JULY 13 2023

Effects of landscape and distance in automatic audio based bird species identification

Panu Somervuo ; Patrik Lauha; Tapio Lokki

J. Acoust. Soc. Am. 154, 245–254 (2023)
https://doi.org/10.1121/10.0020153

JASA ARTICLE

Effects of landscape and distance in automatic audio based bird species identification

Panu Somervuo,[1,a)] Patrik Lauha,[1] and Tapio Lokki[2]
[1]*Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland*
[2]*Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*

ABSTRACT:
The present work focuses on how the landscape and distance between a bird and an audio recording unit affect automatic species identification. Moreover, it is shown that automatic species identification can be improved by taking into account the effects of landscape and distance. The proposed method uses measurements of impulse responses between the sound source and the recorder. These impulse responses, characterizing the effect of a landscape, can be measured in the real environment, after which they can be convolved with any number of recorded bird sounds to modify an existing set of bird sound recordings. The method is demonstrated using autonomous recording units on an open field and in two different types of forests, varying the distance between the sound source and the recorder. Species identification accuracy improves significantly when the landscape and distance effect is taken into account when building the classification model. The method is demonstrated using bird sounds, but the approach is applicable to other animal and non-animal vocalizations as well.
 https://doi.org/10.1121/10.0020153
(Received 2 January 2023; revised 27 June 2023; accepted 28 June 2023; published online 13 July 2023)
[Editor: Amanda M. Lauer] Pages: 245–254

I. INTRODUCTION

Automatic animal sound identification is an active research topic, and lately it has become more important due to the growing number of passive acoustic monitoring (PAM) applications (Furnas and Callas, 2015; Gibb *et al.*, 2019; Piña-Covarrubias *et al.*, 2019; Priyadarshani *et al.*, 2018; Shonfield and Bayne, 2017; Sugai *et al.*, 2019; Sugai *et al.*, 2020).

Sound propagation and attenuation have been studied extensively in physics (Attenborough, 2002; Bass *et al.*, 1984; Bullen and Fricke, 1982; Harris, 1966; Price *et al.*, 1988). There are also studies on how they are related to bird song and its observations (Snell-Rooda, 2012; Yip *et al.*, 2017a; Yip *et al.*, 2017b). In open space under ideal conditions, sound propagates equally to all directions, and the energy is distributed to a surface of a sphere. The surface area of a sphere grows quadratically as a function of distance, and the sound energy attenuates 6 dB per doubling of distance. However, in addition to the distance effect, the properties of the medium affect the sound propagation. These include factors such as air temperature, pressure, and humidity. Moreover, real environments are not just open spaces, but there are various types of objects like trees, rocks, soft ground, hard ground, etc., that affect how the sound waves are absorbed and reflected. For example, a simple environmental effect, such as an echo, can make a drastic difference to the characteristics of a sound. The goal in this work is not to construct any physical model of how the sound attenuates between the source and the recording unit, but rather to measure empirically the effects of various environments and distances and then find out how they affect the accuracy of automatic bird species identification. Based on earlier studies, it is already known that distance affects recognition accuracy (Haupert *et al.*, 2023; Knight and Bayne, 2019; Leseberg *et al.*, 2022; Shaw *et al.*, 2022). Dabelsteen (1993) has studied how the song of *Turdus merula* (common blackbird) degraded by quantifying attenuation, signal-to-noise ratio, and blur ratio. There is also work that tries to estimate the distance between a bird and a recording unit based on the sound data (Darras *et al.*, 2018).

PAM relies on the use of autonomous recording units (ARUs). They can be placed at any site where humans can go, but in contrast to human listeners, they are able to perform recording continuously as long as their batteries last. Data can be easily collected in huge volumes, and this creates the challenge of how to analyze the content of the recordings. Manual annotation and labeling are slow, and therefore, there is a great need to automate the process. There are many alternatives in machine learning for building a classifier, but for audio based bird species identification, the most widely used methods at the moment are based on convolutional neural networks (CNNs) (Goëau *et al.*, 2016; Kahl *et al.*, 2021; Lasseck, 2018). Training data for these classifiers consist of previously collected and labeled data. Bird vocalizations can be extracted from public databases, such as xeno-canto (Xeno-canto Foundation, 2005) or Macaulay Library (Macaulay Library, 2021). One problem

[a)]Electronic mail: panu.somervuo@helsinki.fi

with this approach is that in many cases, the environment where the training data have been recorded does not match the environment where the model is going to be used to classify new audio data. For example, data in xeno-canto may have been recorded with a high-quality directional microphone with a short distance between the recorder and the bird. In PAM and ARU, however, data are typically recorded with omnidirectional microphones, and the distance between the bird and the recorder varies.

Like any neural network, CNNs require training data to optimize their parameters before they can be used for classifying new data. In general, neural networks learn to associate input patterns to given labels during the training process. The key challenge in the training is how to make the network generalize well to unseen input patterns. Bird song has variation on many levels due to the geographical location, environment, and individual. Training data consist of recorded vocalizations of birds, and in practice, the amount of available data is always limited. Therefore, no practical set of recordings can contain all possible variations of bird sounds. Consequently, if the limited variation of the training data is not taken into account, the network can easily become overfitted to the fixed number of training data patterns. Overfitting means that the network learns to map the given training samples perfectly against the given output labels but does not generalize its performance to unseen data. A typical approach to solving this problem is to add perturbations to the training data. This is called data augmentation. The goal is to avoid overfitting and improve generalization. One of the challenges of neural network training is to find the right type of variations to add in the training data.

