Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?

Eklavya Sarkar$^{1,2}$ and Mathew Magimai-Doss$^1$

$^1$Idiap Research Institute, Martigny, Switzerland
$^2$Ecole polytechnique fédérale de Lausanne, Switzerland

{eklavya.sarkar, mathew}@idiap.ch

Abstract

Self-supervised learning (SSL) models use only the intrinsic structure of a given signal, independent of its acoustic domain, to extract essential information from the input to an embedding space. This implies that the utility of such representations is not limited to modeling human speech alone. Building on this understanding, this paper explores the cross-transferability of SSL neural representations learned from human speech to analyze bio-acoustic signals. We conduct a caller discrimination analysis and a caller detection study on Marmoset vocalizations using eleven SSL models pre-trained with various pretext tasks. The results show that the embedding spaces carry meaningful caller information and can successfully distinguish the individual identities of Marmoset callers without fine-tuning. This demonstrates that representations pre-trained on human speech can be effectively applied to the bio-acoustics domain, providing valuable insights for future investigations in this field.

Index Terms bio-acoustics, self-supervised learning, caller discrimination and detection, representation learning.

1. Introduction

The study of animal vocalizations, or bio-acoustics, has progressed significantly in recent years due to approaches inherited from machine learning and deep learning [1]. However, most of these are supervised approaches, which require large amounts of labeled data, which is often scarce in bio-acoustics. Self-supervised representation learning (SSL) has emerged as a powerful tool in speech processing to leverage unlabeled data by pre-training models to solve pretext tasks using surrogate labels created from the structure inherent to the data itself. Given an acoustic waveform signal as input, an SSL model uses said labels and the pretext task to train and iteratively optimize its learning objective. The information encoded in the representations can vary depending on the selected learning objective, which can be roughly categorized into generative and discriminative approaches. Generative methods try to either reconstruct masked acoustic frames [2, 3, 4], or predict future frames using an auto-regressive framework [5, 6]. Discriminative approaches either learn by contrastive learning, i.e. discriminating positive samples from negative ones [7, 8], or else by predicting pseudo-labels of discrete masked regions [9, 10, 11] or the output of specific hidden layers [12]. The representations learnt from the chosen SSL model can then be further fine-tuned to a wide range of speech downstream tasks, which have yielded state-of-the-art results on the SUPERB benchmark [13].

Self-supervised learning only utilizes the intrinsic structure of unlabeled data without any reliance on domain-specific knowledge, such as human speech production, to capture essential information about the input data, and extract high-level representations in an embedding space. Thus, the utility of such representations may not only be restricted for modeling human speech, as demonstrated by recent works on other acoustic domains such as music [14, 15] and biomedical signals [16, 17]. Given this understanding, and the fact that both humans and animals have a voice production system, our objective is to investigate the cross-transferability of representations learned from human speech for analyzing animal vocalizations.

To that end, we conduct an animal caller detection study on Marmoset (Callithrix jacchus) vocalizations, and demonstrate its applicability through means of eleven different SSL models pre-trained with different pretext tasks. Our study also aims to provide practical benefits to biologists and ethologists by providing a framework to distinguish individual identities within the same animal species, which is an understudied topic in bio-acoustics and a much harder problem than across-species classification [1]. Some previous works has explored birdsong detection [18] and bio-acoustic event detection [19] using contrastive learning, however, the generalization of SSL models to animal vocalizations has largely remained unexplored. To the best of our knowledge, no previous study has looked into caller detection by utilizing the embedding space learnt by pre-training on human speech.

2. Study Design

This section presents the study design to systematically investigate the cross-transferability of representations learned from human speech for animal caller detection. Specifically, we design a study with the following research questions:

1. How discriminative are the embedding spaces of SSL models pre-trained on human speech?
2. Can we systematically detect individual Marmoset callers using said embedding space?

The remainder of the section presents the dataset, research framework, and selection of SSL models for our investigations.

