A WEB-BASED GAME FOR COLLECTING MUSIC METADATA

Michael I Mandel
Columbia University
LabROSA, Dept. Electrical Engineering
mim@ee.columbia.edu

Daniel P W Ellis
Columbia University
LabROSA, Dept. Electrical Engineering
dpwe@ee.columbia.edu

ABSTRACT

We have designed a web-based game to make collecting descriptions of musical excerpts fun, easy, useful, and objective. Participants describe 10 second clips of songs and score points when their descriptions match those of other participants. The rules were designed to encourage users to be thorough and the clip length was chosen to make judgments more objective and specific. Analysis of preliminary data shows that we are able to collect objective and specific descriptions of clips and that players tend to agree with one another.

1 INTRODUCTION

The easiest way for people to find music is by describing it with words. Whether this is finding music through a friend, browsing through a large catalog to a particular region of interest, or finding a specific song, verbal descriptions, although imperfect, generally suffice. While there are notable community efforts to verbally describe large corpora of music, these efforts cannot sufficiently cover new, obscure, or unknown music. It would be useful in these cases to have an automatic music description system. The most straightforward such system would base its descriptions wholly on the audio and to build it requires human generated descriptions on which to train computer models.

While writing on music abounds, in such forms as record reviews and music blogs, some of it describes aspects of the audio itself, while the rest describes the music's social and communal trappings. The boundary between the two is difficult to determine, and varies with the specificity of the description. Broad discussions of genre or style generally focus more on the social aspects of music, while specific descriptions of a passage focus more on the audio. Also, as the amount of music being described at once increases, heterogeneity becomes an increasingly serious problem; shorter clips are more likely to be homogeneous, making the link between language and music more definite. The descriptions most useful for a computer system describing musical audio, then, are those of short segments of music, which we call clips.

Thus in this project, we endeavor to collect ground truth about specific, objective aspects of music by asking humans to describe clips in the context of a web-based game [1]. Such a game entertains people while simultaneously collecting useful data. Not only is the data collected interesting, but the game itself is novel. Many decisions went into designing the game-play including the rules for scoring, the way in which clips are chosen for each player, and the ways in which players can observe each other.

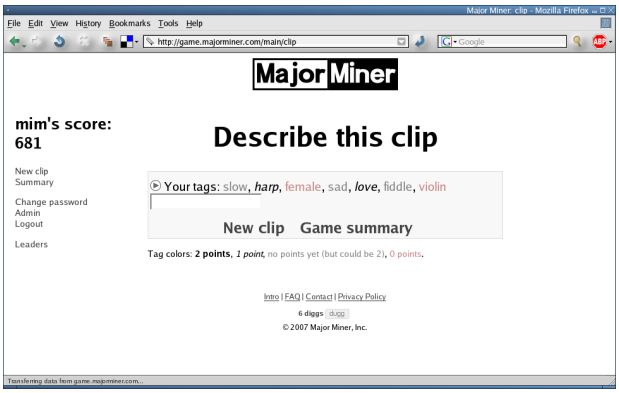

Figure 1. A screenshot of the game in progress. The user describes a 10 second clip of an unknown song. Italicized descriptions have scored 1 point, red descriptions 0 points, and gray descriptions have scored no points immediately, but will score 2 points when verified by another user.

Here is an example of a how a player, mim, experiences the game, see Figure 1 for a screenshot of the game in progress. First he requests a new clip to be tagged. This clip could be one that other players have seen before or one that is brand new, he does not know which he will receive. He listens to the clip and describes it with a few words: *slow*, *harp*, *female*, *sad*, *love*, *fiddle*, and *violin*. The words *harp* and *love* have been used earlier by exactly one other player, so each of these words scores mim one point. In addition, the player who first used each of those words scores two points. The words *female* and *violin* have already been used by at least two players, so they score mim zero points. The words *sad* and *fiddle* have not been used by anyone before, so they score no points immediately, but have the potential to score two points for mim at some later time should another player subsequently use one.

[1] The game is available to play at: http://game.majorminer.com

The player then goes to his game summary, an example of which can be seen in Figure 2. The summary shows both clips that he has recently seen and those that he has recently scored on, e.g. if another user has agreed with one of his tags. It also reveals the artist, album, and track names of each clip and allows the user to see another user's tags for each clip. In the figure, the other user has already scored two points for describing the above clip with *bass*, *guitar*, *female*, *folk*, *violin*, *love*, and *harp*, but has not scored any points yet for *acoustic*, *country*, *drums*, or *storms*. When he is done, mim logs out. The next time he logs in, the system informs him that three of his descriptions have been used by other players in the interim, scoring him six points while he was gone.

