Parameter-Efficient Learning (PEL) for Speech and Language: Adapters, Prompts, and Reprogramming

Dr. Huck Yang
Amazon Alexa Speech

Section 1
(PEL foundation + theory)
1hr

Dr. Pin-Yu Chen
IBM Research

Section 2
(model reprogramming)
1hr

Prof. Hung-Yi Lee
NTU

Section 3
30 min + 30 min

Cheng-Han Chiang
NTU

Section 3 (a)

Kai-Wei Chang
NTU

Section 3 (b)

Parameter-Efficient Learning for Speech Processing

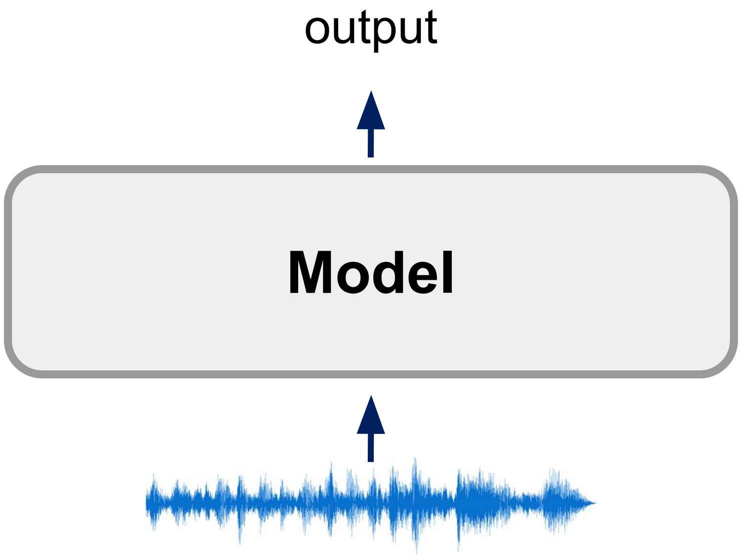
Presenter: Kai-Wei Chang
(National Taiwan University)

Outline

- Adapter Tuning for Speech Processing
  - Language Adaptation
  - Adapters for self-supervised speech models
- Prompting for Speech Processing
  - Prompting Speech Decoding Model
  - Prompting Speech Generation language Model

Adapters

Adapters Tuning for Speech Processing

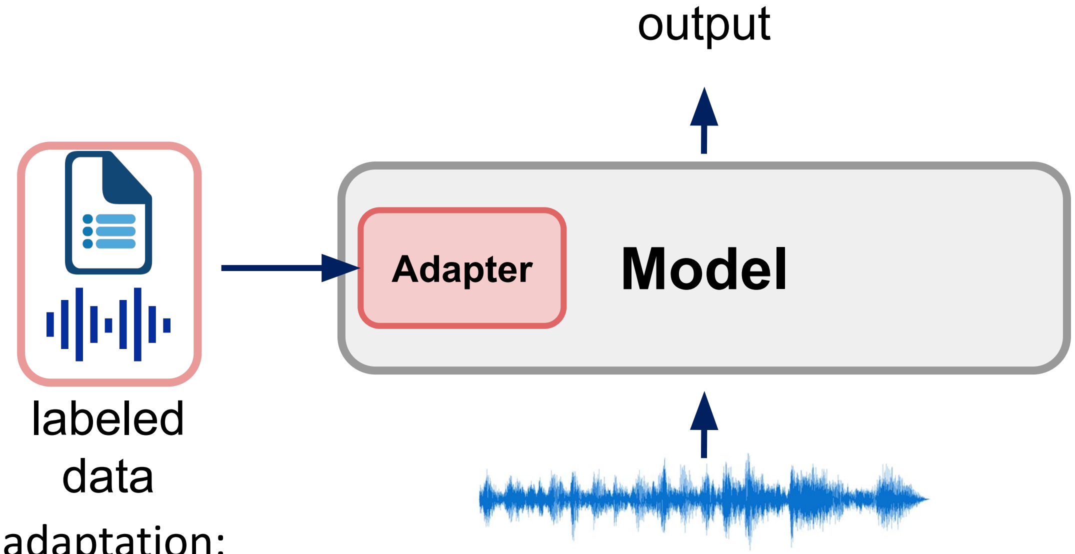


<u>Backbone model</u>:

- Multilingual Model
- Representation model
- …

Adapters Tuning for Speech Processing

Use labeled data to fine-tune adapters



Backbone model:

- Multilingual Model
- Representation model
- …

Domain adaptation:

- Language adaptation
- Speaker adaptation
- Task adaptation
- …

Adapters

Language Adaptation

1. Speech Recognition
2. Speech Translation

Adapters for SSL Model

1. Continual Learning
2. Task Adaption

Adapters

Language Adaptation

- Multilingual speech recognition system
- Multilingual speech translation system

Adapters

Language Adaptation

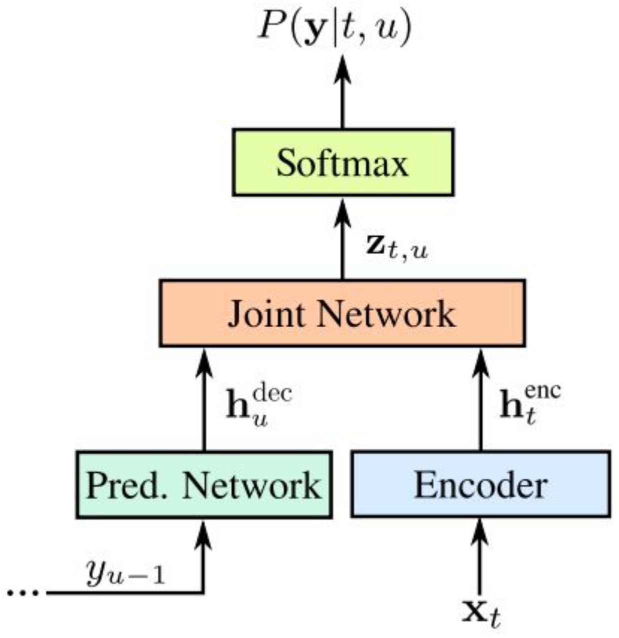
- Multilingual speech recognition system
- Multilingual speech translation system

Adapter for Multilingual Speech Recognition

- RNN-T - Multilingual ASR
- Backbone RNN-T is trained on all languages (9 Indic Languages)

Table 1: *Number of utterances in train and test sets*

Language	Train	Test	Language	Train	Test
Hindi	16M	6.3K	Tamil	1.8M	5.5K
Marathi	4.1M	6.1K	Malayalam	1.5M	9.2K
Bengali	3.9M	3.6K	Kannada	1.2M	1.1K
Telegu	2.4M	2.7K	Urdu	443K	511
Gujarati	2.2M	7.5K	**Total**	33M	43K



RNN-Transducer

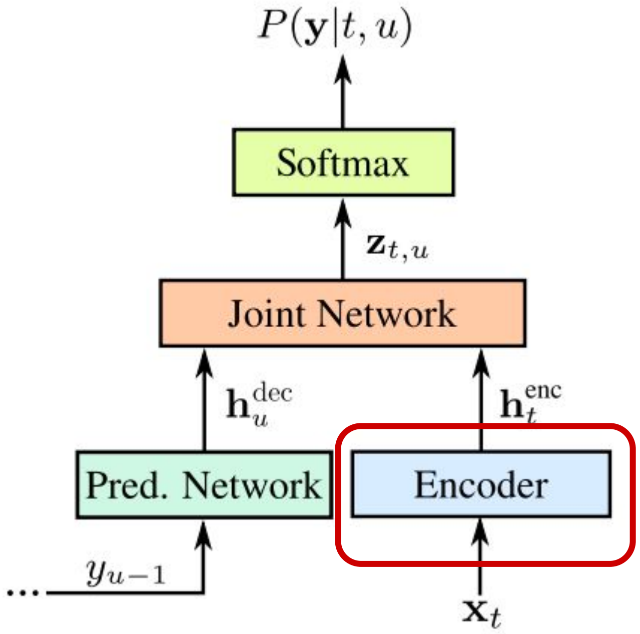
Kannan, Anjuli, et al. "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model." *Proc. Interspeech 2019* (2019): 2130-2134. **(INTERSPEECH 2019)**

Adapter for Multilingual Speech Recognition

- RNN-T - Multilingual ASR
- Backbone RNN-T is trained on all languages (9 Indic Languages)

Table 1: *Number of utterances in train and test sets*

Language	Train	Test	Language	Train	Test
Hindi	16M	6.3K	Tamil	1.8M	5.5K
Marathi	4.1M	6.1K	Malayalam	1.5M	9.2K
Bengali	3.9M	3.6K	Kannada	1.2M	1.1K
Telegu	2.4M	2.7K	Urdu	443K	511
Gujarati	2.2M	7.5K	**Total**	33M	43K

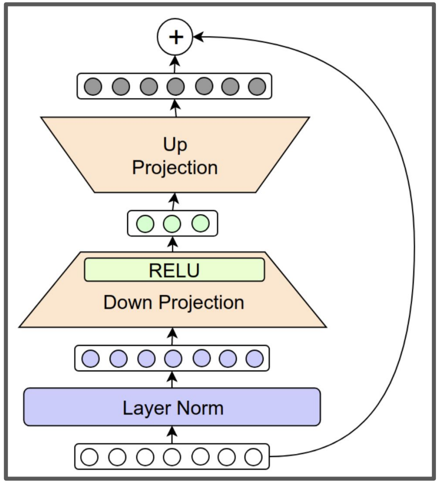


<u>RNN-Transducer</u>

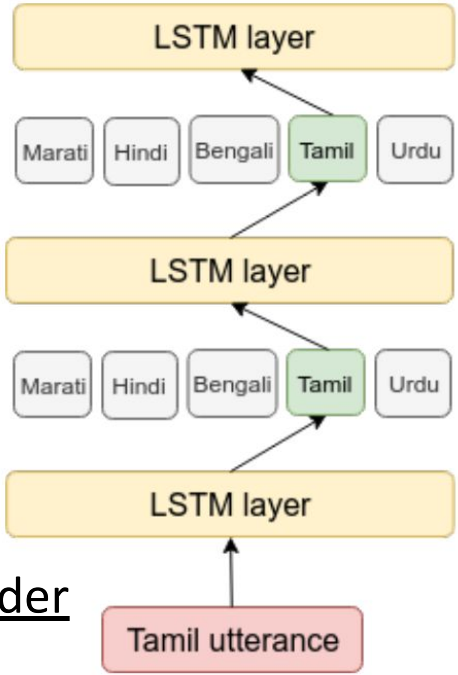
Kannan, Anjuli, et al. "Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model." *Proc. Interspeech 2019* (2019): 2130-2134. **(INTERSPEECH 2019)**

