Game With A Purpose To Collect Home Audio Data

Bryan Kim
Carnegie Mellon University
Pittsburgh, PA 15213, USA
bryank1@andrew.cmu.edu

Chenchen Tan
Carnegie Mellon University
Pittsburgh, PA 15213, USA
chenchet@andrew.cmu.edu

Yi Cheng
Carnegie Mellon University
Pittsburgh, PA 15213, USA
ycheng3@andrew.cmu.edu

Shuo Wang
Carnegie Mellon University
Pittsburgh, PA 15213, USA
shuow1@andrew.cmu.edu

Zixuan Li
Carnegie Mellon University
Pittsburgh, PA 15213, USA
zixuanli@andrew.cmu.edu

Yifeng Shi
Carnegie Mellon University
Pittsburgh, PA 15213, USA
yifengs@andrew.cmu.edu

Ruoxi Li
Carnegie Mellon University
Pittsburgh, PA 15213, USA
ruoxil@andrew.cmu.edu

Jessica Hammer
Carnegie Mellon University
Pittsburgh, PA 15213, USA
hammerj@andrew.cmu.edu

CHI PLAY EA '19, October 22–25, 2019, Barcelona, Spain
ACM 978-1-4503-6871-1/19/10.
https://doi.org/10.1145/3341215.3356270

Abstract
Human computational games, also known as GWAP (games with a purpose), have a history of generating a large amount of annotated visual data. In this work, we explore extending GWAP design to generate large datasets of annotated audio data collected in the home environment. Collecting data in the home presents unique challenges around privacy and comfort; processing audio data requires segmentation as well as labeling and validation. However, the home setting also affords unique opportunities as a gameplay enhancer. This work presents three prototypes, each targeted to a different phase of audio metadata generation, that use the home setting in different ways.

Author Keywords
Game With a Purpose; Human Computational Games; Audio Data; Game Design

CCS Concepts
•Human-centered computing → **Human computer interaction (HCI); Collaborative and social computing systems and tools;** *Computer supported cooperative work;*

Introduction
In this project, we explore new game-based methods for producing large annotated datasets of audio collected at home. Having such a dataset available to train home-based

Procedural phases of collecting audio data
1. Capture: Players capture sound from their environment.
2. Segmentation: Players mark the beginning and end of a sound of interest within an audio clip.
3. Labeling: Players label a sound through text or tags. Also known as annotation phase.
4. Validation: The correctness of the label is validated, typically through consensus from two or more players.
5. Reward: Players receive proper game rewards to their activity to keep players inside the game loop, increasing productivity and accuracy of data.

Table 1: Five identified phases for engaging with audio data in a GWAP

AI technologies has the potential to unlock new designs.

Creating such a dataset is a challenge on several fronts. First, sound event detection is relatively less mature than speech detection or the production of annotated visual datasets [11, 3]. Second, for most people, identifying an audio source from sound alone is a difficult challenge. For example, consider a knocking sound; are they knocking on a door or a chair? Is it wood or metal? Or is the sound being faked using foleying techniques that make it sound more authentic than the real thing would be? Third, there are substantial privacy and ethical concerns related to collecting audio data from the home, as suggested by recent concerns about Alexa and other voice agents [13]. Finally, users must find these activities *useful* and *acceptable* to do in the home, which means they must not violate the users' idea of what home is for.

At the same time, these challenges are themselves the source of design opportunities. We can break down the complexity of audio metadata generation into different stages, and design games that target each stage separately, to reduce the complexity of the challenge. This approach is what we call "suite of game" approach. In this work we share our design process, and show three early designs that take up the challenge of audio metadata games. We look forward to playing future versions of these games with users, and evaluating the quality of the data they produce. In the meantime, we hope that our design approach (creating a suite of related games, and designing for the affordances of home) can inspire others working in this space.

Background

Purpose of Audio Datasets

One major use for annotated audio datasets is to train and develop AI technologies. A publicly available, large-scale sound database with diverse sounds, annotated with strong labels, can accelerate sound-based AI research and development. Consider the case of well-being in the home. A sound-trained AI can detect sounds that mean *something is wrong* such as glass breaking, severe coughing, alarms, and more. Furthermore it can detect *something is missing*, or sounds that are conspicuous only when absent, such as the door opening and closing in the morning. Another use case of AI might be earplugs that an selectively filter out ambient noises common in the home environment to ensure a good night's rest, while still allowing through unusual or urgent noises, such as a baby crying or a window breaking. These approaches can be extended to address other types of problems.

