

Listening with generative models

by

Maddie Cusimano

Submitted to the Department of Brain and Cognitive Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author
Department of Brain and Cognitive Science
June 12, 2022

Certified by
Josh H. McDermott
Associate Professor
Thesis Supervisor

Accepted by
Mark Harnett
Chairman, Department Committee on Graduate Theses

Listening with generative models

by

Maddie Cusimano

Submitted to the Department of Brain and Cognitive Science
on June 12, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis extends classic traditions in perception by leveraging contemporary tools to build and apply rich generative models that describe what we hear. First, I present a hierarchical Bayesian auditory scene synthesis model to address the perceptual organization of sound into sources and events. We aimed to bridge between classical auditory scene analysis phenomena and everyday sounds, asking whether common generative principles could explain auditory scene analysis in both cases. We tested the model by having it listen to a variety of auditory scene analysis illusions and found that its judgments matched those of human listeners. Applied to everyday sounds, the model infers valid perceptual organizations. Also, due to its interpretability, the model’s failures with everyday sounds were informative: they reveal the necessity of peripheral representations of periodicity, a more expressive model of spectra, and sources that compose multiple sound-generating processes. The next projects address alternative scene analysis problems of everyday physical understanding from sound. We developed methods for the ecological sound synthesis of a set of common object interactions: brief impact sounds and sustained scraping and rolling sounds. Our synthesis combines physical simulation from perceptually relevant variables with a statistical model of material. Listeners perceive our synthesized sounds to be realistic and as conveying various physical variables. I discuss future directions for developing inference for these physics-inspired models, learning sound synthesizers, and generating illusions. Given the variety of structured latent-variable generative models investigated through these projects, I conclude by exploring how multiple world models might interact in perception.

Thesis Supervisor: Josh H. McDermott
Title: Associate Professor

Acknowledgments

Why am I so lucky? Thank you first to my advisor Josh McDermott. It has been such a pleasure to listen to the world together. You are an exemplar of staying grounded through chaos and of how to continually grow in service of supporting those around you. It has been an honour to work with my committee, whose research and mentorship directly inspired this thesis: Josh Tenenbaum for introducing me to a language to describe experience, Ted Adelson for phenomenological and philosophical explorations, and Dan Ellis for a balance of positivity and challenging questions. So many other scholars have supported me, to name a few: at the University of Toronto, thank you to Claude Alain, Mark Taylor, Farzaneh Hemmasi, and Jesse Gell-Redman for supporting me to come to graduate school; and to Andrea van Doorn, Jan Koenderink and Seth Riskin, for the inspiration to do science motivated by how it is to see.

The Lab for Computational Audition is the only place where this thesis could have sprouted, first through its alchemy of Wiktor Młynarski's passionate clarity, Max Siegel's encyclopedic knowledge of cognitive science, James Traer's equally encyclopedic knowledge of acoustics, and Richard McWalter's investigations into auditory scene analysis (not to mention his kindness). It has also been the perfect place to wrap it up: thank you to Andrew Franci for always understanding where I'm at, Vinayak Agarwal for your lighthearted and thoughtful collaboration, Malinda McPherson for never giving up on getting the framing right, Jenelle Feather for helping me to consider computational modeling from a more integrative viewpoint, Ajani Stewart for fresh playlists, Bryan Medina for chill river-walks, and Fernanda De La Torre for encouraging me to speak from what I love. Thanks to the whole lab for shaping my work and for all the funny sounds.

I'm also grateful to the wider communities that raised me, including Kelsey Allen who made MIT feel like home and helped me see its wide open possibility, Matthias Hofer and Tyler Brooke Wilson who created so many spaces to connect; and Alex Lew who taught me so much about Bayesian inference while also helping me feel welcome and capable in that space. To Rodrigo Ochigame, Ara Ortiz, Joaquim de Souza, Kate Powell, Dre Cetra, Georgia Hewitt, Cathy Hoffman, Amanda Nash, Taymar Pixleysmith, John Churchill, Dustin DiPierna, Dan Brown, SJ Klein, Teresa Yeh, Irene Zhou, and Lucas Foglia: thank you for the unexpected adventures. Thank you to the Earth Species Project for reminding me the world is utterly sonically alive.

Thank you to my family for their constancy: to my parents for letting me read everything when I was younger; to my siblings, for holding all the things I tend to drop; and to my nonni, especially to my Nonna Carmelina for the espresso pot that helped me finish this thesis.

Sarah, Dae and Luke: here's to many more epiphanies.

Thank you to all the musicians who kept me going while I wrote this thesis. To the rivers and lakes that kept us cool, and to the creatures who kept us company and kept it simple.

Contents

1	Introduction	17
1.1	Experimental phenomenology	18
1.2	Signal synthesis	21
1.3	Bayesian generative models	22
1.4	Contributions	24
2	Bayesian auditory scene synthesis	29
2.1	Introduction	29
2.2	Results	33
2.2.1	Overview	33
2.2.2	Generative model	34
2.2.3	Inference	42
2.2.4	Model results on classic ASA phenomena	44
2.2.5	Model comparisons	57
2.2.6	Model results on everyday sounds	62
2.3	Discussion	68
2.3.1	Relation to prior models	69
2.3.2	Limitations	72
2.3.3	Future directions	74
2.3.4	Inference	77
2.3.5	Other sensory modalities	81
2.4	Supplementary Figures	82
2.5	Methods	94

2.5.1	Model	94
2.5.2	Inference	106
2.5.3	Hypothesis optimization and comparison	106
2.5.4	Sequential inference	108
2.5.5	Enumerative inference for simulated psychophysical experiments	114
2.5.6	Classic ASA phenomena	116
2.5.7	Network comparisons	133
2.5.8	Model alternatives	142
2.5.9	Everyday sound experiments	142
2.6	Open source media credits	148
3	A statistical model of material for synthesizing contact sounds	149
3.1	Introduction	150
3.2	Source-filter model of impacts	154
3.2.1	Modal synthesis of object Impulse Responses (IRs)	155
3.2.2	Effect of impact physics	156
3.3	Perception of synthetic impacts	157
3.3.1	Experiment 1. Realism of synthetic impact sounds	157
3.3.2	Experiment 2. Perception of material	159
3.3.3	Experiment 3. Perception of mass	160
3.4	Sustained contacts	161
3.4.1	Contact force for sustained contacts	162
3.4.2	Variation of IRs over contact location	164
3.5	Perception of synthetic scraping	165
3.5.1	Experiment 5. Realism of synthetic scraping sounds	166
3.5.2	Experiment 6. Perception of motion	166
3.6	Discussion	168
3.7	Conclusion	170
4	Synthesizing sustained contact sounds	171
4.1	Introduction	171

4.2	Scraping Sound Synthesis Model	174
4.2.1	Contact force for scraping	174
4.2.2	Trajectory of the scraping object	175
4.2.3	Effect of macroscopic normal force	177
4.2.4	Morphing impulse responses	179
4.3	Perception of Synthetic Scraping	180
4.3.1	Experiment 1. Realism of synthetic scraping sounds	181
4.3.2	Experiment 2. Perception of motion from scraping sounds	182
4.4	Rolling Sound Synthesis Model	183
4.4.1	Contact force for rolling	185
4.4.2	Impulse responses	187
4.5	Perception of synthetic rolling	187
4.5.1	Experiment 3. Realism of synthetic rolling sounds	187
4.6	Discussion	188
4.7	Conclusion	191
5	Conclusion	193
5.1	Methodology behind generative models	193
5.1.1	Enabling factors	193
5.1.2	Challenges	195
5.1.3	Imagining new starting points	197
5.2	Themes for perception	200
5.2.1	More on structured generative models	200
5.2.2	More on illusion generation	201
5.2.3	Integrating structure and statistics	202
5.2.4	A multiplicity of world models in perception	204

List of Figures

1-1	Everyday examples of auditory scene analysis. Each row corresponds to a different 4-second audio clip that I recorded just before or during graduate school. The first column displays sound pressure waves, which are picked up by our auditory system when they cause the ear drum to vibrate. The second column contains my description of the sound, with a focus on what I perceive to be ‘streams’ of sound arising from distinct sources. Top: Termini Imerese, Sicily, 2015. Middle: Cambridge, Massachusetts, 2016. Bottom: Toronto, Canada, 2021. Sound examples at https://mcdermottlab.mit.edu/mcusi/thesis/ . . .	19
2-1	Examples of generative models.	31
2-2	Components of the generative model, illustrated with everyday recorded sound examples.	38
2-3	Prior distribution $P(S)$, illustrated with recorded natural sound examples.	39
2-4	Model results on masking and filling-in phenomena.	49
2-5	Model results on simultaneous grouping phenomena.	52
2-6	Model results on sequential phenomena.	56
2-7	Comparisons with alternative models on classic ASA phenomena. . .	61
2-8	Model results on everyday sounds	64
2-9	Event structure as implemented by the non-stationary kernel. . . .	82
2-10	Scenes sampled from prior $P(S)$, rendered as cochleagrams.	83
2-11	Example of sequential inference algorithm with tone sequence. . . .	84
2-12	Example of sequential inference algorithm with frequency modulation.	85

2-13	Experiment 2 data per pre-mixture sound (n=7).	86
2-14	Examples of recognizable source inferences in Experiment 1.	87
2-15	Examples of unrecognizable source inferences in Experiment 1, due to periodicity mismatch in the mixture and inferred source.	88
2-16	Examples of inferences when there is a quiet tone in noise.	89
2-17	Oversegmentation deviations that occur when various sound-generating processes within a single premixture clip have some common causal factor, involving sequences.	90
2-18	Oversegmentation deviations that occur when various sound-generating processes within a single premixture clip have some common causal factor, involving excitations that occur simultaneously.	91
2-19	Impact sounds illustrate issues with model spectrum and amplitude (1).	92
2-20	Impact sounds illustrate issues with model spectrum and amplitude (2).	92
2-21	Inference errors due to the local nature of stochastic gradient descent.	93
3-1	We synthesize sounds by (top) a generative model of impact and sus- tained contacts. (Upper-middle) Object Impulse Responses are syn- thesized by sampling modes from empirical distributions. (Lower- middle) Impact forces are modelled via a spring model. (Bottom) Sustained contacts are modelled via measured surface textures and location-dependent IRs.	153
3-2	Discrimination of real vs. synthetic impact sounds (Exp 1). Dashed line denotes chance performance.	158
3-3	Material discrimination from synthetic impact sounds (Exp 2). Left: Confusion matrices of the presented material and participant responses. Right: Correlation of the confusion matrices of various synthetic sounds with that of the recorded impacts.	160
3-4	Mass discrimination with real and synthetic impact sounds (Exp 3). .	161

3-5	Everyday textures measured with the confocal microscope. Surface area is 7.3 mm by 10 mm. From-left: 100 grit sandpaper; 60 grit sandpaper; wood; vinyl tile.	162
3-6	Scraping motions. Left: Measured position traces of scraper over surface, for three different types of motion. Right: Absolute velocity measurements.	163
3-7	Discrimination of real vs. synthetic impact sounds scraping sounds (Exp 5). Dashed line indicates chance performance.	167
3-8	Motion discrimination from synthetic scrape sounds (Exp 6). (Left) Confusion matrices of presented motion pattern and the human responses. (Right) Correlations of the confusion matrices of synthetic sounds with the correlation matrix of recorded sounds.	167
4-1	Complete scraping synthesis model. Yellow boxes: inputs to model. Green boxes: intermediate representations computed from inputs. Blue box: sound waveform computed from contact force and net impulse response.	173
4-2	Trajectory S of the scraper point. The trajectory is determined by the surface depth profile ($z(x)$, shown in grey) and the normal force N . Larger normal forces (red) produce more extreme scraper trajectories than smaller normal forces (yellow).	176
4-3	Macroscopic forces on the scraper when undergoing human-produced simple harmonic motion (as when scraping a held object back and forth on a surface). The normal force (red vector) varies with position due to the variation in the applied force $F_{applied}$	178

4-4	Experiment 1: Realism of synthesized scraping. A) Participants rated the realism of 7 different renderings of an object scraped back-and-forth over a surface, using a MUSHRA paradigm. Each rendering was via a different synthesis method. Participants were given a text description of the scraping event. B) Results of Experiment 1, showing mean realism for each synthesis method. Error bars plot SEM.	182
4-5	Experiment 2: Motion recognition from scraping. A) Participants listened to a sound and selected the path traced by the scraping object. B) Sounds were generated for each of five possible motions. C) Confusion matrices for each of the eight conditions. D) Similarity of the confusion matrix of each synthesized condition compared to that for the recorded sounds.	184
4-6	Synthesis of contact force for rolling. Yellow boxes are inputs to the synthesizer, which are combined to yield the contact force. The contact force is combined with location-specific impulse responses in the same manner as for scraping (shown in Figure 1).	185
4-7	Experiment 3: Realism of synthesized rolling. A) Participants rated the realism of 7 different renderings of a ball rolling down or up an inclined surface, using a MUSHRA paradigm. Each of the 7 renderings was from a different synthesis method. The ball and surface varied in material. B) Results of Experiment 3, showing mean realism for each synthesis method. Error bars plot SEM.	188

List of Tables

2.1	Classic auditory scene analysis experiments and illusions	45
2.2	Discrete priors and temporal normal-gamma hyperprior. The temporal hyperprior is used for both the rest and duration latent variables. . .	101
2.3	Gaussian process hyperprior parameter values. The mean μ is uniformly distributed (with units listed under GP type). The inverse soft-plus of the variance σ and lengthscale ℓ are each normally distributed. Samples of σ and ℓ are bounded to reasonable values. σ is in the GP-units, while ℓ is in seconds for excitation variables and in ERB for spectrum. β is a positive constant that implements the non-stationary kernel. As is typical for GPs, ϵ^2 is added to the diagonal of the covariance matrix for numerical stability. The last three columns indicate the first, second, and third quartile of sampled σ and ℓ distributions based on 5000 samples. *Note: We manually set the value of β to 1.0 because the scenes in the textures dataset only had one event. . . .	102
5.1	Perceptual objectives for illusion generation. S_e is the scene description that encodes the illusory sound X . S_p is the perceived scene description.	202

Chapter 1

Introduction

We often experience our sensory fields as composed of a multiplicity of distinct wholes. When we listen, we perceive streams of sound that seem to arise from distinct sources. Through our felt senses, we experience our body as distinct from the surfaces of our surroundings. And even when we look into a uniform field of light or a simple array of regularly spaced dots, our visual field constellates into various forms. This last example suggests that the forms we perceive are not evident solely in the energy picked up by our sensory receptors. How then do our perceptual systems structure experience?

This emphasis on perceiving ‘structured wholes’, or *perceptual organization*, is usually traced back to Gestalt psychology, a major branch of psychology in the early twentieth century (Wagemans, 2015). The various schools of Gestalt psychology focused on describing perception in terms of holistic units that people naturally perceived, rather than the ‘elementary sensations’ of structural psychology. Their methodology, called experimental phenomenology (Albertazzi, 2015; Koenderink, 2015; Thines et al., 2013), led to an abundance of discoveries about structured wholes in perception.

Whereas the foundations of Gestalt psychology were essentially phenomenological, about the structure of visual awareness in itself, more recently perceptual organization has been cast as scene analysis (Bregman, 1994). Scene analysis refers to the perceptual process of analyzing a field of sensory data in terms of causes (Lewicki et al., 2014). That is, the perceptual system constructs an explanation of what causes

generated the observed sensory data, and these causes presumably correspond to the meaningful cohesive structures that we perceive. This perspective has emphasized the connections between perceptual organization and the physical and ecological processes that generate and structure sensory data, and thus has increasingly focused on complex, naturalistic signals.

Still, a key part of understanding perceptual organization continues to be: what are the structured wholes that we perceive? In sensory ecology, one might say this question means characterizing the carriers of significance' in an organism's Umwelt, their perceptual world (Agamben, 2004). In Marr's framework for vision (Marr, 1982), this question is echoed in his "computational level": what are the goals of perception, its 'output'? And for Gibson's ecological approach, it might imply identifying the invariants that allow constancy in perception (Warren, 2021). This thesis takes as its case study the perceptual organization of our intricate and lively world of sound, exploring computational frameworks for describing the world we hear. In the rest of the Introduction, I will briefly expand upon three strands of research – experimental phenomenology, signal synthesis, and Bayesian generative models – focusing on auditory perception where possible. Each treat the perception of structured wholes from a unique perspective that has influenced this thesis. I will end by outlining our approach and contributions.

1.1 Experimental phenomenology

Much of the classic work in visual perceptual organization comes from Gestalt psychology. Central to their approach were compelling abstract images for which putative perceptual principles could be directly pointed out in an observer's experience. These images experientially illustrated holistic, natural language visual concepts such as groups, contours, objects, surfaces, and backgrounds (Nakayama et al., 1995). These percepts typically involve spatial regions of the visual field that cohere as a meaningful whole, as well as their relations in depth. One rich example that illustrates many of these percepts is figure-ground organization, which refers to the perception

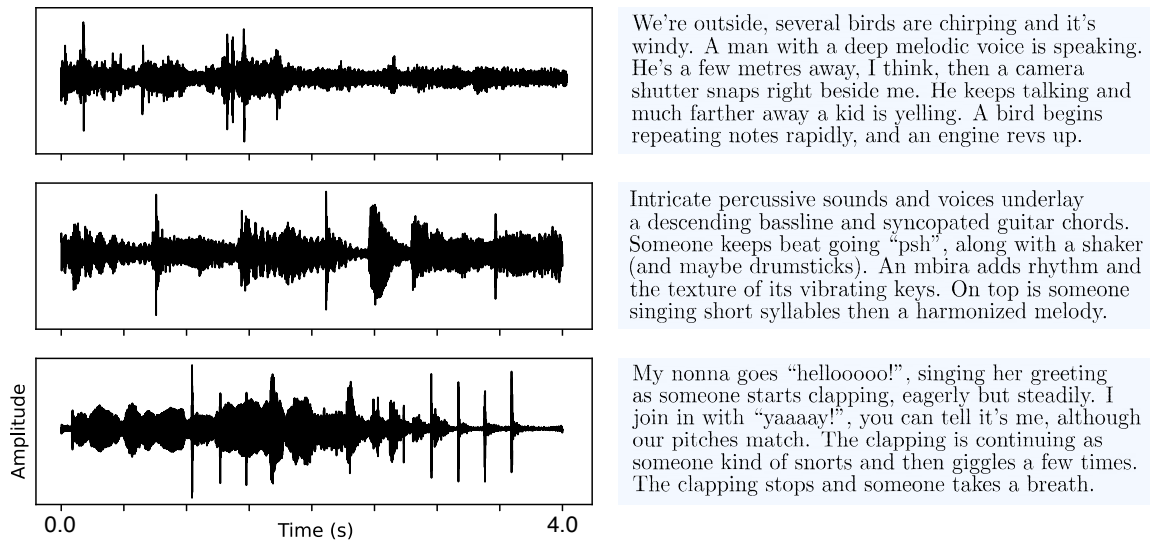


Figure 1-1: Everyday examples of auditory scene analysis. Each row corresponds to a different 4-second audio clip that I recorded just before or during graduate school. The first column displays sound pressure waves, which are picked up by our auditory system when they cause the ear drum to vibrate. The second column contains my description of the sound, with a focus on what I perceive to be ‘streams’ of sound arising from distinct sources. Top: Termini Imerese, Sicily, 2015. Middle: Cambridge, Massachusetts, 2016. Bottom: Toronto, Canada, 2021. Sound examples at <https://mcdermottlab.mit.edu/mcusi/thesis/>.

of one region (a ‘figure’) in front of an adjacent region that appears as a background. The background appears to continue behind the occluding surface of the figure, called ‘amodal completion’. As such, the boundary between the two regions appears as a contour that defines the shape of the figure rather than the background. Various image factors, such as the size of the regions and their convexity, affect which region is seen as the figure. Similar factors were discovered and articulated for visual grouping of image elements, in what are known as the Gestalt principles of perceptual grouping. For example, these include the principles of proximity, common fate, similarity, and good continuation. The kinds of structured wholes explored in this way also extended beyond relatively simple visual groups. Later, Michotte used experimental phenomenology to argue that foundational structured wholes in vision included animacy and causality (Scholl & Tremoulet, 2000; Thines et al., 2013). For instance, when a shape moves and then stops upon contacting a second shape, which begins to move, observers perceive a causal relationship of the first shape launching the second

into motion.

The Gestalt approach for visual perceptual organization heavily influenced work on the perceptual organization of sound. In his seminal book on auditory scene analysis, Bregman (1994) defined a holistic distinct auditory entity, termed a ‘stream’, analogously to visual groups: as regions of a “neural spectrogram” that cohere as a whole. Bregman (2005) wrote about the importance of a phenomenological approach, and both the auditory phenomena and corresponding principles were similar in spirit to the Gestalt ones: carefully designed synthetic sounds that illustrated natural language principles describing how sounds were grouped into streams. But for all these methodological similarities, Bregman advocated for understanding perceptual organization as scene analysis, identifying limitations in the Gestalt approach. He interpreted the grouping principles as an instance of ‘psychophysical complementarity’ (Shepard, 1981) with ecological validity (e.g., early work in visual scene statistics, Brunswik and Kamiya, 1953); that is, as adaptations of the auditory system to the physical and ecological processes that generate sound. Therefore, Bregman presented the grouping principles along with intuitions about sound generation in the world. For example, the principle of common onset reflects the intuition that a single sound-producing event can produce many sound components simultaneously. Therefore, if such components are detected, they should be grouped.

The strength of the Gestalt approach is evident by its continuing broad influence (Brooks, 2015). Researchers were able to discover a diverse set of foundational phenomena in both vision and hearing. My perspective is the influence comes from (1) that their visual concepts could be distilled into a simple, abstract image to make them directly accessible in experience. Beyond their compelling immediacy and ability to spark curiosity, this distillation seems to suggest the relevance and robustness of the putative principles for perception, and (2) the corresponding experiential and conceptual interpretability of the natural language principles encourage scientific analogy-making and intuitive connections to other domains. However, as recognized by Bregman and others, these natural language principles also had their limitations. Although the phenomena seem to be suggestive of principles underlying everyday per-

ceptual organization and could be intuitively connected to causal processes, natural language principles could not obviously be applied to complex or everyday sounds. Their open-endedness and imprecision could even obscure difficulties and contradictions in doing so. Subsequent work in computational auditory scene analysis sought to close this gap, usually by implementing grouping principles to cluster bottom-up features of audio. As expanded upon in Chapter 2, in practice these attempts revealed the difficulty of this approach, leaving the issue unresolved.

1.2 Signal synthesis

Moving beyond the pure tones and white noise bursts of the classic perceptual organization demonstrations in hearing, the synthesis of complex signals has been used to investigate more naturalistic percepts. In his 1993 papers on ecological perception, “What do we hear in the world?” and “How do we hear in the world?”, Gaver presents a physics-inspired approach to sound synthesis as a method for discovering what meaningful entities we hear. He proposes to describe the structured processes in how sounds are physically generated and to build compelling sound synthesis algorithms explicitly based on those processes. To illustrate his approach, Gaver (1993b) proposes a hierarchical sound ontology. At the first level, the ontology has different sound-producing states of matter: air sounds, water sounds, sounds of interacting solids. The second level described the temporal manner of sound production (e.g. dripping as a brief water sound versus flowing as a continuous water sound; impact as a brief solid interaction versus scraping as a continuous solid interaction). He also included various kinds of compositions to build from these levels to more complex sounds, such as the sound of stirring liquid as combining brief solid impacts and continuous flowing water sounds. To formalize this physically-inspired natural language ontology, Gaver (1993a) then presents physics-inspired sound synthesis algorithms for various aspects of this ontology (e.g. dripping sounds). He describes an iterative process of refining the synthesis algorithms, determining which causal factors are necessary for expressive synthesized sounds and which simplifications maintain the

synthesized sounds’ perceptual fidelity. He proposes that this process can reveal what structure is perceived in complex, naturalistic sounds.

More recent examples of applying signal synthesis to auditory perception have innovated extensions and alternatives to Gaver’s methodology. Traer and McDermott (2016) synthesized reverberation with real-world statistical regularities to reveal how those regularities enabled listeners to have distinct percepts of sources and space. In another interesting counterpoint to Gaver’s physics-based sound synthesis, McDermott and Simoncelli (2011) developed a realistic synthesis of sound textures from the time-averaged statistics of peripheral auditory representation. While not using physics-based representations, the process of testing the synthesis algorithm is similar to Gaver’s and demonstrates the centrality of statistical representations for many of the continuous sounds that Gaver was concerned with.

These three examples demonstrate the strengths of using sound synthesis to investigate complex, everyday sounds. Compared to the Gestalt approach, it tends to take more direct inspiration from acoustics in order to propose meaningful entities in perception. The diversity of causal factors in sound generation have not been studied in classic perceptual organization, which has mainly studied the perception of distinct sources. Designing an effective sound synthesis algorithm is non-trivial. While very physically-detailed sound synthesis methods exist for some kinds of sounds (e.g. finite element models), they do not present an obvious route to refining a model of perception as described by Gaver. Also, a sound synthesis algorithm on its own does not provide a model of perception, which can take a sound as input and returns a percept.

1.3 Bayesian generative models

Bayesian generative models of perception emphasize the structured processes that generate sensory data, while providing a formalism to infer ‘scene descriptions’ from data (Kersten & Schrater, 2002). At its core, a generative model describes how a set of random latent variables (the scene, S) interact to generate sensory data (X).

Defining prior probabilities $P(S)$ as well as the probability of generating data from a certain scene, the likelihood $P(X|S)$, is the basis for approximating the probability of a scene given observed data, the posterior distribution $P(S|X)$.

In a related spirit to Gaver’s work, early work in Bayesian models of perception used physical models of optics to describe image generation. This work demonstrated that structured interactions between causal variables in optics is reflected in structured percepts. A classic example from Bloj et al. (1999) contains a model of how shape, illumination, and surface reflectance interact to produce light with a particular luminance and hue, and shows that human judgments of colour accord with Bayesian inference in this model. It has also been argued that the structures explored in generative models of perception have focused too narrowly on considering concepts from optics, and that learning generative models from natural data may provide insight into perceptually relevant latent factors (Fleming & Storrs, 2019).

Bayesian inference has also been applied to perceptual organization problems, where the structure of a scene (i.e., which causes are present) must also be inferred. In vision, Weiss (1998) presented a comprehensive model aimed at estimating multiple groups of motion in an image, called ‘smoothness in layers’. The scene descriptions consist of a number of surfaces, each with a motion that tends to be slow and smooth. The surface motions generate a motion vector at each pixel, which can then be used to generate the intensity values of a subsequent frame. This model provided a unified understanding for a number of previously disparate classic results in visual motion perception. Weiss also applied this framework to some natural video sequences (Weiss, 1997). Other work has addressed other aspects of visual perceptual organization with similar principles, although typically for abstract images or with symbolic input (Froyen et al., 2015; Gershman et al., 2016; Körding et al., 2007; Lake et al., 2015; Zhu, 1999). In hearing, Bayesian generative models of audio date back to the earliest days of computational auditory scene analysis (Weintraub, 1985; for a review, see Ellis, 2006).

However, the applications of Bayesian inference to human perception have been limited by the practical intractability of inferring the posterior distribution (Ellis,

2006). Inference typically involves search over the space of scene descriptions, which can be vast for expressive models. The kinds of scenes considered are thus often chosen to be more simplistic because constraining the scene descriptions can make inference more tractable. With some exceptions (e.g., Lake et al., 2015; Weiss, 1998), prior applications of Bayesian inference to human perception have been limited to few- or fixed-dimensional domains (e.g., Bloj et al., 1999; Fischer and Peña, 2011; Knill and Saunders, 2003; Kulkarni et al., 2015; Saunders and Knill, 2001; Stocker and Simoncelli, 2006; Turner, 2010; Weintraub, 1985; Weiss et al., 2002; Yildirim et al., 2016), operated on symbolic data rather than actual sensory signals (e.g., Barniv and Nelken, 2015; Froyen et al., 2015; Gershman et al., 2016; Körding et al., 2007; Larigaldie et al., 2021), or faced intractable inference issues that prevented them from being fully evaluated (e.g., Ellis, 1996; Nix and Hohmann, 2007).

Despite these challenges, there are many strengths of a Bayesian modeling approach to perception. This approach obviously complements the signal synthesis approach and I think it could also provide a good formal complement to experimental phenomenology, due to its interpretability in describing structured wholes (i.e., world models). In practice, Bayesian models have provided principled unifying accounts of disparate phenomena. Bayesian inference provides a flexible formalism to understand perception in terms of various kinds of generative processes. We discuss more benefits of this approach in Chapter 2.

1.4 Contributions

The overarching contribution of this thesis is to extend these classic traditions by leveraging contemporary tools to build and apply rich generative models of audio that describe what we hear. The two broad challenges are (1) designing generative models that account for human perception and (2) making search in Bayesian inference tractable. As just outlined, many of the principles implemented in this thesis can be traced to classical ideas in vision, where inference in generative models has long been proposed to underlie perception, but where computational systems that

embody this approach have historically been a challenge to make work. The study of auditory perception shares many of the same deep themes; yet, our intuition was that relatively simple generative models can account for many everyday sounds, making the approach more tractable.

This is lucky for me because I really love sound. I started graduate school by recording the sounds of my everyday life, listening to large corpora of soundscapes (Truax, 1978) and documenting what I heard. This eventually morphed into wide-ranging explorations of sound-generating processes, with much time spent throwing found objects down staircases in Building 46. Upon being introduced to computational approaches in perception, these projects crystallized into the various flavours of generative world descriptions that we explore and combine here (signal-based, physics-inspired, and statistical as detailed more below; cf. deep vs. shallow representations in Morgenstern and Kersten, 2017).¹

Chapter 2, on **Bayesian auditory scene synthesis** (BASS), begins by addressing classic perceptual organization problems in hearing. Inspired by prediction-driven computational auditory scene analysis (Ellis, 1996), scenes are modelled as comprising a set of sources, with each source emitting events to produce sounds (Figure 2-2). I describe our excitation-filter source models as having ‘signal-based’ latent variables, in that they are abstract signal attributes (e.g. amplitude) that apply to a wide range of sounds, rather than being based on specific physical processes. We aimed to bridge between perceptual organization in illusions and the perceptual organization of everyday sounds, asking whether common generative principles could explain auditory scene analysis in both cases. We tested the model by having it listen to a variety of classic auditory scene analysis phenomena and found that its judgments matched those of human listeners. We found that neural networks that lack structured constraints could not explain human perception to the same extent (Figure 2-7). When we applied the generative model to everyday sounds, it could often infer perceptually plausible perceptual organizations (Figure 2-8,2-14). Due to the interpretability

¹I also see these flavors as related to Gaver’s distinction between ‘musical listening’ (listening to sounds) and ‘everyday listening’ (listening to the world).

of the model, its failures were instructive: for instance, they point to the potential perceptual relevance of composite, causally-linked sound-generating processes in perceptual organization. This work, completed in collaboration with Luke Hewitt, is currently being prepared for journal submission.

The next chapters address some of the limitations of the BASS model by developing ecological sound synthesis of object interactions. Chapters 3 and 4 address alternative scene analysis problems of everyday physical understanding from sound, and are influenced by Gaver’s physics-inspired sound synthesis. Chapter 3, on **A statistical model of material for synthesizing contact sounds**, presents a statistical model of object material based measurements from recorded contact sounds. We use this model in combination with a physics-based model of impacts and scraping to synthesize contact sounds (Figure 3-1). We find that human listeners can infer the latent variables of our model (e.g. material, mass, motion) from our synthesized sounds better than alternatives. This work is a collaboration with James Traer, and is published as a conference paper at DAFx 2019 where I am a second author. Chapter 4, on **Synthesizing sustained contact sounds**, presents improvements to the scraping model and introduces a synthesis of rolling sounds (Figure 4-1,4-5). This work is a collaboration with Vinayak Agarwal and James Traer, and is published as a conference paper at DAFx 2021 where I am a second author. My primary contributions to this body of work include developing the scraping synthesis algorithm and how it composed with the material model. The ongoing plan for this body of work is to develop a rich sound synthesis toolkit with which we can synthesize causally evocative sound sequences, along with methods for Bayesian inference. Unfortunately, the effort of making inference succeed in the BASS model meant I set aside the task of inference in the physics-inspired models. Some preliminary work that models scene analysis with reverberant impact sounds is reported in Hu et al. (2019).

With these projects, I feel I haven’t gotten close to describing the world I hear. But, by pushing the classic idea of generative models of perception as far as I could, I hope to offer tangible computational systems that inspire debate, experiment, and iteration. In this spirit, the conclusion in Chapter 6 covers the ‘behind-the-scenes’

methodology of building generative models, relationships between perceptual systems, and peregrinations through themes for perception.

Chapter 2

Bayesian auditory scene synthesis

This project is a collaboration with Luke Hewitt.

We often experience sound as multiple streams arising from distinct sources, such as a cat purring amid the pitter-patter of a rainy day. Yet this meaningful structure is not explicit in the sound data that arrives at the ear. Illusions generated with synthetic sounds suggest that general principles underlie auditory perceptual organization. However, the form of such principles and whether they apply to everyday listening remains unclear. Here, we formalize candidate principles as generative constraints in a Bayesian analysis-by-synthesis model. The model accounts for perceptual judgments on a variety of classic illusions. It can also infer perceptually valid sources from audio recordings. In contrast, an assortment of contemporary deep neural networks which succeed on naturalistic source separation tasks do not “hear” illusions like humans. These results illustrate the explanatory strength of the analysis-by-synthesis approach for auditory perception.

2.1 Introduction

Our perceptual experience is comprised of meaningful structures, from the surfaces and bodies we see comprising our visual field, to streams of sound that we hear as arising from distinct sources. But the incoming sensory data alone does not specify the structure we perceive. Illusions show that out of many possible ways to interpret

sensory inputs, human observers tend toward particular perceptual organizations. To bridge between ambiguous sensory data and meaningful perceptual structure, our perceptual systems must impose constraints on perceptual organization.

Gestalt principles were an early approach to characterizing constraints on visual perceptual organization (Brooks, 2015), with analogues in auditory perception (Barker et al., 2005; Cooke & Ellis, 2001). These principles typically describe how features of sensory data constrain grouping. For instance, in auditory perception, the principle of “common onset” states that sound components that begin at the same time should be grouped. In modern incarnations of the Gestalt approach, (Bregman, 1994) many grouping cues were conceived as part of “scene analysis”, the process of analyzing a field of sensory data in terms of its causes, ecologically defined (Adelson & Pentland, 1996; Lewicki et al., 2014). In this view, perceptual constraints are enabled by causal processes that generate sensory data, which manifest regularities as a result. For example, the principle of common onset reflects the intuition that a single event would produce many sound components simultaneously, leading to a regularly recurring feature in sound – so if such components are detected, they should be grouped. These verbally-stated principles can be made precise with computational implementations that detect the cue features in sensory signals (Cooke & Ellis, 2001; Elder & Goldberg, 2002; Ellis, 1994; Field et al., 1993; Fowlkes et al., 2007; Geisler et al., 2001; Krishnan et al., 2014; Młynarski & McDermott, 2019). Such systems have been used to explain targeted perceptual phenomena but in practice fall short of accounting for perceptual organization (Froyen et al., 2015), reflective of the difficulty in specifying particular features that are predictive of perception. For instance, global scene interpretations often alter how locally computable features are perceived (Bloj et al., 1999; Bregman, 1978b; Knill & Kersten, 1991; McDermott, 2004; McDermott et al., 2001; Warren, 1970).

An alternative to specifying diagnostic cue features is to directly specify constraints on how sensory data is generated. Generative constraints can be expressed as signal synthesis models (Figure 2-1): for instance, a model of how a plucked string produces multiple simultaneous audio frequencies related by the harmonic series, or of

how light reflects off curved surfaces to create shading and shadows evident as luminance gradients in the resulting image. Scene analysis can then be explicitly defined in terms of causes specified in perceptual generative models. Generative constraints are brought to bear through Bayesian inference, which provides the mathematical basis for inverting a generative model to infer latent causes from data.

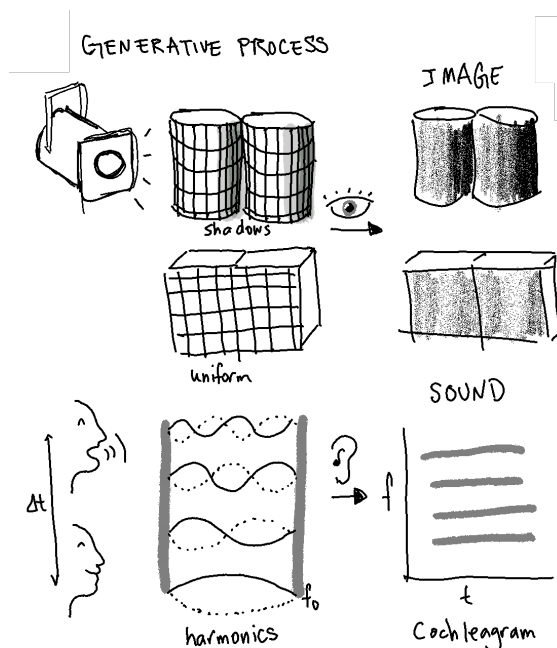


Figure 2-1: Examples of generative models.

Generative models hold promise for overcoming the limitations of grouping cues (Ellis, 1996) because they determine how combinations of causes (the “global” interpretation of a scene) impact the presence of local features. However, such inference typically involves a search through the space of possible causes to find those that are plausible given the data. In practice, this search is computational intractable for all but the simplest models, which has limited their scope for explaining human perception. With few exceptions (Lake et al., 2015; Weiss, 1998), prior applications of Bayesian inference to human perception have been limited to few- or fixed-dimensional domains (Bloj et al., 1999; Fischer & Peña, 2011; Knill & Saunders, 2003; Kulkarni et al., 2015; Saunders & Knill, 2001; Stocker & Simoncelli, 2006; Turner, 2010; Weintraub, 1985; Weiss et al., 2002; Yildirim et al., 2016), operated on symbolic data

rather than actual sensory signals (Barniv & Nelken, 2015; Froyen et al., 2015; Gershman et al., 2016; Körding et al., 2007; Larigaldie et al., 2021), or faced intractable inference issues that prevented them from being fully evaluated (Ellis, 2006; Ellis, 1996; Nix & Hohmann, 2007).

Here we consider auditory scene analysis (ASA) as a case study in perceptual organization. Despite many proposed conceptual approaches to modeling ASA, we have lacked a comprehensive account of ASA phenomena (Szabó et al., 2016; cf. similar concerns in vision, Froyen et al., 2015). One longstanding challenge is the difficulty of integrating structured models with raw audio signals as input – most existing models of ASA are restricted to symbolic rather than acoustic input (Barniv & Nelken, 2015; Larigaldie et al., 2021; Mill et al., 2013). Because of this, it is unclear whether putative principles for auditory perceptual organization which apply to synthetic sounds could also extend to explaining ASA in everyday sounds (Deike et al., 2014). Recent deep neural networks have had success in a variety of naturalistic source separation tasks from raw audio, but have not been intended to model human hearing, and so have not yet been systematically compared to human perception.

In this paper we revisit the idea of perception as causal inference (Adelson & Pentland, 1996; Kersten & Schrater, 2002), leveraging technical developments of the last few years to render Bayesian inference in world models newly approachable. To build a rich, structured model of auditory scenes, we use a generalization of graphical models called probabilistic programs (Ghahramani, 2015). To implement search, we take an “analysis-by-synthesis” approach, assessing bottom-up proposals about potential causes via top-down “synthesis” in the generative model (Barker et al., 2005; Yuille & Kersten, 2006). This strategy combines the benefits of fast pattern recognition and Bayesian inference. Finally, to make analysis-by-synthesis tractable, we use new engineering tools. In particular, we use deep learning to make bottom-up proposals (“amortized inference”, Stuhlmüller et al., 2013) and assess them top-down by taking advantage using stochastic variational inference to approximate the posterior, leveraging a differentiable generative model (Kingma & Welling, 2014; Kucukelbir et al., 2017).

Unlike previous symbolic models, our model can be evaluated on any sound signal, enabling it to be tested on both classic illusions and natural sounds. But in contrast to the pure task-optimized deep neural networks that we examine here as baselines, our model outputs probabilities over meaningful, hierarchical scene descriptions (including sources, their properties, and the sound events they produce) based on its generative constraints. We show that our model accounts for a large variety of classic ASA illusions and infers perceptually valid sources from naturalistic audio. In contrast, other contemporary deep learning systems did not account for human perception. Our results demonstrate a tractable analysis-by-synthesis framework for perceptual organization that bridges traditional psychophysics and natural signals.

2.2 Results

2.2.1 Overview

To investigate the extent to which auditory scene analysis can be explained by a generative model of sound, we considered how to simply describe and render a variety of sounds. The resulting generative model provides a hypothesis for the perceptual constraints underlying auditory scene analysis. The model is accompanied by an analysis-by-synthesis inference procedure that we built to search the model space, enabling us to evaluate the model on audio signals. To test how well the generative model aligns with human perceptual constraints, we assessed the model on a wide range of sounds with diverse perceptual organizations. First, we tested the model on a set of classic auditory scene analysis illusions. These illusions are widely thought to elucidate principles of human hearing and thus provide a clear starting set of phenomena that any model should account for. To establish baseline performance on our illusion set using alternative scene analysis strategies, we compared how the illusions were perceived by our model and by an assortment of contemporary deep neural network source separation systems. We also tested the role and importance of different aspects of the model by assessing the effect of selective model “lesions”.

Finally, we tested the model on a set of everyday sounds.

2.2.2 Generative model

Any theory of perception as causal inference depends on assumptions about how causes generate data. Generative models give these assumptions explicit meaning by providing a direct specification of perception’s “world model”, including: 1) the structure of possible latent causes with associated prior probabilities, and 2) a description of how these causes generate data, determining the likelihood of the observed data under each possible cause. Together, these components provide probabilistic constraints on perceptual organization that determine the Bayesian posterior distribution $P(\text{cause}|\text{data})$.

Our proposed model is inspired by observations of generative principles in everyday sounds, balanced by simplicity to enable tractability in inference. Because everyday auditory scenes can include variable numbers and kinds of sources, we defined the model as a probabilistic program to allow for this variable structure. The program consists of two components, which work together to fully specify the generative constraints:

1. A sampling procedure which generates a hierarchical symbolic description of a scene, S , in terms of sources and the events they emit. The sampling procedure defines the prior distribution $P(S)$ over this space of possible causes of sound, providing constraints on what sources and events are probable *a priori*.
2. A renderer which uses this symbolic scene representation S to generate an audio signal. The rendered sound can, in conjunction with a noise model, be used to evaluate the model likelihood $p(X|S)$ for any observed sound X . The likelihood assesses how likely a scene description is to generate a particular sound.

Given an observed sound waveform X , the sampling procedure and renderer induce a posterior distribution over auditory scenes $p(S|X)$ by Bayes’ rule. The most probable scene descriptions for an observed sound can then in principle be found via

inference (searching through scene descriptions to find a description with high posterior probability). We begin by qualitatively describing the generative constraints that structure the scene description, then further formalize the probabilistic model, and then describe the inference procedure.

Scene description, S

Figure 2-2A shows examples of recorded everyday sounds. We took inspiration from such sounds and how they are produced in order to develop a flexible and widely applicable symbolic description of sound sources, while also keeping its structure minimal so as to facilitate inference. The resulting model is considerably more expressive than previous generative models for auditory scene analysis, and can generate simple approximations of many everyday sounds. The following three observations informed the model’s construction.

First, a substantial variety of everyday sounds can be described as coming from three broad classes (Figure 2-2A): noise, harmonic, and whistle sounds. Noise-like sounds are commonly produced by turbulence in air or fluids, or when large numbers of sounds superimpose to form textures (McDermott & Simoncelli, 2011; Misra et al., 2009); brief impact sounds are also often well described as short snippets of noise. Periodic sources produce harmonic sounds; these range from spoken vowels (Stevens, 2000) to creaking doors (Thoret et al., 2013). Whistles are commonly created by air flow and resonance (Henrywood & Agarwal, 2013) and are produced in a variety of ways by animals across taxa (Beckers et al., 2003; Riede et al., 2017; Wilczynski et al., 1984). Although some complex sound sources can involve more than one of these sound types, whether sequentially (e.g. speech) or simultaneously (e.g., blowing a clarinet creates a noisy breath sound and a harmonic tonal sound, Serra and Smith, 1990), many natural sounds are dominated by a single sound type.

Second, many sounds can be described as being produced by multiple discrete, dynamic “events”, corresponding to when a source supplies energy to produce sound. For example, the squealing sound produced by rubbing glass starts and stops over time, corresponding to relatively discrete time intervals when force is being applied

to the glass surface. Within those intervals, the sound continuously changes in fundamental frequency. The resulting events may be temporally extended over seconds (e.g., a kettle whistling) or more transient (e.g., the chirps of a bird), and can change in their properties from event to event (e.g., a dog panting in and out).

Third, the events emitted by a source may reflect regularities in a variety of attributes, including event timing, fundamental frequency, amplitude, and spectral shape. For example, some sources tend to produce many regular events in quick succession (e.g., rubbing glass), while others produce long events that continuously vary (e.g., coins shaking). And while the chirps of a bird may vary smoothly in fundamental frequency, a bee buzzing varies its frequency rapidly and erratically in time. Both stay in a relatively narrow range of frequency space.

These three observations are a starting point for a generative model of auditory scenes. In our model, a scene description consists of sources which each emit a sequence of discrete, non-overlapping events. A scene can contain any number of sources, which each can emit any number of events (Figure 2-2B, “sample n_s ”). A source generates one distinct type of sound: noise, harmonic, or whistle (Figure 2-2B, “sample type”). Sources of the same type can differ from each other in the attributes noted above, for example, in their spectral shape, in how loud they tend to be, or in how often they tend to emit events (characterized by the setting of the knobs in Figure 2-2B). These tendencies cause regularities across events produced by the same source (Figure 2-2C, “sample events”). Single events are dynamic in time, e.g. changing in frequency or amplitude (Figure 2-2C, events panel). Given a scene description S , the renderer generates the sound emitted by each source (rendered sounds in Figure 2-2C) and sums them to produce the scene sound (Figure 2-2D).

The sound types in our model are parametrized as “excitation-filter combinations” commonly proposed for natural sound generation (McDermott et al., 2012; Taylor & Reby, 2010; Van Den Doel et al., 2001) (Figure 2-2C, excitation-filter split in the events panel). Sound is produced by an “excitation” that supplies sound energy and a “filter” that determines the spectral shape; the excitation comes in one of three varieties determined by the sound type. Sounds produced by very different physical

mechanisms (e.g. the beating wings of a bee versus the friction of rubbing glass) may still be best described by the same excitation type, and therefore can have the same parametrization in the model. A noise source will produce aperiodic excitation with a pink ($1/f$) spectrum and an amplitude that varies over time. A harmonic source produces periodic excitation: a sum of harmonically related sinusoids with a time-varying fundamental frequency and amplitude. A whistle sound similarly produces a periodic excitation, but only generates the first harmonic. In whistle sounds, the filter can be considered fixed, and so we omit it for simplicity.

A source becomes “active” for a bounded time interval to emit an event, and is silent while it “rests” in between events (Figure 2-2C, active/rest in the events panel). An event has an onset when the source’s excitation begins (discretely), a finite active duration during which the excitation can vary, and an offset when the excitation ends. Thus, while active, events are additionally defined by their filter and time-varying excitation (Figure 2-2C, events panel). Source-level variables determine regularities in the timing, excitation, and filter of the events emitted by a source. The timing of events depends on the specific setting of the temporal source-level variables (represented by the pink knobs in Figure 2-2B and C); similarly, the trajectory of the excitation and the shape of the spectrum depend on the corresponding source-level variables (blue knobs in Figure 2-2B and C).

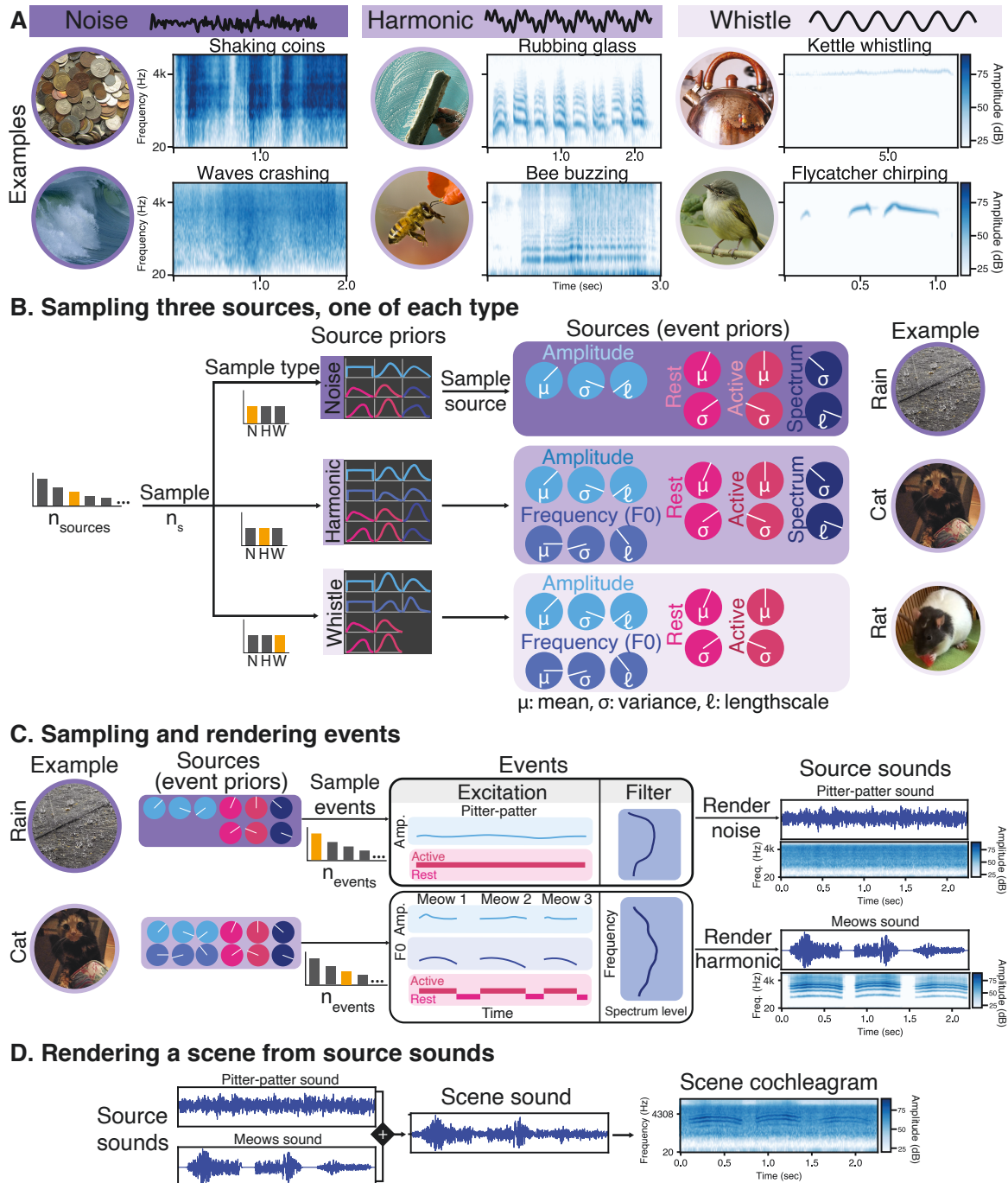


Figure 2-2: Components of the generative model, illustrated with everyday recorded sound examples.

