forum acusticum 2026
8-12 September · Graz, Austria

SOUND SYNTHESIS AND PHYSICS-BASED DATA AUGMENTATION FOR THE DETECTION OF EURASIAN LYNX CALLS

Guillaume Dutilleux[1], Vegard Skagestad[1,2], Marius Brodersen[1,3]

[1] *Acoustics Group, Department of Electronic Systmes, NTNU - Norwegian University of Science and Technology, Trondheim, Norway*
[2] *Now with Q-Free ASA, Trondheim, Norway*
[3] *Now with Elliptic Labs, Oslo, Norway*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

In Scandinavia, the Eurasian lynx (*Lynx lynx*) is a hunted species whose populations must be monitored. We present a deep-learning-based software detector of E. lynx calls. It can be applied to long-term audio recordings carried out during the breeding season where the species produce loud and far-reaching calls. To our knowledge, this is the first published software detector for this species. Considering the large data sets required by deep learning, to compensate for the dearth of E. lynx recordings, we harnessed sound synthesis and data augmentation. Starting from a statistical analysis of the handful of available recordings of lynx calls, we designed a spectrogram synthesizer based on additive sound synthesis. The diversity in the synthetic data set was further extended using a subset of an engineering outdoor noise prediction model to simulate the received spectrograms as a function of topography, source-receiver geometry and other site-specific parameters. A one-channel neural network inspired from LeNet-5 was trained on this synthetic data. The E. lynx detector trained from scratch achieved 95.4 % accuracy. This novel approach of repurposing noise prediction schemes should be valuable in other applications of supervised machine learning for a wide range of outdoor sounds.

Keywords: *data augmentation, data synthesis, deep learning, Eurasian lynx, acoustic monitoring*

1. Introduction

The Eurasian lynx (*Lynx lynx*) is a carnivore species with a wide distribution area from Western Europe to East Asia [1], [2]. In Norway, the Eurasian lynx is a hunted species with a small population of about 420 individuals in 2023 [3]. This makes it vulnerable to rapid extinction [4]. Considering the situation of the species, monitoring is important.

Passive acoustic monitoring with field recorders is an established technique providing information in a fast and cost-effective way [5]. However, in long-term recordings animal vocalizations might only occur sporadically, and thereby only make up a tiny fraction of the total recording time. Therefore, automated processing of the recordings is necessary to extract the animal vocalizations of interest.

In the case of the Eurasian lynx, since the species is both rare, elusive and not very vocal outside a short breeding season, there are very few published recordings of lynx calls at the time of writing. In this paper, we describe the design of a deep-learning-based detector for the most common vocalization of the Eurasian lynx. Our aim was to design a custom neural network and to train it from scratch. Owing to the shortage of available recordings, our approach was to synthesize spectrograms and to harness data augmentation . In machine learning, data augmentation helps avoid overfitting, compensates for imbalanced data sets and improves the capacity of models to generalize [6]. In contrast to the application-agnostic approaches taken in the literature on data augmentation [6], [7] except occasionally in indoor spaces [8], here we wanted to test physics-based data augmentation by harnessing outdoor sound propagation as a parametric filter.

The remainder of this paper is organized as follows. Chapter 2 describes the target calls for the species of interest, the production of a dataset and the design of a detector. Chapter 3 presents the results of spectrogram synthesis and of the detector. These results are discussed in chapter 4. Chapter 5 concludes this paper.

