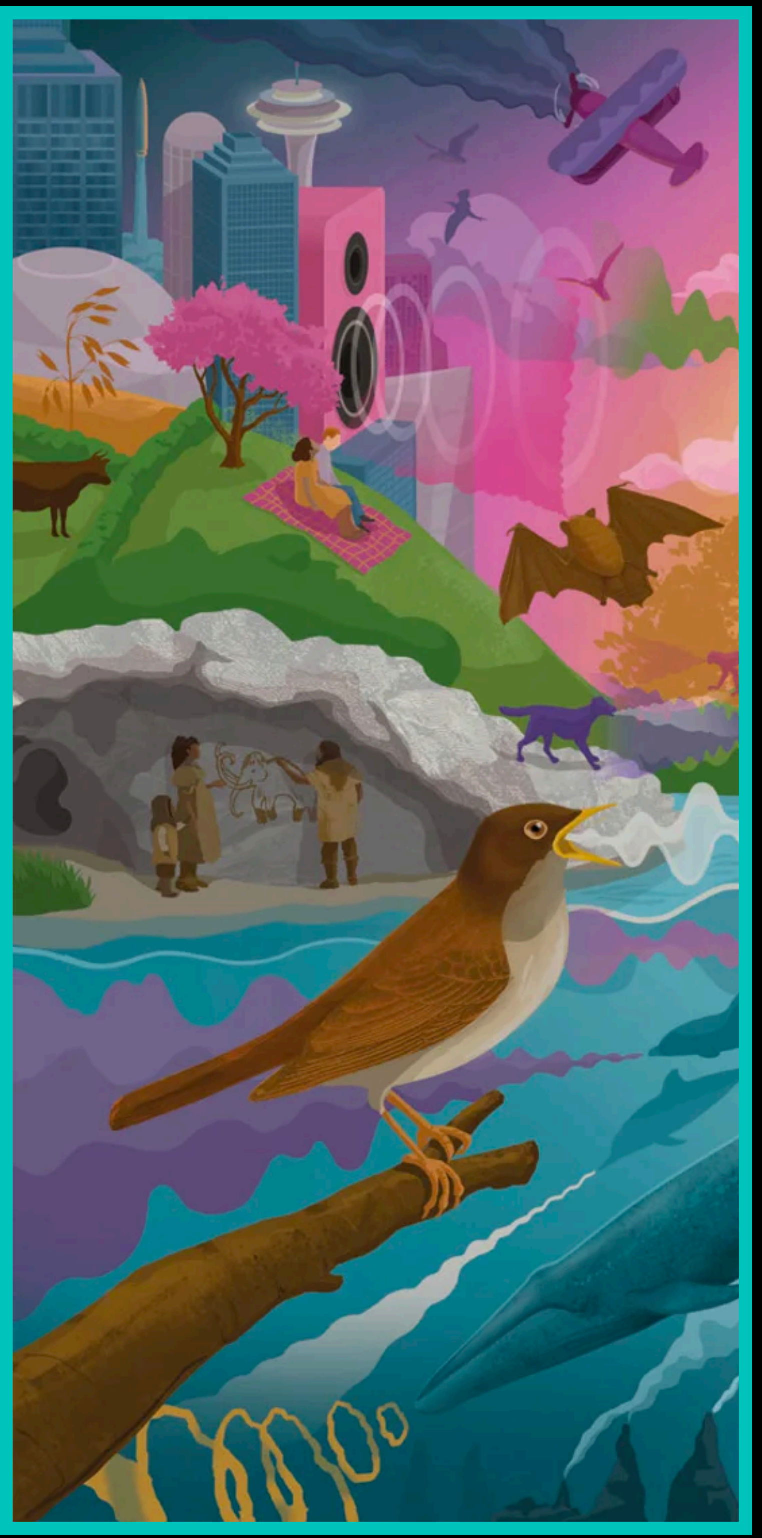


Comparing Self-Supervised Learning Models Pre-Trained on Human Speech and Animal Vocalizations for Bioacoustics Processing

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

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

IEEE ICASSP 2025

April 2025

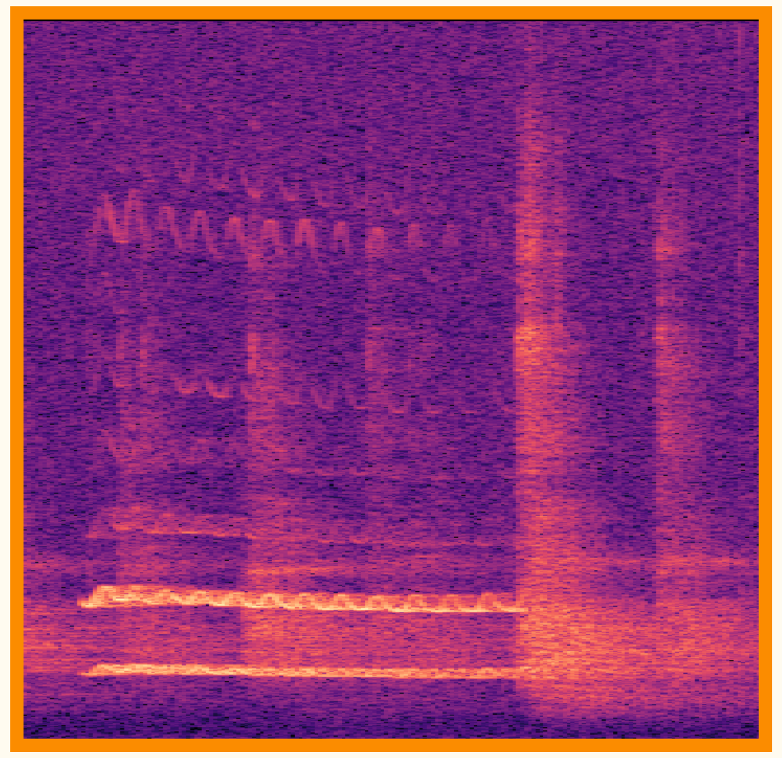
Introduction

Computational Bioacoustics

Computational Bioacoustics

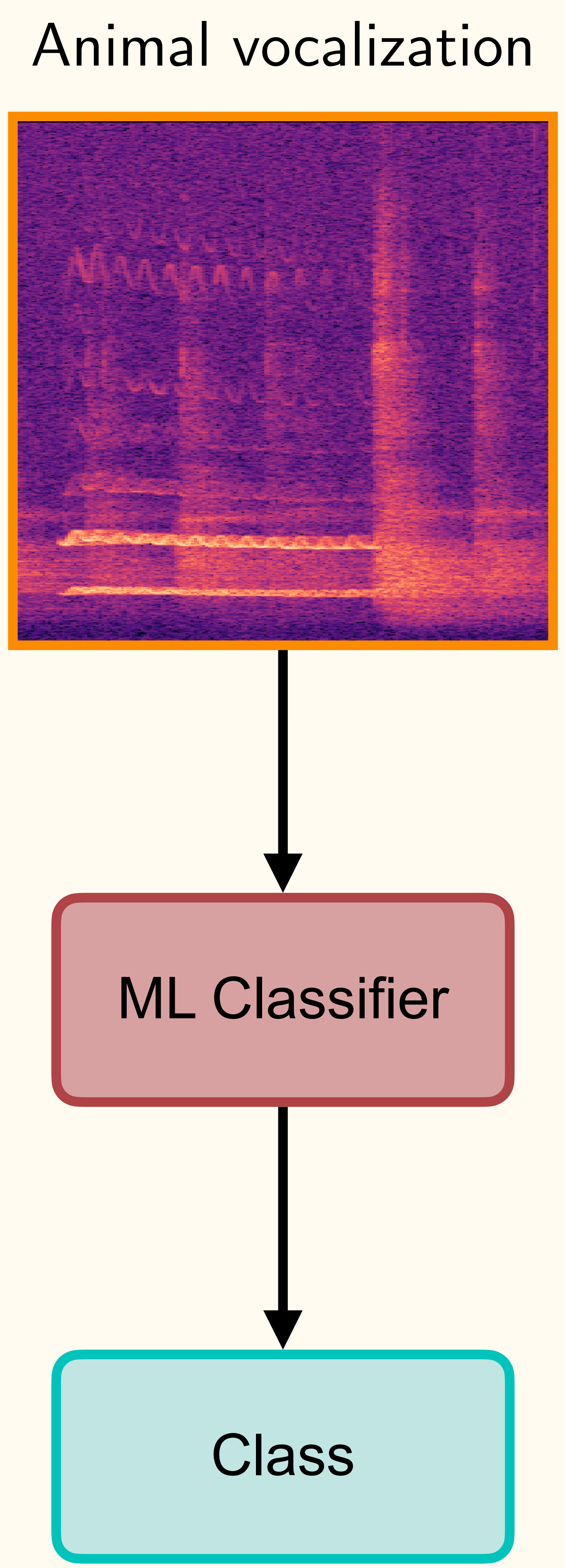
- **What**: study of animal sounds and communication.
  - Plays a role in ecological and evolutionary research, providing insights into animal communication, biodiversity, and the origins of language.

Animal vocalization



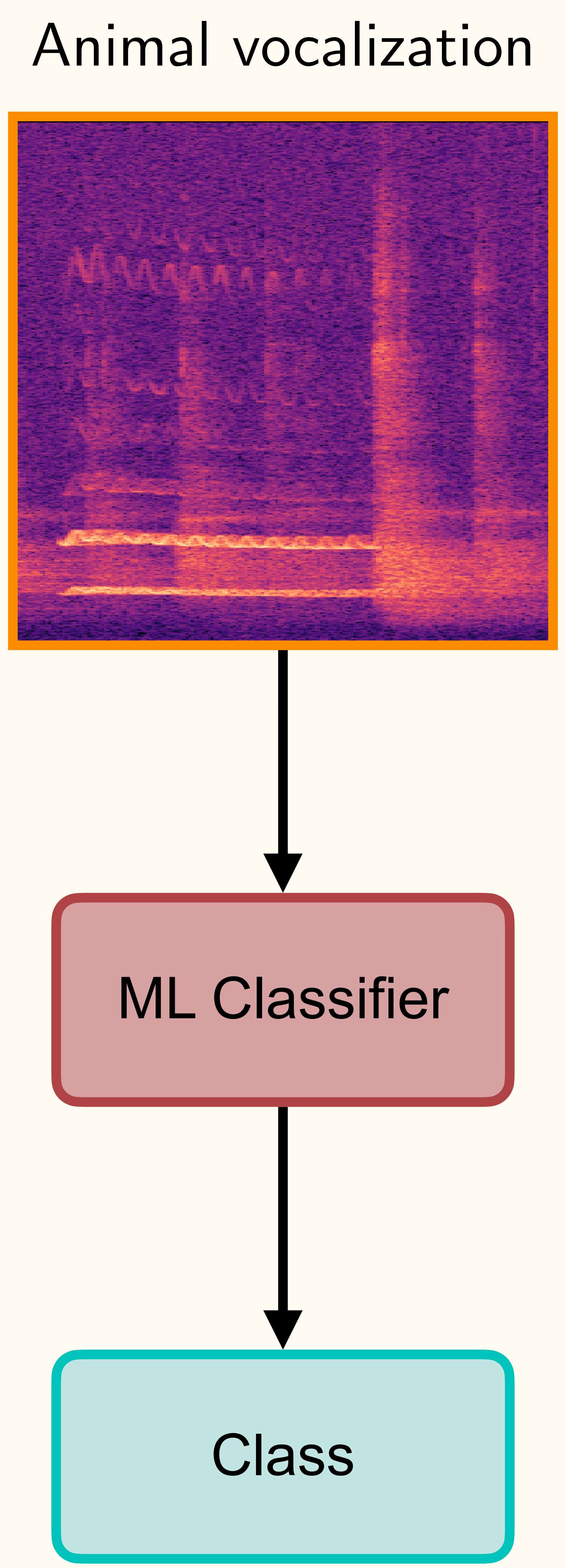
Computational Bioacoustics

- **What**: study of animal sounds and communication.
  - Plays a role in ecological and evolutionary research, providing insights into animal communication, biodiversity, and the origins of language.
- **Tasks**: call detection and classification, caller identification, and species recognition.



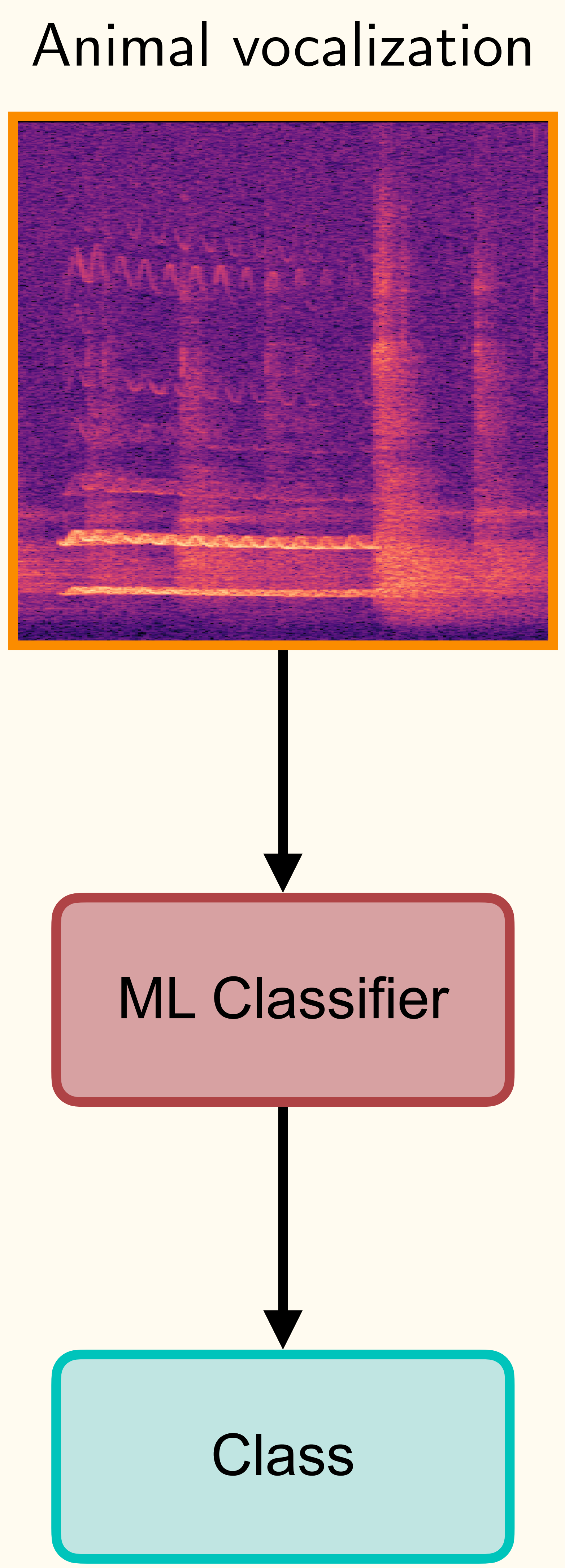
Computational Bioacoustics

- **What**: study of animal sounds and communication.
  - Plays a role in ecological and evolutionary research, providing insights into animal communication, biodiversity, and the origins of language.
- **Tasks**: call detection and classification, caller identification, and species recognition.
- **Challenges**: scarce, noisy, difficult to collect and annotate.



Computational Bioacoustics

- **What**: study of animal sounds and communication.
  - Plays a role in ecological and evolutionary research, providing insights into animal communication, biodiversity, and the origins of language.
- **Tasks**: call detection and classification, caller identification, and species recognition.
- **Challenges**: scarce, noisy, difficult to collect and annotate.
- **Progress**: In recent years advances in ML has addressed challenges. Notably ...



Transferability of Self-Supervised Learning Representations

Transferability of Self-Supervised Learning Representations

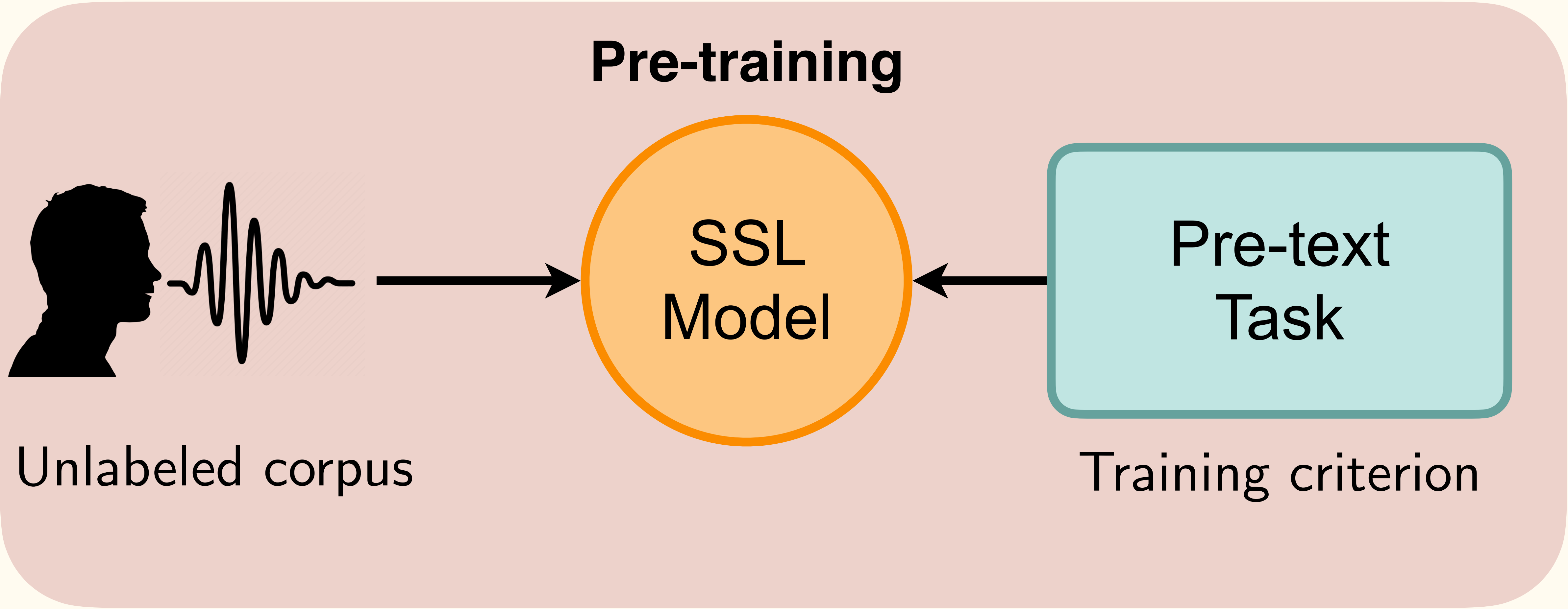
- **Pre-trained foundation models** shown impressive transferability to bioacoustics signals, significantly advancing the field.