Adapter for Multilingual Speech Recognition

- RNN-T - Multilingual ASR
- Fine-tune adapter modules for each language



RNN-T Encoder



Adapter for Multilingual Speech Recognition

Word Error Rate (↓)

Exp	Model	Hindi	Marathi	Beng.	Telugu	Gujarati	Tamil	Mala.	Kann.	Urdu	Avg
A0	Multilingual RNN-T	18.5	26.2	43.9	49.3	55.3	40.1	69.7	60.8	70.1	48.2
A1	A0 + language vector	16.0	17.6	22.8	23.5	24.3	22.2	46.6	20.5	17.3	22.8
A2	A0 + sampling	22.3	29.8	41.1	45.9	43.9	37.7	64.6	55.4	48.1	43.2
A3	A1 + sampling @ 60K	18.7	18.8	24.0	24.6	24.3	25.0	47.8	21.4	17.7	24.7
A4	A1 + sampling	16.2	17.8	24.1	25.1	24.2	22.9	48.9	24.6	20.4	24.9
A5	A1 + adapters	**15.9**	**17.1**	**21.5**	**23.2**	**24.0**	**21.6**	**45.8**	**18.7**	**16.0**	**22.6**

- language vector: concatenate a one-hot vector at the input of encoder network

Adapter for Multilingual Speech Recognition

Word Error Rate (↓)

Exp	Model	Hindi	Marathi	Beng.	Telugu	Gujarati	Tamil	Mala.	Kann.	Urdu	Avg
A0	Multilingual RNN-T	18.5	26.2	43.9	49.3	55.3	40.1	69.7	60.8	70.1	48.2
A1	A0 + language vector	16.0	17.6	22.8	23.5	24.3	22.2	46.6	20.5	17.3	22.8
A2	A0 + sampling	22.3	29.8	41.1	45.9	43.9	37.7	64.6	55.4	48.1	43.2
A3	A1 + sampling @ 60K	18.7	18.8	24.0	24.6	24.3	25.0	47.8	21.4	17.7	24.7
A4	A1 + sampling	16.2	17.8	24.1	25.1	24.2	22.9	48.9	24.6	20.4	24.9
A5	A1 + adapters	**15.9**	**17.1**	**21.5**	**23.2**	**24.0**	**21.6**	**45.8**	**18.7**	**16.0**	**22.6**

- The performance is further improved with adapters
- the adapter parameters for each language is only 2% (2.5M parameters) of the backbone model (120M parameters)

Adapters

Language Adaptation

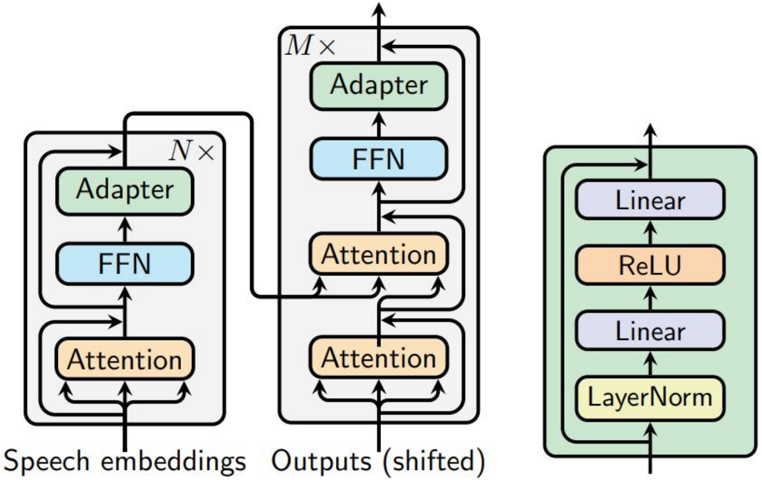
- Multilingual speech recognition system
- Multilingual speech translation system

Adapter for Multilingual Speech Translation

- Language-specific adapters can enable a fully trained multilingual ST model to specialize in multiple language pairs

Backbone: Transformer model

- Encoder-Decoder architecture
- Encoder: 12 layers
- Decoder: 6 layers



Le, Hang, et al. "Lightweight Adapter Tuning for Multilingual Speech Translation." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2021. **(ACL2021)**

Adapter for Multilingual Speech Translation

BLEU Score. The higher, the better

	Dict	D	d	Adapter ENC	Adapter DEC	Finetune ENC	Finetune DEC	# params (M) trainable/total	de	es	fr	it	nl	pt	ro	ru	avg	
								Training data (hours)	408	504	492	465	442	385	432	489		
1	mono		-	-	-	-	-	8×31.1/8×31.1	22.16	30.42	27.92	22.92	24.10	27.19	21.51	14.36	23.82	Bi.
2	multi		-	-	-	-	-	32.1/32.1	22.37	30.40	27.49	22.79	24.42	27.32	20.78	14.54	23.76	Multi.
3	multi	256	64	-	✓	-	-	8×0.2/33.7	22.32	30.50	27.55	22.91	24.51	27.36	21.09	14.74	23.87	
4	multi		64	✓	✓	-	-	8×0.6/36.9	22.75	31.07	28.03	23.04	24.75	28.06	21.20	14.75	24.21	
5	multi		128	-	✓	-	-	8×0.4/35.3	22.45	30.85	27.71	23.06	24.57	27.52	20.93	14.57	23.96	
6	multi		128	✓	✓	-	-	8×1.2/41.7	22.84*	31.25*	28.29*	23.27*	24.98*	28.16*	21.36*	14.71	24.36	
7	multi		-	-	-	-	✓	8×14.6/8×32.1	23.49	31.29	28.40	23.63	25.51	28.71	21.73	15.22	24.75	
8	multi		-	-	-	✓	✓	8×32.1/8×32.1	23.13*	31.39*	28.67*	23.80*	25.52*	29.03*	22.25*	15.44*	24.90	

Refine a multilingual speech translation system with adapters

- Speech to text translation system
  (from English to 8 target languages)

- Bilingual ST model > Multilingual ST model

Adapter for Multilingual Speech Translation

BLEU Score. The higher, the better

	Dict	D	Adapter d	Adapter ENC	Adapter DEC	Finetune ENC	Finetune DEC	# params (M) trainable/total	de	es	fr	it	nl	pt	ro	ru	avg	
								Training data (hours)	408	504	492	465	442	385	432	489		
1	mono		-	-	-	-	-	8×31.1/8×31.1	22.16	30.42	27.92	22.92	24.10	27.19	21.51	14.36	23.82	
2	multi		-	-	-	-	-	32.1/32.1	22.37	30.40	27.49	22.79	24.42	27.32	20.78	14.54	23.76	Multi.
3	multi	256	64	-	✓	-	-	8×0.2/33.7	22.32	30.50	27.55	22.91	24.51	27.36	21.09	14.74	23.87	
4	multi		64	✓	✓	-	-	8×0.6/36.9	22.75	31.07	28.03	23.04	24.75	28.06	21.20	14.75	24.21	
5	multi		128	-	✓	-	-	8×0.4/35.3	22.45	30.85	27.71	23.06	24.57	27.52	20.93	14.57	23.96	
6	multi		128	✓	✓	-	-	8×1.2/41.7	22.84*	31.25*	28.29*	23.27*	24.98*	28.16*	21.36*	14.71	24.36	Adapter
7	multi		-	-	-	-	✓	8×14.6/8×32.1	23.49	31.29	28.40	23.63	25.51	28.71	21.73	15.22	24.75	
8	multi		-	-	-	✓	✓	8×32.1/8×32.1	23.13*	31.39*	28.67*	23.80*	25.52*	29.03*	22.25*	15.44*	24.90	Fine-tune

Refine a multilingual speech translation system with adapters

- Multilingual ST are further refined on each language pair with adapter
- Adapter: 1.2M parameters / Backbone model: 30M parameters

