What and How Can Speech Models Hear?

Evaluation and Analysis of Speech SSL Models' Hearing Abilities.

Yassir EL ATTAR

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

INTRODUCTION & MOTIVATION

Speech self-supervised models have emerged as a central focus in contemporary research. The unprecedented achievements of NLP foundation models serve as an inspiration for advancing speech processing techniques. This study examines the auditory capabilities of speech models to evaluate their comparative advantages and limitations.

OBJECTIVES

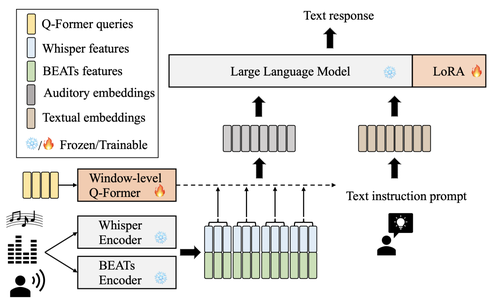
The objectives of this study are:

- To evaluate the auditory processing mechanisms in speech self-supervised models
- To analyze the hierarchical feature representation in speech and speaker models
- To analyze a large-scale evaluation of Speech FMs using (Dynamic-) SUPERB frameworks

SALMONN'S ABILITIES

SALMONN: Bridging Large Language Models and Audio Perception[2]

1. Model's architecture:



2. Activation Tuning unlocks the model's emergent abilities:

Method	ASR↓	En2Zh↑	AAC↑	PR↓	ER↑	MC↑	OSR↓	SV↑
w/o Activation	(2.1, 4.9, 9.1)	34.4	25.6 | 47.6	4.2	0.63	3.5, 22.1	20.7	0.93
w/ Activation	(2.1, 4.9, 10.0)	33.1	24.0 | 40.3	4.2	0.69	5.5, 21.8	23.0	0.94
Reference Value	(2.2, 5.1, 9.2)	38.9	25.0 | 48.5	3.1	0.81	6.1, 21.5	7.6	-

(a) Results of the level 1 tasks.

Method	En2De↑	En2Ja↑	KE↑	SQQA↑	SF↑	Story↑	SAC↑
w/o Activation	19.7	22.0	0.30	0.19 (0.29)	0.33 (0.77)	7.77 (0.00)	0.02 (0.04)
w/ Activation	18.6	22.7	0.32	0.41 (0.98)	0.41 (0.99)	82.57 (1.00)	0.50 (0.73)
Reference Value	16.5	15.6	0.31	0.77 (1.00)	0.46 (1.00)	-	-

(b) Results of the level 2 and level 3 tasks.

3. Example of "Speech Audio Coreasoning" task:

(A man asks "Can you guess where I am right now", and explosions and gunshots can be heard in the background.)

USER: Please answer the speaker's question in detail based on the background sound.

SALMONN: Based on the background sound, it seems like the speaker is in a war zone or a combat situation. The sound of gunfire and explosions can be heard in the background. The speaker is asking if the listener can guess where they are.

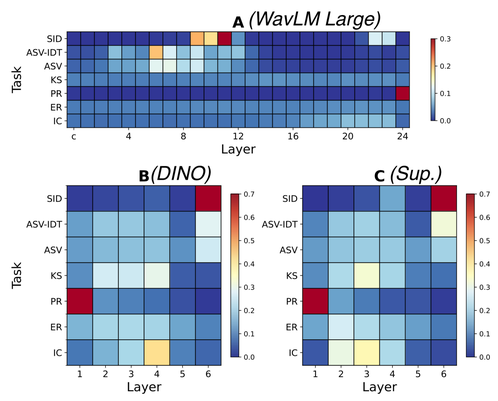
SPEAKER VS. SPEECH MODELS

Comparison of models' layers representations of different speech-auditory tasks[1]

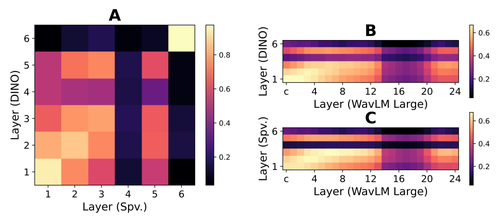
1. General performance of Speaker vs. Speech Models:

Upstream group	Upstream model	Model size	Pre-training data	SID Acc↑	ASV-IDT (ASV) EER↓	KS Acc↑	PR PER↓	ER Acc↑	IC Acc↑
Baseline	FBANK	-	-	0.1	13.7 (10.3)	8.3	96.9	29.0	4.5
Speech SSL model	HuBERT BASE [3]	94M	LS	66.6	4.9 (5.5)	96.6	5.7	65.1	98.2
	HuBERT LARGE [3]	315M	LL	89.4	4.3 (5.6)	94.9	4.3	66.0	98.4
	WavLM BASE [4]	94M	LS	62.2	5.1 (5.7)	96.3	5.6	64.9	98.5
	WavLM BASE+ [4]	94M	LL + GS + VP	72.3	3.9 (5.2)	97.5	4.4	68.4	**98.8**
	WavLM LARGE [4]	315M	LL + GS + VP	92.9	2.8 (5.1)	**98.0**	**4.3**	**69.3**	98.7
Speaker model	DINO	15M	VC	97.6	2.8 (8.0)	96.7	33.8	59.8	88.5
	Supervised	15M	VC	**98.5**	**2.2 (4.8)**	96.3	34.1	60.4	84.1

