On the Utility of Speech and Audio Foundation Models for Marmoset Call Analysis

Eklavya Sarkar[1,2], *Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Switzerland
[2]Ecole polytechnique fédérale de Lausanne, Switzerland
{eklavya.sarkar, mathew}@idiap.ch

Abstract

Marmoset monkeys encode vital information in their calls and serve as a surrogate model for neuro-biologists to understand the evolutionary origins of human vocal communication. Traditionally analyzed with signal processing-based features, recent approaches have utilized self-supervised models pre-trained on human speech for feature extraction, capitalizing on their ability to learn a signal's intrinsic structure independently of its acoustic domain. However, the utility of such foundation models remains unclear for marmoset call analysis in terms of multi-class classification, bandwidth, and pre-training domain. This study assesses feature representations derived from speech and general audio domains, across pre-training bandwidths of 4, 8, and 16 kHz for marmoset call-type and caller classification tasks. Results show that models with higher bandwidth improve performance, and pre-training on speech or general audio yields comparable results, improving over a spectral baseline.

Index Terms: bioacoustics, call-type and caller classification, speech and audio, bandwidth.

1. Marmoset Vocalizations

Non-human vocal communication, such as bioacoustics, i.e. the study of animal vocalizations, is rapidly advancing through the advent of machine learning and the correlated progress in human speech processing [1]. Common marmosets (*Callithrix jacchus*) are of particular interest due to their highly vocal nature, acoustically diverse call repertoire, and acute auditory capabilities. Their extensive vocalizations are rooted in a complex social system, and are thus able to encode a range of information, such as group affiliation, sex [2], population, dialect [3], and even individual caller identity [4, 5], over a number of social and emotional states [6, 7]. Their remarkable vocal adaptability also allows them to modify the duration [8], intensity [9], complexity [10], or timing [11] of their calls. These vocal characteristics align them closely with human speech properties, such as care-giving to infants, turn-taking [12], and categorical perception of sounds [13], and make them into a well-suited surrogate model for understanding the vocal communication of non-human primates among biologists [14] and neuroscientists [15].

In the literature, the automatic analysis of marmoset vocalizations, i.e. call-type, caller identity, or sex classification, has been conducted by leveraging signal processing features alongside traditional machine learning classifiers. Early work demonstrated that k-NN, SVM, and optimal path forest classifiers achieved notable success over multilayer perceptrons (MLPs), Adaboost, and logistic regression, especially with small, specific datasets [16]. Research exploring a variety of audio and spectral feature representations, such as signal energy, zero crossing rate, spectral rolloff, and MFCCs, indicated that integrating different feature could enhance the system's performance on synthetically augmented vocal datasets [17]. Recent studies have also explored leveraging deep learning based techniques. Using convolutional neural networks to process spectrograms for simultaneous vocalization detection, call-type classification, and caller identification was found to outperform separate models for each task [18]. Statistics of log-mel filter-bank energies used as input for recurrent neural networks (RNNs) were shown to improve the detection and classification of calls over SVM or MLPs [19]. Self-supervised learning (SSL) frameworks, which create surrogate labels from the data, were used with the aim of leveraging the large quantities of unlabeled data for birdsong detection [20] and bioacoustic event detection [21].

A novel study demonstrated that neural representations derived from models pre-trained on human speech through SSL could distinguish individual marmoset caller identities [22]. The authors argued that SSLs only learn the intrinsic structure of the unlabeled input signal, typically through a masking-based pre-text training task, to capture essential information independently of any domain-specific knowledge, such as human speech production, and thus can be cross-transferred across different acoustic domains, such as bioacoustics. Building on these findings, our paper investigates the utility and limitations of such pre-trained foundation models for the purpose of marmoset call analysis, with a focus on the following key points:

1. **Classification**: We investigate whether such models can be effectively leveraged for marmoset call analysis tasks, namely call-type and caller classification, which, to the best of our knowledge, has not yet been demonstrated. Additionally, while [22] focused solely on caller detection in a binary framework, we extend the scope to a multi-class approach.
2. **Bandwidth**: Given that these models are typically pre-trained at a bandwidth of 8 kHz, we address their mismatch with the biological vocalization and auditory range of marmosets, predominantly concentrated in the 5–10 kHz spectral region [23], and thus evaluate their capability to accurately represent marmoset calls. By examining models pre-trained across varying bandwidths, we aim to evaluate their effectiveness in adequately representing marmoset calls, and seek to clarify how model bandwidth influences their classification.
3. **Pre-training domain**: It remains unclear how models pre-trained on human speech compare to trained on other acoustic domains for accurately capturing marmoset call characteristics. We examine representations produced by different pre-training sources, such as human speech and general audio, across supervised and self-supervised learning frameworks, against a spectral baseline to identify the most suitable pre-

github.com/idiap/speech-utility-bioacoustics

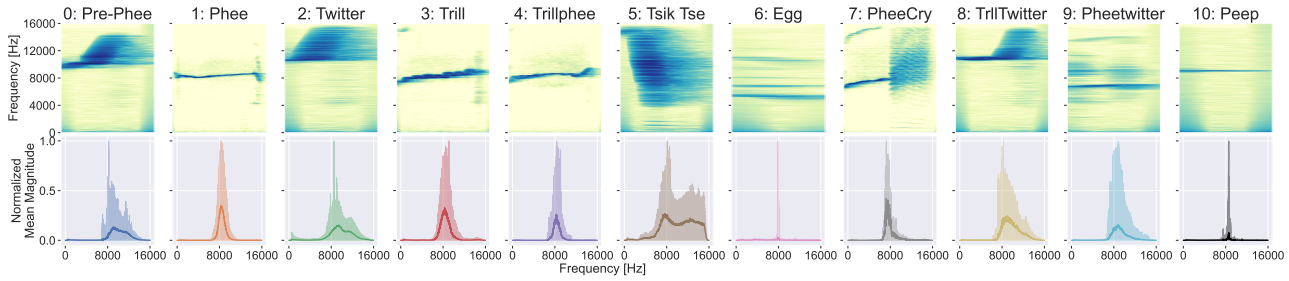

Figure 1: *Marmoset vocalizations with a 16 kHz bandwidth.* Top: *Spectrograms of a* single *call-type vocalization.* Bottom: *The mean spectrum for* all *vocalizations per call-type across the dataset, normalized. Shaded areas indicate* $\pm$ *1 std from the mean spectrum.*

training source for cross-domain bioacoustic signal analysis. The rest of the paper is organized as follows: Section 2 gives the study's methodology, section 3 & 4 present a call similarity and classification analysis. Section 5 finally concludes the paper.

2. Methodology

2.1. Dataset and Tasks

For our study, we used the InfantMarmosetsVox (IMV) dataset [22], which contains 72,921 labelled marmoset vocalization segments (totalling to 464 minutes), sampled at 44.1 kHz, across ten marmoset individuals and contains eleven marmoset call-types. Table 1 presents the data distribution in function of the call-types and callers. For our experiments, we divide the dataset into a *Train*, *Val*, and *Test* sets, following a random 70:20:10 split. We denote call-type and caller identity multi-class classification as CTID and CLID respectively.

Table 1: *InfantMarmosetsVox dataset statistics.*

ID	Call-type	Count	Caller ID	Count
0	Peep (pre-phee)	1283	0	15521
1	Phee	27976	1	8648
2	Twitter	36582	2	13827
3	Trill	1408	3	5838
4	Trillphee	728	4	5654
5	Tsik Tse	686	5	3522
6	Egg	1676	6	4389
7	Pheecry (cry)	23	7	2681
8	TrllTwitter	293	8	6387
9	Pheetwitter	2064	9	6454
10	Peep	202	-	-
	Total	**72921**	**Total**	**72921**

Figure 1 gives the visualizations of all call-types as well the density distribution of the spectrums across the entire dataset. Frequencies below 500 Hz are nullified purely for visualization to eliminate any low-frequency noise. We can observe that information starts at around 7-8 kHz for most calls in this dataset.

