Cognitive Content Curation: An Audio Summarization Tool Driven by Principles of Auditory Cognition

ANONYMOUS AUTHOR(S)

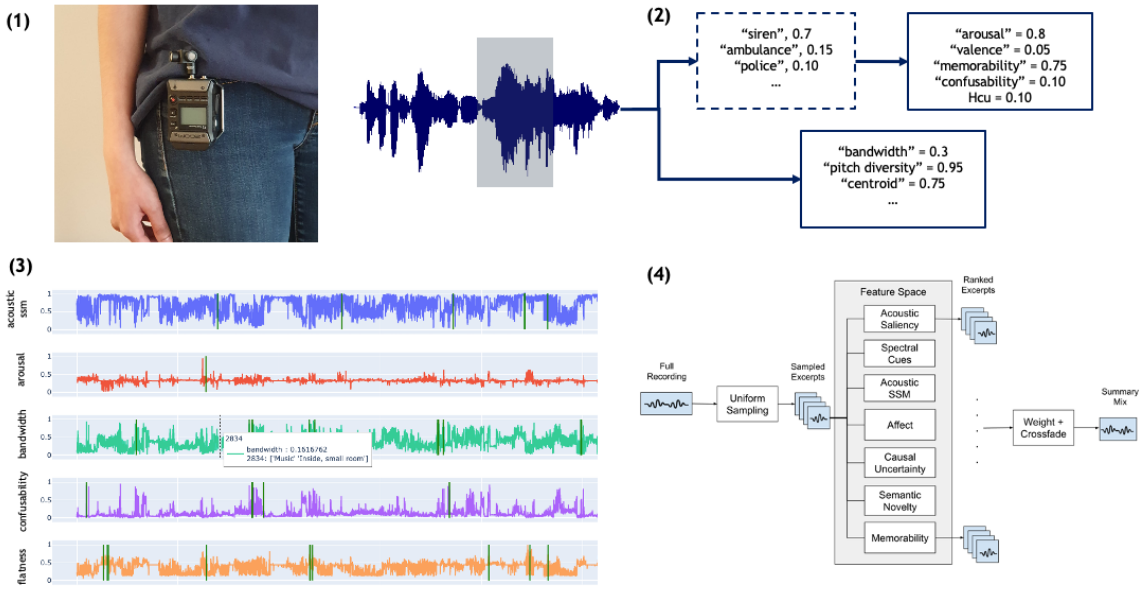

Fig. 1. Our audio summarization tool takes real-world, lifelogged audio (1) as input, uses cognitive principles to extract acoustic and semantic feature information from the audio (2, 3), and reassembles the highest or lowest scoring excerpts to form "summaries" (4). The method allows for summaries that elicit specific subjective, perceptual responses in users, and results in listening experiences that are emotional, compelling, and engaging.

As we move towards an increasingly IoT-enabled ecosystem, we find it easier than ever before to capture vast amounts of audio data. These databases of recorded sound require tools that can curate content to support experience-oriented user goals – from a background soundtrack for studying, to a summary of salient events over the course of a day, to an immersive soundscape for evoking nostalgia. We introduce an audio summarization system that generates various aesthetic outputs by drawing on principles of auditory cognition. We evaluate the system based on 60-second audio presentations that are automatically extracted and mixed from lifelogged audio recordings captured by participants over a span of 1-3 weeks, and highlight trends in the relationships between the system's cognition-inspired feature space and human perception of the

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

presentations. Through this work, we suggest re-thinking traditional approaches to the task of audio summarization to both consider experiential outcomes and to incorporate cognitively motivated principles.

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

Additional Key Words and Phrases: Audio, Summarization, Content Curation, Cognition, Perception, Deep Learning

ACM Reference Format:
Anonymous Author(s). 2018. Cognitive Content Curation: An Audio Summarization Tool Driven by Principles of Auditory Cognition. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Over the last decade, we have witnessed a massive shift towards ubiquity and capacity in audio capture. Low-power, always-on audio sensing technology in our homes and our phones makes it easy to record hours of uninterrupted data; the rapidly falling cost of storage infrastructure makes it even easier to archive. From lifelogged recordings to environmental monitoring databases, however, this trend has resulted in a paradox for consumption – the more audio there is, the harder it is for the average user to interact with the it, and the less likely they are to do so.

While research has produced many strategies for condensing large audio corpora into representations that are distilled in time, these approaches are frequently motivated by the goal of maximizing information in the output representation [3, 12, 17, 21, 29, 47, 48]. We argue that such approaches do not enable other, subjective modalities of engagement with or explorations of a specific body of audio, such as aesthetic or emotional modalities. For example, we find many means of distilling a body of audio so as to preserve only speech or detect new events in a soundscape, as may be useful in a memory aid or surveillance task, but few means to use the audio to create an appropriate background track for sleeping or studying, or for evoking nostalgia of a place and time. In addition, we note that most of the aforementioned systems distill audio by optimizing for a single, well-defined objective function; they do not represent tunable systems with exposed input parameters that can be reconfigured to generate perceptually diverse outcomes. We intuit that this shortcoming is driven by feature space design – these systems are largely built on feature representations that quantify statistical properties of the audio itself (spectral derivatives, event detection, audio quality, frame diversity, etc), instead of being rooted in the science of how the human brain perceives and processes sound.

We attempt to address this gap by constructing an audio summarization system with a feature space that is driven by the principles of auditory processing. Over the last few decades, seminal works in the empirical psychology literature have introduced the notion of "ecological" or "everyday" listening, suggesting that the processing of the sounds we interact with in our day-to-day environments are driven in part by various aspects of sounds themselves – by an interplay of high-level, “gestalt" properties of the sound source (i.e., its ambiguity, emotional impact, familiarity, or likelihood to conjure a mental image) and low-level, acoustic properties (i.e., sound loudness, pitch, timbre, etc) [13, 14]. Further work suggests that gestalt processing, most often in the form of sound source identification, is likely to dominate over acoustic processing [2, 6, 16, 20, 36, 37, 39]. Moreover, the relative importance of gestalt processing is hinted at in [4, 40], wherein researchers attempt to construct simple statistical models of auditory memory from measurements of these higher- and lower-level principles. In this work, we draw inspiration from the feature space of these models, and consider incorporating a selection of the aforementioned principles into the feature space of an audio summarization system – we expect that altering the relative weighting of these features at the input can produce a diversity of outcomes along emotional and aesthetic axes.

We are then left with the task of designing features that reflect these principles to the best extent possible, while still enabling an automated system. In [4, 40], acoustic information could be extracted directly from the

audio data; however, gestalt concepts required thousands of human annotations, performed by a crowd-sourced community. To solve this problem, we rely in this work on recent advances in Deep Neural Networks (DNNs) to mimic gestalt understanding, and boostrap feature extractors from labels in the small dataset presented in [4] to allow for gestalt concept estimation in real-world audio.

Combining these insights, we present an audio summarization system that draws on principles in auditory cognition to generate summaries of varying aesthetics from extended audio recordings. Specifically, we contribute the following:

(1) We construct a feature space derived from the gestalt, intrinsic memorability (which we collectively refer to as gestalt), and acoustic features examined in [4, 40], relying on DNNs to enable computational equivalents of the features which previously demanded human annotations.
(2) We build a tool that labels audio recordings with these features, surfaces short clips along the extremes of the features, and combines them in chronological order to generate audio summaries.
(3) We evaluate our system in the context of more than 800 hours of "lifelogged" audio recordings, wherein 10 participants each collect high-fidelity, first-person recordings over the span of 1 to 3 weeks, and provide quantitative and qualitative feedback regarding the summaries generated from their own data.
(4) We demonstrate the utility of our gestalt feature implementations in the context of audio summarization by showing that (1) acoustic and gestalt features statistically surface different content in long-term recordings; (2) that gestalt features are the strongest drivers of perceptual responses in study participants; and (3) that the presented system results in listening experiences that are immersive, intimate, artistically compelling, emotionally intense, and are suggestive of various user-highlighted use cases geared towards well-being.

