A PROXIMITY GRID OPTIMIZATION METHOD TO IMPROVE AUDIO SEARCH FOR SOUND DESIGN

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 organize their sound libraries either with dedicated applications (often featuring spreadsheet views), or with default file browsers. Content-based research applications have been favoring cloud-like similarity layouts. We propose a solution combining the advantages of these: after feature extraction and dimension reduction (Student-t Stochastic Neighbor Embedding), we apply a proximity grid, optimized to preserve nearest neighborhoods between the adjacent cells. By counting direct vertical / horizontal / diagonal neighbors, we compare this solution over a standard layout: a grid ordered by filename. Our evaluation is performed on subsets of the One Laptop Per Child sound library, either selected by thematic folders, or filtered by tag. We also compare 3 layouts (grid by filename without visual icons, with visual icons, and proximity grid) by a user evaluation through known-item search tasks. This optimization method can serve as a human-readable metric for the comparison of dimension reduction techniques.

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 records the missing sound and tags these recordings as accurately as possible, identifying many physical (object, source, action, material, location) and digital (effects, processing) properties. When it comes to looking for sounds in such collections, successive keywords can help the user to filter down the results. But at the end of this process, hundreds of sounds can still remain for further review. This creates an opportunity for content-based information retrieval approaches and other means for presenting the available content. From these observations, we elicited the following research question: can content-based organization complement or outperform context-based organization once a limit is reached when filtering by tag?

This work partly addresses this question and presents a solution to interactively browse collections of textural sounds after these have been filtered by tags.

We organize sounds in a two-dimensional map using content-based features extracted from their signal. These features are mapped to two visual variables. First, the position of the sample on the screen is obtained after applying dimension reduction over the features followed by a proximity grid that structures items on a grid which facilitates navigation and visualization, in particular by reducing the cluttering. The organization of the samples on the grid is optimized using a novel approach that preserves the proximity on the grid of a maximum of nearest neighbors in the original high-dimensional feature space. Second, the shape of the sample is designed to cue one important content-based feature, the perceptual sharpness (a measure the "brightness" of the sound).

This approach is evaluated through a known-item search task. Our experiments provide one of the first positive result quantitatively showing the interest of MIR-based visualization approaches for sound search, when then proper acoustic feature extraction, dimension reduction, and visualization approaches are being used.

The paper is organized as follows. First, in section 2, we examine the landscape of existing systems dedicated to browsing files in sound design. We then describe how we designed our system in Section 3. In section 4, we describe our evaluation approach, experiments and obtained results. We finish by summarizing our contributions and provide an glimpse of future research directions.

2. BACKGROUND

This section provides a review of the literature and empirical findings on systems for sound design, and outlines some results and gaps that motivated this work.

Systems for mining sounds, particularly for sound design, are actually rather scarce. These may however share some similarities with systems targeted to the management of music collections, in particular in the content-based processing workflow that allows to organize the audio files. A comprehensive survey on these aspects has been proposed by Casey et al. [4]. We nevertheless believe that the design of the user interface of each system class might benefit from different cues from information visualization and human-computer interaction, and that major progress is still possible in all these areas.

2.1 Research-grade systems

The work presented in [18] underlines that few published research provide accurate usability evaluations on such systems, beyond informal and heuristic ones. The author justifies that this may have occurred because complementary research communities have actually been evolving essentially in separate silos. These include the music information retrieval and the human-computer interaction communities. In that work, 20 systems with auditory display are nevertheless reviewed and compared, including 2 audio browsers that are presented hereafter.

Sonic Browser focused on information visualization [7], and later approached content-based organization through the Marsyas framework [3]. A 2D starfield display allows to map the metadata of audio files to visual variables. Its *HyperTree* view consists in a spring-layout hierarchical graph visualization for browsing the file tree of sound collections. They qualitatively evaluated these views with 15 students through timed tasks and a questionnaire [2]; and their system against the Microsoft Windows 2000 explorer through a think-aloud protocol with 6 students [7].

SoundTorch, the most recent content-based audio browser, has been designed by people aware of audio engineering practices [11]. It relies on Mel-Frequency Cepstral Coefficients (MFCCs) as features, clustered with a Self-Organizing Map (SOM) but initialized with smooth gradients rather than randomly, so that the horizontal axis corresponds to a tonal-to-noisy continuum and the vertical axis to pitch increase / dull-to-bright. In addition to cueing in the variety of content through the position of the nodes corresponding to sounds, *SoundTorch* makes use of the node shape to convey additional information: the temporal evolution of the power of the signal is mapped to a circle.

