Manipulating Causal Uncertainty in Sound Objects

Tal Boger
Yale University
New Haven, CT
tal.boger@yale.edu

Ishwarya Ananthabhotla
MIT Media Lab
Cambridge, MA
ishwarya@media.mit.edu

Joseph A. Paradiso
MIT Media Lab
Cambridge, MA
joep@media.mit.edu

ABSTRACT

Causal uncertainty – how sure we are in what produced a sound that we are listening to – is a fundamental aspect of auditory cognition. It is known to be a driver of affect perception, attention, and memory, among other processes. Here, we present an optimization pipeline that systematically manipulates a sound object's intrinsic causal uncertainty by applying a set of acoustic transforms, such as scaling a sound's pitch, amplitude, playback speed, etc. The optimization estimator attempts to produce parameter values for these transforms that modify a sound's causal uncertainty (H_{cu}), as measured by the prediction confidence of an audio classification neural network, while minimizing changes to the resulting prediction labels and transform magnitudes. We then conduct a listening test with N=20 participants to confirm that the causal uncertainty changes resulting from our proposed procedure align with human perception. Though a simple approach, this work demonstrates a first step towards generative audio systems that operate along cognitive dimensions, with powerful implications for user experience design.

KEYWORDS

auditory perception, auditory cognition, sound manipulation, causal uncertainty, optimization

ACM Reference Format:
Tal Boger, Ishwarya Ananthabhotla, and Joseph A. Paradiso. 2021. Manipulating Causal Uncertainty in Sound Objects. In *Proceedings of AudioMostly 2021 - a conference on interaction with sound (AudioMostly 2021).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

When we listen to a sound, the way that our minds subconsciously process what we hear depends on the interaction between low-level, acoustic properties of the sound, as well as higher-level, semantic properties [5, 6, 9, 10, 12, 13, 16, 20, 21, 24].

Research shows that we employ both sets of properties to make complex inferences about the world around us. For example, if we hear a dog barking, we might notice that the pitch of the bark is low, but that its amplitude relative to the other sounds in our periphery is high, thereby drawing our attention. At the same time, we may process more abstract features about the sound, such as its emotionality, which leads us to believe that the dog is not a threat as we take a walk. Perhaps most importantly, ecological sound psychology research demonstrates that one of the primary processes that our mind engages in when interfacing with a sound object is attempting to estimate its source, or *cause* [4, 19]. When asked to describe the sound, for instance, we might say "a dog barked," suggesting that we have immediately inferred the cause of the sound.

This high-level aspect of auditory cognition – causal estimation – is known to play an important role in sound understanding. When treated as an intrinsic property, *causal uncertainty*, or how apparent the source of a sound is, has been shown to be a powerful indicator of how likely we are to remember the sound, attend to it, or respond to it emotionally [4, 19, 23].

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

To this end, we present a method for changing a sound's causal uncertainty by optimization over perturbations in its acoustic properties. Unlike in the ecological audition or psychology literature, we cannot practically compute causal uncertainty by human label annotation and consensus [1, 4, 19, 23]. Instead, following the proposal in [2], we use the uncertainty of a pre-trained audio classification model released by Google, called YAMNet [1] [15], as a proxy for human causal uncertainty (see Section 2). To the best of our knowledge, this is the first known attempt at manipulating causal uncertainty in a structured fashion. Our early results point towards the possibility of using more generalizable learning methods (e.g., methods that can scale to multiple sounds and learn to use a wider range of manipulation strategies) with significant implications for experiences in sound interaction.

Our key contributions in this work are as follows:

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

AudioMostly 2021, September 01–03, 2021, Trento, Italy

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

(2) We apply the procedure to a selection of environmental sounds, quantitatively describe the results of the optimization in terms of convergence and the distribution of changes in acoustic features across sound classes.
(3) We demonstrate the effectiveness of the approach by conducting user listening tests, and show that listeners reliably perceive changes in causal uncertainty, matching those from our optimization procedure, in a sound comparison experiment.