Through this work, we suggest a novel paradigm for designing media summarization systems, and gesture towards a broader framework for capitalizing on elements of cognition to enable novel forms of media interaction.

2 RELATED WORK

The work presented in this paper lies at the intersection of HCI, auditory cognition, and deep learning. We discuss the canonical literature from each community that motivates and informs this work.

2.1 Approaches to Summarization

Content curation of large multimedia collections has been a problem of interest to the HCI community for many years, and has generated a significant body of literature. One approach condenses a large volume into a smaller presentation by selecting and recombining a subset of the content – this has been referred to as summarization [10, 32]. Summarization efforts in multimedia typically focus on building systems with a singular objective function, with the aim of maximizing information [3, 10, 12, 17, 21, 29, 32, 47, 48] or representativeness of/similarity to the larger media body [7, 8, 18, 38] in the output presentation. Common sub-themes of these goals include frame diversity and quality [10, 32], event detection [27, 31, 33, 50], and preserving distinct speech [46, 48]; less frequently, we find examples of more abstract goals, such as memory retention or a qualitative assessment that a summary is "good" – however, these systems are often explicitly constructed using the aforementioned information criteria as input features [26, 34, 50]. Our system aims to address some of the gaps in this literature by (1) considering a set of outcomes beyond information maximization to generate summaries that allow users to engage with their soundscapes from an aesthetic, emotional, therapeutic, or sound-design standpoint; and by (2) incorporating principles of perception and cognition into the feature design space to enable these outcomes.

2.2 Principles of Auditory Cognition

A review of auditory perception and taxonomy research reveals that listeners typically conceive of sounds they encounter in the language of higher level semantics first, and only when a sound's source object becomes

ambiguous – or *causually uncertain* – do they tend to resort to acoustic features for distinction. This idea was introduced in [13, 14] with Gaver's model of "ecological listening", suggesting that our consumption of and interaction with everyday sounds is primarily driven by our ability to estimate the source of the sound and the physical interactions that resulted in its production, and secondarily by acoustic and spectral properties. This dichotomy and hierarchy was reinforced by several later studies – [6, 20, 37] suggest that listeners rely on sound source and context/ location identification prior to acoustic features in categorization tasks; [2, 16, 36, 39] suggest that soundscapes with living/ organic elements (humans, animals, etc) are perceived differently and elicit different emotional responses than soundscapes with purely inorganic sounds, and point to the role that source attribution plays in determining a listener's response. In [5, 35], researchers attempt to quantify causal uncertainty (*Hcu*) from a listener's perspective, describe its complex relationship with a sound's typicality, familiarity, and ecological frequency, and demonstrate the role that the measure plays in sound organization and clustering tasks.

The interplay of higher and lower level processing is further corroborated by neurological observations. Studies of Event Related Potentials (ERPs) demonstrate that the earliest layers of our pre-conscious auditory processing rely on the gestalt semantics of the auditory objects we encounter, 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 the sound of a farm animal unexpectedly appearing in a series of urban sounds) [24, 42, 45, 49]. Measurement results in these works indicate that pre-conscious processing, attention, and memory of two events that might sound very similar – a snare drum and a gunshot, for instance – will vary drastically despite very close acoustic signatures.

In [4, 40], researchers first explored the interaction between higher- and lower-level concepts in ecological listening in a controlled setting. They assembled a dataset of 400 sounds called the *HCU400 Dataset* that intentionally spanned the spectrum of source ambiguity, and associated with them thousands of human annotated labels for high-level concepts (such as emotionality, ease of visualization, and an estimate of causal uncertainty), measures of intrinsic "memorability" and "confusability" resulting from thousands of iterations of a sound memory game, and feature values for low-level concepts (such as pitch, loudness, spectral salience, spectral self-similarity, etc). In attempting to understand the role that these features played in auditory memory formation, the researchers underscored what previous research suggests – that while auditory cognition is incredibly complex, high-level semantic features play a crucial role in the processes of perception, attention, and memory.

For our summarization system, we chose to create a feature space that is designed based on the conclusions from this literature and the values and scores assigned to sound objects in in [4, 40]. As shown in Figure 1(2), we wish to annotate sound excerpts with not only acoustic information, but also with semantic concepts that do not on their own predict user perceptual response, but capture shared conceptual information in the human listening experience that is difficult to estimate from acoustic information alone. While translating the acoustic feature implementations to an automated system is trivial, we highlight a need for an alternate means of extracting gestalt feature information from audio directly, to eliminate the expense and impracticality of human-in-the-loop annotation.

2.3 Deep Learning Representations for High-level Auditioning

In order to computationally replicate the gestalt feature association in [4, 40], we note that automatically estimating sound source labels and associated prediction uncertainty is a useful intermediate step, reflecting the role of sound source identification in auditory processing identified in the cognitive science literature [13, 14]. Here, we suggest using pre-trained neural networks as this intermediary.

Over the last several years, the deep learning community has made significant advancements in the development of audio classification and sound event detection systems [1, 19, 25, 41, 44], built using datasets of varying sizes, label quality, input and output representations. One of the most important research contributions to the problem

is Google's VGGish network [23], trained on their recently released AudioSet database [15], which consists of millions of labeled, isolated audio snippets taken from Youtube. The VGGish classification network, which boasts an AUC score of .92 and .90 respectively for video-level labeling and audio event detection respectively, was determined after the authors of the work applied several of the top-performing convolutional neural network (CNN) topologies for image classification to sound classification. In the version of the source code released to the public, the VGGish network estimates a 128-dimensional embedding vector from an approximately one-second audio sample, providing a means to conceptually relate the sound sources defining the audio snippet to other sound sources. In our system, we rely on this released version from Google, called YAMNet[1], to estimate source labels and embedding vectors from audio input.

3 FEATURE IMPLEMENTATIONS

Drawing from the gap identified in Section 2.1, the principles of audition highlighted in Section 2.2, and employing the infrastructure discussed in Section 2.3, we create seven classes of feature extractors to operate on raw ambient audio streams as inputs to our summarization system: Affect, Memorability, Causal Uncertainty, Semantic Novelty, Acoustic Saliency, Acoustic Self-similarity, Spectral Cues. These extractors produce scalar feature curves associated with the audio as a function of time. The choice of feature classes reflect the feature space in [4, 40], and we extend the work by presenting computational methods to generalize the gestalt feature classes to unseen and unannotated audio. Below, we summarize the motivation for each feature class and detail its implementation.

3.1 Bootstrapping Affect and Memorability

To construct a general approach for labeling sound objects with scores for "Memorability" and "Confusability" [40] and "Arousal" and "Valence" [4], we propose a strategy for mapping the human-labeled scores from the limited HCU400 dataset to Google's AudioSet ontology. To perform the bootstrapping, we use the pre-trained YAMNet to obtain the top 10 predictions for each sound in HCU400. Then, for every class label l in the AudioSet ontology, we wish to obtain a score S_l which represents a scalar gestalt feature value (for example, arousal). For every occurrence k of label l as a prediction in the HCU400 set, we compute S_l with the general form:

$$S_l = \frac{\sum_{k \in K} F_\theta(p^k(l)) \cdot S_k}{\sum_{k \in K} F_\theta(p^k(l))}$$

where F represents a heuristic weighting function with the parameters θ, and $p^k(l)$ represents the YAMNet prediction probability associated with l at the k-th instance. However, using this method allows us to derive gestalt/ memorability scores for only a subset of AudioSet labels, since not every label is represented in the HCU400 prediction set. Therefore, we then apply intuitive rules for propagating scores along the ontology hierarchy – child labels receive scores of parent labels, and parent labels receive mean scores of child labels – and finally assign median scores to every label we still cannot bootstrap.