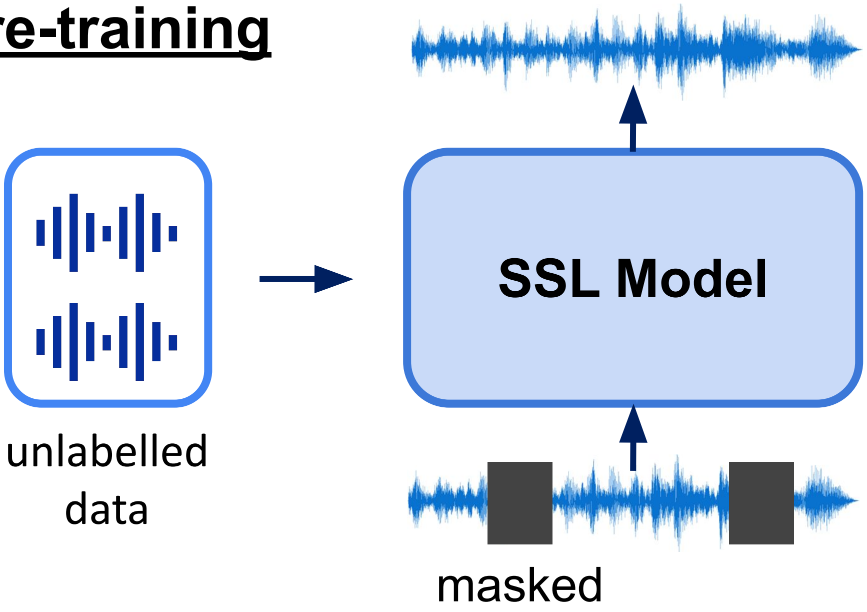
Adapters

Self-supervised Learning Speech Model

- Continual Learning
- Task Adaptation

Adapter for SSL Speech Models

Pre-training



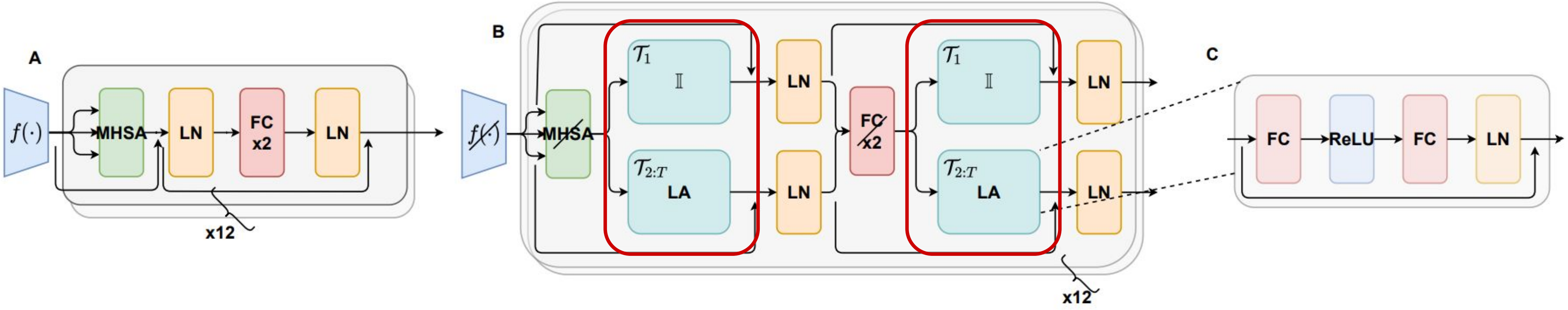
- **wav2vec 2.0 -** contrastive
- **HuBERT / WavLM -** masked prediction

...

Baevski et.al., NeurIPS'20
Hsu et.al., TASLP Volume 29

Adapter for SSL Speech Models

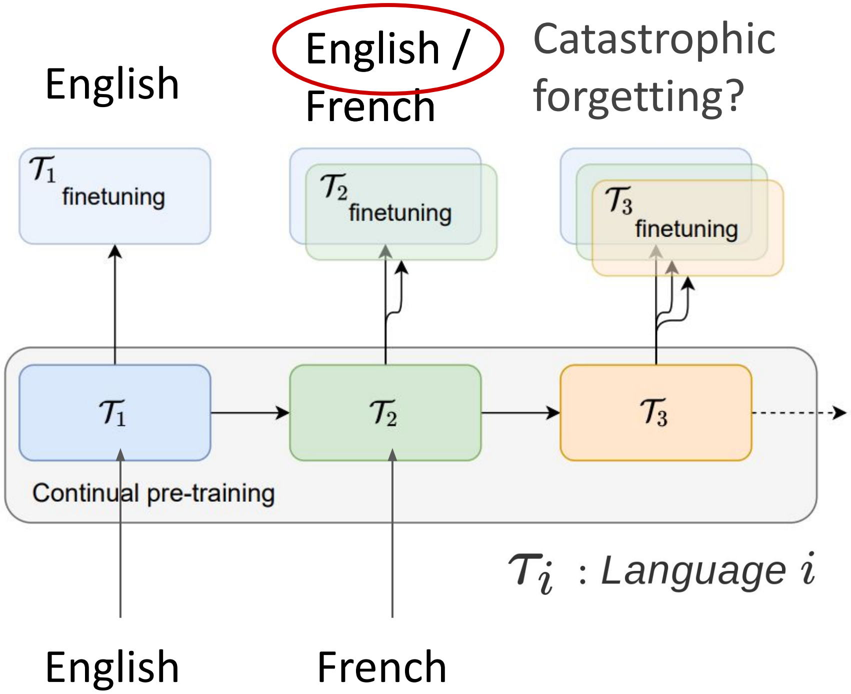
Wav2Vec 2



- Original Wav2Vec2 is pre-trained on English dataset
- Adapters for different languages : French, Spanish
- Continual pre-train the Wav2Vec 2 but only update the adapters

Kessler, Samuel, Bethan Thomas, and Salah Karout. "An adapter based pre-training for efficient and scalable self-supervised speech representation learning." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022. **(ICASSP 2022)**

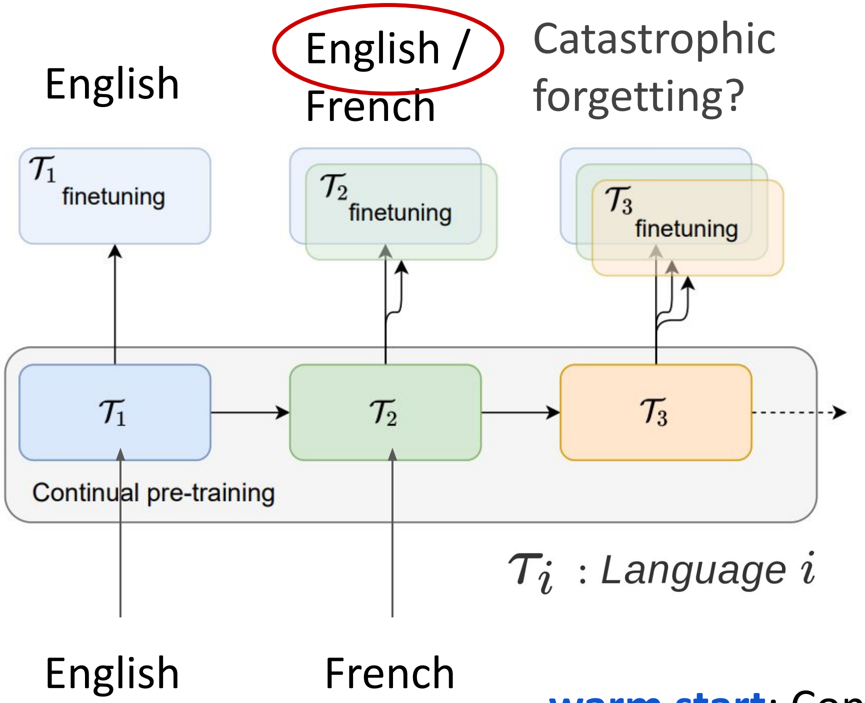
Adapter for SSL Speech Models

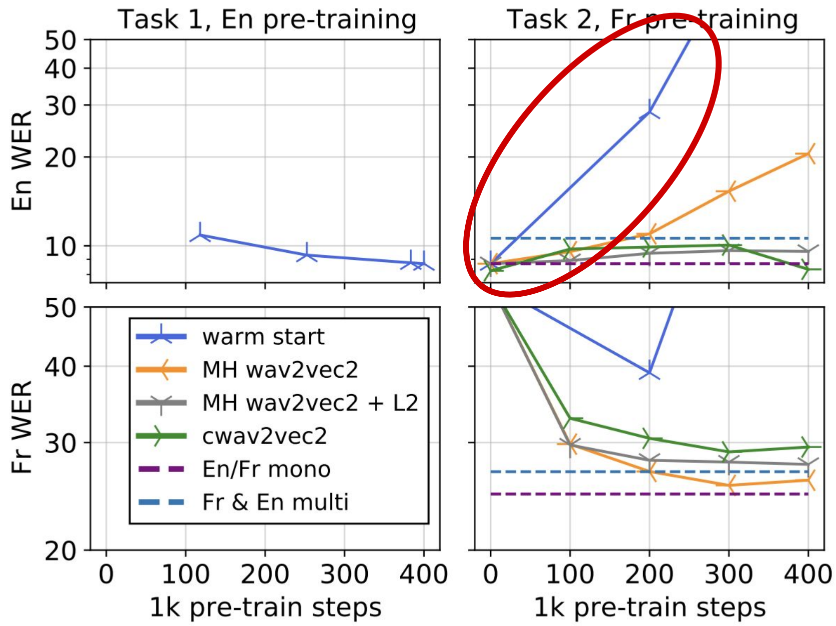


Adapter for SSL Speech Models

WER: The lower, the better

Catastrophic forgetting



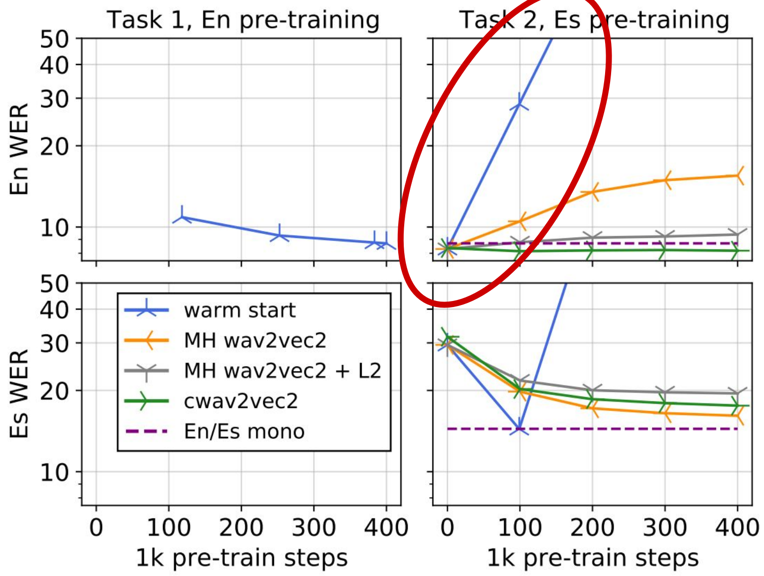


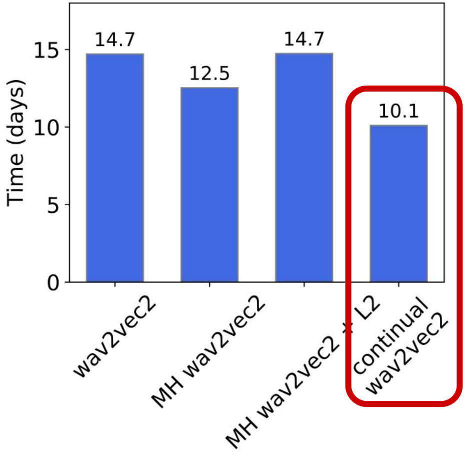
warm start: Continuing training the wav2vec 2 on new language

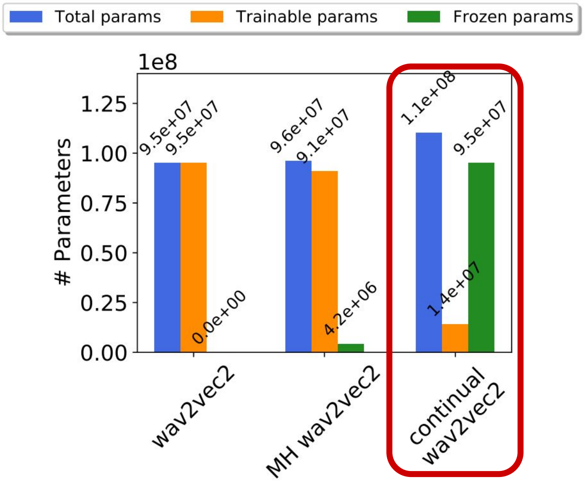
cwav2vec 2: Tuning the adapter for each language

Adapter for SSL Speech Models

WER: The lower, the better **Catastrophic forgetting**







Same trend can be observed when adapting to Spanish speech.

