Carnegie Mellon University

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. **Universality for Speech Representation Learning**
2. **Current Progress**
    1. *Universal Language Processing*
        1. **Analysis** of multilingual speech representation learning
        2. **Extension** of speech representation to the multilingual world
    2. *Universal Task Handling*
        1. **Enhancement** in universal speech 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 long time

- Early works follow the study in psychoacoustics

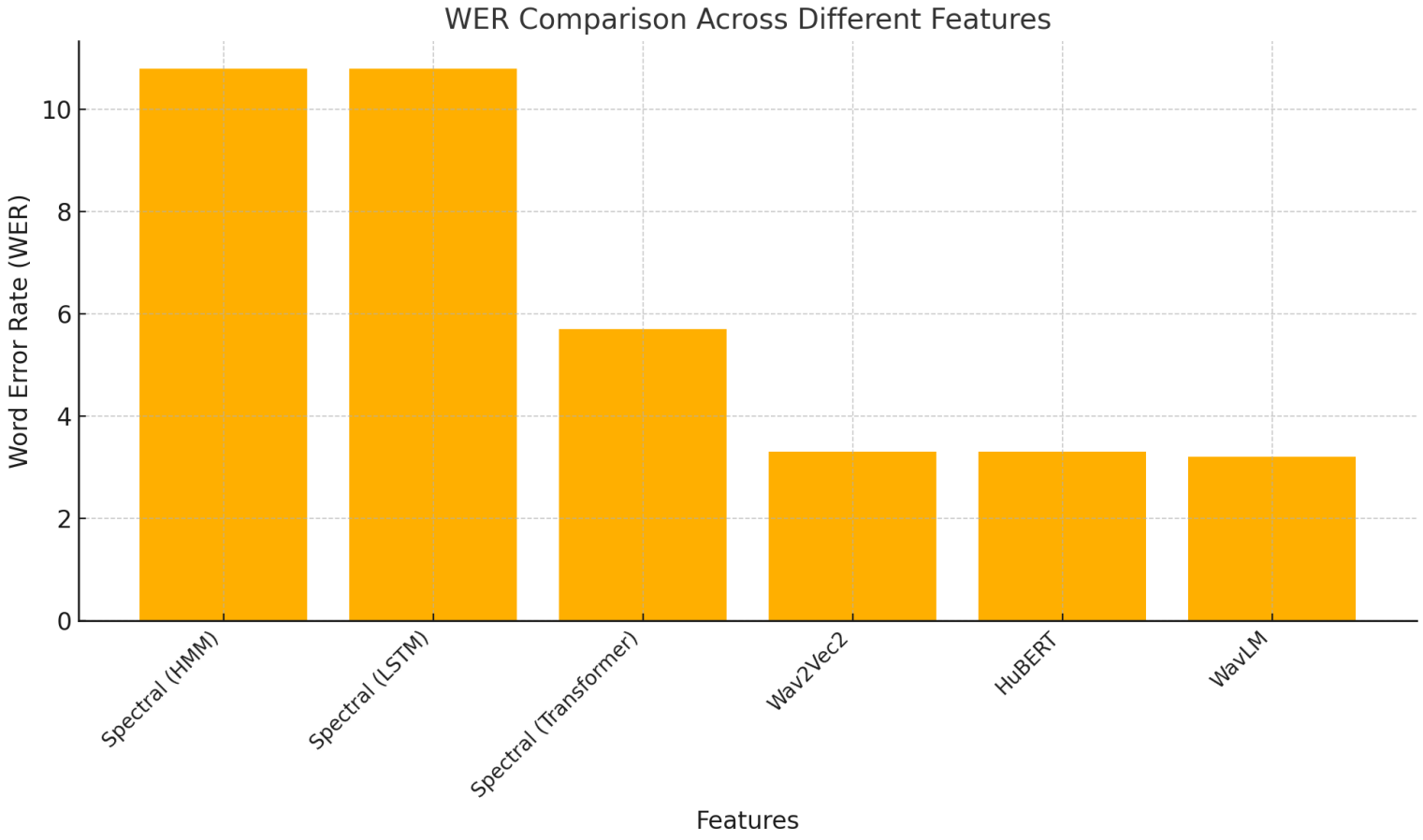
  - Human ear is a frequency acceptor that is stimulated based on the energy at related frequency bands

  - Can we align the representation to a format that human are accustomed to processing?
    - Use **Fourier Transform** to extract speech representation!

From Spectral Features to Neural Representations

- A 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)

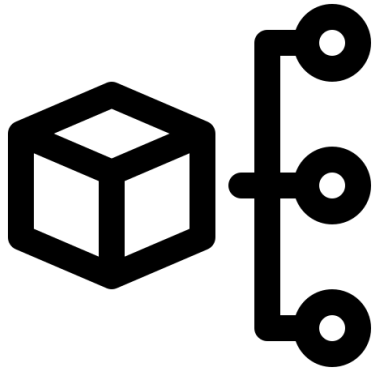
LibriSpeech (Test-other set)

(Wang et al. 2022, WavLM)

Significant improvements after using neural representations

Categorization of Speech Representation

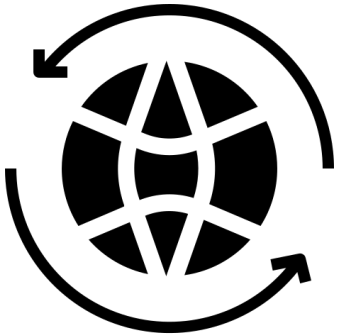
- Define a single-channel speech signal as $\mathbf{S} = [s_1, s_2, ..., s_{L_s}] \in \mathbb{R}^{1 \times L_s}$
  - Denote the representation extraction model as $f(\cdot)$
  - Continuous Frame-wise $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T]$
  - Discrete Frame-wise $\mathbf{C} =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



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

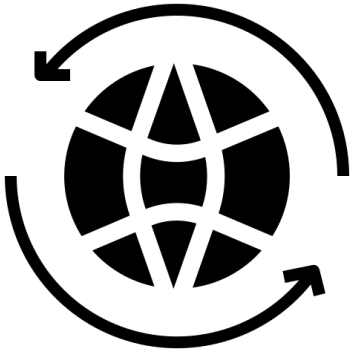
Challenges in Speech Representation

- 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

- Challenges posed by the rich diversity of the world's languages
  - Phonetic complexity and rare sounds
  - Tonal and prosodic variations
  - Dialectal and accented speech variability
  - Low-resource and endangered languages
  - Cross-language speech-to-speech tasks

Universal Language Processing

- Challenges posed by the rich diversity of the world's languages
  - Phonetic complexity and rare
  - Tonal and prosodic variation
  - Dialectal and accented spee
  - Low-resource and endangered languages
  - Cross-language speech-to-speech tasks

Establishing Shared Representation:
a **universal "language"** for speech modeling

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 Representation for Multiple Tasks

- Universal representation: a common “language” for speech modeling
- Transitioning to Multiple Tasks
  - Separate models or unified models ?
  - With shared representation to skip the challenges in starting from scratch or learn separate feature sets

Universal Representation for Multiple Tasks

- Shared representation
- Transitioning to multiple
  - Separate models or unif
  - With shared representat
    learn separate feature se

