AudioMetro: directing search for sound designers through content-based cues

Christian Frisson, Stéphane Dupont, Willy Yvart, Nicolas Riche, Xavier Siebert, Thierry Dutoit
numediart Institute, University of Mons
Boulevard Dolez 31, 7000 Mons, Belgium
{christian.frisson; stephane.dupont; willy.yvart; nicolas.riche; xavier.siebert; thierry.dutoit}@umons.ac.be

ABSTRACT

Sound designers source sounds in massive collections, heavily tagged by themselves and sound librarians. For each query, once successive keywords attained a limit to filter down the results, hundreds of sounds are left to be reviewed. *AudioMetro* combines a new content-based information visualization technique with instant audio feedback to facilitate this part of their workflow. We show through user evaluations by known-item search in collections of textural sounds that a default *grid* layout ordered by filename unexpectedly outperforms content-based similarity layouts resulting from a recent dimension reduction technique (Student-t Stochastic Neighbor Embedding), even when complemented with content-based glyphs that emphasize local neighborhoods and cue perceptual features. We propose a solution borrowed from image browsing: a proximity grid, whose density we optimize for nearest neighborhood preservation among the closest cells. Not only does it remove overlap but we show through a subsequent user evaluation that it also helps to direct the search. We based our experiments on an open dataset (the OLPC sound library) for replicability.

Categories and Subject Descriptors

H5.1 [**Multimedia Information Systems**]: Evaluation / methodology; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces: Graphical user interfaces; H5.5 [**Information interfaces and presentation**]: Sound and Music Computing: Systems

General Terms

Design, Experimentation.

Keywords

Media browsers, sound effects, visual variables, music information retrieval, content-based similarity, known-item search

AM'14, October 01–03 2014, Aalborg, Denmark.

ACM 978-1-4503-3032-9/14/10 $15.00.
http://dx.doi.org/10.1145/2636879.2636880

Figure 1: Setup of the third known-item search evaluation of sound layouts. By hitting a buzzer with her left hand, the tester is to submit the potential target hovered by her right hand on a touchpad.

1. INTRODUCTION

Sound designers source sounds in massive collections, heavily tagged by themselves and sound librarians. If a set of sounds to compose the desired sound effect is not available, a Foley artist, that can be the sound designer herself/himself, records the missing sound(s), and both will tag these recordings as accurately as possible, identifying many facts from physical (object, source, action, material, location) and digital (effects, processing) properties. When it comes to looking for sounds in such collections, for each query, once successive keywords helped the user to filter down the results, but attained a limit, hundreds of sounds are left to be reviewed.

We elicited the following research questions. Can content-based organization be beneficial once a limit is reached when filtering sounds by tag? Are there different search behaviors among users in 2D presentations of results?

This paper presents 4 within-subject summative experiments meaning to solve these questions and whose evaluations lead to a solution to interactively browse collections of textural sounds after these have been filtered by tags. The first evaluated the benefits of a content-based *cloud* obtained by plotting coordinates after dimension reduction. The second investigated the advantages gained when complementing such a representation with content-based glyphs visualizing each sound. The third aimed at understanding whether expert users would perform differently. The fourth introduced a new information visualization technique taking into account all these considerations. Experiments were based on an open dataset and known-item search evaluation.

2. BACKGROUND

We first investigated browsers sound designers may opt for to organize their sound collections: generic for files, more specific for media files, and dedicated for sound design.

2.1 File browsers

From our contextual inquiry we noticed that sound designers also make use of simple browsers, such as the default provided by the operating system, optionally associated to a spreadsheet to centralize tags. Depending on their daily operating system of choice, everyday users of computers may be differently accustomed to any of the common layouts offered by file browsers. For instance the Apple OSX Finder allows to switch between layouts named "icons", "list", "columns" and "Cover Flow". Several reasons may motivate users to switch to such layouts: the nature of the files, the type of search mode (exploratory, directed), the presence of a hierarchy in the file structure, comparing files through metadata.

Fitchett et al. [11] studied different aspects to improve navigation-based file retrieval interfaces. They introduce three design-based factors: 1) Icons Highlights (IH) to increase the visual salience of some items, 2) Hover Menus (HM) to provide shortcuts across menus to locate items deeper in the hierarchy and 3) Search Directed Navigation (SDN) to provide search-based guidance. They show that interfaces with IH and HM best suited for frequently accessed items and SDN for infrequent ones.

2.2 Media browsers

While observing systems for different media types, image and video, that have been under deeper consideration in the fields of multimedia information retrieval (MIR), human-computer interaction (HCI), we notice that such research has long been clustered first by media type (visual media versus audio), then communities. Such boundaries have been evolving recently [23]. One distinction with visual media browsers is that these often present media collections in grids, due to the rectangular nature of the associated files. A very recent work providing cascading background references is the *Panopticon* system [17]. However some works such as the seminal *Film Finder* by Ahlberg and Shneiderman [1] proposed a more scattered visualization, coining the term *starfield display*, with the difference it didn't display rectangular thumbnails of videos but dots of coordinates defined by metadata analysis. The most generalized output modality for presentation interfaces (for non-impaired users) being visualization, audio media, that are non-visual by nature, need to be transduced for instance into visual representations to facilitate their analysis. While visual media items can be browsed by being looked at, audio items need be triggered for playback.