Transferability of Self-Supervised Learning Representations

- **Pre-trained foundation models** shown impressive transferability to bioacoustics signals, significantly advancing the field.
- Notably, SSL models pre-trained on **human speech** (WavLM, HuBERT, wav2vec2, etc.) have shown remarkable success[1-5] in bioacoustics classification tasks.

[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., O*n the Utility of Speech and Audio Foundation Models for Marmoset Call Analysis* (2024). Proc. of Interspeech.
[3] Sarkar et al. *On Feature Representations for Marmoset Vocal Communication Analysis* (2025). Bioacoustics Journal.
[4] Cauzinille et al. *Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures* (2024). Proc. of Interspeech.
[5] Abzaliev et al. *Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification* (2024). Proc. of LREC-COLING.

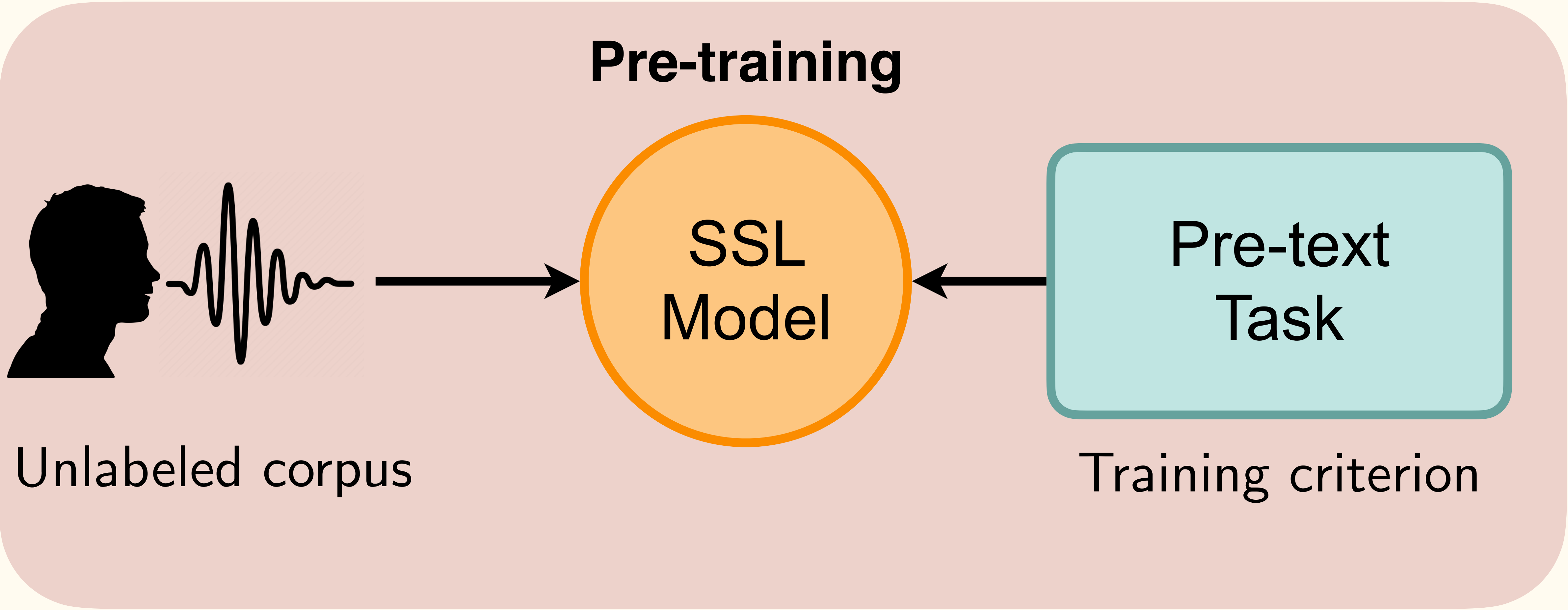
Transferability of Self-Supervised Learning Representations



- These models leverage large volumes of unlabeled data, prevalent in bioacoustics, by creating surrogate labels based on the intrinsic structure of the audio data, and then solving pre-text tasks designed to learn salient representations.

[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., O*n the Utility of Speech and Audio Foundation Models for Marmoset Call Analysis* (2024). Proc. of Interspeech.
[3] Sarkar et al. *On Feature Representations for Marmoset Vocal Communication Analysis* (2025). Bioacoustics Journal.
[4] Cauzinille et al. *Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures* (2024). Proc. of Interspeech.
[5] Abzaliev et al. *Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification* (2024). Proc. of LREC-COLING.

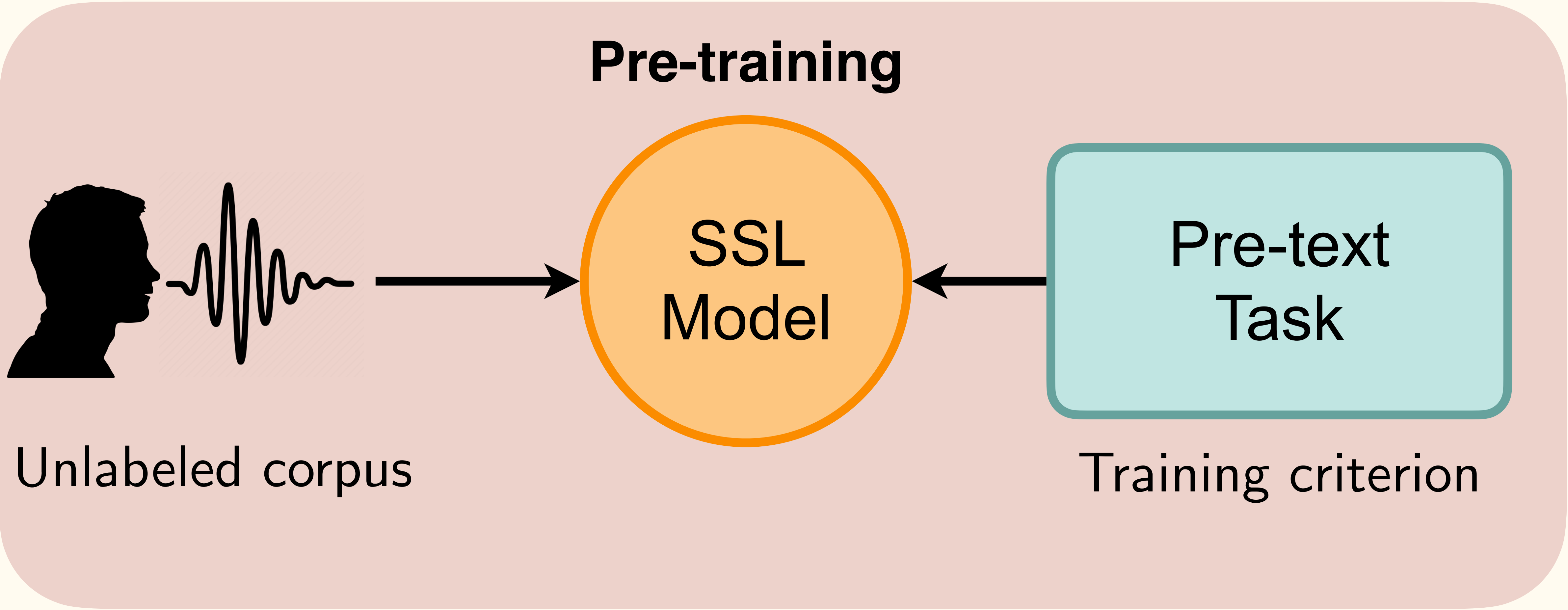
Transferability of Self-Supervised Learning Representations



- Given the domain-agnostic nature of the SSL pre-training tasks, SSL models have been effective in transferring from speech to bioacoustics, without even the need for domain 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., O*n the Utility of Speech and Audio Foundation Models for Marmoset Call Analysis* (2024). Proc. of Interspeech.
[3] Sarkar et al. *On Feature Representations for Marmoset Vocal Communication Analysis* (2025). Bioacoustics Journal.
[4] Cauzinille et al. *Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures* (2024). Proc. of Interspeech.
[5] Abzaliev et al. *Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification* (2024). Proc. of LREC-COLING.

Transferability of Self-Supervised Learning Representations



- Given the domain-agnostic nature of the SSL pre-training tasks, SSL models have been effective in transferring from speech to bioacoustics, without even the need for domain fine-tuning.
- SSL essentially serve as powerful, general-purpose feature extractors for a wide range of downstream tasks.

SSL Pre-Training Domain

Research Question 1

Fine-Tuning on Human Speech

Research Question 2

RQ1: SSL Pre-training Domain

RQ1: SSL Pre-training Domain

- **While** SSL models pre-trained on speech have shown strong transferability to bio tasks, recent works have explored directly PT'ing on bioacoustic data.

RQ1: SSL Pre-training Domain

- **While** SSL models pre-trained on speech have shown strong transferability to bio tasks, recent works have explored directly PT'ing on bioacoustic data.
- **Motivation** behind pre-training on animal data is that these models may better capture species-specific vocal patterns and other properties unique to animal sounds.

RQ1: SSL Pre-training Domain

- **While** SSL models pre-trained on speech have shown strong transferability to bio tasks, recent works have explored directly PT'ing on bioacoustic data.
- **Motivation** behind pre-training on animal data is that these models may better capture species-specific vocal patterns and other properties unique to animal sounds.
- **However**, given that SSL PT'ing is designed to learn general, domain-agnostic features, it's not yet clear whether PT'ing directly on bioacoustics provides any significant benefit over SSLs PT'd on speech.

RQ1: SSL Pre-training Domain

- **While** SSL models pre-trained on speech have shown strong transferability to bio tasks, recent works have explored directly PT'ing on bioacoustic data.
- **Motivation** behind pre-training on animal data is that these models may better capture species-specific vocal patterns and other properties unique to animal sounds.
- **However**, given that SSL PT'ing is designed to learn general, domain-agnostic features, it's not yet clear whether PT'ing directly on bioacoustics provides any significant benefit over SSLs PT'd on speech.
- **Therefore**, we systematically compare SSL models PT'd on human speech against those on animal calls, and evaluate their performance bioacoustic processing across a variety of datasets & tasks.

RQ2. Fine-Tuning on Human Speech

RQ2. Fine-Tuning on Human Speech

- **SSL representations** have shown strong performance on bio tasks without requiring FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls & identities.

RQ2. Fine-Tuning on Human Speech

- **SSL representations** have shown strong performance on bio tasks without requiring FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls & identities.
- **However**, FT'ing in a supervised framework often forces the model to learn novel & specialized patterns.
  - Such as phonetic distinctions and temporal structures → typically leading to performance gains.

RQ2. Fine-Tuning on Human Speech

- **SSL representations** have shown strong performance on bio tasks without requiring FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls & identities.
- **However**, FT'ing in a supervised framework often forces the model to learn novel & specialized patterns.
  - Such as phonetic distinctions and temporal structures → typically leading to performance gains.
- As human speech and animal calls both encode structured vocal and linguistic information for communication, SSL models **fine-tuned** on speech recognition (ASR) may **provide** an **additional inductive bias**, enhancing the model's ability to recognize complex features in bio data.

RQ2. Fine-Tuning on Human Speech

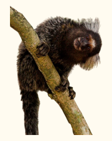
- **SSL representations** have shown strong performance on bio tasks without requiring FT'ing.
  - Indicating their extracted latents can capture acoustically rich information.
  - Capable of distinguishing animal calls & identities.
- **However**, FT'ing in a supervised framework often forces the model to learn novel & specialized patterns.
  - Such as phonetic distinctions and temporal structures → typically leading to performance gains.
- As human speech and animal calls both encode structured vocal and linguistic information for communication, SSL models **fine-tuned** on speech recognition (ASR) may **provide** an **additional inductive bias**, enhancing the model's ability to recognize complex features in bio data.
- **Therefore**, we explore whether fine-tuning PT'd SSLs on human speech tasks, such as ASR, can improve models' capability to process animal calls by capturing the subtle spectro-temporal characteristics, which may otherwise remain under-represented in general SSL pre-training.