1.1 Previous work

A number of authors have explored the link between music and text, especially Whitman. In [7], Whitman and Ellis train a system for associating music with noun phrases and adjectives from a collection of reviews from the All Music Guide and Pitchfork Media. This work is based on the earlier work described in [8]. More recently, [5] have used a naive Bayes classifier to both annotate and retrieve music based on an association between the music and text. This work is inspired by similar work in the field of image retrieval, such as [1, 3].

In [4], the authors describe the "Musicseer" system for collecting ground truth about artist similarity, one aspect of which was a game. In this work, users chose which of a list of artists was most similar to a goal artist. The game was to link a starting artist to a goal artist with as short a chain of intermediate similar artists as possible. By performing this forced choice of the most similar artist from a list, triplets of relative similarity were collected, which could then be used to infer a full similarity matrix.

The "ESP Game" described in [6] asks pairs of players to describe the same image. Once both players agree on a word, they score a certain number of points and move on to the next image. The players attempt to agree on as many images as possible within a time limit. While previous data-collection games had maintained data integrity by forcing players to choose from pre-defined responses, this game first popularized the idea of allowing any response on the condition that it was verified.

2 GAME DESIGN

We designed the game with many goals in mind, informing each piece of the game. Our main goal was to encourage users to describe the music thoroughly. This goal shaped the design of the rules for scoring. Our second goal was for the game to be fun for both new and veteran users. Specifically, new users should be able to score points immediately, and veteran users should be rewarded with opportunities to score more points. This goal informed the method for introducing new clips into the game.

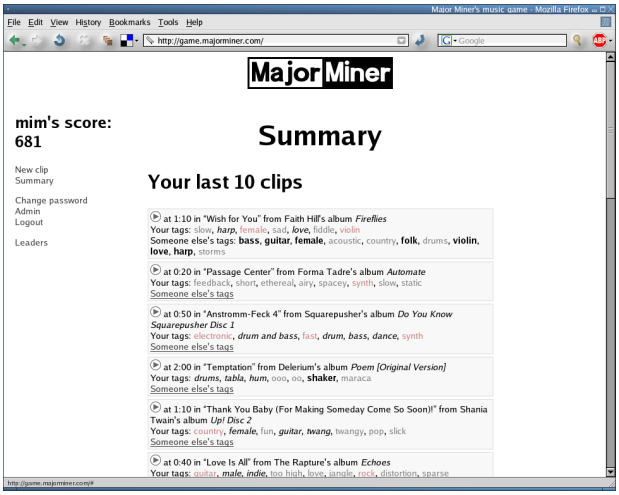

Figure 2. A screenshot of the player's game summary. The artist, album, track, and offset into that track are listed for clips the user has recently seen or scored on. The user can also see their own tags and those of another user.

Other lesser goals inform the architecture and implementation. The first of these is to avoid the problem of a cold start. Specifically, when the game is launched it should be easy for new and old users to score points and a single user should be able to play any time he or she wants, without the need for other online users. Another lesser goal is to avoid the possibility of cheating, collusion, or other manipulations of the scoring system or, worse, the data collected. Our final goal was to make the game accessible to as many people as possible, implementing it as a standard web page, without requiring any special plugins, installation, or setup.

While many games team one user with another, ours in a sense teams one user with all of the other users who have ever seen a particular clip. When a player is paired with a single cooperator, it is possible that the two players could be at very different skill levels or have different familiarities with the clip under consideration, detracting from the fun of both. It is also possible that when the game is first starting out, only one player might be online at a time. Although this problem can be solved by playing back recorded games, new clips can only be introduced through paired play. The non-paired format allows the most creative or expert users to cooperate with each other asynchronously, since obscure descriptions that the first uses will be available for the second. By introducing new clips systematically, the non-paired format also avoids some of the complications of starting the website, the so called "cold start." These benefits of non-paired games come at the price of vulnerability to asynchronous versions of the attacks that afflict paired games.

