Transferability of Learnt Speech Representations for Decoding Non-Human Vocal Communication

Eklavya Sarkar

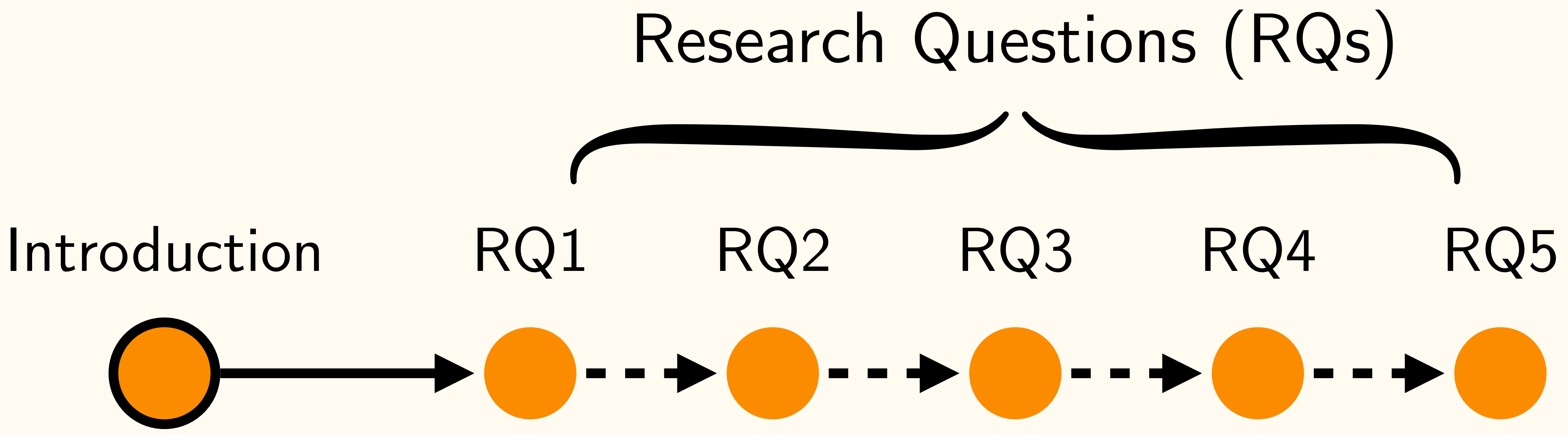
EPFL PhD Defense
8th August 2025

Thesis directors: Dr. J.M. Odobez and Dr. M. Magimai-Doss.
Committee members: Prof. V. Cevher, Prof. D. Van der Ville, Dr. M. Cernak, Dr. M. Miron.

Table of Contents

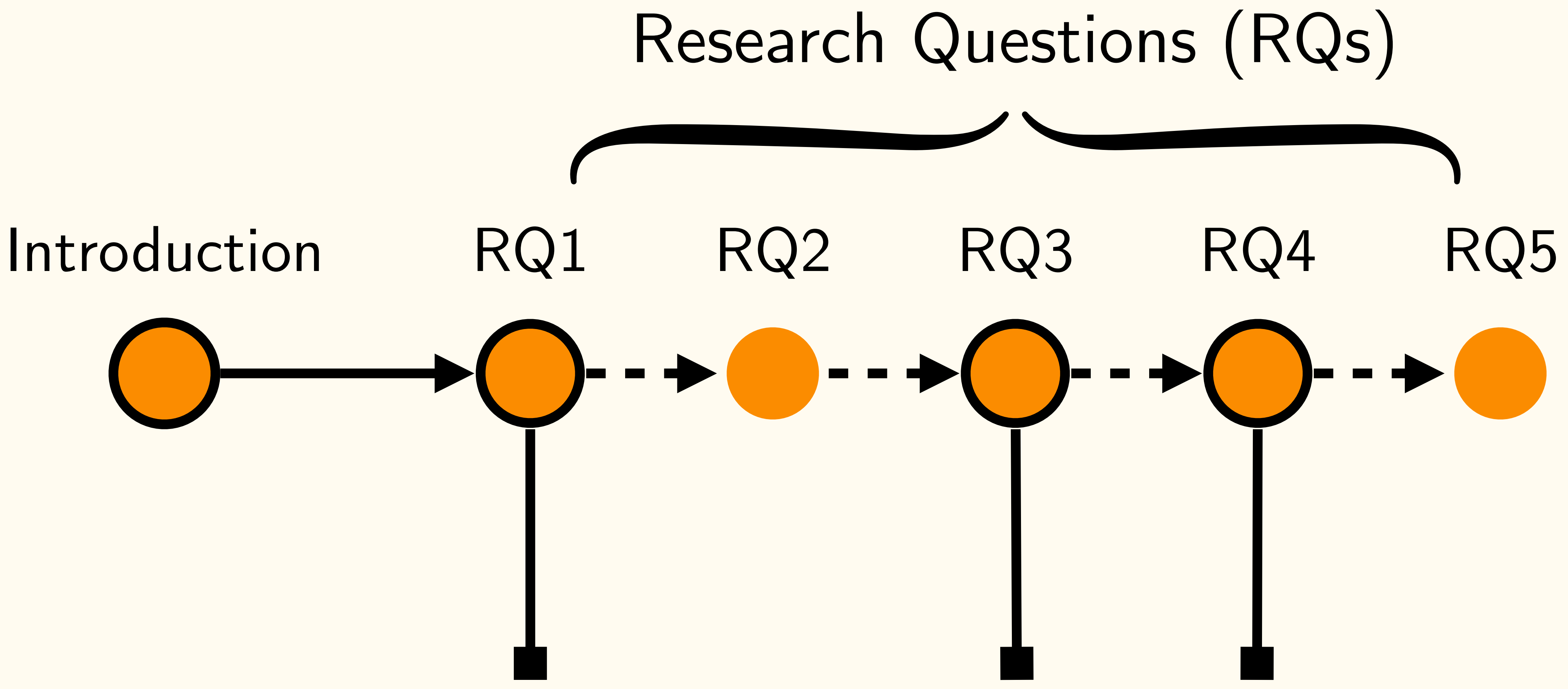
Table of Contents

Table of Contents



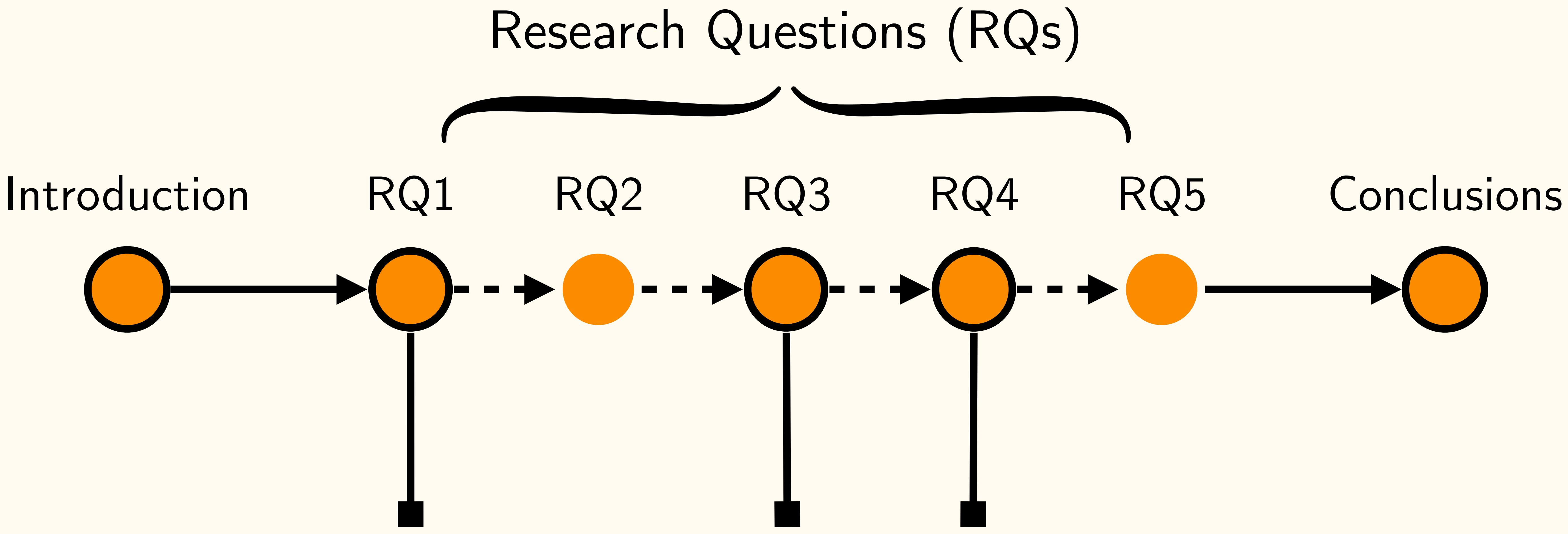
- **Introduction**:
  - Problem
  - Motivations
  - Goals
- **Breadth**: Overview of thesis contributions

Table of Contents



- **Introduction**:
  - Problem
  - Motivations
  - Goals
- **Breadth**: Overview of thesis contributions
- **Depth**:
  - RQ1. Transferability of speech representations.
  - RQ3. Pre-training domain analysis.
  - RQ4. Fine-tuning analysis.

Table of Contents



- **Introduction**:
  - Problem
  - Motivations
  - Goals
- **Breadth**: Overview of thesis contributions
- **Depth**:
  - RQ1. Transferability of speech representations.
  - RQ3. Pre-training domain analysis.
  - RQ4. Fine-tuning analysis.
- **Conclusions**

Bioacoustics

AI for Non-Human Animal Vocal Communication

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

Bioacoustics

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

Bioacoustics

What:

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

Bioacoustics

What:

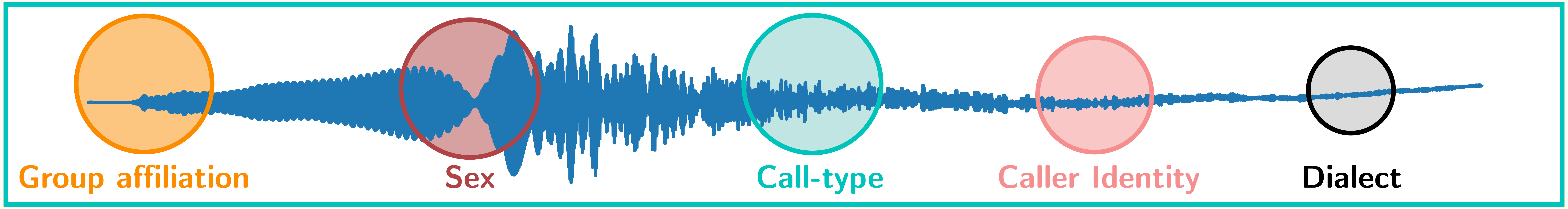
- Study of animal sounds.

Animal vocalization (Marmoset's *twitter*)

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.

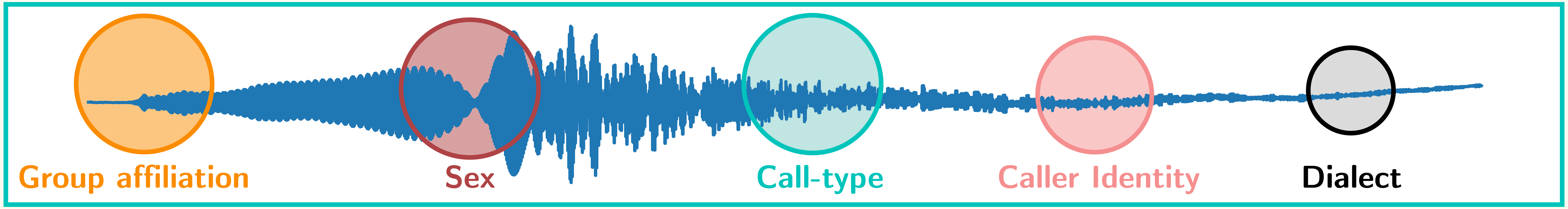


Animal vocalization (Marmoset's *twitter*)

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to ‘decode’ animal calls to gain insights into their vocal communication.



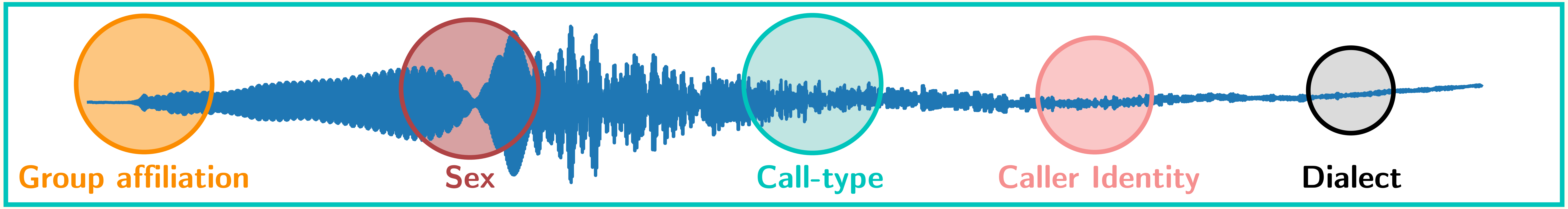
Animal vocalization (Marmoset’s *twitter*)

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to 'decode' animal calls to gain insights into their vocal communication.

Why:



Animal vocalization (Marmoset's *twitter*)

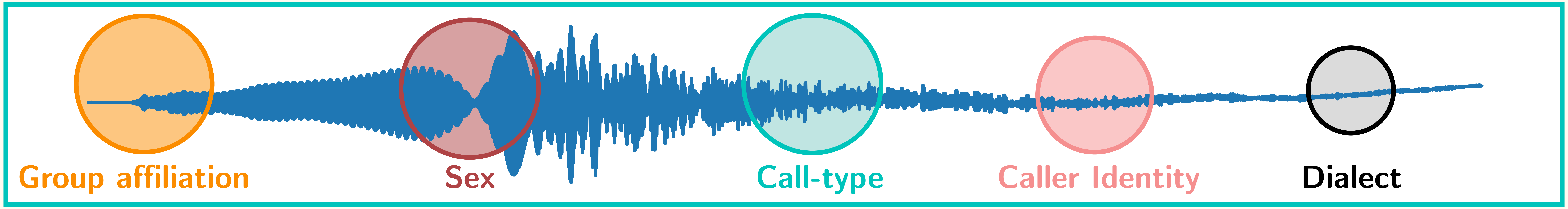
Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to ‘decode’ animal calls to gain insights into their vocal communication.

Why:

Applied:



Animal vocalization (Marmoset’s *twitter*)

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to 'decode' animal calls to gain insights into their vocal communication.

Earth Species Project

Why:

Applied:

- Conservation and biodiversity monitoring.

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to 'decode' animal calls to gain insights into their vocal communication.

Why:

Applied:

- Conservation and biodiversity monitoring.
- Develop tools to support biologists, linguists, and ethologists in their research.

Earth Species Project

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to ‘decode’ animal calls to gain insights into their vocal communication.

Earth Species Project

Why:

Applied:

- Conservation and biodiversity monitoring.
- Develop tools to support biologists, linguists, and ethologists in their research.

Fundamental:

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to ‘decode’ animal calls to gain insights into their vocal communication.

Earth Species Project

nccr evolving language

Why:

Applied:

- Conservation and biodiversity monitoring.
- Develop tools to support biologists, linguists, and ethologists in their research.

Fundamental:

- Evolutionary origins of language.

Bioacoustics

What:

- Study of animal sounds.
- Vocalizations encode a range of information.
- Bioacoustics aims to ‘decode’ animal calls to gain insights into their vocal communication.

Earth Species Project

Why:

Applied:

- Conservation and biodiversity monitoring.
- Develop tools to support biologists, linguists, and ethologists in their research.

Fundamental:

- Evolutionary origins of language.
- Deepen our understanding of communication in the non-human natural world.

Bioacoustics Studies and Challenges

Bioacoustics Studies and Challenges

How ?

Bioacoustics Studies and Challenges

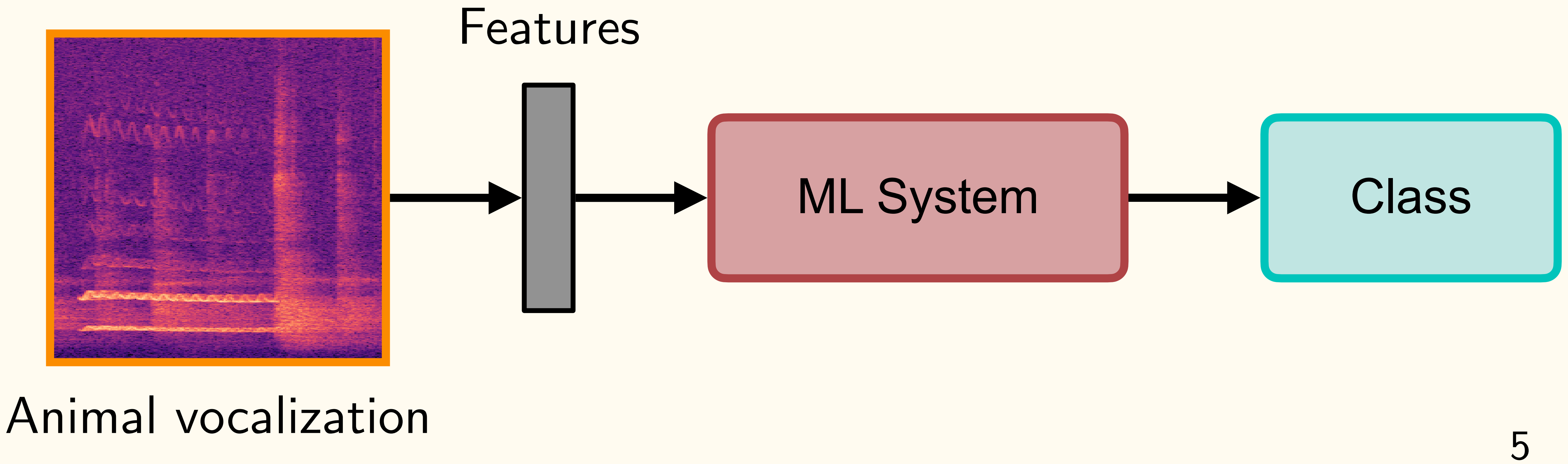
How ?

- Studies took inspiration from human speech feature representations:

Bioacoustics Studies and Challenges

How ?

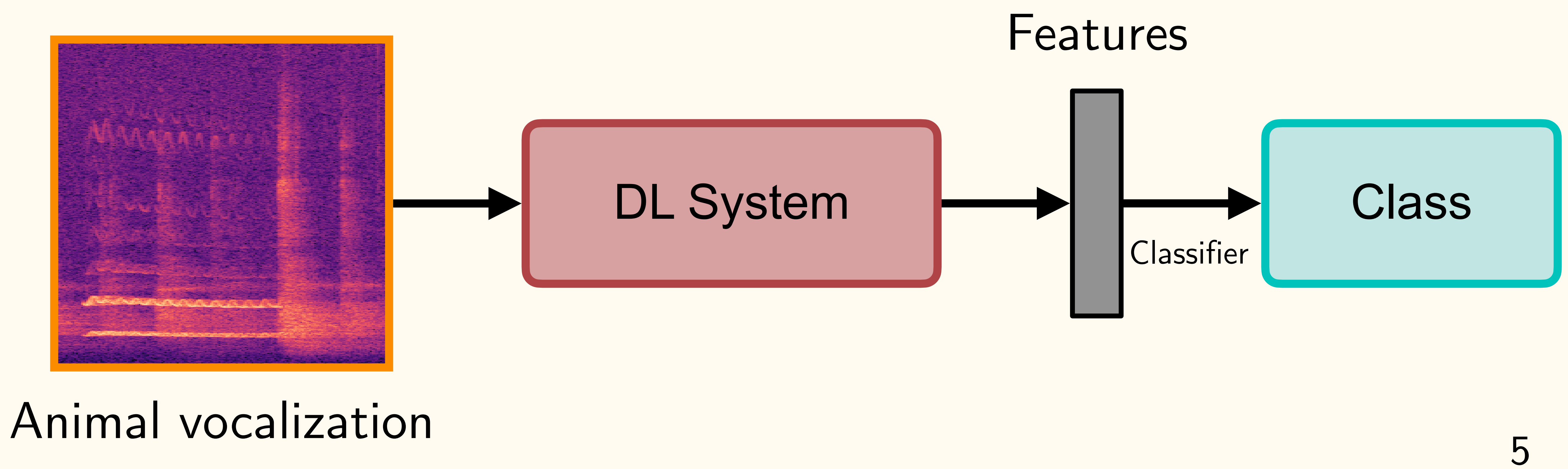
- Studies took inspiration from human speech feature representations:
  - LP coefficients, MFCCs.
  - HCTSA, C22.



Bioacoustics Studies and Challenges

How ?

- Studies took inspiration from human speech feature representations:
  - LP coefficients, MFCCs.
  - HCTSA, C22.
- Re-purposing DL architectures originally developed for speech tasks for bioacoustics has shown some success[1,2].



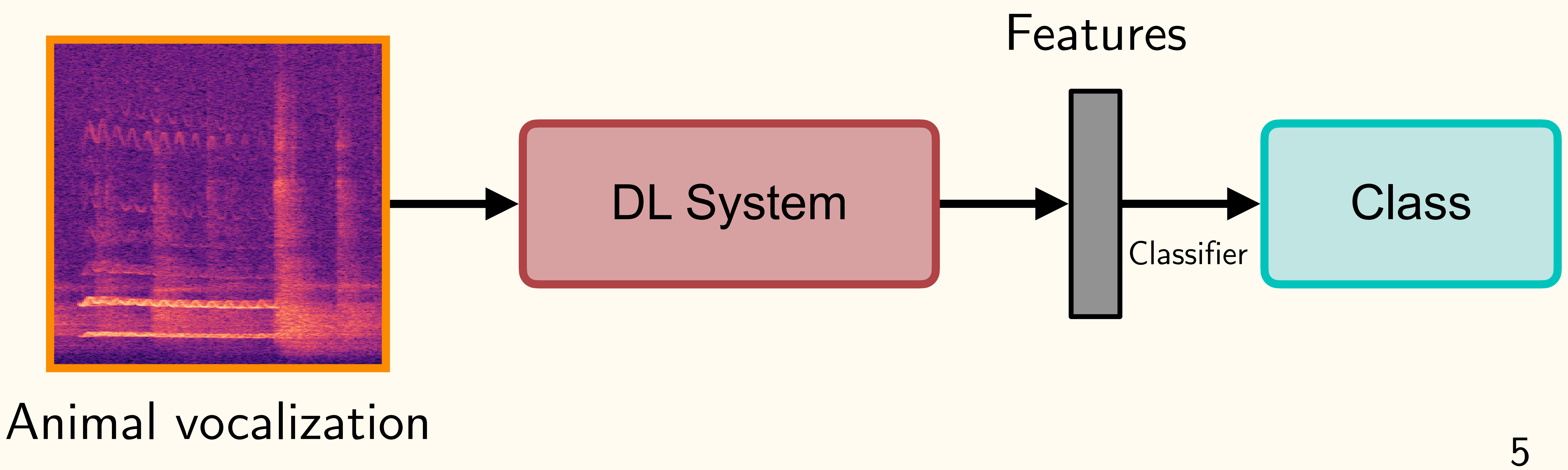
[1] Stowell et al., 2019; Sainburg, Thielk, and Gentner, 2020.
[2] Y.-J. Zhang et al., 2018; E. Coffey et al., 2019; Bergler et al., 2019.

Bioacoustics Studies and Challenges

How ?

- Studies took inspiration from human speech feature representations:
  - LP coefficients, MFCCs.
  - HCTSA, C22.
- Re-purposing DL architectures originally developed for speech tasks for bioacoustics has shown some success[1,2].

Challenges:



[1] Stowell et al., 2019; Sainburg, Thielk, and Gentner, 2020.
[2] Y.-J. Zhang et al., 2018; E. Coffey et al., 2019; Bergler et al., 2019.

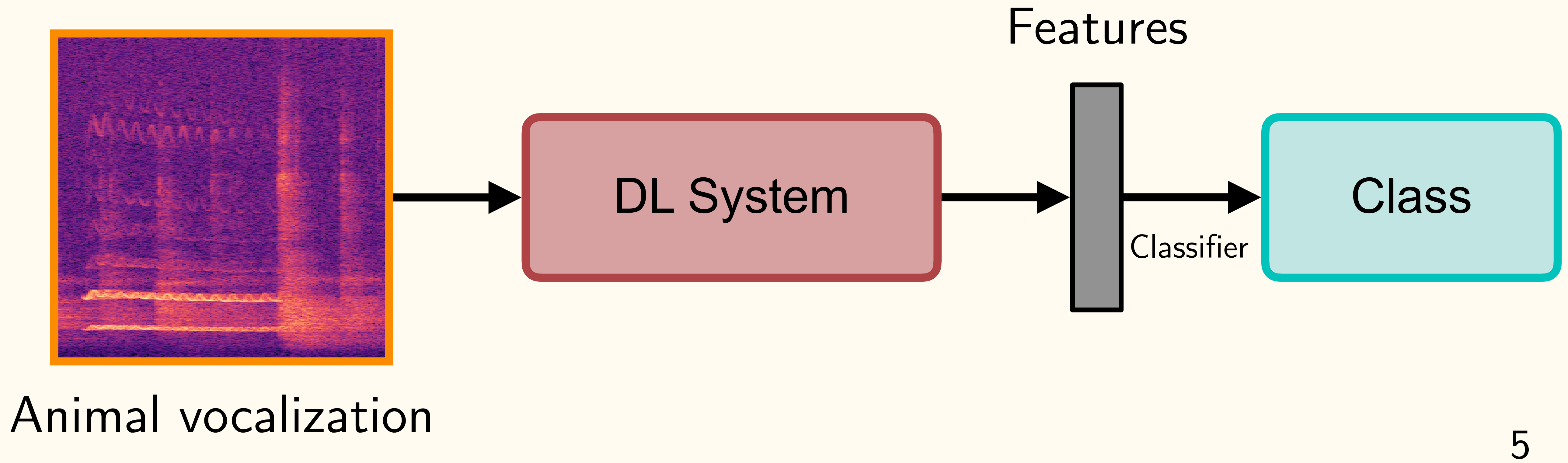
Bioacoustics Studies and Challenges

How ?

- Studies took inspiration from human speech feature representations:
  - LP coefficients, MFCCs.
  - HCTSA, C22.
- Re-purposing DL architectures originally developed for speech tasks for bioacoustics has shown some success[1,2].

Challenges:

- Limited understanding of animal vocal communication.



[1] Stowell et al., 2019; Sainburg, Thielk, and Gentner, 2020.
[2] Y.-J. Zhang et al., 2018; E. Coffey et al., 2019; Bergler et al., 2019.

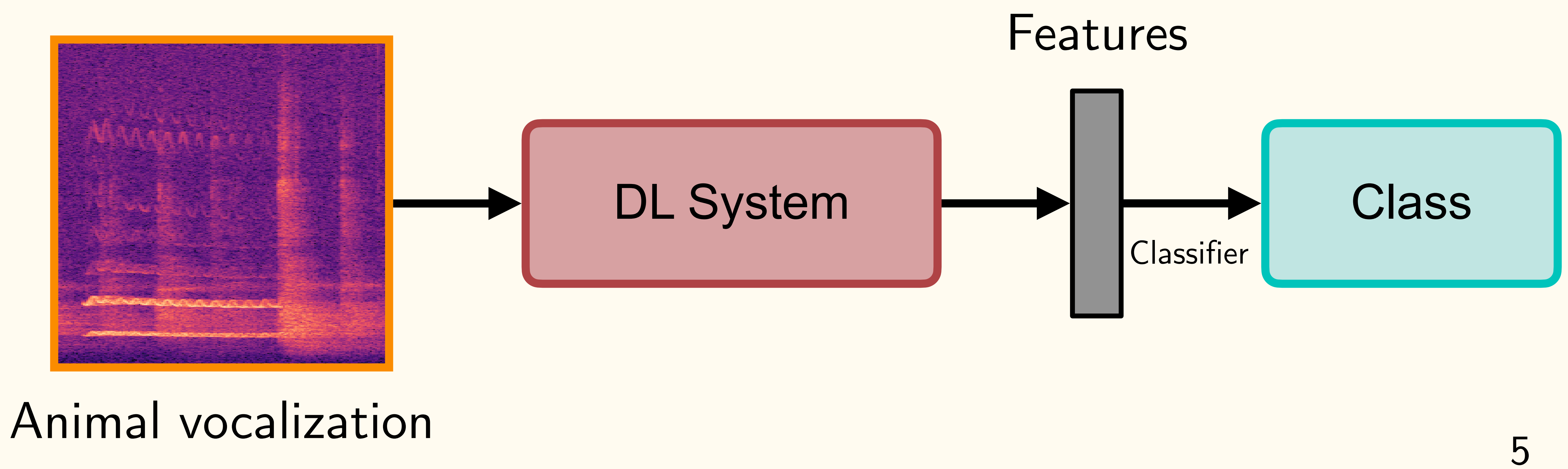
Bioacoustics Studies and Challenges

How ?

- Studies took inspiration from human speech feature representations:
  - LP coefficients, MFCCs.
  - HCTSA, C22.
- Re-purposing DL architectures originally developed for speech tasks for bioacoustics has shown some success[1,2].

Challenges:

- Limited understanding of animal vocal communication.
- Lack of prior knowledge on relevant acoustic information for animal calls.



[1] Stowell et al., 2019; Sainburg, Thielk, and Gentner, 2020.
[2] Y.-J. Zhang et al., 2018; E. Coffey et al., 2019; Bergler et al., 2019.

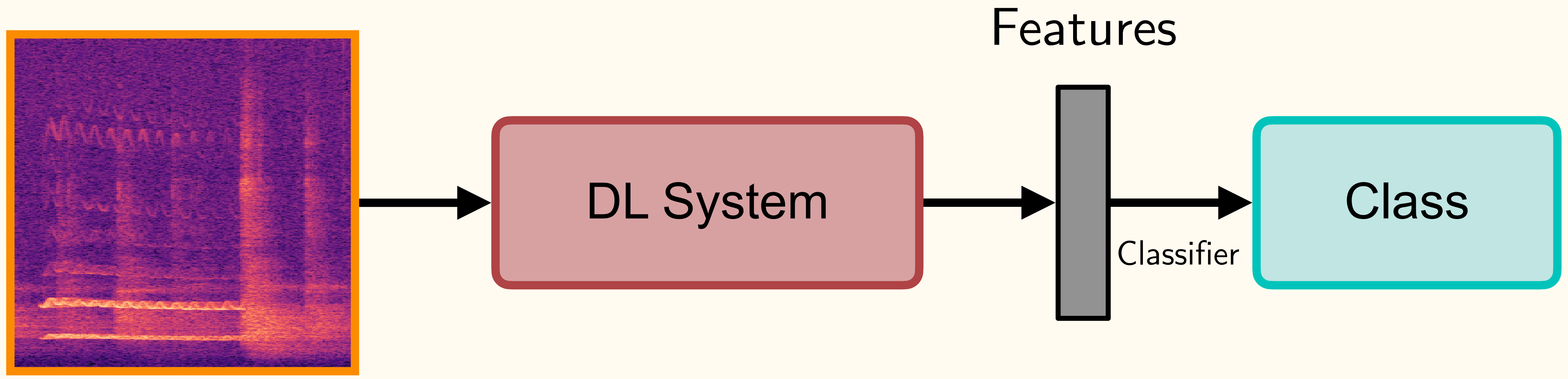
Bioacoustics Studies and Challenges

How ?

- Studies took inspiration from human speech feature representations:
  - LP coefficients, MFCCs.
  - HCTSA, C22.
- Re-purposing DL architectures originally developed for speech tasks for bioacoustics has shown some success[1,2].

Challenges:

- Limited understanding of animal vocal communication.
- Lack of prior knowledge on relevant acoustic information for animal calls.
- Under-resourced labeled data.



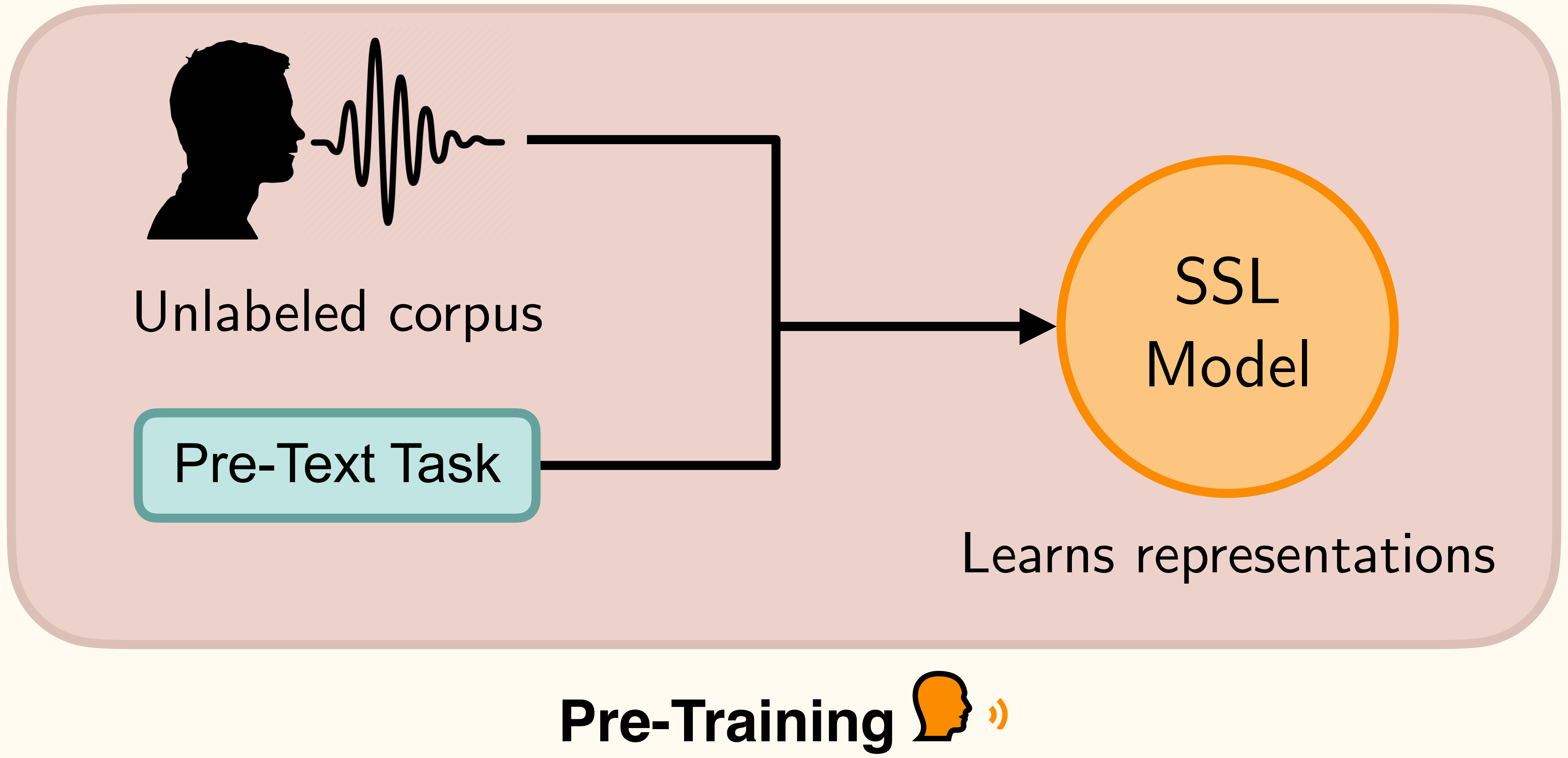
Animal vocalization

[1] Stowell et al., 2019; Sainburg, Thielk, and Gentner, 2020.
[2] Y.-J. Zhang et al., 2018; E. Coffey et al., 2019; Bergler et al., 2019.

Self-Supervised Learning Models (SSLs)

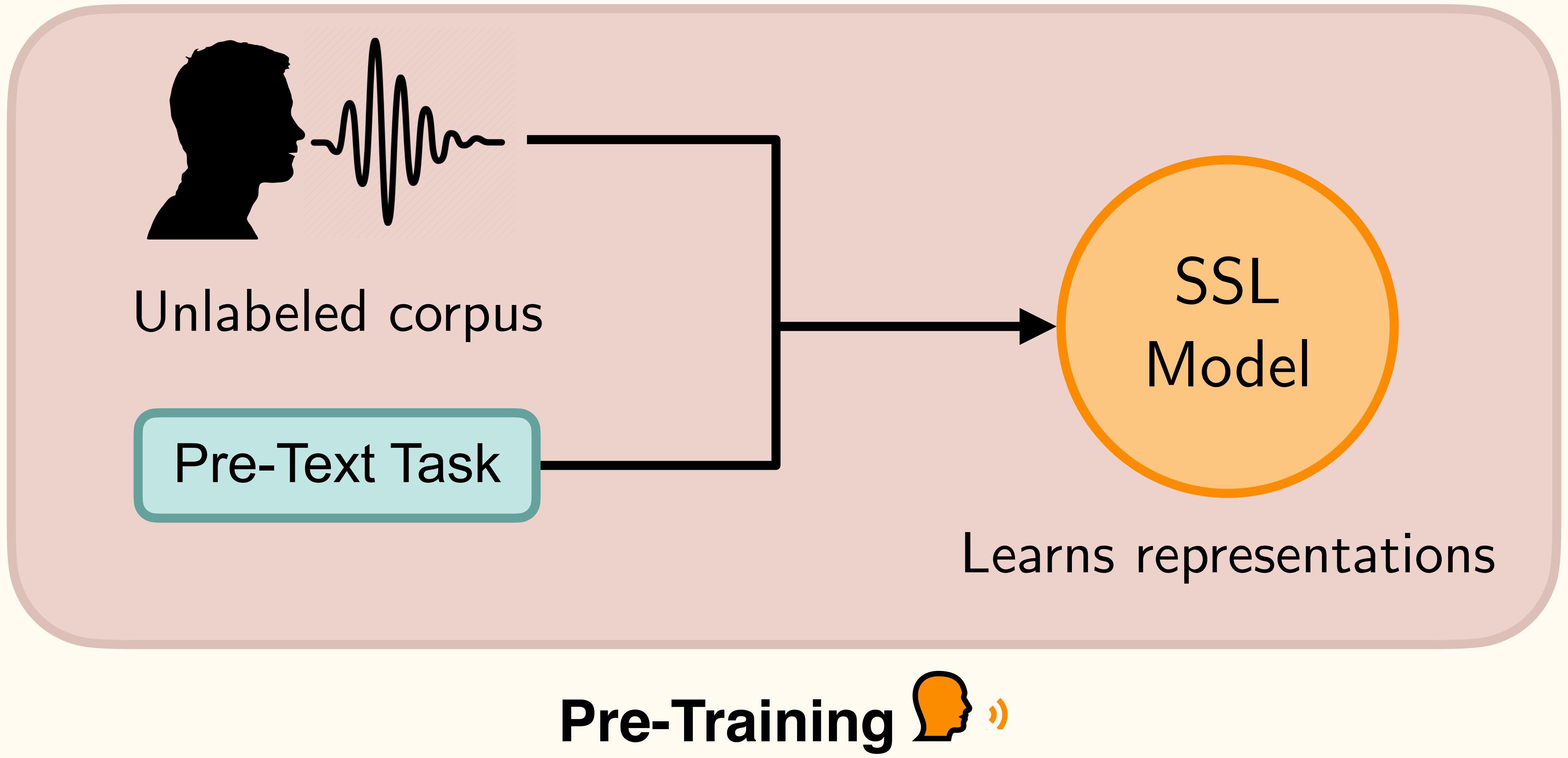
Self-Supervised Learning Models (SSLs)

- SSL models learn representations directly from the raw acoustic input.



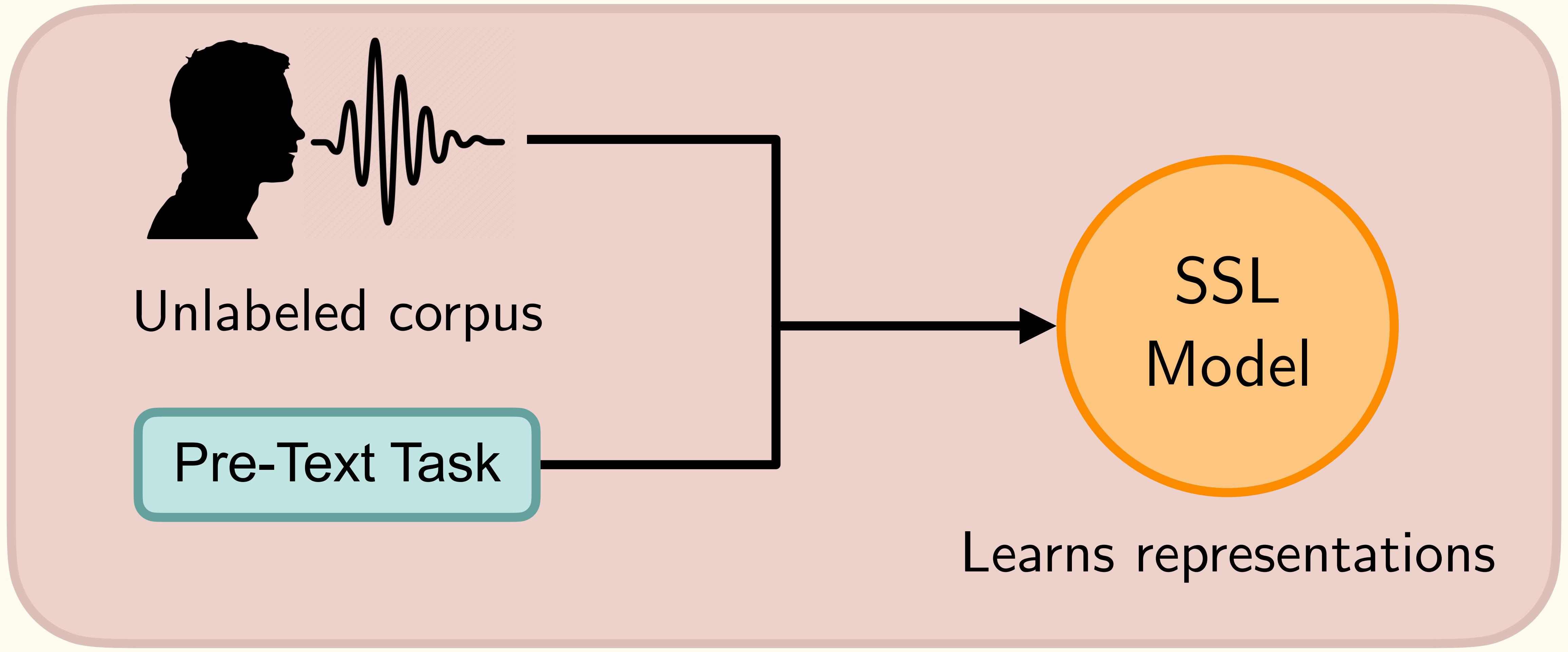
Self-Supervised Learning Models (SSLs)

- SSL models learn representations directly from the raw acoustic input.
- Can leverage unlabelled data and learn general representations.



Self-Supervised Learning Models (SSLs)

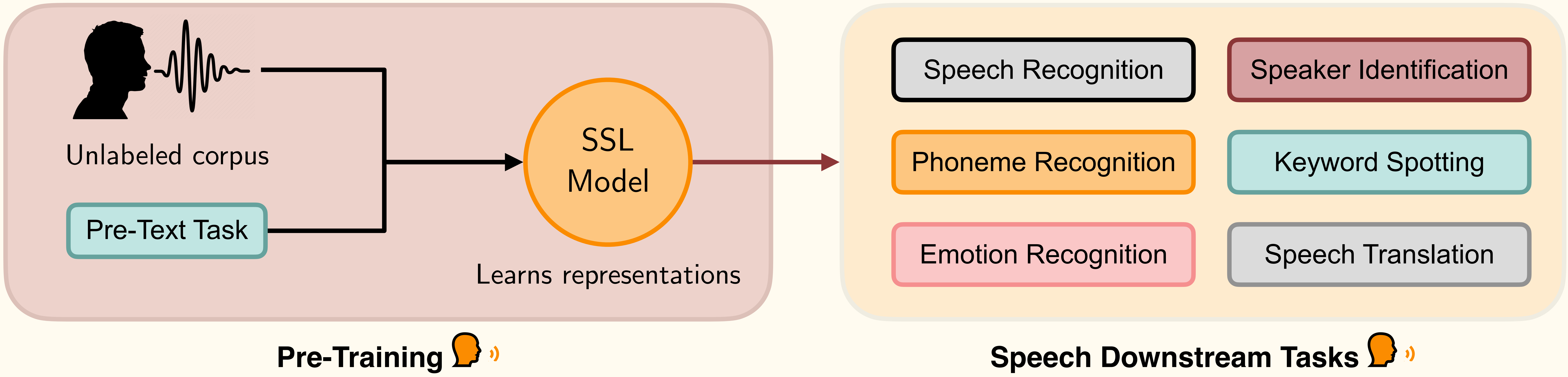
- SSL models learn representations directly from the raw acoustic input.
- Can leverage unlabelled data and learn general representations.
- Has successfully shown state-of-the-art results on speech downstream tasks.



Pre-Training

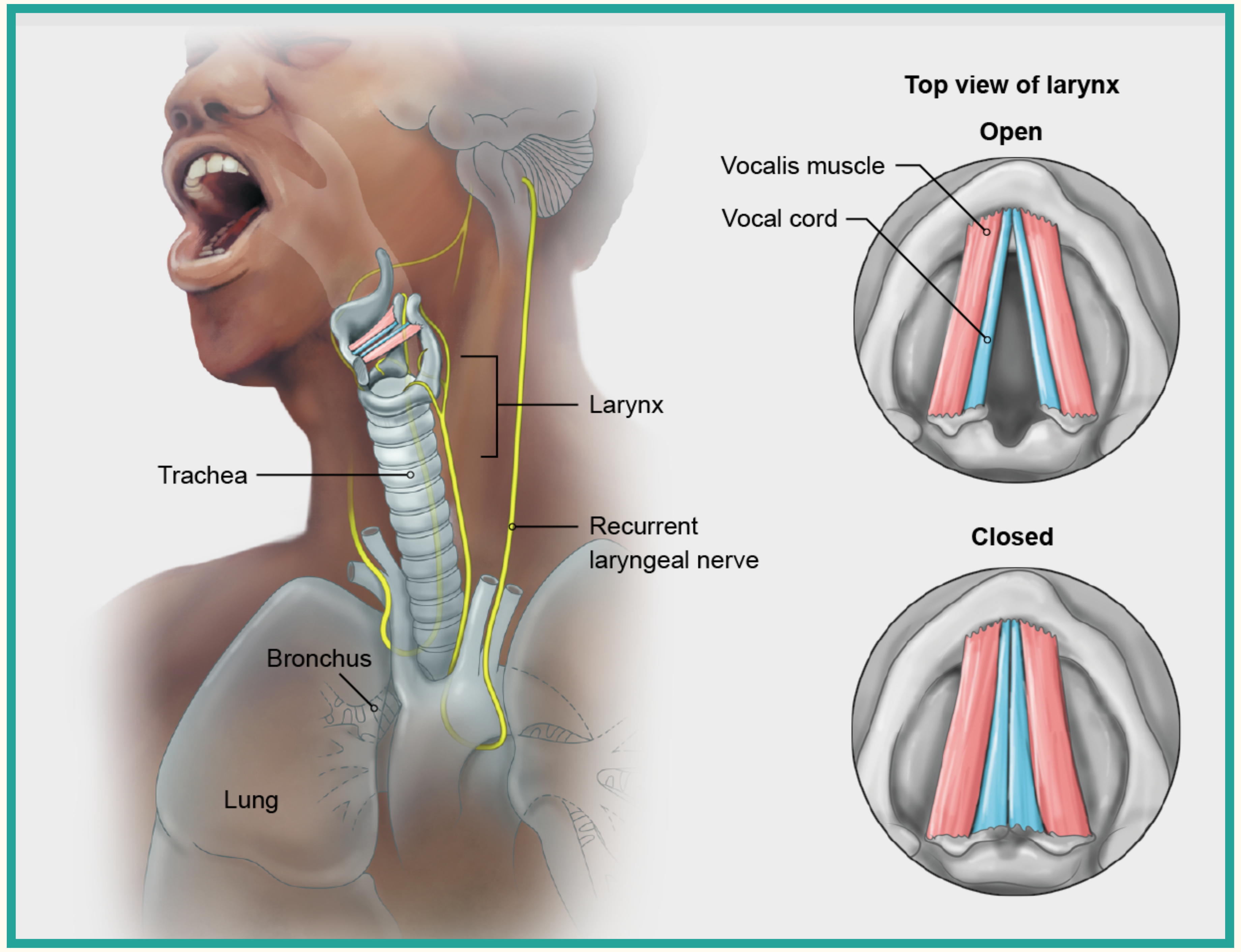
Self-Supervised Learning Models (SSLs)

- SSL models learn representations directly from the raw acoustic input.
- Can leverage unlabelled data and learn general representations.
- Has successfully shown state-of-the-art results on speech downstream tasks.



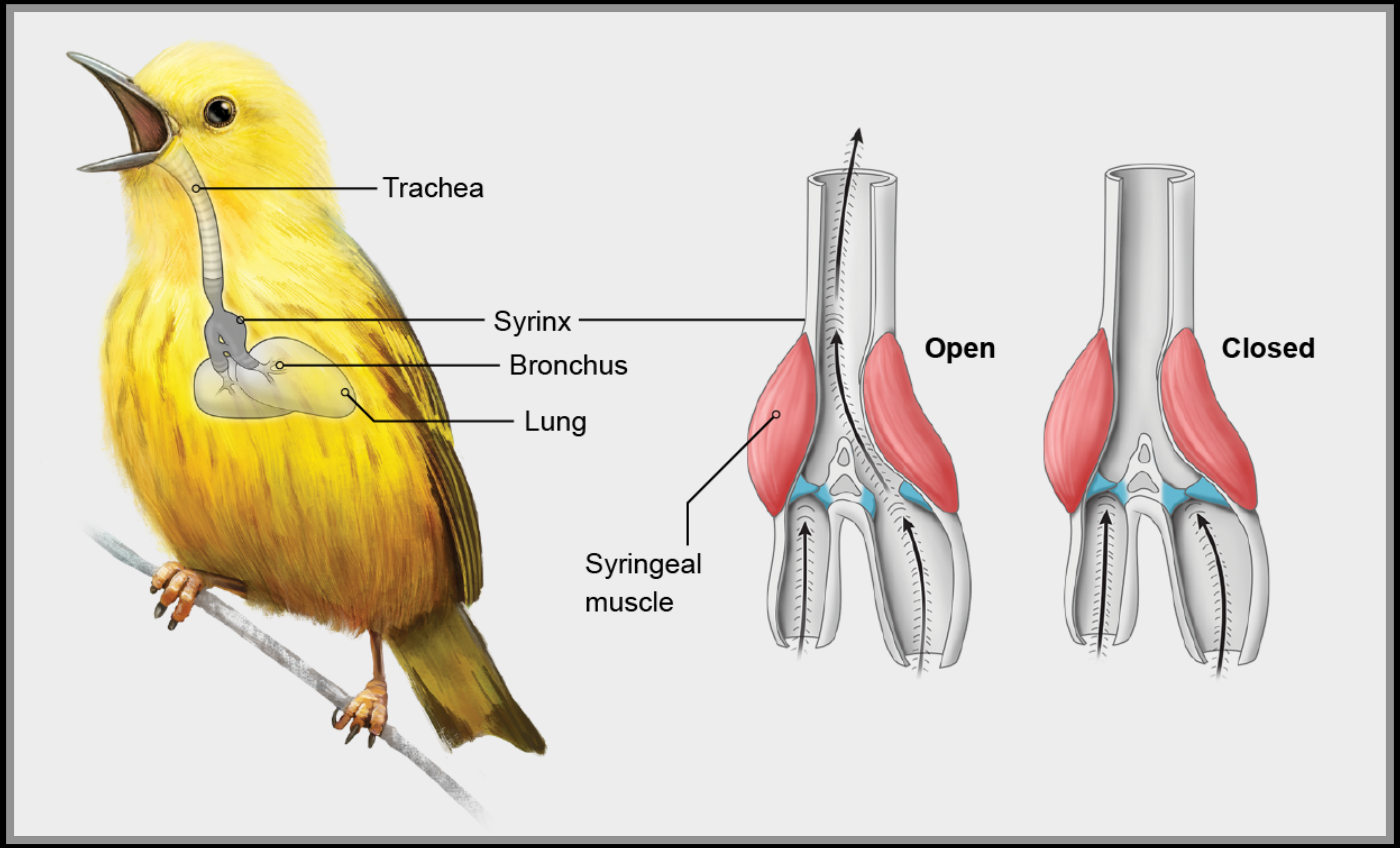
Humans

Larynx



Birds

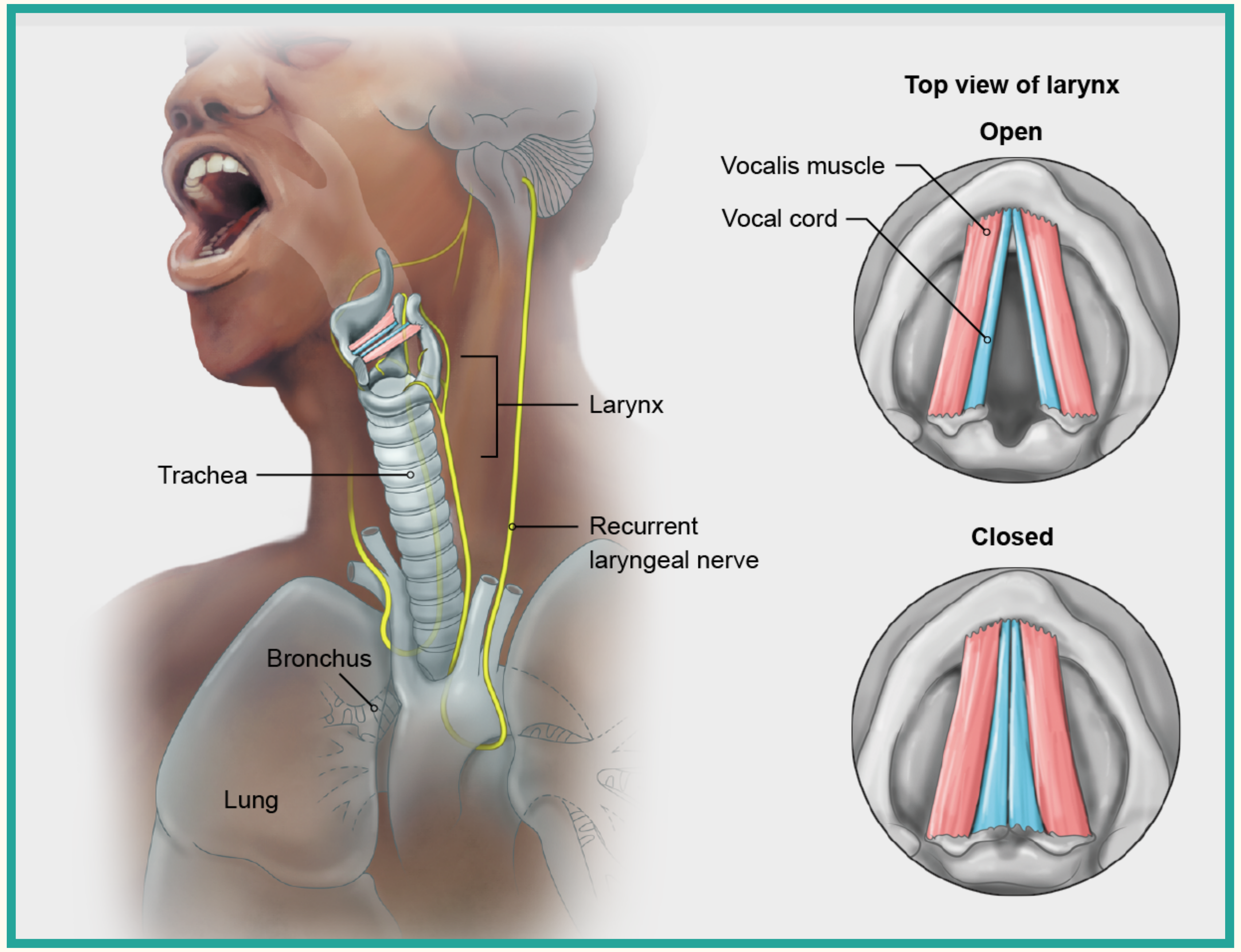
Syrinx



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

Humans

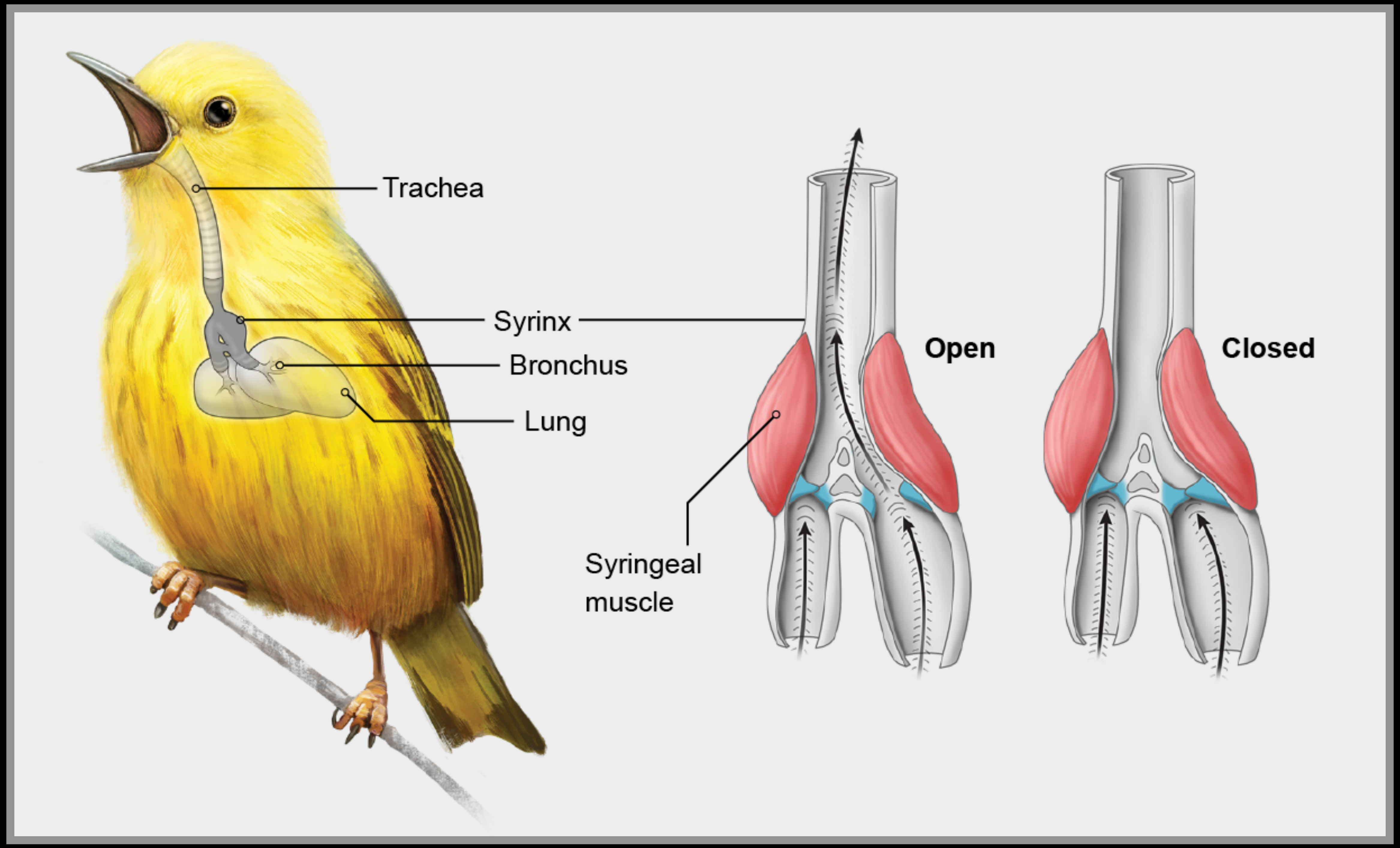
Larynx



Commonality: a production (and perception) system.