- Perform **speech recognition** by decoding from shared speech representations
- **Generate speech** by conditioning on the same universal features that represents diverse languages
- Enable **seamless speech-to-speech translation** by directly transferring information through the shared representation without intermediate language-specific steps

Universal Representation for Multiple Tasks

- Shared representation
- Transitioning to multiple
  - Separate models or unif
  - With shared representat
    learn separate feature se

- Achieving Efficient Development and Enhancing Generalization
- Facilitating Transfer and Reuse
- Achieving Consistency and Efficiency

Speech Tasks

- Understanding focus:
  - speech recognition, spoken language understanding, speech translation
- Generation focus:
  - speech synthesis, voice conversion, singing voice synthesis, speech enhancement

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:

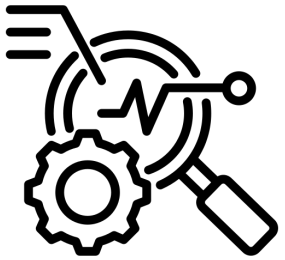
$$\text{argmax}_{\theta_{\text{und}},\theta_{\text{gen}},f_{\text{uni}}(\cdot)}[\mathbb{E}_{\text{und}}[\log p_{\theta_{\text{und}}}(\mathbf{Y}_{\text{und}}|\mathbf{X}_{\text{und}}, f_{\text{uni}}(\cdot))] + \mathbb{E}_{\text{gen}}[\log p_{\theta_{\text{gen}}}(\mathbf{Y}_{\text{gen}}|\mathbf{X}_{\text{gen}},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

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

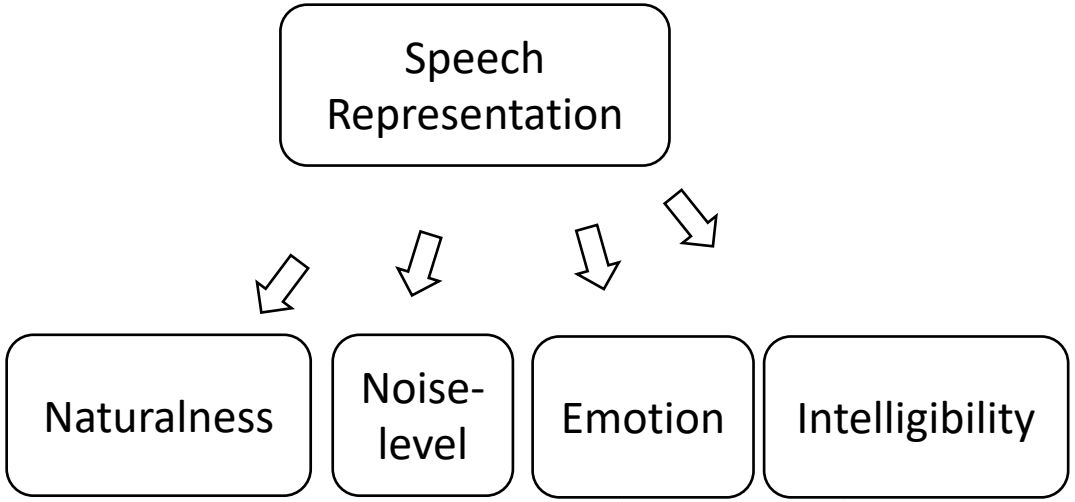
To Evaluate Speech Representation

- Evaluation with the task-specific transformation
- Universal language processing and universal task handling:
  - Targets for representations that are general, flexible, and capable of serving multiple tasks
  - Foundational feature for tasks: multilingual speech recognition, multilingual speech synthesizer, speech-to-speech translation etc.

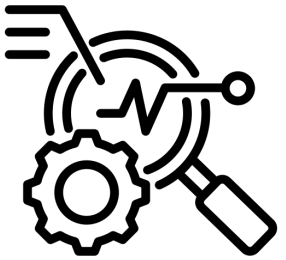
Towards Universal Evaluation for Speech



- Utilizing a unified framework to evaluate the speech representation on
  - Naturalness
  - Intelligibility
  - Noise Level
  - Emotion
  - …



Universal Evaluation Framework



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

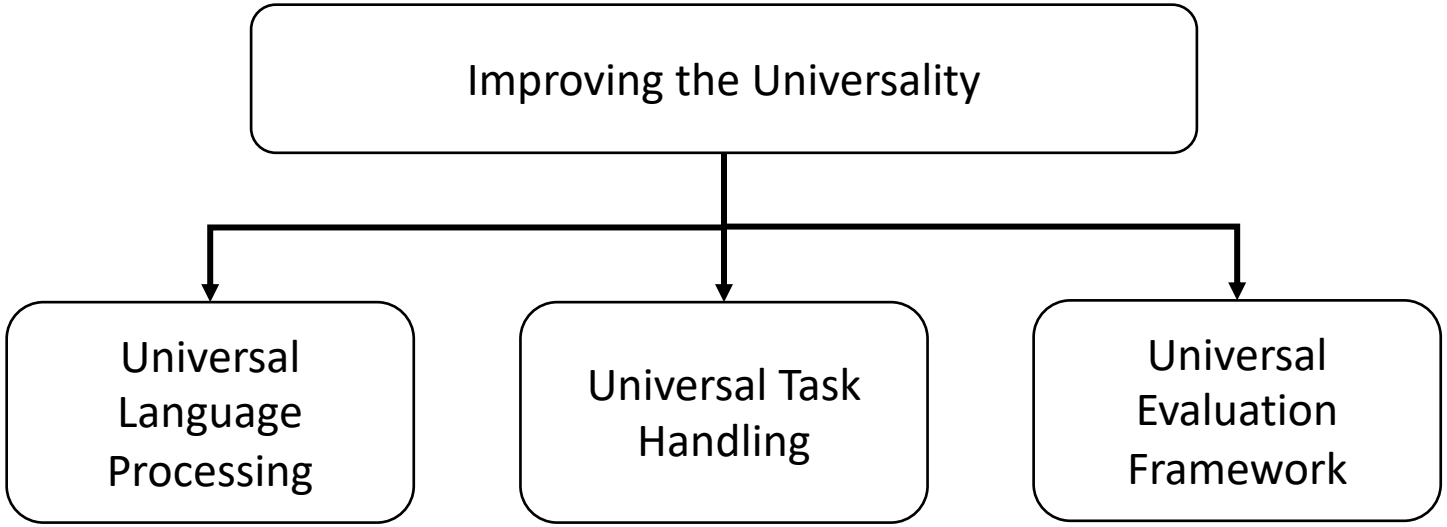
- The focus of single-domain is:

