Audio Engineering Society

Conference Paper

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

The Intrinsic Memorability of Everyday Sounds

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

Responsive Environments, MIT Media Lab

Correspondence should be addressed to Ishwarya Ananthabhotla (Ishwarya@mit.edu)

ABSTRACT

Of the many sounds we encounter throughout the day, some stay lodged in our minds more easily than others; these may serve as powerful triggers of our memories. In this paper, we measure the memorability of everyday sounds across 20,000 crowd-sourced aural memory games, and then analyze the relationship between memorability and acoustic cognitive salience features; we also assess the relationship between memorability and higher-level gestalt features such as its familiarity, valence, arousal, source type, causal certainty, and verbalizability. We suggest that modeling these cognitive processes opens the door for human-inspired compression of sound environments, automatic curation of large-scale environmental recording datasets, and real-time modification of aural events to alter their likelihood of memorability.

1 Introduction

For a sound to enter our memory, it is first unconsciously processed by a change-sensitive, gestalt neural mechanism before passing through a conscious filtering process [1, 2, 3]. We then encode this auditory information via a complex and variable procedure; frequently we abstract our experiences into words, though we also utilize phonological-articulatory, visual/visuospatial, semantic, and echoic memory [4, 5]. Different types of memory may also drive more visceral forms of recollection and experience; non-semantic memory, for example, may underpin powerful recollection and nostalgia similar to those reported with music [6].

In this work, we map out the features of everyday sounds that drive their memorability using an auditory memory game. As a recall experiment, we hypothesize that it can provide useful insights into cognitive models for auditory capture and curation. Additionally, we design the task such that it is beyond the capacity of our working and echoic memory and engages long-term memory cognitive processes [7, 8]. With this work we hope to illuminate the role of top-down features – imageability, emotionality, causal certainty, and familiarity – in auditory memory. Using state-of-the-art cognitive saliency models, we also explore the relative importance of low-level acoustic descriptors against high-level conceptual ones for memory formation. To our knowledge, this is the first general treatment of auditory memorability that combines low-level

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

auditory salience models with multi-domain, top-down cognitive gestalt features. This work enables more accurate models of auditory memory, and represents a step toward cognitively-inspired compression of everyday sound environments, automatic curation of large-scale environmental recording datasets, and real-time modification of aural events to alter their likelihood of memorability.

2 What Influences Memorability?

Many factors influence the cognitive processes underlying human aural processing and storage. Research shows a complicated interdependence between attention, acoustic feature salience, source concept salience, emotion, and memory; furthermore, verbal, pictorial, and phonological-articulatory mnemonics can have a significant impact on sound recall tasks.

Neuroscience research supports the idea that gestalt auditory pre-processing is followed by attentive filtering prior to conscious perception [1, 2]. These gestalt representations incorporate both 'bottom-up' and 'top-down' processes – sounds that are contextually novel only based on their acoustic features, as well as sounds that are only conceptually novel, lead to distinct and measurable variations in unconscious event-related potentials (ERPs) [9]. These data motivate the need to incorporate high-level conceptual features and low-level acoustic features relative to a sound context for even simple models of auditory processing, attention, and memory.

The stored gestalt representation of the current sound context – necessary to explain change-driven ERPs – can be thought of as the first stages of auditory memory [3]. This immediate store, known as 'echoic memory', starts decaying exponentially by 100ms after a sound onset [10]. Measurements of ERPs suggest immediate storage of rhythmic stimuli on the order of 100 ms with a resolution as low as 5 ms [11]; other studies have shown this immediate store is complimented with an additional echoic mechanism that lasts several seconds [12, 13]. On these time-scales, our auditory system compresses its perceptual representation of textures based on time-averaged statistics [14].

These principles have been used to design 'bottom-up' cognitive saliency models [15, 16]. While other time-averaged low-level features have been used to quantify sound similarity [17], saliency models are now common in practical applications [18, 19]. Although the above work does not include higher-level gestalt processing, a few researchers have successfully combined low-level saliency modeling with a focused, task-specific top-down cognitive model [20, 21, 22]. These models aren't designed to generalize outside of their domain, however.

In general, high-level 'top-down' features have been an area of intense study that begins in the 1950's, when Colin Cherry demonstrated that his subjects noticed their name – and no other verbal content – spoken by a secondary speaker in a shadowing test [23]. Besides the ongoing work in verbal auditory processing, research into non-verbal stimuli and auditory memory also provides us with insight into the role of conceptual abstraction in modulating attention and memory in a more general sense.