2.1. Dataset

For our study, we requested and used the marmoset dataset collected and labeled by [20]. The dataset contains audio recordings of eleven different marmoset calltypes, such as Twitters, Phees, and Trills, manually annotated using the Praat tool. 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. The start and end time, call type, and marmoset identity of each vocalization are provided, labeled by an experienced researcher. The data contains 350 files of precisely labelled 10-minute audio recordings across all caller classes. We downsample the data to 16

kHz.
kHZ, remove all segments labeled as ‘silence’ and ‘noise’, and only keep the vocalization segments, amounting to a total of 464 minutes over 72,921 vocalization segments, with a mean and median length of 383 ± 375 ms and 127 ms respectively. Figure 1 shows the imbalanced distribution of vocalizations per caller, color coded by calltype. We divide the entire data into training, validation, and test sets, named Train, Val, and Test respectively, following a 70:20:10 split. This distribution allows us to train models on a sufficiently large dataset while ensuring that we have sufficient data for model evaluation and validation. Train is used to train the models, Val to tune hyperparameters, and Test to evaluate the final performance of the trained models on unseen data.

![Figure 1: Vocalization per callers grouped by call-type.](image)

**2.2. Caller-Groups**

For our study, neural embeddings are extracted from the pre-trained SSL models by giving the Marmoset vocalizations as input for the purpose of caller detection. The log distribution of vocalization lengths in this dataset, depicted in Figure 2, exhibits a bimodal structure consistent with prior findings [21, 22]. However, the same figure also illustrates that the vocalization segments in this dataset are predominantly short, with a median segment length of around 125 ms. Considering the lack of prior knowledge for this task, we took inspiration from i-vector and x-vector based speaker verification systems, where utterance lengths considerably longer than a short-term window size are modeled to achieve high performance [23, 24]. More precisely, in order to effectively model each caller while accounting for the low vocalization segment length as well as to explore the acoustic variations within each caller, we first split all the vocalization embeddings by caller. Then, in order to maintain the chosen 70:20:10 split ratio of our data sets, we divide the embeddings of each caller sequentially into a fixed number of groups, hereafter referred to as ‘caller-groups’. We set the number of said groups to 100 for Train, and proportionally scale for Val and Test. This results in a total of 1000, 280, and 140 groups across all callers for Train, Val, and Test sets, respectively.

2.3. Embedding Spaces

We carry out caller discrimination analysis and caller detection studies by computing the first and second order statistics of the SSL embeddings in the caller-groups. For this purpose, we select eleven pre-trained SSL models from the SUBERB leaderboard [13] based on the different pretext tasks seen in Section 1, and use the S3PRL toolkit [13] to extract the embeddings. Table 1 lists the chosen models, along with their number of parameters $P$ in millions, and the dimension $D$ of the last layer embedding. All the models have been pre-trained on the LibriSpeech (LS) corpus, except Modified-CPC which is pre-trained on the Libri-Light (LL) corpus.

![Table 1: Selected pre-trained SSL models on human speech. $P$ indicates the number of parameters in millions, and $D$ corresponds to the dimension of the last layer embedding.](image)

3. Caller Discrimination Analysis

This section presents a discrimination analysis of SSL embedding spaces for the purpose of marmoset caller distinction. For this study we only use the Train portion of the data.

In order to conduct this analysis on our data, we first model the embedding spaces of each caller-group with a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with mean $\mu$ and diagonal covariance matrix $\Sigma$, resulting in a total of 100 multivariate Gaussians for each caller.

Subsequently, we compute the inter-caller and intra-caller distances by comparing the multivariate Gaussian distributions. Specifically, for inter-caller distances, we calculate a total of 100 · 100 pairwise distances for each pair of callers. For inter-caller distances, we compute a total of $\binom{100}{2}$ distances. To compute the distance between the Gaussians of a pair of caller-groups, we use two measures, namely the Kullback-Leibler (KL) divergence and Bhattacharyya distance, both of which produce distances in the range of $[0, +\infty)$. It is to be noted that the latter provides a symmetric measure while the former does not.

