Improving the Universality of Speech Representation Learning

Jiatong Shi
Carnegie Mellon University

Committee

Shinji Watanabe (Chair)
David R. Mortensen
Chris Donahue
Fernando Diaz
Yossi Adi (Hebrew University)

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Speech → Speech Representation

- Speech Signal (Waveform)
  - (assume a single microphone case)
  - 1-dimensional array

  - This raw format
    - takes down the **air pressure information faithfully**
    - but **cannot be easily transformed** into properties understandable by human

Extracting Speech Representation

- The study of extracting speech representations has been ongoing for a <u>long time</u>

- Early works follow studies in psychoacoustics

Extracting Speech Representation

- The study of extracting speech representations has been ongoing for a <u>long time</u>
- Early works follow the study in psychoacoustics

Figures from https://www.uaudio.com/blog/how-the-ear-works/

Extracting Speech Representation

- The study of extracting speech representations has been ongoing for a <u>long time</u>

- Early works follow the study in psychoacoustics
  - Human ear is a frequency acceptor that is stimulated based on the energy at related frequency bands
  - Can we align the representation to the format where human is used to process?

Extracting Speech Representation

- The study of extracting speech representations has been ongoing for a <u>long time</u>

- Early works follow the study in psychoacoustics
  - Human ear is a frequency acceptor that is stimulated based on the energy at related frequency bands
  - Can we align the representation to a format that human are accustomed to processing?
    - Use **Fourier Transform** to extract speech representation!

Conventional Spectral Speech Representation

- Use Fourier Transformer to convert raw waveform into spectral features

Facts:

- Local stability in waveform
- Sensitivity to
  - High vs. low frequencies
  - High vs. low time resolution

Spectral speech representations:

- Use short-time Fourier transform with a sliding window (typical 20-60ms window length + 10-30ms window shift)
- Use filter banks to simulate the sensitivity of human ear to reduce redundant information

From Spectral Features to Neural Representations

- A new paradigm of feature representation -> Learn from data
  - A **further extension** of signal features
  - To **enhance the performance of downstream tasks** with enhanced representation
    - Successful practice in other domains such as natural language processing, computer vision or other multi-modal works

Example about the Feature Switch

Speech Recognition evaluated by word error rate (WER)

LibriSpeech (Test-other set)

(Wang et al. 2022, WavLM)

Significant improvements after using neural representations

Categorization of Speech Representation

- Speech representations can be categorized based on data types and granularity into the following groups:
  - Continuous Frame-wise
  - Discrete Frame-wise
  - Continuous Utterance-wise

Categorization of Speech Representation

- Define a single-channel speech signal as $\mathbf{S} = [s_1, s_2, ..., s_{L_s}] \in \mathbb{R}^{1 \times L_s}$
  - Denote the representation extraction model as $f(\cdot)$
  - Continuous Frame-wise $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T]$
  - Discrete Frame-wise $\mathbf{C} = [c_1, c_2, ..., c_T]$
  - Continuous Utterance-wise $\mathbf{h}$

L_s, T are length indexes for speech and frame-wise representation

Categorization of Speech Representation

- Given the data types and granularity, we categorize speech representation into:
  - Continuous Frame-wise → beneficial for **understanding tasks** involving rich information
  - Discrete Frame-wise → beneficial for **generation tasks** because they simplify prediction
  - Continuous Utterance-wise → beneficial for **general summarization** of audio properties

Challenges in Speech Representation

- Universality
  - To meet the increasing needs of speech foundation models
  - To ensure the technology is accessible and usable by all communities
  - To enhance the understanding of the real-world, rooted in the intrinsic complexity and variability of speech as a modality

Focuses of the Thesis

- Three kinds of universalities in speech representation learning
  - Diverse linguistic landscapes → Universal language processing
  - Task fragmentation → Universal task handling
  - Evaluation universality → Universal evaluation framework

Universal Language Processing

- Denote language $l \in \mathcal{Z}$ $\mathcal{Z}$ is the set of all world languages

- The monolingual model is optimized for

$$\text{argmax}_{f_l(\cdot)} p(\mathbf{H}_l | \mathbf{S}, l).$$

- For universal language processing, we aim to shift the paradigm to encompass all world languages:

$$\text{argmax}_{f_{\mathcal{Z}}(\cdot)} p(\mathbf{H}_{\mathcal{Z}} | \mathbf{S}).$$

Universal Task Handling

- We use $\mathbf{X}, \mathbf{Y}$ to represent expected I/O for the corresponding tasks.
  - For understanding tasks, we have $\mathbf{X}_{\text{und}}, \mathbf{Y}_{\text{und}}$
  - For generation tasks, we have $\mathbf{X}_{\text{gen}}, \mathbf{Y}_{\text{gen}}$
- The target of the problem is to:

$$\text{argmax}_{\theta_{\text{und}},\theta_{\text{gen}},f_{\text{uni}}(\cdot)}[\mathbb{E}_{\text{und}}[\log p_{\theta_{\text{und}}}(\mathbf{Y}_{\text{und}}|\mathbf{X}_{\text{und}}, f_{\text{uni}}(\cdot))] + \mathbb{E}_{\text{gen}}[\log p_{\theta_{\text{gen}}}(\mathbf{Y}_{\text{gen}}|\mathbf{X}_{\text{gen}}, f_{\text{uni}}(\cdot))]]$$

$\theta_{\text{und}}, \theta_{\text{gen}}.f_{\text{uni}}(\cdot)$ are downstream model for understanding and generation, and the universal speech representation extractor, respectively

Universal Evaluation Framework

- Denote a domain in speech profiling $b \in \mathcal{B}$ $\mathcal{B}$ is the set of speech profiling dimensions

- The focus of single-domain is:

$$\mathrm{argmax}_{\theta_b} p(y_b|\mathbf{S}, f(\cdot))$$

- The target of the problem is to:

$$\mathrm{argmax}_{\theta} \sum_{b \in \mathcal{B}} \mathbb{E}_b(\log p_\theta(y_b|\mathbf{S}, f(\cdot))).$$

Improving Universality

Improving Universality

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Universal Language Processing

- Evaluation of speech representation
  - Not explicitly related to a specific task
    - Word error rate for speech recognition
    - Signal-to-noise ratio for speech enhancement/separation
    - Mean opinion score for speech synthesis
  - One of the key challenges of speech representation learning

Universal Language Processing

- Evaluation of speech representation

(Yang et al. 2021, SUPERB)

  - Speech Universal PERformance Benchmark (SUPERB)

    - A framework to evaluate speech representation by using downstream tasks

    - A wide-coverage of different tasks in:

RECOGNITION | DETECTION | SEMANTICS | SPEAKER | PARALINGUISTICS | GENERATION

Issues with SUPERB

- Designed primarily for English speech
- Does not address the growing interest in applying speech representation for multilingual scenarios

- With the call to supporting universal language processing, we propose:

SUPERB

Multi**L**ingual **S**peech processing
Universal **PER**formance **B**enchmark

Shi, Jiatong et al. "ML-SUPERB: Multilingual Speech Universal PERformance Benchmark." IS2023.

