Improving Bird Identification using Multiresolution Template Matching and Feature Selection during Training

Mario Lasseck

Animal Sound Archive
Museum für Naturkunde Berlin
Mario.Lasseck@mfn-berlin.de

Abstract. This working note describes methods to automatically identify a large number of different bird species by their songs and calls. It focuses primarily on new techniques introduced for this year's task like advanced spectrogram segmentation and decision tree based feature selection during training. Considering the identification of dominant species, previous results of the LifeCLEF Bird Identification Task could be further improved by 29%, achieving a mean Average Precision of 59% (mAP). The proposed approach ranked second place among all participating teams and provided the best system to identify birds in soundscape recordings.

Keywords: Bird Identification · Multimedia Information Retrieval · Spectrogram Segmentation · Multiresolution Template Matching · Feature Selection

1 Introduction

Automated acoustic methods of species identification can serve as a useful tool for biodiversity assessments. Within the scope of the LifeCLEF 2016 Bird Identification Task researchers are challenged to identify 999 different species in a large and highly diverse set of audio files. The audio recordings forming the training and test data set are built from the Xeno-canto collaborative database (www.xeno-canto.org). A novelty in this year's challenge is the enrichment of the test data set by including a new set of soundscape recordings. These soundscapes are not targeting any specific species during recording and can contain an arbitrary number of singing birds. To establish reliable acoustic methods for assessing biodiversity it is essential to improve the automated identification of birds in general but especially within these soundscape recordings. An overview and further details about the Bird Identification Task are given in [1]. The task is among others part of the LifeCLEF 2016 evaluation campaign [2]. Some methods referred to in the following sections are further developments of approaches already successfully applied in previous identification tasks. A more detailed description of these approaches can be found in [3,4,5].

2 Feature Engineering

Two main categories of features (the same as last year) were used for training and prediction: matching probabilities of species-specific 2D spectrogram segments (see 2.1 Segment-Probabilities) and acoustic features extracted with openSMILE (see 2.2 Parametric Acoustic Features). For this year's task a large number of new Segment-Probability features were added for training by extracting new sound segments from audio files using the following two methods.

Re-segmentation of large segments. Using the automated segmentation method of spectrograms described in [3] some of the extracted segments turned out to be quite large and in some cases not very useful for template matching – especially when processing audio files with a lot of background noise, low signal to noise ratio or many overlapping sounds. To overcome this problem all segments having a duration longer than half a second or a frequency range greater than 6 kHz were treated as separate spectrogram images and re-segmented again with a slightly different image preprocessing technique. The preprocessing steps for these too large segments differed in the following ways from the original preprocessing of the spectrogram image:

- transform spectrogram into a binary image via *Median Clipping* by setting each pixel to 1, if it is 3.5 (instead of 3) times the median of its corresponding row AND column, otherwise to 0
- apply binary closing with structuring element of size 2x2 (instead of 6x10) pixel
- no dilation (instead of binary dilation with structuring element of size 3x5 pixel)
- apply median filter with window size of 4x4 (instead of 5x3) pixel
- remove small objects if smaller then 10 (instead of 50) pixel
- enlarge segments in each direction by 10 (instead of 12) pixels

Basically the image preprocessing was adjusted to be more sensitive and to capture smaller sound components and species-specific sub-elements within larger song structures and call sequences. Figure 1 visualizes an example of the new segmentation method.

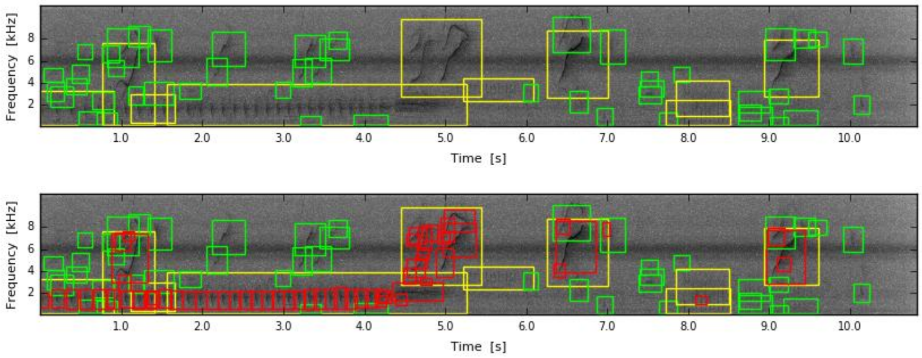

Fig. 1. Spectrogram re-segmentation example (MediaID: 8). **top**: initial segmentation (large segments marked in yellow), **bottom**: via re-segmentation of large segments extracted additional segments in red

With re-segmentation of previously segmented files 1,671,600 new segments were extracted and subsequently used for template matching to generate additional Segment-Probability features.