Adapters saves <u>computation time</u> and <u>storage</u>

Adapters

Self-supervised Learning Speech Model

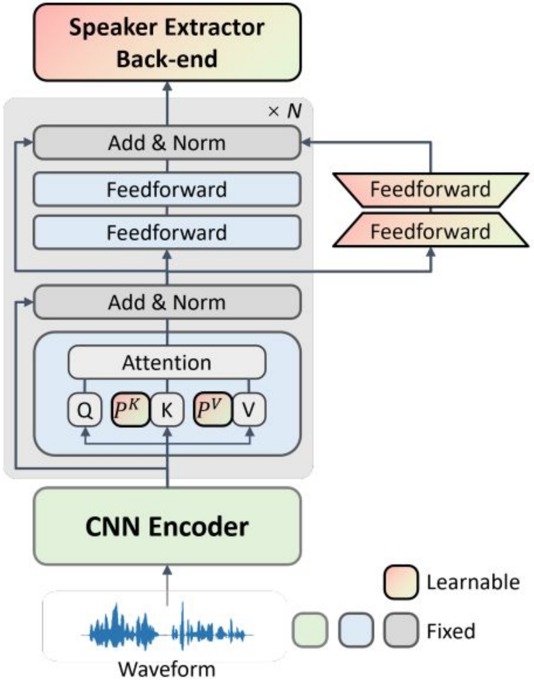
- Continual Learning
- Task Adaptation

Adapters for Speaker Verification

Downstream model:
Speaker extractor model

Self-supervised
Speech model

Tuning adapters and speaker extractor model on top of speech representation model



<u>Mix-And-Match Adapter</u>:
Prefix-tuning + Adapter

Peng, Junyi, et al. "Parameter-efficient transfer learning of pre-trained Transformer models for speaker verification using adapters." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. **(ICASSP 2023)**

Adapters for Speaker Verification

Pre-trained Model: **WavLM BASE+**, Back-end: **MHFA**		
Full fine-tuning	94.7M+2.2M	0.66
Full fine-tuning [LM-FT] [11]	94.7M+2.2M	0.59
Fixed	0.0M + 2.2M	1.45
Bottleneck Adapter	4.7M + 2.2M	0.78
Prefix Tuning	3.6M + 2.2M	1.15
MAM Adapter	5.4M + 2.2M	0.72
MAM Adapter [LM-FT]	5.4M + 2.2M	0.61

Fine-tuning

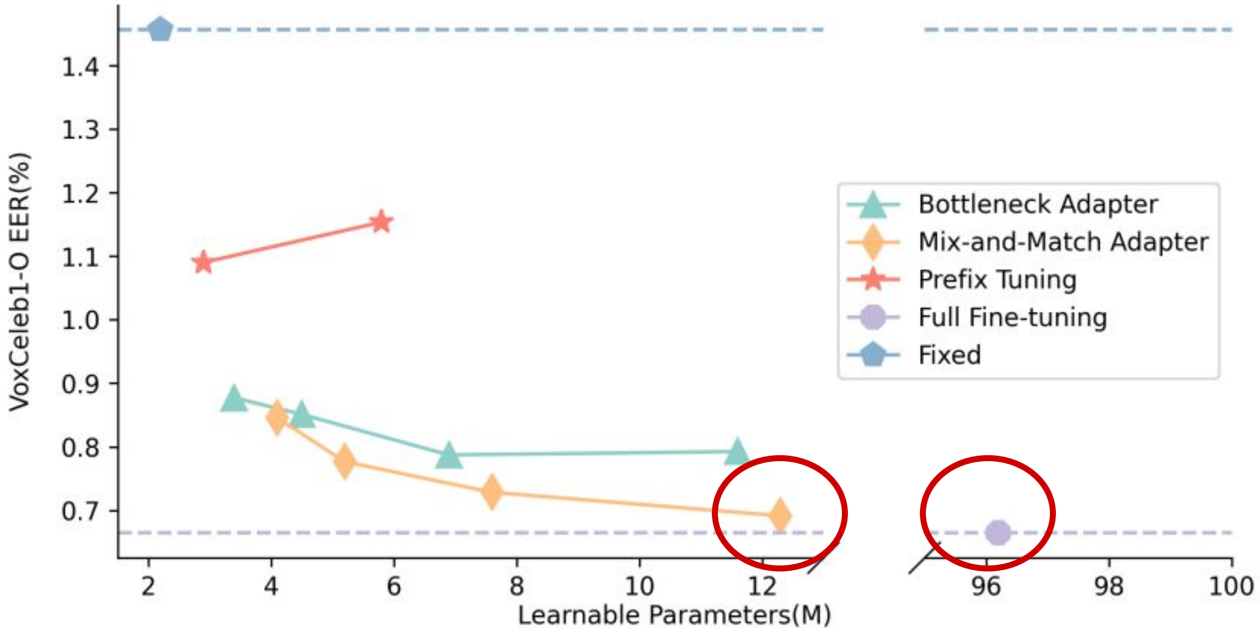
Fixed

Mix-And-Match Adapter

LM-FT: Large margin fine-tuning

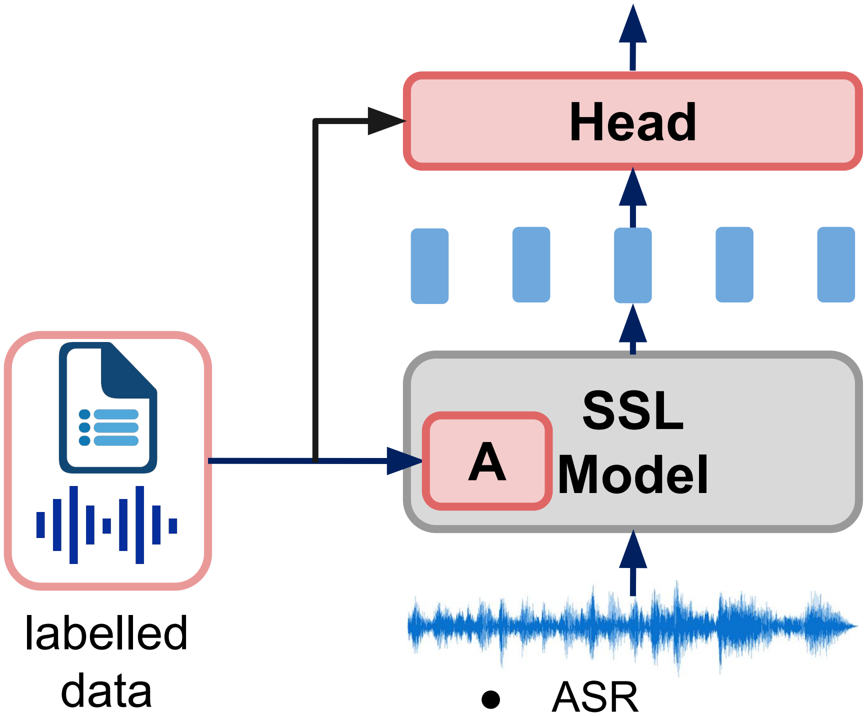
Equal Error Rate (EER). The lower, the better

Adapters for Speaker Verification



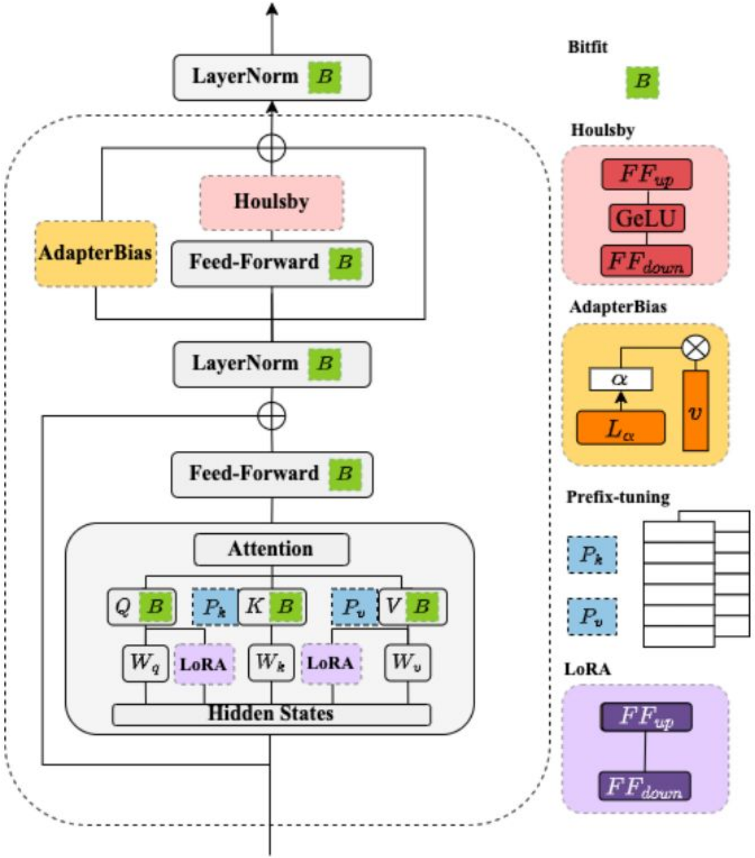
Fusing prefix tuning and bottleneck adapters can perform comparable to full fine-tuning with much fewer parameters (about 8 times fewer parameters)

Adapters for Multiple Tasks



- ASR
- Speaker Id.
- Intent Cls.

