Universal Representation: Modeling, Application and Evaluation

Jiatong Shi

Carnegie Mellon University, Language Technologies Institute

jiatongs@cs.cmu.edu

Content

- Brief Summary
- Research Roadmap
- Future Envision

Education

- Carnegie Mellon University (2021 – present)
  - PhD in Language Technologies Institute
  - Advisor: Dr. Shinji Watanabe
- Johns Hopkins University (2019 – 2021)
  - MS. in Computer Science
  - Advisor: Dr. Shinji Watanabe
- Remin University of China (2015 – 2019)
  - BS. in Computer Science + BA. in Fintech
  - Advisor: Dr. Qin Jin, Dr. Wei Xu, and Dr. Wei Du

Diverse Experiences

- 77 publications in top speech/NLP/ML journals/conferences
  - 20 first-authored papers, 2600+ citations
  - Mentored 17 students with 16 first-authored papers
  - Extensive experiences in **wide-range** of speech tasks
    - Automatic Speech Recognition (ASR)
    - Text-to-speech (TTS)
    - Speech Translation (ST)
    - Speech-to-speech Translation (S2ST)
    - Spoken Language Understanding (SLU)
    - Speech coding (SC)
    - Speaker Diarization (SD)
    - Singing voice synthesis (SVS)

Diverse Experiences (Cont'd)

- **Open-source** contribution
  - Maintainer and major contributor in
    - ESPnet (maintainer, 8.5k stars)
  - Key-feature contributors in
    - S3PRL (2.3k stars)
    - ParallelWaveGAN (1.6k stars)
    - AudioGPT (10k stars)
    - Fairseq (30.5k stars)

- **Benchmark** and **Challenge** Organizers
  - Speech Universal PERformance Benchmark (SUPERB) [SLT2022]
  - Multilingual SUPERB [ASRU 2023 & Interspeech 2025]
  - Discrete speech challenge [Interspeech 2024]
  - Singing voice conversion challenge [ASRU2023]
  - Simultaneous speech translation challenge [IWSLT 2022-2024]
  - Singing voice deepfake detection challenge [SLT2024]

Notable Achievement

- 2024
  - Best paper award at Interspeech 2024
  - Honorable mention demo award at ACMMM 2024
  - Best paper honorable mention at Responsible speech workshop at Interspeech 2024
  - 1st Place at discrete speech challenge (SVS track)
  - 2nd Place at discrete speech challenge (ASR track)
- 2023
  - Best paper finalist at 2023 ASRU
- 2022
  - Best paper finalist at 2022 SLT
  - Best project at 2022 SLT Hackathon (as a mentor)
  - CMU Presidential Fellowship
  - 1st Place at IWSLT 2022 (dialectal speech translation track)
  - 7th Place at AI Song Contest 2022

Research Topic: Universal Representation

- General concept:
  - Towards an **eco-system** for universal representation for various speech tasks

