Compression by Content Curation: An Approach to Audio Summarization Driven by Cognition

[AUTHOR INFORMATION REDACTED]

As we move towards an increasingly IoT-enabled ecosystem, we find that it is easier than ever before to capture vast amounts of audio data. However, there are many scenarios in which we may seek a "compressed" representation of an audio stream, consisting of an intentional curation of content to achieve a specific presentation outcome – a background soundtrack for studying or working; a summary of salient events over the course of a day; or an aesthetic soundscape that evokes nostalgia of a time and place. In this work, we present a novel, automated approach to the task of "compression" by content curation, built upon the tenets of auditory cognition, attention, and memory. We first expand upon the experimental findings in [1] and [21] – which demonstrate the relative importance of higher-level gestalt and lower level spectral principles in determining auditory memory – to design corresponding computational feature implementations enabled by auditory saliency models, deep neural networks for audio classification, and spectral feature extraction. We then develop a tool to form several 30-second binaural presentations from eight-hour ambient audio recordings captured at our institution by surfacing and mixing clips that fall at the extrema of the aforementioned feature axes. We finally conduct an evaluation with n=50 participants to illustrate the relationship between our cognitively-inspired feature space and a user's perception of the resulting presentation. Through this work, we suggest rethinking traditional paradigms of compression in favor of a content selection approach that is goal-oriented and modulated by principles of auditory cognition.

CCS Concepts: • **Human-centered computing → Sound-based input / output**; **Auditory feedback**;

Additional Key Words and Phrases: audio, summarization, compression, cognition, curation, salience, memory, attention

ACM Reference Format:
[Author Information Redacted]. 2018. Compression by Content Curation: An Approach to Audio Summarization Driven by Cognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 154 (December 2018), 13 pages. https://doi.org/10.1145/3287032

1 INTRODUCTION

In a crowded room, the whisper of our name has the power to immediately grab our attention. Fascinatingly, this occurs despite our inability to recall anything the whisperer had uttered just before.

This anecdote belies a powerful feature of cognition– the lowest levels of our pre-conscious auditory processing rely on high-level, gestalt semantics of the auditory objects we encounter. Studies of Event Related Potentials (ERPs) demonstrate this, as pre-attentive characteristics of these neurological signals are invoked in response to changes in both low-level acoustic changes (like a sudden loud noise) as well as high-level semantic ones (like an animal name unexpectedly appearing in a list of fruits) [12, 22, 23, 25].

A review of auditory perception and taxonomy research reveals that people do indeed conceive of sounds they encounter in the language of higher level semantics first, and only when a sound's source object becomes ambiguous– or *causally uncertain*– do they tend to resort to acoustic features for distinction [9]. In previous work,

Author's address: [Author Information Redacted].

2474-9567/2018/12-ART154 $15.00
https://doi.org/10.1145/3287032

the authors of [1] collected thousands of human labels about the high-level concepts that affect perception (such as a sound's emotionality, ease of visualization, or causal uncertainty (Hcu)) for 400 sounds. These sounds were intentionally curated to span from the easily recognized (low Hcu) to the extremely ambiguous (high Hcu). In [21], they mapped out the relative ease with which each of these sounds is recalled as a function of the high-level feature labels obtained in [1] and low-level spectral features, to analyze the roles of these features in auditory memory formation. The authors' results underscore the complexity of auditory cognition and the importance of high-level semantic features when modeling attention and memory.

Models that can extract these high-level features from raw audio in the general case have traditionally been intractable. It follows that general computational models of auditory attention and memory have been out of reach. Fortunately, recent advances in deep learning have demonstrated success at representing large, general classes of human knowledge with compressed embeddings [17, 18]. We've also seen the emergence of deep learning models that extract hundreds of labeled classes from raw audio [11]; sentiment models that successfully classify sentiment in images and text [24, 26]; and optimization techniques that allow these sophisticated models to run with a small memory and power impact on affordable hardware in near real-time [2, 16]. Combined with low-level acoustic salience models also common in the literature (see Section 2), these tools give us the ability to build a complete palette of cognitively meaningful analysis tools for audition.

In this paper, we extend the earlier exploration of auditory cognition with state-of-the-art machine learning techniques to enable a novel audio summarization approach that is based on both high-level and low-level cognitive principles. Specifically, our contributions in this work are as follows:

(1) We develop computational feature implementations of the human-annotated higher-level and lower-level principles of audition detailed in [1, 21], employing machine learning techniques to represent higher-level principles for the first time.
(2) We build these features into an automated tool designed to surface clips from hours of ambient audio recordings that are, for example, semantically notable (a new sound object), acoustically salient (loud or dissonant), emotional, or more or less memorable (based on the work in [1, 21]).
(3) We evaluate the approach by collecting three sets of 8-hour ambient audio recordings from contrasting locations at our institution, and apply the tool to this data to generate several 30-second binaural presentations according to the feature strategies mentioned in (2). We test an assortment of the presentations in a user evaluation study with n=50 participants, who are asked to rank the mixes according to several perceptual classes – i.e., whether a mix is appropriate for socializing, might be distracting from focused work, or serves as a remembrance for an entire day. We demonstrate a statistically significant relationship between our feature strategy space and a user's perception of a generated presentation.

