SoundNet: Investigating a Language Composed of Environmental Sounds

Xiaojuan Ma
Princeton University
35 Olden Street
Princeton NJ 08544, USA
xm@cs.princeton.edu

Christiane Fellbaum
Princeton University
35 Olden Street
Princeton NJ 08544, USA
fellbaum@princeton.edu

Perry R. Cook
Princeton University
35 Olden Street
Princeton NJ 08544, USA
prc@cs.princeton.edu

ABSTRACT

Auditory displays have been used in both human-machine and computer interfaces. However, the use of non-speech audio in assistive communication for people with language disabilities, or in other applications that employ visual representations, is still under-investigated. In this paper, we introduce SoundNet, a linguistic database that associates natural environmental sounds with words and concepts. A sound labeling study was carried out to verify SoundNet associations and to investigate how well the sounds evoke concepts. A second study was conducted using the verified SoundNet data to explore the power of environmental sounds to convey concepts in sentence contexts, compared with conventional icons and animations. Our results show that sounds can effectively illustrate (especially concrete) concepts and can be applied to assistive interfaces.

Author Keywords

SoundNet, environmental sound, assistive technologies.

ACM Classification Keywords

H.5.1 Multimedia Information Systems: Audio input/output; H.5.2 User Interfaces: Auditory (non-speech) feedback; H.1.2 User/Machine Systems: Human factors, Human information processing.

General Terms

Design, Experimentation, Human Factors, Languages

INTRODUCTION

Non-verbal sounds, such as fire alarms and car horns, can be used to attract attention and deliver specific messages. Currently, people have researched the use of audio in communication in two major areas. First, in Human-Computer Interfaces (HCIs), auditory icons [14] and earcons [5,6] use nonspeech audio (either natural sounds or synthetic sounds) to convey computer events. Second, in industry, audio solutions have been offered as HCIs in many settings where the visual system may not provide an effective interface in a specific environment or task, such as aircraft operation [20], power plant monitoring [28], and interfaces for the blind. However, little research has investigated the use of non-speech environmental sounds in assistive technologies to convey concepts as an extension for natural languages. An auditory language could be useful in situations where spoken languages fail to communicate effectively, as for people with language disabilities or language barriers.

Although pictorial representations have been largely used in assistive communications in such situations [12,18,24], and many concepts can be evoked with images, others can be suggested more clearly and unambiguously with sounds. Therefore, sounds may be a good complementary mode of non-verbal communication, and assistive devices might use them in conjunction with pictures. We distinguish three advantages of sounds. First, some concepts may simply not be imageable. For example, the sound for “thunder” was easily identified, yet it is difficult to imagine a picture of thunder (unlike lightening). Second, fine-grained distinction in some cases is more easily made with sounds: the sound for “sneezing” and “coughing” can be easily discriminated, but pictorial representations cannot clearly distinguish the two related activities. Third, events like “tuning (a radio)” or “rewinding (a movie),” which unfold over time, are more difficult to represent in a static image.

To study the potential of sound in assistive technologies, we explored the use of natural audio to communicate familiar and frequently occurring concepts. We built SoundNet, a lexical database enhanced with environmental sounds. SoundNet could help people with language problems to receive and express information. An example is a multimodal dictionary deployed on a mobile device. One possible scenario is that of an aphasic individual suffering from a cold and trying to convey to a nurse or doctor symptoms like “sneezing” and “coughing” by means of the dictionary. Conversely, a healthcare practitioner may create for the patient an association between a pill bottle on the table with a symptom evoked by the sound she plays from the dictionary. In all cases, sounds supplement but do not fully replace visual or verbal communication.

We are fully aware of the limitations of sounds as a means of communication. They include the fact that sounds, unlike images, require a specific sequence and longer display / processing time; many concepts are not audioable at all. Therefore, we conducted two online studies to address the

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
Copyright 2010 ACM 978-1-60558-929-9/10/04....$10.00.

questions 1) what kinds of words are "audioable" (representable by environmental sounds) and 2) how effective are sound clips in terms of illustrating concepts in daily communications. The first study collected a large number of human-generated semantic labels for the "soundnails" (short audio representations of concepts) in our database that were used to verify the concept-sound associations in SoundNet. The second study explored how well our soundnails can convey concepts as verbal fillers in common phrases, compared with icons/animations (traditionally used in assistive devices) and a baseline condition (purely guessing from the context). Our results suggest that there are many concepts that soundnails can effectively evoke, in some cases better than icons/animations. Thus, SoundNet has the potential to support communication in assistive systems for people with language disabilities and language barriers.

BACKGROUND AND RELATED WORK

Visual Languages in Assistive Technology

Assistive technologies have traditionally used iconic stimuli to illustrate concepts for people with language disabilities [23]. Many existing communication devices continue this convention [12,18]. Research [16,17] has shown that other visual representations such as web images, animations, and videos can evoke clear concepts. However, little work has been done on using auditory languages in such applications.

Audio Alerts, Monitoring, & Secondary Audio Systems

In industry, auditory systems have been applied to environments where instant attention, alert, and non-visual or non-tangible interactions are needed, for instance, airplanes [4,20], nuclear power plants [28], and the monitoring and problem diagnosing system in a factory [15].

Auditory Icons and Earcons

SonicFinder [14] is a computer interface exploring the use of auditory icons, which maps everyday sounds to computer events. Sounds like bouncing and breaking were used to convey computer events analogous to concrete events. Work on auditory icons has continued [19,13]. Auditory icons restrict to conveying computer events. Although we use natural sounds like auditory icons, our work differs in that we extend the auditory vocabulary to concepts from daily life, and we are targeting a potential user population with language problems. Earcons are nonverbal structured audio patterns intended to provide information about objects, operations, status, and interactions in computer interface elements such as menus and alerts [5,6]. However, earcons are not sounds that people are familiar with outside the specific computer environment, and thus they require learning and memorization. Earcons are less natural and accessible than auditory icons [13].

