WavLM: Large Scale Unsupervised Pre-Training for Various Speech Tasks

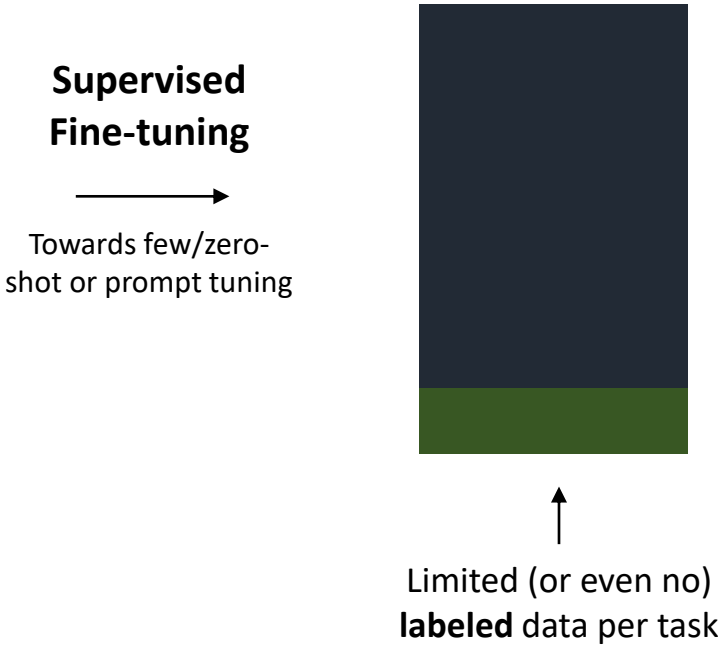
Presenter: Yu Wu

2022-02-28

Microsoft Research Asia

New AI Paradigm: (Self-supervised) Pre-trained Models

Examples: BERT, GPT-3



Downstream AI Tasks:

- Language understanding and generation
- Image classification, Object detection
- Speech recognition

Pre-trained Model Task-specific Model (Layer)

An effective way to leverage unlabeled data
Excellent Results on limited resource tasks

Focus of Previous Work

Previous Speech Pre-train Models

Content Tasks:

Speech Recognition, Speech Translation

Goal of WavLM

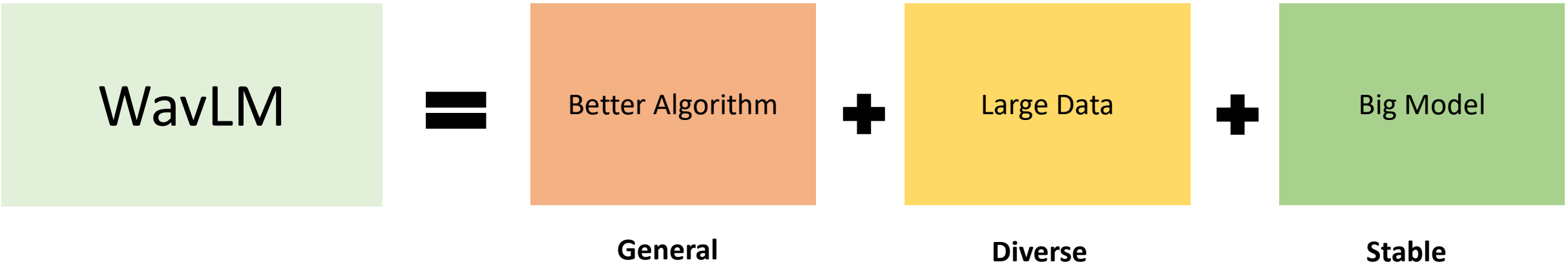
WavLM

Denoising Tasks
Speech Enhancement, Speech Separation

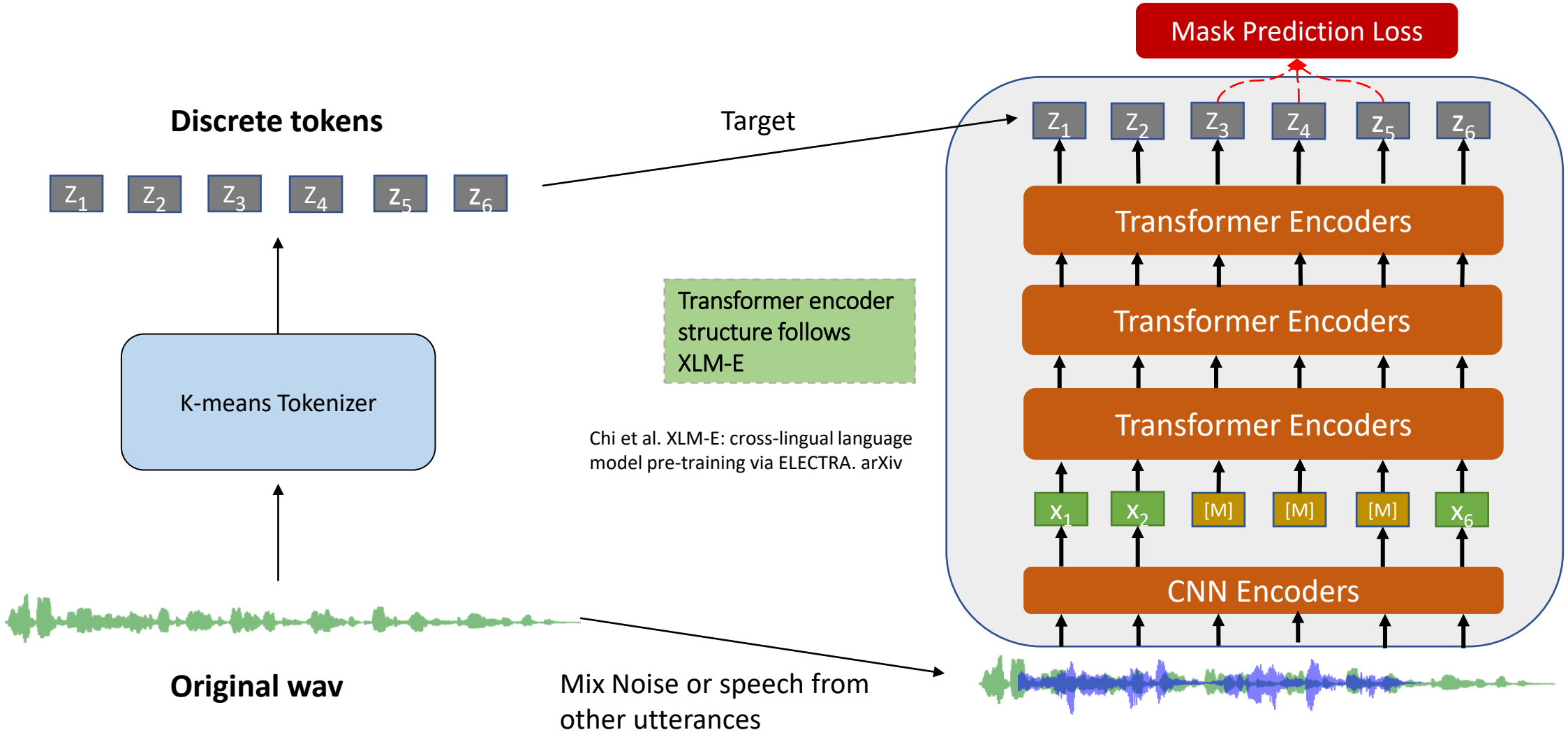
Content Tasks
Speech Recognition, Speech Translation

Speaker Tasks
Speaker Diarization, Speaker Recognition

Can we propose a model for full stack speech processing?


