AUDIO RECOGNITION IN THE WILD: STATIC AND DYNAMIC CLASSIFICATION ON A REAL-WORLD DATABASE OF ANIMAL VOCALIZATIONS

Felix Weninger and Björn Schuller

Institute for Human-Machine Communication, Technische Universität München
80290 München, Germany
weninger@tum.de

ABSTRACT

We present a study on purely data-based recognition of animal sounds, performing evaluation on a real-world database obtained from the Humboldt-University Animal Sound Archive. As we avoid a preselection of friendly cases, the challenge for the classifiers is to discriminate between species regardless of the age or stance of the animal. We define classification tasks that can be useful for information retrieval and indexing, facilitating categorization of large sound archives. On these tasks, we compare dynamic and static classification by left-right and cyclic Hidden Markov Models, recurrent neural networks with Long Short-Term Memory, and Support Vector Machines, as well as different features commonly found in sound classification and speech recognition, achieving up to 81.3 % accuracy on a 2-class, and 64.0 % on a 5-class task.

Index Terms— Sound Classification, Bioacoustics, Audio Pattern Recognition

1. INTRODUCTION

In the field of bioacoustics, a multiplicity of approaches exist for classifying animal sounds. Often they are used in order to monitor populations of certain species, such as whales [1] or birds [2], thereby suiting the algorithms to the special characteristics of the animal vocalizations involved. However, more recently, with increasing efforts invested in digitization of sound archives, increasing attention is being paid to general frameworks for audio classification that can be useful in indexing and search procedures. For example, in [3], an effective indexing algorithm for animals with curve-like harmonic vocalizations, 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 [4], which will be subsequently referred to as 'HU-ASA database'.

In this paper, we do not directly aim at the domain of information retrieval, but our study rather relates to previous work done on sound classification. On the other hand, sound classification, especially into coarse categories that can be robustly discriminated, can be of use in a preselection step for those approaches tailored to certain classes of sounds, and it can facilitate the work of specialists working on categorization of sound databases. In the past, Support Vector Machine (SVM)-based static classification of audio files, using segment-wise functionals (e. g., mean and standard deviation) was proposed in [5], and used for animal sounds in [6]. Other approaches frequently employ a dynamic classification process, e. g., by Hidden Markov Models (HMMs) [7] or, in the animal vocalization domain, by neural networks [8]. However, to our knowledge, there does not exist a comparative study on the performance of static and dynamic classification of animal vocalizations. Hence, a major contribution of our study is to evaluate SVMs, HMMs with different topologies, and recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) on the HU-ASA database. Additionally, we will compare traditional cepstral features which are commonly used in sound classification to an enhanced feature set derived from speech emotion recognition, which arguably is also an instance of sound classification.

The remainder of this paper is structured as follows: first, we describe in detail the evaluation framework for sound classifcation that we have derived from the HU-ASA database, in Sec. 2. Then, we describe our classification and acoustic feature extraction methods in Sec. 3 before presenting our experimental results in Sec. 4 and concluding with an outlook in Sec. 5.

2. EVALUATION DATABASE

Our evaluation database builds on the HU-ASA database, which is a large archive of animal vocalizations annotated with the species and additional metadata, including recording conditions and type of vocalization for each audio file. We obtained all the 1 418 audio files available in MP3 encoding from the archive[1] as of mid 2010. For each species, we automatically annotated the audio files with (biological) class (e. g., *Aves*, *Mammalia*), order (e. g., *Passeriformes*, *Primates*), and family (e. g., *Felidae*, *Canidae*) according to the Linnaean taxonomy (rank-based biological classification), retrieved for each species from Wikipedia. The majority of the available recordings consist of birds (*Aves*) and mammals (*Mammalia*), as shown in Tab. 1. 'Others' include *Sauropsida*, *Hexapoda*, and recordings that could not be annotated automatically due to missing information in the encyclopedia. The total recording length of the files was 20 423 s (5 h 40 min 23 s). *Amphibia*, *Insecta*, and *Reptilia* were not included in our further experiments due to a small number of instances (see Tab. 1).

