Cognitive Audio Interfaces: Mediating Sonic Information With an Understanding of How We Hear

Ishwarya Ananthabhotla, David B Ramsay, Clement Duhart, Joseph A Paradiso

Department: Head
Editor: Name, xxxx@email

Cognitive Audio Interfaces: Mediating Sonic Information with an Understanding of How We Hear

Ishwarya Ananthabhotla*
Responsive Environments, MIT Media Lab

David B. Ramsay*
Responsive Environments, MIT Media Lab

Clément Duhart
Responsive Environments, MIT Media Lab
Léonard de Vinci Pôle Universitaire, Research Center

Joseph A. Paradiso
Responsive Environments, MIT Media Lab

***Abstract*—**
In a world of rich, complex, and demanding audio environments, intelligent systems can mediate our interaction with the sounds around us – both to enable meaningful, aesthetic experiences and to transition work from humans to computational agents. Drawing from several years of our research, we suggest that the design of such systems must be driven by a deep understanding of auditory cognition. In this article, we discuss two concrete approaches we take towards cognition-informed interface design – one that begins with sounds themselves to form explicit, contextualized, cognitive models, built on the foundations of large-data parsing infrastructure; and one that begins with the individual, built from intuition surrounding the influence of cognitive state on perception. We point towards an unexplored and compelling future at their intersection.

INTRODUCTION

It's a Friday evening and you are at an office party, chatting away with the colleague standing nearest to you. It is loud and crowded, and you are focused on tuning in to your colleague, suppressing the vibrant voices and conversations around you. Suddenly, you turn around, compelled to address the source of the sound that has

 Published by the IEEE Computer Society

momentarily distracted you — someone has just mentioned your name. Fascinatingly, despite your inability to recall anything the speaker has said prior to this moment, it is clear that the utterance has caught your attention.

This anecdote underscores the complex, hierarchical nature of auditory cognition. In William Gaver's seminal 1993 work, he introduces the notion of "everyday listening", suggesting that our interaction with ambient sounds in our day-to-day environments — unlike speech understanding or musical listening — is determined by an interplay of *gestalt processing* and *acoustic processing*: gestalt, whereby our perception of a sound is shaped by its semantic associations, emotions, sound-producing actions and events, spatial congruence, and location context; and acoustic, whereby our understanding of sounds is driven by texture, timbre, pitch, intensity, or other spectral cues [1]. Over the last two decades, research efforts in empirical psychology, cognitive science, and neural imaging have reinforced these early ideas, elucidating the role that these concepts play in *how* we hear: what we perceive, what we attend to, and what we remember. For example, studies involving the measurement of Event-Related Potentials (ERPs) in the brain demonstrate that pre-attentive responses are invoked due both to semantic novelty (like an animal sound appearing amidst a series of urban sounds) and acoustic novelty (like a sudden, loud sound) [2]. Furthermore, a listener's intuition about the cause of a sound (the sound source) and contextual cues dominates their use of acoustic cues in audio categorization tasks, opting for the latter only when the source is deemed uncertain [3]. Cognitive tests have demonstrated that the emotionality of a sound plays a significant role in memory formation, even when the source is unclear to the listener [4].

To understand even the most basic aspects of how sounds interact with our cognition, we must consider what caused the sound, how ambiguous that cause is, expectations related to that sound in a given context, and, in virtual environments, the realism of different aspects of the rendering process. These aspects of *sounds themselves* are fundamental in shaping how they will be perceived.

This is only part of the story, however. While both gestalt and acoustic properties factor into the way a sound alters our experience or the likelihood that it rises to the level of our conscious recognition, our individual cognitive state also factors heavily into our experience [5]. Our level of focus, alertness, and excitement modulate the likelihood that we'll notice important background sounds, the contextual or spatial incongruence between sound objects, or changes in a sound we expect to occur. If an individual is afraid, for example, they are more likely to pay attention to sounds they perceive as threatening, or to interpret otherwise neutral sounds as threatening. If an individual is busy at work, focused on studying or reading, they may not attend to sounds that might otherwise be considered extremely salient, such as alarms or notifications. These large, dynamic perceptual shifts depend on an individual's momentary, internal cognitive state.

While the literature has shed light on such structural aspects of auditory cognition, the embodiment of these principles in HCI systems has proven elusive. We find ourselves in an increasingly rich and complex auditory world of greater social contact, urban density, and access to novel experiences. Our interactions with this world are enabled by ubiquitous infrastructure that allows us to capture, store, and stream more audio than ever before. In our research, the systems that mediate our interaction with these sonic environments often share two broad objectives — (1) to offload tedious labor to intelligent computational entities, such as the task of searching for instances of wildlife in large databases of ecological recordings, and (2) to create meaningful, aesthetic experiences, such as facilitating a sense of immersion in nature or producing a "summary soundtrack" from an ambient recording that evokes nostalgia of a time and place.

