Semantic browsing of sound databases without keywords

Grégoire Lafay[1], **Nicolas Misdarris**[2], **Mathieu Lagrange**[1], **Mathias Rossignol**[2]
(gregoire.lafay@irccyn.ec-nantes.fr)

IRCCyN, Ecole Centrale de Nantes [1] *IRCAM, STMS, UPMC* [2]

n this paper, we study the relevance of a semantic organization of sounds for easing the browsing of a sound database. For such task, the semantic organization is traditionally hold thanks to a keyword selection process. However, various issues of written language like word polysemy, ambiguities translation issues may bias the browsing process.

We consider here two display of sounds that are organized by considering underlying semantic information that draws a hierarchy. For the sake of comparison, we also consider a display whose organization of sound in the plane is based on acoustic cues. Those three displays are evaluated in terms of search speed in a crowd sourcing experiment. Results

Audio content management and display, Semantic sound data mining

0 Introduction

With the growing capability of recording and storage, the problem of indexing large databases of audio has recently been the object of much attention [18]. Most of that effort is dedicated to the automatic inference of indexing metadata from the actual audio recording [19, 17]; in contrast, the ability to browse such databases in an effective manner has been less considered.

Most media assets management are based on keyword-driven queries. The user enters a word which characterizes the desired item, and the interface presents him with items related to this word. The effectiveness of this principle is primarily based on the typological structure and nomenclature of the database. However, for databases and more specifically for databases of sounds, issues arise:

1. Sounds, as many others things, can be described in many ways. Sound may be designated by their sources (a car door), as well as by the action of those sources (the slamming of a car door) or their environments (slamming a car door in a garage) [8, 11, 2]. Designing an effective keyword-based search system requires an accurate description of each sound, which has to be adaptable to the sound representation of each user.
2. Pre-defined verbal descriptions of the sounds made available to the users may potentially bias their browsing and final selection.
3. Localization of the query interface is made difficult as the translation of some words referring to qualitative aspects of the sound such as its ambiance is notoriously ambiguous and subject to cultural specificities.
4. Unless considerable time and resources are invested into developing a multilingual interface, any system based on verbal descriptions can only be used with reduced performance by non-native speakers of the chosen language.

To circumvent those issues, not relying on keywords is desirable. That said, semantics, which is traditionally conveyed through written text, may be important for easing browsing.

To study this issue, we consider in this paper several means of displaying sounds without relying on any textual representation. One, considered as a reference baseline, do not rely on any semantic information as the sounds are organized according to their acoustical properties, *i.e.* time averaged spectral features. The second one

Their effectiveness is studied with a search-based task whose aim is find a target sound by browsing the database using the display under evaluation.

The paper is organized as follows: in Section 1 previous work on the topic of sound database browsing is reviewed. Then, the corpus used in this study is described in Section 2. the three displays under evaluation are next described in Section 3. The crowd sourcing test used to compare those displays is presented in Section 4. the outcomes of this experiment are discussed in Section 5.

1 Previous work

Most of the research effort in Sound Design and Music Information Retrieval (MIR) communities is focused on acoustical based indexing and browsing [16, 15]. Typically, the items (sound effects or pieces of music) are modeled by processing the digital audio waveform using some signal processing pipeline in order to get a compact description for each items [5]. Then, statistical projections or embedding techniques, like Principal components analysis (PCA), Multi Dimensional Scaling (MDS) [13, 4], Self Organizing Maps (SOM) [12] and the like are used to project the items in a two or three dimensional space while preserving as much as possible distances among items. One of the advantage of such an approach is its ability to scale to very large databases [14] as it does not need any kind of manual annotations and allows to search by similarity efficiently according to acoustical properties.

Though, such acoustical models are inherently subject to observation noise and biases. Selecting the most relevant features to achieve the correct projection of the data may only be performed by an expert user. If done *a priori* by an expert or by some above cited dimensionality reduction technique, the induced bias can strongly limit the user in its ability to . In that respect, semantic tags, if available for the data at hand have the advantage of implicitly structuring the similarity space, thus eventually easing the browsing process even if the actual tags are not – as in this study – exposed to the user.

