning your favourite CD on your guitar. Or maybe, to be honest, you're really going to learn it on your vocal cords, and then transfer it to your instrument. (V001, 12:28)

Because it appears first, teachers must consider this task easy enough for beginners to understand, and that their students are likely to achieve a successful result. Additionally, their common choice of *Twinkle, Twinkle, Little Star* has a widely-recognized melody (to English language speakers—see, e.g., Humpal 1998; Looi et al. 2003) and does not require students to prepare before making attempts of their own. Further, because no interactions with recordings are necessary to complete this exercise, a student need not worry about their lack of familiarity with (or access to) purpose-built tools that can facilitate the repetitive listening required to refresh their memory.

However, we find it curious that this exercise is relegated to the beginning of one's journey of by-ear learning. That is, we did not observe teachers recommending that students continue this practice afterwards, or adapt it for different material. Do teachers think that people are unable to remember the melodies of pop songs? Research suggests otherwise, as we have seen in Section 2.4.2, Section 2.4.3, and Section 2.4.4.

6.4.2 Playing Pop Music From Memory

We encountered two piano teachers in Section 5.3.4 who demonstrated they could learn a pop song using only long-held memories of its recording. In both cases the teachers appeared to have developed the same kind of memory that one might have of a nursery rhyme.

The primary difference between pop songs and nursery rhymes is that a commercial recording often contains a full band, with singing combined with one or more instruments. By contrast, a nursery rhyme was likely learned during early childhood (Humpal 1998), as an unaccompanied melody sung by (and possibly along with) a teacher. However, despite the *richness* of information in a pop recording, we know from our survey of the literature in Section 2.4.2 that such auditory memories can be formed intentionally. Additionally, we know that these memories might even allow the correct key from the recording to be recalled (Section 2.4.3).

We can see from the two piano teachers that the salient melody notes can be readily accessed from these long-term auditory memories, and the teachers could also recognize whether or not chords were correct. Also, despite the nature of tonal working memory—that it has a limited lifespan, and is easily disrupted—the short-term memory of the melody notes can seemingly be restored with ease from the long-held auditory memory. Therefore, it appears rather robust, and notes are unlikely to be lost while a musician locates them on the instrument.

Despite not seeing this method used in other videos, working entirely from memory seems like a practice worth replicating. That is, internalizing the vocal melodies of popular songs, or the notes in a solo, to a point that they can be recalled as readily as nursery rhymes. To accomplish this kind of practice, musicians would need a way to form memories of pop music recordings that persist for much longer than their short-term memory can hold it.

6.4.3 Forming Long-Term Memories

As we have discussed in Section 2.4.2, instilling such memories is possible; this is demonstrated by Kubit & Janata, (2022) who induced INMI in their participants to improve their long-term memory of musical sequences. To achieve the same outcome, designers can provide users with a mechanism to create bespoke *earworms* that will help them form long-term memories of segments from songs they wish to learn.

Designers should ensure that users can define segments that are aligned to the locations of beats (as discussed earlier in Section 6.3.3) so that loops play seamlessly, and in time with the original recording. Additionally, designers may want to consider generating these segments automatically—for example, from a verse, chorus, or solo—by employing techniques such as music structure analysis and segmentation (see, e.g., McFee & Ellis 2014a; Nieto & Bello 2016).

Once segments are defined, designers should make them available in contexts where users would normally listen to music—for example, while doing chores, or commuting. However, their playback should employ the methods we described in Section 2.5.2 to play the segments on repeat at a slower rate. Such a modification may help users remember complex passages from a solo that are played too quickly for them to sing or remember in full. Ideally, designers should strive to make learning these phrases as easy for users to remember as nursery rhymes.

Designers may want to consider additional steps to prevent playback from becoming tedious for users to listen to—for example, limiting the amount of repetition. Designers can also inject novelty into the playback in a way that may help improve the users memory by encouraging recall. Using a technique used by Kubit & Janata (2022) to test their participants' memory of musical sequences, designers can periodically insert brief periods of silence that causes the user to generate imagery for missing portions of the segment while anticipating what comes next. By also using knowledge of the beat locations in the original recording, these periods of silence can be inserted such that they preserve the rhythmic context of the loops.

6.4.4 Summary

Musicians that are asked to play nursery rhyme melodies entirely from memory appear to face little struggle with their tonal working memory capacity. Ideally, musicians may want to try and replicate this practice with melodies or solos learned from pop recordings.

To facilitate this activity, we recommend that designers consider offering users a way to listen to repeating segments of the pop recording so they can proactively form long-term memories while they are not actively learning from the recording. Additionally, by manipulating complex segments such that they play more slowly, designers can offer a method that makes learning solos as easy as playing a nursery rhyme from memory.

6.5 Assistive Technology for Retaining Notes

Not everybody can accurately recall a single pitch; those who can, remember it for a short time. However, a pitch must be retained in memory long enough that a musician can locate the note on their instrument. Continuing to sing the note appears to help compensate for this, but it is not always possible for musicians to do so.

Here we explain how designers can provide tools that help their users who experience these limitations. Specifically, we discuss how a pitch can be transformed from one that is briefly audible into a pitch that can be heard indefinitely while the user seeks the note on their instrument.

6.5.1 Pitch Working Memory Limitations

In the tonal working memory studies we discussed in Section 2.4.1, not all participants exhibited perfect scores when they were asked to remember a single pitch. However, teachers in video lessons often assumed students could do this, as we saw in Section 5.3.2. Specifically, teachers assumed that students could recognize whether a pitch they heard moments before was played correctly on their instrument.

To identify whether a note played on the instrument is correct, the target pitch must be held in the musician's tonal working memory (Section 2.4.1). Effectively, the musician needs to recall the aural image of the correct pitch so that it can be compared to what they are hearing from the instrument. When they are unable to do so, musicians may try to reproduce the target pitch vocally while comparing both of the sounds they are hearing at once, as we have seen in Section 5.3.4.

6.5.2 Vocal Pitch Imitation Deficiency

Unfortunately, the literature we discussed in Section 2.4.1 suggests that people who lack a strong tonal working memory may also be unable to sing pitches accurately, and vice versa. Additionally, a large minority of the population may lack the ability to reproduce pitches accurately using their voice. Such limitations in singing and tonal working memory abilities would certainly make it challenging for musicians to learn from recordings by ear.

We *also* noticed in both of our studies that saxophone players had no such luxury to sing a note continuously, regardless of their singing ability—their breath was necessary to produce notes on the instrument. Combined with those who cannot accurately imitate pitches with their voice, there is a segment of musicians who are unable to compensate for a lack of tonal working memory capacity.

Fortunately, we saw in Section 2.4.1 that both musicians and non-musicians seem to be adept at accurately reproducing pitches *mechanically*, provided the target pitch is still audible. That is, people can make adjustments to an instrument-like device and match a pitch they are hearing at the same time.

6.5.3 Assisted Note Retention

Based on our findings, we recommend that system designers offer playback facilities that allow musicians to hear pitches for longer durations than they are audible in the original recording. That is, users must be given ample time to hear the pitch while seeking its match on their instrument. Such a feature has uses beyond that of compensating for a user's limited tonal working memory or inability to sing. By increasing their duration, notes played quickly can be held *frozen* while the user takes their time doing any of the following: identifying the note that is salient, singing it for retention, or locating the note on their instrument.

