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

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

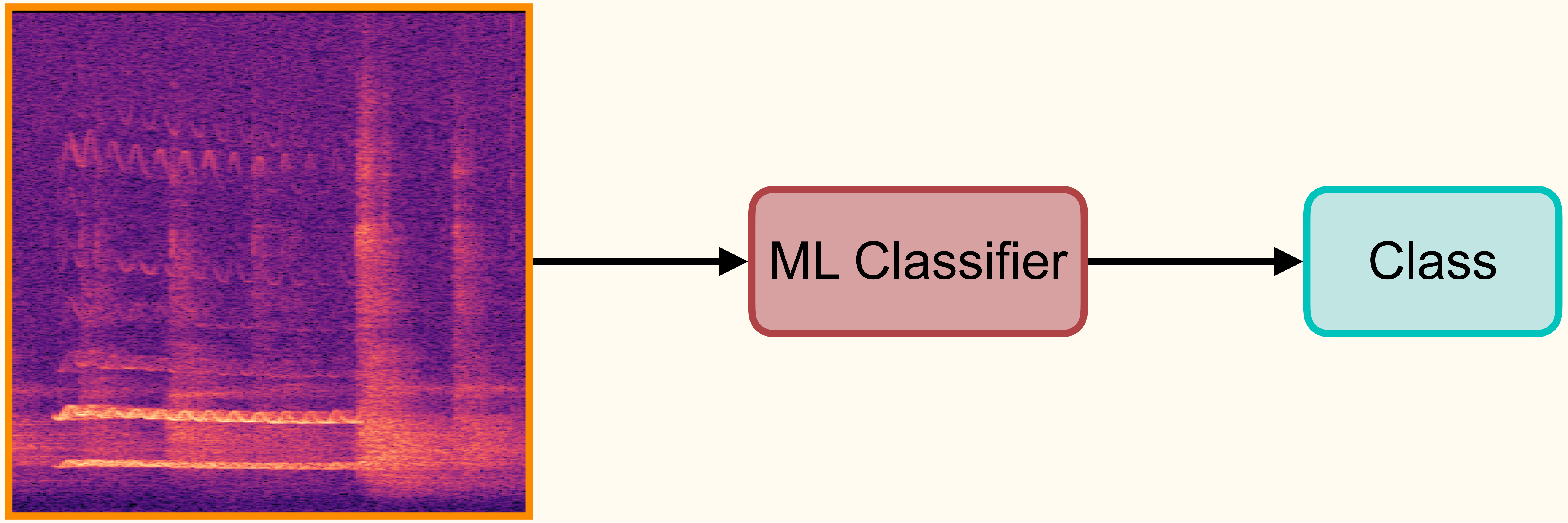
[1] Idiap Research Institute, Switzerland
[2] Ecole polytechnique fédérale de Lausanne, Switzerland

VIHAR 2024
ISCA Interspeech 2024 Satellite Event

September 2024

Introduction

- Bioacoustics a growing field in ML and a theme of Interspeech 2024.
- Tasks typically involve *classification*, *detection*, *denoising* of an animal call.



Animal vocalization

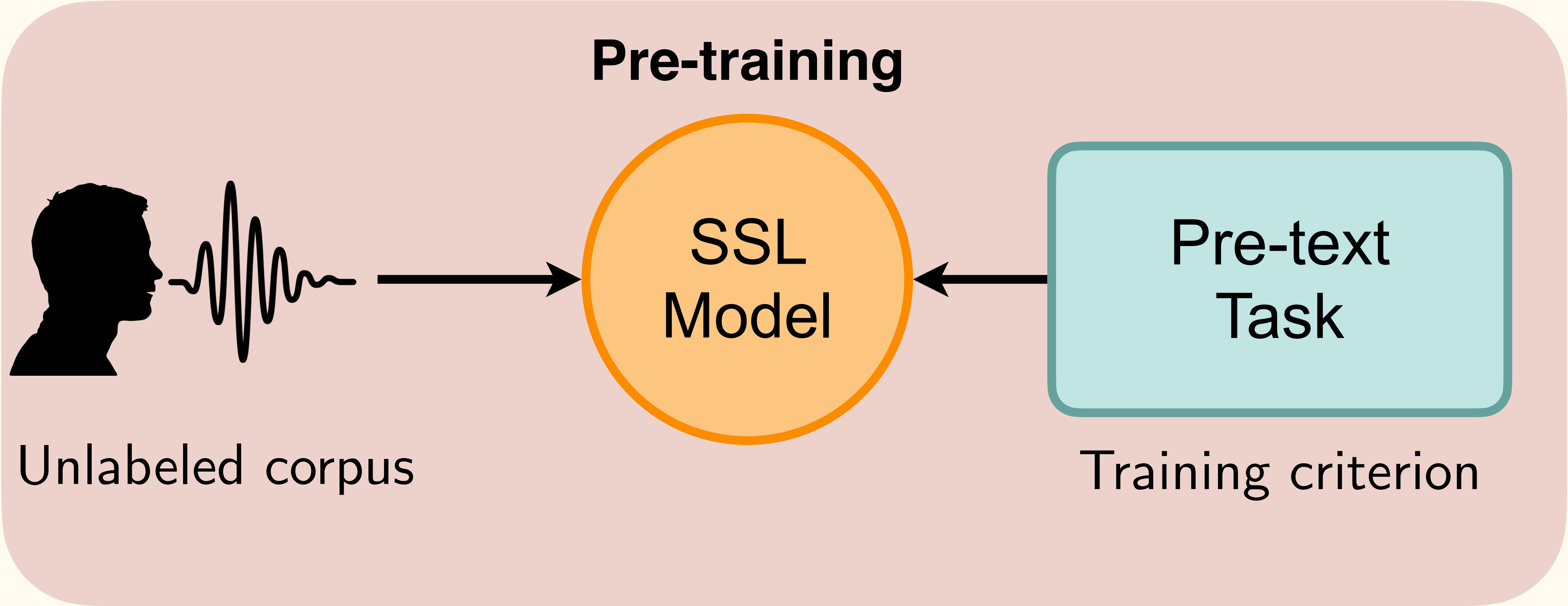
Introduction

Introduction

- Recent trend has been to leverage SSL models pre-trained on **human speech** (WavLM, HuBERT, wav2vec2, etc.) for processing bioacoustics signals[1-3]:
  - PT models are able to classify call-types, individual identities, sex, even without downstream fine-tuning.

[1] Sarkar et al. *Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?* (2023). Proc. of Interspeech.
[2] Sarkar et al. *On Feature Representations for Marmoset Vocal Communication Analysis* (2024). Idiap-Internal-RR.
[3] Cauzinille et al. *Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures* (2024). Proc. of Interspeech.
[4] Abzaliev et al. *Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification* (2024). Proc. of LREC-COLING.

Introduction

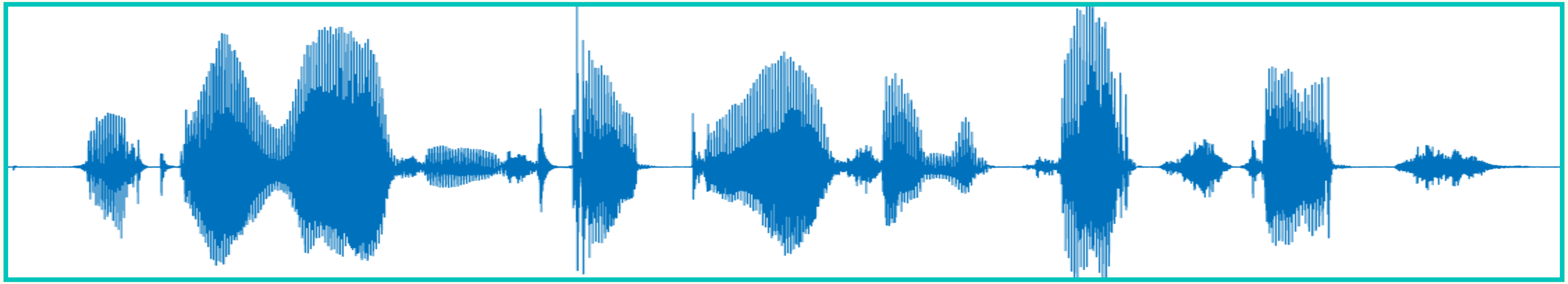


- Since SSLs only learn the intrinsic structure of unlabeled input through a masking pre-text task, they are able to capture essential information independently of any domain-specific knowledge, and thus can be transferred to other acoustic domains.

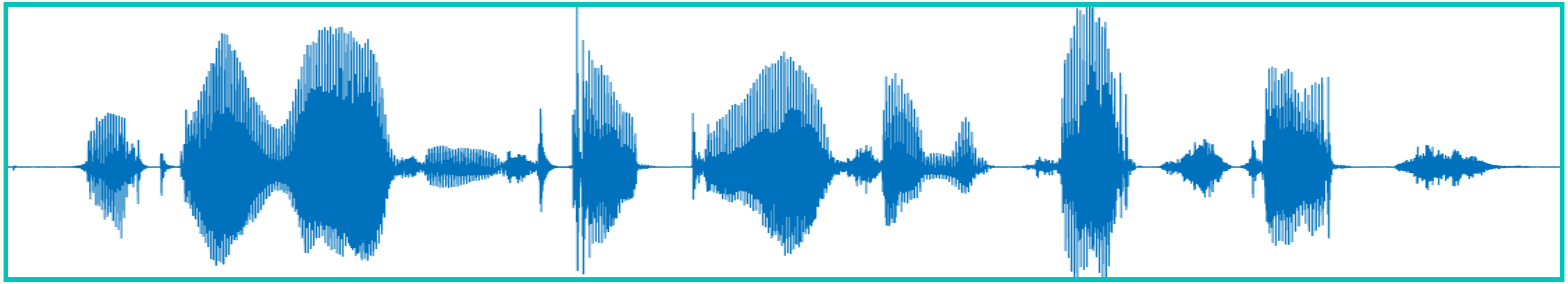
[1] Sarkar et al. *Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?* (2023). Proc. of Interspeech.
[2] Sarkar et al. *On Feature Representations for Marmoset Vocal Communication Analysis* (2024). Idiap-Internal-RR.
[3] Cauzinille et al. *Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures* (2024). Proc. of Interspeech.
[4] Abzaliev et al. *Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification* (2024). Proc. of LREC-COLING.

Marmoset Vocalizations

Marmoset Vocalizations



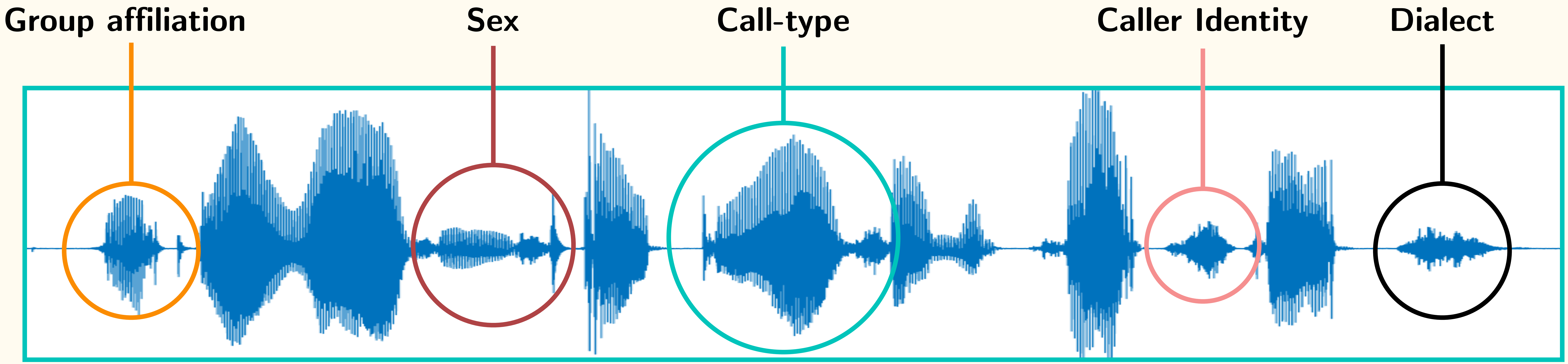
Marmoset Vocalizations



Common marmosets (*Callithrix jacchus*) are of particular interest due to:

- Highly vocal nature rooted in a complex social system.

