ENVIRONMENTAL SOUNDS AS CONCEPT CARRIERS FOR COMMUNICATION

Xiaojuan Ma, Christiane Fellbaum, Perry R. Cook

Computer Science Department, Princeton University
35 Olden St., Princeton, NJ 08544, USA
{xm, fellbaum, prc}@cs.princeton.edu

ABSTRACT

Sonification, the use of nonspeech audio to represent data and information, has been applied to industrial systems and computer interfaces via mechanisms such as auditory icons and earcons. In this paper, we explore a different application of sonification, which is to facilitate communication across language barriers by conveying commonly used concepts via environmental auditory representations. SoundNet, a linguistic database enhanced with natural nonspeech audio, is constructed for this purpose. The concept-sound associations which are building blocks of SoundNet were validated through a sound labeling study conducted on Amazon Mechanical Turk. We determine the factors that cause a sound to evoke a concept. We examine which aspects of the proposed auditory representations are evocative, and what kinds of confusions may occur. Our results show that sounds can effectively illustrate some concepts, especially those related to concrete entities and actions, and thus can be utilized in assistive communication applications.

1. INTRODUCTION

In everyday life, nonspeech audio such as car horns and fire alarms has been widely used to convey specific information (e.g. alert to danger). People use sound to imply other commonly known messages as well. For instance, people sometimes fake a cough to signify that someone is uncomfortable or ill, and in comedy shows we often hear audience laughing in the background indicating that it is supposed to be a funny scene. These are all examples of sonification, "the use of nonspeech audio to convey information" [21].

Current research on sonification mainly focuses on two areas, industrial human-machine interactions [37][5][29] and computer interfaces (e.g. auditory icons [13] and earcons [6][7]). However, little work has explored the use of environmental sounds to evoke concepts for communication.

Natural language is the primary mode of communication between humans. A concept, whether it is about an entity or an event, concrete or abstract, is encoded in a linguistic form, and can be expressed verbally through words and sentences both within a language and across languages. However, language as a message carrier fails when links between concepts and their linguistic forms are missing, in situations like people trying to communicate through an unfamiliar language, people learning a new language, and people with language impairment. When the associations between words and concepts are either not yet established or corrupted, it is impossible to retrieve information via a language. To bridge language barriers, non-linguistic modalities have been explored to assist comprehension and expressions of concepts. Compared to visual languages [23][24][22][32], less attention has been given to language support through nonspeech auditory stimulus.

One disadvantage that auditory representations have over pictures is that sound requires time to play and has to be played in sequence [38]. Many concepts do not produce a (a distinctive) sound. However, there are still cases where a sound can evoke a concept even better than a picture. For example, "thunder" (unlike lightning) and "chirp" (unlike bird) are harder to visualize; "coughing" and "sneezing" can be distinguished more easily by sounds than by pictures; and "tuning a radio" can be better portrayed via a sound unfolding over time than a static picture.

To explore the use of environmental sounds as concept carriers across language barriers, we built SoundNet, a lexical network which consists of associations between concepts and short environmental sounds, and can be employed in applications like multimodal dictionaries for language learners or people with language disorders to look up concepts for communication (e.g. relaying symptoms to doctors or ordering food). A sound labeling study was conducted to verify the concept-sound associations established in SoundNet. Analysis of our results addresses issues such as what kinds of concepts can be expressed by a nonspeech sound, what aspects of a sound can be perceived, and which sounds are confusable, and guides the improvement of SoundNet.

2. BACKGROUND WORK

2.1. Sonification

Sonification refers to the use of acoustic signals to illustrate data and information. Compared to visualization, audio has been found to have the advantages of evoking temporal characteristics and showing transformation over time [19][25][26]. Furthermore, auditory display does not require users to direct their visual attention, and thus is suitable for eyes-free environments.

Sonification techniques have been applied to catching attention/alerting, and depicting changes in data by the shift of sound frequencies and intensity. Examples of such auditory systems include audio alert/monitoring and guidance systems for airplanes [5][29], nuclear power plants [37], factories [17], and scientific data analysis [30]. There are also attempts to use sound patterns on computer interfaces (e.g. earcons and auditory icons).

2.2. Earcons and Auditory Icons

Earcons are nonspeech synthetic audio patterns designed to provide information about objects, operations, status, and interactions on computer interfaces via auditory features like pitch, rhythm and volume [6][7]. People are not familiar with synthetic sounds and their assigned meanings, and thus the use of earcons requires learning. Compared to earcons, auditory icons are more natural since they encode computer events with everyday sounds

We thank the Kimberley and Frank H. Moss '71 Princeton SEAS Research Fund for supporting our project.

[13]. For example, the sound of throwing into a trashcan is used to indicate the deletion of a computer file. Additional work on auditory icons includes [15][27][13].

Both earcons and auditory icons aim to represent specific information, mainly on computer interfaces. Earcons and auditory icons are metaphor or analogy, instead of a direct translation of the everyday experience embedded in the sounds.

2.3. Everyday Listening

Everyday listening is the perception of auditory events (e.g. the characteristics of the sources of the sounds, their position and interactions), in contrast to musical listening, which captures the pitch, loudness, timbre, and changes of the sounds [16]. [20][18][36] have shown that people can identify significant aspects of environmental sounds from their experience. For instance, people can tell a car engine sound from footsteps on a wooden floor, and detect if the car is approaching or departing. Auditory icons utilize everyday listening to illustrate computer events with sounds from real life with similar effects.

