Investigating self-supervised speech models' ability to classify animal vocalizations:

The case of gibbon's vocal identity

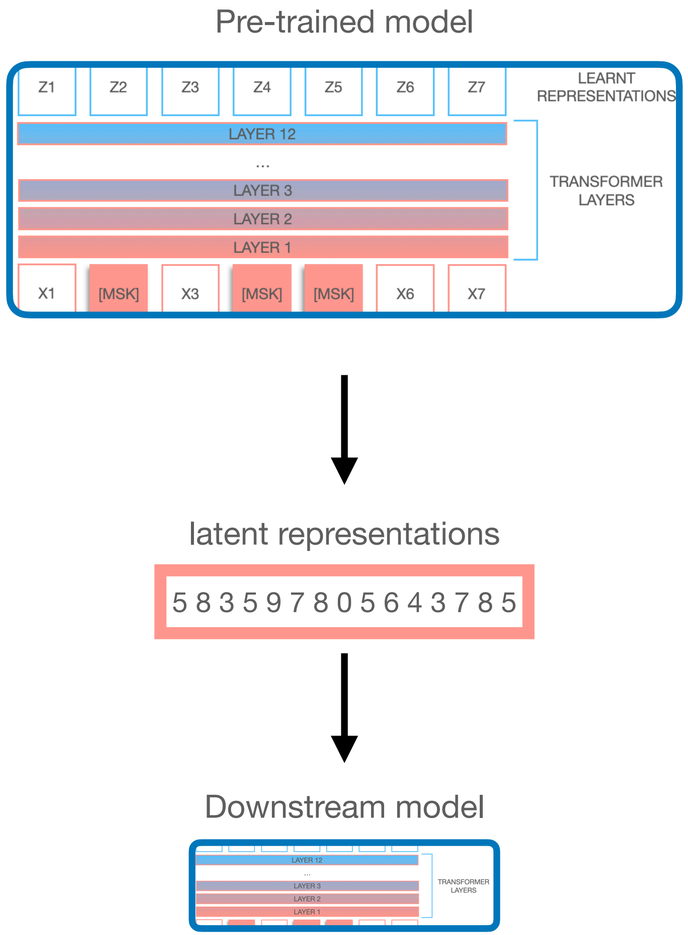
Self-supervised pre-training

- Self supervision :
  - BERT - GPT
  - ResNET - MAE
  - Wav2vec - HuBERT Audio-MAE - APC - CNNs…

text

images

sound



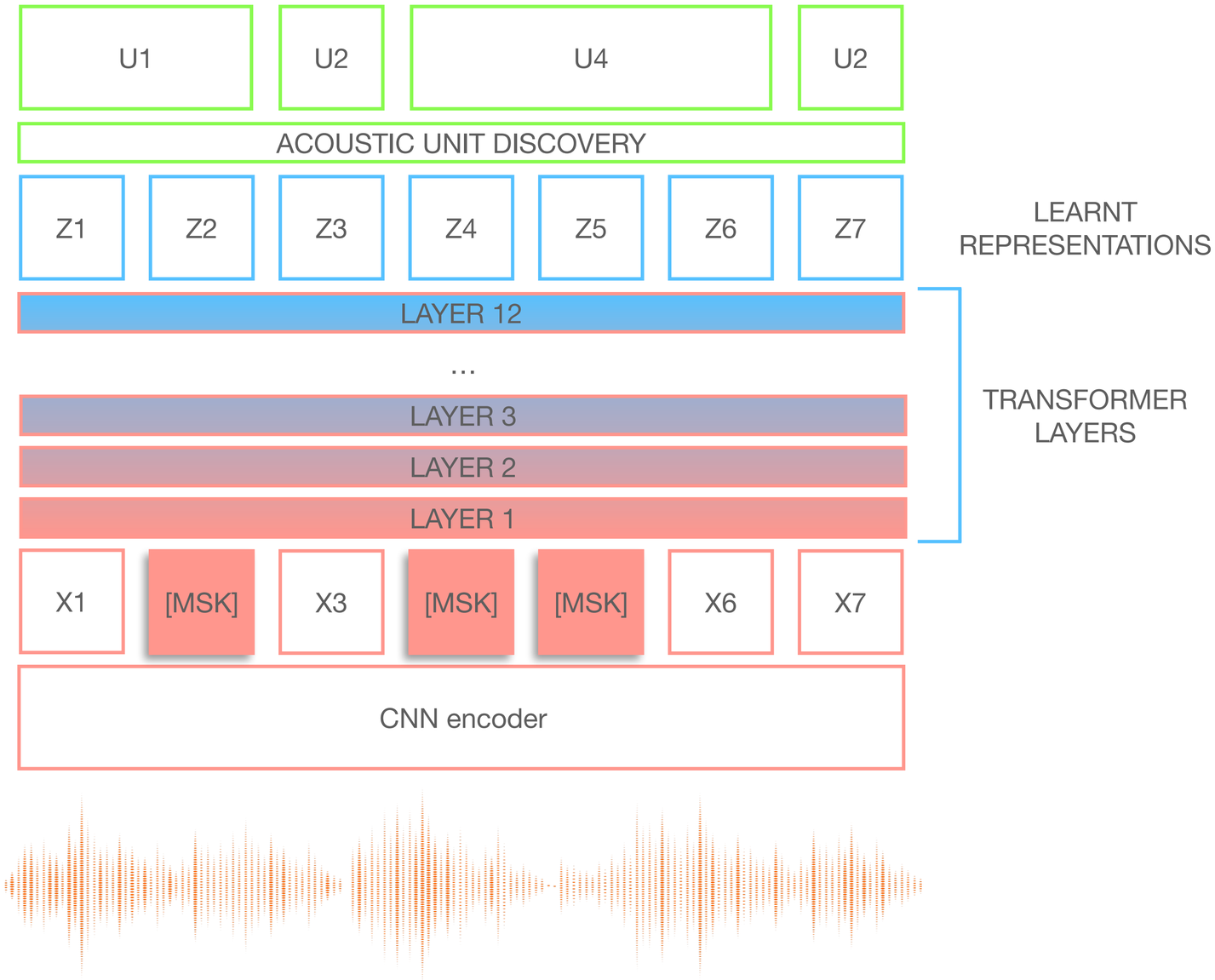
SSL acoustic representations

SSL speech model

Trained on **raw speech**.