In pursuit of these objectives, we suggest that a deep understanding of auditory cognition must underpin system design. We posit that the first of these objectives can lie within the domain of statistical models that can scale over large datasets, applying technical analysis to sounds to extract cognitive relationships that are common across all users, regardless of individual mental state. The second can be built on a foundation of a designer's intuition about user perception and how it interacts with mental state and holistic

experience; designers either attempt to influence the user's mental state or consider it a feature that supports their intended goal. Literature in psychology and in the cognitive sciences is abundant with well-established causal links that help us form a scaffolding for how to advance our intuition in these domains, but distilling this information and integrating it into the interfaces of the future is a challenging exercise.

In this article, we discuss the practical artifacts we have built to embody these ideas. We discuss our early explorations in isolating, identifying, and quantifying sound percepts in the context of real-world, ubiquitous audio; we discuss the potential for novel interfaces driven by implicit design choices reflecting intuition about individual perception and mental state; and we finally discuss the explicit models we have built to aid in the design of shared experiences, which serve as a strong prior for individualization of a sonic environment. We close by pointing towards a compelling technological future that exists at the horizons of advances in cognitive modeling. Through this article, we suggest the value of a cognition-informed approach to interface design in the face of pervasive audio, and hope to initiate a dialogue surrounding the scope for research within this emerging sub-field.

1. Parsing Large-Scale, Ubiquitous Audio Datasets

As audio infrastructure has become cheaper and smaller, we are now able to collect, stream, and manipulate large sets of audio data from natural environments. These large datasets are an attractive resource for new kinds of analysis that can aid our efforts in deriving the statistical relationships between our auditory worlds and our cognition, and enabling, in turn, more intelligent interfaces. This analysis requires us to extract and isolate sound objects and events, identify them robustly, and measure their influence on common aspects of perception. Here, we provide illustrations of these analyses in two very different auditory environments.

Tidmarsh Audio

As part of a major reclamation project of commercialized marshland, the Tidmarsh Living Observatory was formed to study and support public understanding of wildlife restoration on a marsh in Southern Massachusetts. Our research group has instrumented this area with hundreds of sensors; twenty-four of those sensors are custom microphones designed for the harsh environment, clustered in different areas of the vast marsh. All of these microphones are streaming their data off-site using a custom API that allows them to be used in real-time. A collection of years of historical audio data can also be easily accessed through this API [6].

Wide-area Source Separation The setting of these twenty-four microphones streams makes identification and reconstruction of the aural landscape challenging – a single real sound, like the call of a single goldfinch, will appear in multiple streams with different delays; simple playback of the microphone recordings as virtual objects will cause artifacts. To correct for this, we have developed an efficient method to translate microphone streams into sound objects by extracting foreground sounds and estimating their location amongst the microphones [7]. Typically source separation and localization is approached with a tight cluster of microphones; addressing this problem for sparse, wide-area microphones results in a unique solution to meet real-world constraints.

Tidzam The most useful information for predicting whether a sound is important and how it will be perceived is to identify what it is. Tidzam is a deep learning classification engine that has been deployed in the marsh to identify common wetland sounds [8]. The Tidzam engine runs continuously on the Tidmarsh microphone streams, and is used to label the sounds of insects, frogs, rain, wind, airplanes, and 15 different species of bird call with 93% accuracy in real-time. The engine and its corresponding web-interface, Tidplay, is used to crowd-source and validate annotations from experts, bootstrapping and refining the classifier with time.

Audio Summarization

The Tidzam and Wide-Area Source Separation projects allow us to isolate and label sound objects in the context of a marsh. Labels don't tell us, however, how a sound is perceived– whether it will catch our attention, whether it is a distraction, how it contributes to subtle changes in our sense

