SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks

Kai-Wei Chang, Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee




1. Motivation
2. Method
3. Experiment & Analysis
4. Discussions

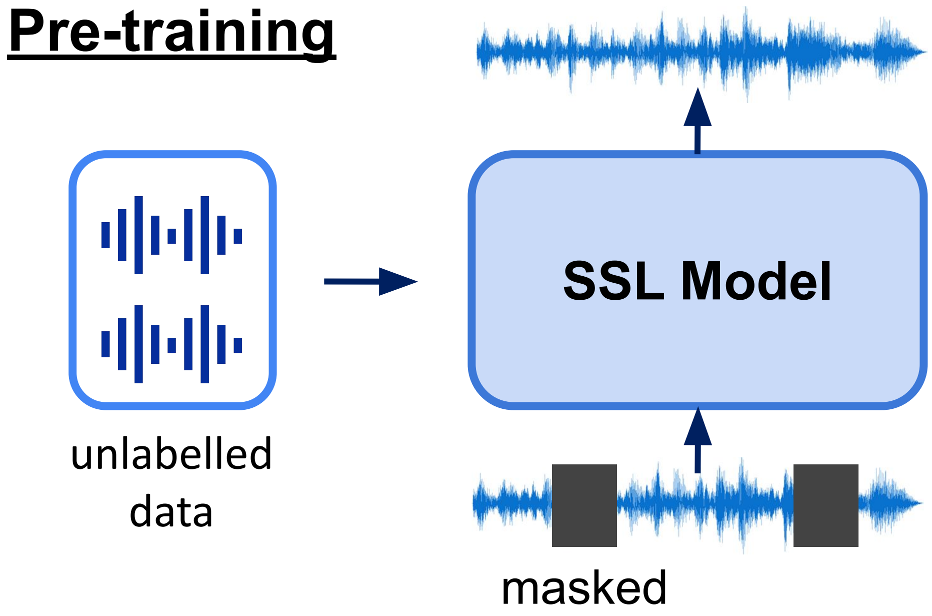
1. **Motivation**
2. Method
3. Experiment & Analysis
4. Discussions

- Pre-train, Fine-tune paradigm
- Prompting paradigm

Motivation

Pre-train, Fine-tune Paradigm

Common practices of using SSL models usually follow the **pre-train, fine-tune paradigm**

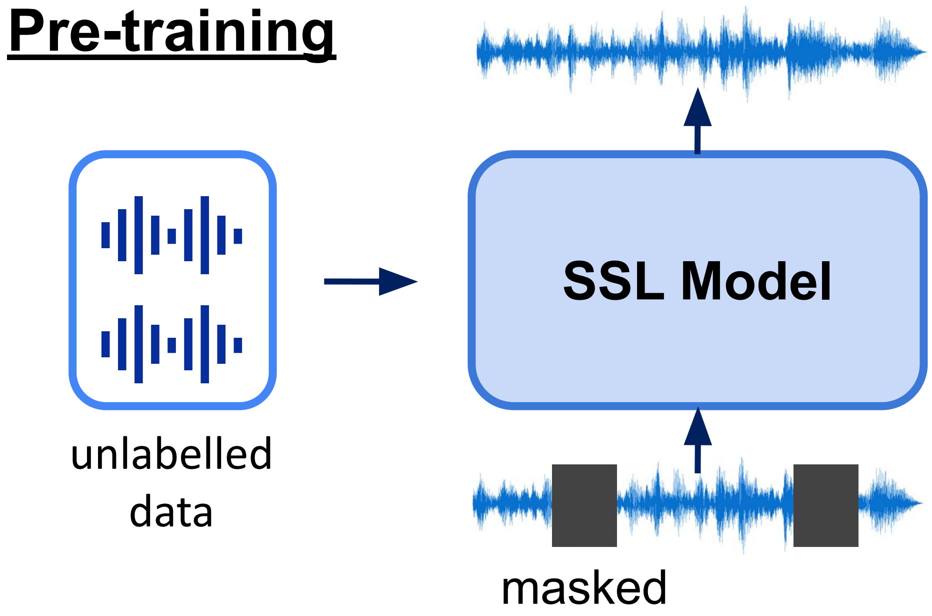


Baevski et.al., NeurIPS'20
Hsu et.al., TASLP Volume 29

Motivation

Pre-train, Fine-tune Paradigm

Common practices of using SSL models usually follow the **pre-train, fine-tune paradigm**



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

...

