What We Should Pay Attention When Going From Domain-specific to Universal

Jiatong Shi

Carnegie Mellon University, Langauge Technologies Institute

jiatongs@cs.cmu.edu

10/27/2024

Content

- Self Introduction
- Modeling: Domain-specific -> Universal -> Domain-specific
  - Thoughts from Multi-resolution HuBERT
- Next move

Myself

- 4th Yeah PhD at WAVLab, LTI, CMU; Advised by Dr. Shinji Watanabe
- Research Interest:
  - Speech representation and its application to diverse downstream tasks (e.g., ASR, TTS, ST, SLU, S2ST, SVS)
- Open-source Activity
  - ESPnet (main contributor);
  - S3PRL;
  - AudioGPT;
  - Fairseq;
  - ParallelWaveGAN;

Full CV

Myself (Cont'd)

- Benchmark and Challenges
  - Speech universal performance benchmark (SUPERB)
    - challenge at SLT2022
  - Multilingual speech universal performance benchmark (SUPERB)
    - challenge at ASRU2023
  - Simultaneous speech translation at IWSLT
    - shared task at IWSLT 2022, 2023, 2024
  - Discrete speech challenge
    - challenge at Interspeech2024
  - Singing voice conversion challenge (SVCC)
    - challenge at ASRU2023
  - Singing voice deepfake detection (SVDD)
    - challenge at SLT2024
  - Dynamic SUPERB (SUPERB focusing on instruction tuning)

Full CV

General trend in recent research community

- From domain-specific to universal
  - Many examples in CV and NLP
  - Similar in speech community:
    - From single-task to multi-task and then to unseen tasks (zero shot)
    - From task-specific representation to general speech representation

Speech Representation

- Speech Representation has been changing fast in recent years
  - Conventional spectral features
    - An elegant formulation of signal processing and psychoacoustics
  - Self-supervised learning representation
  - Neural codecs

My Examples

- Speech universal performance benchmark (SUPERB)
  - Targeting various speech tasks with a unified representation
- We go beyond that to more diverse cases
  - AV-SUPERB (SUPERB with audio-visual components)
  - ML-SUPERB (Multilingual SUPERB with multilingual focuses) -> I’m in charge of
  - Dynamic-SUPERB (Instruction-tuning scenario)

From Universal -> ?

- Given the trend of going universal,
  - most works tends to report a general number, while missing the domain-specific differences
  - what can we do to include the concept of domain-specification in a general modeling?

Motivation of Multi-resolution HuBERT

- Speech signal (single-channel):
  - a sequence of samples in a **high** sampling rate (e.g., 16k, 24k, 48k)

- The local stability
  - Information of speech is stable in **local sequences**
  - Feature extractions happens to **downsample** speech into a sequence of vectors to represent locational information.
    - Conventional: Linear spectrogram, Mel spectrogram, Mel cepstral coefficients (10-30ms per frame)
    - Learnable filters: Convolutional networks, which can be applied in conventional speech downstream tasks or self-supervised models such as wav2vec2, HuBERT (20ms per frame)

Motivation (Cont'd)

- While we generally want everything in a **uniform** representation, we tends to view things from **different perspective**

- In speech, we have

  Signal sample -> short-term frequency components -> phonemes -> syllable -> word -> sentence -> paragraph

Motivation (Cont'd)

- While we generally want everything in a **uniform** representation, we tends to view things from **different perspective**

- A straightforward method is to adding constraints:

  - Add constraints to limit the model's access to corresponding information

Jiatong Shi[1], *Yun Tang*[2], *Hirofumi Inaguma*[2], *Hongyu Gong*[2], *Juan Pino*[2], *Shinji Watanabe*[1]

[1]Language Technologies Institute, Carnegie Mellon University [2]Meta AI

Does resolution matters? (a pilot study in Interspeech 2023)

- Utilizing HuBERT with **different** resolutions

Table 1: *Configurations of the pre-trained HuBERT with multi-resolutions. The convolution (Conv.) module is represented in [(kernel-size, stride)* layer-number].*

ID	Res.(ms)	Param.	Conv. Module
A	20	94.7	(10,5)*1 + (3,2)*4 + (2,2)*2
B	40	95.2	(10,5)*1 + (3,2)*4 + (2,2)*3
C	100	97.3	(10,5)*2 + (3,2)*4 + (2,2)*2

Exploration on HuBERT with Multiple Resolutions

Jiatong Shi[1], Yun Tang[2], Hirofumi Inaguma[2], Hongyu Gong[2], Juan Pino[2], Shinji Watanabe[1]