Comparison to other Benchmarks

- ML-SUPERB, at the time of proposing, is the most comprehensive benchmark supporting **143 languages equally**

Benchmarks	Supported Languages	Citation
Lebenchmark	2	(Evain et al. 2021)
IndicSUPERB	12	(Javed et al. 2023)
XTREME-S	102	(Conneau et al. 2022)
ML-SUPERB	**143**	(Shi et al. 2023)

Dataset Formulation

- A collection of public databases

- Processed uniformly into 10-minute and 1-hour collections with an additional consideration in few-shot learning

Dataset	Hours	Normal Languages (!23)	Few-shot Languages (20)
10-min	37.43	~10min x 240 (lang, data)	5 utt. x 20 lang
1-h	222.46	~1h x 240 (lang, data)	5 utt. x 20 lang
Dev	41.82	~10min x 240 (lang, data)	~10min x 31 (lang, data)
Test	44.97	~10min x 240 (lang, data)	~10min x 31 (lang, data)

Dataset Formulation

- A collection of public databases

- Processed uniformly into 10-minute and 1-hour collections with an additional consideration in few-shot learning

Dataset	Hours	Normal Languages (!23)	Few-shot Languages (20)
10-min	37.43	~10min x 240 (lang, data)	5 utt. x 20 lang
1-h	222.46	~1h x 240 (lang, data)	
Dev	41.82	~10min x 240 (lang, dat	
Test	44.97	~10min x 240 (lang, data)	~10min x 31 (lang, data)

Smaller Dataset → Efficient Evaluation

Tracks Design

- Monolingual Track --> Monolingual ASR
- Multilingual Track
  - Multilingual ASR
  - LID
  - Multilingual ASR + LID

- Evaluated with
  - Character error rate (CER)
  - Phone error rate (PER)
  - Prediction accuracy (ACC)

Model	Params (M)	Pre-trained Languages
wav2vec2-base	95	1
wav2vec2-large	317	1
robust-wav2vec2	317	1
wav2vec2-base-23	95	23
wav2vec2-large-23	317	23
XLS-R 53	317	53
XLS-R 128	317	128
HuBERT-base	95	1
HuBERT-large	317	1
HuBERT-base-cmn	95	1
HuBERT-large-cmn	317	1
mHuBERT-base	95	3

Initial models in the benchmark

General Performance Estimation

- Denote
  - $\mathcal{T}$ as the set of tasks
  - $\mathcal{I}$ as the metric set for corresponding task

$$\mathrm{SUPERB}_{\mathrm{score}}(f) = \frac{1000}{|\mathcal{T}|} \sum_{t}^{\mathcal{T}} \frac{1}{|\mathcal{I}_t|} \sum_{i}^{I_t} \frac{\mathrm{score}_{t,i}(f) - \mathrm{score}_{t,i}(\mathrm{FBANK})}{\mathrm{score}_{t,i}(\mathrm{SOTA}) - \mathrm{score}_{t,i}(\mathrm{FBANK})}$$

ML-SUPERB Results (10min)

SSL Model	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
		Normal	Few-shot	Normal	Normal	Normal	Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
FBANK	72.1	62.4	58.3	11.1	35.9	62.0	58.9	0.0
wav2vec2-base	44.2	43.0	45.7	54.4	66.9	40.6	44.2	755.2
wav2vec2-large	42.0	42.6	45.8	30.9	54.6	45.5	50.3	598.3
robust-wav2vec2	44.4	40.1	45.4	50.8	33.1	38.6	44.9	680.3
wav2vec2-base-23	49.2	37.7	43.4	58.7	45.1	37.2	44.3	735.7
wav2vec2-large-23	42.0	42.1	44.3	1.1	21.8	43.4	46.1	433.8
XLSR-53	49.5	33.9	43.6	6.6	45.6	33.4	43.2	528.8
XLSR-128	39.7	29.2	40.9	66.9	55.6	28.4	42.1	947.5
HuBERT-base	42.8	39.8	44.5	61.2	71.5	39.2	43.8	831.9
HuBERT-large	**38.2**	44.4	48.2	46.5	55.4	45.6	49.3	678.7
HuBERT-base-cmn	43.1	40.8	45.4	49.3	75.1	37.7	43.5	779.0
HuBERT-large-cmn	39.4	42.6	45.8	39.5	66.4	41.9	45.2	715.4
mHuBERT-base	41.0	40.5	45.6	52.4	46.6	36.8	44.2	746.2

ML-SUPERB Results (10min)

SSL Model	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
		Normal	Few-shot	Normal	Normal	Normal	Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
FBANK	72.1	62.4	58.3	11.1	35.9	62.0	58.9	0.0
wav2vec2-base	44.2	43.0	45.7	54.4	66.9	40.6	44.2	755.2
wav2vec2-large	42.0	42.6				45.5	50.3	598.3
robust-wav2vec2	44.4	40.1				38.6	44.9	680.3
wav2vec2-base-23	49.2	37.7				37.2	44.3	735.7
wav2vec2-large-23	42.0	42.1				43.4	46.1	433.8
XLSR-53	49.5	33.9				33.4	43.2	528.8
XLSR-128	39.7	29.2	40.9	66.9	55.6	28.4	42.1	947.5
HuBERT-base	42.8	39.8	44.5	61.2	71.5	39.2	43.8	831.9
HuBERT-large	**38.2**	44.4	48.2	46.5	55.4	45.6	49.3	678.7
HuBERT-base-cmn	43.1	40.8	45.4	49.3	75.1	37.7	43.5	779.0
HuBERT-large-cmn	39.4	42.6	45.8	39.5	66.4	41.9	45.2	715.4
mHuBERT-base	41.0	40.5	45.6	52.4	46.6	36.8	44.2	746.2

- HuBERT-models are better?
- **Large models** are better

ML-SUPERB Results (10min)

SSL Model	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
		Normal	Few-shot	Normal	Normal	Normal	Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
FBANK	72.1	62.4	58.3	11				
wav2vec2-base	44.2	43.0	45.7	54				
wav2vec2-large	42.0	42.6	45.8	30				
robust-wav2vec2	44.4	40.1	45.4	50				
wav2vec2-base-23	49.2	37.7	43.4	58				
wav2vec2-large-23	42.0	42.1	44.3	1				
XLSR-53	49.5	33.9	43.6	6				
XLSR-128	39.7	29.2	40.9	66.9	55.6	28.4	42.1	947.5
HuBERT-base	42.8	39.8	44.5	61.2	71.5	39.2	43.8	831.9
HuBERT-large	**38.2**	44.4	48.2	46.5	55.4	45.6	49.3	678.7
HuBERT-base-cmn	43.1	40.8	45.4	49.3	75.1	37.7	43.5	779.0
HuBERT-large-cmn	39.4	42.6	45.8	39.5	66.4	41.9	45.2	715.4
mHuBERT-base	41.0	40.5	45.6	52.4	46.6	36.8	44.2	746.2

- Model trained with **more languages** are better
- Large models are not **necessarily better**
- Robustness is also important to consider

After ML-SUPERB

- Following the release of ML-SUPERB,
- we extend the activity to a public call for challenges in ASRU2023
  - Receiving 12 model submissions and 54 new language contributions
  - Expanding the challenge to **154 languages**

The geographical distribution of the new 54 languages

Shi Jiatong et al. “Findings of the 2023 ML-SUPERB Challenge: Pre-Training and Evaluation over More Languages and Beyond.” ASRU 2023.