2.3 Browsers for sound design

Systems and methods for music information retrieval have been investigated into further depth, for instance in a comprehensive survey by Casey et al. [5]. Here we focus on a subset of the sound designers workflow: browsing. Thus we won't evaluate solutions for breeding, compositing, morphing sounds. We categorized these browsers between commercial and research-grade systems.

2.3.1 Commercial systems

A pioneering application is *SoundFisher* by company Muscle Fish [27], start-up of scientists from the field of audio retrieval. Their application allowed to categorize sounds along acoustic features. The browser offers several views: a detail of sound attributes in a spreadsheet, a tree of categories resulting from classification by example (a set of sounds input by the user), a scatter plot with one feature per axis.

AudioFinder by Iced Audio[1] mimics personal music managers such as Apple *iTunes*: on top a textual search input widget allows to perform a query, a top pane proposes a hierarchical view similar to the "column" view of the Finder to browse the file tree of the collection, a central view features a spreadsheet to order the results along audio and basic file metadata, a left pane lists saved results like playlists. A bottom row offers waveform visualizations and the possibility to apply audio effect processing to quickly proof the potential variability of the sounds before dropping these into other applications such as digital audio workstations.

Soundminer HD[2] provides a similar interface, plus an alternative layout, *3D LaunchPad*, that allows similarly to the Apple *Finder CoverFlow* view to browse sounds (songs) by collection (album) cover, with the difference that the former is a 2D *grid* and the latter a 1D rapid serial visualization.

Other companies facilitating creativity such as Adobe with *Bridge*[3] provide more general digital asset managers that are accessible through their entire application suite. Besides browsing, these tools may also offer batch processing and *cloud* capabilities. These focus on production-required capabilities and seem to avoid content-based functionalities.

2.3.2 Research-grade systems

In her PhD thesis [24], Stewart compared 20 systems with auditory display including 2 audio browsers presented thereafter. She underlines that a few published works on such systems provide accurate usability evaluations what may be due to a divide between HCI/MIR research communities.

Sonic Browser, by Fernström and Brazil, focused on information visualization [9], and later approached content-based organization through the Marsyas framework [4]. A 2D *starfield display* allows to map the metadata of audio files to visual variables. They evaluated their system qualitatively and positively against the Microsoft Windows 2000 explorer through a think-aloud protocol with 6 students [9]. We couldn't trace reports of usability evaluations of the contribution of content-based organization in their system, however they showed that multiple stream audio feedback significantly outperformed single-stream [10].

The *SoundTorch* content-based audio browser has been designed by Heise et al. [15, 16]. It is the only directly-related work to provide a user evaluation, quantitative with known-/described-item search tasks, comparing positively with 15 users their system against a list-based application. It is not clear from this comparison whether *SoundTorch* outperforms the list-based application because of its content-based or interactive abilities, particularly its instant playback of multiple sound streams. Moreover, they chose to randomize the sound list. Ordering by filename would form a comparison baseline closer to existing solutions.

[1] `http://www.icedaudio.com`
[2] `http://www.soundminer.com`
[3] `http://www.adobe.com/creativesuite/bridge.html`

CataRT by Schwarz "mosaices" sounds into small fragments for concatenative synthesis. A 2D scatter plot allows to browse the sound fragments, with user-definable features assigned to the axes. The authors recently evaluated distribution algorithms [18] to optimize the spreading of sounds in a scatter plot and to opening new perspectives for non-rectangular interfaces (the circular *reacTable*) and complex geometries (physical spaces to sonify). To our knowledge no user study has yet been published about this browser, but planned [18].

3. EVALUATION METHOD

3.1 Known-item search

When reviewing results of queries past keyword filtering, sound designers may not necessarily picture target sounds accurately in their head or be able to name further characteristics, leaning closer towards exploratory search. That said, we need a method to quantitatively estimate the efficiency of a sound browsing system to assess if our research efforts are in the right tracks. Known-item search tasks consist in displaying a media element or fragment to a tester and requiring her/him to find back the target inside a collection or long record. Task success and retrieval time can be measured and serve as metrics. The TRECVID evaluation of video browsers had been including known-item search tasks between 2010 and 2012, but with text-only descriptions of targets, and the analysis doesn't take user interaction much into account. In 2012, Schoeffmann and Bailer launched the *Video Browser Showdown*, a live competition, that addresses these issues [22]. The MIREX evaluation for music browsers hasn't proposed a known-item search track so far [7]. Font's work about sound browsers deliberately rejected investigating on time and speeds, claiming people have different search behaviors [12].