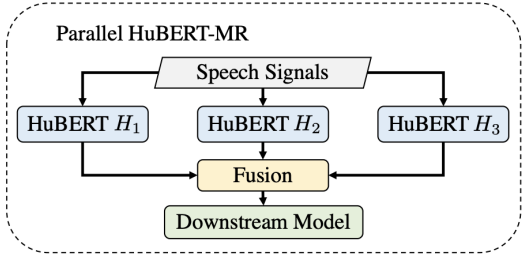
[1]Language Technologies Institute, Carnegie Mellon University [2]Meta AI

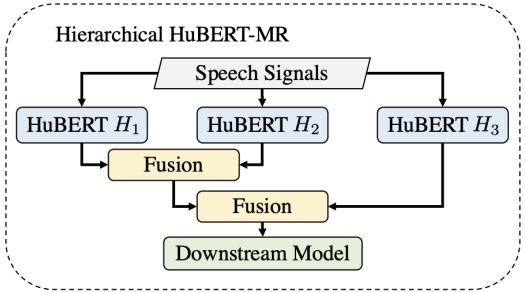
Does resolution matters? (a pilot study in Interspeech 2023)

- Utilizing HuBERT with **different** resolutions
- Use **simple** method to combine HuBERT features

Table 1: *Configurations of the pre-trained HuBERT with multi-resolutions. The convolution (Conv.) module is represented in [(kernel-size, stride)* layer-number].*

ID	Res.(ms)	Param.	Conv. Module
A	20	94.7	(10,5)*1 + (3,2)*4 + (2,2)*2
B	40	95.2	(10,5)*1 + (3,2)*4 + (2,2)*3
C	100	97.3	(10,5)*2 + (3,2)*4 + (2,2)*2





Does resolution matters? (a pilot study in Interspeech 2023)

- Get **better performance over (7/9)** than HuBERT in frozen setting in SUPERB

Model	Res.(ms)	PR(↓)	ASR(↓)	ER(↑)	IC(↑)	SID(↑)	SD(↓)	SV(↓)	SE(↑)	ST(↑)
HuBERT	20	5.41	6.42	**64.92**	98.34	81.42	5.88	5.11	**2.58**	15.53
wav2vec2	20	5.74	6.43	63.43	92.35	75.18	6.08	6.02	2.55	14.81
HuBERT-MR-P	(100,40,20)	**4.83**	**5.48**	63.76	**98.51**	**83.23**	**5.75**	**5.10**	2.55	**16.18**

Jiatong Shi[1], *Yun Tang*[2], *Hirofumi Inaguma*[2], *Hongyu Gong*[2], *Juan Pino*[2], *Shinji Watanabe*[1]

[1]Language Technologies Institute, Carnegie Mellon University [2]Meta AI

Does resolution matters? (a pilot study in Interspeech 2023)

- With fine-tuning, still **better**

Model	Res.(ms)	WER(↓)
HuBERT	20	7.73(3.81)
HuBERT	40	12.38(4.90)
HuBERT	100	98.37(97.87)
HuBERT-MR-P	(100,20)	6.99(3.70)
HuBERT-MR-P	(40,20)	7.13(3.75)
HuBERT-MR-P	(100,40,20)	6.53(3.61)
HuBERT-MR-H	(100,20)	6.59(3.59)
HuBERT-MR-H	(40,20)	7.01(3.71)
HuBERT-MR-H	(100,40,20)	**6.11**(**3.31**)

Does resolution matters? (a pilot study in Interspeech 2023)

- **Reduce** computation costs as well

Table 5: *Fine-tuning results on Librispeech-100h (comparison with other models). Results with language model rescoring are in brackets. * indicates the large model setting. The unlabeled column shows the number of hours used for SSL pre-training. HuBERT-MR-H is discussed in Sec 2.2. Noted that the HuBERT base model with 20ms is our baseline.*

Model	Res.(ms)	Unlabeled(h)	Param.(M)	MACs(G)	FLOPs(T)	WER(↓)
HuBERT	20	960	94.7	1669	3.34	7.73(3.81)
wav2vec2	20	960	95.0	1669	3.34	6.54(4.33)
wav2vec2*	20	60K	317.4	4326	8.66	5.90(3.45)
HuBERT*	20	60K	316.6	4324	8.66	**5.40(2.82)**
HuBERT-MR-H	(100,40,20)	960	298.4	3454	6.91	6.11(3.31)

Motivation of Multi-resolution HuBERT

- Data preparation challenge
- We found
  - HuBERT with multiple resolutions can have benefits to many of the tasks
  - However, pre-training multiple HuBERTs is time-consuming and cannot be applied to real-applications…

Motivation of Multi-resolution HuBERT

- We find some similar ideas of understanding sequencial information from different resolutions:
- Hourglass Transformer
  - https://arxiv.org/pdf/2110.13711.pdf
- Dynamic Token Poolking
  - https://arxiv.org/pdf/2211.09761.pdf

