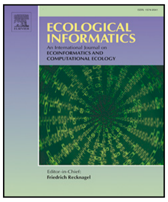
Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Computational methods for detecting insect vibrational signals in field vibroscape recordings

Matija Marolt[a,*], Matevž Pesek[a], Rok Šturm[b], Juan José López Díez[b,c], Behare Rexhepi[b,d], Meta Virant-Doberlet[b]

[a] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana, 1000, Slovenia
[b] National Institute of Biology, Department of Organisms and Ecosystems Research, Večna pot 121, Ljubljana, 1000, Slovenia
[c] University of Ljubljana, Biotechnical Faculty, Jamnikarjeva 101, Ljubljana, 1000, Slovenia
[d] University of Ljubljana, Biotechnical Faculty, Department of Biology, SubBioLab, Večna pot 111, Ljubljana, 1000, Slovenia

ARTICLE INFO

Dataset link: https://github.com/matijama/vibroscape, https://github.com/matijama/vibroscape-code

Keywords:
Vibroscape
Ecotremology
Deep learning
Automatic classification
Biotremology
Insects

ABSTRACT

The ecological significance of vibroscape has been largely overlooked, excluding an important part of the available information from ecosystem assessment. Insects rely primarily on substrate-borne vibrational signalling in their communication, which is why the majority of terrestrial insects are excluded from passive acoustic monitoring. The ability to monitor the biological component of the natural vibroscape has been limited due to a lack of data and methods to analyse the data. In this paper, we evaluate the use of deep learning models to automatically detect and classify vibrational signals from field recordings obtained with laser vibrometry. We created a dataset of annotated vibroscape recordings of meadow habitats, containing vibrational signals categorized as pulses, harmonic signals, pulse trains, and complex signals. We compared different deep neural network architectures for the detection and classification of vibrational signals, including convolutional and transformer models. The PaSST transformer architecture, which was fine-tuned from a pre-trained checkpoint demonstrated the highest performance on all tasks, achieving an average precision of 0.79 in signal detection. For signals with more than one hour of annotated data, the classification models achieved instance-based F1-scores above 0.8, enabling automatic analysis of activity patterns. In our case study, where 24-hour field recordings were analysed, the trained models (even those with lower precision) revealed interesting activity patterns of different species. The presented study, together with the dataset we publish with this paper, lays the foundation for further analysis of the vibroscape and the development of automated methods for ecotremological monitoring that complement passive acoustic monitoring and provide a comprehensive approach to ecosystem assessment.

1. Introduction

In the last decade, ecoacoustics emerged as a discipline that uses animal sounds to monitor biodiversity and gather ecological information (Sueur and Farina, 2015; Ross et al., 2023; Besson et al., 2022). Passive acoustic monitoring (PAM) has been successfully applied in terrestrial and aquatic ecosystems, as described in Sugai et al. (2019), Linke et al. (2018), Miksis-Olds et al. (2018); however, it is limited to animal groups and species that produce air-borne or underwater sounds (Ross et al., 2023). Insects are the most numerous and diverse group of terrestrial animals (Stork, 2017), and they are essential for the functioning and stability of ecosystems (e.g. Prather et al. 2013, Risch et al. 2018, Eisenhauer et al. 2023). Severe long-term decline of insects reported over the last decade (reviewed in Sánchez-Bayo and Wyckhuys, 2019; Wagner, 2020; Wagner et al., 2021) has also revealed large gaps in our knowledge of insect ecology, species distribution, population dynamics and community composition. Insect monitoring has now been recognized as one of the priorities in biodiversity assessment, and PAM has been highlighted as a useful new monitoring tool (van Klink et al., 2022). While such an approach is applicable to insects that emit air-borne or underwater sounds (e.g. orthopterans, cicadas and freshwater insects), most terrestrial insects are not covered as the majority rely on communication through substrate-borne vibrational signals (Cocroft and Rodríguez, 2005; Virant-Doberlet et al., 2023).

Hidden to the human senses and thus to our general awareness, substrate-borne vibrational signalling is one of the most common forms of animal communication, used by over 240,000 animal species, including vertebrates (Cocroft and Rodríguez, 2005; Cocroft et al., 2014).

* Corresponding author.
E-mail address: matija.marolt@fri.uni-lj.si (M. Marolt).

https://doi.org/10.1016/j.ecoinf.2025.103003
Received 18 June 2024; Received in revised form 6 January 2025; Accepted 7 January 2025
Available online 17 January 2025

Only in the last decade has the importance of vibrational communication been recognized (Cocroft et al., 2014) and the increased awareness has led to the emergence of biotremology as a discipline that studies vibrational behaviour (Hill and Wessel, 2016; Hill et al., 2019). Substrate-borne vibrations are ubiquitous in nature (Hill, 2009) and as soundscape (Pijanowski et al., 2011a,b), vibroscape is an important part of the environment. It has been defined as the collection of all substrate vibrations present in the environment and includes biological, geophysical and anthropogenic components (Šturm et al., 2019; Šturm et al., 2021, 2022). While the main source of biological vibrations are vibrational signals used for communication, they also include incidental vibrations generated by other body movements (e.g. locomotion, feeding) and substrate vibrations induced by air-borne sounds (Šturm et al., 2019; Šturm et al., 2021; Choi et al., 2024). Recently, ecotremology has been introduced as a discipline that aims to study vibroscape to assess biodiversity and ecosystem functions and propose more effective conservation plans (Šturm et al., 2022). Because of its potential, ecotremology is now increasingly considered as part of the insect biodiversity monitoring toolkit (van Klink et al., 2024). The usefulness of using ground-transmitted substrate vibrations to identify larger mammals has already been demonstrated (e.g. Szenicer et al. 2022, Parihar et al. 2021, Brickson et al. 2023).

The monitoring of small, species-rich and highly diverse plant-dwelling insects represents a major challenge for ecotremological monitoring. Although the number of vibroscape studies related to arthropod communities has been small (Šturm et al., 2021; Akassou et al., 2022; Choi et al., 2024), they have shown that species-specific vibrational signals can be identified and that vibroscape composition reflects changes in arthropod communities. One of the major challenges in these studies was the identification of vibrational signals and characterization of vibrational communities. Manual identification is very time-consuming and computational methods for automatic classification and identification of insect vibrational signals have not yet been tested on vibroscape recordings in the field. Automatic detection and recognition of vibrational signals is a challenge. Due to the enormous number and diversity of species emitting vibrational signals and the lack of reference libraries, most of vibrational signals encountered in field recordings are unknown. Furthermore, incidental vibrations are often hard to distinguish from vibrational signals used for communication. In addition, vibrational signals are subject to unpredictable degradation during transmission through the substrate (Virant-Doberlet et al., 2023) and vibrations from geophysical and anthropogenic sources overlap the frequency range of vibrational signals (Šturm et al., 2021).

In this paper, we present a study on the use of computational methods for the automatic detection and classification of vibroscape field recordings. We compare a number of deep learning approaches for this task, analyse their performance, and show how the recognition models can support the analysis of such recordings. To test the suitability of the developed AI models, we applied them to 24-hour vibroscape recordings to estimate the daily pattern of vibrational activity. To stimulate further work, we also provide a dataset of annotated field recordings used in the study.[1]

2. Related work

Computational methods for animal vocalization detection and classification for a variety of species have emerged in recent years, driven on the one hand by the development of affordable sensors and sensor networks that produce large amounts of data, and on the other hand by the development of deep learning-based machine learning methods. Stowell (2022) provides an excellent and comprehensive overview of the use of deep learning in computational bioacoustics, focusing on the various aspects of the use of machine learning in bioacoustics and the more technical aspects of recent methods. For a detailed overview of the field, we therefore refer the reader to the referenced paper.