One such abstraction – emotionality – is known to have a powerful effect on cognitive processing and memory formation [24]. For music, recall has been shown to improve with positive valence, high arousal sound events [6], though recent research has called the significance of arousal into question [25]. In noise pollution research, the high-level perception of human activity is considered 'pleasant' (more positively valenced) regardless of low-level acoustic features [26]. In general, the emotional impact of a sound is correlated with the clarity of its perceived source [27], though sounds can have emotional impact even without a direct mapping to an explicit abstract idea [28].

For recognition and recall memory exercises, verbalizing a sound or naming a sound (both of which may engage phonological-articulatory motor memory) is the most common and successful strategy [29]. This semantic abstraction has overshadowing effects, though; verbal descriptions can distort recollection of the sound itself, degrading recognition performance without altering confidence [30]. Some researchers specifically isolate and study echoic memory, separating it from naming (a process that doesn't involve outward verbalization) using homophonic sound sources to ensure subjects are not relying internally on a naming mechanism [31]. In everyday life, though, we naturally rely on a complex mixture of echoic (perceptual), phonological-articulatory (motor), verbal (semantic), and visual memory [4, 5].

We base our work on [27], in which the authors present the HCU400 dataset. The dataset consists of a curated selection of everyday sounds to include high-level

features that may influence a sound's memorability– most notably, its causal certainty (the degree to which a sound implies a clear, unambiguous source, denoted as H_{cu}), the implied source itself as determined by crowd-sourced workers, and its acoustic features. This data includes ratings for the valence and arousal of each sound, its familiarity, and how easily it conjures a mental image (features that have strong correlations with H_{cu}). Combined, these data provide insight into a sample's emotionality, as well as the ease by which it can be stored in semantic or visual memory. Embeddings based on location relationships of sound sources ('at-location', 'located-near') provide additional insight into the conceptual distinctness of a sound compared to the contextualizing sounds of a soundscape, and can serve as a first-order proxy for ecological exposure.

In this paper, we explore the relationship between both low-level acoustic features and high-level conceptual features with the memorability of sound, in a context that engages long-term memory processes. We hope that a thorough analysis of the major low- and high-level features from the literature might lay the foundation for a practical, generalized model of auditory memory.

We set out to test a few important hypotheses, namely:

1. The cognitive processing of sound is similar enough across people that trends in recall across sound samples will be measurable and robust across users.

2. Higher-level gestalt features will be most predictive of successful recall performance. We see from the literature that naming and emotionality have very strong effects in similar tasks– we expect sounds with low H_{cu} (easy to name their source) and strong valence/high arousal to be the most memorable. High H_{cu} (uncertain) sounds elicit weaker emotions, reinforcing this effect.

3. Low-level acoustic feature information will marginally predict memory performance. Gestalt features are not easily mapped to low level feature space (and we expect gestalt features to dominate); however, the literature suggests a measurable, second-order contribution from low-level perceptual saliency modeling.

4. The likelihood of a sound eliciting a false memory will be best predicted by its conceptual familiarity as well as by low-level acoustic features.

5. The context a sound is presented against will have a marginal but measurable impact on whether it is recalled. In other words, we expect emotional and unambiguous sounds to be the most memorable regardless of presentation, but when a sound stands out against the immediately preceding sounds, we hypothesize that it will be slightly more memorable.

3 Samples and Feature Generation

Audio samples for this test were taken from the HCU400 dataset [27], and standard low-level acoustic features were extracted from each sample based on prior precedent [17]. We used default configurations from three audio analysis tools: Librosa [32], pyAudioAnalysis [33], and Audio Commons [34], which include basic features (i.e. spectral spread) as well as more advanced timbral modeling. We supplement these features with additional summary statistics like high/mid/bass energy ratios, percussive vs. harmonic energy, and pitch contour diversity.

Over the last decade there have been advances in cognitive models that can determine the acoustic salience of sound, inspired by the neuroscience of perception [16, 19]. Here we follow the procedure proposed by [15], applying separate temporal, frequency, and intensity kernels to an input magnitude spectrogram to produce three time-frequency salience maps. Figure 1 shows a comparison of temporal salience between two sound samples in the HCU400 dataset with highly contrasting auditory properties. From these maps, we compute a series of summary statistics to be used as features.

High-level, top-down features were taken from [27] and include causal uncertainty (H_{cu}), the cluster diameter of embedding vectors generated from user-provided labels (quantifying source agreement or source location), familiarity, imageability, valence, and arousal.