2.1 Scoring

The design of the game's scoring rules reflects our first goal, to encourage users to thoroughly describe clips; to be original, yet relevant. To foster relevance, users only score

points when other users agree with them. To encourage originality, users are given more points for being the first to use a particular description on a given clip. Originality is also encouraged by giving no points for a tag that two players have already agreed upon.

In the currently used form of the rules, the first player to use a particular tag on a clip scores two points when it is verified by a second player, who scores one point. Subsequent players do not score any points for repeating that tag. These point allocations (2, 1, 0) need not be fixed, but could be changed depending on participation and the rate at which new music is introduced into the game. The number of players who score points by verifying a tag could be increased to increase overall scoring and the point amounts allotted for each type of scoring could also be changed to alter the general style of play. We have found, however, that this simple scoring scheme sufficiently satisfies our goal of encouraging thoroughness. One concern with this system is that later users could be discouraged if all of the relevant descriptions have already been used by two other players. By carefully choosing when clips are shown to players, however, we can avoid this problem and use the tension created by the scoring rules to inspire originality without inducing frustration.

The game has high-score tables, listing the top 20 scoring users over the past day, the past week, and over all time. The principal payoff of the game may be the satisfaction of reaching some standing in these tables. Including the short-term-based tables gives even new users some chance to see their names in lights.

2.2 Picking clips

When a player requests a new clip to describe, we have the freedom to choose one that can help avoid the cold start effect and make the game fun for both new and experienced users. In order for a player to immediately score on a clip, another player must already have seen it. We therefore maintain a pool of clips that have been seen by at least one player and so are ready to be scored on. For new players, we draw clips from this pool to facilitate immediate scoring. For experienced players, we usually draw clips from this pool, but sometimes pick new clips in order to introduce new clips into the pool. While these clips do not allow a player to score immediately, they do offer the opportunity to be the first to use many tags, thus scoring more points when others agree.

In order for players to smoothly transition from being considered new to experienced, we define a parameter γ as the ratio of the number of clips that that user has seen to the total number of unique clips that have ever been shown by the system. The probability of choosing a brand new clip for a user is then γ and the probability of picking a clip that has already been seen is $1 - \gamma$. A brand new user, then, has $\gamma = 0$ and is therefore guaranteed to see a scorable clip. A user who has seen all of the old clips (which is only asymptotically possible under this scheme), has $\gamma = 1$ and thus is guaranteed to see a new clip.

When a clip has been seen by too many players, it will become difficult for subsequent players to score on. Thus, the game is more fun for players when they see clips that have not been seen many times before. Currently, previously-heard clips are re-used by choosing the ones that have been heard the fewest times, typically only once. However, if there are many new users playing, the seen-once pool can become depleted, pushing the system to select clips that have been seen twice or more.

2.3 Revealing labels

Part of the fun of the game is the communal interaction of seeing other users' descriptions. It also acclimatizes new users to the words that they have a better chance of scoring on. These other responses can only be revealed to a player after he or she has finished labeling a given clip, otherwise the integrity of the data and the scoring would be compromised. We would like a way to reveal other users' tags without giving away too much information or opening any security holes.

To solve this problem, we reveal the tags of the first player who has seen a clip, a decision that has many desirable consequences. This person is uniquely identified and remains the same regardless of how many subsequent players see the clip. It insures that the same tags are shows to every user who requests them for a clip and that repeated requests by the same user will be answered identically, even if others have tagged the clip between requests. The only person who sees a different person's labels is that first tagger, who instead sees the tags of the second tagger.

As described in the previous section, the first player to tag a particular clip is more likely to have more experience with the game. These descriptions are good examples for other players as an experienced user will generally be good at describing clips, will have thought about the process, will know what words others are likely to agree with, and will know what sort of formatting to use. Thus their tags can serve as a good example to others.

Also, in order to avoid introducing extra-musical context that might bias the user, we only reveal the name of the artist, album, and track after a clip is finished being labeled. This focuses the user much more on describing the sounds immediately present and less on a preconception of what an artist's music sounds like. It is also interesting to listen to a clip without knowing the artist and only afterward compare the sound to your preconceptions.