3.2 Open dataset

The One Laptop Per Child (OLPC) sound library[4] was chosen so as to make the following tests easily reproducible, for validation and comparison perspectives, and because it is not a dataset artificially generated to fit with expected results.It is licensed under a Creative Commons BY license (requiring attribution), while current datasets from MIREX face copyright issues [7]. In over 8GB of digital storage space, it contains 8458 sound samples. Sound designers collection nowadays contain 10 or 100 times more. Several volunteers contributed sub-libraries, 90 to be precise. It is to be noted, especially for subset libraries curated by Berklee containing Foley sound design material, that within a given subset most samples seem to have been recorded, if not named, by a same author per subset. It is thus frequent to find similar sounds named incrementally.These are likely to be different takes of a recording session on a same setting of sounding object and related action performed on it. Ordering search results by tag filtering in a list by path and filename as would a standard file browser do, will thus imprint local neighborhoods to the list.

[4] `http://wiki.laptop.org/go/Free_sound_samples`

4. STUDY 1: CONTENT-BASED CLOUD

We first wanted to investigate how content-based organization would assist browsing sounds in a *cloud* where coordinates induce similarity, versus a *grid* ordered by filename.

4.1 System

A usual multimedia information retrieval workflow consists in several steps, that can be offline such as indexation and clustering, and as much close to realtime such as dimension reduction and changing parameters of a visual representation.

A first step is feature extraction. We based our selection of features from the work of Dupont et al. [8] since their evaluation considered textural sounds. We refer to their paper for more details. In short we used a combination of derivatives of and statistics (standard deviation, skewness and/or kurtosis) over Mel-Frequency Cepstral Coefficients (MFCC), a regular feature in most MIR systems, and Spectral Flatness (SF) correlated to noisiness. Grill et al [14] aimed at defining features correlated to perceived characteristics of sounds that can be named or verbalized through *personal constructs*, for instance *high-low*. One application of this work is to simplify the user interface of MIR systems by making the choice of features more understandable by users not expert in signal processing. Following their results, we also made use of Perceptual Sharpness that is highly correlated to the perceived brightness of the sound and that was the most correlated to one feature present in the YAAFE feature extraction library [19]. We decided not to perform segmentation on the sounds, for instance to remove silence or adapt to variations in homogeneity, while this is an expected step. Our test collections feature textures of short length and steady homogeneity.

Another important step is dimension reduction. We opted for Student-t Stochastic Neighborhood Embedding (t-SNE) [8, 13, 25]. In short, this method aims at preserving high-dimensional neighbors in a lower-dimensional projection (here 2D) by estimating the probability of each pair of sounds to be neighbors. One emergent result in applying this dimension reduction technique to textural sounds is that takes recorded from the same sound source with slight variations are almost always neighbors in the 2D representation. As input we used a linear combination of all the aforementioned features with the same unit weight. An undesirable artifact from t-SNE is that the 2D positions are initialized randomly, making the representation of a given sound collection variable over time, what works against the human memory especially for exploratory search. We solved this issue by choosing to initialize these positions with the two first axes of a Principal Component Analysis (PCA) of the features, probably not providing an optimal solution, but making it repeatable.

Displaying such a representation results in a scatter plot or *starfield display*. Neither waveforms nor filenames were displayed, to have testers concentrate the visual memory on layouts, and to avoid taking into account the time spent in understanding tags semantically. In their paper about *collection understanding* that they oppose to information retrieval, Chang et al [6] argue how scrolling can be a time-consuming and burdensome interaction technique. We chose to disable panning and zooming.

4.2 Apparatus

Tests where performed on an 15" Apple MacBook Pro laptop (late 2008 model) with a resolution of 1440×900, the test application always fullscreen. For auditory display an Echo AudioFire 4 soundcard wired to a pair of Genelec 8020 CPM powered loudspeakers were used. A 3Dconnexion Space Navigator 3D mouse was repurposed as "buzzer" to submit the closest sound to the pointer as target by bumping on the device, instilling a game feel we believed would have testers concentrated to meet task deadlines. The number of fingers sensed by the multitouch trackpad was accessed through the Apple OSX Multitouch private framework: a 1-finger touch activates the looped playback of the sound represented by the closest node to the pointer/finger. Otherwise noted, this apparatus is maintained in subsequent experiments.

4.3 Participants

Testers were recruited from colleagues, experts in computer vision or speech audio analysis, working in a digital signal processing lab. Among the 19 testers (2 female), 2 were known to be color blind. Most of them are knowledgeable of machine learning methods (some work with dimension reduction techniques such as PCA) and scientific visualization.