2.2. Models and Feature Representations

For our study, we select four distinct frameworks for feature representations $\mathcal{F}$: hand-crafted (HC) features derived through signal processing techniques, neural representations obtained via self-supervised learning (SSL), pre-trained on either human speech or general audio, and features generated through supervised learning (SL) models pre-trained on general audio. These frameworks are summarized in table 2. We extract the features from these frameworks by giving the marmoset calls as input.

Table 2: *# Parameters* P *and feature dimension* D *of selected models, pre-trained on AudioSet (AS) or LibriSpeech (LS).*

$\mathcal{F}$	Corpus	P	D	Type
C22 [24]	-	-	24	HC
WavLM [25]	LS	94.38M	1536	SSL
BYOL [26]	AS	5.32M	2048	SSL
PANN [27]	AS	8.08M	2048	SL

Hand-crafted: The Highly Comparable Time-Series Analysis (HCTSA) framework, used for interpreting diverse time series data, extracts 7700 features through signal processing methods, such as LPC [28]. It has been applied to diverse tasks such as birdsong discrimination [29], ecosystem monitoring [30], and marmoset caller identification [5]. Despite its broad applicability, HCTSA's computational demands and feature redundancy are significant limitations. The CAnonical Time-series CHaracteristics (Catch22/C22), a steamlined subset of HCTSA, provides high performance with minimal redundancy across numerous classification problems [24]. We extend this feature set to a final dimension of $D = 24$ by appending the first and second order statistics, and use it as our spectral baseline.

SSL pre-trained on human speech: Following the approach in [22], we use feature representations from SSL models trained on human speech, extending it to both call-type and caller identity classification. We select the WavLM base model, pre-trained on the 960-hour LibriSpeech dataset, based on its effectiveness in marmoset call detection as well as its versatility in speech processing tasks as demonstrated in the SUPERB challenge [31]. For each layer, feature representations of length 768 are extracted for each frame. Then, they are transformed into fixed-length utterance-level representations by computing and aggregating first and second order statistics across the frame-axis, resulting in a final representation of length $D = 1536$.

SL pre-trained on general audio: Expanding marmoset call analysis literature, we utilize embeddings from models pre-trained on the AudioSet (AS) dataset, which includes audio event classes such as environmental sounds, musical instruments, and human and animal vocalizations. Specifically, we choose the *AudioNTT2020* model from the BYOL-A architecture [26], extracting embeddings from its final fully connected layer of length $D = 2048$. Inputs are processed into log-mel spectrograms, adhering to the spectral parameters detailed in the original study, i.e. a 8 kHz bandwidth, 64 ms window size,

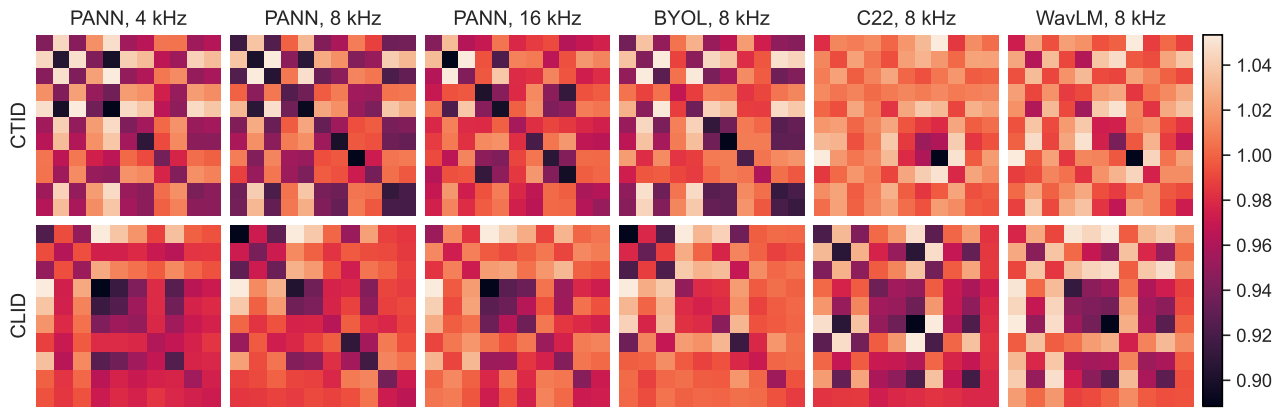

Figure 2: *Pairwise mean cosine distances matrices for features $\mathcal{F}$ at different bandwidths for call-types (CTID) and callers (CLID). Diagonal entries represent intra-class distances, and off-diagonal the inter-class. Darker regions indicate higher similarity.*