2 Dataset

The interface framework requires a pre-organization of the sound dataset it has to display. This organization is based on semantic considerations. A sound is characterized by a tag describing the source of the sound (*man-yelling*, *car-passing*). Sounds are then grouped into classes according to their tags (*car* > *car-passing*; car > *car-starting*). Those classes are in turn packed into classes until high level classes describing broad concepts are reached (*traffic* > *car* > *car-passing*). The sound dataset is organized into a hierarchical structure of semantic classes as described in Figure 1. The different levels of this hierarchy are called semantic levels. Strictly speaking, the sound samples of the dataset are the leaf semantic levels. All the other classes are represented by a *prototype sound* that best characterizes the sounds belonging to the class.

That description implies that there are as many leaf classes as sounds in the dataset, which would be unrealistic for large datasets. We therefore propose, in that case, to adapt the organization by considering the leaf classes as *collections* of semantically similar sound samples. Thus, two sounds of *male-yelling* would be grouped into a single leaf class with the tag *male-yelling*. The leaf class would then also have a *prototype sound* being the most representative item of the different *male-yelling* sounds belonging to the leaf class.

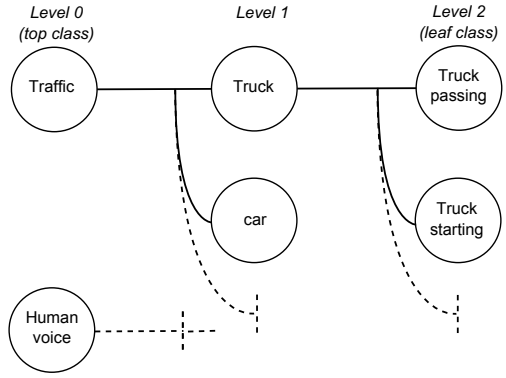

Fig. 1. Semantic hierarchical structure of the dataset of urban environmental sounds

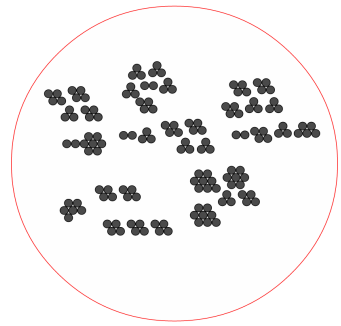

Fig. 2. Full Display (FD) with a non visible hierarchical organization of semantic classes

In order for the semantic hierarchical structure to be perceptually valid, the *tags* describing the classes were chosen from the names of sound categories found by studies addressing environmental auditory scenes perception [11, 2, 6]. In cognitive psychology, sound categories may be regarded as intermediaries between collective sound representations and individual sensory experiences [6]. It is our belief that using such category names to build the hierarchical structure makes the latter perceptually motivated, and thus meaningful for the users.

3 Displays

In this section, two listening oriented displays, called respectively Progressive Display (PD) and Full Display (FD) are presented. Both interfaces 1) allow users to explore a sound dataset without any written textual help and 2) base their display upon the hierarchical structure of the dataset. The two interfaces have been designed in order to see whether a progressive top-down display of the hierarchical structure helps the user explore the dataset.

As shown on Figures 2 and 3, both interfaces are based on the same principle of distributing in a 2D space the