In this work, it is shown how the effect of a specific landscape and distance can be added to the training data of the neural network. The benefit of the proposed approach is that even when the network is trained using data recorded elsewhere, the effect of any specific landscape can be added computationally. This requires measuring impulse responses in the new environment, which is much faster to do than playing back all the bird sounds in the new environment and recording them. The structure of this paper is to first explain how the impulse response is measured in a real environment, how the landscape effect can be introduced to a recorded bird sound, and how the neural network is trained with the augmented data. Models that have been trained with the landscape and distance effect data are compared against the baseline model using both synthetic test data (bird sounds convolved with the impulse response of the landscape) and playback data from the real environment (bird sounds played from the loudspeaker).

II. MATERIALS AND METHODS

The methods include measuring the landscape specific impulse responses and applying them to a set of existing bird recordings. Impulse responses are estimated based on recorded sweep signals, which are sinusoidal signals rising exponentially in time. Figure 1 shows the pipeline for training a landscape specific neural network for identifying bird species based on the proposed method.

A. Impulse response recording sites

Audio recordings were collected on two days with different climatic conditions. The first recording included three locations at Nuuksio, Southern Finland, on September 22, 2022. One of the locations was an open field, and the two other locations were young mixed forest and old coniferous forest. The forests were relatively flat, although the old coniferous forest had shallow (about 1 m deep) pits on the measurement line. The young mixed forest had a small hill on the right at 50 m distance. Both forests were in their natural states, and they included trees with varying diameters and heights. The heights of the taller trees were between 20 and 30 m.

During the measurements, background noise level L_{Aeq} was 29 dB, temperature 8 °C, humidity 79%, wind 1 m/s, and air pressure 1023 hPa. Climatic measures were obtained from the Finnish Meteorological Institute. Distance between

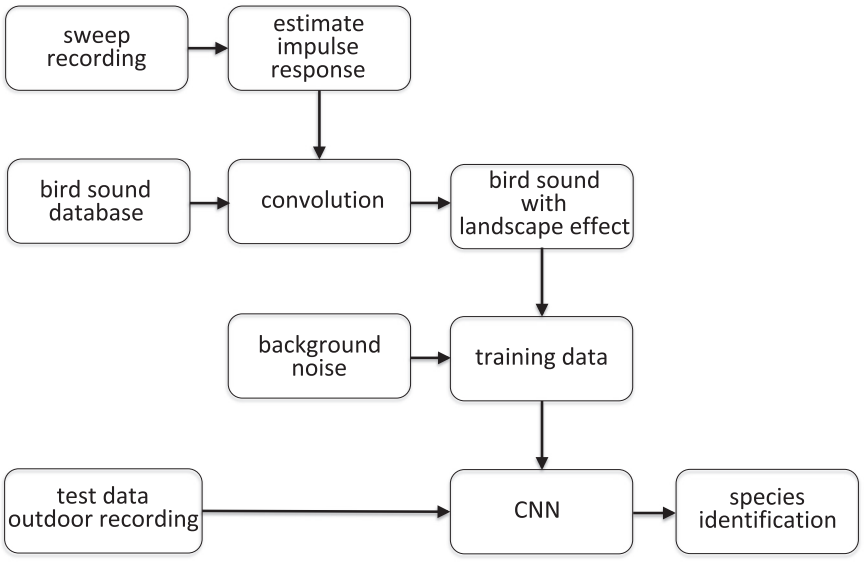

FIG. 1. Flow chart of the proposed method to apply a landscape effect to an existing bird sound and train a neural network that is better adjusted to the sound characteristics of a given landscape.

https://doi.org/10.1121/10.0020153

the sound source and the recording unit was measured by a knotted cord. Recordings of sweeps were performed at distances from 5 to 90 m at every 5 m, totaling 18 recordings per site.

The second set of measurements took place in the same young mixed forest at Nuuksio, Southern Finland, on March 7, 2023. The recordings included measuring the playbacks of bird songs in addition to the sweeps. Distance between the sound source and the recording unit varied from 20 to 90 m at 10 m intervals. In the beginning of the recordings, temperature was −8 °C, humidity 86%, and wind 1 m/s, and in the end, temperature was −5 °C, humidity 67%, and wind 3 m/s. The ground was covered with snow (depth 42 cm), and there was snow on the branches of the trees.

B. Sound source

An LS01 (ACOEM Group, Limonest, France) loudspeaker (on the left in Fig. 2) was located on a stand 1.2 m above the ground. A 19-s long exponentially rising sinusoidal sweep was played covering the frequencies from 50 Hz to 20 kHz. Sound pressure was $L_{\mathrm{Aeq}} = 85$ dB measured at 3 m distance. Due to the clipping of sound in a recording unit at the two closest source positions, which was found only after the first landscape was measured (young forest), in the two other landscapes (open field and old forest), the gain was reduced by 12 dB when the distance between the loudspeaker and the recording unit was 5 m and by 6 dB when the distances were 10 and 15 m. The level differences were compensated back after the computation of impulse responses.

C. ARUs

This study used AudioMoths (Open Acoustic Devices) (Hill *et al.*, 2018), which are small devices with a micro-electro-mechanical systems (MEMS) microphone. MEMS microphones are typically inexpensive, but they have a large frequency range, and in addition to birds, frogs, and mammals, they can be used to record bat sounds. The biggest disadvantage of a MEMS microphone is the signal-to-noise ratio, which can be considerably lower compared to more expensive microphones. The sampling frequency of AudioMoth can be set between 16 and 384 kHz. Each recording creates a file in WAV format where a sample value is represented by 16 bits. Technical information about AudioMoths can be found on the manufacturer's web page. Also, a GitHub repository (Lapp, 2021) contains information regarding frequency response, polar pattern, effect of different types of cases, and measurements in room, grassland, and forest environments.

FIG. 2. (Color online) The applied battery driven omnidirectional sound source and AudioMoth microphones with and without the case.

