On feature representations for marmoset vocal communication analysis

Eklavya Sarkar[a,b,*], Kaja Wierucka[c], Alexandra B. Bossard[d], Judith Burkart[d], Mathew Magimai.-Doss[a]

[a]*Idiap Research Institute, Martigny, 1920, Switzerland.*
[b]*Ecole polytechnique fédérale de Lausanne, 1015, Switzerland.*
[c]*Behavioral Ecology and Sociobiology Unit, German Primate Center, Leibniz Institute for Primate Research Göttingen, 37077, Germany.*
[d]*Department of Comparative Language Science, Institute of Evolutionary Anthropology, and Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich, 8050, Switzerland.*

Abstract

The acoustic analysis of marmoset (*Callithrix jacchus*) vocalizations is often used to understand the evolutionary origins of human language. Currently, the analysis is largely carried out in a manual or semi-manual manner. Thus, there is a need to develop automatic call analysis methods. In that direction, research has been limited to the development of analysis methods with small amounts of data or for specific scenarios. Furthermore, there is lack of prior knowledge about what type of information is relevant for different call analysis tasks. To address these issues, as a first step, this paper explores different feature representation methods, namely, HCTSA-based hand-crafted features Catch22, pre-trained self supervised learning (SSL) based features extracted from neural networks trained on human speech and end-to-end acoustic modeling for call-type classification, caller identification and caller sex identification. Through an investigation on three different marmoset call datasets, we demonstrate that SSL-based feature representations and end-to-end acoustic modeling tend to lead to better systems than Catch22 features for call-type and caller classification. Furthermore, we also highlight the impact of signal bandwidth on the obtained task performances.

Keywords: bioacoustics, marmoset call analysis, feature representation, call-type classification, caller identification, sex classification.

1. Introduction

The advancements in human speech processing have also accelerated and impacted research in non-human communication, such as bioacoustics, i.e. the study of animal sounds. Common marmosets (*Callithrix jacchus*) have recently gained prominence as a valuable research model among non-human primates. This is primarily due to their exceptional vocal abilities, which are rooted in their highly complex social behavior and cooperative breeding system [1, 2]. They possess extensive vocal repertoires used in various social situations [3, 4], and their vocalizations have the capacity to encode a wide range of information, such as population, group affiliation, sex, and even individual identity [5, 6, 7, 8, 9, 10]. These vocalizations are not limited to simple tonal signals but also encompass complex calls with multiple frequency components, some of which are within the ultrasonic range [11]. Moreover, marmosets have been observed to exhibit remarkable vocal adaptability. They can alter the duration [12], intensity [12, 13, 14], complexity [15], or timing [16, 15] of their calls, even when faced with disruptions in their environment that occur after the initiation of a call [14]. While these properties make marmosets an intriguing subject for the study of communication processes, they also pose a significant challenge when attempting to automate the analysis of their vocalizations. The literature on automatic marmoset vocalization analysis is relatively sparse.

Turreson et al. compared different classification methods for marmoset 'call-type' classification using linear prediction coefficients as feature representation, and found that on a small data setup of 30 samples per call-type, k-NN, SVM and optimal path forest algorithms yield better performance than multilayer perceptron, Adaboost, and logistic regression [17]. Wisler et al. investigated different feature representations, namely, audio features (statistics based on energy entropy, signal energy, zero crossing rate, spectral rolloff, spectral centroid, and spectral flux), mel-frequency cepstral coefficients (MFCCs), and Teager energy operator-based features for marmoset vocalization and call-type detection [18]. On a synthetic dataset, created by taking a small set of calls and augmenting it with background noise and acoustic events, it was found that feature level combination leads to better performance. Verma et al. investigated discovering of different patterns in marmoset calls through unsupervised learning. Specifically, they developed an HMM-based approach to segment and cluster marmoset vocalizations into discrete units through multi-resolution and multi-rate analysis of the signal [19]. In [20], it was demonstrated that marmoset vocalizations and call-types can be better detected/classified by feeding statistics of log-mel-filter bank energies as input to recurrent neural networks, when compared to feeding it to SVM or multilayer perceptrons. In the scenario of analyzing recordings obtained from a pair of marmosets, [21] investigated a

*Corresponding author.
Email addresses: eklavya.sarkar@idiap.ch (Eklavya Sarkar, kwierucka@dpz.eu (Kaja Wierucka), alexandra.bosshard@uzh.ch (Alexandra B. Bossard), judith.burkart@aim.uzh.ch (Judith Burkart), mathew@idiap.ch (Mathew Magimai.-Doss)

deep learning approach where a spectrogram was fed as input to a convolutional neural network to jointly perform vocalization, detection, call type classification and caller detection. It was found that joint modeling yielded better performance than training systems individually for each task in this scenario. Recently, Highly Comparable Time-Series Analysis (HCTSA) features have been used to model source (caller) identification through an Adaboost-based hierarchical approach for marmosets [10], as well as for 14 mammalian species [22]. In the bioacoustics field, breakthroughs in self-supervised learning (SSL), which leverages unlabeled data by creating surrogate labels from the data's inherent structure, has led to works which explore birdsong detection [23] and bioacoustic event detection [24] by pre-training with a contrastive learning approach. In that direction, a study using different SSLs pre-trained on human speech, demonstrated that neural embeddings extracted in such a framework can also distinguish marmoset callers [25].

