FEW-SHOT BIOACOUSTIC EVENT DETECTION: A NEW TASK AT THE DCASE 2021 CHALLENGE

Veronica Morfi[1], *Inês Nolasco*[1], *Vincent Lostanlen*[2], *Shubhr Singh*[1],
Ariana Strandburg-Peshkin[3,4], *Lisa Gill*[5], *Hanna Pamuła*[6], *David Benvent*[7], *Dan Stowell*[8]

[1] Centre for Digital Music (C4DM), Queen Mary University of London, London, UK
[2] CNRS, Laboratoire des sciences du numérique de Nantes (LS2N), Nantes, France
[3] Dept. of Biology & Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Germany
[4] Dept. for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Germany
[5] BIOTOPIA Naturkundemuseum Bayern, Munich, Germany
[6] AGH University of Science and Technology, Kraków, Poland
[7] Cornell Lab of Ornithology, Cornell University, Ithaca, NY, US
[8] Tilburg University, Tilburg, The Netherlands; Naturalis Biodiversity Centre, Leiden, The Netherlands

ABSTRACT

Few-shot bioacoustic event detection is a novel area of research that emerged from a need in monitoring biodiversity and animal behaviour: to annotate long recordings, that experts usually can only provide very few annotations for due to the task being specialist and labour-intensive. This paper presents an overview of the first evaluation of few-shot bioacoustic sound event detection, organised as a task of the DCASE 2021 Challenge. A set of datasets consisting of mammal and bird multi-species recordings in the wild, along with class-specific temporal annotations, was compiled for the challenge, for the purpose of training learning-based approaches and for evaluation of the submissions in a few-shot labelled dataset. This paper describes the task in detail, the datasets that were used for both development and evaluation of the submitted systems, along with how system performance was ranked and the characteristics of the best-performing submissions. Some common strategies that the participating teams used are discussed, including input features, model architectures, transferring of prior knowledge, use of public datasets and data augmentation. Ranking for the challenge was based on overall performance of the evaluation set, however in this paper we also present results on each of the subsets of the evaluation set. This new analysis reveals submissions that performed better on specific subsets and gives an insight as to characteristics of the subsets that can influence performance.

Index Terms— Few-shot learning, bioacoustics, sound event detection, DCASE challenge

1. INTRODUCTION

The task of bioacoustic event detection refers to the retrieval of animal vocalizations in terms of onset and offset times. Thus, it shares a common methodology with other sound event detection (SED) contexts, such as offices [1], homes [2], city streets [3], and high-security spaces [4]. Yet, the application domain of bioacoustics is particularly challenging for SED, in part because of the high diversity of possible recording conditions and of vocalisation types [5]. For this reason, the field of machine learning for bioacoustics remains divided into many subfields: birds [6], land mammals, marine mammals [7], and so forth.

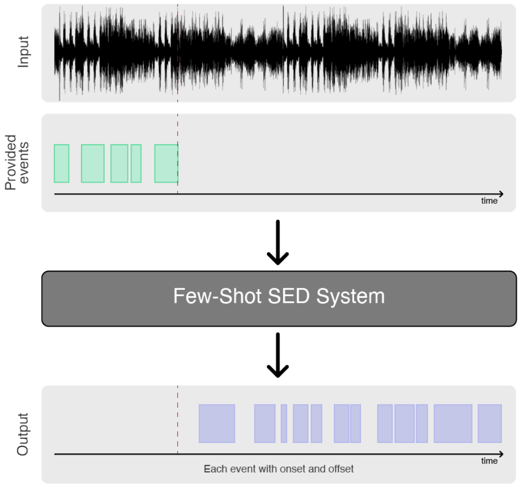

Figure 1: Overview of the proposed few-shot bioacoustic event detection task at the DCASE 2021 challenge. Green and purple rectangles represent labeled and predicted events, respectively.

The past decade witnessed the surge of deep convolutional networks (CNNs) in the time–frequency domain, which have the potential to outperform feature engineering. However, a supervised CNN for SED requires a predefined taxonomy of acoustic events as well as hundreds of annotated examples per class. Yet, collecting a large training set of animal vocalizations is not always feasible in practice, because species are unequally abundant [8]; audio annotation is costly and time-consuming [9]; and, more fundamentally, the taxonomy may vary depending on the use case [10].

