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]

In 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 held thanks to a keyword selection process. However, various issues with written language like word polysemy, ambiguities translation issues may bias the browsing process.

We consider here displays of sounds that are organized by considering the 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 demonstrate the usefulness of the use of implicit semantic organization to display sounds both in terms of search speed and efficiency of learning.

Audio content management and display, Semantic sound data mining

0 Introduction

With the growing capability of recording and storage devices, the problem of indexing large databases of audio has recently been the object of much attention [16]. Most of that effort is dedicated to the automatic inference of indexing metadata from the actual audio recording [17, 15]; 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 best 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, several 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) [7, 9, 2]. Designing an effective keyword-based search system requires an accurate description of each sound, which has to be tunable 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.

Thus, we consider in this paper several means of displaying sounds without relying on any textual representation. The first 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 and third displays have their spatial organization based on the on a predefined hierarchical semantic organization. And those last two displays differs by huw the semantic is used.

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

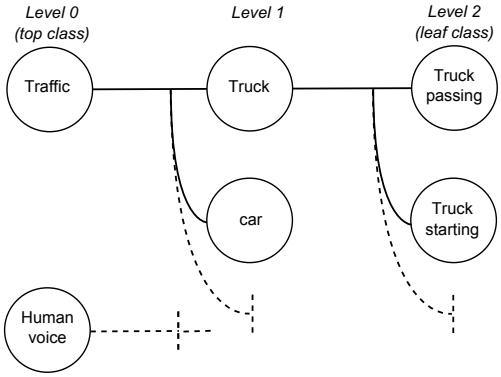

Fig. 1. Semantic hierarchical structure of the dataset of urban environmental sounds

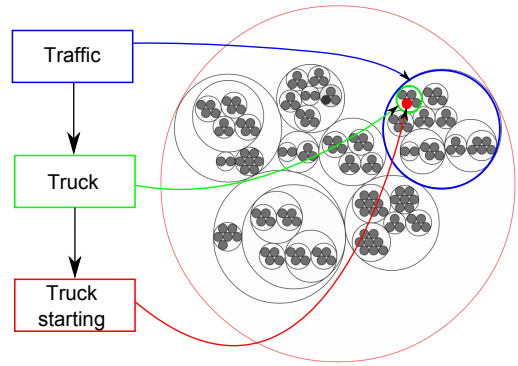

Fig. 2. Semantic displays are based on the semantic hierarchical structure of the dataset

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 [14, 13]. 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) [11, 4], Self Organizing Maps (SOM) [10] 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 [12] as it does not need any kind of manual annotations and allows the user 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 access the data searched for. 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 – explicitly exposed to the user.

2 Dataset

The dataset considered in this study is composed of 149 urban environmental sound events. It has a semantic organization that is as follows: a sound is characterized by a tag describing the physical source of the sound (*man-yelling*, *car-passing*). Sounds are then hierarchically 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 broader concepts are reached (*traffic* > *car* > *car-passing*). The sound dataset is thus organized into a hierarchical structure of semantic classes as described in Figure 1. 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.

In order for the semantic hierarchical structure to be perceptually valid, the *tags* that describe the classes are chosen from sound categories presented in studies addressing environmental auditory scenes perception [9, 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 can thus help the users for efficiently browsing the dataset.

3 Displays

The aim of the displays described in this section is to allow the user to efficiently browse the above presented dataset of sounds explore without any written textual help. For each display, a sound is graphically represented by a filled circle whose, once clicked, plays the actual sound. The spatial organization is specific to each display.

As a reference, an Acoustic Display (AD) provides a spatial representation based on acoustic descriptors. For describing each sound, Mel-Frequency Cepstrum Coefficients (MFCCs) are computed with standard parameter settings (13 lowest quefrency coefficients are

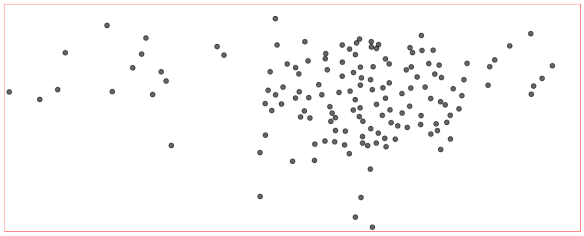

Fig. 3. Acoustical Display (AD)

kept out of 40). The Euclidean distance between time averaged features is then computed between each sounds. A non metric multidimensional scaling with Kruskal's normalized stress is then considered to project the data into two dimensions according to this distance, see Figure 3.