By contrast, we explore the use of environmental sounds to convey everyday concepts to facilitate communication across language barriers. The intended concepts are directly linked to sources, locations, and actions involved in the sound events, and thus can be evoked through everyday listening. People working in an unfamiliar language environment, or people learning a new language, or people with low literacy, or people with language disabilities face difficulties in daily communication due to their failure to comprehend and/or produce languages. [8][9] have shown that many people with language disorders still maintain the ability to identify natural sounds. This suggests that everyday listening is viable for both healthy populations with limited language skills and language-impaired populations. Nonspeech audio has potential to assist language comprehension.

In the following sections, we describe SoundNet, a lexical network enhanced with environmental audio representations, its construction, and its effectiveness in conveying common concepts.

3. SOUNDNET

SoundNet is an environmental sound-enhanced lexical database. Different from auditory icons and earcons, the SoundNet backend vocabulary consists of hundreds of concepts (in English) used frequently in daily communications. The concepts are interlinked through semantic relations inherited from WordNet [10]. Each data unit in SoundNet (structure shown in Figure 1) has three components: a concept represented as a synonym set (synset) with its definition, an **audioability** (we define as "the ability for a concept to be conveyed by an environmental sound") rating, and a soundnail (a five-second non-speech sound) if audioable.

3.1. Vocabulary Generation

The SoundNet vocabulary consisting of commonly used concepts is based on the glossary of Lingraphica [22], a communication support device for people with aphasia. We extracted 1376 words in base form from the Lingraphica vocabulary.

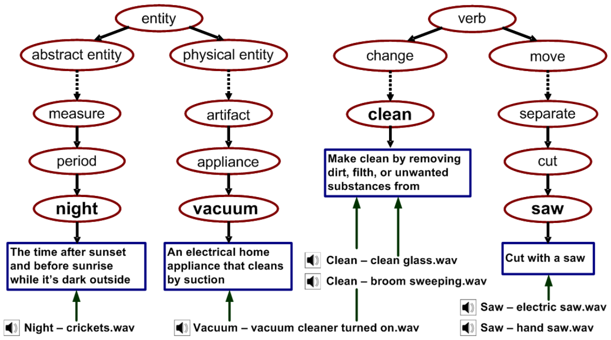

entity
abstract entity
physical entity
measure
artifact
period
appliance
night
vacuum
The time after sunset and before sunrise while it's dark outside
An electrical home appliance that cleans by suction
Night – crickets.wav
Vacuum – vacuum cleaner turned on.wav
verb
change
move
clean
separate
Make clean by removing dirt, filth, or unwanted substances from
cut
saw
Cut with a saw
Clean – clean glass.wav
Clean – broom sweeping.wav
Saw – electric saw.wav
Saw – hand saw.wav

Figure 1. Structure of SoundNet

Score	Justification	e.g.	Script
0	cannot make sound or be used to produce sound and cannot be evoked by sound	"month"	N/A
1	can make sound or be used to produce sound, but cannot be evoked by sound	"fruit"	biting an apple
2	can make sound or be used to produce sound, and may be able to be evoked by sound, meaning the sound could be ambiguous	"pen"	pen writing on paper
3	can make sound or be used to produce sound, and can be evoked by sound (the sound is distinctive)	"dog"	dog barking

Table 1. Audioability four-point rating scale and examples.

However, not every word on initial Lingraphica list could be illustrated by a sound. As a second step, we brought in sound track labels from the BBC Sound Effects Library [4] to pull out words with potential good sound-concept correspondence, since the BBC library is the major environmental sound provider for SoundNet. All the BBC sound captions were decomposed into individual words. The same process was applied to the raw BBC vocabulary as to that of Lingraphica. A list of 1368 words was generated. Its overlap with the Lingraphica list contained 211 nouns, 68 verbs, 27 adjectives, and 16 adverbs. This became the core vocabulary of SoundNet, with each word assigned to its most frequent sense and part of speech according to WordNet.

3.2. Audioability Ratings

Before attempting to create auditory representations for each concept in SoundNet vocabulary, their **audioability** was assessed on a four-point scale (Table 1). Five raters assigned each concept an audioabilty score, and for the ones with a non-zero rating, wrote a script of sound scene that could be used to evoke the intended concept. Two additional raters helped judge and finalize the audioability ratings and scripts. Overall, 184 out of 322 words were considered audioable (score > 1), and their associated sounds were selected based on the scripts.

3.3. Soundnail Creation

The three sources of the environmental audio clips employed in SoundNet include the BBC Sound Effect library (about 2/3 of the representations), Freesound [12] and FindSounds [11]. For three practical reasons, we constructed 5-second auditory illustrations called **soundnails** from the original sound files. First, most of the original sounds are dozens of seconds to several minutes in length, requiring significant listening/processing time. Second, the sound scenes with multiple events are often too complex to evoke individual concept. Third, the sounds, especially the BBC high quality stereo clips, are too large to store in mobile devices and to play on the Internet (the main interfaces to our database).

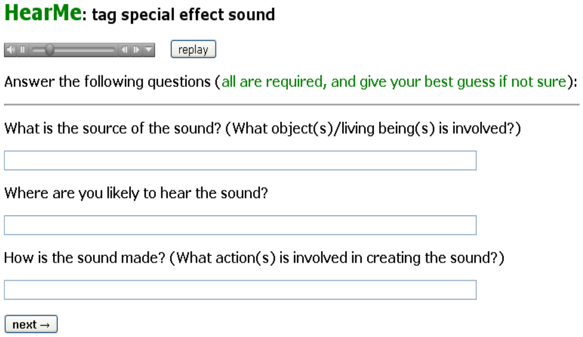
We first down-sampled all selected clips to 16kHz, 16 bit mono. We picked the 16kHz sample rate because it has been conventionally used in speech recognition, and the sample rate used in many games (especially mobile games) is between 8kHZ and 22.05kHz. A pilot study [31] also verified that people can recognize sound events at the 16kHz sample rate.

