A GAME-BASED APPROACH FOR COLLECTING SEMANTIC ANNOTATIONS OF MUSIC

Douglas Turnbull[1]**, Ruoran Liu**[1]**, Luke Barrington**[2]**, Gert Lanckriet**[2]
Dept. of Computer Science and Engineering[1]
Dept. of Electrical and Computer Engineering[2]
University of California, San Diego

ABSTRACT

Games based on *human computation* are a valuable tool for collecting semantic information about images. We show how to transfer this idea into the music domain in order to collect high-quality semantic information about songs. We present *Listen Game*, a online, multiplayer game that measures the semantic relationship between music and words. In the normal mode, a player sees a list of semantically related words (e.g., instruments, emotions, usages, genres) and is asked to pick the best and worst word to describe a song. In the freestyle mode, a user is asked to suggest a new word that describes the music. Each player receives realtime feedback about the agreement amongst all players. We show that we can use the data collected during a two-week pilot study of Listen Game to learn a *supervised multiclass labeling* (SML) model. We show that this SML model can annotate a novel song with meaningful words and retrieve relevant songs from a database of audio content.

1 INTRODUCTION

Collecting high-quality, semantic annotations of music is a difficult and time-consuming task. Previous methods have included hand-labeling music [3, 9], conducting surveys [14, 6, 8] and text-mining web documents [7, 15]. Each approach has drawbacks: human annotation methods are time consuming, costly, and as such, do not scale when attempting to annotate large music collections. Information mined automatically from web documents is often inconsistent with a true semantic description of the audio content.

To collect large amounts of high quality annotation data at low cost, we propose using web-based games. von Ahn et. al. have created a suite of games (ESP Game [11], Peekaboom [13], Phetch [12]) for collecting semantic information about images. These 'games with a purpose' offer users an engaging platform for competition and collaboration while also collecting useful data about the image content. This data analysis technique is called *human computation* because it harnesses the collective intelligence of a large number of human participants to solve a task that can not easily be automated. Using a game-based approach, a population of users can solve large problems (i.e., labeling all the images on the Internet) using voluntary contributions from individuals (i.e., playing a game to label a single image.)

In this paper, we describe *Listen Game*, a multi-player, web-based game designed to collect associations between audio content and words. We show that this game is a powerful tool for collecting semantic music information by using the collected data to build a music information retrieval (MIR) application. In previous work [8], we presented a computer audition system that can automatically both *annotate* novel music with semantically meaningful words and *retrieve* relevant songs from a large database. Our system learns a supervised multi-class labeling (SML) model [1] by training on a set of audio content labeled with semantic annotations. We use the data collected from Listen Game to train our SML model. We then quantitatively evaluate the quality of the data by examining the accuracy of the SML model on the tasks of music annotation and retrieval.

2 COLLECTING MUSIC ANNOTATIONS

A supervised learning approach to semantic music annotation and retrieval requires a large corpus of song-word associations. Early work in music classification (by genre [9, 5], emotion [4], instrument [2]) either used music corpora hand-labeled by the authors or made use of existing song metadata. While hand-labeling generally results in high quality labels, it does not easily scale to hundreds of labels per song over thousands of songs. Companies such as Pandora [14] employ dozens of musical experts whose full-time job is to tag songs with a large vocabulary of musically relevant words but, unfortunately, have little incentive to make their data publicly available.

In [15], Whitman and Ellis collect a large number of web-documents and summarize their content using text-mining techniques. From web-documents associated with *artists*, they learned binary classifiers for musically relevant words by associating words in the documents with the artists' songs. In previous work [7], we mined expert music reviews associated with *songs* and demonstrated that we could learn a supervised multi-class labeling (SML) model over a large vocabulary of words. While web-mining is a more scalable approach than hand-labeling, we found that the data collected was of low quality since the extracted words did not necessarily provide a good description of a song. In general, when writing reviews of songs, albums or artists, authors do not make explicit decisions about the relevance of each single word. In addition, many reviews contain social, historical or opinionated information that is not related to the song's audio content [15].