From the biological classification, we derived two tasks that we found suitable for audio pattern recognition, considering the number of available instances and the 'discriminability' (in an informal sense) of the vocalizations. The tasks are shown in Tab. 2. The first (2-class) task is to discriminate between songbirds (*Passeriformes*) and other birds (*Non-Passeriformes*), the latter including – sorted by number of instances – the orders *Anseriformes*, *Charadriiformes*, *Galliformes*, *Psitacciformes*, *Gruiformes*, and 24 other orders with often a very

This work was supported by the Federal Republic of Germany through the German Research Foundation (DFG) under the grant no. SCHU 2508/2-1 ("Non-Negative Matrix Factorization for Robust Feature Extraction in Speech Processing").

[1] http://www.tierstimmenarchiv.de/

(biol.) class	# inst.	min	mean	max	Σ
Aves	868	2.4 s	14.8 s	64.7 s	12 210 s
Mammalia	487	1.0 s	14.7 s	37.7 s	6 954 s
Amphibia	27	1.8 s	19.6 s	65.9 s	529 s
Reptilia	7	11.2 s	22.5 s	39.6 s	157 s
Insecta	19	2.3 s	16.0 s	30.1 s	287 s
Other	10				133 s
Σ	1 418				20 423 s

Table 1: Number of instances, as well as min(imum), mean, max(imum), and total recording length (Σ) of the audio files, by the biological class of the species in the HU-ASA database.

class	# inst.
Passeriformes	282
Non-Passeriformes	586
Σ	868
Primates	90
Canidae	43
Felidae	62
Σ	1 063

Table 2: Number of instances in the 2-class (*Passeriformes* / *Non-Passeriformes*) and 5-class tasks defined on the HU-ASA database.

low number of instances. Furthermore, to define an arguably more complex task, we added the mammals (*Mammalia*) of the families *Felidae* and *Canidae*, as well as the instances of the biological order *Primates* to the 2-class task, resulting in a 5-class problem as shown in Tab. 2. Both of these tasks are challenging due to the real-world nature of the database. In particular, instances of one class comprise different types of vocalizations of the same species, depending on the situation and stance (i. e., aggression or warning calls), as well as animals of different age, from young to full-grown.

3. METHODOLOGY

3.1. Classifiers

We evaluated static classification by SVMs with polynomial kernel, as well as dynamic classifiers, including two different topologies of HMMs, as well as a LSTM-RNN. An HMM topology commonly found in sound classification is a left-right HMM: assuming N states in total, state transitions are allowed from state $i = 1, \ldots, N-1$ to states i and $i+1$, following a strictly linear topology. This topology appears to be naturally suited to phoneme recognition in human speech, modelling transitions from one phoneme to the other, and repetition of acoustic patterns according to the speech frequency. However, it can be argued that in contrast to human speech, animal vocalizations are highly repetitive, motivating the usage of a *cyclic* topology, where in addition to the transitions in the left-right HMM, an additional transition from state N to the first state 1 is allowed. In our experiments we fixed $N = 8$.

An alternative architecture for dynamic sound classification is built on recurrent neural networks (RNNs). For instance, in [8], a feedforward multilayer perceptron was proposed for classifying animal vocalizations. In contrast to basic feedforward neural networks, recurrent connections from the output to the input provide a RNN with a kind of memory, which may influence the network output in the future. Although RNNs have access to past (and future) information, the range of context is limited to a few frames due to the vanishing gradient problem: the influence of an input value decays or blows up exponentially over time. This problem is circumvented by extending the nonlinear units to LSTM memory blocks, containing linear memory units, whose internal state is maintained by a recurrent connection with constant weight 1.0, and multiplicative gate units to control input, output, and internal state. Hence, during training, the network automatically learns when to store, use, or discard information acquired from previous inputs or outputs. This makes LSTM-RNNs useful for the task considered in this study, as the required amount of context is unknown a priori and would otherwise have to be determined experimentally for each of the classes to discriminate. In various signal processing applications, including emotion recognition [9] and note onset detection [10], they have been proven useful for purely data-based discriminative learning of sequence labeling tasks, thereby often outperforming more traditional sequence classifiers such as HMMs.