Sound recordings in the present study were made using six AudioMoth 1.1.0 devices. Their firmware was updated to AudioMoth-Firmware-Basic 1.8.1, which was the most recent version available at the time of the experiment. Sampling rate was 48 kHz, and medium gain was used. The devices were tied to a tree trunk at a level of 1.2 m above ground. Four of the devices were used without a case, and two of them were inside an AudioMoth IPX7 case (see Fig. 2). The recording angles of the devices without a case were 0°, 60°, 120°, and 180° measured from the center of the tree toward the sound source. Two devices with a case had angles 0° (in front of the tree) and 180° (at the back of the tree).

D. Estimation of impulse responses

Based on the input signal and the recorded output signal, the way that the input signal changes due to the environment can be calculated. When the environment is modeled as a time-invariant linear system, the effect is characterized by an impulse response. The impulse responses were measured with the swept sinusoid technique (Farina, 2000). In brief, the method works as follows. The exponentially rising sinusoid in time is recorded, and the impulse responses are obtained by convolving the recordings with the time domain inverse of the excitation. Since the frequency response of the exponential sweep drops 3 dB per octave, the amplitude of the time inverse must be modified to obtain the flat frequency response. This method is widely used in room acoustical measurements and results in a high signal-to-noise ratio. Moreover, it allows the removal of possible harmonic distortion components of the sound source.

For 3 landscapes, 18 distances, and 6 microphones, the measurements result in a total of 324 impulse responses. Due to the clipping of sound in one landscape for the two shortest distances, 312 impulse responses were used from the first set of measurements. The additional measurements with snow conditions from one landscape included seven distances and six microphones. Due to the failure of one microphone, 35 impulse responses were used from that set.

Impulse responses were calculated and applied using MATLAB's floating point resolution without discretizing and saving them to WAV files in between. The lengths of the impulse responses were limited to be 500 ms since the tails contained mostly noise based on manual inspection.

E. Adding the effect of landscape and distance to a bird sound

The effect of landscape and distance on the sound is obtained by computing the convolution between the bird sound and the landscape-distance specific impulse response. The calculation was performed in the frequency domain, and inverse Fourier transform was used to get back a time domain signal. The number of time points in the FFT was a power of 2, and FFT length was set to be longer than the length of the input signal by using zero padding in the original input signal. The result of the inverse FFT was truncated to have the same length as the original input signal.

F. Bird audio data

To test the species identification, 2020 audio recordings were picked from xeno-canto (Xeno-canto Foundation, 2005). They covered 101 different species (20 recordings per species). As indicated earlier, $312 + 35$ landscape-distance effects were applied to each record via convolution. The xeno-canto recordings used and their detailed classification results are listed in supplementary material.[1] Playback data included 202 audio recordings that were randomly selected from the set of 2020 xeno-canto samples, so that each species was represented by two recordings. Playback data were played at multiple distances between the loudspeaker and the recording unit.

Training data were the same that had been used in Lauha *et al.* (2022) and contained 1000 vocalizations for each species. Ideally, the initial training data should be free of any environmental effects, as the purpose of the proposed method is to add the effect of a desired landscape. Since the training data in this study were from Macaulay Library (2021) field recordings, they inevitably contained some effect of their original recording environment.

G. CNN

For testing the accuracy of bird identification, a CNN was used (Lauha *et al.*, 2022). Its input consisted of a 3-s time window of audio signal that was converted into a 129 $\times$ 128 pixel spectrogram, and its output gave a probability for 101 Finnish bird species. The CNN contained four convolutional layers and two dense layers, and it was trained with PYTHON libraries Keras and TensorFlow for ten epochs using an RMSprop optimizer, a learning rate of 0.0001, and a binary cross-entropy loss function. Augmentations such as horizontal and vertical stretching, shifting, and masking were applied during training.

There were six different versions of the original network. All models were trained from scratch with the same hyperparameters. The training data of the first network consisted of the original bird recordings. Five other networks were trained using the data with different synthetic landscape effects. There were three networks corresponding to summer conditions (impulse responses from three locations from September) and two networks corresponding to winter conditions (impulse responses from one location from March). The difference between the two networks trained with March impulse responses was how the background noise level was added in the training data. In the first model, the noise level was random from a prior distribution (winter CNN1), and in the second model, the noise level was determined based on which distance effect was applied to the training data (winter CNN2). For simplicity, it was assumed that the source signal energy decreases 6 dB per doubling of distance, and the background noise level was set by visual investigation of the spectrograms so that the noise had negligible effect for the training data mimicking the effect of 20 m. These values were not estimated from playback data in order not to overfit the training data to the test data. Background noise was sampled from the mixture of white and pink noise. Within each landscape specific network, training data with different distance effects were merged.

When including the landscape and distance effects in the training data, the impulse responses of only one microphone were used (the microphone that was without AudioMoth case in front of the tree toward the sound source). From the first set of impulse response measurements, to represent different distances evenly in the training data, distances were put into three categories. In the first category, there were impulse responses from 5, 15, and 25 m; in the second category, there were impulse distances corresponding to 35, 45, and 55 m; and in the third category, there were impulse responses corresponding to 65, 75, and 85 m. During training, one impulse response was randomly selected from each category so that there were three different modifications for each original training sample. From the second set of impulse response measurements, the training data contained the effect of 30, 50, and 70 m. When creating test data, all impulse responses were applied covering all measured distances.

H. Statistical analysis

Species identifications were performed for the original data, the data to which landscape-distance effects had been added, and playback data. The classification was deemed correct if the output with the highest probability corresponded to the species label of the input sample. Accuracy was calculated by dividing the number of correct classifications by the total number of recordings. The 95% confidence interval (CI) of the classification accuracy was calculated using normal approximation. Accuracy drop as a function of distance was investigated by fitting a linear model.

Mean frequency for each individual bird sound was calculated from the spectrogram. Mean frequency of each time frame was computed as an energy-weighted sum of frequency bins, and the average over the entire bird sound duration was calculated as a frame based energy-weighted mean frequency. Accuracy drop as a function of mean frequency was investigated by fitting a linear model.

