A Video Browsing Interface for Collecting Sound Labels using Human Computation in SoundsLike

Jorge M. A. Gomes, Teresa Chambel and Thibault Langlois
LaSIGE, Faculcy of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal

Keywords: Browsing, Audio, Music, Soundtrack, Video, Tagging, Human Computation, Gamification, Engagement.

Abstract: Increasingly, movies and videos are becoming accessible as enormous collections over the Internet, and in social media, demanding for new and more powerful ways to search and browse them, based on video content analysis and classification techniques. The lack of large sets of labelled data is one of the major obstacles for the Machine-Learning techniques that are used to build the relevant models. This paper describes and evaluates SoundsLike, an interactive web application that adopts a Human Computation approach through a Game With A Purpose to engage users in movie soundtrack browsing and labelling, while maintaining or improving the entertaining quality of the user experience.

1 INTRODUCTION

Nowadays, video and audio have a strong presence in human life, being a massive source of entertainment. The evolution of technology has enabled the fast expansion of media and social networks over the internet, giving rise to huge collections of videos and movies accessible over the internet and on interactive TV. These multimedia collections are tremendously vast and demand for new and more powerful search mechanisms that may benefit from video and audio content based analysis and classification techniques. Some researchers (Daniel and Chen, 2003) pointed the importance for the development of methods to extract interesting and meaningful features in video to effectively summarize and index them at the level of subtitles, audio and video image. Once this information is collected, we can use it for a better organization and access of the individual and collective video spaces.

The approach described in this paper was designed for the VIRUS (Video Information Retrieval Using Subtitles) project (Langlois et al., 2010) . This project aims to provide users the access to a database of movies and TV series, through a rich graphical interface. MovieClouds (Gil et al., 2012), an interactive web application was developed, adopting a tag cloud paradigm for search, and exploratory browsing of movies in different tracks or perspectives of their content (subtitles, audio events, audio mood, and felt emotions). The back-end is based on the analysis of: video image, and especially audio and subtitles, where most of the semantics is expressed. In the present paper, we focus on the analysis and classification of the audio track for the inherent challenge and potential benefit if addressed from a game perspective to involve the users.

In this context, our objective is to provide overviews of the audio and indexing mechanisms to access the video moments containing audio events (e.g. gun shots, telephone ringing, animal noises, shouting, etc.) and moods to users. To this end, we build statistical models for such events that rely on labelled data. Unfortunately these kinds of databases are rare. The building of our own dataset is a huge task requiring many hours of listening and manual classification - often coined the “Cold Start” problem. If we have models that perform relatively well, we could use the models to collect data similar to the audio events we want to detect, and human operators would be asked to label (or simply verify the classification assigned automatically) a reduced amount of data. This idea of bringing the human into the processing loop became popular with the rise of applications referred to as Games With A Purpose (GWAP), using Gamification and a Human Computation approach. These games have a goal which is collecting data from the interaction with human users. Gamification can be defined as the use of game design elements in non-game contexts (Deterding et al., 2011). These elements can be designed to augment and complement the entertaining qualities of movies, moti-

Gomes J., Chambel T. and Langlois T..
A Video Browsing Interface for Collecting Sound Labels using Human Computation in SoundsLike.
DOI: 10.5220/0004700804720481
In *Proceedings of the 9th International Conference on Computer Graphics Theory and Applications* (GRAPP-2014), pages 472-481
ISBN: 978-989-758-002-4

vating and supporting users to contribute to the content classification, combining utility and usability aspects (Khaled, 2011). Main properties to aim for include: Persuasion and Motivation, to induce and facilitate mass-collaboration, or crowd sourcing, in the audio labeling; Engagement, possibly leading to increased time on the task; Joy, Fun and improved user experience; Reward and Reputation inspired in incentive design.

This paper describes an application that innovates upon previous Human Computation applications both in terms of entertainment aspects and in terms of the definition of the game in order to stimulate the interest of the user in labelling audio in movies, allowing us to collect data that will help solve the cold start problem. In the scope of this paper, the terms "tag" and "label" represent non-hierarchical keywords or terms assigned to a piece of information with the purpose of describing its contents and help finding it through browsing and searching mechanisms.

In the following sections we present, discuss and evaluate design options of our approach. Section 2 makes a review of most relevant related work, section 3 shows the differentiating aspects of the SoundsLike approach from the related work. section 4 introduces the interaction design, while section 5 highlights the gamification process and design options, and section 6 presents the user evaluation. The paper ends in section 7 with conclusions and plans for future work.

2 RELATED WORK

The idea of creating entertainment applications in order to collect data from the user is not new. The approach, identified as Games With A Purpose (GWAP) and Human Computation have been first used in the ESP application (Von Ahn and Dabbish, 2004) and then refined in various systems whose objective is to collect data about image classification, music labelling, mood identification in music, etc. The main incentive for pursuing this approach is the fact that the state-of-the art in Machine Learning algorithms has not reached a level of performance that allows to solve the task automatically. Most Machine Learning techniques rely on the availability of labelled datasets.