Hence, we additionally took into account LSTM networks for classification, which had one hidden layer with 100 LSTM memory cells. The size of the input layer was equal to the number of features, while the size of the output layer was equal to the number of classes to discriminate. Its output activations were restricted to the interval $[0; 1]$ and their sum was forced to unity by normalizing with the softmax function. Thus, the normalized outputs represent the posterior class probabilities.

3.2. Acoustic Features

As a basic feature set we extracted the Mel frequency cepstral coefficients (MFCCs) 1–12 along with energy and their first (δ) and second order ($\delta\delta$) regression coefficients. MFCCs are commonly found in speech processing applications, but have also been successfully used in various audio classification tasks [5–7]. They appear particularly suited to a *general* framework for audio classification as they capture a broadband frequency range, and integrate a perceptual weighting of individual frequency bands as performed by the human ear. Furthermore, in [7] they were found superior to the MPEG-7 spectral projection features as used in [3] for sound classification using HMMs. Note that the regression coefficients allow to integrate past and future context information. The resulting 39-dimensional feature set will be denoted by 'MFCC'.

For static classification of entire signals, it is necessary to reduce the time-varying frame-wise acoustic features to functionals. In [5], mean and standard deviation were proposed; however, especially applications in the paralinguistic domain, tend to employ a broader range of functionals, including extremes and higher-order moments [11]. Furthermore, in this field, often additional features including zero-crossing rate (ZCR), fundamental frequency (F0) and harmonics-to-noise ratio (HNR), are used. We investigated the usefulness of the feature set used for the INTERSPEECH 2009 Emotion Challenge [11], as described in Tab. 3, since it can be argued that emotion recognition is an instance of sound classification, and especially that the aforementioned features could allow to discriminate between animals with voiced and unvoiced sounds. We will denote the functionals of the the 32 low-level descriptors (LLD) by 'IS09-func'. To be able to separately evaluate classifier and feature performance, we also investigated the functionals listed in Tab. 3 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. For best transparency of results, all feature sets were extracted with our open-source feature extractor openSMILE [12].

LLD	Functionals
(δ) ZCR	mean
(δ) RMS Energy	standard deviation
(δ) F0	kurtosis, skewness
(δ) HNR	extremes: value, rel. position, range
(δ) MFCC 1–12	linear regression: offset, slope, MSE

Table 3: INTERSPEECH 2009 Emotion Challenge feature set (IS09-func): low-level descriptors (LLD) and functionals.

4. EXPERIMENTS

Classifiers were evaluated using stratified 10-fold cross validation, creating the folds with the Weka toolkit [13], using the default random seed of 0 for easy reproducibility. In each iteration, 10 % of the data were used for evaluation, and another 10 % for validation whenever needed, e. g., for neural network training.

4.1. Classifier Training

HMMs were trained using the common Expectation-Maximization (EM) algorithm. After six initial iterations, additional Gaussian mixtures were added consecutively and re-estimated during four EM iterations, until the final models had 16 Gaussian mixtures for each state. For network training, supervised learning with early stopping was used as follows: we initialized the network weights randomly from a Gaussian distribution ($\mu = 0, \sigma = 0.1$). Then, each sequence in the training set of each fold was presented frame by frame to the network. To improve generalization, 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 using resilient propagation. To prevent over-fitting, the performance (in terms of classification error) on the validation set was evaluated after each training iteration (epoch). Once no improvement over 20 epochs had been observed, or 100 training epochs had elapsed, the training was stopped and the network with the best performance on the validation set was used as the final network. SVMs were trained using Sequential Minimal Optimization (SMO) on standardized features (zero mean and unit variance), using a complexity constant of 0.1. All other parameters correspond to the default in the Weka toolkit for easy reproducibility.

As we are generally dealing with highly unbalanced classification tasks, the training set for each fold was upsampled for both the LSTM-RNN and SVM classifiers, i. e., all training instances of each minority class were copied until a near-uniform class distribution on the training set was achieved. Note that balancing has no effect on the training of the HMMs by the EM algorithm, as each class is modeled by an individual HMM.