Extracting more features by segmenting files with low average precision. Besides re-segmentation of segmented files, additional files from the training set were chosen for segment extraction. However, instead of a random selection, a small number of files were chosen for each species (approx. 2 to 4) by selecting the ones having the lowest average precision score calculated during cross validation in previous training steps. This was done in two iterations (with new training and feature selection steps in between) increasing the number of features by additional 1,375,928 Segment-Probabilities.

2.1 Segment-Probabilities

For each species an individual feature set was formed by sweeping all segments related to that particular species over the spectrogram representations of all training and test recordings. The features were extracted via **multiresolution template matching** followed by selecting the maxima of the normalized cross-correlation [6]. In this context, multiresolution has a double meaning. On one hand it is referring to the time and frequency resolution of the spectrogram image itself. For 492,753 segments (already used for the BirdCLEF 2014 identification task) a time resolution of Δt = 11.6 ms (approx. 86 pixel per seconds) and a frequency resolution of Δf = 43.07 Hz (approx. 23 pixel per kHz) was used. For all other and newly extracted segments both time and frequency resolution was halved through downsampling the spectrogram image by a factor of 2. On the other hand the template matching can be also interpreted as multiresolution in terms of time and frequency range or size of the different spectrogram patches. Because further re-segmenting large segments, matching is performed for both: larger sound combinations (song syllables, call sequences) and smaller, rather fine-grained sound sub-elements (song elements, single calls).

2.2 Parametric Acoustic Features

Besides Segment-Probabilities, for some models also parametric acoustic features were used for prediction. To extract these features the openSMILE Feature Extractor Tool [7] was utilized again. The configuration file originally designed for emotion detection in speech signals was adapted to capture the characteristics of bird sounds. It first calculates 57 low-level descriptors (LLDs) per frame, adds delta (velocity) and delta-delta (acceleration) coefficients to each LLD and finally applies 39 statistical functionals on all, via moving average smoothened, feature trajectories.
The all in all 73 LLDs consist of: 1 time domain signal feature (zero crossing rate), 39 spectral features (Mel-spectrum bins 0-25; 25%, 50%, 75% and 90% spectral roll-off points; spectral centroid, flux, entropy, variance, skewness, kurtosis and slope; relative position of spectral minimum and maximum), 17 cepstral features (MFCC 0-16), 6 pitch-related features (F0, F0 envelope, F0 raw, voicing probability, voice quality, log harmonics-to-noise ratio computed from the ACF) and 10 energy-related features (logarithmic energy as well as energy in frequency bands: 150-500 Hz, 400-1000 Hz, 800-1500 Hz, 1000-2000 Hz, 1500-4000 Hz, 3000-6000 Hz, 5000-8000 Hz, 7000-10000 Hz and 9000-11000 Hz). To summarize an entire recording, statistics are calculated from all LLD, velocity and acceleration trajectories by 39 functionals including e.g. means, extremes, moments, percentiles and linear as well as quadratic regression. In total this sums up to 8541 (73·3·39) features per recording. Further details regarding openSMILE and the features extracted for bird identification can be found in the openSMILE book [8] and the OpenSmileForBirds_v2.conf configuration file [9].

3 Training and Feature Selection

The classification task was transformed to 999 one-vs-rest multi-label regression tasks. This way the number of selected features could be optimized separately and independently for each species during training. For each audio file in the training set the target function was set to 1.0 for the dominant species and 0.5 for all background species. Ensembles of randomized decision trees (ExtraTreesRegressor [10]) of the scikit-learn machine learning library were used for training and prediction [11].

