Overview of BirdCLEF 2018: monospecies vs. soundscape bird identification

Hervé Goëau, Stefan Kahl, Hervé Glotin, Robert Planqué, Willem-Pier Vellinga, Alexis Joly

Overview of BirdCLEF 2018: monospecies vs. soundscape bird identification

Hervé Goëau[1], Stefan Kahl[4], Hervé Glotin[2], Robert Planqué[3], Willem-Pier Vellinga[3], and Alexis Joly[5]

[1] CIRAD, UMR AMAP, Montpellier, France herve.goeau@cirad.fr
[2] Université de Toulon, Aix Marseille Univ, CNRS, LIS, DYNI team, Marseille, France herve.glotin@univ-tln.fr
[3] Xeno-canto Foundation, The Netherlands, {wp,bob}@xeno-canto.org
[4] Chemnitz University of Technology stefan.kahl@informatik.tu-chemnitz.de
[5] Inria/LIRMM ZENITH team, Montpellier, France alexis.joly@inria.fr

Abstract. The BirdCLEF challenge offers a large-scale proving ground for system-oriented evaluation of bird species identification based on audio recordings of their sounds. One of its strengths is that it uses data collected through Xeno-canto, the worldwide community of bird sound recordists. This ensures that BirdCLEF is close to the conditions of real-world application, in particular with regard to the number of species in the training set (1500). Two main scenarios are evaluated: (i) the identification of a particular bird species in a recording, and (ii), the recognition of all species vocalising in a long sequence (up to one hour) of raw soundscapes that can contain tens of birds singing more or less simultaneously. This paper reports an overview of the systems developed by the six participating research groups, the methodology of the evaluation of their performance, and an analysis and discussion of the results obtained.

Keywords: LifeCLEF, bird, song, call, species, retrieval, audio, collection, identification, fine-grained classification, evaluation, benchmark, bioacoustics, ecological monitoring

1 Introduction

Accurate knowledge of the identity, the geographic distribution and the evolution of bird species is essential for a sustainable development of humanity as well as for biodiversity conservation. The general public, especially so-called "birders" as well as professionals such as park rangers, ecological consultants and of course ornithologists are potential users of an automated bird sound identifying system, typically in the context of wider initiatives related to ecological surveillance or biodiversity conservation. The BirdCLEF challenge evaluates the state-of-the-art of audio-based bird identification systems at a very large scale. Before BirdCLEF started in 2014, three previous initiatives on the evaluation of acoustic bird species identification took place, including two from the SABIOD[6]

[6] Scaled Acoustic Biodiversity http://sabiod.univ-tln.fr

group [3,2,1]. In collaboration with the organizers of these previous challenges, the BirdCLEF 2014, 2015, 2016 and 2017 challenges went one step further by (i) significantly increasing the species number by an order of magnitude, (ii) working on real-world social data built from thousands of recordists, and (iii) moving to a more usage-driven and system-oriented benchmark by allowing the use of metadata and defining information retrieval oriented metrics. Overall, these tasks were much more difficult than previous benchmarks because of the higher confusion risk between the classes, the higher background noise and the higher diversity in the acquisition conditions (different recording devices, contexts diversity, etc.).
The main novelty of the 2017 edition of the challenge with respect to the previous years was the inclusion of *soundscape recordings* containing time-coded bird species annotations. Usually xeno-canto recordings focus on a single foreground species and result from using mono-directional recording devices. Soundscapes, on the other hand, are generally based on omnidirectional recording devices that monitor a specific environment continuously over a long period. This new kind of recording reflects (possibly crowdsourced) passive acoustic monitoring scenarios that could soon augment the number of collected sound recordings by several orders of magnitude.
For the 2018-th edition of the BirdCLEF challenge, we continued evaluating both scenarios as two different tasks: (i) the identification of a particular bird specimen in a recording of it, and (ii), the recognition of all specimens singing in a long sequence (up to one hour) of raw soundscapes that can contain tens of birds singing simultaneously. In this paper, we report the methodology of the conducted evaluation as well as an analysis and a discussion of the results achieved by the six participating groups.