Birds

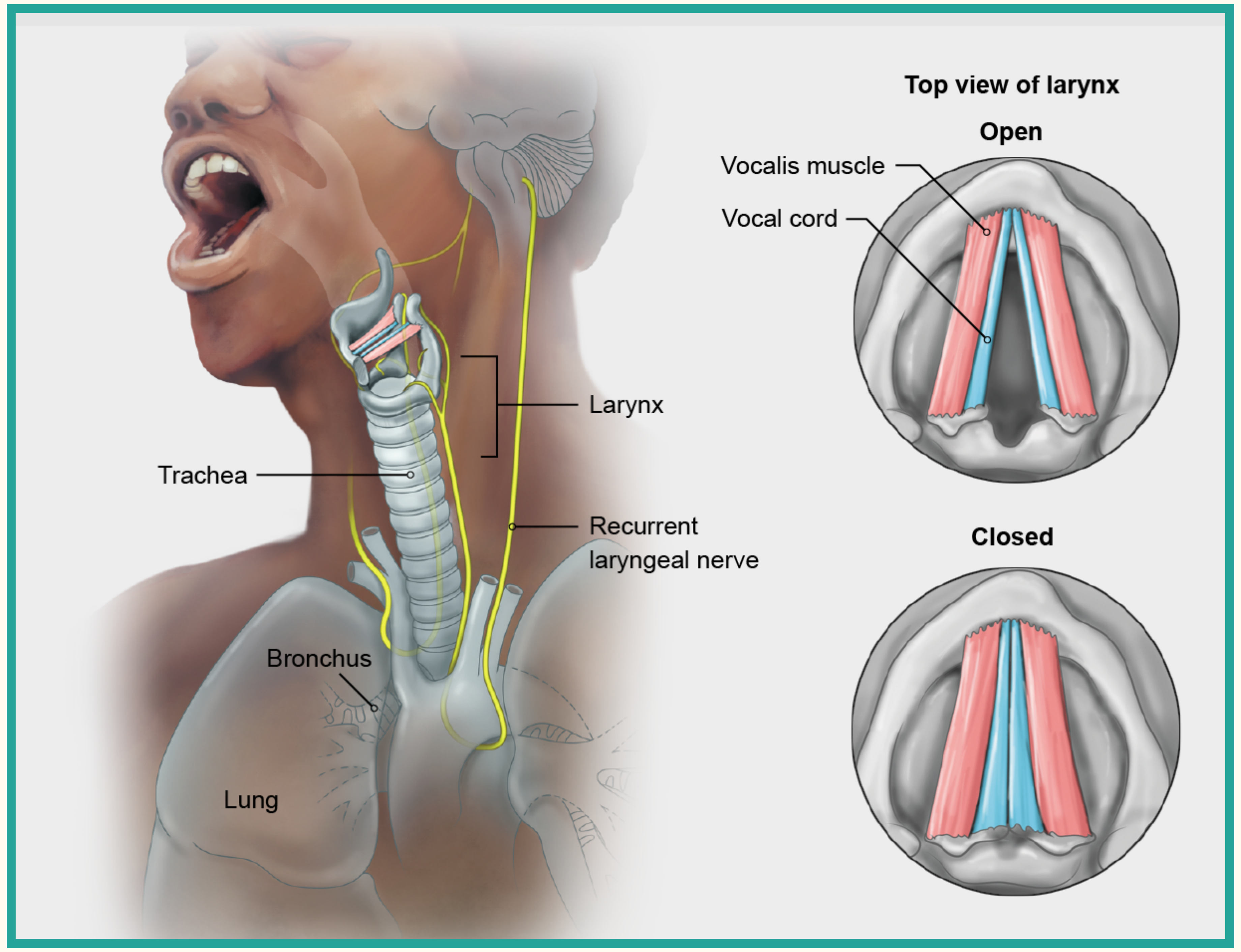
Syrinx



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

Humans

Larynx

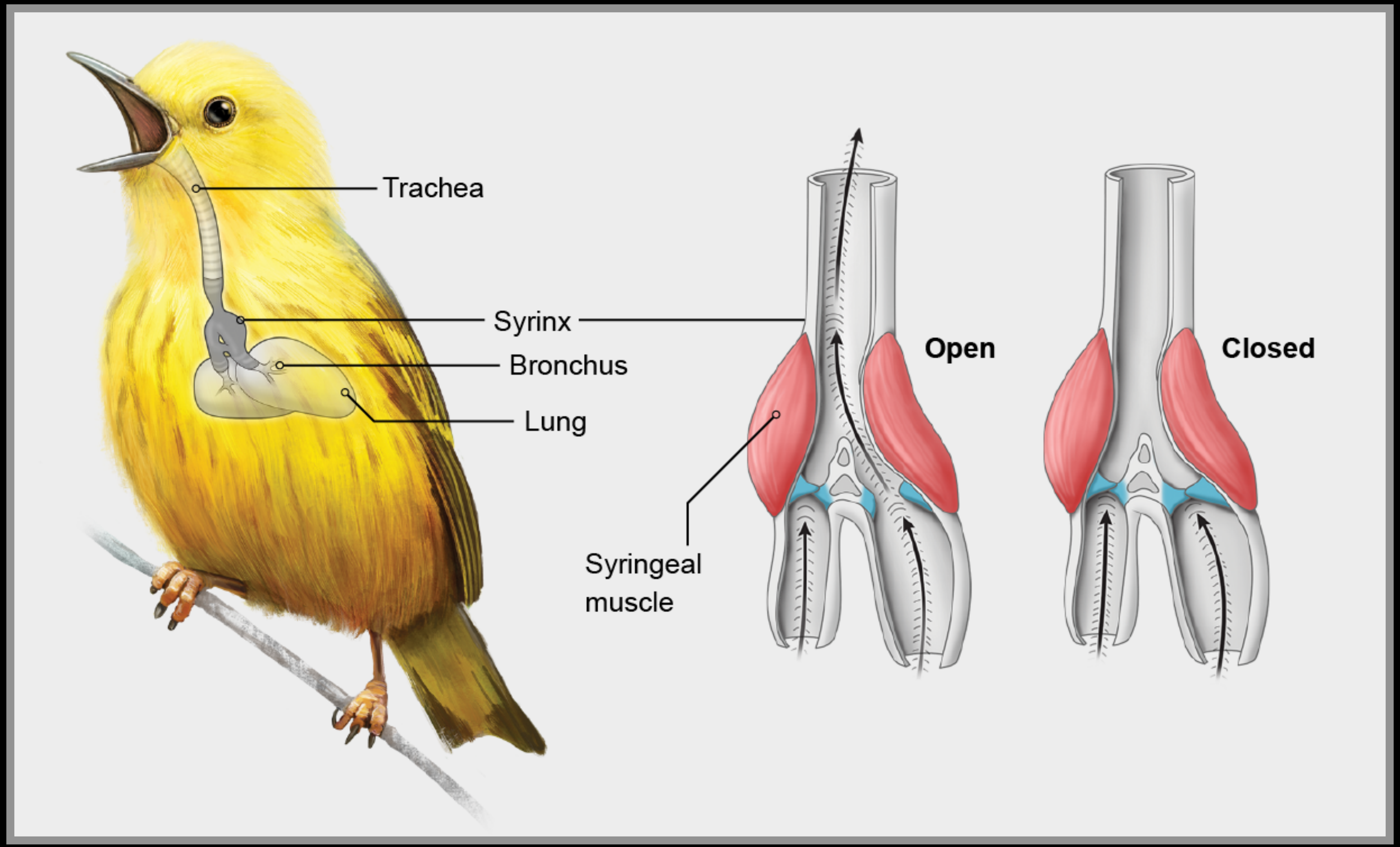


Commonality: a production (and perception) system.

Enables: communication through *structured* acoustic signals.

Birds

Syrinx



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

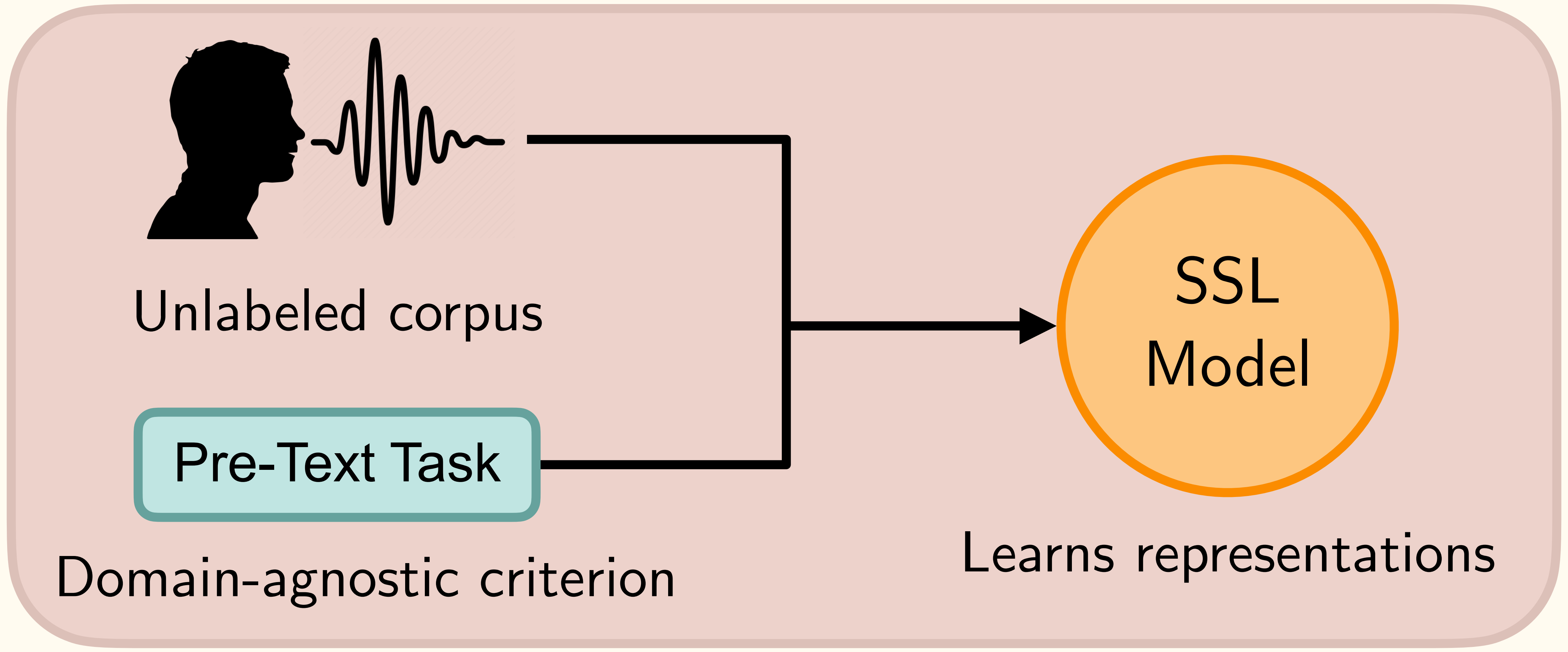
Thesis Hypothesis

Thesis Hypothesis

- Human and animal vocalizations are inherently **structured** signals that encode meaning.

Thesis Hypothesis

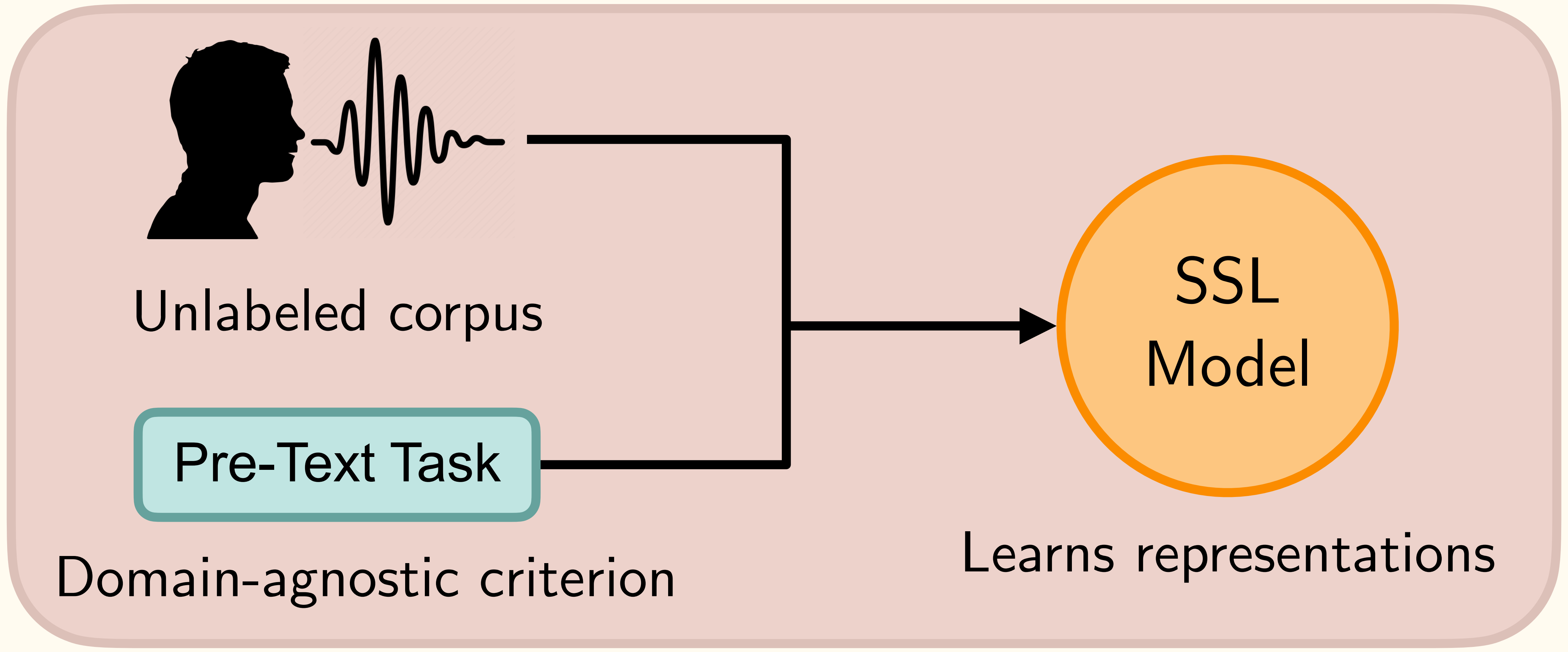
- Human and animal vocalizations are inherently **structured** signals that encode meaning.
- SSLs do *not explicitly incorporate prior knowledge* about underlying production systems.



Pre-Training

Thesis Hypothesis

- Human and animal vocalizations are inherently **structured** signals that encode meaning.
- SSLs do *not explicitly incorporate prior knowledge* about underlying production systems.
- Learns to identify the intrinsic **structure** in the acoustic signal.

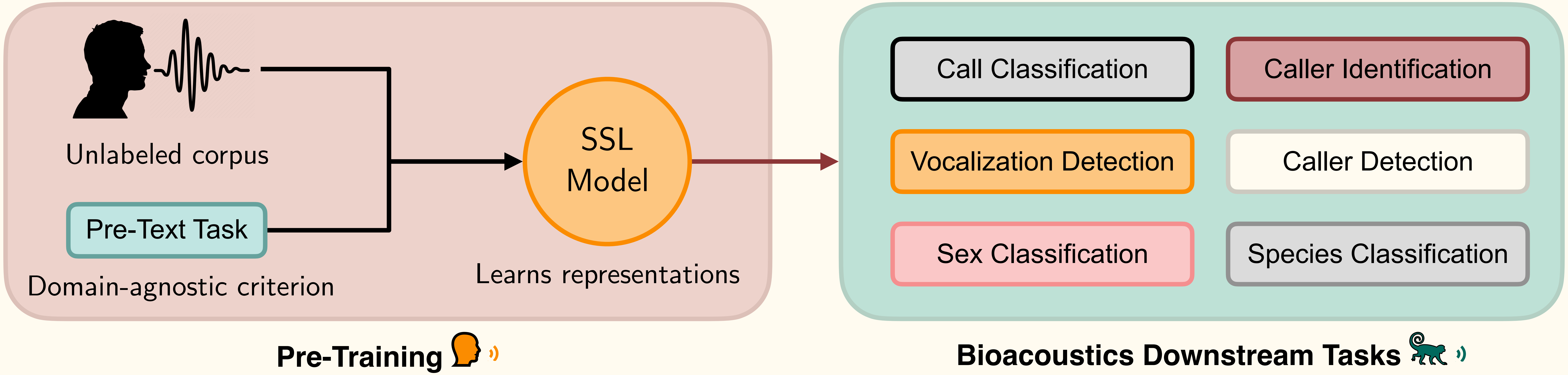


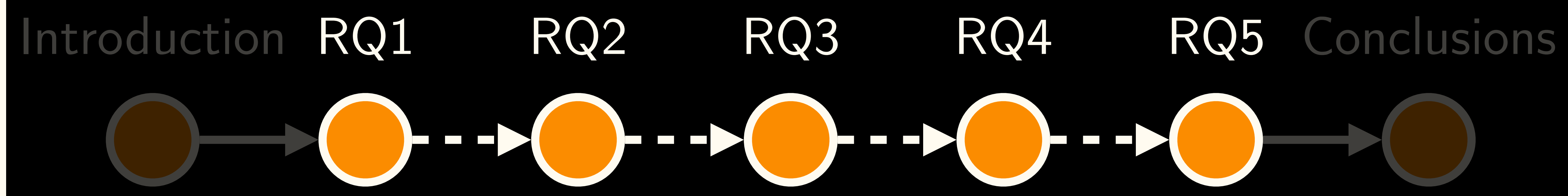
Pre-Training

Thesis Hypothesis

- Human and animal vocalizations are inherently **structured** signals that encode meaning.
- SSLs do *not explicitly incorporate prior knowledge* about underlying production systems.
- Learns to identify the intrinsic **structure** in the acoustic signal.

➡ Hypothesis: speech representations learnt in a SSL framework, can transfer to the bioacoustics domain, and help decode animal vocalizations.





Thesis Contributions

Overview of Thesis
Research Questions (RQs)

RQ1. Transferability
RQ2. Bandwidth
RQ3. Pre-Training Domain
RQ4. Fine-Tuning
RQ5. Sequential Structure

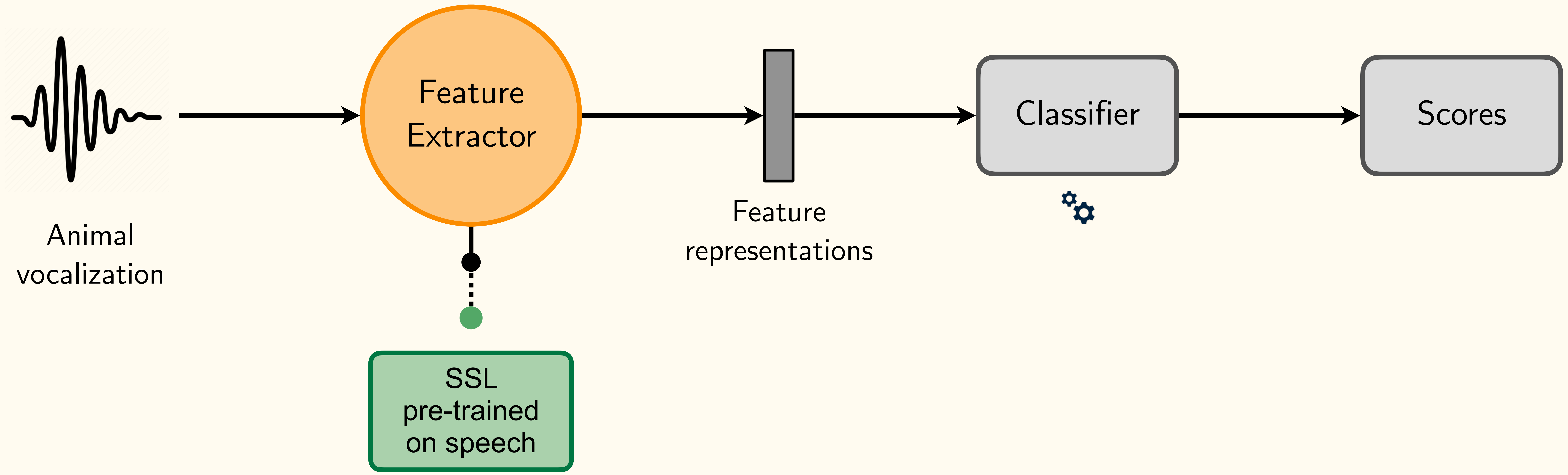
Thesis Contributions 1: Transferability of SSL Representations

Thesis Contributions 1: Transferability of SSL Representations

Can representations learnt from human speech through SSLs be transferred to bioacoustic tasks?

Thesis Contributions 1: Transferability of SSL Representations

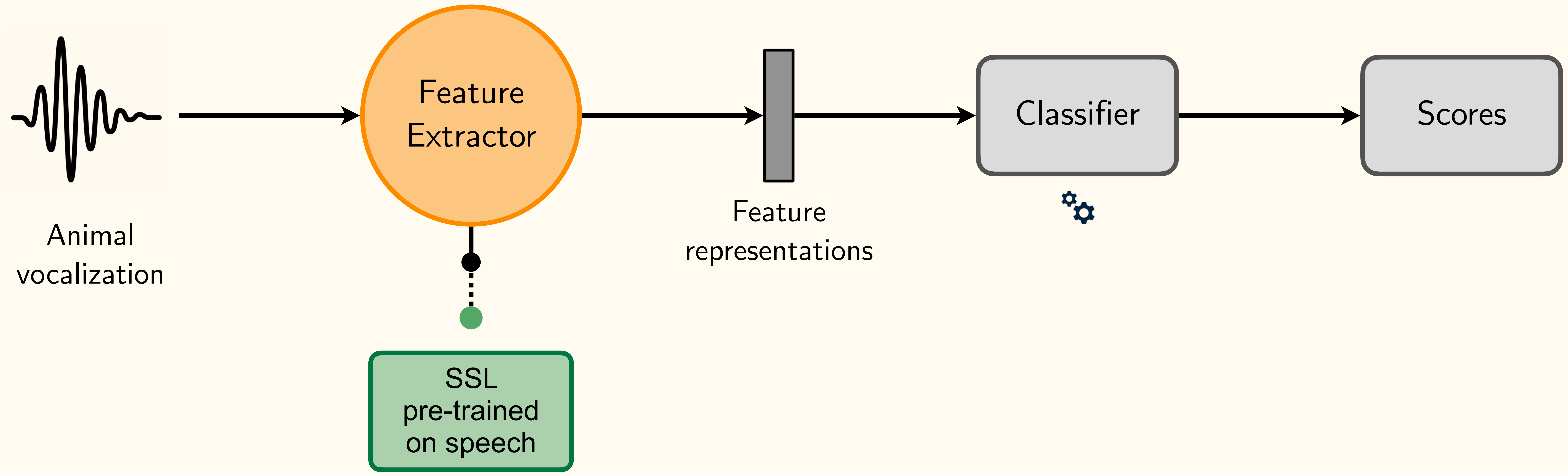
Can representations learnt from human speech through SSLs be transferred to bioacoustic tasks?



Thesis Contributions 1: Transferability of SSL Representations

Can representations learnt from human speech through SSLs be transferred to bioacoustic tasks?

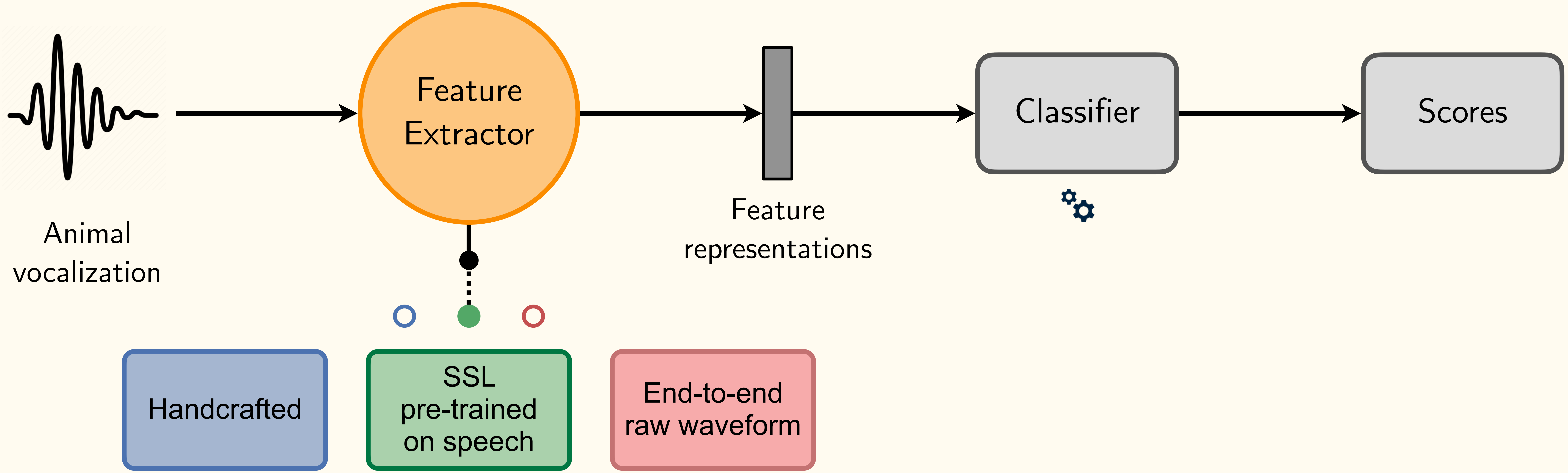
- To what extent?



Thesis Contributions 1: Transferability of SSL Representations

Can representations learnt from human speech through SSLs be transferred to bioacoustic tasks?

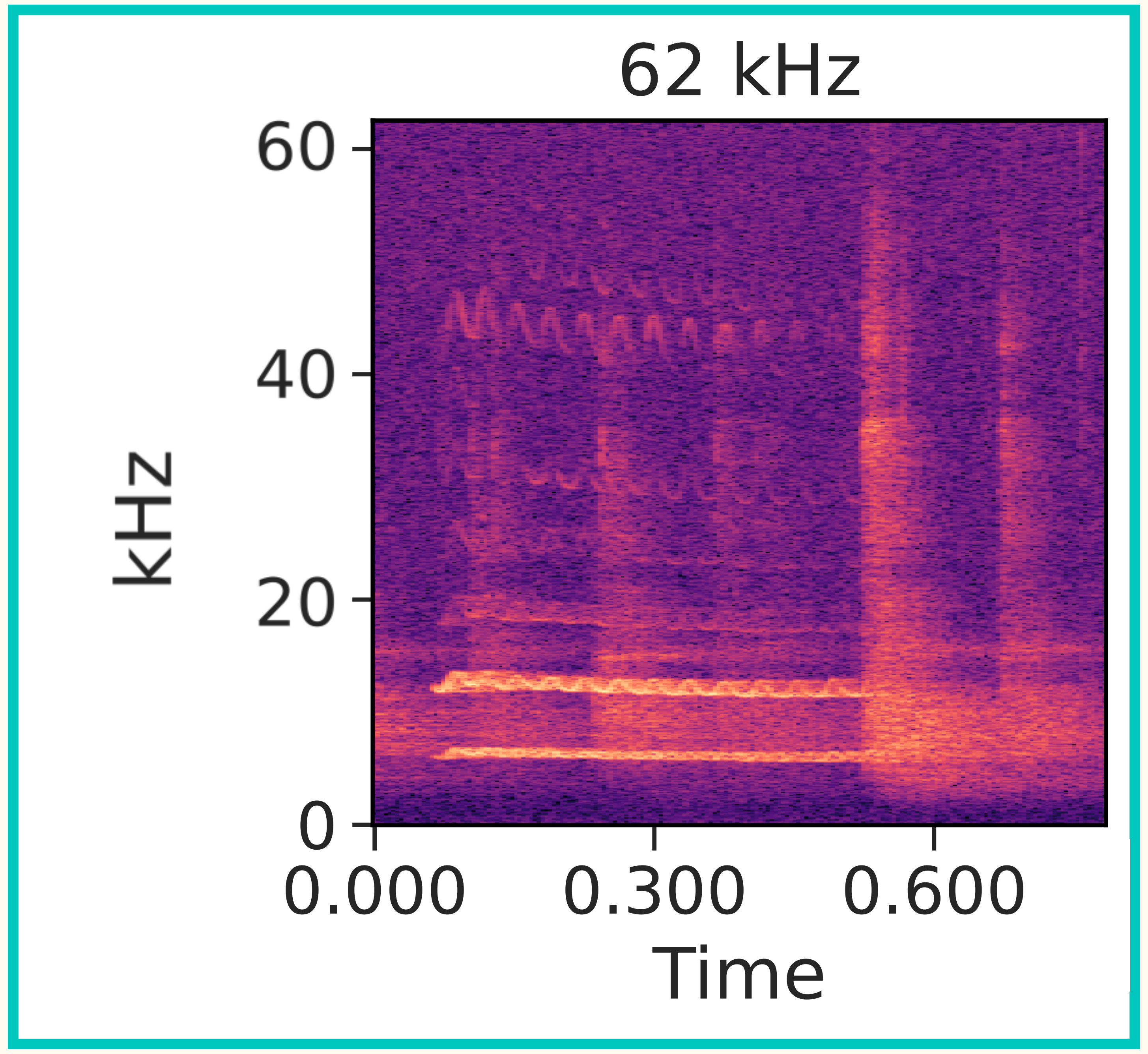
- To what extent?
- How do SSLs features compare to handcrafted features or end-to-end models ?



Thesis Contributions 2: Model Bandwidth Mismatch

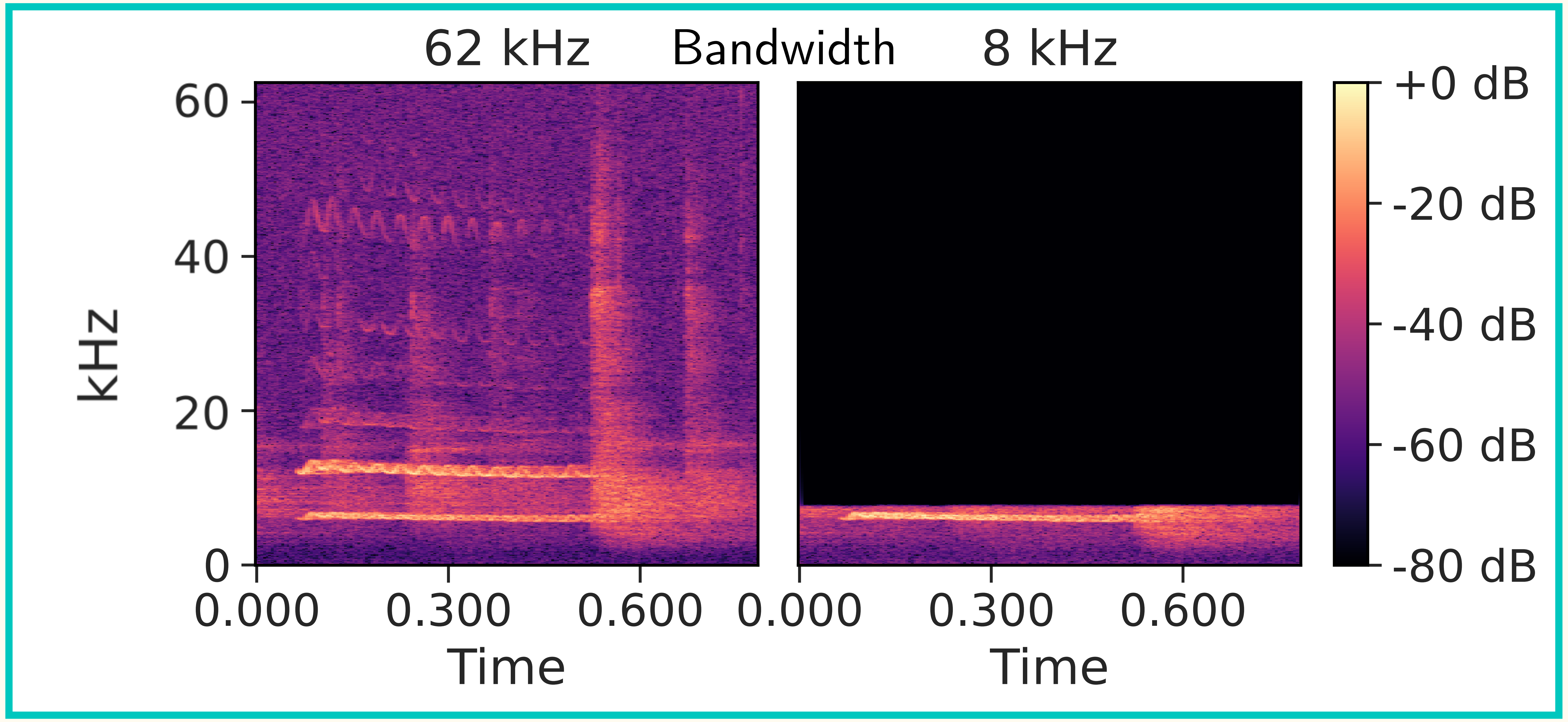
Thesis Contributions 2: Model Bandwidth Mismatch

- Animal vocalizations can go in high frequency ranges compared to human speech.



Thesis Contributions 2: Model Bandwidth Mismatch

- Animal vocalizations can go in high frequency ranges compared to human speech.
- Speech SSLs typically pre-trained at 8 kHz bandwidth.

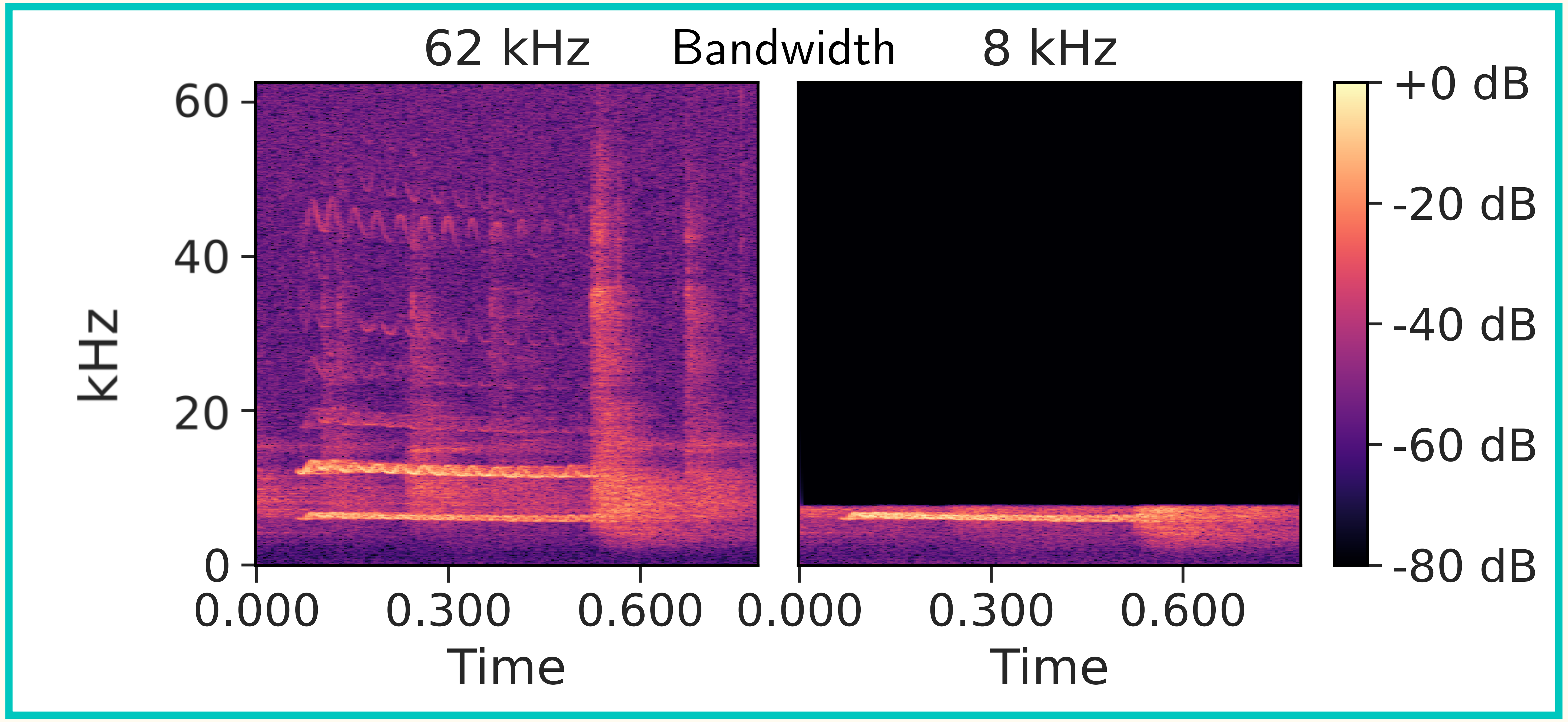


Bandwidth = Sampling Rate / 2

Thesis Contributions 2: Model Bandwidth Mismatch

- Animal vocalizations can go in high frequency ranges compared to human speech.
- Speech SSLs typically pre-trained at 8 kHz bandwidth.

How does this bandwidth mismatch between humans and animals affect this transfer?

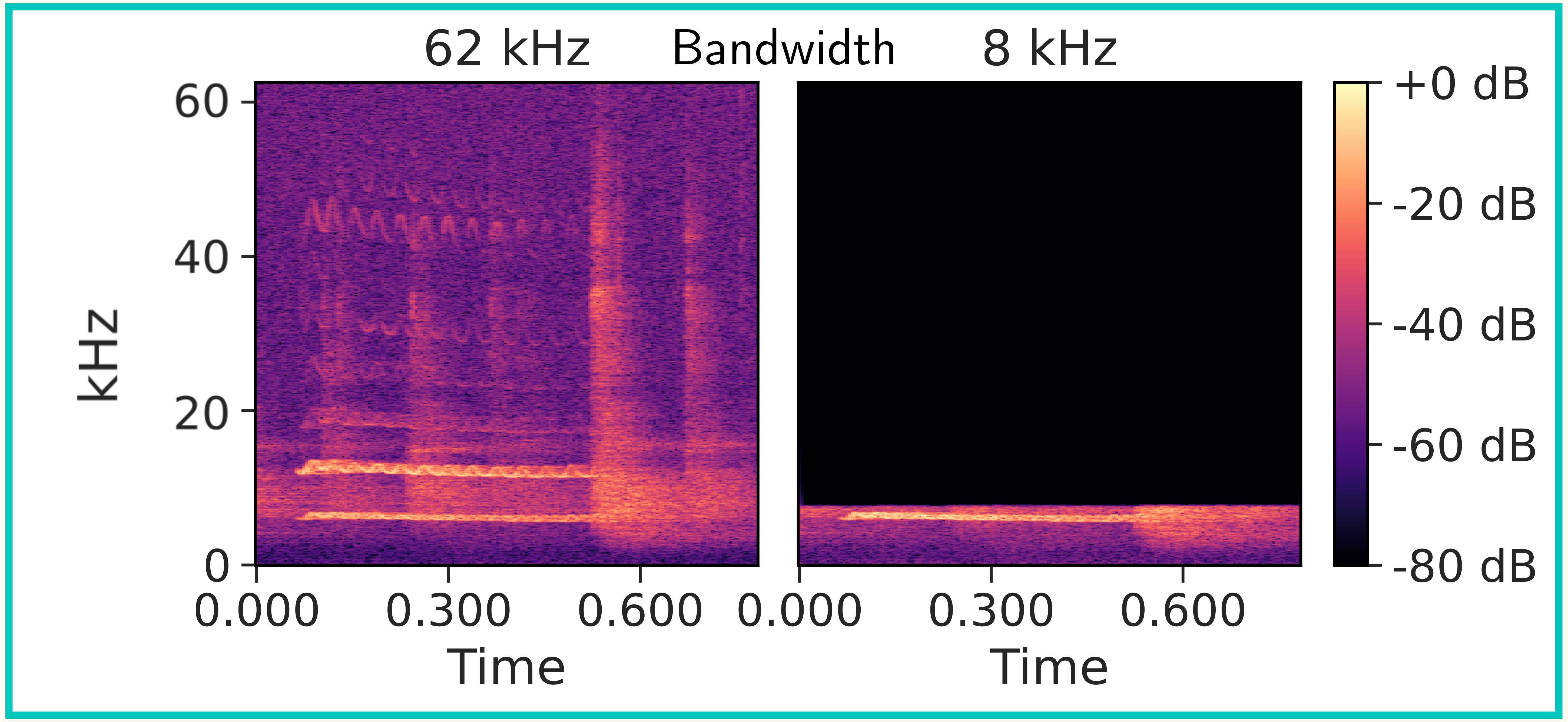


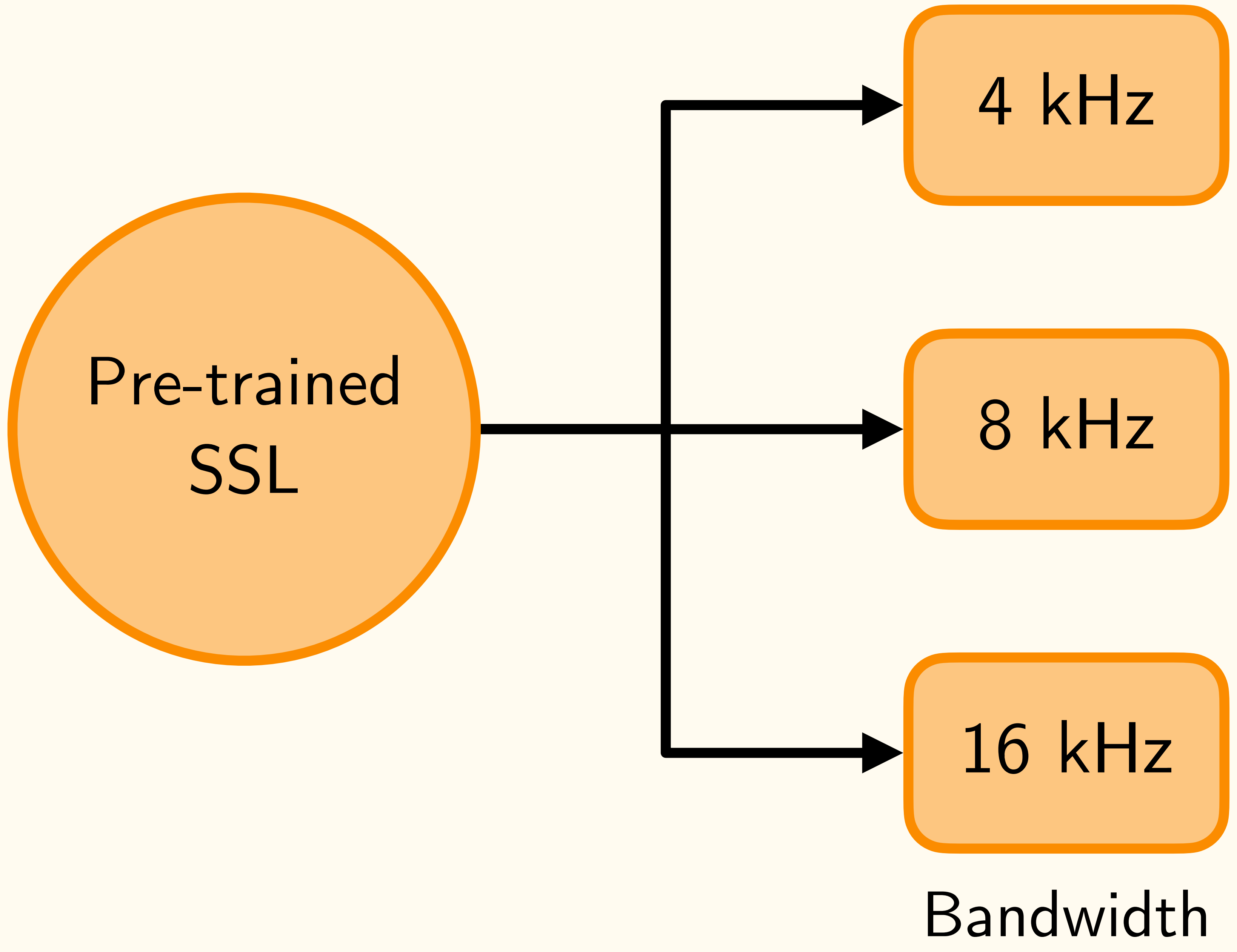
Bandwidth = Sampling Rate / 2

Thesis Contributions 2: Model Bandwidth Mismatch

- Animal vocalizations can go in high frequency ranges compared to human speech.
- Speech SSLs typically pre-trained at 8 kHz bandwidth.

How does this bandwidth mismatch between humans and animals affect this transfer?





Bandwidth = Sampling Rate / 2

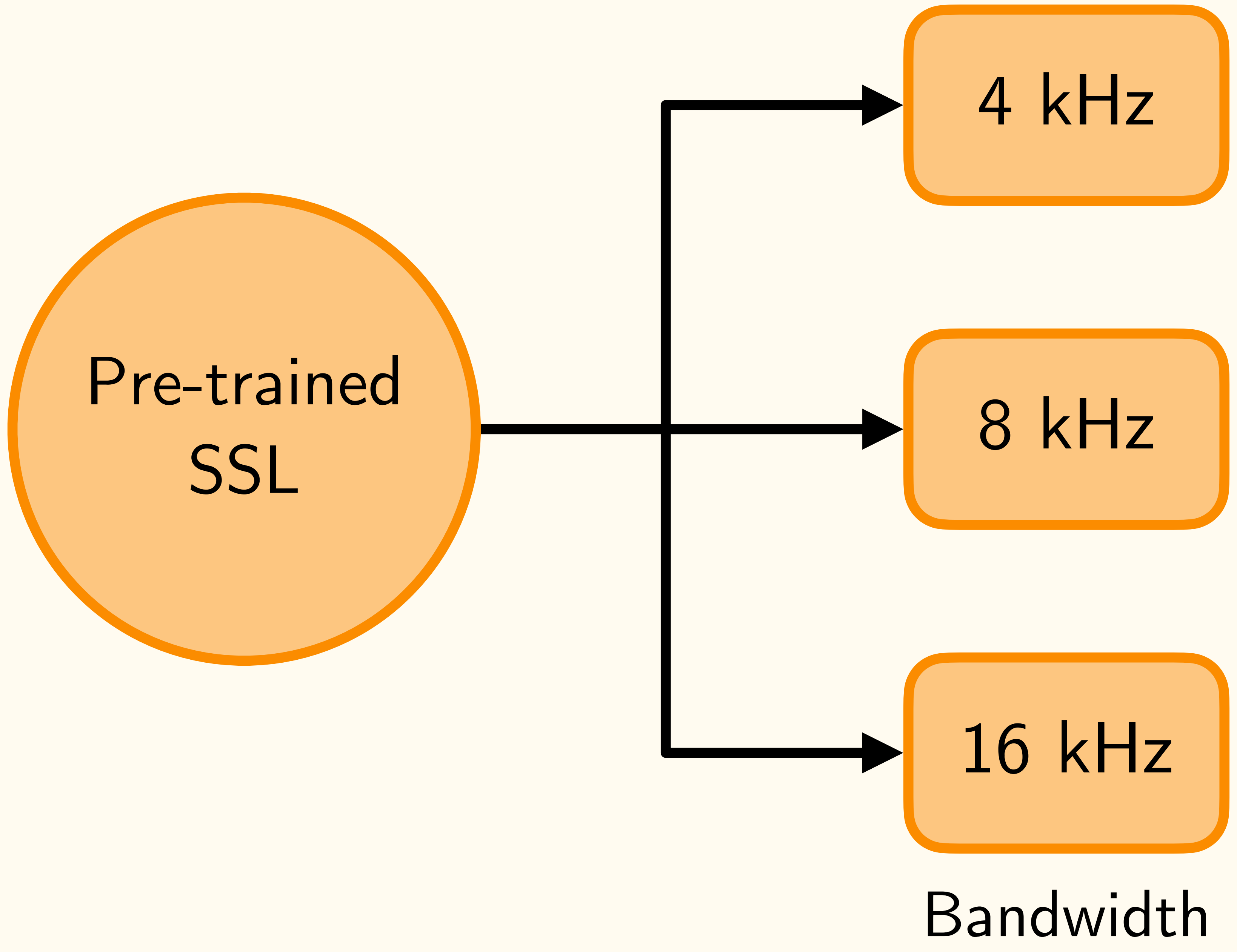
Thesis Contributions 2: Model Bandwidth Mismatch

- Animal vocalizations can go in high frequency ranges compared to human speech.
- Speech SSLs typically pre-trained at 8 kHz bandwidth.

How does this bandwidth mismatch between humans and animals affect this transfer?

Finding

Bandwidth size correlates directly with the performance, increasing monotonically.



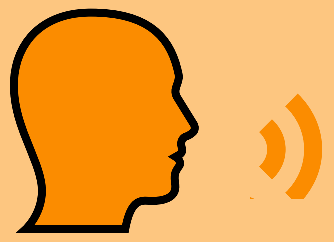
Thesis Contributions 3: Model Pre-Training Domain

Thesis Contributions 3: Model Pre-Training Domain

Is this transferability limited to speech models?

Pre-training

on human speech



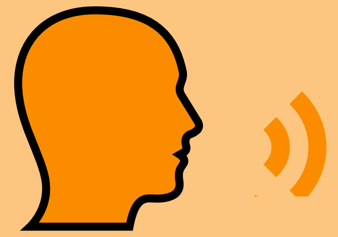
Thesis Contributions 3: Model Pre-Training Domain

Is this transferability limited to speech models?

- SSL pre-training is designed to learn general, domain-agnostic features.

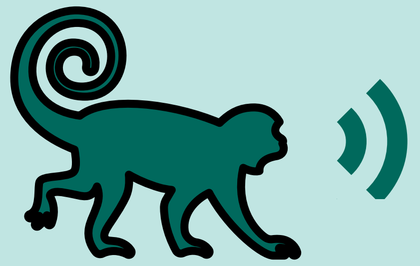
Pre-training

on human speech



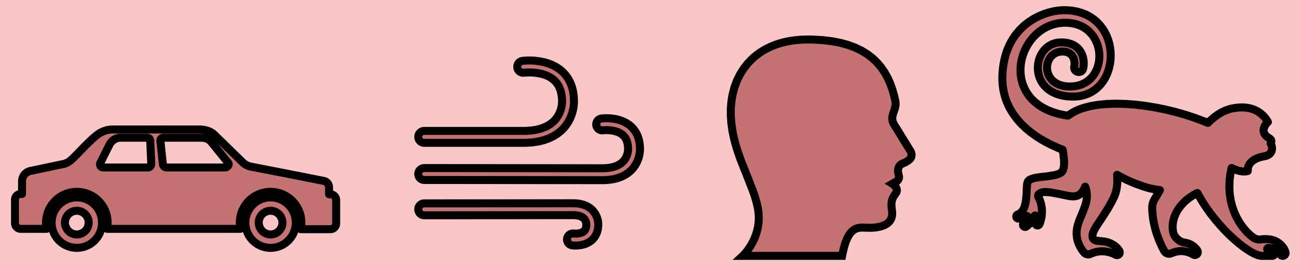
Pre-training

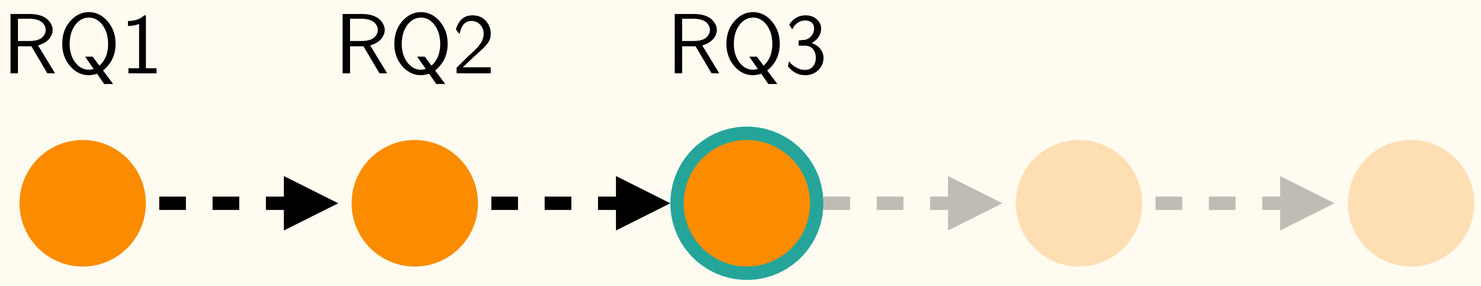
on bioacoustics



Pre-training

on general audio





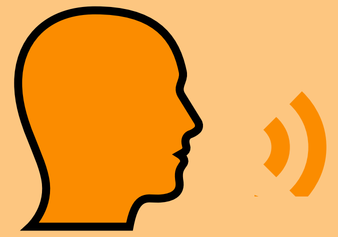
Thesis Contributions 3: Model Pre-Training Domain

Is this transferability limited to speech models?

- SSL pre-training is designed to learn general, domain-agnostic features.
- Can representations learnt from other domains also exhibit this transferability?

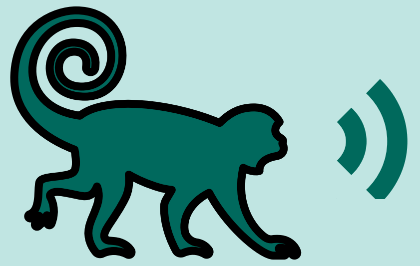
Pre-training

on human speech



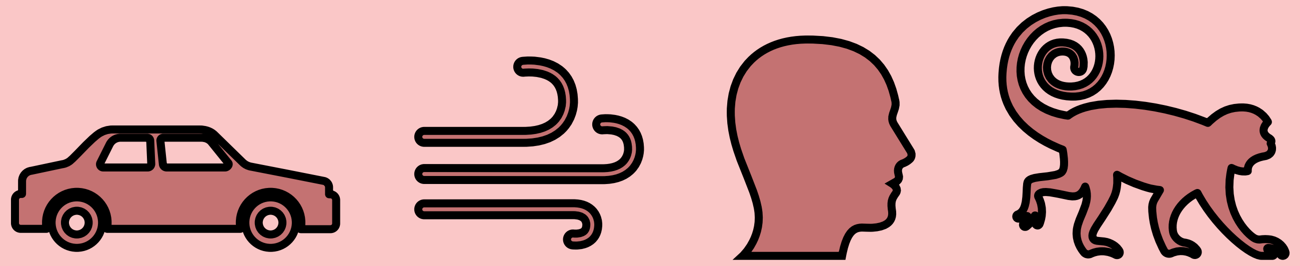
Pre-training

on bioacoustics



Pre-training

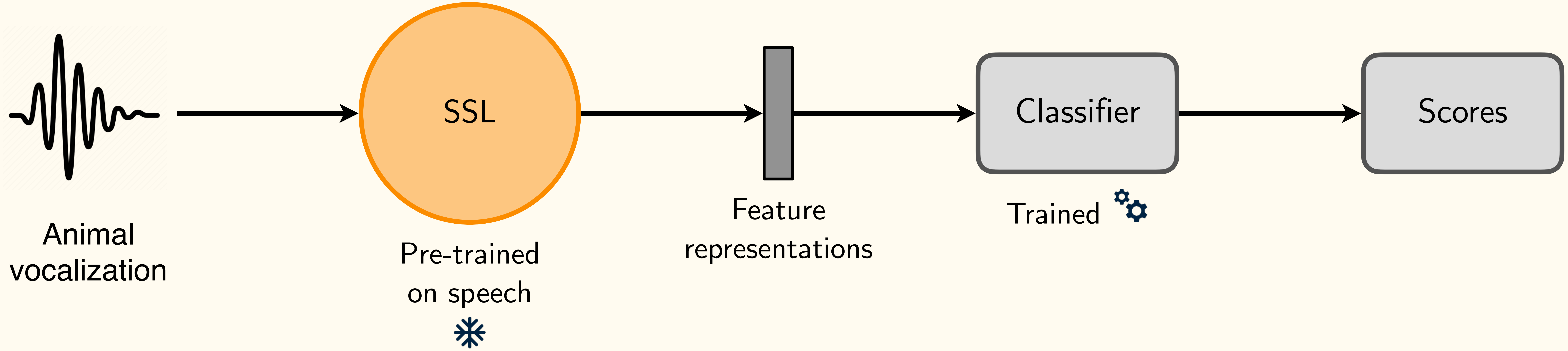
on general audio



Thesis Contributions 4: Model Adaptation

Thesis Contributions 4: Model Adaptation

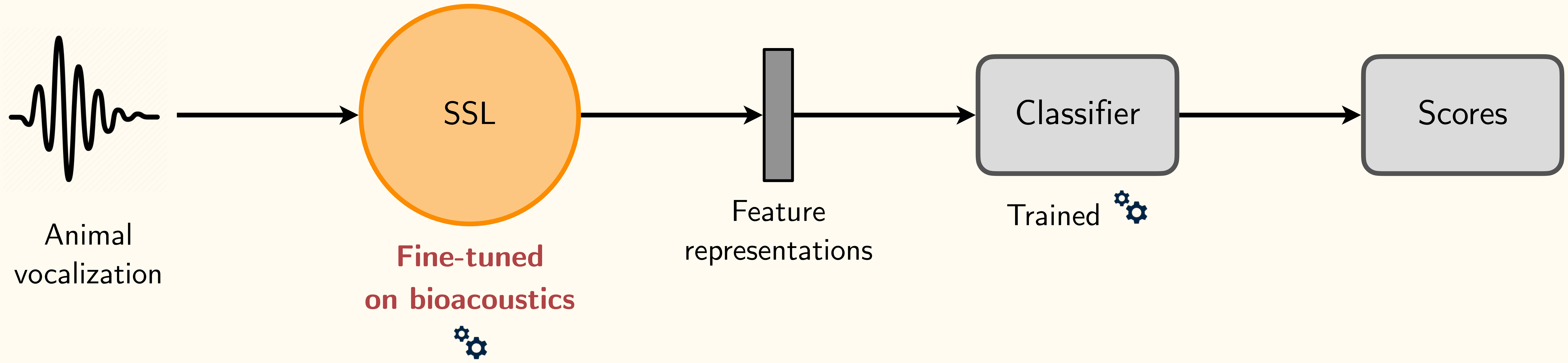
- So far: extracted features from frozen pre-trained models.



Thesis Contributions 4: Model Adaptation

- So far: extracted features from frozen pre-trained models.

Can adaptation of these pre-trained SSL models further improve the transferability?



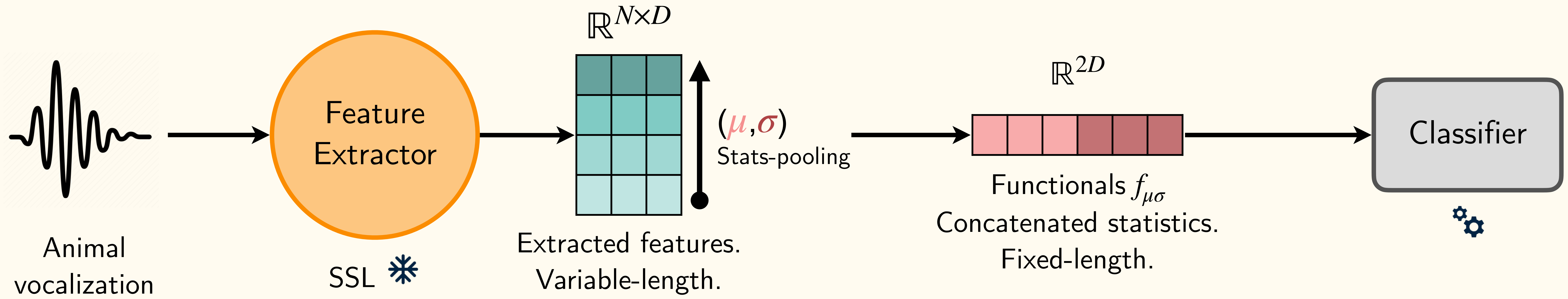
Thesis Contributions 5: Leveraging Sequential Structure

Thesis Contributions 5: Leveraging Sequential Structure

How can we capture the sequential structure of animal vocalizations?

Thesis Contributions 5: Leveraging Sequential Structure

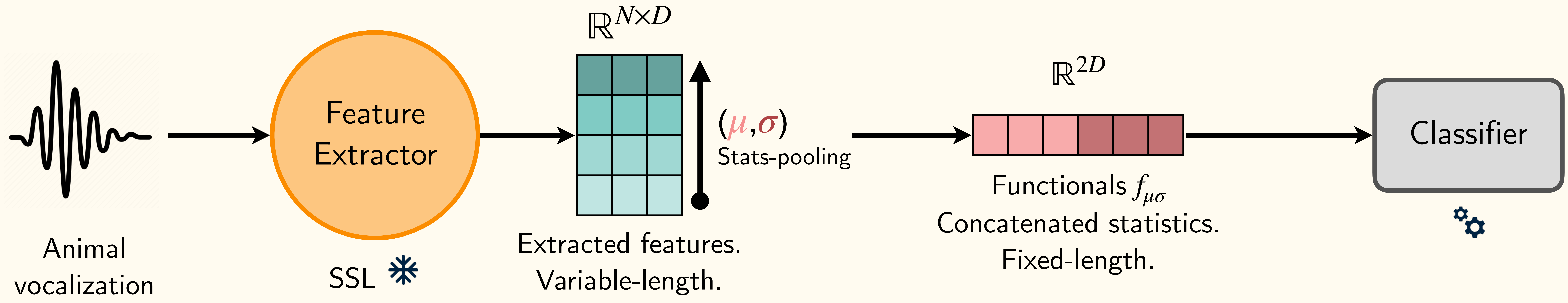
How can we capture the sequential structure of animal vocalizations?



Thesis Contributions 5: Leveraging Sequential Structure

How can we capture the sequential structure of animal vocalizations?

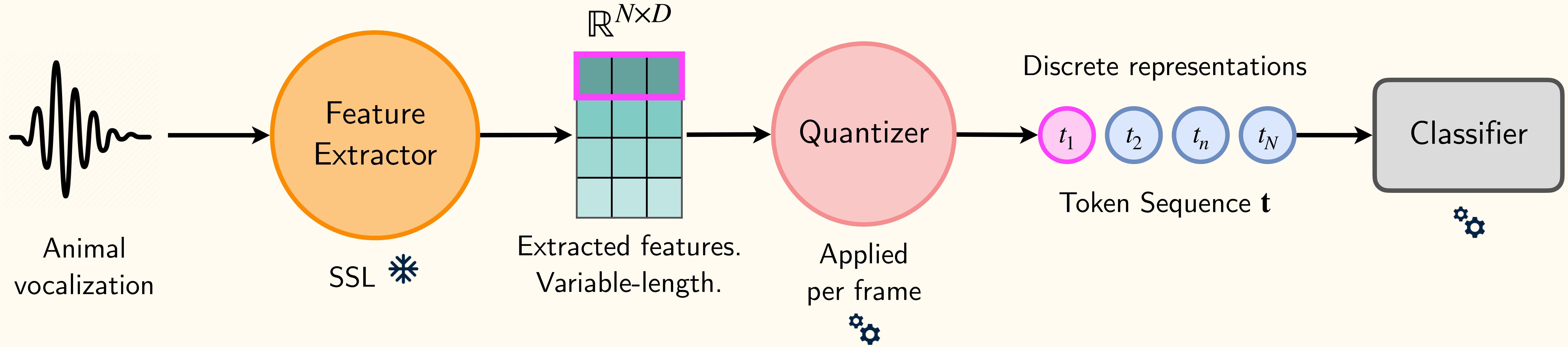
- Each vocalization treated like an unordered collection of frame-level features.



Thesis Contributions 5: Leveraging Sequential Structure

How can we capture the sequential structure of animal vocalizations?

