RECOGNISING ACOUSTIC SCENES WITH LARGE-SCALE AUDIO FEATURE EXTRACTION AND SVM

Jürgen T. Geiger[1], *Björn Schuller*[2,1], *Gerhard Rigoll*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Institute for Sensor Systems, University of Passau, Germany

geiger@tum.de

ABSTRACT

This work describes our contribution to the IEEE AASP Challenge on classification of acoustic scenes. From the 30 second long highly variable recordings, spectral, cepstral, energy and voicing-related audio features are extracted. A sliding window approach is used to obtain statistical functionals of the low-level features on short segments. SVM are used for classification of these short segments, and a majority voting scheme is employed to get a decision for the whole recording. On the official development set of the challenge, an accuracy of 73 % is achieved. A feature analysis using the t-statistic showed that mainly Mel spectra were the most relevant features.

Index Terms— Computational auditory scene analysis, acoustic scene classification, audio feature extraction, computer audition

1. INTRODUCTION

Recognising the acoustic background is known as *acoustic scene classification* and can be counted to the field of *computational auditory scene analysis*. Typically, several different sound sources contribute to the scene, making it a complex combination of different acoustic events. The IEEE AASP Challenge on classification of acoustic scenes provides a test bed for comparison of different systems for acoustic scene classification. The employed corpus consists of a development set and a test set, having ten different classes with ten recordings of 30 s each (in both sets), whereby classification is carried out with cross-calidation on each set.

Previous work on acoustic scene classification investigated the application of various spectral, energy and voicing-related features, in combination with neural networks [1]. In [2], we showed how, in the case of small amounts of training data, new acoustic events can be *learned* by a system. The CLEAR evaluation provided a testbed for different systems for the detection and classification of acoustic events [3].

This work describes our contribution to the IEEE AASP Challenge on classification of acoustic scenes. From the recordings, a number of spectral, cepstral, energy and voicing-related audio features are extracted. A sliding window approach is used to obtain statistical functionals of the low-level features on short segments of several seconds. SVM (Support Vector Machines) are used for classification of these short segments, and a majority voting scheme is employed to get a decision for the whole recording. On the official development set of the challenge, an accuracy of 73 % is achieved. It is expected, that the results generalise well to the test set.

This research was supported by the ALIAS project (AAL-2009-2-049) co-funded by the EC, the French ANR and the German BMBF.

2. METHODOLOGY

This section describes the employed techniques for feature extraction and classification.

2.1. Feature Extraction

The recordings of the acoustic scenes contain a high number of different sound sources of different nature. In order to extract all relevant information, we employ a large feature set of audio features. It was originally designed for speech processing, but fits well general audio analysis owing to its many spectral and further descriptors. The employed features can be grouped into cepstral, spectral, energy-related and voicing features. In previous works on the classification of acoustic events, such as [4], similar audio features have been used. In order to foster reproducibility, we use our open-source feature extraction toolkit openSMILE [5]. The employed feature set is the official openSMILE `emo_large.conf` feature set. All low-level descriptors (LLD) are listed in Table 1. From the low-level descriptors, 39 statistical functionals computed after adding delta and acceleration coefficients. The functionals include values such as mean, standard deviation, percentiles and quartiles, linear regression functionals, or local minima/maxima related functionals. Finally, all features are normalised. In total, the number of features sums up to 6 669 (57 LLD, δ, $\delta\delta$ $\times$ 39 functionals).

To better capture the non-stationary nature of the scenes, we employ a windowing approach when computing the statistical functionals. The whole recording is splitted into (overlapping) windows with a length of several seconds, and the functionals are computed for each of those segments. Experiments on the development set showed that a segment length of 4 s and 50 % overlap is the most stable configuration. These segments can capture the different acoustic events contributing to the acoustic scenes. This window approach is used on the training data and on the test data. Thus, models are trained with 1 200 training instances per class (80 training recordings $\times$ 15 windows per recording).