Two semantically oriented displays are then proposed, called respectively Semantic Display (FSD) and Progressive Semantic Display (PSD), respecively shown on Figures 4 and 5. Both display consider the hierarchical structure of the dataset to organize sounds. Each sound class is represented by a filled circle. Circles are then packed together according to the hierarchical semantic organization of the dataset, as shown on Figure 2. 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. Depending on the display, the user can whether acces directly to each leaf class or has to go through intermediate levels of the hierarchy.

More precisely with the FSD display, users can directly visualize the whole hierarchy, down to the leaf classes. On contrary, with the PSD display, 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. In this display, the leaf classes are distributed in the same manner as FSD. Thus, the spatial configuration of the unfolded version of PSD, which may be obtained after discovering all the classes and subclasses, is equivalent to the one of FSD, as shown on Figures 4, 5 and 2.

The latter interface is introduced in order to study if a progressive top-down display of the hierarchical structure helps the user explore and understand the dataset hierarchy.

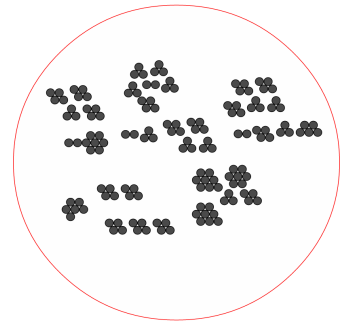

Fig. 4. Full Semantic Display (FSD): implicit hierarchical organization of semantic classes

4 Experiment

4.1 Objective

During this experiment, the three displays are compared. By comparing AD and xSD (the two semantic displays), the goal is to evaluate the relative efficiencies of semantic based and acoustic based spatial configurations. By comparing FSD and PSD, we study the impact of an enforced hierarchical exploration of the corpus by the user. The goal is to check if constraining the user to first browse the high levels of the hierarchy helps him to grasp and memorize the spatial configuration and the organization of the sound classes.

4.2 Experimental protocol

We evaluate and compare those three displays with a crowdsourcing web-based experiment [1] for which we adopt a between subject approach, *i.e.* a subject can only test one display. In this experiment, subjects are asked to successively retrieve 13 target sounds among the 149 of the whole dataset. The target sounds are selected such as there are at least two target sounds in each top-level class of the semantic hierarchical structure of the dataset. To reduce order effects, target sounds are randomly presented to each subject.

First, the subject click a *Play target sound* button to listen to the target sound. Then, the subject may replay the target sound as many times as he likes. A timer starts when the subject clicks on a circle for the first time. When the target sound is found, subject puts an end to the 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

[1] The test is available at `http://soundthings.org/research/speedSoundFinding/index.html`, each of the 3 displays is also available by appending a digit from 1 to 3 to *index* in the above cited url.

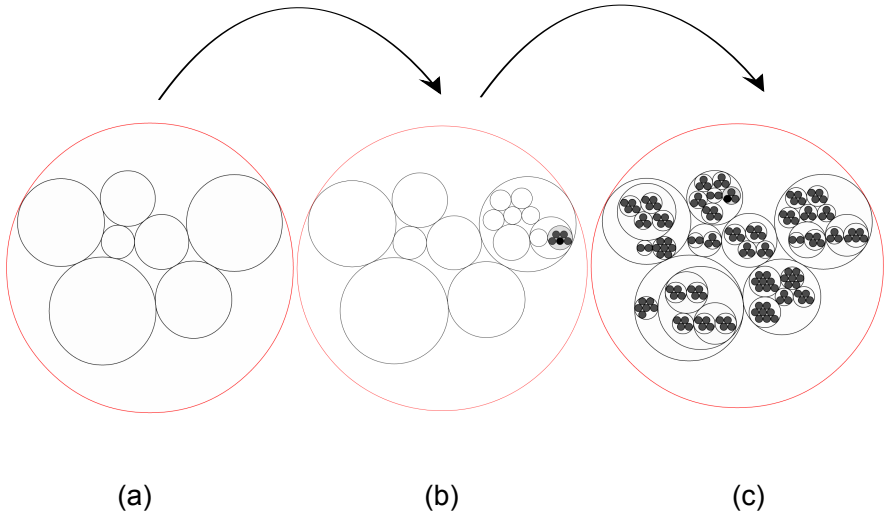

Fig. 5. Progressive Semantic Display (PSD): explicit hierarchical organization of semantic classes. (a) initial folded version; (b) partly folded version; (c) unfolded version

message appears, and the experiment goes on with the same target sound.