Adapters for Multiple Tasks



- BitFit: Tuning the bias
- AdapterBias: Using a neural networks to generate bias
- Houlsby: Normal down proj., up proj. adapter
- Prefix-tuning
- LoRA

Chen, Zih-Ching, et al. "Exploring efficient-tuning methods in self-supervised speech models." *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023. **(SLT2022)**

Adapters for Multiple Tasks

Backbone: HuBERT

			Phoneme Recognition	Speaker Diarization	Speaker Identification	Slot Filling	Intent Classification	Keyword Spotting
Method	Params	ASR	PR	SD	SID	SF	IC	KS
FT	94.7M	6.35	**2.45**	9.32	66.48	84.87	99.10	95.87
Baseline	0	7.09	7.74	7.05	64.78	86.25	96.39	95.32
Houlsby	0.60M	5.88	3.00	**4.00**	**87.71**	85.87	**99.60**	97.17
AdapterBias	0.02M	**5.54**	4.19	5.48	77.38	86.60	99.50	97.30
BitFit	0.10M	9.34	4.23	5.13	83.68	**87.40**	99.50	**97.33**
LoRA	0.29M	6.94	8.74	7.39	62.90	86.25	96.57	96.59
Prefix	0.10M	6.56	4.18	8.17	71.87	85.85	99.31	97.05

- HuBERT + Adapters: The representation of the last layer is fed into the downstream head.
- Fine-tuning only achieve the best performance on Phoneme Recognition
- Different adapter shows different characteristics

Adapter Summary

- Adapters for language adaptation
  - multilingual speech recognition system
  - multilingual speech translation system

- Adapters in Self-supervised speech models
  - continual learning
  - task adaptation

More adapter works…

ga642381 / **Speech-Prompts-Adapters** Public

Pin

Code · Issues · Pull requests · Actions · Projects · Wiki · Security · Insights · Settings

main · 1 branch · 0 tags · Go to file · Add file · Code

ga642381 Update README.md · ed1b86f last week · 11 commits

README.md · Update README.md · last week

README.md

Speech-Prompts-Adapters

This Repository surveys the paper focusing on **Adapters** and **Prompting** methods for **Speech Processing**.

Navigation

- ICASSP 2023 Tutorial Information
- **Adapters** for Speech Processing
- **Prompting** for Speech Processing
- **Reprogramming** and Prompting
- **Parameter Efficient Learning** Methods
- Contact

Adapters and Prompting for Speech Processing

Adapters for Speech Processing

Title	Authors	Modality	Task	Link
A Parameter-Efficient Learning Approach to Arabic Dialect Identification with Pre-Trained General-Purpose Speech Model	Srijith Radhakrishnan et al.	Speech	Dialect Identification	Interspeech 2023
CHAPTER: Exploiting Convolutional Neural Network Adapters for Self-supervised Speech Models	Zih-Ching Chen et al.	Speech	[Multiple]	arXiv 2022
Parameter Efficient Transfer Learning for Various Speech Processing Tasks	Shinta Otake et al.	Speech	[Multiple]	arXiv 2022
Parameter-efficient transfer learning of pre-trained Transformer models for speaker verification using adapters	Junyi Peng et al.	Speech	Speaker Verification	arXiv 2022
Exploring Efficient-tuning Methods in Self-supervised Speech Models	Zih-Ching Chen et al.	Speech	[Multiple]	SLT 2022
DRAFT: A Novel Framework to Reduce Domain Shifting in Self-supervised Learning and Its Application to Children's ASR	Ruchao Fan, Abeer Alwan	Speech	ASR	Interspeech 2022
Speaker adaptation for Wav2vec2 based dysarthric ASR	Murali Karthick Baskar et al.	Speech	ASR	Interspeech 2022
Adaptive multilingual speech recognition with pretrained models	Ngoc-Quan Pham et al.	Speech	ASR	Interspeech 2022
An Adapter Based Pre-Training for Efficient and Scalable Self-Supervised Speech Representation Learning	Samuel Kessler et al.	Speech	ASR	ICASSP 2022
Efficient Adapter Transfer of Self-Supervised Speech Models for Automatic Speech Recognition	Bethan Thomas et al.	Speech	ASR	ICASSP 2022
Scaling End-to-End Models for Large-Scale Multilingual ASR	Bo Li et al.	Speech	ASR	ASRU 2021
Meta-Adapter: Efficient Cross-Lingual Adaptation With Meta-Learning	Wenxin Hou et al.	Speech	ASR	ICASSP 2021
Exploiting Adapters for Cross-Lingual Low-Resource Speech Recognition	Wenxin Hou et al.	Speech	ASR	TASLP 2021
Lightweight Adapter Tuning for Multilingual Speech Translation	Hang Le et al.	Speech	Speech Translation	ACL-IJCNLP 2021
Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech	Katrin Tomanek et al.	Speech	ASR	EMNLP 2021
Multilingual Speech Recognition with Self-Attention Structured Parameterization	Yun Zhu et al.	Speech	ASR	Interspeech 2020
Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model	Anjuli Kannan et al.	Speech	ASR	Interspeech 2019

