Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures

Jules Cauzinille, Benoît Favre, Ricard Marxer, Dena Clink, Abdul Hamid Ahmad, Arnaud Rey

Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal identity

Jules Cauzinille[1], *Benoit Favre*[1], *Ricard Marxer*[1], *Dena Clink*[2], *Abdul Hamid Ahmad*[3], *Arnaud Rey*[1]

[1]Aix-Marseille University, Université de Toulon, CNRS, France
[2]Cornell University, Ithaca, NY, USA
[3]Universiti Malaysia Sabah (UMS), Kota Kinabalu, Sabah, Malaysia

jules.cauzinille@lis-lab.fr

Abstract

With the advent of pre-trained self-supervised learning (SSL) models, speech processing research is showing increasing interest towards disentanglement and explainability. Amongst other methods, probing speech classifiers has emerged as a promising approach to gain new insights into SSL models out-of-domain performances. We explore knowledge transfer capabilities of pre-trained speech models with vocalizations from the closest living relatives of humans: non-human primates. We focus on classifying the identity of northern grey gibbons (*Hylobates funereus*) from their calls with probing and layer-wise analysis of state-of-the-art SSL speech models compared to pre-trained bird species classifiers and audio taggers. By testing the reliance of said models on background noise and timewise information, as well as performance variations across layers, we propose a new understanding of the mechanisms underlying speech models efficacy as bioacoustic tools.

Index Terms: transfer learning, self-supervised learning, computational bioacoustics, probing

1. Introduction

In recent years, automatic speech processing research has been undergoing a slight change of paradigm, partly embodied in the use of transformer models and SSL [1]. SSL is now adopted as a state-of-the-art (SOTA) approach relying on the pre-training of models with large unannotated datasets, and challenges pre-existing benchmarks in a wide range of speech processing tasks [2, 3]. Pre-trained SSL models not only show notable performance gains compared to supervised methods, they also address a major limitation in automatic speech processing: dealing with small and scarcely labeled datasets. This advantage lies in their ability to extract latent representations which contain varied phonetic and acoustic information such as prosody, emotional cues, vocal identity, etc. [2, 4]. This makes SSL speech models particularly interesting candidates for out-of-domain classification of "speech-like" data. The closely related field of computational bioacoustics, i.e., the automatic processing of acoustic data produced by animals, is showing similar interest in the development of pre-trained foundation models. In bioacoustics, the detection and classification of species, call-types or individual vocal signatures is usually tackled with dataset- and species-specific supervised classifiers [5]. Yet, novel methods are emerging to extract meaningful latent representations from publicly available pre-trained bioacoustic classifiers [6, 7, 8] and unsupervised methods [9, 10]. Researchers are progressively turning to transfer learning, mostly as a solution to the scarcity of annotated bioacoustic datasets and the necessity to efficiently process large quantities of unlabeled animal vocalization recordings with SOTA approaches [11, 12, 13]. In this context, bioacoustics and speech processing will benefit from further investigation and comparison of the mechanisms underlying the use of pre-trained speech models for bioacoustic tasks. To better understand speech-based SSL models representations of non-human vocalizations, we propose a set of experiments focused on northern grey gibbons (*Hylobates funereus*), a species of singing apes described in 2.2. We evaluate the ability of pre-trained speech models to extract the vocal signature of individual gibbons by testing the following hypotheses:

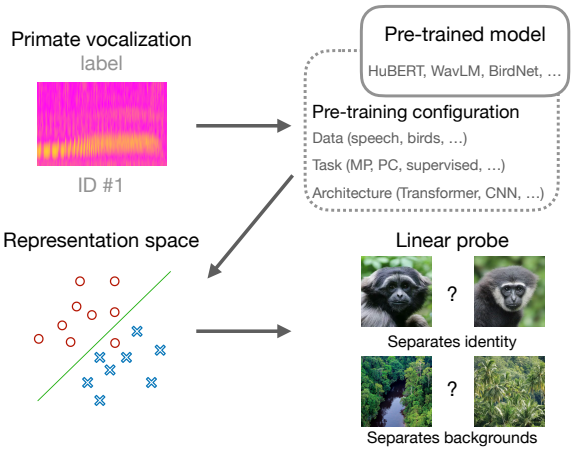

Figure 1: *Overview of the approach: we probe pre-trained models for primate identification based on vocalizations, according to pre-training configuration (data, task, architecture).*

- **H1**: pre-trained speech models are better suited at capturing primate vocal signatures compared to pre-trained bird-species classifiers or audio-tagging models
- **H2**: pre-trained speech models specialize along layers, resulting in better bioacoustic probing performance in initial layers
- **H3**: bird-species classifiers and audio-tagging models tend to rely on background noise information rather than vocal identity features to solve the task
- **H4**: specific SSL models architectures rely on temporal rather than spectral information to recognize the identity of gibbons from their vocalizations.