However, in the existing works, there are three main limitations. First, most of the studies have been carried out on small data sets. Second, these studies have been conducted on datasets intended for specific scenarios. Due to a lack of validation, it is unclear whether the methods studied on one dataset would scale to another. Third, there is limited prior knowledge about what type of information is relevant for different call analysis tasks. There is a need to overcome these limitations to advance the development of automatic analyses of marmoset vocalizations. The present paper is a step in that direction with a specific focus on feature representations for automatic marmoset call analyses, where we investigate three prominent feature representation methods, namely, (a) hand-crafted features, (b) self-supervised learning-based representations, and (c) end-to-end acoustic modeling, on three different marmoset call datasets and three different tasks (call type, caller identity, and caller sex classification).

The paper is organized as follows. Section 2 presents the different datasets, tasks, and investigated feature representations. Section 3 and 4 present the studies and analysis of the results respectively. Finally section 5 concludes the paper.

2. Methodology

2.1. Datasets and tasks

We conduct investigations on three different marmoset datasets, denoted as D_1, D_2, and D_3, respectively. D_2 and D_3 contain vocalizations produced by adult individuals, while D_1, InfantMarmosetsVox, originates from infant marmosets [25]. Consequently, D_1 is expected to encompass different call types, likely characterized by higher frequencies compared to those in D_2 and D_3. Furthermore, D_2 and D_3 are gathered from the same colony, while D_1 was obtained from a different one. All the datasets consist of audio recordings of marmosets vocalizations segments, collected and hand-labeled with the start and end time by experienced researchers. In addition to call-type and caller identity annotations of each vocalization provided for all three datasets, D_1 and D_2 also include information about the sex of the vocalizing individual. For more details regarding the datasets, the reader is referred to Appendix A.

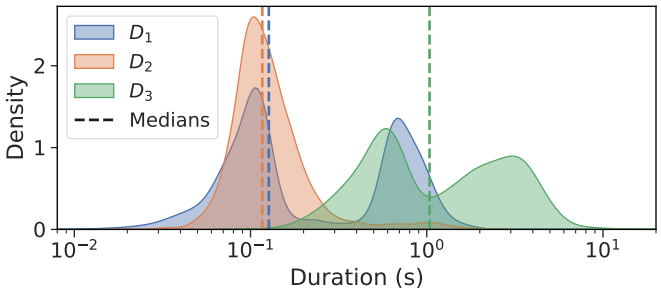

Figure 1: Log distribution of vocalization lengths per dataset. The medians are calculated over the entirety of each dataset.

We discard any segments labeled as 'silence' and 'noise', and only keep the vocalization segments. The log distribution of the vocalization lengths of the three datasets is presented in Figure 1. We can observe that D_1 has the shortest median vocalization length at 127 ms, with D_2 and D_3 at 175 and 1037 ms respectively. Based on the given annotations, we define multi-class tasks, specifically call-type, caller, and sex classification, henceforth referred to as CTID, CLID, and SID respectively. Table 1 gives the number of vocalization segments S, their total duration length L, the native sampling rates, as well as the number of classes n_c for each task across datasets.

Table 1: S indicates the number of data samples, L the sum of all vocalizations segment durations (in minutes), and SR the native sampling rate of the given data (kHz). n_{task} is the number of classes of each task-dataset permutation.

$\mathcal{D}$	S	L	SR	n_{CTID}	n_{CLID}	n_{SID}
D_1 [25]	73K	464	44.1	11	10	-
D_2 [26]	14K	37	300	7	8	2
D_3	5K	138	125	12	8	2

2.2. Feature representations

We investigate the following feature representations:

1) *Hand-crafted features*: Highly Comparable Time-Series Analysis (HCTSA) is an interpretable signal processing-based framework that has been demonstrated to be useful for diverse time series application domains [27]. In this framework, a set of 7700 features are extracted by characterizing the signal by different time series analysis methods, such as, linear correlation, modeling fitting (e.g., autoregressive moving average analysis, GARCH), wavelet analysis, extraction of information theoretic measures, which then is combined with feature selection to build statistical models for the end task. In the literature, these features have been investigated for behavioural birdsong discrimination [28], automated acoustic monitoring of ecosystems [29], as well as marmoset caller identification [10]. One of the challenges of HCTSA approach is computational complexity and involves an evaluation of many similar features. In a recent work, CAnonical Time-series CHaracteristics (Catch22) features, a subset of the HCTSA feature set has been proposed which exhibit a strong performance across 93 real-world time-series classification problems, but are also minimally redun-

dant [30]. In this work, we investigate the Catch22 features, denoted as C22.

