Noise Modeling for Bioacoustics with Contrastive Predictive Coding

Vipul Sharma Randall Balestriero

Abstract

Bioacoustics data suffers from a lack of fine-grained labels and high noise. We propose a method based on Contrastive Predictive Coding (CPC) (Oord et al., 2019) that aims for noise robustness by contrastive learning with stochastic noise augmentations of the audio, and avoids the need for labels. Further, we modify CPC to enforce consistency across time in the latent space like (Bardes et al., 2024). We test our method on the Cornell Bird Challenge dataset (Howard et al., 2020; Gupta et al., 2021) with data of 100 bird species. Our method performs poorly compared to Supervised Learning due to a limited variety of audio augmentations caused by the engineering limitations of scaling the augmentations.

1. Introduction

Bioacoustics is the study of animal communication, generally by passive acoustic monitoring (PAM). It has wide applications ranging from the analysis of coral reef health (Lin et al., 2023; 2021), species conservation efforts (Teixeira et al., 2019; Penar et al., 2020), and the study of language evolution across species (Meyer, 2004; Fitch, 2017).

The increase in high-quality and cheap passive audio recording tools has led to large-scale public bioacoustics datasets like (Vellinga & Planqué, 2015). This allows for the application of AI models that have enjoyed some success (Kahl et al., 2021; Williams et al., 2024).

But such methods are fundamentally limited by the nature of the data and niche bioacoustics tasks requiring specialized knowledge. Progress in the image and text domains has been due to large, carefully curated datasets (Deng et al., 2009; Liu et al., 2024). Even when using internet-scale datasets, image and text modality have natural labels that can be extracted from the data or obtained for cheap by human annotation. Further, the downstream tasks for vision and language models can be easily evaluated by humans. Bioacoustics data, on the other hand, suffers from high noise, sample variety, and low label information. The bioacoustics tasks often require training and aren't amenable to annotation or evaluation by humans, even if there were an economic interest, which is rare.

This poses a challenge for traditional AI methods. Our work seeks to address this problem by building on advances in self-supervised learning for audio and bioacoustics denoising (Oord et al., 2019; Ryan et al.; Barnhill et al., 2024; Zhang & Li, 2022; Kahl et al., 2021; Moummad et al., 2024).

We propose a methodology that uses stochastic noise augmentations and contrastive learning for a model to learn without labels and become invariant to noise.

2. Motivation

Our methodology is grounded in two ideas drawn from the literature. Firstly, recent work has shown that the human auditory system stores noise characteristics over time, instead of learning to filter it out (Hicks & McDermott, 2024). Such noise models are hypothesized to persist lifelong and might be crucial for auditory scene analysis and the segregation of background and foreground noise.

(Dapello et al., 2021) also found that stochastic injection of Gaussian noise makes models of the human auditory system more robust and aligned with the human auditory system. They posit that noise and stochasticity are integral to biologically realistic models of the human sensory-perceptual system.

Based on this, our method proposes learning a noise model by contrastive learning with different stochastic realizations of natural background noise across the positive views of a sample.

Secondly, as shown in Figure 1 from (Dubova & Sloman, 2023), literature on the training dynamics of Deep Neural Networks argues for a direct dependence between the model's ability to generalize and its sensitivity to noise. Such work claims that as a model "overgeneralizes" and moves from "memorization" of the training data to "extrapolation" from the data, it becomes invariant to irrelevant features (”noise”), and equivariant to relevant features (”con-

Code available at: https://github.com/rbalestr-lab/SSL-Bioacoustics/tree/main.

Experimental data used for Table 1 available at: https://wandb.ai/vipul/bioacoustics/table.

cepts"), decreasing the need to memorize each realization of such features and being able to compose them (Ito et al.; Wang et al., 2021; Dunion et al., 2023).

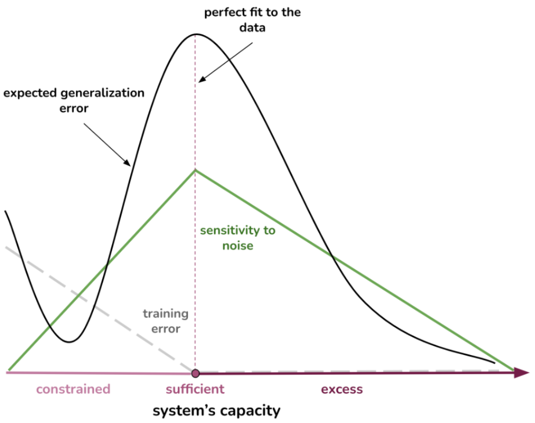


Figure 1. The relationship between noise sensitivity and training dynamics. Figure is from (Dubova & Sloman, 2023).