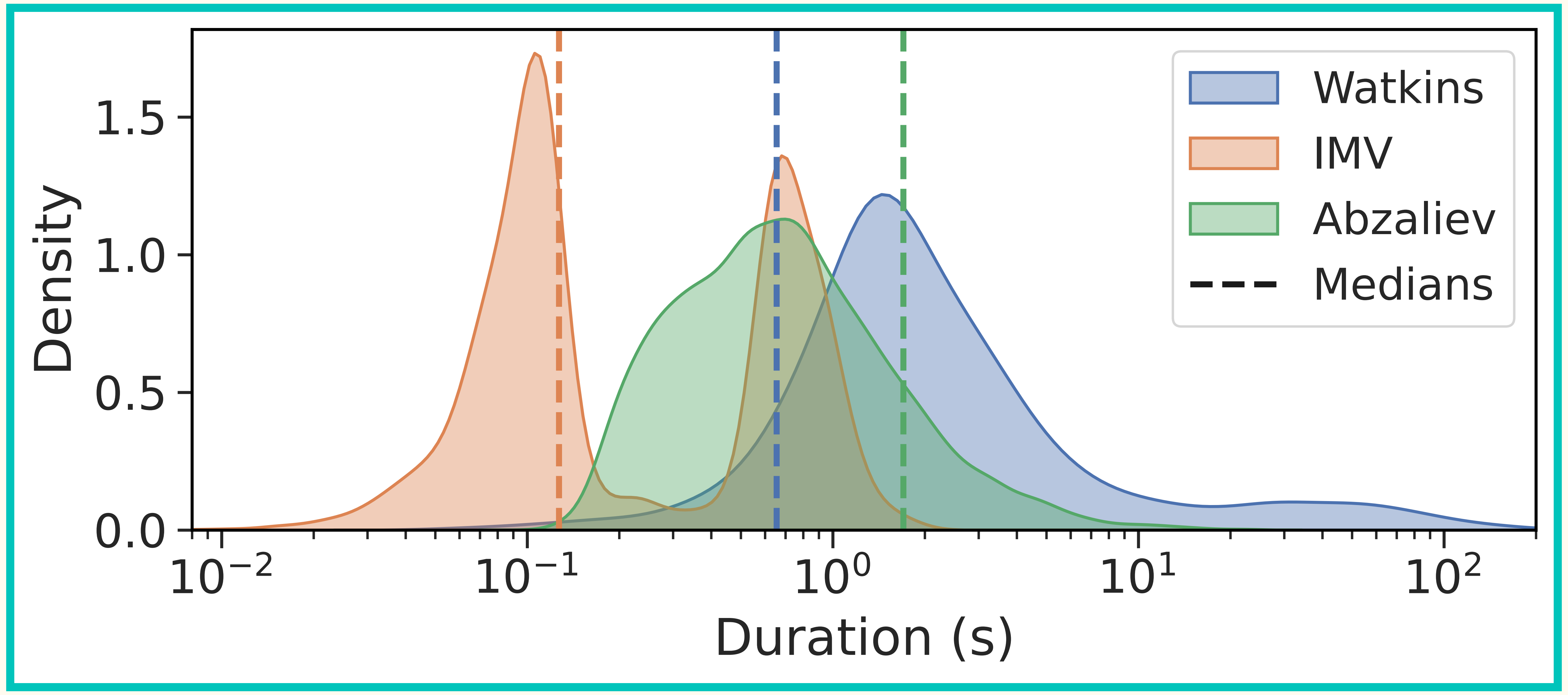
Contents

Experimental Setup

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

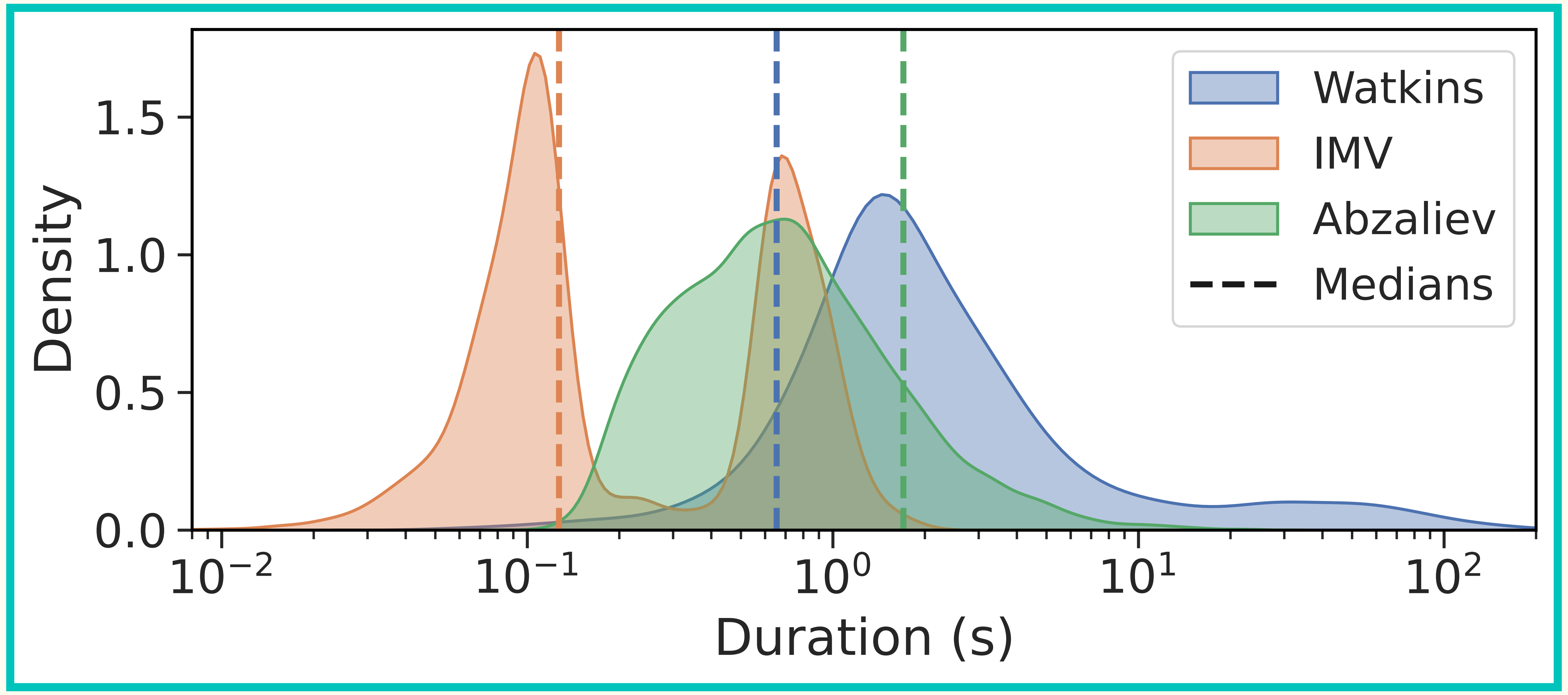
Dataset	**# Samples**	$\boldsymbol{L}$	**SR**	$\boldsymbol{n_c}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313



Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

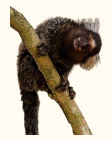
Dataset	**# Samples**	L	**SR**	n_c	μ	σ
Watkins	1,697	295	–	32	1701	71245
IMV	72,920	464	44.1	11	127	375
Abzaliev	8,034	137	48	14	655	1313

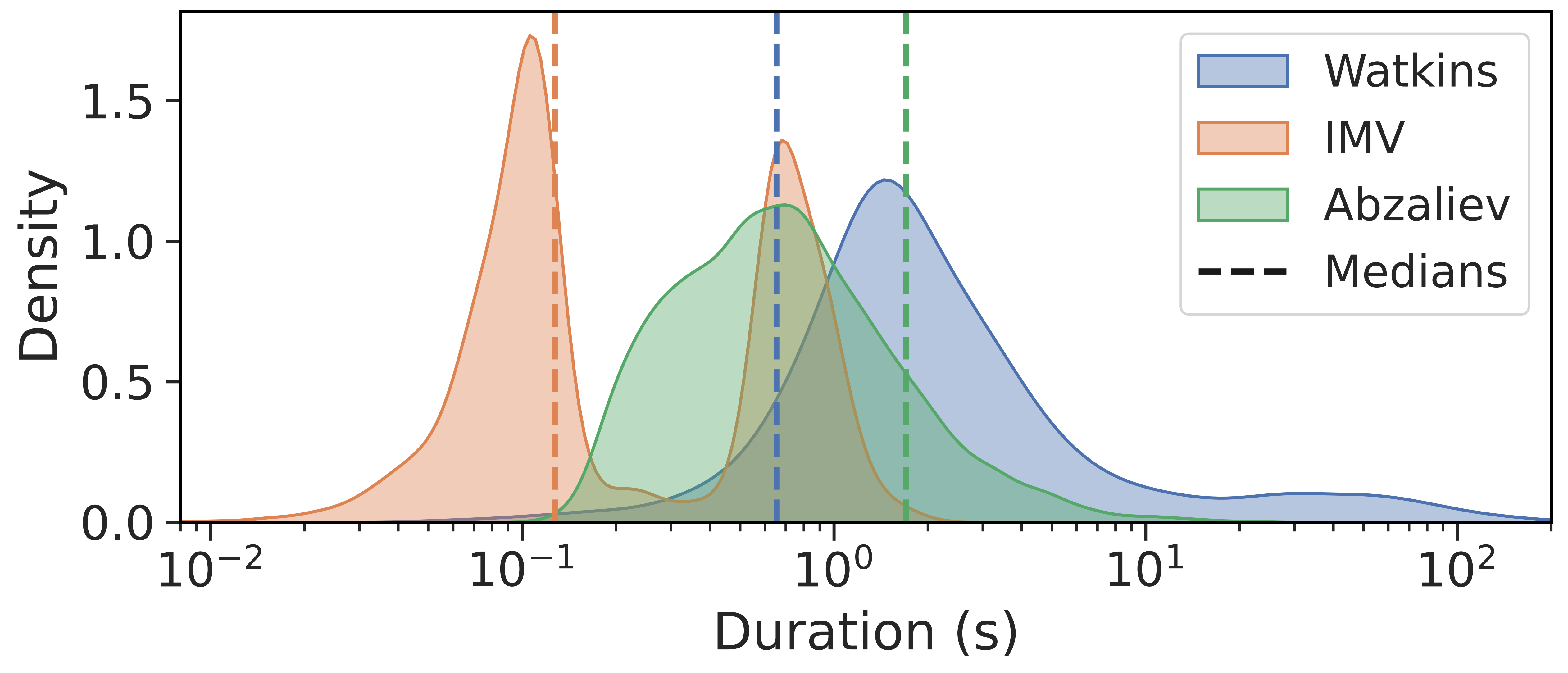


- **Watkins**:

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

Dataset	**# Samples**	L	**SR**	n_c	μ	σ
Watkins 	$1,697$	295	–	32	1701	71245
IMV 	$72,920$	464	44.1	11	127	375
Abzaliev 	$8,034$	137	48	14	655	1313



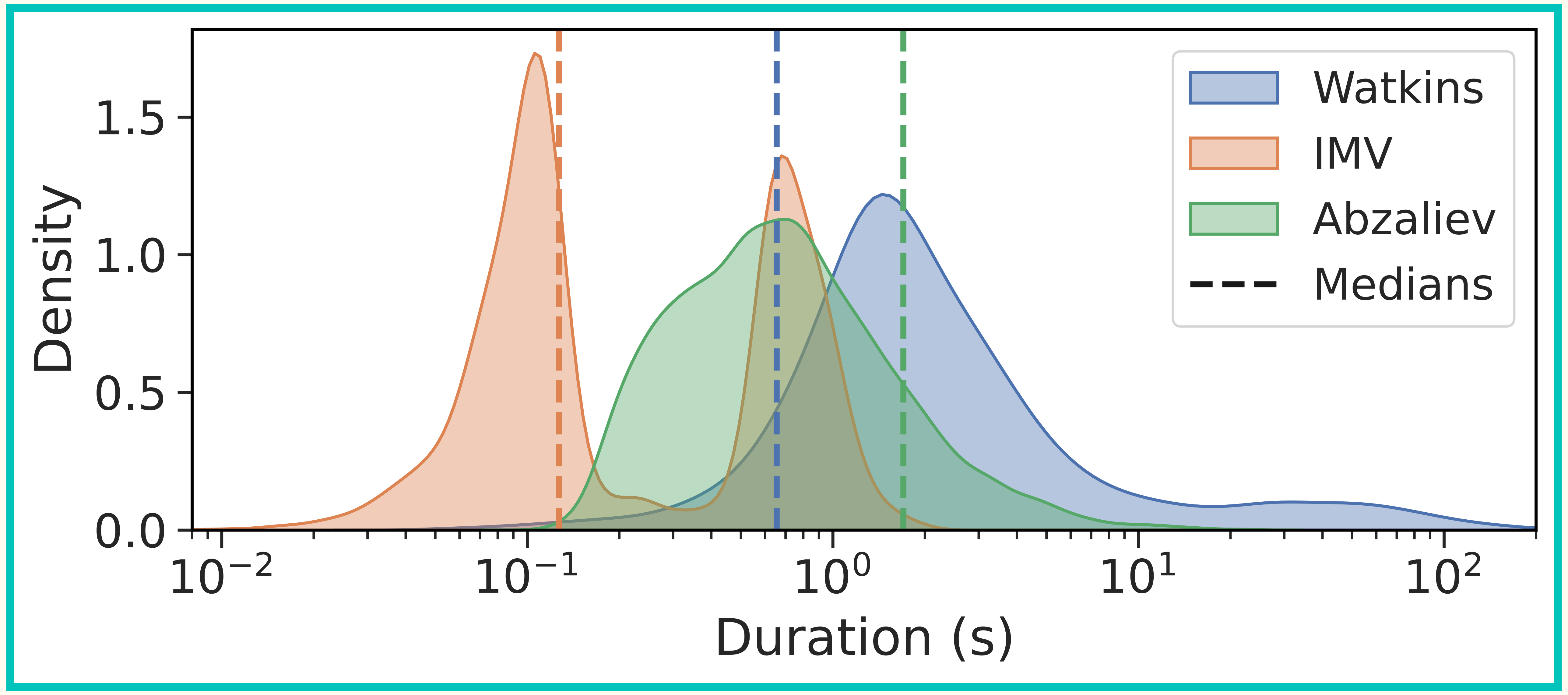
- **Watkins**:

- Marine mammals recordings.

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

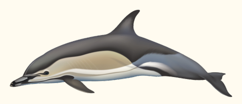
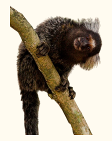
Dataset	**# Samples**	$\boldsymbol{L}$	**SR**	$\boldsymbol{n_c}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
Watkins	1,697	295	–	32	1701	71245
IMV	72,920	464	44.1	11	127	375
Abzaliev	8,034	137	48	14	655	1313

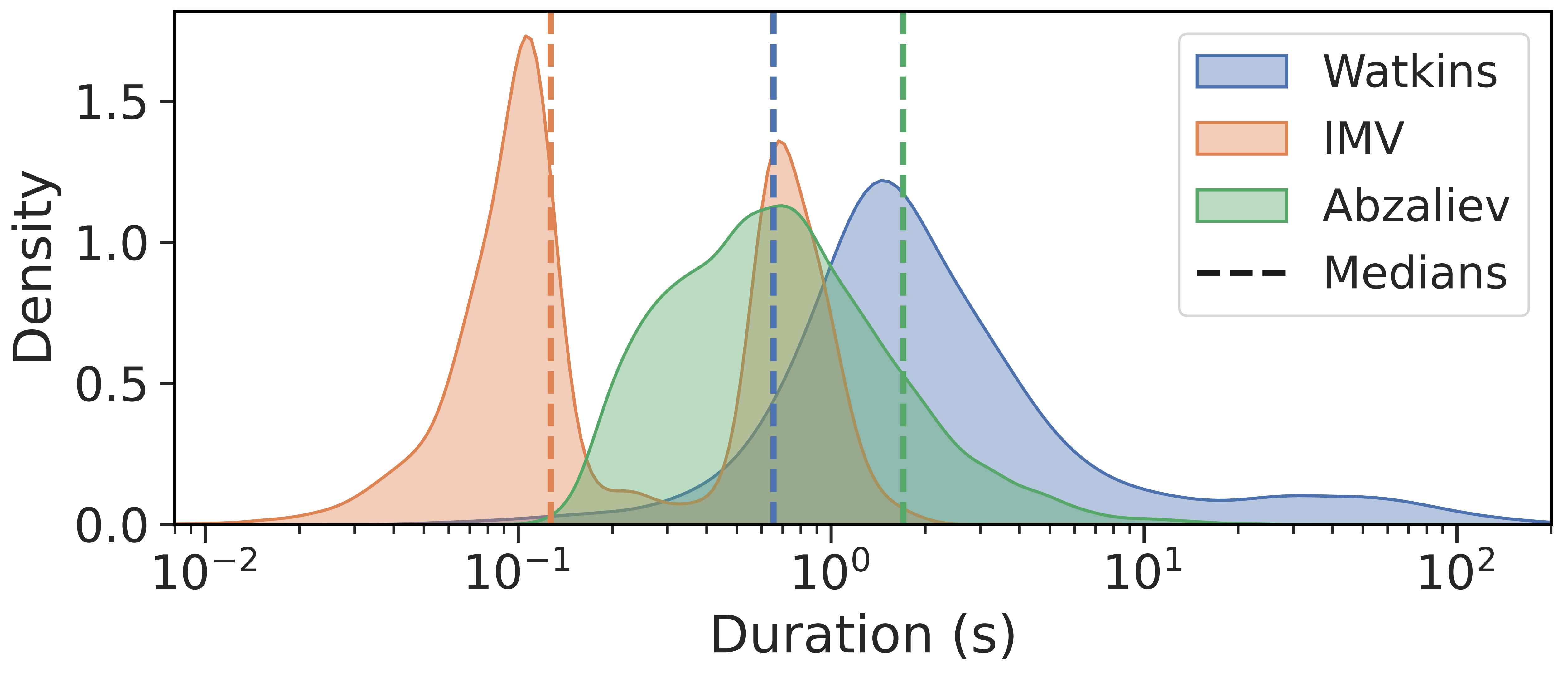


- **Watkins**:
- Marine mammals recordings.
- Multi-species vocalizations, rich acoustic variety, high variance in length.

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

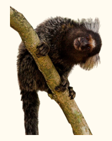
Dataset	# Samples	L	SR	n_c	μ	σ
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313

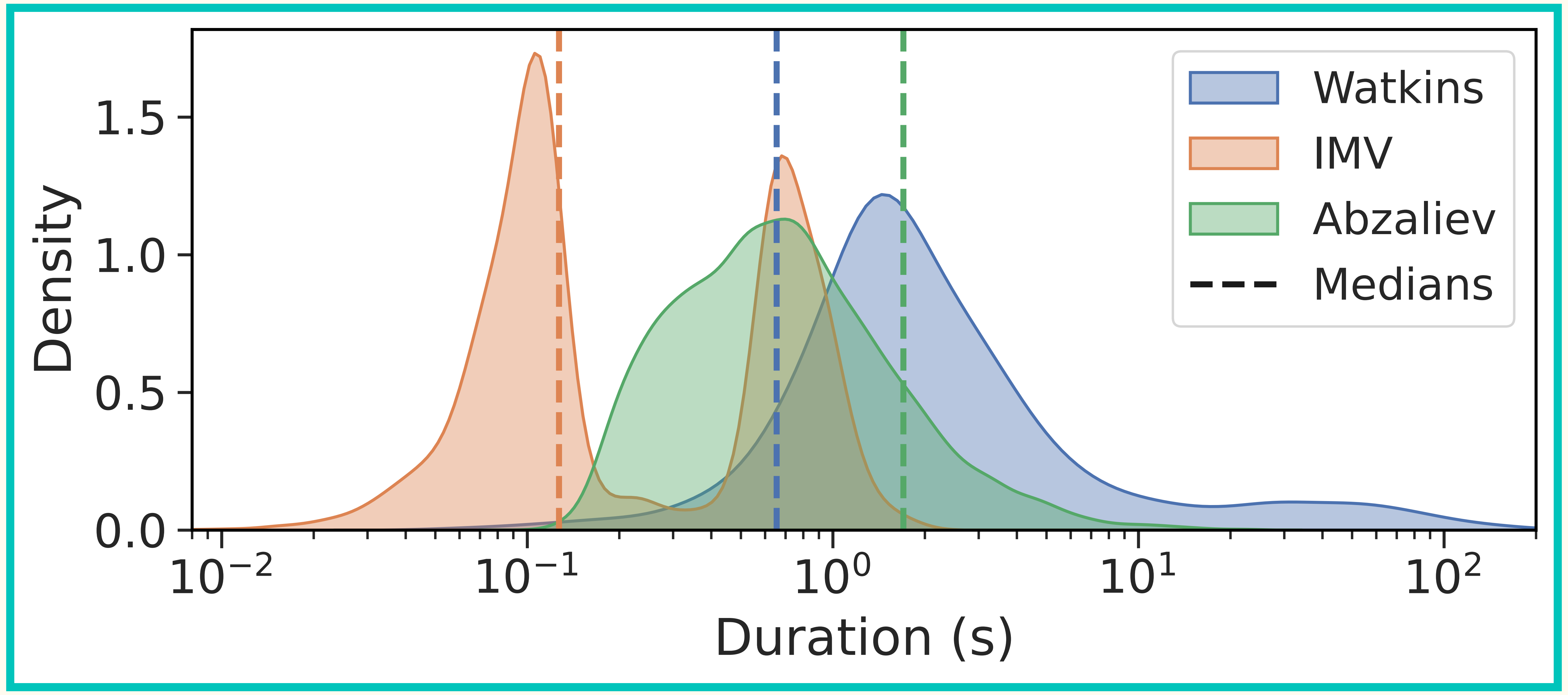


- **Watkins**:
  - Marine mammals recordings.
  - Multi-species vocalizations, rich acoustic variety, high variance in length.
- **InfantMarmosetsVox (IMV)**:

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

Dataset	**# Samples**	L	**SR**	n_c	μ	σ
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313

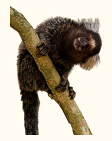


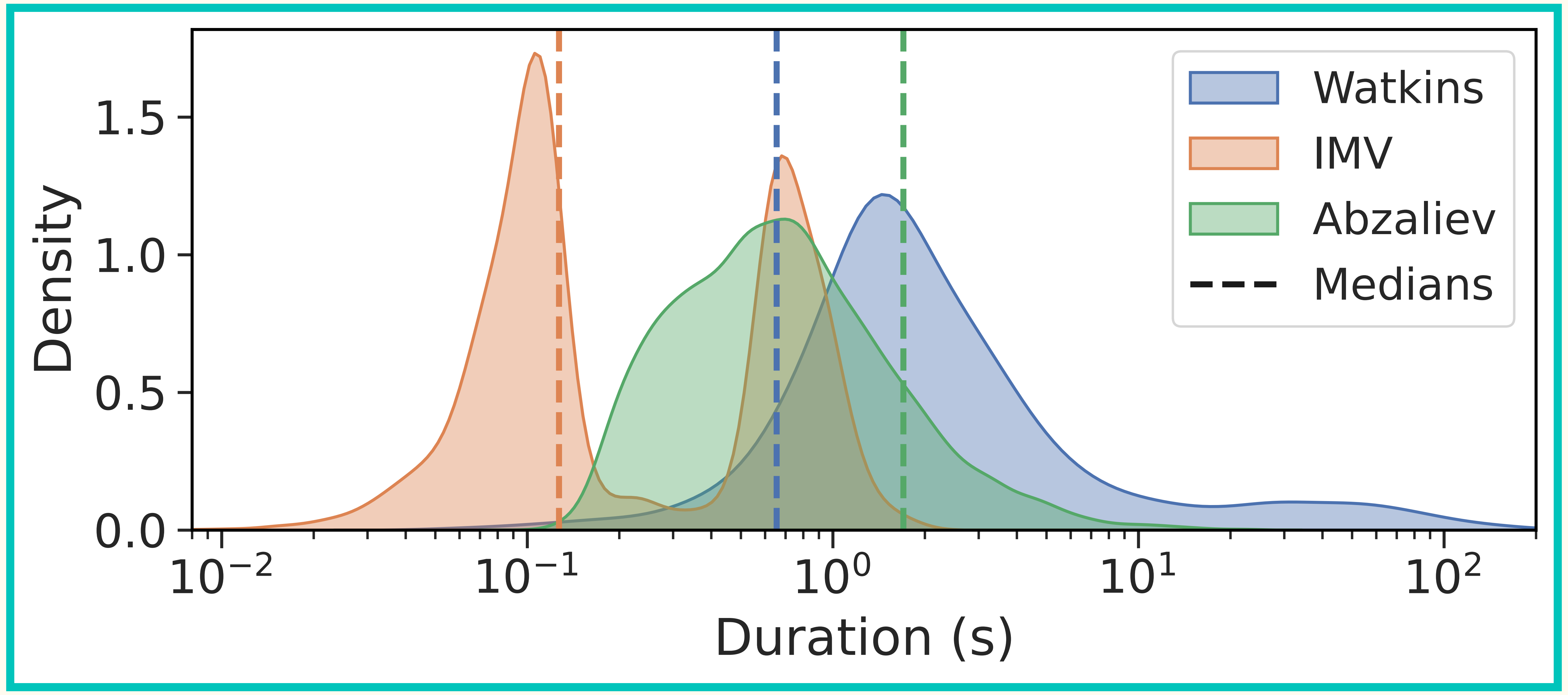
- **Watkins**:
  - Marine mammals recordings.
  - Multi-species vocalizations, rich acoustic variety, high variance in length.

- **InfantMarmosetsVox (IMV)**:
  - Complex social system.

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

Dataset	**# Samples**	L	**SR**	n_c	μ	σ
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313

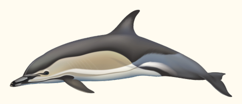
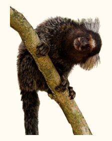


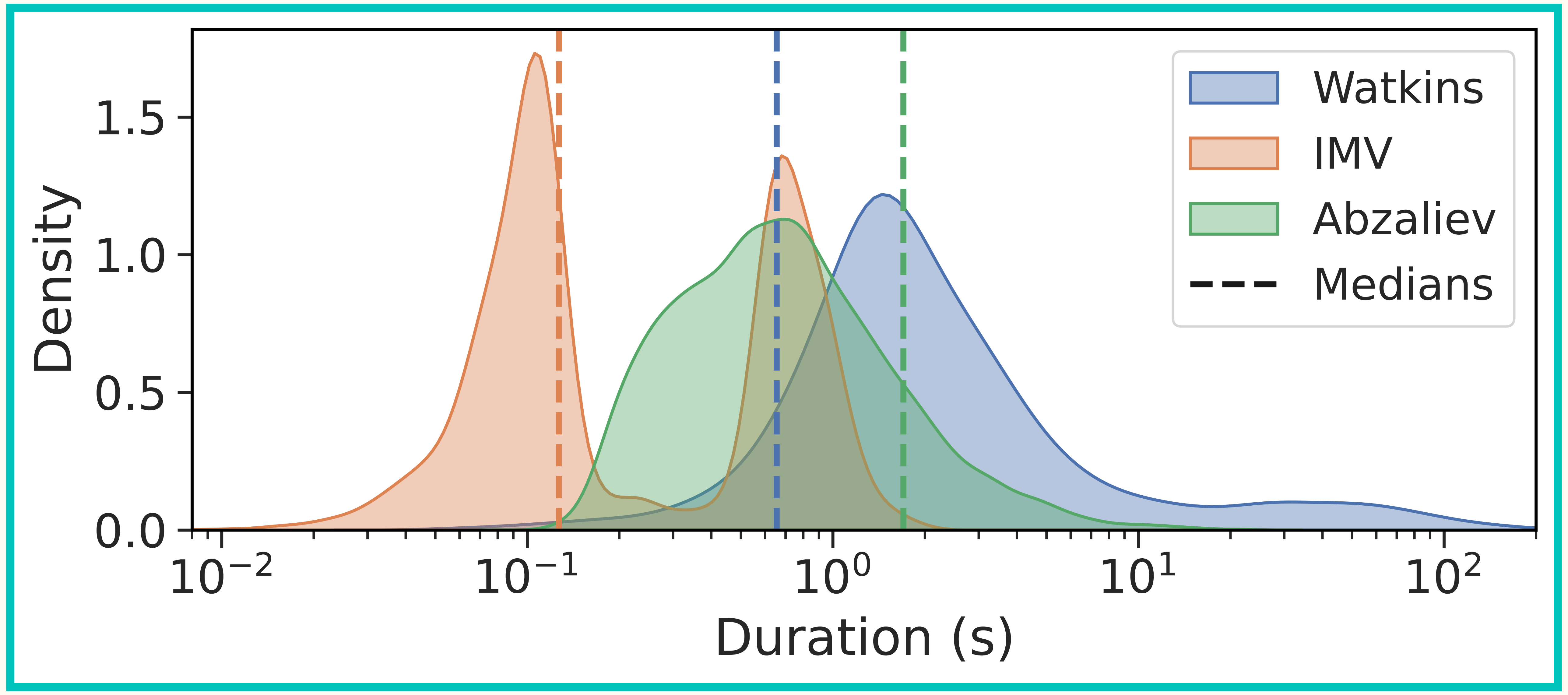
- **Watkins**:
  - Marine mammals recordings.
  - Multi-species vocalizations, rich acoustic variety, high variance in length.

- **InfantMarmosetsVox (IMV)**:
  - Complex social system.
  - Encode critical information in calls.

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

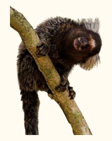
Dataset	**# Samples**	L	**SR**	n_c	μ	σ
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313

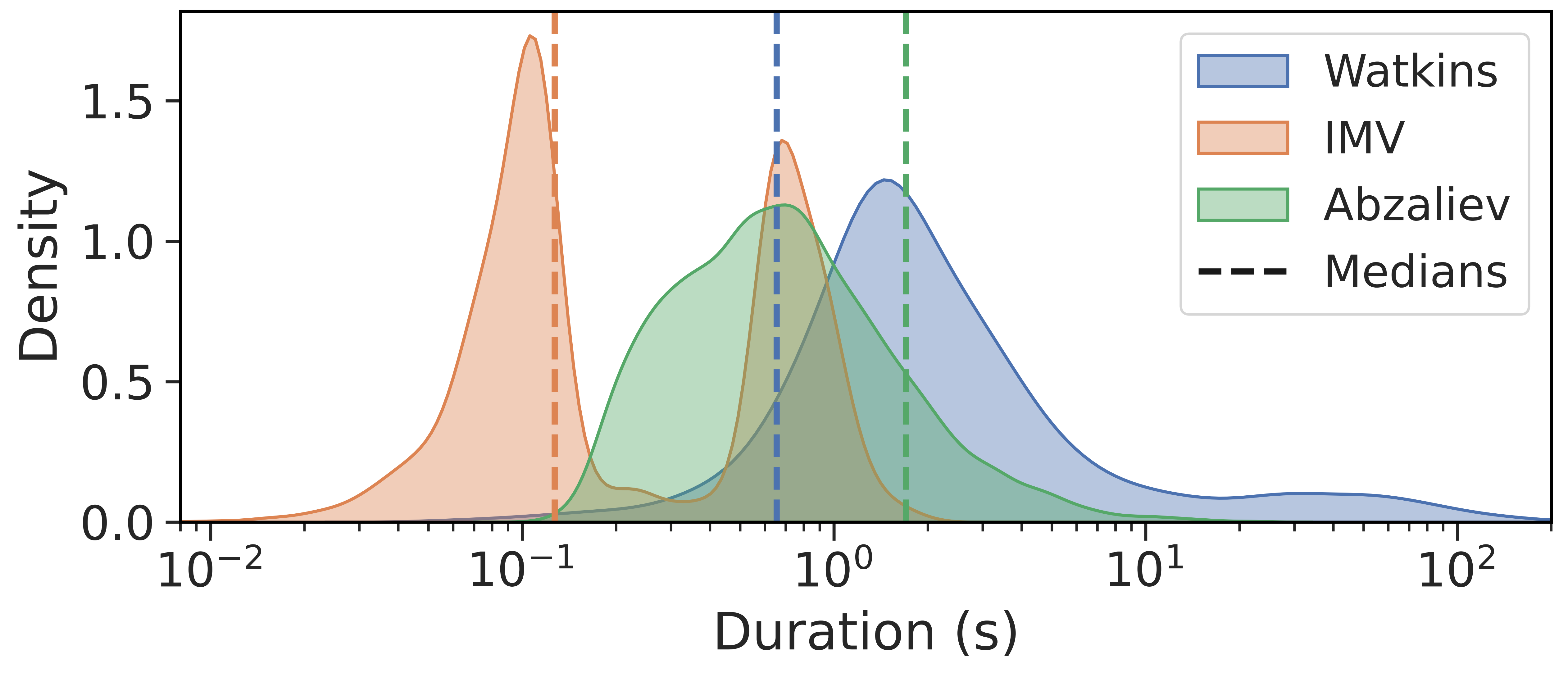


- **Watkins**:
  - Marine mammals recordings.
  - Multi-species vocalizations, rich acoustic variety, high variance in length.
- **InfantMarmosetsVox (IMV)**:
  - Complex social system.
  - Encode critical information in calls.
- **Abzaliev**:

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

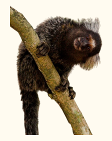
Dataset	**# Samples**	$\boldsymbol{L}$	**SR**	$\boldsymbol{n_c}$	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313

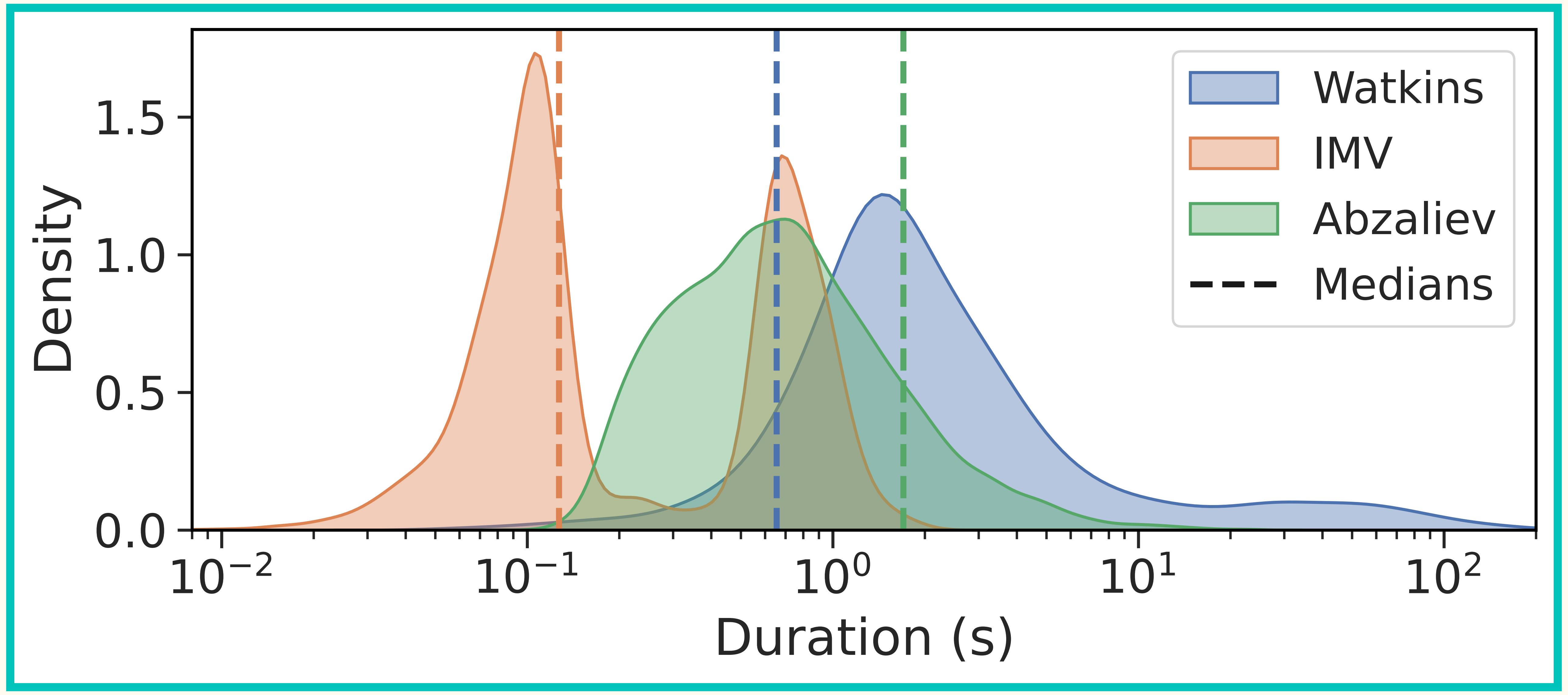


- **Watkins**:
  - Marine mammals recordings.
  - Multi-species vocalizations, rich acoustic variety, high variance in length.
- **InfantMarmosetsVox (IMV)**:
  - Complex social system.
  - Encode critical information in calls.
- **Abzaliev**:
  - Novel dog dataset.

Datasets

L denotes the total length [minutes], n_c the number of classes, SR the sampling rate [kHz], μ the median length [ms].

Dataset	# Samples	L	SR	n_c	μ	σ
Watkins 	1,697	295	–	32	1701	71245
IMV 	72,920	464	44.1	11	127	375
Abzaliev 	8,034	137	48	14	655	1313



- **Watkins**:
  - Marine mammals recordings.
  - Multi-species vocalizations, rich acoustic variety, high variance in length.
- **InfantMarmosetsVox (IMV)**:
  - Complex social system.
  - Encode critical information in calls.
- **Abzaliev**:
  - Novel dog dataset.
  - Various types of barks.

Models and Feature Representations

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	Corpus	P	D	TL	Type
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

\# Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	**Corpus**	P	D	**TL**	**Type**
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	Corpus	P	D	TL	Type
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1] All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT'd on human speech.

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	**Corpus**	$\boldsymbol{P}$	$\boldsymbol{D}$	**TL**	**Type**
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT'd on human speech.
- SSL PT+FT'd on human speech[1].

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	**Corpus**	$\boldsymbol{P}$	$\boldsymbol{D}$	**TL**	**Type**
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT'd on human speech.
- SSL PT+FT'd on human speech[1].
- Fusion.

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	**Corpus**	P	D	**TL**	**Type**
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT'd on human speech.
- SSL PT+FT'd on human speech[1].
- Fusion.

Classifier:

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	Corpus	P	D	TL	Type
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT’d on human speech.
- SSL PT+FT’d on human speech[1].
- Fusion.

Classifier:

- MLP: 3x [Linear, LN, ReLU] + Linear.