Equations 1 and 2 respectively provide the formulas for calculating the KL divergence $D_{KL}$ and Bhattacharyya distances $D_{BC}$ between two multivariate Gaussian distributions $\mathcal{N}_1$ and $\mathcal{N}_2$:

$$D_{KL}(\mathcal{N}_1, \mathcal{N}_2) = \int \mathcal{N}_1(x) \log \frac{\mathcal{N}_1(x)}{\mathcal{N}_2(x)} \, dx$$

$$D_{BC}(\mathcal{N}_1, \mathcal{N}_2) = \sqrt{\sum_{i=1}^{D} \sum_{j=1}^{D} \frac{2(\mu_i - \mu_j)^2}{\mu_i + \mu_j}}$$

where $\mu_i$ and $\mu_j$ are the means of the two distributions, and $\Sigma_i$ and $\Sigma_j$ are their respective covariance matrices.
Within the KL divergence, the mean vector $\mu$, covariance matrix $\Sigma$, determinant $|\Sigma|$, and dimensionality $d$ are utilized. Meanwhile, the Bhattacharyya distance uses the arithmetic mean of the covariance matrices $\Sigma_f$ and $\Sigma_g$ as $\Sigma$.

\[
D_{KL}(f||g) = \frac{1}{2} \log \left( \frac{|\Sigma_f|}{|\Sigma_g|} \right) + \text{Tr} \left( \Sigma_g^{-1} \Sigma_f \right) + (\mu_f - \mu_g)^T \Sigma_g^{-1} (\mu_f - \mu_g) - d \\
\]

\[
D_{BC}(f||g) = \frac{1}{8} (\mu_f - \mu_g)^T \Sigma_g^{-1} (\mu_f - \mu_g) + \frac{1}{2} \log \left( \frac{|\Sigma_f|}{|\Sigma_g|} \right) \\
\]

Once we have computed the distribution of distances for all the SSL embedding spaces, we can visualize them through a heatmap. Figure 3 shows the distance matrix for WavLM’s embedding space, where the diagonal entries represent the intracaller distances and the off-diagonal correspond to the inter-caller distances. In an ideal scenario, one would expect the intra-class distances between distributions to be smaller than the inter-class ones, which is not entirely the case in our results. Nevertheless, for callers with a larger amount of available data, we can observe good discrimination when compared to callers with a lower amount of data, as in the case of Caller 1 and Caller 3 vs. Caller 8. We observe that the distances exhibit similar patterns for all other SSL embeddings, which suggests these embeddings provide similar information for the caller discrimination task. Taken together, the analysis suggests that the SSL embeddings do carry information for distinguishing most callers to a certain extent. However, accomplishing this simple with a linear classifier may be a challenging task.

4. Caller Detection Study

4.1. Classifiers

Based on the insights of our caller discrimination analysis, we proceed to classify the statistics computed over the callergroups for the task of caller detection in a 5 fold cross-validation (CV) framework. We concatenate the mean and variance of the Gaussians into a single functional vector, and use them as our fixed-length representations for classification.

We use Random Forest (RF), Ada Boost (AB), Support Vector Machines (SVM), and Linear SVM (LSVM) algorithms to classify the computed functional vectors. The difference between Linear SVM and SVM with a linear kernel lies in the former’s utilization of a squared hinge-loss, while the latter employs a regular hinge-loss.

To determine the most robust classification technique, we employ the grid search methodology with F1-Macro score as the optimization criterion, integrated into the Scikit-learn toolkit. We tune the hyperparameters for each fold, across the train and validation sets over the search space given in Table 2.