Our methodology uses this observation and attempts to accelerate the generalization of a model and its robustness to noise by using noise as an augmentation and hence a "feature" that a contrastive learning model can become invariant to by design. We hope that such explicit noise modeling would reduce the need for a large scale of data with its inherent variety, for the model to learn noise invariance and concept equi-variance. This could lead to sample-efficient learning with an intelligent design of "noise" augmentations in contrastive self-supervised learning.

3. Methods

3.1. Bird Song Dataset

Most bird song datasets are sourced from Xeno-Canto (Vellinga & Planqué, 2015), which is the largest public repository of bioacoustics data across thousands of bird, frog, grasshopper, bat, and land mammal species. We use a subset of this dataset that was first used for the Cornell Bird Challenge (CBC) 2020 (Howard et al., 2020), and then filtered by (Gupta et al., 2021). The authors filter this dataset to balance the number of samples per species and to ensure each sample is of enough length. Sadly, they don't provide a preprocessed dataset, but only the procedure. Hence, our version of the dataset differs from theirs, even though we followed the same procedure.

The original CBC dataset contains data for 264 bird species. To ensure a balanced dataset, we only retain species with at least 100 samples of audio. Further, we filtered out the samples shorter than 7 seconds to ensure completeness of the bird calls. This led to 100 bird species of data. Unlike (Gupta et al., 2021), who have 15,032 samples in their dataset, we have 9470 samples across the 100 bird species.

This dataset was split into train, validation, and testing splits in the ratio of 80/10/10, giving us 7534 samples for the train, 939 for the validation, and 997 for the testing split.

3.2. Preprocessing

We use the first 3 seconds of a sample for pretraining, except for the "static supervised" method in Table 1, which used the first 7 seconds. Note that (Gupta et al., 2021) used 7 seconds of audio across all methods, but as discussed in the limitations and challenges section, that wasn't feasible due to engineering challenges for our methods that are pretrained on sequential data.

The first 3 seconds of audio were resampled at 32 KHz, and then a sliding window of size 1 second, and a hop length of 0.5 seconds was applied to create 5 overlapping patches per audio sample. Each patch was converted to a mel-spectrogram with 128 mel filter banks, a window size for Fast Fourier Transform of 2048, and a hop length of 512. The "static supervised" method also followed the same preprocessing procedure except for the sliding window step. It had a single spectrogram corresponding to 7 seconds of an audio sample.

Further, each complex-valued mel-spectrogram was processed into an RGB image after conversion to a power spectrogram and normalizing to decibels with the maximum power of each spectrogram as the reference.

3.3. Noise Dataset for Stochastic Augmentation

Inspired by BirdNet (Kahl et al., 2021), we create a noise dataset for augmentation during training from the training dataset's non-salient clips.

The training dataset was filtered based on the Xeno-Canto rating of audio quality (A-E, with A being the best, and E being the worst), and for a length of at least 14 seconds. The last 7 seconds of the audio were used as noise samples to avoid overlap between the pretraining input and the noise augmentation.

Further, to enrich this noise dataset, we manually scraped Xeno-Canto for soundscape recordings containing ambient sounds. Most noise in bird song datasets consists of rain, wind, and waterfalls. We scraped Xeno-Canto for remarks containing this information. The scraped audio was also split into 7-second non-overlapping samples.

We thus obtained a novel natural noise dataset containing 97 samples across the soundscape and training samples. To the best of our knowledge, no such standard dataset exists for bird songs.

For stochastic noise augmentation, we randomly select from the noise dataset and apply the noise to the training sample in the waveform representation before converting the training data to a spectrogram as detailed earlier. Figure 3 shows an example of noise augmentation with different SNR values. Note that the noise acts as a background noise, and doesn't interfere with the bird call itself, which can be seen in the waveform and the pitch representation of the signal.

3.4. Pretraining Method

As seen by the lack of clustering of labels in Figure 2, there isn't an apparent structure in the training data even in the spectrogram representation. Note that each point in the cluster represents a non-overlapping window of the audio samples. This implies that K-Means wasn't able to learn the temporal structure across an audio sample, as it's not even able to cluster the different windows of the same audio sample together.