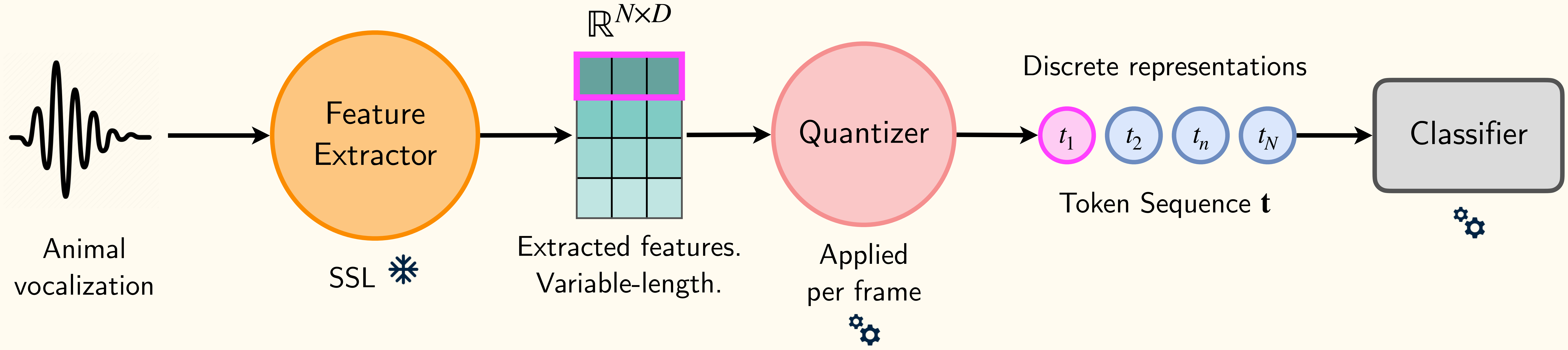
- Each vocalization treated like an unordered collection of frame-level features.



Thesis Contributions 5: Leveraging Sequential Structure

How can we capture the sequential structure of animal vocalizations?

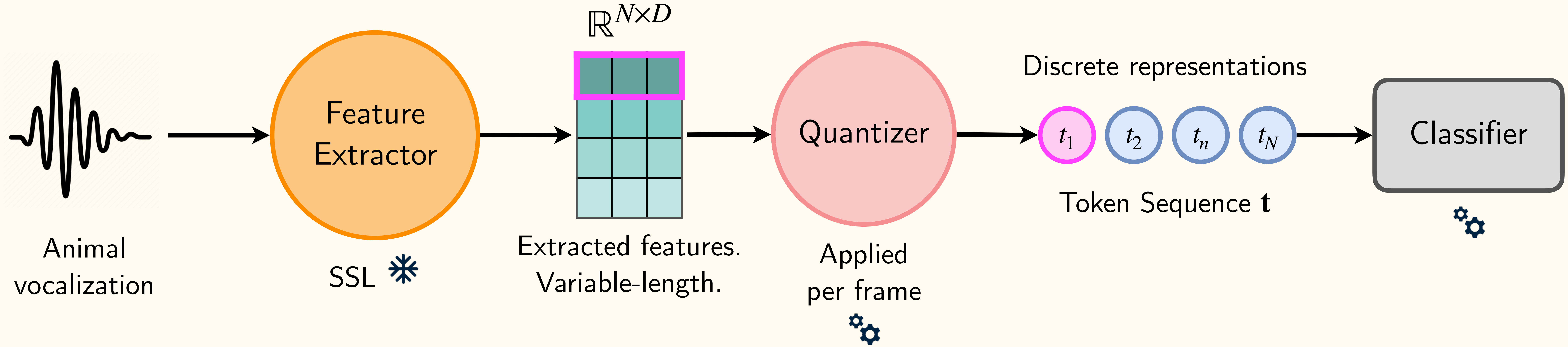
- Each vocalization treated like an unordered collection of frame-level features.
- Can discrete token representations leverage temporal information?



Thesis Contributions 5: Leveraging Sequential Structure

How can we capture the sequential structure of animal vocalizations?

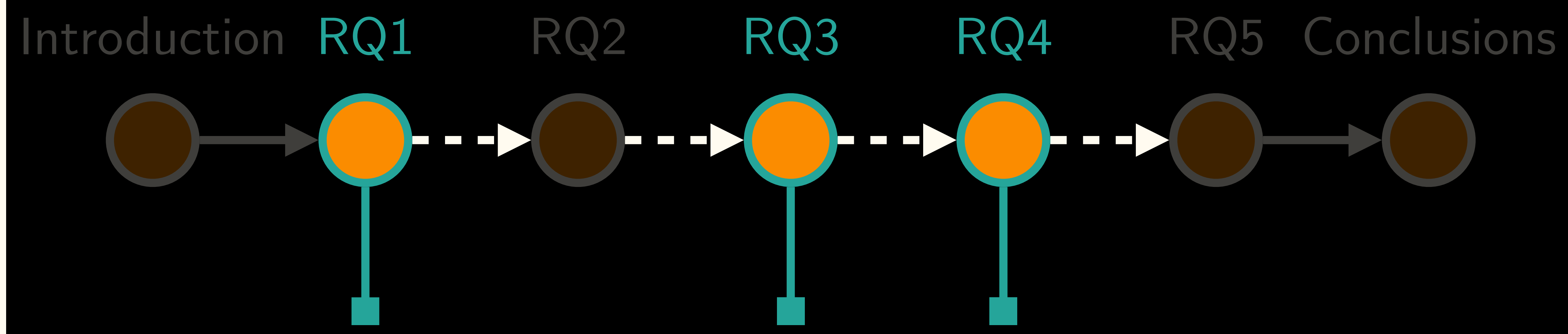
- Each vocalization treated like an unordered collection of frame-level features.
- Can discrete token representations leverage temporal information?



Finding

Token sequence representations are weaker than the stats-pooled representations.

Thesis Contributions



RQ1. Transferability

RQ2. Bandwidth

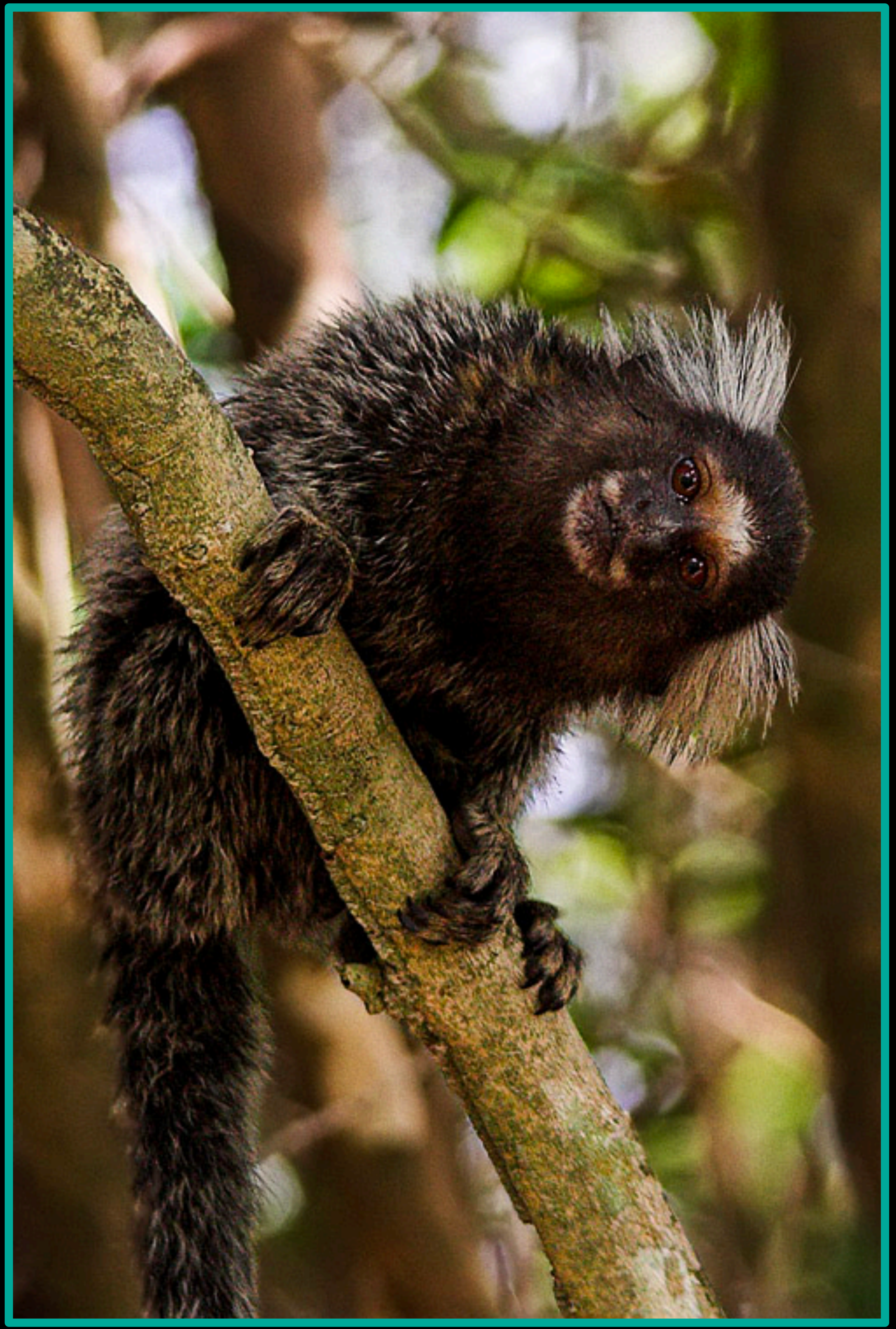
RQ3. Pre-Training Domain

RQ4. Fine-Tuning

RQ5. Sequential Structure

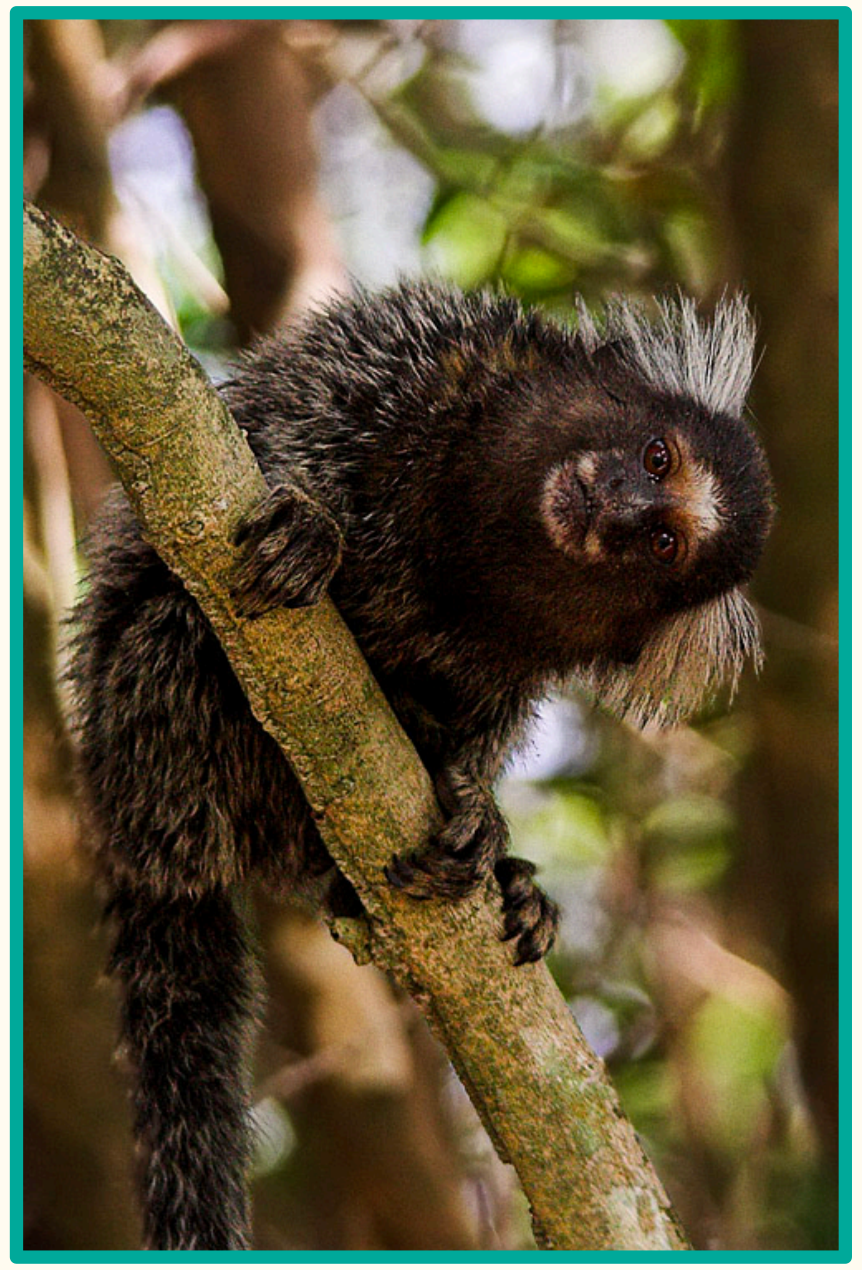
Marmosets

Vocalizations, Datasets, Tasks



Carmem A. Busko. *Callithrix jacchus*, Wikipedia.

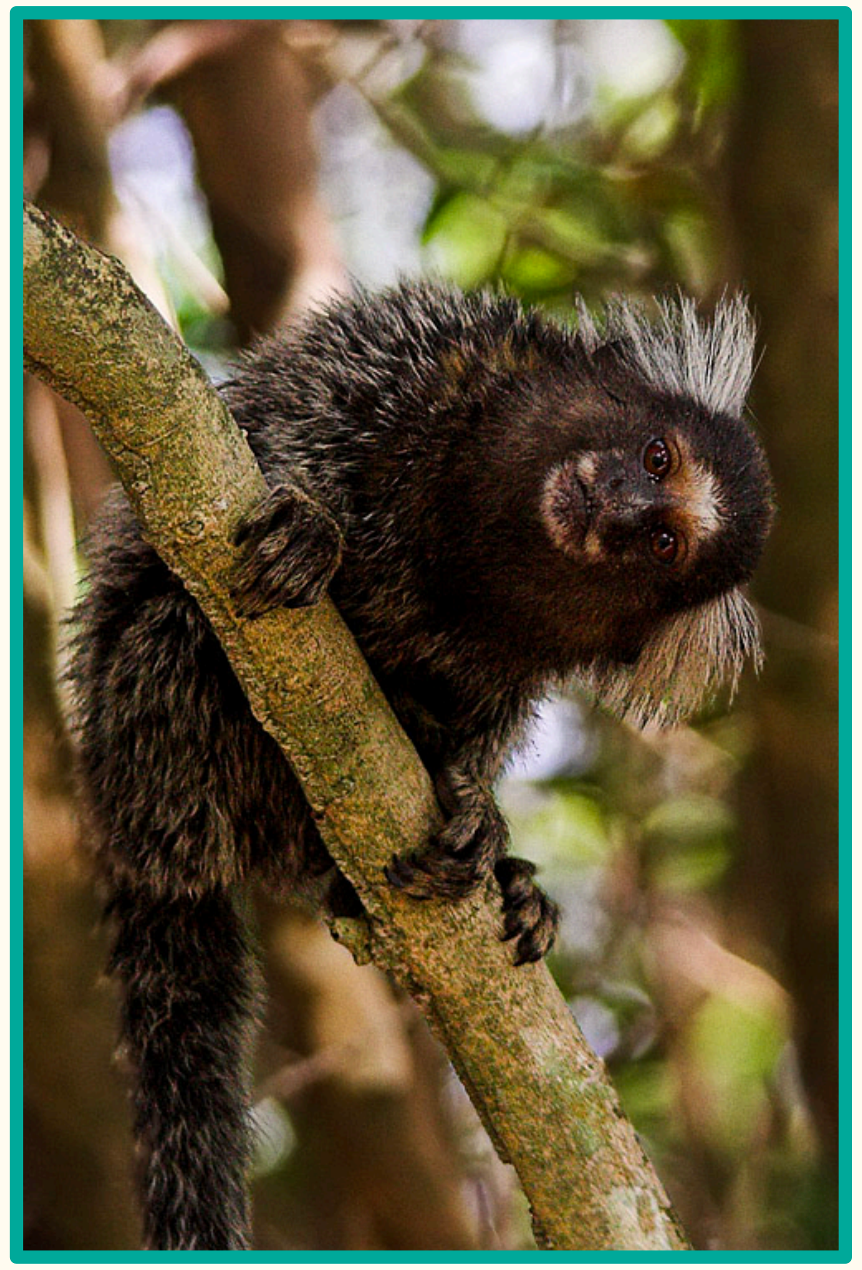
Marmoset Vocalizations



Carmem A. Busko. *Callithrix jacchus*, Wikipedia.

Marmoset Vocalizations

- Highly vocal nature rooted in a *complex social system*.

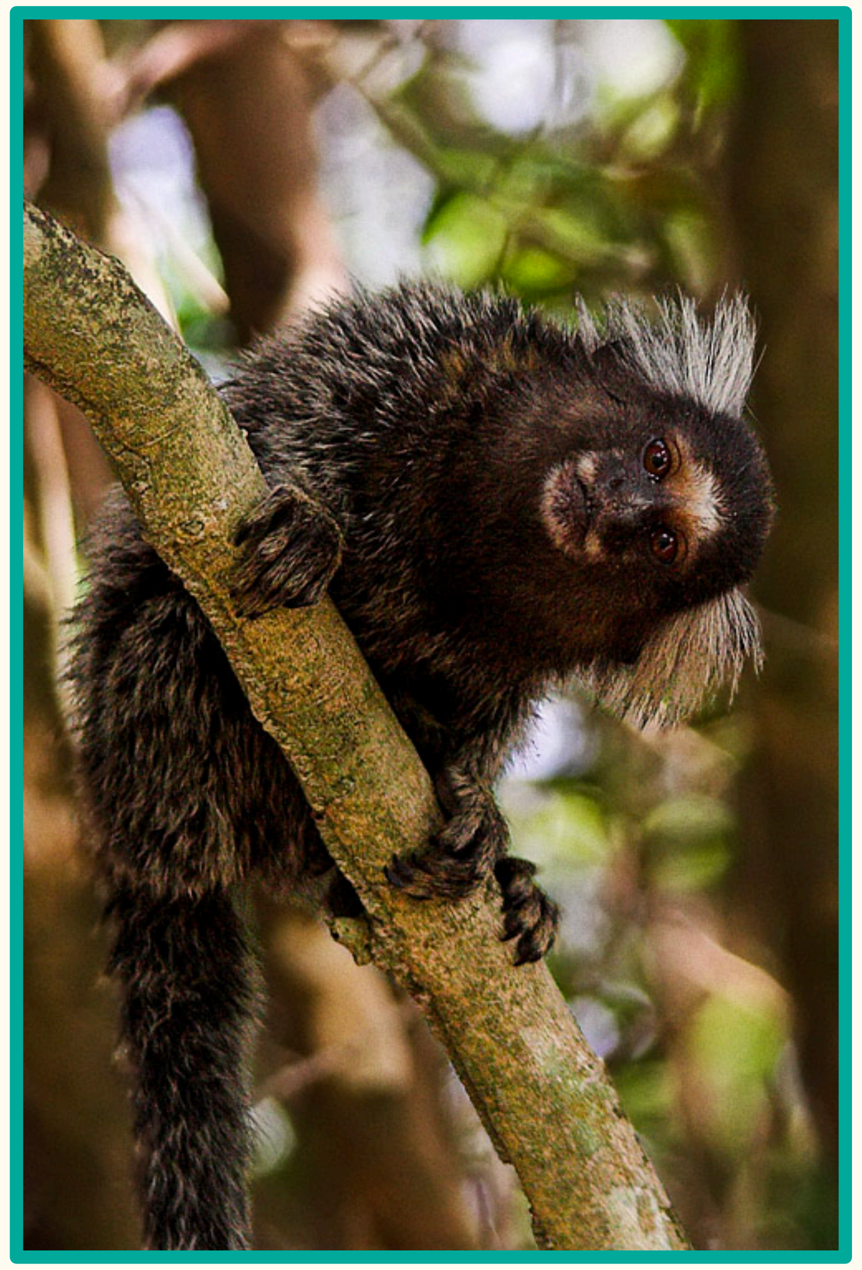


Carmem A. Busko. *Callithrix jacchus*, Wikipedia.

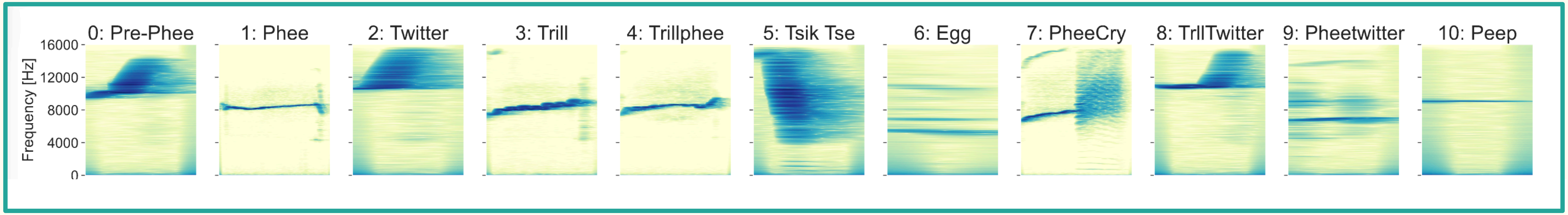
Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalizations

- Highly vocal nature rooted in a *complex social system*.
  - Acoustically diverse call repertoire.



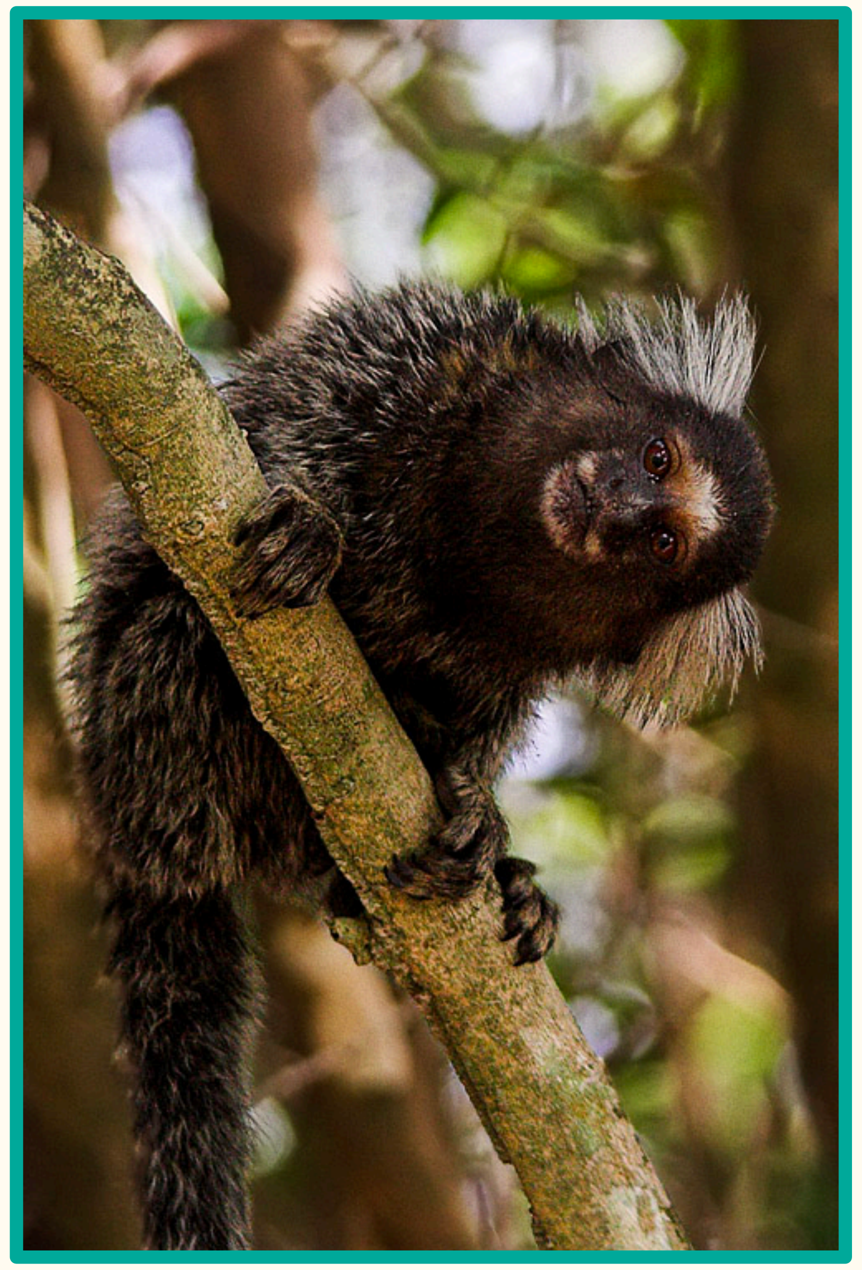
Carmem A. Busko. *Callithrix jacchus*, Wikipedia.



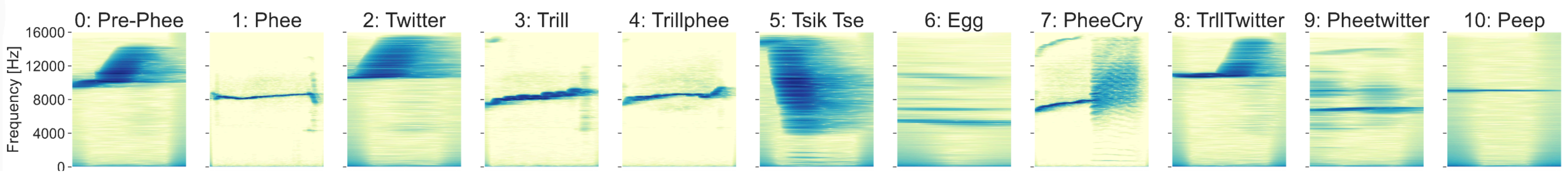
Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalizations

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



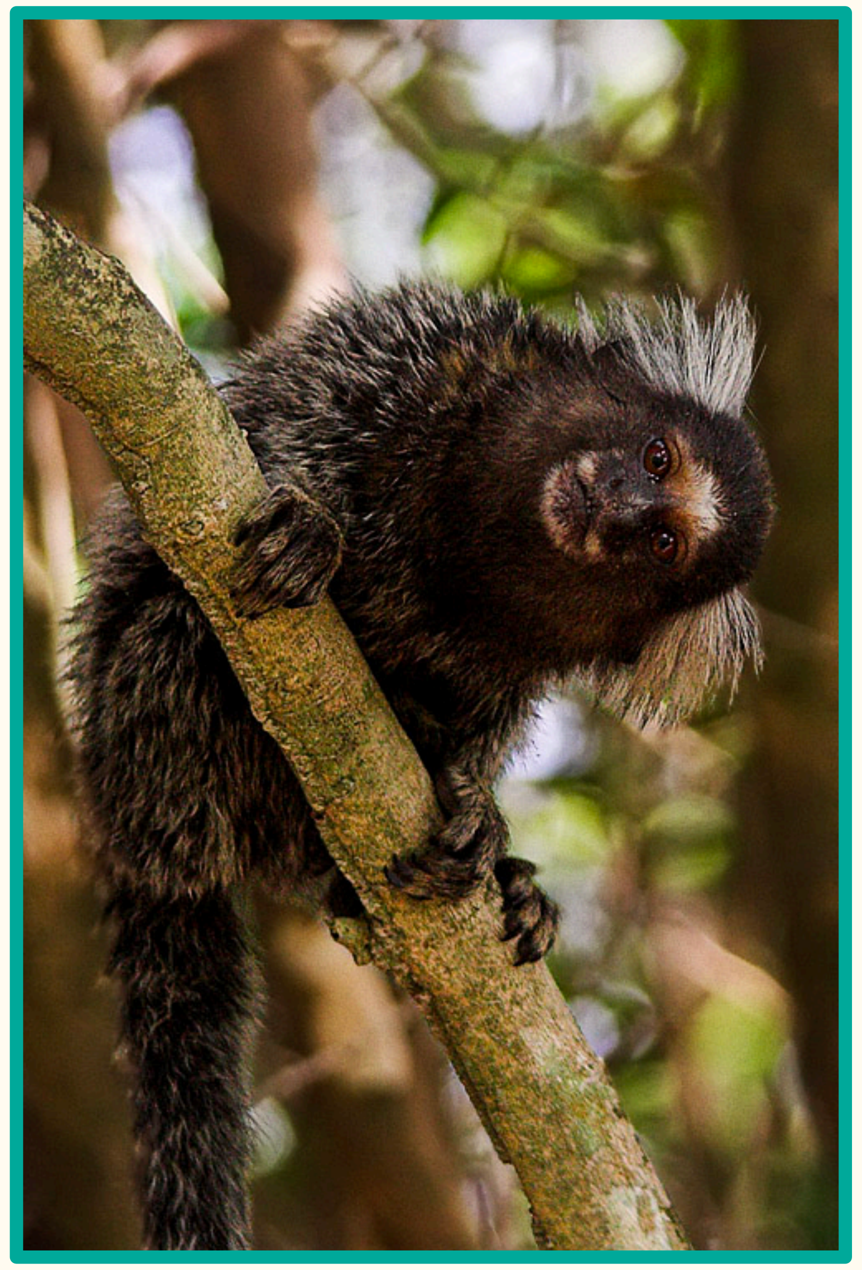
Carmem A. Busko. *Callithrix jacchus*, Wikipedia.



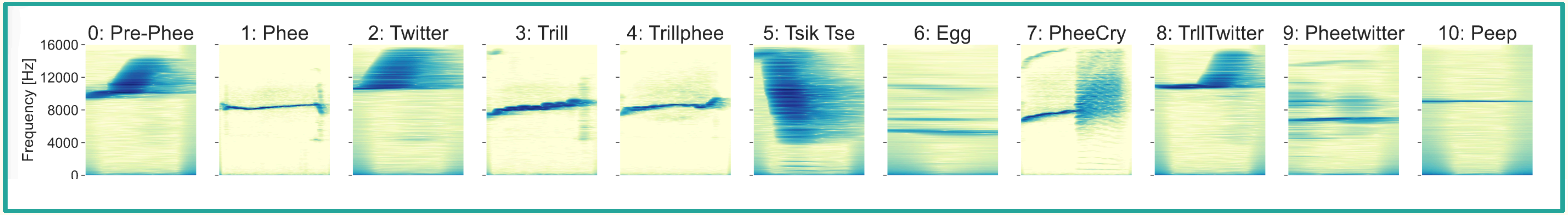
Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalizations

- Highly vocal nature rooted in a *complex social system*.
  - Acoustically diverse call repertoire.
  - Ability to encode a range of information.
- Remarkable *vocal adaptability* allows them to modify their calls:



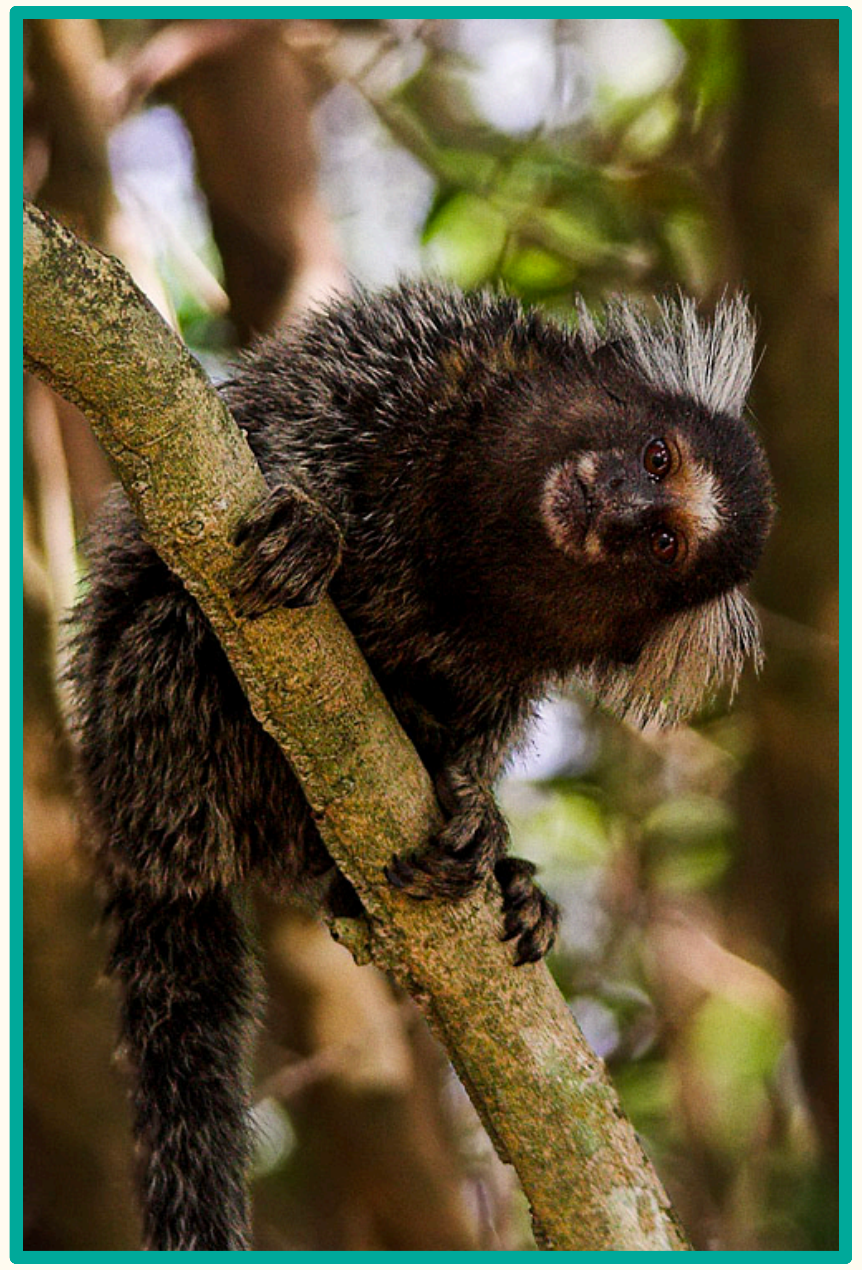
Carmem A. Busko. *Callithrix jacchus*, Wikipedia.



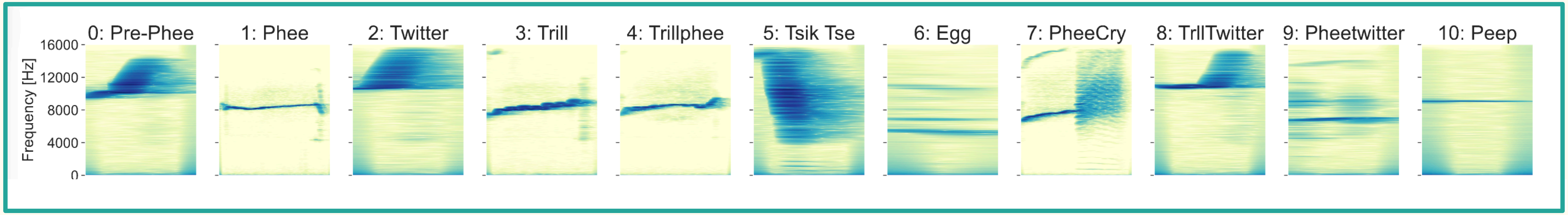
Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalizations

- Highly vocal nature rooted in a *complex social system*.
  - Acoustically diverse call repertoire.
  - Ability to encode a range of information.
- Remarkable *vocal adaptability* allows them to modify their calls:
  - Duration - Complexity
  - Intensity - Timing



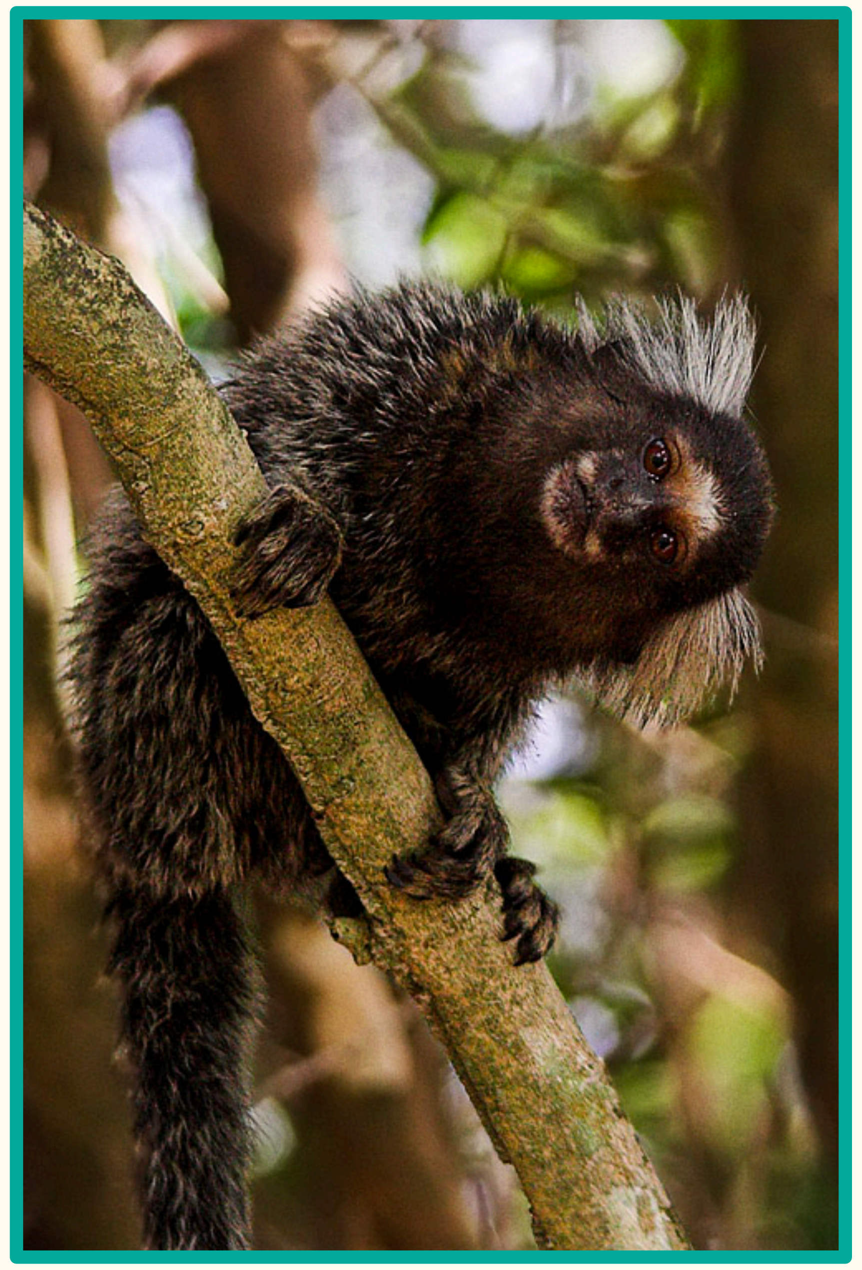
Carmem A. Busko. *Callithrix jacchus*, Wikipedia.



Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalizations

- Highly vocal nature rooted in a *complex social system*.
  - Acoustically diverse call repertoire.
  - Ability to encode a range of information.
- Remarkable *vocal adaptability* allows them to modify their calls:
  - Duration - Complexity
  - Intensity - Timing
- *Vocal characteristics* align them closely with human speech properties:

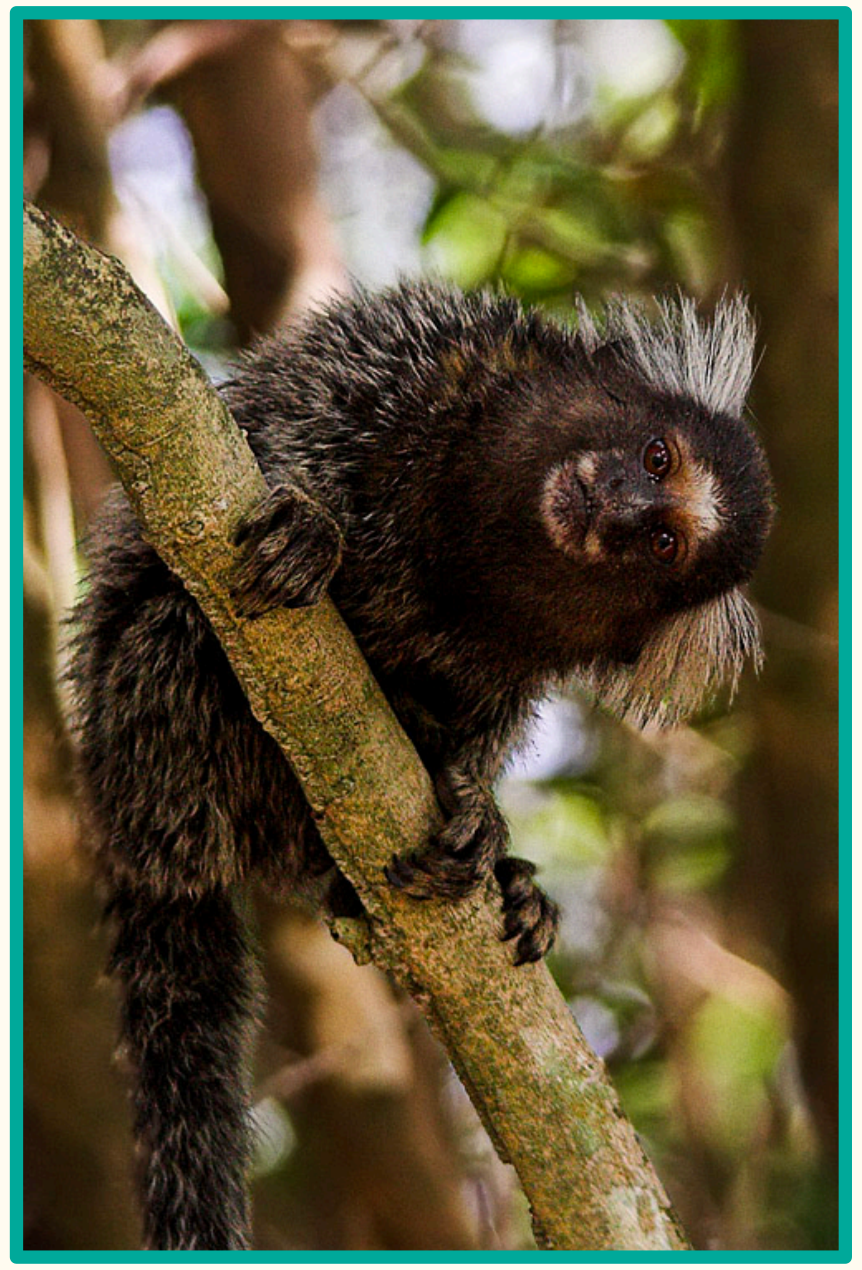


Carmem A. Busko. *Callithrix jacchus*, Wikipedia.

Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalizations

- Highly vocal nature rooted in a *complex social system*.
  - Acoustically diverse call repertoire.
  - Ability to encode a range of information.
- Remarkable *vocal adaptability* allows them to modify their calls:
  - Duration - Complexity
  - Intensity - Timing
- *Vocal characteristics* align them closely with human speech properties:
  - Turn-taking - Categorical perception of sounds
  - Care-giving to infants - Cooperative breeding



Carmem A. Busko. *Callithrix jacchus*, Wikipedia.

Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

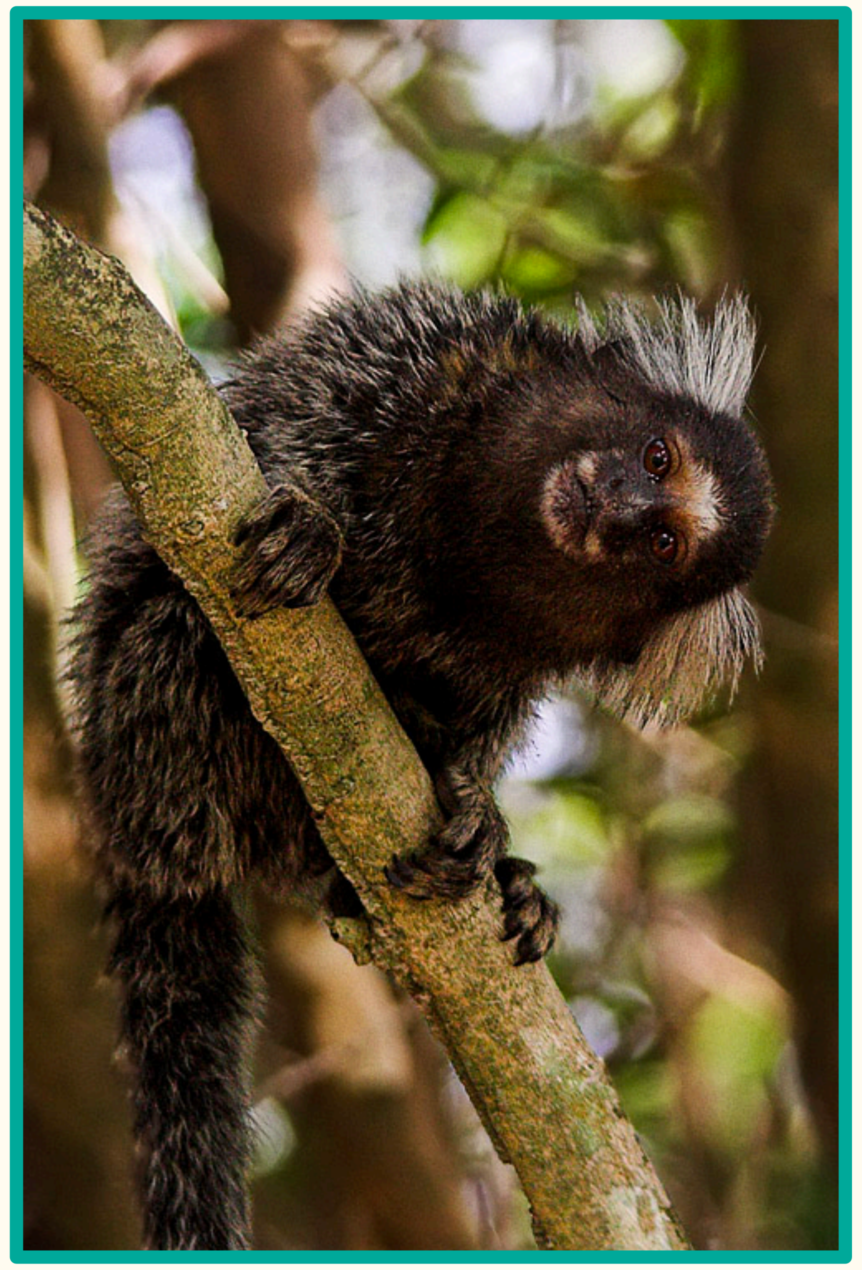
Marmoset Vocalizations

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

- Remarkable *vocal adaptability* allows them to modify their calls:
  - Duration - Complexity
  - Intensity - Timing

- *Vocal characteristics* align them closely with human speech properties:
  - Turn-taking - Categorical perception of sounds
  - Care-giving to infants - Cooperative breeding

➡ Valuable surrogate model for studying the evolutionary origins of human speech.



Carmem A. Busko. *Callithrix jacchus*, Wikipedia.

Agamaite, J. A. et al. *A quantitative acoustic analysis of the vocal repertoire of the common marmoset.* (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.

Marmoset Vocalization Datasets

Marmoset Vocalization Datasets

- Recorded from cages with fixed mic.

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

Marmoset Vocalization Datasets

- Recorded from cages with fixed mic.
- Manually annotated by researcher.

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

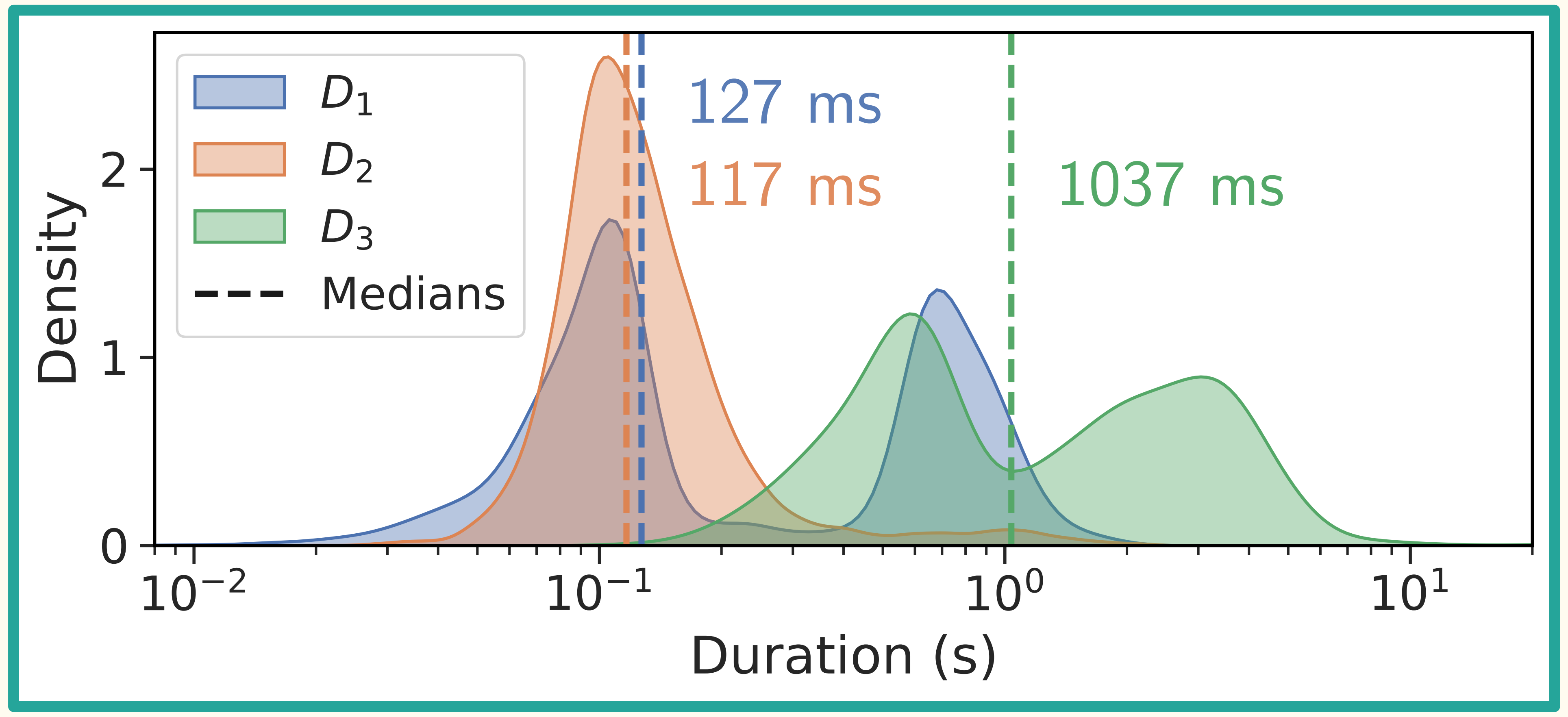
Marmoset Vocalization Datasets

- Recorded from cages with fixed mic.
- Manually annotated by researcher.
- 3 marmoset datasets (D_1, D_2, D_3).

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

D	Dataset	S	L
D_1	IMV	72,920	464
D_2	Bosshard	13,808	37
D_3	Wierucka	4,901	138

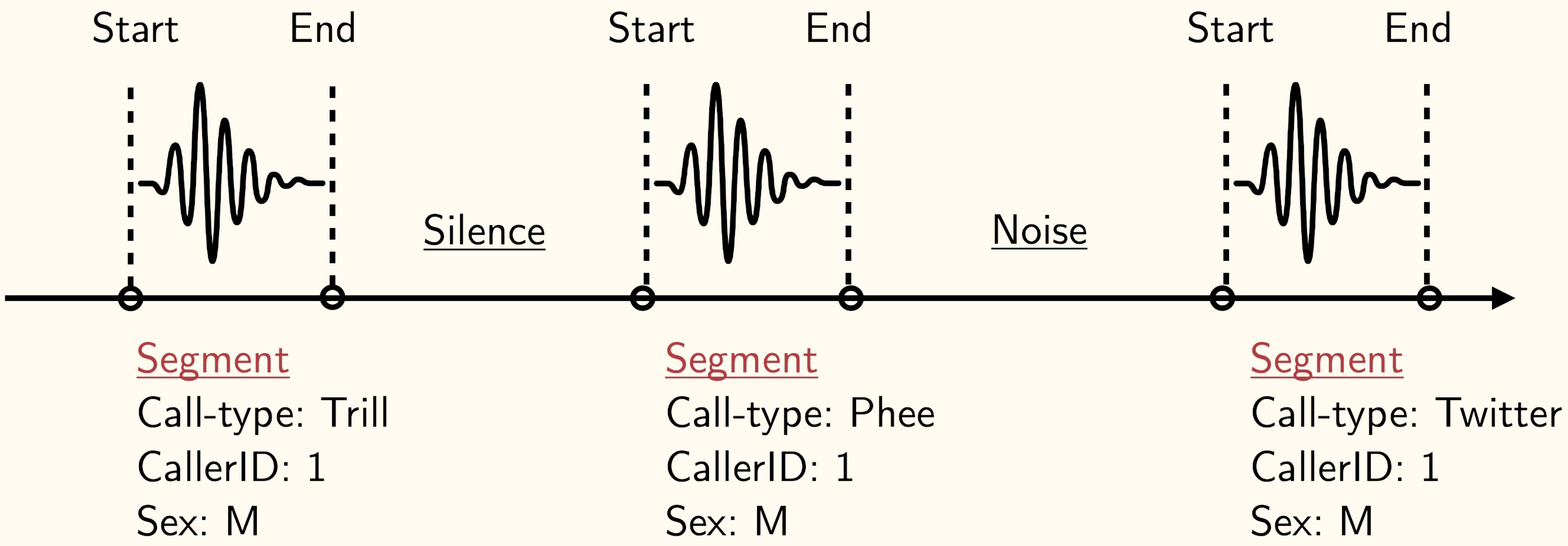
S: number of samples, L: total length [minutes].



Marmoset Vocalization Tasks

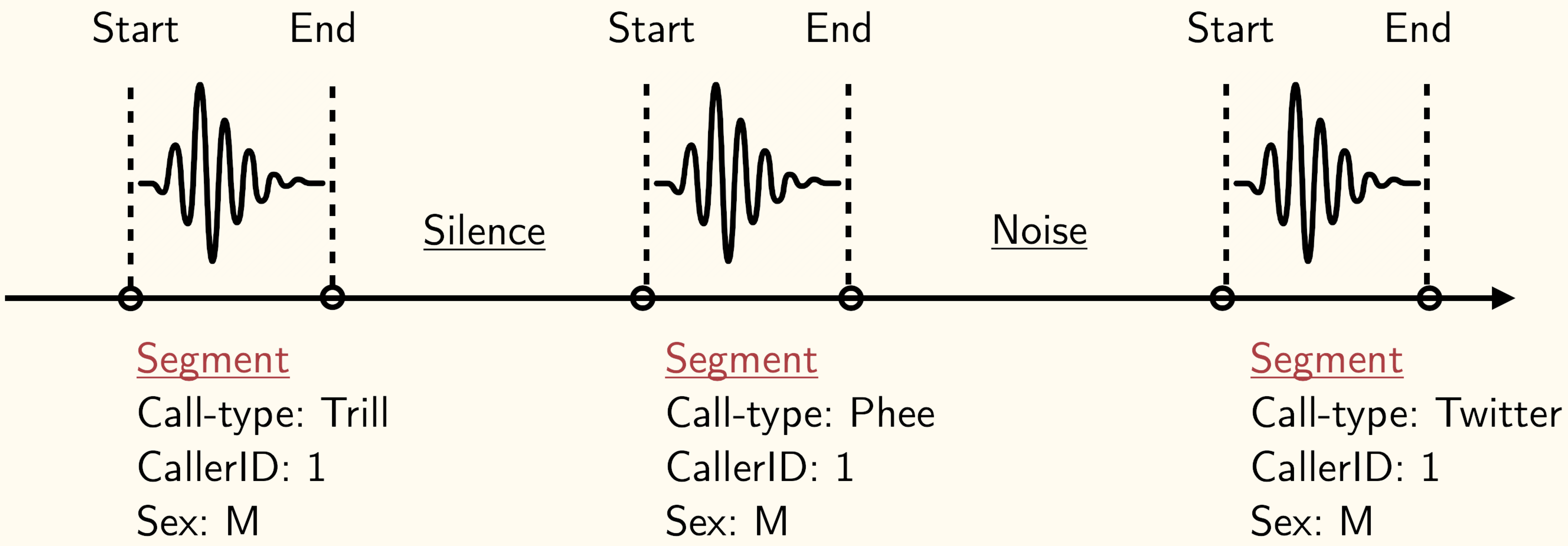
Marmoset Vocalization Tasks

- Data pre-segmented:



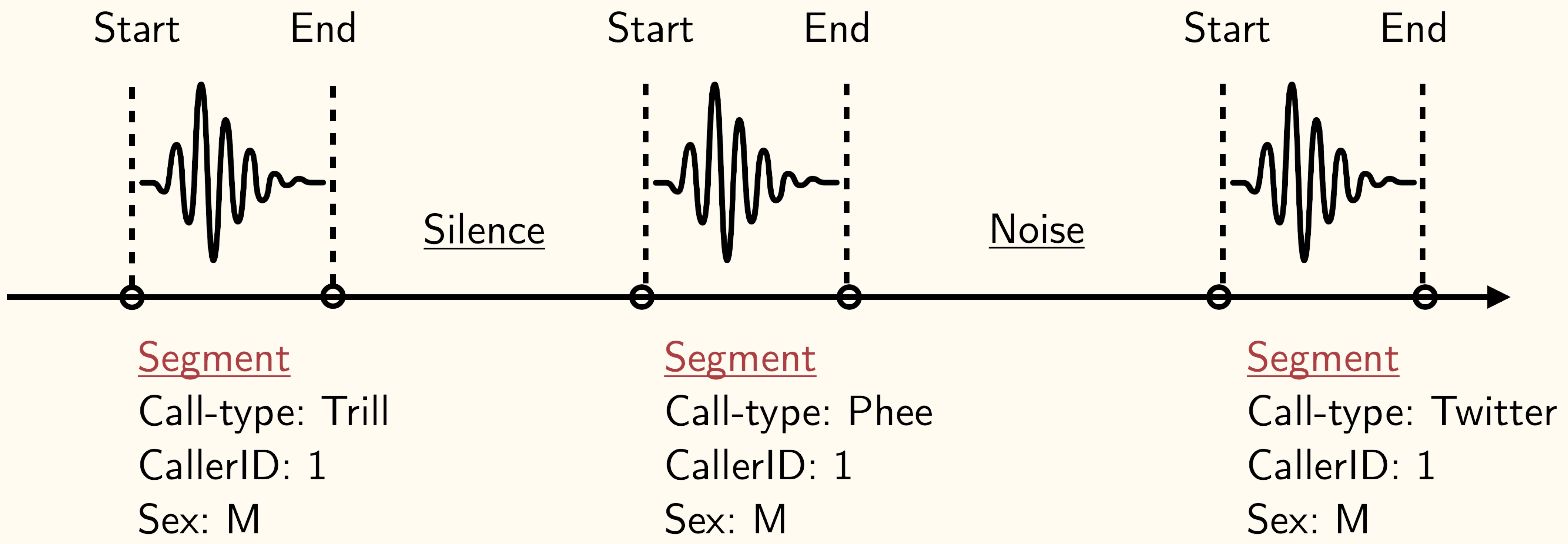
Marmoset Vocalization Tasks

- Data pre-segmented:
    - Vocalization detection not needed.



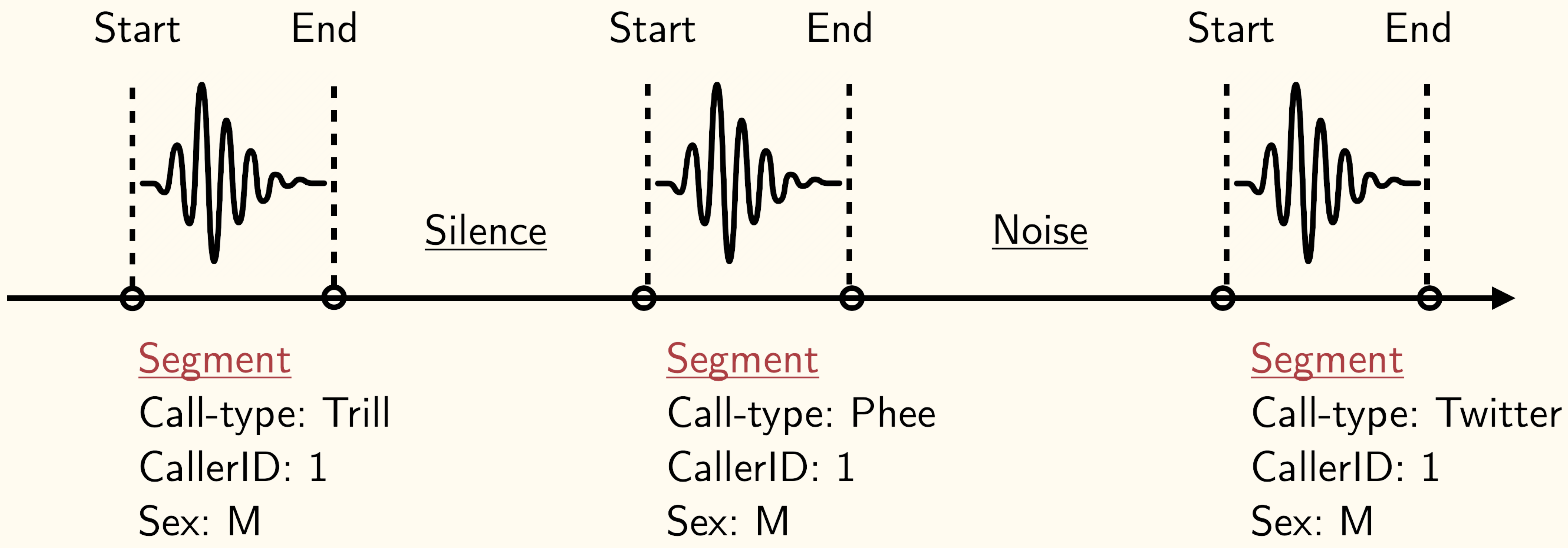
Marmoset Vocalization Tasks

- Data pre-segmented:
    - Vocalization detection not needed.
    - Removed silence and noise.



