HCU400: AN ANNOTATED DATASET FOR EXPLORING AURAL PHENOMENOLOGY THROUGH CAUSAL UNCERTAINTY

Ishwarya Ananthabhotla, David B. Ramsay*, and Joseph A. Paradiso*

MIT Media Laboratory, Cambridge, MA.

ABSTRACT

The way we perceive a sound depends on many aspects– its ecological frequency, acoustic features, typicality, and most notably, its identified source. In this paper, we present the HCU400: a dataset of 402 sounds ranging from easily identifiable everyday sounds to intentionally obscured artificial ones. It aims to lower the barrier for the study of aural phenomenology as the largest available audio dataset to include an analysis of causal attribution. Each sample has been annotated with crowd-sourced descriptions, as well as familiarity, imageability, arousal, and valence ratings. We extend existing calculations of causal uncertainty, automating and generalizing them with word embeddings. Upon analysis we find that individuals will provide less polarized emotion ratings as a sound's source becomes increasingly ambiguous; individual ratings of familiarity and imageability, on the other hand, diverge as uncertainty increases despite a clear negative trend on average.

Index Terms— auditory perception, causal uncertainty, affect, audio embeddings

1. MOTIVATION

Despite a substantial body of literature, human auditory processing remains poorly understood. In 1993, Gaver introduced an ecological model of auditory perception based on the physics of an object in combination with the class of its sound-producing interaction [1]. He suggests that everyday listening focuses on sound sources, while musical listening focuses on acoustic properties of a sound, and that the difference is experiential. Current research has corroborated this distinction– studies show that listeners primarily group sounds by category of sound-source, sometimes group sounds by location/context, and only in certain conditions favor groupings by acoustic properties [2, 3]. Recent work with open-ended sound labeling demonstrates that limited categorization tasks may encourage more detailed descriptions along valence/arousal axes (i.e. for animal sounds) or using acoustic properties (i.e. for mechanical sounds) if sound-source distinctions are too limited for the categorization task [4].

It has been suggested that non-verbal sounds from a living source are processed differently in the brain than other physical events [5]. Symbolic information tends to underly our characterization of sounds from humans and animals (i.e. yawning, clapping), while acoustic information is relied on for other environmental sounds [6, 7, 8]. Furthermore, in [9] Dubois et al. demonstrated that, for complex scenes, the perception of pleasant/unpleasantness was attributed to audible evidence of human activity instead of measurable acoustic features.

It is clear from the above research that any examination of sound phenomenology must start with a thorough characterization of a sound's interpreted cause. In many cases however, a sound's cause can be ambiguous. In [10] Ballas introduced a measure of causal uncertainty (H_{cu}) based on a large set of elicited noun/verb descriptions for 41 everyday sounds: $H_{cui} = \sum_j^n p_{ij} log_2 p_{ij}$. (For sound i, p_{ij} is the proportion of labels for that sound that fall into category j as decided by experts reviewing the descriptions). He shows a complicated relationship between H_{cu} and the typicality of the sound, its familiarity, the average cognitive delay before an individual is able to produce a label, and the ecological frequency of the sound in his subjects' environment. H_{cu} was further explored in [11] using 96 kitchen sounds. Lemaitre et al. demonstrated that H_{cu} alters how we classify sounds: with low causal uncertainty, subjects cluster kitchen sounds by their source; otherwise they fall back to acoustic features.

In this paper, we introduce the HCU400 dataset– the largest dataset available for studying everyday sound phenomenology. In this dataset, we include 402 sounds that were chosen to (1) capture common environmental sounds from everyday life, and (2) to fully sample the range of causal uncertainty. While many of the sounds in our dataset are unambiguous, over 100 of the sounds are modified to intentionally obscure their source– allowing explicit control of source-dependent effects.

As part of the dataset, we include high-level emotional features corresponding to each sound's valence and arousal, in line with previous work on affective sound measurement [12]. We also account for features that provide other insights into the mental processing of sound– familiarity and image-

*Equal contribution. The authors would like to thank the AI Grant for their financial support of this work.

ability [13, 14]. We explore the basic relationships between all of these features.