Perception of Environmental Sounds

In some cases, non-speech audio perception may be impaired together with speech perception for people who have had a stroke or brain injury, because the process may share certain channel and brain regions [21]. But evidence [7,8] has shown that many people with impaired language still retain the ability to recognize environmental sounds. This suggests that for both language-impaired populations and for healthy speakers with compromised linguistic comprehension, environmental sounds have the potential of conveying concepts and assisting language comprehension.

Scavone et al. [22] investigated how people perceive and categorize (by sound) a set of short interactive sounds. Other work such as the Freesound Project [11] collected labels for recorded sounds from human volunteers. Our work differs from previous research in that we evaluate the efficacy of audio to convey concepts from both linguistic and auditory perspectives. Furthermore, through two large online studies (> 2000 subjects in the sound labeling study, and about 240 in the Sounds as Carriers for Communication study), we collected human-generated semantic labels (free form) and interpretations (in sentence contexts) of short soundnails, which were verified and can be used to extend SoundNet. Audio studies using such large subject populations are novel in the field of Assistive Technologies.

BUILDING SOUNDNET

SoundNet is an environmental sound-enhanced lexical database. It consists of 211 nouns, 68 verbs, 27 adjectives, and 16 adverbs. All are frequently used English words. Each data unit includes a synonym set, an audioability rating and, for audioable data, a soundnail; the data are interlinked via semantic relations from WordNet [9].

Vocabulary

The original source of the SoundNet vocabulary is the glossary of Lingraphica [11], a commercial communication device developed by the Lingraphicare Company for people with aphasia. Lingraphica includes common words from different parts of speech and phrases for constructing sentences for everyday communication. After eliminating symbols and duplicates, and stemming, 1376 words were extracted from the Lingraphica vocabulary. However, we could not assume that each word on the list could be represented by a sound, a property we call **audioable**.

To better establish the sound-concept correspondence, we included the sound clip labels from BBC Sound Effects Library [3], which constitute the majority of the environmental sounds used in SoundNet. A list of 1368 words (after filtering out non-linguistic symbols and function words and stemming inflected forms) was generated from the BBC sound captions. The overlap between the Lingraphica and BBC word collections became the core vocabulary of SoundNet. Each word in the core vocabulary was assigned to its most frequent sense and part of speech as reflected in WordNet.

Audioability Rating

The next step in constructing SoundNet was to assess the audioability, which we define as "the ability for a concept to be conveyed by an environmental sound," of the words in our vocabulary. A group of five raters provided audioability scores in a four-point scale on the basis of the ability to produce sound or to be evoked by a sound. For words that are audioable (a score of 2 or 3), each of the raters wrote a scene script that could be used to evoke the intended concept. Two additional judges joined the discussion to finalize the audioability ratings and scripts. Overall, 184 out of 322 words were voted to be audioable. The scripts guided us in selecting associated sounds.

Soundnails

Over two thirds of the 184 audioable words had a representative sound in the BBC library that aligned with the rater scripts. Two other sources of environmental sounds, Freesound [11] and FindSounds [10], were checked to fill in the missing word/sounds.

However, there are three major problems with the original sound clips. First, most of them range from 10 seconds to several minutes in length. It takes time to listen to them and they are therefore not suitable for an instant communication support setting. Second, most of the sounds were recorded from a complex sound scene or event. This could distract people from focusing on a particular sound source or action that we want to depict. Third, the BBC sounds are high quality stereo and too large to store, especially for mobile devices. To address these problems, we extracted and created five-second soundnails from the original tracks.

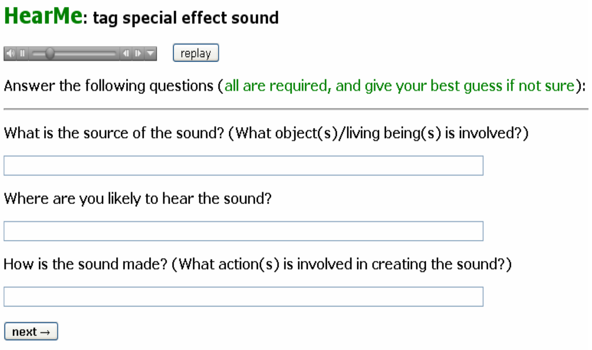
The original sounds were first downsampled to 16kHz, 16 bit mono to reduce the size of the sound files while maintaining their quality so that people can still recognize sound scenes. We chose the 16kHz sample rate based on the fact that it is a conventional sample rate for speech recognition; many video games use 11.025 or 22.05kHz for their sound effects. A pilot study [22] also proved that people could identify and categorize sounds in the 16kHz sample rate. The downsampled sound clips were then randomly chopped into 5-second fragments (the number of fragments was proportional to the original length of the track). Five seconds is a length sufficient to depict a sound source or a complete sound event, and not too long to listen to if used in a communication setting.

The 5-second fragments were grouped into three to four clusters using the K-Means algorithm based on different audio features [25] extracted from them. The fragment which was the closest to the center of each cluster was chosen as a candidate soundnail for the intended concept. In the last round, our group manually examined all the candidates and assigned the one that was the most representative to the target concept.

A total of 327 soundnails were generated for 184 words. Some words were associated with more than one soundnail, each of which was from a different domain. For example, "fire alarm," "burglar alarm," and "car alarm" sounds were all used to illustrate the concept "alarm."

STUDY1: COLLECTING HUMAN SOUNDNAIL LABELS

The current SoundNet vocabulary consisted of 211 nouns, 68 verbs, 27 adjectives, and 16 adverbs, among which 184 were determined to be audioable and associated with 327 soundnails. An online study was conducted to collect people's judgments on what concepts the given soundnails convey. The words in the labels that people agreed on were compared to the initial concept assigned in SoundNet.



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 1. Sound labeling experiment interface.

Study1: Design and Interface

The sound labeling study was carried out via the Amazon Mechanical Turk (AMT) platform [2], which allows people all over the world to post and participate in online surveys. The 327 soundnails were randomly grouped into 32 Human Intelligence Tasks (HITs) with 10 to 11 sounds each. After listening to the soundnail (which automatically plays when the web interface (Figure 1) is loaded), participants were asked to provide free form answers to three questions about the sound source, location of the sound, as well as the activities involved in the creation of the sound. After they finished labeling all the sounds in one HIT, participants submitted their work to AMT. The submission was checked both automatically and manually. Once approved, the participants received payment for their work.