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	Corpus	P	D	TL	Type
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT'd on human speech.
- SSL PT+FT'd on human speech[1].
- Fusion.

Classifier:

- MLP: 3x [Linear, LN, ReLU] + Linear.
- Training: 30 epochs, cross-entropy.

Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	**Corpus**	$\boldsymbol{P}$	$\boldsymbol{D}$	**TL**	**Type**
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Models and Feature Representations

4 neural representations:

- SSL PT'd on animal vocalizations.
- SSL PT'd on human speech.
- SSL PT+FT'd on human speech[1].
- Fusion.

Classifier:

- MLP: 3x [Linear, LN, ReLU] + Linear.
- Training: 30 epochs, cross-entropy.
- Metric: Unweighted Average Recall.

\# Parameters P [M] and feature dimension D of selected models. LS represents LibriSpeech and AS is AudioSet.

$\mathcal{F}$	**Corpus**	P	D	**TL**	**Type**
AVES-Bio	FSD, AS, Bio	94.68	768	12	PT
HuBERT	LS 960	94.68	768	12	PT
W2V2	LS 960	95.04	768	12	PT
W2V2-100h	LS 960	95.04	768	12	PT+FT
W2V2-960h	LS 960	95.04	768	12	PT+FT
WLM	LS 960	94.38	768	12	PT
WLM-100h	LS 960	94.38	768	12	PT+FT

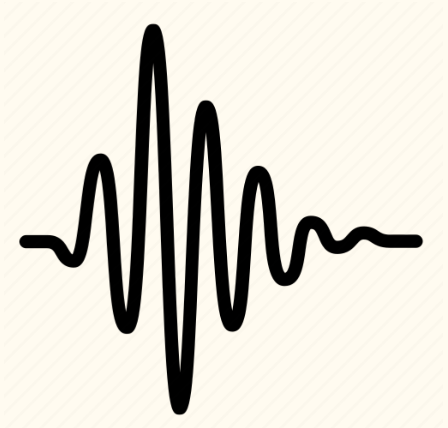
[1]All fine-tuned models are obtained from HuggingFace, namely from the `facebook`, `microsoft`, and `patrickvonplaten` repositories.

Classification Pipeline

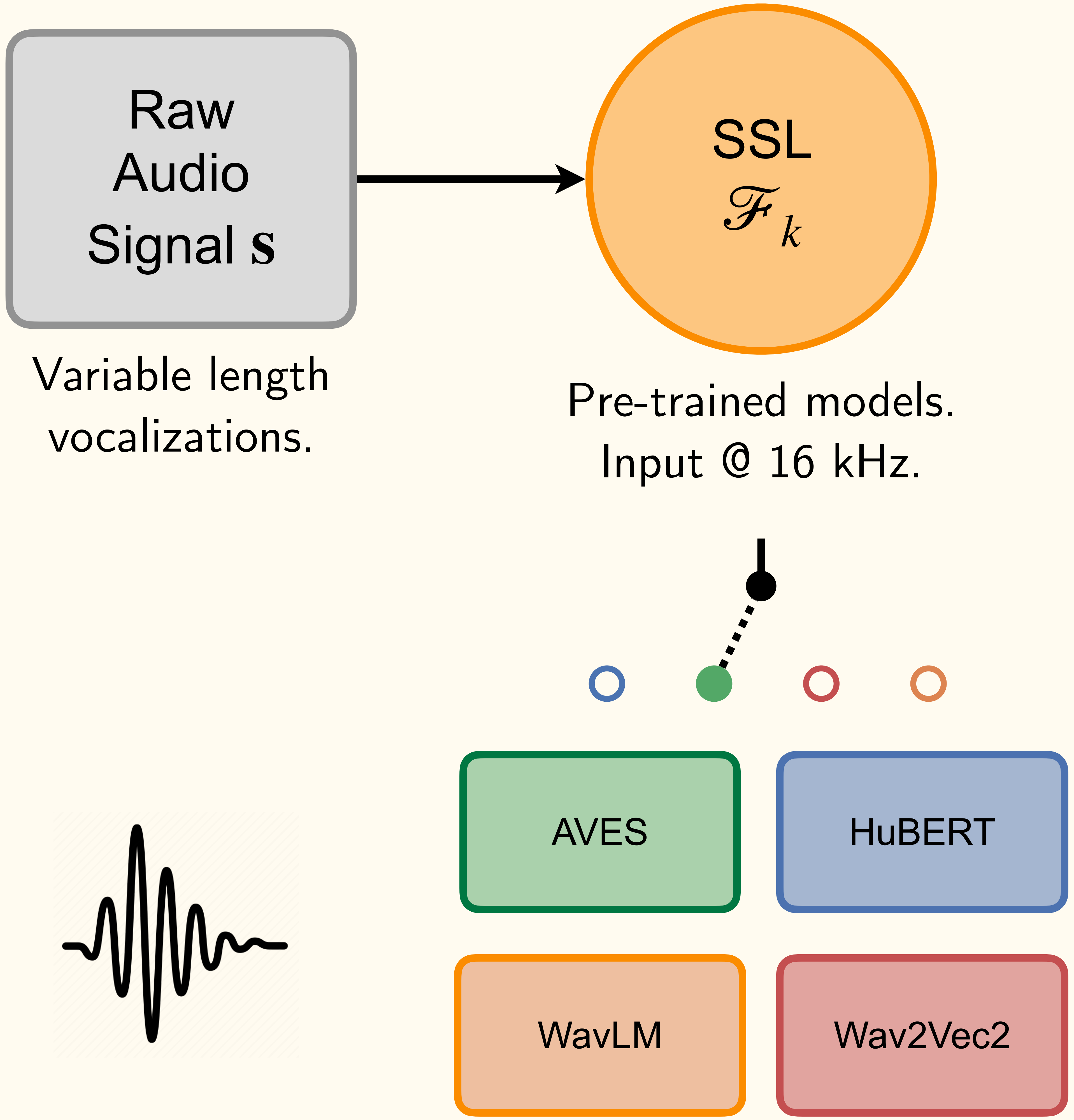
Classification Pipeline

Raw
Audio
Signal $\mathbf{s}$

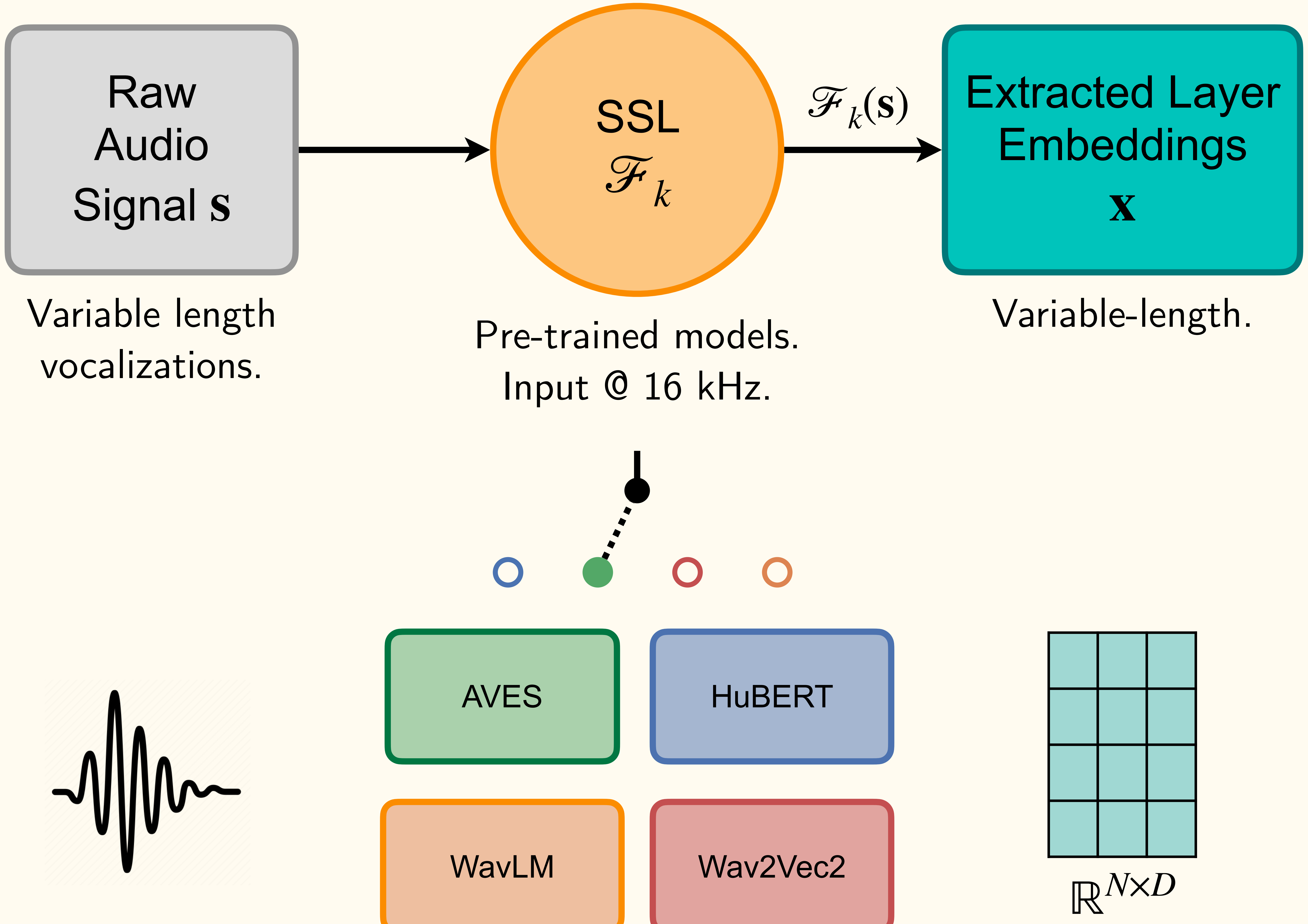
Variable length
vocalizations.