10 ms hop size, and 64 mel bins spanning from 60 to 7800 Hz.

SL pre-trained on general audio: We further investigate feature extraction from large-scale networks pre-trained for general audio pattern recognition. The *CNN14* model from the *PANN* network [27] is chosen, with pre-trained weights applied at three different bandwidths: 4, 8, and 16 kHz. This model employs a balanced sampling strategy across AudioSet's sound classes and also processes input vocalizations into spectrograms to extract log-mel filterbanks. For a bandwidth of 16 kHz, window and hop sizes are set to 1024 and 320 samples, respectively, and proportionally halved for 8 and 4 kHz. The model utilizes 64 mel bands, spanning from 50 Hz and to the Nyquist frequency. Embeddings of length $D = 2048$ are extracted from the linear layer preceding the final classification layer.

3. Call Similarity Analysis

This section presents a pairwise similarity analysis of the selected features on the *Train* set to identify any discernible patterns or correlations for given the vocalizations. Specifically, we investigate how variations in the bandwidth of the pre-trained models affect the similarity distribution of intra-class embeddings, and examine any distinctions between models pre-trained on speech against general audio. To compare the features, which are high-dimensional vectors, we use the cosine distance defined as $\text{sim}(\boldsymbol{x_1}, \boldsymbol{x_2}) = 1 - (\boldsymbol{x_1} \cdot \boldsymbol{x_2} \,/\, \|\boldsymbol{x_1}\| \cdot \|\boldsymbol{x_2}\|)$, bounded in $[0, 2]$. Two features are identical when their cosine distance is 0, orthogonal at 1, and opposite at 2. For WavLM, we select the first layer, and only use the first half of the extracted features, corresponding to the mean values averaged frame-wise.

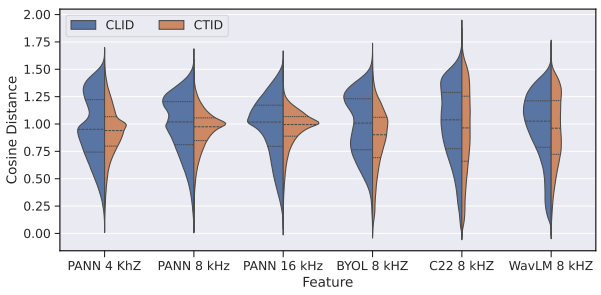

Figure 3: *Distribution of pairwise cosine distances.*

Figure 3 presents the overall distribution of pairwise distances. The distrtibutions are overlapping, centering around a median distance of 1 for all representations, suggesting a lack of clear correlation or similarity within the embeddings generated. Figure 2 further delineates the distributions into distance matrices for each feature set, where diagonal and off-diagonal entries correspond to intra-class and inter-class distances respectively. In an ideal scenario, embeddings from the same call-type or caller would exhibit closer distances, where as embeddings from different classes would have a higher dissimilarity.

We can observe that the models pre-trained on general audio datasets (BYOL and PANN) yield more distinct peaks and diagonals, on figures 3 and 2 respectively, compared to those pre-trained on human speech (WavLM) or the handcrafted baseline (Catch22). This distinction is more pronounced for call-types than for caller identification. This is expected, given that the call-types are spread across caller classes (a caller produces different calls, while a call can come from any caller). Although these patterns indicate some level of class-specific clustering, the distribution of distances largely show that the features are highly orthogonal. The similarity analysis thus indicates minimal feature correlation, and suggests that classifying these vocalizations with a simple linear classifier would be challenging, as there is no clear linear separability between the classes.