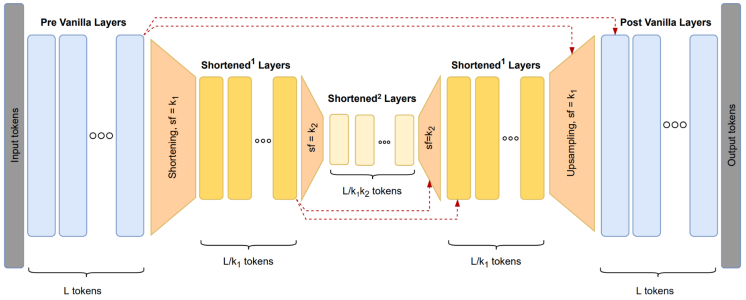
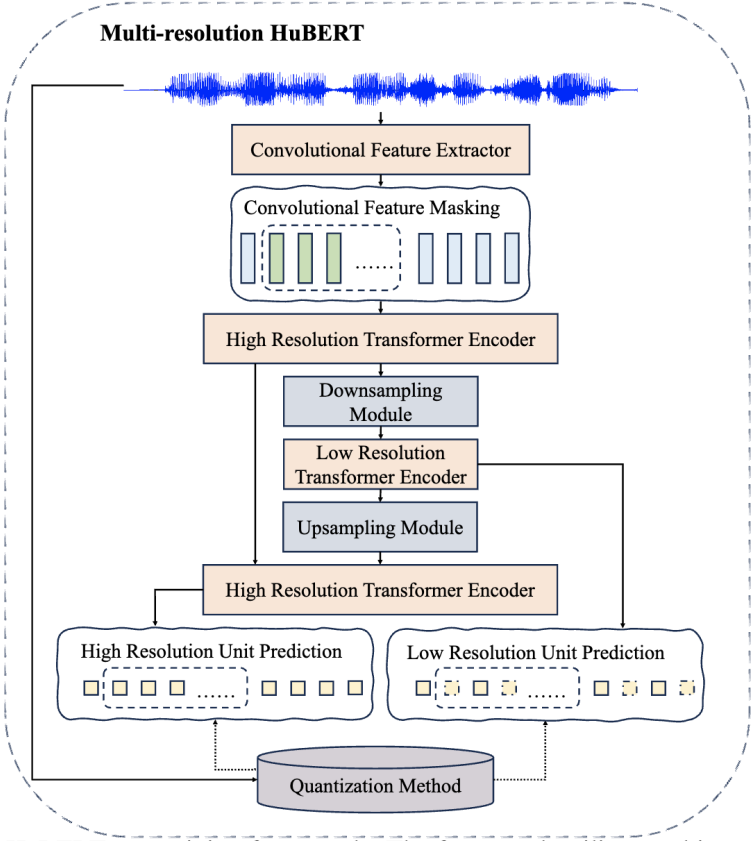


Figure 2: Hourglass - a high-level architecture overview. The arrows denote residual connections.

Multi-resolution HuBERT

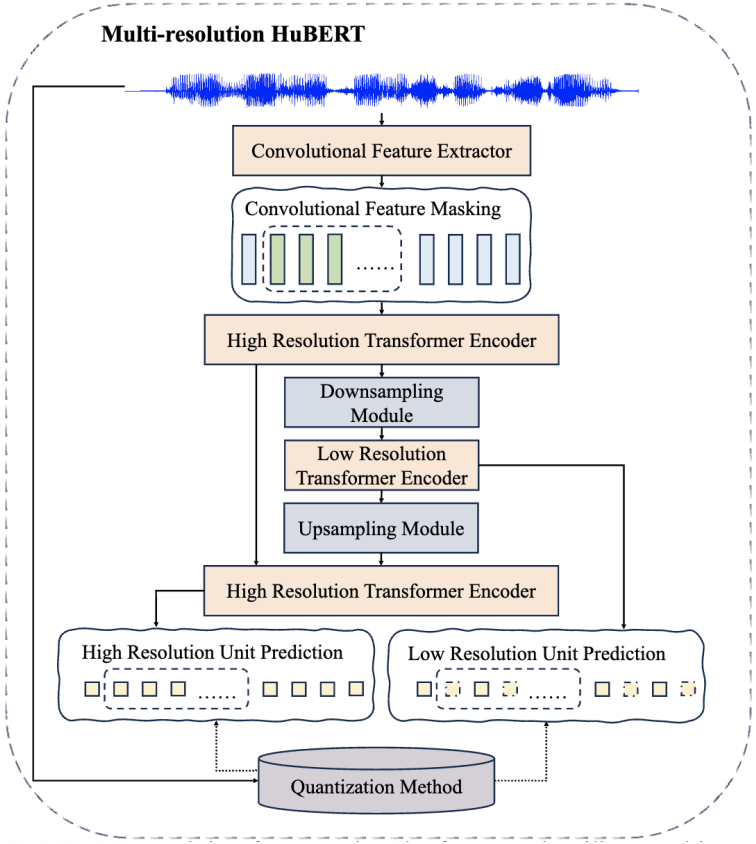
- By considering the motivations, we propose multi-resolution HuBERT
- Use the two-resolution case



