Towards “Gestalt” Computation in Sound

Ishwarya Ananthabhotla, David Ramsay, Joseph Paradiso
MIT Media Lab, Cambridge, USA
`{ishwarya, dramsay, joep}@media.mit.edu`

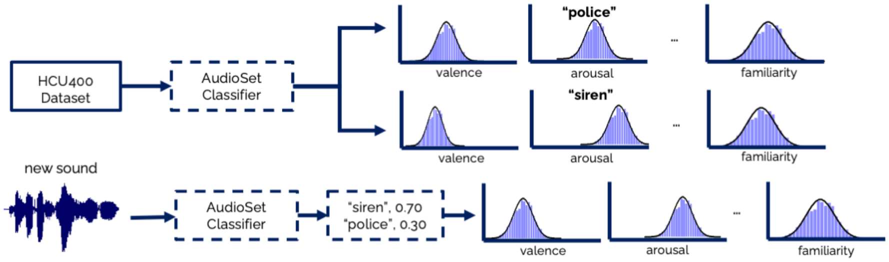

Figure 1: We first bootstrap gestalt property scores to all labels in the AudioSet ontology by running a classifier on the HCU400 dataset *(top)*; we then estimate gestalt property scores for unseen audio examples by first predicting AudioSet labels, and then combining the scores associated with these labels, weighted by the prediction uncertainty *(bottom)*

1 Motivation

Over the last few decades, neuroscientists, cognitive scientists, and psychologists have made strides in understanding the complex processes that define the interaction between our minds and the sounds around us. Their research suggests that we interpret and contextualize sounds through a nuanced dynamic between semantics and acoustics, responding pre-consciously to both the spectro-temporal dynamics of a sound object and the inherent emotionality, memorability, and notions of causal uncertainty we ascribe to it [1–8]. Building statistical models of sound understanding that are aware of these latter, higher-level aspects of cognition could be used to power a suite of compelling, futuristic, and creative experiences in sound - via automatic annotation, manipulation, and generation of audio content. One could imagine using these models to more readily design automated tools for creating audio to accompany film or podcast media that reflects the aesthetic of the content; create immersive soundscapes in virtual reality designed to aid in the practice of mindfulness and wellbeing; and construct intricate sound design work for art installations intended to guide collective perception towards a certain emotional state.

A characterization of the cognitive impacts of sound objects is highly subjective [9] and ultimately demands personalized modeling. Here, however, we suggest that even general purpose, non-personalized estimators that capture crowd-scale information about intrinsic, semantic properties of sound – often referred to in the auditory psychology literature as *gestalt* properties – can be a useful preliminary step for creating experiences. This approach also proves challenging; the cost and difficulty of creating sizeable datasets containing annotations of gestalt properties currently prohibits the use of end-to-end learning to construct these estimators. Thus, in the face of small data, there is a need for effective input representations that are built on top of years of auditory psychology research.

In this paper, we present a simple paradigm for estimating a set of gestalt properties – such as valence and arousal, imageability, causal uncertainty, and memorability – from unseen, real-world audio, using a probabilistic bootstrapping approach that employs an AudioSet [10, 11] classification network as an intermediary. The overall idea behind our approach is that, given limited quantities of annotations of these properties in sound cognition datasets, we can create a robust estimator by first mapping these properties to semantic classes obtained via large, pre-trained networks. We believe this approach is valuable in that it benefits from advancements in sound classification research – the outcomes improve as neural networks for classification improve in performance and label ontologies

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

become more nuanced – but also in that it mimics a well-observed cognitive process [1, 2]: that sound understanding begins with sound source identification, and the uncertainty surrounding it.

2 Approach

To illustrate our approach (summarized in Figure 1), we consider the HCU400 and memorability datasets presented in [12, 13], and aim to scale the hand-labeled annotations of six gestalt properties – arousal, valence, imageability, familiarity, memorability, and confusability – to unseen audio. We achieve this by building a probabilistic mapping between these scores and the 600+ labels in the AudioSet ontology. To build this mapping, we employ a pre-trained AudioSet classification network[1] to obtain the top k label predictions from each audio sample in the HCU400 dataset. Then, we capture the correlation between each AudioSet class and gestalt property. To do this, for a given label and gestalt property, we fit a Gaussian distribution to the property scores across all of the sounds with that label, wherein each set of scores (representing the set of human annotations per sound) is weighted according to the network's label uncertainty. While the labels associated with the sounds in the HCU400 dataset are too sparse to fully cover all of the AudioSet labels, we can exploit the existing class relationships in the AudioSet ontology to meaningfully impute our estimates of the new gestalt properties to the uncharacterized labels: parents adopt mean scores of children, children inherit parent scores. Example results from the complete process can be seen in Table 1.

To calculate gestalt property scores for unseen audio examples, this process can effectively be inverted: the distributions associated with the top k AudioSet labels are combined– weighted by prediction uncertainty– to obtain mean and variance estimates for the unseen audio. Any number of similar weighting heuristics can be applied, contingent on the application context.

Throughout this approach, we treat the uncertainty of the pre-trained AudioSet model as a proxy for human uncertainty in sound source identification. As shown in [14] (reproduced in Figure 2) plotting the audio classes from the HCU400 dataset against the human-rated and artificial causal uncertainty measures demonstrates that it is a reasonable proxy. We use this notion implicitly in the bootstrapping process as we weight the contribution from different instances in the HCU400 dataset by the network prediction uncertainty, mimicking the role of causal uncertainty as the fulcrum between semantic and acoustic processing [1, 2]. We can also use this notion explicitly as a tool for constructing experiences (see Section 3).