Limitations of ML-SUPERB

- The benchmark settings are strictly constrained to self-supervised learning (SSL) pre-trained models
  - While efficient, the benchmark is not sufficiently generalizable to various settings (Zaiem et al. 2023; Arora et al. 2024)
- This motivates benchmarking with more **flexible constraints**

Introduction of ML-SUPERB 2.0

- We revisit the ML-SUPERB:
  - By **relaxing its fixed constraints**
  - By **enriching its evaluation metrics** to focus on **robustness** across languages and **variations** across datasets

Shi, Jiatong et al. “ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets.” IS 2024.

Introduction of ML-SUPERB 2.0 (Cont'd)

- Specifically, we investigate **four new scenarios**:
  - Large downstream models
  - SSL model fine-tuning
  - Efficient model adaptation strategies
  - Supervised pre-trained models

Experimental Design (Evaluation)

- Base metrics:
  - Accuracy for LID
  - Character error rate (CER) for ASR in two subsets (normal and few-shot setting)

- **<u>Enhanced evaluation:</u>**

  - Macro-average over languages/datasets instead of micro-average CER

    - A more reasonable workaround for variations in <u>sentence length, language types (symbolic or alphabetic), and domain differences</u>.

  - Standard deviation of CER across languages

  - Measure CER of the worst-performing languages

  - Measure CER across datasets for the same languages

Effect of Introducing Four Scenarios

Scenarios	Details	Accuracy	CER (Normal)
Original ML-SUPERB	MMS + Transformer CTC	90.3	24.7 ± 12.3
Large Downstream	MMS + E-Branchformer ATT-CTC	95.2	16.6 ± 11.8
SSL Model Fine-tuning	MMS + 9-14 layers partial fine-tuning CTC	**95.6**	**15.5 ± 10.3**
Efficient Model Adaptation	MMS + LoRA + Transformer ATT-CTC	94.2	18.7 ± 11.5
Supervised Pre-trained Model	Whisper Encoder + Transformer CTC	91.7	21.0 ± 12.5

Compared to the original ML-SUPERB, we observe **better performance** for LID and ASR across **ALL configurations** (normal setting)

CTC: the ASR model trained with a connectionist temporal classification objective
ATT-CTC: the ASR model trained with a CTC/Attention hybrid objective

Model Ranking given Different Configurations

- In the original ML-SUPERB,
  - XLS-R reaches better performance in LID
  - MMS achieves better performance in ASR

- However, when considering **different** training settings, the ranking of upstream models can be **different**

Model Ranking given Different Configurations (Large Downstream Models)

	Transformer	Conformer	E-Branchformer
CTC	XLS-R	MMS	XLS-R
ATT-CTC	MMS	MMS	MMS

XLS-R wins

MMS wins

Compared to the original ML-SUPERB, the performance of XLS-R and MMS **depends on the choice of the downstream model**

Model Ranking given Different Configurations (Model Fine-tuning)

	Bottom	Middle	Top
CTC	MMS	MMS	MMS
ATT-CTC	MMS	MMS	MMS

XLS-R wins

MMS wins

Compared to the downstream model configuration,
XLS-R and MMS **rank differently** when considering fine-tuning approaches

Model Ranking given Different Configurations (Efficient Model Adaptation)

	LoRA	Adapter
CTC	XLS-R	XLS-R
ATT-CTC	MMS	XLS-R

XLS-R wins

MMS wins

Compared to previous experimental settings,
XLS-R and MMS **rank differently** when considering efficient model adaptation approaches

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

How to Extend Speech Representation for Universal Language Processing?

- Two routes to extend the representation:
  - External route: combining information from different speech representations
  - Internal route: improving information utilization within a single speech representation

External Route → Representation Fusion

- A straightforward idea to enhance the representation
  → combine various sources of speech representation

- At the time when there were limited multilingual neural speech representation models, spectral representations

  - **are domain and language agnostic** and have been used in many previous multilingual speech models
  - are **less vulnerable to domain shifts**

Berrebbi, Dan et al. ”Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation” IS 2022.

Fusion of Spectral and Neural Representations

- Learnable combination

$$f_{\text{FUSE}}(\mathbf{S}) = \textsc{Linear}(\textsc{Transform}(f_{\text{spectral}}(\mathbf{S}), f_{\text{neural}}(\mathbf{S}))$$

- A range of ”TRANSFORM” methods
  - Concatenation
  - Convolutional Fusion
  - **Co-Attention based Fusion**

Co-attention Fusion

- Define the query, key, and value as

$$i \in \{\text{spectral}, \text{neural}\}$$

$$Q_{\mathrm{i}} = f_{\mathrm{i}}(\mathbf{S})W_{\mathrm{i}}^{\mathrm{Q}}, \quad K_{\mathrm{i}} = f_{\mathrm{i}}(\mathbf{S})W_{\mathrm{i}}^{\mathrm{K}}, \quad V_{\mathrm{i}} = f_{\mathrm{i}}(\mathbf{S})W_{\mathrm{i}}^{\mathrm{V}}$$

$$\mathbf{h}_i = \textsc{SoftMax}\left(\frac{Q_i \cdot K_i}{\sqrt{D}}\right)V_i + f_i(\mathbf{S})$$

$$f_{\textsc{Fuse}}(\mathbf{S}) = \textsc{Linear}(\textsc{Concat}(\mathbf{h}_{\text{spectral}}, \mathbf{h}_{\text{neural}}))$$

Experiments on Low-resource Languages

- Totonac:
  - ISO-639-3 code: toc
  - An endangered Mexican language, belonging to the Nahuatl language family
  - 10-hour training data
- Arabic
  - in CommonVoice Dataset
  - 20-hour training data
- Use a monolingual neural speech representation:
  - HuBERT-large, pre-trained on LibriLight with 60k hours of English speech

Experiments of Representation Fusion

Model	Totonac-CER	Arabic-CER
Spectral Representation	17.2	15.4
HuBERT-large Representation	14.2	8.1
Linear Fusion	14.0	6.6
Convolutional Fusion	13.9	7.2
Co-Attention Fusion	**13.4**	**5.4**

- **Significant improvements** compared to using a single representation
- **The co-attention-based** fusion achieves the best results

After the initial fusion investigation,

- More and more multilingual speech representation emerge in the community
  - XLS-R (Babu et al. 2022), MMS (Pratap et al. 2023), WavLabLM (Chen et al. 2023), SSHR (Xue et al. 2023), etc.

  (Many are submissions to the ML-SUPERB challenge)

- Can we further combine the knowledge for multilingual scenarios?

A Straightforward Extension

- Fusion of multiple multilingual neural speech representation

Complexity of Neural Representation Fusion

- Fusion of two representations with rich information is difficult
  - The fusion is not always helpful due to this complexity
    (as shown in later experiments)
- Using multiple large speech representation extractors is inefficient

An Efficient Representation Fusion Technique

- Two-stage modeling

Srivastava, Tejes et al. “FFUSE: Efficient Self-Supervised Feature Fusion for E2E ASR in Multilingual and Low Resource Scenarios” IS 2024.