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

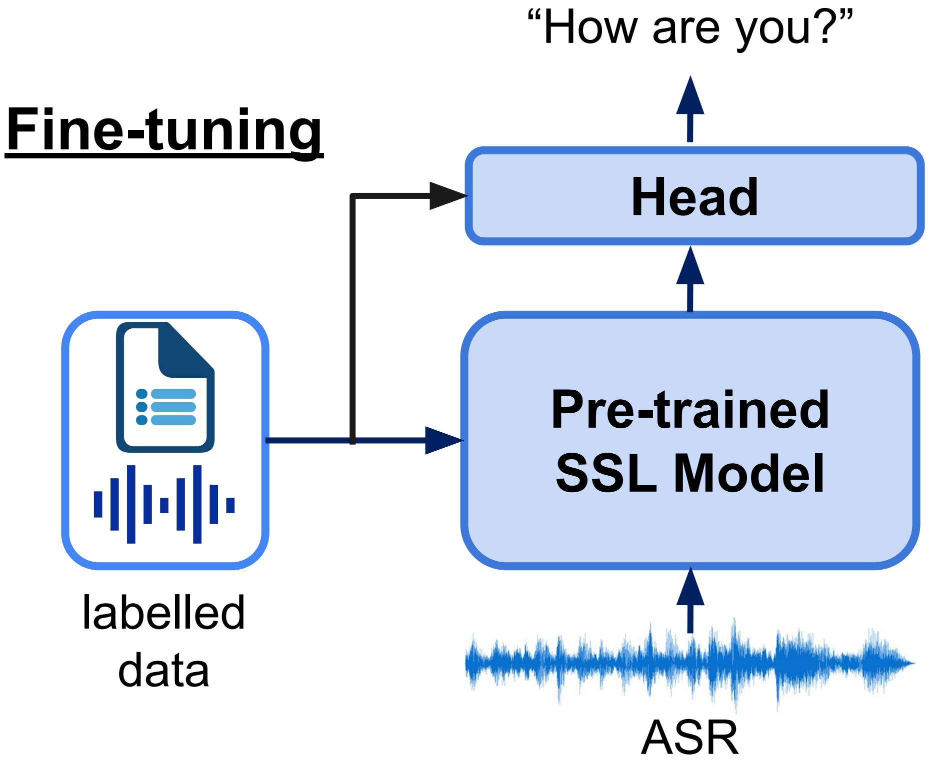
Motivation

Pre-train, Fine-tune Paradigm

For a **downstream task** (ASR):