4.4 Design

This controlled laboratory experiment was set up in a small meeting room, free of visual clutter. After a short 4-task introductory training phase, each tester performed 6 tasks, each time-limited to 60 seconds, passing automatically to the next after the deadline. Each task was associated to one in 6 collections of 64 elements, filtered by tag from the whole OLPC library, using the following keywords: *ball, bell, glass, metal, scrape, water*. A power of two was chosen as collection size to make sure that the *grid* visualization would be symmetric, so as not to induce pathways, assuming that a "crooked" ordering resulting from another collection size making the last row unequal in regard to the others would stimulate users to browse the view in the western reading order, from the top-left corner. To trim down the search results that would exceed 64 items for some of the keywords, files presenting iterative patterns in their filenames were first dismissed, since these would be similar "recording takes" as stated earlier. This reduces similarity neighbors artifacts from filename ordering. Collections were displayed either in a *grid* or a cloud, in a permuted order. The targets were chosen offline without qualitative evaluation by their index import number in regard to the collection so as to be spread over the collections: these targets weren't heard or visualized during the preparation, so as not to favor any task.

The pointer was programmatically moved to the left mid-height position, systematically at the beginning of each task, near a countdown showing the time left to perform the task.

Each session was thus identical for all testers, as illustrated in Figure 2, and included sequentially: 1) warming up with simple known-item search tasks to get acquainted with the setup; 2) performing a larger set of known-item search tasks, with times and mouse path logged, while receiving aloud the first impressions from the testers.

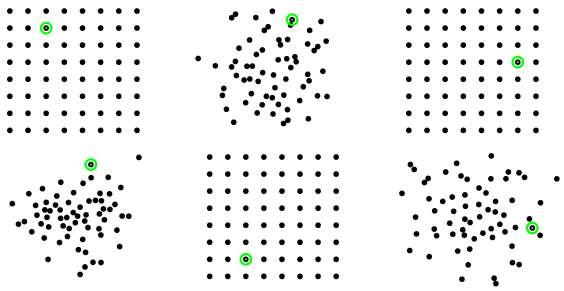

Figure 2: Visual organization and target location for each task. Pairs of subsets filtered by keyword and layouts (left-right, top-bottom): *ball grid*, *bell cloud*, *glass grid*, *metal cloud*, *scrape grid*, *water cloud*.

4.5 Results

From the temporal and spatial variables logged, many quantitative metrics could be extracted. Here we report an analysis on: the percentage of succeeded/failed tasks, the time until the target was successfully found and "buzzed", the distance browsed for each task, the average speed of each task, patterns in the mouse paths. We introduce a user score allowing to compare the results of all users for each view with vectors of metrics of equal sizes, whether or not each target has been successfully retrieved: $\sum_i t_{\max} - t_i$ where i is the number of tasks, $t_{\max}$ is the time limit, t_i the time of target finding at task i. The Shapiro-Wilk test statistic (W) and its p-value can be computed to evaluate the normality of distribution of results: *grid* scores look normal (W=0.98, p=0.94), *cloud* scores more uniform (W=0.91, p=0.07). Normality is not assumed for at least one of the two paired groups, plus we have a small sample size (n<25), we will use Mann-Whitney's u-test for further comparisons. The following table reports the results of two-sample u-tests of the null hypothesis of equal medians of these variables between both views. From this table, we can notice that p-values are quite higher than 0.05, showing that results per variable do not drastically differ between views.

Variable	p-value	Z	M_{grid}	M_{cloud}
Success time (s)	0.66	-0.43	**30.72**	32.03
Distance (cells)	0.95	-0.06	9.19	**9.03**
Speed (cells per s)	**0.37**	0.89	**0.31**	**0.32**

Table 1: Mann-Whitney u-tests (*grid* > *cloud*) of all variables for test 1: p-value, statistic (Z), means per view (M_{view}). Bold means better.

More tasks were successful with *grid* (45/57) than with *cloud* (37/57): clues might be present in the pathways browsed by the mouse pointer and the localization of each target in each task. Some tasks were missed for both views: this test can be considered difficult. Successful targets were submitted on average at half of the deadline, which corresponds to the probable time taken for listening the first second of each sound sequentially. The average speed of browsing for both views match: users seem to be willing to complete tasks with both views without favoring any.

Visual organizations induced patterns of spatial browsing. *Cloud* was browsed more in a brushed way. A good strategy seems to be either looping around the periphery or diving straight inside, in both cases with an added oscillation. Users were more likely to intersect their previous ways with *cloud*. For *grid*, the shortest paths seem to have been more serendipitous than strategical. Past these random chances, the "lawn mowing path" seems to be the fastest and most efficient way and was naturally chosen by most users. Starting from the top left corner, line by line, but with a difference to western text reading in that audio lines were read alternatively forwards and backwards. Fewer users used a transposed version, mowing the map vertically, thus progressing in columns. Kerry Rodden uses the terms *systematic* to describe a browsing approach such as the "lawn mower's path" and *haphazard* for more "sketchy" pathways [21].

5. STUDY 2: CONTENT-BASED GLYPHS

As content-based positioning of sounds on a map appeared to be inefficient vs a simple solution, we investigated whether adding content-based glyphs to represent sounds would help.