2.4 Strategy

There is a good strategy for effectively labeling a clip. When presented with a new clip, the player does not know which tags have already been applied to it. Trying one of the more popular tags will reveal how many times that tag has been used and thus the approximate number of times the clip has been seen. If the popular tag has never been used or has been used only once, the player can apply other popular tags and collect points relatively easily. If the tag has already been used twice, however, most other

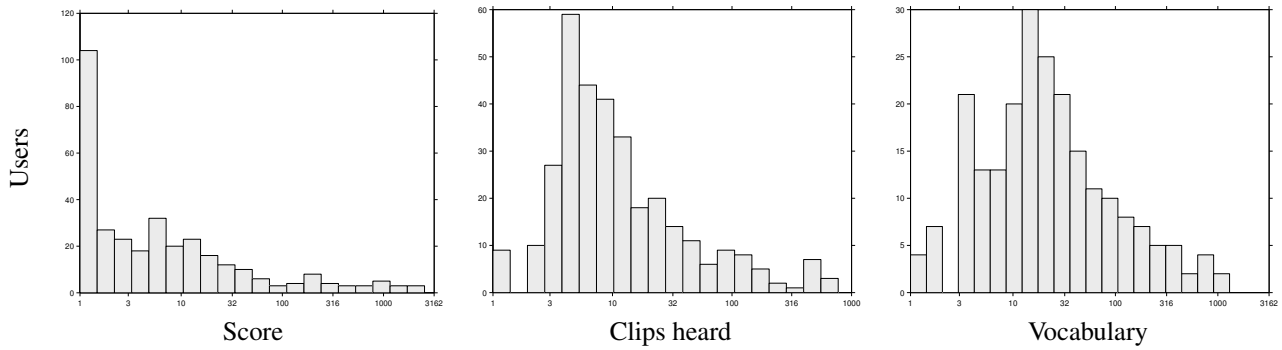

Figure 3. Histograms of user data

popular tags have also probably been used twice and aren't worth trying. The player must then decide whether to be more original or go on to another clip.

This clip-wise strategy leads to two overall strategies. The first is to be as thorough as possible, scoring two types of points. First, the user will score many one-pointers by agreeing with extant tags while preventing future listeners from scoring on them. Second, any new tags will be ready to score two points when used by subsequent listeners. The second strategy is to listen to as many clips as possible, trying to use popular tags on clips that haven't been seen before. This strategy is also viable, as it sets up the possibility of being the first to describe a clip a certain way, scoring two points.

While having a large number of clips with popular labels is worthwhile, in-depth analysis is more useful for us. To encourage breadth of description, we could add a cost to listening to clips or to posting tags. Similarly, we could post the high scores in terms of the number of points scored per clip listened to or the number of points scored per tag. These scoring changes encourage users to be more parsimonious with their tagging and listening.

There are a few possible exploits of the system that we have guarded against. The first is collusion between two users, or even the same person with two different usernames. Two users could theoretically communicate with each other their tags for particular clips and score on all of them. We thwart this attack by making it difficult for users to see a particular clip of their choosing and by adding a refractory period after a user has seen a clip when it cannot be seen again. Since users can only see their most recent clips, we also never have to refer to clips by an absolute identifier, only by relative positions in the recently seen and recently scored lists, making it more difficult to memorize which clips have been seen.

The other exploit is the repeated use of the same tag or tags on every clip, regardless of the music. This is easily detected and the offending account or accounts can be disabled. This sort of behavior can also be discouraged by adding a cost to picking a new clip or by incorporating efficiency into the high score ranking, both of which are diminished by tags unrelated to the music itself. To further prevent these attacks, we could limit the frequency with which a user may use a particular tag, for example forbidding the repetition of tags from the previous clip.

3 DATA ANALYSIS

Data from the game are collected in the form of triplets of clips, tags, and users. These triplets can be assembled into a large, binary, third order tensor, which can be analyzed as it is or by summing over some dimensions and considering the relationships within the remaining subset. Triplets are only included in this tensor when they have been verified by at least one other user, although stricter verification requirements could be used if desired.

To examine the relationship between tags and clips, we sum over the users. This creates a matrix with tags along one dimension, clips along the other, and entries containing the number of times any user has applied that tag to that clip. Some tags will be more popular than others, and some clips will have more labels than others, but the variation in their associations carries the most information. This matrix provides the basic information necessary to build models that associate words with music and vice versa. This analysis assumes that the more players who use a particular tag on a clip, the more relevant it is, which could be refined by weighting users differently.