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

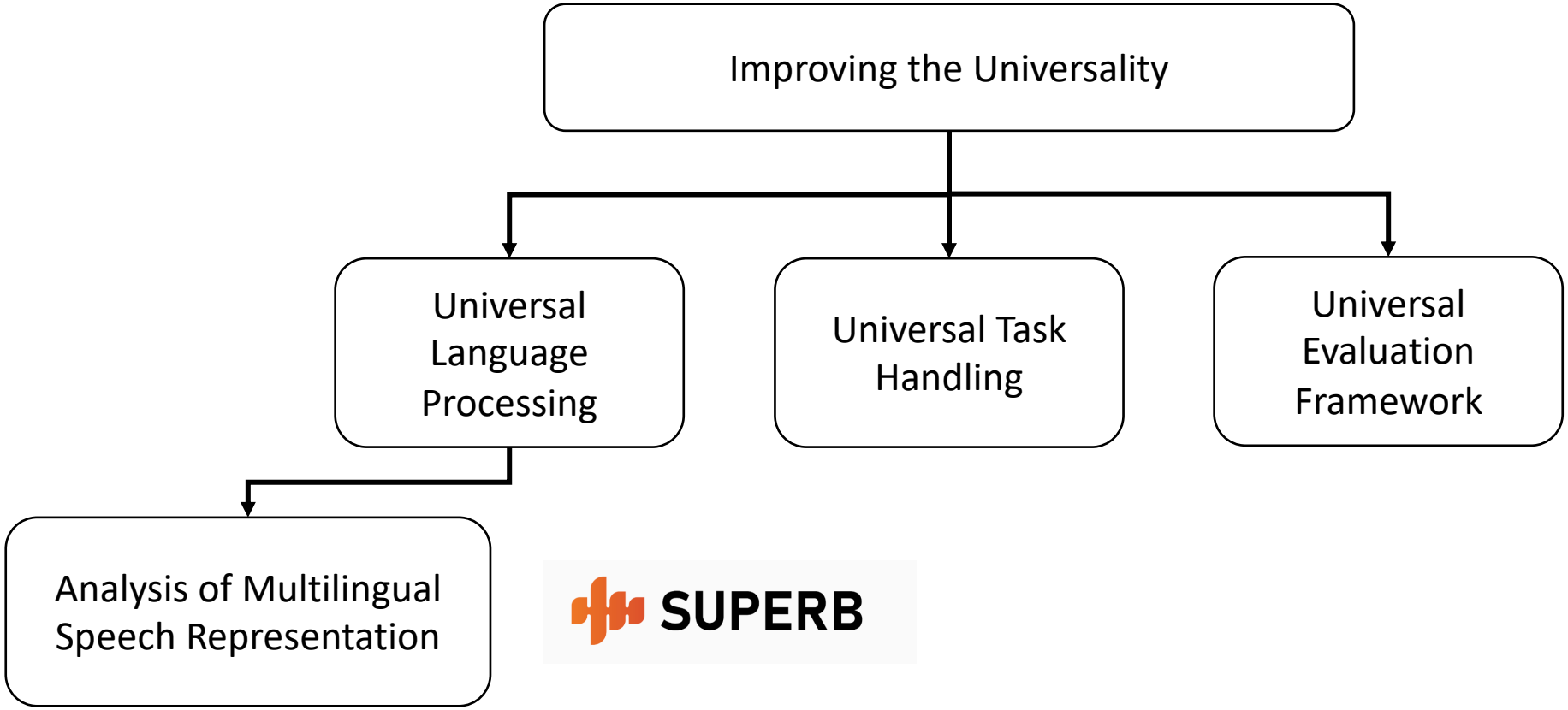
- The target of the problem is to:

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

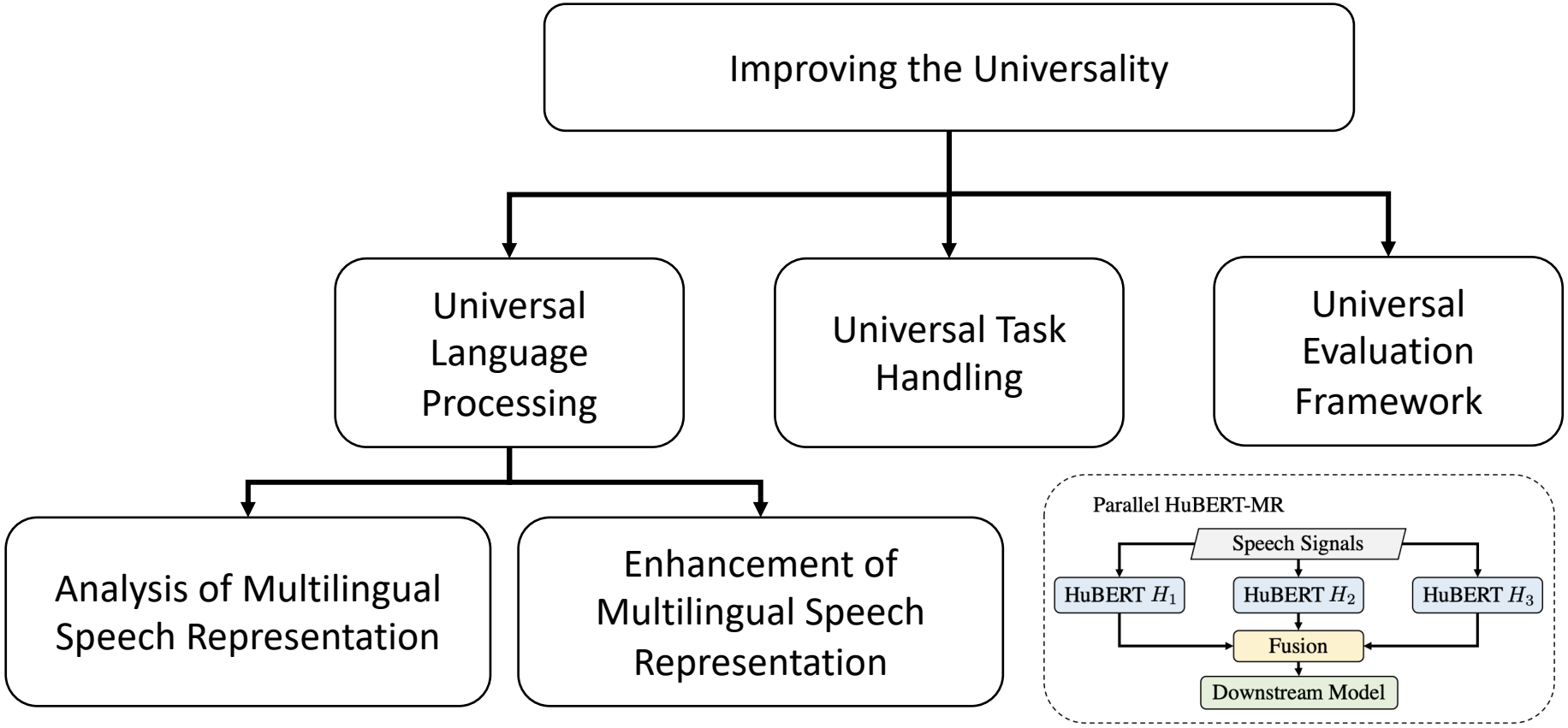
Improving Universality



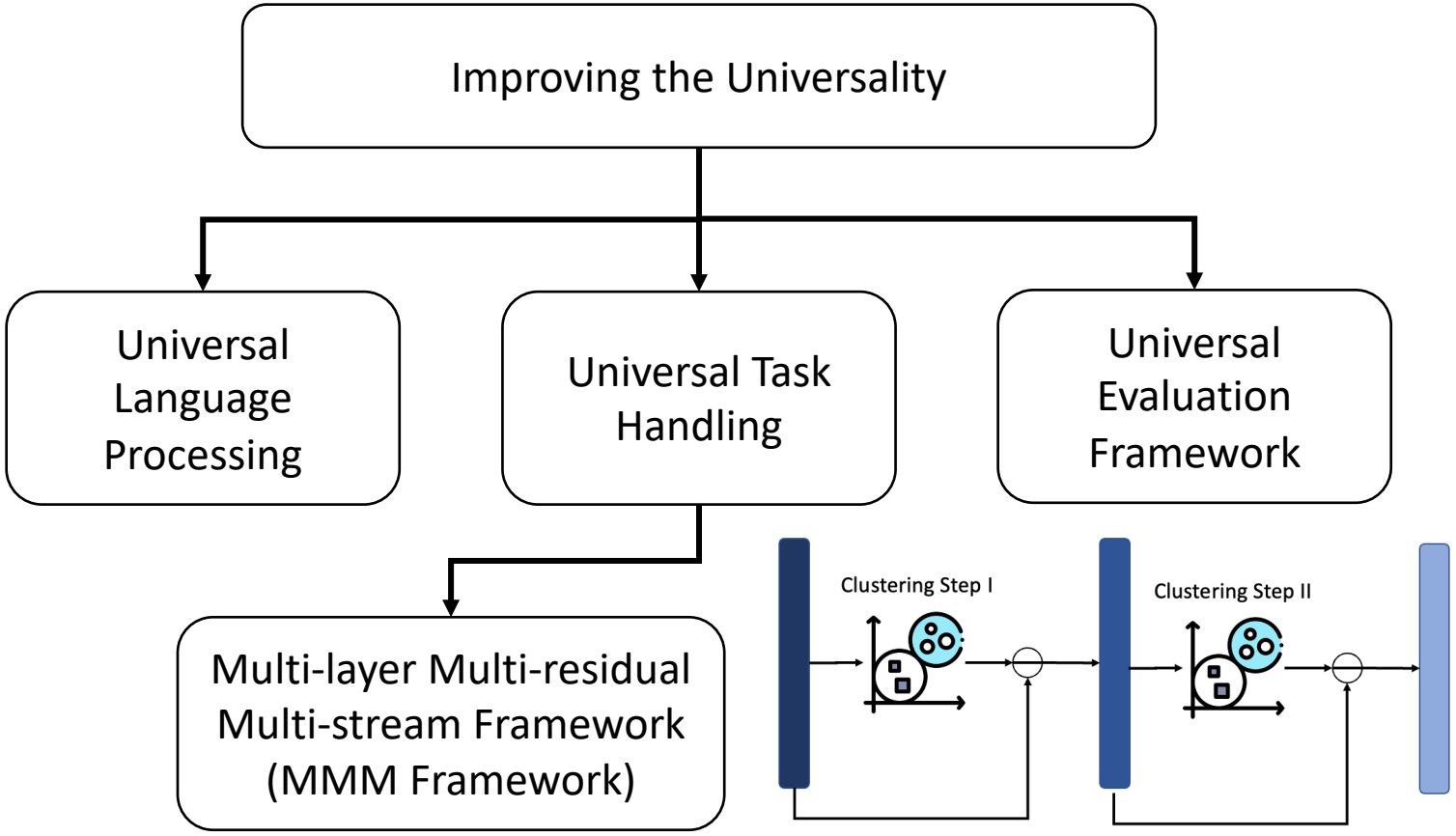
Improving Universality



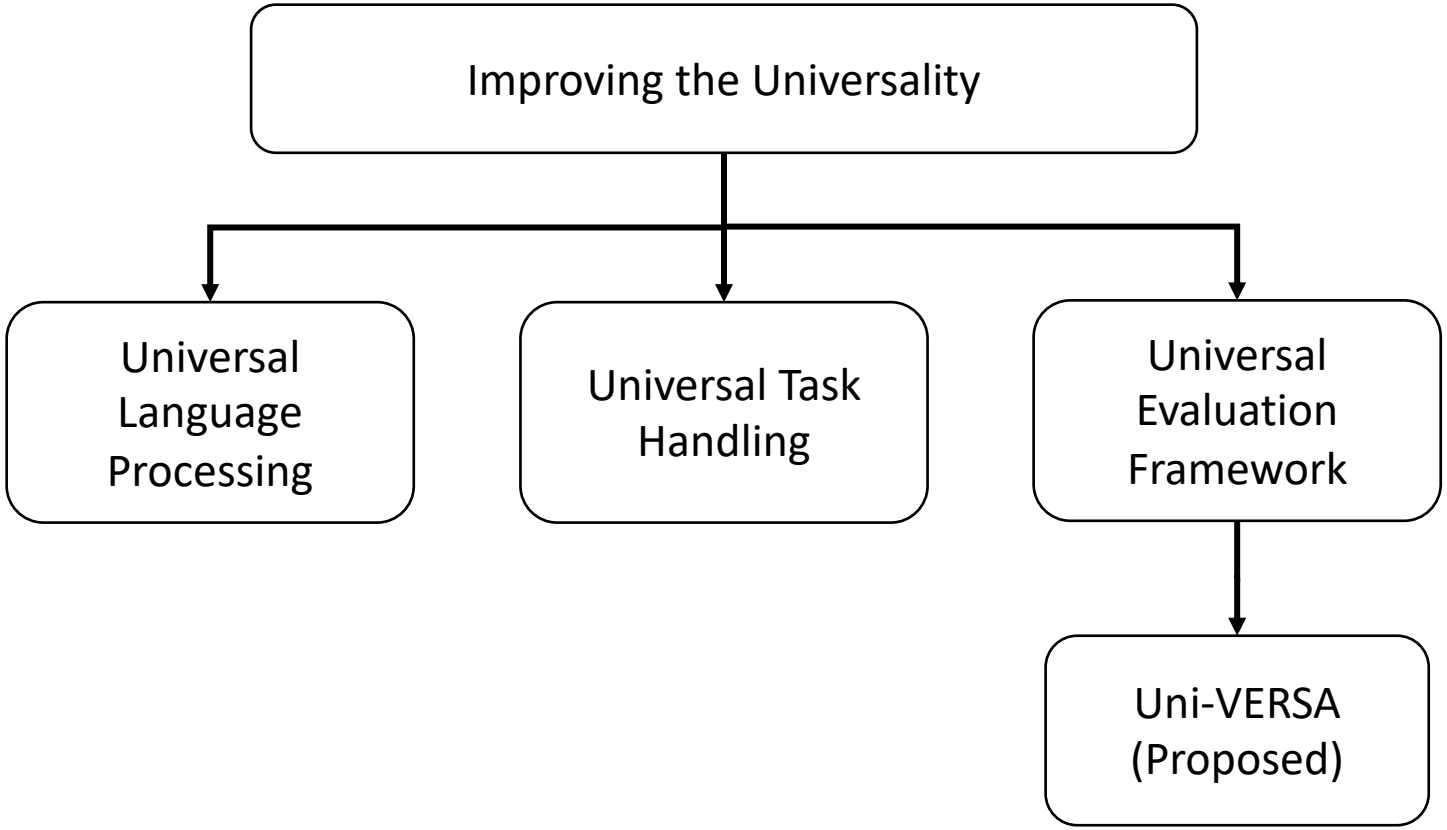
Improving Universality



Improving Universality



Improving Universality

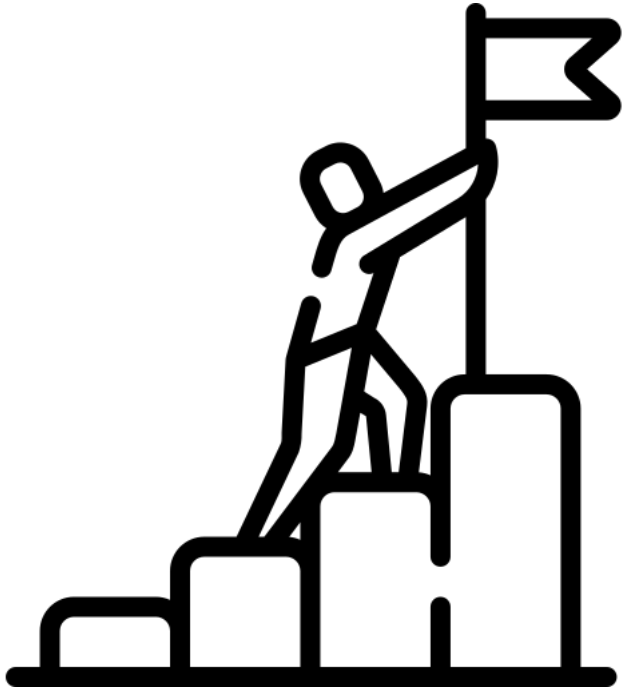


Outline

1. **Universality for Speech Representation Learning**
2. **Current Progress**
    1. *Universal Language Processing*
        1. **Analysis** of multilingual speech representation learning
        2. **Extension** of speech representation to the multilingual world
    2. *Universal Task Handling*
        1. **Enhancement** in universal speech tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

Universal Language Processing

- Analysis of speech representation
  - Not explicitly related to a specific task
    - Check the transcription from speech recognition
    - Check the signal quality from speech enhancement/separation
    - Check the generated speech for speech synthesis
  - One of the key challenges of speech representation learning



Universal Language Processing

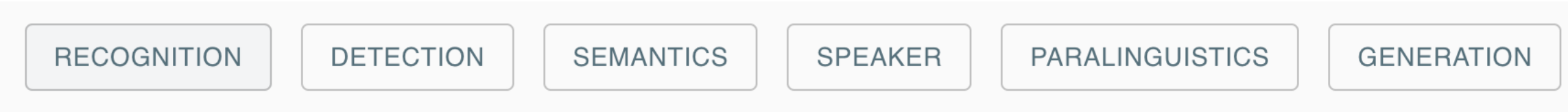
- Analysis of speech representation

(Yang et al. 2021, SUPERB)

  - Speech Universal PERformance Benchmark (SUPERB)

    - A framework to evaluate speech representation by using downstream tasks

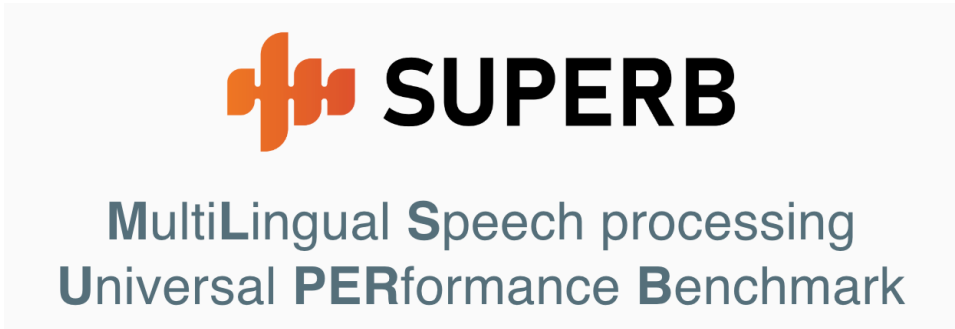
    - A wide-coverage of different tasks in:



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:



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

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 Analysis

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

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

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