It is the only related work to provide a quantitative user evaluation. They positively evaluated known- and described-item search tasks comparatively to a list-based application. A dozen of users were involved. However, it is not clear from this comparison whether the approach outperforms the list-based application because of its content-based capabilities, or else because of its interactive abilities (particularly its instant playback of closely-located nodes in the map), or both. Moreover, it has been chosen to randomize the sound list. Sound designers either buy commercial sound libraries that are tagged properly and named accordingly, or else record their own. They also usually spend a significant amount of time to tag these libraries. Therefore, to our opinion, a more realistic baseline for comparison should be a basic ordering by filename.

CataRT is an application developed in the Max/MSP modular dataflow framework, that "mosaices" sounds into small fragments for concatenative synthesis. A 2D scatter plot allows to browse the sound fragments, assigning features to the axes. The authors recently applied a distribution algorithm that optimizes the spreading of the plotted sounds by means of iterative Delaunay triangulation and a mass-spring model, so as to solve the non-uniform density inherent to a scatter plot, and open new perspectives for non-rectangular interfaces such as the circular *reacTable* and complex geometries of physical spaces to sonify. To our knowledge, no user study has yet been published for this tool. It is however claimed as future work [12].

In summary, it appears that no evaluation have been proposed previously on the specific contribution of content-based analysis to the efficiency of sound search. This is a gap we started to address in this work.

2.2 Commercial systems

It is worth mentioning here that some commercial systems, some making use of content-based approaches, have also been proposed, although no quantitative evaluation of those can be found in the literature. A pioneering application is *SoundFisher* by company Muscle Fish [21], start-up of scientists that graduated in the field of audio retrieval. Their application allowed to categorize sounds along several acoustic features (pitch, loudness, brightness, bandwidth, harmonicity) whose variations over time are estimated by average, variance and autocorrelation. Sounds are compared from the Euclidean distance over these features. The browser offers several views: a detail of sound attributes (filename, samplerate, file size...) in a spreadsheet, a tree of categories resulting from classification by example (the user providing a set of sounds), and a scatter plot to sort sounds along one feature per axis.

A second product, *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 creative applications.

A major product, *Soundminer HD* [2], provides a similar interface, plus an alternative layout named *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 technique.

Other companies facilitating creativity such as Adobe with *Bridge* [3] provide more general digital asset management solutions that are accessible through their entire application suite. These focus on production-required capabilities and seem to avoid content-based functionalities.

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.

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

3. OUR SOLUTION

Our system blends knowledge gained from the fields of multimedia information retrieval (content-based organization), human-computer interaction (usability evaluation) and information visualization (visual variables).

3.1 A multimedia information retrieval pipeline

One first step is feature extraction. For sound and music, a large variety of temporal and/or spectral features have been proposed in the literature [4, 15]. We based our features set from [6] since their evaluation considered textural sounds. In short, we used a combination of derivatives of and statistics (standard deviation, skewness and/or kurtosis) over MFCCs and Spectral Flatness (SF). We did not perform segmentation as our test collections contain textural sounds of short length and steady homogeneity.

Another important step is dimension reduction. From our perspective, one of the most promising approach is Stochastic Neighborhood Embedding (SNE) using Student-t distributions (t-SNE) [13]. It has been previously qualitatively evaluated on sound collection visualization [6, 9]. The method has an interesting information retrieval perspective, as it actually aims at probabilistically preserving high-dimensional neighbors in a lower-dimensional projection (2D in our work), and actually maximizes continuity (a measure that can intuitively be related to recall in information retrieval) in the projected space. One emergent result is that recordings from the same sound source with only slight variations are almost always neighbors in the 2D representation, as the recall is high. Another popular but older approach for dimensionality reduction are SOMs. In [14], it has been compared with most recent techniques, and in particular the Neighbor Retrieval Visualizer (NeRV, a generalization of SNE). SOMs produced the most trustworthy (a measure that can intuitively be related to precision in information retrieval) projection but the NeRV was superior in terms of continuity and smoothed recall. As SNE is a special case of NeRV where a tradeoff is set so that only recall is maximized, we infer from those results that SNE is a better approach for our purposes than SOM. Qualitative evaluations of different approaches applied to music retrieval have been undertaken [19]: Multidimensional Scaling (MDS), NeRV and Growing SOMs (GSOM). Users described MDS to result in less positional changes, NeRV to better preserve cluster structures and GSOM to have less overlappings. NeRV and presumably t-SNE seem beneficial in handling cluster structures.

