Sound Stories: A context-based study of listening to augmented soundscapes

BLANK FOR BLIND REVIEW

Research Highlights

- Overview of research on listening in complex everyday situations
- Background on listening to auditory displays in ambient intelligent spaces
- Large participant study using *Think Aloud* protocol
- Innovative semantic categories of auditory perception
- Guidelines for design of audio-augmented HCI in everyday spaces

Abstract

With an increasing number of everyday operations and communications becoming both automated and autonomous, ambient intelligent soundscapes are transforming to accommodate additional sonic feedback, and with it, new frameworks of listening. While this type of research and design of audio-augmented technology isn’t new, the impact pre-existing acoustic environments upon listeners’ sense-making activities is rarely considered holistically. Much of the study into the design of effective auditory displays focuses on perceptual acuity and correct source identification, often at the expense of understanding the context of meaning-making. This paper presents a study involving 70 participants who listened to unidentified audio recordings of two archetypal everyday urban sound environments naturally containing artificial signals as well as typical sounds. Using a *ThinkAloud* protocol we investigated listeners’ approaches to meaning-making in both semantic and temporal dimensions. Through a semantic content analysis, we articulate five aspects of sonic meaning-making: *spatial, descriptive, experiential, associational and narrative*. We further analyse the use of these perceptual elements on a temporal plane, in order to investigate how listeners construct a narrative of what they hear in *real-time*, naturally evolving as each subsequent sound event is interpreted. Results suggest that while listeners attend to sound events and spatial characteristics of a sound environment at the beginning of a new listening situation, as the soundscape unfolds they utilize associations and familiarity in order to place individual sounds into increasingly coherent narratives. Finally, we suggest that this approach could provide sound designers and human computer interaction specialists with a model for investigating the context aspects of a soundscape more holistically, allowing them to evaluate the effect of any new designed sounds prior to introduction into real-world environments.

Introduction

The role of sound in designing effective interactive technologies and interactive experiences has by now made it to the mainstream of human computer interaction (HCI), evidenced by numerous case studies and prototypes (Paay &Kjeldskov 2008; Bakker, van der Hoven & Eggen 2012; Isherwood & McKeowan 2017). There is a wealth of approaches to the design of sonic signals intersecting HCI and the field of auditory display design[1], however there is still a distinct gap between perceptual research in listening and usability studies of audio-augmented technologies. While Gestalt theory is well accepted in HCI, it is as yet not sufficiently explored when applied to sound design. Specifically, there is still sparse primary ethnographic work into everyday listening for technological or augmented everyday contexts, such as virtual reality (VR) settings, augmented mobile reality applications (AR), and Internet of Things (IoT) technologies. For the purposes of this paper we define augmented everyday contexts as *spaces that are both heavily used (thus associated with typical user expectations and behaviors), and at the same time increasingly contain smart and ambient technologies that rely on sonic signals competing with natural pre-existing soundscapes.* Examples include modern shopping centers, airports, hospitals, “smart” spaces, “smart” homes, mixed-space business buildings; as well as the introduction of augmented reality wearables (e.g. Google Glass, Bose AR[2] or Magic Leap One[3]) into otherwise analogue everyday settings. Context-aware design serves to support both usability and user experience considerations as e.g. in the case of smart homes or shopping centers, but also safety and accessibility considerations in spaces of higher risk profile such as e.g. airports or hospitals.

Since designed sounds are rarely presented in isolation and instead overlap with other complex sensory stimuli, it can be difficult to predict what different listeners are attending to at any point in a given everyday situation. This makes *context* a crucial area of listener-based studies in human-computer interaction, whether to assist auditory interface design, ecological evaluation of soundscapes, or practical usability of complex (smart) environments. This paper details a study which explores the way listeners make sense of what they hear, specifically focusing on how this process is fluid and based on an internal narrative constructed in real time with the aide of auditory contextual cues. The purpose of our discussion is to give interaction and interface designers a novel take on listener experience in augmented environments. Studies in the evaluation of everyday soundscapes use approaches ranging from card sorting exercises with pre-prescribed sound categories, to retrospective accounts of sound experiences; day-long journaling, with sound being presented to participants as audio recordings or in a real-life context (Kogan *et al.,* 2016; Craig *et al.,* 2017). Kogan et al.’s (2016) study was set up to actually poll participants about their listening experiences in the moment through a multidimensional questionnaire organized around several components including familiarity, mental and physical state, listening to sound sources, assessment, extra-auditory perception, as well as listener

[1] As evidenced in decades of publications, projects, models and software by the International Community for Auditory Displays (www.icad.org)

[2] https://developer.bose.com/bose-ar

[3] https://www.magicleap.com/magic-leap-one

expectations and coherence. While the conclusions are complex, the method ambitiously combines considerations of perceptual and extra-perceptual factors. Craig *et al.*'s (2017) study attempted to capture the "level of activity" or "listening state" of people as they go about their day towards understanding listeners' fluctuating attitudes towards their soundscapes. Such studies tend to get at many of the socio-cultural levels of listening and associations that listeners make between sound and place, as well as connotations attached to sounds themselves, and the activities that produce them. Our study, conversely, aims to investigate how listeners re-compose in their mind the practical, informational and aesthetic components of a given sound environment, particularly in cases where spaces are augmented to contain important *designed* sonic cues. In this sense our study falls within the rich tradition of HCI design ethnographies (Jordan & Henderson 1995) from which we can draw guidelines for user-oriented interaction. First we present some core literature in listening perception studies as well as common frameworks in soundscape evaluation design; next we describe the study of everyday listening to two common signal-rich environments; lastly we offer a discussion of our results specifically highlighting the temporal rendering of complex listening experience; and from there we articulate several HCI guidelines relevant to the design of audio-augmented technologies, and in particular, audio-augmented ambient intelligent systems deployed in common everyday public spaces.

Common Factors in Ecological Listening Studies

Some of the common factors that characterize complex auditory environments, and impact listeners involve sound masking (occlusion, including via personal audio), deliberate sound design, space augmentation, architectural acoustics, and many other individual factors (Park 2017). Studying how listeners make sense of typical urban auditory environments filled with acoustic, electronic and technological sounds can provide insight into the role of contextual association in everyday listening and help generate guidelines for better design (Wright, McCarthy & Meekison 2003; Hespanhol & Tomitsch 2015). In terms of conventions that aid listeners in responding to sound, two of the most important are *volume* and *pitch*, where a high sound pressure level indicates dominance, as do temporally impulsive sounds (predominantly in the mid to upper frequency range); quiet, continuous low-pitched sounds can normally be ignored (McAdams 1993). Unfortunately, when in a shared environment there are often a large number of devices all generating sound, on top of the pre-existing auditory environment. It is common to have multiples of the same devices producing sound in close proximity: self-operating tills or automatic teller machines, music, as well as personal mobile devices. In this proliferation of signals it is increasingly important to understand contextual listening and meaning-making in order to both design and evaluate sonic effects, auditory displays and the acoustics of built spaces in dense urban centres.

