

## Chapter 12

# Applications in Intelligent Sound Analysis

*If you develop an ear for sounds that are musical it is like developing an ego. You begin to refuse sounds that are not musical and that way cut yourself off from a good deal of experience.*

—John Cage

Apart from the more specific types of sound considered so far—speech and music—general sound can also carry relevant information. This is, however, a considerably less researched field up to-date. Most prominent in this area are the tasks of acoustic event detection (AED) and classification (AEC) [1] that can be subsumed under the area of computational auditory scene analysis (CASA) [2]. For these tasks international evaluation campaigns exist that have mostly seen HMM and SVM approaches with various acoustic features [1]. Fields of application include media retrieval [3] including affective content analysis [4] or human-machine and human-robot interaction [5], animal vocalisation recognition [6], and monitoring of industrial processes [7]. Mostly, closed-set recognition is addressed, i.e., training and testing classes are the same. Recently, however, also open-set recognition is faced, the so-called novelty detection [8, 9].

As before, examples of application have been chosen for illustration of obtainable performances and methods employed. Three applications have been chosen to cover a good variety of the above named use cases: Firstly, recognition of animal vocalisation [10], then, acoustic event classification including unsupervised learning to exploit the availability of sheer infinite amounts of sound on the Internet [11], and finally prediction of the emotion evoked in human listeners of sound [12] in analogy to the sections on speech and music.

## 12.1 Animal Vocalisations

As a first example of application in more general sound analysis, we will consider animal voices instead of human voices. The following application investigates the recognition of animal vocalisations ‘in the wild’ [10].

In the field of bioacoustics, a multiplicity of approaches exists for classifying animal sounds, for example to monitor populations of certain species, such as whales [13] or birds [14]. More recently, increasing efforts are invested in digitisation of sound archives. Similarly as in the case of MIR, this demands for efficient indexing and retrieval methods. For example, in [15], an effective indexing algorithm for animals with curve-like harmonic vocalisations, such as various species of birds, was presented and evaluated on bird songs contained in the Animal Sound Archive (“Tierstimmenarchiv”) of the Humboldt-University of Berlin [16]. This data set will be referred to as ‘HU-ASA database’ in the ongoing. In the past, SVM-based static classification using segment-wise functionals [17] (e.g., mean and standard deviation) was proposed for animal sounds classification [18]. Alternatively, dynamic classification, e.g., by HMMs [19] or by suited neural networks [6] is reported successful in the literature. Hence, we will consider SVMs, HMMs with different topologies, and LSTM recurrent neural networks on the HU-ASA database in the ongoing.

### 12.1.1 HU-ASA Database

The evaluation database builds on the large HU-ASA database of animal vocalisations. It is annotated with the species and additional meta-data such as recording conditions and the type of vocalisation for each audio file. 1418 audio files are available in MP3 encoding. These were obtained from the on-line archive.<sup>1</sup> Per species, the audio files with (biological) class were automatically annotated (e.g., *Aves*, *Mammalia*), order (e.g., *Passeriformes*, *Primates*), and family (e.g., *Felidae*, *Canidae*) according to the Linnaean rank-based biological classification as retrieved from Wikipedia.<sup>2</sup> The majority of the HU-ASA’s instances consist of bird (*Aves*) and mammal (*Mammalia*) sounds, as shown in Table 12.1. The class ‘Others’ include *Sauropsida*, *Hexapoda*, and recordings without automatic annotation, where according information was missing in Wikipedia. The total audio duration is 20 423 s (5 h 40 min 23 s). *Amphibia*, *Insecta*, and *Reptilia* were not included in the described experiments given their sparseness (cf. Table 12.1).

Two tasks of practical interest were derived from the biological classification, as shown in Table 12.2. The first (2-class) task aims at classification of songbirds (*Passeriformes*) versus non-songbirds (*Non-Passeriformes*). Non-songbirds include by number of instances the orders *Anseriformes*, *Charadriiformes*, *Galliformes*, *Psittaciformes*, *Gruiformes*, and 24 other orders—often with sparse instances.

---

<sup>1</sup> <http://www.tierstimmenarchiv.de>, accessed mid 2010.

<sup>2</sup> <http://www.wikipedia.org>

**Table 12.1** Number of instances, as well as min(imum), mean, max(imum), and total recording length ( $\Sigma$ ) of the audio files by the biological class of the species in the HU-ASA database

(Biological) Class	# Instances	Duration [s]			
		Min	Mean	Max	Sum
Aves	868	2.4	14.8	64.7	12210
Mammalia	487	1.0	14.7	37.7	6954
Amphibia	27	1.8	19.6	65.9	529
Reptilia	7	11.2	22.5	39.6	157
Insecta	19	2.3	16.0	30.1	287
Other	10				133
Sum	1418				20423

**Table 12.2** Distribution of instances in the 2-class (*Passeriformes*/*Non-Passeriformes*) and 5-class tasks as defined on the HU-ASA database

Class	# Instances
Passeriformes	282
Non-Passeriformes	586
Sum	868
Primates	90
Canidae	43
Felidae	62
Sum	1063

The more complex 5-class task adds mammals (*Mammalia*) of the families *Felidae* and *Canidae*, as well as the instances of the biological order *Primates* (cf. Table 12.2). A particular challenge arises from the real-world nature of the database: vocalisations of the same species often vary considerably, depending on the situation and stance (i.e., aggression or warning calls), and age of the animals, from young to full-grown. The recordings are further corrupted by background noises—even of other animal species.