2 Tasks description

2.1 Task1: monospecies (monophone) recordings

The goal of the task is to identify the species of the most audible bird (*i.e.* the one that was intended to be recorded) in each of the provided test recordings. Therefore, the evaluated systems have to return a ranked list of possible species for each of the 12,347 test recordings. Each prediction item (i.e. each line of the file to be submitted) has to respect the following format:

<MediaId;ClassId;Probability;Rank>

Each participating group was allowed to submit up to 4 *run files* providing the predictions made from 4 different methods. The use of any of the provided metadata complementary to the audio content was authorized. It was also allowed to use any external training data but at the condition that (i) the experiment is entirely re-producible, i.e. that the used external resource is clearly referenced and accessible to any other research group in the world, (ii) participants submit

at least one run without external training data so that we can study the contribution of such resources, (iii) the additional resource does not contain any of the test observations. It was in particular strictly forbidden to crawl training data from: www.xeno-canto.org.

The dataset was the same as the one used for BirdCLEF 2017 [4], mostly based on the contributions of the Xeno-Canto network. The training set contains 36,496 recordings covering 1500 species of central and south America (the largest bioacoustic dataset in the literature). It has a massive class imbalance with a minimum of four recordings for Laniocera rufescens and a maximum of 160 recordings for Henicorhina leucophrys. Recordings are associated to various metadata such as the type of sound (call, song, alarm, flight, etc.), the date, the location, textual comments of the authors, multilingual common names and collaborative quality ratings. The test set contains 12,347 recordings of the same type (mono-phone recordings). More details about that data can be found in the overview working note of BirdCLEF 2017 [4].

The used evaluation metric is the Mean Reciprocal Rank (MRR). The MRR is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries ordered by probability of correctness. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer. The MRR is the average of the reciprocal ranks for the whole test set:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i}$$

where $|Q|$ is the total number of query occurrences in the test set.

Mean Average Precision was used as a secondary metric to take into account the background species, considering each audio file of the test set as a query and computed as:

$$mAP = \frac{\sum_{q=1}^{|Q|} AveP(q)}{Q},$$

where $AveP(q)$ for a given test file q is computed as

$$AveP(q) = \frac{\sum_{k=1}^{n} (P(k) \times rel(k))}{\text{number of relevant documents}}.$$

Here k is the rank in the sequence of returned species, n is the total number of returned species, $P(k)$ is the precision at cut-off k in the list and $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant species (i.e. one of the species in the ground truth).

2.2 Task2: soundscape recordings

The goal of the task was to localize and identify all audible birds within the provided soundscape recordings. Therefore, each soundscape was divided into

segments of 5 seconds, and a list of species accomapnied by probability scores had to be returned for each segment. Each prediction item (i.e. each line of the *run file*) had to respect the following format:

<MediaId;TC1-TC2;ClassId;Probability>

where probability is a real value in [0;1] decreasing with the confidence in the prediction, and where TC1-TC2 is a timecode interval with the format of hh:mm:ss with a length of 5 seconds (e.g.: 00:00:00-00:00:05, then 00:00:05-00:00:10).

Each participating group was allowed to submit up to 4 run files built from different methods. As for the monophone task, participants were allowed to use the provided metadata and to use external training data at the condition that the experiment is entirely re-producible and not biased.

The training set provided for this task was the same as that for the monophone task, *i.e.* 36,496 monophone recordings coming from Xeno-canto and covering 1500 species of Central and South America. Complementary to that data, a validation set of soundscapes with time-coded labels was provided as training data. It contained about 20 minutes of soundscapes representing 240 segments of 5 seconds and with a total of 385 bird species annotations. The test set used for the final blind evaluation contained about 6 hours of soundscapes split into 4382 segments of 5 seconds (to be processed as separate queries). Some of them were stereophonic, offering possibilities of source separation to enhance the recognition. More details about the soundscape data (locations, authors, etc.) can be found in the overview working note of BirdCLEF 2017 [4]. In a nutshell, 2 hours of soundscapes were recorded in Peru (with the support of Amazon Explorama Lodges within the BRILAAM STIC-AmSud 17-STIC-01 and SABIOD.org project) and 4,5 hours were recorded in Columbia by Paula Caycedo Rosales, ornithologist from the Biodiversa Foundation of Colombia and an active Xeno-canto recordist.