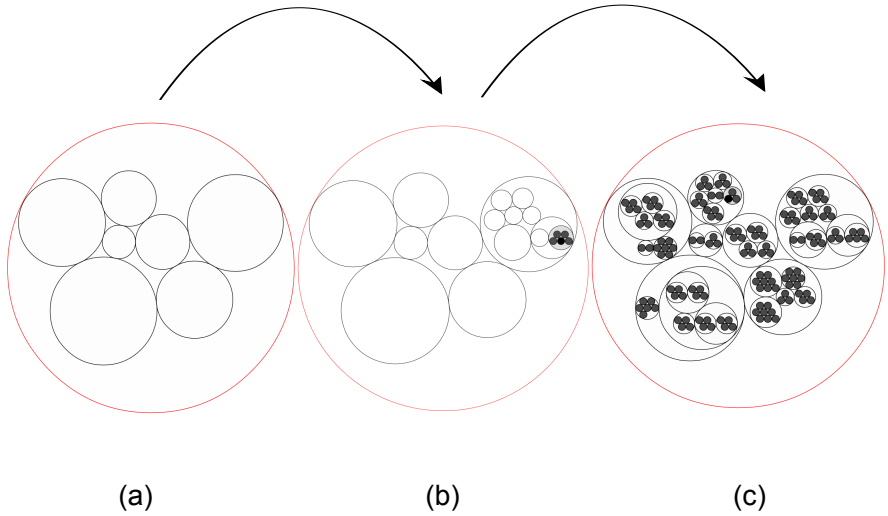

Fig. 3. Progressive Display (PD) with a visible hierarchical organization of semantic classes: (a) initial folded version; (b) partly folded version; (c) unfolded version

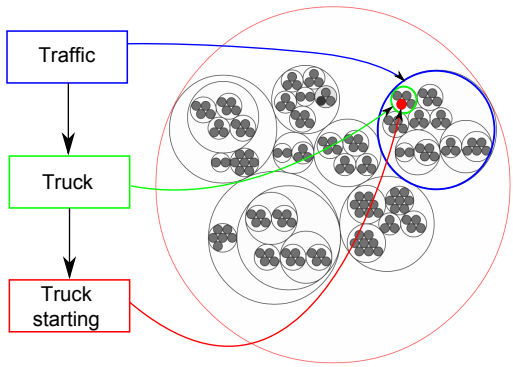

Fig. 4. Spatial configuration of the Progressive Display (PD) based on the semantic hierarchical structure of the dataset

hierarchical structure of the sound elements of the dataset. Each sound class is represented by a circle. Circles are packed together according to the hierarchical semantic organization of the dataset, as shown on Figure 4. Thus, subclasses belonging to the same class are close to each others. Circle packing functions of the D3.js (Data-Driven Documents) javascript library [1] are used to distribute the sound classes in the space.

The way in which a user visualizes the hierarchical organization varies with the interface:

- *PD*: users have access to the intermediate semantic levels of the hierarchy. Upon first using PD, they observe circles representing the top classes of the semantic hierarchical structure of the dataset. When users click on a circle, they hear the sound prototype of the class and the subclasses are progressively revealed, represented by white circles contained in the circle that has just been clicked. The same action repeats itself until the user reaches the leaf classes of the hierarchy. The leaf classes are represented with small gray circles, indicating that there is no subclass to discover. Thus the PD has a constrained exploration system. When a user click on a circle, sub-circles are automatically revealed to him in a gradual way. Each time a sub-circle is automatically revealed, its sound prototype is played. Users may stop the discovery process by clicking on an other circle.
- *FD*: users can directly visualize the whole hierarchy, down to the leaf classes. Those leaf classes are distributed in the same manner as PD. In that sense, the spatial configuration of the unfolded version of PD, which may be obtained after discovering all the classes and subclasses, is similar to that of FD, as shown on Figures 2, 3 and 4.

4 Experiment

4.1 Objective

During this test, three interfaces are compared:

- PD, which provides a visible hierarchical organization of semantic classes;
- FD, which provides a non-visible hierarchical organization of semantic classes;
- an Acoustic Display (AD) providing a 2D representation based on acoustic descriptors. In this case, the spatial configuration is computed by 1) describing the sounds with mel-frequency cepstrum coefficients (MFCCs) and 2) using a non metric multidimensional scaling with Kruskal's normalized stress to compute sound positions

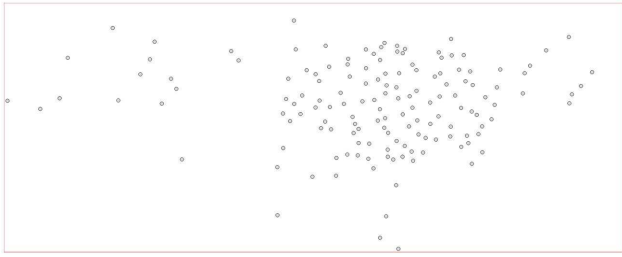

Fig. 5. The Acoustical Display (AD) computed using a non metric multidimensional scaling on MFCCs based acoustic descriptors

in a 2D space. An example of AD can be seen on Figure 5.