Time stretching functionality is already available, but offers only limited help in this regard. For example, a sixteenth note played at a tempo of 120 bpm, would still only be audible for 500 ms when slowed to a quarter of the original speed—often the slowest available playback rate. This provides very little time for the musician to compare what they are playing to what they hear in the recording. Looping playback around a single note helps, but the repeated *attack* of the note onset creates audible artifacts that may distract from the pitch itself.

Instead, we recommend that designers try to *synthesize* the note(s) from the recording, allowing the user to hear them indefinitely. That is, capturing the *timbre* of the recording during that instant of time, and generating a new sound that maintains the characteristics of the original. This can be achieved in many ways: using granular synthesis (Schnell & Schwarz 2005; Schnell et al. 2000), or by synthesizing sinusoids (Serra & Smith 1990). The latter method, or one that extracts harmonic elements from transients (e.g., Juillerat & Hirsbrunner 2017) can help reduce the impact of un-pitched percussive sounds that may *muddy* the spectrum. Of course, each approach has its strengths and weaknesses that the designer must weigh. For example, identifying and synthesizing sinusoids discards timbral information from the original recording that the user may still wish to hear.

We also recommend that designers combine this approach with the navigation enhancements we discuss in Section 6.3.3. Specifically, allowing users to advance note-by-note, or by using subdivisions of the beat grid.

6.5.4 Summary

As we have seen, musicians may struggle to learn songs by ear when they have a tonal working memory deficiency. Additionally, they cannot always compensate for this limitation by trying to sing notes while seeking them on the instrument.

What we recommend is that designers provide a mechanism for users to *hang on* to the sound of notes for much longer than they are heard in the recording, giving them ample opportunity to find them on their instrument. Using a mixture of synthesis techniques, and those that power the foundational time-stretching effect, designers can provide benefits for users that extend beyond compensating for their limitations, such as assisting with salient note identification.

6.6 Assistive Technology for Salient Note Identification

In a recording of music performed by a full band, it can be difficult to identify the notes that are salient to the musician. By *identification*, we refer to the act of recognizing one or more notes from among a full band's performance that the musician wishes to play, and not the playing or naming of those notes. For example, listening for notes played by a rhythm guitarist, whose role is to support the rest of the band—including a lead singer, who will be placed more prominently in the song's mix.

What we suggest here is that designers offer methods that help users achieve this aural identification. Specifically, we recommend that salient notes should be made audible in isolation, and that designers offer users a method to carry out the identification of these notes in a more systematic way.

6.6.1 Listening for Pitches Can Be Challenging

To learn music from a recording, a musician needs to hear which notes they need to play. We have seen in Section 5.3.3 that identifying salient notes in a full band recording may not be straightforward—because of timbral differences between instruments, and variations in mixing techniques. Additionally, notes that are played too quickly are not audible for long enough that the musician can concentrate on identifying the salient ones.

As explained by the teacher in V001, musicians may face challenges when trying to copy notes that are played on instruments different from theirs. For example, a guitar player trying to copy the notes of a saxophonist, or learning guitar chords to be played on a piano. These differences in *timbre*—the harmonic properties that characterize individual instruments—can pose challenges for musicians when attempting to recognize notes.

When a song is *mixed*, the audibility of individual instruments is fixed by the mixing engineer. They control the volume of each instrument, and use tools like equalization and compression to effectively *sculpt* their spectral contribution to the recording. These modifications, combined with psychoacoustic masking effects, can cause notes to seemingly *disappear* in the presence of louder sounds, or when different instruments play the same notes (Wichern et al. 2016).

6.6.2 Isolating Salient Notes

In Section 5.3.3 we saw teachers recommending that students practice focused listening—placing their attention on specific instruments, causing others to seemingly blur into the background. We also observed that students were told to sing the salient notes they hear the recording, presumably to extract the notes so that the student could confirm what they heard was correct.

In a perfect world, musicians would be able to access the original *stems*, or *tracks* from the mixing desk (or software) that was used to produce the recording they are learning from. That way, they could lower the volume of, or mute, all except for the instrument playing the notes they are trying to copy.

Today, such control is possible using music source separation, which aims to *de-mix* the audio into individual stems (e.g., Nakano & Goto 2023; Rouard et al. 2023; Schaffer et al. 2022). While these techniques are impressive, they are still imperfect. For example, they are often trained only to separate a limited subset of the instruments—vocals, bass, drums, and "the rest of the band"—and there are significant barriers to obtain sufficient data to train state-of-the-art models (Pereira et al. 2023). For example, one would require access to a large number of original multitrack recordings from the commercial studios that produced them. One could conceivably commission the production of music for such a training data set, but it is unlikely to be large enough, or capture the variety of mixing and mastering techniques that have evolved over many decades.

6.6.3 Auditioning Salient Notes

Designers may wish to consider a more straightforward alternative that allows users to extract salient notes themselves. That is, the user would be provided with a mechanism to produce the salient notes in isolation, and compare their own notes with the recording.

Such a feature could work as follows. The user marks a segment of the recording with one or more notes that they find difficult to isolate. Then, the software alternates between playing the segment, and recording the user singing the notes they heard. Once they think they have the notes, they can disarm the recording, and playback will continue to alternate between the original, and the recording of the musician's singing. The user may continue to arm and disarm the recording function until they feel they have worked out the correct notes. To facilitate this, designers must ensure that the user's recorded interpretation of the notes is synchronized with the original.

If they cannot, or choose not to sing, the musician could instead produce notes using their instrument. However, this requires that they are already competent enough to play what they hear, and hence are much further along in their learning. For these musicians, this feature is still quite useful—it allows them to check that what they have played is correct.

When musicians can neither sing nor play the notes on their instrument, designers should instead offer a *proxy* that allows musicians to produce audible notes mechanically. This is similar to an activity we saw in Chapter 4, where musicians created sheet music in software, and used a built-in sequencer to play their notation entries and evaluate whether they sounded correct. However, designers should ensure that the entries are synchronized with the recording when they are played. For example, users could place notes on a timeline that is defined by the recording. During playback, those entries would be made audible in unison with the recording, and the user would have an opportunity to alternate between hearing the two.

6.6.4 Summary

Musicians sometimes face recordings that make it difficult to hear the notes they want to learn. Through focused listening exercises, and by singing what they hear, musicians can improve their chances of identifying these tough-to-hear notes.

Designers can improve the experience of users by making the salient notes audible in isolation, or by offering an interface that helps users identify salient notes themselves. By alternating between playback of the recording, and those notes the musician has produced—with their voice, instrument, or using a proxy—users could hear their interpretation of the notes juxtaposed with the recording, and judge whether they captured the correct ones.

6.7 Developing Foundational Skills

Just as there appears to be a foundational task sitting at the core of the by-ear learning sub-tasks, there is a core set of skills that musicians should have in place before they can copy notes from a recording. Specifically, musicians need to know when a pitch matches those heard in the recording, and they should also be able to accurately match a pitch using their voice.

We see an opportunity for designers to create tools that help users develop these core skills. Whether they exist as standalone products or they are integrated into existing ones, designers can usher prospective users towards learning by ear, or improve the skills of those who already use their products.

6.7.1 The Foundational Skills of Budding By-Ear Learners

In Section 5.3.2 we saw that teachers often assumed their students possessed certain prerequisite skills. Among these skills were recognizing when a note played on the instrument matched the pitch heard in the recording, or being able to reproduce pitches by singing or humming them. These skills are foundational to the task of learning by ear, because they are necessary when copying notes from a recording—a foundational element of the by-ear learning sub-tasks (Section 6.1.5). However, we know from Section 2.4.1 that not everybody possesses these skills. Either people lack them completely, or there is room for improvement. For example, some people may only reproduce pitches vocally with middling accuracy, or they can only remember one pitch at a time.