Sound attenuation was calculated by fitting a linear function between $10\log_{10}(P)$ and $\log_2(r)$, where P is the energy of the signal divided by its duration, and r is the

https://doi.org/10.1121/10.0020153

distance between the source and the recorder. The resulting slope gives the relative increase in power in dB when doubling the distance. A negative slope indicates attenuation.

III. RESULTS

The present study provides three main results. The first shows that both landscape and distance have an effect on the accuracy of automated bird identification. Second, the accuracy can be improved by accounting for these effects when training the classification model. Third, landscape and distance have different effects for different species.

A. Distance effects

Investigation of impulse responses shows how landscape and distance affect the original signal. The recorded impulse responses contain all sounds captured by recording units, i.e., reflections from trees, ground attenuation, etc. Frequency responses were calculated for each recording unit based on the Fourier transform of the impulse response. Ideally, if there were no effects, the magnitude spectra would be flat, meaning that no frequencies are attenuated or amplified differently. Naturally, the spectra contain the spectrum of the sound source and the recording unit, in addition to the effect of the landscape. Figure 3(A) shows frequency responses of all six recording units, two with and four without a case during recording. The biggest difference between the two units is visible at low frequencies, where it can be seen that the AudioMoth case markedly attenuates frequencies below 2 kHz. At high frequencies, on the other hand, there seems to be a small gain for the units with the case. Figures 3(B) and 3(C) illustrate the effect of the young forest landscape at different distances. For these plots, the recorders in front of the tree were used. The distance effect where high frequencies are attenuated more compared to low frequencies (starting from 1 kHz) is most clearly visible in Fig. 3(B), which shows the magnitude spectrum of the impulse response corresponding to a 90-m distance.

Sound attenuation was close to 6 dB per doubling of distance when the energy was calculated from sweep outputs. The slope of the fitted curve was −5.98, and R^2 was 0.93. The slope was −6.32, and R^2 was 0.96 when fitting the curve to playback data.

The effect of distance in a forest landscape is shown in the spectrogram of a typical song pattern of *Fringilla coelebs* (common chaffinch) (Fig. 4). The original sound has strong harmonic components due to the close range recording. The details of the spectrogram get blurred as the distance effect gets stronger. It is also visible how high frequencies attenuate faster than low frequencies.

B. Classification accuracy

Figure 5 shows the effects of three landscapes on the classification accuracy of data with synthetic landscape effects. Open field had the smallest effect, as expected, and young and old forests had a stronger distance effect. In old forest, the accuracy dropped drastically after 15 m. This is partly explained by the shape of the landscape: There were small dips and hillocks along the recording line. After training the model to take into account the landscape-distance effects, the accuracies became highly similar between all landscapes, and the distance effect almost disappeared. For the field, young forest, and old forest data with the baseline model, the negative slopes of the accuracy drop curves were 1.6, 3.0, and 3.5, respectively, units being percentage per 10 m. When landscape effect was included in the model, the corresponding accuracy drops were 0.5, 0.3, and 0.4.

Table I shows the species identification accuracies for all test data and model combinations with synthetic effects. The best accuracy was obtained using the model that had been trained with the data from the same domain as the test data. Using the original model, all data with landscape effects were identified considerably worse compared to the original data. After adding the landscape effect to the training data, the accuracies from landscape specific models were very close to the original accuracy of 90%. Both forest data were identified almost equally well with the model specific to their own domain and the model from another forest type. Original data were identified equally well with the baseline model and the field effect model.

The identification accuracies of playback data are shown in Fig. 6. All models that had been trained with

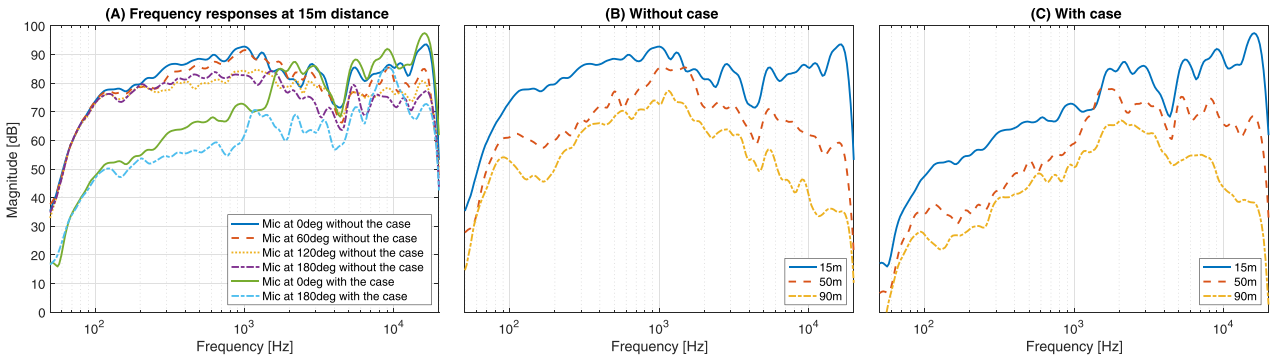

FIG. 3. (Color online) Frequency responses of AudioMoth devices around the tree (A). The effect of young forest at three distances without (B) and with a case (C). Devices were located in front of the tree (angle 0° toward the sound source).