Among these systems, the ESP game is a game where two players that cannot communicate are asked to label the same image and are rewarded when the labels used are the same. If the players consider the image too difficult they may pass. This approach is named output-agreement, because the reward is received when the labels (the output) used by the players correspond. An important aspect of the game is the use of taboo words that is a list of (up to six) words that are not allowed in the description. Taboo word lists are automatically generated and correspond the terms that have already been used to describe the image. The objective is to force players to use less obvious terms and provide a richer description of the image. Other GWAP systems using the *output-agreement* paradigm include Peekaboom (Von Ahn et al., 2006) a system to help locating objects in an image. In all cases, the game is played simultaneously by two players that cannot communicate.

Several GWAP are dedicated to gathering labels relative to audio data (mainly music). The originality of HerdIt (Barrington et al., 2009) is that being installed on the Facebook platforms, it benefits from the Social Netwoking effect. This game provides a multi-player experience, contrasting with the most common two-player experience. The MoodSwings game (Morton et al., 2010) aims to label songs according to their mood. It is a two-user game where users are asked to indicate the mood by clicking on a two dimensional plane showing the traditional valence-arousal axes. The originality of the approach is that tagging occurs in a continuous way while players interact. Another GWAP for music is TagaTune (Law et al., 2007), proposing a different approach. This two player game is based on *input agreement*. The two players, that cannot communicate, listen to the same sound clip and propose labels that describe it. Each player sees the labels proposed by the other player and the round finishes when players indicate if they are listening to the same piece of music. The agreement is therefore on the input and not on the output as in the previous ones. MajorMiner (Mandel and Ellis, 2008) is a single player game based on the output-agreement paradigm that asks users to assign labels to music clips. Unlike other games, MajorMiner is asynchronous.

Concerning the aspect of user-interfaces, the authors of the games previously mentioned opted for minimalistic features. Several aspects can be identified: 1) the item (music clip) has often no graphical representation (HerdIt, MajorMiner, MoodSwing), in this case the music clip is automatically played and the users cannot interfere (stop/pause/continue) nor can they see the progress of the listening. In the case of TagATune, more traditional media player controls are presented. 2) Labels are represented by simple words (no decoration) in MajorMiner and TagATune or by buttons or in the case of HerdIt, by floating bubbles. When labels are chosen by the user, they are entered using a text box and buttons (or bubbles) are used in case of predefined labels. 3) The use of

colours differ between interfaces. In TagATune only two colours are used (purple and white) with gradients in order to produce an aesthetic effect. The MoodSwing game does not use labels and presents a 2D plane coloured with a pink-blue-yellow gradient along the valence axis, user's choices are represented by red and yellow blobs. The bubbles in HerdIt are coloured but the colouring scheme seems to be random. Finally, the MajorMiner game uses very few colours, for differentiating levels of confidence of each label (italicised font is also used for this purpose). 4) Scores are represented either by a numerical value (MajorMiner, TagATune) or using a thermometer metaphor (HerdIt, ESP game).

3 THE SOUNDSLIKE APPROACH

Regarding the interface, we opted for a much richer interface compared with previous applications. The users are given a lot of contextual information about the audio excerpt they are asked to label. This is justified by the fact that arbitrary sound excerpts are likely to be more difficult to identify than music pieces. The objective is also to provide a joyful experience to the user and not only focus on the data collecting task. The context given to the user is three fold: 1) a temporal context through three timelines with different time scales; 2) the "real" visual context is given by the video that corresponds to the scene where the excerpt comes from; and 3) a context in given in the space of labelled excerpts by showing a dynamic force-directed graph with nodes corresponding to similar excerpts the user can listen to, before deciding on the labels. A different approach is used to represent labels. Most of the games opt between closed set of choices (HerdIt) and free labels entered from the keyboard (TagATune, MajorMiner). From the data-collecting task point of view, this choice has an impact on the total number of labels collected (120 for HerdIt up to 70,908 for TagATune). The TagATune case has a unique feature (it shares with the ESP game): the use of *taboo words* that force players to enlarge the set of labels.

Our approach is different, since we adopt a mixture of both approaches by suggesting labels that were previously assigned to similar sound clips and allowing the users to propose their own labels. This way, we hope to collect a set of labels that is as limited as possible, but that also allows the users to propose their own labels . Suggested labels are presented through a tag cloud. These tags are labels already assigned to excerpts that are in the neighbourhood of the current sample, according to a similarity measure on the audio data (it is out of the scope of this paper to describe the similarity measures, further details can be found in (Langlois and Marques, 2009)). This neighbourhood corresponds to the elements shown in the graph described previously.

