Hearing Abilities of Foundation Models

Evaluation and Analysis of How and What Can Foundation Models Hear.

- Current Topics in Speech Technology

Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung

Yassir El Attar

Taha Rasouli

INTRODUCTION AND MOTIVATION

The motivations behind this study are:

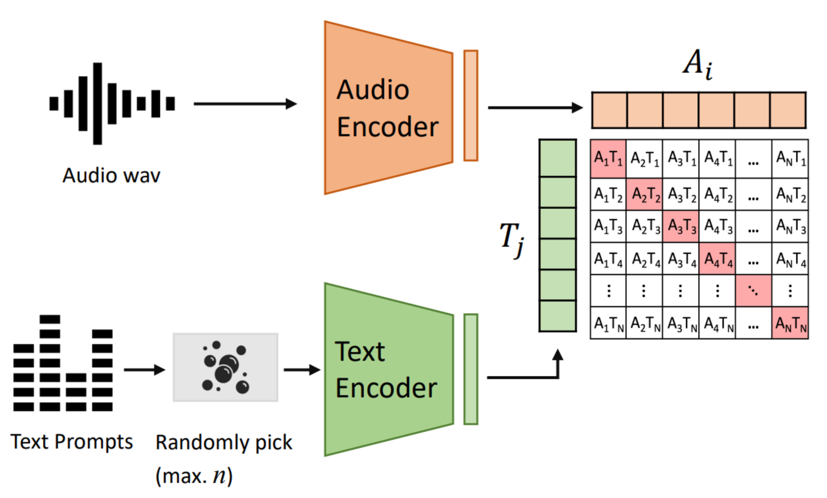
- Foundation models enhance ***hearing*** with self-supervised learning inspired by NLP.
- Hearing includes ***indentifying speech***, ***object attributes***, and ***sound event order***.
- Models combine ***audio*** and ***text*** for classification, retrieval, and generation.
- The study explores ***evolution***, ***challenges***, and ***innovations*** in ***auditory*** models.

The objectives of this study are:

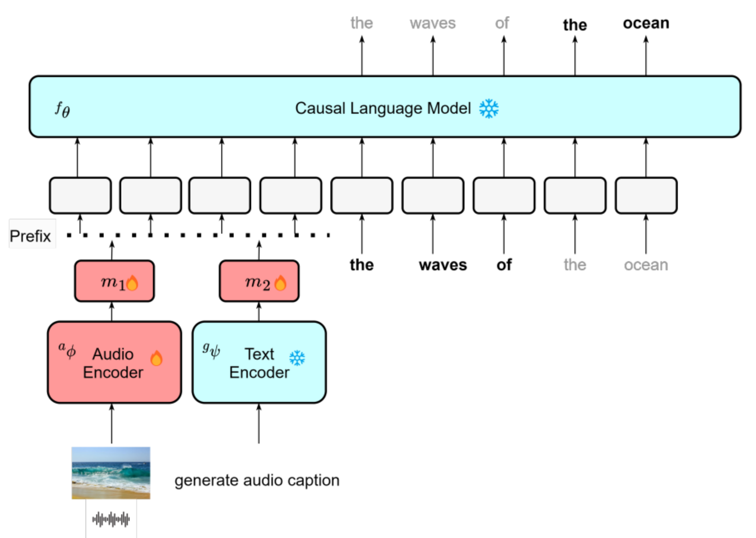
- To analyze the SoTA audio-language models' architecture and training mechanisms
- To analyze the hierarchical feature representations across speech and speaker tasks
- To analyze a large-scale evaluation of Speech FMs using (Dynamic-) SUPERB frameworks
- To study the extent to which models are able to hear.