Generating Audio Datasets

Despite ongoing advances in machine learning, it is currently not possible to automate the generation of large audio datasets [11, 3, 17]. Current audio dataset collection practices focus on some parts of the spectrum: automated home sound detection [8], classification of polyphonic sounds [11], and datasets suitable for machine learning [10, 12]. However, it is impossible for a machine to cover the whole spectrum which includes capturing audio, producing a huge amount of annotated audio data, and validating the accuracy of the taxonomy. Creating a usable audio dataset, therefore, means incorporating human expertise into the segmentation and labeling process; in turn, this requires validation, as humans may disagree, make mistakes, or deliberately introduce wrong information. Once humans are involved, the systems must also take into account methods for *keeping* them involved and engaged, such as rewarding them for participation [16].

Extracted Themes
Party Games
Can play with family and friends
Progression
Focuses on level design
Manipulating Audio
Controlling an existing audio as a gameplay
Moving Through Space
Physically move around to capture audio
Creating Virtual Object
Real world sound input creates virtual object
Physical Interaction
Physical action as a game mechanics
Competitive
Competition through Rewards and points
Gaining Other Skills
Educational gameplay
Player Investment
Take time to get emotionally involved
Relationship Building
Collaboration game
Player Contribution
Players generates content as a gameplay
Imagination
Players generates content as a gameplay
Suspense
Puts players into intense situation

Table 2: List of extracted themes from affinity diagram. Total of 13 themes from affinity diagram.

Games with a Purpose

Games with a Purpose (GWAP), also known as human computation games, leverage the wisdom of the crowd to accomplish tasks that are easy for humans but hard for computers. By making the task playful, the game motivates player participation. For example, *The ESP Game* reveals the same image to two players, and asks them to guess what the other person has written to describe it. If they agree, that word or phrase is then used to annotate the picture. Repeating the same image with other pairs of players, the computer eventually builds up a detailed label [18]. Other well-known GWAPs include *Artigo*, a variation of *the ESP Game* that asks players to annotate art piece, and *FoldIt*, which asks players to help scientists fold proteins [19, 4].

Some existing GWAPs support the creation of audio metadata. *TagATune* asks players to describe sounds and music; when multiple players agree on a description, it becomes a tag for the audio segment [7]. *TagATune* follows *The ESP Game* design, which generates strong labels. However, it lacks a segmentation phase since players take longer time to label audio while they can almost instantaneously label images [7]. Another existing audio GWAP, *Sonic Home* asks players to collect environmental audio, which is then used to train a sound recognition model for daily activities [9]. In *Sonic Home*, players own a virtual house and they can furnish the house with in-game items by collecting environmental sounds. Specifically, the system makes users evaluate label correctness, sound volume, and the presence of silent sound segments. *Sonic Home* creates high quality annotated audio data, but lacks in productivity of data creation.

Contexts of GWAP Play

Collecting audio data *in the home* is a unique challenge, because homes have different social norms and expectations than other social contexts. Most existing GWAPs are location-agnostic; they can be played anywhere that they player has access to a device. However, some GWAPs are context-specific. For example, *Urbanopoly* is a mobile, location-aware GWAP aimed at verifying, correcting and collecting data about "venues" in the urban environment [2]. The gameplay does not always take place *at* a specific venue that a player wants to assess, but it does has several mini-games that are only playable if the player has been to the place or is currently there (e.g. taking a picture of the venue). This requirement led the team to design interactions for a mobile device, such as creating mini-games that can be easily played on a small touch screen, and minimizing need for typing [2].

Using the home as an environment for recording creates unique design space. Platforms become an important tool to specify the location where players play the game. Platforms like PC, Virtual Reality headsets, TV consoles are ideal when designing home audio games, but these platforms should consider the approach and quality of collecting audio. Privacy and security are also issues since it requires recording one's private space.

Methods

We set out to identify new approaches to GWAP design that would a) accommodate the needs of working with audio data, and b) address the specificity of the home context. In order to do this, we used a structured ideation process. Our team divided into three groups. One group developed game concepts based on verbs [1], a second group built on different flow diagrams for how the five phases (Table 1) could produce a game loop, and a third ideated based on existing

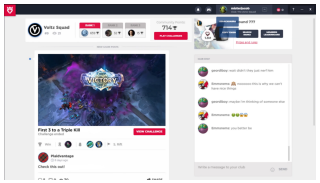

Figure 1: Capture prototype: Reference (IKON Challenge)

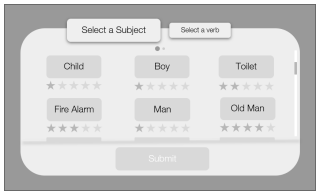

Figure 2: Capture prototype: Prompt selection (Streamer's screen)