Multi-resolution HuBERT

- By considering the motivations, we propose multi-resolution HuBERT

- Notable architectural changes
  - Sampling Module
  - Multi-resolution Objectives



Shu-wen Yang[1], Po-Han Chi[1*], Yung-Sung Chuang[1*], Cheng-I Jeff Lai[2*], Kushal Lakhotia[3*], Yist Y. Lin[1*], Andy T. Liu[1*], Jiatong Shi[4*], Xuankai Chang[6], Guan-Ting Lin[1], Tzu-Hsien Huang[1], Wei-Cheng Tseng[1], Ko-tik Lee[1], Da-Rong Liu[1], Zili Huang[4], Shuyan Dong[5†], Shang-Wen Li[5†], Shinji Watanabe[6], Abdelrahman Mohamed[5], Hung-yi Lee[1]

Experiments – SUPERB Evaluation

- SUPERB evaluation
  - Phone recognition (PR) – Phone error rate (PER)
  - Speech Recognition (ASR) – Word error rate (WER)
  - Intent Classification (IC) – Accuracy (ACC)
  - Keyword spotting (KS) – Accuracy (ACC)
  - Slot-filling (SF) – F1 and Character error rate (CER)
  - Speech Translation (ST) – BLEU
  - Speech enhancement (SE) – STOI and PESQ
  - Speech Separation (SS) – Si-SNR

- Experimental settings
  - Frozen Representation + Shallow downstream models

Experiments – SUPERB Evaluation

Model	Understanding							Enhancement		
	PR(↓)	ASR(↓)	IC(↑)	KS(↑)	SF-F1(↑)	SF-CER(↓)	ST(↑)	SE-STOI(↑)	SE-PESQ(↑)	SS(↑)
HuBERT-base	5.40	6.42	98.34	96.30	88.53	25.20	15.53	**0.94**	2.58	9.36
HuBERT-base^{+}	4.56	6.34	98.39	96.46	89.12	23.10	16.33	0.93	2.55	9.72
HuBERT-large	3.54	3.62	98.76	95.29	89.81	21.76	20.01	**0.94**	2.64	10.45
HuBERT-large*	3.59	**3.53**	98.73	97.70	89.88	22.51	20.02	**0.94**	2.65	10.61
mono-base	4.16	5.76	98.68	96.49	88.96	23.59	16.94	**0.94**	2.55	9.92
mono-large	**3.15**	3.78	**98.76**	**97.76**	**90.57**	**20.60**	**21.05**	**0.94**	**2.67**	**10.97**

- The proposed multi-resolution HuBERT almost **excels over all the tasks except for ASR** (which might be because CTC cannot handle well the repeat upsampling for low-resolution for weighted sum.

Experiments – SUPERB evaluation

Model	Understanding	Enhancement	General
HuBERT-base	861.2	98.20	670.4
HuBERT-base^{+}	876.9	150.2	695.2
HuBERT-large	932.6	456.0	813.4
HuBERT-large*	936.2	501.5	827.5
mono-base	885.8	195.0	708.7
mono-large	**949.7**	**609.5**	**864.6**

When, use SUPERB score, our proposed method has shown significant improvements in both understanding and enhancement tasks.