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

3 Applications

We give several examples from our prior and on-going research about the value of this approach in providing rudimentary tools for constructing human-meaningful experiences.

In one project, we invite participants to "lifelog" their surroundings with audio recorders for several weeks, and review the audio via short presentations created by using the gestalt property estimators to curate and summarize the audio. We find that the gestalt property analysis drives the creation of audio digests that participants find far more emotional and intimate than those generated through acoustic means alone [14]. Examples of the system's output can be auditioned here.

In another project, we design an optimization procedure to manipulate acoustic properties of a sound example in order to modulate causal uncertainty, and demonstrate that we can do this reliably in a series of perceptual tests [15]. Audio examples can be found here.

In additional on-going work, we create personalized soundscapes from generic, ambient audio recordings to facilitate specific cognitive states (such as a state of focus or relaxation). We layer a learning model on top of the gestalt property analysis to drive a generative engine toward an individual's preference. A demo of the system can be found here.

[1] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

4 Ethical Implications

Pre-conscious sound perception and cognition, as mentioned in the text, is highly idiographic; it may be a function of one's experiences, culture, exposure to sonic environments, and physiology, amongst other factors. Models built to capture shared, crowd-scale notions in sound cognition reflect sampling biases: the datasets herein may be unrepresentative of wider U.S. or global populations and/or fail to capture culturally relevant information, making them less applicable to those from different backgrounds.

References

[1] William W Gaver. How do we Hear in the World? Explorations in Ecological Acoustics. *Ecological Psychology*, 5(4):285–313, 1993.

[2] William W Gaver. What in the World do we Hear?: An Ecological Approach to Auditory Event Perception. *Ecological psychology*, 5(1):1–29, 1993.

[3] James A Ballas. Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology: human perception and performance*, 19(2):250, 1993.

[4] Michael M Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers. Confrontation naming of environmental sounds. *Journal of clinical and experimental neuropsychology*, 22(6):830–864, 2000.

[5] Oliver Bones, Trevor J Cox, and William J Davies. Distinct categorization strategies for different types of environmental sounds. Euronoise, 2018.

[6] Bruno L Giordano, John McDonnell, and Stephen McAdams. Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds. *Brain and cognition*, 73(1):7–19, 2010.

[7] Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16, 2010.

[8] Annett Schirmer, Yong Hao Soh, Trevor B Penney, and Lonce Wyse. Perceptual and conceptual priming of environmental sounds. *Journal of cognitive neuroscience*, 23(11):3241–3253, 2011.

[9] Stephen Ed McAdams and Emmanuel Ed Bigand. Thinking in sound: The cognitive psychology of human audition. In *Based on the fourth workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society.* Clarendon Press/Oxford University Press, 1993.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN Architectures for Large-scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[12] Ishwarya Ananthabhotla, David B Ramsay, and Joseph A Paradiso. HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2019.

[13] David B Ramsay, Ishwarya Ananthabhotla, and Joseph A Paradiso. The Intrinsic Memorability of Everyday Sounds. In *AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[14] Ishwarya Ananthabhotla, David Bradford Ramsay, and Joseph A Paradiso. Cognitive Content Curation: An Audio Summarization Tool Driven by Principles of Auditory Cognition. *Under Review*.

[15] Tal Boger, Ishwarya Ananthabhotla, and Joseph Paradiso. Manipulating causal uncertainty in sound objects. In *Proceedings of ACM Audio Mostly*, 2021.

A Appendix

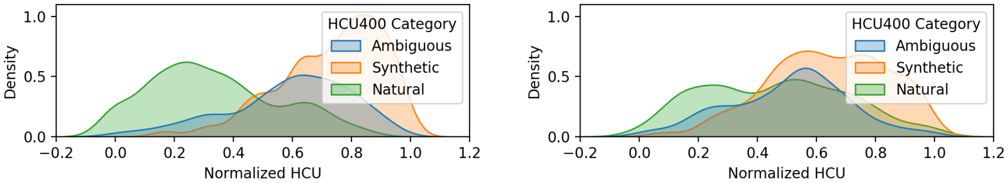

Figure 2: Plots showing the distribution of sounds in the HCU400 dataset labelled with their original categories from [12], namely “Natural", “Ambiguous", or “Synthetic". We compare the separation of these “Natural" and “Synthetic" classes via the human annotated Hcu metric (left) with the proposed, neural network-based approach (right).

Top Scores: "Memorability"	**Top Scores: "Confusability"**
Guitar	Rain on surface
Wail, moan	Pink noise
Fire alarm	Ocean
Baby cry, infant cry	Vibration
Crying, sobbing	Traffic noise, roadway noise
Cough	Idling
Singing	Stream
Whistling	Fire
Chuckle, chortle	Typewriter
Belly laughter	Wind
Baby laughter	Rustling Leaves
Ambulance (siren)	Thump. Thud
Sneeze	Electric shaver, electric razor
Top Scores: “Arousal”	**Top Scores: “Valence”**
Skidding	Acoustic Guitar
Machine gun	Strum
Ambulance (siren)	Wind Chime
Emergency vehicle	Chuckle, chortle
Toot	Giggle
Train horn	Laughter
Fire alarm	Flute
Vehicle horn, car horn, honking	Cello
Growling	Classical Music
Doorbell	Waterfall
Ringtone	Bird call, bird song
Boom	Rain on surface
Roar	Church bell

Table 1: Sound category labels from the AudioSet [11] ontology with top scores for Memorability, Confusability, Arousal, and Valence, determined by bootstrapping from the small HCU400 dataset [12, 13].