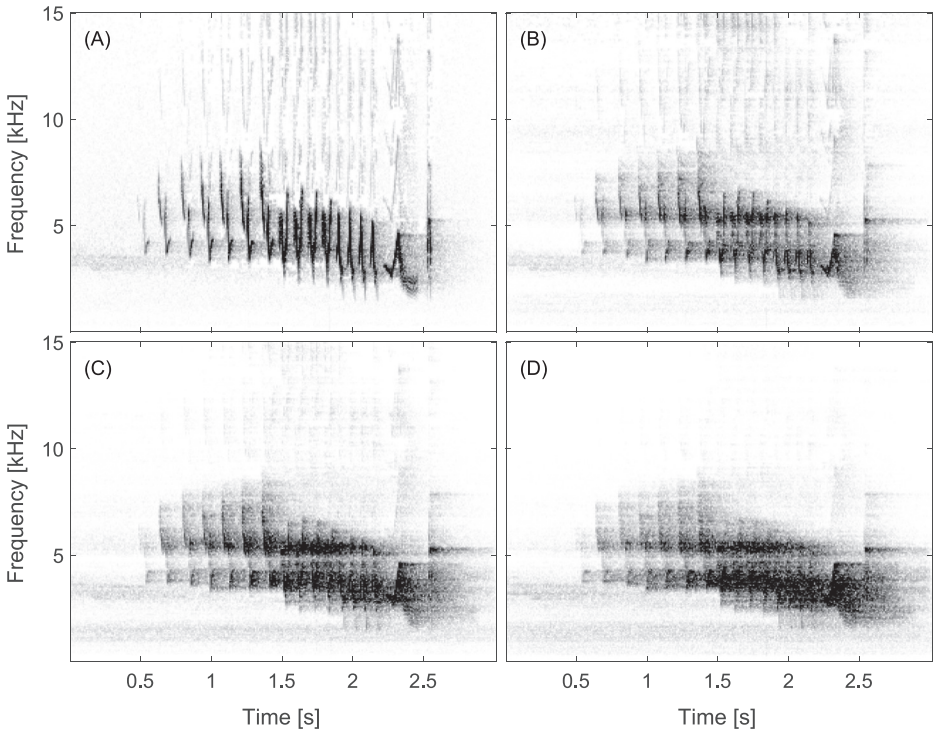

FIG. 4. Examples of landscape and distance effects. Original sound of *F. coelebs*, xeno-canto record XC383547 (A), and the resulting sound with the effect of young forest with distance of 15 m (B), 50 m (C), and 90 m (D).

synthetic landscape effects gave higher accuracies compared to the baseline model. The best performing model was based on the impulse responses that had been measured with identical conditions that occurred when the playback data were recorded (winter CNN). The results also show that it is beneficial to mimic the degradation of signal-to-noise ratio as a function of distance (winter CNN2 vs winter CNN1).

C. Species diversity

The results shown previously were averages over all 101 species. When looking at the results coming from the baseline model, there were noticeable differences between different species. For some birds, such as *Locustella naevia* (common grasshopper warbler) and *Dryocopus martius* (black woodpecker), the distance effect was hardly noticeable in the identification accuracy in all landscapes, whereas for *Sylvia atricapilla* (blackcap) and *Clangula hyemalis* (long-tailed duck), the accuracy dropped in all landscapes as a function of distance. For *Cygnus cygnus* (whooper swan) and *Carduelis carduelis* (goldfinch), the accuracy dropped in both forest landscapes but not in the field landscape. When using landscape specific models, all accuracy drops became smaller. There was no significant correlation between the accuracy drop as a function of distance and mean frequency of the sound.

IV. DISCUSSION

When studying the effect of distance and sound degradation in relation to bird sounds, most studies in the

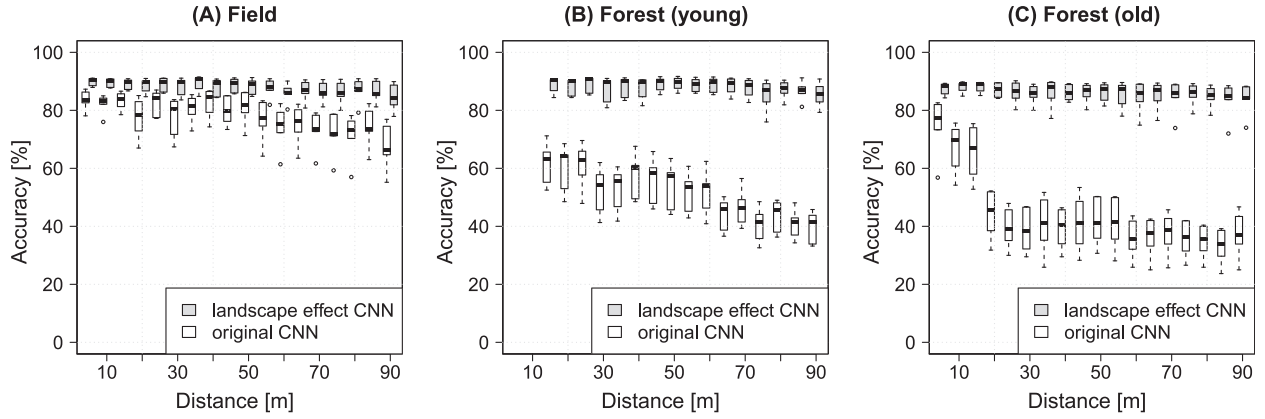

FIG. 5. Neural network based species identification accuracies for bird sounds with synthetic landscape and distance effects. Results for landscape specific data are shown separately in (A)–(C). The landscape specific neural network in each panel has seen the effect of only one microphone at distances 5, 15, 25, 35, 45, 55, 65, 75, and 85 m via training data. Boxplots show the variation between six recording units. For each of them, the accuracies have been averaged over all 101 bird species. Landscape effect CNN refers to the model that has been trained with the synthetic landscape and distance effects, and original CNN refers to the baseline model.

https://doi.org/10.1121/10.0020153

TABLE I. Species identification accuracies (percent) with 95% CIs for original and landscape effect data using original and landscape specific models. The highest accuracy in each case is in bold.