  - Stage1: Fusion stage

An Efficient Representation Fusion Technique

- Two-stage modeling
  - Stage2: Prediction stage

Experiments for the Efficient Fusion

- Experiments in two folds
  - Low-resource scenarios
    - Evaluated on Mexican endangered languages (Yoloxochitl Mixtec, a tonal language; and Totonac Nahuatl, a morphological complex language)
  - Multilingual scenarios
    - Evaluated with ML-SUPERB

Results on Low-resource Scenarios

Model	Mixtec-WER	Mixtec-CER	Nahuatl-WER	Nahuatl-CER
MMS	22.6	10.4	48.2	16.6
XLS-R	23.0	10.4	44.9	14.7
MMS + XLS-R (Fusion)	24.1	10.7	43.8	14.0
MMS -> XLS-R	21.8	9.8	46.2	15.4
XLS-R -> MMS	22.6	10.1	44.6	13.9

While the use of representation fusion can help the ASR model performance for Nahuatl, it does not help to Mixtec

Suggesting **<u>the complexity of representation fusion for ASR</u>**

Results on Low-resource Scenarios

Model	Mixtec-WER	Mixtec-CER	Nahuatl-WER	Nahuatl-CER
MMS	22.6	10.4	48.2	16.6
XLS-R	23.0	10.4	44.9	14.7
MMS + XLS-R (Fusion)	24.1	10.7	43.8	14.0
MMS -> XLS-R	21.8	9.8	46.2	15.4
XLS-R -> MMS	22.6	10.1	44.6	13.9

However, the proposed methods can **always show benefits** from the source speech representation

Results on Low-resource Scenarios

Model	Mixtec-WER	Mixtec-CER	Nahuatl-WER	Nahuatl-CER
MMS	22.6	10.4	48.2	16.6
XLS-R	23.0	10.4	44.9	14.7
MMS + XLS-R (Fusion)	24.1	10.7	43.8	14.0
MMS -> XLS-R	21.8	9.8	46.2	15.4
XLS-R -> MMS	22.6	10.1	44.6	13.9

During inference, the proposed model can **be more efficient** compared to direct fusion due to less inference with speech representation extractor

Results on ML-SUPERB

SSL Model	Info	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
			Normal	Few-shot	Normal	Normal	Normal	Few-shot	
		CER/PER	CER	CER	ACC	ACC	CER	CER	
MMS	Baseline	**33.8** / 30.5	29.9 / 24.6	35.0 / 35.4	62.3 / 84.3	73.8 / 88.9	29.5 / 24.5	34.9 / 35.4	1051.7 / 1029.2
XLS-R	Baseline	39.5 / 30.5	28.9 / 21.6	41.4 / 39.1	65.4 / 87.2	76.9 / 90.4	28.7 / 22.0	41.3 / 38.3	957.6 / 1018.1
MMS + XLS-R	Fusion (Stage 1)	35.6 / **27.9**	27.0 / **20.6**	38.9 / 36.6	81.3 / 86.3	78.1 / 91.9	26.8 / **20.5**	38.7 / 36.9	1095.2 / **1068.4**
MMS → XLS-R	Prediction (Stage 2)	38.9 / 29.4	**26.5** / 22.3	**33.8 / 35.4**	83.4 / 84.9	**82.9 / 92.1**	**26.3** / 21.6	**33.8 / 34.0**	**1152.3** / 1066.5
XLS-R → MMS	Prediction (Stage 2)	38.6 / 28.9	28.3 / 21.0	39.3 / 39.7	**83.9 / 90.4**	76.8 / 91.0	28.1 / 21.0	39.3 / 39.1	1054.8 / 1037.0

We observe **similar trends as the low-resource scenarios** in ML-SUPERB, especially for multilingual processing in few-shot speech recognition

Detailed look into Parameter Efficiency

Model	Num. Param. (M)	Δ Num. Param. (M) (MMS Baseline)	Δ Num. Param. (M) (XLS-R Baseline)	Δ SUPERB Score (MMS Baseline)	Δ SUPERB Score (XLS-R Baseline)
MMS (M)	322	/	/	/	/
XLS-R (X)	324	/	/	/	/
M + X (Fusion)	641	+319	+317	+43.5	+137.6
M → X	324	**+2**	/	**+100.6**	/
X → M	324	/	**+0**	/	**+97.2**

The proposed method enjoys **comparable (or better) performance** to the fusion model with **a much less need to more network parameters**.

Internal Route → Multi-resolution

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

- The local stability
  - Information of speech is in **local sequences** of samples
  - 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)
    - Learnable filters: Convolutional networks, which can be applied in conventional speech downstream tasks or self-supervised models such as wav2vec2, HuBERT (20ms)

Internal Route → Multi-resolution

- In multilingual world,
  - different languages can be presented in different phoneme layout, suggesting a **diverse temporal complexity**
  - modeling multilingual speech in multi-resolution could be a **natural fit to** enhance the speech representation learning in the scenario

Does resolution matter?

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

Configurations of the pre-trained HuBERT with different resolutions. The convolution (Conv.) module is represented in [(kernal-size, stride) * layer-number]

ID	Res. (ms)	Param. (M)	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

Shi, Jiatong et al. “Exploration on HuBERT with Multiple Resolution” IS 2023.

Does resolution matter?

- 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**

A range of tasks includes phone recognition (PR), speech recognition (ASR), emotion recognition (ER), intent classification (IC), speaker identification (SID), speaker verification (SV), speech enhancement (SE), speech translation (ST)

Further Extension

- HuBERT with multiple resolutions → multi-resolution HuBERT
  - achieving multi-resolutional processing within a single network

Shi, Jiatong et al. "Multi-resolution HuBERT: Multi-resolution Speech Self-Supervised Learning with Masked Unit Prediction." ICLR2024.

Comparison to other HuBERT models

- With a pre-trained model trained on Voxpopuli (100k hours, 17 languages)
  - A model with the base size (~100M parameters) can reach **better performances** to other HuBERT-based models trained on English or multilingual setups

SSL Model	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
		Normal	Few-shot	Normal	Normal	Normal	Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
HuBERT-base	42.8 35.3	39.8 / 31.4	44.5 / 42.7	61.2 / **86.1**	**71.5 / 86.0**	39.2 / 30.9	43.8 / 41.8	831.9 / 884.9
HuBERT-large	**38.2** / 32.2	44.4 / 37.7	48.2 / 43.5	46.5 / 64.1	55.4 / 77.7	45.6 / 35.1	49.3 / 42.2	678.7 / 783.6
mHuBERT-base	41.0 / 33.0	40.5 / 33.4	45.6 / 43.6	52.4 / 72.5	46.6 / 70.9	36.8 / 29.7	44.2 / 43.1	746.2 / 812.7
MR-HuBERT-base	38.3 / **30.6**	**34.1 / 27.5**	**39.6 / 38.9**	**64.0** / 85.1	69.9 / 84.4	**34.4 / 28.0**	40.9 / **36.6**	**957.2 / 986.8**

Results are shown in (10min track / 1h track)