***Corresponding author**: guillaume.dutilleux@ntnu.no.*

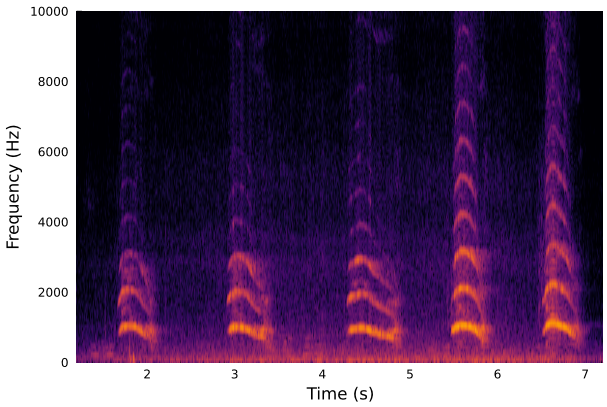

Figure 1: Spectrogram of a typical Eurasian lynx mew sequence. Source: *Tierstimmenarchiv*- [10] : `TSA:Lynx_lynx_S_921_1_1`.

2. Materials and methods

2.1. Target vocalization

The Eurasian lynx has four different calls [9]. Among them, the *mew* is the most common call found in sound libraries. Mews occur especially during the mating season, from February to April [1]. The mews are powerful long-distance calls that are produced both by the male and the female. They are the most likely vocalizations to be detected by a passive recorder for the species. Therefore, the mew was chosen as the target of the detector. Arguably the mew can be described as a superposition of multiple frequency-modulated harmonics. Each typically resemble an "inverted U" in the spectrogram with some degree of asymmetry (Figure 1). Mews are typically produced in sequence.

2.2. Audio data sources

2.2.1. Existing data

Recordings of Eurasian lynx calls were contributed by 6 persons / organizations in Europe (Czech Republic, France, Germany) in February 2022. They represent 143 mews and 250 s of vocalizations in total. Assuming a window length of 4 seconds, this represents less than 60 distinct spectrograms. This is insufficient to train a deep learning model from scratch.

The spectrograms of these recordings were analyzed in Audacity (The Audacity Team, 1999-2026) release 3.7) to extract statistics on (1) the number of mews per call, (2) the duration of a mew, (3) the mew rate within a call, the (4) minimum and (5) maximum frequency of the fundamental and (6) the number of harmonics. 46 mews could not be used to estimate anything else than the number of mews per call and the call rate, because of clipping, poor signal-to-noise ratio, or because they were obviously taken at large distance, and thereby altered by sound propagation.

When designing the model, there were very few recordings of Eurasian lynx calls available in the on-line nature sound archives. Six recordings could be retrieved from the MNHN *Sonothèque* (Paris, France), two from the *Tierstimmenarchiv* (Berlin, Germany). The full potential of the latter could not be exploited because the archive was shut down due to a cyberattack for most of our project.

Since it was desirable to simulate different signal-to-noise ratios, we used existing recordings (PCM, 16-bit resolution, 32 kHz sampling frequency) made in Trondheim at the outskirts of a small forest in October 2023 with an Audiomoth (Open Acoustics Devices, Oxford, UK) as a source of background noise.

2.2.2. Data collection

To build the validation and the test sets we recorded real lynx calls at Namsskogan *Familiepark* (Trones, Norway), a resort that hosted three Eurasian lynxes (two females and one male). A Wildlife Acoustics (Maynard, MA, USA) type SM4 acoustic recorder was installed on a tree about 5 m away from the enclosure. Continuous recording (PCM, 16 bits, 44.1 kHz) was carried out in February-March 2022 and 2023. Recordings were also made at Langedrag *Dyrepark* (Nesbyen, Norway), a zoo that hosted nine E. lynxes placed in two enclosures. A Frontier Labs (Yeerongpilly, Queensland, Australia) type Bar-LT recorder was attached to one of the fences delineating an enclosure and recorded continuously (PCM, 16 bits, 44.1 kHz) from 18:00 to 10:00 UTC+1 daily in the late winter 2023. Both enclosures were within the detection range of the recorder.