Regarding the game aspects, we made some choices that differentiate our approach. First our application is oriented towards the classification of any kind of sound (and not only music). The audio samples presented to the user are four seconds long compared with 30 seconds or more for other games. There are several reasons for choosing shorter excerpts: 1) audio excerpts from video may contain any kind of sound and shorter samples are easier to identify; 2) it is more unlikely to have a large number of different sound events facilitating the objective labelling; 3) the hard part of the Information Retrieval process is the interpretation of low level features. It is where we need Human Computation. If we were able to reliably identify sounds in four seconds excerpts, it would be relatively easy to extend to longer samples by combining the output of four seconds chunks; and 4) it will allow us to distribute the collected data freely. Since previously cited games make use of copyrighted music material, it makes the diffusion of the labelled database to the research community more problematic. The second aspect is that in our game, people play asynchronously (we share this characteristic with the MajorMiner game). This means that it is not necessary to have several users on-line at the same time to start a game. Other benefits of this approach are: 1) cheating becomes harder. It has been observed in the case of the TagATune game that users communicate using labels (labels *yes* and *no* are among the most frequently used labels); and 2) users are rewarded while off-line. If a label the user proposed is re-used by others, he earns points - an incentive to return to the game.

4 INTERACTION DESIGN

SoundsLike is the part of MovieClouds (Gil et al., 2012) devoted to soundtrack interactive browsing and labeling. It integrates gaming elements to induce and support users to contribute to this labeling along with their movie navigation, as a form of GWAP. The interactive user interface was designed with the aim of suggesting this opportunity to play and contribute, by allowing listening to the audio in the context of the movie, by presenting similar audios, and suggesting tags. Next we present main design options for SoundsLike, exemplified by the labeling of a sound excerpt in the context of a movie navigation. In Figure 1, the

user is in the Movies Space View, where movies can be searched, overviewed and browsed, before one is selected to be explored in more detail, in the Movie View (Fig.1a-1b). Tag clouds were adopted in both views to represent summaries or overviews of the content in five tracks (subtitlesaudio events, soundtrack mood, and felt emotions), for the power, flexibility, engagement and fun usually associated with them.

After selecting the Back to the Future movie in Figure 1b) (top right), it plays in the Movie View in Figure 2a), with five timelines for the content tracks showed below the movie, and a selected overview tag cloud of the content presented on the right (audio events in the example), synchronized with the movie that is playing and thus with the video timeline, to emphasize when the events occur along the movie. From the timelines, users may select which time to watch in the video.

Playing SoundsLike. After pressing the SoundsLike logo (Fig.2a-b), a challenge appears: an audio excerpt is highlighted in three audio timelines with different zoom levels below the video, and represented, to the right, in the center of a non-oriented graph displaying similar excerpts, with a field for selection of suggested or insertion of new tags below, to classify the current audio excerpt, and hopefully earn more points. By presenting the surrounding neighbours, and allowing to listen to entire audio excerpts and watch them, SoundsLike was designed to support the identification of the current audio excerpts to be labelled.

Movie Soundtrack Timelines. Three timelines are presented below the video (Fig 2:b-d and Fig.3): the top one represents the entire soundtrack or video timeline for the current movie; the second one presents a zoomed-in timeline view with the level of detail chosen by the user, by dragging a marker on the soundtrack timeline; and the third one, also zoomed-in, presents a close-up of a chosen excerpt as an audio spectrogram. We designed the audio representation to be similar to the timelines already used in MovieClouds (Gil et al., 2012) for content tracks (figure 1a)), with three levels of detail. Audio excerpts are segments from the entire soundtrack, represented here as rectangles. In all the timelines (and the graph): the current excerpt to be classified is highlighted in blue, while grey represents excerpts not yet classified, green (and yellow) refer to excerpts classified before (and skipped) by this user.

The relation between each of the timelines is reinforced by colour matching of selections and frames: the colour of the marker in the soundtrack timeline (white) matches the colour of the zoomed-in timeline frame, and the colour of the selected audio excerpt (blue for the current audio) matches the colour of the spectrogram timeline frame. In addition, the current position in each of the timelines is presented by a vertical red line that moves synchronized along the timelines.

Audio Similarity Graph. To help the classification of the current audio excerpt, the excerpt is displayed and highlighted in the center of a connected graph, representing the similarity relations to most similar audio excerpts in the movie (Fig 2:b-d and Fig.4). Being based on a physical particles system, the nodes repel each other tending to spread the graph open, to show the nodes, and the graph may be dragged around with an elastic behaviour (Fig.4b). The similarity value between two excerpts is expressed as the Euclidean distance between audio histograms computed from extracted features (further details can be found in (Langlois and Marques, 2009)), and it is translated to the graph metaphor as screen distance between two nodes. The shorter the edge is, the most similar excerpts are. The nodes use the same colour mapping as the timelines, and the current excerpt has an additional coloured frame to reinforce if it was already classifying or skipped. The users can hover on each node to quickly listen to the audio excerpts. On click, the corresponding movie segment is played. For additional context, on double click the audio becomes the current audio excerpt to be classified. This is particularly useful, if the users identify this audio better that the one they were trying to identify, and had not classified it before, allowing to earn points faster; and to select similar audios after they have finished the current classification.

