Lexical and Perceptual Grounding of a Sound Ontology

Anna Lobanova, Jennifer Spenader, and Bea Valkenier

Artificial Intelligence Department, University of Groningen, the Netherlands
<{a.lobanova|j.spenader|b.valkenier}@ai.rug.nl>

Abstract. Sound ontologies need to incorporate source unidentifiable sounds in an adequate and consistent manner. Computational lexical resources like WordNet have either inserted these descriptions into conceptual categories, or make no attempt to organize the terms for these sounds. This work attempts to add structure to linguistic terms for source unidentifiable sounds. Through an analysis of WordNet and a psycho-acoustic experiment we make some preliminary proposal about which features are highly salient for sound classification. This work is essential for interfacing between source unidentifiable sounds and linguistic descriptions of those sounds in computational applications, such as the Semantic Web and robotics.

Key words: Ontology, source unidentifiable sounds, sound features, WordNet, psycho-acoustic experiment

1 Sounds without Identifiable Sources

Bumps, rattles and rumbles: languages are filled with expressions to name sounds that we cannot identify according to their origin, the most common way to describe a sound. The ability to describe and distinguish between sounds is an essential cognitive skill, as sounds are one of our major sources of information about our environment.

In computational applications, ontologies are valuable resources for relating different concepts, often for the purpose of inference. For example, it is important to know that a *melody* is a part of a *song* which in turn is a kind of musical piece. In particular with the Semantic Web, having cognitively grounded ontologies available to serve as the backbone of search engines is more necessary than ever before. For source identifiable sounds, existing semantic ontologies are often already sufficient, e.g. for describing something as *the sound of a car engine* or *the sound of running water* the hierarchies for the source concepts *car engine* and *running water* are already present. Additionally, for source identifiable sound names like *a scream*, *a bark* or *a whinny*, the sounds can be integrated into classifications already present, e.g. dogs or horses.

The challenge is dealing with source unidentifiable sounds such as *click, clink, plop, thud, screech* and *rattle*. These sounds cannot be categorized by linking them to a source concept.

Our aim is to discover what features of source unidentifiable sounds are perceived as relevant to their classification, and to use them to develop an ontological structure for source unidentifiable sounds that captures the way in which listeners perceive them. Further, we are interested in if and how linguistic patterns might support an ontological structure. The lexical means available to describe sounds may offer clues to the features most salient to their classification.

In section 2 we discuss a number of examples of source unidentifiable sounds that seem to fall into different feature groups, in section 3 we discuss previous ontological attempts, first looking at some work on source identifiable sounds, and finally focusing on WordNet [1]. We show that WordNet does not organize source unidentifiable sounds in a consistent and sufficient way for computational applications. In section 4 we present a psycho-acoustic experiment we conducted in order to examine which features humans use when classifying sounds and whether some features are more salient than others. Based on these results in section 5 we propose that some features are more salient than others and when identified correctly they can be used to structure sounds in an ontology.

2 Features of Source Unidentifiable Sounds

There are sounds with a clear source, and sounds where the source is not clear or not known at all. The source typically functions as its description. For example, bells toll, horns toot and knocking can be an effect of fingers touching the surface of a door. While sound source identification in the examples above is relatively easy, the sound of *swish* is not, since it can be produced by fallen leaves and the wind (nature), curtains (material), or by a gramophone record (a plastic object). Further, the sound of *whack* can be a result of almost anything from someone's hand (body part) to wings of birds (animal part). And what about sounds like *a thunk*, *a whiz* or *a throb*?

At least three different perspectives can describe sounds. For source unidentifiable sounds, source based descriptions are obviously not possible. Sounds can also be described according to their acoustic properties. This has the advantage of being entirely objective, but has the disadvantage of potentially being completely incompatible with the way in which humans perceive sounds.[1] Since our aim is to make a classification that will allow humans to categorize and relate sounds they perceive to other sounds, the third perspective, descriptions based on perceptually relevant aspects of sounds, seems most promising.

But what are the perceptually relevant aspects of source unidentifiable sounds? We began by listening to a large number of source unidentifiable sounds and identifying salient features that seemed to help characterize the sounds, finally identifying five features:

1. **Repetitiveness**: A drum is repetitive, a sigh is not.

[1] This is then analogous to the correct biological classification of a tomato as a berry, while most people would consider it to be a vegetable, and expect to find it among the vegetables in the grocery store.

2. **Continuousness**: If the sound is interrupted by silences it is not continuous. Drumming is (−)continuous while a sigh is (+)continuous. Repetitive sounds can be continuous, e.g. the toll of a bell, or not continuous, e.g. tapping.
3. **Duration**: If the sound is produced by an ongoing process it exhibits a durational aspect. A sigh, for example, is produced by the ongoing flow of air, while a click is not.
4. **Harmonicity**: Has to do with how pleasant a sound is. The toll of a bell is much more harmonic than a sigh or the beat of a drum.
5. **Pitch**: Has to do with whether or not a sound can show pitch variation. For example, a sigh or swish doesn't, but a screech or ring does.