This observation led to our choice of Contrastive Predictive Coding (CPC) (Oord et al., 2019) as a pretraining method, which promises to learn temporal information across time steps in audio signals.

Further, following the recent success in learning temporal information in the latent space (Bardes et al., 2024), we modified CPC to maximize similarity across time steps in the embedding space of a projector network. As shown in Figure 4, for each time step, we maximize similarity with k future time steps.

The future time steps with an audio sample act like positive samples, and the same time steps for other samples in our batch act like negative samples for our sample-based self-supervised pretraining loss.

This leads to a natural formulation of the InfoNCE loss, as in (Oord et al., 2019):

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{1}$$

3.5. Model Architecture

Following (Gupta et al., 2021), to allow fair comparison across methods, we used a CNN-LSTM architecture for our encoder, and an MLP for our projector network.

The CNN-LSTM consists of 9 convolution layers with max pooling and batch normalization, followed by 2 LSTM layers. The projector network consists of a single MLP layer with an output embedding dimension of 128.

All the methods were trained for 75 epochs with an Adam optimizer, and a batch size of 32 with a learning rate of 0.0001, same as (Gupta et al., 2021).

4. Results

Table 1 contains the results of 4 types of experiments that we ran. Firstly, we use a baseline of ResNet-18 and ResNet-50 (He et al., 2015) trained on the spectrogram representation of 7 seconds of audio. This represents the information that can be learned without any explicit autoregressive modeling to extract the temporal structure in the data. Even though ResNets don't have any autoregressive components, the input representation is a spectrogram with time as the x-axis, allowing for extraction of temporal information by a convolutional neural network which explains the high accuracy.

Then, we train our CNN-LSTM on spectrograms of sliding windows of 3 seconds audio sample. Surprisingly, even when trained on less than half the length of the static supervised method, the autoregressive CNN-LSTM matches the accuracy of the static supervised method, with less than half the number of model parameters.

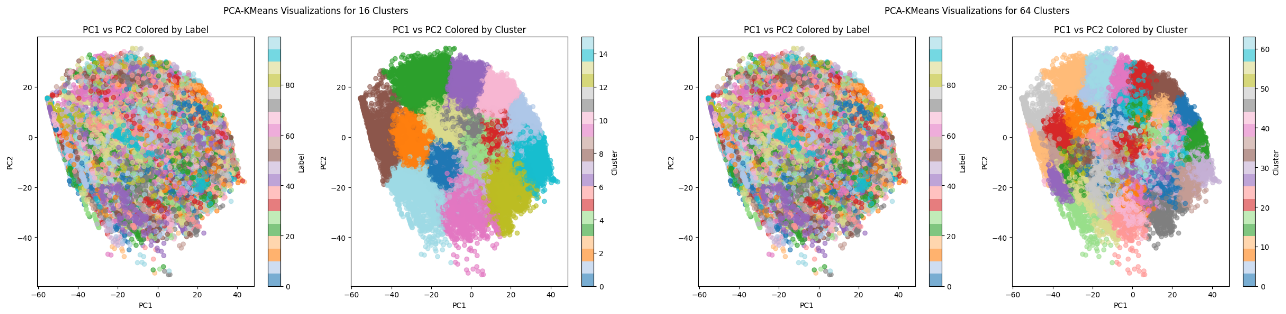
After establishing a supervised learning baseline, which differs from the baseline of (Gupta et al., 2021) given our dataset differences, we performed SSL experiments with our modified CPC methodology. The SSL experiments used the same augmentations as the supervised experiments except for the addition of noise in the waveform representation of the training audio samples.

As elaborated in the limitations and challenges section, due to a lack of supported versatile augmentations, CPC performs poorly compared to supervised learning. CPC needs more ablations with different augmentations to match supervised learning performance.

Interestingly, we see a meaningful trend in the performance of CPC with noise augmentation. It is reasonable to assume that the training dataset has a lot of noise. We also see the evidence of this in the poor K-Means clustering in Figure 2 where there are no clusters based on the class label, or the whole audio sample. Hence, we used stochastic noise as an augmentation with SNR of 10 and 0.1 to learn noise invariant representations.