A) The three basic sound types with natural sound examples. The examples demonstrate a variety of amplitude and frequency modulation, spectral shapes, and temporal patterns. B) A scene consists of any number of sources. Each source belongs to one sound type, which defines how the source is parametrized and rendered, and determines which hyperpriors the source is sampled from. These source variables define distributions over events. C) A source can emit any number of events. Events consist of active and rest intervals, amplitude, and depending on the sound type, possibly a spectrum and/or fundamental frequency. These event attributes are sampled from a source, which consists of a set of parametrized distributions. These events are rendered into sound waveforms by multiplying an excitation and filter (cochleagrams are shown for visualization). D) The sounds generated by independent sources sum together to create a scene sound, which is transformed into a cochleagram for model inference. This scene cochleagram is the basis for the likelihood function. Sound examples at <https://mcdermottlab.mit.edu/mcusi/thesis/>.

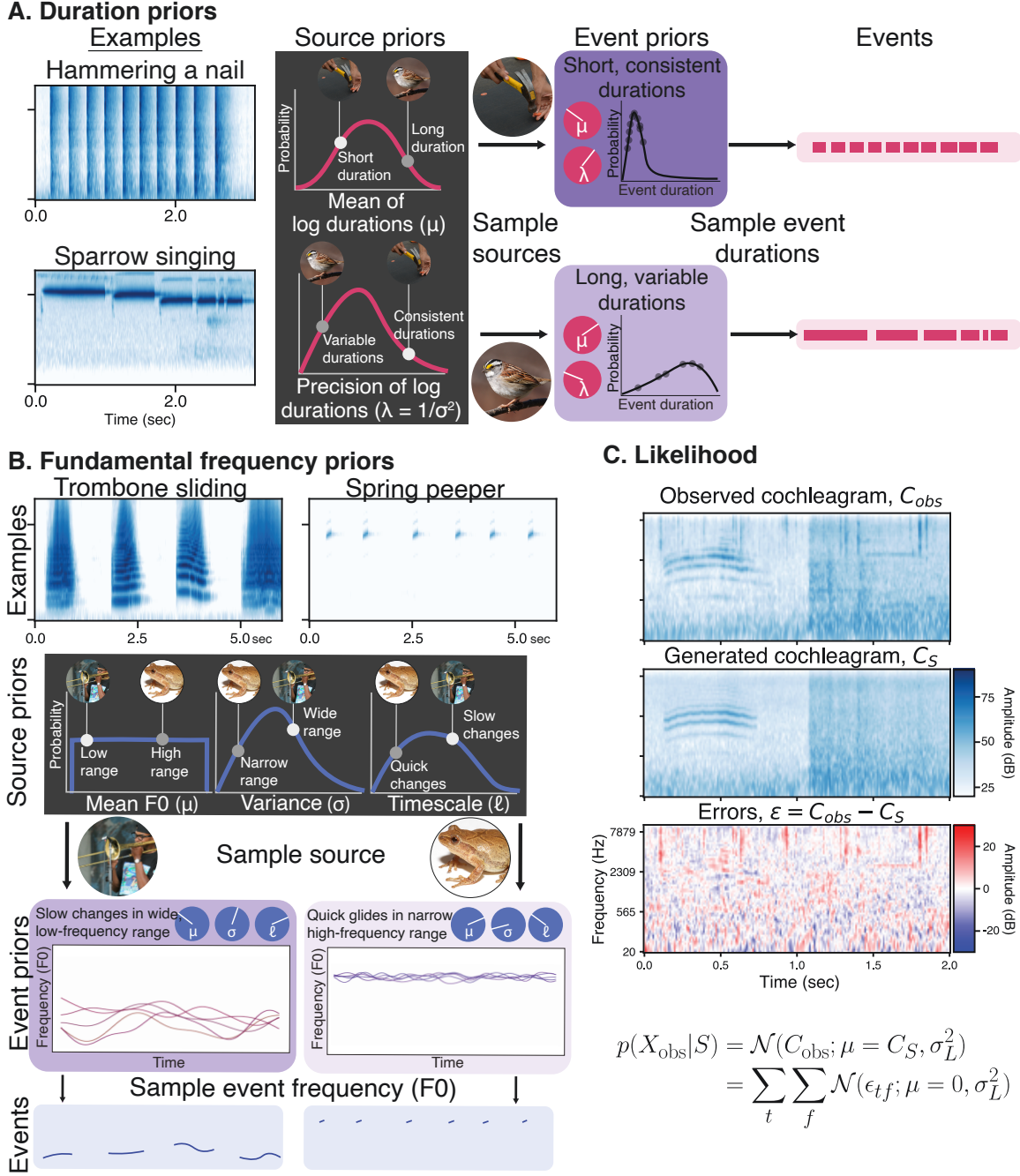


Figure 2-3: Prior distribution $P(S)$, illustrated with recorded natural sound examples.

A) Hierarchical priors on event duration (also used for the “rests” between events). The two recorded sounds have very different duration regularities: hammering nails results in short, consistent impacts, while the song of a white-throated sparrow comprises longer tones of variable duration. To capture such regularities, we use a hierarchical model. Within a single source, durations are sampled from a log normal distribution to ensure non-negativity. The mean and variance of this event-level prior are sampled from a normal-gamma conjugate prior. Different samples from the normal-gamma distribution capture different source regularities: the hammer impacts reflect a low mean and high precision, while the sparrow song reflects a high mean and low precision. The parameters of the source-level distributions are fit to recorded sounds. B) Hierarchical priors on fundamental frequency, an example of the Gaussian process prior used for excitation and filter variables. The mean and kernel parameters are source variables that are sampled from hyperpriors, allowing different sources to tend toward different characteristic excitation trajectories and filter shapes. C) To calculate the likelihood, the observed and generated cochleagrams are compared under a Gaussian noise model. Note the errors in the high-frequency clicks. Sound examples at <https://mcdermottlab.mit.edu/mcusi/thesis/>.

Prior, $P(S)$

The qualitative properties of the generative model described in the preceding section are made precise using a prior probability distribution over scene description, $P(S)$. To reflect the structure of the scene description (in which different sources emit events), $P(S)$ is hierarchical. There are prior distributions over sources for each sound type, and each source is itself defined by a parametrized prior distribution over events. Sources are assumed to be independent, whereas events emitted by a single source are sampled sequentially so that each event depends on the events preceding it.

This hierarchy is able to account for a variety of sound regularities, as illustrated in Figures 2-3A and 2-3B. Figure 2-3A illustrates how the model can express sources with different temporal regularities. The short, consistent impacts of hammering nails is represented as a small value of μ (average duration) paired with a large value of λ (precision). By contrast, the song of a white-throated sparrow comprises mostly longer notes of more variable duration (large μ and small λ). Our model expresses this variety by drawing each source from a ‘source prior’ that specifies the model’s distribution over likely values of μ and λ . The durations of all events produced by the source then follow a distribution governed by these source variables (in particular, a log-normal distribution parametrized by μ and λ). A source with a large μ will tend to produce longer events, while a source with large λ will tend to produce events of variable duration. The generative process for the temporal “rest” variable has the same structure, with different values specifying the shape of the source priors. Figure 2-3B shows the analogous hierarchical generative process for the time-varying fundamental frequency of the source excitation. The structure of this generative process is the same for the time-varying amplitude and spectrum of the source filter, but with different values defining the priors. To determine the parameters of the priors, we fit each one to a dataset of sound textures (for noise) and speech, musical instruments, and birdsong (for periodic sounds), comprising approximately one minute of audio each.

As shown in Figure 2-3A and B, these distributions can express a diversity of regularities that we observe in everyday sounds. These choices (a normal-gamma prior over the temporal variables, and Gaussian processes over the multivariate spectrum and excitation trajectory; see Methods) also enable the use of differentiable sampling procedures and efficient inference platforms that have been developed for common distributions (Gardner et al., 2018) thereby facilitating inference.

The overall model is depicted graphically in Figures 2-2 and 2-3, and is described in detail in the Methods. Supplementary Figure 2-10 shows sounds sampled from $P(S)$.

Likelihood, $P(X|S)$.

Given S , the symbolic description of the auditory scene, the renderer generates the soundwave produced by each source. The sound waveforms corresponding to all sources are summed to produce the scene wave, X_S .

The likelihood tells us how likely it is that a particular scene will generate an observed sound. To compute the $P(X_{\text{obs}}|S)$, the scene waveform X_S rendered from S must be compared to the observation X_{obs} . To do so, the two waveforms are first converted to a cochleagram C , a time-frequency representation of sound that approximates the filtering properties of the human ear, and are then compared under a Gaussian noise model (Figure 2-3C). That is, the likelihood is the probability that the observed cochleagram C_{obs} is a noisy measurement of the rendered cochleagram C_S :

$$\begin{aligned} p(X_{\text{obs}}|S) &= \mathcal{N}(C_{\text{obs}}; \mu = C_S, \sigma_L^2) \\ &= \sum_t \sum_f \mathcal{N}(\epsilon_{tf}; \mu = 0, \sigma_L^2) \quad \text{where } \epsilon = C_{\text{obs}} - C_S \end{aligned}$$

Human discrimination is often constrained by cochlear filtering (Jayant et al., 1993), providing a more suitable representation for matching the observation than the sound waveform. In addition to approximating the input at the human ear, the cochleagram representation of sound is useful as it often varies smoothly with respect to the

continuous latent variables of our model, facilitating gradient-based inference. The likelihood weighting σ_L is a free parameter (fixed to be the same value across all inferences).

2.2.3 Inference

By Bayes’ rule, the prior $P(S)$ and likelihood $P(X|S)$ induce a posterior distribution $P(S|X)$ for an observed sound X , that is, $P(S|X) \propto P(S)P(X|S)$. The posterior describes which hypothesized scenes are more likely explanations of the observed data. From this viewpoint, the computational goal of perception is to uncover the most likely explanation (or set of explanations) of the data under its world model.

Finding the hypotheses that are most likely under the posterior distribution involves two main challenges. First, posterior inference requires solving a difficult search problem. The space of scene descriptions is vast because the generative model is expressive, so it is difficult to find good hypotheses to evaluate. Second, the posterior distribution is always multimodal. Even if we could find regions of scene space that correspond to good hypotheses, there would be multiple modes that would need to be compared since some are bound to only be local optima. To successfully compare the modes, we must evaluate how much posterior probability mass corresponds to each mode. However, because hypotheses are high-dimensional and contain complex dependencies between the various random variables, it is difficult to estimate the probability landscape surrounding a mode closely enough in order to ensure that all of the probability mass covered by a mode will be accounted for.

How can a perceptual system solve these challenges? Intuitively, the observed sound contains clues about likely components of a good hypothesis, pointing us to which parts of the scene space are relevant to explore. For instance, if we recognized that part of a sound was aperiodic, we could guess that the scene description should include a noisy event at that time. However, this bottom-up, local pattern recognition may be unreliable with respect to the global scene interpretation. We might need to consider the global context, for instance, to recognize that a harmonic sound was also present in that interval, but masked. These considerations motivate the analysis-by-

synthesis approach. Analysis-by-synthesis incorporates the strengths of bottom-up pattern recognition to rapidly propose local variables, while leveraging the robustness of Bayesian inference over the full scene.

Specifically, we use amortized inference in neural networks (Dasgupta et al., 2018; Stuhlmüller et al., 2013) to propose local event variables (Step 1) which can be combined into global scene hypotheses (Step 2). These steps comprise the bottom-up aspect of analysis-by-synthesis. These scene hypotheses are then assessed top-down with stochastic variational inference in the generative model (Kingma & Welling, 2014; Kucukelbir et al., 2017). Hypotheses are built up sequentially in time, with the most promising intermediate hypotheses at each round expanded upon to explain successive intervals of sound (Step 4). The full **sequential inference** process is as follows (see examples in Supplementary Figure 2-11 and 2-12):

1. **Events proposal:** A neural network proposes candidate events, which initialize event variables in the hypotheses. This “amortized” inference allows us to benefit from existing state-of-the-art machine learning architectures (Wu et al., 2019) to find good hypotheses. The neural network is trained on data produced by the generative model.

Then, to build up the hypotheses sequentially in time beginning with the earliest candidate event, steps 2-4 are iterated until all candidate events have been assessed.

2. **Source construction:** Inspired by sequential Monte Carlo approaches (Nix & Hohmann, 2007) and similar to (Ellis, 1996; Mill et al., 2013), candidate events corresponding to the current timestep are combined into scene descriptions through simple update rules (add candidate event to existing source, create new source with candidate event, leave out candidate event). This results in a set of alternative hypotheses for the sound up to a particular moment in time. By building up hypotheses sequentially, we avoid searching a combinatorially large number of partitions of events into sources. Nevertheless, there may still be too many hypotheses to maintain computational tractability, so they are prioritized for hypothesis optimization with a set of heuristics (see Methods).

3. **Hypothesis optimization:** These hypotheses are refined using gradient-based optimization. Specifically, for each hypothesis, a guide distribution is optimized to best approximate a mode of the posterior distribution using variational inference. Variational inference allows us to benefit from a fully differentiable generative model, jointly optimizing all of the continuous latent variables in a hypothesis in order to fit each mode as closely as possible.
4. **Scene selection:** The posterior odds, approximated with importance sampling based on the optimized guide distributions, are used to compare the alternative hypotheses. A set of hypotheses with the highest posterior probability are selected for the next round of source construction.

This procedure results in a set of complete scene descriptions which best explain the full observed sound. The number of sources and events, the type of each source, and the events emitted by each source are all automatically inferred. Sequential inference is therefore general-purpose and can be applied to any audio, including both classic synthetic illusions and everyday sounds.

2.2.4 Model results on classic ASA phenomena

The model was first evaluated on a wide range of classic auditory scene analysis phenomena, intended as a representative sample of well-established findings in this domain (Table 2.1). We simulated the experiment associated with each phenomenon in the model. Illusions often simply involve a subjective judgment of perceptual organization; in such cases, we compared the commonly reported percept with the highest probability scene hypotheses under the model (found through sequential inference). For other phenomena, we simulated published psychophysical experiments for comparison with human judgments (using enumerative inference). The experiments in question often queried participants about a limited set of specific scene descriptions (e.g., in a two-alternative forced choice paradigm). When such experimental constraints were present, we replaced bottom-up search (steps 1 and 2) by enumeration

Table 2.1: Classic auditory scene analysis experiments and illusions

	Citation	Result	Experiment
Masking and perceptual filling-in	Auditory induction Warren et al., 1972	Perceived temporal continuity of sounds through masker depends on masker level	Masking and continuity thresholds
	Homophonic continuity Bregman and Ahad, 1996	Abrupt but not gradual amplitude changes cause segregation	Subjective report
	Spectral completion McDermott and Oxenham, 2008	Analogous perceptual completion in frequency domain	Subjective matching of spectra
	Co-modulation masking release, Hall et al., 1984	Detection threshold of tones decreases as bandwidth of co-modulated noise increases	Tone detection (2AFC)
Simultaneous grouping	Mistuned harmonic Moore et al., 1986	Mistuned harmonic component segregates from complex tone	Subjective report (1AFC)
	Frequency modulation McAdams, 1984	Frequency-modulated harmonics segregate	Subjective report
	Asynchronous onsets Darwin and Sutherland, 1984	Harmonic component with asynchronous onset of offset segregates	Vowel categorization
	Cancelled harmonics Hartmann and Goupell, 2006	Amplitude gated harmonic component segregates from complex tone	Pitch matching
Sequential grouping	Frequency proximity Tougas and Bregman, 1985	Crossing tone sequence segregates into non-overlapping bouncing streams	Subjective report
	Bistability Van Noorden, 1975	Increasing frequency difference and decreasing interstimulus interval increases segregation, with a region of bistability	Subjective report
	Cumulative repetition Bregman, 1978a	Segregation increases with repetition	Subjective report
	Effects of context Bregman, 1978b	Context can promote segregation or grouping	Subjective report
	Bregman and Rudnick, 1975	of two tones	

over the experimentally-defined hypotheses. Those hypotheses were refined with variational inference and compared (as in steps 3 and 4) to yield an experimental response. Sound examples can be found at <https://mcdermottlab.mit.edu/mcusi/thesis/>.

Masking and “filling-in”

When sources produce sounds that overlap in time and frequency, sufficiently intense sounds can obscure the presence of less intense sounds. That is, if a less intense sound is added to a sufficiently intense sound, the less intense sound will not be heard; this everyday occurrence is termed “masking” (Warren et al., 1972). In such cases, the

addition of the less intense sound does not alter the peripheral auditory representation to a detectable extent. However, the perceptual interpretation can be modulated by context. For instance, a noise flanked by tones could equally well consist of two short tones adjacent to the noise, or a single longer tone overlapping the noise, that happens to be masked when the noise is present. Warren et al. (1972) showed that listeners hear this latter interpretation as long as the noise is intense enough to have masked the tone were it to continue through the noise. Listeners thus perceptually “fill-in” a continuous, quieter tone behind the noise: a stable percept, despite physical interruption by the masker. We asked whether the model also infers events that are not explicit in the sound when evidence is consistent with masking.

Auditory induction

In Warren et al. (1972), tones of various frequencies were either *embedded* in noise (to measure masking) or *alternated* with noise (to measure continuity). Listeners’ masking threshold corresponded to the level at which the embedded tones became audible, and their continuity threshold corresponded to the level at which the alternated tones sounded continuous (Figure 2-4A, schematic). There was a close correspondence between these thresholds across tones of different frequencies (Figure 2-4A, top plot). The noise was missing energy in a “notch” around 1000 Hz in order to test whether overlap in the frequency domain was necessary for listeners to perceive continuity, which was confirmed by the drop in masking and continuity thresholds for the 1000 Hz tones. This effect is replicated in our model (Figure 2-4A, bottom plot).

Homophonic continuity

The model was also tested on a variant of the continuity illusion involving amplitude-modulated noise (Bregman and Ahad, 1996, Figure 2-4B). When an initially soft noise undergoes a sudden rise in intensity, listeners perceive the initial source as continuing unchanged behind a distinct, louder noise burst (Warren et al., 1972). In contrast, if the amplitude modulation occurs gradually and reaches the same peak, listeners instead hear a single source changing in intensity. In accordance with human listeners,

sequential inference in the model discovered two sources as the model’s preferred explanation when the noise amplitude changes abruptly (over 1 ms), and one source when the amplitude changes gradually (over 250 ms).

Spectral completion

Analogous phenomena occur over the frequency spectrum, dubbed “spectral completion”. In McDermott and Oxenham (2008), listeners heard a long masker noise, which overlapped with a brief target noise halfway through its duration (Figure 2-4C, schematic in top left). The spectrum of the target was ambiguous because the middle band of its spectrum could be masked. Listeners were asked to adjust the level of the middle band of a comparison noise (Figure 2-4C, schematic in bottom left), until the comparison perceptually matched the target. Listeners chose the level of the middle band to be well above its audibility threshold, suggesting that they perceptually filled in the masked portion of the target. Several variations on this basic stimulus configuration revealed how this perceptual “filling-in” was affected by context and masker levels (Figure 2-4C, top plots). The pattern of judgments across these stimuli was replicated in our model (Figure 2-4C, bottom plots).

Co-modulation masking release

Another masking-related grouping phenomenon occurs when masking noise is co-modulated (Hall et al., 1984). Coherently modulated noise bands produce lower tone detection thresholds than unmodulated noise, “releasing” the tone from masking, hence the term “co-modulation masking release” (CMR). One experimental approach for studying CMR compares detection thresholds for co-modulated relative to unmodulated noise maskers centered on a target tone as the masker bandwidth widens. In contrast to unmodulated noise, for which thresholds grow and the level off as bandwidth increases, co-modulated noise produces thresholds that decrease for sufficiently wide bandwidths – the far outskirts of the noise evidently group with the noise around the tone, helping listeners perceptually separate the tone and the noise. The model shows this same effect (Figure 2-4D), although the modulation frequency

at which it is most evident is lower than that for human listeners (Schooneveldt & Moore, 1989).

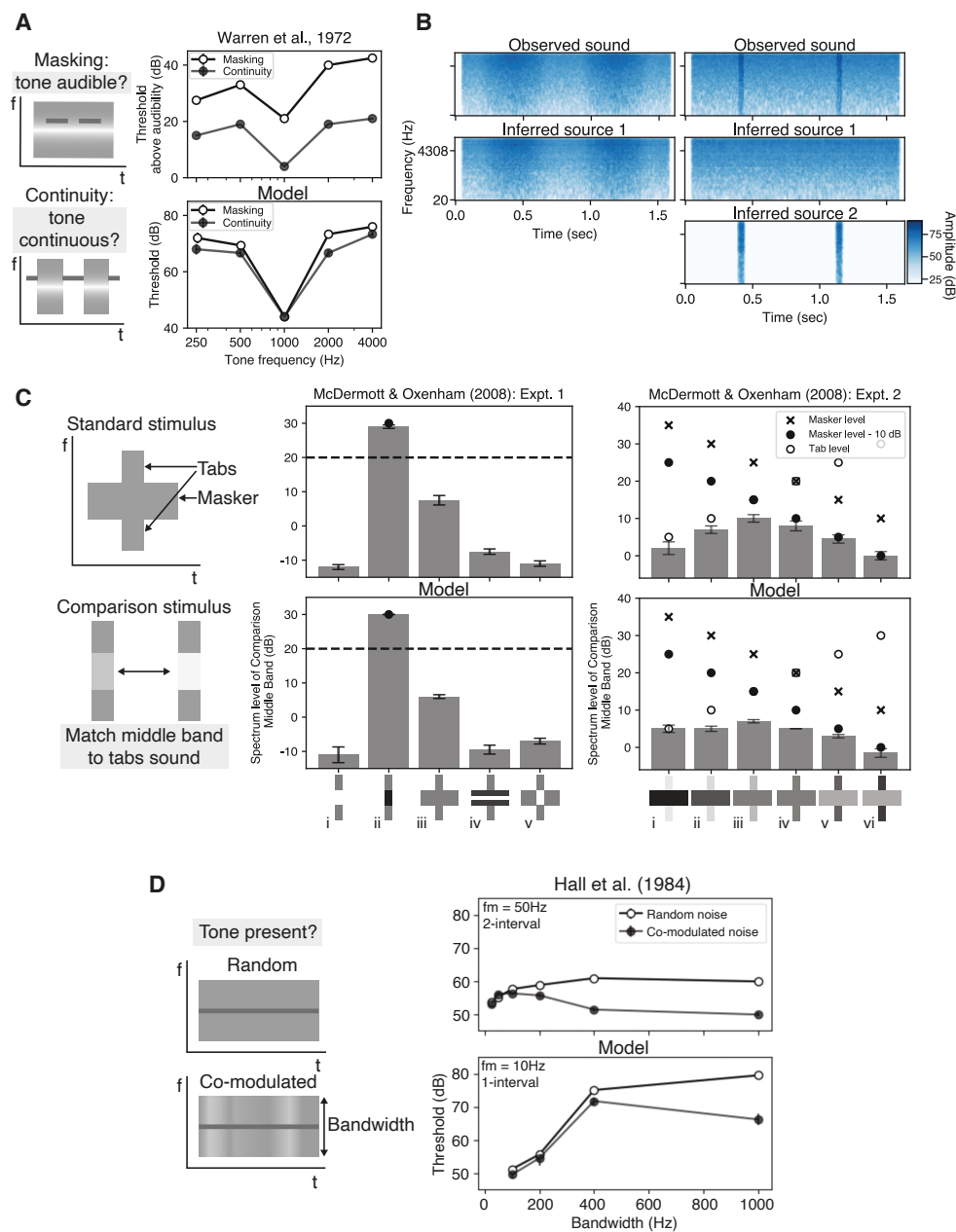


Figure 2-4: Model results on masking and filling-in phenomena.

A) Auditory induction. Schematic: tasks for the masking and continuity condition. Thin lines depict tones, grey rectangle with a gradient depicts notched noise (with notch at 1000 Hz). Top plot: average thresholds from 15 human listeners from Warren et al. (1972). Bottom plot: model results thresholds over 3 inference runs. Like human listeners, the model's thresholds drops at the notch.

B) Homophonic continuity (Bregman & Ahad, 1996). When the amplitude envelope changes abruptly, the model infers a steady source continuing behind a second source. Top plots: human results from McDermott and Oxenham (2008), averaged over 8 participants. Bottom plots: model results, averaged over 5 inference runs. The model shows a similar pattern of results to human listeners.

C) Spectral completion. Schematic shows the experimental paradigm, in which a comparison stimulus is adjusted until it sounds like the target 'tabs' source in the standard stimulus. Shade of grey indicates amplitude. Top plots: human results from McDermott and Oxenham (2008), averaged over 8 participants. Bottom plots: model results, averaged over 5 inference runs. The model shows a similar pattern of results to human listeners.

D) Co-modulation masking release. Schematic of the two noise type conditions, and the model task. Presented with tone superimposed on a noise of varying bandwidth, the model judged whether a tone was present. Top plot: average thresholds from 5 participants (Hall et al., 1984). Bottom plot: model thresholds averaged over 9 stimuli for each (bandwidth, noise type) pair. Like human listeners, the difference between the random and co-modulated threshold increases with bandwidth for the model. All error bars show ± 1 standard error across listeners for the human results and across inference runs for the model results.

Simultaneous grouping

Another set of classic perceptual organization problems pertain to whether simultaneous sound components are perceived as integrated or segregated. We tested whether the model could account for simultaneous grouping phenomena across a variety of harmonic stimuli.

Mistuned harmonic

Listeners tend to hear tonal components with harmonically related frequencies as a single perceptual entity. When the frequency of one tonal component deviates from the harmonic series, the component is heard to stand out as a separate tone. Moore et al. (1986)’s experiment measured the threshold at which this separation occurred, expressed as a percent of the harmonic frequency (Figure 2-5A). Across multiple fundamental frequencies and harmonic numbers, the model also showed a measurable mistuned harmonic effect, albeit with higher thresholds than expert listeners in the experiment (plausibly due to the limited frequency information in the cochleagram representation used to compute the likelihood).

Frequency modulation

Modulating the frequency of a subset of tonal components in parallel also causes them to perceptually separate from an otherwise harmonic tone (McAdams, 1984). To test whether the model showed a similar effect, we created a sound in which the odd-numbered tonal components had a constant fundamental frequency and the even-numbered components were frequency-modulated in parallel. Sequential inference discovers two harmonic sources as the model’s preferred explanation for these stimuli (Figure 2-5B).

Asynchronous onsets

What is perceived when components are harmonically related but start and end at different times? This situation can be considered a “cue conflict”, with harmonicity

favouring grouping and asynchrony favouring segregation. By what logic are these different types of information integrated to produce a final percept? Verbal descriptions of grouping cues cannot answer this question. We propose that source models provide the basis for such integration, and asked whether the model integrates information in the same way as listeners.

Darwin and Sutherland (1984) examined the case where the components are harmonic but do not share a common onset. To determine whether a tonal component of a harmonic sound was perceptually grouped with the others, they used judgments of vowel quality (in particular, whether a sound was perceived as /I/ or /e/). Vowels differ in the frequency of their “formants”, which are peaks in the spectral envelope: the harmonic tone perceived as the vowel /I/ has a lower frequency first formant than /e/. An added tonal component would shift the vowel’s formant if it were grouped with the rest of the vowel, thereby changing its category.

A *basic* continuum of seven stimuli was constructed by shifting the first formant of a 125 Hz harmonic tone from 375 Hz to 500 Hz. This continuum was perceived as shifting from /I/ to /e/. A new *shifted* continuum was then constructed by adding a 500 Hz pure tone overlapping with each harmonic tone of the basic continuum. The phoneme boundary occurred earlier in the shifted continuum (lower nominal first-formant value) compared to the basic continuum. Last, two new *early-onset* continua were constructed. These new continua were constructed by adding a 500 Hz pure tone to each stimulus of the basic continuum, but now the pure tone onset was 32 ms or 240 ms before the harmonic tone. Even though the frequencies and amplitudes were physically identical to those of the shifted continuum, listeners perceived the early-onset continua to have a similar phoneme boundary as the basic continuum. These results indicated that the pure tone was not integrated with the harmonic tone when their onsets differed. To test this effect in our model, we added a final vowel categorization step after inference, based on empirical vowel formant distributions (Hillenbrand et al., 1995). Our model replicated the similarity of the phoneme boundaries for the *basic* continuum and the *early-onset* continua (Figure 2-5C).

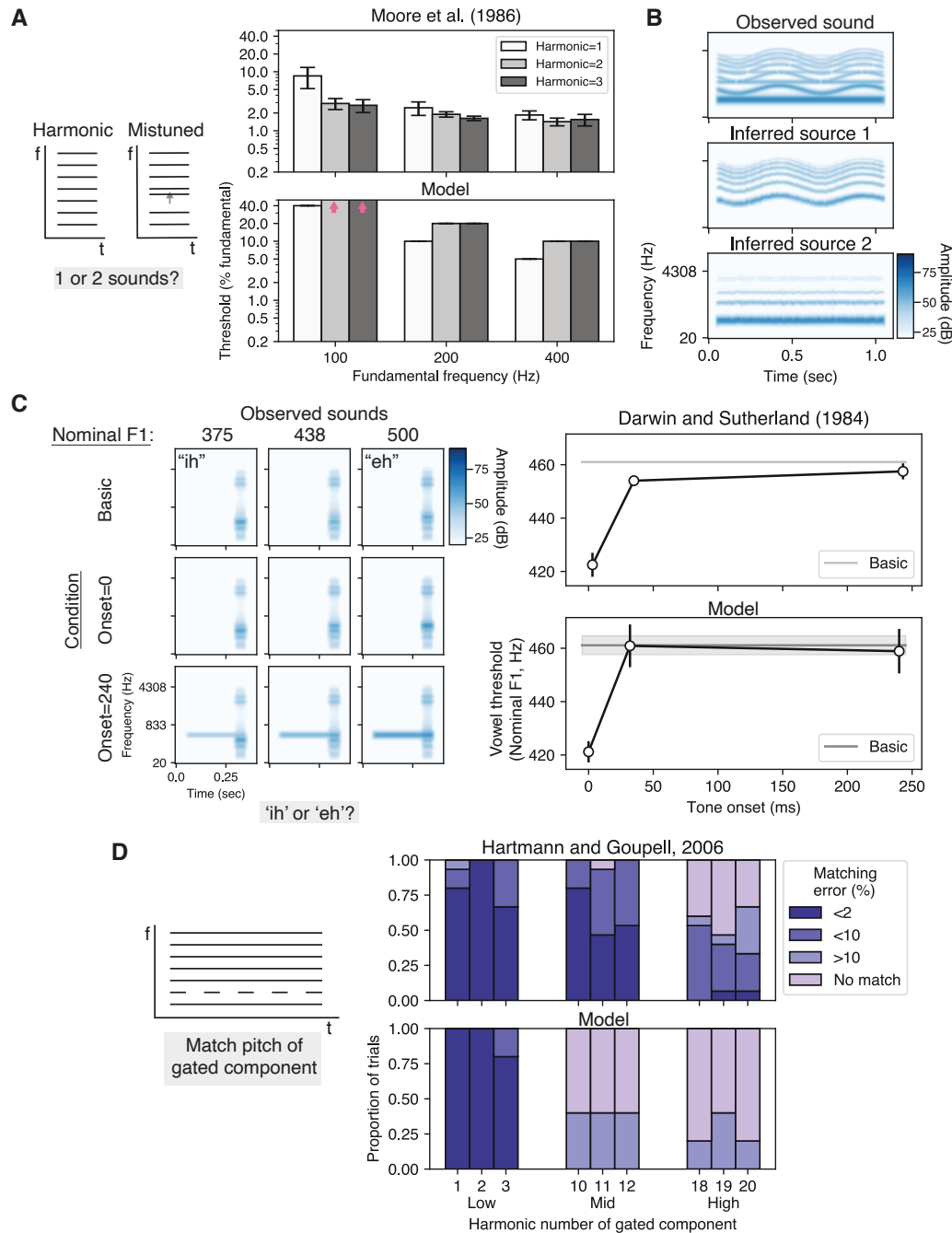


Figure 2-5: Model results on simultaneous grouping phenomena.

A) Harmonic mistuning. Schematic: one-interval task where listeners hear a complex tone that may be mistuned, and judge whether they heard one or two sounds. Top plot: average threshold as a percent of the fundamental frequency, averaged across four listeners (Moore et al., 1986). Bottom plot: average model threshold. The model shows a mistuned harmonic effect, albeit with higher thresholds than human listeners. For the upper harmonics of the 100 Hz tone, we could not find a threshold up to 50% (pink arrows). The higher thresholds plausibly reflect the limits of the frequency information available in the cochleagram used for inference. B) Frequency modulation. The model separates the modulated from the unmodulated components. C) Onset asynchrony. The left show example stimuli from the experiment. The basic row is created by shifting the first formant of the vowel from 375 to 500 Hz. The onset=0 and onset=240 stimuli are both created by adding a 500 Hz tone to the basic row, but with different onsets. Participants judge whether the vowel sounds like “ih” or “eh”. Top right: vowel boundary averaged over 6 human listeners from Darwin and Sutherland (1984). Bottom right: model vowel boundaries averaged over ten inference runs. The model replicates the effect of onset asynchrony seen for human listeners. D) Cancelled harmonics. Schematic shows the stimulus and task. Top plot: error distribution for matching task in human listeners (Hartmann & Goupell, 2006); bottom plot: model error distribution. The model replicates human performance for low frequency harmonics, and shows a similar trend towards more error at higher frequencies. All error bars indicate ± 1 standard error, over listeners for human results and over inference runs for the model.

Cancelled harmonics

The “cancelled harmonics” stimulus (Houtsma et al., 1988) (Figure 2-5D, left) provides another example of cue conflict: for the duration of a harmonic tone, one of its components is gated off and on over time. Although this gated component remains harmonically related to the rest of the tone at all times, it perceptually stands out as a separate component. Hartmann and Goupell (2006) asked listeners to match the frequency of gated tonal components that had different harmonic numbers. They reported that participants could closely match the frequency of the gated component up to the tenth harmonic. In the model, sequential inference discovered a whistle source with a well-matched fundamental frequency but only for low harmonic numbers (Figure 2-5D), qualitatively replicating this effect. The quantitative discrepancy again seems plausibly explained by limitations in the frequency information available in the cochleagram representation used for inference.

Sequential grouping

A final set of ASA results concerns how sound events are grouped or segregated into sequences over time. To assess whether the generative model could also account for sequential grouping, it was evaluated on a set of classic results in the perceptual organization of tone sequences.

Frequency proximity

To demonstrate the role of frequency proximity in sequential grouping, Tougas and Bregman (1985) interleaved rising and falling tone sequences, producing the “X” pattern apparent in Figure 2-6A. Listeners found it difficult to hear rising or falling trajectories in the mixture. Instead, listeners heard the higher frequency tones as segregated from the lower frequency tones, producing two sequences that “bounce” and return to their starting points. Listeners are able to hear the crossing explanation only when the pure tones in the rising trajectory are replaced by harmonic tones. The model replicates these findings: sequential inference discovers the bouncing organi-

zation as the model’s preferred explanation for the fully pure tone sequence, and the crossing organization in the latter sequence (Figure 2-6A).

Bistability

Van Noorden (1975) demonstrated bistability in auditory perceptual organization with the now-classic “ABA sequence”. This sequence comprises a repeating set of three tones, where the first and last have the same frequency. Depending on several stimulus parameters as well as the experiment instructions, participants typically hear one of two dominant potential organizations: 1) all the tones are grouped together to produce a galloping rhythm (“one-stream”), and 2) the A tones are grouped separately from the B tones, such that there are two sequences which each produce an isochronous rhythm (“two-streams”) (Figure 2-6B). van Noorden showed that increasing the frequency difference between A and B (Δf), and increasing the temporal rate of tones ($\Delta \tau$), both increase reports of the “two-streams” percept. In the low Δf region, the “one-stream” percept is inevitable, while for high Δf and low $\Delta \tau$ the “two-streams” percept is inevitable. For intermediate settings of these variables, the percept is bistable. The model accounts for these trends, as reflected in the log odds of the two explanations. Overlapping with Van Noorden’s intermediate region, the model log odds are close to zero, reflecting bistability in the model posterior (Figure 2-6C).

Cumulative repetition

Another factor affecting the preferred perceptual organization of the ABA sequence is the number of repetitions. When an ABA triplet is repeated multiple times over several seconds, listeners increasingly tend toward the “two-streams” percept (Bregman, 1978a). In Thompson et al. (2011), listeners listened to an ABA sequence and could indicate whether they heard one or two sources at any time during the sequence. Participant responses were averaged and binned to reveal an increase in two source responses as participants heard more repetitions of the triplet over time. This trend is reflected in the model, where the posterior odds increase over time to favour

“two-streams” over “one-stream” (Figure 2-6D).

Effects of context

Two complimentary experiments have shown that even when the relative timing and frequency between two tones are fixed, the surrounding context can alter whether the tones are grouped or not (Figure 2-6E and F) (Bregman, 1978b; Bregman & Rudnick, 1975).

In the first experiment, the frequency separation and timing between tones A and B remain constant, but they may be perceived in the same or different source depending on the frequencies of context tones X and Y (Bregman, 1978b) (Figure 2-6E, top). In this experiment, there were two kinds of tone sequences: ‘isolate’ sequences where X and Y were in a separate frequency range from A and B, and ‘absorb’ sequences where all tones were in the same frequency range. Listeners rated whether it was easy to hear tones A and B as a separate pair for four ‘isolate’ sequences and three ‘absorb’ sequences. The model log odds compare the hypothesis when A and B are in their own source versus when they are in separate sources. The pattern of log odds differentiates the two types of sequences, qualitatively replicating listeners’ ratings (Figure 2-6E, bottom).

In the second experiment, the frequency of the captor tones (“C”) affects whether listeners hear the distractor tones (“D”) group with the target tones (“T”) (Bregman & Rudnick, 1975) (Figure 2-6F, top). Listeners had to judge whether the two tones of the target move upward or downwards in pitch. They were more accurate when the frequency of the captor tones was close to the frequency of the distractor tones (at 1460 Hz), presumably because the distractor tones then group with the captor tones, making it easier to hear out the target tones. We evaluated the tendency of the model to group the target and distractor tones, by measuring the model log odds comparing the hypothesis when the target tones are in their own source versus when the distractor and target tones are grouped together. The log odds across the captor conditions reflect the increase in accuracy for human listeners (Figure 2-6F, bottom).

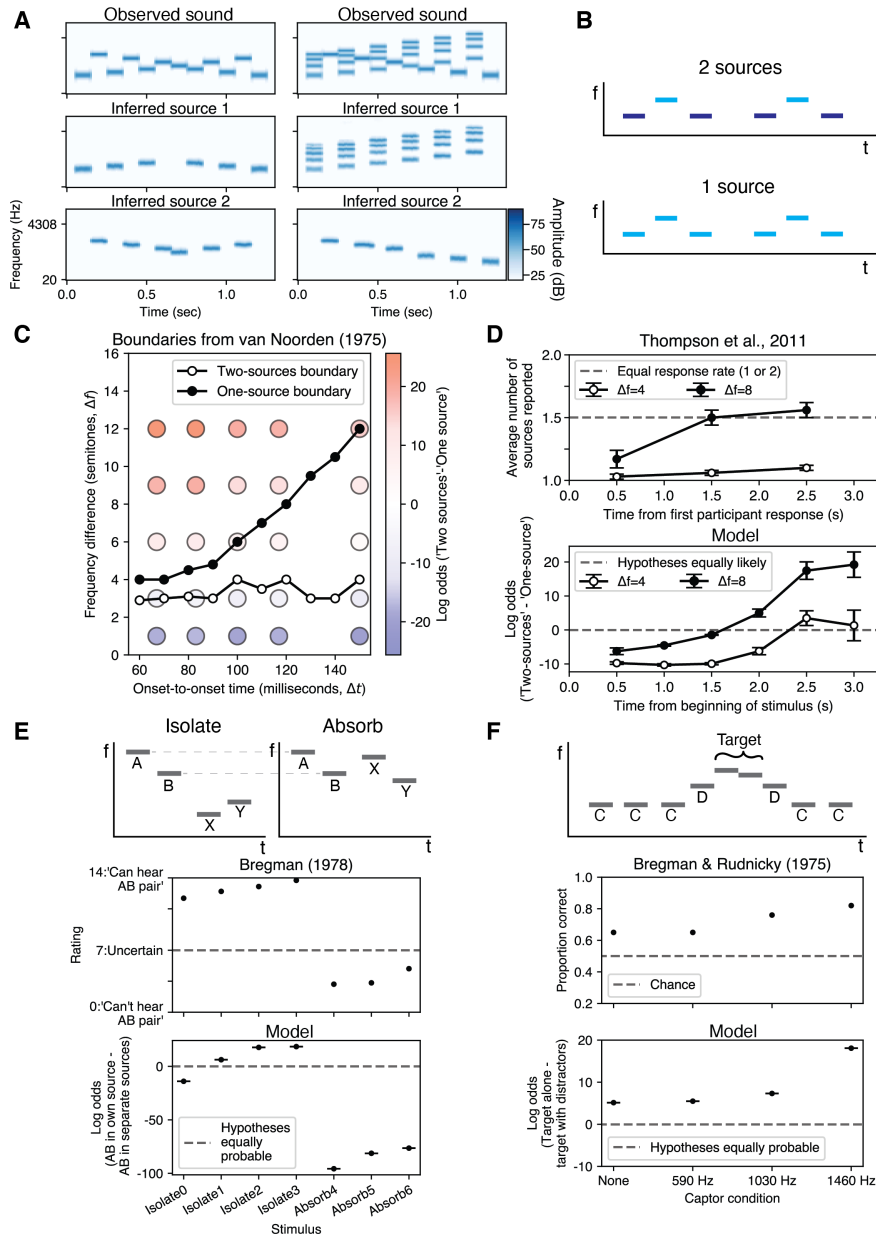


Figure 2-6: Model results on sequential phenomena.

A) Frequency proximity (Tougas & Bregman, 1985). Model prefers grouping tones with nearby frequencies (column 1), unless there is a spectral difference (column 2), matching human perception. B) Two perceptual organizations of the triplet sequence (Thompson et al., 2011; Van Noorden, 1975). ‘2-sources’: listeners hear low tones grouped together as separate from high tones. ‘1-source’: listeners hear all tones grouped. C) Bistability. Van Noorden (1975) measured listeners’ ‘two-sources’ boundary (threshold below which listeners do not hear two-sources) and one-source boundary (threshold above which participants do not hear one-source). Coloured dots: average model log odds at different stimulus settings (red=prefers 2-sources, blue=prefers 1-source). The region of bistability for humans overlaps with log odds close to zero, indicating bistability in the model posterior. D) Cumulative repetition. In Thompson et al. (2011), listeners indicated at any time during the sequence if they heard one or two sources. Top plot: average human responses, which increase in 2-source responses with more stimulus repetition. Bottom plot: model log odds for the 2- vs. 1-source explanation (positive: prefers 2-sources, negative: prefers 1-source), replicating the preference for two-sources with more repetitions. D) Effects of context (1). Top row: schematic tone sequences for each stimulus type. A and B tones are the same in both stimuli. In ‘isolate’ stimuli, X and Y occupy a different frequency range to A and B. In ‘absorb’ stimuli, they occupy an overlapping frequency range. Second row: average participant ratings for the 7 stimuli from Bregman (1978b) (0: confident that A and B could not be heard as a separate pair, 7: unconfident, 14: confident that A and B could be heard as a separate pair). Bottom row: model log odds (positive=prefers A and B in their own source, negative=prefers A and B in separate sources). Log odds are more positive for ‘isolate’ stimuli, just as they are rated more highly by human listeners. E) Effects of context (2). Top row: schematic tone sequence (C: captor tones, D: distractor tones at 1460 Hz). Second row: average participant accuracy on an interval judgment task (Bregman & Rudnick, 1975). Third row: model log odds (positive=prefers target in own source, negative=prefers target grouped with distractors). The log odds increases as the captors get closer in frequency to the distractors, similar to the pattern of human accuracy. All error bars indicate ± 1 standard error, over listeners for human results and over inference runs for the model.

2.2.5 Model comparisons

We have shown that our model comprehensively replicates a diversity of phenomena in auditory perceptual organization. However, it remains unclear what aspects of its structure are important for these results. We address this question through a series of model comparisons. First, we demonstrate the difficulty of matching human perception by evaluating an alternative model class that lacks highly structured constraints – deep neural networks trained on recorded audio. Second, we test alternative versions of our model in order to clarify their role in obtaining these results.

Neural networks

We evaluated an assortment of contemporary source-separation networks on the set of classic ASA phenomena in the previous section. Since previous work has demonstrated that deep neural networks reproduce key aspects of human behaviour and neural processing in tasks such as sound localization and word recognition (Franssen & McDermott, 2022; Kell et al., 2018; Saddler et al., 2021), this model class might reasonably serve as a baseline for matching human judgments. We used seven published models trained on natural sounds chosen to span a diversity of architectures, tasks, supervised and unsupervised training regimes, and natural sound datasets (Cosentino et al., 2020; Pariente et al., 2020; Stöter et al., 2019; Uhlich & Mitsufuji, 2020; Wisdom et al., 2021; Wisdom et al., 2020). Rather than inferring latent symbolic scene structure as in the generative model, these networks are designed to take in a mixture waveform and output premixture waveforms. For each network, we obtained premixture estimates for each experimental stimulus and illusion.

To run a network on a psychophysical task in which it must choose between alternative scene descriptions, we took its judgment to be the scene whose rendering produced the minimum mean squared error from the networks’ output, as measured on a cochleagram (see Methods). Finally, we quantified the overall dissimilarity between the network and human judgments averaged across experiments and illusions. We first compared the neural networks with the model results described in Section 2.2.4.

As shown in Figure 2-7A, no neural network could replicate the generative model’s match to human perception across the broad range of ASA phenomena tested here (neural networks: grey bars, model: blue bar). These results illustrate the difficulty of matching human perception on classic ASA phenomena – human-like results do not fall out of contemporary source separation systems trained on natural sounds.

To evaluate our model in the same way that we evaluated the source separation networks, we inferred sources for all the experimental stimuli (rather than comparing the experimenter-specified hypotheses, which cannot be evaluated using a conventional source separation model). For each experiment, we rendered source sounds from an inferred scene and analyzed them in the same way as for the neural networks. This inference alternative (Figure 2-7A, pink bar) is still more similar to human perception than any of the neural networks tested, showing that the generative model produces human-like results even without experimental constraints during inference.

We also assessed whether a source separation network would produce better results when trained on samples produced by the generative model (Figure 2-7A, gray bar without hashing). The resulting network also had a worse match to human perception than the generative model. This result demonstrates the difficulty of fully amortized inference, highlighting the utility of our analysis-by-synthesis approach. It also suggests that the dissimilarity of the previously published networks to human perception is not wholly explained by the data these networks were trained on.

Model alternatives

To assess the contribution of the structured constraints expressed by the generative model, we devised a set of alternative models that targeted different levels of its structure. The first two model alternatives addressed the hierarchical hyperpriors over sources.

MAP sources. To test whether the generative model’s hierarchical hyperpriors were necessary, we removed a set of the hyperprior distributions over sources. Instead of those distributions, each source had parameters equal to the distribution’s

mode. Specifically, we removed the temporal hyperpriors (Figure 2-3A) and the variance and lengthscale hyperpriors over f_0 , amplitude, and filter shape (Figure 2-3B). We found that restricting the variability in possible source parameters reduced the model’s similarity to human perception (Figure 2-7B), affecting several ASA phenomena including homophonic continuity, auditory induction, and onset asynchrony. This result suggests that an expressive distribution over sources is important to explain perceptual organization.

Uniform distributions over variance and lengthscale. The generative model’s variance and lengthscale hyperpriors were originally fit to a small set of everyday sounds. This means that sources with excitation trajectories and spectral shapes whose variances and lengthscales match those sounds are more probable. We changed these hyperprior distributions to be uniform, making a wider range of source parameters equally likely and therefore making a wider range of excitation trajectories and spectral shapes equally likely. We found that this alteration also decreased the model’s overall similarity to human perception, though substantially less so than the MAP sources alternative (Figure 2-7B). This result indicates that the model’s similarity with human perception is driven more by the model’s overall structure than by the fine-tuning of the hyperpriors to everyday sounds.

The next two model alternatives test the role of the event priors, specifically the Gaussian process priors over continuous event variables (fundamental frequency, amplitude, and spectrum). We altered the covariance kernels of these Gaussian processes, which encode tendencies in how functions change over time (frequency, amplitude) or frequency (spectrum).

Spectral swap. In the original model, we used different kernels to describe spectra of different sound types: a smooth mean-reverting kernel for harmonic spectra, and a non-smooth mean-reverting kernel for noise spectra (see Methods, Equation 2.23). We swapped these kernels to examine the impact of these assumptions. We found the spectral swap reduced the model’s match to human perception for two phenomena: spectral completion and onset asynchrony (Figure 2-7C). These phenomena both rely

on some amount of spectral “filling-in”: in both cases, one sound masks a part of another sound’s spectrum. However, spectral completion uses bandpass noise bursts and onset asynchrony uses speech-like harmonic tones. This result suggests that filling-in may depend on relatively specific assumptions about the spectra of different types of sounds.

Stationary covariance. The original model can express the case where a single source modifies its sound-generating process between events, even if these events occur in immediate succession (e.g., a dog panting in and out; Supplementary Figure 2-9). To implement this, the Gaussian process kernels for the excitation trajectories include a non-stationary term (see Methods, Equation 2.21). When removing this non-stationary term, we found a reduced match to human perception for sequential grouping phenomena: this alternative model showed the same trends as the original model but its quantitative match was worse, consistently favoring the “two-stream” percept even when human listeners do not (Figure 2-7D). This result highlights one possible function of the event level priors in explaining perceptual organization, namely, representing discontinuities in sound-generating processes.

In summary, these alternatives demonstrate the necessity of the different components of the model. Comprehensively explaining many ASA phenomena in a single model required a hierarchical, expressive model of sources and events. The model also provides a refined understanding of key assumptions about sound generation, as highlighted by the kernel alternatives.

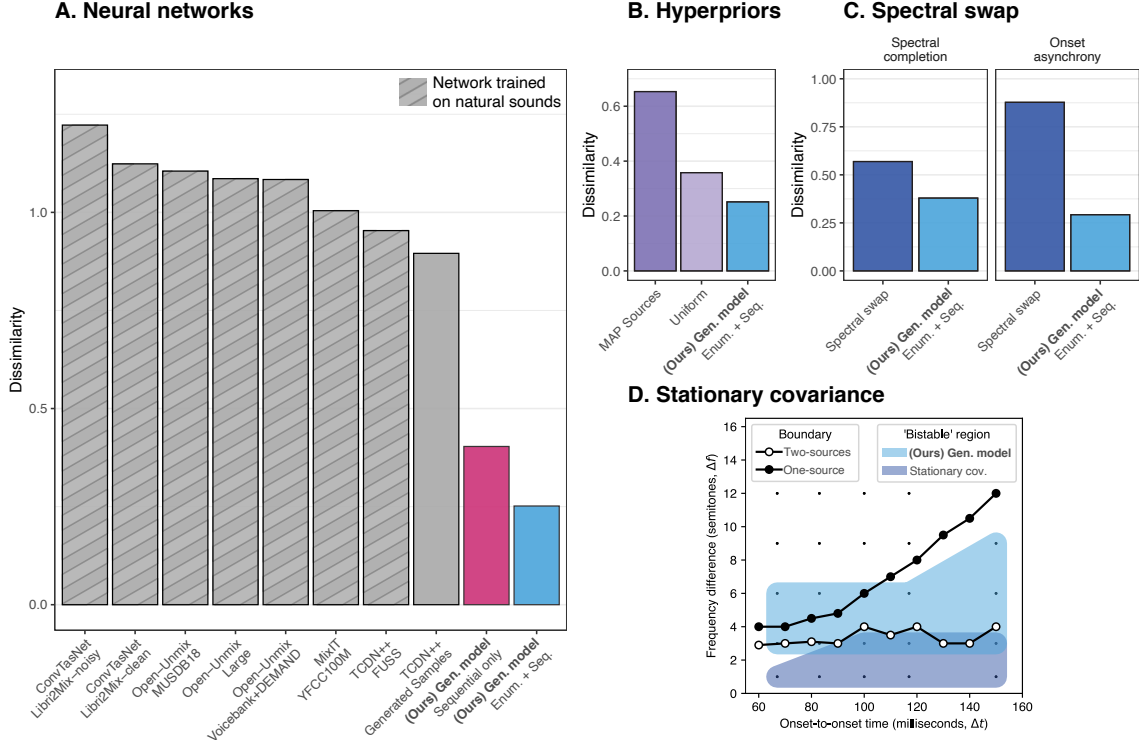


Figure 2-7: Comparisons with alternative models on classic ASA phenomena.