Synchronized Views. The views are synchronized in SoundsLike. Besides the synchronization of the timelines (section 4), the selection of audio excerpts in both timelines and graph synchronize. This is achieved by adopting, in both views: the same and updated colours for the excerpts; and by presenting, on over (to listen), a bright red coloured frame to temporarily highlight the location of the excerpts everywhere. In addition, when an audio excerpt is selected to play a frame is set for the excerpt in the graph node and timelines and this time also in the video (in a brownish tone of red) to reinforce which excerpt is playing (Fig.2c).

Labelling, Scoring and Moving On. To label the current audio excerpt – the actual goal – the users choose one or more textual labels, or tags, to describe

Figure 1: MovieClouds Movie View navigation: a) unique tag cloud for 5 content tracks; b) 5 separated tagclouds for each track. Some tags selected and highlighted in the movies where they appear (top).

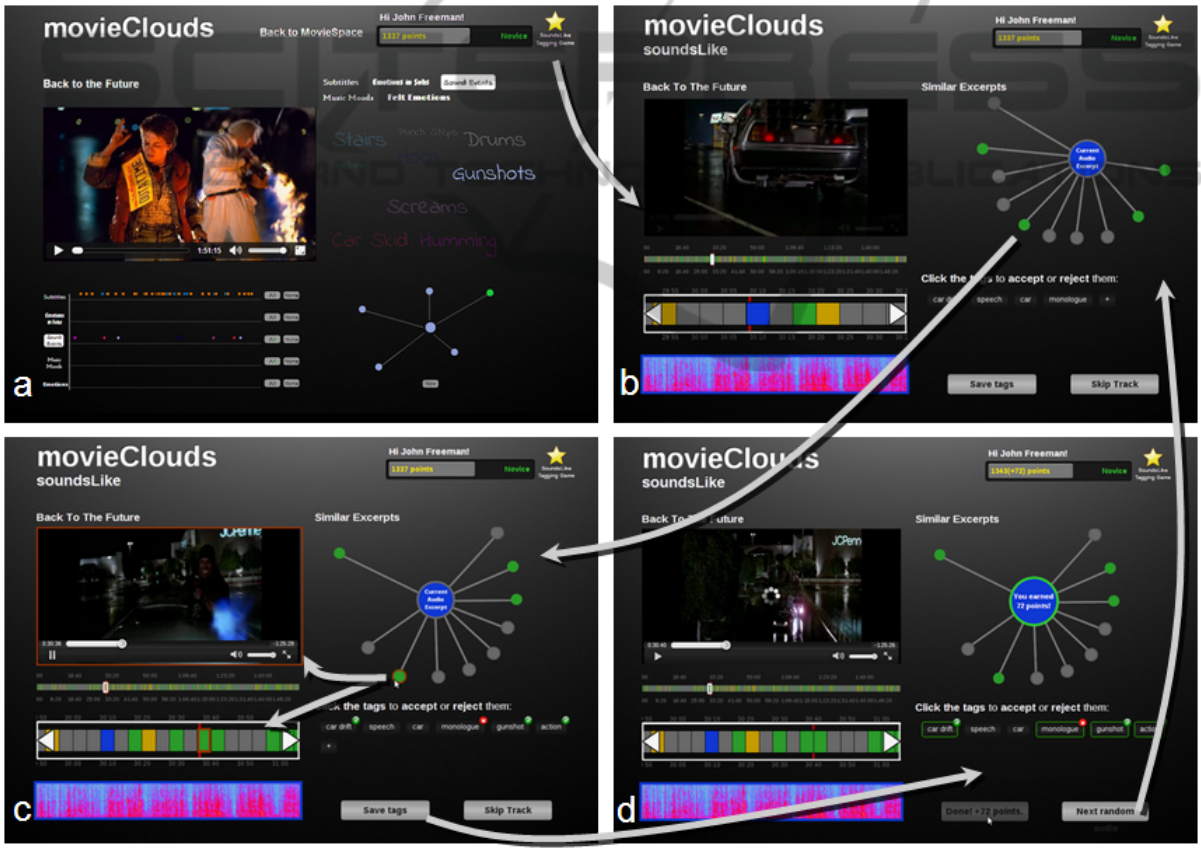

Figure 2: SoundsLike interaction: a) audio events track in MovieClouds Movie View; b) selection (click) of a neighbour excerpt; c) to be played as video, providing more context to identify the audio (excerpt highlighted with brownish red frame on the graph, timelines and video; d) saving tags, winning from labeling, and choosing to play again and tag one more.

