Computer audition for continuous rainforest occupancy monitoring: the case of Bornean gibbons' call detection

Panagiotis Tzirakis, Alexander Shiarella, Robert Ewers, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Tzirakis, Panagiotis, Alexander Shiarella, Robert Ewers, and Björn W. Schuller. 2020. "Computer audition for continuous rainforest occupancy monitoring: the case of Bornean gibbons' call detection." In *Interspeech 2020, Shanghai, China, 25-29 October 2020*, edited by Helen Meng, Bo Xu, and Thomas Zheng, 1211–15. ISCA. https://doi.org/10.21437/interspeech.2020-2655.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Computer Audition for Continuous Rainforest Occupancy Monitoring: The Case of Bornean Gibbons' Call Detection

Panagiotis Tzirakis[1*], *Alexander Shiarella*[1*], *Robert Ewers*[1], *Björn W. Schuller*[1,2]

[1] GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[2] EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

panagiotis.tzirakis12@imperial.ac.uk

Abstract

Auditory data is used by ecologists for a variety of purposes, including identifying species ranges, estimating population sizes, and studying behaviour. Autonomous recording units (ARUs) enable auditory data collection over a wider area, and can provide improved consistency over traditional sampling methods. The result is an abundance of audio data – much more than can be analysed by scientists with the appropriate taxonomic skills. In this paper, we address the divide between academic machine learning research on animal vocalisation classifiers, and their application to conservation efforts. As a unique case study, we build a Bornean gibbon call detection system by first manually annotating existing data, and then comparing audio analysis tool kits including end-to-end and bag-of-audio-word modelling. Finally, we propose a deep architecture that outperforms the other approaches with respect to unweighted average recall. The code is available at: https://github.com/glam-imperial/Bornean-Gibbons-Call-Detection

Index Terms: Primate vocalisations, animal vocalisations, Bornean gibbon detection, bioacoustics, deep learning

1. Introduction

Bioacoustic data is used by ecologists to study a wide range of taxonomic groups, including birds, dogs, insects, amphibians, primates, bats, and cetaceans [1, 2, 3, 4, 5]. Sound data is used to infer species distributions, population abundance, and movement patterns, and to gain information on life cycles, mating behaviours, and invasive species. It can even be related to environmental characteristics like temperature [6]. The resulting insights inform not only the management of vocal species, but also their ecosystems, with soundscape change shown to act as an effective early indicator of ecosystem disturbance [7, 8, 9]. This information enables scientists to understand and model ecosystems at various scales, and informs how policy makers and conservationists manage species [10].

Despite its importance, baseline data on ecosystems is largely constrained by current collection and processing tools [11]. For terrestrial animal monitoring specifically, the most common survey methods continue to be transect/plot sampling and capture-mark-recapture methods [12]. Traditional methodologies, like these, require that scientists or technicians be physically present to collect information visually, auditorily, or via close-range sensing tools. Hence, the amount of data collected is restricted by the amount of resources they are able to commit to fieldwork.

*Equal contribution.

Together with motion sensing cameras, remote sensing drones, wireless sensor networks, and environmental DNA collection, autonomous recording units (ARUs) are part of a set of technology ecologists are applying to address the limitations of traditional field data collection [11]. Advances in integrating solar-power and network technology into ARUs suggests the potential to make passive acoustic monitoring (PAM) even more pervasive, reducing the need for regular visits to replace batteries or collect data stored in memory [13]. Recorders are able to monitor consistently for long periods, and can even be placed in remote or dangerous locations by helicopter or drone. Additionally, the presence of an ARU is less likely to affect animal behaviour than a human observer.

While recording devices (targeted and autonomous) provide an abundance of both passively and actively collected data, the utility of that information is limited without automated processing. Until recently, automatic quantification/filtering of ARU data was mostly used for soundscape analysis and species that vocalise at unique frequency ranges (e. g., whales and bats). However, advances in machine learning have made finer-scale data processing possible and applicable to a wider variety of animals.