As more noise was added, the performance on validation and test set decreases. This indicates our stochastic noise augmentation doesn't reflect the noise distribution of the validation and testing splits. We need more ablations to verify this, and create a better noise augmentation that reflects the noise in the data. Perhaps we would benefit from a controlled noise ablation like (Hendrycks & Dietterich, 2019) where we first manually control the amount of deterministic noise in the dataset to evaluate the performance of stochastic noise augmentation.

Sadly, as discussed in the limitations section, there are fundamental challenges which need to be overcome before such



(a) K=16

(b) K=64

Figure 2. K-Means clustering of the first 2 Principal Components of the flattened training spectrograms. The left panel shows clusters colored by true label, and the right panel shows coloring by cluster ID. K-Means could not extract any class label information from the spectrograms.

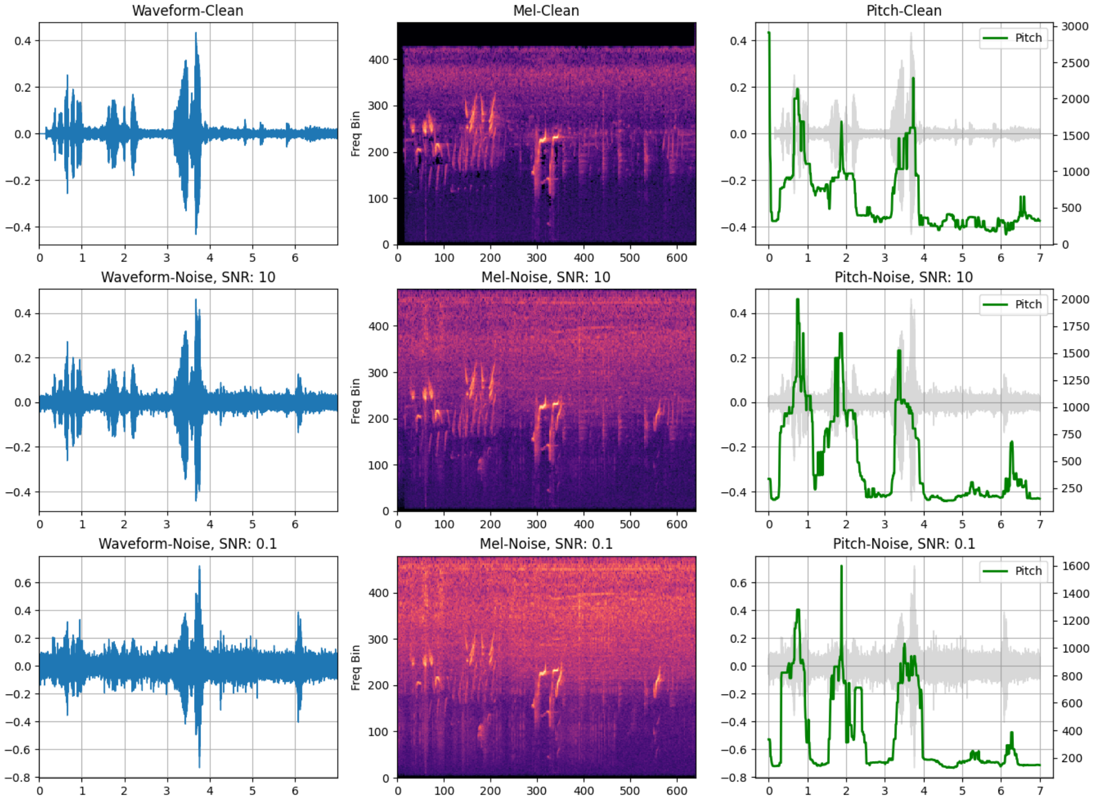


Figure 3. The waveform, mel-spectrogram, and pitch for a training example. The top row corresponds to clean, the middle row to noise augmentation with SNR=10, and the bottom row to noise augmentation with the same noise but SNR=0.1.

