VocalSketch: Vocally Imitating Audio Concepts

Mark Cartwright
Northwestern University, EECS Department
mcartwright@u.northwestern.edu

Bryan Pardo
Northwestern University, EECS Department
pardo@northwestern.edu

ABSTRACT

A natural way of communicating an audio concept is to imitate it with one's voice. This creates an approximation of the imagined sound (e.g. a particular owl's hoot), much like how a visual sketch approximates a visual concept (e.g a drawing of the owl). If a machine could understand vocal imitations, users could communicate with software in this natural way, enabling new interactions (e.g. programming a music synthesizer by imitating the desired sound with one's voice). In this work, we collect thousands of crowd-sourced vocal imitations of a large set of diverse sounds, along with data on the crowd's ability to correctly label these vocal imitations. The resulting data set will help the research community understand which audio concepts can be effectively communicated with this approach. We have released the data set so the community can study the related issues and build systems that leverage vocal imitation as an interaction modality.

Author Keywords

user interaction; vocal imitation; audio software; data set

ACM Classification Keywords

H.5.2 User Interfaces: Input devices and strategies, Interaction styles; H.5.5 Sound and Music Computing: Methodologies and techniques

INTRODUCTION

In audio production applications, users currently communicate audio concepts to the software using low-level technical parameters (e.g. setting parameters on a music synthesizer). For audio search applications (e.g. finding the right door "slam" from a library of 100,000 sound effects), users typically rely on keyword descriptions stored with the audio.

It can be very difficult for a user to communicate via low-level technical parameters if they are unfamiliar with these parameters. When people communicate audio concepts to other people (rather than software), they typically do so using methods such as descriptive language and vocal imitation [10]. While descriptive language can be an effective way of communicating audio concepts [14, 10], audio concepts without broadly agreed upon labels, e.g. most sounds produced by music synthesizers, are difficult to describe with language [10].

When words fail, we vocally imitate the desired sound. In doing so, one approximates the target audio concept by mapping its pitch, timbre, and temporal properties to properties of the voice. Vocal imitation has been shown to be useful for "vocal sketching" in collaborative design [4]. In recent years, the community has begun building interfaces that use vocal imitation for a variety of tasks. These include searching audio databases for particular songs [6], drum loops [7], or sound effects [1], controlling a music synthesizer [2, 13, 8] and editing audio [12].

Given the recent interest in this area, it is important for interface designers to understand what kinds of sounds may be successfully reproduced by vocal imitation so that appropriate interfaces can be built.

Lemaitre et al have investigated vocal imitation from several angles [9, 10]. In [9], participants freely categorized vocal imitations of 12 identifiable kitchen sound recordings into 4 categories. More recently, Lemaitre et al investigated both vocal imitation and descriptive language as a means of communicating audio concepts [10]. This study used a set of 36 sounds (9 from each of the following categories: 1. *identifiable complex events*, 2. *elementary mechanical*, 3. *artificial sound effects*, 4. *unidentifiable mechanical sounds*). They found vocal imitations no more effective than descriptive language when the sound source is identifiable by name (categories 1 and 2). Vocal imitations were much more effective than descriptive language when the sound sources were not identifiable by name (categories 3 and 4). While their work is excellent, the number of sounds and participants is small, limiting the generalizability of their results.

Eitz et al performed a collection of crowd-sourced visual sketches of 250 everyday objects [3]. The authors of that study created two tasks on Amazon's Mechanical Turk: a sketching task and a recognition task, and they also released their data set to the public for others to use. Inspired by Eitz's work, we collected labels and "vocal sketches" from a large number of participants over Amazon Mechanical Turk. We built on Lemaitre's work, using a much larger set of sounds and larger population to obtain more generalizable results.

METHODOLOGY

In this work, we focus on vocal imitation to communicate audio concepts defined primarily by their timbre. We assembled a diverse set of 240 audio concepts with a wide range of timbres. Most audio concepts were defined by both a sound label (e.g. a 1-4 word description, e.g. "barking dog") and a short (up to 10 seconds) sound recording of a specific instance of the audio concept. 40 of the audio concepts did not have sound labels (discussed later). We then collected thousands

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702387

of vocal imitations from workers on Amazon's Mechanical Turk. A second set of workers described the vocal imitations and matched them to their referent audio concept. All workers were first required to pass a simple listening test.