Pretext task : **masked prediciton, predictive coding**

Fine-tuned for Automatic Speech Recognition (ASR)



Out-of-domain abilities

Heggan et al., 2024 - Yang et al., 2021 *SUPERB Benchmark*

- Speech tasks:
  - CONTENT: Phoneme Recognition / Automatic Speech Recognition
  - SPEAKER: Speaker Identification / Speaker Diarization
  - SEMANTICS: Intent Classification
  - PARALINGUISTICS: Emotion Recognition
- Other:
  - ENVIRONMENTAL: Audio-Tagging
  - BIOACOUSTICS: Species Classification

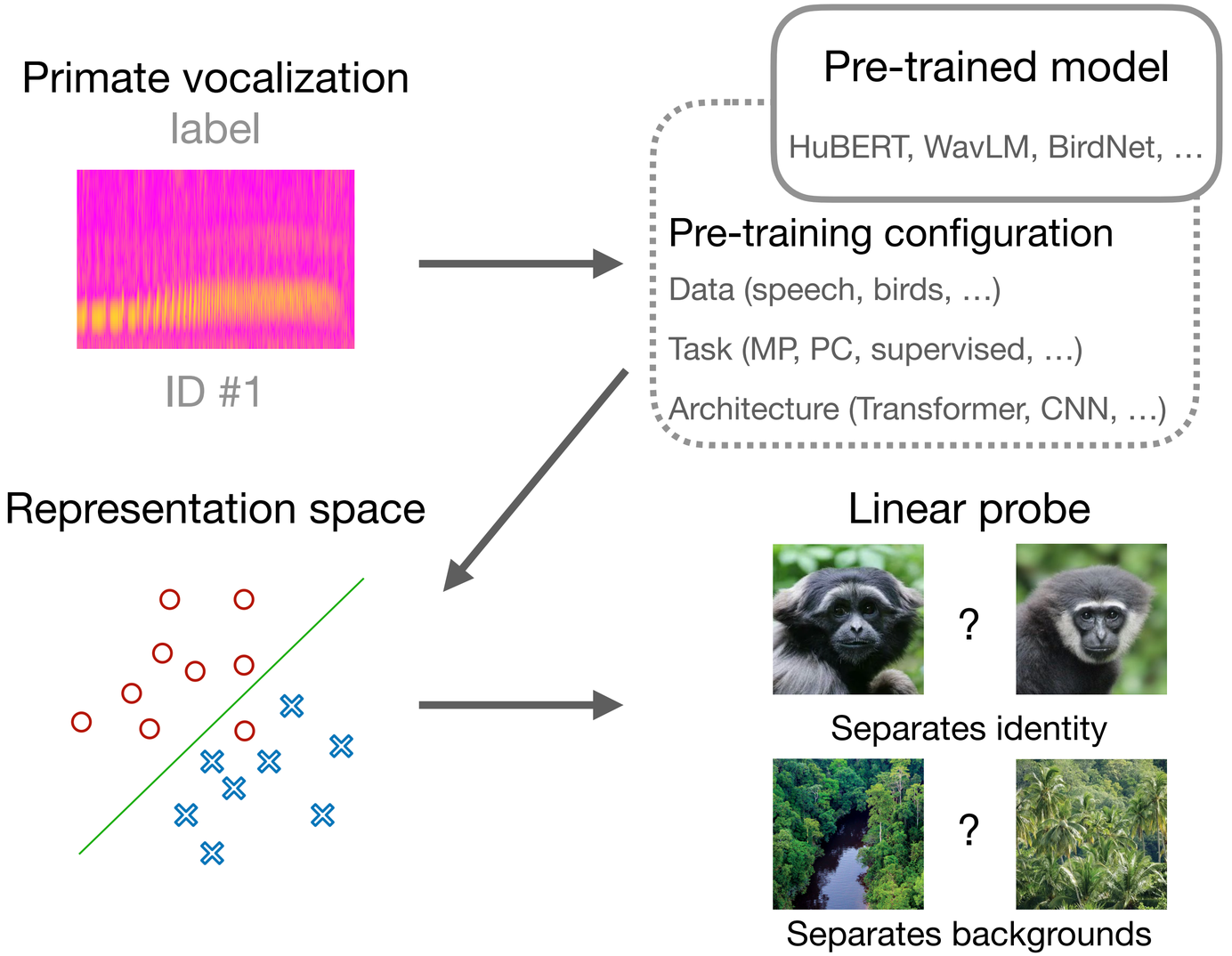
Computational bioacoustics

- Similar context
- Large datasets (PAM)
- SOTA supervised learning (task- and data-specific)
- Difficult annotations / bias
- Interest for unsupervised - SSL
- “foundation models”