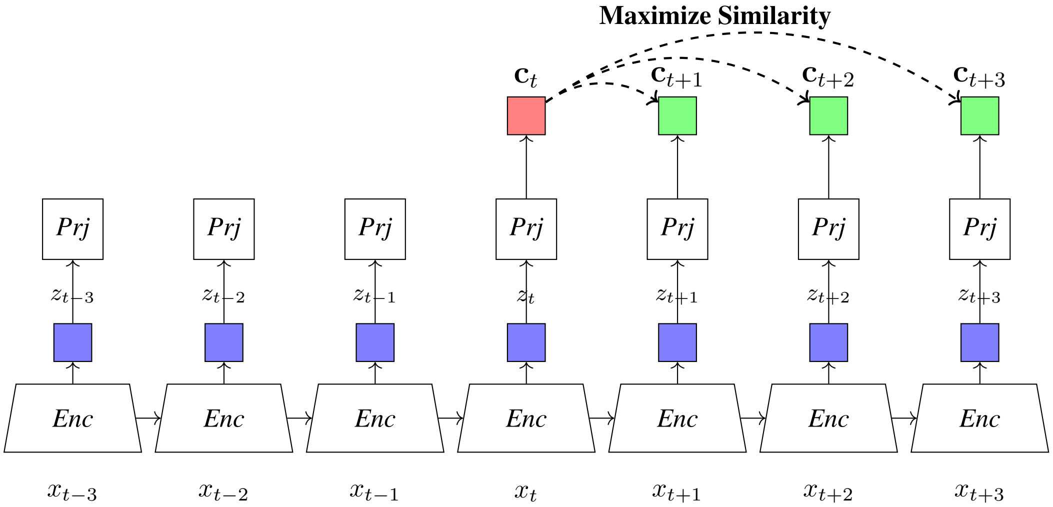


Figure 4. Our modifications on Contrastive Predictive Coding (CPC). We use a CNN-LSTM for the encoder (*Enc*), and an MLP for the projector (*Prj*). Every time step acts as the context (c_t) for the next k time steps, where k is window size -1.

ablations can be run.

5. Limitations and Challenges

The performance of Self-Supervised pre-training clearly lags behind supervised pre-training. Unfortunately, unlike image and text modality data, audio data still has a lot of engineering challenges that need to be overcome before the experiments can be scaled in a university setting or with constrained resources.

When doing autoregressive modeling with audio data, the preprocessing and data augmentation operations are heavily compute-bound. For example, when we use a 7-second-long audio input with a sliding window size of 0.5s and a hop size of 0.25s, each epoch takes around 20 minutes to run with n=7534 samples and our CNN-LSTM model; the bottleneck being the data preprocessing and data augmentation stage. This leads to unreasonable pretraining times in a university setting.

TorchAudio recognizes this difficulty and offers CUDA support for the transforms (tor), but PyTorch is fundamentally limited in CUDA support within its data loader. The PyTorch data loaders need to be rewritten to support CUDA operations in the forked data loading workers so that we can benefit from TorchAudio CUDA support. Any other workaround leads to under-utilization of the multiprocessing capabilities of the data loader, and hence increased training time due to memory access latency, and serial compute operations of TorchVision transforms, which only run on CPU and per-sample.

Once the fundamental engineering problems in PyTorch are solved, we can utilize many more data augmentations to increase SSL performance. Currently, our modified CPC method is unable to learn good data representations by contrastive learning. This is highly likely due to a lack of variety in data augmentations.

Even augmentation with noise was insufficient to extract representation from data, as the noise is mostly ambient sounds like a waterfall, which doesn't alter the signal fundamentally but just adds white noise-like distortion on the spectrogram.

Beyond data augmentations, we could explore different SSL methods for acoustics data like Shuffle and Learn (Misra et al., 2016). But again, such pretraining methods are currently unsuitable for use with long audio inputs because of engineering limitations.

6. Future Work

A clear line of future work is to experiment with more time and frequency augmentations in our SSL pipeline. SpecAugment (Park et al., 2019) is a popular augmentation in bioacoustics that could be used once the engineering challenges are overcome.

Further, we could use UrbanSound and VROOM datasets (Gupta, 2023) as a source for noise augmentations to increase the variety of augmentations in our pretraining. A benchmark like ImageNet-C (Hendrycks & Dietterich, 2019) could be established for bioacoustics with different types and levels of noise corruption by taking advantage of the taxonomy of noise created by UrbanSound.

(Barnhill et al., 2024) has also shown that a major challenge in designing animal-independent bioacoustics models is

Table 1. Linear probing accuracies for bird species classification on the validation (n=939) and the test dataset (n=997). Results are aggregated across 3 seeds. The models were trained on 3-second audio samples (n=7534), except for static supervised, which was trained on spectrograms of 7-second audio samples.