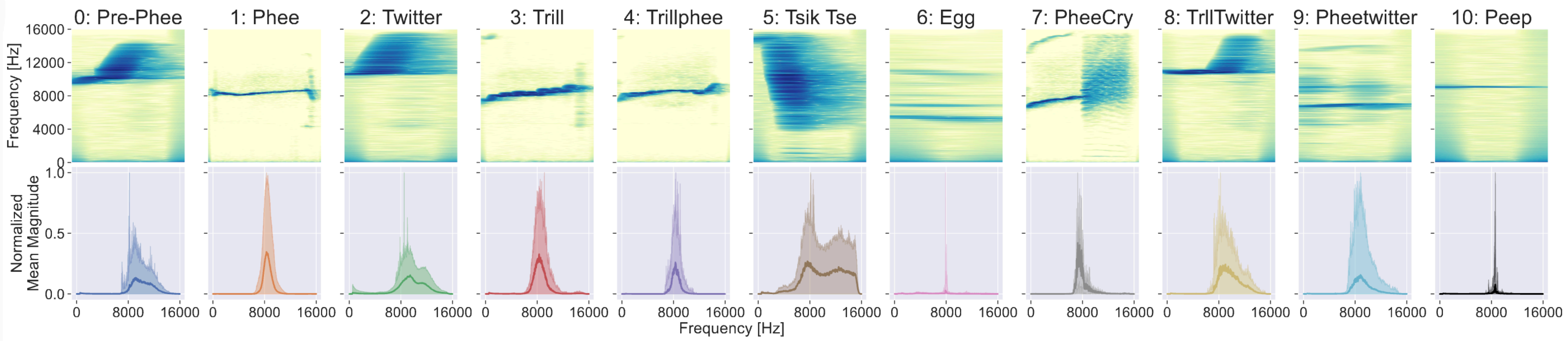
Marmoset Vocalizations



Common marmosets (*Callithrix jacchus*) are of particular interest due to:

- Highly vocal nature rooted in a complex social system.
- Ability to encode a range of information.

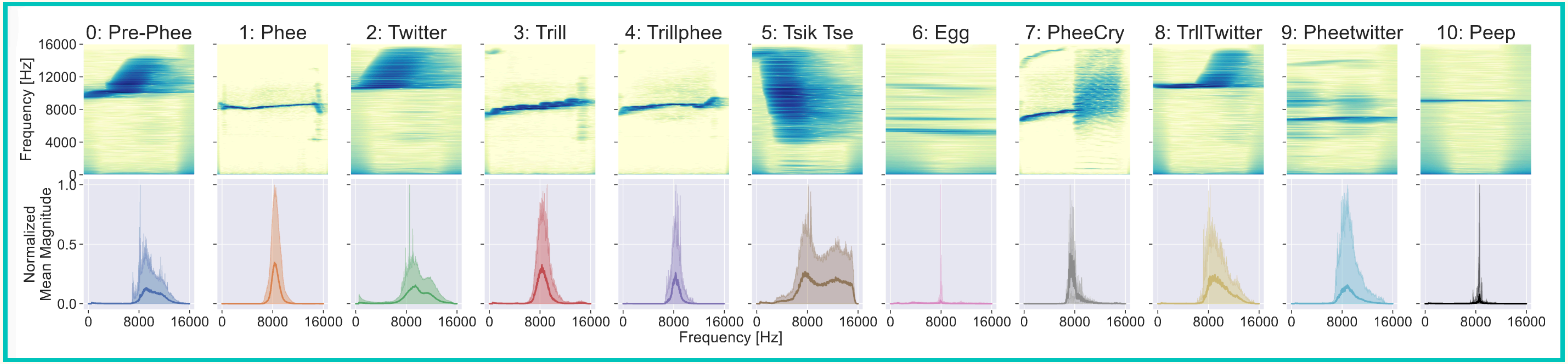
Marmoset Vocalizations



Common marmosets (*Callithrix jacchus*) are of particular interest due to:

- Highly vocal nature rooted in a complex social system.
- Ability to encode a range of information.
- Acoustically diverse call repertoire.

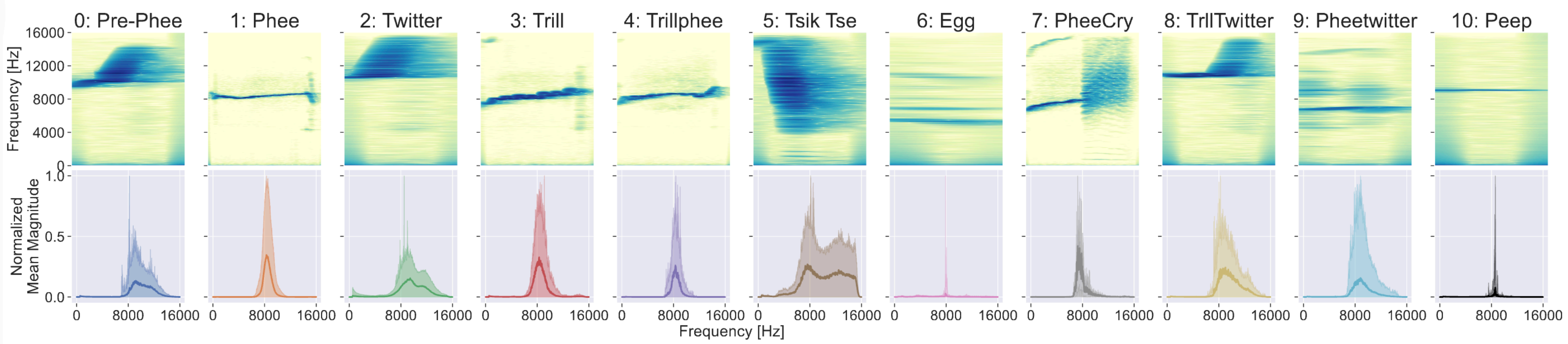
Marmoset Vocalizations



Their remarkable vocal adaptability also allows them to modify their call's:

- Duration
- Intensity
- Complexity
- Timing

Marmoset Vocalizations



Vocal characteristics align them closely with human speech properties:

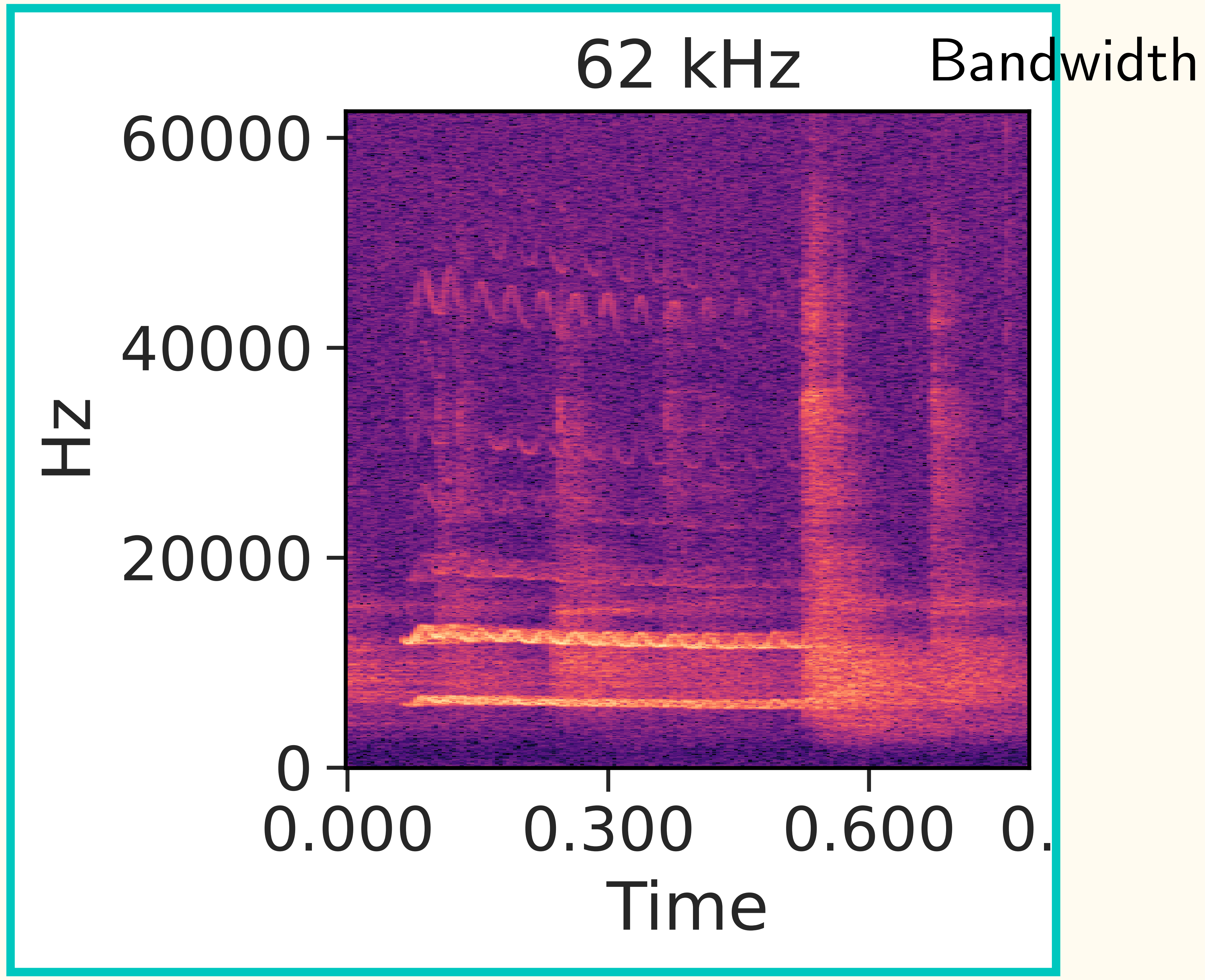
- Turn-taking
- Care-giving to infants
- Categorical perception of sounds

Marmoset Vocalizations

A well-suited surrogate model for
understanding the evolutionary origins of human vocal communication
among biologists and neuroscientists.