To investigate these hypotheses we propose a probing methodology aimed at evaluating the latent representations of pre-trained speech models by: 1) comparing their performance with other pre-trained solutions, 2) assessing performance variations across layers, 3) evaluating their reliance on the acoustic context of the vocalisations and 4) on temporal segments of varying granularity.

2. Methods

2.1. Linear probing

Probing aims at understanding what type of information a pre-trained model can extract and how said information is encoded within the model parameters. We hereby compute feature embeddings from 10 pre-trained models described in Section 2.3 and assess their ability to classify the identity of 10 gibbon female callers from their songs. An overview of the pipeline can be seen in Figure 1.

We use linear probing to test the ability of pre-training to create representation spaces linearly separable according to a task [14]. Given a training dataset $D = \{\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}\}$ composed of N couples (audio recording, primate identity), let $M_l(\boldsymbol{x})$ be the activations of a model M after layer l. A probe is trained to minimize the cross-entropy loss $CE()$ over the softmax of a linear projection of the activations, a logistic regression with $\boldsymbol{W}$ being a trainable parameter matrix.

$$\underset{\boldsymbol{W}}{\text{minimize}} -\frac{1}{N}\sum_{i=1}^{N} \text{CE}\left(\boldsymbol{y}_i, \text{softmax}\left(M_l(\boldsymbol{x}_i)^T \boldsymbol{W}\right)\right) \quad (1)$$

In this process, the original model parameters are not modified, allowing to assess properties of pre-trained representations instead of the effects of initialization which full fine-tuning could uncover. Accuracy on the probing classification task attests of these properties. The code and dataset are available for replication at `https://github.com/jcauzi/GibbonID`.

2.2. Task and dataset

We probe a range of pre-trained models on the task of primate identity classification, that is the capacity to identify primates from a set of known individuals after training from recordings of their calls.

Our experiments focus on northen grey gibbons from North Borneo. Gibbons (Family *Hylobatidae*) are singing apes known for their species- and sex-specific loud calls. Their reliance on long-distance vocal communication makes them excellent candidates for bioacoustics research. Most species of gibbons engage in duets where adult male and females emit alternating vocal output. The female contribution to the duet, known as the great call is composed of several short acoustic units (*ie* notes). It is a highly stereotyped tonal call with important frequencial modulations (within 0.2-2 kHz) and has been shown to contain individual signatures in many gibbon species including grey and lar gibbons [15, 16]. This northern grey gibbons dataset is comprised of focal recordings carried out in their natural habitat between January 2013 and September 2016 in Sabah, Malaysia. As gibbons are a territorial species, we used a combination of recording location and group composition to distinguish different gibbon groups and manually identified female great calls to be considered as individually labeled clips. The full dataset contains 1090 recordings from 91 individuals, segmented by hand around single great call utterances of around 13.6 seconds.

The number of recordings per individual varies from 1 to 47 and only 10 individuals have more than 25 recordings, motivating a 10-way classification task. Many effects can bias models for the task of identity classification, such as background soundscape, microphone distance, atmospheric conditions, data scarcity, etc. In order to limit such biases, we construct a training/test set comprised of the 10 females with 25 recordings, and select a subset of around 5 test samples per class that were recorded either on different days or on a different time of the day compared to the training samples. The training set is thus comprised of 189 samples. This test set, named **Full** in the following experiments, is comprised of 61 samples with an average of 6 samples per class. In addition to this, we address the problem of background noise. This issue is especially important when carrying out individual recognition in bioacoustics where large amounts of information can often be extracted from the acoustic signature of an animal's territory rather than from its vocalizations [17]. We thus extract clips which do not contain vocalizations from portions of the unsegmented recordings. This provides a "background noise test set" labeled in terms of individuals but containing only the background noise from their specific territory. We consider a model's performance on this background test set to be highly correlated with its tendency to rely on the acoustic signature of the individual's territory rather than its vocal features. This particular test set is comprised of 7 segments of about 4 seconds for each of the 10 classes and is named **Background** in the experiments. Finally, we create a third test set comprised of $\simeq 0.5$ seconds single note occurrences from each great call. As great calls are comprised of several consecutive and accelerating notes, we define a note as a unique vocal unit separated by short silences, similarly to the definition in [18], and select $\simeq 0.5$ seconds-long notes from each call. This time-wise ablation allows testing the reliance of models on temporal and contextual information in their representation of vocal signatures. In experiments, this test set is named **Notes**.

