Automatic classification of environmental sounds

Janto Skowronek, Monika Kappelmann, Martin McKinney

Philips Research Laboratories, Eindhoven, The Netherlands, Email: janto.skowronek@philips.com

Introduction

Automatic audio classification provides a core technology for a wide variety of audio processing systems, such as speech recognition or music content management. Because environmental sounds (env-sounds) bear information that helps humans to identify and assess daily-life situations, automatic recognition of these sounds widens the field of possible applications. In multimedia processing for instance, such classification schemes are used for the automatic detection of action scenes in movies. In the field of health-care systems, modern hearing aids can automatically adapt their parameters based on a classification of the current auditory scene.

Currently there are two main types of algorithms that deal with env-sounds: Either they differentiate between speech, music and background noise including env-sounds (e.g. in hearing aids [1]), but without identifying particular env-sounds. Or they try to detect a limited number of sound events that are typical for the considered application (e.g. movie scene detection [2]). In contrast to these approaches we were interested in a detailed classification of different env-sound classes. The research questions in this study were: (1) What categories of env-sounds can be defined? (2) What type of features can be used for classifying env-sounds?

Experimental Study

We performed some experiments in which we classified a database containing env-sounds. First we implemented a classification algorithm with the two standard stages: feature extraction and pattern recognition.

We computed three feature sets: *multimedia features*, *hearing aid features* and *ASA features*. *Multimedia features* [3] were 9 low-level features such as spectral centroid etc, used in many audio classification algorithms. *Hearing aid features* [1] were 10 features that are implemented in hearing aids, which have the already mentioned automatic parameter setting. These features embody amplitude statistics as well as amplitude and frequency fluctuations of the audio signal. *ASA features* were 7 features that have been developed in [1] and refer to the research field of auditory-scene-analysis (ASA). Bearing the processes in mind how a human analyzes an auditory scene, these features mainly represent the harmonicity of a signal as well as the distribution of onsets.

The pattern recognition stage consisted of a standard classifier using quadratic discriminant analysis. The audio database comprised 1200 env-sounds and was randomly split into 90% training and 10% test material.

Feature set	Classification performance
Multimedia	50 %
Hearing aid	38 %
ASA	29 %

Table 1: Summary of classification results.

In order to train and test the classifier, we had to assign the sounds to reasonable classes. For that purpose two subjects listened to all sounds and put similar sounds into the same class. Due to the high variability of env-sounds, we applied different criteria for the class labelling: perceived duration and temporal structure, pitch and spectral structure, meaning and typical surroundings of occurrence. Since we ended up with a high number of classes, which often contained only a few sounds, we further grouped the classes into five general categories:

1. Human / animal articulation: e.g. bird twitter, human laughing, crowd noise
2. Human / animal activity: e.g. human steps, working with non-electrical tools
3. Machine activity: e.g. automotives, electromechanical devices, construction noise
4. Indications of a changing environment: e.g. sounds caused by opening/closing things, explosions
5. Other / non-assigned environments: e.g. nature surroundings

Table 1 shows the percentage of correct classification achieved with the three feature sets. Independently of the feature set used, performance is quite poor, though the features have been proven successful in other classification experiments, in which the distinction of noise (including env-sounds) from other classes (speech, music) was tested [1, 3]. And as expected we found a high variability of sounds within the single classes.

Both facts led us to the conclusion that the performance was limited by the class definitions rather than by the features used.

Theoretical Study

In order to get more insight into the problem of defining better env-sound classes, we analyzed how we had defined the classes in the experiments and what might have been the causes that prevented good classification.

When developing a classification algorithm, the class definitions are usually determined by the particular application. But in our experiments we had no constraints for

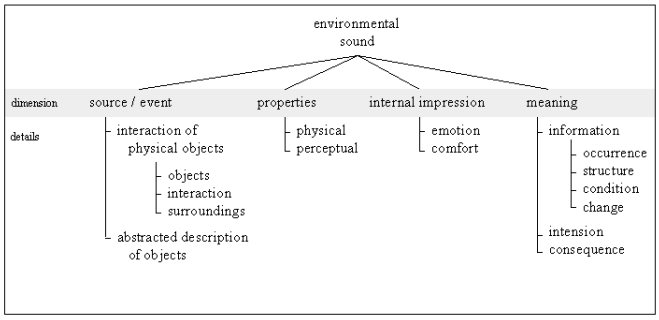

Figure 1: Taxonomy for describing environmental sounds.

that, since we had no particular application in mind. For such an *open* situation we concluded that class definitions should fulfil two requirements: (1) The classes should refer to the same *type of properties* and (2) the assignment of sounds to those classes should be non-ambiguous. As an example, a classifier with the two classes "trucks passing by" and "annoying sounds" violates both requirements: the first one because one class refers to physical objects while the other refers to emotional attributes; the second one because sounds of by-passing trucks are usually also annoying.

