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

Eklavya Sarkar[1,2], Mathew Magimai Doss[2]

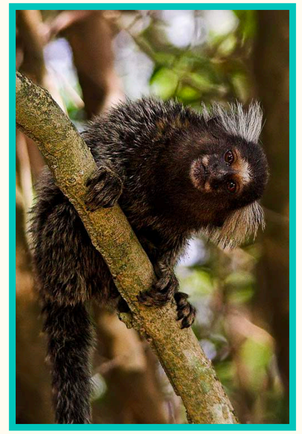
[1] Idiap Research Institute, Switzerland
[2] Ecole polytechnique fédérale de Lausanne, Switzerland

ISCA Interspeech 2023

June 2023

1. Introduction

Bio-Acoustics



Topic:

- Study of animal vocalizations.
- Research has progressed in recent due to approaches inherited from ML/DL.

Issues:

- Labeled data scarcity.
- Lack of domain knowledge.
- Understudied topic.

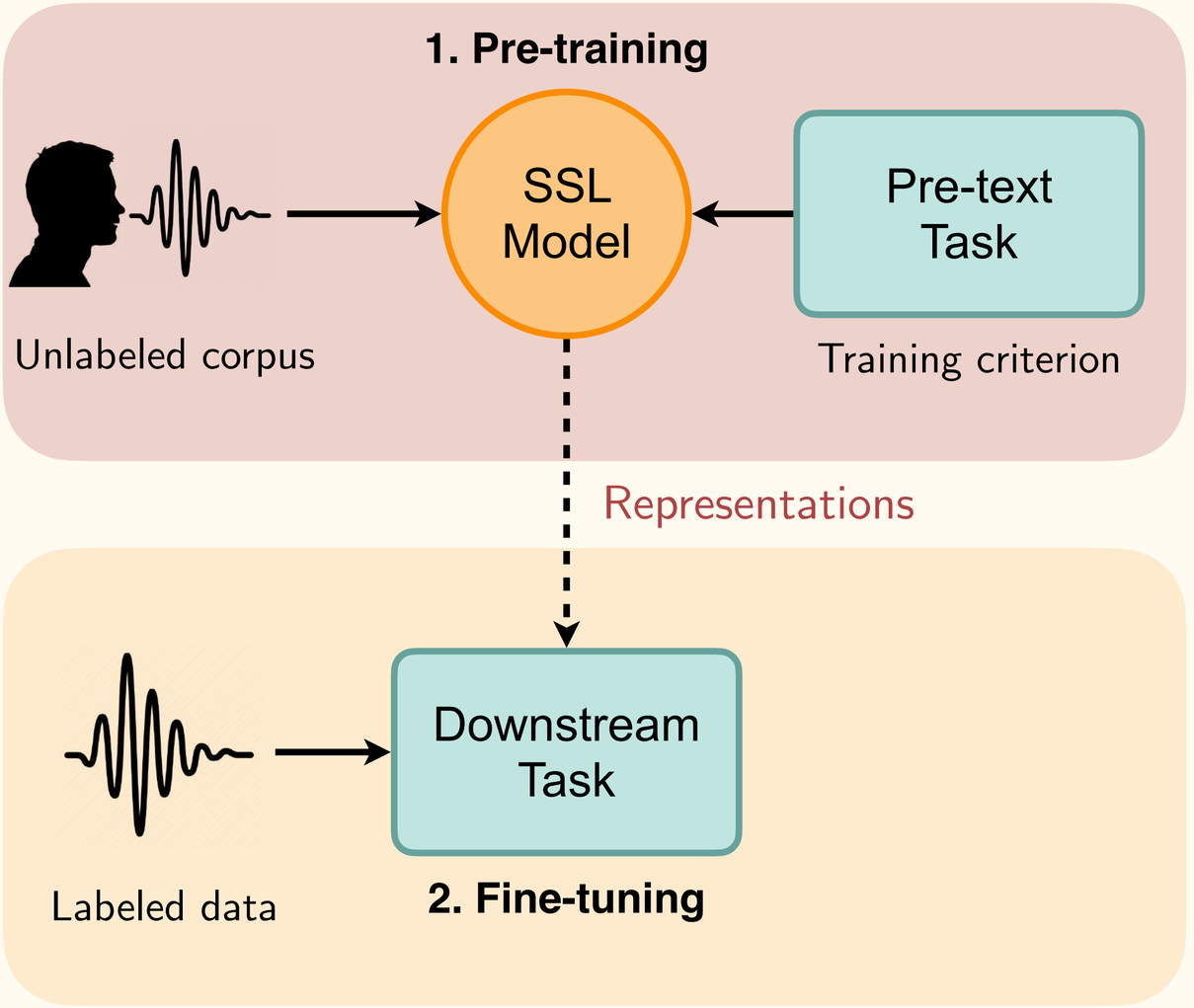
- Self-supervised learning has emerged as a way of leveraging unlabeled data.

Self-Supervised Learning Framework

Pre-training:

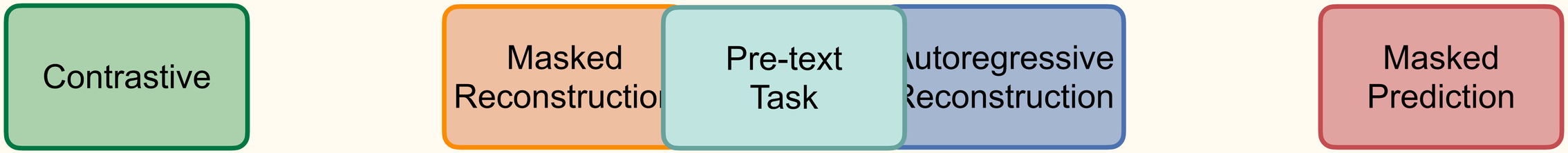
- Create surrogate labels
  from unlabeled data
  based on the pre-text task.
- Optimize its learning objective.
- Goal: learn useful representations.

- Network infers intrinsic structure.
- No knowledge is explicitly provided
  (eg speech production mechanism).
- Utility not limited to modeling speech.



Self-Supervised Pre-Training Objectives

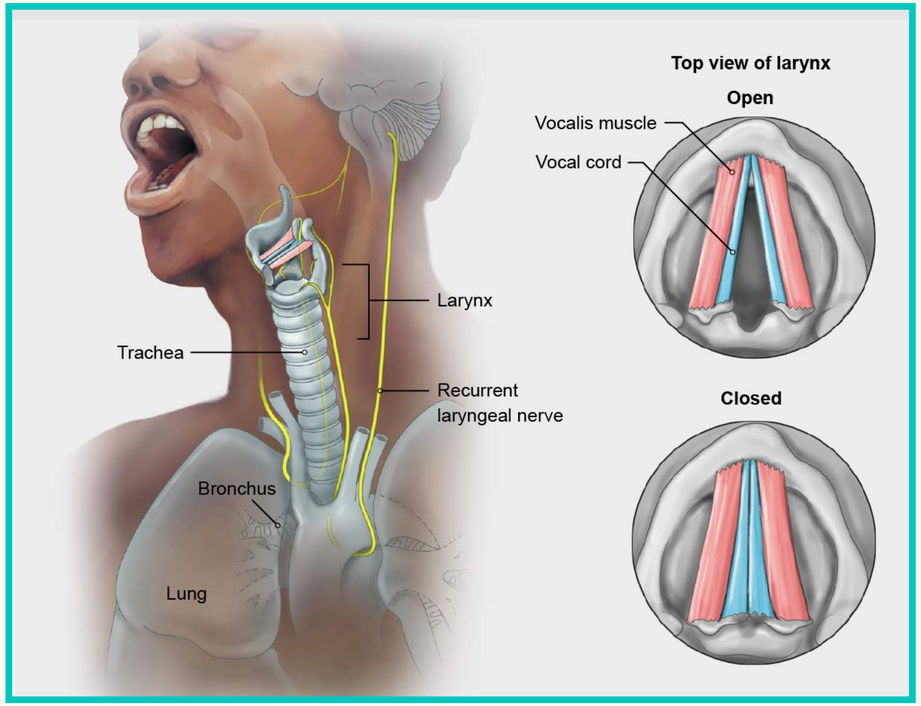
- The information encoded in the SSL representations can vary depending on *learning objective* (among other elements).
- These can be roughly categorized into the four approaches given below.
- This framework has yielded SOTA results on the SUPERB (speech) benchmark.



Balestriero et al. *A Cookbook of Self-Supervised Learning*. (2023). Meta AI.

Humans

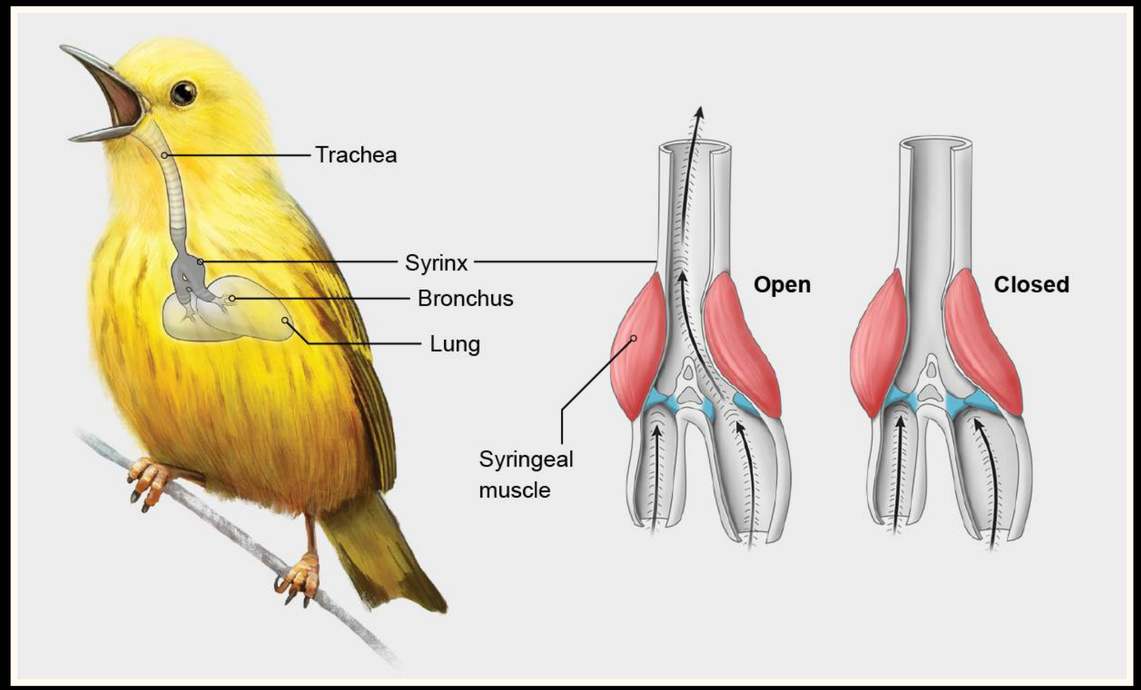
Larynx



Humans and animals share a commonality: they both have voice production mechanism.