The need to reproduce pitches vocally is not called upon once a musician has a well-developed tonal working memory. For example, we have observed in both our studies that musicians could seek notes on the instrument without singing them at the same time. However, we know from Section 5.3.4 that studies suggest tonal working memory and vocal pitch matching are related skills. Therefore, it is plausible that by developing someone's ability to reproduce pitches vocally, it could improve their tonal working memory—further improving the chance of success when learning by ear.

Note that reproducing a pitch accurately and singing well are two very different skills, and the latter is unnecessary for instrumentalists who wish to learn by ear. In some lesson videos, we observed teachers claiming they had poor singing voices, but they still demonstrated the ability to reproduce pitches, as we saw in Section 5.3.2. Additionally, the kind of pitch reproduction that is called upon while learning by ear does not require absolute precision. Specifically, it is acceptable to sing notes one or more octaves away from what is heard in the recording in order to fit one's vocal range, as we have seen in Section 5.3.4.

If musicians have deficiencies in pitch working memory or vocal pitch reproduction, they can benefit from what we proposed in Section 6.5. However, for them to most efficiently learn by ear, and also improve their musicianship, it is ideal that musicians work towards building these foundational skills.

6.7.2 Evaluating Tonal Working Memory Comparisons

To measure one's ability to identify that a tone is correct, we recommend that designers create tests inspired by the literature discussed in Section 2.4.1. In one kind of test, users would be played two tones separated by a delay, and asked whether the second matches the first. In the other kind of test, designers can present users with a target tone, then—after some delay—ask them to match its pitch.

In the simpler test configuration, designers are measuring the user's baseline ability to recognize that a probe tone matches the target. Consistent failures in this test would reveal that a user is unable to tell whether the note they played on their instrument matches the one heard in the recording, and should preclude the user from performing further tests discussed below.

The more difficult test is one that measures the user's tonal working memory *accuracy*—the user's ability to correctly match a target tone held in memory. We recommend that designers consider offering as many as three methods for matching a pitch in this way: mechanically, vocally, and using their instrument.

For mechanical pitch-matching, designers can offer users a virtual instrument with continuous pitch adjustments, measuring the difference between their entry and the target pitch. Vocal pitch matching would record the user's singing, using the same evaluation techniques researchers used in the studies we saw in Section 2.4.1. Finally, instrumental pitch matching would be largely identical, though its evaluation may need to be adjusted such that designers account for inevitable tuning deviations between instruments.

By executing the above tests, designers can help their users understand whether they can accomplish the critical task of knowing when they correctly matched a pitch from a recording. Designers may wish to integrate such critical tests in their ear-learning software, in a manner similar to the "you must be *this* tall to ride" signs found at amusement parks. If users consistently fail these tests, they could be encouraged to practice the requisite skills, or consider employing assistive methods like those discussed in Section 6.5.

6.7.3 Training Vocal Pitch Imitation

If a user demonstrates any deficiency in their tonal working memory, designers may wish to offer tools that allow users to practice matching pitches without asking them to remember anything. For example, designers could present users with a continuously playing target tone that the user is asked to match. By providing visual feedback, designers can indicate to users how close they are to the target pitch.

Based on what we learned from the studies in Section 2.4.1, training vocal pitch imitation could help users develop baseline tonal working memory skills. By creating the necessary sensorimotor connections to produce tones accurately, users who engage in this training may improve their ability to encode tonal memories.

6.7.4 Summary

Alongside the technology that aids a musician learning recordings, there are opportunities for tools that teach those core skills that by-ear learning requires. For example, one's ability to recognize if a pitch is correct or not, or to reproduce pitches by singing, humming, or whistling. While we do not know the prevalence of such limitations among musicians, it is worth considering the ways designers can help them reach the necessary baseline.

Specifically, we recommend that designers offer training for the core skills of by-ear learning: building a suite of tests that attempt to develop a musician's baseline level of tonal working memory, and their ability to reproduce a pitch using their voice.

6.8 Next Steps and Future Work

We have provided a set of design recommendations that are all based on the results of studies we conducted in Chapter 4 and Chapter 5, as well as the literature on memory that we discussed in Section 2.4. However, each of these recommendations call for additional study to validate and shape the design of what we have proposed.

For example, in Section 6.4, we proposed that designers should allow musicians to extract beat-aligned segments of songs that can be heard on repeat away from the learning context. In doing so, designers would foster the creation of long-term memories of notes. However, we are unable to offer guidance in terms of the duration of these segments, or how many repetitions are necessary to provide the desired result. While our proposed practice resembles the study design of Kubit & Janata (2022), the two are not entirely the same; participants heard sequences with the same duration, and they were not asked to *reproduce* what they remembered a week later. Therefore, product designers and academics should conduct studies that test how well musicians would perform in a by-ear learning context; researchers might ask musicians to learn segments of varying duration, and test whether they can reproduce notes from the segment on their instrument after some time has passed.

In most cases, our design suggestions are presented in the context of augmenting existing technology; e.g., our suggestion in Section 6.3 to limit the region of playback, or describing in Section 6.5 how a segment of the recording could be re-synthesized to play continuously. Testing the efficacy of these human-recording interactions would best be carried out by modifying existing tools that are used to learn by ear. In an academic setting, both new and existing users of the tool could be recruited to see how effectively they can learn notes from recordings. Such a study would certainly require participation from product vendors to allow the necessary alterations to their products, and fortunately such partnerships are commonplace in academia.

6.9 Summary

It is important that the designers of by-ear learning tools foster those musicians who choose to learn from recordings. Rather than evolving these products towards the direction of producing automatic sheet music, designers should instead continue to embrace this recording-driven method of learning.

We have shown how copying notes from a recording lies at the core of learning by ear. Musicians must do it while learning melodies, solos, bass lines, chords, and even the key of the song. The key implication of this finding is that designers should focus on making sure that their tools allow musicians to effectively copy notes—even if they lack the baseline tonal working memory skills. There is a clear opportunity for designers to optimize their technological tools by leveraging an understanding of the musician's memory. By doing so, designers not only help musicians remember notes in the recording, but also make their tools accessible to those who cannot (yet) do so.

In this chapter we have described six aspects of by-ear learning that can either be improved upon, or built in to all-new products and features. First, designers can facilitate the process of focused listening, helping users develop familiarity with the recording of a song. Second, we recommend that designers offer playback controls that consider the musician's limited working memory, and allow them to navigate the recording in musically meaningful ways. Third, designers can use ideas from the neuroscience and psychology literature to usher users towards remembering musical sequences over longer periods. Fourth, we explained how designers can help those users who have difficulty retaining and singing notes they hear in the recording. Fifth, we describe ways that users can be assisted during the process of isolating salient notes within a busy recording of a full band. Sixth, we describe some opportunities for designers to possibly develop the foundational skills that may be limited or missing in those musicians who wish to learn music from recordings.

All of our recommendations in this chapter are well-grounded in existing literature, and also our observations of lessons and experienced practitioners. However, each of these recommendations also provide direction for researchers and practitioners to conduct future work in the form of prototyping, and evaluating the effectiveness of these interventions through user studies.

Chapter 7

Conclusion

In this thesis, I have demonstrated the value of conducting an online video-based study to generate design recommendations. Specifically, the insight derived from two such studies can be applied today to improve upon the experience of those musicians who learn by ear from recordings (Chapter 2).