As previous work suggests, when trying to make sense of what is being heard in a mixed environment, people often attempt to group sounds together (Maffiolo et al. 1999, Cain 2008, Isherwood &McKeown 2017). If sound events are spatially, dynamically and temporally similar (Bizley & Cohen, 2013; Bregman, 1993) then there is a strong likelihood that they belong to the same source. Another gestalt listening approach is to identify whether the sound might be associated with a given environment contextually (Woodcock, Davies & Cox 2017; Roddy & Bridges 2018): e.g. a high-pitched beep in a supermarket denoting an item being scanned, or the beeps and alert signals on a transit vehicle indicating stops, doors opening and other relevant information. It is when a sound is not recognized in a typical environment that a meaning must be intuited (before it can be safely ignored). Context here is the main factor towards generating this meaning, and it is this mechanism that we explore through a *semiotic-temporal analysis*.

Methods in Soundscape Evaluation

Everyday listening to urban soundscapes has been the topic of a number of studies focused on the assessment and evaluation of auditory environments (Gaver 1993, Dubois *et al.* 2006, Woodcock *et al.* 2017) including in the field of HCI (Park 2017, Roddy & Bridges 2018). Needless to say, a focus on listeners' ability to evaluate and react to common urban soundscapes is an invaluable way to inform the design of audio-augmented ambient technology and ubiquitous computing. Cain *et al.* (2008) propose that contextual listening starts from an activity-centric standpoint. Craig, Moore and Knox (2016) also recognize the effects of the temporal dimension of sound. Throughout the day a listener's stance, interpretation and reaction to sound may shift many times. Maffiolo *et al.'s* (1999) study that asks participants to 'sort' soundscapes finds two patterns of categorization relevant here: 'event sequences' referring to soundscapes that contain mutually dependent and identifiable elements that listeners recognize, and 'amorphous sequences' where sounds are less easily distinguishable and possibly unrelated. Giordano *et al.* (2010) further find that evaluation of sounds from living sources is based on "sound-independent semantic information" whereas sounds from non-living sources is connected to their physical properties. Everyday listening is of course further compounded by sound's implications for action. In a comprehensive review of listening frameworks and taxonomies Tuuri and Eerola (2012) note that perception and action are "intertwined together since 1) the meanings of an environment are structured through embodied subject – environment interactions, and 2) meaning-structures such as action-sound couplings are organized in terms of directly perceivable action-relevant values, i.e. *affordances*" (p.138). In considering not only perception-action but also subjective, intentional and emotional context to auditory perception, Tuuri and Eerola underscore the importance of the gestalt experiential domain, which they suggest works as an "embodied resonator" (p.145) between sensation and relevant action. With all this in mind, when it comes to audio embedded in everyday augmented environments not all sounds necessarily require action and it is important that they are designed so; thus we can distinguish between the design of sounds meant to blend into the gestalt of listening and sounds that are meant to invite active action-perception; notwithstanding individual listeners' affective states. So while the above studies tell us a lot about how people tend to approach certain sound types and sonic situations, they don't necessarily tell us how that happens in real time, outside a focus on action, and specifically how in-the-moment classifications shift, morph and evolve in the early stages of listening to an unfolding auditory environment of acoustic, electroacoustic and digital sounds. This is our aim in this study. We explore how listeners strategically "place" sounds in archetypal urban contexts through focusing and re-focusing their attention in real time and putting sonic elements they hear into a meaningful narrative.

Method and Study Design

In order to address the temporal aspects of everyday listening, outside a focus on action, we formulated the following research question and exploratory hypothesis:

RQ: How do individuals parse out and mentally reconstruct soundscapes in augmented everyday contexts?

EH: We are interested in finding out whether interpretation of auditory environments as a temporal process first involves sound and spatial source identifications, which are subsequently influenced by associational complex referencing in order to create cohesive, contextual micro-narratives for the perceiver.

Study Materials

Following a pilot project with ten participants this study was intended to increase the sample size to a level that would provide us with both statistical significance and more qualitative data to help establish temporal patterns of everyday listening. We invited 70 participants to listen to two different sound environments over headphones, and describe what they were hearing in real-time using a *ThinkAloud* protocol. *ThinkAloud* is a phenomenological research and data-gathering protocol often used in HCI and usability evaluation, as well as in psychology and a range of social sciences (Lewis 1982). It involves asking users to narrate their thoughts, impressions or real-time descriptions of a process while participating in that process. While not without its drawbacks, this protocol was selected for its ability to provide us with a more granular understanding of listening as it unfolds in time in a textual form that can then undergo content classification. The study was conducted in a laboratory environment rather than in a real world setting in order to ensure repeatability and constrain to a degree the rather complex nature of listening-in-context. It was considered relevant to deprive participants of any visual cues, which might affect their responses when trying to make aural sense of and construct a narrative about what they were attending to. Participants were recruited for the study through face-to-face requests at a university campus; 80% were male, 20% female, with 94% being students, and 6% staff; a sample of convenience. Eighty six percent were in the age range of 18-24, with the remaining 14% being 25 or older. All participants believed themselves to have normal hearing for their age, and we did not anticipate (nor does research suggest) that any prior musical experience would influence perception of everyday-type soundscapes. We presented each participant with two stereo recordings captured by the first author, in a consistent order. One of the recordings was captured just inside the entrance to a retail bank, adjacent to a row of three ATMs (Automated Teller Machine), all of which were being utilised by customers. The audio file was 2 minutes and 12 seconds in length, and recorded at 44.1 kHz, 16 bit (uncompressed Red book CD audio quality). The second recording was captured at a grocery store, specifically a check-out queue, where customers were paying for their purchases. The audio recording was 2 minutes and 13 seconds long, and again recorded using Red Book audio quality. Stereo recordings were made rather than binaural audio due to issues with the wide variation in head shapes/sizes and how this can affect sounds presented in front of a listener's head, as well as sound sources and localisation. Stereo recordings translate more accurately for a wider range of listeners (Pike 2013); however, future work would benefit from the utilization of binaural audio. Participants listened to both recordings through Beyerdynamic DT 770 Pro 250 ohm closed reference headphones (Beyerdynamic, 2017) with an average SPL of 75 dBA. The use of headphones rather than stereo display was necessary here in order to capture the spoken *ThinkAloud* audio separately from the audio environments presented to participants (for transcription purposes). While having only two audio recordings can be seen as a limitation, the two selected spaces are representative of a range of other common urban environments of mixed acoustic design and signal-rich information architecture. The aim was to compare and contrast two archetypal everyday spaces that participants were familiar with in a general way (none of the participants had first-hand experiences with these sound environments). At the same time the two soundscapes were distinct enough so as to produce potentially different listening interpretations and meaning-making strategies. Familiarity with these sound environments was not a distinct point of measurement or classification outside a qualitative discussion; however, it is likely that familiarity (or lack thereof) will influence temporal meaning-making to some degree so it is a worthwhile consideration for a future study that involves more soundscape examples.

Study Procedure

Participants were informed that the study involved listening to everyday situations with auditory displays, in order to gain a better understanding of designed sounds and soundscapes. In order to generate a *ThinkAloud* protocol, the main question we posed participants was: "Tell me what you're hearing as you listen. Try to identify and describe each sound you're hearing." Participants were not provided any information about what the auditory environments in the recordings were, nor were they given any visual reference of them, however we expect that all participants were reasonably familiar with the typology of the two spaces. Immediately following the *ThinkAloud* portion of the session, a short semi-structured interview was conducted with each participant to capture any further thoughts, associations or impressions that listeners were able to share. At the end of the session, which typically took 20 minutes, participants were informed where the recordings had taken place, and any questions that they had were answered in an informal discussion. All of the *ThinkAloud* comments made during the listening sessions were recorded, along with the original recordings, so that they could be transcribed in context, and subsequently annotated using a semantic content anlaysis. Study sessions were mixed to stereo so that the right channel represented the auditory environment being described, and the left contained the *ThinkAloud* comments. The second author conducted the experiments and transcribed the comments. The first author annotated the transcriptions, after first working with the second author to establish inter-rater reliability (IRR), as described below.