In the following, we provide an overview of several recent works that focus on the detection and classification of insects, as these are also our target species. As mentioned by Stowell, most recent works on detection and classification employ convolutional neural networks (sometimes with recurrent modules) to analyse audio spectrograms (or related features) to detect and classify the target species. Transformer architectures are not yet widely used, mainly because they require very large datasets for training. The trained models are usually species-specific and cannot be transferred to other species without retraining, although few-shot learning models are getting noticed (Nolasco et al., 2023). For example, to monitor the activity of pollinating insects and woodpeckers, Folliot et al. (2022) trained a convolutional neural network on audio spectrograms to automatically detect the sounds of flying insects' buzzing and woodpeckers' drumming. They used the output of the model to estimate the seasonality, diel pattern, climatic breadth and distribution of the monitored species. A comparison of classical and deep learning methods for mosquito detection and classification was presented in Yin et al. (2023), while Faiß and Stowell (2023) show that the use of an adaptive and waveform-based preprocessing front-end improves the performance of deep learning methods for automatic insect recognition.

All of the above approaches perform detection on acoustic airborne stimuli, while works dealing with vibrational signals transmitted over solid media are rarer. Bhairavi et al. (2020) and Mankin et al. (2021) provide an overview of pest control approaches, also looking at techniques that rely on piezoelectric sensors, accelerometers and laser vibrometers. Rigakis et al. (2021) introduced the TreeVibes system, which uses vibroacoustic sensors and convolutional networks to detect borers in trees, while Mankin et al. (2021) showed that the same system could be used to detect the rice weevil in grain. Zhang et al. (2023b) introduced a convolutional neural network TrunkNet to detect vibrational signals of a specific trunk borer larvae species. The signals were collected by piezoelectric sensors, MFCC features were extracted, and a five-level convolutional architecture was trained to detect the target signals. In their follow-up work (Zhang et al., 2023a), they showed that convolutional+transformer modules can be used to enhance the vibrational signals of trunk-boring insects and improve classification accuracy. Korinšek et al. (2019) presented ideas for a proof-of-concept solution for detecting vibrational signals in laser vibrometry recordings, and our initial results in automatically analysing such recordings were first presented in Marolt et al. (2022).

3. Materials and methods

3.1. Dataset

The data used for our experiments contain laser vibrometry recordings collected over a period of several years. Most of them contain recordings of vibroscape in its natural environment (hay meadows), a subset also contains laboratory recordings of two species.

Field recordings of vibroscape were conducted in 2016, 2017, 2018 and 2023 on a eutrophic lowland hay meadow in the Ljubljana Moors, Slovenia (coordinates: N 45° 56′42.40'', E 14° 20′09.21''). The vibroscape was recorded with a portable Doppler laser vibrometer (Polytec PDV 100) and stored on a laptop equipped with an external sound card (Sound Blaster SBX) and Raven Pro 1.5 software with a sampling rate of 44.1 kHz and a resolution of 16 bits. To record the vibroscape, we aimed the laser beam precisely at a small piece of reflective foil attached to a plant. The sound files in .wav format were automatically saved every 10 min. The recording site had access to mains power and the devices were powered directly via an extension cable. The recordings were made in stable weather conditions to avoid damaging the equipment (Šturm et al., 2021).

[1] https://github.com/matijama/VibroScape

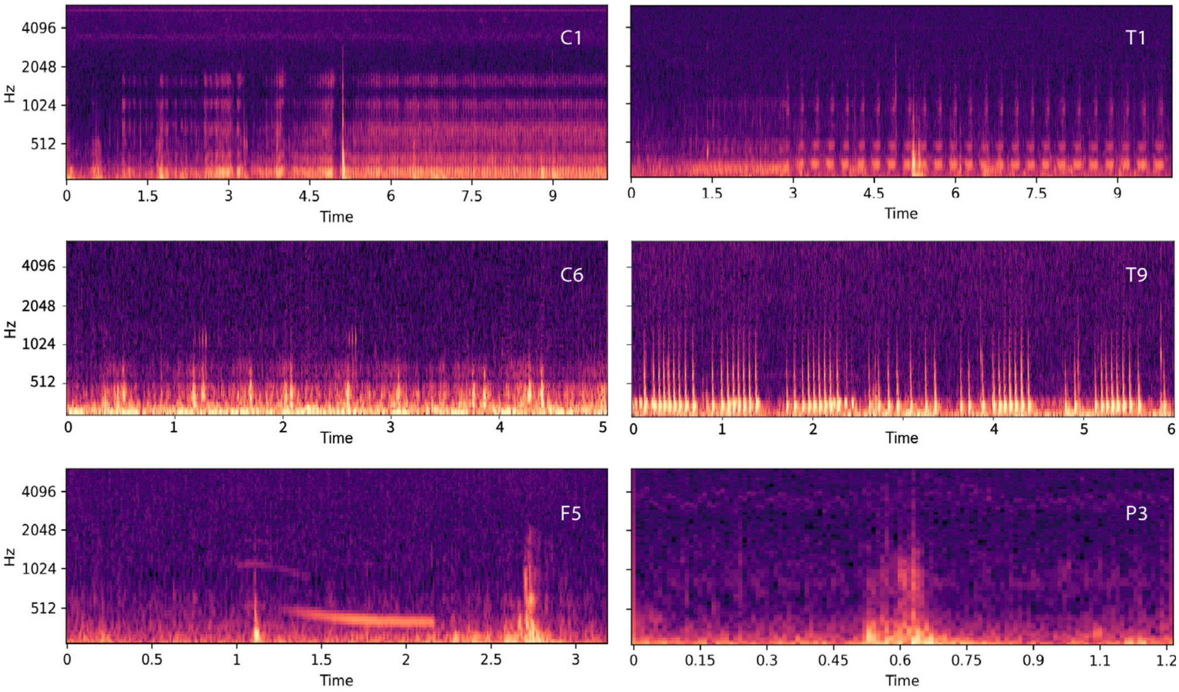


Fig. 1. Different vibrational signals: complex (C1, C6), train (T1, T9), harmonic (F5) and pulse (P3).

Table 1
Dataset by vibrational signal category. Columns two to five show the number of annotated signals, the number of distinct species, the total duration of the annotations in hours and the average signal duration (with standard deviation) in seconds.

Cat.	Count	VSTs	Dur. (h)	Avg. d. (s)
C	7411	17	13.1	6.3 ± 5.5
T	5390	33	11.3	7.5 ± 15.3
P	1527	6	0.1	0.3 ± 0.3
F	937	17	0.6	2.4 ± 4

Table 2
Six common signal types. Columns 2-4 show the number of annotated signals, the total duration of the annotations in hours and the average signal duration (with standard deviation) in seconds.

VST	Count	Dur. (h)	Avg. d. (s)
T1	1683	8.1	17.4 ± 23.8
C1	1490	6.0	14.5 ± 4.4
C2	3290	5.1	5.6 ± 2.4
T9	977	1.2	4.5 ± 5.1
C6	328	0.5	5.5 ± 6.9
T3	387	0.3	2.9 ± 1.1

In contrast, laboratory recordings of leafhoppers from the genus *Aphrodes* were collected over several years at the National Institute of Biology in Slovenia. These recordings were made during behavioural trials using a portable Doppler laser vibrometer (Polytec PDV 100). The laser beam was directed at a reflective foil attached to individual herbaceous plant cuttings with leafhoppers present. The laboratory recordings have a better signal-to-noise ratio than those obtained in the natural environment.

