From speech to primate vocalizations: Self-supervised deep learning as a comparative approach

Jules Cauzinille[*1,2,3], Benoît Favre[1,2], Ricard Marxer[1,2], and Arnaud Rey[1,3]

[*]Corresponding Author: jules.cauzinille@lis-lab.fr
[1]ILCB, Aix-Marseille University, Marseille, France
[2]LIS, Aix-Marseille University, Marseille, France
[3]CRPN, Aix-Marseille University, Marseille, France

Within the recent deep learning revolution, *transformer* architectures and *pre-trained self-supervised models* opened up many perspectives for the study of linguistics and animal communication. These state-of-the-art tools efficiently address a wide range of applications in monitoring animal behavior through sound (Stowell, 2022; Kahl, Wood, Eibl, & Klinck, 2021; Hagiwara, 2023) or in assisting humans with language related tasks. The increasing scientific interest generated by this revolution raises the following question: can acoustic deep learning be leveraged as a scientific tool in the study of the evolution of language?

We propose a novel methodology involving the use of deep learning models as comparative toolkits by testing their ability to jointly process speech and non-human vocal communication. This approach relies on the disentanglement of self-supervised learning (SSL) pre-trained models, i.e., computer models trained on large unlabeled datasets. SSL models were introduced in the field of computer vision (Jing & Tian, 2021) as a way to leverage the extensive availability of image data. They rely on the assumption that pre-training a first model to encode and extract information from large collections of raw data can benefit secondary models specialized in downstream tasks on smaller-sized datasets. By applying this method to acoustic data, researchers were able to develop efficient speech processing models, outperforming most purely *supervised* solutions (Mohamed et al., 2022; Yang et al., 2021). SSL models trained on speech datasets show high performance on an array of tasks (Evain et al., 2021) and learn to encode different levels of linguistic information during pre-training without the need for supervision. For instance, Pasad, Shi, and Livescu (2023) showed that low-level acoustic information tends to be encoded in the initial layers of these models while high-level phonemic or lexical information is mostly encoded in deeper layers.

By adapting the SSL approach to bioacoustic tasks, we develop transfer learning experiments aimed at understanding how much information speech-based models are able to extract from non-human vocalizations. We focus our preliminary experiments on non-human primates, more specifically apes, as our closest

living relatives provide a unique opportunity to explore the evolutionary basis of our vocal communicative behavior. We rely on models pre-trained on human speech (Hsu et al., 2021; Schneider, Baevski, Collobert, & Auli, 2019) to perform primate-related bioacoustic tasks and compare them to models pre-trained on other taxa such as birds (Kahl et al., 2021), or general acoustic data such as music, video soundtracks, etc. (Huang et al., 2022; Kong et al., 2020). The tasks include vocal identity recognition, detection of vocalizations in natural contexts and call-type classification.

We define three main approaches to test the knowledge transfer capabilities of SSL models from speech to primate vocalizations. The **probing** approach consists in using pre-trained models as feature extractors. Said features are then "probed" with logistic regression to disentangle the type of information they extracted from primate vocalizations. Good performance on a given task shows that the information needed to answer the task was successfully extracted during pre-training and is linearly separable within the model's representations. The **fine-tuning** approach involves further training SSL models on small datasets to improve their performance on the downstream task. It can show how much more training data a model needs to efficiently extract information from primate vocalizations. Finally, to ensure true knowledge transfer from human to other primates, a third method involves **parameter-efficient fine-tuning** (PEFT) and **adversarial reprogramming** (Elsayed, Goodfellow, & Sohl-Dickstein, 2018; Zheng et al., 2023). Both methods allow keeping the pre-trained weights of the original model untouched by training additional "filters" for primate-related tasks.

Preliminary experiments consist in recognizing vocal signatures of individual gibbons (*Hylobates funereus*). The probing method shows that the initial layers of speech-based models are capable of extracting sufficient information to classify the individual voices of 10 female gibbons with up to 95% accuracy. This result outperforms models pre-trained on birdsongs, which seem to heavily rely on recognizing the background noise of recordings rather than the primate's vocal signature. Additionally, we demonstrate the ability of some speech models to recognize gibbon's vocal identities from the temporal dynamics of their song rather than the anatomical specificities of their voices. Finally, when the fine-tuning method is applied, further performance gains can be observed, even in few-shot learning setups.

This type of result helps us examine divergences and similarities between speech and primate vocalizations from a deep learning perspective. They show how speech-based pre-training may be at an advantage when dealing with primate vocal communication by transferring knowledge from one to the other. In general terms, our experiments test for the validity of deep transfer learning as a scientific tool in the study of the origins of language from a comparative standpoint. Future experiments will focus on extending previously mentioned methods to other tasks and primate species.

Acknowledgments

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). It was also supported by the COMPO ANR project (#ANR-23-CE23-0031) and the HEBBIAN ANR project (#ANR-23-CE28-0008).

All gibbon recordings and annotations were provided by Dena J. Clink (K. Lisa Yang Center for Conservation Bioacoustics and Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA). More information can be found in Clink, Kier, Ahmad, and Klinck (2023) and Clink, Bernard, Crofoot, and Marshall (2017)

References

Clink, D., Kier, I., Ahmad, A., & Klinck, H. (2023). A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution*, *11*.

Clink, D. J., Bernard, H., Crofoot, M. C., & Marshall, A. J. (2017). Investigating Individual Vocal Signatures and Small-Scale Patterns of Geographic Variation in Female Bornean Gibbon (Hylobates muelleri) Great Calls. *International Journal of Primatology*, *38*(4), 656–671.

Elsayed, G. F., Goodfellow, I., & Sohl-Dickstein, J. (2018). *Adversarial Reprogramming of Neural Networks.* arXiv. (arXiv:1806.11146 [cs, stat])

Evain, S., Nguyen, H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., & Besacier, L. (2021). LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *INTERSPEECH 2021: Conference of the International Speech Communication Association.* Brno, Czech Republic.

Hagiwara, M. (2023). Aves: Animal vocalization encoder based on self-supervision. In *Icassp 2023 - 2023 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 1-5).

Hsu, W.-N., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *PP*, 1-1.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., & Feichtenhofer, C. (2022). Masked autoencoders that listen. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 28708–28720). Cur-

ran Associates, Inc.

Jing, L., & Tian, Y. (2021). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(11), 4037-4058.

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, *61*, 101236.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2880-2894.

Mohamed, A., Lee, H. yi, Borgholt, L., Havtorn, J., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T., & Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1179–1210.

Pasad, A., Shi, B., & Livescu, K. (2023). Comparative layer-wise analysis of self-supervised speech models. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. interspeech 2019* (pp. 3465–3469).

Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, *10*.

Yang, S., Chi, P., Chuang, Y., Lai, C., Lakhotia, K., Lin, Y., Liu, A., Shi, J., Chang, X., Lin, G., Huang, T., Tseng, W., Lee, K., Liu, D., Huang, Z., Dong, S., Li, S., Watanabe, S., Mohamed, A., & Lee, H. (2021). Superb: Speech processing universal performance benchmark. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5). International Speech Communication Association.

Zheng, Y., Feng, X., Xia, Z., Jiang, X., Demontis, A., Pintor, M., Biggio, B., & Roli, F. (2023). Why adversarial reprogramming works, when it fails, and how to tell the difference. *Information Sciences*, *632*, 130–143.