Experiments – SUPERB Evaluation

	PR	ASR	IC	KS	SF	ST	SE	SS
15	0.0773	0.0041	0.1588	0.0924	0.0051	0.0169	0.0384	0.0147
14	0.4173	0.0661	0.1228	0.0911	0.0191	0.2904	0.0335	0.0017
13	0.1124	0.3551	0.1085	0.0850	0.2418	0.2637	0.0296	0.0040
12	0.0490	0.0870	0.1039	0.0901	0.0489	0.0364	0.0372	0.0063
11	0.0040	0.0066	0.0796	0.0867	0.0061	0.0027	0.0435	0.0001
10	0.1806	0.0052	0.0613	0.0440	0.0070	0.1527	0.0123	0.0316
9	0.0592	0.2029	0.0928	0.0763	0.3045	0.0365	0.0210	0.0015
8	0.0024	0.2189	0.0760	0.0923	0.2165	0.0072	0.0196	0.0043
7	0.0019	0.0179	0.0415	0.0896	0.0159	0.0048	0.0240	0.0052
6	0.0001	0.0038	0.0278	0.0577	0.0050	0.0047	0.0338	0.0026
5	0.0414	0.0037	0.0245	0.0424	0.0088	0.0258	0.0111	0.0962
4	0.0251	0.0041	0.0246	0.0381	0.0106	0.0219	0.0270	0.0979
3	0.0096	0.0055	0.0260	0.0395	0.0175	0.0246	0.1269	0.1171
2	0.0066	0.0076	0.0257	0.0353	0.0323	0.0424	0.2116	0.1893
1	0.0130	0.0116	0.0262	0.0397	0.0611	0.0693	0.3304	0.4276

(a) **mono-base** (*base* MR-HuBERT)

	PR	ASR	IC	KS	SF	ST	SE	SS
13	0.0000	0.0007	0.0993	0.1359	0.0013	0.0073	0.0528	0.0001
12	0.4250	0.0194	0.1121	0.1057	0.0068	0.1503	0.0332	0.0270
11	0.3487	0.3853	0.1779	0.1343	0.2315	0.4271	0.0314	0.0095
10	0.1315	0.2937	0.2756	0.1542	0.3878	0.1887	0.0284	0.0058
9	0.0660	0.2001	0.1725	0.1181	0.1901	0.0460	0.0274	0.0066
8	0.0109	0.0645	0.0635	0.1091	0.0751	0.0096	0.0324	0.0190
7	0.0001	0.0044	0.0149	0.0391	0.0053	0.0071	0.0358	0.0375
6	0.0000	0.0011	0.0132	0.0309	0.0016	0.0089	0.0505	0.0728
5	0.0000	0.0011	0.0107	0.0336	0.0021	0.0121	0.0751	0.0776
4	0.0000	0.0025	0.0139	0.0276	0.0052	0.0146	0.1302	0.1308
3	0.0008	0.0044	0.0180	0.0381	0.0143	0.0183	0.1326	0.1003
2	0.0074	0.0087	0.0140	0.0253	0.0325	0.0349	0.1381	0.1505
1	0.0096	0.0141	0.0143	0.0481	0.0464	0.0750	0.2322	0.3624

(b) HuBERT-base

Mono-base (1-5 layer high resolution, 6-10 layer low resolution, 11-15 layer high resolution)

- Low resolution significantly contributes to PR, ASR, and SF
- Also there are some implications of inexplicit speech disentanglement in SE and SS (the low resolution layer weights are minimized)

Experiments – SUPERB Evaluation

	PR	ASR	IC	KS	SF	ST	SE	SS
27	0.0143	0.0187	0.1490	0.0402	0.0198	0.0099	0.0402	0.0038
26	0.0132	0.0203	0.0910	0.0477	0.0104	0.0008	0.0185	0.0000
25	0.0706	0.0184	0.0689	0.0463	0.0034	0.0224	0.0178	0.0000
24	0.0553	0.0525	0.0804	0.0441	0.0100	0.0480	0.0188	0.0000
23	0.0000	0.0447	0.0423	0.0406	0.0165	0.1074	0.0148	0.0000
22	0.0000	0.0723	0.0274	0.0360	0.0597	0.0167	0.0151	0.0000
21	0.0000	0.1221	0.0307	0.0345	0.0843	0.0047	0.0164	0.0000
20	0.0001	0.0425	0.0146	0.0328	0.0234	0.0019	0.0169	0.0000
19	0.0001	0.0175	0.0100	0.0324	0.0049	0.0037	0.0147	0.0000
18	0.4774	0.1652	0.0286	0.0348	0.1541	0.3636	0.0492	0.0405
17	0.0000	0.0186	0.1090	0.0471	0.0051	0.0038	0.0196	0.0000
16	0.0000	0.0325	0.0738	0.0441	0.0258	0.0104	0.0201	0.0000
15	0.0000	0.1085	0.0827	0.0424	0.1188	0.0008	0.0221	0.0000
14	0.0000	0.0888	0.0317	0.0399	0.0594	0.0004	0.0189	0.0000
13	0.0000	0.0007	0.0114	0.0341	0.0001	0.0003	0.0218	0.0000
12	0.0000	0.0003	0.0053	0.0305	0.0001	0.0004	0.0180	0.0000
11	0.0000	0.0007	0.0055	0.0294	0.0003	0.0007	0.0131	0.0000
10	0.0128	0.0574	0.0024	0.0304	0.0962	0.1766	0.0251	0.0724
9	0.3556	0.0912	0.0024	0.0307	0.2845	0.2194	0.0335	0.6435
8	0.0000	0.0015	0.0166	0.0438	0.0007	0.0003	0.0628	0.0000
7	0.0000	0.0016	0.0147	0.0391	0.0008	0.0004	0.0613	0.0000
6	0.0000	0.0011	0.0122	0.0350	0.0005	0.0005	0.0595	0.0117
5	0.0000	0.0012	0.0125	0.0338	0.0006	0.0005	0.0902	0.0400
4	0.0000	0.0022	0.0169	0.0337	0.0014	0.0007	0.0673	0.0161
3	0.0000	0.0035	0.0161	0.0332	0.0026	0.0009	0.0948	0.0923
2	0.0000	0.0068	0.0185	0.0320	0.0064	0.0015	0.0760	0.0515
1	0.0006	0.0094	0.0252	0.0316	0.0103	0.0033	0.0736	0.0281