The dataset was manually annotated by experts (R.Š., J.J.L.D., B.R.) who marked the time/frequency range of each vibrational signal they encountered by listening to the recordings and inspecting the corresponding spectrograms. Since there are no comprehensive public reference libraries of vibrational signals, the experts classified the signals based on their distinct temporal and spectral characteristics (Šturm et al., 2021). They identified four main vibrational signal types (VSTs): pulse (P), which contains short broadband or harmonic pulses, harmonic (F), which contains longer signals with a clear harmonic structure, train (T) with regularly repeating pulses or harmonic structures, and complex (C), which contains at least two of the previous types with a relatively well-defined structure. Some of the signals could be attributed to individual species, e.g. leafhoppers *Aphrodes makarovi* (C1), *A. bicincta* Dragonja (T1), *Anoscopus serratulae* (C2), *Megophthalmus scanicus* (T3), while others were identified only by their type and number (e.g. T9), as the species producing the signal is not known. Examples of the different signal types are shown in Fig. 1, the content of the dataset is summarized in Table 1.

In total, the dataset contains about 91 h of annotated recordings, of which 10 h are laboratory recordings and the rest were recorded in the field. Of the 91 h, around 25 h are labelled as vibrational signals. Most of the VSTs belong to class C, followed by T, P and F. On average, train signals are the longest and P the shortest. The duration of T-type signals also varies greatly, as these signals have no clear structure or fixed duration. The signals of several commonly found species are summarized in Table 2, and it is not surprising that they belong to the two most common categories C and T.

Most signals are in a predominantly low-frequency range (below 3000 Hz). Field recordings contain a high level of vibrational noise from a variety of sources, including wind, air-borne sounds (e.g. birdsong), animal movements and human activity. As a transmission medium, plants act as low-frequency filters with discrete resonances (Polajnar et al., 2012), and in a grassland habitat, a random distribution of plants with different geometries and selective frequency filtering characteristics impose unpredictable effects on the signal structure (Šturm et al., 2021). An example is shown in Fig. 2, which shows a laboratory and a field recording of a T1 signal. Various noise sources and differences in the frequency characteristics are clearly recognizable.

3.2. Models

In this paper, we compare different models for the automatic detection and classification of vibrational signals. We focus on deep learning methods, which are currently the dominant approach in related works. Since our dataset is not very large, we investigate the usability of deep

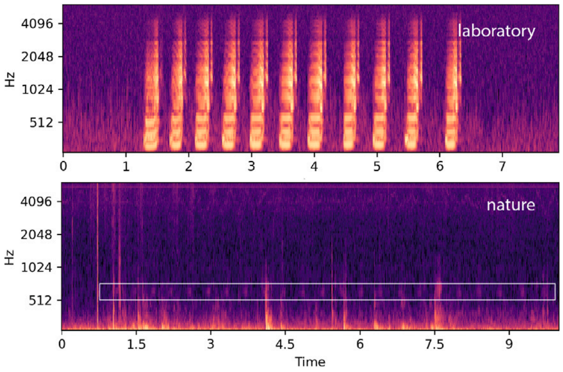


Fig. 2. Signal of T1, recorded in the laboratory on a plant on which the insect signalled, and in the natural environment at an unknown distance from the insect. In the latter case, the signal only occupies a narrow frequency band around 600 Hz due to filtering by the plants (the signal is marked by the white rectangle).

audio embeddings as well as transfer learning of models pretrained on other domains. We also investigate whether using a transformer-based architecture improves performance over convolutional networks. We compare four approaches for this task.

We investigate the use of deep audio embeddings with OpenL3 embeddings (Cramer et al., 2019) extracted with a Look, Listen, and Learn Net, a deep convolutional network trained by self-supervised learning of audio-visual correspondences in videos. The authors provide several embedding models, some trained on music clips and others trained on sounds in natural acoustic environments; all clips are from the AudioSet (Gemmeke et al., 2017). The model encodes 1 s long audio segments. Since we need a larger context to successfully detect and classify our signals (see Section 5), we extract consecutive embeddings within each context window of size w_c with a step size of s_o. On the decoder side, we use a pooling layer to summarize the embeddings over the length of the context window. We compared several pooling layer types, including temporal average pooling, statistics pooling, and self-attention (Safari et al., 2020). The pooling layer is followed by a fully connected layer with batch normalization and ReLU activation, and a final linear classification layer. We illustrate the architecture of the OL3 network in Fig. 3.

We also evaluate three different neural network architectures for this task. All three use the same input representation, namely the mel-scaled spectrogram, which represents the audio signal within the context window of size w_c.

The first model (TCSC) is a convolutional network based on 1D time-channel separable convolutions as proposed by Kriman et al. (2020). A 1D time-channel separable convolutional block consists of a 1D depthwise convolutional layer that processes each frequency channel individually but across time frames, and a pointwise convolutional layer that processes each time frame independently but across frequency channels. These layers are followed by batch normalization and the ReLU activation function. We included five such blocks of increasing convolution kernel sizes (11, 13, 15, 19, 23) in our model, with the middle three blocks containing additional residual connections, as shown in Fig. 4. The blocks are followed by a final pointwise convolution layer that gathers information across all frequency channels and feeds the decoder. For more details on the individual blocks, we refer the reader to the original paper. The time-channel separable convolutions help to keep the size of the network small; the entire network has about 71k trainable parameters. The decoder is the same as in the OL3 network.

The second model is a Large-Scale Pretrained Audio Neural Network for Audio Pattern Recognition (PANN) (Kong et al., 2020). We opted for a much larger network with 80 million trainable parameters based on the CNN14 architecture. Since the network is too large to be successfully trained on our small dataset, we used transfer learning and initialized the encoder part of the network from a model pre-trained on the AudioSet tagging task. We then fine-tuned the entire network for our task.

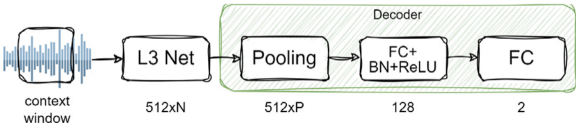


Fig. 3. The OL3 network based on OpenL3 deep audio embeddings.

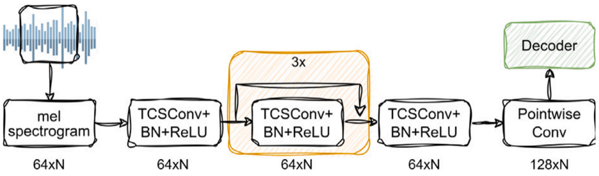


Fig. 4. The TCS network based on 1D time-channel separable convolutions.

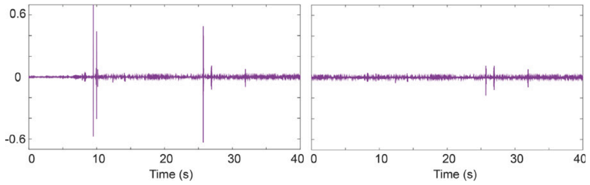


Fig. 5. Clicks in the recorded signal (left) and the signal after preprocessing.

To evaluate the performance of transformer networks, the third model is based on the Patchout faSt Spectrogram Transformer model (PaSST) (Koutini et al., 2022) with 85M trainable parameters. Similar to the PANN, we initialized the model with a network pre-trained on the AudioSet tagging task and then fine-tuned it with our dataset.

4. Computational model development

In this section, we present the details of dataset preparation, model training and testing.