Problem: Bandwidth

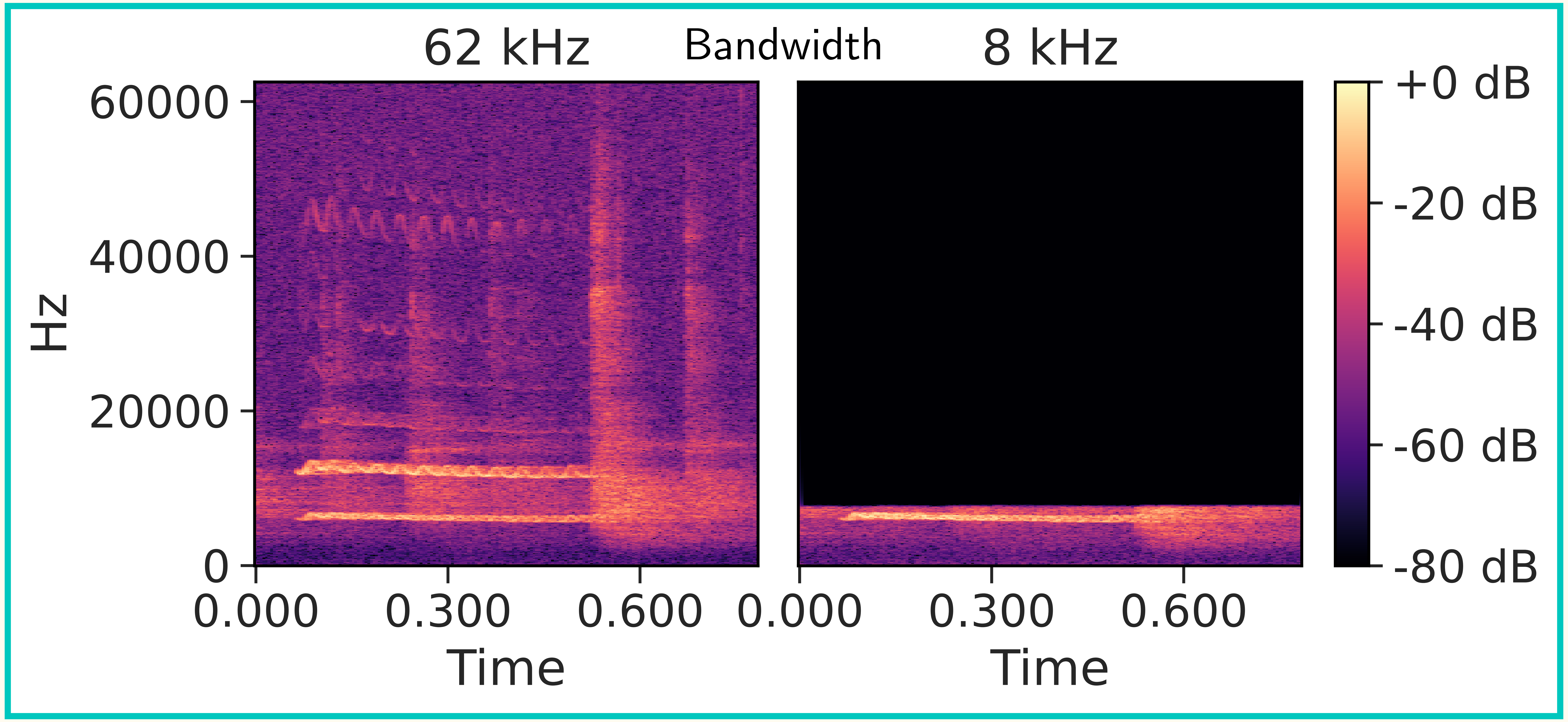
Problem: Bandwidth



Bandwidth = Sampling Rate / 2

Problem: Bandwidth

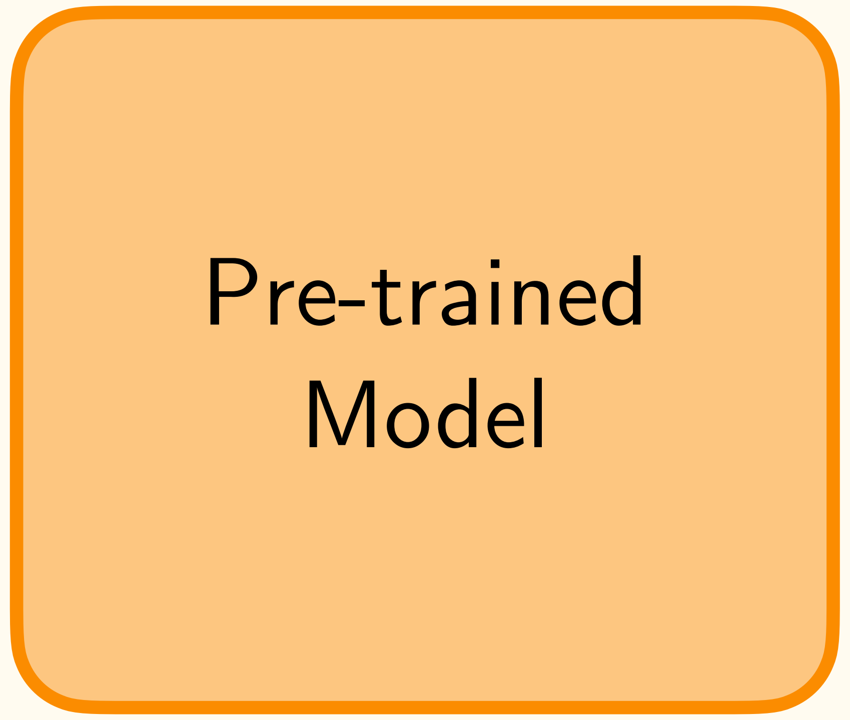
- Models typically pre-trained at 8 kHz bandwidth (16 kHz sampling rate).
- Mismatch with the biological vocalization range of animals.



Bandwidth = Sampling Rate / 2

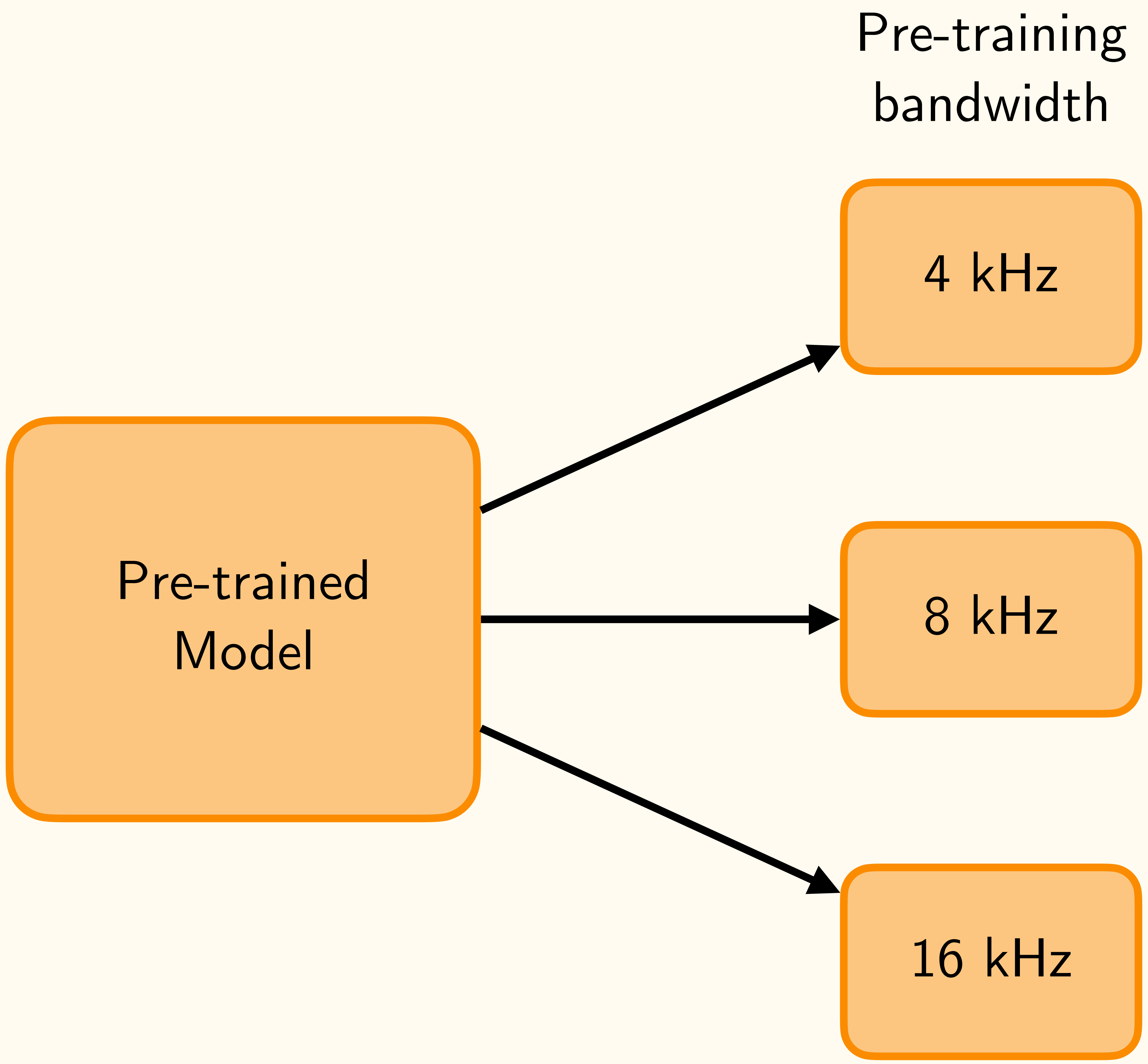
Problem: Bandwidth

- Examine models pre-trained across varying bandwidths.
- Aim to evaluate their effectiveness in adequately representing marmoset calls, and seek to clarify how model bandwidth influences their classification.



Problem: Bandwidth

- Examine models pre-trained across varying bandwidths.
- Aim to evaluate their effectiveness in adequately representing marmoset calls, and seek to clarify how model bandwidth influences their classification.



Problem: Pre-Training Domain

- The influence of the pre-training domain for accurately capturing marmoset call characteristics remains unclear.
- Examine representations produced by different pre-training domains to identify the most suitable pre-training source for cross-domain bioacoustic signal analysis.

General Audio

vs

Human Speech

vs

Hand-crafted

Methodology

Dataset Recording

Dataset Recording

- Used a dataset from a previous paper[1].

[1] *Zhang et al., Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks.* (2018). The Journal of the Acoustical Society of America.
Sarkar et al., *Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?* (2023). Proc. of Interspeech.

Dataset Recording

- Used a dataset from a previous paper[1].
- Inside a 2-layer cage.

Yun et al. Modeling Parkinson's disease in the common marmoset (Callithrix jacchus): Overview of models, methods, and animal care (2023). Laboratory Animal Research.

[1] *Zhang et al., Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks.* (2018). The Journal of the Acoustical Society of America.
Sarkar et al., *Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?* (2023). Proc. of Interspeech.

Dataset Recording

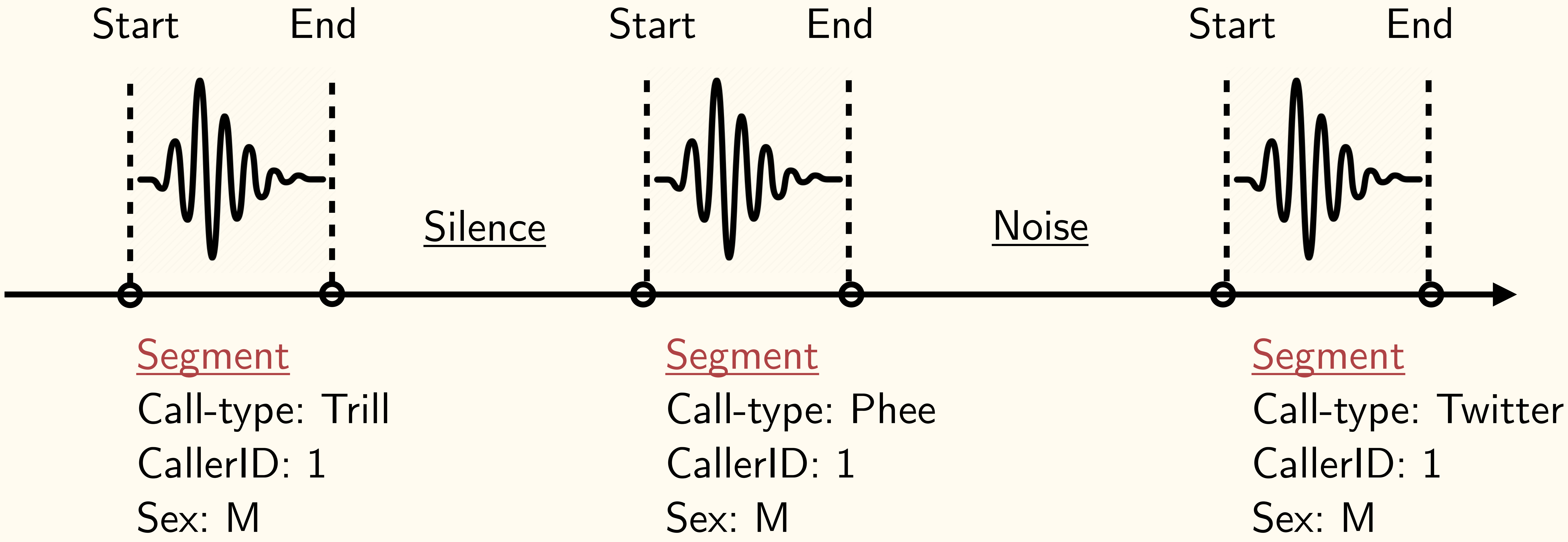
- Used a dataset from a previous paper[1].
- Inside a 2-layer cage.
- Recorded individually with a fixed microphone @ 44.1 kHz without external interference.