5.1 System

Ware's book offer great explanations and recommendations to use visual variables to support information visualization tailored for human perception [26]. Grill et al.'s approach was to map many perceptual audio features to many visual variables (position, color, texture, shape), in one-to-one mappings [13]. They chose to fully exploit the visual space by tiling textures: items are not represented by a distinct glyph, rather by a textured region. In a first attempt to discriminate the contribution of information visualization versus media information retrieval in sound browsing, we opted here for a simpler mapping. Iterating on the previous system, we mapped the mean over time of perceptual sharpness to the Value in the Hue Saturation Value (HSV) space of the node color for each sound, normalized against all sounds in each collection. We used the temporal evolution of perceptual sharpness to define a clockwise contour of the nodes. To compute the positions, perceptual sharpness was also added to the features from the former iteration of the system, intuiting it would gather closer items that are similar visually from their glyph representation. Why perceptual sharpness? By undertaking an online evaluation based on a repertory *grid* method asking testers to rate sounds by choosing continuous values on scales of perceived features, Grill et al. reported brightness to be one of the most salient statistically and also the most correlated to one of the audio feature algorithms offered by YAAFE: perceptual sharpness.

5.2 Participants

16 participants (1 female) of average age 21.9 (+/-2.2) years old were recruited from students in Engineering (Digital Signal Processing) during two afternoons of group work, volunteering during breaks. 8 had corrected vision.

5.3 Design

This controlled laboratory experiment was setup in a small control room embedded in a larger room where other students were working in groups, but without mutual visibility and limited sound interference. After a short 4-task introductory training, each tester performed 10 tasks, each time-limited to 60 seconds, passing automatically to the next.

A single collection of 150 elements was chosen from a subset of the OLPC collection: the *Berklee Sampling Archive Volume 7: noises (mechanical and industrial)*. Tasks toggled between layouts: *grid* without glyphs and *cloud* with glyphs, in an interleaved order.

Targets were chosen randomly at runtime, however the random seeds weren't initialized at startup, therefore task sequences were similar daily between tests since the application was restarted for each participant. A post-test questionnaire was submitted to participants to collect demographic data and qualitative feedback. No financial reward was provided, but chocolate was offered.

5.4 Results

We again compute the Shapiro-Wilk test statistic (W) and its p-value to evaluate the normality of distribution of results: neither *grid* (W=0.85, p<0.01) nor *cloud* (W=0.89, p=0.04) scores look normal. We will thus again use Mann-Whitney's u-test for further comparisons. The following table reports the results of two-sample u-tests of the null hypothesis of equal medians of these variables between views.

Variable	p-value	Z	M_{grid}	M_{cloud}
Success time (s)	0.04	-1.78	**33.94**	40.21
Distance (cells)	0.02	-1.98	**4.82**	6.14
Speed (cells per s)	**0.19**	-0.86	**0.15**	**0.15**

Table 2: Mann-Whitney u-tests (*grid* > *cloud*) of all variables for test 2: p-value, statistic (Z), means per view (M_{view}). Bold means better.

This time, *cloud* (with glyphs) was significantly slower than *grid* and required a smaller distance to reach the targets. Participants browsed both layouts at a similar speed, inclining us to claim that no layout seem to have been favored for instance by participants who would have guessed which results we expected. For the music information retrieval community these results may be considered negative since a simple baseline solution outperforms a complex system with a layout obtained from a recent dimension reduction technique, carefully chosen feature extraction, both evaluated algorithmically in previously mentioned references, and glyph representation aiming at supporting audition with vision from perceptual cues. Several potential reasons may be explored: tasks may be too hard to complete within the time limit, the system should be evaluated by expert users (in sound auditioning), the sound textures may be too complex and uncanny to non experts, better methods to cue sound similarity visually should be investigated.

6. STUDY 3: EXPERT STUDENTS

From the negative results obtained from the previous experiment, we targeted a different population sample closer to expert users (in sound auditioning), doubled the task deadline, and instilled a competition mood by announcing before the test session a give away of a prize to the best overall score.

6.1 System

The system from the previous experiment was employed.

6.2 Participants

27 participants (6 female) of average age 21.3 (+/-2.2) years old were recruited from students in Audiovisual Communication, during two days. 13 had corrected vision. All the participants have studied audiovisual communication practices such as sound design and film edition.

6.3 Design

The *Great CHI'97 Browse-Off forum* [20] can be considered as a pioneering evaluation of browsers in a live competition setting. Its differs from our scope since their datasets consisted in hierarchical structured data. Since 2012, Schoeffmann and Bailer have been organizing a yearly interactive evaluation of video browsers through known-item search tasks in a competitive ambience, as special session of the Multimedia Modeling conference, later promoted to collocated workshop. The most recent evaluation available, of the 2013 session, explains its design into further extent [22].