2) *Pre-trained self-supervised learning (SSL) based features*: Inspired from the recent study presented in [25], we investigate the use of feature representations extracted from pre-trained SSL neural networks trained on human speech for marmoset call analysis. We extend the investigations from caller detection to call type, caller ID and sex classification. Furthermore, contrary to the previous work [25], which focused only on the last transformer layer representation, in this work we investigate representations obtained from all the transformer layers to gain insight which level of layer representations are informative for marmoset call analysis.

3) *End-to-end acoustic modeling*: With advances in deep learning, acoustic modeling approaches have emerged in speech and audio processing where raw signal can be modeled to learn task-dependent information from the signal in an end-to-manner with minimum prior knowledge [31, 32, 33, 34]. Such approaches hold potential for advancing marmoset call analysis, as they could help not only in addressing the lack of reliable task-dependent prior knowledge challenge, but also in gaining insight into the task relevant acoustic information learned by such trained networks through analysis [34, 35, 36]. The insight gained could then be further validated through linguistic studies. Motivated by these aspects, we investigate this approach.

A sub-challenge that arises when analyzing marmoset calls is the range of frequency information to be modeled. More precisely, the fundamental frequencies (typically corresponding to the peak frequency) of adult marmoset vocalisations span a range of 6-13 kHz, depending on the *call-type* [3]. However, as can be seen in Table 1, datasets are collected at varying sampling frequencies. Furthermore, the SSL neural networks are typically pre-trained on speech signal of 8 kHz bandwidth (i.e., 16 kHz sampling frequency). As part of the investigation, we thus also study the impact of sampling rate (SR) on marmoset call analysis tasks.

3. Experimental Study

3.1. Systems

For each task, we divided all datasets into training, validation, and test sets, named *Train*, *Val*, and *Test* respectively, following a 70:20:10 split ratio, in order to train models on a sufficiently large number of samples, while ensuring sufficient data points for model evaluation and validation. *Train* is used to train the models, *Val* to tune any hyperparameters, and *Test* to evaluate the trained models on unseen data. We then developed the following systems for each task on each dataset to investigate the aforementioned feature representations:

1) We used *pycatch22* to extract a feature vector $\boldsymbol{x} \in \mathbb{R}^{1\times D}$ (denoted as C22) for each utterance, where $D = 24$, and feed it to a multilayer perceptron (MLP) with three hidden layers of 128, 64, and 32 number of hidden units, respectively. The classifier is trained for 30 epochs, using a batch size 16 and learning rate $\eta = 1e-3$.

2) As it is challenging to investigate all the different types of pre-trained SSL feature representations across all tasks and datasets, we simply chose WavLM [37], as it was found to yield strong performance on the task of marmoset caller detection [25], been found to scale well to different human speech processing tasks in the SUPERB challenge [38]. For each layer, we extracted frame-by-frame variable-length feature representations $\boldsymbol{x} \in \mathbb{R}^{N\times D}$, where $D = 768$ and N the variable number of frames (contingent on the vocalization length). We then converted these embeddings into utterance-level fixed-length representations $\boldsymbol{f}_{\mu\sigma} \in \mathbb{R}^{1\times 2D}$ (denoted as WLM), by computing and concatenating the first and second order statistics across the frame axis on the extracted features. An MLP of same three layer architecture as *C22* is then trained with the fixed length feature as input.

3) We trained a convolutional neural network (CNN) based end-to-end acoustic modeling system (denoted as E2E) that takes a raw waveform as input and classifies to the output classes. Following the literature in speech processing [39, 40, 41], the E2E system consists of four convolution layers followed by an adaptive pooling layer and two hidden layers. The E2E system is optimized with a cross-entropy cost function with an early stopping criteria. Further details of the architecture are provided in the Appendix B.

In the case of *C22*, we developed systems at native sampling frequency and downsampled acoustic signals: 16 kHz for D_1, 60 and 16 kHz for D_2, and 60 and 16 kHz for D_3. In the case of WLM, we developed systems with signals downsampled to required pre-training sampling rate of 16 kHz. For E2E system, D_2 and D_3 signals were downsampled to 60 and 16 kHz. To evaluate the systems we used Unweighted Average Recall (UAR) as the metric to account for any class imbalance.

3.2. Results

Table 2 shows the performances of systems based on different feature representations. For the sake of clarity, only the best layer and worst layer performances are reported for WLM. Figure 2 presents the layer-wise performances for all tasks on all datasets for WLM. Note that layer 0 corresponds to the output embedding of the CNN encoder, where as the other 12 refer to the outputs of the transformer encoder layers. The performances are all above chance level, i.e. $100/n_c$, for all systems.