Yun et al. Modeling Parkinson's disease in the common marmoset (Callithrix jacchus): Overview of models, methods, and animal care (2023). Laboratory Animal Research.

[1] *Zhang et al., Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks.* (2018). The Journal of the Acoustical Society of America.
Sarkar et al., *Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?* (2023). Proc. of Interspeech.

Dataset Recording

- Data manually annotated by an experienced researcher:
  - Vocalization segments: [Start, End, Call-type, CallerID, Sex].
  - Removed any silence and noise segments.



Dataset

- 73k vocalization segments (7.7 hours).
- 11 call-types & 10 caller classes.

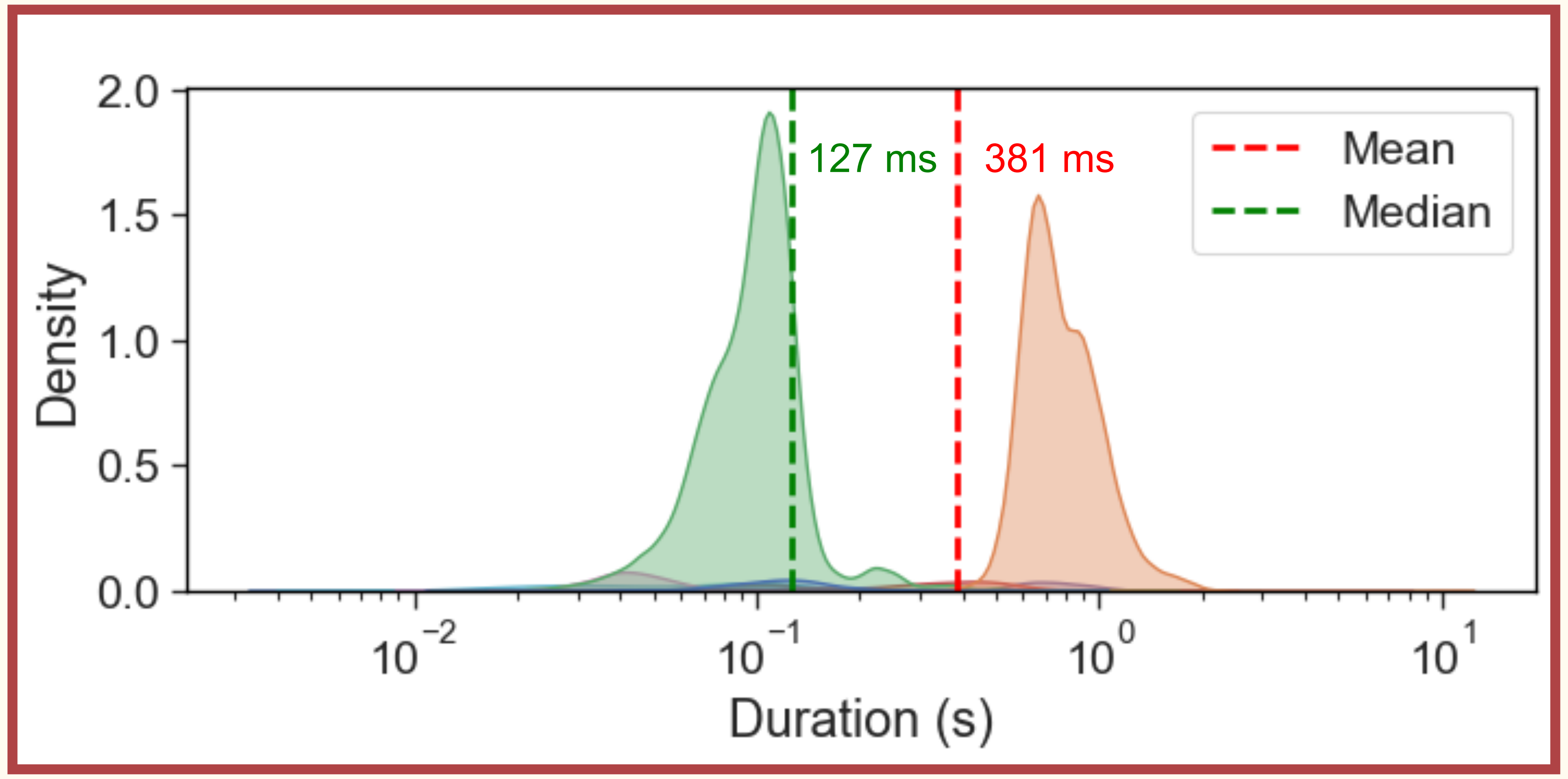
InfantMarmosetsVox dataset statistics

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

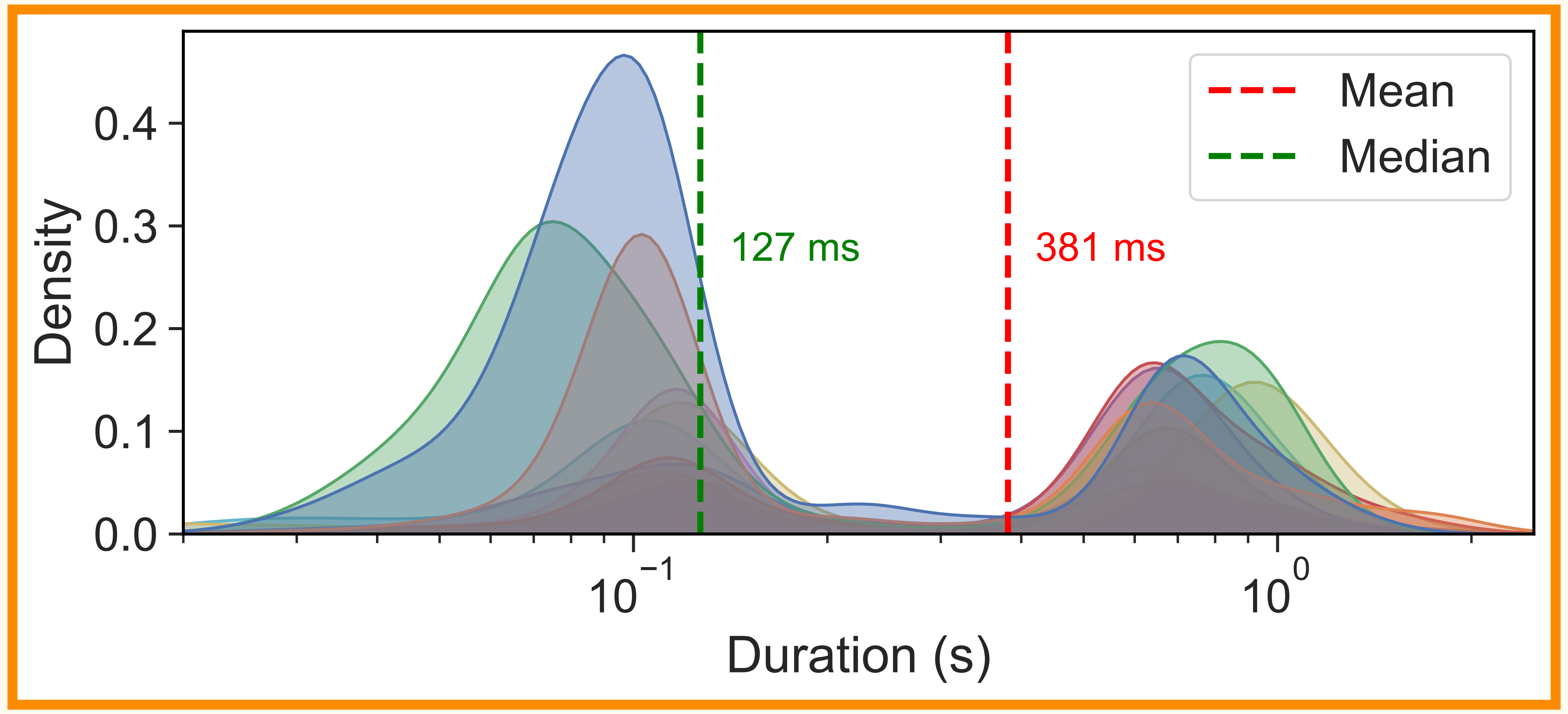
Caller ID	Count
0	15521
1	8648
2	13827
3	5838
4	5654
5	3522
6	4389
7	2681
8	6387
9	6454
-	-
Total	**72921**

Dataset

- 73k vocalization segments (7.7 hours).
- 11 call-types & 10 caller classes.
- Predominantly short (127 ms median).
- Tasks:
  - Call-type classification (CTID).
  - Caller classification (CLID).
- Protocol: *70:20:10* split *Train:Val:Test*.
- Metrics: Unweighted Average Recall (UAR) to account for class imbalance.



Log distribution of vocalization lengths for call-types.



Log distribution of vocalization lengths for callers 1-10.

Models and Feature Representations

Num. of parameters P and feature dimension D of selected models, pre-trained on AudioSet (AS) or LibriSpeech (LS).

	$\mathcal{F}$	Corpus	P	D	Type
Handcrafted (spectral) baseline →	C22 [1]	-	-	24	HC
Pre-trained on human speech →	WavLM [2]	LS	94.38M	1536	SSL
Pre-trained on general audio →	BYOL [3]	AS	5.32M	2048	SSL
Pre-trained on general audio →	PANN [4]	AS	8.08M	2048	SL

[1] Lubba et al., *Catch22: Canonical Time-Series Characteristics*, (2019). Data Mining and Knowledge Discovery.
[2] S. C. et al., *WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing*, (2022) IEEE Journal of Selected Topics in Signal Processing.
[3] Niizumi et al., *Byol for audio: Self-supervised learning for general-purpose audio representation.* (2021). IEEE International Joint Conference on Neural Networks (IJCNN).
[4] Kong et al., *PANN: Large-scale pretrained audio neural networks for audio pattern recognition.* (2020). IEEE/ACM Transactions on Audio, Speech, and Language Processing.