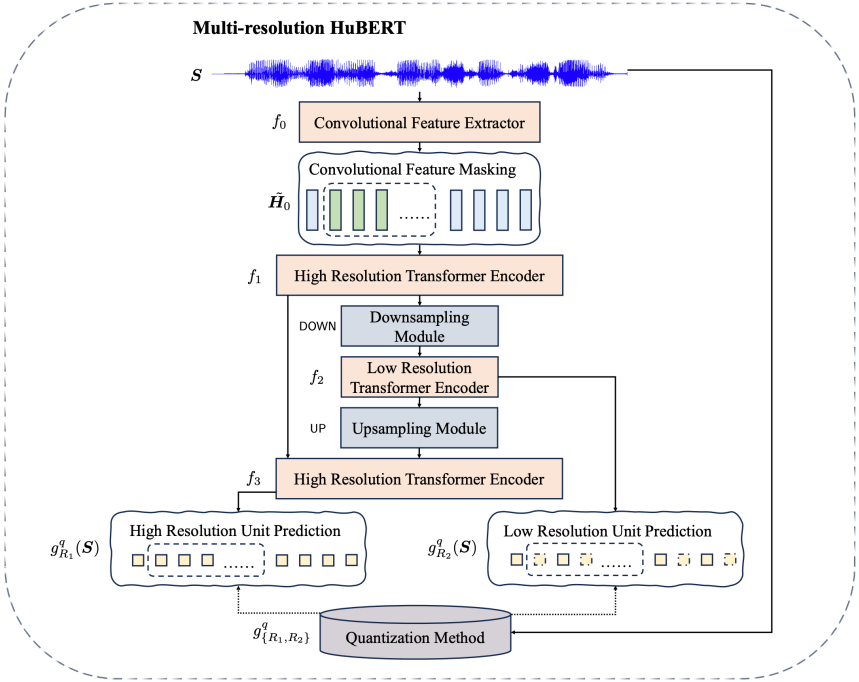
- Major components:
  - Modeling
    - Multi-resolution modeling (MR-HuBERT, SingOMD)
    - Multi-level modeling (MMM, ESPnet-Codec, TokSing)
  - Application
    - Speech recognition, text-to-speech, speech-to-text/speech translation, speech coding, spoken language understanding, singing voice synthesis
  - Evaluation
    - Representation evaluation (SUPERB, ML-SUPERB)
    - Spoken language model (Dynamic SUPERB)
    - Generative audio/speech/music evaluation (VERSA, VERSA-v2)

Universal Representation (Modeling)

- Go universal with multi-level multi-resolutional processing

MR-HuBERT (ICLR2024, Spotlight)

- Use multi-resolution architecture for speech self-supervised learning



Universal Representation (Modeling)

- Go universal with multi-level multi-resolutional processing

Multi-resolution HuBERT [ICLR2024, Spotlight]

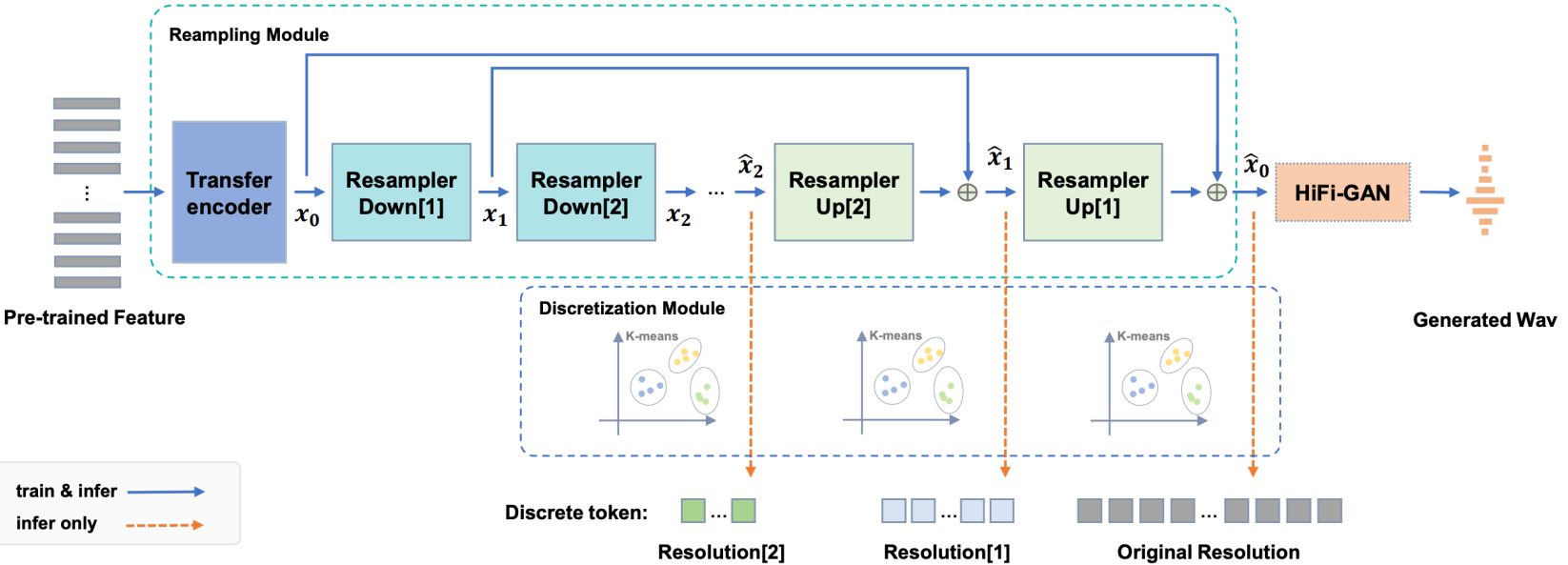
- Use multi-resolution architecture for speech self-supervised learning

