Cognitive Audio: Enabling Auditory Interfaces with an Understanding of How We Hear

Ishwarya Ananthabhotla

S.B., Massachusetts Institute of Technology (2015)
M.Eng., Massachusetts Institute of Technology (2016)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences at the Massachusetts Institute of Technology

February 2022

Author . . . .
Program in Media Arts and Sciences, School of Architecture and Planning
January 5, 2022

Certified by . . . .
Joseph A. Paradiso
Alexander W. Dreyfoos (1954) Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by . . . .
Tod Machover
Academic Head, Program in Media Arts and Sciences

Cognitive Audio: Enabling Auditory Interfaces with an Understanding of How We Hear

Ishwarya Ananthabhotla

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
on January 5, 2022, in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Media Arts and Sciences

Abstract

Over the last several decades, neuroscientists, cognitive scientists, and psychologists have made strides in understanding the complex and mysterious processes that define the interaction between our minds and the sounds around us. Some of these processes, particularly at the lowest levels of abstraction relative to a sound wave, are well understood, and are easy to characterize across large sections of the human population; others, however, are the sum of both intuition and observations drawn from small-scale laboratory experiments, and remain as of yet poorly understood. In this thesis, I suggest that there is value in coupling insight into the workings of auditory processing, beginning with abstractions in pre-conscious processing, with new frontiers in interface design and state-of-the-art infrastructure for parsing and identifying sound objects, as a means of unlocking audio technologies that are much more immersive, naturalistic, and synergistic than those present in the existing landscape. From the vantage point of today's computational models and devices that largely represent audio at the level of the digital sample, I gesture towards a world of auditory interfaces that work deeply in concert with uniquely human tendencies, allowing us to altogether re-imagine how we capture, preserve, and experience bodies of sound – towards, for example, augmented reality devices that manipulate sound objects to minimize distractions, lossy "codecs" that operate on semantic rather than time-frequency information, and soundscape design engines operating on large corpora of audio data that optimize for aesthetic or experiential outcomes instead of purely objective ones.

To do this, I aim to introduce and explore a new research direction focused on the marriage of principles governing pre-conscious auditory cognition with traditional HCI approaches to auditory interface design via explicit statistical modeling, termed "Cognitive Audio". Along the way, I consider the major roadblocks that present themselves in approaching this convergence: I ask how we might "probe" and measure a cognitive principle of interest robustly enough to inform system design, in the absence of immediately observable biophysical phenomena that may accompany, for example, visual cognition; I also ask how we might build reliable, meaningful statistical models from the resulting data that drive compelling experiences despite inherent noise, sparsity, and generalizations made at the level of the crowd.

I discuss early insights into these questions through the lens of a series of projects centered on auditory processing at different levels of abstraction. I begin with a discussion of early work focused on cognitive models of lower-level phenomena; these exercises then inform a comprehensive effort to construct general purpose estimators of gestalt concepts in sound understanding. I then demonstrate the affordances of these estimators in the context of application systems that I construct and characterize, incorporating additional explorations on methods for personalization that sit atop these estimators. Finally, I conclude with a dialogue on the intersection between the key contributions in this dissertation and a string of major themes relevant to the audio technology and computation world today.

Thesis Supervisor: Joseph A. Paradiso
Title: Alexander W. Dreyfoos (1954) Professor of Media Arts and Sciences

This doctoral thesis has been examined by the following committee:

Thesis Supervisor . . . .

Joseph A. Paradiso
Alexander W. Dreyfoos (1954) Professor of Media Arts and Sciences
MIT Media Lab

Member, Thesis Committee . . . .

Sebastian Ewert
Research Lab Lead, Spotify, Inc
Lecturer, Queen Mary University of London

Member, Thesis Committee . . . .

Poppy Crum
Chief Scientist, Dolby Laboratories
Adjunct Professor, Stanford University

Acknowledgments

to Joe P, for giving me wings to fly, for sharing with me a love for sound, for leading from the front with humanity and integrity, and for never saying never;

to Sebastian, for your wealth of knowledge, your methodical outlook, and your compassion and kindness, that together summed to one of the most positive mentorship experiences a student could ever ask for;

to Poppy, for massive waves of inspiration, for pointers to roads less traveled, for listening to and believing in my ideas when few from the tech world could be convinced;

to Tod, for unending encouragement, for showing me and reminding me that bold, beautiful ideas can live somewhere between the objective and the subjective, and that they are ideas worth exploring.

to the ResEnv crew – Gershon, Nan, Spencer, Brian, David, Juliana, Artem, Patrick, Elena, Caroline, Don, Ariel, Irmandy, Devora, Fangzheng, Mark, Ali, Clément, Océane, Carlos – for being soldiers of science and creativity, bravely exploring unique, uncharted territory while never failing to be present for one another; for becoming family to the new student who walked through the doors of E14-548 for the first time five years ago, quiet, shy, and unsure of herself;

to David, for being my research partner-in-crime and the relentless force behind so much of this work; for the weeks we tortured our groupmates while testing out the sounds for our dataset, for our conference travel adventures, and for our weekly research ~~check-in~~ therapy sessions, without which I would not have ever made it to the finish line;

to Amna, for taking care of me, for celebrating my South Asian identity, for giving me reminders of home;

to my friends from the lab – to Ishaan, Nikhita, Jaya, Neil, Nikhil, Ash, Rebecca, Charles, OoF friends, and so many others – for the laughs, the late nights, the foodcam meals and third floor coffee breaks, the submission proof-reads and career debates, and the reassurance that research is never more important than the bigger things in life;

to the MAS staff, present and past – Linda, Keira, Monica, Lily, Mahy, Sarra and others – for doing everything possible to support my research quest, and for catching me when I fell;

to mentors and collaborators and students from beyond the walls of the lab – Prof. Anantha from MIT EECS, Vamsi, Owen, and Paul from Reality Labs Research, Sebastian and Tristan of Spotify, Tal

and Sree Harsha, my talented UROPs, Srikanth and the fellowship team from Apple – for the gift of your time and perspectives.

to all of the teachers who've been guideposts along my journey of learning over the last 28 years – teachers who showed me the beauty of mathematics and the elegance of physics, teachers who encouraged me to build and break and ask freely along the way, teachers who taught me to fall in love with language and history, poetry and music, philosophy and spirituality, and who taught me that no scientific endeavor could truly exist without these things, the things that make us human – for moulding me, and for inspiring me, one class at a time.

to my friends from various communities in and around MIT and Boston, who have constantly pushed me to become a better person alongside becoming a better researcher –

to Jaya, Hosea, William, Abby, Tally, Sally, Mandy, Imane, Tariq, Emma, Jen, and so many others who helped bring ATHack to life each spring over the last seven years, for your dedication, your spirit of camaraderie, and your passion for accessibility;

to the staff and students of WMBR, for the stories you helped me collect and let me tell, for the hours I spent wandering through the aisles of your audio library, for letting me blast Indian classical music over public airwaves late at night;

to my lifelong friends from the MIT Ohms, for the perfect harmonies and well-timed drops that could occasionally be found amidst the laughter, the chaos, and the learning;

to my YAs, for reminding me to seek personal growth, for helping me grow my self-esteem, and for keeping my audio mixing, cable wrapping, and mic boom-ing skills sharp;

to Felicia, Vivian, Anisha, and Jaya, for keeping an eye out for me from day one, for always listening, and for never giving up on me.

to Mom and Dad, for doing everything in your power to help me achieve my dreams;

to Bhavani, for laughing with me, for crying with me, for praying with me, and above all else, for singing with me;

to my Maama and Ammamma, my rocks here on earth, for teaching me faith and fortitude in equal measure;

to my Thathas, my two stars watching from skies, for your unconditional love, and for never failing to tell me that you were proud;

to all of my aunts and uncles and cousins, for raising me as a village, for making home a place filled with warmth and happiness, for being my biggest and loudest cheerleaders, and for believing in me until the very end;

thank you, for everything.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Auditory Interfaces and Created Experiences

It is only apt that we begin this narrative with an attempt to frame the term *auditory interfaces*, which is used liberally over the pages of this thesis. In the context of this work, we can think of an interface as a technological embodiment that stands between us – or more specifically, our minds – and the sounds that we consume. This definition is perhaps much broader than what might otherwise come to mind; in this definition, I include technologies that mediate the sounds that naturally exist around us, and technologies that synthesize and modulate entirely new sonic environments, and technologies that inhabit the ambiguous space between the two. In this definition, I include physical hardware with active and passive components that we might treat as appendages to our bodies, such as headphones or virtual reality headsets; algorithms that we encounter on our daily forays into the digital world, such as those for audio enhancement or compression; tools meant for the creation, production, and curation of audio, such as digital audio workstations and compositional applications; sound installations and public art designed with the intent to educate or amuse; and sonic "displays" that allow us to tap into auditory environments that are vastly different from the ones physically surrounding us. Moreover, I suggest that we can think of all interfaces as begetting of an *experience* – from feeling connected to a loved one when listening to his or her voice over a phone call, to enjoying a deep sense of focus and productivity while using noise-cancelling headphones and listening to a favorite soundtrack, to losing ourselves on an immersive virtual tour in an alternate sonic reality – which is a term I use (equally liberally) to a describe a uniquely human internalization of an audio stimulus, after it has been mediated by an interface and distilled to some representation by our cognitive hardware, comprised of our ears, our pre-conscious, and then conscious minds.

These sorts of interfaces and associated experiences are ubiquitous in today's world, as are the forces of research and development that are seeking to push them to new frontiers. These forces consider new sensory modalities and new interaction paradigms, greater input-output fidelity and greater resources for compute, and new form factors that better suit our dynamic environments and lifestyles. What they rarely consider, I argue, is a computational paradigm centered about our collective knowledge – both formal and intuitive – of *how* we, as humans, hear. The *how*, while not exclusive of it, refers to far more than the anatomical structures which serve to transduce physical stimuli in the environment to electrical signals that travel to our neural circuitry[1]; it also refers to the higher order behaviors that allow us to identify whether the barking sound emanating from the distance belongs to a dog that is threatening or innocuous; whether the series of otherwise spectrally indistinguishable sonic events we have just heard correspond to someone ascending or descending the stairs; it refers to our ability to unconsciously respond to the sound of our name being whispered on the opposite end of a loud and crowded room, despite the fact that we are actively attending to an alternate conversation; and it refers to our ability to recognize the voice of a person on the phone whom we've met only once before. These behaviors, and the many more that will be discussed in detail over the pages of this work, together form the nuts, bolts, and gears of our auditory machinery. If we are able to deeply understand and computationally mimic aspects of this machinery – whether individual phenomena or whole subsystems at once – I believe we will have created the potential to unlock a landscape of technologies that are unknown to us as of yet.

For just a moment, let us suspend the disbelief that stems from our experience with the state of auditory interfaces at present, and imagine enhanced interfaces – or entirely new ones – that point, in turn, to enhanced experiences, as a result of having tapped into an understanding of human hearing. Let us imagine, for instance, headphones that do not simply uniformly cancel all environmental sounds, but that instead delete or modify select sounds in our periphery that are most likely to distract us from our work; or soundscape engines for virtual reality applications that estimate regions in time, frequency, and space that are expected to draw our attention, and optimize activity and rendering fidelity accordingly; or search and curation

[1] The knowledge of which has, increasingly, made its way into commonplace audio devices and infrastructure, from headphones to codecs.

engines for large bodies of audio, such as recordings from ecological monitoring efforts or lifelogging databases, that operate along emotional or aesthetic dimensions rather than objective, informational ones; or a new generation of audio "codecs" that are informed by auditory semantics, such as sound sources, musical structures, and perceived emotions, instead of redundancy in time and frequency. At the core of each of these futuristic technologies is a model that maps sound to *sound interpretation* at some level of our cognitive pipeline, used in a predictive capacity to drive an experience that is in synergy with human listening tendencies – some shared, and some subjective. These examples illustrate the power that would be awarded to technologists if such models were constructed; they also clearly illustrate the challenges associated with doing so, including defining and scoping phenomena of interest, scaling models to diverse demographics, and creating models that exhibit flexibility and malleability over time and towards individual preferences.

This thesis is about asking if it is possible, and if so, what it takes, to make small strides towards this futuristic landscape, towards pushing the auditory interfaces we are familiar with into new territory, as a way of catalyzing compelling experiences we hadn't yet imagined; this thesis is about forming concrete representations of key ideas in the body of knowledge that theorizes about the way we hear as humans, so that they may drive these experiences; and this thesis is about a journey of conjecture, experimentation, and evaluation at the little-known crossroads between several of these ideas and the trade of statistical modeling, so that we might arrive at a set of tools that allows us to construct these representations, both in the context of the application spaces chosen for exploration in this work, and in the context of those that might present themselves well beyond the pages of this thesis.

1.2 *How We Hear*

> Imagine that you are a participant in a psychology experiment on the perception of complex sounds. Your task is to listen to a series of sounds and write down a brief description of what you hear.
>
> "A single-engine propeller plane flying past," you write in response to the first sound, pleased with yourself for providing so much detail.
>
> The experimenter, on the other hand, is not pleased. He says, with some irration, "No, no, no. Write down what you *hear*, not what you think it is."

> "But I heard a propeller plane fly past," you object. "I didn't think about it; that's what I heard."
>
> "You may not have thought about it consciously," he retorts, "but you didn't hear an airplane, you heard a quasi-harmonic tone lasting approximately 3 seconds with smooth variations in the fundamental frequency and the overall amplitude. That's what I want you tell me about."
>
> "I don't understand," you persist, though a little hesitantly. "I didn't *hear* whatever it is you said. I heard a propeller plane."
>
> The experimenter signs and explains patiently, "No, you interpreted the sound as a propeller plan by matching the incoming stimulus with representations stored in your memory. I'm not interested in how people interpret sounds; that's a job for cognitive science. I'm interested in how you hear the sound itself. Now try again.."
>
> - Adapted from William Gaver's "How Do We Hear in the World?",
> Ecological Psychology, p. 286, 1993.

Over the last several decades, the psychology, neuroscience, and cognitive science communities have made massive strides in uncovering the mysteries and complexities of the interaction between our minds and the sounds around us. Researchers agree that this interaction is governed by an intricate interplay of conscious and pre-conscious processing, and use the phrase "auditory cognition" frequently as an umbrella term for many sub-processes that have been examined for years in highly constrained, laboratory environments – psychoacoustics, spatial localization, scene analysis, auditory attention, memory, causal reasoning, and language understanding, to name hardly a few. Some of these research disciplines take a *bottom-up* approach, choosing to begin by studying the aspects of our response to sonic stimuli that is a function of sounds themselves; others take a *top-down* approach, studying a sonic stimulus as something that is always interpreted in the larger context of the world knowledge, rationale, and consciousness of the interpreter.

For the moment, let us zoom into a smaller region of this multi-tiered and multi-dimensional research space, comprised of *pre-conscious* processes and abstractions at the lowest levels of human auditory cognition – we choose to begin our study here, as the literature suggests that the abstractions that live within this space have the valuable property of being largely shared across human listeners, and can be meaningfully associated with the sound content itself, allowing us to examine them with a bottom-up approach. In this regime, notions that stem from our understanding of lower-level processes, such as those within the studies of psychoacoustics and spatial localization – notions like "we don't hear high-frequencies very well" or "the shape of our head and ears informs how we use sound information to locate objects" – are well supported by neuroscientific observations. More abstract notions that are derived from our understanding of higher-level processes, such as ecological listening and short-term, pre-conscious memory – notions like "our ability to name and picture a sound influences our ability to remember it", or "hearing the sound of a farm animal amidst a scene of urban sounds is likely to draw our attention!" – are intuitive notions, and are still shared, but have more poorly understood neural underpinnings.

The satirical and fictitious exchange printed at the start of this section, originally scripted in William Gaver's seminal paper, playfully illustrates the dynamics and tension between these two classes of abstractions, and the schools of research surrounding them. Technically, neither the participant nor researcher is at fault. In the dialogue, the researcher describes the sound in terms of its spectral properties, which, psychoacousticians would argue, is a representation formed by a segment of the auditory pathway at the lowest levels of interaction between a physical stimulus and conscious internalization of a sound – and therefore, *is* what the participant is hearing. The participant presents the case, in line with the literature on ecological listening, that the fundamental unit of human hearing should be treated as one that is more complex than a series of spectral cues or even a composition of them, or one that is *gestalt*. This research argues that as a human, even in the absence of conscious thought, one reaches first for the source of a sound, attempts to label it, to embed it quickly in the model of the world in sound that one already possesses; and only under circumstance that one is unable to succeed in this classification process would one begin to implicitly rely on the tools expressed in the language of the researcher[2].

At the crux of this work is the gap that persists when this dialogue, or the larger dichotomy that it represents, meets the field of computational modeling. To most readers, the participant's response is intuitive and human-like, relative to that of the researcher. Therefore, capturing information at the level of abstraction expressed by the participant would appear to be a valuable tool for constructing human-meaningful experiences. However, it is also noticeably more difficult to define, capture, and interpret from a sound stimulus than the description provided by the researcher. This discrepancy finds an analog in the computational modeling domain, as will be discussed

[2] This is, in fact, true, and shown explicitly in an experiment we conduct ourselves, described in Chapter 4.

in detail in Chapter 2 – while the community has made sizeable contributions in terms of general purpose spectral representations, cochlear models, frequency masking models, and other models of low-level phenomena in the pre-conscious realm, there is a scarcity of systematic, scalable work targeting gestalt ideas.

1.3 *Cognitive Audio: A New Research Paradigm*

This dissertation presents the idea of *Cognitive Audio*, a name that I've given both to the series of individual explorations in this work as well as the larger research trajectory that they together shape. At the core of the idea is a rethinking of the traditional paradigms enacted by today's auditory interfaces – paradigms of audio capture, representation, replay, and retrieval, amongst others – by capitalizing on the principles of pre-conscious auditory cognition. As hinted at thus far, I suggest that the key to enabling this is the explicit construction of statistical models of cognitive phenomena, with a focus on gestalt ideas from the ecological listening literature that are now tractable given advances in deep learning that enable robust audio segmentation and classification.

As one might imagine, probing, quantifying, and synthetically approximating any of the subset of the processes discussed in Section 1.2 in the wild is non-trivial; much of the structure that has been outlined in the literature about our auditory pathways stems from empirical studies conducted in closed-door, carefully controlled experimental settings. Below, I discuss the biggest obstacles to building generalized statistical models of auditory phenomena, which serve as the most significant research questions motivating this work.

The problem of data acquisition: How do we query or "probe" a cognitive principle of interest, in order to obtain data labels? As we consider pre-conscious processes at greater and greater levels of abstraction, such as auditory attention or memory, self-report is rarely an option; and where one might rely on bio-physical markers as ground-truth for processes in visual cognition (like using the movement of the eyes as a marker for visual salience), analogs for the auditory domain are not immediately apparent. What do we use as an observable marker when attempting to capture more abstract phenomena in sound perception? How do we choose probes that are applicable to real-world audio or environments as opposed to only in experimental setting, so that we may obtain sufficient data?

The problem of data scale and robustness: As we approach more complex, semantic ideas in sound understanding, all our models are likely to suffer from a similar fate – models constructed from individual-first datasets have a greater potential to drive personalized experiences; however, if collected in realistic, longitudinal scenarios, this data is likely to have gaps and sizeable uncertainty. Datasets gathered en masse via crowd-sourcing platforms or large-scale laboratory assessments, on the other hand, are more likely to produce more robust labels; they are more likely also to only account for trends observable across the entire population, blissfully ignorant of the behaviors of demographic sub-groups or individuals that subtly influence auditory cognition. What is the value embedded in the different balances of this tradeoff, and are there strategies that might point us to a meaningful middle ground?

This dissertation is a critical examination of these questions through the lens of a series of projects that I have developed and contributed to, centered about different phenomena in auditory cognition. In Chapter 2, I provide a brief survey of relevant background material, covering examples of auditory interfaces that begin to dance with but fall short of fully embracing cognitive ideas, detailing the state-of-the-art in the machine learning infrastructure that makes this research possible, and discussing the progress in auditory modeling at present, highlighting a gap where this research exists. In Chapter 3, I begin with a simpler version of the research problem, and examine case studies in cognitive modeling for lower-level auditory phenomena. I translate the tools and techniques that emerge from this work to gestalt ideas in Chapter 4, and discuss the process of building general purpose models of these ideas from the ground up, including the construction of custom datasets and intermediate representations. In Chapter 5 and 6, we apply these models towards specific experiences in sound, and evaluate the unique affordances they provide; we also consider the point where general purpose models fail to meet the nuanced preferences of individuals, and illustrate a personalization paradigm that serves as an example of the bridge between the two. Finally, in Chapter 7, we conclude with a discussion of the intersection between the contributions in this work and a string of themes that represent active areas of research in the audio technology and computation world today.

2

Background

2.1 Cognitively-inspired Interfaces

The Responsive Environments group has a longstanding tradition of creating audio technologies that edge forward the state-of-the-art and consider new user interaction paradigms altogether. In particular, the group has experimented with several examples of novel audio interfaces that begin to foray into cognitive and perceptual territory, drawing inspiration from the sound cognition literature on sensory augmentation, extension, confusion, transpresence, and incongruence. For instance, as a part of a series of work merging artistic practice with the newest frontiers in HCI, Gershon Dublon and Rebecca Kleinberger developed *PhoxEars* — a device consisting of a helmet with two parabolic microphones attached as “ears”, whose positions a user can independently control with joysticks [4]. Based on a custom bone conduction headset, the user-controlled ears overlay highly directional sound sources on top of the user’s natural experience of the soundscape. The evolution of this work led the researchers to develop a more comprehensive system, known as *HearThere* [5]. HearThere users wear a bone-conduction headset that overlays virtual sounds, sourced from a real-world environment instrumented with microphones and other sensors, over their natural hearing; the HearThere headset uses a combination of GPS and head tracking to render these sounds as though they are coming from their real-world locations. Because HearThere combines distributed sound capture in a dynamic environment with auditory AR presentation, the device goes beyond creating a traditional AR "layer" towards creating an experience of extended hearing, the first of its kind; the device infers a user’s attentional state by taking into account stillness, eye movement, and neural signals. PhoxEars and HearThere can be considered sensory prostheses that toy with well understood axioms of auditory attention – a user's natural ability to hear is supplemented and heightened intuitively, but users experience sensory confusion as a result; it is difficult for them to tell which sounds are real and which are virtual. The prosthesis provides the user access to an additional perceptual layer, to which they may choose to direct their conscious attention; when the user turns their visual attention to a particular area of the instrumented site, for instance, sounds from that area (normally too far to hear) organically blend over their usual hearing.

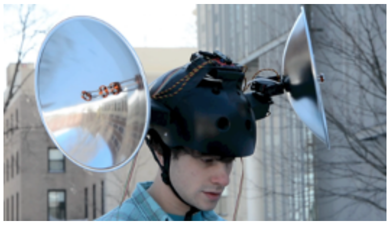


Figure 2.1: The PhoxEars project, an example of an auditory interface that begins to explore cognitive ideas.

In another project from the group, called SoundSignaling, we introduce a platform for notification delivery (such as from email, social media, or SMS) via subtle, stylistic manipulations in a personal corpus of music [6]. The system injects genre-specific modifications — such as adding harmonies to a jazz standard, adding extra layers of rhythm to a blues track, or altering the tempo of a classical piece — at varying levels of conspicuity to a stream of music in real-time. SoundSignaling is an example of design by cognitive heuristics: it operates on the implicit assumption that attentional load modulates awareness of incongruence, an idea borrowed from Stroop's famous colored text experiments [7] and explorations of auditory and visual switching costs [8]; here, the magnitude of "incongruence" is intuited by the designer based on music theory and studies in musical perception[1]. Quantitative and anecdotal data from in-the-wild, long-term studies support the conclusion that SoundSignaling reduces task-switching cost and mediates the intrusion of everyday notifications as a function of cognitive load.

Figure 2.2: The SoundSignaling project incorporates ideas from the cognitive science literature, though only implicitly.

All three of the above examples interface deeply and intuitively with individual perception, working in tandem with a user's attention to support a holistic experience. However, in order to create interfaces that give us more control over or allow us to engage more directly with an individual's experience, we require underlying models that are informed by more than just assumptions and heuristics. This demands that we walk one step further to create *explicit* models of cognitive phenomena – consisting of a structured representation that affords inferential power – rather than the *implicit* connection between cognitive ideas and experimental devices that are demonstrated in these examples. In the following section, I discuss the machine learning tools available to the auditory community that render the construction of these explicit models tractable.

[1] Audio examples can be found at `https://resenv.media.mit.edu/soundsignaling/`

2.2 *Deep Learning Infrastructure*

A precursor to the construction of cognitive models that tackle gestalt ideas is the ability of the model, or the pre-processing pipeline that feeds in to it, to parse, isolate, and identify a *sound object* – much like in the intuitive response provided by the participant in Section 1.2. Tools to undertake this prerequisite step are now under active development in the deep learning community. For instance, several research efforts have focused on deep learning architectures for classification and tagging of environmental sounds, spanning datasets that reflect both urban and nature soundscapes [9, 10, 11, 12, 13]. These efforts have typically been centered about large scale convolutional networks [13], as well as transformer architectures in more recent years [10, 11], trained under fully supervised conditions. Several of the datasets used for training in these works, like Google's AudioSet [2], offer hierarchical ontologies of the associated label set, which serve as valuable, knowledge-graph-like representations of semantic relationships in sound from a human's perspective. Additionally, there has been significant progress in the domains of machine-driven sound-event detection and segmentation [14, 15, 16], localization [17], and environmental audio source separation [18]. In order to consider more dynamic, noise-laden environments and sound stimuli, some recent work has moved towards constructing small-scale, efficient implementations of classification neural networks that can operate on board wearable, always-on devices [19], and other work has demonstrated classification, source separation, and volumetric resynthesis capabilities in highly constrained, complex auditory settings, like at a wetland site instrumented for ecological monitoring purposes [20, 21].

Overall, the advancements in these areas continue to bridge the gap between machine listening and human listening, in two ways that are particularly useful for the construction of gestalt cognitive models: firstly, this infrastructure provides the tools to convert streams of sound into isolated sound objects, or *percepts*, so that we may construct bottom-up models that reflect gestalt phenomena as a composite of the cognitive impact of these individual percepts; secondly, it is working towards repositories of world knowledge in sound – relationships between label categories, sound-producing actions that map to labels, etc – and building tools that can apply that knowledge predictively at a rudimentary level equivalent to the capabilities of a human child. For instance, these tools are able to make a guess as to the source of a sound, and express uncertainty in that estimation that stems from their underlying knowledge structure.

I discuss this research here for two reasons: the first is that I capitalize on existing versions of some of this infrastructure already in this dissertation, to construct cognitive models that capture gestalt ideas; the second is that many of the early ideas presented in this work can be expanded upon and extended to other application spaces if deployed in conjunction with this infrastructure, and the rapid advances we expect to see in it over the next several years.

2.3 *Auditory Models*

Auditory Model	Examples	Built on Cognitive Principles?	Scales to Real-World Audio?
T-F Representations: Cochleagrams, Mel-Spec, Correlograms	[22, 23, 24]	✓	✓
Audio Coding (Spectral Masking)	[25, 26]	✓	✓
Auditory Saliency Maps	[27, 28]	✓	✓
Bottom-up Attention Models	[29]	✓	✗
Auditory Scene Understanding	[30, 31]	✓	✗
Source Separation	[32, 33, 34]	✗	✓
Audio Classification	[35]	✓	✓
Sound Semantic Relations	[36, 37]	✗	✓
Affect Estimation	[38, 39]	✓	✗
Memorability	[40]	✓	✗

Table 2.1: An overview of computational models of audition at different levels of abstraction and complexity.

In this section, I provide a survey of the existing computational models of audition that loosely map to the pre-conscious, environmental sound-focused region of human auditory cognition that we are interested in in this work. As shown in Table 2.1, I choose several state-of-the-art examples from the machine learning and hearing sciences domains, and roughly organize them in order of increasing complexity and abstraction, moving from top to bottom. Neither the chosen model topics nor the research examples associated with each topic are comprehensive, but are selected as representative examples. I also include several model topics that may not have been constructed with the intention of modeling human audition, but can be considered fruitful first steps towards that aim.

For each auditory model type, I evaluate it along two dimensions. I first ask whether the model is *built on cognitive principles*, wherein I evaluate whether or not a model has explicitly been constructed using human anatomical data, user behavior data collected in empirical research experiments (such as in a psychology study), or human-supplied data annotations; in the case of the latter, I am interested in whether the model handles subjectivity and rater diversity, if applicable. I then ask whether the model *scales to real-world data*, evaluating a model for its generalizability to unseen environmental data, as a function of having both analytic and predictive capabilities, and in the case of statistical models, having been trained on a sufficiently diverse and representative dataset, having an explainable intermediate structure for wider inference, or both.

In the case of the models capturing the lowest level cognitive phenomena, such as perceptually motivated time-frequency representations [22, 23, 24], models of frequency masking and psychoacoustics present in audio coding technology [25, 26], and multi-resolution, kernel-based auditory saliency implementations derived from the visual saliency literature [27, 28] , we find that all are built on cognitive principles – the shared characteristics of the principles they represent far outweigh the subtle differences stemming from demographic diversity[2] – and that as largely a series of signal transforms, readily scale to all audio. As we move up the ladder of abstractions, models of sound-driven attention and scene understanding [30, 31, 29] tightly adhere to the mechanics outlined in the experimental hearing sciences literature – for instance, on the grouping of time-frequency units into auditory cues – but current implementations are demonstrative rather than inferential, and are therefore not applicable to unseen audio data. Further up the ladder, we find scalable statistical models that aim to tackle source separation [32, 33, 34] or build an understanding of semantic relationships in sound [36, 37], but note that the training data is not built on human behavior data – it is built, rather, on tasks that may have coarse cognitive analogs, such as learning to estimate individual tracks from audio mixtures, or learning the equivalence of audio signals under statistical perturbations. Finally, we find models for affect and memorability estimation from audio [38, 39, 40]. While more recent modeling efforts in this arena are statistical and result in predictive capability, we classify them as being unable to scale to real-world data for two reasons – (1) the datasets that are used, due to the cost and complexity of human annotation of these phenomena, are small and unrepresentative of a real-world sound ontology, and the resulting models fail to generalize; and (2) almost all of these efforts, both models and corresponding datasets, fail to capture annotator spread. This is a critical idea in this research – while we are interested in capturing shared trends, as we move towards modeling more complex ideas, we edge upon the terrain of subjectivity. Therefore building datasets with point objectives, such as for emotion recognition, and then building models to overfit to those point estimates, quickly becomes meaningless. What we seek instead are models that express *uncertainty* in these regimes, or quantify the

[2] For instance, audio coding in telephony applications creates satisfaction in intelligibility for most people.

shape of the annotation distribution, which can be treated as desired in the context of a downstream application.

In Table 2.1, we are looking for models that are both built on cognitive principles and scale to real-world audio. Overall, we find that most of the models that meet this criteria exist in the lower-level regime, with audio classification [35] being a notable exception[3]. Taking advantage of audio classification models as pre-requisite infrastructure (as discused in Section 2.2), I contribute in this work to the gap highlighted by Table 2.1 by ultimately attempting to construct scalable, generalizable models of gestalt ideas from experimental data, and demonstrating their utility towards the advancement of auditory interfaces.

[3] Though one may argue about the cognitive validity of and inherent diversity in sound source annotations.

3

Explorations in Cognitive Modeling

In this chapter, we take a step back from the task of modeling gestalt phenomena, and begin with modeling exercises involving lower-level perceptual phenomena. From our results in this arena, we begin to form a set of computational strategies for tackling the research questions posed in Section 1.3, that we will ultimately extend to gestalt phenomena in the sections that follow. Specifically, this chapter will cover the following projects and associated insights:

Cognitive Models with Large-scale Data: Loss Functions for Limited Capacity Inference We discuss a model that mimics an MP3 codec as an approximation for simple auditory perception behaviors, such as spectral masking and filtering, and uses the model as an error metric for neural source separation tasks. The work is a demonstration of choosing an approximate mechanism to serve as a data labeling oracle (the codec) and constructing a model under a fully supervised scenario.

Adapting to Sparse Cognitive Labels: Distance Metrics for HRTF Localization We present a model built to compare HRTFs using an error metric that is relevant to human spatial localization, as opposed to computing the error on the basis of the spectral difference. Unlike in the previous work, perceptual data labels are more difficult to acquire, and the model must be constructed in conjunction with non-perceptual data using a domain adaptation and statistical testing approach.

3.1 *Cognitive Models with Large Data: Loss Functions for Limited Capacity Inference*

There is an increasing demand for generative audio applications that are powered by neural networks with lower model complexity and fewer free parameters. This ensures that they can be run, for example, on power- or memory-constrained devices such as mobile phones or wearables. To train such neural networks, there is a significant push in the research community to develop better error metrics for audio representations. Today, much of the state-of-the-art for audio-outputting neural networks relies on sample level euclidean distances [41, 42], which have been shown to correlate poorly with notions in human listening [43]. In light of the mentioned constraints, one option may be to use error functions that are designed to reflect auditory perception, so as to optimize limited neural network capacity only towards perceptually relevant signal components. However, two key requirements must be met in designing such an error function, both of which have historically rendered this regime of research challenging in the deep learning community [44]– (1) the assumptions about perception that are represented by the error function must be sufficiently generalizable and applicable to all listeners; and (2) the error function must be fully differentiable, so as to maintain training compatibility for any downstream task.