In order to assist participants in the development of their system, a baseline code repository and a validation dataset were shared with the participants. The validation package contained 20 minutes of annotated soundscapes split into 5 recordings taken from last year's test dataset. The baseline repository [7] was developed by Chemnitz University of Technology and offered tools and an example workflow covering all required topics such as spectrogram extraction, deep neural network training, audio classification on field recordings and local validation (more details can be found in [9]).

The metric used for the evaluation of the soundscape task was the classification mean Average Precision ($cmAP$), considering each class c of the ground

[7] https://github.com/kahst/BirdCLEF-Baseline

truth as a query. This means that for each class c, all predictions with $ClassId = c$ are extracted from the run file and ranked by decreasing probability in order to compute the average precision for that class. Then, the mean across all classes is computed as the main evaluation metric. More formally:

$$cmAP = \frac{\sum_{c=1}^{C} AveP(c)}{C}$$

where C is the number of classes (species) in the ground truth and $AveP(c)$ is the average precision for a given species c computed as:

$$AveP(c) = \frac{\sum_{k=1}^{n_c} P(k) \times rel(k)}{n_{rel}(c)}.$$

where k is the rank of an item in the list of the predicted segments containing c, n_c is the total number of predicted segments containing c, $P(k)$ is the precision at cut-off k in the list, $rel(k)$ is an indicator function equaling 1 if the segment at rank k is a relevant one (*i.e.* is labeled as containing c in the ground truth) and $n_{rel}(c)$ is the total number of relevant segments for class c.

3 Participants and methods

29 research groups registered for the BirdCLEF 2018 challenge and 6 of them finally submitted a total of 45 runs (23 runs for task1: monophone recordings and 22 runs for task2: soundscape recordings). Details of the methods used and systems evaluated are collected below (by alphabetical order) and further discussed in the working notes of the participants [6,10,12,8,11]:

Duke, China-USA, 8 runs [6]: This participant designed a bi-modal neural network aimed at learning a joint representation space for the audio and the metadata information (latitude, longitude, elevation and time). It relies on a relatively shallow architecture with 6 convolutional layers for the audio and a few full-connected layers aimed at learning features from the meta-data and combining them with the audio features into a single representation space. A softmax is then used for the classification output. Concerning the monophone subtask, DKU SMIIP run3 uses the bi-modal model whereas DKU SMIIP run 2 uses only the audio-based part. DKU SMIIP run 3 is a fusion of both runs. DKU SMIIP run 4 relies on a ResNet model as for comparison with the proposed model. DKU SMIIP run 5 is a combination of all models. Concerning the soundscape subtask, DKU SMIIP run1 uses the bi-modal model, DKU SMIIP run2 uses an ensemble of two bi-modal models (one with data augmentation and one without data augmentation). DKU SMIIP run3 is a fusion and run1 and run2. DKU SMIIP run4 is a fusion of all models including the ResNet.

ISEN, France, 4 runs: This participant used the *Soundception* approach presented in [13] and which was the best performing system of the previous edition

of BirdCLEF. It is based on an Inception-v4 architecture extended with a time-frequency attention mechanism.

