A novel interface for audio based sound data mining find a better title that put the focus on semantic versus acoustic organizations

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* [2]

n this paper, the design of a web interface for audio-based sound data mining is studied. The interface allows the user to explore a sound dataset without any written textual hint. Dataset sounds are grouped into semantic classes which are themselves clustered to build a semantic hierarchical structure. Each class is represented by a circle distributed on a two dimensional space according to its semantic level. Several means of displaying sounds following this template are presented and evaluated with a crowdsourcing experiment.

audio-based sound data mining, listening oriented user interface, crowdsourcing experiment

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 [**?**]. Most of that effort is dedicated to the automatic inference of indexing metadata from the actual audio recording [**?**, **?**]; in contrast, the ability to browse such databases in an effective manner has been less considered. The temporal aspect of sounds has been studied in [**?**] and the use of multidimensional projection of audio features in [**?**].

Typically, a sound data mining interface is based on keyword-driven queries. The user enters a word which characterizes the desired sound, and the interface presents him with sounds related to this word. The effectiveness of this principle is primarily based on the typological structure and nomenclature of the database. However, some issues arises from this paradigm:

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) [**?**, **?**, **?**]. 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 selections.
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 ambience 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 avoid those issues, we propose in this paper several means of displaying sounds without relying on any textual representation. Their effectiveness is studied with a search-based task whose aim is to listen to a target sound, and browse the database using the evaluated display to find this target sound as fast as possible. The proposed displays are first described along with the dataset used for evaluation. We then explain the chosen validation protocol, before finally presenting and discussing performance results.

1 Interface framework

1.1 Dataset structure

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

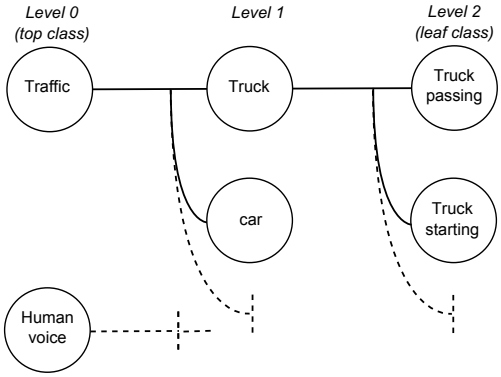

Fig. 1. Semantic hierarchical structure of the dataset of urban environmental sounds

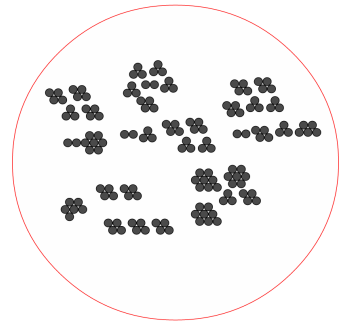

Fig. 2. Full Display (FD) with a non visible hierarchical organization of semantic classes

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 **??**. 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.

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 [**?**, **?**, **?**]. In cognitive psychology, sound categories may be regarded as intermediaries between collective sound representations and individual sensory experiences [**?**]. 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.

1.2 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.

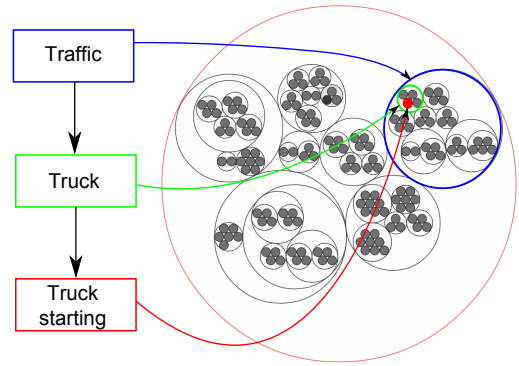

Fig. 4. Spatial configuration of the Progressive Display (PD) based on the semantic hierarchical structure of the dataset

As shown on Figures **??** and **??**, both interfaces are based on the same principle of distributing in a 2D space the 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 **??**. Thus, subclasses belonging to the same class are close to each others. Circle packing functions of the D3.js (Data-Driven Documents) javascript library [**?**] 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