4. Classification Analysis

Based on the insights of our similarity analysis, we aim to evaluate the saliency of the extracted representations, and proceed to classify them using a simple, non-linear MLP, for the multi-class classification tasks. We implement three blocks of [Linear, LayerNorm, ReLU] layers, with 128, 64, and 32 number of hidden units respectively, followed by a final linear layer to obtain the posterior probabilities. To evaluate the performance we used Unweighted Average Recall (UAR) as the metric to account for any class imbalance. To obtain robust results, we employ the grid search methodology with *Val* UAR score as the optimization criterion. We train the classifier for 30 epochs with cross-entropy loss, and search for the optimal hyperparameters values of η and batch-size across $2^{[5-9]}$ and [1e-3, 1e-4] respectively for each feature–task permutation on *Train* and *Val*. The optimization consists of Adam and a η-scheduler of factor 0.1 and patience of 10 epochs. Lastly, for WavLM, we classify each of the encoder layers [0–13] to identify the optimal layer.

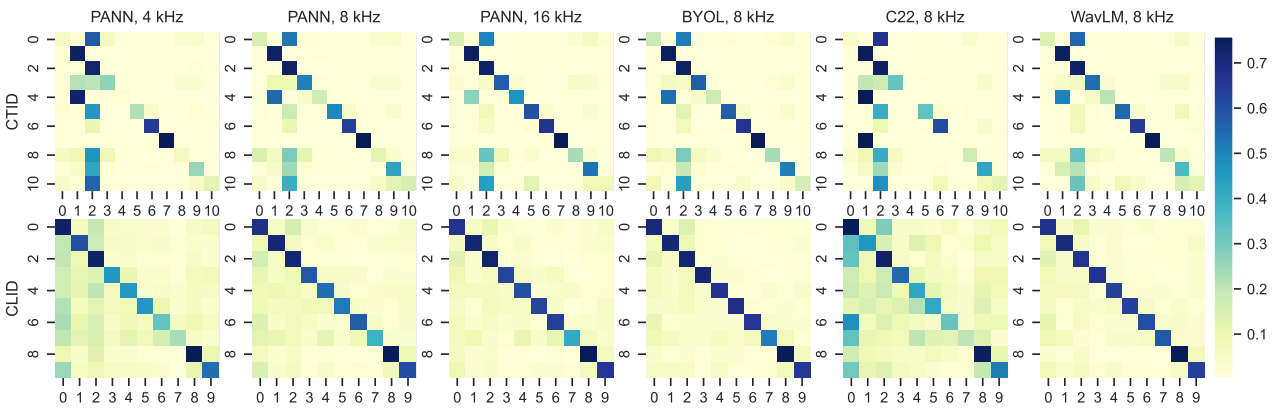

Figure 4: *Normalized confusion matrices with row indices representing true class labels. Darker diagonals signify higher performance.*

Figure 5 presents the layer-wise scores for WavLM, normalized per task to a [0, 1] range. We can observe that the lower layers are clearly much more salient representations for both tasks compared to higher layers. Based on these results, we use the best individual WavLM layers for our two tasks.

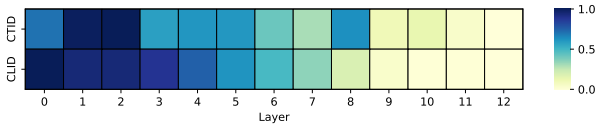

Figure 5: *Layer-wise UAR scores of WavLM features, normalized per task. Darker regions indicate a higher performance.*

Table 3a) summarizes the classification results of the different feature sets at an 8 kHz bandwidth (BW). Random performance is given as 100 over the number of classes. Notably, BYOL features outperform the other features, for both CTID and CLID, despite having fewer parameters than WavLM and PANN, while C22 proves to be the overall weakest representation. WavLM shows the highest difference in performance across tasks. Meanwhile, table 3b) highlights the impact of pre-training bandwidth for salient representations on PANN features. The results clearly show that the bandwidth size correlates directly with the performance, increasing monotonically. Particularly, PANN features at 16 kHz achieve the highest performance across all features and BWs for CTID. BYOL embeddings at 8 kHz notably outperform PANN at 16 kHz for CLID. The best scores for both tasks are also closely matched in value.