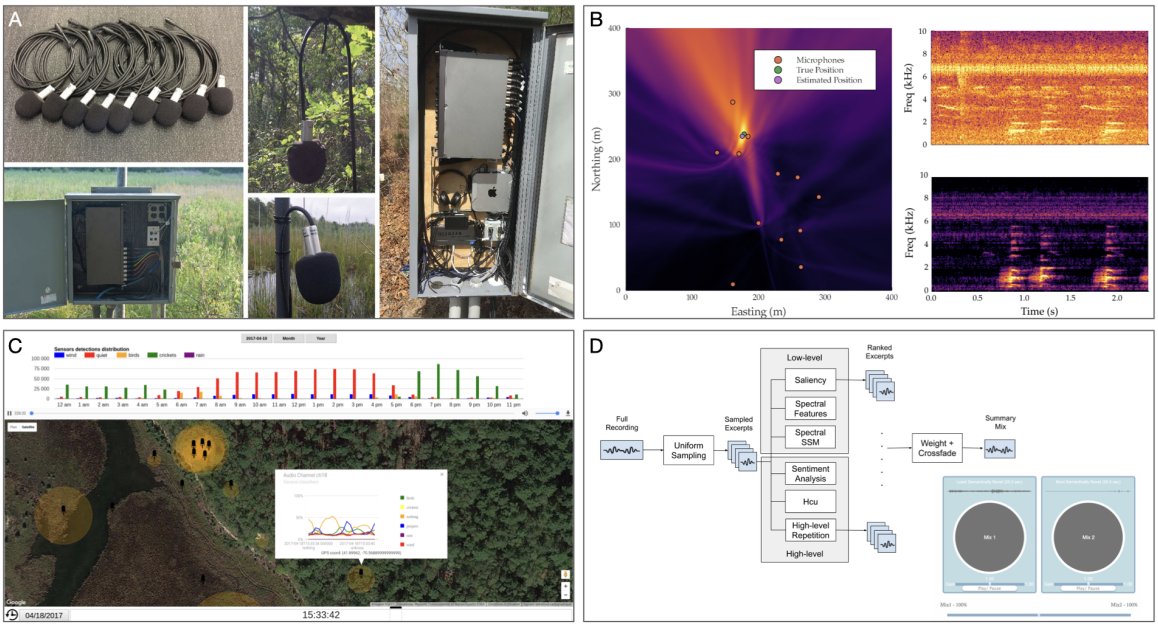

Figure 1. From Microphone Streams To Sound Objects: the Tidmarsh project includes 24 microphones installed over a large wetland area (A). We use these microphone streams to identify, isolate, and locate foreground sounds (B); the image shows a frog recording in the upper right, a probability map of its location amongst the microphones on the left, and the results of a low-rank subspace extraction process to isolate it from the background in the lower right. We also run a real-time classifier called Tidzam on all streams (C). Finally, as part of our Audio Summarization work, we attempt to classify the perceptual impact of different sounds given sound labels and acoustic information (D).

of envelopment and presence, and whether it alters our emotional and cognitive states.

The Audio Summarization project attempts to measure these aspects of perception for everyday sounds across typical human contexts. In this work, we combine gestalt features – based on sound labels and measures of semantic relatedness – with acoustic analysis, to curate extended ambient audio recordings into short "summaries". We then ask human listeners to assign perceptual descriptors to the resulting presentations, such as whether the summary might be useful as a background track for studying or sleeping, or be considered nostalgic, or simply distracting. In this way, we can map the semantic and acoustic analysis to perceptual outcomes. We find that the tool surfaces diverse collections of sounds across the long-term environmental recordings that point to different regions of our perceptual outcome space.

The Future of Parsing Ubiquitous Audio

We've reviewed here our effort to deconstruct and parse large-scale audio datasets in ways that map directly to human perception; we translate microphone streams to sound objects, locate them in space, label them, and explore some basic characteristics of how they interact with perception. This type of analysis can be used to design virtual and real experiences of spaces– highlighting specific types of sounds or emotions, augmenting our real-time perception in natural environments, and curating large swaths of content in ways that were once reserved for human listeners.

Valuable future work takes one of two forms; generalizing and scaling the tools to sense, separate, and label foreground sound objects, and advancing the state of perceptual prediction for sounds and sound contexts.

In the first case, sensing infrastructure can take many new and interesting forms. Low-power wearables with embedded microphones and lightweight neural networks will soon be able

to instantly and accurately identify sound sources in the wearer's environment; however there is still a lot of effort required to take traditional successes in this space, like Google's Audioset networks for sound recognition [9], and turn them into scalable solutions for sound identification that can be engineered to preserve accuracy in constrained computational contexts with less ideal input data. Scalable source separation and recognition continues to be an area of research.

In the second case, we require new techniques to gauge the impact of a sound in isolation and in context, and new modeling strategies that perform well with the irreducible uncertainty and noise that define any attempt to quantify human experience. We will discuss both our attempts to tackle this problem and projections for future work in this space in later sections.

Ultimately, virtual and augmentative reality devices will be able to perform real-time spatial sound analysis and rendering, isolating sound objects for input into a statistical model. Powerful combinations of such technologies will be able to monitor and gauge our responses to these sounds, both actively and passively, in order to build closed-loop systems that are capable of predicting and mediating the effects of our sonic environment in real-time.

2. Frontiers in User Interface Design

The above projects demonstrate a methodical approach to breaking down our auditory experience, beginning with sound percepts – this assumes a static, shared notion of human perception. This type of approach, however, fails to capture the dynamics we experience as a result of our individual cognitive state – our attention, our emotions, our mental load, and other similar factors. We present two examples from our work which showcase the ways in which traditional notions of interfaces can be driven in new directions based on an intuitive understanding of how perception itself can be modified or extended depending on cognitive state — by creating a hyper-awareness of our own faculties, pushing them to their limits, or exploring what it takes to fool them – and identifying, as a result, untapped spaces for new behaviors to emerge.