2.2. Classification

SVM as implemented in the Weka toolkit [6] are used for classification. They are very well suited for this problem because of the small

Cepstral features (13)
MFCC 0 – 12
Spectral features (35)
Mel-Spectrum bins 0–25, zero crossing rate, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, centroid, relative position of spectral maximum and minimum
Energy features (6)
logarithmic energy, energy in bands from 0 – 250 Hz, 0 – 650 Hz, 250 – 650 Hz, 1 – 4 kHz, 3010 – 9123 Hz
Voicing-related features (3)
F0 (subharmonic summation (SHS) followed by Viterbi smoothing), F0 envelope, probability of voicing

Table 1: 57 cepstral, spectral, energy and voicing-related aoustic low-level descriptors (LLD).

number of classes and small amount of training data per class. SVM with a linear kernel and complexity 1.0 are used. They are trained with the sequential minimal optimisation (SMO) algorithm and the windowed training data.

Classification is performed on the windowed test data. Each of the 4 s long windows is separately fed to the classifier to recognise one part of the scene. Thereby, for each test instance, 15 classification results are obtained. In order to get one decision for the whole instance, a majority voting scheme is employed. A majority vote is made over all separate classification results to decide for the recognised class. Weighting the single classification results by their confidence (obtained by fitting the output of the SVM to a logistic regression model) brought no improvement, and thus, the majority vote is not weighted.

3. EXPERIMENTS

All experimental results reported here are obtained with the official challenge development set. This dataset contains 100 recordings of 10 different classes. Experiments are conducted using a 5-fold cross validation. In the first fold, files 1 and 2 of each class are used for testing and the rest for training. Files 3 and 4 are used for testing in the second fold, 5 and 6 in the third fold, etc. Final results are obtained by summing up results for each fold.

3.1. Results

Without using the window approach, an accuracy of 60 % is obtained. Segmenting the training and test data into windows with a length of 4 s and a window shift of 2 s results in an increased accuracy of 73 %. Table 2 shows results for experiments with two different feature sets: using only MFCCs (13 MFCCs + delta + acceleration coefficients and their functionals) or using the full feature set (MFCC, spectral, energy, voicing). Both systems are tested with the simple feature extraction approach where features are computed over the whole recording, or with the windowing approach.

	simple	window
MFCC	50 %	67 %
MFCC + Spectral + Energy + Voicing	60 %	73 %

Table 2: Accuracies for two different feature sets (MFCC vs. all features), using the simple feature extraction method or the window approach.

window length (s)	2	3	4	5	6
accuracy (%)	71	67	73	66	64

Table 3: Accuracies for different window lengths, keeping the window shift constant at 2 s.

The results in Table 2 demonstrate that the proposed feature set improves the system performance compared to only MFCC features. Furthermore, the employed window approach can succesfully increase the accuracy.

The influence of the window length on the classification accuracy can be seen in Table 3. With a window shift of 2 s, smaller window sizes generally lead to better accuracy, whereby the best result is achieved with a window length of 4 s.

Table 4 shows the confusion matrix for the best-performing system, using all proposed features and the employed window approach. Some classes (*bus*, *busystreet*) are recognised with 100 % accuracy, while for others (*park, restaurant, tube*), scores as low as 40 % are obtained. Most confusions are made between the classes *park* and *quietstreet* or *restaurant* and *supermarket*. The recordings of the classes *park* and *quietstreet* are partly very similar. Recordings of the class *tube* and *tubestation* contain a high variability, depending on the actual occuring acoustic events. Therefore, these classes are confused with several other classes.

It has to be noted that, given the small size of the dataset, a large improvement in accuracy is needed to make it a significant improvement. When results in the order of 60 % are achieved, the accuracy has to be improved by roughly 12 % to be significant. Significance was evaluated using a one-sided z-test and a p-value of 5 %.