Comparison to Other Multilingual Models

SSL Model	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
		Normal	Few-shot	Normal	Normal	Normal	Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
HuBERT-base	42.8 / 35.3	39.8 / 31.4	44.5 / 42.7	61.2 / **86.1**	**71.5 / 86.0**	39.2 / 30.9	43.8 / 41.8	831.9 / 884.9
HuBERT-large	**38.2** / 32.2	44.4 / 37.7	48.2 / 43.5	46.5 / 64.1	55.4 / 77.7	45.6 / 35.1	49.3 / 42.2	678.7 / 783.6
mHuBERT-base	41.0 / 33.0	40.5 / 33.4	45.6 / 43.6	52.4 / 72.5	46.6 / 70.9	36.8 / 29.7	44.2 / 43.1	746.2 / 812.7
wav2vec2-base-23	49.2 / 35.1	37.7 / 32.0	43.4 / 42.2	58.7 / 71.9	45.1 / 66.3	37.2 / 30.9	44.3 / 43.0	735.7 / 798.0
wav2vec2-large-23	42.0 / 34.2	42.1 / 35.3	44.3 / 42.4	1.1 / 64.2	21.8 / 49.7	43.4 / 35.2	46.1 / 43.1	433.8 / 724.9
XLS-R-53	49.5 / 34.9	33.9 / 26.9	43.6 / 40.6	6.6 / 87.1	45.6 / 76.9	33.4 / 28.6	43.2 / 44.6	528.8 / 894.0
XLS-R-128	39.7 / **30.6**	**29.2 / 22.0**	40.9 / 39.3	**66.9 / 87.9**	55.6 / 85.6	**28.4 / 22.9**	42.1 / 42.4	947.5 / **996.0**
MR-HuBERT-base	38.3 / **30.6**	34.1 / 27.5	**39.6 / 38.9**	64.0 / 85.1	69.9 / 84.4	34.4 / 28.0	**40.9 / 36.6**	**957.2** / 986.8

Comparison to Other Multilingual Models

SSL Model	Mono	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB Score
	AS						Few-shot	
	CER						CER	
HuBERT-base	42.8						43.8 / 41.8	831.9 / 884.9
HuBERT-large	**38.2**						49.3 / 42.2	678.7 / 783.6
mHuBERT-base	41.0						44.2 / 43.1	746.2 / 812.7
wav2vec2-base-23	49.2						44.3 / 43.0	735.7 / 798.0
wav2vec2-large-23	42.0						46.1 / 43.1	433.8 / 724.9
XLS-R-53	49.5 / 34.9	33.9 / 26.9	43.6 / 40.6	6.6 / 87.1	45.6 / 76.9	33.4 / 28.6	43.2 / 44.6	528.8 / 894.0
XLS-R-128	39.7 / **30.6**	**29.2 / 22.0**	40.9 / 39.3	**66.9 / 87.9**	55.6 / 85.6	**28.4 / 22.9**	42.1 / 42.4	947.5 / **996.0**
MR-HuBERT-base	38.3 / **30.6**	34.1 / 27.5	**39.6 / 38.9**	64.0 / 85.1	69.9 / 84.4	34.4 / 28.0	**40.9 / 36.6**	**957.2** / 986.8

Compared to XLSR-128, the proposed model can reach comparable performance (slightly better in 10min, slightly worse in 1h), with

- **⅓ model size**
- **¼ pre-trained data**
- **⅕ language coverage**

Summary

- In this section, we focus on universal language processing for speech representation learning, specifically we:
  - Present the **multilingual SUPERB (ML-SUPERB)** for evaluating speech representation in multilingual scenarios
  - Use **representation fusion techniques** to fuse representations to achieve better speech representation for multilingual and low-resource speech processing
  - Adopt **multi-resolution framework** to enhance the speech representation learning framework for multilingual processing

Improving Universality

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Achieving Universal Task Handling

- Transitioning to discrete representation
  - Could potentially a possible approach to achieve **universal task handling**

Transitioning to Discrete Representation

- Benefits from discrete representation

- Scaling up (e.g., TTS with VALL-E (Wang et al. 2023))

	Conventional TTS Systems	VALL-E
Intermediate Representation	Continuous spectral representation	Audio codec code
Objective Function	Continuous Regression	Language Model
Training Data	< 600 hours	60k hours
In-context Learning	❌	✅

Transitioning to Discrete Representation

- Benefits from discrete representation

- Improved storage efficiency (e.g., discrete ASR (Chang et al. 2023))

Data Format	Data size (bits)
Raw waveform	16 x 16000 x T
Acoustic Features	32 x D x 100 x T
Self-supervised learning representation	32 x 1024 x 50 x T
Discrete tokens	**12 x 50 x T**

Transitioning to Discrete Representation

- Benefits from discrete representation
- Enhanced efficiency in training (e.g., singing synthesis with TokSing (Wu et al. 2024))

Convergence speed comparison between Discrete-based SVS and SVS based on Mel spectrogram

Transitioning to Discrete Representation

- Benefits from discrete representation

- Enhanced efficiency in computation (e.g., speech enhancement with CodecFormer (Yip et al. 2024))

Model	GMACs	Training Time (h/epoch)
SepFormer	77.3	2.7
CodecFormer	**1.5**	**1.0**

Transitioning to Discrete Representation

- Benefits from discrete representation
- Potential for integration with various modalities (e.g., Multimodal LLM with AnyGPT (Zhang et al. 2024))

Down to Speech Tasks

- A common split of speech tasks are:
  - Understanding focus:
    - speech recognition, spoken language understanding, speech translation
  - Generation focus:
    - speech synthesis, voice conversion, singing voice synthesis, speech enhancement

From Continuous to Discrete

- In previous section, we mostly focus on self-supervised learning (SSL) speech representation
  - Pros:
    - Leveraging **large amounts** of unlabeled speech data
    - Exhibits strong performance **in understanding tasks**
  - Cons:
    - Scalability issues in **storage and computation**
    - Challenges in **integration** with other modalities (e.g., text)

From Continuous to Discrete

- In previous section, we mostly focu
  speech representation
  - Pros:
    - Leveraging **large amounts** of unlabeled
    - **Good performance** in understanding tas
  - Cons:
    - Scalability issues in **storage and computation**
    - Difficulties in **integration** with other modalities (e.g., text)

Recent works have shown **reasonable benefits** of using discrete speech representation to avoid/mitigate the issues. (Chang et al. 2023; Hayashi et al. 2020; Polyak et al. 2021; Lee et al. 2022, Zhang et al. 2023)

Introduction: From Continuous to Discrete

- Two mainstream approaches for discrete speech representation
  - SSL-based units
    - Clustering over SSL representation
  - Neural codecs
    - Quantization over waveform reconstruction

From Continuous to Discrete

- Two mainstream approaches for discrete speech representation
  - SSL-based units
    - Pros: A good tradeoff between efficiency and effectiveness
    - Cons: Worse performance than continuous SSL representation and less detailed acoustics information for speech generation.