Feature Selection during Training. Feature importance returned by the ensemble of decision trees was cumulated during training and used to rank individual features. The importance of each feature is determined by the total reduction of the mean squared error brought by that particular feature. After a complete training pass, including cross validation, the number of features was reduced by keeping only the N highest scoring and therefore most important features. The number N of features kept for the next training iteration was set to select 85% of the best features from the previous iteration.

Different percentages were tested (75% to 90%) to find a good compromise between time of training and finding the optimal number of features. After the time consuming feature reduction procedure (the number of training iterations was repeated until there were only 5 features left to predict each species) the optimal number and best performing features per species were selected by finding either the maximum of the Area Under the Curve (AUC) or alternatively the maximum mAP score calculated over the entire training set. Figure 2 shows two examples of resulting AUC (with and without background species), mAP and R2 (coefficient of determination) score trajectories when successively discarding 15% of the least important features. The maximum of each evaluation criteria is marked with a red square. The features used in the corresponding training iteration (maximum of AUC or mAP score) were then chosen for predicting the test files.

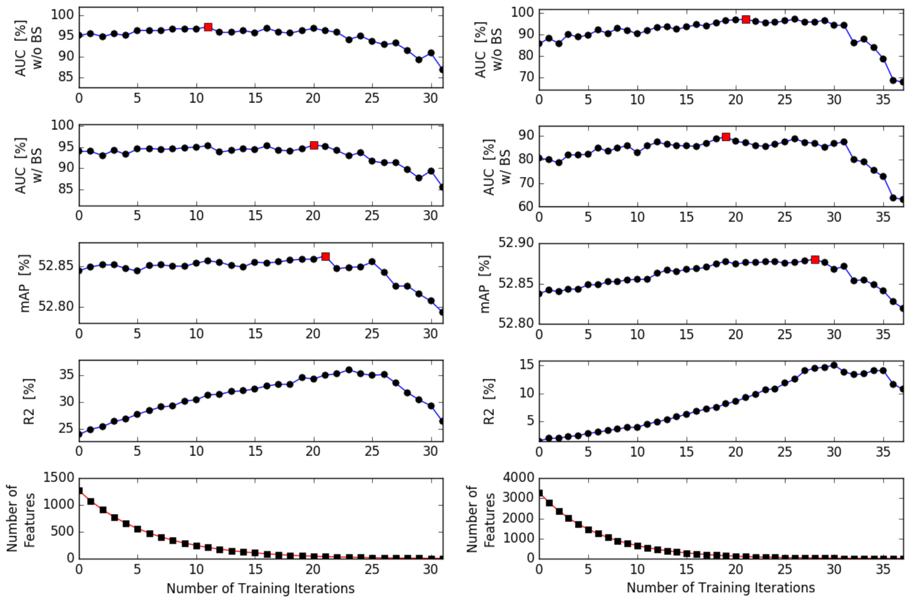

Fig. 2. Progress of AUC, mAP and R2 scores during feature selection for **left**: Scytalopus latrans (SpeciesID: lzezgo) and **right**: Psarocolius decumanus (SpeciesID: cxyhrl)

4 Submission Results

In Table 1 results of the submitted runs are summarized using two evaluation statistics: mean of the Area Under the Curve calculated per species and mean Average Precision on the public training and the private test sets. For all runs no external resources and only audio features (features extracted from audio files) were used for training and prediction.

Table 1. Performance of submitted runs (without | with background species)

	Public Training Set		Test Set	Test Soundscapes
Run	**Mean AUC [%]**	**mAP [%]**	**mAP [%]**	**mAP [%]**
1	96.4 | 93.5	64.1 | 61.9	58.5 | 51.9	13.7
2	96.2 | 92.1	62.4 | 58.2	39.9 | 33.6	-
3	96.9 | 92.2	74.5 | 70.1	45.6 | 39.6	13.0
4	97.1 | 92.5	74.7 | 70.2	55.1 | 47.2	12.9

Run 1. For the best performing first run just a single model was used. This model was trained using only a small but highly optimized selection of Segment-Probabilities (as described in the previous section). For this run, features were selected per species by optimizing the mAP score on the training set. A total of 125,402 features (with a minimum of 20, a maximum of 1833 and an average of 126 features per species) were used to predict all species in the test files.