Each down-sampled sound clip was then randomly divided into five to over a hundred 5-second fragments in proportion to the length of the original clip. To pick out a representative soundnail for the given concept, the fragments were grouped into three to four clusters (based on sound scene complexity) using K-Means algorithms using six audio features, including means and standard deviations of RMS Energy, Spectral Centroid, Spectral Flux, 50% and 80% rolloff, and MFCCs [33]. The fragments closest to the center of each cluster automatically became soundnail candidates. We review all candidates which captured different distinctive parts of the original sound scene and picked out the most appropriate one to illustrate concepts in SoundNet. For example, 5-second fragments from the sound “Lines AND Tones, 3 STD Rings, Phone Answered With Pip” were clustered into “connecting,” “ringing,” and “ringing and picked up.” The representative from “ringing and picked up” was assigned to the concept “call: get or try to get into communication by telephone.”

A total of 327 soundnails were generated for the 184 audioable concepts in SoundNet. It is not a one to one mapping (Figure 1). Certain concepts are associated with more than one sound. For example, two sounds “electric saw” and “hand saw” are assigned to the verb “saw (cut with a saw).” On the other hand, some soundnails are used to depict multiple concepts. For instance, the soundnail “vacuum cleaner turned on” is assigned to both “vacuum” (noun) and “clean” (verb). As suggested by previous research [2][3], the number of options and the ease of mental image generation may affect people’s performance on sound naming. Most of the soundnails were normalized in volume, except for those that explicitly needed to have higher or lower volume, such as the soundnail for “distance”.

4. STUDY: SOUNDNAIL COMPREHENSION

Before applying SoundNet to assistive communication systems, we need to investigate if the soundnails effectively convey the pre-assigned concepts or cause confusions, and try to determine guidelines to generate more evocative auditory representations. This can be extended to more general research questions: what kinds of concepts can be evoked by a natural sound? What kinds of sounds are distinctive enough to evoke a concept? What kinds of miscomprehension may appear in everyday listening and what introduces the confusion? To address these questions, we designed and conducted a large-scale study to collect human-generated semantic labels for the nonspeech soundnails on the Amazon Mechanic Turk (AMT) platform [1]. Compared to a well-controlled lab experiment, an online study is faster, less expensive, and can access a larger number of participants more easily, despite the lack of knowledge of participants’ background and behaviours. We inserted several safeguarding methods to ensure the quality of the online study.



HearMe: tag special effect sound

Answer the following questions (all are required, and give your best guess if not sure):

What is the source of the sound? (What object(s)/living being(s) is involved?)

Where are you likely to hear the sound?

How is the sound made? (What action(s) is involved in creating the sound?)

Figure 2. Sound labeling experiment interface.

4.1. Study Design

Our goal was to determine whether, and in which cases, specific responses (nouns, verbs, adjectives, and adverbs) can be generated from auditory perception of a soundnail. Since people tended to label a sound with its source(s) in a free tagging study [14][36], we collected answers to three questions about each soundnail, so as to encourage people to generate as much information across different parts of speech as possible:

1) What is the source of the sound? (What object(s)/living being(s) is/are involved?)
2) Where are you likely to hear the sound?
3) How is the sound made? (What action(s) is/are involved in creating the sound?)

4.2. Study Environment and Participants

Figure 2 shows the web-based experiment interface. The sound automatically starts to play once the page is loaded. Subjects could replay the sound as desired. They need to submit responses regarding the source(s), location(s), and interaction(s) involved in the sound production. The study was posted on Amazon Mechanical Turk (AMT), a web service provided by Amazon, where people all over the world can post or take part in online surveys with an Amazon account.

In our sound labelling study, the 327 soundnails were randomly divided and grouped into 32 Human Intelligence Tasks (HITs), with 10 to 11 sounds in each. A HIT is the basic unit for task submission and payment. The average completion time of a HIT of tagging 10 to 11 soundnails is 14.64 minutes, not too long to get tired and lose focus. We requested at least 100 people to label each HIT, and no participants could work on the same HIT twice. It took 97 days to complete the experiment. Although we have no access to participants’ identity and demographic information, we were able to record their geographic locations. Over 2,000 people from 46 countries took part in the study, which implies that our results had universal validity. Individual responses and completion time was logged.

4.3. Quality Control

A pilot study was carried out to test the experiment interface with 22 undergraduate students. Each soundnail was tagged by five to eight students, and a post-study questionnaire gathered feedback on the design of the study. Adjustments such as auto-playing of the sound and phrasing of the questions were made accordingly.

Since we have no control over participants' behavior in the AMT study, quality-guarding schemes were applied the study:

1) **Hardware/software preparation**: The hardware (speakers or a headphone) and software (proper plugin to play the sounds) requirements were specified on the welcome page of the experimental interface. Instructions and links were provided to help with the study setup.
2) **Embedded checks**: Participants needed to correctly fill out a sequence of letters and numbers presented in an auditory **"captcha"** to login the actual study. A **training sound** was played at the beginning of each HIT. It demonstrated what kinds of sound would be played and how to answer the three questions. Participants were asked to fill out corresponding text fields as instructed as a practice. These mechanisms checked the quality of the sound system, and ensure that it was a human listening, not a computer script. Participants also get a chance to learn about the interface and tasks.
3) **Label validation**: Once the answers were submitted, non-lexical responses such as "09j1h" were automatically eliminated. To further filter out irrelevant words like "hello," the responses were compared to the labels collected from the undergraduate student pilot study. Responses with less than 50% overlap were rejected. Finally, we manually reviewed the remaining responses and kept the valid ones.