Birds

Syrinx



Michael B. Habib, 2020. *Fossils Reveal When Animals Started Making Noise.* Scientific American 326, 1, 42-47, Jan 22.

Motivation

- Given this understanding, our objective is to:
- ▸ Investigate the cross-transferability of pre-trained SSL representations learned from human speech for analyzing animal vocalizations.
- Previous works has explored birdsong detection[1] and bio-acoustic event detection[2] using contrastive learning.
- However, the generalization of SSL representations to animal vocalizations has largely remained unexplored.

[1]Saeed et al., *Contrastive learning of general-purpose audio representations*, ICASSP, 2021.
[2]Bermant et al., *Bioacoustic Event Detection with Self-Supervised Contrastive Learning*, BioRxiv, 2022.

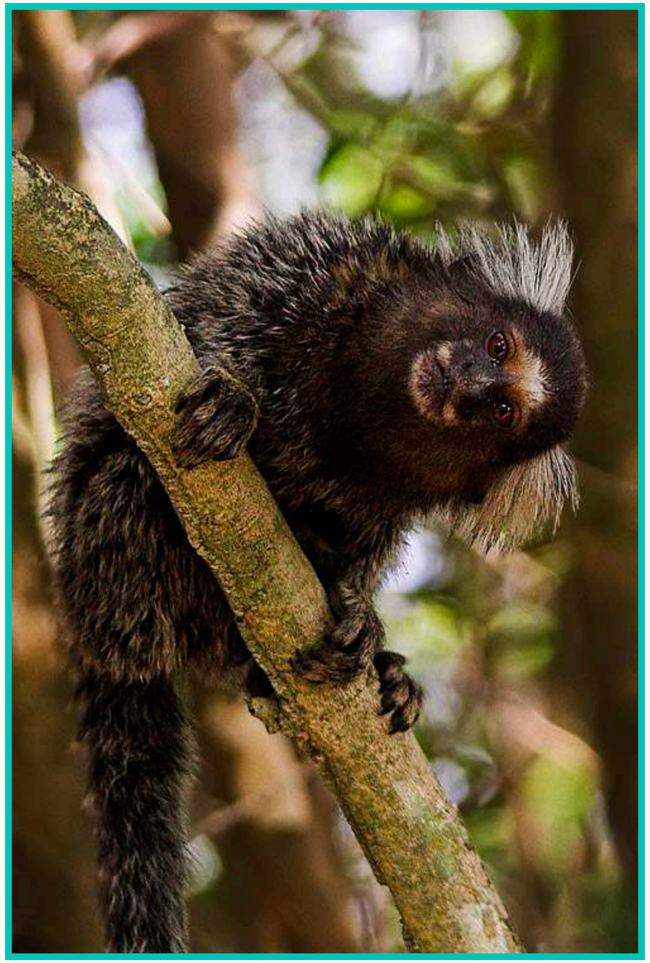
Research Questions

We design studies for following research questions:

1. How discriminative are the embedding spaces of SSL models pre-trained on human speech?

2. Can we systematically detect individual animal callers using said embedding spaces ?

For this study we focus on marmosets.

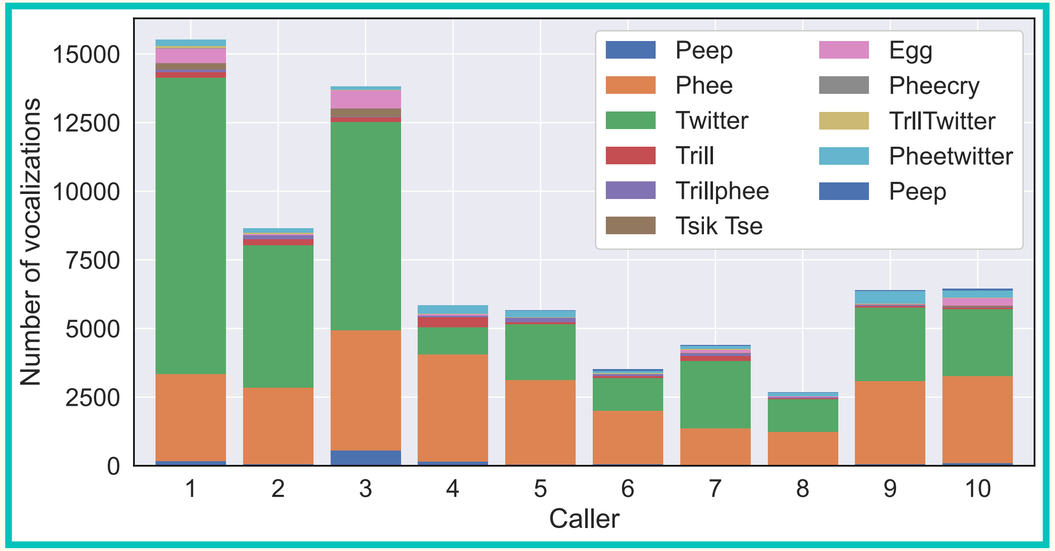


Marmoset

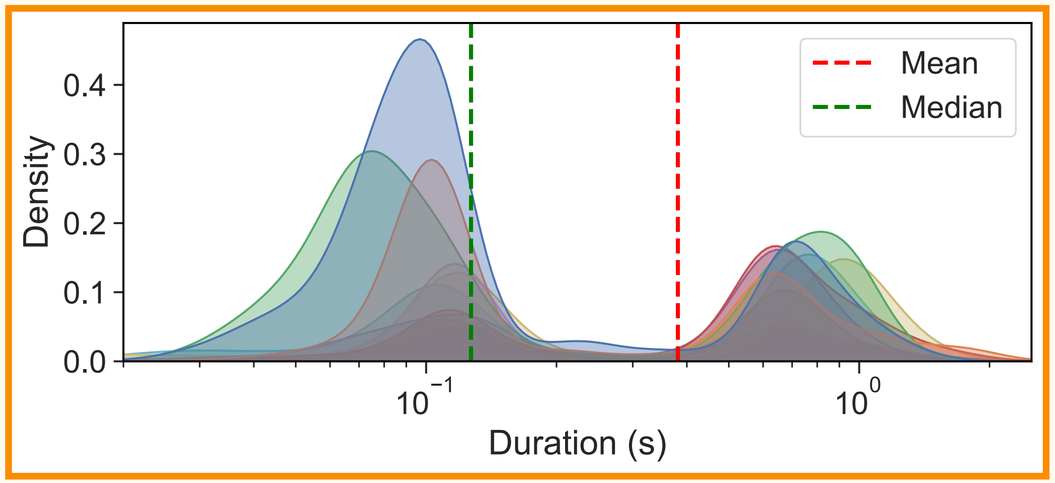
2. Study Design

Dataset

- We used a marmoset dataset collected and labeled by a previous paper[1].
- The dataset contains audio vocalization segments of eleven different marmoset call-types (*Twitters*, *Phees*, *Trills*, etc.) as well as caller identities (1-12).
- 73k vocalization segments (7.7 hours).
- Imbalanced distribution of vocalizations per caller, color coded by call-type.



Vocalization per callers grouped by call-type.



Log distribution of vocalization lengths for callers 1-10.

Zhang YJ, et al. Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks. JASA. 2018.

Embedding Spaces

- 11 selected SSL models.
- Pre-trained on human speech.
- 4 different pre-text tasks.

Model	Corpus	P	D	Pretext Objective
APC	LS 360	4.11	512	Autoreg. Recon.
VQ-APC	LS 360	4.63	512	Autoreg. Recon.
NPC	LS 360	19.38	512	Masked Recon.
Mockingjay	LS 100	21.33	768	Masked Recon.
TERA	LS 100	21.33	768	Masked Recon.
Mod-CPC	LL 60k	1.84	256	Contrastive
Wav2Vec2	LS 960	95.04	768	Contrastive
Hubert	LS 960	94.68	768	Masked Pred.
DistilHubert	LS 960	27.03	768	Masked Pred.
WavLM	LS 960	94.38	768	Masked Pred.
Data2Vec	LS 960	93.16	768	Masked Pred.

LS is LibriSpeech, and LL is Libri-Light.

P indicates the number of parameters in millions.

D corresponds to the last layer embedding's dimension.

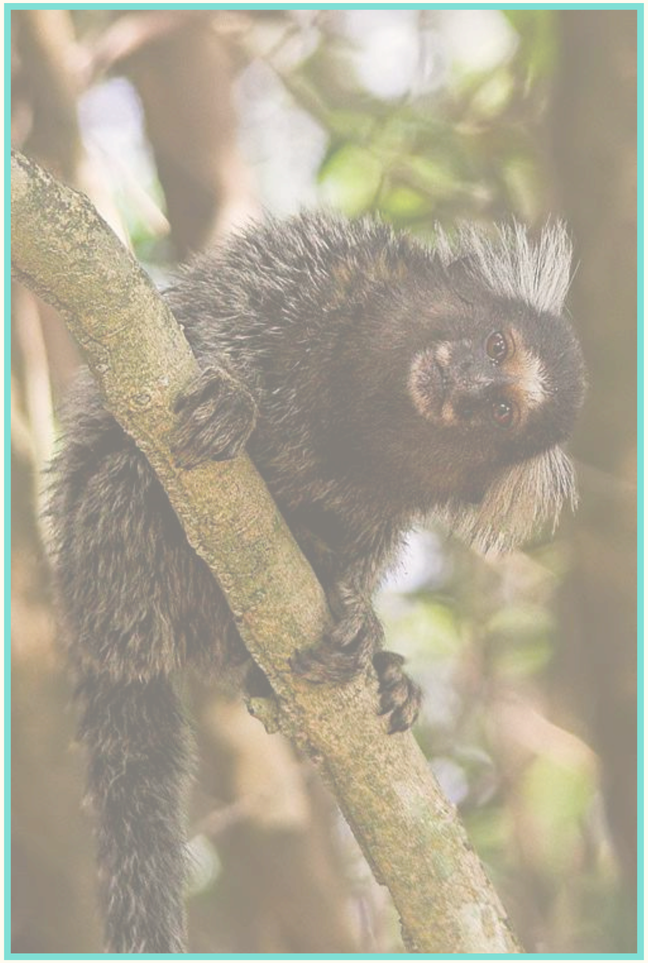
3. Caller Discrimination Analysis

Research Questions

We design a study with the following research questions:

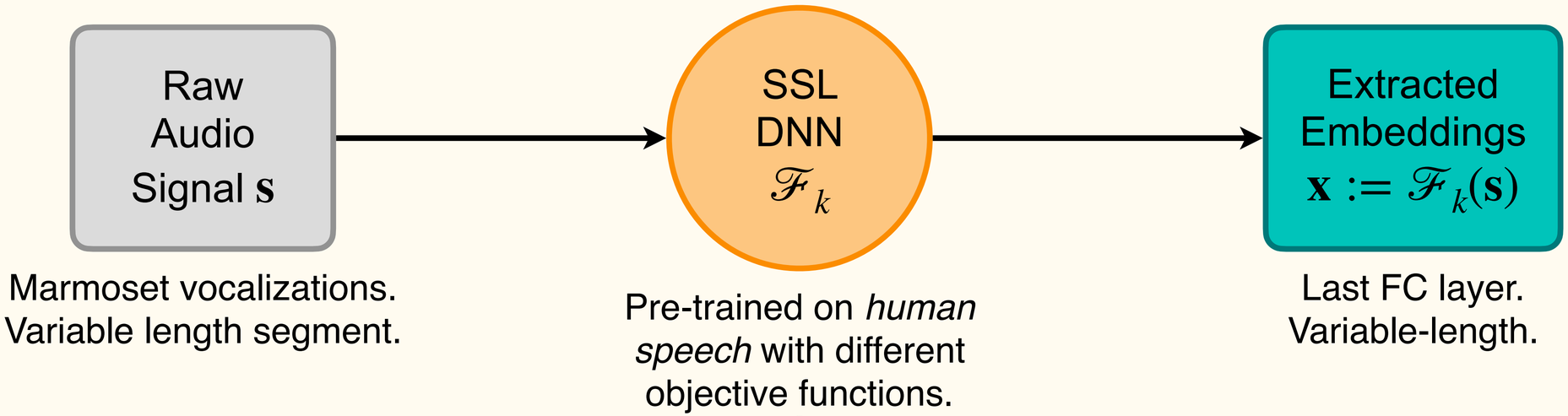
1. How discriminative are the embedding spaces of SSL models pre-trained on human speech?

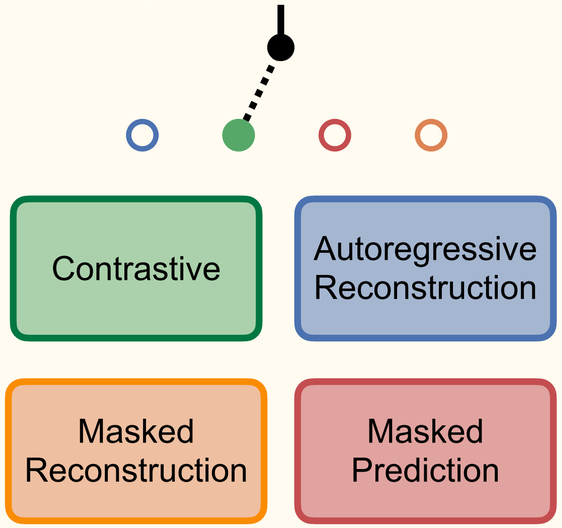
2. Can we systematically detect individual Marmoset callers using said embedding spaces ?

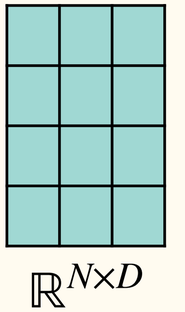


Marmoset

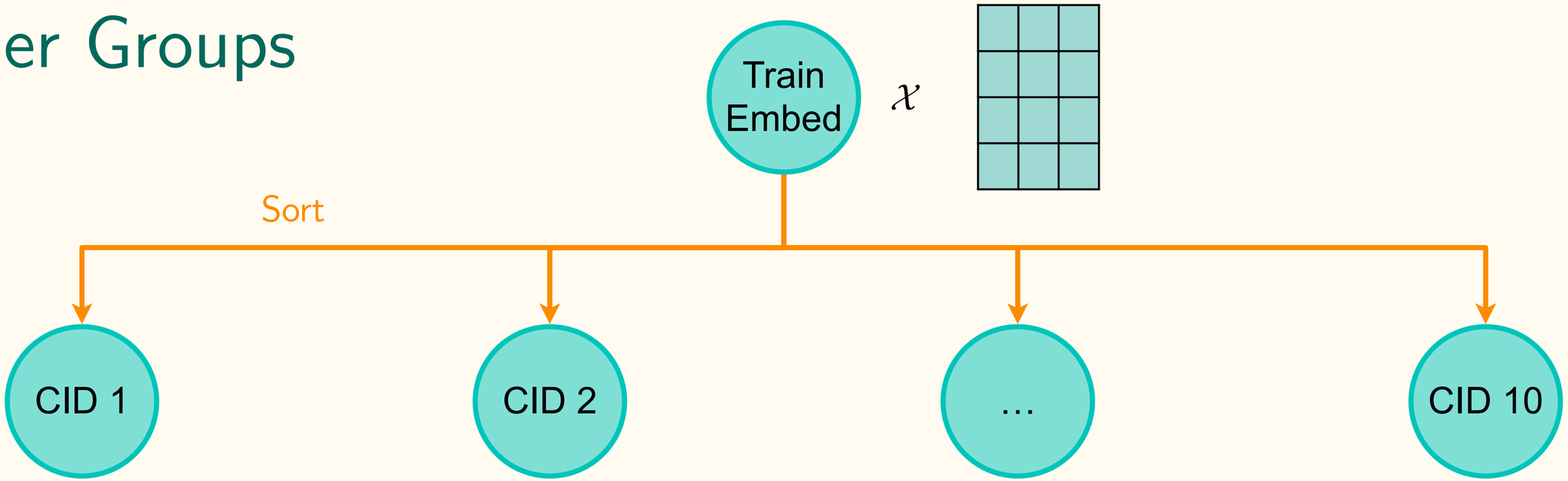
Pipeline


Raw
Audio
Signal $\mathbf{s}$
SSL
DNN
$\mathscr{F}_k$
Extracted
Embeddings
$\mathbf{x} := \mathscr{F}_k(\mathbf{s})$
Marmoset vocalizations.
Variable length segment.
Pre-trained on *human speech* with different objective functions.
Last FC layer.
Variable-length.

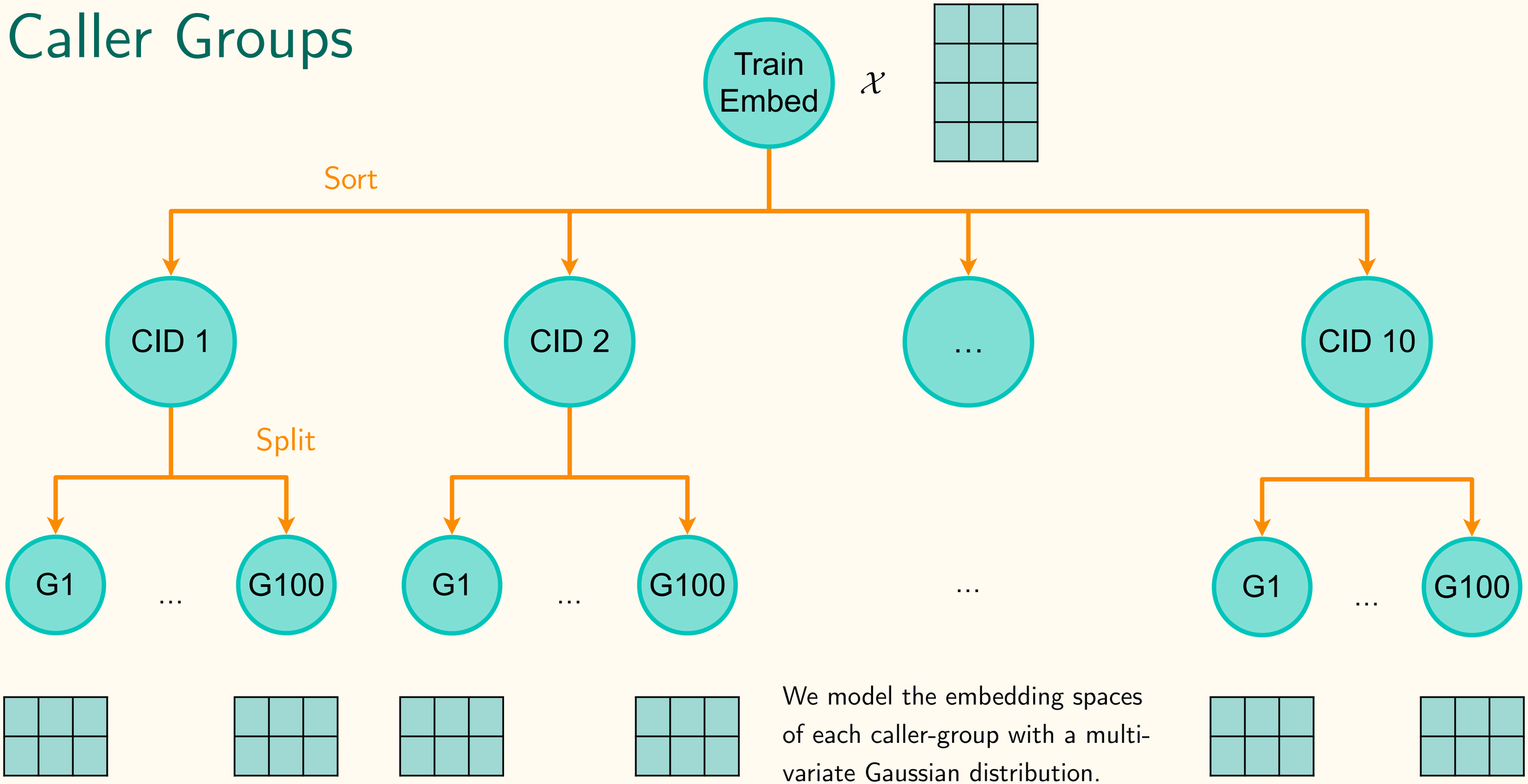

Contrastive
Autoregressive Reconstruction
Masked Reconstruction
Masked Prediction