Lastly, we wish to have a generalized means of using these results to predict the gestalt/ memorability score S_R associated with any new audio sample R. Assuming that YAMNet's top N predicted labels are $l_0, \ldots, l_{N-1}$ for a sample R, which have associated gestalt scores $S_0, \ldots, S_{N-1}$, we compute S_R with the general form:

$$S_R = \frac{\sum_{n \in N} J_\phi(p(l_n)) \cdot S_n}{\sum_{n \in N} J_\phi(p(l_n))}$$

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

where J represents a weighting function with the parameters ϕ, and $p(l_n)$ represents the YAMNet prediction probability associated with label l_n for sound sample R. In practice, parameters θ and ϕ form a linear function with a value of zero below a given cutoff.

As an illustration of the results of the bootstrapping process, Table 1 gives examples of some top scoring AudioSet labels for Arousal, Valence, Memorability, and Confusability.

Top Scores: "Memorability"	**Top Scores: "Confusability"**
Guitar	*Rain on surface*
Wail, moan	*Pink noise*
Fire alarm	*Ocean*
Baby cry, infant cry	*Traffic noise, roadway noise*
Singing	*Typewriter*
Chuckle, chortle	*Wind*
Top Scores: "Arousal"	**Top Scores: "Valence"**
Skidding	*Acoustic guitar*
Machine gun	*Wind chime*
Ambulance (siren)	*Chuckle, chortle*
Fire alarm	*Laughter*
Growling	*Classical Music*
Ringtone	*Church bell*

Table 1. Sound category labels from the AudioSet [15] ontology with top scores for Memorability, Confusability, Arousal, and Valence, determined by bootstrapping from the small HCU400 dataset [4, 40].

As we might expect, the sound labels with the highest scores for valence represent concepts that humans might consider to be calming or positive in sentiment; those with the highest scores for arousal are what one might consider frightening, alarming, or jarring; for confusability, we see labels that might be considered "background" or "ambiance"; and those with the highest scores for memorability are a combination of the previous dimensions, in addition to sounds that are organic and human-centered.

3.2 Causal Uncertainty

As suggested in Sections 2.2 and 2.3, causal uncertainty from a listener's point of view underlies much of our understanding of gestalt listening. Here, we approximate the causal uncertainty annotation in [4] by using neural network uncertainty as a proxy for human uncertainty. We use the probability assigned by YAMNet to the top performing AudioSet class and average it across the prediction frames in the duration of a sound clip. We find that, as shown in Figure 2, this method separates the most causally uncertain sounds ("Synthetic") from the least causally uncertain sounds ("Natural") from the HCU400 dataset, akin to the original annotated measure.

3.3 Acoustic Self-similarity

The literature in the cognitive sciences [5, 13, 14] emphasizes 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, both at a gestalt and acoustic level. Here, we assess acoustic-level repetition through a measure of self-similarity, first employed in the context of audio by [9]. To obtain a score revealing how similar an audio excerpt is relative to itself and all other sampled audio excerpts, we compute a magnitude Short-time Fourier Transform (STFT) for each excerpt with 512 FFT bins and a hop size of 512 samples. To decrease the computational overhead, each STFT is then smoothed along the time axis with a window width of 10 units, and down-sampled along the same axis by

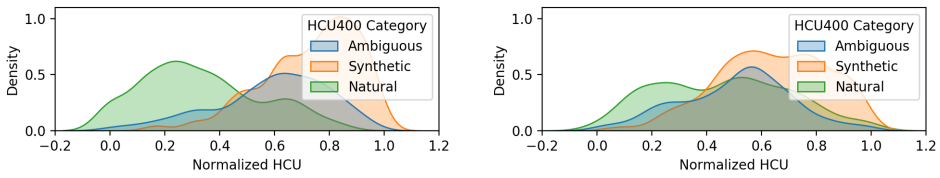

Fig. 2. Plots showing the distribution of sounds in the HCU400 dataset labelled with their original categories from [4], namely "Natural", "Ambiguous", or "Synthetic". We compare the separation of these "Natural" and "Synthetic" classes via the human annotated Hcu metric (left) with the proposed, neural network-based approach (right).

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. We then obtain a novelty curve by summing along one of the matrix axes; 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.

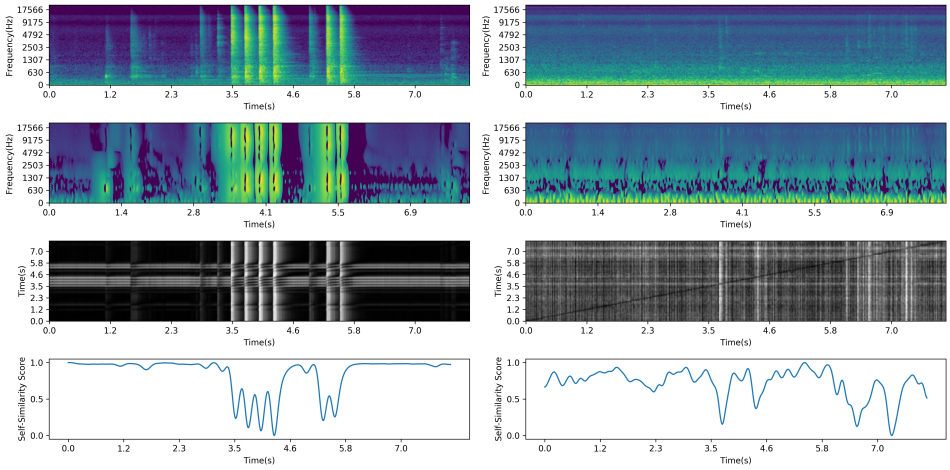

Fig. 3. We illustrate the process for computing Acoustic Self-similarity and Acoustic Saliency for two contrasting sound excerpts, given by their STFT representations (top). We show the saliency intensity map (second row), the calculated self-similarity matrix (third row), and the resulting self-similarity score (bottom).

3.4 Semantic Novelty

We extend the importance of the notion of repetition from [5, 13, 14] to gestalt processing as well, and posit that novelty on the semantic level also represents a valuable control in our feature space. We therefore compute a high-level self-similarity measure, calculated with the same algorithm as for low-level spectral self-similarity, except based on the centroid of each audio excerpt's YAMNet embeddings (where the centroid is the mean of the embeddings over all of the frames in an excerpt). Given the conceptual value of the embeddings as discussed in [23], we expect that audio excerpts whose embedding centroids stand out in this context represent contrasting semantic information relative to the other excerpts being compared.

3.5 Acoustic Saliency

Auditory saliency has been loosely defined as a measure of novelty, surprise, or attention [11, 28, 30]. While these definitions are arguably gestalt concepts, most widely used models of saliency operate on spectral input, often employing kernel convolution methods derived from vision research to identify "salient" audio regions as a function of time and frequency. For this reason, we treat auditory salience as a low-level perceptual feature, and hypothesize that measures of auditory salience can be used to estimate change or contrast in an audio stream stemming from acoustic analysis alone.

To obtain a measure of auditory salience for a frame of audio, we build upon the implementation in [40], inspired by the original model proposed by [30]. 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 and the total energy appearing in each map, and equally weight these statistics from each map. Figure 3 provides an example of resulting salience maps for two different audio excerpts.

