forum acusticum 2026
8-12 September · Graz, Austria

DISENTANGLING ANIMAL VOCALISATIONS AND ANTHROPOGENIC SOUND TO ASSESS NOISE EFFECTS ON BIODIVERSITY

Dick Botteldooren[1], Hannes Beheydt[1], Ben De Meurichy[1], Vinicius Donisete Lima Rodrigues Goulart[2]

[1] *Noise and Health, institute of Biomedical Engineering, Ghent University, Belgium*
[2] *Federal University of Minas Gerais, Belo Horizonte, Brazil*

Abstract

Passive acoustic monitoring (PAM) has become a widely used tool for biodiversity assessment, particularly for monitoring vocalising animal species. However, in soundscapes dominated by natural or urban background noise, or in the presence of human voices, reliable detection and classification of animal vocalisations becomes challenging for widely used pre-trained classifiers, such as BirdNET, and even expert auditory analysis. In contrast, when assessing the impact of anthropogenic noise, such as the noise of high-altitude aircrafts, intense biological activity —such as bird dawn chorus— can interfere with monitoring.

To address these challenges, we investigate artificial intelligence–based sound separation as a preprocessing step for both biodiversity monitoring and environmental noise assessment. Two approaches are explored. The first approach, inspired by speech enhancement techniques, employs complex ideal ratio masks trained with a U Net architecture to separate bird vocalisations from natural or urban background sounds. The results show that this effectively improves bird sound separation and improves the accuracy of automatic classification.

The second approach is based on a recently proposed transformer based Task-aware Unified Source Separation (TUSS), which allows token controlled sound extraction. By fine tuning the pre-trained model for task specific targets, such as aircraft noise or animal vocalisations, the system accuracy increased.

Keywords: *Passive Acoustic monitoring, biodiversity, disentangling, anthropogenic sound*

***Corresponding author**: dick.botteldooren@ugent.be.*

1. Introduction

Assessing animal biodiversity typically involves labour intensive field observations. The emergence of low-cost sensor technologies has made it possible to expand monitoring in both space and time without proportionally increasing human effort. For vocalising species, passive acoustic monitoring (PAM) offers such a solution. Compared to camera traps, PAM has the benefit of functioning well under challenging visual conditions, for example, in dense forests, low-light environments, or when animals are simply too small to be reliably detected visually. Traditionally, PAM has relied on acoustic indices such as the acoustic complexity index (ACI), bioacoustic index (BIO), acoustic diversity index (ADI) and the normalised difference soundscape index (NDSI). However, these biodiversity indices are strongly affected by traffic noise [1], [2], which does not necessarily reflect actual changes in biodiversity.

With the increasing power of artificial intelligence (AI), PAM methods have been improved to identify which species are present. Tools such as BirdNet [3], birdMAE [4], and SoundProtoPNet [5] can now be used to detect and classify bird vocalisations. The underlying AI models used by these methods are quite heterogeneous, and their resilience to interference from anthropogenic noise may vary. In addition, the typical acoustic properties of vocalisations differ between species, which can lead to varying degrees of susceptibility of species detection to masking by anthropogenic sound, such as road traffic noise.

Thus, when assessing the effect of anthropogenic noise on biodiversity through PAM, it should be ensured that PAM itself is not affected by this noise. In this contribution, we assess two approaches for separating animal vocalisations from anthropogenic sound prior to biodiversity assessment. The first approach is inspired by the use of complex ideal ratio masks (cIRM) in speech recognition [6]: a mask is predicted that

when applied to a spectrogram improves speech recognition both by machines and humans (after inverse transform to the time signal). The second approach attempts to separate the recorded sound stream in its components using a transformer architecture. Here we opt for a recently proposed generic model after comparing several alternatives: Task-aware Unified Source Separation (TUSS) [7].