Run 2. The second run was submitted quite early as an interim result and is therefore not worthy of being discussed here. It was actually supposed to be replaced by the submission of another run averaging the predictions of several different models. Unfortunately uploading could not be completed before the submission deadline.

Run 3. For the third submitted run blending of different models followed by post-processing was used as described in [5]. Predictions from all models created during training as well as predictions from the two best performing models submitted last year were included (in total 24 models). Some models used Segment-Probabilities or openSMILE features only, others a combination of both. Also different feature sets were used with the number of features included for training and prediction optimized regarding either AUC (with and without using background species) or mAP score.

Run 4. The fourth run also used blending to aggregate model predictions. But unlike run 3 only those predictions were included that after blending resulted in the highest possible mAP score calculated on the entire training set (13 models including the best model from 2015).

Figure 3 visualizes the official scores of the LifeCLEF 2016 Bird Identification Task. The here proposed approach ranked second place among all teams (MarioTsaBerlin Run 1 & 4) and provided the best system to identify birds in soundscape recordings.

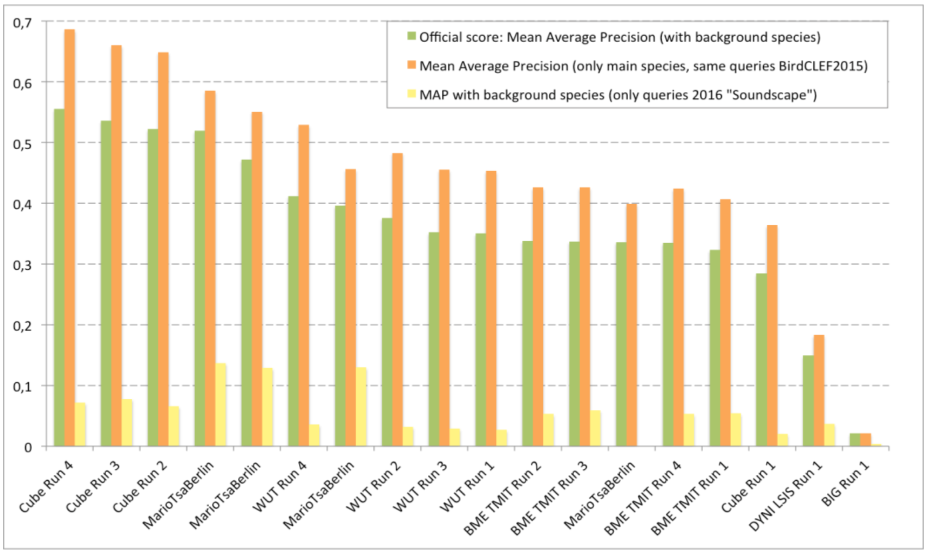

Fig. 3. Official scores of the LifeCLEF 2016 Bird Identification Task. The above described methods and submitted runs belong to MarioTsaBerlin.

5 Discussion

Interestingly, for the best performing submission (Run 1) just a single model was used with only one category of features. Although using a good selection of features for this model one would expect that blending several models with different feature sets would perform better than just a single one. One possible explanation for the comparatively weak results achieved with blending could be the inclusion of the best combined models from 2015. Those combined and post-processed predictions already showed a fairly high overfitting on the training set and blending was perhaps done in favor of these predictions at the cost of the maybe better generalizing new models.

On the other hand achieving an improvement by almost 30% on the mAP score with a single model (25% if taking background species into account) clearly shows that the techniques introduced this year could be applied very successfully. They also seem to complement each other quite well. Extracting additional, fine-grained spectrogram segments for template matching by re-segmenting larger segments captures typical sub-elements of songs or call sequences. Matching these sub-elements can give better identification performance than matching larger song structures, especially if those show a high variability between different individuals of the same species. The downside of the new segmentation method is a collection of many redundant or even useless segments e.g. when dealing with noisy recordings or overlapping sounds from other species or sources. However, the proposed feature selection method can compensate for that by successively discarding irrelevant features during training.