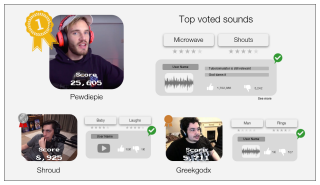

Figure 3: Capture prototype: Leader board

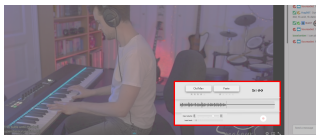

Figure 4: Capture prototype: Audio capture (Viewers' screen)

game genres. Taken together, the three groups generated over 250 game concepts. The team then used an affinity diagramming process to extract themes (see Table 2) from the game concepts.

In parallel with this work, we collaborated with audio data experts to create a set of metrics for evaluating our game concepts. Working with them, we identified that there would be different metrics for different phase of the process. Each procedural phases (Table 1) were evaluated: quality, accuracy, productivity, playability, and efficiency. While a game might be able to target many of these metrics, sometimes the metrics were in tension with one another. For example, a competitive game might increase the productivity of the labeling phase, but reduce quality in the capture phase. Making the game perfect for all of the phases would not be possible.

This realization led us to our final design approach: to create a suite of games that would operate on the same audio dataset. Each game could focus on a smaller number of stages of the metadata generation process, and target the pleasures of play toward a smaller set of metrics.

We returned to our set of themes with two key ideas in mind. First, we wanted to develop ideas that would target *specific* phases, instead of trying to do the end-to-end process in a single game. Second, we wanted to prioritize ideas that took advantage of the home context. To accomplish this, we created 20 game pitches for each of the 13 themes, building on our 250+ original ideas and amplifying them with our further insights. For each of these pitches, team members independently rated how well they were targeted toward each phase of the audio GWAP process. After discussion among team members, the group selected one concept per phase that took advantage of some aspect of the home context.

Capture: Exploiting Multiple Devices

One affordance of the home is that homes typically have multiple computing devices accessible to the player. In addition to phones that they carry with them, users may be able to access tablets, game consoles, laptop or desktop computers, home audio assistants, and more. We can take advantage of this affordance for data capture, where one device serves as a motivator for activity while another is used as a portable recorder.

Drawing inspiration from *IKON* (Figure 1), a gaming challenge platform for streamers and viewers, we have created an early prototype of *Sounds Like Duck*, a Twitch-enabled game played by streamers and viewers together. During downtimes in streamed gameplay, the Twitch streamer creates a prompt for their audience by selecting two words: a subject and a verb (see Figure 2). The subjects and verbs are dynamically provided by the system; for example, if a particular sound is underrepresented in the database, then related subjects and verbs will be more likely to appear as higher score for the streamer, which will represented in a leader board (Figure 3). Viewers then use their phones or recording device to capture audio from inside their house that responds to the prompt the streamer has chosen (see Figure 4). Their captured audio is uploaded to a moderated queue. Sounds that violate the rules of the game are discarded; sounds that pass the moderation check are forwarded to the streamer, who must try to guess what they are hearing and how it responds to the (slightly off-beat, not able to be taken literally) prompt. The streamer may continue to guess sounds as long as valid options remain in the queue, or they may select a new prompt for data capture.

This design meets a need for streamers. Streamers need help entertaining their audiences while they take small breaks, such as for a snack; while audience members go

Figure 5: Segmentation prototype: Reference (Secret Office Kissing Game)

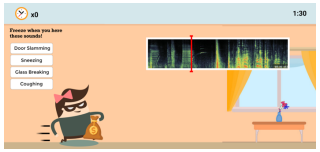

Figure 6: Segmentation prototype: Main gameplay screen

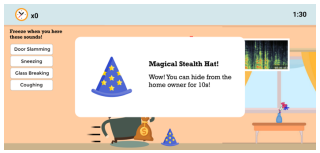

Figure 7: Segmentation prototype: Interaction feedback

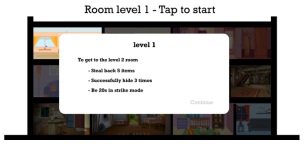

Figure 8: Segmentation prototype: Map selection

hunting for sounds, the streamer can relax. This design also meets audience members' desire to have the streamer pay attention to them [14, 5]. Having the streamer listen to your sound and react to it is potentially highly motivating. With 140 million unique Twitch viewers per month, this game has the potential to reach large audiences in a diverse range of households. For data capture, this is an enormous advantage.