Analytical Framework – Semantic Content Analysis

In order to analyse the everyday listening process that participants engaged in when listening to the two respective soundscapes we conducted a content analysis of participant transcripts, sorting their commentary about sounds and sonic events roughly into *spatial*, *semiotic* and *experiential* categories. This was done through a systematic annotation, validated via an inter-rater reliability (IRR) process (Armstrong et al. 1997). The choice of categories reflects not only some of the general directions in everyday listening research outlined above, but also the very real logistical needs for recognition and appropriate interpretation of designed sonic cues in common urban environments. In other words, we needed to understand how listeners integrate incidental and designed sonic elements within a rich and complex auditory space. Based on existing schemata of sonic categories, we generated an analytical framework from within the data. Following an extensive process of listening through participants' descriptions, and taking into consideration prior work in both perceptual science and applied acoustics (Bregman 1993; Gaver 1993; Kogan et al., 2016; Craig et al., 2017), the research team arrived at an analytical framework that distinguishes between *descriptive, associational, narrative, spatial, and experiential* dimensions of listener experience. This framework bears necessary resemblance to Tuuri and Eerola's (2012) denotative dimensions including causal, empathetic, functional, and semantic ways of listening; however they depart in order to stay cohesive in relation to a narrative temporal framework of source identification and perception. Table 1 below describes these categories, along with examples from respondents' transcripts illustrating each dimension. Coding reliability for the consistent application of these categories was established via an inter-coder reliability (IRR) process where two coders (the first and second authors) applied the mutually agreed upon coding schema (Table 1) to four randomly selected participant transcripts and then the annotations were compared. Accepted standards for qualitative IRR agreement stipulate that coders have to be consistent in their application of annotations at least 70% of the time on average for the schema and coding process to be considered reliable (Krippendorff, 1995; Burla et al., 2008). The IRR test average of all four tests was established as 72% and along with a follow-up discussion it was deemed satisfactory to proceed.

Table 1: Semantic codes, definitions and examples

In order to foreground a temporal (rather than strictly semantic) analysis we separated the written transcripts of participant sessions into 30-second chunks. This allowed us to account for the types of words and categories that participants were most likely to use in each time segment. We then applied the content analysis codes to the transcripts in the same 30-second intervals in order to visualize the temporal matrix of codes used (see Figure 2). The selection of time chunks was a pragmatic choice for our 2-minute sound files, and allowed a way of identifying temporal patterns without making the time sections too granular or too coarse. In a previous short paper, we reported some preliminary results pointing to the idea that listeners start with more descriptive and spatial markers, and move on to more associational and narrative components as they gain a better sense of the auditory environment (BLANK). At that time, we worked with a sample of 10 participants, which we have, for this paper, expanded to a much higher number of 70. The aim was to test out whether this hypothesis (that listeners start with spatial and descriptive categories and move on to associational and narrative categories with time) holds true on a larger scale.

Results

Figure 1 below shows the average percentage of each type of sonic characteristic that all 70 participants used across both conditions. We can see that descriptive elements dominated, however associational qualifiers – whether related to the space, sound sources, or implied events and activities – also comprised a significant amount of participant comments. In the next section we look at how usage of these characteristics is distributed over time in the listening tasks, and we offer an analysis of findings.

Figure 1: Breakdown of semantic codes by average used by all 70 participants, both conditions.

Figure 2: Combined (Bank and Grocery store) absolute instances of semantic categories in 30-second intervals, with trend lines for narrative and associational categories.

Looking at a temporal breakdown of semantic sound categories, spatial references indeed decreased from 25% to 11% of *ThinkAloud* comments as the soundscape unfolded in its entirety, in both listening conditions (see Table 2). At the same time, as hypothesized, associational and narrative comments increased during the span of the listening task, as indicated by the trend lines in Figure 2. Descriptive references (e.g. naming sound sources or sound events directly) remained relatively consistent in all time intervals with a slight decrease from 45% to 34% of references overall (Table 2). Experiential categories and reference to sonic qualities were similarly distributed throughout each time interval, with an average of 6%, suggesting that their use was event-based (when a sound made a qualitative/palpable impression) rather than part of a deliberate cognitive listening process. Associational categories, while present throughout all time intervals, did increase over time from 11% to 23% of all sound categories named. Narrative categories – putting several sounds into a coherent situational narrative – are almost non-present in the first two time intervals and show a definitive increase in the last two time chunks. Spatial-associational references (attempting to name location or type of space) show a marked decrease in the last two intervals (3% and 4% of all comments respectively). After applying the Jonckheere Terpstra Test to the combined results (bank and grocery store) the majority of the parameters can be considered temporally significant in terms of the sequence trend: Spatial (JT= 53679, p=0.004), Spatial Associative (JT= 43614.5, p<0.001), Descriptive (JT= 51381, p<0.001), Associative (JT= 67397.5, p<0.001) and Narrative (JT= 76165, p<0.001). Descriptive Associative (JT= 59996, p=0.548) and Experiential (JT= 61045, p=0.172) categories were not significant in their ordered difference.

Table 2: Combined (Bank and Grocery store) percentage of semantic categories in 30-second intervals.

In addition to global patterns in participant listening responses across the two conditions, it is worth diving into the temporal breakdown of several key sonic categories separated by soundscape condition, Figs 3-5 below. While both environments are common urban spaces signal-rich with informational beeps and alerts, there were nuanced differences in the contextual character of each space, providing different levels of sonic information. A qualitative look into participants' interpretation of each condition shows some notable distinctions. Most listeners (over 75%) quickly identified the grocery store as a "shop", a "market," or "supermarket" – within the first 30 seconds, which could explain why we see a diminishing amount of spatial references after the first interval in that particular audio condition (see Figure 3). The increase in the last time interval refers largely to associative spatial comments describing the movement and proximity of people in space. In contrast, most participants interpreted the bank soundscape to be a range of generalized spaces including an "office space", a "parking lot", a "warehouse," a "transportation centre or subway," even "airport." As Figure 3 shows, the bank audio listening condition generated an almost even percentage of spatial category words throughout the duration of the experience, speaking to a continued effort to both describe the space and to identify its function.

Figure 3: Percentage of spatial sound categories used by participants in each 30-second interval.

With associational and narrative category references we see a definite temporal pattern emerging, again respective to each listening condition (see Figures 4 and 5). While associational references in the grocery story soundscape outnumbered those in the bank soundscape, both demonstrate a clear tendency to increase with each time interval. In particular, narrative references – the stringing together of three or more sound sources into a story – show an even more marked increase in the full duration of each soundscape; and again, an overall higher presence of narrative references in the grocery store soundscape condition (Fig 5). Given that the grocery store ambience was denser and contained more individuated elements of activity, we see overall higher numbers of narrative and associational comments. The bank ambience, conversely, was uneventful until the last time period when we hear the machine dispersing money bills for someone, clicking and beeping as it does so: hence the highest amount of associational comments in the bank soundscape are located in the last 30 seconds.