Prompting for Speech Processing

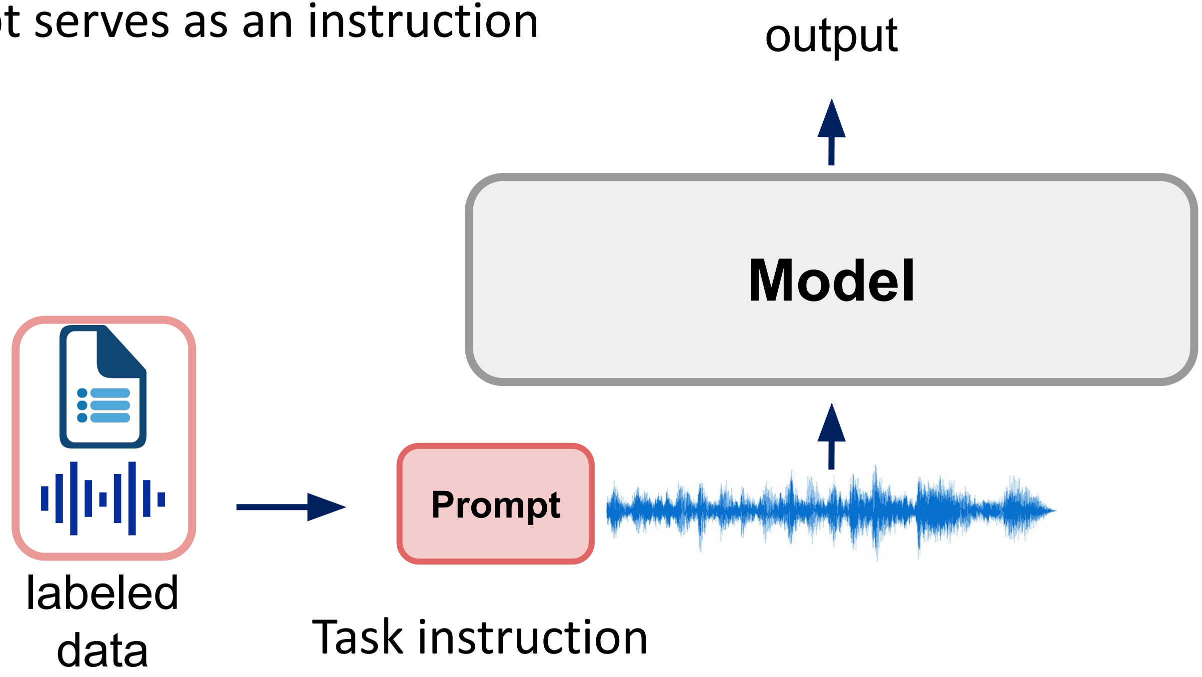
Title	Authors	Modality	Task	Link

https://github.com/ga642381/Speech-Prompts-Adapters

Prompting

Prompting for Speech Processing

- Use labeled data to fine-tune Prompt
- The prompt serves as an instruction



Prompting

Prompt Speech Decoding Model

1. Prompting Whisper

Prompt Speech Generation Model

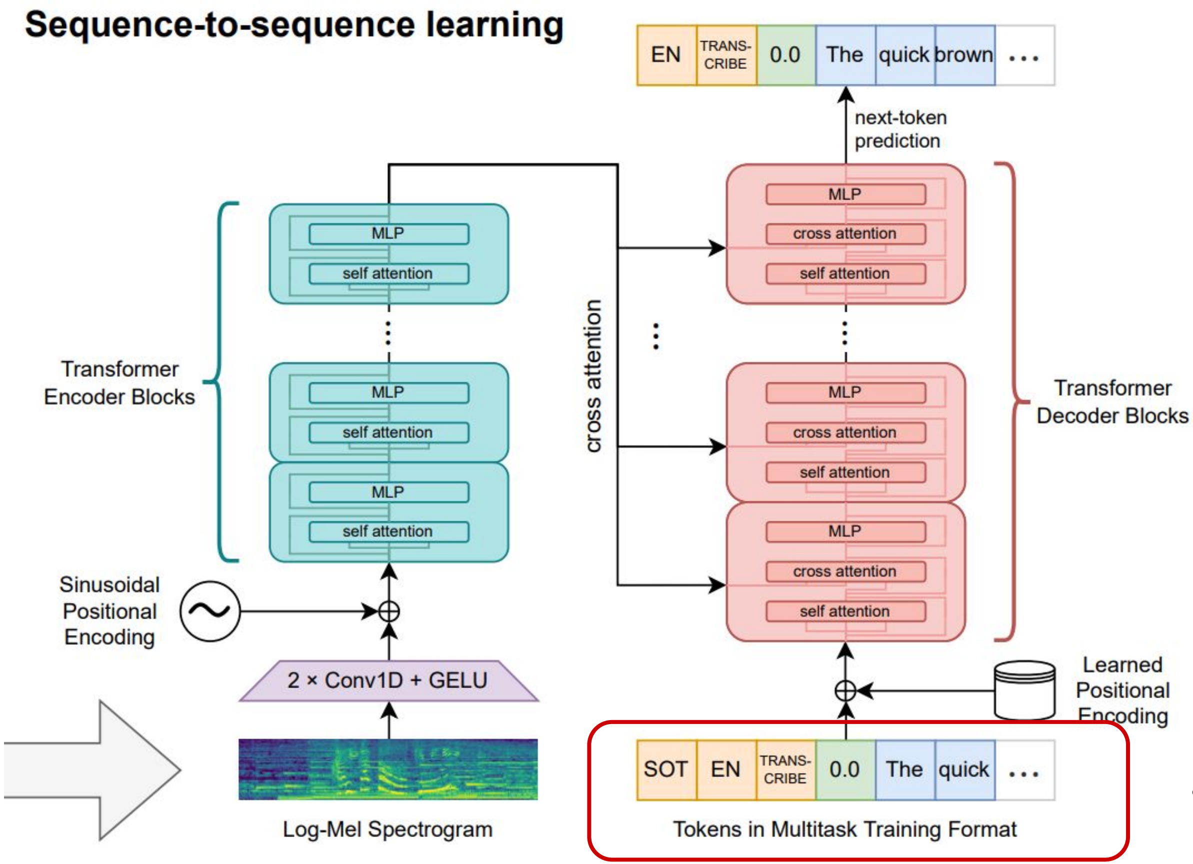
1. Prompting Generative SpeechLM

Prompting

Speech Decoding Model

- Prompting Whisper for multiple new tasks

Whisper model with multitask learning



Whisper is supervised trained in a multitask learning manner

- LID
- Speech recognition
- Speech translation

Multitask learning

Language tags: <|en|> <|zh|>,...

Task tags: <|asr|> <|st|>

Prompting Whisper Model

Prompt design for whisper model to perform multiple tasks

<table>
<tr><th>Task</th><th>Language(s)</th><th>Default prompt</th><th>Our proposed prompt</th><th>Improvement</th></tr>
<tr><td>AVSR</td><td>En</td><td><|sot|><|en|><|asr|></td><td><|sop|>CLIP retrie.<default></td><td>10%</td></tr>
<tr><td>CS-ASR</td><td>Zh+En</td><td><|sot|><|zh|>or<|en|><|asr|></td><td><|sot|><|zh|><|en|><|asr|></td><td>19%</td></tr>
<tr><td>ST</td><td>En→Ru</td><td><|sot|><|ru|><|st|></td><td><|sot|><|ru|><|asr|></td><td>45%</td></tr>
</table>

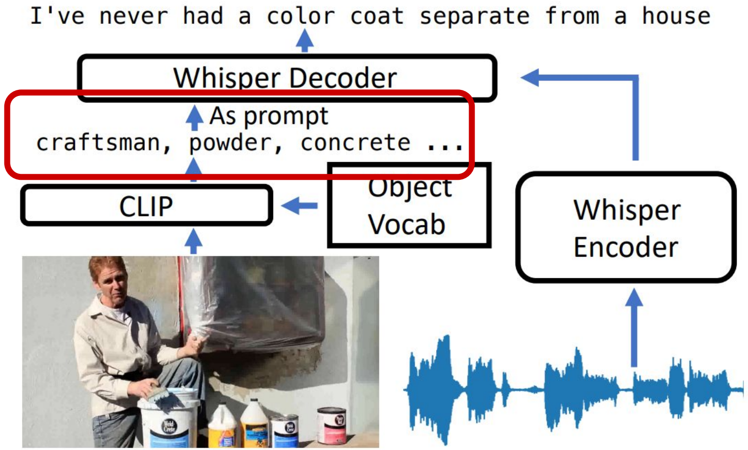
- AVSR: Audio visual speech recognition
- CS-ASR: Code-switched ASR
- ST: En-> X Speech translation

Peng, Puyuan, et al. "Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization." ***(INTERSPEECH 2023)***.

Prompting Whisper Model

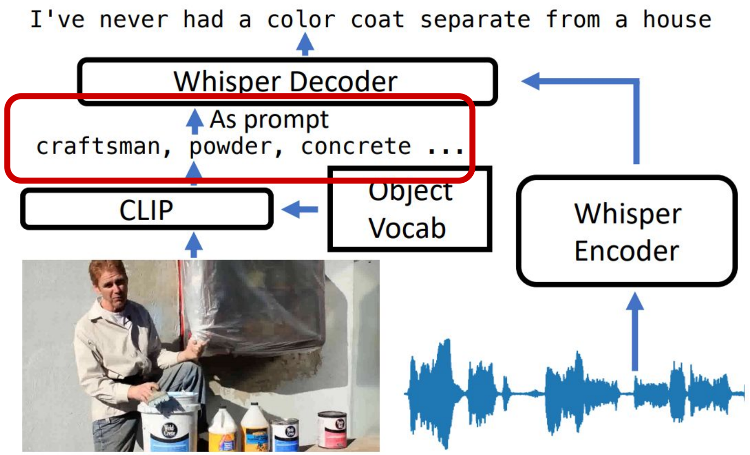
<u>Audio Visual Speech Recognition</u>

Providing Whisper with visually-conditioned prompt



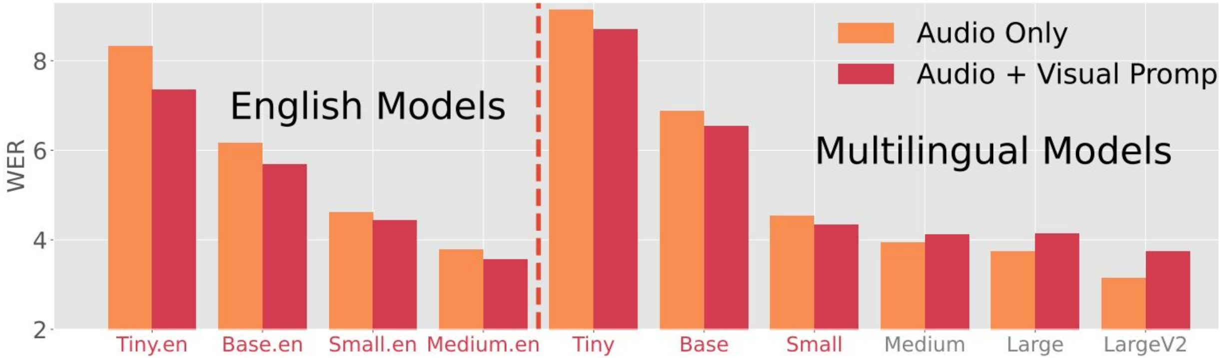
Prompting Whisper Model

Audio Visual Speech Recognition



Providing Whisper with visually-conditioned prompt

Visual information helps when Whisper model is not large



Prompting Whisper Model

Code Switched Speech Recognition

Dataset	Lang. prompt.	Zh CER	En WER	CS MER	Total MER
ASCEND	<|zh|>	**16.3**	93.1	33.1	32.6
	<|en|>	90.4	**31.5**	80.1	78.9
	default	17.0	31.8	26.6	22.1
	concat	16.6	31.8	**25.0**	**21.3**
SEAME	<|zh|>	26.3	97.4	43.3	46.7
	<|en|>	99.3	**33.8**	86.9	82.2
	default	27.1	85.5	43.2	45.3
	concat	**25.9**	44.7	**38.4**	**36.9**

default

concat

default

concat

- default: let Whisper perform LID first then perform speech recognition
- concat: <|sot|>**<|en|><|zh|>**<|asr|>

Prompting Whisper Model

<u>Zero-shot En-X speech translation</u>

Whisper has never perform ST on these language pairs.

Category	Approach	En→De	En→Ru	En→Fr
Supervised	w2v2+mBART [30]	32.4	20.0	23.1
	E2E Transformer [35]	27.2	15.3	11.4
Unsupervised	Chung et al. [36]	-	-	12.2
	Cascaded [30]	22.0	10.0	15.4
	E2E (w2v2+mBART) [30]	23.8	9.8	15.3
Zero-shot	Escolano et al. [33]	6.8	-	10.9
	T-Modules* [34]	23.8	-	32.7
	Whisper w/ default prompt	0.4	8.8	0.8
	Whisper w/ our prompt	18.1	12.8	13.1

- default: <|sot|><|ru|>**<|st|>**
- proposed prompt: <|sot|><|ru|>**<|asr|>**

Prompting

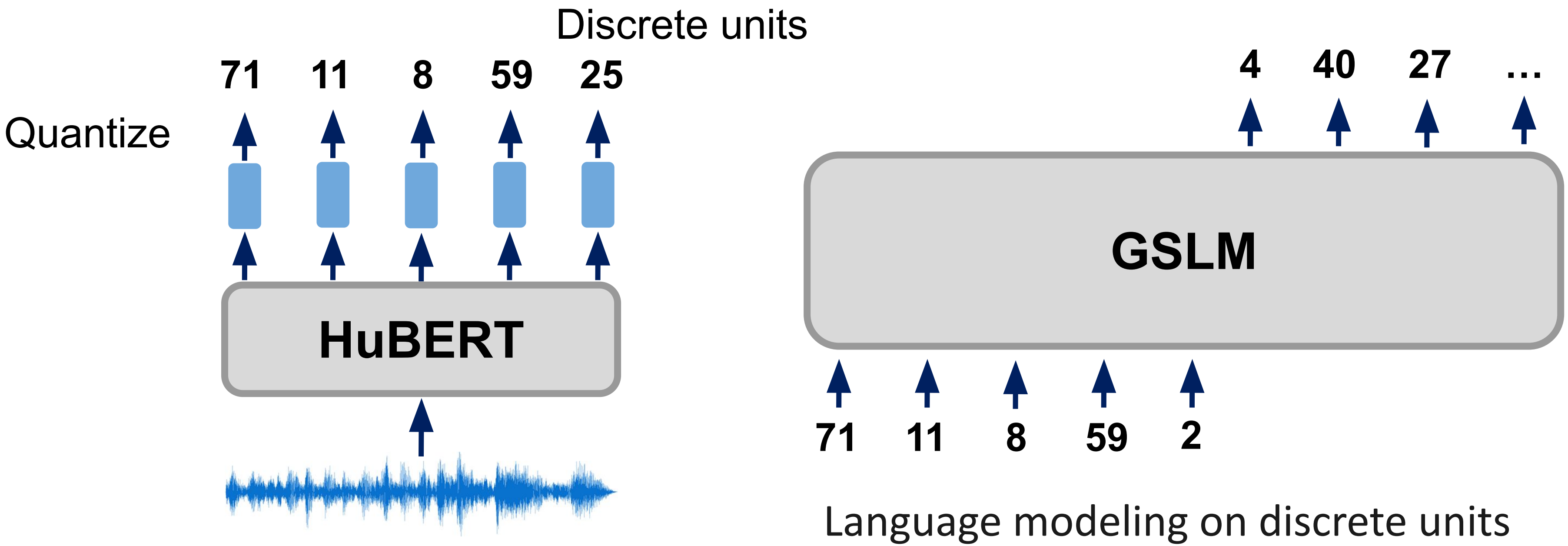
Speech Generation Language Model

- Prompting Generative Speech LM for multiple tasks

Generative Spoken Language Model (GSLM)