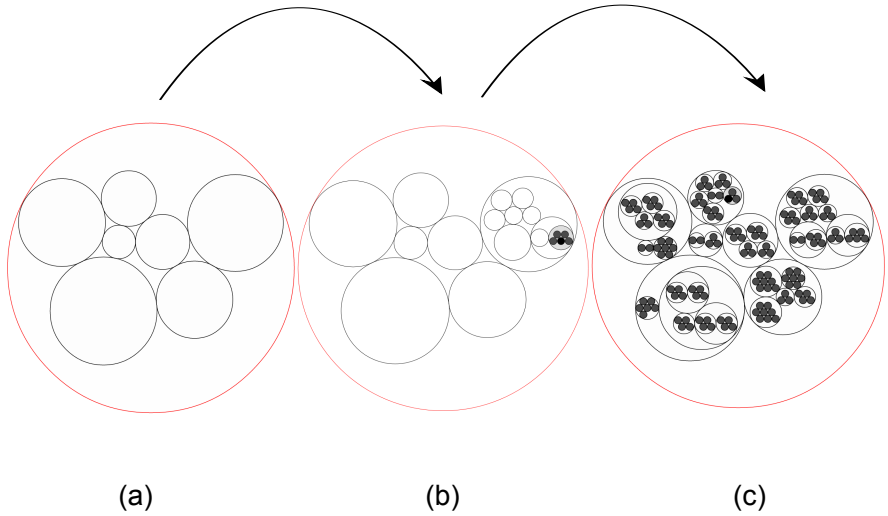

Fig. 3. Progressive Display (PD) with a visible hierarchical organization of semantic classes: (a) initial folded version; (b) partly folded version; (c) unfolded version

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 **??**, **??** and **??**.

2 Validation test

2.1 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 in a 2D space. An example of AD can be seen on Figure **??**.

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.

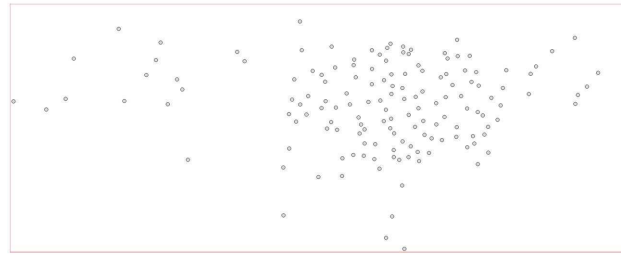

Fig. 5. The Acoustical Display (AD) computed using a non metric multidimensional scaling on MFCCs based acoustic descriptors

2.2 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.

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

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.

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.

2.3 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 duration searches, which is the *absolute duration* minus the break times between two target sound searches, is called *the relative duration*.
- the name of each sound which has been heard.
- the time at which each sound has been heard.

2.4 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.

2.5 Participants

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

Table 1. Standard deviations of the number of heard sounds per subject, with and without outliers.

Interface	PD	FD	AD
with outliers	141	155	315
without outliers	139	140	146

Table 2. Standard deviations of the *relative duration* per subject, with and without outliers.

Interface	PD	FD	AD
with outliers	353	277	520
without outliers	363	249	273

Table 3. Standard deviations of the *absolute duration* per subject, with and without outliers.

Interface	PD	FD	AD
with outliers	1401	339	4034
without outliers	408	327	273

3 Results

3.1 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 [**?**][**?**]. 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 [**?**]. As this method is not robust to the presence of isolated extreme observations (as it is often the case for crowdsourcing experiment), a method using the inter-quartile range (IQR) proposed by [**?**] is used in this paper. With this approach, an observation is considered to be an outlier if it is more than $3 * IQR$ higher than the third quartile or more than $3 * IQR$ lesser than the first quartile. For normalized distribute data, the IQR method remove less than 0.00023% of the data whereas the STD method remove 4.6% of the data [**?**]. This methods is applied to the following list of parameters:

- durations of each target sound search
- average duration of target sound searches
- maximum duration of target sound searches
- *relative duration*
- *absolute duration*
- number of heard sounds for each target sound search
- average number of heard sounds
- maximum number of heard sounds
- total number of heard sounds

Using the IQR method, 4 subjects are detected as outliers and removed from the analysis. 1 subject used PD, 1 subject FD, and 2 subjects AD. 2 subjects are detected by observing the *absolute duration* (4 and 12 hours), 1 subject by observing the total number of heard sounds (1800 heard sounds, roughly 12 times the total size