Speech pre-training → Primate bioacoustics

Linear probing

- Linear regression
- Frozen representations
- Accuracy

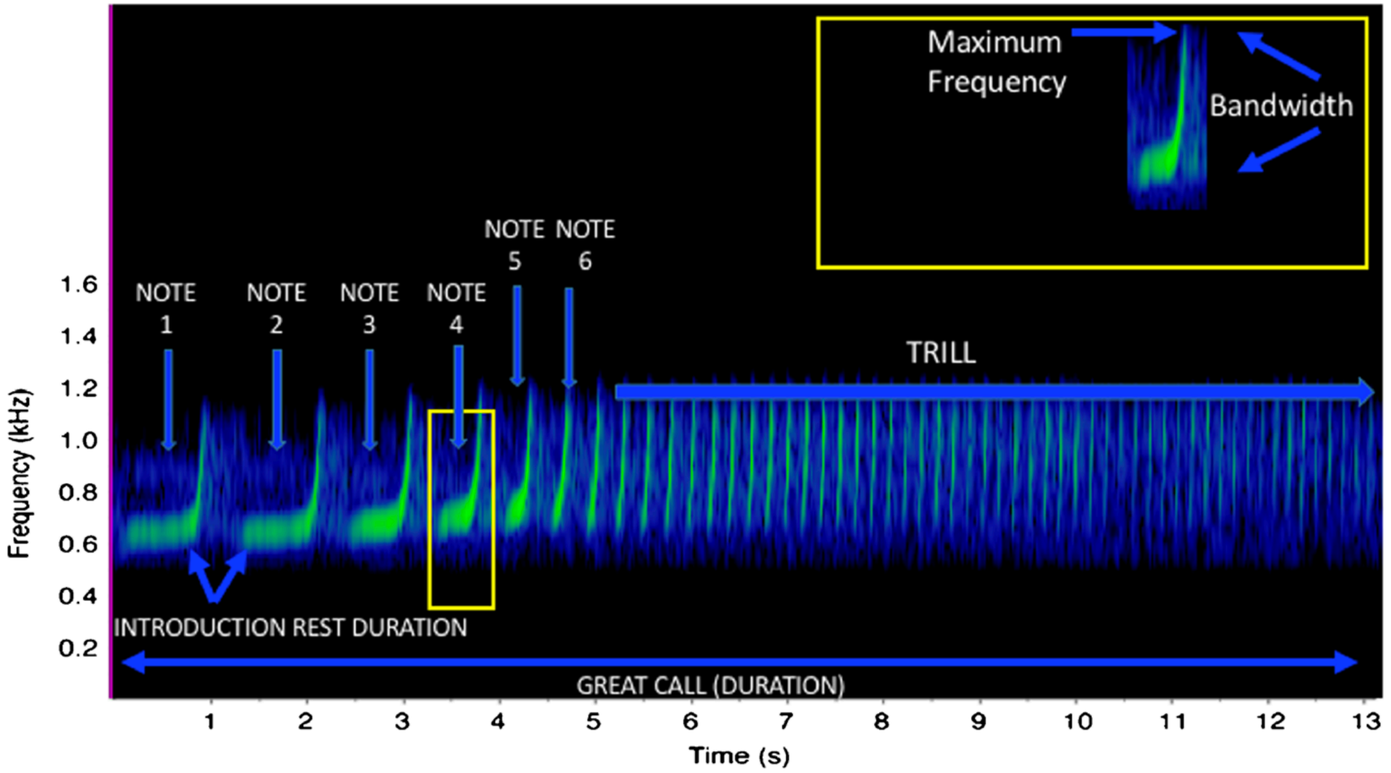


Northern Grey Gibbons

- Mammals > Primates > Apes
- Malaysia - North Borneo
- Singing apes - long distance calling
- Vocalizations > Duets > Great Calls
- Identity labels

Dataset

- Stereotyped call
- Important frequencial modulations
- 91 females
  1 to 47 recordings
- Test = 10 females
  25 recordings



H1: Comparison

Comparing models

- Size
- Architecture
- Pre-training data
  - Speech models
  - Bird species classifiers
  - Audio-taggers
- Baselines (chance - MFCC)

Table 1: *Pre-trained model characteristics.*
number of Transformer Layers (n TL) - Masked Prediction (MP) - Predictive Coding (PC) - Convolutional Neural Network (CNN)

Model	Genre	Hours	Arch.	Task	#param.	Embedding	Window
[20] HuBERT Large	speech	60k	24 TL	MP	317M	1280	20 ms
[20] HuBERT Base	speech	960	12 TL	MP	95M	768	20 ms
[21] UniSpeech SAT Large	speech	94k	24 TL	MP + speaker	317M	1280	20 ms
[22] Wav2vec 2.0 Large	speech	53k	24 TL	contrastive PC	317M	1280	20 ms
[23] WavLM Large	speech	94k	24 TL	MP + robustness	90M	1280	20 ms
[24] APC	speech	360	3 TL	autoregressive PC	4M	512	10 ms
[7] Google perch	bird	10k	CNN	supervised	20M	1280	5000 ms
[6] BirdNET 2.3	bird	4k	CNN	supervised	10M	1280	3000 ms
[25] Audio-MAE AST	general + speech	50k + 960	MAE	MP	90M	768	20 ms
[26] Vggish	general	5k	CNN	supervised	10M	128	96 ms