2 RELATED WORK

Following the experimental psychology literature [4, 19], previous work has defined and quantified causal uncertainty in a dataset called HCU400 [1]. The authors of the work curated a set of approximately 400 sounds that intentionally span the spectrum of source ambiguity (from natural, environmental sounds to artificially synthesized sounds), and obtained crowd-sourced annotations for the labels corresponding to each sound. They proposed a metric estimating the cluster density of a sound's labels in a word embedding space as a measure of causal uncertainty, or H_{cu}, and assign this measure to each sound in the dataset. Higher H_{cu} corresponds to a more causally uncertain sound.

In this initial work, the authors discuss how causal uncertainty affects other high-level properties of sound. They find that H_{cu} affects emotion (as higher H_{cu} sounds have less polarized emotion ratings), familiarity, and imageability.

Later work used this dataset in a memory game to quantify how H_{cu} and other features interact in sound object-driven auditory memory [23]. In the work, researchers showed that H_{cu} was among the most important features in predicting a sound's intrinsic memorability and "confusability" (likelihood to be selected as a false positive target). This work adds to the sound psychology literature in suggesting the significance of causal uncertainty in auditory cognition.

Finally, the authors in [2] suggest that the human annotation method for calculating H_{cu} for a previously unseen sound sample, which entails crowd-sourcing tens of labels and performing cluster analysis in a knowledge graph space, is not scalable to real-world, real-time applications. They propose a weak proxy that uses the probabilities associated with the class predictions of an audio classification model as an analog to human uncertainty in estimating the cause of a sound sample, which we adopt in this work.

In this work, we aim to extend methods to quantify causal uncertainty by presenting a strategy to manipulate it. Specifically, we seek to morph sounds towards a target H_{cu}. To this end, methods for altering sounds have progressed significantly in recent years. New large-scale, statistical approaches using neural style transfer, generative adversarial networks, and other deep learning techniques have produced impressive results in the domains of music, speech, and environmental sounds [7, 8]. However, we find that such approaches require large datasets and significant compute to achieve training stability and convergence, especially in the context of the proposed task, which demands subtle changes to create ambiguity without significantly altering, adding, or removing sound objects or events. We further expect statistical methods to pose challenges in stability and complexity, because our method for estimating H_{cu} from a sound excerpt also requires a pre-trained neural network which may suffer from a lack of adversarial robustness.

As a simpler, alternative approach that serves as a proof-of-concept, we take inspiration from the work in [3, 17, 18], which presented a per-image optimization pipeline to modify the visual memorability of face images. The problem in [3, 17, 18] is analogous to ours, as both problems require similarly tight control over semantic and lower-level properties. Their method succeeded in optimizing images of faces to become more or less intrinsically memorable, and the results of their approach were verified in a perceptual task. Here, we take a similar approach in designing an optimization problem, adapting the optimization space and cost function to reflect the relevant semantic and acoustic properties of audio.

3 METHODS

3.1 Overview

Our proposed method operates as follows, as shown in the illustration in Figure 1. The method takes as input a target H_{cu} value and a sound excerpt, and applies a Gaussian process regression-based Bayesian optimization strategy [14], a "blackbox" optimization framework, to determine parameter values for a small, fixed set of acoustic transforms. To evaluate the parameter values, the acoustic feature transforms are applied to the input sound, and a cost function is computed on the result at each iteration of the optimization. We utilize a blackbox approach because our cost function is expensive to compute and not differentiable. For every sound that we wish to manipulate, we apply this optimization for a fixed number of calls and examine the result with the lowest cost as the output. We examine the individual components of this optimization process in the sections below.

3.2 Optimization Parameters

We first define the parameters we are optimizing. To begin with a simple formulation, we create a constrained search space of select low-level audio features. We selected these features and their parameter ranges based on their definitions in a popular sound editing toolchain known as SoX,[2] which we also use to implement the transforms. Table 1 shows the low-level feature values we optimize over, along with the range of their search spaces.