Finally, we introduce word embeddings as a clustering technique to extend the original H_{cu}, and apply it to the free response labels we gathered for each sound in the dataset. Deep learning has provided a new tool to represent vast amounts of semantic data in a highly compressed form; these techniques will likely make it possible to model and generalize source-dependent auditory processing phenomena. The HCU400 represents a first step in that direction.

2. DATASET OVERVIEW

The HCU400 dataset consists of 402 sound samples and 3 groups of features: sound sample annotations and associated metadata, audio features, and semantic features. It is available at `github.com/mitmedialab/HCU400`.

2.1. Sourcing the Sounds

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

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

To source unfamiliar/ambiguous sounds, we include a handful of digitally synthesized samples in addition to artificially manipulated everyday sounds. Our manipulation pipeline applies a series of random effects and transforms to our existing samples from the former categories, from which we curated a subset of sufficiently unrecognizable results. Effects include reverberation, time reversal, echo, time stretch/shrink, pitch modulation, and amplitude modulation.

2.2. Annotated Features

We began by designing an Amazon Mechanical Turk (AMT) experiment in which participants were presented with a sound chosen at random. Upon listening as many times as they desired, they then provided a free-text description alongside likert ratings of its familiarity, imageability, arousal, and valence (as depicted by the commonly used self-assessment manikins [12]). The interface additionally captured metadata such as the time taken by each participant to complete their responses, the number of times a given sound was played, and the number of words used in the free-text response. Roughly 12000 data points were collected through the experiment, resulting in approximately 30 evaluations per sound after discarding outliers (individual workers whose overall rankings deviate strongly from the global mean/standard deviation). A reference screenshot of the interface and its included questions can be found at `github.com/mitmedialab/HCU400`.

2.3. Audio Features

Low level features were extracted using the Google VGGish audio classification network, which provides a 128-dimensional embedded representation of audio segments from a network trained to classify 600 types of sound events from YouTube [15]. This is a standard feature extraction tool, and used in prominent datasets. A comprehensive set of standard features extracted using the OpenSMILE toolkit [**?**] is also included.

2.4. Semantic Features

A novel contribution of this work is the automation and extension of H_{cu} using word embeddings and knowledge graphs. Traditionally, these are used to geometrically capture semantic word relationships; here, we leverage the "clustering radius" of the set of label embeddings as a metric for each sound's H_{cu}.

We employed three major approaches to embed each label: (1) averaging all constituent words that are nouns, verbs, adjectives, and adverbs– a common/successful average encoding technique [16]– (2) choosing only the first or last noun and verb, and (3) choosing a single 'head word' for each embedding based on a greedy search across a heavily stemmed version of all of the labels (using the aggressive Lancester Stemmer [17]). In cases where words are out-of-corpus, we auto-correct their spelling, and/or replace them with a synonym from WordNet where available [18]. Labels that fail to cluster are represented by the word with the smallest distance to an existing cluster for that sound (using WordNet path-length). This greedy search technique is used to automatically generate the group of labels used in the H_{cu} calculation. Both Word2Vec [19] and Conceptnet Numberbatch [20] were tested to embed individual words.

After embedding each label, we derived a 'cluster radius' score for the set of labels, using the mean and standard deviation of the distance of each label from the centroid as a baseline method. We also explore (k=3) nearest neighbor intra-cluster distances to reduce the impact of outliers and increase tolerance of oblong shapes. Finally, we calculate the sum