4 Measuring Memorability

In order to quantify memorability, we drew inspiration from work in [35], which used an online memory game to determine the features that make images memorable. We designed an analogous interface for the audio samples in the HCU400 dataset; this interface can be found at `http://keyword.media.mit.edu/memory`. The game opens with a short auditory phase alignment-based assessment [36] to ensure that

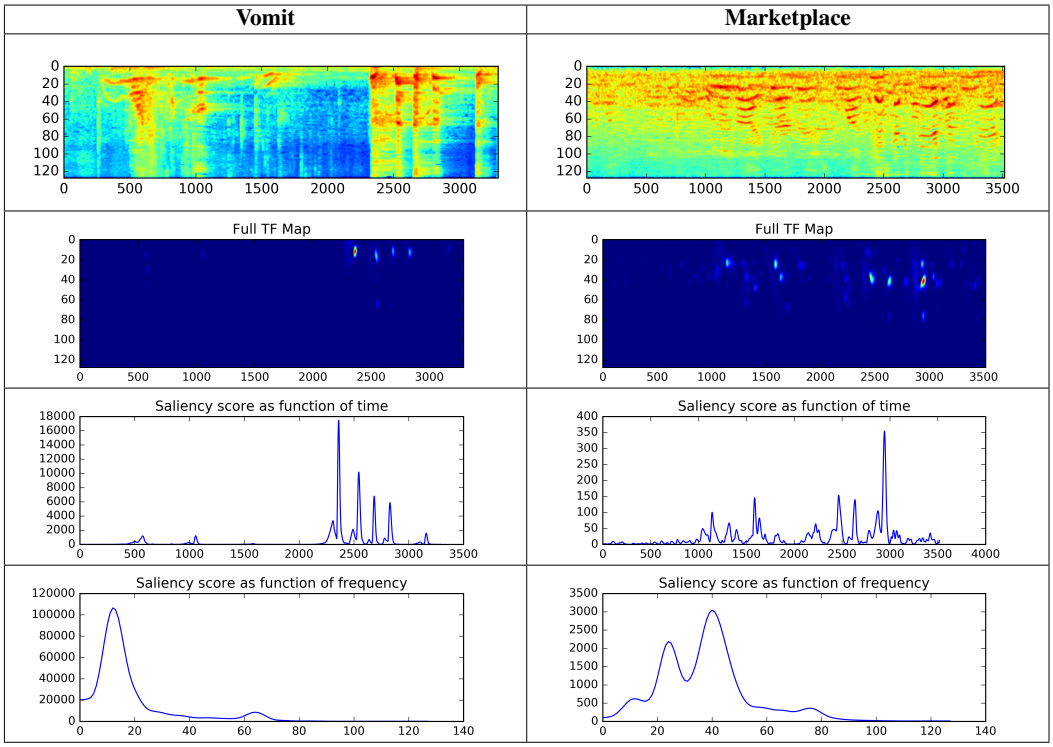

Fig. 1: A table demonstrating the auditory salience model based on [15] applied to two contrasting audio samples in the HCU400 dataset. The resulting salience scores (bottom) are summarized and used as features in predicting memorability.

participants are wearing headphones, followed by a survey that captures data about where they spend their time (urban vs. rural areas, the workplace vs home, etc). Participants are then presented with a series of 5 second sound clips from the HCU400 dataset, and are asked to click when they encounter a sound that they've heard previously in the task. At the end of each round consisting of roughly 70 sound clips, the participant is provided with a score. Screenshots of the interface at each stage are shown in Figure 2.

By design, each round of the game consisted of 1-2 pairs of *target sounds* and 20 pairs of *vigilance sounds*. *Target sounds* were defined as samples from the dataset that were separated by exactly 60 samples– the sounds for which memorability was being assessed in a given round. The *vigilance sounds*, pairs of sounds that were separated in the stream by 2 to 3 others, were used to ensure reliable engagement throughout the task following the method in [35]. Roughly 20,000 samples were crowd-sourced on Amazon Mechanical Turk such that a single task consisted of a single round in the game. Individual workers were limited to no more than 8 rounds to ensure that target samples were not repeated. Rounds that failed to meet a minimum vigilance score (>60%) or exceeded a maximum false positive rate (>40%) were discarded.

5 Summary of Participant Data

We recruited 4488 participants, consisting of a small (<50) number of volunteers from the university community and the rest from Amazon Mechanical Turk. Our survey data shows that our participants report a 51/37/12% split between urban, suburban, and rural communities. We see weak trends in the average time per location reported for each community type– urbanites self-report spending less time at home, in the kitchen, in cars, and watching media on average. Rural participants report spending more time in churches and in nature. Using KNN clustering and silhouette analysis, we find four latent clusters – students (590 users), office workers (1250 users), home-makers (1640 users),

