A hierarchical approach to content-based classification of environmental sounds using a predefined taxonomy

Steffen Kortlang[1], Jens Schröder[1], Danilo Hollosi[1], Jörn Anemüller[2] and Birger Kollmeier[1,2]
[1] *Fraunhofer IDMT / Hearing, Speech and Audio Technology, Oldenburg*
[2] *University of Oldenburg, Institute of Physics, Medical Physics*

Abstract

Motivated by the success in the field of automatic music genre classification and document classification, a hierarchical approach for content-based classification of environmental sounds with a predefined, tree-structured taxonomy is presented in this paper. The application is evaluated with respect to classification accuracy and compared with the flat approach. With average classification rates over 90%, better results than in the literature can be achieved, whereat on average the hierarchical approach provides even better outcomes than the flat. The results allow an assessment of the usability and limitations of such a system also for other classification scenarios.

Introduction

Audio classification usually implies the automatic assignment of an audio signal to predefined categories that are arranged in a flat (or direct) way. In contrast, hierarchical classification schemes assume a classification problem to be separable into smaller, independent and in general easier-to-solve tasks. This involves the advantage, that the natural systematization is considered. Furthermore, this approach affords an opportunity to make node-specific settings, such as a special choice of features or parameters. Misclassifications may occur at lower levels of the hierarchy. Thus, at least a rough classification is possible. On the other hand, the implementation contains higher complexity as well as high computational costs. The main disadvantage contains the fact that error rates are multiplied along the paths of the hierarchy.

Classification System

The implemented Matlab-based classification system is presented in this section. It provides a short overview of the audio taxonomy, the extracted features, the feature subset selection and the classifier used. A detailed description of the system is given in [1].
The obtained taxonomy contains a total number of 9 classes of environmental sounds, that were partly collected from several databases and partly new recorded. They are shown in Figure 1 (rectangular boundary). According to the proposed categorization of environmental sounds by [2], the sounds are separated into continuous sounds, impacts and harmonic sounds (such as vocalizations or signals). The features for the classification are extracted frame-based with a window size of 25 ms and an offset of 10 ms. In addition to the classical MFCCs, temporal features, energy features, harmonic features and perceptual features are considered for each of these frames. The spectral features are calculated not only for the whole spectrum, but also for 21 bark and 6 octave bands, so that a huge amount of 476 instantaneous features are available. As those descriptors are temporal modelled (calculation of delta features, lowpass filtered and standard deviation features), a total number of 1904 features are computed for each window. In a last step, the features are normalized. Because of the limited computational power and the curse of dimensionality, a dimension reduction of the D-dimensional feature space F seems inevitable. Commonly, this reduction is realized by a feature transformation (Linear discriminant analysis), where linear combinations of the original features are computed. Therefore, the physical meaning of the descriptors gets lost. In addition, all the features have to be computed. In contrast, the feature selection algorithms are seeking for a smaller feature space F' with the $D' < D$ most relevant and non-redundant descriptors. The minimum-redundancy-maximum-relevance feature-selection by [3], based on the mutual information I is implented. The cost function J in equation (1) considers a maximal statistical correlation between the features f_i in the feature subset F' and the class label $\vec{c}$ and a min-

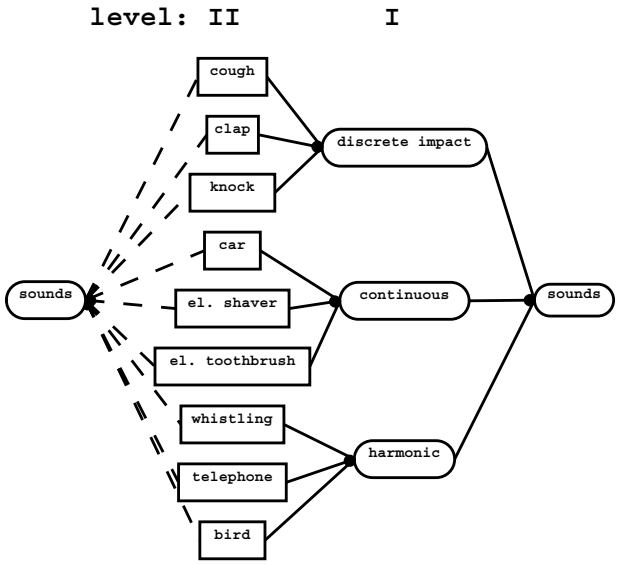

Figure 1: Audio taxonomy for the presented classification system. The flat approach (on the left) is indicated by dashed, the hierarchical one (on the right) by solid lines.

imal mutial information among the features themselves:

$$J = \frac{1}{D'} \sum_{f_i \in F'} I(f_i; \vec{c}) - \frac{1}{D'^2} \sum_{f_i, f_j \in F'} I(f_i, f_j) \qquad (1)$$

The features are selected in a sequential feed-forward step. Therefore, the descriptors for each node and tree level are selected within the training process in such a way that the overall error-rate is minimized. In a first approach, a number of $D' = 10$ features are considered for each judgement stage in the hierarchy. In the training process, an individual Gaussian classifier (3-GMM) is developed and trained with the node-specific features for each node in the tree, so that each class is modelled in the feature space in terms of a reference. The file-based posterior probabilities for each class are calculated in the test phase. In the hierarchical approach, the (independent) probabilities are multiplied accumulatively among the paths in the hierachy. Finally, a hard decision is enforced with Bayes classification.