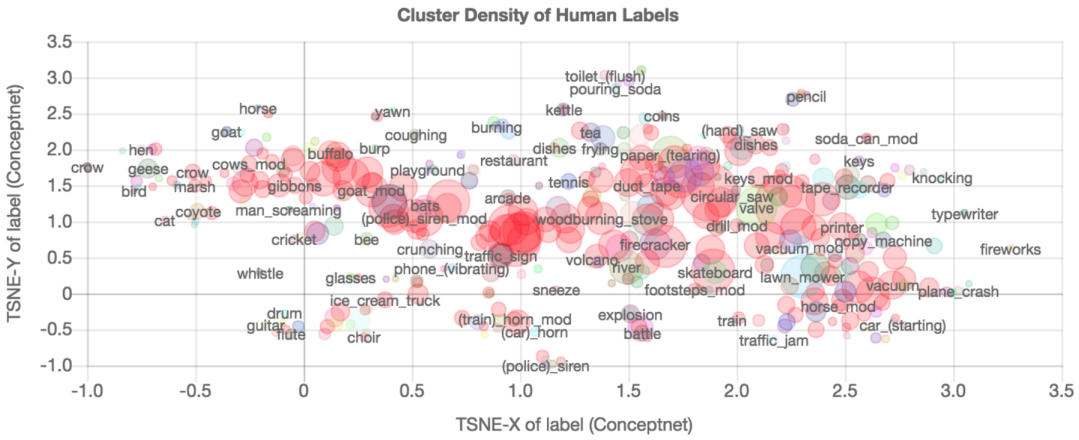

Fig. 1. Average ConceptNet embedding where the radius represents our H_{cu} metric; red bubbles and the '_mod' suffix are used to indicate sounds that have been intentionally modified.

of weighted distance from each label subgroup to the largest 'head word' cluster– a technique which emphasizes sounds with a single dominant label.

We also include a location-based embedding to capture information pertaining to the likelihood of concept co-location in a physical environment. In order to generate a co-location embedding, we implement a shallow-depth crawler that operates on ConceptNet's location relationships ('Located-Near', 'Located-At', etc) to create a weighted intersection matrix of the set of unique nouns across all our labels as a pseudo-embedding. Again, we derive the centroid location and mean deviation from the centroid of the labels (represented by the first unique noun) for a given sound sample.

Given the number of techniques, we compare and include only the most representative pipelines in our dataset. All clustering approaches give a similar overall monotonic trend, but with variations in their derivative and noise. Analysis of cluster labels in conjunction with scores suggests that a distance-from-primary-cluster definition is most fitting. Most embedding types are similar, but we prefer ConceptNet embeddings over others because it is explicitly designed to capture meaningful semantic relationships.

Our clustering results from a Processed ConceptNet embedding are plotted in Figure 1. Intentionally modified sounds are plotted in red, and we see most sounds with divergent labeling fall into this category. Sounds that have not been modified are in other colors– here we see examples of completely unambiguous sounds, like human vocalizations, animal sounds, sirens, and instruments.

3. BASELINE ANALYSIS AND DISCUSSION

First, we find that the likert annotations are reliable amongst online workers, using a split ranking evaluation adapted from [21]. Each of the groups consisted of 50 % of the workers, and

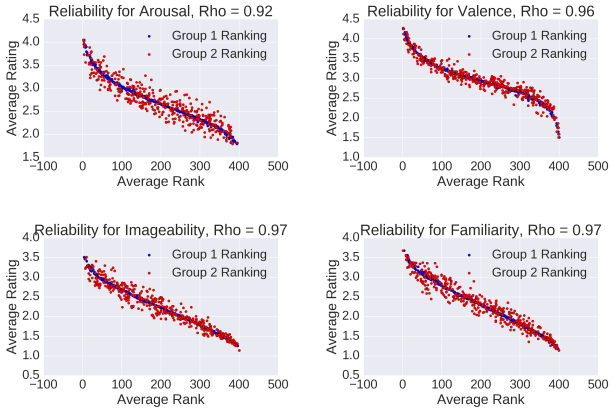

Fig. 2. Split ranking correlation plots and Spearman rank coefficient values for the four likert annotated features.

the mean ranking was computed after averaging N=5 splits. The resulting spearman rank coefficient value for each of the crowd-sourced features is given in Figure 2. This provides the basis for several intuitive trends in our data, as shown by Figure 4 – we find a near linear correlation between imageability and familiarity, and a significant correlation between arousal and valence. We also find a strong correlation between imageability, familiarity, time-based individual measures of uncertainty (such as such "time to first letter" or "num of times played"), and the label-based, aggregate measures of uncertainty (the cluster radii and H_{cu}).