2.3. Pre-trained models

In order to test hypotheses H1-H4, we harness a set of pre-trained models to cover a range of conditions along the kind and size of training data as well as model architectures and pre-training tasks. In particular, we use six speech models built for automatic speech transcription with capabilities in out-of-domain classification or specifically tailored for speaker identification (HuBERT [19], UniSpeech, wav2vec 2.0 [21], WavLM, APC [23]). All speech models are trained on mask-based pretext tasks to recover masked targets from left and right sound contexts. Wav2vec 2.0 and APC use contrastive losses and learn to extract masked targets representations from local features while HuBERT, WavLM and UniSpeech-SAT are trained with classification losses. Their targets are represented as discrete cluster IDs coming from intermediary layers of earlier model iterations. Except for APC and HuBERT base, those models are trained on relatively large quantities of speech, spanning different recording conditions. We also select two bird species classification models that were shown to perform well in out-of-domain bioacoustic tasks [12] (Google perch [7] and BirdNET [6]). They are trained to recognize $\simeq 1000$ species from recordings of bird vocalizations and cover a wide range of soundscapes. Finally, we use two audio-tagging models tailored for music tagging, acoustic scene classification or audio-event detection. VGGish [25], trained on AudioSet, is a popular option for bioacoustic classification [12] and Audio-MAE [24] is a SOTA audio-tagging model partly trained on speech data and Youtube audios.

Given that models diverge in embedding sizes and frame rates, we apply mean pooling of the activation vectors over entire sound segments prior to passing them to the probing model, resulting in a single input vector for each example. The characteristics of each model are summarized in Table 1.

In the experiments, we compare probing results to two sim-

Table 1: *Pre-trained model characteristics.*
number of Transformer Layers (n TL) - Masked Prediction (MP) - Predictive Coding (PC) - Convolutional Neural Network (CNN)

Model	Genre	Hours	Arch.	Task	#param.	Embedding	Window
[19] HuBERT Large	speech	60k	24 TL	MP	317M	1280	20 ms
[19] HuBERT Base	speech	960	12 TL	MP	95M	768	20 ms
[20] UniSpeech SAT Large	speech	94k	24 TL	MP + speaker	317M	1280	20 ms
[21] Wav2vec 2.0 Large	speech	53k	24 TL	contrastive PC	317M	1280	20 ms
[22] WavLM Large	speech	94k	24 TL	MP + robustness	90M	1280	20 ms
[23] APC	speech	360	3 TL	autoregressive PC	4M	512	10 ms
[7] Google perch	bird	10k	CNN	supervised	20M	1280	5000 ms
[6] BirdNET 2.3	bird	4k	CNN	supervised	10M	1280	3000 ms
[24] Audio-MAE AST	general + speech	50k + 960	MAE	MP	90M	768	20 ms
[25] Vggish	general	5k	CNN	supervised	10M	128	96 ms

ple baselines: chance which is 0.1 for the 10-way classification, and 12 MFCC-vectors extracted each 25 ms time windows and mean pooled before probing [13].

3. Results

Table 2: *Probing accuracy on full segments (Full), background noise (Background), and single notes (Notes). Bayes Factor (BF) - (**Best**, **second best**, lowest) - Transformer models results correspond to their second layer and to the last layer for APC and Audio-MAE.*

Model	Full	Background	Notes	BF
HuBERT Large	**0.95**	*0.12*	*0.13*	**7.76**
HuBERT Base	0.72	0.21	0.17	3.39
UniSpeech-SAT	0.87	0.19	0.14	4.64
Wav2vec 2.0 Large	0.86	0.14	0.44	**5.96**
WavLM Large	**0.94**	0.22	0.62	4.23
APC	0.75	0.14	0.14	5.44
Google perch	0.87	**0.69**	–	*1.26*
BirdNET 2.3	0.87	**0.63**	–	1.39
Audio-MAE AST	**0.95**	0.43	**0.82**	2.21
Vggish	*0.66*	0.43	–	1.52
MFCC	0.82	0.22	**0.94**	3.66
Chance	0.10	0.10	0.10	1.00