### 12.1.2 Methodology

Static classification by SVMs bases on linear kernel SVM. For dynamic classification, two topologies of HMMs and LSTM RNNs are compared. A typical HMM topology in audio (and general sequence) classification is a linear (left-right) layout: With  $N$  as the number of states in total, state transitions are allowed from state  $i = 1, \dots, N - 1$  to states  $i$  and  $i + 1$ . However, animal vocalisations are often highly repetitive, motivating the usage of a *cyclic* topology. In such a layout a transition from state  $N$  to the first state is added. In the following experiments the number of states is fixed to  $N = 8$  basing on a series of evaluations.

As for neural networks, e.g., a feedforward MLP was used for classifying animal vocalisations in [6]. To enhance the neural network paradigm by extended memory capabilities, LSTM networks are considered here with one hidden layer of 100 LSTM memory cells. The size of the input and output layers was equal to the number of features and classes to discriminate. Softmax functions were applied to the output activations, and the resulting values were normalised to the sum one to provide posterior class probabilities.

MFCCs 1–12 along with energy and their first ( $\delta$ ) and second order ( $\delta\delta$ ) regression coefficients were chosen as features for frame-level classification due to their suitability across a multiplicity of Intelligent Audio Analysis tasks [17–19]. In [19], these features were found superior to the MPEG-7 spectral projection features as used in [15] for sound classification with HMMs. The overall 39-dimensional feature set will be denoted by ‘MFCC’.

For static classification of larger audio chunks, functionals are applied. In [17], mean and standard deviation were proposed. The functionals considered in the ongoing also include extremes and higher-order moments [20]. Additional LLDs for include HNR, pitch and ZCR by using openSMILE’s (cf. Sect. 6.5, [21]) INTER-SPEECH 2009 Emotion Challenge set [20], as described in Table A.1. This choice could allow to discriminate between animals with voiced and unvoiced sounds. The functionals of the 32 LLD will be denoted by ‘IS09-func’. For better comparability of classifier paradigms less dependent of the acoustic features used, the functionals listed in Table A.1 were also computed only from the MFCCs 1–12 along with energy; this feature set will be called ‘MFCC-func’. The IS09-func and MFCC-func feature sets consist of 384 and 312 features, respectively.

### 12.1.3 Performance

Ten-fold SCV is used for evaluation with partitioning by the Weka toolkit [22] with the default random seed of 0 for easy reproducibility. 10% of the data were used for evaluation, and 10% for validation whenever needed, e.g., for neural network training. HMMs were trained by the EM algorithm: Gaussian mixtures were consecutively added and re-estimated after six initial iterations until 16 Gaussian mixtures were reached for each state. For network training, supervised learning with early stopping and MVN was used. The network weights were initialised randomly from a Gaussian distribution ( $\mu = 0$ ,  $\sigma = 0.1$ ). Then, each training sequence was presented frame by frame. For improved generalisation ability, the order of the input sequences was determined randomly, and Gaussian noise ( $\mu = 0$ ,  $\sigma = 0.3$ ) was added to the input activations. The network weights were iteratively updated by resilient propagation. Further, the performance (in terms of WA) on the validation set was evaluated after each training epoch. Training was stopped in case of no improvement over 20 epochs or after a total of 100 training epochs. Then, the network with the best performance on the validation set was selected as the final network. SVMs were trained using SMO and a complexity constant of 0.1 on MVN processed features.

**Table 12.3** Results of the 2-class and 5-class tasks of the HU-ASA database with various classifiers and feature sets

Classifier	[%] Features	2-class		5-class	
		UA	WA	UA	WA
SVM	IS09-func	69.0	72.0	46.4	57.2
SVM	MFCC-func	73.9	75.6	42.2	56.0
Left-right HMM	MFCC	79.0	79.8	47.3	63.4
cyclic HMM	MFCC	79.0	79.6	49.5	64.0
LSTM	MFCC	80.0	81.3	41.1	62.3

The training set was up-sampled for each fold for the LSTM-RNN and SVM classifiers. This was done by copying training instances of minority classes until a near-uniform class distribution is achieved. This step was not necessary in the case of HMMs, as each class is learnt by an individual model, and classification is performed with HMMs and the maximum likelihood criterion, i.e., class priors, were not used in the decision rule. For classification with the LSTM RNN each sequence in the test set was presented frame by frame to the input layer, and each frame was assigned to the class with the highest probability as indicated by the output layer. Then, a majority vote over the frame-level decisions was made to label the sequence.