This year also deep learning techniques were successfully applied to the BirdCLEF dataset [12]. By using convolutional neural networks (CNNs) the best performing system achieved a mAP score of almost 70% when ignoring background species. It outperformed the here described approach by 17%, or 7% when also identifying all background species (see Fig.3 Cube Run 4). For soundscape recordings, however, the technique proposed in this paper achieved a 76% better performance than the best run using CNNs. Although identification performance for the new introduced test set was generally low among all teams, in the case of soundscapes, template matching seems to be better suited. The matching of rather small templates is not so much affected by surrounding sound events (e.g. coming from many simultaneously vocalizing animals) and therefore can create features more robust to various background noises. Compared to the black box architecture of a neuronal network classifier, using template matching and decision tree based feature selection also has some additional advantages. By visually or acoustically examining the most important and best discriminating sound elements of a species (typical calls, syllables or song phrases) one can gain a better insight into its sound repertoire and learn more about its call or song characteristics. The following figures visualize sound elements most suitable to identify a certain species. Each spectrogram segment is positioned at its original frequency position within a box representing the frequency range of 0 to 11025 Hz. More figures and additional material can be found at [13].

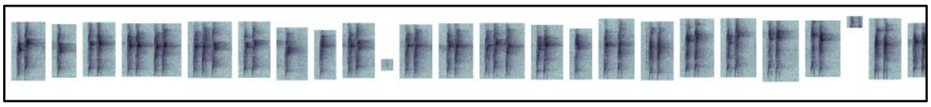

Fig. 4. Pallid Spinetail / Cranioleuca pallida, (ID: aiwvzm)

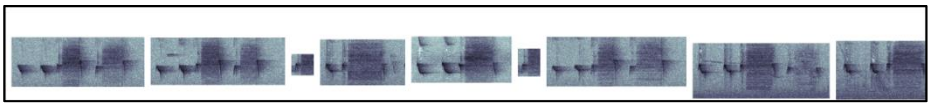

Fig. 5. Streak-headed Antbird / Drymophila striaticeps, (ID: alouyq)

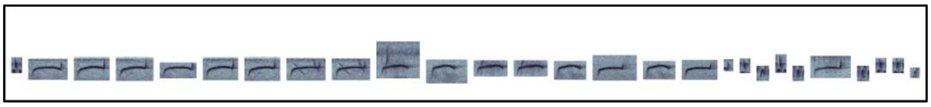

Fig. 6. Southern Beardless Tyrannulet / Camptostoma obsoletum, (ID: armfvy)

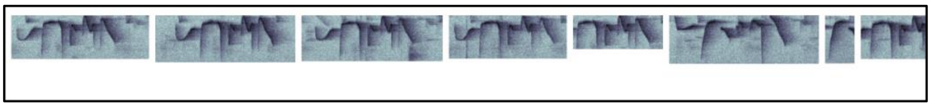

Fig. 7. Chestnut-capped Brush Finch / Arremon brunneinucha, (ID: ayfewp)

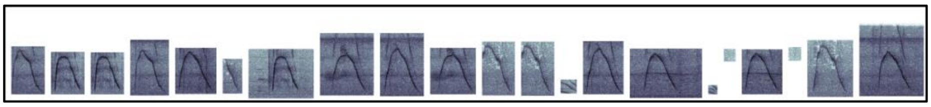

Fig. 8. Bare-faced Curassow / Crax fasciolata, (ID: bgqyzr)

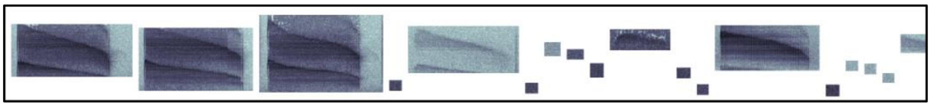

Fig. 9. Rufous-headed Pygmy Tyrant / Pseudotriccus ruficeps, (ID: bgtgbo)

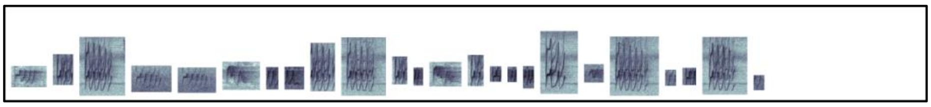

Fig. 10. Yellow-throated Flycatcher / Conopias parvus, (ID: bpufoz)