Drawing from our results, we suggest that this cognitively-inspired approach enables us to re-think the way that extensive audio captured in changing environments (such as for lifelogging, ecological monitoring, etc) is condensed or summarized, and demonstrate a content curation-oriented methodology to achieve aesthetic presentations for a variety of intents: from a summary of memorable events to an aid that recovers forgettable but semantically novel moments; from background ambiance devoid of distraction to a foreground filled with emotion.

2 RELATED WORK

Automated "compression" of multimedia content is certainly not a new problem, especially under the framing of audio summarization. Several prior contributions in this space have used simple time and frequency features (such as pitch and pause) [10] or more complex spectral decomposition and filtering [7, 8, 27] to select excerpts of interest from a larger recording, which are then assembled together. In the realm of music, thumbnail generation has often been considered a holy-grail task; work by [5], for example, tackles this problem by performing

self-similarity analysis on a spectral representation. This class of work has typically capitalized on statistical variations in distributions of audio features to identify excerpts that should be present in a summary; they do not, however, exploit our knowledge of human perception, attention, and memory to drive the selection process.

The task of forming acoustical summaries using cognitive salience models is first attempted in [19]. The authors employ the approach first suggested in [15] and further explored in [6, 13, 14], which entails the generation of salience time-frequency maps from perceptually motivated kernels convolved with a spectrogram at different resolutions, to select samples from urban and environmental recordings that are most easy for users to associate with a specific location. While this is a significant initial step, our work expands the boundaries of this problem in two ways: first, we aim to reach beyond the space of spectral features or low-level salience to tap into human gestalt processing, through the use of deep networks for audio event classification. Secondly, and more broadly, we aim to explore the relationship between our feature space and the response elicited in users when these features are used to curate content. We posit that manipulating these gestalt and acoustic features in accordance with our understanding of attention and memory (shaped by [21]) might allow us to generate a "compressed" audio presentation that, for example, serves as a background track for studying, an ambient audio for socializing, evokes nostalgia, or simply records salient auditory events over the course of a day. Over the course of this work, we discuss both the novel suite of feature analysis techniques that enable these aims and the causal relationships between the generated content and a user's perception.

3 COGNITIVELY-INSPIRED CONTENT CURATION

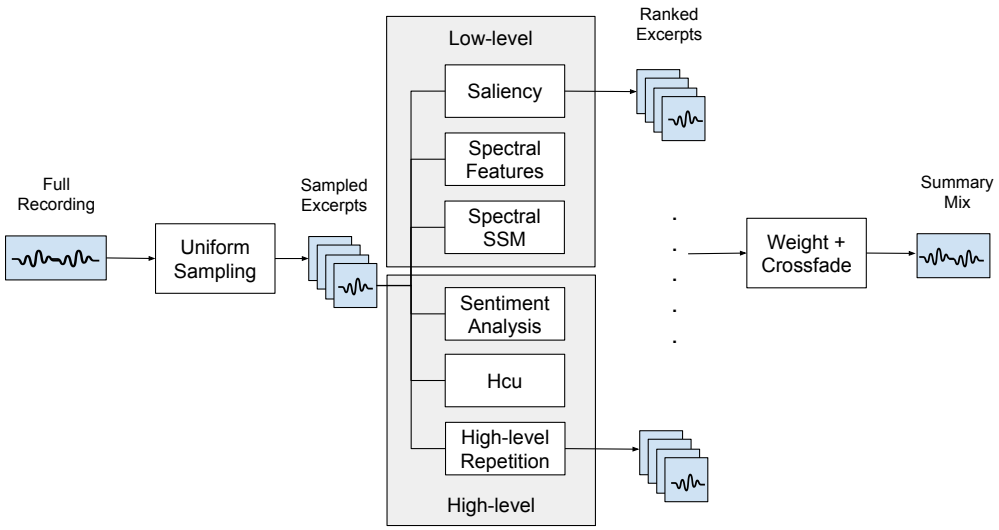

Fig. 1. An illustration of the cognitive analysis and summary generation tool.

3.1 Overview

An overview of our approach is as follows: from an uninterrupted audio recording, we select 3-second audio excerpts at equally-spaced time intervals throughout the recording, each with the same pre-determined duration.

We then extract the value of each of our features (detailed below) for each excerpt in the set. Our implementation outputs a ranking ordered by feature value assigned, and a set of configuration parameters optionally determine the relative weighting of each feature ranking, which are combined to select the subset of excerpts forming the final audio presentation (see Section 3.3 for more details). The selected excerpts are finally cross-faded in chronological order and output as a single track. Figure 1 provides a detailed illustration of the data flow in the system.