This data set will help address the following questions: What types of audio concepts can be effectively communicated between people via vocal imitation? What are the acoustic characteristics of audio concepts that can be effectively communicated between people? The answers to these questions will clarify what applications this method of communication can be used for. The human-provided labels of the vocal imitations form a performance baseline for automated systems that search for relevant audio based on vocal imitations.

Audio Concept Set

Our audio concept set contains four subsets: *everyday*, *acoustic instruments*, *commercial synthesizers*, and *single synthesizer*. These subsets were chosen with two goals in mind: 1) diversity and 2) applicability to potential applications for vocal imitation (e.g. sound design).

The ***everyday*** subset is a set of 120 audio concepts assembled by Marcell et al [11] for confrontation naming applications in cognitive psychology (i.e. applications in which participants are asked to identify sounds). This set contains a wide variety of acoustic events such as sounds produced by animals, people, musical instruments, tools, signals, and liquids. The set includes a recording and label (e.g. "brushing teeth", "jackhammer") for each of the audio concepts.

The ***acoustic instruments*** audio concept subset consists of 40 primarily orchestral instruments [5]. Each sound recording in this subset is of a single note played on musical pitch C (where applicable) at on octave chosen to be appropriate for the range of each particular instrument. The sound labels for the audio concepts are instrument names and any short notes on the playing technique (e.g. "plucked violin").

The ***commercial synthesizers*** subset consists of 40 recordings of a variety of synthesizers in Apple Inc's Logic Pro music production suite with various popular synthesis methods. This let us explore people's ability to recognize and reproduce sounds that they could not necessarily name and and had not had many years of exposure to (unlike "brushing teeth"). Each recording was created from a "factory preset" (well-crafted settings chosen and named by Apple Inc) and consists of a single note played on musical pitch C (the octave varied according to the intended pitch range of the preset). The labels for this subset are the names of the factory presets, (e.g. "resonant clouds", "abyss of despair").

The ***single synthesizer*** subset consists of 40 recordings of a single 15-parameter subtractive synthesizer (with some limited FM and AM capabilities). Each recording consists of a note played on musical pitch C (the octave varied depending on the parameter settings). This subset was included because we know the parameter settings used, and we have the source code for this synth. This data could be used to learn mappings between vocal imitation features, referent audio features, and synthesis parameters, which is of use to researchers of new music synthesis tools and interfaces. Since these recordings were not derived from presets and are difficult to describe with language, no labels exist for these sound recordings.

Vocal Imitations of Audio Concepts

We designed two tasks for Mechanical Turk in which participants recorded a vocal imitation in response to a stimulus.

The first task addressed the use case where a user seeks a general audio concept (e.g. any church bells). Participants were given a *sound label* (e.g. the text "church bells") from our audio concept set and asked to "imagine a short (less than 10 seconds) sound produced by the following description." Next they were asked to "indicate the degree of confidence you have in your ability to imagine the sound" on a discrete scale. They were then given a simple recording interface and asked to "imitate the imagined sound with your voice as closely as possible." They were told to avoid using conventional onomatopoeia (e.g. "meow"), but that whistles, coughs, clicks, and other mouth noises were okay. Before continuing to the next step they were required to listen again to their recording and to indicate how satisfied they were with the recording. Participants were allowed to rerecord their vocal imitations unlimited times before proceeding. Discarded imitations were saved as "drafts" on our server.

The second task addressed the use case where a user seeks to reproduce the exact sound of a specific instance of an audio concept (e.g. the sound of specific church bells). This task was similar to the first task, but instead of imitating the imagined sound of a description, participants were asked to listen to a reference *sound recording* (e.g. a recording of church bells) and to imitate it with their voice as closely as possible. They were then required to listen to both the reference recording and their own recorded imitation and indicate their satisfaction with the imitation. They were allowed to rerecord their vocal imitations unlimited times until satisfied. They were then asked to "describe the identity of the source of the reference audio" (using less than 5 words) and to indicate their confidence in their description on a discrete scale.