H1: Comparison

Full test set

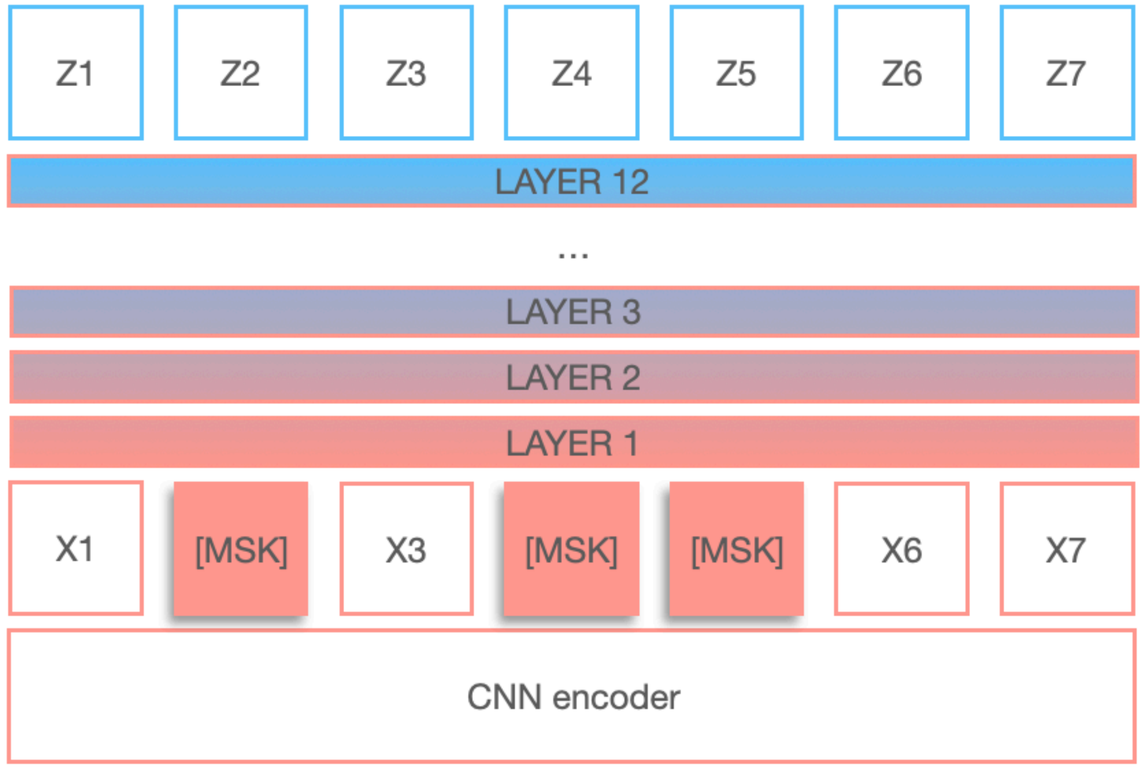
- Selection of 10 individuals
- Test (10%)
- Train (80%) - 10 random seed trainings
- Frozen representation extraction
- Accuracy

H2: Layer-wise analysis

- The last layer approach
- Layer-wise variation

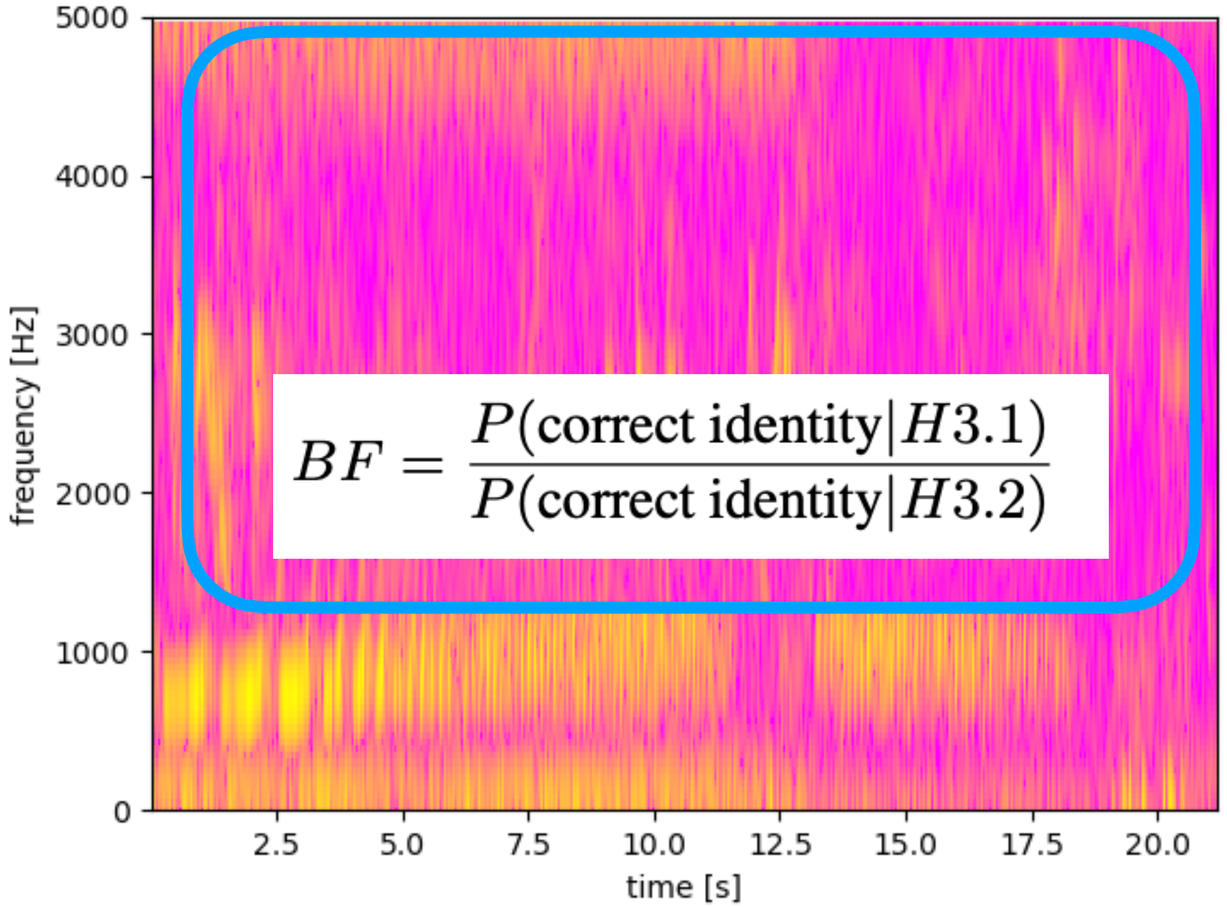
(Pasad et al. 2023)

- Initial layers performance



H3: Background noise

- Manual segmentation
- What is background noise?
  - Channel
  - Territorial signature
  - Birds
- The Bayes Factor