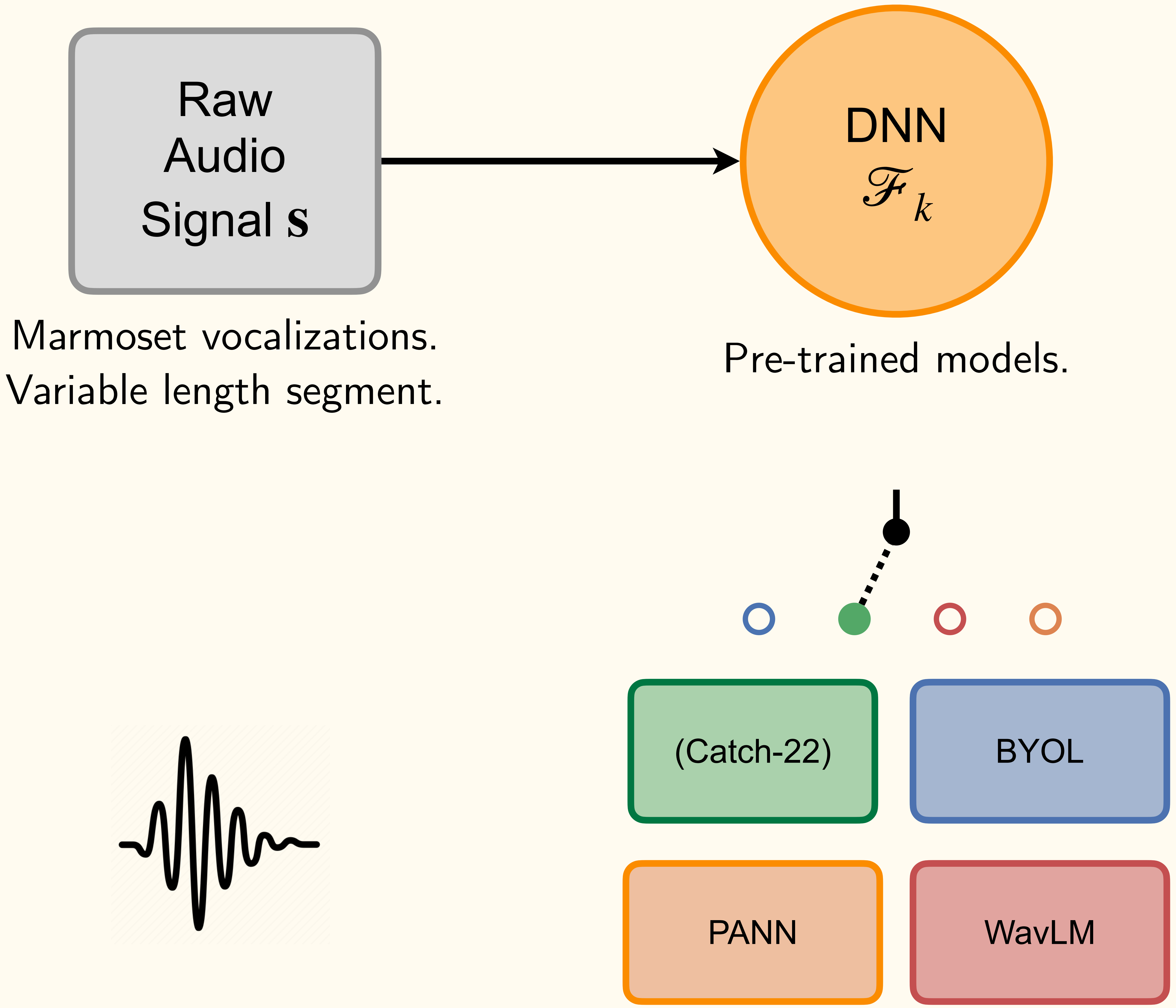
Feature Extraction

Feature Extraction

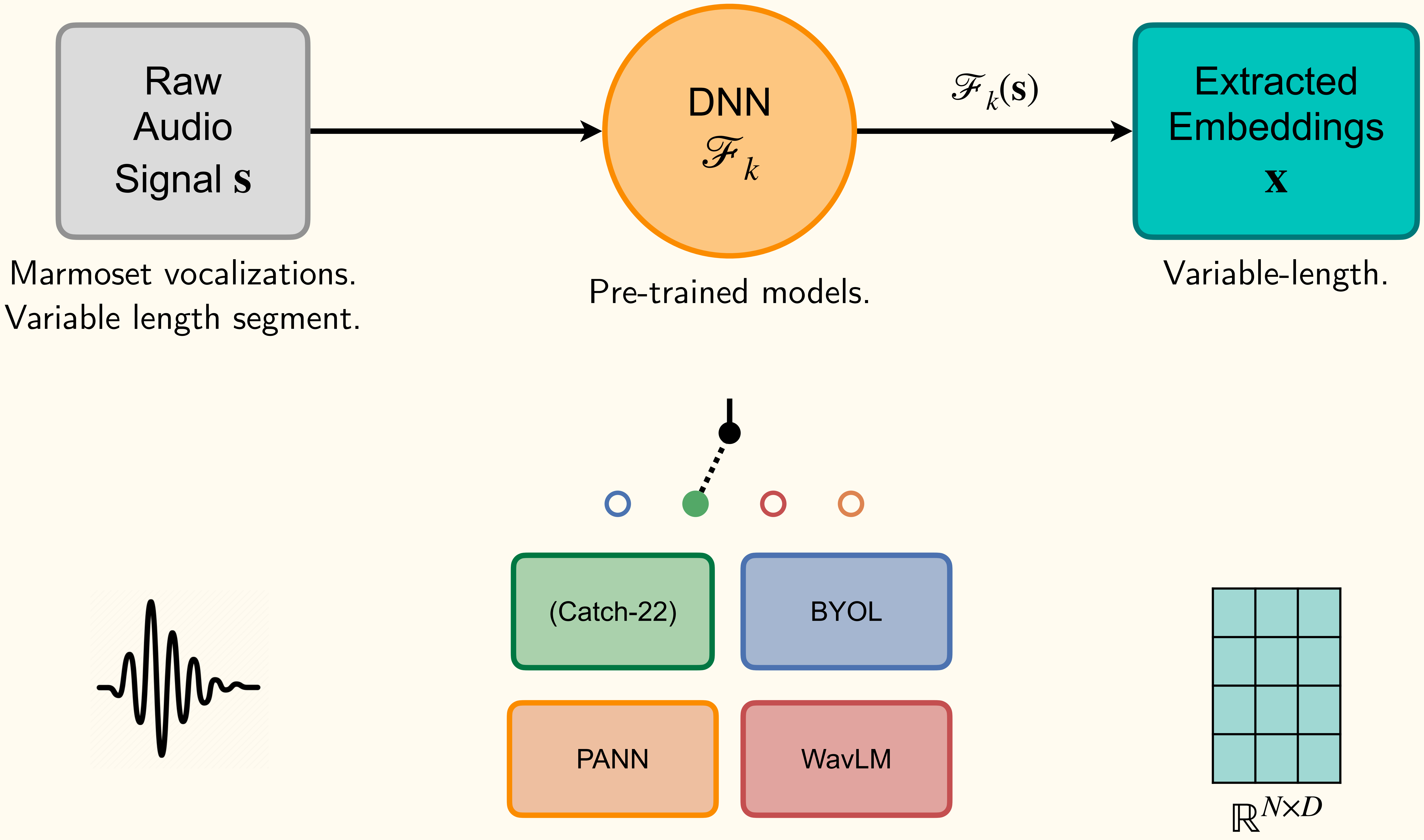
Raw
Audio
Signal $\mathbf{s}$

Marmoset vocalizations.
Variable length segment.