PhoxEars and HearThere

As a part of a series of work merging artistic practice with the newest frontiers in HCI, we developed PhoxEars — a device consisting of a helmet with two parabolic microphones attached as "ears", whose positions a user can independently control with joysticks [10]. Based on a custom bone conduction headset, the user-controlled ears overlay highly directional sound sources on top of the user's natural experience of the soundscape.

The evolution of this work led us to develop a more comprehensive system, known as HearThere [11]. HearThere users wear a bone-conduction headset that overlays virtual sounds over their natural hearing when they are physically present at Tidmarsh. These virtual sound sources are sourced through the Tidmarsh infrastructure, and are therefore real sounds from the marsh that have been identified and located; the HearThere headset uses a combination of GPS and head tracking to render these sounds as though they are coming from their real-world locations.

This can be considered a sensory prosthesis much like a pair of glasses– a user's natural ability to hear is supplemented and extended instinctively. Moreover, users wearing HearThere experience sensory confusion; it is difficult for them to tell which sounds are real and which are virtual. This prosthesis simply provides the user access to an additional perceptual layer, to which they may naturally direct their conscious attention; when the user turns their visual attention to a particular area of the marsh, for instance, sounds from that area (normally too far to hear) organically blend over their usual hearing.

Both projects richly inform the extension of underlying cognitive principles to future interactions — for example, allowing users to control when and how they direct and heighten their perception on demand creates a way to measure their intention. This iterative process can also be used to understand the value of different rendering methods– which techniques preserve a sense of realism? Which techniques preserve the underlying sense without realism? As we codify intuitive practices into explicit ones, we can design more reactive, interesting, and technically advanced versions of projects like PhoxEars or

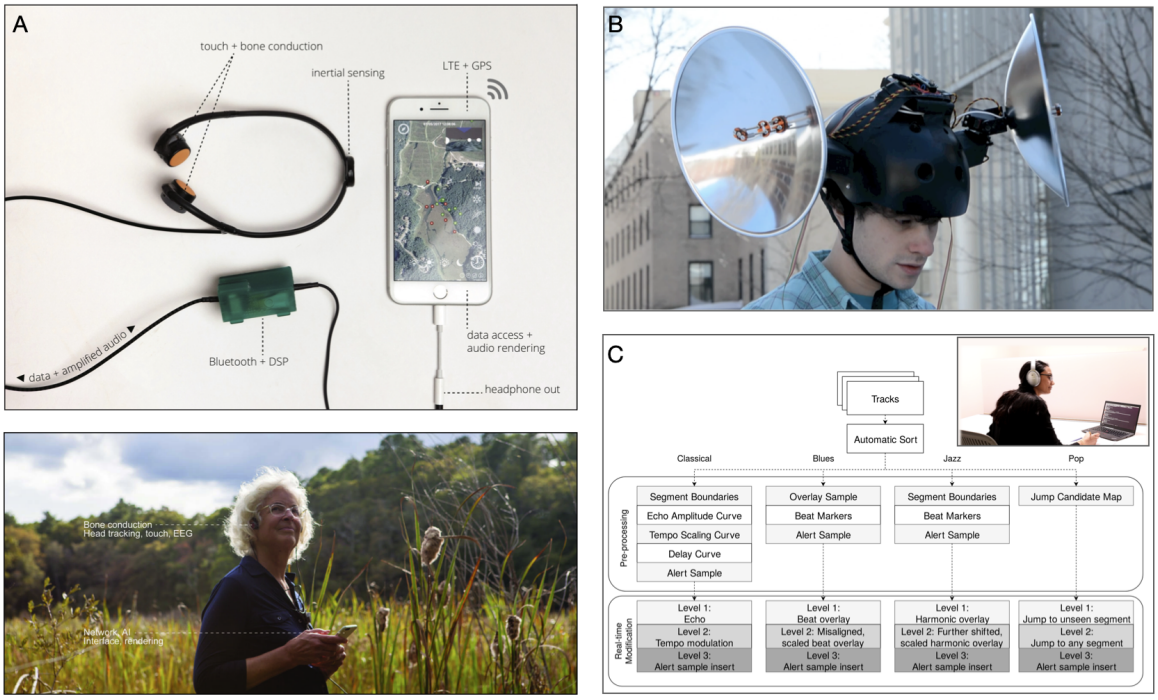

Figure 2. User Interfaces at the Perceptual Boundary: the HearThere project (A) can be used in the marsh environment to naturally extend a user's hearing in a way that requires no conscious effort and is organically managed by our auditory attention. This follows the earlier PhoxEars project (B) which also extended auditory perception through overlay, this time using parabolic microphones on user-controlled motorized gimbals. A final example is the Sound Signaling project (C), which introduces real-time alterations to your music library as a notification tool. These changes are less likely to be noticed by a focused or preoccupied user– fundamentally incorporating their mental state into the design of this notification UI.

