Crossing the Species Divide: Transfer Learning from Speech to Animal Sounds

Jules Cauzinille[1], *Marius Miron*[2], *Olivier Pietquin*[2], *Masato Hagiwara*[2], *Ricard Marxer*[3], *Arnaud Rey*[2], *Benoit Favre*[1]

[1]Aix-Marseille University, France [2]Earth Species Project, USA
[3]Université de Toulon, France

***Abstract*—Self-supervised speech models have demonstrated impressive performance in speech processing, but their effectiveness on non-speech data remains underexplored. We study the transfer learning capabilities of such models on bioacoustic detection and classification tasks. We show that models such as HuBERT, WavLM, and XEUS can generate rich latent representations of animal sounds across taxa. We analyze the models' properties with linear probing on time-averaged representations. We then extend the approach to account for the effect of time-wise information with other downstream architectures. Finally, we study the implication of frequency range and noise on performance. Notably, our results are competitive with fine-tuned bioacoustic pre-trained models and show the impact of noise-robust pre-training setups. These findings highlight the potential of speech-based self-supervised learning as an efficient framework for advancing bioacoustic research.**

***Index Terms*—Computational bioacoustics, Transfer learning, Self-supervision, Linear probing**

1. INTRODUCTION

Self Supervised Learning (SSL) of speech features has led to state-of-the-art performances on linguistic and paralinguistic tasks [1]. SSL methods leverage large volumes of unlabeled data by solving general-purpose pretext tasks and learn contextual and meaningful acoustic representations from the inherent structure of speech. One valid question, given the massive amounts of available speech data, is whether these performances extend to the comparatively under-resourced analysis of non-human animal vocalizations. Recent research has shown that SSL speech models can reach high performance on a range of tasks such as species recognition [2], caller identification [3], or call-type classification [4], [5]. Many such studies have investigated a specific selection of these models with different perspectives, methods, and datasets but lack a comprehensive understanding of how pre-trained speech representations transfer across tasks and taxa.

Drawing from previous work, we thus propose a formal study of speech features' transferability:

- We benchmark speech models on a set of 10 different bioacoustic tasks, spanning a diversity of taxa, including birds, mammals, and insects.
- We explore three comparable speech models, showing that robustness to noise and multilingual pre-training are key aspects of bioacoustic transferability.
- We compare a set of probing approaches, including linear probing as well as recurrent neural networks, and introduce time-weighted averaging of representations to leverage their contextual nature.
- We confirm the superiority of shallow transformer layers over deeper ones in the context of speech features transferability.
- We further analyze the effect of background noise and vocal frequency ranges on a selected dataset.

2. METHODS

This study unfolds as a set of bioacoustic knowledge transfer experiments from three pre-trained SSL speech models. Regardless of the selected downstream approach, we initially extract pre-trained representations from each layer of frozen speech models. These representations are subsequently given as train, development, and test sets to a downstream model. The observed performances are taken as a proxy of the speech model's ability to extract the acoustic information necessary to accomplish the task (see Figure 1). We extend this method, referred to as linear probing [6], to both non-linear and time-wise analysis of pre-trained representations through other downstream approaches described in Section 2.3. The code is available for replication on GitHub at https://github.com/jcauzi/speech2bio.git.

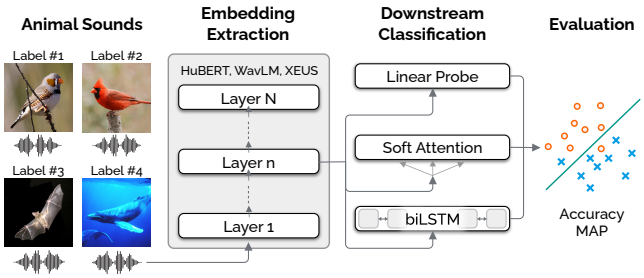

Fig. 1: Workflow of the transfer learning method

2.1. Tasks and datasets

The 11 datasets used in this experiment are publicly available as the BEANS benchmark [7]. They include bioacoustic classification and detection tasks, namely "Watkins" [8] (31 aquatic mammal species classification), "CBI" [9] (264 bird species classification), "Bats" [10] (10 individuals Egyptian fruit bats classification), "Humbug" [11] (14 mosquito species classification), "Dogs" [12] (10 individuals dogs classification), "Dcase" [13] (20 bird and mammal species detection), "Enabirds" [14] (34 bird species detection), "Hiceas" [15] (Minke whale detection), "Gibbons" [16] (three gibbon call-types detection), "RFCX" [17] (24 bird and frog species detection) and "ESC-50" (50 environmental and animal sounds classification).