Evaluation

The data sets are divided into train and test data in the ratio of 2:1 with a 3-fold cross-validation. The confusion matrix in figure 2 (top) show the classification results for the flat approach at layer II. Similar classes (especially impacts) are confused. Furthermore, the car tends to be classified as knocking and the toothbrush as a bird. The average classification rate amounts to $(92.18 \pm 2.06)\%$. For the hierarchical classification, the average recognition rate increases to $(95.72 \pm 2.45)\%$, as adjacent classes can better be distinguished. The results are summarized in table 1. The average classification rates at level I indicate, that the advantages of the hierarchical approach yet occur at level I. Furthermore, the independant classification rates for the hierarchical approach are given. They are comparable for both levels.

Table 1: Average classifiaction rates for the 3 classes on level I and the 9 classes on level II using the flat and hierarchical approach (independent vs. accumulative). The standard deviation is applied to the iterations of the 3-fold cross-validation.

level	approach		
	flat	hier(ind)	hier(acc)
I	**95.45%**± 1.87%	98.09%± 1.35%	**98.09%**± 1.35%
II	**92.18%**± 2.06%	97.39%± 1.37%	**95.72%**± 2.45%

Conclusions

The presented classification system achieves relatively high accuracies. Regardless, a reliable comparison across other classification systems in the literature is difficult because of different classes and data sets. The hierarchical approach produces even better results compared to the flat system, but this surely depends on the user-defined taxonomy. The hierarchical structuring of the classes in audio classification delivers special benefit in distinguishing similar classes. A sufficient quantity of features, an efficient feature selection algorithm and a reasonable taxonomy is to be recommended in order to exploit the full potential of a hierarchical approach. This makes it a promising base for further developments.

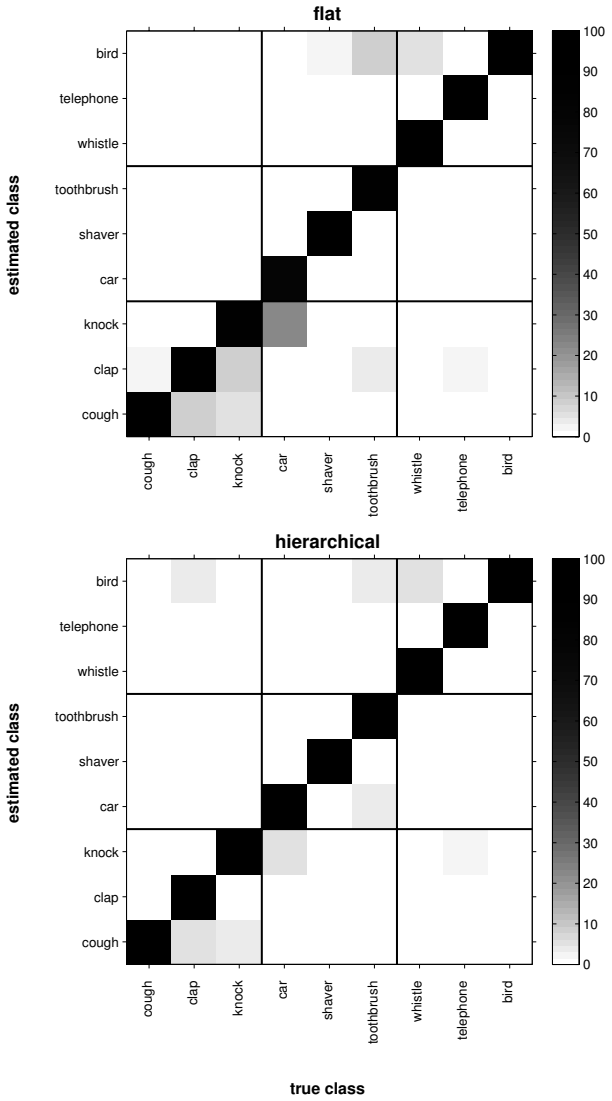

Figure 2: Confusion matrizes for the flat (top) and the hierarchical (bottom) approach at layer II. While the diagonal elements presents a correct decision, the misclassifications can be seen in the non-diagonal fields.

References

[1] Kortlang, S.: Ein hierarchisches Modell zur inhaltsbezogenen Audio-Klassifikation. Masterarbeit. CvO Ossietzky Universität Oldenburg, 2011

[2] Gygi, B.; Kidd, G.; Watson, C.: Similarity and categorization of environmental sounds. Perception & Psychophysics Bd. 69 (2007), 839–855

[3] Peng, H. ; Long, F. ; Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence Bd. 27 (2005), 1226–1238