In an early collaboration with Dr. Sebastian Ewert of Spotify's MIQ research team, I presented a strategy for optimizing the performance of limited capacity networks by training a secondary network to emulate a low bit-rate codec [44, 26]. In the work, we make the assumption that, given its ubiquity in modern telecommunications infrastructure, an MP3 codec is a robust approximation of the lowest-level aspects of our pre-conscious cognition, representative of principles such as logarithmic hearing, filtering, spectral masking, etc. In training a neural network to approximate a codec, which includes complex non-linearities in its contemporary form, we are able to arrive at a fully differentiable approximation that can be used directly with a downstream training task. To demonstrate the efficacy of the proposed approach, we consider two commonplace audio learning tasks – vocal source separation and speech denoising – and demonstrate that the loss function outperforms a traditional spectral ℓ_2 measure in listening tests for sample outputs from parameter-constrained neural networks.

3.1.1 Related Work

Developing objective functions that incorporate principles of perception is not a well-explored area. Some attempts have been made to approximate metrics used in existing perceptual evaluation toolkits (e.g., STOI [45] and PESQ [46]), such as in [47, 48, 49]. These metrics, however, are either not differentiable functions, thus requiring numerical approximations for back propagation which is highly inefficient, or can be represented as differentiable functions with the consequence of being limited to rather simple models. Most recently, the authors in [50] suggested a perceptual weighting derived from psychoacoustic models applied to a mean-squared-error objective function and highlighted improvements in the performance of small scale neural networks. While this work provides a foundational step in exploring the intersection of psychoacoustic objective functions and limited capacity networks, we note that it does not incorporate subjective listening tests as a part of the evaluation, and employs a per-spectrum calculation of the global masking threshold from the PAM-1 model, which is a non-differentiable approximation.

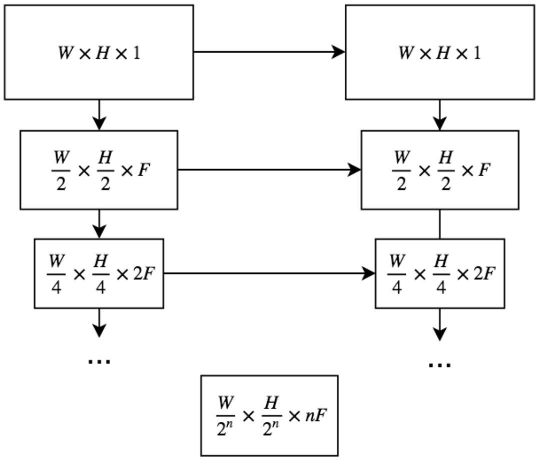

$W \times H \times 1$
$W \times H \times 1$
$\frac{W}{2} \times \frac{H}{2} \times F$
$\frac{W}{2} \times \frac{H}{2} \times F$
$\frac{W}{4} \times \frac{H}{4} \times 2F$
$\frac{W}{4} \times \frac{H}{4} \times 2F$
...
...
$\frac{W}{2^n} \times \frac{H}{2^n} \times nF$

Figure 3.1: An illustration of the u-network architecture used for our separation and loss networks.

3.1.2 Modeling Approach

To describe our approach, let f_Θ denote a function representing a neural network with parameters Θ. Our aim is to train f_Θ to maximize performance for a specific audio separation task, while taking resource constraints for Θ into account. We consider noise removal in speech and vocal separation as applications; since they are source separation tasks, we refer to f as the *separation network* in the following. In this context, f_Θ will operate on short snippets of magnitude spectrograms, with $X_M \in \mathbb{R}^{F \times N}$ denoting the input mixture and $X_S \in \mathbb{R}^{F \times N}$ the desired output for the target source. Given this notation, a baseline speech noise removal or vocal separation method can be trained in a supervised fashion using a standard ℓ_1 loss:

$$\Theta^* = \underset{\Theta}{\text{argmin}} \underset{(X_M, X_S)}{\mathbb{E}} ||f_\Theta(X_M) \odot X_M - X_S||_1, \tag{3.1}$$

where $\odot$ denotes the Hadamard product and $f_\Theta(X_M) \in [0,1]^{F \times N}$ represents a mask to be applied to X_M (resembling Wiener filtering).

For our method, we follow [51] and replace the ℓ_1 term with a new expression that takes human perception into account. To this end, we define a second function g_Φ, which we train to approximate the operation of an audio codec. More precisely, let X denote a snippet of a magnitude spectrogram for an audio signal and let X_C be the corresponding representation for the signal after applying a codec, we train g_Φ to approximate the codec via:

$$\Phi^* = \underset{\Phi}{\text{argmin}} \underset{(X, X_C)}{\mathbb{E}} ||g_\Phi(X) - X_C||_1.$$

This way, we can construct a new supervised loss $\tilde{L}$

$$\tilde{L}(X, Y) := ||g_{\Phi^*}(X) - g_{\Phi^*}(Y)||_1$$

and by replacing the ℓ_1 term in Eq. 3.1 with $\tilde{L}(f_\Theta(X_M) \odot X_M, X_S)$ we obtain a first version of a loss that removes signal components that are perceptually less relevant before computing the actual comparison. We refer to g_Φ as the *loss network* in the following.

While $\tilde{L}$ can work, we observed in practice slow convergence and sometimes even instabilities during training. Therefore, we incorporate ideas found useful in the image domain [52, 53], where trained classifiers were used as losses, which is conceptually related to our approach. More precisely, let $g^m_\Phi(X)$ denote the output of the m-th layer of the multilayer network g_Φ. In this context, $g^m_\Phi(X)$ corresponds to representations or features the network extracts intermittently to fulfill its task, i.e. the input signal is represented at various semantic levels. Thus, we can compare the two inputs not only at the final output layer but also at additional semantic levels. In [52, 53], this was shown to considerably stabilize the use of such a loss and we observed similar behaviour in our setting as well. Our proposed perceptual loss is thus defined as:

$$L_{\mathcal{M}}(X, Y) := \sum_{m \in \mathcal{M}} \lambda_m ||g^m_{\Phi^*}(X) - g^m_{\Phi^*}(Y)||_1, \tag{3.2}$$

where λ_m are weights to adjust the importance of individual layers and $\mathcal{M} \subset \{1, \ldots, M\}$, where M is the number of layers. In practice, we first train for 10 epochs with $\lambda_m = 1$, and then set λ_m

$= \frac{1}{||g^m_{\Phi^*}(X) - g^m_{\Phi^*}(Y)||_1}$ for the remainder of the training to equally weight the contribution of the selected layers, following the suggestion in [52].

The architectures for our loss and separation networks closely follow the U-Net architecture described in [54, 32], as shown in Figure 3.1. Similar to wavelets, the architecture is designed to represent the signal at multiple scales, via a series of down- and up-sampling blocks, which are implemented as convolutional or transposed convolutional layers with stride. As demonstrated in [32], the addition of skip connections between the layers enables the network to focus on higher-level semantics at higher layers, while still being able to access low-level information to reconstruct the signal as needed. This architecture was found useful for various tasks, including source separation [32] and lyrics alignment [55]. One may observe that using a U-Network architecture also for the loss network does not directly emulate the typical encoder-decoder structure that is characteristic of an audio codec, as the presence of skip connections circumvents the introduction of a true information bottleneck. In other words, we do not choose a network that would imitate an audio codec also on the architecture side. In particular, as the MP3 compressed audio data is already limited in information compared to the original audio, there is no need to introduce a separate information bottleneck in the network itself, which would limit the network's capacity to reproduce the audio codec faithfully. Instead, we use specific regularizers to provide a balance between approximation accuracy for the audio codec and smoothness of the function described by the loss network – we found this to be essential to be able to back-propagate through the loss network in a meaningful way. We use different configurations of this architecture in our experiments, which are given in Table 3.1.

3.1.3 *Experiments*

Parameter	**Loss Network**	**Speech Denoising Network Loss Configuration Experiment**	**Speech Denoising Network Model Capacity Experiment**	**Vocal Separation Network Model Capacity Experiment**
Number of Layers	6	2	{1,1,1,2,2}	{2,2,3,4,5}
W	128	128	128	128
H	512	512	512	512
F	28	1	{1,2,4,2,4}	{1,4,2,2,4}
Batch Normalization	All layers	All layers	All layers	All layers
Dropout	50% (first 3 upsampling layers)	50% (first 3 upsampling layers)	50% (first 3 upsampling layers)	50% (first 3 upsampling layers)
Kernel Size (Downsampling)	(5,5), Stride=2	(5,5), Stride=2	(5,5), Stride=2	(5,5), Stride=2
Kernel Size (Upsampling)	(5,5), Stride=2	(5,5), Stride=2	(5,5), Stride=2	(5,5), Stride=2
Activation	*ReLu*, *sigmoid* in final layer	*ReLu*, *sigmoid* in final layer	*ReLu*, *sigmoid* in final layer	*ReLu*, *sigmoid* in final layer
Learning Rate	0.0001	0.001	0.001	0.001
Decay	5e-6	5e-6	5e-6	5e-6
Batch Size	32	16	16	16
Optimizer	Adam	Adam	Adam	Adam

Table 3.1: A list of the model architecture parameters and hyperparameters used in training the loss and separation networks for all experiments.

We conducted a series of experiments to investigate the benefit of our proposed loss strategy for limited capacity networks in the context of the speech denoising and vocal separation tasks. We choose these tasks due to their relevance to on-device applications – examples include speech enhancement for phone calls and song identification based on lyric transcription. For the former task, we used the dataset first presented in [56], selecting the 56 speaker corpus. To increase the difficulty of the task, we select only those examples where the speech and noise are mixed at 0dB SNR, and sub-divide these examples into training, validation, and test sets of approximately 4000, 1200, and 600 samples respectively. For vocal separation, we employ the MUSDB18 dataset [57], which consists of pairs of mixes and corresponding stems for entire songs. We choose the mixture stem as the noisy input X_M, and attempt to predict the vocal stem as the clean output X_S.

Model Type	**Num of Parameters (Speech)**	**Num of Parameters (Vocals)**
P1	54	188
P2	107	1,949
P3	213	2,411
P4	575	9,683
P5	1,949	153,653

Table 3.2: Number of trainable parameters associated with each limited-capacity configuration.

This dataset is sub-divided into train, validation, and test sets consisting of 100, 25, and 25 tracks respectively, with each track being several minutes in length. All of the speech segments/ music tracks are downsampled to 22050Hz, magnitude spectrograms are computed with a window size of 1024 and a hop size of 512 samples, and are broken into non-overlapping snippets of size 128. Some speech segments are simply tiled if they do not meet this minimum input width of the separation network.

We select and fix the loss network parameters as in Table 3.1, and then choose five different sets of parameters determining model capacity for the separation networks performing each of the two tasks, denoted P1, P2, P3, P4, P5. In this context it should be noted that the size of the loss network does not contribute to the model capacity for the separation network associated with each experiment – the loss network is only used during training to improve the performance of the separation network at inference time. A set of parameters P is determined by the values for W and F in the U-Network (see Fig. 3.1), and are given in Table 3.1; the corresponding total number of trainable parameters is shown in Table 3.2. Note that the values for W and F for a given model type P may not be identical between the two tasks; state-of-the-art results in music separation tasks have been achieved with significantly larger networks than those needed for speech denoising tasks. We choose a range of model capacities whose extremes still demonstrate meaningful outputs, and discuss results from a few points sub-sampled in this range.

We begin by training the loss network in a fashion similar to our previous work, and utilize the dataset detailed in [51] consisting of lossless music tracks paired with their 16kbps MP3 coded counterparts; we pre-process the training examples to a sample rate of 22050Hz (as we intuit that perception will be influenced by higher frequency spectral detail in speech and music), a window size of 1024 and a hop size of 512, and use an ℓ_1 loss with early stopping to terminate training. Once this is complete and coding behavior is verified on the test set as in [51], this network is fixed without any further training. We then proceed to train the combined system of the separation network in each configuration P with the loss network, using Eq. 3.2 applied to an optimal subset of layers from the loss network[1]. We additionally train the separation network in each configuration using an ℓ_1 loss to illustrate the respective performance improvement over the loss used in state-of-the-art systems for source separation, such as [32] and [33]. Our training is performed on a single GPU machine, using early stopping to terminate training; each experiment takes approximately 8-10 hours and 4-6 hours for the speech denoising and vocal separation tasks respectively.

We finally generate several examples from the test set for each configuration by applying the phase of the input mixture to the predicted output and inverting the resulting spectrogram. We evaluate the outcomes by conducting an online listening test, recruiting 20 participants in a crowd-sourced experiment for a small fee. We found that performing an actual listening test yielded more reliable results compared to approximative metrics such as PESQ or STOI. Each task in a study consisted of an A/B/X evaluation of a sample track or speech sample comparing our proposed loss metric to the baseline, corresponding to a particular model capacity type P. Each participant evaluated the same five speech samples/ tracks for each of the five configurations in a random order, for a total of 25 comparisons. For the speech task, participants were asked to select the sample that was more intelligible; for the vocal separation task, participants were asked to select the sample where the vocals were more distinct and stronger compared to the background track; in both cases, participants could select "I Don't Know" if they were unable to decide.

[1] The experiment conducted to identify this configuration is detailed in [44].

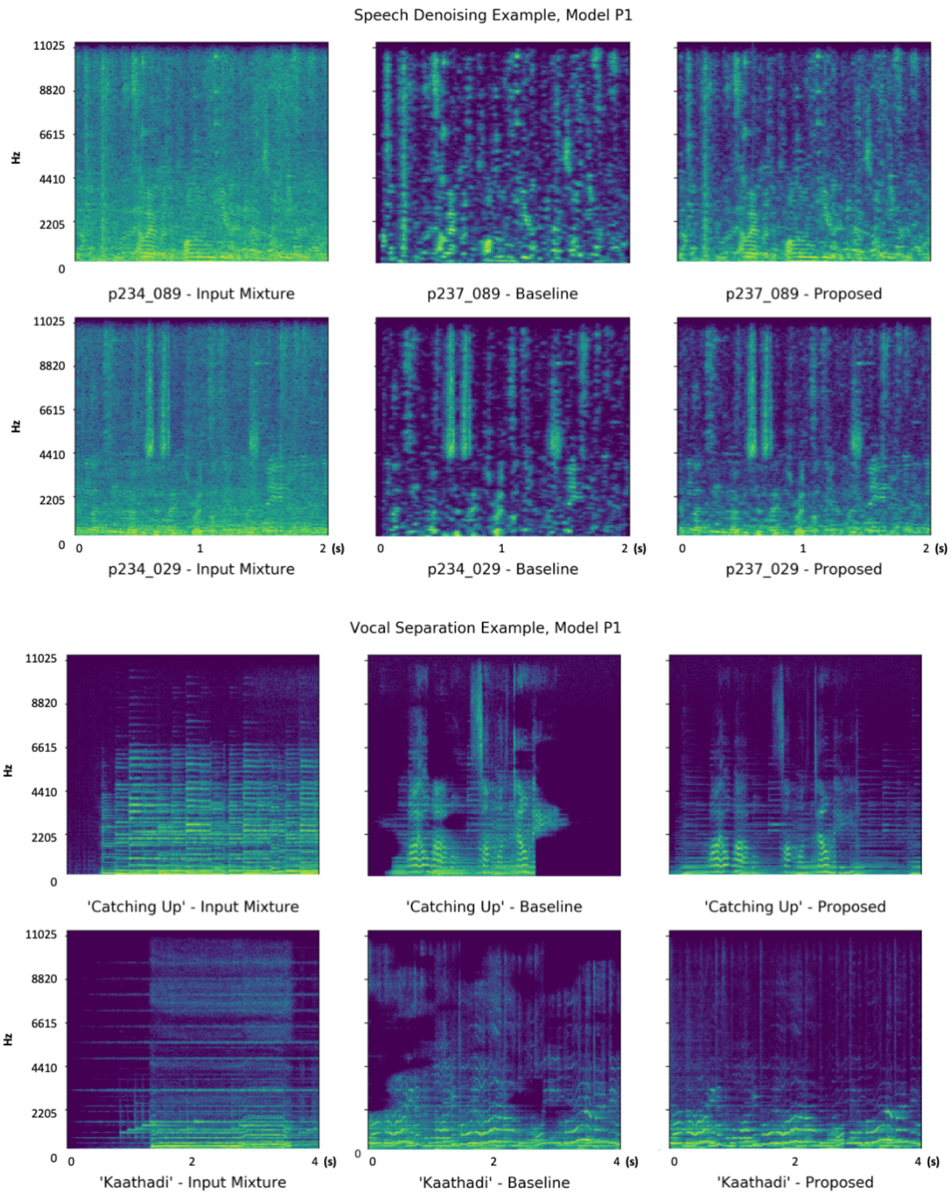


Figure 3.2: Magnitude spectrogram examples comparing the output of P1 models trained with the baseline (center) and custom loss (right), referenced against the input (left), for the speech denoising (top) and vocal separation (bottom) tasks.

3.1.4 *Results*

Audio samples from both the experiments can be found at the repository accompanying the paper[2]. In Figure 3.3, we summarize the quantitative results from our listening experiments by plotting the rate of selection of a sample associated with our proposed loss over the baseline loss strategy, as a function of model capacity for both the speech denoising and vocal separation tasks. We observe that the likelihood that a sample generated using a network trained with our proposed loss is preferred over a sample from the baseline procedure

[2] `https://ishwaryaanant.github.io/small-network-perceptual-loss/`

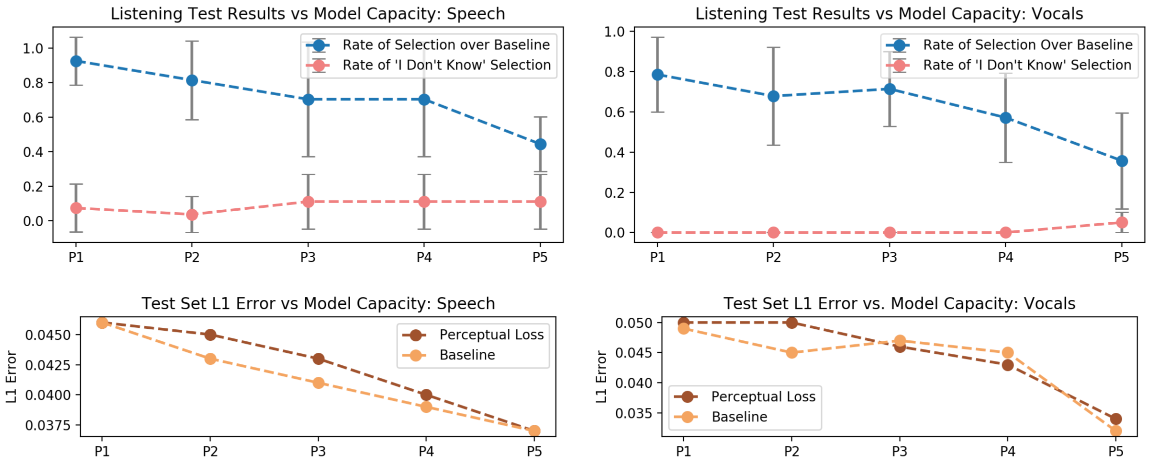


Figure 3.3: Results from the listening tests comparing the proposed loss with the baseline for different model capacity configurations; (Top) For both the speech and vocal separation task, use of the proposed loss leads to better performance for lower capacity models; (Bottom) The ℓ_1 metric resulting from the loss case closely follows or is greater than that which results from the baseline case, suggesting that the loss network optimizes for a different set of constraints.

decreases as model capacity increases, for both tasks; the likelihood of a participant choosing option "X" ("I Don't Know") also increases with model capacity for both tasks. For example, we see that for Model P1, 80% or more of the participants were likely to choose the sample associated with our proposed loss for both tasks. Conversely, this number falls to 50% or less for Model P5. We also note the inter-rater variance by the error bars in both cases. While this variance is roughly constant for the vocal separation tasks, we see a significant drop in the variance for configurations P1 and P5 in the speech denoising tasks, indicating high confidence in rater agreement on preferring the proposed loss sample (P1) or the baseline sample (P5). Taken together, this behavior suggests that perceptual gains are afforded by our proposed objective function particularly in the case of the smallest source separation networks, while performance converges to the baseline with an increase in model capacity.

Additionally in Figure 3.3, we plot the final test set ℓ_1 error for both training procedures as a function of model capacity. We show that the ℓ_1 error for the baseline case tightly follows the perceptual loss case; this suggests that our loss strategy is not simply a form of a regularizer that leads to better ℓ_1 optimization, but an error metric that optimizes for a different set of aims.

In Figure 3.2, we show spectrograms for sound samples from the test set corresponding to both the speech denoising and vocal separation tasks. Visually inspecting the samples provides an interesting observation – that the spectrogram resulting from the perceptual loss strategy appear to be "noisier" than their baseline counterparts, or that the noise appears in different time-frequency regions than in the baseline. This suggests that our proposed loss enables the network to optimize for regions of the spectrogram that more strongly influence our audition and ignore other regions, rather than optimize uniformly across the spectrogram – particularly in the case of limited capacity networks.

3.1.5 *Summary*

The work is an illustration of simple first steps in explicit cognitive modeling; an MP3 codec is treated as a compact and broadly applicable approximation of psychoacoustic principles, and a fully supervised approach is followed to derive a statistical model of them. While there is both noise resulting from the codec's representation of individualized psyachoacoustics and the model's representation of the codec, a clear advantage in this work is the sheer abundance of perceptual labels – for any audio sample, we can readily compute its coded counterpart. But what of the case where this isn't possible, where labeled data is sparse relative to the modeling strategy that we wish to use?

3.2 *Adapting to Sparse Cognitive Labels: Distance Metrics for HRTFs*

A Head-related Transfer Function (HRTF) parameterizes the relationship between the spatial position of a sound source and the signal received by an individual at either ear. A requirement for achieving realism in virtually rendered soundscapes is a robust reproduction of a user's unique HRTF, allowing him/ her to localize sound sources accurately in the virtual environment. However, an HRTF is highly person-specific, and is determined by one's anthropometric features, such as ear shape, head circumference and torso size [58]. The cost and complexity of HRTF measurement systems has resulted in a significant body of recent work surrounding the automated estimation of personalized HRTFs from representations of anthropometric features, such as photographs of the ears or 3D scans of the head and torso [59, 60, 61], and methods to adapt or select from a pool of generic HRTFs that are most appropriate for a given user (as in HRTF selection) [62, 63, 64].

Both these paradigms of personalization require a means to compare two HRTFs and quantify the distance between them. This has traditionally been achieved by error functions that are computed as log-scaled l_p norms directly on the filter representations in the magnitude frequency domain, known as Spectral Difference Errors (SDE).

However, while robust and easy to implement, this class of error metrics is not always well-correlated to perceptual errors. In a collaborative effort with researchers Dr. Vamsi Krishna Ithapu and Dr. Owen Brimijoin at the Facebook Reality Labs (formerly Oculus Research) Audio Team, I sought to develop a more effective error metric for HRTF comparison that is better aligned with localization perception than SDE measures. To achieve this, we attempted to design a statistical model that relates a spectral HRTF representation to a spatial location, according to human perception, allowing us to then use any spherical distance measure in spatial location as an error metric. A naive approach to this problem would entail an approach similar to the previous project on mimicking psychoacoustic principles in a loss function – it would require collecting data in an experimental setting wherein (1) high-fidelity HRTFs are acoustically measured for each participant, (2) participants are presented with a series of sound sources rendered at different spatial locations using these measurements, and (3) are asked to relay their perception of the location of the sources; this data would then be used to create a fully supervised model relating an HRTF to a perceived spatial location. In practice, however, collecting comprehensive amounts of data involving personalized HRTF measurement and listening/ localization experiments is costly and inefficient. In the absence of such large datasets, learning models derived solely from small, sparse, noisy datasets are likely to be unstable and generalize poorly – a well-known challenge in classical data learning contexts [65, 66].

To this end, we proposed and demonstrated a more flexible framework for constructing an HRTF comparison error metric that incorporates localization perception. We suggested a methodology which requires (1) constructing a model that is first built on large amounts of informative, non-perceptual data, which constitutes a "prior" on the relationship between an HRTF spectrum and its corresponding spatial location; (2) fine-tuning this model to reflect sparse, noisy perceptual observations from experiments like the hypothetical setting described above, which we consider the "posterior" model; and (3) computing measures of statistical significance as a function of spatial location between the prior and posterior model to inform further collection of perceptual data. We demonstrated this idea in practice by constructing a neural network model designed to predict a spatial location, parameterized by azimuth and elevation, from a left-right pair of HRTF magnitude frequency responses. The model is first trained on a large database of acoustic HRTF measurements, and further fine-tuned with a small set of observations from a spatial localization experiment using a transfer learning approach.

3.2.1 Datasets

Measured HRIRs: ($\mathcal{H}$) consists of a database of acoustic far-field head-related impulse response (HRIR) measurements captured from 123 subjects. The 201-point, 48KHz-sampled HRIRs were captured along a 612 point spherical grid (denoted by $\mathcal{G}$) of directions with 36 equally-spaced azimuth locations and 17 equally-spaced elevation locations. For the purposes of this work, we used the corresponding magnitude spectra HRTFs, normalized by the maximum spectral energy across all of the individual's measured responses.

Localization Test: ($\mathcal{L}$) is formed by the results of a listening test conducted to evaluate localization perception, using a subset of 30 individuals from $\mathcal{H}$. The participants were presented with a series of virtual sound sources, rendered using their measured HRTF, through a pair of headphones. The test was performed in a quiet room, and the participants were seated at the center of a spherical dome mounted at the center of the room. The participants were asked to identify the sound source location by pointing with a head-mounted laser pointer to the location on the dome where they perceived the sound to be coming from. This location was registered as azimuth and elevation angles relative to the initial front direction. Further details regarding the setup, experiment protocol, and spatial processing necessary for synthesizing the virtual sound sources can be found in [67].

3.2.2 Modeling Approach

We propose a computational framework that predicts the perceived spatial location of the sound from a given pair of far-field left and right HRTFs. To do this, we first constructed a learning model that directly relates this HRTF pair to its measured source location, relying only on $\mathcal{H}$; we then modified this trained model appropriately using the perceptual data in $\mathcal{L}$. This two stage framework allowed us to validate the efficacy of the trained model with and without the perceptual data, implicitly providing insight into the biases induced by the perceptual feedback. We utilized ideas from deep learning (and more precisely, feed forward convolutional networks) to design this model.

We denote the left and right HRTFs corresponding to azimuth θ and elevation ϕ by $h^L_{\theta,\phi}$ and $h^R_{\theta,\phi}$ respectively, defined over frequency. We constructed a learning model M that maps these signals to θ and ϕ. Ideally, this is a regression prediction problem from continuous inputs to continuous outputs; however, keeping in mind the sparse structure of the spatial grid $\mathcal{G}$, and the fact that a discrete output or target space is desirable for neural network training (as highlighted in audio applications like [68, 69]), we transformed the prediction problem into a classification one. The outputs are denoted by $y \in [0,1]^{612}$, where y_i represents the i^{th} direction from the grid $\mathcal{G}$.

At prediction time, given a new pair of HRTFs, the vector entry with the greatest probability in the prediction $\hat{y}$ represents the estimated source location, denoted by $(\hat{\theta}, \hat{\phi})$. These predictions from M can be used to construct an error metric for downstream tasks. Given two HRTFs h_i and h_j from some unknown locations, M estimates $(\hat{\theta}_i, \hat{\phi}_i)$ and $(\hat{\theta}_j, \hat{\phi}_j)$, which can then be used to compute the distance metric $d(|\hat{\theta}_i - \hat{\theta}_j|, |\hat{\phi}_i - \hat{\phi}_j|)$. $d(\cdot)$ can take the form of any appropriately chosen angular distance metric in the spherical domain.

To explicitly account for the influence of perceptual feedback from $\mathcal{L}$, as mentioned earlier in this section, we first trained M with $\mathcal{H}$, followed by a fine-tuning training phase with the data from $\mathcal{L}$, resulting in a model denoted $\tilde{M}$. Details regarding the architecture, design, learning, and optimization strategies for M can be found in the supplement. Training the models followed the standard learning criterion and cross-validation principles employed in the deep learning literature [70, 35].

3.2.3 *Experiments and Results*

Without Perceptual Feedback In this first set of experiments, we examined the performance of M without the inclusion of perceptual data L. The model achieves an absolute test set classification accuracy of 65.8 %, with a mean distance error (absolute difference in predicted angle) of 0.67° (σ = 2.7°) and 4.7° (σ = 12.7°) in azimuth and elevation respectively. As a more meaningful representation of performance, however, we also report the 1-bin Tolerance (1BT) measure, which corresponds to the model's accuracy in predicting the true spatial location of HRTFs within one neighboring grid location (where one grid location has a resolution of 10°). In Figure 3.4, we show both the 1BT and absolute accuracy as a function of space. For simplicity, we report this measure along the elevation and azimuth axes independently, aggregating the opposite axis. We note slightly decreased performance in elevation classification at the extremes, likely due to measurement noise. We also observe lower performance in azimuth prediction overall as compared to elevation prediction likely due to smaller cartesian spacing in azimuth at higher elevations.

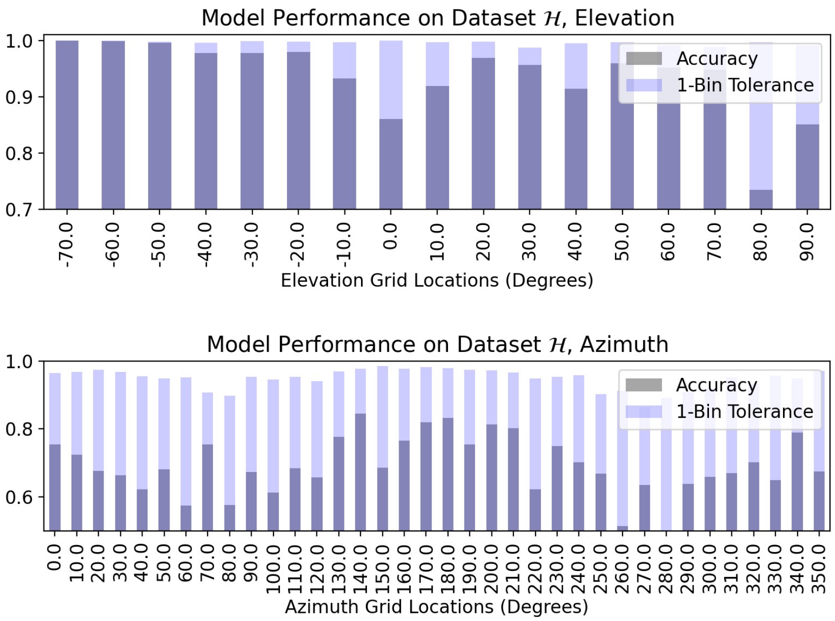


Figure 3.4: Accuracy and 1BT for elevation (top) and azimuth (bottom), aggregated over azimuth and elevation respectively.

We next attempted to understand how the distance metric derived using M behaves in comparison to SDE. We did this using the following procedure: given an azimuth θ_0, we chose two possible grid locations along the elevation axis (ϕ_i and ϕ_j), and selected two HRTFs, $h^{L/R,P_1}_{\theta_0,\phi_i}$ and $h^{L/R,P_2}_{\theta_0,\phi_j}$ belonging to two random individuals P_1 and P_2 from the database $\mathcal{H}$. Using these, we computed a simple distance measure in the output space of M, namely:

$$L_M = |\hat{\phi}_i^{P_1} - \hat{\phi}_j^{P_2}| \tag{3.3}$$

where $\hat{\phi}_i$ and $\hat{\phi}_j$ are predicted by M. We additionally computed an SDE measure from the magnitude frequency HRTFs. For the right HRTF, this is defined as:

$$L_{SDE_R} = \frac{1}{N} \sum_{n}^{N} |20 \cdot \log_{10}(h^{R,P_1}_{\theta_0,\phi_i}[n]) - 20 \cdot \log_{10}(h^{R,P_2}_{\theta_0,\phi_j}[n])| \tag{3.4}$$

which we repeat separately for the left HRTF; N is the number of frequency bins, equal to the tap length of the measured HRIRs (see Section 3.2.1). We averaged these measures (L_M, L_{SDE_R}, L_{SDE_L}) across all pairs of subjects in the test partition of $\mathcal{H}$ (approximately 50 subjects) for a given location, and repeated the procedure for every pos-

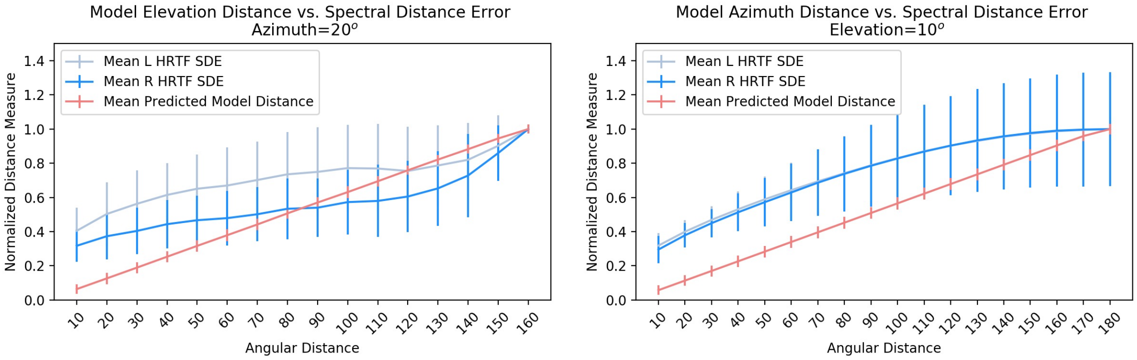
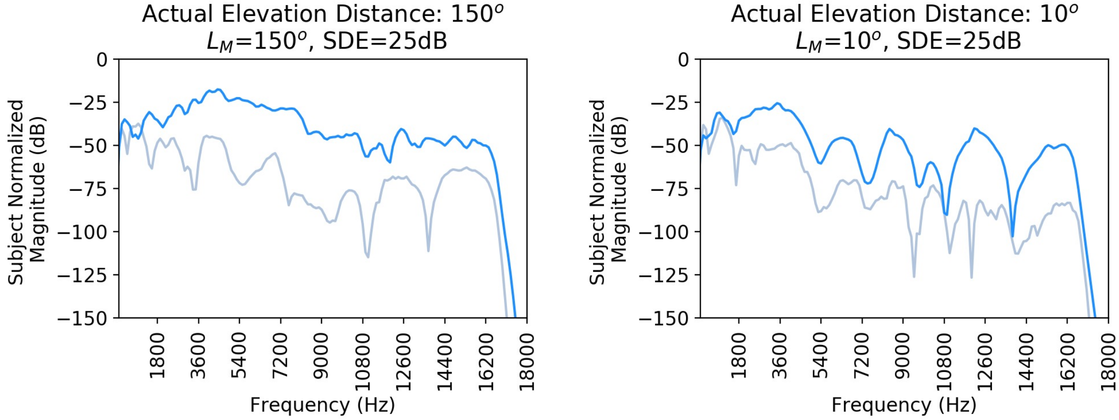


Figure 3.5: We show a comparison between L_M and L_{SDE} for several pairs of HRTFs from randomly sampled subjects in $\mathcal{H}$; the vertical bars give the standard deviation across subject pairs. We show trends in elevation for a selected azimuth location (left), and trends in azimuth for a selected elevation location (right).

sible location along the fixed axis, choosing a few values in azimuth and elevation for the fixed axis. The results of this process are shown in Figure 3.5, with trends in elevation for a fixed azimuth shown in the left image and trends in azimuth for a fixed elevation shown in the right image. We note that L_M is linear and monotonic with increasing distance in elevation and azimuth. On the other hand, while L_{SDE} for the ipsilateral ear is monotonic with increasing distance, and it lacks linearity. We also show that L_M displays significantly less inter-person variability than L_{SDE}. To ensure that these trends are robust and do not result from sampling noise, we treat each curve in Figure 3.5 as a 2D distribution in angular distance and subject pair, and compute a two-sample multivariate t-test on L_M and the mean of L_{SDE_R} and L_{SDE_L}. A T^2 statistic with a p-value < 0.05 suggests that the two distributions are unrelated. Taken together, these demonstrate the utility of our proposed metric – while SDE may reflect variance in distance at a course spatial resolution, our proposed metric is more robust for fine-grained angular distance comparisons, and is more robust to inter-personal spectral differences.