In all bar plots, a value closer to zero indicates the model makes judgments more similar to human perception. A) Human dissimilarity of source separation neural networks. The hatched grey bars indicate neural networks trained on source separation tasks using corpora of natural sounds (labelled as network/dataset). The blue bar indicates our model as assessed in section 2.2.4, with a mix of enumerative and sequential inference, depending on the experiment/illusion. We also assessed our model using purely sequential inference so as to obtain judgments in the same way as for the networks (pink solid bar). Our model is still more similar to human perception in this case. The solid gray bar, ‘TCDN++/Generated Samples’, shows the result for a source separation network trained on samples from our generative model. Its results remains much less similar to those of humans than the two cases of inference in the generative model. B) Models with alternative hyperpriors. Fixing all sources to the mode of the hyperpriors decreased dissimilarity more than changing the hyperpriors to uniform, which increases the variability in the prior over sources. C) Comparing the simulated results of van Noorden (1975) under the full model and a lesioned generative model with a stationary covariance kernel for frequency and amplitude trajectories. As in Figure 2-6B, human boundaries are plotted from van Noorden (1975). Small black dots indicate stimulus parameters for which we obtained model judgments. For each model, the shaded area (‘bistable’ region) covers the stimulus parameters for which the log odds fell between -12 and 12 nats. For all parameters, the lesioned model favours the “two-source” explanation more than the original model. Whereas the bistable region of original generative model overlaps with the area between the human thresholds, the lesioned model has a bistable region below humans’ two-source boundary (i.e., where human listeners cannot hear the two source percept).

2.2.6 Model results on everyday sounds

The structure of the generative model was inspired by our observations of everyday sounds. However, for reasons of both tractability and explanatory parsimony, we prioritized general principles of sound generation rather than high-fidelity sound synthesis which could flawlessly reproduce naturalistic sounds. Samples from the model are thus typically not fully naturalistic in appearance. Nonetheless, the model can be applied to any sound waveform, allowing us to test it on everyday sounds. Can the same generative principles that explain auditory scene analysis in experimental settings also explain the perceptual organization of naturalistic sounds? We evaluated the generative model on naturalistic sound mixtures from the Free Universal Sound Separation dataset (FUSS) (Wisdom et al., 2021), with the aim of elucidating the gap between understanding auditory scene analysis with synthetic sounds and everyday sounds.

As the basis for quantifying the generative model’s match to human perception of naturalistic sound mixtures, we identified four key ways that the model’s inferred sources could deviate from human perceptual organization:

1. **Unrecognizability:** the model infers a source which people do not hear in the mixture
2. **Absence:** the model omits a source that people do hear in the mixture
3. **Over-segmentation:** the model segregates sounds into distinct sources, but people hear these sounds as coming from a single source
4. **Over-combination:** the model combines sounds into a single source, but people hear these sounds as coming from distinct sources

We aimed to understand how often and in what circumstances the generative model’s inferences deviated from listeners’ perceptual organization. Experiment 1 addresses the first deviation type, identifying which model-inferred sources are unrecognizable

to human listeners in the mixtures. Experiment 2 addresses the other three deviations: absence, over-segmentation, and over-combination. In addition to Supplementary Figures 2-14–2-21 which depict rendered model inferences, we provide an online repository that includes audio through my thesis webpage (<https://mcdermottlab.mit.edu/mcusi/thesis/>).

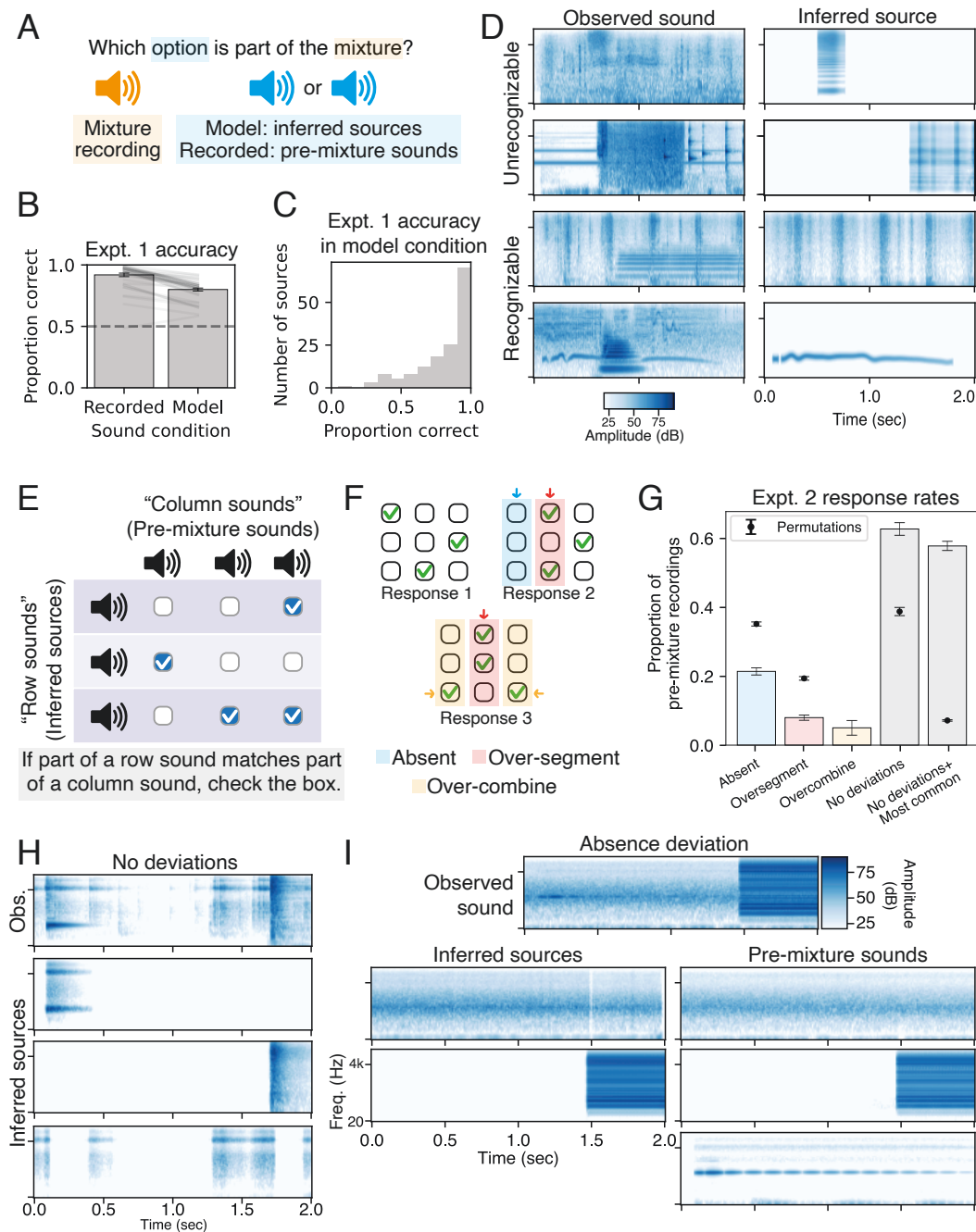


Figure 2-8: Model results on everyday sounds

A) Experiment 1 tests if inferred sources are recognizable in the mixture. B) Mean results by condition for Experiment 1, $n=45$. Accuracy in the model condition is above chance. C) Histogram of proportion of participants responding correct per inferred source in the 'model' condition. It is skewed left, indicating that unrecognizability errors occurred for a minority of inferred sources. D) Examples of source inferences with low and high recognizability in Experiment 1. E) Experiment 2 paradigm. F) Examples of tallying up deviations per premixture sound. Response 1 has no deviations tallied. Response 2 has an absence deviation for column 1, and an oversegmentation error for column 2. Response 3 has an overcombination deviation each for columns 1 and 3, and an oversegmentation deviation for column 2. G) Mean proportion of pre-mixture recordings for each deviation type, averaged across participants ($n=7$). Chance level computed from random permutations of participant responses, constrained to have one checkmark per row. H) Inferred sources with no deviations from the premixture sounds. I) Sound with an absence deviation from the premixture sounds.

Experiment 1.

The purpose of experiment 1 was to identify when the model inferred sources which human listeners did not recognize in the mixture. On each two-alternative forced choice trial, participants heard three sounds (Figure 2-8A). First, they heard one of the two-second sound mixtures. They then listened to two potential sources and chose which was part of the mixture. On ‘recorded’ condition trials, both potential sources were premixture sounds used in the FUSS dataset: the correct answer was a premixture sound for the mixture presented in the trial, while the incorrect answer was an unrelated premixture sound. On ‘model’ condition trials, both potential sources were inferred sources: the correct answer was a source inferred from the trial mixture, while the incorrect answer was inferred from a different mixture.

We found that most of the model’s inferred sources were recognizable in the mixture. As shown in Figure 2-8B, performance in both conditions was far better than chance (recorded 95%CI=[0.89, 0.95], model 95%CI=[0.78, 0.82]). For each inferred source, we computed the proportion of participants that correctly recognized the source as present in the mixture. A histogram of this proportion is skewed left (Figure 2-8C), indicating that unrecognizability errors were limited to a modest number of inferred sources. Examples of unrecognizable and highly recognizable source inferences are shown in Figure 2-8D. These results indicate that the model often successfully infers sources that emit sounds that are recognizable in the mixture.

Experiment 2.

The purpose of experiment 2 was to identify when the model deviates from perceptual organization via absence, over-segmentation and over-combination of sources. We relied on a matching task between the inferred sources and premixture sounds, depicted in Figure 2-8E. On each trial, participants listened to inferred sources and premixture sounds from the same sound mixture, which were arranged in a grid. Participants were instructed to place a checkmark whenever part of a “row sound” (inferred source) matched part of a “column sound” (premixture sound). We excluded both the pre-

mixture sounds and model inferences which listeners could not reliably recognize as part of the mixture in Experiment 1, and therefore required that participants place at least one checkmark for each inferred source.

This setup allowed us to tally instances of when the inferred sources deviated from each premixture sound, in terms of perceptual organization (Figure 2-8F). If participants provided a one-to-one response for the inferred and premixture sounds, no deviations were tallied. If a column did not have any checkmarks, this was tallied as one absence deviation from the corresponding premixture sound. If a column had more than one checkmark, this was tallied as an over-segmentation deviation for the corresponding premixture sound. Finally, if a row had more than one checkmark, this was tallied as an over-combination deviation for each checked premixture sound. Although premixture sounds might not always reflect human perceptual organization of the sound mixture, they often effectively pointed to a coherent perceptual entity in the mixture, and so provided a convenient perceptual reference for participants to describe how sound is allocated in the inferred scene.

Listeners' judgments had good inter-rater reliability, computed for inferred/premixture sound pairs ($ICC(2, 1) = 0.80$, $95\%CI=[0.77, 0.83]$). The average results across participants are shown in Figure 2-8G. To obtain chance performance for each statistic, we randomly permuted each participant's responses subject to the task constraint that there be one checkmark in each row, then averaged over permuted participants. The mean proportion of premixture sounds having a one-to-one correspondence with an inferred source exceeded chance level ('no deviations', $p<0.001$ by permutation test). Moreover, the mean proportion of premixture sounds which had a one-to-one correspondence and for which a participant responded with the most common response was only slightly reduced, showing that participants tended to respond consistently when they choose one-to-one correspondences ('no deviations+most common', $p<0.001$ by permutation test). By contrast, the mean proportions of premixture sounds with an absence or an oversegmentation were significantly less than chance ($p<0.001$ by permutation test). Finally, overcombination errors occurred, with a small mean proportion of premixture recordings with an overcombination deviation

(two-tailed 95%CI=[-0.006,0.111]). Figures 2-8H and 2-8I depict examples from two response categories. These results demonstrate that the model can infer valid perceptual organizations from everyday sounds.

Qualitative investigation.

Overall, listeners could recognize a majority of the sources inferred from natural sound mixtures (examples in Supplementary Figure 2-14), and many model inferences had reliable one-to-one correspondences with the source recordings used to generate the mixture (results by premixture sound in Supplementary Figure 2-13). To better understand the model’s successes and failures, we qualitatively examined participants’ judgments for individual sound mixtures, linking them to acoustic structure and sound category labels. While some failures inevitably reflect the difficulty of inference (see Discussion), many point to generative principles underlying everyday sounds that are missing from our model, and that could be productive directions to explore in the future.

Cochleagram input. One commonality to many of the least recognizable inferred sources was a mismatch in the periodicity of the inferred source and sound mixture. In one instance, the model explains a noisy fireworks explosion with a harmonic source (Supplementary Figure 2-15). This failure is plausibly due to the cochleagram used for the likelihood representation – this representation was chosen for computational efficiency, but limits the resolution with which periodicity can be measured. We suspect this impoverished input may also cause absence deviations, for example with quiet tones in noise (Supplementary Figure 2-16).

Hierarchical sources. Sources are independent in our model, but this assumption is insufficient for everyday sounds. We find oversegmentation deviations that occur when various sound-generating processes within a single premixture clip share a common cause (Supplementary Figure 2-17), such as the different sounds produced by brief and extended strokes of chalk over a board. Related oversegmentation deviations occur with causally linked sources of different sound types (Supplementary Figure 2-18), such as breath noise along with pitched sound of a flute note. These deviations

suggest that our model should express sources with a hierarchy between multiple sound-generating processes.

Diversity of everyday sound spectra. Premixture sounds of impacts reveal the inadequacy of our generative model of spectrum and amplitude to explain everyday sounds. We find that the model oversegments frequency components of impact sounds that decay at different rates (Supplementary Figure 2-19), because it cannot explain frequency-dependent decay as being due to a single source, despite the frequent occurrence of this type of sound in the environment. The model also sometimes overcombines perceptually distinct impact sounds. For instance, in the example shown in Supplementary Figure 2-20, a sequence of bounce sounds is combined with a violin sound, because the model cannot represent the spectral differences that distinguish the two types of sounds. These results point to an oversimplistic model of source spectra, and how these spectra change over time.

2.3 Discussion

Inspired by everyday sounds, we built a probabilistic generative model that describes sources which emit events to produce sound mixtures. To address the challenges of Bayesian inference inherent to inferring sources from audio, we combined deep learning with search: a neural network made event proposals, which were sequentially combined into sources and evaluated with stochastic variational inference. The model qualitatively matched human perception on a variety of classic auditory scene analysis phenomena. In contrast, contemporary source separation networks could not account for perceptual organization, whether trained on datasets of recorded audio or samples from the generative model, revealing the difficulty of matching perception comprehensively in the absence of a generative model. Selective model “lesions” revealed the importance of a distribution over sources that allows for variation in source properties. Despite the model’s simplicity, experiments with human listeners showed that the model could also explain the perceptual organization of many natural sound mixtures. The model failures were instructive, revealing the need for specific

additional generative principles. Some deviations additionally pointed to perceptual phenomena that have not been extensively studied (e.g. the perception of hierarchical structure in sound sources) and thus constitute directions for future work. Our model provides a relatively comprehensive computational explanation of classic results in auditory scene analysis, while also being applicable to natural sounds, and explaining aspects of their perceptual organization.

2.3.1 Relation to prior models

Modeling perception with Bayesian inference is a longstanding enterprise that has inspired successful applications and productive debate. Yet, the difficulties of specifying a rich world model and performing inference from sensory signals has limited applications to human perception (Ellis, 2006; Ma, 2012). To meet these challenges, we applied new tools to integrate and extend key elements of previous work into a unified and functional computational system. Although our model is not a full account of auditory scene analysis, it provides an example of how theories of perception could in principle be instantiated and tested.

Structured models

The basic generative structure of our model posits that parametrized sources produce sound via discrete events. Among previous models that explicitly considered discrete events, most treated them as pre-determined symbolic input to be grouped (Barniv & Nelken, 2015; Larigaldie et al., 2021; Mill et al., 2013). These models typically were intended to explain the sequential grouping of tones, but cannot be applied to raw audio (where the events are not explicit), and thus cannot be evaluated on natural sounds. Our model is more closely related to earlier work on computational auditory scene analysis (Barker et al., 2005; Ellis, 1996), containing events as one level of a hierarchical description of auditory scenes that is inferred from audio. We extended this work by both specifying a rich, probabilistic relationship between sources and events (the hyperprior over sources accounts for a diversity of event regularities while

continuous event variables have Gaussian process priors) and by making inference work in a rich generative model, which was not possible in earlier efforts along these lines. One previous model used Gaussian processes to model time-varying audio for scene analysis, but did not attempt to model sources with events (Turner, 2010). We also developed novel modeling commitments in order to explain auditory scene analysis phenomena, including source-specific kernels and non-stationarity between events. Finally, we integrated these features within a probabilistic program which could express an arbitrary number of sources and events. Enabled by a new generation of computational tools, the resulting world model is substantially more expressive than prior work.

Clustering

Some computational approaches to auditory scene analysis instead frame perceptual organization as clustering bottom-up features of audio (Chen et al., 2017; Hershey et al., 2016). This approach has often been combined with bottom-up features derived from neurophysiology (Brown & Cooke, 1994; Chakrabarty & Elhilali, 2019; Elhilali & Shamma, 2008; Krishnan et al., 2014; Wang & Brown, 1999), but it remains difficult to identify features whose clustering is comprehensively predictive of perceptual organization. Our approach shares some abstract similarities in combining a bottom-up processing component with a top-down inference stage, but a key difference is that the bottom-up component of our analysis-by-synthesis algorithm preserves uncertainty, outputting proposals for event latent variables that can be refined or rejected during inference. Maintaining uncertainty over the event proposals was critical to robust inference, particularly with everyday sounds.

Machine hearing

We compared our model to source separation neural networks from machine hearing. These particular networks were not developed with the intention of replicating human perception, but this broad class of models has had success in explaining human auditory perception in various other domains (Francl & McDermott, 2022; Kell

et al., 2018; Saddler et al., 2021), and analogous models trained on source separation might be envisioned to account for aspects of human auditory perceptual organization. Broadly speaking, the source separation networks were trained to reconstruct a set of premixture sounds from a mixture sound. We found that a suite of networks with varied training datasets, supervision, and architectures did not comprehensively match human perception as well as our model. The match to human perception was worse even for a source separation network trained on samples from our generative model. It is possible that source separation provides the wrong task constraints, or that these particular networks are not good enough at source separation in order to sufficiently constrain their representations. But the results underscore the challenge of accounting for human perceptual organization, and suggest the benefits of coupling a generative model with neural networks.

In addition to better accounting for perception in the domain examined here, generative models confer several desirable traits as compared to pure neural network systems. First, generative models provide a language for specifying structured and detailed scene descriptions. Without an explicit generative model to create input data labelled with latent variables, human annotators are required to create data needed for supervised training, which can be inefficient. Moreover, the labels that are convenient for humans to annotate may be limited, e.g. to verbal class labels. In contrast, our generative model uses a structured hierarchy of latent variables to describe a source: multiple source-level variables along with a sequence of events that the sources generate, which each include their own generative parameters (e.g. temporal trajectories). The source-level latent variables alone would likely be methodologically difficult for a human annotator to label. In addition, as demonstrated with model lesions (Figure 2-7), we can manipulate the probability distribution over scenes to help understand the assumptions that underlie the model’s inferences. Such manipulations of training data can be paired with neural network models of perception (Frans & McDermott, 2022), but are facilitated by a model to generate the data. Second, our model explains human perception with a set of interpretable principles. As shown in the everyday sounds experiments, this interpretability helps to identify what is miss-

ing in the model and how to address its shortcomings. Since the model is composed of meaningful parts, we can augment it by adding more meaningful parts. For example, to create a composite source that emits both noises and harmonics, we could combine the noise and harmonic renderer and define how their latent variables co-vary. It would also be possible to add entirely new generative modules if needed, such as reverberation that filters the source sounds. Finally, learning and structured models are not mutually exclusive. We discuss the possibility of combining the benefits of both approaches in the Future directions.

2.3.2 Limitations

Likelihood

A major limitation of our model is the cochleagram likelihood representation. We found that the model quantitatively deviated from human results for phenomena that plausibly depend on either a high-resolution representation of frequency (e.g. harmonic mistuning) or time (co-modulation masking release), which likely reflects the limited resolution of the cochleagram. This deficiency was especially apparent for periodic sounds, which also limited the model’s performance with many everyday sounds. Some previous work supplemented a cochleagram with an explicit periodicity representation for this reason (e.g., Brown and Cooke (1994), Elhilali and Shamma (2008), and Ellis (1996)). We deviated from this previous work for reasons of efficiency (in terms of speed and memory for iterative computation). Some alternative representations that might better capture periodicity include the correlogram (Slaney & Lyon, 1993), sparse periodicity acoustic features (Josupeit et al., 2020), wavelet scattering transforms (Lostanlen et al., 2018), wefts (Ellis & Rosenthal, 1995) or multi-scale cochleagrams (Engel, Hantrakul, et al., 2020). Another possibility would be to use representations learned by a neural network to extract periodicity from a high-fidelity simulated cochlear representation (Saddler et al., 2021). But the challenge of defining appropriate mid-level representations for the likelihood extends beyond periodicity. For example, the cochleagram also seems poorly suited for generative inference on

sound textures (McDermott & Simoncelli, 2011), the details of which will affect the cochleagram-based likelihood despite being inaccessible to human listeners (McDermott et al., 2013).

Model structure

The generative model is limited by a relatively simplistic model of spectrum and amplitude. The model lacks event-linked structure in amplitude (e.g., amplitude decay after an impact) as well as a time-varying spectrum (e.g. frequency-dependent decay). This limitation was apparent in the everyday sound experiments, and although it sufficed for many classic auditory scene analysis phenomena, it also prevented us from modeling some classical results, notably the asymmetry between onset and offset asynchrony (Darwin & Sutherland, 1984) and the perceptual segregation caused by repetition (McDermott et al., 2011). Research into the spectral factors that affect perceptual organization has mainly focused on the sequential grouping of musical instrument notes (Gregory, 1994; Wessel, 1979). Our model exposes the extent to which everyday sounds include a much broader set of spectral structures, necessitating a more sophisticated and/or multifaceted spectral model (see Future directions).

Inference

Many of the model limitations reflect choices made to facilitate inference. Yet inference was still not completely robust. The everyday sound experiments revealed three classes of inference failures. First, variational inference generally could not recover when the event proposals from the neural network were far from correct, for example proposing the wrong sound type or missing a proposal for a whistle in noise (Supplementary Figures 2-15,2-16). Second, gradient descent could get caught in local minima, for example, making it difficult to adjust the fundamental frequency of a harmonic tone or continue a quiet sound behind a masker (Supplementary Figure 2-21). Third, the set of hypotheses maintained during sequential inference can become very similar to one another over time, a problem known as ‘degeneracy’. Discarding alternative explanations too early in the timecourse of a sound may set up inference

to fail later, when subsequent evidence would favour those explanations.

However, maintaining multiple hypotheses also meant that inference required massive amounts of computation, to an extent that may be unrealistic for a biological system (see Methods). For these reasons, we do not consider our current inference algorithm to provide a mechanistic or algorithmic-level explanation of perceptual organization. Rather, it is a means to implement the computational-level explanation provided by our model (Marr, 1982). Inference is nonetheless a bottleneck for this type of modeling, and an important open scientific issue in perception and neuroscience. The combined efficiency and efficacy of perception demands explanation from both an algorithmic and neuroscientific perspective (see Future Directions).

2.3.3 Future directions

Hierarchical organization

Our model results on everyday sounds illustrate a nuance of perceptual organization which the model fails to fully capture, in which sources are perceived to contain hierarchical structure. Consider a rhythm played by the bass, snare and cymbal of a drumkit, or a composite sound with simultaneous periodic and aperiodic components (e.g. a breathy flute note). The components within such sounds can be perceived simultaneously as distinct and linked. We suggest these scenarios are akin to visual hierarchical grouping (Baylis & Driver, 1993; Froyen et al., 2015; Palmer, 1977). Sound ontologies for auditory scene analysis have advocated for multi-layered scene descriptions to describe such scenarios (Gaver, 1993b; Nakatani & Okuno, 1998), but we know relatively little about the perception of such hierarchical structure. There has been some consideration of such hierarchical structure in sequences of tones (Larigaldie et al., 2021), and in music instrument synthesis (Engel, Hantrakul, et al., 2020; Serra & Smith, 1990). But we know relatively little about the perception of hierarchical structure in ecological sounds (e.g. a glass bouncing then shattering, Warren and Verbrugge (1984)), particularly in the context of scene analysis.

One open related question is whether perceptual organization in these cases is

based on physical-acoustic interpretations. For example, our model produces an over-segmentation deviation for the sounds of a ball that bounces on the floor and then strikes a metallic surface (Supplementary Figure 2-18C). The acoustical structure of the two resulting acoustic events is dissimilar, but humans hear an integrated causal sequence. One way to account for such examples is with “schema-based” auditory scene analysis, in which the auditory system learns to group patterns of sound which recur in the environment, as is thought to aid the streaming of music, speech and arbitrary sound patterns (Billig et al., 2013; Dowling, 1973; Woods & McDermott, 2018). But it is also possible that the auditory system constructs more specific causal models, for instance representing sounds from physical events using internal models of physics (Gaver, 1993b; Giordano & McAdams, 2006; McAdams et al., 2010), as is thought to occur in vision (Gerstenberg et al., 2021; Scholl & Tremoulet, 2000; Yildirim, Siegel, et al., 2020). Physics-based sound synthesis should enable additional work in this direction (Agarwal et al., 2021; Rocchesso & Fontana, 2003; Traer et al., 2019).

Sound textures

Another notable model shortcoming occurred for sound textures (McDermott & Simoncelli, 2011). In one suggestive example, the model explained a mixture of temporally overlapping textures (running water and brushing teeth) as a single noise source, whereas we could perceive them to be two separate streams. The model is limited both by a noise model that does not capture all of the statistical regularities that differentiate natural textures (McDermott & Simoncelli, 2011), and by a likelihood function that is applied to the cochleagram rather than to a statistical representation akin to that thought to determine human texture perception (McDermott et al., 2013). But this limitation also exposes the need to better understand the role of texture in auditory scene analysis. The auditory system is known to ‘fill-in’ textures when they are masked (McWalter & McDermott, 2019), indicating that multiple textures can be heard at the same time, and to selectively average sound elements attributed to a texture (McWalter & McDermott, 2018). But we know little about the conditions in

which humans segregate some texture mixtures, pointing to an important direction for both future experimental and modeling work.

Alternative scene descriptions

The limitations imposed by the model’s scene descriptions raise the question of what an ideal model’s descriptions should contain, which in turn raises open questions about the content of human perceptual experience. Our model’s scene descriptions were relatively abstract and signal-based, with a somewhat arbitrary division into noises, harmonic sounds, and whistles (although overlapping with categories used in Burger et al., 2012; Ellis, 1996; Gemmeke et al., 2017; Misra et al., 2009). These could be enriched and extended in many ways, but as discussed above, another possibility is for the model’s latent descriptions to more closely relate to physical descriptions of sound production (Gaver, 1993b). One related question is whether the source models in our heads are closer to a flat hierarchy of many source models that each describe a fairly specific set of sounds (impacts, speech, woodwind instruments, textures, modern electronic sounds, etc.), as opposed to a deep, unified hierarchy in which a modest number of primitives are composed to create more complex sounds. In addition to more descriptive source models, an ideal model of scenes should include other aspects of sound generation, such as reverberation (Traer & McDermott, 2016), spatial effects (Franci & McDermott, 2022), and other generative factors in ecological acoustics (Grinfeder et al., 2022; Traer et al., 2021; Wallach et al., 1949; Zahorik & Wightman, 2001).

Learning

Another alternative for designing source models is to learn them. One promising possibility would be to combine learned statistical components into a structured generative model. For example, one could swap out our classic Bayesian priors (Gaussian processes) over excitation and spectra for autoregressive neural networks (Wu et al., 2022) trained to synthesize natural sounds through the model’s renderer. The discrete sound types could also be swapped out for a more flexible learned representation. The

excitation could instead be modeled as a point in a high-dimensional space that can express a combination of types: for example, a set of parameters could specify the noisy part of the excitation and another set of parameters could specify the periodic part. Starting with a uniform prior over this continuous space, we could infer scene descriptions for a variety of everyday sounds. This could ideally result in clusters of inferred type parameters resulting from the structure of everyday sounds. We could then fit a mixture distribution to these inferences, to use as the new prior over type parameters. After updating the prior, we could then refine the inferences of the scene descriptions for everyday sounds. Iterating on this process until convergence would result in a prior over type parameters. Integrating learned components into a structured model would allow the model to capture a richer class of sounds accurately while still maintaining interpretability (Feinman & Lake, 2020), but the tractability of this extension would depend on the efficiency of inference (Ellis, 2020).

2.3.4 Inference

If Bayesian models are to be taken as mechanistic explanations of the human auditory system, then its key algorithmic formalism - search through a hypothesis space - must be shown to be plausible. We propose that modeling everyday sound mixtures is a useful exercise towards this goal, because we found that listening in experiments versus more everyday, unconstrained conditions tend to have different inferential demands. Typically, psychophysical experiments are designed to be ambiguous with respect to just a few reasonable perceptual hypotheses, in order to isolate the effect of a single variable. In experiments, and in most classic illusions, listeners are instructed to report one of these hypotheses, and are given practice to help them get accustomed to doing so. Because the hypotheses are delimited to a small set, the inferential difficulty lies in accurately estimating the posterior distribution: it may be difficult to estimate the relative probability mass between the two hypotheses because they can be different in subtle ways (eg. the presence of a quiet tone, or the placement of a formant). In contrast, when inferring explanations for everyday sound mixtures, there are many more plausible hypotheses in a model rich enough to explain them. It

appears as if the local information is less diagnostic in such settings, but the overall global scene interpretation is less ambiguous (as noted for visual inference by Yuille and Kersten, 2006), particularly with respect to the discrete latent variables. This means that adjudicating between hypotheses is easier than for typical experimental settings, but that search is much harder. We thus think that testing models on everyday sounds is critical for future progress, as it exposes challenges that are less apparent in classical experimental settings.

Search

As mentioned above, we found that maintaining plausible alternative hypotheses for a sound over time was important for successful inference. We suggest that solving the challenge of narrowing search while maintaining multiple hypotheses is central for a mechanistic account of perceptual organization. For our model, using an amortized inference network to propose events was necessary for tractability - the network narrows the search space to a small subset of all possible events. But because these event proposals were combined into scene hypotheses, assessing all of their combinations is intractable when there are many event proposals (as was typically the case for everyday sound mixtures). Search relied on simple heuristics to prioritize smaller scene descriptions, as well as substantial parallel computing resources. If perception is solving a similar search problem in real-time, it seems likely to utilize more efficient procedures to search this combinatorially large space. We suggest that more plausible algorithmic accounts could be discovered by learning procedural knowledge. For instance, over many inference trials, we could observe which combinations of events proposals are successful versus unsuccessful. These could then be used to learn how to prioritize the assessment of certain combinations (Cusumano-Towner & Mansinghka, 2018; Gothoskar et al., 2021), in place of the simple hand-designed heuristics that we used here. Other aspects of the search algorithm could also be replaced by learnable components (e.g. the stochastic gradient descent used for hypothesis optimization Andrychowicz et al. (2016)). An ambitious future direction would be to replace the entire search algorithm with a fully learned procedure, as has been done for simple

compositional graphics programs (Eslami et al., 2016).

Time

Our model was designed to perform joint inference over the sources and events that produced an entire observed sound (2 seconds in our everyday sound experiments). However, the application of this procedure to longer sounds raises two challenges, due to the long-range temporal dependencies that arise. First, the computational cost of inference increases with duration, prohibiting us from running our model on longer sounds. In addition to increased computation, the posterior becomes higher-dimensional and thus requires more optimization to estimate. Second, human perceptual judgements are sensitive to context within a local window, and do not integrate evidence from an arbitrarily long time horizon (Carlyon et al., 2009; McWalter & McDermott, 2018). The extent of this ‘context’ remains an open scientific problem. Both of these issues could be addressed by explicitly including memory in a scene analysis model. One possibility is to modify the sequential inference procedure such that variables eventually stop being actively inferred and become fixed as a memory trace, no longer affected by further observations but potentially informing future variables. Such an approach could enable investigation of the timescales for which inference improves in complex scenes, which could clarify whether the temporal extent of memory and postdiction are adapted to the scale of temporal dependencies in natural scenes, or whether they mainly reflect resource constraints.

Another challenge for the future lies in understanding the perception of time. Our model currently cannot explain listeners’ inability to judge relative timing across sources that are perceived as distinct (Bregman & Campbell, 1971; Micheyl & Oxenham, 2010). It is unclear how such effects would arise from a generative model like ours, in which the scene description treats the perceptual timing of events as objectively reflecting their timing in the stimulus.

Attention

The richness of our model structure combined with its ability to be applied to everyday sounds provides a formalism with which to investigate attention. Attention can be construed as a mechanism that selectively refines parts of a perceptual hypothesis in order to deal with the intractability of inference in complex scenes (Whiteley & Sahani, 2012). One simple option to refine a scene hypothesis is to increase the dimensionality of the parametrization of the approximate variational posterior. For example, increasing the time-resolution in the variational posterior over the excitation trajectories would enable inference of finer-grained detail in those trajectories. Along these lines, attention could be instantiated in a model like ours by increasing/decreasing the complexity of the approximate variational posterior distribution for attended/unattended sources, respectively. Implementing attention in this way could help explain psychophysical benefits of attention (Woods & McDermott, 2015) in a normative framework, while yielding predictions for the effects of attention in naturalistic scenes. Generative models could also provide a formal hypothesis for the targets of “object-based attention” (Alain, Arnott, et al., 2000; Shinn-Cunningham, 2008), providing behavioral predictions about how attention spreads within a source or traverses across different levels of a hierarchical scene description.

Automatically generated illusions

A generative model of perception can both listen to and produce sound. One intriguing application is to use the model to automatically generate illusions by inferring the stimulus properties that would produce a desired percept. Illusion generation would correspond to an additional layer of inference (Chandra et al., 2022). One could formalize common illusion paradigms, such as multistability or conflicting global and local percepts, into objectives for optimization, providing a powerful additional model test.

2.3.5 Other sensory modalities

Many of the principles implemented in this paper can be traced to classical ideas in vision, where inference in generative models has long been proposed to underlie perception, but where computational systems that embody this approach have historically been a challenge to make work. Our work demonstrates the utility of auditory perception as a case study in perceptual inference. Audition has the advantage that relatively simple generative models can account for many everyday signals, making the approach tractable. However, the general principles and questions explored here are not modality-specific, and the time seems ripe to reapply this framework to other perceptual domains and to multi-sensory perception. Applying such an interpretable framework could give new insight into relationships between the worlds sensed by our perceptual systems.

2.4 Supplementary Figures

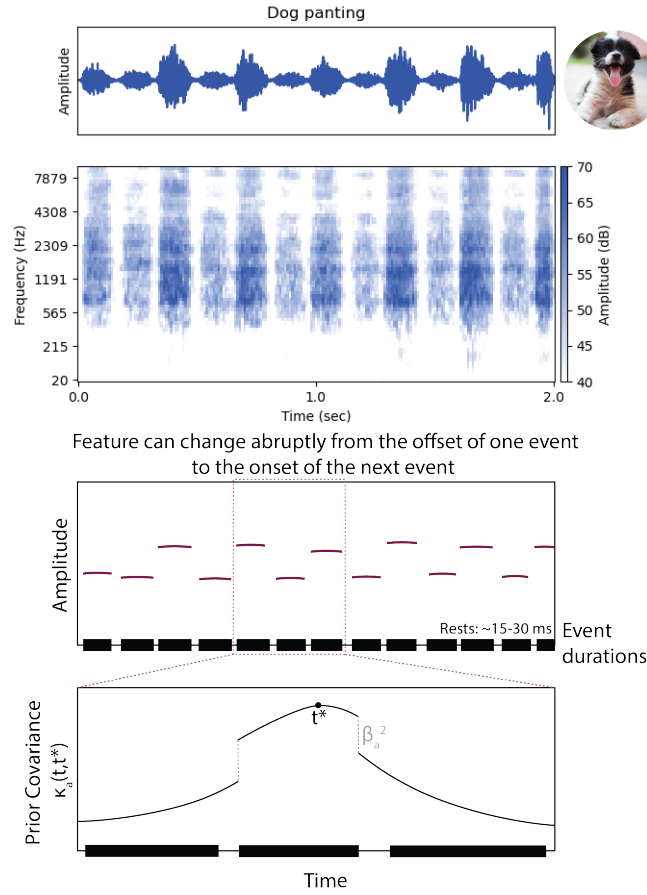


Figure 2-9: Event structure as implemented by the non-stationary kernel.

The generative model accounts for when a single source slightly changes its sound-generation process. Top two panels: sound and cochleagram of a dog panting, alternating between the in- and out-breath. Middle: Schematic latent variable description of the sound. The red lines indicate the amplitude trajectory, y-axis is amplitude level and x-axis is time. The black rectangles at the bottom depict the each event duration. Bottom: Non-stationary prior covariance kernel. Black rectangles at the bottom depict event durations. Timepoints are more correlated within an event then between events.

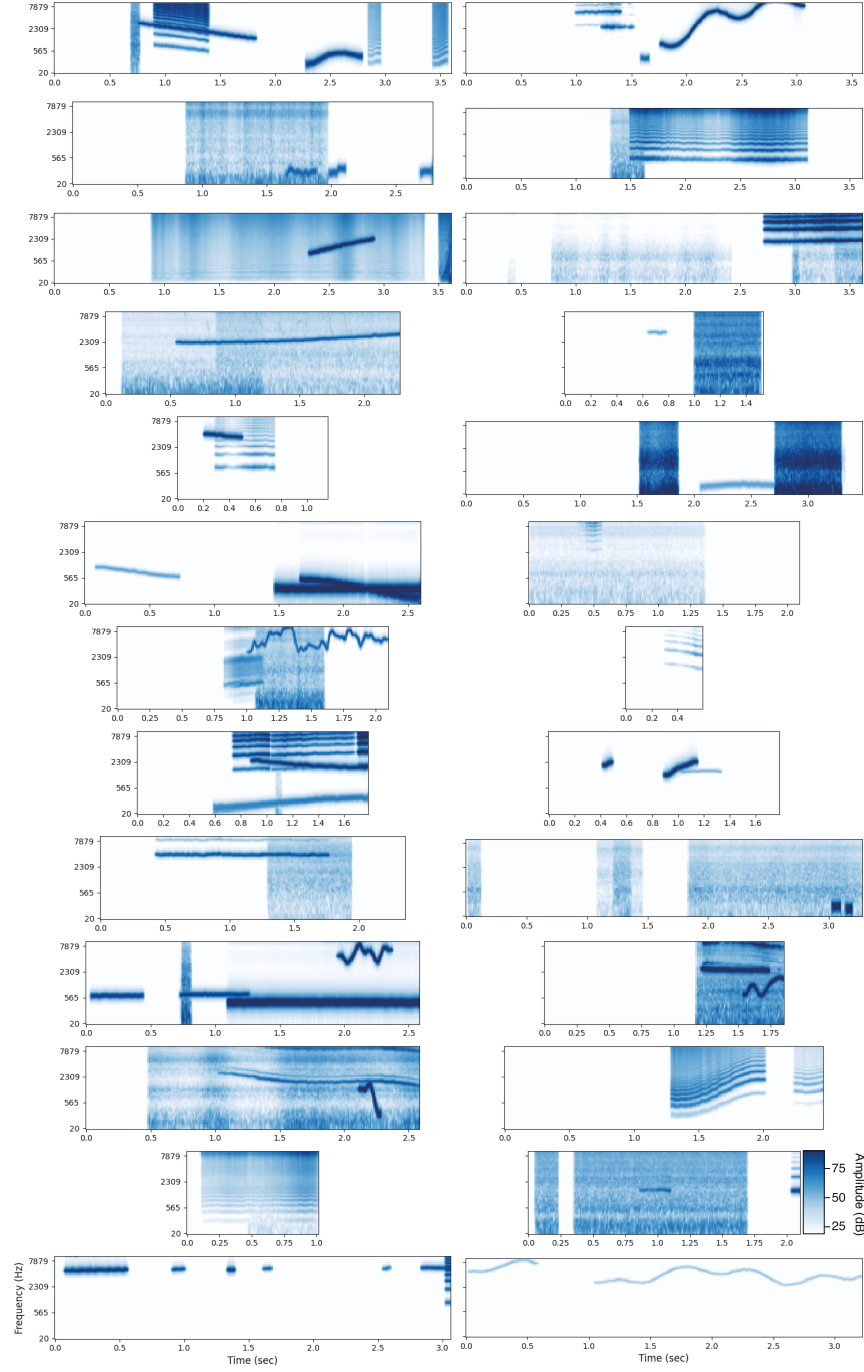


Figure 2-10: Scenes sampled from prior $P(S)$, rendered as cochleagrams.

Scenes sampled from prior show variation in the number of sources and events, event timing, frequency and amplitude modulation, and spectral shape. These sounds demonstrate the model's expressivity, while also revealing its simplifying assumptions (e.g. time-symmetry, raised cosine ramps for all events, etc.).

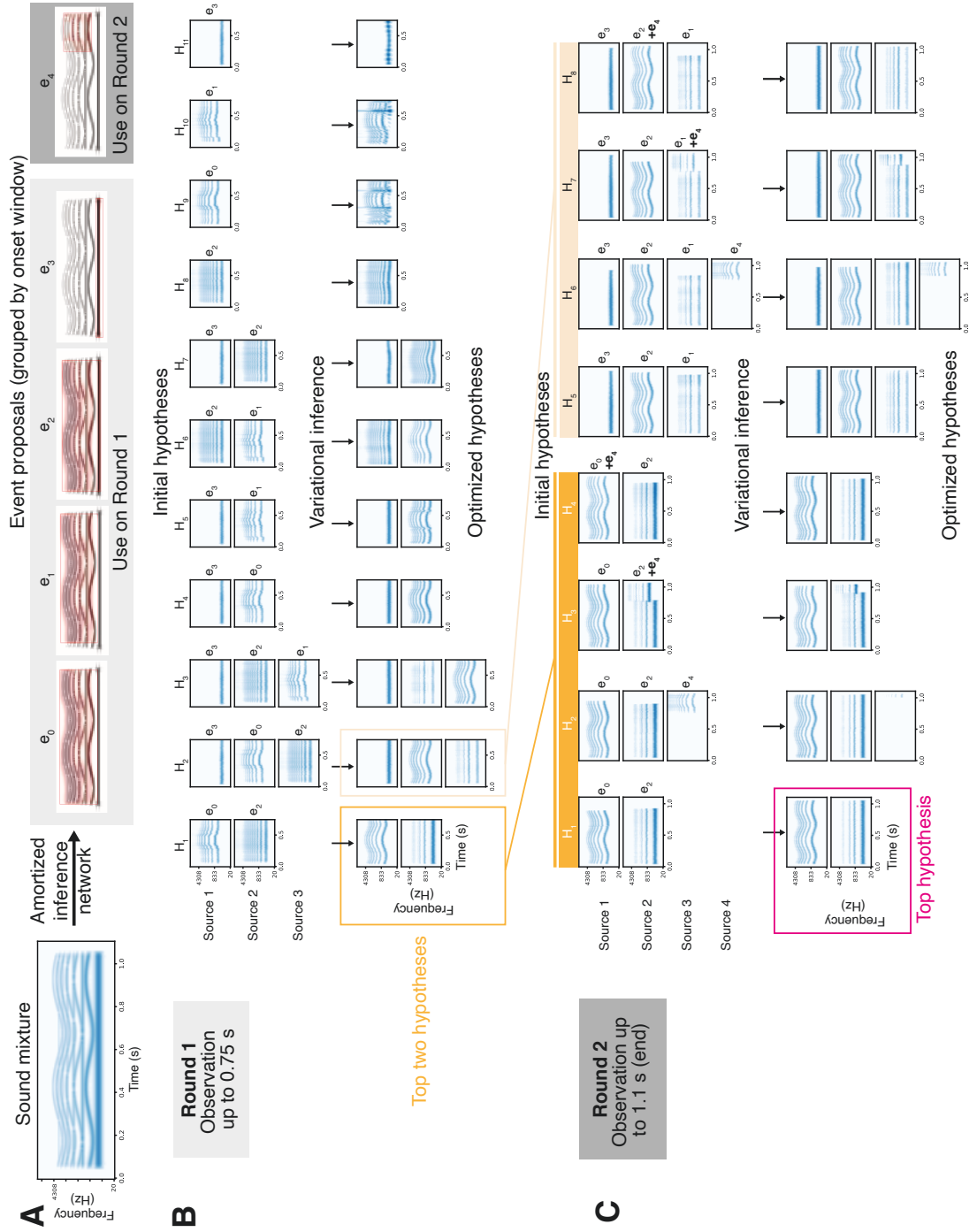


Figure 2-12: Example of sequential inference algorithm with frequency modulation.

A) Event proposal network. The observed sound mixture is 1.1 seconds long. Given the sound mixture as input, the amortized inference network returns a set of five event proposals. Each event proposal is depicted as a red mask overlaid on the sound mixture cochleagram, and is associated with a set of event latent variables (not shown). Event proposals with nearby onsets are considered in the same round of sequential inference. B) Round 1 of sequential inference. On Round 1 of inference, the first 0.75 seconds of the observation are considered and the first four event proposals are utilized. Initial hypotheses are constructed by combining different sets of event proposals subject to the heuristic constraints described in Section 2.5.4, resulting in eleven initial hypotheses ($\{H_i\}$; note that all of the actual hypotheses used during inference are depicted here, in contrast to Supplementary Figure 2-11). The initial hypotheses are optimized with variational inference, resulting in optimized event-level variables (as reflected by the change in the rendered cochleagrams) and source-level variables (not shown). The two hypotheses with the highest posterior probability are used in source construction for Round 2. C) The second round of sequential inference. The full observation is considered and therefore this is the last round. Initial hypotheses are constructed by using H_1 and H_2 of round one (dark and light grey, respectively), potentially combining them with event proposal e_4 . The initial hypotheses are optimized with variational inference. After variational inference, the hypothesis with the highest posterior probability is selected.

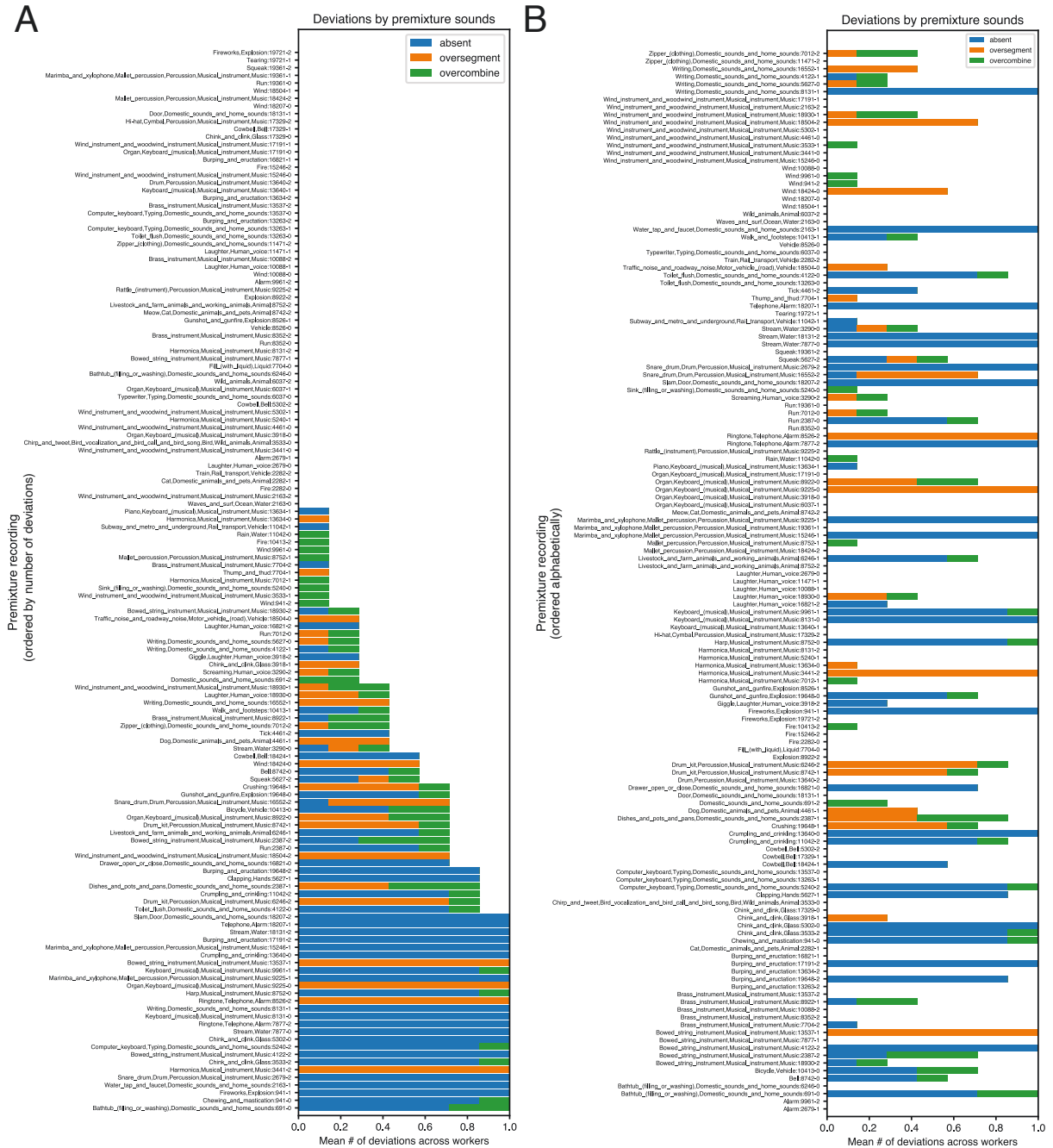


Figure 2-13: Experiment 2 data per pre-mixture sound (n=7).

Zoom on digital copy to see labels. A) Results by premixture clip in Experiment 2. Each bar displays the average number of deviations across workers, tallied for a particular premixture clip. The bars are ordered by total number of deviations, increasing from top to bottom. B) Same data, but ordered alphabetically by FUSS category label.