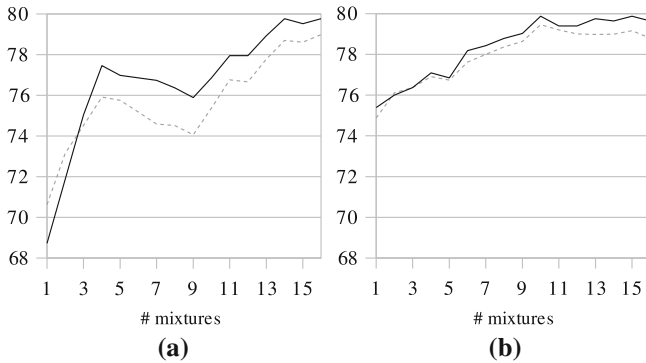
Table 12.3 depicts results by UA and WA for the 2-class and 5-class tasks of the HU-ASA database, as defined in Table 12.2. Always deciding for the majority class leads to WA and UA of 55.1 % and 20.0 % (5-class task), and 67.5 % and 50.0 % (2-class task).

In SVM classification on the 2-class task, the MFCC-func feature set outperforms the IS09-func set in terms of WA by 3.6 % absolute, being significant at the 5 % level (one-tailed  $z$ -test). However, the IS09-func feature leads to a significantly higher UA (4.4 % absolute improvement) for the 5-class task. Both types of HMMs outperform static classification by SVM. Further, the cyclic HMM is superior to the left-right HMM justifying the made assumption of partly quasi-periodic vocalisations. Yet, this observation is not significant on the 5 % level. To explain this, the estimated ‘cycle probability’  $a_{N,1}$  of the HMMs is shown for each class, on average across the ten folds, in Table 12.4. There, the cycle probabilities are around 28 % in the models for songbirds (*Passeriformes*) and primates, but below 10 % for *Felidae*.

The additional LLDs from Table A.1 as input features for the HMMs could not improve the above results. The impact of a varying number of Gaussian mixtures

**Table 12.4** Cycle probabilities  $a_{N,1}$  after training of the cyclic HMMs for comparison among each other given for each class in the 5-class task, averaged over ten folds

Class	$a_{N,1}$ [%]
Passeriformes	28.1
Non-Passeriformes	17.2
Canidae	14.2
Felidae	9.9
Primates	28.0



**Fig. 12.1** UA and WA on the HU-ASA database by 8-state HMMs with left-right and cyclic topologies, depending on the number of mixtures per state. Solid line: WA, dashed line: UA [10] **a** left-right HMM, 2-class task, **b** cyclic HMM, 2-class task

for the HMMs is selectively shown in Fig. 12.1 for the 2-class task. Interestingly, the cyclic HMM performs better than the left-right HMM for a small numbers of mixtures. Further, the UA on the 5-class task seems to be largely unaffected by the number of mixtures. This is surprising given that, ML classification partially compensates for the unequal class distribution. LSTM RNNs outperform—not significantly ( $p > 5\%$ )—the HMMs on the 2-class task. Yet, they have the lowest UA for the 5-class task. Additional variation of the network layout may change this behaviour. However, the lower performance for the 5-class is likely partly owing to the sparseness of the non-bird classes as LSTM RNN have a comparably high demand of training data.

### 12.1.4 Summary

In this section, an evaluation framework was shown for a challenging real-world database of animal vocalisations. The performances of static and dynamic classifiers, including LSTM networks, were compared. Dynamic classification provided higher accuracy. In the comparison of ‘standard’ MFCC features with an enhanced feature set containing pitch and voicing information no clear preference could be determined. Further evaluations in this direction are needed to reveal the relevance of different LLD and functional types for the classification of animal vocalisations.

From a classifier point of view, a hierarchical classification framework, e.g., by combining the songbird/non-songbird classifier with a bird song recogniser could be attempted.

## 12.2 Acoustic Events

In the next application of sound analysis, baseline results for the recognition of sound events are given. At the same time, this shall serve as an example of the usage of unlabelled data—sound event archives exist in masses on the Internet and can be exploited in semi-supervised learning even if no labels are given [11].

Recently, there is increasing interest in sound event classification in the field of acoustic signal analysis. This comes, apart from interest for application in multimedia search based on sound, as it is one of the key components to acoustically analyse environments, e.g., in surveillance [23, 24], monitoring of people in need of care, or detecting, and classifying sources of interest in real time [25]. There is also a benefit for humanoid and general robots [26] if they are able to better understand their acoustic environment. Finally, speech and music enhancement may be improved given a reliable identification of disturbing sound events. So far, most of research efforts in this direction base on rather prototypical and small databases with less than or around 1 000 instances (e.g., as in [24, 27–32]), or a few thousands of instances [26, 32, 33].

In this section, we will focus on sound events classification in a large scale database, covering sound classes that reach from nature (such as animals) over human beings (i.e., people) to artificial sounds (i.e., office, musical instruments, noise makers, and vehicles) as was introduced in Sect. 5.3.3.