Besides, we propose in this paper an approach to reduce the possible overlappings in t-SNE. An undesirable artifact of the original t-SNE approach however comes from the optimization procedure, which relies on gradient descent with a randomly initialized low-dimensional representation. It creates a stability issue, where several runs of the algorithm may end up in different representations after convergence. This works against the human memory. We thus initialized the low-dimensional representation using the two first axes of a Principal Component Analysis (PCA) of the whole feature set.

3.2 Mapping audio features to visual variables

Displaying such a representation results in a scatter plot or starfield display. We address two shortcomings: 1) clusters of similar sounds might not be salient, and 2) this visualization technique may cause overlap in some areas. *SonicBrowser* [7], that we analyzed in the previous section, and the work of Thomas Grill [9], dedicated to textural sounds, approached the first issue by mapping audio features to visual variables. Ware's book [20] offer great explanations and recommendations to use visual variables to support information visualization tailored for human perception. Thomas Grill's approach was to map many perceptual audio features to many visual variables (position, color, texture, shape), in one-to-one mappings.

3.2.1 Content-based glyphs as sound icons

Grill et al. designed a feature-fledged visualization technique mapping perceptual qualities in textural sounds to visual variables [9]. 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. 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 the Values for all sounds in each collection. A sense of brightness is thus conveyed in both the audio and visual channels through perceptual sharpness and value. We also used the temporal evolution of perceptual sharpness to define a clockwise contour of the nodes, so that sounds of similar average brightness but different temporal evolution could be better discriminated. To compute positions, perceptual sharpness was also added to the feature selection, intuiting it would gather items that are similar visually. The choice of perceptual sharpness was motivated by another work of Grill et al. [10]: they aimed at defining features correlated to perceived characteristics of sounds that can be named or verbalized through *personal constructs*. *High-low*, or brightness of the sound, was the construct the most correlated to an existing feature: perceptual sharpness.

3.2.2 A proximity grid optimizing nearest neighbors

For the removal of clutter in 2D plots, two major approaches exist: reducing the number of items to display, or readjusting the position of items. In our context, we want to display all the items resulting of search queries by tag filtering. For this purpose, we borrow a method initially designed to solve the problem of overlap for content-based image browsing [16]: a proximity grid [1]. Their work is heavily cited respectively for the evaluation of multidimensional scaling techniques [1] and as a pioneering application of usability evaluation for multimedia information retrieval [16], but almost never regarding the proximity grid. To our knowledge, no audio or music browser approached this solution.

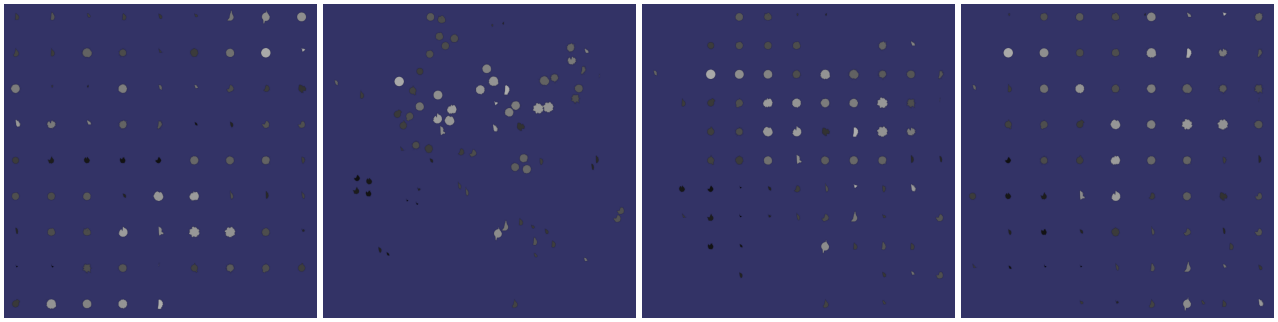

Figure 1. Different layouts with glyphs for the same sound collection filtered by keyword "water", from left to right: "album", "cloud", "metro", and most dense proximity grid.