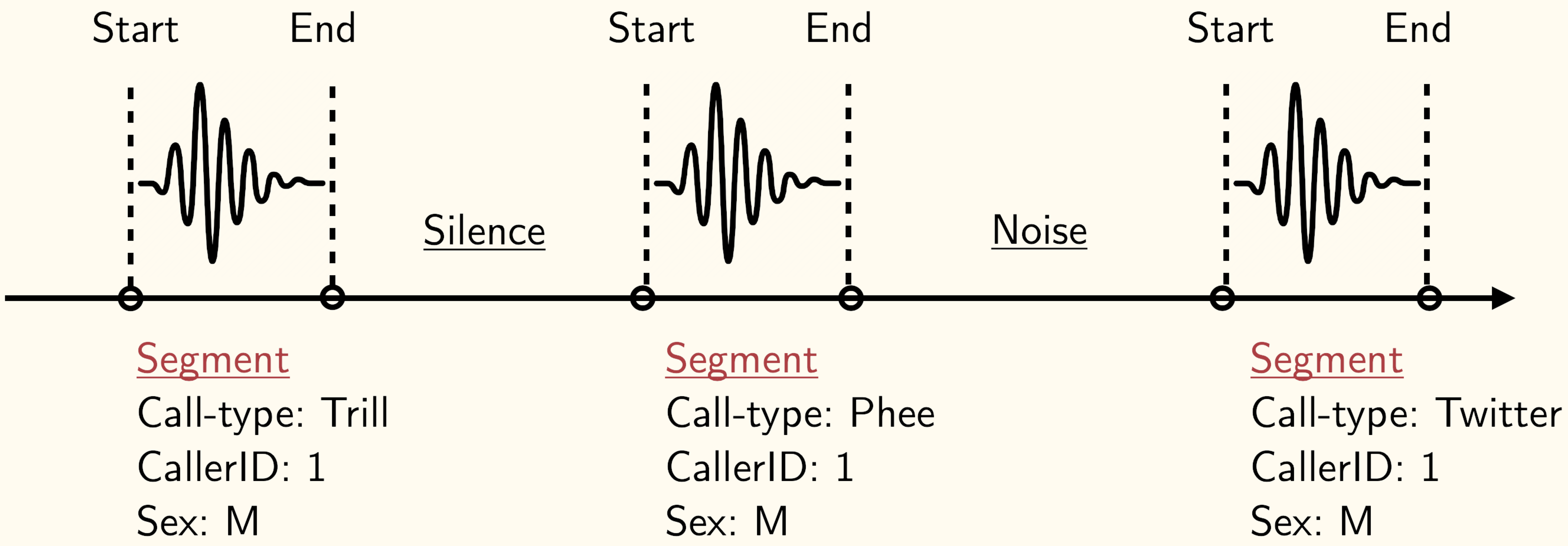
Marmoset Vocalization Tasks

- Data pre-segmented:
  - Vocalization detection not needed.
  - Removed silence and noise.
- 3 classification tasks:



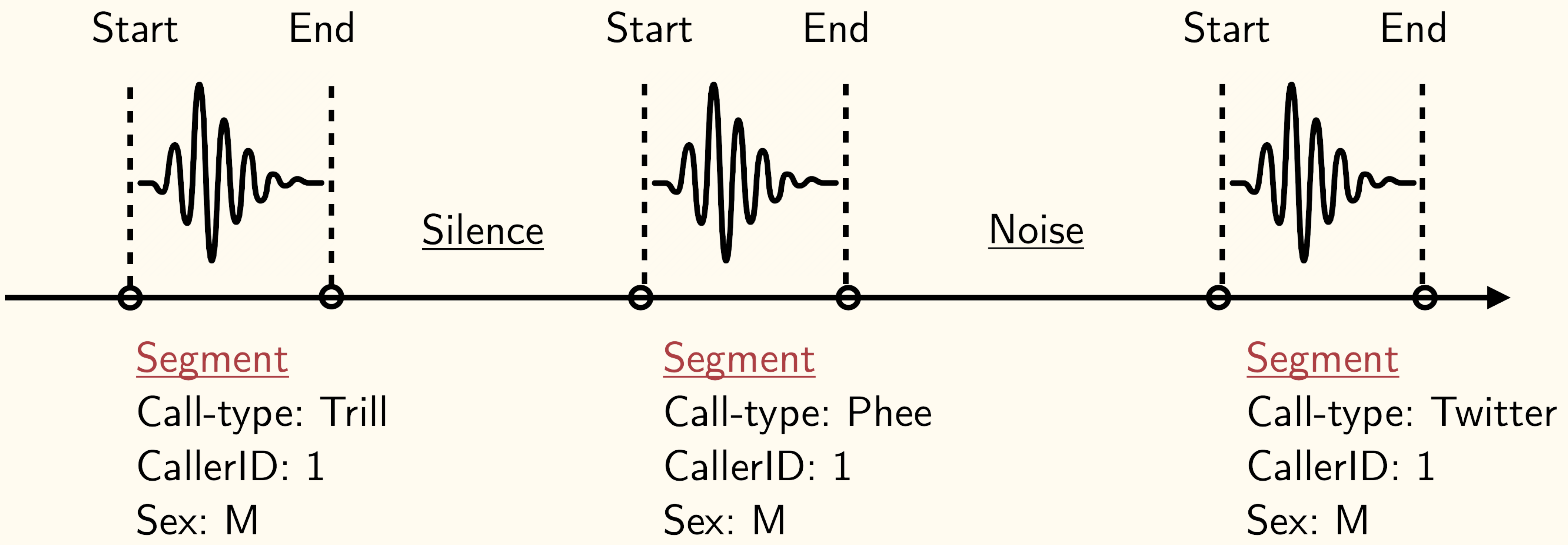
Marmoset Vocalization Tasks

- Data pre-segmented:
  - Vocalization detection not needed.
  - Removed silence and noise.
- 3 classification tasks:
  - CTID: Call-type identification.



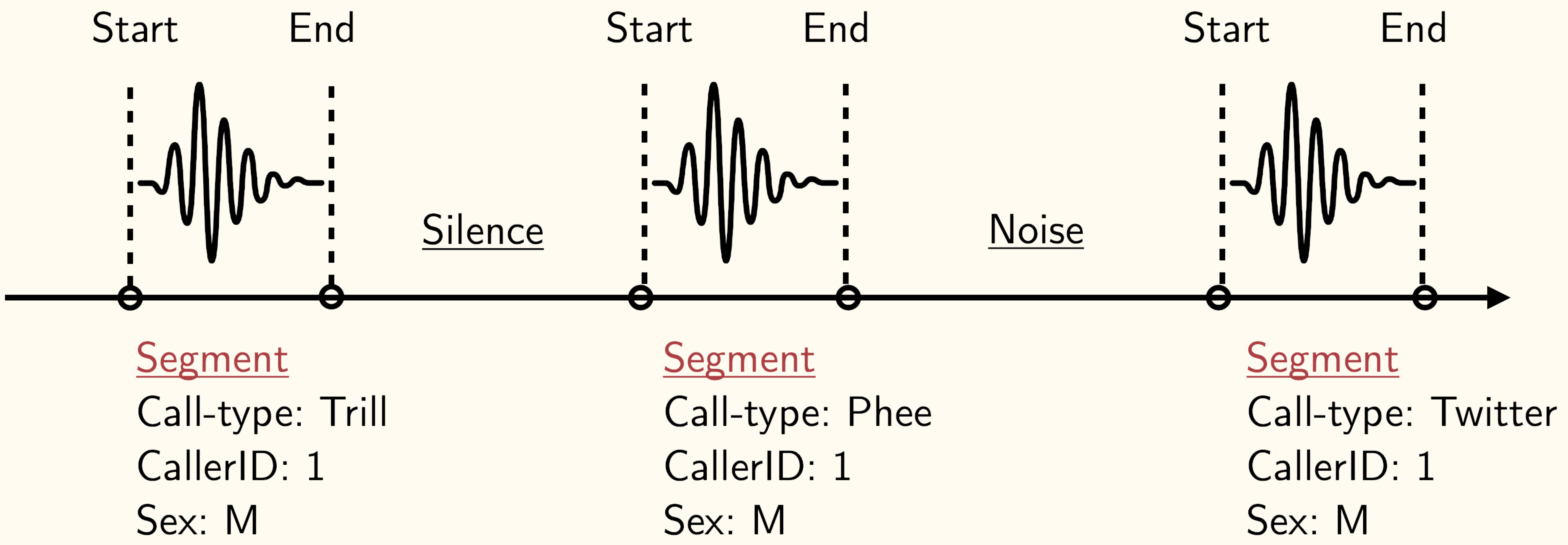
Marmoset Vocalization Tasks

- Data pre-segmented:
  - Vocalization detection not needed.
  - Removed silence and noise.
- 3 classification tasks:
  - CTID: Call-type identification.
  - CLID: Caller identification.



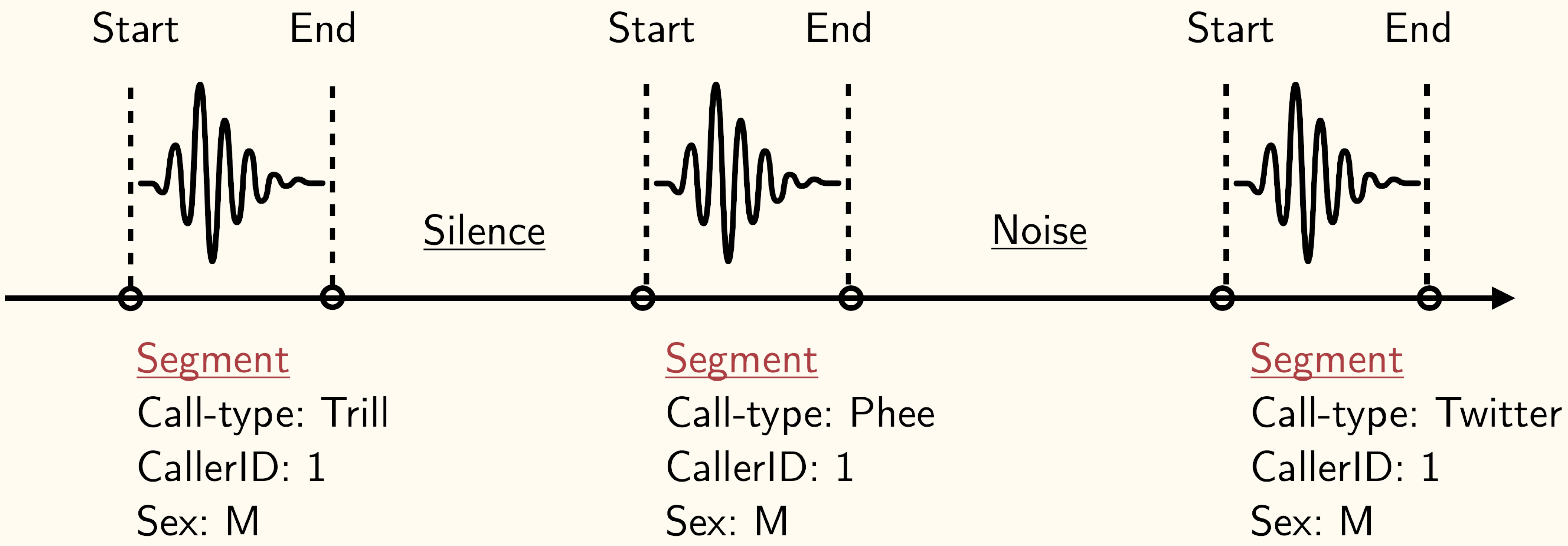
Marmoset Vocalization Tasks

- Data pre-segmented:
  - Vocalization detection not needed.
  - Removed silence and noise.
- 3 classification tasks:
  - CTID: Call-type identification.
  - CLID: Caller identification.
  - SID: Sex identification.



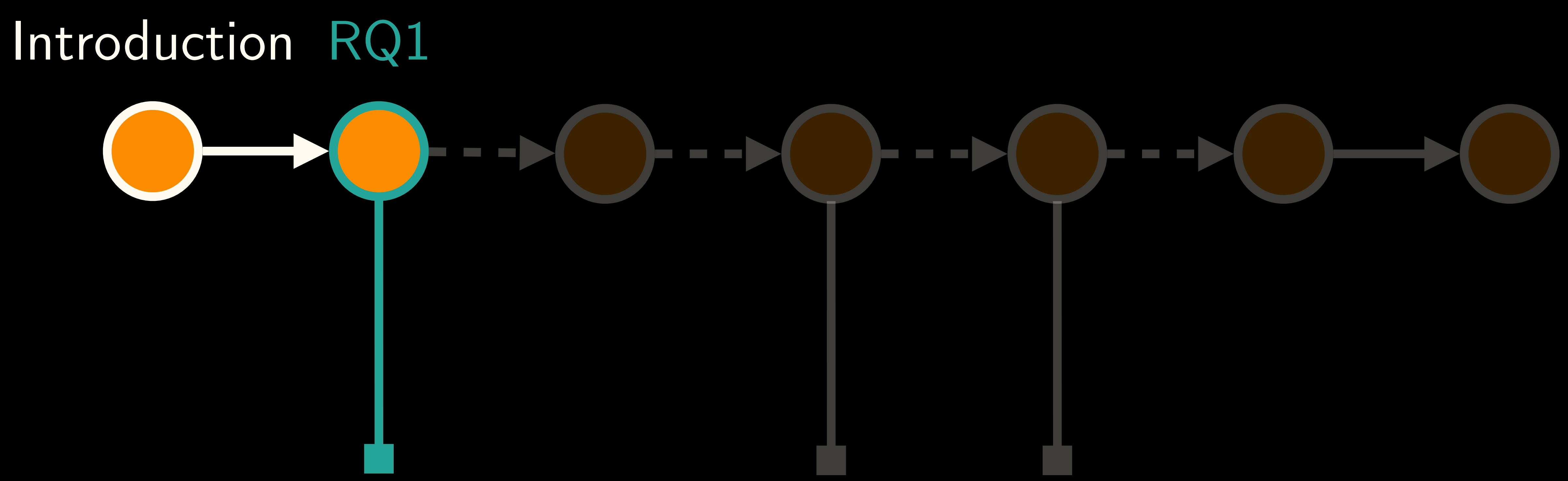
Marmoset Vocalization Tasks

- Data pre-segmented:
  - Vocalization detection not needed.
  - Removed silence and noise.
- 3 classification tasks:
  - CTID: Call-type identification.
  - CLID: Caller identification.
  - SID: Sex identification.



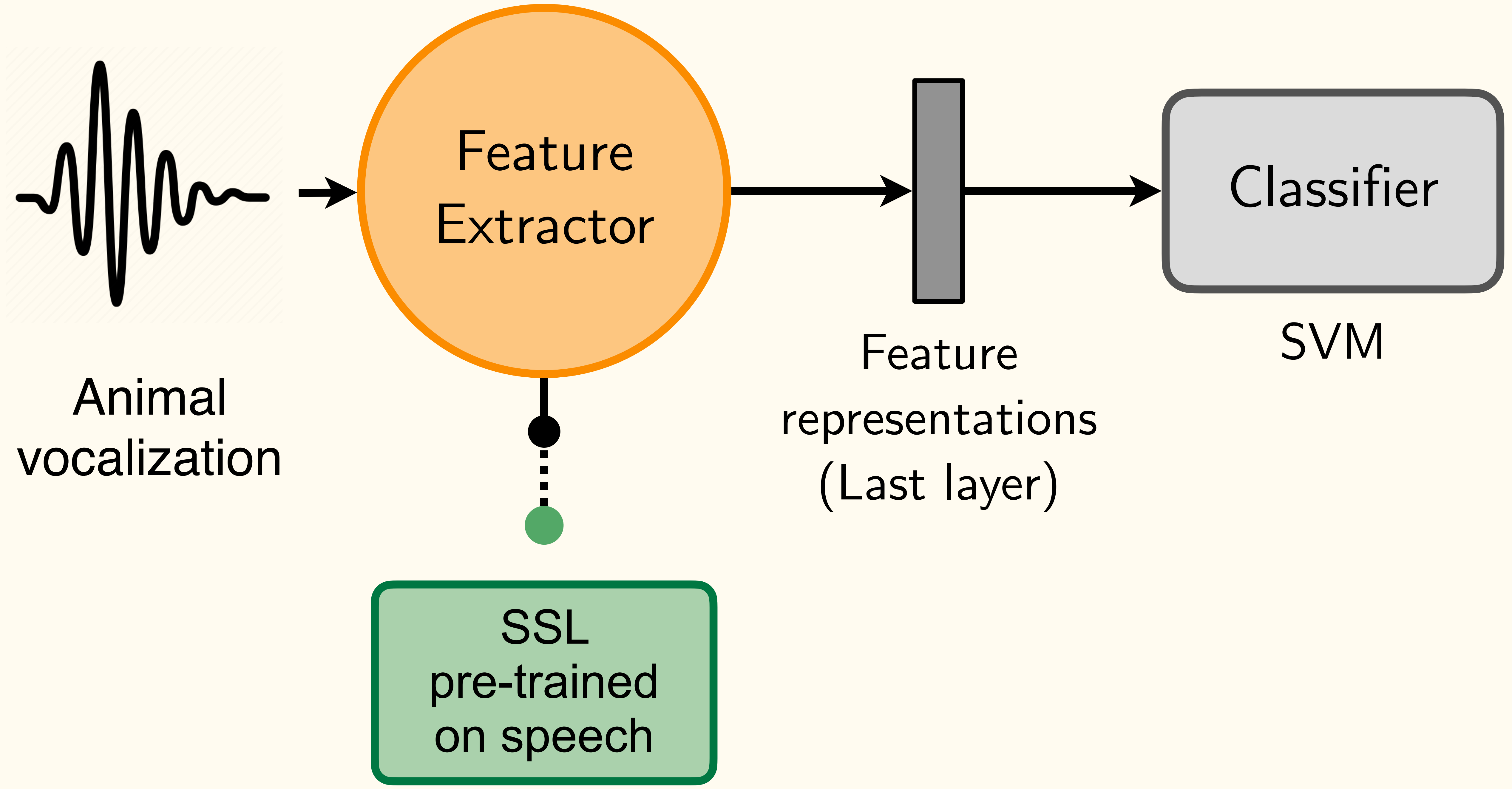
D	n_{CTID}	n_{CLID}	n_{SID}
D_1	11	10	–
D_2	7	8	2
D_3	12	8	2

Number of classes per task.



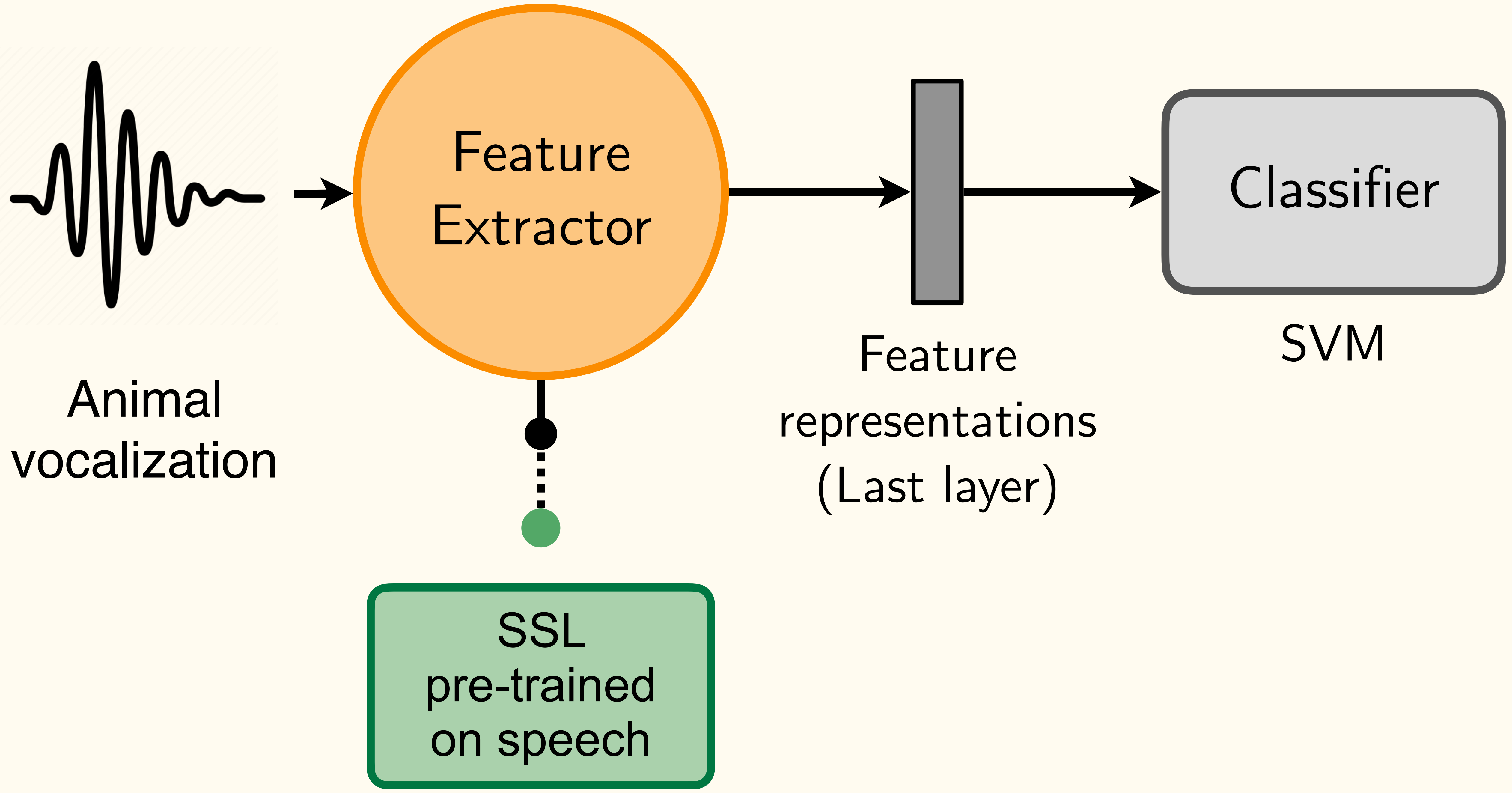
RQ1. Transferability of SSL Representations

SSL Embedding Spaces



SSL Embedding Spaces

- 11 selected SSL models pre-trained on speech.

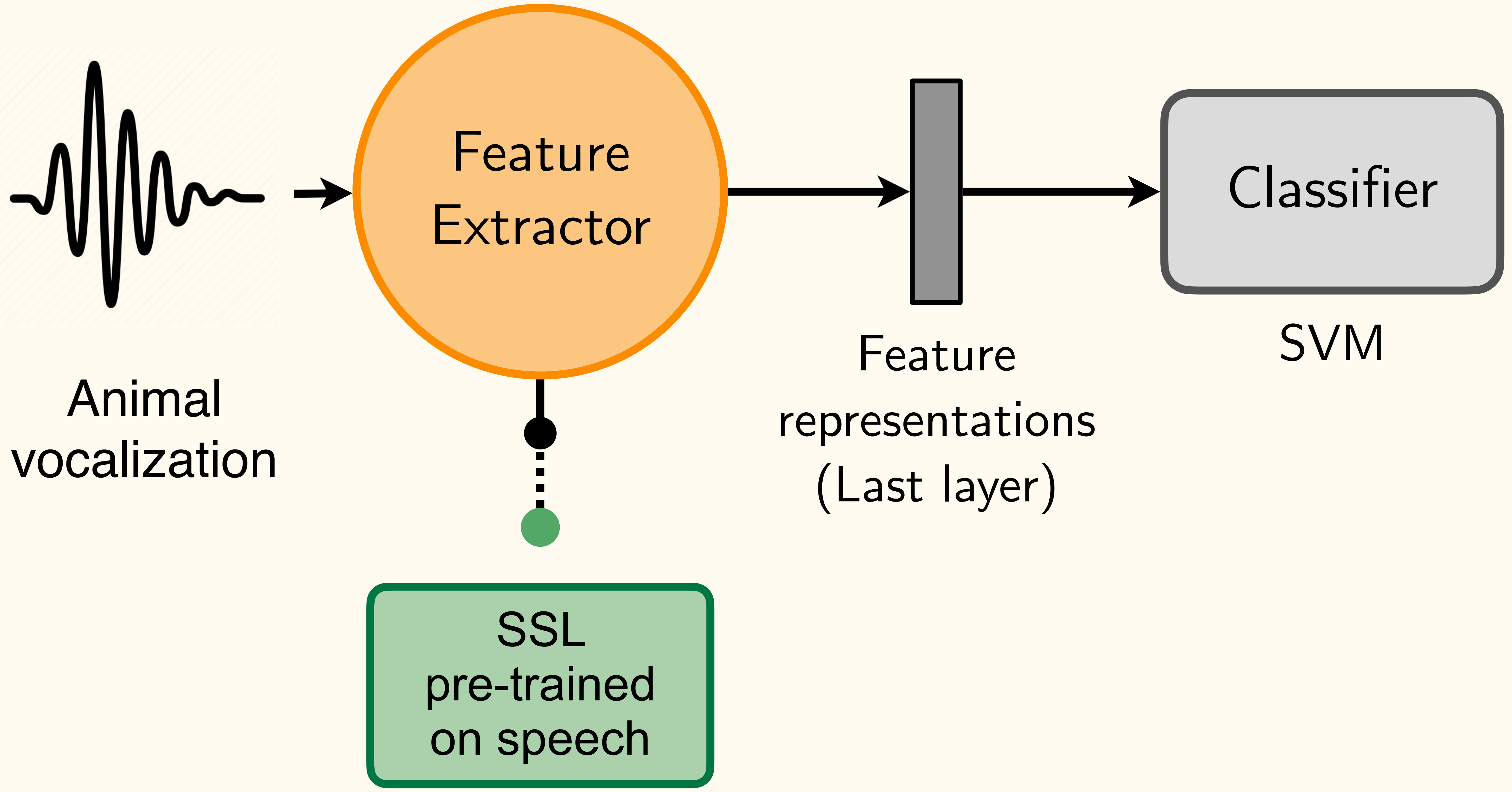


Model	Corpus
APC	LS 360
VQ-APC	LS 360
NPC	LS 360
Mockingjay	LS 100
TERA	LS 100
Mod-CPC	LL 60k
Wav2Vec2	LS 960
Hubert	LS 960
DistilHubert	LS 960
WavLM	LS 960
Data2Vec	LS 960

LS: LibriSpeech, LL: Libri-Light.

SSL Embedding Spaces

- 11 selected SSL models pre-trained on speech.
- Pre-trained using different types of pre-text tasks.

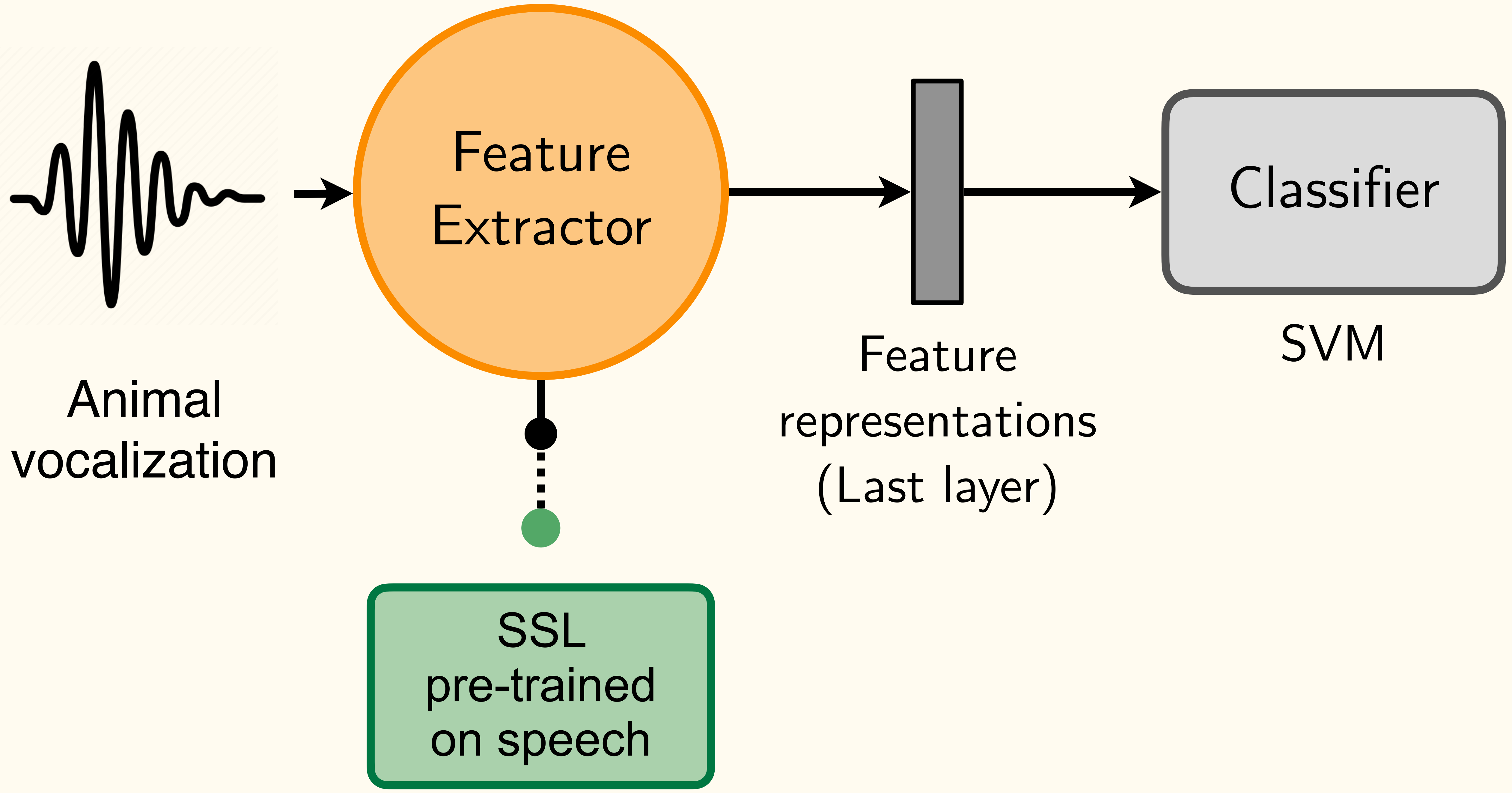


Model	Corpus
APC	LS 360
VQ-APC	LS 360
NPC	LS 360
Mockingjay	LS 100
TERA	LS 100
Mod-CPC	LL 60k
Wav2Vec2	LS 960
Hubert	LS 960
DistilHubert	LS 960
WavLM	LS 960
Data2Vec	LS 960

LS: LibriSpeech, LL: Libri-Light.

SSL Embedding Spaces Results

- 11 selected SSL models pre-trained on speech.
- Pre-trained using different types of pre-text tasks.
- Classify segments using SVM.



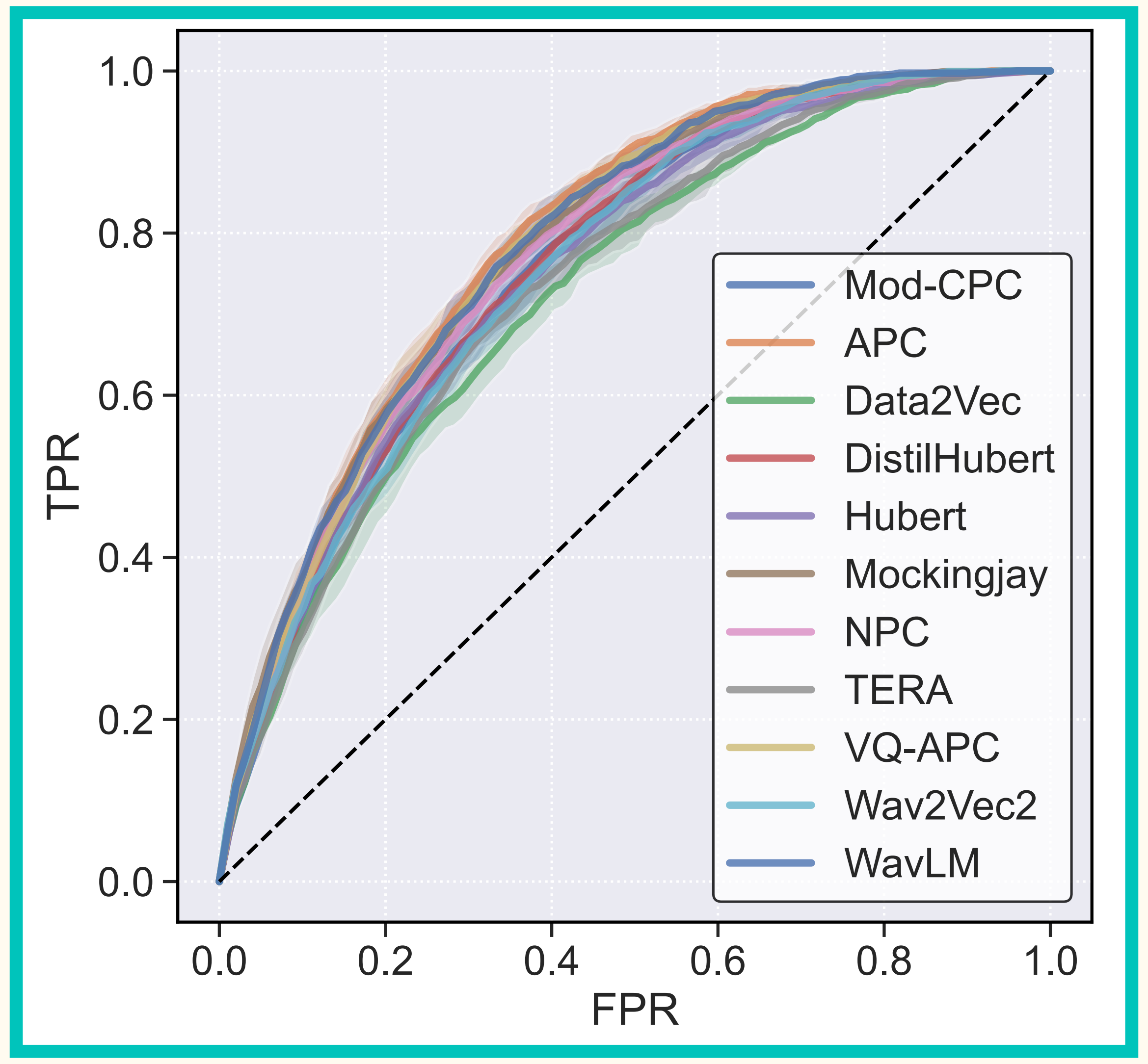
Caller detection task on D_1 (binary problem).

Model	Corpus	SVM
APC	**LS 360**	**79.16**
VQ-APC	LS 360	78.45
NPC	LS 360	77.32
Mockingjay	LS 100	78.44
TERA	LS 100	74.03
Mod-CPC	LL 60k	75.96
Wav2Vec2	LS 960	75.85
Hubert	LS 960	75.64
DistilHubert	LS 960	76.26
<u>WavLM</u>	<u>LS 960</u>	<u>78.60</u>
Data2Vec	LS 960	73.04

LS: LibriSpeech, LL: Libri-Light.
Macro AUC scores [%] on *Test* with 5-fold CV.

SSL Embedding Spaces Results

- 11 selected SSL models pre-trained on speech.
- Pre-trained using different types of pre-text tasks.
- Classify segments using SVM.
  - Representations capable of classifying animal calls.



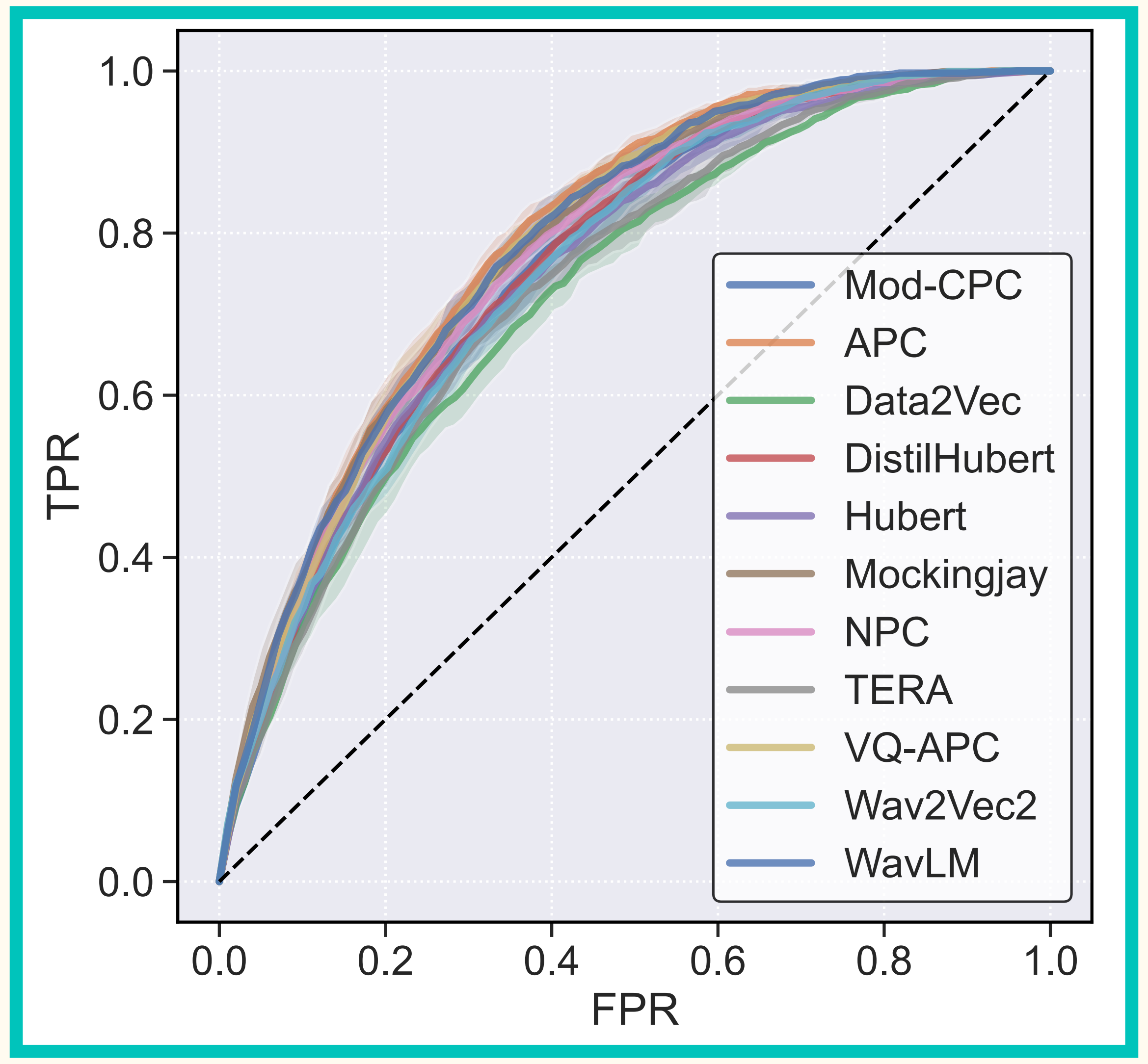
Caller detection task on D_1 (binary problem).

Model	Corpus	SVM
APC	**LS 360**	**79.16**
VQ-APC	LS 360	78.45
NPC	LS 360	77.32
Mockingjay	LS 100	78.44
TERA	LS 100	74.03
Mod-CPC	LL 60k	75.96
Wav2Vec2	LS 960	75.85
Hubert	LS 960	75.64
DistilHubert	LS 960	76.26
<u>WavLM</u>	<u>LS 960</u>	<u>78.60</u>
Data2Vec	LS 960	73.04

LS: LibriSpeech, LL: Libri-Light.
Macro AUC scores [%] on *Test* with 5-fold CV.

SSL Embedding Spaces Results

- 11 selected SSL models pre-trained on speech.
- Pre-trained using different types of pre-text tasks.
- Classify segments using SVM.
  - Representations capable of classifying animal calls.
  - WavLM: competitive results in speech and bioacoustics → used in follow-up work.



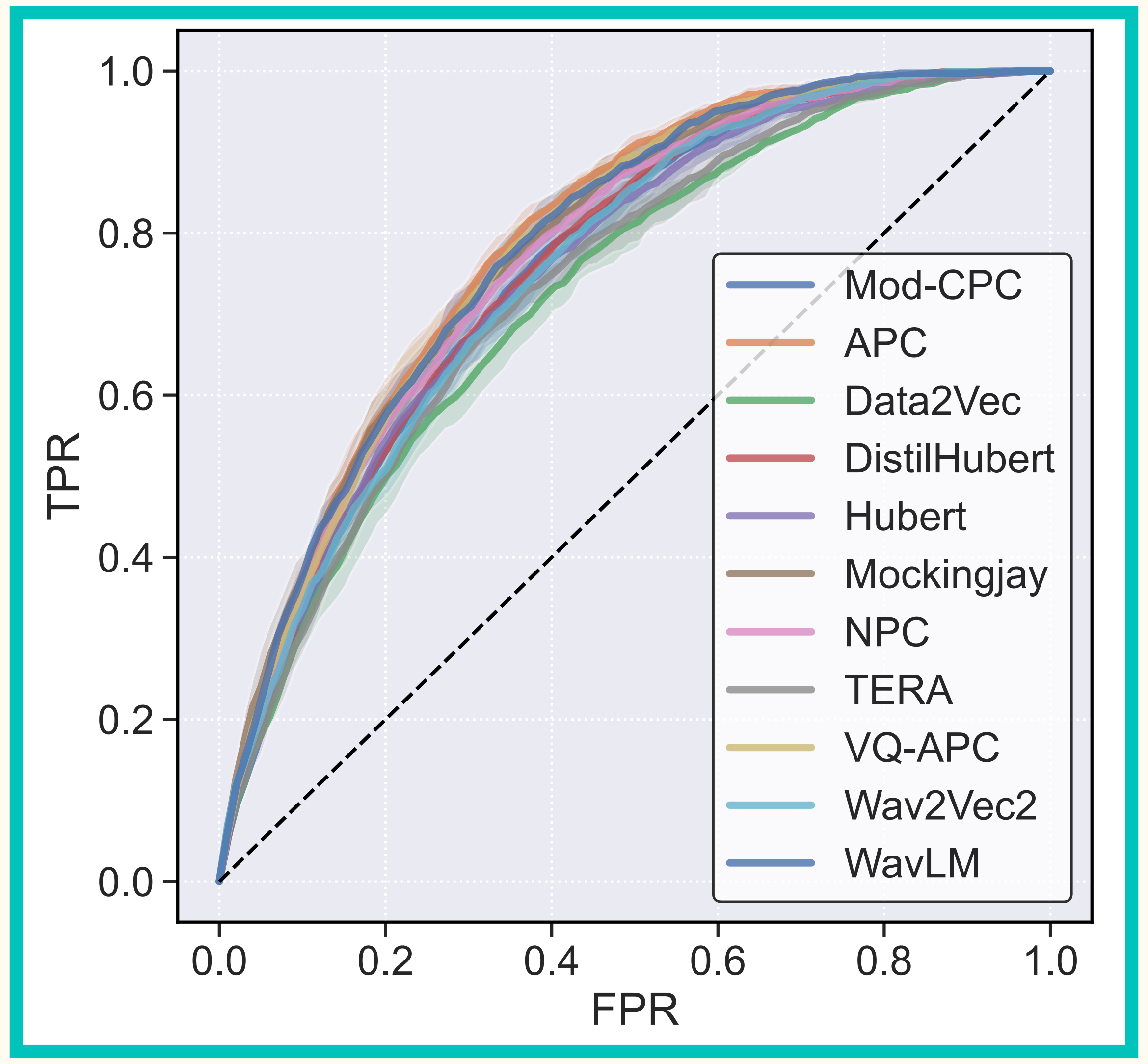
Caller detection task on D_1 (binary problem).

Model	Corpus	SVM
APC	**LS 360**	**79.16**
VQ-APC	LS 360	78.45
NPC	LS 360	77.32
Mockingjay	LS 100	78.44
TERA	LS 100	74.03
Mod-CPC	LL 60k	75.96
Wav2Vec2	LS 960	75.85
Hubert	LS 960	75.64
DistilHubert	LS 960	76.26
WavLM	LS 960	78.60
Data2Vec	LS 960	73.04

LS: LibriSpeech, LL: Libri-Light.
Macro AUC scores [%] on *Test* with 5-fold CV.

SSL Embedding Spaces Results

- 11 selected SSL models pre-trained on speech.
- Pre-trained using different types of pre-text tasks.
- Classify segments using SVM.
  - Representations capable of classifying animal calls.
  - WavLM: competitive results in speech and bioacoustics → used in follow-up work.
- Limitations:
  - Last layer.
  - Single dataset.

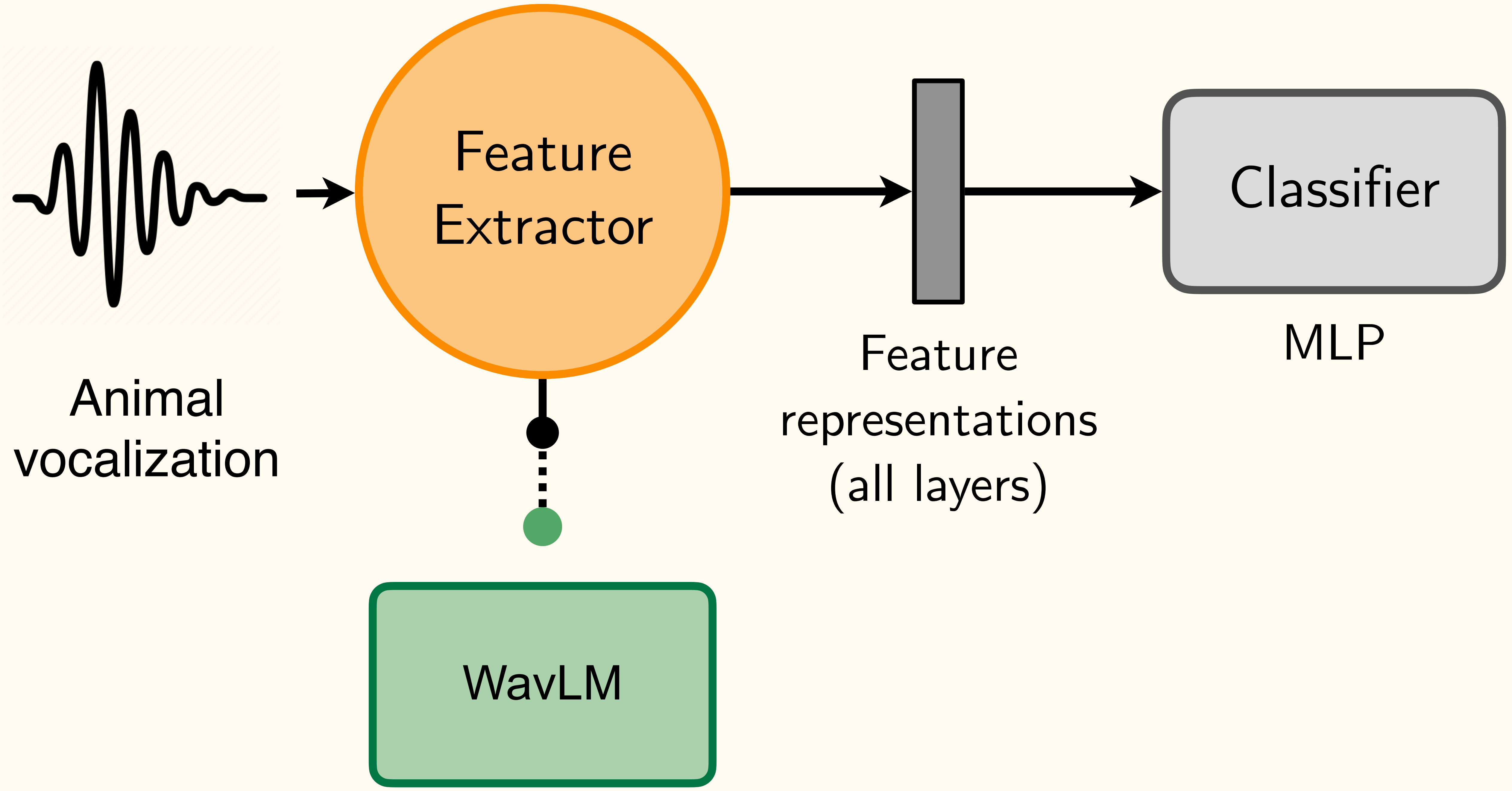


Caller detection task on D_1 (binary problem).

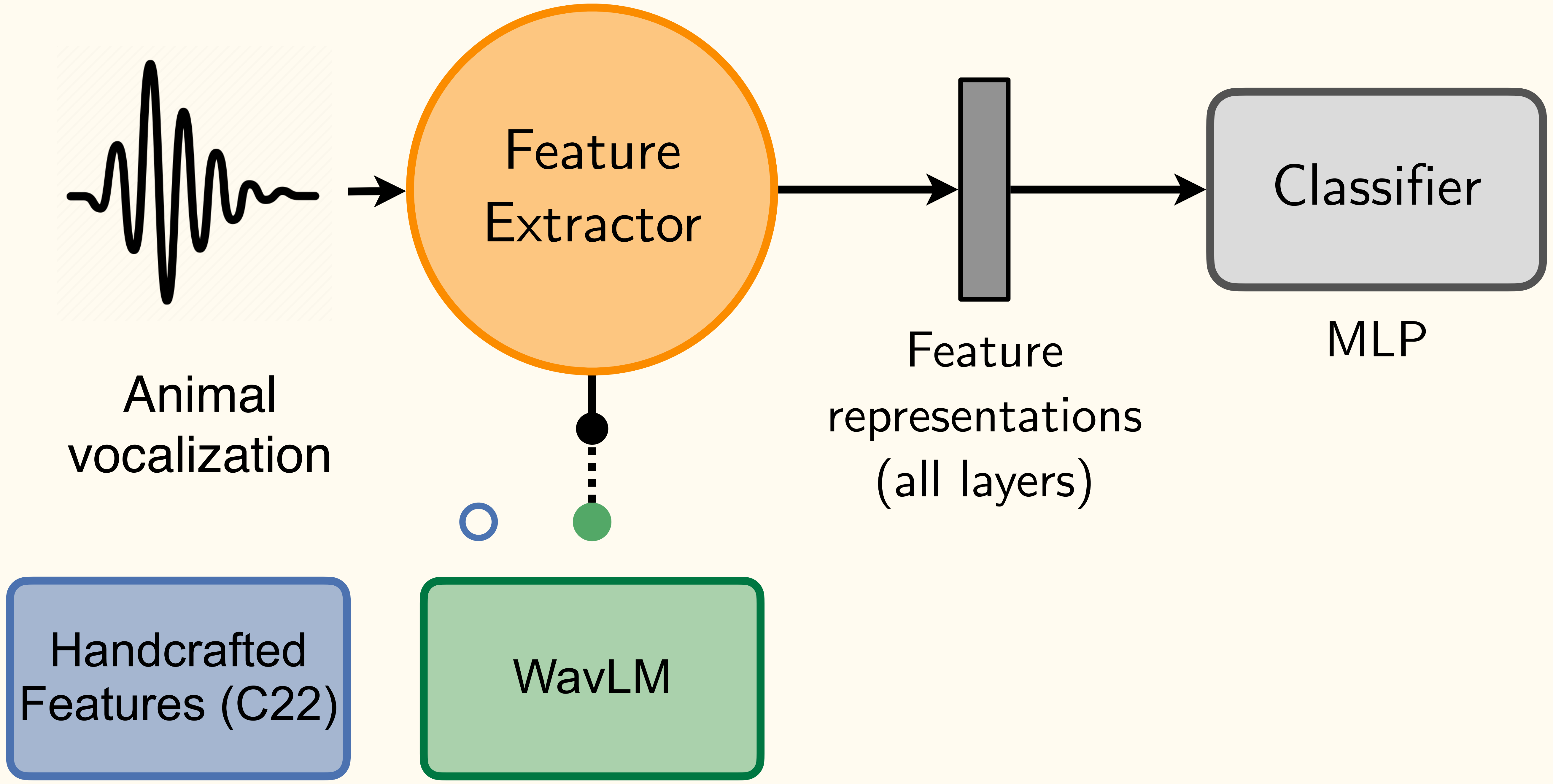
Model	**Corpus**	**SVM**
APC	**LS 360**	**79.16**
VQ-APC	LS 360	78.45
NPC	LS 360	77.32
Mockingjay	LS 100	78.44
TERA	LS 100	74.03
Mod-CPC	LL 60k	75.96
Wav2Vec2	LS 960	75.85
Hubert	LS 960	75.64
DistilHubert	LS 960	76.26
<u>WavLM</u>	<u>LS 960</u>	<u>78.60</u>
Data2Vec	LS 960	73.04

LS: LibriSpeech, LL: Libri-Light.
Macro AUC scores [%] on *Test* with 5-fold CV.

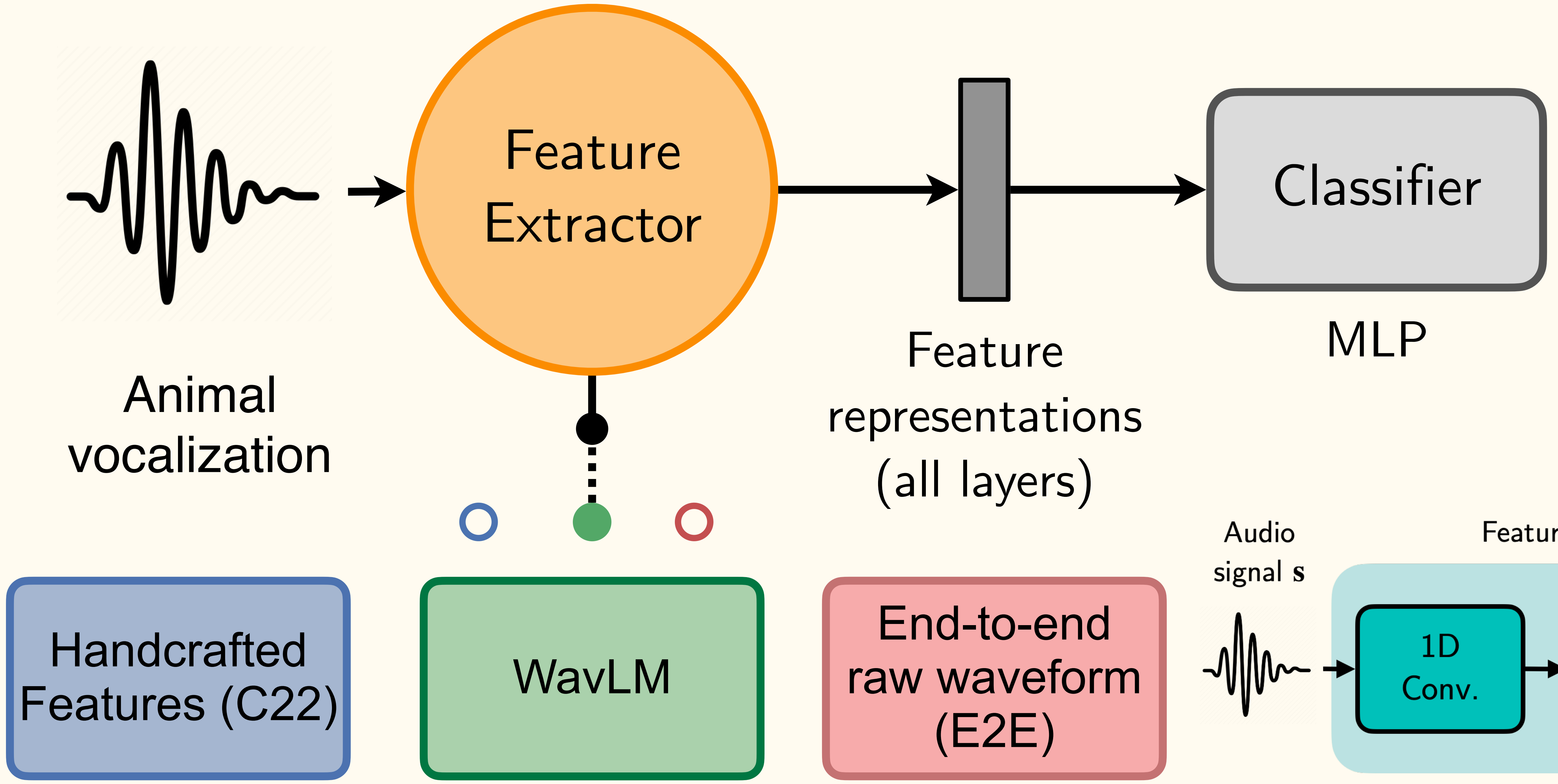
Classification Results

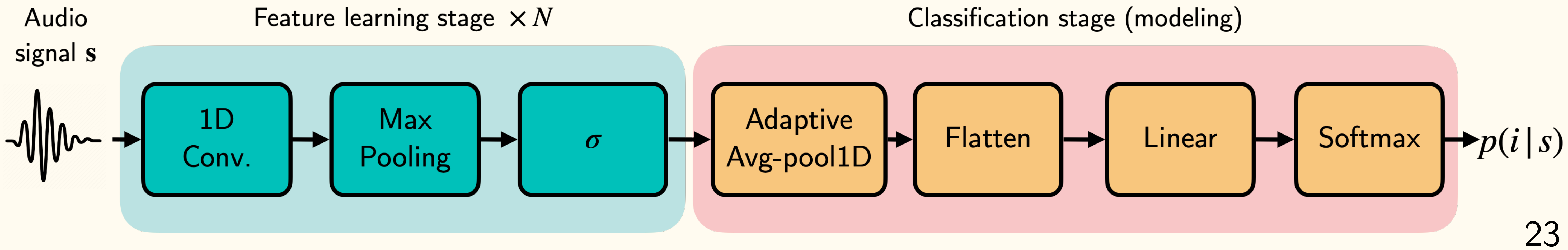


Classification Results



Classification Results

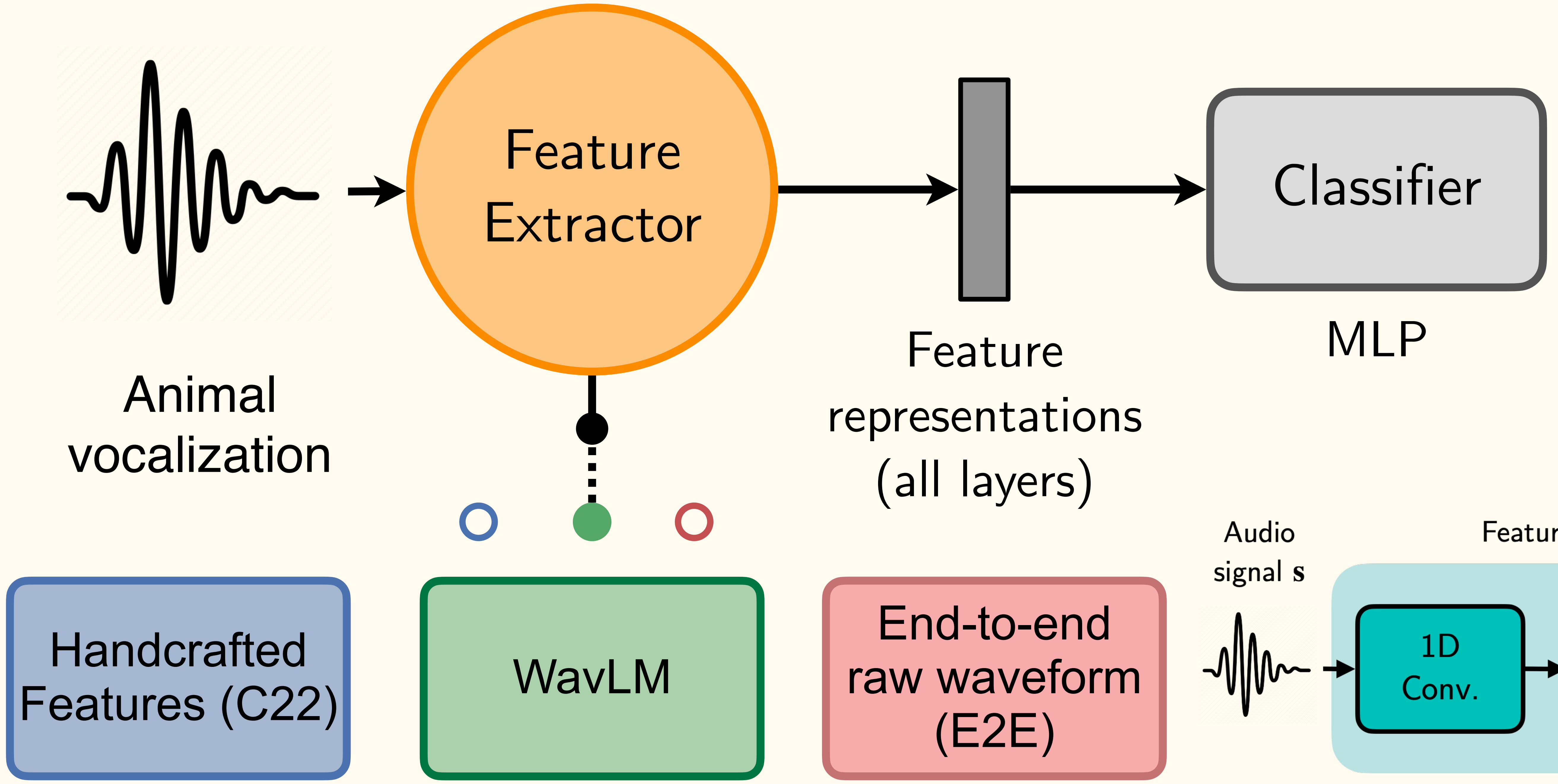


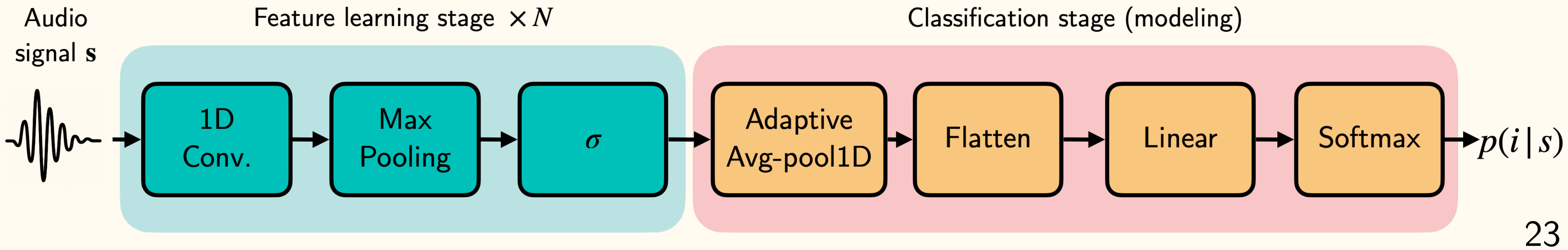


Classification Results

UAR [%] scores on *Test* on features at 16 kHz.
Best layer's results are shown for WavLM.