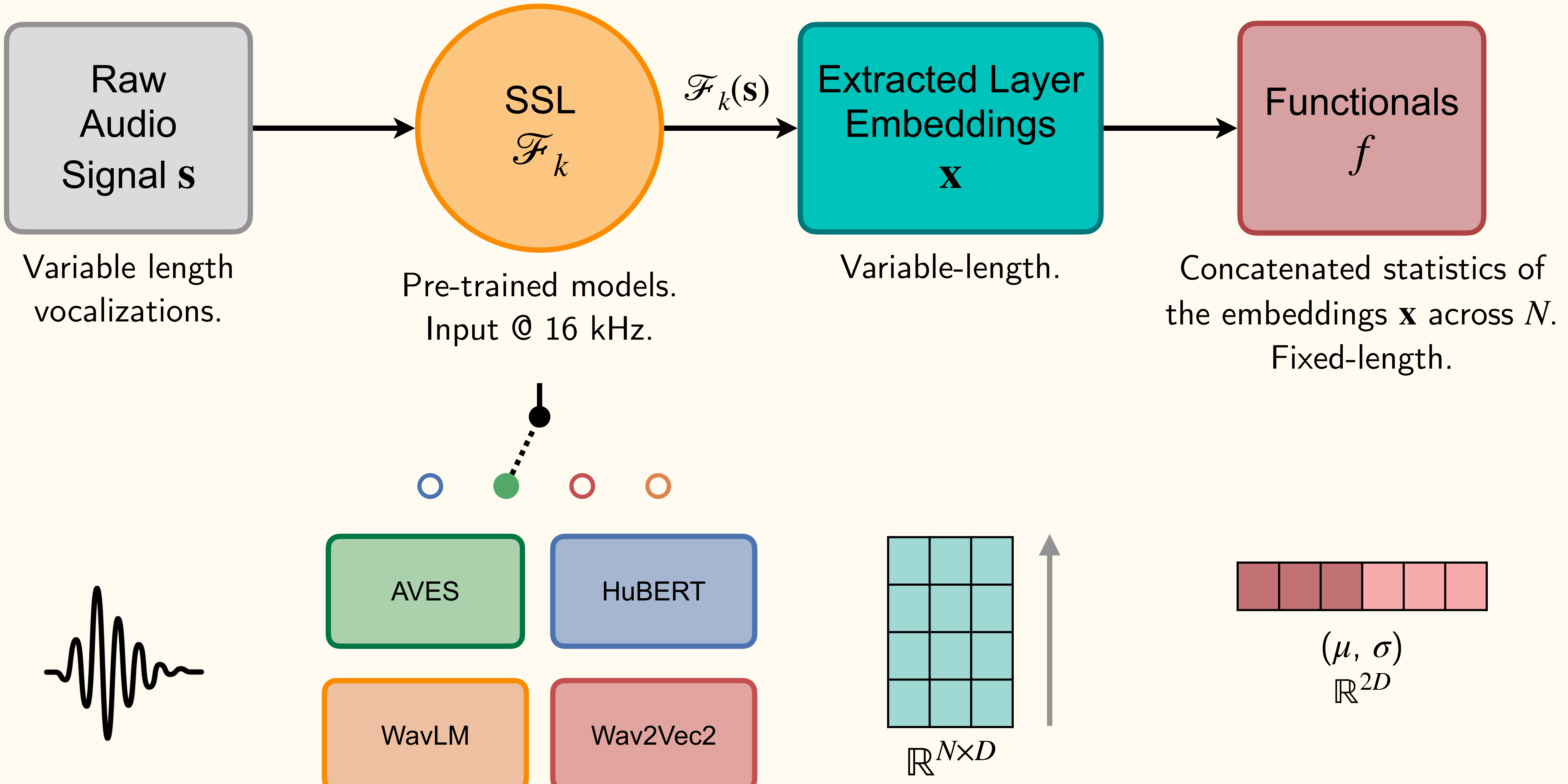
Classification Pipeline



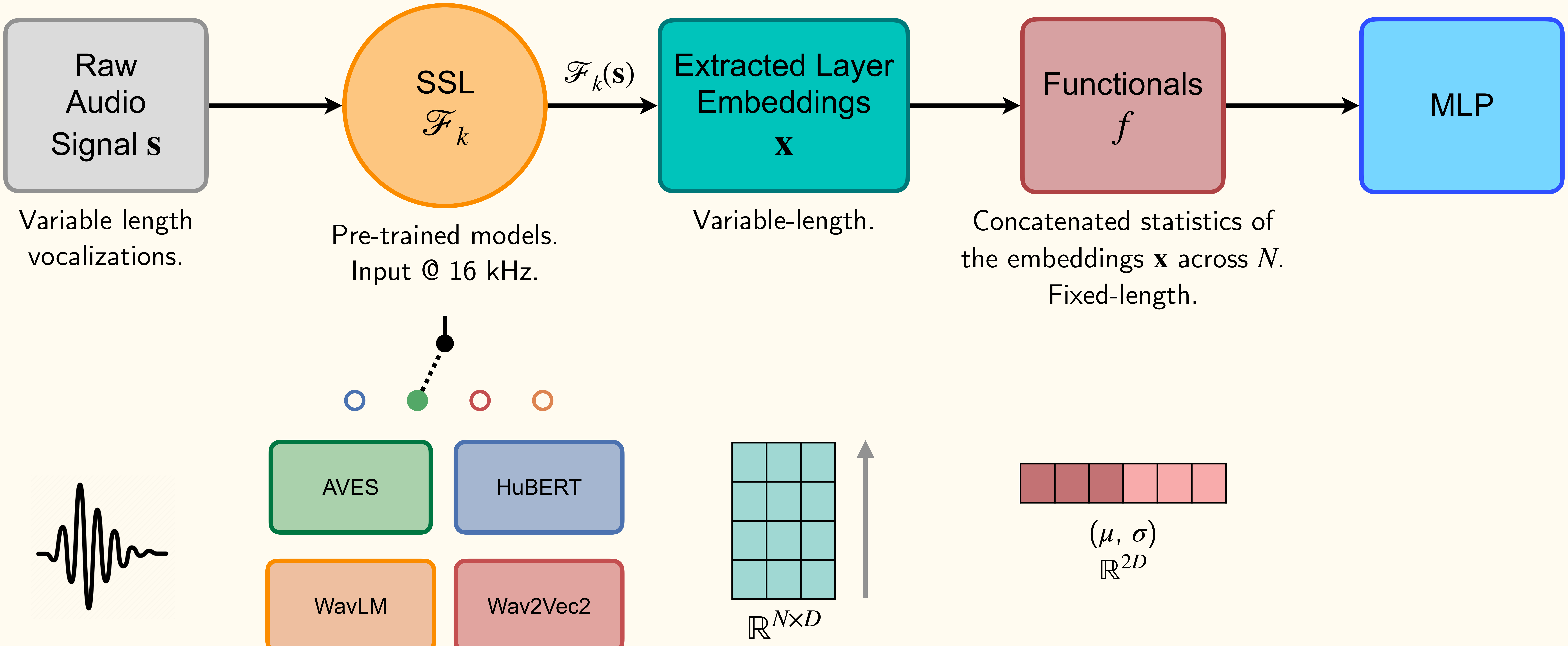
Classification Pipeline



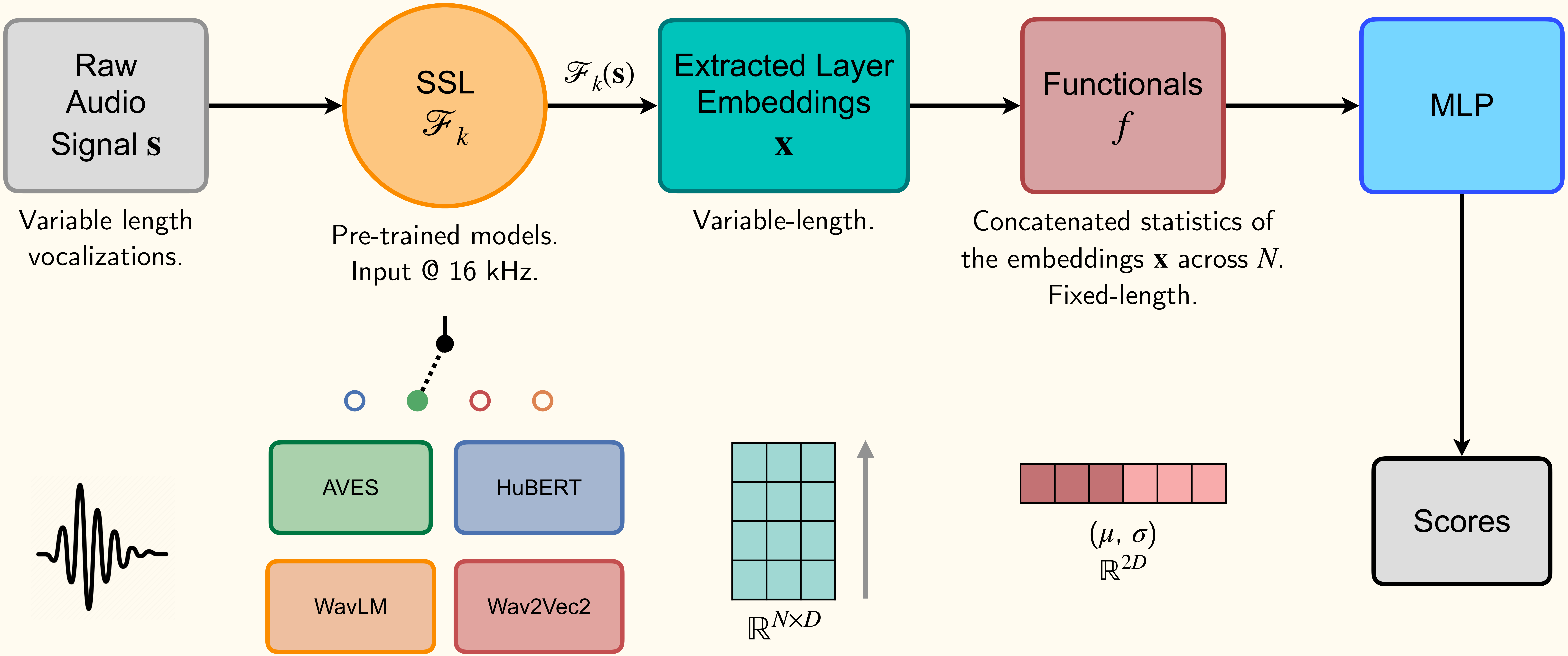
Classification Pipeline



Classification Pipeline

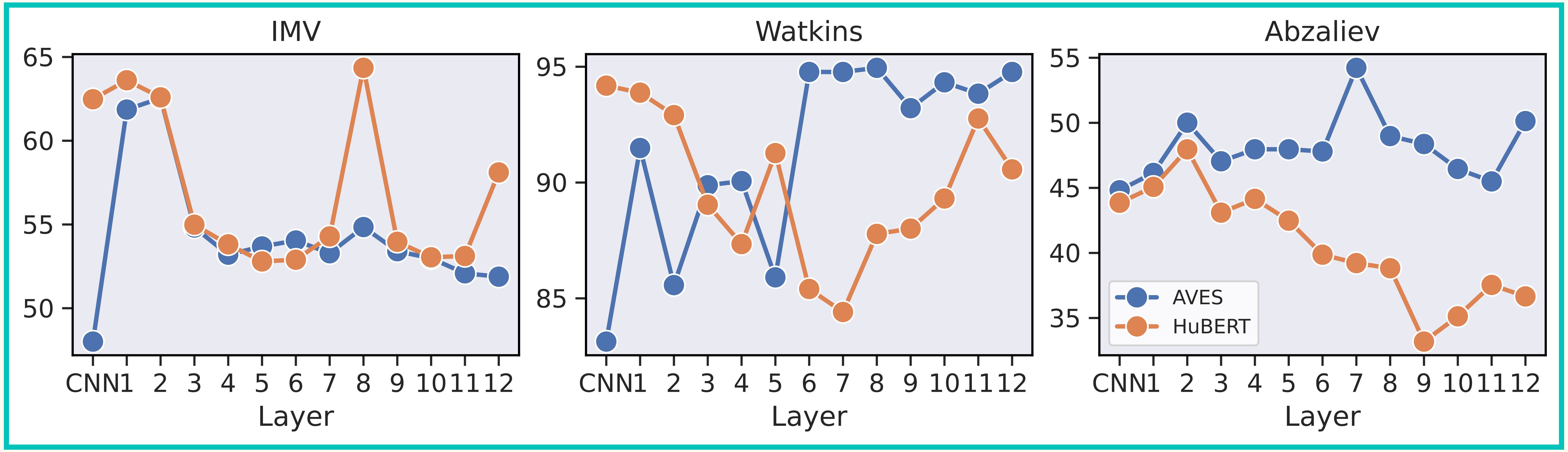


Classification Pipeline



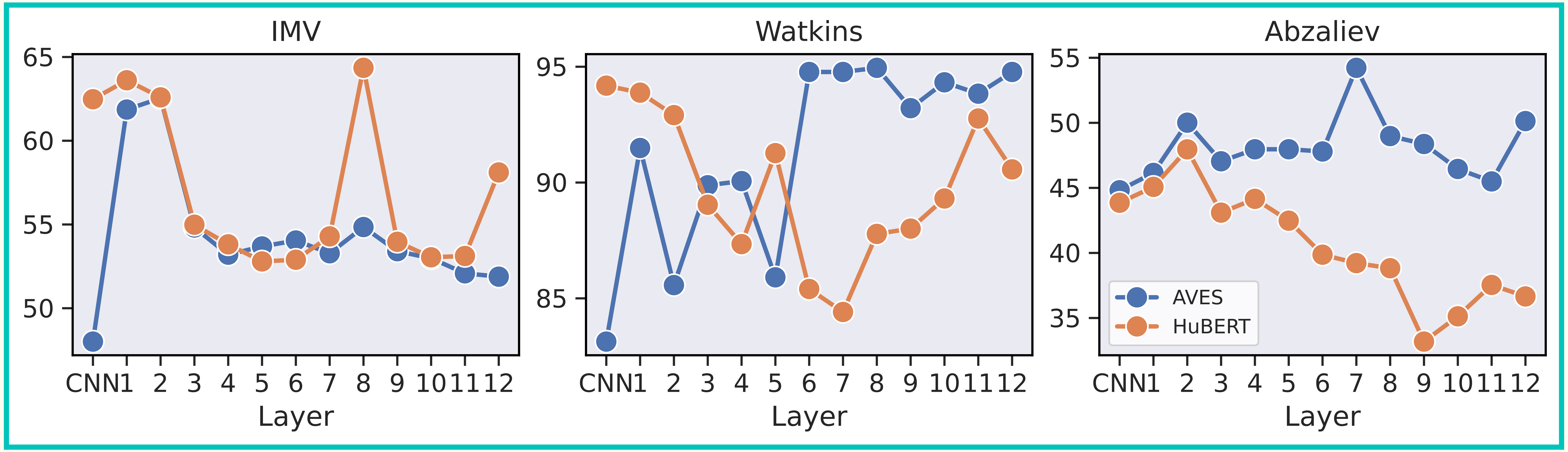
Experiments & Analysis

A. Pre-Training Domain Analysis



Layer-wise performance of AVES (•) against HuBERT (•).

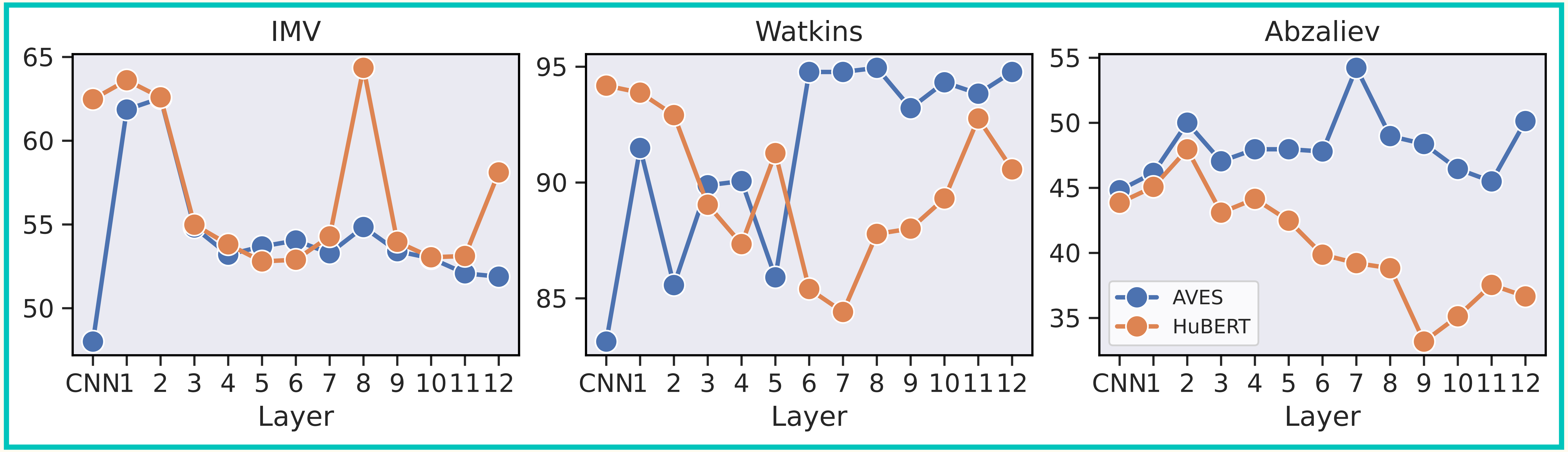
A. Pre-Training Domain Analysis



Layer-wise performance of AVES (•) against HuBERT (•).

- **IMV**: HuBERT > AVES in the initial and final layers. Both models show that initial layers are important - trend not limited to speech models.

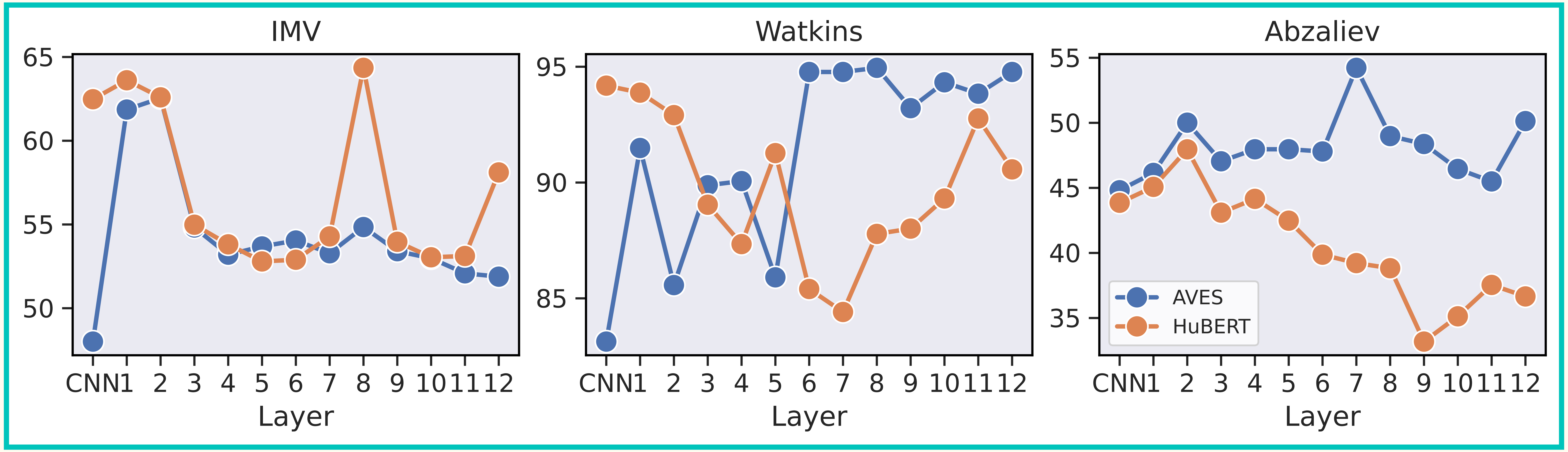
A. Pre-Training Domain Analysis



Layer-wise performance of AVES (•) against HuBERT (•).

- **IMV**: HuBERT > AVES in the initial and final layers. Both models show that initial layers are important - trend not limited to speech models.
- **Watkins**: AVES's initial layers are not as salient as later ones, where as HuBERT's middle layers are conversely the least useful.

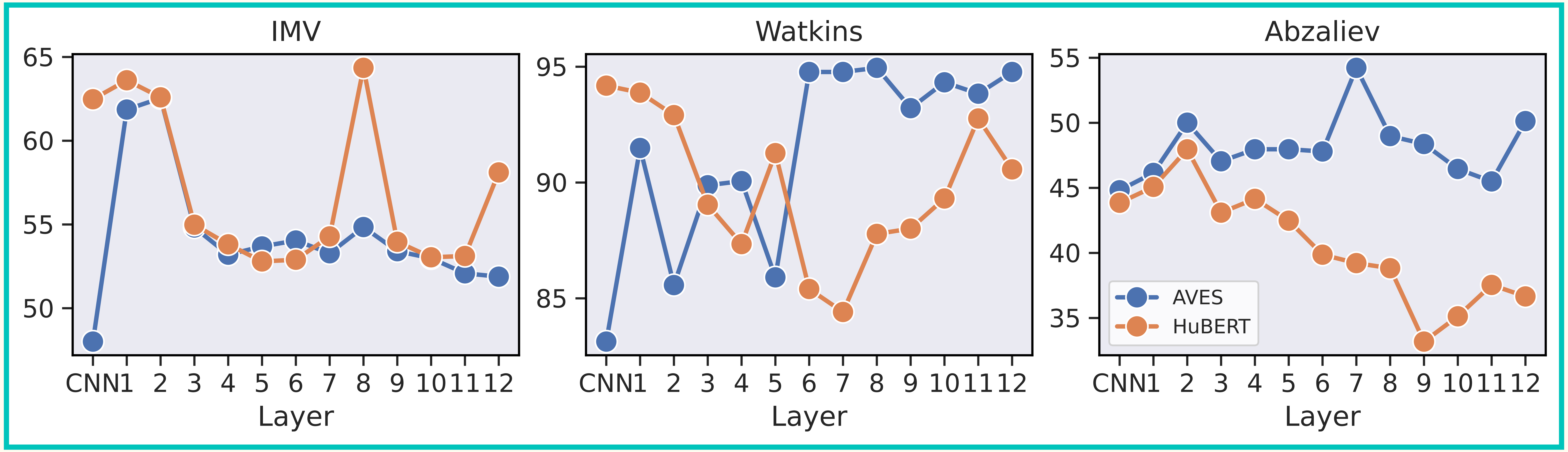
A. Pre-Training Domain Analysis



Layer-wise performance of AVES (•) against HuBERT (•).

- **IMV**: HuBERT > AVES in the initial and final layers. Both models show that initial layers are important - trend not limited to speech models.
- **Watkins**: AVES's initial layers are not as salient as later ones, where as HuBERT's middle layers are conversely the least useful.
- **Abzaliev**: AVES better overall. Initial and later layers contributing comparably. HuBERT doesn't scale well, follows downwards trend as IMV.