Method	Model	Val Acc (Top-1)	Val Acc (Top-5)	Test Acc (Top-1)	Test Acc (Top-5)
Static Supervised (7s)	ResNet-18	23.94%	47.42%	22.81%	46.73%
Static Supervised (7s)	ResNet-50	**29.00**%	**52.82**%	27.18%	51.28%
Sequential Supervised	CNN-LSTM	26.81%	51.44%	**27.31**%	**52.38**%
CPC	CNN-LSTM	1.58%	7.72%	1.30%	6.73%
CPC with Noise SNR=10	CNN-LSTM	1.23%	6.43%	1.16%	6.26%
CPC with Noise SNR=0.1	CNN-LSTM	1.02%	5.52%	1.06%	5.14%

to learn unique representations of noise corresponding to each species. Each species has its own range of sound frequencies for communication, and they inhabit a wide variety of environments with their unique noise sources. This makes creating a foundational bioacoustics model a big challenge. Even within birds, the lack of a common model in the annual iterations of BirdClef (noa) is evidence of the challenging nature of the problem.

(Zhang & Li, 2022) provides a denoising dataset with ground truth of noise in bird call spectrograms as a mask. This allows a feature attribution analysis (Fucci et al., 2025) of our trained models to analyze their sensitivity to noise by calculating overlap in the attribution map and the ground truth noise and signal maps. Explainable-AI for audio modality is an underexplored field with fundamental open questions because of the versatility of data type and representation, unlike image and text (Fucci et al., 2024). Analyzing the noise robustness of speech models is a good step in this direction, as it's applicable irrespective of the task or domain of the audio data.

References

BirdCLEF+ 2025. URL https://kaggle.com/birdclef-2025.

Supported Features — Torchaudio 2.5.0.dev20241105 documentation. URL https://docs.pytorch.org/audio/main/supported_features.html.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. Revisiting feature prediction for learning visual representations from video, 2024. URL https://arxiv.org/abs/2404.08471.

Barnhill, A., Noeth, E., Maier, A., and Bergler, C. ANIMAL-CLEAN – A Deep Denoising Toolkit for Animal-Independent Signal Enhancement. In *Interspeech 2024*, pp. 632–636. ISCA, September 2024. doi: 10.21437/Interspeech.2024-1151. URL https://www.isca-archive.org/interspeech_2024/barnhill24_interspeech.html.

BizhuWu. BizhuWu/ShuffleAndLearn_pytorch, April 2024. URL https://github.com/BizhuWu/ShuffleAndLearn_PyTorch. original-date: 2021-06-01T10:03:06Z.

Cauzinille, J., Favre, B., Marxer, R., Clink, D., Ahmad, A. H., and Rey, A. Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures. In *Interspeech 2024*, pp. 132–136. ISCA, September 2024. doi: 10.21437/Interspeech.2024-1096. URL https://www.isca-archive.org/interspeech_2024/cauzinille24_interspeech.html.

Dapello, J., Feather, J., Le, H., Marques, T., Cox, D. D., McDermott, J. H., DiCarlo, J. J., and Chung, S. Neural population geometry reveals the role of stochasticity in robust perception, 2021. URL https://arxiv.org/abs/2111.06979.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dubova, M. and Sloman, S. J. Excess Capacity Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45), 2023. URL https://escholarship.org/uc/item/49w48008.

Dunion, M., McInroe, T., Luck, K. S., Hanna, J. P., and Albrecht, S. V. Temporal Disentanglement of Representations for Improved Generalisation in Reinforcement Learning, February 2023. URL http://arxiv.org/abs/2207.05480. arXiv:2207.05480 [cs].

Fitch, W. T. Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review*, 24 (1):3–33, February 2017. ISSN 1531-5320. doi: 10.3758/s13423-017-1236-5. URL https://doi.org/10.3758/s13423-017-1236-5.

Fucci, D., Savoldi, B., Gaido, M., Negri, M., Cettolo, M., and Bentivogli, L. Explainability for speech models: On the challenges of acoustic feature selection. In Dell'Orletta, F., Lenci, A., Montemagni, S., and Sprugnoli, R. (eds.), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pp. 373–381, Pisa, Italy, December 2024. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL https://aclanthology.org/2024.clicit-1.45/.

Fucci, D., Gaido, M., Savoldi, B., Negri, M., Cettolo, M., and Bentivogli, L. Spes: Spectrogram perturbation for explainable speech-to-text generation, 2025. URL https://arxiv.org/abs/2411.01710.