Figure 4: Percentage of associational categories spoken by participants in each 30-second interval.

Figure 5: Percentage of narrative categories spoken by participants in each 30-second interval.

Discussion

To begin with some caveats, listening in everyday contexts – technologically augmented or not – is a complex and situational activity. As Tuuri and Eerola note, a sound source can elicit multiple simultaneous listening attentions as listening can "incorporate a multitude of intentions" (2012, p.147). It is important to note that our study design influenced, to a degree, the results by letting users create their own understanding of sonic environments without the multimodal stimuli of vision. Yet precisely due to the complexity of listening, reducing the study conditions to audio only, and the analytical framework to imaginative aspects of source identification allowed us to draw some stable conclusions about temporal patterns of everyday listening. Namely that in complex environments, which contain incidental and designed sounds, people interpret what they hear in relation to the context they have constructed during the early phases of the listening. This kind of social context of listening would be determined in part, of course, by each listener's familiarity, emotional state, and connotation with each space presented. To that end, a discussion of the subtle differences between the two soundscape environments provided is in order. Based on short individual discussions with participants following the *ThinkAloud* task it appears that the bank environment presented a more abstract soundscape with familiar yet difficult to place sound signals, lack of clear human references, or definitive contextual events. Participants spent more time in this condition describing individual sound sources, commenting on spatial characteristics, and free-associating sonic elements into possible designation, functionality, and purpose of the space. The grocery store ambience contained more characteristic elements that were more frequently occurring and easier to apprehend, thus generating more associative-type references, which made it possible to place individual sounds together into a coherent story of likely events, e.g. moving trolleys of groceries, scanning items at the till, occasionally making a mistake. This is consistent with Maffiolo *et al's.* (1999) study, which suggests that listeners apprehend unfolding soundscapes as either 'amorphous' or 'event' sequences, and that determines whether listeners perceive the soundscape as a "collection of individual sounds" (Kuwano *et al.* 2002) or as *unfolding narratives* dependent on the event, place and activity. Listening in time is different from soundscape identification or assessment, however, in that listeners don't so much "analyse" an auditory scene (Bregman, 1993), but rather experience it contextually (Cain, 2008) employing *source identification, associative sorting, aesthetic, functional* and *spatial* approaches to re-constructing the soundscape into a coherent narrative.

Moreover, as our study suggests, these approaches are employed in a specific order, and depend on the contextual nature of the soundscape itself (see Fig 6). First listeners make spatial and descriptive identifications, gradually interweaving more associational, complex references, leading up to micro-narrative identifications. The more obvious and defined the soundscape – the more associative and complex narrative responses it evokes in listeners earlier on; conversely, the more abstract or amorphous a soundscape, the longer listeners focus on distinctive descriptive and spatial references, as they aim to first identify the nature and function of the location before being able to construct event narratives out of the sounds heard.

Figure 6. Conceptual schema of the temporal evolution of listening comprehension in context.

Looking at the content of the *ThinkAloud* transcripts in combination with the short follow-up interviews with participants, there are further correlations between spatial and descriptive identifications that would be invaluable for HCI design: specifically, once listeners believe they have identified the space, place and function, they tend to tune in to hearing and interpreting specific sounds as *belonging to that place*. For instance, in the case of the bank soundscape, over 25 (out of 70) participants who believed the space was a large warehouse rather than a bank proceeded to identify related sound events as ones that would typically be found in a warehouse: machinery, footsteps, people “doing jobs,” printing, echoes, reverberation, forklifts, logging info on keypads, the reverse signal beeping of trucks. In reality, these sounds were traffic noise from the busy road outside, an ATM area where people were withdrawing money from stations, the automated counting of bills, and the printing of a receipt. Sound events – such as the ringing of a mobile phone – were ignored as ones not belonging to a warehouse space. One participant who thought the space was a large multi-level parking lot described the phone ring as someone calling their friend to tell them they had found a parking spot. Listeners who managed to identify the bank soundscape as an indoor business area proceeded to hear doors opening and closing, people walking and chatting in various proximity to the recorder; those listeners also placed the sound of beeps and cell phones as normal and familiar occurrences in this type of environment.

Guidelines for audio-augmented HCI in everyday spaces

> *From the Gestalt perspective, new information is seen as organised and bridged to prior knowledge to form an organised whole, and it is the combination of the context that something sits in as well as our prior knowledge that allows us to interpret what we are looking at or listening to (Paay & Kjedskov, 2008).*

Our findings have particularly important implications for any sort of sound design of informational displays located in everyday environments, as well as for highly mediatized soundscapes such as virtual reality (VR), mobile augmented reality (AR), or audio-enabled Internet of Things (IoT). Our findings point to the holistic, contextual, and ultimately memory / association-driven perception of everyday sound. In other words, designers not only need to understand the *interface* but all the other sounds that it has to contend with in the experiential domain, all of which come to form the *listening context* within which sound sources are perceived and evaluated; not to mention design has to contend with potential changes of the listening context as an internal narrative evolves with more stimuli introduced. This can lead to a fluidity of meaning as demonstrated in our study via real-time *ThinkAloud* comments; meaning that is indeed difficult to capture in design guidelines. First it is important to establish a new sound's role and fit in the gestalt listening context of a space as that would determine whether listeners will approach it from a reflective, denotative, connotative or experiential position (Tuuri & Eerola, 2012), or a combination thereof. It is also important to note the conditions when meanings remain stable versus when they shift in response to contextual factors. Participants literally mis-heard sounds that weren't there, or ignored sounds that *were* there in favour of the sonic reality they constructed while listening to an unfolding situation: e.g. ignoring cell phone ring when it didn't fit into a work setting, or consistently recognizing error beeps at the store register (due to their unique temporal pattern and widespread familiarity). Unlike visual displays, which create finite worlds of information, an auditory (only) display engages the listener to work harder to re-create a mental picture of a space, as well as relevant events and actors in that space. With more and more auditory information displays entering the experiential domain of people's everyday lives, considering the ecology of these signals as part of a wider sonic ecology is going to increasingly important. Based on what we learned, we offer several key considerations and guidelines for designers who seek to understand everyday sound perception before they introduce 'augmented' informational displays such as alerts, announcements, sonifications, and the like, into common public situations:

Table 3. Considerations and guidelines for narrative-oriented sound design for augmented environments