3.2. Feature Analysis

In order to understand the contribution of different features to the classification result, we performed a feature analysis. For each of the employed 6 669 features, a score is computed using a t-test. The t-statistic is computed for each pair of acoustic classes and is summed up over all pairs to obtain a single score for each feature. This score is computed for each fold, using the training data of this fold. Summing up the scores over all folds results in a feature ranking. Using this feature ranking, another set of experiments is conducted, starting with the 10 best features and gradually adding more features until the whole feature set is used. Figure 1 shows the results of this experiment.

It can be seen that already with 150 audio features, an accuracy of 65 % is obtained. Adding more features does not improve the accuracy. Only with 6 000 features or the full feature set, the accuracy improves. The top 150 features contain mostly Mel spectra (116, whereby lower-order components are represented more often), but also energy (14), MFCC (14, only from the 12th component), spectral flux (2) and position of spectral minimum (4). Most of the functionals are equally represented in the top 150 features. However, it

	bus	*busystreet*	*office*	*openairmarket*	*park*	*quietstreet*	*restaurant*	*supermarket*	*tube*	*tubestation*
bus	**10**	0	0	0	0	0	0	0	0	0
busystreet	0	**10**	0	0	0	0	0	0	0	0
office	0	0	**9**	0	0	0	0	1	0	0
openairmarket	0	0	0	**9**	0	0	0	1	0	0
park	0	0	0	0	**5**	5	0	0	0	0
quietstreet	0	0	0	1	2	**7**	0	0	0	0
restaurant	0	0	0	2	0	0	**4**	4	0	0
supermarket	0	0	0	0	0	0	1	**8**	0	1
tube	0	1	0	0	0	0	1	2	**5**	1
tubestation	1	0	0	0	1	1	1	0	0	**6**

Table 4: Confusion Matrix of the development data for the proposed system, achieving an accuracy of 73 %.

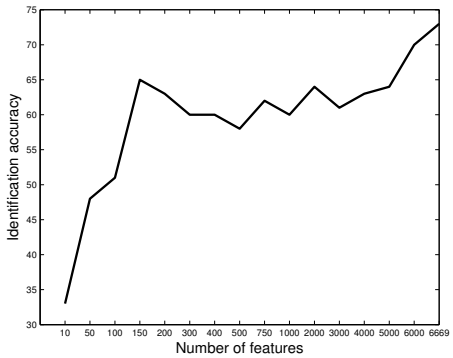

Figure 1: Influence of number of features on the accuracy.

stands out that 88 of them are variants of a mean value (e. g., mean of absolute values, mean of non-zero values, quadratic mean).

4. CONCLUSIONS

We presented our approach to the IEEE AASP Challenge on classification of acoustic scenes. Using large-scale audio feature extraction and SVMs, an accuracy of 60 % is obtained on the development set of the challenge. This is an improvement compared to the result of 50 % when only MFCC features are used. A feature analysis showed that Mel spectra were an important factor for this improvement. Furthermore, the other energy and spectral features helped to better capture the information in the acoustic scenes. Segmenting the data into windows of $4s$ (and 50 % window overlap) resulted in an increased accuracy of 73 %. The employed window approach makes the system more stable, since the contributing acoustic events are better captured in these small windows. Some acoustic scenes (*park, restaurant, tube, tubestation*) are difficult to recognise due to the high variability in the class and the similarity between the different classes.

Future work includes applying source separation techniques such as non-negative matrix factorisation (NMF) to better recognise the different sources contributing to the scene.

5. REFERENCES

[1] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 20, no. 1-2, pp. 61–79, 1998.

[2] J. T. Geiger, M. A. Lakhal, B. Schuller, and G. Rigoll, "Learning new acoustic events in an hmm-based system using map adaptation," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 293–296.

[3] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*. Springer, 2007, pp. 311–322.

[4] A. Temko and C. Nadeu, "Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering," in *Proc. ICASSP*, Philadelphia, PA, USA, 2005, pp. 502–505.

[5] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.