# PRYING EARS: A CRITICAL RESPONSE TO MASS AUDIO SURVEILLANCE

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Master of Arts

in

Digital Musics

by Aaron Karp

Guarini School of Graduate and Advanced Studies

Dartmouth College

Hanover, New Hampshire

April, 2019

Examining Committee:

(chair) Michael Casey, Ph.D.

Aden Evens, Ph.D.

Ben Vida, M.F.A.

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

# Abstract

Mass audio surveillance, conducted by state or corporate entities, is one of the most significant threats to personal privacy in the digital age. This thesis seeks to reconcile the technology of the practice with theory in order to form a holistic picture of mass audio surveillance for future discussion and action. First, theories of machine listening are developed through an artistic lens using different abstractions of human listening and frameworks from perceptual biology. These provide a context within which a theory of mass audio surveillance is situated. These findings serve as the conceptual backbone for the Acoustic Counterfeit Machine, a system designed to "hide" sounds from methods of mass audio surveillance through acoustic masking, altering the sounds' semantic content while maintaining their semantic context. The system is explored in technical detail and evaluated to show its effectiveness; it is then demonstrated and postulated as part of a future art installation design. The installation seeks to communicate the combination of theory and technology central to this thesis, placing my work within a broader political context. A hyper-focus on the technology of mass audio surveillance can lead to a rebuttal built on flawed foundations, while overemphasis on theory can lead to a conception at odds with reality. As this thesis seeks to demonstrate, it is through a unification of the two that an effective response to mass audio surveillance can begin to be built.

# Acknowledgments

I'd like to express my sincere gratitude to my thesis committee: Michael Casey, Aden Evens, and Ben Vida. Special thanks to my brother, Jonathan Karp, for suggesting the topic of this thesis and reminding me to take a deep breath every once and a while. Thanks to the current and former graduate students of the Dartmouth Digital Musics program- including Hunter Brown, Clara Allison, Christy Rose, Lloyd May, Dominic Coles, Xanthe Kraft, Camilla Tassi, and Beau Sievers- for all their help and camaraderie. Thanks to Alexis Hill, Hedra Rowan, and Sun Chang for their friendship, and to Amey Zhang for keeping my head in the clouds. I am indebted to Michelle Lou for teaching me to create music in the manner that most speaks to me, and a special apology to my countless violin and viola teachers for never practicing. Thank you to the nameless programmers on Stack Overflow who's problem-solving was essential to the success of my code. Finally, special thanks to my parents, Benjamin and Margie Karp, for their continual love and support, and for always encouraging my curiosity.

# Contents

# List of Figures

# 1

# Introduction

In contemporary life, personal surveillance is ubiquitous. External monitoring of individuals can take many different forms, most commonly perpetrated by state or corporate actors. These two systems may seem far from equivalent, but with advances in technology and a paralleled increase in access through the internet, both of these forms have developed a similar structure: one of mass automated surveillance. For the purpose of this thesis, mass surveillance refers to as any close observation or monitoring of sections or totalities of society that happens on a scale that necessitates automation.

Of the stated forms of mass surveillance, none seem more surprising to the general public than the state surveillance apparatus. With recent leaks from the NSA and CIA, the existence

of the domestic surveillance apparatuses of global powers has become undeniably clear. The reach of these programs has been shown to encompass all of society, essentially operating on a principle of assumed suspicion, allowing the state to conduct close observation of anyone at any time.

A number of papers, books, artistic endeavors, and other projects have been created to point out and explore the nature of video surveillance. With the advent of modern CCTV in the 1970s and their widespread adoption through the ’80s and early ’90s, the rise of video surveillance has been an easily observable phenomenon. In contrast, audio surveillance has a less visible history. Relatively little has been done academically or artistically regarding audio surveillance.

It is easy to point out all of the cameras around you at any given point. Right now I am visible to my phone, my computer, and two security cameras placed on the ceiling of the library. It is comparatively much more difficult to isolate sources of audio surveillance. This is for one obvious but important distinction: microphones don’t rely on line-of-sight. While the four aforementioned devices could be listening, so too could the computers and phones of everyone within fifty feet of me. The rise of audio surveillance is difficult for the average person to passively notice partly because of how invisible it is.

When asked how to combat video surveillance, the simplest answer is of course to hide. A video camera, while an imposing presence, loses its potency the moment you close a door or step behind a wall. This is not the case when it comes to audio surveillance. Microphones are designed to take advantage of the distributive quality of sound, which is an acoustic property that is difficult to subvert.

The goal of this project is two-fold. First, I will discuss audio surveillance as a means of machine listening, exploring its place in society and the implications of its presence by creating a description of the phenomenon through an artistic lens. Then I will present—both technically and in an artistic context—a prototype system for "hiding" from audio surveillance using acoustic masking, walking through its design, construction, and effectiveness.

# 2

# Background

"Most people do not mind if a machine records all their words like an attentive parent or friend and renders language concretely infinite and searchable. What people mind is being hurt or being killed, or being made to feel ugly, or being made poor and then being punished for it." - Hannah Black, *Social Life* [20]

## 2.1 A Tool of Oppression

The first commonly understood example of mass surveillance in the United States occurred during World War II, when the FBI and Military Intelligence surveilled the written and telephone communications of some 10 million German-Americans [73]. Immediately following in the 1950's, suspected communists were placed under state surveillance during the era known as "McCarthyism" [4]. During the Vietnam War and the Civil Rights Movement, the NSA, CIA, and FBI regularly monitored the communication of journalists [103], civil rights leaders [18], and congresspeople [14]. The Gulf Wars brought about widespread domestic phone surveillance [98], which ultimately paled in comparison to the domestic surveillance undertaken after September 11th, 2001 and the passing of the Patriot Act. In addition to removing court protections for surveillance (meager as they were) [65], telecommunications companies were directly aiding the state surveillance effort [12]. This was also the first time the public became aware that the data from these wiretaps, which were previously thought to have been used only in transience, were being stored for years after their capture.

In 2011, a WikiLeaks dossier pointed to an even greater expansion of the modern surveillance apparatus, revealing the "multi-billion dollar industry" that existed exclusively to support state surveillance [74]. The Edward Snowden NSA leaks in 2013 brought the PRISM program to light, which was started in 2007 with the express purpose of intercepting and storing internet communications and had been operating under broad warrantless directives aimed at US citizens [15, 47]. This program extracted raw data, including audio, directly from "the central servers of nine leading U.S. Internet companies" [40].

There's a through line one can attempt to draw about the history of surveillance in the United States as being a progression from primarily targeted to primarily generalized. Certainly, as the technology has allowed for broader surveillance, the state has used it for such. What was once (pre-1940's) a practice focused on the individual is now a practice applied to the broad public. This is first and foremost due to the advances in technologies

available to the state. The monitoring of German-American citizens during WWI involved immense efforts, "requiring legions of postal workers to physically examine some 30 million first-class letters and 350,000 badge-carrying vigilantes to perform shoe-leather snooping on immigrants, unions, and socialists of every sort" [73]. Any reassurance gained from considering the manpower cost of surveillance is dashed in the digital age. The collection of data had started to be automated as soon as data storage technologies were created, starting with innovations in magnetic tape made in the 1930s.

The biggest changes coming into the 2000s were an exponential increase in storage capabilities and an automation of the analysis of the data. A warehouse of tapes in the '70s would fit in a small room in the '90s and on a single portable hard drive now. Any previous need to be diligent and considerate about what data was being collected and what data was being stored left the discussion twenty years ago. If being selective wasn't necessary, why wouldn't the NSA record every communication sent by every person? As made public in the Snowden leaks, that is exactly what the government has done at an increasing rate and depth.

By focusing on the expanding reach of surveillance, however, one can lose sight of a more essential common thread, which is that surveillance is used as a tool by those with power against those without. In every instance of contemporary mass surveillance this has been the case. This fact was easier to ignore for a general populace when it was just the German-Americans or the outspoken Civil Rights leaders who were under watch. As surveillance has grown to encompass the entirety of the population, it is forcing that entirety to come to terms with the realization that they are *all* in fact vulnerable under the current state and capitalist structures.

This is not to say that some are not still more targeted than others: black and brown people, individuals of lower socio-economic class, and the politically outspoken are all still targeted by surveillance at a greater concentration. As is often the case, the best example of how the state treats the less privileged and vulnerable can be found in its abuse of

incarcerated individuals. On January 30, 2019, George Joseph and Debbie Nathan published an article in the *Intercept* uncovering the widespread use of voice print technology in U.S. prisons [58]. Multiple states have systems in place that record and store all conversations held over prison phones. That data is then used to create individual voice prints of both prisoners and anyone calling in to talk to prisoners. In brief, voice print technology uses machine learning to build statistical models of an individual's vocal properties, allowing a machine to identify someone (with varying accuracy) from the sound of their voice. The recordings used to create these models are often obtained either without consent or under threat of punishment. The prints are then used to track both incarcerated and non-incarcerated individuals, mapping relationships between current and former prisoners, as well as family members, case workers, and anyone else who happens to be caught in the technology's wide net. All of the technology behind this dystopian practice is privatized and operated by such companies as Securus Technologies and Global Tel Link, which offer no transparency as to the security or specifics of their practices.

While it doesn't fit into the exact same mold as state surveillance, corporate surveillance serves a complementary goal. In the past five years, the commercial push for "always-on" listening devices, such as Amazon Echo and Google Home, has vastly reduced the possibility for surveillance free spaces. Conservative estimations put these "smart speakers" in 16% of American households in January of 2018 [80], with that number ballooning to 20% just three months later [81]. Another report by the same journalist take from RBC Capital Markets stated in December of 2018 that "41% of U.S. consumers now [own] a voice activated speaker" [82]. While this number could be exaggerated, Nielson gave an estimate in late 2018 of "24 percent of US households owning a smart speaker" [90], which is still a staggering statistic. The recordings made from these speakers, which were only thought to be heard by algorithms, are reviewed by human employees at Amazon, Google, and Apple, as shown in an April 2019 report by *Bloomberg* [36]. Even if one believes wholeheartedly in the benevolence of megacorporations it is impossible to have confidence in the security

of such devices to protect against government or private infiltration. In mid-2018 there was a alleged "bug" reported by researchers that allowed for continuous speech transcription without user's knowledge or consent using an Amazon Echo [62], and an almost identical report was made for the Google Home Mini six months earlier [24]. It can't be overstated how large the contributions of these corporations are to the legitimization of the surveillance state. The fastest way to ensure the collapse of any mass resistance to surveillance is to normalize surveillance itself as a helpful technology, thus transforming it from a human rights violation to an accepted necessary evil.

Mass surveillance additionally operates as a ruthless tool of expansive capitalism. In 2015, artist and writer Hannah Black published "Social Life" in Texte Zur Kunst [20], where she explored the relationship of social media to surveillance, capitalism, and human connection. In social media systems, user's information, including both demographics and their posts themselves, are used by corporations as the basis for revenue streams. Black writes, "Like the capitalist dream of robot workers, the direct conversion of life into value is a fantasy about the full negation of labor." Under social media, all activity is valued, and therefore a form of labor; similarly, under a system of mass surveillance, speaking is labor, and speech has value. The "capitalist dream" Black refers to creates a contradiction, though one that is in line with capitalism, where all speaking is labor and yet no one are conscious laborers, allowing surveillance systems to create capitol with no consideration or recompense as to its source. Speaking of "the black womanists", Black writes, "surely black women have felt the pain of being valuable often enough." The transference of existence to value is at the core of oppression, and is facilitated through forms of continual surveillance.

State and corporate mass surveillance is a monumental structure that shows no sign of slowing down. State practices face little opposition from either party, and regulatory agencies are either not equipped or not interested in curtailing corporate abuses. Surveillance is a part of everyday life in the United States—alongside most of the world—and a lack of political opposition to its presence can make resistance seem impossible. And this is, of

course, a necessary goal of oppression: to make the oppressed feel that their circumstances are the inevitable result of some grand architecture and are too great to overcome.

This thesis is not a road map towards organized political resistance; it seeks to elucidate the practices of mass audio surveillance. By naming, defining, and exploring the components of such practices, forms of personally, socially, and politically effective resistance to surveillance can be constructed.

## 2.2 Listening and the Acousmatic

Audio surveillance at a mass scale necessitates the use of machine listening, a term which encompasses any machine that uses a microphone to "hear" sound. While this can include very basic functionality, in the context of modern mass audio surveillance it refers to computers that run complex algorithms to analyze sound and make decisions based on the results. The concept of machine listening is central to an understanding of mass audio surveillance, both from an ontological and a technical perspective. I will be approaching the topic from two angles, the conceptual and the practical, with the ultimate goal of connecting them into a unified description of machine listening in a surveillance context.

Machine brains listening with machine ears operate on a fundamentally different framework than a human model of perception. There are, however, significant similarities between machine listening and Pierre Schaeffer's concept of *reduced listening* [85]. I recruit reduced listening and the acousmatic experience as a jumping-off point in describing machine listening.

Schaeffer's concept of the acousmatic came from his explorations of musique concrète. This was a musical practice pioneered by Schaeffer that used spliced audio recordings and synthesized sound to create sonic elements of unknowable origin. He and his colleagues at Studio d'Essai, in Paris, would make recordings of any sound, from rivers to trains to conversations, and physically cut the tape into small segments. These segments would then

be used in isolation or spliced together to create new composites [79]. The segmentation was done at a small enough scale to make it virtually impossible to decipher the origin of the component sources. Without any discernible point of departure, sounds could hypothetically exist without historical or personal context. These sounds were called acousmatic, a term which came to mean any sound disambiguated from a direct source. Schaeffer considered acousmatic music to possess immense artistic possibility, allowing listeners to focus on the sound itself rather than its place in the world.

In musique concrète, "The emphasis was placed on listening; the ear would have to train itself to hear these new musical values unique to the sonic materials deployed" [59, p. 17]. This new kind of listening was categorized as "reduced listening", one of four modes of listening Schaeffer outlines in *Traité des objets musicaux: essai interdiscipline*. Reduced listening is characterized as listening to sounds' morphological attributes without care to source, cause, or spatial location.

While Pierre Schaeffer coined the term in musical vernacular, discussion of reduced listening and the acousmatic has continued in various forms. Denis Smalley has argued for an expansion of the listening practice in electronic and electroacoustic music [92], as well as a return to and emphasis on spatial identification within acousmatic sound [94]. The many effects of the switch to recordings as a primary mode of listening (and acousmatic sound as the primary mode of hearing music) have been explored by John Young [107] and Eric Clarke [30], among many others. The relationship between the acousmatic and the physical has been explored by Simon Emmerson [43]. A recent book by Brian Kane titled "Sound Unseen: acousmatic sound in theory and practice" [59] has thrown the very foundation for acousmatic sound into question. Here Kane argues that the specific framing of Pythagorean origin in which Schaeffer defines acousmatic sound is historically inaccurate, and that a broader context for acousmatic sound and reduced listening exists in numerous forms.

For a machine, a being seemingly without external bias, it may appear quite clear that all sound is acousmatic, and all listening reduced. This is true to greater and lesser extents.

With Schaeffer's modes of listening as existing in different interpretive contexts, the question of machine listening ultimately comes to who can access which contexts. All of these works of theory rely on the human listener as one of their foundational elements, and thus they all make certain assumptions about access to contexts that do not necessarily hold true when applied to machines.

In December of 2018, Technosphere Magazine released a dossier focused on the phenomenon of machine listening and its numerous consequences for art and human aurality [69]. The nine works that compose the dossier each take a different approach to machine listening. Florian Hecker explicitly explored the difference between human and machine audition in his piece *1935*, emphasizing the heightened ability of machines to hear certain aspects of sound and music [53]. In her piece *Calm can only make it false (Noise Floor)*, Yoneda Lemma marks the key difference between machine and human listening as one of abstraction [63], which is the ability to perceive from multiple perspectives, whether cultural or structural or temporal. This is not dissimilar from the concept of context discussed above, where the ability to attune oneself to different contextual spaces marks an essential difference between human and machine listeners. The guest curator of the dossier, Stefan Maier, combines a theory of machine listening with one of machine learning, examining Google's *WaveNet* project as an example of the results when these two technologies are combined [70].

In my exploration of machine listening I will be piecing together multiple theories, using Brian Kane's "Sound Unseen" as a starting point for Schaeffer's ideas and their rebuttals, to create a definition of machine listening that is acousumatic-adjacent, but not altogether parallel.