Examples of recognizable inferred sources in Experiment 1

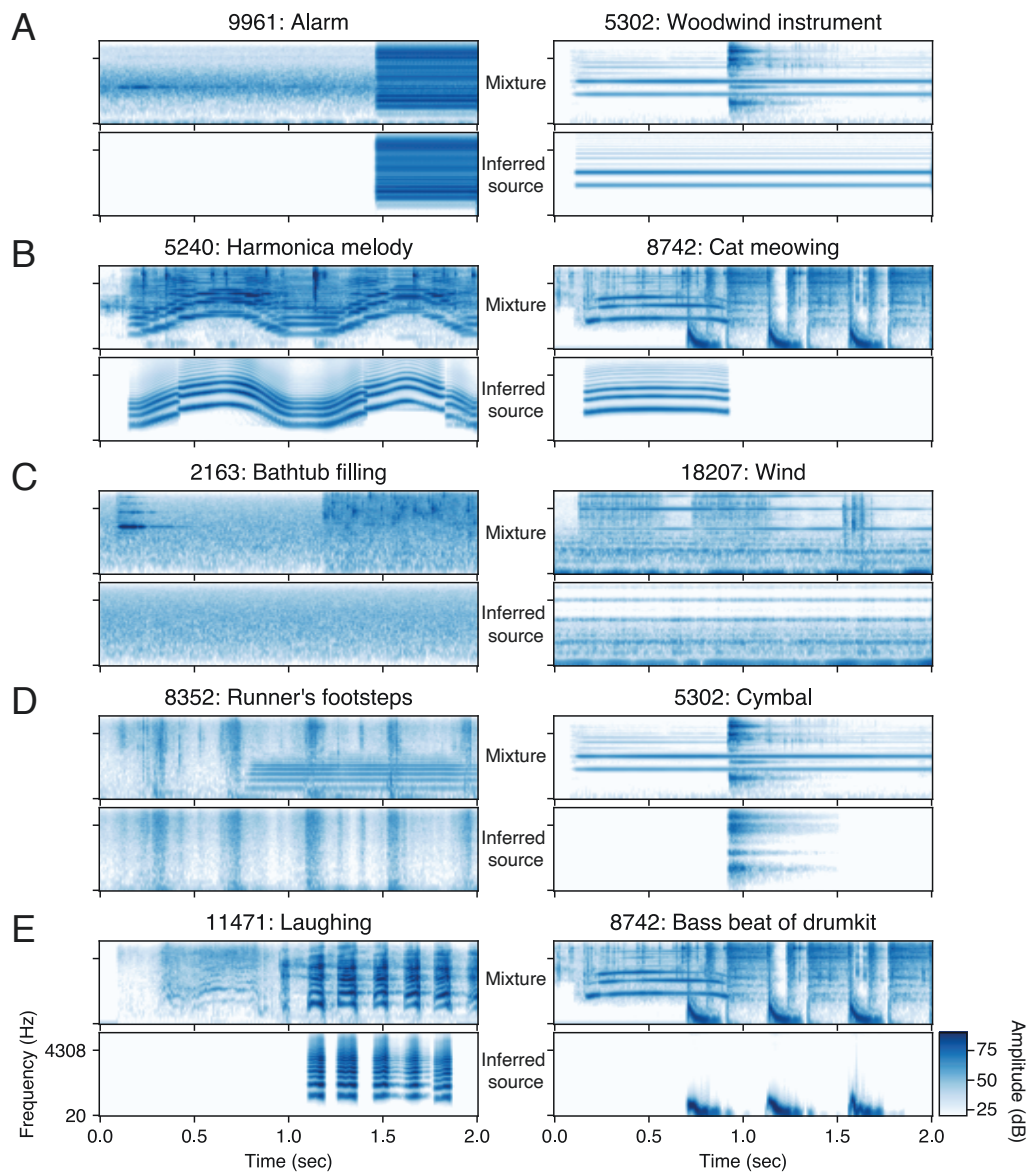


Figure 2-14: Examples of recognizable source inferences in Experiment 1. Title numbers refer to FUSS filename. A) Harmonic sources, with fairly static fundamental frequencies. B) Harmonic sources with dynamic fundamental frequencies. C) Background noises with relatively static amplitude. D) Broadband amplitude-modulated noises. E) Sound sequences.

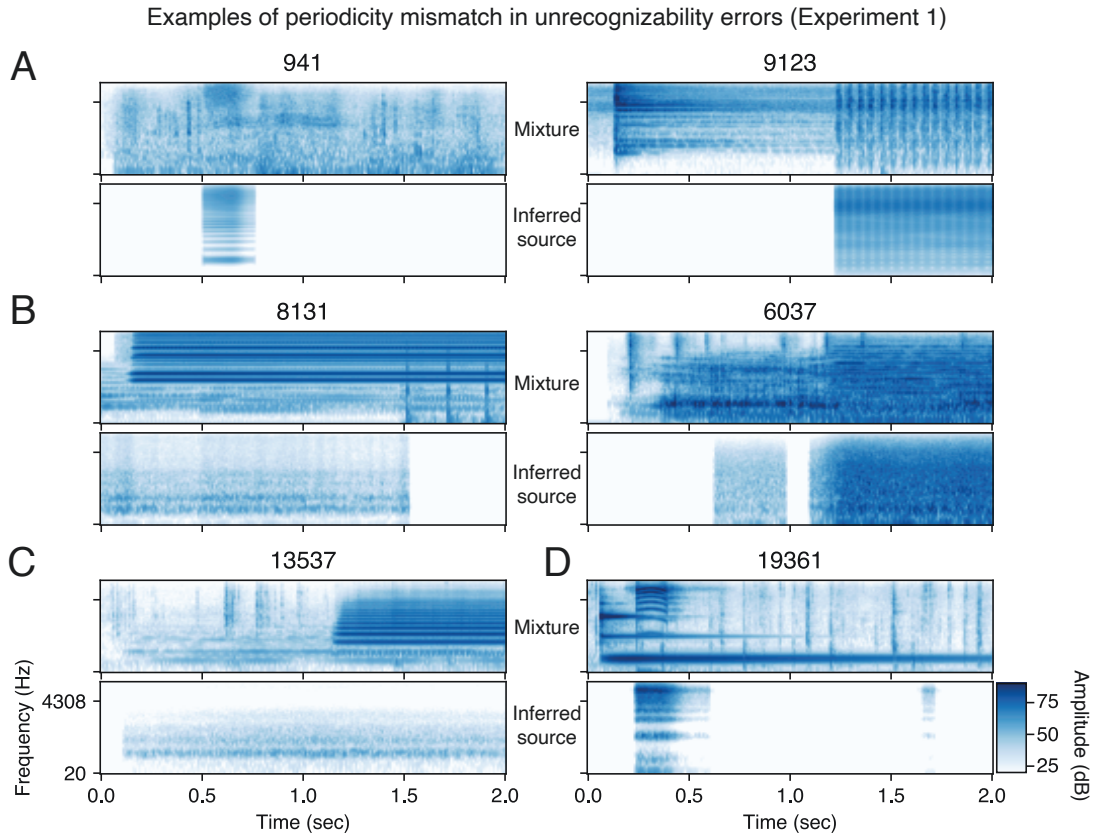


Figure 2-15: Examples of unrecognizable source inferences in Experiment 1, due to periodicity mismatch in the mixture and inferred source.

A) Noisy sounds are explained as harmonics. 941 contains a firework exploding during the duration of the inferred source. 9123 contains a clacking tool during the duration of the inferred source. B) Low-frequency instrument sounds with overlapping notes are explained as noises rather than periodic sounds. 8131 contains overlapping piano tones and 6037 contains a chord on the organ. C) Low-frequency instrument sound with just a single note is still explained as a noise rather than a periodic sound. 8752 contains a single note on the cello being bowed. D) High-frequency harmonic squeak is explained as a noise in mixture 19361.

Absence errors in Experiment 2 involving quiet tones in noise

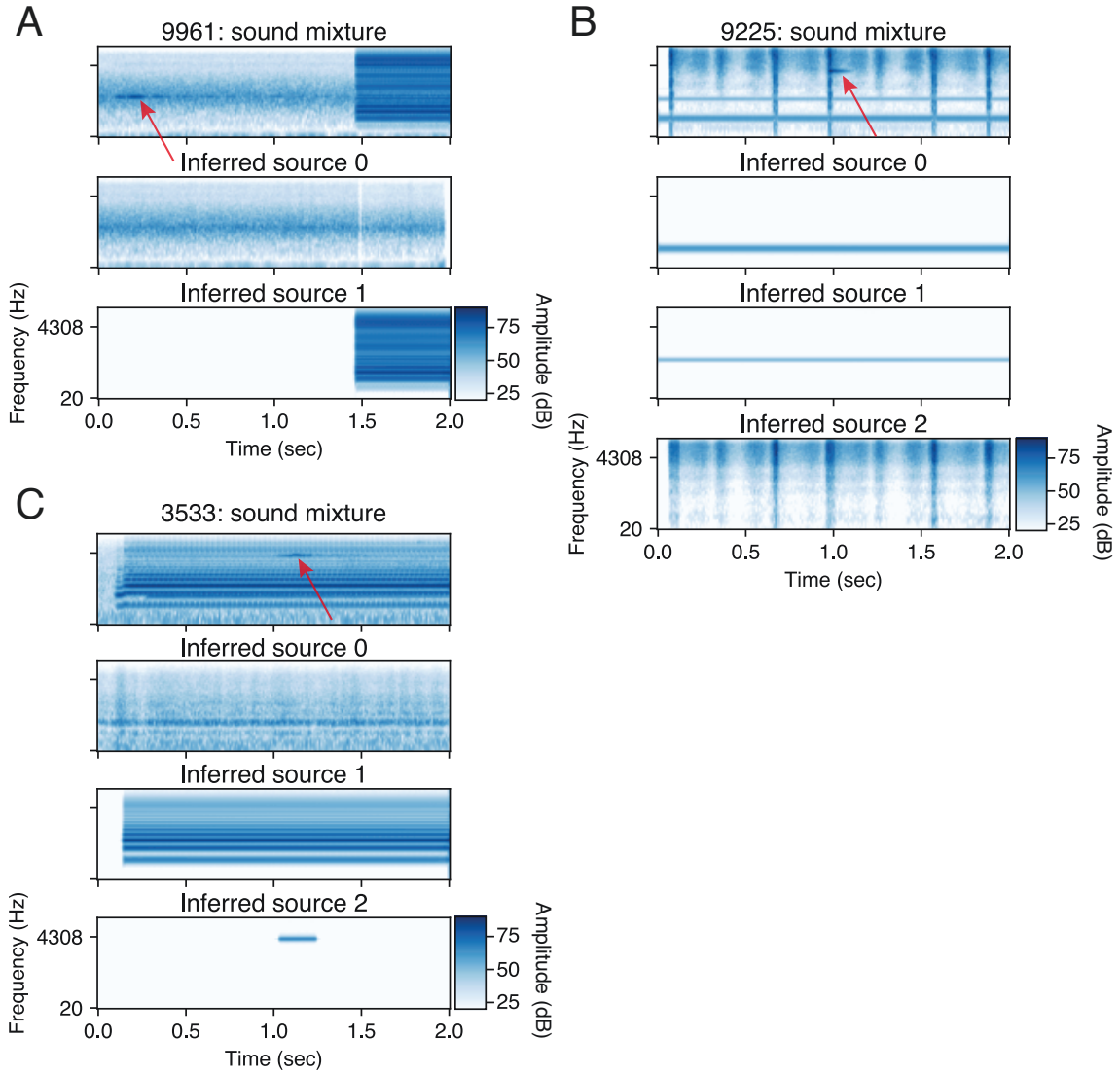


Figure 2-16: Examples of inferences when there is a quiet tone in noise. Red arrows indicate where the tones occur in the sound mixtures. A, B) Examples where the model fails. C) Example where the model succeeds. It is not impossible for the model to detect tones in noise, but it is likely that the lack of a periodicity representation hinders its ability.

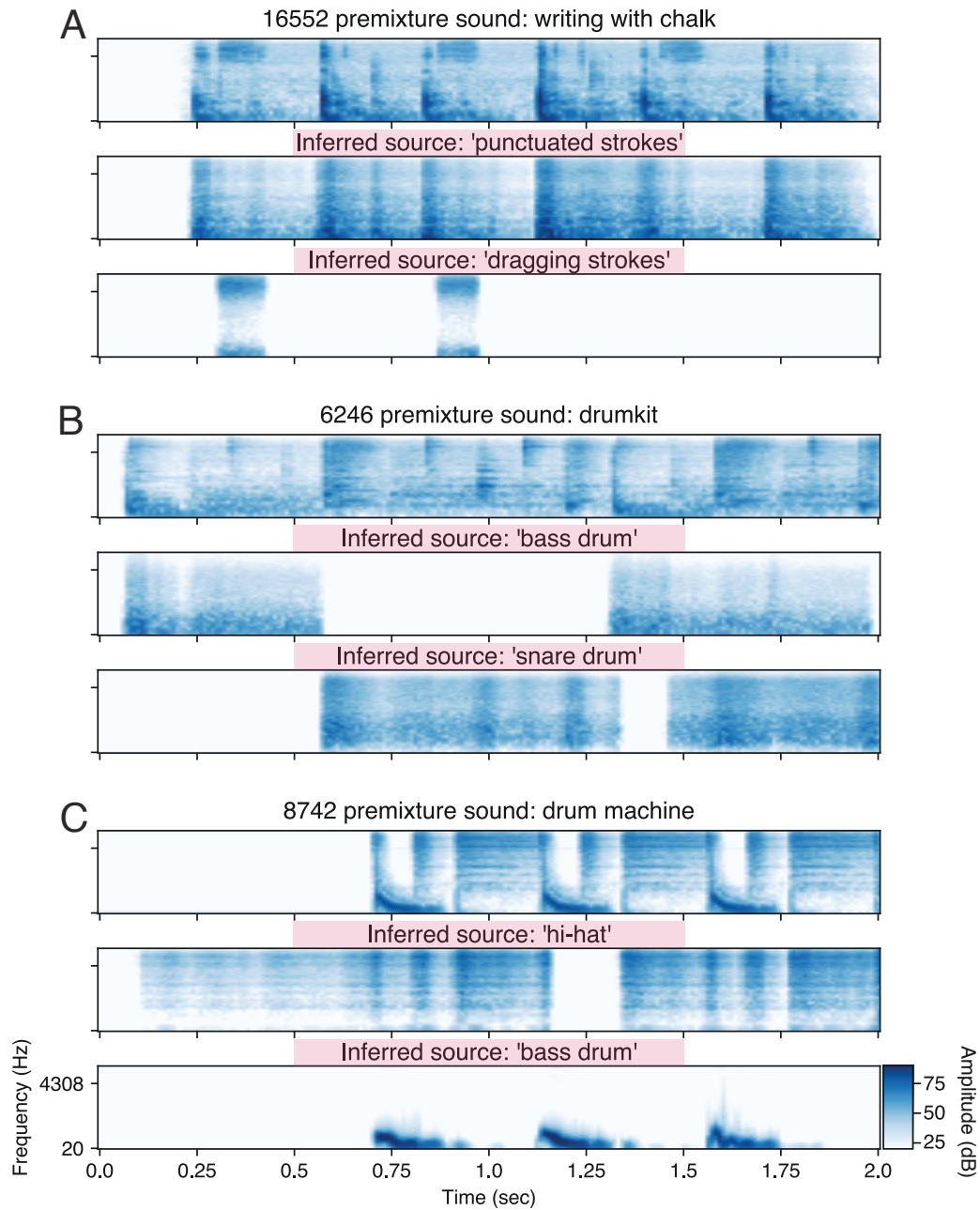


Figure 2-17: Oversegmentation deviations that occur when various sound-generating processes within a single premixture clip have some common causal factor, involving sequences.

A) An oversegmentation deviation for a premixture clip of writing on a chalk board, in which dragging the chalk across the board is punctuated by shorter strokes. The sustained contact sounds and punctuated contact sounds are explained by separate sources. B) A similar oversegmentation deviation occurs in a premixture clip of a drumkit, for which the model infers a noise source for the bass drum and a noise source for the snare. C) Another drum machine in which the hi-hat and bass are inferred separately, is also linked to an oversegmentation deviation. We suggest that percepts for these sound examples are akin to hierarchical grouping in vision, and such hierarchy cannot be captured by the generative model.

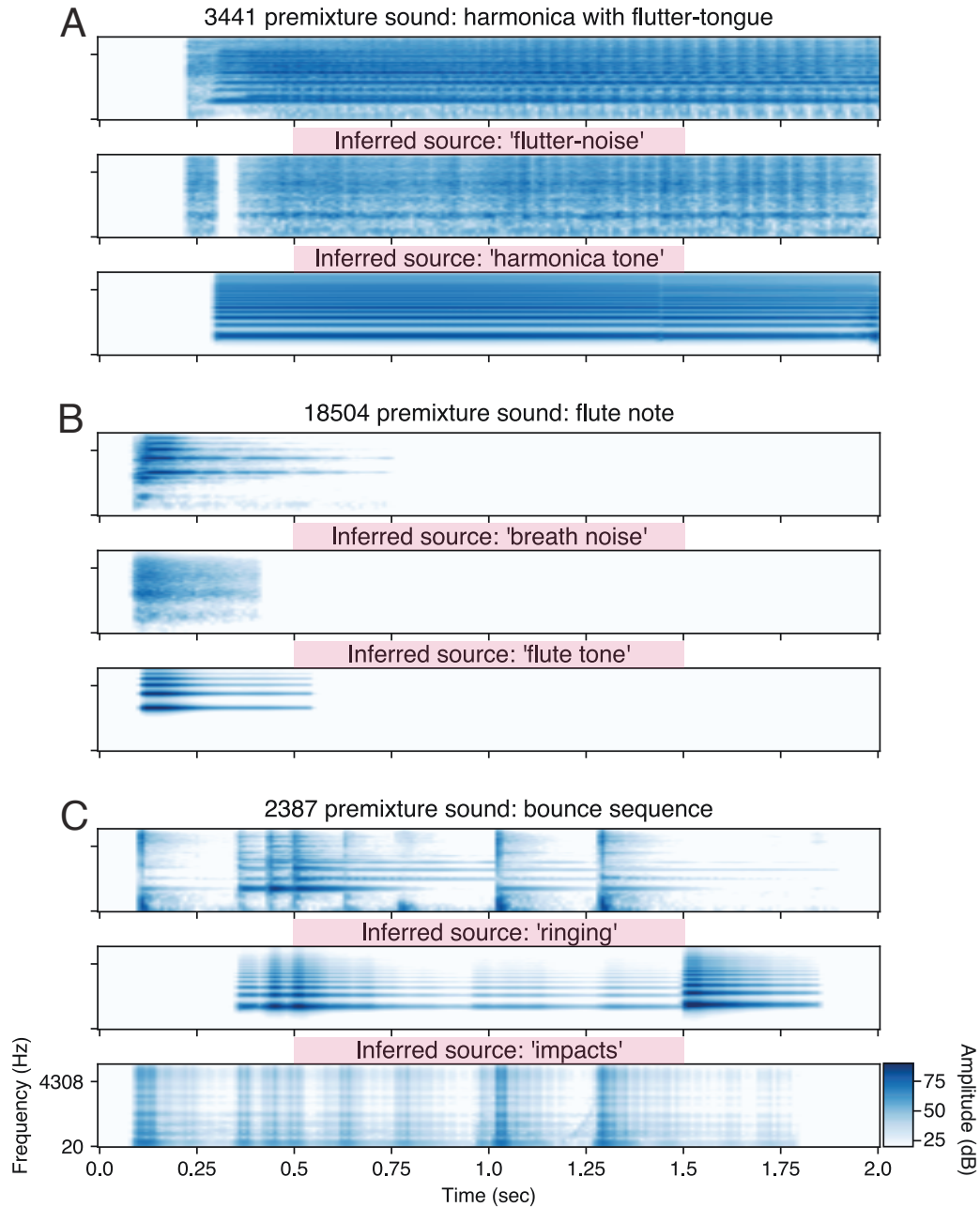


Figure 2-18: Oversegmentation deviations that occur when various sound-generating processes within a single premixture clip have some common causal factor, involving excitations that occur simultaneously.

A) Pre-mixture clip of a harmonic played with fluttersong, in which the player rolls their tongue while playing a pitched note. The model separates this into the noisy sound of the tongue-roll and the pitched sound of the harmonica tone. B) Pre-mixture clip has a flute note. The model separates the breath noise from the pitched tone. C) Pre-mixture clip is a series of bounces. On the second to fifth bounce, the object hits a resonant metal surface (you can see the modal resonances in the sound). The model separates the sequence into ringing sounds and the impact ‘thuds’. The event at the end of the ringing source (starting at 1.5-sec) corresponds to a sound in the mixture that is not in this pre-mixture clip (see Supplementary Figure 2-20). We again suggest that percepts for these sound examples are akin to hierarchical grouping in vision, and such hierarchy cannot be captured by the generative model.

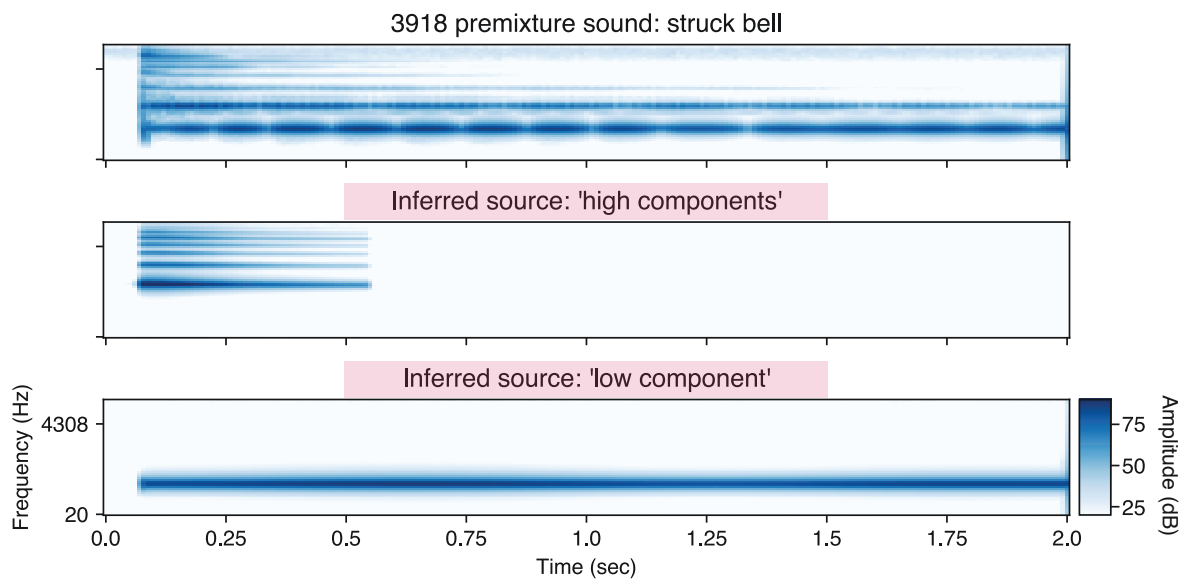


Figure 2-19: Impact sounds illustrate issues with model spectrum and amplitude (1). In this premixture impact sound of a struck bell, the model oversegments higher and lower frequency sound components which decay at different rates. Other issues for the model include the bell sound's inharmonicity and its transient.

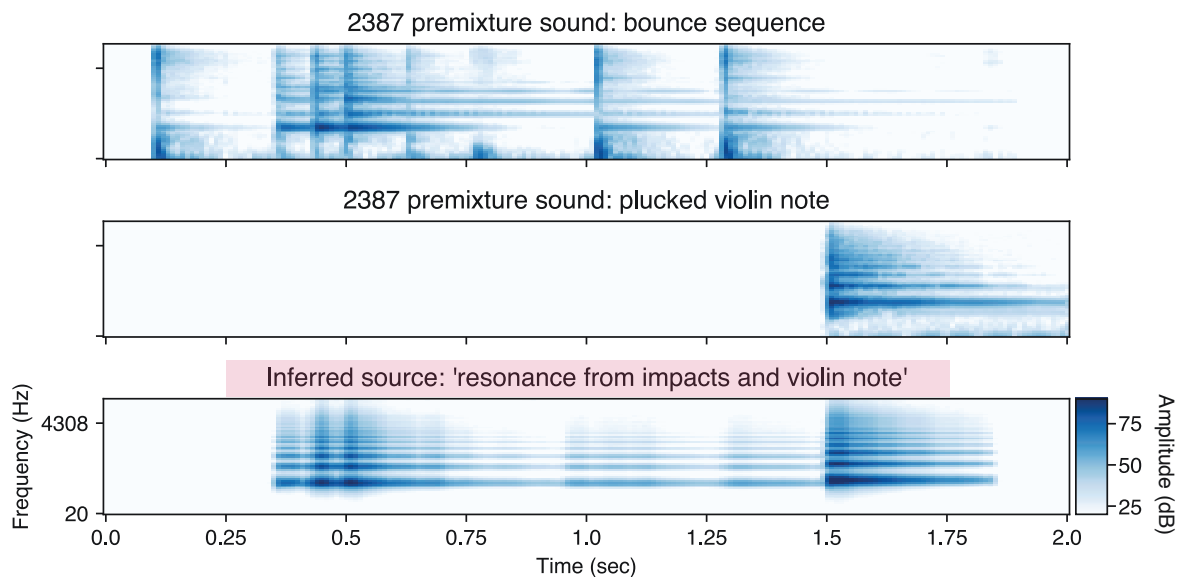


Figure 2-20: Impact sounds illustrate issues with model spectrum and amplitude (2). An overcombine deviation occurs when the model sequentially groups the resonances of an impacted metal object with a plucked string that occurs much later in the mixture. Since both sources have similar fundamental frequencies, spectral envelopes, and amplitudes, the model cannot separate these perceptually distinct sources.

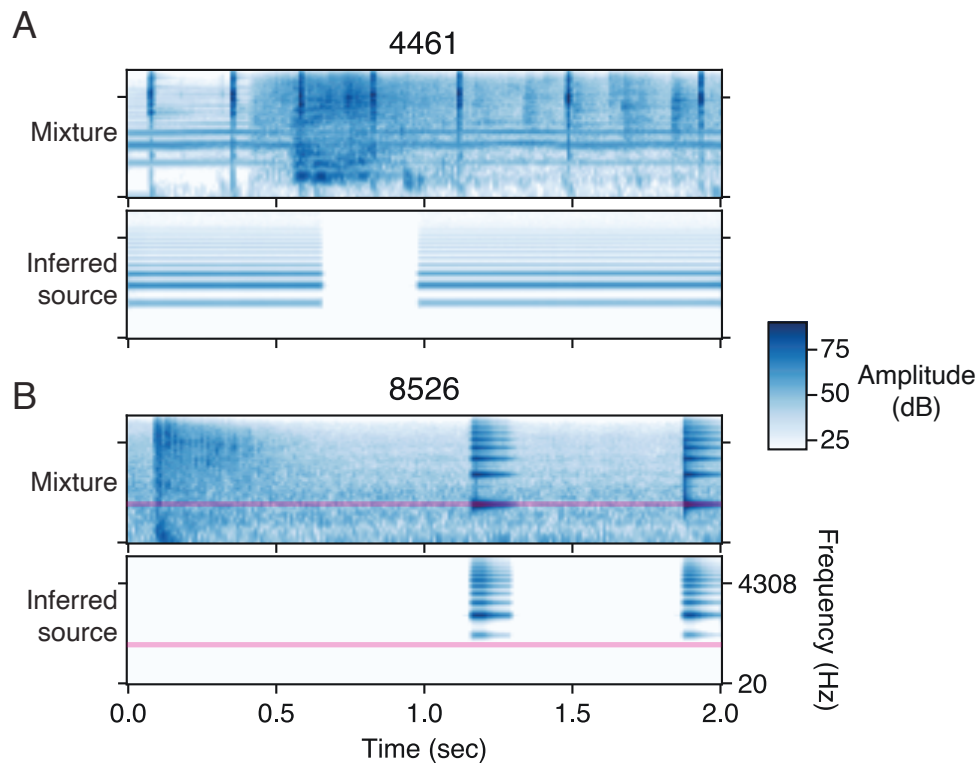


Figure 2-21: Inference errors due to the local nature of stochastic gradient descent. A) The harmonic tone is not inferred to continue across the masker, although it is perceived by human listeners to continue. Because evidence for the harmonic tone is masked, this is a local optima for gradient descent. B) The pink lines indicate the lowest frequency component of the complex tone in the mixture, which does not align with the lowest frequency component in the inferred source. This is a local optima because smoothly shifting the fundamental frequency of the inferred events downward will decrease the likelihood.

2.5 Methods

Code is available through my thesis webpage at <https://mcdermottlab.mit.edu/mcusi/thesis/>.

2.5.1 Model

Priors, $P(S)$.

For the scene sampling procedure that defines $P(S)$, see Algorithm 1. A full scene description S is composed of a number of sources n . Each source s_i has a sound type T_i , and is composed of the number of events it emits m_i , a set of source variables Θ_i , and a set of m_i event descriptions E_i .

$$S = (n, \{s_i\}_{i=1\dots n}), \quad \text{where} \quad s_i = (m_i, T_i, \Theta_i, E_i) \quad \text{and} \quad E_i = \{e_{ij}\}_{j=1\dots m_i} \quad (2.1)$$

The sources are independent, meaning $p(\{s_i\}_{i=1\dots n}) = \prod_{i=1}^n p(s_i)$. Each source specifies a distribution over events, which is dependent on the source variables and sound type. Thus, $p(s_i)$ further factorizes into a hierarchical model:

$$p(S) = p(n) \prod_i^n p(m_i, T_i, \Theta_i, E_i) \quad (2.2)$$

$$= p(n) \prod_i^n p(E_i, \Theta_i, m_i | T_i) p(T_i) \quad (2.3)$$

$$= p(n) \prod_i^n p(E_i | \Theta_i, m_i) p(\Theta_i | T_i) p(T_i) p(m_i) \quad (2.4)$$

The hierarchical factorization into event priors $p(E_i | \Theta_i, m_i)$ and source “hyperpriors” $p(\Theta_i | T_i)$ expresses that different sources will have different tendencies when emitting events. To generate scenes with any number of sources and events, $p(n)$ is a Poisson distribution and $p(m_i)$ is a Geometric distribution. The Poisson distribution $p(n)$ specifies the probability of some number of sources occurring during a known time

interval, where the timings of the first event of each source are considered to be independent. The Geometric distribution $p(m_i)$ describes the probability of how many events are emitted in total before the source ceases. The sound types are equally likely, with $p(T_i)$ as a uniform categorical distribution.

Now we examine the event priors in more detail. Each event e_{ij} consists of a set of temporal variables $\boldsymbol{\tau}_{ij}$ which define its onset and offset, as well as time-varying excitation variables (amplitude $\mathbf{a}_{ij}(t)$ and, if periodic, fundamental frequency $\mathbf{f}_{ij}(t)$) which depend on the timing of events $\boldsymbol{\tau}_i$.

$$e_{ij} = (\boldsymbol{\tau}_{ij}, \mathbf{a}_{ij}(t), \mathbf{f}_{ij}(t)) \text{ for periodic sources,} \quad e_{ij} = (\boldsymbol{\tau}_{ij}, \mathbf{a}_{ij}(t)) \text{ for noisy sources} \quad (2.5)$$

Unlike sources, events are not independent from each other. Rather, events are sampled sequentially, such that each event depends on the events before it. For example, the full event prior for a periodic source is:

$$p(E_i | \Theta_i, m_i) = p(\boldsymbol{\tau}_i | \Theta_i, m_i) p(\mathbf{a}_i(t) | \Theta_i, m_i) p(\mathbf{f}_i(t) | \Theta_i, m_i) \quad (2.6)$$

$$= p(\{\boldsymbol{\tau}_{ij}\}_{j=1:m_i} | \Theta_i, m_i) p(\{\mathbf{a}_{ij}(t)\}_{j=1:m_i} | \Theta_i, m_i, \boldsymbol{\tau}_i) p(\{\mathbf{f}_{ij}(t)\}_{j=1:m_i} | \Theta_i, m_i, \boldsymbol{\tau}_i) \quad (2.7)$$

$$(2.8)$$

with individual events conditioned on the preceding events:

$$p(\{\boldsymbol{\tau}_{ij}\}_{j=1:m_i} | \Theta_i, m_i) = \prod_{j=1}^{m_i} p(\boldsymbol{\tau}_{ij} | \Theta_i, m_i, \boldsymbol{\tau}_{i,j-1}) \quad (2.9)$$

$$p(\{\mathbf{a}_{ij}(t)\}_{j=1:m_i} | \Theta_i, m_i, \boldsymbol{\tau}_i) = \prod_{j=1}^{m_i} p(\mathbf{a}_{ij}(t) | \Theta_i, m_i, \{\mathbf{a}_{ik}(t)\}_{k=1:j-1}, \{\boldsymbol{\tau}_{ik}\}_{k=1:j}) \quad (2.10)$$

$$p(\{\mathbf{f}_{ij}(t)\}_{j=1:m_i} | \Theta_i, m_i, \boldsymbol{\tau}_i) = \prod_{j=1}^{m_i} p(\mathbf{f}_{ij}(t) | \Theta_i, m_i, \{\mathbf{f}_{ik}(t)\}_{k=1:j-1}, \{\boldsymbol{\tau}_{ik}\}_{k=1:j}) \quad (2.11)$$

To constrain the events to be non-overlapping in time, we reparametrize an event’s onset τ_{ij}^{on} and offset τ_{ij}^{off} as the silent “rest” interval τ_{ij}^R preceding e_{ij} and its active interval τ_{ij}^D . If the durations τ_{ij}^R and τ_{ij}^D are non-negative then the events will not overlap. Thus we model the rest and duration as:

$$\boldsymbol{\tau}_{ij} = (\tau_{ij}^R, \tau_{ij}^D, \tau_{ij}^{\text{on}}, \tau_{ij}^{\text{off}}) \quad (2.12)$$

$$\tau_{ij}^R \sim \text{LogNormal}(\mu_i^R, \lambda_i^{R-1}) \quad (2.13)$$

$$\tau_{ij}^D \sim \text{LogNormal}(\mu_i^D, \lambda_i^{D-1}) \quad (2.14)$$

$$\tau_{ij}^{\text{on}} = \tau_{ij-1}^{\text{off}} + \tau_j^R \quad (2.15)$$

$$\tau_{ij}^{\text{off}} = \tau_{ij}^{\text{on}} + \tau_{ij}^D \quad (2.16)$$

The only exception is the first onset which is sampled uniformly from the duration of the scene.

The sampled event timings will reflect source regularities as captured by the parameters of the LogNormal priors. These parameters are source-level variables, that is $(\mu_i^R, \lambda_i^R, \mu_i^D, \lambda_i^D) \in \Theta_i$. The hyperprior of each pair is a normal-gamma conjugate prior which is shared across all sound types. These source hyperpriors allow the model to account for a wide range of temporal regularities, as occur in natural sounds: for example, a source could tend to have long events (μ_i^D is high), events occurring in rapid succession (μ_i^R is low), or a wide variety of event durations (λ_i^D is low).

To instantiate source regularities for dynamic event excitations, we induce temporal correlations both within and between events. Specifically, we model the time-varying excitation variables using one-dimensional Gaussian Process (GP) priors (Rasmussen & Williams, 2005). GPs are distributions over functions $g(x)$, for which any finite set of function values $\{g(x_1), \dots, g(x_n)\}$ has a multivariate normal distribution. GPs are thus characterized by their mean function $\mu(x)$ and a kernel function $\kappa(x, x')$ that outputs the prior covariance between $g(x)$ and $g(x')$. Given these functions and a vector of timepoints \mathbf{t} which lie within the events (i.e., $\tau_{ij}^{\text{on}} < t < \tau_{ij}^{\text{off}}$), we can sample the trajectory of the time-varying excitation variables $\mathbf{a}(\mathbf{t})$ and $\mathbf{f}(\mathbf{t})$. In

particular, the time-varying excitation trajectory of the source's first event is sampled from a GP with a constant mean function (e.g., for amplitude $\mu_i^a(t) = \mu_i^a$) and non-stationary kernel function $\kappa_i(\tau_i)$; we call this \mathcal{GP}_1 . Taking the excitation amplitude $\mathbf{a}_{i,1}(\mathbf{t})$ of the first event of source i ($e_{i,1}$) as an example:

$$\mathbf{a}_{i,1} \sim \mathcal{GP}_1(\mu_i^a, \kappa_i^a(\boldsymbol{\tau}_i)) \quad (2.17)$$

The following events are sampled from a GP obtained by conditioning \mathcal{GP}_1 on the actual sampled excitation trajectories of the previous events. Therefore, the prior distribution over the excitation amplitude $\mathbf{a}_{ij}(\mathbf{t})$ of event e_{ij} is:

$$\mathbf{a}_{ij} \sim \mathcal{GP}_{j|1:j-1}(\mu_i^a, \kappa_i^a(\boldsymbol{\tau}_i) \mid \{\mathbf{a}_{ik}\}_{k=1:j-1}) \quad (2.18)$$

The parameters of the mean and kernel functions are source variables that define the regularity reflected in the source's events. The mean constant μ_i determines the central tendency of the trajectory (e.g. quiet vs. loud). The kernel κ_i determines the shape of the trajectories which are likely under the GP prior. For time-varying excitation trajectories, we implement a non-stationary kernel as follows:

$$\kappa_a(t_1, t_2; \boldsymbol{\tau}_i) = \text{SE}(t_1, t_2; \sigma_a, \ell_a) + \sum_{j=1}^m \text{NS}(t_1, t_2; \beta_a; \boldsymbol{\tau}_{ij}) \quad (2.19)$$

$$\text{SE}(t_1, t_2; \sigma_a, \ell_a) = \sigma_a^2 \exp\left(-\frac{(t_1 - t_2)^2}{\ell_a^2}\right) \quad (2.20)$$

$$\text{NS}_{\boldsymbol{\tau}_{ij}}(t_1, t_2; \beta_a) = \begin{cases} \beta_a, & \text{if } \tau_{ij}^{\text{on}} < t_1, t_2 < \tau_{ij}^{\text{off}} \\ 0, & \text{otherwise} \end{cases} \quad (2.21)$$

The first term of κ is a standard squared exponential kernel (SE) that instantiates a prior favouring smoothly-varying excitation trajectory both within and across events, which is a reasonable assumption for natural sounds. The second term is a non-

stationary kernel (NS) which specifies higher covariance for trajectory values that occur within the same event, compared to across events. This non-stationary kernel can express the case where a single source slightly modifies its sound-generating process between events. In natural sounds, this occurs in a variety of ways: to provide a few examples, a dog panting is quieter on the in-breath than the out-breath, and a flute can discretely switch pitches between notes with a button-press.

The filter shape, which is a function of frequency, has a one-dimensional stationary GP prior. Like the excitation trajectories, the prior over the spectra of harmonic sources uses an SE kernel, exhibiting a tendency to vary smoothly. Because of the prevalence of bandpass noise with sharp cutoffs, we used a Ornstein-Uhlenbeck (OU) kernel for noise source spectra. Like the SE kernel, the OU kernel exhibits a tendency to revert to the mean, but in contrast it does not tend to be smooth. Regardless of sound type, the sampled filter \mathbf{H} is shared across all events.

$$\mathbf{H} \sim \mathcal{GP}(\mu_H, \kappa_H) \quad (2.22)$$

$$\kappa_H(\omega_1, \omega_2) = \begin{cases} \text{SE}(\sigma_H, \ell_H), & \text{if harmonic} \\ \text{OU}(\sigma_H, \ell_H), & \text{if noise} \end{cases} \quad (2.23)$$

$$\text{OU}(\omega_1, \omega_2; \sigma_H, \ell_H) = \sigma_H^2 \exp\left(-\frac{|\omega_1 - \omega_2|}{\ell_H}\right) \quad (2.24)$$

The sampled trajectories and filters will reflect source regularities, as captured by the parameters of the GPs. The mean μ as well as the kernel variables σ and ℓ of each GP are source-level variables. The means are sampled from an appropriately scaled uniform distribution. The inverse softplus of σ and ℓ are distributed normally (to maintain positivity and numerical stability). These source priors allow the model to account for a wide range of regularities in the excitation trajectories and the filters, as observed in natural sounds. For example a source could have a flat spectrum (low σ_H) or a spectrum with widely spaced peaks (high σ_H and high ℓ_H).

Prior parameters

The prior parameters used for the classic ASA phenomena are listed in Tables 2.2-2.3.

Discrete priors. There are two discrete priors, the prior over the number of sources and the prior over the number of events in a source. In practice, we found the specific settings of these parameters had little effect on the results.

Temporal hyperpriors. The parameters of the normal-gamma hyperpriors for the duration and rest variables were chosen to be only weakly informative, covering a large range (durations and rests used to create the classic ASA sounds range from 40ms to 2s). We restricted the rest and duration hyperpriors to be the same as each other, and the same across all sound types. In practice, we found the specific settings of these parameters had little effect on the results.

Gaussian process hyperpriors. Since the Gaussian process kernel hyperpriors encode more complex constraints, we fit them to natural sounds. To fit a given hyperprior, we simultaneously inferred scene descriptions for a set of natural sounds. Each scene description was constrained to use a single source of a specified sound type. The variable hyperprior parameters were shared across all scenes, instantiating a learnable distribution over the corresponding source-level variables. The parameters were inferred by variational inference, along with the source- and event-level variables, to maximize the marginal likelihood of the full set of sounds. The inferred parameters were then fixed as constants for the experiments with classic ASA illusions and natural sounds.

We chose publicly available datasets of recorded sounds for which the dominant sound type was obvious and it was possible to condition on the secondary, event-level variables (onset, offset and fundamental frequency where applicable). Conditioning on known secondary event variables facilitated stable inference of the hyperprior parameters, since a single source may be inadequate to explain real world audio (e.g., due to the presence of background noise in addition to the focal source). For noises, we used a subset of the sound texture dataset used by (McDermott & Simoncelli, 2011) which were obviously aperiodic. For periodic sources, we used three different datasets that

covered a variety of sounds: speech, music, and bioacoustics. For speech, we used the Pitch Tracking Database from Graz University of Technology (PTDB-TUG), which provides recorded audio of speakers saying English sentences and pitch trajectories extracted from corresponding laryngograph signals (Pirker et al., 2011). For music, we used the University of Rochester Multi-Modal Music Performance (URMP) dataset, which contains recorded audio of various instruments playing classical music along with corresponding MIDI and pitch tracks (Li et al., 2018). For bioacoustics, we used Synth Birds Database (SynthBirdsDB) which contains recorded audio from a variety of bird species, with either harmonic or pure tone vocalizations (O’Reilly & Harte, 2017). This dataset was compiled to develop pitch-tracking for bird vocalizations, so we used a subset of the data for which accurate pitch tracks were provided.

Specifically, we fit an inverse softplus normal hyperprior each for the variance σ and lengthscale ℓ kernel variables of a GP prior. For time-varying excitation trajectories with an additional non-stationary component in the kernel, we also fit a single value of the non-stationary parameter β , shared across all sounds. Due to the computation required for fitting many scenes simultaneously, we randomly selected a subset of two-second clips from the appropriate dataset(s) to fit each set of hyperpriors (approximately one minute total). This small amount of data is sufficient, as we fit only 2-3 variables per hyperprior.

We sought to maximize the variability in the set of sounds for each hyperprior, while ensuring that the model was appropriate for the included sounds. The amplitude and spectrum hyperpriors for the noise sound type were simultaneously fit to the sound textures dataset. The amplitude and spectrum hyperpriors for the harmonic sound type were simultaneously fit to URMP and SynthBirdsDB. We omitted PTDB-TUG here because the formants of speech could not be well-modeled with the model’s constant spectrum constraint. The amplitude hyperprior for the whistle sound type was fit to clips from SynthBirdsDB containing pure tone bird vocalizations, omitting harmonic speech and music. Last, the fundamental frequency hyperprior was fit on the speech, music, and bioacoustics dataset, and shared between the whistle and harmonic sources. We could use all three datasets here because we could directly fit

Table 2.2: Discrete priors and temporal normal-gamma hyperprior. The temporal hyperprior is used for both the rest and duration latent variables.

Variable	Class	Distribution
n	Discrete	Poisson(rate = 1 sources/sec)
m	Discrete	Geometric($p = 0.5$)
(μ, λ)	Temporal	Normal-Gamma($\mu_0 = -1.0, \lambda_0 = 0.5, \alpha_0 = 2.5, \beta_0 = 1.0$)

a GP to the pitch tracks rather than to the recorded audio. It turns out that using uniform hyperpriors for these variables instead of fitting them has a relatively small effect on the model’s similarity to human perception (see Results: Model alternatives).

For all hyperpriors, we manually set the bounds of the uniform distributions over the mean (the spectrum means were fixed at zero). The uniform distributions were bounded to cover a natural range of values (e.g., fundamental frequency could span the range of frequencies audible to humans and below the Nyquist limit; amplitude was bounded below by the quietest audible amplitude) and to capture the range of means inferred during hyperprior fitting.

We conducted the everyday sound experiments first (section 2.2.6), using hyperpriors which included parameters that were not fit to recorded audio. The hyperpriors were modified so that all were fit to recorded audio (as just described above) before conducting the classic ASA experiments (reported in sections 2.2.4 and 2.2.5). The use of different hyperpriors for these two sets of experiments was unintentional. Time constraints prevented us from re-running the everyday sounds experiments with the hyperpriors fit to recorded audio, but we plan to conduct all experiments again with shared hyperpriors.

Likelihood, $P(X|S)$

To compute the likelihood, the sampled scene description S must be rendered into the resulting sound X . The scene description first was sampled according to Algorithm 1. The amplitude and frequency trajectories were sampled at 10 ms intervals for the classic auditory scene analysis stimuli and at 20 ms for the everyday sounds for increased memory efficiency (\mathbf{a} in dB, \mathbf{f} in ERB). The filter shape was sampled at 0.3

Table 2.3: Gaussian process hyperprior parameter values. The mean μ is uniformly distributed (with units listed under GP type). The inverse softplus of the variance σ and lengthscale ℓ are each normally distributed. Samples of σ and ℓ are bounded to reasonable values. σ is in the GP-units, while ℓ is in seconds for excitation variables and in ERB for spectrum. β is a positive constant that implements the non-stationary kernel. As is typical for GPs, ϵ^2 is added to the diagonal of the covariance matrix for numerical stability. The last three columns indicate the first, second, and third quartile of sampled σ and ℓ distributions based on 5000 samples. *Note: We manually set the value of β to 1.0 because the scenes in the textures dataset only had one event.

Source type	GP type	Kernel	Datasets	Distribution parameters	Bounds	Q_1	Q_2	Q_3
W, H	F0 (ERB)	SE +NS	PTDB-TUG, URMP, SynthBirdsDB	μ	(3, 33)			
				σ	(3.6, 3.0)	(0.1, 33)	2.05	3.86
				ℓ	(−0.76, 4.0)	(0.01, 20)	0.13	0.69
				β	0.92			5.83
				ϵ	0.1			2.54
W	Amp. (dB)	SE +NS	SynthBirdsDB (pure tone only)	μ	(0, 120)			
				σ	(1.2, 4.4)	(1, 50)	2.21	3.85
				ℓ	(3.6, 0.4)	(0.01, 10)	3.10	3.69
				β	0.46			5.95
				ϵ	0.1			4.29
H	Amp. (dB)	SE +NS	URMP, SynthBirdsDB (harmonic only)	μ	(0, 120)			
				σ	(7.2, 1.9)	(0.1, 50)	5.87	7.25
				ℓ	(−1.9, 0.7)	(0.01, 10)	0.07	0.14
				β	5.3			8.61
				ϵ	0.5			0.27
H	Spec. (dB)	SE	URMP, SynthBirdsDB (harmonic only)	μ	(12., 5.0)	(0.1, 50)	9.42	12.66
				σ	(3.6, 8.5)	(0.01, 33)	1.28	5.34
				ℓ				16.09
				ϵ	0.5			10.43
				ϵ	0.5			
N	Amp. (dB)	SE +NS	Textures	μ	(0, 80)			
				σ	(1.8, 0.9)	(0.1, 50)	1.28	1.94
				ℓ	(−2.1, 2.3)	(0.01, 10)	0.05	0.18
				β	1.0*			2.70
				ϵ	0.5			0.57
N	Spec. (dB/Hz)	OU	Textures	μ	(7.1, 1.0)	(0.1, 50)	6.23	7.10
				σ	(12., −0.02)	(0.1, 33)	12.03	12.48
				ℓ				7.95
				ϵ	0.5			12.94
				ϵ	0.5			

ERB intervals (\mathbf{H} in dB/Hz for noises and dB for harmonics). To render the sounds specified by these sampled latent variables, we generated the sound corresponding to each sampled event in each source. Then, we concatenated the events and silent intervals to construct each source sound. Finally, we summed the source sounds to produce the final sound mixture.

Events of all sound types were differentiably rendered. Each event type was generated by combining an initial excitation with spectral and/or temporal amplitude modulation (filter). For whistle events, we generated a pure tone that was frequency-modulated according to \mathbf{f} . To amplitude modulate the tone, it was windowed with half-overlapping cosine windows and then each window was scaled by the corresponding amplitude in \mathbf{a} . For noise events, we generated a pink noise sample. This noise sample was frozen with a fixed random seed to enable differentiability. The noise was then amplitude modulated as above. Finally, a log-spaced, cosine-shaped, half-overlapping filterbank scaled by \mathbf{H} was multiplicatively applied in the frequency domain. For harmonic events, we generated a set of 200 pure tones at the harmonic frequencies specified by \mathbf{f} . Any portions of the pure tones that exceeded the Nyquist limit were set to zero to prevent aliasing. We scaled the amplitudes of each tone so that they were pink with respect to the fundamental, and then summed them to produce a complex tone. The sampled spectrum \mathbf{H} was shifted in frequency such that its first channel aligned with fundamental frequency \mathbf{f} at all timepoints. Then, spectral and amplitude modulations were applied as above. Sigmoid on-ramps and off-ramps were applied to each event (duration=18 ms), with the sampled onsets and offsets corresponding to the maximum of the ramps. Sigmoid ramps were chosen for their differentiability. We rendered all sounds at 20 kHz.

We used a differentiable implementation of gammatone-based spectrograms for our likelihood representation (Ellis, 2009). A conventional spectrogram is calculated and then the frequencies are combined into gammatone-like channels. This particular human-like time-frequency representation was chosen because of the speed of computation during stochastic variational inference, which requires thousands of iterations. We used a window size of 25 ms, hop size of 10 ms, 64 filters with a half-ERB width,

and a lower threshold of 20 dB. The sampled scene gammatonegram was compared to the observed gammatonegram under an isotropic Gaussian noise model. The standard deviation on the noise model was fixed to $\sigma = 10$ across all inferences.