Semi-supervised learning will be used to have the machine by itself label additional data instances as “there is no data like more data” and human labelling can easily become tedious and is expensive. Given a sufficiently robust automatic sound event classification system, unlabelled data can be classified and used in an iterative re-training process. Unlabelled sound data is practicably available in ‘infinite’ amounts: Recordings of real-life audio can be easily collected and typically contain various kinds and huge numbers of sound events [34]. Further, audio data can be added from the Internet. The semi-supervised adaptation of AMs and LMs in ASR [35, 36] and affective speech analysis [37] demonstrates that addition of unlabelled training data can lead to improvements in accuracy of classification systems. However, typically at least twice or sometimes up to around ten times as much unlabelled data is needed as compared to labelled data. Thus, AEC is shown in this book as an example for semi-supervised learning to improve a sound event classifier.

### 12.2.1 Methodology

openSMILE’s (cf. Sect. 6.5, [21]) ‘AVEC 2011’ set as shown in Table A.1 in the Annex is used for AEC. It consists of 1 941 features, composed of 25 energy and spectral related LLD x 42 functionals, 6 voicing related LLD x 32 functionals, 25 delta coefficients of the energy/spectral LLD x 23 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features.

As classifier, Random Forests as ensemble of decision trees are used. This choice is motivated by their good ability to cope with large feature spaces, as feature sub-spaces are randomly assigned to the trees in the forest. A good configuration proved to be 30 trees, and 150 randomly assigned features for each tree. For further reproducibility besides using an open-source feature extractor and the FindSounds database (cf. Sect. 5.3.3) that can be retrieved from the Internet, the classifier implementation provided by the Weka toolkit [22] is chosen again.

### 12.2.2 Performance

Considering the imbalance of instances among the classes, UA will be the measure of primary interest. Further, WA is partly provided in addition, as well as recall, precision, and  $F_1$ -measure. The experiments base on random partitioning of the FindSounds database into three stratified folds to provide two training and one completely disjoint testing set. The first fold (F1, 5 646 instances) is always used with its original manually assigned labels for training. The second fold (F2, 5 646 instances) is used either without its original labels (F2<sub>U</sub>) or with these labels (F2) to be able to compare to using this fold in a semi-supervised or supervised manner for training. The third and last fold (5 645 instances) is always used for testing. Random partitioning is carried out with Weka's default random seed.

Table 12.5 shows the occurred confusions for seven categories of sound event classification using the original labels training on fold one and two and testing on the third fold. This is the 'best case' given the entirely supervised learning with utmost data and serves as upper benchmark. Most confusions can be explained well by common sense, such as those of sounds from people with sounds of animals or sounds from vehicles with sounds of noise makers.

**Table 12.5** 'Best case' confusions when automatically classifying seven sound categories on the FindSounds database with original labels for both training folds F1 and F2 (cf. line 'supervised F1 + F2' in Table 12.6)

Truth [#]	Classified as						
	People	Animals	Nature	Vehicles	Noisemakers	Office	Instruments
People	564	153	11	26	17	25	50
Animals	126	717	7	35	23	20	18
Nature	18	35	157	42	44	10	6
Vehicles	37	37	26	476	86	15	45
Noisemakers	22	43	36	77	372	72	48
Office	29	37	1	16	111	364	31
Instruments	32	33	6	31	47	16	1 395
Confusions	264	338	87	227	328	158	198



**Table 12.6** Recall for seven sound categories and UA/WA with un-/supervised learning on the FindSounds database

	[%]	UA	WA	People	Animals	Nature	Vehicles	Noisemakers	Office	Instruments
supervised	F1	61.1	67.0	61.7	68.2	39.7	60.2	52.7	57.9	87.2
semi-supervised	2·F1 + F2 <sub>U</sub> <sup>1</sup>	63.1	68.5	61.7	72.5	47.4	61.8	51.9	58.4	87.9
supervised	F1 + F2	66.5	71.7	66.7	75.8	50.3	65.9	55.5	61.8	89.4

To establish a reference if the fold two data is not used at all, let us now consider exclusively fold one with its original labels for training (line “supervised F1” shown in Table 12.6) and fold three for testing. Then, for semi-supervised learning, fold one with the original manually assigned labels and fold two without the original labels, but labelled automatically by a system which was trained on fold one with application of diverse strategies (line “semi-supervised” in the same table). Testing is again carried out on fold three. Finally, the upper benchmark of using both folds with the original labels is shown in the table (line “supervised F1 + F2”)—again with fold three for testing.