3.2 Feature Implementations

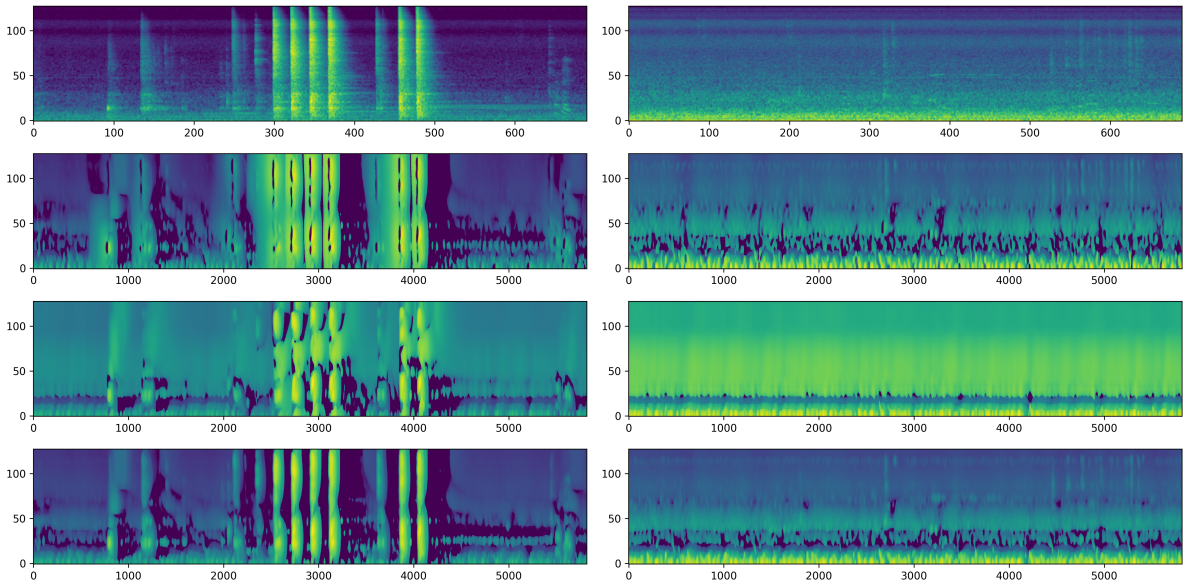

Fig. 2. Examples of saliency maps for two different audio samples; Top to bottom: Original spectrogram, temporal map, frequency map, intensity map. An increase in saliency is indicated by a lighter coloring.

3.2.1 Auditory Salience. To obtain a measure of auditory salience, we build upon the implementation mentioned in [21], inspired by the original model proposed by [15]. The saliency model consists of temporal, frequency, and intensity kernels convolved against a melspectrogram at multiple resolutions to obtain three time-frequency maps. We summarize the information from the saliency maps by computing the peak energy, the total energy, and the number of 2D peaks (local maxima obtained from a wavelet-based peak finder) appearing in each map for every sample. A final ranking for each of these three features is obtained by applying an equal weighting to the statistics computed for each kernel type. Figure 2 provides an example of the resulting salience maps for two different audio excerpts.

3.2.2 Spectral Self-similarity. Though not an explicit component of the previous work in [21] or [1], we infer from literature in the cognitive sciences [4, 9] that the notion of repetition (often framed as ecological frequency) is a driving force in determining the way an audio presentation is received or attended to. From a low-level standpoint, we assess repetition through a measure of self-similarity, first employed the context of audio by [5]. To obtain a score revealing how similar an audio excerpt is relative to itself and all other sampled audio excerpts, we first compute a magnitude STFT for each excerpt with a 512 FFT bins and a hop size of 512 samples.

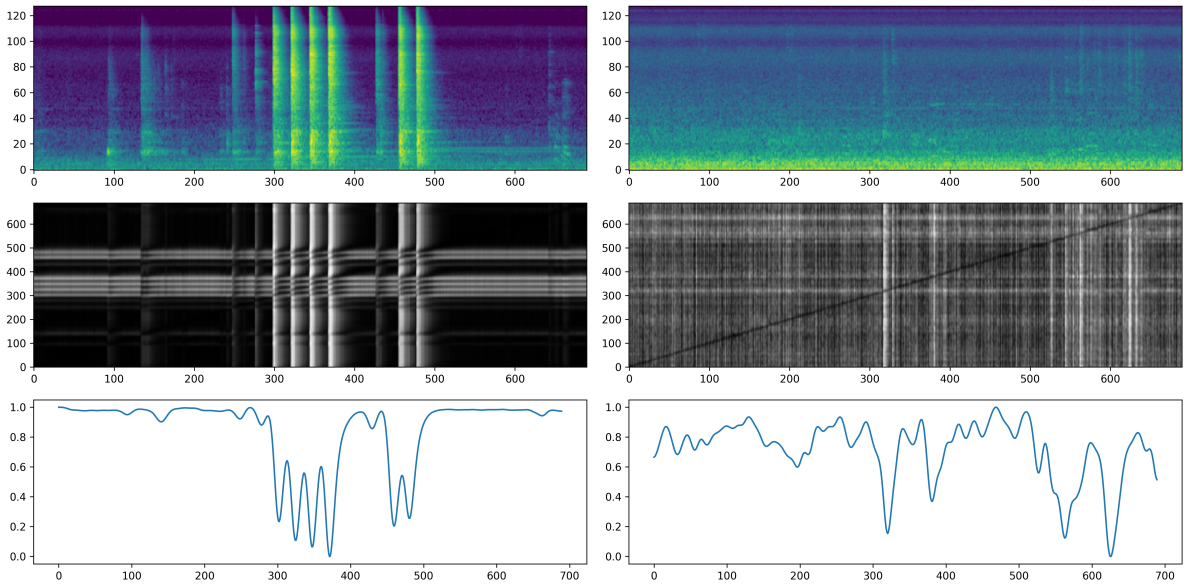

Fig. 3. An example of the self-similarity analysis procedure; we show two different sets of multiple excerpts resulting from the uniform sampling procedure, which are downsampled and concatenated along the time dimension as detailed in Section 3.2.2. Top: A concatenation of the spectrograms of each of the excerpts in the set; Middle: self-similarity matrix computed from the concatenated representation; Bottom: novelty curve derived from the self-similarity matrix.

