LARGE-SCALE AUDIO FEATURE EXTRACTION AND SVM FOR ACOUSTIC SCENE CLASSIFICATION

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 a system for acoustic scene classification using large-scale audio feature extraction. It is our contribution to the Scene Classification track of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (D-CASE). The system classifies 30 second long recordings of 10 different acoustic scenes. From the highly variable recordings, a large number of spectral, cepstral, energy and voicing-related audio features are extracted. Using a sliding window approach, classification is performed on short windows. SVM are used to classify these short segments, and a majority voting scheme is employed to get a decision for longer recordings. On the official development set of the challenge, an accuracy of 73 % is achieved. SVM are compared with a nearest neighbour classifier and an approach called Latent Perceptual Indexing, whereby SVM achieve the best results. A feature analysis using the t-statistic shows that mainly Mel spectra are the most relevant features.

***Index Terms*—** Computational auditory scene analysis, acoustic scene recognition, feature extraction

1. INTRODUCTION

Recognising the acoustic background is known as *acoustic scene classification* and can be counted to the field of *computational auditory scene analysis* [1]. Typically, several different (overlapping) sound sources contribute to the scene, making it a complex combination of different acoustic events.

Previous work on acoustic scene classification investigated the application of various spectral, energy and voicing-related features, in combination with neural networks [2]. In [3], a system for acoustic scene recognition is described and evaluated. That system uses various audio features and a nearest neighbour (NN) classifier. The system for acoustic scene recognition described in [4] used Support Vector Machines (SVM), embedded in a hierarchical or parallel framework. The CLEAR evaluation provided a testbed for different systems for the detection and classification of acoustic events [5], which is a related problem. In [6], we showed how, in the case of small amounts of training data, new acoustic events can be *learned* by a system. One possible application of acoustic scene classification is a system as described in [7], where cyclist's routes are recognised using scene recognition techniques.

In the scene classification track of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (D-CASE), systems for acoustic scene recognition are compared. The employed corpus (divided into a development set and a non-public test set) is categorised into 10 different classes of (mostly outdoor) acoustic scenes.

This contribution describes our method for acoustic scene classification. From the recordings, a large 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 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 test set of the challenge, an accuracy of 69 % is achieved. Furthermore, we compare our approach to a method based on Latent Perceptual Indexing [8] and to a NN classifier as it was used in [3]. The employed database, audio features and classification methods are described in Section 2. Experimental results are presented in Section 3, and some conclusions are given in Section 4.

2. METHODOLOGY

2.1. Database

For evaluation of our system, we employ the official dataset of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events [9]. Thereby, we use only the data of the scene classification track. This dataset contains 30 s recordings of various acoustic scenes, categorised into ten different classes. For each of the ten classes, the database contains ten recordings, summing up to 100 recordings and 50 minutes total length. In addition, for the challenge, the systems were evaluated with a non-public testset containing similar data. Sounds were recorded with a high-quality binaural recording system, whereby the portability and subtlety of the system allowed to obtain unobstructed everyday recordings with relative ease. Since the recordings were performed with binaural microphones on the ears of a person, the head-related transfer function (HRTF) of that person is intrinsically incorporated.

2.2. Feature Extraction

Prior to feature extraction, the stereo recordings are mixed down to mono. This causes a loss of information, while, on the other hand, we simulate realistic, simple conditions with devices like mobile phones. In order to foster reproducibility, we use our open-source feature extraction toolkit openSMILE [10]. Since the recordings of the acoustic scenes contain a high number of different sound sources of different nature, a large set of different audio features is extracted in order to extract all relevant information. The employed feature set is the official openSMILE `emo_large.conf` feature set. This feature set was originally designed for speech processing, but fits well general audio analysis owing to its many spectral

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

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 acoustic low-level descriptors (LLD).

and further descriptors. In previous works on the classification of acoustic scenes and events, such as [3, 11], similar audio features have been used. All low-level descriptors (LLD), which are listed in Table 1, are extracted every 10 ms from 25 ms frames. The employed features can be grouped into cepstral, spectral, energy-related and voicing features. In addition to MFCC and Mel spectra, spectral roll-off points and other spectral features contribute to a comprehensive description of the spectrum. Furthermore, a number of energy-related features is computed. Since the recordings in the test data contain a considerable portion of speech, our features also include a small set of voicing-related features. From the 57 low-level descriptors, 39 statistical functionals are 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, whereby the statistics of the training set are used to normalise the test set as well. In total, the number of features sums up to 6 669 (57 LLD, δ, $\delta\delta \times 39$ functionals).

2.3. Classification

To better capture the non-stationary nature of the scenes, classification is performed on smaller windows. Each recording is split into (overlapping) windows with a length of several seconds, and the statistical functionals are computed for all LLD in those segments. In [3, 8], a window length of 1 s is applied. However, our experiments showed that longer windows lead to better results. These segments can capture the different acoustic events contributing to the acoustic scenes. This windowing is performed on the training data and on the test data. Thus, models are trained with a larger number of training instances per class (80 training recordings $\times$ N_w windows per recording). Classification is performed on the windowed test data. Each of the N_w windows is separately fed to the classifier to recognise one part of the scene. In order to get one decision for the whole instance, a majority voting scheme is employed. Weighting the single classification results by their confidence (which, in the case of SVM as classifier, is obtained by fitting the output of the SVM to a logistic regression model) brought no improvement, and thus, the majority vote is not weighted.