$\mathbb{R}^{N \times D}$

Caller Groups

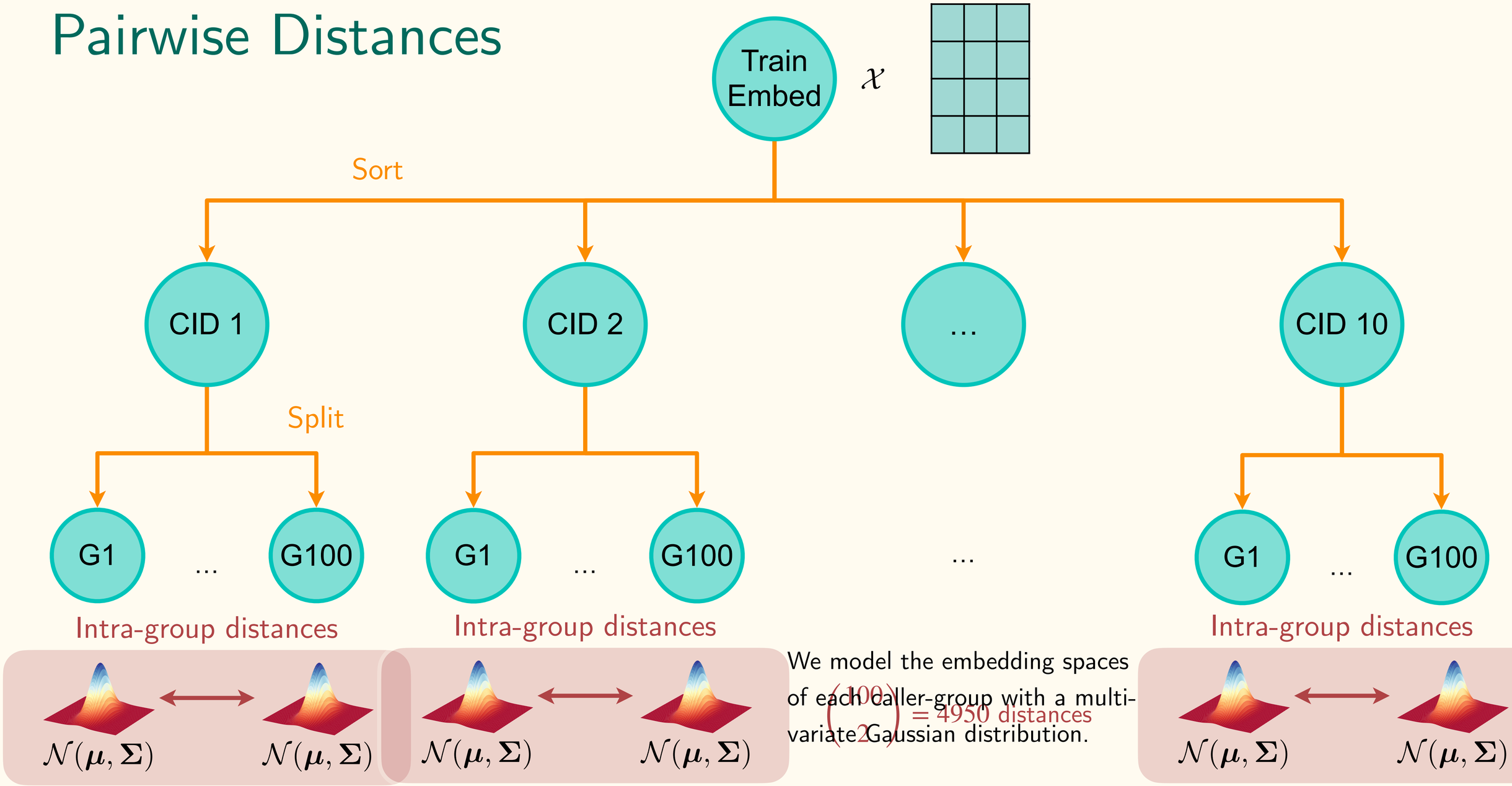


Caller Groups



We model the embedding spaces of each caller-group with a multi-variate Gaussian distribution.

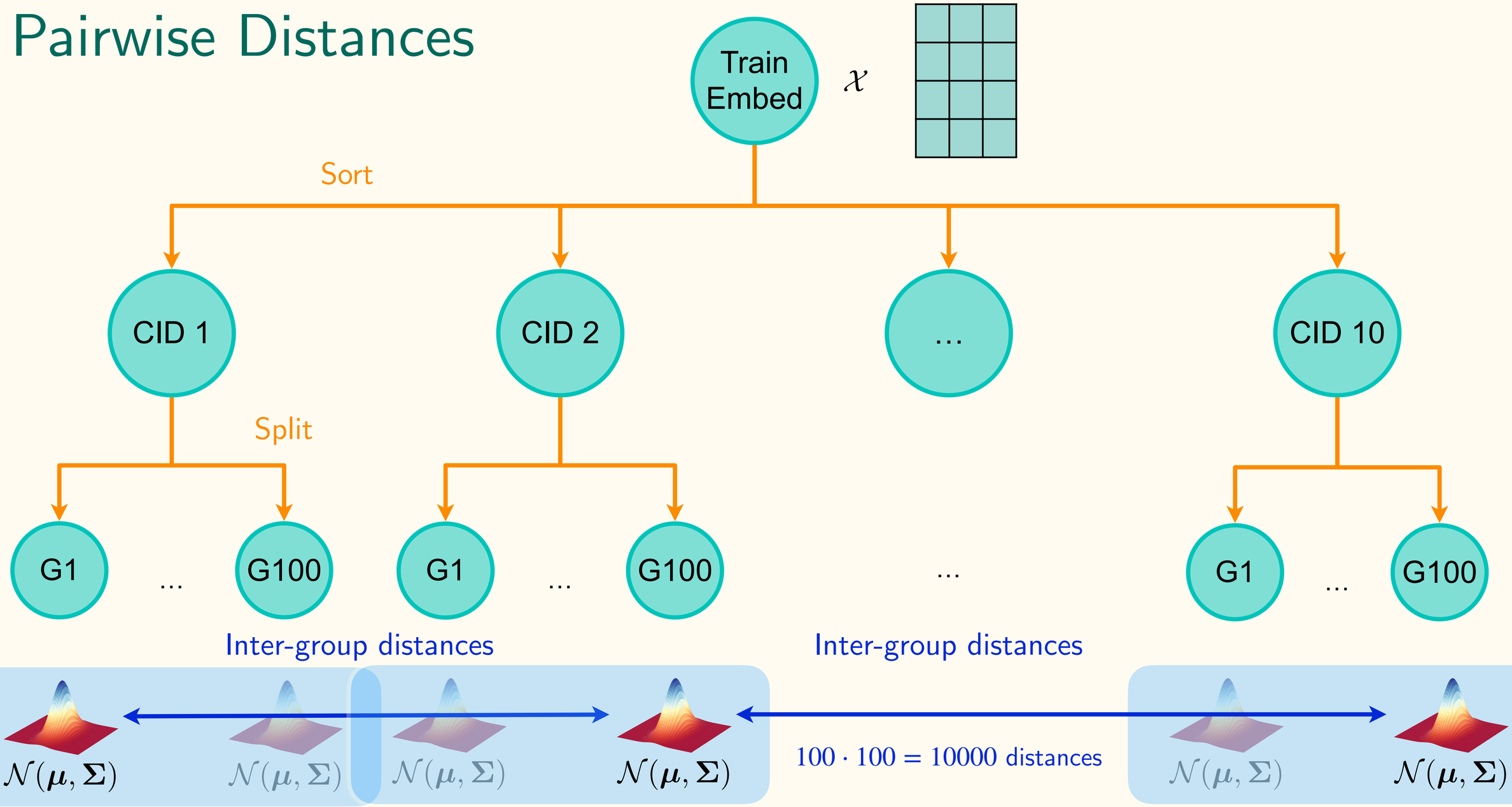
Pairwise Distances



We model the embedding spaces of each caller-group with a multivariate Gaussian distribution.

$\binom{100}{2} = 4950$ distances

Pairwise Distances



Pairwise Distances

Distance measures between two multivariate Gaussian distributions $\mathcal{N}_f$ and $\mathcal{N}_g$.

- KL-Divergence:

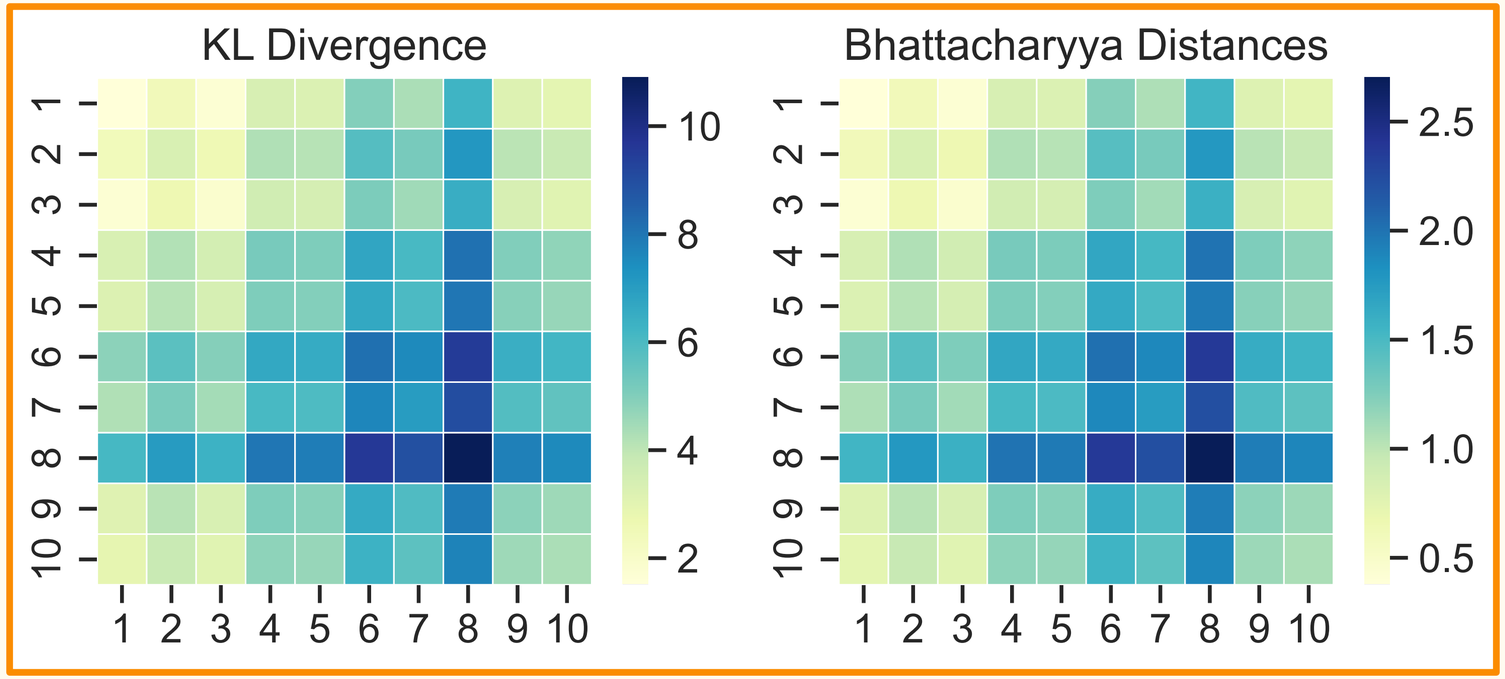
$$D_{\mathrm{KL}}(f||g) = \frac{1}{2}\Big(\log\frac{|\boldsymbol{\Sigma_g}|}{|\boldsymbol{\Sigma_f}|} + \mathrm{Tr}(\boldsymbol{\Sigma_g}^{-1}\boldsymbol{\Sigma_f}) + (\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T\boldsymbol{\Sigma_g}^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g) - d\Big)$$