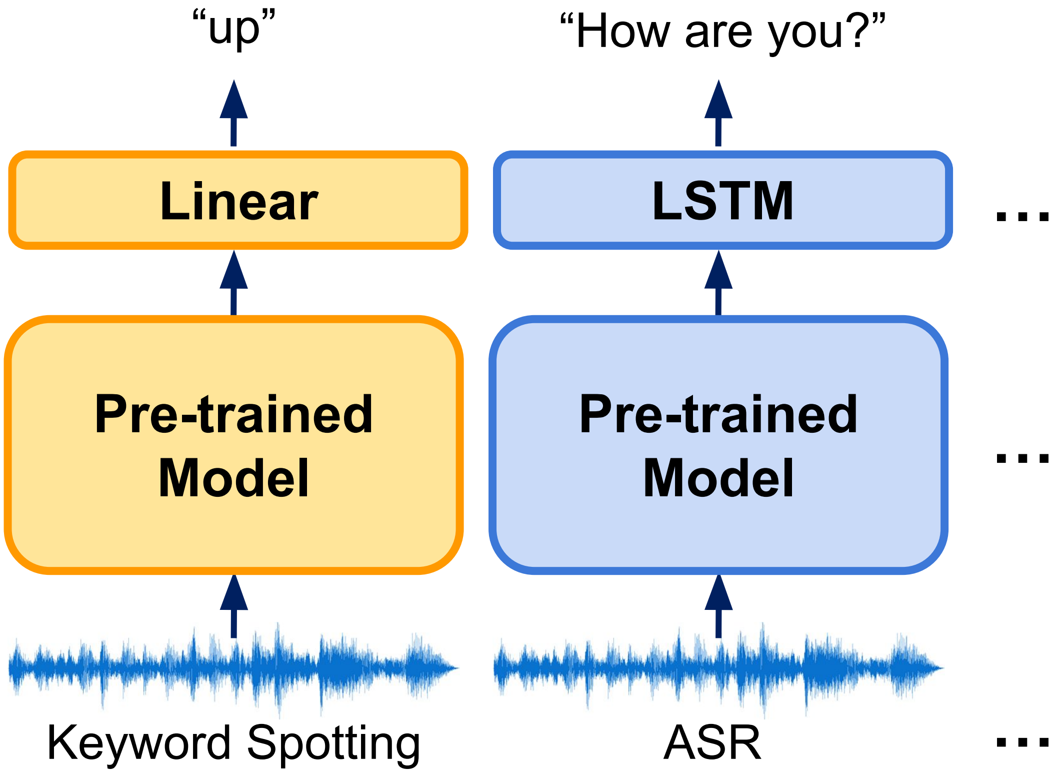
1. design a downstream head

2. fine-tune the head and the pre-trained model



Motivation

Pre-train, Fine-tune Paradigm

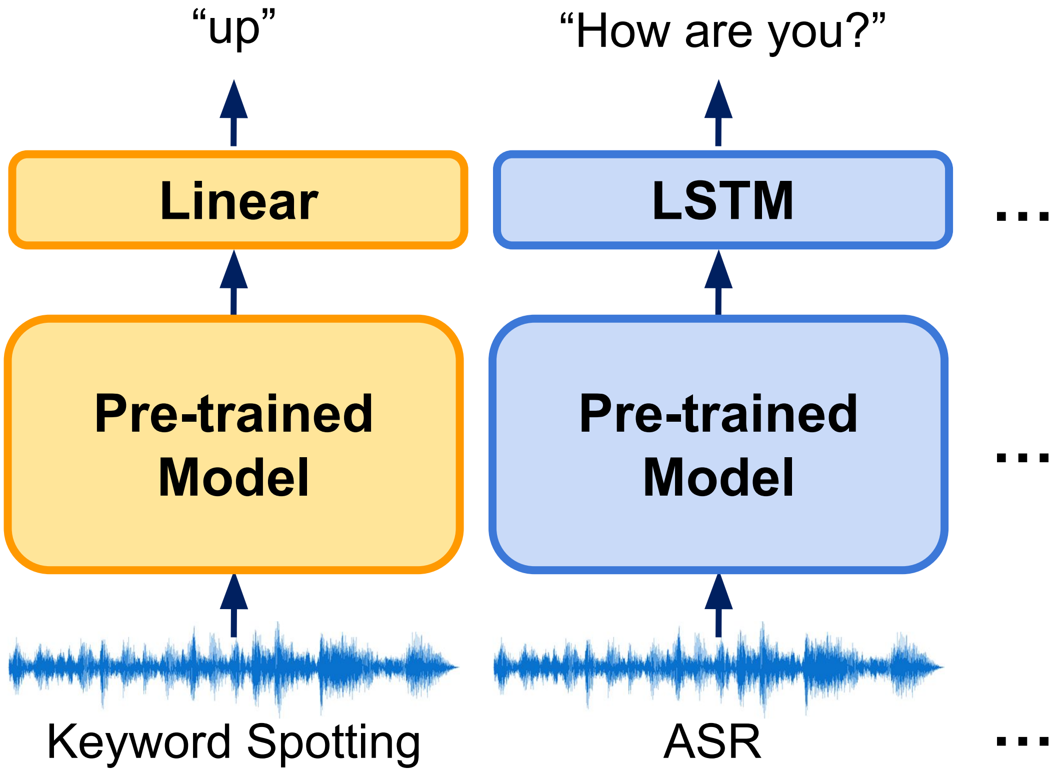


If there are lots of tasks…

- **design a head**
- **fine-tune the model**
- **save the parameters**

Motivation

Pre-train, Fine-tune Paradigm



If there are lots of tasks…

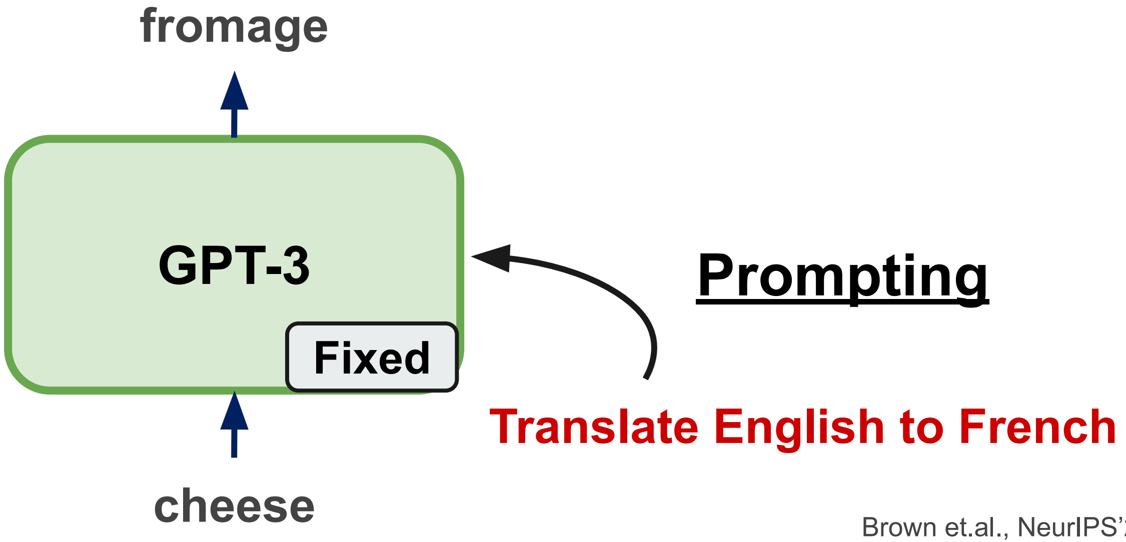
- **human labor**
- **computation cost**
- **storage cost**

Motivation

Prompting Paradigm

Prompting: make the model condition on the “prompt” and directly generate the output for the downstream task.