A proximity grid consists in adapting the coordinates of each item of a 2D plot to magnetize these items on an evenly-distributed grid. Basalaj proposed several variants to compute a proximity grid: greedy methods with spiral search to find empty cells and *empty/swap/bump* strategies to assign items to cells; an *improved greedy* method replacing spiral search by shortest distance estimation; a *"squeaky wheel" optimization* using simulated annealing, and a *genetic algorithm* [1]. We implemented the simplest greedy method with all strategies. To determine the order of the items to assign to cells, we used the fast minimum spanning tree algorithm implementation from the machine learning library *mlpack* of Boruvka's dual-tree based on k-dimensional trees [5]. Applied in high dimension of the audio features, the empty strategy starts with shortest edges while it is the opposite for swap and bump strategies, according to Basalaj. We opted for a simplification: a spiral search always turning clockwise and starting 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 node 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 estimate 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.

For our experiments described in the next section, we prepared the collections by qualitative selection of the optimal grid resolution based on the amounts of horizontal, vertical and 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. Not all collections presented a proximity grid resolution that outperformed a simple grid by filename in terms of neighbor preservation.

4. EXPERIMENTS

4.1 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 when testing machine learning algorithms. It is licensed under a Creative Commons BY license (requiring attribution). It contains 8458 sound samples, 90 sublibraries combine diverse types of content or specialize into one type, among which: musical instruments riffs or single notes, field recordings, Foley recording, synthesized sounds, vocals, animal sounds. 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, for instance *Metal on the ground [n]* with n varying from 1 to 4. 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, similarly to a standard file browser, will thus imprint local neighborhoods to the list.

4.2 Evaluation method

We chose to perform a qualitative and quantitative evaluation: qualitative through a feedback questionnaire, quantitative through known-item search tasks as popularized recently for video browsers by the Video Browser Showdown [17]. In the context of audio browsers, for each task the target sound is heard, the user has to find it back as fast as possible using a given layout. Font's thesis compared layouts for sound browsing: *automatic* (PCA), *direct mapping* (scatter plot) and *random map* [8]. Time and speeds were deliberately not investigated, claiming that people employ different search behaviors.

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

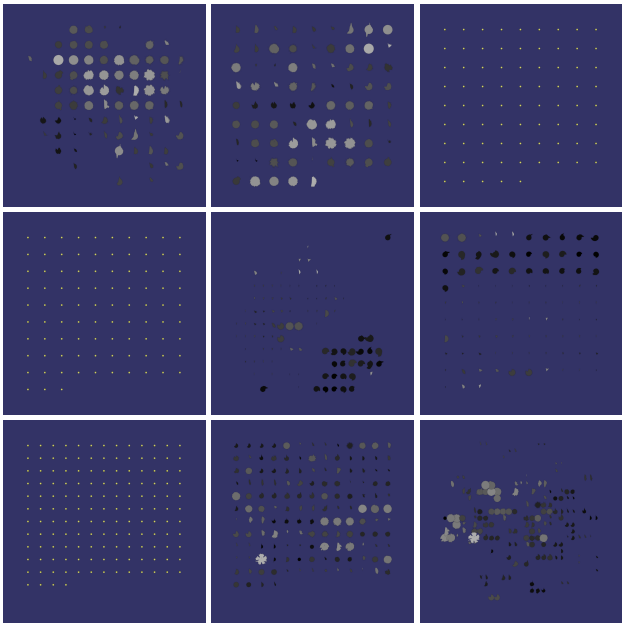

Figure 2. Sequence of tasks in the last experiment. In rows subsets of the One Laptop Per Child (OLPC) sound library filtered by keyword, respectively "water", "spring", "metal". In columns: permutations of layouts.

4.3 Design

We undertook four experiments: the first comparing *grid* and glyph-less *cloud* layouts motivated us to add glyph representations (*cloud* was outperformed), the second and third confirmed that a proximity grid was to be investigated (*cloud* still outperformed), the last validated these choices. We recorded several metrics (success times, pointer distances and speeds, audio hovers) and ratings from feedback questionnaires. Here we only report the last experiment and only analyze times taken to successfully find targets.

The fourth experiment was designed as a within-subject summative evaluation. Figure 2 shows the exact sequence of tasks presented to the users. An additional collection was used for training tasks with each layout.

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 land 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).

4.4 Participants and apparatus

16 participants (5 female) of mean age 28 (+/- 6.3) each performed 9 tasks on 3 different collections. Besides 2 subjects, all the participants have studied or taught audiovisual communication practices (sound design, film edition). They were asked which human sense they favored in their work (if not, daily) on a 5-point Likert scale, 1 for audition to 5 for vision: on average 3.56 (+/- 0.60). All self-rated themselves with normal audition, 10 with corrected vision.