Algorithm 1 Scene sampling procedure

procedure SAMPLESCENE

$n \sim \text{Poisson}(\lambda_n)$
for $i = 1, \dots, n$ **do**
 $\theta_i^{\text{type}} \sim \text{Categorical}(\text{whistle, noise, harmonic})$
 $s_i = \text{SampleSource}(\theta_i^{\text{type}})$

procedure SAMPLESOURCE(θ^{type})

$m \sim \text{Geometric}(\lambda_m)$ \triangleright Number of events
 $\mu^{\mu_P}, \sigma^{\mu_P}, \alpha^{\lambda_P}, \dots = \text{hyperparams}(\theta^{\text{type}})$ \triangleright Hyperprior parameters for sources of type θ^{type}

$\mu_P \sim \text{Normal}(\mu^{\mu_P}, \sigma^{\mu_P})$ \triangleright Variables for priors over event timings
 $\lambda_P \sim \text{Gamma}(\alpha^{\lambda_P}, \beta^{\lambda_P})$
 $\mu_D \sim \text{Normal}(\mu^{\mu_D}, \sigma^{\mu_D})$
 $\lambda_D \sim \text{Gamma}(\alpha^{\lambda_D}, \beta^{\lambda_D})$

$\mu_a \sim \text{Uniform}(a^{\mu_a}, b^{\mu_a})$ \triangleright Amplitude GP parameters
 $\sigma_a \sim \text{invSoftplusNormal}(\mu^{\sigma_a}, \sigma^{\sigma_a})$
 $\ell_a \sim \text{invSoftplusNormal}(\mu^{\ell_a}, \sigma^{\ell_a})$
 $\beta_a = \beta_{T,a}$ \triangleright Constant for sound type
 $\kappa_a(t_1, t_2; \tau) = \text{SE}(\sigma_a, l_a) + \sum_{j=1}^m \text{NS}(\beta_a; \tau_j)$ \triangleright Nonstationary kernel using timings τ

if $\theta^{\text{type}} \in \{\text{whistle, harmonic}\}$ **then**
 $\mu_f \sim \text{Uniform}(a_f, b_f)$ \triangleright Parameters for f_0 GP
 $\sigma_f \sim \text{invSoftplusNormal}(\mu^{\sigma_f}, \sigma^{\sigma_f})$
 $\ell_f \sim \text{invSoftplusNormal}(\mu^{\ell_f}, \sigma^{\ell_f})$
 $\beta_f = \beta_{T,f}$ \triangleright Constant for sound type
 $\kappa_f(t_1, t_2; \tau) = \text{SE}(\sigma_f, l_f) + \sum_{j=1}^m \text{NS}(\beta_f; \tau_j)$ \triangleright Nonstationary kernel using timings τ

if $\theta^{\text{type}} \in \{\text{noise, harmonic}\}$ **then**
 $\mu_H \sim \text{Uniform}(a^{\mu_H}, b^{\mu_H})$ \triangleright Parameters for filter GP
 $\sigma_H \sim \text{invSoftplusNormal}(\mu^{\sigma_H}, \sigma^{\sigma_H})$
 $\ell_H \sim \text{invSoftplusNormal}(\mu^{\ell_H}, \sigma^{\ell_H})$
 $\kappa_H(\omega_1, \omega_2) = (\theta^{\text{type}} == \text{harmonic}) ? \text{SE}(\sigma_H, \ell_H) : \text{OU}(\sigma_H, \ell_H)$
 $\mathbf{H} \sim \mathcal{GP}(\mu_H, \kappa_H)$ \triangleright Filter (common to all events)

\triangleright Sample event variables

$\text{onset}_1 \sim \text{Uniform}(0, \tau_{\text{scene}})$

for $j = 1, \dots, m$ **do**

$\text{offset}_j \sim \text{LogNormal}(\mu_D, \lambda_D^{-1}) + \text{onset}_j$
 $\tau_j = (\text{onset}_j, \text{offset}_j)$

\triangleright Sample amplitude and f_0 from GPs conditioned on previous events

$\mathbf{a}_j \sim \mathcal{GP}_{j|1:j-1}(\mu_a, \kappa_a(\tau) \mid \mathbf{a}_{1:j-1})$

if $\theta^{\text{type}} \in \{\text{whistle, harmonic}\}$ **then** $\mathbf{f}_j \sim \mathcal{GP}_{j|1:j-1}(\mu_f, \kappa_f(\tau) \mid \mathbf{f}_{1:j-1})$

if $j \neq m$ **then** $\text{onset}_{j+1} \sim \text{LogNormal}(\mu_P, \lambda_P^{-1}) + \text{offset}_j$

2.5.2 Inference

We used two modes of inference. Both enumerative and sequential inference involve optimizing and comparing hypotheses, but they differ in how they determine the hypotheses in the first place. In enumerative inference, which is more appropriate for the constrained setup of psychophysics experiments, we directly assess the posterior probability for alternative experimenter-defined hypotheses. In sequential inference, which is more appropriate to model unconstrained listening with illusions and everyday sounds, we build up hypotheses sequentially using candidate events from an amortized inference network. Sequential inference can also be used in psychophysics experiments (as utilized for the network comparisons). As demonstrated in the network comparisons, sequential inference results in similar performance to enumerative inference.

Ultimately, the posterior probabilities computed by the generative model should adjudicate between the different hypotheses. We do not want to prematurely exclude hypotheses based on our inference algorithms, so we follow the principle of least commitment: keeping as many hypotheses as possible while maintaining computational tractability.

We first explain the process of hypothesis optimization and comparison which is common to both modes of inference. Then we explain sequential and enumerative inference in turn.

2.5.3 Hypothesis optimization and comparison

A good hypothesis H will correspond to a region containing a “mode” of the full posterior. A hypothesis is thus characterized by a specific setting of the structural variables (e.g. the number of sources and events), as well as an approximate setting of continuous variables (e.g. source variables, onset timings). This corresponds to the organization of specific events into sources, while only approximately specifying the low-level features of the events. For instance, a particular hypothesis could specify a single source with one high and one low whistle, but there may be some variance

around the exact frequencies of the whistle. To compute the posterior probability of this hypothesis requires marginalizing over the continuous variables.

Many applications of Bayesian inference to perception opt for point-estimates of the posterior mode (i.e., MAP inference), which sometimes bypasses this challenging integral. However, this integral is unavoidable in the case of perceptual organization, because scenes with potentially different dimensionalities cannot be compared by their posterior densities (which have different units). Instead, we need to integrate each hypothesis into a posterior probability mass so that we can compare across dimensionalities.

Formally, each hypothesis H_i corresponds to a region of scene space ($H_i : S \in \mathcal{S}_i$), and the hypotheses are compared based upon their posterior odds ratio (i.e., relative posterior mass contained within the regions). For two hypotheses and an observed sound D , the posterior odds is:

$$\frac{p(H_i|D)}{p(H_j|D)} = \frac{p(H_i, D)}{p(H_j, D)} \quad (2.25)$$

Since H_i contains many specific scenes $S \in \mathcal{S}_i$, computing the marginal probability $p(H_i, D)$ requires integrating over \mathcal{S}_i .

$$p(H_i, D) = \int_{S \in \mathcal{S}_i} p(S)p(D|S) \mathrm{d}S \quad (2.26)$$

To approximate this integral, we use importance sampling. In importance sampling, we use a guide distribution q_i corresponding to hypothesis H_i to take samples.

$$p(H_i, D) = \mathbb{E}_{q_i} \frac{p(S)}{q_i(S)} p(D|S) \quad \text{where } \mathcal{S}_i \text{ is support of } q_i \quad (2.27)$$

$$\approx \frac{1}{K} \sum_{k=1}^K \frac{p(S_k)}{q_i(S_k)} p(D|S_k) \quad \text{where } S_k \stackrel{\text{iid}}{\sim} q_i \quad (2.28)$$

The success of this method is determined by how closely the guide distribution q_i approximates a mode of the posterior. In order to derive a good guide distribution, we start with an initial guide distribution which is either 1) based on the amortized neural network output in sequential inference or 2) based on experimenter-defined hypotheses in enumerative inference. In both cases, this initial guide distribution is then refined using automatic differentiation variational inference. The form of the guide distribution is mean-field, except for the vector variables (e.g. time-varying amplitude) which have a Gaussian Process prior. For these vector variables, we use a variational inducing point framework (Hensman et al., 2015). Each iteration of gradient descent, we optimize q_i with respect to the standard variational objective,

$$\mathcal{L}_K(H_i, D) = \mathbb{E}_{q_i} \log \left(\frac{p(S)}{q_i(S)} p(D|S) \right) \quad (2.29)$$

This constitutes a lower bound on the log marginal probability, $\mathcal{L}_K(H_i, D) \leq \log p(H_i, D)$. We used Adam to implement stochastic gradient descent. Each iteration of gradient descent used a batch of 10 samples from q . The number of iterations varied in sequential and enumerative inference, and are specified in the next section. The learning rates for the different latent variables were scaled so that the gradient steps taken in any direction were approximately the same size. We also used a scheduler to automatically decrease the learning rate if the variational lower bound plateaued.

2.5.4 Sequential inference

We provide examples of sequential inference for two different sound mixtures in Supplementary Figures 3 and 4.

Events proposal

A segmentation network proposes candidate events from the observed sound for use in later stages of sequential inference. Event proposals are depicted in Supplementary Figures 2-11A and 2-12A, where they are illustrated as a red mask overlaid on the sound mixture cochleagram. Although not directly depicted, a candidate event

also includes its estimated source type and event-level latent variables (onset, offset, amplitude, f0, spectrum). In later stages, these event-level variables are combined into full scene hypotheses (shown in Supplementary Figures 2-11B and 2-12B/C).

Architecture. The segmentation network is based on a publicly available image segmentation network, *Detectron2* (Wu et al., 2019). We use the standard *Detectron2* Generalized R-CNN/FPN architecture, but modified the training and test procedures to adapt *Detectron2* for use with sounds, as explained below.

Training dataset: We sample a dataset from the generative model to derive input/output training pairs. The basis for the input is a scene cochleagram, C_S . The outputs include event-level latent variables $\{e_{ij}\}$, a set of cochleagrams rendered from each event in the scene $\{C_{e_{ij}}\}$, a set of binary masks of the cochleagram for each event $\{M_e\}$. Because data generation for training the segmentation network does not need to occur iteratively, we used a more computationally expensive and detailed cochlear model to compute C_S and $\{C_{e_{ij}}\}$ Feather et al., 2022. Rather than sampling from the model with hyperpriors as defined by the natural sound datasets, we use uniform priors over all hyperprior parameters to increase data diversity and thereby improve transfer (called domain randomization, Billot et al., 2021; Engel, Swavely, et al., 2020; Tobin et al., 2017). The input to our network is composed of three channels: 1) C_S scaled to image values 0-255, 2) a binary channel indicating where C_S has no energy above threshold, and 3) a binary channel full of ones. The third channel enabled us to tell the network about the edges of the cochleagram. For sounds, the top and bottom edges of the cochleagram indicate the ends of the frequency range and they are not arbitrary as they are in images. Without this channel, we found that the network would assign events at these edges of the cochleagram, presumably due to spurious features created by automatic padding.

Training objective. Using these three channels, the network is then trained to recover the set of binary event masks $\{M_e\}$ and to recover sound type and the event-level latent variables $\{e_{ij}\}$ with a custom objective. The objective combines *Detectron2*'s standard composite segmentation loss, reconstruction loss of the event cochleagrams $\{C_{e_{ij}}\}$, and prediction error for the event-level latent variables $\{e_{ij}\}$.

Detectron2 also gives a confidence score for each output event proposal. This network confidence score (or associated rank) is listed above each event mask in Supplementary Figure 2-11A.

We conducted the everyday sound experiments first (section 2.2.6), but found that the network hyperparameters we chose were not appropriate for the classic ASA experiments, as the network was inadequately sensitive to masked tones. We therefore modified the network hyperparameters before conducting the classic ASA experiments (reported in sections 2.2.4 and 2.2.5), by (a) increasing the batch size from 10 to 20; (b) removing the event cochleagram reconstruction loss from the training objective; and (c) increasing the limit on cochleagram amplitude from 120 to 180 dB, when scaling the cochleagram input to image values. The use of different networks for these two sets of experiments was unintentional. Time constraints prevented us from re-running the everyday sounds experiments with the modified network, but we plan to conduct all experiments again with a shared network.

Test procedure. At test time, we input a scene cochleagram into the segmentation network in order to determine a list of candidate events for the entire sound. In addition to our custom outputs for latent variables, we modified *Detectron2*'s outputs to better suit the audio domain. Instead of outputting binary masks, we retrieve the soft masks computed before *Detectron2*'s standard binary-thresholding to allow partial overlap of the resulting candidate events. As for machine vision networks, there are typically a large number of candidate events, possibly including duplicates. To cut down on the number of candidate events, we modified *Detectron2*'s standard intersection-over-union threshold (IoU) in two ways. First, we computed the following IoU threshold for the soft masks of events e_i and e_j :

$$\text{IoU}_{ij}^{\text{soft}} = \frac{\sum_{t,f} \min(M_i, M_j)}{\sum_{t,f} \max(M_i, M_j)} \quad (2.30)$$

Second, when two candidate events were both harmonics, we computed this IoU on the estimated fundamental frequencies rather than the soft masks. This allows highly overlapping harmonic sources to be distinguished more effectively. Then, proceeding

from the highest- to lowest-confidence candidate event, we calculate the IoU of the current event with all events that have already been included. If the event exceeds an IoU threshold of 0.5 with any of the previously chosen events, it is discarded. This process results in a list of candidate events with corresponding latent variables for an entire sound, ordered by network confidence.

In summary, our custom adaptations of *Detectron2* are:

1. Self-supervised training on generative model samples, using domain randomization
2. Custom input channels to indicate the edges of the cochleagram
3. Losses to include the estimation of sound type and event-level latent variables, along with each event mask
4. Obtaining soft masks at test time to allow overlap
5. Modifying the IoU threshold to accommodate soft masks and harmonic sounds

Source construction

In source construction, candidate events are combined to create scene descriptions that each specify a set of sources and the sequence of events they emit. Inspired by particle filtering (Doucet et al., 2001), each round j of the inference procedure considers a progressively longer duration of the observed sound, providing a set of likely scene descriptions that explain the sound up to time t_j . The result of the first round of source construction is depicted in Supplementary Figure 2-12B (“Initial hypotheses”). Each initial hypothesis is a full scene, with one or more sources composed of candidate events, along with initial source-level latent variables.

One round of source construction involves 1) any candidate events that have estimated onsets $t_{j-1} < \tau^{on} \leq t_j$, and 2) the existing scene descriptions selected on the previous round (Supplementary Figures 2-11 and 2-12). To create a new scene description, one or more of these candidate events may be added to an existing scene or used to start a new scene. An event can be added to a new source, or it can be added to an old source as long as the sound type matches and it only minimally

overlaps with the other events in the source ($\text{IoU} \leq 0.05$). The new scene description is constructed and then passed onto the hypothesis optimization step. Specifically, the scene description S_i is used to initialize the means of a guide distribution q_i , corresponding to hypothesis H_i . Supplementary Figure 2-11B depicts this process over several rounds. Similarly, Supplementary Figure 4 2-12C depicts how the existing scene descriptions selected on Round 1 are combined with the last remaining event proposal on Round 2 (“ $+e_4$ ”) to create a new set of initial hypotheses.

These simple update rules can result in a large number of scene hypotheses if there are many candidate events within the interval $[t_{j-1}, t_j]$. On some rounds of sequential inference, heuristics were necessary to limit which scene hypotheses were optimized for further consideration. These heuristics essentially favored smaller scene descriptions or ranked hypotheses based on the outputs of the segmentation network (IoU and confidence). They were as follows:

1. Only consider a maximum number of candidate events on each round (max=5). Prefer candidate events with high network confidence.
2. Only consider a maximum number of scenes per round (max=100 for classic ASA illusions, max=15 for natural sounds). Prefer scene hypotheses for which events have high network confidence and low IoU between the old and new event(s).
3. After a given round (round=2), only consider hypotheses which utilize the events inferred in previous rounds.
4. Do not consider scene hypotheses with more than a maximum number of sources (max=4).
5. After the first round, do not consider scene hypotheses which add more than a maximum number of new sources (max=1).
6. Do not consider scene hypotheses which add more than a maximum number of new events (max=3) on a single round.

It would be ideal to comprehensively test the impact of these heuristics, but everyday sound inference is intractable without them.

Hypothesis optimization and selection.

Supplementary Figures 2-12B and C depict hypothesis optimization. The “initial hypotheses” are optimized with variational inference, resulting in “optimized hypotheses” that have optimized event-level variables (as reflected in the change in the rendered cochleagrams) and optimized source-level variables (not shown). We first used 250 iterations of variational inference as an initial assessment of the hypotheses. Then we ranked the hypotheses by their variational lower bound (Equation 2.28) and kept only the top five (not shown in figures). Then, for these five hypotheses, we ran additional iterations of variational inference, for a total of 2000 steps for everyday sounds and 4000 steps for classic ASA sounds. The top two hypotheses are selected to be used in the next iteration of source construction. Supplementary Figure 2-11B shows optimized and selected hypotheses over several rounds of sequential inference, demonstrating how the final top hypothesis for the observation was successively built up through several rounds.

Cleanup proposals

For the everyday sounds, we generated a few extra sets of hypotheses after completing inference on the entire sound. These hypotheses were meant to alleviate posterior degeneracy and test alternative minima that were difficult to reach through local gradient based search. They were generated by (1) removing a source, (2) removing an event, (3) changing a source’s sound type, (4) merging sources, and (5) merging events. In practice, we found that these cleanup proposals had only a minimal effect. This is possibly because the discovered hypotheses had already undergone multiple rounds of optimization during sequential inference, making it unlikely that new hypotheses could be optimized as effectively.

Computational resources

Sequential inference required considerable computational resources. Over the course of a single sound, hundreds of hypotheses could be tested. All hypotheses were

optimized on a single GPU for time ranging from 2 minutes to an hour. Therefore, sequential inference for a single sound could potentially take on the order of tens or hundreds of GPU hours. This computationally intensive inference was enabled by parallel computing.

2.5.5 Enumerative inference for simulated psychophysical experiments

In enumerative inference, rather than proposing hypotheses sequentially based on the data, we directly defined hypotheses corresponding to the choices provided to a human participant in a psychophysical experiment. A single experiment could require a participant to choose between explanations of sound that are structurally distinct in our model, for example:

- Tone sequences: the arrangement of events into one or two sources (streams)
- Co-modulation masking release, Mistuned harmonic: the presence or absence of a whistle event

In contrast, two experiments involve reports based on properties of the sound that our model treats as continuous latent variables, specifically:

- Spectral completion: the spectrum of the target sound
- Onset asynchrony: the spectrum peaks in order to make a vowel judgment

To perform enumerative inference, we first create hypotheses corresponding to each distinct “structure” to be considered in the experiment (such as the arrangement of elements into sources). For each of these structural hypotheses, we then optimize the continuous latent variables corresponding to the sources and events defined by the hypothesis (such as the spectrum of each source in the hypothesis). Specifically, we use gradient descent to optimize a “guide distribution” to approximate the posterior distribution over continuous latent variables. As this optimization can land

in local optima, we repeat it from multiple initializations. Finally, we use these optimized guide distributions to calculate the relevant psychophysical quantities. To choose between different structural hypotheses, we compare them based on marginal probability, which is estimated via importance sampling through the optimized guide distributions (and maximizing over all initializations). To extract continuous latent variables from a hypothesis, we simply take the expectation of the optimized guide distribution (weighting initializations by their importance-sampled marginal probability, which in practice was often similar between initializations).

For each hypothesis, we initialized the continuous event-level latent variables so that the initial rendered scene sound would be a relatively close match to the observed sound. This was done on a case-by-case basis based on our knowledge of the stimulus parameters and the generative models. For instance, the onsets of the source events were set equal to the onsets of the stimulus components, and the frequencies of the events were set equal to the frequencies of the stimulus components. However, the rendered stimulus was typically not exactly faithful to the original stimulus, because it was often not obvious how to set the generative model parameters by hand to exactly replicate aspects of the stimulus. A simple example of this is the fact that the model used fixed duration onset and offset ramps for events, whereas these varied somewhat in shape and duration in the experimental stimuli. The best fitting onset and offset times thus varied somewhat depending on these ramps and the initialization was not always optimal. It was thus critical to optimize these hypotheses to give the model the best chance of explaining the experimental stimulus. In addition, in order to estimate the probability associated with each hypothesis, we had to optimize the continuous source-level latent variables and estimate posterior variances on all latent variables. For brevity, we omit the long list of generative parameters used for the hypothesis initializations in each experiment, but they are available in the Github repository for the experiments (available through my thesis webpage at <https://mcdermottlab.mit.edu/mcusi/thesis/>).

Hypothesis optimization

In enumerative inference, we used 8000 steps of variational inference for a single hypothesis. We found that more steps of variational inference were needed to sufficiently estimate the posterior distribution, especially for long sounds such as the tone sequences.

Computational resources

Similar to sequential inference, enumerative inference required considerable computational resources. Depending on the experiment, there were 1-10 hypotheses for each sound. All hypotheses were optimized on a single GPU for time ranging from 30 minutes to two hours. Therefore, enumerative inference to test all the hypotheses for a single sound could take on the order of tens of GPU hours. Again, this computationally intensive inference was enabled by parallel computing.

2.5.6 Classic ASA phenomena

Overview

We simulated a set of classic psychophysical experiments and illusions in our generative model. For the results reported in the main text figures, we used a mix of enumerative or sequential inference. For all illusions, we used sequential inference to simulate unconstrained listening. For all psychophysical experiments except one, we used enumerative inference, on the grounds that the experiments explicitly asked participants to choose between different perceptual interpretations. The one exception was the cancelled harmonics experiment, where it seemed more appropriate to use sequential inference (see below).

With enumerative inference, we ran the inference procedure on all hypotheses for a stimulus 3-10 times with different random seeds. Each run of the procedure yielded a distribution over the continuous variables for each hypothesis. This distribution could be used to estimate the perceived value of one of the continuous variables (e.g. the spectrum level of part of a source). The distribution could also be integrated

out to estimate the marginal probability of that hypothesis (e.g., to choose between alternative hypotheses). For each run of the inference procedure, we computed the quantity to be plotted (this varied from experiment to experiment, see below). We then calculated the mean response across inference runs and calculated standard error bars.

Some experiments (auditory induction, co-modulation masking release, onset asynchrony, and mistuned harmonic) required estimating the threshold of a stimulus-generation parameter at which the model preferred one explanation to the other. While it would be natural to define this threshold as the point at which the posterior probability crosses 50%, this estimator is not robust to the noisy probability estimates produced by our stochastic inference procedure (and does not necessarily produce a unique value). Instead, we defined the threshold more robustly based on the integral of the posterior with respect to the stimulus-generation parameter. Specifically, if H_0 is to be preferred for low parameter values and H_1 for high parameter values, we define the threshold τ (uniquely) with respect to stimulus-generation parameter values μ_X such that:

$$\sum_{\mu_X < \tau} p(H_1|X) = \sum_{\mu_X > \tau} p(H_0|X)$$

All sounds were generated at 20kHz. Stimulus levels for the model experiments are specified in dB relative to an arbitrary model reference value.

Auditory induction

Stimuli. We adapted the experimental stimuli from Experiment 3 of ref. Warren et al. (1972). In the original experiment, participants adjusted the level of a tone until it was just audible in noise (to measure masking thresholds) or until it just sounded discontinuous, if alternating with noise bursts (to measure continuity thresholds). In both conditions, noise was played throughout the adjustment process. The experiment included 15 participants who had not participated in previous experiments in

the paper.

Because our model was not set up to actively adjust the parameters of the experimental stimuli like participants could, we instead generated stimuli spanning a range of tone levels (44-80dB in 4dB increments) and then simulated model responses at each level. We tested model responses for pure tones at 250, 500, 1000, 2000 and 4000 Hz, replicating all the frequencies in the experiment except 8000 Hz (due to its presence on the very edge of the cochleagram representation we used). For both conditions, we generated pink noise with an order-206 FIR filter with a one-octave notch. The filter’s 3dB down points were at 722 and 1367 Hz, and it reached its maximum attenuation of 40dB between 900-1200 Hz. After filtering, the level of the noise was set to 80 dB.

To generate the stimuli for the masking condition, three pure tones were embedded in a noise masker. The masker was a single noise burst of duration 1.8-s with 10-ms raised cosine onset and offset ramps. Each tone was 300-ms long with 10-ms raised cosine ramps. The tones were separated by 300-ms interstimulus intervals. The first tone began 150-ms after the onset of the noise burst, such that the last tone ended 150-ms before the offset of the masker. To generate the stimuli for the continuity condition, the tones were alternated with short noise bursts. Five 300-ms noise bursts were generated with an interstimulus interval of 300-ms. The first 150-ms tone began 150-ms after the offset of the first noise burst. The next two noise bursts were alternated with 300-ms tones. The last tone started immediately after the fourth noise burst and ended 150-ms before the onset of the last noise burst. The noise and the tones each had 10-ms onset and offset ramps, and overlapped in the region where they were ramped (i.e., were cross-faded). All stimuli were padded with 50ms of silence. In total, there were 50 stimuli resulting from this process (5 frequencies \times 10 levels). The above process was replicated 10 times with different exemplars of noise for each replication of inference, increasing the stimulus set size to 500 (10 exemplars of 50 stimuli).

Analysis (enumerative inference). We used a one-interval, two-alternative forced choice (2AFC) task to measure thresholds. For each stimulus in the masking condition

(corresponding to a tone level/frequency pair), the model compared the hypotheses that tones were present versus absent. The tone-absent hypothesis was initialized with a noise with a similar spectrum to the stimulus masking noise (as described above, we attempted to set the parameters of the generative model to match the noise, but this did not produce an exact match, and was optimized during the model inference to match the stimulus as best possible). The tone-present hypothesis initialized this noise along with the tones at the appropriate level, frequency and timing for the stimulus. For each stimulus in the continuity condition, the model compared the hypothesis that there was a continuous single tone versus four discontinuous tones. Each hypothesis was initialized with noise bursts of appropriate duration and spectrum, and tones of constant level. For each condition, this resulted in a log odds curve as a function of tone level for each tone frequency, which we used to calculate thresholds.

Homophonic continuity

Stimuli. We extracted two 1.5-s clips from Track 32 of Bregman and Ahad (1996), titled “Homophonic continuity and rise time”. Both clips were composed of continuous bandpass white noise (0-8 kHz), but had different amplitude envelopes. The initial amplitude of both sounds was a quarter of the peak amplitude, and both sounds rose and fell in amplitude twice. The first clip was gradually modulated, rising linearly to a peak over 252 ms and then falling to the initial amplitude over 252 ms. The second clip was abruptly modulated, rising to peak amplitude in 1 ms, maintaining peak amplitude for 32 ms, and then falling to the initial amplitude in 1 ms. We downsampled the original clips to a sampling rate of 20kHz. Each sound was padded with 50-ms of silence.

Analysis (sequential inference). The test of whether the model succeeds in this illusion is whether the abruptly modulated noise is interpreted as two separate sources despite the gradually modulated noise being interpreted as a single source. Therefore, we only used sequential inference for these sounds and visually assessed the results.

Spectral completion

Stimuli. We reproduced the experimental stimuli from Experiments 1 and 2 of ref. McDermott and Oxenham (2008). In the original experiment, participants adjusted the spectrum level of the middle frequency band of a short comparison noise until it sounded as similar as possible to the target noise. The starting level of the middle band of the comparison noise was set randomly between -10 and 30 dB spectrum level. The final spectrum level of the middle band was reported as an average across participants. The experiments included 8 participants between the ages of 18 and 30 years old.

Stimulus generation was identical to that in the original experiments. Each stimulus was generated by combining bandpass white noise bursts of various spectrum levels and pass bands. The standard stimulus consisted of “masker” and a “target”. The masker was 750 ms in duration and had a pass band of 500-2500 Hz. The target contained a lower tab (100-500 Hz) and an upper tab (2500-7500 Hz). The target was 150 ms in duration, and started 300 ms after the onset of the masker. To generate each noise burst, we first set all spectral domain magnitude coefficients outside the pass band to zero, and then performed an inverse fast Fourier transform. Each noise burst had 10-ms raised cosine ramps. We additionally padded each stimulus with 100ms of silence.

For Experiment 1, we generated five stimuli. The tabs in Experiment 1 were set to a spectrum level of 20 dB/Hz. Stimulus i and ii only contained the short target sound. In stimulus i, the middle band was silent. In stimulus ii, the middle band had a spectrum level of 30 dB/Hz. Stimulus iii was the standard stimulus. Stimulus iv had a masker with a spectral gap that extended from 600-2080 Hz and its spectrum level was increased so that its overall level was equal to that of the masker in the standard stimulus. Stimulus v had a masker which stopped at the tab onset and began again at the tab offset.

For Experiment 2, we generated six variations on the standard stimulus in which the spectrum level of the tabs and maskers were varied in opposite directions. The

tab and masker levels for these six stimuli were (5,35), (10, 30), (15, 25), (20, 20), (25, 15), and (30, 10) dB/Hz.

There were a total of 11 stimuli across both experiments. We replicated the stimulus generation process 10 times with different exemplars of noise for each replication of inference, increasing the stimulus set size to 110 (10 exemplars of 11 stimuli).

We also generated each comparison stimulus with a range of spectrum levels for the middle band. For Experiment 1, we generated the comparison stimulus with middle band spectrum levels spanning -15-40 dB/Hz in steps of 2.5 dB/Hz, and tab spectrum levels at 20 dB/Hz. This resulted in a total of 22 comparison stimuli. For Experiment 2, we generated each comparison stimulus with spectrum levels spanning -5-15 dB/Hz in steps of 2.5 dB/Hz. For each comparison stimulus, the tab level matched that of the corresponding target stimulus. This resulted in 8 comparison stimuli for each of 6 target stimuli, for a total of 48 comparison stimuli. We replicated this process 10 times with different exemplars of noise for each replication of inference, increasing the comparison stimulus set size to 700 (10 exemplars of 70 comparison stimuli across both experiments).

Analysis (enumerative inference). For each stimulus, we optimized a single structural hypothesis that was designed based on the stimulus generation parameters. For the first two stimuli in Experiment 1, the structural hypothesis contained a single noise source corresponding to the target. For the rest of the Experiment 1 stimuli and all stimuli in Experiment 2, the structural hypothesis contained two noise sources that corresponded to masker and target. For 1v, the source corresponding to the masker contained two events; otherwise, each source contained one event. We note that these structural hypotheses accord with the number of sources and events found by sequential inference except in the case of 1iv, where the masker is accounted for by two simultaneous sources. Nevertheless, for 1iv, the inferred middle band spectrum level is comparable in either inference case.

The structural hypothesis was initialized with multiple settings of the middle band spectrum level of the target. The hypothesis was otherwise initialized to match the masker and tab level in the stimulus. This emulated the variable starting level of the

middle band in the experiment. For Experiment 1, the initial spectrum levels were -20, 0, 10, 20, and 30 dB/Hz. For Experiment 2, the initial spectrum levels were -5, -2.5, 0, 5, 10, 12.5, 20 dB/Hz. After variational inference, for each initialization, we selected the batch of sampled scenes with the best score (described above in Section 2.5.3). We averaged the tab spectrum over the batch. Then, we took a weighted average of the tab spectra (arising from the different initializations) using their importance-sampled marginal probability. Then, we computed the mean squared error between this average spectrum and the inferred latent spectra for each comparison stimulus (described above). The model’s judgment was selected to be the spectrum level of the comparison stimulus which minimized the error. Averaging over runs of inference with different random seeds provided the plotted value and standard error bars.

Co-modulation masking release

Stimuli. We adapted the experimental stimuli from Experiment 1 of ref. Hall et al. (1984). In the original experiment, participants were asked to detect a tone in bandpass noise that varied in bandwidth. The different spectral regions of the noise were either co-modulated or had random amplitude envelopes. Participants’ thresholds were measured using a two-interval, 2AFC procedure. The experiment included five highly experienced participants with two hours of training in all experimental conditions.

For each bandwidth, we generated a co-modulated noise burst and a random noise burst. The random noise burst was generated from white noise (with power from 0-10 kHz, i.e. up to the Nyquist frequency). To generate co-modulated noise, we multiplied the white noise by a low-pass noise with power between 0-10 Hz. Then, we bandpass filtered both noises to the appropriate bandwidth, centered on 1000 Hz, using a fourth-order Butterworth filter and forward-backward filtering. After filtering, we confirmed that the absolute spectrum level of the co-modulated and random noises were within 1 dB/Hz of 40 dB/Hz within the passband. We trimmed each noise to 400-ms and applied 50-ms raised cosine ramps. We then generated a set of 1 kHz tones with a range of levels (40-85 dB in steps of 5 dB). The tones were 400-ms with 50-ms raised

cosine ramps. We added the tones to each noise, for a total of 20 stimuli for each noise bandwidth (2 noise conditions \times 10 tone levels). We repeated this process for 100, 200, 400, and 1000 Hz bandwidths, resulting in 80 stimuli. Finally, we repeated this process ten times with different exemplars of noise for each replication of inference, increasing the stimulus set to 800 sounds.

The major difference between the original experiment stimuli and our model simulation is that the frequency cutoff of the low-pass noise for the model experiment was 10 Hz, instead of the original 50 Hz. We used slower amplitude modulations in the stimuli for the model experiment because the relatively coarse temporal resolution of the cochleagram representation limited the rate of modulations that could be resolved. We also omitted the 25 and 50 Hz bandwidth conditions due to the frequency resolution of our cochleagram.

Analysis (enumerative inference). For each stimulus, we compared the hypotheses that a tone was present versus absent (one-interval, 2AFC task). For all hypothesis initializations, we initialized the noise with the appropriate amplitude envelope. The tone absent hypothesis only contained the noise burst. For the tone present hypothesis, we made two initializations in order to aid in finding the best setting of the continuous latent variables for this hypothesis. One initialization had the tone matching the tone level in the stimulus. The other had a “quiet” tone initialized at 0 dB. For each bandwidth and noise condition, we computed the log odds as a function of tone level then computed the detection threshold.

Mistuned harmonic

Stimuli. We reproduced a subset of the experimental stimuli from ref. Moore et al. (1986). In the original experiment, participants were presented with a complex tone and asked to indicate whether they heard a single sound (with one pitch) or two sounds (a complex tone and a pure tone). On half of trials, the tone was harmonic and on the other half one frequency component was mistuned. The authors then calculated the degree of mistuning which was necessary to detect two sounds. The

experiment included four participants, the authors and one volunteer, who were all highly experienced with psychoacoustics and with the task of detecting mistuning in harmonic complexes.

We generated 400-ms complex tones with a fundamental frequency of 100, 200 or 400 Hz. The complex tones had equal amplitude harmonics, each with a level of 60 dB. Tones at 100 and 200 Hz included harmonics 1-12 and tones at 400 Hz included harmonics 1-10. Each tone was given 10-ms raised cosine onset and offset ramps. Based on these harmonic complex tones, we created complex tones where one component was mistuned. We mistuned harmonics 1-3 by 5, 10, 20, 30, 40, or 50% of the fundamental frequency. Including the in-tune harmonic complexes (0% mistuning), this resulted in 63 stimuli (3 fundamental frequencies \times 3 harmonic numbers \times 7 mistuning levels). Every stimulus was padded with 50-ms of silence.

Analysis (enumerative inference). For each stimulus, we compared the hypothesis that there was one harmonic source versus the hypothesis that there was one harmonic source and one whistle source (one-interval, 2AFC task). For the single source hypothesis, we initialized a single harmonic source with the stimulus parameters of the in-tune harmonic complex. For the two source hypothesis, we made two initializations in order to aid in finding the best setting of the continuous latent variables for this hypothesis. Both initializations had a whistle source at the mistuned frequency and a harmonic source with the fundamental frequency of the stimulus, but they varied in the relative energy of the whistle and the component within the harmonic source that corresponded to the mistuned harmonic number. In the first initialization, the component corresponding to the mistuned harmonic was attenuated by 30 dB in the harmonic source, and the whistle source was set to 60 dB. In the second initialization, the component corresponding to the mistuned harmonic was only attenuated by 6 dB, and the whistle source was set to 50 dB. For each fundamental frequency and harmonic number, we computed the log odds as a function of mistuning percent then computed the detection threshold.

Frequency modulation

Stimuli. We based our stimulus design on the classic demonstration first described in ref. McAdams (1984). We generated a 1-s complex tone in which the odd harmonics had a steady fundamental frequency and the even harmonics were coherently frequency modulated. The even harmonics began in a harmonic relationship with the odd harmonics, but were immediately frequency modulated at a rate of 2 Hz, with a maximum frequency change of 70 Hz. The fundamental frequency of the odd harmonics was 300 Hz. The stimulus contained harmonics 1-12, with levels of 75, 63, 56, 52, 48, 48, 42, 46, 38, 45, 35, 39 dB respectively. The entire stimulus had 20-ms raised cosine ramps, and was padded with 50-ms of silence.

Analysis (sequential inference). The test of whether the model succeeds in this illusion is whether the modulated components are discovered as a separate source. Therefore, we only used sequential inference for these sounds and visually assessed the results. Sequential inference for this sound is depicted in Supplementary Figure 2-12.

Asynchronous onsets

Stimuli. We reproduced a subset of the stimuli from Experiment 1 in ref. Darwin and Sutherland (1984). In the original experiment, participants heard short speech-like sounds and categorized them as either /I/ or /e/. The experiment included six participants, who had practice with the control conditions.

We generated four of the continua from the original experiment. The first was the ‘basic’ continuum, which was composed of seven vowels for which the first formant was varied from 375 to 500 Hz in equal steps. The other three continua were created by adding 500-Hz tones to the basic continuum. The first ‘onset=0’ continuum was created by adding a tone with the same onset and offset as the vowel. The onset of the tone in the other two continua was asynchronous with the vowel, either 32-ms before or 240-ms before the vowel. To generate the vowel sounds, we used a publicly available Python interface of the Klatt speech synthesizer, which was used in the

original experiment (Klatt, 1980; Sprouse, 2013). To create the basic continuum, we generated seven 60-ms long vowels with a fundamental frequency of 125 Hz and overall level of 56 dB. In each vowel, formants 2-5 were centered at 2300, 2900, 3800, and 4600 Hz respectively. The bandwidths of these formants were unspecified in the original paper and were set to the Klatt synthesizer defaults (70, 150, 200 and 200 Hz for formants 2-5, respectively). The bandwidth of the first formant was 70-Hz and kept constant in each vowel. The first formant frequency varied across vowels, with values of 375, 296, 417, 438, 459, 480 and 500 Hz. These values are referred to as the ‘nominal first formant frequencies’ for the other continua. Each vowel had 16-ms linear onset and offset ramps. To generate the tones to add to each vowel in the basic continuum, we first measured the level of the 500 Hz component in each vowel. For each vowel, we generated a pure tone that was 6-dB higher in level than the 500-Hz component. This pure tone constructively interfered with the 500 Hz component of the vowel to produce a 9.5 dB increment from the original level at 500 Hz. After adding 16-ms linear onset and offset ramps to the tone, we added the tone to the vowel so as to produce the desired onset difference. This process resulted in 28 stimuli (4 continua \times 7 vowels). All stimuli had the same overall duration, and were zero-padded so that the vowel had the same absolute onset time in each stimulus. The stimulus with the largest onset asynchrony was zero-padded with 50-ms of silence.

Analysis (enumerative inference). For each stimulus, we found the frequencies where the first two “formant” peaks occurred in the inferred spectrum of the harmonic source. We then computed classification probabilities according to empirical first and second formant distributions measured by Hillenbrand et al. (1995). This enabled us to compute the vowel boundary (classification threshold) in terms of the nominal first formant frequency.

For each stimulus in the basic and ‘onset=0’ continua, we optimized a single structural hypothesis with a harmonic source only, initialized with the stimulus parameters (in experiments not reported here, we found that including an additional structural hypothesis, a harmonic source and a simultaneous 500-Hz whistle source, gives the same result). For the asynchronous onset continua, we optimized a single

structural hypothesis (a whistle source and a harmonic source) from two initializations. The first initialization included a harmonic source with the ‘onset=0’ spectrum and a whistle that ended at the beginning of the harmonic source. This corresponded to the possibility that the overlapping 500-Hz component is grouped with the harmonic source. The second initialization included a harmonic source with the basic spectrum and a whistle that ended synchronously with the harmonic source. This corresponded to the possibility that the 500-Hz component of the harmonic event is ‘captured’ by the asynchronous whistle.

We then selected a subset of the data in Hillenbrand et al. (1995) corresponding to the vowels /I/ and /e/. To compensate for potential differences between the speakers in the original vowel set and those which the synthetic vowel stimuli were modeled on, we normalized both the empirical formant distribution and inferred formant distribution (Adank et al., 2004) by z-scoring. We z-scored the inferred formants using the mean and standard deviation computed across all conditions and seeds. We z-scored the empirical formants using the mean and standard deviation of the whole selected subset of formants. We then separately computed the mean and covariance of the z-scored /I/ and /e/ formant distributions. To derive classification probabilities for the model inferences, we computed the probability of each z-scored, inferred formant pair under two normal distributions with the empirical means and covariances. For each continuum, this provided a classification probability as a function of nominal first formant frequency. Finally, we computed the vowel boundary threshold. We took a mean of the vowel boundary over the different runs of the inference procedure and computed the standard error.

Cancelled harmonics

Stimuli. We adapted the stimuli from Experiment 1 of ref. Hartmann and Goupell (2006). In the original experiment, listeners matched the frequency of a comparison tone to their percept of a gated harmonic within a harmonic complex tone. The experiment included four male listeners between the ages of 21-65; two were the authors.

We created harmonic complex tones with fundamental frequencies spanning 190 to 210 Hz in five equal steps (in Hz). The duration of each tone was 750-ms, with 10-ms raised cosine onset and offset ramps. Each tone contained harmonics 1-30. The harmonics were each 45 dB and added in sine phase. For each fundamental frequency, we created a set of stimuli each with a different gated harmonic component (harmonics 1-3,10-12,18-20). We gated the component with 10-ms raised cosine ramps to create five tones of 100-ms each. The onset of the first tone and the offset of the last tone were aligned to those of the harmonic complex. Therefore, the interstimulus interval between the tones was 62.5-ms. The entire stimulus was padded with 50-ms of silence. This process led to a total of 45 stimuli (5 fundamental frequencies \times 9 harmonic numbers).

The stimuli for the model experiment were shorter than those for the original experiment. The original experiment used 9.1-s long harmonic complexes with four tones of 1.3-s. Inference with this stimulus duration would have been prohibitively computationally expensive, so we instead used 750-ms long harmonic complexes with five tones of 100-ms.

Analysis (sequential inference). The test of whether the model succeeds in this illusion is whether it discovers any whistle sources that correspond to the gated components. Therefore, we used sequential inference, after which we selected the top-scoring hypothesis which contained any whistle sources. For any whistle source in that scene, we recorded its source-level mean fundamental frequency (μ_f). We took the average μ_f across all whistle sources if there were more than one (because this typically corresponded to some of the gated tones being assigned to a different whistle source than others, but with similar frequencies). If no hypothesis contained a whistle source, we recorded this as a ‘no-match’ trial, as in the original experiment. For each trial (corresponding to a harmonic number and fundamental frequency), we calculated the percent matching error as a proportion of the gated component frequency. We also recorded the number of ‘no-match’ trials for each harmonic number summed across fundamental frequencies. We plotted the distribution of errors in Figure 2-5D.

Frequency proximity

Stimuli. We adapted two tone sequences from Experiment 2 of Tougas and Bregman (1985), which are similar to the demonstrations in Track 17 of Bregman and Ahad (1996) titled “Failure of crossing trajectories to cross perceptually”. In the original sequences, tones in an ascending sequence are alternated with tones in a descending sequence. In the first clip, all the tones in both sequences are pure tones. In the second clip, the ascending sequence is composed of harmonic tones. Listeners judge whether they can hear the ascending and descending sequences, or whether they hear two ‘bouncing’ sequences.

We interleaved an ascending and descending sequence, each composed of 6 tones with frequencies evenly spaced on a log-frequency scale (ascending: 400, 504, 635, 800, 1008, 1270 Hz; descending: 1600, 1270, 1008, 635, 504, 400 Hz). Each tone was 100-ms in duration with 8-ms onset and offset ramps, with adjacent tones played back-to-back. In the first version (where all tones were pure tones), all tones were 70 dB. In the second version (where the ascending sequence contained harmonic complex tones), the harmonic complex tones contained the first four harmonics of the fundamental frequency (which were the pure tone frequencies from the first version). The pure tones were 73 dB while each component of the harmonic tone was 67 dB. Each sequence was padded with 50 ms of silence.

Analysis (sequential inference). Since the stimuli from this experiment are often used as standalone demonstrations, we tested whether “bouncing” sequences are discovered in the first sequence and “crossing” sequences are discovered in the second sequence. Therefore, we only used sequential inference for these sounds and visually assessed the results. Sequential inference for the first (pure-tone) sequence is depicted in Supplementary Figure 2-11.

Bistability

Stimuli. We adapted the classic ABA sequences used in Track 3 of Bregman and Ahad (1996), titled “Loss of rhythmic information as a result of stream segregation”,

in order to estimate the stimulus parameters Δf and Δt that lead to one- or two-source perceptual organizations. These effects were first measured by Experiment 2.3.2 in ref. Van Noorden (1975), but this experiment used 80-s sequences that are computationally infeasible for our model, and so we generated shorter sequences.

In the typical ABA tone sequence, three tones are followed by a silence of the same duration as the tone onset-to-onset interval. The first and third tone in the triplet (A) have the same frequency, which can be different from the frequency of the second tone (B). We generated versions of such sequences in which the onset-to-onset interval was 67, 83, 100, 117, or 150 ms. The A tone was always 1000 Hz, and the B tone was 3, 6, 9, or 12 semitones higher in frequency. The tones were all 70-dB and 50-ms long with 10-ms raised cosine onset and offset ramps. The ABA triplet was repeated four times. The stimuli were padded with silence so their total duration was equal (3.1 s). This resulted in 25 stimuli (5 frequency intervals \times 5 time intervals).

Analysis (enumerative inference). For each tone sequence, we compared two hypotheses: one stream (all tones in one source) versus two streams (high tones in one source, low tones in a separate source). We initialized each hypothesis with a set of whistle events with the stimulus frequencies and timings that were organized into the appropriate sources. After variational inference, we computed the log odds of the hypotheses for each sequence. We took the average log odds across all runs of the inference procedure and computed the standard error.

Cumulative repetition

Stimuli. We adapted the tone sequences used in ref. Thompson et al. (2011). In the original experiment, listeners heard a 12.5-s ABA sequences. At any point during each sequence, they could freely indicate whether they were hearing one or two sources. The experiment included eight listeners between 23-57 years old.

Listeners in Thompson et al. (2011) could respond at any point during a sequence. To obtain an analogous measure of the effect of time on the model’s inferred perceptual organization, we instead evaluated the model for multiple sequences, each with a

different number of repetitions. We generated two sets of ABA sequences (with the tone arrangement described in Section 2.5.6), one with a frequency difference Δf of eight semitone and one with a frequency difference of four semitones. The A tone was 500 Hz. The tones were 50-ms in duration with 10-ms raised cosine onset and offset ramps, and with 125-ms onset-to-onset intervals. Each ABA triplet was thus 500-ms in duration. We generated sequences with 1-6 repetitions of the ABA triplet. This resulted in 12 tone sequences total ($2 \Delta f \times 6$ sequence durations).

Analysis (enumerative inference). The initialization of inference, and the analysis, was identical to Section 2.5.6.

Effects of context (1)

Stimuli. We adapted the stimuli used in Task 1 of Experiment 2 in ref. Bregman (1978b). In the original experiment, listeners first heard the standard AB tone pair in isolation and then 12 repetitions of the ABXY sequence. They rated whether the standard was audible as a separate pair in the sequence. The experiment included 16 young adult participants.

We generated the seven ABXY sequences in the original experiment, but only repeated them four times (because it was computationally prohibitive to run inference on longer stimuli). For the four “isolate” stimuli, the ABXY frequencies were (2800, 1556, 600, 333), (600, 333, 2800, 1556), (2800, 2642, 1556, 1468), and (333, 314, 600, 566) Hz, respectively. For the three “absorb” stimuli, the ABXY frequencies were (2800, 1556, 2642, 1468), (600, 333, 566, 314), and (2800, 600, 1468, 314) Hz, respectively. Each tone was 100-ms with 10-ms sine-squared onset and offset ramps, and 10-ms silences between tones. In the original experiment, all tones were 70-dB except for 333 and 314 Hz, which were 77-dB (to equate the loudness of the tones, based on equal loudness contours at 70 phon). This adjustment was unnecessary for the model, so we presented all tones at 70-dB. In the original human experiment, the sequence was faded in and out to prevent participants from adopting strategies based on the beginning or end of the sequence. This was not an issue with the model, and so all cycles of the sequence were presented at the same level.

Analysis (enumerative inference). For each tone sequence, we compared two sets of hypotheses. In both sets, all hypotheses contained one source with a pair of tones, and one or two other sources. The first set contained the two hypotheses in which A and B were paired in their own source (one with X and Y in a second source, and one with X and Y in different sources). The other set contained the eight hypotheses in which A and B were in separate sources (each with a different assignment of X and Y to the sources containing A and B, or to separate sources). We initialized each hypothesis with a set of whistle events with the stimulus frequencies and timings that were organized into the appropriate sources. After variational inference, we summed the marginal probabilities within each set and then calculated the log odds for each sequence. We averaged the log odds across all runs of the inference procedure and computed the standard error.

Effects of context (2)

Stimuli. We adapted the stimuli used in Task 1 of Experiment 2 in ref. Bregman (1978b). In the original experiment, listeners first heard a standard tone pair and then a longer sequence that contained a “target” tone pair, comprising tones of the same frequency. The target tones could be in the same order as the standard, or in reverse order. The longer sequence also contained two “distractor” tones, one immediately preceding the target pair and one following the target pair. Participants judged whether the standard and the target had the same order. The experiment included 13 participants from ages 16 to 26 years.

There were 4 stimuli in this experiment, each corresponding to a different “captor” condition that varied in the presence of captor tones that might cause distractor tones to segregate from the target tones. In the “none” captor condition, there were only distractor tones and target tones in the long sequence. In the other captor conditions, the distractor and target tones were in the same configuration, preceded by 3 captor tones and followed by 2 captor tones. The first target tone had frequency 2200-Hz and level 60-dB. The second target tone had frequency 2400-Hz and level 60-dB. The frequency of the distractor tones was 1460-Hz and their level was 65-dB. The

frequencies and levels of the captor tones in the three conditions with captors were (590-Hz, 63-dB), (1030-Hz, 60-dB) and (1460-Hz, 65-dB) respectively. The duration of each tone was 70-ms (compared to 45-ms in the original experiment), with 7-ms on-ramps and 5-ms off-ramps. Each target tone and the first distractor was preceded by 9-ms of silence and followed by 0-ms of silence. The captor tones and second distractor tone were preceded by 9-ms of silence and followed by 64-ms of silence.

Analysis (enumerative inference). For each tone sequence, we compared the pair of hypotheses where the target tones were in their own source to the hypothesis where they were grouped with the distractors. The pair of hypotheses with the target tones in their own source included the hypothesis where the captors and distractors were grouped and the hypothesis where the captors and distractors were segregated. Within this pair, we summed the marginal probabilities. Then we took the log odds of the targets-alone hypotheses and the targets-plus-distractors hypothesis. We averaged the log odds across all runs of the inference procedure and computed the standard error.

2.5.7 Network comparisons

We selected a set of source-separation neural networks to compare to our model. Selection was based on 1) public availability of pre-trained weights, 2) good performance in machine hearing competitions and 3) the goal of spanning a variety of training methods, tasks, network architectures and natural sound datasets. These criteria yielded a set of seven networks:

1. ConvTasNet for two-speaker separation, trained on LibriMix (Cosentino et al., 2020; Pariente et al., 2020)
2. the same ConvTasNet trained with background noise
3. TDCN++ for open domain sound separation trained on FUSS (Wisdom et al., 2021)

4. Open-Unmix RNN for music separation trained on MUSDB18 (Stöter et al., 2019)
5. Open-Unmix RNN for music separation trained on a much larger, private dataset (Stöter & Liutkus, 2021)
6. Open-Unmix RNN for speech enhancement trained on the Voicebank+DEMAND corpus (Uhlich & Mitsufuji, 2020)
7. MIXIT network trained on YFCC100m (Wisdom et al., 2020) (MIXIT is the only network optimized with an unsupervised training objective)

To compare how well these networks and our model matched human perception, we quantified the dissimilarity with human perception of classic ASA phenomena. For a given input soundwave, the source-separation networks output a set of soundwaves (estimates of the premixture sounds) rather than the symbolic scene description provided by our model. To simulate psychophysical experiments, we devised a way to obtain psychophysical judgments from the output soundwaves.