HearThere.

SoundSignaling

In SoundSignaling, we introduce a platform for notification delivery (such as from email, social media, or SMS) via subtle, stylistic manipulations in a personal corpus of music [12]. The system will inject genre-specific modifications — such as adding harmonies to a jazz standard, adding extra layers of rhythm to a blues track, or altering the tempo of a classical piece — at varying levels of conspicuity to a stream of music in real-time.

SoundSignaling is an example of design by cognitive heuristics: it operates on the implicit assumption that attentional load modulates awareness of incongruence, an idea borrowed from Stroop's famous colored text experiments [13] and explorations of auditory and visual switching costs [14]; here, the magnitude of "incongruence" is intuited by the designer based on music theory and studies in musical perception. Quantitative and anecdotal data from in-the-wild, long-term studies support the conclusion that SoundSignaling reduces task-switching cost and mediates the intrusion of everyday notifications as a function of cognitive load.

Most participants noted that personalization would be an important facet for the success of such interfaces in the long run. For example, the original design – applied to participants' personal music preferences indiscriminately – evinced both that we habituate to repeated violations of our musical expectations with increased time and use, and that the intensity with which such a violation will draw our attention varies on an individual basis across occupational contexts and musical selection. This result highlights the power of an interface built upon an understanding of how focus-related cognitive states mediate perception,

just as it surfaces the limitations of implicit, heuristic-driven cognitive models.

The Future of Interface Design

The above projects interface deeply and intuitively with individual perception, working in concert with a user's attention and focus to support a holistic experience. We expect that most of the interesting work of the future will combine similarly advanced real-time spatial rendering hardware with models of perception and cognitive state. In the limit, these models will be informed by contextual analysis of environmental data, as we saw in Section 1. It's easy to imagine, for instance, HereThere enabling super-sensitivity to certain types of sounds in the environments– insects, frogs, a particular species of bird– or even to remix the soundscape in a way that promotes a certain mood.

In order to create interfaces that give us more control over or allow us to engage more directly with an individual's experience, we require underlying models that are informed by more than just assumptions and heuristics. These models must contain an explicit structure that can account for contextual information and is capable of evolving with uncertain observations of cognitive phenomena. How do we enable our PhoxEars device, for example, to naturally guide us to regions in our spatial periphery that we identify to be of interest in a way that doesn't disrupt the sense of immersion? How do we devise "musical modifications" that are most appropriate for one's taste in music, that evolve in response to one's increasing mental fatigue with the passage of time in a day?

3. Explorations in Constructing Cognitive Models

Thus far in the discussion, we have alluded to the unifying role of cognitive modeling – in Section 1 we've suggested that the infrastructure needed to extract and analyze individual sound objects while beginning to examine their influence on our perception is a necessary predecessor to more complex models that capture aspects of cognition at greater levels of abstraction; and in Section 2, we've hypothesized that the future of user interfaces to the auditory world rests on our ability to advance the capabilities of these very same models, catalyzing an evolution from implicit design choices to explicit ones. Future tools and interfaces will require powerful predictive insight into the interactions between soundscapes and our internal experience that follows from rigorous and explicit modeling. In this section, we shine a spotlight on the notion of cognitive modeling in isolation, to examine it in greater depth.

Our early explorations have shaped our thinking about the nature of ideal cognitive models, in terms of both the learning methodologies and data labeling strategies best suited to constructing them. We began our work with an investigation of principles at the lowest rungs of the ladder of cognition, which include aspects of unconscious processing such as psychoacoustics, attention, and memory. Here, we describe two preliminary modeling efforts which employ supervised learning strategies and present creative, effective ways of "probing" the cognitive state of interest to obtain labeled data. We then address the expected shortcomings of these approaches in real-world audio contexts, and delineate a likely path forward.

Modeling Auditory Memory

Our group began a research effort in 2018 to develop statistical models to predict memorability from audio features [15], [16]. While the literature indicates that auditory memory is an incredibly complex phenomenon, we were motivated to work towards coarse, general purpose models that might serve as a baseline for enabling futuristic audio interfaces — augmentative systems, for example, that could manipulate sound objects in the environment to make them more or less memorable, or artistic virtual reality platforms that could create immersive, spatial sound presentations that are likely to be remembered long after the experience is over. Drawing from the theory surrounding ecological listening, we hypothesized that causal uncertainty and other gestalt sound properties might also play a critical role in short-term memory formation. To quantify this relationship, we constructed a dataset of 400 sounds, intentionally curated to span the spectrum of source ambiguity. For each sound, we obtained thousands of crowd sourced annotations for the perceived source of the sound, and likert scale