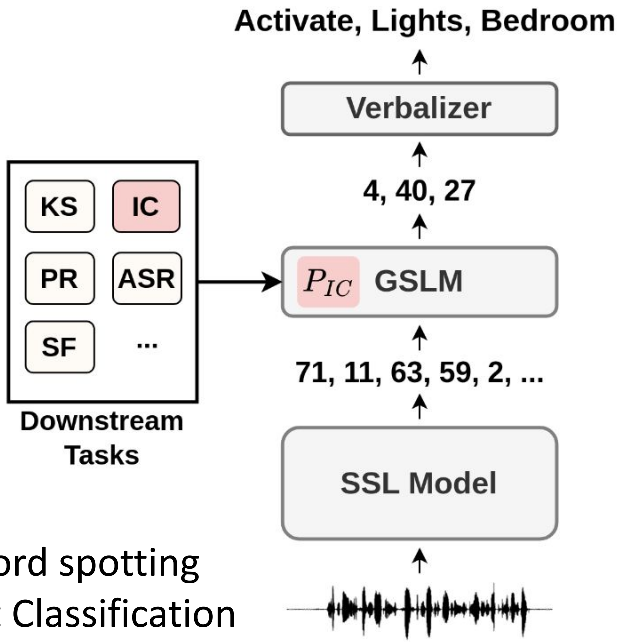
Prompting Generative Spoken Language Model (GSLM)

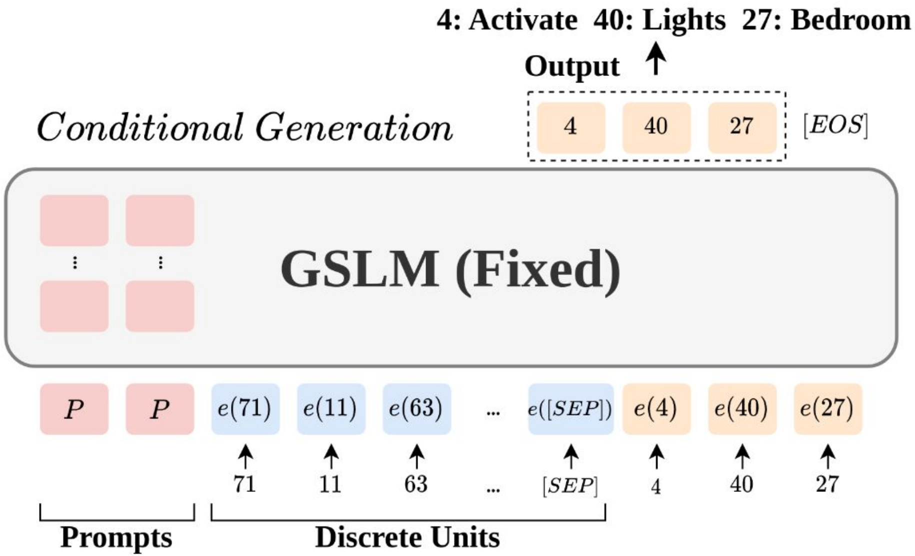
Chang, Kai-Wei, et al. "An exploration of prompt tuning on generative spoken language model for speech processing tasks." *(INTERSPEECH2022)*



Language modeling on discrete units

SpeechPrompt: Prompting Speech LM





- Keyword spotting
- Intent Classification
- Phoneme Recognition
- ASR
- Slot Filling

Prompts are trainable parameters

Chang, Kai-Wei, et al. "An exploration of prompt tuning on generative spoken language model for speech processing tasks." ***(INTERSPEECH2022)***

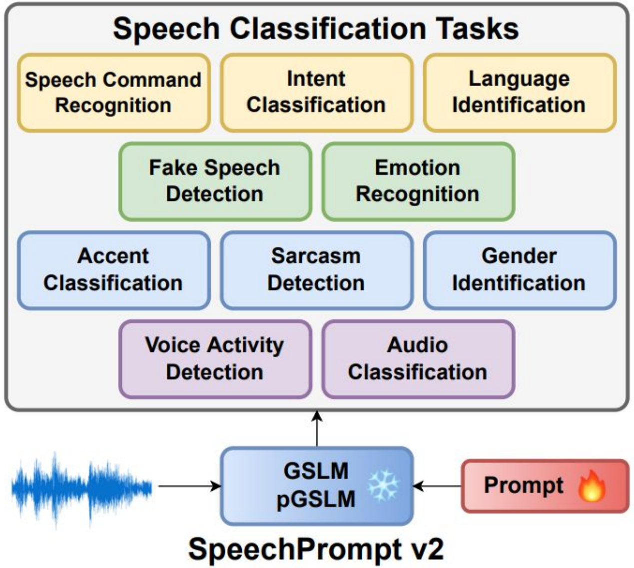
SpeechPrompt: Prompting Speech LM

- Speech classification tasks
  - Keyword Spotting (KS)
  - Intent Classification (IC)
- Sequence generation tasks
  - Speech Recognition (ASR)
  - Slot Filling (SF)

Scenarios	KS		IC	
	Acc ↑	#	Acc ↑	#
HuBERT-PT	95.16	0.08M	**98.40**	0.15M
FT-LM	94.03	151M	97.63	151M
FT-DM	**96.30**	0.2M	98.34	0.2M
CPC-PT	**93.54**	0.05M	**97.57**	0.05M
FT-LM	93.48	151M	95.62	151M
FT-DM	91.88	0.07M	64.09	0.07M

Scenarios	ASR		SF		#
	WER ↓	CER ↓	F1 ↑	CER ↓	
HuBERT-PT	34.17	26.14	66.90	59.47	4.5M
FT-LM	26.19	16.80	80.58	40.15	151M
FT-DM	**6.42**	**1.48**	**88.53**	**25.20**	43M
CPC-PT	59.41	37.12	65.25	60.84	4.5M
FT-LM	35.61	17.90	**79.34**	**42.64**	151M
FT-DM	**20.18**	**5.25**	71.19	49.91	42.5M

SpeechPrompt: Prompting Speech LM



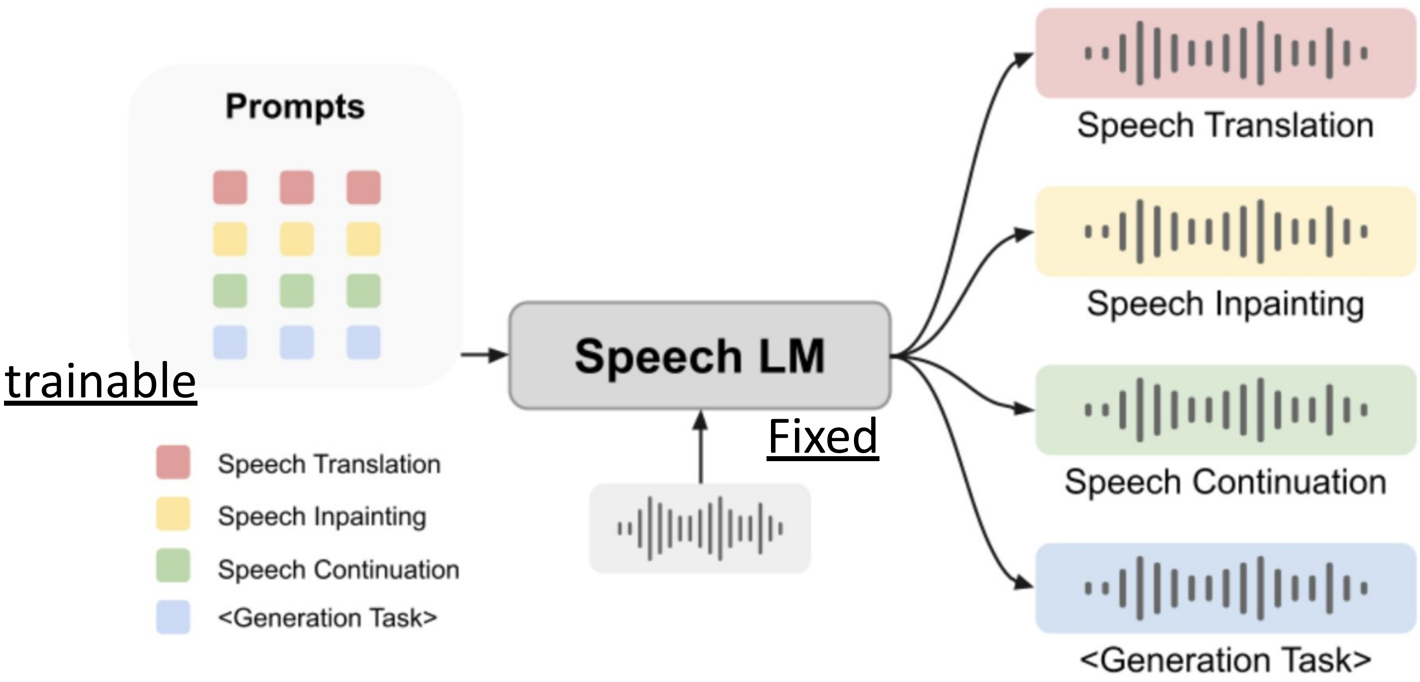
Explore various Speech LM

Task	Metric	Dataset	Language	GSLM	GSLM+	pGSLM	pGSLM+
SCR	ACC (↑)	Google SC v1	En	94.5	94.6	94.3	**94.7** (-3.9)
		Grabo SC	Du	92.4	**92.7** (-6.2)	17.5	19.6
		Lithuanian SC	Lt	93.2	**95.5** (+3.7)	90.9	79.5
		Arabic SC	Ar	99.7	**100.0** (+1.1)	85.6	92.6
IC	ACC (↑)	Fluent SC	En	97.2	97.3	98.1	**98.2** (-1.5)
LID	ACC (↑)	Voxforge	En, Es, Fr De, Ru, It	90.9	**94.2** (-5.6)	81.8	80.4
FSD	EER (↓)	ASVspoof	En	18.5	13.5	**13.1** (+10.6)	18.3
ER	ACC (↑)	IEMOCAP	En	42.1	44.3	49.9	**50.2** (-29)
AcC	ACC (↑)	AccentDB	En	78.9	83.4	86.5	**87.1** (-12.4)
SD	F1 (↑)	MUStARD	En	55.0	77.8	74.4	**78.7** (+13.1)
		MUStARD++	En	74.0	**75.2** (+10)	52.7	58.2
GID	F1 (↑)	VoxCeleb1	En	86.2	87.3	**91.6** (-6.7)	86.2
VAD	ACC (↑)	Google SC v2 & Freesound	En	96.6	96.9	**98.3** (-0.5)	98.1
AuC	ACC (↑)	ESC-50	✖	9.0	**37.5** (-59.5)	20.3	27.0