For classification, each sample is assigned one or more labels, and evaluation is reported in terms of accuracy. Detection implies identifying subsections of interest and their labels from long recordings with a sliding window approach. Each produced segment may contain multiple labels. Detection evaluation is reported in terms of Mean Average Precision (mAP). All datasets are first down-sampled to the required 16 kHz sample rate of pre-trained speech models. They are split into predefined train, development, and test sets with a 6:2:2 ratio, as in [7]. Although all tasks are not directly comparable due to the nature and balance of their labels, dataset sizes, or acoustic environments, they allow a broad overview of speech models' performances in an array of bioacoustic scenarios.

2.2. Pre-trained models

SSL pre-training consists in leveraging large non-labeled data to pre-train foundation models optimized to extract general-purpose latent representations. With or without fine-tuning, these representations can

be given to so-called downstream models, or classification heads, to achieve competitive performance on tasks with small labeled datasets [18]. In speech processing, HuBERT [19], WavLM [20], and XEUS [21] are SSL models originally designed for downstream speech recognition. Additionally, they show high performance on paralinguistic tasks such as speaker identification or emotion recognition, among others [1]. Both HuBERT and WavLM were successfully implemented in similar bioacoustic experiments [2]–[5], demonstrating performance that matches or surpasses that of bioacoustic models such as AVES [22], as well as pre-trained bird species classifiers [23] and general-purpose audio taggers. Further details on the three models' architectures are presented in Table 1.

Table 1: Pre-trained model specificities

Model	Pre-train Data	Arch.	Task	#param.
HuBERT large	60k hours English only	24 TL	MP	317M
WavLM large	94k hours English only	24 TL	MP+NR	90M
XEUS	1M+ hours 4k+ languages	18 EBL	MP+NR +Rev	577M

TL = transformer layers, EBL = E-Branchformer layers, MP = masked prediction, NR = noise robustness, Rev = reverberation robustness.

2.3. Downstream setups

We use linear probing [24], i.e., training a shallow downstream model on a given layer of the SSL models, to assess the linear separability of bioacoustic information in their latent representation space. Our first experiment involves training a linear layer on **Time-Averaged** (T-A) representations as

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t, \tag{1}$$

where $x_t \in \mathbb{R}^d$ is the embedding of the t-th time frame in the full representation (all model representations are 2-D tensors with embeddings of size 1024 for each 20-millisecond frame). This method allows extracting a single input vector of size 1024 for each example in the dataset and is a preferred option in similar transfer learning experiments [1]. We hypothesize that averaging latent representations, particularly in such out-of-domain contexts, might end up diluting relevant time-wise information, potentially leading to suboptimal performance for long sound segments.

In a second experiment, we thus introduce **Time-Weighted Averaging** (T-WA), a form of soft attention that consists of adding a layer of learnable attention weights before averaging all frame representations to better leverage their variability over time. Raw attention scores are obtained as

$$\alpha_t = w_\alpha^\top x_t + b_\alpha, \tag{2}$$

where $w_\alpha \in \mathbb{R}^d$ and $b_\alpha \in \mathbb{R}$ are learned parameters. We then compute a weighted average of feature representations as

$$\alpha' = \text{softmax}_T(\alpha), \tag{3}$$

$$\tilde{x} = \sum_{t=1}^{T}(\alpha'_t)x_t, \tag{4}$$

where α_t determines the contribution of each frame. This weighted representation is given to a linear layer as in the Time-Averaged (T-A) version. The attention weights and the linear probe are trained simultaneously on a single loss. Note that this only adds 1024+1 parameters to the downstream model compared to T-A, no matter the input sound length.