From Continuous to Discrete

- Two mainstream approaches for discrete speech representation
  - Neural codecs
    - Pros: Enhanced expressiveness in audio quality preservation
    - Cons: Limited semantic information with short-context modeling --> relative worse performance in understanding tasks (e.g., speech recognition)

SSL Units and Neural Codec

- Limited comparative analysis between the two types
- Notably
  - SSL units typically operate within a **single-stream**
  - Neural codecs are in **multi-stream** naturally

Introduction: SSL Units and Neural Codec

- Limited comparative analysis between the two type
- Notably
  - SSL units typically operate in a **single-stream**
  - Neural codecs are in **multi-stream** naturally

Two Enhancement Approaches

- **Enhance SSL units for generation tasks**
- Enhance neural codecs for understanding tasks

MMM Framework for SSL Units

- We enhance SSL units by including more information →
  - A **multi-layer multi-residual multi-stream (MMM)** framework for multi-stream SSL units
- Using MMM framework, we conduct extensive analysis and show that:
  - MMM-based discrete speech units **elevate the performances** by **a large margin**
  - While **maintaining good speech understanding knowledge**, MMM-based units can **achieve comparable or better performance** to neural codec-based approaches in **speech generation tasks**

Shi, Jiatong et al. "MMM: Multi-Layer Multi-Residual Multi-Stream Discrete Speech Representation from Self-supervised Learning Mode.l" IS 2024.

MMM Framework

- Two base setups for the MMM-framework
  - Multi-stream unit extraction from a single layer
  - Multi-stream unit extraction from multiple layers

Multi-stream Unit Extraction (Single Layer)

- Iteratively residual clustering over single-layer SSL representation

Multi-stream Unit Extraction (Multi-layer)

- Clustering over multi-layers

Source Multi-layer Representation

Discrete Multi-layer Representation

Application of MMM-based Discrete Units

- Input Scenario
  - Streamwise (layer-wise) weighted summation over multi-stream embeddings
- Output Scenario
  - Parallel prediction of multi-stream discrete units

Strategies of Multi-Layer Selection

- Discretization over all layers of SSL models are not efficient
- To balance quantization efficiency and performance, we
  - Train a model with representation from all layers, employing learnable layer-wise weighted sum
  - Select layers with top weights

Experiments – Speech Recognition

- Experimental Setup
  - Dataset: The official challenge dataset of discrete speech challenge (ASR)
    - A combination of ML-SUPERB multilingual 1-hour set + LibriSpeech train-clean-100h
    - Evaluation: character error rate (CER) for ML-SUPERB and word error rate (WER) for Librispeech
  - SSL models: WavLM (Chen et al. 2022)
  - K-means: 30% training data into K=500 clusters
  - Downstream model: the same model as challenge baseline

Experiments – Speech Recognition

SSL Base	Single-layer Stream	Multi-layer Stream	Total Streams	Librispeech WER	ML-SUPERB CER
WavLM	1	1	1	6.3	22.8
Encodec	-	-	8	15.9	35.9
WavLM	2	1	2	5.9	21.4
WavLM	1	4	4	5.0	20.8
WavLM (MMM)	2	4	8	**4.7**	**19.5**

The proposed MMM framework demonstrated significantly **better recognition** performance over single-stream scenarios

Experiments – Speech Recognition

SSL Base	Single-layer Stream	Multi-layer Stream	Total Streams	Librispeech WER	ML-SUPERB CER
WavLM	1	1	1	6.3	22.8
Encodec	-	-	8	15.9	35.9
WavLM	2	1	2	5.9	21.4
WavLM	1	4	4	5.0	20.8
WavLM (MMM)	2	4	8	**4.7**	**19.5**

Also, multi-stream from both single layer and multi-layer settings demonstrate **complimentary benefits**

Experiments – Speech Resynthesis (Vocoder)

- Experimental Setup
  - Dataset: The official challenge dataset of discrete speech challenge (Vocoder)
    - A filtered Expresso dataset
  - Evaluation: mel cepstral distortion and UTMOS (Saeki et al. 2022)
  - SSL model: HuBERT-base model
  - K-means: whole training data into K=500 clusters
  - Downstream model: discrete HiFi-GAN vocoder (Lee et al. 2022)

Experiments – Speech Resynthesis (Vocoder)

SSL (Model)	Single-layer Stream	Multi-layer Stream	Total Streams	MCD	UTMOS
HuBERT	1	1	1	7.19	2.27
Encodec	-	-	8	**3.91**	3.18
HuBERT	2	1	2	6.79	2.89
HuBERT	1	4	4	5.12	3.10
HuBERT (MMM)	2	4	8	4.54	**3.22**

In vocoder cases, MMM-based SSL units have **much better quality** than single-stream SSL units.

Experiments – Speech Resynthesis (Vocoder)

SSL (Model)	Single-layer Stream	Multi-layer Stream	Total Streams	MCD	UTMOS
HuBERT	1	1	1	7.19	2.27
Encodec	-	-	8	**3.91**	3.18
HuBERT	2	1	2	6.79	2.89
HuBERT	1	4	4	5.12	3.10
HuBERT (MMM)	2	4	8	4.54	**3.22**

Similar to speech recognition, **complimentary benefits** are observed when single-layer stream and multi-layer stream are used.

Experiments – Speech Resynthesis (Vocoder)

SSL (Model)	Single-layer Stream	Multi-layer Stream	Total Streams	MCD	UTMOS
HuBERT	1	1	1	7.19	2.27
Encodec	-	-	8	**3.91**	3.18
HuBERT	2	1	2	6.79	2.89
HuBERT	1	4	4	5.12	3.10
HuBERT (MMM)	2	4	8	4.54	**3.22**

While MMM-based SSL units **has less matched** spectrogram measured in MCD, it has **slight better perceptual quality** in UTMOS.

Experiments – Text-to-speech

- Experimental Setup
  - Dataset: The official challenge dataset of discrete speech challenge (TTS)
    - LJSpeech
  - Evaluation: mel cepstral distortion, WER from Whisper (Radford et al. 2023) and UTMOS (Saeki et al. 2022)
  - SSL models: HuBERT-base model
  - K-means: whole training set into K=500 clusters
  - Downstream model: discrete VITS model that predicted SSL units + discrete HiFi-GAN vocoder

Experiments – Text-to-speech

SSL (Model)	Single-layer Stream	Multi-layer Stream	Total Streams	MCD	WER	UTMOS
HuBERT	1	1	1	7.19	8.1	3.73
Encodec	-	-	8	**7.01**	7.8	4.01
HuBERT	2	1	2	7.11	8.0	3.79
HuBERT	1	4	4	7.25	7.7	4.06
HuBERT (MMM)	2	4	8	7.15	**7.7**	**4.15**

In text-to-speech, MMM-based framework also shows **steadily benefits** over the single stream case

Experiments – Text-to-speech

SSL (Model)	Single-layer Stream	Multi-layer Stream	Total Streams	MCD	WER	UTMOS
HuBERT	1	1	1	7.19	8.1	3.73
Encodec	-	-	8	**7.01**	7.8	4.01
HuBERT	2	1	2	7.11	8.0	3.79
HuBERT	1	4	4	7.25	7.7	4.06
HuBERT (MMM)	2	4	8	7.15	**7.7**	**4.15**

Compared to Encodec, MMM-based HuBERT units has **better perceptual quality** from both downstream ASR word error rate and UTMOS

Two Enhancement Approaches

- Enhance SSL units for generation tasks
- **Enhance neural codecs for understanding tasks**

Continual Pre-training for Speech Codec LLM

- Adapt neural codecs to downstream tasks
  - Conduct in-context learning via additional continual pre-training step
  - Use speech large-language model pre-training tasks as objectives

Shi, Jiatong et al. “Balancing Speech Understanding and Generation: an Investigation of Continual Pre-training for Codec-based Speech LLM on Translation Tasks.” In Submission.