By comparing PD and FD, the effect of a visible hierarchy on the user is investigated. The goal is to check if forcing the user to browse the high levels of the hierarchy helps him to understand and learn the spatial configuration and the organisation of the sound classes. By confronting PD and FD with AD, the relative efficiencies of semantic based and acoustic based spatial configurations are compared.

4.2 Experimental protocol

We choose to test and compare the three displays through a crowdsourcing experiment. Here is the link to access the experiment web page [1].

Subjects are asked to successively retrieve 13 target sounds in a dataset of 149 urban environmental sound events. The target sounds are distributed such as there are at least two target sounds in each top-level class of the semantic hierarchical structure of the dataset. To minimize order effects, target sounds are randomly presented to each subject.

To listen to a target sound, the subject has to click a *Play target sound* button. Subject may replay the target sound as many times as they like. A timer is started when the subject clicks on a circle for the first time.

When the target sound is found, subject puts an end to its search by clicking on the *Click if found* button. This action 1) stops the timer and 2) loads a new target sound. If the subject does not find the correct target sound, an error message appears, and the experiment continues.

During the experiment, two indications are communicated to the subject:

- *The research state*: "pause" if the subject is currently not looking for a target sound (at the beginning of the experiment, or between two target sounds); "in progress" if the subject is currently looking for a target sound.
- *Remaining target sounds*: the number of target sounds which remain to find.

[1] http://217.70.189.118/soundthings/speedSoundFinding/

Table 1. Criterion used to detect outliers

criterion	1	2	3	4
outlier 1 (PD)	—	—	—	x
outlier 2 (PD)	x	—	—	—
outlier 3 (AD)	—	—	—	x
outlier 4 (AD)	x	x	x	—
outlier 5 (AD)	x	—	—	—

The experiment ends when all the target sounds have been found.

It is most important to note that PD do not pack up at each target sound search. When a circle is revealed, it remains visible during the whole experiment.

4.3 Data Collection

Three sets of data are collected during the experiment:

- the total duration of the entire experiment. It includes breaks between two target sound searches and it is called the *absolute duration*.
- the duration of each search. The sum of the 13 search durations, which is the *absolute duration* minus the break times between two target sound searches, is called the *total search duration*.
- the name of each sound which has been heard.
- the time at which each sound has been heard.

4.4 Apparatus

A crowd sourcing approach has been adopted. The experiment was designed to be supported by the *chrome-browser* web navigator. The link to the experiment has been sent to the subjects via three mailing list being *music-ir, auditory* and *uuu-IRCAM* (internal IRCAM mailing list). Subjects were allowed to perform the experiment once, and on one interface only. Data were automatically collected server-side at the end of the experiment. Subjects were asked to use headphones. All the presented sounds were normalized to the same root mean square (RMS) level.

4.5 Participants

60 subjects have completed the experiment, 20 for each interface.

4.6 Statistical analysis

We use one-way ANOVA and two samples t tests at the 5% significance level to test for statistical significance. All statistical analysis are performed using native functions of the computing environment MATLAB.

5 Results

5.1 Outlier detection

Outlier detection is an important step of any crowdsourcing experiment as experimenters do not control the experimental environment in which the

subjects perform the experiment [10][3]. A widely used method to detect outlier in human-computer interaction studies is to consider as outlier an observation which deviates of at least ± 2 standard deviation (STD) from the average [10]. As this method is not robust to the presence of isolated extreme observations (as it is often the case for crowdsourcing experiment), we follow the method proposed by proposed by [10] and used the inter-quartile range (IQR). With this approach, an observation is considered to be an outlier if its value is more than $3 * IQR$ higher than the third quartile or more than $3 * IQR$ lesser than the first quartile. This methods is applied to the four following criterion :

1. the *absolute duration*: to detect subjects who took abnormally long breaks between searches.
2. the *total search duration*: to detect subjects who took an abnormally long time to complete the experiment.
3. number of heard sounds: to detect subjects who had to hear an abnormally high number of sounds to complete the experiment.
4. the maximum number of times a target sound had to be heard before being found: to detect subjects who had difficulty to recognize the target sounds.