All training procedures were conducted on predefined dataset splits from [7] for 100 epochs. Hyper-parameters such as batch size are empirically determined for each dataset by comparing metrics on the validation set. We test three learning rates (1e-5, 5e-5, and 1e-4) with an Adam optimizer and weight decay for each experiment and display the best observed result for the best layer. Classification tasks are conducted with cross entropy loss and evaluated in terms of accuracy. Multi-label detection tasks are conducted with BCE loss and evaluated in terms of mAP. Random baselines are computed by selecting the highest possible score between predictions of only the most frequent label (for highly unbalanced datasets) or random predictions.

3. RESULTS

Table 2 shows accuracy and mAP of the best performing layer on the validation set for each dataset, model, and linear probing approach, as well as the results from fully fine-tuned AVES bio [22], a Vggish model, and the best supervised CNN-based approach taken from the BEANs benchmark [7]. Underlined results inform on performance gains between our solution and previous BEANs results.

The results show the intrinsic ability of pre-trained speech representations to encode enough information to resolve bioacoustic tasks across taxa with linear probing. All models show variable results across layers. Depending on the dataset, performances increase in the initial layers, reach a maximum between layer 3 and 11 for HuBERT, layer 4 and 15 for WavLM, and layer 2 and 6 for XEUS, then consistently drop in deeper layers. This behavior was previously observed in similar experiments [2], [3], [5] and is consistent with speech models' layer-wise acoustic information encoding as described in [25]. Both WavLM and XEUS show higher probing performance than HuBERT, which we discuss in Section 4.1. Time-Weighted Averaging (T-WA) representations marginally improve results compared to T-A ones.

In such cross-species transfer learning experiments, one could make the hypothesis that the phylogenetic proximity of a species with humans results in similarities between the vocalizations of said species and speech. This can be based on the resemblance of vocal tracts, anatomical sizes, vocal frequencies, degrees of sequentiality, etc. Although the tasks presented here are difficult to compare due to differences in dataset sizes, annotations, label distribution, and acoustic environments, there seems to be no effect of species proximity with humans on the observed performances. Pre-trained speech models show a general ability to encode bioacoustic information regardless of the species.We thus extend this experiment to account for the effect of time, noise, and pre-training setup in an attempt to better understand the mechanisms underlying such results.

4. DISCUSSION AND PERSPECTIVES

We identify four factors that may account for performance variations between models and datasets. The effect of background noise is discussed in Sections 4.1, and 4.3, species vocalization overlap is investigated in Section 4.1, time-related factors in Section 4.2 and frequency range mismatch with speech in Section 4.3.

4.1. Model comparison

WavLM and XEUS include noise robustness solutions by being pre-trained with dynamic mixing of artificial noise and speech overlap (as well as simulated reverberant conditions for XEUS). This improves resilience to background noise, a significant limiting factor in bioacoustic tasks, since animal recordings typically have much lower sound-to-noise ratios than speech datasets. Although all three

Table 2: Probing Results

	accuracy					mAP					accuracy
Datasets	Watkins	CBI	Bats	Humbug	Dogs	Dcase	Enabirds	Hiceas	Gibbons	RFCX	ESC-50
HuBERT T-A	0.89	0.28	0.62	0.76	0.79	0.32	0.48	0.54	0.19	**0.14**	0.72
WavLM T-A	**0.90**	0.34	0.68	0.79	0.84	0.33	**0.51**	0.54	0.18	0.10	0.76
XEUS T-A	0.89	0.25	0.73	**0.79**	0.86	0.30	0.44	0.54	0.26	0.08	**0.81**
HuBERT T-WA	0.85	0.37	0.74	0.75	0.90	0.33	0.44	0.58	0.23	0.13	0.57
WavLM T-WA	0.84	**0.38**	0.76	0.74	0.93	**0.39**	0.44	**0.64**	**0.32**	0.12	0.65
XEUS T-WA	0.78	0.28	**0.77**	0.75	**0.95**	0.32	0.47	0.52	0.28	0.10	0.68
AVES bio ft	0.88	0.60	0.75	0.81	0.95	0.39	0.56	0.63	0.28	0.13	0.77
Vggish	0.85	0.44	0.74	0.81	0.91	0.34	0.54	0.46	0.15	0.14	0.77
CNN (best)	0.80	0.57	0.74	0.70	0.76	0.22	0.46	0.30	0.31	0.09	0.59
random	0.07	0.01	0.10	0.35	0.22	0.03	0.48	0.27	0.02	0.01	0.02