WavLM
=
Better Algorithm
+
Large Data
+
Big Model
General
Diverse
Stable

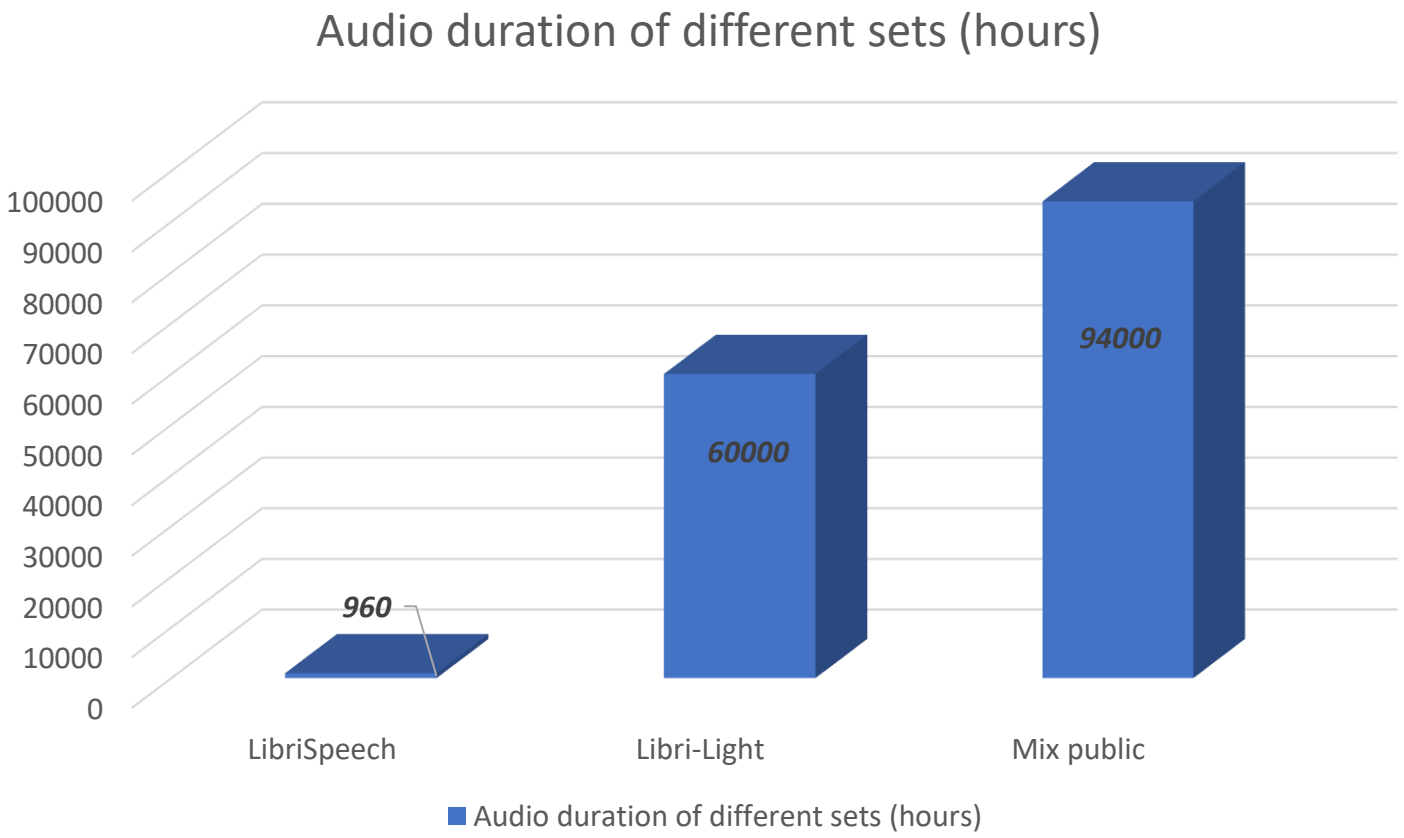
WavLM: Masked Speech **Prediction and Denoising**



Ablation Study: Algorithm

	Speaker Identification ACC↑	Speaker Verification EER↓	Speaker Diarization DER↓	Phoneme Recognition PER↓	ASR WER↓	Keyword Spotting Acc↑	Query by Example Spoken Term Detection MTWV↑	Intent Classification Acc↑	Slot Filling (Slot type) F1↑	Slot Filling (Slot value) CER↓	Emotion Recognition Acc↑
WavLM-Base	85.49	4.49	4.65	4.86	6.13	96.79	0.087	98.63	89.38	22.86	65.94
-structure change	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60
-denoising	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55

Pre-Training Data



Public unlabeled data = Libri-Light (60kh) + VoxPopuli (24kh) + GigaSpeech (10kh)