Data	Original Model	Field Model	Young forest Model	Old forest Model
Original	**90.4** ± 1.3	90.2 ± 1.3	81.5 ± 1.7	84.4 ± 1.6
Field	77.2 ± 1.8	**87.3** ± 1.5	81.0 ± 1.7	81.3 ± 1.7
Young forest	52.0 ± 2.2	66.0 ± 2.1	**87.9** ± 1.4	86.0 ± 1.5
Old forest	48.1 ± 2.2	63.3 ± 2.1	84.0 ± 1.6	**86.1** ± 1.5

literature have focused only on a single species. In a study by Leseberg *et al.* (2022), the similarity score between the calls of *Pezoporus occidentalis* (night parrot) was calculated based on template matching with a binary template. The score was defined as a difference between the mean amplitude detected in the template's "on" and "off" cells. The score decreased exponentially as a function of distance. In a study by Knight and Bayne (2019), the calls of *Chordeiles minor* (common nighthawk) were recorded at near, mid range, far, and mixed distances. Distance was a significant factor for the variation of correct classification at near and mid range recorders, but it explained only a small portion of variation at mid range and far distances. In that study, the classification was based on HMM. In the present study, CNN was used as a classifier. Visual investigation of accuracy drop versus distance did not suggest any strong exponential decrease. There was a significant drop in accuracy after 15 m in the old forest, but the most likely explanation for this was the shape of the landscape: There were hillocks that absorbed the sound. For other landscapes, the visual investigation suggested a linear trend in the accuracy drop. The most likely reason for this is that the present study included distances only up to 90 m. For longer distances, an exponentially decreasing function would be a more useful and appropriate choice.

In a study by Haupert *et al.* (2023), white noise was played and recorded at different locations along a 100-m transect. It was shown that the detection distance can be predicted knowing the contribution of each attenuation factor, the coefficient of attenuation of the habitat, the ambient sound pressure level, and the amplitude and frequency bandwidth characteristics of the transmitted sound. In our study, we have used impulse responses that contain the information on how different frequency components are attenuated and what kind of reverberation effect the landscape introduces. The only remaining parameter to be set in order to mimic the real sound is the ambient noise and its level. The presence of ambient noise has a larger proportional effect when the distance between the microphone and the target sound source increases. When mimicking the sound coming from a more distant source by applying the convolution with the corresponding impulse response, the decreased signal-to-noise ratio can be taken into account when training the neural network. This improves the classification accuracy as indicated in the results of playback data [see Fig. 6(B)]. One could further argue that the training data should be specific to the type of recording device because there are noticeable differences between different microphone types (Darras *et al.*, 2020).

There were interesting differences between the results of different bird species. It is generally known that high frequencies attenuate faster along a distance than low frequencies. According to Snell-Rooda (2012), signals of lower frequency, narrower bandwidth, and longer duration are more detectable in environments with high sound absorption. That study found evidence that warbler species with higher mean absorption were more likely to have narrow bandwidth songs. Therefore, the initial speculation to explain the results of the present study was that the identification of bird species whose vocalizations are at low frequencies would be less affected than the identifications of

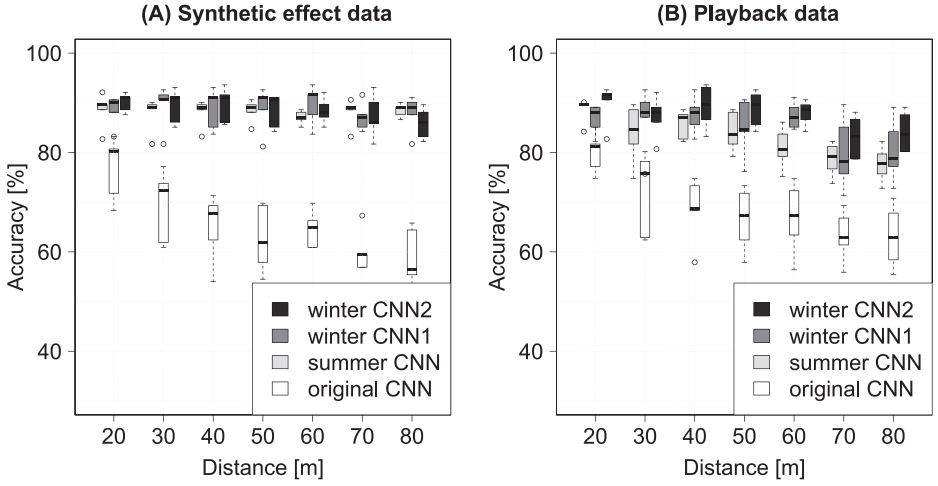

FIG. 6. Species identification accuracies for synthetic effect data and playback sounds covering 101 bird species recorded in winter conditions. The results are shown for the baseline model and models trained with synthetic landscape effects specific to the recording site (young forest). Summer CNN is the model for which the impulse responses were measured without snow, and winter CNN is the model for which the impulse responses were measured in winter conditions with snow. The latter model has two versions differing in how the background noise level was set. Boxplots show the variation between five recording units.

other bird species. Another initial thought was that the degree to which the signal is spread into different frequencies might explain how much the species identification suffers from the distance effect. If the bird sound has a narrow frequency range, although its intensity will be reduced if the sound is at a high frequency, the pattern of the sound should remain the same, and therefore, the distance would not degrade the identification. In contrast, if the song pattern covers a large frequency range, some part of it will be dampened more than other parts, and that type of perturbation might be more challenging for a computer to recognize when it has been trained to recognize only clean sounds recorded at short distances. Despite some effort, the authors of the present study were not able to find any simple reasons why some species were more affected by the distance effect than other species. A complicating fact is that although it can be seen from the magnitude spectrum of the impulse response which frequencies will be attenuated, it is not necessarily known which parts of the spectrum are key factors for species identification from the neural network's point of view.