By taking a "suite of games" approach, we do not have to incorporate validation into this game. This in turn means that we can create humorous prompts that players must respond to obliquely. We can accept player silliness or even attempts to game the system, because the goal of this game is to seed the system with a diverse set of audio data. In turn, this choice allows us to fit the game into the raucous culture of Twitch [6].

Segmentation: Soundscape Control

Homes are typically a space where people can control their environment, and use it to express their identity. This allows home residents to agree that game audio can be broadcast in the space. We take advantage of this to design a game for segmenting sounds, using a shared audio broadcast to all local participants along with a mobile game interface for each participant.

Drawing inspiration from *Secret Office Kissing Game* (Figure 5), a one-button game where players must kiss their office crush without being caught by the boss, the team created mobile game, *Sally Stealth*. The eponymous game protagonist is a burglar. The player controls her movement; they hold down the screen to move and release it to stop. The player's goal is to have Sally move only when there are distracting sounds. If she moves when there is silence, then she will be caught because of the noise she makes. The player receives two cues to help them decide when Sally should move. First, they can hear an audio file played out loud in the home, for all participants to hear. Second, they can see a sound spectrum visualization on their own interface that helps them decide when to move (Figure 6). The object of the game is to keep Sally from being caught for as long as possible; she will accumulate treasures periodically as long as she remains free. A good streak of stealth movement will grant Sally a power-ups, such as invisibility hat (Figure 7). The player can advance to the next level with more complex sounds to deal with (Figure 8).

By using a one-button design, this game is simple and accessible. Like one-button games in the "endless runner" genre, the game can be played with audio files of arbitrary length. The only limiting factor is the player's ability to avoid making a mistake (moving when there is silence). Because the game is multiplayer, the player is motivated to move when it is safe to do so, so that their opponent doesn't get ahead of them. Trash talk between players in the same space can enhance this aspect of the experience. However, this game could also be adapted to single-player use, for example by adding an AI opponent.

By taking a "suite of games" approach, we can have the player focused on identifying the beginning and end of sound events. We can capture a set of time stamps and validate their accuracy across multiple players working with the same file, so that we can identify when particular sounds start and end. However, we do not have to ask the player to interrupt their segmentation gameplay to label what the sounds are. This can be addressed with other prototypes working from the partly annotated dataset.

Validation: Party Time!

The home is a place where people nurture and maintain their social relationships. Shared cooperative gameplay, whether in the form of multiplayer digital games or board

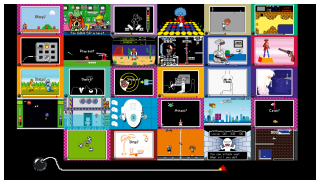

Figure 9: Validation prototype: Reference (Warioware)

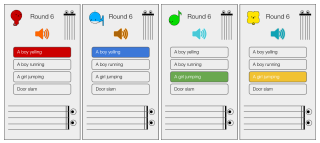

Figure 10: Validation prototype: Harmony gameplay screen

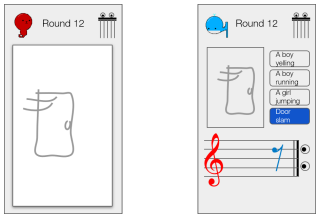

Figure 11: Validation prototype: Catch Mind gameplay screen

game nights, are one important way that homes are used socially. We can leverage the presence of multiple people seeking to play together in the design of our audio GWAP for the home. This aspect of home is a particularly good match for labeling and validation games, since most validation processes depend on agreement across multiple users.

Drawing inspiration from *Warioware* (Figure 9), a fast paced mini-game compilation game, the team came up with a four player local party game played on mobile devices, called *Hear Here!*. Like *Warioware*, *Hear Here* is a fast-paced game that contains multiple rounds of mini-games. Each mini-game uses different playful techniques to elicit tags from players, or to establish consensus about the validity of existing tags. For example, in *Harmony*, all four players listen to a sound and have five seconds to select a tag (see Figure 10); if they all select the same tag, they have won the mini-game. In *Catch Mind*, one player listens to an audio clip and selects a tag that describes it only through hand-drawn image. Other players must try to select the same tag with given image (see Figure 11).

Hear Here allows the integration of physical and digital game mechanics and leverages players' existing social connections; prior work suggests that multiplayer GWAP are motivating and may increase accuracy [15]. With co-present players there is, of course, the risk of collusion, particularly in the service of humor. Our design addresses this issue by interspersing dummy rounds with known correct answers to detect whether the team is spamming consensus. Data from each group can also be compared to data from other groups to increase validity.