For classification, we propose to use SVM. The implementation of the Weka toolkit [12] is used. SVM are very well suited for this problem because of the small number of classes and small amount of training data per class. We employ SVM with a linear kernel and complexity 1.0. They are trained with the sequential minimal optimisation (SMO) algorithm using the windowed training data.

For comparison, a NN classifier is tested. Preliminary experiments showed that NN performed better than k-NN. As a distance function, the Euclidean Distance is used for the NN classifier. Since the features are normalised, this distance function gives better results compared to, e. g., the cosine distance.

2.4. Latent Perceptual Indexing

In addition to SVM and NN, we implement an approach for acoustic scene classification based on Latent Perceptual Indexing (LPI), as presented in [8].

With our employed SVM or k-NN approach, due to the windowing of training and test data, the contributing parts of each acoustic scene are recognised separately. Thereby, all occuring acoustic events are processed and classified on their own. With a majority voting, a decision for the whole recording is made. This approach ignores the overall composition of a sound scene, recognising only the occurring sources, without using a higher-level decision logic. Such a decision logic could take account of all occurring sounds and decide on the acoustic scene based on the mixture of single acoustic events.

LPI is an approach where the classification of the acoustic scene is made based on the composition of the contributing sounds. Each recording is represented as a vector in a latent perceptual space. First, a clustering (using K-means) of all (windowed) training data is performed to obtain a number of reference clusters, that is higher than the number of occurring classes. Then, for each recording, a *bag of feature vectors* is computed using the windows. The recording is then transferred into the latent space by counting the occurrence of each of the reference clusters for this recording. The dimension of this latent space is the number of reference clusters. Thus, all training and test recordings are each described by a vector in the latent space. Transformed test recordings are then classified using a NN classifier and the cosine vector similarity. Preliminary experiments showed that, in the latent space, a NN classifier is as good as using SVM. To obtain a more fine-grained representation in the latent space, smaller window sizes are used in the LPI method, compared to the SVM or NN classifier.

One disadvantage of LPI compared to SVM is that LPI requires more training data. In the SVM approach, classification is performed on the window level, whereby the training data are windowed as well. Thus, in the SVM approach, from N training recordings, $N \times N_w$ training instances are available. On the contrary, the LPI approach represents each training recording as a single vector in the latent space, and therefore the number of training instances is equal to the number of training recordings. In our experimental validation, there are only small amounts of training data, which is probably not enough for the LPI method.

3. EXPERIMENTS

This section describes our experimental setup and results. All implemented systems are evaluated with the development set of the database of acoustic scenes from the D-CASE challenge as described in Section 2.1. For evaluation, 5-fold cross validation is

	simple	window
MFCC	50 %	68 %
All	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)	1	2	3	4	5	6
ACC (%)	62	71	67	73	66	64

Table 3: Accuracies for different window lengths, keeping the window shift constant at 2 s (except window length 1 s, where the shift is 1 s.). SVM is used as classifier, with the full feature set.

performed, which is the official protocol of the D-CASE challenge. The evaluation measure is the average accuracy (ACC in %) over all folds, whereby additionally, the 95 % confidence interval is reported. It has to be noted that, given the small size of the dataset, the minimum significant improvement is relatively large. 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 %.

3.1. Windowing Approach

As a first experiment, we analysed the influence of the windowing approach. In this experiment, all 6 669 features or only the MFCC features are used in combination with a SVM classifier. Table 2 lists the result for those two feature sets, either using the window approach (4 s windows with 50 % overlap) or performing classification on the whole recordings. It can be seen that with MFCC features and without the windowing method, an accuracy of 50 % is obtained, which can be considered a rough baseline. Segmenting training and test data into 4 s windows and making a majority vote over single window classification results increases the accuracy to 68 %. On top of that, adding the other energy, spectral, and voicing-related features improves the result (not significantly) to 73 %.

Next, we analyse the influence of the window length. Results (using the whole feature set and SVM) are shown 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. This is in contrast to findings in [3], where a window size of 1 s was found to be optimal. In that study, more training data were used, which made it possible to apply a finer resolution of the data. Furthermore, the database was divided into more acoustic classes, which made such a finer resolution necessary in order to distinguish between these classes.

3.2. SVM and NN Results

We tested different feature configurations when comparing SVM and NN classifiers. In addition to using the full set of all LLD and all functionals, smaller feature sets are obtained by taking only the MFCC features and/or using only mean and variance instead of all functionals. Table 4 shows results for experiments with different feature configurations. The best performance (73 %) is obtained with the full feature set and SVM as classifier, while using only MFCC features (68 %) is slightly worse. Reducing the set of functionals to only the mean and variance of each LLD leads to a degradation in performance to 64 %, for both feature sets. These interesting results show that the additional features (compared to only MFCC) are only relevant when being combined with the large set of functionals. For the NN classifier, acceptable results were only obtained with the reduced set of functionals, resulting in an accuracy of 50 % for all LLD and 39 % for MFCC. Interestingly, for the NN classifier, there is a larger difference in the performance of MFCC vs. all LLD.