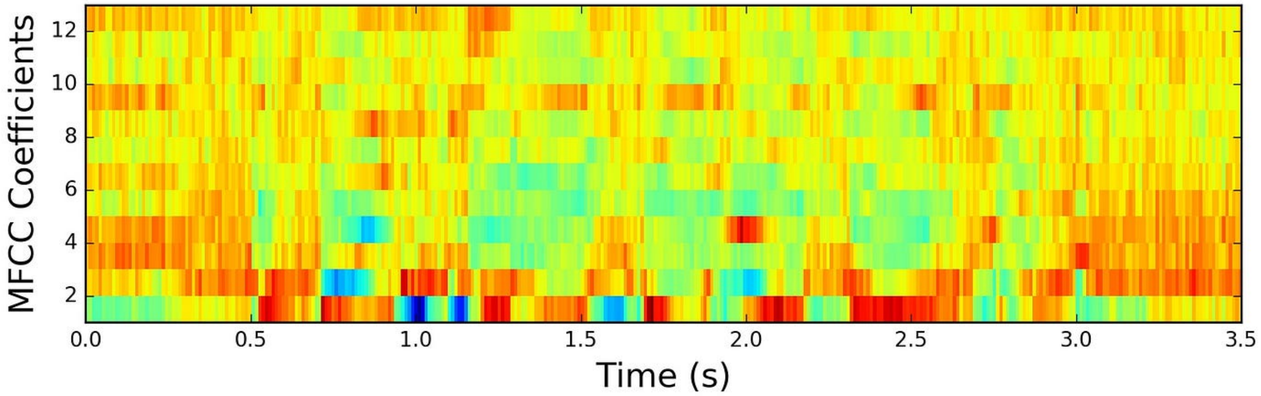
H4: Temporal information

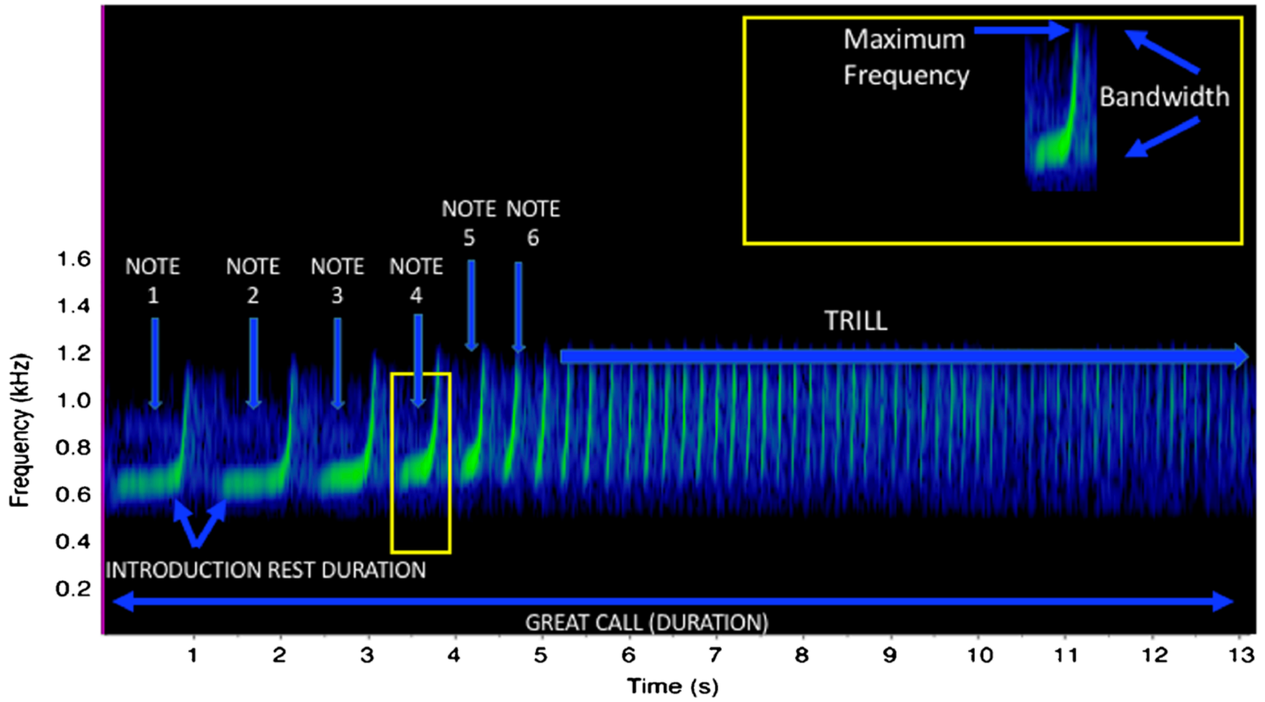
- What is a vocal signature ?
  - Spectral information
  - Temporal information

Gamba 2017 - Charlton et al. 2020

Fitch et al. 2002 - Terleph et al. 2015 - Bradbury 2011

- Segmented single notes
- Performance drop as an indicator of temporal feature extraction