it, in the region below the graph (Figs 2b-d and 5). Tags may be selected from a list of suggested tags, or introduced by the user (Figs. 2a and 2b) as a sequence of comma separated tags. Choosing a tag either to select, reject or ignore, is done through clicks till the desired option is on (Fig. 5). The choices are saved only when the save button is pressed (Fig.2b-c). At any time, the user can skip an excerpt and choose another one, or simply leave the game. The score will remain registered. When choices are submitted, the current excerpt changes colour to green in the similarity graph and the timelines, displaying

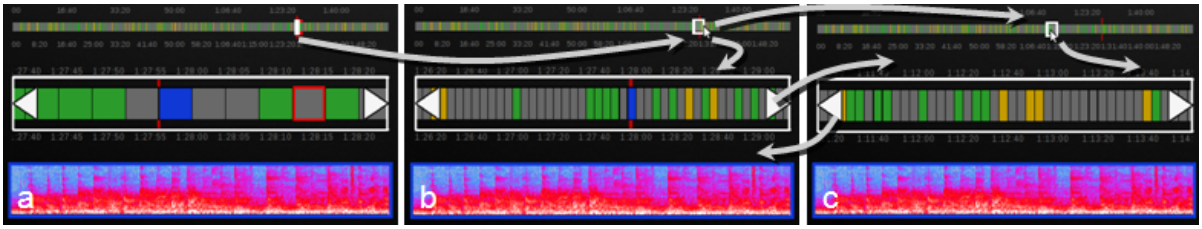

Figure 3: Timelines: a) video timeline, zoom timeline and current audio spectrogram; b) zooming out in the zoom timeline by dragging the marker open in the video timeline; c) dragging the marker right to make the zoom timeline move ahead in time. The arrows allow to do the same. The markers are always syncronized on the timelines.

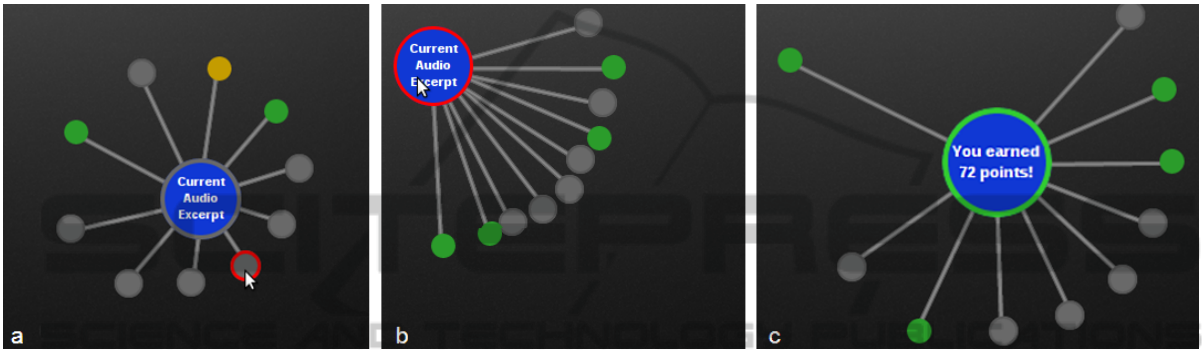

Figure 4: Similarity Graph: a) listening to a neighbour (on over); b) dragging the graph around and comparing distances; c) earning points from saving tags to label the audio excerpt (current audio frame becomes green, meaning it was already labeled by this user).

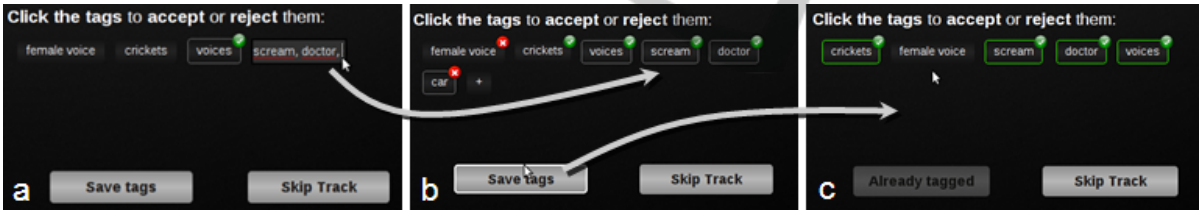

Figure 5: Labeling the Audio excerpt: a) two tags suggested, one inserted (the one with the frame), and introducing new ones in a comma separated list; b)accepting(v) and rejecting(x) tags; c) labelled audio with four tags introduced by the user and a suggested tag ignored.

the earned score at the central node of the graph that is also enlarged (Fig.2d). Now, the user may use the Next Sound button to randomly navigate to another segment in the same movie, or may choose a specific excerpt from the similarity graph or from the timeline (Fig.2c-b).

5 GAME ELEMENTS OVERVIEW

This section presents main design options to induce and support users contributing to the movies soundtracks labelling, with a gamification approach.