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

- Large models are not **necessarily 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

- **Robustness** is also important to consider

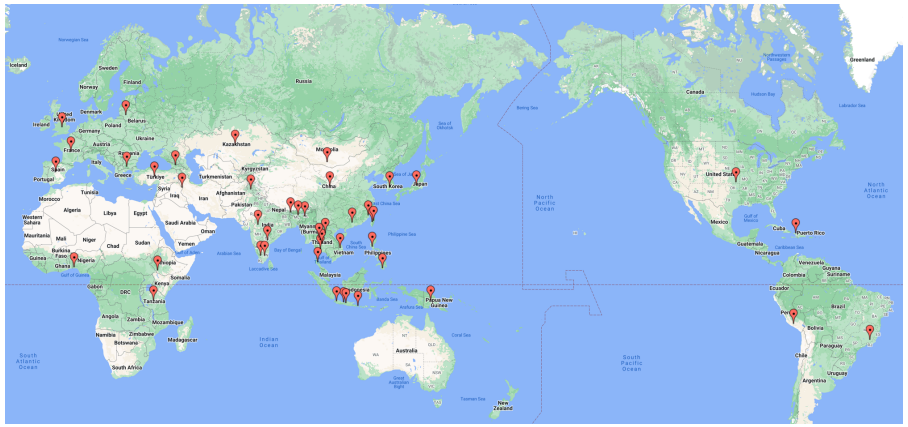
ML-SUPERB Results (10min)

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						0.3	598.3
robust-wav2vec2	44.4						4.9	680.3
wav2vec2-base-23	49.2						4.3	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 has served as an **efficient and comprehensive** framework of analyzing speech representation, using by more than 20 recent multilingual speech representation works.

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)

- **Enhanced evaluation:**

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

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

  - 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

Effect of Introducing Four Scenarios

Scenarios	Details	Accuracy	CER (Normal)
Original ML-SUPERB			3
Large Downstream			3
SSL Model Fine-tuning			3
Efficient Model Adaptation			5
Supervised Pre-trained Mo			5

Further extend to **diverse practical use cases** for speech representation learning by focusing on the downstream performance

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

Outline

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

Multi-resolution for Multilingual Speech Processing

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

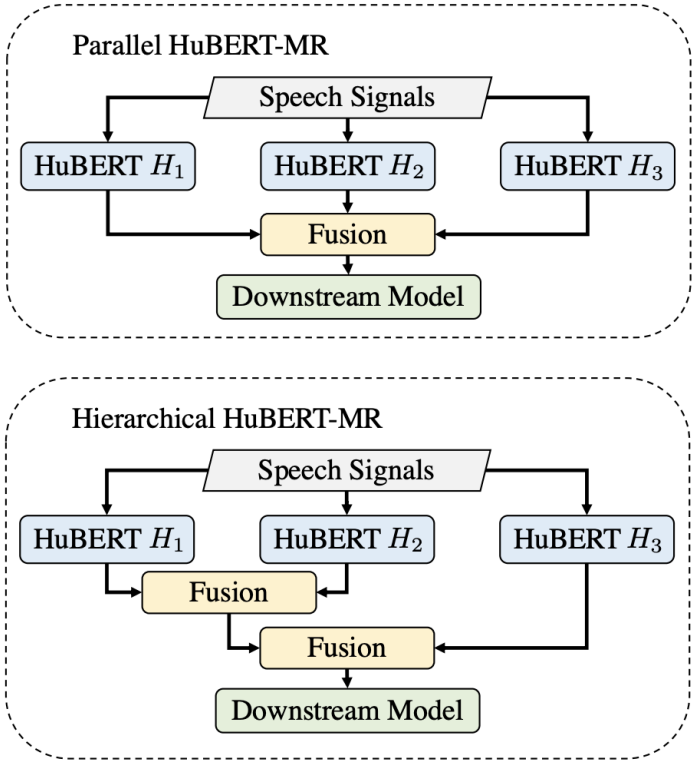
- The local stability
  - Information of speech is in **local sequences** of samples
  - Feature extractions happens to **downsample** speech into a sequence of vectors to represent locational information.
    - Conventional: Linear spectrogram, Mel spectrogram, Mel cepstral coefficients (10-30ms)
 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