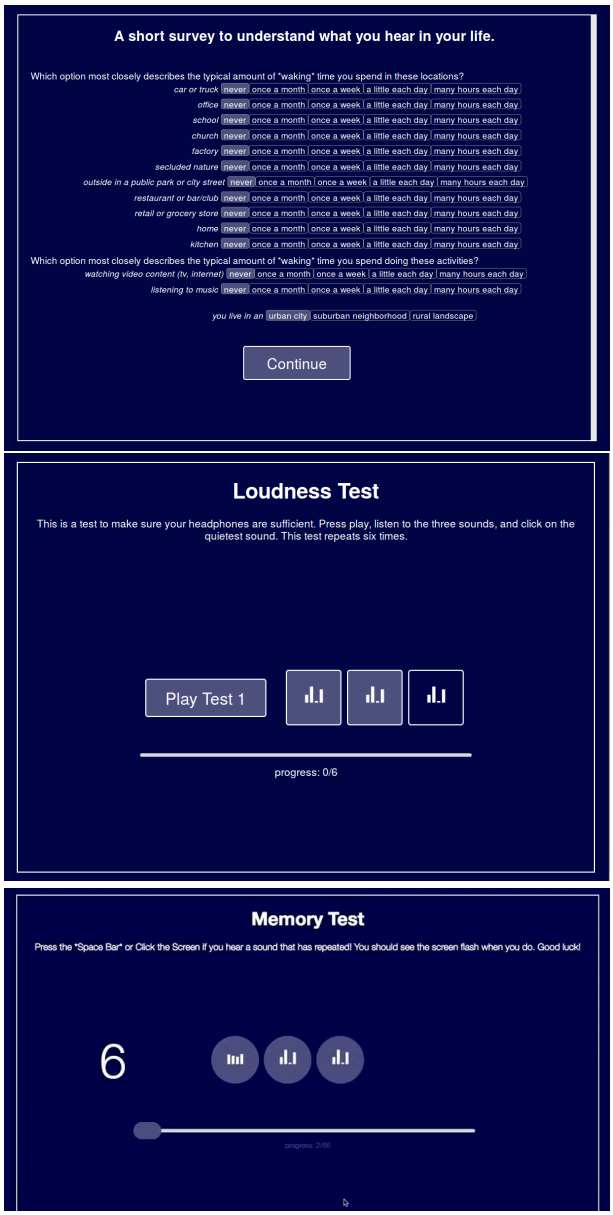

Fig. 2: Screenshots of the auditory memory game interface presented to participants as a part of our study. The game can be found at `http://keyword.media.mit.edu/memory`.

and none of these (1010 users). Split-rank comparisons between groups did not reveal meaningful differences in results across user groups; we speculate any differences due to ecological exposure of sounds between environments is not consistent or influential enough at this group level to alter performance.

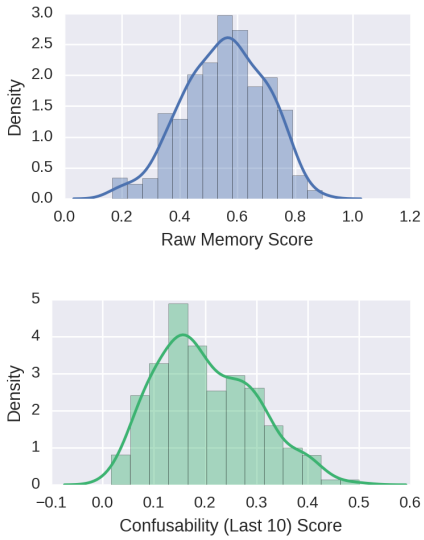

Fig. 3: Top: A histogram of the raw scores for each sound – they were successfully remembered and identified about 55% of the time on average, with a large standard deviation; Bottom: A histogram of "confusability" scores for each sound, with an average score of about 25%.

6 Summary of Memory Data

The raw memorability score M for each sound is simply computed as the number of times it was correctly identified as the target divided by the number of its appearances. However, this does not account for the likelihood that the sound will be falsely remembered (i.e. clicked on without a prior presentation). We additionally compute a "confusability" score C_{10} for each sound sample, defined as the false positive rate for sounds when they fall close to the second target presentation (i.e. in the last ten positions of the game). We can thus derive a "normalized memory score" represented by $M - C_{10}$. In attempting to understand auditory memory, we consider both what makes a sound memorable *and* what makes a sound easily mis-attributed to other sounds, whether those sounds are encountered in our game or represent the broader set of sounds that one encounters on a habitual basis. We therefore model both normalized memorability and confusability in this work.

We confirm the reliability of both the normalized memory scores and the confusability scores across participants by performing a split ranking analysis similar to [35] with 5 splits, shown in Figure 4 with their respective Spearman correlation coefficients. This confirms that memorability and confusability are consistent, user-independent properties.

In Table 1, we show a short list of the most and least memorable and confusable sounds in our dataset as a function of the normalized memorability score and confusability score.