Using the IQR method, 5 subjects are detected as outliers and removed from the analysis. 2 subjects used PD and 3 subjects AD. The tables 1 show what criterion are use to detect outliers. 3 subjects are detected by observing the *absolute duration*. They spent respectively 50 minutes, 5 hours and 17.5 hours to complete the experiment. 1 subject is detected by observing the total search duration (46 minutes), and an other by observing the total number of heard sounds (1800 heard sounds, roughly 12 times the total size of the corpus). And lastly 2 subjects are detected by observing the maximum number of times they had to hear a target sounds before finding it (4 and 8 times).

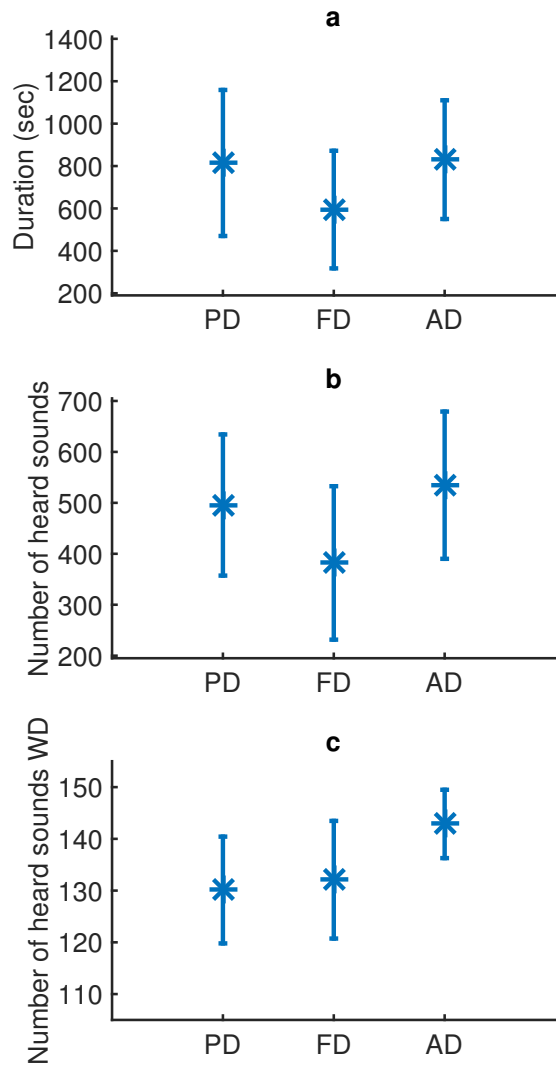

Fig. 6. Average and standard deviation for (a) the total search durations, (b) the total search durations and (c) the total search durations

5.2 Interface efficiencies

To characterize the interface efficiencies, three set of collected data are assessed:

- the *total search duration*
- the number of heard sounds
- the number of heard sounds without duplication. By "without duplication" we mean that, if a same sound prototype is heard 10 times during the 13 searches of the experiment, it counts only for one.

The two first data help us qualify the notion of efficiency by considering the time and the number of clicks needed to achieve the task (ie. reach the target). The goal for those values is to be as low as possible. The third data allows us to measure the selectivity of the interfaces. A low number of heard sounds without duplication indicates that subjects understood the spatial organisation of the dataset, and use this knowledge to improve their searches. In contrary, a high number of heard sounds without duplication suggest that the subject did not understood the way circles are organized in space, and tends to play all the sounds at each search. The maximum number of heard sounds without duplication is the corpus size: 149 sounds.

The type of interface have a significant effect on the the *total search durations* ($F[2,52] = 3.63; p = 0.03$), the number of heard sounds ($F[2,52] = 5.65; p < 0.01$) and the number of heard sounds without duplication ($F[2,52] = 8.65; p < 0.01$).