For semi-supervised learning, the confidence of the Random Forests—the percentage of trees agreeing on the class—is taken into account. Evaluated confidence levels are  $> 0.7$  and  $> 0.8$ . This is needed to suppress data likely labelled wrong by the machine. Two additional strategies are investigated: up-sampling of the originally labelled data to emphasise more on definitely correctly labelled data and repeated iteration of the semi-supervised learning process. Table 12.7 shows the UA of up to three iterations of semi-supervised learning, i.e., repeated re-labelling of the unlabelled data in fold two using all fold one data and selected fold two data in training with labels from the last iteration, and only using instances with sufficient confidence level. Without up-sampling (1·F1), a gain is also obtained (62.0% vs. 61.6% UA for confidence level  $> 0.7$ , and 63.0% vs. 62.1% UA for confidence level  $> 0.8$ ). However, one notices that the benefit of iteration is limited, as UA partly begins to decrease after the third iteration. A larger number of iterations did not lead to improvements (not shown in numbers). Finally, the up-sampling and iterating strategies are

**Table 12.7** UA of iterative semi-supervised learning on the FindSounds database with minimum confidence values 0.7 and 0.8 combined with up-sampling or not up-sampling of originally labelled data

UA [%]	Confidence level			
	$>0.7$		$>0.8$	
	F1	2·F1	F1	2·F1
$F2_U^1$	61.6	63.1	62.1	62.5
$F2_U^2$	62.0	62.2	63.0	62.6
$F2_U^3$	62.0	61.7	62.6	63.2

2·F1: up-sampling (doubling up) fold 1 instances;  $F2_U^1$ ,  $F2_U^2$ ,  $F2_U^3$ : first, second, and third iteration of semi-supervised learning

combined expecting synergies. Looking at line “2·F1” in Table 12.6, up-sampling improves over the baseline setting in four out of six cases. Table 12.6 also shows detailed results for the case up-sampling by copying (2·F1) and confidences higher than 0.7.

Looking again at UA values in Table 12.7, as one would expect, the best average result is obtained using the original labels and data of fold one and fold two for training (66.5% UA). Then, semi-supervised learning significantly (one-sided z-test,  $p < 0.05$ ) boosts the performance of sound event classification by an increase in UA of 2% absolute over not using fold two data at all. This boost is almost half the one achieved by supervised training (5.4%) with all data over only using fold one. The nature class being the most sparse one, benefited most from semi-supervised learning. This effectively demonstrates the potential gain of semi-supervised learning for exploitation of unlabelled audio data.

### 12.2.3 Summary

The potential of semi-supervised learning on a large scale AEC task was investigated. In the result, adding unlabelled data with high classifier confidence level to the human-labelled training data can enhance recognition performance. Up-sampling of originally labelled data and iterating the semi-supervised learning process both boosted classification accuracy in the experiments by emphasising on originally labelled data. Combining both strategies gradually increases the advantage of semi-supervised learning. As one would expect, accuracy of semi-supervised learning is below the gain that can be expected when adding labelled data of the same amount. Yet, given the considerable efforts and costs involved in human labelling of thousands of instances and the large amounts of sound event data publicly available makes consideration of semi-supervised learning a promising approach in future machine-based sound analysis.

Future efforts could continue to focus on agglomeration of huge amounts of unlabelled sound event data and its application in analysis of real-life sound streams—ideally in combination with blind audio source separation.

## 12.3 Emotion

Similarly to the analysis of speech and music, where we first looked at ‘what’ was being said or played before looking at the affective side of speech and music, one can also attempt to automatically predict the emotion a sound event is likely to evoke in a listener. This will be the last application example presented in this book. It was first introduced in [12].

In fact, literature on emotion recognition from the acoustic channel—be it the emotion a listener thinks is contained or that she or he feels when listening—, is

dominated by studies dealing with speech [20, 38], and next follows music [39]. However, as shown in the last two sections, there is a rich variety of sounds besides speech and music in a real acoustic environment. These sounds certainly are also loaded with emotional connotation for a human listener. As an example, the shrill sound of a fire alarm would be less pleasant than the gentle sound of waves drilling the sand beach to the majority of listeners. In fact, listeners feed back emotion to any sound they are listening to in their daily life. This is independent of the kind of sound and its subjective or objective nature. Sound perception is thus linked with emotional response: New-borns' first attempts to overcome anxiety are centred on sound making [40]. Thus, for future intelligent systems it may be useful or relevant to understand emotion connotated with general sound. In 'sound information retrieval' emotional content may help in the design and dubbing of audio plays and films. For example, one might look for a furious door slam or a spooky door creek, etc. Research in this direction is utmost limited up to the present day: The only work besides the work by Schuller et al. is the very recent one presented in [41] basing on 120 clips of the BBC Sound Effects Library labelled in three affective dimensions. The approach uses mean and standard deviation per one second of 12 MFCC features as acoustic feature information. In this section, the focus is set on sound emotion recognition in realistic conditions.

A crucial problem is the lack of specialised sound databases for emotion research. There are some freely accessible sound databases [42], but usually without emotional labelling. The Emotional FindSounds database, which was described in Sect. 5.3.3 solves these issues. In emotion recognition from speech, emphasis is usually put on the subject's expressed emotion rather than listeners' emotions evoked by sound. This is more mixed for music emotion recognition. In fact, knowledge upon the emotion elicited on the listener side may help identify human reaction ahead. In this section, 'sound emotions' refer to the listeners' induced emotions.