A third approach uses surveys to collect semantic information about music. Moodlogic [6] customers annotate music using a standard survey containing questions about genre, instrumentation, emotional characteristics, etc. We used a similar approach [8] to collect the CAL500 data set of 500 songs, each of which has been annotated using a vocabulary of 174 words by a minimum of three people. Data collection took over 200 person-hours and resulted in approximately 300,000 individual word-song associations. Using a survey produced higher quality annotations than the web data but required that we pay test subjects for their time. Furthermore, surveys are tedious and time consuming. Despite financial motivation, test subjects quickly tire of lengthy surveys, resulting in inaccurate annotations.

Human computation games motivate players to generate reliable annotations based on incentives built into the game. In the ESP Game [11] for example, a pair of unacquainted players are partnered up and each shown the same image. Both players are asked to "type what your partner is thinking". Since they have no means of communicating, players invariably type words that have something to do with the common image they see. When two people *independently* suggest the same word to describe an image, the annotation is assumed to be reliable.

Human computation games also address the issue of collecting *lots* of data by turning annotation into an entertaining task. The ESP Game has gathered over 10 million image annotations. Games build a sense of community and loyalty in users and can be highly addictive. Statistics from the ESP Game highlight that some people played in multiple 40 hour per week spans. Since they require little maintenance and run 24 hours a day, games can constantly collect new information from multiple players. Developing human computation games for annotating music is a useful approach for collecting semantic information. We believe that this approach has the potential for large-scale success because people enjoy talking about, sharing, discovering, arguing about and listening to music.

3 LISTEN GAME

Image annotation often makes objective binary associations between an image and the objects ('sailboat'), scene information ('landscape'), and visual characteristics ('red') it represents. Our human computation game broaches the subjectivity inherent in many semantic labels that could be applied to music by allowing users to share their opinions, rather than be judged as correct or incorrect. Listen Game collects the strength of association between a word and a song, rather than an all-or-nothing binary label.

3.1 Description of Game Play

Listen Game (www.listengame.org) is a multi-player, online, music annotation game. Players listen to a common piece of music, select good and bad semantic labels and get realtime feedback on the selection of all other players. In a regular round (Figure 1.a), the game server selects a 15-second music clip (chosen from 250 popular western songs) and six words or phrases associated with a semantic category (e.g., instrumentation, usage, genre). The words are randomly chosen from a predefined 174-word vocabulary used in the CAL500 survey [8]. Each player's game client, loaded in a standard web browser, plays the clip and displays the category and the words in a randomly permuted order (to avoid order bias). The player then chooses both the best word to describe the clip and the worst word to describe the clip. Once the choices are committed, the game client displays instant feedback on the choices made by all other players. A player's score, S, is determined by the amount of agreement between the player's choices and the choices of all other players:

$$\begin{aligned} S &= 100 * (\text{fraction in agreement with best word}) \\ &\quad + 100 * (\text{fraction in agreement with worst word}). \end{aligned}$$

A player plays 7 regular rounds plus a *freestyle* round (Figure 1.b) where the game client plays a *preview* clip and displays a semantic category. The player is asked to enter a word or phrase that is an appropriate description of the preview clip. In the next regular round, the same music clip is played and the player's suggested word is presented as one of the possible annotations which other players may select as the best or worst word. Using these novel words from freestyle rounds, Listen Game can automatically grow the predefined vocabulary of musically relevant terms. Upon finishing 8 rounds, a game summary displays the player's score, the songs played, and various game statistics.

