TAGATUNE: A GAME FOR MUSIC AND SOUND ANNOTATION

Edith L. M. Law, Luis von Ahn, Roger B. Dannenberg, Mike Crawford
Carnegie Mellon University
School of Computer Science

ABSTRACT

Annotations of audio files can be used to search and index music and sound databases, provide data for system evaluation, and generate training data for machine learning. Unfortunately, the cost of obtaining a comprehensive set of annotations manually is high. One way to lower the cost of labeling is to create *games with a purpose* that people will voluntarily play, producing useful metadata as a by-product. TagATune is an audio-based online game that aims to extract descriptions of sounds and music from human players. This paper presents the rationale, design and preliminary results from a pilot study using a prototype of TagATune to label a subset of the FreeSound database.

1 INTRODUCTION

Human computation, the idea of channeling the collective human presence over networks to solve difficult AI problems, has had great success. One of the first realizations of this idea is an online game called ESP [10], later adopted as the Google image labeler, where two players collaborate to label images on the internet. The result is one of the largest existing databases of images with labeled content.

More than half of the nation now has access to the internet, and 42% of those with access play games online [6]. The so-called *games with a purpose* take advantage of this burgeoning human interest in online games to solve important technological problems. This paper presents the rationale and iterative design of a new game called *TagATune*, which elicits from human players descriptions of sounds and music (collectively referred to as tunes).

2 RATIONALE

TagATune is a human computation tool that is capable of gathering perceptually meaningful descriptions for audio data that are agreed upon by multiple players. Such data are useful for several purposes described below.

Large audio databases have become invaluable resources for listeners, sound designers and composers. Current audio retrieval systems are primarily text-based, relying on accurate and comprehensive annotation of the data. However, keywords that describe a particular audio file are often subjective, based on one person's opinion [11].

TagATune has the potential to produce better labels at lower cost because the labor is essentially free and the validity of each label is confirmed by many players.

Researchers have long been interested in what makes a tune *warm* or *cold*, *scary* or *pleasant*, *bright* or *dull*, *happy* or *sad* and to what extent. For example, ongoing studies attempt to relate the perceptions [2, 7] of sounds to their acoustical and physical properties. A better understanding of auditory perception can yield important insights for retrieval tasks and auditory-enhanced interfaces. Part of the labeled data collected by TagATune is based on shared perception of sounds and are useful towards this kind of psychoacoustic or phenomenological research.

Another application is creating CAPTCHAs for the visually impaired. As a line of defense against bots, many websites now require humans to pass a visual CAPTCHA [9], which usually involves reading distorted characters against cluttered backgrounds. Common image- and text-based CAPTCHAs are inaccessible to the visually impaired. In addition, audio CAPTCHAs is a much needed alternative to visual CAPTCHAs, which have been successfully broken by various algorithms [4, 5]. Audio CAPTCHAs based on recorded speech exist, but as automatic speech recognition systems improve, labeled audio may offer more effective alternatives.

TagATune differs from two recently developed music annotation games, MajorMiner [3] and The Listen Game [8], in several ways. First, in TagATune, players are immersed in an audio environment that is not limited to music, but also a variety of sound clips, which are essential for the production of usable sound CAPTCHAs. Second, instead of playing offline against a database [3] or simultaneously against multiple players [8], players of TagATune are paired with a partner with whom they must collaborate to label a given tune. This enhanced level of rapport shown to be critical in the popularity of the ESP game. Finally, in addition to tags, TagATune collects *comparative information* about sounds and music, which allows for more discriminative audio search.

3 TAGATUNE: A PROTOTYPE

A prototype of TagATune is built in order to experiment with the game mechanics and to evaluate the usability and reception of the game. TagATune prototype is a simplified version of the game which uses sound clips only, lacking a blind-accessible interface, safeguards against cheating and inappropriate content, and other text matching

and verification capability such as spell checking and synonym matching. In addition, the prototype is a simulated two-player game. In it, players have the illusion that they are playing with human partners, whereas in reality, they are playing with a simulated player, or a *bot*, whose moves and timing come from recorded sequences of previous gameplay. The use of a bot is an indispensable component of the game, since it ensures that every player is paired when there is an odd number of them.