Ghani, B., Denton, T., Kahl, S., and Klinck, H. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1): 22876, December 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-49989-z. URL https://www.nature.com/articles/s41598-023-49989-z. Publisher: Nature Publishing Group.

Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., and Ferres, J. L. Comparing recurrent convolutional neural networks for large scale bird species classification. *Scientific Reports*, 11(1):17085, August 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-96446-w. URL https://www.nature.com/articles/s41598-021-96446-w. Publisher: Nature Publishing Group.

Gupta, P. pranavgupta2603/SimCLR-UrbanSound8K, December 2023. URL https://github.com/pranavgupta2603/SimCLR-UrbanSound8K. original-date: 2023-12-16T12:00:15Z.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

Hicks, J. M. and McDermott, J. H. Noise schemas aid hearing in noise. *Proceedings of the National Academy of Sciences*, 121(47):e2408995121, 2024. doi: 10.1073/pnas.2408995121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2408995121.

Howard, A., Klinck, H., Dane, S., Kahl, S., tom denton, and Denton, T. Cornell birdcall identification. https://kaggle.com/competitions/birdsong-recognition, 2020. Kaggle.

Huzaifah, M. Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks, June 2017. URL http://arxiv.org/abs/1706.07156. arXiv:1706.07156 [cs].

Igarashi, T., Saito, Y., Seki, K., Takamichi, S., Yamamoto, R., Tachibana, K., and Saruwatari, H. Noise-Robust Voice Conversion by Conditional Denoising Training Using Latent Variables of Recording Quality and Environment. In *Interspeech 2024*, pp. 2750–2754. ISCA, September 2024. doi: 10.21437/Interspeech.2024-972. URL https://www.isca-archive.org/interspeech_2024/igarashi24_interspeech.html.

Ito, T., Klinger, T., Schultz, D. H., Murray, J. D., Cole, M. W., and Rigotti, M. Compositional generalization through abstract representations in human and artificial neural networks.

Kahl, S., Wood, C. M., Eibl, M., and Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, March 2021. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2021.101236. URL https://www.sciencedirect.com/science/article/pii/S1574954121000273.

Kumar, S., Li, J., and Zhang, Y. Vision Transformer Segmentation for Visual Bird Sound Denoising. In *Interspeech 2024*, pp. 122–126. ISCA, September 2024. doi: 10.21437/Interspeech.2024-1412. URL https://www.isca-archive.org/interspeech_2024/kumar24_interspeech.html.

Lim, C.-y., Shin, H.-s., Kim, J.-h., Heo, J., Koo, K.-W., Kim, S.-b., and Yu, H.-J. Improving Noise Robustness in Self-supervised Pre-trained Model for Speaker Verification. In *Interspeech 2024*, pp. 2665–2669. ISCA, September 2024. doi: 10.21437/Interspeech.2024-1630. URL https://www.isca-archive.org/interspeech_2024/lim24_interspeech.html.

Lin, T.-H., Akamatsu, T., Sinniger, F., and Harii, S. Exploring coral reef biodiversity via underwater soundscapes. *Biological Conservation*, 253:108901, 2021. ISSN 0006-3207. doi: https://doi.org/10.1016/j.biocon.2020.108901. URL https://www.sciencedirect.com/science/article/pii/S0006320720309599.

Lin, T.-H., Sinniger, F., Harii, S., and Akamatsu, T. Using soundscapes to assess changes in coral reef social-ecological systems. *Oceanography*, issue$_volume, March 2023. URL.$

Liu, Y., Cao, J., Liu, C., Ding, K., and Jin, L. Datasets for large language models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2402.18041.

Meyer, J. Bioacoustics of human whistled languages: an alternative approach to the cognitive processes of language. *Anais da Academia Brasileira de Ciências*, 76:406–412, June 2004. ISSN 0001-3765, 1678-2690. https://doi.org/10.1590/S0001-37652004000200033. URL https://www.scielo.br/j/aabc/a/C7c5W8SQmKmGxgdV4T5JbRJ/?lang=en. Publisher: Academia Brasileira de Ciências.

Misra, I., Zitnick, C. L., and Hebert, M. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification, July 2016. URL http://arxiv.org/abs/1603.08561. arXiv:1603.08561 [cs].