Features	func.	class.	ACC (%)
All LLD	All	SVM	73 ± 5
MFCC 0-12	All	SVM	68 ± 5
All LLD	mean, var	SVM	64 ± 9
MFCC 0-12	mean, var	SVM	64 ± 9
All LLD	mean, var	NN	50 ± 12
MFCC 0-12	mean, var	NN	39 ± 6

Table 4: Results for different features, functionals (func.) and classifiers (class.) with accuracy (ACC) in % and 95 % confidence interval.

	bus	*busyst.*	*office*	*open.*	*park*	*quietst.*	*rest.*	*superm.*	*tube*	*tubest.*
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 5: Confusion Matrix of the development data for the proposed system, achieving an accuracy of 73 %.

Table 5 shows the confusion matrix for the best-performing system, using all proposed features with SVM 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.

3.3. LPI Approach

For the LPI approach, the best results were obtained with MFCC features and all functionals. Furthermore, a smaller window size led to better results. We used 1 s windows without overlap. The results for different numbers of clusters are shown in Table 6.

The best result (46 ± 10 %) is obtained with 500 clusters. Generally, the performance is similar to the NN approach. Considering that the classification is performed on the whole recordings instead of windowed data, the accuracy is also comparable to the SVM system with MFCC and without the window approach. Generally

# cluster	10	20	50	100	200	500	1000
ACC (%)	32	36	44	42	43	46	44

Table 6: Results for the LPI approach with MFCC features for different number of clusters.

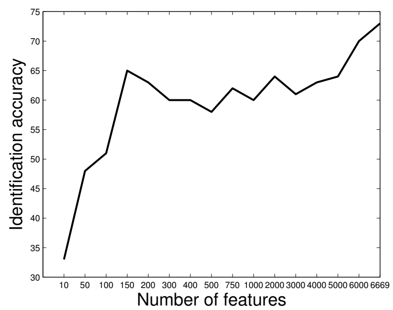

Figure 1: Influence of number of features on the accuracy (note the nonlinear scale of the horizontal axis).

speaking, such an LPI approach requires more training data to deliver better results.

3.4. Feature Analysis

In order to better 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. For this analysis, SVM is chosen as a classifier. Fig. 1 shows the results of this experiment. Generally, with increasing number of features, the accuracy increases, with an outlier at 150 features and 65 % accuracy. This could either be a (statistically not significant) outlier, or it could be the case that the subsequent features are worse and therefore deteriorate the performance. 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 12 th component), spectral flux (2) and position of spectral minimum (4). Comparing this result to Table 4, where MFCC achieved a similar performance, the conclusion is that MFCC and Mel spectra perform equally well in this task. Most of the functionals are equally represented in the top 150 features. However, it 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). The results of the feature analysis are in line with the results presented in Table 4, showing the performance gain of the employed large feature set. Already with a comparably small feature set (MFCC or Mel spectra, with mean and variance as functionals), a relatively good performance is achieved. Adding more LLD and functionals leads to a small but substantial improvement.

3.5. Test set

Our best system configuration (all LLD, all functionals, SVM classifier on 4 s windows with 50 % overlap) achieves an accuracy of 69 $\pm$ 12 % on the non-public test set of the employed corpus. This result shows that our system generalises well to a previously unseen testset. Although the accuracy is slightly worse than on the development set (73 %), it can be concluded, that, even with such a large feature set, there is no overfitting of the system to the development set.

4. CONCLUSIONS

We presented and evaluated a system for acoustic scene classification. Using large-scale audio feature extraction and SVM, an accuracy of 73 % is obtained on the development set of the D-CASE challenge. A feature analysis showed that Mel spectra are among the most important features and perform equally well compared to MFCC. Furthermore, the other employed energy and spectral features helped to better capture the information in the acoustic scenes. 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.

5. REFERENCES

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley interscience, 2006.

[2] 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.

[3] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, “Computational auditory scene recognition,” in *Proc. ICASSP*, Orlando, FL, USA, 2002.

[4] H. Jiang, J. Bai, S. Zhang, and B. Xu, “Svm-based audio scene classification,” in *Proc. Natural Language Processing and Knowledge Engineering (NLP-KE)*. IEEE, 2005, pp. 131–136.

[5] 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.

[6] 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.

[7] B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, and L. Paletta, “Acoustic geo-sensing: Recognising cyclists’ route, route direction, and route progress from cell-phone audio,” in *Proc. ICASSP*, Vancouver, Canada, 2013.

[8] O. Kalinli, S. Sundaram, and S. Narayanan, “Saliency-driven unstructured acoustic scene classification using latent perceptual indexing,” in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2009.

[9] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*. Technical Report, Queen Mary University of London, 2013.

[10] 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.

[11] 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.

[12] 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.