5. RESULTS AND ANALYSIS

Over 100 (up to 174) tags were collected for each soundnail in the AMT study. They are mostly in the form of sentences or short phrases. We extracted concepts out of the raw answers following the process described in Section 5.1, and a quantitative measure called sense score was computed to assess people's agreement. Section 5.2 presents the validation of concept audioability. Section 5.3 explores the influence on audioability of two linguistic properties, concreteness and parts of speech. Section 5.4 looks at three main aspects of everyday listening: source, location, and action. Section 5.5 provides a detailed discussion of confusion errors in soundnail perception.

5.1. Data Processing

The processing procedure of raw responses collected in the AMT study was similar to that for the BBC sound captions. Sentence and short phrase were broken up into "bags of words," with function words such as "a" and "or" removed. Remaining content words were checked in WordNet [10] for validity. If not found, they were transformed back to the base form using a Natural Language Toolkit stemmer [28] and then assigned to proper sense. For example, "woods" were kept while "pens" was changed to "pen." All misspellings were corrected manually.

For each soundnail, we counted how often each word appeared across all labellers. This number is referred as the **word count.** Because people may use different words to express the same idea, we further group lexicons with same or very similar meanings into units called **sense set**. By its nature, words from the same synonym set were always in the same sense set (e.g. "child" and "kid"). Other relations between words in sense set include hypernym (superordinate), hyponym (subordinate), meronym (part), holonym (whole), instance, etc. Words in a sense set could have different parts of speech. For instance, the "rain" sense set includes "rain" (n.), "rain" (v.), and "rainy" (adj.). To be distinguished from individual **words**, a sense set is referred as a **label** in the following sections. If not specified, the evaluations described below are all label (sense set) instead of word-based. The most frequent word within a sense set (from WordNet) was used as the representative for reference.

The word count of a sense set is the sum over all word counts of its members. Since a word count depends on the number of labellers and thus cannot be compared across sounds, a relative score, referred as **sense score** is calculated for each sense set per sound. It is the average number of times a sense set (label) is generated for a soundnail across all labellers.

sense score = word count of a sense set / number of labelers

The sense score indicates the strength of people's agreement on a label. The estimate of the highest sense score is 3, meaning that every labeller used the label in answers to all three questions. A sense score of 0.5 means 50% of the participants generate the label (sense set) once, and a score of 2 means each person entered the label twice on average. For each soundnail, the sense set receiving the highest sense score (**top sense score**) is considered as the **most agreed-on label**.

5.2. Audioability

For each soundnail, we compared its pre-assigned concepts in SoundNet to the most agreed-on label obtained in the AMT study. Table 2 shows the top five and bottom five soundnails based on the sense scores of intended concepts. A test for homogeneity of variances showed that sense scores for intended concepts and most agreed-on labels came from the same normal distribution. It suggests that if the pre-assigned concepts are strongly audioable (with a rating 3 in the parentheses), they are likely to be agreed-on by labelers. In contrast, people tend to generate a different more audioable concept if the intended one is less evocative.

Sound	Assigned	S.S.	Agreed-on	S.S.
cat_meowing	cat (3)	2.53	cat (3)	2.53
train_choochoo	train (3)	2.46	train (3)	2.46
telephone_ring	phone (3)	2.43	phone (3)	2.43
horm_carHorn	horn (3)	2.42	horn (3)	2.42
baby_happy	baby (3)	2.36	baby (3)	2.36
empty_waterOut	empty (2)	0	water (3)	1.68
teapot_waterFill	teapot (1)	0	water (3)	1.71
speed_carTurnFast	speed (2)	0	car (3)	1.71
skip_tapeForward	skip (1)	0	projector(3)	1.81
cracker_eatCrunch	cracker(2)	0	eat (3)	1.91

Table 2. The five most and least effective soundnails with audioability ratings and sense score (S.S.) for their pre-assigned concept and the most agreed-on labels.

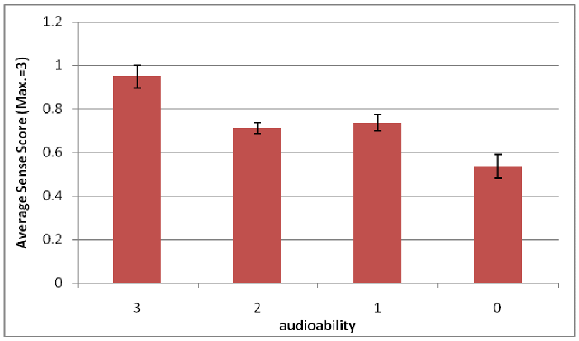


Figure 3. Comparison of audioability ratings and sense score.

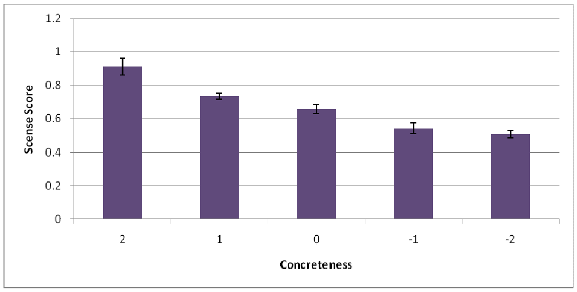


Figure 4. Comparison of concreteness and sense score.