Moummad, I., Serizel, R., and Farrugia, N. Self-Supervised Learning for Few-Shot Bird Sound Classification, January 2024. URL http://arxiv.org/abs/2312.15824. arXiv:2312.15824 [cs] version: 3.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pp. 2613–2617, September 2019. 10.21437/Interspeech.2019-2680. URL http://arxiv.org/abs/1904.08779. arXiv:1904.08779 [eess].

Penar, W., Magiera, A., and Klocek, C. Applications of bioacoustics in animal ecology. *Ecological Complexity*, 43:100847, 2020. ISSN 1476-945X. https://doi.org/10.1016/j.ecocom.2020.100847. URL https://www.sciencedirect.com/science/article/pii/S1476945X19301606.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. WHISPER: Robust Speech Recognition via Large-Scale Weak Supervision, December 2022. URL http://arxiv.org/abs/2212.04356. arXiv:2212.04356 [eess].

Runia, T. F. H., Snoek, C. G. M., and Smeulders, A. W. M. Repetition Estimation. *International Journal of Computer Vision*, 127(9):1361–1383, September 2019. ISSN 1573-1405. 10.1007/s11263-019-01194-0. URL https://doi.org/10.1007/s11263-019-01194-0.

Ryan, P. SingingData/Birdsong-Self-Supervised-Learning, March 2022. URL https://github.com/SingingData/Birdsong-Self-Supervised-Learning. original-date: 2020-06-16T02:52:46Z.

Ryan, P., Takafuji, S., Yang, C., Wilson, N., and McBride, C. Using Self-Supervised Learning of Birdsong for Downstream Industrial Audio Classification.

Salamon, J., Jacoby, C., and Bello, J. P. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, Orlando Florida USA, November 2014. ACM. ISBN 978-1-4503-3063-3. 10.1145/2647868.2655045. URL https://dl.acm.org/doi/10.1145/2647868.2655045.

Sheikh, M. U., Abid, H., Shafique, B. S., Hanif, A., and Khan, M. H. Bird Whisperer: Leveraging Large Pre-trained Acoustic Model for Bird Call Classification. In *Interspeech 2024*, pp. 5028–5032. ISCA, September 2024. 10.21437/Interspeech.2024-1623. URL https://www.isca-archive.org/interspeech_2024/sheikh24_interspeech.html.

Teixeira, D., Maron, M., and van Rensburg, B. J. Bioacoustic monitoring of animal vocal behavior for conservation. *Conservation Science and Practice*, 1(8):e72, 2019. ISSN 2578-4854. 10.1111/csp2.72. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/csp2.72. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/csp2.72.

Torrence, C. and Compo, G. P. A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, January 1998. ISSN 0003-0007, 1520-0477. 10.1175/1520-0477(1998)079¡0061:APGTWA¿2.0.CO;2. URL http://journals.ametsoc.org/doi/10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2.

Ullah, A., Ragano, A., and Hines, A. Reduce, Reuse, Recycle: Is Perturbed Data Better than Other Language Augmentation for Low Resource Self-Supervised Speech Models. In *Interspeech 2024*, pp. 77–81. ISCA, September 2024. 10.21437/Interspeech.2024-396. URL https://www.isca-archive.org/interspeech_2024/ullah24_interspeech.html.

Vellinga, W.-P. and Planqué, R. The xeno-canto collection and its relation to sound recognition and classification. In *Conference and Labs of the Evaluation Forum*, 2015. URL https://api.semanticscholar.org/CorpusID:18544013.

Wang, T., Yue, Z., Huang, J., Sun, Q., and Zhang, H. Self-Supervised Learning Disentangled Group Representation as Feature. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18225–18240. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/97416ac0f58056947e2eb5d5d253d4f2-Abstract.html.

Williams, B., van Merriënboer, B., Dumoulin, V., Hamer, J., Triantafillou, E., Fleishman, A. B., McKown, M., Munger, J. E., Rice, A. N., Lillis, A., White, C. E., Hobbs, C. A. D., Razak, T. B., Jones, K. E., and Denton, T. Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics, 2024. URL https://arxiv.org/abs/2404.16436.

Zhang, Y. and Li, J. BirdSoundsDenoising: Deep Visual Audio Denoising for Bird Sounds, October 2022. URL http://arxiv.org/abs/2210.10196. arXiv:2210.10196 [cs].