Dataset	Feature	CTID	CLID	SID
D_1	C22	37.72	34.54	N/A
	WavLM	**60.10**	**67.47**	N/A
	E2E	53.03	59.94	N/A

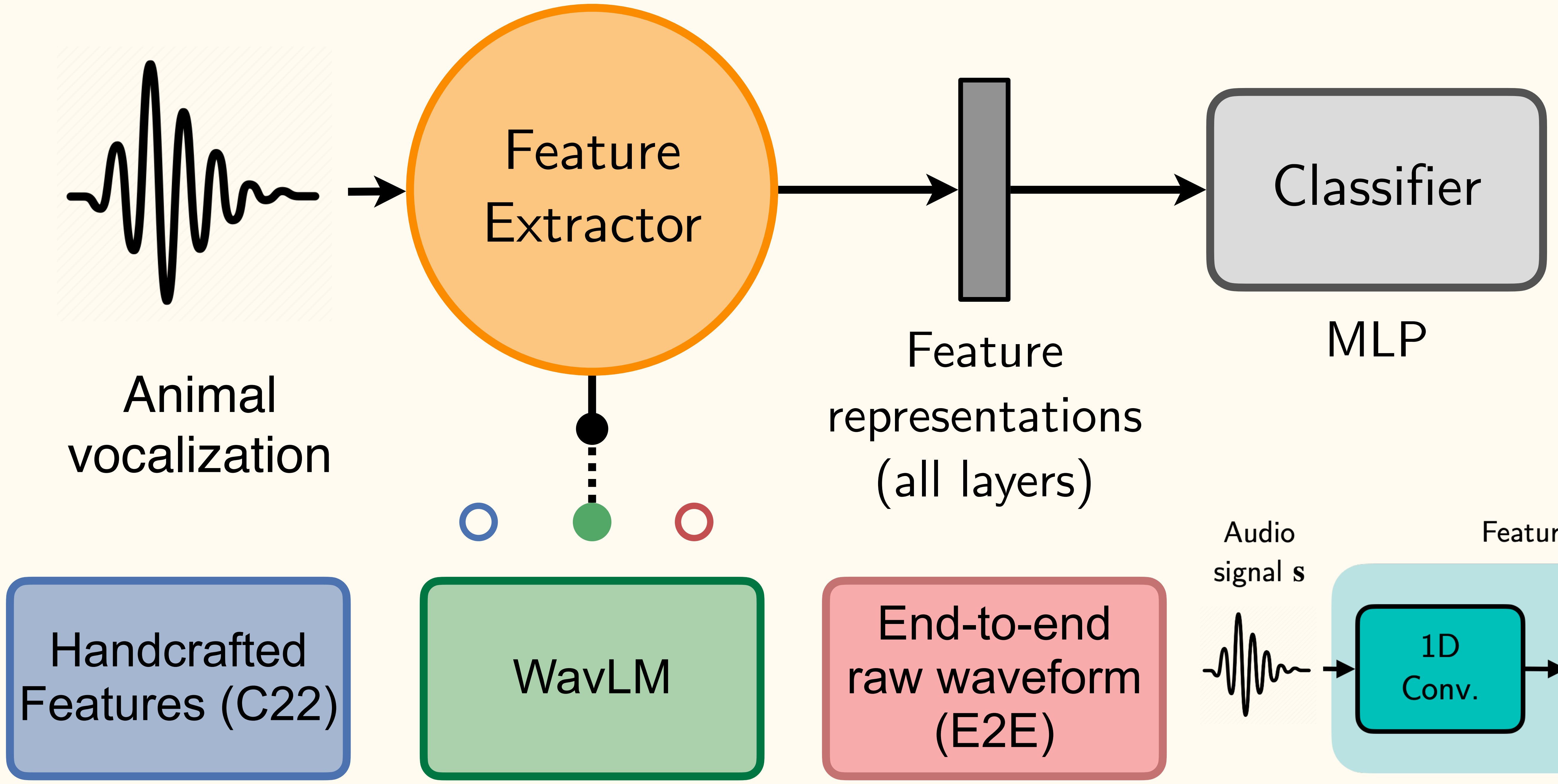


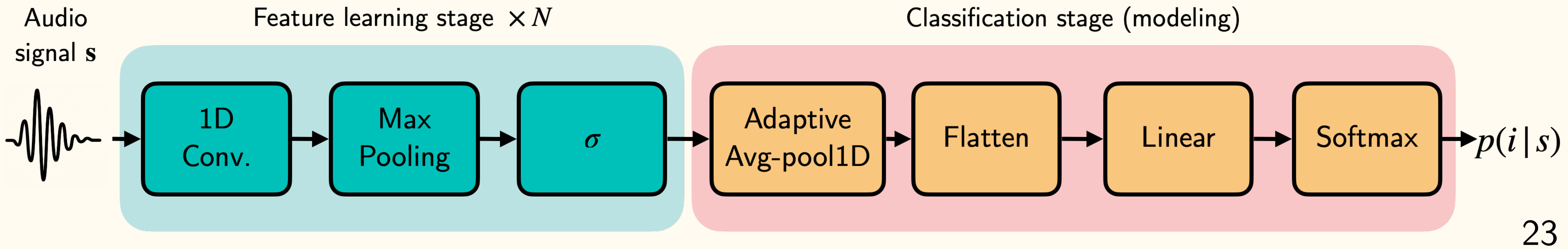


Classification Results

UAR [%] scores on *Test* on features at 16 kHz.
Best layer's results are shown for WavLM.

Dataset	Feature	CTID	CLID	SID
D_1	C22	37.72	34.54	N/A
	WavLM	**60.10**	**67.47**	N/A
	E2E	53.03	59.94	N/A
D_2	C22	35.65	35.32	58.14
	WavLM	**56.77**	**46.05**	**63.80**
	E2E	37.65	36.21	60.15



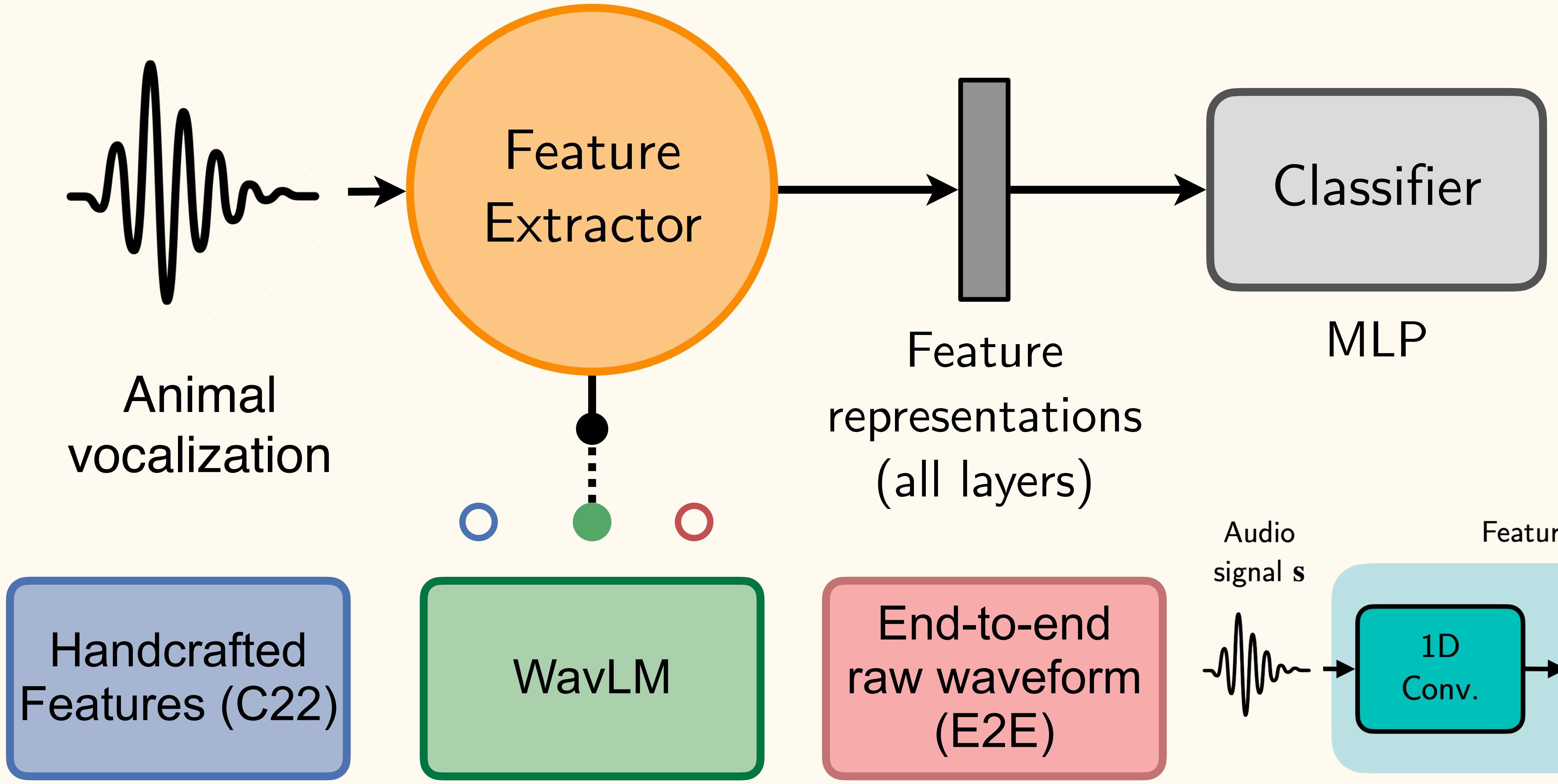


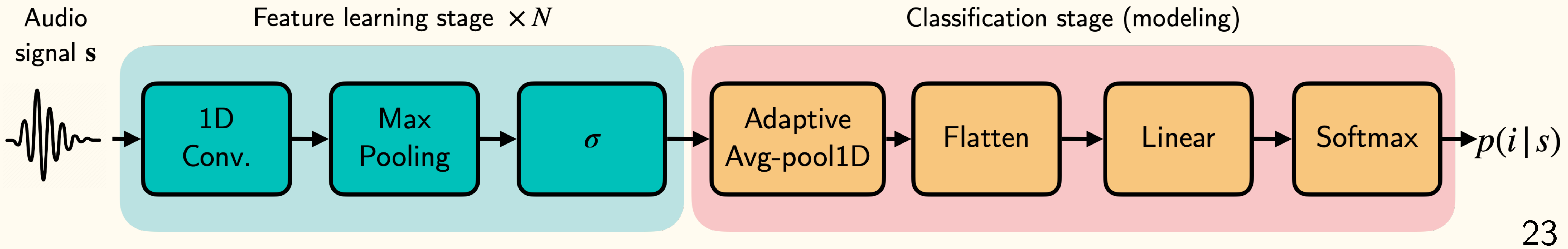
Classification Results

- WavLM: ‘best’ layer yields robust performances on all 3 datasets and tasks.

UAR [%] scores on *Test* on features at 16 kHz.
Best layer’s results are shown for WavLM.

Dataset	Feature	CTID	CLID	SID
D_1	C22	37.72	34.54	N/A
	WavLM	**60.10**	**67.47**	N/A
	E2E	53.03	59.94	N/A
D_2	C22	35.65	35.32	58.14
	WavLM	**56.77**	**46.05**	**63.80**
	E2E	37.65	36.21	60.15
D_3	C22	52.59	39.43	57.32
	WavLM	**80.38**	**55.58**	**74.26**
	E2E	66.24	31.31	56.59



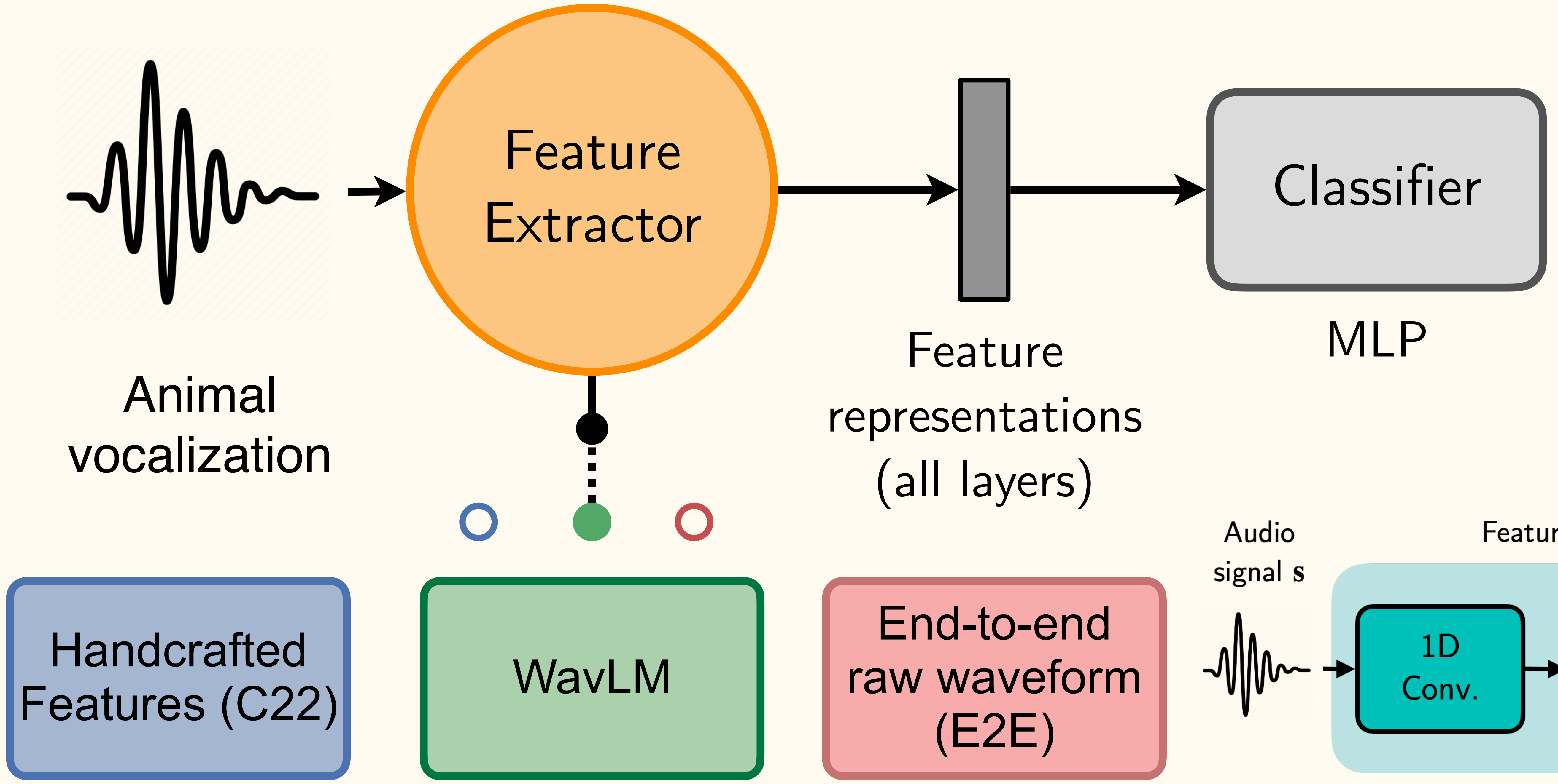


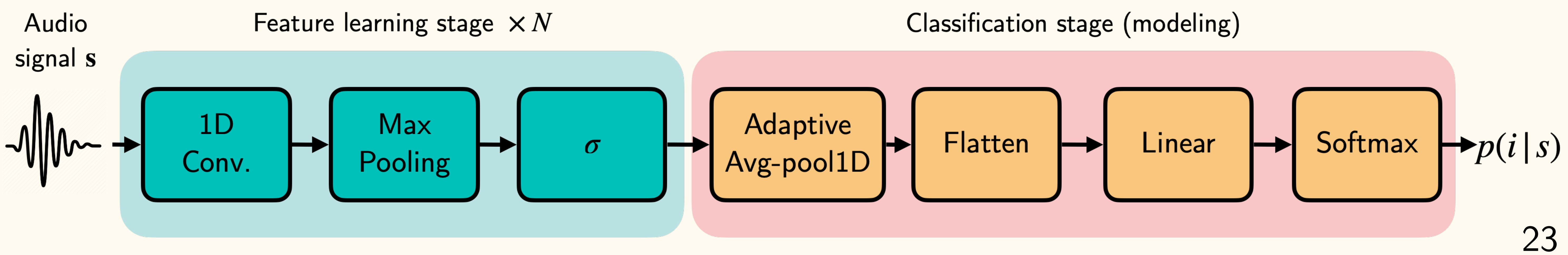
Classification Results

- WavLM: ‘best’ layer yields robust performances on all 3 datasets and tasks.
- What about other layers ?

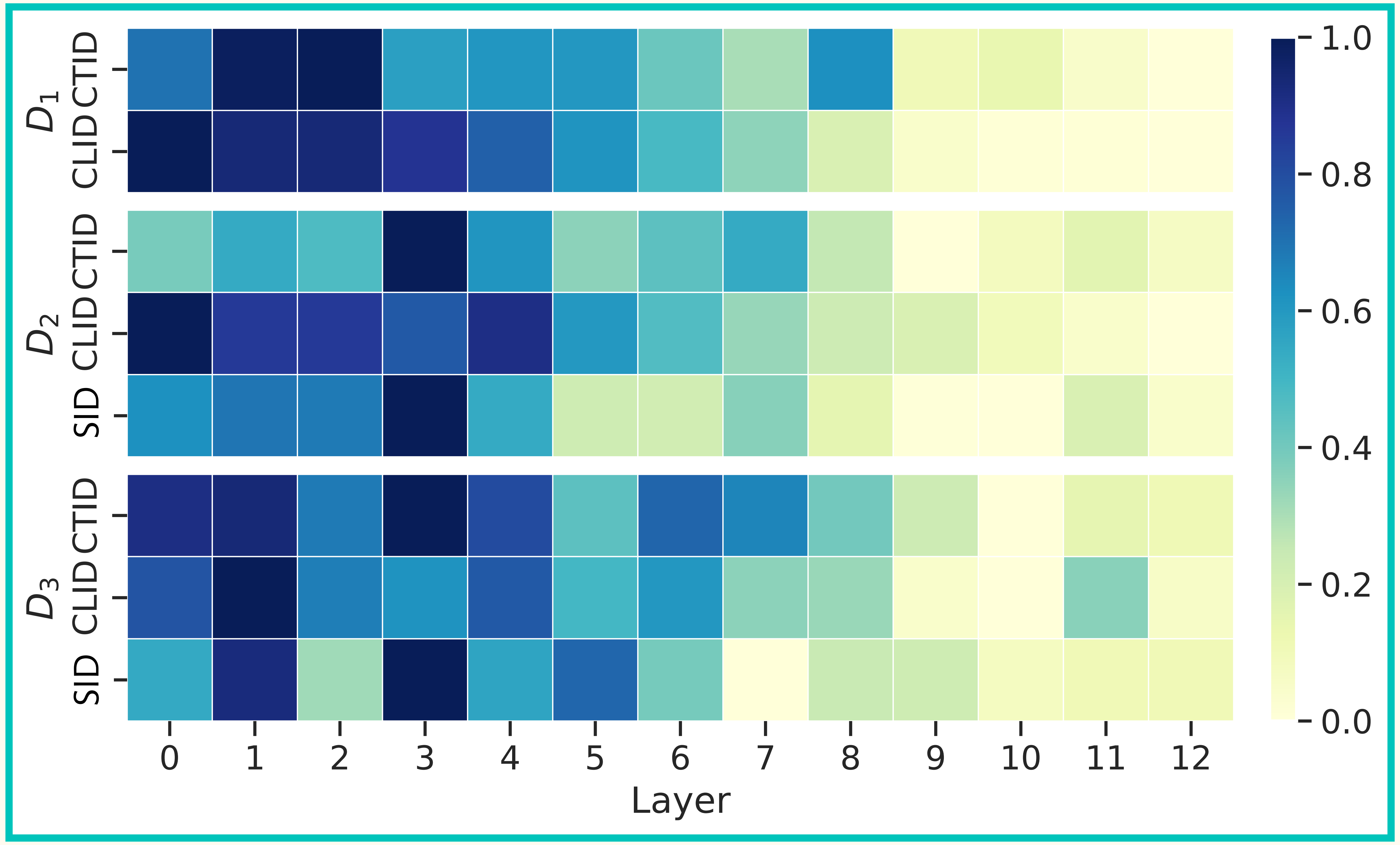
UAR [%] scores on *Test* on features at 16 kHz.
Best layer’s results are shown for WavLM.

Dataset	Feature	CTID	CLID	SID
D_1	C22	37.72	34.54	N/A
	WavLM	**60.10**	**67.47**	N/A
	E2E	53.03	59.94	N/A
D_2	C22	35.65	35.32	58.14
	WavLM	**56.77**	**46.05**	**63.80**
	E2E	37.65	36.21	60.15
D_3	C22	52.59	39.43	57.32
	WavLM	**80.38**	**55.58**	**74.26**
	E2E	66.24	31.31	56.59





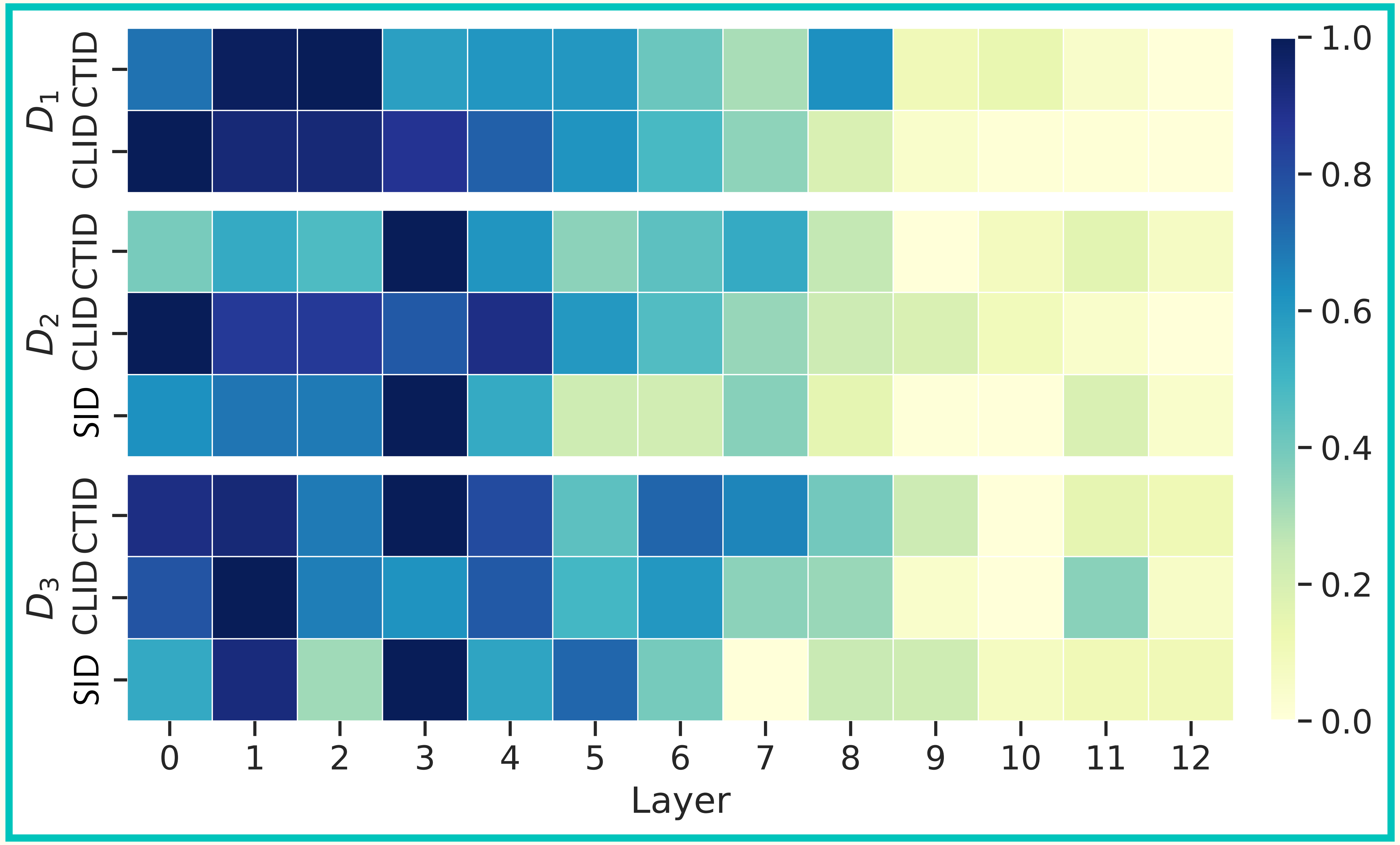
WavLM Layer Analysis



- Layer-wise UAR scores of WavLM features, normalized [0,1] per task.
- Layer 0 corresponds to the output of the CNN encoder.
- Darker regions indicate a higher performance.

WavLM Layer Analysis

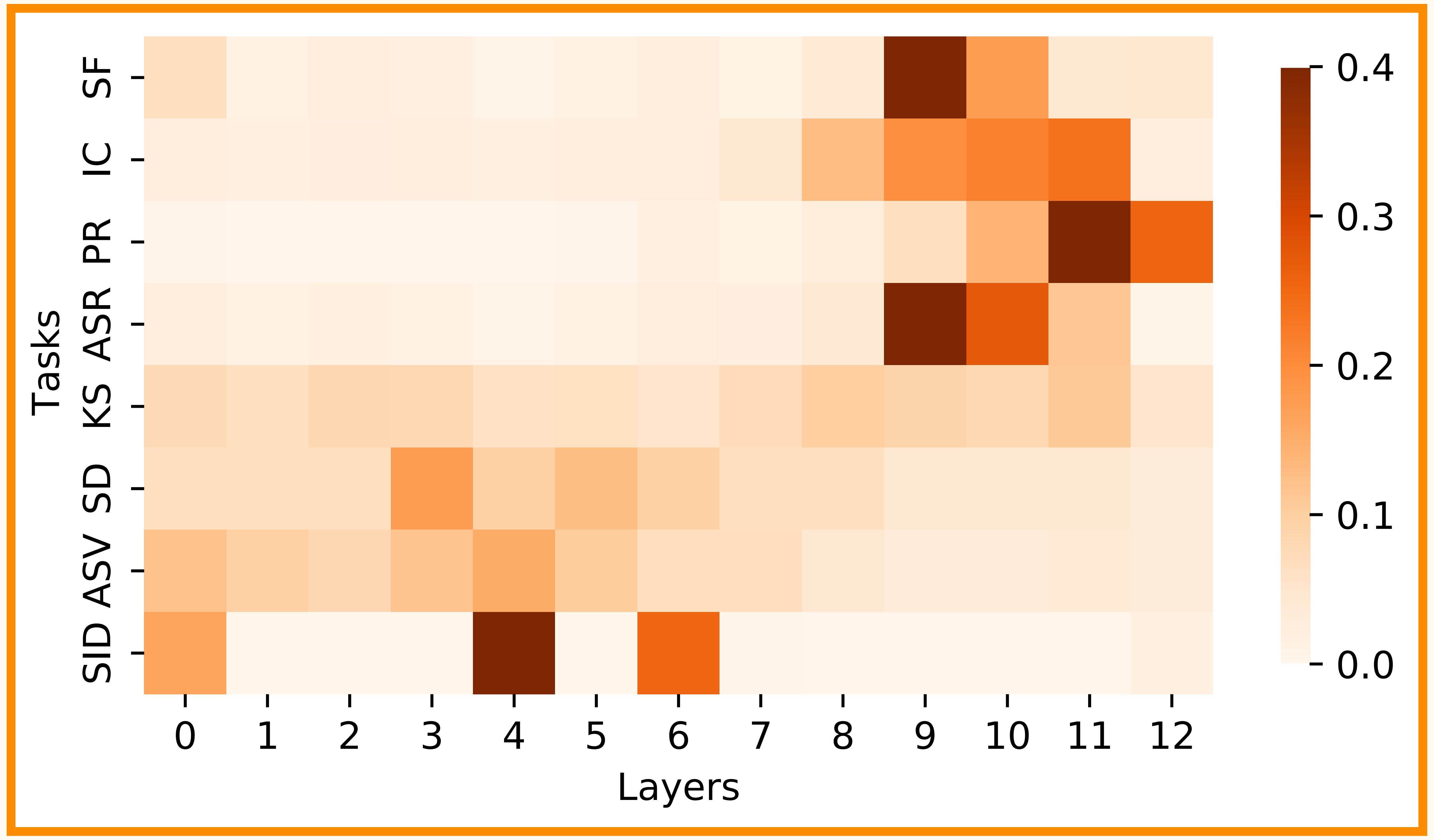
- **Trend**: lower layers are more salient representations.



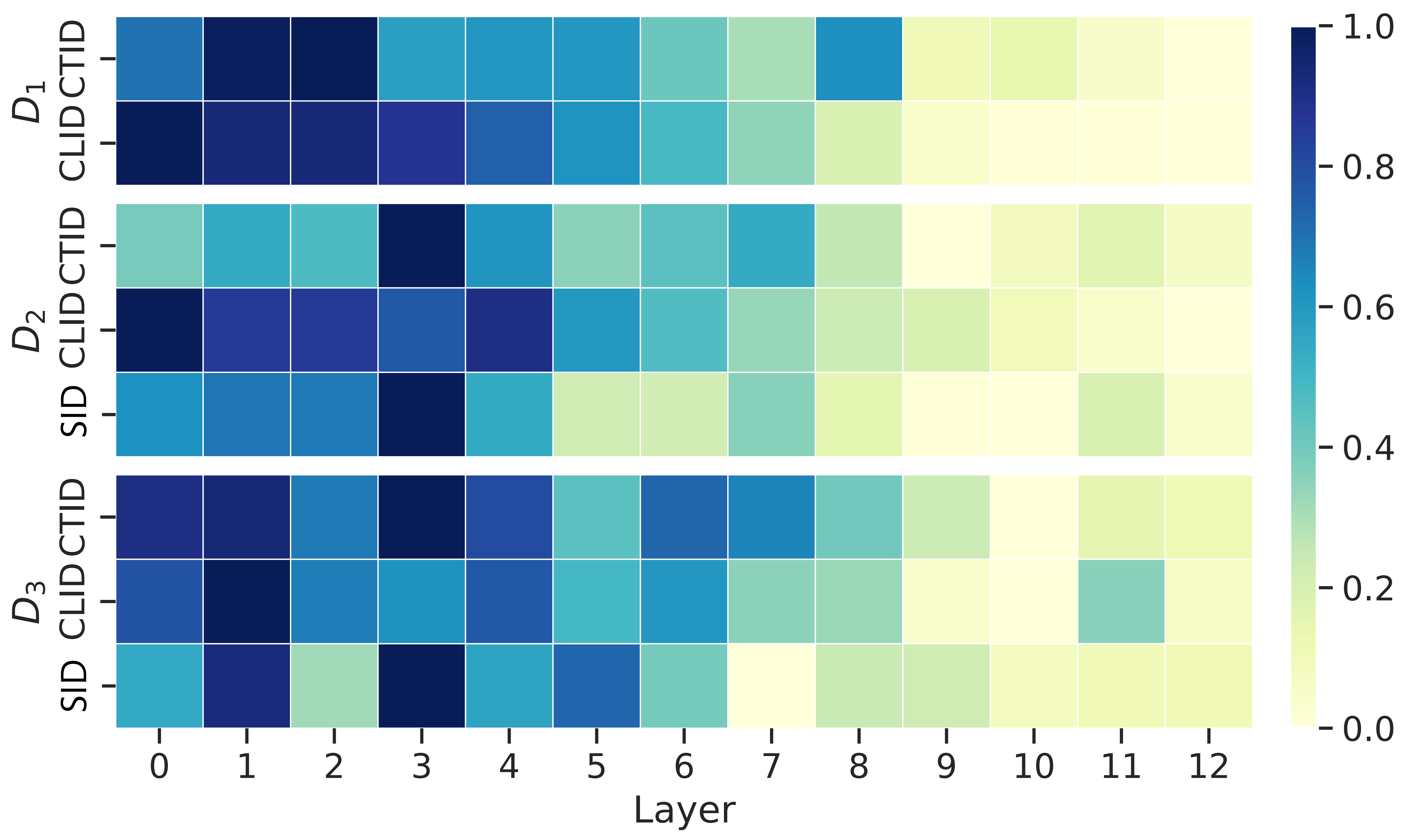
- Layer-wise UAR scores of WavLM features, normalized [0,1] per task.
-Layer 0 corresponds to the output of the CNN encoder.
- Darker regions indicate a higher performance.

WavLM Layer Analysis

- **Trend**: lower layers are more salient representations.



- WavLM layer importance distribution per task.
- Softmax normalization per row.
- WavLM base+ model.

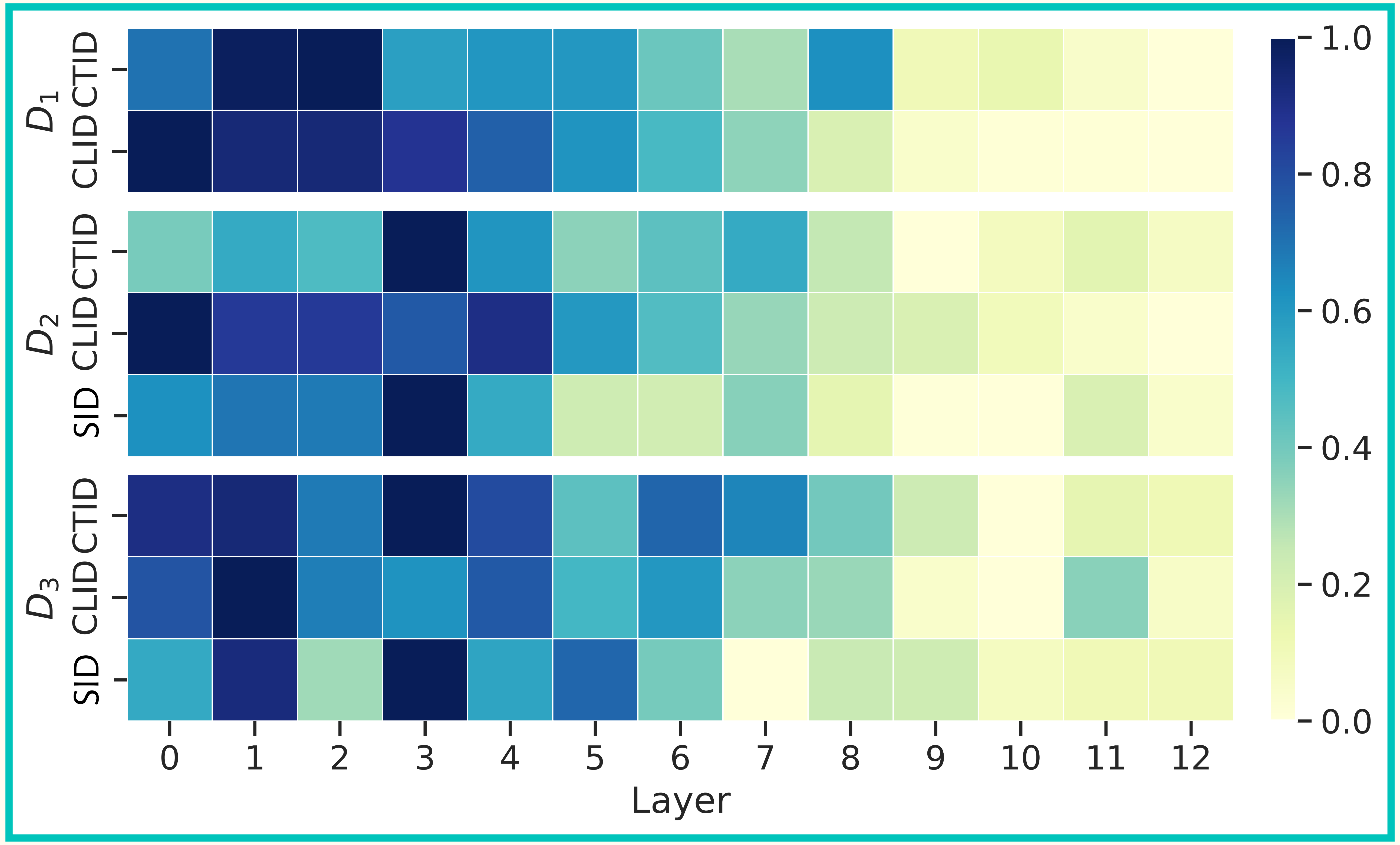


- Layer-wise UAR scores of WavLM features, normalized [0,1] per task.
-Layer 0 corresponds to the output of the CNN encoder.
- Darker regions indicate a higher performance.

[1] *Chen et al., WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.* (2022). IEEE Journal of Selected Topics in Signal Processing.

WavLM Layer Analysis

- **Trend**: lower layers are more salient representations.
- WavLM: lower layers tend to capture fundamental acoustic features; later layers perform on linguistic tasks[1].

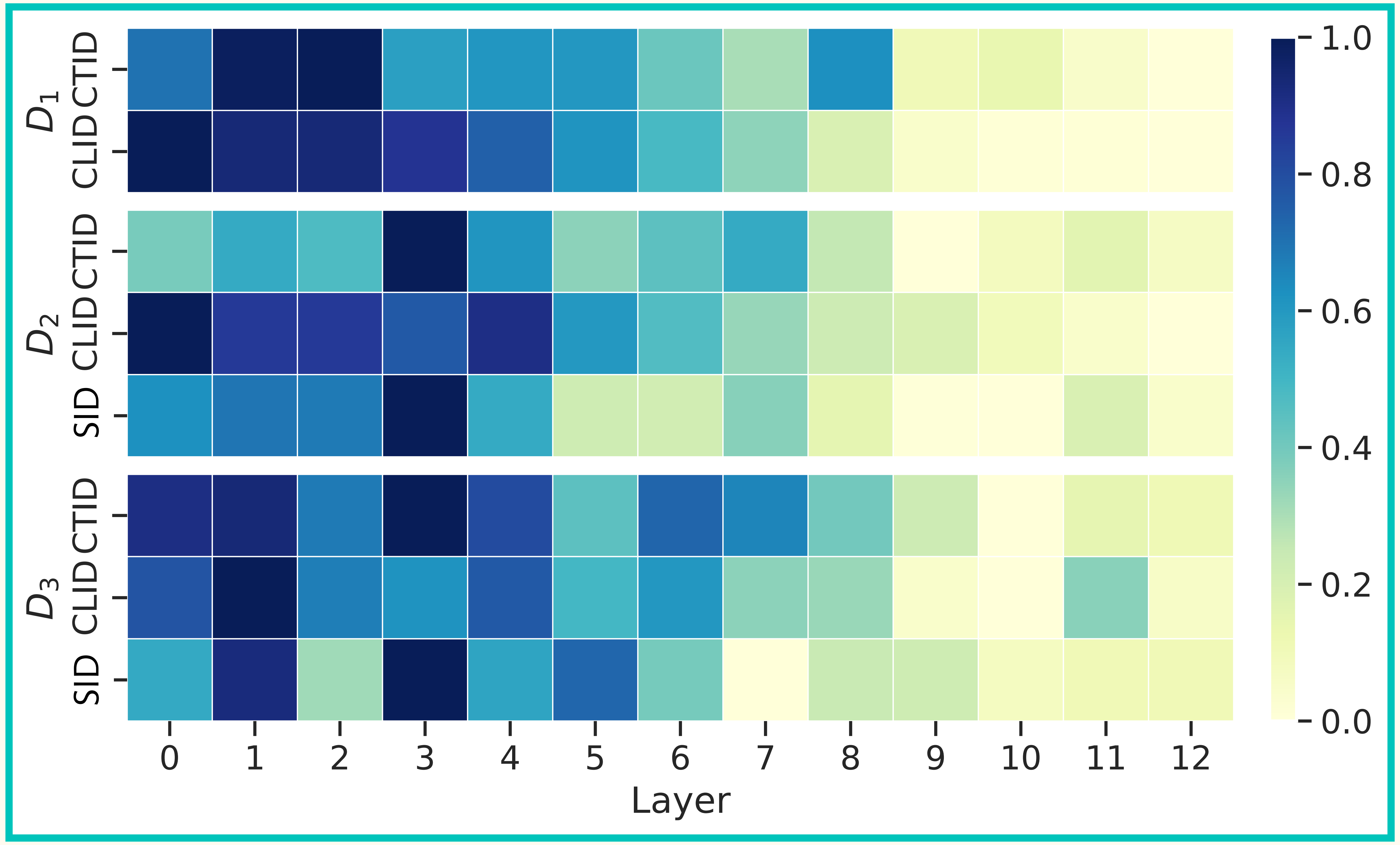


- Layer-wise UAR scores of WavLM features, normalized [0,1] per task.
-Layer 0 corresponds to the output of the CNN encoder.
- Darker regions indicate a higher performance.

[1] *Chen et al., WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.* (2022). IEEE Journal of Selected Topics in Signal Processing.

WavLM Layer Analysis

- **Trend**: lower layers are more salient representations.
- WavLM: lower layers tend to capture fundamental acoustic features; later layers perform on linguistic tasks[1].
  - Lower layers: generalize better to other acoustic domains, e.g. marmoset calls.

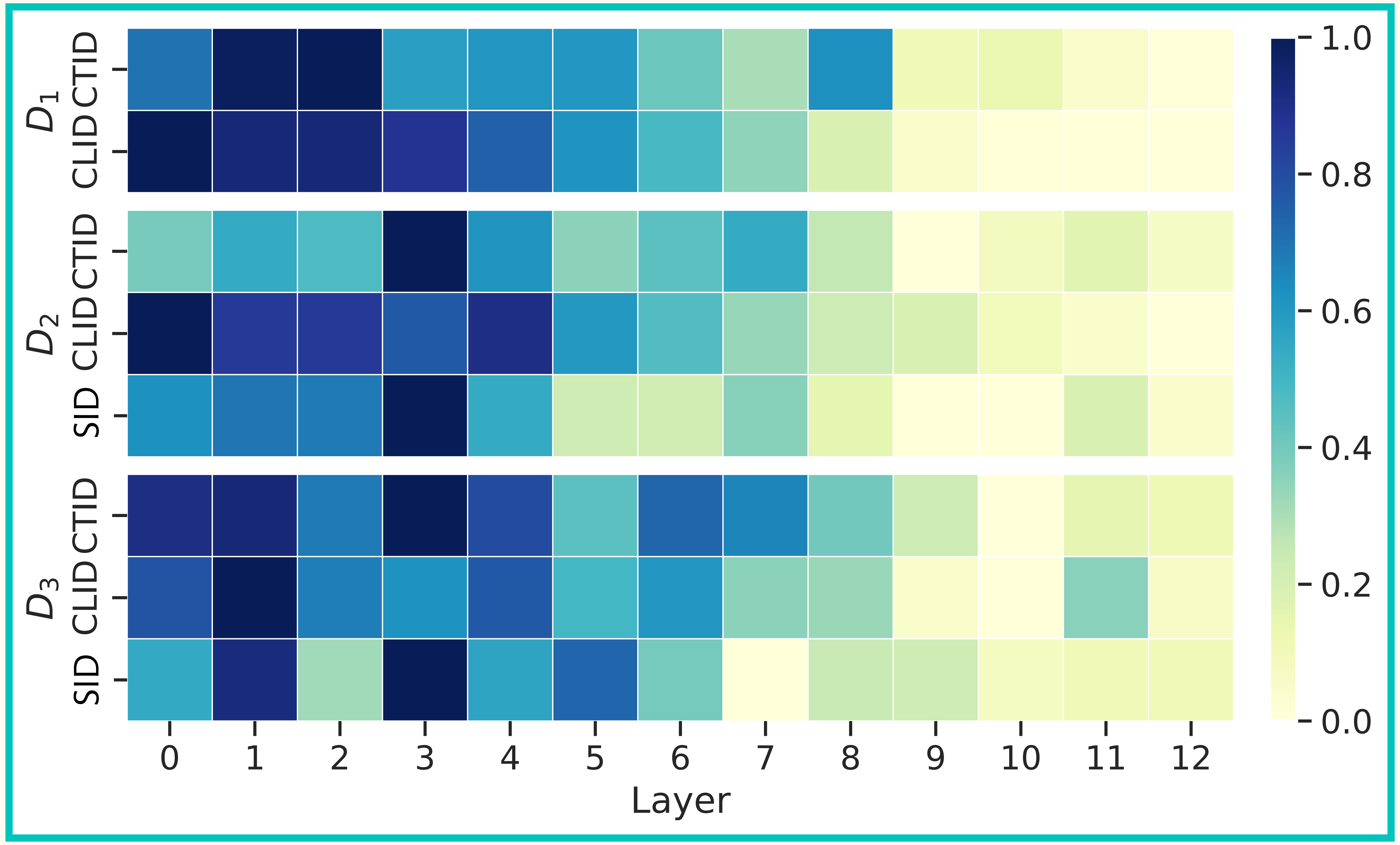


- Layer-wise UAR scores of WavLM features, normalized [0,1] per task.
-Layer 0 corresponds to the output of the CNN encoder.
- Darker regions indicate a higher performance.

[1] *Chen et al., WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.* (2022). IEEE Journal of Selected Topics in Signal Processing.

WavLM Layer Analysis

- **Trend**: lower layers are more salient representations.
- WavLM: lower layers tend to capture fundamental acoustic features; later layers perform on linguistic tasks[1].
  - Lower layers: generalize better to other acoustic domains, e.g. marmoset calls.
  - Later layers: appear more specialized for human speech, and consequently much less transferable to bioacoustics.

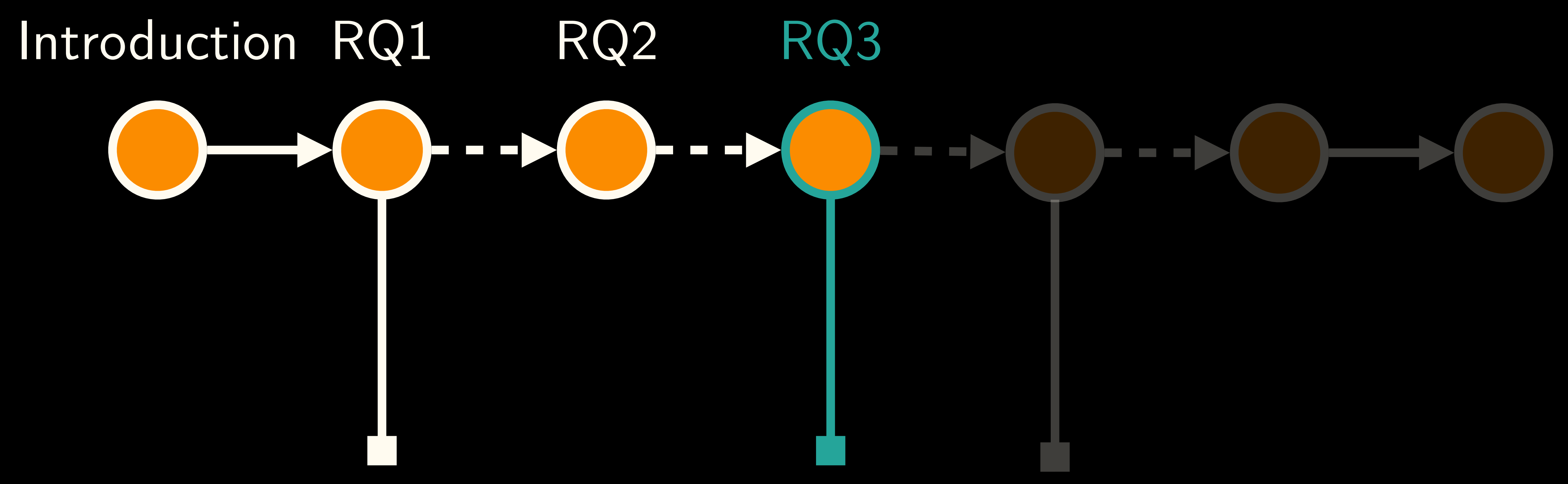


- Layer-wise UAR scores of WavLM features, normalized [0,1] per task.
-Layer 0 corresponds to the output of the CNN encoder.
- Darker regions indicate a higher performance.

[1] *Chen et al., WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.* (2022). IEEE Journal of Selected Topics in Signal Processing.

Key Takeaways

- Representations of speech SSLs can classify bioacoustics vocalizations, even without fine-tuning.
- Lower layers of these SSLs are significantly more salient than later layers for the conducted bioacoustics tasks.

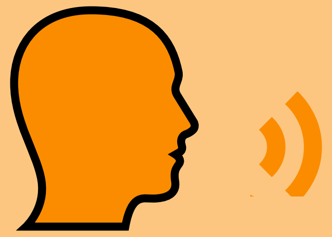


RQ3. Pre-Training Domain Analysis

Pre-Training Domain - Feature Representations

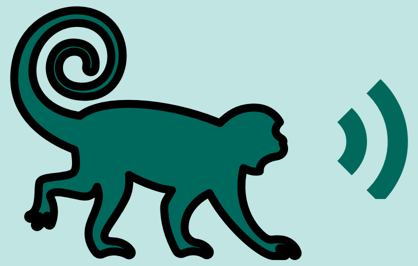
Pre-training

on human speech



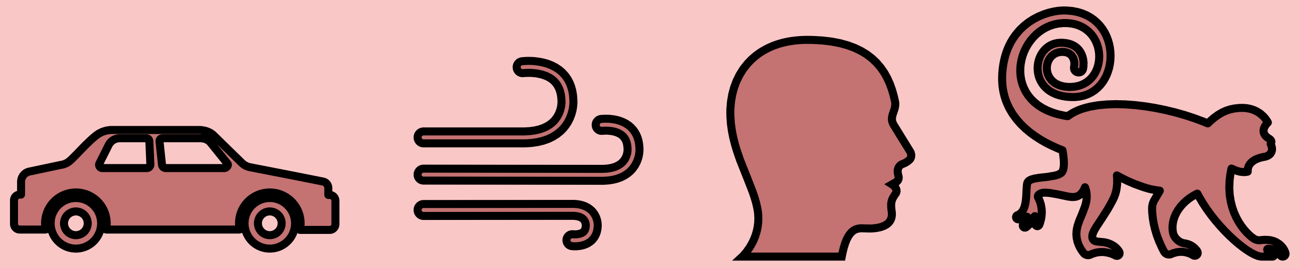
Pre-training

on bioacoustics



Pre-training

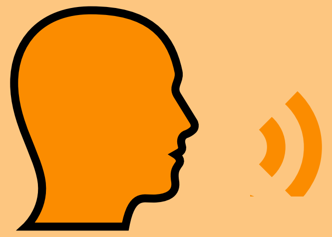
on general audio



Pre-Training Domain - Feature Representations

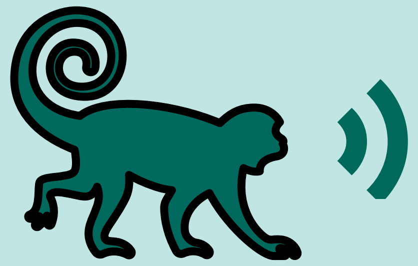
Pre-training

on human speech



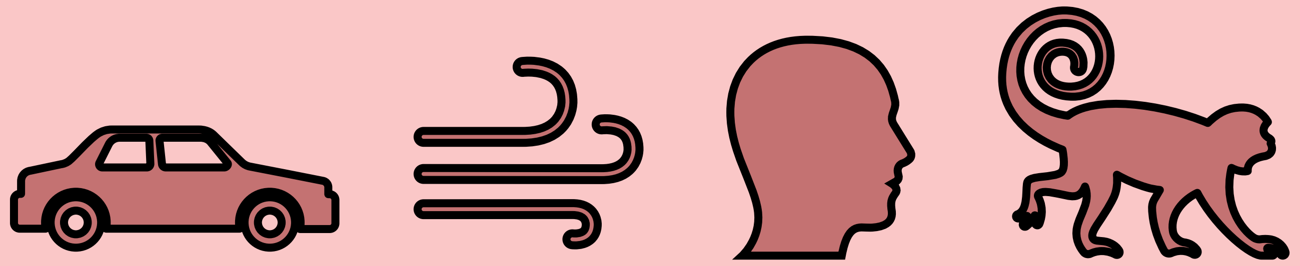
Pre-training

on bioacoustics



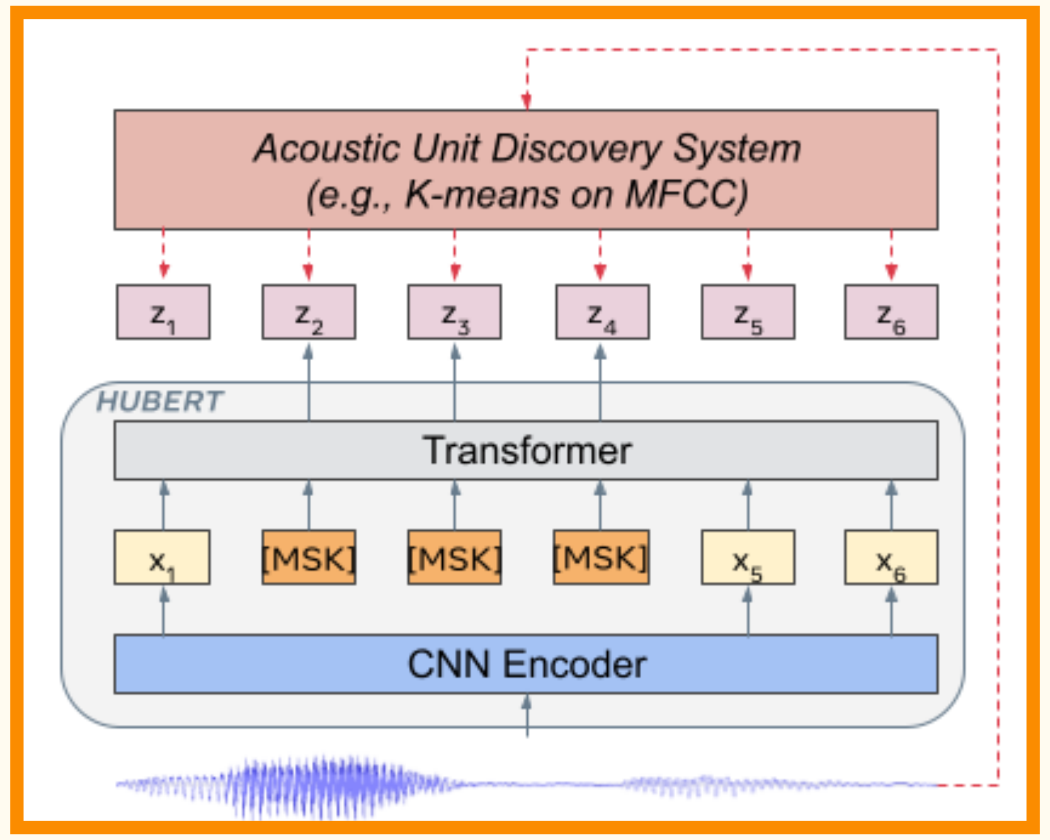
Pre-training

on general audio



HuBERT:

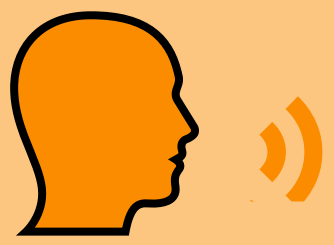
- Librispeech 960h.
- Similar pre-training as WavLM.



Pre-Training Domain - Feature Representations

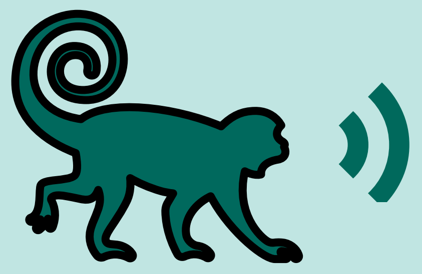
Pre-training

on human speech



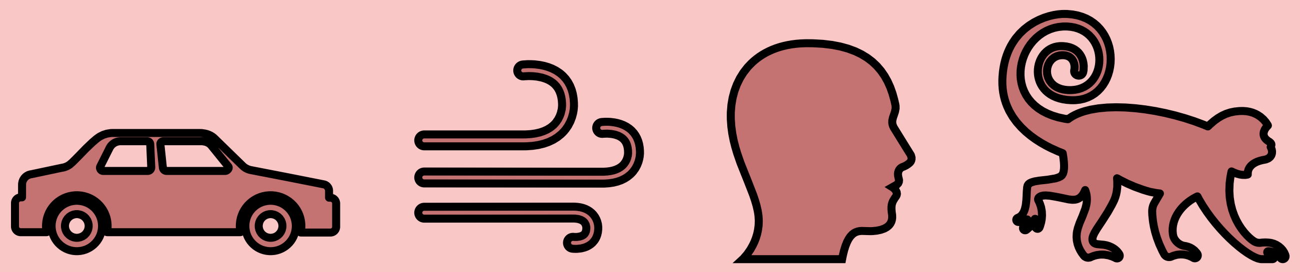
Pre-training

on bioacoustics



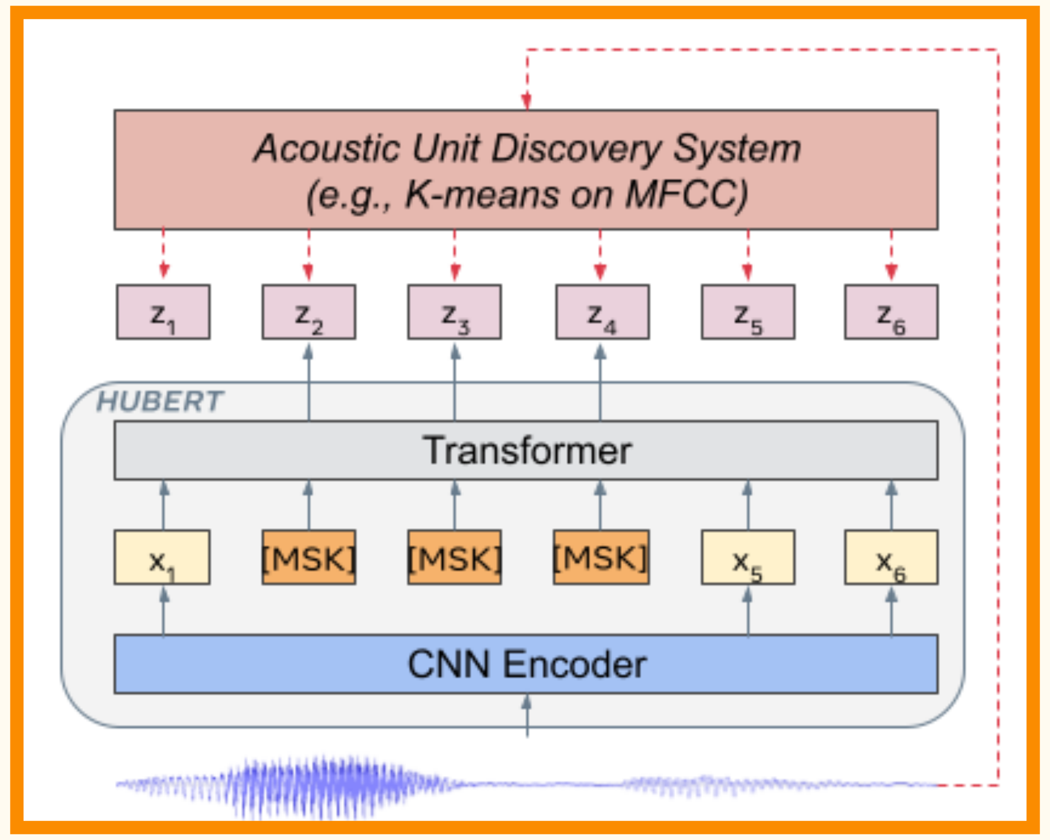
Pre-training

on general audio



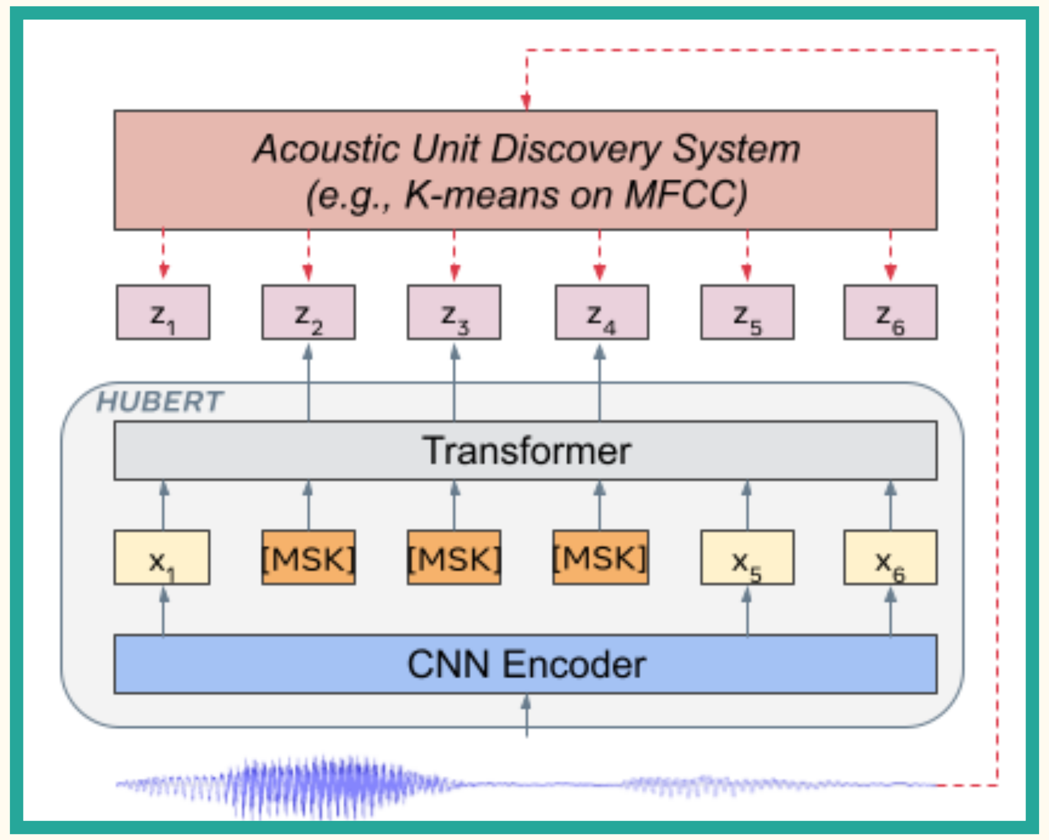
HuBERT:

- Librispeech 960h.
- Similar pre-training as WavLM.