MODELS ARCHITECTURES

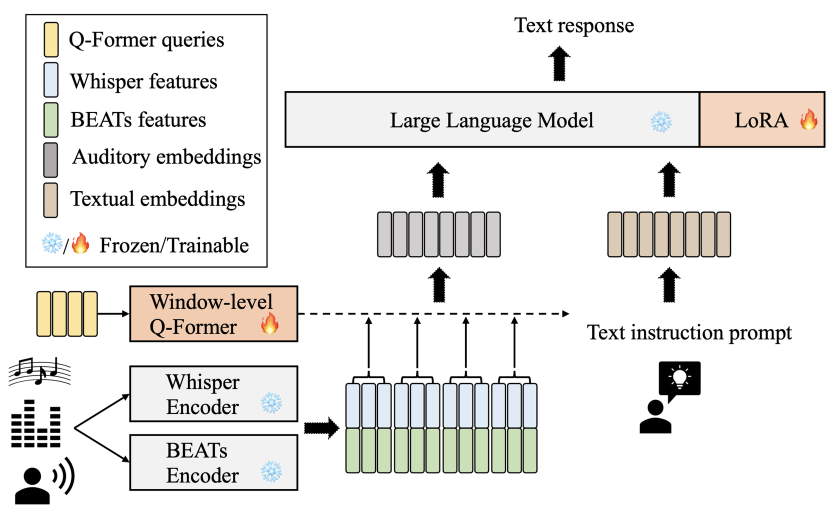
1. **CLAP's Architecture** ***(Elizalde et al., 2023)*** **:**



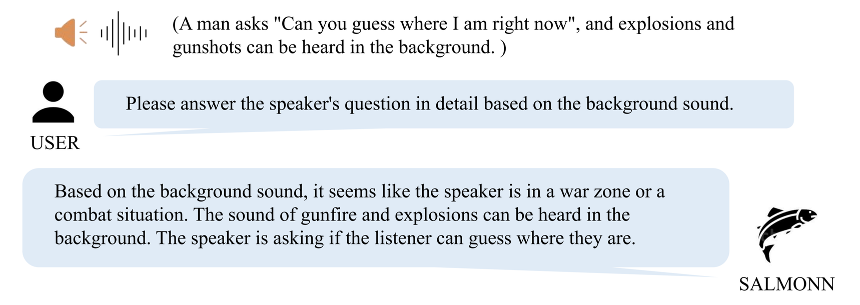
2. **Pengi's Architecture** ***(Singh et al., 2024)*****:**



3. **SALMONN's Architecture** ***(Tang et al., 2024)*****:**



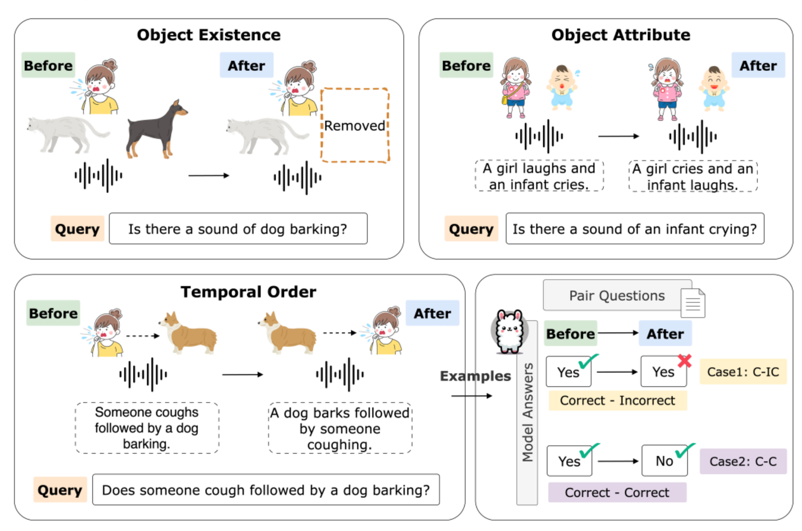
4. **Example of SALMONN** ***(Tang et al., 2024)*****:**