The audio used in the TagATune prototype are sound clips provided by the FreeSound Project. Within two months of its release, the project had "attracted over 2,300 members with over 1,650 sound contributions totaling more than 300 minutes of sound" (freesound.iua.upf.edu) and tens of thousands of downloads. Clips can be field recordings or synthesized audio containing a wide range of possible sounds, including music, rhythm, effects, ambience noise and speech. The collection of 23,084 sounds in the database vary between 0 and 4,522s in duration. For the prototype, 100 sound clips that are approximately 10 seconds long were randomly chosen. Limiting the amount of data makes it feasible to evaluate manually the quality and universality of the descriptions independent players submitted for each sound.

3.1 Design

One way to assure that the description of a sound is meaningful is to have two independent players agree on a particular description, a mechanism that has proven to work well labeling images in the ESP Game [10]. Similar to ESP, players of TagATune are not asked to describe the sound, but told to guess what their partners are thinking. TagATune is played by partners, who are paired randomly and anonymously from a pool of available players.

The partners are given three minutes to come up with agreed descriptions for as many sounds as possible. In each round, a sound is randomly selected from the database and presented to the partners. Controls are available for stopping or replaying the sound.

A main difference between TagATune and ESP is that what people hear in a sound is often more subjective, ambiguous, and imaginative than what they see in an image. The same sound can be perceived to come from very different objects. For example, a "hissing" sound can come from a cat, a tea boiler or a snake [1]. While natural sounds can be described in terms of their (imagined) source, music is more abstract, and even the descriptive terms for music are ambiguous. In order to elicit from players specific and well-defined annotation of the audio, TagATune displays a *category word* which specifies what kind of description the game is seeking. The eight category words used in the prototype are *Object*, *Place*, *Action*, *Color*, *Mood*, *Movie Genre*, *Opposite* (describe what the sound is *not*), and *Freebie* (unrestricted description).

Players enter guesses on what they think their partners hear. The players' guesses so far are displayed to remind them of their previous guesses, allowing them to strategically plan their next move. Players can also choose to pass on a difficult sound. Both partners must pass before the next sound is presented. Finally, the partners are not allowed any communication with each other, although they are notified of their partners' activities.

A description becomes an official tag that can be used for search when it is agreed upon by *enough* people, the threshold of which will depend on game statistics.

3.2 Lessons Learned

The effectiveness of TagATune to collect meaningful descriptions of sounds rides on two factors — that the game is fun enough to attract players on a regular basis, and that individuals agree, to a certain degree, on the descriptions of a given sound, even in an open-ended category such as color and movie genre. The prototype enables us to evaluate the extent to which the current game mechanics achieve these two goals, and what design changes are necessary to fill the gap.

Over the course of 5 days, 54 people signed up to play the prototype game. The system recorded sequences of gameplay, including every description and passed trial, and their associated timestamps. At the end of the evaluation period, players were asked for feedback concerning their impression of the prototype.

3.2.1 TagATune as a Game

Fun is a vague concept that is difficult to characterize, since many different elements in a game come together to create the specific experience. The design of this game focuses on three of these elements, namely its ability to create for each player a sense of competence, a pleasant and interesting sensory experience, and the opportunity to connect with their partners. Table 1 shows the average rating (on a five point scale) to questions related to the enjoyability of the game.

Did you find this game enjoyable?	3.3
Did you like playing with your partner?	3.5
Are you likely to play this game again?	3.5

Table 1. How enjoyable is TagATune? Average rating on the scale of 1 to 5, provided by 11 players who filled in the survey at the end of the game. In this scale, 1=Not at all, 3=Somewhat, 5=Very.