Experiments correspond to training a probe on the 10-way classification task and reporting accuracy when activations from a given model layer are used as input. Model-wise results are reported at the second layer in transformer-based models, and the last layer of APC and Audio-MAE AST, after inspecting best layer-wise performance on a validation set (20% of the train set). Given the small dataset size and constraints for selecting test sets, we resort to statistical bootstrapping: probes are trained on samples drawn from 80% of the available training data, and we report mean performance on the test set over 10 folds.

Main result Accuracy on the **Full** test set show that all models perform better than chance, reaching high accuracy (as seen in Table 2). This suggests that representation spaces contain significant information about gibbon vocal signatures. Most models yield better performance than the MFCC baseline, except for VGGish and APC which do not seem to benefit from the corresponding pre-training settings. HuBERT Large, WavLM and Audio-MAE AST show the highest accuracy. Speech models result in higher accuracy than bird-song models, supporting hypothesis H1. We further exercise scrutiny on those results to check whether hypotheses are validated.

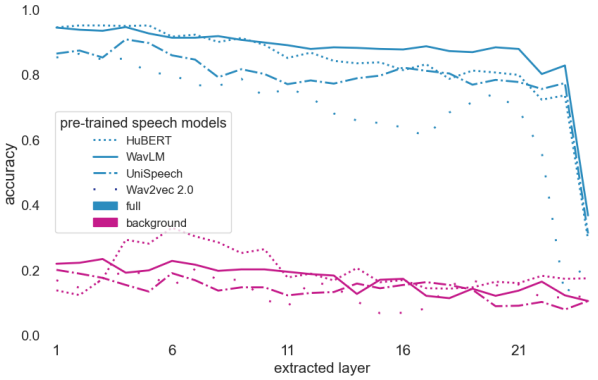

Figure 2: *Layer-wise probing performance of 4 "Large" speech models (**Full** and **Background** test sets).*

Layer depth in speech models This experiment addresses H2, the effect of layer depth on specialization. According to Figure 2, the initial layers of pre-trained speech models result in the best probing accuracy. Except for HuBERT base, which performances stay relatively consistent between all layers (from 0.69 to 0.84 accuracy for the 8^{th} and 12^{th} layers respectively), a significant drop can be seen in the last three to four layers of every Large model, suggesting for strong specialization on human speech. WavLM, besides being the second most performant speech model on this task, shows lower layer-wise accuracy decrease, with a significant drop only happening in its 20th layer. This might be explained by the specific noise-robust training setup of WavLM compared to other models [22]. In general terms, these results indicate that using the last layer of speech-based transformer models without fine-tuning for feature extraction in bioacoustic tasks may be a sub-optimal solution.

Effect of background noise Channel, and in particular background noise, is a potentially strong source of bias in identity classification. Results on the **Background** test set show that speech-only models such as HuBERT tend to rely less on background for classifying gibbon identity than models pre-trained to recognize bird species or to reconstruct acoustic signal from general audio. To measure this effect, we compare performance on the **Background** and **Full** test sets using a Bayes Factor

[26]. This metric corresponds to the ratio between the probability to accurately classify an individual given a recording of its vocalizations and the probability to do so given its territorial background only. Given two sub-hypotheses: H3.1—a model accounts for both vocal features and background noise, and H3.2—a model only accounts for background noise information to predict identity. We formalize the Bayes Factor as:

$$BF = \frac{P(\text{correct identity}|\text{H3.1})}{P(\text{correct identity}|\text{H3.2})} \quad (2)$$

Those probabilities can be estimated with the accuracy from probing the models on the **Full** and the **Background** test sets. Following [26], we consider $3 < BF < 10$ to be moderate evidence for the hypothesis that a model does not significantly rely on background noise to predict identity. All models with speech-only pre-training outperform the rest in terms of BF, with HuBERT Large reaching the highest ratio, and MFCCs improving compared to all non-speech models. Figure 2 shows relatively stable background representations among layers of speech models.