4.1. Preprocessing

When examining the vibroscape recordings taken in the natural environment, we found that they often contain high-amplitude clicking sounds (see Fig. 5). We do not know whether the clicks are caused by the vibrometers or a sudden plant movement, but they interfere with the later stages of dataset preparation, especially with volume normalization and dynamic range compression. Therefore, we have implemented a simple procedure to remove the clicks.

To emphasize the clicks, the recorded signal is first filtered with a high-pass filter at 2 kHz, as the clicks contain strong high-frequency components that are otherwise not present in the signal. We estimate the position of the clicks by finding all outliers exceeding 20x standard deviation within a 200 ms sliding window over the filtered signal. Since each click can manifest itself in multiple strong peaks, we search the original signal for additional outlier points exceeding 5x standard deviation in the 10 ms neighbourhood of each click. We replace the found outlier points by a shape-preserving piecewise cubic spline interpolation of the signal. The procedure is repeated twice to capture peaks masked by high-amplitude clicks. A signal before and after preprocessing is shown in Fig. 5.

Table 3
Model performance across all tasks on the test datasets. The number of trainable parameters for each model is given in column 2, the three performance metrics (average precision, F1 frame-based and F1 instance-based) are given in columns 3-5.

Model	Par.	AP	F1 fr.	F1 in.
OL3	131k	0.65	0.61	0.66
TSCS	71k	0.70	0.65	0.70
PANN	80M	0.62	0.70	0.70
PANN tr.	80M	0.74	0.72	0.77
PaSST	86M	**0.75**	**0.75**	**0.80**

4.2. Datasets

To assess how well deep models can detect and classify vibrational signals, we evaluated them on a series of binary decision tasks. For *signal detection*, we trained a model to detect the presence of any type of vibrational signal (CFPT) in each context window. To evaluate which vibrational signal type can be better identified, we trained separate models to recognize the presence of the two most common categories: *T* and C. Finally, to assess how well individual species can be recognized, we trained separate models for the six most common VSTs: T1, T9, T3, C1, C2, C6. For each task, we prepared a separate dataset containing clips labelled as the target class or as *background*. The background class represents either background noise or other signals that do not belong to the target class.

The datasets were divided into a development (80%) and a test (20%) dataset using stratified random sampling to have approximately the same proportion of target and background classes in both sets. To avoid including samples recorded in close temporal proximity into both datasets, we put signals recorded within the same hour of the day in either the development or the test dataset, but not both.

4.3. Training

We trained the OL3 model on 48 kHz audio files, since the embedding network was trained on 48 kHz data. For the other three models, we used 16 kHz audio files as this allowed for faster data augmentation and mel spectrogram computation during training. Signal in each context window was normalized to compensate for the large differences in signal levels within and between recordings. We extracted mel spectrogram features with a 20 ms window and 10 ms hop length, using 64 mel bands from 70–5000 Hz. To improve robustness to loudness variations, we processed the mel features with adaptive per-channel energy normalization (sPCEN), which applies short-term automatic gain control to every frequency subband Zeghidour et al. (2020). Although PCEN originates from the speech domain, it is increasingly used in bioacoustics (Lostanlen et al., 2019; Cramer et al., 2020; Stowell, 2022; Faiß and Stowell, 2023).

We split the development datasets into training (90%) and validation (10%) subsets. Since they can be very unbalanced (e.g., there are only 1.2 h of T9 signals in all 91 h of recordings), we balanced each training epoch by including all samples of the target class and randomly sampling the background class to achieve a 1 : 4 ratio. We used the AdamW optimizer with a weight decay of 10^{-4} and a maximum learning rate of 10^{-4}. We trained all models for 8000 steps with a batch size of 256. We scheduled the learning rate to increase linearly for 400 steps (warm-up), stay at the maximum for 1600 steps and then decay with a ratio of 0.5.

4.3.1. Data augmentation

Large deep models tend to overfit the training data, especially with small datasets, so data augmentation is an important part of the training process (Abayomi-Alli et al., 2022). We used the following waveform-based data augmentation techniques (we randomly selected zero or more techniques and their parameters to augment each data sample):

Background noise: background noise is added to the signal. The noise is either Gaussian or comes from a dataset with background sounds (Thiemann et al., 2013; Richey et al., 2018; Ko et al., 2017; Wichern et al., 2019). The signal-to-noise ratio is randomly selected in the range from 12 to 50 dB.

Filtering: to mimic the effect of filtering by the plants, we randomly select a filter type (bandpass, bandstop, lowpass, highpass) and the respective cutoff frequencies to filter the signal.

Equalization: a parametric equalizer is used to randomly boost or cut seven frequency bands from −12 to +12 dB.

Gain: the signal gain is changed randomly between −12 and +12 dB.

Mixup: we add a randomly selected signal from the same batch to the target signal (Zhang et al., 2018).

Context window shift: since the signal annotations are usually longer than the context window, the position of the context window within the annotated region is selected at random.

In addition, we use two spectrogram-based augmentation techniques for the three models that use the mel spectrogram input representation:

Time stretching: we stretch the spectrogram along the time axis by a random factor between 0.85 and 1.15.

Specaugment: we use SpecAugment (Park et al., 2019) to mask up to 8 frequency bins and 32 time frames.

4.4. Testing

Each trained model was evaluated on a test set containing recordings that were not used for training, as explained in Section 4.2. For each recording, the context window was shifted over the entire duration with a step size of s_t seconds and the prediction of the model was interpreted as the probability of observing the target class at the centre of the context window. We evaluated the accuracy of such predictions given expert annotations and used the following performance measures:

Average precision (AP): the measure estimates the area under the precision–recall curve and is calculated as a weighted average of the precision values at each decision threshold, using the increase in recall from the previous threshold as weight. We use the model output calculated with the selected step size over the entire test set to compute the score (frame-based metric).

F1 score (F1 fr.): provides a balance between model precision and recall, using a fixed target-background threshold of 0.5. As with average precision, the metric is calculated from the classifications of the entire test set (frame-based metric).

Instance-based F1 score (F1 in.): we treat each annotation of a vibrational signal as a separate signal instance. We consider an instance to be correctly recognized if at least 1/3 of its duration is classified as the target class. As with the frame-based F1 score, we use a fixed target-background threshold of 0.5.

5. Results and discussion

All results presented in this section represent the average performance of five models trained with different train-validation splits. Unless otherwise stated, the value of the parameters described in Section 3.2 were: $w_c = 3$ s, $s_o = 0.5$ s, $s_t = 0.5$ s.

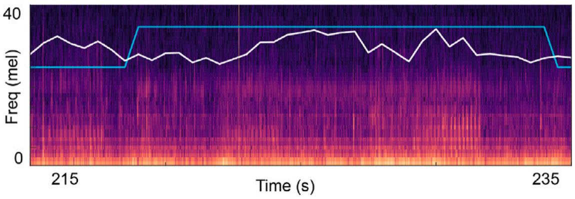


Fig. 6. Mel spectrum of a complex signal with overlaid reference annotations (blue) and model output (white). The model labels the signal only partially correctly, mainly because the signal changes from the beginning to the end, so that the beginning is missed. Although not all parts of the signal were recognized correctly, the entire instance is considered correctly recognized. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1. Models

We compare four different deep models for detection and classification of vibrational signals (see Section 3.2). We trained and evaluated all four models with the same parameters and on the same datasets. Results are shown in Table 3.