AVES-Bio:

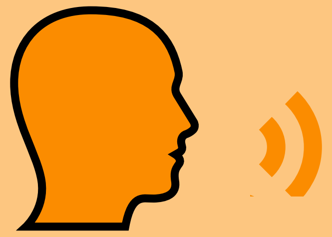
- FSD50K, AS, VGGSound.
- 360 hours of *animal* classes.



Pre-Training Domain - Feature Representations

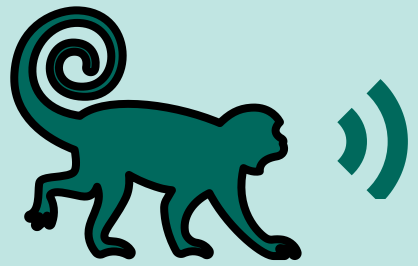
Pre-training

on human speech



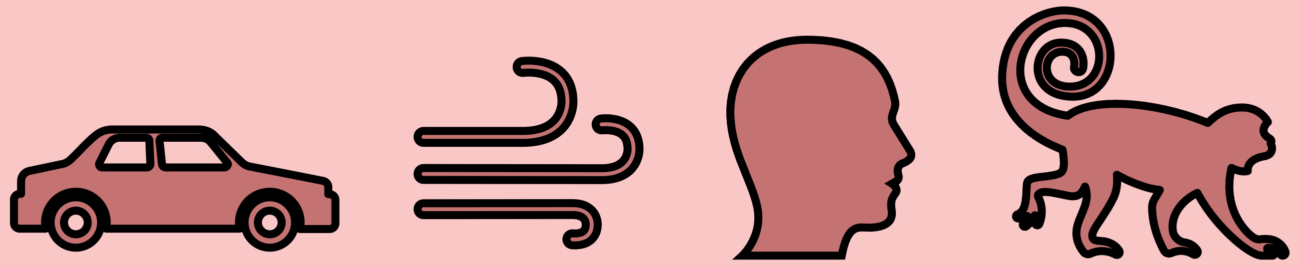
Pre-training

on bioacoustics



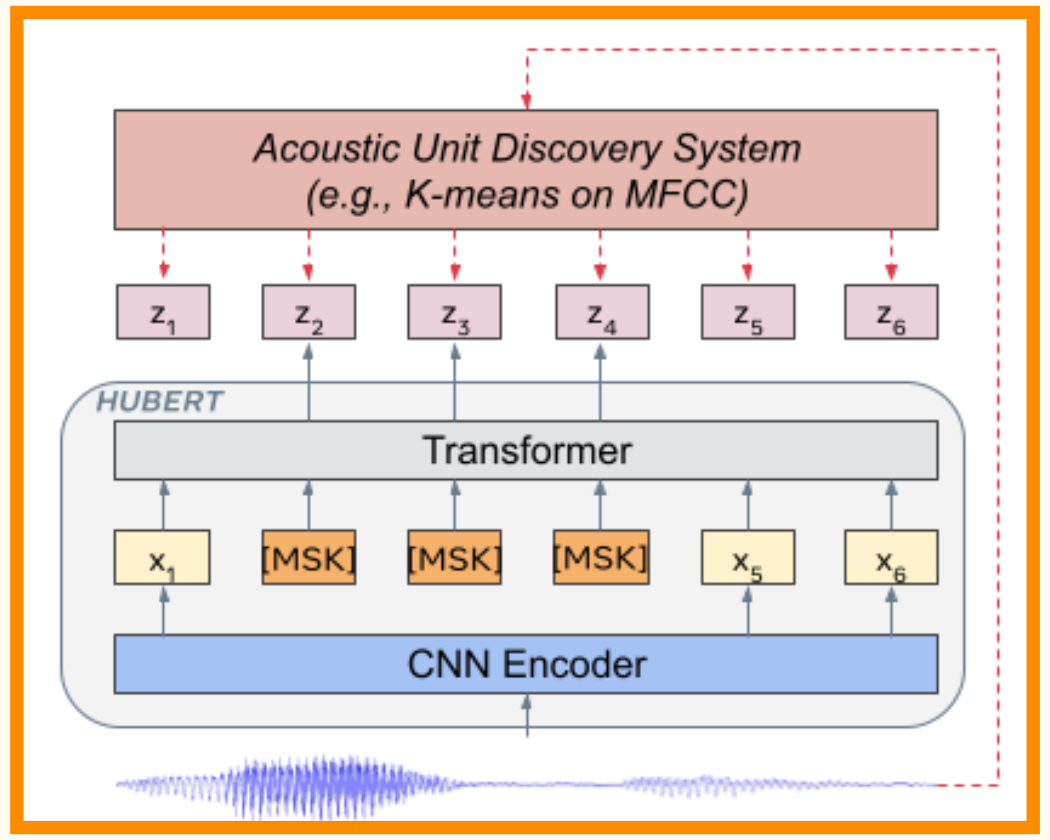
Pre-training

on general audio



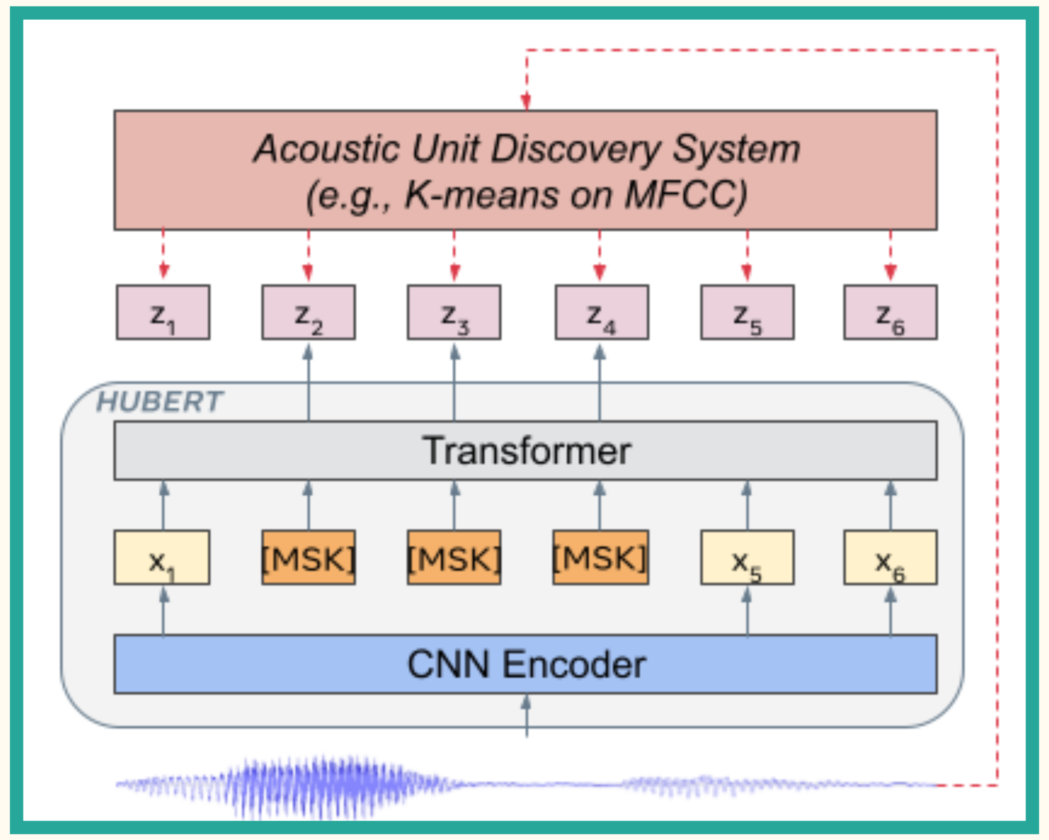
HuBERT:

- Librispeech 960h.
- Similar pre-training as WavLM.



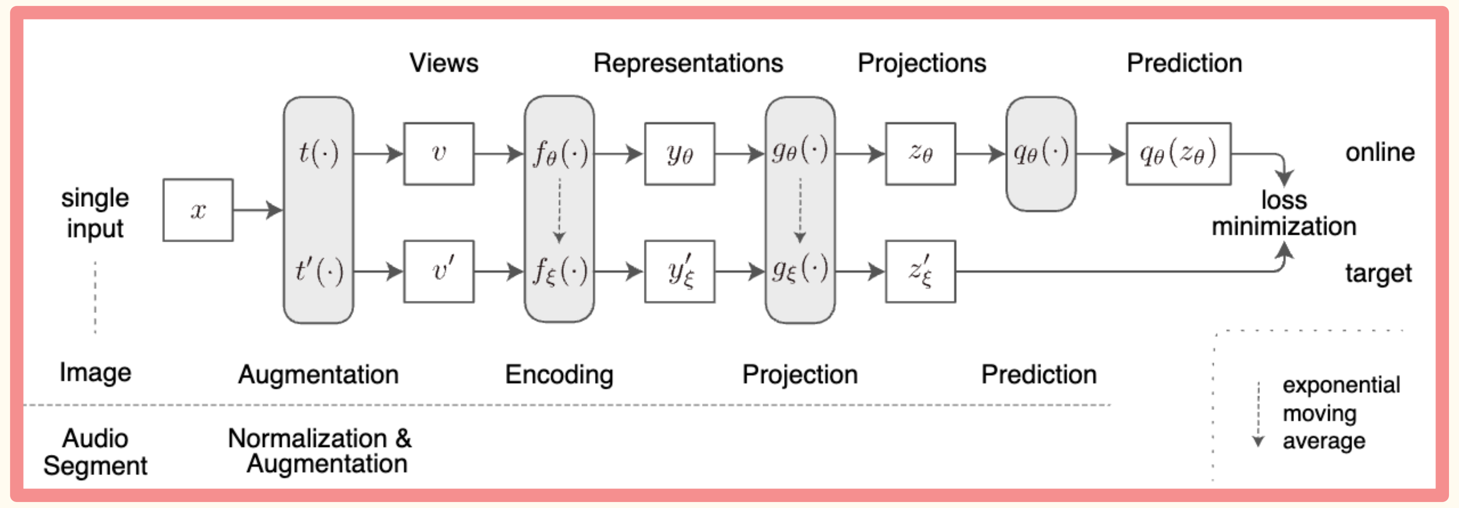
AVES-Bio:

- FSD50K, AS, VGGSound.
- 360 hours of *animal* classes.



BYOL:

- Full AudioSet.
- Different architecture.



Pre-Training Domain Analysis

UAR scores [%] on D_1 *Test*. Best layer scores are shown.

$\mathcal{F}$	Type	Corpus	CTID
Chance	-	-	9.09
AVES	SSL	FSD, AS, VGG-S	62.54
HuBERT	SSL	LS960	**64.35**

Pre-Training Domain Analysis

- Marginal difference in performance - can vary on datasets and contexts.

UAR scores [%] on D_1 *Test*. Best layer scores are shown.

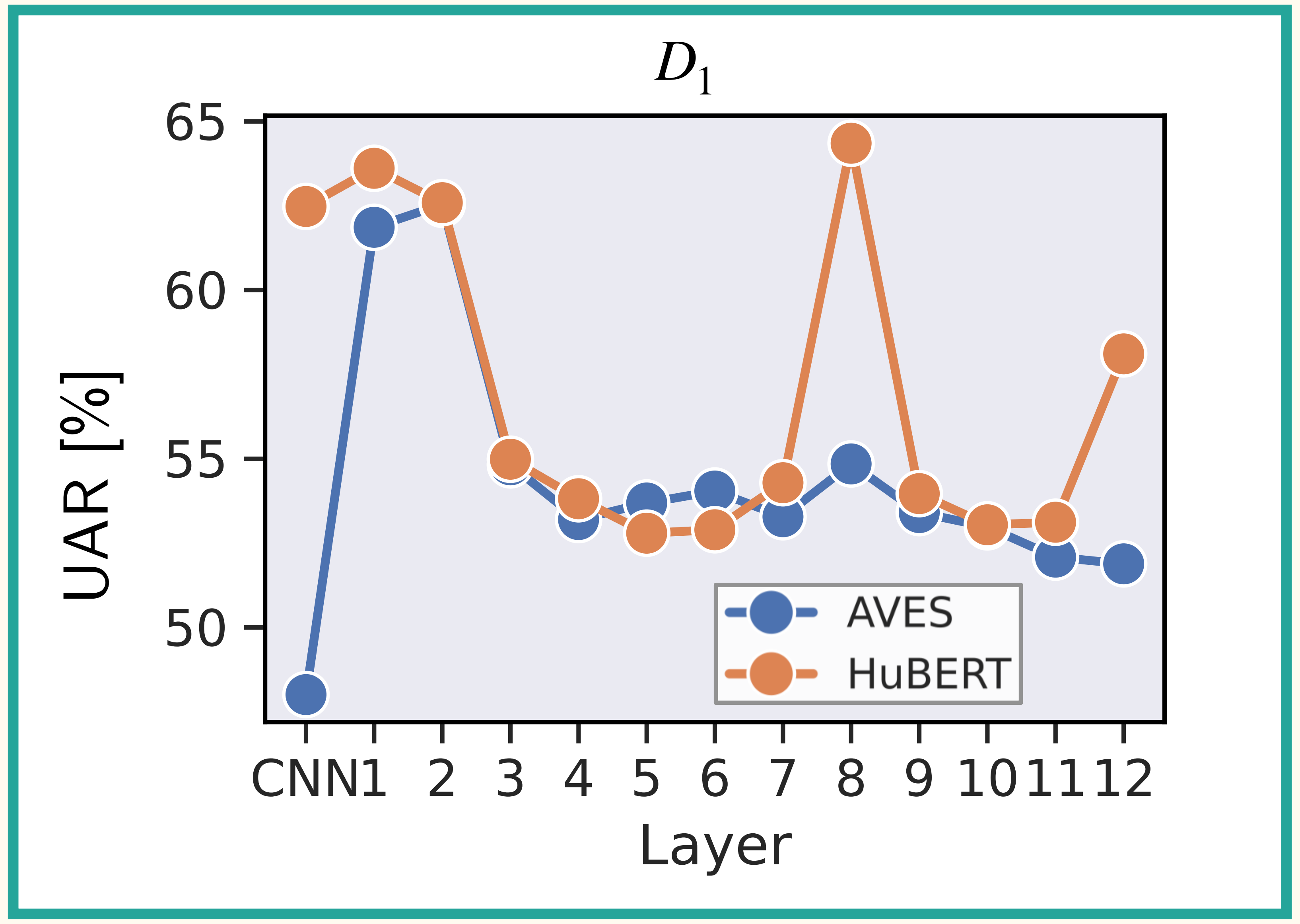
$\mathcal{F}$	**Type**	**Corpus**	**CTID**
Chance	-	-	9.09
AVES	SSL	FSD, AS, VGG-S	62.54
HuBERT	SSL	LS960	**64.35**

Pre-Training Domain Analysis

- Marginal difference in performance - can vary on datasets and contexts.
- AVES & HuBERT both show that initial layers are important.

UAR scores [%] on D_1 *Test*. Best layer scores are shown.

$\mathcal{F}$	Type	Corpus	CTID
Chance	-	-	9.09
AVES	SSL	FSD, AS, VGG-S	62.54
HuBERT	SSL	LS960	**64.35**



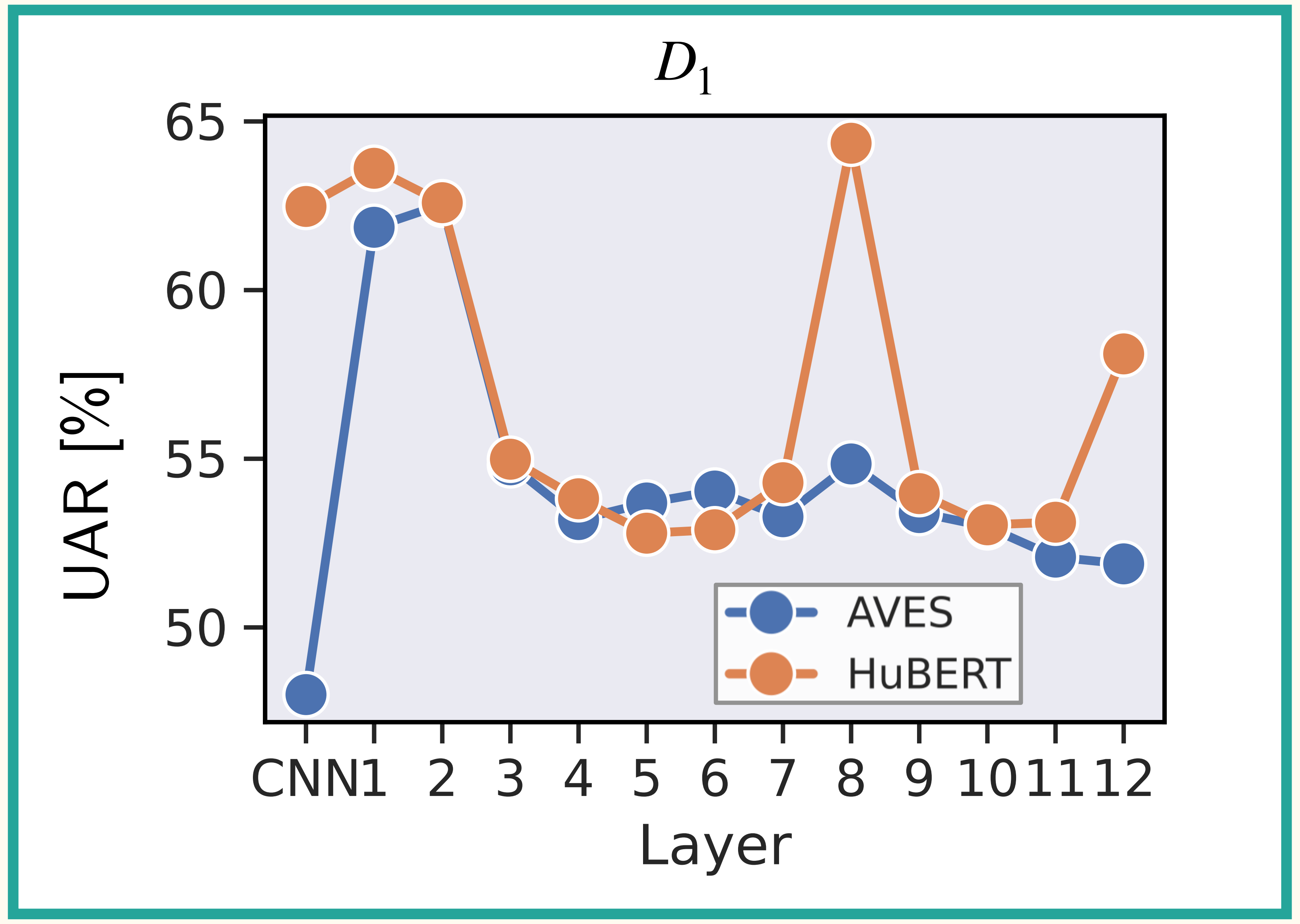
Layer-wise performance scores.

Pre-Training Domain Analysis

- Marginal difference in performance - can vary on datasets and contexts.
- AVES & HuBERT both show that initial layers are important.
  - Trend not limited to speech models.

UAR scores [%] on D_1 *Test*. Best layer scores are shown.

$\mathcal{F}$	Type	Corpus	CTID
Chance	-	-	9.09
AVES	SSL	FSD, AS, VGG-S	62.54
HuBERT	SSL	LS960	**64.35**



Layer-wise performance scores.

Pre-Training Domain Analysis

- Marginal difference in performance - can vary on datasets and contexts.
- AVES & HuBERT both show that initial layers are important.
  - Trend not limited to speech models.
- All 3 SSLs yield comparable results despite differences in pre-training domain, architecture, and objective.

UAR scores [%] on D_1 *Test*. Best layer scores are shown.

$\mathcal{F}$	**Type**	**Corpus**	**CTID**
Chance	-	-	9.09
AVES	SSL	FSD, AS, VGG-S	62.54
HuBERT	SSL	LS960	**64.35**
BYOL	SSL	AS	63.64

Pre-Training Domain Analysis

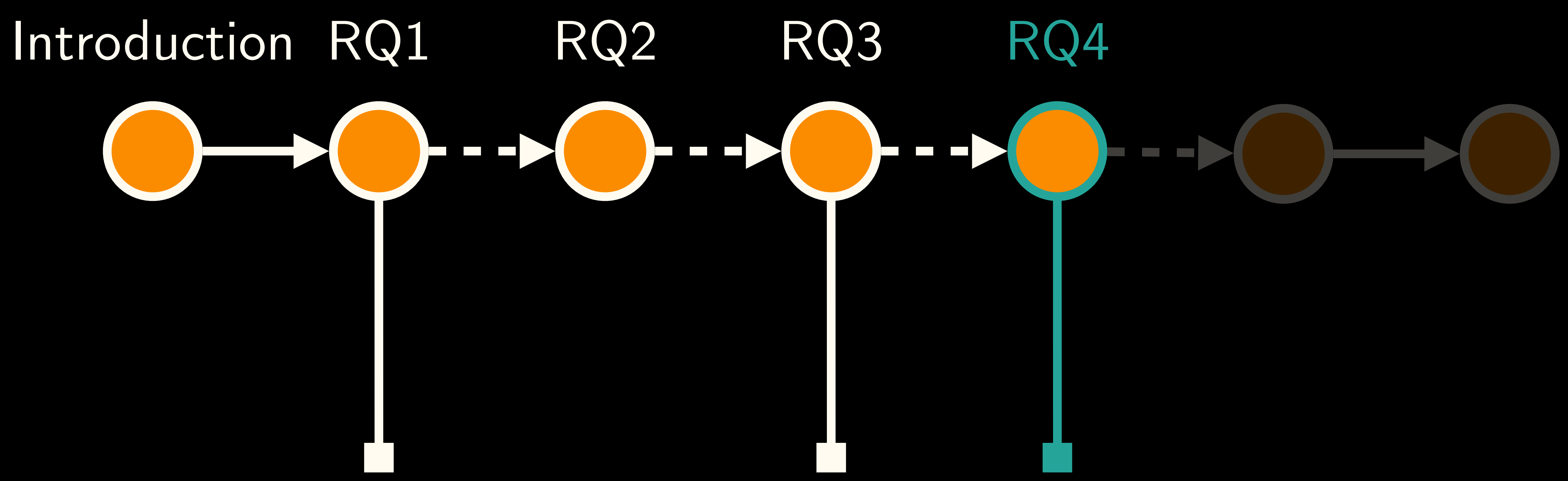
- Marginal difference in performance - can vary on datasets and contexts.
- AVES & HuBERT both show that initial layers are important.
  - Trend not limited to speech models.
- All 3 SSLs yield comparable results despite differences in pre-training domain, architecture, and objective.

UAR scores [%] on D_1 *Test*. Best layer scores are shown.

$\mathcal{F}$	**Type**	**Corpus**	**CTID**
Chance	-	-	9.09
AVES	SSL	FSD, AS, VGG-S	62.54
HuBERT	SSL	LS960	**64.35**
BYOL	SSL	AS	<u>63.64</u>

Key Takeaway

Self-supervised pre-training itself that allows these models to learn general representations with cross-domain transferability.

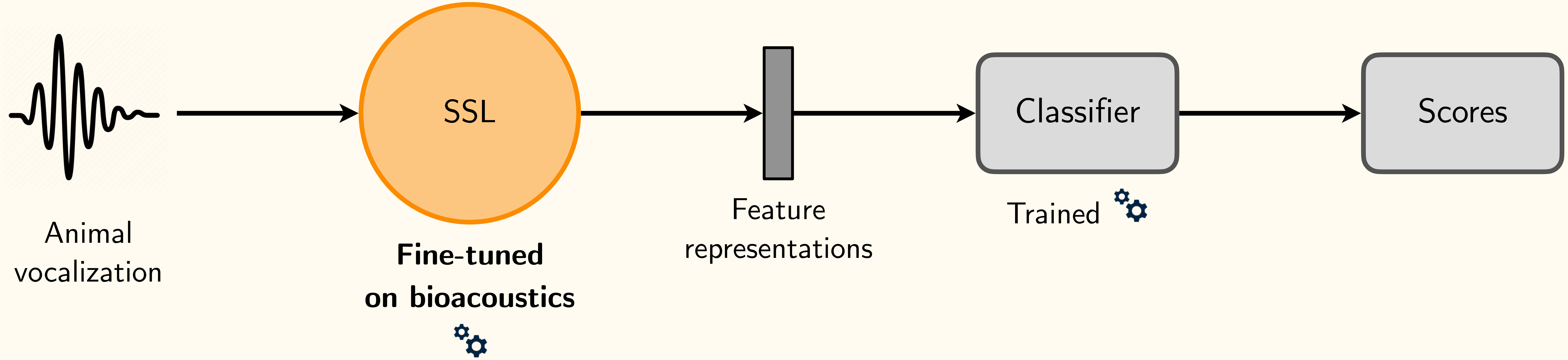


RQ4. Model Adaptation

Model Adaptation

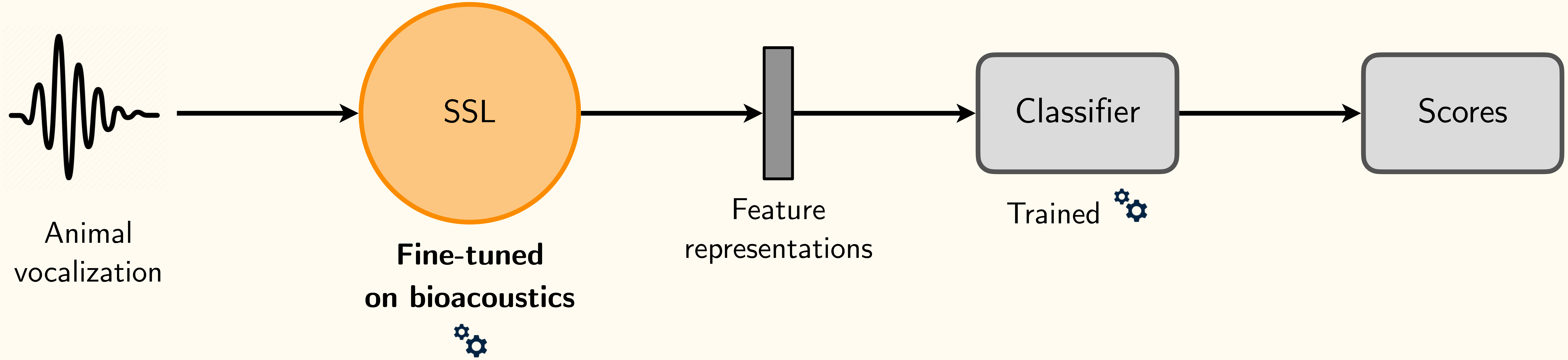
Model Adaptation

- Investigate: does fine-tuning the same SSL models directly on the downstream bioacoustic data yields better results ?



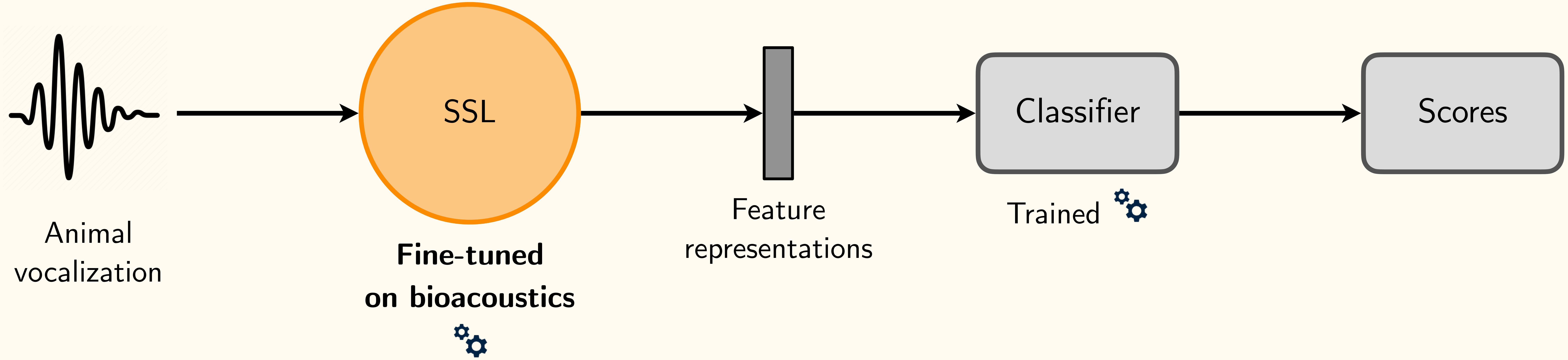
Model Adaptation

- Investigate: does fine-tuning the same SSL models directly on the downstream bioacoustic data yields better results ?
  - Adapt HuBERT and AVES.



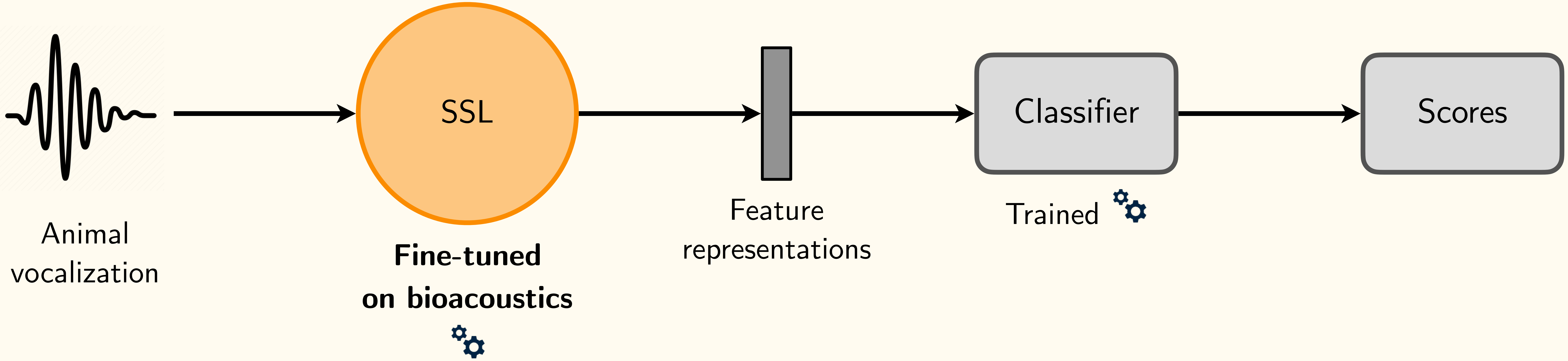
Model Adaptation

- Investigate: does fine-tuning the same SSL models directly on the downstream bioacoustic data yields better results ?
  - Adapt HuBERT and AVES.
  - Focus only on CTID.



Model Adaptation

- Investigate: does fine-tuning the same SSL models directly on the downstream bioacoustic data yields better results ?
  - Adapt HuBERT and AVES.
  - Focus only on CTID.
- Multiple studies: matrix selection, layer selection strategy, *fine-tuning strategy*.

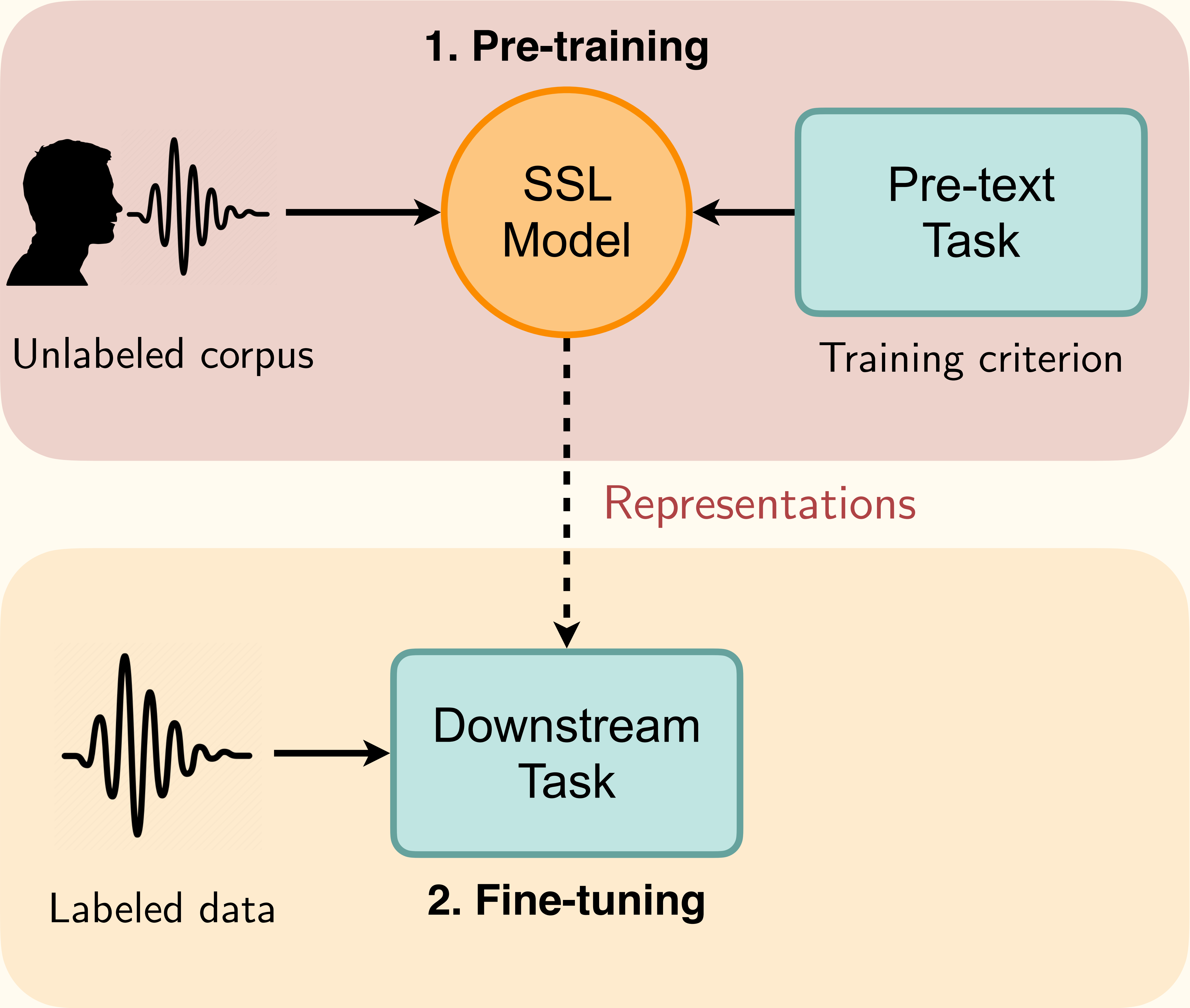


Parameter Efficient Fine-Tuning and Low-Rank Adaptation

[1] Aghajanyan et al., *Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine*-Tuning, (2021) ACL-IJCNLP.
[2] Hu, E.J. et al., *LoRA: Low-Rank Adaptation of Large Language Models* (2022). International Conference on Learning Representations.

Parameter Efficient Fine-Tuning and Low-Rank Adaptation

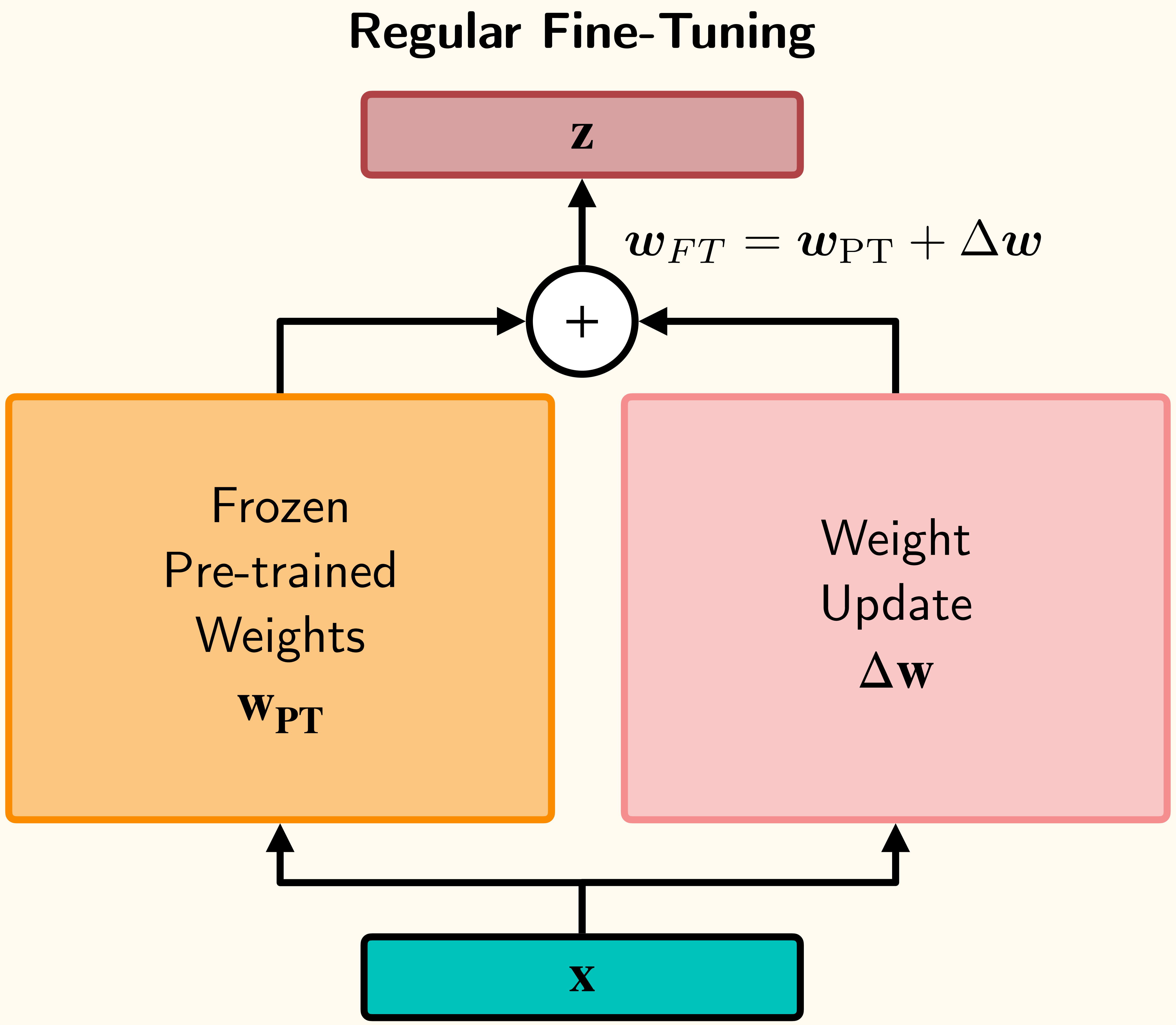
- Fine-tuning on a downstream task: 2nd step of the SSL framework.



[1] Aghajanyan et al., *Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine*-Tuning, (2021) ACL-IJCNLP.
[2] Hu, E.J. et al., *LoRA: Low-Rank Adaptation of Large Language Models* (2022). International Conference on Learning Representations.

Parameter Efficient Fine-Tuning and Low-Rank Adaptation

- Fine-tuning on a downstream task: 2nd step of the SSL framework.
- Full fine-tuning: entire parameter set updated → computationally expensive and requires large quantities of data.

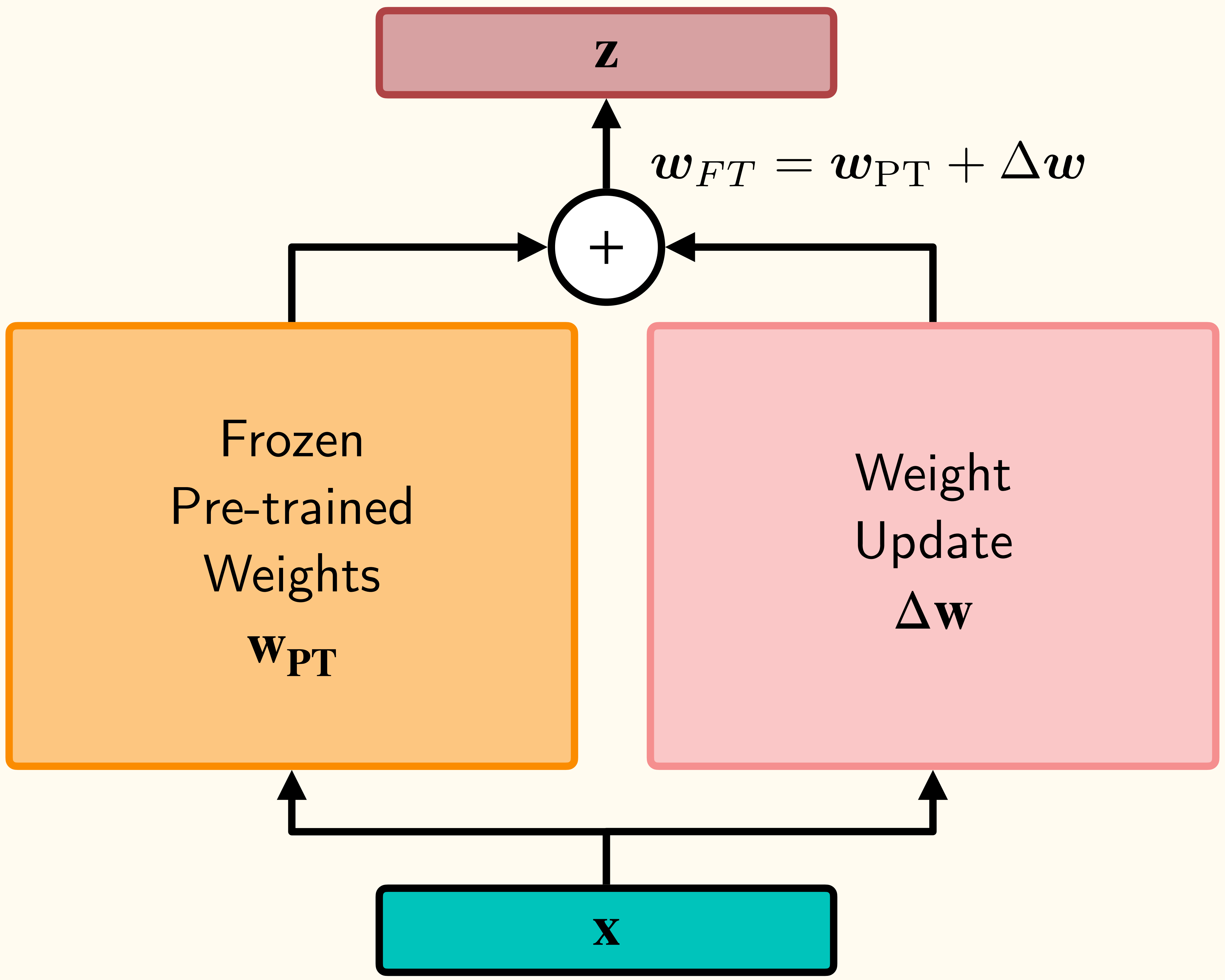


[1] Aghajanyan et al., *Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine*-Tuning, (2021) ACL-IJCNLP.
[2] Hu, E.J. et al., *LoRA: Low-Rank Adaptation of Large Language Models* (2022). International Conference on Learning Representations.

Parameter Efficient Fine-Tuning and Low-Rank Adaptation

- Fine-tuning on a downstream task: 2nd step of the SSL framework.
- Full fine-tuning: entire parameter set updated → computationally expensive and requires large quantities of data.
- PEFT approach: strategically update only a small subset → reduced cost.

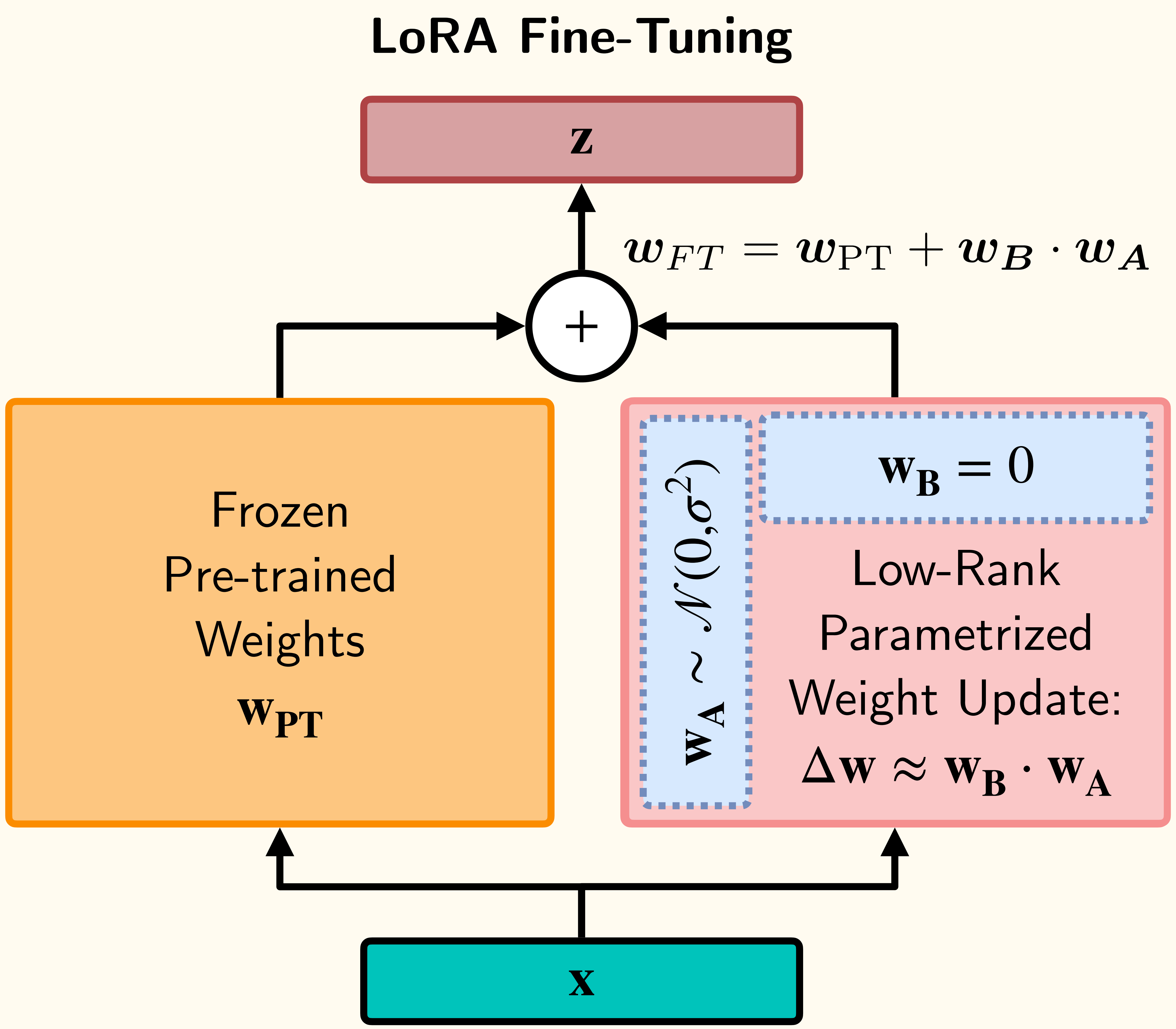
Regular Fine-Tuning



[1] Aghajanyan et al., *Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine*-Tuning, (2021) ACL-IJCNLP.
[2] Hu, E.J. et al., *LoRA: Low-Rank Adaptation of Large Language Models* (2022). International Conference on Learning Representations.

Parameter Efficient Fine-Tuning and Low-Rank Adaptation

- Fine-tuning on a downstream task: 2nd step of the SSL framework.
- Full fine-tuning: entire parameter set updated → computationally expensive and requires large quantities of data.
- PEFT approach: strategically update only a small subset → reduced cost.
  - Low-Rank Adaptation (LoRA): approximate Δw with 2 smaller matrices.



[1] Aghajanyan et al., *Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine*-Tuning, (2021) ACL-IJCNLP.
[2] Hu, E.J. et al., *LoRA: Low-Rank Adaptation of Large Language Models* (2022). International Conference on Learning Representations.

Adaption - Fine-Tuning Strategy

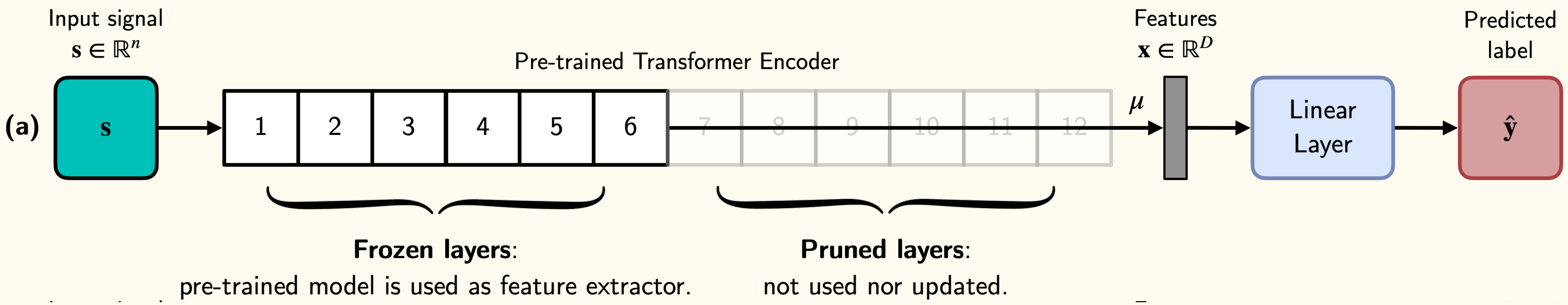
Adaption - Fine-Tuning Strategy

3 scenarios:

Adaption - Fine-Tuning Strategy

3 scenarios:

(a) Linear probing.

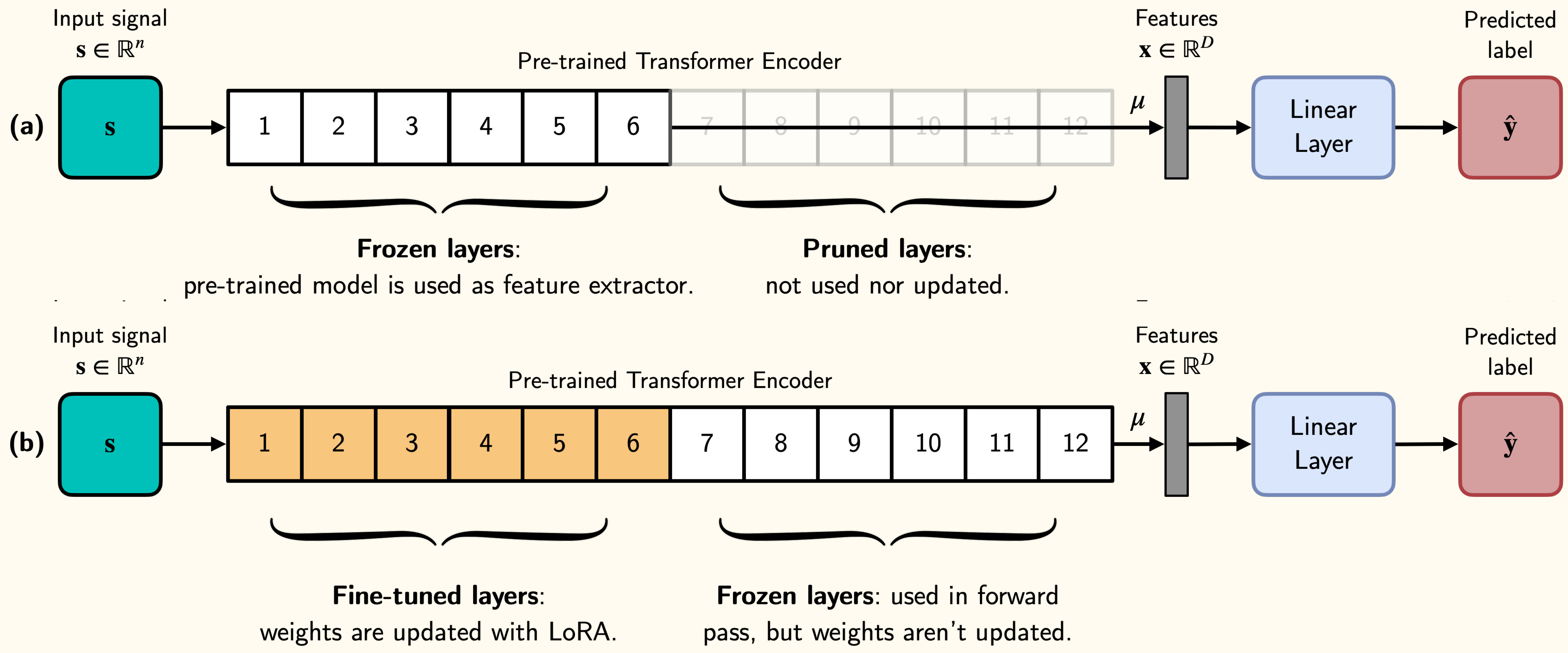


Adaption - Fine-Tuning Strategy

3 scenarios:

(a) Linear probing.

(b) LoRA + Freeze.



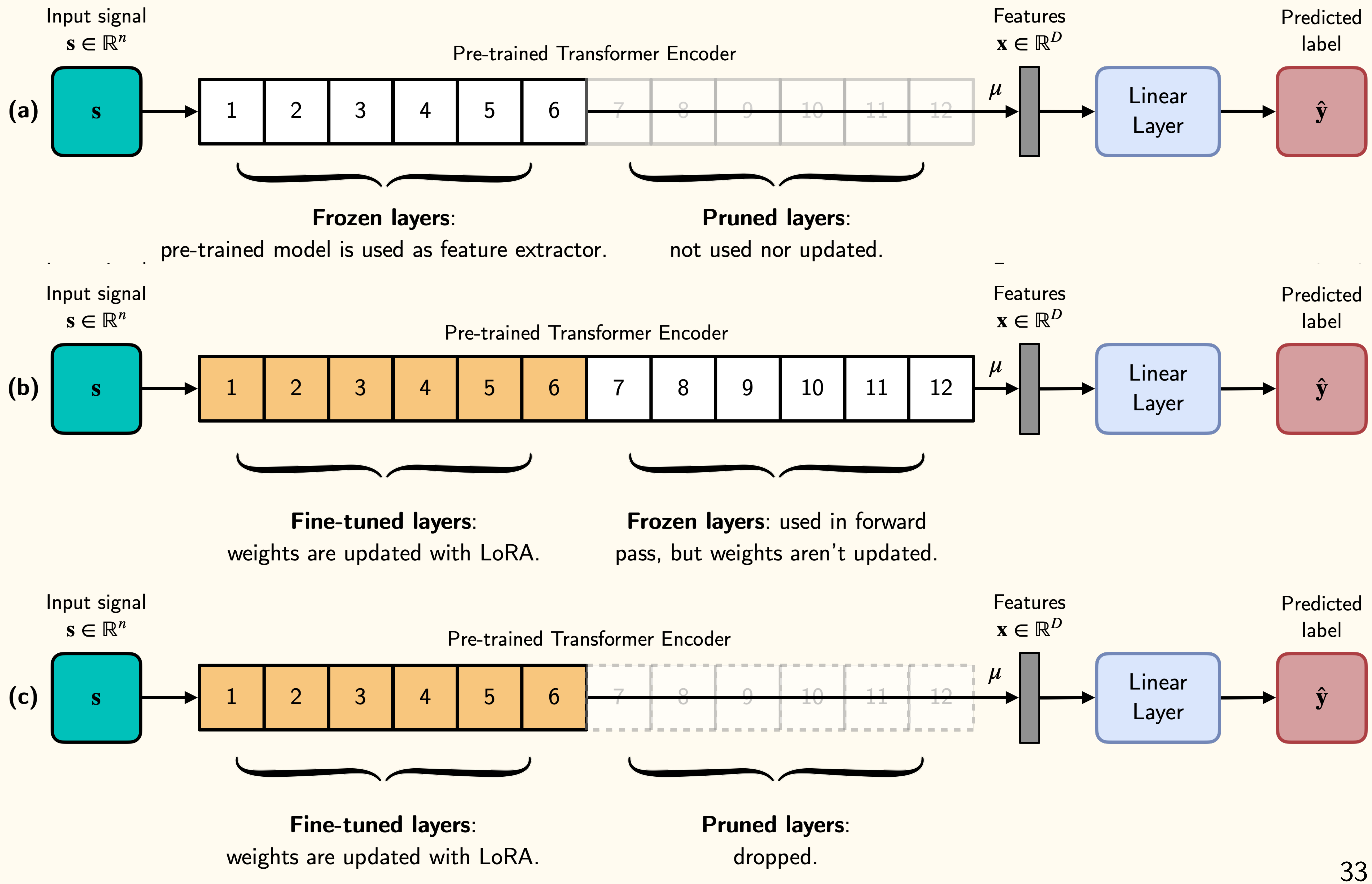
Adaption - Fine-Tuning Strategy

3 scenarios:

(a) Linear probing.

(b) LoRA + Freeze.

(c) LoRA + Drop.



Adaption - Fine-Tuning Strategy

3 scenarios:

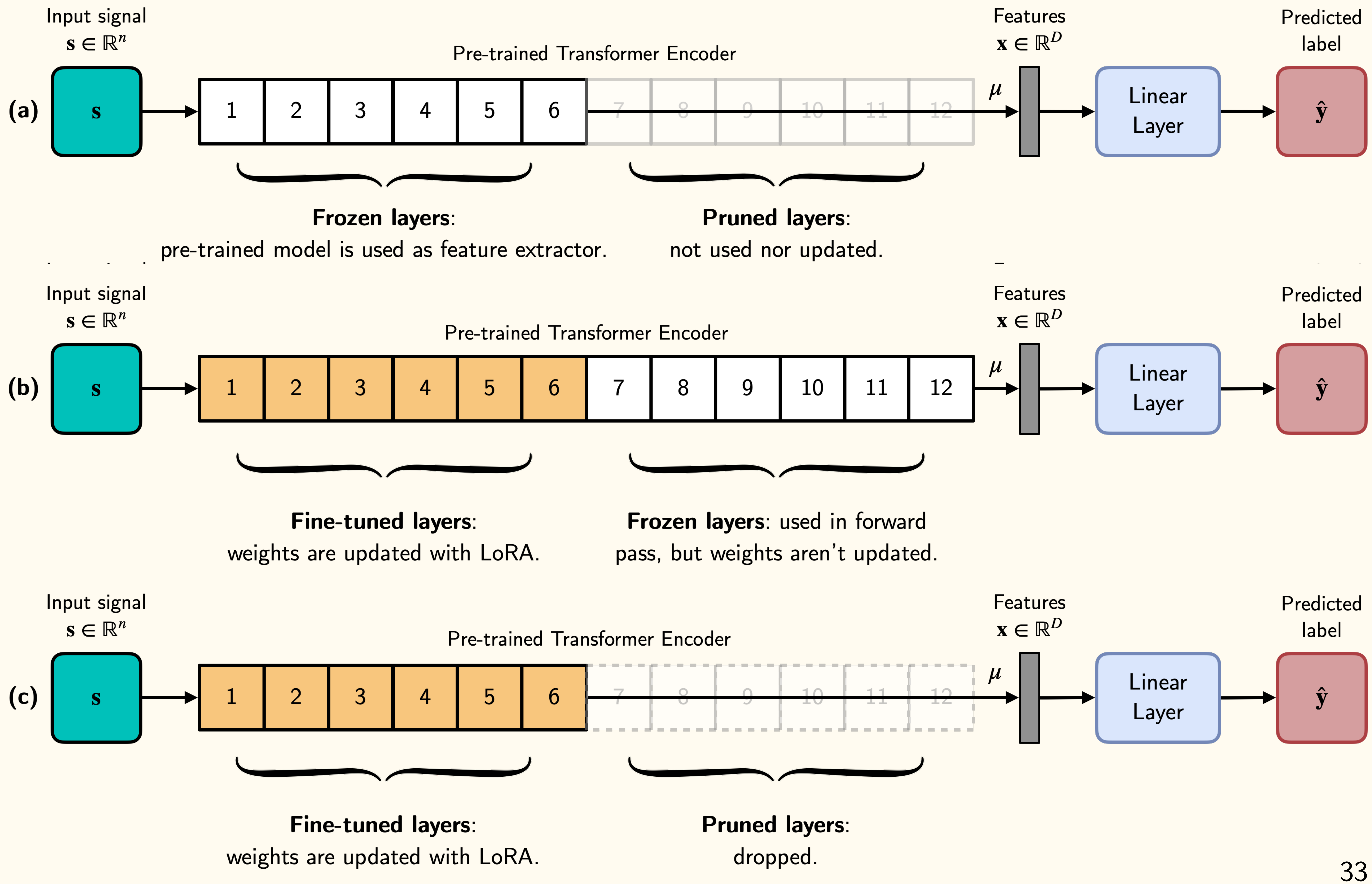
(a) Linear probing.

(b) LoRA + Freeze.

(c) LoRA + Drop.

Aims:

- Does LoRA improve over linear probing?



Adaption - Fine-Tuning Strategy

3 scenarios:

(a) Linear probing.