Video-based studies like ours are time-consuming—especially when they are conducted by very few researchers. However, the limited financial investment and ease of access makes such studies attractive for independent researchers or very small teams. In Chapter 3 we discussed the advantages, practices, and pitfalls of running an online video study, and methods that we and others have employed to save time. We also presented some of the unique ethical considerations that are necessary when conducting such studies.

By analyzing a collection of 18 videos obtained from YouTube (Chapter 4), we developed a baseline understanding of the ways musicians interact with recordings as they learn by ear, and generated hypotheses to guide future work. We found that musicians who produced sheet music did not actually use it while learning, intentional familiarization with a recording may help with learning them, and that musicians employed varying techniques for retaining the notes they heard in memory.

We then analyzed a collection of 29 lesson videos (Chapter 5) to help characterize the by-ear learning process, and explain the variations in technique that we observed in the first study. Our findings revealed differences in the way by-ear learning is taught, and we identified a collection of sub-tasks that comprise the by-ear song learning process. Using what we learned from neuroscience and psychology research in Chapter 2, we identify how the process of learning by ear calls upon a musician's memory, and discuss how some teachers acknowledge where memory plays a role.

In Chapter 6, we presented a conceptual model of the by-ear learning sub-tasks based on our findings, and showed that each of them are built upon copying individual notes from a recording. Effectively, by creating technology that facilitates this activity, designers can facilitate the whole process of learning songs by ear. Among our recommendations for designers, we suggest offering tools that develop a musician's familiarity with recordings, and describe how to assist musicians that have difficulty retaining notes in tonal working memory. All of our recommendations can be built today using existing techniques.

While all of our design recommendations are grounded in observations and research, *specific* designs must be implemented and studied to determine whether these interventions have the intended impact; ideally, in partnership with the companies who produce tools for musicians who learn by ear. Our work creates a number of possible directions for future research to evaluate their effectiveness.

Our design recommendations are driven entirely by real-world observations that came from user-uploaded videos on YouTube. Our studies generated hypotheses, triggered additional research, helped us understand by-ear learning, and revealed opportunities for designers to improve the musician's toolbox for doing so.

References

Lilach Akiva-Kabiri, Tomaso Vecchi, Roni Granot, Demis Basso, and Daniele Schön. Memory for tonal pitches: A music-length effect hypothesis. *Annals of the New York Academy of Sciences*, 1169:266 – 269, 2009. doi:10.1111/j.1749-6632.2009.04787.x.

Erman Altunisik, Yasemin E. Firat, and Yeliz K. Keceli. Content and quality analysis of videos about multiple sclerosis on social media: The case of YouTube. *Multiple Sclerosis and Related Disorders*, 65:104024, 2022. doi:10.1016/j.msard.2022.104024.

Lisa Anthony, YooJin Kim, and Leah Findlater. Analyzing user-generated YouTube videos to understand touchscreen use by people with motor impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1223–1232. 2013.

Bill Barich. Still truckin'. *New Yorker*, page 96, October 11 1993.

Ava Bartolome and Shuo Niu. A literature review of video-sharing platform research in HCI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, University of Minnesota, Minneapolis, MN, United States; Clark University, Worcester, MA, United States, 2023. doi:10.1145/3544548.3581107.

Corey H. Basch, Charles E. Basch, Kelly V. Ruggles, and Rodney Hammond. Coverage of the Ebola Virus Disease epidemic on YouTube. *Disaster Medicine and Public Health Preparedness*, 9(5):531–535, 2015. doi:10.1017/dmp.2015.77.

Corey H. Basch, Isaac Chun-Hai Fung, Rodney N. Hammond, Elizabeth B. Blankenship, Zion Tsz Ho Tse, King-Wa Fu, Patrick Ip, and Charles E. Basch. Zika virus on YouTube: An analysis of English-language video content by source. *Journal of Preventive Medicine and Public Health*, 50(2):133–140, 2017. doi:10.3961/jpmph.16.107.

Corey H. Basch, Grace C. Hillyer, Zoe C. Meleo-Erwin, Christie Jaime, Jan Mohlman, and Charles E. Basch. Preventive behaviors conveyed on YouTube to mitigate transmission of COVID-19: Cross-sectional study. *JMIR Public Health and Surveillance*, 6(2):e18807, 2020. doi:10.2196/18807.

Roger E. Beaty, Chris J. Burgin, Emily C. Nusbaum, Thomas R. Kwapil, Donald A. Hodges, and Paul J. Silvia. Music to the inner ears: Exploring individual differences in musical imagery. *Consciousness and Cognition*, 22(4):1163–1173, 2013. doi:10.1016/j.concog.2013.07.006.

Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1046, 2005. doi:10.1109/TSA.2005.851998.

H. Stith Bennett. *On Becoming a Rock Musician*. University of Massachusetts Press, Amherst, 1980.

H. Stith Bennett. Notation and identity in contemporary popular music. *Popular Music*, 3:215–234, 1983. doi:10.1017/s026114300000163x.

Magdalena Berkowska and Simone Dalla Bella. Acquired and congenital disorders of sung performance: A review. *Advances in Cognitive Psychology*, 5(1):69–83, 2009. doi:10.2478/v10053-008-0068-2.

Rachel M. Bittner, Magdalena Fuentes, David Rubinstein, Andreas Jansson, Keunwoo Choi, and Thor Kell. `mirdata`: Software for reproducible usage of datasets. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands, 2019.

Mark Blythe and Paul Cairns. Critical methods and user generated content: The iPhone on YouTube. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1467–1476. 2009.

Sebastian Böck and Markus Schedl. Enhanced beat tracking with context-aware neural networks. In *Proceedings of the 14th International Conference on Digital Audio Effects*, 2011.

Sebastian Böck, Florian Krebs, and Gerhard Widmer. A multi-model approach to beat tracking considering heterogeneous music styles. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 603–608. Taipei, Taiwan, 2014.

Katya Borgos-Rodriguez, Kathryn E. Ringland, and Anne Marie Piper. Myautsomefamilylife: Analyzing parents of children with developmental disabilities on YouTube. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019. doi:10.1145/3359196.

John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 633–638. Miami, Florida (USA), 2011.

Patricia Shehan Campbell. Of garage bands and song-getting: The musical development of young rock musicians. *Research Studies in Music Education*, 4:12–20, 1995.

Kathy Charmaz. *Constructing Grounded Theory*. Sage Publications, London, 2014.

Souti Chattopadhyay, Denae Ford, and Thomas Zimmermann. Developers who vlog: Dismantling stereotypes through community and identity. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33, 2021. doi:10.1145/3479530.

Tsung-Ping Chen and Li Su. Harmony transformer: Incorporating chord segmentation into harmony recognition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 259–267. Delft, The Netherlands, 2019.

Tian Cheng and Masataka Goto. Transformer-based beat tracking with low-resolution encoder and high-resolution decoder. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, pages 466–473. Milano, Italy, 2023.

Ji Young Cho and Eun-Hee Lee. Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *The Qualitative Report*, page 157, 2014. doi:10.46743/2160-3715/2014.1028.

Taemin Cho and Juan P Bello. A feature smoothing method for chord recognition using recurrence plots. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 651–656. Miami, Florida (USA), 2011.

Taemin Cho and Juan P Bello. On the relative importance of individual components of chord recognition systems. *IEEE Transactions on Audio, Speech and Language Processing*, 22(2):477–492, 2014. doi:10.1109/TASLP.2013.2295926.