The transformer architecture, fine-tuned from a pretrained checkpoint (PaSST), shows the best performance on all 7 tasks described in Section 4.2. Its performance outperforms the convolutional approaches. Among these, the PANN fine-tuned from a pretrained model (PANN tr.) performs best, although not by a large margin over the convolutional TSCS model, which has 1000 times fewer parameters. Training a PANN model from scratch does not yield any improvements, as the dataset is too small to effectively train such a large model. Furthermore, the OpenL3 embeddings (OL3) do not seem to capture all the relevant features required to successfully classify the vibrational signals. The choice of pooling layer does not have a large impact on the results. We used statistics pooling in the results reported for TSCS and OL3.

5.2. Tasks

We show the accuracy of the PaSST model for the seven tasks described in Section 4.2 in Table 4.

For the detection of vibrational signals (CFPT), the model achieves an average precision of 0.79. A closer look at the results shows that the model (unsurprisingly) more accurately detects the complex (C) and train (T) signals, which are more common in the dataset. It is also quite successful with harmonic (F) signals, but fails with the pulse (P) signal type. Pulse signals are very short and can easily be mistaken for noise, so they are rarely recognized correctly (instance F1 is 0.37).

The recognition of the two most common signal categories (C and T) shows that although the mean average precision is very similar for both, there are large differences in the instance-based F1 measure - it increases for C signals compared to the frame-based measure and decreases for *T* signals. The better result in instance detection for complex signal types can be attributed to the variability of complex signals. They are often composed of different parts, only some of which can be recognized by the model. In addition, the parts may be separated by periods of inactivity, which are attributed to the background even though they are annotated with the target class. The model thus recognizes an instance (higher instance score), but not all segments within the instance (lower frame-based score). An example is given in Fig. 6.

For the train signal category, the instance measure is worse than the frame-based measure. This is mainly due to the many short T-signals that the model fails to detect. This has a stronger effect on the instance-based measure (whole instance missed) than on the frame-based measure (one or two frames missed). An example can be found in Fig. 7.

The accuracy of the recognition of the individual signals depends on the total duration of the signals in the dataset. With more than one hour of annotations, the accuracy of the models exceeds the instance F1 value of 0.8, so that the models are already useful as a tool for manual signal inspection or for the automatic analysis of diurnal and seasonal cycles in large datasets. With less than one hour of data, performance is worse. When training a model to detect the C6 signal (0.5 h of labelled data), the F1 instance measure only reached 0.4. Nevertheless, the model provided interesting results when analysing daily activity patterns, which were confirmed by manual inspection.

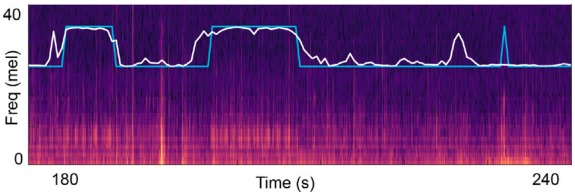


Fig. 7. Mel spectrum of a train signal with superimposed reference annotations (blue) and model output (white). The model correctly labels the longer signals, but overlooks a very short signal and also makes a false positive error. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Performance of the PaSST model on the seven tasks. The total duration (in hours) of target signals in the training set is given in column 2, followed by the three performance metrics on the test set (AP, F1 fr. and F1 in.).

Task	Dur. (h)	AP	F1 fr.	F1 in.
CFPT	21.5	0.79	0.73	0.76
T	10.0	0.70	0.68	0.54
C	10.8	0.69	0.74	0.82
T1	6.6	0.81	0.73	0.85
T3	0.3	0.48	0.51	0.67
T9	1.0	0.90	0.86	0.83
C1	5.1	0.68	0.76	0.85
C2	4.2	0.65	0.75	0.92
C6	0.1	0.57	0.61	0.52

Table 5
Influence of context sizes on model accuracy.

Context (s)	AP	F1 fr.	F1 in.
4	**0.75**	**0.75**	0.79
3	**0.75**	**0.75**	**0.80**
2	**0.75**	**0.75**	0.79
1	0.72	0.70	**0.80**

5.3. Context size and dynamic range compression

The size of the context window limits the length of the signal that the model processes in each step. The smaller the window, the shorter the signal that the model processes. Since many vocalizations are periodic, using windows shorter than one period leads to poor classification. On the other hand, the temporal localization of the model predictions is less accurate if the context size is too large, since the model can detect the signal anywhere within the context window, so the signal boundaries can be extended, as can also be observed in Fig. 7. We therefore trained models with different context sizes to find an optimal value. The results for the PaSST model are shown in Table 5. As can be seen, short context sizes degrade per frame accuracy, while larger sizes result in similar performance. For our experiments, we chose a context size of three seconds as it represents a good compromise. Interestingly, three seconds is also used in other animal vocalization studies, e.g. for bird call recognition (Kahl et al., 2021).

Vibroscape recordings show large amplitude fluctuations in both the vibrational signals and the surrounding noise sources. The amplitudes can vary depending on the location of the signal in relation to the

Table 6
Comparison of dynamic compression methods. The first column indicates whether the signal was normalized before processing, and the second indicates the dynamic range compression method used.

Norm.	Dyn.	AP	F1 fr.	F1 in.
no	log	0.70	0.68	0.69
yes	log	0.73	0.71	0.77
no	sPCEN	**0.75**	**0.75**	0.79
yes	sPCEN	**0.75**	**0.75**	**0.80**

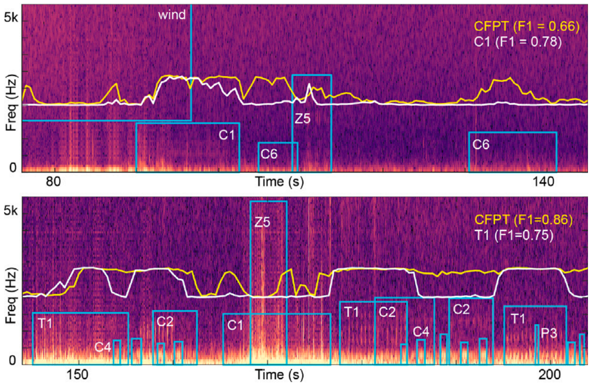


Fig. 8. Classification of two different field recordings. Outputs of different models (yellow and white lines) are superimposed on the corresponding spectrograms, with blue boxes indicating the reference annotations of the signals. The upper image shows the output of a detection model (CFPT, yellow) and a C1 model (white), the lower image shows the output of the detection model and a T1 model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
The distribution of false positive classifications for species-specific models. The ratio of false positive activations compared to all positive activations is shown in column two, followed by the distribution of the false positives across the four signal categories, labelled noise (N) and background (Bg).

Model	FPs	C	T	F	P	N	Bg
T1	0.13	0.25	0.06	0.07	0.03	0.23	0.36
T3	0.33	0.30	0.31	0.10	0.00	0.12	0.17
T9	0.07	0.26	0.01	0.00	0.00	0.25	0.48
C1	0.19	0.16	0.27	0.02	0.01	0.17	0.37
C2	0.12	0.21	0.14	0.01	0.01	0.15	0.48
C6	0.5	0.16	0.14	0.00	0.02	0.35	0.32

recording position and on the characteristics of the plants carrying the signal. To reduce these variations, we compared two dynamic range compression methods: standard log-mel spectrogram compression and adaptive per-channel energy normalization (sPCEN) (Zeghidour et al., 2020). We also tested whether normalizing the vibrational signal before spectrogram calculation affects the detection. The results are summarized in Table 6.

Using normalization as a preprocessing step always improves the results as the maximum signal level is equalized in all examples. sPCEN also brings a significant performance improvement by learning channel- and time-dependent compression parameters. We have also experimented with replacing the entire input representation (mel-sPCEN) with the learnable frontend for audio classification LEAF (Zeghidour et al., 2020), which also learns filterbank parameters. LEAF has been shown to be beneficial in some studies (Faiß and Stowell, 2023), but in our experiments it did not improve performance, a result shared by others (Schlüter and Gutenbrunner, 2022).