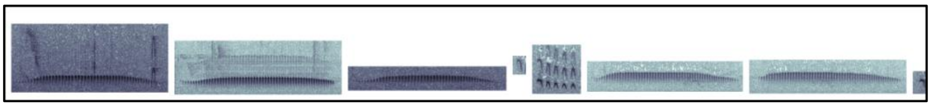

Fig. 11. Caatinga Antwren / Herpsilochmus sellowi, (ID: cuwfkj)

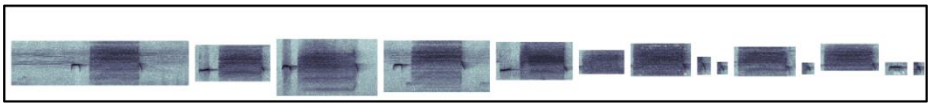

Fig. 12. Ochre-rumped Antbird / Drymophila ochropyga, (ID: dhzucj)

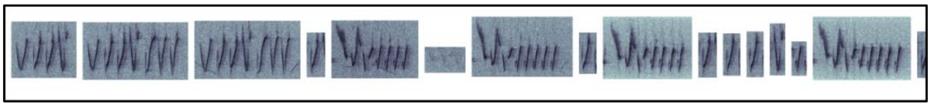

Fig. 13. Yellow-headed Brush Finch / Atlapetes flaviceps, (ID: dxtjbh)

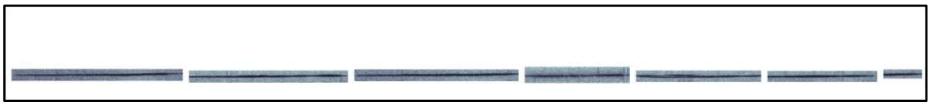

Fig. 14. Black-and-gold Cotinga / Tijuca atra, (ID: eofsmg)

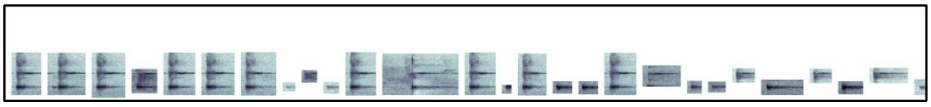

Fig. 15. Bare-throated Bellbird / Procnias nudicollis, (ID: fcojdk)

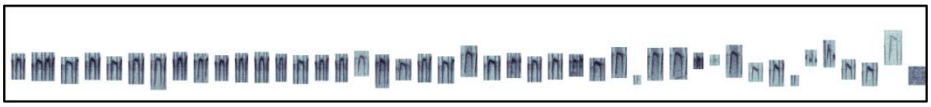

Fig. 16. Itatiaia Spinetail / Asthenes moreirae, (ID: gshgib)

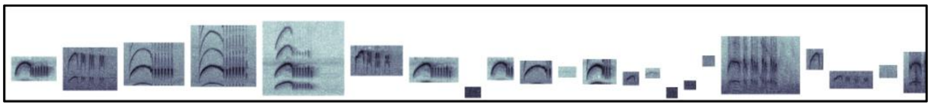

Fig. 17. Cliff Flycatcher / Hirundinea ferruginea, (ID: hdtboj)

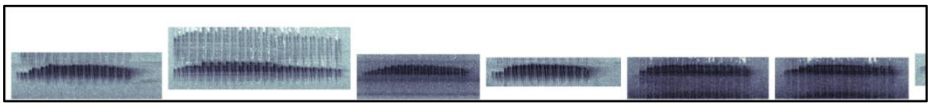

Fig. 18. Scalloped Antbird / Myrmeciza ruficauda, (ID: hgmqff)

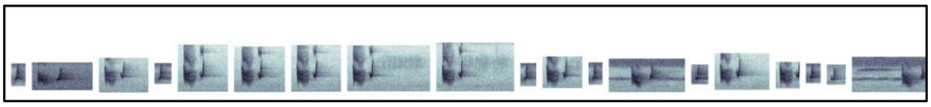

Fig. 19. Dwarf Tyrant-Manakin / Tyranneutes stolzmanni, (ID: iyshfg)

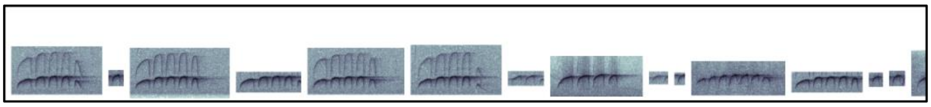

Fig. 20. Eastern Sirystes / Sirystes sibilator, (ID: jiuopg)