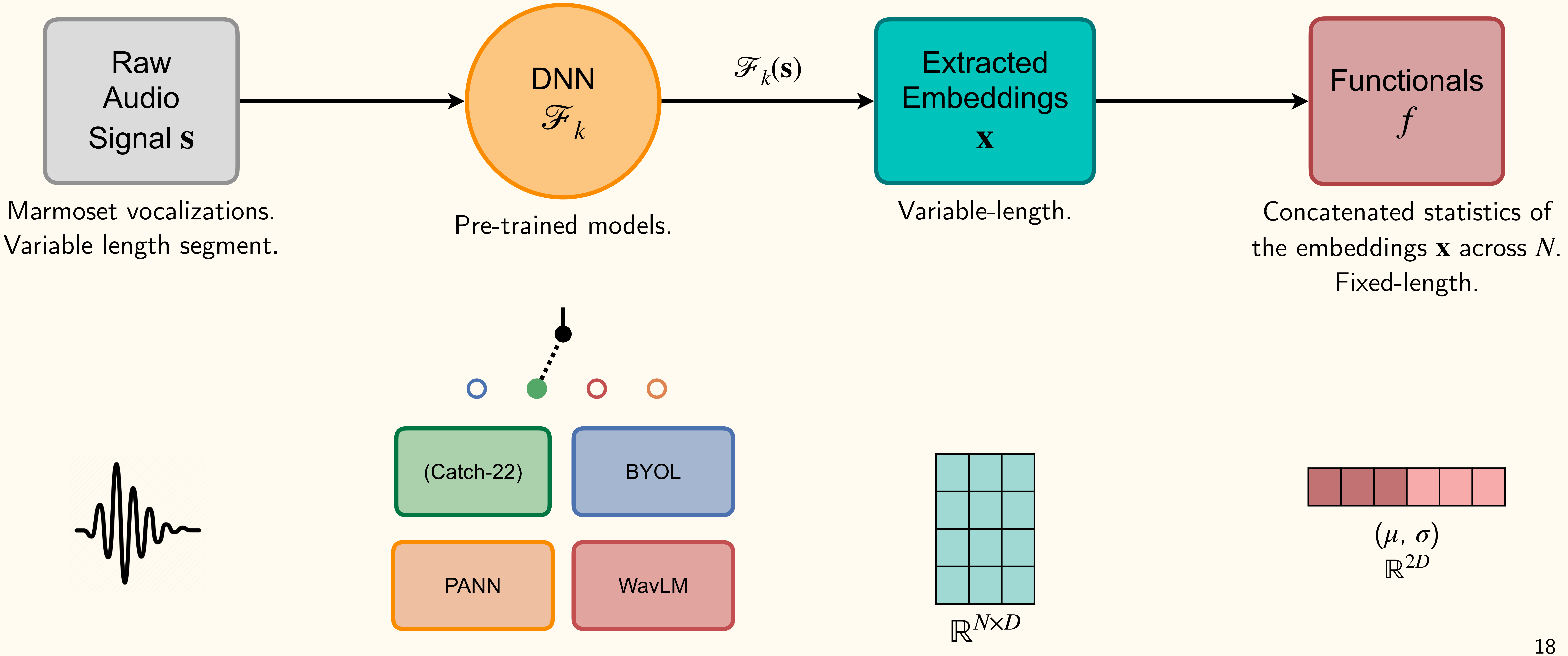
Feature Extraction



Feature Extraction



Feature Extraction

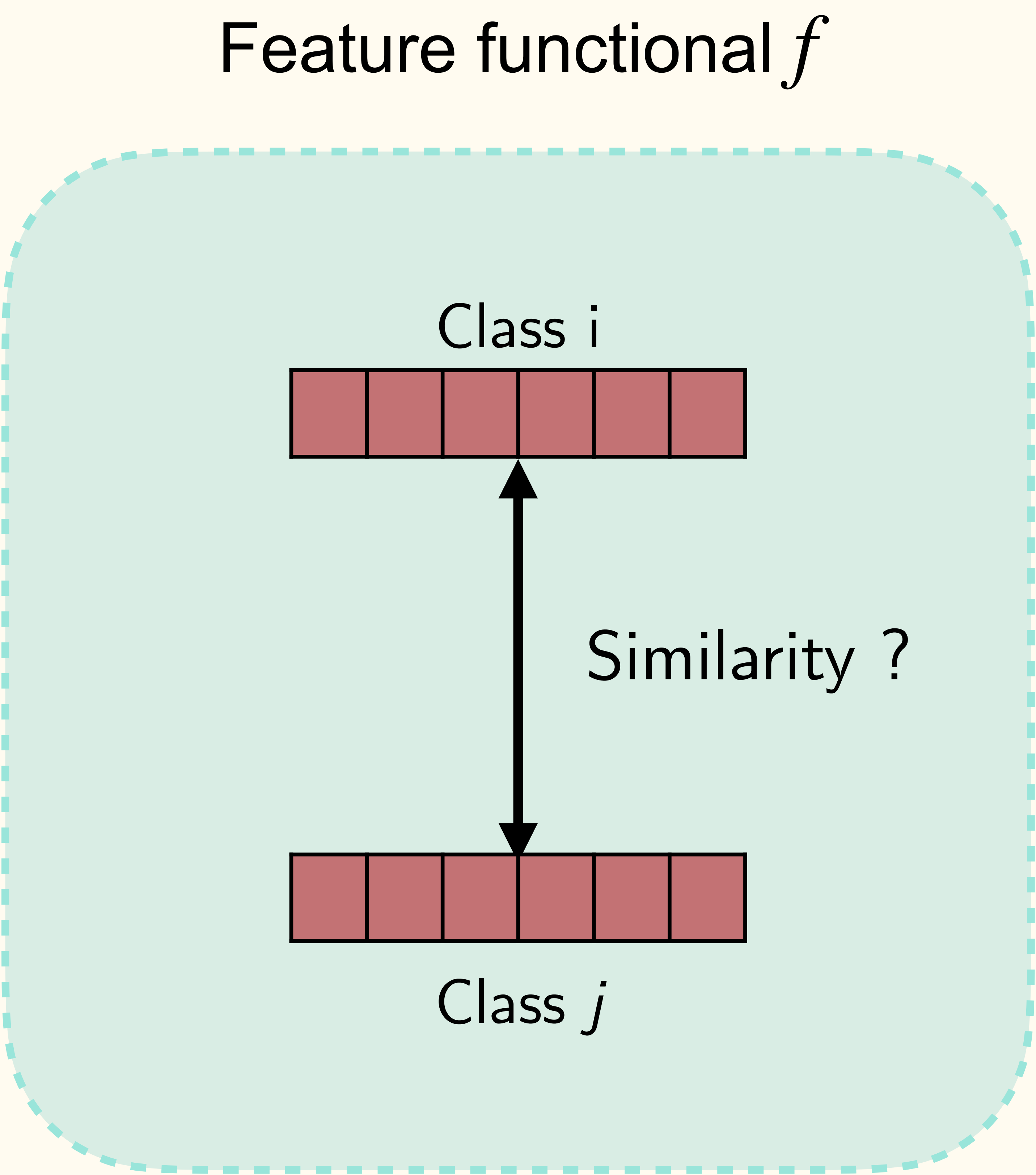


Call Similarity Analysis

Call Similarity Analysis

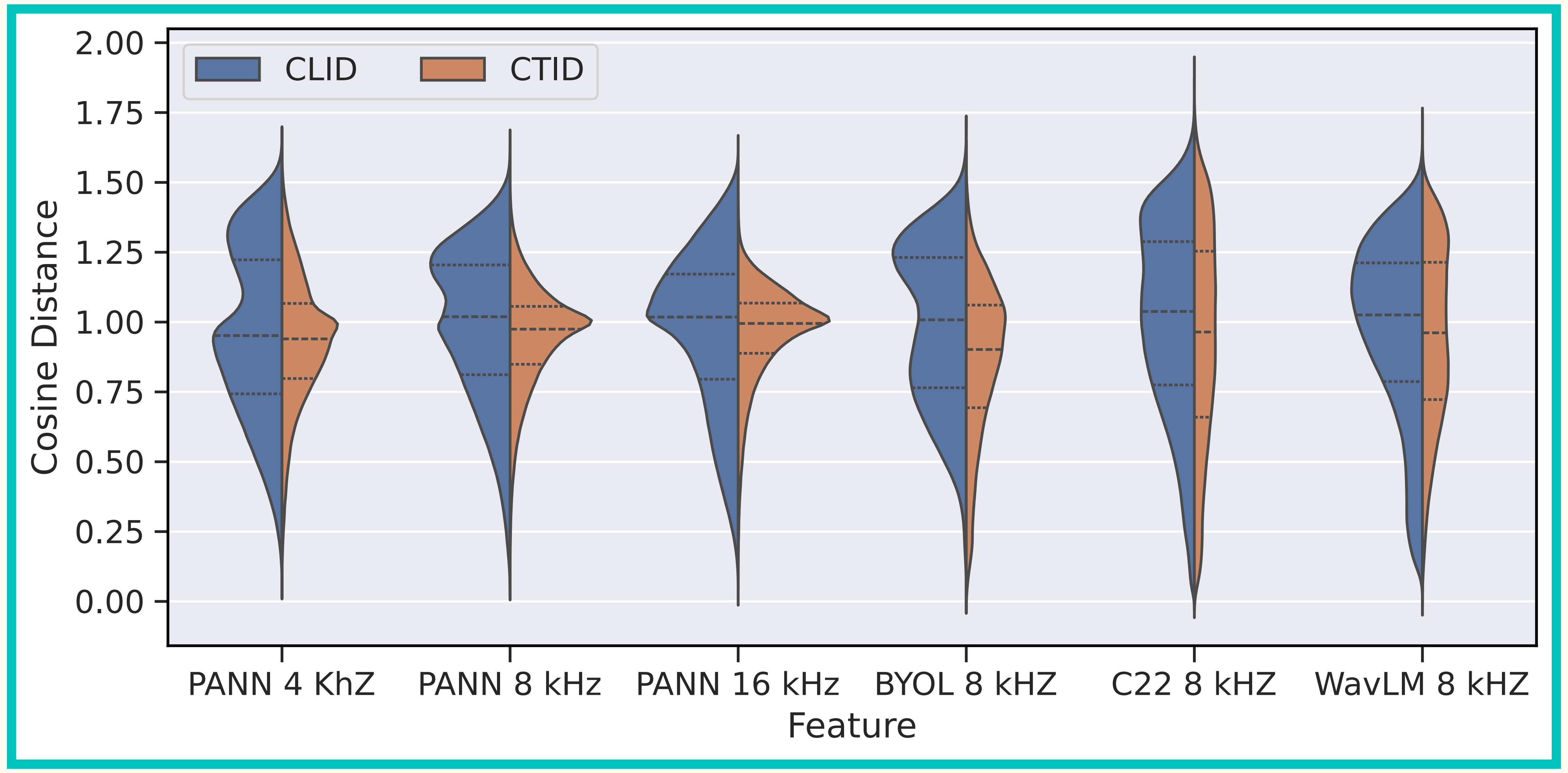
- Do variations in the bandwidth affect the similarity distributions of the intra-class embeddings ?

- Do we see any distinctions between the models pre-trained on speech vs. general audio ?



Call Similarity Analysis

- Distributions centered around a median distance of 1 for all features.
  - Suggests a lack of clear correlation or similarity within the embeddings generated.

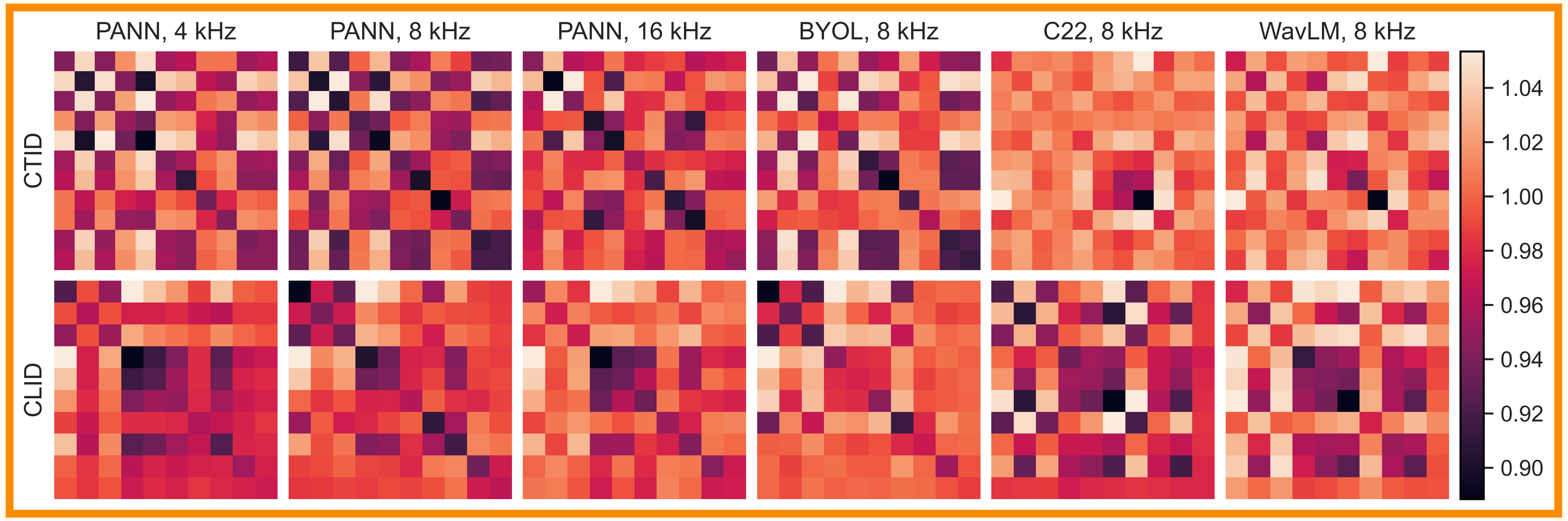


General distribution of pairwise cosine distances [0-2] on *Test*.

sim(f1, f2) = 0 → Identical.
sim(f1, f2) = 1 → Orthogonal.
sim(f1, f2) = 2 → Opposite.

Call Similarity Analysis

- Can delineate distributions into distance matrices.
- Ideal scenario: intra-class distances smaller than inter.

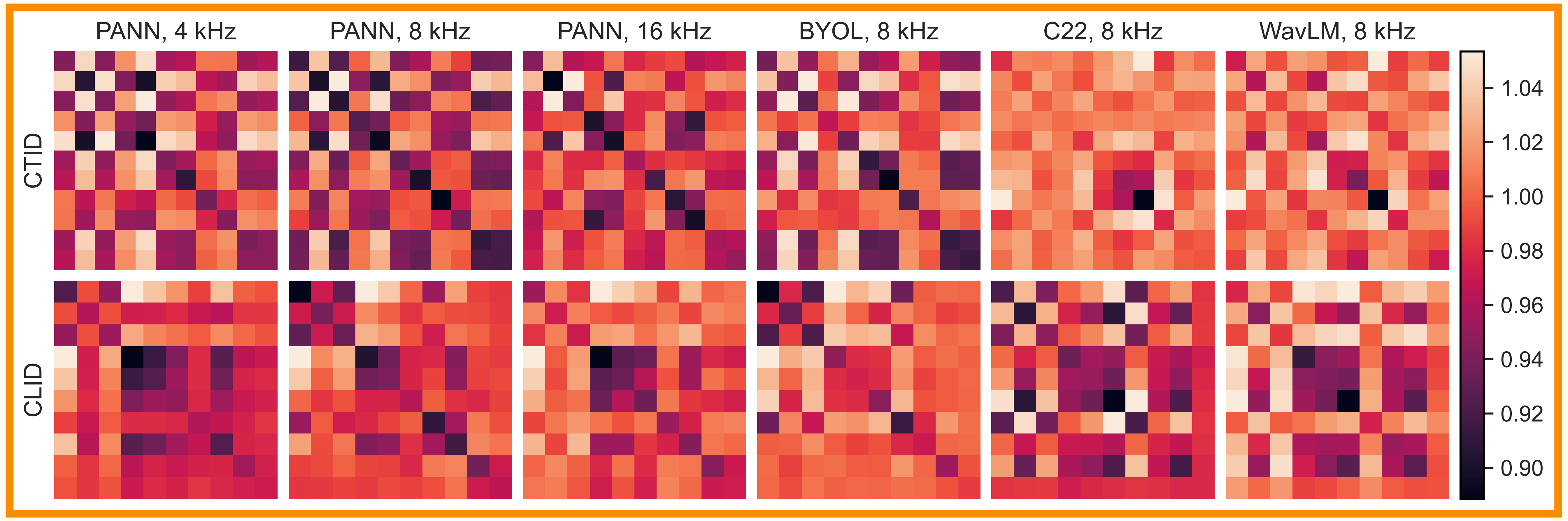


Pairwise mean cosine distances [0-2] matrices.

Diagonal: intra-class distances
Off-diagonal: inter-class distances.
Darker: higher similarity.

Call Similarity Analysis

- Models PT'd on general audio (BYOL and PANN) yield more distinct diagonals than those PT'd on speech (WavLM).
- Marginal level of class-specific correlation, but mostly features seem to be highly orthogonal.
- No clear linear separability. Challenging to classify ?