MfN, Germany, 8 runs [10]: This participant trained an ensemble of convolutional neural networks based on the Inception-V3 architecture applied to mel-scale spectrograms as input. The trained models mainly differ in the pre-processing that was used to extract the spectrograms (with or without high-pass filter, sampling rate value, mono vs. stereo, FFT parameters, frequency scaling parameters, etc. Another particularity of this participant is that he uses intensive data augmentation both in the temporal and frequency domain. About ten different data augmentation techniques were implemented and evaluated separately through cross-validation ablation tests. Among them, the most contributing one is indisputably the addition of background noise or sounds from other files belonging to the same bird species with random intensity, in order to simulate artificially numerous context where a given species can be recorded. Other augmentations seem not to contribute as much taken individually, but one after one, point after point, they lead to significant improvements. Data augmentation most notably included a low-quality degradation based on MP3 encoding-decoding, jitter on duration (up to 0.5 sec), random factor to signal amplitude, random cyclic shift, random time interval dropouts, global and local pitch shift and frequency stretch, as well as color jitter (brightness, contrast, saturation, hue). MfN Run 1 for each subtask included the best single model learned during preliminary evaluations. These two models mainly differ in the pre-processing of audio files and choice of FFT parameters. MfN Run 2 combines both models, MfN Run 3 added a third declination of the model with other FFT parameters, but combined the predictions of the two best snapshots per model (regarding performance on the validation set) for averaging 3x2 predictions per species. MfN Run 4 added 4 more models and earlier snapshots of them, reaching a total combination of 18 predictions per species. No additional metadata was used except for the elimination of species based on the year of introduction in the BirdCLEF challenge.

OFAI, Austria, 7 runs [12]: This participant carefully designed a CNN architecture dedicated to birds sounds analysis in the continuity of its previous work described in [5] (the *sparrow* model). The main architecture is quite shallow with a first block of 6 convolutional layers aimed at extracting features from mel-spectrograms, a species prediction block aimed at computing local predictions every 9 frames, and a temporal pooling block aimed at combining the local predictions into a single classification for the whole audio excerpt. Several variants of this base architecture were then used to train a total of 17 models (with or without ResNet blocks instead of classical convolutional layers, different temporal pooling settings, with or without background species prediction). Complementary to audio-based models, this participant also studied the use of metadata-based models. In total, 24 MLPs were trained and based on four main variables: date, elevation, localization and time. The different MLPs mainly dif-

fer in the used variables (all, only one, all except one, etc.) and various parameter settings.

TUC MI, Germany, 10 runs [8]: All runs by this participant were conducted thanks to the baseline BirdCLEF package provided by Chemnitz University [9]. They ensemble different learning and testing strategies as well as different model architectures. Classical deep learning techniques were used, covering audio-only and metadata assisted predictions. Three different model architectures were employed: First, a shallow, strictly sequential model with only a few layers. Secondly, a custom variation of the WideResNet architecture with multiple tens of layers and thirdly a very slim and shallow model which is suited for inference on low-power devices such as the Raspberry Pi. The inputs for all three models are 256 x 128 pixel mel-scale log-amplitude spectrograms with a frequency range from 500 Hz to 15 kHz. The dataset is pre-processed using a bird activity estimator based on median thresholds similar to previous attempts of this participant [7]. The most successful run for the monospecies task was an ensemble consisting of multiple trained nets covering different architectures and dataset splits. The participant tried to estimate the species list for the soundscape task based on time of the year and location using the eBird database. Despite the success of this approach in last year's attempt, the pre-selection of species did not improve the results compared to a large ensemble. Finally, the participant tried to establish a baseline for real-time deployments of neural networks for long-term biodiversity monitoring using cost-efficient platforms. The participant proposes a promising approach to shrinking model size and reducing computational costs using model distillation. The results of those runs using the slim architecture are only a fraction behind the scores of large ensembles. All additional metadata and code are published online, complementing the baseline BirdCLEF package.

ZHAW, Switzerland, 8 runs [11]: In contrast to every other submission, the participants evaluated the use of recurrent neural networks (RNN). Using time-series as inputs for recurrent network topologies seems to be the most intuitive approach for bird sound classification. Yet, this method did not receive much attention in past years. Despite the limitations of time and computational resources, the experiments showed that bidirectional LSTMs are capable of classifying bird species based on two-dimensional inputs. Tuning RNNs to improve the overall performance seems to be challenging, although works from other sound domains showed promising results. The participants noted that not every design decision from other CNN implementations carry their benefit over to a RNN-based approach. Especially dataset augmentation methods like noise samples did not improve the results as expected. The results of the submitted runs suggest that an increased number of hidden LSTM units has significant impact on the overall performance. Additionally, data pre-processing and detection post-filtering impacts the prediction quality. Longer input segments and LSTMs with variable input length should be subject to future research.