Our goal was not just to gather labels for the sounds but to determine whether, and in which cases, specific aspects of the soundnails evoked responses. Thus, instead of acquiring a single label, we collected answers to three targeted questions. We hypothesize that in some cases, the location, the source, or the manner of the sound production is salient, but perhaps not all of these. We also wanted to see in which cases not all of the words in the label were named by the subjects. For example for the "walking on snow" soundnail, "walking" and "footstep" were generated, but not "snow," suggesting that the location was not audioable here.

Study 1: Quality Control

Since Amazon Mechanical Turk does not reveal information about the participants and all tasks were completed over the Internet, we had no knowledge of the background of the workers nor the quality of their work. To control the quality of the collected labels and to prevent the use of scripts or robots that can automatically fill out web forms, we embedded mechanisms and checkpoints in the interface as well as in the submission approval process.

Auditory Captcha and Training Sound

Since an appropriate web browser plug-in was needed to play the sounds, we listed both the required software (audio players) and hardware (speakers or a headset) in the first page of the study. The instruction page provided a step-by-step walk-through of the study. In order to start, participants logged in by typing the keywords revealed in an auditory captcha (short sequence of spoken letters and numbers).

This ensured that sounds played properly and that participants listened and paid attention. At the beginning of each HIT, a demo sound and spoken example answers to the three questions were played. People were asked to put down the answers as instructed. This step, too, helped to ensure that a human (not a robot) performed the task.

Pilot Study and Ground Truth Labels

Our system could filter out invalid responses such as "YYYY" and "08gv2" by automatically checking for valid words in WordNet [9], however, irrelevant answers such as "hello" and "OK" could not be eliminated. To determine the relevance of the submitted responses to the content of the sound, we ran a pilot study with 25 undergraduate students. Each soundnail was tagged by eight to nine students, and those labels became the ground truth data for comparing to the online responses. If over half of the labels in a HIT had some words appearing in the tags provided by the undergraduate students, we considered the submission as acceptable. Our group also reviewed the responses and flagged the ones considered as invalid.

Study 1: Data Process and Evaluation Metrics

After 97 days of data collection, we obtained at least 100 (up to 174) human semantic labels for each of the 327 soundnails. The raw responses were mostly in sentence format, and they were transformed into word-level data, in a process similar to that applied to the BBC sound file names. Each sentence was broken down into a bag of words; function words like "the" and "or" were removed. Subsequently, misspellings were corrected and stemming was performed based on the word's presence in WordNet. Thus, "woods" (forest) was left unchanged but "dogs" and "dragging" were normalized to "dog" and "drag," respectively.

We calculated "**word count**" (the total number of times a word appeared in labels for a sound across all labelers) for each word. In general, the more people use a word in their descriptions regarding the sound source(s), location(s), and interaction(s), the stronger the word is connected to what the sound portrays. However, people may use different words denoting the same concept, for example "plane," "airplane," and "aeroplane." In those cases, words in the same sense were grouped into units called "**sense sets**." The members of a sense sets may come from different parts of speech (e.g. "rain (n.)," "rain (v.)," and "rainy (adj.)"). The most commonly used word of each sense set was used as the representative of that whole set, and referred as a "**label**" in the following sections, to distinguish it from an individual word. The word count of a sense set is the sum of word counts over all its members. However, word counts varied as the number of labelers changed, and therefore could not be used directly for comparison across all sounds. Thus, the following metrics were introduced for the measurements and comparison.

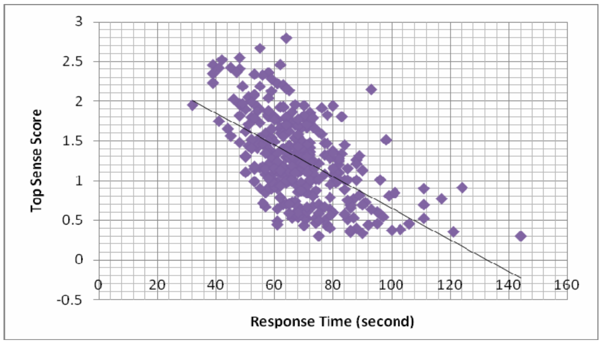


Figure 2. Correlation between response time and top sense score for each soundnail.

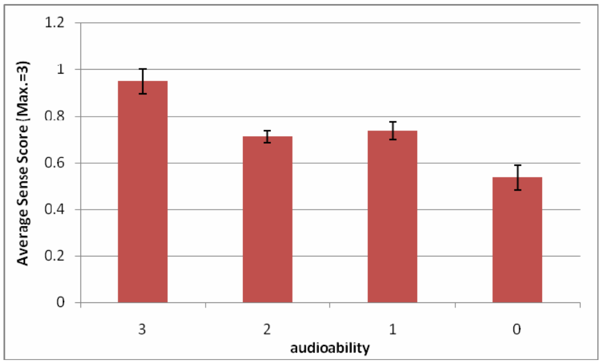


Figure 3. Comparison of audioability ratings and sense score.

- **Sense score**: the number times per person that words in a sense set were used to describe a sound, equal to the total word count of a sense set divided by the number of people who labeled the sound. The maximum sense score was "3", suggesting that for every labeler the concept appeared in answers for all three questions. Among all soundnail sense sets, the one with the **top sense score** was the concept that people most agreed on.
- **Response time**: the time (logged by interface) between the sound starting to play and the participant submitting his/her answers. Although the response time could be affected by factors like how quickly a sound is perceived, how long a sentence was used to describe the sound, how fast he/she typed, and so on, it still can reflect whether or not people had difficulty identifying a sound. Figure 2 shows that the more distinctive a sound, the less time was needed for people to listen and respond.

Study 1: Results and Analysis

For each sound, we collected sense sets with a sense score no less than 0.25 (meaning that at least 25% of the labellers came up with words in this sense set once). In general, target words that were highly audioable (audioability rating of 3) received a significantly high sense score (Figure 3).