Figure 6 (a) shows the average and standard deviations of the *total search durations* for the three interfaces. FD seems to perform better than the other interfaces (FD-PD: $p = 0.04$; FD-AD: $p = 0.02$), whereas PD and AD seem to have similar results (PD-AD: $p = 0.88$).

We found similar results for the numbers of heard sounds (Figure 6 (b)). FD significantly outperforms the other interfaces (FD-PD: $p = 0.02$; FD-AD: $p < 0.01$),

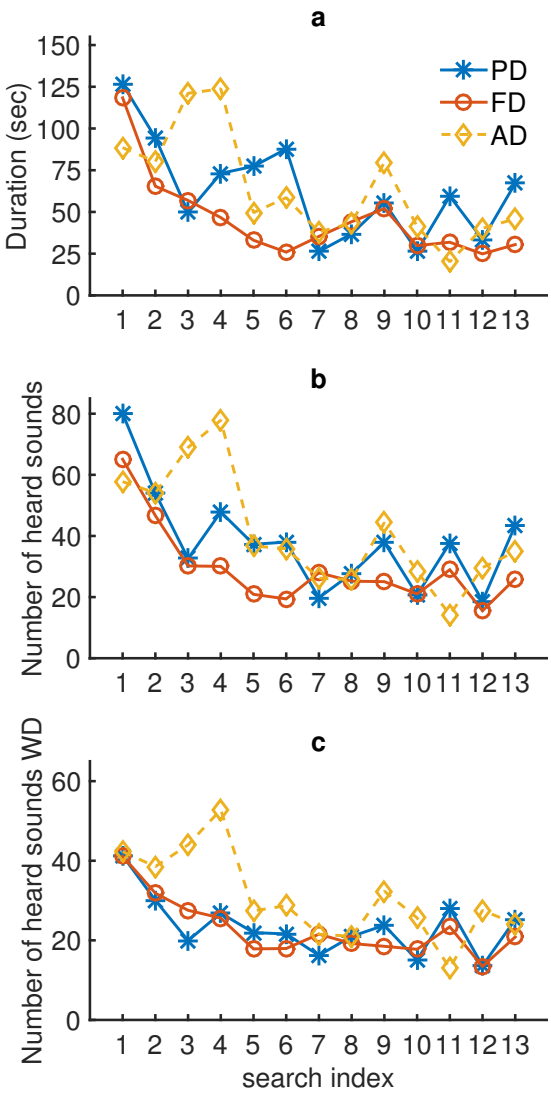

Fig. 7. Medians of (top) the numbers of heard sounds at each target sound search, (middle) the numbers of heard sounds without duplication (WD) at each target sound search, (bottom) the relative durations at each target sound search

whereas PD and AD show similar outcomes (PD-AD: $p = 0.42$).

Figure 6 (c) shows the results for the number of heard sounds without duplication. This time the results of AD are significantly lower than those of both PD and FD (AD-PD: $p < 0.01$; AD-FD: $p < 0.01$). For AD, 75% of the subjects heard more than 138 sounds, and 25% heard 148 sounds, that is almost the entire database. Considering PD, 75% of subjects heard less than 133 sounds, against 144 for FD. This time PD and FP seem to perform equally (PD-FD: $p = 0.58$)

According to those results, a hierarchical organization of the dataset based on semantic values (PD and FD) allows users to retrieve the 13 target sounds 1) quicker, and 2) by listening to a smaller amount of sounds than an organization based on acoustic descriptors (AD). But those two effects are significantly compromised when users have to parse the entire hierarchy to reach the first target sound, as it the case for PD. It seems that imposing a graphical representation of the hierarchy disturbs or confuses the user instead of allowing him to learn the spatial organization of the classes.

5.3 Learning phenomenon

We now study if and how users progressively acquire knowledge about the spatial organization of the classes. To do that, variations of the data over the searches are assessed. Three sets of collected data are used:

- the duration of each target sound search
- the number of heard sounds for each target sound search
- the number of heard sounds for each target sound search without duplication.