Continual Pre-training for Speech Codec LLM

Simply use <u>shared embeddings</u> for speech and text tokens

Continual Pre-training for Speech Codec LLM

Continual pre-training with the **<u>next token prediction</u>** objective

Pre-training Data Formulation

- Task-oriented pre-training data
  - Tasks composition:
    - **speech continuation** (next speech token prediction)
    - **language modeling** (next text token prediction)
    - **ASR** (given speech tokens and prompts, predict text tokens)
    - **TTS** (given prompt speaker tokens, text tokens, and prompts, predict speech tokens)
    - **ST** (given speech tokens, predict translated text tokens)
    - **T2ST** (given prompt speaker tokens, text tokens, and prompts, predict speech tokens)

With a focus on **translation** tasks

Experimental Design

- Task-oriented pre-training data
  - 140,000 hours of English and Mandarin speech (collection of open-source datasets), along with
    - Transcriptions
    - Translations
  - Translations are generated from a pre-trained machine translation model if no ground truth is in original sources

Downstream Tasks

- Two examples in downstream tasks with additional fine-tuning
  - **Speech translation (ST)**
    - Datasets: CoVoST2 (En -> Zh & Zh -> En) and GigaST (En -> Zh)
    - Evaluation: BLEU
  - **Speech-to-speech translation (S2ST)**
    - Dataset: GigaST (En -> Zh)
    - Evaluation: ASR-BLEU (with the Whisper-large ASR) and UTMOS

Experimental Results (ST)

Pre-training	CoVoST2 (En -> Zh)	CoVoST2 (Zh -> En)	GigaST (En -> Zh)
No initialization	25.5	5.8	30.4
Text LLM initialization	28.9	9.9	33.2
Speech pre-training	24.8	5.4	33.1
Speech & text pre-training	**33.1**	**16.1**	**37.5**

Significant improvements in speech-to-text translation observed when using **continual pre-training**

Experimental Results (S2ST)

Pre-training	[illegible]	[illegible]	[illegible] (En -> Zh)
No initialization	[illegible]	[illegible]	[illegible].4
Text LLM initial[illegible]	[illegible]	[illegible]	[illegible].2
Speech pre-training	[illegible]	5.4	33.1
Speech & text pre-training	[illegible]	16.1	37.5

For speech-to-speech translation, **without** speech continual pre-training, it is hard to conduct end-to-end training

Pre-training	GigaST (ASR-BLEU)	GigaST (UTMOS)
No initialization	/ (Cannot converge)	/ (Cannot converge)
Text LLM initialization	/ (Cannot converge)	/ (Cannot converge)
Speech pre-training	28.0	3.41
Speech & text pre-training	33.4	3.66

Experimental Results (S2ST)

Pre-training	[illegible]	[illegible]	[illegible]En -> Zh)
No initialization			[illegible].4
Text LLM initial			[illegible].2
Speech pre-training		5.4	33.1
Speech & text pre-training		16.1	37.5

Jointly considering speech and text in continual pre-training can **align the codec token to textual semantics** for end-to-end speech-to-speech translation.

Pre-training	[illegible]ST (ASR-BLEU)	GigaST (UTMOS)
No initialization	[illegible]annot converge)	/ (Cannot converge)
Text LLM initialization	/ ([illegible]annot converge)	/ (Cannot converge)
Speech pre-training	28.0	3.41
Speech & text pre-training	33.4	3.66

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Challenges in Cross-modal Tasks

- Take spoken language understanding as an example:

Challenges in Cross-modal Tasks (Cont'd)

- Previous works addressing modalities mismatch
  - Deliberation modules (Huang et al. 2020; Chung et al. 2021; Arora et al. 2022)
  - Two stage decoding (Lai et al. 2021)
  - K-means clustering for "tokenization" (Lin et al. 2022)
    - Conduct K-means clustering over hidden representations from speech SSL
    - In training/inference, use cluster ID for each speech SSL feature frames as the input to text pre-trained models
    - The mismatch between cluster IDs from speech SSL and the expected input for text SSL necessitates fine-tuning

Challenges in Cross-modal Tasks (Cont'd)

- Previous works addressing modalities mismatch
  - Deliberation modules (Huang et al. 2020; Chung et al. 2021; Arora et al. 2022)
  - Two stage decoding (Lai et al. 2021)
  - K-means clustering for "tokenization" (Lin et al. 2022)

- However, they
  - come with **complicated** **training/inference** procedures
  - need **further fine-tuning** **with supervision** for downstream tasks

Method for Cross-modal Application (Cont'd)

- Previous works to fix the modalities mismatch
  - Deliberation modules (Huang et al. 2020; Chung et al. 2021; Arora et al. 2022)
  - Two stage decoding (Lai et al. 2021)
  - K-means clustering for "tokenization" (Lin et al. 2022)

- However, they
  - come with **complicated** **training/inference** procedures
  - need **further fine-tuning** **with supervision** for downstream tasks

- Proposed method:
  - Connecting representations **unsupervisedly** between two modalities using unsupervised ASR

Shi, Jiatong et al. "Bridging Speech and Text Pre-trained Models with Unsupervised ASR." ICASSP 2023.

Method for Cross-modal Application (Cont'd)

An unsupervised speech recognizer (UASR) is used to connect speech representation extractor and text pre-trained model.

Method for Cross-modal Application (Cont'd)

- **Unsupervised Property**
- **Enhanced textual benefits**
- **Optional Fine-tuning**

Experiments

- Freeze setting
  - Follow the exact setting of SUPERB benchmark
  - Test on
    - SLU tasks: Intent classification (IC), emotion recognition (ER), speech translation (ST)
    - Other tasks: speech recognition (ASR), phoneme recognition (PR), and speaker identification (SID)
- Fine-tuning setting
  - Follow (Lin et al. 2022), fine-tuning the textual model while freezing the other modules
  - Test on SLU tasks
    - SLURP for IC and slot-filling (SF) → accuracy score and slot-type F1
    - IEMOCAP for ER → accuracy score
    - CoVOST2 for ST → BLEU score
    - NMSQA for spoken question answering (SQA) →area overlapping score (AOS) and frame F1 (FF1)

Experimental settings

- Speech SSL: pre-trained wav2vec2 (large version) on LibriLight (Beavski et al. 2020)
- UASR: wav2vec-U 2.0 trained with Fairseq (Liu et al. 2022)
- Pre-trained text model: PhonemeT5 (PT5) (Hsu et al. 2022)
- Model candidates (main results)
  - A: utilizes only the speech SSL
  - B: employs K-means as a connector (follow Lin et al. 2022)
  - C: uses UASR as a connector

Freeze setting

SSL	PR (↓)	ASR(↓)	IC(↑)	ST(↑)	SID(↑)
FBANK	82.01	23.18	9.65	2.23	20.06
wav2vec2-large	5.51	3.79	94.38	13.01	**83.69**
+ UASR	**4.53**	**3.76**	94.33	13.00	82.85
+ UASR + PT5	4.68	3.97	**94.88**	**13.53**	82.74

- **Significant gains** are observed with the proposed method **in IC and ST.**

Fine-tuning setting

	Connector	Text Model	IC(↑)	SF(↑)	ER(↑)	ST(↑)
A	/	/	82.82	65.82	66.35	22.1
B	K-means	PT5	86.41	71.70	72.61	22.5
C	UASR	PT5	**87.10**	**74.03**	**73.57**	**24.3**

- **Using the connector with fine-tuning** can significantly improve **all SLU tasks**
- Using UASR **further improves** the performances compared to the one using K-means

Experiments on SQA

	Connector	Text Model	SQuAD		OOD	
			AOS(↑)	FF1(↑)	AOS(↑)	FF1(↑)
(Lin et al. 2022)	K-means	Longformer	49.1	55.9	/	/
Ours	UASR	PT5	**65.8**	**69.7**	**42.1**	**46.2**

SQuAD and OOD are two test sets for NMSQA dataset.