### ***12.3.1 Methodology***

The audio feature set used is the openSMILE toolkit's 'AVEC 2011' set with 1941 features as shown in Table A.1 in the Annex and as was used in the last section for AEC. For recognition, random subspace meta-learning is used again owing to its good generalisation properties—the sounds are highly varied and require this feature. The base classifier is a decision tree. Based on experience, trees are not pruned. A subspace size of 0.05 is chosen, which means that 97 features out of the 1941 are assigned by random to each tree in the forest. The forest is grown from 500 trees [12]. The labelling and the feature extractor including the configuration are available for reproduction.<sup>3</sup> This principle was kept by again deciding for Weka for the implementation of the trees.

---

<sup>3</sup> Available at <http://www.openaudio.eu>

**Table 12.8** Automatic regression results by CC with different types of gold standard

CC		# trees		
		100	200	500
Arousal	EWE	0.611	0.608	0.606
	median	0.553	0.555	0.548
	mean	0.558	0.563	0.559
Valence	EWE	0.458	0.469	0.473
	median	0.446	0.449	0.454
	mean	0.467	0.484	0.485

EWE, median, and mean in ten-fold SCV. The number of trees is varied

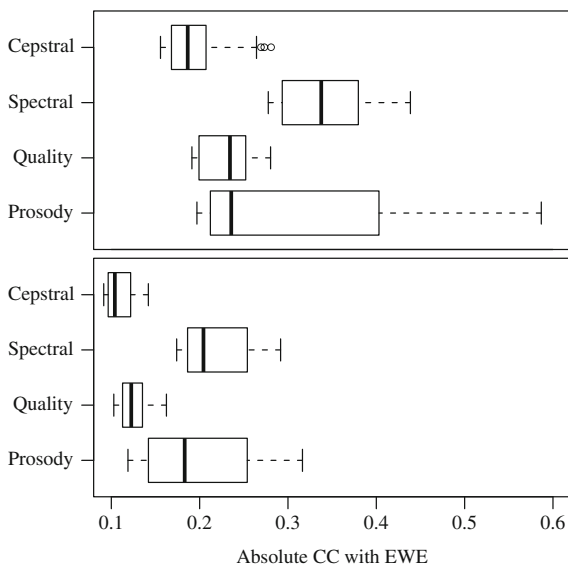
### 12.3.2 Performance

A ten-fold SCV—again with reproducible partitioning by Weka’s default random seed—is carried out on the emotionally tagged partition of the FindSounds database as introduced in Sect. 5.3.3. Table 12.8 shows the CCs for arousal and valence employing the Evaluator Weighted Estimator (EWE), median, and mean to establish a gold standard by merging the evaluation results of the four evaluators. In this table, numbers of trees in the forest are additionally varied. Visibly, the regression of sound emotion performs well with CCs of around 0.61 (arousal) and up to 0.49 (valence) when evaluating on the EWE. The tendency that arousal is the ‘easier’ task is well in line with experience from speech and music emotion analysis based on acoustics [20, 43]. CC as evaluated on EWE usually exceeds the other two methods of gold standard establishment—mean and median. Median is found on the other end of the scale probably due to its instability when evaluators show huge disagreement. In Table 12.9 the CC and its relation to sound category is highlighted for one exemplary configuration. There, arousal prediction is roughly stable across sound categories. As for valence, especially *Noisemakers* and *Nature* can be identified well above others in

**Table 12.9** Automatic regression results by CC per sound category for EWE and 500 trees in ten-fold SCV

Class	CC	
	Arousal	Valence
	0.601	0.474
Animals	0.643	0.448
Musical instruments	0.516	0.217
Nature	0.688	0.589
Noisemaker	0.579	0.778
People	0.604	0.048
Sports	0.682	0.198
Tools	0.590	−0.057
Vehicles	0.579	0.279

**Fig. 12.2** Boxplots of the 30 highest absolute CCs of features with the EWE. Features are grouped in four cover classes for arousal (top) and valence (bottom). The ‘Quality’ group contains voicing probability, log HNR, jitter, and shimmer based features. ‘Prosody’ groups loudness, F0, and ZCR [12]



terms of CC. In comparison with the gold standard as was shown in Fig. 5.8, one may argue that the regressor is not only implicitly recognising the sound category. In fact, the values of valence for *Noisemakers* are rather widespread despite considerable differences in the mean valence.

As there exists practically no experience on feature relevance for this particular task, it seems worth to have a look at this issue. The 30 best features were ranked by their CC with the EWE as gold standard. The result is shown as boxplots per dimension for the groups cepstral, spectral, ‘sound quality’ in analogy to voice quality, and prosody in Fig. 12.2. Independent of arousal or valence, spectral features are the most relevant group. Interestingly, the best single feature is prosody-related for these two dimensions. From the full list of the 30 best features (not shown) the following is found: Arousal is highly correlated with loudness, and loudness features almost reach the CC with the EWE of the learnt regressor. The highest CC is observed for the root quadratic mean of loudness (0.587).