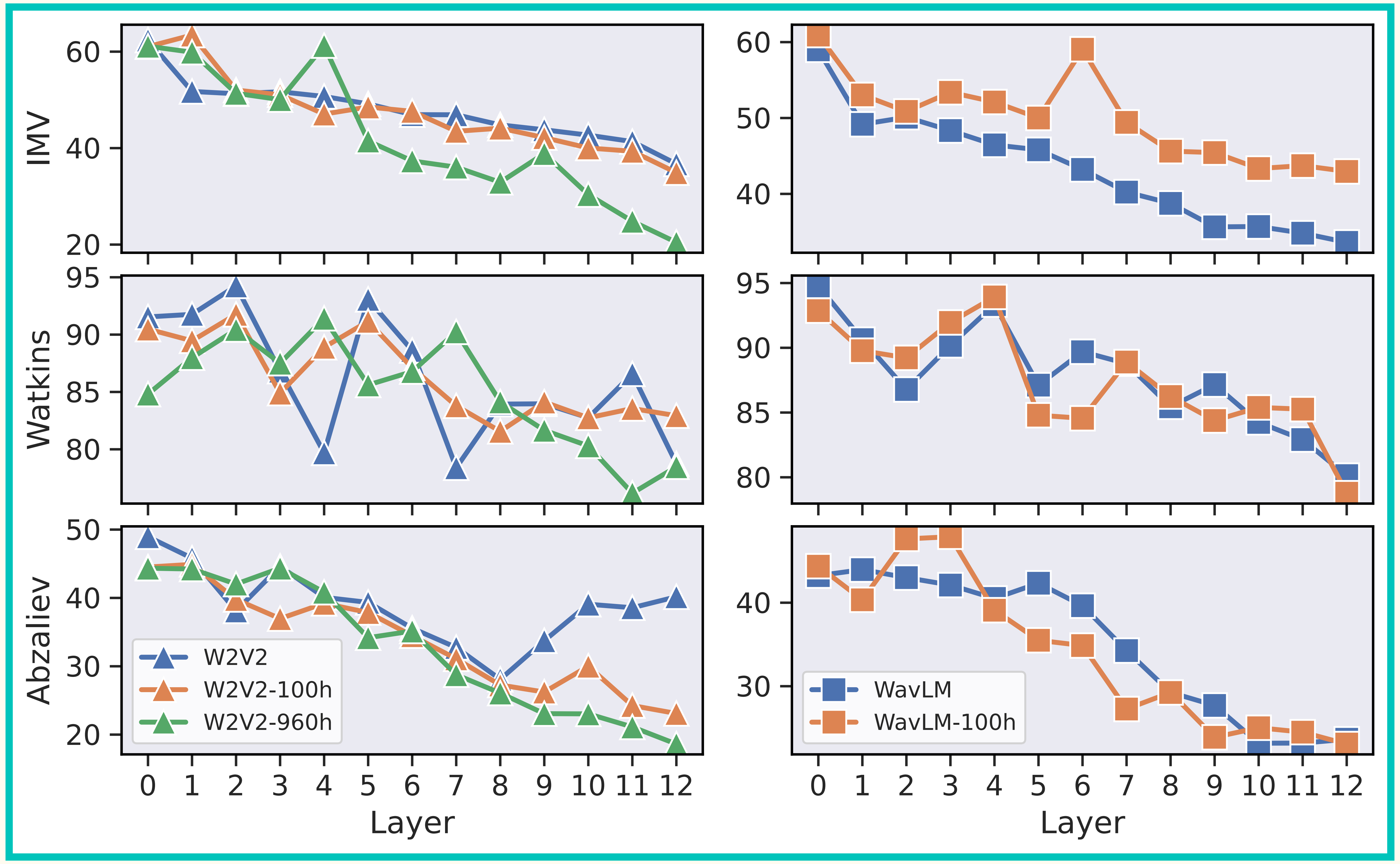
A. Pre-Training Domain Analysis



Layer-wise performance of AVES (•) against HuBERT (•).

- **IMV**: HuBERT > AVES in the initial and final layers. Both models show that initial layers are important - trend not limited to speech models.
- **Watkins**: AVES's initial layers are not as salient as later ones, where as HuBERT's middle layers are conversely the least useful.
- **Abzaliev**: AVES better overall. Initial and later layers contributing comparably. HuBERT doesn't scale well, follows downwards trend as IMV.
- **Overall**: Results indicate that pre-training on bioacoustic data can provide marginal improvements in some datasets/contexts.

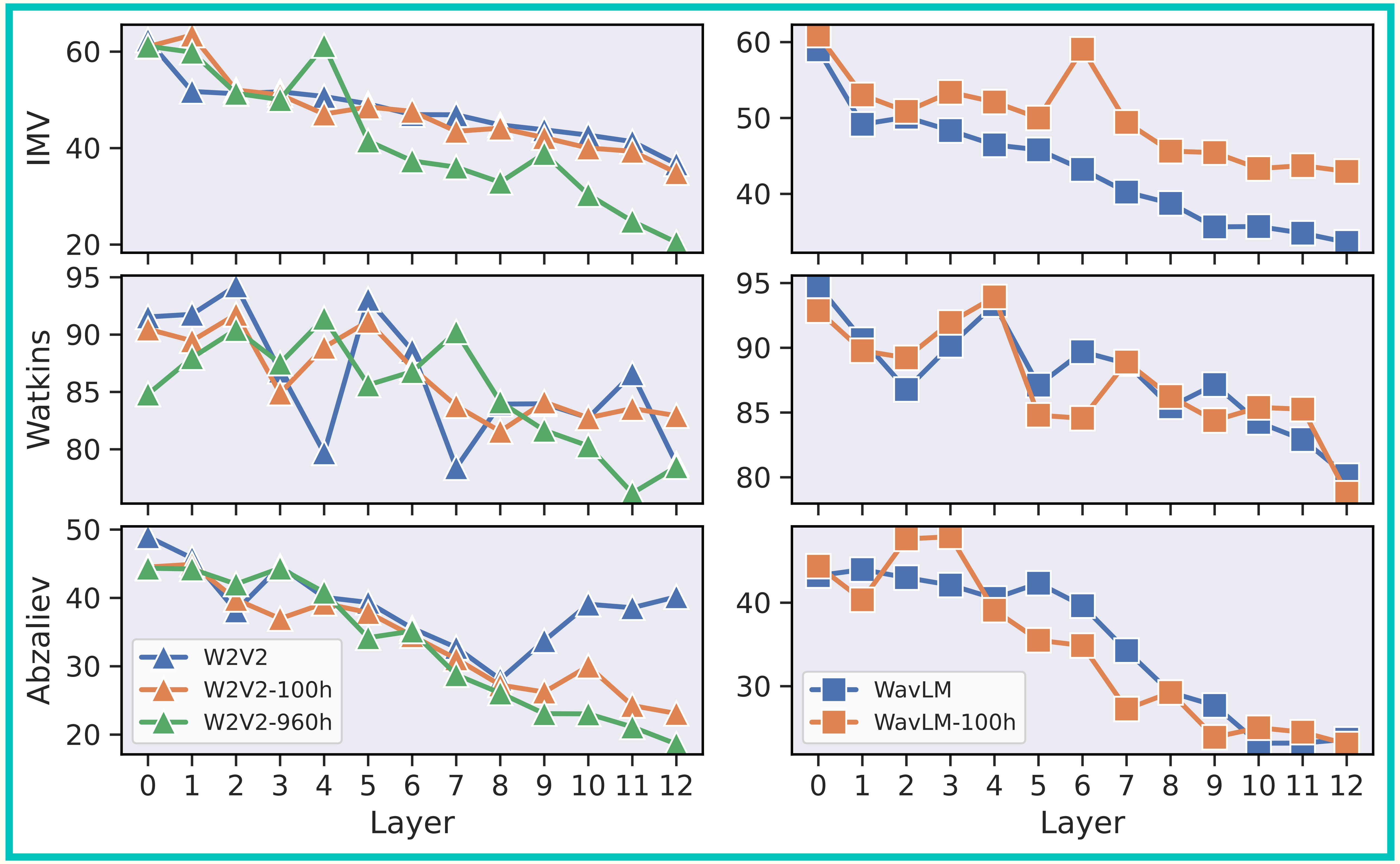
B. Fine-Tuning Analysis



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

B. Fine-Tuning Analysis

Fine-tuning yields mixed effects across both models and datasets.

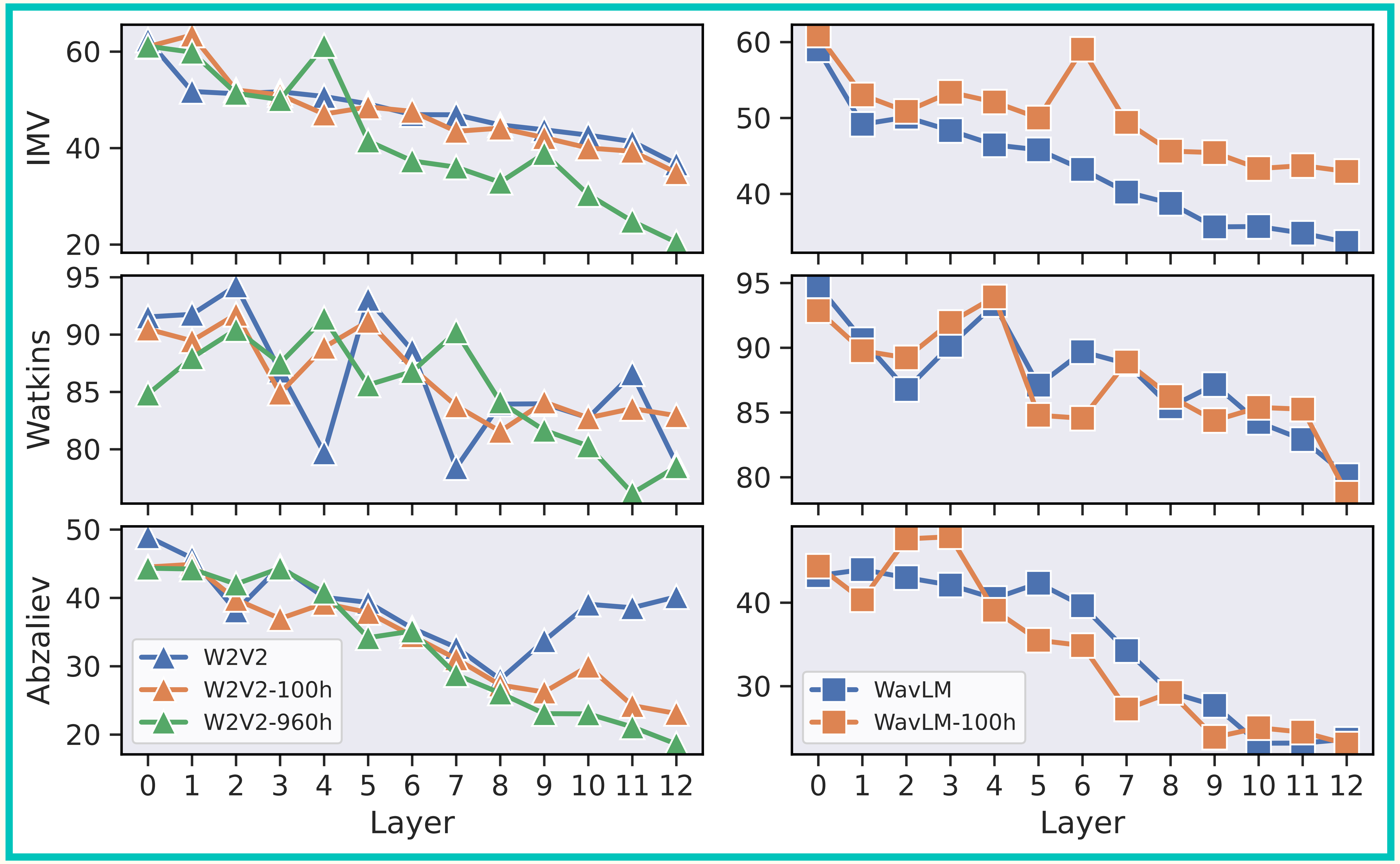


UAR of W2V2 (▲) and WLM (■) against their FT'd versions.

B. Fine-Tuning Analysis

Fine-tuning yields mixed effects across both models and datasets.

- FT models do not consistently outperform their base counterparts, particularly in W2V2-960h, with performance gains being marginal at best.

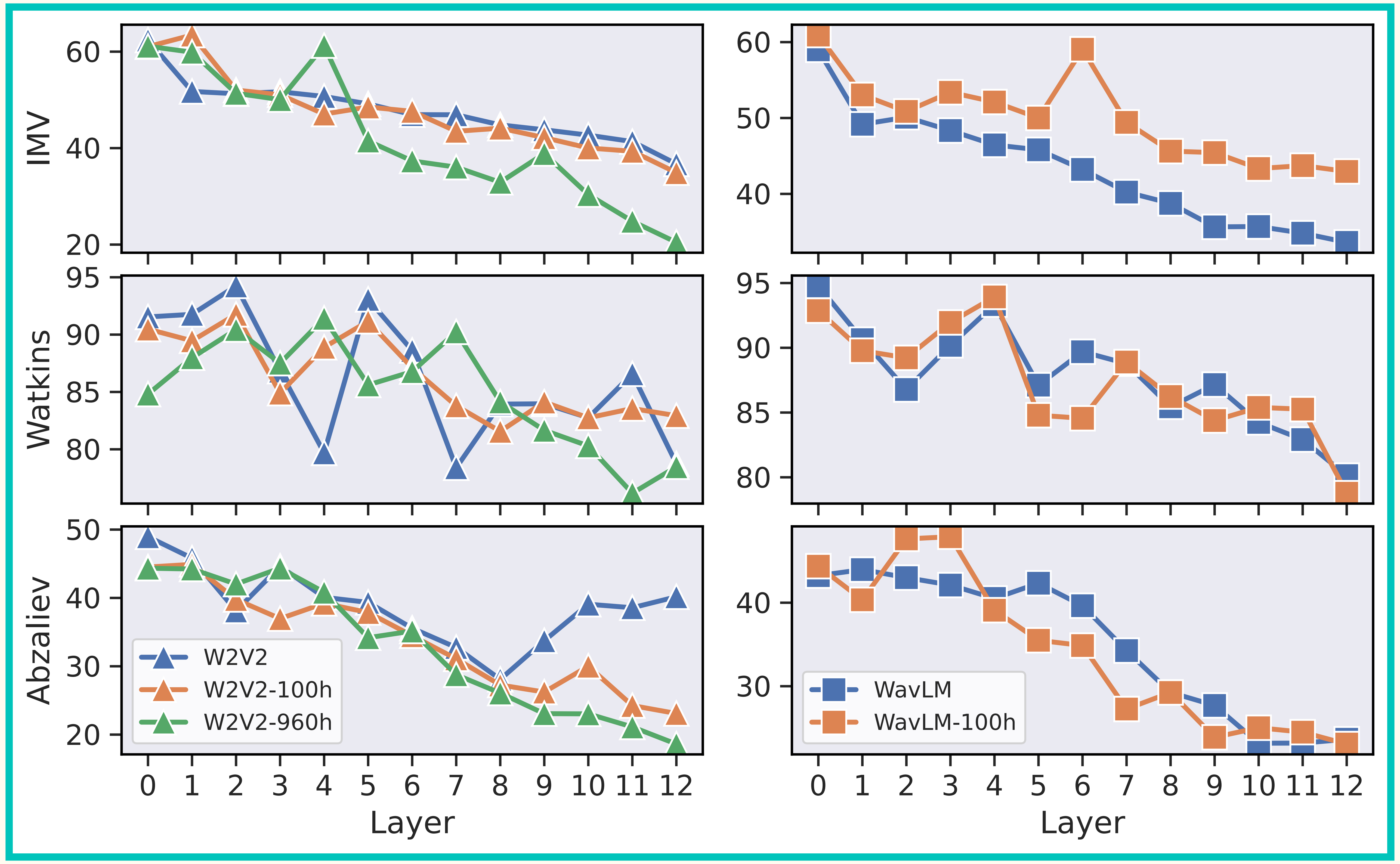


UAR of W2V2 (▲) and WLM (■) against their FT'd versions.

B. Fine-Tuning Analysis

Fine-tuning yields mixed effects across both models and datasets.

- FT models do not consistently outperform their base counterparts, particularly in W2V2-960h, with performance gains being marginal at best.
- Notably, FT'ing on more speech data, such as the 960h-W2V2, sometimes leads to a decline in performance in later layers, as seen on IMV and Abzaliev.

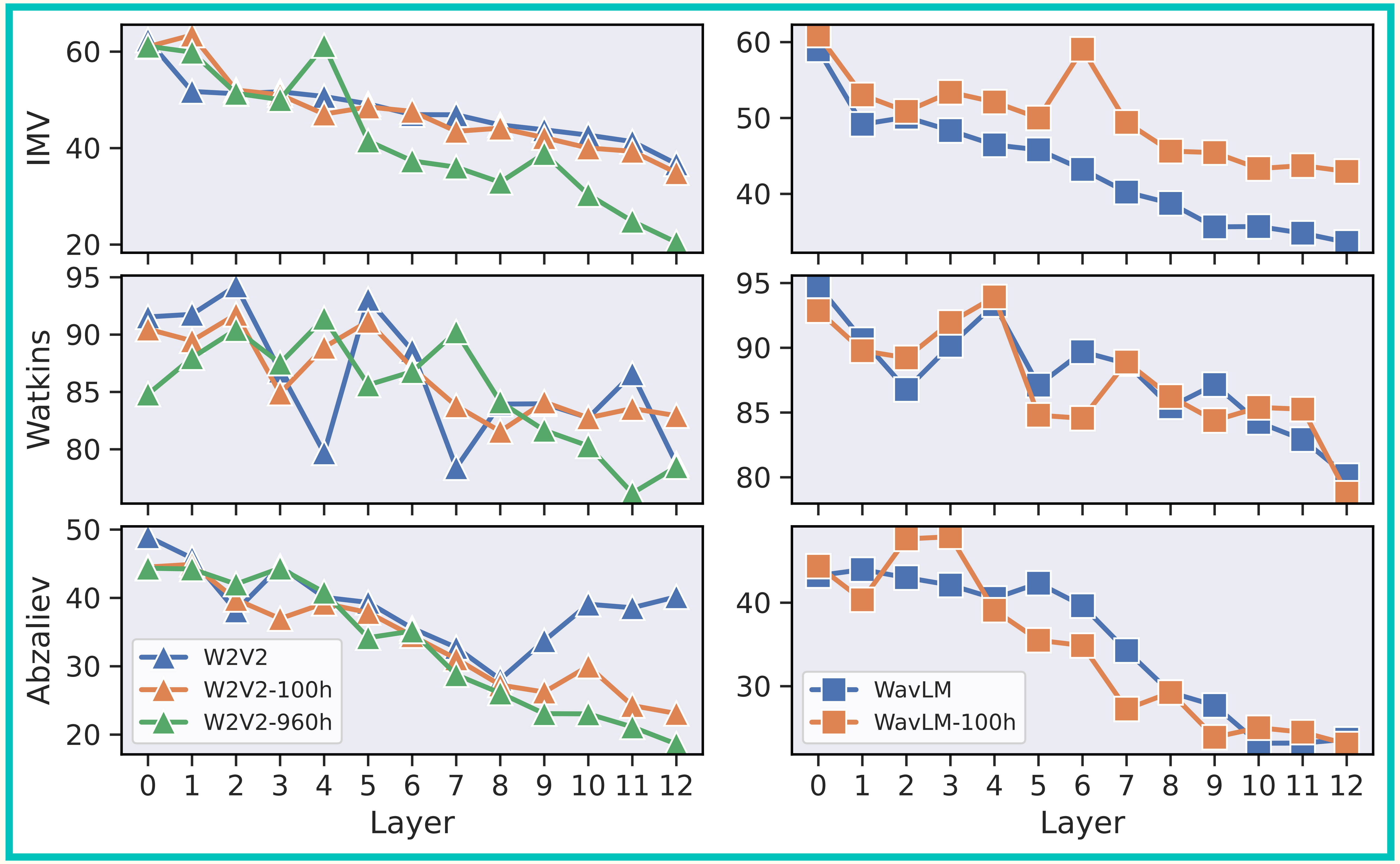


UAR of W2V2 (▲) and WLM (■) against their FT'd versions.