5.4. Error analysis

In Fig. 8 we show how the models perform on typical field vibroscape recordings. Two recordings are shown that contain a variety of overlapping vibrational signals as well as noises including wind and animal movement (Z5). Both images show the output of the signal detection model (CFPT, yellow), as well as species-specific models C1 (top) and T1 (bottom), both shown in white.

Overall, the classifications are mostly correct (frame-based F1 scores are given in the figure). All models are affected by noise, which can lead to either false positive activations (wind and animal movements, Fig. 8 upper image) or false negative activations (animal movements, Fig. 8 lower image). Errors are also caused by plant filtering, which alters the signal, and by signal overlap, both of which can cause the species-specific models to either not label a signal or to label it incorrectly (the first C2 label, Fig. 8 bottom image). Sometimes the beginning of a signal is not recognized immediately, but sometimes the labels are not placed very precisely, as can be seen in the first C2 label in the bottom image, which should start earlier.

The analysis of the false positive activations of the species-specific models (Table 7) shows that false positives often occur in noisy passages. The last two columns of the table show the proportion of false positives in regions where noise is either explicitly labelled (N, e.g. as animal movements, wind, birds or other sources) or is just background noise (Bg). As the background class is more common than labelled noise, errors also more commonly occur on background. Models that recognize the train signal type (T1, T9) also frequently activate on complex signals (about 25% of false positive errors), but do not make errors within their own signal category as often, while models that recognize complex signals (C1, C2) make about the same amount of errors with other C or T signals.

In laboratory recordings, where noise is minimal and there is no overlap between the signals, the frame-based F1-measure of the species-specific models C1 and T1 (only these two were recorded under laboratory conditions) is above 0.95.

6. Case study

The aim of the case study was to investigate the distribution of vibrational signals of different species over entire 24-hour periods on several days.

6.1. Case study dataset

The case study dataset contained 24-hour field recordings collected in 2016 and 2017 in the same meadow with the same protocol and equipment as described in Section 3.1. This means that the laser beam of a laser Doppler vibrometer (Polytech PDV 100) was precisely aimed at a small piece of reflective foil attached to a plant. 10-minute recordings were saved with Raven Pro 1.5 on a laptop equipped with an external sound card (Sound Blaster SBX) with a sampling rate of 44.1 kHz and a resolution of 16 bits. The dataset consists of continuous recordings between July 16–20, 2016 and July 5–8, 2017, a total of 8 days. It includes 1123 10-minute audio files in .wav format, totalling 187 h of recordings, slightly less than 8 full days, as some technical issues led to short dropouts.

These recordings are not annotated, so they were not included in our training and testing data and were not analysed elsewhere prior to our case study.

6.2. Experiment

All the case study dataset recordings were fed through each of the nine PaSST models (see Table 4) with a step size of 0.5 s. Outputs of each model indicate the presence of the vibrational signal it detects. They were averaged over one minute periods and the daily activities visualized as shown in Fig. 9.

The number of detections was unevenly distributed over the 24-hour period, revealing a diel variation in signalling activity, with the highest

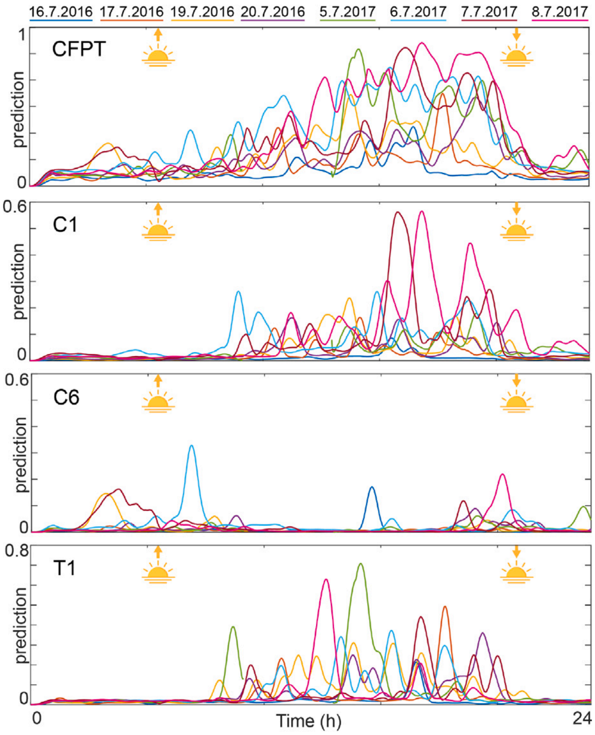


Fig. 9. Smoothed output of four models (CFPT, C1, C6 and T1) for eight 24-hour recordings. The recording dates are presented in different colours, as shown on top of the figure. Average sunrise (5:25) and sunset (20:50) times are indicated with icons. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

activity during the day between 10:00 and 20:00 (Fig. 9). A similar pattern in the vibroscape composition of meadows has already been described when considering only six short, manually labelled intervals throughout the day (Šturm et al., 2021). Automated screening of the vibroscape allowed us to determine the activity more precisely, providing new insights into the diel variation of the vibroscape composition as well as the partitioning of vibrational communication space. We were able to identify consistent activity patterns between species (Fig. 9). In particular, VST C6, attributed to so far unidentified species, was regularly registered during the night and around sunrise and sunset, but only rarely during the day when other VSTs were predominant. In contrast to the soundscape, the vibroscape is invisible to the human senses and such patterns, which are crucial for conducting ecotremological monitoring, could so far only be determined by time-consuming manual analyses.

7. Conclusion

In the paper we presented a comparison of different deep architectures for the task of detecting and recognizing vibrational signals in laser vibrometry recordings created in natural environments. With the paper, we also publish the dataset used for our experiments. To our knowledge, this is the first such study of field vibroscape recordings and by publishing the dataset we hope to stimulate interest in automatic vibroscape analysis, as it has proven to be a challenging task. We have shown that adapting a pretrained transformer model leads to models that are already useful for automatic signal detection and classification of common species, and can be used by domain experts to facilitate inspection of recordings and statistical analysis of diurnal cycles.

Field recordings are noisy, and we have shown that noise has a large impact on performance, so increasing the robustness of the models to various noise sources should improve their performance. Noise suppression of the recordings themselves would also be beneficial to facilitate expert review, as the vibrational signal is often drowned out by ambient noise. Unsupervised and few-shot learning architectures will be explored to better categorize species that are not very common and also to find new signal types in recordings.

The information extracted from the biological vibroscape component provides the opportunity to comprehensively assess ecosystems (Šturm et al., 2022). We hope that this study, together with our dataset, will provide a starting point for other researchers to accelerate the implementation of vibroscape monitoring in ecosystem assessment.

CRediT authorship contribution statement

Matija Marolt: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Matevž Pesek:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Rok Šturm:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Juan José López Díez:** Visualization, Validation, Investigation, Data curation. **Behare Rexhepi:** Investigation, Data curation. **Meta Virant-Doberlet:** Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Acknowledgements

This research was partly funded by the Slovenian Research and Innovation Agency, within the project *Vibroscape: discovering an overlooked world of vibrational communication*, grant number J1-3016, research core funding Structure and function of ecosystems (P1-0255) and project Ecotremology – Investigating ecological interactions through vibroscape (Z1-50018).