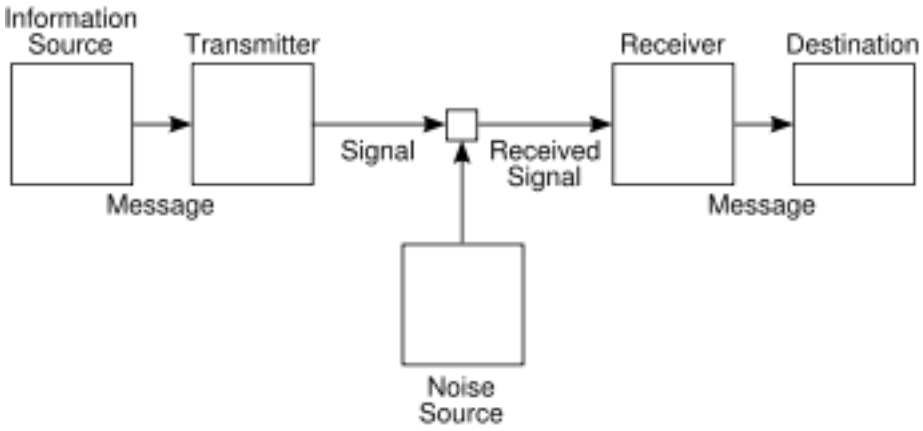



**Figure 2.1:** Claude Shannon's model of all communication systems (source: https://commons.wikimedia.org/wiki/File:Shannon_communication_system.svg

## 2.3 Claude Shannon and the Structure of Communication

Machine listening as a theory cannot exist isolated from considerations of its more technical elements. Machine listening is a subset of explorations that falls under the overarching description of machine perception. Machine perception (and, consequently, listening) as a practice is built on the foundations of information theory. This field seeks to describe the flow of information in a generalized sense mathematically.

The creator of information theory was Claude Shannon. In 1948, at the age of 32, he published "A Mathematical Theory of Communication", a two-part article that both raised and answered almost every question about the definitions and limits of information communication [88]. While applied to the problem of signal processing, the theory is easily extrapolated to other domains. The points raised in these articles—later appropriately published under the name "*The* Mathematical Theory of Communication"—are numerous and all consequential. To understand why surveillance and machine perception operate as they do I will briefly describe the core elements of Shannon's theory.

Shannon outlines six elements in any communication, laid out in Figure 2.1. First is the Information Source, where the message is conceived. The message is then encoded and sent by means of a Transmitter. The message travels across a Channel where it is

affected by a Noise Source. The second half of the diagram is the mirror of the first, with the message being decoded by the Receiver and finally handed to the Destination. An important concept to understand is that the starting message is not the received message. Though the transmission and reception are exactly inverse functions, the addition of a source of noise alters the signal in transit, changing it by some non-zero amount.

In the context of mass audio surveillance, the diagram can be interpreted as follows. The Information Source is the source of the sound being observed, such as an individual who is speaking aloud. The Message is whatever content is trying to be conveyed; it is the core meaning of the communication, and is not necessarily the same as the speech itself. The Transmitter is the speech itself, meant to convey the message from one party to another. The Signal is the physical vibrations in the air that describe the sound being observed. The Noise Source can be many things: other sounds in the space, sound dampening material, sound reflecting material, and anything else that affects the acoustic vibrations in the physical domain. The modified signal then reaches the Receiver, which is in this case a microphone. That receiver then goes through a nested communication system with the signal being sent to and processed by a machine. The final Message is interpreted by the machine and, if deemed appropriate, sent to a final Destination of a human listener.

This diagram fundamentally describes any and all forms of communication. Drawing from this structure, Shannon outlines three "problems" of communication that, taken together, describe the "success" of any communication. They are as follows:

> Level A: How accurately can the symbols of communication be transmitted? (the Technical problem)
>
> Level B: How precisely do the transmitted symbols convey the desired meaning? (the Semantic problem)
>
> Level C: How effectively does the received meaning affect conduct in the desired way? (the Effectiveness problem)

The goal of hiding sound from audio surveillance algorithms is ultimately to disrupt conduct in the receiver. However, that doesn't necessarily imply that the best point of aggravation would be at Level C, the effectiveness problem. While the effectiveness problem is closest to the receiver, that also makes it much more difficult to use as the location for any intervention.

There have certainly been examples of anti-surveillance work done within different levels, but those tend to be fall under significant projects undertaken by state or corporate entities. Anti-surveillance technologies created by individuals or small groups usually operate at a point prior to even level A. For example, anti-face-recognition prosthetic masks [11] and makeup [50] seek to alter the Signal itself, essentially operating as an intentional Noise Source. It is at this point that the power most clearly rests within an individual's grasp, and it is here that the system I design seeks to hold its intervention.

## 2.4 Anti-Surveillance Practices

Technologically-mediated audio surveillance has existed since the creation of the first commercial telegraphs in 1844 [13]. Resistance to surveillance practices rose immediately and remained through the 20th century, but it has always represented a minority of the public, and has a noted absence among those in power. People's responses to the problem of surveillance have taken many forms throughout history. I will briefly describe recent examples of artistic and technological anti-surveillance works to situate within them my anti-surveillance system, which exists both as a demonstrative technology and a reflective art installation.

A significant amount of artistic work has been done in the area of sousveillance. First used by researcher Steve Mann, sousveillance is defined as observing the observer. As seen in Figure 2.2, sousveillance often represents the individual recording a situation in which they are participating; this is essentially a reversal of the surveillance organization. There

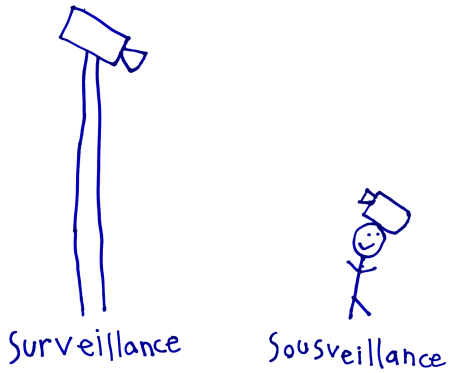



**Figure 2.2:** The concept of sousveillance, as drawn by Steve Mann's six-year-old daughter. (source: https://upload.wikimedia.org/wikipedia/commons/7/7f/SurSousVeillanceByStephanieMannAge6.png

are many kinds of sousveillance, which vary in their purpose and the relationship between the surveyor and the subject. Sousveillance has been used in a variety of contexts, and is one of the most common forms of direct critique on surveillance.

Perhaps the earliest notable work using sousveillance is Andy Warhols' *Outer and Inner Space* [102], released in 1966. This film shows Edie Sedgwick answering interview questions while a pre-recorded tape of her answering the questions plays on a monitor next to her. The two faces are turned towards each other from the camera's perspective. What results is the illusion of Sedgwick observing her own observation. While this film has only been screened on rare occasions, it is an early example of a critique on modern surveillance through the lense of sousveillance.

In 2001 Steve Mann created HeartCam [71]. This device was a bra with two "surveillance domes" as the cups. The domes would take photographs at a speed corresponding to the wearer's heart rate. When someone assailed the wearer their heart rate would involuntarily increase and more photos would be taken of the assailant. This was sousveillance on an individual level, meant to invert the male gaze.

In 2014, ten artists working with photography created pieces approaching surveillance from a variety of directions as part of *Watching You, Watching Me*, a collection published as number 22 in the series Moving Wall from the Open Society Foundations Documentary

Photography Project [5]. This has been one of the largest public exhibitions of surveillance art, and a number of the works included take a sousveillance approach. *It's Nothing Personal*, by Mari Bastashevski, explored the economic world of surveillance. She used materials from "international electronic surveillance companies" and installed them around photographs of corporate security enterprises [16]. *Qaddafi Intelligence Room*, a series by Edu Bayer, depicts a number of rooms used by Muammar al-Qaddafi's intellignce agents, "where they spied on emails and chat messages with the help of technology Libya acquired from the West" [17]. Both of these are examples of broader sousveillance, documenting surveillance practices by looking at where they happen.

Sousveillance can be an incredibly effective technique in certain respects. Recording the "watching" itself allows for a level of direct accountability. Recording is evidence, and when addressing abuses by state surveillance powers it's necessary to have as much and as irrefutable evidence as possible. If the objective is to observe such abuses, then sousveillance is a logical path to go down. If the end goal cannot be achieved through observation, however, sousveillance as a methodology falls short.

Also part of *Watching You, Watching Me* is a series titled *Thousand Little Brothers*, by Hasan Elahi [41]. Elahi was erroneously tied to terrorists by the FBI, and was put under an extended investigation. In response, he photographed thousands of "mundane details from his daily life" and mailed them to the the FBI on a weekly basis. This piece offers a form of direct confrontation with the surveillance state; its combative and actionable methods provide a useful framework for an artistic practice grounded in the reality and physicality of surveillance.

A common approach for surveillance art is addressing physical location and GPS data. In terms of the many realities of a surveillance state, location tracking is quite commonly practiced and relatively simple to implement and understand, which makes it easy to duplicate for the artist or layman. At the same time, location tracking elicits strong responses. People view their locations as sacred information, though, like most surveillance

technology, location tracking has been normalized by framing it as a helpful feature. By this point most phones are passively tracking location at all times [100, 106].

Mont-réel, created in 2015 by Eva Clouard, displays the artist's location on a gallery display in realtime [31]. Visitors could watch as Eva traveled across the city, gaining insight into how a relatively small piece of information can imply significant personal connections, such as where she lives, where she works, and where her friends are. This practice of personal exploitation is frequent in surveillance art. Because of its intrusive and violent nature, it is certainly safer for an artist to mimic such practices on themselves instead of others.

As with any scientific or technical subject, some of surveillance art chooses to focus on the technology itself, perhaps to the detriment of the human consideration. In 2017, *Hansel and Gretel* opened at the Park Avenue Armory in New York City [104]. A collaboration between Chinese activist Ai Weiwei and Swiss architects Jacques Herzog and Pierre de Meuron. The installation was in two parts. The first involved a camera-based setup that tracks the position of participators. Their previous positions are projected onto the floor, creating a ghost-like fading image of their movement over time. The second part used facial recognition software to give the participant a picture of themselves taken from the first room. This installation gives the audience an impression of how advanced the technology for surveillance is, but does nothing to question its use. The technology itself is the centerpiece, and the result is that surveillance "is mostly reduced from threat to mildly educational fun" [95].

While a technical focus has the potential to harm an artistic project by emphasizing the technology over the meaning and purpose of its use and impact, much valuable work has been done within a solely technical domain centered around the creation of anti-surveillance tools. A presently common example is the rise in usage of messaging apps with default end-to-end encryption, such as Signal [8] and WhatsApp [25]. These use algorithms that ensure a much safer path for a message to travel along to prevent snooping. Secure web-browsing can be

attained through the use of Tor [57, 105] for anonymity and Ghostery [2] to prevent tracking services to access your information. Mobile anti-surveillance tools are just as plentiful. such as the Android IMSI-Catcher Detector (AIMSICD) [28], an open-source android application that notifies users if their phone connects to a false mobile tower (a device used by local, state, and federal law enforcement to intercept cellular communication [9]).

Physical tools are another subset of anti-surveillance technology. The anti-face-recognition prosthetic masks [11] and makeup [50] mentioned previously are examples of personal-use physical technologies. Clothing has been the medium of choice for many anti-surveillance designers, with such examples as the Jammer Coat [1] and the Anti-Surveillance Coat [86]. NSA whistle blower Edward Snowden designed an iPhone case that displays otherwise hidden information about whether your phone may be transmitting data that's unsecured [72].

For technology to be effective, access and education are essential. Matt Mitchell founded Crypto Harlem, an organization that seeks to educate individuals in Harlem about anti-surveillance technologies [34]. A simple fact that is often ignored by technologists and entrepreneurs is that mass surveillance is not a equally-applied force. As stated previously, it is a system of structured oppression, and as such it disproportionally targets the poor, communities of minorities, and individuals of color. While this is a topic too immense to cover in this text, Jason Nance's 2016 article "Student surveillance, racial inequalities, and implicit racial bias" [75] provides a thorough study of surveillance as a racialized system. Additionally, Arun Kundnani's and Deepa Kumar's 2015 article "Race, surveillance, and empire" [60] delves deeply into the social history of surveillance, providing a broader context for today's surveillance state.

With all of these works of art and functional tools, there is a glaring historical lack of consideration for the acoustic. Anti-surveillance has a rich history, but also a clear bias towards the visual. Audio surveillance has existed long before video, but when video surveillance was created, most surveillance commentary switched focus to the more "advanced" technique. Even in other non-visual modes of surveillance, such as location tracking, works

critiquing them are generally presented within a strictly visual context. Despite audio surveillance technology and usage making giant leaps in the past thirty years, there has been very little work done in that time with a purely auditory focus. Ultimately, the installation design presented in this thesis was created to accentuate the technological capabilities of mass audio surveillance and simultaneously present a world in which such surveillance is not inevitable, but is a force that can and should be resisted.

# 3

# Machine Ears

"Listening puts us in the world" - Stephen Handel, *Listening: An Introduction to the Perception of Auditory Events* [49]

In order to adequately approach the topic of mass audio surveillance, it is first necessary to explore the foundation on which it rests: machine listening. In an effort to place mass audio surveillance within a broader context, this chapter seeks to build models that describe machine listening in its different forms, walking through their definitions and implications. First, a brief overview of models of human listening will be given, and their relationships to machine listening will be questioned. Using the form of these human models, three different

axes of machine listening will then be presented, followed finally by a reduction of those theories to the practice of mass audio surveillance.

## 3.1 Human Listening

In 1967, Pierre Schaeffer outlined four modes of listening that one can experience. They are: *ouïr*, *comprendre*, *écouter*, and *entendre*. These modes provided the foundation for acousmatic theory, and are fundamental to the Schaefferian model of auditory perception.

*Ouïr* is the most basic biological listening mode. It is pure perception, but only obtained in an unconscious, inattentive way. While the other modes require some level of focus, "Ouïr provides that which is *passively* 'given to me in perception'" [59, p. 27]. *Comprende* is listening to sound in reference to a system of signs and symbols. For example, speech falls under *comprende* because the sound is heard as a particular reference to external symbols, that of a language. *Écouter* is concerned with the identification of sound relative to its environment. A listener tuned into *écouter* would be using sonic and spacial properties to classify a sound as coming from a specific source and cause: "It is an information-gathering mode" [59, p. 27]. Lastly, *entendre* separates sounds from those properties that relate to source, cause, and space. *Entendre* centers around "a sound's morphological attributes" [59, p. 28], and serves as the primary basis for Schaeffer's reduced listening.

While these four modes are presented as distinct, in practice they almost always happen in concert. To attempt to listen to speech while ignoring the tacit understanding that it comes from a human would be impossible. Likewise, listening exclusively to morphological attributes is beyond the realistic capabilities of most people, with notable exceptions in cases of abnormal attention, such as dyslexia [26, 42]. However, attuning oneself more to a certain mode over another is possible, and is what Schaeffer desired to invoke with musique concrète and the acousmatic experience.

Pierre Schaeffer's modes of listening are the most established, but they are far from the

only theory of listening. In 1993, Michel Chion published *Audio-vision*, and in it he proposed a simplified version of Schaeffer's modes of listening, reducing the possibilities to just three. The first, *causal listening*, is "listening to a sound in order to gather information about its cause (or source)" [29, p. 26]. This encompasses all possible techniques and methodologies for determining a sound source, from unique identification, such as an individual speaker, to generalized identification, such as saying "this sound is metallic". The second category is that of *semantic listening*, which refers to any form of comprehension rooted in translation, that of a language or code. The final category is *reduced listening*, a directed form derived from Schaeffer's *entendre*. "Reduced listening takes the sound... as itself the object to be observed instead of as a vehicle for something else" [29, p. 29]. This is the description of reduced listening most often understood and referenced, at least within English texts. Chion goes on to explicate the requirements and benefits of a reduced listening practice, arguing that it is both practical and useful for enhancing one's understanding of sound.

There is also the problem of time, a notion which remains largely unaddressed in Schaeffer's and Chion's modes of listening. While their modes are reasonable when taking a "snapshot" of a listening experience, they are inconsistent with a temporal understanding of sound. Denis Smalley's *spectromorphology* is an attempt to reconcile this. Spectromorphology is a description of frequency content (the spectrum) over time. This innovation in understanding allowed for descriptions of music theory that couldn't be appropriately analyzed under more traditional methods. Some examples of spectromorphological diagrams can be seen in Figure 3.1. Through visual metaphor, these diagrams directly show a sound's relationship to time. For example, in sub-figure 3, the composite shape on bottom represents a sound that grows from a minimal place, contains some staggered transition, and ends with a growth and sudden release. The power in combining such simple shapes to represent complex sounds is elucidated when looking at more complicated examples, presented in the numerous possibilities shown in sub-figure 6. Smalley says that spectromorphology explicitly solves two problems introduced in reduced listening. First, practicing reduced

listening makes reinterpreting sound *with* its external contexts more difficult. And secondly, "microscopic perceptual scanning tends to highlight less pertinent, low-level, intrinsic detail such that the composer-listener can easily focus too much on background at the expense of foreground" [93, p. 5].