Taemin Cho, Ron J. Weiss, and Juan P. Bello. Exploring common variations in state of the art chord recognition systems. volume Proceedings of the Sound and Music Computing Conference (SMC), pages 1–8, 2010.

Justin Clark. Systematic reviewing. In *Methods of Clinical Epidemiology*, pages 187–211. Springer, Berlin, Heidelberg, 2013.

Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States, 2008.

Kate Covington. The mind's ear: I hear music and no one is performing. *College Music Symposium*, 45:25–41, 2005. doi:10.2307/40374518.

Andy Crabtree and Tom Rodden. Domestic routines and design for the home. *Computer Supported Cooperative Work (CSCW)*, 13(2):191–220, 2004. doi:10.1023/b:cosu.0000045712.26840.a4.

Andy Crabtree, David M. Nichols, Jon O'Brien, Mark Rouncefield, and Michael B. Twidale. Ethnomethodologically informed ethnography and information system design. *Journal of the American Society for Information Science*, 51(7):666–682, 2000. doi:10.1002/(sici)1097-4571(2000)51:7<666::aid-asi8>3.0.co;2-5.

Emily Dao, Andreea Muresan, Kasper Hornbæk, and Jarrod Knibbe. Bad breakdowns, useful seams, and face slapping: Analysis of VR fails on YouTube. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. doi:10.1145/3411764.3445435.

Trevor de Clercq. The Nashville Number System: A framework for teaching harmony in popular music. *Journal of Music Theory Pedagogy*, 33:3–28, 2019.

Amalia De Götzen, Nicola Bernardini, and Daniel Arfib. Traditional (?) implementations of a phase-vocoder: The tricks of the trade. In *Proceedings of the COST G-6 Conference on Digital Audio Effects*, Verona, Italy, 2000.

Diana Deutsch. Absolute pitch. In *The Psychology of Music*, pages 141–182. Elsevier, 2013.

Yue Ding, Kathleen Gray, Alexander Forrence, Xiaoqin Wang, and Juan Huang. A behavioral study on tonal working memory in musicians and non-musicians. *PLoS ONE*, 13 (8), 2018. doi:10.1371/journal.pone.0201765.

Simon Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects*, page 17, Montréal, Canada, 2006.

Chris Donahue, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, pages 485–492. Bengaluru, India, 2022.

Jonathan Driedger and Meinard Müller. A review of time-scale modification of music signals. *Applied Sciences*, 6(2):57, 2016. doi:10.3390/app6020057.

Chris Duxbury, Mike Davies, and Mark B Sandler. Improved time-scaling of musical audio using phase locking at transients. In *Audio Engineering Society Convention 112*. Audio Engineering Society, Munich, Germany, 2002.

Daniel P. W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36(1):51–60, 2007. doi:10.1080/09298210701653344.

Joel E. Fischer, Andy Crabtree, Tom Rodden, James A. Colley, Enrico Costanza, Michael O. Jewell, and Sarvapali D. Ramchurn. “Just whack it on until it gets hot”. Working with IoT data in the home. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2016.

Will Gibson. All you need is guitar pedals: The communicative construction of material culture in YouTube product reviews. *Discourse, Context & Media*, 49:100626, 2022. doi:10.1016/j.dcm.2022.100626.

Emilia Gómez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3):294–304, 2006. doi:10.1287/ijoc.1040.0126.

David Gonçalves, Manuel Piçarra, Pedro Pais, João Guerreiro, and André Rodrigues. “My Zelda cane”: Strategies used by blind players to play visual-centric digital games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Hamburg, Germany, 2023.

Lucy Green. *How Popular Musicians Learn*. Routledge, London, 2017.

Emma B. Greenspon and Peter Q. Pfordresher. Pitch-specific contributions of auditory imagery and auditory memory in vocal pitch imitation. *Attention, Perception, and Psychophysics*, 81(7):2473 – 2481, 2019. doi:10.3758/s13414-019-01799-0.

Emma B. Greenspon, Peter Q. Pfordresher, and Andrea R. Halpern. Pitch imitation ability in mental transformations of melodies. *Music Perception*, 34(5):585–604, 2017. doi:10.1525/MP.2017.34.5.585.

Stephen B. Groce. Occupational rhetoric and ideology: A comparison of copy and original music performers. *Qualitative Sociology*, 12(4):391–410, 1989. doi:10.1007/BF00989399.

Peter Grosche and Meinard Müller. Computing predominant local periodicity information in music recordings. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, page 5. IEEE, 2009. doi:10.1109/aspaa.2009.5346544.

Peter Grosche and Meinard Müller. Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6): 1688–1701, 2011. doi:10.1109/tasl.2010.2096216.

Andrea R. Halpern. Memory for the absolute pitch of familiar songs. *Memory & Cognition*, 17(5):572–581, 1989. doi:10.3758/bf03197080.

Andrea R. Halpern and Peter Q. Pfordresher. What do less accurate singers remember? pitch-matching ability and long-term memory for music. *Attention, Perception, and Psychophysics*, 84(1):260–269, 2022. doi:10.3758/s13414-021-02391-1.

Denise Harrison, Margaret Sampson, Jessica Reszel, Koowsar Abdulla, Nick Barrowman, Jordi Cumber, Ann Fuller, Claudia Li, Stuart Nicholls, and Catherine M Pound. Too many crying babies: A systematic review of pain management practices during immunizations on YouTube. *BMC Pediatrics*, 14:134, 2014. doi:10.1186/1471-2431-14-134.

Denise Harrison, Shokoufeh Modanloo, Ashley Desrosiers, Louise Poliquin, Mariana Bueno, Jessica Reszel, and Margaret Sampson. A systematic review of YouTube videos on pain management during newborn blood tests. *Journal of Neonatal Nursing*, 24(6): 325–330, 2018. doi:10.1016/j.jnn.2018.05.004.

Christopher Harte. *Towards automatic extraction of harmony information from music signals*. PhD thesis, Queen Mary, University of London, 2010.

Alana N. Hawkins and Ashley J. Filtness. Driver sleepiness on YouTube: A content analysis. *Accident Analysis & Prevention*, 99:459–464, 2017. doi:10.1016/j.aap.2015.11.025.

Juan Pablo Hourcade, Sarah L. Mascher, David Wu, and Luiza Pantoja. Look, my baby is using an iPad! an analysis of YouTube videos of infants and toddlers using tablets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1915–1924, Seoul, Korea, 2015. doi:10.1145/2702123.2702266.

Timothy L. Hubbard. Auditory imagery: Empirical findings. *Psychological Bulletin*, 136 (2):302–329, 2010. doi:10.1037/a0018436.

Dan Huckabee. *How to Figure Out Music From Recordings (DVD)*, 2004. Musician's Workshop. Austin, Texas (USA).

Marcia Earl Humpal. Song repertoire of young children. *Music Therapy Perspectives*, 16 (1):37–42, 1998. doi:10.1093/mtp/16.1.37.

Sean Hutchins and Isabelle Peretz. A frog in your throat or in your ear? searching for the causes of poor singing. *Journal of Experimental Psychology: General*, 141(1):76–97, 2012. doi:10.1037/a0025064.

IFPI. Engaging with music, 2022. URL `https://www.ifpi.org/wp-content/uploads/2022/11/Engaging-with-Music-2022_full-report-1.pdf`.

Kjell-Gunnar Johansson. What chord was that? A study of strategies among ear players in rock music. *Research Studies in Music Education*, 23(1):94–101, 2004.