Data availability

The data used in the study is available on GitHub (https://github.com/matijama/vibroscape). Source code for the training the deep models is also available on GitHub (https://github.com/matijama/vibroscape-code).

References

Abayomi-Alli, O.O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., Misra, S., 2022. Data augmentation and deep learning methods in sound classification: A systematic review. Electronics 11 (22), http://dx.doi.org/10.3390/electronics11223795, URL: https://www.mdpi.com/2079-9292/11/22/3795.

Akassou, I., Zapponi, L., Verrastro, V., Ciolli, M., Mazzoni, V., 2022. Extending the vibroscape to agroecosystems: investigating the influence of abiotic factors and monitoring insect vibrational signaling. PeerJ 10, http://dx.doi.org/10.7717/peerj.14143.

Besson, M., Alison, J., Bjerge, K., Gorochowski, T.E., Høye, T.T., Jucker, T., Mann, H.M., Clements, C.F., 2022. Towards the fully automated monitoring of ecological communities. Ecol. Lett. 25, 2753–2775. http://dx.doi.org/10.1111/ele.14123.

Bhairavi, K.S., Bhattacharyya, B., Manpoong, N.S., Das, P.P.G., Devi, E.B., Bhagawati, S., 2020. Recent advances in exploration of acoustic pest management: A review. J. Entomol. Zool. Stud 8, 2056–2061.

Brickson, L., Zhang, L., Vollrath, F., Douglas-Hamilton, I., Titus, A.J., 2023. Elephants and algorithms: a review of the current and future role of AI in elephant monitoring. J. R. Soc. Interface 20, http://dx.doi.org/10.1098/rsif.2023.0367.

Choi, N., Miller, P., Hebets, E.A., 2024. Vibroscape analysis reveals acoustic niche overlap and plastic alteration of vibratory courtship signals in ground-dwelling wolf spiders. Commun. Biol. 7, http://dx.doi.org/10.1038/s42003-023-05700-6.

Cocroft, R.B., Gogala, M., Hill, P.S.M., Wessel, A., 2014. In: Cocroft, R.B., Gogala, M., Hill, P.S., Wessel, A. (Eds.), Fostering Research Progress in a Rapidly Growing Field, vol. 3, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 3–12. http://dx.doi.org/10.1007/978-3-662-43607-3_1.

Cocroft, R.B., Rodríguez, R.L., 2005. The behavioral ecology of insect vibrational communication. BioScience 55, 323–334.

Cramer, A.L., Lostanlen, V., Farnsworth, A., Salamon, J., Bello, J.P., 2020. 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, pp. 901–905. http://dx.doi.org/10.1109/ICASSP40776.2020.9052908.

Cramer, A.L., Wu, H.H., Salamon, J., Bello, J.P., 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 3852–3856.

Eisenhauer, N., Ochoa-Hueso, R., Huang, Y., Barry, K.E., Gebler, A., Guerra, C.A., Hines, J., Jochum, M., Andraczek, K., Bucher, S.F., Buscot, F., Ciobanu, M., Chen, H., Junker, R., Lange, M., Lehmann, A., Rillig, M., Römermann, C., Ulrich, J., Weigelt, A., Schmidt, A., Türke, M., 2023. Ecosystem consequences of invertebrate decline. Curr. Biol. 33, 4538–4547.e5. http://dx.doi.org/10.1016/j.cub.2023.09.012.

Faiß, M., Stowell, D., 2023. Adaptive representations of sound for automatic insect recognition. PLoS Comput. Biol. 19 (10), 1–22. http://dx.doi.org/10.1371/journal.pcbi.1011541.

Folliot, A., Haupert, S., Ducrettet, M., Sèbe, F., Sueur, J., 2022. Using acoustics and artificial intelligence to monitor pollination by insects and tree use by woodpeckers. Sci. Total Environ. 838, 155883.

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 776–780.

Hill, P.S., 2009. How do animals use substrate-borne vibrations as an information source? Naturwissenschaften 96, 1355–1371. http://dx.doi.org/10.1007/s00114-009-0588-8.

Hill, P.S.M., Virant-Doberlet, M., Wessel, A., 2019. What is Biotremology? In: Hill, P.S.M., Lakes-Harlan, R., Mazzoni, V., Narins, P.M., Virant-Doberlet, M., Wessel, A. (Eds.), Biotremology: Studying Vibrational Behavior. Springer International Publishing, Cham, pp. 15–25. http://dx.doi.org/10.1007/978-3-030-22293-2_2.

Hill, P.S., Wessel, A., 2016. Biotremology. Curr. Biol. 26 (5), R187–R191. http://dx.doi.org/10.1016/j.cub.2016.01.054.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: A deep learning solution for avian diversity monitoring. Ecol. Inform. 61, 101236, URL: https://api.semanticscholar.org/CorpusID:232359415.

van Klink, R., August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F., Høye, T.T., Jongejans, E., Menz, M.H., Miraldo, A., Roslin, T., Roy, H.E., Ruczyński, I., Schigel, D., Schäffler, L., Sheard, J.K., Svenningsen, C., Tschan, G.F., Wäldchen, J., Zizka, V.M., Åström, J., Bowler, D.E., 2022. Emerging technologies revolutionise insect ecology and monitoring. Trends Ecol. Evol. 37, 872–885. http://dx.doi.org/10.1016/j.tree.2022.06.001.

van Klink, R., Sheard, J.K., Høye, T.T., Roslin, T., Nascimento, L.A.D., Bauer, S., 2024. Towards a toolkit for global insect biodiversity monitoring. Phil. Trans. R. Soc. B 379, http://dx.doi.org/10.1098/rstb.2023.0101.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 5220–5224. http://dx.doi.org/10.1109/ICASSP.2017.7953152.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio, Speech Lang. Proc. 28, 2880–2894. http://dx.doi.org/10.1109/TASLP.2020.3030497.

Korinšek, G., Tuma, T., Virant-Doberlet, M., 2019. Automated vibrational signal recognition and playback. Biotremology: Stud. Vib. Behav. 149–173.

Koutini, K., Schlüter, J., Eghbal-zadeh, H., Widmer, G., 2022. Efficient training of audio transformers with patchout. In: Ko, H., Hansen, J.H.L. (Eds.), Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association. Incheon, Korea, 18-22 September 2022, ISCA, pp. 2753–2757. http://dx.doi.org/10.21437/INTERSPEECH.2022-227.

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., Zhang, Y., 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6124–6128.

Linke, S., Gifford, T., Desjonquères, C., Tonolla, D., Aubin, T., Barclay, L., Karaconstantis, C., Kennard, M.J., Rybak, F., Sueur, J., 2018. Freshwater ecoacoustics as a tool for continuous ecosystem monitoring. Front. Ecol. Environ. 16, 231–238. http://dx.doi.org/10.1002/fee.1779.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., Bello, J.P., 2019. Robust sound event detection in bioacoustic sensor networks. PLoS One 14 (10), e0214168.

Mankin, R., Hagstrum, D., Guo, M., Eliopoulos, P., Njoroge, A., 2021. Automated applications of acoustics for stored product insect detection, monitoring, and management. Insects 12 (3), 259.

Marolt, M., Šturm, R., López Díez, J.J., Pesek, M., 2022. Who's shaking? : on using machine learning to detect vibrational signals in laser vibrometry recordings. In: Biotremology Abstract Book. p. 28.

Miksis-Olds, J.L., Martin, B., Tyack, P.L., 2018. Exploring the ocean through sound. Acoust. Today 14, 26–34.

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., et al., 2023. Learning to detect an animal sound from five examples. Ecol. Inform. 77, 102258.