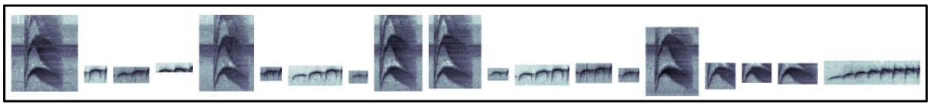

Fig. 21. Large-tailed Antshrike / Mackenziaena leachii, (ID: jpjrlh)

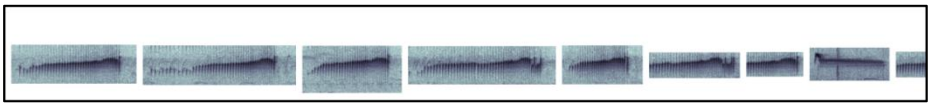

Fig. 22. Slender-billed Inezia / Inezia tenuirostris, (ID: myrlln)

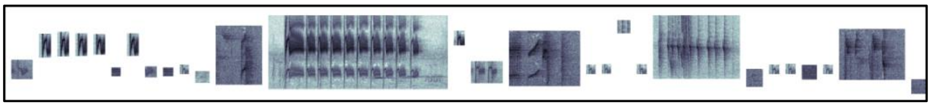

Fig. 23. White-throated Hummingbird / Leucochloris albicollis, (ID: orktkw)

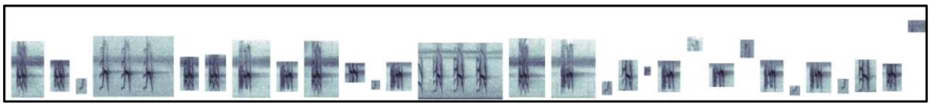

Fig. 24. Orange-breasted Thornbird / Phacellodomus ferrugineigula, (ID: osqzxt)

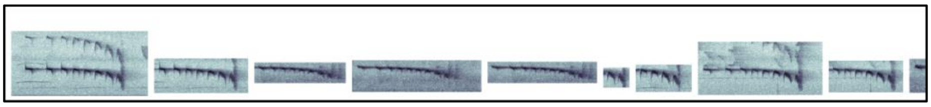

Fig. 25. Southern Chestnut-tailed Antbird / Myrmeciza hemimelaena, (ID: ptwyjr)

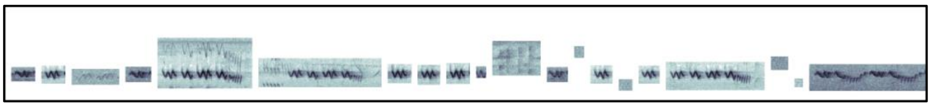

Fig. 26. Ashy-headed Greenlet / Hylophilus pectoralis, (ID: szaokc)

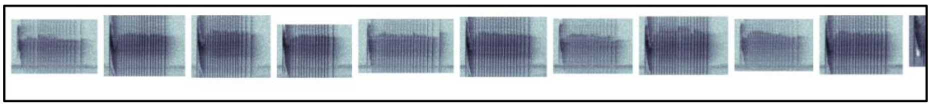

Fig. 27. Cinnamon Flycatcher / Pyrrhomyias cinnamomeus, (ID: uissyt)

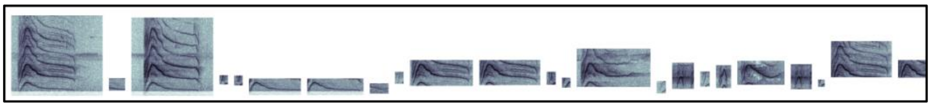

Fig. 28. Roadside Hawk / Rupornis magnirostris, (ID: uphahj)

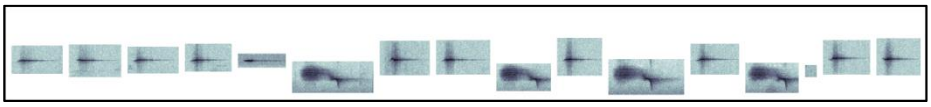

Fig. 29. Black-collared Jay / Cyanolyca armillata, (ID: vnkcgy)