3.6 Spectral Cues

We finally include a suite of spectral features representing low-level, acoustic properties of a sound object that determine our interaction with a sound in the presence of weak semantic information. Because they can be implemented computationally without any further approximation, we simply replicate the spectral cues (pitch diversity, harmonic-percussive ratio, centroid, bandwidth, etc) detailed in [40].

4 SUMMARIZATION TOOL

We present a description of our audio summarization tool that employs these feature classes to generate short audio presentations from extended recordings. An overview of the tool is as follows: we first select 3-second audio excerpts at equally-spaced time intervals throughout the body of audio, which is a concatenation of all recorded audio (over 1-3 weeks) with appropriate padding to ensure an excerpt does not span multiple recording files. We then extract the value of each of our features for each excerpt in the set. Our implementation outputs a ranking of all of the excerpts ordered by feature value assigned, and depending on the feature strategy preference specified, a subset of excerpts are selected for the final presentation. The selected excerpts, about 30 excerpts for a 1-minute summary, are finally cross-faded in chronological order and output as a single track. Figure 1(4) provides a detailed illustration of the data flow in the system.

The excerpts can be selected in one of two ways, which we call “feature strategies":

(1) Top or Bottom Feature: a single feature strategy is specified, and excerpts are simply drawn from the top or bottom of the ordered ranking (for example, taking the top 30 acoustically salient samples is referred to as “Most Acoustically Salient", or taking the bottom 30 semantically novel clips is referred to as “Least Semantically Novel").
(2) Baseline: a naive benchmark strategy for comparison within our listening experiments, where 30 excerpts are simply chosen at equally spaced intervals in time from the set of excerpts being analyzed, without any feature extraction and ranking.

5 PARTICIPANT EVALUATION

To understand the value of our system and the underlying feature space, we believe it is most appropriate to evaluate it in the context of the motivating applications – using lifelogged, first-person audio recordings – rather than generic databases of soundscape recordings which may be constrained in sound object content and demonstrate no personal relationship with a listener. For a more realistic study, we opt for an in-the-wild study conducted over the span of several months.

In our evaluation, participants were provided with wearable, high-fidelity stereo recorders (shown in Figure 1(1)) to capture their sonic environments for as many hours as possible during waking hours for 1 (minimum) to 3 (maximum) weeks. After signing up for the study and providing consent, participants were delivered their recorders in a contactless fashion, provided with detailed explanations on the use of the device and ethical best practices in a video call with the researchers, and given instructions on uploading their recorded audio to a secure server in our lab accessible only by the authors of this research. Significant precautions were taken to ensure the privacy of study participants and the individuals in their environments: based on guidance from a student law clinic associated with our institution, participants were allowed to pause and restart the recorder at any point during the day to avoid capturing sensitive content, and were required to obtain consent from individuals who could be identified in the recordings and wear the recording device in plain sight; the researchers also were not permitted to audition raw participant recordings at any point, and required explicit consent to audition generated summaries.

After the recordings were captured and uploaded by each individual, raw recordings were processed using the summarization tool, and were used to generate 16 summaries (see Figure 5) from the pool of possible feature strategies that we hypothesized would map to unique perceptual outcomes, in addition to the baseline strategy. The generated summaries were automatically embedded in an individualized survey, alongside several series of questions: for each summary, participants were asked to assign relevant perceptual descriptors from a pre-determined list ("calming", "nostalgic", "social", etc), and provide ratings on perceived emotional intensity, sense of intimacy, and positive or negative sentiment. Participants also provided lengthy qualitative descriptions of their listening experiences using the system, responding to guiding questions such as "What surprised you most or least about what you heard?" and "Did you find listening to the summaries to be an immersive experience?" The full set of questions in the survey is reproduced in Appendix A for reference.

We recruited N=10 participants (4 females, 6 males, aged between 25 and 65), via public advertisement at our institution and the surrounding community. The participants included undergraduate and graduate students, young professionals, and faculty members, who were associated with a spread of living spaces (dormitories, shared apartments, suburban independent housing with large families), working conditions (remote desk work, office space desk work, physical laboratory work), and experience with audio and music (from no inclination towards or experience with sound recording or production to semi-professional audio engineers and musicians). Participants collectively provided over 800 hours of audio, recording for 4-6 hours per day on average. We provide audio examples of summaries formed with different feature strategies at [*link redacted: please see supplemental files for sound examples*].

6 EXPLORING GESTALT-ACOUSTIC CLIP OVERLAP

We suggest that the value and novelty of this system is a feature space that incorporates gestalt auditory understanding; however, it is possible that low-level features correlate so strongly with high-level ones that gestalt analysis is redundant for summarization tasks. To assess this aspect of our summarization system, we analyze whether high-level features independently surface novel content as compared to low-level features. To do this, we compute the percentage of overlap between clips that rank in the top and bottom 1 percentile of the entire pool of excerpts from a single participant's data, per feature strategy. We then average these results across all 10 participants, shown in Figure 4.

From the heatmap, we see very little overlap across the two classes of features (high-level on the bottom right of the grid, and low-level on the top left)[2]. We do see evidence of intra-class overlap, such as between spectral bandwidth, centroid, and flatness, and between valence and memorability, which aligns with intuition. The results

[2]Random overlaps have a very small likelihood, as all participants have upwards of 10,000 excerpts in their audio pool.

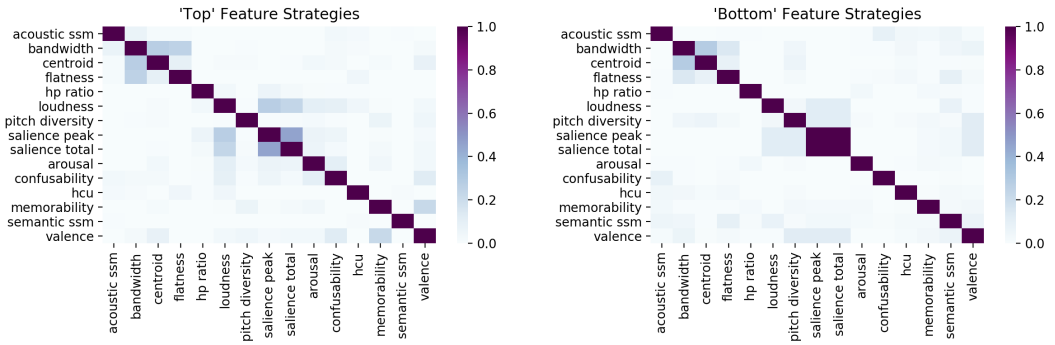

Fig. 4. Percentage of overlap between the "Top" (left) and "Bottom" (right) 1st percentile-ranked excerpts from each participant's recordings, averaged across all participants.

suggest that the introduction of gestalt principles to the task of audio summarization is a valuable contribution, and extends methods relying on spectral processing alone.

7 LINKING FEATURE STRATEGIES TO PERCEPTUAL OUTCOMES

We next examine how our chosen feature strategies map to perceptual descriptors across participants. In Figure 5, we show the relative contributions of the 16 feature strategies towards a perceptual goal, given by the frequency that a particular descriptor was selected for a particular feature strategy. The darkened bars give the most significant drivers of a descriptor (if one exists), computed using a modified (mean-absolute-deviation) z-score [22]. We see that there are 16 descriptors for which these drivers exist (for instance, "calming", "familiar", "distracting", "summary of events"), and 4 for which they do not ("busy", "surprising", "stressful", "eerie"). Of the former 16, we find that for 12 descriptors, gestalt feature strategies are the top performers. Finally, we find that apart from the descriptor "calming", the baseline strategy of selecting clips without any feature analysis is not a driver of any other descriptor.