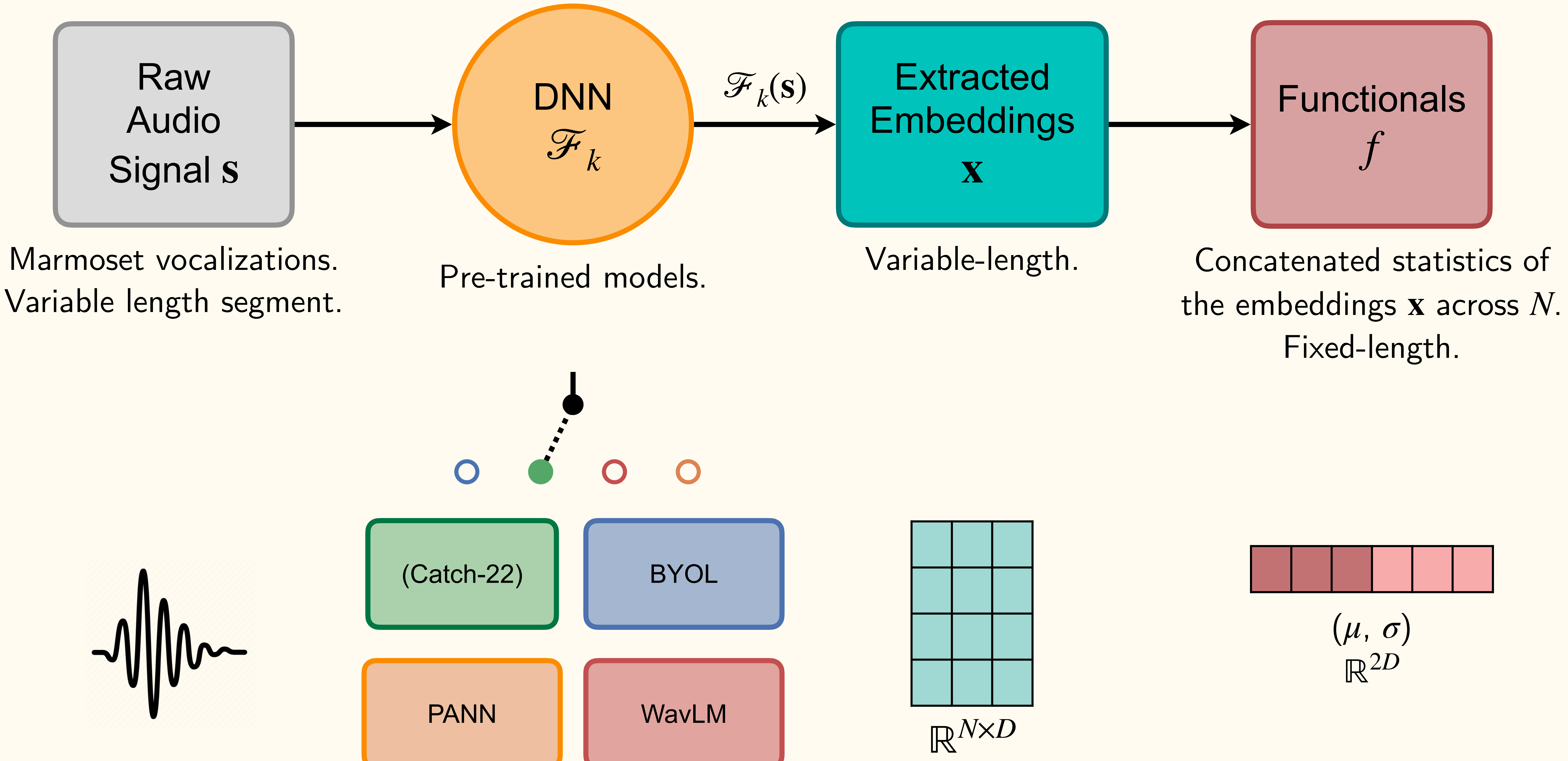


Pairwise mean cosine distances [0-2] matrices.

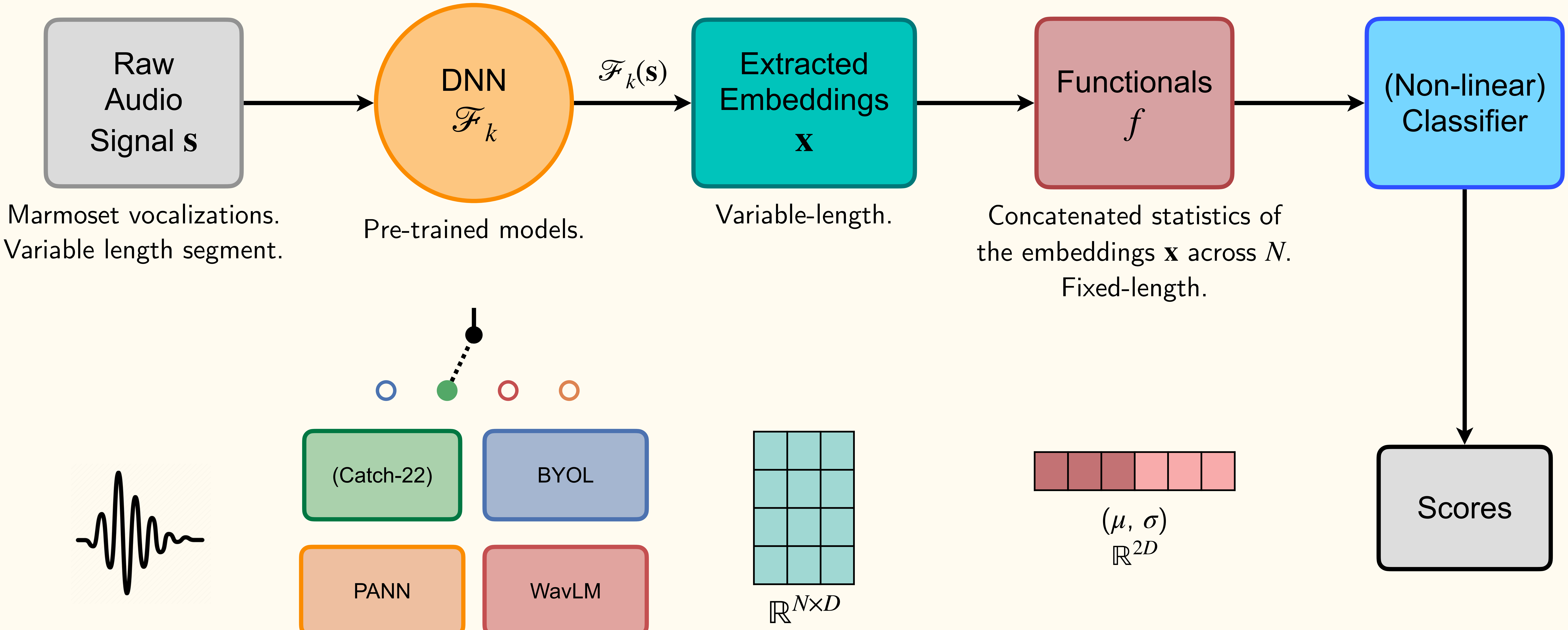
Diagonal: intra-class distances
Off-diagonal: inter-class distances.
Darker: higher similarity.

Classification Analysis

Classification



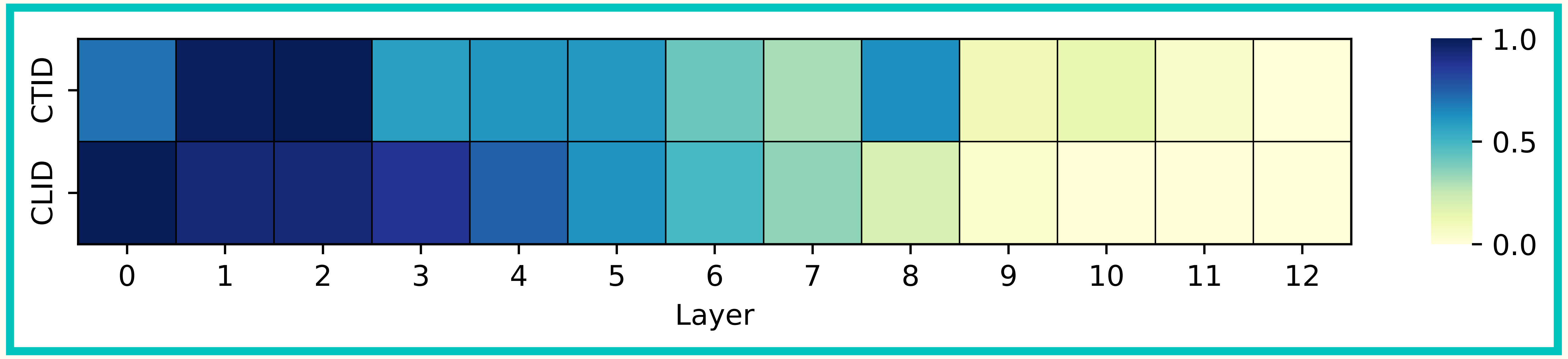
Classification



Classification Analysis

For WavLM: we classify each layer.

- Lower layers are clearly much more salient representations for both tasks compared to higher layers.
- Higher layers: modeling phonotactic information ?
- We use the best individual WavLM layers for our two tasks.



Layer-wise UAR scores of WavLM features, normalized [0,1] per task. Darker regions indicate a higher performance.

Classification Analysis

(a) Results of features @ 8 kHz BW.

- BYOL outperforms the others, for both CTID and CLID.
- Despite having fewer params than WavLM & PANN.
- Hand-crafted C22 is the overall weakest representation.
- WavLM shows highest difference in performance across tasks.

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

UAR scores [%] on *Test* for pre-trained features F.
Random performance = 100 / # classes.
For WavLM, the best layer's score is given.

Classification Analysis

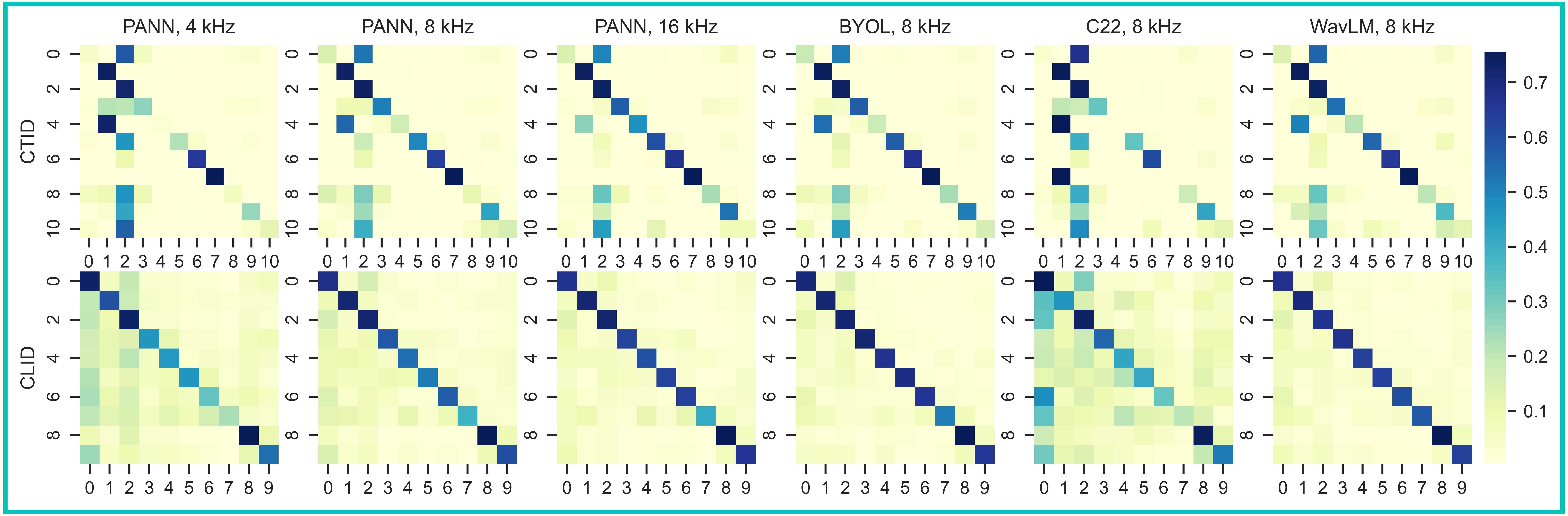
(b) Impact of bandwidth during pre-training.