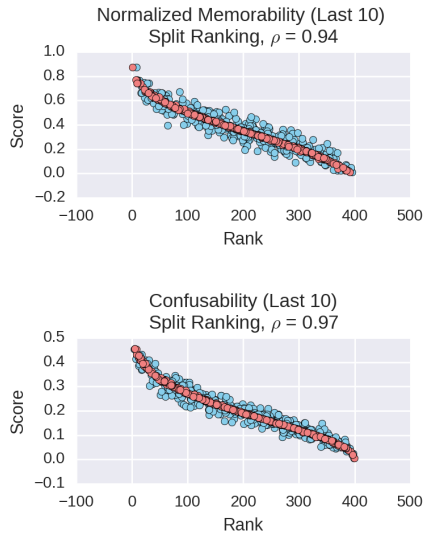

Fig. 4: The results of the split-ranking analysis for the normalized memorability score and confusability score, using 5 splits; The Spearman coefficient correlations demonstrate the reliability of these scores across study participants, enabling us to model both metrics in the later parts of the work.

7 Feature Trends in Memorability and Confusability

We consider two objectives – (1) to determine the relationship between individual features and our measured memorability and confusability scores, and (2) to determine the relative importance of these features in predicting memorability and confusability. To address the former, we provide the resulting R^2 value after applying a transform learned using support vector regression

Most Memorable	**Least Memorable**
man_screaming.wav	*morphed_firecracker_fx.wav*
woman_screaming.wav	*truck_(idling).wav*
flute.wav	*morphed_turkey2_fx.wav*
woman_crying.wav	*morphed_airplane_fx.wav*
opera.wav	*morphed_metal_gate_fx.wav*
yawn.wav	*morphed_shovel_fx.wav*
Most Confusable	**Least Confusable**
garage_opener.wav	*clock.wav*
lawn_mower.wav	*morphed_335538_fx.wav*
washing_machine.wav	*phone_ring.wav*
rain.wav	*woman_crying.wav*
morphed_tank_fx.wav	*woman_screaming.wav*
morphed_printing_press_fx.wav	*vomit.wav*

Table 1: A list of the most and least memorable and confusable sounds from the HCU400 dataset.

(SVR) for each individual feature. For the latter, we use a sampled Shapely value regression technique in the context of SVR– that is, we first take N random features (N between 1 and 10) and perform an SVR to predict memorability or confusability scores for our 402 sounds and the calculated R^2 of the fit. We then measure the change in R^2 as we append every remaining feature to the model, each individually. The largest average changes over 10k models are reported in table 2. This technique is robust to complex underlying nonlinear relationships from feature space to predicted metric as well as feature collinearity. We find that the strongest predictors of both memorability and confusability are the measures of imageability (how easy the sound is to visualize) and its causal uncertainty. Memorability is dominated by high level, gestalt features, with only one lower level feature ('pitch diversity') in the ten most important features. Low level features, including those derived from the auditory salience models, play a more significant role in determining confusability.

The absolute R^2 values indicate that no individual feature is a significant predictor of memorability by itself. This implies a complex causal interplay in feature space, which we explore further in the set of plots presented by Figure 5. In each plot, we show a distribution of feature values for the 15% of sounds that are most memorable or least confusable (blue) contrasted against the least memorable or most confusable sounds (red). We first consider the effect of H_{cu} and valence on memory– low memorability and high confusability sounds exhibit a similar trend of high causal uncertainty and neutral valence (Column 1). In Column 2, we consider imageability and familiarity ratings, shown to be strongly collinear in [27]. Here, their re-

lationship to memorability and confusability diverge; while both are positively correlated with memorability, *neutral* ratings are the stronger predictor of confusability. This suggests that we are most likely to confuse sounds if they are loosely familiar but neither strictly novel nor immediately recognizable. Finally, Column 3 reveals a discernible decision boundary in low-level feature space for confusability which doesn't exist in its memorability counterpart. The relative importance of low-level salience features, here represented by spectral spread, aligns with intuition– we hypothesize that, in the absence of strong causal uncertainty or affect feature values, our perception of sounds is driven by their spectral properties.

8 Per-game Modeling of Memorability

The aural context in which a sound is presented, which includes ecological exposure as well as the immediate preceding sounds in our audition task, may influence the memory formation process. The literature supports the notion that, given a context, unexpected sounds are more likely to grab our attention and engage memory [9]. To understand this effect in our test, we ran two studies based on a 5 sound context (approximating the limits of semantic working memory) and a 1 sound context (approximating the limits of echoic memory).

Table 3 shows the results of a model trained to predict whether the target in each game will be successfully recalled. This model was trained with the most memorable and least memorable sounds only (15th/85th percentiles) with a 5-fold cross-validation process, and results are reported on a 15% hold-out test set.