The results also highlight trends that hint at the complexity of emotional response in sound perception that this work begins to uncover. For instance, contrary to intuition, sound clips with significant pitch diversity are found to be "calming" or "comforting"; percussive sounds ("Least Harmonic Percussive Ratio") are found to be relaxing; the most "nostalgic" summaries are comprised of sound objects with labels that are easily identifiable and intrinsically memorable, *as well as* those that are acoustically unique and diverse in tonal content. We intuit that a personal relationship with the audio being summarized (reflecting back on first-person recordings) is a significant functional force in these relationships. We examine this further in Section 10.

Noting the overlap in top performing feature strategies across some descriptors, we perform a clustering analysis to identify orthogonal archetypes in the perception-feature strategy space. To do this, we first compute an affinity matrix between descriptors using the jaccard index of intersecting feature strategies as the affinity measure (considering only descriptors and feature strategies with significance given by their z-scores), and perform a simple agglomerative clustering. The results are given in Table 2, with union of the feature strategies shown beside each descriptor cluster. These clusters suggest that we may be able to explicitly map the system's underlying feature space to different perceptual archetypes that hold across multiple individuals; more importantly, however, on the level of an individual listener, we suggest that these clusters are useful "initializations" for personalized summary generation, wherein the relative weights between feature strategies are refined based on user feedback (see Section 10).

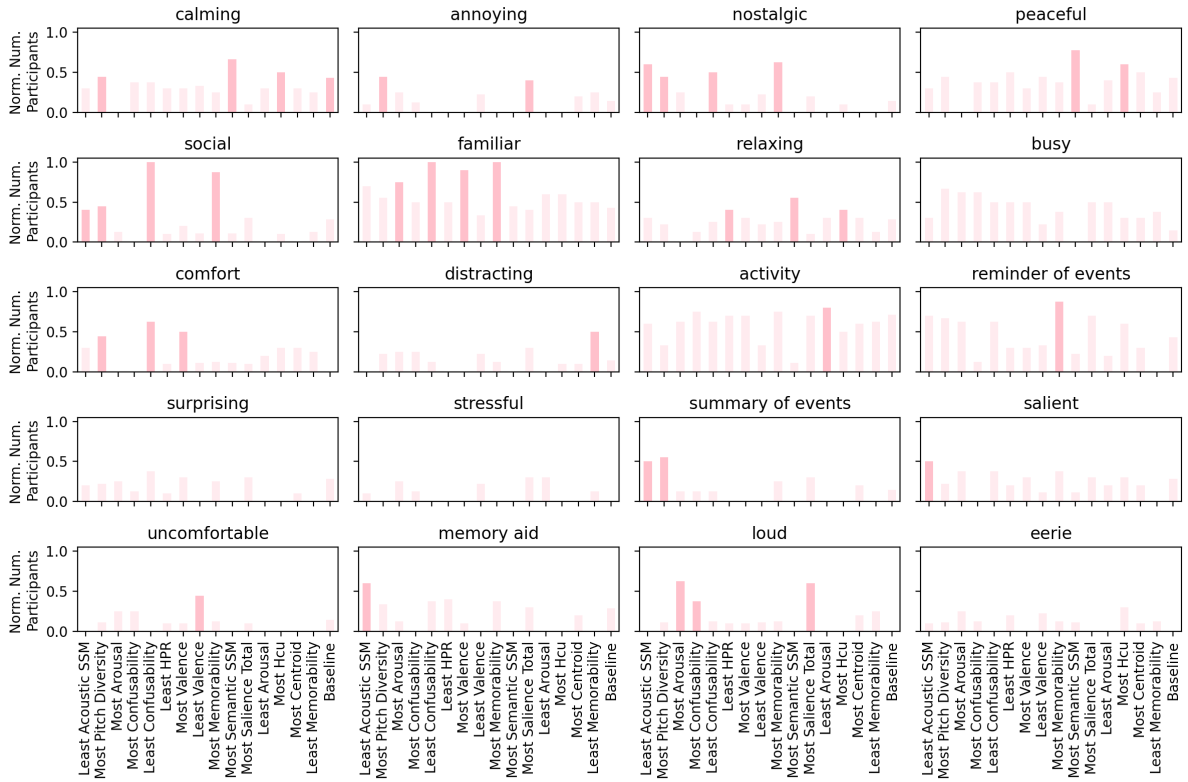

Fig. 5. The relative contribution of each individual feature strategy in driving the selection of a descriptor, given by the likelihood that the descriptor was marked by a participant in association with summaries made with a feature strategy. Darkened bars highlight positive outliers identified by 2 median absolute deviations above the median (modified z-scores).

	Descriptors	Feature Strategies
1	familiar, comfort	Least Confusability, Most Arousal, Most Memorability, Most Pitch Diversity, Most Valence
2	calming, peaceful, relaxing	Baseline, Least HPR, Most Hcu, Most Pitch Diversity, Most Semantic SSM
3	loud	Most Arousal, Most Confusability, Most Salience Total
4	nostalgic, social, summary of events	Least Acoustic SSM, Least Confusability, Most Memorability, Most Pitch Diversity
5	salient, memory aid	Least Acoustic SSM
6	uncomfortable	Least Valence
7	reminder of events	Most Memorability
8	distracting	Least Memorability
9	annoying	Most Pitch Diversity, Most Salience Total
10	activity	Least Arousal

Table 2. The results of an Agglomerative Clustering applied to the affinity matrix describing the intersection in dominant feature strategies between descriptors. Each cluster is described by the descriptor or new group of descriptors, and the union of the dominant feature strategies.

8 EXAMINING INTIMACY, SENTIMENT, EMOTIONALITY

We look next at the distribution of likert ratings provided by participants in response to each summary. Participants were asked to rate each summary on the scale of its (1) emotional intensity, (2) associated sentiment (positive/ negative), (3) intimacy and familiarity (see Appendix A for full text), and the results are shown in Figure 6. We use a non-parametric Kruskal-Wallis test with post-hoc Dunn test comparisons to examine differences between pairs of feature strategies, and observe that the most significant ($p < 0.05$) drivers of emotional intensity, positive sentiment, and intimacy are summaries comprised of clips that score highest in memorability and valence, and lowest in confusability. We note that all three feature strategies are gestalt, and are based on scoring models bootstrapped from human annotations in the HCU400 dataset. We do also find that certain low-level feature strategies are high-performing relative to others in each assessment category – for instance, least acoustic self-similarity for emotionality ($p < 0.05$) and intimacy ($p < 0.05$), and most salience for intimacy ($p < 0.05$). We suggest that this hints at the ability of these feature extractors to capture human-meaningful information at scale that points to subjective, aesthetic experiences, rendering them useful for our summarization task.