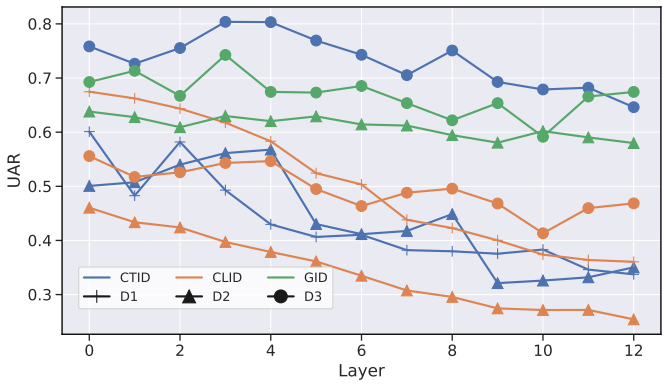

Figure 2: Layer-wise UAR scores for WLM for all tasks and datasets. The layers follow the same indexing as [37].

Table 2: UAR scores on *Test* on features $\mathcal{F}$. WavLM's best and worst layer's score is given. For each dataset, the best score across features is bolded per task.

$\mathcal{D}$	$\mathcal{F}$	SR	CTID	CLID	SID
D_1	C22	44.1	51.04	47.58	N/A
		16	37.72	34.54	N/A
D_1	WLM	16	60.10	67.47	N/A
			33.74	36.05	N/A
D_1	E2E	44	**68.32**	**74.12**	N/A
		16	53.03	59.94	N/A
D_2	C22	300	37.68	43.56	**66.24**
		60	32.50	35.52	63.38
		16	35.65	35.32	58.14
D_2	WLM	16	**56.77**	46.05	63.80
			32.11	25.42	57.98
D_2	E2E	60	42.03	**49.78**	62.36
		16	37.65	36.21	60.15
D_3	C22	125	64.32	43.19	62.80
		60	65.67	45.50	61.22
		16	52.59	39.43	57.32
D_3	WLM	16	**80.38**	**55.58**	**74.26**
			64.62	41.33	59.14
D_3	E2E	60	65.31	47.92	60.73
		16	66.24	31.31	56.59

Ignoring the sampling frequency aspect, it can be observed that E2E yields the best performances for D_1's CTID and CLID tasks. For D_2, WLM yields best performance for CTID, E2E for CLID, and C22 for SID. On both D_1 and D_2, we can observe that WLM yields competitive systems, however in the case of D_3, WLM's third layer representations consistently yield the best performance across all the tasks (see Figure 2), and outperform C22 and E2E. Although WLM yields competitive performances on D_1 and D_2, it is difficult to systematically compare to C22 or E2E as different layers yield best performance for different tasks.

Furthermore, it can be observed that the 16 kHz SR performance is generally inferior across different datasets and tasks for C22 and E2E. This finding is in line with the understandings in the literature gained by analysis of different call types which showed that most marmoset call types extend into frequencies above 8 kHz [3]. This implies that, with an 8 kHz bandwidth, certain vital information for specific call types might be lost, rendering it increasingly challenging, if not impossible, for the classifier to accurately categorize certain calls. Indeed, it can be observed that C22 systems yield superior performance with the native SR compared to 16 kHz for all datasets. This emphasizes that higher frequencies are likely to contain valuable information. A comparison between C22, WLM and E2E at 16 kHz sampling frequency demonstrates the potential of SSL based feature representations learned on human speech.

It is worth noting that a recent, independent study explored representations learned from other acoustic domains such as general audio, which includes audio event classes such as environmental sounds, musical instruments, and human and animal vocalizations. They demonstrated on D_1 that increasing the pre-training bandwidth of a PANN model [42], pre-trained on the AudioSet dataset with log-mel spectrogram inputs, improved performance on both CTID and CLID tasks [43]. However, the study didn't explicitly disentangle whether these improvements resulted from the increased bandwidth itself, the spectrogram-based inputs, or from the inclusion of some animal vocalizations in the pre-training dataset. This distinction still remains an important open question for future investigations.

4. Analysis

4.1. Layer-wise linear performance analysis

In Figure 2, it can be observed that lower layer representations tend to yield better systems. To further ascertain that, we carried out layer-wise classification performance of the same tasks using a simple linear classifier (single layer perceptron). Figure 3 shows the results independently normalized per-task to a [0, 1] range. It can be observed that the lower layers are much more salient representations for all three tasks across all datasets when compared to higher layers. A possible explanation is that, because WavLM's CNN encoder operates directly on the raw waveform, the early layers capture fundamental *acoustic* features and can leverage spectro-temporal variations relevant to tasks such as speaker identification and verification [37]. Thus, these lower layers inherently generalize better to other acoustic domains, such as marmoset vocalizations. In contrast, the later layers – shown to perform well on *linguistic* tasks, such as speech or phoneme recognition – appear more specialized for human speech and consequently much less transferable to bioacoustics, resulting in lower performance. We can also observe that there is no consistent optimal layer for each task type across the datasets.