To decrease the computational overhead, each STFT is then smoothed along the time axis with a width of 10 units, and down-sampled along the same axis by a factor of 10. After concatenating the magnitude STFTs from all of the excerpts along the time dimension to create a single 2-dimensional representation, we compute a self-similarity matrix as the cosine distance between each pair of 512 unit time vectors in the new representation. More formally, for a concatenated STFT representation S with time dimension T and frequency dimension F, the self-similarity matrix is defined as:

$$SSM(n,m) = \frac{\vec{S(n)} \cdot \vec{S(m)}}{||S(n)||||S(m)||}, n, m \in T \tag{1}$$

We then obtain a novelty curve by summing along the time or frequency axis; the total novelty within the bounds of a single excerpt is assigned as the self-similarity score to that excerpt. This process is illustrated by Figure 3.

3.2.3 Spectral Feature Extraction. We additionally include the low-level spectral features found in [21] as features in our suite. These cover (1) the ratio of harmonic to percussive energy; (2) the bass/mid and treble/mid energy ratios; and (3) the variation and range of the pitch contour identified using a low-pass filtered f0 extraction technique.

3.2.4 Gestalt Features via Deep Learning. As determined in [21], there are many high level descriptors of sound objects that dominate our perception, attention, and memory. In this paper, we develop computational equivalents of three of these features using deep learning techniques– Causal Uncertainty (Hcu), Sentiment, and High-Level Repetition.

In order to quantify these high-level descriptors, the first step must be to identify and label sound objects from the raw audio stream. This is possible using Google's AudioSet network, which is trained on a large YouTube corpus to classify hundreds of sound sources. Unfortunately, Google has not yet released a full end-to-end version of this network, opting instead to release the first half of it– called VGGish– which simply embeds incoming raw audio as a 128-dimension vector. This vector is designed to capture and disambiguate features that correspond to semantically meaningful information. We use it to transform the raw audio into a representation that should primarily capture when the *objects* or *sound sources* in a scene change (and by how much) as opposed to when and how much the sounds themselves change.

We additionally used DataArt's state-of-the-art publicly available implementation of the full AudioSet network that is trained using Google's VGGish and Youtube-8M dataset [20]. However, we found that this classifier was not always robust, and struggled to correctly label the sounds in our environmental recordings. Given this limitation, we decided to use the native VGGish embedding where possible.

Employing this infrastructure, our final high-level features are as follows:

(1) Hcu, calculated using the L2 spread between VGGish embeddings of dozens of overlapping segments in a 3-second excerpt. The more diverse the semantic information in a short period of time, the higher the Hcu.
(2) High level semantic novelty (or High-level repetition), calculated with the same algorithm as for low-level spectral self-similarity, except based on the centroid of each frame's VGGish embeddings. Frames whose embeddings stand out in this context represent novel semantic information about the sounds in the raw audio.
(3) Sentiment or affect in the audio. This is the most speculative, as it requires accurate classification labels from the DataArt AudioSet model. In this instance, we take all of the possible classes from our the model, feed the text label for each into the senti-wordnet model to identify its sentiment score [3], and scale that score by the model's class probability for that label to get a final sentiment estimate. In this way, we average over many noisy observations, incorporate many uncertain classifications into the model, and derive a simple probabilistic score for the likely affect in each frame.

3.3 Excerpt Ranking and Selection

In our setup, we extract low- and high-level features from about 4,800 3-second excerpts in a given 8 hour recording (resulting in an analysis of approximately 4 hours of audio per recording). The tool is then tasked with the selection of 15-20 excerpts to make a 30 second auditory presentation, where the excerpts are assembled together with crossfades.

To make the selections, the excerpts are ranked with respect to a feature or set of features that a user may specify. In the case that any single higher or lower-level feature from our suite is specified, excerpts are simply drawn from the top or bottom of the ordered ranking (i.e. taking the top 15 most salient or the top 15 most semantically novel clips); in the case that assessing excerpts by overall "memorability" is specified, the tool performs a linear weighted average of the relevant features, following the model presented in [21]. We use the shapely coefficients from this work as the weights and give high- and low-level features equal value, with a small modification wherein our metric for Hcu is exclusively used to represent the high-level principles to ensure robust predictions.

Finally, we incorporate a "baseline" strategy into the tool, where 15-20 excerpts are simply chosen at equally spaced intervals in time from the set of excerpts being analyzed, without any feature extraction and ranking. This serves as a benchmark for comparison in our evaluation study (see Section 5).

4 DATASET

To evaluate our approach, we first form our dataset by recording 8 hours of audio (from 9am to 5pm) at 3 different locations across our institution. The recording venues included a laboratory space inhabited by approximately 15 students spread across desks, workbenches, and a meeting space; a public "atrium" space with heavy traffic and conversation during the mid-morning and afternoon; and a setting just outside of an academic building, at a street intersection with frequent vehicle traffic and construction activity.