In the following subsections, we present results of our analysis from different perspectives. First, how do the labels that participants agreed on correspond to the intended

Case	Sound	Target	Agreed
1)	Phone, ring and pick up	phone	phone
1)	Baby, crying	baby	cry
2a)	Knock, on the door	knock	door
2a)	Heart, heart beating	heart	beat
2b)	Bag, zipping	bag	zipper
2b)	Ride, horse riding	ride	horse
3)	Turn, right turn signal	turn	clock
3)	Chair, chair squeaks	chair	door
4)	Umbrella, open umbrella	umbrella	match
4)	Saucepan, hiss	saucepan	water

Table 1. Examples of cases of how well sounds convey concepts.

concepts? Second, we examine the influence of factors like parts of speech, concreteness, and imageability. Third, what role, if any, do possible cultural and linguistic differences among the participants play?

Target Words vs. Most agreed-on Labels

For each soundnail, the initial concept (target word) assigned in SoundNet was compared to the label (sense set) that labellers agreed on the most. The results can be put into the following four cases, exemplified in Table 1:

1) The target word was in the most agreed-on sense set, confirming that the sound (90 sounds in this category) succeeded in conveying the intended concept and has the potential to assist language communication.
2) The label with the highest agreement (different from the target word) matches the sound description (given in the sound file name). It showed that although different from what was intended, the sound (150 sounds in this category) was distinctive enough to illustrate a concept. There were two subcategories: 2a) the participants focused on different objects or aspects related to the sound; 2b) the intended (abstract) concept requires extra linkage to the sound scene.
3) People agreed on a concept but it was completely unrelated to the sound scene. There were 52 sounds in this category. It suggested that the soundnail was similar to the sound associated with the agreed-upon label, meaning that the sound could be communicatively effective.
4) People showed no agreement on identifying the sound, suggesting these sounds do not clearly and unambiguously illustrate a concept. Thirty-five sounds fell into this category.

Of course, cases 2 to 4 may simply suggest problems with the scripting and sound selection. Further investigation on why people generated those labels can lead us to refine and extend SoundNet.

Parts of Speech

Figure 4 shows the sense scores for target words and the most agreed-on labels for different parts of speech. If the members of a sense set represented multiple parts of speech, its sense score was counted for each of the parts of speech. Results showed that it was significantly more likely for people to associate a sound with a noun than with a verb, an adjective or adverb (for target words: $F(3,204) = 3.296$, $p = 0.022$, $\eta^2 = 0.7673$). Table 2 shows the pairwise comparison between the target word part of speech and the part of speech of the most agreed-on label. About 80% of sounds for a noun concept were labeled as a noun, while half of the sounds for a verb and almost all sounds for adjectives and adverbs were labeled using a part of speech other than the intended one. This is consistent with interpretation of pictorial representations [25].

However, the parts of speech people produced changed as they answered different questions (Table 3). Since a sound source can be a person, a thing, or an action/event, mainly nouns and some verbs were used. Responses to the location(s) of the sound contained fewer verbs in proportion, and a few adverbs indicating positions were introduced. On the contrary, the "how the sound was made" questions focused on the interaction involved, and thus a lot more verbs appeared in the descriptions.

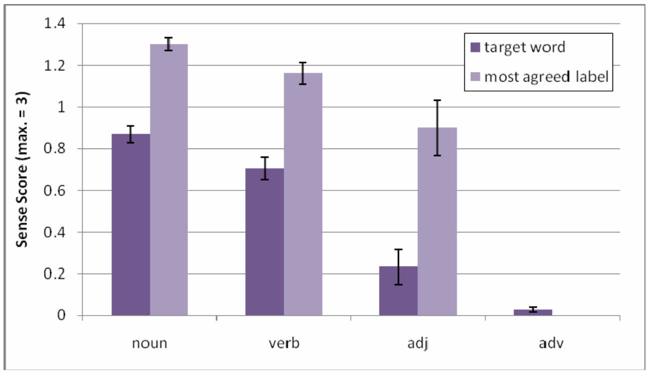


Figure 4. Comparison of sense score of target words and most agreed-on labels from different parts of speech.

Target POS	Agreed POS	count	Target POS	Agreed POS	count
Noun	Noun	231	Adj.	Noun	14
	Verb	56		Verb	2
	Adj.	4		Adj.	2
	Adv.	0		Adv.	0
Verb	Noun	38	Adv.	Noun	6
	Verb	39		Verb	1
	Adj.	0		Adj.	0
	Adv.	0		Adv.	0

Table 2. Pairwise comparison between parts of speech of the target words and those of the most agreed-upon labels.

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.

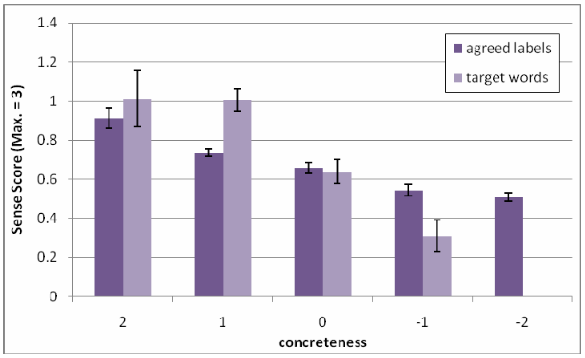


Figure 5. Comparison of sense score of target words and agreed-on labels at different concreteness level.

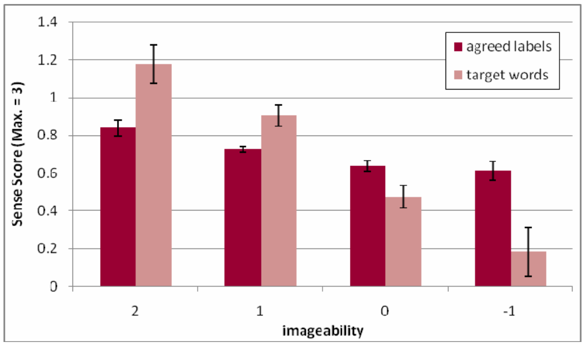


Figure 6. Comparison of sense score of target words and agreed-on labels at different imageability level.

	words	Df	F	p-value	η^2
CNC	target	1, 221	25.60	< 0.01	0.9624
	agreed	1, 702	33.60	< 0.01	0.9710
IMG	target	1, 221	36.06	< 0.01	0.9730
	agreed	1, 731	21.18	< 0.01	0.9550

Table 4. ANOVA results on concreteness (CNC) and imageability (IMG) for target words and most agreed labels.