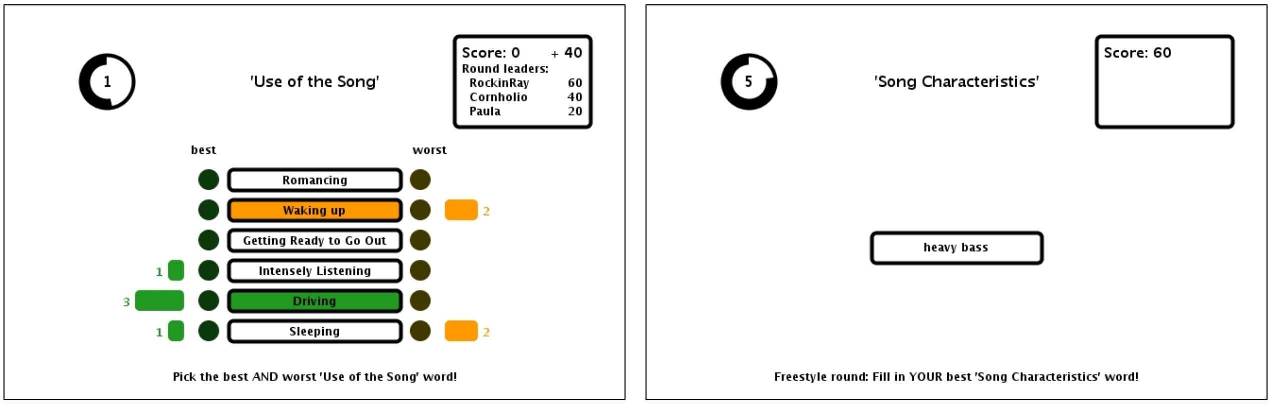
3.2 Quality of Data

While individual best/worst choices by players are binary, the aggregate song-word associations are not binary. One may interpret them as real-valued weights, proportional to the percentage of players who agree that a word does (or does not) describe a song. We calculate the semantic weight w as a function of the 'best' votes, 'worst' votes and potential votes (the number of times a song-word pair is presented to any player):

$$w = \begin{cases} 0, & \text{if \#(Best) - \#(Worst)} < 2 \\ w', & \text{otherwise} \end{cases}$$

$$w' = \max\left(0, \left[\frac{\#(\text{Best}) - \#(\text{Worst})}{\#(\text{Potential Votes})}\right]\right).$$

For a song-word pair to be reliable, we require that at least two people make the association in any given round. We would hope that with more data, we could raise the threshold for agreement significantly.



Pick the best AND worst 'Use of the Song' word!

Freestyle round: Fill in YOUR best 'Song Characteristics' word!

(a) Normal Round: players select best and worst words to describe the song (b) Freestyle Round: players enter their own word to describe the song

Figure 1: Screenshots of Listen Game

4 SUPERVISED MULTICLASS LABELING (SML)

We use the semantic song-word associations collected using Listen Game to train a SML model. The SML model was developed by Carneiro et al. [1] for the tasks of image annotation and retrieval. In [8], we showed how to use the SML model to learn a Gaussian mixture model (GMM) over an audio feature space for each word in a predefined vocabulary. We estimate these ‘word-level’ GMMs by combining ‘song-level’ GMMs (one trained on the feature vectors extracted from a single song) using the mixture hierarchies expectation-maximization algorithm (MH-EM) [10]. We also extended MH-EM to allow for real-valued *semantic weights*, rather than binary labels. While binary labels are quite natural for images where the majority of words are associated with objective semantic concepts, music is more subjective. For example, two listeners may not always agree that a song is representative of a certain genre or emotion. Listen Game directly reflects this notion by recording the votes of a large group of users on the best and worst words to describe a song. Using our *weighted* MH-EM algorithm, we learn GMMs that reflect the strength of the semantic associations between words and songs. We refer the reader to [8] for a full explanation of this system, as well as other details related to audio feature extraction and semantic representation.