Figure 7 (a) shows the evolution of the average durations of each target sound search observed for PD, FD and AD. It is interesting to note that both for PD and FD, the maximum value is observed for the first search, whereas it is observed for the fourth search for AD. Moreover if the curve profiles of PD and FD seem to progressively decrease and are very similar, the one of AD is much more irregular. Similar results are found observing the evolution of the average numbers of heard sounds with and without duplication (Figures 7 (b) and (c))

These results tend to indicate that PD and FD facilitate the learning of the spatial configuration, as the search durations and the numbers of heard sounds at each search decrease over time. Although curves for PD and FD have similar profiles, FD seem to better perform as users of FD where able to find the target sounds faster by clicking on fewer circles.

6 Conclusion

In this paper, two displays allowing users to explore a sound dataset without written textual help are presented. The interfaces distribute sounds represented by circles on a 2D space. The spatial organisation is driven by semantic features. The two graphical displays are assessed and compared to a third listening based interface in which spatial configuration depends upon acoustic features. The tests consist in data retrieval tasks. The Full-Display (FD), that allows users to directly visualize the leaf classes of the semantic hierarchical structure, proves to be the most effective interface for the task.

Two main conclusions may be derived from this experiment. First, a spatial configuration based on semantic features is more effective to retrieve target sounds than a spatial configuration based on acoustic features. Second, an imposed visualisation of the semantic hierarchical structure of the dataset does not help user to understand and learn the spatial configuration of the semantic class, but instead disturbs the navigation.

7 Acknowledgements

Research project partly funded by ANR-11-JS03-005-01.

8 REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[2] A. Brown, J. Kang, and T. Gjestland. Towards standardization in soundscape preference assessment. *Applied Acoustics*, 72(6):387–392, 2011.

[3] S. Buchholz and J. Latorre. Crowdsourcing preference tests, and how to detect cheating. In *INTERSPEECH*, pages 3053–3056, 2011.

[4] P. Cano, M. Kaltenbrunner, F. Gouyon, and E. Batlle. On the Use of FastMap for Audio Retrieval and Browsing. *ISMIR*, 2002.

[5] G. Coleman. Mused: Navigating the personal sample library. *Proc. ICMC, Copenhagen, Denmark*, 2007.

[6] D. Dubois, C. Guastavino, and M. Raimbault. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta Acustica united with Acustica*, 92(6):865–874, 2006.

[7] J. D. Gibbons and S. Chakraborti. *Nonparametric statistical inference*. Springer, 2011.

[8] O. Houix, G. Lemaitre, N. Misdariis, P. Susini, and I. Urdapilleta. A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52, 2012.

[9] M. Kobayashi and C. Schmandt. Dynamic Soundscape: mapping time to space for audio browsing. *ACM SIGCHI Conference*, 8, 1997.

[10] S. Komarov, K. Reinecke, and K. Z. Gajos. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 207–216. ACM, 2013.

[11] M. Niessen, C. Cance, and D. Dubois. Categories for soundscape: toward a hybrid classification. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 2010, pages 5816–5829. Institute of Noise Control Engineering, 2010.

[12] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. *Computer Music Journal*, 28(2):49–62, 2004.

[13] D. Schwarz and N. Schnell. Sound search by content-based navigation in large databases. In *Sound and Music Computing (SMC)*, pages 1–1, 2009.

[14] D. Schwarz, N. Schnell, and S. Gulluni. Scalability in content-based navigation of sound databases. In *International Computer Music Conference (ICMC)*, pages 1–1, 2009.

[15] S. Streich and B. S. Ong. A music loop explorer system. In *Proceedings of the International Computer Music Conference (ICMC), Belfast, Northern Ireland*, 2008.

[16] G. Tzanetakis. Musescape: An interactive content-aware music browser. In *Proc. Conference on Digital Audio Effects (DAFX)*, 2003.

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

[18] E. Wold, T. Blum, D. Keislar, and J. Wheaten. Content based retrieval of audio. *IEEE Multimedia*, 1996.

[19] T. Zhang and C. Kuo. Hierarchical classification of audio data for archiving and retrieving. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pages 1–4, 1999.

THE AUTHORS