Parihar, D.S., Ghosh, R., Akula, A., Kumar, S., Sardana, H.K., 2021. Seismic signal analysis for the characterisation of elephant movements in a forest environment. Ecol. Inform. 64, http://dx.doi.org/10.1016/j.ecoinf.2021.101329.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. Interspeech 2019.

Pijanowski, B.C., Farina, A., Gage, S.H., Dumyahn, S.L., Krause, B.L., 2011a. What is soundscape ecology? An introduction and overview of an emerging new science. Landsc. Ecol. 26, 1213–1232. http://dx.doi.org/10.1007/s10980-011-9600-8.

Pijanowski, B.C., Villanueva-Rivera, L.J., Dumyahn, S.L., Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N., 2011b. Soundscape ecology: The science of sound in the landscape. BioScience 61, 203–216. http://dx.doi.org/10.1525/bio.2011.61.3.6.

Polajnar, J., Svenšek, D., Čokl, A., 2012. Resonance in herbaceous plant stems as a factor in vibrational communication of pentatomid bugs (Heteroptera: Pentatomidae). J. R. Soc. Interface 9 (73), 1898–1907.

Prather, C.M., Pelini, S.L., Laws, A., Rivest, E., Woltz, M., Bloch, C.P., Toro, I.D., Ho, C.K., Kominoski, J., Newbold, T.A.S., Parsons, S., Joern, A., 2013. Invertebrates, ecosystem services and climate change. Biological Rev. 88, 327–348. http://dx.doi.org/10.1111/brv.12002.

Richey, C., Barrios, M.A., Armstrong, Z., Bartels, C., Franco, H., Graciarena, M., Lawson, A., Nandwana, M.K., Stauffer, A., van Hout, J., Gamble, P., Hetherly, J., Stephenson, C., Ni, K., 2018. Voices obscured in complex environmental settings (VOICES) corpus. arXiv:1804.05053.

Rigakis, I., Potamitis, I., Tatlas, N.A., Potirakis, S.M., Ntalampiras, S., 2021. TreeVibes: Modern tools for global monitoring of trees for borers. Smart Cities 4 (1), 271–285.

Risch, A.C., Ochoa-Hueso, R., van der Putten, W.H., Bump, J.K., Busse, M.D., Frey, B., Gwiazdowicz, D.J., Page-Dumroese, D.S., Vandegehuchte, M.L., Zimmermann, S., Schütz, M., 2018. Size-dependent loss of aboveground animals differentially affects grassland ecosystem coupling and functions. Nat. Commun. 9, http://dx.doi.org/10.1038/s41467-018-06105-4.

Ross, S.R., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N., Sethi, S.S., Wood, C.M., Burivalova, Z., 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Funct. Ecol. 37, 959–975. http://dx.doi.org/10.1111/1365-2435.14275.

Safari, P., India Massana, M.À., Hernando Pericás, F.J., 2020. Self-attention encoding and pooling for speaker recognition. In: Interspeech 2020, 21st Annual Conference of the International Speech Communication Association: Virtual Event. Shanghai, China, 25-29 October 2020, International Speech Communication Association (ISCA), pp. 941–945.

Sánchez-Bayo, F., Wyckhuys, K.A.G., 2019. Worldwide decline of the entomofauna: A review of its drivers. Biol. Cons. 232, 8–27. http://dx.doi.org/10.1016/j.biocon.2019.01.020.

Schlüter, J., Gutenbrunner, G., 2022. Efficientleaf: A faster learnable audio frontend of questionable use. In: 2022 30th European Signal Processing Conference. EUSIPCO, IEEE, pp. 205–208.

Stork, N.E., 2017. How many species of insects and other terrestrial arthropods are there on earth? Annu. Rev. Entomol. 63, 31–45. http://dx.doi.org/10.1146/annurev-ento-020117.

Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152.

Šturm, R., Díez, J.J.L., Polajnar, J., Sueur, J., Virant-Doberlet, M., 2022. Is it time for ecotremology? Front. Ecol. Evol. 10, http://dx.doi.org/10.3389/fevo.2022.828503.

Šturm, R., Rexhepi, B., López Díez, J.J., Blejec, A., Polajnar, J., Sueur, J., Virant-Doberlet, M., 2021. Hay meadow vibroscape and interactions within insect vibrational community. IScience 24 (9), 103070. http://dx.doi.org/10.1016/j.isci.2021.103070.

Sueur, J., Farina, A., 2015. Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics 8, 493–502. http://dx.doi.org/10.1007/s12304-015-9248-x.

Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W., Llusia, D., 2019. Terrestrial passive acoustic monitoring: Review and perspectives. BioScience 69, 5–11. http://dx.doi.org/10.1093/biosci/biy147.

Szenicer, A., Reinwald, M., Moseley, B., Nissen-Meyer, T., Muteti, Z.M., Oduor, S., McDermott-Roberts, A., Baydin, A.G., Mortimer, B., 2022. Seismic savanna: machine learning for classifying wildlife and behaviours using ground-based vibration field recordings. Remote. Sens. Ecol. Conserv. 8, 236–250. http://dx.doi.org/10.1002/rse2.242.

Thiemann, J., Ito, N., Vincent, E., 2013. The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings. J. Acoust. Soc. Am. 133 (5_Supplement), 3591.

Virant-Doberlet, M., Stritih-Peljhan, N., Žunič-Kosi, A., Polajnar, J., 2023. Annual review of entomology functional diversity of vibrational signaling systems in insects. Annu. Rev. Entomol. 2023 68, 2022. http://dx.doi.org/10.1146/annurev-ento-120220.

Šturm, R., Polajnar, J., Virant-Doberlet, M., 2019. Practical issues in studying natural vibroscape and Biotic noise. In: Hill, P.S., Lakes-Harlan, R., Mazzoni, V., Narins, M.P., Virant-Doberlet, M., Wessel, A. (Eds.), Springer, p. 526, Wagner, D.L., 2020. Insect declines in the anthropocene. Annu. Rev. Entomol. 65, 457–480. http://dx.doi.org/10.1146/annurev-ento-011019-025151.

Wagner, D.L., Grames, E.M., Forister, M.L., Berenbaum, M.R., Stopak, D., 2021. Insect decline in the anthropocene: Death by a thousand cuts. Proc. Natl. Acad. Sci. USA 118, http://dx.doi.org/10.1073/PNAS.2023989118.

Wichern, G., Antognini, J.M., Flynn, M., Zhu, L.R., McQuinn, E., Crow, D., Manilow, E., Roux, J.L., 2019. WHAM!: Extending speech separation to noisy environments. In: Interspeech. URL: https://api.semanticscholar.org/CorpusID:195776451.

Yin, M.S., Haddawy, P., Ziemer, T., Wetjen, F., Supratak, A., Chiamsakul, K., Siritanakorn, W., Chantanalertvilai, T., Sriwichai, P., Sa-ngamuang, C., 2023. A deep learning-based pipeline for mosquito detection and classification from wingbeat sounds. Multimedia Tools Appl. 82 (4), 5189–5205.

Zeghidour, N., Teboul, O., de Chaumont Quitry, F., Tagliasacchi, M., 2020. LEAF: A learnable frontend for audio classification. In: International Conference on Learning Representations.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations.

Zhang, H., Li, J., Cai, G., Chen, Z., Zhang, H., 2023a. A CNN-based method for enhancing boring vibration with time-domain convolution-augmented transformer. Insects 14 (7), 631.

Zhang, X., Zhang, H., Chen, Z., Li, J., 2023b. Trunk borer identification based on convolutional neural networks. Appl. Sci. 13 (2), 863.