Effect of call structure An interesting question is whether primate identity is conveyed through low-level features of the acoustic signal driven by vocal tract morphology, such as pitch and timbre, or via call structure, such as the duration and arrangement of individual notes. In order to assess H4, we probe the models on the **Notes** test set which removes high-level temporal information. As Google perch, BirdNET and VGGish take segments longer than 0.5 seconds as input they are not included in the single notes tests. Results tend to show that Audio-MAE, as well as WavLM and wav2vec 2.0 to a lesser extent, reach accuracies higher than chance without the need for temporal information. Interestingly, UniSpeech and HuBERT are unable to classify individuals from short segments, suggesting their potential reliance on temporal features of the individual calls. Unsurprisingly, MFCCs are particularly well suited to recognize vocal signatures from short sound segments since they do not encode long term structure. Additionally, averaging MFCCs across full segments introduces noise between notes within the extracted features.

4. Discussion, prior work & limits

As stated in hypothesis H1, we show the superior ability of SSL speech models at transferring knowledge from their pre-training dataset to non-human primate vocalizations for the automatic classification of individual vocal signatures. The observed superior performances of said models may be explained by our following three hypotheses.

Regarding hypothesis H2, our experiments show that the information needed to answer such a task is predominantly encoded in the initial transformer layers of speech models. Deeper layers tend to show lower accuracies, potentially as a result of their increasing specialization on the extraction of more lexical and semantic human speech features [27]. This may explain the divergences in our results compared to similar experiments which limit their usage of pre-trained speech models to the extraction of representations from their last transformer layer [11, 13]. As minor layer-wise performance variation and lower performances in general can be seen in small models (namely HuBERT Base and APC), we also note the importance of the number of parameter and pre-training dataset size for the resolution of the task.

In accordance with hypothesis H3, pre-trained speech models are particularly adapted to extract meaningful information from primate vocalizations rather than their territorial background noise. Pre-trained bird species classifiers and general audio-tagging models, although relatively performant on the task, are seemingly relying on the extraction of eco-acoustic information (i.e., the background forest sounds and territorial acoustic signatures). This advantage of speech models can be explained by various factors. In terms of pre-training data, the phylogenetic link between humans and non-human primates could be indirectly leveraged by speech models extracting acoustic features common to both species. Vocal signatures are highly influenced by individual physical properties and anatomical variability [28, 29] in both humans and non-human primates. Therefore, said features in primates may be easier to capture by models trained on speech rather than birds or other types of acoustic data. Yet, the spectral information needed to recognize vocal features proper to an individual should also be available in single notes which, as we have shown, do not provide sufficient information for some pre-trained speech models. We thus hypothesize that HuBERT and UniSpeech, contrary to audio-tagging models such as Audio-MAE, rely on the temporal dynamics of gibbon vocalizations to extract their emitter's identity. In addition to spectral features due to anatomical variation, temporal dynamics and non-linearities in the vocal production mechanism can, in fact, lead to significant differences in individuals' calls [30, 16, 31].

Similarly, the high performance observed in pre-trained bird models for background noise classification (hypothesis H3), could be explained by the presence of forest environments in their pre-training dataset. We advocate for similar testing of background effects in experiments involving the use of bird species classifiers for bioacoustic transfer learning [12]. As both models learn to extract meaningful information from forest acoustic environments, they apparently rely on such information for the task at hand. Furthermore, gibbon recordings are partly "contaminated" by bird vocalizations. As specific bird species can be encountered in overlapping territories with individual gibbons, bird species classifiers could be relying on background bird vocalizations to solve the task, as indicated by their low Bayes Factor. Similar yet less significant dynamics can also be observed with Audio-MAE and VGGish, both trained on the varied set of sounds from AudioSet [32].

From a computational perspective, a link should be made between the architecture and training objectives of speech models compared to bird classification and audio-tagging. Our results show that the use of masked modeling to learn discrete speech units with contextualized representations can be utilized on non-speech data, at least for our specific task and dataset. The observed efficacy of this approach combined with the extensive availability of speech data should thus be explored as a potential asset to be leveraged in computational bioacoustics.

Our work is limited in that the variability of each model in terms of size, embedding dimension and pre-training data can be a source of bias in their comparison. Yet, retraining models in comparable setups is a particularly tedious task demanding large computational resources. This work aims at showing how the development of specific tailor-made datasets and controlled probing experiments can help overcoming these limitations. Further exploration of the aforementioned hypotheses will also imply conducting similar experiments on additional species, tasks and models.