Figure 4 shows the classifier's performance through confusion matrices. We can again clearly observe the monotonic improvement in CTID classification performance for PANN features as the bandwidth increases. We also notice a prevalent trend of false positives for call-type ID 2 (Twitter) across all feature sets, especially against IDs 0, 8, and 10, attributable to its high occurrence in the dataset and broad spectral range [32, 33]. The CLID results contain distinctly fewer misclassifications, which aligns with expectations since the call-types are spread among the different callers classes. The exception is C22, which yields the weakest performance. Caller classes with higher data volumes (IDs 0 and 2) perform better compared to the others. Finally, a clear improvement in performance correlated with bandwidth is seen for PANN features, as with CTID.

Table 3: *UAR scores [%] on Test for pre-trained features $\mathcal{F}$. WavLM's best layer's score is given.*

Section	$\mathcal{F}$	BW	CTID	CLID
(a)	Random	-	9.09	10
	C22	8	41.96	35.62
	WavLM	8	59.99	67.47
	BYOL	8	**63.64**	**68.30**
	PANN	8	58.54	56.02
(b)	PANN	4	46.27	41.10
	PANN	8	58.54	56.02
	PANN	16	**69.09**	**65.39**

5. Summary and Conclusion

This paper investigated the utility and limitations of foundations models, pre-trained on human speech or general audio, which have not been demonstrated for marmoset call-type and caller identity multi-class classification. To that end, we conducted and validated two studies across three lines of investigation.

First we conducted a call similarity analysis, which revealed that the features extracted from these models lacked linear separability within or across classes. Then, we conducted a classification study which demonstrated that a non-linear classifier can still achieve substantial performance, and highlighted that a larger bandwidth directly correlates with improved performance. Classification of call-types also appeared to be more sensitive to bandwidth changes than caller identities. Additionally, the pre-training domain of speech and general audio showed comparable performances, with a distinct improvement over handcrafted features. Finally, we obtained close best performance for both call-type and caller classification tasks.

In conclusion, our findings underscore the potential of leveraging pre-trained foundation models for bioacoustic signals, particularly when the model's bandwidth aligns with the biological auditory and vocal range of the studied species. Future collaborative work with biologists and linguistics researchers could explore the biological implications of these results, especially in understanding the evolutionary aspects of marmoset vocal behaviour and their perceptual processing, to bridge the gap between computational models and biological insights in non-human vocal communication research.

6. Acknowledgements

This work was funded by Swiss National Science Foundation's NCCR Evolving Language project (grant no. 51NF40_180888).

7. References

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022.

[2] J. L. Norcross and J. D. Newman, "Context and gender-specific differences in the acoustic structure of common marmoset (callithrix jacchus) phee calls," *American journal of primatology*, vol. 30(1), p. 37–54, 1993.

[3] Y. Zürcher and J. M. Burkart, "Evidence for dialects in three captive populations of common marmosets (callithrix jacchus)," *International Journal of Primatology*, vol. 38, no. 4, pp. 780–793, 2017.

[4] J. BS, H. DHR, and C. CK, "The stability of the vocal signature in phee calls of the common marmoset, callithrix jacchus," *American journal of primatology*, vol. 31(1), pp. 67–75, 1993.

[5] N. Phaniraj, K. Wierucka, Y. Zürcher, and J. M. Burkart, "Who is calling? optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers," *Journal of Royal Society Interface*, 2023.

[6] G. Epple, "Comparative studies on vocalization in marmoset monkeys (hapalidae)," *Folia Primatol (Basel)*, vol. 8, no. 1, pp. 1–40, 1968.

[7] R. Seyfarth and D. Cheney, "Signalers and receivers in animal communication," *Annual review of psychology*, vol. 54, pp. 145–73, 02 2003.