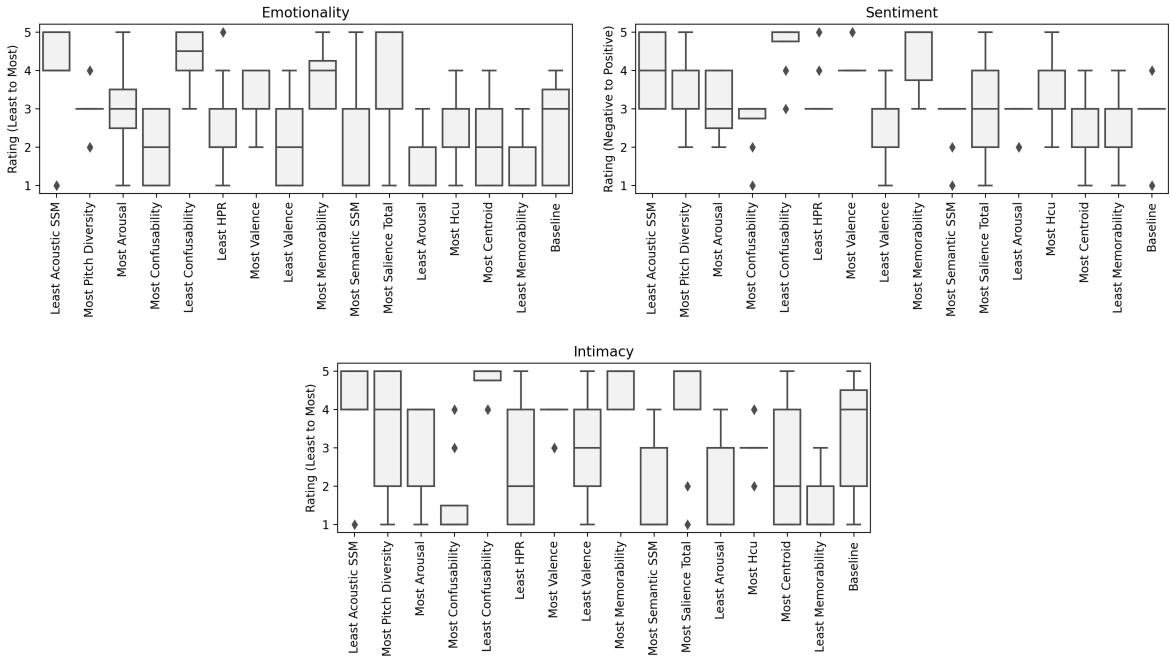

Fig. 6. The likert response values to Q1, Q2, and Q4 (see Appendix A) assigned by participants to summaries of each feature strategy.

9 ANECDOTAL RESPONSES TO LISTENING EXPERIENCES

Finally, in order to understand the affordances of our summary system that are more difficult to capture quantitatively, we examine the free text responses provided by participants at the end of their custom surveys. Participants were provided guiding questions (see Appendix A), but were free to provide any comments or reflections on the listening experience that came to mind. Several participants chose to provide further commentary by video calls

with the researchers regarding specific summaries in their surveys, which have been recorded and transcribed with the participant's consent.

Below, we highlight several themes from the responses and their supporting commentary. The commentary is marked with an anonymized participant ID, as well as the feature strategies associated with the summaries described by the participant, if applicable. As an overview, we uncover four major conclusions:

(1) our system succeeds in creating *impactful experiences* over *informative summaries*, aligned with the motivation presented at the start of the work
(2) participants are unanimous in their willingness to use the system, and suggest a diversity of application contexts and use paradigms
(3) the output of our system is found to be engaging, emotional, immersive, and intimate
(4) participants offer feedback on production aspects, reinforcing some design choices and suggesting improvements in others

We note that this anecdotal evidence alone is not sufficient to disentangle the value of the system design choices (such as the size of the audio excerpts, the cross-fading heuristic, the use of the stereo field, etc) from the algorithmic choices (the incorporation of gestalt information in the feature space). We treat this commentary as positive reinforcement for *both* aspects of this work, and we intuit the value of the latter aspect especially when the commentary is coupled with the trends observed in the quantitative analysis in Sections 7 and 8.

Immersion and Intimacy

Several participants explicitly described a sense of presence in time or place when reviewing their audio summaries. For instance:

> *"[Listening to the summaries] was really engaging, I think the quality and the spatial nature of it make it feel really impactful. I actually think because the sounds seem to move around spatially a bit between some of the clips, it feels more like a fly on the wall observing from different perspectives, and that really makes it feel like you're peeking into something... there is definitely a feeling of being there in the space in that time."* - P1

We noted that participants experienced this sense of immersion from summaries curated to have human-oriented content (via gestalt feature strategies) as well as those that combined notable but inorganic content from a sound context:

> *"Ones that brought me back to a moment in time or had talking made me feel transported. The kitchen ones also felt like I was making food again."* - P4, **[Least Acoustic SSM, Most Salient, Least HPR]**

> *"One of my tracks seemed to highlight laughter in particular, and it was pretty neat. It's hard to capture memories of how or what makes people––or even if they laughed. Also hearing someone's laugh is like hearing them in a very innocent state that I found interesting."* - P5, **[Most Memorable, Most Valence]**

> *"..just hearing someone's voice made me reflexively smile, which was surprising. It was actually even quite immersive to hear strictly ambient noises that had been recorded, like keyboard typing and doors closing. "* - P5, **[Most Memorable, Least Confusable, Least Arousal, Most Salient]**

The participants also suggested that the compelling nature of the listening experience stemmed from a sense that the content was very personal; while this is due in part to the nature of the recording exercise (a wearable recorder that was capturing audio indiscriminately), participants highlight the role of the "sensitive editing", or the automated sound object analysis, extraction, and curation process:

> *"The more unusual immersion was in the intimate mixes, again stemming from the sensitive editing*[3] *and the pervasive recording that managed to capture scenes that are usually off-limits... I am not sure whether those examples would also feel immersive for other people, or if the sense of immersion comes from recognizing the strange little details of one's own life..."* - P9

Triggering Memories without Explicit Documentation

A common theme that emerged from the participant responses was the unprompted recollection of events and occurrences tied to the summary content:

> *"..this was an almost bizarre (in a good way) experience of feeling the connection between audio and my memory. Especially with people's voices and laughs. If you asked me what I did in the past few weeks before hearing this, I'm not sure I would have written down what I am now enjoying my memory of."* - P5, **[Most Memorability, Least Confusability]**
>
> *"The sound was often cut too short for me to recognize the situation, but at the same time I was surprised how much I remember from the snippet of sound of what happened in that situation... I find the auditory experience to be very stimulating and the fact that there are no images triggers memories in a much more emotional way for me, because it is vague and ambiguous."* - P7
>
> *"In the best examples, I was surprised by what felt like creative editing and curation: with almost comedic timing, tying domestic activities together with a well-cut sneeze or hoot. In those examples, I also found what felt like the closest representation of my sense of home life: intimate, detailed snippets that were just long enough to remind me of a bigger picture."* - P9

As suggested by the quotations, most participants were aware that they could not precisely pinpoint the events that were taking place or the actions that they were engaged in when the audio in the summaries was captured. Most, however, went on to suggest that the curation process was effective in transforming this notion into an emotional listening experience, eliciting feelings of nostalgia and positive sentiment associated with past memories.

New Perspectives on Everyday Life

Depending on the feature strategy used to curate the summary, participants highlighted the ability of the presentation to offer different perspectives on the same sound context, often allowing participants to reexamine sound elements they had not noticed when recording:

> *"I was surprised by the level of environmental noise in [the] recordings, combined together like in the recording they really make me feel stressed. It makes clear how my brain filters information about my environment and my experiences. When I hear the situations without the visuals it highlights different aspects that was not noticeable to me before."* - P7, **[Least Arousal, Most Semantic SSM, Least Memorable, Most Hcu]**
>
> *"It's just fascinating too to see what clips appear – some of them are really clearly tied to specific events (like a video I watched or a specific project I was working on), some evoke a common behavior I do (going for walks), some I couldn't even identify because they're just background sounds I filter out (or have headphones in)."* - P1, **[Least HPR, Most Semantic SSM, Most Salience, Least Arousal]**
>
> *"[I was surprised by..] the amount of laughter I encountered over the course of the study; how eerie my workday 'silence' sounded like; how my laugh is more high-pitched than I remember."* - P3, **[Most Memorable, Least Arousal, Most Pitch Diversity]**

[3]Participants were not informed of how the summaries were prepared.