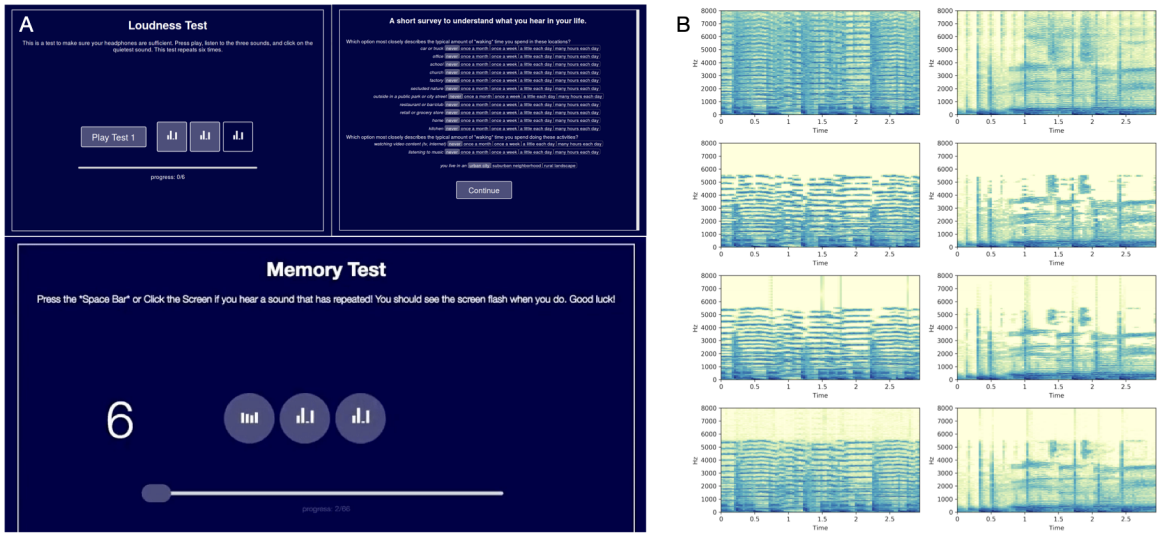

Figure 3. Exploring Cognitive Models: to begin to explore auditory memory, we had thousands of people play a custom audio memory game using a set of 400 curated, everyday sounds; the web interface of the game is shown in (A). Our work also extends to psychoacoustically motivated loss functions for training audio neural networks; in (B) we show the performance of the loss function via original, lossless spectrograms (top row), the compressed and reconstructed versions that preserve perceptually meaningful information (third and bottom rows), and the ideal result from an MP3 codec (second row).

ratings for arousal, valence, sound familiarity, and likelihood of eliciting an image in the listener's mind ("imageability"), pairing them with several extracted spectral features. We also included features based on pre-existing models of our neural response to salient stimuli. After aggregating the source labels into a measure of causal uncertainty based on a novel word-relatedness technique, we created an online memory game to assess sound memorability, and built a regression model to map the gestalt and spectral features to memory outcomes. The data reflects what one might expect — certain sounds are repeatably memorable across individuals (i.e. "man screaming", "women crying", "opera" and "flute"), and certain sounds are not (i.e. causally ambiguous sounds like "truck idling", or synthetically generated sounds). Furthermore, while gestalt features were the best predictors of memorability, auditory memory processes are not well approximated using a simple weighted combination of our feature annotations.

Modeling Listening Perception

In a parallel effort, we also developed a model of perception at the psychoacoustic level, capturing well-understood elements of pre-attentive listening such as frequency masking, the irrelevance of high-frequency content, etc [17]. We envisioned a model that could be used as an error function for optimizing supervised components of upstream tasks, such as real-time speech denoising or audio source separation, by serving as a transform to enable a comparison of a predicted and ground truth sound sample in the perceptual domain. To achieve this, we chose to "sample" a low bit-rate mp3 codec as an approximation of psychoacoustic principles in the average listener, training a supervised model on examples of compressed and lossless track pairs. In this generic form, we show the utility of the model as a means to improve the performance of select upstream tasks as measured by subjective listening tests.

We believe these and similar modeling efforts will form the cornerstone of audio interfaces of the future; both for next generation off-line compression and search of soundscapes based on the principles of human-perception, as well as for real-time analysis and control of our auditory landscape to direct our attention, promote our focus, and strengthen our memory. We will see "sensory prostheses" based on these principles in the near future, at the convergence of

recent advances in low-power, real-time silicon for (1) beamforming and source separation, (2) embedded machine learning, and (3) in-situ noise cancellation and source modification.

The Limitations of Crowd-aggregate, Supervised Modeling

These efforts do suffer, however, from similar limitations: firstly, both models are examples of statistical relationships built by harnessing the power of the crowd, and do not capture individual differences in aural experience. Users that are surveyed for ground truth labels find themselves in constantly shifting, real-world contexts and cognitive states; even if the participants from our modeling exercises above were all in a similar mindset, we know that memory formation is highly dependent on an individual's experience, and listening perception varies from person to person based on age, gender, genetics, and other personal factors. How do we extend these sampling strategies to exhibit relevance on a personal level?