For each sound mixture, we defined “standard sets” of sources: each is a set of sounds rendered from a scene hypothesis for that sound mixture. These were typically a set of sounds that reflect what human listeners hear (e.g., a pair of standard sounds for frequency modulation: one sound with the frequency-modulated harmonics and one sound with the steady harmonics) or sounds rendered from the experimentally-defined hypotheses of enumerative inference, and are described below for each ASA phenomenon. For example, in the bistability experiment, there were two standard sets with two sounds each. The first represents the two stream hypothesis: one sound has all the high frequency tones and the other sound has all the low frequency tones. The second represents the one stream hypothesis: one sound is the mixture with both high and low tones, and the other sound is silence (because the networks always output at least two sources). The standard sets were compared to network outputs using one of four methods to obtain a correlation coefficient with human results, based on what seemed most appropriate for each ASA phenomenon.

Method 1, Comparing network to standard cochleagrams. For illusions (homophonic continuity, frequency modulation, frequency proximity), we computed the maximum Pearson correlation between the cochleagram values of network outputs and the single set of standard sources that human listeners are considered to hear.

The next two methods could involve determining the network’s preference for one hypothesis over another. To do so, we computed the L2 distance between each the network outputs $N = \{N_i\}$, and each standard set $S_k = \{S_{kj}\}$, in cochleagram space. We found the correspondence of network outputs i to standards j which minimized this L2 distance for each standard set. For experiments which compared two hypotheses, we then compared these minimal distances to each standard set, to provide a measure of the network’s preference for hypothesis H_1 over hypothesis H_2 , specifically:

$$\text{Preference}(H_1 > H_2) = \frac{d(N, S_2) - d(N, S_1)}{d(S_1, S_2)} \quad (2.31)$$

$$d(N, S_k) = \min_{i,j, i \neq j} ||N_i - S_{kj}||_2 \quad (2.32)$$

where the denominator is the minimal distance between each set of standards, used to normalize the distance metric.

Method 2, Comparing network- and human- perceptual judgments. For each stimulus, the network output is used to determine a continuous “perceptual” judgment. We computed the Pearson correlation between the set of network judgments and the human judgments for all stimuli. In the case of the remaining tone sequences (bistability, cumulative repetition, and effect of context), the network judgment is its preference for one explanation over another. In the case of spectral completion and cancelled harmonics, the judgment is an estimate of continuous variable (spectrum level and frequency). Note that for cancelled harmonics, defining a standard set was not required.

Method 3, Comparing network- and human- thresholds. For auditory induction, co-modulation masking release, and mistuned harmonic (which measured thresholds), we used the change in network preference as a function of an experimental parameter to estimate a threshold (using the threshold method as described in Section 2.5.6), and then compared the network thresholds to human thresholds. For onset asynchrony, we used the change in an acoustic property (formant frequencies) of the preferred network output to estimate a threshold. The neural networks sometimes did not change the sign of their preference across all stimulus values. In these cases, we assigned the network judgment just above or below the range of stimulus parameters, analogous to setting a threshold to a ceiling or floor value when it cannot be measured in a human participant (we then used Spearman rather than Pearson correlation because of the resulting truncated distribution of judgments).

In summary, the three methods to obtain a correlation are:

1. Pearson correlation across pixels (between cochleagrams)
2. Pearson correlation across stimuli (between continuous judgments)
3. Spearman correlation across stimulus parameters (between thresholds)

For each experiment, we also computed a baseline result that used the input mixture sounds as “outputs”, that is, without any source separation. We then computed the dissimilarity for the experiment as:

$$\text{Dissimilarity} = \frac{1 - r}{1 - r_{\text{baseline}}}$$

where r is the appropriate correlation coefficient as described above, normalized by the baseline dissimilarity so that the dissimilarity of zero corresponds to a perfect correlation and a dissimilarity of one corresponds to the baseline correlation. Note that the dissimilarity can exceed 1 if the model results deviate more from the human results than the baseline. We then averaged the dissimilarity across all experiments. These results are shown as the hatched grey bars in Figure 2-7A.

We computed two versions of our model’s results for use in the dissimilarity comparisons.

- **Enumerative plus sequential inference**, as outlined in Section 5.6. The enumerative inference method of obtaining results is arguably most similar to what human participants do when they complete experiments, but was only possible with our model, as it leverages the generative model to constrain inference. This result is shown as the blue bar in Figure 2-7A.
- **Sequential inference only**: For every ASA phenomenon, we used sequential inference to obtain most likely scene description. We analyzed the sounds rendered from this description exactly as for the sounds output by the source separation networks. This is arguably the fairest comparison to the source separation networks, but would be expected to produce a worse match to human results. This result is shown as the pink bar in Figure 2-7A.

We also trained an additional source separation network on samples from the generative model. We chose to use the TCDN++ network because it was designed for open domain sound separation with more than two premixture sounds, which is most similar to our model. We used the same network architecture as the TCDN++ reported in Wisdom et al. (2021), but trained several versions with different hyperparameters (batch size, learning rate) and datasets (varying dataset size, higher density of events than the prior or not, and the use of domain randomization in which samples were obtained from uniform distributions instead of the model’s prior). From these several TCDN++ networks, we selected the one with the highest similarity to human results. This network was trained on a dataset consisting of approximately 180-h of sound mixtures ($n=327564$) until the validation error converged (268000 iterations, batch size=20, learning rate= $1e-5$). The sound mixtures were samples from our generative model fit to natural sounds, all limited to contain 1-4 sources. This result is shown as the solid grey bar in Figure 2-7A.

In the sections that follow, we define the standard source sounds for each experiment that were compared to the source separation network outputs to yield simulated

experimental results, as well as the procedure used to obtain the experiment result.

Auditory induction

For each continuity stimulus, we generated two pairs of standard sources. The first pair contained a sound with the noise bursts on their own and a sound with the (discontinuous) tones on their own. The second pair contained a sound with the noise bursts and a sound with a continuous constant-amplitude tone. For each masking stimulus, we also generated two pairs of standards. The first pair included the input mixture itself (no segregation of tones) and silence. The second pair included a sound with the tones only and a sound with the noise only.

For each condition and tone frequency, the preference measure defined in Equation 2.31 quantified whether continuity was preferred over discontinuity as a function of tone level and whether the tone was detected or not, as a function of tone level. Using Method 3, we obtained the Spearman correlation between the network and human thresholds.

Homophonic continuity

For the clip with the abrupt amplitude change, we generated a pair of standard sources: a sound comprising just the two short, louder noise bursts, and a sound comprising the long, quiet noise. For the clip with the gradual amplitude change, the standard was simply the stimulus itself. Using Method 1, we computed the Pearson correlation between the time-frequency bins of the cochleagrams of the network outputs and the standards, and selected the best match.

Spectral completion

The standards were the comparison stimuli described in Section 2.5.6 (Methods: Spectral Completion stimuli), with tab levels that matched the target and varying middle band spectrum levels. We found the standard which minimized the distance with any network output (finding the network output that best captured the target). The network’s judgment for a stimulus was chosen to be the middle band spectrum level

of the selected standard. Using Method 2, we correlated the network and human judgments of the spectrum level.

Co-modulation masking release

The two pairs of standard sources were (1) the input mixture itself and silence and (2) one sound with the noise only and one sound with the tone only. For each noise type and noise bandwidth, the preference measure quantified whether the tone was detected or not as a function of tone level. Using Method 3, we obtained the Spearman correlation between the network and human thresholds.

Mistuned harmonic

The two pairs of standard sources were (1) the input mixture itself and silence and (2) one sound with the harmonic components only and one sound with the mistuned tone only. For each fundamental frequency and harmonic index, the preference measure quantified whether the tone was detected or not as a function of mistuning. Using Method 3, we obtained the Spearman correlation between the network and human thresholds.

Frequency modulation

We generated a pair of standards: a sound with only the frequency-modulated components and a sound with only the constant-frequency components. Using Method 1, we computed the Pearson correlation between the time-frequency bins of the cochleagrams of the network outputs and the standards, and selected the best match.

Asynchronous onsets

The two standards were the ‘basic’ vowel and the ‘onset=0’ vowel with the same nominal first formant frequency as the input stimulus. We selected the network output which minimized the distance to either standard (finding the network output that best captured the vowel). Then, we estimated the first and second formant

frequencies from the selected output using linear predictive coding (Markel & Gray, 1976; Snell & Milinazzo, 1993). The rest of the analysis followed the same steps as for the model with enumerative inference in order to derive vowel thresholds. Using Method 3, we computed the Spearman correlation between the network and human vowel thresholds.

Cancelled harmonics

The logic for the analysis of this experiment is based on the idea that if the network successfully separated out a whistle source, then the hypothetical whistle source should have a higher amplitude at one frequency than the rest. We computed the spectrum of each network output and found the peaks in the spectrum that were spaced at least 95% of the fundamental frequency apart. We then computed the amplitude ratio of the highest peak to the second highest peak. We selected the network output with the highest peak-to-peak amplitude ratio. The network’s judgment was selected to be the frequency of the highest peak in this output. We then calculated the proportion of trials for which the percent error was less than or equal to two percent, as a function of harmonic number. Using Method 2, we correlated this proportion with the human results. This method is conceptually similar to using standards with pure tones of varying frequencies, but it allowed us to obtain a more precise pitch judgment.

Frequency proximity

For the pure-tone stimulus, the pair of standard sources was the low-frequency bouncing melody and the high-frequency bouncing melody. For the alternating pure-complex tone stimulus, the pair of standard sources was the ascending melody and the descending melody. Using Method 1, we computed the Pearson correlation between the time-frequency bins of the cochleagrams of the network outputs and the standards, and selected the best match.

Bistability

For each tone sequence stimulus, we generated standards for the two-source and one-source explanation. The pair of two-source standards were one sound with all the high-frequency tones and another sound with all the low frequency tones. The standard set for the one-source explanation was the mixture sound itself and silence. We found the minimal distance between any set of network outputs and each standard set. We then computed the difference in distance to the two-source standard set versus the one-source standard. For each setting of Δf and Δt , this provided a measure of whether two sources were preferred over one source.

For each tone sequence stimulus, we generated two pairs of standards, for the two-source and one-source explanation. The standard set for the two-source explanation contained one sound with all the high-frequency tones and another sound with all the low frequency tones. The standard set for the one-source explanation contained the mixture itself and silence. The preference measure quantified whether two sources were preferred over one source for each setting of Δf and Δt . To compare with human data, we labeled each (Δf , Δt) point as above the human one-source threshold (1), in the bistable region (0.5), or below the human two-source threshold (0). Using Method 2, we computed the Pearson correlation between the network preferences and the human judgments as reflected in these labels.

Cumulative repetition

We computed the network preferences as for bistability, but correlated them with the human proportion of two-source responses.

Effects of context (1 and 2)

Both of these experiments included comparing hypotheses, for which multiple sequences corresponded to a hypothesis. Therefore, for each hypothesis, we generated sets of standards corresponding to the sets of structural hypotheses described in Section 2.5.6 (Effects of context 1 and 2, analysis). Since multiple sets of standards

corresponded to a single hypothesis, we took the minimum distance across all sets of standards to calculate the distance of the network outputs to a hypothesis. The preference measure quantified which hypothesis was preferred (Expt. 1: A and B paired in their own source versus a non-AB pairing; Expt. 2: target tones in their own source versus grouped with distractors). Using Method 2, we computed the Pearson correlation between the network preferences and the human judgments.

2.5.8 Model alternatives

We assessed the four model alternatives as described in Section 2.5.6. (1) For the MAP sources lesion, we set each set of source parameters to the modes of the temporal hyperpriors and each of the variance and lengthscale hyperpriors, specific to that source’s type. (2) For the uniform lesion, we set the distributions over variance and lengthscale to a uniform distribution for each Gaussian Process (frequency trajectory, amplitude trajectory, and spectral shape). (3) For the spectral swap lesion, we used an Ornstein-Uhlenbeck kernel for the harmonic source model and a squared-exponential kernel for the noise source model, and then remeasured the hyperpriors as described in Section 2.5.1. We only tested ASA results which had a noise or harmonic source in them because the tone sequences are not affected by this lesion. (4) For the stationary covariance lesion, we only altered the whistle source in order to investigate the effect of the non-stationary on tone sequence grouping. We fixed the non-stationary kernel parameter β to zero and re-measured the hyperpriors as described in Section 2.5.1.

2.5.9 Everyday sound experiments

Experiments were run online using Amazon Mechanical Turk. Prior to each experiment, potential participants gave consent and indicated that they were wearing earphones or headphones. They used a calibration sound to set their volume to a comfortable level. Participants were initially screened with a short experiment to check that they were wearing earphones or headphones (Woods et al., 2017). If participants failed the headphone check, they were compensated and did not continue to

the main experiment. All experiments were approved by the Committee on the use of Humans as Experimental Subjects at the Massachusetts Institute of Technology, and were conducted with the informed consent of the participants.

Stimuli

To assess the ability of the model to infer perceptually valid scenes from naturalistic sounds, we sourced a small subset of sounds from the Free Universal Sound Separation dataset (FUSS; Wisdom et al. (2021)). The FUSS dataset contains mixtures generated by adding together audio clips of everyday sounds and then simulating reverberation. FUSS was designed for open domain source separation, with each premixture clip derived from one of over 300 sound categories.

The mixture clips in FUSS are 10-s long, composed of 1-4 sounds from the FSD50K dataset with simulated reverberation. There is always one background sound in each mixture clip, defined to be a sound which extends the entire duration of the clip. We randomly selected 50 2-s clips from the training set of FUSS, subject to a few constraints. The main constraint was that each mixture clip should contain three pre-mixture recordings of at least 200-ms in duration. Second, although sounds in FUSS are not explicitly labeled, we recovered labels from FSD50K for each pre-mixture sound. This allowed us to exclude four categories out of the 357 categories included in FUSS: “Speech”, “Scratching (performance technique)”, “Mechanisms”, and “Human group actions”. We excluded speech and scratching because we knew that our model would be poorly suited to the variable spectra that occurs in these sounds. We excluded Mechanisms, and Human group actions because their names suggested that they contained more than one perceptual stream. We used sequential inference to obtain full scene descriptions for each sound mixture, and then rendered each inferred source sound into audio using the maximum a posteriori scene description. We used these rendered audio signals in the experiments. In addition, the original mixture clips were used in Experiment 1, and their corresponding pre-mixture clips were used in Experiments 1 and 2.

For both experiments, the maximum level across experimental stimuli was set to

9 dB and 13 dB below the calibration level for Experiments 1 and 2 respectively. We maintained the relative levels of the recorded audio and the inferred sources. We excluded 16 model sounds out of 166 because they did not reach a threshold sound level and would have been inaudible over typical headphones. Pilot participants reported that a few sounds remained which were difficult to hear (presumably low-frequency sounds).

Experiment 1 procedure

For the main experiment, we split the 50 mixture sounds into two halves. For each participant, one of these splits was randomly assigned to the model condition, and the other was assigned to the recorded audio condition. On each trial, participants heard a mixture sound followed by two additional sounds that they had to choose between. Participants were instructed to select which of the two sounds was part of the initial mixture. On recorded sound trials, the correct sound was a pre-mixture sound from the mixture, and the incorrect sound was a pre-mixture sound from a different mixture randomly chosen from FUSS. We selected the incorrect option to not share a class label with any of the premixture sounds in the mixture for that trial. On model sound trials, the correct sound was a source inferred by the model from the mixture. The incorrect sound was a source inferred by the model from a mixture from the other split of mixture sounds. Participants were told that the correct answer could either be the exact sound from the mixture or a computer imitation of a sound present in the mixture. Participants could listen to the sounds multiple times, and did not receive feedback. The exact sounds chosen for the incorrect options were not varied across participants, a choice which should be investigated in future versions of this experiment.

Experiment 1 participants

60 participants were recruited through Amazon Mechanical Turk. The main experiment included 10 catch trials, which were not included in the main analysis. The catch trials comprised an independent set of recorded sound trials and were the same

across all participants. 8 participants failed the headphone check and 7 participants answered less than 7 out of 10 catch trials correctly. The data from these 15 participants was removed. We analyzed the data of the remaining participants (N=45, 30 male, 15 female, 0 non-binary based on self-report, mean age = 39.4 years, S.D. = 11.4 years).

Experiment 1 sample size

The experiment reported here was a pilot, which can be used to inform sample size for future replications.

Experiment 1 analysis/statistics

For Figure 2-8B, we computed the percent correct for each participant in each condition (recorded, model) by taking the proportion over trials. The figure reports the average and standard error over all participants (regardless of the half of the data they were assigned to). We also report the two-tailed 95% confidence interval over all participants for each condition, using the function `DescrStatsW.tconfint_mean` from the Python statsmodels module version 0.13.2. For Figure 2-8C, we computed the percent correct for each model inference by taking the proportion of correct answers across participants who completed a trial with that sound.

Experiment 2 procedure

Upon successful completion of the headphone check, participants proceeded to the experiment instructions and a set of six practice trials. Participants received feedback on the first four practice trials. The last two practice trials did not have feedback. If participants did not correctly answer the last two practice trials, they were compensated and did not proceed to the main experiment.

Each trial presented two sets of sounds: source sounds inferred by the model (“row sounds”) and pre-mixture sounds (“column sounds”), which were arranged in the headers of a grid. Participants were instructed to mark the corresponding checkbox within the grid if any part of a row sound matched part of a column sound. Participants

were told that the purpose of the task was to evaluate a computer sound-synthesis algorithm. They were warned that the computer-generated sounds were meant to imitate the column sounds but that they might not always be perfect resemblances. They were explicitly told that multiple row sounds might match to the same column sound or vice versa, or that a column sound might not have any matches (the practice trials also demonstrated these possibilities). Participants were not allowed to proceed to the next trial until they placed at least one checkmark for each row sound (see reasoning in next paragraph). Participants were allowed to listen to the sounds as many times as they wanted and in any order.

Participants performed 49 trials (plus ten catch trials - see below), corresponding to all of the mixture sounds. We excluded any model sounds and premixture sounds for which the average performance of participants in Experiment 1 did not exceed 70%. This corresponded to excluding 33 model sounds (out of 150) and 6 premixture sounds (out of 134). Premixture sounds were usually unrecognizable because they were completely masked in the mixture sound. We reasoned that only marginally recognizable model inferences were covered by the ‘unrecognizability’ error and may not be easily matched to any premixture sound. An alternative would have been to include all model sounds but allow participants to leave some without checkmarks; however, we thought this would result in lower data quality from participants skipping trials.

Experiment 2 sample size

The experiment reported here was a pilot, which can be used to inform sample size for future replications.

Experiment 2 participants

A distinct group of 10 participants were recruited through Amazon Mechanical Turk. The main experiment included 10 catch trials, which were not included in the main analysis. The catch trials had the same three audio recordings for the row and column sounds, and the correct answer was to match a sound with only itself. 2 participants

failed the headphone check and 1 participant answered less than 7 out of 10 catch trials correctly. The data from these 3 participants was removed. The remaining participants (N=7, 5 male, 2 female, 0 non-binary based on self-report, mean age = 41.4 years, S.D. = 9.5 years).

Experiment 2 analysis/statistics

Of the four deviations detailed in section 2-8, Experiment 2 was meant to measure absence, oversegmentation, and overcombination deviations. We tallied the total number of these three deviations per premixture sound for each participant. This allowed us to compute the five quantities displayed in Figure 2-8G. Bars 1-3, labeled ‘Absent’, ‘Oversegment’, and ‘Overcombine’, were computed as the proportion of premixture sounds with each of the corresponding deviation type, averaged across participants. Bar 4, labeled ‘No deviations’, was computed by finding the proportion of premixture sounds for which a participant checked only one model inference and then averaging that proportion across participants. Bar 5, labeled ‘No deviations + Most common’, is based on a subset of the proportion indicated by Bar 4. We determined which response (i.e. pattern of checkmarks across model inferences) was most common for each premixture sound. This allowed us to compute the average proportion of premixture sounds for which a participant checked only one model inference and for which that concurred with the most common response. We plotted the standard error across participants.

To derive the chance level for each of these average proportions (except overcombine deviations - see below), we randomly permuted our data 5000 times within each worker’s responses, subject to the constraint that each row contains at least one checkmark. We then computed the same statistics on the permuted datasets. We plotted the average proportion across mixture sounds, permutations and workers as well as the standard error across workers. We use a permutation test to determine statistical significance: we report the quantile of the empirical average proportions (across workers) with respect to the distribution of 5000 permuted average proportions (across workers). The proportion of overcombine deviations was an exception

because by design they occur at the same rate in the permuted and original data. Instead, we computed the two-tailed 95% confidence interval of the proportion of overcombine deviations (same method as in Experiment 1).

2.6 Open source media credits

Sounds. **Figure 2-2:** ftpalad (glass squeaking, source: Freesound), fst180081 (bee against window, source: Freesound), Bob MacGuire (olive flycatcher, source: Macaulay Library ML195787), shelbyshark (tea kettle boil then whistle, source: Freesound) **Figure 2-3:** inspectorj (hammering nails close, source: Freesound), Cornell Lab of Ornithology (white throated sparrow, source: Macaulay Library), phonosupf (trombone glissandi, source: Freesound), anonymous author (spring peeper, also known as *Pseudacris crucifer*; source: Wikipedia).

Images. **Figure 2-2:** Michael Sander (coins, source: Wikimedia), Jon Sullivan (waves, source: Wikimedia), Gary L. Clark (flycatcher, source: Wikimedia), Louise Docker (bee, source: Wikimedia), Aqua Mechanical (rubbing glass, source: Flickr), maliciousfairy (tea kettle, source: Flickr). **Figure 2-3:** robbiesaurus (trombone player, source: Wikimedia), brian.gratwicke (spring peeper, source: Wikimedia), public domain (hammering, source: Wikimedia), Cephas (sparrow singing, source: Wikimedia). **Supplementary Figure 2-9:** Ivan Radic (cute puppy panting, source: Flickr).

Chapter 3

A statistical model of material for synthesizing contact sounds

This work was first reported as a conference paper at DAFX2019. Citation: Traer, J., Cusimano, M., & McDermott, J. H. (Sep 2019). A perceptually inspired generative model of rigid-body contact sounds. The 22nd International Conference on Digital Audio Effects (DAFx-19).

Contact between rigid-body objects produces a diversity of impact and friction sounds. These sounds can be synthesized with detailed simulations of the motion, vibration and sound radiation of the objects, but such synthesis is computationally expensive and prohibitively slow for many applications. Moreover, detailed physical simulations may not be necessary for perceptually compelling synthesis; humans infer ecologically relevant causes of sound, such as material categories, but not with arbitrary precision. We present a generative model of impact sounds which summarizes the effect of physical variables on acoustic features via statistical distributions fit to empirical measurements of object acoustics. Perceptual experiments show that sampling from these distributions allows efficient synthesis of realistic impact and scraping sounds that convey material, mass, and motion.

3.1 Introduction

The sounds that enter the ear are collectively determined by the physical processes that generate the acoustic waveform. Sound generation by rigid bodies is a classic physics problem and the processes by which material parameters (e.g. material, mass, motion) affect acoustic waveforms have been well characterized (Fletcher & Rossing, 2012; Helmholtz & Ellis, 1875; Morse & Ingard, 1986; Rayleigh, 1896). Typically, physical sound synthesis is done by modelling in detail the relevant processes which lead to the generation of a sound. For example, rigid bodies are modelled as a mesh-grid of masses on springs (Cadoz, 1979; Cadoz et al., 1993; Cadoz et al., 1984; O’Brien et al., 2002; van den Doel & Pai, 1996; Zheng & James, 2011), or decomposed into small segments over which wave equations can be solved by Finite-Element or Boundary-Element-Methods (FEM/BEM) (Bilbao, 2009; James et al., 2006; Manocha & Lin, 2009). These models yield a set of resonant modes from which contact sounds can be synthesized. In practice such models require computing physical interactions at very small spatiotemporal scales, and are thus computationally expensive.

Humans perceive sounds in terms of physical variables (Gaver, 1993b; Rocchesso & Fontana, 2003), and these perceptual abilities might inform sound synthesis approaches. When we hear the sound of a fork dropped upon a wooden table, we can make judgments about the size (Carello et al., 1998; Grassi, 2005; Tucker & Brown, 2002), material (Avanzini & Rocchesso, 2001; Giordano & McAdams, 2006; Klatzky et al., 2000) and motion of the fork (Lemaitre & Heller, 2012). However, our discrimination abilities are limited. It is not clear that humans can tell a fork from a knife in such a case, for instance, let alone the detailed geometry of the fork. Indeed, perceptual experiments indicate that humans can infer broad material differences (e.g. metal vs wood) from contact sounds, but are less accurate for more precise judgments (e.g. distinguishing metal from glass) (Giordano & McAdams, 2006).

The coarse-grained nature of human material judgments suggest material perception is insensitive to mode properties within some tolerance. Exactly what tolerance

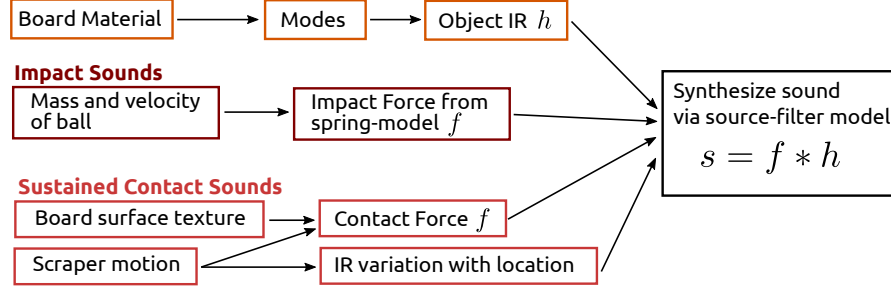
remains an open question, but it suggests that synthetic modes need not have a detailed correspondence to those of an actual object to yield compelling sounds. We hypothesize that the auditory system infers coarse-grained material parameters from statistical properties of modes, rather than their precise details. For example, consider again the sound of a fork dropped upon a table. Although fine-grained features (e.g. the thickness of the handle, the length of the tines, the narrowing of the neck, etc.) may affect individual modes, we see little evidence that humans infer such subtle features. However, coarse-grained physical features, which are crucial to inferring scene properties like material and size, will affect all the modes and thus are likely to be reflected in the modal statistics.

Rather than attempt to simulate the physical process in fine-grained detail, we measure statistics of modes from real-world impact sounds and use these distributions as the building blocks for sound synthesis via a source-filter model (in which a time-varying force is convolved with the object impulse response). We synthesize sounds from both impacts and sustained frictional forces (Fig. 3-1). As with our statistical model of modes, the impact forces are parametrized only by coarse-grained properties: mass, stiffness, and velocity. For scraping sounds, the force is generated through a texture quilting algorithm (Efros & Freeman, 2001), reflecting listeners’ perception of summary statistics as opposed to fine-grained temporal detail in sound textures (McDermott et al., 2013).

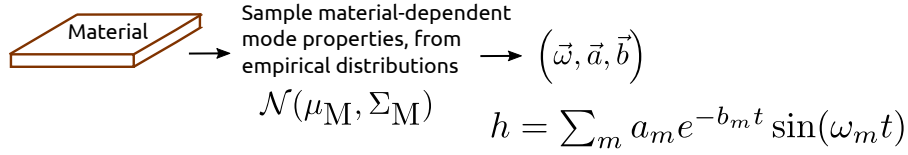
Our approach yields compelling renditions of sounds via a fast and efficient process. As with other similar approaches (Aramaki et al., 2010; Pruvost et al., 2015), it is thus ideal for use in physics engines used in modern computer games and simulations. Such engines store a set of attributes for rigid-bodies to compute how they will move (e.g. mass, elasticity, frictional coefficients, a grid-model of the geometry, etc.) and to compute their appearance under lighting (e.g. diffuse and specular reflectance profiles, visual surface statistics, etc.). As conventional sound synthesis is slow, current engines rely on memory intensive sample banks of pre-recorded or pre-computed sounds to be played on contact. Our synthesis model only requires a simple texture model and low-dimensional representations of coarse physical features, such as are

already encoded for motion and visual appearance. From these crude features and a sample bank of mode distributions (e.g. wood, metal, plastic, ceramic, etc.), our synthesis algorithm can rapidly generate a range of realistic and unique contact sounds. Here we show that impact sounds generated in this way convey mass and material to listeners as well as recordings of real sounds. Scraping sounds derived from these mode distributions are also realistic and convey motion trajectories.

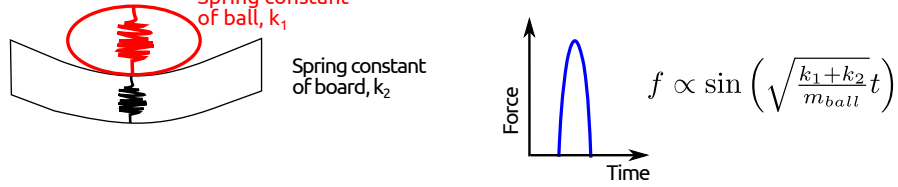
Sound Synthesis



Object Impulse Response (IR)



Impact Force



Sustained Contact Force

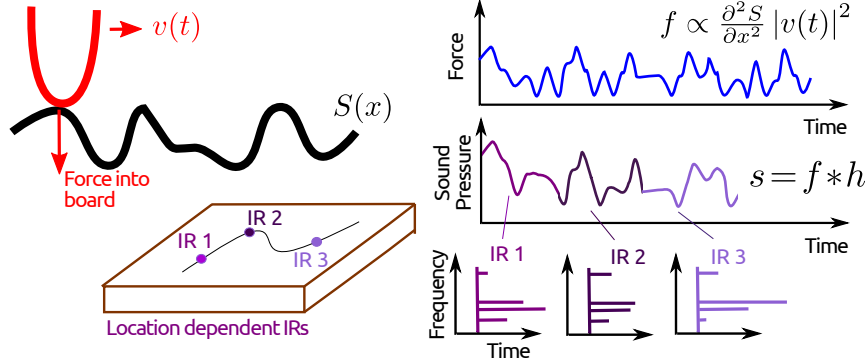


Figure 3-1: We synthesize sounds by (top) a generative model of impact and sustained contacts. (Upper-middle) Object Impulse Responses are synthesized by sampling modes from empirical distributions. (Lower-middle) Impact forces are modelled via a spring model. (Bottom) Sustained contacts are modelled via measured surface textures and location-dependent IRs.

3.2 Source-filter model of impacts

Our model is inspired by the well-known source-filter model (Conan, Derrien, et al., 2014)

$$s(t) = f(t) * [h_1(t) + h_2(t)] \quad , \quad (3.1)$$

where $s(t)$ is the sound entering a listener’s ear, $f(t)$ is the contact-force between two objects and $h_j(t)$ is the impulse response (IR) of the j th object. Past sound synthesis techniques have computed high-resolution IRs with large grid models such as finite-element or boundary-element techniques (James et al., 2006; McAdams et al., 2010; Raghuvanshi & Lin, 2006), solved analytically for the resonant modes of an object of known material and shape (Klatzky et al., 2000; Lutfi, 2008; Lutfi & Stoelinga, 2010; van den Doel & Pai, 1996), or fit parameteric models of mode parameters to measured impacts (Aramaki et al., 2010). The grid solutions are flexible but require significant computational power. The analytical modal solutions allow fast synthesis but only apply to a small subset of rigid bodies.

We approximate, as have others before (Ren et al., 2013), object IRs via the summation of a broadband transient “click” and a set of exponentially decaying sinusoids corresponding to the resonant modes of the object

$$h(t) = h_T(t) + \sum_m^M 10^{(a_m - b_m t)/20} \cos(\omega_m t) \quad , \quad (3.2)$$

where h_T is the transient, and (a_m, b_m, ω_m) are the onset power, decay rate and angular frequency of the m th mode. The transient can be described via a set of decaying noise-bands:

$$h_T(t) = \sum_n^N 10^{(\alpha_n - \beta_n t)/20} \nu_n(t) \quad , \quad (3.3)$$

where ν_n is a time-series of random noise filtered by the n th Equivalent-Rectangular-Bandwidth (ERB) filter of a cochleagram decomposition, and (α_n, β_n) are the onset power and decay rates of this channel. Under our model, an object IR can be com-

pletely described by $2N + 3M$ parameters, to precisely determine the shape of the transient and the modes. Throughout this work we use $N=30$ and $M=15$, which we found to be sufficient for compelling resynthesis.

Our preliminary experiments suggest several broad perceptual trends: (1) perception of material properties is dominated by a small number of powerful modes; (2) changes to the properties of weaker modes are barely noticeable; (3) slight changes to the most powerful modes are detectable, but the resulting sound is perceived as a different exemplar of a similar object or the the same object struck in a different location; (4) altering the transient but not the modes, has a minimal effect on perceived material. All of these perceptual trends suggest that human perception of object properties (i.e. material, size, shape) are primarily predicated upon the statistics of the most powerful object resonant modes.

3.2.1 Modal synthesis of object Impulse Responses (IRs)

To test our hypothesis that human judgments of object properties are based on mode statistics, we seek to synthesize impact sounds which match the modal statistics of real-world impacts, but are otherwise unconstrained (such that the exact mode parameters are different). We began by measuring the mode statistics from real-world objects.

To measure resonant modes, we recorded the sounds of a large number of materials being struck by small pellets. We estimated the resonant modes of each impact via an iterative procedure of spectrogram matching: (1) we obtained the frequency channel of the spectrogram of the impact sound with the maximum power; (2) we synthesized an initial synthetic impact with an exponentially decaying sinusoid at that frequency; (3) we adjusted the mode properties (frequency, onset power and decay rate) to minimize the mean-squared error between the spectrograms of the recording and the synthetic; (4) we subtracted the synthetic spectrogram from the original (removing the mode we just measured). We then repeated the procedure 14 times, yielding parameters for the 15 most powerful modes. After fitting the modes we repeat this procedure using exponentially decaying noise-bands instead of sinusoidal modes to fit

the properties of the transient.

For each material, we recorded multiple impacts at different locations on multiple objects. We pooled together modes from multiple objects and characterized the mode statistics by fitting a multivariate Gaussian distribution to the resulting collection. We similarly fit distributions to the transient decay parameters.

To generate a synthetic IR, we sample both mode and and transient properties from our empirically measured distributions.

$$\begin{aligned} (a_i, b_i, \omega_i) &\sim \mathcal{N}(\mu_M, \Sigma_M) \text{ for mode } i \\ (\vec{\alpha}, \vec{\beta}) &\sim \mathcal{N}(\mu_T, \Sigma_T) \quad , \end{aligned} \tag{3.4}$$

where (μ_M, Σ_M) are the mean and covariance of the three mode properties, conditioned upon the required object or material, and (μ_T, Σ_T) are the analogous mean and covariance of the transient subband properties. We used rejection sampling to ensure that the average frequency spacing between sampled modes was within 10% of that measured from recordings of the material. Because the mode statistics are computed offline prior to synthesis, all that needs to be encoded at time of sound synthesis are material labels which index distributions of IR properties.

To simulate multiple contacts of the same object we sample from the distributions once, and then randomly perturb mode onset powers (standard deviation=20% mean mode power) for each later impact. This emulates the fact that impacts in different locations differentially excite the same modes. We found empirically that either sampling from the distribution twice or repeating the exact same set of mode parameters produced unrealistic sounds (Lloyd et al., 2011).

3.2.2 Effect of impact physics

To synthesize an impact sound, we also need to compute the contact force, to be convolved with the object IR [Eq. (3.1)]. We approximate the contact force using a simple spring-model, in which the force acting on either object is proportional to the displacement of the surface at the point of contact. This yields the force between two

objects as a half-wavelength of a sinusoid

$$f(t) = \begin{cases} \sin\left(\sqrt{\frac{k}{m}}t\right) & 0 < t < \frac{\pi m}{k} \\ 0 & \text{otherwise} \end{cases}, \quad (3.5)$$

where v is the velocity at impact, m the mass of the pellet and k a spring constant determined by the materials of the board and ball. Note that as the mass tends to zero, the time of contact between the two materials tends to zero and the contact force tends towards a Dirac-delta function. This observation partly justifies the use of small pellet impact recordings to approximate the object impulse response. (We note that future work should also investigate alternative methods of measuring the impulse response.)

To synthesize impact sounds, we convolve a synthesized IR from Eq. (3.4) with the contact force described in Eq. (3.5). All that needs to be encoded at the time of impact are labels of object mass, velocity, and material labels, which determine both the spring constants and the distributions from which modes are sampled. Except for parameters of the mode distributions, these features are already included in physics engines.

3.3 Perception of synthetic impacts

To assess our impact synthesis model we played both recorded and synthesized sounds to listeners and asked them to judge: (1) realism; (2) material; and (3) mass of the colliding objects. All perceptual experiments were conducted over Amazon’s Mechanical Turk platform. A standardized test was used to ensure participants were wearing headphones (Woods et al., 2017).

3.3.1 Experiment 1. Realism of synthetic impact sounds

We first sought to test whether our synthetic sounds were compelling renditions of real-world impacts. If our synthesis method neglected sound features to which the

brain is sensitive, the synthetic sounds should be recognizable as fake.

Participants were presented with a pair of impact sounds and identified which was the real recording. In all trials, one sound was a real-world recording of a ball dropped on a resonant object, and one a synthetic impact generated via our model or a model that was ‘lesioned’ in some way, by omitting the transient component of the IR, or by omitting the modes from the IR. The conditions of the experiment were (1) full synthetic model; (2) Modes only, without transient; (3) Transient only, without modes; (4) Time-reversed synthetics. The sound in the final condition were clearly synthetic, which serves to ensure task comprehension.

The results (Fig. 3-2) show that listeners could not distinguish sounds from either the full or lesioned models from real-world recordings, demonstrating that our method of impact sound synthesis yields plausible sounds. The chance performance for the lesioned models presumably reflects the fact that the resulting sounds remained realistic even though the lesion altered the quality of the sounds. As participants were good at identifying the Time-Reversed sounds it is clear they understood the task. Poor performance in the other conditions thus reflects the success of the synthesis.

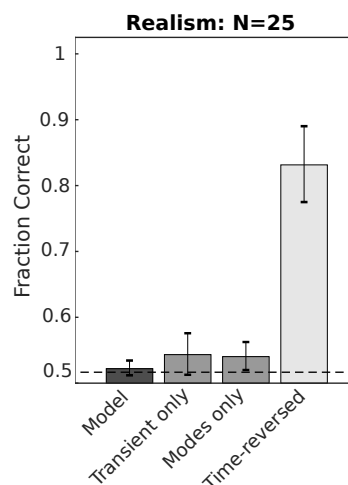


Figure 3-2: Discrimination of real vs. synthetic impact sounds (Exp 1). Dashed line denotes chance performance.

3.3.2 Experiment 2. Perception of material

Having demonstrated that our synthetic impact sounds are realistic, we sought to test whether they convey appropriate physical parameters to listeners. We first tested whether listeners can recognize the material of a struck resonant object.

Participants heard a single impact sound and were asked to identify the material of the struck object from one of four possible categories: metal, ceramic, wood or cardboard. Participants were told that the striking mallet was effectively noiseless and that many different objects of each material class were used, of a range of different sizes, shapes and sub-material (i.e. metal contained steel, tin, aluminium etc.; wood contained poplar, pine, oak etc.)

With real-world recordings, participants were excellent at distinguishing hard materials (metal or ceramic) from soft materials (wood or cardboard) but made errors within the hard or soft categories (Fig. 3-3). This result is consistent with prior studies (Giordano & McAdams, 2006). Sounds from our synthesis model - both with and without the transient - yielded a similar pattern of success and failures. Without modes, or with shortened modes, human judgments were strongly biased towards softer materials. With lengthened modes, judgments were biased towards harder materials, particularly metal. This demonstrates that our model - particularly the mode statistics - have captured the acoustical features that humans use to judge material classes from impact sounds. The correlation of the confusion matrices for the full model and recorded sounds was 0.72.

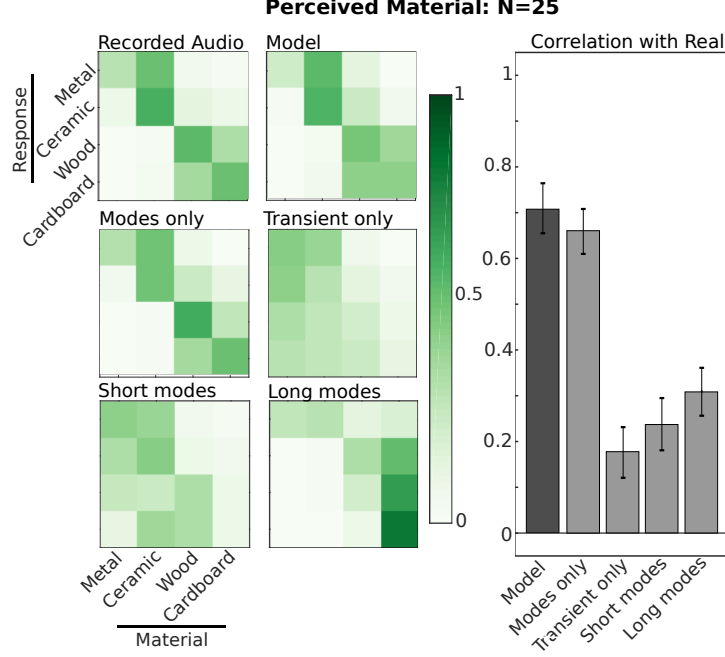


Figure 3-3: Material discrimination from synthetic impact sounds (Exp 2). Left: Confusion matrices of the presented material and participant responses. Right: Correlation of the confusion matrices of various synthetic sounds with that of the recorded impacts.

3.3.3 Experiment 3. Perception of mass

We next sought to test whether our synthetic sounds convey the mass of the striking mallet to listeners. Participants heard two impact sounds, one of a small wooden pellet (0.7 g) dropped onto an object, and one of a larger wooden ball (7.6 g) dropped onto the same object. Participants were asked to identify which of the two balls was heavier. To generate synthetic sounds the synthetic IRs were convolved with two different contact forces to emulate different ball masses, as shown in Eq. (3.5). The impact levels were not normalized but retained the relative variation in power level induced by the difference in impact force (i.e. the coefficient in Eq. (3.5) and the amplitude of the IR). All recordings and simulations were made with balls dropped from the same height (8 cm), but participants were not explicitly told this.

Since we do not know k , the spring constant, we cannot compute the contact force [Eq. (3.5)]. Instead we estimate k from the recorded impact sounds. Since both balls are the same material, we assume $k_{\text{large}} = k_{\text{small}}$, which means the ratio between

the contact times for the two balls is $m_{\text{large}}/m_{\text{small}}$. We set the contact time of the larger ball to be 10.9 times that of the smaller ball. We then iteratively adjusted the contact time of the smaller ball, until it produced a match between the average spectral centroid of the synthetic sounds and of the corresponding impact recordings.

The results (Fig. 3-4) show that humans perform very well at this task, both with real-world recordings and with synthetic sounds. This demonstrates that humans are sensitive to the filtering effect described by the contact force and can use this acoustic information to estimate the mass of the striking mallet. Participants showed a small performance decrement in the conditions where modes were shortened or excluded altogether, suggesting that humans are using modes, in addition to the sound level and spectral centroid, to estimate mass. The results suggest that our synthetic sounds convey mass as well as real-world recordings.

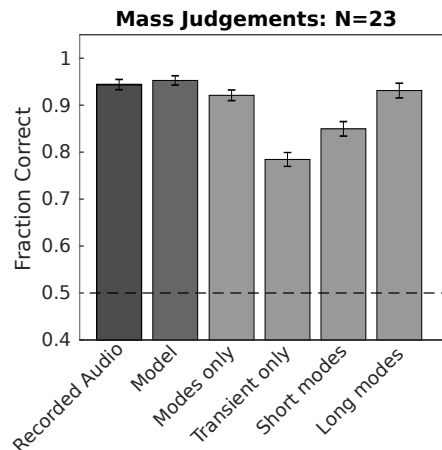


Figure 3-4: Mass discrimination with real and synthetic impact sounds (Exp 3).

3.4 Sustained contacts

To test the generalizability of our impulse response distributions, we next consider sustained contacts such as made by two objects scraping across each other. Similar to Ren et al. (2010), we again use the source-filter model of Eq. (3.1) but both the force and object IRs are more complicated than for impact sounds. The contact force $f(t)$ is generated by a series of small collisions as the scraper moves across the surface

of the scraped object, and is thus a function of the downward force applied to the scraper, the surface texture depth, and the scraper speed (Fig. 3-1, bottom). The object IR changes with scraper position $x(t)$, and thus, as the scraper moves across the surface, the IR becomes a time-varying function $h_{\text{surface}}(x(t))$. We describe these models of force and IR in more detail below. Despite the simplicity of this model, our results suggest that it yields plausible scraping sounds which convey motion of the scraper.

3.4.1 Contact force for sustained contacts

To model the force between scraper and surface we start with several simplifying assumptions: that the external force applied to the scraper F_p is constant and applied vertically downwards, and that the probe follows the surface exactly without any slip or bounce, such that the probe height $z(t)$ at time t , is given by the surface elevation $S(x)$ at the probe location x . For now we consider a transect across the surface so x is a one-dimensional variable, though the following analysis applies easily to a 2D treatment.

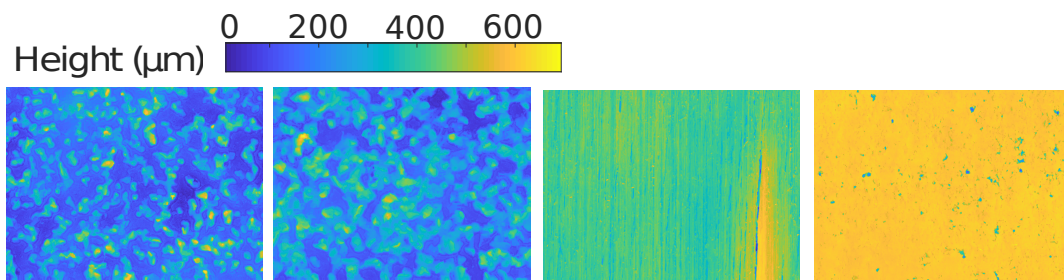


Figure 3-5: Everyday textures measured with the confocal microscope. Surface area is 7.3 mm by 10 mm. From-left: 100 grit sandpaper; 60 grit sandpaper; wood; vinyl tile.

We first consider the vertical component of the force. Under our assumptions, the change in vertical force applied to the surface can be derived from the vertical

acceleration of the probe, which, as the probe follows the surface, is given by

$$\begin{aligned} f_v(t) &= m_p \ddot{z} \\ &= m_p \frac{\partial^2 S}{\partial x^2} |v(t)|^2 \quad , \end{aligned} \quad (3.6)$$

where m_p is the mass of the probe and $v(t)$ is the horizontal velocity of the probe.

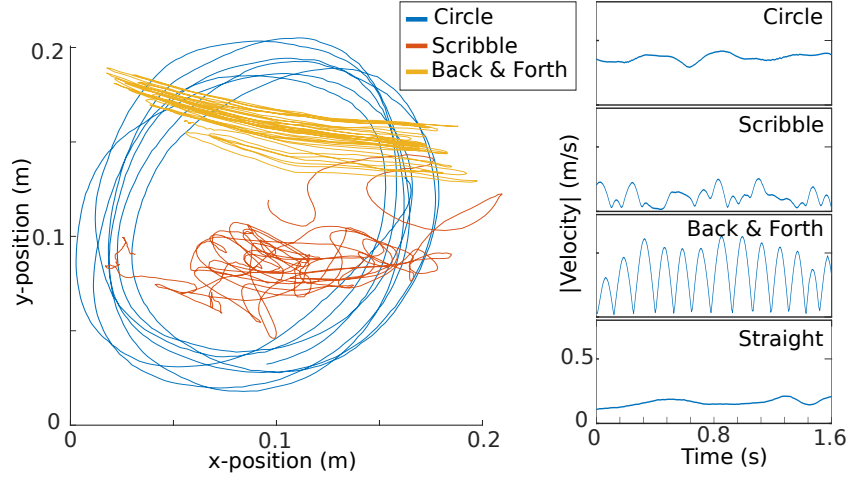


Figure 3-6: Scraping motions. Left: Measured position traces of scraper over surface, for three different types of motion. Right: Absolute velocity measurements.

We next consider the frictional force tangential to the surface. We model this as proportional to the probe speed raised to the power of an exponential factor γ , giving

$$f_h(t) \propto \left(\left| v(t) \frac{\partial S}{\partial x} \right| \right)^\gamma \quad , \quad (3.7)$$

where the partial derivative with respect to x accounts for the difference between the speed of scraping across the surface, which is the important factor, and the horizontal speed $|v(t)|$. Assuming that these forces are independent acoustic processes which give rise to waveforms that add incoherently, the total force imparted onto the object is then given by

$$\begin{aligned} f(t) &= f_v + f_h \\ &= m_p \frac{\partial^2 S}{\partial x^2} v(t)^2 + A \left(\left| v(t) \frac{\partial S}{\partial x} \right| \right)^\gamma \quad , \end{aligned} \quad (3.8)$$

where A and γ are unknown constants which titrate the importance of shear friction versus vertical forcing. We explore the role of these factors by listening to synthetic scrape sounds from a range of values. We have neglected the constant downward force term F_p which, though present, does not create any sound.

To obtain S , we measured the surface texture of several real objects using a scanning confocal microscope (Keyence VK-X260K). In these experiments, we used a micro-scale depth map of a small section of a wood block (Fig. 3-5). These are relatively small matrices (1600 pixels by 2300 pixels), which render the surface with horizontal resolution of $5.6\ \mu\text{m}$ and vertical resolution of $0.1\ \text{nm}$. Based on perceptual results concerning auditory texture perception, we expect that the perceptually important properties of such textures are statistical (McDermott & Simoncelli, 2011). Therefore, to define S , we use one-dimensional quilting to generate a texture from a measured depth map (Efros & Freeman, 2001), sampling a series of single rows and concatenating them. In future work, we plan to synthesize these surfaces statistically from coarse-level variables, in the same spirit as our distribution over impulse responses.

In addition to a depth map, the synthetic scraping force requires ecologically plausible velocity profiles of scraping motions. To probe the mechanics of typical human scraping movements, we measured the velocity and position profiles of several scraping movements using an optical tracking system (OptiTrack V120:Trio; Fig. 3-6). We use these recorded trajectories in the reported synthesis. However, in informal experiments, we found that the quality of sound synthesis was not heavily dependent on a precise match to the recorded data. Future work will include simple statistical models of these trajectories.

3.4.2 Variation of IRs over contact location

Object IRs depend upon the location being struck, and thus to simulate scraping we model this variation of modal properties with probe location. To informally assess the variability of mode properties as a function of impact location, we compared impact recordings we had made with different strike locations and found the variation in

mode properties to be moderate. To emulate such changes with synthetic IRs, we synthesized a single canonical IR from our model [Eq. (3.4)], with properties $(\mathbf{a}_o, \mathbf{b}, \omega)$, and simulated a number of location specific IRs by adding some noise to the mode powers

$$(\vec{a}, \vec{b}, \vec{\omega}) = (\vec{a}_o + \vec{\epsilon}, \vec{b}, \vec{\omega}) \quad , \quad (3.9)$$

where \vec{a}_o is the original vector of mode powers sampled from our model and $\vec{\epsilon}$ is a Gaussian noise vector. This gives a set of IRs with similar but varying modes, which crudely emulate an object of arbitrary shape struck in various locations.