We address this problem by introducing *few-shot bioacoustic event detection* as a new task to the DCASE 2021 challenge. In contrast to traditional deep learning approaches that use a large amount of data to train models, the key idea behind few-shot learning is to build accurate models with less training data [11]. More specifically, few-shot learning is usually studied using N-way-k-shot classification, where N denotes the number of classes and k the number of known examples for each class. Figure 1 illustrates the functioning of the system in deployment. After being trained on the first $k = 5$ occurrences of an event of interest, the system detects all the remaining occurrences of the same event in the rest of the recording.

Diverse approaches have been used to address the few-shot learning problem for classification, with no consensus on the best. Some use prior knowledge about similarity between sounds by computing embeddings (learnt representation spaces) while training and discriminate between unseen classes [11], while others exploit prior knowledge about the structure of the data by using augmentation to synthesize new data [12]. Finally, some approaches can learn models with parameters that can be fine-tuned to smaller datasets [13]. All of the above approaches deal with classification tasks in a few-shot learning setting and there is still much to be learnt in the field of few-shot SED; especially in concern to bioacoustic events.

While typical SED models must be retrained from scratch for each new use case, this few-shot formulation aims at learning generic representations of bioacoustic sounds. We encourage the community to develop an open-set SED system which bioacoustics practitioners will use on their own data after a modest amount of annotation, i.e., identifying the first k examples for each sound type.

2. DATASETS

A *development dataset* was provided for the task when the challenge was launched, consisting of predefined training and validation sets to be used for system development.[1] The development set consists of datasets from multiple sources with audio recordings and associated reference annotations in a task-specific format. More specifically, for the training set multi-class temporal annotations were provided for each recording as: positive (POS), negative (NEG) and unknown (UNK), while for the validation set single-class temporal annotations (POS/UNK) were provided for each recording.

A separate *evaluation set* was kept for evaluating the performance of the systems.[2] It consists of datasets from multiple sources. During the task five event annotations were provided for each of the recordings for the class of interest. The developed systems had to use those five annotated events and then learn to detect the same type of events throughout the rest of the recording.

Table 1 presents an overview of all the datasets in the development and evaluation sets, with information about the microphones used during recording, number of audio files, total time duration of the set, number of labels and number of annotated events.

BirdVox-DCASE-10h (BV): The BirdVox-DCASE-10h (BV) contains five audio files from four different autonomous recording units, each lasting two hours. These autonomous recording units are all located in Tompkins County, NY, US. They follow the same hardware specification: the Recording and Observing Bird Identification Node (ROBIN) developed by the Cornell Lab of Ornithology [14]. All recordings were acquired in 2015, during the fall migration season. An expert ornithologist, Andrew Farnsworth, has annotated flight calls from four families of passerines, namely: American sparrows, cardinals, thrushes, and New World warblers. The annotator found 2,662 flight calls from 11 different species in total. These flight calls have a duration in the range 50–150 milliseconds and a fundamental frequency in the range 2–10 kHz.

Hyenas (HT, HV): Spotted hyenas are a highly social species that live in "fission-fusion" groups where group members range alone or in smaller subgroups that split and merge over time, using a variety of types of vocalizations to coordinate with one another. Spotted hyena vocalization data were recorded on custom-developed audio tags designed by Mark Johnson and integrated into combined GPS/acoustic collars (Followit Sweden AB) by Frants Jensen and Mark Johnson. Collars were deployed on female hyenas of the Talek West hyena clan at the MSU-Mara Hyena Project (directed by Kay Holekamp) in the Masai Mara, Kenya as part of a multi-species study on communication and collective behavior. Recordings used as part of this task contain a variety of different vocalisations which were identified and classified into types based on the established hyena vocal repertoire [15]. The HT subset of the hyena recordings and their accompanying annotations were used as part of the development set, while the HV subset of recordings and their annotations were used as part of the validation. There is no overlap between the vocalisations annotated in the two sets. Field work was carried out by Kay Holekamp, Andrew Gersick, Frants Jensen, Ariana Strandburg-Peshkin, and Benson Pion; labeling was done by Kenna Lehmann and colleagues.