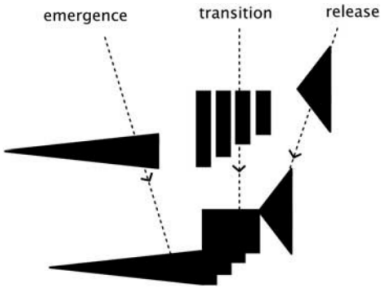



**Figure 3.** Sound unit – *emergence, transition, release*.

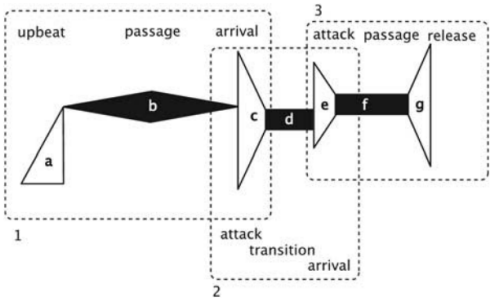



**Figure 5.** Dual functionality in a *morphological string*.

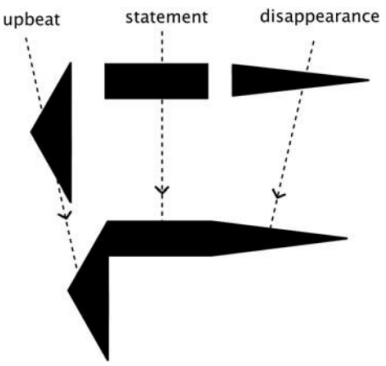



**Figure 4.** Sound unit – *upbeat, statement, disappearance*.

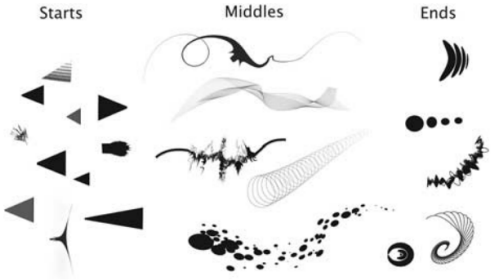



**Figure 6.** Starts, middles and ends.

**Figure 3.1:** Four examples of spectromorphological sonic structural concepts. Source: *The Visual Sound-Shapes of Spectromorphology: an illustrative guide to composition* [21]

## 3.2 Axes of Machine Listening

While there exist different conceptions of machine listening as a technological or cultural phenomenon, none satisfactorily address it through a more artistic lens. For Schaeffer, Chion, and Smalley, their modes of human listening were developed in concert with their desires as a composer. Their audiences are humans, and understanding how they hear is

beneficial to them as music-makers, and can afford new compositional possibilities. In audio surveillance, the machines are the listeners, and one could create and compose sound for their ears. As environmental "extraneous" sound can be thought of as musical in certain contexts or to certain listeners, so too can sounds such as speech be considered compositions to ears designed to hear them as such. I think the field of machine listening could benefit from a consideration of possible modes of machine listening, as has been explored for human listeners.

Perhaps the modes of human listening explored above can find direct application onto machine listening. Both Schaeffer and Smalley's formulations of human listening rely on two primary assumptions. The first is intentionality. Within these modes of listening lies the ability to attune oneself to one over another, often called "attention" or "salience". Attending to certain modes could be an unconscious act, but it could also be an act over which one has conscious control. The second assumption is some degree of cultural knowledge or understanding. This is most obvious with Chion's *semantic listening* or Schaeffer's *comprende*, but it is certainly present in the other modes as well. Being able to place a sound within an environment requires some biological or cultural understanding of the environment and its sonic qualities. Likewise, identifying sound sources from their products requires some prior knowledge of the relationships between certain sound sources and their sonic products.

These two assumptions become potentially problematic when applied to machines. Intentionality does not exist in the same way for a machine listener as it does for a human one. Machines are able to effect action, but the parameters and biases of those choices to act are determined exclusively by their human creators, rather than indirectly by biology and culture. Broader understanding and context is a more difficult issue to parse, but it falls under the same category as intentionality. Machines only have the contexts given to them by their human creators.

Interestingly, this lack of context gives rise to another potential consequence. P. F.

Strawson, in his book *Individuals* [96], posits a world that exists purely in the auditory domain. Strawson states that with the destruction of all other senses comes the destruction of space, and without other contextual information, it would be impossible to separate distance from volume. For example, one couldn't possibly know if a sound is far away or just quiet. Brian Kane summarizes Strawson's argument in *Sound Unseen*, stating with without space, there can be only qualitative descriptions of sound, never quantitative, numerical ones. One couldn't say that two sounds "are the same", or come from the same source; one could only say that two sounds "are alike". The result is that "A purely auditory world, surprisingly enough, turns out to be a world where types or universals, rather than particulars, are primary" [59, p. 145]. To a contextless machine, this would imply that broad classification might be possible, but specific identification is unattainable.

While accepting Strawson's theory as fact might make a reduction of machine listening easier, his philosophy isn't grounded in the reality of acoustics. As John Pierce states in his chapter titled "Hearing in Time and Space" in *Music, Cognition, and Computerized Sound*, "without visual clues we do sense the direction and even the distance of sound sources" [32, p. 89]. The human auditory system is designed in such a way as to facilitate spatial identification of sounds through a complex decoding of the difference in information between our two ears. If a machine listening system has only a single microphone, this collapse of space becomes more relevant, though even then it's not a guarantee. This is not to say that there aren't instances where spatial identification is not possible, even with two ears. In fact, there are a number of ways to trick the ear to hear sound as coming from a falsified direction or distance. But these auditory illusions no more discount the spatialization of aurality as MC Escher's works disprove vision's ability to discern depth. The fact that it's possible at all to perceive space from sound alone seems to discount Strawson's theory and Kane's interpretation of it.

That being said, a slight adjustment of Strawson's theory could pull it back to help contextualize aspects of machine listening. Strawson takes an understanding of the world and

splits it into two categories: qualitative description and quantitative description. As described above, in humans and many machine listening systems, both of these methods are accessible. While this is true for an understanding of sonic comparison, there is another category of understanding that a purely auditory world cannot contain: *extra-aural identification*. With only quantitative comparison, one can identify two sounds as coming from the same source, but one cannot reasonably identify what that source is. The ability to assign external non-sonic identity to a source requires additional senses for corroboration. This adjusted approach to Strawson's theory will help in an understanding of the possibilities and non-possibilities of machine listening.

Another key component in placing machine listening outside of human listening is the role of attention. In practice, human listeners are naturally attuned to certain listening modes over others, but there exists within them the capacity to listen in any mode. And, with training, one can shift the relative importance of the modes to prioritize one over the other, as Schaeffer proposes with the reduced listening practice. This capacity is notably restrained in machines. In any construction analogous to a Schaefferian model, a machine can be designed to have access to only a single listening mode, or prioritize one mode over all others. It is possible to design a machine listening system that listens at multiple hierarchies simultaneously, and most state of the art listening systems do this to some extent. However, the innate human ability to switch and alter listening modes can not be assumed to be present in an all machine listeners.

"Listening" is a verb; in "machine listening", the machine commits an action. Actions do not exist in isolation, but rather follow purpose and intention, whether conscious or unconscious. Importantly, the actions are not the machine's own, but are a result of human construction. Schaeffer's modes of listening can't be directly used to understand machine listening because they rely on a separation of the conscious intention from the unconscious one, and likewise a separation of intention from passive experience. For a machine listener, there is no unconscious, and there is no passive experience. Every act of machine listening is

deliberate, even if its subject happens to be unintended. Chion's modes more closely address the intentionality of the machine listener, but they fail to accurately reflect the real-world functions and intentions of a machine listener. Smalley's spectromorphology holds potential for understanding the technical workings of machine listening systems, but it again fails in reflecting the context in which machine listening takes place.

An illustrative framework for understanding the rise and development of machine listening can be found in J.J. Gibson's theory of ecological optics. In brief, this theory posits that "the environmental niche determines the structure of an animal and its sense" [46, p. 150]. Ecological optics ran counter to the general belief of animal vision at the time, and, though certain aspects of it are no longer in fashion within the perceptual theory community, was hugely influential in our understanding of perceptual systems. The development of vision systems was in response to what Gibson called *affordances*. Affordances exist outside the animal in its environment, and "the affordances of things are what they furnish, for good or ill, that is, what they *afford* the observer" [46, p. 154]. The adaptations of perceptual systems, then, was all about being attuned to affordances, as "organisms need to be attuned to affordances before they can exert their power to shape actions" [46, p. 155].

Gibson's theory provides ample reflection for machine perception. As "beings" who's every adaptation was created for an express purpose, the evolutionary science behind ecological optics parallels the development of machine systems. And, just like for animals, it is essential to any analysis of machine listening to consider the space in which machines exist as well as their purpose for existing. "We must remember that understanding perception requires the joint study of an organism and its environment" [46, p. 155].

By bringing into focus the human intent behind machine listening systems, we also align ourselves with the original model of communication theory as dictated by Claude Shannon. A basic assumption of all forms of communication, as laid out by Shannon in 1948 and discussed more extensively in the Background, is that they carry a message for the purpose of affecting conduct. This is the basis for Level C, the effectiveness problem, which again

reads as follows: how effectively does the received meaning affect conduct in the desired way? From ecological optics we've extracted that all forms of machine listening exist to serve express purposes, and from Shannon we can realize that these purposes are always in the form of affecting *human action* in some way.

An understanding of different common conceptions of human listening and a framework for organizing the intentionality of machines and their creators prepare us to undergo an examination of machine listening in a certain direction. However, we will find that limiting ourselves to only this conception does not adequately cover the spectrum of potential ontologies of machine listening. In the end, different structuring principles are used to create models, and models are always reductions of the world. This is of course what makes them useful, but reductions do not apply equally across all realistic scenarios. With theories of the many different organizing principles discussed above, I present three different axes of machine listening, each providing a new angle for examination. While none are more *right* than the others, some might prove more useful for this thesis and the contextualization of mass audio surveillance.

### 3.2.1 Technologically Determined

One possible structuring principle for differentiating categories of machine listening is the technology necessary to execute different listening functions. This topology allows for a loosely chronological basis for machine listening. It also makes clear a unique trait of machine listening as compared to human listening: the different modes are quite strictly constructive, in that each development in machine listening relies and builds on the previous ones, much like an evolutionary process. A technology-centered approach additionally gives access to a direct questioning of P. F. Strawson's theory. In application, starting from a place of broad, contextless ears, each addition of potential context seems to bring an increasing ability to identify and specify.

### Without Listening

Imagine a microphone plugged into a magical tape recorder with an infinitely long tape. This is the simplest form of a listening machine possible. The microphone observes and the recorder scribes. This system knows nothing and does not attempt to interpret in any form. While certain information about the recording itself is known, such as its sample rate, tape length, and the mic specifications, this information is limited to the act of recording rather than the subject of the recording. It exists solely to observe and record, given no external context and needing none.

There is a strong similarity between *without listening* and Schaeffer's ouïr. Both exist in an unconscious, or perhaps pre-conscious, space. The physical response of one's eardrum to a vibration in the air requires no thought, and is completely autonomic. The primary difference is that, in human listeners, ouïr is not manageable as a distinct action. While one can certainly hear sounds and not think about them, it is impossible to intentionally sever the connection between the ear and the brain. At any moment a sound physically experienced can turn into a sound consciously experienced, through no effort on the part of the listener. This is not the case with a machine. It is possible to create a machine (such as the one described above) that exists only within the bounds of *without listening*, and can perform no other function without external intervention.

*Without listening* serves as the foundation for all digital recording technology, and all modern recorded material. Music, radio, speeches, and countless other messages come out of this mode. It is completely without external context, more a means of transmission than anything else.

### Analytic Listening

Let's connect our ear to a simple brain: our machine is now a microphone hooked up to a computer. The microphone and the tape can record information; with a computer at the other end, the machine can now look at the information. And it does so in the format most

natural to computers: numbers. To a machine, all information is numbers, which is a simple fact with far-reaching consequences to its listening capabilities.

The simplest way a machine can interpret sound data is through numerical analysis. While this would be a large leap from *without listening* in humans, performing mathematical operations on data is the simplest thing a computer can do. Most complex calculations are composed of intricately layered simple problems of addition and multiplication, and thus most methods of analysis take relatively similar amounts of technical complexity to execute. There are a wide variety of forms this analysis can take, and a wide variety of tasks it can be applied to. For example, the relative loudness of a sound can be calculated using the formula $dB = 20 * log10(amplitude)$. The pitch could be analyzed through a spectrogram by calculating the Fourier transform of the signal using the equation $F(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}dx$. A series of simple equations can be put together to detect onsets in the sound signal, a method that could form the basis of rhythmic tracking or phoneme recognition. The machine could calculate the Mel Frequency Cepstral Coefficients, which gives a rough description of the timbre of a sound.

At this level, the machine listener possesses the ability to numerically describe the properties of the sound. What is important for this listening mode is that there is only numerical analysis of the self-contained audio data; the only insights revealed come from within the audio signal itself. While *without listening* contained strong similarities to a biological listener, *analytic listening* contains the computational power of the ear. The correspondence between the stereocilia of our inner ears and a logarithmic representation of the audible frequency range means that this level of frequency analysis is largely done pre-consciously in humans as well. The other forms of analysis are less clear, with timbre and rhythm taking place somewhere in between our ears and higher thinking processes [78, 99].

*Analytic listening* forms the basis of methods for various scientific analyses and creative musical sonic transformations. Even what is probably the simplest form of analysis, that of "volume" determination, is used in nearly every piece of sound technology one would ever

interact with. Numerical analysis is the machine equivalent of basic neural processing, and is similarly fundamental to any type of sonic "understanding".

### Relational Listening

We are now ready for a step that is almost always implicit in human listeners, but has not yet been covered in the previous two categories. Our brain now contains knowledge, or a computational approximation of it. I am defining computational knowledge as having access to information that originated outside of a computer's enclosed system. This is rather abstract, but only by necessity, as it can take near infinite forms.

One of greatest possibilities gained through *relational listening* is that of comparison. What previously existed only in isolation can now be mathematically compared and placed within a context. For example, with outside knowledge a machine listener can say that, given a frequency spectrogram of a piece of audio, the sound likely originated from a human voice. Sounds can be compared directly with other sounds, allowing the possibility of finding sounds that are similar to other sounds. In other words, with comparison comes qualitative analysis, which is possibly a much more "human" interpretation than isolated self-contained forms of analysis.

With the allowance of external communication, the creation of shared knowledge bases becomes available. A machine listener can become part of a network, and that network can share information with other networks. The creation of such databases removes previous limitations on locality and access, allowing machine listening systems to be vastly greater in size than previously possible.

### Learned Listening

The final piece of technology that paves the way for a new kind of listening is that of machine learning. This technology, often thought of as a machine approximation of "intelligence", generally refers to any algorithm that can expand its knowledge and improve itself. In the

same way that a least squares regression line is used in introductory statistics to approximate the relationship between simple data, machine learning uses statistical methods to try and model nonlinear relationships between data. The primary function of machine learning is to build an accurate representation of data in order to precisely predict the "outcome" of future input.

In *relational listening*, our machine was given the possibility of context. With *learned listening*, context becomes the central component to the process of inference, which is at the heart of machine learning. While machine learning allows for some kind of understanding of complex relationships, machine learning systems still have to be given data by humans, and that data carries with it implications about present and missing contexts. For example, an algorithm trained to predict health outcomes of patients diagnosed with heart disease might be trained using all the patients' health data but exclude information about their income, which could have a huge impact on their ability to make significant lifestyle changes. The ability to consider information not provided is simply not an option with machine learning, and thus, while *learned listening* is the closest to a human cultural listening that a machine can be, but it is still lacking in essential ways.

Examples of *learned listening* can be found in corporate enterprises, government projects, and academic research. Modern text-to-speech technology relies on machine learning, as does music recommendation, and speaker identification. Nonlinear predictions are useful in any number of big-data contexts, both broad and personal.

### 3.2.2 Purpose Driven

A disadvantage of technologically determined machine listening is its inconsideration of the human. Rather than springing forth via spontaneous generation, machines with the capability to listen are created by humans. Focusing only on the technology biases the space of machine listening towards a machine understanding, where cultural and human context isn't important. Lumping machine listening into categories of technology orients

musique concrète as closer to political propaganda than spectralism. While these points of comparison may be interesting to explore, they fail to capture the human, which is essential to the consideration of audio surveillance.

To enforce the human element of machine listening systems, we will go back to Gibson's ecological optics. If machine listeners are created to serve a purpose, what is that purpose? Perhaps organizing machine listening by human intentions yields a more practically robust and inclusive categorization. In such a topology, the ghost hand of the programming is made evident, and machine listening can be accurately discussed as a tool used with intentions.

One possible category in this formulation could be *to remember*. Here is where we would find machine listening employed for recording music, home videos, speeches, or events. The purpose of these methods of listening is to capture their product for posterity, either personally or at a larger cultural level. Such sounds could be used for later enjoyment or as cultural artifacts, but their capture itself is the goal of this category. Another category might be *to identify*. Here lie many types of analysis, used to calculate different qualities of a sound. Audio fingerprinting, source separation, and music analysis tools would fit in this category. The properties of the sounds are information in of themselves, and this category seeks to quantify them for extraction and comparison.

A final category might be *to understand*. Whereas identification seeks to place sounds in a sonic context, understanding tries to find a broader, extra-aural context for the creation and result of sonic events. An example of this could be a machine listening systems created to understand environmental conditions in a forest. There is significant evidence for the effects of different medical conditions on the human voice [83], and a system that automatically detected these and suggested diagnoses would certainly be an attempt *to understand*. An automated speech recognition system would fall under identification, but a system that includes some non-syntactic inference about the topic being discussed would be in this category.