2. Separating Animal Vocalisations using Complex Ideal Ratio Mask (cIRM)

In earlier work, the cIRM technique was extended to extract sound events from background noise [8]. This methodology is adapted here to extract a specific type of event: animal vocalisation. The cIRM estimates and ideal filter to extract the vocalisation spectrogram from the mixture. To this end, a U-Net [9] convolutional neural network architecture is trained. The U-Net combines convolutional layers for encoding to a more abstract representation and deconvolution to reconstruct while so-called skip connections assure that fine detail is maintained. The target mask that is used for training is derived by comparing the spectrogram of the target to the spectrogram of the mixture. Therefore it depends on characteristics of backgrounds as well as on characteristics of the target sounds. The U-Net used in this work contains 23 convolutional layers organised in six encoder-decoder layers with skip connections. All hidden layers use 64 filters, allowing skip connections via addition rather than concatenation. Each convolution layer (except the output) is followed by batch normalisation and ELU activation. The total parameter count is approximately 2.05 million.

2.1. Datasets and training

Bird vocalisation recordings were downloaded from Xeno-Canto [10], the largest open database of bird audio. Only recordings in the highest Xeno-Canto quality class that also scored above 0.5 confidence in BirdNET [3] bird sound recognition, were retained. Because Xeno-Canto recordings contain residual background noise, a stationary spectral gating step [11] was applied to the foregrounds before training. This resulted in 20 000 vocalisations.

Two sets of background sounds were derived from Freesound [12] using keyword searches and automated background filtering [8] (minimum loudness 50 dBFS; RMS standard deviation ¡ 4 dB; no peaks more than 8 dB above mean). After manual verification, a large-scale natural set of 30 000 segments were assembled. An anthropogenic set of 3500 segments (street ambience, traffic, cafes, train stations, etc.) was collected using 37 targeted keywords.

Training was performed with a classical Adam optimiser (learning rate = 0.001, decay 0.97/epoch) with mean squared error for mixtures of bird vocalisation and backgrounds with signal to noise ratio (SNR) ranging from -20 dB to + 20 dB. Note that SNR is defined here as the ratio of the 95 percentile of absolute level of bird sound to the RMS (root-means-square) value of background. This resulted in two models, the *Nature* model based on natural backgrounds and the *Urban* model based on anthropogenic backgrounds.

3. Extracting Animal Vocalisations and Aircraft Noise using TUSS

TUSS (Task-aware Unified Source Separation) [7] is a model that is designed for universal sound separation based on learned task conditioning via prompt embeddings. TUSS is a time-frequency domain model that uses a TF-Locoformer [13] backbone for feature extraction and separation. The latter combines the success of transformers for combining different time scales with the local resolution provided by convolution layers.

The key distinguishing design of TUSS is its use of N learnable prompt embeddings $\{\mathbf{P}_n\}_{n=1}^{N}$ that condition the model's behaviour at inference time, enabling it to handle multiple, potentially contradictory separation tasks with a single set of weights. The prompts are prepended to compact representation $\mathbf{Z}$ in frequency time that groups spectra to 64 auditory bands, along the temporal axis to form an augmented sequence $\mathbf{Z}' = [\mathbf{P}, \mathbf{Z}]$, which is then processed by a cross-prompt module consisting of four TF-Locoformer blocks. Self-attention over this joint prompt-plus-mixture sequence contextualises each prompt against the mixture and vice versa. After the cross-prompt module, the updated prompt features $\tilde{\mathbf{P}}_n$ are split out and each separately gates the updated mixture feature $\tilde{\mathbf{Z}}$ via element-wise multiplication, directing the network's attention towards each target source independently. These per-source features are further refined by two shared TF-Locoformer blocks in the conditional TSE module, before a shared band-wise decoder converts each $\tilde{\mathbf{Z}}_n$ back to a waveform via inverse STFT, producing N separated signals $\hat{s}$. The pretrained base model supports 8 general-purpose prompts (*speech*, *sfx* (sound effects), *sfx-mix*, *drums*, *bass*, *vocals*, *other instruments*, *music-mix*), which indicates a need for fine-tuning for the environmental sound tasks at hand.