Multi-resolution for Multilingual Speech Processing

- 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)

Switching to Multilingual Scenario

- With a pre-trained model trained on Commonvoice (60k hours, 96 languages)
  - A model with the base size 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
HuBERT-MR	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)

Switching to Multilingual Scenario

- With a pre-trained model trained on Commonvoice (60k hours, 96 languages)
  - A model … uBERT-based models t…

Continuous evaluation demonstrates great potential of using the multi-resolution framework for multilingual representation learning.

SSL Model	Monoli… ASR					…shot		SUPERB Score
	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
HuBERT-MR	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)

Summary

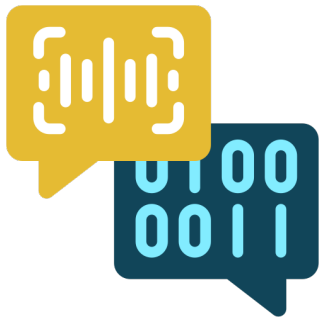
- In this section, we focus on universal language processing for speech representation learning, specifically we:
  - Present the **multilingual SUPERB (ML-SUPERB)** for analyzing speech representation in multilingual scenarios
  - Adopt the **multi-resolution framework** to enhance the speech representation learning framework for multilingual processing

Outline

1. **Universality for Speech Representation Learning**
2. **Current Progress**
    1. *Universal Language Processing*
        1. **Analysis** of multilingual speech representation learning
        2. **Extension** of speech representation to the multilingual world
    2. *Universal Task Handling*
        1. **Enhancement** in universal speech tasks
3. **Proposed work** – Universal Evaluation Framework
4. **Timeline**

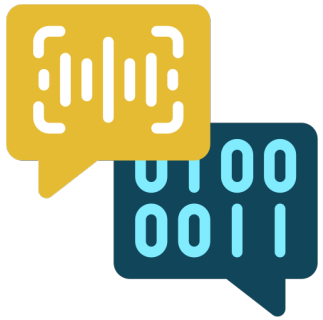
Achieving Universal Task Handling

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



Achieving Universal Task Handling

- Transitioning to discrete representation
  - Could potentially a possible approach to achieve **universal task handling**
    - Continuous features are **difficult** to use for the generation tasks, when using as the predicting targets



Transitioning to Discrete Representation

- **Scaling up**
  - TTS with VALL-E (Wang et al. 2023)
- Improved storage efficiency
- Enhanced efficiency in training
- Enhanced efficiency in inference
- Potential for integration with various modalities

Transitioning to Discrete Representation

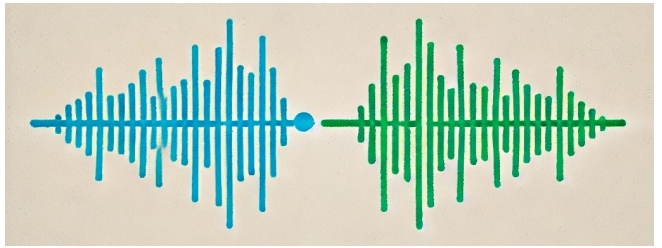
- Scaling up
- **Improved storage efficiency**
  - Discrete ASR (Chang et al. 2023)
- Enhanced efficiency in training
- Enhanced efficiency in inference
- Potential for integration with various modalities

Transitioning to Discrete Representation

- Scaling up
- Improved storage efficiency
- **Enhanced efficiency in training**
  - singing synthesis with TokSing (Wu et al. 2024)
- Enhanced efficiency in inference
- Potential for integration with various modalities

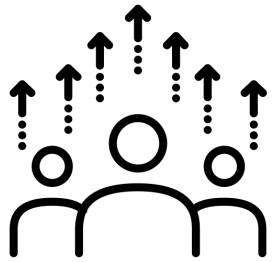
Transitioning to Discrete Representation

- Scaling up
- Improved storage efficiency
- Enhanced efficiency in training
- **Enhanced efficiency in inference**
  - speech enhancement with CodecFormer (Yip et al. 2024)
- Potential for integration with various modalities



Transitioning to Discrete Representation

- Scaling up
- Improved storage efficiency
- Enhanced efficiency in training
- Enhanced efficiency in inference
- **Potential for integration with various modalities**
  - Multimodal LLM with AnyGPT (Zhang et al. 2024)



Discrete Speech Representation

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

Discrete Speech Representation

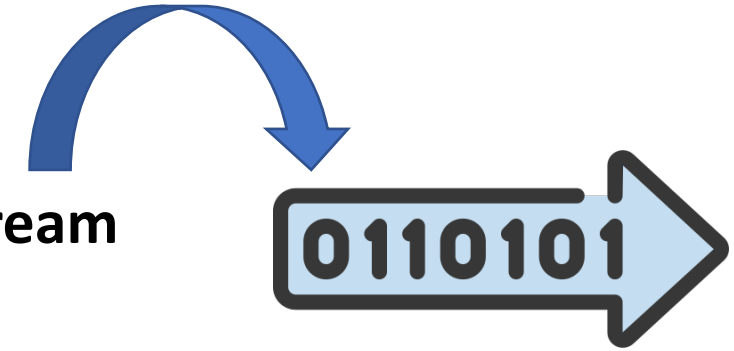
- Two mainstream approaches for discrete speech representation
  - SSL-based units
    - Pros: A good tradeoff between efficiency and effectiveness
    - Cons: Worse performance than continuous SSL representation and less detailed acoustics information for speech generation.