5. Acknowledgements

This research was supported by the Convergence Institute ILCB (ANR-16-CONV-0002), the COMPO ANR project (#ANR-23-CE23-0031) and the HEBBIAN ANR project (#ANR-23-CE28-0008).

6. References

[1] 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 Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[2] S. Yang, P. Chi, Y. Chuang, C. Lai, K. Lakhotia, Y. Lin, A. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee, D. Liu, Z. Huang, S. Dong, S. Li, S. Watanabe, A. Mohamed, and H. Lee, "Superb: Speech processing universal performance benchmark," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. International Speech Communication Association, 2021, pp. 1–5.

[3] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7927–7931.

[4] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "On the Transferability of Large-Scale Self-Supervision to Few-Shot Audio Classification," Feb. 2024, arXiv:2402.01274 [cs, eess] version: 1.

[5] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, 2022.

[6] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

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

[8] M. Hagiwara, "AVES: Animal Vocalization Encoder based on Self-Supervision," Oct. 2022, arXiv:2210.14493 [cs, eess].

[9] P. Best, S. Paris, H. Glotin, and R. Marxer, "Deep audio embeddings for vocalisation clustering," *PLOS ONE*, vol. 18, no. 7, pp. 1–18, 07 2023.

[10] M. Thomas, F. H. Jensen, B. Averly, V. Demartsev, M. B. Manser, T. Sainburg, M. A. Roch, and A. Strandburg-Peshkin, "A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations," *Journal of Animal Ecology*, vol. 91, no. 8, pp. 1567–1581, 2022.

[11] E. Sarkar and M. Magimai.-Doss, "Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?" in *Proc. INTERSPEECH 2023*, 2023, pp. 1189–1193.

[12] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, p. 22876, Dec. 2023.

[13] M. W. Lakdari, A. H. Ahmad, S. Sethi, G. A. Bohn, and D. J. Clink, "Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons," *Ecological Informatics*, vol. 80, p. 102457, May 2024.

[14] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 311–315.

[15] D. J. Clink, H. Bernard, M. C. Crofoot, and A. J. Marshall, "Investigating Individual Vocal Signatures and Small-Scale Patterns of Geographic Variation in Female Bornean Gibbon (Hylobates muelleri) Great Calls," *International Journal of Primatology*, vol. 38, no. 4, pp. 656–671, Aug. 2017.

[16] T. A. Terleph, S. Malaivijitnond, and U. H. Reichard, "Lar gibbon (Hylobates lar) great call reveals individual caller identity," *American Journal of Primatology*, vol. 77, no. 7, pp. 811–821, Jul. 2015.

[17] D. Stowell, T. Petrusková, M. Šálek, and P. Linhart, "Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions," *Journal of the Royal Society Interface*, vol. 16, no. 153, p. 20180940, Apr. 2019.

[18] E. Clarke, U. Reichard, and K. Zuberbühler, "The Syntax and Meaning of Wild Gibbon Songs," *PloS one*, vol. 1, p. e73, Feb. 2006.

[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/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, 2021.

[20] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[22] 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 Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[23] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3497–3501.

[24] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked Autoencoding Audio Spectrogram Transformer," in *Proc. Interspeech 2022*, 2022, pp. 2438–2442.

[25] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[26] S. Andraszewicz, B. Scheibehenne, J. Rieskamp, R. Grasman, J. Verhagen, and E.-J. Wagenmakers, "An introduction to bayesian hypothesis testing for management research," *Journal of Management*, vol. 41, no. 2, pp. 521–543, 2015.

[27] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[28] B. D. Charlton, K. Pisanski, J. Raine, and D. Reby, *Coding of Static Information in Terrestrial Mammal Vocal Signals*. Cham: Springer International Publishing, 2020, pp. 115–136.

[29] M. Gamba, L. Favaro, A. Araldi, V. Matteucci, C. Giacoma, and O. Friard, "Modeling individual vocal differences in group-living lemurs using vocal tract morphology," *Current Zoology*, vol. 63, no. 4, pp. 467–475, 03 2017.

[30] W. Fitch, J. Neubauer, and H. Herzel, "Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production," *Animal Behaviour*, vol. 63, pp. 407–418, 03 2002.

[31] J. Bradbury and S. Vehrencamp, *Principles of Animal Communication*. Sinauer, 2011.

[32] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.