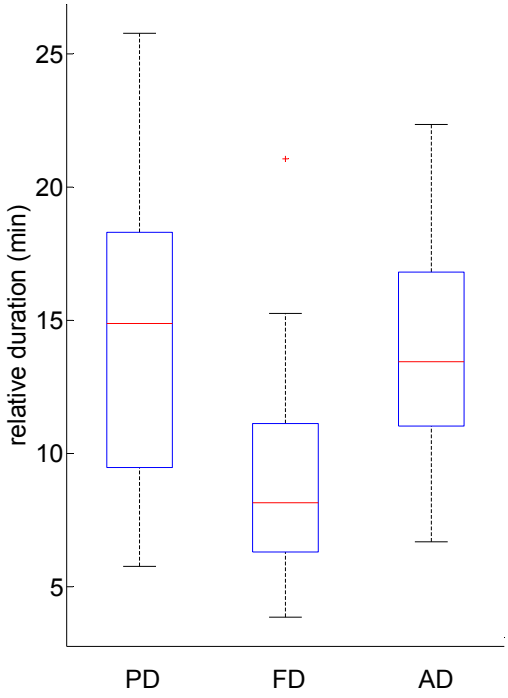

Fig. 6. Boxplot representing the distributions of the relative durations for the PD, FD and AD

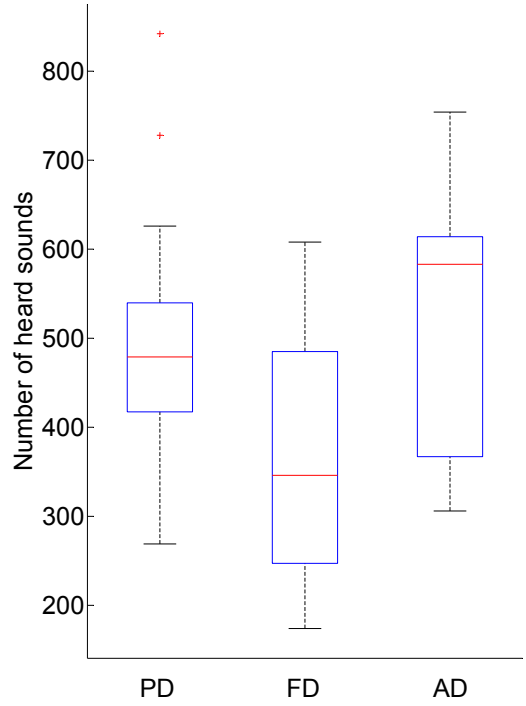

Fig. 7. Boxplot representing the distributions of the numbers of heard sounds for the PD, FD and AD

of the corpus) and 1 subject by observing the number of heard sound for the first target sound search (321 heard sounds, including 21 times the target sound).

The tables **??**, **??** and **??** measure the effect of the presence of the outliers on the standard deviations of three observed data being the total number of heard sounds, the *relative duration* and the *absolute duration*. The removal of the outliers have important effects on the data distributions, specially for AD.

3.2 Interface efficiencies

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

- the *relative 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.

Concerning the *relative durations*, distributions of the data are displayed on Figure **??** for the three interfaces. FD seems to perform better than the other interfaces, whereas PD and AD seem to have similar results. To refine the analysis, a two sided Wilcoxon rank sum test is considered. It is a non parametric statistical test which tests the null hypothesis that two set of observed data originate from distributions having equal median [**?**]. As expected, FD is significantly better than the other interfaces (FD-PD: $p = 0.0142$; FD-AD: $p = 0.028$) and there is no statistical differences between PD and AD (PD-AD: $p = 1$).

Distributions of the numbers of heard sounds are displayed on Figure **??** for the three interfaces. Results are similar of those observed for the *relative durations*. FD significantly outperforms the other interfaces (FD-PD: $p = 0.0115$; FD-AD: $p = 0.018$), whereas PD and AD show similar outcomes (PD-AD: $p = 0.3699$).

Lastly, Figure **??** displays the distributions of the number of heard sounds without duplication. This time the results of AD are significantly lower than those of both PD

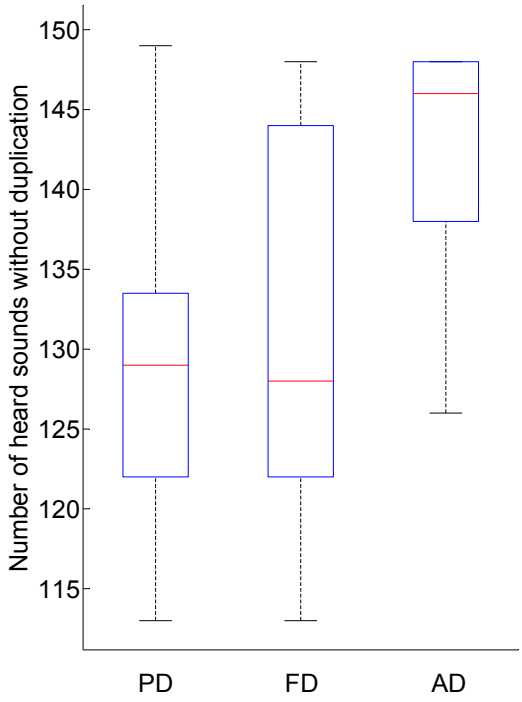

Fig. 8. Boxplot representing the distributions of the numbers of heard sounds without duplication for the PD, FD and AD