(a) **mono-large** (*large* MR-HuBERT)

	PR	ASR	IC	KS	SF	ST	SE	SS
25	0.9777	0.8306	0.0297	0.0466	0.7935	0.9266	0.0152	0.7418
24	0.0000	0.0035	0.0654	0.0966	0.0049	0.0003	0.0097	0.0000
23	0.0160	0.0110	0.0603	0.0617	0.0100	0.0030	0.0129	0.0000
22	0.0050	0.0193	0.0719	0.0632	0.0101	0.0275	0.0147	0.0000
21	0.0000	0.0144	0.0725	0.0561	0.0180	0.0283	0.0195	0.0000
20	0.0000	0.0070	0.0672	0.0445	0.0201	0.0035	0.0169	0.0000
19	0.0000	0.0101	0.0624	0.0361	0.0211	0.0011	0.0154	0.0000
18	0.0000	0.0275	0.0549	0.0305	0.0347	0.0003	0.0150	0.0000
17	0.0000	0.0206	0.0468	0.0281	0.0214	0.0002	0.0156	0.0000
16	0.0000	0.0040	0.0355	0.0303	0.0046	0.0002	0.0175	0.0000
15	0.0000	0.0016	0.0317	0.0327	0.0028	0.0002	0.0200	0.0000
14	0.0000	0.0339	0.0296	0.0324	0.0216	0.0002	0.0214	0.0000
13	0.0000	0.0029	0.0291	0.0357	0.0044	0.0002	0.0226	0.0000
12	0.0000	0.0004	0.0283	0.0388	0.0012	0.0002	0.0260	0.0000
11	0.0000	0.0001	0.0279	0.0376	0.0004	0.0002	0.0312	0.0002
10	0.0000	0.0000	0.0299	0.0344	0.0004	0.0002	0.0383	0.0042
9	0.0000	0.0000	0.0294	0.0284	0.0003	0.0003	0.0398	0.0080
8	0.0000	0.0001	0.0286	0.0281	0.0006	0.0003	0.0413	0.0095
7	0.0000	0.0001	0.0271	0.0295	0.0008	0.0003	0.0826	0.0250
6	0.0000	0.0002	0.0287	0.0333	0.0011	0.0004	0.1049	0.0141
5	0.0000	0.0005	0.0276	0.0318	0.0017	0.0004	0.0739	0.0312
4	0.0000	0.0008	0.0287	0.0358	0.0024	0.0006	0.0958	0.0196
3	0.0001	0.0015	0.0279	0.0327	0.0034	0.0007	0.0786	0.0594
2	0.0003	0.0035	0.0275	0.0321	0.0062	0.0015	0.0575	0.0524
1	0.0008	0.0062	0.0314	0.0431	0.0142	0.0035	0.1136	0.0346

(b) `HuBERT-large`

Mono-large (1-9 layer high resolution, 10-18 layer low resolution, 19-27 layer high resolution)

- Mono-large
  - High resolution emphasis: enhancement tasks SE, SS
  - Low resolution emphasis: understanding tasks PR, ASR, SF, and ST
  - Equitable emphasis: IC and KS for sentence classification