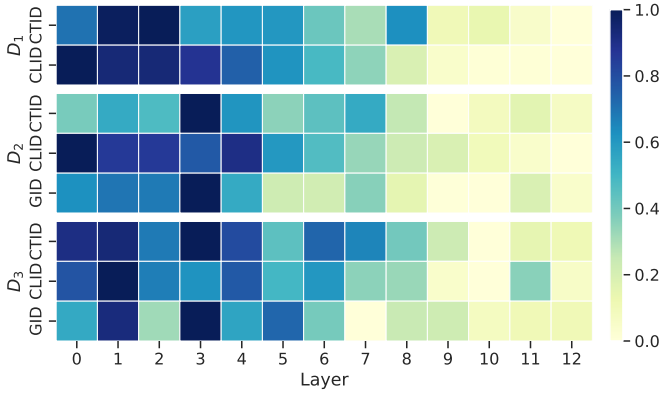

Figure 3: Layer-wise UAR scores of WLM features modeled by single layer perceptron. The scores are normalized independently per task. Darker regions indicate higher performance.

4.2. Frequency response of learnt convolution filters

We analyzed the frequency response of the first learnt convolution layer filters of E2E systems by estimating the cumulative frequency response F_{cum} as [36]:

$$F_{cum} = \sum_{k=1}^{n_f} \frac{F_k}{\|F_k\|_2}, \quad (1)$$

where n_f denotes the 128 filters in the first convolution layer (see Appendix) and F_k denotes discrete Fourier transform of filter k over 2048 DFT points.

Figure 4 shows the cumulative frequency response for each task per dataset at an SR of 16 kHz, and 44.1 or 60 kHz. With a 8 kHz bandwidth (left half), it can be observed that the emphasis is on frequencies 4-5 kHz and above irrespective of the task. As the bandwidth of the signal is increased (right half), it can be observed that emphasis is also given to higher frequency regions such as around 10 kHz or above. These observations further corroborate previous findings that most marmoset calls occupy frequency ranges beyond 8 kHz [3], and also explain the improved performance obtained with higher bandwidth signals. In addition, we observe that for different tasks the learned filters give emphasis to different frequency regions. A detailed analysis of the spectral information learned is part of our future work. Taken together, the analysis indicates that the E2E framework inspired from speech processing can be scaled to marmoset call analysis.

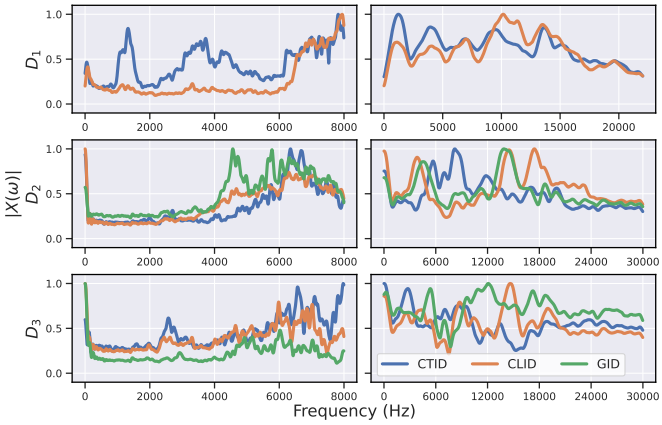

Figure 4: Cumulative frequency response per task on all datasets. Sampling rate: 16 kHz (left), and 44.1 or 60 kHz (right).

5. Conclusions

This paper explored different feature representations or learning methods, namely handcrafted feature Catch22, SSL feature representation WLM, and end-to-end acoustic modeling (E2E) for analyzing marmoset calls. Our investigations on three different datasets demonstrate that end-to-end acoustic modeling and SSL feature representations yield better systems than handcrafted Catch-22 features for call-type classification and caller identification, while also achieving comparable performances for sex identification at a common sampling rate. As a by-product, our studies demonstrated that (a) the utility of pre-trained SSL models on human speech can be extended to call-type and sex, besides caller discrimination and (b) end-to-end acoustic modeling methods developed for speech processing can be scaled for marmoset call analysis. Our study raises a few pertinent questions such as: (a) with limited signal bandwidth how are SSL features informative about marmoset calls? (b) what kind of task specific spectral information is learned by the E2E systems?, and (c) how to combine the different approaches for improving marmoset call analysis? Furthermore, in this work we only investigated feature representations that directly modeled the raw input waveform. However, recent bioacoustic studies on bats, birds, and rodents have leveraged spectrogram-based methods [44, 45, 46, 47]. Whether such approaches can offer distinct advantages over the waveform-based methods for marmoset vocal communication analysis remains to be determined. Our future work will investigate these questions.