### 3.2.3 Action Oriented

While it's headed in the right direction, a purpose driven framework for machine listening stops one step short. Intents may well be satisfactory for certain tasks, but machine listening is ultimately an act of communication. There exists sound, or some aspect of sound, in the world that humans use machine listening to access. In this respect, machine listening is a kind of translation, communicating audition to human participants in formats of varying directness from transliteration to syntactic mapping to poetic interpretation. As Claude Shannon showed, communication exists to affect action, and this interpretation of intention allows us to more accurately pinpoint the guiding force behind the machine systems: a force that can be human, but can also be cultural, incorporated, and political. The recognition of action as the driving motivation allows us to directly interrogate the desired effects of machine listening. The following categories are not mutually exclusive, but rather seek to try and envelop as much of machine listening as possible.

#### Consumption

Some machine listening systems exist to encourage consumption. These systems are born out of capitalism, created to push products to generate value. The machine listening systems themselves are not the goal, but serve as a kind of indirect advertisement for their products.

A clear example of a consumption-driven machine listening system would be a piece of music recommendation software, the backbone of companies like Spotify, Pandora, and Google Play Music. These programs are built to analyze and relate sounds for the express purpose of encouraging continued use of their respective platforms. The music chosen is meant to be enjoyable, but only insofar as user satisfaction equates to product consumption and capital gain. One could cynically argue that applied machine listening in music production software is also for consumption, and though this case may be easier to make for some musics than others, I don't seek to contain all of machine listening's creative applications to this category.

### Learning

A second purpose for machine listening is to promote learning of some kind. For example, this could be in a scientific context, encompassing projects like environmental soundscape studies or linguistic pronunciation research. This category would also include medical applications, where machine listening is used for automated analysis of various body signals in an attempt to track and predict the body's physical conditions. If the outcome of machine listening is an attempt to learn for some advancement of knowledge, then it fits in this category.

With that in mind, this is a somewhat disingenuous category. After all, science does not truly seek to learn for learning's sake, but follows the agendas of the people conducting the research and the people funding them. A soundscape study could be conducted in an urban environment under the guise of studying environmental noise levels, when it could ultimately serve the function of protecting local industry from noise ordinances or encouraging population flight towards a "quieter" part of town. Learning is an action, but it is never without other motivations, though certainly not always negative. What this category seeks to include are examples where the ultimate action-oriented goal is separate from the third and final category.

### Control

Another function for machine listening is to facilitate the control of people through disenfranchisement and a negation of their agency. This is communication as a structure that gives power: not abstract power, but direct power over another. Power can take many shapes, but essential to any modern model of control is information. Some physical control could be taken without a control of information, and potentially even be held as such, but in any instance outside of brute threat of violence, power over information is necessary. Just as eyes let you see people's actions, ears let you hear their discussions, plans, and lives. If one's goal is to control the actions of others, machine listening can exist as a tool to translate

their personal, interpersonal, and embodied information to that end.

## 3.3 Placing Mass Audio Surveillance

Audio surveillance can take many forms. While the application of specific technologies can enable new possibilities for surveillance, it is clear that surveillance can, and does, exist at all levels of technology. The key separation between audio surveillance and mass audio surveillance, however, is the application of automation, whether that takes place in the collection of data, its analysis, or its inference. This automation has several effects, all of which serve to further mass surveillance's goal of controlling speech, movement, and existence.

In the world of technology and innovation, automation is used as a way to defer morality. Decisions are made by the machines, and in instances where machines make immoral decisions, they are forgiven—and their designers at most scolded—because of their perceived amorality. While this is an essential part of automation, it seems almost irrelevant when discussing surveillance, since the aforementioned decisions have already been established as moral by those in power. It's true that in public perception this deferment might be inconsequential. It's also true that arguing for the correct attribution of the immorality of these decisions is possibly unimportant when the problem being addressed is immoral in of itself. If one believes that surveillance is wrong to perpetrate, then the specific entity that perpetrates it doesn't make much of a difference. The replacement of human ears with machine ones does not change the base morality of surveillance, but it does change its scope, which places its morality into even starker relief.

While morality may be maintained, the logic underlying surveillance might not hold up for machine listening. As discussed previously, P. F. Strawson hypothesized that, in a purely auditory world, only qualitative identities are possible. The adjustment to that theory added that, while both qualitative and quantitative comparisons are possible, non-sonic (or

extra-aural) identification remains inaccessible. This is antithetical to the purpose of mass audio surveillance, where algorithms devoid of other senses attempt to identify specifics, from language to environments to individuals. From Strawson's perspective, such specificity is categorically impossible with only aurality. Yet it appears that machine listeners are used to identify; the inclusion of methods of signal processing certainly give numerical, quantitative information.

The advantages gained from giving machine listening systems complex technology are found only in being able to speak with increased specificity about the *sound itself*. It is undeniable that machine listening systems can determine incredibly minuscule, accurate details about sounds. Increasingly, we're even seeing systems that can calculate larger, macro-scale sound information, which is a historically much more difficult task. But this information does not reveal anything outside of the domain of sound. Inference beyond aurality cannot be accomplished through aurality alone, and machines are not exempt from this constraint.

Looking at machine listening systems, it is apparent that these algorithms are being used as though they are. One can use an algorithm to approximate a sound's timbre by calculating its Mel Frequency Cepstral Coefficients, but the next step, identifying the instrument creating the timbral markers, is not. That being said, an approximation can be made, which is what machine learning relies on; in the end, all machine learning is a game of maximizing probability rather than establishing truths. Yet a reasonable approximation for instrument identification becomes an unreasonable one when applied to a valuation of human intent and a determining factor of human life.

Audio surveillance is an example of using technology for a function for which it is ill-equipped. As a means of enabling the control of peoples, mass audio surveillance takes advantage of the fundamental human right and capacity for sonic communication and misuses technology to place capitol value on that communication. This oppressive act is only possible through the use of technology and its application in machine listening. Having

built some theoretical foundation on which to stand, the specifics of audio surveillance can be examined in practice. Through repurposing the same technology used in mass audio surveillance, one can begin to counteract its existence.

# 4

# The Acoustic Counterfeit Machine

Having explored the theoretical foundations of mass audio surveillance, finding its hyper-focus on identity and reliance on machine listening, an opportunity arises to use the properties of the surveillance system against itself. This chapter contains the implementation of an anti-surveillance tool inspired by that idea. Named the Acoustic Counterfeit Machine, or ACM, this system is designed to hide speech from methods of mass audio surveillance, and to do so in such a way as to not arouse suspicion. The ACM is built around a neural network that matches acoustically similar sounds between source speech and a database. After a brief overview, its complete structure is described in detail and its effectiveness evaluated.

## 4.1 System Overview

An immediate problem in attempting to build an anti-surveillance system is that surveillance systems, corporate or state, are unavailable for public perusal. While we cannot know with absolute certainty how mass audio surveillance systems operate, we can take an educated guess using a combination of the recent government leaks [15, 47], the public effects of surveillance [97], and cutting-edge technologies of signal processing and machine learning [35, 44, 101].

In a 2014 paper titled "Audio Surveillance: a Systematic Review", Marco Crocco et al. present a thorough overview of the technology of audio surveillance used specifically for automated contexts [35]. They outline four primary steps of audio surveillance: background subtraction, event classification, object tracking, and situation analysis. For each category, they present the most relevant foundational and contemporary research, giving an in-depth look into the details of surveillance technologies. The authors are quite clear on their intention for publishing such a review: "the present [survey] is specifically targeted to automated surveillance, highlighting the target applications of each described methods and providing the reader tables and schemes useful to retrieve the most suited algorithms for a specific requirement" [35, p. 1]. This explicitly pro-surveillance attitude isn't surprising or unique within the scientific literature. At the time of this writing, the first four pages of Google Scholar results for the term "audio surveillance" contain exactly one paper not designed for the purpose of executing or enhancing surveillance. The one upside of this general scientific appreciation for surveillance is that academic research in the subject is public.

The following is a generic architecture of such a system, as adapted from "Audio Surveillance: a Systematic Review" [35]:

1. Audio is indiscriminately captured from countless microphones

2. Audio undergoes background subtraction, isolating the foreground sounds from envi-

ronmental noise

3. Audio is analyzed in real-time to determine the nature of its content as a classification problem, with such categories as speech, music, environment, etc.

4. Audio that matches the class attempting to be surveilled is further analyzed in a class-respective manner

5. The relevance of this further analysis is calculated by an algorithm

6. Any results deemed appropriately relevant are sent to a corresponding database or human observer

Based on the paper by Crocco et al., an example of a mass audio surveillance system looking for keywords in speech might read as follows: first, sound that microphones receive is analyzed using a deep neural network to assign it a class. These neural networks are computer algorithms trained to take input they've never seen before and give reasonable output based on their prior experience. In this case, the neural network would receive incoming audio and, based on its training data, attempt to place it within a general category. Example classes might be: speech, noise, tones, rhythmic patterns, etc. Importantly, it is likely that all such systems would also include a class for "other". Any audio that's classified as "speech" is decoded using an automatic speech recognition (ASR) algorithm. The ASR output would be scanned for any instances of the specified keywords. If any are found, the transcript and audio is sent to a database with tags specifying the keywords heard, the location and time of the recording, and any other relevant information. This is the first point in the process where it is likely a human would have access to the data. This is largely because of the pure volume of data being collected and analyzed. If observers were actively listening in at an earlier step, the whole of the United States would have to be employed to surveil itself.

The goal of the Acoustic Counterfeit Machine (ACM) from a technical perspective is to hide sounds from methods of mass audio surveillance. But what does it mean to *hide*

from a machine listening system as described above? Specifically, to hinder the ability for a surveillance system to perform the fourth step, when the audio is analyzed in a class-respective manner (ASR in the case of speech).

There are a few reasons to not attempt intervention at previous or later steps. Firstly, a *direct stoppage* of any of the steps past the first would require a software installation on the machine either collecting or processing the audio. Access to the machine processing the audio is rarely an option; it's possible the processing may happen on the device doing the recording (e.g. a phone or laptop), but it's next-to-impossible to know if that's the case, or if the audio is instead being sent elsewhere to process. Trying to disrupt the recording process of the microphone would be reasonable if one desires to secure a single device from being used for surveillance, but loses feasibility when considering the wide array of microphones available for listening. Both of these also assume an unbreakable security and robustness of a software solution. While this assumption underlies much of modern-day life, it is tenuous at best; major software hacks and database leaks are almost weekly news stories. If a team of professional security experts have difficulty with those issues, I would hesitate to offer a solution myself.

The Acoustic Counterfeit Machine is designed to mask audio in the physical domain, before it ever reaches a microphone. This gives it the advantage of being a self-contained system that cannot be accessed through any external connection. The key idea of my system that makes it unique compared to other attempts at masking is that it seeks to alter the audio content to confuse step three while *not* switching the classification of step two. In other words, if the target audio is speech, my goal is to make the microphone hear *other speech*, rather than noise, silence, or any other sound.

Why is this important at all? For example, physics tells us that there's no better acoustic mask than pure white noise, which (theoretically) contains all possible frequencies at once. As early as 1950 we knew that, if one placed a speaker in a room and played incredibly loud white noise, it would be impossible for any algorithm to reasonably reconstruct human

speech from the resulting mix [51]. This is certainly true at multiple steps in the process as well, as an algorithm designed to classify sound would never classify white noise as speech. However, such an algorithm would likely have a class corresponding to "other" or "suspicious" sounds. A sudden shift in an environment from silence or speech to loud white noise would almost certainly be categorized as immediately suspicious in some regard, and would be flagged for further examination. By the automated nature of mass audio surveillance, this highly effective masking only serves to draw attention to the fact that sound is being hidden. By making an acoustic mask that maintains the contextual domain of the original sound, my system can effectively hide sounds without broadcasting their non-presence.

The construction of the ACM is shown in Figure 4.1. The architecture, which I will explain in depth in the subsequent sections, is briefly outlined below:

1. A comparative database is created to calculate approximate matches between the spectral features of spoken words

2. That database is used to generate training data to an LSTM (Long Short Term Memory) neural network, designed to match spectrally similar audio in a real-time context

3. Live input is fed through the neural network and an ideal mask is calculated

4. The calculate mask is played over a speaker, with a delay of <15ms between the reception of the input audio and the sonification of the matching mask

## 4.2 Approximate Matching Database

The purpose of the neural network in the ACM is to take incoming speech sound and convert it to *similar sounding* speech sound. To train a neural network, one needs a huge amount of data, and this network needs that data in the form of an input sound and a corresponding

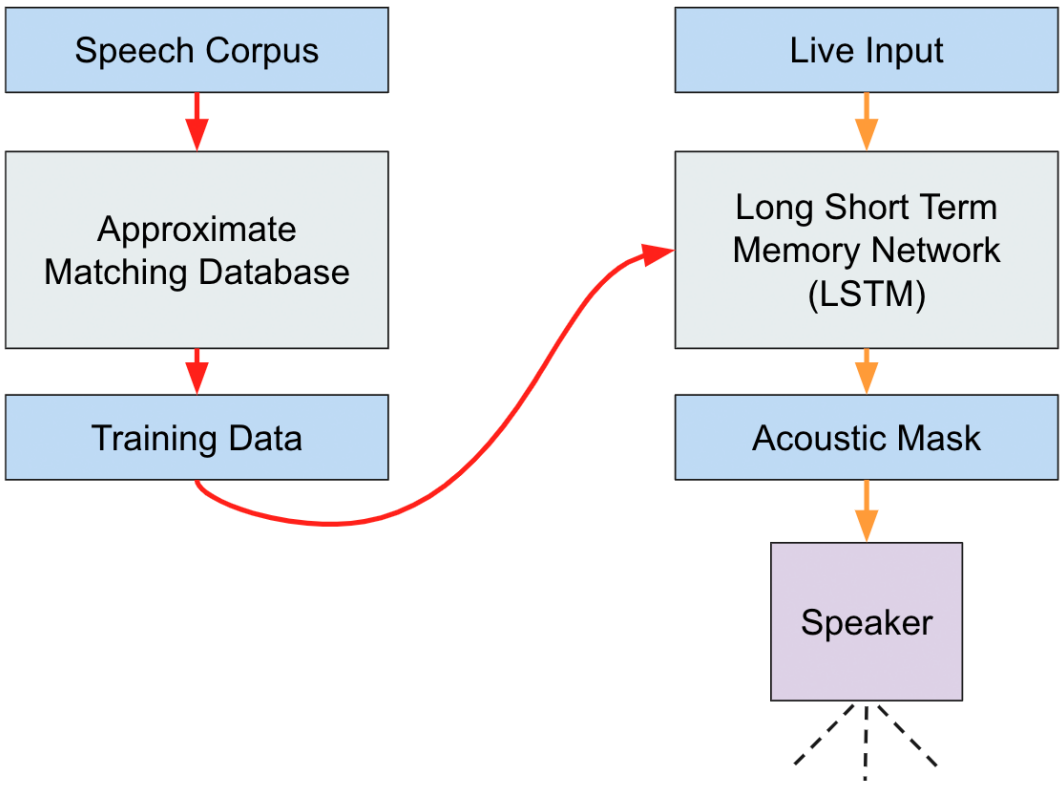



**Figure 4.1:** An outline of the Acoustic Counterfeit Machine architecture. First, a corpus of speech data is used to populate an approximate matching database. This is used to generate training data to a neural network. In a real-time context, live input is fed through the neural network and the output is used to generate an ideal acoustic mask, which is then played through a speaker.

"matched" sound. There does not exist a database of matched sounds, nor does there exist a front-facing database interface that allows one to compare new sounds to the ones found in the database using arbitrary feature sets. Thus, it quickly became clear that I needed to create my own database construction that allows for flexible and speedy sonic matching.

The first step in this process was procuring a corpus of recorded audio. Because my system was meant to operate on speech at a reasonable perceptual time-scale, I wanted to create the database (and thus the whole system) around individual spoken words. I used a compilation of six different data sets, all of which are available for free download in the FLAC file format. The list was from a compilation by Shtooka, a multilingual database of audio recordings of words and sentences [7]. Three of the data sets I used were sourced

directly from Shtooka (titled: *eng-balm-emmanuel, eng-balm-judith, eng-balm-verbs*), one was from the Wikimedia Foundation's pronunciation database (titled: *eng-wcp-us*), and two were from WimsEdu (titled: *eng-wims-mary, eng-wims-mary-num*). The files are from four different speakers with the exception of the Wikimedia Foundation data set, which has many. These six data sets were the entirety of the available corpus that met the requirements of being individual audio files of single spoken English words. In total, the complied corpus consists of 14,449 audio files, each corresponding to a single spoken word. The spoken words were not necessarily distinct, meaning that there could be (and certainly are) multiple recordings of the same word spoken by different individuals. This number of files may seem excessive, but in the context of speech training its quite minimal. For a commercial system, one would want as many speakers as possible for as many words as possible, ideally numbering in the hundreds-of-thousands, if not millions. For example, Google's Speech-to-Text transcription software has the entirety of YouTube as only one of many sources to draw data from. In addition, live systems can retrain as they operate; any product that uses Google Speech-to-Text is constantly sending their data back to retrain and improve the networks. For the scope of this version of the ACM, however, $14,449$ files were enough to train and operate with reasonable accuracy and efficiency.