Experiments – SUPERB Evaluation

	PR	ASR	IC	KS	SF	ST	SE	SS
27	0.0143	0.0187	0.1490	0.0402	0.0198	0.0099	0.0402	0.0038
26	0.0132	0.0203	0.0910	0.0477	0.0104	0.0008	0.0185	0.0000
25	0.0706	0.0184	0.0689	0.0463	0.0034	0.0224	0.0178	0.0000
24	0.0553	0.0525	0.0804	0.0441	0.0100	0.0480	0.0188	0.0000
23	0.0000	0.0447	0.0423	0.0406	0.0165	0.1074	0.0148	0.0000
22	0.0000	0.0723	0.0274	0.0360	0.0597	0.0167	0.0151	0.0000
21	0.0000	0.1221	0.0307	0.0345	0.0843	0.0047	0.0164	0.0000
20	0.0001	0.0425	0.0146	0.0328	0.0234	0.0019	0.0169	0.0000
19	0.0001	0.0175	0.0100	0.0324	0.0049	0.0037	0.0147	0.0000
18	0.4774	0.1652	0.0286	0.0348	0.1541	0.3636	0.0492	0.0405
17	0.0000	0.0186	0.1090	0.0471	0.0051	0.0038	0.0196	0.0000
16	0.0000	0.0325	0.0738	0.0441	0.0258	0.0104	0.0201	0.0000
15	0.0000	0.1085	0.0827	0.0424	0.1188	0.0008	0.0221	0.0000
14	0.0000	0.0888	0.0317	0.0399	0.0594	0.0004	0.0189	0.0000
13	0.0000	0.0007	0.0114	0.0341	0.0001	0.0003	0.0218	0.0000
12	0.0000	0.0003	0.0053	0.0305	0.0001	0.0004	0.0180	0.0000
11	0.0000	0.0007	0.0055	0.0294	0.0003	0.0007	0.0131	0.0000
10	0.0128	0.0574	0.0024	0.0304	0.0962	0.1766	0.0251	0.0724
9	0.3556	0.0912	0.0024	0.0307	0.2845	0.2194	0.0335	0.6435
8	0.0000	0.0015	0.0166	0.0438	0.0007	0.0003	0.0628	0.0000
7	0.0000	0.0016	0.0147	0.0391	0.0008	0.0004	0.0613	0.0000
6	0.0000	0.0011	0.0122	0.0350	0.0005	0.0005	0.0595	0.0117
5	0.0000	0.0012	0.0125	0.0338	0.0006	0.0005	0.0902	0.0400
4	0.0000	0.0022	0.0169	0.0337	0.0014	0.0007	0.0673	0.0161
3	0.0000	0.0035	0.0161	0.0332	0.0026	0.0009	0.0948	0.0923
2	0.0000	0.0068	0.0185	0.0320	0.0064	0.0015	0.0760	0.0515
1	0.0006	0.0094	0.0252	0.0316	0.0103	0.0033	0.0736	0.0281

(a) **mono-large** (*large* MR-HuBERT)

	PR	ASR	IC	KS	SF	ST	SE	SS
25	0.9777	0.8306	0.0297	0.0466	0.7935	0.9266	0.0152	0.7418
24	0.0000	0.0035	0.0654	0.0966	0.0049	0.0003	0.0097	0.0000
23	0.0160	0.0110	0.0603	0.0617	0.0100	0.0030	0.0129	0.0000
22	0.0050	0.0193	0.0719	0.0632	0.0101	0.0275	0.0147	0.0000
21	0.0000	0.0144	0.0725	0.0561	0.0180	0.0283	0.0195	0.0000
20	0.0000	0.0070	0.0672	0.0445	0.0201	0.0035	0.0169	0.0000
19	0.0000	0.0101	0.0624	0.0361	0.0211	0.0011	0.0154	0.0000
18	0.0000	0.0275	0.0549	0.0305	0.0347	0.0003	0.0150	0.0000
17	0.0000	0.0206	0.0468	0.0281	0.0214	0.0002	0.0156	0.0000
16	0.0000	0.0040	0.0355	0.0303	0.0046	0.0002	0.0175	0.0000
15	0.0000	0.0016	0.0317	0.0327	0.0028	0.0002	0.0200	0.0000
14	0.0000	0.0339	0.0296	0.0324	0.0216	0.0002	0.0214	0.0000
13	0.0000	0.0029	0.0291	0.0357	0.0044	0.0002	0.0226	0.0000
12	0.0000	0.0004	0.0283	0.0388	0.0012	0.0002	0.0260	0.0000
11	0.0000	0.0001	0.0279	0.0376	0.0004	0.0002	0.0312	0.0002
10	0.0000	0.0000	0.0299	0.0344	0.0004	0.0002	0.0383	0.0042
9	0.0000	0.0000	0.0294	0.0284	0.0003	0.0003	0.0398	0.0080
8	0.0000	0.0001	0.0286	0.0281	0.0006	0.0003	0.0413	0.0095
7	0.0000	0.0001	0.0271	0.0295	0.0008	0.0003	0.0826	0.0250
6	0.0000	0.0002	0.0287	0.0333	0.0011	0.0004	0.1049	0.0141
5	0.0000	0.0005	0.0276	0.0318	0.0017	0.0004	0.0739	0.0312
4	0.0000	0.0008	0.0287	0.0358	0.0024	0.0006	0.0958	0.0196
3	0.0001	0.0015	0.0279	0.0327	0.0034	0.0007	0.0786	0.0594
2	0.0003	0.0035	0.0275	0.0321	0.0062	0.0015	0.0575	0.0524
1	0.0008	0.0062	0.0314	0.0431	0.0142	0.0035	0.1136	0.0346