2.3. Spectrogram synthesis

A mew was modelled analytically as follows:

$$m(t) = A(t) \sum_{i=0}^{n} a_i \sin(\phi_i(t)),\ t \in [t_0, t_1] \qquad (1)$$

where $t_0 = 0$ s without any loss of generality and $0.34 \leq t_1 \leq 0.64$ s. Moreover, the tuple $(a_i, \phi_i(t) = (i+1)\phi_0(t))$ with relative amplitude and phase at instant t define the i-th harmonic. In Eq. (1), the phase of the fundamental ϕ_0 was defined by a third-order polynomial of t that was fitted to the spectrogram of the fundamental of a mew.

The amplitude modulation function $A(t)$ was defined as a piecewise linear function according to the classical *attack-decay-sustain-release* ADSR model used in analog musical synthesizers [11]. The function was fitted to the typical envelope of a mew that features a relatively short attack-decay phase followed by a longer sustain-release one. As implied by Eq. (1), the same ADSR window was applied on all harmonics.

As expected and visible in Fig. 1, the energy of the harmonics is a decreasing function of frequency. The decay pattern defined by $a_i, i \in \{1, \ldots, N\}$ was fitted at the center of a single mew. $n = 14$ was used in Eq. (1). While higher harmonics do exist in E. lynx mews at short range, they are above the frequency range of the spectrogram used by our detector as an input.

Eq. (1) corresponds to a single mew. This mew will be repeated N times to form a complete E. lynx call. This can be obtained by a simple convolution

$$x(t) = \sum_{j=1}^{N} m(t) * \delta\left(t - ((j-1)T + \Delta_j)\right) \quad (2)$$

where $N \in \{1, 2, 3\}$ for a 4-s window is the number of mews, $\delta()$ is the Dirac distribution, T is the average period between the starts of two successive mews and Δ_j is a random variable that accounts for deviations from the perfect periodicity T in the sequence of mews. For the possible values of N, $T > 1.3$ s.
A sampling rate of 22 kHz was assumed. The computed spectrograms (buffer size 512 with 50% overlap) span over 4 s in the time domain and between 0 and 10 kHz in the frequency one.

2.4. Data augmentation

Many methods have been developed to predict outdoor sound propagation [12], [13]. It is beyond the scope of this paper to review them. Considering that animal communication occurs in complex environments and that computation time should be minimized if it desirable to generate a large amount of augmented data, engineering noise prediction methods like the ISO 9613-2 standard [14] seem to be the most suitable ones.

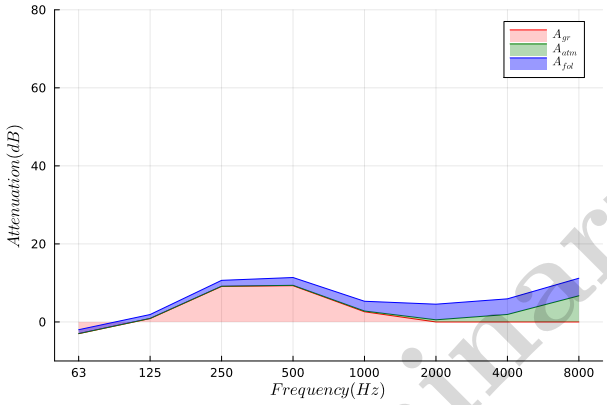

Figure 2: Components of A_{excess} 50 m from the source. Source height: 0.5 m, receiver height: 2 m, temperature: 10°C, relative humidity: 60%. Propagation on flat absorbing ground in a forest.

A call will be produced from a point S in an outdoor environment. It will be received at another point R. According to ISO 9613-2, further neglecting directivity of the animal's vocal organ, and assuming that S and R are in line of sight, the signal will be received with the following total *excess attenuation* A_{excess} in dB for the k-th octave band f_k:

$$A_{\text{excess}}(f_k) = A_{\text{atm}}(f_k) + A_{\text{gr}}(f_k) + A_{\text{fol}}(f_k) \quad (3)$$

where A_{atm} is the atmospheric attenuation, A_{gr} the attenuation due to ground reflections and A_{fol} the attenuation due to propagation through foliage. In this model, attenuation depends on source and receiver height, horizontal distance, ground absorption,