We consider these features to be highly salient but it is not clear which features are most relevant for classifying linguistic terms for source unidentifiable sounds. This can be analogous to an initial classification of animals where it would be determined that whether or not an animal could fly is a less salient feature than whether or not they give birth to leave offspring, since the latter distinguishes mammals from birds, while the former only distinguishes birds like penguins from e.g. robins. In order to reliably determine which features are more salient than others we will need to do some psycho-acoustic experiments. But first let's see what classification attempts have already been made.

3 Proposed Ontologies for Sounds

3.1 Previous Work on Sound Ontologies

Most of the work on sound classification is done in the area of sound and speech recognition. There is no consistency as to which criteria should be used when distinguishing different sounds. [4], for example, divided sounds into 3 classes: speech, music and sound texture. [4] does not give a clear definition of sound textures but some examples include the sounds of a copy machine, fish tank babbling, waterfall, applause, and so on, in other words, source identifiable sounds that are not music or speech. In a psycho-acoustic experiment [4] asked the participants to cluster sound textures in order to find out which features people find salient. Participants differed radically in their classifications. One possible explanation is that some participants used the *source* of the sound as the main feature, while others used such perceptual features as *periodicity* and *smoothness* leading to different classifications.

[3] proposed to make a sound ontology where sounds were grouped according to their acoustic features into such sound classes as music and speech. All individual sounds in these groups could be listed together with their attributes (that is acoustic features) like *frequency*, *timber* or *rhythm*, and connected with each other by the ontological relationships *part-of* and *isa*. The main aim of such an ontology was to provide enough information about the features and to enable sound segregation from an input sound mixture. Because of its specific purpose, the attributes, or features listed in the ontology are acoustic in

nature (e.g. AM/FM modulation, power spectrum, formant) and no lexical information is given. Similarly, [5] looked at 13 acoustic features in their real-time computer models in order to see whether these features can help to distinguish between music and speech sounds. They report that the best model used only 3 out of 13 features (namely, 4 Hz Modulation Energy, Var Spectral Flux and Pulse Metric). These results suggest that some features are more salient than others. The formal properties of these salient features might have some overlap with the basic perceptual features we used in our psycho-acoustic experiment. For example, modulation energy is related to the *loudness* and *repetitiveness*, spectral "Flux" might have to do with the *continuity* and pulse metric might overlap with our feature *repetitiveness.* However, it is possible that humans use very different features, and it is not clear how well these features carry over to source unidentifiable sounds.

3.2 WordNet's Sound Classification

The WordNet ontology includes source identifiable as well as source unidentifiable sounds. Unlike traditional dictionaries where all words and their meanings are enumerated in the alphabetical order, WordNet was originally based on psycholinguistic principles trying to capture the way words and meanings are represented in humans. We concentrate on WordNet because it is currently the most widely used lexical resource in computational linguistics. All words are organized in the so-called synsets, or sets of synonymous words, hierarchically organized via such semantic relations as hyponymy and hypernymy. Each lexical entry has a definition and often an example of use. WordNet contains four categories: nouns, verbs, adjectives and adverbs but our main focus is on the nouns describing sounds. Nouns belong to one of nine hierarchies, each associated with a top level concept called a unique beginner. Since the same lexical string can have more than one meaning and belong to different synsets, it can occur in several different hierarchies.

Sounds are organized in the WordNet according to their senses and not their features. The string *sound* has 8 senses but only 6 are relevant: - sound$_1$: the particular auditory effect produced by a given cause; - sound$_2$: auditory sensation: the subjective sensation of hearing something; - sound$_3$: mechanical vibrations transmitted by an elastic medium; - sound$_4$: the sudden occurrence of an audible event; - sound$_5$: the audible part of a transmitted signal; - sound$_6$: (phonetics) an individual sound unit of speech without concern as to whether or not it is a phoneme of some language;

Sound$_1$, sound$_5$ and sounds$_6$ occur in the hierarchy with the unique beginner *Abstraction*, sound$_2$ is in the hierarchy *Psychological Feature*, sound$_3$ is in the hierarchy *Phenomenon* and sound$_4$ is in the hierarchy *Event.* It is important to point out that WordNet does not distinguish between source identifiable and unidentifiable sounds per se because its main point is to represent sounds as to the main concepts they imply, for example, whether it is an instance of such basic cognitive process as sensation (such as *music*) or whether it is an occurrence of an audible event (such as *drumbeat*). This kind of approach seems to be insufficient because of the inconsistencies it causes.