(ft: fine-tuning) Best speech model's layer performance is in **bold**, best performance between speech and other is underlined.

models show additional differences in their pre-training data (see Table 1), the noise robustness of WavLM and XEUS seems to play an important role in their higher performances. WavLM, in particular, with significantly fewer parameters, outperforms both other models in six out of the 11 tasks. Similarly to background noise, overlapping animal vocalizations may also harm performance in some datasets. For Enabirds and RFCX, the observed low performancescan be linked to a significant amount of interspecies vocalization overlap present in both datasets. The enabirds dataset is comprised of recordings of bird choruses, which can contain over six species labels in a single sound sample. Similarly, rfcx recordings are drawn from tropical forest environments with high multi-species presence. WavLM and XEUS pre-training with artificial mixup of two speech utterances seems to offer insufficient robustness to disentangle simultaneously vocalizing species. Conversely, pre-trained bioacoustic models such as Perch [26] and Birdnet [23] have been shown to increase multi-label classification performance with multi-species mixup training.

The multilingual pre-training of XEUS may also explain some of the performance improvements in 4 out of 11 tasks, as English-only pre-training datasets are inherently less acoustically variable than multilingual ones. However, this assumption would need further investigation, as XEUS is also trained on significantly more data compared to both other models.

4.2. Time information

Excluding enabirds and rfcx, two datasets showing relatively low performance in general, T-A representations consistently outperform T-WA on datasets with sound segments shorter than 3.98 seconds. T-A representations do not entirely remove time-wise information, as each frame representation still contains contextual information from previous and future frames, allowing good performance on short sound segments despite averaging. Yet, T-WA results indicate that relying on a learned selection of relevant sound frames according to their representation is an important feature for datasets containing longer segments.

To further analyze the effect of temporal information and linear separability on the observed performances, we replace linear probes with Echo State Networks (ESN) [27] or Bidirectional Long Short-Term Memory networks (biLSTM). Both can extract nonlinear information from entire representations without risking a loss of time-sensitive information. ESNs, derived from reservoir computing, mostly consist of an RNN with randomly initialized fixed weights. They are a good approximation of a recurrent non-linear solution that does not demand learning additional parameters compared to linear probing. BiLSTMs, on the other hand, significantly augment the number of parameters but should produce a strong topline solution with sufficient amounts of training data.

Interestingly, ESNs always showed worse or on-par performance compared to linear probing on all datasets. BiLSTMs showed some improvements for HuBERT humbug (0.77 accuracy with a two-layer biLSTM), HuBERT gibbons (0.26 mAP with a single-layer biLSTM), XEUS enabirds (0.5 mAP with a single-layer biLSTM) and XEUS gibbons (0.35 mAP with a single-layer biLSTM) but mostly underperformed linear probes from either XEUS or WavLM. In conclusion, linear probes applied to time-averaged or time-weighted latent representations seem to outperform more complex recurrent models in bioacoustic tasks, as they may strike a better balance between signal extraction and overfitting.

4.3. Frequency shift and noise addition

From elephant rumbles to bat calls, many animal vocalizations sit below or above the human speech range. We design an ablation experiment to account for the effects of frequency range on the Egyptian fruit bats dataset. Pre-trained speech models are trained on relatively clean recordings of speech at 16 kHz, thus ranging from 0 to 8,000 Hz, with most of the energy between 90 and 2,500 Hz. This could lead to some degree of overfitting to clean signals centered around this particular frequency range, hindering performance on higher-pitched or noisy signals. We investigate whether pitch downshifting of high-pitched bat calls results in improved performance by reducing their frequency towards the speech range. We progressively lower the pitch of the recordings at 16 kHz with factors of 0.125, 0.25, and 0.5 by resampling, then interpolating them with the `rate` function from `pysox`. Surprisingly, Figure 2 shows that this lowers the probe's accuracy, regardless of the model's layer used. We also observed this tendency in similar experiments not included in this manuscript: time-stretching as well as pitch-shifting the original 250 kHz samples. In these cases, adding high-frequency information from ranges above 8 kHz did not improve performance either. These results show that speech models may be robust to high frequencies, potentially through their modeling of fricatives. Additionally, our pitch shifting method involves resampling and interpolation, which may result in unnatural audio quality when performed at such extreme ratios.