We used an Apple Macbook Pro Late 2013 laptop with 15-inch Retina display, with a RME FireFace UCX sound card, and a pair of Genelec active loudspeakers. A 3Dconnexion Space Navigator 3D mouse was repurposed into a buzzer to submit targets hovered by the touchpad, with audio feedback of the closest node to the pointer.

4.5 Results

A one-way ANOVA shows that there is a quite significant difference between views within subjects on success times (p=.02), more on self-reported ratings of efficiency (p<.001) and pleasurability (p<.001). Mean and standard deviations are compared in table 1. A Tukey multiple comparisons of success times means at a 95% family-wise confidence level on layouts shows that *metro* outperformed *grid* (p=.01), but *album* was not significantly better than *grid* (p=.34) or worse than *metro* (p=.26).

	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 1. Mean (standard deviations) of evaluation metrics

4.6 Discussion

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 only content-based icons and simple ordering by filename. A content-based layout comes at a greater computational cost but brings significant improvements.

5. CONCLUSION, FUTURE WORKS

We proposed 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. Through a usability evaluation of known-item search tasks, we showed that this solution was more efficient and pleasurable than a grid of sounds ordered by filenames.

An improvement to this method would require to investigate all blocks from the multimedia information retrieval data flow. First, other features tailored for sound effects should be tried. Second, we have noticed that some of the first high-dimensional nearest neighbors are positioned quite far away in 2D, already past dimension reduction. Reducing pairwise distance preservation errors may be an investigation track.

6. 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. This work has been partly funded by the Walloon Region of Belgium through GreenTIC grant SONIXTRIP.

7. REFERENCES

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

[2] Eoin Brazil. Investigation of multiple visualisation techniques and dynamic queries in conjunction with direct sonification to support the browsing of audio resources. Master's thesis, Interaction Design Centre, Dept. of Computer Science & Information Systems University of Limerick, 2003.

[3] Eoin Brazil, Mikael Fernström, George Tzanetakis, and Perry Cook. Enhancing sonic browsing using audio information retrieval. In *Proceedings of the International Conference on Auditory Display (ICAD)*, 2002.

[4] 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 *Proceedings of the IEEE*, volume 96, 2008.

[5] Ryan R. Curtin, James R. Cline, Neil P. Slagle, William B. March, P. Ram, Nishant A. Mehta, and Alexander G. Gray. MLPACK: A scalable C++ machine learning library. *Journal of Machine Learning Research*, 14:801–805, 2013.

[6] Stéphane Dupont, Thierry Ravet, Cécile Picard-Limpens, and Christian Frisson. Nonlinear dimensionality reduction approaches applied to music and textural sounds. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2013.

[7] Mikael Fernström and Eoin Brazil. Sonic browsing: An auditory tool for multimedia asset management. In *Proceedings of the 2001 International Conference on Auditory Display*, 2001.

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

[9] Thomas Grill and Arthur Flexer. Visualization of perceptual qualities in textural sounds. In *Proceedings of the Intl. Computer Music Conference*, ICMC, 2012.

[10] Thomas Grill, Arthur Flexer, and Stuart Cunningham. Identification of perceptual qualities in textural sounds using the repertory grid method. In *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*, ACM, 2011.

[11] Sebastian Heise, Michael Hlatky, and Jörn Loviscach. Soundtorch: Quick browsing in large audio collections. In *125th Audio Engineering Society Convention*, 2008.

[12] Ianis Lallemand and Diemo Schwarz. Interaction-optimized sound database representation. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, 2011.

[13] Joshua M. Lewis, Laurens van der Maaten, and Virginia de Sa. A behavioral investigation of dimensionality reduction. In N. Miyake, D. Peebles, and R. P. Cooper, editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012.

[14] Kristian Nybo, Jarkko Venna, and Samuel Kaski. The self-organizing map as a visual neighbor retrieval method. In *Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM)*, 2007.

[15] G. Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proc. of the Intl. Symposium on Music Information Retrieval (ISMIR)*, 2007.

[16] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI. ACM, 2001.

[17] Klaus Schoeffmann, David Ahlström, Werner Bailer, Claudiu Cobârzan, Frank Hopfgartner, Kevin McGuinness, Cathal Gurrin, Christian Frisson, Duy-Dinh Le, Manfred Fabro, Hongliang Bai, and Wolfgang Weiss. The video browser showdown: a live evaluation of interactive video search tools. *International Journal of Multimedia Information Retrieval*, pages 1–15, 2013.

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

[19] Sebastian Stober, Thomas Low, Tatiana Gossen, and Andreas Nürnberger. Incremental visualization of growing music collections. In *Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR)*, 2013.

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

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