[8] H. Brumm, K. Voss, I. Köllmer, and D. Todt, "Acoustic communication in noise: regulation of call characteristics in a new world monkey," *Journal of Experimental Biology*, vol. 207, no. 3, pp. 443–448, 01 2004.

[9] S. J. Eliades and X. Wang, "Neural correlates of the lombard effect in primate auditory cortex," *Journal of Neuroscience*, vol. 32, no. 31, pp. 10 737–10 748, 2012.

[10] T. Pomberger, C. Risueno-Segovia, J. Löschner, and S. R. Hage, "Precise motor control enables rapid flexibility in vocal behavior of marmoset monkeys," *Current biology*, vol. 28(5), p. 788–794, 2018.

[11] S. Roy, C. T. Miller, D. Gottsch, and X. Wang, "Vocal control by the common marmoset in the presence of interfering noise," *Journal of Experimental Biology*, vol. 214, no. 21, pp. 3619–3629, 11 2011.

[12] D. Takahashi, A. Fenley, and A. Ghazanfar, "Early development of turn-taking with parents shapes vocal acoustics in infant marmoset monkeys," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, p. 20150370, 05 2016.

[13] M. S. Osmanski and X. Wang, "Perceptual specializations for processing species-specific vocalizations in the common marmoset (Callithrix jacchus)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, no. 24, p. e2221756120, 2023.

[14] K. Worley and al., "The common marmoset genome provides insight into primate biology and evolution," *Nature Genetics*, vol. Nat Genet. 2014 Aug;46(8):850-7., p. 850–857, 07 2014.

[15] H. Okano, A. Miyawaki, and K. Kasai, "Brain/minds: brain-mapping project in japan," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 370, 05 2015.

[16] H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque, "Machine learning algorithms for automatic classification of marmoset vocalizations," *PLOS ONE*, vol. 11, pp. 1–14, 09 2016.

[17] A. Wisler, L. J. Brattain, R. Landman, and T. F. Quatieri, "A Framework for Automated Marmoset Vocalization Detection and Classification," in *Proc. Interspeech 2016*, 2016, pp. 2592–2596.

[18] T. O. et al., "Deep convolutional network for animal sound classification and source attribution using dual audio recordings," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 654–662, 2018.

[19] Y. Zhang, J. Huang, N. Gong, Z. Ling, and Y. Hu, "Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks," *The Journal of the Acoustical Society of America*, vol. 144, pp. 478–487, 07 2018.

[20] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. of ICASSP*, 2021, pp. 3875–3879.

[21] P. C. Bermant, L. Brickson, and A. J. Titus, "Bioacoustic Event Detection with Self-Supervised Contrastive Learning," *bioRxiv*, 2022.

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

[23] M. S. Osmanski, X. Song, Y. Guo, and X. Wang, "Frequency discrimination in the common marmoset (callithrix jacchus)," *Hearing Research*, vol. 341, pp. 1–8, 2016.

[24] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics," *Data Mining and Knowledge Discovery*, 2019.

[25] S. C. et al., "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, 2022.

[26] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul 2021.

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

[28] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of The Royal Society Interface*, vol. 10, no. 83, 2013.

[29] A. Paul, H. McLendon, V. Rally, J. T. Sakata, and S. C. Woolley, "Behavioral discrimination and time-series phenotyping of birdsong performance," *PLOS Computational Biology*, vol. 17, no. 4, pp. 1–21, 04 2021.

[30] S. S. Sethi, "Automated acoustic monitoring of ecosystems," Ph.D. dissertation, Imperial College London, UK, 2020.

[31] S. wen Yang et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[32] A. L. Pistorio, B. Vintch, and X. Wang, "Acoustic analysis of vocal development in a new world primate, the common marmoset (callithrix jacchus)," *Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1655–1670, Sep 2006.

[33] J. Agamaite, C. Chang, M. Osmanski, and X. Wang, "A quantitative acoustic analysis of the vocal repertoire of the common marmoset (callithrix jacchus)," *Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2906–2928, Nov. 2015.