By summing over tags, we can examine the number of tags each user applied to each clip that they saw. Some clips might be easier to describe than others, these would have a preponderance of tags from all users. Clips that elicit more descriptions on average from all users might be considered more describable or informative. An automatic predictor of this quality in new clips could be used to find interesting clips to browse through or to summarize a song. Some users might also be more verbose than others, these users would have a preponderance of tags for all clips. Furthermore, certain users might have more experience with certain types of music, making their descriptions more specific and possibly numerous. Hopefully, the players' familiarities will complement one another and we will be able to collect both expert and outsider descriptions of all of our music.

By summing over clips, we can see which words are used most frequently and by which users. This sense of the vocabulary of our music can be used to construct language models. It can also be used to detect equivalent

words, such as *hip hop* and *hip-hop*, which are generally used by different users, but occasionally are both used by a player attempting to score more points. The line between synonyms and equivalent words is not clear and more thought must be put into separating the two. Once the distinction is made, however, a dictionary of equivalent words could avoid this problem and the related problem in which *synth* does not score with *synthesizer*.

4 RESULTS

At the time of this paper's writing, the site had been live for about 1 month, in which 335 users had registered. A total of 2080 clips had been labeled, being seen by an average of 5.85 users each, and described with an average of 25.87 tags each, 4.14 of which had been verified. See Table 1 for some of the most frequently used descriptions.

We built the system as a web application using the Ruby on Rails framework. The user needs only a browser and the ability to play mp3s, although javascript and flash are helpful and improve the game playing experience. The page and the database are both served from the same Pentium III 733 MHz with 256MB of RAM. This rather slow server can sustain tens of simultaneous users.

The type of music present in the database affects the labels that are collected, our music comes from four sources. By genre, the first, and biggest source, contained electronic music, drum and bass, post-punk, brit pop, and indie rock. The second contained more indie rock and hip hop. The third contained pop, country, and more mainstream contemporary rock. And the last contained mostly jazz. Much of the music is from independent or more obscure bands, hopefully diminishing the biases that come with recognition of an artist or song.

Certain patterns are observable in the collected descriptions. As can be seen in Table 1, the most popular tags describe genre, instrumentation, and the gender of the singer, if there are vocals. People do use descriptive words, like soft, loud, quiet, fast, slow, and repetitive, but less frequently. Emotional words are even less popular, perhaps because they are difficult to verbalize in a way that others will likely agree with. There are hardly any words describing rhythm, except for an occasional *beat*.

Since any tag is allowed, users can and do use the names of artists they recognize. For example, *cure* has been verified 12 times, *bowie* 8 times, and *radiohead* 6 times. Only five of the clips verified as *bowie* were actually performed by David Bowie, however, the other three were performed by Gavin Friday, Suede, and Pulp. One need not take these descriptions literally, they could just be indicating a similarity between that particular clip from Suede's song "New Generation" and David Bowie's overall style, whatever that might be. These comparisons could indicate the artists to use as the anchors in an anchor space of artist descriptions [2]. Such a system would describe new artists by their musical relationships to well known artists.

Another valid and easily verified description of a clip is its lyrical content, if it is decipherable. Ten seconds are generally enough to include a line or two of lyrics, which the user then must distill down to a one or two word description. This has the added benefit of isolating some of the more important words from the lyrics, since players want to make their descriptions easy for others to match. Currently, *love* seems to be the most popular lyric word, with 14 verified uses.

Label	Verified	Used	Users
drums	690	2288	76
guitar	646	2263	113
male	536	1727	62
rock	488	1802	136
synth	380	1369	51
electronic	365	1299	87
pop	301	1157	100
bass	295	1232	63
female	273	998	67
dance	263	963	71
techno	176	664	65
piano	140	598	73
rap	133	555	86
electronica	119	503	47
hip hop	114	409	65
jazz	108	522	102
vocal	103	571	52
synthesizer	101	503	31
slow	89	453	57
80s	82	378	49
beat	71	424	47
voice	67	463	28
fast	65	318	44
vocals	60	471	34
british	59	301	43

Table 1. The 25 most popular tags. Three measures of tag popularity are provided: the number of clips on which it was verified by two players, the total number of times it was used, including unverified uses and uses by more than two players, and the number of players who have ever used it.