- Bhattacharyya Distance:

$$D_{BC}(f||g) = \frac{1}{8}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_g) + \frac{1}{2}\log(\frac{|\Sigma|}{\sqrt{|\Sigma_f||\Sigma_g|}})$$

Results

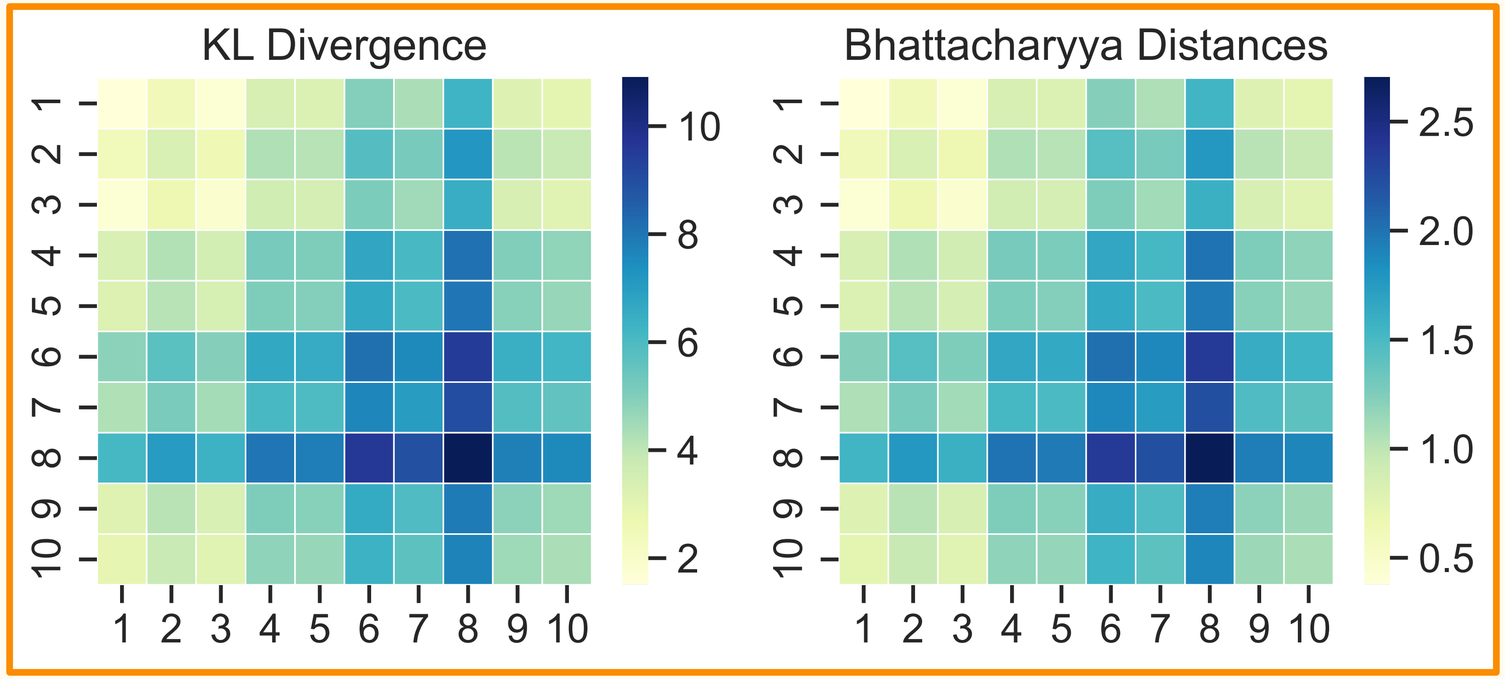
- We can visualize the computed distribution of distances through a heatmap.



Distance matrix of callers in WavLM's embedding space.
Darker regions indicate higher dissimilarity.

Results

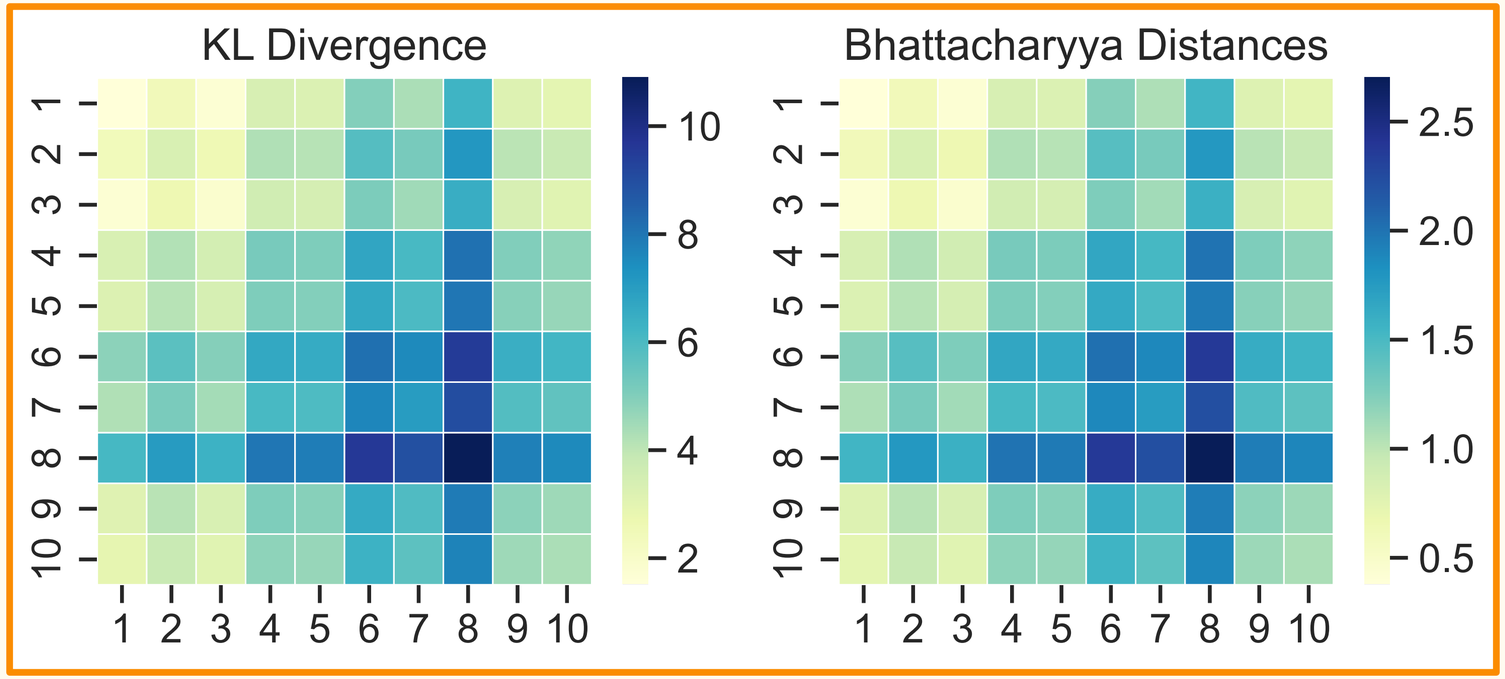
- Diagonal: intra-caller distances
- Off-diagonal: inter-caller distances.



Distance matrix of callers in WavLM's embedding space.
Darker regions indicate higher dissimilarity.