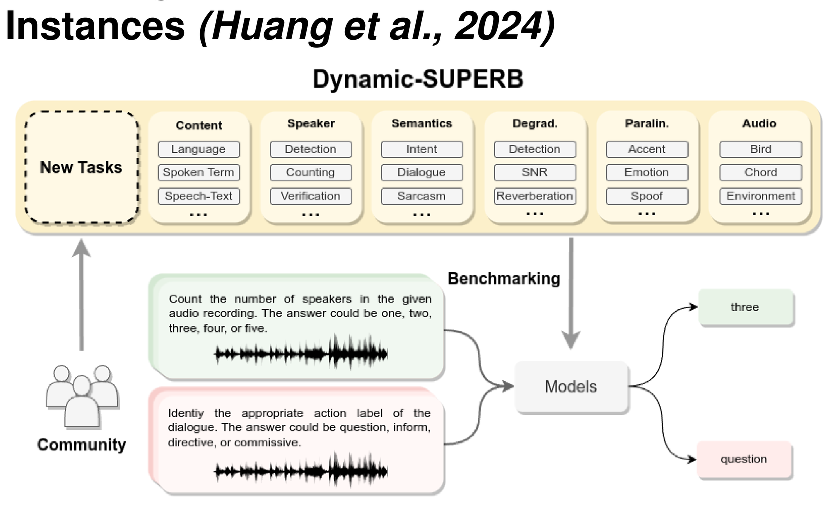
Task: "Speech-Audio Coreasoning"

SPEECH MODELS BENCHMARKS

Hallucinations in Audio Models: A Multi-Task Study ***(Kuan & Lee, 2024)***



Introducing the First Dynamic, Collaborative Benchmark for Speech Instruction Tuning: Covering 33 Tasks and 55 Evaluation Instances ***(Huang et al., 2024)***



Dynamic-SUPERB Results:

1. Accuracy on **seen** tasks:

Model	CNT	SPK	SEM	DEG	PRL	AUD
BERT-GSLM	66.3	49.1	47.2	68.2	52.7	-
Whisper	95.3	47.9	55.5	71.1	49.4	-
ImageBind-LLM	64.3	54.7	47.6	78.7	59.8	-
Whisper-LLM	77.6	91.7	55.7	91.0	66.3	-
ASR-ChatGPT	-	-	-	-	-	-
Random	49.9	40.2	41.0	45.9	67.1	-

2. Accuracy on **unseen** tasks:

Model	CNT	SPK	SEM	DEG	PRL	AUD
BERT-GSLM	0.0	32.8	5.3	41.6	12.6	0.0
Whisper	14.4	58.0	13.8	55.4	8.5	0.8
ImageBind-LLM	15.7	45.4	24.7	47.6	20.6	35.7
Whisper-LLM	8.7	60.6	20.9	59.0	6.6	15.9
ASR-ChatGPT	65.0	40.1	69.3	43.5	22.9	9.8
Random	11.8	50.2	33.1	43.1	21.0	23.4

Subset of the full results of a Large-Scale evaluation of 33 Models Across 15 tasks of SUPERB ***(Yang et al., 2024)*****:**