4 Results

The results achieved by all the evaluated system are displayed on Figure 1 for the monospecies recordings and on Figure 2 for the soundscape recordings. The main conclusion we can draw from that results are the following:

The overall performance improved significantly over last year for the mono-species recordings but not for the soundscapes: The best evaluated system achieves an impressive MRR score of 0.83 this year whereas the best system evaluated on the same dataset last year [13] achieved a MRR of 0.71. On the other side, we do not measured any strong progress on the soundscapes. The best system of MfN this year actually reaches a c-mAP of 0.193 whereas the best system of last year on the same test dataset [13] achieved a c-mAP of 0.182.

Inception-based architectures perform very well: As previous year, the best performing system of the challenge is based on an Inception architecture, in particular the Inception v3 model used by MfN. In their working note [10], the authors report that they also tested (for a few training epochs) more recent or larger architectures that are superior in other image classification tasks (ResNet152, DualPathNet92, InceptionV4, DensNet, InceptionResNetV2, Xception, NasNet). But none of them could meet the performance of the InceptionV3 network with attention branch.

Intensive data augmentation provides strong improvement: All the runs of MfN (which performed the best within the challenge) made use of intensive data augmentation, both in the temporal and frequency domain (see section 3 for more details). According to the cross-validation experiments of the authors [10], such intensive data augmentation allows the MRR score to be increased from 0.65 to 0.74 for a standalone Inception V3 model.

Shallow and compact architectures can compete with very deep architectures: Even if the best runs of MfN and ISEN are based on a very deep Inception model (Inception v3), it is noteworthy that shallow and compact architectures such as the ones carefully designed by OFAI can reach very competitive results, even with a minimal number of data augmentation techniques. In particular, OFAI Run 1 that is based on an ensemble of shallow networks performs better than the runs of ISEN, based on an Inception v4 architecture.

Using metadata provides observable improvements: Contrary to all previous editions of LifeCLEF, one participant succeeded this year in improving significantly the predictions of its system by using the metadata associated to each observation (date, elevation, localization and time). More precisely, OFAI Run 2 combining CNNs and metadata-based MLPs achieves a mono-species MRR of 0.75 whereas OFAI Run 1, relying solely on the CNNs, achieves a MRR of 0.72. According to the cross-validation experiments of this participant [12], the most contributing information is the localization. The elevation is the second most informative variable but as it is highly correlated to the localization, it does not provide a strong additional improvement in the end. Date and then Time are the less informative but they do contribute to the global improvement of the MRR.

The brute-force assembling of networks provides significant improvements: as for many machine learning challenges (including previous BirdCLEF editions), the best runs are achieved by the combination of several deep neural networks (*e.g.* 18 CNNs for MfN Run 4). The assembling strategy differs from a participant to another. MfN rather tried to assemble as much networks as possible. MfN Run 4 actually combines the predictions of all the networks that were trained by this participant (mainly based on different pre-processing and weights initialization), as well as snapshots of these models recorded earlier during the training phase. The gain of the ensemble over a single model can be observed by comparing MfN Run 4 ($MRR = 0.83$) to MfN Run 1 ($MRR = 0.78$). The OFAI team rather tried to select and weight the best performing models according to their cross-validation experiments. Their best performing run (OFAI Run 3) is a weighted combination of 11 CNNs and 8 metadata-based MLPs. It allows reaching a score of $MRR = 0.78$ whereas the combination of the best single audio and metadata models achieves a score of $MRR = 0.69$ (OFAI Run 4).