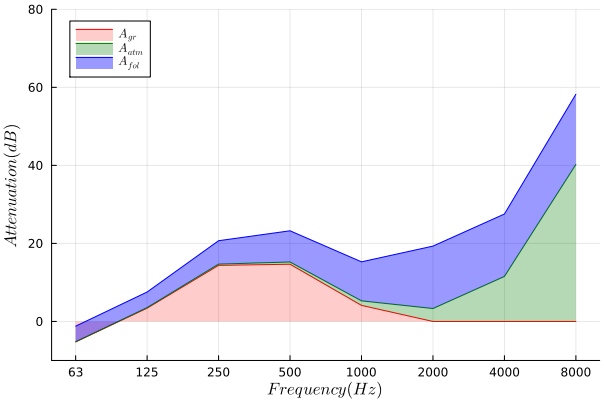

Figure 3: Components of A_{excess} 300 m from the source. The other parameters are unchanged compared to Fig. 2.

relative humidity, temperature. It also depends on propagation distance below the canopy, if any. A_{gr} is responsible for the so-called *gound dip* in the left half of the spectrum. A_{atm} is a strongly increasing function of frequency. Fig. 2 and 3 illustrate the three components of A_{excess} in Eq. (3) 50 m and 300 m away from the source.

All these parameters were modelled as uniform random variables within specified boundaries. The ground was assumed to be flat. For any given set of parameters, the attenuation is a piecewise constant function of frequency. Only the excess attenuation was considered in data augmentation because spherical divergence does not depend on frequency and the detector takes normalized spectra as inputs. A_{excess} is subtracted from the spectrogram of the synthetic call obtained according to the procedure described in section 2.3.

To further augment the data, background noise was taken into account. The recordings collected in Trondheim (see section 2.2.1) were used to produce spectrograms at five different signal-to-noise ratios from -10 to 10 dB in 5-dB steps.

In addition to this physics-based data augmentation `ColorJitter` offered by the PyTorch API was also used. `ColorJitter` randomly changes the brightness, contrast, saturation and hue of the spectrograms.

2.5. Machine learning

2.5.1. Data sets

The *training* set for the Eurasian lynx detector consisted of 2444 noise spectrograms (*non lynx*) and 2444 spectrograms of synthesized E. lynx calls. The *validation* set consisted of 470 spectrograms computed from real-world recordings. 234 are labeled as *lynx* and 236 as *not lynx*, resulting in a balanced data set. Most of the recordings are from the Namsskogan 2022 campaign (see section 2.2.1). The *test* set is perfectly balanced, containing 357 spectrograms in each class. All the E. lynx detector test set spectrograms were taken from the Langedrag 2023 campaign (Sec. 2.2.1).

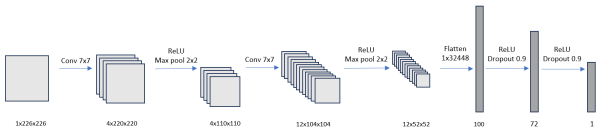

Figure 4: Architecture of the Eurasian lynx detector.

2.5.2. Model architecture

The machine learning model we designed is a relatively shallow convolutional neural network composed of two convolutional layers, three linear layers, two max pooling layers and two dropout layers. It is inspired by LeNet-5 [15]. Figure 4 displays the model architecture. The first layer is a 2D convolutional layer with one input channel, as the spectrograms are converted to 220×220 grayscale images in the pipeline. The final layer features a single output and uses Binary Cross Entropy With Logits Loss. This is a utility from PyTorch which combines a sigmoid activation function and the binary cross entropy loss function. Therefore, there is no need for implementing a separate activation function after the final layer. Our custom model depends on 3254809 trainable parameters.