5 EVALUATION OF LISTEN GAME DATA

Previously we used a survey to collect the CAL500 data set of semantic weights between 500 songs (by 500 unique artists) and 174 words [8]. The 174 words are part of an hierarchical vocabulary with six high-level semantic categories: genre, emotion, instrumentation, vocal characteristic, general song characteristics, and usage. We determine the ‘strength of association’ for these 87,000 word-song pairs by averaging the response of multiple individuals who annotated the song using a standard survey [8].

More recently, we conducted a two-week pilot study of Listen Game. We reduced the vocabulary to 120 words by eliminating ambiguous and less well known words. For the experiments reported in Section 5.2, we require that each word has been used to describe a minimum of five songs in the corpus, further reducing this vocabulary to 82 words. A randomly selected set of 250 songs from the CAL500 data set were used in the game. Players for Listen Game were recruited using emails to the authors’ friends and families, a mass email to a Music-IR list and word-of-mouth referrals.

During the two-week study, we collected the *Listen250* data set: 26,000 annotations (best and worst votes) of 250 songs using 120 words from 440 unique players. 20 players played more than 30 eight-round games and five (including one of the authors) played more than 100 games. In the freestyle round, players suggested 775 new words not from the original 120-word CAL500 vocabulary. Some standouts include subgenres (‘psychedelic’, ‘lounge’), usages (‘good for a hangover’,‘cooking’), adjectives (‘airy’, ‘fun loving’) and slang (‘agro’, ‘moshing’).

5.1 Qualitative Analysis

In Table 1, we present human- and machine-generated annotations of two songs. Human annotations are sum-

Table 1: “Musical MadLibs”. Annotations generated directly using semantic weights collected by Listen Game and automatically using the Listen250 SML model.

Norah Jones - Don’t Know Why

Generated using Listen Game data

This is **cool jazz**, **soul** song that is **mellow** and **positive**. It features **female vocal**, **piano**, **bass**, and **breathy**, **aggressive** vocals. It is a song **with a light beat** and **with a catchy feel** that you might like to listen to while **studying**.

Automatically Generated using SML model

This is **soft rock**, **jazz** song that is **mellow** and **sad**. It features **piano**, **synthesizer**, **ambient sounds**, and **monotone**, **breathy** vocals. It is a song **with a slow tempo** and **with low energy** that you might like to listen to while **studying**.

Rick James - Super Freak

Generated using Listen Game data

This is **R&By**, **funk** song that is **positive** and **cheerful**. It features **male vocal**, **piano**, **acoustic guitar**, and **high-pitched**, **aggressive** vocals. It is a song **with a catchy feel** and **with a changing energy level** that you might like listen to to while **at a party**.

Automatically Generated using SML model

This is **popy**, **R&B** song that is **not mellow** and **cheerful**. It features **sequencer**, **synthesizer**, **male vocal**, and **spoken**, **rapping** vocals. It is a song **that is very danceable** and **with a synthesized texture** that you might like to listen to while **at a party**.

marized by ranking words within each semantic category according to the semantic weights calculated by Listen Game. This results in labeling 'Don't know why' by Norah Jones as both 'Cool Jazz' and 'Soul' though these may not be the *best* genres to describe this song. 'Cool Jazz' was selected by multiple players in a round where there happened to be no truly relevant words. After many rounds, the semantic weight of words appearing in rounds with no clear choice would be reduced by votes for relevant words. We consider the Listen250 data set to be sparse, since there have only been on average two 'potential votes' for each of the 20,500 song-word pairs. The second set of annotations in Table 1 are automatically produced by the SML model trained using Listen250 data.

5.2 Quantitative Evaluation

We use per-word precision and recall (pwPrecision and pwRecall) metrics to measure annotation performance. We annotate each song with a fixed number of words, picked by the SML model. For each word w in our vocabulary, $|w_H|$ is the number of songs that have word w in the "ground truth" annotation, $|w_A|$ is the number of songs that our model annotates with word w, and $|w_C|$ is the number of "correct" words that have been used both in the ground truth annotation and by the model. pwRecall is $|w_C|/|w_H|$ and pwPrecision is $|w_C|/|w_A|$. The reported values in Table 2 are found by averaging these metrics over all the words in the vocabulary. While trivial models can easily maximize one of these measures (e.g., by labeling all songs with a certain word or, instead, none of them), achieving excellent precision and recall simultaneously requires a truly valid model.