Next, valence is correlated with loudness as well, but not as strongly and negatively, which seems intuitive, as loud sounds are likely unpleasant. The highest absolute CC with the EWE can be reported for the third quartile of loudness ( $-0.316$ ). Spectral flux also shows good (negative) CC, i.e., large spectral variations seem to be perceived as unpleasant: The CC of the inter quartile range 1–2 of spectral flux is  $-0.292$ . Finally, spectral harmonicity is negatively correlated: Apparently quasi-sinusoidal sounds are unpleasant. The CC of 50% up-level time of harmonicity is  $-0.241$ .

### 12.3.3 Summary

Automatic recognition of emotion evoked by general sound events was shown and found in the rough range of typical dimensional speech and music emotion recognition when operating in high realism comparable to the results in Sect. 10.4.2 and Sect. 11.7. The sound events considered here were completely independent of each other and often of lower acoustic quality. Spectral features were shown to be most important as a group after individual prosodic features for this task.

Future efforts may aim at creation of larger sound emotion resources, e.g., by crowd sourcing or similar. Deeper analysis of feature relevance per sound category will also shed more light on optimal acoustic feature spaces. Finally, multi-task learning of the sound category and the evoked emotion seems a promising approach to improve both tasks as was suggested in speech and music processing before.

## References

1. Temko, A., Nadeu, C., Macho, D., Malkin, R., Zieger, C., Omologo, M.: Acoustic event detection and classification. In: Waibel, A., Stiefelwagen, R. (eds.) *Computers in the Human Interaction Loop*, pp. 61–73. Springer, London (2009)
2. Wang, D., Brown, G.: *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press (2006)
3. Huang, Q., Cox, S.: Using high-level information to detect key audio events in a tennis game. In: *Proceedings INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pp. 1409–1412. Makuhari, Japan, Sept 2010. ISCA
4. Xu, M., Chia, L., Jin, J.: Affective content analysis in comedy and horror videos by audio emotional event detection. In: *Proceedings 6th IEEE International Conference on Multimedia and Expo, ICME 2005*, p. 4. Amsterdam, The Netherlands, IEEE, July 2005
5. Okuno, H., Ogata, T., Komatani, K., Nakadai, K.: Computational auditory scene analysis and its application to robot audition. In: *Proceedings of the International Conference on Informatics Research for Development of Knowledge Society Infrastructure*, pp. 73–80. IEEE (2004)
6. Gunasekaran, S., Revathy, K.: Content-based classification and retrieval of wild animal sounds using feature selection algorithm. In: *Proceedings of International Conference on Machine Learning and Computing (ICMLC)*, pp. 272–275. IEEE Computer Society, Bangalore, India, Feb 2010
7. Wan, C., Mita, A.: An automatic pipeline monitoring system based on PCA and SVM. *World Acad. Sci. Eng. Technol.* **45**, 90–96 (2008)
8. Bach, J., Anemuller, J.: 11th Annual Conference of the International Speech Communication Association, pp. 2206–2209. ISCA, Makuhari, Japan, Sept 2010
9. Geiger, J.T., Lakhall, M.A., Schuller, B., Rigoll, G.: Learning new acoustic events in an hmm-based system using map adaptation. In: *Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*, pp. 293–296. ISCA, Florence, Italy, Aug 2011
10. Weninger, F., Schuller, B.: Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. In: *Proceedings of 36th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*, pp. 337–340. IEEE, Prague, Czech Republic, May 2011