In the present study, multiple distance effects were pooled in the training data of a single model. Alternatively, one could train a distance specific model for each landscape. It is tempting to speculate that if the distance to the bird were known at the moment of recording its sound, that information could improve the classification. Birds seem to be able to assess the distance based on the degradation of the sound (McGregor and Krebs, 1984). Cues for distance perception include reverberation, overall amplitude, and relative intensities of frequencies (Naguib and Wiley, 2001). These have been found in studies of both humans and birds. The familiarity of a sound and the knowledge of the properties of the transmission path also play a role. In a study by Darras *et al.* (2018), human listeners estimated distances based on the recordings. To automate the distance estimation, a straightforward engineering solution would be to use multiple microphones and estimate the source location based on the differences between the times of arrival of the signal. One can also use a neural network for the task (Adavanne *et al.*, 2019).

In the present study, there was a separate model for each landscape. An interesting question is whether there are benefits if the effects of different environments are pooled in a single model. In this case, the training data would consist of samples representing multiple types of distance effects in multiple types of environments. In practice, the locations where ARUs are placed are known, so in principle, it is easy to measure the impulse responses in those specific places and train a specific model for each individual site. However, even when keeping the effects of completely different environments separate, there could be benefits in pooling the impulse responses among similar environments so that there are more measurements available. In the present work, the impulse responses were measured at several distances but only along a single line. The more measurements from the same environment, the better. Also, in the present work, the measurements were done only at the ground level, but it would be interesting to expand the locations of impulse sources into three-dimensional (3D) space covering also different heights. How many measurements are needed to cover the specific site depends on the characteristics of the landscape. The results of the present study indicated some robustness against different orientations of the recording units compared to the direction of the sound source. The six AudioMoths were located at different positions around the tree, and the training data were created by applying the effect of only one microphone in front of the tree. Also, the classification accuracies for winter data using the summer landscape model were only slightly lower compared to those of the winter model.

Bird recordings for which the synthetic effects are applied may already contain the effects of their original recording environments. To make the sounds as authentic as possible, i.e., to make the sound as close as it would be in a real environment, the original bird sound should be as clean from the effect of its original environment as possible. The best candidates would be good quality recordings obtained with directional microphones. Most of the recordings that were used in the present study were of good quality, although the amount of background noise in them varied. In close range recordings with a directional microphone, it may be assumed that recordings have only a minimal effect of environment. However, since none of the Macaulay Library samples have been recorded in a free field, they are not completely anechoic. As the results show, the method seems to be quite robust to the variations of the original training data quality, but we have not investigated the limits of the method, i.e., how noisy an input sample could still be useful. The ideal type of training data to which to apply the impulse response based effects would be free of any original environment effects.

One issue that was not covered in the present study was the presence of multiple bird species vocalizing simultaneously. The proposed way that the distance effect is introduced to any sound also enables mimicking of cases where two or more birds are located at different distances. With a single bird source, sound level normalization is not critical, since typically, the input is normalized anyway before being fed into a neural network. However, if many birds vocalize at the same time, in addition to the way that different frequencies are attenuated as a function of distance, also the sound level can be set for each bird species. In practice, all possible combinations are not possible to simulate in data augmentation, but the more realistic the variation included in the training data is, the better performing model will be the outcome.

V. CONCLUSIONS

The present work has introduced a way of adding landscape and distance effect to the original sound. The proposed method is especially suitable for PAM applications in fixed sampling sites. Impulse responses measured in these

sites can be applied to an unlimited amount of bird species and vocalizations to create a comprehensive set of bird sounds in the specific environment. Sweeps can be easily recorded from multiple distances and directions, whereas recording playbacks multiple times would be practically impossible for any sufficiently large set of vocalizations. Measuring a 19-s sweep allows a dense spatial sampling grid to characterize the landscape.

An interesting topic to study in the future would be to do the inverse, i.e., to remove the effect of the recording environment in the sound. It would be straightforward if the distance is known and the distance specific impulse response is available. This would give a method to purify sounds by applying deconvolution. In practice, there are many challenges, e.g., knowing the distance based on the sound alone and handling the presence of multiple birds vocalizing simultaneously. With several AudioMoths at decent distances from each other, the triangularization could be used to locate the birds, separate different sound sources, and estimate the recording distances.

Another interesting direction for future studies would be to investigate how much variation there is in the impulse responses, both within individual sites and between different landscapes.

ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 856506). Special thanks to Bess Hardwick for checking the language of the manuscript.

[1]See supplementary material at https://doi/org/10.1121/10.0020153 for classification results of data with landscape and distance effects.

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (**2019**). "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE J. Sel. Top. Signal Process. **13**(1), 34–48.

Attenborough, K. (**2002**). "Sound propagation close to the ground," Annu. Rev. Fluid Mech. **34**(1), 51–82.

Bass, H., Sutherland, L., Piercy, J., and Evans, L. (**1984**). "Absorption of sound by the atmosphere," in *Physical Acoustics: Principles and Methods*, edited by W. P. Mason and R. N. Thurston (Academic, Orlando, FL), Vol. 17, pp. 145–232.

Bullen, R., and Fricke, F. (**1982**). "Sound propagation through vegetation," J. Sound Vib. **80**(1), 11–23.

Dabelsteen, T. (**1993**). "Habitat-induced degradation of sound signals: Quantifying the effects of communication sounds and bird location on blur ratio, excess attenuation, and signal-to-noise ratio in blackbird song," J. Acoust. Soc. Am. **93**(4), 2206–2220.

Darras, K., Deppe, F., Fabian, Y., Kartono, A. P., Angulo, A., Kolbrek, B., Mulyani, Y., and Prawiradilaga, D. (**2020**). "High microphone signal-to-noise ratio enhances acoustic sampling of wildlife," PeerJ **8**, e9955.