In this paper, we aim to examine and address limitations in model creation by building a Bornean gibbon call detection system. Part of our motivation for studying gibbons stems from the fact that they share a number of voice perception abilities as humans [14, 15, 16]. We build our system for the Stability of Altered Forest Ecosystems (SAFE) Project, as its solar-powered, mobile connected ARU network provides a unique opportunity to use machine learning for large-scale, continuous ecological monitoring [13]. To this end, we develop a graphical user interface (GUI) tool for rapidly annotating vocalisation data. For call detection, we propose a new convolutional recurrent neural network (RNN) architecture that uses a Mel-spectrogram as input. We compare our method with two recently proposed toolkits for audio representation and analysis, namely the End2You [17] and the openXBOW [18] toolkits. The End2You toolkit is an end-to-end method using the raw waveform [19, 20], and the openXBOW creates a bag-of-audio-words representation using hand-crafted audio features such as low-level descriptor (LLD) features as input. Our proposed architecture outperforms in terms of Unweighted Average Recall (UAR) the other two methods.

2. Related Work

To detect animal sounds, researchers often split audio into short segments and treat each segment as a binary classification prob-

http://dx.doi.org/10.21437/Interspeech.2020-2655

lem. The 2017 and 2018 DCASE bird challenges [21] define bird call detection as just that, requesting entries that detect bird calls, using short segments of standard length, and a binary label indicating the presence or absence of a bird call [22]. This method has been successfully applied in ecosystem monitoring [23, 24].

In studies where individual calls need to be differentiated, a binary classifier will not suffice. Options for finer-scale temporal or spectral labelling include (in increasing complexity) detecting event onsets, monophonic segmentation, polyphonic segmentation, time-frequency boxes, and time-frequency blobs or sinusoids. Stowell et al. [22] summarise these output formats, along with their advantages and disadvantages.

The primary difficulty of these more complex methods is the time required for producing accurately labelled training and testing data. Some studies have attempted to address this issue by using crowd/citizen science, though such methods produce labels with varied accuracy [25]. Alternatively, Fanioudakis and Potamitis [26] show the potential for deep autoencoders (U-nets) to provide location information on weakly labelled data, which could be used as a pre-processing method for training.

The lack of large, even weakly labelled datasets, limits the ability for researchers to use supervised machine learning methods for ecological acoustic monitoring. High performing audio classification methods like convolution neural networks (CNNs) require a sizeable volume of data to train without overfitting. Moreover, in ecological monitoring, an increased amount of data may be needed to account for soundscape variation. Often, field recorders can differ in the quality of the audio they collect, and the vocalisations of target species may be distorted by distance, other animal sounds, or significant environmental noise, such as wind and rain.

3. Dataset

The data used in this study were collected for the purposes of the SAFE Project. [1] The project's site spans over 8 000 hectares of land with 12 recorders in Malaysian Borneo, including Virgin Jungle Reserve (VJR), logged and fragmented forest, and palm oil plantation. For our purposes, we use machine learning approaches to detect Bornean gibbon (*Hylobates muelleri*) calls in the SAFE Project's ARU data. The Bornean gibbon is one of 18 species of ape in the family Hylobatidae. While the SAFE Project audio data contains vocalisations from a wide range of taxonomic groups, we chose this species because:

- Gibbon calls are loud and distinct, making their identification in audio easier for a non-expert surveyor.
- Gibbon calls are variable, allowing this study to test model generalisability across call types.
- Despite considerable resources dedicated to monitoring primates, little research is done on the automatic classification of primate vocalisations.
- As with most gibbon species, Bornean gibbons are endangered, with a 50 percent decrease in population over the past three generations.
- Gibbon vocalisation tends to follow circadian patterns, with male/female duets occurring regularly in the morning. This characteristic makes developing a training dataset from unlabelled audio easier.

[1] https://www.safeproject.net

3.1. Annotation

There are several approaches to annotating animal vocalisations such as presence/absence, polyphonic segmentation, time-frequency blobs, time-frequency boxes etc [22]. While some of these, like bounding-box annotations, provide greater resolution, most studies label fixed-sized clips based on presence or absence. This simpler approach has the benefit of decreasing labelling time and requiring less precision in annotation, the latter of which can be difficult for soft calls, noisy environments, and non-expert listeners.

The behavioural ecology of gibbons makes higher resolution annotation largely impractical. Gibbons tend to travel in large groups, vocalise at the same time, and display a variety of call lengths. Localising single calls, such as by using a spectrogram bounding-box method, would likely be subjective and inaccurate. Therefore, in this study, we focused on detecting gibbon occupancy rather than quantifying call abundance.