Ablation Study: Data

	Speaker Identification ACC↑	Speaker Verification EER↓	Speaker Diarization DER↓	Phoneme Recognition PER↓	ASR WER↓	Keyword Spotting Acc↑	Query by Example Spoken Term Detection MTWV↑	Intent Classification Acc↑	Slot Filling (Slot type) F1↑	Slot Filling (Slot value) CER↓	Emotion Recognition Acc↑
WavLM-Base 960h	85.49	4.49	4.65	4.86	6.13	96.79	0.087	98.63	89.38	22.86	65.94
WavLM-Base 60kh	85.54	4.53	4.5	4.49	5.93	96.85	0.071	98.52	89.57	22.8	66.07
WavLM-Base 94kh	**89.42**	**4.11**	**3.5**	**3.92**	**5.6**	**97.37**	**0.098**	**99**	**90.58**	**21.2**	**68.65**

- For a base model, only increasing data amount shows limited gains.
- ***The diversity matters!***

SUPERB Evaluation

SUPERB

HELP

NEWS PAPER CODE TASKS RULES LEADERBOARD CHALLENGE SUBMIT LOGIN

All Const. Less. Un.

Task Collections

Paper Challenge Public Challenge Hidden Dev

Method	Name	Description	URL	Rank ↑	Score ↑	Rank-P ↑	Score-P ↑	PR public ↓	KS public ↑	IC public ↑	SID public ↑	ER public ↑	ASR public ↓	QbE public ↑	SF-F1 public ↑	SF-CER public ↓	SV public ↓	SD public ↓
WavLM Large	Microsoft	M-P + VQ ...		18.9	1145	6.1	3.61	3.06	97.86	99.31	95.49	70.62	3.44	8.86	92.21	18.36	3.77	3.24
WavLM Base+	Microsoft	M-P + VQ ...		17.7	1106	12.7	11.68	3.92	97.37	99	89.42	68.65	5.59	9.88	90.58	21.2	4.07	3.5
WavLM Base	Microsoft	M-P + VQ ...		15.9	1019	11.45	10.76	4.84	96.79	98.63	84.51	65.94	6.21	8.7	89.38	22.86	4.69	4.55
HuBERT Large	paper	M-P + VQ	-	15.1	919	4.1	2.9	3.53	95.29	98.76	90.33	67.62	3.62	3.53	89.81	21.76	5.98	5.75
wav2vec 2.0 Large	paper	M-C + VQ	-	14.8	914	3.9	2.88	4.75	96.66	95.28	86.14	65.64	3.75	4.89	87.11	27.31	5.65	5.62
HuBERT Base	paper	M-P + VQ	-	14.45	941	10.25	9.94	5.41	96.3	98.34	81.42	64.92	6.42	7.36	88.53	25.2	5.11	5.88
FaST-VGS+	Puyuan P...	FaST-VG...	-	12.9	809	5.9	3.72	7.76	97.27	98.97	41.34	62.71	8.83	5.62	88.15	27.12	5.87	6.05
wav2vec 2.0 Base	paper	M-C + VQ	-	11.85	818	8.7	8.61	5.74	96.23	92.35	75.18	63.43	6.43	2.33	88.3	24.77	6.02	6.08
DistilHuBERT	Heng-Jui ...	multi-task ...	-	11.1	717	15.6	30.54	16.27	95.98	94.99	73.54	63.02	13.37	5.11	82.57	35.59	8.55	6.19
DeCoAR 2.0	paper	M-G + VQ	-	10.5	722	8.5	8.03	14.93	94.48	90.8	74.42	62.47	13.02	4.06	83.28	34.73	7.16	6.59
wav2vec	paper	F-C	-	8.9	529	12.55	16.25	31.58	95.59	84.92	56.56	59.79	15.86	4.85	76.37	43.71	7.99	9.9
vq-wav2vec	paper	F-C + VQ	-	7	422	9.8	-5.53	33.48	93.38	85.68	38.8	58.24	17.71	4.1	77.68	41.54	10.38	9.93
APC	paper	F-G	-	5.8	392	16.05	87.25	41.98	91.01	74.69	60.42	59.33	21.28	3.1	70.46	50.89	8.56	10.53
VQ-APC	paper	F-G + VQ	-	5.75	377	14.25	72.1	41.08	91.11	74.48	60.15	59.66	21.2	2.51	68.53	52.91	8.72	10.45
NPC	paper	M-G + VQ	-	5.4	386	12	19.94	43.81	88.96	69.44	55.92	59.08	20.2	2.46	72.79	48.44	9.4	9.34
modified CPC	paper	F-C	-	5.3	278	15.6	113.94	42.54	91.88	64.09	39.63	60.96	20.18	3.26	71.19	49.91	12.86	10.38
TERA	paper	time/freq ...	-	3.5	150	8.7	-141.81	49.17	89.48	58.42	57.57	56.27	18.17	0.13	67.5	54.17	15.89	9.96
PASE+	paper	multi-task	-	2.45	149	10.55	-56.14	58.87	82.54	29.82	37.99	57.86	25.11	0.72	62.14	60.17	11.61	8.68
Mockingjay	paper	time M-G	-	1.15	54	1.75	-93.58	70.19	83.67	34.33	32.29	50.28	22.82	0.07	61.59	58.89	11.66	10.54