Data Overview

Including all draft imitations, there were 10750 vocal imitations recorded by 248 unique participants. All of the submitted imitations (i.e. non-"drafts") were listened to by one of the authors, and any recordings lacking a vocal imitation or of poor recording quality were removed. The remaining recordings form the subset discussed in the remainder of the paper. This subset contains 4429 vocal imitations (2418 from the *sound recording* stimulus task, 2011 from the *sound label* stimulus task) recorded by 185 unique participants. Of the 175 participants that completed the survey, 100 identified as male / 75 as female, and their mean reported age was 31.8 (SD=8.5). The median number of imitations per participant was 4 (min=1, max=204). There were at least 10 (max of 11) vocal imitations collected for each of the 240 sound recordings and 200 sound labels.

Human Recognition of Vocal Imitations

To establish a human baseline level for the effectiveness of vocal imitations for communication, we had Mechanical Turk workers perform several identification tasks.

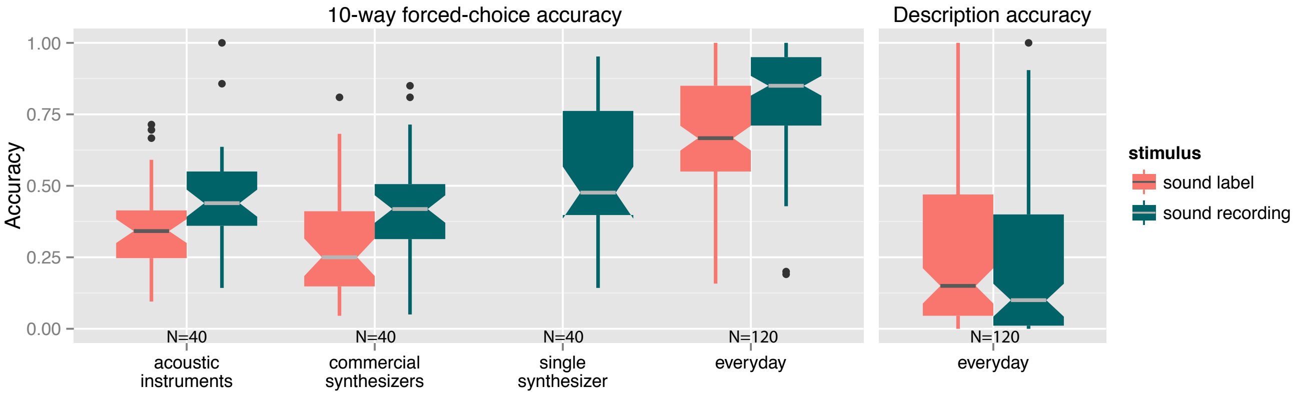

Figure 1. Human recognition accuracy of vocal imitations. The boxes extend from the lower to upper quartiles of the data. The horizontal lines in each box are the medians, and the notches around the median are 95% confidence intervals.

For both the 2418 vocal imitations produced in response to reference sound recordings (a recording of a jackhammer) and the 2011 vocal imitations produced in response to descriptive labels (the word ”jackhammer”), participants were presented a randomly selected vocal imitation. They could play this imitation unlimited times. They were asked to describe what it was an imitation of, using five words or less. They were then asked to indicate their degree of confidence in this free-response description on an 5-level scale.

Then, if the vocal imitation had been produced in response to a sound recording, the participant was presented 10 recordings drawn from the same audio concept subset (e.g. if it was an *everyday* sound they were presented 10 distinct everyday sounds): the referent recording and 9 random distractors. After hearing each recording in its entirety at least once, they were presented a 10-way forced choice to identify the recording the imitation was based on. Lastly, they had to indicate their confidence in their choice on a 5-level scale. Participants could play all of the recordings unlimited times.