In Figure 3.6, we provide an example to illustrate the affordances of the proposed metric. On the left, we show two magnitude HRTFs from the ipsilateral ear of two subjects which were measured 150° apart in elevation, at a fixed azimuth location of 20°; on the right, we show another pair from the same two subjects and azimuth location representing a difference of only 10° in elevation. L_M predicts a value of 150° and 10° respectively, while L_{SDE} reports 25dB for *both* pairs.

With Perceptual Feedback In a second set of experiments, we explored the role of perceptual data in shaping predictions across spatial locations. To do this, we applied M and $\tilde{M}$ to the test partition of $\mathcal{L}$, and compared the performance of the two models. As an exhaustive approach, we performed an iterative hypothesis test comparing


Actual Elevation Distance: 150^o L_M=150^o, SDE=25dB

Actual Elevation Distance: 10^o L_M=10^o, SDE=25dB

Figure 3.6: A comparison of two pairs of right-HRTFs measured from the same two subjects at the same azimuth location (20°); the first pair (left) were measured 150° apart, and the second pair were 10° apart. However, both pairs result in the same L_{SDE} value.

the two distributions of model predictions for each possible grid location along either the elevation or azimuth axis. This results in a measure of confidence describing whether the two distributions were drawn from the same underlying distribution. We suggest that combining this information with the models' performance as indicated by the 1BT measure provides insight into the value of perceptual data as a function of space, and we provide an illustrative example for discussion.

In Figure 3.7, we plot the 1BT measure for each possible location along the elevation axis, above the p-values (plotted as 1 - p-value) resulting from the iterative hypothesis test and the number of perceptual observations available for each location from $\mathcal{L}$. An analogous plot for this analysis in azimuth can be found in the supplement. In spatial regions where the hypothesis test shows statistical significance, and $\tilde{M}$ has outperformed M, such as where $\phi = 10°$, we draw the conclusion that perceptual observations provide critical information; in regions where the improved performance of model $\tilde{M}$ is not supported by statistical significance, such as where $\phi = 0°$, we conclude that perceptual observations do not afford additional information over that already captured by M. However in spatial regions where the hypothesis test does not show statistical significance and very few perceptual observations have been captured relative to other locations, such as where $\phi = -40°$, there is not enough certainty to draw either conclusion; instead, we suggest that this is a useful spatial heuristic to inform the collection of perceptual observations in future iterations of participant experiments.

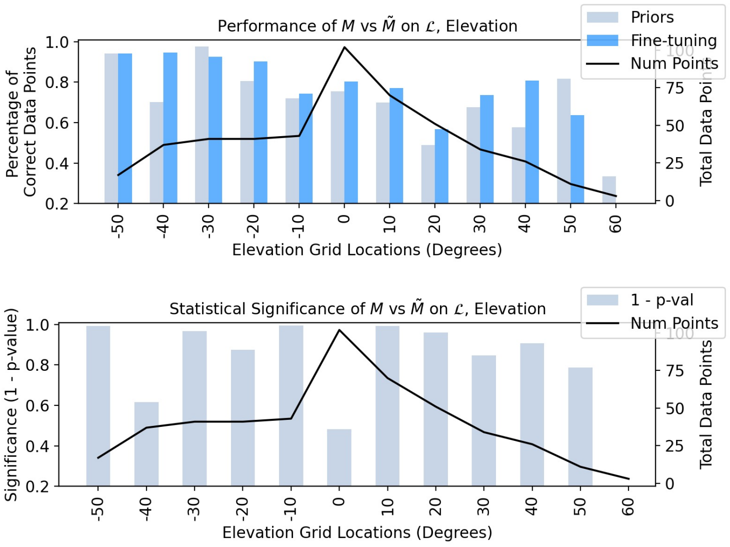


Figure 3.7: We give a comparison of the performance of M and $\tilde{M}$ on the elevation axis via the 1BT measure (top); we juxtapose this with the p-values from the iterative hypothesis test comparing the two distributions (bottom).

3.2.4 *Summary*

As an illustrative example of an exercise in cognitive modeling, two key takeaways can be drawn from the work:

(1) There is value in choosing an appropriately scaled and rich dataset representing a reasonable causal assumption to form the "prior" that is later adapted to small-scale perceptual data. In the case of this project, for example, it is reasonable to assume that HRTF measurement location is very coarsely correlated with perceived location, and sizeable datasets consisting of human HRTF measurements are easily available for research purposes [71]. The results from the project also go on to show that the error metric obtained directly from the "prior" data already outperformed SDE measures in terms of monotonicity and linearity with spatial distance. From a machine learning standpoint, this is not a technically novel concept, and is well explored in the literature on transfer learning, domain adaptation, pre-training, etc. However, forming a suitable prior more broadly requires both insight into the application domain and intuition about alignment with cognition.

(2) Measures of statistical significance, such as hypothesis testing, are required to quantify the affordances of domain adaptation strategies to small-scale perceptual datasets, and to provide feedback to the perceptual label acquisition process. In this work, the performance of the fine-tuned model on unseen perceptual data provides insight into model uncertainty as a function of space; by obtain-

ing measures of confidence for the difference between priors and posteriors at a particular location in space, we draw conclusions regarding the spatial regions where perceptual data does or does not provide critical information over non-perceptual information, and where more perceptual data needs to be collected to be able to draw such a conclusion.

3.3 *Conclusions*

In this chapter, we've discussed two example projects that are suggestive of two important ideas in auditory cognition modeling. The first is that we need to be creative in identifying tractable, computationally-friendly proxies for cognitive phenomena so that we can quantify them robustly at scale – we demonstrate this idea in Section 3.1, where we trained a deep learning model to emulate an MP3 codec. The second is that domain adaptation is a practical approach to building models of cognitive phenomena when the available dataset is small, noisy, and sparse. However, it is important that the domain adaptation approach is complemented by an analysis of uncertainty within smaller datasets or a measure of uncertainty that is propagated through to the final, adapted model, just as hypothesis tests are used to estimate the significance of the perceptual feedback in Section 3.2. We revisit both of these ideas in our discussion of gestalt modeling in Chapter 4.

4

Modeling Gestalt Phenomena

In this chapter, I consider the insights and learnings from the problems tackled in Chapter 3 as I move towards exploring shared notions in auditory cognition that are more complex the ones presented thus far, wherein the challenges of reasonable "probing" and label uncertainty make themselves more apparent, in addition to problems of data scale.

As an early foray into exploring these challenges, and as the primary contribution in this thesis, I attempt to construct statistical models that reflect higher-level, semantic concepts in sound understanding that are derived from the literature on ecological listening and short-term, pre-conscious memory. Following the nomenclature in the auditory psychology literature, I collectively refer to these concepts as gestalt principles. One might argue, of course, that the set of concepts that constitute gestalt principles can be quite large, and that it would be rather bold to assume that all of the information required to estimate these principles is contained entirely within the audio signal. However, attempting to capture pre-conscious ideas (see Section 1.2) means that we wish to begin by working with principles that, as we identify through crowd-sourced annotations, are largely shared and agnostic to the diversity in individuals' processing, world-experiences, sound exposure, and other such factors that influence auditory cognition. It is important to note that these properties (or estimates of them) on their own are not likely to say very much about an *individual's* gestalt response to a sound object that has been presented to them; but they are extremely valuable building blocks for constructing downstream interfaces which afford meaning at the level of the individual, which we will discuss further in Chapter 5 and Chapter 6.

All of the work described in this chapter is a close collaboration with fellow graduate student David Ramsay[1]; to the best of our knowledge, this is the first end-to-end effort to bridge the study of the gestalt and acoustic principles of everyday sound objects with statistical models.

In the sections ahead, I first present an overview of the background literature that highlights the cognitive principles we wish to examine from a modeling standpoint. I then discuss the curation, annotation, and validation of two datasets that we construct to support our modeling efforts. I next describe a probabilistic bootstrapping strategy that we develop to create robust estimators for the gestalt properties in these datasets, which are limited to several hundred isolated sounds, in the face of unseen, real-world audio. Finally, while the approach is well-motivated and demonstrates potential, I discuss the possible shortcomings of the approach given our choice of intermediate structure for bootstrapping, and suggest that we must consider these limitations when connecting the estimators to downstream, user-facing auditory interfaces.

4.1 *Ecological Listening and Memory*

A review of auditory perception and taxonomy research reveals that listeners typically conceive of sounds they encounter in the language of higher level semantics first, and only when a sound's source object becomes ambiguous – or *causally uncertain* – do they tend to resort to acoustic features for distinction. As mentioned in Section 1.2, this idea was introduced in [72, 73] with Gaver's model of ecological listening, suggesting that our consumption of and interaction with everyday sounds is primarily driven by our ability to estimate the source of the sound and the physical interactions that resulted in its production, and secondarily by acoustic and spectral properties. This dichotomy and hierarchy was reinforced by several later studies – [74, 75, 76] suggest that listeners rely on sound source and context/ location identification prior to acoustic features in categorization tasks; [77, 78, 79, 80] suggest that soundscapes with living/ organic elements (humans, animals, etc) are perceived differently and elicit different emotional responses than soundscapes with purely inorganic sounds, and point to the role that source attribution plays in determining a listener's response. In [81, 82], researchers attempt to quantify causal uncertainty (*Hcu*) from a listener's perspective, describe its complex relationship with a sound's typicality, familiarity, and ecological frequency, and demonstrate the role that the measure plays in sound organization and clustering tasks.

[1] Statement of contributions: David Ramsay and I jointly designed the data collection procedure and infrastructure for the HCU400 dataset and accompanying memorability scores. The subsequent work that builds on these results, including the general purpose estimators, audio summarization project (Chapter 5) and personalization experiments (Chapter 6), is my own, though with frequent input and constructive feedback from David. I am grateful for his involvement in and support of this work.

The interplay of higher and lower level processing is further corroborated by neurological observations. Studies of Event Related Potentials (ERPs) demonstrate that the earliest layers of our pre-conscious auditory processing rely on the gestalt semantics of the auditory objects we encounter, as pre-attentive characteristics of these neurological signals are invoked in response to changes in both low-level acoustic changes (like a sudden loud noise) as well as high-level semantic ones (like the sound of a farm animal unexpectedly appearing in a series of urban sounds) [83, 84, 85, 86]. Measurement results in these works indicate that pre-conscious processing of, attention towards, and memory of two events that might sound very similar – a snare drum and a gunshot, for instance – will vary drastically given different semantic interpretations despite very close acoustic signatures.

The literature also suggests that these concepts play a role in auditory memory formation. For a sound to enter our memory, it is first unconsciously processed by a change-sensitive neural mechanism before passing through a conscious filtering process [85, 86, 83]. We then encode this auditory information via a complex and variable procedure; frequently we abstract our experiences into words, though we also utilize phonological-articulatory, visual/visuospatial, semantic, and echoic memory [87, 88]. Different types of memory may also drive more visceral forms of recollection and experience; non-semantic memory, for example, may underpin powerful recollection and nostalgia similar to that reported with music [89]. Research shows a complicated interdependence between attention, acoustic feature salience, source concept salience, emotion, and memory; furthermore, verbal, pictorial, and phonological-articulatory mnemonics can also have a significant impact on sound recall tasks [28, 90, 91, 92, 87, 88]. The research suggests that sounds that are both contextually novel based on their acoustic features, as well as sounds that are only conceptually novel (i.e., differ from the surrounding objects in semantic terms) are likely to trigger neural mechanisms that are responsible for encoding into memory.

4.2 HCU400

To begin to capture these higher-level ideas, and explore the dichotomy between them and traditional, lower-level sound properties as a function of the notion of causal uncertainty, we constructed the HCU400 dataset[2] [1]. At the time that the HCU400 dataset was introduced, it was largest dataset available for studying everyday sound phenomenology. The dataset includes 402 sounds that were chosen

[2] An interactive demo of the dataset can be found at: `https://resenv.media.mit.edu/memory-dataset/demo.html`

to (1) capture common environmental sounds from everyday life, and (2) to fully sample the range of causal uncertainty. While many of the sounds in the dataset are unambiguous, over 100 of the sounds are modified to intentionally obscure their source– allowing explicit control of source-dependent effects.

As part of the dataset, we include high-level emotional features corresponding to each sound's valence and arousal, in line with previous work on affective sound measurement [93]. We also account for features that provide other insights into the mental processing of sound– familiarity and imageability [94, 95]. We also introduce word embeddings as a clustering technique to extend the original H_{cu}, and apply it to the free response labels we gathered for each sound in the dataset.

4.2.1 Dataset Construction

The HCU400 dataset consists of 402 sound samples and 3 groups of features: sound sample annotations and associated metadata, audio features, and semantic features. It is freely available for public use[3].

Sourcing the Sounds All sounds in the dataset are sourced from the Freesound archive[4]. We built tools to rapidly explore the archive and re-label sound samples, searching for likely candidates based on tags and descriptions, and finally filtering by star and user ratings. Each candidate sound was split into 5 second increments (and shorter sounds were extended to 5 seconds) during audition.

A major goal in our curation was to find audio samples that spanned the space from "common and easy to identify" to "common but difficult to identify" and finally to "uncommon and difficult to identify". We explicitly sought an even distribution of sounds in each broad category (approximately 130 sounds) using rudimentary blind self-tests. In sourcing sounds for the first two categories, we attempted to select samples that form common scenes one might encounter, such as *kitchen, restaurant, bar, home, office, factory, airport, street, cabin, jungle, river, beach, construction site, warzone, ship, farm,* and *human vocalization*. We avoided any samples with explicit speech.

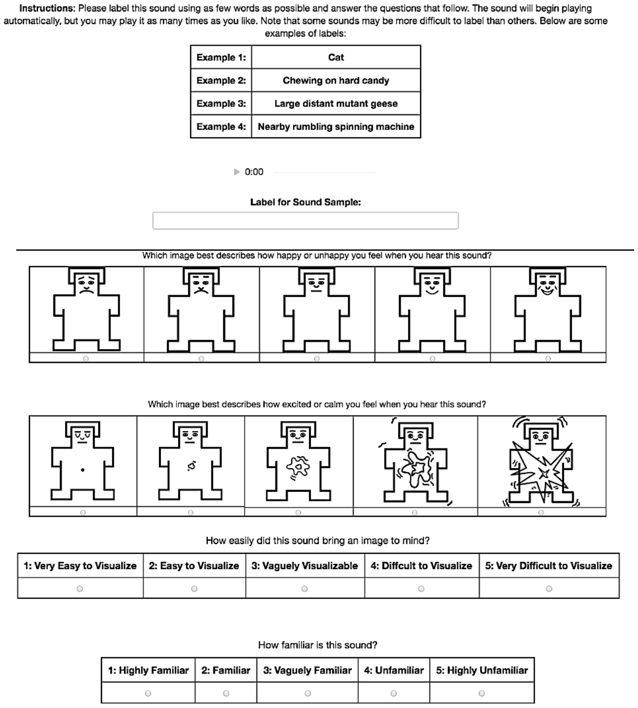
To source unfamiliar and ambiguous sounds, we include digitally synthesized samples in addition to artificially manipulated everyday sounds. Our manipulation pipeline applies a series of random effects and transforms to our existing samples from the former categories, from which we curated a subset of sufficiently unrecogniz-

[3] http://github.com/mitmedialab/HCU400

[4] https://freesound.org

able results. Effects include reverberation, time reversal, echo, time stretch/shrink, pitch modulation, and amplitude modulation.

Annotated Features We began by designing an Amazon Mechanical Turk (AMT) experiment in which participants were presented with a sound chosen at random. Upon listening as many times as they desired, they then provided a free-text description alongside likert ratings of its familiarity, imageability, arousal, and valence (as depicted by the commonly used self-assessment manikins [93]). The interface additionally captured metadata such as the time taken by each participant to complete their responses, the number of times a given sound was played, and the number of words used in the free-text response. Roughly 12000 data points were collected through the experiment, resulting in approximately 30 evaluations per sound after discarding outliers (individual workers whose overall rankings deviate strongly from the global mean/standard deviation). A screenshot of the interface can be seen in Figure 4.1



Instructions: Please label this sound using as few words as possible and answer the questions that follow. The sound will begin playing automatically, but you may play it as many times as you like. Note that some sounds may be more difficult to label than others. Below are some examples of labels:

Example 1:	**Cat**
Example 2:	**Chewing on hard candy**
Example 3:	**Large distant mutant geese**
Example 4:	**Nearby rumbling spinning machine**

▶ 0:00

Label for Sound Sample:

Which image best describes how happy or unhappy you feel when you hear this sound?

○ ○ ○ ○ ○

Which image best describes how excited or calm you feel when you hear this sound?

○ ○ ○ ○ ○

How easily did this sound bring an image to mind?

1: Very Easy to Visualize	**2: Easy to Visualize**	**3: Vaguely Visualizable**	**4: Diffcult to Visualize**	**5: Very Difficult to Visualize**
○	○	○	○	○

How familiar is this sound?

1: Highly Familiar	**2: Familiar**	**3: Vaguely Familiar**	**4: Unfamiliar**	**5: Highly Unfamiliar**
○	○	○	○	○

Figure 4.1: The AMT interface used to collect the crowd-sourced annotations which form the HCU400 dataset.

Hcu Features A novel contribution of this work is the estimation of H_{cu} using word embeddings and knowledge graphs, applied to the set of free-text labels accompanying each audio sample. Traditionally, these graphs are used to geometrically capture semantic word relationships; here, we leverage the "clustering radius" of the set of label embeddings as a metric for each sound's H_{cu}.

We employed three major approaches to embed each label: (1) averaging all constituent words that are nouns, verbs, adjectives, and adverbs– a common average encoding technique [96]– (2) choosing only the first or last noun and verb, and (3) choosing a single "head word" for each embedding based on a greedy search across a heavily stemmed version of all of the labels (using the aggressive Lancester Stemmer [97]). In cases where words are out-of-corpus, we autocorrect their spelling, and/or replace them with a synonym from WordNet [98] where available. Labels that fail to cluster are represented by the word with the smallest distance to an existing cluster for that sound (using WordNet path-length). This greedy search technique is used to automatically generate the group of labels used in the H_{cu} calculation. Both Word2Vec [99] and Conceptnet Numberbatch [100] were tested to embed individual words.

After embedding each label, we derived a "cluster radius" score for the set of labels, using the mean and standard deviation of the distance of each label from the centroid as a baseline method. We also explore (k=3) nearest neighbor intra-cluster distances to reduce the impact of outliers and increase tolerance of oblong shapes. Finally, we calculate the sum of weighted distance from each label subgroup to the largest "head word" cluster– a technique which emphasizes sounds with a single dominant label. We also include a location-based embedding to capture information pertaining to the likelihood of concept co-location in a physical environment. In order to generate a co-location embedding, we implement a shallow-depth crawler that operates on ConceptNet's location relationships ('Located-Near', 'Located-At', etc) to create a weighted intersection matrix of the set of unique nouns across all our labels as a pseudo-embedding. Again, we derive the centroid location and mean deviation from the centroid of the labels (represented by the first unique noun) for a given sound sample.

All clustering approaches give a similar overall monotonic trend, but a qualitative analysis of cluster labels in conjunction with scores suggests that a distance-from-primary-cluster definition is most fitting. We focus on ConceptNet embeddings over others in the subsequent discussion, because it is explicitly designed to capture meaningful semantic relationships in language *and* in other domains that reflect world knowledge.

Our clustering results using a ConceptNet embedding are plotted

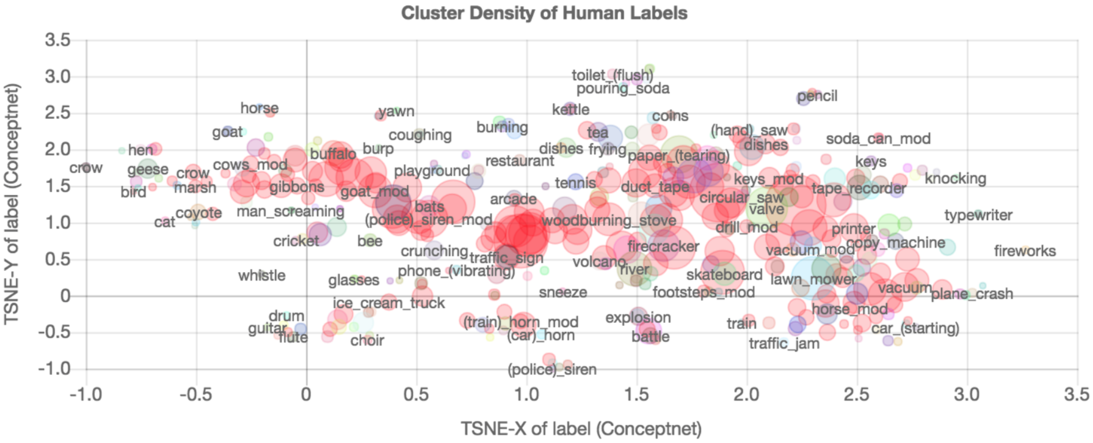


Figure 4.2: Average ConceptNet embedding where the radius represents our H_{cu} metric; red bubbles and the "_mod" suffix are used to indicate sounds that have been intentionally modified.

in Figure 4.2, and a qualitative example of user labels is given in Table 4.1 Intentionally modified sounds are plotted in red, and we see most sounds with divergent labeling fall into this category. Sounds that have not been modified are in other colors, reflecting the cluster size– here we see examples of completely unambiguous sounds, like human vocalizations, animal sounds, sirens, and instruments.

Typing	**Modified Chair Sliding**
Cluster Radius = 6.3	Cluster Radius = 8.7
typing on a keyboard	*bowling ball*
Typing on keyboard	*electric tube*
typing	*PVC pipe building pressure and release*
typing	*Error message on computer*
Typing on a keyboard	*High Speed Frisbee*
someone typing	*beer mug sliding on bar*
Someone typing on keyboard	*driving a car*
keyboard	*toy car hitting wall*
typing	*filling up a tub*
...	...

Table 4.1: Example labels provided by annotators for two audio samples in the HCU400 set, which result in different cluster radii.

4.2.2 *Baseline Analysis*

First, we find that the likert annotations are reliable amongst online workers, using a split ranking evaluation adapted from [101]. Each of the groups consisted of 50% of the workers, and the mean ranking was computed after averaging N=5 splits. The resulting spearman rank coefficient value for each of the crowd-sourced features is given in Figure 4.3. This provides the basis for several intuitive trends in our data, as shown by Figure 4.4 – we find a near linear correlation between mean imageability and familiarity, and a significant inverse correlation between mean arousal and valence. We also find a strong correlation between imageability, familiarity, time-based individual measures of uncertainty (such as such "time to first letter" or "num

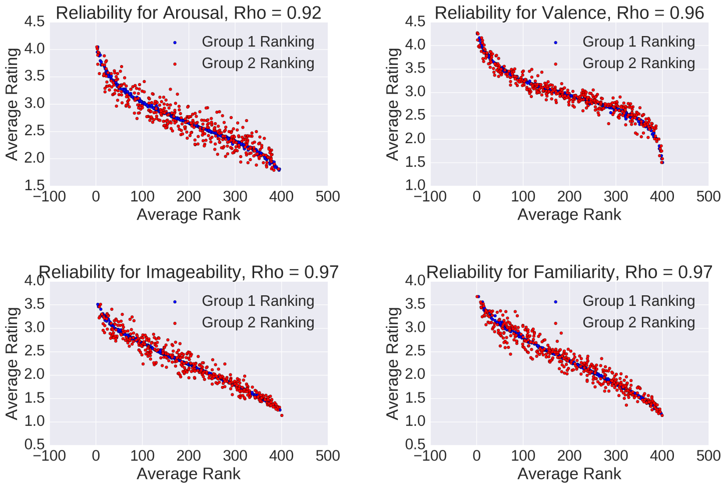


Figure 4.3: Split ranking correlation plots and Spearman rank coefficient values for the four likert annotated features.

of times played"), and the label-based, aggregate measures of uncertainty (the cluster radii and H_{cu}).

We next see strong evidence of the value of word embeddings as a measure of causal uncertainty – the automated technique aligns well with the split of modified/ non-modified sounds (see Fig. 4.2) and a qualitative review of the data labels. Furthermore, we use this data to explore the causal relationship between average source uncertainty and individual assessment behavior. In Figure 4.5, we plot the distributions of pairs of features as a function of data points within the 15th (red) and greater than 85th (blue) percentile of a single conceptnet cluster metric. It confirms a strong relationship between the extremes of the metric and individual deliberation (bottom right), as reported by [102]. We further find that more ambiguous sounds have less extreme emotion ratings (top right); the data suggest this is not because of disagreement in causal attribution, but because individuals are less impacted when the source is less clear (bottom left). This trend is not true of imageability and familiarity, however; as sounds become more ambiguous, individuals are more likely to diverge in their responses (top center). Regardless, we find a strong downward trend in average familiarity and imageability scores as the source becomes more uncertain (top left).

4.3 *Intrinsic Memorability*

We next set out to obtain annotations for the intrinsic memorability of each audio sample in the HCU400 dataset [5][3]. In order to quantify memorability, we drew inspiration from work in [101], which

[5] An interactive demo of the dataset can be found at" `https://resenv.media.mit.edu/memory-dataset/demo.html`

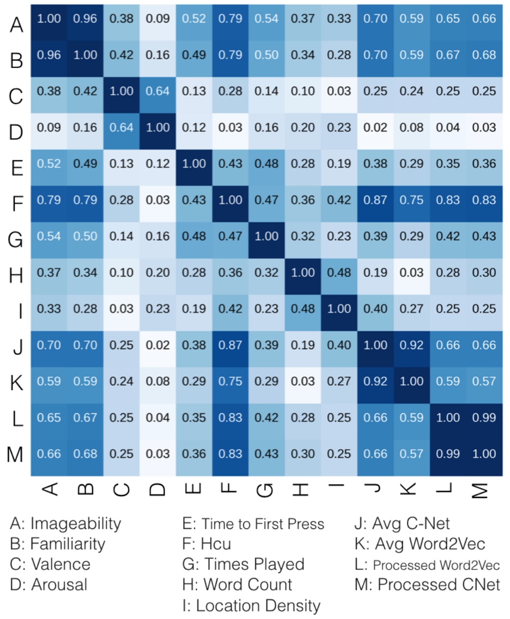


Figure 4.4: Correlation Matrix displaying the absolute value of the Pearson correlation coefficient between the mean values of annotated features, metadata, and four representative word embedding based clustering techniques.

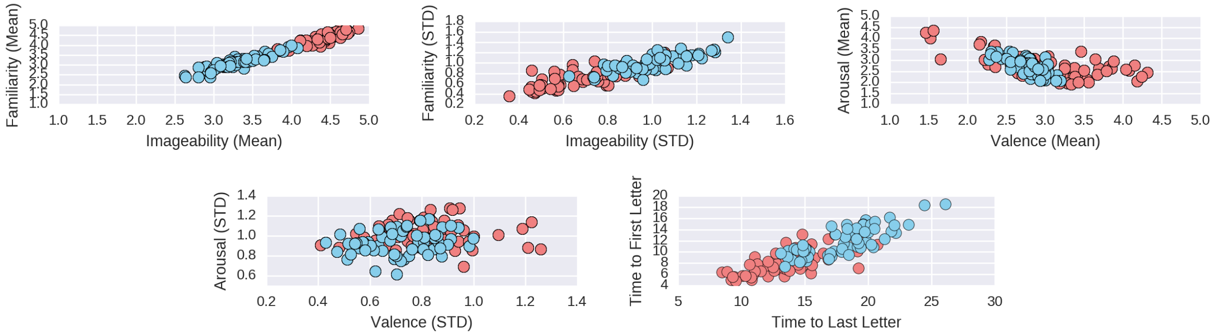
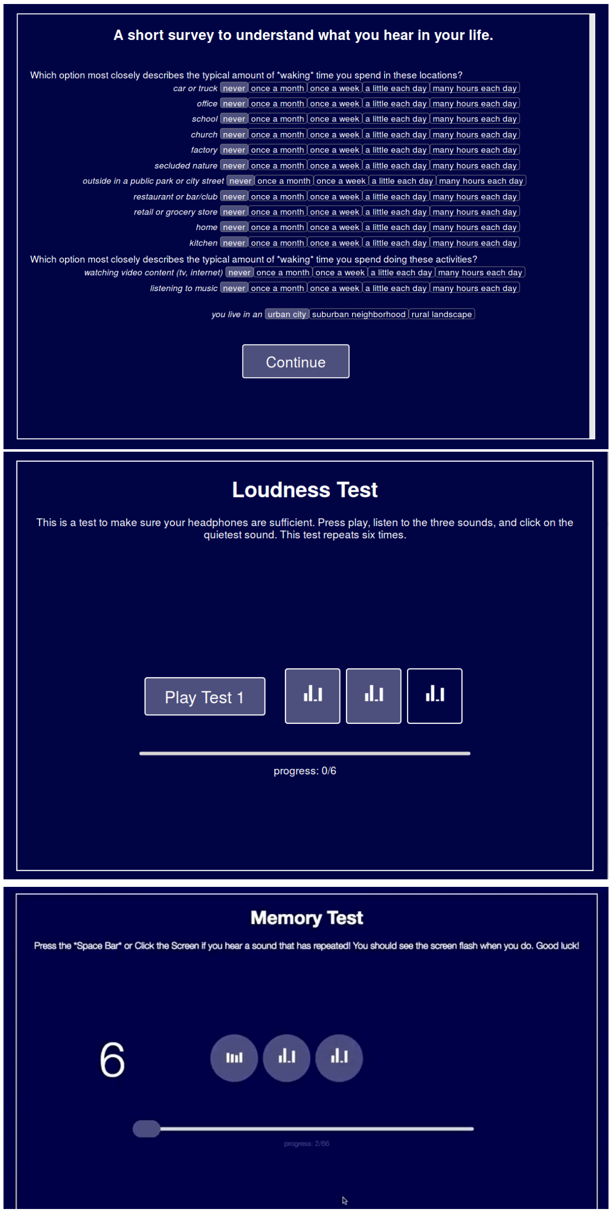


Figure 4.5: Feature distributions grouped by extremes in the "Processed CNET" cluster metric; red points represent data at $\leq$ 15th percentile (low H_{cu}); blue dots are $\geq$ 85th percentile (high H_{cu}).

used an online memory game to determine the features that make images memorable. We designed an analogous interface for the audio samples in the HCU400 dataset. The game opens with a short auditory phase alignment-based assessment [103] to ensure that participants are wearing headphones, followed by a survey that captures data about where they spend their time (urban vs. rural areas, the workplace vs home, etc). Participants are then presented with a series of 5 second sound clips from the HCU400 dataset, and are asked to click when they encounter a sound that they've heard previously in the task. At the end of each round consisting of roughly 70 sound clips, the participant is provided with a score. Screenshots of the interface at each stage are shown in Figure 4.6.

By design, each round of the game consisted of 1-2 pairs of *target sounds* and 20 pairs of *vigilance sounds*. *Target sounds* were defined as



A short survey to understand what you hear in your life.