2.5.3. Training

Each model was trained using three different seeds from a random number generator to improve reproducibility when doing hyperparameter tuning. The training set was shuffled during training. In other words, the order of images fed into the model was different for each epoch without seeds. For each new combination of hyperparameters, the machine learning model has been trained three times, each time with a given seed. Combinations which provided good results for all of the three seeds were deemed more promising. During hyperparameter tuning, the metrics that were monitored are accuracy on the validation set, loss on the training set and loss on the validation set to spot hyperparameters that provided best results. Different kernel sizes for the convolutional layers were tested, ranging from 3x3 to 7x7. The latter worked best. When it comes to the optimizer, AdamW with a weight decay coefficient [16] outperformed Adam.

Epochs:	5 - 20
Batch size:	32
Dropout rate:	0.9
Convolution kernel size:	7x7
Optimizer:	AdamW

Table 1: Final hyperparameters of the machine learning model after tuning.

3. Results

3.1. Spectrogram synthesis

Figure 5 shows one example of a synthesized spectrogram for the Eurasian lynx, according to the procedure described in section 2.3. See [17] for more examples.

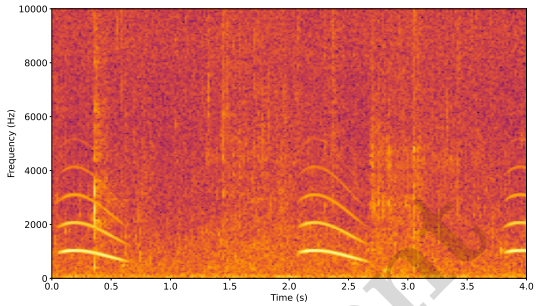

Figure 5: Synthesized fragment of E. lynx call with 10 dB SNR.

3.2. Detector performance

Table 2 displays three hyperparameter configurations from the experimental approach for the E. lynx detector, with corresponding accuracy percentage on the validation set for each seed.

Batch size:	32	32	32
Dropout rate:	0.5	0.9	0.9
Optimizer:	Adam	Adam	AdamW
Learning rate:	0.001	0.001	0.0011
Weight decay:	0	0	0.01
Conv. kernel size:	5x5	5x5	7x7
Val. accuracy seed 1:	84.0	86.0	93.4
Val. accuracy seed 2:	77.7	87.2	80.4
Val. accuracy seed 3:	78.5	84.0	86.8

Table 2: Three configurations from the experimental approach for the model.

The third model has the best validation accuracy, and also the best accuracy on the test set. This is a good result, because it shows that there is a correlation between validation accuracy and test accuracy, and therefore also a correlation between the validation set and the test sets. Similar validation and test set is important to obtain effective hyperparameter tuning. Figure 6 displays the confusion matrix for the best model on the test set. The number of false positives for the third model is extremely low. Almost all the errors are false negatives.

4. Discussion

4.1. Data generation

When comparing the synthesized spectrograms to real ones, the most visible difference is the occasional presence of a reverberation tail in the real calls. In addition, the real spectrograms typically contain energy between the harmonics. This is not accounted for by

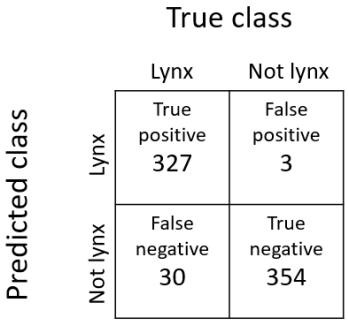

Figure 6: Confusion matrix for the best model on the test set. Accuracy: 95.4 %. Precision: 0.99. Recall: 0.92.

our synthesizer. In that respect, the additive synthesis is a limitation here. There is not enough space here to present a thorough comparison between spectrograms of real and synthetic calls. Fig. 1 and 5 are merely one example of each. More details can be found in [17]. The performance of the detector suggests that the synthetic data used for training the model accounts for the variability in the real calls.