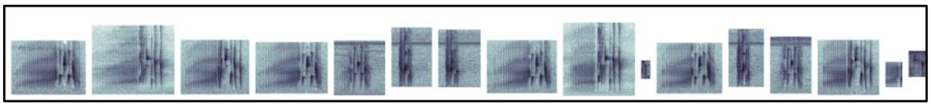

Fig. 30. Tawny-crowned Pygmy Tyrant / Euscarthmus meloryphus, (ID: wckepf)

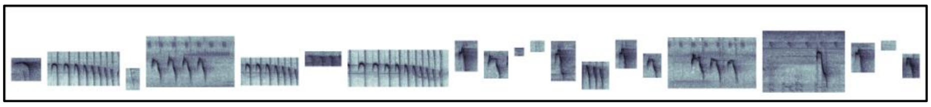

Fig. 31. Dot-winged Antwren / Microrhopias quixensis, (ID: xyvbwf)

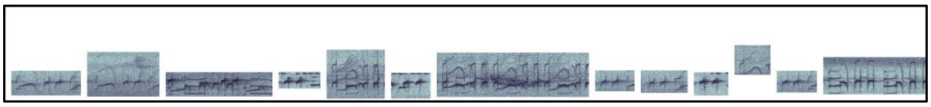

Fig. 32. Brown-banded Puffbird / Notharchus ordii, (ID: ynwbeg)

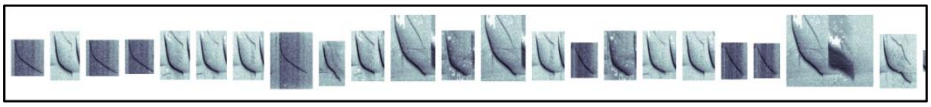

Fig. 33. Rufous-collared Sparrow / Zonotrichia capensis, (ID: yyjrms)

Acknowledgments. I would like to thank Hervé Glotin, Hervé Goëau, Willem-Pier Vellinga, Alexis Joly and Henning Müller for organizing this task and the Xeno-Canto foundation for nature sounds as well as the French projects Floris'Tic (INRIA, CIRAD, Tela Botanica) and SABIOD Mastodons for their support. I also want to thank the BMUB (Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit), the Museum für Naturkunde and especially Dr. Karl-Heinz Frommolt for supporting my research and Wolfram Fritzsch for providing me with additional hardware resources.

References

1. Goëau H, Glotin H, Planqué R, Vellinga WP, Joly A (2016) LifeCLEF Bird Identification Task 2016, In: CLEF working notes 2016
2. Joly A, Goëau H, Glotin H et al. (2016) LifeCLEF 2016: multimedia life species identification challenges, In: Proceedings of CLEF 2016
3. Lasseck M (2013) Bird Song Classification in Field Recordings: Winning Solution for NIPS4B 2013 Competition, In: Glotin H. et al. (eds.). Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod.org/nips4b, joint to NIPS, Nevada, dec. 2013: 176-181
4. Lasseck M (2015a) Towards Automatic Large-Scale Identification of Birds in Audio Recordings, In Lecture Notes in Computer Science Vol.9283: pp 364-375
5. Lasseck M (2015b) Improved Automatic Bird Identification through Decision Tree based Feature Selection and Bagging, In: Working notes of CLEF 2015 conference
6. Lewis JP (1995) Fast Normalized Cross-Correlation, Industrial Light and Magic
7. Eyben F, Weninger F, Gross F, Schuller B (2013) Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, In: Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013, doi:10.1145/2502081.2502224
8. http://www.audeering.com/research/opensmile
9. http://www.animalsoundarchive.org/RefSys/LifeCLEF2015
10. Geurts P et al. (2006) Extremely randomized trees, Machine Learning, 63(1), 3-42
11. Pedregosa F et al. (2011) Scikit-learn: Machine learning in Python. JMLR 12, pp. 2825-2830
12. Sprengel E (2016) Audio Based Bird Species Identifcation using Deep Learning Techniques, In: Working notes of CLEF 2016 conference
13. http://www.animalsoundarchive.org/RefSys/LifeCLEF2016