We next see strong evidence of the value of word embeddings as a measure of causal uncertainty – the automated technique aligns well with the split of modified/ non-modified sounds (see Fig. 1) and a qualitative review of the data labels. Our measure also goes one step beyond H_{cu}, as the cluster

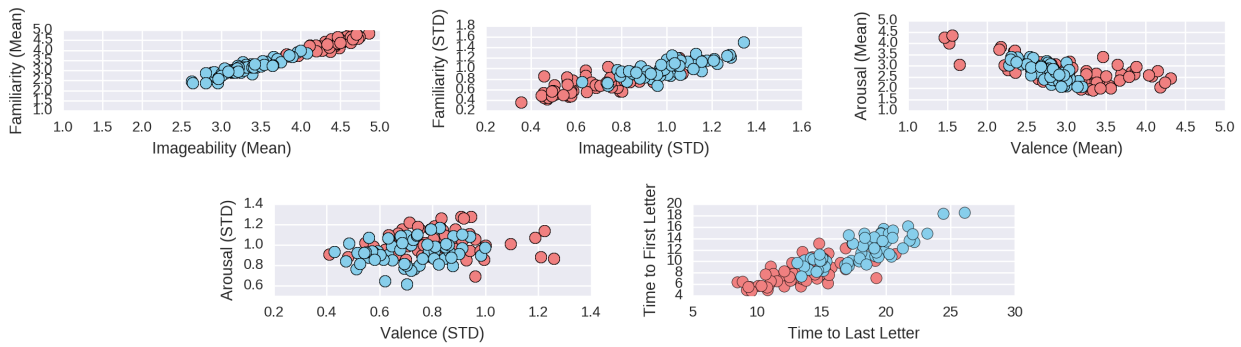

Fig. 3. Feature distributions grouped by extremes in the "Processed CNET" cluster metric; red points represent data at $\leq$ 15th percentile (the most labeling agreement and least ambiguous); blue dots are $\geq$ 85th percentile (high H_{cu}).

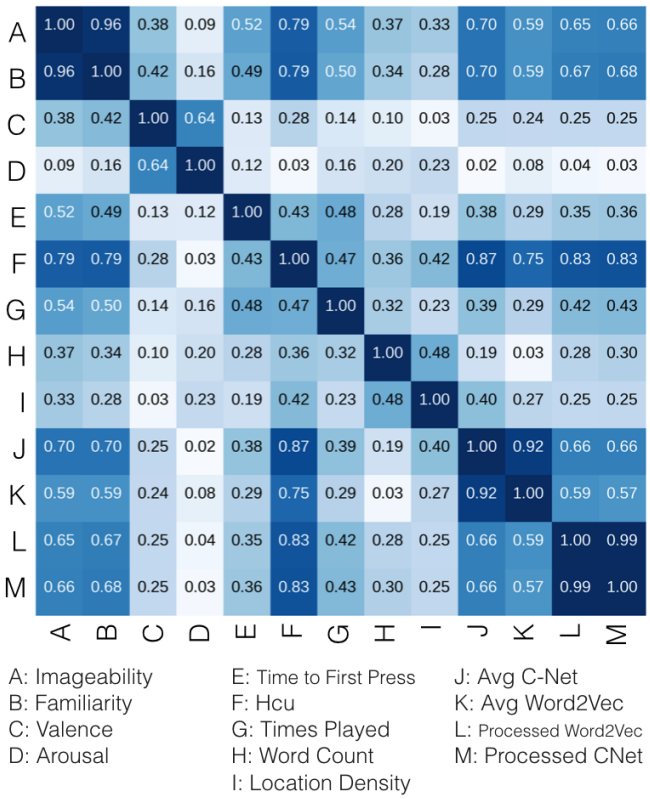

Fig. 4. Correlation Matrix displaying the absolute value of the Pearson correlation coefficient between the mean values of annotated features, metadata, and four representative word embedding based clustering techniques.

centroid assigns representative content to the group of labels. Initial clustering of sounds by their embedded centroids reveals a relationship between clusters and emotion rankings when the source is unambiguous, which could be generalized to predict non-annotated sounds (i.e., sirens, horns, and traffic all cluster together and have very close positive arousal and negative valence rankings; similar kinds of trends hold for clusters of musical instruments and nature sounds).