Data Availability

The Dataset D_1 and a corresponding PyTorch Dataloader is publicly available on Zenodo[1], with reference number `10130104`. In addition, the datasets D_2 and D_3 are both available from the corresponding authors upon request.

CRediT Authorship Contribution Statement

The following paragraph details the CRediT (Contributor Roles Taxonomy) designations for each author:

ES: Conceptualization, Data Curation, Methodology, Investigation, Formal Analysis, Writing – original draft Preparation, Writing – review and editing, Resources. **KW**: Data curation, Formal Analysis, Writing – review and editing, Resources. **ABB**: Data Curation, Writing – review and editing, Resources. **JB**: Writing – review and editing, Supervision, Resources, Funding acquisition. **MMD**: Conceptualization, Methodology, Writing – original draft, Writing – review and editing, Supervision, Resources, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Generative AI Disclosure Statement

The authors acknowledge using OpenAI's ChatGPT (models `GPT-4o` and `o1`) solely for language editing and improving readability. All results, analyses, interpretations, conclusions, and overall research findings presented in this paper remain exclusively the original work of the authors.

[1] `https://zenodo.org/records/10130104`

Acknowledgements

This work was funded by the Swiss National Science Foundation's (SNSF) National Centre of Competence in Research (NCCR) Evolving Language project (grant 51NF40_180888).

Appendix A. Data description

Dataset D_1 is an extended version of the dataset used in the study on marmoset call type discrimination by Zhang et al. [20]. This version, entitled InfantMarmosetsVox, was used in the recent work on marmoset caller discrimination using SSL features [25]. The audio was recorded from five pairs of infant marmoset twins, each recorded individually in two separate sound-proofed recording rooms at a sampling rate of 44.1 kHz. Additionally, marmosets were recorded individually without communication with other marmosets and the intervention from experimenters. The audio recordings were manually annotated using the Praat tool by an experienced researcher. For each vocalization, the start and end time, call type, and marmoset identity have been provided. The data consists of 11 different marmoset calltypes, namely, peep (pre-phee), phee, twitter, trill, trillphee, tsik tse, egg, pheecry (cry), trllTwitter, pheetwitter, and peep. The data contains 350 files of precisely labelled 10-minute audio recordings across all ten caller classes.

D_2 consists of 102 labelled 10-min focal audio recordings of common marmoset calls recorded in six behavioural contexts. A pair of marmosets was either separated or in the same enclosure, with preferred food either freely available for the focal individual or not. Each of the 8 subjects was recorded on 16 separate occasions. Most of the calls were given in bouts as holistic single call units, and thus, a call-type unit was defined as a call bout with call elements which were not further apart than 0.5s, as per existing literature [3, 48]. We only used the segments labelled as single call elements, i.e. not split up in bouts, to avoid data overlap and duplication. The dataset consists of 7 calls, namely alarm, ek, food, phee, trill, tsk, and twitter. The audio recordings were manually annotated by using Avisoft SASLab Pro (Avisoft Bioacoustics, Feb. 2017) to narrowly label the start and end of each call-type. The data was collected under Swiss legislation and licensed by Zurich's cantonal veterinary office (license ZH 223/16 and ZH 232/19).

D_3 was collected from 6 target adult common marmosets, 3 male and 3 female, housed at the University of Zurich. Two additional non-target individuals were also included in the dataset, summing to 8 individuals in total. The data consists of 12 calls classes: phee, trill, food call, tsk, low tsk (tsk with a peak frequency of approximately 7-9 kHz), twitter (sequence), ek, phee sequence (multiple phees), low tsk sequence (multiple low tsks), ek sequence (multiple eks), food call sequence (multiple food calls). All procedures were done in accordance with Swiss legislation and were licensed by Zurich's cantonal veterinary office (license ZH223/19). For each recording, two individuals (one male and one female) were placed in adjacent wire cages and recorded simultaneously in 15-minute intervals with two UltraSoundGate 116H recorders coupled with an Avisoft CM16/CMPA condenser microphone (Avisoft Bioacoustics, Germany), each set to a different gain to capture both low and high amplitude calls with a sampling rate of 125kHz. A total of 12 recordings, spread over 7 months, were made for each target individual. Caller identity was labeled in real time using Avisoft-RECORDER USGH (Avisoft Bioacoustics, Germany). The labelling of the calls' exact start and end points was carried out through a visual examination of the spectrograms. For inclusion in subsequent analyses, calls needed be distinctly visible on the spectrogram, devoid of any interference from other calls, and readily classifiable into specific call-type categories.