Meerkats (MT, ME): Meerkats are a highly social mongoose species that live in stable social groups and use a variety of distinct vocalisations to communicate and coordinate with one another. The meerkat vocal repertoire has been well characterized based on previous research, allowing calls to be reliably classified by human labellers [16, 17]. Recordings used in this task were acquired at the Kalahari Meerkat Project (Kuruman River Reserve, South Africa; directed by Marta Manser and Tim Clutton-Brock), as part of a multi-species study on communication and collective behavior. Recordings of the development set (MT) were recorded on small audio devices (TS Market, Edic Mini Tiny+ A77, 8 kHz) integrated into combined GPS/audio collars which were deployed on multiple members of meerkat groups to monitor their movements and vocalisations. Recordings of the evaluation set (ME) were recorded by an observer following a focal meerkat with a Sennheiser ME66 directional microphone (44.1 kHz) from a distance of less than 1 m. Recordings were carried out during daytime hours while meerkats were primarily foraging and include several different call types. Field work was carried out by Ariana Strandburg-Peshkin, Baptiste Averly, Vlad Demartsev, Gabriella Gall, Rebecca Schaefer and Marta Manser. Audio recordings were labeled by Baptiste Averly, Vlad Demartsev, Ariana Strandburg-Peshkin, and colleagues.

Jackdaws (JD): Jackdaws are corvid songbirds that usually breed, forage and sleep in large groups, but form a pair bond with the same partner for life. They produce thousands of vocalisations per day, but many aspects of their vocal behaviour remain unexplored due to the difficulty in recording and assigning vocalisations to specific individuals. In a multi-year field study (Max-Planck-Institute for Ornithology, Seewiesen, Germany), wild jackdaws were equipped with small backpacks containing miniature voice recorders (Edic Mini Tiny A31, TS-Market Ltd., Russia) to investigate the vocal behaviour of individuals interacting with their group and behaving freely in their natural environment. The JD dataset contains a 10-minute on-bird sound recording (22050 Hz) of one male jackdaw during the breeding season 2015. Field work was conducted by Lisa Gill, Magdalena Pelayo van Buuren and Magdalena Maier. Sound files were annotated by Lisa Gill, based on a previously established video-validation in a captive setting [18].

Polish Baltic Sea bird flight calls (PB): The PB dataset consists of six 30 minute recordings of bird flight calls recorded along the Polish Baltic Sea coast. The recordings are the excerpt from Hanna Pamuła's project, focused on the acoustic monitoring of birds migrating at night along the Polish Baltic Sea coast. Three autonomous recording units were used with the same hardware set-

[1] `https://doi.org/10.5281/zenodo.4543504`

[2] `https://doi.org/10.5281/zenodo.4864755`

	Dataset	mic type	# audio files	total duration	# labels (excl. UNK)	# events (excl. UNK)
Development Set: Training	BV	fixed	5	10 hours	11	2,662
	HT	mobile	3	3 hours	3	435
	MT	mobile	2	70 mins	4	1,234
	JD	mobile	1	10 mins	1	355
Development Set: Validation	HV	mobile	2	2 hours	2	50
	PB	fixed	6	3 hours	2	260
Evaluation Set	ME	handheld	2	20 mins	2	70
	ML	various	17	20 mins	17	1,035
	DC	fixed	13	105 mins	3	967

Table 1: Information on each dataset.

tings (Song Meters SM2, Wildlife Acoustics, Inc). They were deployed close to each other (<100m) - near the lake, on the dune, and on the forest clearing - to provide diverse acoustic background. The recordings were acquired during the 2016, 2017 and 2018 fall migration seasons. The passerines night flight calls were annotated by Hanna Pamuła. The PB dataset is part of the development set used for validation. In each recording only one bird species is the target class: song thrush, *Turdus philomelos* (3 recordings); blackbird, *Turdus merula* (3 recordings). Each recording contains 22–93 calls in the 8–400 milliseconds range. The usual fundamental frequency range for calls of the chosen species is 5–9 kHz.

Macaulay Library (ML): The Macaulay Library is a digital archive of images, videos, and sounds from animals.[3] As of 2021, it contains 175k audio recordings from 10k species of birds and 2k species of amphibians, fish, mammals and insects. These recordings are contributed by amateur and professional recordists around the world, and the catalogue is maintained by the Cornell Lab of Ornithology. For the DCASE 2021 challenge, one author (DB) curated 17 recordings from the Macaulay Library and annotated them in terms of animal vocalizations. Each recording contains calls from a different species: 14 terrestrial mammals (not including hyena or meerkat) and 3 birds (not including passeriformes). The average duration of each recording is of the order of one minute and the number of calls per minute varies in the range 10–150.