The raw audio is recorded in each location using a single microphone capturing 4 channels of audio in an Ambisonic-A format. We then convert this audio to a binaural rendering using Facebook's Spatial Audio Workstation, to create spatial presentations that can be evaluated over personal headphones. Throughout our evaluation, we perform the feature ranking analysis on the audio from a single omni-directional channel taken from the raw ambisonic configuration; to create the final mixes, however, we choose the corresponding, time-aligned excerpts from the binaural rendering, to eliminate the possible distortion effects of a generic Head-related Transfer Function on the spectral content used for feature extraction.

5 EVALUATION

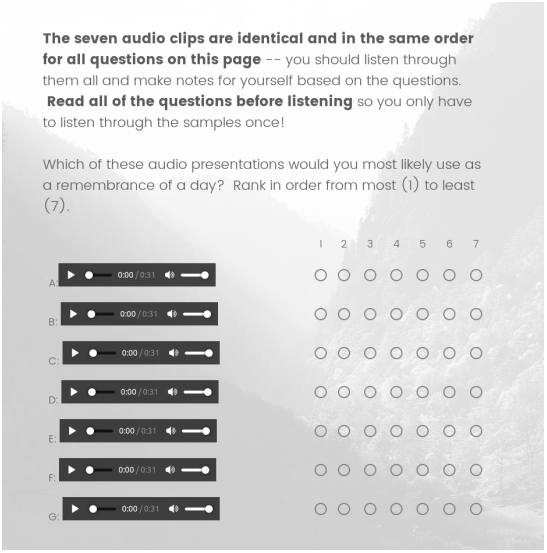

Fig. 4. Screenshot of the evaluation survey.

To understand the relationship between the cognitive features implemented in our tool and the various perceptual goals that the generated audio might align with, we conduct a survey requiring participants to rank audio mixes from each of the three recording locations according to a set of questions. More specifically, we choose the set of questions summarized by the keywords given in Table 1 (the full questionnaire is given in Table 3) and hypothesize as to the dominant feature strategies that might produce mixes most in line with the questions. We then take inspiration from these hypotheses to select a set of 7 binaural mixes per recording location, which participants must rank from most to least likely in terms of accordance with each question. It is important to note that the set of mixes is intentionally not identical across recording locations; this is largely motivated by a trade-off between audio summaries generated accurately from a computational standpoint and the need for more

diverse, relevant presentations given the location. We present a detailed discussion of this experimental design decision in Section 7.1.

Goal	Feature Hypothesis
Remembrance	Most Memorable
Notable	Most Memorable, Most Semantically Novel
Relaxing	Least Salient Energy
Distracting	Most Salient Energy
Socializing	Most Affect
Evolution	Baseline

Table 1. The feature strategies we hypothesize are mostly likely to produce the binaural presentations that are suitable for the listed goals (see Table 3 for the corresponding survey questions); these inspire our final per-location mix selections, given in Table 2

.

"Atrium" Space	Outside	Laboratory Space
Most Memorable	Most Memorable	Most Memorable
Most Semantically Novel	Most Semantically Novel	Most Semantically Novel
Most Acoustically Self-similar	Most Acoustically Self-similar	Most Acoustically Self-similar
Least Acoustically Self-similar	Most Semantically Novel	Most Semantically Novel
Most Salient Energy (Total)	Most Salient Energy (Peak)	Most Salient Energy (Total)
Most Affect	Least Salient Energy (Peak)	Most Affect

Table 2. The final selections of the feature strategies corresponding to the binaural presentations corresponding to each recording venue.

Keyword	Survey Question
Remembrance	Which of these audio presentations would you most likely use as a remembrance of a day?
Notable	Which of these audio presentations would you most likely use as a compilation of notable events in a space?
Relaxing	Which of these audio presentations would you most likely use as a background track for sleeping/ relaxing?
Distracting	Which of these audio presentations would be most distracting to focused work?
Socializing	Which of these audio presentations would be most comfortable for socializing?
Evolution	Which of these audio presentations best captures the evolution of the environment over the course of the day?

Table 3. A reference mapping between the goal keywords we use throughout the work and the survey questions to which they correspond.

Table 2 shows the final set of mixes (defined by feature strategy) chosen per recording location, and a screenshot of the evaluation interface can be found in Figure 4. We recruited 75 participants to complete the evaluation – a small number (<10) were recruited from our institution on a voluntary basis, and the remaining were recruited from Amazon Mechanical Turk for a small fee per survey. The audio mixes pertaining to each feature strategy and recording location and an interactive demo can be found at [redacted].