We further explore noise robustness by mixing bat vocalizations with natural noise sources [28]–[32] at Signal-to-Noise Ratios (SNRs) of 0, –5, –10, and –20 dB. Note that these are extreme factors for speech applications but quite common in bioacoustics. As shown in Figure 3, the accuracy decreases with the SNR in a similar way for all three models. Interestingly, performance remains above chance even at extreme values such as –20 dB, likely because the added environmental

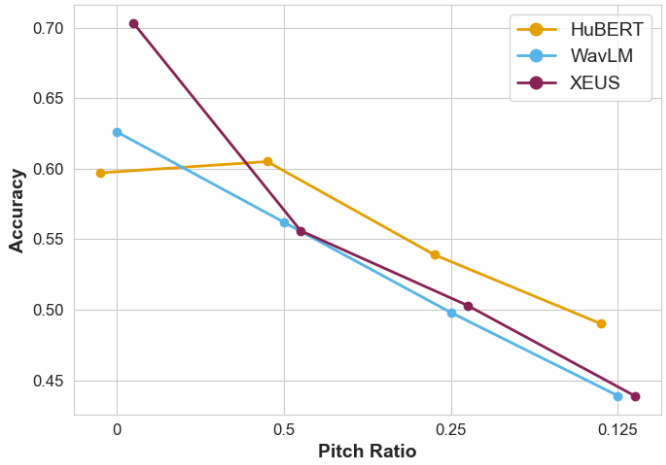

Fig. 2: Performance for the Egyptian fruit bats dataset on the 10th layer with pitch shifting (T-A)

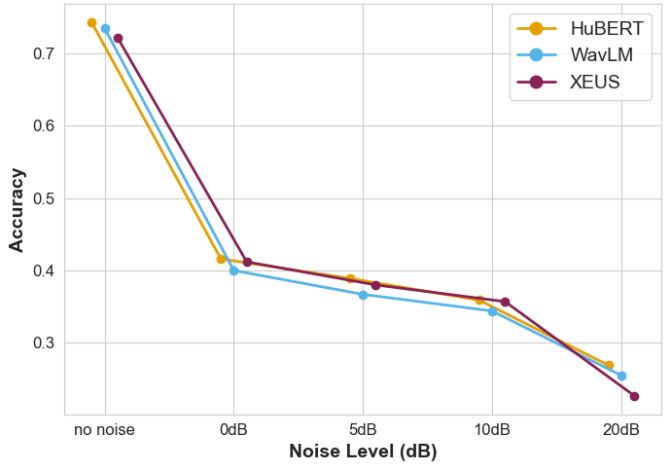

Fig. 3: Performance for the Egyptian fruit bats dataset on the 10th layer with noise addition (T-A)

noise does not fully mask the high frequencies of bat calls. However, since the original recordings are already noisy, further reducing SNRs yields severely degraded samples, suggesting that the robustness of speech models may reach its limit under such conditions.

Both these results indicate some degree of robustness of speech models to bioacoustic transfer learning while partly explaining the variability of their performance across scenarios. Low performance on datasets such as CBI, Gibbons, RFCX, and Enabirds could be explained both by the difficulty of the task or label distribution and because of the significantly low SNRs of the recordings.

4.4. Comparing pre-trained models with random initializations

To further examine the effect of pre-training on speech models' performance, we test the intuitive hypothesis stating that randomly initialized models should show random baseline performance (as only the pre-training phase should result in model parameters able to capture bioacoustic information). Interestingly, a randomly initialized HuBERT model achieves slightly better-than-chance performance on bioacoustic tasks (with an average of $\approx +10\%$) when evaluated using linear probing. While the model lacks any learned knowledge of speech or acoustic structure, this result aligns with a growing body of evidence suggesting that untrained deep learning models can provide non-trivial representations due to their inherent inductive biases. This also indicates that most of the discriminative capabilities of SSL models stem from their speech-based pretraining phase, although a small fraction of their effectiveness on bioacoustic tasks may be attributed to structural priors [33], [34].