Concreteness and Imageability

Research [27] suggested that concrete words and words that are highly imageable are easier to name and categorize based on pictorial representations than abstract words. Figure 5 and 6 show that concept recall via auditory representations followed the same rule. Sense score dropped significantly as concreteness and imageability (based on the MRC Psycholinguistic Database [26]) went down for both target words and most agreed labels (Table 4). This indicates that, in general, concrete concepts and concepts that can be easily illustrated by a picture are more likely to be conveyable by an environmental sound.

National Background of Participants

People from 46 countries and regions participated in the sound labeling study. Table 5 lists the countries with more than 10 participants. In Table 6, the average length of valid tags (removing all function words) and average response times were compared. Significant differences were found in both cases (length of tags: $F(8, 1867) = 86.114$, $p < 0.01$, $\eta^2 = 0.9885$; response time: $F(8, 1867) = 11.833$, $p < 0.01$, $\eta^2 = 0.9221$). The results revealed that the response time did not correlate with the length of tags, suggesting that other factors such as proficiency of English may be involved. Even with those differences, responses to the soundnails from people in different countries were similar.

Country	Workers	Country	Workers
United States	1344	Macedonia	15
India	465	Bahamas	12
United Kingdom	49	Philippines	12
Canada	48	Germany	11
Egypt	24	Others	55

Table 5. Examples of country and worker counts for Study 1.

Country	Tag Length (words)		Response Time (sec.)	
	Mean	SD	Mean	SD
United States	12.57	1.49	61.07	14.65
India	11.19	1.52	88.30	26.10
United Kingdom	10.86	4.39	64.04	42.24
Canada	12.10	4.00	48.98	26.59
Egypt	18.39	5.62	120.5	144.33

Table 6. Valid tag length and response time across countries.

Accuracy	Number	Example
1	29	buy, day, hair, smoke, travel, etc.
2	27	boat, chalk, fast, rain, walk, etc.
3	24	alarm, bird, cough, ice, print, etc.
4	7	baby, cat, dog, horn, phone, etc.

Table 7. Selection of target words at different accuracy level in the sound labeling study.

STUDY 2: SOUNDS AS CARRIERS FOR COMMUNICATION

Our sound labeling study showed that 89% of the SoundNet soundnails can convey a concept, and a third evoked the intended concepts. The question arose as to how effective these environmental soundnails are when used to communicate information in a context of common phrases.

A second study "Sounds as Carriers for Communication" was designed and conducted to explore answers to the following questions. First, will context improve the performance of soundnails? In the sound labeling study, 46% of our soundnails evoked concepts that were directly related to the sound scenes but differed from those we intended. It is possible that clues such as parts of speech could direct people's attention to the target. Second, how well do auditory representations perform compared to pictorial representations? Pictures have long been used in assistive technologies. If we want to apply the data in SoundNet to systems that support communication, we need to verify their effectiveness compared to the use of icons.

Study 2: Data Preparation

The goal of the study was to investigate how well people could interpret sentences in which words are replaced by soundnails based on SoundNet's audio-concept associations. It merely aims to explore how sounds can convey certain concepts when compared to icons and/or animations. Our work so far constitutes constructing and testing a new audio lexicon. Thus it is a proof of concept, not a user study for a specific population. Eighty-seven target words with different ratings from the sound labeling study were selected (Table 7). They covered all cases listed in Table 1.

The phrases used in the study came from the Ageless Project [1]. Ageless Project is a blog forum for senior people who fall into the same age span as our ultimate target population, people with aphasia. The posts in Ageless project reflect popular topics among the elderly, and thus is a good reflection of the topics important to the aphasic population and their everyday communication needs. Sentences with the selected words were crawled. Thirty-six phrases were picked and paraphrased if they were too long. Each phrase was of the length five to twelve words, and had one to four target words embedded.

Study 2: Methodologies

Design

In the Sounds as Carriers for Communication study, we introduced two other modes for comparison. One mode used icons (for nouns and adjectives) and animations (for verbs) from Lingraphica. Those iconic representations have been used for almost 20 years in assistive devices to help people with aphasia to compose phrases for language rehabilitation, and therefore, are valid for comparison. In addition, a baseline mode which shows a gap in place of the target word tested how much information the context provided. Figure 7 shows the example phrase "It is written in the book." in the three different modes.

Unlike pictures, which can appear at the same time, sounds in a phrase need to be played in sequence. To ensure the proper order, all of the phrases were turned into Flash files, which displayed the words one after another. The interval was one second for context words, and five seconds (the length of the soundnails and animations) for the words replaced with one of the modes. It helped to estimate how much time people spent on interpreting the missing words.

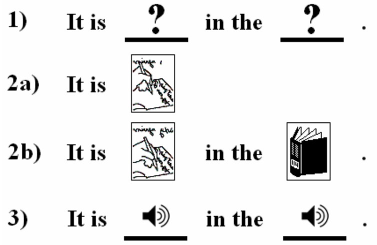
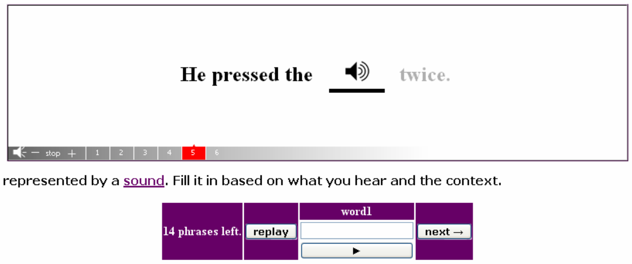


Figure 7. Phrases with target words replaced by 1) blank, 2a) and 2b) animation, 2c) icon, and 3) soundnail.


He pressed the twice.

represented by a sound. Fill it in based on what you hear and the context.

Figure 8. Sounds as Words for Communication interface.