BIOTOPIA Dawn Chorus (DC): Many bird species produce vocalisations during the entire day, but their vocally most active period by far usually occurs around dawn. This natural phenomenon is called *dawn chorus*. The Dawn Chorus project is a worldwide citizen science and arts project bringing together amateurs and experts to experience and record the dawn chorus at their doorstep.The DC dataset used as part of the evaluation set stems from dawn chorus recordings, made using Zoom H2 recorders at 44100 Hz, at three different locations in Southern Germany (Haspelmoor, Munich's Nymphenburg Schlosspark, and Nantesbuch), by Moritz Hertel and Rudi Schleich. The vocalisations of three target species were annotated by LG (Common cuckoo, *Cuculus canorus*: 6 files, ca. 9 minutes, 543 labels; European robin, *Erithacus rubecula*: 3 files, ca. 43 min, 381 labels; Eurasian wren, *Troglodytes troglodytes*: 3 files, ca. 50 min, 268 labels).

3. BASELINE METHODS

We propose two systems as baselines to measure submitted methods performance with. One is an approach commonly used in bioacoustics based on spectrogram cross-correlation and the other is a deep learning approach based on prototypical networks [11].

3.1. Template Matching

Our first baseline is a spectrogram cross-correlation method, based on scikit-image's `match_template` function that uses fast, normalized cross-correlation to find instances of a template in an image, returning values ranged between -1.0 and 1.0, with higher values corresponding to higher correlation. Our few-shot template matching method computes cross-correlation across the time axis between each of the events (shots) provided for a file and the rest of the recording. A different detection threshold is set for each audio file based on the max value of the cross-correlation results between the shots provided. Peak picking is performed on the results of the template matching algorithm, with any peak above the threshold corresponding to the center of a detected event in that recording. Borders of the predicted event are computed based on the length of the shot it was correlated with. Predictions from all shots of a recording are collapsed into a single binary prediction vector which will produce the final events predicted for the class of interest.

3.2. Prototypical Network

Our second baseline is based on prototypical networks [11]. The goal of prototypical networks and episodic training is to learn a classifier which can adapt quickly to new classes with only a few examples. Each episode of the training is configured as a N*-way-*k*-shot classification*, where N denotes the number of classes and k the number of known samples per class. A mini batch is sampled from the training set and split into a support set consisting of k labelled samples, $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$ where $x_i \in \mathbb{R}^D$ and $y_i \in \{1, 2, \ldots, N\}$ is the corresponding label, with the remaining samples comprising the query set Q. Prototypical networks compute an M-dimensional class prototype $c_n \in \mathbb{R}^M$, through an embedding function $f_\phi : \mathbb{R}^D \rightarrow \mathbb{R}^M$ with learnable parameters ϕ. In our baseline $D = 128$ and $M = 64$.

We compute a prototype for each class as the mean of the embedded support points belonging to it. Then, for each sample x_q from the query set, a distance function is used to calculate the Euclidean distance of x_q from each prototype, following which a softmax function over the distances produces a distribution over the classes. Learning proceeds by minimizing the negative log probability $J(\phi) = -\log p_\phi(y_q = n|x_q)$ over the true class k via stochastic gradient descent.

During evaluation, we adopt a binary classification strategy inspired by [26]. The first 5 positive (POS) annotations are used for calculation of positive class prototype and the rest of the audio file is treated as the negative class, based on the assumption that the positive class is relatively sparse in the recording. We randomly sample from the negative class to calculate the negative prototype. Each

[3] Official website: `https://www.macaulaylibrary.org/`

Rank	Team name	Evaluation set: F-score % (97.5% confidence interval)	Validation set: F-score %	DC F-score %	ME F-score %	ML F-score %
1	Zou_PKU [19]	**38.4** (36.2 - 40.6)	55.3	20.6	68.0	67.3
2	Tang_SHNU [20]	38.3 (36.1 - 40.5)	51.4	25.6	61.5	43.3
3	Anderson_TCD [21]	35.0 (33.1 - 37.0)	26.2	19.9	56.6	56.8
4	Baseline_TempMatch	34.8 (32.6 - 37.1)	2.0	**32.2**	47.1	29.5
5	Cheng_BIT [22]	23.8 (21.9 - 25.7)	46.3	10.6	53.5	**78.8**
6	Baseline_PROTO	20.1 (18.2 - 21.9)	41.5	8.5	**72.7**	55.7
7	Zhang_uestc [23]	16.8 (15.5 - 18.2)	54.4	8.1	45.1	29.9
8	Johannsmeier_OVGU [24]	15.2 (13.7 - 16.7)	**58.6**	6.5	64.3	35.8
9	Bielecki_SMSNG [25]	8.4 (7.1 - 9.7)	51.8	3.1	56.3	51.4