We slightly adapted our previous test design towards such methods. We doubled the time deadline from to 120s. We modified the test application to display a realtime score below the time countdown, computed as follows: $\sum_i t_{\max} - t_i - \sum p_r - \sum p_f$ where i is the number of tasks, $t_{\max}$ is the time limit, t_i the time of target finding at task i, p_r the number of target rehears commanded by users, p_f the number of false submissions A prize in cash was awarded to the contestant with the best score, announced a few days before the tests, and proportional to the number of participants to invite these to invite challengers. This experiment was setup in a corner of small cafeteria as illustrated in Figure 1.

6.4 Results

We again compute the Shapiro-Wilk test statistic (W) and its p-value to evaluate the normality of distribution of results: this time both *grid* (W=0.97, p=0.49) and *cloud* (W=0.92, p=0.03) scores look normal. We will thus use Student-t tests for further comparisons, unpaired due to the existence of failed tasks. Table 6.4 reports the results of two-sample u-tests of the null hypothesis of equal medians of these variables between both views.

Variable	p-value	Z	M_{grid}	M_{cloud}
Success time (s)	**0.02**	-2.04	**50.18**	56.29
Distance (cells)	**0.06**	-1.54	7.89	**7.84**
Speed (cells per s)	**0.94**	1.54	**0.15**	**0.14**

Table 3: Mann-Whitney u-tests (*grid* > *cloud*) of all variables for test 3: p-value, statistic (Z), means per view (M_{view}). Bold means better.

Even with this preferred population sample closer to experts, *cloud* (with glyphs) still remains significantly slower than *grid*. Again speeds were similar between layouts. Testers in the current experiment were slower in absolute time but faster in relative time (to the task deadline) than testers in the previous experiments.

Several users complained that some sounds were overlapping in the *cloud* layout. This must be addressed in following iterations, by adjusting the node sizes or positions.

7. STUDY 4: THE METRO LAYOUT

From results obtained with the previous experiments, we posited that a layout with regular geometry such as the *grid* directs the search pathway and helps user keep a visual track of their progress in screening collections. We iterated our system with an overlap-free layout combining such practicalities of *grid* and local similarity neighborhoods of *cloud*.

7.1 System

We borrow a method initially designed to solve the problem of overlap for content-based image browsing [21]: a proximity *grid* [2]. It is quite important to stress that their work is heavily cited respectively for the evaluation of multidimensional scaling techniques [2] and as a pioneering application of usability evaluation for multimedia information retrieval [21], but almost never for its *proximity grid* approach, the only such reference we found that actually applies this method is the seminal *PhotoMesa* image browser [3]. To our knowledge, no audio browser approached this solution.

A proximity *grid* consists in adapting the coordinates of each item of a 2D plot to magnetize these items on an evenly-distributed grid. We implemented the simplest greedy method with the empty strategy described in [2]. We opted for a simplification: the spiral search for empty cells was always turning clockwise and started above the desired cell, while it is recommended to choose the rotation and first next cell from exact distance computation between the actual coordinates of the sound item and the desired cell. The minimal side of a square *grid* is the ceil of the square root of the collection size, providing the most space efficient density. To approximate a least distorted grid, the collection size can be taken as *grid* side. To come up with a tradeoff between density and neighborhood preservation, we estimated the number of high-dimensional nearest neighbors (k=1) preserved in 2D at a given *grid* resolution simply by counting the number of pairs in adjacent cells. We distinguish the amounts of horizontal and vertical and diagonal neighbors since different search patterns may be opted by users: mostly horizontal or vertical for people accustomed respectively to western and non-western reading order, diagonal may be relevant for grids of light density. The obtained layout resembles a *metro* map.

7.2 Apparatus

The tests were undertaken on an Apple Macbook Pro Late 2013 laptop with 15-inch Retina display and resolution of 3360×2100, with a RME FireFace UCX sound card, and the same pair of Genelec active loudspeakers.

7.3 Participants

16 participants (5 female) of mean age 28 (+/- 6.3) were recruited from the same population as the previous experiment, including teachers this time. All self-rated themselves with normal audition, 10 with corrected vision.

7.4 Design

We prepared 3 collections filtered by tag from the whole OLPC dataset: *water* (77 sounds), *spring* (93) and *metal* (147). An additional smaller collection was used for training tasks with each layout.

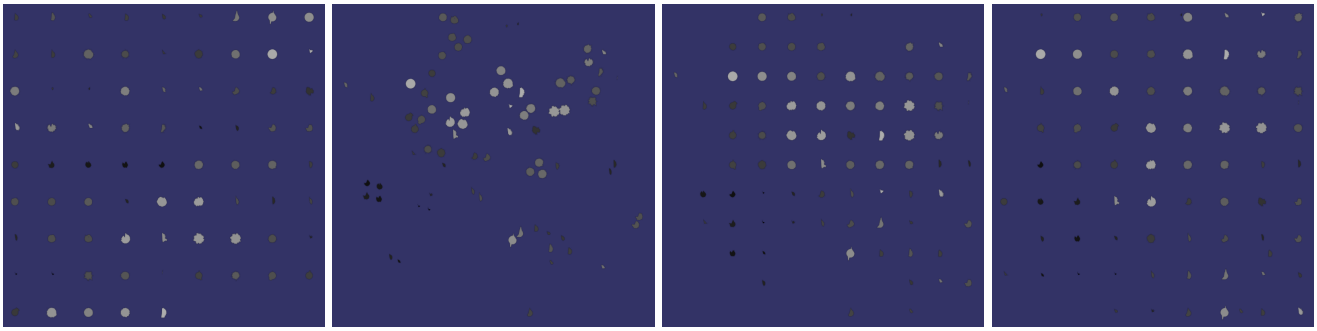

Figure 3: Different layouts with glyphs for the same sound collection filtered by keyword "water", from left to right: *album*, *cloud*, *metro*, *proximity grid*.