With all the sounds in one place, the next step was defining the structure of the database. The most similar class of problems to draw solutions from are those of audio fingerprinting. Audio fingerprinting is the task of finding or creating a sound's unique fingerprint, or lower-dimensional identifying characteristic set, which can then be compared to other fingerprints for quick and efficient searching. A simple diagram illustrating a standard audio fingerprinting setup can be seen in Figure 4.2. Matching sound to other sound is a difficult task as an audio file cannot be described by a single number or a simple vector. I wanted to have the flexibility of attempting to match files based on a number of different audio features, so a matching algorithm was needed that could adapt to map different high-dimensional feature spaces. Additionally, a strict audio fingerprinting database wouldn't work for this

task, since its purpose is to find *approximate* matches rather than exact ones. What I settled on for the approximate mapping was using locality-sensitive hashing, or LSH. Hashing is the process of converting input data to an identifying data tag of a set size. Hashing is used in many different technologies, though it's most-often discussed in relation to cryptography. The purpose of locality-sensitive hashing is to reduce the dimensionality of a data set while attempting to preserve statistical likelihood that similar data in the higher-dimensional space will end up nearby each other in the lower-dimensional space. LSH has a strong history of use in matching contexts where efficient searching is of a high priority [68, 84, 91].

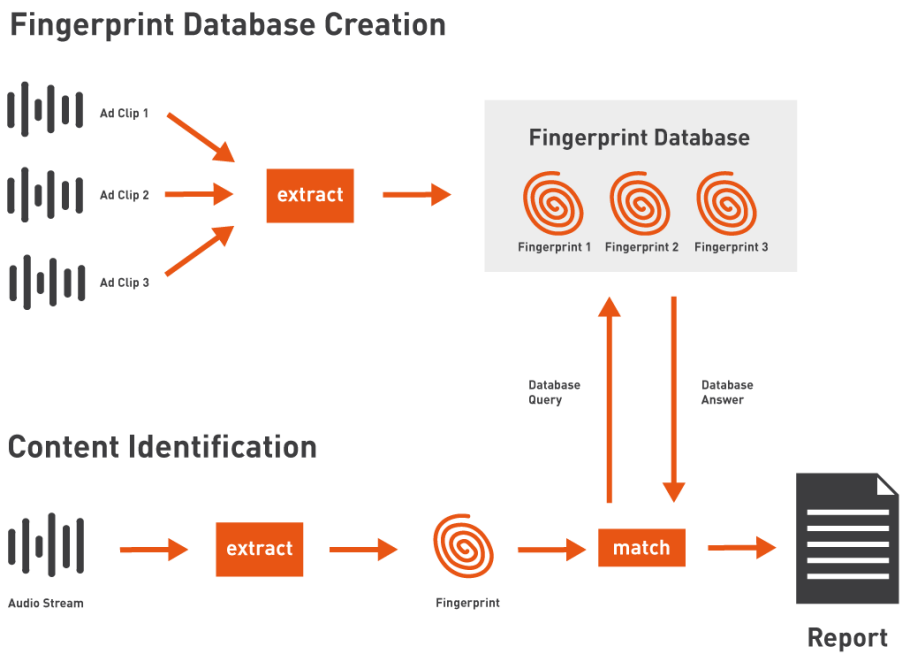



**Figure 4.2:** A simple commercial audio fingerprinting system. Source: https://www.mufin.com/company/technology/

Now that the database structure was organized, the database needed to be created and tested. Functionality was created to allow for different feature mapping of the database, including mel-scaled spectrograms (henceforth called mel spectrograms), mel-frequency cepstral coefficients, and root-mean-squared energies. Loosely, mel spectrograms are a

description of frequency content, biased towards the non-linear representations of frequency in human speech [23]; mel-frequency cepstral coefficients (MFCCs) are a description of timbral qualities [76]; root-mean-squared (RMS) energies are a description of volume [64]. The database was built and tested with each of these features, but ultimately the mel spectrogram was chosen. This is partly because it contains more speech-variant information compared to an MFCC or RMS, and partly because the final step of the system, the acoustic masking, was chosen to be frequency-reliant. To make the masking process as simple as possible, sounds with similar perceptual spectral content to the live input were desired, and so the machine learning system needed training data that corresponded to that relationship. While standard frequency could have been used, mel-scaled spectrograms represent *quefrency*-domain information, where quefrency is the spectrum of the log of a waveform in the time-domain [56]. As quefrency is a closer approximation of the frequency banding in human speech, the database was set up to match across the quefrency domain.

Of the 14,449 files available, a random selection of 10,000 were used to build the approximate matching database. The remaining 4,449 were queried against the database as input, and their corresponding matches were recorded. To test the database, the 10,000 original files were also queried as input. They all returned their original audio fingerprint, and thus constitute "exact matches", confirming that the LSH function was properly used. A selection of the 4,449 "inexact matches" were compared to their matching pairs using a distance calculation on their respective mel spectrograms. They were found to be very close to their calculated matches, further validating the LSH functionality. Subjectively, the inexact matches shared many similar properties of the original source files, often both corresponding to a similar pattern of consonance and vowels, especially with regard to plosives, as well as a similar tonal inflection.

At the end of the database creation, two text files were written: one that stored the file names of the "exact matches" (which match onto themselves), and one that stored the file names of the "inexact matches" along with their corresponding matched file (one of the

10,000). This construction allowed the training of the machine learning algorithm to be tested under ideal (exact) and non-ideal (inexact) conditions.

## 4.3 Neural Network

With an approximate matching database, it's possible to match incoming sound to a previously-identified one. So why does the matching portion of the system not just stop there? The primary issue with only using an approximate matching system is about time. Possibly the greatest unsolved mystery with computational audio analysis is how to consider time. In a generalized sense, audio can carry important information in a variety of timescales, even simultaneously. For example, in a recording of the first movement of Beethoven's Fifth Symphony, one could only consider time at a minute scale, where the most salient information would be found regarding pitch envelopes and attack timings. At a slightly larger timescale, one could analyze the piece on the level of eighth notes and build a harmonic road map of the pitch progressions. Zoomed out even more, one could attempt to group phrases together and infer some qualities about the unique manner in which Beethoven starts and ends a musical phrase. Finally, looking at the entire movement as a single sonic entity might lend to fascinating insights about the regurgitation and rewiring of musical components within the movement.

It's rare that issues of timing are even that simple, however. The above example would be relatively trivial if we were analyzing a midi piano roll, where each note came at an exact time and distances between rhythms were exactly measured. In a live recording of the symphony, there are frequent tempo changes, ritardandos, accelerandos, and other forms of rubato. While one's chosen timescale might accurately segment the first two notes from each other, it is likely it would be incredibly offset by halfway through the opening motif.

These issues carry over to speech, which is the chosen medium of this system. Ideally, the design is such that there is some built-in understanding of a *single word*. Yet there

is clearly no set length for what a single word sounds like. The 14,449 recordings vary in length from 0.5 seconds to 3 seconds. While this may seem like a small difference to a human brain, it's incredibly challenging for a computational system to understand data at such different timescales. This is the reason why the system has at its center a neural network, and why a Long Short Term Memory (LSTM) architecture was chosen.

Because of their ability to operate on longer and variant timescales, LSTMs have long been a popular choice among audio researchers. In 2000, F.A. Gers and J. Schmidhuber suggested an exploration into using LSTMs for musical rhythmic analysis in their paper "Recurrent nets that time and count" [45]. Shortly thereafter, a number of studies using LSTM for music generation tasks were published with supportive and encouraging results [38, 39]. More recently, LSTMs have been used to identify larger musical structures [37], model and compose polyphonic music [48, 66, 67], perform automated speech recognition [48], and analyze music's emotional affect [33].

An LSTM is a widely used sub-type of a recurrent neural network. While neural networks are designed to learn non-linear relationships between inputs and outputs [52], they don't generally have any concept of *memory*. Each input is separate from each other input, and so a standard neural network considers each in isolation. Any task that relies on some sequential or cause-effect relationship can't be naturally understood by a neural network. This presents a huge problem in any signal analysis, since earlier signal information is generally needed to understand later signal information.

Recurrent neural networks (RNNs) seek to solve this problem by introducing the idea of persistence [22]. Information from an RNN node is used as part of the input to the following node. With the ability to pass information from one time step to the next, RNNs have a kind of proto-memory, allowing them to use past information to help in their predictions. In a classic RNN, however, that memory is fairly limited [19]. If a network needs information that it received three cycles ago that's alright, but if it needs to learn over a longer period then the information is lost too quickly [54]. Additionally, traditional neural networks have

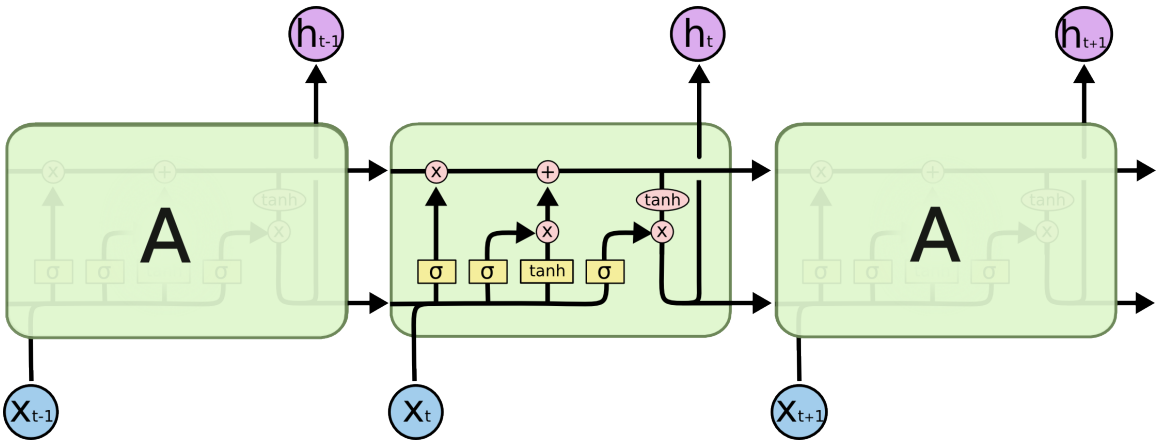



**Figure 4.3:** A diagram of the structure of an LSTM module. The yellow rectangles represent neural network layers and the red circles represent mathematical operations. $X_t$ is a time-slice input, $h_t$ is an output, $\sigma$ is a sigmoid layer, and $\tanh$ is a tanh layer. $x$ and $+$ are multiplication and addition, respectively. Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

less flexibility with *varying* timescales, where certain data is longer or shorter than other data.

Long short term memory networks (LSTMs) are designed to solve both of those problems. Their structure, outlined in Figure 4.3, is designed specifically to hold onto information for a very long time. First described in a 1997 paper by S. Hochreiter and J. Schmidhuber [55], the structure of LSTMs is more complicated than a standard RNN, with four layers instead of just one, but the important aspect to understand is that information is largely persistent from cell to cell. The horizontal arrow across the top section of the cell shows that information is passed directly from a previous state to the following state, with only slight modification from the current cell's state. This is the key feature that allows for longer sustained memory.

The task I used the LSTM for was to classify incoming audio as belonging to a matched audio class, as determined by the approximate matching system. The training and testing data was generated by first selecting random classes according to how many were desired for training in proportion to the total number of exact vs. inexact matched sounds. In other words, with exact matches constituting approximately 70% of all available sounds, there is a 70% chance that any randomly chosen class will be an exact match, and a 30% chance it

will be an inexact match. The maximum number of classes I could have chosen was 10,000, as any inexact match would be classified as "belonging to" one of the exact matched classes. While working on the system, the task was originally one of binary classification, attempting to tell the difference between only two audio samples. When that was sufficiently successful the task was expanded to an increasing number of classes.

The concatenated training signal then needs to be processed and segmented in accordance with the LSTM structure. LSTMs take as input sequences of data, where each sequence is often called a batch. Though I tried different ways of approaching this step, since the system is run in a real-time setting, the training data needed to be formatted as it would be in the real-time setting. I went through and chopped the audio into segments approximately $0.279$ seconds long that overlapped with a hop size of approximately $0.003$ seconds. The spectrogram of each segment is calculated individually, then scaled to fit between $0$ and $1$. The spectrogram is transposed so the time-steps, which were previously columns, are now rows. Finally, the first $n$ time steps are put in a list, representing a single sequence of data. Correspondingly, a single class value from the concatenated class list is added to a training data set, representing the matched class for the first time step of the audio segment. This process is iteratively calculated over each segment in the extended audio file until two final sets of data are created: a list of time-step-oriented spectrogram sequences and a list of matching class identifications. This is the final form of the training data for the LSTM.

An equal number of instances of each selected audio file were loaded for the training data, and their order was randomly chosen. For the majority of trials, $30$ copies of each audio file were loaded in a randomized order. The audio was concatenated into a single audio file of extended length. As each instance was loaded into the training data it underwent a series of semi-random transformations. Each of the transformations were meant to both boost the data—a practice which increases the noise in a given data set, which can both help in neural network training accuracy and in avoiding over-fitting—and make the system more robust to real-world acoustic possibilities that the system may encounter. First, the

files were time-stretched (with maintained pitch) using Librosa's *stretch* function, which uses a phase vocoder time-stretch method. The quantity and direction of the stretch was in accordance with a random Gaussian centered at $1.0$ with a standard deviation of $0.1$.

Next, a random signal to noise ration (SNR), determined by a random dB value between $0$ and $60$, was achieved with the addition of Brown noise. As opposed to white noise, which is a signal with all possible frequencies present at equal amplitude, brown noise has an amplitude rolloff of $-3dB$ per octave. Brown noise was chosen because of its similarity to arbitrary environmental noise. Finally, the signal was modulated with a linear amplitude ramp on the beginning and the end of the file, the length of the ramps being determined by random numbers up to a max length of $0.25$ seconds. This was to make the system more robust to potential differences in vocal attack or decay. For example, "happy" can be spoken with a variety of different pronunciations of "h", ranging from hard to soft. However, since it's likely the rest of the word would be pronounced the same regardless, all of those instances should ideally map to the same audio file of someone saying "happy". Once all transforms were applied, the resulting signal was added to the previous concatenated signals. Simultaneous to this process, a single concatenated list was created containing the correct matched class identification at each time step.

The chosen classes along with their transformation parameters were saved in a text file so that later instances could track and use the same data for bug-fixing and optimizing.

At this point, there are two large lists of numbers for the training data: one of audio signals and one of corresponding class identifications. The data formatting is not yet finished for two reasons. First, the training data must be in the form of mel spectrograms instead of audio, as this was how the task was structured and how the approximate matching database was organized. Secondly, a long list of data cannot automatically be segmented into separate sub-sequences of data. LSTMs can take many different forms of input, but essential to their application is the use of sequences of data. This process, detailed below, can be seen in the *batch* function in Listing 4.4.

The concatenated training audio file is split and walked through with a window size of $6144$ (or $2048 * 3$) frames and a hope size of $64$ frames. At each step, the window of audio is right-padded (concatenated on the time-wise end of the window) with $2048$ frames of $0$s. A longer window for the spectrogram calculation allows for better estimation of low-frequency values, as the FFT algorithm requires a signal to exhibit a number of repetitions before it is recognized as a frequency presence [87]. Allowing the signal to complete repetitions before the end of the spectrogram window is essential, so while 0-padding doesn't add any frequency information, it ultimately increases frequency resolution. The reason a longer audio window isn't taken is because the length of the audio window is equal to the amount of delay between input and calculation in the real-time version of the system. Thus, the smaller the audio window taken, the shorter the delay. The window size remains as small as it is because human speech doesn't generally fall below $85$ Hz [61]. Additionally, padding the signal to a total length of $2048 * 4$ results in a signal of length a power of $2$, which greatly speeds up the spectrogram calculation.

Once the sub-signal is padded to the appropriate length, the mel spectrogram is calculated on the sub-signal. The mel spectrogram operates with $128$ mels, a window size of $2048$, and a hop length of $512$. While lower mel values can be sufficient for certain audio tasks, and are indeed much more space- and time-efficient, for high frequency resolution tasks such as speech, a full 128 mels worked best, as shown both in theory and in my testing. A sequence of spectrogram information is then created using shingling, a term originating from internet search algorithms but which has significant advantages over other methods when applied in audio contexts [27]. The spectrogram is transposed, such that the time-steps, which were previously represented in the spectrogram's columns, are now represented in its rows, with the frequency information in its columns. Though not a part of shingling, the data is then energy normalized to ensure robustness to general differences in energy amplitude. The data sequence is created by appending the first $n$ rows of the transposed mel spectrogram into one list. Different values of $n$ were tested, and ultimately a sequence length of $n = 10$

samples ( 0.3 seconds of audio) proved to result in the highest accuracy.

```
def batch(signal, matches, hop_length = 512/8):

    signal_batch_length = 2048*3
    data = []
    classes = []
    batched_frames = []
    cur_frame_count = 0
    num_to_add  = signal_batch_length

    while cur_frame_count < len(signal):
        batched_frames.extend(signal[cur_frame_count : cur_frame_count +
    num_to_add - 1])
        recent_signal = np.asarray(batched_frames)
        recent_signal = np.pad(recent_signal, (0, 2048), 'constant',
   constant_values=(0.0,0.0))
        spec = get_spectrogram(recent_signal, 22050, n_mels=n_mels,
   display=False)
        transposed = spec.T
        scaler = MinMaxScaler(feature_range=(0, 1))
        transposed = scaler.fit_transform(transposed)
        comp_cols = []
        for i in range(0,seq_length):
            comp_cols.append(transposed[i])
        data.append(comp_cols)
        classes.append(int(matches[cur_frame_count-1]))

        # Prepare variables for next iteration
        batched_frames = batched_frames[int(hop_length)-1:]
        cur_frame_count += num_to_add
        num_to_add = int(hop_length)

    return data, classes
```

**Figure 4.4:** Function to split a batch of training data into sequences of correctly ordered spectrogram data.

The above steps are taken to calculate each sequence of training data. Once a sequence is calculated, it's added to a growing list of all calculated sequences. A single value for the class corresponding to the sequence's last value is added to a corresponding list of class sequence data. Then the audio window is shifted over by 64 frames and another sequence is calculated.