We assign these sampled IRs to points along a motion trajectory, and interpolated between them in waveform space to give a smoothly varying surface IR, $h_{\text{surface}}(x(t))$. When the scraper was at a position between the defined centerpoints, the impulse response was a linear combination of the impulse responses with weights proportional to the relative distances from the scraper to the centerpoints. We ignore the contribution of the scraper to the impulse response, assuming that it is damped by the hand in which it is held.

3.5 Perception of synthetic scraping

To assess the efficacy of our scraping synthesis model, we played both recorded and synthesized sounds to listeners and asked them to judge: (1) realism; and (2) the shape of the scraper’s position trajectory. As in section 3.3, all experiments were conducted online using Amazon’s Mechanical Turk platform, and a standardized test was used to ensure participants were wearing headphones (Woods et al., 2017). In each experiment, in addition to testing lesioned forms of our own synthesis model, we compare our model to the one other scraping synthesis method that we are aware has been tested psychophysically (Thoret et al., 2014). Thoret et al. generated low-pass filtered white noise whose amplitude and filter cutoff increased with increasing velocity, and showed that several motion trajectories could be accurately judged from

the resulting sounds.

3.5.1 Experiment 5. Realism of synthetic scraping sounds

Participants were played a pair of scraping sounds and asked to identify which was the real recording. In all trials, one sound was a real-world recording of chopstick scraping a board, and one a synthetic scrape generated via our model or a lesioned version thereof. The synthetic conditions of the experiment were generated via (1) the full model, using measurement-based surface textures and varied IRs; (2) measured depth map and just a single IR; (3) pink noise depth map and varied IRs; (4) white noise with varied filter cutoff from (Thoret et al., 2014); and (5) velocity-gated white noise, which is silent when the chopstick moves more slowly than a threshold, but otherwise constant. Condition (5) is clearly synthetic and serves to ensure the participants understand the task.

The results (Fig. 3-7) show that the full synthesis model, though not perfectly realistic, frequently fools listeners. However, using a time-varying impulse response does not improve realism over filtering with a single synthetic impulse response. A synthetic noise depth map also produced comparably realistic sounds. Our sounds were less obviously synthetic than those of Thoret et al. (2014), but one caveat is that the comparison recordings were produced by a narrow scraping probe. We suspect that condition (4), with its flat broadband spectrum, may be more appropriate for modeling scrapes produced by heavier objects with large contact surface area (e.g. pushing a heavy box over tile). The gated white-noise is easily recognized as synthetic by the participants, demonstrating that they understood the task.

3.5.2 Experiment 6. Perception of motion

Participants were presented with a single scraping sound and asked to choose the scraping trajectory from four choices: "circular", "back-and-forth", "scribble", or "straight". Participants heard both real-world recordings and synthetic sounds derived from a real-world motion. The motion trajectories used to generate synthetic

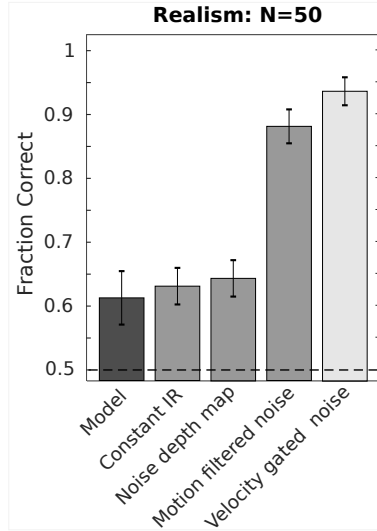


Figure 3-7: Discrimination of real vs. synthetic impact sounds scraping sounds (Exp 5). Dashed line indicates chance performance.

scrapes were matched in speed to the scrapers used to make the real-world recordings.

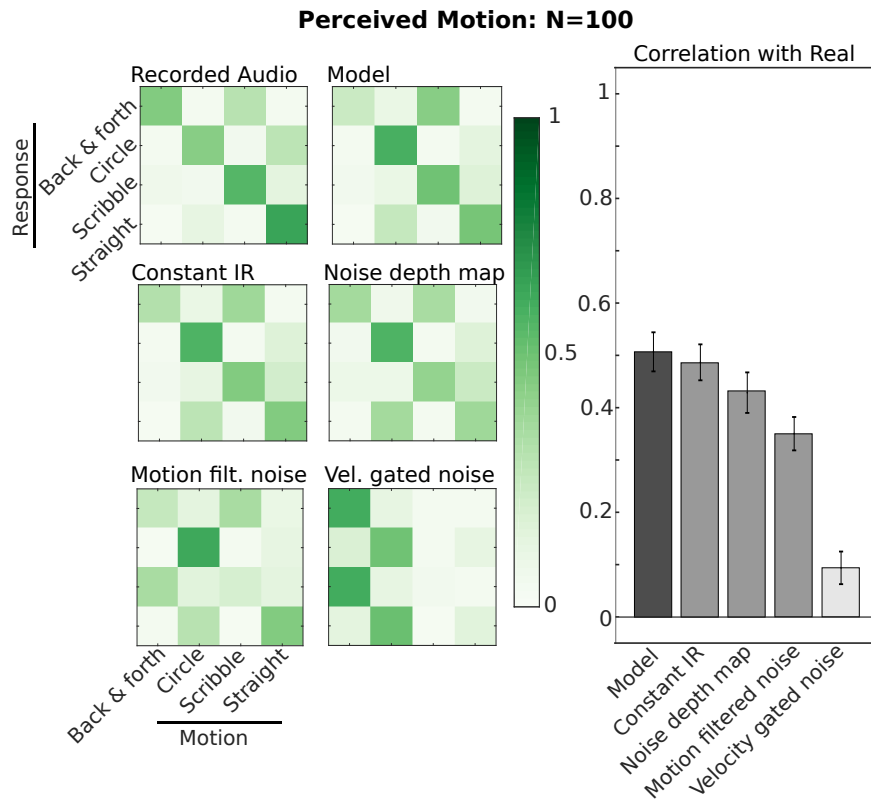


Figure 3-8: Motion discrimination from synthetic scrape sounds (Exp 6). (Left) Confusion matrices of presented motion pattern and the human responses. (Right) Correlations of the confusion matrices of synthetic sounds with the correlation matrix of recorded sounds.

As shown in Fig. 3-8, motion judgments for synthetic scrapes were similar to those for real-world scrape recordings. In both cases participants were correct most of the time, but misjudged "straight" motions to be "circular", both of which have velocity profiles without zero points. When judging either "back-and-forth" or "scribble" sounds, the full model and its lesioned variants led to more "scribble" judgments. This result could reflect the greater scattering of contact position around the surface in scribbling compared to other motions. Although we attempted to simulate this positional change with changing IRs, the full model and the constant IR model were comparable for both realism and motion, suggesting that we did not successfully capture this informative spatial variation.

3.6 Discussion

Our synthesis model is fast because it only models the effects of a small number of physical variables (material, mass, velocity etc.). It is evident from daily life that humans can infer more than just the variables we have described from contact sounds. Impact sounds contain cues to shape, size, and hollowness, as well as to the environmental reverberation (Traer & McDermott, 2016). Some physical variables explored in our impact model can also be conveyed by frictional sounds (e.g. material) but this remains to be explored in future work. Furthermore, friction sounds are not limited to scraping, but rather include other interactions such as rubbing, brushing and sliding. Future investigations into how these interactions produce sound, and into human sensitivity to their properties, will hopefully suggest extensions to a better and more nuanced synthesis algorithm.

The current version of our synthesis model requires some physical measurements of real-world objects: statistical distributions of object IRs conditioned upon material parameters; and surface structures. In contrast to the fully statistical approach in Cavaco and Lewicki (2007), we incorporate the IR distributions into physics-based audio synthesis of impacts (Gaver, 1993a; Ren et al., 2013). These distributions are not reported here but the dataset is currently being expanded upon to include additional

objects and materials in preparation for a future journal submission. We also hope to be able to synthesize these intermediate representations from physical variables. Our impact experiments with altered IRs demonstrated that lengthening or shortening the resonant modes caused listeners to rate the synthetic materials as “harder” or “softer” materials, consistent with physical models (Giordano & McAdams, 2006; Klatzky et al., 2000; Lutfi, 2008), but did not diminish their realism. This suggests that we should be able to synthesize IRs for novel objects without having to measure them first, permitting sound synthesis for a much larger range of objects. Similar generalizations should be possible for the forcing functions used to generate scraping sounds. As with perception of acoustic textures (McDermott et al., 2013), it is likely that humans are insensitive to the fine-grained temporal details of the contact force we use to synthesize scrapes. Presumably we can synthesize such a contact force directly from a texture model (McDermott & Simoncelli, 2011), enabling sound synthesis for a wider and more diverse range of objects without costly and time-consuming measurements.

Our impulse response model, while derived from statistics of impact sounds, can successfully contribute to the synthesis of relatively realistic scraping sounds. However, it appears that this model does not accurately capture the spatial covariance between impulse responses over a surface. Our full model and lesioned model with a single IR perform equally well, and neither are yet on par with real recordings, both in terms of realism and in the conveyed motion (Fig. 3-7, Fig. 3-8). Future investigations will include measurements and modelling of this variation in impulse responses based on position, as well as comparing modes measured from scraping sounds with those from impacts. The other component of the scraping synthesis is an excitation force based on quilted textures of measured depth maps. Several authors have treated scraping as a noisy source paired with a modal filter. Some model the friction force as $1/f^\beta$ noise (Conan, Thoret, et al., 2014; Ren et al., 2010; van den Doel et al., 2001), while others use a statistical model of densely-packed impact events (Lagrange et al., 2010). In the experiments explored here, the utilization of real-world measurements did not improve realism or motion inference. However, it remains possible that constraining a more sophisticated model of surface texture with these measurements

could be useful, particularly in judgments of material and surface roughness. Last, modeling the scraping excitation in more detail (for instance by including the scraper radius) may result in increased fidelity and expressivity of the synthesized sounds.

The model we have presented is similar in some respects to that of Conan et al. (Conan, Derrien, et al., 2014), who used statistics of contact forces to synthesize rolling sounds. We also utilize a statistical approach, but model the sounds of impacts and scraping, using statistics of the resonant modes of objects. We also found that we could use a linear model for contact forces. By contrast, Conan et al. found that a non-linearity in impact force (namely that the duration of impact should change with impact force) was required to induce realistic rolling sounds. In Chapter 4, we investigate whether there are perceptual benefits to sound synthesis with a more detailed model of impact forces for sustained contact sounds.

3.7 Conclusion

We have presented a fast and efficient method for synthesis of contact sounds - inspired by both physics and perception. The method generates object IRs by sampling resonant modes from distributions fitted to empirical measurements from example impact sounds. The method then convolves the IRs with contact force simulated with a simple physics model of either impacts or sustained scrapes. Despite the simplicity of the model, perceptual listening tasks demonstrate that the synthetic sounds are realistic and convey basic physical information as well as recorded sounds. These results suggest that our model has captured many of the acoustic features that matter for perception of physical contact sounds, despite neglecting a great deal of physical information about the sound sources.

Chapter 4

Synthesizing sustained contact sounds

This work was originally reported as a conference paper at DAFX20in21. Citation: Agarwal, V., Cusimano, M., Traer, J., & McDermott, J. H. (Sep 2021). Object-based synthesis of scraping and rolling sounds based on non-linear physical constraints. The 24th International Conference on Digital Audio Effects (DAFx-21).

Sustained contact interactions like scraping and rolling produce a wide variety of sounds. Previous studies have explored ways to synthesize these sounds efficiently and intuitively but could not fully mimic the rich structure of real instances of these sounds. We present a novel source-filter model for realistic synthesis of scraping and rolling sounds with physically and perceptually relevant controllable parameters constrained by principles of mechanics. Key features of our model include non-linearities to constrain the contact force, naturalistic normal force variation for different motions, and a method for morphing impulse responses within a material to achieve location-dependence. Perceptual experiments show that the presented model is able to synthesize realistic scraping and rolling sounds while conveying physical information similar to that in recorded sounds.

4.1 Introduction

Collisions, scraping, and rolling are commonplace in daily life, and the sounds they produce convey information about physical events in the world. Often this informa-

tion is uniquely available via sound. For instance, the visual evidence that an object is in contact with another is often ambiguous, but objects in contact will make sound if they move.

The synthesis of contact sounds is accordingly important for a wide range of applications including virtual reality, game engines, auditory displays, and the training of machine perception systems. Such synthesis requires mapping physical variables (object materials, shapes, and motions) to sound. These generative models of sound could also contribute to theories of perception, which involves inferring physical causes from sound (Gaver, 1993b), potentially by inverting internal generative models (Kersten & Yuille, 2003).

Contact sound synthesis is possible via detailed but expensive physical simulations (Bilbao, 2009; Cadoz et al., 1993; James et al., 2006; Manocha & Lin, 2009; O’Brien et al., 2002; Raghuvanshi & Lin, 2006; Zheng & James, 2011). However, many applications require more efficient synthesis, as might be enabled by lower-dimensional characterizations of objects. Previous methods have either used signal-based heuristics (Conan, Thoret, et al., 2014; van den Doel et al., 2001) or physics-inspired intuitive modelling (Avanzini et al., 2005; Reiss et al., 2019; Serafin, 2004; Stoelinga, 2007). These models are able to convey some intended physical parameters, but thus far have not been fully perceptually convincing.

In the previous chapter, we introduced a method to synthesize impact sounds by characterizing material-specific impulse response distributions, sampling from them, and convolving them with simple spring-based models of impact forces (Traer et al., 2019). In that paper, we extended the method to synthesize scraping sounds using force derived from a surface depth map, but the sounds produced were not fully realistic, sounding excessively rough and not producing a fully accurate sense of motion.

Here we introduce a new method to synthesize scraping and rolling sounds from object properties. The key innovations are a nonlinearity in the calculation of the contact force from the surface depth map and the use of normal forces that more accurately reflect typical scraping motions. These are combined with a method for

generating more realistic location-dependent impulse responses via morphing within the same material, similar to that proposed in (Pruvost et al., 2015), and a periodic force produced by rolling objects due to misalignment of the center-of-mass with the geometric center (Rath & Rocchesso, 2005; van den Doel et al., 2001). We term the approach ‘object-based’, as it relies on perceptually-relevant macroscopic properties of objects and their motions.

The resulting sounds are substantially more realistic and recognizable than those from previous methods. We first describe and evaluate our scraping synthesis. We then show that the scraping synthesis can be extended to generate realistic rolling sounds. You can listen to demos of sounds from all our experiments through my thesis webpage (<https://mcdermottlab.mit.edu/mcusi/thesis/>).

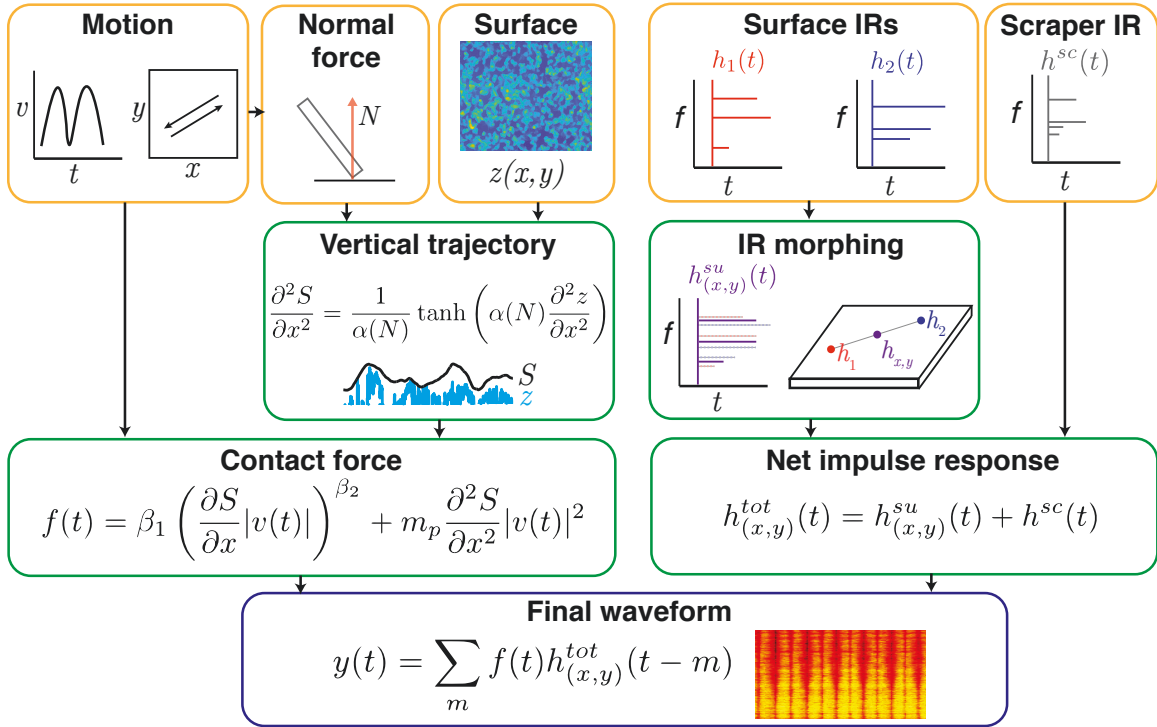


Figure 4-1: Complete scraping synthesis model. Yellow boxes: inputs to model. Green boxes: intermediate representations computed from inputs. Blue box: sound waveform computed from contact force and net impulse response.

4.2 Scraping Sound Synthesis Model

We use a source-filter approach depicted in Figure 4-1. Similar to (Conan, Thoret, et al., 2014; Ren et al., 2010; Traer et al., 2019), we generate a sound waveform by the convolution of a contact force function and a net impulse response (IR). The contact force $f(t)$ is obtained by modelling the microscopic trajectory of the scraper on the scraped surface. We use a weighted sum of the impulse responses of both the scraper and the scraped surface to get the net impulse response of the interaction. To reflect the variation in object resonances with position, the impulse response of the surface changes continuously over the motion. Thus, the overall impulse response is a function of time. We describe the force, impulse response and audio waveform calculation in more detail below.

4.2.1 Contact force for scraping

The contact between scraper and surface causes the two bodies to undergo continuous micro-collisions. Previous authors have sampled the friction force as $1/f^\beta$ noise (Ren et al., 2010; van den Doel et al., 2001) or from a distribution over concentrated impact events (Conan, Thoret, et al., 2014; Lagrange et al., 2010; Lee et al., 2010). In our model, the only major assumption is that the scraper follows the surface without leaving the surface (e.g., which could in reality occur due to inertia or bouncing). We also assume that the masses of the scraper and surface do not change over the course of the motion.

The microscopic interaction force is modelled in two parts - a vertical force f_v (from pressing the scraper down) and a horizontal force f_h (from horizontal micro-collisions with surface asperities). We also calculate an additional macroscopic normal force (§4.2.3), which modulates the microscopic interaction force.

The vertical force f_v is estimated as:

$$f_v(t) = m_p \ddot{S}(x, y) \quad (4.1)$$

$$= m_p \left(\frac{\partial^2 S(x, y)}{\partial x^2} |v_x(t)|^2 + \frac{\partial^2 S(x, y)}{\partial y^2} |v_y(t)|^2 \right) \quad (4.2)$$

where m_p is the mass of the scraper, $S(x, y)$ is the scraper's vertical trajectory over the surface, and $v_x(t)$ and $v_y(t)$ are the velocities of the scraper in the x and y directions respectively (which define the plane of the surface).

We assume the horizontal contact force f_h to originate from the horizontal collisions of the scraper and steep asperities of the surface texture. Based on our previous work (Traer et al., 2019), we assume the following relation:

$$f_h(t) = \beta_1 \left| v_x(t) \frac{\partial S(x, y)}{\partial x} + v_y(t) \frac{\partial S(x, y)}{\partial y} \right|^{\beta_2} \quad (4.3)$$

where the partial derivatives with respect to the position (x, y) reflect the slopes of the surface texture and determine the intensity of the micro-impacts. The velocities v_x and v_y determine the density of these impacts in time. β is a free parameter that we set to 1.

Since the waveforms due to the horizontal and vertical force components are assumed to add incoherently (i.e., as arising from independent acoustic processes), the total force is given as a sum, $f_{tot}(t) = f_v + f_h$. We note that $f_{tot}(t)$ is the interaction force at the surface and acts on both the scraper and the surface as an action-reaction pair.

4.2.2 Trajectory of the scraping object

One of the most critical components of our model is the vertical trajectory $S(x, y)$ of the scraper near the surface. Most previous models have assumed the scraper's vertical trajectory to be the same as the surface depth profile $z(x, y)$ (Conan, Thoret, et al., 2014; Traer et al., 2019). In our model, we use $z(x, y)$ to derive the estimated vertical trajectory of the scraper but additionally incorporate nonlinear physical constraints

on the scraper-surface interaction. As in previous work, we assume the scraper to be in contact with the surface at a single point.

We obtained surface depth profiles for several everyday materials using a scanning confocal microscope (Keyence VK-X260K, horizontal resolution of $5.6 \mu\text{m}$, vertical resolution of 0.1 nm) (Traer et al., 2019). In the future, we plan to synthesize surface depth maps as textures using relevant texture statistics of measured depth profiles (McDermott & Simoncelli, 2011).

The assumption that the scraper exactly follows the surface depth profile z is physically unrealistic for two reasons: 1) the forces involved would be exceptionally high near the locations on the surface where the slopes change rapidly, and 2) both the surface and scraper materials would have to be abnormally elastic.

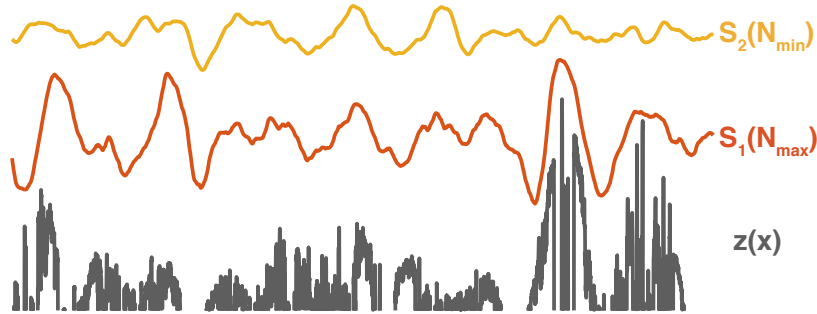


Figure 4-2: Trajectory S of the scraper point. The trajectory is determined by the surface depth profile ($z(x)$, shown in grey) and the normal force N . Larger normal forces (red) produce more extreme scraper trajectories than smaller normal forces (yellow).

To solve these issues, we chose to constrain the allowed curvatures of the scraper trajectory $S(x, y)$ to limits that are physically realizable given bounded applied force. The applied force bounds the curvature of the turns that the scraper can take over the surface asperities, limiting its acceleration and hence the second derivative of S (Figure 4-2). To achieve this, we made a heuristic decision to use a tanh non-linearity with limits dictated by the elasticity of the material and the macroscopic normal force

N .

$$\frac{\partial^2 S(x, y)}{\partial x^2} = \frac{1}{\alpha_x(N)} \tanh \left(\alpha_x(N) \frac{\partial^2 z(x, y)}{\partial x^2} \right) \quad (4.4)$$

$$\frac{\partial^2 S(x, y)}{\partial y^2} = \frac{1}{\alpha_y(N)} \tanh \left(\alpha_y(N) \frac{\partial^2 z(x, y)}{\partial y^2} \right) \quad (4.5)$$

α_x and α_y change the limits of the non-linearity depending on the normal force:

$$\alpha(N) = (1 - \nu)\alpha_{max} + \nu\alpha_{min} \quad (4.6)$$

$$\nu = \left(\frac{N - N_{min}}{N_{max} - N_{min}} \right)^\zeta \quad (4.7)$$

N_{max} and N_{min} are determined by the variation in normal force described in the next section; for constant N we substitute fixed values for α_x and α_y . The current synthesis process is dependent on a judicious choice of these parameters. In the future, we aim to estimate these parameters from the applied vertical force and hardness of the material.

We note that this procedure is not equivalent to low-pass filtering. Rather, it is a form of soft clipping which introduces high-frequency components. Therefore, we apply a Gaussian moving average following the non-linearity to remove these high-frequency components. The window size in each dimension was proportional to α_x and α_y , with an average half window size of 5 samples at a sampling rate of 44.1kHz. We note that future work should investigate the effects of the non-infinitesimal surface area of contact.

4.2.3 Effect of macroscopic normal force

The applied force dictates how closely the scraper tracks the surface, which in our model is reflected in the force-dependent nonlinearity parameters (α_x, α_y) that influence the modulation of the normal force (Figure 4-2). The applied force can vary over the scraping trajectory, in part due to constraints on how the human hand applies forces. For example, consider back-and-forth scraping with a cylindrical object held

by hand (Figure 4-3). This action can be approximated as simple harmonic motion with an angular frequency ω . We calculate the resulting normal force assuming that 1) the applied force is along the length of the cylinder 2) the force is proportional to the horizontal acceleration of the cylinder, and 3) the cylinder is at an angle θ from the horizontal. As a result, the force is maximal when close to the torso and minimal when furthest away. The normal forces at both ends of the trajectory are then given by:

$$N_{max} = \frac{mg + \omega^2 mL \tan \theta}{1 - \mu \tan \theta} \quad (4.8)$$

$$N_{min} = \frac{mg - \omega^2 mL \tan \theta}{1 - \mu \tan \theta} \quad (4.9)$$

This normal variation is characteristic of back-and-forth scraping motion (e.g. from sanding or scrubbing). The normal force variation influences the scraping sound via α_x and α_y as described in the previous section. For each motion trajectory, we similarly modelled the normal force variation assuming a scraping cylinder held at a fixed angle to the surface.

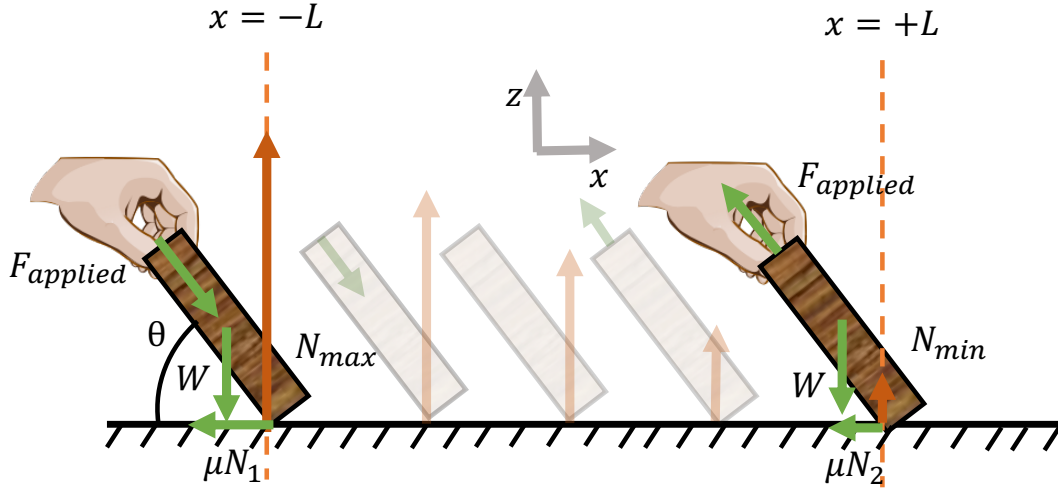


Figure 4-3: Macroscopic forces on the scraper when undergoing human-produced simple harmonic motion (as when scraping a held object back and forth on a surface). The normal force (red vector) varies with position due to the variation in the applied force $F_{applied}$.

4.2.4 Morphing impulse responses

The impulse responses on the surface depend on the location on the surface currently being excited. Based on informal observations of recordings of scraping sounds, we found that the modal frequencies and their powers varied smoothly with location. To generate an impulse response for an arbitrary surface location, we thus morph from one measured impulse response to another by smoothly interpolating between both the modal frequencies and powers (similar to Pruvost et al., 2015), in contrast to our previous work where we cross-faded impulse responses (Traer et al., 2019).

We used impulse responses measured in previous work (Traer et al., 2019) at different locations on the surfaces of objects. We then extracted the mode parameters of the 50 strongest modes using sinusoidal modelling, giving us the frequency f_i of each mode and its time-dependent mode power $A_i(t)$. To obtain the mode parameters for an impulse response at a location intermediate between two measured impulse responses we logarithmically interpolated between the measured modes in both frequency and amplitude. The impulse response for the location was then synthesized as a superposition of decaying sinusoids using these morphed parameters.

As the scraper applies a force on the surface, the surface applies an equal and opposite reaction force on the scraper. If the scraper is not damped, it also contributes to the overall sound because of the effect of this forcing on the scraper’s resonances. To account for this, we sum the synthetic impulse response calculated using morphed parameters with the measured impulse response of the scraping object weighted by η , which determines the relative amplitude of the surface and scraper impulse responses.

Since each position on the surface has a different impulse response, and because the scraper moves continuously, the traditional linear view of the sound waveform being the result of a convolution of force and a fixed impulse response does not hold. To accommodate a different impulse response at each time instant, we calculate the waveform $y(t)$ using a modified convolution where the impulse response is a function

of time:

$$h_{x(t),y(t)}^{su}(\tau) = \begin{cases} \sum_{i=1}^{50} A_i(\tau) \sin(2\pi f_i \tau) & 0 \leq \tau \leq t_0 \\ 0 & \text{otherwise} \end{cases} \quad (4.10)$$

$$f(t) = \begin{cases} f_{total}(t) & 0 \leq t \leq t_1 \\ 0 & \text{otherwise} \end{cases} \quad (4.11)$$

$$y(t) = \begin{cases} \sum_{m=1}^t h_{x(t),y(t)}^{tot}[t-m]f[m] & 0 \leq t \leq t_0 \\ \sum_{m=1}^{t_0} h_{x(t),y(t)}^{tot}[t_0-m]f[t+m-t_0] & t_0 \leq t \leq t_1 \\ \sum_{m=1}^{t_1+t_0-t} h_{x(t),y(t)}^{tot}[t_0-m]f[t+m-t_0] & t_1 \leq t \leq t_1+t_0 \end{cases} \quad (4.12)$$

where $h_{x(t),y(t)}^{tot}(t)$ is the impulse response (with duration t_0) at time t when the scraper is at $(x(t), y(t))$, and f_{total} is the total scraping force near the surface (with duration t_1).

4.3 Perception of Synthetic Scraping

To assess whether our synthesis model produces perceptually compelling scraping sounds, we conducted two psychophysical experiments in which listeners made judgments about synthesized scraping sounds. We compared our synthesis model to several ‘lesioned’ versions of the model, in which we omitted model components in order to test their perceptual relevance. For baselines, we also compared our model to our implementations of previous synthesis methods based on (1) physical synthesis (Traer et al., 2019), (2) signal manipulation (Conan, Thoret, et al., 2014), and (3) minimalist modulated noise that has previously been shown to elicit accurate motion judgments (Thoret et al., 2014). These experiments were conducted online using Amazon’s Mechanical Turk platform, using a standardized test to verify that participants were wearing headphones (Woods et al., 2017). We have previously found that online participants can perform about as well as in-laboratory participants (McPherson &

McDermott, 2020; McWalter & McDermott, 2019; Woods & McDermott, 2018) provided basic steps are taken to maximize the chances of reasonable sound presentation by testing for earphone/headphone use and to ensure compliance with instructions.

4.3.1 Experiment 1. Realism of synthetic scraping sounds

We first tested whether our model could convincingly render descriptions of physical scraping events, by asking listeners to rate the realism of synthetic scrapes. Participants were presented with a text description of the scraping event, which specified the material of the scraper, its motion, and the material of the surface (Fig. 4-4A, top). The scraper material was either PVC or poplar wood and the surface material was basswood, poplar wood, or ceramic. The motion was one of five variants: slow back-and-forth, fast back-and-forth, circular, short single scrape, and long single scrape. Each listener evaluated a total of 30 trials comprising a fully-crossed set of these parameters (2 scraper materials x 3 surface materials x 5 motions). For each description, listeners rated the realism of seven synthetic sounds, each generated with a different method. The methods were (1) the full model, (2) a lesioned model using only the surface impulse response ($\eta = 0$), (3) a lesioned model lacking variation in the normal force (omitting §4.2.3) (4) a lesioned model without a non-linearity to constrain the trajectory curvatures (omitting §4.2.2), (5-7) the baseline models. All seven sounds were presented in a single trial using a MUSHRA paradigm (Fig. 4-4A). The order of the sounds was randomized on each trial. We did not include audio recordings in this experiment because they typically have distinguishing features that do not relate to scraping (e.g. room noise).

The results (Fig. 4-4B) show that our model produces scraping sounds that are more realistic than previous scraping synthesis methods. The non-linearity lesion caused the biggest decrease in the realism. The ‘Only surface IR’ and normal force lesions produced relatively small impairments of the realism, potentially because there exist alternative realistic physical explanations of those sounds (e.g. damping of the scraper).

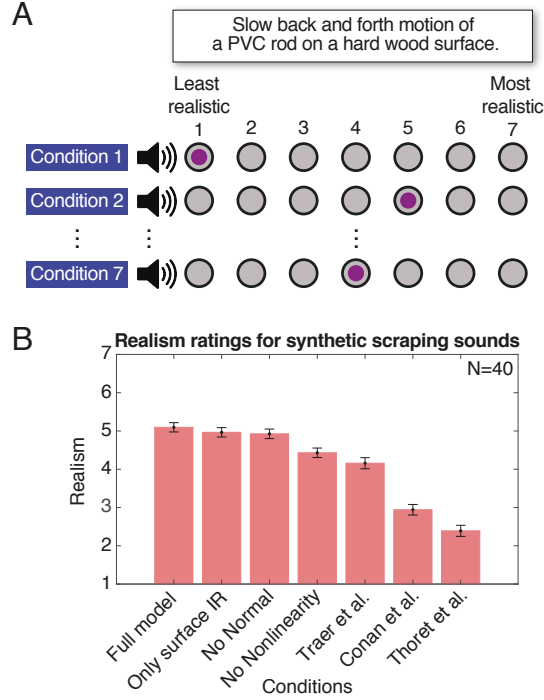


Figure 4-4: Experiment 1: Realism of synthesized scraping. A) Participants rated the realism of 7 different renderings of an object scraped back-and-forth over a surface, using a MUSHRA paradigm. Each rendering was via a different synthesis method. Participants were given a text description of the scraping event. B) Results of Experiment 1, showing mean realism for each synthesis method. Error bars plot SEM.

4.3.2 Experiment 2. Perception of motion from scraping sounds

We next asked whether participants could infer the motion that produced the scrape. In addition to testing whether the synthesis conveys an important physical parameter, this experiment provides a more fine-grained examination of the effect of the various lesions as it allows analysis of confusions between different motions.

Listeners were presented with a single scraping sound and then asked to choose the most probable path (Fig. 4-5A) traced by the object from five options: 1) a scribble, 2) four scrapes in a straight line, 3) four back-and-forth scrapes, 4) a single long line, or 5) a circle (Fig. 4-5B). The scribble motion which was obtained by optical tracking (Traer et al., 2019) and the other motions were modelled using ideal trajectories. We chose these motions because they provide pairs which share a velocity profile, but differ spatially: back-and-forth and four-in-one line, and the circle and single

line. Based on previous literature (Thoret et al., 2014) we expected that listeners would make mistakes within these pairs but would make accurate judgments between them. We tested the seven synthesis conditions of Experiment 1 as well as real audio recordings.

As expected, the confusion matrices (Fig. 4-5C) show that listeners experienced confusions when listening to recorded as well as synthesized audio. However, the confusion patterns for our full model are more similar to those of recorded audio than were the other synthesis methods. We quantified this similarity between the confusion matrix for recorded audio $C_{recorded}$ and each synthesis method $C_{synth,i}$ as follows:

$$Similarity(i) = 1 - \frac{\|C_{recorded} - C_{synth,i}\|_2}{\|C_{recorded}\|_2} \quad (4.13)$$

As shown in Figure 4-5D, the full model achieves the highest similarity to listeners’ judgments on recorded audio. The ‘Only surface IR’ model performs similarly to the full model, while the normal force and non-linearity lesions greatly decrease the similarity. These results indicate that capturing naturalistic variation in the normal force and constraining the trajectory curvature are important for the perceptual estimation of motion from scraping.

4.4 Rolling Sound Synthesis Model

To synthesize rolling sounds, we use a similar source-filter model where the filter changes with time due to the change in the position of the rolling object, but with a contact force that differs from that for scraping. Pure rolling is a unique form of sustained contact where the surface is excited without the point of contact on the object being in relative motion with the surface. Previous efforts to synthesize rolling sounds have proposed a force term arising from the offset of the object’s centre of mass from its geometric center (due to slight deviations from perfect sphericity). We hypothesized that there would be an additional contribution from a scraping-like term

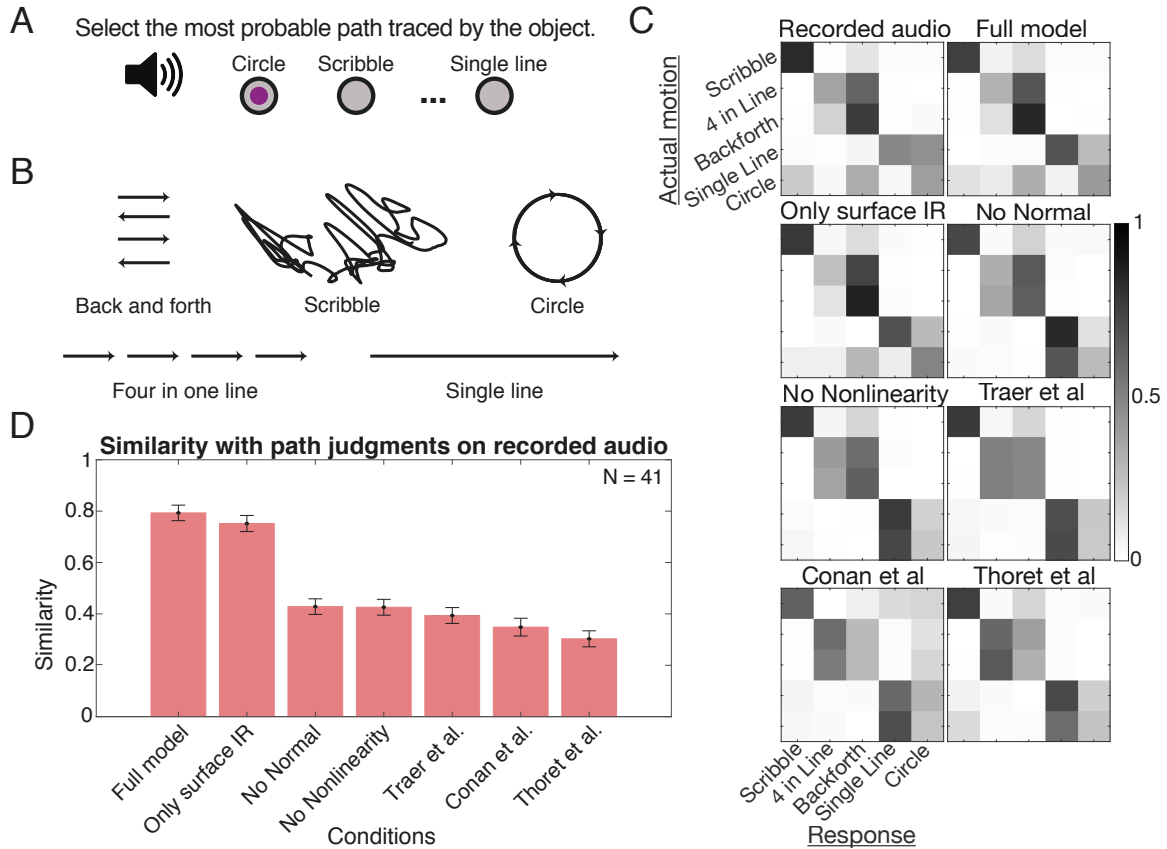


Figure 4-5: Experiment 2: Motion recognition from scraping. A) Participants listened to a sound and selected the path traced by the scraping object. B) Sounds were generated for each of five possible motions. C) Confusion matrices for each of the eight conditions. D) Similarity of the confusion matrix of each synthesized condition compared to that for the recorded sounds.

due to the catching and release between surface asperities that would depend on the trajectory of the point in contact. Figure 4-6 depicts a schematic of the overall contact force for rolling. The first two terms reflect the same calculation used in scraping, and the third term is unique to rolling.

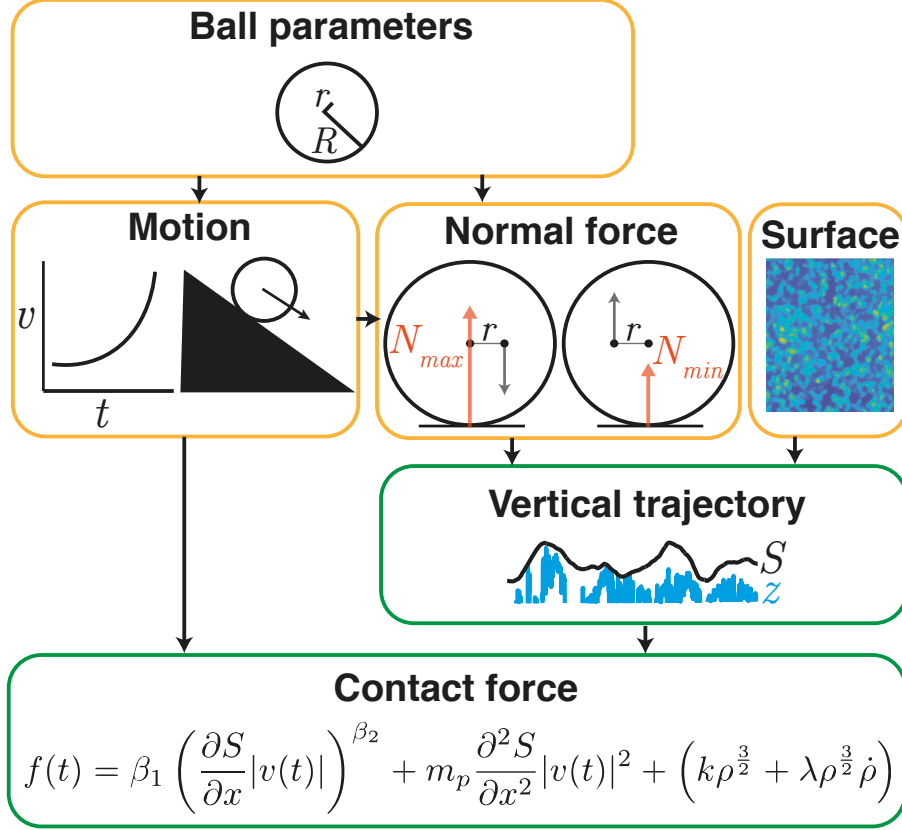


Figure 4-6: Synthesis of contact force for rolling. Yellow boxes are inputs to the synthesizer, which are combined to yield the contact force. The contact force is combined with location-specific impulse responses in the same manner as for scraping (shown in Figure 1).

4.4.1 Contact force for rolling

We used a rolling-specific force term similar to that previously proposed (Rath & Rocchesso, 2005; van den Doel et al., 2001), including a non-linear dissipative component. Our implementation of this force differed from prior work in being based on the trajectory of the ball, derived as in our scraping model as a nonlinear function of the surface depth map:

$$f_{roll}(\rho, \dot{\rho}) = k\rho^{3/2} + \lambda\rho^{3/2}\dot{\rho} \quad (4.14)$$

where ρ denotes penetration depth (the deviation of the ball center of mass from its mean vertical position), k denotes the equivalent spring constant of the ball material

and λ is a dissipation constant which we set by ear. ρ and its derivative are given by

$$\rho = R - r \cos \frac{x}{R} + S(x, y) \quad (4.15)$$

$$\dot{\rho} = \frac{r}{R} \dot{x} \sin \frac{x}{R} + \dot{x} \frac{\partial S(x, y)}{\partial x} \quad (4.16)$$

where x is the horizontal position of the center of mass:

$$x = R\theta - r \sin \theta \quad (4.17)$$

with r the distance between the center of mass and the geometric center and R the mean radius of the ball.

The total contact force is given by

$$f(t) = \beta_1 \left| \frac{\partial S}{\partial x} v(t) \right|^{\beta_2} + m_p \frac{\partial^2 S}{\partial x^2} |v(t)|^2 + (k\rho^{3/2} + \lambda\rho^{3/2}\dot{\rho}) \quad (4.18)$$

The relative contribution of the scraping and rolling terms likely depend on how much slip is present in the rolling motion and the roughness of the materials, and can be adjusted using the existing free parameters β_1 , β_2 , k and λ .

For all three force terms, the vertical trajectory $S(x, y)$ of the point of contact on the surface is determined as in the scraping model. A tanh non-linearity was used to constrain the curvatures of the vertical trajectory based on the amount of normal force. The only difference is that the normal force varies periodically over the rolling interaction due to the offset between the center of mass and the geometric center of the sphere, being maximal when the centre of mass moves downwards, and minimal when it moves up (Figure 4-6). This variation in the normal force affects the penetration of the ball into the surface, which we modeled by varying the non-linearity parameter α and the subsequent smoothing.

4.4.2 Impulse responses

Location-dependent impulse responses were synthesized in the same way as for scraping. An impulse response for the rolling object is added to the surface impulse response, with the relative weighting depending on the ball material. At present we set the relative weighting by hand. In the rolling sound recordings we sought to emulate in our experiments, the surfaces were planks, and typically had a much higher contribution to the overall sound than the balls. If modeling the sound of a ball rolling on a floor, which is typically damped, the weighting would instead upweight the contribution of the ball.

4.5 Perception of synthetic rolling

To evaluate the synthesis model, we asked human listeners to rate the realism of synthetic sounds generated in various ways, using the MUSHRA paradigm from Experiment 1. We again compared our model to lesioned models with omitted components and to two previous baselines using types of physics-based (Rath & Rocchesso, 2005) and signal-based synthesis (Conan, Thoret, et al., 2014).

4.5.1 Experiment 3. Realism of synthetic rolling sounds

In each trial, online participants were presented with a test description of a rolling event which specified the material of the rolling object and the surface, incline of the surface and the motion of the object (Figure 4-7A). The ball material was ceramic, glass or wood. The surface material ceramic, bass wood or poplar wood. The surface had a gradual incline in all the sounds and the ball was rolled up or down the incline. The different synthesis conditions were (1) full model, lesioned models with (2) only the surface IR contribution, (3) a constant surface IR (omitting the location-dependent morphing), (4) only the ball IR contribution, (5) no nonlinearity in the calculation of contact force from the depth map, and (6-7) baseline methods (Conan, Thoret, et al., 2014; Rath & Rocchesso, 2005).

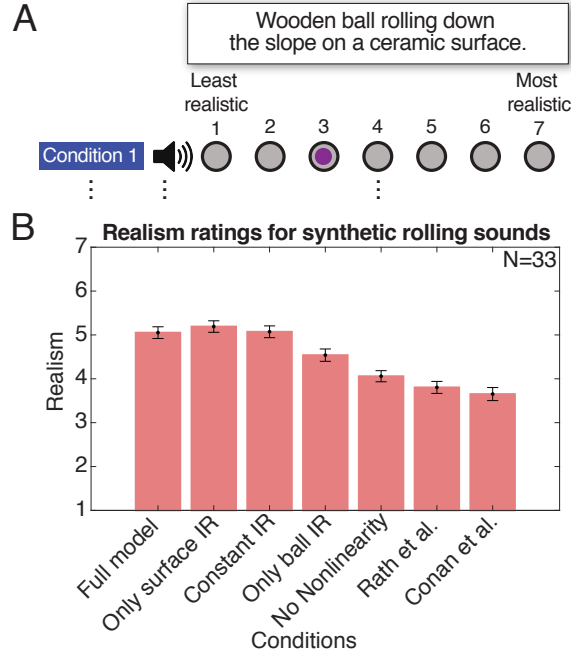


Figure 4-7: Experiment 3: Realism of synthesized rolling. A) Participants rated the realism of 7 different renderings of a ball rolling down or up an inclined surface, using a MUSHRA paradigm. Each of the 7 renderings was from a different synthesis method. The ball and surface varied in material. B) Results of Experiment 3, showing mean realism for each synthesis method. Error bars plot SEM.

As shown in Figure 4-7B, our model produced more realistic rolling sounds than the previous baseline models. The surface resonance was more important than the ball resonance for overall realism (‘Only Surface IR’ vs. ‘Only Ball IR’). As with the scraping sounds of Experiment 1, the non-linearity had a significant impact on the realism of rolling sounds (without it, sounds were unrealistically rough). In contrast, morphing the impulse responses did not significantly contribute to the realism.

4.6 Discussion

We developed a novel method for synthesizing scraping and rolling sounds and evaluated it with perceptual experiments. The model is object-based, in the sense of depending on relatively macroscopic properties of objects and their motions. The key innovations compared to previous methods are the introduction of a nonlinearity in the relationship between contact force and the surface depth map and the use

of normal forces that more accurately reflect scraping by biological organisms. We combine these two ideas with variants of several previous proposals for synthesis in this domain. The resulting synthesis is substantially more realistic than previous methods.

Our experiments explored the contribution of different components of the synthesis model. We found that the non-linearity used to constrain the curvatures of the vertical trajectory of the scraper (and thus the resulting contact force) was critical. Without the nonlinearity, the contact force is unrealistically high and the sounds are unrealistically rough.

The normal force variation was also important for compelling synthesis. Its effect was most evident in the motion estimation experiment, where it was necessary to convey motion comparable to actual audio recordings. Motion recognition for the full model containing this normal force variation qualitatively matched that for audio recordings, despite using ideal velocity profiles which do not follow empirical relations for human-generated profiles (Thoret et al., 2014).

Several synthesis components did not produce a clear benefit in some of the experiments, but this appears to be a limitation of the experiments rather than of the synthesis. For instance, realism judgments for scraping did not distinguish between the full model and the lesioned model without the normal force variation. This is plausibly because sounds without variation in the normal force correspond to an alternative physically possible situation in which the normal force does not change over the path. Participants would rate the sound as realistic because they can envision a physical scenario that could have produced the sound, even though that scenario deviates from the one that was intended. The same issue may explain why there was little effect of omitting the contribution of the scraper impulse response, as this could correspond to a situation with damping.

We suspect that the failure to demonstrate a benefit of location-dependent impulse response morphing (Experiment 3; Fig. 4-7) is also a limitation of the realism judgments we used. There are realistic physical explanations for both the cases where impulse responses change with position and where they do not. For example, rolling

on a damped floor would not lead to location-dependent impulse responses, whereas rolling on a small wooden plank yields a much more appreciable location-dependence. Overall, our subjective sense is that all components of the synthesis that we tested here contribute meaningfully to the resulting sound, but additional work is needed to better understand and demonstrate the situations in which each component of the synthesis is most perceptually relevant.

The model presented here can be further improved in several respects. At present several parameters of the model are set by hand. For instance, we currently lack an empirical relationship between the non-linearity parameter and the normal force. We plan to investigate the impact of the softness/hardness of the materials empirically and to base this parameter on empirical measurements in the future. At present we also set the balance between the rolling and scraping components of rolling sounds by hand, when this should ultimately be determined by physical parameters of the motion and surface characteristics. It also seems likely that the perception of scraping and rolling is based on summary statistics (McDermott et al., 2013) that capture surface roughness (as opposed to detailed representations of the underlying depth maps). Synthesis could thus plausibly be based on depth maps synthesized from summary statistics, which would make for a more parsimonious model.