For example, one of the problems with sound representation in WordNet is that when two sounds are on the same low level of a hierarchy, the difference between them can only be elicited from their definitions. Consider sounds *throbbing* and *knocking*, two terminal sister leaves of the *Event* hierarchy with sound_4. According to the WordNet, *throbbing* is a sound "with a strong rhythmic beat", and *knocking* is a sound of "knocking as on a door or in an engine or bearing". Only these definitions provide information about the quality of the sound *throbbing* (namely, that this sound is strong and rhythmic), and about the source of the sound *knocking* (namely, a door or an engine).

However, in some cases the way of distinction between sister terms is not possible since some of them (especially sounds that belong to the same synset) share the same definition. For example, both *click* and *clink* belong to the same synset, hence, share the same hierarchy and base type. They are also described by the same definition of "a short light metallic sound" and no example of use is given. Likewise, a "plop-and-a-plunk" problem is rather an evident example of the inconsistent representation of sounds at the lower levels of sound hierarchies. Sounds *plop* and *plunk* occur in the same *Event* hierarchy, however, while *plunk* is a direct hyponym of sound_4, *plop* is linked to sound_4 indirectly via *noise*. There is no clear criterion to consider *plop* (defined as "the noise of a rounded object dropping into liquid without splash") as a hyponym of *noise* (defined as "sound of any kind (especially unintelligible or dissonant sound)") while *plunk* (defined as "a hollow twanging sound") as its sister.

In summary, although the idea of organizing sounds as to the possible mental representation of sounds is very appealing, the current state of affairs in WordNet proves to be inconsistent and insufficient.

One of the plausible ways to proceed is to look at the definitions of sounds more closely. As has been mentioned above, some definitions provide enough information for distinguishing one sound from another. Namely, *throbbing* is rhythmic, *clicks* and *clinks* are short and light and *plunk* is hollow and twanging. These descriptive words seem to be very good indicators of how people perceive and describe sounds (as in "I heard a short click" because they represent the perceptual features of sounds. These descriptive words are what we use as the basis for our psycho-acoustic experiment we present in the next section.

4 Experiment

Are the features identified by the experimenters in section 2 perceived by listeners as relevant to classifying sounds? Participants were presented with three sounds and asked to choose the sound that differs most from the other two.

4.1 Method and Materials

We used 26 questions consisting of three sounds each as stimuli. All of the sounds are real life sounds taken from the Auvidis sound library [2]. We were careful to choose sounds for which the source was difficult, or not possible to determine. All files were cropped to a uniform duration of 80 milliseconds. After initial selection, we decided which features characterize each sound sample. The stimuli can be subdivided into three types. Type Simple (16/26) questions consisted of two sounds similar on all features and one sound that differed on one or more of these features from the other two. In the example below sound S2 differs on feature F2:

S1: F1(+)F2(+)... S2: F1(+)F2(-)... S3: F1(+)F2(+)...

This type of questions will be used to test whether participants actually perceive the features. If this is the case, the sound that differs on one feature will be determined as the least similar. In questions from Type Complex (7/26) a set of three sounds consists of two pairs, where one sound belongs to both pairs. Within each pair the sounds share features. Since this set-up creates a conflict of several features, participant's choice will show which feature is more dominant. For example,

S1: F1(+)F2(-)... S2: F1(+)F2(+)... S3:F1(-)F2(+)...

The sounds S1 and S2 share feature F1 and the sounds S2 and S3 share feature F2. This type of question will be used to test whether one feature is more salient than another. If, for example, feature F2 is more salient than feature F1, sound S4 will be experienced as being more different, because it does not share this feature. Finally Type Control consisted of three questions (3/26) where two out of three sounds were exactly the same. These control questions tested whether participants were paying attention. Additionally five times during the experiment participants were asked to explain their choice in comments.

The experiment was done online and results were stored in a database. 31 adult native Dutch speakers took part in the experiment. Two of the participants were excluded for reporting hearing problems and one participant was excluded for giving a wrong answer on one of the control questions.

4.2 Results and Discussion

Questions of type Simple were meant to test whether people were sensitive to the features identified by the experimenters. From the 16 type Simple questions ten questions were answered as expected, the stimulus that differed on the feature dimension identified by the experimenters was chosen significantly more often than chance (χ^2, p-value 0.001). However, in four cases participants consistently chose a stimulus different from the stimulus predicted by the experimenters (χ^2, p-value 0.001). The participants agreed on which sound was different but this was not what the experimenters predicted from the features identified. Since all five features identified were presented in the type Simple stimuli answered as predicted as well as in the type Simple stimuli not answered as predicted the results are difficult to interpret. What is, however, striking is the high degree of agreement among participants as to which sound was different (cf. [4]'s results), suggesting that if the correct salient features could be identified, what sound participants will judge as different should be predictable.