The proposed method is **significantly better** than the **previous state-of-the-art model** in NMSQA dataset.

Summary

- In this section, we focus on universal task handling for speech representation learning, specifically we:

- Examine approaches to achieve universal speech modeling by

  - **Enhancing SSL-based units** for generation tasks
    - **MMM framework**
  - **Enhancing neural codecs** for understanding tasks
    - **Continual training for speech language modeling**

- Propose **the UASR connector approach** for achieving cross-modal adaptation.

Improving Universality

Outline

1. **Background**
2. **Current Progress**
   1. *Universal Language Processing*
      1. **Evaluation** of multilingual speech representation learning
      2. **Extension** of speech representation to the multilingual world
   2. *Universal Task Handling*
      1. **Enhancement** in universal speech tasks
      2. **Adaptation** to cross-modal tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Proposed Directions

- Audio profiling and descriptive evaluation
  - To meet the growing need for
    - **standardization**
    - **scalability**
    - **practical utility**

  in analyzing and applying utterance-level audio representations

Preparation: VERSA + ESPnet

- VERSA (Versatile Evaluation for Speech and Audio)
  - Targets a general interface for speech and audio evaluation
  - A collection of conventional/recent automatic quality evaluation metrics

Shi, Jiatong et al. “VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music” In submission

Preparation: Metrics in VERSA

Dependency	Example Evaluation Metrics
Dependent	MCD, F0-RMSE, F0-CORR, SI-SNR, CI-SDR, PESQ, STOI, WARPQ, VISQOL, SpeechDiscreteBLEU, SpeechDiscreteDistance, SpeechBERT, etc.
Independent	DNSMOS, UTMOS, PLCMOS, SingMOS, Torch-squim (PESQ, STOI, SI-SNR), Sheet-SSQA, PAM, SpoofS, SWR, etc.
Non-matching	CER/WER (ESPnet/ESPnet-OWSM, Whisper),, CLAP-score, Log-WMSE, EMO-SIM, APA, LLR, NOMAD, LLR, etc.
Distributional	FAD, KID, Coverage, Density, KLD

Check full list and details about metrics in the paper!

Comparison to existing open-source toolkits

Toolkit	Github Repo	Evaluation-Focus	Supported Metrics
ESPnet	espnet/espnet	❌	16
Amphion	open-mmlab/Amphion	❌	15
Stable-Audio-Metrics	Stability-AI/stable-audio-metrics	✅	3
AudioLDM-Eval	haoheliu/audioldm_eval	✅	9
VERSA	shinjiwlab/versa	✅	**63 (711 variants)**

Compared to other open-source toolkits, VERSA is dedicated to speech and audio evaluation with **extensive integration** of both conventional and recent evaluation metrics.

(Check our paper for a full comparison at https://arxiv.org/abs/2412.17667)

Proposed Directions – the Task of Uni-VERSA

- A unified framework for utterance-level representation, including
  - Benchmarking for evaluating utterance-level representation
  - Pseudo-labeling framework for learning utterance-level representation
  - With extension to universal domains → speech, music and general audio

Proposed Directions – the Task of Uni-VERSA

- A unified framework for utterance-level representation, including

Timeline

- Jan. 2025: Thesis Proposal
- Feb. 2025 – May. 2025: Designing the Uni-VERSA Framework for Speech, Music, and General Audio
- May. 2025 – Sep. 2025: Improving the Uni-VERSA Framework
- Sep. 2025 – Dec. 2025: Thesis Writing

Summary of Contributions

Related papers (1st author or equally contributed)

1. Shi, Jiatong et al. “ML-SUPERB: Multilingual Speech Universal PERformance Benchmark.” Interspeech 2023.
2. Shi Jiatong et al. “Findings of the 2023 ML-SUPERB Challenge: Pre-Training and Evaluation over More Languages and Beyond.” ASRU 2023.
3. Shi, Jiatong et al. “ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets.” Interspeech 2024.
4. Berrebbi, Dan et al. ”Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation” Interspeech 2022.
5. Srivastava, Tejes et al. “FFUSE: Efficient Self-Supervised Feature Fusion for E2E ASR in Multilingual and Low Resource Scenarios” Interspeech 2024.
6. Shi, Jiatong et al. “Exploration on HuBERT with Multiple Resolution” Interspeech 2023.
7. Shi, Jiatong et al. “Multi-resolution HuBERT: Multi-resolution Speech Self-Supervised Learning with Masked Unit Prediction.” ICLR2024.
8. Shi, Jiatong et al. “MMM: Multi-Layer Multi-Residual Multi-Stream Discrete Speech Representation from Self-supervised Learning Mode.l” Interspeech 2024.
9. Wu, Yuning et al. “TokSing: Singing Voice Synthesis based on Discrete Tokens.” Interspeech 2024.
10. Shi, Jiatong et al. “Balancing Speech Understanding and Generation: an Investigation of Continual Pre-training for Codec-based Speech LLM on Translation Tasks.” In Submission.
11. Shi, Jiatong et al. “Bridging Speech and Text Pre-trained Models with Unsupervised ASR.” ICASSP 2023.
12. Shi, Jiatong et al. “Enhancing Speech-to-Speech Translation with Multiple TTS Targets” ICASSP 2023.
13. Shi, Jiatong et al. “VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music” In submission.

Summary of Contribution (Cont'd)

- Open-source activities
  - Most of the discussed works are open-sourced
  - Toolkits:
    - ESPnet tasks
      - Diarization, ESPnet-ST-V2, EURO (unsupervised ASR), ESPnet-Codec, ESPnet-Muskits (singing voice synthesis), ESPnet2-TTS, ESPnet-SPK
    - VERSA
    - S3PRL
      - Focusing on speech self-supervised representation
    - ParallelWaveGAN
      - Focusing on vocoder modeling (especially discrete-based vocoder)
    - AudioGPT
      - Focusing on centralized demonstration of downstream speech/music tasks
    - Fairseq
      - Multiresolution HuBERT
  - Dataset
    - Speech related: ML-SUPERB, Totonac Nahuatl, Yoloxochitl Mixtec, Highland Puebla Nahuatl
    - Singing voice related: SingMOS dataset, SVDD challenge data (CtrSVDD), ACE-Opencpop and ACE-KiSing, KiSing

Thanks for your attention!

Unless otherwise specified, figures are created by the author. Icons are sourced from https://www.flaticon.com/ or generated by DALL-E