Conclusions

Gestalt theory is a readily accepted concept within computing and HCI, but not sufficiently explored when it comes to the auditory, and absolutely essential for any form of sound design, especially for VR, AR and IoT technologies. Listening is a unique experience, co-created within the physical and semiotic context of the soundscape that it occurs in. Sound is temporal by nature and the interpretation of sounds unfolds in time. In this study we explore two extremely common urban environments – the bank and the grocery store – in an effort to show how individuals parse out and mentally reconstruct the contents of these soundscapes. Specifically, we are interested in these environments inasmuch as they already contain many electronic confirmatory feedback signals that typical urban dwellers experience on a daily basis. With removing the visual reference this approach intends to replicate to a small degree the way in which new sounds are interpreted in signal-rich environments with clutters of visual, haptic, and physical information. The current state-of-the-art sound design of informational cues does not take into account the holistic and individually-experienced sonic narratives in which these sonic cues are placed. Thus, a takeaway from our work towards better sonic design is 1) considering how a user might form the sonic context of a given situation in the early stages of listening; 2) counting on source identifications to shift in time based on additional cues, prior knowledge, or spatial information; and so 3) ensuring that a sound has associative qualities within its intended ecology and supports a narrative, even a micro-narrative, within that space. With the Internet of Things and augmented/mixed reality gradually becoming more ubiquitous in everyday environments (and in the pockets of individual consumers), sounds that are informational could enrich acoustic spaces rather than add layers of incomprehensible density. For instance, a typical home of the (near) future might contain several listening and auditory devices including a personal mobile device but also a voice assistant, and individual smart technologies that produce sonic alerts – fridge, dishwasher, etc. Currently these auditory displays are designed in isolation from the gestalt of a modern home and everyday experience. Still users tend to acclimatize to their habitual environments and ecologies of sound. But what happens when public spaces become increasingly augmented with IoT, 'smart' and responsive technologies, or when more of us use augmented reality (AR) devices in public space? A better design for such informational sonic accents would be for each sound to be designed to help users make sense of the cohesive whole, rather than follow the canonical conventions of individualized auditory displays of this type.

What we showcase with this study is how a focus on temporal analysis rather than on categorization tasks can reveal much about the process by which everyday listeners interpret everyday sound-augmented acoustic environments. As our analysis demonstrates, when listeners try to make sense of the world around them they construct narratives so that everything 'belongs' to the world that they believe they are inhabiting. A lot of this relies predictably on source identification, which is why so many descriptive comments were elicited during the study. Yet, after a *minimal* sufficient number of sources was recognized, users then proceed to associate sound sources with each other placing them within an *imaginative storyline*. Situational references to place interchange with descriptive elements to attempt to render spatial characteristics in line with the imagined acoustic environment. Memory and association work hand in hand with source identification as the listening experience unfolds. These findings provide not only a unique angle into temporal aspects of everyday listening, but also concrete insights into the development of any designed sounds and soundscapes, particularly those that feature invisible or occluded sound sources. Finally, this work shines a light into some otherwise internal and unconscious processes of decoding soundscapes, thus adding important qualitative contributions to the study of aural perception across the disciplines. Future work will and should include a wider variety of everyday augmented contexts, the use of binaural recordings, and an attention to participants' familiarity with these everyday settings, as well as attention to affective and empathetic aspects of listening.

References

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. Sociology, 31(3), 597Y606.

Bakker, S., Elise van den Hoven & Berry Eggen (2012). Acting by hand: Informing interaction design for the periphery of people's attention, *Interacting with Computers*, Volume 24, 3(1): 119–130.

Beyerdynamic. (2017). DT 770 PRO: Closed reference headphone for control and monitoring applications. Retrieved August 10, 2017, from http://europe.beyerdynamic.com/shop/hah/headphones-and-headsets/studio-and-stage/studio-headphones/dt-770-pro.html

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nat Rev Neurosci*, *14*(10), 693–707.

Bregman, A. S. (1993). Auditory scene analysis: hearing in complex environments. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: the cognitive psychology of human audition* (pp. 10–36). New York: Oxford University Press.

Burla, L. et al. (2008) From Text to Codings: Intercoder Reliability Assessment in Qualitative Content Analysis. Nursing Research, 57(2): 113-117.

Cain, R., Jennings, P., Adams, M., Bruce, N., Carlyle, A., Cusack, P., *et al.* (2008). SOUND-SCAPE: a framework for characterising positive urban soundscapes. In Proceedings of EuroNoise, Paris, June 29–July 4.

Craig, A., Moore, D., & Knox, D. (2017). Experience sampling: Assessing urban soundscapes using in-situ participatory methods. *Applied Acoustics*, *117*, 227–235. https://doi.org/10.1016/j.apacoust.2016.05.026

Droumeva, M., & McGregor, I. (2012). Everyday Listening to Auditory Displays: Lessons from Acoustic Ecology. In *18th International Conference on Auditory Display (ICAD2012)* (pp. 52–59). Atlanta, Georgia: International Conference on Auditory Display (ICAD).

Dubois, D., Guastavino, C., & Raimbault, M. (2006). A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta Acustica United with Acustica*, *92*(6), 865–874. https://doi.org/10.1038/

Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data* (2nd ed.). Boston: MIT Press. ISBN 0-262-05029-3.

Gaver, W.W. (1993). How do we hear in the world? Explorations in ecological acoustics. *Ecol Psychol*, 5(4): 285–313.

Giordano, B. L., McDonnell, J., & McAdams, S. (2010). Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds. *Brain and Cognition*, *73*(1), 7–19. https://doi.org/10.1016/j.bandc.2010.01.005

Gygi, B., Kidd, G.R., Watson, C.S. (2007). Similarity and categorization of environmental sounds. *Percept Psychophys*, 69(6): 839–55.

Isherwood, SJ & McKeown, JD (2017). Semantic congruency of auditory warnings. *Ergonomics*, 60 (7). pp.1014-1023.

Hespanhol, L. & M. Tomitsch (2015). Strategies for Intuitive Interaction in Public Urban Spaces, *Interacting with Computers*, Vol. 27 No. 3, pp. 311-326.

Houix, O., Lemaitre, G., Misdariis, N., Susini, P., & Urdapilleta, I. (2012). A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, *18*(1), 52–80. https://doi.org/10.1037/a0026240

Kogan, P., Turra, B., Arenas, J. P., & Hinalaf, M. (2017). A comprehensive methodology for the multidimensional and synchronic data collecting in soundscape. *Science of The Total Environment*, *580*, 1068–1077. https://doi.org/10.1016/j.scitotenv.2016.12.061

Krippendorff, K. (1995) On the reliability of unitizing continuous data. *Sociological Methodology, Vol.* 25: 47–76.

Lewis, C. H. (1982). Using the "Thinking Aloud" *Method In Cognitive Interface Design* (Technical report). IBM. RC-9265.

Maffiolo, V., Castellengo M, Dubois D. (1999). Qualitative judgments of urban soundscapes. *INTER-NOISE and NOISE-CON congress and conference proceedings*, Institute of Noise Control Engineering, pp. 1251–4.

Marcell, M., Malatanos, M., Leahy, C., & Comeaux, C. (2007). Identifying, rating, and remembering environmental sound events. *Behavior Research Methods*, *39*(3), 561–569.

McAdams, S. (1993). Recognition of sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in Sound: The Cognitive Psychology of Human Audition* (pp. 146–198). Oxford: Oxford University Press.

Niessen, M. E. (2010). *Context-Based Sound Event Recognition*. University of Groningen.

Paay, J., & Kjeldskov, J. (2008). Understanding the user experience of location based services: five principles of perceptual organisation applied. *Journal of Location Based Services*, 2(4), 267-286. DOI: 10.1080/17489720802609328

Park, T. H. (2017). Mapping Urban Soundscapes via Citygram. In Seeing Cities Through Big Data, Springer Geography (pp. 491–513). https://doi.org/10.1007/978-3-319-40902-3

Pike, C. (March 4, 2013). BBC - Research and Development: Listen Up! Binaural Sound. Retrieved April 20, 2019, from https://www.bbc.co.uk/blogs/researchanddevelopment/2013/03/listen-up-binaural-sound.shtml

Roddy, S. & B. Bridges (2018). "Sound, Ecological Affordances and Embodied Mappings in Auditory Display" In M. Filimowicz, V. Tzankova (eds.), New Directions in Third Wave Human-Computer Interaction: Volume 2 - Methodologies, Human–Computer Interaction Series. Springer International Publishing.