Mean average precision (meanAP) and mean area under the receiver operating characteristic (ROC) curve (meanAROC) metrics measure retrieval performance. We calculate average precision (AP) by moving down the ranked list of retrieved test songs and averaging the precisions at every point where we correctly identify a new song. An ROC curve plots the true positive rate as a function of the false positive rate as we move down the ranked list of songs. The area under the ROC curve (AROC) is upper bounded by 1.0. Random guessing results in AROC of 0.5. MeanAP and meanAROC are found by averaging AP and AROC across all words in our vocabulary.

Table 2 compares the performance of three SML models: Listen250 trained using 225 songs annotated using Listen Game, CAL250 and CAL500 trained using 225 and 450 songs respectively, annotated using responses to surveys. We evaluate all models with the CAL500 data using 10-fold cross-validation. All differences are significant (paired t-test with $\alpha = 0.05$) with the exception of pwRecall and pwPrecision between CAL250 and CAL500. As expected, the models CAL250 and CAL500, trained on survey data produce better annotation and retrieval performance than the model Listen250 trained with sparser game data. The model CAL500, trained on more songs, achieves better retrieval performance than CAL250.

We would expect the performance of all models, but especially Listen250, to improve with both more training songs and more accurate estimates of the word-song relationships. For example, we noticed an improvement in meanAROC for Listen250 from 0.640 to 0.661 during the last 4 days of our two-week pilot study during which time we collected approximately 35% more data. By the end of our pilot study, we had shown each of our 20,500 word-song pairs only twice to a player.

Table 2: Model evaluation. The semantic information for CAL models was collected using a survey. The Listen model was trained on data collected by Listen Game. Each song is annotated with 8 words.

Model	Annotation		Retrieval	
	pwRecall	pwPrecision	meanAP	meanAROC
Random	0.092	0.058	0.188	0.501
Listen250	0.188	0.289	0.368	0.661
CAL250	0.215	0.333	0.410	0.701
CAL500	0.224	0.338	0.429	0.722

Acknowledgements

We thank D. Torres, S. Dubnov, O. Lang, and M. Yazdani for helpful comments. We also thank everyone who played and provided feedback on Listen Game. This work is supported by NSF IGERT DGE-0333451 and NSF grant DMS-MSPA 062540922.

6 REFERENCES

[1] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[2] S. Essid, G. Richard, and B. David. Inferring efficient hierarchical taxonomies for music information retrieval tasks: Application to musical instruments. *ISMIR*, 2005.

[3] M. Goto. AIST annotation for RWC music database. *ISMIR*, 2006.

[4] T. Li and M. Ogihara. Detecting emotion in music. *ISMIR*, 2003.

[5] M. F. McKinney and J. Breebaart. Features for audio and music classification. *ISMIR*, 2003.

[6] T. Sulzer. Moodlogic. http://www.moodlogic.com, 2007.

[7] D. Turnbull, L. Barrington, and G. Lanckriet. Modelling music and words using a multi-class naïve bayes approach. *ISMIR*, 2006.

[8] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic description using the CAL500 data set. *SIGIR*, 2007.

[9] G. Tzanetakis and P. R. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 7 2002.

[10] N. Vasconcelos. Image indexing with mixture hierarchies. *IEEE CVPR*, pages 3–10, 2001.

[11] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM CHI*, 2004.

[12] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In *ACM CHI Notes*, 2006.

[13] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A game for locating objects in images. In *ACM CHI*, 2006.

[14] T. Westergren. Music genome project. www.pandora.com, 2007.

[15] B. Whitman and D. Ellis. Automatic record reviews. *ISMIR*, 2004.