Results

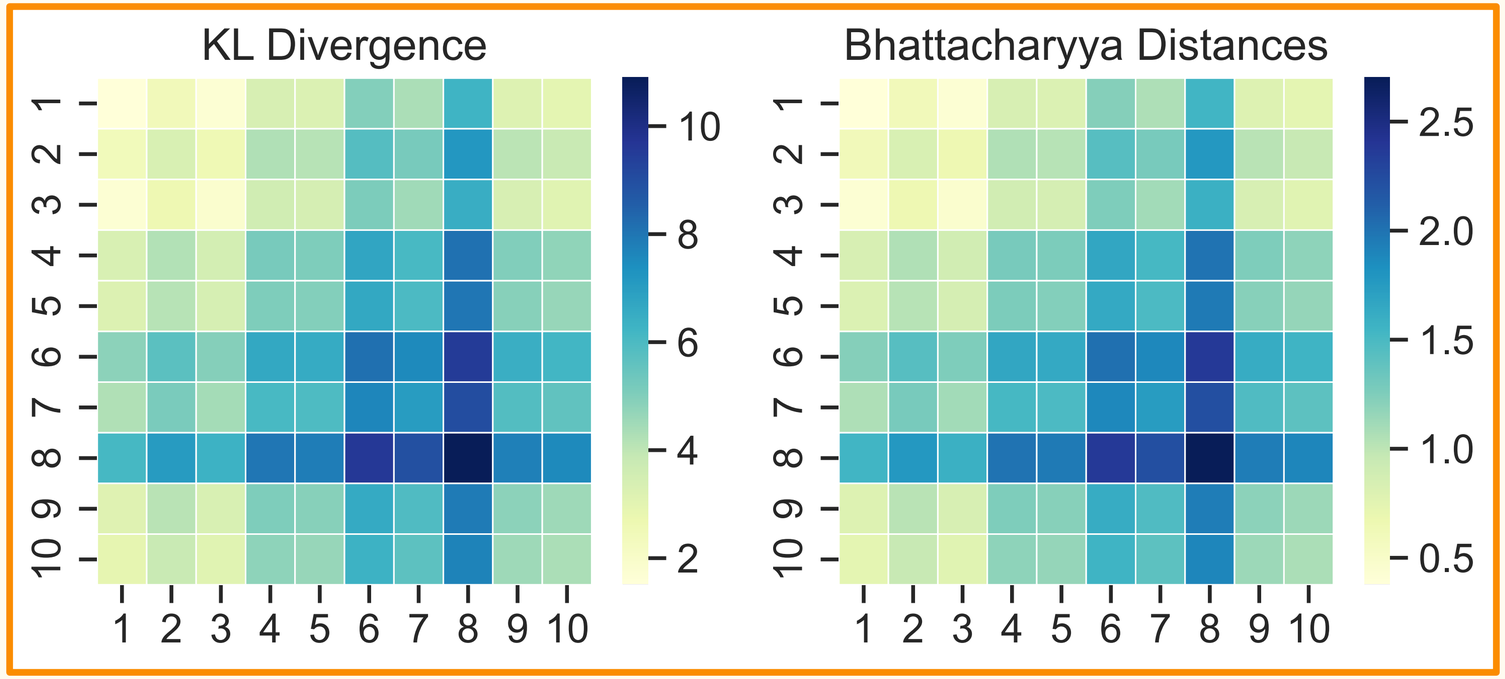
- Ideal scenario: the intra-class distances to be smaller than the inter-class ones
- Not entirely the case in our results.



Distance matrix of callers in WavLM's embedding space.
Darker regions indicate higher dissimilarity.

Results

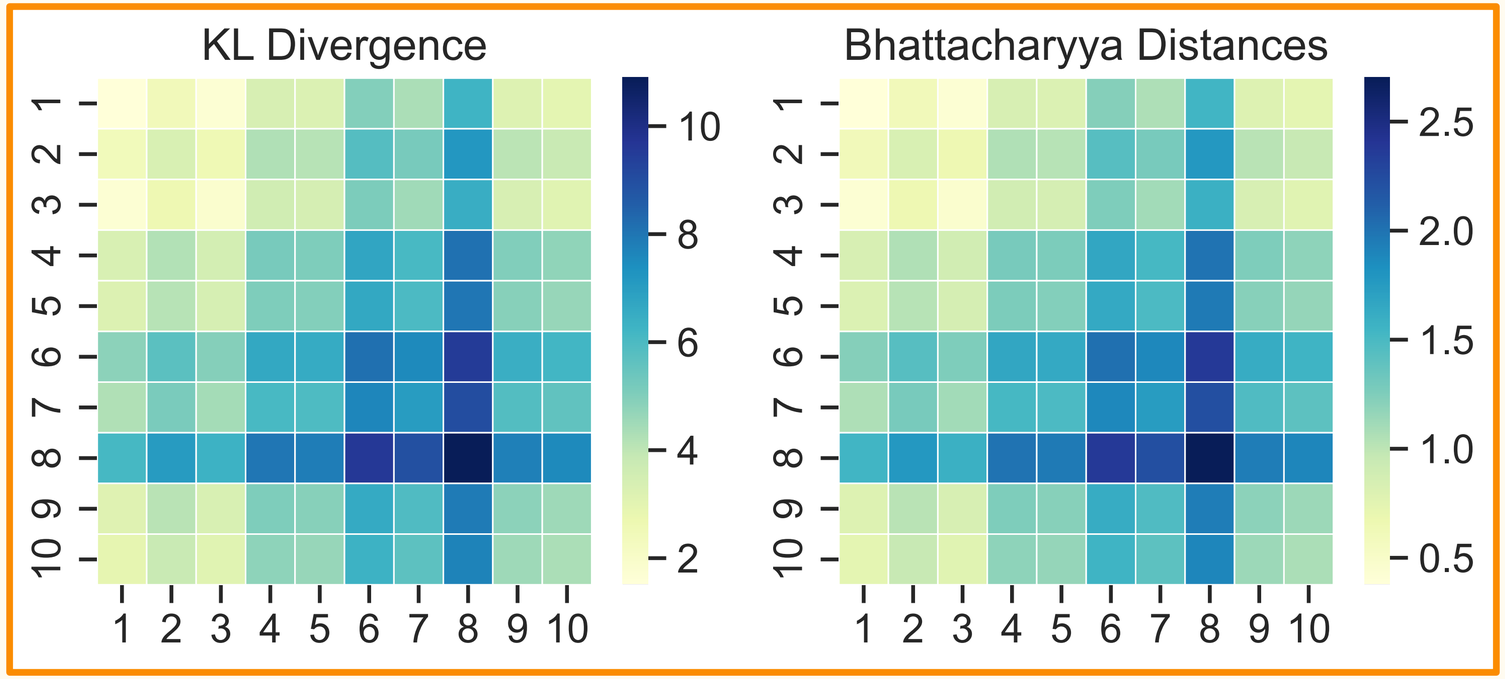
- Nevertheless, for callers with a larger amount of available data (Caller 1-3), we observe good discrimination when compared to callers with a lower amount of data (Caller 8).



Distance matrix of callers in WavLM's embedding space.
Darker regions indicate higher dissimilarity.

Results

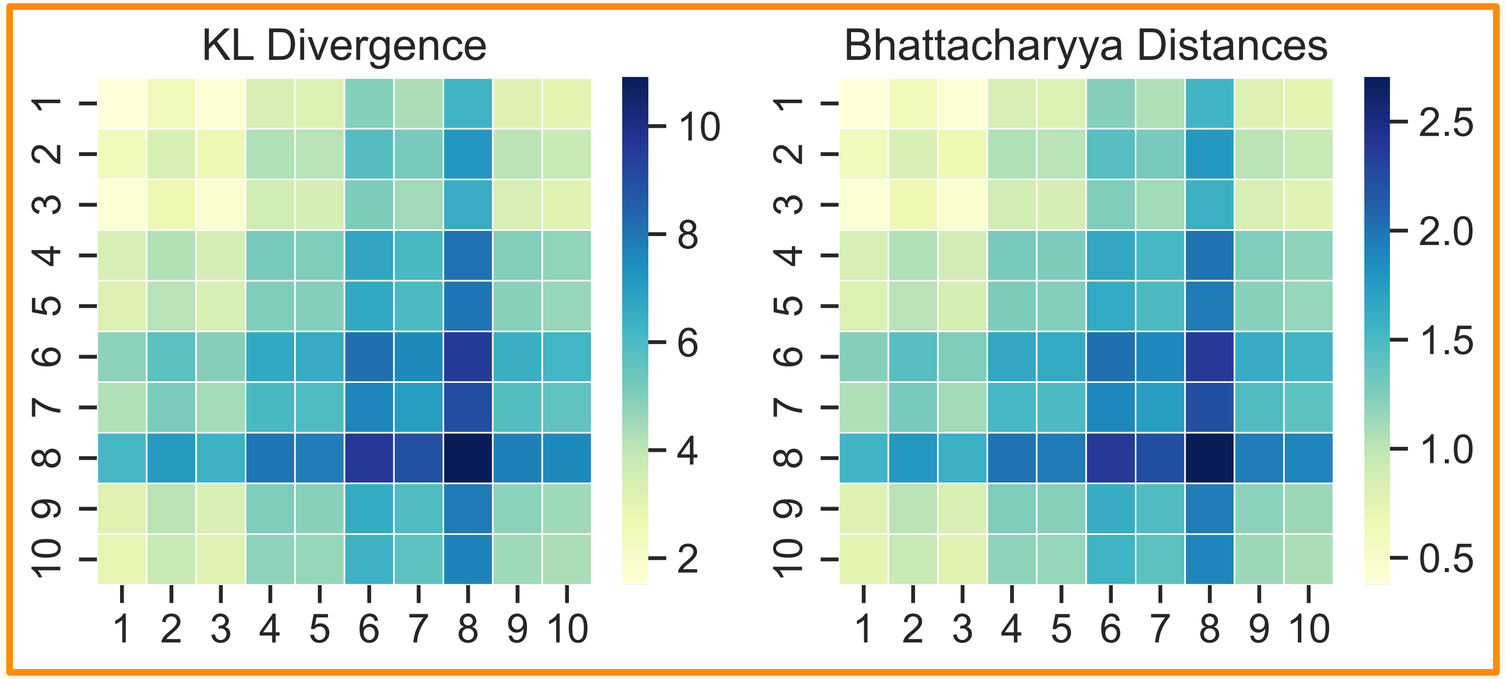
- Analysis suggests that the SSL embeddings do carry information for distinguishing marmoset callers to a certain extent.



Distance matrix of callers in WavLM's embedding space.
Darker regions indicate higher dissimilarity.

Results

- Accomplishing this with a simple linear classifier may still be a challenging task.



Distance matrix of callers in WavLM's embedding space.
Darker regions indicate higher dissimilarity.