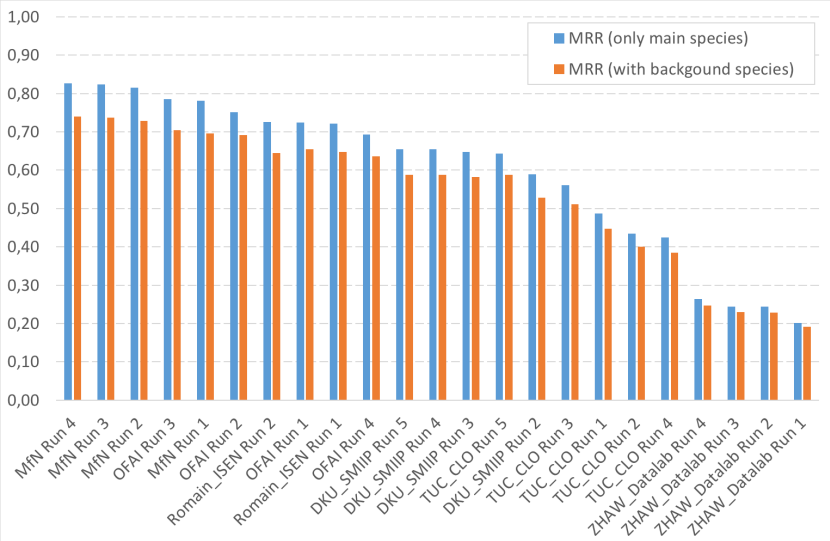

Fig. 1. BirdCLEF 2018 monophone identification results - Mean Reciprocal Rank.

5 Conclusion

This paper presented the overview and the results of the LifeCLEF bird identification challenge 2018. It confirmed the results of the previous edition that inception-based convolutional neural networks on mel spectrograms provide the best performance. Moreover, the use of large ensembles of such networks and of

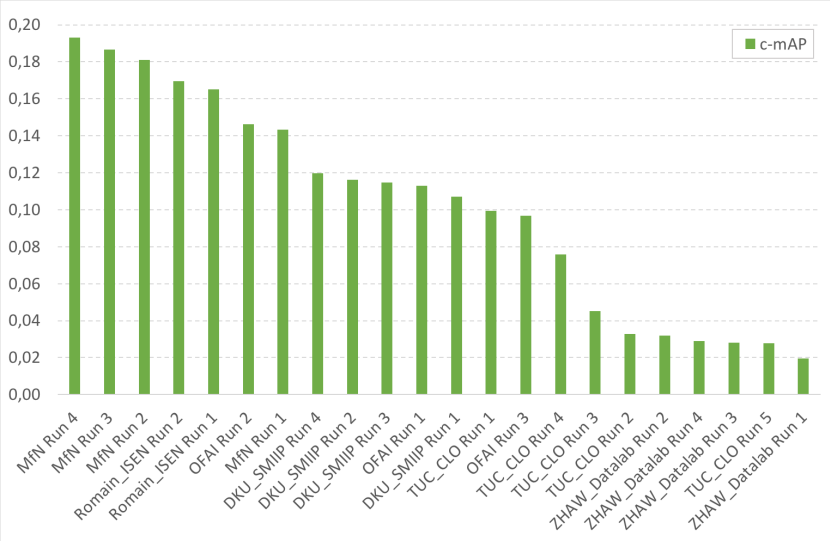

Fig. 2. BirdCLEF 2018 soundscape identification results - classification Mean Average Precision.