Using my Summaries

Participants were nearly unanimous in their interest in using the summarization system again in the future. For instance, respondents said:

> *"[I] absolutely [would use these audio presentations]. I'm actually quite surprised at how interesting and fun and engaging listening to these clips was, and I think it captured really interesting tidbits of life in a really unique way."* - P1

> *"I'd probably use it every few days. The information content was low though.. But I loved hearing a quasi-musical or sound-art gist of my days.. I would use this more as an experience rather than anything to derive information from - but it was indeed a fun experience, and if I listen to this a year from now, the nostalgia quotient will probably be quite a bit higher!"* - P10

Users also detailed specific use cases and application scenarios to which they would want to see the system applied, with a focus on reflection, wellbeing, and mindfulness practice:

> *"Especially would use it to capture a certain extended experience, like living abroad, an internship, or a summer with family."* - P3

> *"I would use the audio presentations like a diary to reflect back on things further in the past than a couple weeks, and I would use it as a daily diary rather than weekly or longer. It would be interesting to use audio presentations for funeral or wedding services, very intimate."* - P4

> *"I'd be very interested to use this tool as 1) a kind of gratitude journal that helps me keep my family and friends in mind, I am also interested in daily or weekly intervals. I do a daily mindfulness meditation practice... I see parallels and think that this could be a very valuable companion exercise to that practice."* - P5

Finally, several users alluded to the fact that given the intimate and personal nature of the summaries, the mechanics of the summary production that enable privacy – selecting short, individually unrecognizable sound snippets that are combined by feature strategy – are of paramount importance to the adaptation of such a system.

> *"If I know that the recordings never leave my recorder/phone and are processed into unrecognizable bits (like the ones I head [sic] in the study) in real time, would make me much more likely use this technology."* - P7

> *"So perhaps intentional, intermittent use of a system like this would be good: a couple of weeks a year just to have a taste of what each year felt like, not from the perspective of the public, curated social feed but from the inside.. a system for documenting my life that is only for ears and no one else's."* - P9

Positives and Limitations of Summary Production

Though not explicitly prompted to do so, users provided feedback regarding the production methodology and audio rendering infrastructure that is used to support the algorithmic choices in the system. Several praised aspects of the production pipeline:

> *"I really liked how the audio was spatialized. That added information about the events that made the summaries tailored to how I experienced them. For example, audio summaries of the workday (a lot of typing) had typing sounds in different parts of space, because of my changing positions relative to my keyboard, which made the end result more artistic"* - P3

> *"[I was surprised by] how well some of the sounds were blended together or laid over each other that sounded natural but I know they didn't occur simultaneously like that."* - P4

> *"The way the sound clips flowed together was surprisingly good - almost like a piece of composed sound-art music.. The stereo field was used well, and some of the clips really did take me on a mini-voyage through the time I had the gear."* - P10

Others expressed preferences for production parameters that would have improved their sense of engagement:

> *"If the clips were longer I think I would have found it a bit more immersive... Some way to unintentionally record would be better for me because the results are surprising and fun."* - P7

In the future, the feedback associated with this theme can be used to explore other options for assembling summaries based on cognitively analyzed and curated sound content. While we choose a simple approach in this work and keep it consistent across users, more complex production techniques (for instance, concatenative synthesis [43]) could enhance the listening experience independently of feature strategy.

10 DISCUSSION

Thus far, we have described a new auditory interface – a tool that can be used to mediate our interaction with the sounds that we capture around us – built with a feature space that augments traditional spectral measures with "gestalt" measures inspired by ideas in auditory cognition. In quantitative tests of user perceptions, we show that these features dominate over or combine with spectral measures to elicit specific aesthetic responses. And in an exploration of qualitative feedback, we demonstrate that these features result in a system that create compelling, moving listening experiences.

Here we address the value and limitations of this work, and suggest several important contributions and open research problems that it offers to the HCI community.

Cognitive Features as Inputs to Auditory User Interfaces Cognitively-inspired "gestalt" feature extractors have the potential to be powerful, as observed in the presented quantitative and qualitative results. In the most reliable mappings between feature strategy and perceptual goal from our study, gestalt feature extractors play the dominant role; our analysis also shows that they surface a statistically different set of audio samples than the acoustic feature extractors alone, implying that the relationship between acoustic and gestalt information is non-trivial and demands greater complexity in modeling. It is important to note that these gestalt extractors cannot themselves be considered predictive models of perceptual response – a sound with a high "Memorability" score does not automatically mean it will be readily remembered by a certain listener! – but instead that they capture collective conceptual information which can be exploited to construct such relationships in the context of an application, as was done in this work. We see fascinating possibilities emerging from this untapped field of interface design that attempts to incorporate cognitive models, such as AR devices that manipulate the sonic environment in real-time to modulate attention or memory, VR environments that process and resynthesize perceptually realistic versions of real-world soundscapes, and sensory prostheses that extend or limit our natural faculties to facilitate new experiences.

Towards Personalized Audio Summaries Despite the broad brush-stroke conclusions that can be drawn from the aggregate participant response, there is still significant variance in the reported summary perception that is a function of an individual's sonic diversity and relationship to his or her sonic environment under personal context. We suggest using the clusters mapped out in Section 7 as "priors" – or *a priori* information that forms a coarse model – between feature strategies and perceptual archetypes, and then considering a closed-loop system that incorporates user feedback to refine feature weights towards specific preferences over extended periods of time. We intend to present approaches to personalizing our summarization system in forthcoming work.

Limitations: Audio Classification Networks While neural networks can bootstrap the translation of gestalt listening principles to system design, there are limitations with current state-of-the-art models. Supervised learning strategies designed to map independent sound events to labels – as the AudioSet model does – do not generalize well to dynamic, real-world sonic contexts, where sound events frequently overlap and vary in signal-to-noise ratio. Classification taxonomies are also often limited. Despite AudioSet's notably large ontology, information that can be extracted from the labels is drastically limited when compared to free-text human

annotation. As the research advances in favor of more naturalistic datasets and unsupervised learning strategies, we can advance the capabilities of this summarization system and other similar interfaces.

Limitations: Discrete Models In Section 3.1, properties of sound labels are estimated as discrete scores from a smaller dataset. More flexbility in the face of imperfect DNN estimators and varying human experiences, however, is afforded by variational models that represent label properties as probabilistic distributions. A detailed technical presentation and evaluation of such an approach is discussed in forthcoming work.

Limitations: User Study The user study methodology chosen for this work allowed for a largely realistic evaluation of the system, but presented certain practical challenges for the participants: discomfort in wearing the recording devices for several hours a day, limited battery life and storage space, privacy concerns, and constant awareness of the device's presence all limited the extent to which a natural, unadulterated sonic environment could be captured. In forthcoming studies, we will consider ways to improve physical properties of the device (i.e., choosing a device with a longer battery life and memory), and attempt to better understand the role of the behavioral factors in participant data.

11 CONCLUSION

In this work, we demonstrate an audio summarization system achieved by cognitively-inspired content curation and remixing. To this end, we develop a suite of computational features that reflect the principles of everyday listening, employing deep learning techniques to estimate gestalt information from unseen audio data for the first time. We use these features to build short summaries from personal lifelogged recordings and show that gestalt feature extractors dominate over or combine with spectral measures to reliably elicit specific perceptual responses, and are the most significant drivers of emotional intensity, positive sentiment, and intimacy in generated audio summaries, suggesting that they are a valuable design contribution. We show through anecdotal data that the summarization interface built from this novel feature space results in emotionally impactful and engaging listening experiences.

More broadly, this work presents an important contribution to the research problem of summarizing large-scale, experience-oriented audio (such as audio from personal recordings or ecological monitoring) with aesthetic aims. We expect that exploring content along subjective dimensions will emerge as an important area of research for the future of recording summarization techniques.