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

Categorical SUPERB score over 10 speech processing tasks

Universal Representation (Modeling)

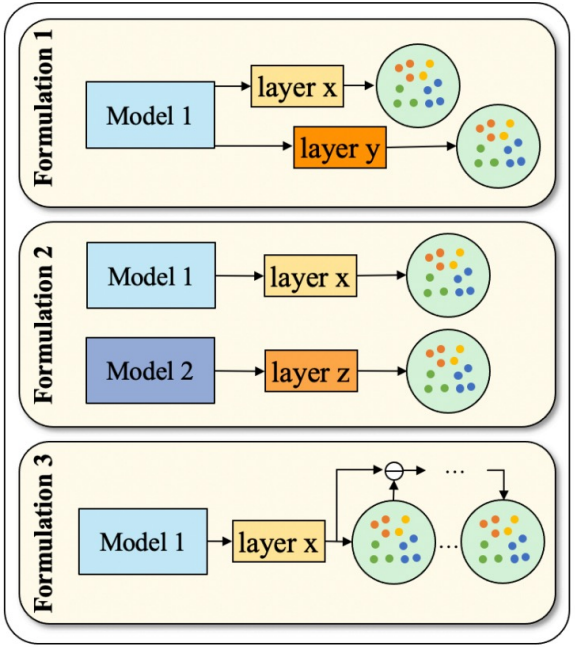
- Go universal with multi-level multi-resolutional processing
  - Sing Oriented Multi-resolution Discrete Representation (SingOMD) [Interspeech 2024, oral]



Universal Representation (Modeling)

- Go universal with multi-level multi-resolutional processing
  - Multi-layer Multi-residual Multi-stream Discrete Speech Representation (MMM)
    - [Interspeech 2024, Oral]
  - Discrete token-based SVS (TokSing)
    - [Interspeech 2024, Oral]

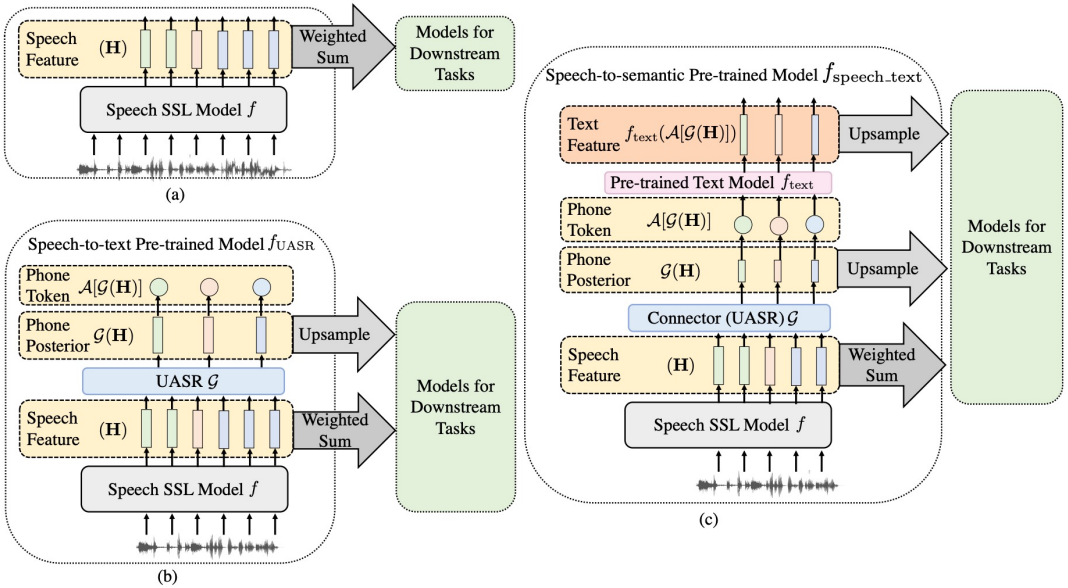
Use **multi-level** information to construct representation for both understanding and generation purposes.