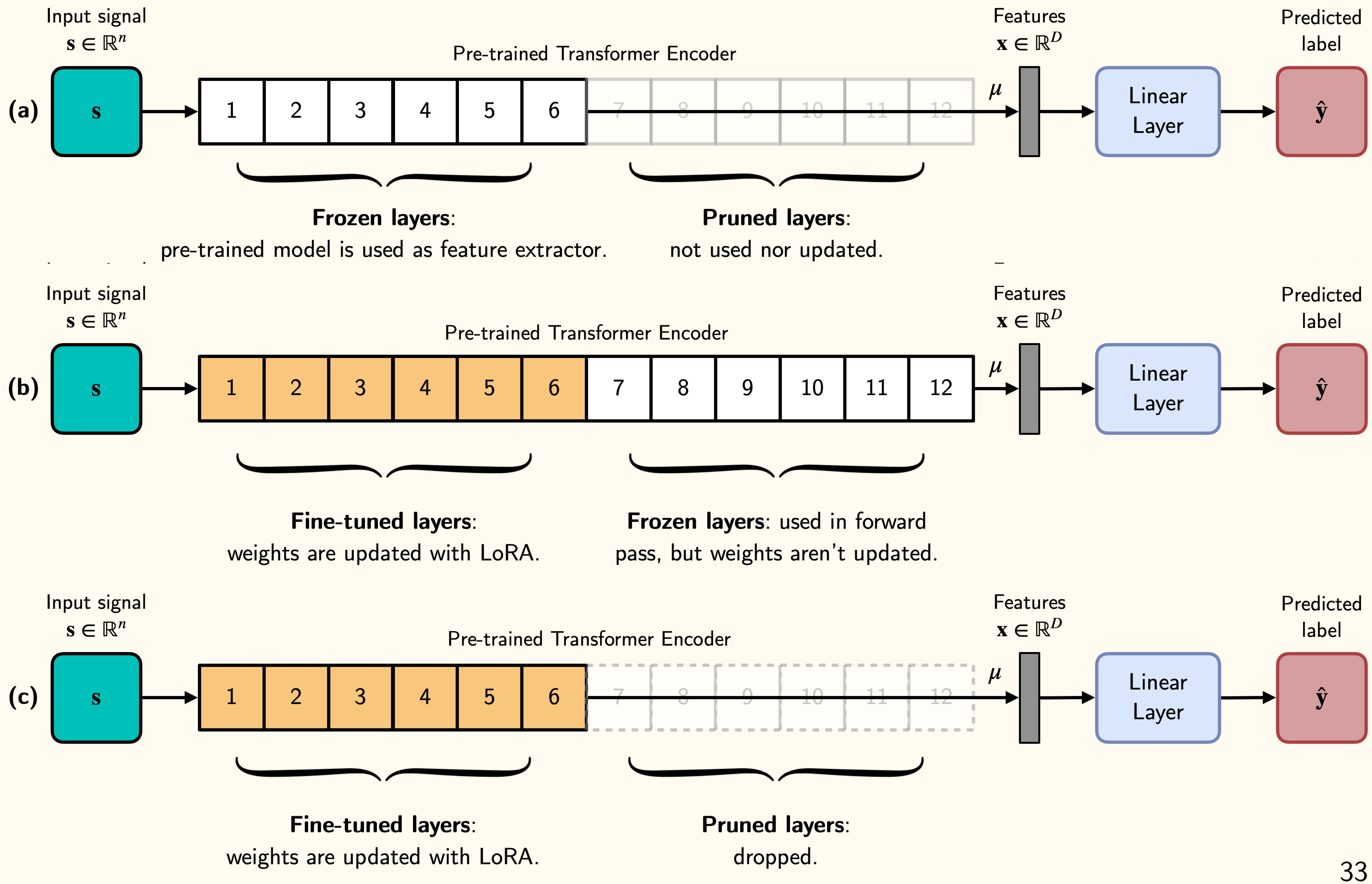
(b) LoRA + Freeze.

(c) LoRA + Drop.

Aims:

- Does LoRA improve over linear probing?

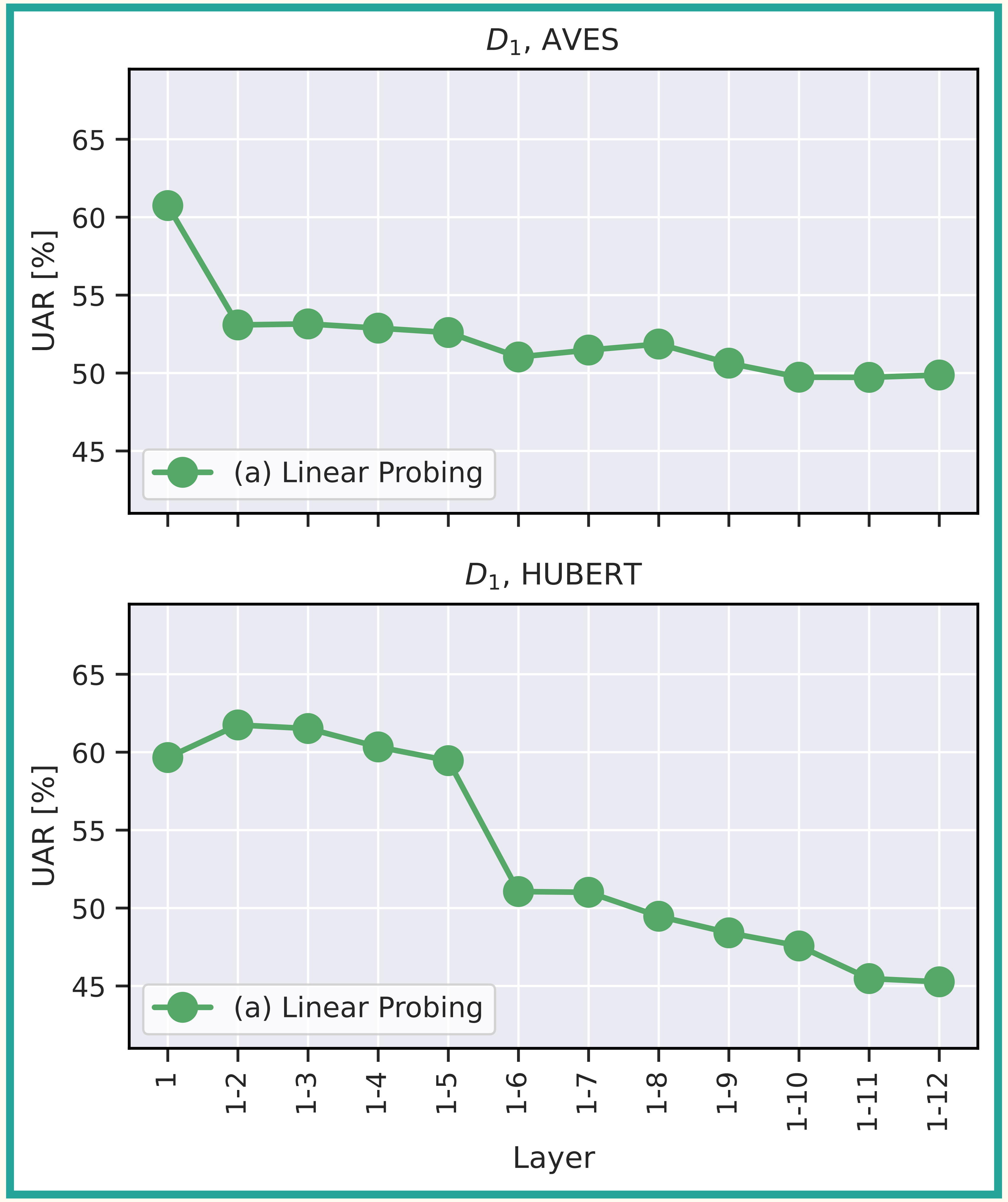
- Any difference between freezing and dropping ?



Adaption - Fine-Tuning Strategy

Adaption - Fine-Tuning Strategy

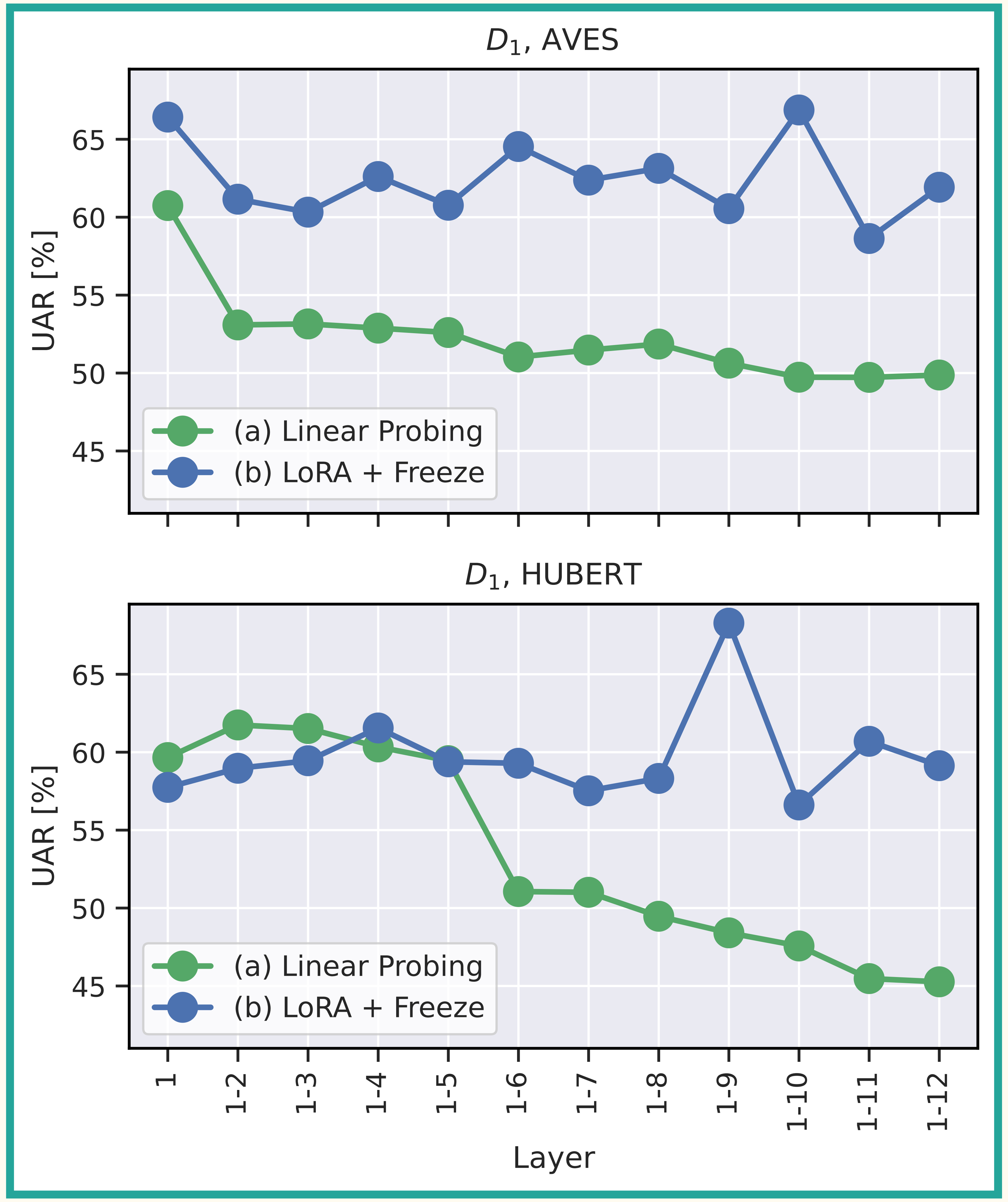
- Linear probing: downwards trend through the layers.



Layer-wise UAR [%] performance on IMV.

Adaption - Fine-Tuning Strategy

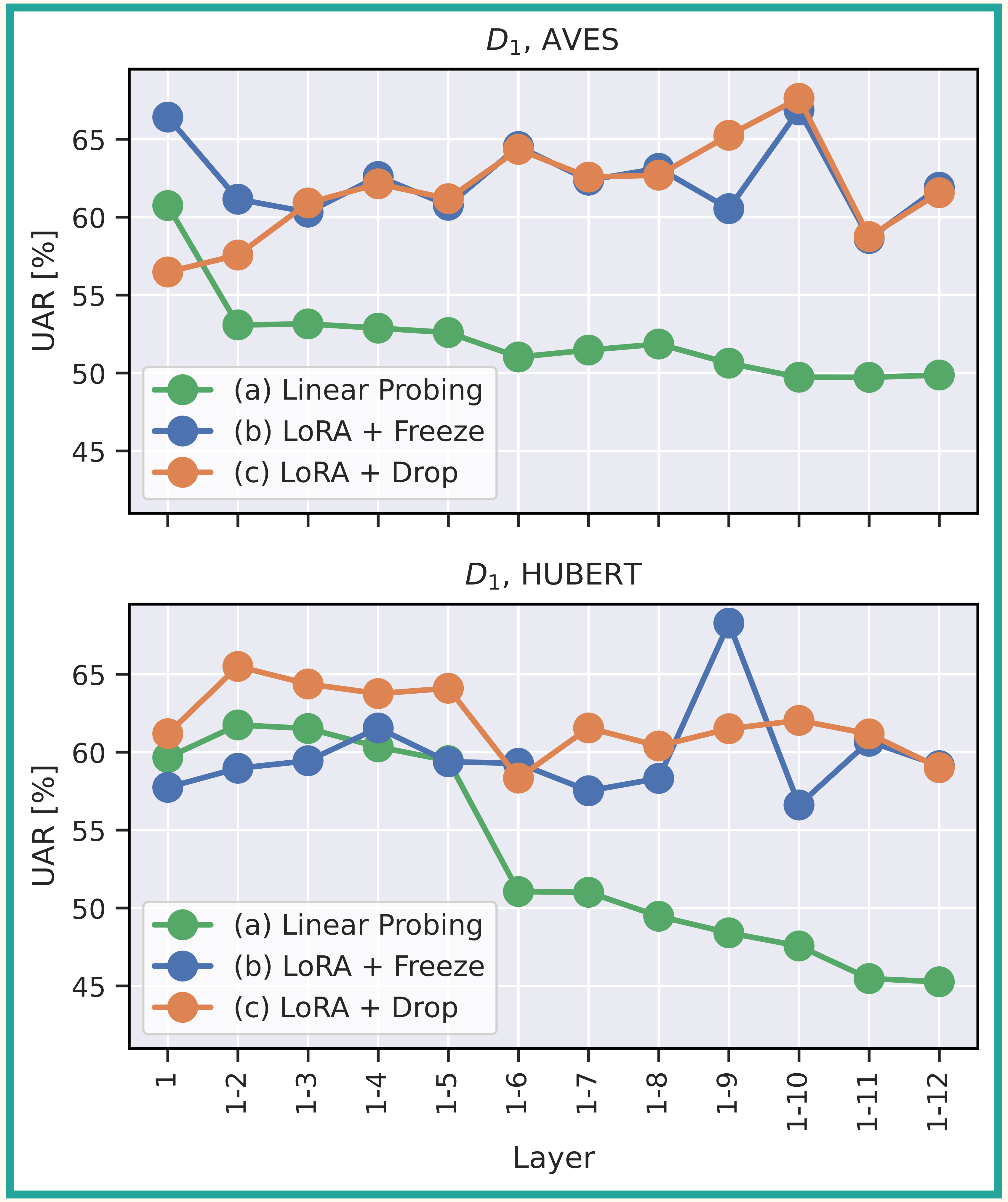
- Linear probing: downwards trend through the layers.
- LoRA fine-tuning: consistently and significantly improves performance across nearly all layers.



Layer-wise UAR [%] performance on IMV.

Adaption - Fine-Tuning Strategy

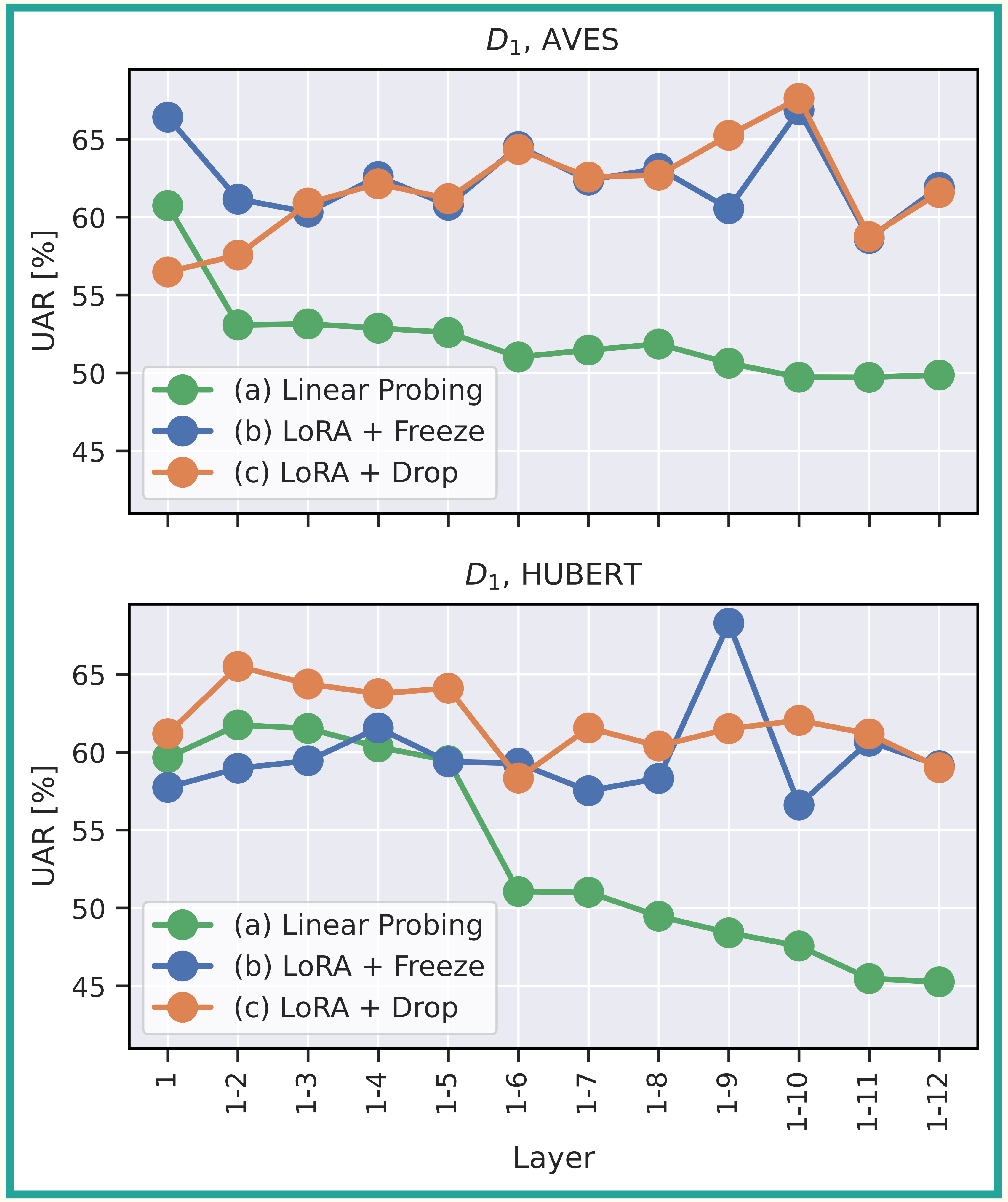
- Linear probing: downwards trend through the layers.
- LoRA fine-tuning: consistently and significantly improves performance across nearly all layers.



Layer-wise UAR [%] performance on IMV.

Adaption - Fine-Tuning Strategy

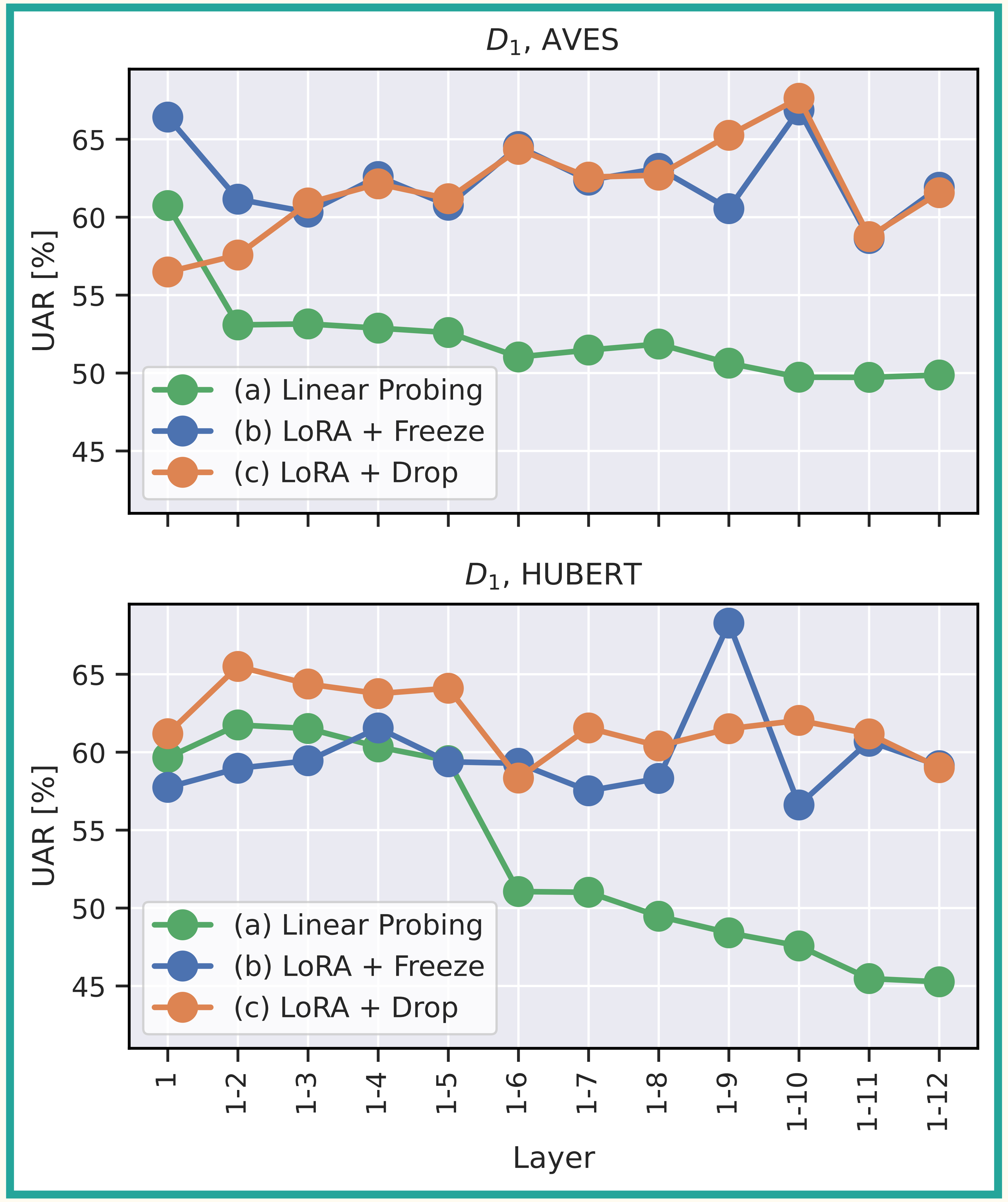
- Linear probing: downwards trend through the layers.
- LoRA fine-tuning: consistently and significantly improves performance across nearly all layers.
- AVES: LoRA models have a general upward trend.



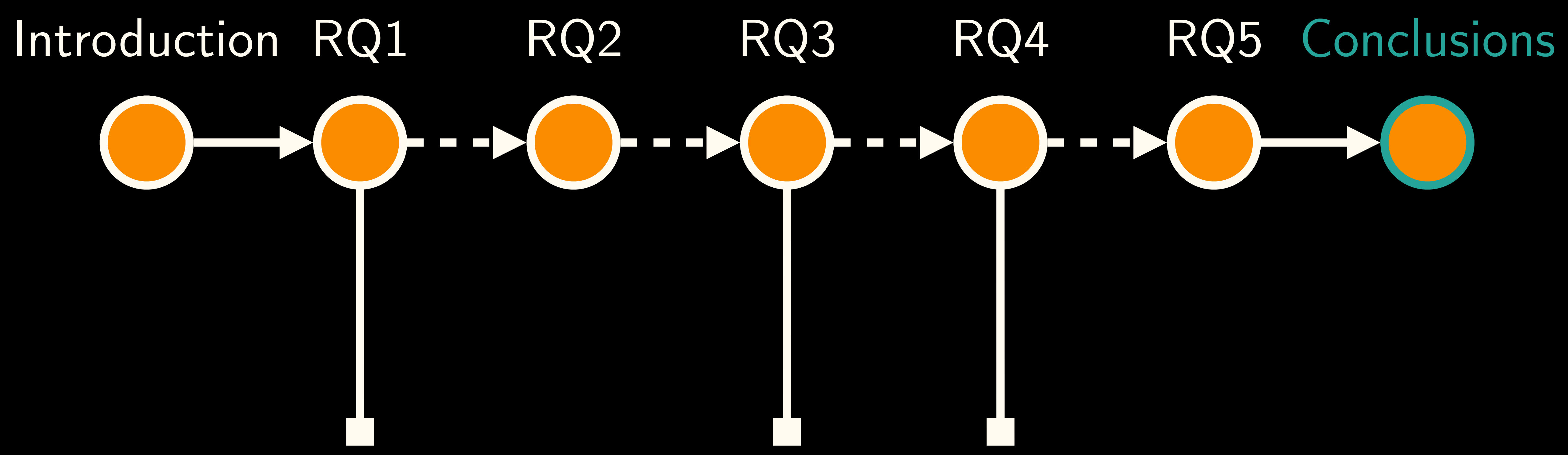
Layer-wise UAR [%] performance on IMV.

Adaption - Fine-Tuning Strategy

- Linear probing: downwards trend through the layers.
- LoRA fine-tuning: consistently and significantly improves performance across nearly all layers.
- AVES: LoRA models have a general upward trend.
- Later layers perform poorly without fine-tuning, but become informative with LoRA adaptation.



Layer-wise UAR [%] performance on IMV.



Conclusions

Conclusions

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.
- Bioacoustics and general audio SSLs performance comparably to speech SSLs.

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.
- Bioacoustics and general audio SSLs performance comparably to speech SSLs.
- Fine-tuning SSLs on the downstream data can lead to improved performances.

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.
- Bioacoustics and general audio SSLs performance comparably to speech SSLs.
- Fine-tuning SSLs on the downstream data can lead to improved performances.

This thesis:

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.
- Bioacoustics and general audio SSLs performance comparably to speech SSLs.
- Fine-tuning SSLs on the downstream data can lead to improved performances.

This thesis:

➡ Establishes that audio SSL models constitute a powerful, domain-agnostic toolkit.

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.
- Bioacoustics and general audio SSLs performance comparably to speech SSLs.
- Fine-tuning SSLs on the downstream data can lead to improved performances.

This thesis:

➡ Establishes that audio SSL models constitute a powerful, domain-agnostic toolkit.

➡ Offers versatile starting point for decoding animal vocal communication.

Conclusions

- Speech SSLs carry meaningful information → distinguish animal vocalizations by call-type, caller identity, and sex.
- Bioacoustics and general audio SSLs performance comparably to speech SSLs.
- Fine-tuning SSLs on the downstream data can lead to improved performances.

This thesis:

➡ Establishes that audio SSL models constitute a powerful, domain-agnostic toolkit.

➡ Offers versatile starting point for decoding animal vocal communication.

➡ Provides practical framework: extendable to new species, recording conditions, and behavioral contexts.

List of Publications I

1. **Sarkar, E.**, Prasad, R., Magimai-Doss, M., *Unsupervised Voice Activity Detection by Modeling Source and System Information using Zero Frequency Filtering*, Interspeech 2022.

2. **Sarkar, E.**, Magimai-Doss, M., *Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers?*, Interspeech 2023.

3. **Sarkar, E.**, Magimai-Doss, M., *On the utility of Speech and Audio Foundation Models for Animal Call Analysis*, 4th International Workshop on Vocal Interactivity In-and-between Humans, Animals and Robots (VIHAR), Interspeech 2024.

List of Publications II

4. Ben Mahmoud, I., **Sarkar, E.**, Manser, M., Magimai-Doss, M., *Feature Representations for Automatic Meerkat Vocalization Classification*, 4th International Workshop on Vocal Interactivity In-and-between Humans, Animals and Robots (VIHAR), Interspeech 2024.
5. **Sarkar, E.**, K. Wierucka, A. B. Bosshard, J. M. Burkart, Magimai-Doss, M., *On Feature Representation for Marmoset Vocal Communication Analysis*, Bioacoustics 2025.
6. **Sarkar, E.**, Magimai-Doss, M., *Comparing Self-Supervised Learning Models Pre-Trained on Human Speech and Animal Vocalizations for Bioacoustics Processing*, ICASSP 2025.

List of Publications III

7. **Sarkar, E.**, Mohammadi, A., Magimai-Doss, M., *Adaptation of Speech and Bioacoustics Models*. Idiap-RR Idiap-Internal-RR-05-2025. Idiap, 2025.

8. **Sarkar, E.**, Magimai-Doss, M., *Leveraging Sequential Structure in Animal Vocalizations*, Idiap-RR Idiap-Internal-RR-06-2025, Idiap, 2025.

Thank you !

Idiap Research Institute

https://eklavyafcb.github.io

eklavya.sarkar@idiap.ch

References

- Stowell, D. (2022). ‘Computational bioacoustics with deep learning: a review and roadmap’. *PeerJ* 10, e13152.
- Sainburg, T. et al. (2020). ‘Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires’. PLoS Computational Biology 16.10, e1008228.
- Zhang, Y. et al. (2018). ‘Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks’. The Journal of the Acoustical Society of America 144, pp. 478–487.
- Coffey, E. et al. (2019). ‘Deep representation learning for orca call type classification’. Scientific Reports 9.1, pp. 1–10.
- Bergler, C. et al. (2019). ‘ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning’. Scientific Reports 9.1, pp. 1–10.
- Agamaite, J. A. et al. ‘A quantitative acoustic analysis of the vocal repertoire of the common marmoset’. (2015). The Journal of the Acoustical Society of America 138(5), pp. 2906–2928.
- Chen et al., ‘WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing’. (2022). IEEE Journal of Selected Topics in Signal Processing.
- Aghajanyan et al., Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning, (2021) ACL-IJCNLP.
- Hu, E.J. et al., LoRA: Low-Rank Adaptation of Large Language Models (2022). International Conference on Learning Representations.

Appendix

FAQ - MLP Classifier

- **Model**: 4-layer MLP

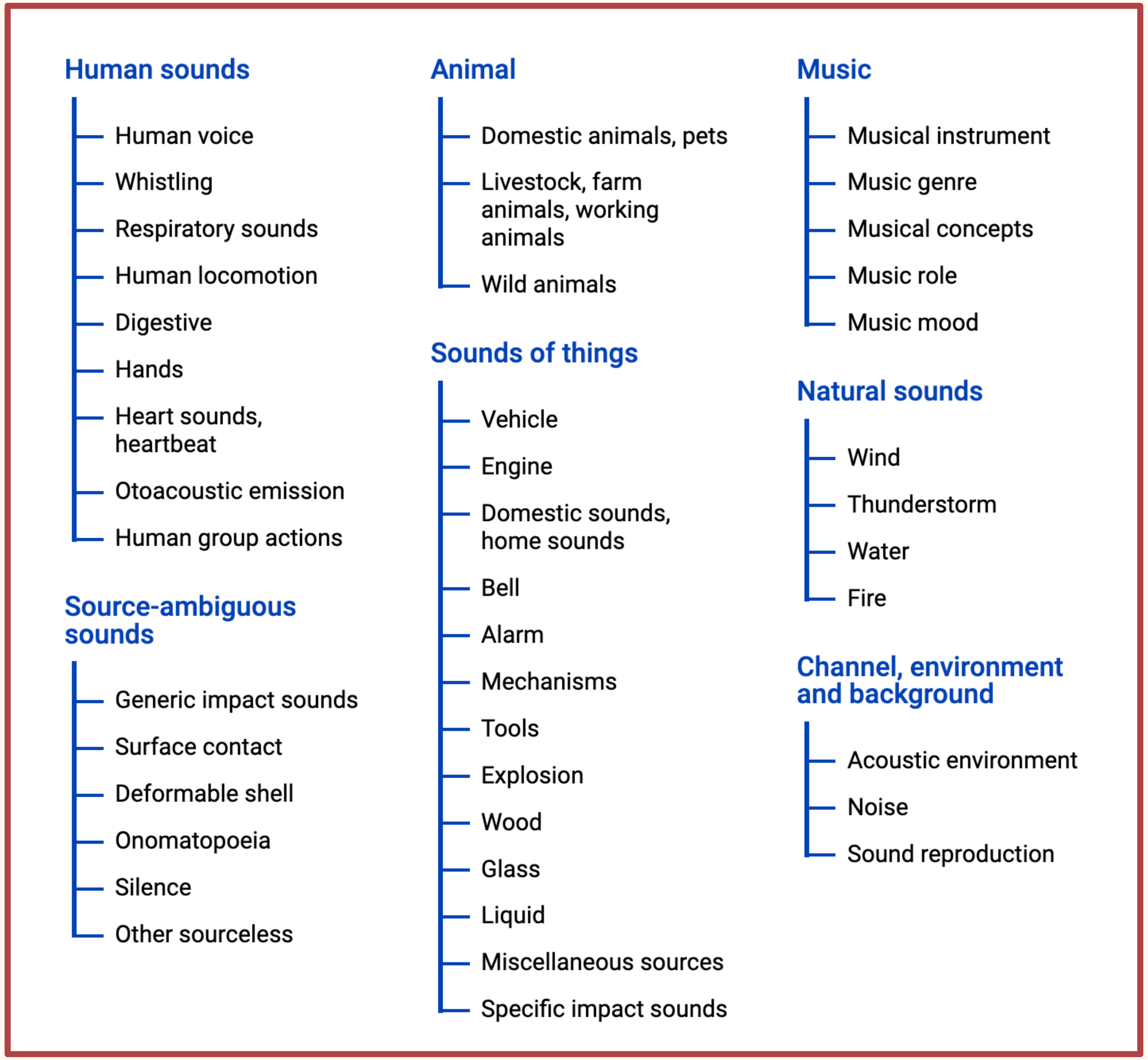
Block	Layers	# Hidden Units	Activation
1	Linear, LayerNorm	128	ReLU
2	Linear, LayerNorm	64	ReLU
3	Linear, LayerNorm	32	ReLU
4	Linear	# classes	

- **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 - AudioSet

Audio event classes such as:

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



AudioSet Dataset Ontology

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)
```

(* → self.fc1)

FAQ - BYOL

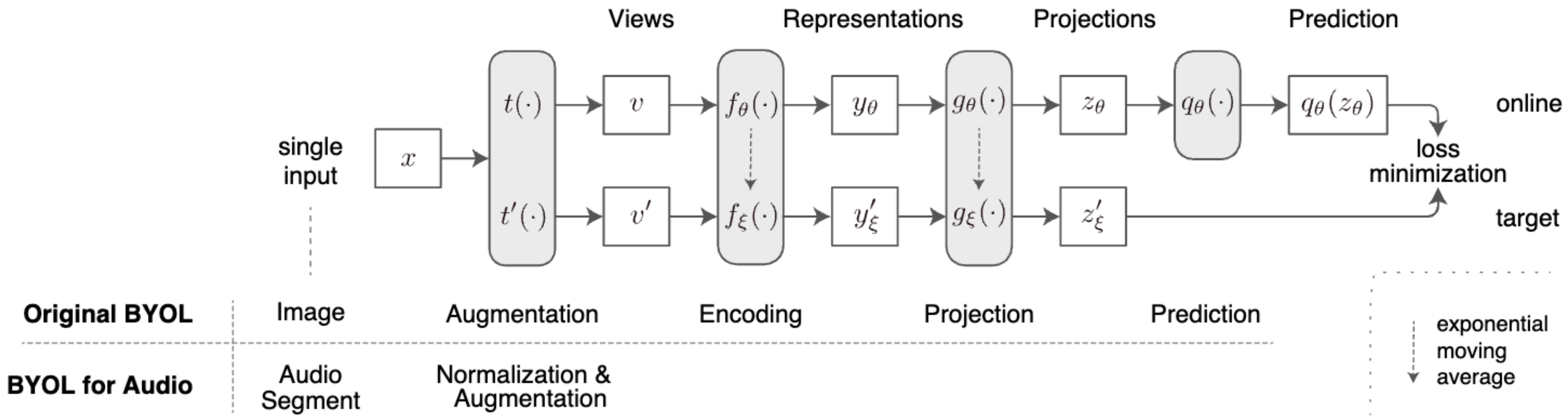


Fig. 2. BYOL and BYOL-A system overview.

- Minimizes distance between two augmented views of the same audio sample.

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)

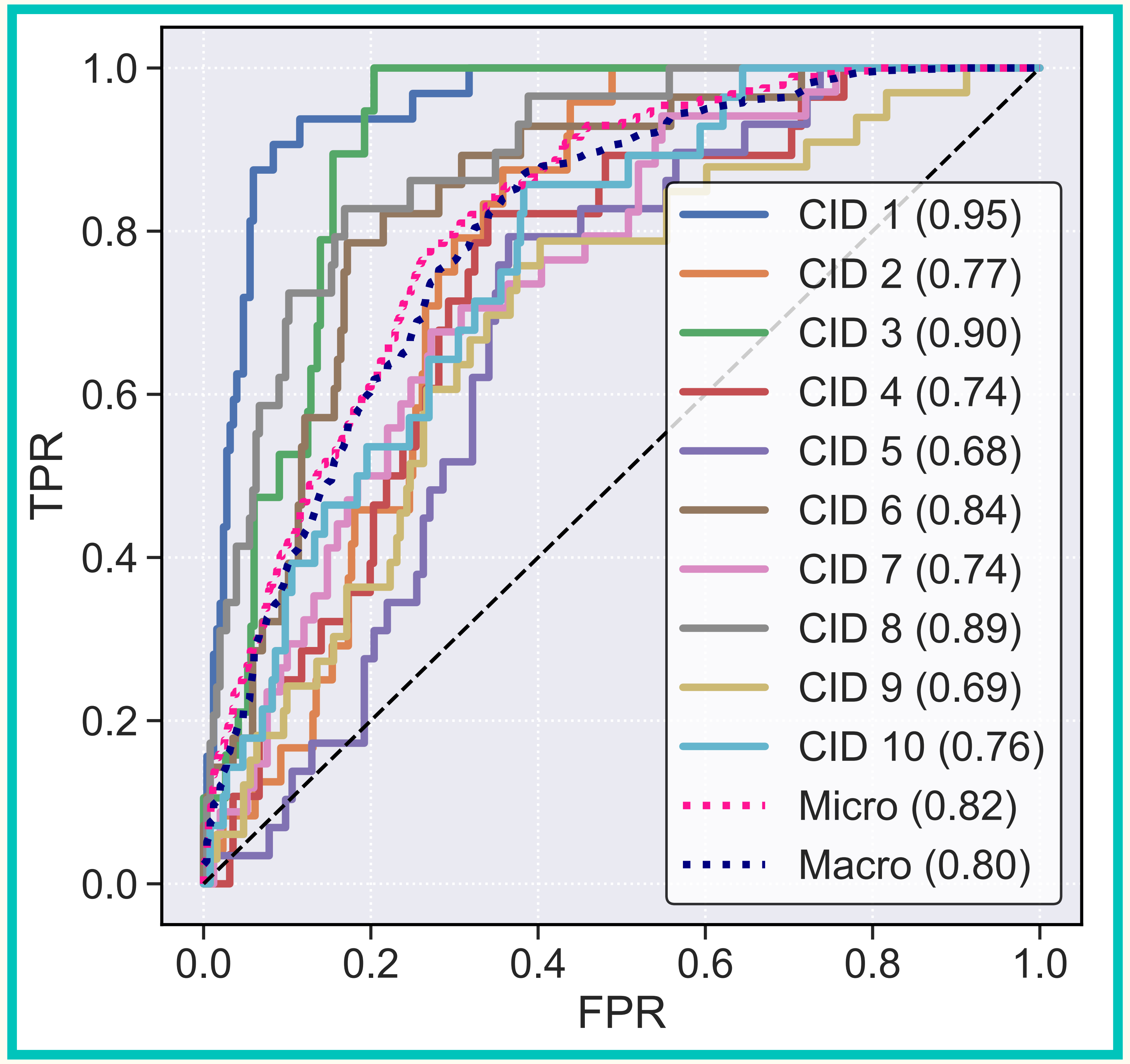
Layer-#	Layer prms.	Output shape	Parameters
Conv2D-1	3x3@64	[B, 64, 64, 96]	640
BatchNorm2D-2		[B, 64, 64, 96]	128
ReLU-3		[B, 64, 64, 96]	0
MaxPool2D-4	2x2,stride=2	[B, 64, 32, 48]	0
Conv2D-5	3x3@64	[B, 64, 32, 48]	36,928
BatchNorm2D-6		[B, 64, 32, 48]	128
ReLU-7		[B, 64, 32, 48]	0
MaxPool2D-8	2x2,stride=2	[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 - Catch-22

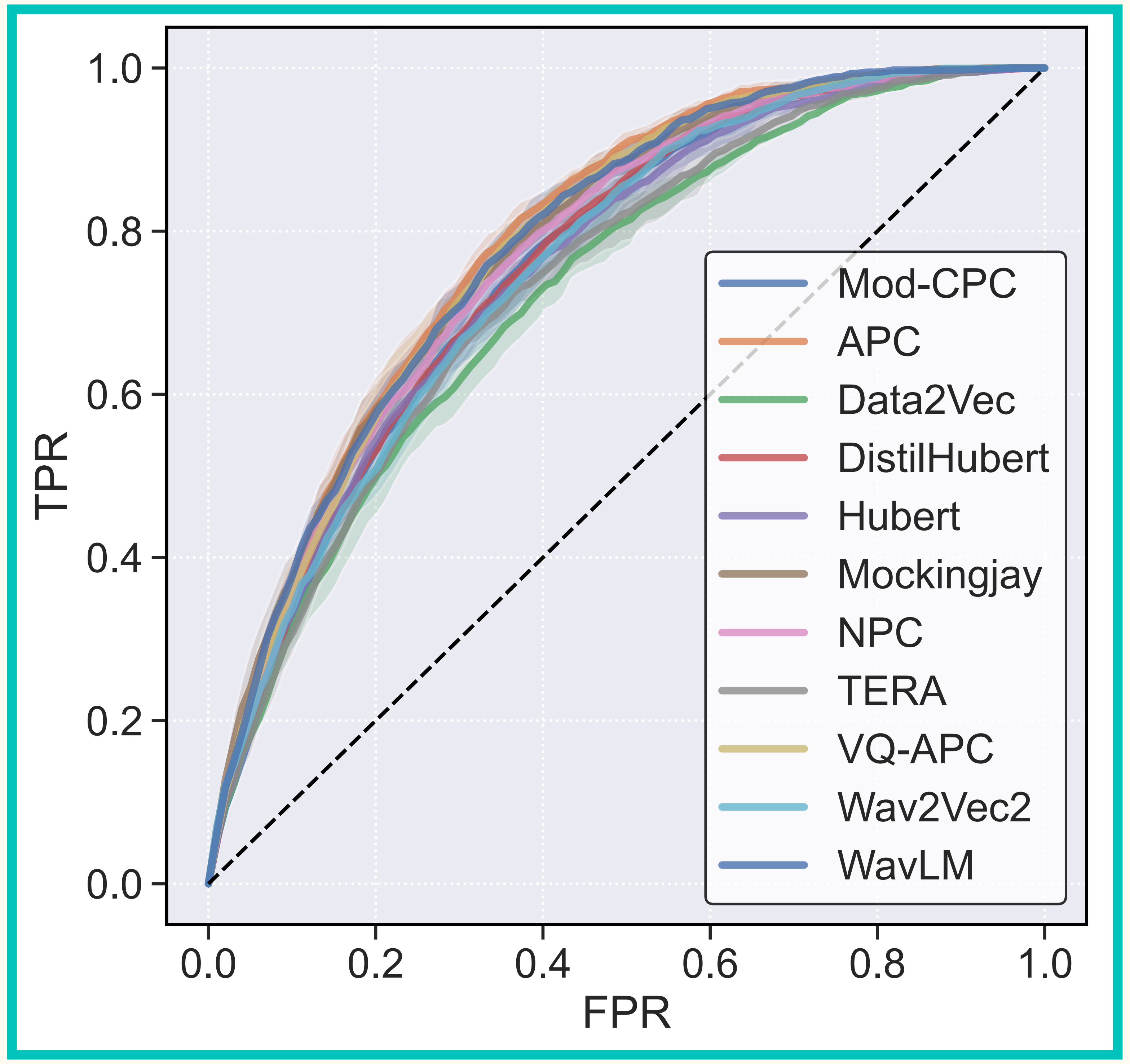
- Subset of *Highly Comparable Time-Series Analysis* (HCTSA):
  - 7700 features through signal processing methods (eg LPC, Wavlet transform).
  - Tested on: birdsongs, ecosystem monitoring, and marmoset caller identification.
  - Significant limitations: computational demands and feature redundancy.

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

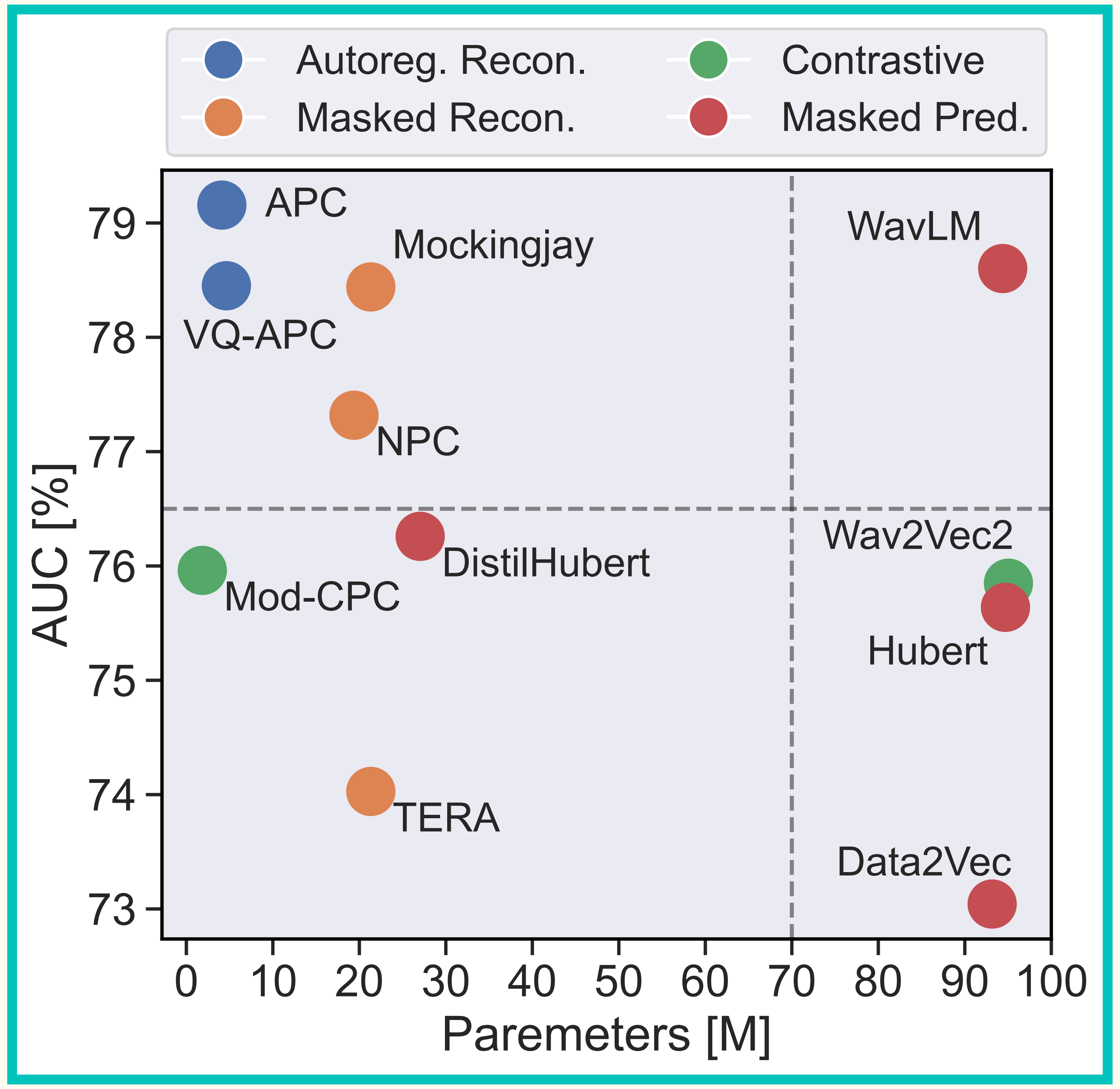
FAQ Transferability of SSLs



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



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

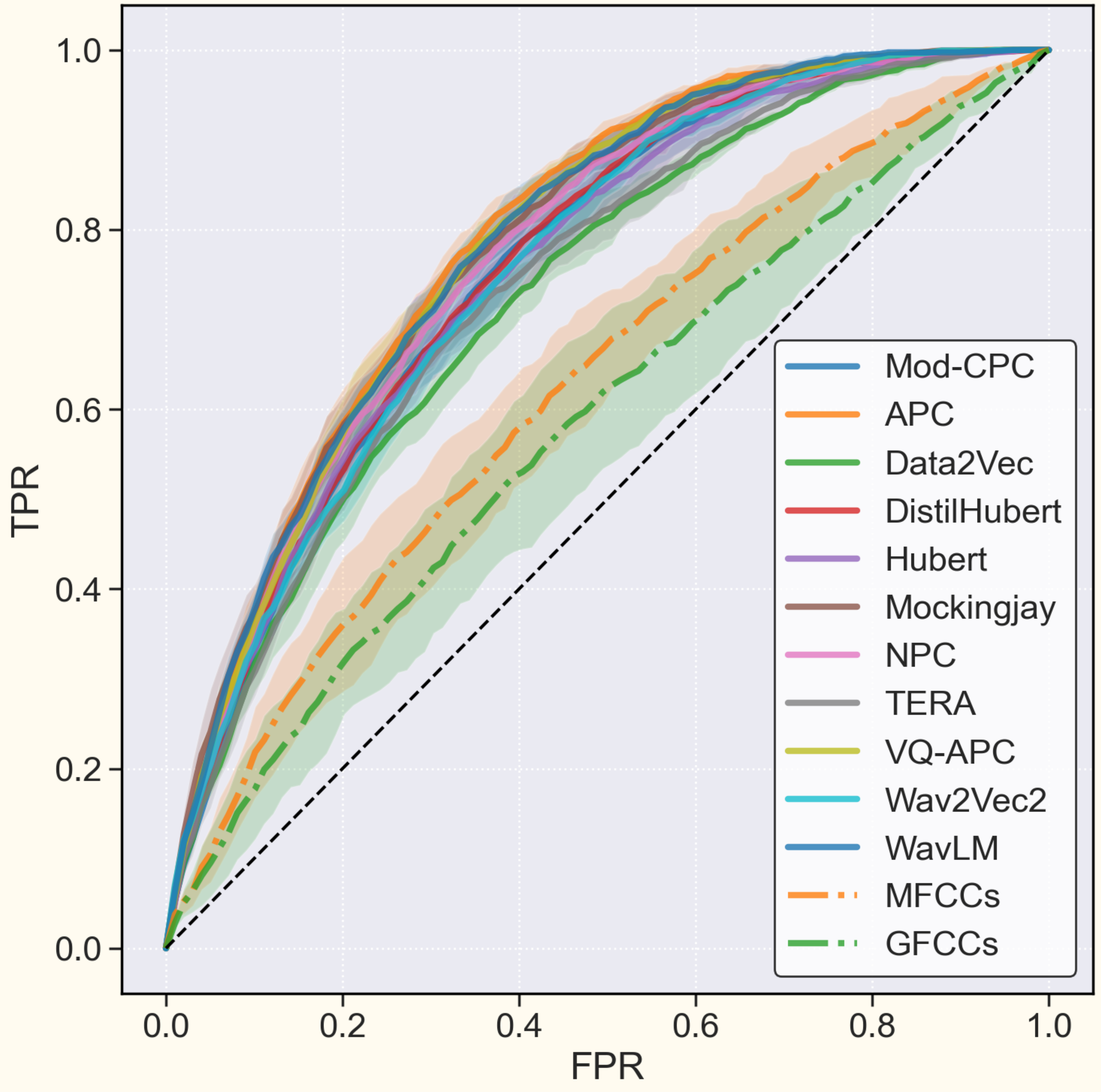


Model size against performance, divided into 4 quadrants.

FAQ Transferability of SSLs - MFCC Baseline

- MFCC:
  - Window size: 15 ms (240 samples)
  - Window shift: 5 ms (80 samples)

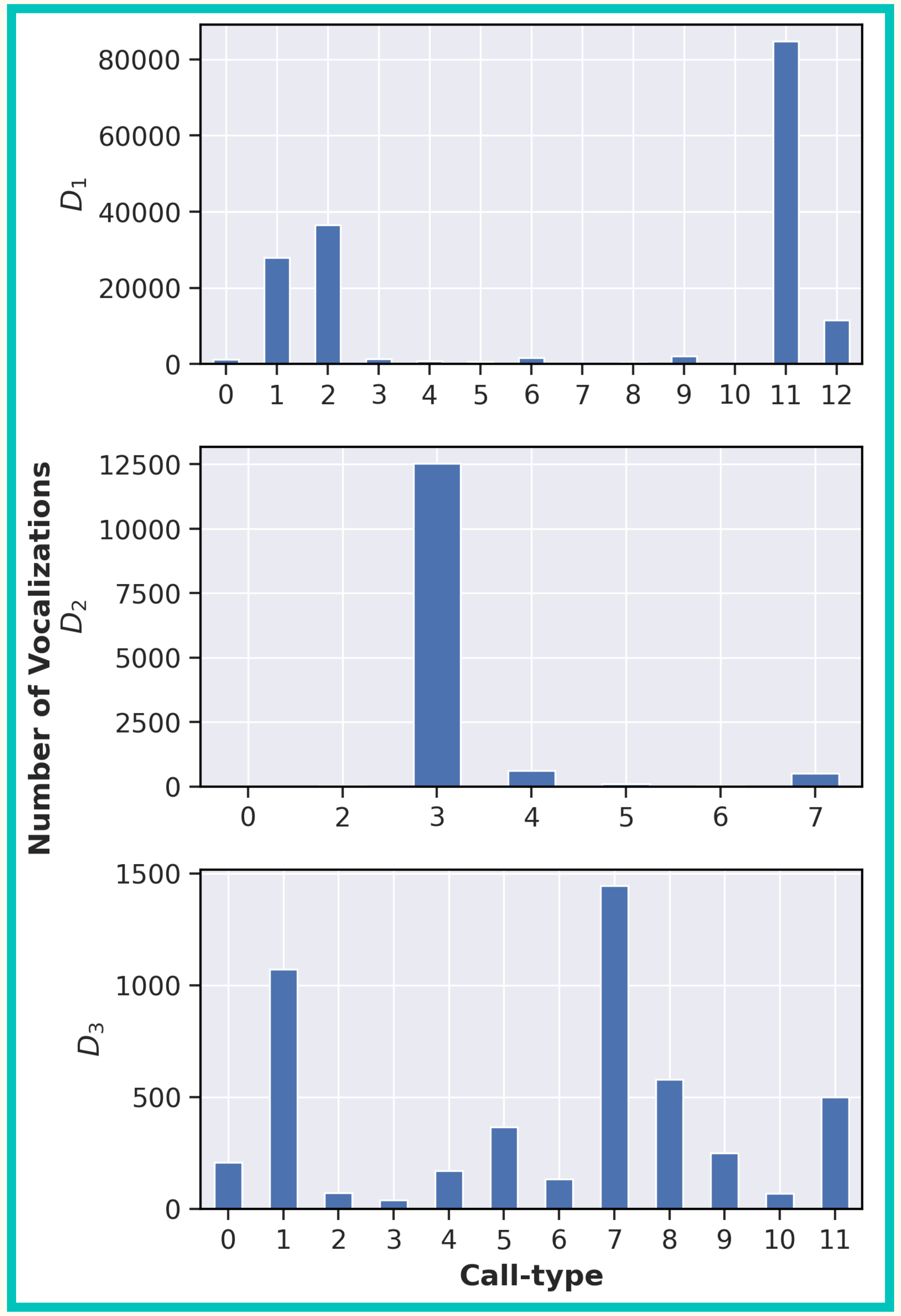
- Weaker performance compared to pre-trained SSL models.



Marmoset Vocalization Task Metrics

Marmoset Vocalization Task Metrics

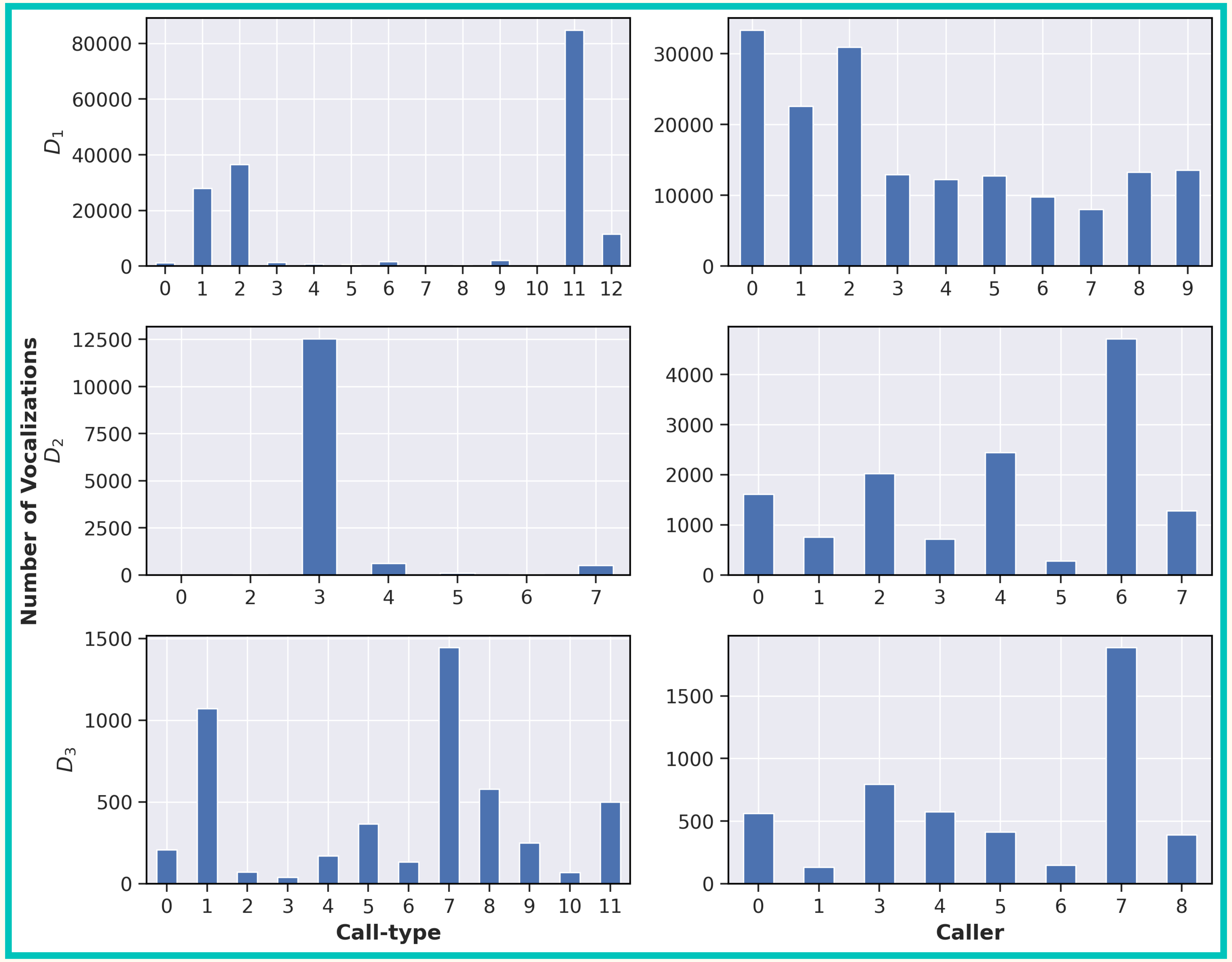
- Imbalanced class distribution !



Dataset class distributions.

Marmoset Vocalization Task Metrics

- Imbalanced class distribution !



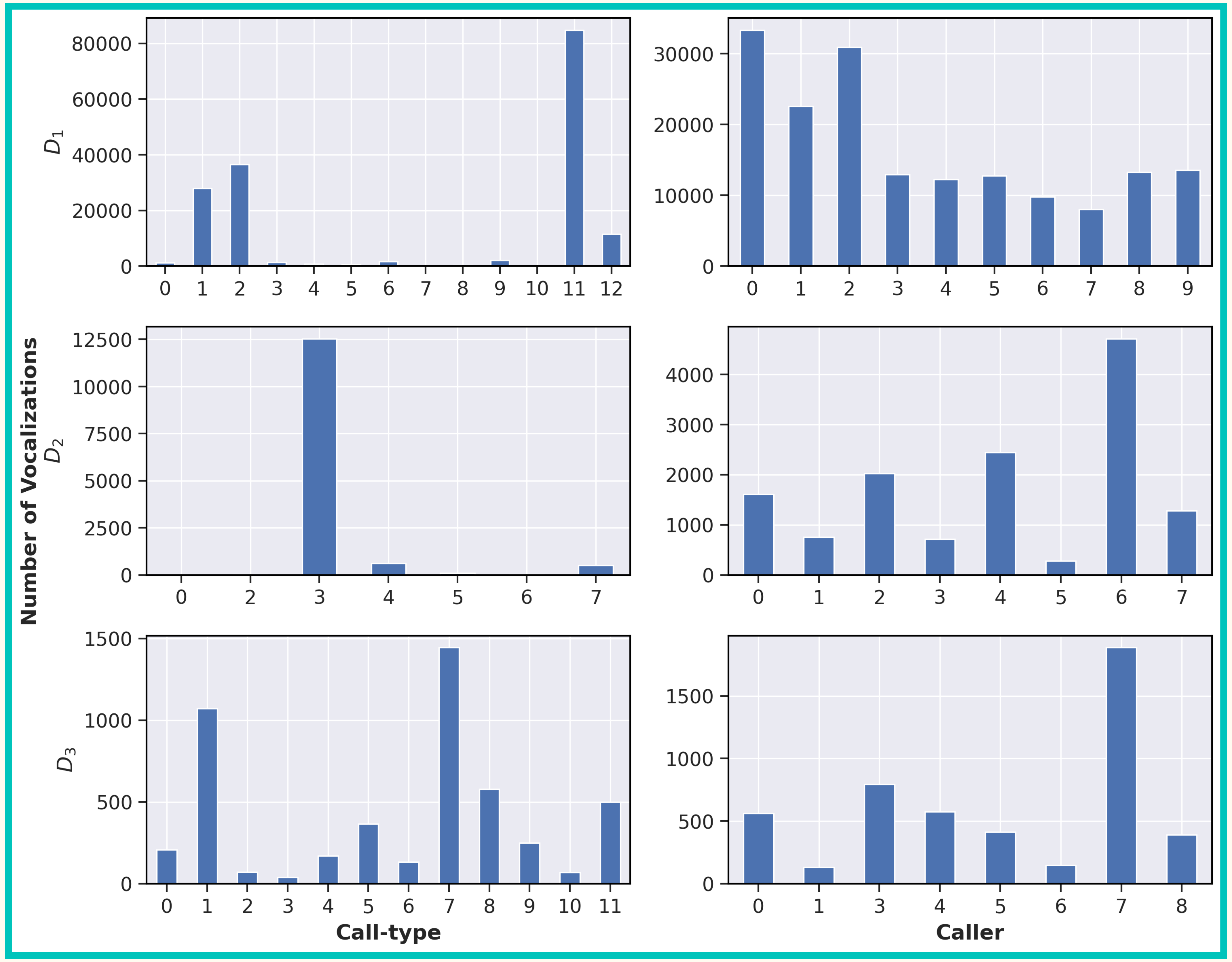
Dataset class distributions.

Marmoset Vocalization Task Metrics

- Imbalanced class distribution !

Metric:

- Unweighted Average Recall (UAR).



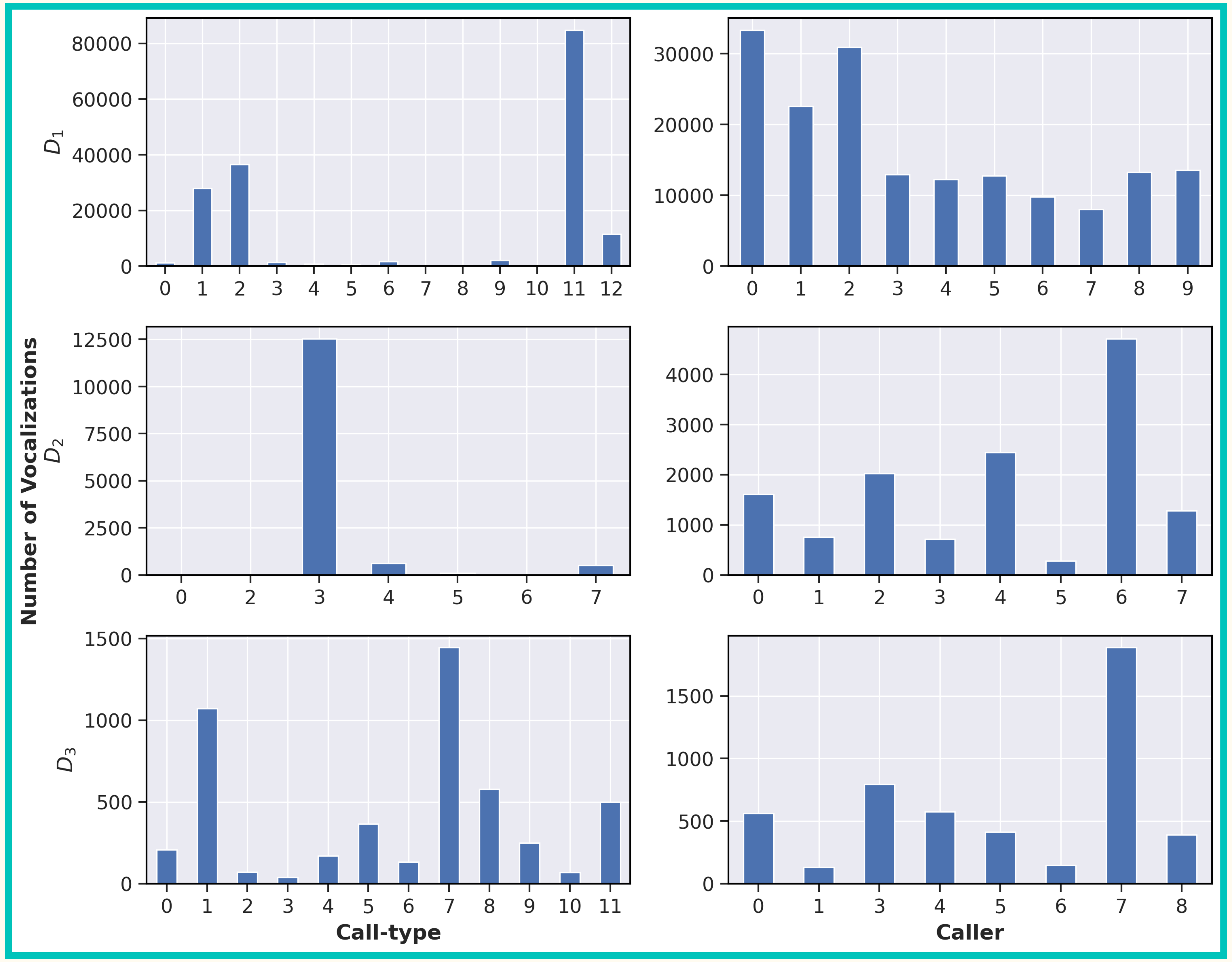
Dataset class distributions.

Marmoset Vocalization Task Metrics

- Imbalanced class distribution !

Metric:

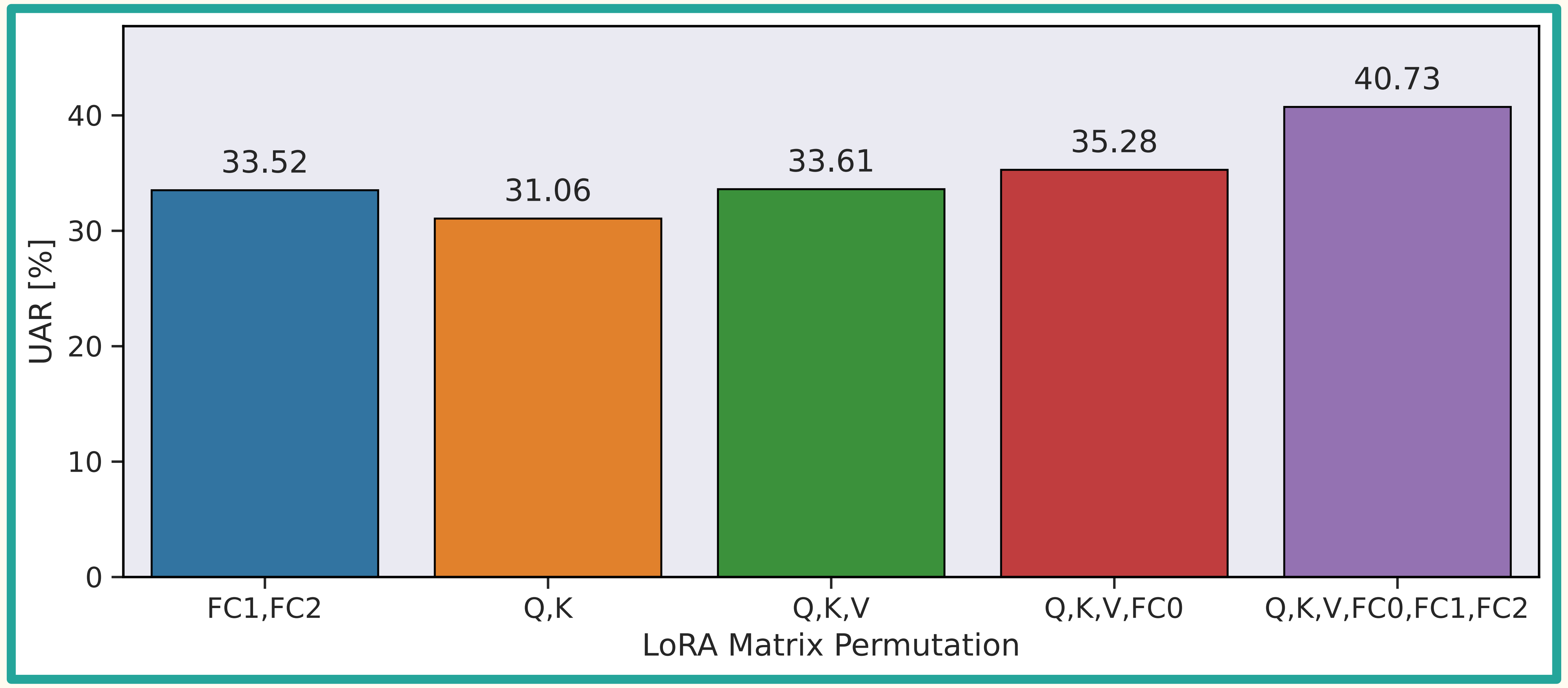
- Unweighted Average Recall (UAR).
- Accounts for class imbalance by treating each class equally.



Dataset class distributions.

FAQ Adaptation - Matrix Selection

- UAR score achieved for each of the five different LoRA adapter matrix configurations.
- Monotonic progression: performance increases as projection modules are tuned.
- Fine-tuning only the query and key projections yields the lowest UAR, with each successive addition leading to higher scores.

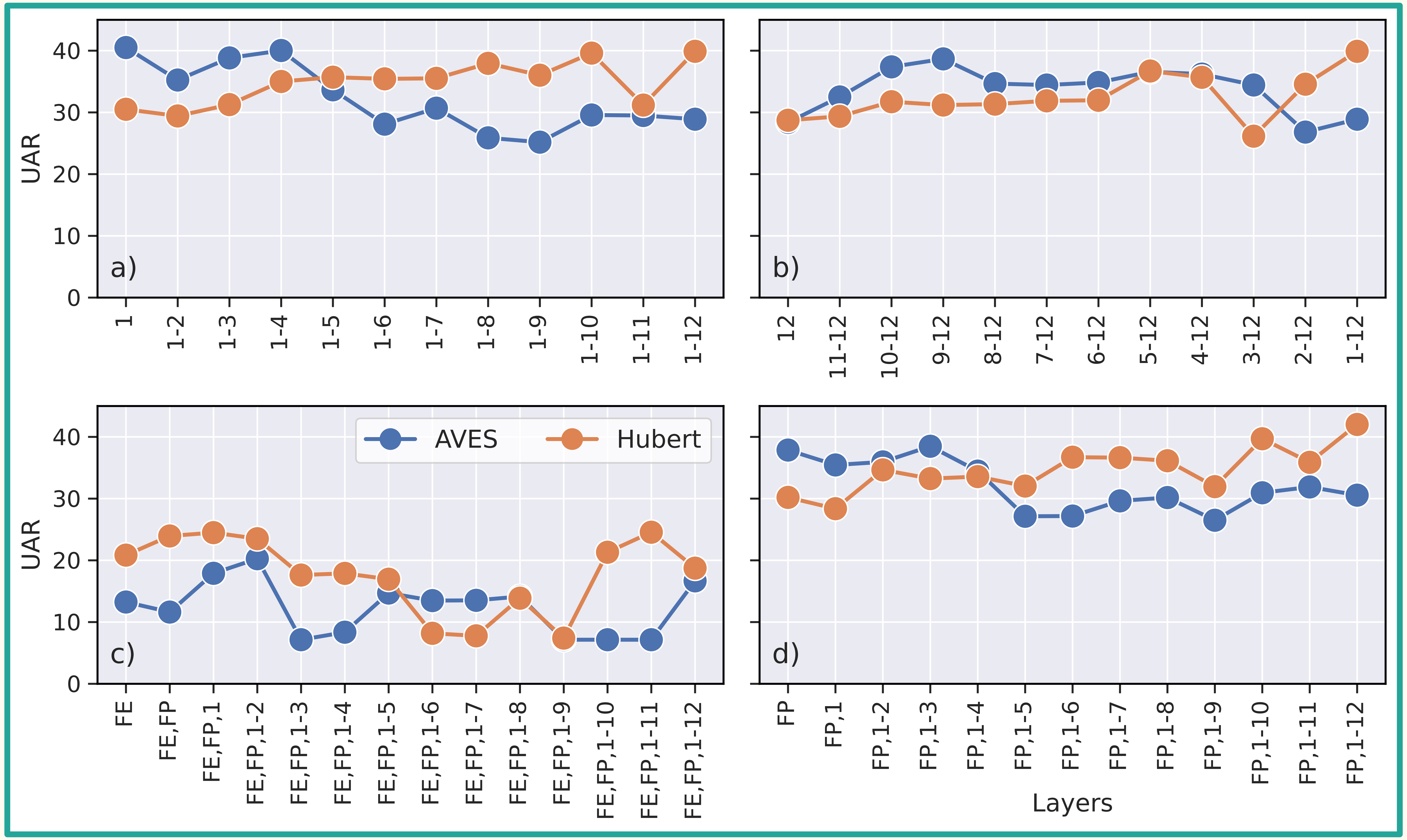


Abzaliev dataset for CTID.

- Best UAR [%] for each LoRA adapter configuration on layers 1–12.
- Fine-tuning all matrices yields the best performance.

FAQ Adaptation - Layer and Module Selection

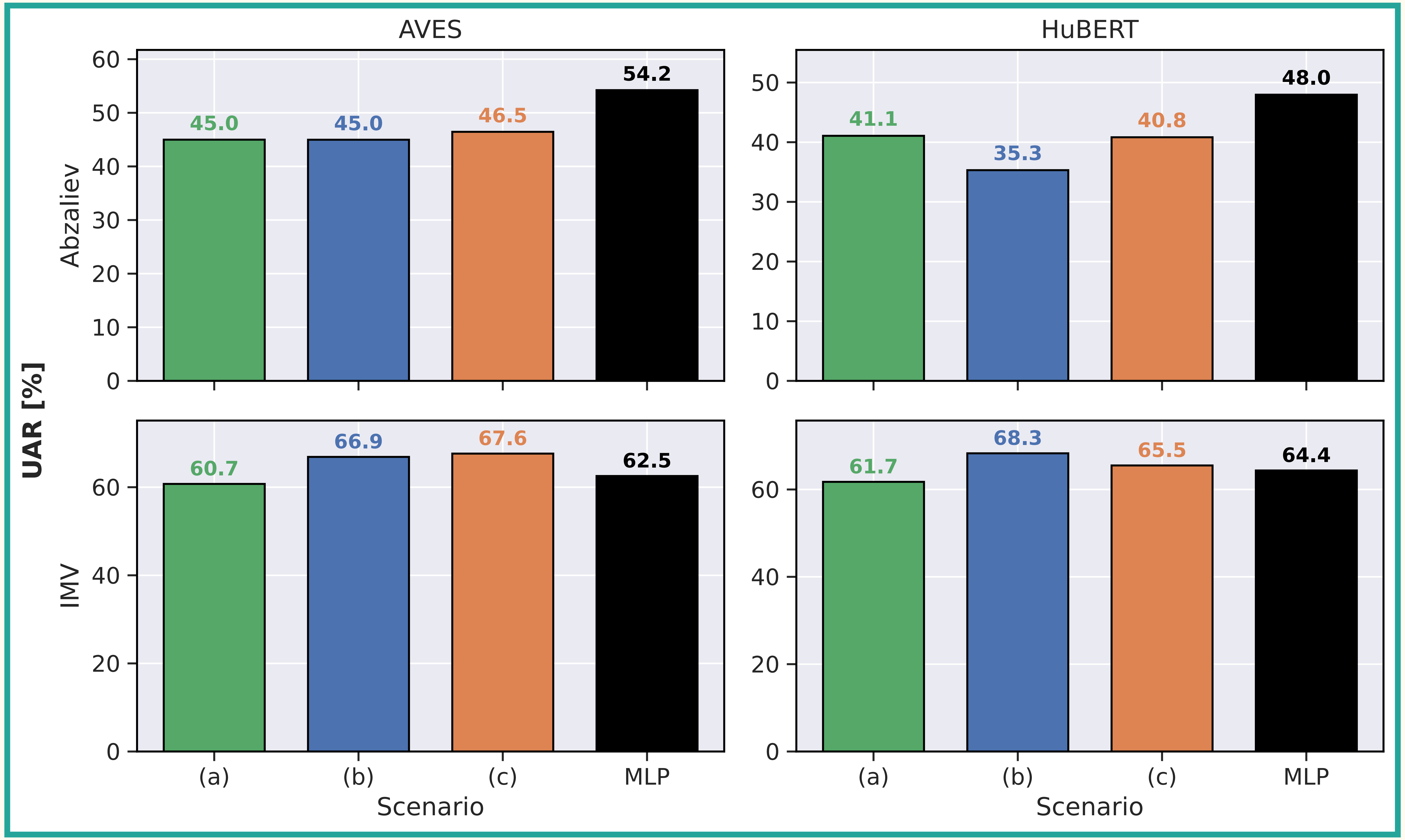
- Fine-tuning the feature extraction (FE) layers severely degrades performance.
- Fine-tuning the feature projection (FP) alone does not improve performance.
- Bottoms-up & top-down layer selection strategies yield similar results.
- Neither AVES nor HuBERT consistently outperforms the other across all layer selections.



Layer selection strategy UAR [%] results.
(a) bottoms-up, (b) top-down, (c) FE + FP + bottoms-up, (d) FP + bottoms-up.

FAQ Adaptation - MLP vs. Linear Layer

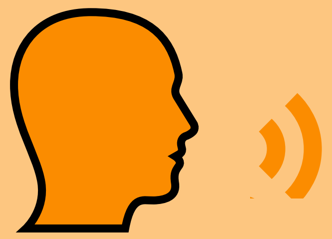
- **Abzaliev**: MLP outperforms single-layer models.
- **IMV**: single-layer models outperform MLP.
  - Cannot draw general conclusions.
  - Increased capacity may help in some cases, it may not be universally beneficial.



Model Adaptation

Fine-tuning

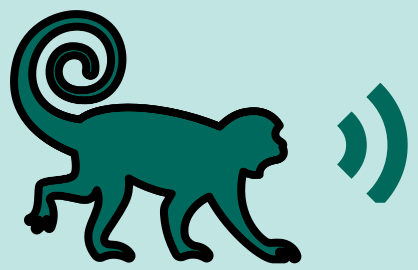
on human speech



Can it provide an additional inductive bias, useful for bioacoustics tasks ?

Fine-tuning

on bioacoustics



Does fine-tuning on the downstream bioacoustic data yields better results ?

FAQ - Fine-Tuning on Human Speech

FAQ - Fine-Tuning on Human Speech

- SSL representations: strong performance on bioacoustics tasks without FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls and identities.

FAQ - Fine-Tuning on Human Speech

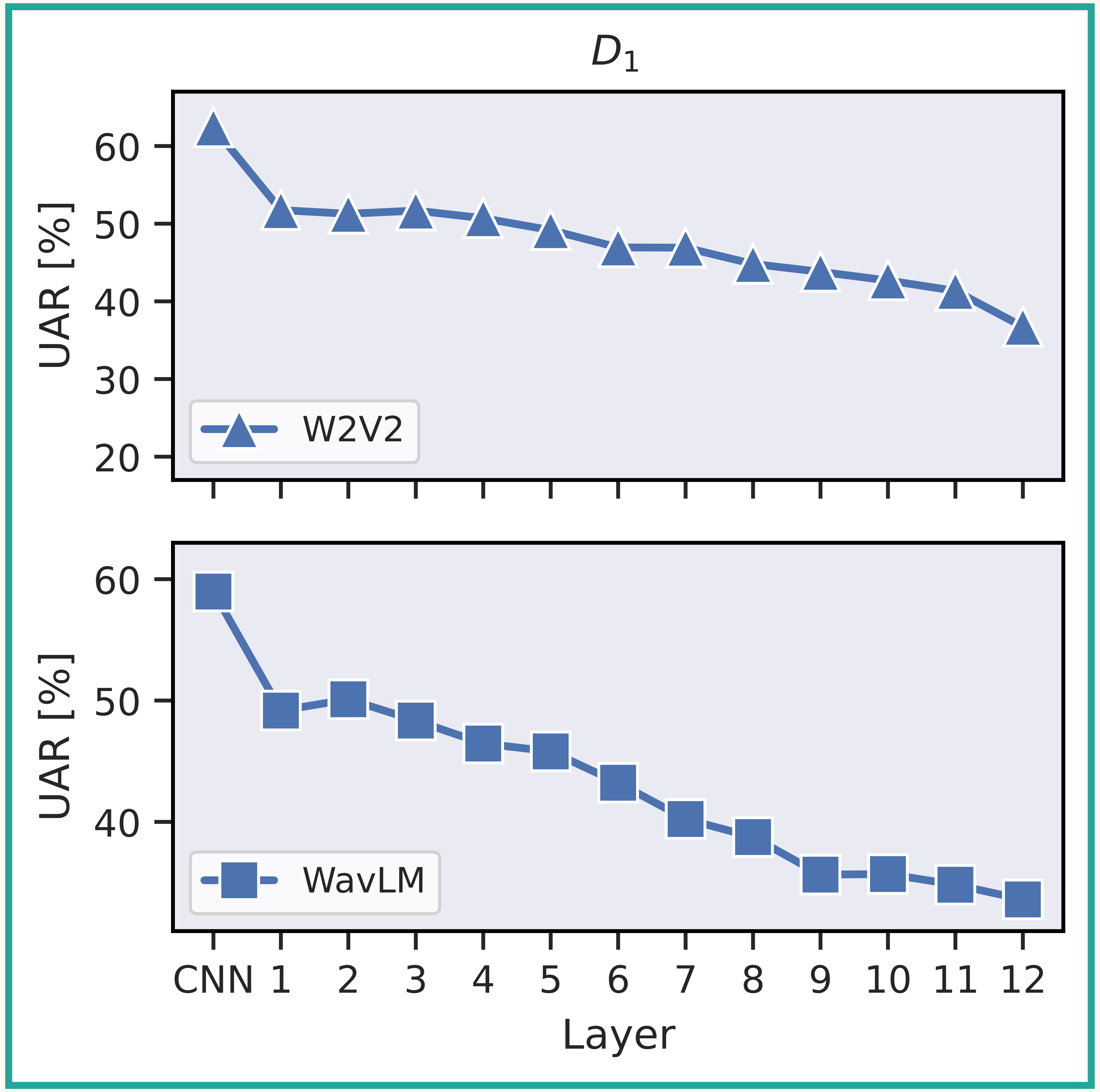
- SSL representations: strong performance on bioacoustics tasks without FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls and identities.
- FT'ing in supervised framework: forces model to learn novel, specialized patterns.
  - Phonetic distinctions and temporal structures → can lead to performance gains.

FAQ - Fine-Tuning on Human Speech

- SSL representations: strong performance on bioacoustics tasks without FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls and identities.
- FT'ing in supervised framework: forces model to learn novel, specialized patterns.
  - Phonetic distinctions and temporal structures → can lead to performance gains.
- As speech and animal calls both encode *structured vocal* and *linguistic* information
  - SSL models *fine-tuned* on ASR may provide an additional inductive bias, enhancing the model's ability to recognize complex features in bioacoustics data.

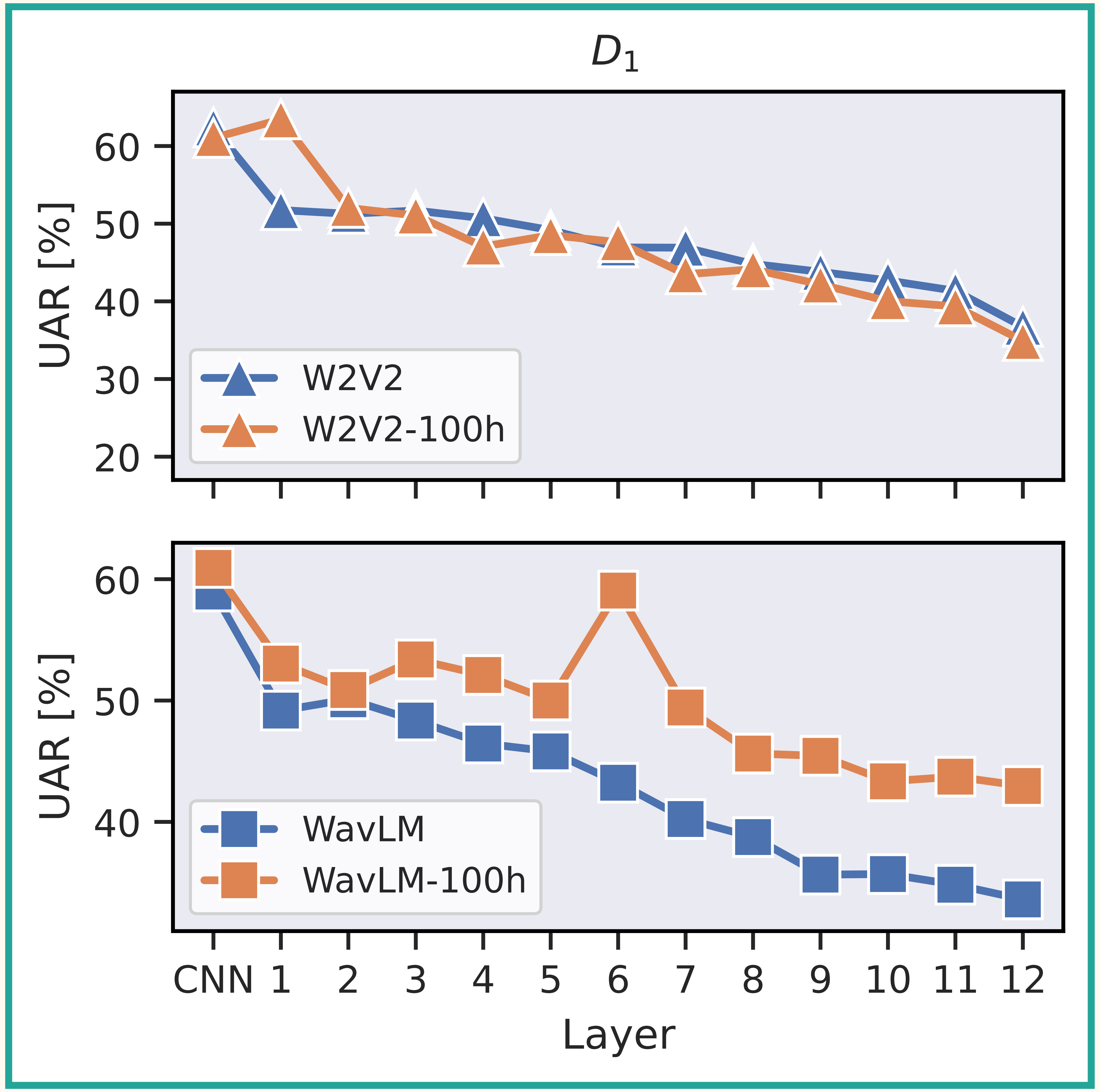
FAQ - Fine-Tuning on Human Speech

FAQ - Fine-Tuning on Human Speech



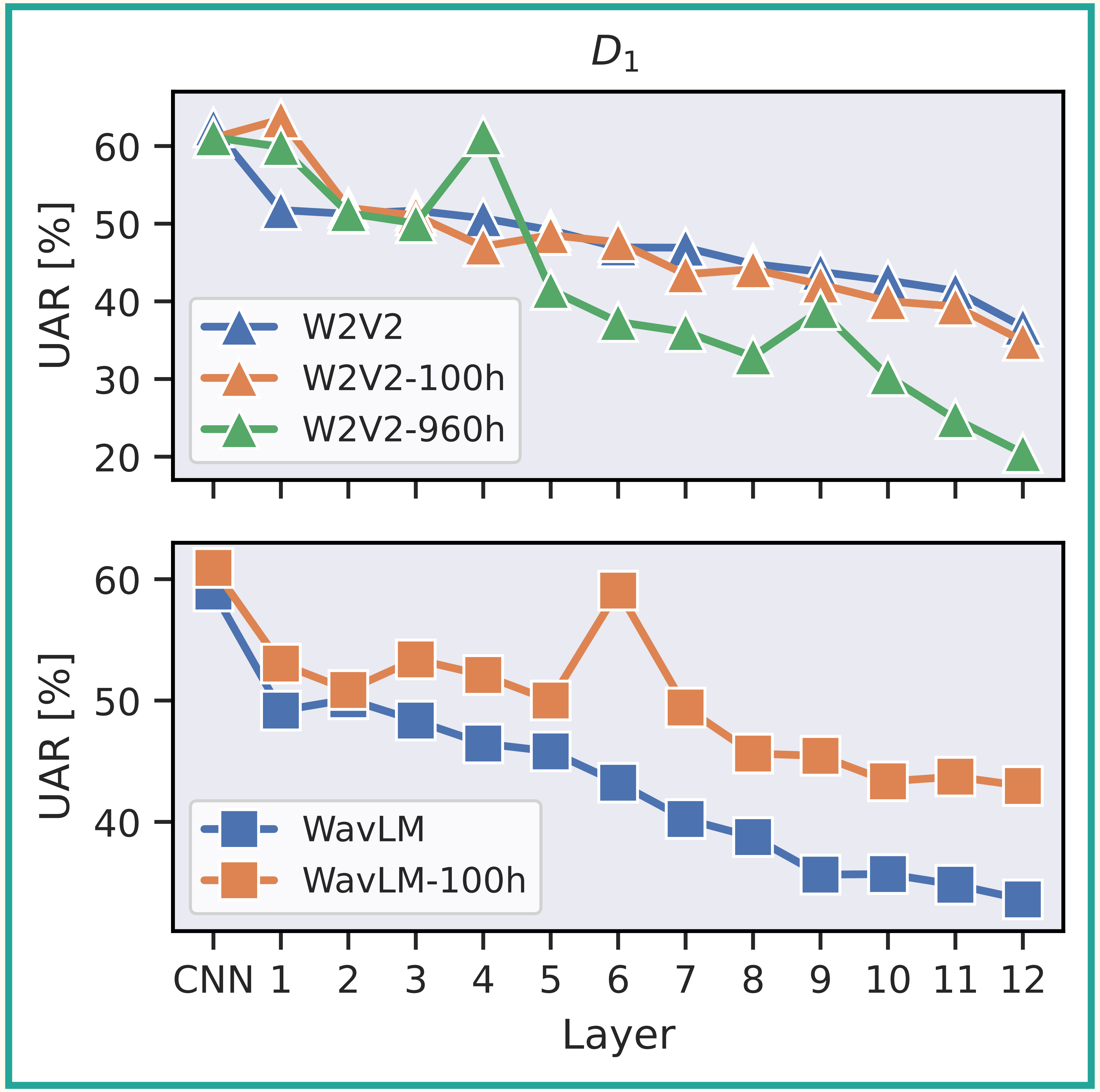
W2V2 (▲) and WLM (■)

FAQ - Fine-Tuning on Human Speech



W2V2 (▲) and WLM (■) against their FT'd versions

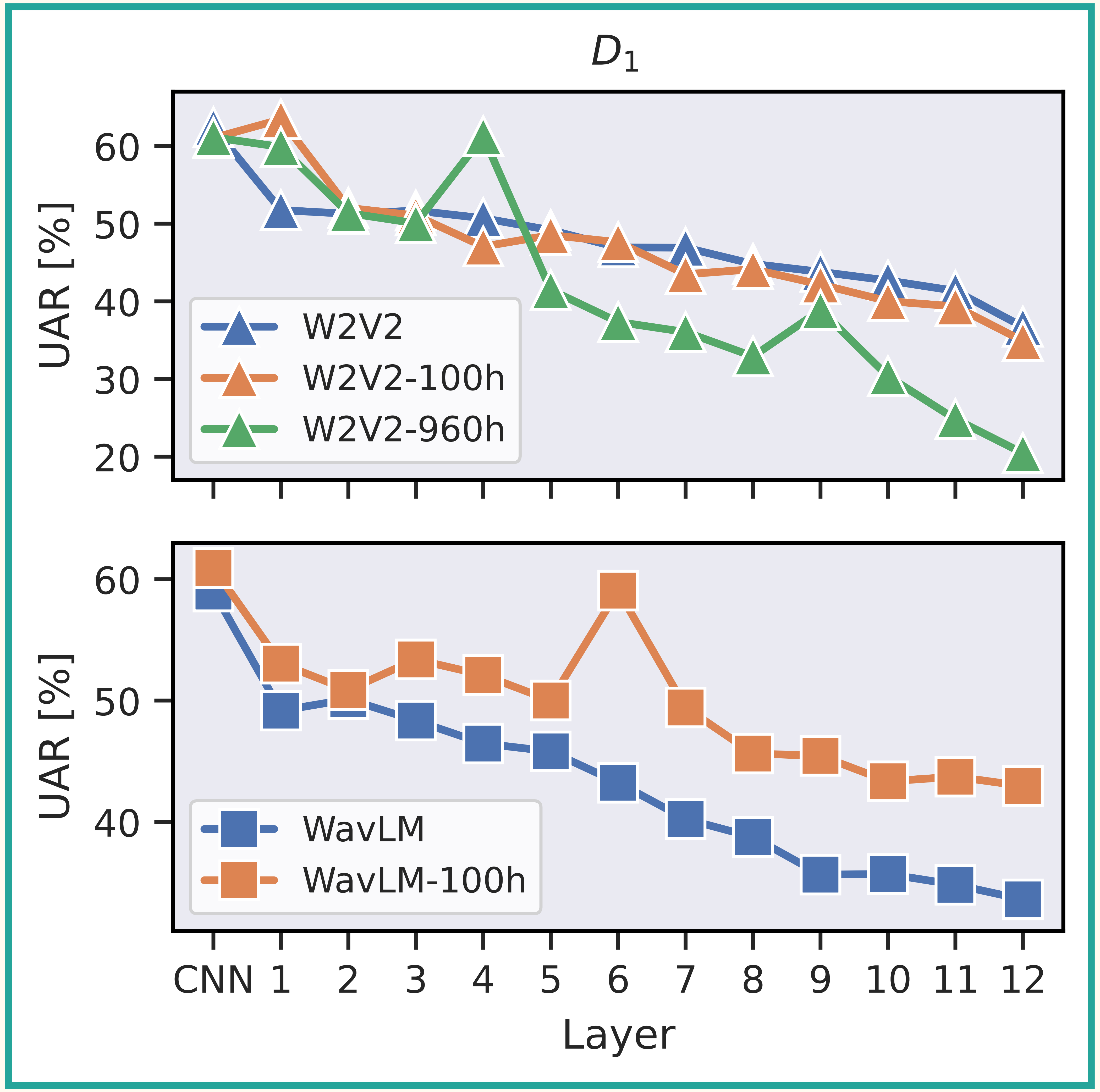
FAQ - Fine-Tuning on Human Speech



W2V2 (▲) and WLM (■) against their FT'd versions

FAQ - Fine-Tuning on Human Speech

Fine-tuning yields mixed effects across both models.

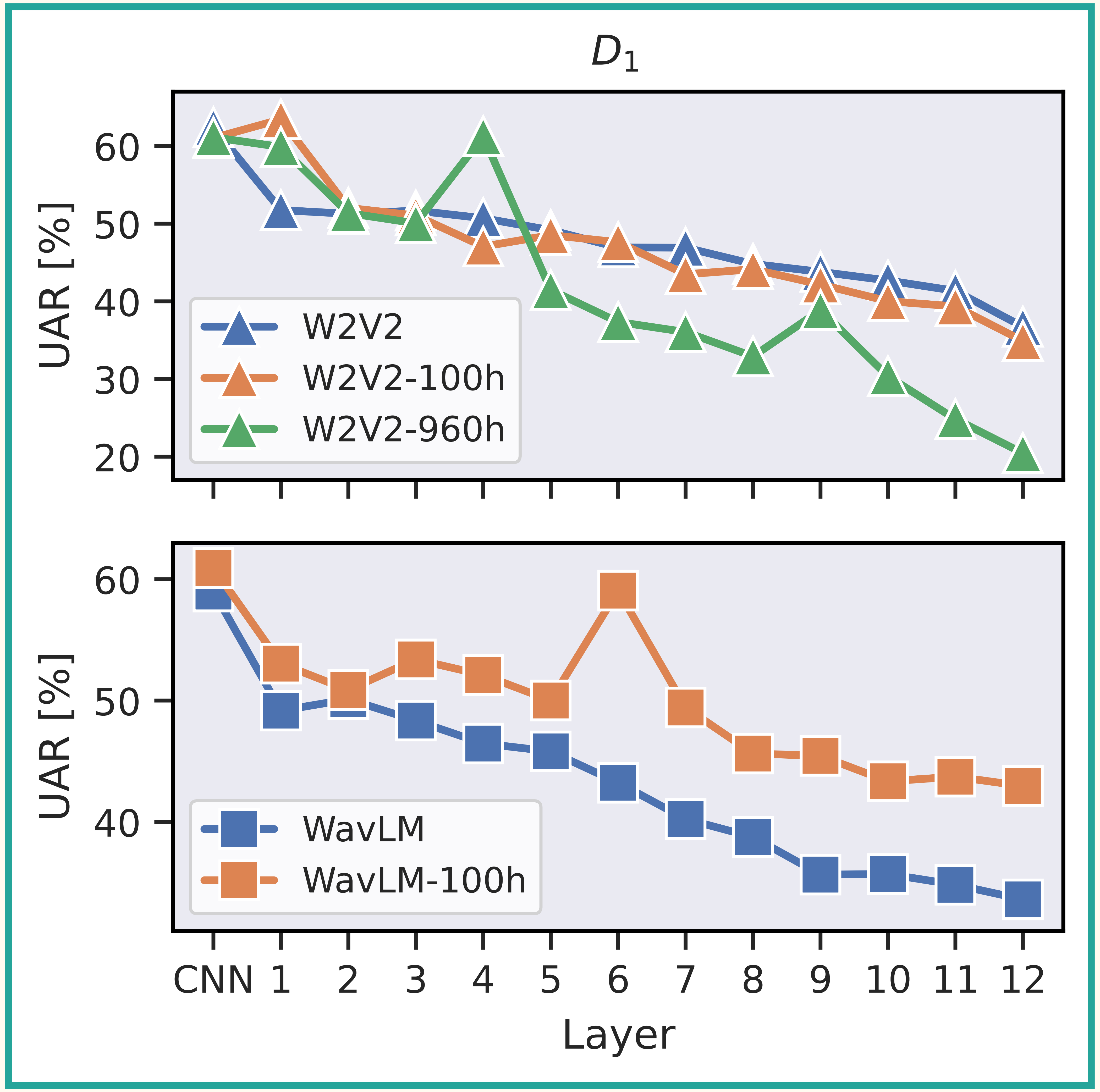


W2V2 (▲) and WLM (■) against their FT'd versions

FAQ - Fine-Tuning on Human Speech

Fine-tuning yields mixed effects across both models.

- FT'd models don't consistently outperform their base ones.

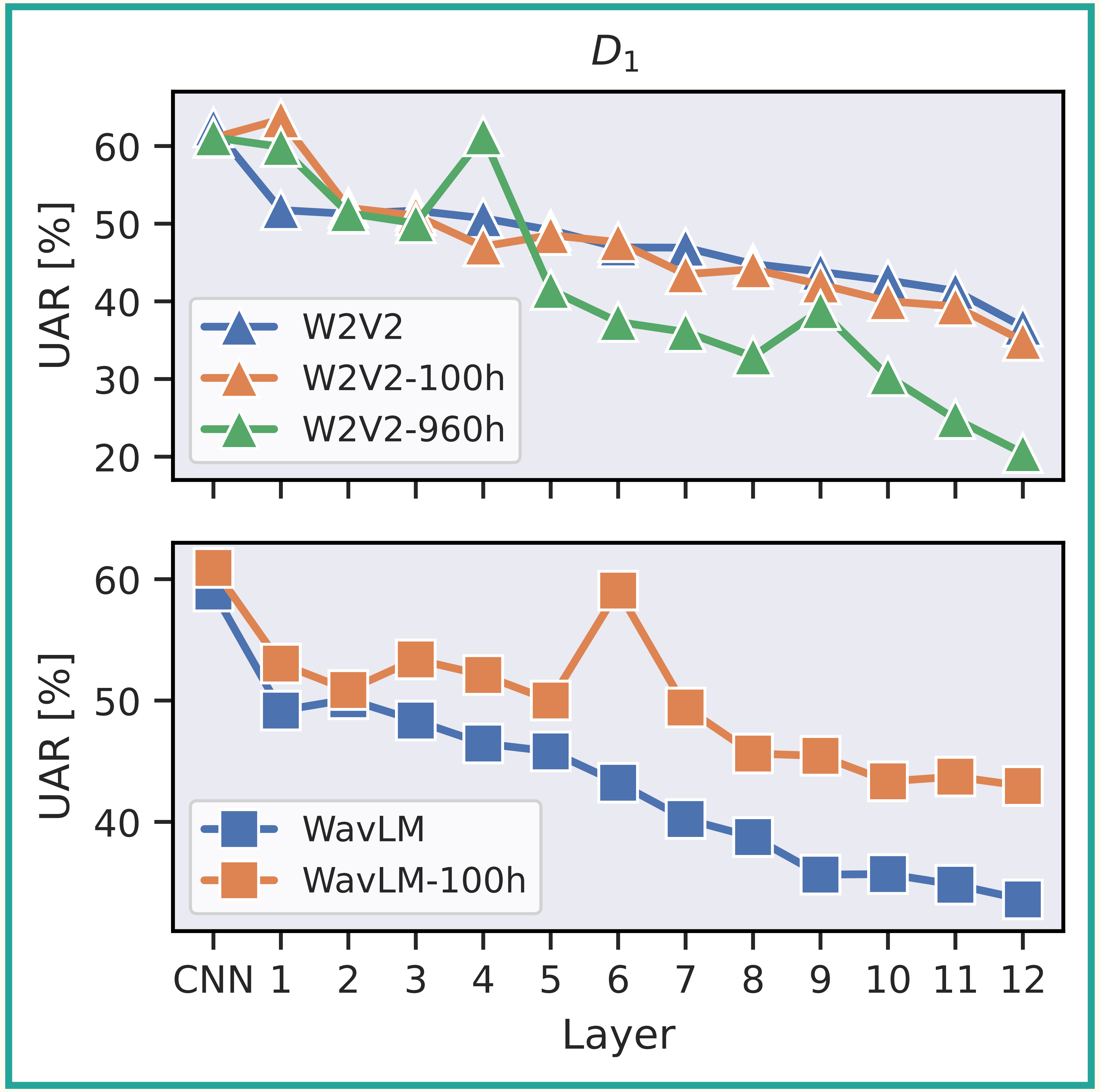


W2V2 (▲) and WLM (■) against their FT'd versions

FAQ - Fine-Tuning on Human Speech

Fine-tuning yields mixed effects across both models.

- FT'd models don't consistently outperform their base ones.
- FT'ing on more speech data can lead to a decline in performance in later layers, e.g. 960h-W2V2.

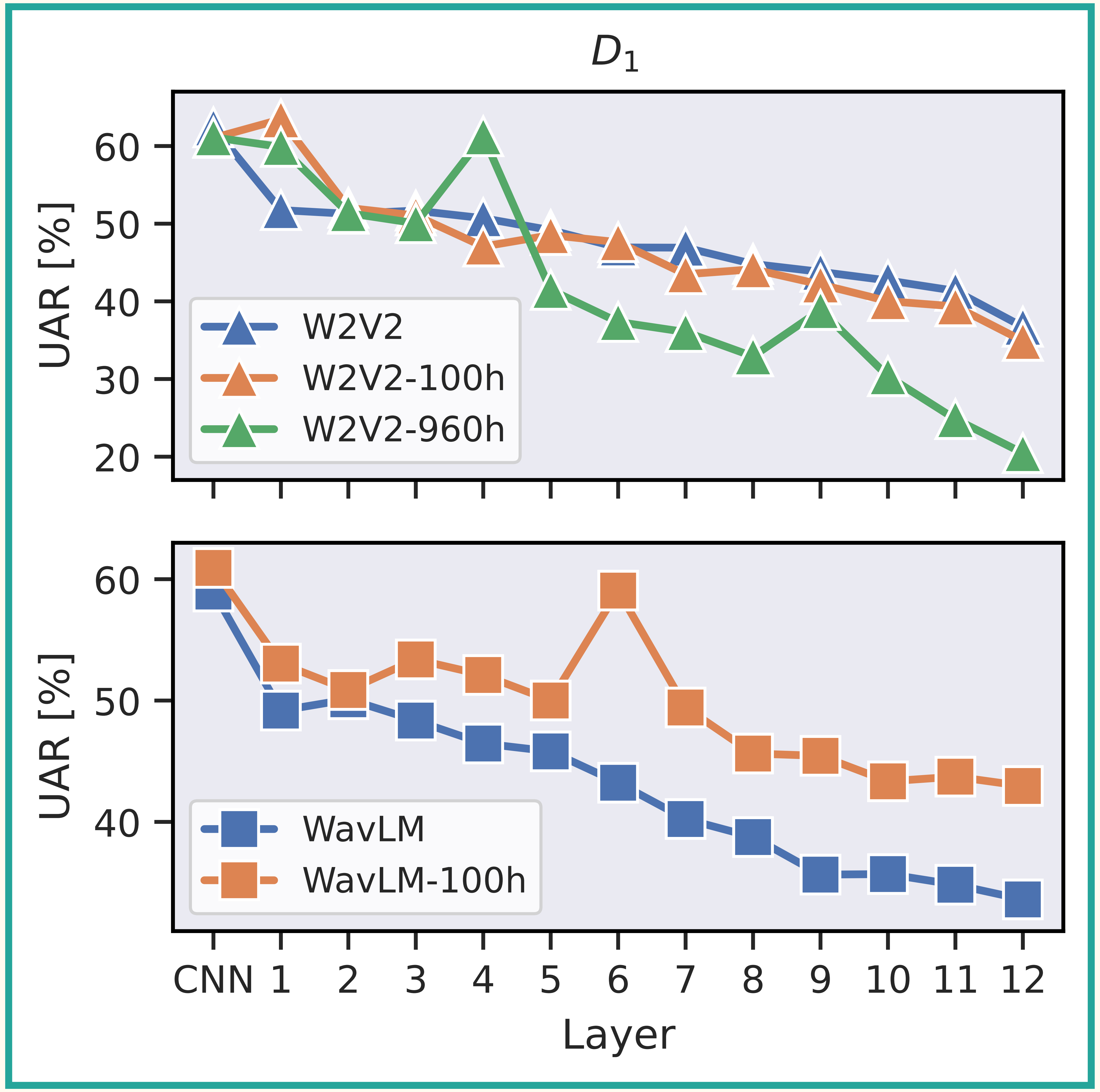


W2V2 (▲) and WLM (■) against their FT'd versions

FAQ - Fine-Tuning on Human Speech

Fine-tuning yields mixed effects across both models.

- FT'd models don't consistently outperform their base ones.
- FT'ing on more speech data can lead to a decline in performance in later layers, e.g. 960h-W2V2.
  - FT on ASR may push models to learn task-specific features that don't generalize well to bioacoustic tasks.



W2V2 (▲) and WLM (■) against their FT'd versions

Comparative Analysis

- Best scores from AVES and HuBERT.
  - HuBERT's representations are robust for CTID tasks across different species.
- Best scores are from the PT category.
  - Fine-tuning PT'd speech models on an ASR does not consistently bring us any advantage over PT'd alone.
  - PT'd representations may already be 'optimized', and FT'ing might not always yield significant benefits.

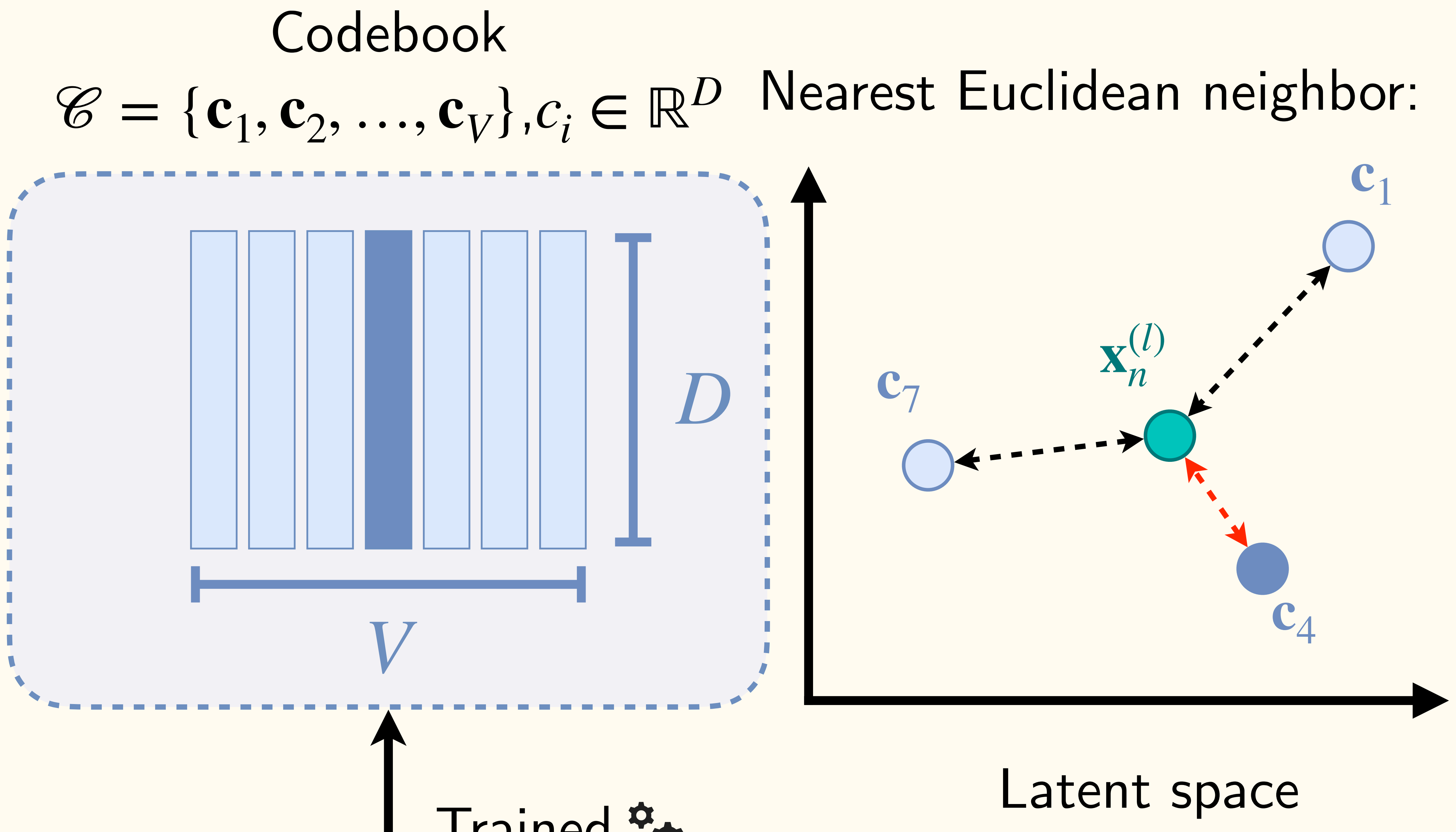
Type	$\mathcal{F}$	IMV
PT	AVES	62.54
	HuBERT	**64.35**
	WavLM	58.98
	W2V2	62.40
PT + FT	WavLM-100h	60.93
	W2V2-100h	<u>63.44</u>
	W2V2-960h	61.25
	Fusion	62.48

UAR scores [%] on the best feature layer, on *Test*.
Best performance is **bolded**, second best is <u>underlined</u>.

FAQ - Vector Quantization Pipeline

Nearest Euclidean Neighbour: $q[\boldsymbol{x}_n^{(l)}] = \underset{i \in \{1,2,\ldots,V\}}{\arg\min} \|\boldsymbol{x}_n^{(l)} - \boldsymbol{c}_i\|_2^2$

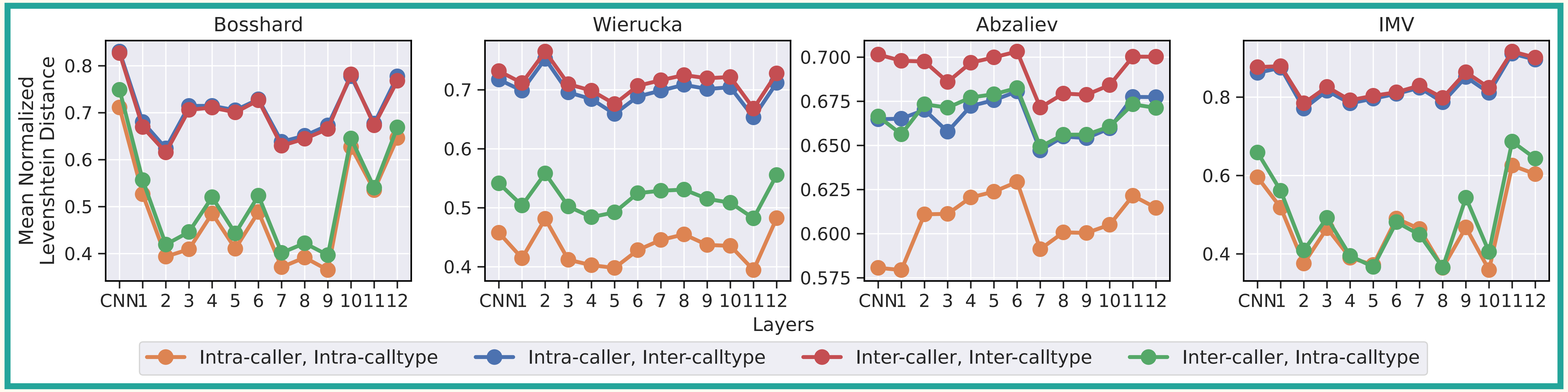
VQ Loss: $\mathcal{L}_{\text{VQ}} = \underbrace{\|\text{sg}[\boldsymbol{x}_n^{(l)}] - \boldsymbol{c}_k\|_2^2}_{\text{Codebook Loss}} + \underbrace{\beta\,\|\boldsymbol{x}_n^{(l)} - \text{sg}[\boldsymbol{c}_k]\|_2^2}_{\text{Commitment Loss}}.$



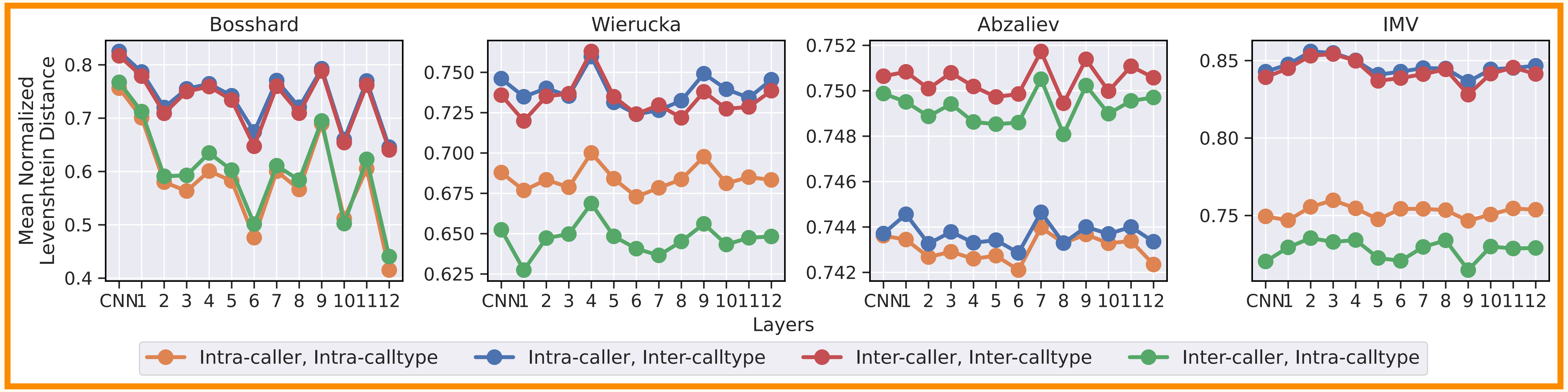
FAQ - Token Sequence Distance Analysis

- Levenshtein distance across token sequences:

VQ



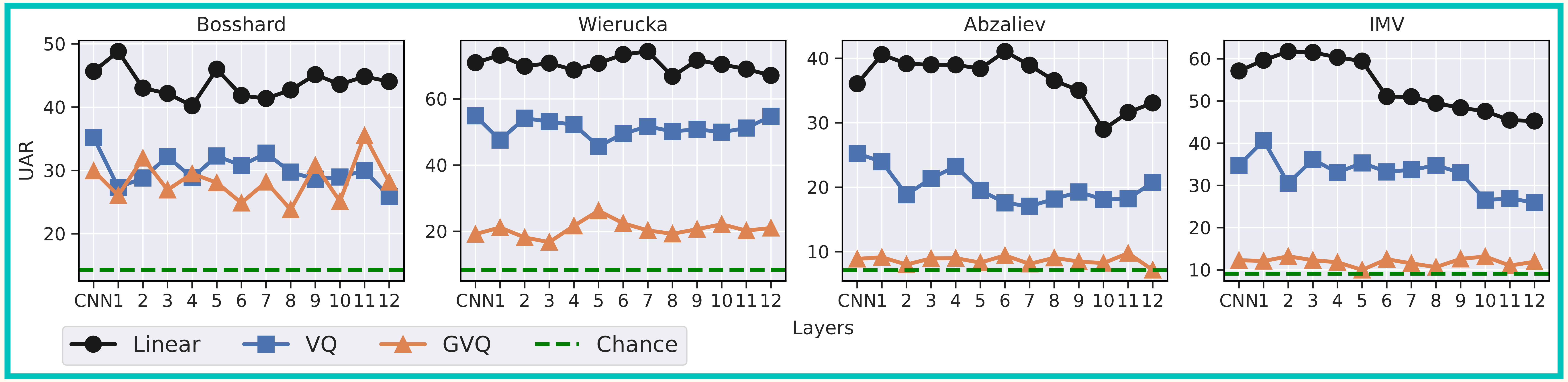
GVQ



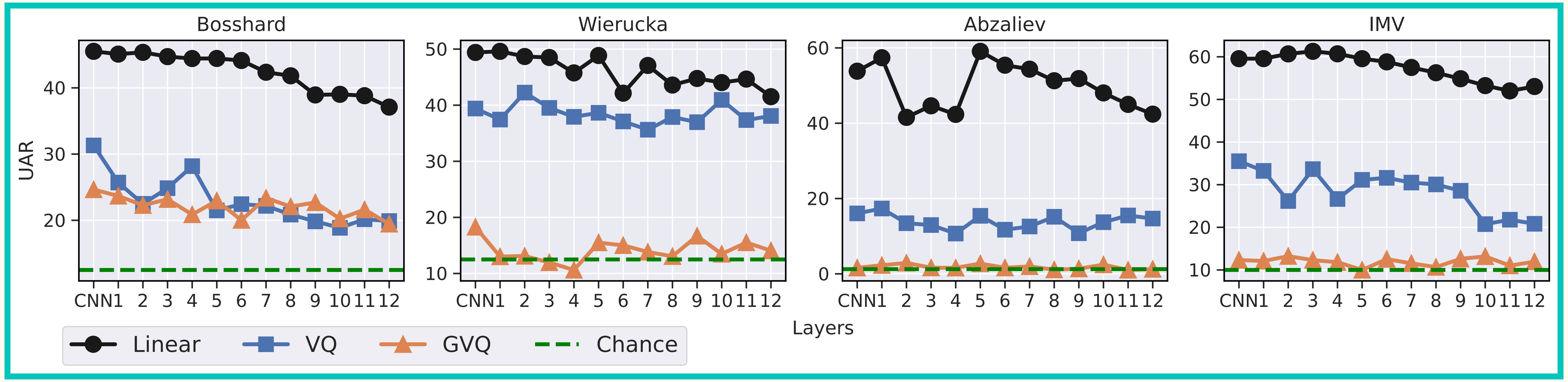
FAQ - Token Sequence Classification

- *k*-NN based sequence classification using Levenshtein distance as metric.

CTID

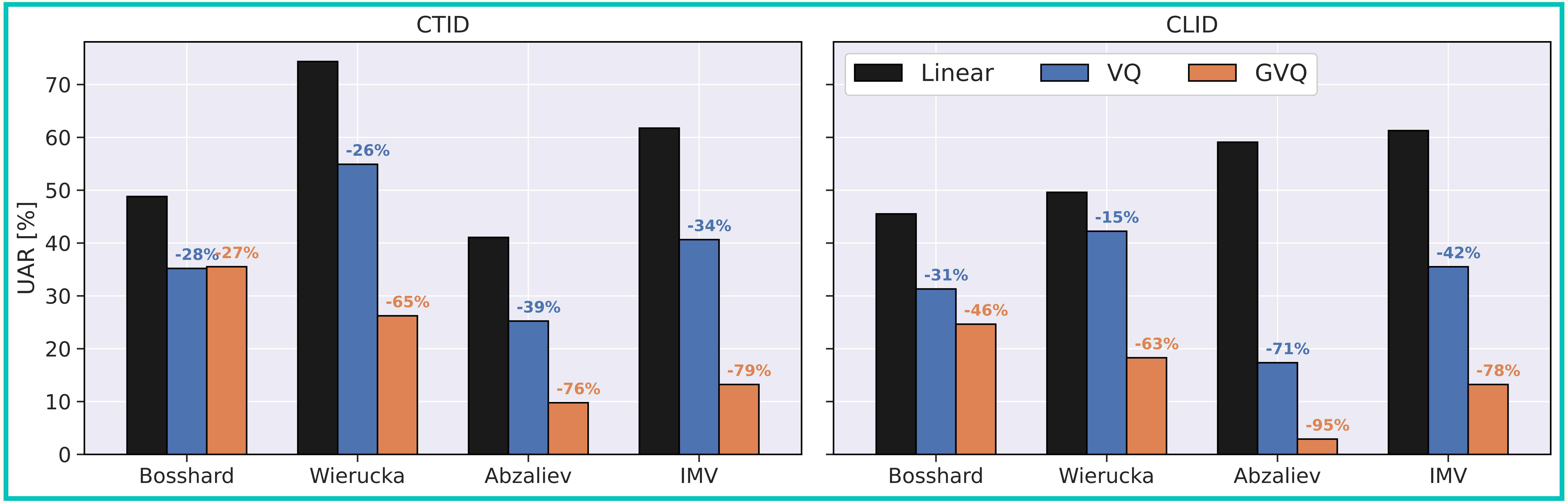


CLID



FAQ - Token Sequence Classification (Best Layers)

- Classification performance drop: linear layer vs token sequences (VQ, GVQ).



Best UAR results across layers for CTID and CLID.