While it would not have been possible to train a model from scratch based solely on the E. lynx audio data available at the time of developing the model, and although the approach combining both spectrogram synthesis and physics-based data augmentation taken here seems to be successful, further work should evaluate the relative contribution of synthesis and augmentation based on a model for attenuation of sound outdoors by carrying out ablation studies, preferably in a scenario where the size of the training set is kept constant. A comparison of our approach with more common audio augmentation techniques should also be carried out.

Although Eq. (1) is expressed in the time domain, it is more appropriate to write about *spectrogram synthesis* than about *sound synthesis*, since the modeling was guided by the desire to mimic spectrograms, because they are the input expected by the detector. Even after fitting the model's parameters on real calls the typical synthetic call in the time domain was a far cry from a real E. lynx sequence of mews. Turning to physical modeling of sound generation would likely lead to more realistic sounds. However, since the input of our model is spectrograms where by definition phase information is lost, physical modeling seems like overkill.

For data augmentation, we relied on a method that delivers attenuation spectra in octave bands. Although this method was simple to implement the decomposition in octave bands is arguably quite coarse when it comes to filtering bioacoustic signals. Further work could consider switching to Nord2000 [18] because of the higher sensitivity of this method to a larger number of physical parameters. This would be beneficial for the diversity of the generated datasets.

4.2. Performance, hyperparameters and data

Table 2 shows tremendous variability in validation accuracy from one seed to another. This is likely attibutable to the limited variability in our purely synthetic training set compared to real calls. A number of parameters are fixed in the synthesis of spectrograms, like the sweep pattern or harmonic decay. The size of the data set may also be a limitation here.

Moreover, a theoretical investigation shows the interaction between neural network size, dropout and learning rates [19]. According to this research, with the value we took for the learning rate, the use of dropout leads to training that is likely to depend on the initial conditions. Other recent research [20] observes that the specialization of neurons will only occur if the dropout rate is significantly smaller than the values we considered when training our neural network. This could also have compromised the stability of the performances.

When evaluating the machine learning performance, it is important to have in mind what kind of data the models was actually trained on. It would be desirable to train the models on data containing vocalizations from not only E. lynx, but also other sympatric species whose vocalizations have spectrograms that are close to those of E. lynx, like some representative of the *Corvus* genus.

We used synthesized data for the E. lynx vocalizations because there was not enough recorded data of these animals available to constitute a proper-sized training set. Even though the Eurasian lynx is rare, it would be challenging to collect enough data for sympatric species as well. Furthermore, developing synthesizers for sympatric species that are likely to appear would also be time consuming. Therefore, considering only two classes was a pragmatic choice. The detector could be used in parallel with classifiers like BirdNET [21] to help mitigate occasional false positives.

5. Conclusion

In this paper, we presented a deep-learning-based Eurasian lynx detector. To cope with the lack of labeled data in this elusive species, we developed dedicated spectrogram synthesis based on a statistical analysis of a small number of recordings and heuristic analytical modeling of the call of the target species.

The training data was further expanded through data augmentation. Here, augmentation was based on a subset of a repurposed engineering noise prediction model. This made it easy to simulate many different realistic outdoor source-receiver scenarios. The LeNet-5-inspired model that we trained from scratch achieved 95.4% accuracy on the test set.

When it comes to designing detectors and classifiers for outdoor sounds, both data synthesis and physics-based data augmentation as described here seem to be relevant approaches for both man-made and biological sounds. Many animal vocalizations in birds and mammals can be approached by additive synthesis. It may be beneficial to replace the engineering noise

prediction method used here by a more advanced one.

References

[1] V. G. Heptner and A. A. Sludskij, *Die Saügetiere der Sowjetunion - Band III: Raubtiere (Feloidea)*, V. G. Heptner and N. P. Naumov, Eds. Jena, Germany: Veb Gustav Fischer Verlag, 1980.