and FD (AD-PD: $p = 8.4910.10^{-4}$; AD-FD: $p = 0.027$). 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. There is no statistical differences between the PD and FD (PD-FD: $p = 0.8607$), indicating that those two interfaces perform equally.

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.

3.3 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 **??** (bottom) displays the evolution of the medians of the durations of each target sound search observed over the subjects 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. If we compare PD and FD, we note that the durations are systematically shorter for FD, except for the search 12. Furthermore, for FD, a threshold of 25 seconds is reached from the search four, whereas it is of 50 seconds for PD.

Figure **??** (top) displays the evolution of the medians of the numbers of heard sounds for each target sound search, observed over the subjects for the three interfaces. If the curve profiles of PD and AD seem to be similar to those respectively observed for the durations, here the maximum value for FD is reached for the third search. Again, values of FD are mostly below those of PD, except for the search index 3,10 and 12. For both PD and FD the curves oscillate from the search four, but those oscillations occur in a range of $9-17$ for FD and $9-30$ for PD. Similar results are found for the evolution of the medians of the numbers of heard sounds without duplication, shown on Figure **??** (middle).

Those 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.

4 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.

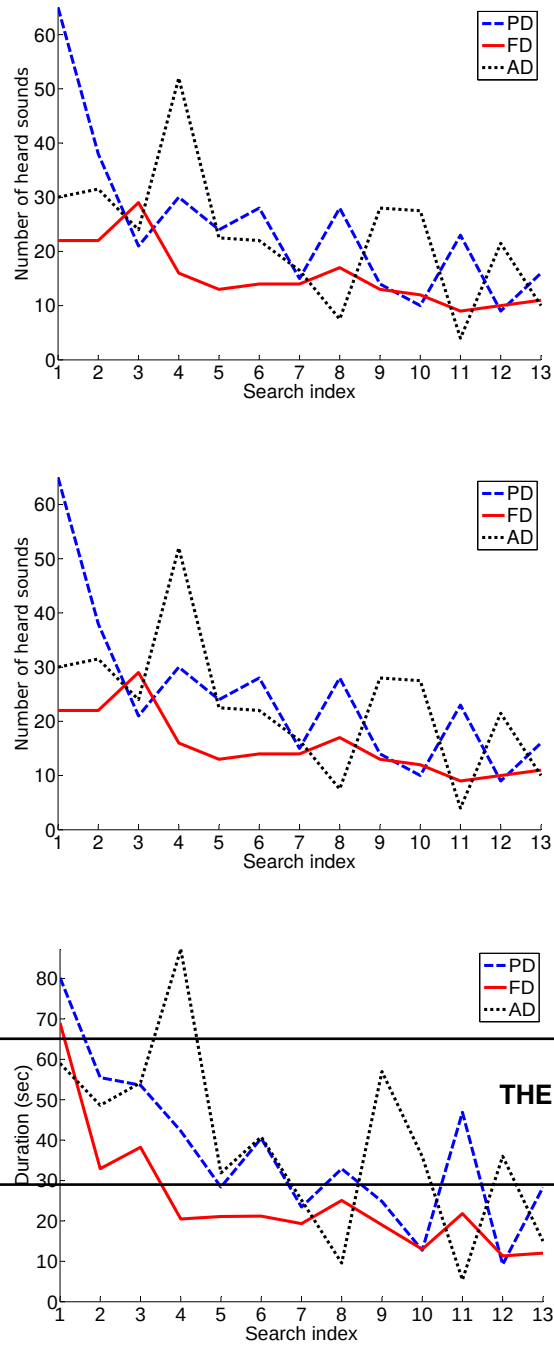

Fig. 9. Medians of (top) the numbers of heard sounds at each target sound search, (middle) the numbers of heard sounds without duplication (WD) at each target sound search, (bottom) the relative durations at each target sound search Gregoire : peux-tu ajouter WD.pdf ? la deuxieme figure

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.

5 Acknowledgements

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

THE AUTHORS