4.2. Results

Classification with HMMs was done by assigning the class corresponding to the model with the maximum likelihood (ML), which is particularly suitable to unbalanced classification tasks, as the a-priori class probabilities do not affect the decision. Classification with the LSTM-RNN was performed as follows: each sequence in the test set was presented frame by frame (in correct temporal order) to the input layer, and each frame was assigned to the class with the highest probability as indicated by the output layer. From the frame-wise decisions, the majority vote was taken as label for the sequence.

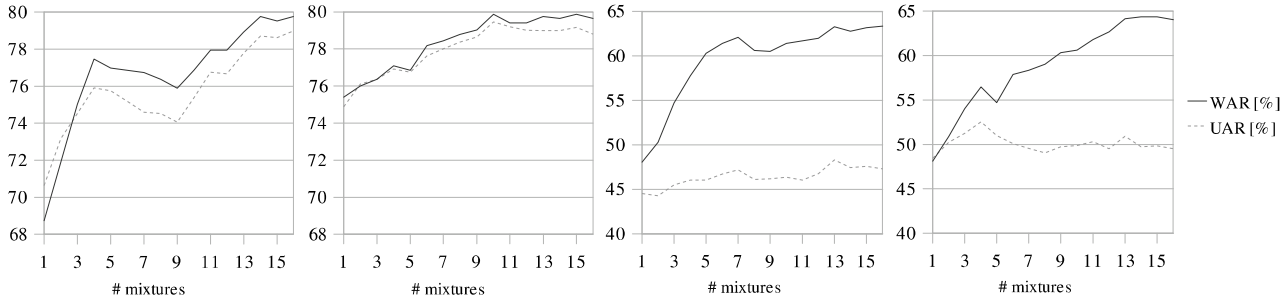
In Tab. 4, we show the unweighted (UAR) and weighted average recall (WAR) on the 2-class and 5-class tasks of the HU-ASA database, as defined in Tab. 2. We adopt UAR as an additional evaluation measure due to the unbalanced data sets. Note that by always deciding for the majority class (*Non-Passeriformes*), a WAR of 55.1 % and a UAR of 20.0 % are obtained on the 5-class task, and a WAR / UAR of 67.5 % / 50.0 % on the 2-class task. In SVM classification on the 2-class task, the MFCC-func feature set outperforms the IS09-func set in terms of WAR by 3.6 % absolute, and this improvement is even significant at the 5 % level, according to a one-tailed z-test. However, the IS09-func feature set seems to deliver significantly higher UAR in the 5-class task (4.4 % absolute improvement).

	[%]	2-class		5-class	
Classifier	**Features**	**UAR**	**WAR**	**UAR**	**WAR**
SVM	IS09-func	69.0	72.0	46.4	57.2
SVM	MFCC-func	73.9	75.6	42.2	56.0
LR-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

Table 4: Results on the 2-class and 5-class tasks of the HU-ASA database, using various classifiers (Sec. 3.1) and feature sets (Sec. 3.2). LR-HMM is a left-right HMM. UAR and WAR denote (un)weighted average recall. 16 Gaussian mixtures per state were estimated for the HMMs.

Comparing to dynamic classification by HMMs, it can be seen that both types of HMMs outperform static classification by SVM, and that the cyclic HMM is in turn slightly superior to the left-right HMM. Yet, the latter observation fails to be significant on the 5 % level. The fact that there is no significant difference in the performance of cyclic and left-right HMM may be explained by examining the estimated transition probabilities of the HMM, in particular the 'cycle probability' $p_{N,1}$, which are shown for each class, on average across the 10 folds, in Tab. 5. These probabilities are generally quite low, indicating that the cyclic connection in the HMM is of lower importance. However, it is notable that the cycle probabilities considerably differ: while they are around 28 % in the models for songbirds (*Passeriformes*) and primates, the probability is below 10 % for *Felidae*. While we also investigated the LLDs from Tab. 3 as features for the HMMs, these could not improve the results. Concluding the discussion of HMMs, the impact of using different numbers of Gaussian mixtures for the HMMs is shown in Fig. 1. Interestingly, the cyclic HMM performs better than the left-right HMM for small numbers of mixtures, and the UAR on the 5-class task seems to be largely unaffected by the number of mixtures, despite the fact that ML classification partially compensates for the unequal class distribution.