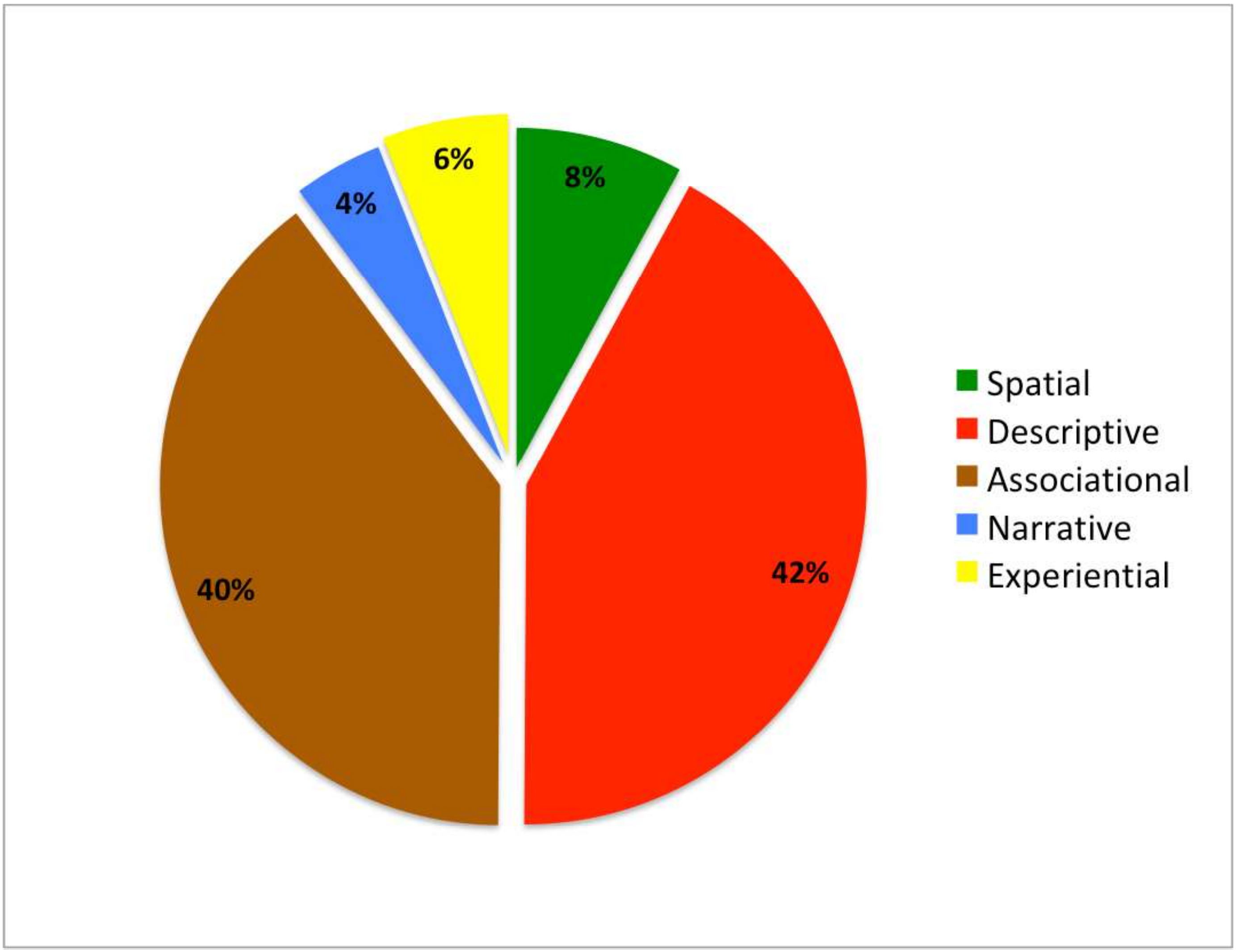
Tuuri, K., & Eerola, T. (2012). Formulating a revised taxonomy for modes of listening. Journal of New Music Research, 41(2), 137-152.

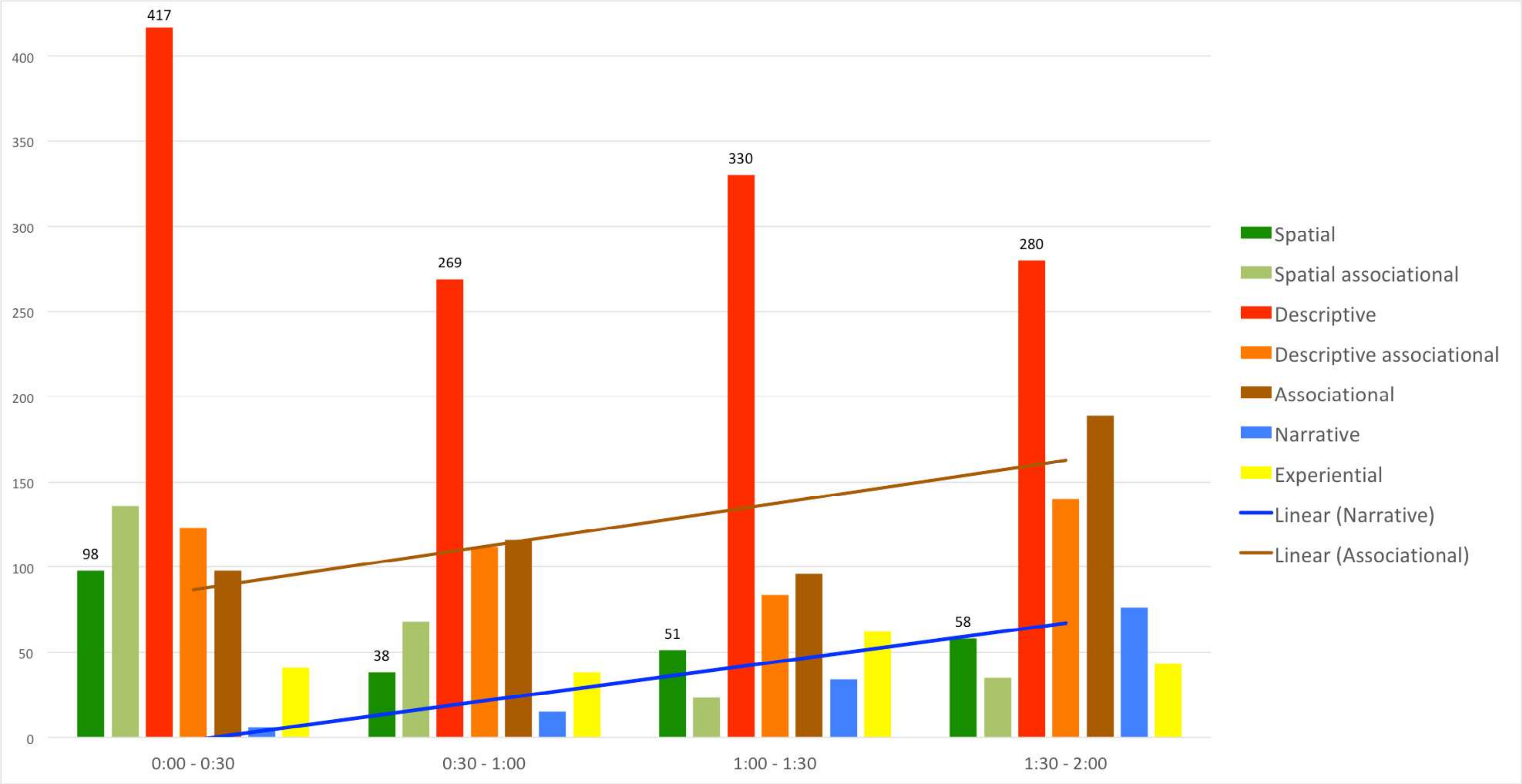
Woodcock, J., Davies, W. J., & Cox, T. J. (2017). A cognitive framework for the categorisation of auditory objects in urban soundscapes. *Applied Acoustics*, *121*, 56–64. https://doi.org/10.1016/j.apacoust.2017.01.027