Assessing User's Skills. Since the labelling of some audio categories is not an obvious task, the users' level of expertise has to be evaluated, to assign a confidence level to their labelling. When users play for the first time, they are presented with audio excerpts for which the confidence level given by the model is very high. These excerpts are chosen frequently in the beginning for evaluating skills and trust of beginners, but they fade with time. The picked control excerpts are seamlessly integrated in the game flow without any visible difference from the other excerpts.

Involving The User. Once the level of expertise of the users is assessed, the system starts to benefit from their skills aware of their confidence level of their contributions. Our models also give us a confidence level for each label, indicating how "difficult" the query is; and an estimate of the similarity between two audio excerpts (independently of labels). Using

this information, we can present audio excerpts and possible labels to the user. During the interaction with the user, it alternates simple and difficult queries to better evaluate the user. When the "correct" labels are unknown, a consensus rule is used between users. Periodically, the new label associations are used to estimate parameters for a new generation of models of the audio backend that will this way benefit from the users' input.

Rewards. Gamification typically involves some kind of reward to the users. Although the needs for achievement or even cash incentives are often considered, these may be felt as controlling and not aligned with the player's culture (Khaled, 2011).

In SoundsLike, each user is assigned points for different achievements: 1) when their labels correspond to an existing consensus by partial or entire textual matching, taking into account their confidence level; 2) when the consensual answer corresponds to a difficult query; and 3) when a label proposed by a user gets confirmed by others the former is rewarded. This way users can see their score increase even when offline, through in-game and email notifications. Such notification may act as an incentive to return to the game later. A further analysis is required to determine how to maintain users engaged, particularly for cases where users are not immediately rewarded and must wait for others to receive points.

6 USABILITY EVALUATION

An evaluation with users has been performed to assess SoundsLike interface, its features, and their perceived usefulness, satisfaction and ease of use. In the process, detected usability problems and user suggestions could inform us about future improvements in the application.

Method and Participants. A task-oriented approach was conducted to guide the user interactions with the different features, while we were observing and taking notes of every relevant reaction or commentary. The evaluation was based on the USE questionnaire (Arnold M. L., 2001) and using a 1-5 Likert scale. At the end of each task, we asked for suggestions and USE evaluation of every relevant featureand to rate the application globally, refer to the aspects or features they liked the most and the least, and classify it with a group of terms from a table representing the perceived ergonomic, hedonic and appeal quality aspects (Hassenzahl et al., 2000). The evaluation had 10 participants (which allows to find most usability problems and perceive tendencies in users acceptance and preferences), with ages ranging from 21 to 44 years (24 mean value), with computer experience and a frequent use of internet.

Results from the USE based evaluation are presented in table 1, by mean and standard deviation values for each feature, performed in the context of the 8 tasks, described next, followed by a global evaluation of SoundsLike.

Assessing SoundsLike Features. In task 1, after reading an introductory sheet, users were presented with the application for the first time, with a random (non tagged) audio excerpt, and asked to identify every element that would represent the current audio excerpt in the interface. Most users were able to quickly identify all the components from the interface, even though some users had difficulties pointing to the current audio segment in the similarity graph, in the first contact.

The second task involved the playback of audio segments from the timeline and graph. We asked the users to play the sound and video of some excerpts. Most users identified a relationship between the graph nodes and the timeline elements and appreciated the interaction between every component during the video playback. We noticed that the quick sound playback while hovering a similar excerpt (T2.1) took an important role in the perception of those elements as audio segments by the users, without the use of additional informational tips. On the other hand, it was not obvious for some users at first time that clicking on every excerpt would play the associated video, although the feature was very much appreciated when learned.

The third task focused on the timeline features, such as the overview of the film provided by the video timeline, scrolling and manipulation of the dynamic zoom. Users were told to play the video audio excerpts in the zoom timeline in a fashion that would require them to manipulate the timeline to attain the task's objectives. All users used the scroll buttons without problems, and the majority quickly perceived the zoom manipulation in the video timeline (T3.2). But most users did not find utility about the audio spectrogram (T3.3), which they reported was due to a lack of knowledge in the area of sound analysis, but recognized its importance to users that would have that knowledge.

The fourth task was meant to introduce the user to the real objective of the similarity graph and navigation by changing (selecting) the current audio excerpt . We simply asked the user to listen to the three

Table 1: USE Evaluation of SoundsLike (scale: 1-5). (M = Mean, Δ = Std. Deviation).