Efficient and high-fidelity synthesis will open the door to new perceptual studies of ecological audition by enabling sound generation for physical events whose properties can be varied in a controlled manner. Relatively simple synthesis methods with input variables that are both physically meaningful and perceptually consequential can also provide the foundation for models of human auditory perception. If prior distributions are determined for the physical variables, the principles of Bayesian inference can be leveraged to invert the synthesis method and thereby estimate latent physical variables from sound (Cusimano et al., 2018; Kersten & Yuille, 2003). We plan to develop a complete contact sound synthesis model (including scraping, rolling, and impacts) within an inference framework that can be used to make predictions about the mechanisms underlying human physical inference. In principle, this could be combined with inverse graphics (Gan et al., 2020; Yildirim, Belledonne, et al.,

2020) to achieve multimodal physical event perception. Bayesian inference could also provide a way to set the values of the free parameters, given corpora of recorded contact sounds.

4.7 Conclusion

We have presented a method to synthesize the sounds of scraping and rolling from physical descriptions of objects (surface depth maps, object impulse responses) and their motions. The method is relatively efficient and produces substantially more realistic scraping and rolling sounds than previous methods.

Chapter 5

Conclusion

My thesis has presented accounts of various domains of auditory perception based on generative models. This conclusion first discusses various “behind-the-scenes” methodological details to provide context for this style of work, mainly focusing on the Bayesian auditory scene synthesis project. Then, I move to discussing more general ideas about perception sparked by this thesis.

5.1 Methodology behind generative models

5.1.1 Enabling factors

Many of the ideas behind the Bayesian auditory scene synthesis model can be traced to classical ideas in vision as well as to important precursors in auditory perception (Ellis, 2006). Here I discuss the reasons of why these ideas can be more fully implemented now.

1. **Basic tools.** It is easy to underrate how much the projects in this thesis were enabled by the development and accessibility of very basic computational tools, including both software and hardware. In one of my first classes in graduate school in 2015, I was introduced to Autograd (Maclaurin, 2016). We were also learning to specify differentiable functions manually, and automatic differentiation was included in the class because it was actively being developed by

one of the class’s teaching assistants, David Duvenaud. At the time, it was a uniquely fully automatic differentiation system. Autograd eventually formed the basis of backproagation in Pytorch, initially released in 2016. It was then only in 2018 that Pytorch released an expressive collection of probability distributions. These distributions also automatically implemented differentiable sampling through the reparametrization trick. GPyTorch, a Python package with which we express our Gaussian process priors and approximate mean-field posterior, was released in mid-2017. The relative ease of iterating on models and inference without having to completely re-implement learning algorithms and distributions was helpful for completing this project. Second, as discussed in Chapter 2, access to a large number of GPUs in parallel was essential for sequential inference. For me, the amount of GPU memory and time required to run inference on a single sound has challenged the (algorithmic) idea of perception as search through a hypothesis space. Rather than scaling models to use more compute, developing more efficient implementations of search is an interesting target for research. For example, amortizing the sequential combination of events into sources (while retaining uncertainty) seems tractable and would vastly improve the efficiency of inference.

2. **Inference.** We used an analysis-by-synthesis approach for inference, which conceptually corresponded well to the two main challenges in inference that we encountered: searching for modes and precisely estimating the posterior distribution. Each challenge was addressed by developments that occurred during or just before I started graduate school: a bottom-up amortized neural network (Wu et al., 2019) and stochastic variational inference (Kucukelbir et al., 2017) respectively. Both methods rely on the technical innovations just discussed and on stochastic gradient descent in particular. Early versions of our model instead relied on Metropolis-Hastings Markov chain Monte Carlo (MCMC), using reversible jump MCMC to infer the scene structure (Cusumano-Towner et al., 2019). We found that it was too difficult to hand-design proposals that jump

between modes in the high-dimensional scene space. For example, to resample an event’s amplitude trajectory could also require resampling the source-level latent variables or a different event’s latent variables. The variety of inference problems presented by different auditory scene analysis phenomena generally required us to write new resampling proposals. Switching to stochastic variational inference allowed us to leave behind the hand-designed proposals and instead use gradient descent over all variables jointly. Stochastic gradient descent was a workable, general-purpose search heuristic; however, many remaining inference issues are due to local minima (see Section 2.3.2).

3. **Model.** Earlier work in Bayesian generative models used graphical models as the basic formalism for model structure (Kersten & Schrater, 2002). Our model has discrete variables that change the dimensionality of the model and integrates a renderer, and so it cannot be expressed as a graphical model. The expressivity of the model instead stems from writing it as a probabilistic program (Ghahramani, 2015). Also, using stochastic variational inference allows us to define a much more flexible posterior approximation for a richer model than is possible with exponential family models in classic variational inference.

5.1.2 Challenges

Taking a comprehensive approach to perceptual organization introduced challenges for both designing the generative model and doing inference. First, to comprehensively explain perception of a variety of illusions and everyday sounds with a single model is non-trivial, requiring conceptual innovations in the generative model. While the basic structure of our model was inspired by everyday sounds, its ability to comprehensively explain psychophysics results required an iterative process of model design and evaluation. A benefit of this approach is that the model’s interpretability generally lends itself to reasoning, such that the final details of the generative process (such as the non-stationarity in time) provide insight into perceptually relevant aspects of sound structure. Second, we had to put together a rather general-purpose

inference algorithm composed of several parts in order to confront a wide variety of inference problems. For example, in the tone sequence demos, the events are quite obvious and the main challenge is to organize events into sources. However, for the perceptual filling-in results, the source organization is quite obvious while the event structure needs to be inferred. We initially solved these various problems independently with different inference algorithms, before understanding how to unify them. After both of these unification processes, it was surprising to me how little conceptual innovation was necessary to apply the model and inference to everyday sounds: the main conceptual addition was the heuristics to prioritize simpler hypotheses for optimization.

Second, building a generative model that can account for and synthesize everyday sounds was difficult both for the BASS model and the ecological synthesis projects. Of course, there is a research process to discovering the relevant generative principles. But moreover, humans listeners are extremely good at knowing when sounds are synthesized. It seems like there are small, perceptually detectable effects in sound generation which are hard to reason about from physical principles or from experience. I think integrating learned components into structured models could give us insight into what kinds of structure we are missing in our synthesis algorithms. Also, this would allow us to integrate different kinds of statistical components beyond the simple material distributions explored in Chapter 3.

Finally, inference remains an effort/time bottleneck for this approach and can limit the expressivity of choices in the generative model. Even now, inference does not always work and is resource-intensive. We have already discussed several aspects of this. It is worth adding that we are now working on developing inference methods for the synthesis models described in Chapters 3 and 4. These models currently deal with more constrained domains (e.g., a single impact event), and initial results indicate that some of the inference methods developed for this thesis extend to those domains.

5.1.3 Imagining new starting points

The causal process behind the design of a generative model and inference algorithm involves a degree of path dependency. In this section, I consider possible alternatives to how the BASS project could have proceeded.

Experimental methods for probing everyday scene analysis

If I were to start this project again, I would probably want to first figure out how to convincingly probe listeners’ perceptual organization in everyday sounds. Our model comparisons in Experiment 1 and 2 are very coarse-grained, using the mixture and premixture sounds as perceptual references. Such experiments could elucidate issues raised in this thesis about hierarchical perceptual organization and provide new constraints for the distinct perceptual entities posited by the model.

Fresh inspiration for the scene descriptions

Another new starting point would be to re-consider the inspiration for the BASS generative model. As described in Chapter 2, the basic design of the BASS model was inspired by everyday sounds, grounded in previous modeling work, and constrained by what was tractable. An initial pilot version of the project involved a few illusions as well as inference on simple everyday sound mixtures (Cusimano et al., 2018). I am curious what model we would have ended up with if we first focused on designing a signal-based generative model for which we could do inference on everyday sounds that we perceive to have ‘single sources’ (if that is indeed what would be discovered by the hypothetical experimental methods just discussed), more in the style of our ecological sound synthesis projects. Another interesting possibility of inspiration would be to review existing sound synthesizers (e.g., for music and audio design) and think about the utility of those as scene representations.

Redesigning source models.

Basic sound types would be a key aspect of the source models to reconsider because they are somewhat arbitrary. Our types (noise, harmonic, whistle) do not exhaustively describe everyday sounds, are not obviously distinct (i.e. harmonic and whistle; hierarchical sources with both aperiodic and periodic components), nor are they grounded in psychophysical results that indicate these are true perceptual primitives. Other systems have made different choices for primitives (eg. transients/clicks in Ellis (1996) and Misra et al. (2009)). However, in Experiment 2 of Section 2.2.6, we also found reason to think at least that the noise and harmonic sound types were reasonable descriptions of perceptual categories: the worst unrecognizability deviations occurred when the model inferred a source with a periodicity mismatch to the mixture (i.e. used a harmonic source rather than a noise source, or vice versa). There are many options for reconsidering the sound types; we discuss a few here to shed light on the kinds of issues that would need to be addressed.

To address the limitations of the BASS model while retaining its discrete signal-based sound types, additional or alternative types are needed as well as ways to compose the primitives. Additional primitives could include transients and inharmonic tones. Alternatively, harmonic and inharmonic tones could be handled by a single source type that describes a collection of unconstrained sinusoids. Presumably, it would be necessary to define the priors such that harmonic sounds are more probable a priori. To compose primitive types, it would also be necessary to define the prior on the covariance between the temporal evolution of each excitation.

Rather than defining compositions of the primitive types, type could instead be construed as a high-dimensional continuous space. All source models would need to be higher-dimensional in order to account for different rendering possibilities, even if the specific source does not produce certain sounds (e.g., contain parameters to specify that the noisy part of the excitation is silent). However, a continuous space of sound types could potentially better describe everyday sounds as well as improve inference by providing a smooth path between types.

Finally, a distinct approach would be to abandon signal-based types and instead consider a separate source model for various categories derived from an ecological ontology of sound. For instance, there could a source model for animal vocalizations and a source model for various kinds of impact sounds. Prior work on sound ontology could be inspiring in this regard (e.g. Gaver, 1993b). One challenge here would be to determine whether the various models should share any latent variables.

Redesigning the Gaussian processes

We chose to use Gaussian processes as a classic Bayesian prior over functions. In the initial versions of the model, for simplicity, we used the standard squared exponential kernel. However, the standard variance and lengthscale parameters may not be the most relevant to characterize a perceptual source. Reconsidering the assumptions to build into the kernel (or learning them; e.g. Cavaco and Lewicki, 2007) from first principles could be insightful, particularly when it comes to incorporating time-asymmetry, event-linked excitation structure (e.g., attack-decay-sustain-release envelopes), changing spectra (e.g., frequency-dependent decay), and repetition. It also may be worth considering other options like state space models.

Likelihood representations

The cochleagram likelihood representation was a major limitation. Starting over, I would try to choose (or figure out how to discover) more informative mid-level audio representations from the start. There are a variety of bottom-up features of audio that researchers have designed over the years, although they may need to be made differentiable or more memory/time-efficient. It is also arguably implausible that a perceptual model would actually go all the way to the soundwave. This choice has consequences in the BASS model, because the model has access to explanations involving the destructive interference of audio that don't seem relevant to perception. To avoid this, a model would have to directly generate mid-level features from its latent variables (rather than generating the soundwave and then computing mid-level features from that). Perhaps given a full generative model of audio and an

algorithm for computing the mid-level representation from audio, a neural network could be trained to go directly from the latent-variable renderer inputs to the mid-level representations. This could be a productive approach for including sound texture in the model.

Inference

One aspect of inference that showed considerable path dependency was our treatment of time. In the first iteration of the project (Cusimano et al., 2018), we used an earlier version of the segmentation network that we eventually included in the final paper. These networks take in the full observation and then output event proposals, which are then used in different rounds of sequential inference. An alternative would be to limit the amortized inference network to considering only the cochleagram frames observed on a single round of sequential inference. For other concerns about our treatment of time, see Section 2.3.4. The main challenge would be to determine how to appropriately integrate context during inference, but it would be interesting to see if treating time more realistically could naturally resolve any inference difficulties.

5.2 Themes for perception

I end with some relatively unstructured thoughts on aspects of perception that I am intrigued by, or confused about, after completing these projects. I sometimes wonder whether these issues stem from an ontological blurriness of whether generative models are a scientific model of the world or a proposal for the internal model within the perceiver.

5.2.1 More on structured generative models

In some ways, I feel that using structured generative models for science is a matter of personal aesthetics that can contribute to methodological diversity. However, here I wanted to add a few points that add to those in the Discussion section of Chapter 2 that explain how structured generative models can complement machine learning

approaches. Phenomenology and natural language are extremely powerful scientific tools in perception, as shown by the lasting impact of Gestalt psychology. While analyzing the results of the everyday sound experiments, I was struck by how clearly Gaver’s 1993 paper foretold the kinds of structure that were being overlooked by mainstream psychophysics work in auditory scene analysis. Of course, phenomenology and natural language have limitations and so it is important to use them in conversation with other methods. I think the generative models in the styles explored in this thesis make good formal complements to a phenomenological approach, because they provide a language to write down what we perceive in a formal but still interpretable way (rather than in natural language) and because of their ability to generate and manipulate signals. For me, developing synthesis methods for non-generative models as well (e.g. Feather et al., 2019) seems key for this same reason.

5.2.2 More on illusion generation

A generative model of perception can both listen to and make sounds. By having one of our generative models listen to its own sounds, we could automatically generate illusions by finding the stimulus properties that would produce a desired percept. It would also be possible to use two separate generative models, one for synthesis and one for listening (e.g., our physics-based synthesizers for generation and the BASS model for listening). These sounds could provide a strong additional model test. This corresponds to an additional layer of inference over the generative parameters of the illusory sound. Chandra et al. (2022) formalized this process as doing inference through Hamiltonian Monte Carlo, by taking advantage of its time-reversibility. In a model that uses stochastic variational inference, an alternating, iterative expectation-maximization style of inference may be more appropriate. Otherwise, in analogy to Chandra et al. (2022), inference would have to be tracked through gradient descent. In Table 5.1, I list some possibilities for interesting perceptual objectives to make illusions. I think this possibility could be especially interesting with physical sound synthesis because it could generate naturalistic auditory illusions.

Percept	Objective	Example
Multistability	$\operatorname{argmax}_{S_e} H(S_p X)$	Bistable material perception.
Conflicting global and local percepts	$\operatorname{argmax}_{S_e} p(S_p = \phi X[t_1 : t_2])$ $\times p(S_p \neq \phi X)$	Continuity illusion where $X[t_1 : t_2]$ is white noise.
Smallest difference in the sound that allows a structural change in the percept	$\operatorname{argmin}_{S_{e_2}} \ X_1 - X_2\ ,$ such that S_{p_1} and S_{p_2} are structurally different.	Move an impact's mode to cause it to group differently.
Separable/inseparable source pairs	$\operatorname{argmax}_{S_{e_1}, S_{e_2}} p(n_s = 1 X_1, X_2)$ $\times \ S_{e_1} - S_{e_2}\ $ or, $\operatorname{argmax}_{S_e} p(n_s = 2 X)$	Discover different textures that sound like 1 source when they are mixed. Find a single texture that sounds like two sources.
Induce different groupings by altering context	Fix events ν_a and ν_b . Optimize two sounds. $\operatorname{argmax}_{S_{e_1}} P(\nu_a \in \text{source}(\nu_b) X_1),$ $\operatorname{argmax}_{S_{e_2}} P(\nu_a \notin \text{source}(\nu_b) X_2)$	Modify a rolling event to capture an impact event from a different source.

Table 5.1: Perceptual objectives for illusion generation. S_e is the scene description that encodes the illusory sound X . S_p is the perceived scene description.

5.2.3 Integrating structure and statistics

In Ted Adelson’s graduate seminar *Touching and Grasping with Soft Fingers* (9.357), one of the questions that came up was, how do you tell that something in your hand is a “single object”? What’s a “single object” experientially? It is not clear to me how the classic idea of a physical object corresponds to perceived wholes in different modalities. Certainly, different perceptual systems share similar issues concerning local measurements versus global meaning. Consider trying to spot a bird as it is occluded by leaves in a tree, picking out that same bird’s call as it is masked by the bubbling of a stream, the feeling of attending to that bird even as it goes silent or tracking fireflies flashing on and off, exploring something that doesn’t fit into your hand all at once or a crayfish only appearing against the riverbed when it moves. However, we often perceive wholes in transformation rather than constancy: smoke that curls and dissipates, ‘objects’ that irreversibly transform in making sound (eg., boiling water, shattered glass, a popped balloon), materials that we explore by changing them (eg., crumpling paper, snapping a stick). One question is, can we clarify the similarities and differences in these problems with generative models, and

how would this lend clarity to what we perceive?

And yet, there is structuring of sensory data by many possibly interacting causal generative processes and at many scales. And so in contrast to the idea of a coherent object, it seems like you can perceive the parts to have parts, et cetera, or that the wholes can be integrated into larger wholes, et cetera. This is why I like the language of ‘structured wholes’. Imagine music playing on the radio in a room with a window open. You could listen to the guitar in the band on the radio, the radio itself, all the sounds inside, all the sounds outside, or the sound of a dog barking down the street. Or, the experience of listening to rain as a coherent sound texture, versus listening to how rain drops sound different on the different surfaces around you. While hierarchy has been discussed in perceptual organization in vision, it has been understudied in hearing (although noted e.g. by Gaver). This may be connected to the account of attention put forward by Whiteley and Sahani (2012), in which attention modulates the complexity of an approximate posterior. Perhaps in addition to being a mechanism that modulates the difficulty of approximate inference, attention is a mechanism to modulate the scale of analysis in an optimal way for how (i.e., at what level of whole) an organism is exploring its environment – bringing some wholes into structural focus. (Attention could be a lot of things: modulating a prior, modulating the incoming data, changing a model, changing an approximate posterior, and it is hard to know which are different in both math and in terms of predictions for behavior.) How could we innovate generative model that centers this perceptual flexibility rather than the more traditional idea of sources?

Similarly, one theme across the different generative models presented is where the model of perception is structured (e.g. because it is physics-inspired) versus where it abstracts from that structure into statistics (called “deep” versus “shallow” representations’ by Morgenstern and Kersten (2017)). There might not be a strong distinction here in perception but it is certainly a choice in modeling. For the physics-inspired sound synthesis projects, one of our initial motivations was to understand how to compose event structure with structured changes in texture statistics, by explicitly integrating spatially-averaged statistical representations of surface structure. In some

ways this was not entirely addressed: we used quilting of actual height samples from the measured surfaces, rather than spatially-averaged texture statistics. However, we did modulate the surface samples (and thus the noisy contact force) through a physically-inspired simulation of motion over a surface. Another statistical representation we used was of object material, but rather than averaging over time or space, we defined statistical distributions over modes measured from recorded audio. Reverberation is another example where the physical detail of the sound-generating process does not seem to matter (Traer & McDermott, 2016). The cases where these statistical representations showed up reminded me of material perception in vision, and how Fleming and Storrs (2019) (in part inspired by material perception) advocated for learning generative models. I am curious to understand the contexts under which various levels of abstraction are represented in perception, an investigation which could perhaps be facilitated by learning structured models of perception or learning within structured models (e.g. Engel, Hantrakul, et al., 2020).

5.2.4 A multiplicity of world models in perception

One question that arises from the various kinds of models explored in this thesis is how perception can “take different perspectives”. For example, Gaver (1993b) distinguishes between musical and everyday listening. He writes:

“The distinction between everyday and musical listening is between experiences, not sounds (nor even psychological approaches). It is possible to listen to any sound either in terms of its attributes or in terms of those of the event that caused it. For instance, while listening to a string quartet, we might be concerned with the patterns of sensation the sounds evoke (musical listening) or we might listen to the characteristics and identities of the instruments themselves (everyday listening). Conversely, while walking down a city street, we are likely to listen to the sources of sounds—the size of an approaching car, how close it is, and how quickly it is approaching—but occasionally we might listen to the world as we do to

music—to the humming pitch of a ventilator punctuated by a syncopated birdcall, to the interplay and harmony of the sounds around us.”

How do we listen in these different ways (if they are indeed distinct) and what does that imply for the structure of wholes in hearing? Intuitively, to focus on musical listening seems to relate to the ‘signal-based’ latent variables in the Bayesian auditory scene synthesis model, while everyday listening seems to relate to the statistical and physics-inspired latent variables of the ecological sound synthesizers. As raised in Chapter 2, what is the relationship of these kinds of world models to each other? One could equally imagine a deep, unified and compositional world model (in which musical listening reflects an intermediate level of latent explanation upon which a full ecological explanation is built; and ‘music’ reflects those sounds for which perception readily emphasizes that intermediate level¹), or a mixture of multiple systems, for example, one for environmental events and another for patterned musical sounds. Either way, the question remains as to how our perceptual systems emphasize different aspects of description in experience. One interesting way to examine this may be to study vocal or synthesizer-aided imitations of environmental sounds: how do human listeners translate the perceived generative structure behind one sound to generating another? This also seems to me to be related to the question of the richness of human perception when it comes to art. It is hard to imagine classic generative models being able to handle such a situation—would a different renderer be needed for every new artist, every new expression? Whether generative models can elucidate the structure underlying the richness of perceptual experience remains to be seen.

¹Thank you to Dan Ellis for this elaboration.

Bibliography

- Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, 116(5), 3099–3107.
- Adelson, E. H., & Pentland, A. P. (1996). The perception of shading and reflectance. *Perception as Bayesian inference*, 409, 423.
- Agamben, G. (2004). *The open: Man and animal*. Stanford University Press.
- Agarwal, V., Cusimano, M., Traer, J., & McDermott, J. (2021). Object-based synthesis of scraping and rolling sounds based on non-linear physical constraints. *2021 24th International Conference on Digital Audio Effects (DAFx)*, 136–143.
- Alain, C., Arnott, S. R. et al. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience*, 5, D202–D212.
- Albertazzi, L. (2015). Philosophical background: Phenomenology. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization*. Oxford University Press.
- Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., & De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29.
- Aramaki, M., Besson, M., Kronland-Martinet, R., & Ystad, S. (2010). Controlling the perceived material in an impact sound synthesizer. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2), 301–314.
- Avanzini, F., & Rocchesso, D. (2001). Controlling material properties in physical models of sounding objects. *Proceedings of the 2001 International Computer Music Conference*.

- Avanzini, F., Serafin, S., & Rocchesso, D. (2005). Interactive simulation of rigid body interaction with friction-induced sound generation. *IEEE transactions on speech and audio processing*, 13(5), 1073–1081.
- Barker, J., Cooke, M., & Ellis, D. (2005). Decoding speech in the presence of other sources. *Speech Communication*, 45(1), 5–25. <https://doi.org/https://doi.org/10.1016/j.specom.2004.05.002>
- Barniv, D., & Nelken, I. (2015). Auditory streaming as an online classification process with evidence accumulation. *PloS one*, 10(12), e0144788.
- Baylis, G. C., & Driver, J. (1993). Visual attention and objects: Evidence for hierarchical coding of location. *Journal of Experimental Psychology: Human Perception and Performance*, 19(3), 451.
- Beckers, G. J., Suthers, R. A., & Ten Cate, C. (2003). Pure-tone birdsong by resonance filtering of harmonic overtones. *Proceedings of the National Academy of Sciences*, 100(12), 7372–7376.
- Bilbao, S. D. (2009). *Numerical sound synthesis*. Wiley Online Library.
- Billig, A. J., Davis, M. H., Deeks, J. M., Monstrey, J., & Carlyon, R. P. (2013). Lexical influences on auditory streaming. *Current Biology*, 23(16), 1585–1589.
- Billot, B., Greve, D. N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A. V., & Iglesias, J. E. (2021). Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution. *arXiv preprint arXiv:2107.09559*.
- Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature*, 402(6764), 877–879.
- Bregman, A. S. (1978a). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3), 380.
- Bregman, A. S. (1978b). Auditory streaming: Competition among alternative organizations. *Perception & Psychophysics*, 23(5), 391–398.
- Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

- Bregman, A. S. (2005). Auditory scene analysis and the role of phenomenology in experimental psychology. *Canadian Psychology/Psychologie canadienne*, 46(1), 32.
- Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of experimental psychology*, 89(2), 244.
- Bregman, A. S., & Rudnicki, A. I. (1975). Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 263.
- Bregman, A., & Ahad, P. (1996). Demonstrations of auditory scene analysis: The perceptual organization of sound.
- Brooks, J. L. (2015). Traditional and new principles of perceptual grouping. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization*. Oxford Library of Psychology.
- Brown, G., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech Language*, (8), 297–336.
- Brunswik, E., & Kamiya, J. (1953). Ecological cue-validity of 'proximity' and of other gestalt factors. *The American Journal of Psychology*, 66(1), 20–32.
- Burger, S., Jin, Q., Schulam, P. F., & Metze, F. (2012). Noisemes: Manual annotation of environmental noise in audio streams.
- Cadoz, C. (1979). *Synthèse sonore par simulation de mécanismes vibratoires. applications aux sons musicaux* (Doctoral dissertation). Institut national polytechnique de Grenoble.
- Cadoz, C., Luciani, A., & Florens, J. L. (1993). Cordis-anima: A modeling and simulation system for sound and image synthesis: The general formalism. *Computer music journal*, 17(1), 19–29.
- Cadoz, C., Luciani, A., & Florens, J.-L. (1984). Responsive input devices and sound synthesis by stimulation of instrumental mechanisms: The cordis system. *Computer music journal*, 8(3), 60–73.

- Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological science*, 9(3), 211–214.
- Carlyon, R. P., Deeks, J. M., Shtyrov, Y., Grahn, J., Gockel, H. E., Hauk, O., & Pulvermüller, F. (2009). Changes in the perceived duration of a narrowband sound induced by a preceding stimulus. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6), 1898.
- Cavaco, S., & Lewicki, M. S. (2007). Statistical modeling of intrinsic structures in impacts sounds. *The Journal of the Acoustical Society of America*, 121(6), 3558–3568.
- Chakrabarty, D., & Elhilali, M. (2019). A gestalt inference model for auditory scene segregation. *PLoS computational biology*, 15(1), e1006711.
- Chandra, K., Li, T.-M., Tenenbaum, J., & Ragan-Kelley, J. (2022). Designing perceptual puzzles by differentiating probabilistic programs. *Proceedings of SIGGRAPH 2022*.
- Chen, Z., Luo, Y., & Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 246–250.
- Conan, S., Derrien, O., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2014). A synthesis model with intuitive control capabilities for rolling sounds. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(8), 1260–1273.
- Conan, S., Thoret, E., Aramaki, M., Derrien, O., Gondre, C., Ystad, S., & Kronland-Martinet, R. (2014). Intuitive synthesizer of continuous-interaction sounds: Rubbing, scratching, and rolling. *Computer Music Journal*, 38(4), 24–37.
- Cooke, M., & Ellis, D. P. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech communication*, 35(3-4), 141–177.
- Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., & Vincent, E. (2020). Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*.

- Cusimano, M., Hewitt, L. B., Tenenbaum, J., & McDermott, J. H. (2018). Auditory scene analysis as bayesian inference in sound source models. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Cusumano-Towner, M. F., & Mansinghka, V. K. (2018). Using probabilistic programs as proposals. *arXiv preprint arXiv:1801.03612*.
- Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). Gen: A general-purpose probabilistic programming system with programmable inference. *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 221–236. <https://doi.org/10.1145/3314221.3314642>
- Darwin, C., & Sutherland, N. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *The Quarterly Journal of Experimental Psychology*, 36(2), 193–208.
- Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, 178, 67–81.
- Deike, S., Denham, S. L., & Sussman, E. (2014). Probing auditory scene analysis. *Frontiers in neuroscience*, 8, 293.
- Doucet, A., De Freitas, N., & Gordon, N. (2001). An introduction to sequential monte carlo methods. *Sequential monte carlo methods in practice* (pp. 3–14). Springer.
- Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive psychology*, 5(3), 322–337.
- Efros, A. A., & Freeman, W. T. (2001). Image quilting for texture synthesis and transfer. *SIGGRAPH: Computer Graphics*.
- Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, 2(4), 5–5.
- Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America*, 124(6), 3751–3771.

- Ellis, D. P. W. (2009). Gammatone-like spectrograms. <http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/>
- Ellis, D. (2006). Computational auditory scene analysis: Principles, algorithms, and applications. In D. Wang & G. J. Brown (Eds.), *Auditory scene analysis: Principles, algorithms, and applications* (pp. 115–146). Wiley-IEEE press.
- Ellis, D. P. W., & Rosenthal, D. F. (1995). *Mid-level representations for computational auditory scene analysis*. Citeseer.
- Ellis, D. P. (1994). A computer implementation of psychoacoustic grouping rules. *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 2.
- Ellis, D. P. (1996). *Prediction-driven computational auditory scene analysis* (Doctoral dissertation). Columbia University.
- Ellis, K. (2020). *Algorithms for learning to induce programs* (Doctoral dissertation). Massachusetts Institute of Technology.
- Engel, J., Hantrakul, L. H., Gu, C., & Roberts, A. (2020). Ddsp: Differentiable digital signal processing. *International Conference on Learning Representations*. <https://openreview.net/forum?id=B1x1ma4tDr>
- Engel, J., Swavely, R., Hantrakul, L. H., Roberts, A., & Hawthorne, C. (2020). Self-supervised pitch detection by inverse audio synthesis. *ICML 2020 Workshop on Self-supervision in Audio and Speech*. <https://openreview.net/forum?id=RlVtYWhsky7>
- Eslami, S., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Hinton, G. E., et al. (2016). Attend, infer, repeat: Fast scene understanding with generative models. *Advances in Neural Information Processing Systems*, 29.
- Feather, J., Durango, A., Gonzalez, R., & McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, 32.
- Feather, J., Leclerc, G., Mađry, A., & McDermott, J. H. (2022). Model metamers illuminate divergences between biological and artificial neural networks. *bioRxiv*.

- Feinman, R., & Lake, B. M. (2020). Learning task-general representations with generative neuro-symbolic modeling. *arXiv preprint arXiv:2006.14448*.
- Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local “association field”. *Vision Research*, *33*(2), 173–193.
- Fischer, B. J., & Peña, J. L. (2011). Owl’s behavior and neural representation predicted by bayesian inference. *Nature neuroscience*, *14*(8), 1061–1066.
- Fleming, R. W., & Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences*, *30*, 100–108. <https://doi.org/https://doi.org/10.1016/j.cobeha.2019.07.004>
- Fletcher, N. H., & Rossing, T. D. (2012). *The physics of musical instruments*. Springer Science & Business Media.
- Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure–ground cues are valid for natural images. *Journal of Vision*, *7*(8), 2–2.
- Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, *6*(1), 111–133.
- Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review*, *122*(4), 575.
- Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., Freitas, J. D., Kubilius, J., Bhandwaldar, A., Haber, N., Sano, M., Kim, K., Wang, E., Mrowca, D., Lingelbach, M., Curtis, A., Feigelis, K., Bear, D. M., Gutfreund, D., Cox, D., ... Yamins, D. L. K. (2020). Threedworld: A platform for interactive multi-modal physical simulation.
- Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., & Wilson, A. G. (2018). Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *arXiv preprint arXiv:1809.11165*.
- Gaver, W. W. (1993a). How do we hear in the world? explorations in ecological acoustics. *Ecological psychology*, *5*(4), 285–313.

- Gaver, W. W. (1993b). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1), 1–29.
- Geisler, W., Perry, J., Super, B., & Gallogly, D. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6), 711–724.
- Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780.
- Gershman, S. J., Tenenbaum, J. B., & Jäkel, F. (2016). Discovering hierarchical motion structure. *Vision research*, 126, 232–241.
- Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. (2021). What happened? reconstructing the past through vision and sound. <https://doi.org/10.31234/osf.io/tfjdk>
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553), 452–459.
- Giordano, B. L., & McAdams, S. (2006). Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119(2), 1171–1181.
- Gothoskar, N., Cusumano-Towner, M., Zinberg, B., Ghavamizadeh, M., Pollok, F., Garrett, A., Tenenbaum, J., Gutfreund, D., & Mansinghka, V. (2021). 3dp3: 3d scene perception via probabilistic programming. *Advances in Neural Information Processing Systems*, 34.
- Grassi, M. (2005). Do we hear size or sound? balls dropped on plates. *Perception & psychophysics*, 67(2), 274–284.
- Gregory, A. H. (1994). Timbre and auditory streaming. *Music Perception*, 12(2), 161–174.
- Grinfeder, E., Lorenzi, C., Sueur, J., & Hauptert, S. (2022). What do we mean by "soundscape"? a functional description. *Frontiers in Ecology and Evolution*, 10, 894232.

- Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectrotemporal pattern analysis. *The Journal of the Acoustical Society of America*, 76(1), 50–56.
- Hartmann, W. M., & Goupell, M. J. (2006). Enhancing and unmasking the harmonics of a complex tone. *The Journal of the Acoustical Society of America*, 120(4), 2142–2157.
- Helmholtz, H. L., & Ellis, A. J. (1875). On the sensation of sound in general.
- Henrywood, R. H., & Agarwal, A. (2013). The aeroacoustics of a steam kettle. *Physics of Fluids*, 25(10), 107101. <https://doi.org/10.1063/1.4821782>
- Hensman, J., Matthews, A., & Ghahramani, Z. (2015). Scalable variational gaussian process classification. *Artificial Intelligence and Statistics*, 351–360.
- Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 31–35.
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, 97(5), 3099–3111.
- Houtsma, A. J. M., Rossing, T. D., & Wagenaars, W. M. (1988). Auditory demonstrations on compact disc. *The Journal of the Acoustical Society of America*, 83(S1), S58–S58. <https://doi.org/10.1121/1.2025424>
- Hu, J., Traer, J., & McDermott, J. H. (2019). Separating object resonance and room reverberation in impact sounds. *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*, 449.
- James, D. L., Barbič, J., & Pai, D. K. (2006). Precomputed acoustic transfer: Output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics (TOG)*, 25, 987–995.
- Jayant, N., Johnston, J., & Safranek, R. (1993). Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10), 1385–1422.

- Josupeit, A., Schoenmaker, E., van de Par, S., & Hohmann, V. (2020). Sparse periodicity-based auditory features explain human performance in a spatial multitalker auditory scene analysis task. *European Journal of Neuroscience*, *51*(5), 1353–1363.
- Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.
- Kersten, D., & Schrater, P. (2002). Pattern inference theory: A probabilistic approach to vision. In R. Mausfeld & D. Heyer (Eds.), *Perception and the physical world*. John Wiley; Sons, Ltd.
- Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*(2), 150–158.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*(3), 971–995.
- Klatzky, R. L., Pai, D. K., & Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence: Teleoperators & Virtual Environments*, *9*(4), 399–410.
- Knill, D. C., & Kersten, D. (1991). Apparent surface curvature affects lightness perception. *Nature*, *351*(6323), 228–230.
- Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision research*, *43*(24), 2539–2558.
- Koenderink, J. (2015). Methodological background: Experimental phenomenology. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization*. Oxford University Press.
- Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, *2*(9), e943.

- Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLOS Computational Biology*, 10(12), e1003985.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*.
- Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. *Proceedings of the ieee conference on computer vision and pattern recognition*, 4390–4399.
- Lagrange, M., Scavone, G., & Depalle, P. (2010). Analysis/synthesis of sounds generated by sustained contact between rigid objects. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3), 509–518.
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.
- Larigaldie, N., Yates, T., & Beierholm, U. R. (2021). Perceptual clustering in auditory streaming. *bioRxiv*.
- Lee, J. S., Depalle, P., & Scavone, G. (2010). Analysis/synthesis of rolling sounds using a source-filter approach. *13th Int. Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria.
- Lemaitre, G., & Heller, L. M. (2012). Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America*, 131(2), 1337–1348.
- Lewicki, M. S., Olshausen, B. A., Surlykke, A., & Moss, C. F. (2014). Scene analysis in the natural environment. *Frontiers in psychology*, 5, 199.
- Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. (2018). Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2), 522–535.
- Lloyd, D. B., Raghuvanshi, N., & Govindaraju, N. K. (2011). Sound synthesis for impact sounds in video games. *Symposium on Interactive 3D Graphics and Games*, PAGE–7.

- Lostanlen, V., Lafay, G., Andén, J., & Lagrange, M. (2018). Relevance-based quantization of scattering features for unsupervised mining of environmental audio. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1), 1–10.
- Lutfi, R. A. (2008). Human sound source identification. *Auditory perception of sound sources* (pp. 13–42). Springer.
- Lutfi, R. A., & Stoelinga, C. N. (2010). Sensory constraints on auditory identification of the material and geometric properties of struck bars. *The Journal of the Acoustical Society of America*, 127(1), 350–360.
- Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in cognitive sciences*, 16(10), 511–518.
- Maclaurin, D. (2016). *Modeling, inference and optimization with composable differentiable procedures* (Doctoral dissertation).
- Manocha, D., & Lin, M. C. (2009). Interactive sound rendering. *2009 11th IEEE International Conference on Computer-Aided Design and Computer Graphics*, 19–26.
- Markel, J. D., & Gray, A. H. (1976). Formulations. *Linear prediction of speech* (pp. 18–41). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-66286-7_2
- Marr, D. (1982). *Vision : A computational investigation into the human representation and processing of visual information*. W. H. Freeman; Company.
- McAdams, S., Roussarie, V., Chaigne, A., & Giordano, B. L. (2010). The psychomechanics of simulated sound sources: Material properties of impacted thin plates. *The Journal of the Acoustical Society of America*, 128(3), 1401–1413.
- McAdams, S. E. (1984). *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university.
- McDermott, J. (2004). Psychophysics with junctions in real images. *Perception*, 33(9), 1101–1127.
- McDermott, J., Weiss, Y., & Adelson, E. H. (2001). Beyond junctions: Nonlocal form constraints on motion interpretation. *Perception*, 30(8), 905–923.

- McDermott, J. H., Ellis, D. P., & Kawahara, H. (2012). Inharmonic speech: A tool for the study of speech perception and separation. *SAPA-SCALE Conference*.
- McDermott, J. H., & Oxenham, A. J. (2008). Spectral completion of partially masked sounds. *Proceedings of the National Academy of Sciences*, *105*(15), 5939–5944.
- McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature neuroscience*, *16*(4), 493–498.
- McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940.
- McDermott, J. H., Wroblewski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences*, *108*(3), 1188–1193.
- McPherson, M. J., & McDermott, J. H. (2020). Time-dependent discrimination advantages for harmonic sounds suggest efficient coding for memory. *Proceedings of the National Academy of Sciences*, *117*, 32169–32180.
- McWalter, R., & McDermott, J. H. (2018). Adaptive and selective time averaging of auditory scenes. *Current Biology*, *28*(9), 1405–1418.
- McWalter, R., & McDermott, J. H. (2019). Illusory sound texture reveals multi-second statistical completion in auditory scene analysis. *Nature communications*, *10*(1), 1–18.
- Micheyl, C., & Oxenham, A. J. (2010). Objective and subjective psychophysical measures of auditory stream integration and segregation. *Journal of the Association for Research in Otolaryngology*, *11*(4), 709–724.
- Mill, R. W., Böhm, T. M., Bendixen, A., Winkler, I., & Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLoS computational biology*, *9*(3), e1002925.
- Misra, A., Wang, G., & Cook, P. R. (2009). Tapestry: A new way to design sound. *Proceedings of the 17th ACM international conference on Multimedia*, 1033–1036.

- Młynarski, W., & McDermott, J. H. (2019). Ecological origins of perceptual grouping principles in the auditory system. *Proceedings of the National Academy of Sciences*, 116(50), 25355–25364.
- Moore, B. C., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *The Journal of the Acoustical Society of America*, 80(2), 479–483.
- Morgenstern, Y., & Kersten, D. J. (2017). The perceptual dimensions of natural dynamic flow. *Journal of Vision*, 17(12), 7–7.
- Morse, P. M., & Ingard, K. U. (1986). *Theoretical acoustics*. Princeton university press.
- Nakatani, T., & Okuno, H. C. (1998). Sound ontology for computational auditory scene analysis. *AAAI-98 Proceedings: Sound Understanding*, 1004–1010. <https://www.aaai.org/Papers/AAAI/1998/AAAI98-142.pdf>
- Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn & D. N. Osherson (Eds.), *An invitation to cognitive science: Visual cognition*. MIT Press.
- Nix, J., & Hohmann, V. (2007). Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE transactions on audio, speech, and language processing*, 15(3), 995–1008.
- O’Brien, J. F., Shen, C., & Gatchalian, C. M. (2002). Synthesizing sounds from rigid-body simulations. *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 175–181.
- O’Reilly, C., & Harte, N. (2017). Pitch tracking of bird vocalizations and an automated process using yin-bird. *Cogent Biology*, 3(1), 1322025. <https://doi.org/10.1080/23312025.2017.1322025>
- Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9(4), 441–474. [https://doi.org/https://doi.org/10.1016/0010-0285\(77\)90016-0](https://doi.org/https://doi.org/10.1016/0010-0285(77)90016-0)

- Pariente, M., Cornell, S., Cosentino, J., Sivasankaran, S., Tzinis, E., Heitkaemper, J., Olvera, M., Stöter, F.-R., Hu, M., Martín-Doñas, J. M., et al. (2020). Asteroid: The pytorch-based audio source separation toolkit for researchers. *arXiv preprint arXiv:2005.04132*.
- Pirker, G., Wohlmayr, M., Petrik, S., & Pernkopf, F. (2011). A pitch tracking corpus with evaluation on multipitch tracking scenario. *Twelfth Annual Conference of the International Speech Communication Association*.
- Pruvost, L., Scherrer, B., Aramaki, M., Ystad, S., & Kronland-Martinet, R. (2015). Perception-based interactive sound synthesis of morphing solids' interactions. *SIGGRAPH Asia 2015 Technical Briefs*, 17.
- Raghuvanshi, N., & Lin, M. C. (2006). Interactive sound synthesis for large scale environments. *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, 101–108.
- Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning*. The MIT Press. <https://doi.org/10.7551/mitpress/3206.001.0001>
- Rath, M., & Rocchesso, D. (2005). Continuous sonic feedback from a rolling ball. *IEEE MultiMedia*, 12(2), 60–69.
- Rayleigh, J. W. S. B. (1896). *The theory of sound* (Vol. 1). Macmillan.
- Reiss, J. D. et al. (2019). Real-time synthesis of sound effects caused by the interaction between two solids. *Audio Engineering Society Convention 146*.
- Ren, Z., Yeh, H., & Lin, M. C. (2010). Synthesizing contact sounds between textured models. *2010 IEEE Virtual Reality Conference (VR)*, 139–146.
- Ren, Z., Yeh, H., & Lin, M. C. (2013). Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, 32(1), 1.
- Riede, T., Borgard, H. L., & Pasch, B. (2017). Laryngeal airway reconstruction indicates that rodent ultrasonic vocalizations are produced by an edge-tone mechanism. *Royal Society Open Science*, 4(11), 170976.
- Rocchesso, D., & Fontana, F. (2003). *The sounding object*. Mondo estremo.

- Saddler, M. R., Gonzalez, R., & McDermott, J. H. (2021). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature communications*, *12*(1), 1–25.
- Saunders, J. A., & Knill, D. C. (2001). Perception of 3d surface orientation from skew symmetry. *Vision research*, *41*(24), 3163–3183.
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. [https://doi.org/https://doi.org/10.1016/S1364-6613\(00\)01506-0](https://doi.org/https://doi.org/10.1016/S1364-6613(00)01506-0)
- Schooneveldt, G. P., & Moore, B. C. (1989). Comodulation masking release (cmr) as a function of masker bandwidth, modulator bandwidth, and signal duration. *The Journal of the Acoustical Society of America*, *85*(1), 273–281.
- Serafin, S. (2004). *The sound of friction: Real time models, playability and musical applications* (Doctoral dissertation). Department of Music, Stanford University.
- Serra, X., & Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, *14*(4), 12–24.
- Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization*. Routledge.
- Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, *12*(5), 182–186.
- Slaney, M., & Lyon, R. F. (1993). On the importance of time-a temporal representation of sound. *Visual representations of speech signals*, 95116.
- Snell, R. C., & Milinazzo, F. (1993). Formant location from lpc analysis data [Actually got from here: <https://www.mathworks.com/help/signal/ug/formant-estimation-with-lpc-coefficients.html>]. *IEEE transactions on Speech and Audio Processing*, *1*(2), 129–134.
- Sprouse, R. L. (2013). Rsprouse/ksyn: Dennis klatt’s speech synthesis system, updated with a python interface. <https://github.com/rsprouse/ksyn>
- Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

- Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature neuroscience*, 9(4), 578–585.
- Stoelinga, C. N. (2007). *A psychomechanical study of rolling sounds* (Doctoral dissertation). ENSTA ParisTech.
- Stöter, F.-R., & Liutkus, A. (2021). *Open-unmix-pytorch umx-l* (Version 1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.5069601>
- Stöter, F.-R., Uhlich, S., Liutkus, A., & Mitsufuji, Y. (2019). Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*. <https://doi.org/10.21105/joss.01667>
- Stuhlmüller, A., Taylor, J., & Goodman, N. (2013). Learning stochastic inverses. *Advances in neural information processing systems*, 26, 3048–3056.
- Szabó, B. T., Denham, S. L., & Winkler, I. (2016). Computational models of auditory scene analysis: A review. *Frontiers in Neuroscience*, 10, 524.
- Taylor, A. M., & Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, 280(3), 221–236.
- Thines, G., Costall, A., & Butterworth, G. (2013). *Michotte’s experimental phenomenology of perception*. Routledge.
- Thompson, S. K., Carlyon, R. P., & Cusack, R. (2011). An objective measurement of the build-up of auditory streaming and of its modulation by attention. *Journal of Experimental Psychology: Human Perception and Performance*, 37(4), 1253.
- Thoret, E., Aramaki, M., Gondre, C., Kronland-Martinet, R., & Ystad, S. (2013). Controlling a non linear friction model for evocative sound synthesis applications. *International Conference on Digital Audio Effects (DAFx)*, XX.
- Thoret, E., Aramaki, M., Kronland-Martinet, R., Velay, J.-L., & Ystad, S. (2014). From sound to shape: Auditory perception of drawing movements. *Journal of Experimental Psychology: Human Perception and Performance*, 40(3), 983–994.
- Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the

- real world. *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 23–30.
- Tougas, Y., & Bregman, A. S. (1985). Crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance*, 11(6), 788.
- Traer, J., Cusimano, M., & McDermott, J. H. (2019). A perceptually inspired generative model of rigid-body contact sounds. *The 22nd International Conference on Digital Audio Effects (DAFx-19)*.
- Traer, J., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48), E7856–E7865.
- Traer, J., Norman-Haignere, S. V., & McDermott, J. H. (2021). Causal inference in environmental sound recognition. *Cognition*, 214, 104627.
- Truax, B. (1978). *Handbook for acoustic ecology*. World Soundscape Project, Simon Fraser University, ARC Publications.
- Tucker, S., & Brown, G. J. (2002). Investigating the perception of the size, shape and material of damped and free vibrating plates. *University of Sheffield, Department of Computer Science Technical Report CS-02-10*.
- Turner, R. E. (2010). *Statistical models for natural sounds* (Doctoral dissertation). UCL (University College London).
- Uhlich, S., & Mitsufuji, Y. (2020). Open-unmix for speech enhancement (umx se). *Zenodo*. <https://doi.org/10.5281/zenodo.3786908>
- Van Den Doel, K., Kry, P. G., & Pai, D. K. (2001). Foleyautomatic: Physically-based sound effects for interactive simulation and animation. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 537–544.
- van den Doel, K., Kry, P. G., & Pai, D. K. (2001). Foleyautomatic: Physically-based sound effects for interactive simulation and animation. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 537–544.
- van den Doel, K., & Pai, D. K. (1996). Synthesis of shape dependent sounds with physical modeling.

- Van Noorden, L. S. (1975). Temporal coherence in the perception of tone sequences. *PhD thesis, Eindhoven University of Technology*.
- Wagemans, J. (2015). Historical and conceptual background: Gestalt theory. In J. Wagemans (Ed.), *Oxford handbook of perceptual organization*. Oxford University Press.
- Wallach, H., Newman, E. B., & Rosenzweig, M. R. (1949). A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, 21(4), 468–468.
- Wang, D. L., & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE transactions on neural networks*, 10(3), 684–697.
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917), 392–393.
- Warren, R. M., Obusek, C. J., & Ackroff, J. M. (1972). Auditory induction: Perceptual synthesis of absent sounds. *Science*, 176(4039), 1149–1151.
- Warren, W. H. (2021). Information is where you find it: Perception as an ecologically well-posed problem. *i-Perception*, 12(2), 20416695211000366.
- Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human perception and performance*, 10(5), 704.
- Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation* (Doctoral dissertation). Stanford University.
- Weiss, Y. (1997). Smoothness in layers: Motion segmentation using nonparametric mixture estimation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 520–526.
- Weiss, Y. (1998). *Bayesian motion estimation and segmentation* (Doctoral dissertation). Massachusetts Institute of Technology.
- Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature neuroscience*, 5(6), 598–604.

- Wessel, D. L. (1979). Timbre space as a musical control structure. *Computer music journal*, 45–52. <https://www.jstor.org/stable/pdf/3680283.pdf>
- Whiteley, L., & Sahani, M. (2012). Attention in a bayesian framework. *Frontiers in human neuroscience*, 6, 100.
- Wilczynski, W., Zakon, H. H., & Brenowitz, E. A. (1984). Acoustic communication in spring peepers. *Journal of Comparative Physiology A*, 155(5), 577–584.
- Wisdom, S., Erdogan, H., Ellis, D. P., Serizel, R., Turpault, N., Fonseca, E., Salamon, J., Seetharaman, P., & Hershey, J. R. (2021). What’s all the fuss about free universal sound separation data? *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 186–190.
- Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., & Hershey, J. (2020). Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems*, 33, 3846–3857.
- Woods, K. J., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology*, 25(17), 2238–2246.
- Woods, K. J., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, 115(14), E3313–E3322.
- Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79(7), 2064–2072.
- Wu, Y., Manilow, E., Deng, Y., Swavely, R., Kastner, K., Cooijmans, T., Courville, A., Huang, C.-Z. A., & Engel, J. (2022). MIDI-DDSP: Detailed control of musical performance via hierarchical modeling. *International Conference on Learning Representations*. <https://openreview.net/forum?id=UseMOjWENv>
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2.
- Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, 6(10), eaax5979.

- Yildirim, I., Siegel, M., & Tenenbaum, J. (2020). Physical object representations for perception and cognition. In D. Poeppel, G. R. Mangun, & M. S. Gazzaniga (Eds.), *The cognitive neurosciences* (6th ed., pp. 399–409). MIT Press.
- Yildirim, I., Siegel, M. H., & Tenenbaum, J. B. (2016). Perceiving fully occluded objects via physical simulation. *Proceedings of the 38th annual conference of the cognitive science society*.
- Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: Analysis by synthesis? *Trends in cognitive sciences*, 10(7), 301–308.
- Zahorik, P., & Wightman, F. L. (2001). Loudness constancy with varying sound source distance. *Nature neuroscience*, 4(1), 78–83.
- Zheng, C., & James, D. L. (2011). Toward high-quality modal contact sound. *ACM Transactions on Graphics (TOG)*, 30, 38.
- Zhu, S.-C. (1999). Embedding gestalt laws in markov random fields. *IEEE transactions on pattern analysis and machine intelligence*, 21(11), 1170–1187.