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

Table 1: Features and parameter ranges for optimization. Features are presented in the order they were applied to the sounds

Note that when optimizing a sound to *lower* its H_{cu} (i.e., making a sound *more certain*), we restrict the gain to [0, 25] dBs. This was

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

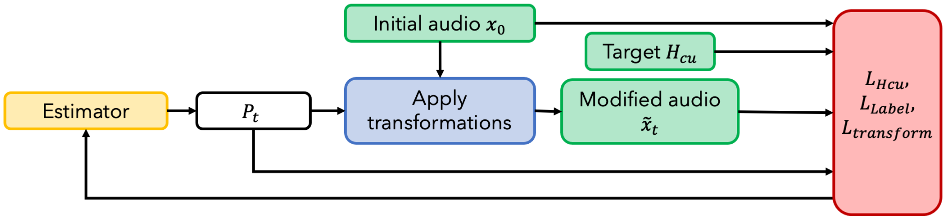

Figure 1: The system takes an audio input X_0 and target causal uncertainty (H_{cu}) value, and estimates parameter values P_t at each time step t for a select set of audio transforms. The transforms are applied to x_0 to generate a modified sound $\tilde{x_t}$ which, along with P_t and the target H_{cu}, are fed to a custom cost function defined over error in H_{cu} (L_{Hcu}), relatedness in sound labels (L_{label}), and transform magnitude ($L_{transform}$).

to ensure that sounds were not becoming “more certain" by being converted to silence, an edge case in our optimization. We also note that this approach can be extended to several other fine-grained audio effects and transforms – examples include equalizers, limiters, compressors, etc. – but we choose to limit our early explorations to the set in Table 1 to achieve reasonable execution times under limited compute.

3.3 Objective Function

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

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

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

3.3.1 *Causal Uncertainty Constraint*

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

$$L_{H_{cu}} = |\text{current } H_{cu} - \text{target } H_{cu}|$$

where H_{cu} is computed following [2] by taking the prediction output from YAMnet on the transformed sound and obtaining the maximum mean probability output.

Then, we weight this term with a λ_1 parameter which is designed to penalize larger values of $L_{H_{cu}}$ more heavily. We defined λ_1 with a step function that monotonically increases as $L_{H_{cu}}$ increases from 0 to 0.5, after which point it is constant.

3.3.2 *Label Constraint*

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

To construct L_{label}, we use node distances in the class ontology of AudioSet, the dataset used to train YAMNet [11, 15]. The ontology presents AudioSet classes in a tree structure. For example, the label “cat" has child nodes “purr" and “meow." We can use this tree structure to our advantage to create an intuitive L_{label} term. If we apply transforms that modify a sound that is initially labeled a “purr," we would want to penalize it much more for becoming a “chainsaw" than for becoming a “meow," given their relative distances on the ontology tree.

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

In defining λ_2, we have two separate cases:

(1) Target H_{cu} > initial H_{cu} (i.e., making a sound less certain)
(2) Target $H_{cu} \leq$ initial H_{cu} (i.e., making a sound more certain)

In case (1), we scale λ_2 according to the current H_{cu}. This is because as we raise H_{cu}, we expect the labels to vary slightly, so we want to relax the total penalty of the $\lambda_2 L_{label}$ term. As such, in case (1), we define:

$$\lambda_2 = \frac{1}{1 + \text{current } H_{cu}}$$

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

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

3.3.3 *Transform Constraint*

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

$$L_{transform} = \frac{1}{6}\left(\frac{|\text{gain}|}{\text{gain}_{max}} + \frac{|\text{pitch}|}{\text{pitch}_{max}} + \frac{|\text{speed} - 1|}{\text{speed}_{max} - 1} + \frac{\text{reverb}}{\text{reverb}_{max}} + \frac{\text{HPF}}{\text{HPF}_{max}} + \frac{\text{LPF}_{max} - \text{LPF}}{\text{LPF}_{max} - \text{LPF}_{min}}\right)$$

and λ_3 is set to a constant scalar.

4 EXPERIMENTS

4.1 Dataset

To evaluate our approach, we apply our optimization method to a selection of sounds from Google's AudioSet dataset [11]. The original dataset consists of 632 classes of sounds, with more than 2 million 10-second sound examples in total; however, we choose a small set of illustrative examples to demonstrate our results. Specifically, we choose four broad categories of environmental sounds – human sounds, animal sounds, nature sounds, and inorganic sounds – and selected pairs of classes of sounds within each category. These include:

- Crying and laughing (human sounds)
- Dog and cat (animal sounds)
- Fire and water (nature sounds)
- Wood and glass (inorganic sounds)

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

4.2 Optimization Targets

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

We allow the optimization for each sound to run for a maximum of 200 iterations using the parameter search ranges and cost function described in Sections 3.2 and 3.3. The initial values for the audio transforms are set to be neutral (zero gain, playback rate of 1, etc.), and the YAMNet model prediction is used to obtain an initial list of the top 10 labels describing the sound.