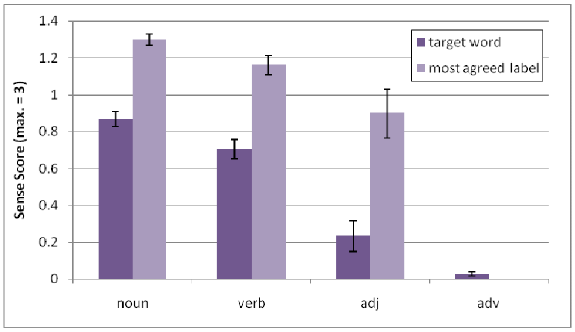


Figure 5. Comparison of sense score of target words and most agreed-on labels from different parts of speech.

Comparison of the audioability ratings and sense scores of the target concepts is shown in Figure 3. ANOVA shows that strongly audioable (rating 3) concepts received a significantly higher sense score, and scores for non-audioable concepts were significantly lower (F(1, 206) = 19.941, p < 0.01).

5.3. Relevant Linguistic Properties

How likely a concept can be evoked by an environmental sound may be affected by its linguistic properties such as concreteness and part of speech. We collected all labels with a sense score no less than 0.25 (meaning that at least 25% of the participants generated the label once), and explored the impact of two lexical properties on their sense scores.

Concreteness: Figure 4 shows that concrete words are easier to name and categorize based on nonspeech sounds, similar to the conclusions for pictures [23][24][35]. Sense scores dropped significantly as the concreteness (based on the MRC Database [34]) went down (F(1,702) = 33.596, p < 0.01).

POS	What	Where	How
Noun	313	323	256
Verb	56	15	134
Adj.	3	2	2
Adv.	0	8	0

Table 3. Comparison of numbers of labels in different parts of speech among answers to the three questions.

	Word Count		Label Count		Top Sense Score	
	Mean	SD	Mean	SD	Mean	SD
What	49.57	1.31	39.99	1.20	0.535	0.013
Where	56.46	0.93	47.64	0.90	0.414	0.010
How	104.48	1.93	86.43	1.83	0.624	0.014

Table 4. Word count, label count, and sense score comparison.

	Source	Location	Action
1	phone	playground	start (a car)
2	baby	road	type
3	chicken/rooster	farm	zip
4	car horn	house/home	honk
5	doorbell	station	ring (phone)
6	cat	school	print
7	bird	office	eat
8	train	store	uncork
9	dog	swimming pool	knock
10	typewriter	kitchen	talk

Table 5. Top 10 recognizable sources, locations, and actions.

Parts of speech: Figure 5 shows the sense score for intended (target) concepts and the most agreed-on labels for different parts of speech. Results showed that it was significantly more likely for people to generate a noun than a verb, and even more than an adjective or adverb (for target words: F(3,204) = 3.296, p = 0.0215, η^2 = 0.7673). About 80% of soundnails intended for a noun, half of those for a verb, and almost all for adjectives and adverbs were most agreed upon as nouns.

However, the parts of speech of responses varied for the different questions (Table 3). Predominantly nouns and some verbs were used for sound sources, since these are usually a person, a thing, or an action/event. Answers to "how the sound was made," mainly about sound production actions, contained many more verbs. On the contrary, fewer verbs and some adverbs (e.g. "outside") appeared in the descriptions of the location(s).

5.4. Sources, Locations, and (Inter)actions

Previous sections discussed at word level which concepts can be evoked by a sound. In this section the data are investigated from the sound perspective: what kinds of sounds are distinctive, and what aspects of the sounds can people perceive.

Table 4 shows the comparison on word counts, sense set (label) counts, and top sense scores among responses to the three questions "what," "where," and "how." In general, significantly more words (F(2,978) = 424.85, p < 0.01) and sense sets (F(2,978) = 331.84, p < 0.01) were generated to describe an interaction than to name a source or location. People are significantly more likely (F(2,978) = 67.668, p < 0.01) to agree on what kinds of actions create sound (60% soundnails) than what object(s) it was (23% soundnails) and where it took place (17% soundnails). Table 5 lists the top 10 sources, locations, and actions that people correctly recognized given the soundnails.

Assigned Type	Resp.Type (by S.S.)	Resp.Type (by %)
source	sound source (1.31)	location (21%)
	sound action (0.76)	similar source (15%)
	similar source (0.59)	sound source (12%)
	location (0.49)	sound action (9%)
source indirect	source indirect (1.87)	location (17%)
	source partial (1.37)	action partial (11%)
	action partial (0.57)	source partial (9%)
source active	source active (1.10)	source active (19%)
	source passive (0.97)	location (18%)
	sound action (0.68)	sound action (12%)
source passive	source passive (1.21)	location (14%)
	source active (0.99)	source active (10%)
	sound action (0.76)	sound action (10%)
location	sound source (1.04)	action partial (17%)
	location (0.74)	location (11%)
	action partial (0.67)	source partial (9%)
action	sound source (1.15)	location (16%)
	sound action (0.87)	sound action (15%)
	similar source (0.63)	similar source (7%)
attribute	sound source (1.17)	similar source (36%)
	sound action (0.99)	sound action (9%)
	location (0.72)	location (8%)
scene	sound source (0.83)	source partial (18%)
	scene (0.81)	location (13%)
	source partial (0.71)	action partial (11%)
	action partial (0.68)	sound source (10%)
time	sound source (1.44)	sound source (20%)
	source partial (0.95)	source partial (16%)
	location (0.69)	location (16%)
	time (0.56)	time (12%)

Table 6. Types of pre-assigned concepts and agreed-on sense sets (Resp.) ranked by sense score (S.S.) and by percentage (%).