		Usefulness		Satisfaction		Ease of Use	
Task	Feature	M	Δ	M	Δ	M	Δ
1.1	**Find the current node** in the **timeline**.	4.4	0.7	4.3	1.1	4.5	0.8
1.2	Find the current node in the **similarity graph**.	2.9	1.4	3.4	1.4	3.7	1.4
2.1	**Play** of the segment's sound.	3.7	1.1	4.2	0.8	4.4	0.7
2.2	Play of the segment's video.	4.7	0.5	4.6	0.7	4.4	0.8
3.1	**Movies Timeline**.	4.4	0.8	4.0	1.3	3.6	1.0
3.2	Zoom timeline.	4.3	1.1	4.2	1.1	4.3	0.9
3.3	Spectrogram timeline.	2.7	1.6	3.5	1.2	4.3	0.8
3.4	Timeline relationships.	4.6	0.5	4.4	0.7	4.0	0.9
3.5	Timeline - Overview.	4.6	0.5	3.9	0.7	3.8	0.8
4.1	**Similarity Graph**.	4.6	0.7	4.0	0.8	4.2	0.9
4.2	Graph dynamism.	3.5	1.2	3.6	0.8	2.5	1.0
4.3	Sound segments colours.	4.6	0.5	3.6	0.8	2.5	1.0
5.1	Choosing suggested **tags**.	4.6	0.5	4.4	0.5	4.3	0.9
5.2	Add new tags.	4.8	0.4	4.5	0.5	4.2	0.8
5.3	The possibility of tag rejection.	4.7	0.7	4.6	0.7	3.4	1.4
5.4	Adding tags fast.	5.0	0.0	4.4	1.0	2.4	1.4
6.1	**Play** of the segment's sound on **tagging context**.	4.8	0.4	4.8	0.4	4.6	0.5
6.2	Play of the segment's video on tagging context.	4.9	0.3	4.7	0.5	4.7	0.5
6.3	Using graph's **similar sounds** for tagging the current sound segment.	4.9	0.3	4.8	0.4	4.7	0.5
7	**In game context**, choosing the most similar is an efficient way of earning points?	4.5	1.3	4.4	1.1	4.5	1.0
8	**Points**' attribution for sound tagging	3.9	0.9	3.4	1.3	4.0	1.2
	SoundsLike Overall Evaluation	**4.4**	**0.5**	**4.2**	**0.6**	**3.9**	**0.6**
	Total (mean)	*4.3*	*0.6*	*4.2*	*0.4*	*4.0*	*0.6*

most similar sounds, select one of them and observe the transition to become the current excerpt. We also evaluated the user's perception about each audio element and their colouring. We observed that most users got the distance metaphor correctly, but they did not move the graph to verify ambiguous cases (when every node stands almost at the same distance), until they got instructed to do so (T4.2) and since the distances did not differ that much, in this case the usefulness was 3.5.

In the fifth task, we introduced tagging to the users, where they could add some tags to audio excerpts and submit the changes to the database. We prepared three cases: one audio excerpt without any kind of tag associated, and two with suggested tags: one case where tags were presented related with the current audio excerpt, the other one with unrelated tags. We observed that the users were able to introduce and accept suggestions (T5.1) without significant problems, but some failed to perceive the tag rejection feature (T5.3) without a proper explanation from the evaluator. Despite the usefulness of the fast tagging feature (comma separated text), without a proper tooltip, users were unable to find and use it without help (T5.4,U:5,S:4.4,E:2.4), but it is a typical feature for more experienced users as a shortcut, very appreciated as soon as you become aware of it.

With every user interface section covered, tasks 6, 7 and 8 were meant to immerse the user inside the labelling experience in the application. Here users were asked to label ten audio excerpts without any restriction, starting on a random excerpt. In this context, we observed that most users found the similarity graph more useful than other components due to their highly similarity and the propagation of previous used tags as suggestions. We also inquired users about the scoring mechanism (T8), and they found it interesting and a great way to stimulate users to participateand around half of them also showed interest in rankings for competition. We also observed a great user's engagement and immersion within the application, and received compliments and suggestions to be discussed and applied in future developments.

Overall. In the end, users were asked to rate SoundsLike globally. The values obtained were fairly high, with values of 4.4 for usefulness, 4.2 for satisfaction and 3.9 for ease of use. This feedback offers a good stimulus for continuing the development and improvement of the application and project. The experience also allowed us to control and witness the rapid and fairly easy learning curve, and to notice that some of the least understood features in the first contact turned out to be the most appreciated. Users provided us with a great amount of suggestions for improvements and some possible new features. Engagement was observed in most users when they were using the application freely to label excerpts without any restriction, during the execution of task 6. They pointed for the interface fluidity as one of the factors contributing for the engagement felt. The most appreciated features pointed out by the users were the similarity graph, the timelines and the scoring system. The least appreciated was the audio spectrogram because some users commented on their lack of expertise in the field of audio analysis.

Table 2: Quality terms to describe SoundsLike. H:Hedonic; E:Ergonomic; A:Appeal (Hassenzahl et al., 2000).