4.3 Perceptual Evaluation

We finally create a listening task to evaluate whether our optimization results reflect human perception. Specifically, we wish to know whether raising or lowering a sound's H_{cu} results in more or less certainty in listener source estimation.[3]

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

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

To conduct the study, we recruited 20 participants from the online crowd-sourcing platform Prolific (for a discussion of the reliability of Prolific's subject pool, see [22]). Each experiment took approximately 25 minutes, and each participant was compensated upon completion of the experiment.

5 RESULTS

5.1 Optimization Results

Figure 2 shows the convergence values of unweighted loss terms (i.e., $L_{H_{cu}}$, L_{label} and $L_{transform}$) for each class in our optimization when raising and lowering causal uncertainty.

We note that our system was much more successful in raising a sound's causal uncertainty than in lowering it, as demonstrated by the lower $L_{H_{cu}}$ values. Conversely, when lowering a sound's causal uncertainty, the optimization pipeline maintained the labels more consistently, as shown by the lower L_{label}. Both raising and lowering causal uncertainty required similarly large transformations to be applied; both have a similar $L_{transform}$.

To listen to samples of the original and manipulated sounds, visit this OSF repository.

5.2 Perceptual Evaluation Results

In our perceptual task, human evaluations aligned with our optimization results. Subjects were able to choose the more causally certain sound (as determined by our proxy H_{cu}) at a rate significantly above chance ($t(19) = 4.46$, $p < 0.0001$, $M = 57.60\%$, 95% CIs = [54.04%, 61.17%]). This was not simply driven by a few subjects performing with very high accuracy; 16 of 20 subjects chose the more causally certain sound over half the time.

The results of each of the three trial types (higher-anchor, higher-lower, and lower-anchor) were significantly different from chance (see Figure 3). On higher-anchor trials, participants had nearly perfect accuracy (90.15%). On higher-lower trials – where the original sound was not presented – participants chose the more causally certain sound 68.96% of the time. Finally, on the lower-anchor trials, participants consistently mistook the original sound as more causally certain than the one with lower H_{cu} (12.81% accuracy).