Which option most closely describes the typical amount of *waking* time you spend in these locations?

car or truck	[✓ never]	once a month	once a week	a little each day	many hours each day
office	[✓ never]	once a month	once a week	a little each day	many hours each day
school	[✓ never]	once a month	once a week	a little each day	many hours each day
church	[✓ never]	once a month	once a week	a little each day	many hours each day
factory	[✓ never]	once a month	once a week	a little each day	many hours each day
secluded nature	[✓ never]	once a month	once a week	a little each day	many hours each day
outside in a public park or city street	[✓ never]	once a month	once a week	a little each day	many hours each day
restaurant or bar/club	[✓ never]	once a month	once a week	a little each day	many hours each day
retail or grocery store	[✓ never]	once a month	once a week	a little each day	many hours each day
home	[✓ never]	once a month	once a week	a little each day	many hours each day
kitchen	[✓ never]	once a month	once a week	a little each day	many hours each day

Which option most closely describes the typical amount of *waking* time you spend doing these activities?

watching video content (tv, internet)	[✓ never]	once a month	once a week	a little each day	many hours each day
listening to music	[✓ never]	once a month	once a week	a little each day	many hours each day

you live in an [✓ urban city] suburban neighborhood rural landscape

Loudness Test

This is a test to make sure your headphones are sufficient. Press play, listen to the three sounds, and click on the quietest sound. This test repeats six times.

Memory Test

Press the *Space Bar* or Click the Screen if you hear a sound that has repeated! You should see the screen flash when you do. Good luck!

Figure 4.6: Screenshots of the auditory memory game interface presented to participants as a part of our study.

samples from the dataset that were separated by exactly 60 samples– the sounds for which memorability was being assessed in a given round. The *vigilance sounds*, pairs of sounds that were separated in the stream by 2 to 3 others, were used to ensure reliable engagement throughout the task following the method in [101]. Roughly 20,000 samples were crowd-sourced on Amazon Mechanical Turk such that a single task consisted of a single round in the game. Individual workers were limited to no more than 8 rounds to ensure that target samples were not repeated. Rounds that failed to meet a minimum vigilance score (>60%) or exceeded a maximum false positive rate

(>40%) were discarded.

Audio samples for this test were taken from the HCU400 dataset [104], and standard low-level acoustic features were extracted from each sample based on prior precedent [105]. We used default configurations from three audio analysis tools: Librosa [106], pyAudioAnalysis [107], and Audio Commons [108], which include basic features (i.e. spectral spread) as well as more advanced timbral modeling. We supplement these features with additional summary statistics like high/mid/bass energy ratios, percussive vs. harmonic energy, and pitch contour diversity. We also include a vision-inspired perceptual saliency model following the procedure proposed by [28], applying separate temporal, frequency, and intensity kernels to an input magnitude spectrogram to produce three time-frequency salience maps. From these maps, we compute a series of summary statistics to be used as features.

High-level features were taken from [1] and include causal uncertainty (H_{cu}), the cluster diameter of embedding vectors generated from user-provided labels (quantifying source agreement or source location), familiarity, imageability, valence, and arousal.

4.3.1 Summary of Participant Data

We recruited 4488 participants, consisting of a small (<50) number of volunteers from the university community and the rest from Amazon Mechanical Turk. Our survey data shows that our participants report a 51/37/12% split between urban, suburban, and rural communities. We see weak trends in the average time per location reported for each community type– urbanites self-report spending less time at home, in the kitchen, in cars, and watching media on average. Rural participants report spending more time in churches and in nature. Using KNN clustering and silhouette analysis, we find four latent clusters – students (590 users), office workers (1250 users), home-makers (1640 users), and none of these (1010 users). Split-rank comparisons between groups did not reveal meaningful differences in results across user groups; we speculate any differences due to ecological exposure of sounds between environments is not consistent or influential enough at this group level to alter performance.

4.3.2 Summary of Memory Data

The raw memorability score M for each sound is simply computed as the number of times it was correctly identified as the target divided by the number of its appearances. However, this does not account for

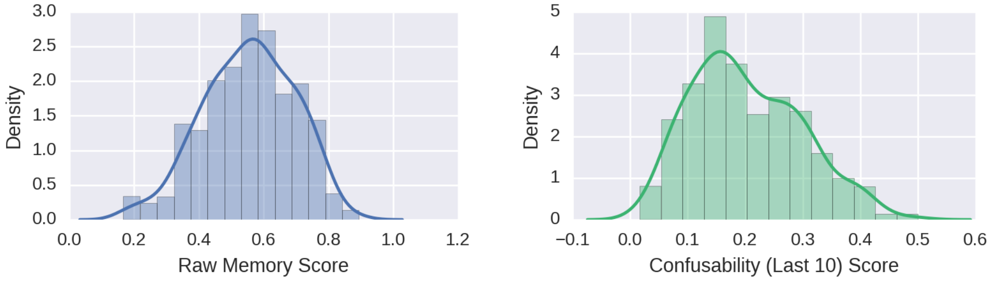


Figure 4.7: A histogram of the raw scores for each sound – they were successfully remembered and identified about 55% of the time on average, with a large standard deviation (left); A histogram of "confusability" scores for each sound, with an average score of about 25% (right).

the likelihood that the sound will be falsely remembered (i.e. clicked on without a prior presentation). We additionally compute a "confusability" score C_{10} for each sound sample, defined as the false positive rate for sounds when they fall close to the second target presentation (i.e. in the last ten positions of the game). We can thus derive a "normalized memory score" represented by $M - C_{10}$. In attempting to understand auditory memory, we consider both what makes a sound memorable *and* what makes a sound easily mis-attributed to other sounds, whether those sounds are encountered in our game or represent the broader set of sounds that one encounters on a habitual basis. We therefore model both normalized memorability and confusability in this work.

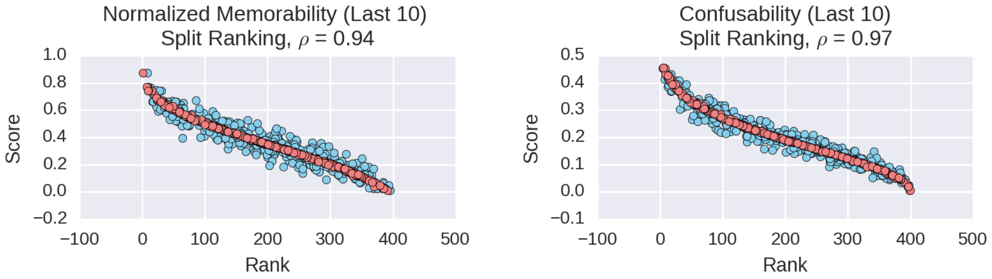


Figure 4.8: The results of the split-ranking analysis for the normalized memorability score and confusability score, using 5 splits; the Spearman coefficient correlations demonstrate the reliability of these scores across study participants, enabling us to model both metrics in the later parts of the work.

We confirm the reliability of both the normalized memory scores and the confusability scores across participants by performing a split ranking analysis similar to [101] with 5 splits, shown in Figure 4.8 with their respective Spearman correlation coefficients. This confirms that memorability and confusability are consistent, user-independent properties.

In Table 4.2, we show a short list of the most and least memorable and confusable sounds in our dataset as a function of the normalized memorability score and confusability score.

Most Memorable	**Least Memorable**
man_screaming.wav	*morphed_firecracker_fx.wav*
woman_screaming.wav	*truck_(idling).wav*
flute.wav	*morphed_turkey2_fx.wav*
woman_crying.wav	*morphed_airplane_fx.wav*
opera.wav	*morphed_metal_gate_fx.wav*
yawn.wav	*morphed_shovel_fx.wav*
Most Confusable	**Least Confusable**
garage_opener.wav	*clock.wav*
lawn_mower.wav	*morphed_335538_fx.wav*
washing_machine.wav	*phone_ring.wav*
rain.wav	*woman_crying.wav*
morphed_tank_fx.wav	*woman_screaming.wav*
morphed_printing_press_fx.wav	*vomit.wav*

Table 4.2: A list of the most and least memorable and confusable sounds from the HCU400 dataset.

4.3.3 Feature Trends in Memorability and Confusability

We consider two objectives – (1) to determine the relationship between individual features and our measured memorability and confusability scores, and (2) to determine the relative importance of these features in predicting memorability and confusability. To address the former, we provide the resulting R^2 value after applying a transform learned using support vector regression (SVR) for each individual feature. For the latter, we use a sampled Shapely value regression technique in the context of SVR– that is, we first take N random features (N between 1 and 10) and perform an SVR to predict memorability or confusability scores for our 402 sounds and the calculated R^2 of the fit. We then measure the change in R^2 as we append every remaining feature to the model, each individually. The largest average changes over 10k models are reported in Table 4.3. This technique is robust to complex underlying nonlinear relationships mapping the feature space to the predicted metric as well as to feature collinearity. We find that the strongest predictors of both memorability and confusability are the measures of imageability and causal uncertainty. Memorability is dominated by high level, gestalt features, with only one lower level feature ('pitch diversity') in the ten most important features. Low level features, including those derived from the auditory salience models, play a more significant role in determining confusability.

The absolute R^2 values indicate that no individual feature is a significant predictor of memorability by itself. This implies a complex causal interplay in feature space, which we explore further in the set of plots presented by Figure 4.9. In each plot, we show a distribution of feature values for the sounds that are most memorable or least confusable (>85th percentile, blue) contrasted against the least memorable or most confusable sounds (<15th percentile, red). We first consider the effect of H_{cu} and valence on memory– low memo-

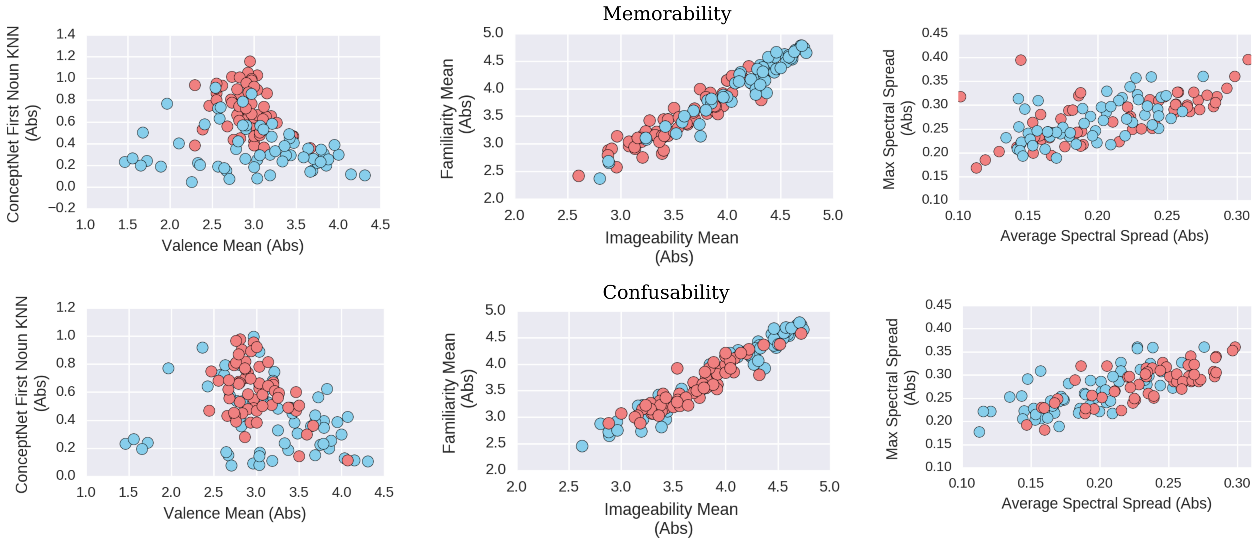


Figure 4.9: Scatter plots showing the changes in distribution of select features based on extremes in memorability (top row) and confusability (bottom row); blue indicates sounds that are most memorable or least confusable (85th percentile); red indicates sounds that are least memorable or most confusable (15th percentile).

rability and high confusability sounds exhibit a similar trend of high causal uncertainty and neutral valence (Column 1); In Column 2, we consider imageability and familiarity ratings, shown to be strongly collinear in [1]. Here, their relationship to memorability and confusability diverge; while both are positively correlated with memorability, *neutral* ratings are the stronger predictor of confusability. This suggests that we are most likely to confuse sounds if they are loosely familiar but neither strictly novel nor immediately recognizable. Finally, Column 3 reveals a discernible decision boundary in low-level feature space for confusability which doesn't exist in its memorability counterpart. The relative importance of low-level salience features, here represented by spectral spread, aligns with intuition– in the absence of strong causal certainty or affect feature values, our perception of sounds is driven by their spectral properties.

4.3.4 *Per-game Modeling of Memorability*

The aural context in which a sound is presented, which includes ecological exposure as well as the immediate preceding sounds in our audition task, may influence the memory formation process. The literature supports the notion that, given a context, unexpected sounds are more likely to grab our attention and engage memory [94]. To understand this effect in our test, we ran two studies based on a 5 sound context (approximating the limits of semantic working memory) and a 1 sound context (approximating the limits of echoic memory).

Top Predictors for Memorability and Confusability					
Memorability			Confusability		
Feature	R^2	Shapely Δ R^2	Feature	R^2	Shapely Δ R^2
Imageability	0.201	0.126	**Imageability**	0.065	0.078
Hcu	0.224	0.125	**Hcu**	0.073	0.078
Familiarity	0.176	0.123	Avg Spectral Spread	0.087	0.078
Valence	0.178	0.120	Peak Spectral Spread	0.037	0.076
Location Embedding Density	0.147	0.117	Peak Energy, Frequency Salience Map	0.059	0.076
Familiarity std	0.103	0.117	**Location Embedding Density**	0.100	0.076
Pitch Diversity	0.084	0.113	Frequency Skew, Frequency Salience Map	0.059	0.076
Imageability std	0.086	0.113	**Arousal**	0.039	0.076
Arousal	0.072	0.112	Peak Energy, Intensity Salience Map	0.044	0.075
Arousal std	0.056	0.111	**Familiarity**	0.045	0.075
Avg Spectral Spread	*0.099*	*0.107*	***Valence***	*0.100*	*0.075*
Timbral Sharpness	*0.094*	*0.091*	*Timbral Roughness*	*0.094*	*0.047*
Max Energy	*0.091*	*0.100*	*Avg Flux, Sub-band 1*	*0.092*	*0.064*
Treble Energy Ratio	*0.090*	*0.020*	*Flux Entropy, Sub-band 1*	*0.091*	*0.061*

Table 4.3: The top performing features from the Shapely regression analysis for both memorability and confusability (gestalt features are bolded); shown are the features ordered by their respective contributions to the R^2 value, with additional features with top performing individual R^2 values appended in italics. The first column indicates the individual predictive power of each feature; the second indicates its relative importance in the context of the full feature set.

Table 4.4 shows the results of a simple Support Vector classification model trained to predict whether the target in each game will be successfully recalled. This model was trained with the most memorable and least memorable sounds only (85th/15th percentiles) with a 5-fold cross-validation process, and results are reported on a 15% hold-out test set.

To begin, a baseline model is trained using the absolute, immutable features of the target sound. Because there are a limited number of sounds in our dataset relative to the number of games, the feature space is redundant and sparse, and we expect the accuracy of this model to converge to the average expected value over our set of sounds. We then introduce contextual features– the *relative difference* (z-score) of target sound features with those of the varying sounds that precede its first presentation in each game– to see if our model improves based on the context of our 50 most meaningful features (from the SVR analysis; 25 high-level and 25 low-level). In both 5- and 1-sound context cases, however, model performance does not improve as we would expect if the context provided additional useful information.

We also run a classifier that *only* uses contextual features, to ensure informative context has not been obscured or subsumed by the absolute features in our first test. We start with a noise baseline, in which contextual features are calculated using a random, incorrect context– these features are still informative as the z-score depends largely on the absolute features of the target sound. We then train the same model with the proper context to assess the difference in performance. There is no improvement when the true context is reintroduced.

This leads us to a meaningful insight, contrary to our hypothesis – context does *not* exert a measurable influence on our results. While context likely does matter in real-world settings, we suspect that our memory game framework indirectly primes participants to expect otherwise surprising sounds. This confirms that our data is the consequence of truly intrinsic properties of the sounds themselves, independent of immediate context *and* participant ecological exposure (as was demonstrated in the split-rank analysis).

Memorability Per-Game Models	
Features	**Accuracy (%)**
Absolute + All 5-Sound Context Feats (working semantic)	68.0
Absolute + Top 50 5-Sound Context Feats	69.1
Absolute Feature Only Baseline (~expected value)	*70.3*
Contextual Only, 5-Sound Context (working semantic)	62.5
5 Sound Context, Noise Baseline	*64.1*
Absolute + All 1-Sound Context Feats (echoic)	68.0
Absolute + Top 50 1-Sound Context Feats	69.5
Absolute Feature Only Baseline (~expected value)	*70.3*
Contextual Only, 1-Sound Context (echoic)	60.0
1 Sound Context, Noise Baseline	*61.3*

Table 4.4: The influence of contextual sounds before the first presentation of the target on our ability to predict recall across games.

4.3.5 *Summary of Findings*

In this leg of the work, we broadly showed that (1) the most important features that contribute to a sound being remembered are gestalt – namely, the sounds with clear sound sources (low Hcu), that are easy to visualize, are familiar, and elicit strong emotions; (2) high Hcu sounds that are not familiar to listeners or are not easy to visualize are more likely to be "confused" and misattributed, and low-level or acoustic features play an important role in predicting this behavior; and (3) these relationships are not influenced by the context of surrounding sounds within the confines of the game, and are intrinsic properties of the sounds themselves. Although this baseline analysis is only cursory, we believe that the publicly available data is rich with a plethora of insights and relationships that we hope many others will investigate and uncover.

4.4 Bootstrapping General Purpose Estimators

As hinted at in the discussion surrounding the preliminary findings from our memory assessment, it is challenging to construct end-to-end models that can estimate gestalt or intrinsic memorability characteristics directly from audio. This stems from both the small-scale nature of the dataset (as a function of the total number of audio examples in the datasets and the number of annotations per example that was feasible to acquire), and the inter-rater diversity amongst the annotations.

Given this, in order to generalize gestalt and memorability characterization to unseen, real-world audio, we require an alternate way to obtain a general purpose estimator. We suggest a simple bootstrapping approach that allows us to construct a model with the *sound source label* as the input, mimicking the role of causal certainty in sound understanding as highlighted by Gaver [72]. To do this, we propose a strategy for mapping the gestalt and memorability scores from the small HCU400 dataset to Google's AudioSet [9, 2] ontology. We choose the AudioSet ontology as it is a set of class labels that accompanies one of the largest available sound object datasets for deep learning research, and is therefore the ontology that lies in the output space of many state-of-the-art audio classification neural networks. Therefore, given a reliable strategy for mapping from a small dataset to a fixed ontology, both the refinement and expansion of the label set as well as advances in audio classification networks from the deep learning community have the power to enable increasingly nuanced cognitive labels for real-world audio.

The problem that's posed here – attempting to build a modeling structure that can generalize from a smaller dataset to diverse, unseen data – is a common problem encountered in machine learning. One method used to tackle this problem is transfer learning, wherein a model first trained on a larger, related dataset is fine-tuned via training on the smaller, target dataset [109]; an instance of this class of strategies is explored in Section 3.2. Another approach entails the use of Few-shot learning (or Low-shot learning), wherein a model relies on an intermediate relational space, metric learning, the synthesis of additional training examples, or improved weight optimization methods tailored to the scenario of limited per-class examples, in order to assign labels to unseen data. Groundbreaking examples of advancements in this area include the development of Matching Networks [110], Prototypical Networks [111], Triplet Networks [112], the Model-Agnostic Meta Learning (MAML) method [113],

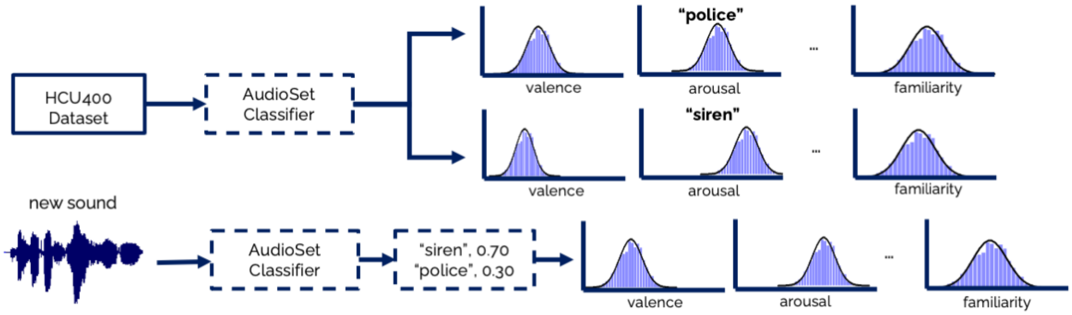


Figure 4.10: A summary of the proposed bootstrapping approach. We first bootstrap gestalt property scores to all labels in the AudioSet ontology by running a classifier on the HCU400 dataset *(top)*; we then estimate gestalt property scores for unseen audio examples by first predicting AudioSet labels, and then combining the scores associated with these labels, weighted by the prediction uncertainty *(bottom)*.

Memory-Augmented Neural Networks [114], and the "Shrinking and Hallucinating" strategy [115].

However, we choose a related but slightly different approach, using the AudioSet ontology as an intermediary in the manner mentioned above, for two reasons: (1) firstly, in our case, we seek an intermediate structure that affords interpretability, which we wish to achieve by drawing structure from the learnings that stem from empirical human psychology research. While it is unclear whether this objective provides advantages in terms of generalizability (see Section 4.4.2), we see value in it when applying the resulting estimators to devices or experiences, as interpretability allows us to more easily identify and understand failure cases. (2) Secondly, we also want a modeling structure that allows us to represent *uncertainty*; uncertainty in this context does not refer to the stochastic measurement of a single ground truth label or quantity, but refers rather to annotator diversity which we wish to capture explicitly and propagate to downstream interfaces or experiences. Several machine learning research efforts which attempt to model gestalt principles – for instance, affect recognition tasks [116] – collapse annotator diversity into point objectives at the dataset preparation or model training phase; we can intuit that constructing models that attempt to estimate these objectives provide little value, especially as the community considers objectives that represent increasingly higher-level aspects of cognition where we expect to observe greater amounts of diversity. Given these two constraints, we believe that the approach we design, which is detailed below, is a suitable option; however, we do explicitly discuss the limitations of the approach in Section 4.4.3.

4.4.1 Approach

To illustrate our approach (summarized in Figure 4.10), we consider the HCU400 and memorability datasets presented in [1, 3], and aim to scale the hand-labeled annotations of six gestalt properties – arousal, valence, imageability, familiarity, memorability, and confusability[6] – to unseen audio. We achieve this by building a probabilistic mapping between these scores and the 600+ labels in the AudioSet ontology. To do this, we first consider audio samples X_k in the HCU400 dataset, with a corresponding set of annotations for a given gestalt feature S_k, which consists of all of the individual crowd-sourced ratings s_k^n collected for that feature and audio sample. Each of these audio samples has an associated set of AudioSet labels. This setup is illustrated in Table 4.5.

Table 4.5: An illustration of the formulation that enables bootstrapping.

HCU400 Dataset	Gestalt Annotations	AudioSet Label Predictions
X_k	$S_k = \{S_k^0, \ldots, S_k^{N-1}\}$	$\{\ldots, l, \ldots\}$
$\ldots$	$\ldots$	$\ldots$
X_{k+1}	$S_{k+1} = \{S_{k+1}^0, \ldots, S_{k+1}^{N-1}\}$	$\{l, \ldots, \ldots, \}$

For a given label l in the AudioSet ontology, we wish to obtain a gaussian estimate $\mathcal{N}_l(\mu, \sigma^2)$ of the associated gestalt feature score[7]. To do this, we compute:

$$\mathcal{N}_l \leftarrow hist\left(\frac{\sum\limits_{k \in K} F_\theta(p^k(l)) * S_k}{\sum\limits_{k \in K} F_\theta(p^k(l))}\right) \tag{4.1}$$

where $p^k(l)$ represents the classification probability associated with label l as a prediction for sound X_k, scaled and normed by a function F_θ, a hyperparameter. We use the probabilities as a "weighting" on the frequency of the set of ratings S_k, and combine all of the ratings across sound samples where l appears as a predicted label into a single data series. We compute a histogram of this data, and estimate a best-fit gaussian, $\mathcal{N}_l$.

Unfortunately, the labels associated with the sounds in the HCU400 dataset are too sparse to fully cover all of the AudioSet labels, so some labels l remain without an estimated $\mathcal{N}_l$. To address this, we can exploit the existing class relationships in the AudioSet ontology to meaningfully impute our estimates of the new gestalt properties to the uncharacterized labels: parents adopt mean scores of children, children inherit parent scores. Example results from the complete label estimation and imputation process can be seen in Table 4.7.

Table 4.6: An illustration of the formulation that enables inference.

Unseen Audio Recording	AudioSet Label Predictions	AudioSet Label Gestalt Scores
X_r	$\{l_0, \ldots, l_{n-1}\}$	$\{\mathcal{N}_{l_0}, \ldots, \mathcal{N}_{l_{n-1}}\}$
$\ldots$	$\ldots$	$\ldots$

[6] The previous section suggests that "confusability" is predicted well by acoustic features; however, we expect to find interesting semantic links as well, since the ontology we use also includes some tags that assign language to acoustic properties – for instance, "rumbling" or "buzzing".

[7] We note that a gaussian distribution does not have the capacity to represent, for instance, bimodal responses; however, we expect the variance to still be a useful measure of dispersion for our illustrative tasks, while keeping the math convenient.

To calculate gestalt property scores for unseen audio examples, this process can effectively be inverted. We consider the setup in Table 4.6, where each new audio example X_r has associated AudioSet labels $\{l_0, \ldots, l_{N-1}\}$, and associated gestalt scores $\{\mathcal{N}_{l_0}, \ldots, \mathcal{N}_{l_{N-1}}\}$ that map to each label for the gestalt feature in question. We compute:

$$\hat{\mathcal{N}}_r = \left(\frac{\sum\limits_{n \in N} J_\phi(p^r(l_n)) \cdot \mathcal{N}_{l_n}}{\sum\limits_{n \in N} J_\phi(p^r(l_n))} \right) \tag{4.2}$$

where $p^r(l_n)$ represents the prediction probability associated with label l_n for example X_r. We use this probability, after scaling via hyperparameter function J_θ and norming, to weight and sum the pre-computed gaussian functions $\mathcal{N}_{l_n}$ assigned to each label l_n, obtain an estimated $\hat{\mathcal{N}}_r$ for the new recording.

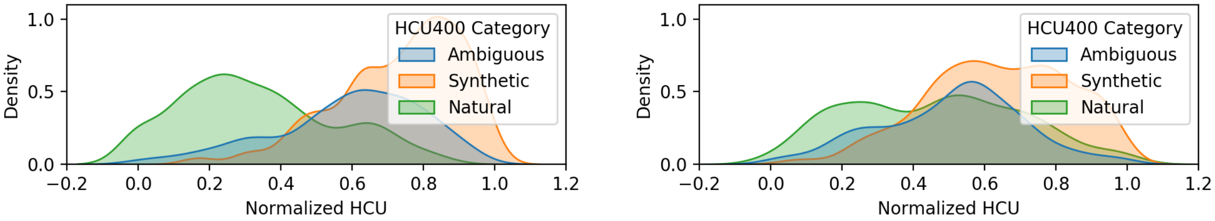


Figure 4.11: Plots showing the distribution of sounds in the HCU400 dataset labelled with their original categories from [1], namely "Natural", "Ambiguous", or "Synthetic". We compare the separation of these "Natural" and "Synthetic" classes via the human annotated Hcu metric (left) with the proposed, neural network-based approach (right).

Throughout this approach, we treat the uncertainty of the pre-trained AudioSet model as a proxy for human uncertainty in sound source identification. As shown in Figure 4.11, plotting the audio classes from the HCU400 dataset against the human-rated and artificial causal uncertainty measures demonstrates that it is a reasonable proxy. We use this notion implicitly in the bootstrapping process as we weight the contribution from different instances in the HCU400 dataset by the network prediction uncertainty, mimicking the role of causal uncertainty as the fulcrum between semantic and acoustic processing [73, 72]. We can also use this notion explicitly as a tool for constructing experiences (see Chapter 7).

The intermediary structure in this approach can be constructed using a spectrum of methods, spanning simple causal intuition derived from auditory psychology literature to rigorous bootstrapping from more extensive datasets onto detailed ontologies. In contrast to traditional transfer or few-shot learning approaches, the structure here has intuitive meaning, and we rely on explicit relationships in label and language space to provide scaffolding for relationships in cognitive understanding space.

Top Scores: "Memorability"	Top Scores: "Confusability"
Guitar	Rain on surface
Wail, moan	Pink noise
Fire alarm	Ocean
Baby cry, infant cry	Vibration
Crying, sobbing	Traffic noise, roadway noise
Cough	Idling
Singing	Stream
Whistling	Fire
Chuckle, chortle	Typewriter
Belly laughter	Wind
Baby laughter	Rustling Leaves
Ambulance (siren)	Thump. Thud
Sneeze	Electric shaver, electric razor

Top Scores: “Arousal”	Top Scores: “Valence”
Skidding	Acoustic Guitar
Machine gun	Strum
Ambulance (siren)	Wind Chime
Emergency vehicle	Chuckle, chortle
Toot	Giggle
Train horn	Laughter
Fire alarm	Flute
Vehicle horn, car horn, honking	Cello
Growling	Classical Music
Doorbell	Waterfall
Ringtone	Bird call, bird song
Boom	Rain on surface
Roar	Church bell

Table 4.7: Sound category labels from the AudioSet [2] ontology with top scores for Memorability, Confusability, Arousal, and Valence, determined by bootstrapping from the small HCU400 dataset [1, 3].

4.4.2 *Baseline Validation*

For the sake of completeness, we describe the results of a simple validation experiment, entailing a test of the bootstrapping approach on a very small, external dataset annotated in the same format as the HCU400 dataset. We use the ESC-50 dataset[8], a publicly available dataset of environmental sounds, and obtain 30 crowd-sourced ratings for each of the likert features in HCU400 – arousal, valence, imagebility, and familiarity[9]. We use our bootstrapped approach to estimate these features in the ESC-50 dataset, and compare it against a very standard zero-shot learning approach that relies on a weighting scheme based on distance in the AudioSet vector embedding space [117]. We report the results as KL-divergences comparing the discrete distributions (the predicted distribution evaluated at the likert intervals, and the annotated ground truth), aggregated across of the sound examples. The results are given in Table 4.8.

While our approach does perform marginally better than the zero-shot approach, we provide these results only to understand whether our approach is within a reasonable ballpark of standard transfer learning or few shot approaches. We do not believe that the value of

[8] `https://github.com/karolpiczak/ESC-50`

[9] Running fresh cycles of memorability games at the rate of the original dataset (approximately 20,000 games) with the ESC-50 dataset was prohibitively expensive and time consuming at this stage, so we exclude memorability and confusability from this analysis.

	Bootstrap HCU400 (train) μ, σ	Bootstrap ESC-50 μ, σ	Zero Shot ESC-50 μ, σ
arousal	0.12, 0.11	0.29, 0.24	0.35, 0.29
valence	0.19, 0.15	0.35, 0.27	0.47, 0.34
familiarity	0.22, 0.18	0.39, 0.22	0.56, 0.34
imageability	0.20, 0.14	0.40, 0.24	0.54, 0.31

Table 4.8: Validation experiment results. The mean and std of the KL divergence is reported in a comparison between the our bootstrapping approach and a zero-shot learning approach on the ESC-50 dataset. The training performance of the bootstrapping method is given for reference.

this work is in finding a state-of-the-art transfer learning technique, but in finding a technique that reflects hearing science research and is interpretable.

4.4.3 *Limitations*

We see two major sources of limitations in the presented approach – one associated with the intermediate structure we choose for bootstrapping, and one associated with our dataset annotation and analysis process. In the former case, we first note that AudioSet has only a finite number of classes to which a classification model may map a new sound; as a result, substantially erroneous gestalt quantities might be estimated for sounds that do not fit neatly into the ontology – examples include synthesized sounds or inorganic sounds whose causality may still be readily apparent to a listener. We also point out that the semantic relationships in the AudioSet ontology that we rely on for the bootstrapping process were not originally curated in an entirely crowd-sourced manner; the relationships are inferred from hyponyms found across a large corpus of internet text and further curated manually by the original researchers – so the extent to which they are reflective of a vast demographic is difficult to discern. We describe a strategy for bridging the gap between these estimators and individual differences in Chapter 6, but having a more representative or more democratically collected structure even at this phase would be beneficial. We hope that the audio research community will consider these challenges and requirements as new ontologies are created and associated classification engines are trained.

In the latter case, while our dataset formation process captures annotator disagreement and coarse trends in annotator demographics, we did not collect the demographic information at sufficient resolution to be able to statistically connect the two – that is, explicitly establish causal links between demographic subgroups and annotation behavior – prior to the estimation phase. We note that having had this capacity would afford us better estimators; therefore, revisiting our small-scale datasets (or those like them) with an eye towards this additional capability would be critical future work.

4.5 *Summary*

The work delineated in this chapter results in datasets that tackle aspects of auditory phenomenology and a set of models that attempt to extend some of the properties examined in the datasets to real-world audio. The more important takeaways from this series of work, however, surround the motivation behind and framework for constructing these data and models. The key ideas are:

- A thorough survey of auditory psychology and neuroscience literature is first conducted and distilled into factors of interest that together would form a valuable sound cognition dataset.
- The dataset is collected at crowd-scale, under the considerations of rater diversity, uncertainty, and unreliability of self-report.
- A structure for generalizing a subset of the factors in the dataset to unseen audio is proposed, using a transfer-learning scaffolding that takes inspiration from the literature on auditory pathways and capitalizes on advancing deep learning infrastructure for sound object segmentation and labeling.