Table 2: F-score results per team on evaluation and validation sets.

query sample is assigned a probability based on the distance from the positive and negative prototype. Onset and offset predictions are made based on thresholding probabilities at a value of 0.5 across the query set. The prediction process for each file is repeated 5 times, with the negative prototype created by random sampling each time. The final prediction probability for each query frame is the average of predictions across all iterations. Finally, post-processing is applied to the outputs in order to remove possible false positives. For each audio file, predicted events with shorter duration than 60% of the duration of the shortest shot provided for that file are removed.

4. EVALUATION AND RESULTS

For the evaluation of this task we employ an event-based F-measure with macro-averaged metric. The main challenge is related to the detection of a match between ground truth events and predicted events. Traditional approaches use onset detection based metrics and fixed-size evaluation windows. Given the great variation between datasets and characteristics of the events we want to detect in this task, these approaches are not suitable. Instead, we use the Intersection over Union (IoU), with 30% minimum overlap to produce a list of possible matches of the predictions. For each ground truth event, a single best match is selected by applying the Hopcroft-Karp-Karzanov algorithm for bipartite graph matching.

In a SED task we can define True Positives (TP) as predicted events that match ground truth events, False Positives (FP) as predicted events that do not match any ground truth events, and False Negatives (FN) as ground truth events that are not predicted. In this task, ground truth events consist of POS events of the class and UNK events that have some uncertainty associated to the assigned class. The procedure we employ is:

1. Apply IoU and bipartite graph matching between predicted events and ground truth POS events only, resulting in TP.
2. Apply IoU and bipartite graph matching between remaining predicted events, that did not match with any POS event, and ground truth UNK events only.
3. Compute FP as the number of predicted events that were not matched to either POS or UNK events.
4. Compute FN as the number of POS ground truth events that were not matched by any predicted event.

This is applied to each dataset in the evaluation set where we compute the F-score metric. The reported results are the harmonic mean over all the datasets, which is appropriate for combining percentage results, and ensures that a system should perform well across all datasets to achieve a strong score.

4.1. Results

DCASE 2021 task 5 had 7 teams participating with a total of 24 submitted systems. F-score results per team are presented in Table 2. All submitted systems adopted prototypical networks. Data augmentation was applied by the majority of the teams with SpecAugment[27] being the most popular choice. All systems rely on some sort of post-processing mechanism designed to removing superfluous predictions and many teams report important improvements in results due to it. Another popular choice was using Perchannel Energy Normalization (PCEN) [28] as acoustic features.

The best ranked system [19] improved over the baseline prototypical approach by applying a transductive inference method, where supplemental information is used to convey more representative prototypes of each category. A mutual learning framework designed to make the feature extraction network more task dependent is also adopted. The system ranked in second place [20] also improved over the prototypical baseline by using additional data from Audioset to train a ResNet for the feature extraction part. They have also adopted embedding propagation (EP) [29], with the objective of smoothing the decision boundaries as a way of increasing the generalisation capabilities of the few-shot system. The third ranking system [21], follows the same approach as the prototypical network baseline, with the main differences being the use of data augmentation and reducing the size of the network. Interestingly, although the results in the validation set are not on par with the other systems, this system outperforms most systems in the evaluation set.

Also of note, the work in [22] uses i-vectors as input features; both submissions in [23] and [24], explicitly create a negative class to model background noise and construct a negative prototype; and in [25], the team opted for combining the prototypical loss, with knowledge distillation and attention transfer loss.