Results

H1 - Full

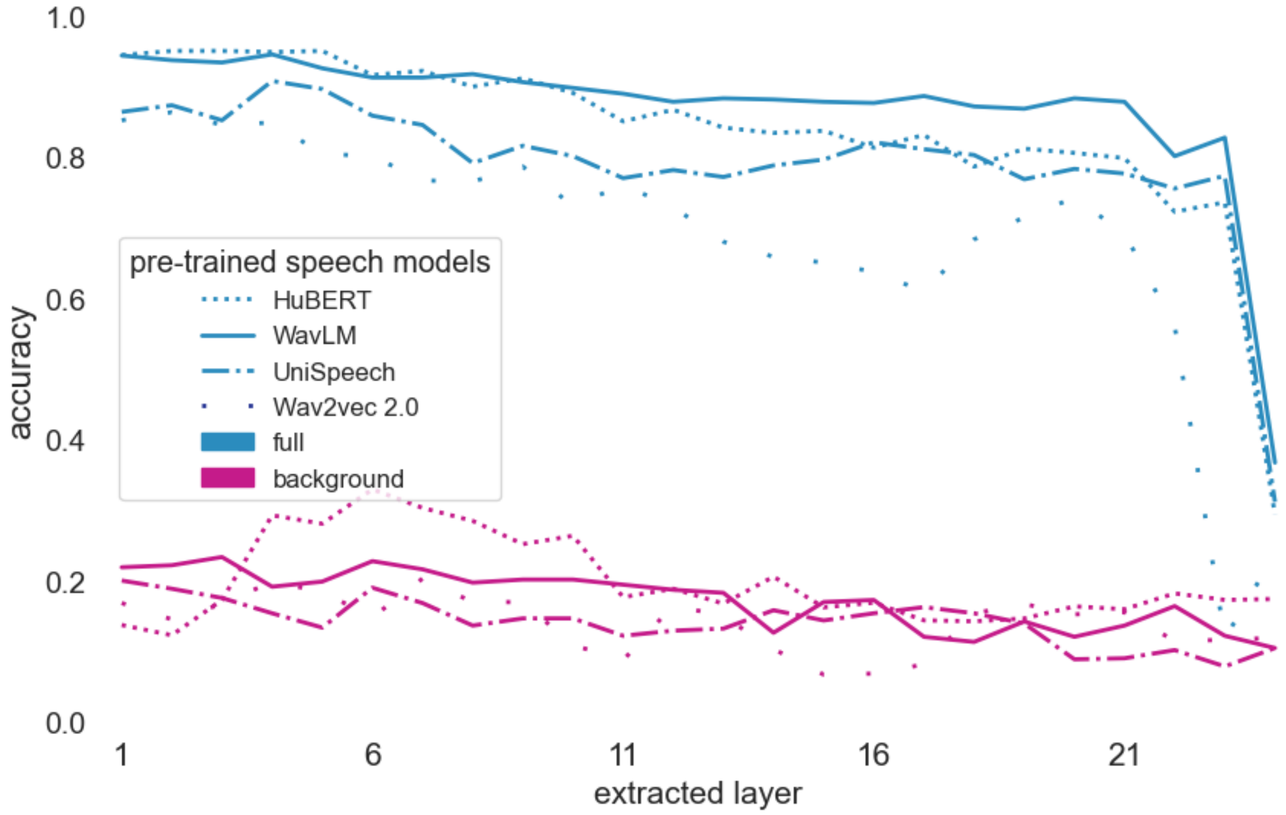
- PT Dataset size - number of parameters
- PT Dataset nature
- Architecture

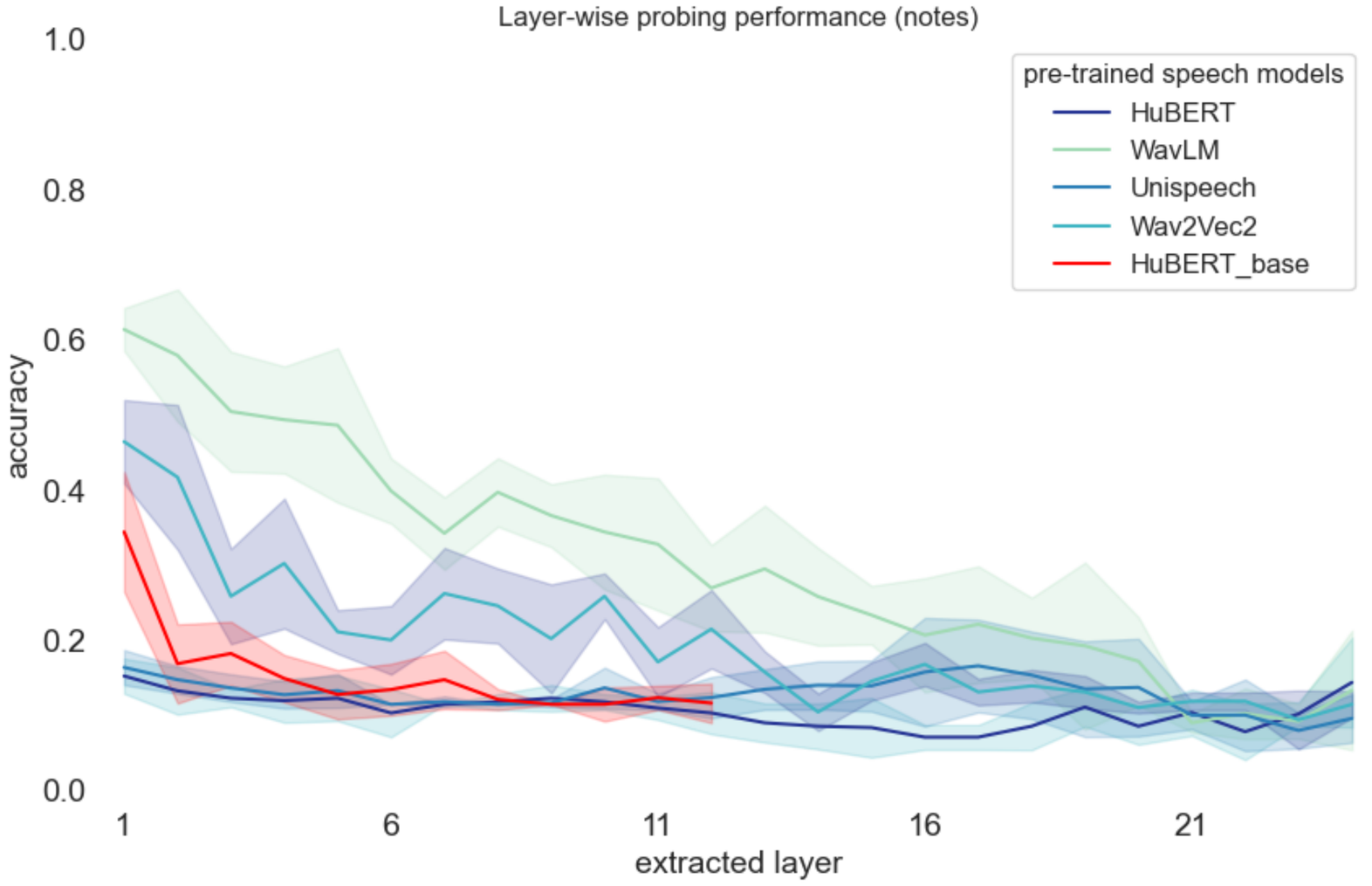
Model	Full
HuBERT Large	**0.95**
HuBERT Base	0.72
UniSpeech-SAT	0.87
Wav2vec 2.0 Large	0.86
WavLM Large	**0.94**
APC	0.75
Google perch	0.87
BirdNET 2.3	0.87
Audio-MAE AST	**0.95**
Vggish	*0.66*
MFCC	0.82
Chance	0.10