B. Fine-Tuning Analysis

Fine-tuning yields mixed effects across both models and datasets.

- FT models do not consistently outperform their base counterparts, particularly in W2V2-960h, with performance gains being marginal at best.
- Notably, FT'ing on more speech data, such as the 960h-W2V2, sometimes leads to a decline in performance in later layers, as seen on IMV and Abzaliev.
  - Suggests FT'ing on speech may push models to learn task-specific features that don't generalize as well to certain bioacoustic tasks.

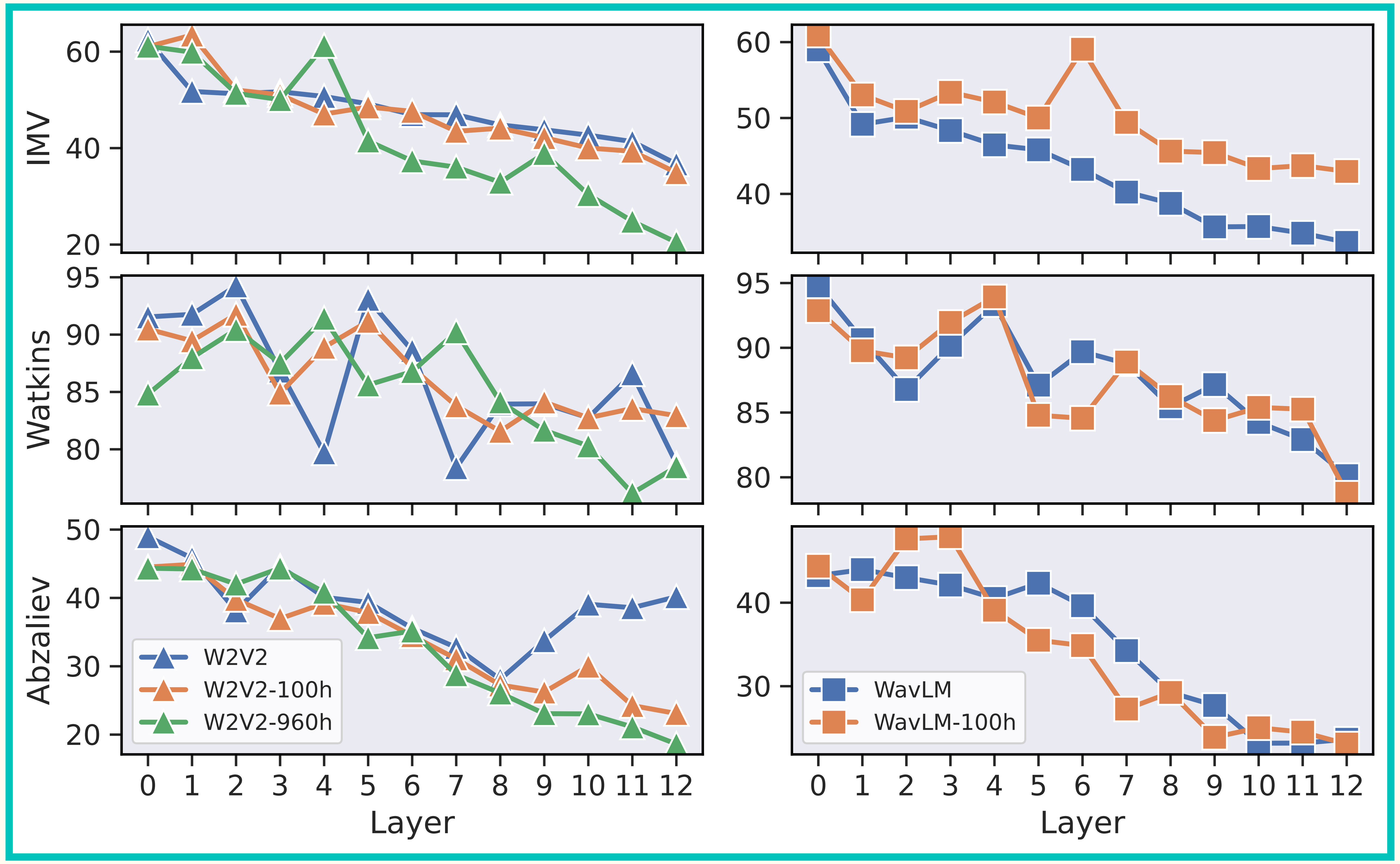


UAR of W2V2 (▲) and WLM (■) against their FT'd versions.

B. Fine-Tuning Analysis

Fine-tuning yields mixed effects across both models and datasets.

- FT models do not consistently outperform their base counterparts, particularly in W2V2-960h, with performance gains being marginal at best.
- Notably, FT'ing on more speech data, such as the 960h-W2V2, sometimes leads to a decline in performance in later layers, as seen on IMV and Abzaliev.
  - Suggests FT'ing on speech may push models to learn task-specific features that don't generalize as well to certain bioacoustic tasks.
- Interestingly, for non-FT models, earlier layers often capture enough general acoustic features to perform adequately.

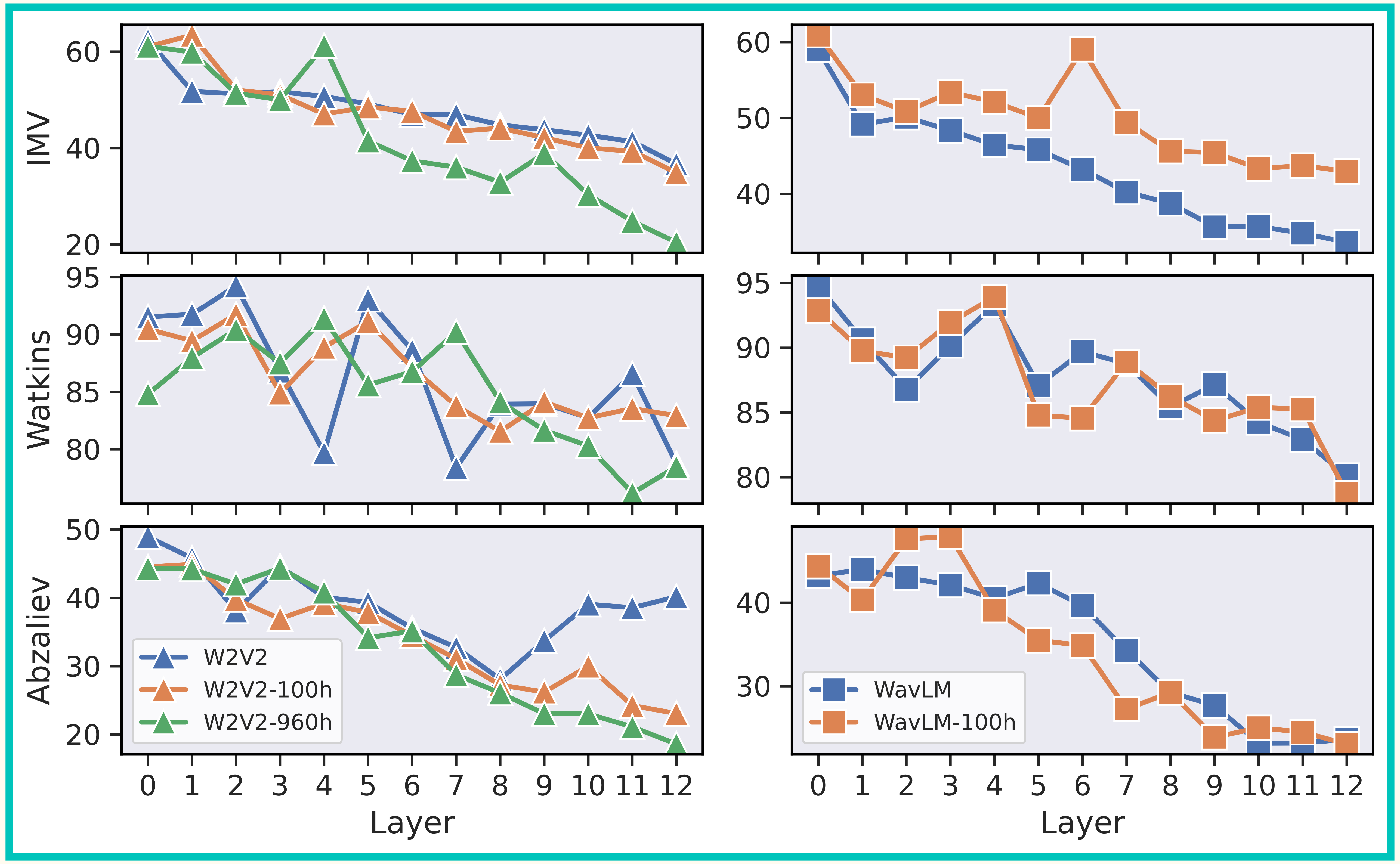


UAR of W2V2 (▲) and WLM (■) against their FT'd versions.

B. Fine-Tuning Analysis

Fine-tuning yields mixed effects across both models and datasets.

- FT models do not consistently outperform their base counterparts, particularly in W2V2-960h, with performance gains being marginal at best.
- Notably, FT'ing on more speech data, such as the 960h-W2V2, sometimes leads to a decline in performance in later layers, as seen on IMV and Abzaliev.
    - Suggests FT'ing on speech may push models to learn task-specific features that don't generalize as well to certain bioacoustic tasks.
- Interestingly, for non-FT models, earlier layers often capture enough general acoustic features to perform adequately.
- However, for fine-tuned models, layer selection becomes more important/necessary, as different layers may capture more specialized representations that could benefit specific certain tasks.



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

C. Comparative Analysis

Type	$\mathcal{F}$	IMV	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

C. Comparative Analysis

- Best scores from AVES and HuBERT.

Type	$\mathcal{F}$	IMV	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

C. Comparative Analysis

- Best scores from AVES and HuBERT.
  - Yield very comparable performances for both IMV and Watkins.

Type	$\mathcal{F}$	IMV	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

C. Comparative Analysis

- Best scores from AVES and HuBERT.
  - Yield very comparable performances for both IMV and Watkins.
  - HuBERT's representations are robust for call classification tasks across different species.

Type	$\mathcal{F}$	IMV	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

C. Comparative Analysis

- Best scores from AVES and HuBERT.
  - Yield very comparable performances for both IMV and Watkins.
  - HuBERT's representations are robust for call classification tasks across different species.
- All the best scores are from the PT category, as well as the second best scores.

Type	$\mathcal{F}$	IMV	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

C. Comparative Analysis

- Best scores from AVES and HuBERT.
  - Yield very comparable performances for both IMV and Watkins.
  - HuBERT's representations are robust for call classification tasks across different species.
- All the best scores are from the PT category, as well as the second best scores.
  - Fine-tuning PT'd speech models on an ASR does not consistently bring us any advantage over PT'd alone.

Type	$\mathcal{F}$	IMV	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

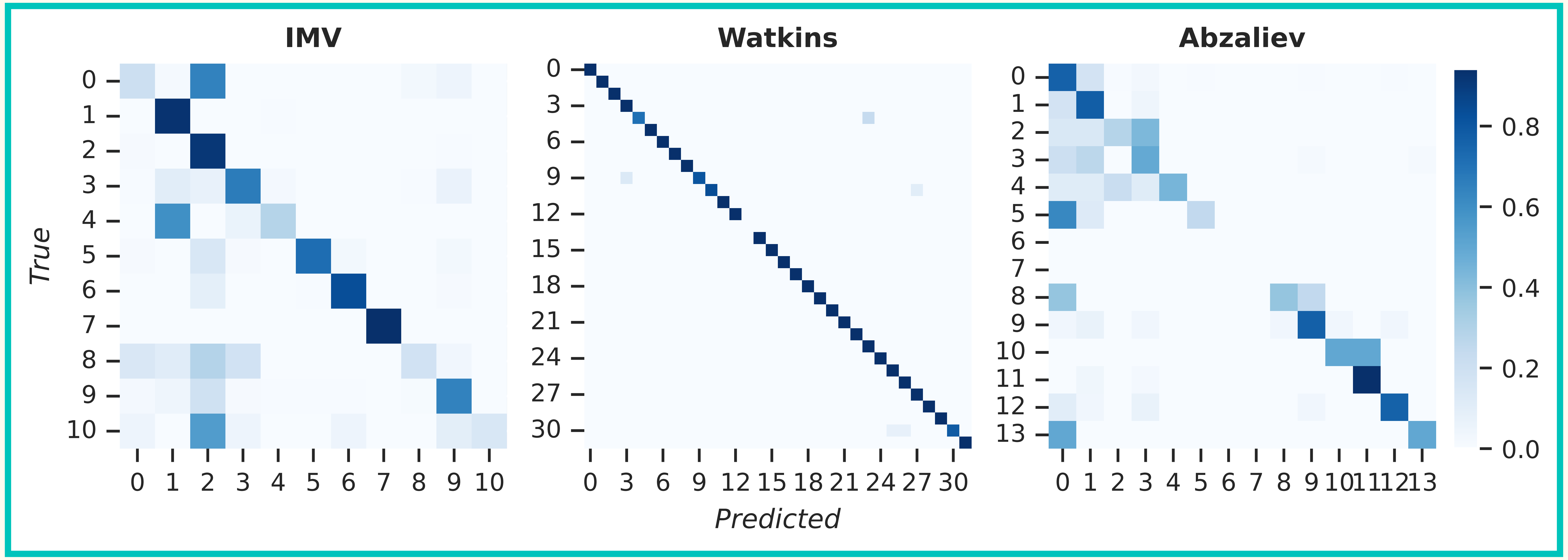
C. Comparative Analysis

- Best scores from AVES and HuBERT.
  - Yield very comparable performances for both IMV and Watkins.
  - HuBERT's representations are robust for call classification tasks across different species.
- All the best scores are from the PT category, as well as the second best scores.
  - 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	Watkins	Abzaliev
PT	AVES	62.54	**94.95**	**54.23**
	HuBERT	**64.35**	94.18	47.96
	WavLM	58.98	<u>94.78</u>	43.97
	W2V2	62.40	94.25	<u>48.95</u>
PT + FT	WavLM-100h	60.93	93.93	47.90
	W2V2-100h	<u>63.44</u>	91.77	44.91
	W2V2-960h	61.25	91.42	44.36
	Fusion	62.48	94.78	48.95

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

C. Comparative Analysis



Confusion matrices of the best feature layers' fusion.

Good general classification alignment.

- **IMV**: False positives for call-type ID 2. High occurrence in dataset. Wide spectral range.
- **Watkins**: Easiest to classify. Clear acoustic/spectral differences. Class ID 13 only had 2 samples.
- **Abzaliev**: Confusion between barks (IDs 0-5): overlapping acoustic features. ID 6 had few samples. ID 7 removed.

Conclusion

Conclusion

Conclusion

- **Summary**: Paper compared SSL models pre-trained on speech and animal calls for bioacoustic tasks.

Conclusion

- **Summary**: Paper compared SSL models pre-trained on speech and animal calls for bioacoustic tasks.

1. Impact of pre-training domains: pre-training on bioacoustic data mostly yields comparable performance to pre-training on speech, but can offer limited advantages in select contexts.

Conclusion

- **Summary**: Paper compared SSL models pre-trained on speech and animal calls for bioacoustic tasks.

1. Impact of pre-training domains: pre-training on bioacoustic data mostly yields comparable performance to pre-training on speech, but can offer limited advantages in select contexts.
2. Impact of fine-tuning PT'd speech models on ASR for animal vocalizations: fine-tuning yielded inconsistent results, suggesting that the general-purpose representations learned during pre-training may already be well-suited for bioacoustic tasks.

Conclusion

- **Summary**: Paper compared SSL models pre-trained on speech and animal calls for bioacoustic tasks.

1. Impact of pre-training domains: pre-training on bioacoustic data mostly yields comparable performance to pre-training on speech, but can offer limited advantages in select contexts.
2. Impact of fine-tuning PT'd speech models on ASR for animal vocalizations: fine-tuning yielded inconsistent results, suggesting that the general-purpose representations learned during pre-training may already be well-suited for bioacoustic tasks.

- **Conclusion**: results highlight the utility of PT speech models for bioacoustic tasks, even without FT.

Source code

Thank you !

Idiap Research Institute

eklavya.sarkar@idiap.ch

Acknowledgments: NCCR Evolving Language, Dr. Humberto Pérez-Espinosa.
Pic. credit: Michael B. Habib, 2020. *Fossils Reveal When Animals Started Making Noise.* Scientific American 326, 1, 42-47, Jan 22.