3.1. Datasets and Fine-Tuning

Bird vocalisations are a primary class of interest (COI) hence a new label was introduced in TUSS. Weights are initialised with the *sfx* class weights and fine-tuning is done on the BirdSet database [14]. This database is particularly useful as it contains recordings of single bird song with limited background.

As a second COI for the separation we focus on airplane noise as this is a frequent temporally focussed disturbance of natural areas. For fine tuning the AeroSonicDB [15] is used. a dataset created for recognition of low flying aircraft. Weights are initialised with the *sfx* class.

Background sound has been curated from Freesound as before but without differentiating between natural and anthropogenic backgrounds. For background sound,

weights are initialised from the *sfx-mix* class in the pretrained TUSS.
The loss function used for fine tuning is the scale invariant signal to noise ration SI-SNR, defined as

$$\mathrm{SI-SNR}(\hat{s}, s) = 10 \log_{10} \frac{\left\| \frac{\langle \hat{s}, s \rangle}{\|s\|^2} s \right\|^2}{\left\| \hat{s} - \frac{\langle \hat{s}, s \rangle}{\|s\|^2} s \right\|^2} \qquad (1)$$

with $\hat{s}$ the estimated signal and s its ground truth. To assure that the COI has enough focus, the loss for COI and background are separately weighted giving COI a higher weight. This illustrates that the method works, but it is fair to mention that this is a fairly easy example.

4. Results on artificial mixtures

In this section, the performance of the disentangling methods on artificial mixtures of independent test sounds from the same datasets as used for training, is evaluated.

4.1. Illustrations of sound separation

Reconstruction of an original bird vocalisation by cIRM is evident from the illustrated in Fig. 1.

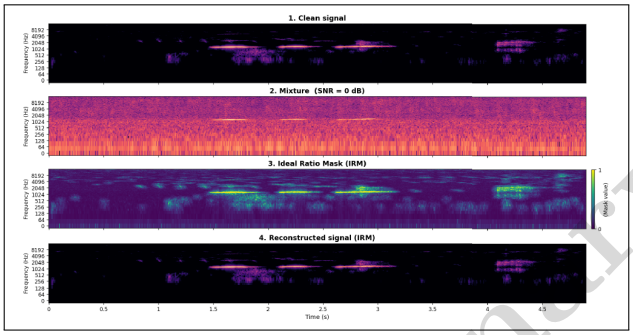

Figure 1: From top to bottom: spectrogram of Xeno-Canto recording XC427277: Undulated Tinamou (*Crypturellus undulatus*); mixture with background; cIRM; reconstructed background-free recording.

The inverse problem of removing animal vocalisation from the measurement of anthropogenic disturbance is handled with TUSS. Fig. 2 shows that the TUSS model with target label *airplane* removes some of the occasional bird tweets.

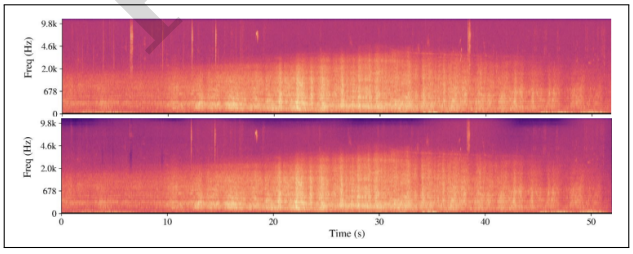

Figure 2: Spectrogram plane pass-by with occasional bird sounds which are eliminated using TUSS in the lower pane.