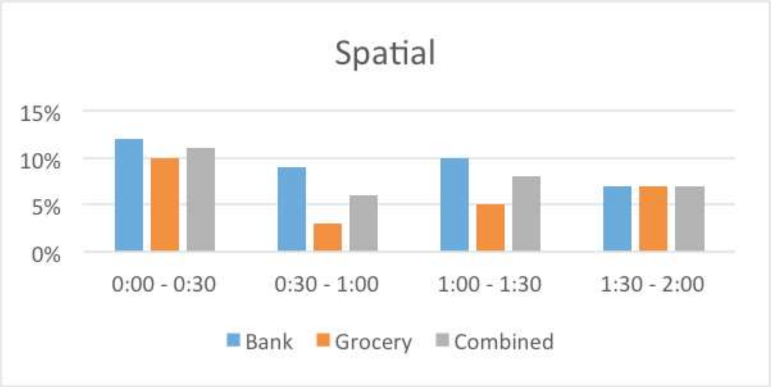
Wright, P., McCarthy, J. & L. Meekison (2003). "Making Sense of Experience." In Mark A. Blythe, Andrew F. Monk, Kees Overbeeke and Peter C. Wright (eds.), *Funology: From Usability to Enjoyment*, (pp.43-53) Kluwer Academic Publishers