We qualitatively selected the optimal *grid* resolution based on the amounts of horizontal / vertical / diagonal adjacent neighbors computed for each resolution between the minimal side and the least distorted approximate, comparing such amounts between a proximity *grid* applied after dimension reduction and a *grid* ordered by filename. It is to be noted that not all collections obtained from other tags presented a proximity *grid* resolution that outperformed a simple *grid* by filename in terms of neighbor preservation.

Each layout was given a nickname: *grid* for the simple *grid* ordered by filename, *album* for its upgrade with glyphs, *metro* for the proximity *grid* of optimal resolution for neighbors preservation. These short nicknames brought two advantages: facilitating their instant recognition when announced by the test observer at the beginning of each task, and suggesting search patterns: horizontal lawn mowing for *grid* and *album*, adjacent cell browsing for *metro*. The *metro* layout was described to users using the metaphor of *metro* maps: items (stations) can form (connect) local neighborhoods and remote "friends" (through *metro* lines usually identified by color). Figure 3 illustrates all these layouts.

We sequenced the tasks for each tester as follows: *water metro*, *water album*, *water grid*, *spring grid*, *spring metro*, *spring album*, *metal album*, *metal grid*, *metal metro*. All collections exhibited several local neighborhoods with at least 3 very similar sounds that would end up close one another on each layout, exactly at the same positions between *grid* and *album*, elsewhere for *metro*. For each given collection, each layout was assigned one of such sounds as target.

We tamed the stress of users by removing the deadline and countdown display, only showing the score.

7.5 Results

We again compute the Shapiro-Wilk test statistic (W) and its p-value to evaluate the normality of distribution of results: this time both *grid* (W=0.93, p=0.23) and *metro* (W=0.93, p=0.22) scores look normal, *album* scores don't (W=0.84, p=0.01). Thus, instead of ANOVA, we use the Kruskal-Wallis rank sum test (chi-square = 5.26 with p=0.07) which shows that there is almost a significant effect of layouts. A Tukey multiple comparisons of success times means at a 95% family-wise confidence level on layouts shows that *metro* outperforms *grid* (p=.01), but *album* is not significantly better than *grid* (p=.34) or worse than *metro* (p=.26).

Table 4 displays for each layout the mean and standard deviation of success times, plus user-reported efficiency and pleasurability.

	grid	*album*	*metro*
Success times (s)	53.0(46.6)	43.1(38.0)	31.3(22.9)
Efficiency [1-5]	1.87(1.01)	3.75(1.00)	4.12(0.96)
Pleasurability [1-5]	2.25(1.18)	3.62(0.81)	4.25(0.86)

Table 4: Mean (standard deviations) of evaluation metrics for test 4.

These positive results open a promising track of investigation with the *metro* layout.

Feature extraction is a one-shot offline process at indexing time. Dimension reduction for layout computation is a process that should be close to real-time so as not to slow down search tasks and that is likely to be performed at least once per query. Decent results can be achieved by combining just content-based glyphs with simple ordering by filename. A content-based layout comes at a greater computational cost.

8. CONCLUSION AND DIRECTIONS FOR FUTURE WORKS

Through four iterations of a usability evaluation based on known-item search tasks, we designed a method to assist sound designers in reviewing results of queries by browsing a sound map optimized for nearest neighbors preservation in adjacent cells of a proximity grid, with content-based features cued through glyph-based representations. We showed that this solution was more efficient and pleasurable than a *grid* of sounds ordered by filenames.

Future tracks can be addressed by MIR and HCI communities. Each step of the MIR dataflow we employed could be improved, starting with improving feature extraction methods tailored for sound textures, correcting distance errors during dimension reduction or designing a new similarity layout from scratch. Regarding HCI considerations, most commercial applications for digital asset management use list or spreadsheet views, allowing different ways of sorting per categories of metadata. Conveying similarity through 1D, rather than 2D as in the current work, may help to adapt such systems directly. A subsequent iteration of this system should be designed to combine such a content-based map representation with a context-based tag view, and allow filtering by tag as a standalone application.

9. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their careful recommendations. We thank all the testers for their time and patience in performing tasks that were sometimes too difficult. We thank the professors from UMONS and UVHC and IRISIB for having let their students take the tests during breaks. This work is partly funded by the Walloon Region of Belgium through the GreenTIC grant SONIXTRIP.