Appendix B. CNN architecture

Table B.3 presents the architecture of the E2E system. The first convolution layer kernel width kW and shift dW was chosen based on the sampling frequency. More precisely, based on the understanding gained from speech studies, we chose those hyper-parameters to strike a balance between the length of the convolution filter and enough pitch cycles being modeled [34]. For 44.1 and 60 kHz sampling frequency, we chose kW = 1 ms and dW = 0.05 ms, respectively. As marmoset calls have fundamental frequency around 5 kHz and above [3], 1 ms signal would be expected to contain around 10 pitch cycles or more. However, for 16 kHz sampling frequency, 1 ms would contain only 16 samples, i.e. at the most 1-2 sample(s) representing each pitch cycle. This may not hinder capturing the pitch frequency information in the marmoset call well. So, for 16 kHz we set kW = 10 ms and dW = 0.5 ms. The training batch size 16 and learning rate of 0.001, same as the MLP classifier for C22 and WLM. The optimization configuration simply consisted of Adam and a dynamic learning rate scheduler which reduces the learning rate η when the selected optimization criterion, in this case *Val* UAR, shows no improvement after 10 epochs.

Table B.3: CNN model parameters. n_f denotes the number of filters, n_{hu} the the number of hidden units, and σ the activation function.

Layer	kW	dW	n_f/n_{hu}	**Padding**	σ
Conv 1	kW	dW	128	-	ReLU
Conv 2	10	5	256	-	ReLU
Conv 3	4	2	512	2	ReLU
Conv 4	3	1	512	1	ReLU
Adapt	-	-	-	-	-
FC 1	-	-	512	-	ReLU
FC 2	-	-	256	-	ReLU
FC 3	-	-	n_c	-	-

References

[1] S. J. Eliades, C. T. Miller, Marmoset vocal communication: Behavior and neurobiology, Developmental Neurobiology 77 (3) (2017) 286–299.

[2] J. M. Burkart, J. E. C. Adriaense, R. K. Brügger, F. M. Miss, K. Wierucka, C. P. van Schaik, A convergent interaction engine: vocal communication among marmoset monkeys, Philosophical Transactions of the Royal Society B: Biological Sciences 377 (1859) (2022) 20210098.

[3] J. A. Agamaite, C. J. Chang, M. S. Osmanski, X. Wang, A quantitative acoustic analysis of the vocal repertoire of the common marmoset (callithrix jacchus), The Journal of the Acoustical Society of America 138(5) (2015) 2906–2928.

[4] B. Bezerra, A. Souto, Structure and usage of the vocal repertoire of callithrix jacchus, International Journal of Primatology 29 (2008) 671–701.

[5] Y. Zürcher, J. M. Burkart, Evidence for dialects in three captive populations of common marmosets (callithrix jacchus), International Journal of Primatology 38 (4) (2017) 780–793.

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

[7] G. P. Newman JD, Noncategorical vocal communication in primates: the example of common marmoset phee calls, Nonverbal vocal communication (eds A Manstead, K Oatley) (1992) 87–101.

[8] M. Rukstalis, J. French, Vocal buffering of the stress response: exposure to conspecific vocalizations moderates, urinary cortisol excretion in isolated marmosets, Hormones and behavior 47 (2005) 1–7.

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

[10] N. Phaniraj, K. Wierucka, Y. Zürcher, J. M. Burkart, Who is calling? optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers, Journal of Royal Society Interface 20 (207) (2023).

[11] B. J, L. JAM, Ultrasonic components of vocalizations in marmosets, Handbook of ultrasonic vocalization (ed. S Brudzynski) (2018) 535–544.

[12] H. Brumm, K. Voss, I. Köllmer, D. Todt, Acoustic communication in noise: regulation of call characteristics in a new world monkey, Journal of Experimental Biology 207 (3) (2004) 443–448.

[13] S. J. Eliades, X. Wang, Neural correlates of the lombard effect in primate auditory cortex, Journal of Neuroscience 32 (31) (2012) 10737–10748.

[14] T. Pomberger, J. Löschner, S. R. Hage, Compensatory mechanisms affect sensorimotor integration during ongoing vocal motor acts in marmoset monkeys, The European journal of neuroscience 52(6) (2020) 3531–3544.

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

[16] S. Roy, C. T. Miller, D. Gottsch, X. Wang, Vocal control by the common marmoset in the presence of interfering noise, Journal of Experimental Biology 214 (21) (2011) 3619–3629.

[17] H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, V. H. C. de Albuquerque, Machine learning algorithms for automatic classification of marmoset vocalizations, PLOS ONE 11 (2016) 1–14.

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

[19] S. Verma, K. Prateek, K. Pandia, N. Dawalatabad, R. Landman, J. Sharma, M. Sur, H. A. Murthy, Discovering Language in Marmoset Vocalization, in: Proc. Interspeech 2017, 2017, pp. 2426–2430.