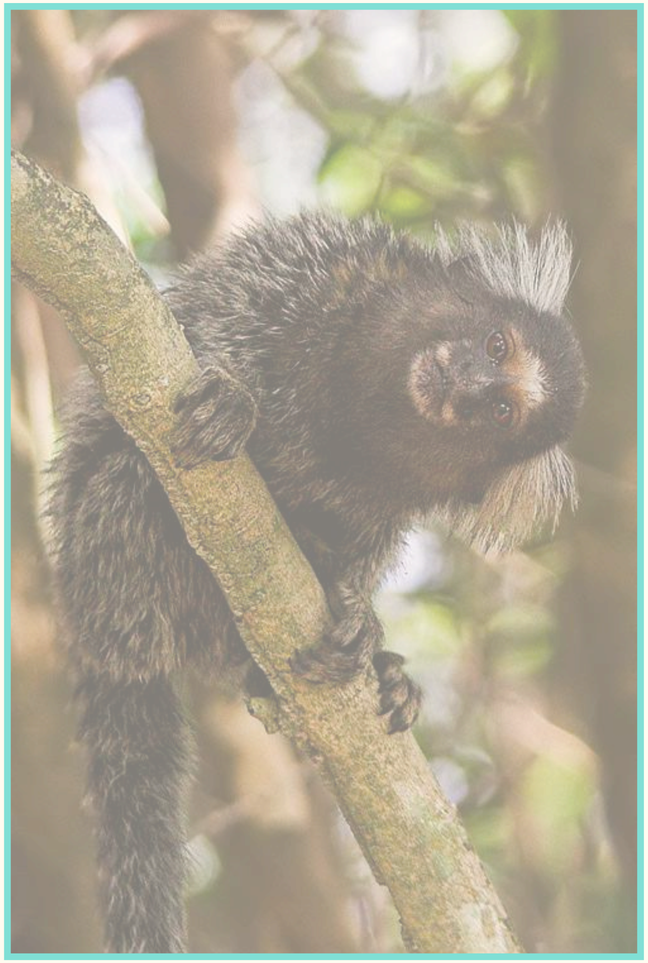
4. Caller Detection Study

Research Questions

We design a study with the following research questions:

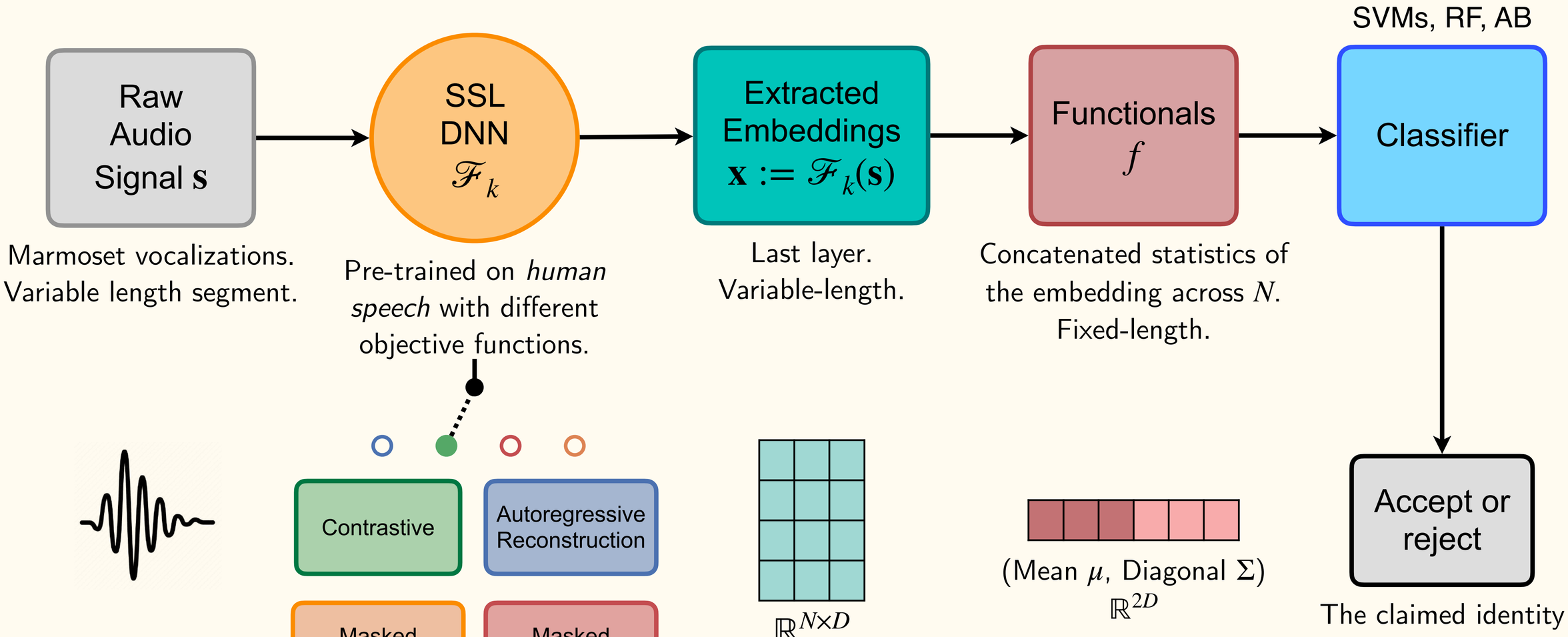
1. How discriminative are the embedding spaces of SSL models pre-trained on human speech?

2. Can we systematically detect individual Marmoset callers using said embedding spaces ?



Marmoset

Caller Detection Pipeline



Caller Detection Study

Classifiers:

- Random Forest (RF).
- Support Vector Machine (SVM).
- AdaBoost (AB).

Framework:

- 5-fold cross-validation.
- Hyper-parameter tuning on each fold.
- Grid Search.

- Task: Caller detection.
- Performance measure: AUC.

Classifier	**Hyperparameters**	**Search space**
RF	# Estimators	[50, 500, 1000, 2000]
	Max # Features	[‘auto’, ‘sqrt’, ‘log2’]
	Criterion	[‘gini’, ‘entropy’]
	Min samples leaf	[1, 2, 4]
AB	Learning rate	[0.1, 0.2, 0.5, 1]
	Algorithms	[SAMME, SAMME.R]
	Max # Estimators	[50, 500, 1000, 2000]
SVM	C	1e[-5, -4, -3, -2, -1, 0]
	Kernel	[RBF, Linear, Polynomial]
	Gamma	[‘scale’, ‘auto’]
LSVM	C	1e[-5, -4, -3, -2, -1, 0]
	Max # Iterations	10000
	Class weights	[‘balanced’, ‘None’]

Search space to find optimal hyper-parameters.

Results and Discussion

- SVM significantly outperforms the other classifiers across all embedding spaces.
- The decision tree-based ensemble methods exhibit comparable performance for most models, consistently outperform LSVM.

➡ Shows that the relationship between features in embedding space and their labels can be modeled by ensemble methods to some degree, but not to the extent of non-linear SVMs.

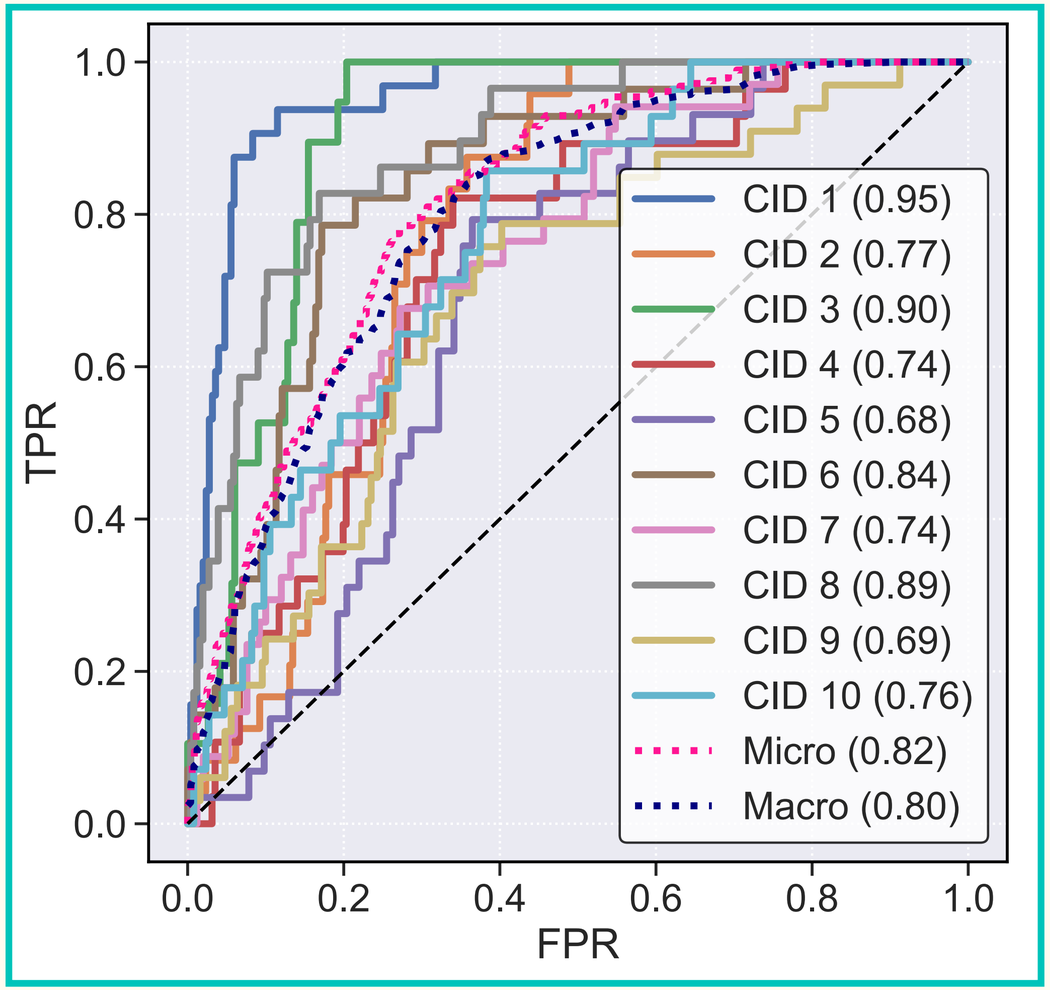
Model	**AB**	**LSVM**	**RF**	**SVM**
APC	71.44	65.18	70.89	79.16
VQ-APC	71.60	65.58	70.04	78.45
NPC	72.61	66.27	71.50	77.32
Mockingjay	72.39	64.43	71.75	78.44
TERA	70.34	64.57	68.43	74.03
Mod-CPC	72.62	64.05	69.81	75.96
Wav2Vec2	74.41	63.94	70.18	75.85
Hubert	71.71	64.14	70.17	75.64
DistilHubert	70.77	65.11	70.34	76.26
WavLM	73.97	65.32	70.74	78.60
Data2Vec	69.81	62.58	68.23	73.04
Average	71.97	64.66	70.19	**76.61**

Macro AUC scores [%] on *Test* with 5-fold CV.

Results and Discussion

- Per-caller detection performance in distinguishing a positive class from the negative instances using SVM on a single *Test* fold.

➡ All callers are systematically distinguished in this binary framework, including the classes with a low amount of data (CID 6–8).