Case	Sound	Intended	Agreed
1)	Phone, ring and pick up	phone	phone
2a)	Knock, on the door	knock	door
2b)	Bag, zipping	bag	zipper
3)	Turn, right turn signal	turn	clock
4)	Umbrella, open umbrella	umbrella	match

Table 7. Examples of how well sounds convey target concepts.

1) **Source**. Results suggest that human and animal sounds are relatively easy to name. However, how fine the distinction is depends on the sound characteristics. For example, most people can identify sea gulls but not chaffinches. For non-living objects and devices, those which are an auditory system themselves like doorbells and those which produce sounds with special temporal patterns are easier to tell.
2) **Location**. Environments in which sounds are made can be identified by the sources and events detected in the scene. For example, traffic sounds may suggest "road" whereas dish clicking sounds may suggest "kitchen."
3) **Action**. Sound-creating actions that people can name include ones that aim to make a sound, like "honk," and ones that represent an operation or a process with unique sounds generated, like "zip" and "start (a car)."

We assigned what aspect of a sound is described (type) to both the intended concepts and the sense sets (score >= 0.25). The comparison results are listed in Table 6, ranked by both sense score and frequency. If a sound is produced by a single object or living being, the source is denoted as "source" (e.g. "bird" for the bird chirping sound). If the sound is created by interaction between two objects (e.g. footsteps on the wooden floor), the one that initiates the action (e.g. feet and shoes) is called "source active," while the other one (e.g. floor) is called "source passive." If the object is not directly related to the sound scene, it is called "source indirect" (e.g. "bag" for the sound of "zipping up"). "sound source" and "sound action" refer to the actual source and action; "source partial" and "action partial" refer to part of the source/action; and "similar source" and "similar action" refer to those generating similar sounds to what were given. Results show that, regardless of what information was expected (whether it is source, location, action, or attribute), many sense sets were related to locations but all with relatively low scores. It suggests that the location information is usually more ambiguous because some sounds can appear in different places. For example, the dish clinking sound occurs in the kitchen or on the dining room table. The sound of someone coughing can happen nearly everywhere with a person. On the contrary, less labels about the source of the sounds were generated, but with high agreement.

5.5. Confusion Errors in Soundnail Perception

We compared for each soundnail the intended concept assigned in SoundNet and the most agreed-on sense set in the AMT study. The results can be categorized in four cases (Table 7):

1) The target concept appeared in the most agreed-on sense set. Soundnails in this category (90 of them) succeeded in conveying the intended concept and has the potential to enhance language comprehension and communication.
2) People agreed on a concept related to certain aspect of the sound, though not the one given in SoundNet. It indicates that the sound is distinctive but people have different focus: 2a) a different object or action; 2b) concrete content in the sound scene while the assigned concept is abstract or not directly reflected. This category has 150 soundnails.
3) Label with highest agreement was completely unrelated to the sound scene (52 sounds in this category). It suggested that those soundnails have some characteristics, but not fine enough to be told apart from similar sound events.
4) People showed no agreement on 35 soundnails, meaning that these sounds are too ambiguous to illustrate a concept.

We further looked into the semantic relations (based on WordNet) between sense set members for each question. This gives us an insight on the causes for confusion, including synonyms (e.g. car-auto mobile), hypernyms (e.g. vehicle-car), hyponyms (e.g. sports car-car), meronyms (e.g. car window-car), holonyms (e.g. window-windowpane), sisters (e.g. truck-car), nephews (e.g. fire truck-car), and instances (e.g. Ford-car). Table 8 shows that over 1/3 of the words in the responses to each question are synonyms to the representative word for the sense set they belong to, around 10% are hyponyms. However, hypernyms and meronyms got relatively higher scores (bold in Table 8). This suggests that people are more likely to recognize a more generic scope of the actual source, location, and action in the sound, or detect part of them. People usually got confused with objects or interactions in the sister or nephew categories, and even with completely unrelated events that cause similar effects or generate similar sounds.

Question	Resp. Type	Percentage	Sense Score
what	**synonym**	**0.3995**	**0.2085**
	hyponym	0.1026	0.0354
	sister	0.0691	0.0411
	hypernym	0.0676	0.0532
	similar sound	0.0386	0.0358
	nephew	0.0338	0.0270
	meronym	**0.0331**	**0.0829**
	instance	0.0286	0.0424
	holonym	0.0193	0.0324
where	**synonym**	**0.3386**	**0.1817**
	hyponym	0.0974	0.0506
	hypernym	**0.0876**	**0.0933**
	meronym	0.0788	0.0448
	nephew	0.0479	0.0429
	sister	0.0435	0.0487
	similar place	0.0411	0.0392
	instance	0.0240	0.0307
	holonym	0.0210	0.0588
how	**synonym**	**0.3490**	**0.2412**
	hyponym	0.0900	0.0430
	sister	0.0638	0.0478
	hypernym	**0.0517**	**0.0637**
	similar sound	0.0459	0.0521
	nephew	0.0398	0.0396
	meronym	**0.0391**	**0.1053**
	instance	0.0315	0.0378
	similar effect	0.0237	0.0444
	holonym	0.0204	0.0331

Table 8. Semantic relations between sense set members.