To this end, we create a GUI tool, where a user can load audio files into a media player (Fig. 1). Similar to doing an avian point count survey, as audio plays, the annotator can add millisecond timestamps to a survey record when hearing vocalisations. Keyboard shortcuts can be used to include custom labels, allowing timestamps to be differentiated by call type and volume. This feature proved to be helpful in later clip extraction steps and for more nuanced analysis of model performance. It could also eventually provide a means to label multiple species simultaneously. Survey records are saved in a relational database and can be edited, exported, and used to extract training data.

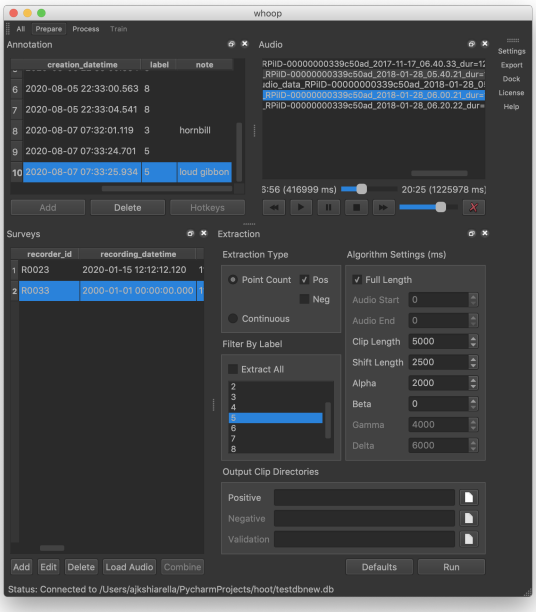

Figure 1: *Proposed GUI tool for manually annotating audio clips.*

The software converts longer audio to training examples using custom parameters for clip length and overlap. To account for both human and computer lag times, we implement an algorithm to extract positive and negative examples with high certainty. This method has a number of benefits, allowing for dy-

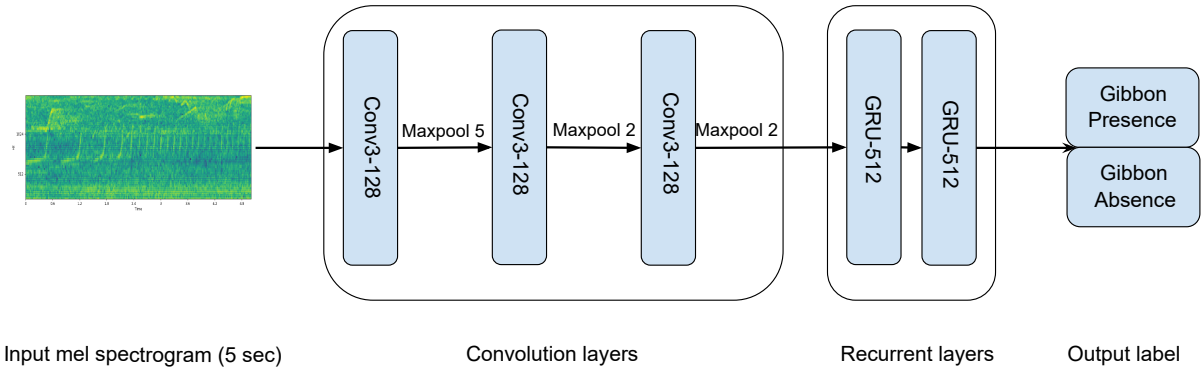

Figure 2: *Proposed deep neural network architecture. The input representation is passed through a convolution neural network before passing to a recurrent neural network with gated recurrent units for the final prediction.*

namic clip size/overlap, varied call location within clips, and continuous (often context-dependent) annotation. However, to assess performance on non-curated data, the validation and test sets were annotated manually for presence/absence on continuous audio split into fixed-sized clips.

We used our tool to annotate and extract gibbon call data. In total, 4 annotators were used. We kept the audio clips where all annotators agree that a gibbon call is present or absent. Each clip was segmented with a window length of 5 s and a percentage overlap of 50 %. In order to create a dataset such that the trained model would be able to generalise to unseen audio clips, we consider the following characteristics for the audio clips: (a) *recorder location*: recorders are placed in a wide range of locations, including VJR, logged forest, palm plantation, and riparian easements, (b) *audio noise*: for example, weather changes and temporal soundscape variation, and (c) *gibbon vocalisation characteristics*: ranging from short, high-pitched vocalisations to the long female great call.