Nicolas Juillerat and Béat Hirsbrunner. Audio time stretching with an adaptive multiresolution phase vocoder. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 716–720, University of Fribourg, Switzerland, 2017. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ICASSP.2017.7952249.

Margot Kelly-Hedrick, Paul H. Grunberg, Felicia Brochu, and Phyllis Zelkowitz. "it's totally okay to be sad, but never lose hope": Content analysis of infertility-related videos on YouTube in relation to viewer preferences. *Journal of Medical Internet Research*, 20 (5):e10199, 2018. doi:10.2196/10199.

Anssi P. Klapuri, Antti J. Eronen, and Jaakko T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1): 342–355, 2006. doi:10.1109/tsa.2005.854090.

Hubert Knoblauch, Ren Tuma, and Bernt Schnettler. Video analysis and videography. In *The SAGE Handbook of Qualitative Data Analysis*, pages 435–449. SAGE Publications Ltd, London, 2014.

Aida Komkaite, Liga Lavrinovica, Maria Vraka, and Mikael B. Skov. Underneath the skin. an analysis of YouTube videos to understand insertable device interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2019. ACM. doi:10.1145/3290605.3300444.

Grace Kong, Heather LaVallee, Alissa Rams, Divya Ramamurthi, and Suchitra Krishnan-Sarin. Promotion of vape tricks on YouTube: Content analysis. *Journal of Medical Internet Research*, 21(6):e12709, 2019. doi:10.2196/12709.

Filip Korzeniowski and Gerhard Widmer. A fully convolutional deep auditory model for musical chord recognition. In *IEEE 26th International Workshop on Machine Learning for Signal Processing*, Salerno, Italy, 2016. doi:10.1109/MLSP.2016.7738895.

Filip Korzeniowski and Gerhard Widmer. End-to-end musical key estimation using a convolutional neural network. *25th European Signal Processing Conference, EUSIPCO 2017*, 2017:966–970, 2017. doi:10.23919/EUSIPCO.2017.8081351.

Nathan B. Kruse and Kari K. Veblen. Music teaching and learning online: Considering YouTube instructional videos. *Journal of Music, Technology & Education*, 5(1):77–87, 2012. doi:10.1386/jmte.5.1.77_1.

Benjamin M. Kubit and Petr Janata. Spontaneous mental replay of music improves memory for musical sequence knowledge. *Journal of Experimental Psychology: Learning Memory and Cognition*, 49(7):1068–1090, 2022. doi:10.1037/xlm0001203.

Amir Lahav, Adam Boulanger, Gottfried Schlaug, and Elliot Saltzman. The power of listening: Auditory-motor interactions in musical training. *Annals of the New York Academy of Sciences*, 1060:189–194, 2005. doi:10.1196/annals.1360.042.

Jean Laroche and Mark Dolson. Phase-vocoder: About this phasiness business. In *Proceedings of the 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, page 17. IEEE, 1997. doi:10.1109/aspaa.1997.625603.

Kyogu Lee. A unified system for chord transcription and key extraction using hidden markov models. *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 245–250, 2007.

Nicolas Legewie and Anne Nassauer. YouTtube, Google, Facebook: 21st century online video research and research ethics. *Forum Qualitative Sozialforschung*, 19(3), 2018. doi:10.17169/fqs-19.3.3130.

Daniel J. Levitin. Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*, 56(4):414–423, 1994. doi:10.3758/bf03206733.

Franklin M. Li, Francheska Spektor, Meng Xia, Mina Huh, Peter Cederberg, Yuqi Gong, Kristen Shinohara, and Patrick Carrington. "It feels like taking a gamble": Exploring perceptions, practices, and challenges of using makeup and cosmetics for people with visual impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New Orleans, Louisiana (USA), 2022.

Franklin Mingzhe Li, Jamie Dorst, Peter Cederberg, and Patrick Carrington. Non-visual cooking: Exploring practices and challenges of meal preparation by people with visual impairments. In *ASSETS 2021 - 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 2021. doi:10.1145/3441852.3471215.

Lassi A. Liikkanen and Kelly Jakubowski. Involuntary musical imagery as a component of ordinary music cognition: A review of empirical evidence. *Psychonomic Bulletin and Review*, 27(6):1195–1217, 2020. doi:10.3758/s13423-020-01750-7.

Lars Lilliestam. On playing by ear. *Popular Music*, 15(2):195–216, 1996. doi:10.1017/S0261143000008114.

Valerie Looi, Catherine Sucher, and Hugh McDermott. Melodies familiar to the australian population across a range of hearing abilities. *Australian and New Zealand Journal of Audiology*, 25(2):75–83, 2003. doi:10.1375/audi.25.2.75.31118.

Silvia Lovato and Anne Marie Piper. "Siri, is this you?": Understanding young children's interactions with voice input systems. In *Proceedings of IDC 2015: The 14th International Conference on Interaction Design and Children*, pages 335–338, 2015. doi:10.1145/2771839.2771910.

Kapil Chalil Madathil, A. Joy Rivera-Rodriguez, Joel S. Greenstein, and Anand K. Gramopadhye. Healthcare information on YouTube: A systematic review. *Health Informatics Journal*, 21(3):173–194, 2015. doi:10.1177/1460458213512220.

Samuel R. Mathias, Leonard Varghese, Christophe Micheyl, and Barbara G. Shinn-Cunningham. Gradual decay and sudden death of short-term memory for pitch. *Journal of the Acoustical Society of America*, 149(1):259 – 270, 2021. doi:10.1121/10.0002992.

Matthew Louis Mauriello, Cody Buntain, Brenna McNally, Sapna Bagalkotkar, Samuel Kushnir, and Jon E. Froehlich. SMIDGen: An approach for scalable, mixed-initiative dataset generation from online social networks. *HCIL Tech Report*, 2018a.

Matthew Louis Mauriello, Brenna McNally, Cody Buntain, Sapna Bagalkotkar, Samuel Kushnir, and Jon E. Froehlich. A large-scale analysis of YouTube videos depicting everyday thermal camera use. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, New York, NY, USA, 2018b. ACM. doi:10.1145/3229434.3229443.

Brian McFee and Juan Pablo Bello. Structured training for large-vocabulary chord recognition. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pages 188–194, Suzhou, China, 2017.

Brian McFee and Daniel P. W. Ellis. Analyzing song structure with spectral clustering. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 405–410, Taipei, Taiwan, 2014a. doi:10.5281/zenodo.1415778.

Brian McFee and Daniel P.W. Ellis. Better beat tracking through robust onset aggregation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 13. IEEE, 2014b. doi:10.1109/icassp.2014.6853980.

Kelsey McKinney. Where did all the saxophones go?, 2017. URL `https://theoutline.com/post/1409/saxophones-in-american-pop-music-history`.

Gary E. McPherson. Cognitive strategies and skill acquisition in musical performance. *Bulletin of the Council for Research in Music Education*, No. 133, The 16th International Society for Music Education: ISME Research Seminar:64–71, 1997. doi:10.2307/40318841.

Matt McVicar, Raul Santos-Rodriguez, Yizhao Ni, and Tijl De Bie. Automatic chord estimation from audio: A review of the state of the art. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(2):556–575, 2014. doi:10.1109/taslp.2013.2294580.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956. doi:10.1037/h0043158.

Tomoyasu Nakano and Masataka Goto. Music source separation with MLP mixing of time, frequency, and channel. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023.