I suggest that this paradigm of model building is more broadly useful, either in considering other, unexplored phenomena in sound cognition, or in working towards generalized models that draw from existing, small-scale datasets. In the forthcoming chapter, I connect these efforts back to the narrative surrounding new and/or enhanced experiences presented at the start of this dissertation, by building and evaluating systems that apply the cognitive models towards select, human-meaningful tasks.

5

Constructing Experiences

We now move on to the discussion of a system built with and around the models that we constructed in the previous chapter, that is suggestive of the sorts of experiences that motivate this research. In this section, we describe a concatenative synthesis engine that employs the gestalt estimation models to automatically create short audio presentations as a means of exploring large corpora of audio data along aesthetic, subjective dimensions. We present the motivation for the system, the role of the cognitive modeling in the system design, a detailed description of the system mechanics, and outcomes and learnings from corresponding user evaluations. In the larger context of this work, we treat the construction of and experimentation with this system as a means to examine (1) the unique affordances of a feature space powered by gestalt ideas over purely acoustic measures, and (2) the magnitude of the impact stemming from the system on the user, and whether or not we can link that impact explicitly to the underlying cognitive models.

5.1 Cognitive Content Curation: A Novel Audio Summarization Tool

Over the last decade, we have witnessed a massive shift towards ubiquity and capacity in audio capture. Low-power, always-on audio sensing technology in our homes and our phones makes it easy to record hours of uninterrupted data; the rapidly falling cost of storage infrastructure makes it even easier to archive. From lifelogged recordings to environmental monitoring databases, however, this trend has resulted in a paradox for consumption – the more audio there is, the harder it is for the average user to interact with the it, and the less likely they are to do so.

While research has produced many strategies for condensing large audio corpora into representations that are distilled in time, these approaches are frequently motivated by the goal of maximizing information in the output representation [118, 119, 120, 121, 122, 123, 124]. We argue that such approaches do not enable other, subjective modalities of engagement with or explorations of a specific body of audio, such as aesthetic or emotional modalities. For example, we find many means of distilling a body of audio so as to preserve only speech or detect new events in a soundscape, as may be useful in a memory aid or surveillance task, but few means to use the audio to create an appropriate background track for sleeping or studying, or for evoking nostalgia of a place and time. In addition, we note that most of the aforementioned systems distill audio by optimizing for a single, well-defined objective function; they do not represent tunable systems with exposed input parameters that can be reconfigured to generate perceptually diverse outcomes. We intuit that this shortcoming is driven by feature space design – these systems are largely built on feature representations that quantify statistical properties of the audio itself (spectral derivatives, event detection, audio quality, frame diversity, etc), instead of being rooted in the science of how the human brain perceives and processes sound.

We attempt to address this gap by constructing an audio summarization system with a feature space that is driven by the principles of auditory processing, incorporating the gestalt estimators described in Chapter 4. We aim to create a feature space that reflects gestalt/ memorability properties and acoustic measures, and expect that altering the relative weighting of these features at the input can produce a diversity of outcomes along emotional and aesthetic axes. Specifically, we contribute the following:

1. We construct a feature space derived from the gestalt, intrinsic memorability (which we collectively refer to as gestalt), and acoustic features examined in Chapter 4, employing the general purpose estimators constructed at the end of the chapter.

2. We build a tool that labels audio recordings with these features, surfaces short clips along the extremes of the features, and combines them in chronological order to generate audio summaries[1].

3. We evaluate our system in the context of more than 800 hours of "lifelogged" audio recordings, wherein 10 participants each collect high-fidelity, first-person recordings over the span of 1 to 3 weeks, and provide quantitative and qualitative feedback regarding the

[1] Audio examples that demonstrate this process on generic, ambient field recordings can be found at `https://ishwaryaanant.github.io/audio-stories/`

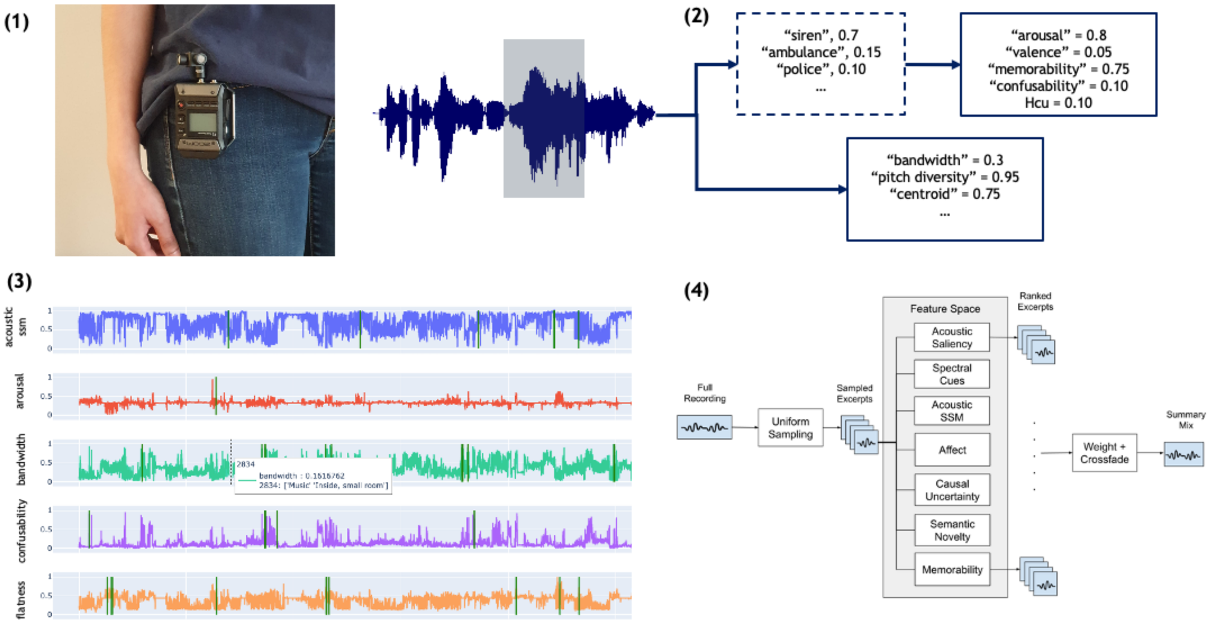


Figure 5.1: Our audio summarization tool takes real-world, lifelogged audio (1) as input, uses cognitive principles to extract acoustic and semantic feature information from the audio (2, 3), and reassembles the highest or lowest scoring excerpts to form "summaries" (4). The method allows for summaries that elicit specific subjective, perceptual responses in users, and results in listening experiences that are emotional, compelling, and engaging.

summaries generated from their own data.

4. We demonstrate the utility of our gestalt feature implementations in the context of audio summarization by showing that (1) acoustic and gestalt features statistically surface different content in long-term recordings; (2) that gestalt features are the strongest drivers of perceptual responses in study participants; and (3) that the presented system results in listening experiences that are immersive, intimate, artistically compelling, emotionally intense, and are suggestive of various user-highlighted use cases geared towards well-being.

Through this work, we suggest a novel paradigm for designing media summarization systems, and gesture towards a broader framework for capitalizing on elements of cognition to enable novel forms of media interaction.

5.1.1 *Related Work*

Content curation of large multimedia collections has been a problem of interest to the HCI community for many years, and has generated a significant body of literature. One approach condenses a large volume into a smaller presentation by selecting and recombining a subset of the content – this has been referred to as summarization

[125, 126]. Summarization efforts in multimedia typically focus on building systems with a singular objective function, with the aim of maximizing information [118, 119, 120, 121, 122, 123, 124, 125, 126] or representativeness of/ similarity to the larger media body [127, 128, 129, 130] in the output presentation. Common sub-themes of these goals include frame diversity and quality [125, 126], event detection [131, 132, 133, 134, 135], and preserving distinct speech [136, 122]; less frequently, we find examples of more abstract goals, such as memory retention or a qualitative assessment that a summary is "good" – however, these systems are often explicitly constructed using the aforementioned information criteria as input features [137, 138, 131]. Our system aims to address some of the gaps in this literature by (1) considering a set of outcomes beyond information maximization to generate summaries that allow users to engage with their soundscapes from an aesthetic, emotional, therapeutic, or sound-design standpoint; and by (2) incorporating principles of perception and cognition into the feature design space to enable these outcomes.

5.1.2 Feature Implementations

We create seven classes of feature extractors to operate on raw ambient audio streams as inputs to our summarization system: *Affect*, *Memorability*, *Causal Uncertainty*, *Semantic Novelty*, *Acoustic Saliency*, *Acoustic Self-similarity*, *Spectral Cues*. These extractors produce scalar feature curves associated with the audio as a function of time. The Affect, Memorability, and Causal Uncertainty implementations draw from the approach described in Section 4.4; the Semantic Novelty implementation examines the audio classification infrastructure itself as a tool for capturing relationships in sound label space; and the Acoustic Saliency, Acoustic Self-Similarity and Spectral Cues implementations are acoustic measures extracted directly from the audio, in part extended from the implementations in the HCU400/ memorability datasets (see Section 4.2 and Section 4.3).

Gestalt Features: Affect, Memorability, and Causal Uncertainty For the purpose of this system, we replicate the process described in Section 4.4 to obtain estimated scores for Arousal, Valence, Imageability, Familiarity, Memorability, and Confusability, and consider feature scores constructed from the mean of the predicted gaussians, using the standard deviation as a measure of confidence in the mean. We also replicate the idea in Section 4.4 regarding the quantification of causal uncertainty, using the probability assigned by YAMNet to the top performing AudioSet class and averaging it across the prediction frames in the duration of the sound clip.

Acoustic Features: Acoustic Self-Similarity, Acoustic Saliency, Spectral Cues We assess acoustic-level repetition through a measure of self-similarity, first employed in the context of audio by [139]. To obtain a score revealing how similar an audio excerpt is relative to itself and all other sampled audio excerpts, we compute a magnitude Short-time Fourier Transform (STFT) for each excerpt with 512 FFT bins and a hop size of 512 samples. To decrease the computational overhead, each STFT is then smoothed along the time axis with a window width of 10 units, and down-sampled along the same axis by a factor of 10. After concatenating the magnitude STFTs from all of the excerpts along the time dimension to create a single 2-dimensional representation, we compute a self-similarity matrix as the cosine distance between each pair of 512 unit time vectors in the new representation. We then obtain a novelty curve by summing along one of the matrix axes; the total novelty within the bounds of a single excerpt is assigned as the self-similarity score to that excerpt.

We posit that novelty on the semantic level also represents a valuable control in our feature space. We therefore compute a high-level self-similarity measure, calculated with the same algorithm as for low-level spectral self-similarity, except based on the centroid of each audio excerpt's YAMNet embeddings (where the centroid is the mean of the embeddings over all of the frames in an excerpt). Given the conceptual value of the embeddings as discussed in [9], we expect that audio excerpts whose embedding centroids stand out in this context represent contrasting semantic information relative to the other excerpts being compared.

Finally, we employ the saliency model and the spectral measures (pitch diversity, harmonic-percussive ratio, centroid, bandwidth, etc) detailed in Section 4.3 as a part of our feature space.

5.1.3 Summarization Tool

We present a description of our audio summarization tool that employs these feature classes to generate short audio presentations from extended recordings. An overview of the tool is as follows: we first select 3-second audio excerpts at equally-spaced time intervals throughout the body of audio, which is a concatenation of all recorded audio (over 1-3 weeks) with appropriate padding to ensure an excerpt does not span multiple recording files. We then extract the value of each of our features for each excerpt in the set. Our implementation outputs a ranking of all of the excerpts ordered by feature value assigned, and depending on the feature strategy preference specified, a subset of excerpts are selected for the final presentation. The selected excerpts, about 30 excerpts for a 1-minute summary, are finally cross-faded in chronological order and output as a single track. Figure 5.1(4) provides a detailed illustration of the data flow in the system.

The excerpts can be selected in one of two ways, which we call "feature strategies":

1. Top or Bottom Feature: a single feature strategy is specified, and excerpts are simply drawn from the top or bottom of the ordered ranking (for example, taking the top 30 acoustically salient samples is referred to as "Most Acoustically Salient", or taking the bottom 30 semantically novel clips is referred to as "Least Semantically Novel").

2. Baseline: a naive benchmark strategy for comparison within our listening experiments, where 30 excerpts are simply chosen at equally spaced intervals in time from the set of excerpts being analyzed, without any feature extraction and ranking.

5.1.4 Participant Evaluation

To understand the value of our system and the underlying feature space, we believe it is most appropriate to evaluate it in the context of the motivating applications – using lifelogged, first-person audio recordings – rather than generic databases of soundscape recordings which may be constrained in sound object content and demonstrate no personal relationship with a listener. For a more realistic study, we opt for an in-the-wild study conducted over the span of several months.

In our evaluation, participants were provided with wearable, high-fidelity stereo recorders (shown in Figure 5.1(1)) to capture their sonic environments for as many hours as possible during waking hours for 1 (minimum) to 3 (maximum) weeks. After signing up for the study and providing consent, participants were delivered their recorders in a contactless fashion, provided with detailed explanations on the use of the device and ethical best practices in a video call with the researchers, and given instructions on uploading their recorded audio to a secure server in our lab accessible only by the authors of this research. Significant precautions were taken to ensure the privacy of study participants and the individuals in their environments: based on guidance from a student law clinic associated with our institution, participants were allowed to pause and restart the recorder at any point during the day to avoid capturing sensitive content, and were required to obtain consent from individuals who could be identified in the recordings and wear the recording device in plain sight; the researchers also were not permitted to audition raw participant recordings at any point, and required explicit consent to audition generated summaries.

After the recordings were captured and uploaded by each individual, raw recordings were processed using the summarization tool, and were used to generate 16 summaries (see Figure 5.3) from the pool of possible feature strategies that we hypothesized would map to unique perceptual outcomes, in addition to the baseline strategy. The generated summaries were automatically embedded in an individualized survey, alongside several series of questions: for each summary, participants were asked to assign relevant perceptual descriptors from a pre-determined list ("calming", "nostalgic", "social", etc), and provide ratings on perceived emotional intensity, sense of intimacy, and positive or negative sentiment. Participants also provided lengthy qualitative descriptions of their listening experiences using the system, responding to guiding questions such as "What surprised you most or least about what you heard?" and "Did you find listening to the summaries to be an immersive experience?" The full set of questions in the survey is reproduced in Appendix A for reference.

We recruited N=10 participants (4 females, 6 males, aged between 25 and 65), via public advertisement at our institution and the surrounding community. The participants included undergraduate and graduate students, young professionals, and faculty members, who were associated with a spread of living spaces (dormitories, shared apartments, suburban independent housing with large families), working conditions (remote desk work, office space desk work, physical laboratory work), and experience with audio and music (from no inclination towards or experience with sound recording or production to semi-professional audio engineers and musicians). Participants collectively provided over 800 hours of audio, recording for 4-6 hours per day on average[2].

5.1.5 *Exploring Gestalt-Acoustic Clip Overlap*

We suggest that the value and novelty of this system is a feature space that incorporates gestalt auditory understanding; however, it is possible that low-level features correlate so strongly with high-level

[2] Audio summary examples can be requested from the researchers, contingent on permission from the participants.

ones that gestalt analysis is redundant for summarization tasks. To assess this aspect of our summarization system, we analyze whether high-level features independently surface novel content as compared to low-level features. To do this, we compute the percentage of overlap between clips that rank in the top and bottom 1 percentile of the entire pool of excerpts from a single participant's data, per feature strategy. We then average these results across all 10 participants, shown in Figure 5.2.

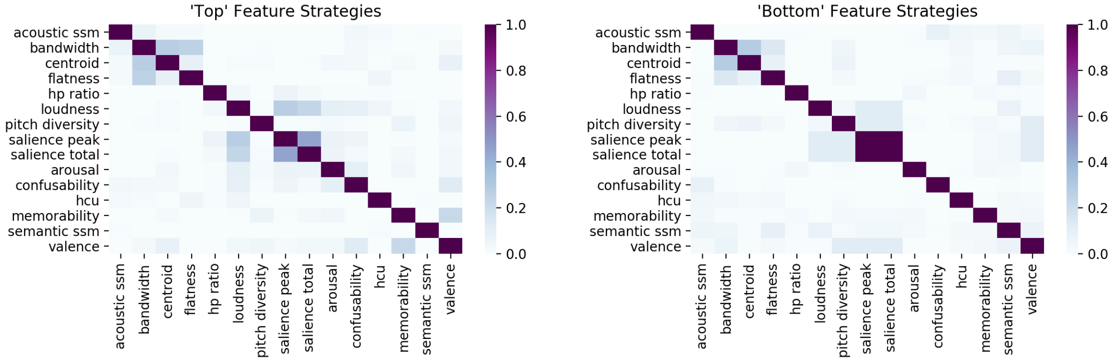


Figure 5.2: Percentage of overlap between the "Top" (left) and "Bottom" (right) 1st percentile-ranked excerpts from each participant's recordings, averaged across all participants.

From the heatmap, we see very little overlap across the two classes of features (high-level on the bottom right of the grid, and low-level on the top left)[3]. We do see evidence of intra-class overlap, such as between spectral bandwidth, centroid, and flatness, and between valence and memorability, which aligns with intuition. The results suggest that the introduction of gestalt principles to the task of audio summarization is a valuable contribution, and extends methods relying on spectral processing alone.

5.1.6 *Linking Feature Strategies to Perceptual Outcomes*

We next examine how our chosen feature strategies map to perceptual descriptors across participants. In Figure 5.3, we show the relative contributions of the 16 feature strategies towards a perceptual goal, given by the frequency that a particular descriptor was selected for a particular feature strategy. The darkened bars give the most significant drivers of a descriptor (if one exists), computed using a modified (mean-absolute-deviation) z-score [140]. We see that there are 16 descriptors for which these drivers exist (for instance, "calming", "familiar", "distracting", "summary of events"), and 4 for which they do not ("busy", "surprising", "stressful", "eerie"). Of the former 16, we find that for 12 descriptors, gestalt feature strategies are the top performers. Finally, we find that apart from the descriptor "calming", the baseline strategy of selecting clips without any feature analysis is

[3] Random overlaps have a very small likelihood, as all participants have upwards of 10,000 excerpts in their audio pool.

not a driver of any other descriptor.

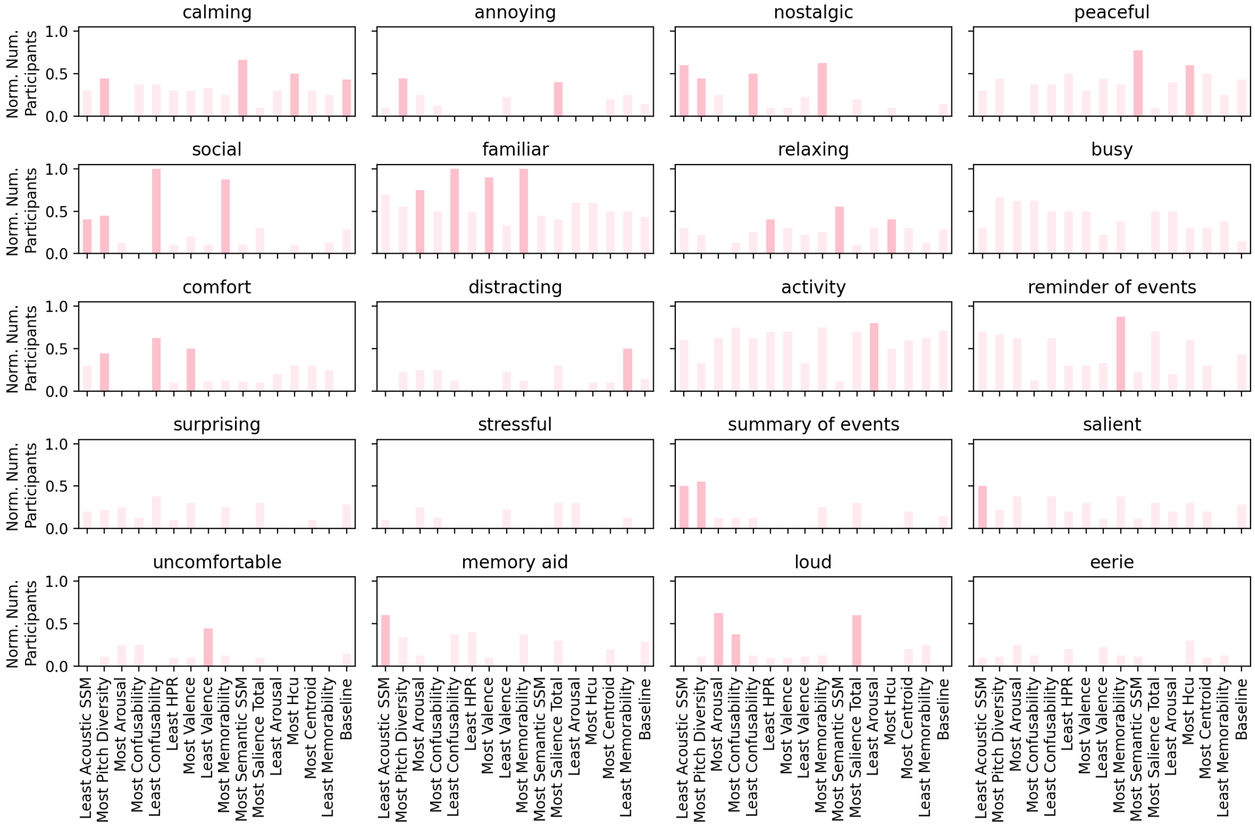


Figure 5.3: The relative contribution of each individual feature strategy in driving the selection of a descriptor, given by the likelihood that the descriptor was marked by a participant in association with summaries made with a feature strategy. Darkened bars highlight positive outliers identified by 2 median absolute deviations above the median (modified z-scores).

The results also highlight trends that hint at the complexity of emotional response in sound perception that this work begins to uncover. For instance, contrary to intuition, sound clips with significant pitch diversity are found to be "calming" or "comforting"; percussive sounds ("Least Harmonic Percussive Ratio") are found to be relaxing; the most "nostalgic" summaries are comprised of sound objects with labels that are easily identifiable and intrinsically memorable, *as well as* those that are acoustically unique and diverse in tonal content. We intuit that a personal relationship with the audio being summarized (reflecting back on first-person recordings) is a significant functional force in these relationships. We examine this further in Section 5.1.9.

Noting the overlap in top performing feature strategies across some descriptors, we perform a clustering analysis to identify orthogonal archetypes in the perception-feature strategy space. To do this, we first compute an affinity matrix between descriptors using the jaccard index of intersecting feature strategies as the affinity measure (considering only descriptors and feature strategies with significance given by their z-scores), and perform a simple agglomerative clustering. The results are given in Table 5.1, with the union of the feature strategies shown beside each descriptor cluster. These clusters suggest that we may be able to explicitly map the system's underlying feature space to different perceptual archetypes that hold across multiple individuals; more importantly, however, on the level of an individual listener, we suggest that these clusters are useful "initializations" for personalized summary generation, wherein the relative weights between feature strategies are refined based on user feedback (see Section 5.1.9).

	Descriptors	Feature Strategies
1	familiar, comfort	Least Confusability, Most Arousal, Most Memorability, Most Pitch Diversity, Most Valence
2	calming, peaceful, relaxing	Baseline, Least HPR, Most Hcu, Most Pitch Diversity, Most Semantic SSM
3	loud	Most Arousal, Most Confusability, Most Salience Total
4	nostalgic, social, summary of events	Least Acoustic SSM, Least Confusability, Most Memorability, Most Pitch Diversity
5	salient, memory aid	Least Acoustic SSM
6	uncomfortable	Least Valence
7	reminder of events	Most Memorability
8	distracting	Least Memorability
9	annoying	Most Pitch Diversity, Most Salience Total
10	activity	Least Arousal

Table 5.1: The results of an Agglomerative Clustering applied to the affinity matrix describing the intersection in dominant feature strategies between descriptors. Each cluster is described by the descriptor or new group of descriptors, and the union of the dominant feature strategies.

5.1.7 *Examining Intimacy, Sentiment, Emotionality*

We look next at the distribution of likert ratings provided by participants in response to each summary. Participants were asked to rate each summary on the scale of its (1) emotional intensity, (2) associated sentiment (positive/ negative), (3) intimacy and familiarity (see Appendix for full text), and the results are shown in Figure 5.4. We use a non-parametric Kruskal-Wallis test with post-hoc Dunn test comparisons to examine differences between pairs of feature strategies, and observe that the most significant ($p < 0.05$) drivers of emotional intensity, positive sentiment, and intimacy are summaries comprised of clips that score highest in memorability and valence, and lowest in confusability. We note that all three feature strategies are gestalt, and are based on scoring models bootstrapped from human annotations in the HCU400 dataset. We do also find that certain low-level feature strategies are high-performing relative to others in each assessment category – for instance, least acoustic self-similarity for emotionality ($p < 0.05$) and intimacy ($p < 0.05$), and most salience for intimacy ($p < 0.05$). We suggest that this hints at the ability of these feature extractors to capture human-meaningful information at scale that points to subjective, aesthetic experiences, rendering them useful for our summarization task.

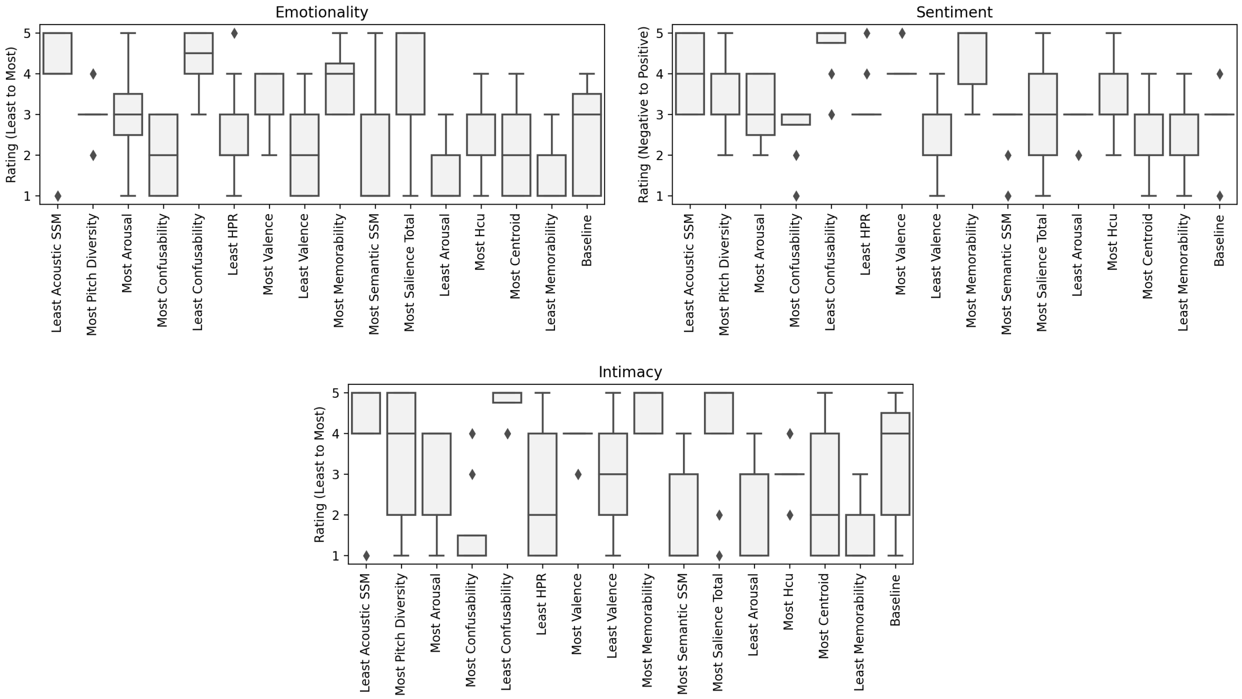


Figure 5.4: The likert response values to Q1, Q2, and Q4 (see Appendix) assigned by participants to summaries of each feature strategy.

5.1.8 *Anecdotal Responses to Listening Experiences*

Finally, in order to understand the affordances of our summary system that are more difficult to capture quantitatively, we examine the free text responses provided by participants at the end of their custom surveys. Participants were provided guiding questions (see Appendix), but were free to provide any comments or reflections on the listening experience that came to mind. Several participants chose to provide further commentary by video calls with the researchers regarding specific summaries in their surveys, which have been recorded and transcribed with the participant's consent.

Below, we highlight several themes from the responses and their supporting commentary. The commentary is marked with an anonymized participant ID, as well as the feature strategies associated with the summaries described by the participant, if applicable. As an overview, we uncover four major conclusions:

1. our system succeeds in creating *impactful experiences* over *informative summaries*, aligned with the motivation presented at the start of the work

2. participants are unanimous in their willingness to use the system,

and suggest a diversity of application contexts and use paradigms

3. the output of our system is found to be engaging, emotional, immersive, and intimate

4. participants offer feedback on production aspects, reinforcing some design choices and suggesting improvements in others

We note that this anecdotal evidence alone is not sufficient to disentangle the value of the system design choices (such as the size of the audio excerpts, the cross-fading heuristic, the use of the stereo field, etc) from the algorithmic choices (the incorporation of gestalt information in the feature space). We treat this commentary as positive reinforcement for *both* aspects of this work, and we intuit the value of the latter aspect especially when the commentary is coupled with the trends observed in the quantitative analysis in Sections 5.1.6 and 5.1.7.

Immersion and Intimacy
Several participants explicitly described a sense of presence in time or place when reviewing their audio summaries. For instance:

> *"[Listening to the summaries] was really engaging, I think the quality and the spatial nature of it make it feel really impactful. I actually think because the sounds seem to move around spatially a bit between some of the clips, it feels more like a fly on the wall observing from different perspectives, and that really makes it feel like you're peeking into something... there is definitely a feeling of being there in the space in that time."* - P1

We noted that participants experienced this sense of immersion from summaries curated to have human-oriented content (via gestalt feature strategies) as well as those that combined notable but inorganic content from a sound context:

> *"Ones that brought me back to a moment in time or had talking made me feel transported. The kitchen ones also felt like I was making food again."* - P4, **[Least Acoustic SSM, Most Salient, Least HPR]**

> *"One of my tracks seemed to highlight laughter in particular, and it was pretty neat. It's hard to capture memories of how or what makes people—or even if they laughed. Also hearing someone's laugh is like hearing them in a very innocent state that I found interesting."* - P5, **[Most Memorable, Most Valence]**

> *"..just hearing someone's voice made me reflexively smile, which was surprising. It was actually even quite immersive to hear strictly ambient noises*

> *that had been recorded, like keyboard typing and doors closing. "* - P5, **[Most Memorable, Least Confusable, Least Arousal, Most Salient]**

The participants also suggested that the compelling nature of the listening experience stemmed from a sense that the content was very personal; while this is due in part to the nature of the recording exercise (a wearable recorder that was capturing audio indiscriminately), participants highlight the role of the "sensitive editing", or the automated sound object analysis, extraction, and curation process:

> *"The more unusual immersion was in the intimate mixes, again stemming from the sensitive editing*[4] *and the pervasive recording that managed to capture scenes that are usually off-limits... I am not sure whether those examples would also feel immersive for other people, or if the sense of immersion comes from recognizing the strange little details of one's own life..."* - P9

Triggering Memories without Explicit Documentation

A common theme that emerged from the participant responses was the unprompted recollection of events and occurrences tied to the summary content:

> *"..this was an almost bizarre (in a good way) experience of feeling the connection between audio and my memory. Especially with people's voices and laughs. If you asked me what I did in the past few weeks before hearing this, I'm not sure I would have written down what I am now enjoying my memory of."* - P5, **[Most Memorability, Least Confusability]**

> *"The sound was often cut too short for me to recognize the situation, but at the same time I was surprised how much I remember from the snippet of sound of what happened in that situation... I find the auditory experience to be very stimulating and the fact that there are no images triggers memories in a much more emotional way for me, because it is vague and ambiguous."* - P7

> *"In the best examples, I was surprised by what felt like creative editing and curation: with almost comedic timing, tying domestic activities together with a well-cut sneeze or hoot. In those examples, I also found what felt like the closest representation of my sense of home life: intimate, detailed snippets that were just long enough to remind me of a bigger picture."* - P9

As suggested by the quotations, most participants were aware that they could not precisely pinpoint the events that were taking place or the actions that they were engaged in when the audio in the summaries was captured. Most, however, went on to suggest that the curation process was effective in transforming this notion into an emotional listening experience, eliciting feelings of nostalgia and positive sentiment associated with past memories.

[4] Participants were not informed of how the summaries were prepared.