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

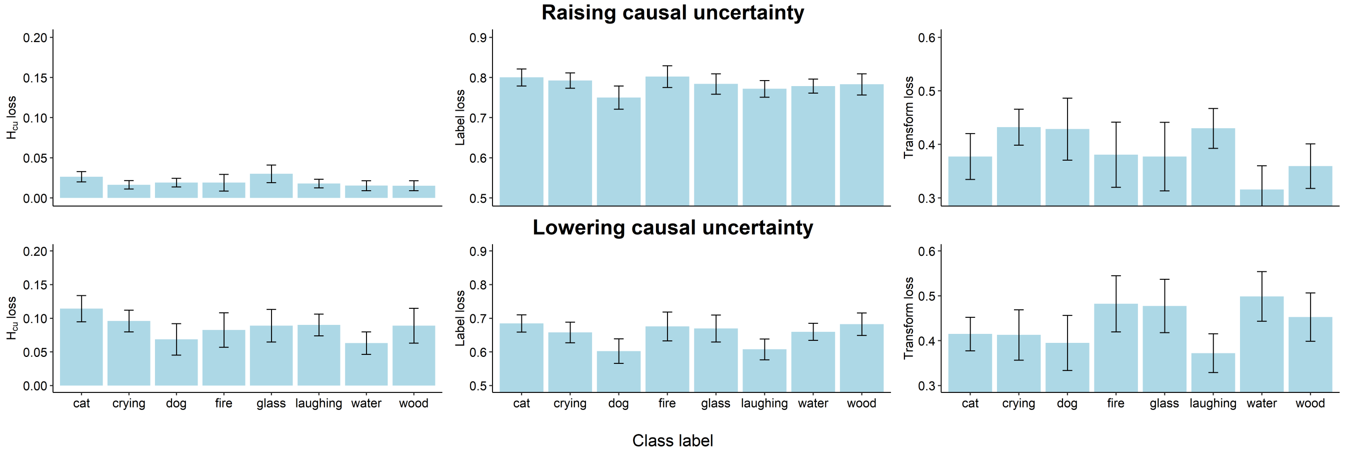

Figure 2: A comparison of unweighted loss terms when raising and lowering H_{cu} for sounds of each class. Error bars are 95% CIs.

The very poor accuracy in the lower-anchor trials has two potential causes. First, it highlights the challenge of making the source of a sound less uncertain, using only a simple set of acoustic tools – we discuss the framing of the task and future strategies for improving performance in Section 6. Secondly, the results demonstrate that participants potentially perceive *any* change to a sound using our effects chain as increasing its source ambiguity, suggesting that the manipulated sounds do not seem natural. We expect that this behavior can be controlled for by more subtle, computationally intensive sound operations (see Section 6).

Finally, we compare the accuracy per class in our perceptual task. In Figure 4, we note that we do not have homogeneous results across classes; for instance, sounds with the labels fire and water – the two natural sound classes – have lower accuracy than the sounds with other class labels. This hints at differences between sounds – either spectral, semantic, or both – that could be further exploited for the manipulation of causal uncertainty, with deeper analysis that could stem from a larger number of data points per trial.

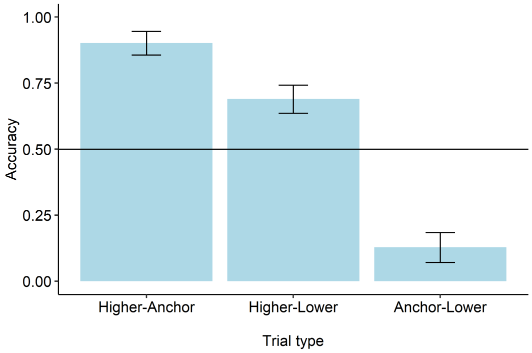

Figure 3: Grouped subject accuracy for each trial type in perceptual task. Error bars are 95% CIs. All groups differ from 0.5 with p < 0.001

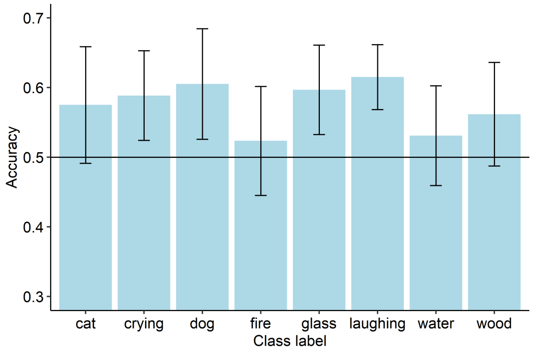

Figure 4: Grouped subject accuracy for each class in the perceptual task, across trial types. Error bars are 95% CIs.

6 DISCUSSION AND FUTURE WORK

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

6.1 Dataset selection

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

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

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

6.2 Modeling approaches

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

Methods other than our blackbox optimization may yield better results, too, given that corresponding changes to the dataset are made. For instance, in a broader problem space employing a larger dataset, deep learning-based approaches may yield better results, as they can change not only low-level features (like we do here) but also semantic features of the sound. Future work may explore the viability of such approaches for this problem.

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

6.3 Meanings of causal uncertainty

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

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

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

7 CONCLUSION

In this work, we present an optimization procedure for manipulating the causal uncertainty of a sound. We select a set of acoustic effects whose parameters we optimize over to alter an input sound, and design a custom objective function to drive the input sound towards a target H_{cu} while minimizing both the magnitude of transformations applied and changes to the original set of labels ascribed to the sample by a pre-trained audio classification network. We demonstrate reasonable convergence errors across sound classes in our test dataset, and show that the results from a perceptual listening task align with our optimization results. Given the important role that sound source causal uncertainty plays in auditory cognition, we believe this work demonstrates immense potential for new paradigms in auditory interface and user experience design.