[2] K. Schmidt, M. Ratkiewicz, and M. K. Konopiński, "The importance of genetic variability and population differentiation in the Eurasian lynx (*Lynx lynx*) for conservation, in the context of habitat and climate change," *Mammal Review*, vol. 41, no. 2, pp. 112–124, Mar. 2011. DOI: 10.1111/j.1365-2907.2010.00180.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1365-2907.2010.00180.x.

[3] M. Tovmo, J. Odden, and E. B. Nilsen, *Antall familiegrupper, bestandsestimat og bestandsutvikling for gaupe i Norge i 2023*, 2023. [Online]. Available: https://hdl.handle.net/11250/3070345, (Accessed: 15.07.26).

[4] J. D. C. Linnell, H. Broseth, J. Odden, and E. Birkeland Nielsen, "Sustainably harvesting a large carnivore? Development of Eurasian lynx populations in Norway during 16 years of shifting policy," *Environmental Management*, vol. 45, pp. 1142–1154, Mar. 2010. DOI: 10.1007/s00267-010-9455-9.

[5] S. Hoefer, D. T. McKnight, S. Allen-Ankins, E. J. Nordberg, and L. S. and, "Passive acoustic monitoring in terrestrial vertebrates: A review," *Bioacoustics*, vol. 32, no. 5, pp. 506–531, 2023. DOI: 10.1080/09524622.2023.2209052. [Online]. Available: https://doi.org/10.1080/09524622.2023.2209052.

[6] S. Wei, S. Zou, F. Liao, and w. lang weimin, "A comparison on data augmentation methods based on deep learning for audio classification," *Journal of Physics: Conference Series*, vol. 1453, no. 1, p. 012 085, Jan. 2020.

[7] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, p. 101 084, 2020.

[8] C. Kim et al., "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google home," in *Interspeech 2017*, 2017, pp. 379–383.

[9] G. Peters, "Acoustic communication in the genus lynx (mammalia: Felidae) - comparative survey and phylogenetic interpretation," *Bonner zoologischer Beiträge*, vol. 38, pp. 315–330, 1987.

[10] Tierstimmenarchiv, *Animal Sound Archive*. [Online]. Available: https://www.tierstimmenarchiv.de/webinterface/?language=english, (Accessed: 17.09.23).

[11] K. Beeman, *Digital Signal Analysis, Editing, and Synthesis*. Springer, 1998.

[12] K. Attenborough and T. van Renterghem, *Predicting outdoor sound*. CRC Press, 2021.

[13] D. A. Bies, C. H. Hansen, and C. O. Howard, *Engineering noise control*, 5th. CRC Press, 2018.

[14] ISO, "Acoustics, Attenuation of sound during propagation outdoors - Part 2: General method of calculation," International Organization for Standardization, Geneva, CH, Standard ISO 9613-2, 1996. [Online]. Available: https://www.iso.org/standard/20649.html.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: https://doi.org/10.1109/5.726791.

[16] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017. arXiv: 1711.05101. [Online]. Available: http://arxiv.org/abs/1711.05101.

[17] M. Brodersen and V. Skagestad, "Machine learning for classifying lynx and Arctic fox calls sound synthesis and physics-based data augmentation," M.S. thesis, NTNU, Trondheim, Norway, Jun. 2024.

[18] B. Plovsing, "Proposal for nordtest method: Nord2000 - Prediction of outdoor sound propagation," DELTA, Hørsholm, DK, Tech. Rep., Jan. 2014.

[19] L. Chizat, P. Marion, and Y. Yesbay, *Phase diagram of dropout for two-layer neural networks in the mean-field regime*, 2025. arXiv: 2510.07554 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2510.07554.

[20] F. Mori and F. Mignacco, *Analytic theory of dropout regularization*, 2025. arXiv: 2505.07792 [stat.ML]. [Online]. Available: https://arxiv.org/abs/2505.07792.

[21] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101 236, 2021, ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2021.101236. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121000273.