	"Atrium" Space	Outside	Laboratory Space
Remembrance	Most Memorable	Baseline	Most Salient Energy
Notable	**Least Acoustically Self-Similar**	Most Semantically Novel	**Most Acoustically Salient Energy**
Relaxing	**Most Acoustically Self-Similar**	**Least Semantically Novel**	Least Semantically Novel
Distracting	**Least Acoustically Self-Similar**	**Most Semantically Novel**	**Most Acoustically Salient Energy**
Social	**Most Memorable**	**Smallest Acoustically Salient Peaks**	Most Memorable
Evolution	**Least Acoustically Self-Similar**	Most Semantically Novel	**Most Semantically Novel**

Table 4. Top-ranked feature strategy for a given goal and recording space; bold if the p-value corresponding to the set of rankings is less than 0.025

	"Atrium" Space	Outside	Laboratory Space
Remembrance	Baseline	Least Semantically Novel	Most Acoustically Self-Similar
Notable	**Most Acoustically Self-Similar**	Least Semantically Novel	**Most Acoustically Self-Similar**
Relaxing	**Most Acoustically Salient Energy**	**Most Semantically Novel**	Least Affect
Distracting	**Least Semantically Novel**	**Smallest Acoustically Salient Peaks**	**Most Acoustically Self-Similar**
Social	**Least Semantically Novel**	**Largest Acoustically Salient Peaks**	Most Acoustically Self-Similar
Evolution	**Most Acoustically Self-Similar**	Most Memorable	**Baseline**

Table 5. Bottom-ranked feature strategy for a given goal and recording space; bold if the p-value corresponding to the set of rankings is less than 0.025

6 RESULTS

Collected survey responses were first filtered by duration and completion percentage (users who took less time than was required to listen through all audio samples were eliminated). We then examined the data for outliers using clustering techniques based on the kendall distance (a pair-wise technique used in ranking analysis). We were left with 50 trustworthy participants.

Our analysis first consisted of a few basic statistics– the average rankings of each feature strategy for a particular goal in a particular space, and the pair-wise probability table to show how frequently one type of feature strategy outperforms another for a goal/location pair. We then further examine the data through marginal plots which show the percentage of times a particular mix was chosen for each rank. We examined data broken out across all seven original rank options; for ease of analysis, we also look at consolidated marginal ranking plots where counts in positions 1-2, 3-5, and 6-7 are grouped together. We use a χ^2 test to make sure the marginal data is significantly different than pure chance.

In Table 4 and Table 5, we show the feature type that best describes the differences in the highest and lowest average rankings. Bold feature types have a p value of < 0.025 compared with random selection. Additionally, we provide several examples of marginal plots to aid our discussion of trends in perception (see Section 7), shown in Figures 5, 6, and 7.

7 DISCUSSION

7.1 Mix Selection and Edge Cases

Before discussing the survey data, we start with our selection criteria for the per-location mixes presented in the questionnaire (Table 1 and Table 2). We immediately found with our tool interesting edge cases that were not useful for our survey– for instance, the "least semantically novel" summary of the laboratory space was composed entirely of doors opening and closing. While this objectively matches expectations of a properly functioning tool (door closing is an extremely common background sound with a clear semantic label), the mix subjectively

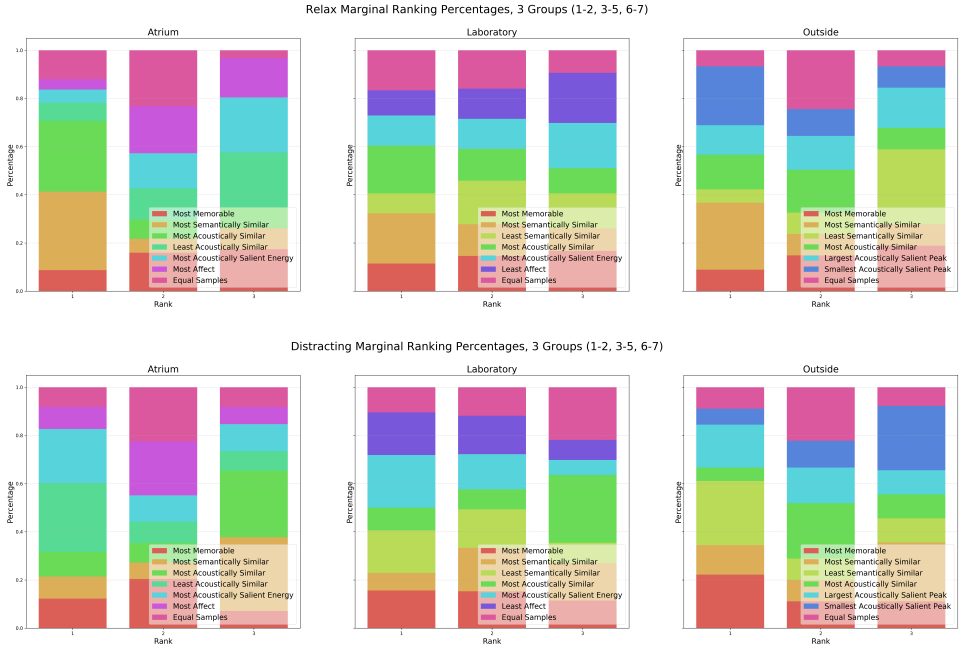

Fig. 5. Marginal Ranking plots for the Relaxing and Distracting goals. These show the percentage that each mixing strategy appears in the top, middle, and bottom ranks as evaluated by users.