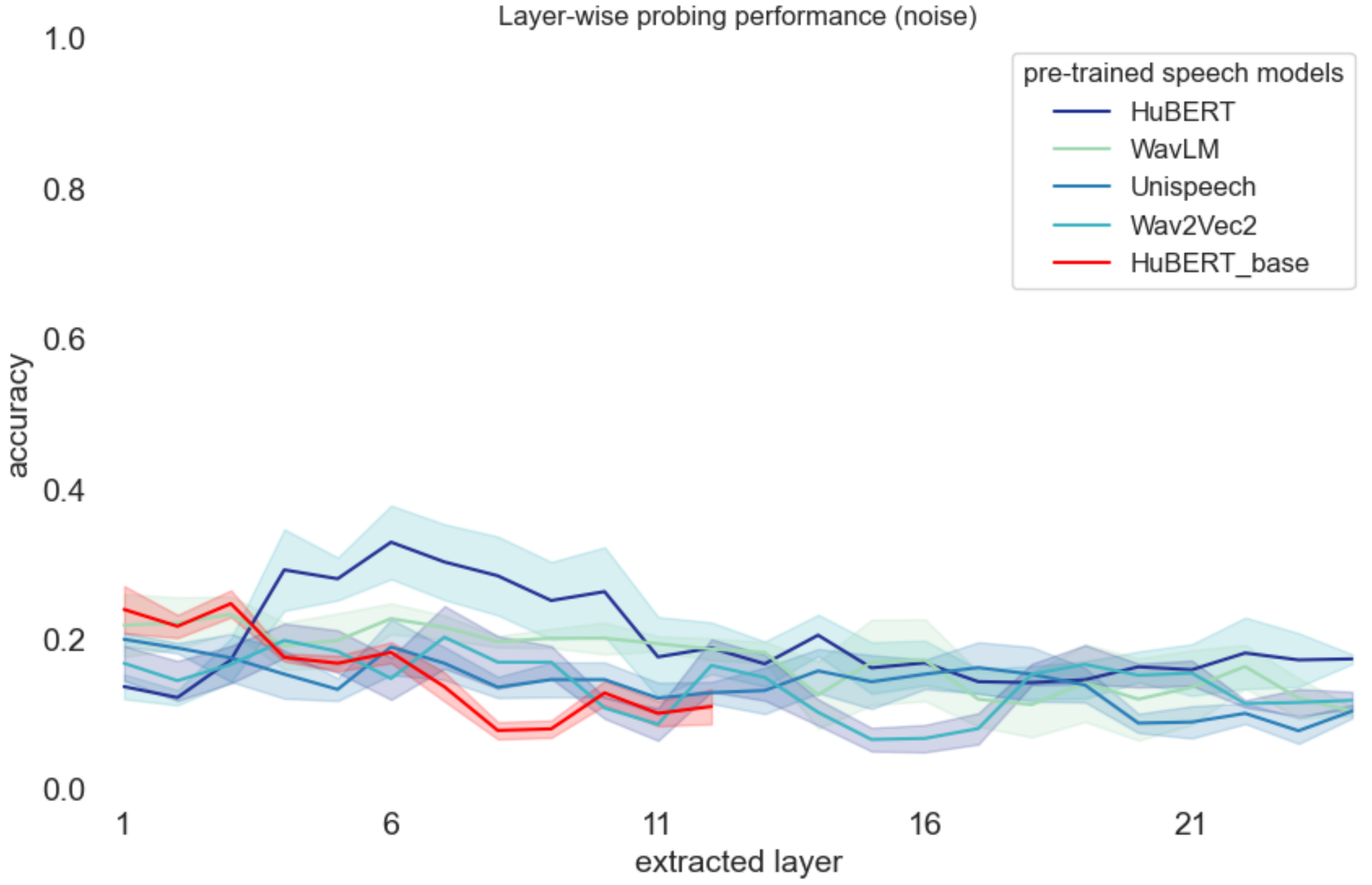
Results

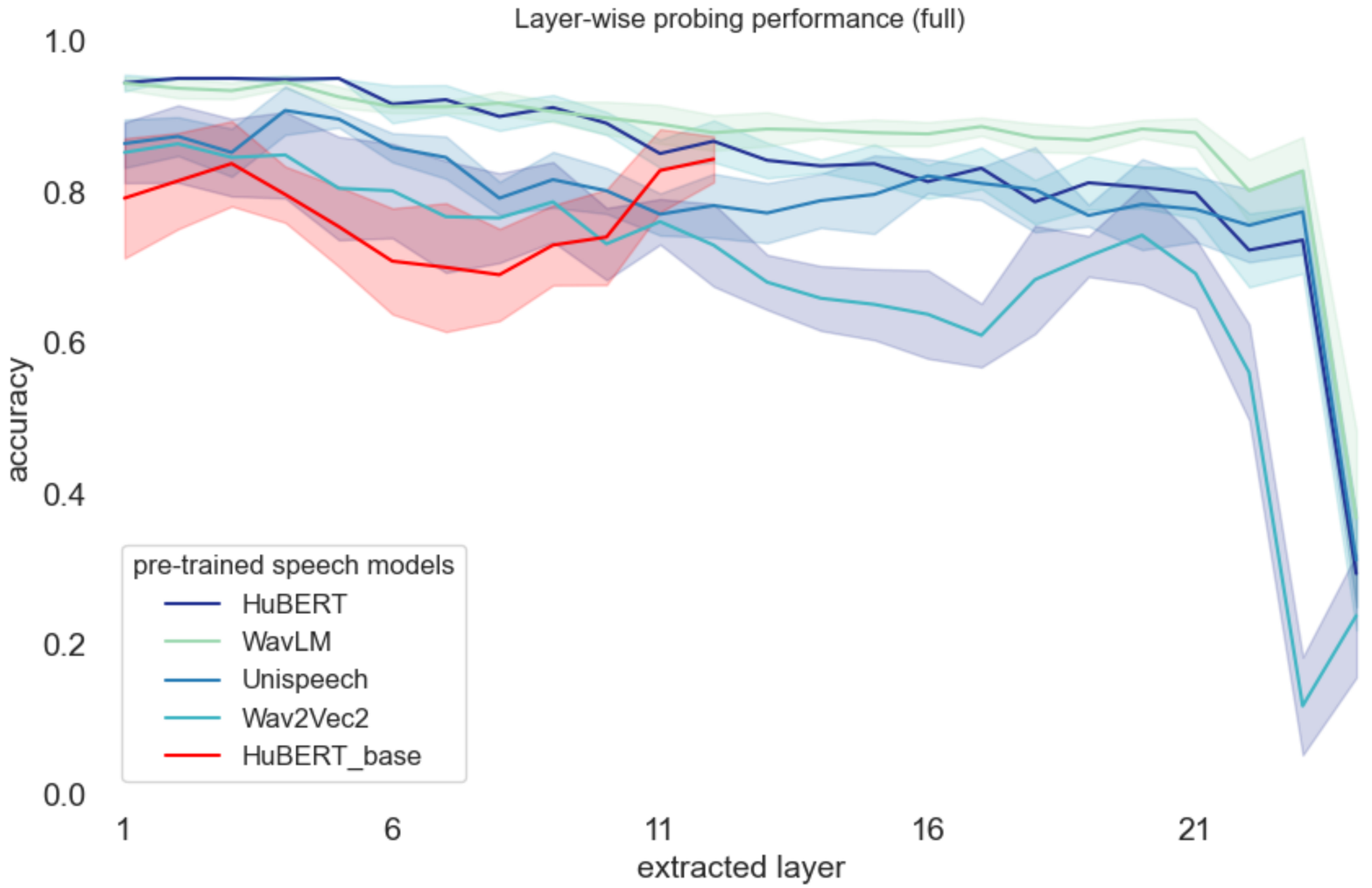
H2 - Layer-wise

- Initial layers - best performance
- last layer - worst performance
- Lowering performances in deeper layers
- No layer effects with background noise




Layer-wise probing performance (notes)
pre-trained speech models
HuBERT
WavLM
Unispeech
Wav2Vec2
HuBERT_base
accuracy
1.0
0.8
0.6
0.4
0.2
0.0
1
6
11
16
21
extracted layer


Layer-wise probing performance (noise)
pre-trained speech models
HuBERT
WavLM
Unispeech
Wav2Vec2
HuBERT_base
accuracy
1.0
0.8
0.6
0.4
0.2
0.0
1
6
11
16
21
extracted layer


Layer-wise probing performance (full)
1.0
0.8
0.6
0.4
0.2
0.0
accuracy
1
6
11
16
21
extracted layer
pre-trained speech models
HuBERT
WavLM
Unispeech
Wav2Vec2
HuBERT_base

Results

H3 - Background

- Speech PT > **BF**
- Bird / Audio PT < *BF*
- Bird PT > Background acc.

Model	Full	Background	BF
HuBERT Large	**0.95**	*0.12*	**7.76**
HuBERT Base	0.72	0.21	3.39
UniSpeech-SAT	0.87	0.19	4.64
Wav2vec 2.0 Large	0.86	0.14	**5.96**
WavLM Large	**0.94**	0.22	4.23
APC	0.75	0.14	5.44
Google perch	0.87	**0.69**	*1.26*