Discrete Speech Representation

- 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)

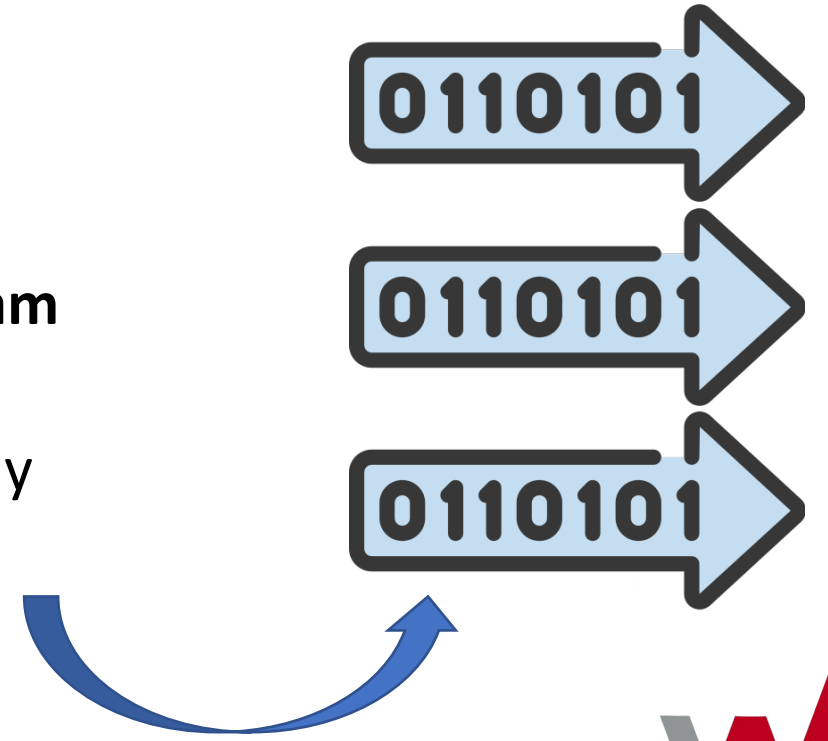
SSL Units and Neural Codec

- Notably
  - SSL units typically operate within a **single-stream**
  - Neural codecs are in **multi-stream** naturally



ISSL Units and Neural Codec

- Notably
  - SSL units typically operate in a **single-stream**
  - Neural codecs are in **multi-stream** naturally



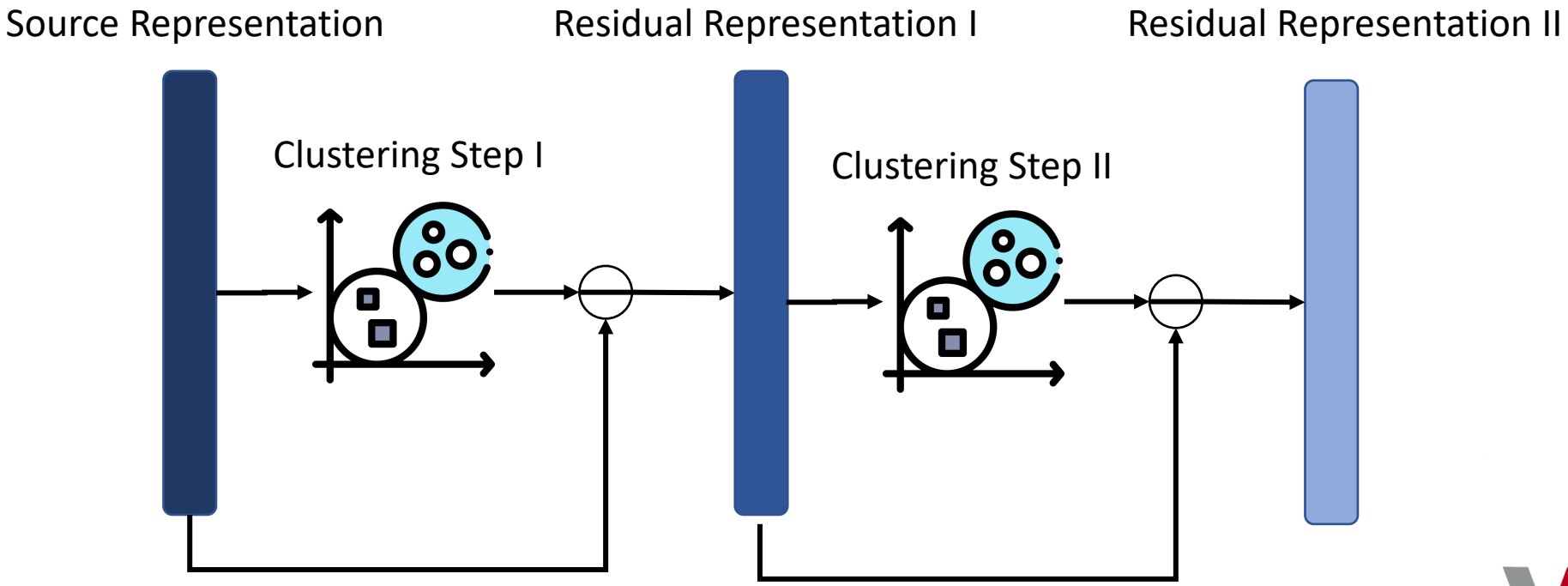
MMM Framework for SSL Units

- 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.

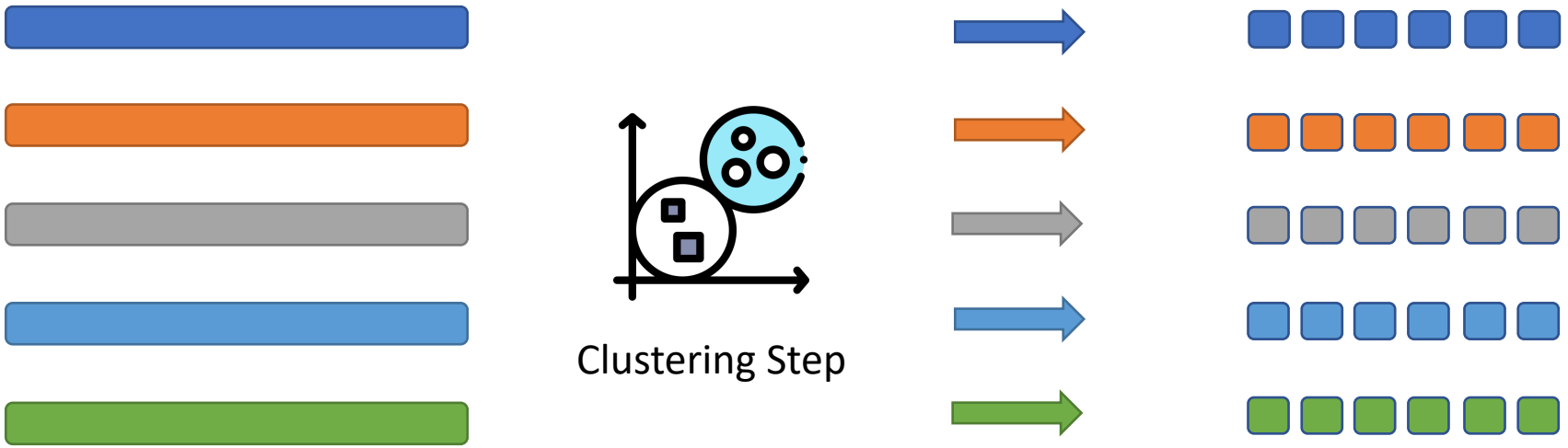
Multi-stream Unit Extraction (Single Layer)

- Iteratively residual clustering over single-layer SSL representation



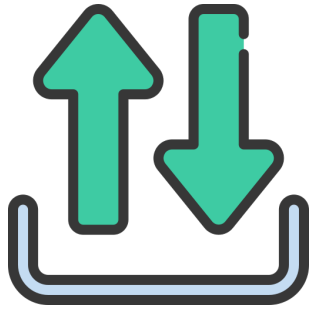
Multi-stream Unit Extraction (Multi-layer)

- Clustering over multi-layers



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



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
    - E-branchformer-based encoder-decoder architecture