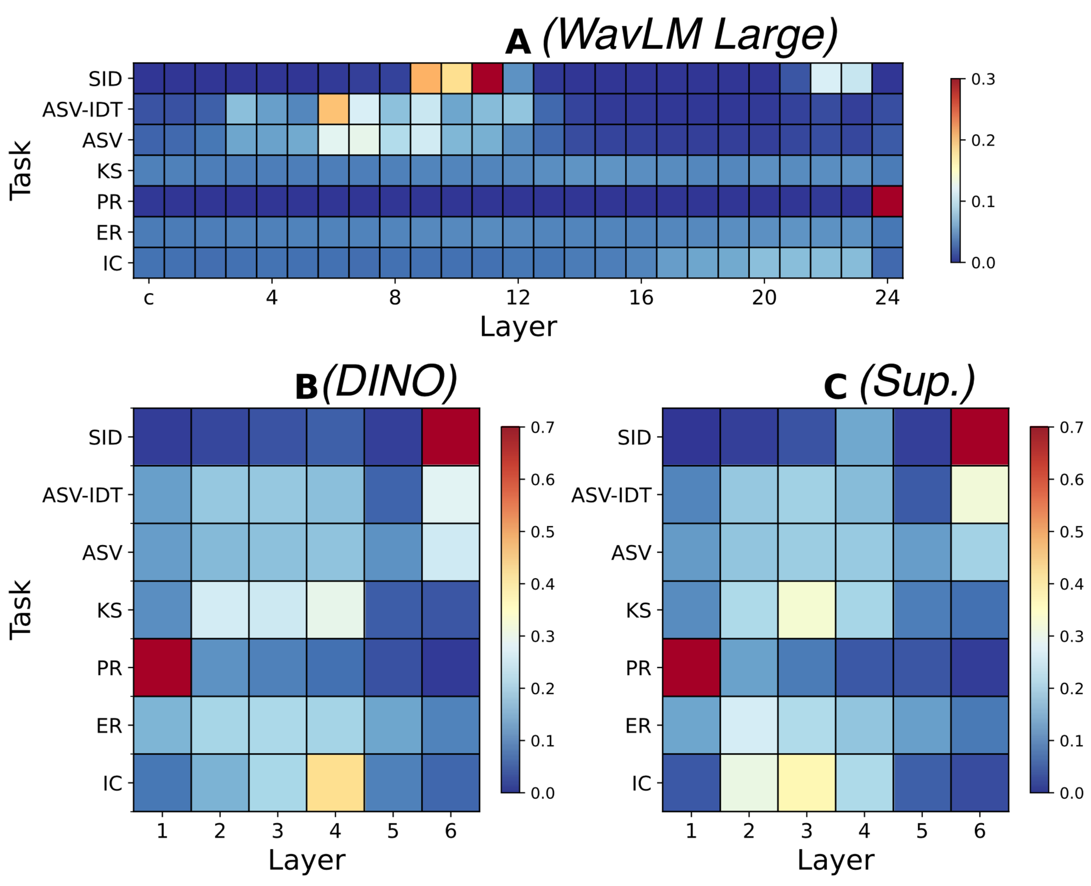
Model	Best Task (Metric: Score)	Worst Task (Metric: Score)	Notes
WavLM Large	ST (BLEU: 21.5)	SE (PESQ: 2.71)	Excels in semantics; weaker in generative tasks.
HuBERT Large	SID (ACC: 90.3)	SS (SI-SDRi: 9.2)	Strong speaker identification; weaker in source separation.
wav2vec 2.0 Large	KS (ACC: 97.6)	SE (PESQ: 2.62)	Strong keyword spotting; weaker in enhancement tasks.
Data2vec Large	VC (ASV-ACC: 99.5)	SE (PESQ: 2.6)	Excellent in voice conversion; weaker in enhancement tasks.
FBANK (Baseline)	OOD-ASR (WER: 53.6)	Most tasks	Provides a baseline for comparison.

$\rightarrow$ Models rely on **pattern recognition** rather than true **semantic comprehension**.

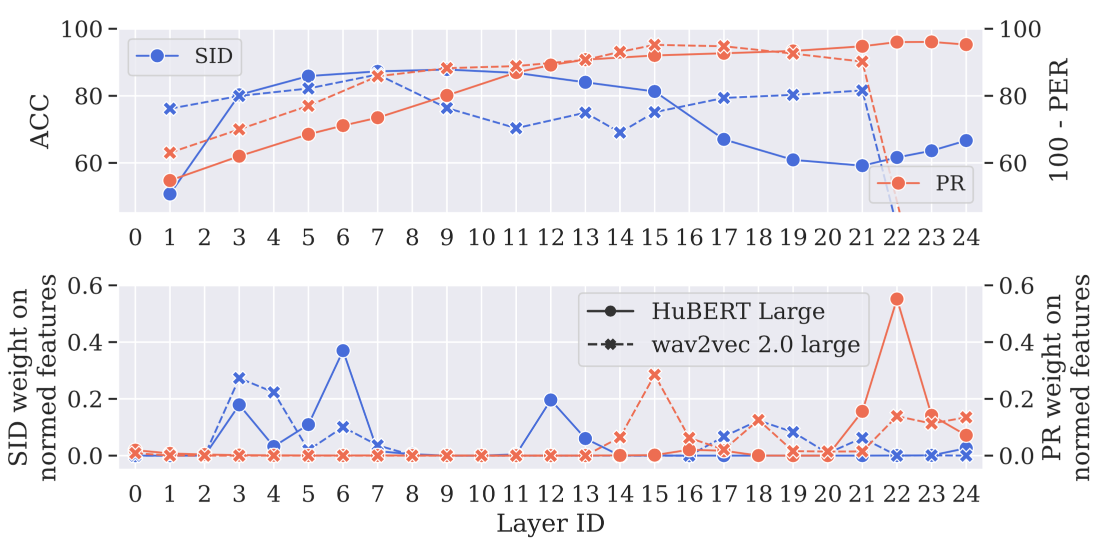
$\rightarrow$ Highlights the need for robust ***instruction understanding*** and ***generalization***.