During the experiment, two visual cues are shown to the subject:

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

The experiment ends when all the target sounds have been found. One shall notice that the PSD display is not reintialized at each change of the target sound. Thus, when a circle is unfolded, it remains visible during the whole experiment.

4.3 Data Collection

Four types 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 crowdsourcing approach is adopted. The experiment is designed to be performed using the *chrome* web browser. The link to the experiment has been sent to three mailing lists, namely *Music-Ir, Auditory* and an internal IRCAM mailing list. Subjects are only 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 sounds of the dataset are 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 environment in which the subjects perform the experiment [8, 3]. A commonly used method to detect outliers in human-computer interaction studies is to consider as outlier an observation which deviates of at least ± 2 standard deviation from the average [8]. though, as this latter method is not robust to the presence of

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	—	—	—

isolated extreme observations (as it is often the case for crowdsourcing experiment), we follow the method proposed by [8] 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 in this study to the four following criteria :

1. the *absolute duration*: to detect subjects who took abnormally long breaks between searches.
2. the *total search duration*: to detect subjects who spent 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 the criteria used to detect the 5 outliers. 3 subjects are detected by considering the *absolute duration*. They spent respectively 50 minutes, 5 hours and 17.5 hours to complete the experiment. 1 subject is detected by considering 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).

5.2 Efficiency

To study the relative efficiency of the three displays, three metrics are considered:

- the *total search duration*
- the number of sounds heard Mathias : on met heard avant ou apres sounds ?
- the number of unique sounds heard. By ”unique” we mean that, if a same sound is heard 10 times during the 13 searches of the experiment, it counts only for one.

The first two metrics quantify 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

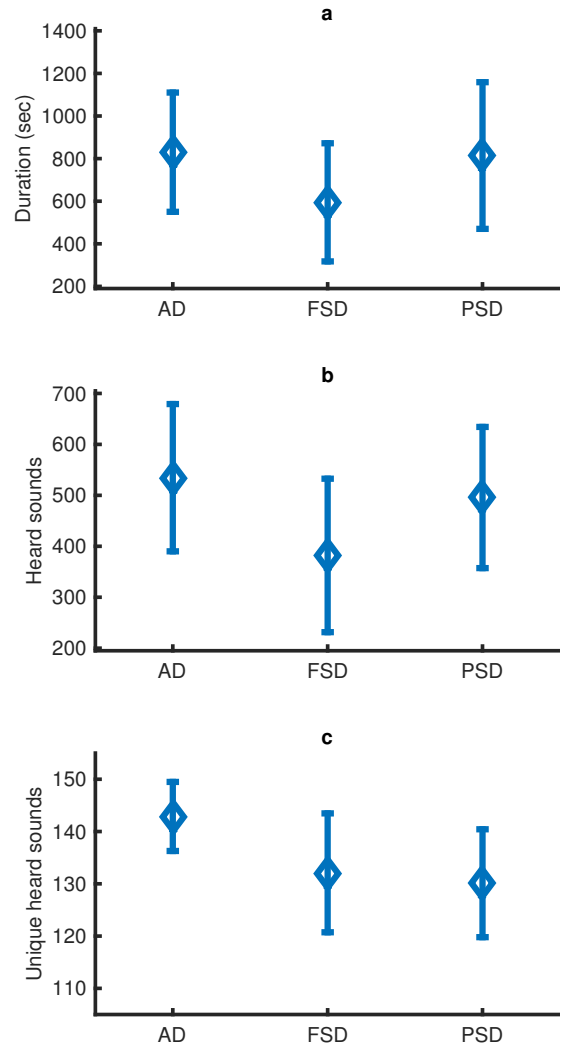

Fig. 6. Average and standard deviation for (a) the total search durations, (b) the number of heard sounds and (c) the number of unique heard sounds

to measure the selectivity of the interfaces. A low number of unique heard sounds 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 sounds are organized in space, and tends to play all the sounds at each search. The maximum number of unique heard sounds is the corpus size: 149 sounds.

The averages and standard deviations for the three metrics are presented on Figure 6. The type of display have a significant effect on the the *total search duration* ($F[2,52] = 3.63; p = 0.03$), the number of heard sounds ($F[2,52] = 5.65; p < 0.01$) and the number of unique heard sounds ($F[2,52] = 8.65; p < 0.01$). Post hoc analysis on the total search time (Figure 6 (a)) shows that FSD perform better than the other interfaces (FSD-PSD: $p = 0.04$; FSD-AD: $p = 0.02$), whereas PSD and AD seem to have similar results (PSD-AD: $p = 0.88$).