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

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**

Outline

1. **Universality for Speech Representation Learning**
2. **Current Progress**
    1. *Universal Language Processing*
        1. **Analysis** of multilingual speech representation learning
        2. **Extension** of speech representation to the multilingual world
    2. *Universal Task Handling*
        1. **Enhancement** in universal speech 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
    - 63 types of metrics with over 700 variants

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

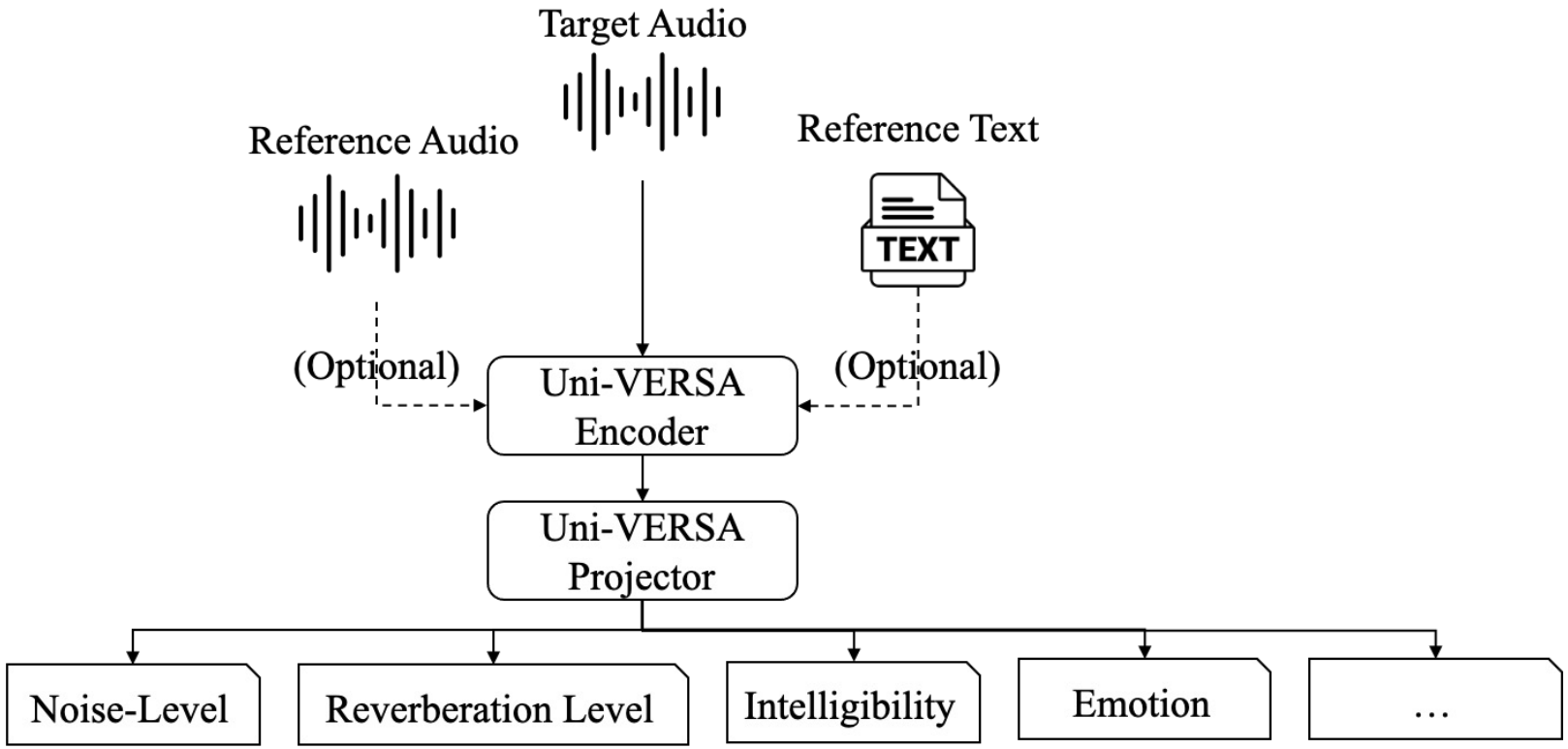
Preparation: Metrics in VERSA

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!

Proposed Directions – the Task of Uni-VERSA

- A unified framework for utterance-level representation



Proposed Directions – the Task of Uni-VERSA

- A unified framework for utterance-level representation, including focuses of
  - **Benchmarking for analyzing speech representation in utterance-level**

- Similar to SUPERB, we target to **various tasks (domains)** for utterance-level representation
- Focus on the ability of achieving multi-domain evaluation **simultaneously**
  - Noise-level
  - Naturalness
  - Reverberation level
  - Speaker identity
  - Emotion
  - …

Proposed Directions – the Task of Uni-VERSA

- A unified framework for utterance-level representation, including
  - **Pseudo-labeling framework for learning representation**

- **Large-scale resources for representation learning**
  - A **comprehensive coverage of pseudo labels** for any audio databases (based on **VERSA**)
  - A collection of the **released databases from recent challenges**
    - VoiceMOS Challenge and Spoofceleb on MOS prediction with TTS/Voice Conversion systems
    - Urgent Challenge on speech enhancement systems
    - Singing Voice Deepfake Detection Challenge (SVDD challenge) with singing voice synthesis/conversion systems
    - …

Proposed Directions – the Task of Uni-VERSA

- A unified framework for utterance-level representation, including
  - **With extension to universal domains → speech, music and general audio**

- Compared to speech related works, music and general audio has **their own difficulties** in evaluating the generated signals
  - Issues with measuring creativity
  - Issues with low-resource data

- Based on the enough knowledge in speech representation, we **can further extend** our research to music and general audio domains for universality in the evaluation.

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

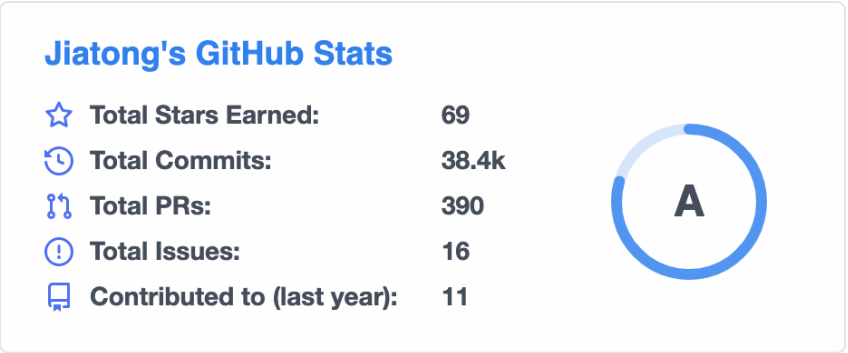
Summary of Contributions

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. Shi, Jiatong et al. “Exploration on HuBERT with Multiple Resolution” Interspeech 2023.
5. Shi, Jiatong et al. “MMM: Multi-Layer Multi-Residual Multi-Stream Discrete Speech Representation from Self-supervised Learning Mode.l” Interspeech 2024.
6. Wu, Yuning et al. “TokSing: Singing Voice Synthesis based on Discrete Tokens.” Interspeech 2024.
7. Shi, Jiatong et al. “Enhancing Speech-to-Speech Translation with Multiple TTS Targets” ICASSP 2023.
8. 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