4.5. Limitations

Comparing performance across datasets must be taken with care due to differences in dataset sizes, tasks, label distribution, etc. We limit our conclusions to relative improvements of specific methods rather than absolute performance differences. Additionally, the use of multiple random seeds would allow getting a stronger sense of performance gaps between models. Here, each training was performed once, from fixed dataset splits provided in the BEANS benchmark. We also advocate for a cautious comparison of our results with AVES-bio, which is based on a significantly smaller model (HuBERT base), pre-trained with in-domain data, and fully fine-tuned on each task. Although our experiments are not performance-oriented, we must mention that fine-tuning of large bioacoustic foundation models with comparable sizes generally outperforms our methods on most of the benchmark [35]–[37]. Conversely, task-specific machine learning solutions and CNNs trained on Mel Frequency Cepstral Coefficient representations are mostly outperformed by probing frozen speech models [7].

5. CONCLUSION

This study shows the effective knowledge transfer capabilities of self-supervised speech models for bioacoustic tasks. In particular, HuBERT, WavLM, and XEUS generate rich representations, resulting in competitive performance on classification and detection tasks across taxa. Our findings indicate that phylogenetic proximity to humans does not influence this transfer, emphasizing the generalizability of speech representations in out-of-domain scenarios. Variability in results across datasets highlights the impact of task complexity and acoustic environments. Notably, the presence of noise greatly affects linear probing results, although the use of noise-robust pre-training methods seems to show an advantage. Temporal information remains crucial, with time-weighted averaging approaches outperforming time-averaging of representations for longer audio samples. However, recurrent networks showed no benefits compared to linear probing, likely due to overfitting issues. Finally, speech models exhibit some degree of robustness to the frequency range of bioacoustic signals, even with high-pitched vocalizations. Overall, self-supervised speech models show strong promise for the development of foundation models in bioacoustics, offering interesting solutions related to data quality and quantity. Leveraging large speech datasets and SSL architectures, future work could focus on developing bioacoustic models partly pre-trained on speech with enhanced noise robustness and improved mixup strategies.

6. ACKNOWLEDGMENTS

This research was supported by the Convergence Institute ILCB (ANR-16-CONV-0002), the Excellence Initiative of Aix-Marseille University (A*MIDEX), the HEBBIAN ANR project (#ANR-23-CE28-0008) and funding from the Earth Species Project.

REFERENCES

[1] J. Shi, S.-H. Wang, W. Chen, M. Bartelds, V. Bannihatti Kumar, J. Tian, X. Chang, D. Jurafsky, K. Livescu, H. yi Lee, and S. Watanabe, "Ml-superb 2.0: Benchmarking multilingual speech models across modeling constraints, languages, and datasets," in *Proc. Interspeech*, 2024.

[2] E. Sarkar and M. Magimai.-Doss, "Comparing self-supervised learning models pre-trained on human speech and animal vocalizations for bioacoustics processing," in *Proc. ICASSP*, 2025.

[3] J. Cauzinille, B. Favre, R. Marxer, D. Clink, A. H. Ahmad, and A. Rey, "Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures," in *Proc. Interspeech*, 2024.

[4] M. de Heer Kloots and M. Knörnschild, "Exploring bat song syllable representations in self-supervised audio encoders," in *Proc. VIHAR*, 2024.

[5] E. Sarkar and M. M. Doss, "On the utility of speech and audio foundation models for marmoset call analysis," in *Proc. VIHAR*, 2024.

[6] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *Proc. ICLR*, 2017.

[7] M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian, "Beans: The benchmark of animal sounds," in *Proc. ICASSP*, 2023.

[8] L. Sayigh, M. A. Daher, J. Allen, H. Gordon, K. Joyce, C. Stuhlmann, and P. Tyack, "The watkins marine mammal sound database: An online, freely accessible resource," *Proc. Meetings on Acoustics*, 2017.

[9] A. Howard, H. Klinck, S. Dane, S. Kahl, and T. Denton, "Cornell birdcall identification," https://kaggle.com/competitions/birdsong-recognition, 2020, kaggle.