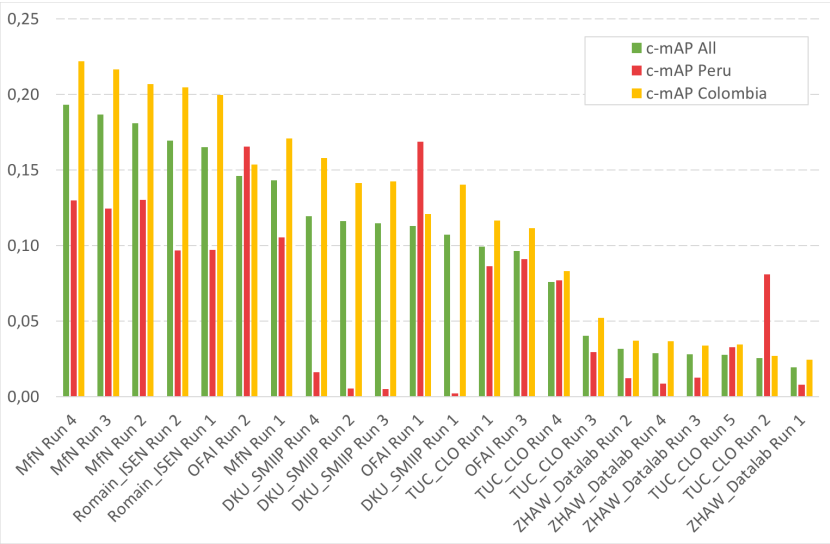

Fig. 3. BirdCLEF 2018 soundscape identification results detailed per country - classification Mean Average Precision.

intensive data augmentation provides significant additional improvements. The best system of this year achieved an impressive MRR score of 0.83 on the typical Xeno-Canto recordings. It could probabaly even be improved by a few points by combining it with a metadata-based prediction model, as shown by the second best participant to the challenge. This means that the technology is now mature enough for this scenario. Concerning the soundscapes recordings however, we did not observe any significant improvement over the performance of last year. Recognizing many overlapping birds remains a hard problem and none of the efforts made by the participants to tackle it provided observable improvement. In the future, we will continue investigating this scenario, in particular through the introduction of a new dataset of several hundred hours of annotated soundscapes that could be partially used as training data.

Acknowledgements The organization of the BirdCLEF task is supported by the Xeno-canto Foundation as well as by the French CNRS project SABIOD.ORG and EADM GDR CNRS MADICS, BRILAAM STIC-AmSud, and Floris'Tic. The annotations of some soundscapes were prepared by the regretted wonderful Lucio Pando of Explorama Lodges, with the support of Pam Bucur, H. Glotin and Marie Trone.

References

1. Briggs, F., Huang, Y., Raich, R., Eftaxias, K., et al., Z.L.: The 9th mlsp competition: New methods for acoustic classification of multiple simultaneous bird species in noisy environment. In: IEEE Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–8 (2013)
2. Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J.: Bioacoustic challenges in icml4b. In: in Proc. of 1st workshop on Machine Learning for Bioacoustics. No. USA, ISSN 979-10-90821-02-6 (2013), `http://sabiod.org/ICML4B2013_proceedings.pdf`
3. Glotin, H., Dufour, O., Bas, Y.: Overview of the 2nd challenge on acoustic bird classification. In: Proc. Neural Information Processing Scaled for Bioacoustics. NIPS Int. Conf., Ed. Glotin H., LeCun Y., Artières T., Mallat S., Tchernichovski O., Halkias X., USA (2013), `http://sabiod.org/NIPS4B2013_book.pdf`
4. Goëau, H., Glotin, H., Planque, R., Vellinga, W.P., Joly, A.: Lifeclef bird identification task 2017. In: CLEF working notes 2017 (2017)
5. Grill, T., Schlüter, J.: Two convolutional neural networks for bird detection in audio signals. In: Signal Processing Conference (EUSIPCO), 2017 25th European. pp. 1764–1768. IEEE (2017)
6. Haiwei, W., Ming, L.: Construction and improvements of bird songs' classification system. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
7. Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M.: Large-scale bird sound classification using convolutional neural networks. In: CLEF 2017 (2017)
8. Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: A baseline for large-scale bird species identification in field recordings. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)

9. Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: Recognizing birds from sound-the 2018 birdclef baseline system. arXiv preprint arXiv:1804.07177 (2018)
10. Lasseck, M.: Audio-based bird species identification with deep convolutional neural networks. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
11. Müller, L., Marti, M.: Two bachelor students' adventures in machine learning. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
12. Schlüter, J.: Bird identification from timestamped, geotagged audio recordings. In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum) (2018)
13. Sevilla, A., Glotin, H.: Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: Working Notes of CLEF 2017 (Cross Language Evaluation Forum) (2017), http://ceur-ws.org/Vol-1866/paper_177.pdf