Intended Concept (bold) and Different Responses	
source	**alarm**: siren, alert, warning, doorbell, clock
	baby: infant, newborn, child, kid, toddler, little
	bottle: container, jar, can, dish, plate, glass
	car: vehicle, engine, motor, truck, bus, motorcycle
	floor: ground, stairs, porch, patio, surface
	movie: film, TV, radio, stereo, videogame
	plastic: wrapper, cellophane, polyethylene, paper
	rain: droplet, storm, hail, downpour, waterfall
	snow: dirt, dry leaves, ice, gravel, mud, twig
	typewriter: copier, fax, printer, computer, xerox
location	**farm**: barn, livestock, ranch, yard, garden, zoo
	hospital: clinic, nursery, daycare, medical center
	kitchen: restaurant, bar, café, cafeteria,lunchroom
	playground: park, court, gym, yard, stadium
	road: street, highway, race track, driveway
	school: class, classroom, college
	store: shop, supermarket, market, mall, retail
	swimming pool: lake, pond, river, ocean, beach
	train station: airport, terminal, platform, bus stop
	workshop: factory, garage, construction site
action	**break**: crack, creak, crush, shatter, smash, crash
	chirp: call, crow, sing, whistle, cackle
	clink: clank, jingle, tinkle, click, chime
	crunch: crackle, crisp, rack, scrap, scratch, break
	eat: bite, chew, munch, masticate, crunch
	jingle: rattle, rustle, fiddle, tinkle, shake
	knock: beat, kick, bang, strike, clap, hit, punch
	rub: scratch, scrub, rip, stretch, twist, squeeze
	pour: drip, fill, leak, trickle, splash, drop
	walk: gallop, run, jump, stomp, climb, jog, trot

Table 9. Examples of confusions generated for the sounds

Table 9 summarizes examples of words (some are confusion errors) people generated for the pre-assigned concepts given soundnails. The bold words are the actual sound source, location, or action. The confusions for sound sources may come from similar materials (e.g. bottle and jar) or textures (e.g. snow and gravel), and similar functions/interactions (e.g. typewriter and computer). The confusions for sound locations can be caused by similar content (e.g. farm and zoo), and similar events (e.g. playground and gym). The confusions for sound-producing actions can result from similar objects involved (e.g. knock and kick) and similar effects they lead to (e.g. crumple and squeeze).

6. DISCUSSION

The results from the soundnail labeling study may guide us towards better creation of nonspeech auditory representations.

It seems while people are focusing on everyday listening (as expected for our purposes), less information from musical listening is utilized. For example, a soundnails (far away foghorn) were used to illustrate "distance." The volume of the sound is much lower than average, but people still tried to indentify the source instead of describing the distance. In another example, the "power down" sound is used to evoke "down." People mostly wrote "videogames," "Sci-Fi," or "synthesized," rather than saying that the pitch and loudness went down.

Abstract concepts are hard to evoke. We have tried to use sounds of a special instance (e.g. the sleigh bell sound for "winter") and the combination of sounds for several concrete components of the abstract event (e.g. a sequence of rooster crowing – clock ticking – crickets chirping to depict "day" (a 24-hour period)), but none was successful. People almost always identify the concrete objects and actions, such as "bell," "rooster," and "crowing."

The effectiveness of different sounds from similar source(s) may vary greatly. For instance, the top sense score for the "saw – hand saw" soundnail is 1.78, while that for "saw – electric saw" is 0.65; the "train – choochoo.wav" sound (steam train whistling) receives a top sense score of 2.48 while the soundnail "train – arriving.wav" gets a score of 1.41. It implies certain sounds are more distinctive and should be selected as the representation.

People's familiarity with the sounds has great impact on their interpretation. This difference may come from 1) Age: younger generations have little exposure to old fashioned devices, and thus have more trouble recognize them. For example, the sense score for the "call – rotary dial.wav" soundnail is much lower with the 25 undergraduate students in the pilot study than in the AMT study. 2) Culture: people from different cultures may associate completely different sounds with the same event/scene. For example, for labelers from China, the "NBC news theme" sound used for the concept "news" may just be a piece of music. 3) Personal experience: people who have never heard an elephant trumpeting are less likely to name the sound correctly.

7. CONCLUSIONS

We presented an attempt to use short environmental sounds to convey concepts for facilitating communication across language barriers. SoundNet, a lexical semantic network enhanced with nonspeech sounds was constructed and evaluated via a large scale sound labeling study conducted on Amazon Mechanical Turk.

Results showed that over 73% of the soundnails evoked a concept that people agreed upon, 37.5% of which matched what was assigned in SoundNet. It was suggested that concrete concepts are easier to name from a sound, and many more nouns were generated than other parts of speech. As perceived in everyday listening, location(s) of a sound is the hardest to specify, whereas actions involved in the sound production are easier to distinguish. Similar materials or textures of the sound sources, similar effects of the interactions, and similar events that take place can all be the cause of confusion. Furthermore, people are more likely to agree on a more generic concept or a specific part of a complex source or action involved in the sound creation.

Overall, distinctive environmental sounds can effectively evoke concepts (nouns and verbs) commonly used in everyday communication, indicating that SoundNet has the potential of assisting communication across language barriers.

8. REFERENCES

[1] Amazon Mechanical Turk. https://www.mturk.com/ 2009.

[2] Ballas, J.A. Common factors in the identification of an assortment of brief everyday sounds. *J. of Experimental Psychology*, 19(2):250–267, 1993.