[10] Y. Prat, M. Taub, E. Pratt, and Y. Yovel, "An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny," *Scientific Data*, 2017.

[11] I. Kiskin, M. E. Sinka, A. D. Cobb, W. Rafique, L. Wang, D. Zilli, B. Gutteridge, T. Marinos, Y. Li, E. W. Kaindoa, G. F. Killeen, K. J. Willis, and S. Roberts, "HumBugDB: a large-scale acoustic mosquito dataset," in *Proc. NeurIPS*, 2021.

[12] S. Yin and B. McCowan, "Barking in domestic dogs: context specificity and individual identification," *Animal Behaviour*, 2004.

[13] V. Morfi, I. Nolasco, V. Lostanlen, S. Singh, A. Strandburg-Peshkin, L. F. Gill, H. Pamula, D. Benvent, and D. Stowell, "Few-shot bioacoustic event detection: A new task at the dcase 2021 challenge." in *Proc. DCASE*, 2021.

[14] L. M. Chronister, T. A. Rhinehart, A. Place, and J. Kitzes, "An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information," *Ecology*, 2021.

[15] NOAA Pacific Islands Fisheries Science Center, "Hawaiian islands cetacean and ecosystem assessment survey (hiceas) towed array data," in *Proc. DCLDE*, 2022.

[16] E. Dufourq, I. Durbach, J. P. Hansford, A. Hoepfner, H. Ma, J. V. Bryant, C. S. Stender, W. Li, Z. Liu, Q. Chen, Z. Zhou, and S. T. Turvey, "Automated detection of hainan gibbon calls for passive acoustic monitoring," *Remote Sensing in Ecology and Conservation*, 2021.

[17] J. LeBien, M. Zhong, M. Campos-Cerqueira, J. P. Velev, R. Dodhia, J. L. Ferres, and T. M. Aide, "A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network," *Ecological Informatics*, 2020.

[18] A. Mohamed, H. yi Lee, L. Borgholt, J. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE J. Sel. Top. Signal Process.*, 2022.

[19] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. Audio, Speech, Lang. Process.*, 2021.

[20] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, 2022.

[21] W. Chen, W. Zhang, Y. Peng, X. Li, J. Tian, J. Shi, X. Chang, S. Maiti, K. Livescu, and S. Watanabe, "Towards robust speech representation learning for thousands of languages," in *Proc. EMNLP*, 2024.

[22] M. Hagiwara, "Aves: Animal vocalization encoder based on self-supervision," in *Proc. ICASSP*, 2023.

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

[24] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," in *Proc. ICASSP*, 2021.

[25] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *Proc. ICASSP*, 2023.

[26] Google Research, "Perch." [Online]. Available: https://github.com/google-research/perch

[27] C. Sun, M. Song, D. Cai, B. Zhang, S. Hong, and H. Li, "A systematic review of echo state networks from design to application," *IEEE Trans. Artificial Intelligence*, 2024.

[28] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE Trans. Audio, Speech, Lang. Process.*, 2021.

[29] J. Salamon and J. Jacoby, Cand Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Multimedia*, 2014.

[30] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *Meetings on Acoustics*, 2013.

[31] J. Abeßer, S. Gourishetti, A. Kátai, T. Clauß, P. Sharma, and J. Liebetrau, "Idmt-traffic: an open benchmark dataset for acoustic traffic monitoring research," in *Proc. EUSIPCO*, 2021.

[32] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019.

[33] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Proc. Advances in Neural Information Processing Systems*, 2008.

[34] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[35] D. Robinson, A. Robinson, and L. Akrapongpisak, "Transferable models for bioacoustics with human language supervision," in *Proc. ICASSP*, 2024.

[36] D. Robinson, M. Miron, M. Hagiwara, and O. Pietquin, "Naturelm-audio: an audio-language foundation model for bioacoustics," 2024.

[37] X. Deng, T. Wan, K. Xu, T. Gao, P. Qiao, D. Feng, and Y. Dou, "Scaling bioacoustic signal pre-training with million samples via mask-modeling," in *Proc. ICASSP*, 2025.