Universal Representation (Application)

- Expanding the usage of representation towards different tasks (using strategy for multi-modal modeling)

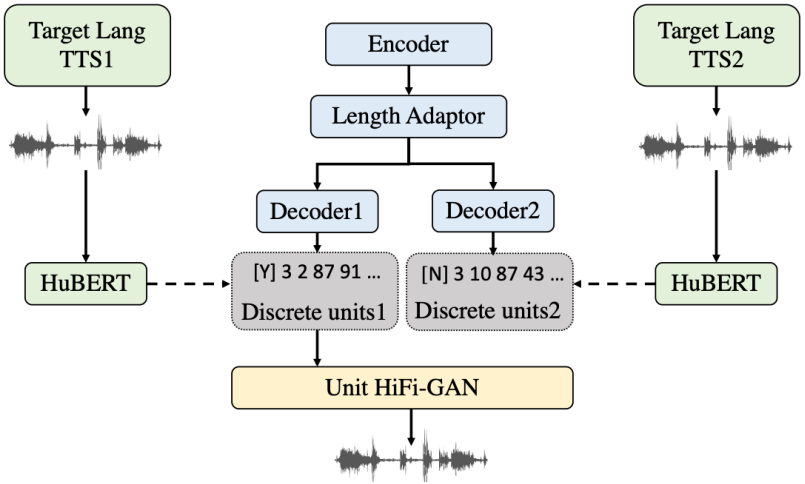
Multi-modal representation connector with unsupervised ASR [ICASSP 2023]



Universal Representation (Application)

- Expanding the usage of representation towards different tasks (using strategy for synthesis purpose)

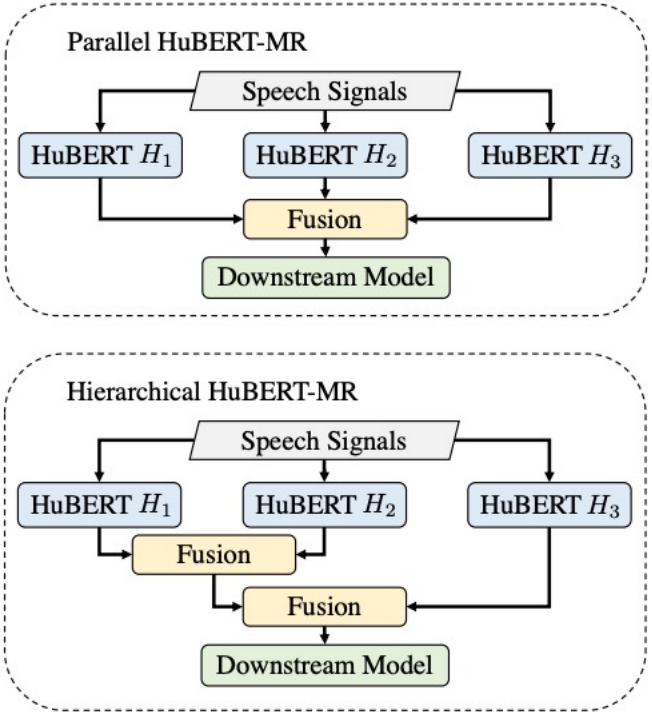
  Discrete representation as regularization terms for network training.
  [ICASSP 2023]



Universal Representation (Application)

- Expanding the usage of representation towards different tasks (using strategy for understanding purpose)