LAYER-WISE ANALYSIS

Comparison of different layers' contribution to the models performance on Speaker vs. Speech tasks ***(Ashihara et al., 2024)*****:**



layer-wise comparison of two speech-SSL Models (HuBERT Large vs. wav2vec 2.0 Large on SID & PR tasks ***(Yang et al., 2024)***



$\Rightarrow$ Tasks favor specific layers (e.g., SE: lower, SID/ER: middle, PR: higher).

$\Rightarrow$ **Layer Weights** do not reliably reflect layer performance;

Layer-wise (*Lowest, Middle, Highest*) Benchmarking of wav2vec 2.0, HuBERT, and Data2vec Models Across 5 Tasks ***(Yang et al., 2024)***

Model	PR-PER	SID-ACC	ER-ACC	VC-MCD	VC-WER	VC-ASV	SE-PESQ	SE-STOI
FBANK	82.01	0.09	35.39	8.47	38.30	77.25	2.55	93.60
wav2vec 2.0 Base	5.74	75.18	63.43	7.50	10.50	98.00	2.55	93.90
wav2vec 2.0 Base - 1	37.10	58.42	59.48	7.79	14.08	95.50	2.51	93.57
wav2vec 2.0 Base - 5	15.72	71.48	60.81	7.36	11.75	97.50	2.47	93.69
wav2vec 2.0 Base - 11	24.42	54.60	57.97	7.32	19.15	97.25	2.40	93.25
wav2vec 2.0 Large	4.76	86.15	65.64	7.63	15.80	97.25	2.52	94.00
wav2vec 2.0 Large - 1	36.95	76.17	60.99	8.06	16.45	85.75	2.55	93.86
wav2vec 2.0 Large - 11	11.13	70.35	65.19	7.71	16.02	94.00	2.57	94.14
wav2vec 2.0 Large - 21	9.80	81.60	62.89	7.80	16.42	91.25	2.54	94.10
wav2vec 2.0 Large - 23	93.23	0.35	52.32	7.50	42.55	97.00	2.11	91.88
HuBERT Base	5.40	81.42	64.92	7.47	8.00	98.50	2.58	93.90
HuBERT Base - 1	42.43	58.16	59.28	7.93	17.00	90.75	2.51	93.68
HuBERT Base - 5	21.05	83.81	62.55	7.38	10.72	99.25	2.55	93.74
HuBERT Base - 11	6.10	69.78	62.72	7.23	11.05	99.00	2.48	93.55
HuBERT Large	3.54	90.33	67.62	7.22	9.00	99.25	2.64	94.20
HuBERT Large - 1	45.25	50.74	58.79	8.16	18.98	75.75	2.56	93.96
HuBERT Large - 11	13.04	86.84	67.47	7.24	10.17	99.50	2.59	93.97
HuBERT Large - 23	3.92	63.63	65.85	7.06	11.43	99.75	2.49	93.51
Data2vec Large	2.55	79.24	66.31	7.02	8.80	99.50	2.56	93.95
Data2vec Large - 1	36.28	52.54	60.69	7.77	15.87	95.00	2.55	93.72
Data2vec Large - 11	6.82	26.14	61.48	6.93	14.72	99.00	2.36	92.98
Data2vec Large - 23	2.62	15.67	60.28	6.87	21.73	99.50	2.36	92.46

CONCLUSION

- Models differ not only in their ***embeddings of auditory features***, but also across ***different tasks***.
- **Activation tuning** is important for harnessing the ***emerging abilities*** of speech models as seen in SALMONN.
- Research could benefit from training on ***multi-modal data*** and ***multi-tasking*** training.
- CLAP and ParaCLAP advanced **audio-text alignment** using **contrastive learning** and paraphrased text.
- Pengi's **generative approach** and SALMONN's **sound event encoder** addressed background audio issues highlighted by *Kuan (2024).*

MORE DETAILS &REFERENCES

`>> Course Title: "Current Topics in Speech Technology" - Lecturer: Prof. Dr. Thang Vu - Semester: Winter24/25 - IMS, University of Stuttgart, Germany.`