Our "suite of games" approach appears here partially. Not only do we assume that *Hear Here* builds on an existing database of diverse, segmented sounds, but we also can use different mini-games within *Hear Here* to target different tasks. By strategically presenting limited options for the next mini-game, groups can be directed to participate in more labeling or more validation, as the needs of the entire dataset change. *Hear Here* also becomes highly scalable, as adding a new mini-game to the existing structure can keep players engaged.

Conclusion

We focused on finding a novel design space by understanding how GWAP/data collection games could be extended to address generating audio metadata based on sounds collected in the home. We analyzed existing games and evaluated the current state of the field. Using this information, we went through a discovery and ideation process, proposed "suite of game" approach to GWAP game design, and created our prototype designs. In this work we describe three prototypes: a data collection game that exploits having multiple devices available, a segmentation game that uses locally broadcast audio, and a labeling-validation game that draws on the home as a social space. In our future work, we will test these prototypes with users and evaluate how well they accomplish their goals. We will also evaluate the advantages and challenges of our "suite of games" approach.

Acknowledgments

Many thanks to our colleagues Kevin Bergen, Janine Louie, Jesse Song, Dustin Stephan, and Xuejun Wang. We thank Bosch Corporate Research and Philips Health for their generous support of our work. In particular, we are grateful to Pete Hill and Ji Eun Kim for their feedback and advice on engaging with this topic.

REFERENCES

1. Anna Anthropy and Naomi Clark. 2014. *A game design vocabulary: Exploring the foundational principles*

behind good game design. Pearson Education.

2. Irene Celino, Dario Cerizza, Simone Contessa, Marta Corubolo, Daniele DellAglio, Emanuele Della Valle, and Stefano Fumeo. 2012. Urbanopoly–A Social and Location-Based Game with a Purpose to Crowdsource Your Urban Data. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 910–913.
3. S. Chu, S. Narayanan, and C. . J. Kuo. 2009. Environmental Sound Recognition With Time—Frequency Audio Features. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 6 (Aug 2009), 1142–1158. DOI: http://dx.doi.org/10.1109/TASL.2009.2017438
4. Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and others. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756.
5. Steven Drucker, Li-wei He, Michael Cohen, Curtis Wong, and Anoop Gupta. 2002. Spectator games: A new entertainment modality of networked multiplayer games. *Microsoft Research* (2002).
6. William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1315–1324.
7. Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. 2007. TagATune: A Game for Music and Sound Annotation.. In *ISMIR*, Vol. 3. 2.
8. Neil Marten. 2017. Home automation sound detection and positioning. (Aug. 8 2017). US Patent App. 09/729,989.
9. Akira Masuda, Kun Zhang, and Takuya Maekawa. 2016. Sonic home: environmental sound collection game for human activity recognition. *Journal of Information Processing* 24, 2 (2016), 203–210.
10. Annamaria Mesaros, Toni Heittola, Emmanouil Benetos, Peter Foster, Mathieu Lagrange, Tuomas Virtanen, and Mark D Plumbley. 2018. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26, 2 (2018), 379–393.
11. Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. Metrics for polyphonic sound event detection. *Applied Sciences* 6, 6 (2016), 162.
12. A. Mesaros, T. Heittola, and T. Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*. 1128–1132. DOI: http://dx.doi.org/10.1109/EUSIPCO.2016.7760424
13. Kate O'Flaherty. 2019. Amazon Staff Are Listening To Alexa Conversations – Here's What To Do. (2019). https://www.forbes.com/sites/kateoflahertyuk/2019/04/12/amazon-staff-are-listening-to-alexa-conversations-heres-what-to-do/ .
14. Joseph Seering, Saiph Savage, Michael Eagle, Joshua Churchin, Rachel Moeller, Jeffrey P Bigham, and Jessica Hammer. 2017. Audience Participation Games: Blurring the Line Between Player and Spectator. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, 429–440.

15. Kristin Siu, Matthew Guzdial, and Mark O Riedl. 2017. Evaluating singleplayer and multiplayer in human computation games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. ACM, 34.

16. Kristin Siu and Mark O Riedl. 2016. Reward systems in human computation games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. ACM, 266–275.

17. MWW Van Grootel, Tjeerd C Andringa, and JD Krijnders. 2009. DARES-G1: Database of annotated real-world everyday sounds. In *Proceedings of the NAG/DAGA International Conference on Acoustics*.

18. Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

19. Christoph Wieser, François Bry, Alexandre Bérard, and Richard Lagrange. 2013. ARTigo: building an artwork search engine with games and higher-order latent semantic analysis. In *First AAAI Conference on Human Computation and Crowdsourcing*.