The final dataset contains a total of 19 370 training set examples, 848 validation set examples, and 1 680 test set examples. The training set contains 9 569 positive and 9 801 negative, 5 s audio clips. Recordings were taken from 4 recorders in areas with similar noise characteristics, ranging from one hectare logged forest to VJR, with recording months spanning November through May. The validation and test sets contain clips from a different area, different recorder, and different morning than the training set. The validation set contains 431 positive and 416 negative clips, whereas the test set contains 884 negative clips and 796 positive clips. We plan to release the data within an Interspeech challenge framework.

4. Model Design

Traditional machine learning approaches first extract features from the raw waveform in order to suppress the background noise, and at the same time, reduce the dimensionality of the input. In the deep learning era, a key operation using CNN is convolution, which in the discrete domain can be defined as follows:

$$(f * x)[n,t] = \sum_{l=-N}^{N} \sum_{k=-T}^{T} f[n,t] \cdot x[n-l, t-k], \qquad (1)$$

where $f[n,t]$ indicates a kernel function, which in our case operates on the raw signal $x[n,t]$, with N frequency bins, and T time steps.

We use CNNs to model spatio-temporal characteristics of the input representation, and the max-pooling operation to reduce the dimensionality of the extracted representation while preserving the necessary statistics of the convolved signal.

While 2D convolutions extract spatiotemporal features by removing background noise, we also utilise RNN models to capture the temporal dynamics in the signal. For our purposes, we use gated recurrent units (GRU) as they have few parameters and fast convergence of the optimisation.

4.1. Proposed model

Our model, which aims to learn the extracted features and the classification step in one jointly trained model, is depicted in Fig. 2 and described below.

Input. First we choose the input representation. Since gibbons share highly similar vocal properties as humans, we extract Mel-spectrograms from the raw audio signal. We utilise 128 Mel-bands, over a 5 sec window, yielding an input representation of size of $128 \times 431 \times 1$.

Convolution layer. We utilise 128 time finite impulse filters of kernel size 3×3 and stride 1×1, with rectified linear unit (ReLU) activation, to extract features using the input representation.

Max-pooling layer. We apply a max-pooling layer of size of 5×1 to decrease the frame rate of the signal and keep the most activated features.

Convolution layer. As in our first convolution layer, the kernel size is 3×3 and stride 1×1, with ReLU activation, to extract a higher-level of abstraction.

Max-pooling. We apply max-pooling of a size of 2 to reduce the dimensionality of the features. In contrast to the first max-pooling layer where we use a large window, in this layer we use a smaller window size.

Convolution layer. We utilise 2D convolutions of kernel size 3×3 and stride 1×1, with ReLU activation and 128 filters.

Max-pooling layer. Our final feature representation is the output of a max-pooling layer of kernel size 2×1 with stride 1×1 for the final output.

Batch normalisation. Due to the high number of parameters our model contains, we use batch normalisation [27] as regularisation after each convolution layer.

Recurrent neural network. Finally, 2-GRU layers are used before the final prediction with 512 hidden units.

5. Experiments

5.1. Experimental Setup

To train the models, we utilised the Adam optimisation method [28], and a fixed learning rate of 10^{-6} throughout all experiments. We used a mini-batch of 25 samples and a gradient norm clipping of 5.0. As discussed earlier, our proposed model uses batch normalisation [27] to regularise our network, such that it will not overfit. To train our models, we use binary cross-entropy loss. Our input representation is computed using a Hann window with the length of the FFT to be 2 048, and a hop length of 512, after subtracting the mean of the raw waveform with 16 kHz sampling rate. Finally, we apply data augmentation to train the models by randomly creating 5 sec long windows with a gibbon call.

5.2. Machine Learning Approaches

We compare our method with End2You (Sec. 5.2.1) and openXBOW (Sec. 5.2.2) which are comparably new toolkits used in the literature such as in the Interspeech Computational Paralinguistics Challenge series, and produce competitive results in several domains such as emotion recognition [19, 29, 30, 31], and others [32, 33, 34].