4.2. Improvement of BirdNet recognition on artificial mixtures

For assessing the impact of anthropogenic sound in animal vocalisation, bird recognition software such as BirdNet might be used. Hence it is useful to explore how adding anthropogenic noise to recordings deteriorates recognition and whether the cIRM can reconstruct this. For this we created mixtures between Xeno-Canto sounds (same species but other recordings than used for training) and the Freesound backgrounds. Xeno-Canto recordings have species labels and these are used as ground truth for the recognition. Recall and falso positives per recording are exchangeable by choosing a confidence threshold for BirdNet classification. Based on a balance between both performance measures, 0.3 chosen here. Fig. 3 shows that for the clean sample (no background added) recogition by BirdNet does not perfectly match the Xeno-Canto label resulting in on average 80% recall and 0.7 false positives per recording. It should however be noted that this is not necessarily due to the accuracy of BirdNet: Xeno-Canto labels are given by contributors and may be faulty but also the recordings can contain secondary vocalisations (e.g. towards the end in the example of Fig. 1).
With decreasing SNR recall drastically drops and neither the *cIRM Nature* nor *cIRM Urban* result in significant overall improvement. False positives increase by adding background and surprisingly this does not depend on the SNR which lead to the hypothesis that the sounds labelled "background" from FreeSound may contain some bird vocalisations that BirdNet picks up, even at low levels. *cIRM Nature* seems to eliminate these false positives as it is trained on natural backgrounds that contain such low level, distal, bird vocalisations. *cIRM Urban* has a similar effect for higher SNR.

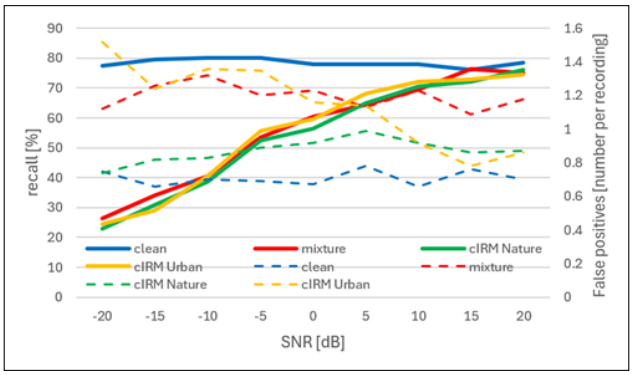

Figure 3: Effect of anthropogenic sound on bird song recognition. Solid line: recall on left axis; dashed line: false positives per recording on right axis

To further explore the limited success of cIRM as a pre-processing step for BirdNet recognition, the test set is split by bird family and improvement in recall is plotted in Fig. 4. It becomes clear that the improvement is strongly dependent on bird families:

for sognbirds (*Passeriformes*), a large family of birds, recall is decreased probably because their song deviates in frequency and temporal pattern from that of most anthropogenic sound and thus BirdNet recognises them easily; the largest benefit (12% increase in recall) of the mask is seen for *Apodiformes* (swifts, treeswifts, and hummingbirds) that vocalise at high frequencies often above 10 kHz and *Pelecaniformes* (such as pelicans, herons, and ibises) produce lower frequency vocalisations with broadband characteristics and limited modulation; Significant improvement is also observed for *Piciformes* (woodpeckers, barbets, and toucans), *Galliformes* (such as a pheasant, quail, or grouse), *Coraciiformes* (kingfishers, bee-eaters, and rollers), and *Accipitriformes* (hawks, eagles, and kites); somewhat surprisingly, cIRM does not improve recognition of *Anseriformes* (waterfowl such as ducks, geese, and swans) although their vocalisation spectrum overlaps with anthropogenic sounds. These observations should however be interpreted with care as the distribution of species in each family contained in the test set is limited and random.

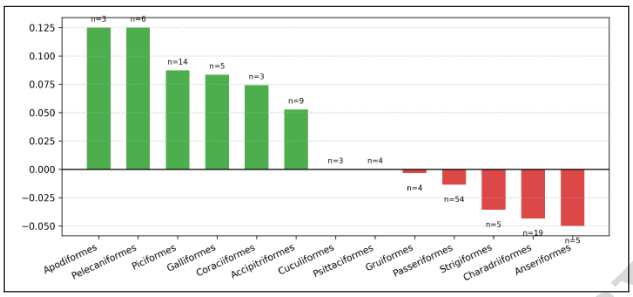

Figure 4: Effect of the cIRM pre-processing on sound recognition (recall) by BirdNet by bird family, for mixtures with anthropogenic sound with SNR between 0 and 20 dB