New Perspectives on Everyday Life

Depending on the feature strategy used to curate the summary, participants highlighted the ability of the presentation to offer different perspectives on the same sound context, often allowing participants to reexamine sound elements they had not noticed when recording:

> *"I was surprised by the level of environmental noise in [the] recordings, combined together like in the recording they really make me feel stressed. It makes clear how my brain filters information about my environment and my experiences. When I hear the situations without the visuals it highlights different aspects that was not noticeable to me before."* - P7, **[Least Arousal, Most Semantic SSM, Least Memorable, Most Hcu]**

> *"It's just fascinating too to see what clips appear – some of them are really clearly tied to specific events (like a video I watched or a specific project I was working on), some evoke a common behavior I do (going for walks), some I couldn't even identify because they're just background sounds I filter out (or have headphones in)."* - P1, **[Least HPR, Most Semantic SSM, Most Salience, Least Arousal]**

> *"[I was surprised by..] the amount of laughter I encountered over the course of the study; how eerie my workday 'silence' sounded like; how my laugh is more high-pitched than I remember."* - P3, **[Most Memorable, Least Arousal, Most Pitch Diversity]**

Using my Summaries

Participants were nearly unanimous in their interest in using the summarization system again in the future. For instance, respondents said:

> *"[I] absolutely [would use these audio presentations]. I'm actually quite surprised at how interesting and fun and engaging listening to these clips was, and I think it captured really interesting tidbits of life in a really unique way."* - P1

> *"I'd probably use it every few days. The information content was low though.. But I loved hearing a quasi-musical or sound-art gist of my days.. I would use this more as an experience rather than anything to derive information from - but it was indeed a fun experience, and if I listen to this a year from now, the nostalgia quotient will probably be quite a bit higher!"* - P10

Users also detailed specific use cases and application scenarios to which they would want to see the system applied, with a focus on reflection, wellbeing, and mindfulness practice:

> *"Especially would use it to capture a certain extended experience, like living abroad, an internship, or a summer with family."* - P3

> *"I would use the audio presentations like a diary to reflect back on things further in the past than a couple weeks, and I would use it as a daily diary rather than weekly or longer. It would be interesting to use audio presentations for funeral or wedding services, very intimate."* - P4

> *"I'd be very interested to use this tool as 1) a kind of gratitude journal that helps me keep my family and friends in mind, I am also interested in daily or weekly intervals. I do a daily mindfulness meditation practice... I see parallels and think that this could be a very valuable companion exercise to that practice."* - P5

Finally, several users alluded to the fact that given the intimate and personal nature of the summaries, the mechanics of the summary production that enable privacy – selecting short, individually unrecognizable sound snippets that are combined by feature strategy – are of paramount importance to the adaptation of such a system.

> *"If I know that the recordings never leave my recorder/phone and are processed into unrecognizable bits (like the ones I head [sic] in the study) in real time, would make me much more likely use this technology."* - P7

> *"So perhaps intentional, intermittent use of a system like this would be good: a couple of weeks a year just to have a taste of what each year felt like, not from the perspective of the public, curated social feed but from the inside.. a system for documenting my life that is only for ears and no one else's."* - P9

Positives and Limitations of Summary Production

Though not explicitly prompted to do so, users provided feedback regarding the production methodology and audio rendering infrastructure that is used to support the algorithmic choices in the system. Several praised aspects of the production pipeline:

> *"I really liked how the audio was spatialized. That added information about the events that made the summaries tailored to how I experienced them. For example, audio summaries of the workday (a lot of typing) had typing sounds in different parts of space, because of my changing positions relative to my keyboard, which made the end result more artistic"* - P3

> *"[I was surprised by] how well some of the sounds were blended together or laid over each other that sounded natural but I know they didn't occur simultaneously like that."* - P4

> *"The way the sound clips flowed together was surprisingly good - almost like a piece of composed sound-art music.. The stereo field was used well, and some of the clips really did take me on a mini-voyage through the time I had the gear."* - P10

Others expressed preferences for production parameters that would have improved their sense of engagement:

> *"If the clips were longer I think I would have found it a bit more immersive... Some way to unintentionally record would be better for me because the results are surprising and fun."* - P7

In the future, the feedback associated with this theme can be used to explore other options for assembling summaries based on cognitively analyzed and curated sound content. While we choose a simple approach in this work and keep it consistent across users, more complex production techniques (for instance, concatenative synthesis [141]) could enhance the listening experience independently of feature strategy.

5.1.9 Discussion

Thus far, we have described a new auditory interface – a tool that can be used to mediate our interaction with the sounds that we capture around us – built with a feature space that augments traditional spectral measures with "gestalt" measures inspired by ideas in auditory cognition. In quantitative tests of user perceptions, we show that these features dominate over or combine with spectral measures to elicit specific aesthetic responses. And in an exploration of qualitative feedback, we demonstrate that these features result in a system that create compelling, moving listening experiences.

Here we address the value and limitations of this work, and suggest several important contributions and open research problems that it offers to the HCI community.

Cognitive Features as Inputs to Auditory User Interfaces Cognitively-inspired "gestalt" feature extractors have the potential to be powerful, as observed in the presented quantitative and qualitative results. In the most reliable mappings between feature strategy and perceptual goal from our study, gestalt feature extractors play the dominant role; our analysis also shows that they surface a statistically different set of audio samples than the acoustic feature extractors alone, implying that the relationship between acoustic and gestalt information is non-trivial and demands greater complexity in modeling. It is important to note that these gestalt extractors cannot themselves be considered predictive models of perceptual response – a sound with a high "Memorability" score does not automatically mean it will be readily remembered by a certain listener! – but instead that they capture collective conceptual information which can be exploited to construct such relationships in the context of an application, as was done in this work.

Towards Personalized Audio Summaries Despite the broad brushstroke conclusions that can be drawn from the aggregate participant response, there is still significant variance in the reported summary perception that is a function of an individual's sonic diversity and relationship to his or her sonic environment under personal context. We suggest using the clusters mapped out in Section 5.1.6 as "priors" – or *a priori* information that forms a coarse model – between feature strategies and perceptual archetypes, and then considering a closed-loop system that incorporates user feedback to refine feature weights towards specific preferences over extended periods of time. We return to this idea in Chapter 6.

Limitations: Audio Classification Networks While neural networks can bootstrap the translation of gestalt listening principles to system design, there are limitations with current state-of-the-art models. Supervised learning strategies designed to map independent sound events to labels – as the AudioSet model does – do not generalize well to dynamic, real-world sonic contexts, where sound events frequently overlap and vary in signal-to-noise ratio. Classification taxonomies are also often limited. Despite AudioSet's notably large ontology, information that can be extracted from the labels is drastically limited when compared to free-text human annotation. As the research advances in favor of more naturalistic datasets and unsupervised learning strategies, we can advance the capabilities of this summarization system and other similar interfaces.

Limitations: User Study The user study methodology chosen for this work allowed for a largely realistic evaluation of the system, but presented certain practical challenges for the participants: discomfort in wearing the recording devices for several hours a day, limited battery life and storage space, privacy concerns, and constant awareness of the device's presence all limited the extent to which a natural, unadulterated sonic environment could be captured. In forthcoming studies, we will consider ways to improve physical properties of the device (i.e., choosing a device with a longer battery life and memory), and attempt to better understand the role of the behavioral factors in participant data.

6

Personalizing from Gestalt Models

In the two previous chapters, we discussed constructing gestalt models that reflect crowd-average ideas about semantics in sound, and presented examples of how these models could be used as they were to construct applications that have the potential to modulate experience in sound at the level of the individual. However, in this chapter, we return to a question left unanswered in Chapter 4. To illustrate it better, let's consider a simplistic example – our gestalt models suggest that an audio clip of a body of water (sounds of a lake, waves at the beach, a waterfall, etc) should have a high positive valence. Having a model that assigns the clip this annotation is powerful, because it is an intrinsic measure that can't readily be estimated from acoustic quantities like pitch or amplitude. However, to go as far as saying that a listener who consumes this clip will experience a sense of calm or positivity after sampling this recording is a poor inference. What if this listener has had a traumatic experience with large bodies of water in the past, and feels frightened upon listening to any related sound? Or perhaps the listener enjoys spending time near bodies of water, and but finds the sound distracting as a result, and chooses mostly to avoid water-centric soundscapes for his or her daily meditation practice. Given this subjectivity and diversity, is it at all possible to bridge the gap between our gestalt feature representations and an individual's interpretation of those representations in the context of a specific application, especially when that individual interpretation is contingent on so much that is immeasurable – such as one's life experience, exposure, sensitivity to and training in sound, and physiology?

In this chapter, we draw on the tools that stemmed from the work in Chapter 3 to explore this idea – namely, we suggest that we can treat gestalt annotation models as a useful prior for personalized applications, particularly because the models are constructed from data that represents the "average" human. These priors can then be adapted into personalized posteriors by sparse, longitudinal, noisy reinforcement observations applied to a probabilistic modeling framework suitable for this type of data. The powerful idea behind this approach is that we use a probabilistic framework to treat "noise" in observations of user preference as a single entity, allowing the system to remain agnostic to the source of the noise and its likely association with multiple factors that we can't realistically account for – such as a person's background, an uncontrolled, naturalistic environment that creates a diversity of biases in cognitive state, or the lack of sensory measures that provide insight into cognitive state to supplement self-reported measures.

To explore and demonstrate this idea, we return to the audio summarization system in Chapter 5, and adapt it specifically to create a system that condenses pre-selected ambient audio recordings into stochastic soundscapes that facilitate a sense of focus and productivity. In this system, we use the learnings from Chapter 5 as guiderails, by choosing a smaller set of feature strategies that we expect will point to the appropriate cognitive outcomes; we then layer a reinforcement learning model on top of these "priors" to attempt to optimize their relative weights to an individual's preferences with very little data and in as natural a setting as possible. In discussing the design of this system and an analysis of the results from a deployment study with N=25 individuals, we are most interested in the following research questions:

- What is an appropriate modeling strategy to handle personalization with regards to this application, where the system must handle noisy, realistic observations stemming from self-report? We use the term "noise" to collectively refer to two phenomena likely to occur over one participant's use of the system – changing baseline cognitive state (and in particular, cognitive load) due to changes in the environment, such as the work one is undertaking or levels of activity in the surrounding space; and changes in interpretation of the audio or the impact it has on cognitive state as a function of stochasticity in the summary assembly process.

- What metrics do we use to understand whether we have truly found an individual's region of preference in the optimization space? How can we combine observations from a test setting with measures of confidence derived from the training process?

- If we use this metric to examine the individuals we are able to personalize separately from those we are not able to personalize, in what ways do the training dynamics and engagement behaviors appear to be different? What can these differences tell us about whether and why personalization is necessary, and the ways in which a system built under the proposed modeling framework might succeed or fail?

6.1 Related Work

Personalization applied to tasks and interventions in the study of HCI is not a new area of research, broadly speaking – we can think of recommendation engines, location-specific services, customized educational technology, and automated adaptations of interaction modalities and user interfaces as illustrations of this intersection [142]. Depending on the context, applications in this space are often powered by a spread of machine learning paradigms – examples include domain adaptation (such as transfer learning or few shot learning with neural networks) [117], bayesian inference [143], traditional or deep reinforcement learning [144], multi-task learning with neural networks [145], or active learning conditioned explicitly on demographic information [146].

While still few in number relative to other application spaces, there an increasing number of applications employing personalization to tasks and interventions in sound perception. For instance, [147] and [148] apply a simple best fit mechanism and a gaussian optimization process respectively to adapt to a user's perception preferences for an EQ tool; [149] demonstrates a gaussian process regression and active learning system to optimize hearing aid parameters for individuals with different assistive needs; and [150] mentions several approaches using neural networks to individualize HRTFs by adapting generic HRTFs under perceptual feedback.

The novelty of this work lies not only in applying a machine learning-driven personalization strategy to previously unexplored terrain in audio perception, but more importantly in demonstrating a personalization paradigm in this field that exhibits a fine-tuning procedure layered on top of prior intuition. Unlike the approaches common in the deep learning literature, we undertake this fine-tuning in a probabilistic manner using a reinforcement learning strategy to account for real-world observation constraints. We suggest that the approach allows us to abstract away some of the sources of "noise" that are not a function of individual preference, and provide an accompanying analysis that illustrates how failure modes can still form if the noise is sizeable and persistent.

6.2 Methods

6.2.1 System Overview

The goal of our system is to create a soundscape from an extended ambient recording that an individual feels is suitable for facilitating a state of focus and productivity – a cognitive state that we choose by drawing from the HCI literature on synthetic atmospheres [151] – by optimizing over the relative weighting of a subset of summarization feature strategies presented in Section 5.1. The subset of features is chosen from the original set of 16 based on the feature strategies that were found to be the most significant drivers of the "familiar, comfort" and "calming, peaceful, and relaxing" clusters (see Table 5.1). To further restrict the optimization space, the union of these feature strategies was whittled down to a set of four strategies per ambient recording, by choosing the most divergent four strategies after a content overlap analysis mimicking Section 5.1.5.

Interaction with the system consists of two phases: in the "training" phase, participants are asked to engage in a task that demands a state of deep focus – such as working on an assignment, writing code, reading a paper, or working at the lab bench – for several minutes while listening to the automatically generated soundscape. Any time and any number of repetitions after the prescribed duration has passed, participants are required to provide a rating along a 5-point semantic differential scale regarding the suitability of the soundtrack for their desired mental state. The system uses this rating to generate another soundscape in a subsequent trial, and the process is repeated for a fixed number of trials. In the testing phase, participants are asked to repeat the process of listening and evaluating, but are informed that all subsequent trials are the system's best guesses as to their preferences and that they are now evaluating the learning outcomes.

6.2.2 Model

As the preference model underlying the interface, let us consider an optimization space that has F features. For each feature, a corresponding weight λ can be chosen from a scale of 0 to 1, and the weights must sum to 1 across features. Therefore, the grid that we use as a visual representation of the optimization space has as many dimensions as degrees-of-freedom, F - 1, and is only valid in the region where the weights are convex. A single point in the space can be represented as:

$$\lambda_p = (\lambda_0, ..., \lambda_{F-1}),\ s.t. \sum_{f \in F} \lambda_f = 1 \tag{6.1}$$

A typical optimization problem would be framed as

$$O := \underset{\lambda_p, p \in P}{\operatorname{argmin}} \mathcal{H}(\lambda_p) \tag{6.2}$$

where P represents the set of all of the valid points in the optimization space, and $\mathcal{H}$ is a function representing a human's preference. In this framing, the human response is perfectly deterministic, and any number of numerical optimization methods could be used to locate the optimal λ_p.

More realistically, however, we wish to construct a framework under the assumption that $\mathcal{H}$ is probabilistic and reflects an underlying preference likelihood, and that we build confidence that likelihood by probing a λ_p repeatedly and inferring information about it from neighboring points. In this vein, there are many traditional probabilistic optimization approaches that one may consider – examples include gaussian process optimization and bayesian optimization [152], markov decision processes [153], etc.

In this work, we diverge slightly from the optimization literature and frame the problem as a reinforcement learning problem. Namely, we consider a Multi-arm Bandit (MAB) problem solved with Thompson sampling [154], and adapt it slightly to the context of the system. We choose this framework over the aforementioned strategies for the following reasons: (1) given a limited number of user trials and perceptual equivalence of the system output for very small perturbations of a given λ_p, we wish to discretize the optimization space, and make assumptions about the spatial independence of λ_p's as a function of the euclidean distances between them; (2) for a given λ_p, we wish to assume an underlying expected probability distribution over preference outcomes (such as "preferred" or "not preferred"), rather than a ground truth outcome with noisy measurements, which is the assumption made in stochastic numerical optimization approaches [155]; and (3) we want to be able to easily incorporate tunable heuristics into the model about spatial relevance and the tradeoff between exploration and exploitation as a function of the number of user trials.

In our application of MAB to this system, we first discretize each dimension of the optimization grid into D weights. Then, each possible grid point is treated as a bernoulli bandit arm, which when evaluated by a human listener ($\mathcal{H}$) after listening to the corresponding soundscape, returns a reward $r \in \{0, 1\}$, indicating whether the soundscape is preferred or not preferred. As in the Thompson sampling approach, each point is represented by a beta distribution, parametrized by pseudocounts α and β. The probability density function (PDF) is given below:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{6.3}$$

where B is the Beta function. The PDF represents the likelihood of the mean expected reward, $\theta_{\lambda_p} = \frac{\alpha_{\lambda_p}}{\alpha_{\lambda_p}+\beta_{\lambda_p}}$, as a function of the ratio of observed positive/ negative rewards and the total number of observations at that point. In our setup, the objective is first to maximize $\sum_{t=1}^{T} r_t$, the cumulative reward obtained over the total number of user trials T, and then to select λ_p that maximizes θ over $p \in P$ as the region of greatest preference for further, fine-grained evaluation.

Our basic algorithm for sampling from the optimization space to present soundscapes to the user and update the probability estimates based on the user-supplied reward observation roughly follows [154], as below:

Algorithm 1: Sampling and updating Bernoulli Bandit with Thompson Sampling

```
for $t \in T$ do
    # sample model
    for $p \in P$ do
        sample $\hat{\theta}_{\lambda_p} \sim f(\alpha_{\lambda_p}, \beta_{\lambda_p})$
    end for

    # select and apply action
    $\lambda_p^t \leftarrow \text{argmax}_{\lambda_p} \hat{\theta}_{\lambda_p}$
    Apply $\lambda_p^t$ to $\mathcal{H}$ and observe $r_t$

    # update distribution
    $(\alpha_{\lambda_p^t}, \beta_{\lambda_p^t}) \leftarrow (\alpha_{\lambda_p^t} + r_t, \beta_{\lambda_p^t} + 1 - r_t)$
end for
```

However, we determine the need to make minor adjustments to this algorithm to account for constraints and requirements stemming from pilot experiments – for instance, we would like to restrict the number of user trials so as to bound the experimental duration to 1-2 hours, to minimize user fatigue; and users preferred to respond with a graded scale as opposed to in a binary format, to better express nuances in their assessment of a soundscape. As a result, we

(1) include a spatial reward update kernel, allowing us to use reward information about a specific point in the grid to make weak inferences about neighbors; and (2) replace the binary response options on the user interface ("I Prefer this Soundscape" vs. "I Don't Prefer this Soundscape") with a 5-point semantic differential scale (from "Strongly Not Preferred" to "Strongly Preferred"), translated to the MAB model as weighted updates to the grid cell psedo-counts. The revised algorithm is as below:

Algorithm 2: Adapted Bernoulli Bandit with Thompson Sampling

```
for $t \in T$ do
    # sample model
    for $p \in P$ do
        sample $\hat{\theta}_{\lambda_p} \sim f(\alpha_{\lambda_p}, \beta_{\lambda_p})$
    end for

    # select and apply action
    $\lambda_p^t \leftarrow \text{argmax}_{\lambda_p} \hat{\theta}_{\lambda_p}$
    Apply $\lambda_p^t$ to $\mathcal{H}$ and observe $r_t$, where $r_t \in \{0, w_0, w_1\}$

    # update distribution
    if likert rating is centered or positive: then
        $(\alpha_{\lambda_p^t}, \beta_{\lambda_p^t}) \leftarrow (\alpha_{\lambda_p^t} + r_t, \beta_{\lambda_p^t})$
        for $n \in N$, where $N$ is the set of neighbors of $p$ do
            $(\alpha_{\lambda_n^t}, \beta_{\lambda_n^t}) \leftarrow (\alpha_{\lambda_n^t} + \gamma r_t, \beta_{\lambda_n^t})$, where $\gamma \in [0, 1]$
        end for
    else
        $(\alpha_{\lambda_p^t}, \beta_{\lambda_p^t}) \leftarrow (\alpha_{\lambda_p^t}, \beta_{\lambda_p^t} + r_t)$
        for $n \in N$, where $N$ is the set of neighbors of $p$ do
            $(\alpha_{\lambda_n^t}, \beta_{\lambda_n^t}) \leftarrow (\alpha_{\lambda_n^t}, \beta_{\lambda_n^t} + \gamma r_t)$ where $\gamma \in [0, 1]$
        end for
    end if
end for
```

This algorithm is applied over T trials, which forms the training phase. In the subsequent test phase, we choose the best performing region from the training phase by selecting $argmax_{\lambda_p} \theta_{\lambda_p}$, and probe it for an observed reward on the semantic differential scale as in the training phase. We additionally probe k random neighbors of this chosen λ_p, which form the set of locations N, in the same manner.

At this stage, it is imperative that we discuss the several assumptions we make regarding the model and the interactivity paradigms; we believe that making these assumptions are useful for the task at hand, but important to consider when examining experimental outcomes:

- We assume that the underlying θ is fixed per λ_p, and therefore that varying reward observations across multiple probes of the same λ_p reflects noise that emanates from changes in the user's cognitive state, not fundamentally shifting preferences.

- Though the user interface contains a graded scale, we model preference as a binary variable, with an expectation over preferring or not preferring. We implement this as a more robust approach in the face of a limited number of trials T.

- We assume that users will treat the semantic differential scale options as spaced equidistantly, though this may not be the case in practice.

- We assume that spatial dependence occurs at small euclidean distances, and assume spatial independence at large euclidean distances.

6.2.3 Experiment

To examine the research questions outlined at the start of the chapter, we construct an evaluation system by applying the model described above to the audio summarization system, as described in Section 6.2.1. Table 6.1 gives the values used for the tunable parameters in the model.

Model Params	**Experimental Value**
D	4
T	15
w_0	2
w_1	4
γ	0.5
N	radius=1
k	2

Table 6.1: Values used for model parameters in the experimental setting.

We choose three audio contexts – a city street, a forest, and a university atrium – and obtain publicly available audio tracks corresponding to each of these locations[1]. Each track is approximately 1-2 hours in duration, recorded in or converted to stereo format, and is downsampled to 44100 Hz for the purpose of the evaluation. Table 6.2 gives the final list of feature strategies that were chosen per context based on the overlap analysis.

A demonstration of the experiment can be found at the accompanying website[2]. We recruit N=25 participants from the science crowdsourcing platform Prolific[3], consisting of an approximately 50:50

[1] https://freesound.org/people/shuraifa/sounds/412817/; https://freesound.org/people/klankbeeld/sounds/506678/; https://freesound.org/people/JonnyThePonny/sounds/232602/

[2] http://audio-mafia.media.mit.edu/summary-study

[3] http://www.prolific.co

Context	Selected Feature Strategies
City	most semantic self-similarity, most arousal, least confusability, most valence
Forest	most arousal, most pitch diversity, most valence, most semantic self-similarity
Atrium	least acoustic self-similarity, most arousal, least confusability, most semantic self-similarity

Table 6.2: Chosen feature strategies per audio context in the experiment.

male to female ratio, with participants spanning the ages of 18 to 47. Though all were required to be fluent in the English language, participants identified as being from at least 9 different countries. With T=15 trials, each experiment takes approximately 1.5 hours to complete, for which we compensate participants approximately $9.

6.3 *Results and Discussion*

In the following section, we present and discuss the implications of several trends that emerge from the experimental data. Unlike a traditional analysis of an experiment that might accompany a novel HCI intervention, this section will not attempt to argue that the system is "successful" in achieving personalization for a statistically significant majority of the user study participants, nor will it provide a generalized analysis of the outcomes as they pertain to the application itself (such as noting that more people who listened to the "forest" soundscapes preferred content with higher "arousal" scores)[4]. Instead, we focus this analysis on attempting to answer the research questions outlined at the start of the chapter, by first determining an appropriate metric to quantify personalization, and then uncovering the participant behaviors and biases that appear to separate out the participants whom we can personalize from those that we can't. We believe the insight is valuable not only in the context of building custom soundscape engines for targeted cognitive states, but more broadly to inform the design of systems that must tackle the complex challenges of user preference adaptation – we present examples of such systems following the results.

In Figure 6.1, we show the distribution of participant test point scores, normalized linearly by the range of the participant's training rating to account for diversity in usage of the response scale. In the case of a traditional intervention, we would consider the training data as the input to the model and the testing data as the output; therefore, we might draw a threshold along the score axis – perhaps at 0.75, to select responses that map loosely to the "Prefer" option

[4] This analysis, however, would be foundational work for smart environment and sound perception research – a great starting point for future graduate thesis work.

on the scale – and suggest that all participants that fall above this threshold should be considered personalized.

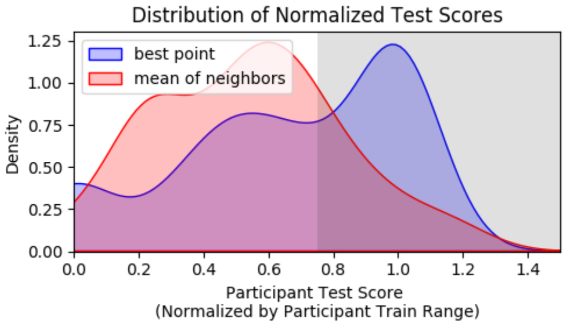


Figure 6.1: Distribution of normalized user test point ratings. The blue curve gives only the ratings in the region with the greatest θ; the red curve considers a mean of this point and the neighbors probed in the test phase.

However, in light of the challenges motivating the chosen modeling framework, we determine that it is not reasonable to take the test point scores at face value, as there is little reason to believe that the test phase evaluations are more resistant to noise than the training phase evaluations. Therefore, we require a measure of confidence to accompany the test scores that can be derived from the intermediate modeling process. In Figure 6.2, we plot the normalized test scores against the expected reward, θ, associated with the test point location. The participants for whom we believe that we have truly identified a region of preference – and for whom we have confidence in this finding – are located towards the top right of the plot, associated with a greater expected reward and normalized test rating at the primary test region. Any heuristic can be chosen to divide the respondents in this metric space into two groups, one representing those for whom we have achieved personalization, and one representing those for whom we have not. For further analysis, we intuitively select a region greater than 0.75 in expected reward and normalized test scores, as this value would map to the "Prefer" mark on the likert scale for most participants[5].

We next examine the behaviors of the personalized group in more detail. Figure 6.3, 6.4, and 6.5 give a visualization of the outcomes of the reinforcement learning process for select participants in the personalized group. Each point on the grid, representing a point λ_p, is shown with a colored circle, representing the final expected reward θ at that point. The region with the greatest θ is additionally marked with a triangle, representing the normalized likert rating assigned to that location by the participant in the testing phase. In sample grids corresponding to the forest context, participants' preferred weighting locations differ, but are roughly constrained to a small region, despite different rating behaviors as a function of the optimization

[5] This, of course, depends on a participant's use of the scale; we can account for conservative responding behavior by normalizing test scores, but the expected reward does not have a direct mapping to rating behavior – it is simply a confidence heuristic.

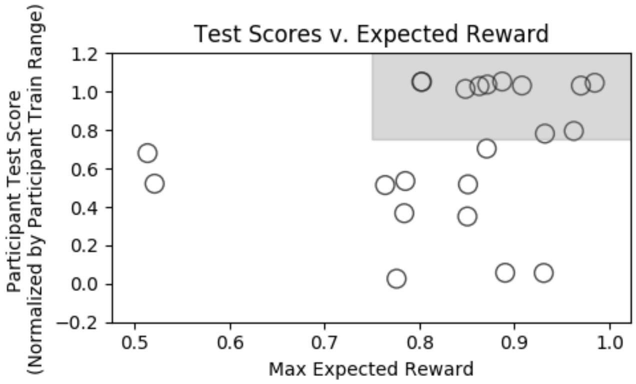


Figure 6.2: Normalized test scores in the primary test region plotted against corresponding θ values. The shaded region indicates the participants selected by our heuristic for personalization.

space; however, in the atrium and city contexts, preferred λ_p's are diverse and spread apart in euclidean space. These plots collectively indicate the *need* for and *value* of personalization, contrary to intuition – even within a context, and for some contexts more than others, participants who demonstrate a reliable preference can display very different preferences. At the accompanying repository[6], we provide audio examples of summaries associated with the most preferred grid locations associated with different contexts and different participants in the personalized group.

We then further investigate the differentiating attributes between the personalized and non-personalized participant groups, focusing on training dynamics. We first note an overall context bias, as shown by Figure 6.6 – independent of the personalization outcomes, the forest context is generally more suitable to the task of focus and productivity when compared to the city and atrium contexts. This is reflected in the across-participant distribution of raw training ratings (Figure 6.6), which shows a skew towards "Preferred" and "Strongly Preferred" in the forest context relative to the other two contexts.

Despite this bias, there are two behaviors that hold across contexts and participants that distinguish the personalized/ non-personalized groups, which hold implications for understanding the circumstances under which in-the-wild systems can adapt to user preferences. The first is shown in Figure 6.7, where we see that across all grid locations presented to a participant – not just the location with the greatest expected reward – participants in the personalized group have much higher test-to-re-test consistency. This suggests that, in the context of this task and the activity they chose to complete while listening to the soundscapes, these participants were more likely to be in a consistent cognitive state, or interpret the changes in the audio content in a consistent manner, or both. We can go on to weakly infer that

[6] http://audio-mafia.media.mit.edu/MAB-personalized-soundscapes/

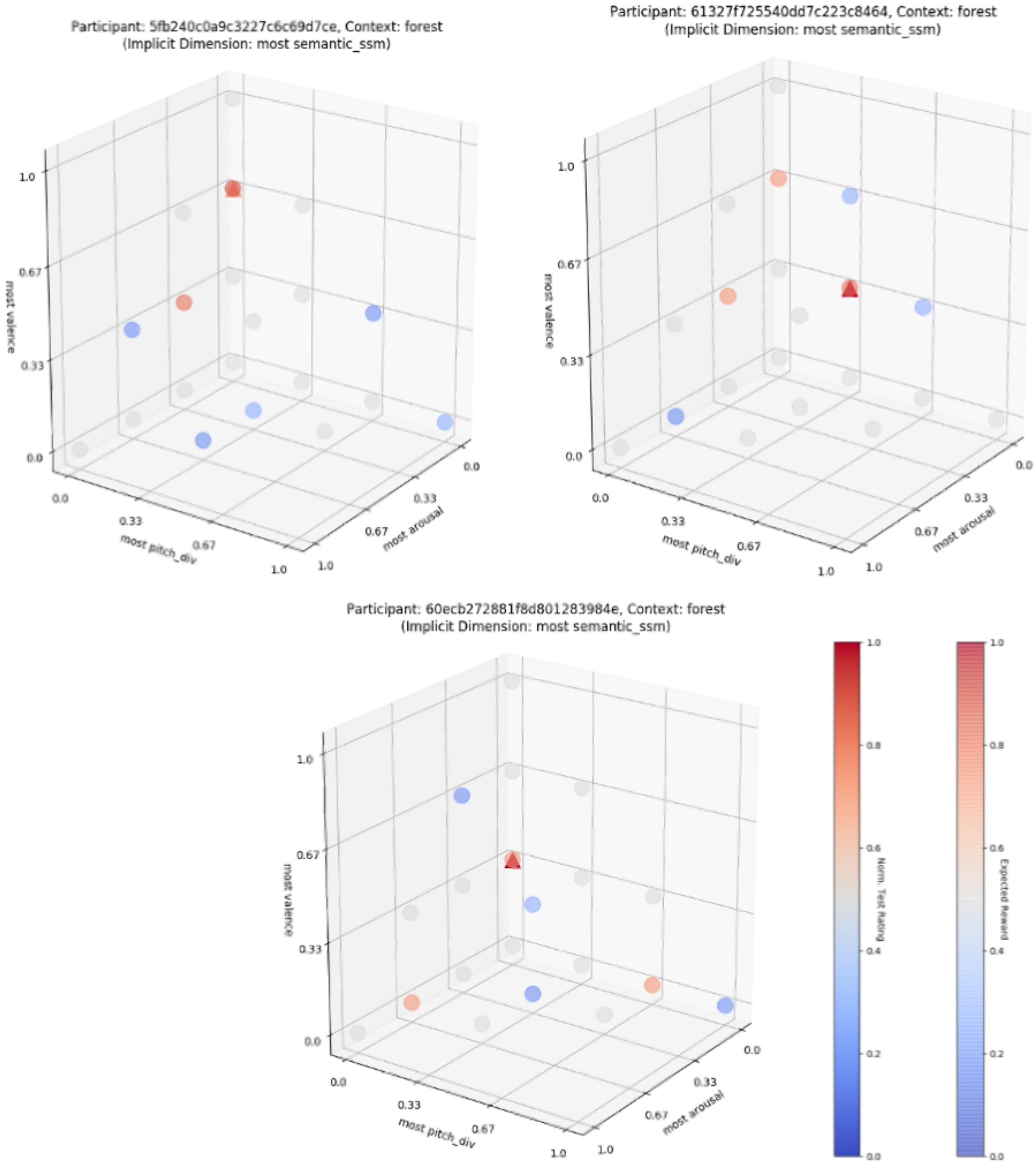
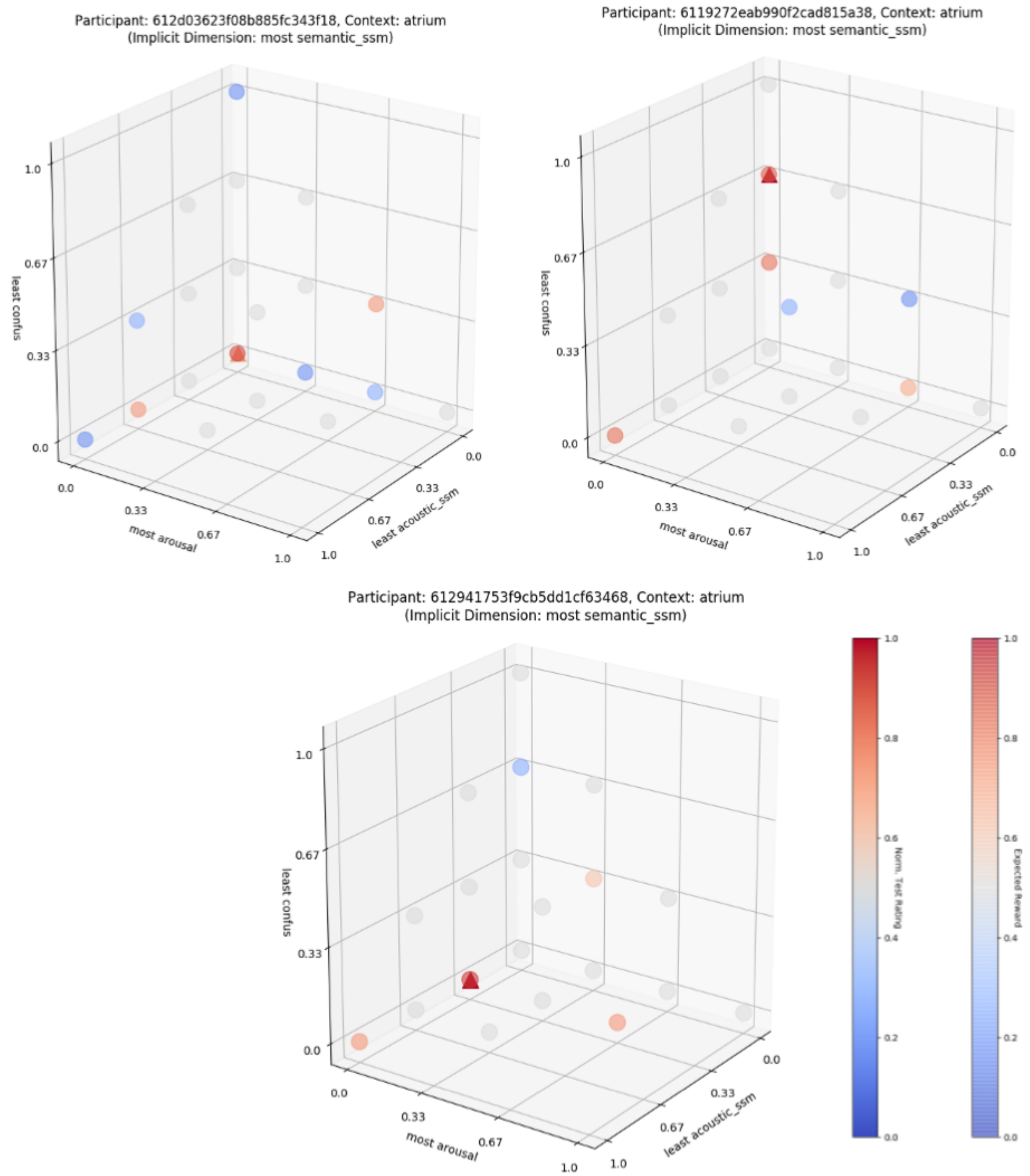