Due to the feedback inherent in the point scoring system and in seeing other users' descriptions, it is possible that the most popular words were determined by the first players. Were the first players using different words, those words might now be the most popular. It would not be difficult to divide the users between separate game worlds in order to test this hypothesis, although we have not yet attempted this. While it is true that the top tags have been quite stable, this could be also be because they are the most appropriate for the music in our database.

When looking at the tags *fast* and *slow*, one notices that some clips are tagged with one, some with the other, and some with neither. It seems that these descriptions are only used when the speed of a clip is notable. The "average" tempo of a song is unremarkable by definition. The absence of such opposing words could then be used in some sense to determine an "average" value for various

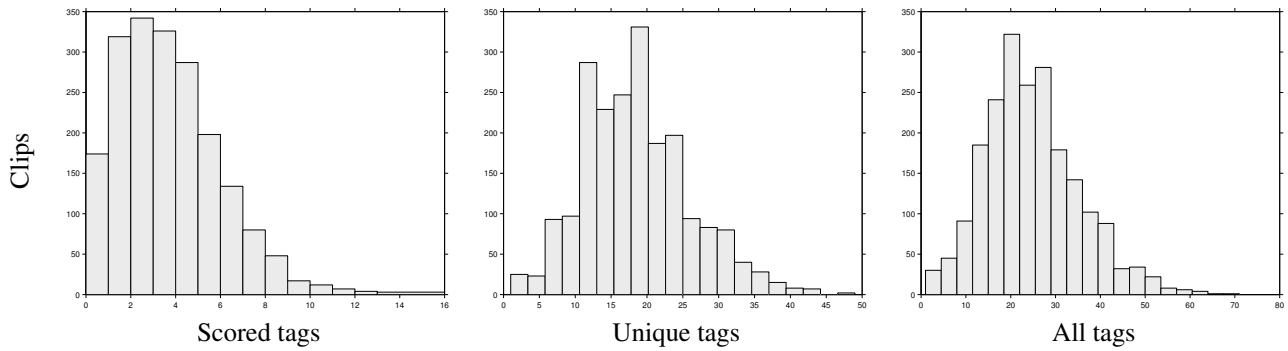

Figure 4. Histograms of clip data

musical characteristics and to identify a clip that is "average" in all of them, to be used as a baseline in music descriptions systems.

5 CONCLUSIONS

We have presented a game for collecting objective, specific descriptions of musical excerpts. The game is, in our estimation, fun, interesting, and thought provoking to play. Preliminary data collection has shown that it is useful for gathering relevant, specific data and that users agree on many characteristics of clips of music.

5.1 Future work

There is much that we would like to do with this data in the future: train models to automatically describe music, analyze the similarities between clips, between users, and between words, investigate ways to combine audio-based and word-based music similarity to help improve both, use automatic descriptions as features for further manipulation, investigate an anchor space built from the data collected here, use descriptions of clips to help determine the structure of songs, and so forth.

Acknowledgments

The authors would like to thank Marios Athineos and Graham Poliner for their help and Neeraj Kumar and Johanna Devaney for valuable conversations. This work was supported by the Fu Foundation School of Engineering and Applied Science via a Presidential Fellowship, and by the Columbia Academic Quality Fund, and by the National Science Foundation (NSF) under Grant No. IIS-0238301. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

6 REFERENCES

[1] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D.M. Blei, and M.I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3(6):1107–1135, 2003.

[2] A. Berenzweig, DPW Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *Proc Intl Conf on Multimedia and Expo (ICME)*, 2003.

[3] G. Carneiro and N. Vasconcelos. Formulating Semantic Image Annotation as a Supervised Learning Problem. In *Computer Vision and Pattern Recognition*, volume 2, pages 163–168, 2005.

[4] D. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proc Intl Symp Music Information Retrieval*, 2002.

[5] Douglas Turnbull, Luke Barrington, and Gert Lanckriet. Modeling music and words using a multi-class naive bayes approach. In *Proc Intl Symp Music Information Retrieval*, October 2006.

[6] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proc SIGCHI conference on Human factors in computing systems*, pages 319 – 326, 2004.

[7] B. Whitman and D. Ellis. Automatic record reviews. In *Proc Intl Symp Music Information Retrieval*, 2004.

[8] B. Whitman and R. Rifkin. Musical query-by-description as a multiclass learning problem. In *IEEE Workshop on Multimedia Signal Processing*, pages 153–156, 2002.