The first requirement indicates that there exist a number of dimensions for class definitions. Note that we do not mean the number of the chosen classes but the different methods of how classes can be defined. Speech signals for example can be classified by language or phoneme or speaker; music signals can be assigned to classes of genres or instruments or artists etc. Bearing our labelling procedure in mind, one can argue that we mixed such dimensions when using the different criteria: signal properties, meaning and physical surroundings.

That means the intuitive mix of the chosen criteria might have violated the above mentioned requirements such that no good classification was in principal possible.

From that analysis we concluded that a systematic overview of such dimensions may help in finding better class definitions. Based on own experience (o.e.) as well as literature knowledge [4, 5, 6], we propose a taxonomy for describing env-sounds (Figure 1). It is intended to serve as a guideline for defining good env-sound classes by using only one of the four identified dimensions:

1. *Source/event*
This dimension refers to the physical object/events that produce the sound. We know that humans tend to describe auditory scenes in terms of sound sources [4]. Actually not single objects but the interaction between at least two objects cause env-sounds [5]. However people tend to describe such interactions with one single object or event (o.e.). For example they would say "a rushing creek" instead of "bursting air bubbles in water, when flowing over little stones in a creek".

2. *Signal properties*
This dimension refers to the properties of the sound signal itself. Class definitions based on *physical* signal properties lead to trivial classification tasks, since then the features have to be simply these properties. But in contrast, classes referring to the *perceived* signal properties can be quite difficult to recognize. Class definitions using this dimension can be based either on perceived similarity (o.e.) or on formal subjective experiments [6] using Semantic Differentials (SD): loud-silent, high-dull etc.

3. *Internal impression*
This dimension refers to how humans feel, when they hear a particular sound. It is commonly known that humans can associate emotions with the sound (e.g. scary sounds) and that they can experience a degree of comfort (e.g. annoying sounds) when they listen to a sound. Again class definitions could be determined with SD's as was done in [6].

4. *Meaning*
We found three issues that the term *meaning* of env-sounds can refer to. (1) Env-sounds can bear some information about objects and events [5]: occurrence, structure, condition, changes. (2) There are sounds which are artificially produced with a certain intension (o.e.), e.g. alarm signals. (3) Humans can associate a possible consequence with the sound [(o.e.),4], e.g. in traffic: truck passing by ⇒ probably a dangerous situation.

Conclusions

This study dealt with a detailed classification of single env-sounds classes. We observed that the definition of categories for env-sounds is not trivial and should be done systematically. Therefore we propose a systematic overview about criteria (dimensions) for class definitions as a guideline for finding good env-sound classes. In future work we will evaluate the practical usefulness of this taxonomy by repeating our experiments with new env-sound categories based on the taxonomy.

References

[1] M.C. Buechler, *Algorithms for sound classification in hearing instruments*, PhD thesis, Swiss Federal Institute of Technology, Zuerich, 2002

[2] S. Moncrieff, C. Dorai, S. Venkatesh, *Detecting indexical signs in film audio for scene interpretation*, ICME, IEEE, Tokyo, 2001.

[3] M.F. McKinney, D.J. Breebaart, *Features for audio and music classification*, 4th ISMIR, Baltimore, 2003

[4] V.T.K. Peltonen, A.J. Eronen, M.P. Parviainen, A.P. Klapuri, *Recognition of everyday auditory scenes: Potentials, latencies and cues*, 110th AES, 2001

[5] M. Rauterberg, M. Motivalli, A. Darvishi, H. Schauer, *Automatic sound generation for spherical objects hitting straight beams based on physical models*, ED-MEDIA 94, Assoc. for the Advancement of Computing in Education, USA, 1994

[6] H.U. Prante, *Modeling judgments of environmental sounds by means of artificial neural networks*, PhD thesis, Technical University Berlin, 2001