Similarly, if the vocal imitation had been produced in response to a sound label, the participant was presented a 10-way forced choice between labels from the same audio concept subset. The task was to choose the label that the vocal imitation had been produced in response to. Participants could play the imitation recording unlimited times.

Data Overview

There were at least two (maximum of three) identification task instances assigned for each vocal imitation we collected. There were a total of 9174 identifications by 329 unique participants. The median number of identifications per participant was 10 (min=1, max=424).

RESULTS

In this section, we provide a brief analysis of the human baseline performance on the identification of vocal imitations.

The authors of the *everyday* subset [11] published guidelines for scoring short descriptions of their sound recordings for binary recognition (e.g. description must contain “both the object (‘door’) and the closing action” for “door closing”). Using these guidelines, we scored our participants’ descriptions of the everyday sounds for recognition. We found that our participants’ mean recognition accuracy across the 120 *everyday* sounds was 0.80 (SD=0.25). This was similar to that of the previous study: mean=0.78 (SD=0.25). Comparing these to the previous study’s results on the same 120 sounds, we obtained a Pearson correlation of r=0.84 (p=0.0), and a paired t-test of t(119)=1.67 (p=0.097). Therefore, our participants via Mechanical Turk performed comparably to lab-consented participants, giving validity to the effort our participants put into the task.

Figure 1 shows the recognition accuracy for the vocal imitations, grouped by audio concept subset and stimulus type. “10-way forced-choice accuracy” refers to the recognition accuracy of the participants’ 10-way forced-choice response in the identification task. For the *sound recording* stimulus vocal imitations, the mean recognition accuracy, broken down by audio concept subset was *acoustic instruments*: 0.45 (SD=0.18), *commercial synthesizers*: 0.42 (SD=0.18), *single synthesizer*: 0.54 (SD=0.22), *everyday*: 0.80 (SD=0.17), and mean accuracy for the *sound label* stimulus vocal imitations was *acoustic instruments*: 0.35 (SD=0.16), *commercial synthesizers*: 0.29 (SD=0.19), *everyday*: 0.68 (SD=0.20). Note that chance performance on all these tasks is 0.1.

Comparing each subset by stimulus type, the mean accuracy of the *sound recording* stimulus is greater than the *sound label* stimulus for all subsets (excluding *single synthesizer* since it lacks sound labels) according to one-sided paired t-tests, $p < 0.01$ in all cases. This difference is likely due participants’ varied interpretations of what a text-based label means.

Due to space constraints, we focus the rest of our analysis on the vocal imitations from the “sound recording” stimulus tasks. The *everyday* sounds were communicated the most effectively with vocal imitations. This may be due to the familiarity but also reproducibility of the *everyday* subset. Within that subset, imitations with the highest accuracy from the sound recording stimuli were typically sounds easily producible by the voice (human and animal sounds - e.g. “yawning”, “wolf”) or those with salient time-varied characteristics (e.g. “police siren”). Those with the lowest recognition accuracy were likely harder to accurately imitate with a single voice. For instance "glass breaking" (accuracy=0.20) has many overlapping small sonic events.

After a Welch's f-test to test for equal means between the four audio concept subsets (p=0.0), we performed a pairwise t-test with Bonferroni correction and found that the difference of 0.12 in recognition rates between the *single synthesizer* (0.54) subset and the *commercial synthesizers* (0.42) subset bordered on statistical significance (p=0.058). Without an in depth acoustic analysis it is hard to establish why, but the data set is available to allow this follow-on work. One hypothesis is that the audio concepts in the *single synthesizer* class typically have simpler but strong modulation characteristics (i.e. salient temporal properties) which may have aided in the imitation and therefore recognition of these sounds.