An important observation from Table 2 is the drop in F-score from the validation to the evaluation set for the majority of the systems. This suggests that the systems are generally dataset sensitive. To highlight this aspect further, we report the F-score results per dataset in the evaluation set. Most systems tend to have a decrease in performance on the DC set, comprised of dawn chorus recordings, while perform better on ME and ML that include mainly mammal vocalisations. This leads to the conclusion that very complex environments, such as dawn chorus, need further techniques to be employed for robust SED. Our template matching baseline improved performance from the validation set to the evaluation set. This is mainly due to template matching not being trained over specific recordings, treating each audio file as a unique task without any knowledge about the rest of the set with performance only depends on the templates (shots) used for cross-correlation.

5. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[2] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," 2019.

[3] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European Signal Processing Conference*. IEEE, 2010, pp. 1267–1271.

[4] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[5] D. Stowell, "Computational bioacoustic scene analysis," in *Computational analysis of sound scenes and events*. Springer, 2018, pp. 303–333.

[6] S. Kahl, M. Clapp, W. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga, and A. Joly, "Overview of BirdCLEF 2020: Bird sound recognition in complex acoustic environments," in *CLEF 2020*, 2020.

[7] F. Frazao, B. Padovese, and O. S. Kirsebom, "Workshop report: Detection and classification in marine bioacoustics with deep learning," 2020.

[8] W.-P. Vellinga and R. Planqué, "The xeno-canto collection and its relation to sound recognition and classification," in *CLEF (Working Notes)*, 2015.

[9] A. E. Méndez Méndez, M. Cartwright, and J. P. Bello, "Machine–crowd–expert model for increasing user engagement and annotation quality," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[10] J. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. P. Bello, "Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 901–905.

[11] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *NIPS*, 2017.

[12] Y.-X. Wang, R. B. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7278–7286, 2018.

[13] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *ArXiv*, vol. abs/1803.02999, 2018.

[14] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: A dataset and benchmark for avian flight call detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 266–270.

[15] K. D. S. Lehmann, "Communication and cooperation in silico and nature," Ph.D. dissertation, Michigan State University, 2020.

[16] M. B. Manser, "The evolution of auditory communication in suricates, suricata suricatta," Ph.D. dissertation, University of Cambridge, 1998.

[17] M. B. Manser, D. A. Jansen, B. Graw, L. I. Hollén, C. A. Bousquet, R. D. Furrer, and A. le Roux, "Chapter six - vocal complexity in meerkats and other mongoose species," ser. Advances in the Study of Behavior, M. Naguib, L. Barrett, H. J. Brockmann, S. Healy, J. C. Mitani, T. J. Roper, and L. W. Simmons, Eds. Academic Press, 2014, vol. 46, pp. 281–310. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128002865000067

[18] D. Stowell, E. Benetos, and L. F. Gill, "On-bird sound recordings: automatic acoustic recognition of activities and contexts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1193–1206, 2017.

[19] D. Yang, H. Wang, Z. Ye, and Y. Zou, "Few-shot bioacoustic event detection = a good transductive inference is all you need," DCASE2021 Challenge, Tech. Rep., June 2021.

[20] T. Tang, Y. Liang, and Y. Long, "Two improved architectures based on prototype network for few-shot bioacoustic event detection," DCASE2021 Challenge, Tech. Rep., June 2021.

[21] M. Anderson and N. Harte, "Bioacoustic event detection with prototypical networks and data augmentation," DCASE2021 Challenge, Tech. Rep., June 2021.

[22] H. Cheng, C. Hu, and M. Liu, "Prototypical network for bioacoustic event detection via i-vectors," DCASE2021 Challenge, Tech. Rep., June 2021.

[23] Y. Zhang, J. Wang, D. Zhang, and F. Deng, "Few-shot bioacoustic event detection using prototypical network with background classs," DCASE2021 Challenge, Tech. Rep., June 2021.

[24] J. Johannsmeier and S. Stober, "Few-shot bioacoustic event detection via segmentation using prototypical networks," DCASE2021 Challenge, Tech. Rep., June 2021.

[25] R. Bielecki, "Few-shot bioacoustic event detection with prototypical networks , knowledge distillation and attention transfer loss," DCASE2021 Challenge, Tech. Rep., June 2021.

[26] Y. Wang, J. Salamon, M. Cartwright, N. J. Bryan, and J. P. Bello, "Few-shot drum transcription in polyphonic music," *CoRR*, vol. abs/2008.02791, 2020. [Online]. Available: https://arxiv.org/abs/2008.02791

[27] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[28] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[29] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–138.