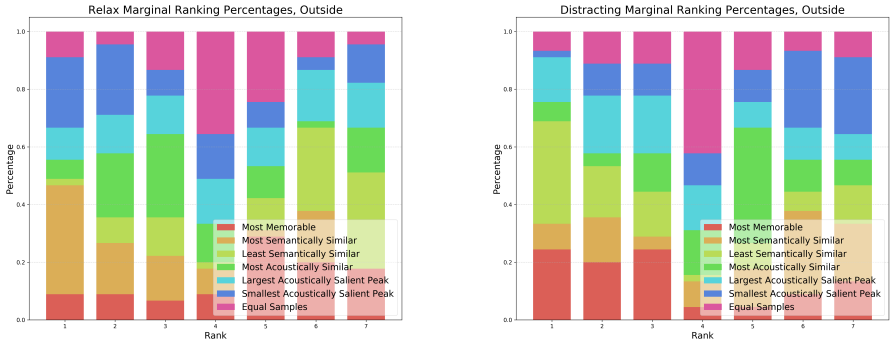

Fig. 6. Marginal Ranking plots for the Relaxing and Distracting goals outside. These show expanded views of the plots in Figure 5 for every rank.

does not serve as a summary or representation of an audio environment along *any* of the examined perceptual dimensions and was likely to skew the results of the study for that particular recording location.

We also found several cases where the "uninteresting" end of the spectrum of low-level features values (i.e. the most acoustically self-similar, least salient, least pitched, etc) tended to converge to the same subset of excerpts

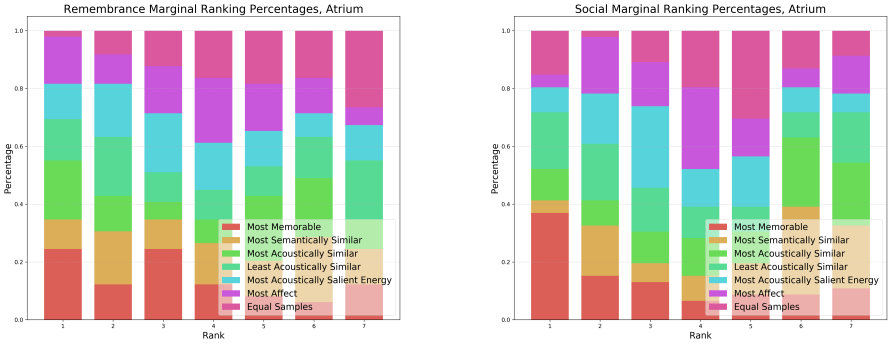

Fig. 7. Full marginal rankings for the "Social" and "Remembrance" goals, in the atrium space. No strong trends are evident outside of a preference for memorability score.

from the analysis set. Interestingly, these scales *diverge* as they move towards the positive end (the most salient or least acoustically similar sounds) of the feature spectrum. To maintain diversity in the selected mixes per recording location and to be able to disambiguate between feature strategies, we intentionally limited overlap in the negative cases for these lower-level features.

Finally, for the "outside" location, there were very few sound sources that were pitched or organically generated– this soundscape was made almost entirely of cars and trucks. For this case, the sentiment or affect feature strategies did not provide mixes representative of corresponding cognitive principles, and were discarded in favor of other feature strategies.

7.2 Trends in Mix Rankings

As demonstrated by Table 4 and Table 5, not every mix resulted in a χ^2 p-value that suggested non-random selection– most notably, none of the "remembrance" rankings for any environment showed a strong p-value. This hints at the complexity behind an ideal model of memorability, which was likely not captured in entirety with a sparse and simple linear combination of our features. We also generally see from these tables that self-similarity in low-level acoustic feature space as well as high-level semantic feature space are the two most important and most useful feature strategies that can be applied to the dataset.

Figure 5 shows the rank-consolidated marginal plots for the two perceptual objectives with the clearest statistical trends– "relaxing" and "distracting". For these objectives, we see the strongest trends in high-level semantic similarity (mixes with consistent sound sources are relaxing), and weaker trends in low-level acoustic similarity (consistent sound sources have similar sound textures, which is also relaxing). Interestingly, in the "outside" condition, we find that the correlation between low- and high- level similarity diverges (some mixes have similar sources but different textures and vice versa) and in these cases high-level semantic similarity (or dis-similarity) emerges as the most useful way to create a relaxing or distracting mix. In environments where many sound objects share low-level features– or in environments where the same sound source may appear with varied texture– high-level semantic features become the only way to disambiguate the interesting from the mundane. The strength of these trends demonstrates that our tool creates clearly superior and inferior mixes along the "relaxing" and "distracting" axes, and we find that high level semantic similarity is the most important driver of that difference.

In Figure 6 we see the full marginal plot of the "relaxing" and "distracting" rankings from the "outside" recording – this confirms in more detail the primacy of semantic similarity or dissimilarity over acoustic features. We also see a hint that our memorability score has captured something useful (though secondarily), with clear boost in the top and bottom three ranks of "distracting" and "relaxing" mixes respectively.

In our analysis of mix rankings in response to the "social" and "remembrance" questions in the study, we see some of the weakest trends (Figure 7. Social rankings seem to favor a diversity of salient, semantically relevant events– in the case of the atrium ranking, we see some evidence for a trend in the memorability mix. In general, these results suggest a complicated relationship we have yet to effectively decompose; however the dominant features have the most clearly direct cognitive parallels (i.e. salience, semantic similarity, memorability, etc).

Finally, the "baseline" mix strategy of equally spaced samples was not a common selection for the "evolution of a space" perceptual objective, despite our hypothesis. It seems– particularly in a short 30 second format– that people are attuned to changes in the soundscape more readily captured by acoustic and semantic self-similarity, despite the fact that a naive sampling strategy is statistically most representative of a recording.