Having a sense of competence at the game is an important source of motivation for players to revisit the game. After having people play the prototype game, it was immediately apparent that finding specific descriptions for a given sound is a difficult task. The main problem is that the randomly chosen *category word* is not always appropriate for a given sound. In order to address this problem, the final game should have a mechanism for finding suitable category words for each sound. Solutions include allowing free form entry of descriptions without a category constraint, allowing one of the two partners to select the category word most appropriate for a given sound,

and eliminating *sound clip / word category* pairs for which players often pass.

TagATune must be able to provide a consistently interesting sensory experience for its players. In particular, it was evident that the uneven quality of the sounds used in the pilot study significantly hindered the experience of the game. One solution is to have partners rate the sound. For example, pairs of players can get extra points by giving the sound a similar rating. This rating information can also be useful to sound designers and music producers, who may be interested in first-impression judgements by the public. Unpleasant sounds can still be labeled as long as their occasional presentation does not degrade with the overall user experience. The same rating system is applicable to music.

3.2.2 TagATune as a Data Collection Tool

The most basic premise of any audio annotation games is that people generally perceive similar things in sounds and music. An encouraging observation from a pen-and-paper trial of the game is that there exist striking similarities among different people's descriptions of an arbitrary audio file. For example, when asked to describe the mood of a fireworks sound clip, more than one person put down words such as "exciting" and "happy." Likewise, for a mellow guitar solo, common colors such as "green" and "yellow" were used.

Table 2 shows some of the tags that were collected by the prototype. Most notably, in contrast to the description given by the author of the sound (first column), the descriptions by the players are usually simpler (single words or short phrases instead of sentences), perceptually more meaningful (high-level concepts conveyed by the sound instead of detailed description of how the sound is made), and more amenable to use as keywords in audio search.

Sound description	Category	Tags
Recorded sound of my mouth with re-verb and echo	place freebie object	forest, wood, jungle frog, cricket insect
field recording of the muehlkreisbahn — a small regional train which connects the northern part of upper austria with the town of linz	action object place	driving, braking car, truck, motorcycle on the bus, restaurant, ship, train station, factory
This drone is similar to "Dronetail 04" but a bit more bright.	mood	alarming, freaky, scary, dark, creepy
rubbed glass, granular synthesis	object	paper

Table 2. Descriptions of four different sounds: author versus players

A second observation is that the descriptions under different categories can help to reinforce each other. For example, the sounds of "frog," "cricket," and "insect" also remind the players of "forest," "wood," and "jungle." Likewise, for the sound that seems to be produced by motor vehicles, the places associated with that sound tend to be where cars can be heard.

Finally, the pilot study showed that some category words are meaningless for certain sounds. 36% of the time, the players opted to pass instead of providing a description for a sound.

4 TAGATUNE: FINAL DESIGN

The design of the final game benefited enormously from the lessons learned from this prototype. The final TagATune design consists of two different rounds - the annotation round and the comparison round. The annotation round is identical to the prototype rounds, except that one of the two players gets to choose the category word for the tune. In addition, each category word is attached with a score indicating its difficulty to describe with respect to the tune. This gives players the flexibility to choose an easier annotation task for fewer points, or a harder task for more points but a higher chance of never arriving at an agreement with their partners.

Figure 1. Preliminary TagATune Interface

The comparison round presents a tune, and asks the players to compare it to one or two other tunes of the same type based on a particular question. There are three kinds of questions. Preference questions ask "which of the two tunes do you prefer?" Similarity questions ask "which of the two tunes is more similar to this third tune?" Perception questions ask "which of the two tunes is more *X*" where X is a description previously provided by the players for one of the tunes. The data collected in the comparison round is more fine-grained because the pairwise comparisons can be converted into a ranking, which describes not only how the audio is perceived by the players, but to *what extent*.