#	Terms		#	Terms	
6	Controllable	H	4	Innovative	H
6	Original	H	4	Inviting	A
5	Comprehensible	E	4	Motivating	A
5	Simple	E	3	Supporting	E
5	Clear	E	3	*Complex*	E
5	Pleasant	A	3	*Confusing*	E
4	Interesting	H	3	Aesthetic	A

At the end of the interview, we prompted users to classify the application with most relevant perceived ergonomic, hedonic and appeal quality aspects from (Hassenzahl et al., 2000) (8 positive and 8 negative terms for each category in a total of 48 terms), as many as they would feel appropriate. Table 2 displays the most chosen terms, with the terms "Controllable" and "Original" on the top with 6 votes each, "Comprehensible", "Simple", "Clear" and "Pleasant" leading after with 5 votes each, followed by "Interesting", "Innovative", "Inviting" and "Motivating" with 4 votes, and "Supporting", "Complex", "Confusing" and "Aesthetic" with 3 votes. Almost all the terms were positive, the most for Ergonomic qualities which are related with traditional usability practices and ease of use. The two most frequent negative terms were "Complex" and "Confusing", although the first is correlated with interesting or powerful applications , and both terms are also opposite to the terms "Simple" and "Clear", that were selected more often.

7 CONCLUSIONS

This paper describes and evaluates SoundsLike, a new Game With A Purpose whose objective is to collect labels that characterize short audio excerpt taken from movies. The interface is designed to entertain the user while pursuing the data collection task. The proposed interface innovates with respect to previous GWAPs with similar objectives by providing a rich context both in terms of temporal aspects through three timelines with different time scales, and in terms of similarities between items by displaying a dynamic force-directed graph where the neighbourhood of the current item is represented. The interface was evaluated through a user study. We observed that users went through a pleasant experience and they found that it was original, controllable, clear and pleasant. They particularly liked the similarity graph and the timeline representation. This also provided us with useful feedback and we were able to identify some usability aspects to improve.

Future work includes: 1) the refinement of the interface based on the feedback received and the perceived issues; 2) the refinement of the scoring mechanisms. For example, a user could be given no points for creating a label and, instead, the creator of a label would be given points when this label was re-used by other users. We could also give medals to the users, that would correspond to the number of created tags that had been reused a given amount of times (like the h-index user for ranking researchers). This would introduce a race among players in order to gain medals, because finding relevant labels is an easy task at the beginning but it is likely to become harder and harder.

ACKNOWLEDGEMENTS

This work is partially supported by FCT through LASIGE Funding and the ImTV project (UTA-Est/MAI/0010/2009).

REFERENCES

Arnold M. L. (2001). Measuring Usability with the USE Questionnaire. In *Usability and User Experience, 8(2).*

Barrington, L., O'Malley, D., Turnbull, D., and Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proc. of the ACM SIGKDD Workshop on Human Computation - HCOMP*, page 7, USA. ACM.

Daniel, G. and Chen, M. (2003). Video visualization. In *IEEE Trans. on Ultrasonics, Ferroelectrics and Frequency Control*, pages 409–416. IEEE.

Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". In *Proc. of the 15th Int. Academic MindTrek Conf.*, page 9, USA. ACM.

Gil, N., Silva, N., Dias, E., Martins, P., Langlois, T., and Chambel, T. (2012). Going Through the Clouds: Search Overviews and Browsing of Movies. In *Proc. of the 16th Int. Academic MindTrek Conf.*, pages 158–165, Finland. ACM.

Hassenzahl, M., Platz, A., Burmester, M., and Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. *Proc. of the SIGCHI conf. on CHI*, 2(1):201–208.

Khaled, R. (2011). Its Not Just Whether You Win or Lose: Thoughts on Gamification and Culture. In *Gamification Workshop at CHI'11*, pages 1–4.

Langlois, T., Chambel, T., Oliveira, E., Carvalho, P., Marques, G., and Falcão, A. (2010). VIRUS: video information retrieval using subtitles. In *14th Int. Academic MindTrek Conf.*, page 197, USA.

Langlois, T. and Marques, G. (2009). Automatic music genre classification using a hierarchical clustering and a language model approach. In *MMEDIA*, pages 188–193. IEEE.

Law, E., Dannenberg, R., and Crawford, M. (2007). Tagatune: a game for music and sound annotation. *ISMIR 2007*.

Mandel, M. and Ellis, D. (2008). A Web-Based Game for Collecting Music Metadata. *Journal of New Music Research*, 37(2):15.

Morton, B. G., Speck, J. A., Schmidt, E. M., and Kim, Y. E. (2010). Improving music emotion labeling using human computation. In *Proc. of the ACM SIGKDD Workshop on Human Computation - HCOMP*, page 45, USA. ACM Press.

Von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *CHI '04*, volume 6, pages 319–326, USA. ACM.

Von Ahn, L., Liu, R., and Blum, M. (2006). Peekaboom: a game for locating objects in images. In *Proc. of the SIGCHI conf. on CHI*, page 55, USA. ACM.