- outperform / competitive with SOTA: 10 / 14 tasks
- trainable parameters: 0.1 M parameters

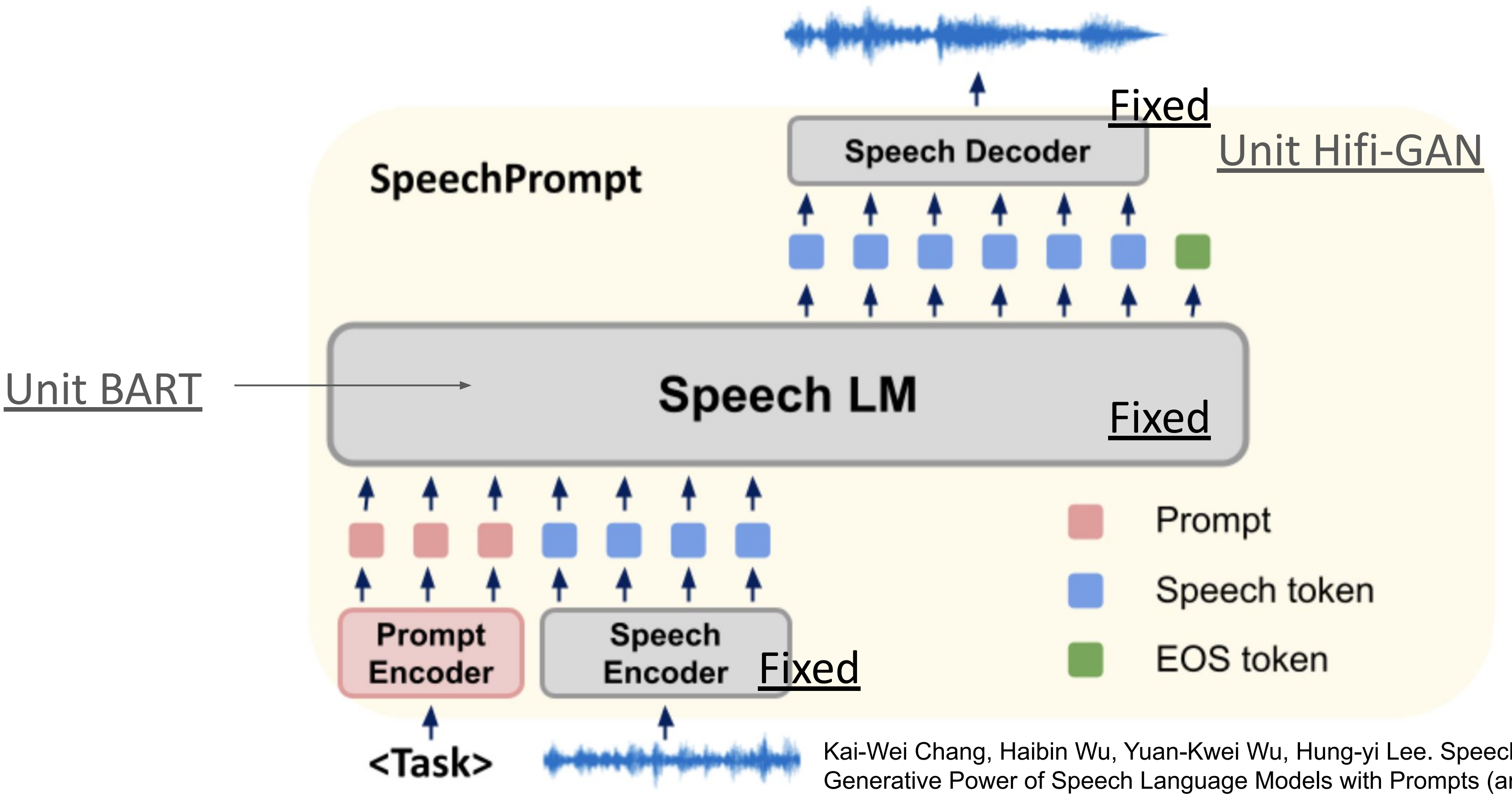
Chang, Kai-Wei, et al. "SpeechPrompt v2: Prompt Tuning for Speech Classification Tasks." *arXiv preprint arXiv:2303.00733* (2023).

SpeechGen: Prompting for Speech Generation



Kai-Wei Chang, Haibin Wu, Yuan-Kwei Wu, Hung-yi Lee. SpeechGen: Unlocking the Generative Power of Speech Language Models with Prompts (arXiv 2023)

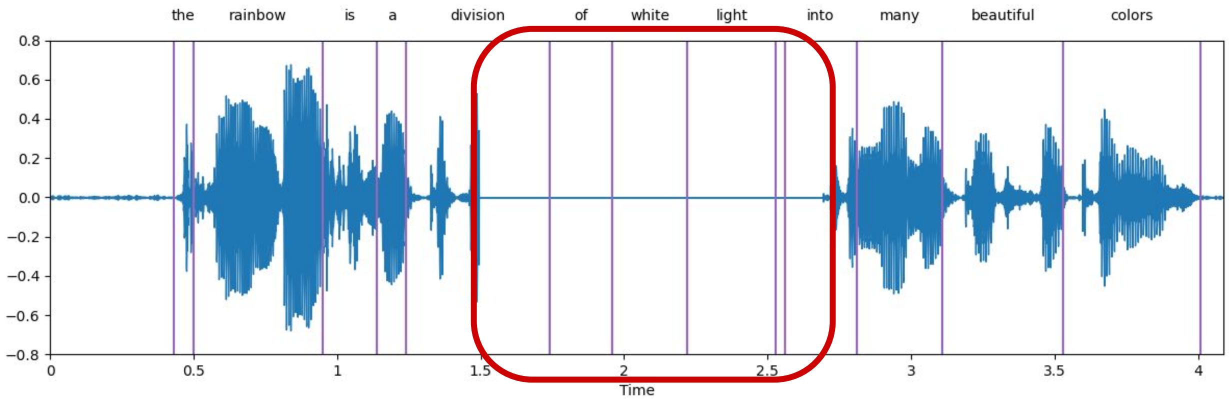
SpeechGen: Prompting for Speech Generation

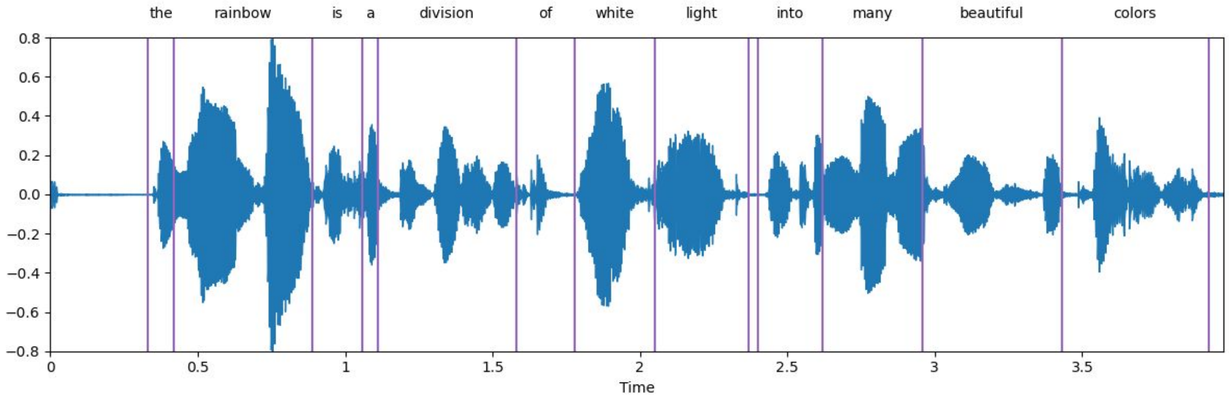


Kai-Wei Chang, Haibin Wu, Yuan-Kwei Wu, Hung-yi Lee. SpeechGen: Unlocking the Generative Power of Speech Language Models with Prompts (arXiv 2023)

SpeechGen for Speech Inpainting

The rainbow is a division of white light into many beautiful colors





SpeechGen for Speech Continuation

Speech Continuation Result
Childless parents widows and helpless orphans broken and controlled by the master and sentence pursuit life apt to paradise.
But these king's witnesses were also put at times into the press yard and charged with the service available on a second charge to them.
And the obvious bulk of the package which he intended to bring to work was confirmed

Black text: seed segment
Red text: Continued by SpeechGen

- Grammatically coherent
- Semantic related

Summary

- Parameter efficient tuning: adapters and prompting
- Adapters for adaptation
  - Language adaptation (e.g. multilingual ASR)
  - Task adaptation (e.g. representation learning)
- Prompting speech model for new tasks
  - Prompting speech decoding models (e.g. Whisper)
  - Prompting speech generative language models (e.g. GSLM, unit BART)
- Learn more: https://github.com/ga642381/Speech-Prompts-Adapters