Preference questions extract from each player a set of pairwise preferences for each tune. From the pairwise

preferences, a complete ranking can be calculated for individual players as well as for pools of players. This information, together with the data gathered from the similarity questions, will allow the game to selectively pick audio that players enjoy, thus motivating them to return to play the game. In addition, TagATune will provide the additional functionality for uploading one's own sounds and music into the game to be annotated by other players. Emerging artists can therefore use our game to test out the potential popularity of their creations. Finally, this data can be used to improve music recommendation and automatic playlist generation.

Perception questions are constrained similarity questions which ask players to compare the extent to which different sounds and music can be described by a particular tag. For example, instead of asking *if* a tune is exciting, the question permits us to ask *how exciting* it is, in comparison to other tunes. Audio search by similarity can benefit enormously from the availability of this type of similarity data.

5 CONCLUSION

In this paper, we have introduced TagATune, an audio-based *game with a purpose* for the annotation of sounds and music. TagATune has the potential to improve audio search, CAPTCHA accessibility, and the understanding of auditory perception.

The results from the prototype experiments are encouraging. Many of the labels of the sounds collected by the game are more descriptive and perceptually meaningful than the descriptions given by the authors of the sounds. 8 out of 11 of the participants who filled in the survey find the game (somewhat, quite, or very) enjoyable, and 10 out of 11 said that they are (somewhat, quite or very likely) to play the game again.

The fact that TagATune is a labeler for audio data rather than images raises a unique set of research questions. One of the most obvious differences is that sounds and music take time to listen to, whereas in ESP, players can almost instantaneously judge the content of an image. To minimize the risk of longer audio files being skipped by players, it may be advantageous to break them up into smaller segments, and serve those segments instead in the game. On the other hand, arbitrarily truncating a tune may render it incomprehensible. In the future, we expect to perform automatic segmentation or perhaps even make segmentation a task for a new game.

As an effective data collection tool, TagATune must first fulfill the requirement of being attractive to its players. The success of TagATune rides on the ability of the system to control the enjoyability and difficulty of the game by pre-determining for any given tune, whether it is enjoyable to listen to and which category words are most appropriate. While TagATune is built to elicit human intelligence to solve the audio segmentation and classification problem, ironically, the game itself must also tackle some of the same problems in order to be attractive to its players.

The full-scale release of the game will be launched along with the GWAP portal (www.gwap.com). The ultimate goal of TagATune is a reusable system that can help generate millions of labels for sounds and music in proprietary and online databases.

6 ACKNOWLEDGMENTS

We are especially grateful to Bram de Jong (FreeSound) for encouragement and providing access to the data necessary for this project. Luis von Ahn is partially supported by a MacArthur Foundation Fellowship.

7 REFERENCES

[1] P. Cano and M. Koppenberger. Automatic sound annotation. *IEEE Workshop on Machine Learning for Signal Processing*, pages 391–400, 2004.

[2] B. Giordano. An annotated bibliography. soundobject.org, 2001.

[3] M. Mandel and D. Ellis. A web-based game for collecting music metadata. In *8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[4] G. Mori and J. Malik. Recognizing objects in adversarial clutter: Breaking a visual captcha. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 134–141, 2003.

[5] G. Moy, N. Jones, C. Harkless, and R. Potter. Distortion estimation techniques in solving visual captchas. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 23–28, 2004.

[6] NTIA. A nation online: How americans are expanding their use of the internet. U.S. Department of Commerce. www.ntia.doc.gov, 2002.

[7] D. Rocchesso, R. Bresin, and M. Fernstrom. Sounding objects. *IEEE Multimedia*, 10(2):42–52, 2003.

[8] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. In *8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[9] L. von Ahn, M. Blum, and J. Langford. Telling humans and computers apart automatically: How lazy cryptographers do ai. *Communications of the ACM*, 47:57–60, feb 2004.

[10] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 319–326, 2004.

[11] E. Wold, T. Blum, and D. Keislar. Content-based classification, search and retrieval of audio. *IEEE Multimedia*, 3:27–36, 1996.