For a small number of classes, the data can be organized just as stated and then fed to the neural network to train. However, as the number of classes increases, the amount of memory

and CPU usage becomes intractable. To solve this problem, I switched to a mini-batch structure of training. The only significant difference in mini-batch learning is that the data is fed to the network in small "batches", instead of all at once. Thus, instead of calculating the entirety of the training sequence at once, a portion of it can be calculated and then fed to the neural network to train before the next portion is calculated. This saves significant memory usage and makes it possible to train with much larger amounts of data. Different numbers of sequences were tested for the size of the mini-batches, but I ultimately settled on $5000$ sequences-per-mini-batch.

I arrived at the technical details of the LSTM architecture by starting with similarly-purposed LSTMs from the scientific literature and adjusting the structure until the network was optimized for my task [CITATIONS NEEDED]. The machine learning python library used was *Keras* [3], which is built on top of *TensorFlow* [10]. The code for the creation of the LSTM model can be seen in Listing 4.5. The network was built as a Sequential structure, which allows for intuitive stacking of neural network layers by letting Keras handle some of the low-level parameterization behind the scenes. The network only has two layers: one visible layer that accepts the input data, and one hidden LSTM layer with $l$ LSTM neurons or blocks. The input data was in the shape of $(10, 128)$, which corresponds to a single sequence of $10$ sequential $128$-dimension mel spectrogram time slices. The number of LSTM neurons, $l$, was changed depending on the number of classes. For training the network with $50$ classes, $350$ neurons were sufficient, and for $MAXVAL$ classes, $MAXVAL$ was the final chosen value. Also in the hidden LSTM layer were two $dropout$ values, one for the standard inputs and one for the recurrent state. These dictate the percentage of values that are lost in the passing and transformations of the data. Both of these values were set to $0.1$.

The output layer has a $softmax$ activation applied. The model was compiled using a $sparse\_categorical\_crossentropy$ loss method, which is a standard choice for classification models where the data values are integers rather than one-hot encodings. The $adam$

```
lstm_out = 350
dropout = 0.1
dropout_r = 0.1
number_outputs = NUM_CLASSES_USED

model = Sequential()
model.add(LSTM(lstm_out, input_shape=(seq_length, n_mels), dropout =
    dropout, recurrent_dropout = dropout_r))  # LSTM input layer
model.add(Dense(number_outputs,activation='softmax'))  # Output layer
model.compile(loss = 'sparse_categorical_crossentropy', optimizer='adam'
    ,metrics = ['accuracy'])
```

**Figure 4.5:** Creation of LSTM model / architecture.

optimizer was chosen and the chosen metric to optimize was accuracy. The final training parameter is the number of epochs, which is the number of times a training set is passed through the network. To increase accuracy but avoid over-fitting, the epoch accuracies were tracked and graphed over time, and 4 epochs gave the best results. The execution of the model's training can be seen in Listing 4.6. An example of the network accuracy throughout the training process can be seen in Figure 4.7. The mini-batch training itself is shuffled, which means the order of the sequences within a mini-batch are randomized to further alleviate over-fitting.

At the conclusion of the batch processing, the model was saved to a file so it could be loaded for the testing and execution of the real-time system.

The system successfully trained models with up to 200 classes, but even with the mini batch processing, the memory requirements with more classes were unfeasible on the server. I migrated the code over to Dartmouth College's High Performance Computer cluster, named Discovery (see: `http://discovery.dartmouth.edu`). After refactoring the code to work as a python script, I wrote batch processes to execute multiple training trials with higher class numbers. The debugging and running of these trials wouldn't have been possible without the help of the Research Computing team at Dartmouth, for whom I am very grateful. With Discovery's resources, I was trained models with class numbers up to

```
batch_length = 5000  # In # of sequences
audio_hop = 64
batch_window = batch_length * audio_hop  # In # of frames

while start_point + batch_window < composite_signal_length:
    end_point = start_point + batch_window

    # Compute batch of training data
    training_data, training_classes = batch(composite_signal[start_point
    :end_point], composite_matches[start_point:end_point])
    training_data = np.array(training_data)

    # Train the model
    history = model.fit(training_data, training_classes, epochs=4,
    batch_size=batch_size, verbose=0, shuffle=True)

    # Adjust frame position for next iteration
    start_point += batch_window
```

**Figure 4.6:** Execution of LSTM model training.

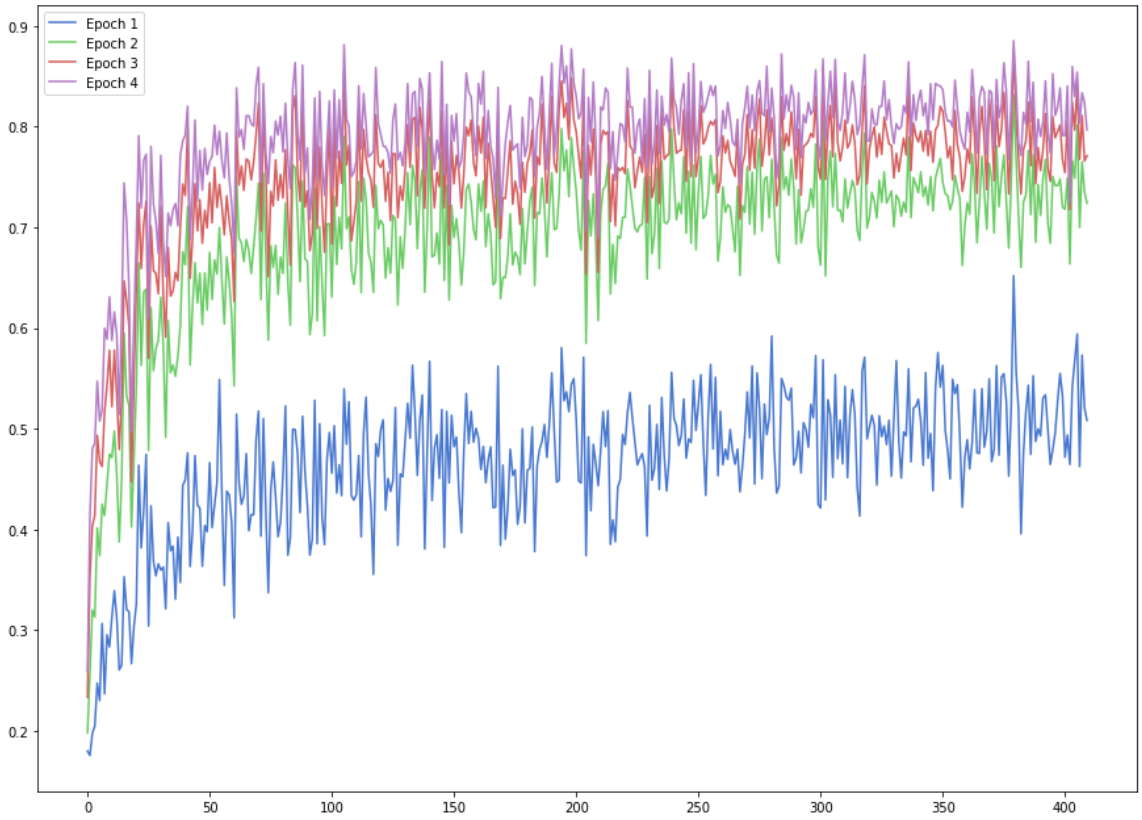



**Figure 4.7:** The accuracy of the LSTM on a 200-class training example. Each epoch is represented with accuracies on the y-axis and each mini-batch execution on the x-axis.

1500. While a higher number would have been possible, it would have taken significantly more time, and 1500 was more than sufficient for the purposes of the installation.

## 4.4 Real-Time Masking and Evaluation

The determination of the specifics of the masking techniques occurred alongside the development of the real-time system. The real-time component of the ACM was developed for the express purpose of use in the art installation, and thus it's architecture were optimized for this specific artistic context. It is quite possible that, if the system were engineered as a strictly technical demonstration, the real-time component would look quite different.

The real-time system is written in Python because of the language's flexibility and the possibility for exact technical correspondence to the methodology used in the training of the neural network. Neural networks are great at seeing data in the same form they've seen it before, but if that form is slightly off they can be wildly inaccurate. If the real-time system had been built in Java, for example, it's possible that the specific mel spectrogram calculations used would slightly differ from the original python Librosa version, which would potentially cause significant errors in the neural network's prediction task.

Because the system needs to operate with as little delay as possible, the most important factor in its architecture and coding is efficiency. The Python Multiprocessing package (`https://docs.python.org/2/library/multiprocessing.html`) was the primary chosen optimization solution. Multiprocessing provides methods for running concurrent operations by putting them in separate processes. It's a common solution to Python optimization problems, as it allows for true concurrency with the advantage of safe shared data structures, as opposed to alternatives like the Threading module.

Building concurrency into the program involved finding which processes can exist as separate entities, and which rely on other processes to complete first. In an unfortunate stroke of luck, it became clear during the organization of the script that three required libraries

for the real-time processing (pyAudio for streaming audio i/o, Librosa for the spectrogram calculation, and Keras for the loading and querying of the LSTM) all break when placed outside of the main process of a program. In other words, the three most essential pieces of the program—and likely the three most time-consuming pieces—must operate in the main process of the program and cannot be paralleled. This made the efficiency of the program a much more challenging problem than would otherwise be the case, and necessitated a rather convoluted series of sub-processes to try and unload as much of the computing power off of the main process as possible.

To determine the necessary acoustic masking, I first conducted a series of qualitative tests. After recording a clean example of sample speech, I then recorded and spliced a corresponding audio track of potential matched speech, attempting to keep output in line with the reality of the ACM. I then tested different ways of combining the two audio files to maximize the perception of the second over the first, keeping the volume of the original track constant, as that will not be changed by the ACM. With the matched audio played at the same volume and simple filtering applied to the original, the matched audio sounds much more prominent. Different filters were tested, and rough EQs, phase shifting, and delays all seemed to effectively obscure the original speech. While exploring different methods of composing the real-time acoustic mask is a desired step of the ACM, time constraints limited the implementation to only the simplest possibility: that of playing the matched sound directly over the source speech.

It is at this point where a problem with an evaluation of the system came about. As defined by its application in the future installation, success of the ACM is defined by its ability to "fool" automated speech recognition (ASR) systems. The purpose of the ACM's specific contextual approach to acoustic masking was designed to also deceive sound classification systems. Because of the proprietary nature of state-of-the-art corporate and governmental software that performs these tasks, sufficient evaluation would require the creation of similar systems from scratch. This is not a trivial task, and due to time constraints,

the scope of the system's evaluation needed to be limited.

Unfortunately, a similar problem surfaced in the ASR evaluation method. Commercial ASR software is incredibly advanced, but licenses to use frameworks like Google Cloud Speech-to-Text or IBM's Watson Speech-to-Text can be quite expensive if not used in a significantly limited setting. There are a few open-source tools for ASR that are well-documented, and I ended up building a custom real-time ASR program with *CMUSphinx* [89]. This database, put out by Carnegie Melon, is the most advanced open-source ASR system available. A framework for working with the database, called *PocketSphinx*, can be easily programmed in Python. Despite its advantages over comparable options, however, *CMUSphinx* and *PocketSphinx* aren't reliably accurate on arbitrary speakers and generic speech recognition. It performs quite well when focused on keyword recognition, or when it can be retrained on text from specific speakers. Though the ASR software is not reliably accurate on unconstrained, completely unedited speech, a difference in the accuracy between unedited speech and masked speech recognition still gives some rough estimate of the ACM's effectiveness.

There were two tests conducted in a sound-isolated environment with two laptops, two microphones, and a single speaker. The first test was run without the ACM activated, and the first laptop simply played a set of $250$ recordings of single words with $2$ seconds of silence in between over the speaker, recording the exact words played as well as their timings. The next test followed the same format but with the ACM activated, playing the original source audio and the real-time mask simultaneously. The other laptop was running a real-time ASR system in *PocketSphinx*, and the output of this system was also written to a file with the transcription timings. The two files were matched up using the timing information and compared. The tests were evaluated by two measures. First was the percentage of exact matches that were obtained, where the ASR output was the exact same as the original input. The second measure was the percentage of approximate matches, where the ASR transcription contained the original input, but didn't need to be *only* the input.

On the set of $250$ words, the unmasked audio was partially matched with an accuracy of $37.6\%$ and the masked audio was partially matched with an accuracy of $12.5\%$. With exact matching only, the unmasked audio had $35.9\%$ accuracy and the masked audio had only $0.9\%$ accuracy. This means that on approximate matching tasks, the ACM was shown to cause a $\sim 3$x decrease in accuracy of the ASR, and on exact matching tasks, the ACM caused a $\sim 40$x decrease in accuracy.

The difference in the approximate match rate and the exact match rate for the masked audio is corroborated by looking at the resulting transcriptions. The majority of the examples of this disparity contained two or three transcribed words, one of which would be the original unmasked word, the others of which would be amalgamations of the original word and the masking audio. For the masked audio, the $12.5\%$ accuracy of the approximate matches certainly isn't as impressive a number as the $0.9\%$ for the exact matches, but it should be emphasized that approximate matches can be incredibly difficult to decipher. If only one-third of a transcribed piece of audio is correct, it should not be said that the transcript is particularly accurate or easily discernible. For practical use against an arbitrary ASR system, the ACM would likely decrease the accuracy of the transcribed audio by an amount somewhere between the $3$x and $40$x results.

These results are incredibly encouraging and demonstrate the potential of the ACM in hiding sounds from methods of mass audio surveillance. In the future, steps can be taken to increase the effectiveness of the system in a number of ways. First, the number of trained audio classes could be increased by a significant factor; ideally there would be on the order of $10,000$ audio classes for the LSTM. A significant roadblock with the masking process is the inefficiency of the used Python libraries, and further investigation into different approaches to the live processing would be necessary moving forward. Once those are addressed, the masking process should contain some form of spectral matching between the matched audio and the live input to better disguise the original sound. As a primary method of evaluation, a customized audio classifier network needs to be made to test whether the perceived resulting

sound does indeed get categorized as "speech".

There are many tweaks that could improve the efficacy of the Acoustic Counterfeit Machine, but its results thus far show the strength in the chosen technologies and methods. As a vehicle to revalue speech, the ACM marries theories of machine listening to the practical reality of audio signal processing, taking advantage of the technologies of mass audio surveillance to pointedly counteract them. With the ACM as an established, working system, it can now be shown for public feedback, paving the way for its function as apart of an art installation.

# 5

# Towards an Installation

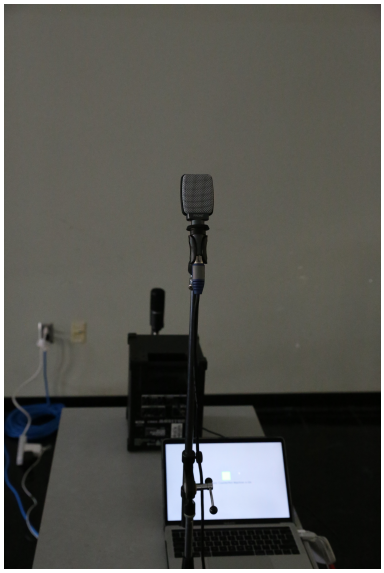
Presented below is first a documentation of a technical demonstration of the Acoustic Counterfeit Machine that took place on April 20th, 2019. The technical aspects of the demonstration and the public feedback to it are explored. Following this is a description and justification of a future art installation design using the ACM, built on the theoretical and political foundations established in Chapter 3.

## 5.1 Technical Demonstration

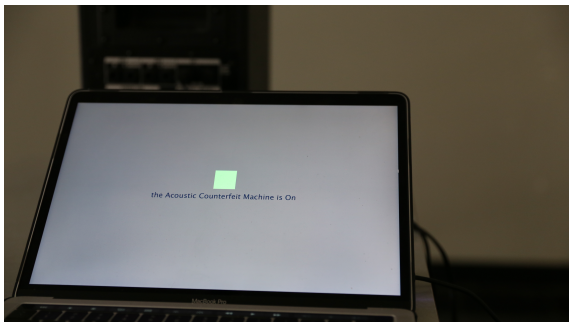
The Acoustic Counterfeit Machine was demonstrated on April 20th, 2019, in the Garage at the Hopkins Center for the Arts at Dartmouth College. The demonstration was designed as a way to showcase the effectiveness of the ACM in a realistic surveillance context, as well as investigate its operation as an artistic process. The primary purpose of the demonstration was to get feedback on the clarity of the ACM in an automated speech recognition context: was it obvious what the ASR was doing, what the ACM was doing, and how the ACM affected the ASR?