11. Zhang, Z., Schuller, B.: Semi-supervised learning helps in sound event classification. In: Proceedings of 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, pp. 333–336. IEEE, Kyoto, Japan, March 2012
12. Schuller, B., Hantke, S., Weninger, F., Han, W., Zhang, Z., Narayanan, S.: Automatic recognition of emotion evoked by general sound events. In: Proceedings of 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, pp. 341–344. IEEE, Kyoto, Japan, March 2012
13. Mellinger, D.K., Clark, C.W.: Recognizing transient low-frequency whale sounds by spectrogram correlation. *J. Acoust. Soc. Am.* **107**(6), 3518–3529 (2000)
14. Härmä, A.: Automatic recognition of bird species based on sinusoidal modeling of syllables. In: Proceedings of ICASSP, vol. 5, pp. 545–548. Hong Kong, April 2003
15. Bardeli, R.: Similarity search in animal sound databases. *IEEE Trans. Multimedia* **11**(1), 68–76 (2009)
16. Frommolt, K.-H., Bardeli, R., Kurth, F., Clausen, M.: The animal sound archive at the Humboldt-University of Berlin: current activities in conservation and improving access for bioacoustic research. *Adv. Bioacoustics* **2**, 139–144 (2006)
17. Guo, G., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. *IEEE Trans. Neural Networks* **14**(1), 209–215 (2003)
18. Mitrovic, D., Zeppelzauer, M., Breiteneder, C.: Discrimination and retrieval of animal sounds. In: Proceedings of Multi-Media Modelling Conference, IEEE, Beijing, China, Jan 2006
19. Kim, H.-G., Burred, J.J., Sikora, T.: How efficient is MPEG-7 for general sound recognition? In: Proceedings of AES 25th International Conference, London, UK, June 2004
20. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* **53**(9/10):1062–1087 (2011) (Special Issue Sensing Emotion and Affect-Facing Realism in Speech Processing)
21. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile—the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 9th ACM International Conference on Multimedia, MM 2010, pp. 1459–1462. ACM, Florence, Italy, October 2010
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. *SIGKDD Explor.* **11**(1), 10–18 (2009)
23. Temko, A., Malkin, R., Zieger, C., Macho, D., Nadeu, C.: Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems. In: Proceedings of the IV Biennial Workshop on Speech Technology, pp. 1–6. Zaragoza, Spain (2006)
24. Clavel, C., Ehrette, T., Richard, G.: Events detection for an audio-based surveillance system. In: Proceedings of ICME, pp. 1306–1309. Amsterdam (2005)
25. Ferguson, B.G., Lo, K.W.: Acoustic cueing for surveillance and security applications. In: Proceedings of SPIE, Orlando, FL, USA (2006)
26. Kraft, F., Malkin, R., Schaaf, T., Waibel, A.: Temporal ICA for classification of acoustic events in a kitchen environment. In: Proceedings of INTERSPEECH, pp. 2689–2692. Lisbon, Portugal (2005)
27. Temko, A., Nadeu, C.: Classification of acoustic events using SVM-based clustering schemes. *Pattern Recogn.* **39**, 682–694 (2006)
28. Zieger, C., Omologo, M.: Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm. In: Proceedings of INTERSPEECH, pp. 115–118. Brisbane, Australia (2008)
29. Heittola, T., Klapuri, A.: TUT acoustic event detection system 2007. In: Proceedings of Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, pp. 364–370. Springer, Berlin, Heidelberg (2008)
30. Ntalampiras, S., Potamitis, I., Fakotakis, N.: Automatic recognition of urban environmental sound events. In: Proceedings of CIP2008, Eurasip, pp. 110–113 (2008)
31. Peng, Y., Lin, C., Sun, M., Tsai, K.: Healthcare audio event classification using hidden markov models and hierarchical hidden markov models. In: Proceedings of ICME, pp. 1218–1221. Piscataway, NJ, USA (2009)

32. Dat, T.H., Li, H.: Probabilistic distance svm with hellinger-exponential kernel for sound event classification. In: Proceedings of ICASSP, pp. 2272–2275. Prague, Czech Republic (2011)
33. Chu, S., Narayanan, S., Kuo, C.-C.J.: Environmental sound recognition with time-frequency audio features. *Trans. Audio Speech Lang. Process.* **17**(6), 1142–1158 (2009)
34. Mesaros, A., Heittola, T., Eronen, A., Virtanen, T.: Acoustic event detection in real life recordings. In: Proceedings of EUSIPCO, Aalborg, Denmark (2010)
35. Hakkani-Tur, D., Tur, G., Rahim, M., Riccardi, G.: Unsupervised and active learning in automatic speech recognition for call classification. In: Proceedings of ICASSP, pp. 429–432. Montreal, Canada, (2004)
36. Tur, G., Stolcke, A.: Unsupervised language model adaptation for meeting recognition. In: Proceedings of ICASSP, pp.173–176. Honolulu, Hawaii, USA (2007)
37. Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B.: Unsupervised learning in cross-corpus acoustic emotion recognition. In: Proceedings of 12th Biannual IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2011, pp. 523–528. IEEE, Big Island, HI, Dec 2011
38. Gunes, H., Schuller, B., Pantic, M., Cowie, R.: Emotion representation, analysis and synthesis in continuous space: a survey. In: Proceedings of the International Workshop on Emotion Synthesis, representation, and Analysis in Continuous space, EmoSPACE 2011, held in Conjunction with the 9th IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, FG 2011, pp. 827–834. IEEE, Santa Barbara, CA, March 2011
39. Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D.: Music emotion recognition: a state of the art review. In: Proceedings of ISMIR, pp. 255–266. Utrecht, The Netherlands (2010)
40. Forrester, M.: Auditory perception and sound as event: theorising sound imagery in psychology. *J. Sound*, <http://www.kent.ac.uk/arts/sound-journal/forrester001.html> (2000)
41. Sundaram, S., Schleicher, R.: Towards evaluation of example-based audio retrieval system using affective dimensions. In: Proceedings of ICME, pp. 573–577. Singapore, Singapore (2010)
42. Gygi, B., Shafiro, V.: Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations. *EURASIP J. Audio Speech Music Process.* pp. 12 (2010). Article ID: 654914
43. Schuller, B., Dorfner, J., Rigoll, G.: Determination of non-prototypical valence and arousal in popular music: Features and performances. *EURASIP J. Audio Speech Music Process.* (Special Issue on Scalable Audio-Content Analysis, 2010) pp. 19 (2010). (Article ID 735854)