Finally, concerning the performance of the LSTM-RNN, there is no clear picture: while it outperforms (yet not significantly, $p > 5\,\%$) the HMMs on the 2-class task both in terms of WAR and UAR, it shows the lowest performance among the classifiers concerning the UAR on the 5-class task, while yielding a significantly ($p < 0.1\,\%$) higher WAR in comparison with the best SVM, and a slightly (not significantly) inferior WAR in comparison to both types of HMMs. Naturally, additional investigations as to the network topology and training parameters would be needed for a thorough evaluation of the LSTM-RNN performance; still, we believe that the observed difference between the 2-class and 5-class tasks can be attributed to insufficient generalization due to the relatively little amount of training data for the *Primates*, *Canidae*, and *Felidae* classes – note that upsampling does not help generalization. Thus, the decision of the LSTM-RNN in the 5-class task remains strongly biased towards the majority class, which results in low UAR.



(a) left-right HMM, 2-class task (b) cyclic HMM, 2-class task (c) left-right HMM, 5-class task (d) cyclic HMM, 5-class task

Fig. 1: Recognition performance in terms of (un)weighted average recall (UAR / WAR) on the HU-ASA database by 8-state HMMs with left-right ((a), (c)) and cyclic ((b), (d)) topologies, depending on the number of mixtures per state.

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

Table 5: Estimated cycle probabilities $p_{N,1}$ for the cyclic HMMs, for each class in the 5-class task, averaged over 10 folds.

5. CONCLUSION

We have proposed an evaluation framework for sound classification based on a challenging real-world database of animal vocalizations, and compared the performances of static and dynamic classifiers, including a novel type of recurrent neural network. Overall, dynamic classification delivered higher accuracy. Notably, no clear picture could be established in the comparison of standard cepstral features with an enhanced feature set containing pitch information – thus, an interesting area for further research will be to further evaluate the relevance of different feature and functional types for the classification of animal vocalizations. Furthermore, we will introduce data-based feature extraction by Non-Negative Matrix Factorization (NMF) to the domain of animal sound classification, but using global optimization constraints instead of simple projections, as done for the MPEG-7 spectral projection features in [7]. Finally, we will evaluate the presented classification systems in a hierarchical classification framework, e. g., by combining the songbird / non-songbird classifier with a bird song recognizer.

6. REFERENCES

[1] D. K. Mellinger and C. W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518–3529, 2000.

[2] A. Härmä, "Automatic recognition of bird species based on sinusoidal modeling of syllables," in *Proc. of ICASSP*, Hong Kong, April 2003, vol. 5, pp. 545–548.

[3] R. Bardeli, "Similarity search in animal sound databases," *IEEE Transactions on Multimedia*, vol. 11, no. 1, pp. 68–76, 2009.

[4] K.-H. Frommolt, R. Bardeli, F. Kurth, and M. Clausen, "The animal sound archive at the Humboldt-University of Berlin: Current activities in conservation and improving access for bioacoustic research," *Advances in Bioacoustics*, vol. 2, pp. 139–144, 2006.

[5] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 209–215, January 2003.

[6] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder, "Discrimination and retrieval of animal sounds," in *Proc. of Multi-Media Modelling Conference*, Beijing, China, January 2006, IEEE.

[7] H.-G. Kim, J. J. Burred, and T. Sikora, "How efficient is MPEG-7 for general sound recognition?," in *Proc. of AES 25th International Conference*, London, UK, June 2004.

[8] S. Gunasekaran and K. Revathy, "Content-based classification and retrieval of wild animal sounds using feature selection algorithm," in *Proc. of International Conference on Machine Learning and Computing (ICMLC)*, Bangalore, India, February 2010, pp. 272–275, IEEE Computer Society.

[9] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing (JSTSP), Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, vol. 4, no. 5, pp. 867–881, October 2010.

[10] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long-short term memory neural networks," in *Proc. of ISMIR*, Utrecht, Netherlands, August 2010.

[11] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication, Special Issue on "Sensing Emotion and Affect – Facing Realism in Speech Processing"*, 2011, to appear.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Florence, Italy, October 2010, ACM.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.