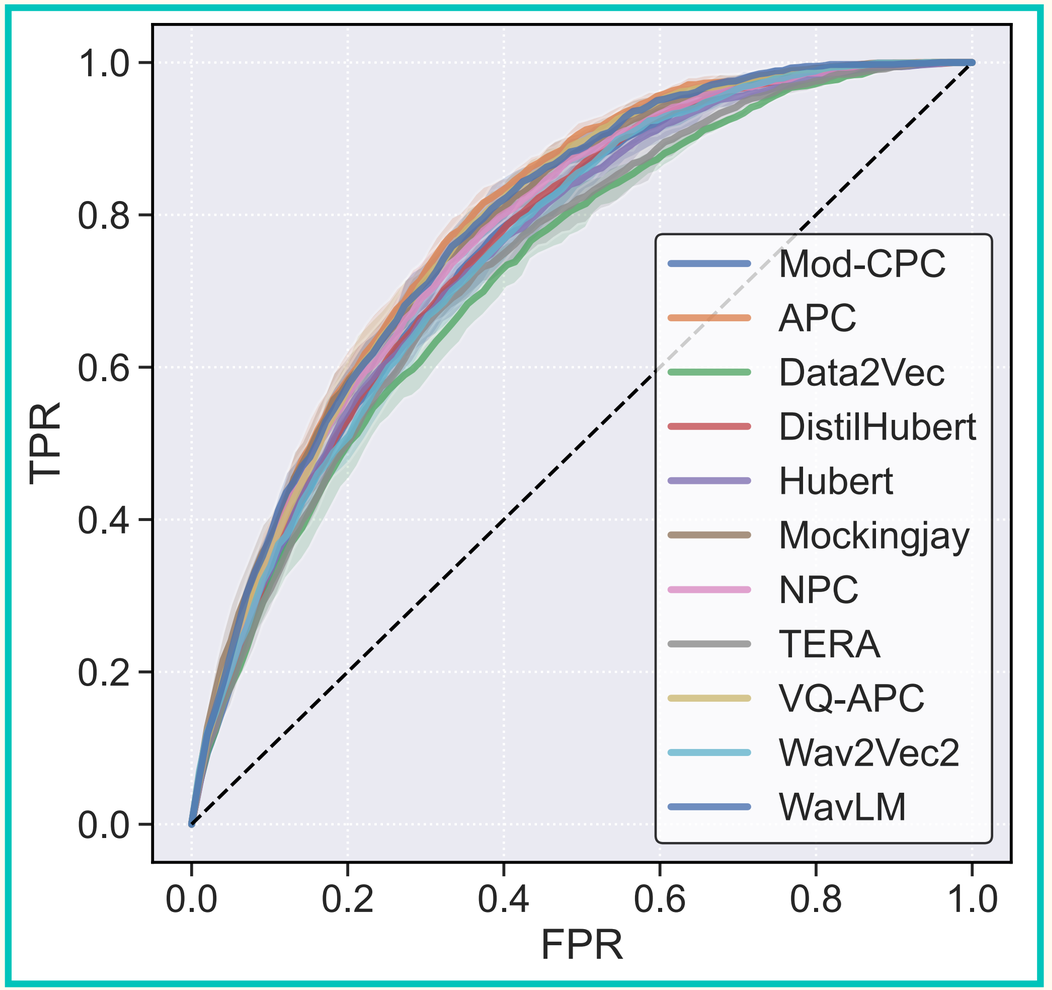
AUC-ROC curves per caller class (CID) for WavLM embeddings using RBF SVM on one fold of Test.

Results and Discussion

- SVM's average performance for each embedding space across the 5 folds.

➡ Embedding spaces of all models are capable of successfully differentiating Marmoset callers.

➡ Indicates that SSL models pre-trained on human speech data can generate salient representations capable of distinguishing animal vocalizations *regardless* of the pre-training criterion.



Macro average ROC curves of all models on *Test* using RBF SVM over all folds. Shaded areas represent ± 1 std over the 5-folds.

Results and Discussion

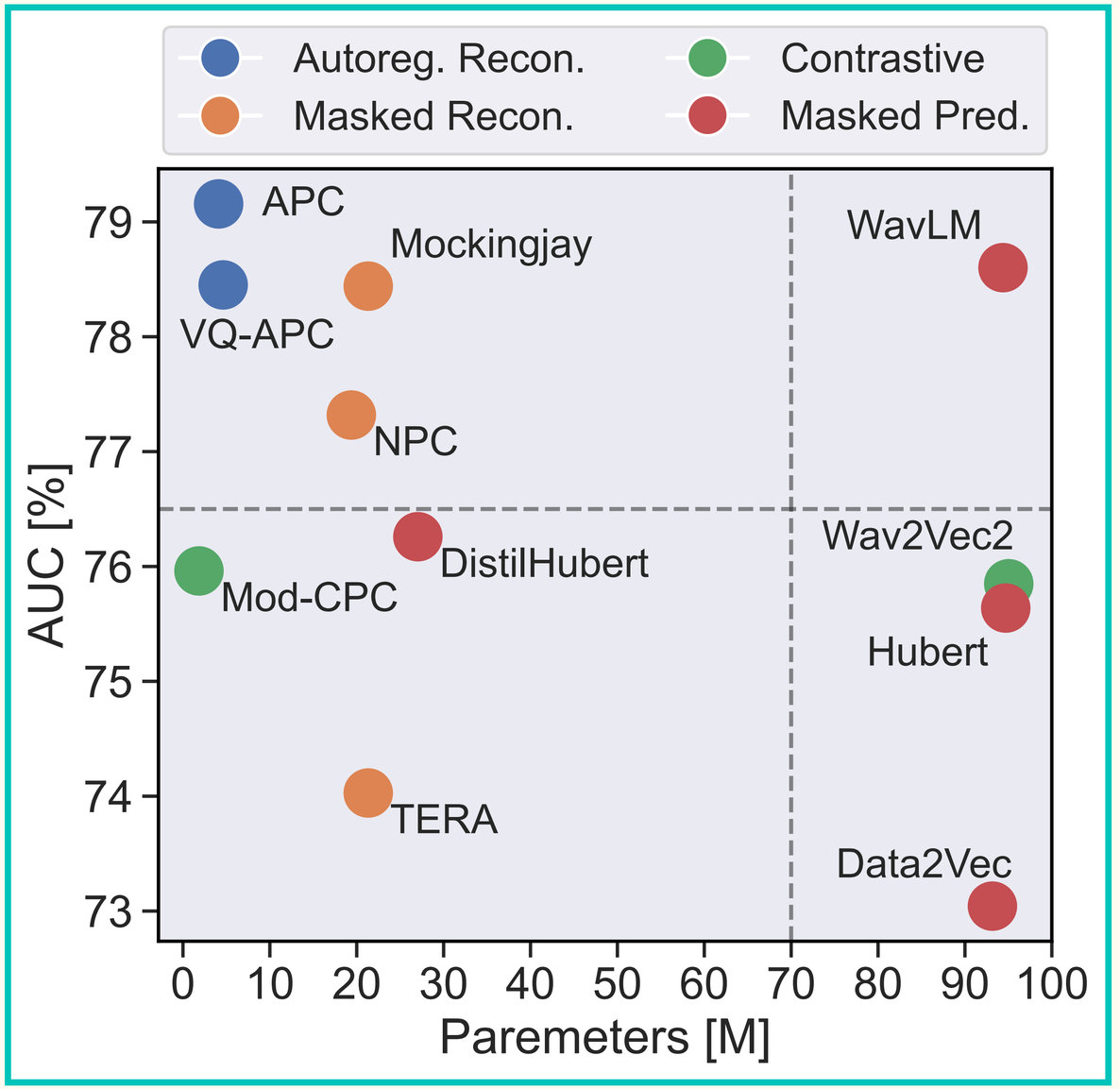
- Relationship between the number of parameters and detection performance for all models.
- No clear pattern.

➡ WavLM's embedding space is more separable than the other masked prediction models.

➡ Both auto-regressive reconstruction models perform exceptionally well with significantly fewer parameters.

➡ All pre-training criteria yields competitive performance, some are more efficient than others, allowing smaller models to perform comparably to larger models.

➡ Data2Vec is weaker than other masked prediction models, despite similar training conditions.



Model size against performance, divided into 4 quadrants.

Summary

- **Aim:** we investigated the applicability of SSL representations, pre-trained on human speech, to analyze animal vocalizations.
- **Findings:** such representations can effectively classify vocalizations in the bio-acoustics domain for tasks such as Marmoset caller detection.
- **Consequence:** findings can greatly benefit bio-acoustics researchers looking to distinguish individual identities within a specific species in their acoustic data.

Limitations and Ongoing Work

- We only looked at embedding spaces. What about traditional spectral features ?
- We only used DNNs pre-trained on human speech. What about other generic audio sets ?
- We only evaluated linear probing. What about fine-tuning on a downstream task ?
- We only looked at marmosets. What about other animals, eg meerkats ?

Thank you !

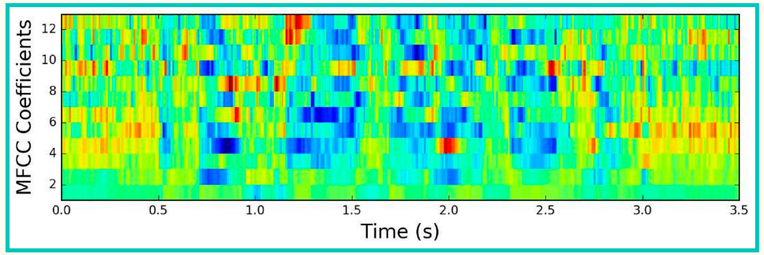

Room 305-8, Idiap Research Institute


www.idiap.ch/~esarkar/


+41 78 82 50 754


eklavya.sarkar@idiap.ch

FAQ - Lack of Baseline (e.g. MFCCs)



- There are ‘no’ prior works in the literature that guide us as to which features are important for caller detection: no baseline feature and approach exist for the task at hand.
- We took inspiration from speech processing, where embedding-based speaker verification systems are becoming the norm, and investigated with SSL neural embedding representations.
- Some SSL models are trained with log-mel spectral features as input.
- Our implemented methods can now serve as baselines. 🙂
- Our focus was not to achieve the best performance.

FAQ - Bias towards certain call-types

- We are constrained by the scarcity of certain call-type classes in the dataset.
- Due to limited data availability, we can't comprehensively investigate this question.
- Even if such a study were carried out on this dataset, it would be challenging to conclude whether the observed differences in performance were due to call-types or data scarcity.
- This issue does not change or invalidate the analysis and findings in our paper.