Score	Justification	Example
0	Completely unrelated response.	wood
1	Word in hypernyms, hyponyms, or sister sets in WordNet	yacht
2	Word from the same synonym set.	**boat (target)**

Table 8. Scoring scales and justifications (target word "boat")

Interface

The study was conducted on Amazon Mechanical Turk. The 36 phrases were divided evenly into nine blocks, and regrouped into 27 HITs. Each HIT contained one block in audio mode, one in icon/animation mode, and one in blank mode. The mode assignment and position were determined using a Latin Square block design. On the interface (Figure 8), the Flash file of a phrase was automatically played. Text fields corresponding to the number of missing words were provided. People were asked to fill in their interpretation of the picture, sound, or gap. They could replay the Flash, as well as individual soundnails in the audio mode. Quality control similar to the sound labeling study was applied. The captcha was also implemented in Flash to ensure that participants had proper software installed to play the Flash files. All of the soundnails were converted to Flash, so that people did not need an extra player for the audio files.

Study 2: Results and Analysis

About 240 people participated in the Sounds as Carriers for Communication study. Each phrase in each mode was interpreted by at least 50 (up to 74) participants. Effects in different representation modes at both word and phrase levels were tested and compared.

Data Processing and Evaluation Metrics

All typed responses were collected, stemmed, and corrected for misspelling. To better assess the data quantitatively, four evaluation metrics were used. A test for homogeneity of variances in the four metrics showed that results in different modes came from the same normal distribution.

- **Accuracy rate**: the percentage of responses matching the target word, including exact matches and words from the same synonym set (e.g. child, kid).
- **Entropy**: the distribution of percentage of word count on different responses. This measures how well people's responses converged. Entropy gives low scores if users agree on a concept and high scores for distributions that are more spread out, which means more words were generated and each has a lower count across all labelers. This takes into account both the total number of different labels (sense sets) that were generated as well as the sense score for each label. Entropy for each sound was computed using the standard equation below, in which p_i was the sense score for label i: $H(p) = -\sum_i p_i \log_2 p_i$
- **Score**: the average score of all responses based on the scale in Table 8. This includes not only synonyms but also words that are similar and meaningful in the context.

Effect of Context, Concreteness, and Imageability

First, the audio mode results from the Sounds as Carriers for Communication study were compared to that from the sound labeling study. As shown in Figure 9, the target words with high sense scores in the previous study were again those with significantly higher accuracy rate than the ambiguous ones (F(1,85) = 37.037, $p < 0.01$).

However, context did provide information for people to identify the sounds or concentrate on intended aspects in many cases. Table 9 lists the 10 words with highest accuracy rate in audio mode as well as their corresponding blank mode accuracy rate. Six out of the ten words had an accuracy level of 1 or 2 in the sound labeling study, and half of them (particularly those at accuracy level 1) had an accuracy rate higher than 0.7 in the blank mode. This meant that people could guess these words quite well purely based on the context. An example is "I will bring an umbrella in case it rains." In other cases, the context suggested the part of speech of the missing word. For example, the "baby crying" sound was used to illustrate the word "cry." In the sound labeling study, many people identified the sound as "baby." The phrase given in the second study was "Her baby ____ a lot …" which indicated that the missing word should be a verb. As a result, people mostly generated "cry" instead of "baby."

Concreteness (F(1, 85) = 4.9204, p = 0.029, $\eta^2 = 0.8311$) and imageability (F(1, 85) = 8.0836, $p < 0.01$, $\eta^2 = 0.8898$) had significant impact on the perception and interpretation of soundnails. With the help of context, the accuracy rate of abstract words was greatly increased (Figure 10). The accuracy rate of words with an average level of concreteness (=0) even approached highly concrete ones. Similar effect was found in imageability (Figure 11).

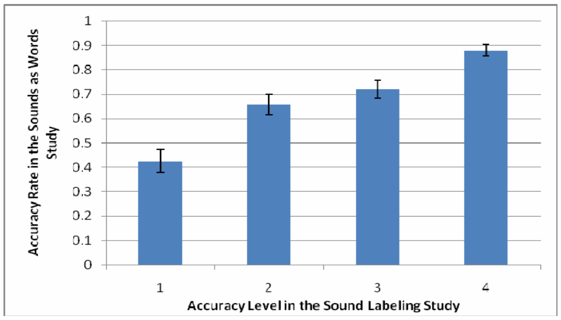


Figure 9. Comparison of accuracy with and without context.

Words	Audio Accuracy	Blank Accuracy	Labeling Accuracy
cough	1.0000	0.1970	3
cat	0.9545	0.1167	4
cry	0.9531	0.3788	2
laugh	0.9531	0.4242	2
dog	0.9508	0.2500	4
rain	0.9394	0.8919	2
wine	0.9342	0.8000	1
night	0.9298	0.7200	1
umbrella	0.9242	0.9054	1
eat	0.9153	0.8571	3

Table 9. Comparison of accuracy in audio and blank modes for the top 10 words with highest auditory accuracy

Word Level Comparison

Table 10 lists the number of different responses (F(2,252) = 117.2420, $p < 0.01$), accuracy rate (F(2,252) = 92.3268, $p < 0.01$), entropy (F(2,252) = 107.3207, $p < 0.01$), and score (F(2,258)=110.50, $p < 0.01$) of audio, icon/animation, and blank mode. In all respects, icon/animation mode performed significantly better. Entropy difference related to parts of speech of target words is significant (F(2.252) = 3.7052, p = 0.026, $\eta_p^2 = 0.7876$, $\eta^2 = 0.0327$), with responses for noun and verb concepts showing higher convergence than those for adjectives. The small eta squared effect size showed that part of speech was not as great a factor as representation mode.

Looking at the details more closely and taking entropy as an example, the results for the words can be divided into groups based on the mode with the best performance (Figure 12). Within the group where the audio mode had lower entropy value (23 words), the audio mode performed significantly better than the icon/animation mode (F(1,30) = 4.6411, p = 0.040, $\eta^2 = 0.8086$). Specifically, the audio mode significantly outperformed icon/animation mode for seven words (Figure 13) in terms of score, and the scores for another 31 words were not significantly different, indicating that certain concepts can be better conveyed by a sound than by an icon or animation.