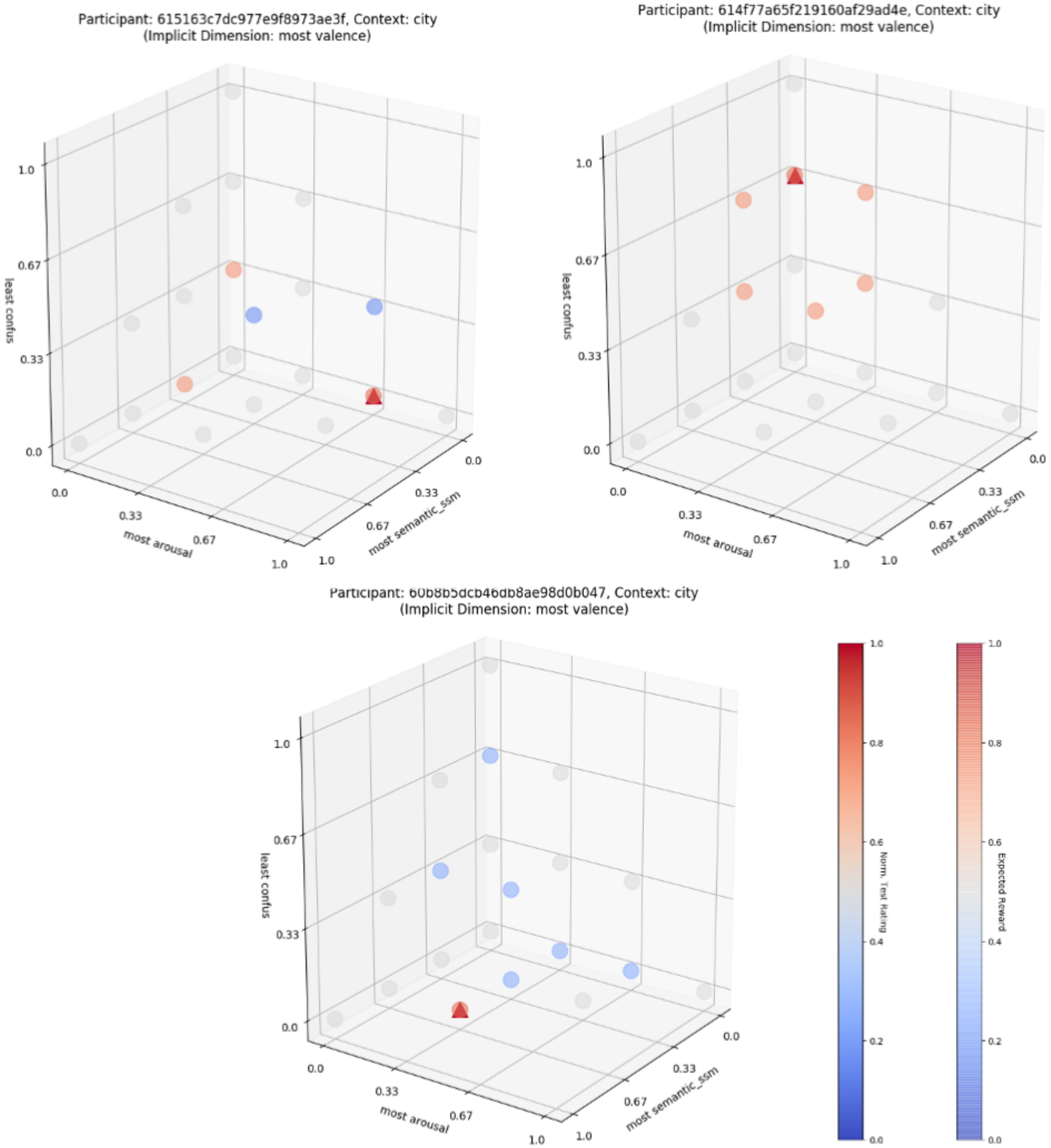
Figure 6.3: Visualization of sample participant rating behaviors, in the "Forest" context.


Participant: 612d03623f08b885fc343f18, Context: atrium
(Implicit Dimension: most semantic_ssm)

Participant: 6119272eab990f2cad815a38, Context: atrium
(Implicit Dimension: most semantic_ssm)

Participant: 612941753f9cb5dd1cf63468, Context: atrium
(Implicit Dimension: most semantic_ssm)

Figure 6.4: Visualization of sample participant rating behaviors, in the "Atrium" context.



Participant: 615163c7dc977e9f8973ae3f, Context: city
(Implicit Dimension: most valence)

Participant: 614f77a65f219160af29ad4e, Context: city
(Implicit Dimension: most valence)

Participant: 60b8b5dcb46db8ae98d0b047, Context: city
(Implicit Dimension: most valence)

Figure 6.5: Visualization of sample participant rating behaviors, in the "City" context.

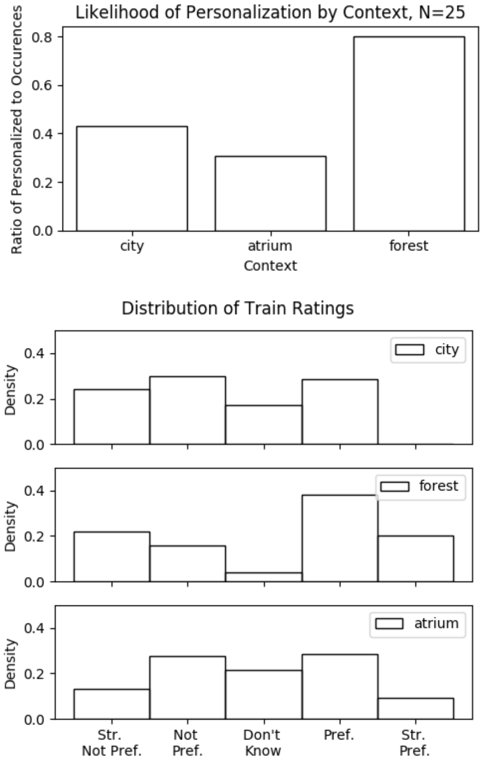


Figure 6.6: Plots describing the baseline bias induced by the context on personalization.

the presence of these behaviors in participants, whether tied to the individual or the environment, increase the likelihood of our system being able to identify a region of preference in the weighting space.

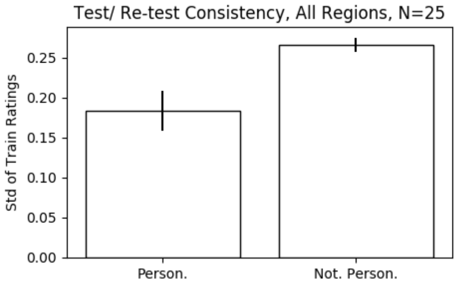


Figure 6.7: A measure of test/ re-test consistency across locations aggregated over participants, reported as the spread of train ratings. The smaller bar indicates higher consistency.

In Figure 6.8, we also see that the responses in the personalized group (given by the circles in black) cluster towards having the most diverse train ratings and expected reward values across the trials they are presented with, as measured by the standard deviation of both quantities. This suggests that the participants we are able to personalize are able to provide polarized opinions, identifying not only regions of preference with high consistency in their ratings, but also regions of strong *dislike*. We can weakly infer that these participants are likely to interpret the task at higher resolution, and are able to translate to their responses subtle changes in their perception of the audio, changes in their perception of cognitive state, or both.

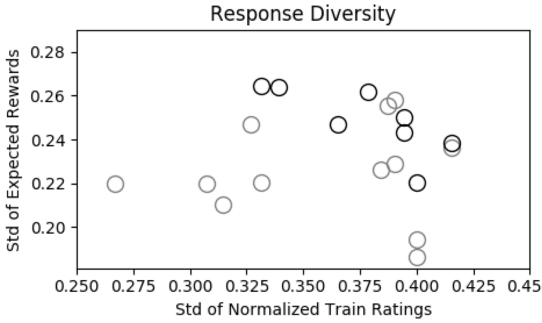


Figure 6.8: A measure of response diversity, reported by the spread of train ratings and expected rewards across grid locations, per participant. Personalized participants are given in black, and non-personalized are shown in gray.

6.4 *Conclusion*

In this chapter, we present a personalization framework that allows us to operate in the face of small quantities of data, while acquiring this data through a single, noisy response modality; as a result, we make strong assumptions about the structure of the beliefs underlying the model, intuit the nature of the noise accompanying the response channel, design an appropriately constrained optimization space, and implement a probabilistic model that allows us to robustly identify local optima. The insights from the analysis provide concrete information, specific to this intervention, about the kinds of user behaviors that allow for personalization and about those that pose a challenge; subsequent iterations of this and similar systems can seek to explicitly minimize these sources of noise. For instance, a pre-training phase can be included to improve response resolution and graded scale interpretation; or means to capture other changes in the environment that influence cognitive state of sound interpretation – like a question field that asks, "How loud is the environment you're working in?" or "What task are you trying to focus on?" – can explicitly be factored into the label confidence heuristic.

The current study does exhibit several limitations and scope for future work. For instance, we employed only a small number of participants each completing only a few trials due to time and cost constraints, which could be extended to develop a better understanding of the modeling strategy. We also chose, in the interest of practicality and realism, self-report as the sole user feedback modality for personalization – considering signals from an array of biophysical sensors could provide additional channels of information to the personalized model, contingent on having the ability to study those modalities first, per individual, in an isolated environment, so as to obtain suitable priors before personalization. If these extensions are explored further, the design choices and validation methods used here could extend to the individualization of several other audio perception tasks; for instance, calibration experiments for perceptually-aligned spatial rendering on augmented and virtual reality platforms; customized mappings for sonification engines (or sound modulation engines, like the SoundSignaling system) that are tailored to individual preferences over time via sparse, longitudinal reinforcement; or the optimization of audio infrastructure attributes (speaker placement, EQ settings, etc) for home media consumption given an individual's psychoacoustic profile and content preference.

7

Towards Gestalt Computation

In this dissertation, we introduce tools for the construction of models that reflect gestalt ideas in sound understanding. We discuss creative means to probe and capture sound understanding phenomena at the level of the crowd, including using a word-embedding space to assess causal uncertainty, a gamified web interface to capture trends in memorability, and a hierarchical label ontology to extrapolate from small-data annotations to dynamic, real-world audio settings; we discuss the value of domain adaptation in this problem space, suggesting the value of a bootstrapping paradigm with interpretable intermediate structure and constructing a secondary modeling structure on top of the crowd-scale models to attempt personalization; and we discuss the representation of annotation diversity as model uncertainty, as we encroach upon the cognitive ideas that sit in the gray area between the shared and subjective. Together, these tools will enable a suite of cognitive models in audition – including and beyond the ones that are documented and demonstrated in this thesis – that will allow for what I call *gestalt computation*. In the idea of gestalt computation, I suggest that the fundamental unit of processing in sound shifts from signal level properties to gestalt properties, an exact analog to the metaphorical dichotomy that is expressed in the dialogue between the researcher and participant in Section 1.2. Via gestalt computation, we can analyze and manipulate audio along gestalt dimensions, as we have already demonstrated in this thesis. But we can also go one step further – we can use gestalt measures as metrics for distance, assigning new meaning to the idea of "loss" that is entertained regularly in acoustics research, and we can synthesize and generate entirely *new* content along gestalt dimensions.

The language used here may insinuate that this notion of gestalt computation is a far-fetched thought; however, in the subsequent section, we describe an example of a simple research effort that is illustrative of the proposed paradigm and the potential it may afford future auditory interfaces.

7.1 *Manipulating Causal Uncertainty in Sound Objects*

As discussed in Section 4.1, research shows that we employ both acoustic and gestalt sound understanding processes to make complex inferences about the world around us. For example, if we hear a dog barking, we might notice that the pitch of the bark is low, but that its amplitude relative to the other sounds in our periphery is high, thereby drawing our attention. At the same time, we may process more abstract features about the sound, such as its emotionality, which leads us to believe that the dog is not a threat as we take a walk. Perhaps most importantly, ecological sound psychology research demonstrates that one of the primary processes that our mind engages in when interfacing with a sound object is attempting to estimate its source, or *cause* [81, 82]. When asked to describe the sound, for instance, we might say "a dog barked," suggesting that we have immediately inferred the cause of the sound.

As a review, this high-level aspect of auditory cognition – causal estimation – is known to play an important role in sound understanding. When treated as an intrinsic property, *causal uncertainty*, or how apparent the source of a sound is, has been shown to be a powerful indicator of how likely we are to remember the sound, attend to it, or respond to it emotionally [81, 82, 3].

Because of the wide range of phenomena in sound understanding that causal uncertainty drives, being able to manipulate a sound object's intrinsic causal uncertainty would prove very useful in audio experience and interface design. For instance, subtly altering the causal uncertainty of objects in a virtual reality soundscape might allow us to steer a listener's attention or focus towards specific spatial regions with time; we could envision augmented reality devices that modify causal uncertainty in the sounds that surround a user during periods of intense cognitive load to minimize distraction or surprise; and we might imagine algorithms that manipulate causal uncertainty in foley sounds used for film soundtrack design in an attempt to achieve heightened emotional impact.

To this end, in a project led by undergraduate researcher Tal Boger (Yale University)[1], we present a method for changing a sound's

[1] Statement of Contributions: Tal was a visiting undergraduate researcher under my mentorship. The research idea and steps to implementation were my work, while Tal contributed the implementation in entirety. The writing in this section is adapted from a publication which was a joint effort. I thank Tal for his excellent work and contribution to this dissertation.

causal uncertainty by optimization over perturbations in its acoustic properties. Unlike in the ecological audition or psychology literature, we cannot practically compute causal uncertainty by human label annotation and consensus, as in the HCU400 dataset. Instead, following the proposal in Section 4.4, we again use the uncertainty of a pre-trained audio classification model released by Google, YAMNet, as a proxy for human causal uncertainty[2]. To the best of our knowledge, this is the first known attempt at manipulating causal uncertainty in a structured fashion. Our early results point towards the possibility of using more generalizable learning methods (e.g., methods that can scale to multiple sounds and learn to use a wider range of manipulation strategies) with significant implications for experiences in sound interaction.

Our key contributions in this work are as follows:

(1) We design an optimization procedure which takes a sound excerpt as input and perturbs select acoustic properties (such as amplitude, pitch, playback speed, etc.) to scale the sound to a desired level of causal uncertainty.

(2) We apply the procedure to a selection of environmental sounds, quantitatively describe the results of the optimization in terms of convergence and the distribution of changes in acoustic features across sound classes.

(3) We demonstrate the effectiveness of the approach by conducting user listening tests, and show that listeners reliably perceive changes in causal uncertainty, matching those from our optimization procedure, in a sound comparison experiment.

7.1.1 *Related Work*

In this work, we aim to extend methods to quantify causal uncertainty by presenting a strategy to manipulate it. Specifically, we seek to morph sounds towards a target H_{cu}. To this end, methods for altering sounds have progressed significantly in recent years. New large-scale, statistical approaches using neural style transfer, generative adversarial networks, and other deep learning techniques have produced impressive results in the domains of music, speech, and environmental sounds [156, 157]. However, we find that such approaches require large datasets and significant compute to achieve training stability and convergence, especially in the context of the proposed task, which demands subtle changes to create ambiguity

[2] This approach, again, has its limitations that stem from the choice of AudioSet as the target ontology and models trained on this ontology as the inference engine. See Section 4.4.3 for an in-depth discussion.

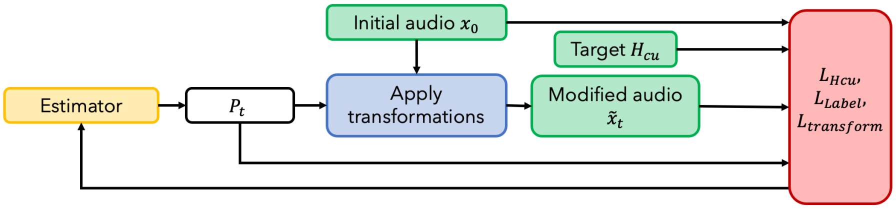


Figure 7.1: The system takes an audio input X_0 and target causal uncertainty (H_{cu}) value, and estimates parameter values P_t at each time step t for a select set of audio transforms. The transforms are applied to x_0 to generate a modified sound $\tilde{x}_t$ which, along with P_t and the target H_{cu}, are fed to a custom cost function defined over error in H_{cu} (L_{Hcu}), relatedness in sound labels (L_{label}), and transform magnitude ($L_{transform}$).

without significantly altering, adding, or removing sound objects or events. We further expect statistical methods to pose challenges in stability and complexity, because our method for estimating H_{cu} from a sound excerpt also requires a pre-trained neural network which may suffer from a lack of adversarial robustness.

As a simpler, alternative approach that serves as a proof-of-concept, we take inspiration from the work in [158, 159, 160], which presented a per-image optimization pipeline to modify the visual memorability of face images. The problem in [158, 159, 160] is analogous to ours, as both problems require similarly tight control over semantic and lower-level properties. Their method succeeded in optimizing images of faces to become more or less intrinsically memorable, and the results of their approach were verified in a perceptual task. Here, we take a similar approach in designing an optimization problem, adapting the optimization space and cost function to reflect the relevant semantic and acoustic properties of audio.

7.1.2 *Methods*

Overview Our proposed method operates as follows, as shown in the illustration in Figure 7.1. The method takes as input a target H_{cu} value and a sound excerpt, and applies a Gaussian process regression-based Bayesian optimization strategy [161], a "blackbox" optimization framework, to determine parameter values for a small, fixed set of acoustic transforms. To evaluate the parameter values, the acoustic feature transforms are applied to the input sound, and a cost function is computed on the result at each iteration of the optimization. We utilize a blackbox approach because our cost function is expensive to compute and not differentiable. For every sound that we wish to manipulate, we apply this optimization for a fixed number of calls and examine the result with the lowest cost as the output. We examine the individual components of this optimization process in the sections below.

Optimization Parameters We first define the parameters we are optimizing. To begin with a simple formulation, we create a constrained search space of select low-level audio features. We selected these features and their parameter ranges based on their definitions in a popular sound editing toolchain known as SoX[3], which we also use to implement the transforms. Table 7.1 shows the low-level feature values we optimize over, along with the range of their search spaces.

Table 7.1: Features and parameter ranges for optimization. Features are presented in the order they were applied to the sounds.

Feature name	**Search range**
Gain	[-25, 25] (dBs)
Pitch	[-250, 250] (hundredths of a semitone)
Playback Speed	[0.5, 1.5] (rate)
Reverberance	[0, 100] (factor)
High-pass filter	[1, 3000] (Hz, cutoff frequency)
Low-pass filter	[5000, 8000] (Hz, cutoff frequency)

Note that when optimizing a sound to *lower* its H_{cu} (i.e., making a sound *more certain*), we restrict the gain to [0, 25] dBs. This was to ensure that sounds were not becoming "more certain" by being converted to silence, an edge case in our optimization. We also note that this approach can be extended to several other fine-grained audio effects and transforms – examples include equalizers, limiters, compressors, etc. – but we choose to limit our early explorations to the set in Table 7.1 to achieve reasonable execution times under limited compute.

Objective Function In our optimization, we minimize the following loss function, which consists of a weighted sum of three terms:

$$L = \lambda_1 L_{H_{cu}} + \lambda_2 L_{label} + \lambda_3 L_{transform} \tag{7.1}$$

where L_x represents a loss term constraining a different aspect of the sound manipulation, and λ_i is a weighting term. Note that each λ_i is defined either as a constant, or a function of the loss terms. We provide the definitions for each term below.

$L_{H_{cu}}$ represents a measure to compare the H_{cu} at the current optimization step ($\hat{H}_{cu}$) with the target H_{cu}, and penalize the error. We define it as:

$$L_{H_{cu}} = |\hat{H}_{cu} - H_{cu}| \tag{7.2}$$

where H_{cu} is computed following [162] by taking the prediction output from YAMnet on the transformed sound and obtaining the maximum mean probability output. Then, we weight this term with a λ_1 parameter which is designed to penalize larger values of $L_{H_{cu}}$ more heavily. We defined λ_1 with a step function that monotonically increases as $L_{H_{cu}}$ increases from 0 to 0.5, after which point it is constant.

[3] http://sox.sourceforge.net/

While changing H_{cu} is our main goal, it is important to do so while preserving the label integrity of the initial sound. After all, any sound can be made causally uncertain by adding noise to the point that it is uninterpretable; doing this, however, results in little utility for achieving control in sound experiences via subtle changes. Therefore, we decide to penalize sounds based on how different their labels are from the labels of the initial sound by introducing L_{label}.

To construct L_{label}, we use node distances in the class ontology of AudioSet, the dataset used to train YAMNet [9, 2]. The ontology presents AudioSet classes in a tree structure. For example, the label "cat" has child nodes "purr" and "meow." We can use this tree structure to our advantage to create an intuitive L_{label} term. If we apply transforms that modify a sound that is initially labeled a "purr," we would want to penalize it much more for becoming a "chainsaw" than for becoming a "meow," given their relative distances on the ontology tree.

Let S_0 be the initial (i.e., before applying any audio transformations) set of top 10 most probable labels, and S_t be the top 10 most probable labels at step t. We define M to be a matrix of all pairwise combinations of labels in $S_0 \cup S_t$. M_{ij} is defined as the number of edges between label i and label j in AudioSet's ontology. For instance, a child-parent relationship consists of a single edge, and a sibling-sibling relationship consists of two edges. For label combinations where no connection exists between nodes, we set the distance equal to one more than the maximum possible number of edges. Then, we define L_{label} as the mean of $M_{ij|i<j}$ (i.e., M is symmetric, and we ignore diagonal entries, where we have the "distance" between a label and itself).

In defining λ_2, we have two separate cases: (1) Target $H_{cu} >$ initial H_{cu} (i.e., making a sound less certain); (2) Target $H_{cu} \leq$ initial H_{cu} (i.e., making a sound more certain). In case (1), we scale λ_2 according to the current H_{cu}. This is because as we raise H_{cu}, we expect the labels to vary slightly, so we want to relax the total penalty of the $\lambda_2 L_{label}$ term. As such, in case (1), we define:

$$\lambda_2 = \frac{1}{1 + \hat{H}_{cu}} \tag{7.3}$$

In case (2), we scale λ_2 according to the current L_{label}. When making a sound more certain, it is crucial to maintain its labels. So, we apply the following very strict penalty for label distance:

$$\lambda_2 = \frac{1}{1 - L_{label} + \varepsilon} \tag{7.4}$$

We lastly design a loss term to penalize large changes in the transform parameters. $L_{transform}$ is a normalized sum of our transforms such that the largest transformation of a specific feature corresponds to a penalty of 1 for that feature. We define it as:

$$L_{transform} = \frac{1}{6}\Bigg(\frac{|gain|}{gain_{max}} + \frac{|pitch|}{pitch_{max}} + \frac{|speed-1|}{speed_{max}-1} + \\ \frac{reverb}{reverb_{max}} + \frac{HPF}{HPF_{max}} + \frac{LPF_{max}-LPF}{LPF_{max}-LPF_{min}}\Bigg) \quad (7.5)$$

and λ_3 is set to a constant scalar.

7.1.3 *Experiments*

Dataset To evaluate our approach, we apply our optimization method to a selection of sounds from Google's AudioSet dataset [2]. The original dataset consists of 632 classes of sounds, with more than 2 million 10-second sound examples in total; however, we choose a small set of illustrative examples to demonstrate our results. Specifically, we choose four broad categories of environmental sounds – human sounds, animal sounds, nature sounds, and inorganic sounds – and selected pairs of classes of sounds within each category. These include: (1) Crying and laughing (human sounds); (2) Dog and cat (animal sounds); (3) Fire and water (nature sounds); (4) Wood and glass (inorganic sounds).

We run our optimization on all examples within the AudioSet balanced partition which list one of these categories, or their children in the ontology, as their primary label. This gave us 323 sounds in total (approximately 40 sounds per class). We downsample the audio to 16000 Hz to allow for compatibility with the YAMNet model, which is embedded in the cost function. This downsampling results in audio with less high-frequency detail, which may result in a narrower scope for subtle manipulations; however, this is a limitation of the network, rather than our approach.

Optimization Targets For each sound in our curated dataset, we apply the optimization to generate both a more uncertain (higher H_{cu}) and a less uncertain (lower H_{cu}) version, with target H_{cu} values of 0.8 and 0.2 respectively. We therefore restricted sounds in our dataset to include only those within an initial H_{cu} range of 0.3 and 0.7. This restriction ensures that we sample sounds from a broad range of ambiguity that require some modification to reach our target H_{cu}.

We allow the optimization for each sound to run for a maximum of

200 iterations using the parameter search ranges and cost function described in Sections 7.1.2 and 7.1.2. The initial values for the audio transforms are set to be neutral (zero gain, playback rate of 1, etc.), and the YAMNet model prediction is used to obtain an initial list of the top 10 labels describing the sound.

Perceptual Evaluation We finally create a listening task to evaluate whether our optimization results reflect human perception. Specifically, we wish to know whether raising or lowering a sound's H_{cu} results in more or less certainty in listener source estimation.[4]

We create a task wherein participants are asked to listen to two sounds and choose the sound for which they have greater certainty in its source. On 1/3 of the trials, the two sounds presented were the unchanged sound (the anchor) and the higher H_{cu} version of that sound, as created by our optimization pipeline. On 1/3 of the trials, the two sounds were the anchor and the lower H_{cu} version of that sound. On the remaining 1/3 of the trials, there was no anchor; the two sounds presented were the higher H_{cu} and lower H_{cu} versions of the same sound. This creates a two-alternative forced-choice task to quantify our success in changing a sound's causal uncertainty.

A single experiment included 48 trials in total, split into 6 blocks of 8. Each block contained one sound sample of each class. The sounds chosen, along with the order of the sounds within-block, the order of the trial types, and the position of the more certain sound, were all randomized within-subject.

To conduct the study, we recruited 20 participants from the online crowd-sourcing platform Prolific (for a discussion of the reliability of Prolific's subject pool, see [163]). Each experiment took approximately 25 minutes, and each participant was compensated upon completion of the experiment.

7.1.4 *Results*

Samples of the original and manipulated sounds can be found at the repository accompanying this work[5].

In our perceptual task, human evaluations aligned with our optimization results. Subjects were able to choose the more causally certain sound (as determined by our proxy H_{cu}) at a rate significantly above chance ($t(19) = 4.46$, $p < 0.0001$, $M = 57.60\%$, 95% CIs = [54.04%, 61.17%]). This was not simply driven by a few subjects per-

[4] You can try the task yourself at `http://audio-mafia.media.mit.edu/hcu_task/`

[5] `https://osf.io/6xmv7/?view_only=c2dd0d00d2064b61ad380fb14e640664`

forming with very high accuracy; 16 of 20 subjects chose the more causally certain sound over half the time.

The results of each of the three trial types (higher-anchor, higher-lower, and lower-anchor) were significantly different from chance (see Figure 7.2). On higher-anchor trials, participants had nearly perfect accuracy (90.15%). On higher-lower trials – where the original sound was not presented – participants chose the more causally certain sound 68.96% of the time. Finally, on the lower-anchor trials, participants consistently mistook the original sound as more causally certain than the one with lower H_{cu} (12.81% accuracy).

The very poor accuracy in the lower-anchor trials has two potential causes. First, it highlights the challenge of making the source of a sound less uncertain, using only a simple set of acoustic tools – we discuss the framing of the task and future strategies for improving performance in Section 7.1.5. Secondly, the results demonstrate that participants potentially perceive *any* change to a sound using our effects chain as increasing its source ambiguity, suggesting that the manipulated sounds do not seem natural. We expect that this behavior can be controlled for by more subtle, computationally intensive sound operations (see Section 7.1.5).

Finally, we compare the accuracy per class in our perceptual task. In Figure 7.3, we note that we do not have homogeneous results across classes; for instance, sounds with the labels fire and water – the two natural sound classes – have lower accuracy than the sounds with other class labels. This hints at differences between sounds – either spectral, semantic, or both – that could be further exploited for the manipulation of causal uncertainty, with deeper analysis that could stem from a larger number of data points per trial.

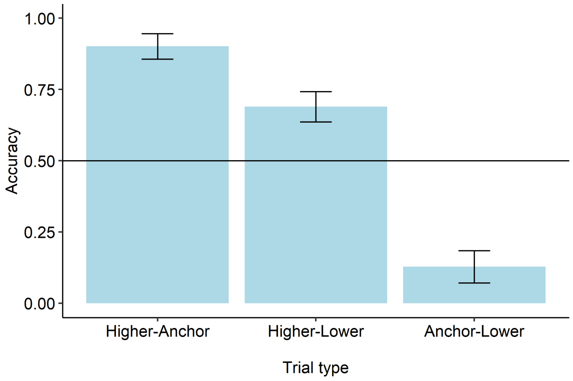


Figure 7.2: Grouped subject accuracy for each trial type in perceptual task. Error bars are 95% CIs. All groups differ from 0.5 with p < 0.001

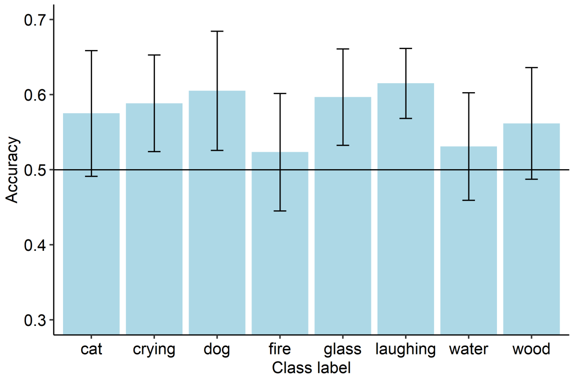


Figure 7.3: Grouped subject accuracy for each class in the perceptual task, across trial types. Error bars are 95% CIs.

7.1.5 *Discussion*

We have shown that our optimization pipeline succeeds in altering a sound's H_{cu} to within a close range of a target while maintaining its labels. These results are then confirmed by a perceptual task, where human judgments begin to align with our optimized results. The results from the perceptual task shed light on interesting areas of further work, especially regarding limitations in our dataset, methods, and the notion of changing H_{cu}.

Dataset Selection In our evaluation, some limitations stem from our sampling approach. For example, several sounds of one class may not be isolated (e.g., a "rain" sound having thunder in the end), which affects both its cognitive properties (as thunder may help one identify the sound as rain) and its transformations (as the same transform affects the uncertainty of rain and thunder differently). On the contrary, we do not perceive sounds in perfectly isolated environments; we perceive them as part of a broader world, which often includes other sounds and properties. Our dataset selection reflects how one would change sounds "in the wild" as opposed to in controlled, isolated environments.

Along with the issue of isolation, our dataset contains a wide variety of sounds within a specific class. Within the class of "dog" sounds, for instance, are sounds of both dog barks and dog cries. While we choose the parent label as the "cause" to demonstrate our approach, a more granular exploration would be a valuable future exercise.

We also evaluated our methods on a small subset of the total classes available in AudioSet. However, we chose both broad categories (e.g., both natural and artificial sounds) and orthogonal classes within our categories (e.g., dog and cat), which point to the generality of our approach. We intend to expand our evaluation to a broader set of AudioSet classes in future explorations.

Modeling Approaches Though we present a simple and robust approach, there are several ways to extend our framework for changing a sound's causal uncertainty. One may consider how adding additional transformations to increase sensitivity, removing transformations to create a more controlled set of changes, or changing the order of transformations may affect results.

Methods other than our blackbox optimization may yield better results, too, given that corresponding changes to the dataset are made. For instance, since our current *Hcu* approximation method does not restrict us to working with small datasets, the most natural extension to this work would entail the use of deep learning-based approaches – this may allow us to move beyond low-level feature manipulation towards semantic content manipulation, and allow for higher resolution, smooth and continuous changes to the *Hcu* attribute of a given sound. Future work may explore the viability of such approaches for this problem.

