Using Morphological Description for Generic Sound Retrieval

Julien Ricard and Perfecto Herrera
Music Technology Group
Pompeu Fabra University
Barcelona, Spain
jricard, pherrera@iua.upf.es

Abstract

Systems for sound retrieval are usually "source-centred". This means that retrieval is based on using the proper keywords that define or specify a sound source. Although this type of description is of great interest, it is very difficult to implement it into realistic automatic labelling systems because of the necessity of dealing with thousands of categories, hence with thousands of different sound models. Moreover, digitally synthesised or transformed sounds, which are frequently used in most of the contemporary popular music, have no identifiable sources. We propose a description framework, based on Schaeffer's research on a generalised *solfeggio* which could be applied to any type of sounds. He defined some morphological description criteria, based on intrinsic perceptual qualities of sound, which doesn't refer to the cause or the meaning of a sound. We describe more specifically experiments on automatic extraction of morphological descriptors.

1 Introduction

When building a sound description system, the first issue to be addressed is defining the aspect of the sound we want to describe, that is the information we want to provide about this sound. A sound can be interpreted in many different ways depending both on the application and on the sound itself. When analysing speech, for example, a system can aim at extracting some information about the speaker (causal description, i.e. description of the origin of the sound), such as whether he is a male or a female or his age, or at grasping the meaning of what is being said (semantic description). In the case of musical signals, semantic description is not well-defined and some specific features can be described (melody, rhythm…).

In sound retrieval systems, the aim of the description is to ease the search for a desired sound by providing a simple representation of it according to one or several description criteria. Typical systems offer textual search facility for sounds labelled according to their origin (e.g. "bird" or "creaking door"). More elaborate systems allow refining the search by adding information at other description levels through more textual labels (e.g. "violin + E# + vibrato" or "voice + sad") or visual representations; in other cases they propose search by similarity, comparing directly some features of a sound example to all sounds in the database and retrieving "the most similar", whatever it may mean.

From our point of view, some important issues need to be addressed when looking at current systems. First, no universal description scheme exists. Applying semantic description to musical sounds or classical musical description to speech wouldn't make much sense. Having a general description scheme would allow handling all types of sound in the same way. Moreover, source recognition systems still don't perform well enough and, in most cases and especially for non-musical sounds, labelling is done manually, which is very time-consuming. A general description scheme, based on a few perceptual properties common to all sounds, could ease the labelling or could even be used directly for the search. Another issue regarding the possibility to decompose sounds in several perceptual dimensions (e.g. roughness or noisiness) would be to retrieve sounds by specifying only one or a couple of them.

2 Schaeffer typo-morphology

In his *Traité des objets musicaux* (Treatise on musical objects) (synthesised and commented by Chion (1983)), Schaeffer (1966) proposes a generalization of what is usually heard as musical sounds (typically notes generated by traditional musical instruments) by considering all kind of *sound objects*, disregarding their origin (electronic sounds, noise, environmental sounds, loops…). After performing some listening experiments, he proposed a sound classification (*typology*), independent from the meaning or the source of the sounds, according to some intrinsic perceptual properties (*morphological* criteria) described below:

- *Mass*: related to the perception of the *pitchiness* of a sound, and then to its spectral distribution.
- *Harmonic timbre*: "the more or less diffuse halo associated to the mass and more generally what allows describing it" (Schaeffer, 1966, p. 516). We interpreted this definition as a finer characterisation of the spectral distribution, often described by analogy to vision: bright/dull, round/sharp…
- *Grain*: defined as the microstructure of sound matter,

such as the rubbing of a bow.

- *Dynamics*: energy temporal evolution.
- *Allure*: amplitude or frequency modulation.
- *Melodic profile*: variation of the pitch sensation.
- *Mass profile*: variation within the mass (e.g. pitched to complex).