Phrase Level Comparison

Phrase level results were similar to the word level. The average score of target words in each phrase was computed, and the icon/animation mode significantly outperformed the audio mode (F(2, 105) = 62.493, $p < 0.01$, Figure 14 green/dark columns). However, there were still five phrases for which the scores in the audio mode was significantly higher than the visual mode (Figure 15, Table 11).

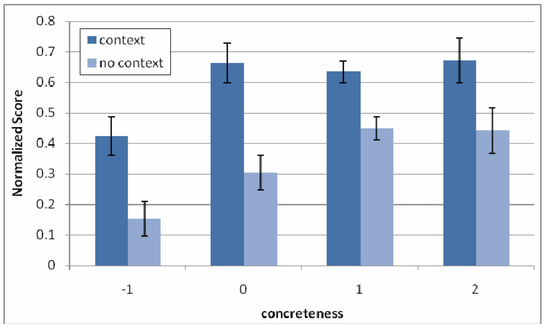


Figure 10. Accuracy at different concreteness level.

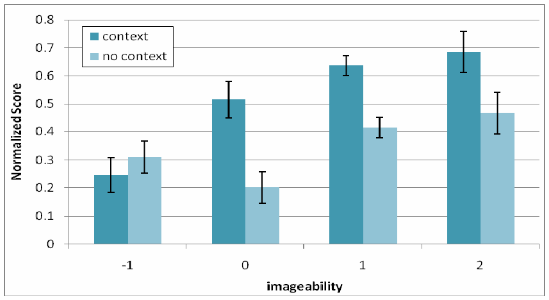


Figure 11. Accuracy at different imageability level.

Metrics	Audio		Icon/Anim.		Blank	
	Mean	Std.	Mean	Std.	Mean	Std.
Count	13.07	7.55	7.33	5.53	25.76	10.72
Accuracy	0.62	0.26	0.74	0.26	0.26	0.21
Entropy	1.72	0.98	1.07	0.84	3.08	0.99
Score	1.33	0.46	1.57	0.47	0.59	0.43

Table 10. Comparison of number of different responses (count), entropy, accuracy, and score in different modes.

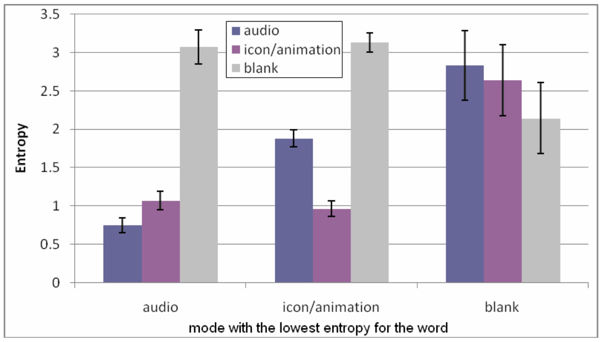


Figure 12. Comparison of entropy in different modes within groups categorized by which mode had the lowest value.

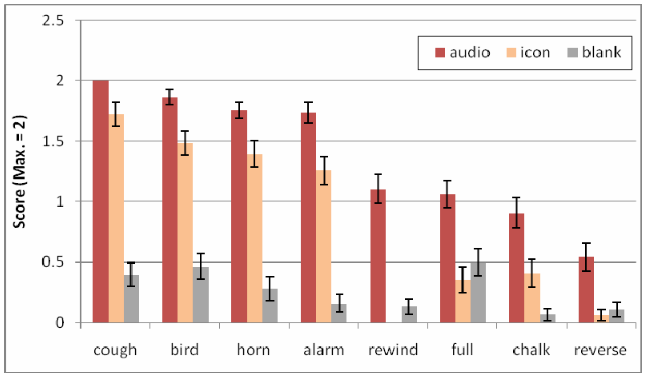


Figure 13. Words for which the audio mode score was significantly higher than the icon/animation mode score.

The response time was computed for each phrase, calculated as the time between phrase loading and the response submission (logged by the interface) minus the time spent on playing sounds for context words. Although the response time could be affected by participants' behavior in the study (for instance, some started to type as soon as the sound began to play, while others waited until the sound finished playing), it still provides a rough estimate of how long people spent on trying to figure out the missing words and typing in the answers. Figure 14 (grey columns) showed that overall, significantly more time was required for the audio mode (F(2,105) = 20.279, p < 0.01), suggesting that unlike pictures, which people can interpret at a glance, sounds may require listening to the entire clip before forming an idea. However, in the audio mode, time spent on words for which people showed low agreement was not significantly longer than that spent on words where people showed high agreement. This suggests that time might be an important feature for auditory representations, whether the sound was recognizable or not.

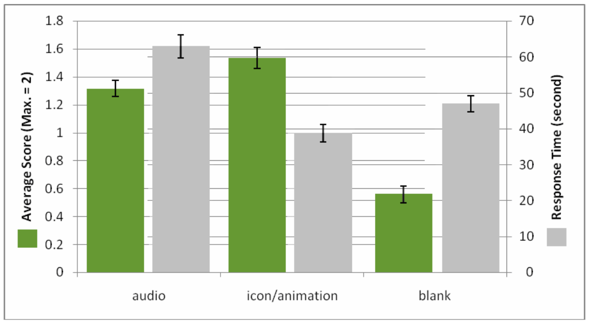


Figure 14. Comparison of phrase score and response time.

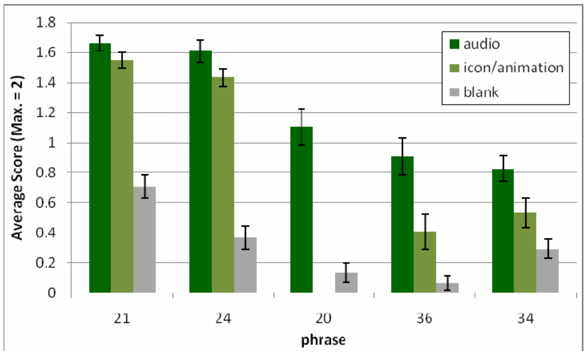


Figure 15. Phrases for which the audio mode score was significantly higher than the icon/animation mode score.