[3] Ballas, J.A. and Sliwinsky, M.J. Causal uncertainty in the identification of environmental sounds. *Tech Report ONR-86-1*, Office of Naval Research, Dept. of Psychology, Georgetown University, Washington, D. C., 1986.

[4] BBC Sound Effects Library. Original CD series. 2009.

[5] Begault, D., Wenzel, E., Shrum, R., and Miller, Joel. A Virtual Audio Guidance and Alert System for Commercial Aircraft Operations. *ICAD'96* 1996.

[6] Blattner, M., Sumikawa, D., and Greenberg, R. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*. 4(1). 1989.

[7] Brewster, S. Using nonspeech sounds to provide navigation cues. *ACM Transaction on Computer-Human Interactions*. 5(3), pp. 224-259. ACM Press. 1998.

[8] Clarke, S., Bellmann, A., De Ribaupierre, F., and Assal, G. Non-verbal auditory recognition in normal subjects and brain-damaged patients: Evidence for parallel processing. *Neuropsychologia*. 34 (6), 587-603. 1996.

[9] Dick, F., Bussiere, J., and Saygm, A. The Effects of Linguistic Mediation on the Identification of Environmental Sounds. Center for Research in Language. 14 (3). 2002.

[10] Fellbaum, C. WordNet: Electronic Lexical Database, A semantic network of English verbs. MIT Press, 1998.

[11] FindSounds. http://www.findsounds.com/. 2008

[12] Freesound Project. http://www.freesound.org/. 2008

[13] Garzonis, S., Jones, S., Jay, T., and O'Neill, E. Auditory Icon and Earcon Mobile Service Notifications: Intuitiveness, Learnability, Memorability and Preferences. In Proc. *CHI'09*. pp. 1513-1522. 2009.

[14] Gaver, W. Everyday listening and auditory icons. Doctoral Dissertation, University of California, San Diego. 1988.

[15] Gaver, W. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction*. 4. 1989.

[16] Gaver, W. What in the World Do We Hear? An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, 5 (1): 1-29. 1993.

[17] Gaver, W., Smith, R., and O'Shea, T. Effective Sounds in Complex Systems: The ARKola Simulation. In Proc. *CHI'91*. pp. 85-90. ACM Press, 1991.

[18] Handel, S. *Listening: An introduction to the perception of auditory events*. Cambridge, MA. MIT Press. 1989.

[19] Hartmann, W. M. Sounds, signals, and sensation: Modern acoustics and signal processing. Springer Verlag. 1997.

[20] Jenkins, J. J. Acoustic information for objects, places, and events. *Persistence and change: Proceedings of the first international conference on event perception*. Hillsdale, NJ: Lawrence Erlbaum Associates. 1985.

[21] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N. and Neuhoff, J. "Sonification Report: Status of the field and research agenda", ICAD technical report, 1999.

[22] Lingraphica. http://www.lingraphicare.com/. 2005

[23] Ma, X., Boy-Graber, J., Nikolova, S., and Cook, P. Speaking Through Pictures: Images vs. Icons. *Proc. ASSETS09.* 2009.

[24] Ma, X. and Cook, P. How Well do Visual Verbs Work in Daily Communication for Young and Old Adults? In *Proc. CHI 2009*, ACM Press, 2009.

[25] Moore, B. C. J. (ed.). *Handbook of perception and cognition*: Vol. 6. Hearing. 1995.

[26] Moore, B. C. J. An introduction to the psychology of hearing. 4th ed. Orlando, FL: Academic Press. 1997.

[27] Mynatt, J. Designing with Auditory Icons: How Well do We Identify Auditory Cues? Proc. *CHI'94*. 269-270. 1994.

[28] Natural Language Toolkit. http://www.nltk.org/. 2009.

[29] Patterson, R, and Milroy, R. Auditory warnings on civil aircraft: The learning and retention of warnings. *MRC Applied Psychology Unit*. Cambridge, England. 1980.

[30] Pereverzev, S. V., Loshak, A., Backhaus, S., Davis, J. C., and Packard, R. E. Quantum oscillations between two weakly coupled reservoirs of superfluid 3He, *Nature* 388, 449-451. 1997.

[31] Scavone, G., Lakatos, S., Cook, P., and Harbke, C. Perceptual Spaces for Sound Effects Obtained with an Interactive Similarity Rating Program. Intl. Symposium on Musical Acoustics, Perugia, Italy. 2001.

[32] Takasaki, T. PictNet: Semantic Infrastructure for Pictogram Communication. In Proc. *Global WordNet Conference 2006.* pp. 279-284. 2006

[33] Tzanetakis, G. and Cook, P. Musical Genre Classification of Audio Signals. In *Proc. IEEE Transaction of Speech and Audio Processing*. 10 (5), 293-302. IEEE Press, 2002.

[34] UWA Psychology. MRC Psycholinguistic Database.2009. http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm.

[35] Van Hell, J. and De Groot, A. Conceptual Representation in Bilingual Memory: Effects of Concreteness and Cognate Status in Word Association. *Bilingualism*, 1(3):193-211. 1998.

[36] Vanderveer, N. J. Ecological acoustics: Human perception of environmental sounds. Dissertation Abstracts International. 40/09B, 4543. 1979.

[37] Visuri, P. J. Multi-variate alarm handling and display. In Proc. *the International Meeting on Thermal Nuclear Reactor Safety*. National Technical Information Service. 1983.

[38] Yost, W., Popper, A, and Fay, R. Auditory Perception of Sound Sources. Springer. 2007