The setup, seen in Figure 5.1, involved one laptop running a customized ASR program. Built on *PocketSphinx*, it was connected to *Processing* via OSC, an internal messaging protocol. *Processing* is a Java-based program used for programming visual arts [6]. The connected system was quite simple: when a phrase was recognized by *PocketSphinx*, it was sent to the *Processing* program, where the phrase appeared on a projected display, seen in Figure 5.1c. The most recent messages are displayed in a vertical list, which scrolls upwards when more messages are received or there is a significant period of time in which no messages are received. The computer is connected to a microphone to try and increase the ASR's accuracy. In front of this computer, directly in front of the participant, is a second computer running the ACM, seen in Figure 5.1b. This computer is connected to a microphone and a speaker, which is pointed towards the first microphone for the ASR program. The ACM is connected to a front-facing GUI with a simple ON/OFF button. When the ACM is off, the audio from the microphone goes straight through the speaker to the ASR and the projection should display the (mostly) correct speech. When the ACM is on, the dry audio still goes from the microphone to the speaker, but the masking audio is played simultaneously over the speaker, and the project should display text quite different from the intended speech.

The technical aspects of the demonstration worked quite well, with the ACM producing very unusual transcriptions on the display. The simple interaction of turning the ACM on and



**(a)** The physical layout of the system.




**(b)** The interface for tunring the ACM on/off.

**(c)** Projection displaying the real-time ASR.

**Figure 5.1:** Three images from the ACM technical demonstration on April 20th, 2019.

off was crucial in people's ability to understand the purpose and use of the ACM. In order to show its effectiveness, it seems essential to show it not working as well; it is not necessarily self-evident. That being said, it seemed to be clear to everyone participating what the ASR was meant to accomplish and, in a general sense, how the ACM is supposed to interact with it. The participants found significant pleasure in the poetics of the altered transcription as it appeared on the wall. They shared the feelings that the spacing and aesthetic choices for the projected text contributed to its evocation of poetry, and I think this placement of the ACM in reference to an artistic context helped it feel enjoyable for those unaware of its technical underpinnings.

## 5.2 Future Realization

In creating a work centered around a piece of technology and a politics around it, there are a few fundamental options for directions one can take. First, one can make a work that is a demonstration of the technology in question. A second option is to make a work that is a presentation of the technology in relation to its politics. Finally, one can create a work that attempts to ask questions about the technology and its politics rather than just display them.

Set up in this order, it seems like the best, most satisfying artistic choice is clearly always the third option. However, when approaching a topic that isn't already in the public consciousness, it can be challenging to jump right into indirectly questioning it in a nuanced manner. Mass audio surveillance is not a subject most people think about regularly, and so trying to question that reality without having that reality first established could create confusion and be a generally unsatisfactory experience. Much of the surveillance artwork exists within that first or second category, which is understandable for addressing a subject that's new to most viewers. For example, Mont-réel, the 2015 work by Eva Clouard which displays the artist's location on a gallery screen in real-time [31], does not do much more than present a function of surveillance technology. But most viewers would be previously unaware of the possibilities inherent in GPS tracking, and thus the exhibition might have been well-served by its clarity and directness.

A good example of surveillance art that envelops all of these categories at once can be found in the work of Trevor Paglen. An American geographer and artist, Paglen primarily uses photography and multimedia installation pieces to thoughtfully observe and critique surveillance practices. His work that perhaps best demonstrates this with a focus on technology is Autonomy Cube, created in 2014 [77], and seen in Figure 5.2. A beautiful visual piece, Autonomy Cube is a series of bare computer elements placed in a glass cube. The computers establish a WiFi network that can be connected to any device in the vicinity. Autonomy Cube routes all connected traffic through Tor, a decentralized network of computers around the world meant to hide the source of web traffic [57]. The cube itself also

**Figure 5.2:** A picture of Trevor Paglen's *Autonomy Cube*, installed in Madrid in 2014. Source: http://www.paglen.com/?l=work&s=cube&i=2

serves as a Tor node, which means it's one of the computers through which other Tor users' traffic might be routed. As an installation, Autonomy Cube serves the dual purpose of an art object and a practical technology that actively helps others simply by being installed. As a physical object, it is a curious relic that encourages further exploration and understanding on the part of the viewer. Interacting with the object automatically sets the participant on a path towards learning about Tor, privacy, and the place of those things in their personal lives. Autonomy Cube accomplishes this without needing a conscious desire to investigate privacy on the part of the viewer; the natural interaction one is led to with the piece does the work of introducing and problematizing the topic with ease.

The installation design for this thesis has gone through many iterations, primarily prompted by questions of what purpose it should serve. Initially, it felt necessary to set the installation up like a technological demonstration of the Acoustic Counterfeit Machine. I thought that trying to introduce audio surveillance, the ACM, *and* question the politics of their situation would be too much for someone to experience, and would come across muddled. After further reduction, I came to realize that the different elements of the situation

could be presented in slightly altered function; their use for surveillance could be implied, but their direct presentation could be simplified to make the installation easier to take in as a single piece. The politics of the situation could be presented as a possibility for inference rather than a direct message. Making the installation a clear and cohesive piece that places the nuance in the subtleties of the interactions was fundamental to the creation of the installation's final iteration.

### 5.2.1 Construction and Manifestation

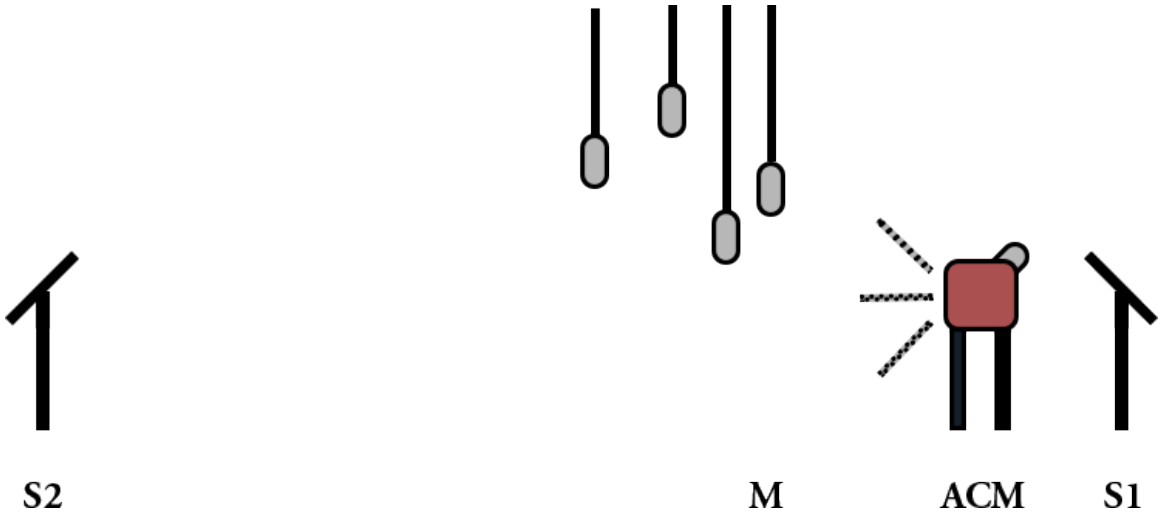



**Figure 5.3:** The proposed installation setup. S1 is the screen for the speaker, ACM is the Acoustic Counterfeit Machine, M is the microphone for the ASR surveillance system, and S2 is the screen for the actor.

The installation is set up as an interaction between two participants mediated by surveillance and anti-surveillance technology. The format of the interaction is like a game of charades. The *speaker*, standing at the S1 station in Figure 5.3, will receive prompts on a screen in front of them. These prompts will be requests for actions, thoughts, feelings, or things. Some examples of the on-screen prompts are:

- Acting like _____
- Being _____

- Feeling like _____
- Seeing _____
- Hearing _____

The prompts are meant to be broad and allow for creative and fun answers. The speaker receives a prompt and states an answer, then presses a button on screen when they've finished. Their speech is heard by one of the microphones hanging from the ceiling—M in Figure 5.3—which is attached to a computer running a real-time automated speech recognition (ASR) system. The calculated text is then sent to the screen placed in front of the *actor*, standing at S2 in Figure 5.3. Their screen will have the prompt followed by the text interpreted by the ASR system, and will read as a set of instructions for them to carry out. The actor will be wearing noise-cancelling headphones, so their only access to the speaker's words is through the on-screen ASR.

On the speaker's screen is also a button to turn on the Acoustic Counterfeit Machine. The machine is placed in between the speaker and the ASR microphone—ACM in Figure 5.3—such that the output of the ACM is played into the ASR microphone. When the ACM is active, the sound interpreted and displayed on the actor's screen will be the acoustically modulated speech. When the ACM is off, the actor's screen should display the correct, unaltered speech as interpreted through the ASR.

### 5.2.2 Actors and Relationships

In any artistic framework there are physical and abstract relationships, and they can be explicitly stated or subtly implied. For example, for a renaissance-era portrait on the wall of a museum, there is the most obvious relationship of the viewer to the painting. There is also the relationship of the painting's subject to their historical context, the frame to the painting, the painting to the rest of the room, the historical context to the contemporary one, and countless others. While it's not always essential to meticulously coordinate all possible

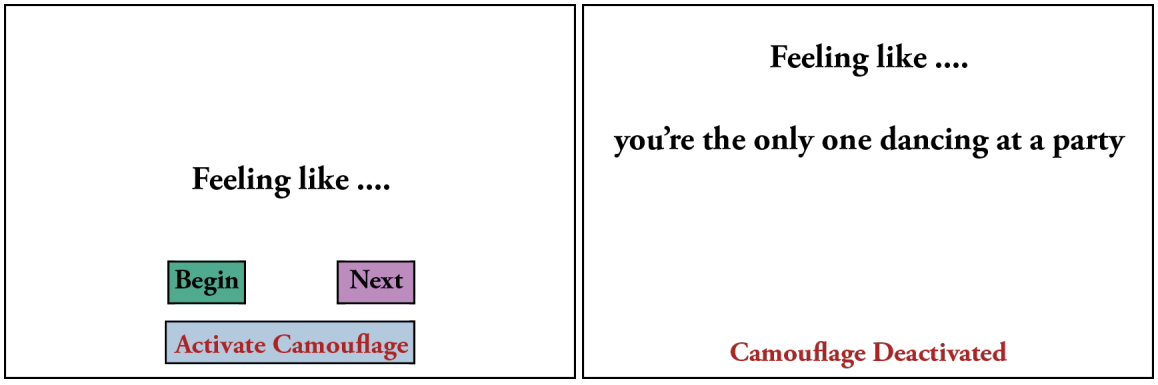
relationships a work may have, considering relationships can provide a useful framework for a work's design, and a lack of consideration can result in a work who's message is contradicted by its methods.

There are four separate conceptual systems in my installation: the human speaker, the human actor, the Acoustic Counterfeit Machine, and the surveillance apparatus. Each of these systems have specific input and output, and the connections between their respective experiences of each others' output give insight into how they might visually and sonically operate. The four systems all operate in reciprocal relationships, and thus there are 12 primary subject-object relationships in the installation. That being said, I will focus on the relationships involving the human participants, as those are the most clearly relevant to the construction and aesthetics of the installation.

The first subject-object relationships are those between the participants and the surveillance apparatus. The microphone of the surveillance apparatus should be perceived by the speaker to be familiar in function; they should recognize it as a symbol of surveillance, seeing the similarity between it and an Amazon Alexa or Google Home, for example. Initially, I thought of the microphone as needing to feel invasive, intimidating, and condescending to the speaker. While an element of this should remain, in the context of the larger game-like interaction, the microphone needs to be clear in function first and foremost. The give-and-take between these two ideas lead to the bare microphone being hung from the ceiling. As the installation is in a sizable room, the vertical height of the microphone gives it a slightly formidable air, while the plainness of its appearance makes it explicitly clear what it is and how it operates. As Figure 5.3 shows, however, there are numerous microphones hanging from the ceiling at different heights and placements. Though there are only four in the figure for visual clarity, the installation should have as many as possible. While only one is connected to the program running the ASR software, their wires all overlap in such a way as to make it impossible to tell which microphones are actually listening. The chaotic, jumbled field of microphones has an oppressive and intimidating nature; the speaker doesn't

know what exactly is listening to them, but their impression of the system is that its large, complex, and *always listening*.

A challenge apparent in discussing the relationships with the surveillance apparatus exists in the trade-off between familiarity and comfortability. The most blatant form of mass audio surveillance most people encounter regularly are "smart speakers". As a primary goal of the installation is to promote recognition of everyday interaction with audio surveillance, it would make sense to either directly use one of these devices, or at least visually model the surveillance apparatus after them. The issue in this strategy is that people don't generally have a negative view of their smart speakers. If someone who'd never confronted audio surveillance was placed in the position of seeing an Amazon Echo performed automated speech recognition with their voice, it is unlikely that their first reaction would be one of concern, fear, or even surprise. After all, ASR is what makes these devices work, and even a layperson would likely understand its integral role in a smart speaker system. While using or evoking a smart speaker would let participants more easily find a point of relation to the installation, conveying that the anti-surveillance tool confuses an Echo would likely only make participants feel that the *tool* is negative, or makes their convenient smart speakers more difficult to use. It is for this reason as well that the surveillance apparatus is in its bare form.




**(a)** S1 display in front of the speaker. **(b)** S2 display in front of the actor.

**Figure 5.4:** Example displays when a prompt has been answered and the ACM is deactivated.

The actor's relationship with the surveillance apparatus is through the on-screen output of the ASR. Given that the interaction is meant to involve mistranslation of a kind, this interaction should also primary be one of clarity. The ASR as a surveillance process should feel clean and easy, both taking little effort and being highly accurate. With this in mind, the display is neat, with simple black text on a white background. The prompt appears at the top of the screen, and the ASR text appears below it, updated live as the speaker and/or the ACM construct their response. An example of what this looks like can be seen in Figure 5.4b.

The next set of relationships are between the participants and the Acoustic Counterfeit Machine. To the speaker, the ACM should appear supportive and exciting. They should recognize the ACM's effect, even when it's not entirely clear what its exact output is. More than just recognition, however, they should *want* to use it and enjoy its results. From the actor's perspective, a similar set of interpretations should be implied. The ACM is a software system that requires the use of a microphone and a speaker, and for a participant to feel that it is one singular object, it is necessary to have its parts appear as one physical form. The ACM will be housed in a wooden box with the microphone protruding on one end and the speaker on the other. The box should be constructed and painted with clean lines and simple aesthetics. The indication of the ACM's operation comes through the displays in front of the two participants. The speaker's screen has an on/off button for the ACM, as seen in Figure 5.4a, and both screens will change color when the ACM is activated, as well as show a small message indicating that fact.

The final set of relationships to explore are those between the two participants. As a game, the installation should encourage the two participants to have fun with the interactions. The speaker should enjoy giving creative instructions when the ACM is off, and when it's on, they should enjoy the disconnect between their words and the actor's response, likely trying to guess the mistranslated message. The actor should enjoy receiving direct, coherent commands as much as they are creative in nature. When the ACM is on, they should have

fun trying to interpret commands that are likely somewhat confusing and grammatically incorrect. The addition of the noise cancelling headphones for the actor to wear encourages the guessing aspect of the interaction, since neither party is aware of the other's version of the message.

Outside of the specific physical systems at play, there is an overarching dichotomy that operates at the conceptual level between the act of exploitation and the act of deception. The systems in place are only present to serve as a structure within which these acts can occur. In initial designs, the primary focus was that surveillance is exploitation and a reduction of agency, and that such a system is not necessary to accept, but can be rebelled and fought against. While this should hopefully still be a conclusion of one's encounter with this installation, central to this final version are concepts of labor and value.

As discussed previously, audio surveillance is a process whereby people's speech is transformed into value. This simultaneous creation of value and immediate rejection of labor at the point of the value's creation results in a model of extreme exploitation. It's not realistic for me to create a system that removes or destroys the surveillance value of speech. However, by modulating it and re-contextualizing it, the value itself is shifted and altered. As a synthesis of the theories of machine listening in Chapter 3, the Acoustic Counterfeit Machine does exactly this. In the installation, when the ACM is off and the surveillance apparatus is working properly, there is a value in the direct translation of the speaker's words. The speaker can say something funny, the actor can act it out, and there is appreciation in the value of that exact correspondence. When the ACM is on, the words are jumbled, and the outcome is unexpected and unique. What is enjoyable about this experience is exactly its *mistranslation*; the process of making the coherent into something else creates an exciting and fun surprise, both in reception and interpretation by the participants. Importantly, this value is not a value that is useful to a machine, a corporation, or the surveillance state. The value of mistranslation is a uniquely human value, and in its presentation, the exploitation of speech as labor not only ceases to be productive, but is turned into an act of joyful humanity.

# Bibliography

[1] Chbl jammer coat. Online at http://www.coop-himmelblau.at/architecture/projects/chbl-jammer-coat/. Citation on page 17.

[2] Ghostery makes the web cleaner, faster and safer! Online at https://www.ghostery.com/. Citation on page 17.

[3] Keras: The python deep learning library. Online at https://keras.io/. Citation on page 54.