10. REFERENCES

[1] C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *Conf. Companion on Human Factors in Computing Systems*, CHI '94. ACM, 1994.

[2] W. Basalaj. *Proximity visualisation of abstract data.* PhD thesis, University of Cambridge, 2000.

[3] B. B. Bederson. Photomesa: A zoomable image browser using quantum treemaps and bubblemaps. In *Proc. of the 14th Annual ACM Symposium on User Interface Software and Technology*, UIST '01. ACM, 2001.

[4] E. Brazil, M. Fernström, G. Tzanetakis, and P. Cook. Enhancing sonic browsing using audio information retrieval. In *Proc. of the Intl. Conf. on Auditory Display (ICAD)*, 2002.

[5] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. In *Proc. of the IEEE*, volume 96, 2008.

[6] M. Chang, J. J. Leggett, R. Furuta, A. Kerne, J. P. Williams, S. A. Burns, and R. G. Bias. Collection understanding. In *Proc. of the 4th ACM/IEEE-CS joint Conf. on Digital Libraries*, JCDL '04. ACM, 2004.

[7] J. S. Downie. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoust. Sci. & Tech.*, 29(4), 2008.

[8] S. Dupont, T. Ravet, C. Picard-Limpens, and C. Frisson. Nonlinear dimensionality reduction approaches applied to music and textural sounds. In *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2013.

[9] M. Fernström and E. Brazil. Sonic browsing: An auditory tool for multimedia asset management. In *Proc. of the 2001 Intl. Conf. on Auditory Display*, 2001.

[10] M. Fernström and C. McNamara. After direct manipulation—direct sonification. *ACM Trans. Appl. Percept.*, 2(4):495–499, Oct. 2005.

[11] S. Fitchett, A. Cockburn, and C. Gutwin. Improving navigation-based file retrieval. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '13. ACM, 2013.

[12] F. Font. Design and evaluation of a visualization interface for querying large unstructured sound databases. Master's thesis, Universitat Pompeu Fabra, Music Technology Group, 2010.

[13] T. Grill and A. Flexer. Visualization of perceptual qualities in textural sounds. In *Proc. of the Intl. Computer Music Conf.*, ICMC, 2012.

[14] T. Grill, A. Flexer, and S. Cunningham. Identification of perceptual qualities in textural sounds using the repertory grid method. In *Proc. of the 6th Audio Mostly Conf.: A Conf. on Interaction with Sound*, ACM, 2011.

[15] S. Heise, M. Hlatky, and J. Loviscach. Soundtorch: Quick browsing in large audio collections. In *125th Audio Engineering Society Convention*, 2008.

[16] S. Heise, M. Hlatky, and J. Loviscach. Aurally and visually enhanced audio search with soundtorch. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2009.

[17] D. Jackson, J. Nicholson, G. Stoeckigt, R. Wrobel, A. Thieme, and P. Olivier. Panopticon: A parallel video overview system. In *Proc. of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13. ACM, 2013.

[18] I. Lallemand and D. Schwarz. Interaction-optimized sound database representation. In *Proc. of the 14th Intl. Conf. on Digital Audio Effects (DAFx-11)*, 2011.

[19] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *Proc. of the ISMIR Conf.*, 2010.

[20] K. Mullet, C. Fry, and D. Schiano. On your marks, get set, browse! In *Extended Abstracts on Human Factors in Computing Systems*, CHI EA. ACM, 1997.

[21] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Does organisation by similarity assist image browsing? In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI. ACM, 2001.

[22] K. Schoeffmann, D. Ahlström, W. Bailer, C. Cobârzan, F. Hopfgartner, K. McGuinness, C. Gurrin, C. Frisson, D.-D. Le, M. Fabro, H. Bai, and W. Weiss. The video browser showdown: a live evaluation of interactive video search tools. *Intl. Journal of Multimedia Information Retrieval*, pages 1–15, 2013.

[23] K. Schoeffmann, F. Hopfgartner, O. Marques, L. Boeszoermenyi, and J. M. Jose. Video browsing interfaces and applications: a review. *SPIE Reviews*, 1(1):1–35, 2010.

[24] R. Stewart. *Spatial Auditory Display for Acoustics and Music Collections.* PhD thesis, School of Electronic Engineering and Computer Science, Queen Mary, University of London, 2010.

[25] S. Stober, T. Low, T. Gossen, and A. Nürnberger. Incremental visualization of growing music collections. In *Proc. of the 14th Conf. of the Intl. Society for Music Information Retrieval (ISMIR)*, 2013.

[26] C. Ware. *Visual Thinking: for Design.* Interactive Technologies. Morgan Kaufmann, 2008.

[27] E. Wold, T. Blum, D. Keislar, and J. Wheaten. Content-based classification, search, and retrieval of audio. *MultiMedia, IEEE*, 3(3):27–36, 1996.