4.3. Joint separation of bioacoustics and disturbance

Using artificial mixtures of backgrounds, aircraft noise, and bird song with signal to noise ratio's ranging from -10 to 10 dB, the applicability of TUSS for detecting both the disturbance and using bird vocalisation for impact assessment is investigated. The model trained for both tasks increases the SI-SNR for aeroplane sounds by 6.4 dB and that for Bird sound by 13.9 dB. For comparison, humans are assumed to be able to obtain 10 dB by attentive listening. Similarly the signal to distortion ration (SDR) increases by 7.1 and 17.4 dB for aeroplane noise and bird song respectively. The fine-tuned TUSS model thus performs better for bird song than for aeroplane sound separation.

To assess the impact on the performance of a downstream classifier two models for general environmental sound classification: PANN [16] and AST [17] will be tested on the aircraft signal. For bird recognition, BirdMAE [4] and AudioProtoPNet [5] are chosen over BirdNet as they are based on the large BirdSet dataset also used for training TUSS. Fig. 5 shows that AST performs very well for aeroplane recognition and that its performance is hardly affected by mixing the aeroplane noise with other sounds. Separation using TUSS destroys this performance. PANN on the other hand performs much less and its performance is negatively affected by mixing the aeroplane sound recordings with other sound. Separating in this case restores the reduction in performance on mixtures partly. A possible reason for the lower performance of PANN on aeroplane sound is its limited temporal scope compared to the transformer-based AST, and hence it has to rely on pre-separation by TUSS.

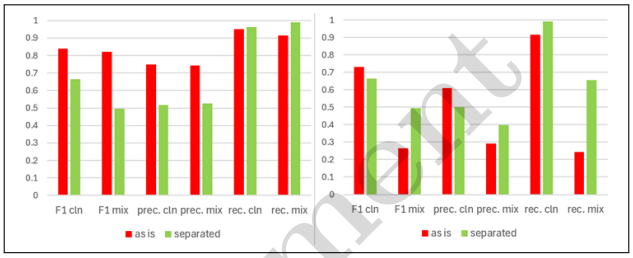

Figure 5: Impact of separation on the performance of downstream aeroplane sound recognition by AST (left) and PANN (right)

For bird song recognition a similar analysis is shown in Fig. 6. As the task is more difficult, it is not unexpected that both models perform well. In line with the previous paragraph, performance also declines when background noise is added (note that we do not distinguish between athropophony and biophony here). Sound separation using TUSS improves performance for both models even when the clean sounds are used. The increase in F1 is mainly due to the strong increase in precision; there are almost no false positives left after the bird sound is separated.

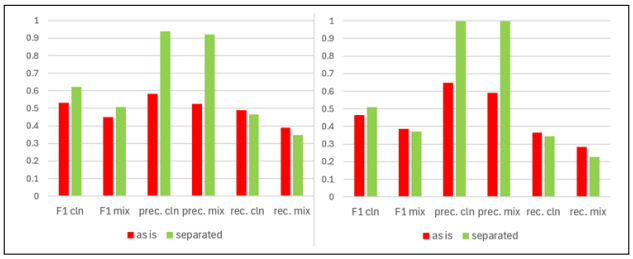

Figure 6: Impact of separation on the performance of downstream bird song recognition by BirdMAE (left) and AudioProtoPNet (right)

5. Results on independent datasets, generalisation

To assess generalisation of the approach the labelled fraction of the Risoux dataset [18] is used. This dataset is particularly rich in bird song and aeroplane sound [19].

In contrast to the results obtained for artificial mixtures, Fig. 7 neither AST nor PANN succeed in reli-

ably detecting aeroplanes in the field recordings and TUSS based separation improves performance considerable, at a small expense of introducing occasional false positives that reduce precision.