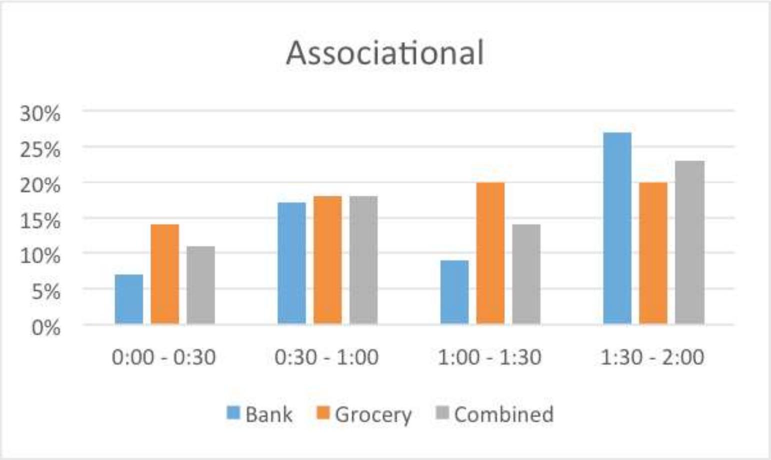
List of Figures

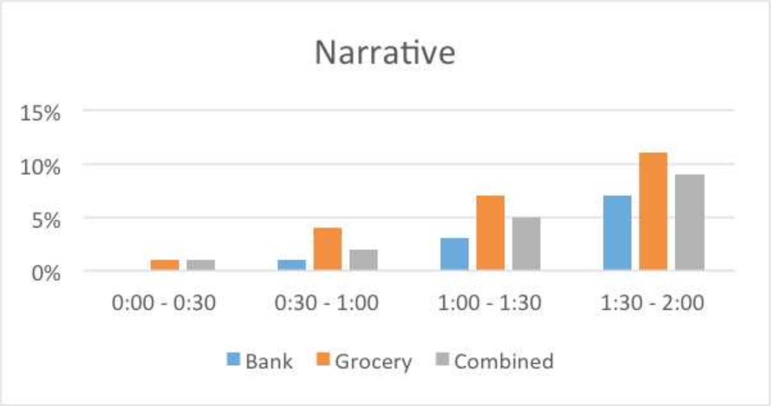
List of Tables

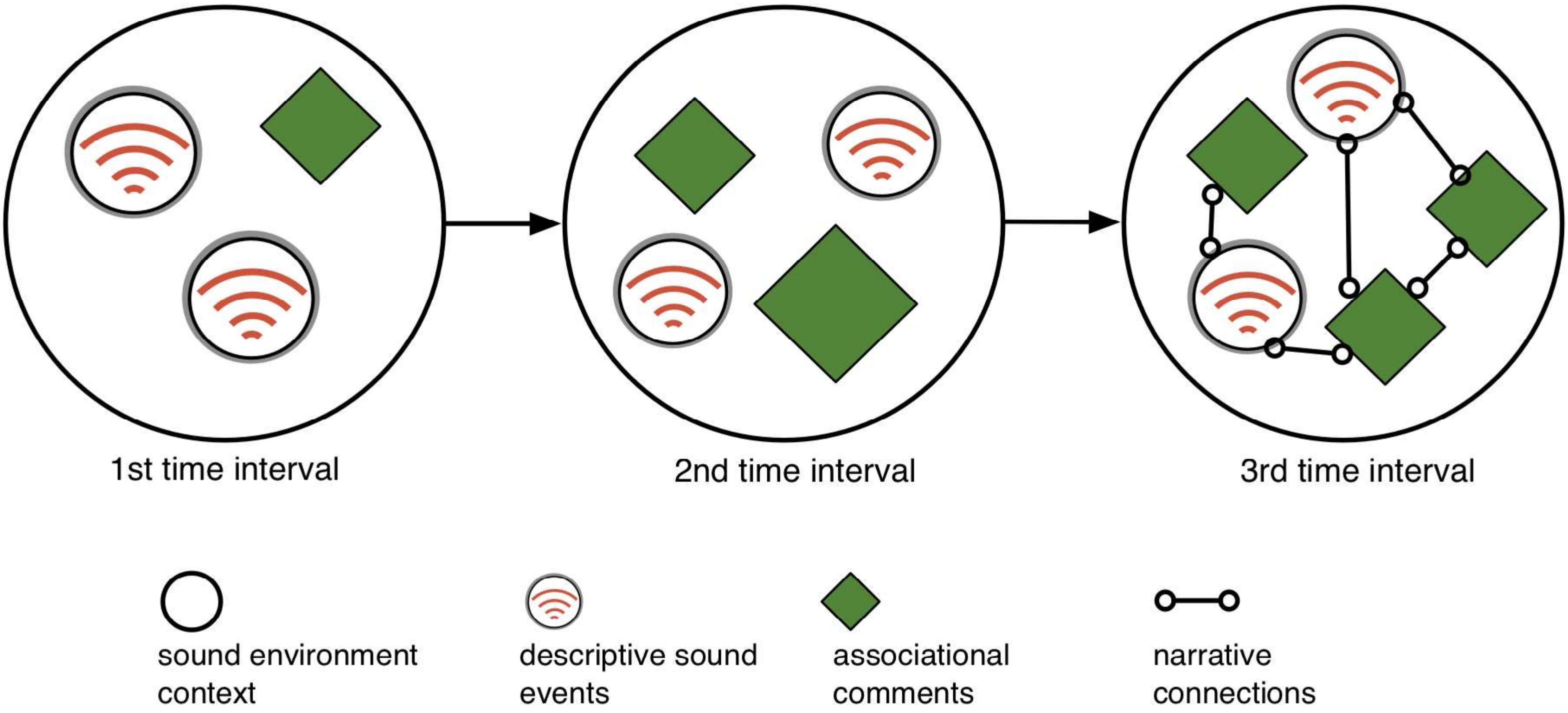

6%
8%
4%
40%
42%
Spatial
Descriptive
Associational
Narrative
Experiential


417
400
350
330
300
269
280
250
200
150
100
98
50
38
51
58
0
0:00 - 0:30
0:30 - 1:00
1:00 - 1:30
1:30 - 2:00
Spatial
Spatial associational
Descriptive
Descriptive associational
Associational
Narrative
Experiential
Linear (Narrative)
Linear (Associational)


Spatial
15%
10%
5%
0%
0:00 - 0:30
0:30 - 1:00
1:00 - 1:30
1:30 - 2:00
Bank
Grocery
Combined


Associational
30%
25%
20%
15%
10%
5%
0%
0:00 - 0:30
0:30 - 1:00
1:00 - 1:30
1:30 - 2:00
Bank
Grocery
Combined


Narrative
15%
10%
5%
0%
0:00 - 0:30
0:30 - 1:00
1:00 - 1:30
1:30 - 2:00
Bank
Grocery
Combined


1st time interval
2nd time interval
3rd time interval
sound environment context
descriptive sound events
associational comments
narrative connections

Aspects of sound comprehension	**Definition (analytical framework)**	**Examples (instances of use by Participant code)**
Spatial (S)	Making specific references to space including proximity, size, architectural features, location, buildings, etc. Words include inside/outside; close/far; big/small space; echo/reverb; left/right/up/down, background/foreground, towards/away	*P7 - I think it was either a factory or an office, but I think it was actually a bigger space than an office, or it might be a corridor in a office, but I, more I think it was actually a factory…because of all the echoes around…* *P60 - It sounds like a smaller space than the last one, because it's not as echoey.*
Spatial – Associational (SA)	Specifically identifying a context/place as a setting, i.e. naming a typical urban space along with spatial characteristics that led to its identification	*P59 - Sounds like a supermarket…Or a train station, ya, a station of some kind.* *P64 - It's like em, almost like a self-serve machine, a supermarket. You hear a lot of chatter in the background.* *P15 - I'm hearing, maybe the sort of inside of a café.*
Descriptive (D)	Identifying and naming specific sounds typically in referring to sound event/action (not source). When source is referenced it is a transparent, non-ambiguous sound, such as "car" or "printer", "woman speaking", etc. These tend to be one-word identifications without further qualifying or associational cues.	*P23 – Lots of loud beeping. Mobile phone.* *P5 – Cars. Beeping of a machine. More beeping. Footsteps. Switches being pressed. More beeps. Mobile phone.* *P57 -Screeching noise. Banging sound. Beeping sound. More talking. Car speeding.*
Descriptive – Associational (DA)	Making associational categorizations that happen to contain a concrete element or be organized around an easily recognizable sound signal e.g. "buttons being pressed". These descriptors still tend to be more concrete than abstract, referring to interactions between two sources or agents.	*P8 – Sounds like a trolley, being wheeled around.* *P49 – sounds like a door's opened. High heel shoes walking about.* *P58 - Something's printing off. Someone in high heels walking down the hall.*
Associational (A)	Making association based on an entire sequence of sounds, using associative language such as "sounds like". Specifically refers to actions or sources that are not concrete, making inferences to the context of the activity.	*P2 – Sounds like a bus of some sort, a vehicle taking off. And then some beeping, which could be, a vehicle reversing or something* *P44 – It's like a cassette tape going in and out*
Narrative (N)	Connecting several (2+) sounds together to build a story of what is happening, specifically using the context in the narrative; interpreting a combination of sounds to put a sequence of events together; a higher, holistic level of associational thinking.	*P25 - It's eh, sounds like a money counter or like a card counter you see in casinos. I hear a receipt machine printing out a receipt. I hear, as if someone is tearing the receipt from the machine.* *P1- Em, next customer's coming along, put their stuff through the scanner, again you can hear the beeps of the scanner…the customer just said they had a bag so I'm assuming the cashier's offered them one…*
Experiential (E)	Describing the quality of sounds as they are experienced; use of onomatopoeia words to refer to timbre and spectral characteristics; reference to any sound parameters: loud/quiet; timbre, pitch, rhythm, etc.	*P43 - Loud noises. Screeching footsteps. Squeaking noise.* *P53 – Squeaking of feet. And a whistling sound.* *P39 – Lots of crashing and bashing… Rustling. Creaking.*

	Spatial	Descriptive	Associational	Narrative	Experiential
0:00 - 0:30	25.20%	58.60%	10.66%	0.65%	4.46%
0:30 - 1:00	16.00%	58.00%	17.68%	2.29%	5.79%
1:00 - 1:30	10.80%	60.80%	14.12%	5.00%	9.12%
1:30 - 2:00	11.20%	51.00%	23.02%	9.26%	5.24%

	S	SA	D	DA	A	N	E
0:00 - 0:30	11%	15%	45%	13%	11%	1%	4%
0:30 - 1:00	6%	10%	41%	17%	18%	2%	6%
1:00 - 1:30	8%	3%	49%	12%	14%	5%	9%
1:30 - 2:00	7%	4%	34%	17%	23%	9%	5%

	S	SA	D	DA	A	N	E
Bank	0.027	0.000	0.061	0.139	0.000	0.000	0.241
Grocery	0.060	0.000	0.001	0.726	0.553	0.000	0.440
Combined	0.004	0.000	0.000	0.548	0.000	0.000	0.172

Considerations	Guidelines
Listening context	Identify the most characteristic sounds (soundmarks) of the space, as they are likely to configure the listening context and influence the way users interpret subsequent or new sounds. Design unique auditory signals that fit coherently into the existing context (including the architectural acoustics), to encourage correct identification.
Narrative listening in space and time	Learn the “story” of a space or situation before augmenting it with audio signals and interactions. Designed audio signals should easily fit into the local story. Think semiotics and gestalt, not just functionality and acoustics.
Sound occlusion in complex everyday situations	Designed sounds should be as informative as contextual acoustic sounds without perceptually or physically masking them (ideally occupy their own frequency niche relevant to the base established soundscape)
Afford subjective interpretations	Interpretation of sonic signals is always partially subjective and depends on interaction with the experiential domain. Design unique signals that could take on symbolic significance for frequent users.
Empathetic + Critical aspects of listening	In addition to informational / reduced listening, users experience emotional associations (especially with tonal sounds) as well as evaluative judgements on sound’s quality or functionality. Design to stimulate both registers of listening to increase seamless identification and if necessary, action.