5.2.1. End2You

End2You is an open-source toolkit implemented in Python which provides capabilities to train and evaluate audio (and other) models in an end-to-end manner, i. e., using raw input. The audio processing model is comprised of two blocks, each one containing a convolution with 40 filters and a max-pooling operation, where the first block is applied in the time domain with size and stride 2, and the second block is applied to the feature maps with size and stride 10. On top of the convolution network, a 2-layer GRU block with 512 units is applied such that the temporal information in the data can be considered. To train the model, we used the binary cross-entropy loss. We performed hyper-parameter optimisation and we show the best results obtained on the validation set.

5.2.2. openXBOW

openXBOW is an open-source crossmodel bag-of-words toolkit, written in Java, to extract a bag of words model from the low-level descriptor (LLD) features (e. g. MFCC). In particular, each LLD feature vector is treated as a point in a hyper-dimensional space. Then, a codebook is extracted either through k-means clustering or random sampling. LLD vectors can be quantised to this codebook as 'audio-words' and the count of these audio-words across frames is used to create a bag-of-words representation. For our purposes, we use openSmile [35, 34, 33] and its pre-defined ComParE 2010 feature set as LLD features, hence fostering reproducibility. We experimented with several classifiers, namely, Naive Bayes, linear Support Vector Machines (SVMs), Random Forest, AdaBoost, and Nearest Neighbour. Basic hyper-parameter optimisation was performed for all classifiers. We found that Random Forest provides the best results on the validation set.

5.3. Results

We compare our proposed model with openSMILE plus openXBOW and End2You on the created dataset. To evaluate the model's performance, we employ the frequently used metric UAR, i. e., the sum of classwise recall divided by the number of classes, for audio analysis.

Results are depicted in Table 1. Our model outperforms the other two on both the validation and test sets. For the test set, our model outperforms End2You with 13 % absolute value and openXBOW with 9 % absolute value. The results are significant with a level of significance of 0.05 in a one-sided z-test. Finally, we should note that both End2You and OpenXBOW produce highly generalisable models, but so does our model with high validation and test scores.

The high performance of our approach shows the potential for machine learning algorithms to transform ecological monitoring, even in dynamic and biodiverse environments. In particular, our method can reduce human data processing times by filtering out data unlikely to contain vocalisations or by sorting clips using prediction confidence to quickly confirm occupancy at a lower temporal resolution.

Table 1: *Results (w. r. t. UAR) on the validation and test sets for End2You, openSMILE + openXBOW, and the proposed approach.*

UAR [%]	Validation	Test
End2You	78.5	80.6
OpenXBOW	82.7	84.8
Proposed	97.1	**93.3**

6. Conclusion

We proposed applying recent advances in intelligent audio analysis to help automate autonomous recording units' data processing in the wild. While a larger variety of labelled data is needed to verify performance across recordings and animal species, the suggested model's performance demonstrates the feasibility of using deep learning for continuous primate occupancy monitoring in a challenging real-world scenario. Furthermore, the audio annotation and extraction pipeline developed can be used to train bespoke call detection systems for other species and study sites. This paper demonstrates the importance of integrating ecological motivations with computer science perspectives throughout dataset curation, model training, and testing phases of building a machine learning tool.

For future work, we plan to incorporate additional species in our workflow and train our model to simultaneously detect and predict animal sounds. To ensure reproducibility and comparability, we further aim to release the data within a gibbon call detection Interspeech challenge event. Lastly, we intend to use transfer learning methods, which have been used for numerous audio processing applications and can be helpful when using small datasets. In particular, we will try to apply the image feature extraction abilities of a model like VGGNet to spectrogram features, and then feed those features to recurrent components.

7. Acknowledgements

The support of the EPSRC Center for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged.

8. References

[1] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack, "Estimating animal population density using passive acoustics," *Biological Reviews*, pp. 287–309, 2013.

[2] S. Hantke, N. Cummins, and B. Schuller, "What is my Dog trying to tell me? The Automatic Recognition of the Context and Perceived Emotion of Dog Barks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5134–5138.

[3] B. W. Schuller, A. Batliner, C. Bergler, F. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristià, A. Seidl, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Proc. INTERSPEECH*, 2019.