Fine-tuning for various downstream tasks

	Vox1-O	Vox-E	Vox-H
ECAPA-TDNN	1.08	1.2	2.127
WavLM Large	**0.383**	**0.480**	**0.986**

Speaker Verification (EER**)**
Train: VoxCeleb2 Dev
Test: VoxCeleb1 Test

	DER
EEND-vector clustering	12.49
WavLM Large	**10.92**

Speaker Diarization (DER**)**
Train-test: CALLHOME

	DNSMOS P.808	DNSMOS P.835_SIG	DNSMOS P.835_BAK	DNSMOS P.835_OVR
LSTM	3.050	3.689	3.687	3.110
WavLM Large	**3.165**	**3.747**	**3.858**	**3.217**

Speech Enhancement (DNSMOS**)**
Train: 10-hour LibriSpeech data
Test: DNS Challenge

	0S	0L	OV10	OV20	OV30	OV40
Conformer	4.5	4.4	6.2	8.5	11	12.6
WavLM Large	**4.3**	**4.2**	**5.0**	**6.3**	**8.2**	**8.8**

Speech Separation (WER for different overlap ratios**)**
Train: WSJ
Test: LibriCSS

	Test-Clean	Test-Other
HuBERT Large	1.9	3.3
WavLM Large	1.8	3.2

Speech Recognition (WER)
Train: 960-hour LibriSpeech data
Test: LibriSpeech

Analysis of SSL Transferability: Loss

- Setting:
  - LibriSpeech 960 -> VoxCeleb Dev ASV task (Transformer + ECAPA_TDNN)
- Compare Different Loss
  - MSE pre-train is to reconstruct raw fbank feature instead of discrete label

	Model	Vox1-O (EER ↓)	Vox1-E (EER ↓)	Vox1-H (EER ↓)
Training from scratch	ECAPA-TDNN	1.01	1.24	2.32
	ECAPA-TDNN + Transformer	3.69	3.71	6.034
ASR Transfer	ASR Pre-train	1.16	1.26	2.43
SSL	HuBERT Pre-train	**0.84**	**0.87**	**1.72**
	Wav2vec 2.0 pre-train	0.973	0.933	1.83
	MSE pre-train	0.979	1.075	1.98

Analysis of SSL Transferability: Quantizers

- Compare Different Quantizers

Model	Vox1-O (EER ↓)	Vox1-E (EER ↓)	Vox1-H (EER ↓)
HuBERT 2nd iter label	**0.84**	**0.87**	1.72
Phone Label	0.867	0.918	1.776
Fbank Clustering	0.883	0.903	**1.675**
VQVAE	0.878	0.939	1.734

- Different quantizers show comparable performance on SV

Conclusion

- WavLM is a general speech processing model for various task
  - Evaluation on SUPERB and other non-ASR tasks

- Investigate the SSL Transferability
  - Masked Speech + Pseudo label prediction is the key

Thanks!