Darras, K., Furnas, B., Fitriawan, I., Mulyani, Y., and Tscharntke, T. (**2018**). "Estimating bird detection distances in sound recordings for standardizing detection ranges and distance sampling," Methods Ecol. Evol. **9**(9), 1928–1938.

Farina, A. (**2000**). "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proceedings of the 110th Audio Engineering Society Convention*, February 19–22, Paris, France.

Furnas, B., and Callas, R. (**2015**). "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," J. Wildl. Manage. **79**(2), 325–337.

Gibb, R., Browning, E., Glover-Kapfer, P., and Jones, K. (**2019**). "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring," Methods Ecol. Evol. **10**(2), 169–185.

Goëau, H., Glotin, H., Vellinga, W., Planqué, R., and Joly, A. (**2016**). "LifeCLEF bird identification task 2016: The arrival of deep learning," in *Proceedings of CLEF: Conference and Labs of the Evaluation Forum*, September 5–8, Evora, Portugal, pp. 440–449.

Harris, C. (**1966**). "Absorption of sound in air versus humidity and temperature," J. Acoust. Soc. Am. **40**(1), 148–159.

Haupert, S., Sèbe, F., and Sueur, J. (**2023**). "Physics-based model to predict the acoustic detection distance of terrestrial autonomous recording units over the diel cycle and across seasons: Insights from an Alpine and a Neotropical forest," Methods Ecol. Evol. **14**(2), 614–630.

Hill, A., Prince, P., Covarrubias, E., Doncaster, C., Snaddon, J., and Rogers, A. (**2018**). "Audiomoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment," Methods Ecol. Evol. **9**(5), 1199–1211.

Kahl, S., Wood, C., Eibl, M., and Klinck, H. (**2021**). "BirdNET: A deep learning solution for avian diversity monitoring," Ecol. Inform. **61**(7), 101236.

Knight, E., and Bayne, E. (**2019**). "Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software," Bioacoustics **28**(6), 539–554.

Lapp, S. (**2021**). "Audiomoth performance testing: A quantitative report of audio recording quality for the audiomoth," https://github.com/kitzeslab/audiomoth-performance (Last viewed January 1, 2023).

Lasseck, M. (**2018**). "Acoustic bird detection with deep convolutional neural networks," in *Detection and Classification of Acoustic Scenes and Events*, edited by D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumble (IEEE, Piscataway, NJ).

Lauha, P., Somervuo, P., Lehikoinen, P., Geres, L., Richter, T., Seibold, S., and Ovaskainen, O. (**2022**). "Domain-specific neural networks improve automated bird sound recognition already with small amount of local data," Methods Ecol. Evol. **13**(12), 2799–2810.

Leseberg, N., Venables, W., Murphy, S., Jackett, N., and Watson, J. (**2022**). "Accounting for both automated recording unit detection space and signal recognition performance in acoustic surveys: A protocol applied to the cryptic and critically endangered night parrot (*Pezoporus occidentalis*)," Austral Ecol. **47**(2), 440–455.

Macaulay Library (**2021**). "Macaulay Library—A scientific archive for research, education, and conservation." https://www.macaulaylibrary.org/ (Last viewed January 1, 2023).

McGregor, P., and Krebs, J. (**1984**). "Sound degradation as a distance cue in great tit (*Parus major*) song," Behav. Ecol. Sociobiol. **16**(1), 49–56.

Naguib, M., and Wiley, H. (**2001**). "Estimating the distance to a source of sound: Mechanisms and adaptations for long-range communication," Anim. Behav. **62**(5), 825–837.

Piña-Covarrubias, E., Hill, A., Prince, P., Snaddon, J., Rogers, A., and Doncaster, C. (**2019**). "Optimization of sensor deployment for acoustic detection and localization in terrestrial environments," Remote Sens. Ecol. Conserv. **5**(2), 180–192.

Price, M., Attenborough, K., and Heap, N. (**1988**). "Sound attenuation through trees: Measurements and models," J. Acoust. Soc. Am. **84**(5), 1836–1844.

Priyadarshani, N., Marsland, S., and Castro, I. (**2018**). "Automated birdsong recognition in complex acoustic environments: A review," J. Avian Biol. **49**(5), jav-01447.

Shaw, T., Müller, S., and Scherer-Lorenzen, M. (**2022**). "Slope does not affect autonomous recorder detection shape: Considerations for acoustic monitoring in forested landscapes," Bioacoustics **31**(3), 261–282.

Shonfield, J., and Bayne, E. (**2017**). "Autonomous recording units in avian ecological research: Current use and future applications," Avian Conserv. Ecol. **12**(1), 42–54.

Snell-Rood, E. (**2012**). "The effect of climate on acoustic signals: Does atmospheric sound absorption matter for bird song and bat echolocation?," J. Acoust. Soc. Am. **131**(2), 1650–1658.

Sugai, L., Desjonquères, C., Silva, T., and Llusia, D. (**2020**). "A roadmap for survey designs in terrestrial acoustic monitoring," Remote Sens. Ecol. Conserv. **6**(3), 220–235.
Sugai, L., Silva, T., Ribeiro, J., and Llusia, D. (**2019**). "Terrestrial passive acoustic monitoring: Review and perspectives," BioScience **69**(1), 15–25.
Xeno-canto Foundation (**2005**). "Xeno-canto: Sharing wildlife sounds from around the world," http://xeno-canto.org (Last viewed January 1, 2023).
Yip, D., Bayne, E., Solymos, P., Campbell, J., and Proppe, D. (**2017a**). "Sound attenuation in forest and roadside environments: Implications for avian point-count surveys," Condor **119**(1), 73–84.
Yip, D., Leston, L., Bayne, E., Sólymos, P., and Grover, A. (**2017b**). "Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data," Avian Conserv. Ecol. **12**(1), 194–214.