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

[21] T. Oikarinen, K. Srinivasan, O. Meisner, J. B. Hyman, S. Parmar, R. Desimone, R. Landman, G. Feng, Deep convolutional network for animal sound classification and source attribution using dual audio recordings, The Journal of the Acoustical Society of America 145 (2) (2018) 654–662.

[22] K. Wierucka, D. Murphy, S. Watson, N. Falk, C. Fichtel, J. León, S. Leu, P. Kappeler, E. Briefer, M. Manser, N. Phaniraj, M. Scheumann, J. Burkart, Same data, different results? evaluating machine learning approaches for individual identification in animal vocalisations, bioRxiv (2024).

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

[24] P. C. Bermant, L. Brickson, A. J. Titus, Bioacoustic Event Detection with Self-Supervised Contrastive Learning, bioRxiv (2022).

[25] E. Sarkar, M. Magimai.-Doss, Can self-supervised neural representations pre-trained on human speech distinguish animal callers?, in: Proc. of Interspeech, 2023, pp. 1189–1193.

[26] A. B. Bosshard, Sequential dynamics in common marmoset vocal strings, Master's thesis, University of Zurich (2020).

[27] B. D. Fulcher, M. A. Little, N. S. Jones, Highly comparative time-series analysis: the empirical structure of time series and their methods, Journal of The Royal Society Interface 10 (83) (2013).

[28] A. Paul, H. McLendon, V. Rally, J. T. Sakata, S. C. Woolley, Behavioral discrimination and time-series phenotyping of birdsong performance, PLOS Computational Biology 17 (4) (2021) 1–21.

[29] S. S. Sethi, Automated acoustic monitoring of ecosystems, Ph.D. thesis, Imperial College London, UK (2020).

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

[31] D. Palaz, R. Collobert, M. Magimai-Doss, Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks, in: Proc. of Interspeech, 2013, pp. 1766–1770.

[32] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: Proc. of ICASSP, 2016, pp. 5200–5204.

[33] R. Zazo, T. N. Sainath, G. Simko, C. Parada, Feature learning with raw-waveform cldnns for voice activity detection, in: Proc. of Interspeech, 2016, pp. 3668–3672.

[34] H. Muckenhirn, M. Magimai.-Doss, S. Marcel, Towards directly modeling raw speech signal for speaker verification using cnns, in: Proc. of ICASSP, 2018, pp. 4884–4888.

[35] H. Muckenhirn, V. Abrol, M. Magimai-Doss, S. Marcel, Understanding and visualizing raw waveform-based cnns, in: Proc. of Interspeech, 2019, pp. 2345–2349.

[36] D. Palaz, M. Magimai-Doss, R. Collobert, End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition, Speech Communication 108 (2019) 15–32.

[37] S. C. et al., WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing, IEEE Journal of Selected Topics in Signal Processing 16 (6) (2022).

[38] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, H. yi Lee, Superb: Speech processing universal performance benchmark, in: Proc. of Interspeech, 2021, pp. 1194–1198.

[39] S. P. Dubagunta, B. Vlasenko, M. Magimai.-Doss, Learning voice source related information for depression detection, in: Proc. of ICASSP, 2019, pp. 6525–6529.

[40] V. S. Nallanthighal, Z. Mostaani, A. Härmä, H. Strik, M. Magimai.-Doss, Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings, Neural Networks 141 (2021) 211–224.

[41] T. Purohit, S. Yadav, B. Vlasenko, S. P. Dubagunta, M. Magimai.-Doss, Towards learning emotion information from short segments of speech, in: Proc. of ICASSP, 2023, pp. 1–5.

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

[43] E. Sarkar, M. Magimai.-Doss, On the utility of speech and audio foundation models for marmoset call analysis, in: 4th International Workshop on Vocal Interactivity In-and-between Humans, Animals and Robots (VIHAR2024), 2024, pp. 7–11.

[44] J. Goffinet, S. Brudner, R. Mooney, J. Pearson, Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires, eLife 10 (2021) e67855. doi:10.7554/eLife.67855.

[45] Z. J. Ruff, D. B. Lesmeister, C. L. Appel, C. M. Sullivan, A convolutional neural network and r-shiny app for automated identification and classification of animal sounds, bioRxiv (2020).

[46] K. R. Coffey, R. G. Marx, J. F. Neumaier, Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations, Neuropsychopharmacology 44 (2019) 859–868.

[47] N. Gu, K. Lee, M. Basha, S. Kumar Ram, G. You, R. H. R. Hahnloser, Positive transfer of the whisper speech transformer to human and animal

voice activity detection, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 7505–7509.

[48] C. T. Snowdon, A. M. Elowson, 'babbling' in pygmy marmosets: Development after infancy, Behaviour 138(10) (2001) 1235–1248.