8 CONCLUSION AND FUTURE WORK

In this work, we demonstrate a novel approach to audio "compression" or "summarization" by means of cognitively-inspired content curation. We develop a suite of computational features that mimic the higher-level and lower-level principles of auditory memory and attention as outlined in prior work, employing deep learning techniques to achieve the former for the first time. We apply these features to form short presentations from lengthy environmental recordings and demonstrate a correlation between our feature space and the perceptual attributes of the auto-generated presentations.

We acknowledge that the approach can benefit from further research in two dimensions – (1), in which the fundamental research pertaining to statistical models of auditory memory and attention presented in [21] is expanded to incorporate and investigate spatialized audio, audio presented in an uncontrolled context, and user-driven input such as bio-sensory signals or an assessment of auditory exposure; and (2), in which the feature design is more explicitly constrained by computation requirements. Given the measures of self-similarity, for example, the current implementation does not immediately lend itself to an on-line, real-time curation process for streaming audio – future work in this regard is necessary to extend the realm of application of this work from tens of hours of audio to hundreds or thousands of hours.

Despite the work to come, we believe that this work presents an important contribution to the research problem of large-scale audio consumption – that given an experience-oriented audio capture context instead of a task-oriented one (such as lifelogging or ecological monitoring instead of surveillance), we might consider compressed or summarized representations with *aesthetic* aims over objective ones; that exploring content along subjective, perceptual dimensions might provide a novel, valuable means of interfacing with hours of recorded audio.

ACKNOWLEDGMENTS

This work was funded in part by the 2019 AI Grant. The authors would like to thank the voluntary study participants for their time and dedication.

REFERENCES

[1] Ishwarya Ananthabhotla, David B Ramsay, and Joseph A. Paradiso. 2019. HCU400: An Annotated Dataset for Exploring Aural Phenomenology through Causal Uncertainty. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

[2] Renzo Andri, Lukas Cavigelli, Davide Rossi, and Luca Benini. 2016. YodaNN: An ultra-low power convolutional neural network accelerator based on binary weights. In *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 236–241.

[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. [n. d.]. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.
[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] Matthew L Cooper and Jonathan Foote. 2002. Automatic Music Summarization via Similarity Analysis.. In *ISMIR*.
[6] Varinthira Duangudom Delmotte. 2012. *Computational auditory saliency*. Ph.D. Dissertation. Georgia Institute of Technology.
[7] Georgios Evangelopoulos, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, A Zlatintsi, and Yair Avrithis. 2008. Movie summarization based on audiovisual saliency detection. In *2008 15th IEEE International Conference on Image Processing*. IEEE, 2528–2531.
[8] Georgios Evangelopoulos, Athanasia Zlatintsi, Georgios Skoumas, Konstantinos Rapantzikos, Alexandros Potamianos, Petros Maragos, and Y Avrithis. 2009. Video event detection and summarization using audio, visual and text saliency. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3553–3556.
[9] 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.
[10] Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin. 1999. Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 489–498.
[11] 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. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://arxiv.org/abs/1609.09430
[12] 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.
[13] Ozlem Kalinli and Shrikanth Narayanan. 2009. Prominence detection using auditory attention cues and task-dependent high level information. *IEEE Transactions on audio, Speech, and language processing* 17, 5 (2009), 1009–1024.
[14] Ozlem Kalinli, Shiva Sundaram, and Shrikanth Narayanan. 2009. Saliency-driven unstructured acoustic scene classification using latent perceptual indexing. In *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*. IEEE, 1–6.
[15] Christoph Kayser, Christopher I Petkov, Michael Lippert, and Nikos K Logothetis. 2005. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology* 15, 21 (2005), 1943–1947.
[16] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. 2015. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530* (2015).
[17] Hugo Liu and Push Singh. 2004. ConceptNetâĂŤa practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.
[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[19] Damiano Oldoni, Bert De Coensel, Annelies Bockstael, Michiel Boes, Bernard De Baets, and Dick Botteldooren. 2015. The acoustic summary as a tool for representing urban sound environments. *Landscape and Urban Planning* 144 (2015), 34–48.
[20] Igor Panteleev. 2017. Devicehive Audio Analysis. (Oct. 2017). https://github.com/devicehive/devicehive-audio-analysis
[21] David B Ramsay, Ishwarya Ananthabhotla, and Joseph A. Paradiso. 2019. The intrinsic memorability of everyday sounds.. In *AES Conference on Immersive and Interactive Audio*.
[22] 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.
[23] Joel S Snyder and Mounya Elhilali. 2017. Recent advances in exploring the neural underpinnings of auditory scene perception. *Annals of the New York Academy of Sciences* 1396, 1 (2017), 39–55.
[24] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1631–1642.
[25] István Winkler, Susan L Denham, and Israel Nelken. 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in cognitive sciences* 13, 12 (2009), 532–540.
[26] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
[27] Athanasia Zlatintsi, Petros Maragos, Alexandros Potamianos, and Georgios Evangelopoulos. 2012. A saliency-based approach to audio event detection and summarization. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 1294–1298.

Received February 2018; revised July 2018; accepted October 2018