3 Computational morphological description

There has been very little research on automatic morphological description. The main work done in that area is the ECRINS project on sound samples audio content description (Geslin, Mullon, and Jacob, 2002). A morphological description scheme is defined, in which sounds are automatically described according to the following descriptors: dynamical profile (amplitude evolution), melodic profile (pitch evolution), attack type, note pitch, spectral distribution, space (sound location and movement) and texture (vibrato, tremolo, and grain). The description can then be refined manually by the user.

The classification allowed by the automatic description is quite limited and it seems possible to complete it by further analysis and by adding new description criteria. In order to further investigate computational morphological description, we specifically considered three criteria: mass, mass profile, and dynamics.

The mass describes the pitchiness of a sound. It is estimated using pitch salience, as defined by Slaney (1994). This descriptor gives an estimation of the signal periodicity by comparing the amplitude of the largest peak to the zero-lag peak (power) in short-term autocorrelations.

The mass profile describes whether the sound mass varies or not. We used the mean and the variance of the smoothed pitch salience for classifying sounds into two mass profile classes: *varying* or *unvarying*.

The dynamics describes the type of amplitude envelope of the sound object. We defined four classes: *unvarying*, *varying-impulse*, *varying-iterative* (sound object with several transients) and *varying-other*. We used four descriptors, derived from the amplitude envelope and chosen intuitively according to specificity of each class:

- A 'balance coefficient', describing how much the centre of gravity of the amplitude envelope is off-centre.
- Size of the middle part of the envelope (between the attack and the release).
- Number of high amplitude derivatives: peaks in the middle part of the amplitude envelope (so that the first attack is not taken into account) above a threshold. Only iterative sounds are supposed to have such peaks.
- Mean values of high amplitude derivatives.

4 Evaluation and discussion

Evaluation of perceptual sound quality is a difficult task. There is no ground truth, such as in source recognition, and some given quality can be perceived differently according to the listener. However, we considered that morphological criteria could be regarded as rather listener-independent: the classes we defined for each criterion seem sufficiently distinguishable for a sound to be quite confidently classified in one rather than in another.

In order to evaluate our system, we built a small database (around 50 sounds for each class) for each morphological criterion. Sounds were manually labeled according to the classes defined. We performed no segmentation so that each file was considered as a sound object, whatever it contained (sequence of musical notes, street atmosphere, dog bark…). The tests were done independently for each morphological descriptor, on half the database using a decision tree trained on the other half. Each test showed around 80% correctly classified sounds, and the rules obtained were the expected ones (for example, sounds with negative balance coefficient, short middle part and no high derivatives were classified as impulse).

Further work includes refining current morphological descriptors and improving our classification models by building a much larger database. In order to get a more complete description of perceptual sound qualities, new descriptors, inspired by remaining Schaeffer's morphological criteria or by other research in sound perception (e.g. grain, modulation features…), should also be added. We are currently working on automatically describing the melodic profile and detecting more amplitude envelope classes.

5 Conclusion

We think that morphological description could be a good complement to traditional source-centred sound retrieval systems. It provides a universal description scheme according to meaningful perceptual sound qualities that could be queried directly or used in addition to other criteria, as a pre-filtering or refining stage for retrieval in large sound database. We showed that a few low-level descriptors allow automatically classifying sounds in a simplified 3-dimensional typo-morphology with good performance. The results obtained are encouraging, and many issues will be considered in order to improve the current system.

Acknowledgements

Part of this research has been supported by the European IST project CUIDADO.

References

Chion, M. (1983). *Guide des objets sonores*. Paris: INA-GRM/Buchet-Chastel.

Geslin, Y., Mullon, P., & Jacob, M. (2002). Ecrins : an audio-content description environment for sound samples. In Nordahl, M. (Ed.), *Proceedings of 2002 International Computer Music Association* (pp. 581–590). Göteborg, Sweden.

Schaeffer, P. (1966). *Traité des objets musicaux*. Paris: Seuil.

Slaney, M. (1994). *Auditory Toolbox: a Matlab toolbox for auditory modeling*. Apple Computer technical report #45, 1994.