[4] Mccarthy era. Online at http://www.trackedinamerica.org/timeline/mccarthy_era/intro/. Citation on page 4.

[5] Moving walls 22. Online at https://www.opensocietyfoundations.org/moving-walls/22. Citation on page 15.

[6] Processing.org. Online at https://processing.org/. Citation on page 63.

[7] Shtooka: A free audio database. Online at http://shtooka.net/index.php. Citation on page 43.

[8] Signal » home. Online at https://signal.org/. Citation on page 16.

[9] Stingray tracking devices: Who's got them? Online at https://www.aclu.org/issues/privacy-technology/surveillance-technologies/stingray-tracking-devices-whos-got-them. Citation on page 17.

[10] Tensorflow. Online at https://www.tensorflow.org/. Citation on page 54.

[11] Urme prosthetic. Online at http://www.urmesurveillance.com/urme-prosthetic/. Citation on pages 13 and 17.

[12] Breaking: Nsa has massive database of americans' phone calls, May 2006. Online at https://consumerist.com/2006/05/11/breaking-nsa-has-massive-database-of-americans-phone-calls/. Citation on page 4.

[13] Audiosurveillance, Aug 2011. Online at https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol4no3/html/v04i3a04p_0001.htm. Citation on page 13.

[14] M. M. Aid and W. Burr. Secret cold war documents reveal nsa spied on senators, Sep 2013. Online at https://foreignpolicy.com/2013/09/25/secret-cold-war-documents-reveal-nsa-spied-on-senators/. Citation on page 4.

[15] J. Ball. Nsa's prism surveillance program: how it works and what it can do, Jun 2013. Online at https://www.theguardian.com/world/2013/jun/08/nsa-prism-server-collection-facebook-google. Citation on pages 4 and 39.

[16] M. Bastashevski. *It's Nothing Personal*. Open Society Foundations, 2014. Online at https://www.opensocietyfoundations.org/moving-walls/22/it-s-nothing-personal. Citation on page 15.

[17] E. Bayer. *Qaddafi Intelligence Room*. Open Society Foundations, 2014. Online at https://www.opensocietyfoundations.org/moving-walls/22/it-s-nothing-personal. Citation on page 15.

[18] A. M. Bedoya. What the fbi's surveillance of martin luther king tells us about the modern spy era, Jan 2016. Online at https://slate.com/technology/2016/01/what-the-fbis-surveillance-of-martin-luther-king-says-about-modern-spying.html. Citation on page 4.

[19] Y. Bengio, P. Simard, P. Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994. Citation on page 48.

[20] H. Black. Social life, Jun 2015. Online at https://www.textezurkunst.de/98/soziales-leben/. Citation on pages 3 and 7.

[21] M. Blackburn. The visual sound-shapes of spectromorphology: an illustrative guide to composition. *Organised Sound*, 16(1):5–13, 2011. Citation on pages vi and 22.

[22] M. Boden. A guide to recurrent neural networks and backpropagation. *the Dallas project*, 2002. Citation on page 48.

[23] B. P. Bogert. The quefrency alanysis of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Time series analysis*, pages 209–243, 1963. Citation on page 46.

[24] S. Burke. Google admits its new smart speaker was eavesdropping on users. Online at https://money.cnn.com/2017/10/11/technology/google-home-mini-security-flaw/index.html. Citation on page 7.

[25] H. Burrell. Whatsapp makes the news quite a bit. why is it encrypted, and what does that mean?, Jul 2018. Online at https://www.techadvisor.co.uk/feature/internet/

how-secure-is-whatsapp-whatsapp-security-encryption-explained-3637780/. Citation on page 16.

[26] S. Casalis, P. Colé, and D. Sopo. Morphological awareness in developmental dyslexia. *Annals of dyslexia*, 54(1):114–138, 2004. Citation on page 20.

[27] M. Casey and M. Slaney. Fast recognition of remixed music audio. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1425. IEEE, 2007. Citation on page 52.

[28] CellularPrivacy. Cellularprivacy/android-imsi-catcher-detector, Sep 2018. Online at https://github.com/CellularPrivacy/Android-IMSI-Catcher-Detector. Citation on page 17.

[29] M. Chion. *Audio-vision: sound on screen*. Columbia University Press, 1994. Citation on page 21.

[30] E. F. Clarke. The impact of recording on listening. *twentieth-century music*, 4(1):47–70, 2007. Citation on page 9.

[31] E. Clouard. *Mont-rÃl'el*. 2015. Online at http://www.artandsurveillance.com/artists-artworks-exhibits/. Citation on pages 16 and 65.

[32] P. R. Cook. *Music, Cognition, and Computerized Sound An Introduction to Psychoacoustics*. MIT Press, 2015. Citation on page 24.

[33] E. Coutinho, F. Weninger, B. W. Schuller, and K. R. Scherer. The munich lstm-rnn approach to the mediaeval 2014" emotion in music'" task. In *MediaEval*, 2014. Citation on page 48.

[34] J. Cox. Matt mitchell is arming underserved communities with anti-surveillance tools, Feb 2017. Online at https://motherboard.vice.com/en_us/article/ezaane/

matt-mitchell-is-arming-underserved-communities-with-anti-surveillance-tools. Citation on page 17.

[35] M. Crocco, M. Cristani, A. Trucco, and V. Murino. Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, 48(4):52, 2016. Citation on page 39.

[36] M. Day, G. Turner, and N. Drozdiak. Amazon workers are listening to what you tell alexa, Apr 2019. Online at https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio. Citation on page 6.

[37] D. Eck and J. Lapalme. Learning musical structure directly from sequences of music. *University of Montreal, Department of Computer Science, CP*, 6128, 2008. Citation on page 48.

[38] D. Eck and J. Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 747–756. IEEE, 2002. Citation on page 48.

[39] D. Eck and J. Schmidhuber. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 103, 2002. Citation on page 48.

[40] M. Ehrenfreund. Edward snowden says he leaked nsa surveillance program (complete coverage), Jun 2013. Online at https://www.washingtonpost.com/world/national-security/edward-snowden-says-he-leaked-nsa-surveillance-program-complete-coverage/2013/06/10/1a6d525e-d1da-11e2-a73e-826d299ff459_story.html?noredirect=on. Citation on page 4.

[41] H. Elahi. *Thousand Little Brothers*. Open Society Foundations, 2014. Online at https://www.opensocietyfoundations.org/moving-walls/22/it-s-nothing-personal. Citation on page 15.

[42] C. Elbro and E. Arnbak. The role of morpheme recognition and morphological awareness in dyslexia. *Annals of dyslexia*, 46(1):209–240, 1996. Citation on page 20.

[43] S. Emmerson. Acoustic/electroacoustic: the relationship with instruments. *Journal of new music research*, 27(1-2):146–164, 1998. Citation on page 9.

[44] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento. Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE transactions on intelligent transportation systems*, 17(1):279–288, 2016. Citation on page 39.

[45] F. A. Gers and J. Schmidhuber. Recurrent nets that time and count. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 3, pages 189–194. IEEE, 2000. Citation on page 48.

[46] I. E. Gordon. *Theories of visual perception*. Psychology Press, 2004. Citation on page 26.

[47] G. Greenwald and E. MacAskill. Nsa prism program taps in to user data of apple, google and others, Jun 2013. Online at https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. Citation on pages 4 and 39.

[48] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017. Citation on page 48.

[49] E. Handelman. *Computer Music Journal*, 14(3):79–81, 1990. Online at http://www.jstor.org/stable/3679964. Citation on page 19.

[50] A. Harvey. Camouflage from face detection. Online at https://cvdazzle.com/. Citation on pages 13 and 17.

[51] J. Hawkins Jr and S. Stevens. The masking of pure tones and of speech by white noise. *The Journal of the Acoustical Society of America*, 22(1):6–13, 1950. Citation on page 42.

[52] S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. Citation on page 48.

[53] F. Hecker. 1935. Online at https://soundcloud.com/hkw/florian-hecker-1935. Citation on page 10.

[54] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001. Citation on page 48.

[55] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. Citation on page 49.

[56] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 93–96. IEEE, 1983. Citation on page 46.

[57] Inc. Tor. Online at http://torproject.org/. Citation on pages 17 and 65.

[58] G. Joseph and D. Nathan. Prisons across the u.s. are quietly building databases of incarcerated people's voice prints, Jan 2019. Online at https://theintercept.com/2019/01/30/prison-voice-prints-databases-securus/. Citation on page 6.

[59] B. Kane. *Sound unseen: acousmatic sound in theory and practice*. Oxford University Press, 2016. Citation on pages 9, 20, and 24.

[60] A. Kundnani and D. Kumar. Race, surveillance, and empire. *International Socialist Review*, (96), Mar 2015. Online at https://isreview.org/issue/96/race-surveillance-and-empire. Citation on page 17.

[61] J.-H. Lee, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee. Speech feature extraction using independent component analysis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1631–1634. IEEE, 2000. Citation on page 52.

[62] R. LeFebvre. Amazon fixed an exploit that allowed alexa to listen all the time, Apr 2018. Online at https://www.engadget.com/2018/04/25/amazon-fixed-exploit-alexa-listen/. Citation on page 7.

[63] Y. Lemma. Anti-worlds, Dec 2018. Online at https://technosphere-magazine.hkw.de/p/9-Anti-Worlds-7uP3HqsVvcre2A8KBwZBAL. Citation on page 10.

[64] A. Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley-IEEE Press, 2012. Citation on page 46.

[65] J. R. LICHTBLAU and ERIC. Bush lets u.s. spy on callers without courts, Dec 2005. Online at https://www.nytimes.com/2005/12/16/politics/bush-lets-us-spy-on-callers-without-courts.html. Citation on page 4.

[66] Q. Lyu, Z. Wu, and J. Zhu. Polyphonic music modelling with lstm-rtrbm. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 991–994. ACM, 2015. Citation on page 48.

[67] Q. Lyu, Z. Wu, J. Zhu, and H. Meng. Modelling high-dimensional sequences with lstm-rtrbm: Application to polyphonic music generation. In *IJCAI*, pages 4138–4139, 2015. Citation on page 48.

[68] M. Magas, M. Casey, and C. Rhodes. mhashup: Fast visual music discovery via locality sensitive hashing. In *ACM SIGGRAPH 2008 New Tech Demos*, SIGGRAPH '08, pages 26:1–26:1, New York, NY, USA, 2008. ACM. DOI 10.1145/1401615.1401641. Citation on page 45.

[69] S. Maier. Machine listening, Dec 2018. Online at https://technosphere-magazine.hkw.de/p/Machine-Listening-kmgQVZVaQeugBaizQjmZnY. Citation on page 10.

[70] S. Maier. Wavenet: On machine and machinic listening, Dec 2018. Online at https://technosphere-magazine.hkw.de/p/1-WaveNet-On-Machine-and-Machinic-Listening-a2mD8xYCxtsLqoaAnTGUbn. Citation on page 10.

[71] S. Mann. Existential technology: Wearable computing is not the real issue! *Leonardo*, 36(1):19–25, 2003. DOI 10.1162/002409403321152239. Citation on page 14.

[72] H. Mascarenhas. Edward snowden designs spy-proof smartphone case to warn if you're being monitored, Jul 2016. Online at https://www.ibtimes.co.uk/edward-snowden-designs-spy-proof-smartphone-case-warn-if-youre-being-monitored-1571928. Citation on page 17.

[73] A. McCoy. Mass surveillance began with world war i, Aug 2014. Online at https://uwpress.wisc.edu/blog/?p=164. Citation on pages 4 and 5.

[74] D. McCullagh. Wikileaks disclosure shines light on big brother, Dec 2011. Online at https://www.cbsnews.com/news/wikileaks-disclosure-shines-light-on-big-brother/. Citation on page 4.

[75] J. P. Nance. Student surveillance, racial inequalities, and implicit racial bias. *Emory LJ*, 66:765, 2016. Citation on page 17.

[76] A. V. Oppenheim and R. W. Schafer. From frequency to quefrency: A history of the cepstrum. *IEEE signal processing Magazine*, 21(5):95–106, 2004. Citation on page 46.

[77] T. Paglen. Autonomy cube. Online at http://www.paglen.com/?l=work&s=cube. Citation on page 65.

[78] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali. Music in our ears: the biological bases of musical timbre perception. *PLoS computational biology*, 8(11):e1002759, 2012. Citation on page 29.

[79] J. Patrick. A guide to pierre schaeffer, godfather of sampling, Oct 2018. Online at https://www.factmag.com/2016/02/23/pierre-schaeffer-guide/. Citation on page 9.

[80] S. Perez. 39 million americans now own a smart speaker, report claims, Jan 2018. Online at https://techcrunch.com/2018/01/12/39-million-americans-now-own-a-smart-speaker-report-claims/. Citation on page 6.

[81] S. Perez. 39 million americans now own a smart speaker, report claims, Jan 2018. Online at https://techcrunch.com/2018/01/12/39-million-americans-now-own-a-smart-speaker-report-claims/. Citation on page 6.

[82] S. Perez. Smart speakers hit critical mass in 2018, Dec 2018. Online at https://techcrunch.com/2018/12/28/smart-speakers-hit-critical-mass-in-2018/. Citation on page 6.

[83] R. Robbins. The sound of your voice may diagnose disease, Jun 2016. Online at https://www.scientificamerican.com/article/the-sound-of-your-voice-may-diagnose-disease/. Citation on page 32.

[84] M. Ryynanen and A. Klapuri. Query by humming of midi and audio using locality sensitive hashing. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2249–2252. IEEE, 2008. Citation on page 45.

[85] P. Schaeffer. *Traité des objets musicaux: essai interdisciplines*. Éditions du Seuil, 2002. Citation on page 8.

[86] M. Schagen and L. Baauw. Project kovr - a wearable countermovement. Online at https://projectkovr.com/designs.html#asctype2. Citation on page 17.

[87] M. Sedláček and M. Titěra. Interpolations in frequency and time domains used in fft spectrum analysis. *Measurement*, 23(3):185–193, 1998. Citation on page 52.

[88] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. Citation on page 11.

[89] N. Shmyrev. Cmusphinx open source speech recognition. Online at https://cmusphinx.github.io/. Citation on page 59.

[90] M. Singleton. Nearly a quarter of us households own a smart speaker, according to nielsen, Sep 2018. Online at https://www.theverge.com/circuitbreaker/2018/9/30/17914022/smart-speaker-40-percent-us-households-nielsen-amazon-echo-google-home-apple-homepod. Citation on page 6.

[91] M. Slaney and M. Casey. Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *IEEE Signal processing magazine*, 25(2):128–131, 2008. Citation on page 45.

[92] D. Smalley. The listening imagination: Listening in the electroacoustic era. *Contemporary Music Review*, 13(2):77âĂŞ107, 1996. DOI 10.1080/07494469600640071. Citation on page 9.

[93] D. Smalley. Spectromorphology: explaining sound-shapes. *Organised sound*, 2(2):107–126, 1997. Citation on page 22.

[94] D. Smalley. Space-form and the acousmatic image. *Organised Sound*, 12(01):35, 2007. DOI 10.1017/s1355771807001665. Citation on page 9.

[95] R. Smith. Watch out: You're in ai weiwei's surveillance zone, June 2017. Online at https://www.nytimes.com/2017/06/08/arts/design/watch-out-youre-in-ai-weiweis-surveillance-zone.html. Citation on page 16.

[96] P. F. Strawson. *Individuals*. Routledge, 2002. Citation on page 24.

[97] B. Sun and S. Velastin. Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems. *Acta Autom. Sin*, 20(3):393–407, 2003. Citation on page 39.

[98] M. Tackett. Technology squeezes fbi wiretapping, Sep 2018. Online at https://www.chicagotribune.com/news/ct-xpm-1992-05-12-9202120073-story.html. Citation on page 4.

[99] M. Thaut, P. Trimarchi, and L. Parsons. Human brain basis of musical rhythm perception: common and distinct neural substrates for meter, tempo, and pattern. *Brain sciences*, 4(2):428–452, 2014. Citation on page 29.

[100] J. Valentino-devries. Uncovering what your phone knows, Dec 2018. Online at https://www.nytimes.com/2018/12/14/reader-center/phone-data-location-investigation.html. Citation on page 16.

[101] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26. IEEE, 2007. Citation on page 39.

[102] A. Warhol. Outer and inner space. *film excerpt downloaded*, 100325, 1965. Citation on page 14.

[103] M. M. WEINER and TIM. Files on illegal spying show c.i.a. skeletons from cold war, Jun 2007. Online at https://www.nytimes.com/2007/06/27/washington/27cia.html. Citation on page 4.

[104] A. Weiwei, J. Herzog, and P. de Meuron. Hansel & gretel, 2017. Citation on page 16.

[105] R. N. Wright. *Financial cryptography: 7th International Conference, FC 2003, Guadeloupe, French West Indies, January 27-30, 2003: revised papers*. Springer, 2003. Citation on page 17.

[106] D. Yanofsky. If you're using an android phone, google may be tracking every move you make, Jan 2018. Online at https://qz.com/1183559/if-youre-using-an-android-phone-google-may-be-tracking-every-move-you-make/. Citation on page 16.

[107] J. Young. Imagining the source: the interplay of realism and abstraction in electroacoustic music. *Contemporary music review*, 15(1-2):73–93, 1996. Citation on page 9.