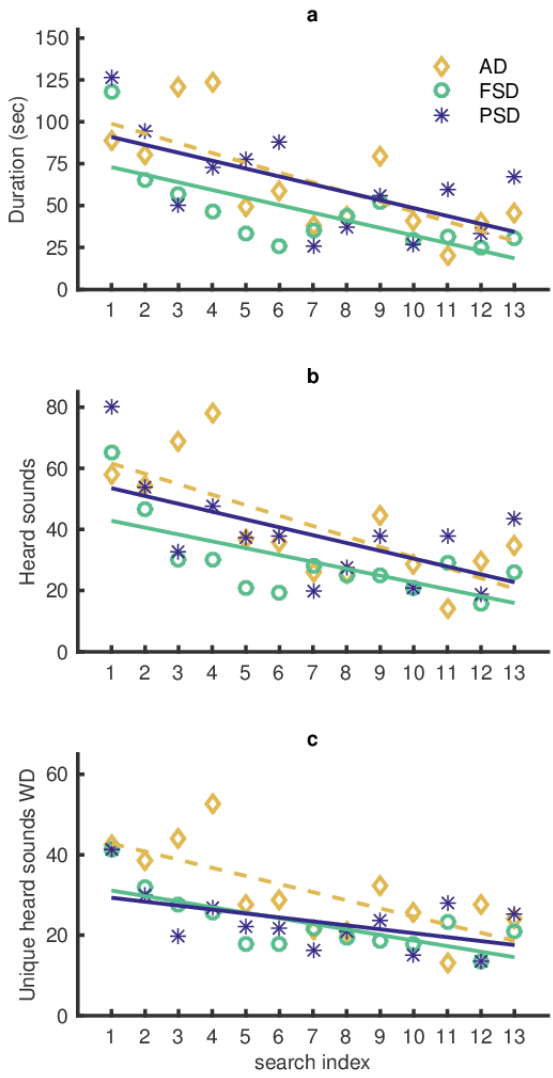

Fig. 7. evolution of (a) the average search durations, (b) the average numbers of heard sounds, and (c) average the numbers of unique heard sounds at each target sound search. Lines are linear regression.

Similar results are found for the numbers of heard sounds (Figure 6 (b)). FSD significantly outperforms the other interfaces (FSD-PSD: $p = 0.02$; FSD-ASD: $p < 0.01$), whereas PSD and AD show similar outcomes (PSD-AD: $p = 0.42$).

Figure 6 (c) shows the results for the number of unique heard sounds. 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 140 sounds, and 25% heard at least 148 sounds, that is almost the entire database. Considering PSD, 75% of subjects heard less than 134 sounds, against 143 for FSD. This time PSD and FSD perform equivalently (PD-FD: $p = 0.58$).

According to those results, a hierarchical organization of the dataset based on semantic values (PSD and FSD) allows the users to retrieve the 13 target sounds 1) more efficiently, and 2) by listening to a smaller amount of sounds than an organization based on acoustic descriptors (AD). However, those two effects are significantly compromised when users have to parse the entire hierarchy to reach the first target sound, as in the PSD display. It seems that enforcing an explicit exploration of the hierarchy disturbs or confuses the user instead of allowing him to learn the semantically motivated 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, the evolution of the above described metrics with respect to the search indexes is considered. 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 unique heard sounds for each target sound search.

Figure 7 (a) shows the evolution of the average durations of each target sound search observed for AD, FSD and PSD. As shown by a linear regression of the data, there is an overall increase of efficiency with respect to the index of the search. For the duration and the number of heard sounds, the AD and PSD perform equivalently, whereas the FSD exhibit better performance. In the case of the number of unique heard sounds, the two semantic displays are equivalent. This latter result suggest that for both semantic displays, users have parsed the same range of the dataset, although users of FSD have done so by listening to fewer sounds and more quickly.

6 Conclusion

In this paper, two displays allowing users to explore a semantically organized 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. Those two semantic displays are assessed and compared to a third display 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.

In the dataset considered in this study, there are as many leaf classes as sounds in the dataset, which would be unrealistic for large datasets. Though the organization can easily be adapted 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.

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] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

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

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

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

[17] 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