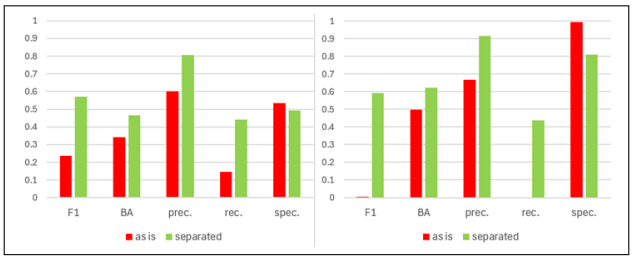

Figure 7: Impact of separation on the performance of downstream aeroplane sound recognition by AST (left) and PANN (right) for the independent Risoux dataset. In addition to F1, precision and recall, balanced approach and specificity are also reported to account for the sparsity of events in 24h recordings

For bird song recognition (Fig. 8), separation prior to recognition does not result in an improvement on the overall metric F1 and BA, neither for BirdMEA nor AudioProtoNet. An increase in specificity is replaced by a decrease in recall, probably because the separation ignores sounds where it is doubtful whether they are bird vocalisations or not. Very low precision underlies the low F1. It should however be noted that the *ground truth* identified by a single person may have missed some bird vocalisations and thus the false positives that reduce precision may actually be true bird vocalisations.

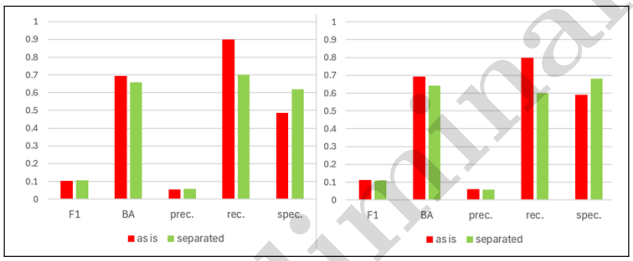

Figure 8: Impact of separation on the performance of downstream bird song recognition by BirdMAE (left) and AudioProtoPNet (right) for the independent Risoux dataset

6. Conclusions

By creating artificial mixtures of bird vocalisations and anthropogenic sounds, we showed that species recognition by state-of-the-art classifiers such as BirdNET degrades as the signal-to-noise ratio (SNR) decreases. On average, source separation based on complex ideal ratio mask (cIRM) estimation did not lead to a significant improvement in recognition performance when BirdNET was applied to the separated bird-vocalisation stream. However, responses varied across species: some species showed improved detection at positive SNRs, whereas others exhibited reduced performance. This heterogeneity may reflect BirdNET's training strategy, which incorporates data augmentation to increase resilience to acoustic interference. For acoustically *difficult* vocalisations, this built-in robustness may be insufficient, allowing the benefits of sound separation to outweigh the distortions introduced by cIRM processing. In contrast, for acoustically *easy* species, these distortions may have a greater impact than the masking effects of the anthropogenic noise, resulting in reduced recall.

Transformer-based source separation of bird vocalisations and anthropogenic noise increased the SNR of both reconstructed streams, improving their perceptual quality and facilitating subsequent manual analysis. The effect on automated classification, however, depended on the performance of the downstream classifier. For synthetic mixtures, aeroplane detection improved for the convolutional PANN model but not for the transformer-based AST model. A possible explanation is that the separation model mitigates PANN's limitations in recognising long, slowly varying acoustic events, whereas AST already achieves high aeroplane recognition accuracy. For bird vocalisations, sound separation consistently improved the precision of BirdMAE and AudioProtoPNet, largely through the elimination of false positive detections.

Evaluation on an independent field-recording dataset demonstrated that the proposed separation approach generalises beyond synthetic mixtures, maintaining and in some cases improving aeroplane detection performance. This improvement may arise because aeroplane sounds recorded in natural environments are often dominated by high-altitude overflights, which are under-represented in the training data of standard sound-event recognisers. For bird vocalisations, however, both evaluated classifiers exhibited unexpectedly low precision. Applying TUSS prior to classification modified the balance between recall and specificity but did not substantially improve precision. The prevalence of apparent false positives may indicate limitations in the ground-truth annotations, warranting further investigation through expert listening. Such analysis could determine whether some detections labelled as false positives in fact contain faint or previously overlooked bird vocalisations. Given the size of the Risoux dataset, the proposed separation framework offers a practical means of reducing the manual effort required for large-scale annotation review and validation.