Sara Nielsen, Mikael B. Skov, Karl Damkjær Hansen, and Aleksandra Kaszowska. Using user-generated YouTube videos to understand unguided interactions with robots in public places. *ACM Transactions on Human-Robot Interaction*, 12(1):1–40, 2023. doi:10.1145/3550280.

Oriol Nieto and Juan Pablo Bello. Systematic exploration of computational music structure research. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 547–553, New York, New York (USA), 2016.

Shuo Niu, Katherine G. McKim, and Kathleen Palm Reed. Education, personal experiences, and advocacy. examining drug-addiction videos on youtube. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28, 2022. doi:10.1145/3555624.

Katy Noland and Mark Sandler. Signal processing parameters for tonality estimation. *Audio Engineering Society - 122nd Audio Engineering Society Convention 2007*, 3:1224–1229, 2007.

Emmett J. O'Leary. The ukulele and YouTube: A content analysis of seven prominent YouTube ukulele channels. *Journal of Popular Music Education*, 4(2):175–191, 2020. doi:10.1386/jpme_00024_1.

Peter Oswald. *High Schoolers' Approaches to Learning Melodies by Ear*. PhD thesis, Temple University, 2022.

Jeni Paay, Jesper Kjeldskov, Mikael B. Skov, and Kenton O'Hara. Cooking together: A digital ethnography. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 1883–1888. 2012.

Jeni Paay, Jesper Kjeldskov, Mikael B. Skov, and Kenton O'hara. F-Formations in cooking together: A digital ethnography using YouTube. In *INTERACT 2013: Proceedings, Part IV*, pages 37–54, Cape Town, South Africa, 2013. Springer.

Jeni Paay, Jesper Kjeldskov, and Mikael B. Skov. Connecting in the kitchen. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, New York, NY, USA, 2015. ACM. doi:10.1145/2675133.2675194.

Hélène Papadopoulos and Geoffroy Peeters. Large-scale study of chord estimation algorithms based on chroma representation and HMM. *CBMI'2007 - 2007 International Workshop on Content-Based Multimedia Indexing, Proceedings*, pages 53–60, 2007. doi:10.1109/CBMI.2007.385392.

Jonggwon Park, Kyoyun Choi, Sungwook Jeon, Dokyun Kim, and Jonghun Park. A bi-directional transformer for musical chord recognition. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 620–627. Delft, The Netherlands, 2019.

Ranjan Pathak, Dilli Ram Poudel, Paras Karmacharya, Amrit Pathak, Madan Raj Aryal, Maryam Mahmood, and Anthony A Donato. YouTube as a source of information on Ebola Virus Disease. *North American Journal of Medical Sciences*, 7(7):306–309, 2015. doi:10.4103/1947-2714.161244.

Michael Quinn Patton. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. Sage Publications, 2014.

Johan Pauwels, Ken O'Hanlon, Emilia Gómez, and Mark B Sandler. 20 years of automatic chord recognition from audio. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 54–63, Delft, The Netherlands, 2019.

Geoffroy Peeters. Chroma-based estimation of musical key from audio-signal analysis. In *Proceedings of the 7th International Society of Music Information Retrieval Conference*. 2006a.

Geoffroy Peeters. Musical key estimation of audio signal based on hidden Markov modeling of chroma vectors. In *Proceedings of the 9th International Conference on Digital Audio Effects*, pages 127–131. Montréal, Canada, 2006b.

Randall G. Pembrook. Interference of the transcription process and other selected variables on perception and memory during melodic dictation. *Journal of Research in Music Education*, 34(4):238–261, 1986. doi:10.2307/3345259.

Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl. MoisesDB: A dataset for source separation beyond 4-stems. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, pages 619–626. Milano, Italy, 2023.

Peter Q. Pfordresher and Steven M. Demorest. The prevalence and correlates of accurate singing. *Journal of Research in Music Education*, 69(1):5–23, 2021. doi:10.1177/0022429420951630.

Peter Q. Pfordresher, Andrea R. Halpern, and Emma B. Greenspon. A mechanism for sensorimotor translation in singing. *Music Perception*, 32(3):242–253, 2015. doi:10.1525/mp.2015.32.3.242.

Judy Plantinga and Laurel J. Trainor. Memory for melody: infants use a relative pitch code. *Cognition*, 98(1):1–11, 2005. doi:10.1016/j.cognition.2004.09.008.

Michael R. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):243–248, 1976. doi:10.1109/TASSP.1976.1162810.

Zdeněk Průša and Nicki Holighaus. Phase vocoder done right. In *Proceedings of the 25th European Signal Processing Conference*, volume 2017, pages 976–980, 2017. doi:10.23919/EUSIPCO.2017.8081353.

Miller Puckette. Phase-locked vocoder. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Accoustics*, page 19, New Paltz, New York (USA), 1995. IEEE. doi:10.1109/aspaa.1995.482995.

Henna A. Qureshi and Züleyha Ünlü. Beyond the paradigm conflicts: A four-step coding instrument for grounded theory. *International Journal of Qualitative Methods*, 19: 1609406920928188, 2020. doi:10.1177/1609406920928188.

Roy V. Rea, Chris J. Johnson, Daniel A. Aitken, Kenneth N. Child, and Gayle Hesse. Dash cam videos on YouTube™ offer insights into factors related to moose-vehicle collisions. *Accident Analysis and Prevention*, 118:207–213, 2018. doi:10.1016/j.aap.2018.02.020.

Bernhard Rieder, Ariadna Matamoros-Fernández, and Òscar Coromina. From ranking algorithms to 'ranking cultures'. *Convergence: The International Journal of Research into New Media Technologies*, 24(1):50–68, 2018. doi:10.1177/1354856517736982.

Axel Röbel. Transient detection and preservation in the phase vocoder. In *Proceedings of the International Computer Music Conference*, Singapore, 2003a.

Axel Röbel. A new approach to transient processing in the phase vocoder. *Proceedings of the 6th International Conference on Digital Audio Effects*, 2003b.

Andrew Robertson, Adam Stark, and Matthew E. P. Davies. Percussive beat tracking using real-time median filtering. In *Proceedings of the 6th International Workshop on Music and Machine Learning*, Prague, 2013.

Dana Rotman and Jennifer Preece. The 'WeTube' in YouTube – creating an online community through video sharing. *International Journal of Web Based Communities*, 6(3): 317, 2010. doi:10.1504/ijwbc.2010.033755.

Dana Rotman, Jennifer Golbeck, and Jennifer Preece. The community is where the rapport is - On sense and structure in the YouTube community. *Proceedings of the 4th International Conference on Communities and Technologies*, 2009:41–49, 2009. doi:10.1145/1556460.1556467.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 12. IEEE, 2023. doi:10.1109/icassp49357.2023.10096956.

Matthew Rueben, Jeffrey Klow, Madelyn Duer, Eric Zimmerman, Jennifer Piacentini, Madison Browning, Frank J. Bernieri, Cindy M. Grimm, and William D. Smart. Mental models of a mobile shoe rack. *ACM Transactions on Human-Robot Interaction*, 10(2): 1–36, 2021. doi:10.1145/3442620.

Margaret Sampson, Jordi Cumber, Claudia Li, Catherine M. Pound, Ann Fuller, and Denise Harrison. A systematic review of methods for studying consumer health YouTube videos, with implications for systematic reviews. *PeerJ*, 1:e147, 2013. doi:10.7717/peerj.147.

Noah Schaffer, Boaz Cogan, Ethan Manilow, Max Morrison, Prem Seetharaman, and Bryan Pardo. Music separation enhancement with generative modeling. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. Bengaluru, India, 2022.