A SURVEY QUESTIONS

Summary Questions

Listen to the following audio presentation generated from the audio you recorded:

Q1: How would you say listening to this presentation made you feel? Select a point along the scale from least to most emotionally evocative. *[Select from a scale of 1 (least) to 5 (most)]*

Q2: How would you describe the sentiment (if any) associated with this presentation? Select a point along the scale of "negative" to "positive". *[Select from a scale of 1 (negative) to 3 (neutral) to 5 (positive)]*

Q3: Which of the following terms best describe the presentation? Select as many as apply. *[Select from – calming, annoying, nostalgic, peaceful, social, familiar, relaxing, busy, comfort, distracting, activity, reminder of events, surprising, stressful, summary of events, salient, uncomfortable, memory aid, loud, eerie]*

Q4: How would you describe how intimate this presentation felt? Select a point along the scale from least to most intimate. *[Select from a scale of 1 (This presentation felt generic, reflecting sounds that could have been recorded by others.) to 5 (This presentation is uniquely mine, reflecting the spaces and events I recorded.)]*

Reflection Guiding Questions

G1: What surprised you most about what you heard? What didn't surprise you?

G2: Do you find listening to the audio to be an immersive experience? Why or why not?
G3: Would you use such audio presentations to review or reflect on your day, week, month, or year? Why or why not?

REFERENCES

[1] Sainath Adapa. 2019. Urban Sound Tagging using Convolutional Neural Networks. *arXiv preprint arXiv:1909.12699* (2019).
[2] Salvatore M Aglioti and Mariella Pazzaglia. 2010. Representing actions through their sound. *Experimental brain research* 206, 2 (2010), 141–151.
[3] Jitendra Ajmera, Om D Deshmukh, Anupam Jain, Amit Anil Nanavati, Nitendra Rajput, and Saurabh Srivastava. 2012. Audio cloud: creation and rendering. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. 277–280.
[4] 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.
[5] 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.
[6] Oliver Bones, Trevor J Cox, and William J Davies. 2018. Distinct categorization strategies for different types of environmental sounds. Euronoise.
[7] Wei Chai and Barry Vercoe. 2003. Music thumbnailing via structural analysis. In *Proceedings of the eleventh ACM international conference on Multimedia*. 223–226.
[8] Matthew Cooper and Jonathan Foote. 2002. Automatic Music Summarization via Similarity Analysis.. In *ISMIR*.
[9] Matthew L Cooper and Jonathan Foote. 2002. Automatic Music Summarization via Similarity Analysis.. In *ISMIR*.
[10] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. 2016. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems* 47, 1 (2016), 65–76.
[11] Varinthira Duangudom Delmotte. 2012. *Computational auditory saliency*. Ph.D. Dissertation. Georgia Institute of Technology.
[12] Duo Ding, Florian Metze, Shourabh Rawat, Peter Franz Schulam, Susanne Burger, Ehsan Younessian, Lei Bao, Michael G Christel, and Alexander Hauptmann. 2012. Beyond audio and video retrieval: towards multimedia summarization. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. 1–8.
[13] William W Gaver. 1993. How do we Hear in the World? Explorations in Ecological Acoustics. *Ecological Psychology* 5, 4 (1993), 285–313.
[14] 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.
[15] 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.
[16] 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.
[17] Carlos-Emiliano González-Gallardo, Romain Deveaud, Eric SanJuan, and Juan-Manuel Torres. 2020. Audio Summarization with Audio Features and Probability Distribution Divergence. *arXiv preprint arXiv:2001.07098* (2020).
[18] Peter Grosche, Meinard Müller, and Joan Serrà. 2013. Towards cover group thumbnailing. In *Proceedings of the 21st ACM international conference on Multimedia*. 613–616.
[19] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2020. ESResNet: Environmental Sound Classification Based on Visual Domain Models. *arXiv preprint arXiv:2004.07301* (2020).
[20] Brian Gygi and Valeriy Shafiro. 2007. General functions and specific applications of environmental sound research. *Frontiers in Bioscience* 12 (2007), 3152–3166.
[21] 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)*. 489–498.
[22] Nathanael A Heckert, James J Filliben, C M Croarkin, B Hembree, William F Guthrie, P Tobias, and J Prinz. 2002. Handbook 151: NIST/SEMATECH e-Handbook of Statistical Methods. (2002).
[23] 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.
[24] 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.

[25] Turab Iqbal, Yin Cao, Qiuqiang Kong, Mark D Plumbley, and Wenwu Wang. 2020. Learning With Out-of-Distribution Data for Audio Classification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 636–640.

[26] Wei Jiang, Courtenay Cotton, and Alexander C Loui. 2011. Automatic consumer video summarization by audio and visual analysis. In *2011 IEEE international conference on multimedia and expo*. IEEE, 1–6.

[27] Yukang Jin, Tong Lu, and Feng Su. 2012. Movie keyframe retrieval based on cross-media correlation detection and context model. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 816–825.

[28] Ozlem Kalinli, Shiva Sundaram, and Shrikanth Narayanan. 2009. Saliency-driven unstructured acoustic scene classification using latent perceptual indexing. In *2009 IEEE International Workshop on Multimedia Signal Processing*. IEEE, 1–6.

[29] Chen-Tai Kao, Yen-Ting Liu, and Alexander Hsu. 2014. Speeda: adaptive speed-up for lecture videos. In *Proceedings of the adjunct publication of the 27th annual ACM symposium on User interface software and technology*. 97–98.

[30] 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.

[31] Seita Kayukawa, Keita Higuchi, Ryo Yonetani, Masanori Nakamura, Yoichi Sato, and Shigeo Morishima. 2018. Dynamic Object Scanning: Object-Based Elastic Timeline for Quickly Browsing First-Person Videos. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[32] Hyun Hee Kim and Yong Ho Kim. 2010. Toward a conceptual framework of key-frame extraction and storyboard display for video summarization. *Journal of the American Society for Information Science and Technology* 61, 5 (2010), 927–939.

[33] Petros Koutras, Athanasia Zlatintsi, Elias Iosif, Athanasios Katsamanis, Petros Maragos, and Alexandros Potamianos. 2015. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. In *2015 IEEE international conference on image processing (ICIP)*. IEEE, 4361–4365.

[34] Huy Viet Le, Sarah Clinch, Corina Sas, Tilman Dingler, Niels Henze, and Nigel Davies. 2016. Impact of video summary viewing on episodic memory recall: Design guidelines for video summarizations. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4793–4805.

[35] 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.

[36] 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.

[37] 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.

[38] Jouni Paulus and Anssi Klapuri. 2006. Music structure analysis by finding repeated parts. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. 59–68.

[39] L Pizzamiglio, T Aprile, G Spitoni, S Pitzalis, E Bates, S D'amico, and F Di Russo. 2005. Separate neural systems for processing action-or non-action-related sounds. *Neuroimage* 24, 3 (2005), 852–861.

[40] 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.

[41] Justin Salamon and Juan Pablo Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 24, 3 (2017), 279–283.

[42] 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.

[43] Diemo Schwarz. 2007. Corpus-based concatenative synthesis. *IEEE signal processing magazine* 24, 2 (2007), 92–104.

[44] Ankit Shah, Anurag Kumar, Alexander G Hauptmann, and Bhiksha Raj. 2018. A closer look at weak label learning for audio events. *arXiv preprint arXiv:1804.09288* (2018).

[45] 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.

[46] Roger CF Tucker, Marianne Hickey, and Nick Haddock. 2003. Speech-as-data technologies for personal information devices. *Personal and Ubiquitous Computing* 7, 1 (2003), 22–29.

[47] Simon Tucker and Steve Whittaker. 2008. Temporal compression of speech: An evaluation. *IEEE transactions on audio, speech, and language processing* 16, 4 (2008), 790–796.

[48] Sunil Vemuri, Philip DeCamp, Walter Bender, and Chris Schmandt. 2004. Improving speech playback using time-compression and speech recognition. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 295–302.

[49] 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.

[50] 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.