2. Weights' comparison of weighted sum across layers:



3. Layer-wise similarity across models:

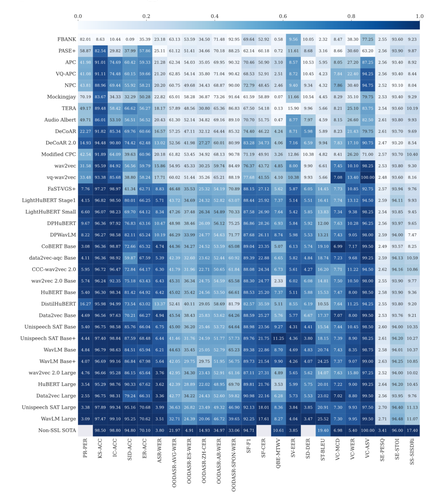


LARGE-SCALE SUPERB BENCHMARKING OF SPEECH MODELS

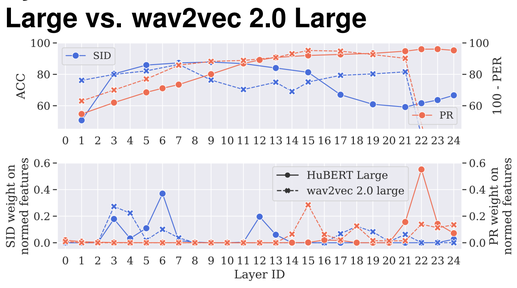
A large scale evaluation of 33 Speech Models across 15 SUPERB tasks[3]

- The study shows that using a pre-trained model with task-specific heads leads to good generalization across speech tasks.
- The proposed multi-tasking framework proves effective, allowing competitive performance without extensive model fine-tuning.
- A long-term maintained platform is provided for deterministic evaluation and collaborative benchmarking.
- In-depth analyses reveal how models process information and demonstrate the benefits of few-shot activation tuning.
- The study confirms the reliability of benchmark results through statistical significance and robust evaluation methods.

Heatmap comparison of 33 foundation models across 15 speech processing tasks



Performance Comparison of Layer Weights by Task and Model: SID and PR in HuBERT Large vs. wav2vec 2.0 Large

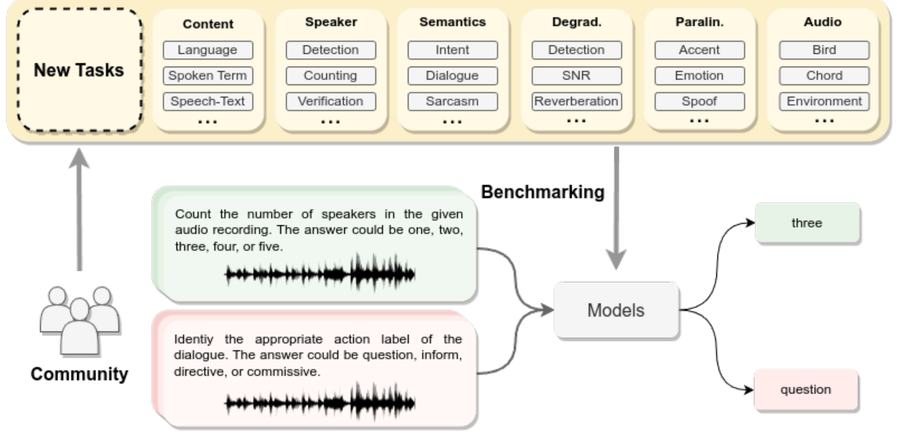


→ The **multi-layer** approach shows the models' ability to capture a **variety of auditory features**.

→ Models exhibit different **"hearing abilities"** for *Speech content*, *Semantics*, and *Speaker traits*.

NEW APPROACH: DYNAMIC SUPERB BENCHMARKING

Introducing the First Dynamic, Collaborative Benchmark for Speech Instruction Tuning: Covering 33 Tasks and 55 Evaluation Instances[4]



Performance results of multiple speech models on seen tasks *(with seen/unseen instructions)* and on unseen tasks:

1. Accuracy of models 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 of models 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

3. Accuracy of models on **seen** tasks with ***seen*** and ***unseen*** instructions:

Model	Instruction	CNT	SPK	SEM	DEG	PRL
BERT-GSLM	Seen	66.5	48.8	47.3	68.2	69.3
	Unseen	65.9	49.2	46.9	60.6	54.6
Whisper	Seen	95.4	48.2	55.4	71.0	65.8
	Unseen	95.0	47.3	55.7	50.5	49.6
ImageBind-LLM	Seen	64.1	55.1	47.3	78.6	75.8
	Unseen	64.8	54.4	48.2	62.8	59.7
Whisper-LLM	Seen	77.3	91.7	55.3	91.1	84.7
	Unseen	78.1	91.7	56.8	84.9	66.4

→ Models often rely on **pattern recognition**. rather than true **semantic comprehension**.

→ The need for a more robust ***instruction understanding*** and ***generalization***. capabilities.

REFERENCES

[1] T. Ashihara, M. Delcroix, T. Moriya, K. Matsuura, T. Asami, and Y. Ijima. What do self-supervised speech and speaker models learn? new findings from a cross model layer-wise analysis, 2024.

[2] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024.

[3] S. wen Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang, T. hsun Feng, P.-H. Chi, Y. Y. Lin, Y.-S. Chuang, T.-H. Huang, W.-C. Tseng, K. Lakhotia, S.-W. Li, A. Mohamed, S. Watanabe, and H. yi Lee. A large-scale evaluation of speech foundation models, 2024.

[4] C. yu Huang, K.-H. Lu, S.-H. Wang, C.-Y. Hsiao, C.-Y. Kuan, H. Wu, S. Arora, K.-W. Chang, J. Shi, Y. Peng, R. Sharma, S. Watanabe, B. Ramakrishnan, S. Shehata, and H. yi Lee. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech, 2024.

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.
- Research could benefit from training on ***multi-modal data***.

MORE DETAILS