Furthermore, we use this data to explore the causal relationship between average source uncertainty and individual assessment behavior. In Figure 3, we plot the distributions of pairs of features as a function of data points within the 15th (red) and greater than 85th (blue) percentile of a single cluster metric ("Processed CNET"). It confirms a strong relationship between the extremes of the metric and individual deliberation (bottom right), as reported by [10]. We further find that more ambiguous sounds have less extreme emotion ratings (top right); the data suggest this is not because of disagreement in causal attribution, but because individuals are less impacted when the source is less clear (bottom left). This trend is not true of imageability and familiarity, however; as sounds become more ambiguous, individuals are more likely to diverge in their responses (top center). Regardless, we find a strong downward trend in average familiarity/imageability scores as the source becomes more uncertain (top left).

4. CONCLUSION

It is known that aural phenomenology rests on a complex interaction between a presumed sound source, the certainty of that source, the sound's acoustic features, its ecological frequency, and its familiarity. We have introduced the HCU400– a dataset of everyday and intentionally obscured sounds that reliably captures a full set of affective features, self-reported cognitive features, timing, and free-text labels. We present a new technique to quantify H_{cu} using the distances between word embeddings of free text labels. Our analysis demonstrates (1) the efficacy of a quantified approach to H_{cu} using word embeddings; (2) the quality of our crowd-sourced likert ratings; and (3) the complex relationships between global uncertainty and individual rating behavior, which offers novel insight into our understanding of auditory perception.

5. REFERENCES

[1] William 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] Michael M Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers, "Confrontation naming of environmental sounds," *Journal of clinical and experimental neuropsychology*, vol. 22, no. 6, pp. 830–864, 2000.

[3] Brian Gygi and Valeriy Shafiro, "General functions and specific applications of environmental sound research.," *Frontiers in bioscience: a journal and virtual library*, vol. 12, pp. 3152–3166, 2007.

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

[5] James W Lewis, Frederic L Wightman, Julie A Brefczynski, Raymond E Phinney, Jeffrey R Binder, and Edgar A DeYoe, "Human brain regions involved in recognizing environmental sounds," *Cerebral cortex*, vol. 14, no. 9, pp. 1008–1021, 2004.

[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*, vol. 73, no. 1, pp. 7–19, 2010.

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

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

[9] Danièle Dubois, Catherine Guastavino, and Manon Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 865–874, 2006.

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

[11] 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*, vol. 16, no. 1, pp. 16, 2010.

[12] Margaret M Bradley and Peter J Lang, "The international affective digitized sounds (IADS-2): Affective ratings of sounds and instruction manual," *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[13] Annett Schirmer, Yong Hao Soh, Trevor B Penney, and Lonce Wyse, "Perceptual and conceptual priming of environmental sounds," *Journal of cognitive neuroscience*, vol. 23, no. 11, pp. 3241–3253, 2011.

[14] Bradley R Buchsbaum, Rosanna K Olsen, Paul Koch, and Karen Faith Berman, "Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory," *Neuron*, vol. 48, no. 4, pp. 687–697, 2005.

[15] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson, "CNN architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

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

[17] D Paice Chris et al., "Another stemmer," in *ACM SIGIR Forum*, 1990, vol. 24, pp. 56–61.

[18] George A Miller, "WordNet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[20] Robert Speer, Joshua Chin, and Catherine Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge.," in *AAAI*, 2017, pp. 4444–4451.

[21] Wilma A Bainbridge, Phillip Isola, and Aude Oliva, "The intrinsic memorability of face photographs.," *Journal of Experimental Psychology: General*, vol. 142, no. 4, pp. 1323, 2013.