In Figure 1, "Description accuracy" refers to the recognition accuracy of the participants' free-response descriptions *of the vocal imitations* in the identification task. We again used the same scoring guidelines described in [11], and therefore we only scored the *everyday* subset, which was the same set used in their work. The mean recognition accuracy was 0.23 (SD=0.27) for the *sound recording* stimulus vocal imitations, and mean=0.27 (SD=0.27) for the *sound label* stimulus vocal imitations. Some audio concepts had a 0% recognition (e.g. "blinds"), while some had a 100% recognition (e.g. "sheep"). While the free-response recognition accuracy is much lower than the forced-choice recognition accuracy, this is to be expected since the participants must describe the imitation without any additional clues. However, when failing to identify the referent audio concept, participants often described similar (e.g. "motorcycle" instead of "lawnmower") or possibly more general (e.g. "horn" instead of "boat horn") concepts. This implies that more information may be needed help to disambiguate or refine certain concepts. In an audio application, this could be achieved by asking the user for additional information.

DISCUSSION

As this work is primarily about providing a data set for the community to use, the analysis in this work is intended to be illustrative rather than comprehensive. From our analysis, it seems *everyday* audio concepts were communicated the most effectively, though in a real application additional information may need to be provided by a user to disambiguate concepts. In the remaining instrumental subsets, audio concepts from the *single synthesizer* subset were communicated the most effectively, but further acoustic analysis is required to determine what enables some audio concepts to be communicated more effectively than others with vocal imitation. The data set in this paper can be obtained at `http://dx.doi.org/10.5281/zenodo.13862`.

CONCLUSION

In this work we presented a novel data set containing crowd-sourced vocal imitations of audio concepts and identifications of those imitations by other humans. This data set will help the research community understand which audio concepts can be effectively communicated with vocal imitation and what the characteristics of these audio concepts are. It is the authors' hope that by studying this data set, the research community will learn how we can apply this natural form of communication to future audio software.

ACKNOWLEDGMENTS

This work was supported by NSF Grant Nos. IIS-1116384 and DGE-0824162.

REFERENCES

1. Blancas, D. S., and Janer, J. Sound retrieval from voice imitation queries in collaborative databases. In *Proc. of Audio Engineering Society 53rd Int'l Conf.* (2014).
2. Cartwright, M., and Pardo, B. Synthassist: Querying an audio synthesizer by vocal imitation. In *Conference on New Interfaces for Musical Expression* (2014).
3. Eitz, M., Hays, J., and Alexa, M. How do humans sketch objects? *ACM Trans. Graph. 31*, 4 (2012), 44.
4. Ekman, I., and Rinott, M. Using vocal sketching for designing sonic interactions. In *Proc. of the 8th ACM Conf. on Designing Interactive Systems* (2010).
5. Fritts, L. University of iowa musical instrument samples, 2012. `http://theremin.music.uiowa.edu/MIS.html`.
6. Ghias, A., Logan, J., Chamberlin, D., and Smith, B. C. Query by humming: musical information retrieval in an audio database. In *Proc. of the Third ACM Int'l Conference on Multimedia* (1995).
7. Gillet, O., and Richard, G. Drum loops retrieval from spoken queries. *Journal of Intelligent Information Systems 24*, 2-3 (2005), 159–177.
8. Janer Mestres, J. *Singing-driven interfaces for sound synthesizers*. PhD thesis, Universitat Pompeu Fabra, 2008.
9. Lemaitre, G., Dessein, A., Susini, P., and Aura, K. Vocal imitations and the identification of sound events. *Ecological Psychology 23*, 4 (2011), 267–307.
10. Lemaitre, G., and Rocchesso, D. On the effectiveness of vocal imitations and verbal descriptions of sounds. *The Journal of the Acoustical Society of America 135*, 2 (2014), 862–873.
11. Marcell, M. M., Borella, D., Greene, M., Kerr, E., and Rogers, S. Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology 22*, 6 (2000), 830–864.
12. Smaragdis, P., and Mysore, G. J. Separation by "humming": User-guided sound extraction from monophonic mixtures. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2009).
13. Stowell, D. *Making music through real-time voice timbre analysis: machine learning and timbral control*. PhD thesis, Queen Mary University of London, 2010.
14. Sundaram, S., and Narayanan, S. Vector-based representation and clustering of audio using onomatopoeia words. In *Proc. of AAAI* (2006).