ACKNOWLEDGMENTS

The authors would like to thank David Ramsay from the Responsive Environments Group at the MIT Media Lab for his comments and guidance on this work.

REFERENCES

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

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

[3] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. 2013. The Intrinsic Memorability of Face Photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323.

[4] James A Ballas. 1993. Common Factors in the Identification of an Assortment of Brief Everyday Sounds. *Journal of experimental psychology: human perception and performance* 19, 2 (1993), 250.

[5] James Craig Bartlett. 1977. Remembering Environmental Sounds: The Role of Verbalization at Input. *Memory & Cognition* 5, 4 (1977), 404–414.

[6] Oliver Bones, Trevor J Cox, and William J Davies. 2018. Distinct Categorization Strategies for Different Types of Environmental Sounds. Euronoise.

[7] Chris Donahue, Julian McAuley, and Miller Puckette. 2019. Adversarial Audio Synthesis. arXiv:1802.04208 [cs.SD]

[8] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. GANSynth: Adversarial Neural Audio Synthesis. https://openreview.net/pdf?id=H1xQVn09FX

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

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

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[12] Bruno L Giordano, John McDonnell, and Stephen McAdams. 2010. Hearing Living Symbols and Nonliving Icons: Category Specificities in the Cognitive Processing of Environmental Sounds. *Brain and cognition* 73, 1 (2010), 7–19.

[13] Brian Gygi and Valeriy Shafiro. 2007. General Functions and Specific Applications of Environmental Sound Research. *Frontiers in Bioscience* 12 (2007), 3152–3166.

[14] Tim Head, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, fcharras, Zé Vinícius, cmmalone, Christopher Schröder, nel215, Nuno Campos, Todd Young, Stefano Cereda, Thomas Fan, rene rex, Kejia (KJ) Shi, Justus Schwabedal, carlosdanielcsantos, Hvass-Labs, Mikhail Pak, SoManyUsernamesTaken, Fred Callaway, Loïc Estève, Lilian Besson, Mehdi Cherti, Karlson Pfannschmidt, Fabian Linzberger, Christophe Cauet, Anna Gut, Andreas Mueller, and Alexander Fabisch. 2018. *scikit-optimize/scikit-optimize: v0.5.2.* https://doi.org/10.5281/zenodo.1207017

[15] 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. 2017. CNN Architectures for Large-scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[16] Koji Inui, Tomokazu Urakawa, Koya Yamashiro, Naofumi Otsuru, Makoto Nishihara, Yasuyuki Takeshima, Sumru Keceli, and Ryusuke Kakigi. 2010. Non-Linear Laws of Echoic Memory and Auditory Change Detection in Humans. *BMC neuroscience* 11, 1 (2010), 80.

[17] Aditya Khosla, Wilma A Bainbridge, Antonio Torralba, and Aude Oliva. 2013. Modifying the Memorability of Face Photographs. In *Proceedings of the IEEE International Conference on Computer Vision*. 3200–3207.

[18] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and Predicting Image Memorability at a Large Scale. In *Proceedings of the IEEE International Conference on Computer Vision*. 2390–2398.

[19] Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini. 2010. Listener Expertise and Sound Identification Influence the Categorization of Environmental Sounds. *Journal of Experimental Psychology: Applied* 16, 1 (2010), 16.

[20] James W Lewis, Frederic L Wightman, Julie A Brefczynski, Raymond E Phinney, Jeffrey R Binder, and Edgar A DeYoe. 2004. Human Brain Regions Involved in Recognizing Environmental Sounds. *Cerebral cortex* 14, 9 (2004), 1008–1021.

[21] Michael M Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers. 2000. Confrontation Naming of Environmental Sounds. *Journal of clinical and experimental neuropsychology* 22, 6 (2000), 830–864.

[22] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

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

[24] Annett Schirmer, Yong Hao Soh, Trevor B Penney, and Lonce Wyse. 2011. Perceptual and Conceptual Priming of Environmental Sounds. *Journal of cognitive neuroscience* 23, 11 (2011), 3241–3253.