- Bandwidth size correlates directly with the performance, increasing monotonically.
- PANN features at 16 kHz achieve the highest performance across all features and BWs for CTID.
- The best scores for both tasks are also closely matched in value.

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**

UAR scores [%] on *Test* for pre-trained features F.
Random performance = 100 / # classes.
For WavLM, the best layer's score is given.

Classification Analysis



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

Summary

Conclusion

- Investigated the utility of foundations models for marmoset call analysis.
  - Showed that a larger bandwidth directly correlates with improved performance.
  - Pre-training on general audio showed improved performance over speech.
- 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.

Thank you !

Idiap Research Institute

https://github.com/idiap/speech-utility-bioacoustics

https://zenodo.org/records/10130104
(Includes PyTorch Dataset & Dataloader !)

eklavya.sarkar@idiap.ch

FAQ - MLP Classifier

- **Model**: 3-layer MLP

Block	Layers	# Hidden Units	Activation
1	Linear, LayerNorm	128	ReLU
2	Linear, LayerNorm	64	ReLU
3	Linear, LayerNorm	32	ReLU
4	Linear	# classes	

- **Training**: 30 epochs, Adam optimizer, η-scheduler factor 0.1, patience 10 epochs.
- **Grid search**: values of batch-size [32, 64 ..., 512] and η across [1e-3, 1e-4].
- **Protocol**: 70:20:10 split of *Train:Val:Test* sets.
- **Metrics**: Unweighted Average Recall (UAR) to account for class imbalance.

FAQ - PANN

- CNN14 Model
- Balanced sampling strategy across AudioSet's classes.
- Embeddings from final FC layer*
- Works on a log-mel base.

PANN models parameters

BW [kHz]	**4**	**8**	**16**
Window Size	256	512	1024
Hopp Size	80	160	320
Mel Bins	64	64	64
F_{min}	50	50	50
F_{max}	4000	8000	16000

PANN Architecture

```
# Spectrogram extractor
self.spectrogram_extractor = Spectrogram()

# Logmel feature extractor
self.logmel_extractor = LogmelFilterBank()

# Spec augmenter
self.spec_augmenter = SpecAugmentation()

# Model
self.bn0 = nn.BatchNorm2d(64)

self.conv_block1 = ConvBlock(in_channels=1, out_channels=64)
self.conv_block2 = ConvBlock(in_channels=64, out_channels=128)
self.conv_block3 = ConvBlock(in_channels=128, out_channels=256)
self.conv_block4 = ConvBlock(in_channels=256, out_channels=512)
self.conv_block5 = ConvBlock(in_channels=512, out_channels=1024)
self.conv_block6 = ConvBlock(in_channels=1024, out_channels=2048)

self.fc1 = nn.Linear(2048, 2048, bias=True)
# self.fc_audioset = nn.Linear(2048, classes_num, bias=True)
```

FAQ - BYOL

- AudioNTT2020 Model
- BYOL-A architecture
- Embeddings from final FC layer*
- Works on a log-mel base.

BYOL models parameters

BW [kHz]	8
Window Size	64
Hopp Size	10
Mel Bins	64
F_{min}	60
F_{max}	8000

BYOL Architecture

TABLE IV
ENCODER NETWORK ARCHITECTURE (2048-D)

Layer-#	Layer prms.	Output shape	Parameters
Conv2D-1	3x3@64	[B, 64, 64, 96]	640
BatchNorm2D-2		[B, 64, 64, 96]	128
ReLU-3		[B, 64, 64, 96]	0
MaxPool2D-4	2x2,stride=2	[B, 64, 32, 48]	0
Conv2D-5	3x3@64	[B, 64, 32, 48]	36,928
BatchNorm2D-6		[B, 64, 32, 48]	128
ReLU-7		[B, 64, 32, 48]	0
MaxPool2D-8	2x2,stride=2	[B, 64, 16, 24]	0
Conv2D-9	3x3@64	[B, 64, 16, 24]	36,928
BatchNorm2D-10		[B, 64, 16, 24]	128
ReLU-11		[B, 64, 16, 24]	0
MaxPool2D-12	2x2,stride=2	[B, 64, 8, 12]	0
Reshape-13		[B, 12, 512]	0
Linear-14	out=2048	[B, 12, 2048]	1,050,624
ReLU-15		[B, 12, 2048]	0
Dropout-16	0.3	[B, 12, 2048]	0
* → Linear-17	out=2048	[B, 12, 2048]	4,196,352
ReLU-18		[B, 12, 2048]	0
$\max(\cdot) \oplus \text{mean}(\cdot)$-19		[B, 2048]	0

FAQ - BYOL

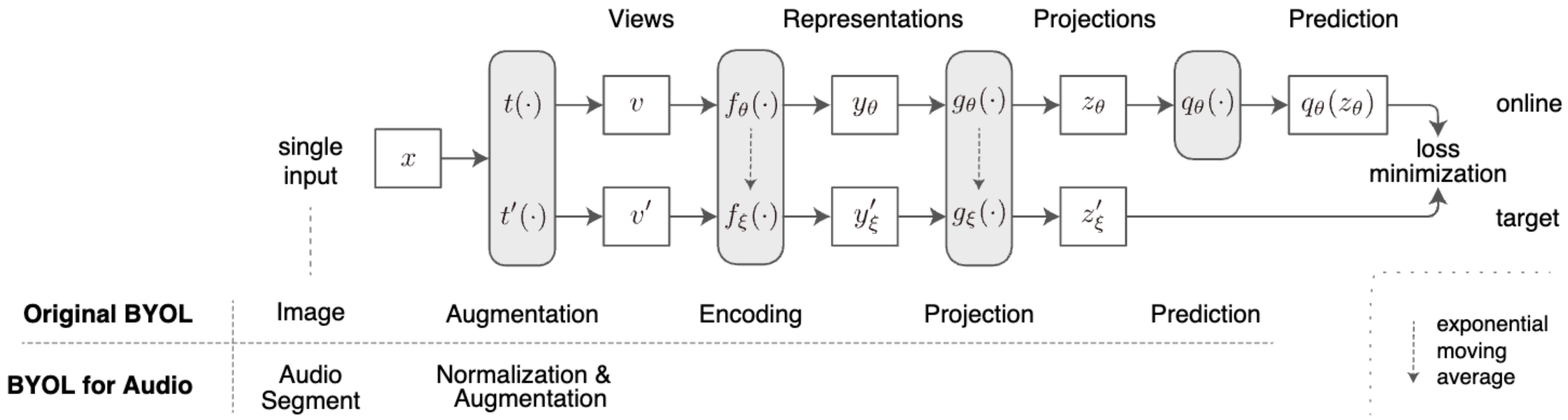


Fig. 2. BYOL and BYOL-A system overview.

FAQ - Catch-22

- Subset of *Highly Comparable Time-Series Analysis* (HCTSA):
  - 7700 features through signal processing methods (eg LPC, Wavlet transform).
  - Tested on: birdsongs, ecosystem monitoring, and marmoset caller identification.
  - Significant limitations: computational demands and feature redundancy.

- Catch-22: steamlined subset of HCTSA.
- High performance with minimal redundancy across many classification problems.
- Add first and second order statics to make it $D = 24$.

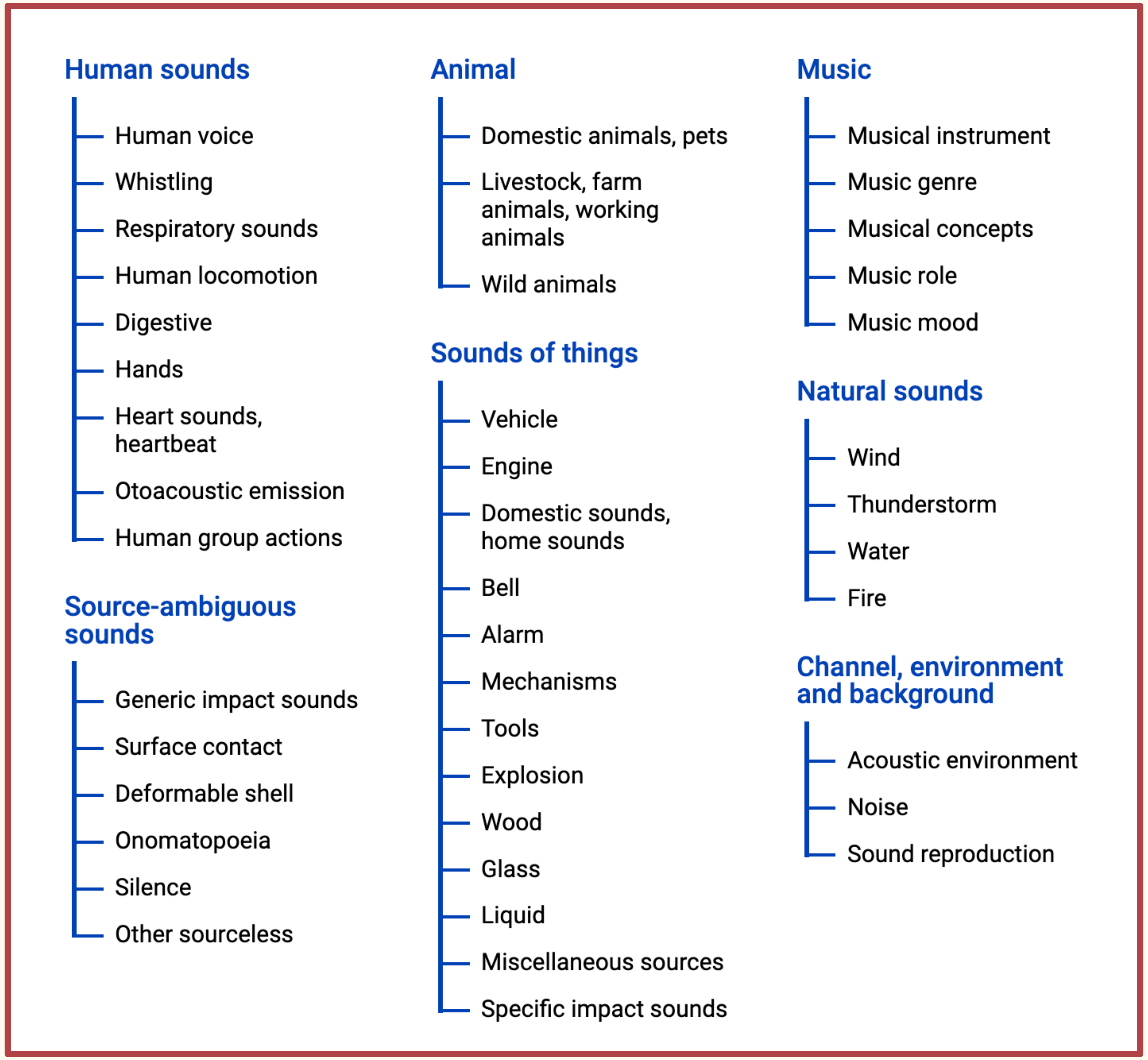
FAQ - WavLM

- Base model.
- Pre-trained on the 960h LibriSpeech.
- 13 encoder transformer layers.

FAQ - AudioSet

Audio event classes such as:

- Environmental sounds.
- Musical instruments.
- Human and animal vocalizations.



AudioSet Dataset Ontology

FAQ - Audio Classification

- Audio classification isn't synonymous to biological acoustic signals analysis like speech, marmoset calls, which contain vocal and linguistic structures.
- Our work shows the utility of BYOL and PANN for Marmoset vocalization analysis along with WLM.