To begin, a baseline model is trained using the absolute, immutable features of the target sound. Because there are a limited number of sounds in our dataset relative to the number of games, the feature space is redundant and sparse, and we expect the accuracy of this model to converge to the average expected value over our set of sounds. We then introduce contextual features– the *relative difference* (z-score) of target sound features with those of the varying sounds that precede its first presentation in each game– to see if our model improves based on the context of our 50 most meaningful features (from the SVR analysis; 25 high-level and 25 low-level). In both 5- and 1-sound context cases, however, model performance does not improve as we would expect if the context provided additional useful information.

We also run a classifier that *only* uses contextual features, to ensure informative context has not been obscured/subsumed by the absolute features in our first test. We start with a noise baseline, in which contextual features are calculated using a random, incorrect context– these features are still informative as the z-score depends largely on the absolute features of the target sound. We then train the same model with the proper context to assess the difference in performance. There is no improvement when the true context is reintroduced.

This leads us to a meaningful insight, contrary to our hypothesis – context does *not* exert a measurable influence on our results. While context likely does matter in real-world settings, we suspect that our memory game framework indirectly primes participants to expect otherwise surprising sounds. This confirms that our data is the consequence of truly intrinsic properties of the sounds themselves, independent of immediate context *and* participant ecological exposure (as was demonstrated in the split-rank analysis).

9 Implications and Conclusion

In this work, we quantify the inherent likelihood that a sound will be remembered or incorrectly confused and confirm that is consistent across user groups. In line with our hypotheses, we show that the most important features that contribute to a sound being remembered are gestalt– namely those sounds with clear sound sources (high H_{cu}), that are easy to visualize, familiar, and emotional. We also show that low H_{cu} sounds that are not familiar or easy to visualize are most likely to be mis-attributed, and low level features play a more important role in predicting this behavior. These relationships are not influenced by context, and are intrinsic properties of the sounds themselves.

To our knowledge, this is the first body of work that combines top-down theories from psychology and cognition with bottom-up auditory salience frameworks to model the memorability of everyday sounds. We posit that the demonstration of memorability as an intrinsic, user and context-independent property of sounds, along with the insights mentioned above, have significant implications for audio technology research – for example, knowing that gestalt features are the primary drivers of memorability might allow us to selectively choose audio samples in a stream or sound environment to be recorded and stored, as a way of mimicking human

Top Predictors for Memorability and Confusability					
Memorability			Confusability		
Feature	R^2	Shapely ΔR^2	Feature	R^2	Shapely ΔR^2
Imageability	0.201	0.126	**Imageability**	0.065	0.078
H_{cu}	0.224	0.125	H_{cu}	0.073	0.078
Familiarity	0.176	0.123	Avg Spectral Spread	0.087	0.078
Valence	0.178	0.120	Peak Spectral Spread	0.037	0.076
Location Embedding Density	0.147	0.117	Peak Energy, Frequency Salience Map	0.059	0.076
Familiarity std	0.103	0.117	**Location Embedding Density**	0.100	0.076
Pitch Diversity	0.084	0.113	Frequency Skew, Frequency Salience Map	0.059	0.076
Imageability std	0.086	0.113	**Arousal**	0.039	0.076
Arousal	0.072	0.112	Peak Energy, Intensity Salience Map	0.044	0.075
Arousal std	0.056	0.111	**Familiarity**	0.045	0.075
Avg Spectral Spread	*0.099*	*0.107*	***Valence***	*0.100*	*0.075*
Timbral Sharpness	*0.094*	*0.091*	*Timbral Roughness*	*0.094*	*0.047*
Max Energy	*0.091*	*0.100*	*Avg Flux, Sub-band 1*	*0.092*	*0.064*
Treble Energy Ratio	*0.090*	*0.020*	*Flux Entropy, Sub-band 1*	*0.091*	*0.061*

Table 2: The top performing features from the Shapely regression analysis for both memorability and confusability (gestalt features are bolded); shown are the features ordered by their respective contributions to the R^2 value, with additional features with top performing individual R^2 values appended in italics. The first column indicates the individual predictive power of each feature; the second indicates its relative importance in the context of the full feature set.