While the proposed sound-stream separation framework narrows the gap between automated analysis and human listening performance, substantial opportunities for further development remain. One promising direction is the integration of multiple classifiers within a consensus-based decision framework, following the harmonisation of species and sound-event ontologies. Such an ensemble approach could leverage complementary classifier strengths and improve the robustness of biodiversity assessments. Another avenue for future

research is the exploitation of spatial information for source separation and recognition, as spatial cues play a key role in human auditory scene analysis. Although this remains challenging because most passive acoustic monitoring (PAM) systems employ single-channel recorders, increasing availability of low-cost microphone arrays may make spatially informed acoustic monitoring a feasible direction for future biodiversity studies.

7. Acknowledgments

This study is part of the PLAN-B project "The Path Towards Addressing Adverse Impacts of Light and Noise Pollution on Terrestrial Biodiversity and Ecosystems', funded by the European Union under the Horizon Europe program (project n°101135308).

References

[1] S. Wang, Y. Duan, R. Cao, J. Feng, J. Ge, and T. Wang, "Road disturbance drives a more simplified soundscape in temperate forests revealed by deep learning and acoustics indices," *Biological Conservation*, vol. 306, p. 111 115, 2025, ISSN: 0006-3207. DOI: 10.1016/j.biocon.2025.111115.

[2] R. L. Ducay and B. S. Pease, "The impact of vehicular noise on acoustic indices within simulated bird assemblage soundscapes," *Bioacoustics*, vol. 33, no. 2, pp. 203–220, 2024. DOI: 10.1080/09524622.2024.2332748.

[3] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101 236, 2021.

[4] L. Rauch, R. Heinrich, I. Moummad, A. Joly, B. Sick, and C. Scholz, "Can masked autoencoders also listen to birds?" *arXiv preprint arXiv:2504.12880*, 2025.

[5] R. Heinrich, L. Rauch, B. Sick, and C. Scholz, "Audioprotopnet: An interpretable deep learning model for bird sound classification," *Ecological Informatics*, vol. 87, p. 103 081, 2025.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016. DOI: 10.1109/TASLP.2015.2512042.

[7] K. Saijo, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Task-aware unified source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025.

[8] T. Grzywalski, D. Botteldooren, Y. Song, and N. Madhu, "Salient sound extraction using deep neural networks predicting complex masks," in *2024 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, 2024, pp. 166–171. DOI: 10.23919/SPA61993.2024.10715626.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[10] W.-P. Vellinga and R. Planqué, "The xeno-canto collection and its relation to sound recognition and classification,"

[11] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, e1008228, 2020.

[12] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13, Barcelona, Spain: Association for Computing Machinery, 2013, pp. 411–412, ISBN: 9781450324045. DOI: 10.1145/2502081.2502245.

[13] K. Saijo, G. Wichern, F. G. Germain, Z. Pan, and J. L. Roux, "Tf-locoformer: Transformer with local modeling by convolution for speech separation and enhancement," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 205–209. DOI: 10.1109/IWAENC61483.2024.10694313.

[14] L. Rauch et al., *Birdset: A large-scale dataset for audio classification in avian bioacoustics*, 2025. arXiv: 2403.10380 [cs.SD].

[15] B. Downward and J. Nordby, *The aerosonicdb (ypad-0523) dataset for acoustic detection and classification of aircraft*, 2023. arXiv: 2311.06368 [cs.SD].

[16] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[17] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[18] S. Haupert, J. Sueur, F. Sèbe, J. Barlet, and M.-P. Reynet, *Db@risoux dataset: A full year of 1-min soundscapes recorded in a cold temperate forest in france*, version 1.0, Zenodo, Feb. 2025. DOI: 10.5281/zenodo.14856482.

[19] E. Grinfeder et al., "Soundscape dynamics of a cold protected forest: Dominance of aircraft noise," *Landscape Ecology*, vol. 37, no. 2, pp. 567–582, 2022.