Eric D. Scheirer. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103(1):588–601, 1998. doi:10.1121/1.421129.

Norbert Schnell and Diemo Schwarz. Gabor, multi-representation real-time analysis/synthesis. In *Proceedings of the International Conference on Digital Audio Effects*, pages 122–126, Madrid, Spain, 2005.

Norbert Schnell, Geoffroy Peeters, Serge Lemouton, Philippe Manoury, and Xavier Rodet. Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA). In *Proceedings of the International Computer Music Conference*, Berlin, Germany, 2000.

Karen Schulze, Stefan Koelsch, and Victoria Williamson. Auditory working memory. *Springer Handbooks*, pages 461–472, 2018. doi:10.1007/978-3-662-55004-5_24.

Katrin Schulze and Stefan Koelsch. Working memory for speech and music. *Annals of the New York Academy of Sciences*, 1252(1):229 – 236, 2012. doi:10.1111/j.1749-6632.2012.06447.x.

Katrin Schulze and Barbara Tillmann. Working memory for pitch, timbre, and words. *Memory*, 21(3):377 – 395, 2013. doi:10.1080/09658211.2012.731070.

Katrin Schulze, Dowling W. Jay, and Tillmann Barbara. Working memory for tonal and atonal sequences during a forward and a backward recognition task. *Music Perception: An Interdisciplinary Journal*, 29, No. 3:255–267, 2012.

Donna L. Schuman, Karen A. Lawrence, and Natalie Pope. Broadcasting war trauma: An exploratory netnography of veterans' YouTube vlogs. *Qualitative Health Research*, 29 (3):357–370, 2019. doi:10.1177/1049732318797623.

Xavier Serra and Julius Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12, 1990. doi:10.2307/3680788.

Alexander Sheh and Daniel P.W. Ellis. Chord segmentation and recognition using EM-trained Hidden Markov Models. In *Proceedings of the 4th International Society for Music Information Retrieval Conference*, Baltimore, Maryland (USA), 2003. doi:10.7916/D8C82KNH.

Bob Snyder. Memory for music. In *The Oxford Handbook of Music Psychology*, pages 167–180. Oxford University Press, 2014.

Lucy Suchman. *Human-machine reconfigurations: Plans and situated actions, 2nd edition*. Cambridge University Press, New York, New York (USA), 2006.

David Temperley. What's key for key? the Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception*, 1999.

Paul ten Have. Ethnomethodology and conversation analysis. In *APA handbook of research methods in psychology: Research designs: Quantitative, qualitative, neuropsychological, and biological (Vol. 2) (2nd ed.)*, pages 131–146. American Psychological Association, Washington, 2023.

Harvey Thornburg, Randal J. Leistikow, and Berger Jonathan. Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1257–1272, 2007. doi:10.1109/TASL.2006.889801.

Michael A. Tollefsrud, Chelsea N. Joyner, Alexandria C. Zakrzewski, and Matthew G. Wisniewski. Not fully remembered, but not forgotten: interfering sounds worsen but do not eliminate the representation of pitch in working memory. *Attention, Perception, and Psychophysics*, 2024. doi:10.3758/s13414-024-02845-2.

Sylvaine Tuncer, Barry Brown, and Oskar Lindwall. On pause: How online instructional videos are used to achieve practical tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2020. ACM. doi:10.1145/3313831.3376759.

Sylvaine Tuncer, Oskar Lindwall, and Barry Brown. Making time: Pausing to coordinate video instructions and practical tasks. *Symbolic Interaction*, 44(3):603–631, 2021. doi:10.1002/symb.516.

Stephen C. Van Hedger, Shannon L.M. Heald, Rachelle Koch, and Howard C. Nusbaum. Auditory working memory predicts individual differences in absolute pitch learning. *Cognition*, 140:95 – 110, 2015. doi:10.1016/j.cognition.2015.03.012.

Stephen C. Van Hedger, Shannon L. M. Heald, and Howard C. Nusbaum. Long-term pitch memory for music recordings is related to auditory working memory precision. *Quarterly Journal of Experimental Psychology*, 71(4):879 – 891, 2018. doi:10.1080/17470218.2017.1307427.

Stephen C. Van Hedger, Noah R. Bongiovanni, Shannon L. M. Heald, and Howard C. Nusbaum. Absolute pitch judgments of familiar melodies generalize across timbre and octave. *Memory & Cognition*, 51(8):1898–1910, 2023. doi:10.3758/s13421-023-01429-z.

Maria Varvarigou and Lucy Green. Musical 'learning styles' and 'learning strategies' in the instrumental lesson: The Ear Playing Project (EPP). *Psychology of Music*, 43(5): 705–722, 2015. doi:10.1177/0305735614535460.

Jasmin L. Vassileva, Jill M. Vongher, Mariellen Fischer, Lisa L. Conant, Robert C. Risinger, Betty Jo Salmeron, Elliot A. Stein, Russell A. Barkley, and Stephen M. Rao. 57. working memory deficits in adults with ADHD. *Brain and Cognition*, 47(1-2):216–219, 2001.

Radu-Daniel Vatavu, Ovidiu-Ciprian Ungurean, and Laura-Bianca Bilius. Interactive public displays and wheelchair users: Between direct, personal and indirect, assisted interaction. In *The 35th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2022. ACM. doi:10.1145/3526113.3545662.

Don Weenink, Raheel Dhattiwala, and David van der Duin. Circles of peace. A video analysis of situational group formation and collective third-party intervention in violent incidents. *The British Journal of Criminology*, 62(1):18–36, 2022a. doi:10.1093/bjc/azab042.

Don Weenink, René Tuma, and Marly van Bruchem. How to start a fight: A qualitative video analysis of the trajectories toward violence based on phone-camera recorded fights. *Human Studies*, 45(3):577–605, 2022b. doi:10.1007/s10746-022-09634-6.

Johann Wentzel, Sasa Junuzovic, James Devine, John Porter, and Martez Mott. Understanding how people with limited mobility use multi-modal input. In *CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2022. ACM. doi:10.1145/3491102.3517458.

Jennifer A. Whitaker, Evelyn K. Orman, and Cornelia Yarbrough. Characteristics of "music education" videos posted on YouTube. *Update: Applications of Research in Music Education*, 33(1):49–56, 2014. doi:10.1177/8755123314540662.

Gordon Wichern, Hannah Robertson, and Aaron Wishnick. Quantitative analysis of masking in multitrack mixes using loudness loss. In *141st Audio Engineering Society International Convention*, 2016.

Matthew G. Wisniewski and Michael A. Tollefsrud. Auditory short-term memory for pitch loses precision over time. *JASA Express Letters*, 3(3):034402, 2023. doi:10.1121/10.0017518.

Robert H. Woody and Andreas C. Lehmann. Student musicians' ear-playing ability as a function of vernacular music experiences. *Journal of Research in Music Education*, 58 (2):101–115, 2010.

Jingwei Zhao, Gus Xia, and Ye Wang. Beat transformer: Demixed beat and downbeat tracking with dilated self-attention. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. Bengaluru, India (Hybrid), 2022.

Qingxiao Zheng, Shengyang Xu, Lingqing Wang, Yiliu Tang, Rohan C. Salvi, Guo Freeman, and Yun Huang. Understanding safety risks and safety design in social VR environments. In *Proceedings of the ACM on Human-Computer Interaction*, volume 7(CSCW1), 2023. doi:10.1145/3579630.