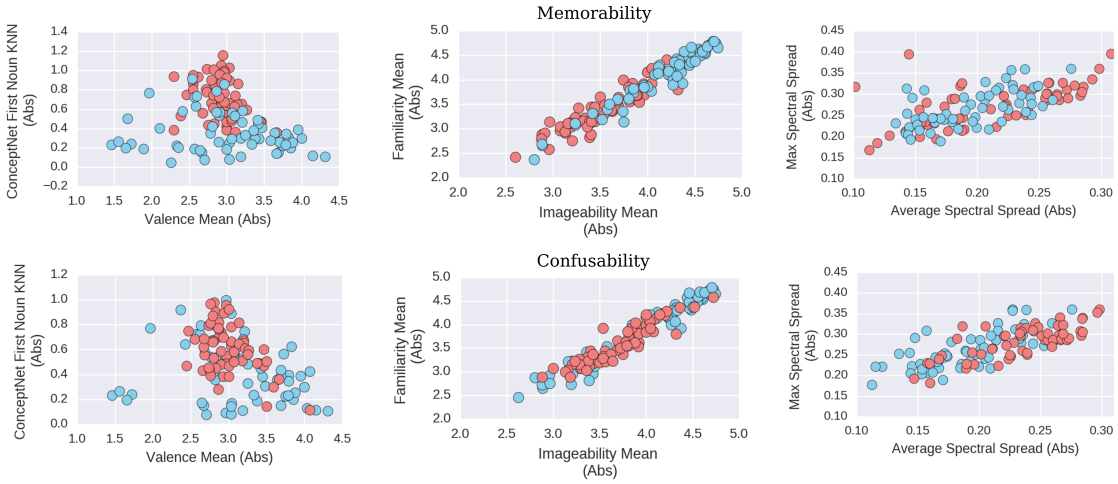

Fig. 5: Scatter plots showing the changes in distribution of select features based on extremes in memorability (top row) and confusability (bottom row); blue indicates sounds that are most (85th percentile) memorable or least (15th percentile) confusable; red indicates sounds that are least memorable or most confusable.

memory to perform compression at a level of abstraction higher than the sample level. An understanding of the most significant predictors of memorability and confusability might also allow us to artificially manipulate our sonic environments to make certain streams of audio more or less memorable, perhaps as a memory aid or a mechanism to eliminate distractions vying for our attention. Looking ahead, we aim to enable many of these applications by translating the principles from this work to an online, real-time model.

References

[1] Snyder, J. S. and Elhilali, M., "Recent advances in exploring the neural underpinnings of auditory scene perception," *Annals of the New York Academy of Sciences*, 1396(1), pp. 39–55, 2017.

[2] Winkler, I., Denham, S. L., and Nelken, I., "Modeling the auditory scene: predictive regularity representations and perceptual objects," *Trends in cognitive sciences*, 13(12), pp. 532–540, 2009.

Memorability Per-Game Models	
Features	**Accuracy (%)**
Absolute + All 5-Sound Context Feats (working semantic)	68.0
Absolute + Top 50 5-Sound Context Feats	69.1
Absolute Feature Only Baseline (~expected value)	*70.3*
Contextual Only, 5-Sound Context (working semantic)	62.5
5 Sound Context, Noise Baseline	*64.1*
Absolute + All 1-Sound Context Feats (echoic)	68.0
Absolute + Top 50 1-Sound Context Feats	69.5
Absolute Feature Only Baseline (~expected value)	*70.3*
Contextual Only, 1-Sound Context (echoic)	60.0
1 Sound Context, Noise Baseline	*61.3*

Table 3: The influence of contextual sounds before the first presentation of the target on our ability to predict recall across games.

[3] Inui, K., Urakawa, T., Yamashiro, K., Otsuru, N., Nishihara, M., Takeshima, Y., Keceli, S., and Kakigi, R., "Non-linear laws of echoic memory and auditory change detection in humans," *BMC neuroscience*, 11(1), p. 80, 2010.

[4] Buchsbaum, B. R., Olsen, R. K., Koch, P., and Berman, K. F., "Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory," *Neuron*, 48(4), pp. 687–697, 2005.

[5] Vaidya, C. J., Zhao, M., Desmond, J. E., and Gabrieli, J. D., "Evidence for cortical encoding specificity in episodic memory: memory-induced re-activation of picture processing areas," *Neuropsychologia*, 40(12), pp. 2136–2143, 2002.

[6] Jäncke, L., "Music, memory and emotion," *Journal of biology*, 7(6), p. 21, 2008.

[7] Ma, W. J., Husain, M., and Bays, P. M., "Changing concepts of working memory," *Nature neuroscience*, 17(3), p. 347, 2014.

[8] Loaiza, V. M., Duperreault, K. A., Rhodes, M. G., and McCabe, D. P., "Long-term semantic representations moderate the effect of attentional refreshing on episodic memory," *Psychonomic bulletin & review*, 22(1), pp. 274–280, 2015.

[9] Schirmer, A., Soh, Y. H., Penney, T. B., and Wyse, L., "Perceptual and conceptual priming of environmental sounds," *Journal of cognitive neuroscience*, 23(11), pp. 3241–3253, 2011.