Secondly, there is irreducible uncertainty in both the observation of cognitive state and relationships inherent to the data itself. For example, the models fail to address the likelihood that the same input can lead to different internal experiences, even for the same person; they also fail to account for variability in the quality of data labels which may stem from unreasonable probes of cognitive state – such as asking, for instance, an individual to reflect on their past experience and articulate largely subconscious feelings.

4. The Future of Cognitive Modeling

Looking ahead, we envision two classes of strategies for creating explicit models that are highly personalized. In the first, we begin by constructing a system with implicit, cognitively motivated heuristics; we then create opportunities for user feedback and system teaching, allowing the original heuristics to be informed and mended with time. Concretely, we reprise the example of the SoundSignaling project. We imagine an extension to the project wherein modifications to the music stream evolve with time to be perfectly tailored to a user's musical choices and cognitive state, based on sparse queries for user preference and supplementary multimodal input to the system that adapt the existing causal structure.

In the second, we consider the scenario where identifying such heuristics is not immediately tractable; here, we sample the crowd en masse to build coarse models representing the average listener, and use the heuristics from these models as "priors", or a priori information that forms the basis for an individualized model. Forthcoming work on our Audio Summarization projects illustrates this paradigm, wherein our priors, the simplistic, crowd-wide mappings between deep learning-enabled sound object labels and perceptual descriptors, are adjusted based on user preferences of the summaries generated by the priors in personal auditory environments.

In both approaches, human labels designed to be reflections of cognition are noisy, sparse, and vary in structure with time and context. Building models directly from these perceptual observations becomes a challenging learning problem with fragile outcomes. Instead, we emphasize the use of domain adaptation strategies – namely active learning, uncertainty sampling, and machine teaching – that are coupled with probabilistic, unsupervised learning methods to construct such models. We also look towards labels that include more than just simple self-report; physiological measurements, behavioral observations, and simple models of the effect of an experience on its later recall can all serve to improve our estimates of our unobservable cognitive worlds.

Intent and Causal Reasoning

Finally, we look to cognitive models that describe more than cognitive state. We imagine constructing models that go on to map cognitive state to *intent* – likelihood estimations of physical actions, of expectations of changes in the environments, of needs or wants. We further imagine models that build upon estimations of cognitive state and intent to *reason* about the environment and its actors, much as our minds do on a second-by-second basis. The HearThere project includes a simple example of intent modeling already – at the heart of the system is a causal link suggesting that head orientation and gaze are indicative of the location into which a user would like to aurally "zoom in". More complex models open many more possibilities. A model that can identify the objects in the soundscape that are drawing your

attention could, in tandem with subtle physical cues, actively mitigate (or amplify!) the sensory confusion resulting from the overlay of the bone conduction audio upon your natural hearing. A model that could draw simple inferences about the expected locations of sound sources given appearances in different audio streams could direct you to a particular place or time in the Tidmarsh database based on your interest in a particular animal species. Such models, while seemingly out of the current realm of possibility, can be constructed from the building blocks of simpler models. We suggest that advances in hierarchical Bayesian techniques and causal inference strategies will be the enablers for models at greater and greater levels of abstractions, in line with [18].

Cognitive models will improve by making the implicit explicit; by moving towards probabilistic modeling techniques that measure uncertainty directly, fit to individual users, and learn from better, more naturalistic labeled data. Low-level perceptual models will inevitably extend up the cognitive ladder, towards modeling of intent and reasoning, leaving us with tools that require simpler, less demanding expressions of agency, and interfaces that anticipate our needs and help curate and extend our organic experience of the auditory world.

5. Conclusion

There are exciting opportunities for parsing ubiquitous audio sensor data in the ways the human mind might, and presenting it to users in ways that take advantage of an understanding of their perception, subconscious, and conscious processing. We present examples from our research illustrating an "audio-first" approach to cognitive interface design by deconstructing sensor data into human-meaningful sound objects, an "individual-first" approach built on intuition surrounding the relationship between cognitive state and perception, and early explorations in statistical modeling which we believe form the foundations for greater complexity and novelty in both interfaces and models themselves. We see great opportunities in virtual environments driven by real-world audio data; in sensory prostheses that support our goals by extending and by mediating our perception; and in advanced tools that allow the user to focus completely on expression of agency rather than menial listening tasks. In all cases, a deep understanding of our cognitive processing, embodied as explicit statistical models, drive every design. We hope that our examples motivate a compelling vision of the future, and form a first step towards a significant body of research to come.