Despite its simplicity, this approach presents a first step in generalized methods for scaling complex properties of sound objects, with powerful implications for user experiences. This optimization methodology can be readily extended to other annotated sound properties – examples include affect and memorability – when coupled with custom or off-the-shelf proxy estimation models that scale to real-world audio.

The Meaning of Reducing Causal Uncertainty The poor accuracy resulting from the lower-anchor trials in our perceptual task raises questions regarding the philosophical meaning of changing a sound's causal uncertainty. *Raising* a sound's causal uncertainty is easy to define and understand, as it simply requires making its source less clear. However, what does it mean to take an already-uncertain sound, and *lower* its causal uncertainty? Seemingly, the opposite of the raising H_{cu} definition applies – lowering a sound's causal uncertainty requires making its source more clear. However, this requires *adding* information to a sound to allow it to be more identifiable, which must be inferred. Our current methods are not well-equipped to achieve this.

Perhaps making a sound less causally uncertain demands a broader set of tools that includes both a suite of subtle, production quality acoustic effects, as well as the insertion or deletion of content on a semantic level. To experiment with the former, we might expand the optimization space to include operations such as multiband equalizers and compressors, band-specific filters, and limiters, without constraining the order of application. To consider the latter, we eventually look to large-scale statistical approaches, such as deep neural networks, in order to learn to generate a wider diversity of natural-sounding excerpts that meet the target H_{cu} constraint.

Nevertheless, any future work requires additional analysis and discussion surrounding the definition of reducing causal uncertainty from the standpoint of cognitive processing and sound understanding.

7.2 *Applications*

I conclude this thesis with a discussion of the technological future that gestalt computation could catalyze, by speculating about its intersection with a few themes of active discussion in the research community today. Some of the ideas that result from this marriage are lofty, futuristic, and are perhaps more worthy of cinematic storytelling than practical research aspirations; some, on the other hand, are but a few years away from being within our grasp.

Compression While audio is becoming easier to capture, cheaper to store, and faster to retrieve by the day, our mindset surrounding the idea of compression hasn't changed in decades. However, having access to gestalt models allows us to restructure these mental models; for instance, rather than being focused on constraints of perceptibility that are drawn from psychoacoustics research, we can consider measuring error, or "loss", in more abstract terms. Imagine a system that considers the portion of an audio recording you are least likely to attend to or remember, and deletes it, instead of the portion that is least likely to be audible; imagine a system that takes the least valent or arousing components of a recording, deletes those objects at compression time, and inserts arbitrary equivalents of those sound objects – from a standard database – at replay time. Much like in the way that natural language processing attempts to distill long transcripts of writing into meaningful, interpretable text bytes are distinct from condensing them by conversion into a .zip file, compression in this sense may lead to new content and new listening behaviors altogether. Incorporating more complex perceptual ideas into compression standards is already being explored in vision[6]; and notably, we are steadily building up the infrastructure that would allow us

[6] https://opensource.googleblog.com/2021/09/using-saliency-in-progressive-jpeg-xl-images.html

to do it in audio – the MPEG-7 standard[7] allows for the marking of time-dependent audio metadata which could be used to tag audio with gestalt information, and the new Dolby Atmos standard[8] is already encouraging audio content creators to treat sounds as individual sound percepts.

Recording Since the days of Edison's famous Montclair tone tests[9], little has changed in our definition of a recording. The idea, still, is to reproduce the soundfield present at a particular location in space, where the recording device was present, in a way that is as faithful to the original as possible. However, as a society, we have learned to listen differently, and learned to alter our expectations for the listening experience – the Edison experiments would be subject to mockery if replicated today in their original fidelity. If this is the case, we might consider changing aspects of the idea of a recording – the idea behind the most common auditory interfaces we encounter today – in ways that facilitate new experiences, create scope for artistic work, or simply serve utilitarian purposes in the face of audio data scale. For instance, gestalt computation can enable the selective capture of content along cognitive dimensions – this points towards systems with intelligent pass-through modes for spaces and devices that cause acoustic isolation, like cars and noise-cancelling headphones; or towards embedded recording devices for acoustic monitoring in complex environments that can operate with a limited power budget, though under the tradeoff of larger prediction error margins that come with small-footprint neural networks. Gestalt computation can enable non-deterministic replay, that plays with the idea of attentional foreground and background as a function of our conscious or subconscious needs – an extension, for instance, on the custom soundscapes that were constructed in Chapter 6, but into a possibly larger combinatorial space consisting of multiple sources or perceptual objectives that are varying over time. And lastly, it can even enable the degradation or erosion of audio content in a manner aligned with theories of memorability, exhibiting artistic notions of ephemerality, creating a sense of nostalgia, or even eliminating content that is a source of discomfort.

Auditory Augmented Reality One of the most promising applications of this work is in the space of auditory augmented reality (AAR). Not many practical realizations of AAR devices exist to date – Gershon Dublon's [5] is one of the earliest – but several major corporate players in the audio space are actively working on forthcoming products[10]. At the crossroads of AAR and gestalt computation, one could imagine customized sonic environments for a wearer that are

[7] https://en.wikipedia.org/wiki/MPEG-7

[8] https://www.dolby.com/technologies/dolby-atmos/

[9] https://blogs.loc.gov/now-see-hear/2015/05/is-it-live-or-is-it-edison/

[10] https://tech.fb.com/inside-facebook-reality-\labs-research-the-future-of-audio/

achieved by selectively amplifying, muting, inserting, and deleting sounds in consideration of the overall content in the scene that is arousing or causally uncertain – a simple of manifestation of this idea would be noise-cancelling headphones that exhibit "intelligent pass-through" behavior; one could imagine modifying environmental sounds in real-time with subtle acoustic perturbations, like in the work in Section 7.1, to optimize for one of these metrics as a function of a wearer's use case; and one could imagine new soundscapes altogether that are generated for a wearer after an assessment of the current environmental dynamics. In particular, there are new perceptual sensibilities afforded by the overlay of virtual sounds on our natural sense of hearing, which is an integral aspect of AAR; this, in turn, forms a perfect environment in which gestalt computation can exist and thrive.

Content Creation Lastly, we discuss perhaps the most obvious application – the creation of new media content via gestalt computation. The gestalt computation paradigm allows for the analysis and synthesis of audio media along cognitive dimensions; this points towards a set of tools that may help creators rapidly prototype content that is appropriate given other guiding heuristics – for instance, assembling soundscapes for an immersive video game, constructing foley sound for a film or podcast, and generating ambience to accompany an interactive art installation. Such tools could follow the creative prototyping method introduced in an older project in the Responsive Environments group, called VisualSoundtrack [164], wherein the current feature axes that are sketched upon are replaced by concepts like familiarity, sound source uncertainty, and memorability.

7.3 The Sounds We Seek

Greg Milner's beautiful 2011 anthropological narrative, *Perfecting Sound Forever* [165], begins with one of my favorite anecdotes. Milner tells the story of Guglielmo Marconi, the father of modern radio and communications infrastructure, and an epiphany that he had towards the end of his life. "The godfather of radio technology decided," Milner writes, "that no sound ever dies. It just decays beyond the point that we can detect it with our ears. Any sound was forever recoverable," he believed – if only we had the right device. And if he had such a device, he was often asked, what would *he* choose to listen to? Marconi, a man known for his rationality, scientific approach, and technical insight, had an answer that surprised many – his dream was to hear the music that was playing on the night the Titanic sank, or listen to the voice of Christ delivering the Sermon on the Mount.

This anecdote is a touching reminder of something that makes us human – rarely, in sound, do we look for that which is objective; we search, instead, for emotions, aesthetics, and sentiment. Rarely do we find ourselves scrubbing through hundreds of hours of audio in search of an uttered phrase or in an attempt to document a precise location or time; we are far more likely to find ourselves smirking unexpectedly at the sound of laughter that erupts after a group jam session; moved to tears at the sound of the voice of a grandparent who has long passed on; or feeling transported to a busy market in our home country, thousands of miles away, at the sound of cars honking at an intersection in just the right way. To build the technologies that can create these experiences for us, again and again and again, is to embrace our cerebral complexity and our love for the subjective. To build these sorts of technologies is to know that there is immense potential in tapping into the mysteries that make us, us.

"The knower of the mystery of sound knows the mystery of the whole universe."

- Hazrat Inayat Khan

Appendix

Audio Summarization - User Study

Participant Procedure

This study has received approval from the MIT Institutional Review Board (IRB) under the protocol #2006000177. The complete application and consent form can be viewed on the MIT IRB platform. The participant procedure is excerpted below:

If you volunteer to participate in this study, we would ask you to do the following things:

1. Provide us a contactless means for providing you with the necessary equipment (a small, pocket sized audio recorder) – for example, your address or place of work for delivery. The device is unused, and we encourage you to wear a face covering/ disposable gloves when collecting the delivery, and to sanitize the product using alcohol based wipes or sprays consisting of at least 70% alcohol before use in light of Covid-19 (following the recommendation here – `https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/cleaning-disinfection.html`).

2. Choose an easy means to secure the device on your person – for example, leave it in your pocket, secure it to your wrist using the provided Velcro straps, etc.

3. Proceed through your day as you would ordinarily, but ensure that the recorder is on and running. When doing so, please note the following:

    (a) You may switch off the recorder at any point during the day, if you do not feel comfortable with recording for a brief period of time.

(b) You may encounter other individuals and record discernable speech that they produce. In this scenario, we ask that you do not record without disclosure and consent (as per Massachusetts law). You may instead (1) inform the individual(s) that you encounter that you are recording the conversation and obtain verbal or written consent to be recorded from them (if the individual is a minor, parental consent and child assent for children over the age of 7 is required); or (2) choose to switch off the recording device temporarily. If speech or other sensitive content is recorded accidentally, you may retroactively remove content from the recording before submitting to the researchers. Sample script to request verbal consent: "I am participating in a research study on capturing sounds in a person's environment, and would like to record this conversation. The content of this conversation will not be reviewed by anyone other than myself and the researchers conducting the study, except as required by law. Do I have your permission to record? If so, please state your name and confirm in a full sentence."

(c) While recording audio is at the user's discretion, we ask that you provide us with at least 3-4 hours of audio per day. If this will not be possible, we ask that you do not participate in the study.

4. Once every 2 days, you will be asked to upload your recordings from the SD card on your device to a secure repository on a server in our lab. You will receive a unique link to this location by email, associated with an anonymous participant ID. Prior to uploading, you may choose to delete sensitive portions of the recordings, though you are not permitted to edit submitted content in any other way. Your recordings will not be accessible by anyone other than you and the aforementioned researchers.

5. You will be asked to continue this exercise for up to 15 days.

6. Shortly after your final upload, you will receive a link by email that points you to a set of audio summaries generated from your recordings, as well as instructions to a short survey to complete as you listen to these summaries. There is no time limit on this exercise, but it should not take you more than 2 hours to complete.

7. After the study is complete, you will be free to retain the device in the near-term. Near the end of the research period (approximately 6 months), you will be requested to return the device via a

contactless procedure if you feel comfortable doing so.

Survey Questions
Listen to the following audio presentation generated from the audio you recorded:
Q1: How would you say listening to this presentation made you feel? Select a point along the scale from least to most emotionally evocative. *[Select from a scale of 1 (least) to 5 (most)]*
Q2: How would you describe the sentiment (if any) associated with this presentation? Select a point along the scale of "negative" to "positive". *[Select from a scale of 1 (negative) to 3 (neutral) to 5 (positive)]*
Q3: Which of the following terms best describe the presentation? Select as many as apply. *[Select from – calming, annoying, nostalgic, peaceful, social, familiar, relaxing, busy, comfort, distracting, activity, reminder of events, surprising, stressful, summary of events, salient, uncomfortable, memory aid, loud, eerie]*
Q4: How would you describe how intimate this presentation felt? Select a point along the scale from least to most intimate. *[Select from a scale of 1 (This presentation felt generic, reflecting sounds that could have been recorded by others.) to 5 (This presentation is uniquely mine, reflecting the spaces and events I recorded.)]*

Reflection Guiding Questions
G1: What surprised you most about what you heard? What didn't surprise you?
G2: Do you find listening to the audio to be an immersive experience? Why or why not?
G3: Would you use such audio presentations to review or reflect on your day, week, month, or year? Why or why not?

Bibliography

[1] I. Ananthabhotla, D. B. Ramsay, and J. A. Paradiso, "HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924, IEEE, 2019.

[2] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, IEEE, 2017.

[3] D. B. Ramsay, I. Ananthabhotla, and J. A. Paradiso, "The Intrinsic Memorability of Everyday Sounds," in *AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.

[4] R. Kleinberger, G. Dublon, J. A. Paradiso, and T. Machover, "PhoxEars: a Parabolic, Head-mounted, Orientable, Extrasensory Listening Device.," in *NIME*, pp. 30–31, 2015.

[5] G. Dublon, *Sensor(y) Landscapes: Technologies for New Perceptual Sensibilities*. PhD thesis, Massachusetts Institute of Technology, 2018.

[6] I. Ananthabhotla and J. A. Paradiso, "Soundsignaling: Realtime, stylistic modification of a personal music corpus for information delivery," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–23, 2018.

[7] J. R. Stroop, "Studies of interference in serial verbal reactions.,"

Journal of Experimental Psychology: General, vol. 121, no. 1, p. 15, 1992.

[8] S. Monsell, "Task Switching," *Trends in Cognitive Sciences*, vol. 7, no. 3, pp. 134–140, 2003.

[9] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN Architectures for Large-scale Audio Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, IEEE, 2017.

[10] P. Verma and J. Berger, "Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions," *arXiv preprint arXiv:2105.00335*, 2021.

[11] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[12] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," *arXiv preprint arXiv:2106.13043*, 2021.

[13] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[14] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," *arXiv preprint arXiv:2107.03649*, 2021.

[15] A. Greco, A. Roberto, A. Saggese, and M. Vento, "Denet: a deep architecture for audio surveillance applications," *Neural Computing and Applications*, pp. 1–12, 2021.

[16] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound event detection and time–frequency segmentation from weakly labelled data," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 777–787, 2019.

[17] K. Guirguis, C. Schorn, A. Guntoro, S. Abdulatif, and B. Yang, "Seld-tcn: sound event localization & detection via temporal

convolutional networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 16–20, IEEE, 2021.

[18] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3879–3888, 2019.

[19] S. Verbitskiy and V. Vyshegorodtsev, "Eranns: Efficient residual audio neural networks for audio pattern recognition," *arXiv preprint arXiv:2106.01621*, 2021.

[20] S. S. F. Russell, *Resynthesizing volumetric soundscapes: Low-rank subspace methods for soundfield estimation and reconstruction*. PhD thesis, Massachusetts Institute of Technology, 2020.

[21] C. Duhart, G. Dublon, B. Mayton, G. Davenport, and J. A. Paradiso, "Deep learning for wildlife conservation and restoration efforts," in *36th International Conference on Machine Learning, Long Beach*, vol. 5, 2019.

[22] D. Baby, A. Van Den Broucke, and S. Verhulst, "A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 134–143, 2021.

[23] M. Slaney, R. F. Lyon, R. Garcia, B. Kemler, C. Gnegy, K. Wilson, D. Kanevsky, S. Savla, and V. G. Cerf, "Auditory measures for the next billion users," *Ear and Hearing*, vol. 41, pp. 131S–139S, 2020.

[24] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, pp. 18–25, Citeseer, 2015.

[25] M. Bosi and R. E. Goldberg, *Introduction to digital audio coding and standards*, vol. 721. Springer Science & Business Media, 2002.

[26] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1518–1525, 2019.

[27] O. Kalinli and S. Narayanan, “Prominence detection using auditory attention cues and task-dependent high level information,” *IEEE Transactions on audio, Speech, and language processing*, vol. 17, no. 5, pp. 1009–1024, 2009.

[28] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, “Mechanisms for allocating auditory attention: an auditory saliency map,” *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[29] E. M. Kaya and M. Elhilali, “Modelling auditory attention,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, p. 20160101, 2017.

[30] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[31] S. E. McAdams and E. E. Bigand, “Thinking in sound: The cognitive psychology of human audition.,” in *Based on the fourth workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society.*, Clarendon Press/Oxford University Press, 1993.

[32] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” 2017.

[33] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” *arXiv preprint arXiv:1806.03185*, 2018.

[34] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35, IEEE, 2016.

[35] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, “Cnn architectures for large-scale audio classification,” in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135, IEEE, 2017.

[36] A. K. Vijayakumar, R. Vedantam, and D. Parikh, “Sound-word2vec: Learning word representations grounded in

sounds," *arXiv preprint arXiv:1703.01720*, 2017.

[37] A. Jansen, M. Plakal, R. Pandya, D. P. Ellis, S. Hershey, J. Liu, R. C. Moore, and R. A. Saurous, "Unsupervised learning of semantic audio representations," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 126–130, IEEE, 2018.

[38] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021.

[39] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert systems with applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[40] D. E. Broadbent, "A mechanical model for human attention and immediate memory.," *Psychological review*, vol. 64, no. 3, p. 205, 1957.

[41] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 577–581, IEEE, 2014.

[42] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2017.

[43] J. Turian and M. Henry, "I'm sorry for your loss: Spectrally-based audio distances are bad at pitch," *arXiv preprint arXiv:2012.04572*, 2020.

[44] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Using a neural network codec approximation loss to improve source separation performance in limited capacity networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020.

[45] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency

weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[46] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, 2001.

[47] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5074–5078, IEEE, 2018.

[48] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech separation system to improve stoi and pesq directly," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5374–5378, 2018.

[49] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, p. 1780–1792, 2018.

[50] K. Zhen, A. Sivaraman, J. Sung, and M. Kim, "On psychoacoustically weighted cost functions towards resource-efficient deep neural networks for speech denoising," *arXiv preprint arXiv:1801.09774*, 2018.

[51] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models," in *Processings of the ACM International Conference on Multimedia (ACM MM)*, 2019.

[52] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 694–711, Springer, 2016.

[54] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings*

of the International Conference on Medical Mmage Computing and Computer-assisted Intervention, pp. 234–241, Springer, 2015.

[55] D. Stoller, S. Durand, and S. Ewert, "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Brighton, UK), pp. 181–185, 2019.

[56] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks.," in *Proceedings Interspeech*, pp. 352–356, 2016.

[57] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," 2017.

[58] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space," in *Audio Engineering Society Convention 107*, Audio Engineering Society, 1999.

[59] S. Ghorbal, R. Séguier, and X. Bonjour, "Process of hrtf individualization by 3d statistical ear model," in *Audio Engineering Society Convention 141*, Audio Engineering Society, 2016.

[60] T. Huttunen, A. Vanne, S. Harder, R. R. Paulsen, S. King, L. Perry-Smith, and L. Kärkkäinen, "Rapid generation of personalized hrtfs," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Audio Engineering Society, 2014.

[61] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha, "P-hrtf: Efficient personalized hrtf computation for high-fidelity spatial sound," in *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 53–61, IEEE, 2014.

[62] K. McMullen, A. Roginska, and G. H. Wakefield, "Subjective selection of head-related transfer functions (hrtf) based on spectral coloration and interaural time differences (itd) cues," in *Audio Engineering Society Convention 133*, Audio Engineering Society, 2012.

[63] D. Schönstein and B. F. Katz, "Hrtf selection for binaural synthesis from a database using morphological parameters," in *International Congress on Acoustics (ICA)*, 2010.

[64] S. Spagnol, "Hrtf selection by anthropometric regression for improving horizontal localization accuracy," *Ieee Signal Processing Letters*, vol. 27, pp. 590–594, 2020.

[65] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[66] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[67] Z. Ben-Hur, D. Alon, P. W. Robinson, and R. Mehra, "Localization of virtual sounds in dynamic listening using sparse hrtfs," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society, 2020.

[68] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[69] K. Yamamoto and T. Igarashi, "Fully perceptual-based 3d spatial sound individualization with an adaptive variational autoencoder," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–13, 2017.

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[71] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The cipic hrtf database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pp. 99–102, IEEE, 2001.

[72] W. W. Gaver, "What in the World do we Hear?: An Ecological Approach to Auditory Event Perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[73] W. W. Gaver, "How do we Hear in the World? Explorations in Ecological Acoustics," *Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.

[74] M. M. Marcell, D. Borella, M. Greene, E. Kerr, and S. Rogers, "Confrontation naming of environmental sounds," *Journal of clinical and experimental neuropsychology*, vol. 22, no. 6, pp. 830–864, 2000.

[75] B. Gygi and V. Shafiro, "General functions and specific applications of environmental sound research," *Frontiers in Bioscience*, vol. 12, pp. 3152–3166, 2007.

[76] O. Bones, T. J. Cox, and W. J. Davies, "Distinct categorization strategies for different types of environmental sounds," Euronoise, 2018.

[77] J. W. Lewis, F. L. Wightman, J. A. Brefczynski, R. E. Phinney, J. R. Binder, and E. A. DeYoe, "Human brain regions involved in recognizing environmental sounds," *Cerebral cortex*, vol. 14, no. 9, pp. 1008–1021, 2004.

[78] B. L. Giordano, J. McDonnell, and S. McAdams, "Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds," *Brain and cognition*, vol. 73, no. 1, pp. 7–19, 2010.

[79] S. M. Aglioti and M. Pazzaglia, "Representing actions through their sound," *Experimental brain research*, vol. 206, no. 2, pp. 141–151, 2010.

[80] L. Pizzamiglio, T. Aprile, G. Spitoni, S. Pitzalis, E. Bates, S. D'amico, and F. Di Russo, "Separate neural systems for processing action-or non-action-related sounds," *Neuroimage*, vol. 24, no. 3, pp. 852–861, 2005.

[81] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds.," *Journal of experimental psychology: human perception and performance*, vol. 19, no. 2, p. 250, 1993.

[82] G. Lemaitre, O. Houix, N. Misdariis, and P. Susini, "Listener expertise and sound identification influence the categorization of environmental sounds.," *Journal of Experimental Psychology:*

Applied, vol. 16, no. 1, p. 16, 2010.

[83] K. Inui, T. Urakawa, K. Yamashiro, N. Otsuru, M. Nishihara, Y. Takeshima, S. Keceli, and R. Kakigi, "Non-linear laws of echoic memory and auditory change detection in humans," *BMC neuroscience*, vol. 11, no. 1, p. 80, 2010.

[84] A. Schirmer, Y. H. Soh, T. B. Penney, and L. Wyse, "Perceptual and conceptual priming of environmental sounds," *Journal of cognitive neuroscience*, vol. 23, no. 11, pp. 3241–3253, 2011.

[85] J. S. Snyder and M. Elhilali, "Recent advances in exploring the neural underpinnings of auditory scene perception," *Annals of the New York Academy of Sciences*, vol. 1396, no. 1, pp. 39–55, 2017.

[86] I. Winkler, S. L. Denham, and I. Nelken, "Modeling the auditory scene: predictive regularity representations and perceptual objects," *Trends in cognitive sciences*, vol. 13, no. 12, pp. 532–540, 2009.

[87] B. R. Buchsbaum, R. K. Olsen, P. Koch, and K. F. Berman, "Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory," *Neuron*, vol. 48, no. 4, pp. 687–697, 2005.

[88] C. J. Vaidya, M. Zhao, J. E. Desmond, and J. D. Gabrieli, "Evidence for cortical encoding specificity in episodic memory: memory-induced re-activation of picture processing areas," *Neuropsychologia*, vol. 40, no. 12, pp. 2136–2143, 2002.

[89] L. Jäncke, "Music, memory and emotion," *Journal of biology*, vol. 7, no. 6, p. 21, 2008.

[90] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[91] J. E. LeDoux, "Emotion, memory and the brain," *Scientific American*, vol. 270, no. 6, pp. 50–57, 1994.

[92] J. C. Bartlett, "Remembering environmental sounds: The role of verbalization at input," *Memory & Cognition*, vol. 5, no. 4, pp. 404–414, 1977.

[93] M. M. Bradley and P. J. Lang, “The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual,” *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[94] A. Schirmer, Y. H. Soh, T. B. Penney, and L. Wyse, “Perceptual and conceptual priming of environmental sounds,” *Journal of cognitive neuroscience*, vol. 23, no. 11, pp. 3241–3253, 2011.

[95] B. R. Buchsbaum, R. K. Olsen, P. Koch, and K. F. Berman, “Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory,” *Neuron*, vol. 48, no. 4, pp. 687–697, 2005.

[96] L. Huang, H. Ji, *et al.*, “Learning phrase embeddings from paraphrases with GRUs,” in *Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora*, pp. 16–23, 2017.

[97] D. P. Chris *et al.*, “Another stemmer,” in *ACM SIGIR Forum*, vol. 24, pp. 56–61, 1990.

[98] G. A. Miller, “WordNet: a lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[99] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[100] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge.,” in *AAAI*, pp. 4444–4451, 2017.

[101] W. A. Bainbridge, P. Isola, and A. Oliva, “The intrinsic memorability of face photographs.,” *Journal of Experimental Psychology: General*, vol. 142, no. 4, p. 1323, 2013.

[102] J. A. Ballas, “Common factors in the identification of an assortment of brief everyday sounds,” *Journal of experimental psychology: human perception and performance*, vol. 19, no. 2, p. 250, 1993.

[103] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott,

"Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[104] I. Ananthabhotla, D. Ramsay, and J. Paradiso, "Hcu400: An annotated dataset for exploring aural phenomenology through causal uncertainty," *International Conference on Acoustics, Speech, and Signal Processing*, 2018. under review; `http://arxiv.org/abs/1811.06439`.

[105] G. Richard, S. Sundaram, and S. Narayanan, "An overview on perceptually motivated audio indexing and classification," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1939–1954, 2013.

[106] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, pp. 18–25, 2015.

[107] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[108] F. Font, T. Brookes, G. Fazekas, M. Guerber, A. La Burthe, D. Plans, M. D. Plumbley, M. Shaashua, W. Wang, and X. Serra, "Audio commons: bringing creative commons audio content to the creative industries," in *Audio Engineering Society Conference: 61st International Conference: Audio for Games*, Audio Engineering Society, 2016.

[109] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[110] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.

[111] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *arXiv preprint arXiv:1703.05175*, 2017.

[112] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*, pp. 84–92, Springer, 2015.

[113] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017.

[114] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," *arXiv preprint arXiv:1605.06065*, 2016.

[115] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3018–3027, 2017.

[116] M. M. Bradley and P. J. Lang, "International affective digitized sounds (iads): Stimuli, instruction manual and affective ratings (tech. rep. no. b-2)," *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida*, 1999.

[117] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2021.

[118] C.-E. González-Gallardo, R. Deveaud, E. SanJuan, and J.-M. Torres, "Audio summarization with audio features and probability distribution divergence," *arXiv preprint arXiv:2001.07098*, 2020.

[119] J. Ajmera, O. D. Deshmukh, A. Jain, A. A. Nanavati, N. Rajput, and S. Srivastava, "Audio cloud: creation and rendering," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 277–280, 2012.

[120] C.-T. Kao, Y.-T. Liu, and A. Hsu, "Speeda: adaptive speed-up for lecture videos," in *Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology*, pp. 97–98, 2014.

[121] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pp. 489–498, 1999.

[122] S. Vemuri, P. DeCamp, W. Bender, and C. Schmandt, "Improving speech playback using time-compression and speech recognition," in *Proceedings of the SIGCHI conference on Human*

factors in computing systems, pp. 295–302, 2004.

[123] S. Tucker and S. Whittaker, “Temporal compression of speech: An evaluation,” *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 4, pp. 790–796, 2008.

[124] D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann, “Beyond audio and video retrieval: towards multimedia summarization,” in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pp. 1–8, 2012.

[125] A. G. Del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, “Summarization of egocentric videos: A comprehensive survey,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 65–76, 2016.

[126] H. H. Kim and Y. H. Kim, “Toward a conceptual framework of key-frame extraction and storyboard display for video summarization,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 5, pp. 927–939, 2010.

[127] M. Cooper and J. Foote, “Automatic music summarization via similarity analysis.,” in *ISMIR*, 2002.

[128] P. Grosche, M. Müller, and J. Serrà, “Towards cover group thumbnailing,” in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 613–616, 2013.

[129] J. Paulus and A. Klapuri, “Music structure analysis by finding repeated parts,” in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pp. 59–68, 2006.

[130] W. Chai and B. Vercoe, “Music thumbnailing via structural analysis,” in *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 223–226, 2003.

[131] A. Zlatintsi, P. Maragos, A. Potamianos, and G. Evangelopoulos, “A saliency-based approach to audio event detection and summarization,” in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 1294–1298, IEEE, 2012.

[132] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and A. Potamianos, “Predicting audio-visual salient events based

on visual, audio and text modalities for movie summarization," in *2015 IEEE international conference on image processing (ICIP)*, pp. 4361–4365, IEEE, 2015.

[133] Y. Jin, T. Lu, and F. Su, "Movie keyframe retrieval based on cross-media correlation detection and context model," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 816–825, Springer, 2012.

[134] S. Kayukawa, K. Higuchi, R. Yonetani, M. Nakamura, Y. Sato, and S. Morishima, "Dynamic object scanning: Object-based elastic timeline for quickly browsing first-person videos," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2018.

[135] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood, "Sensecam: A retrospective memory aid," in *International Conference on Ubiquitous Computing*, pp. 177–193, Springer, 2006.

[136] R. C. Tucker, M. Hickey, and N. Haddock, "Speech-as-data technologies for personal information devices," *Personal and Ubiquitous Computing*, vol. 7, no. 1, pp. 22–29, 2003.

[137] H. V. Le, S. Clinch, C. Sas, T. Dingler, N. Henze, and N. Davies, "Impact of video summary viewing on episodic memory recall: Design guidelines for video summarizations," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 4793–4805, 2016.

[138] W. Jiang, C. Cotton, and A. C. Loui, "Automatic consumer video summarization by audio and visual analysis," in *2011 IEEE international conference on multimedia and expo*, pp. 1–6, IEEE, 2011.

[139] M. L. Cooper and J. Foote, "Automatic music summarization via similarity analysis.," in *ISMIR*, 2002.

[140] N. A. Heckert, J. J. Filliben, C. M. Croarkin, B. Hembree, W. F. Guthrie, P. Tobias, and J. Prinz, "Handbook 151: Nist/sematech e-handbook of statistical methods," 2002.

[141] D. Schwarz, "Corpus-based concatenative synthesis," *IEEE*

signal processing magazine, vol. 24, no. 2, pp. 92–104, 2007.

[142] M. Augstein, E. Herder, and W. Wörndl, *Personalized human-computer interaction.* Walter de Gruyter GmbH & Co KG, 2019.

[143] J. R. Finkel and C. D. Manning, "Hierarchical bayesian domain adaptation," in *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pp. 602–610, 2009.

[144] M. M. Afsar, T. Crump, and B. Far, "Reinforcement learning based recommender systems: A survey," *arXiv preprint arXiv:2101.06286*, 2021.

[145] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[146] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, 2018.

[147] A. T. Sabin and B. Pardo, "A method for rapid personalization of audio equalization parameters," in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 769–772, 2009.

[148] J. B. Nielsen, B. S. Jensen, T. J. Hansen, and J. Larsen, "Personalized audio systems—a bayesian approach," in *Audio Engineering Society Convention 135*, Audio Engineering Society, 2013.

[149] J. B. Nielsen, J. Nielsen, B. Sand Jensen, and J. Larsen, "Hearing aid personalization," 2013.

[150] C. Guezenoc and R. Seguier, "Hrtf individualization: A survey," *arXiv preprint arXiv:2003.06183*, 2020.

[151] N. Zhao, A. Azaria, and J. A. Paradiso, "Mediated atmospheres: A multimodal mediated work environment," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–23, 2017.

[152] P. I. Frazier, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

[153] B. Givan and R. Parr, "An introduction to markov decision processes," *Purdue University*, 2001.

[154] D. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling," *arXiv preprint arXiv:1707.02038*, 2017.

[155] J. Schneider and S. Kirkpatrick, *Stochastic optimization*. Springer Science & Business Media, 2007.

[156] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," 2019.

[157] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," 2019.

[158] W. A. Bainbridge, P. Isola, and A. Oliva, "The intrinsic memorability of face photographs.," *Journal of Experimental Psychology: General*, vol. 142, no. 4, p. 1323, 2013.

[159] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva, "Modifying the memorability of face photographs," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3200–3207, 2013.

[160] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2390–2398, 2015.

[161] T. Head, MechCoder, G. Louppe, I. Shcherbatyi, fcharras, Z. Vinícius, cmmalone, C. Schröder, nel215, N. Campos, T. Young, S. Cereda, T. Fan, rene rex, K. K. Shi, J. Schwabedal, carlosdanielcsantos, Hvass-Labs, M. Pak, SoManyUsernamesTaken, F. Callaway, L. Estève, L. Besson, M. Cherti, K. Pfannschmidt, F. Linzberger, C. Cauet, A. Gut, A. Mueller, and A. Fabisch, "scikit-optimize/scikit-optimize: v0.5.2," Mar. 2018.

[162] I. Ananthabhotla, D. B. Ramsay, and J. A. Paradiso, "Cognitive Content Curation: An Audio Summarization Tool Driven by Principles of Auditory Cognition," *Under Review*.

[163] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, 2017.

[164] I. Ananthabhotla and J. A. Paradiso, "Visualsoundtrack: An approach to style transfer in the context of soundtrack prototyping," in *ICMC*, 2017.

[165] G. Milner, *Perfecting Sound Forever: An Aural History of Recorded Music*. Le Castor Astral éditeur, 2017.