[10] Lu, Z., Williamson, S., and Kaufman, L., "Behavioral lifetime of human auditory sensory memory predicted by physiological measures," *Science*, 258(5088), pp. 1668–1670, 1992.

[11] Nishihara, M., Inui, K., Morita, T., Kodaira, M., Mochizuki, H., Otsuru, N., Motomura, E., Ushida, T., and Kakigi, R., "Echoic memory: investigation of its temporal resolution by auditory offset cortical responses," *PloS one*, 9(8), p. e106553, 2014.

[12] Cowan, N., "On short and long auditory stores." *Psychological bulletin*, 96(2), p. 341, 1984.

[13] Alain, C., Woods, D. L., and Knight, R. T., "A distributed cortical network for auditory sensory memory in humans," *Brain research*, 812(1-2), pp. 23–37, 1998.

[14] McDermott, J. H., Schemitsch, M., and Simoncelli, E. P., "Summary statistics in auditory perception," *Nature neuroscience*, 16(4), p. 493, 2013.

[15] Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K., "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, 15(21), pp. 1943–1947, 2005.

[16] Delmotte, V. D., *Computational auditory saliency*, Ph.D. thesis, Georgia Institute of Technology, 2012.

[17] Richard, G., Sundaram, S., and Narayanan, S., "An overview on perceptually motivated audio indexing and classification," *Proceedings of the IEEE*, 101(9), pp. 1939–1954, 2013.

[18] Schauerte, B., Kühn, B., Kroschel, K., and Stiefelhagen, R., "Multimodal saliency-based attention for object-based scene analysis," in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pp. 1173–1179, IEEE, 2011.

[19] Kalinli, O., Sundaram, S., and Narayanan, S., "Saliency-driven unstructured acoustic scene classification using latent perceptual indexing," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*, pp. 1–6, IEEE, 2009.

[20] Kalinli, O. and Narayanan, S. S., "Combining task-dependent information with auditory attention cues for prominence detection in speech,"

in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[21] Kalinli, O. and Narayanan, S., "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Transactions on audio, Speech, and language processing*, 17(5), pp. 1009–1024, 2009.

[22] Marchegiani, M. L., "Top-Down Attention Modelling in a Cocktail Party Scenario," 2012.

[23] Cherry, E. C., "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, 25(5), pp. 975–979, 1953.

[24] LeDoux, J. E., "Emotion, memory and the brain," *Scientific American*, 270(6), pp. 50–57, 1994.

[25] Eschrich, S., Münte, T. F., and Altenmüller, E. O., "Unforgettable film music: the role of emotion in episodic long-term memory for music," *BMC neuroscience*, 9(1), p. 48, 2008.

[26] Dubois, D., Guastavino, C., and Raimbault, M., "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta acustica united with acustica*, 92(6), pp. 865–874, 2006.

[27] Ananthabhotla, I., Ramsay, D., and Paradiso, J., "HCU400: An Annotated Dataset for Exploring Aural Phenomenology through Causal Uncertainty," *International Conference on Acoustics, Speech, and Signal Processing*, 2018, under review; `http://arxiv.org/abs/1811.06439`.

[28] Quirin, M., Kazén, M., and Kuhl, J., "When nonsense sounds happy or helpless: the implicit positive and negative affect test (IPANAT)." *Journal of personality and social psychology*, 97(3), p. 500, 2009.

[29] Bartlett, J. C., "Remembering environmental sounds: The role of verbalization at input," *Memory & Cognition*, 5(4), pp. 404–414, 1977.

[30] Mitchell, H. F. and MacDonald, R. A., "Remembering, Recognizing and Describing Singers' Sound Identities," *Journal of New Music Research*, 40(1), pp. 75–80, 2011.

[31] Conrad, R., "The developmental role of vocalizing in short-term memory," *Journal of Memory and Language*, 11(4), p. 521, 1972.

[32] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O., "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, pp. 18–25, 2015.

[33] Giannakopoulos, T., "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," *PloS one*, 10(12), 2015.

[34] Font, F., Brookes, T., Fazekas, G., Guerber, M., La Burthe, A., Plans, D., Plumbley, M. D., Shaashua, M., Wang, W., and Serra, X., "Audio Commons: bringing Creative Commons audio content to the creative industries," in *Audio Engineering Society Conference: 61st International Conference: Audio for Games*, Audio Engineering Society, 2016.

[35] Bainbridge, W. A., Isola, P., and Oliva, A., "The intrinsic memorability of face photographs." *Journal of Experimental Psychology: General*, 142(4), p. 1323, 2013.

[36] Woods, K. J., Siegel, M. H., Traer, J., and McDermott, J. H., "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, 79(7), pp. 2064–2072, 2017.