(b) HuBERT-large

- While the layer weight distribution presents a quite good alignment with the tasks, it also indicated the implicit speech disentanglement (or focus differences in the multi-resolution HuBERT)
  - This model might be able to extend to be an interpretable framework for understanding self-supervised learning in speech

MULTI-RESOLUTION HUBERT: MULTI-RESOLUTION SPEECH SELF-SUPERVISED LEARNING WITH MASKED UNIT PREDICTION

Jiatong Shi[1]*, **Hirofumi Inaguma**[2], **Xutai Ma**[2], **Ilia Kulikov**[2], **Anna Sun**[2]
[1] Language Technologies Institute, Carnegie Mellon University; [2] Meta AI
jiatongs@cs.cmu.edu
{hirofumii, xutaima, kulikov, annaysun}@meta.com

9 ACKNOWLEDGEMENT

We extend our heartfelt gratitude to Juan Pino, Paden Tomasello, Changhan Wang, Andy Chung, Ning Dong, Hongyu Gong, and Maha Elbayad for their invaluable advice and unwavering support throughout this project. Their insights and expertise have been indispensable to this work.

Special recognition is owed to Yun Tang and Shinji Watanabe. Their contributions, particularly in the formative stages of our research, have been instrumental. Their guidance in shaping our initial research idea has set a strong foundation for the entirety of this work.

Related works on the concept of going domain-specific

- Tejes Srivastava+, **Jiatong Shi**, William Chen, and Shinji Watanabe. ”EFFUSE: Efficient Self-Supervised Feature Fusion for E2E ASR in Multilingual and Low Resource Scenarios”. (Interspeech 2024)
- Yuxun Tang+, Yuning Wu+, **Jiatong Shi**, and Qin Jin. ”SingOMD: Singing Oriented Multi-resolution Discrete Representation Construction from Speech Models”. (Interspeech 2024)
- Yuning Wu+, Chunlei Zhang, **Jiatong Shi**, Yuxun Tang+, and Qin Jin. ”TokSing: Singing Voice Synthesis based on Discrete Tokens”. (Interspeech 2024)
- **Jiatong Shi**, Chan-Jan Hsu, Holam Chung, Dongji Gao, Paola Garcia, Shinji Watanabe, Ann Lee, and Hung-yi Lee. ”Bridging Speech and Text Pre-trained Models with Unsupervised ASR” (ICASSP 2023)
- **Jiatong Shi**, Yun Tang, Ann Lee, Hirofumi Inaguma, Changhan Wang, Juan Pino, and Shinji Watanabe. ”Enhancing Speech-to-Speech Translation with Multiple TTS Targets” (ICASSP 2023)

- + student mentored

Next move

- We've drawn our attention to discrete space, given its compatibility to LLM integration.
  - Modeling:
    - **Jiatong Shi**, Xutai Ma, Hirofumi Inaguma, Anna Sun, and Shinji Watanabe. "MMM: Multi-Layer Multi-Residual Multi-Stream Discrete Speech Representation from Self-supervised Learning Model". (Interspeech 2024)
    - **Jiatong Shi**, Chunlei Zhang, Jinchuan Tian, Junrui Ni, Hao Zhang, Shinji Watanabe, and Dong Yu. "Balancing Speech Understanding and Generation Using Continual Pre-training for Codec-based Speech LLM" (To be on ArXiv soon)
  - All-in-one infrastructure:
    - **Jiatong Shi**, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu+, Yuxun Tang+, Massa Baali, Dareen Alharthi, Dong Zhang, Ruifan Deng, Tejes Srivastava+, Haibin Wu, Alexander Liu, Bhiksha Raj, Qin Jin, Ruihua Song, and Shinji Watanabe. "ESPnet-Codec: Comprehensive Training and Evaluation of Neural Codecs for Audio, Music, and Speech". (Incoming SLT 2024)
  - Challenge
    - Xuankai Chang, **Jiatong Shi**, Jinchuan Tian, Yuning Wu+, Yuxun Tang+, Yihan Wu, Shinji Watanabe, Yossi Adi, Xie Chen, and Qin Jin. 2024. "The Interspeech 2024 Challenge on Speech Processing Using Discrete Units". (Interspeech 2024)
  - Evaluation:
    - VERSA in ESPnet-codec paper and its extension to be released soon

Thank you for listening!