[4] X. Dong, N. Yan, and Y. Wei, "Insect sound recognition based on convolutional neural network," in *Proc. International Conference on Image, Vision and Computing (ICIVC)*, 2018, pp. 855–859.

[5] V. A. Tatsis, C. Tjortjis, and P. Tzirakis, "Evaluating data mining algorithms using molecular dynamics trajectories," *International journal of data mining and bioinformatics*, pp. 169–187, 2013.

[6] A. Frankel, C. Clark, L. Herman, and C. Gabriele, "Spatial distribution, habitat utilization, and social interactions of humpback whales, megaptera novaeangliae, off hawai'i, determined using acoustic and visual techniques," *Canadian Journal of Zoology*, pp. 1134–1146, 1995.

[7] J. Sueur, A. Farina, A. Gasc, N. Pieretti, and S. Pavoine, "Acoustic indices for biodiversity assessment and landscape investigation," *Acta Acustica united with Acustica*, pp. 772–781, 2014.

[8] H. Slabbekoorn, "Songs of the city: noise-dependent spectral plasticity in the acoustic phenotype of urban birds," *Animal Behaviour*, pp. 1089–1099, 2013.

[9] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, "What is soundscape ecology? an introduction and overview of an emerging new science," *Landscape ecology*, pp. 1213–1232, 2011.

[10] I. Vaughan and S. Ormerod, "Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data," *Conservation biology*, pp. 1601–1611, 2003.

[11] W. Turner, "Sensing biodiversity," *Science*, pp. 301–302, 2014.

[12] S. Heinicke, A. K. Kalan, O. J. Wagner, R. Mundry, H. Lukashevich, and H. S. Kühl, "Assessing the performance of a semi-automated acoustic monitoring system for primates," *Methods in Ecology and Evolution*, pp. 753–763, 2015.

[13] S. S. Sethi, R. M. Ewers, N. S. Jones, C. D. L. Orme, and L. Picinali, "Robust, real-time and autonomous monitoring of ecosystems with an open, low-cost, networked device," *Methods in Ecology and Evolution*, pp. 2383–2387, 2018.

[14] P. Belin, "Voice processing in human and non-human primates," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, pp. 2091–2107, 2006.

[15] R. Adolphs, "The social brain: neural basis of social knowledge," *Annual review of psychology*, pp. 693–716, 2009.

[16] C. I. Petkov, C. Kayser, T. Steudel, K. Whittingstall, M. Augath, and N. K. Logothetis, "A voice region in the monkey brain," *Nature neuroscience*, pp. 367–374, 2008.

[17] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2You – The Imperial Toolkit for Multimodal Profiling by End-to-End Learning," *arXiv preprint arXiv:1802.01115*, 2018.

[18] M. Schmitt and B. Schuller, "openXBOW: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, pp. 3370–3374, 2017.

[19] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[20] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1301–1309, 2017.

[21] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, pp. 368–380, 2019.

[22] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *Proceedings 26th IEEE international workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[23] A. F. Raine, M. Boone, M. McKown, and N. Holmes, "The breeding phenology and distribution of the band-rumped storm-petrel oceanodroma castro on kaua'i and lehua islet, hawaiian islands," *Marine Ornithology*, vol. 45, pp. 73–82, 2017.

[24] D. A. Miller, L. A. Weir, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and T. R. Simons, "Experimental investigation of false positive errors in auditory species occurrence surveys," *Ecological Applications*, vol. 22, pp. 1665–1674, 2012.

[25] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8769–8778.

[26] L. Fanioudakis and I. Potamitis, "Deep networks tag the location of bird vocalisations on audio spectrograms," *arXiv preprint arXiv:1711.04347*, 2017.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter *et al.*, "Muse 2020–the first international multimodal sentiment analysis in real-life media challenge and workshop," *arXiv preprint arXiv:2004.14858*, 2020.

[30] P. Tzirakis, S. Zafeiriou, and B. Schuller, "Real-world automatic continuous affect recognition from audiovisual signals," in *Multimodal Behavior Analysis in the Wild*, 2019, pp. 387–406.

[31] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 907–929, 2019.

[32] S. Hantke, M. Schmitt, P. Tzirakis, and B. Schuller, "Eat- the icmi 2018 eating analysis and tracking challenge," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 559–563.

[33] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[34] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," *Proc. Interspeech 2018*, pp. 122–126, 2018.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.