Among questions of type Complex, five out of seven answers were significantly different from a uniform distribution. However, these results do not indicate that one feature was consistently considered to be more dominant than the other feature, so not much can be concluded about which features might be more dominant than the others. In four cases participants chose the sound exhibiting the most features (χ^2, p-value 0.001). These sounds may be considered more complex, and *sound complexity* might also be a salient feature.

Participants' explanations about their choices were not always easy to interpret. For example, participants did not report that a sound was chosen because it exhibited "a different tone color". Instead they reported that the sound they chose was "more sharp" or the sound was "more dull". All these descriptions were interpreted and labeled. In most of the cases the reports were consistent with the chosen sound. For example, the participants who chose the third sound for a given question gave another description than the participants that chose the first sound. The features that were determined by the experimenters were all mentioned at least once. Furthermore participants referred to *tone color*, *changing through time* and *on-/offset* characteristics.

There are three possible explanations why we didn't obtain the clear results we had hoped for. First, it could be that many of the stimuli were too complex, making it hard to compare. We removed a number of questions because we thought they were too simple but that may have been a mistake. Second, it could be that additional features play key roles. Features mentioned by the participants might be a good starting point to look for other salient characteristics in future work. Third, it's possible that the features are hierarchically ordered but in such a way that some of our stimuli sets made it difficult to compare, or led to comparisons in a way we did not expect.

But because people were quite consistent in their evaluation of sounds, and because this to a certain degree was similar to our expectations we are quite optimistic that further experiments with more stimuli will help us determine the actual hierarchical characteristics of the features.

5 Conclusions

How can we use our observations about the shortcomings of sound classification in WordNet and the experimental results to propose a classification for source unidentifiable sounds?

If we examine the features we have studied again, *pitch*, *duration*, *harmonicity*, *continuity* and *repetitiveness*, what characteristics do these have compared to features we chose not to focus on? One important characteristic is that the values of these features are consistent across all tokens of a given type of sound type. Taking each sound term, such as *clink*, *rattle* or *plonk* as a type, the feature *absolute pitch*, which we did not choose to study, conspicuously does not have these characteristics. A *clink* could have a high pitch or a low pitch, and both would still be instances of *clinks*. The same goes for *rattle*: high-pitched rattle or a low-pitched rattle are both possible rattle tokens. Based on this, even though absolute pitch might be a salient feature for classification of some sounds in the experiment, its ability to vary among tokens of the same type make it an inappropriate choice for classification.

The feature *pitch* we used has to do with having or not having pitch, so e.g. *swish*, *thump* and *gurgle* are all examples that are $(-)$*pitch*, and *pitch* seems to remain consistent for all tokens of each of these types. The experimental results also suggest this is a salient feature, and could be used to split sound types into two sets.

Examining a number of sound types, long *durations* seem to be consistent across e.g. *rattle* or *hum* tokens, while short *durations* are also characteristics of *clinks* and *plonks.* Further *repetitiveness* and *continuous* seem to be associated with *a long duration.* Thus it seems that these might be lower branches. Both these features also seem to be consistent among tokens of the same type.

But which feature would make a better initial split: *pitch* or *duration*? Unfortunately, the results from the experiment were not clear enough to allow us to make this decision, and more tests are needed.

The function of the feature *harmonicity* is also not clear. It might be a feature allowing us to split the set of sounds with short duration into $+$*harmonic* (e.g. *pling* or *booing*) from those that seem to be $(-)$*harmonic* (e.g. *click* or *plonk*), but it might be necessary to do a classification experiment to see if subjects would agree with this division.

As for the other features pointed out by experiment participants, such as e.g. *tone color*, we will have to do more research. Our results certainly suggest that the choice of salient features in a sound ontology has to be empirically grounded. It seems possible to make principled decisions to structure source unidentifiable sounds in an ontology, a result that should be useful for many applications, both those under development such as for searching media on the Semantic Web, and applications in the future, e.g. robots that can describe sounds they heard as humans would.

References

1. Fellbaum, C. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, (1998).
2. Mercier, D. *Sound Library.* AUVIDIS, (1989).
3. Nakatani, T., Okuno H. G. *Sound Ontology for Computational Auditory Scene Analysis.* AAAI/IAAI, pp:1004-1010, (1998).
4. Saint-Arnaud, N. *Classification of Sound Textures.* M.S. Thesis in Media Arts and Sciences, Massachusetts institute of Technology, (1995).
5. Scheirer, E., Slaney, M. *Construction and Evaluation of a Robust Multifeature Speech/music Discriminator.* Proc. ICASSP-97, Munich, (1997).