BirdNET 2.3	0.87	**0.63**	1.39
Audio-MAE AST	**0.95**	0.43	2.21
Vggish	*0.66*	0.43	1.52
MFCC	0.82	0.22	3.66
Chance	0.10	0.10	1.00

Results

H4 - Single notes

- MFCC > Notes acc.
- Audio-MAE, WavLM, wav2vec also pick up spectral information
- HuBERT and UniSpeech seem to rely on call structure

Model	Full	Background	Notes	BF
HuBERT Large	**0.95**	*0.12*	*0.13*	**7.76**
HuBERT Base	0.72	0.21	0.17	3.39
UniSpeech-SAT	0.87	0.19	0.14	4.64
Wav2vec 2.0 Large	0.86	0.14	0.44	**5.96**
WavLM Large	**0.94**	0.22	0.62	4.23
APC	0.75	0.14	0.14	5.44
Google perch	0.87	**0.69**	–	*1.26*
BirdNET 2.3	0.87	**0.63**	–	1.39
Audio-MAE AST	**0.95**	0.43	**0.82**	2.21
Vggish	*0.66*	0.43	–	1.52
MFCC	0.82	0.22	**0.94**	3.66
Chance	0.10	0.10	0.10	1.00

Limits and perspectives

- No retraining = biased comparison
  - Probing on custom test sets
- Adding models (but not any model)
- More tasks = more tests = more dataset
- Other species = delving into the phylogenetic hypothesis
- Pre-training a bioacoustics model
  (AVES - Hagiwara et al. 2022)

Thank you :)

Centre de Recherche en Psychologie et Neurosciences