4.2. Evaluation Metrics

To evaluate the effectiveness of our proposed approach for the given task, we present the area under the curve (AUC) scores, which provide a evaluation of the performance of all the classifiers in correctly classifying the positive instances against negative. For SVM it is computed pairwise using a ‘one-vs-one’ methodology, while for the other classifiers it is calculated in a binary ‘one-vs-rest’ framework, by averaging the AUC scores for each class against all others.

4.3. Results and Discussion

Table 3 summarizes the performance of the different classifiers on all the embedding spaces. The results show that SVM
significantly outperforms the other classifiers across all embedding spaces. The decision tree-based ensemble methods, AdaBoost and Random Forest, exhibit comparable performance for most models, and consistently outperform Linear SVM. This suggests that the relationship between the features in the embedding space and their labels is likely to be complex and non-linear, which can be modelled by ensemble methods to some degree, but not to the extent of non-linear SVMs.

Figure 4a) shows the classifier classification performance in distinguishing a positive class from the negative instances using SVM on a single Test fold. We can observe that all callers are systematically distinguished in this binary framework, including the classes with a low amount of data (CID 6–8).

Figure 4b) visualizes SVM’s average performance for each embedding space across the 5 folds, with the shaded areas representing ± 1 std. The results clearly demonstrate that the embedding spaces of all models are capable of successfully differentiating Marmoset callers, indicating that SSL models pre-trained on human speech data can generate salient representations capable of distinguishing animal vocalizations regardless of the pre-training criterion.

Figure 4c) illustrates the relationship between the number of parameters and classification performance for each embedding space. The plot is divided into four quadrants to highlight differences in performance. Interestingly, WavLM’s embedding space is found to be more separable than the other masked prediction models, indicating that its masked speech denoising task may be more effective in capturing animal caller identification information than Hubert’s masked speech modeling. Surprisingly, both auto-regressive reconstruction based models perform exceptionally well with significantly fewer parameters. These findings suggest that while all pre-training criteria can yield competitive performance, some may be more efficient than others, allowing models with simpler architectures and fewer parameters, such as APC and AQ-APC, to perform comparably to larger models like WavLM. Finally, we observe that Data2Vec is not as successful as the other masked prediction based models, despite the same number of pre-training hours, corpus and comparable number of parameters. While it has shown to outperform the other masked prediction models in human speech, it seems to clearly learn weaker representations for the task of domain adaptation.

5. Conclusions

This paper investigated the applicability of self-supervised representations, pre-trained on human speech through different approaches, to analyze vocalizations in the bio-acoustics domain. To that end, we conducted and validated two lines of investigation on Marmoset calls in a caller detection framework.

We first conducted a caller discrimination analysis study on the training data to examine the linear separability of eleven pre-trained embedding spaces by splitting the training data into caller-groups, and then calculating the intra-group and inter-group distances through a multivariate Gaussian distribution framework. The results showed that all spaces exhibited similar distance patterns, and that distinguishing marmoset callers is possible with a linear classifier but only to a certain extent.

For our second investigation, we conducted a caller detection study to analyze whether the embedding spaces of said caller-groups can be systematically distinguished by class. We trained four classifiers to predict the classes of the caller-groups in 5 fold cross-validation framework. The results show that we can effectively distinguish all Marmoset callers, including those with low data, in a binary classification framework. The results also show that non-linear SVMs are able to most accurately model the non-linear relationship between the features of the embedding space. Finally, we observe that although all embedding spaces seem effective at the caller detection task, some learning objectives may be more efficient than others.

In summary, our research demonstrates that self-supervised representations pre-trained on human speech can effectively classify vocalizations in the bio-acoustics domain for tasks such as Marmoset caller detection, even without fine-tuning. These findings can greatly benefit bio-acoustics researchers looking to distinguish individual identities within a specific species in their acoustic data. Additionally, we anticipate that further fine-tuning of these models on relevant bio-acoustics downstream tasks can improve performance. Therefore, we plan to investigate the impact of model size on performance after fine-tuning, and also explore adapting the embedding spaces for other tasks like call-type classification in our future work.

6. Acknowledgments

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