21	I heard some horn sound outside my door at night.
24	The fire alarm went on while I was cleaning the house.
20	I rewound the movie several times.
36	We have run out of chalks.
34	I am too full after having so many crackers.

Table 11. Phrases for which the audio mode score was significantly higher than the icon/animation mode score.

DISCUSSION

There were a few interesting facts observed in the studies.

- "If I see the word, I'll say, of course, it is the sound associated with an umbrella." The undergraduate students who participated in the pilot sound labeling study stated that given a sound-label pair, the association is often easily established , but given only the sound, retrieving the concept can be difficult.
- Familiarity with the sounds was also a factor that can impact people's interpretation of the soundnails. For example, one of the soundnails that was assigned to "telephone" was the dialing sound of an old style rotary dial telephone. Results showed that very few undergraduate students accurately identify the sound, whereas this soundnail receive a top sense score of 0.7265 in the Amazon Mechanical Turk study. This suggests that young people who may not be familiar with such a phone fail to recognize the source of the sound.
- An essential question is how to illustrate abstract concept with sounds. When trying to evoke the word "day" with a sound playing rooster crewing, clock ticking, and crickets chirping in sequence, most people put down "rooster"

even though the phrase was "It took a day to have the refrigerator fixed." Similarly, in an attempt to illustrate the concept "down" with the "power down" sound, almost nobody named this concept in the labeling study. Although they are closer to auditory icons, some kinds of sounds seem similar to earcons, and may require learning.

CONCLUSIONS AND FUTURE WORK

In this paper, we introduced SoundNet, a lexical network extended with environmental sounds. SoundNet provides a vocabulary of common words with an audioability rating, as well as a five-second soundnail if the word was considered audioable. The audioability property can be automatically scaled based on the semantic similarity of concepts. SoundNet carries great potential for facilitating assistive technologies with auditory representations of everyday concepts, and could be used to aid people with language disorders to receive and express information.

A large scale online study was run to collect semantic human labels on the source(s), location(s), and interaction(s) of 327 soundnails. A further study "Sounds as Carriers for Communication" was conducted to evaluate the efficacy of environmental sound representations in daily phrase context in comparison to icons and animations. Results showed that although the icon/animation mode had better performance overall, there were seven concepts for which the audio mode had significantly higher scores, while there were another 31 words for which the auditory and visual modes were not significantly different. This suggests that audio has advantages in conveying certain concepts over visual stimuli and may be able to utilize in assistive systems.

We next plan to look at combined auditory and visual cues in language comprehension. We will continue to refine and extend SoundNet, and explore applications in assistive technologies using SoundNet.

ACKNOWLEDGMENTS

We thank the Princeton Aphasia Project and SoundLab for assistance in study design and execution, and the Kimberley and Frank H. Moss '71 Research Fund for sponsorship.

REFERENCES

1. Ageless Project. http://jenett.org/ageless/. 2009.
2. Amazon Mechanical Turk. https://www.mturk.com. 2009.
3. BBC Sound Effects Library. http://www.sound-ideas.com/bbc.html. 2009.
4. Begault, D., Wenzel, E., Shrum, R., and Miller, Joel. A Virtual Audio Guidance and Alert System for Commercial Aircraft Operations. *ICAD'96,* 1996.
5. Blattner, M., Sumikawa, D., and Greenberg, R. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*. 4(1), pp. 11-44. 1989.
6. Brewster, S. Using nonspeech sounds to provide navigation cues. *ACM Transaction on Computer-Human Interactions*. 5(3), pp. 224-259. 1998.
7. 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.
8. Dick, F., Bussiere, J., and Saygm, A. The Effects of Linguistic Mediation on the Identification of Environmental Sounds. Newsletter of the Center for Research in Language. 14 (3). University of California, San Diego. 2002.
9. Fellbaum, C. WordNet: Electronic Lexical Database, A semantic network of English verbs. 1998.
10. FindSounds. http://www.findsounds.com/. 2008.
11. Freesound Project. http://www.freesound.org/. 2008.
12. Lingraphica. http://www.lingraphicare.com/. 2009.
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. The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction*. 4, pp. 67-94. 1989.
15. Gaver, W., Smith, R., and O'Shea, T. Effective Sounds in Complex Systems: The ARKola Simulation. In Proc. *CHI'91*. pp. 85-90. 1991.
16. Ma, X., Boy-Graber, J., Nikolova, S., and Cook, P. Speaking Through Pictures: Images vs. Icons. In *Proc. ASSETS09.* 2009.
17. Ma, X. and Cook, P. How Well do Visual Verbs Work in Daily Communication for Young and Old Adults? In *Proc. CHI 2009*, 2009.
18. Mayer-Johnson. http://www.dynavoxtech.com/. 2009.
19. Mynatt, J. Designing with Auditory Icons: How Well do We Identify Auditory Cues? In Proc. *CHI'94*. pp 269-270. 1994.
20. Patterson, R, and Milroy, R. Auditory warnings on civil aircraft: The learning and retention of warnings. *MRC Applied Psychology Unit*. Cambridge, England. 1980.
21. Saygm, A., Dick, F., Wilson, S., Dronkers, N., and Bates, E. Neural Resources for Processing Language and Environmental Sounds: Evidence from Aphasia. Brain. 126(4), 928-945. 2003
22. 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.
23. Steele R., Weinrich M., Wertz R., Kleczewska, M., and Carlson, G. Computer-based Visual Communication in Aphasia. *Neuropsychologia*. 27(4). pp. 409-426. 1989.
24. Takasaki, T. PictNet: Semantic Infrastructure for Pictogram Communication. In Proc. *Global WordNet Conference 2006*. pp. 279-284. 2006
25. 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.
26. UWA Psychology. MRC Psycholinguistic Database. http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm. 2009.
27. 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.
28. Visuri, P. J. Multi-variate alarm handling and display. In Proc. *the International Meeting on Thermal Nuclear Reactor Safety*. National Technical Information Service. 1983.