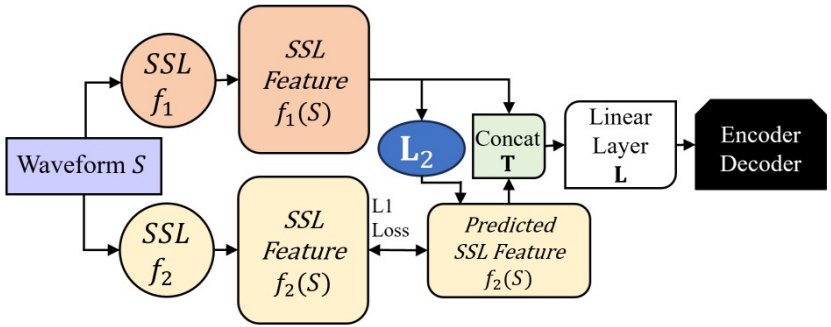
Strategies of using HuBERT with multi-resolution for various tasks
[Interspeech 2023]



Universal Representation (Application)

- Expanding the usage of representation towards different tasks (using strategy for multilingual purpose)

Efficient feature fusion for multilingual speech recognition (EFFUSE)
[Interspeech 2024, Best paper award]



Universal Representation (Evaluation)

- Representation evaluation in universal scenarios
  - Speech Universal PERFormance benchmark [Interspeech 2021]
    - Self-supervised Representation
  - Multilingual SUPERB [Interspeech 2023]
    - General speech representation (supervised/unsupervised) in multilingual scenarios
  - Dynamic SUPERB [ICASSP 2024]
    - Speech language modeling -> a unified solution for systematic evaluation in general speech systems with instruction tuning

Universal Representation (Evaluation)

- Representation evaluation with a special focus on speech in universal scenarios
  - VERSA (together with ESPnet-Codec) [SLT2024]
    - A comprehensive evaluation interface for evaluation
  - VERSA-v2 (ongoing)
    - A universal evaluation model with generative audio analysis

Brief review of recent speech language model

- A big trend together with the release of GPT-4o
  - Use additional module to inject speech/audio/music into textual LLM
    - MuLlaMA, GAMA, SALMONN, Qwen2-Audio, Llama 3.1, Audio-Flamingo, WavLLM, SALMONN-OMNI
    - Pros: easier to **maintain** the textual LLM performance
    - Cons: difficult to involve **paralinguistics** into modeling (tones, expression, emotion, environment understanding, etc.)
  - Use speech representation to joint model with text
    - SLAM, AudioLM, X-LLM, Pengi AnyGPT, AudioPaLM, VoxtLM, SpiritLM. Moshi
    - Pros: easier to use **all information** in audio
    - Cons: difficult to **maintain** the textual LLM performance
      - (sequence complication)

Brief review of recent speech language model

Major issue of recent works:
- Difficulty in balancing understanding and generation performance
- Difficulty to get aware of paralinguistics/general audio
- Difficulty to evaluate generation tasks
- Difficulty to keep efficient

ONN-

- Cons: difficult to **maintain** the textual LLM performance
  - (sequence complication)

Future with Representation

- Current needs in speech representation learning
  - Going **balanced** -> balance between understanding and generation
  - Going **expressive** with **real** ear and mouth -> be aware of paralinguistic information (understand and generate)
  - Going **evaluated** -> comprehensive + easy to use objective metrics
  - Going **fast** -> streaming + lightweight (save for LLM reasoning body)

Future with Representation

- More than GPT-4o in audio/speech?

  - More understanding in the general audio world beyond the content (paralinguistics and other audio/music information)
  - Better feedback systems to lead the targets of generation
  - More specialization in target domain with easy extension.

- Going **fast** -> streaming + lightweight (save for LLM reasoning body)

Future envision

- Thoughts from my research experiences
  - Multi-level/multi-resolution information in representation for diverse use cases
  - Strategic customization for targeted domain usage
  - Complete automatic evaluation for better guidance of the system

Thank you for listening!