REFERENCES

1. W. W. Gaver, "What in the World do we Hear?: An Ecological Approach to Auditory Event Perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.
2. A. Schirmer, Y. H. Soh, T. B. Penney, and L. Wyse, "Perceptual and Conceptual Priming of Environmental Sounds," *Journal of Cognitive Neuroscience*, vol. 23, no. 11, pp. 3241–3253, 2011.
3. M. M. Marcell, D. Borella, M. Greene, E. Kerr, and S. Rogers, "Confrontation Naming of Environmental Sounds," *Journal of Clinical and Experimental Neuropsychology*, vol. 22, no. 6, pp. 830–864, 2000.
4. M. Quirin, M. Kazén, and J. Kuhl, "When Nonsense Sounds Happy or Helpless: the Implicit Positive and Negative Affect Test (IPANAT)," *Journal of Personality and Social Psychology*, vol. 97, no. 3, p. 500, 2009.
5. S. Da Costa, W. van der Zwaag, L. M. Miller, S. Clarke, and M. Saenz, "Tuning in to Sound: Frequency-selective Attentional Filter in Human Primary Auditory Cortex," *Journal of Neuroscience*, vol. 33, no. 5, pp. 1858–1863, 2013.
6. B. Mayton, G. Dublon, S. Russell, E. F. Lynch, D. D. Haddad, V. Ramasubramanian, C. Duhart, G. Davenport, and J. A. Paradiso, "The Networked Sensory Landscape: Capturing and Experiencing Ecological Change Across Scales," vol. 26, pp. 182–209, MIT Press, 2017.
7. S. F. Russell, *Resynthesizing Volumetric Soundscapes: Low-rank Subspace Methods for Soundfield Estimation and Reconstruction*. PhD thesis, Massachusetts Institute of Technology, 2020.
8. C. Duhart, G. Dublon, B. Mayton, and J. Paradiso, "Deep Learning Locally Trained Wildlife Sensing in Real Acoustic Wetland Environment," in *International Symposium on Signal Processing and Intelligent Recognition Systems*, pp. 3–14, Springer, 2018.
9. S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN Architectures for Large-scale Audio Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, IEEE, 2017.

10. R. Kleinberger, G. Dublon, J. A. Paradiso, and T. Machover, “PhoxEars: a Parabolic, Head-mounted, Orientable, Extrasensory Listening Device.,” in *NIME*, pp. 30–31, 2015.
11. G. Dublon, *Sensor(y) Landscapes: Technologies for New Perceptual Sensibilities*. PhD thesis, Massachusetts Institute of Technology, 2018.
12. I. Ananthabhotla and J. A. Paradiso, “SoundSignaling: Realtime, Stylistic Modification of a Personal Music Corpus for Information Delivery,” in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, pp. 1–23, 2018.
13. J. R. Stroop, “Studies of interference in serial verbal reactions.,” *Journal of Experimental Psychology: General*, vol. 121, no. 1, p. 15, 1992.
14. S. Monsell, “Task Switching,” *Trends in Cognitive Sciences*, vol. 7, no. 3, pp. 134–140, 2003.
15. I. Ananthabhotla, D. B. Ramsay, and J. A. Paradiso, “HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924, IEEE, 2019.
16. D. B. Ramsay, I. Ananthabhotla, and J. A. Paradiso, “The Intrinsic Memorability of Everyday Sounds,” in *AES International Conference on Immersive and Interactive Audio*, Audio Engineering Society, 2019.
17. I. Ananthabhotla, S. Ewert, and J. A. Paradiso, “Using a Neural Network Codec Approximation Loss to Improve Source Separation Performance in Limited Capacity Networks,” in *Proceedings of the International Joint Conference on Neural Networks*, 2020.
18. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” *Behavioral and brain sciences*, vol. 40, 2017.

Ishwarya Ananthabhotla is a 4th year PhD candidate in the Responsive Environments Group at the MIT Media Lab, and is interested in problems at the intersection of auditory cognition and machine learning. Ishwarya has an SB and M.Eng in Electrical Engineering and Computer Science from MIT, and can be contacted at ishwarya@media.mit.edu.

David B. Ramsay is a 4th year PhD candidate in the Responsive Environments Group at the MIT Media Lab studying techniques to measure cognitive phenomena and systems that can respond to those measurements. David has a BS in Electrical Engineering and a BA in Music from Case Western Reserve University, as well as an MS from the MIT Media Lab. He can be contacted at dramsay@media.mit.edu.

Clément Duhart is an affiliated researcher in the Responsive Environment Group at the MIT Media Lab. He is currently head of the De Vinci Innovation Center in Paris where he explores modern technologies in human computer interaction, wearable and sensor networks, soft robotics and artificial intelligence with his Masters and PhD students. He can be contacted at duhart@media.mit.edu.

Joseph A. Paradiso joined the MIT Media Laboratory in 1994, where he is the Alexander W. Dreyfoos (1954) Professor in Media Arts and Sciences. He is currently serving as the associate academic head of the MAS Program, and also directs the Media Lab’s Responsive Environments Research Group, which explores how sensor networks augment and mediate human experience, interaction and perception. He can be contacted at joep@media.mit.edu.