In NLP, prompting technology has been widely used.

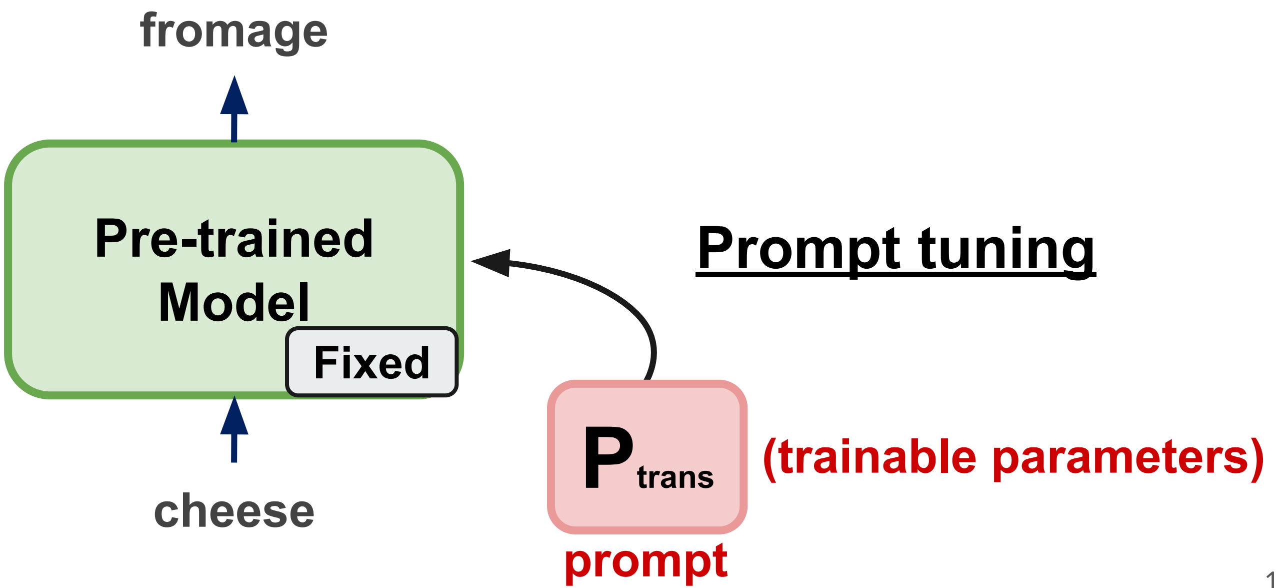


Brown et.al., NeurIPS’20

Motivation

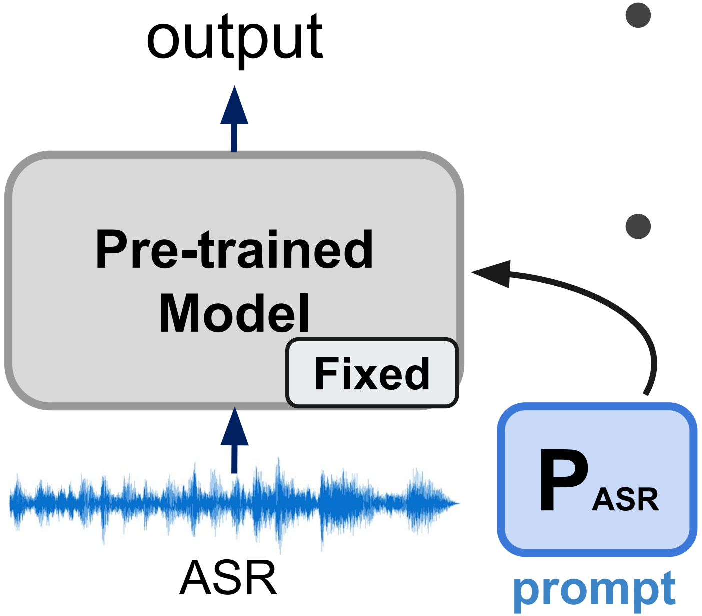
Prompting Paradigm

Prompt-tuning: The prompts are trainable parameters. It can achieve better performance than the prompts using real words



Motivation

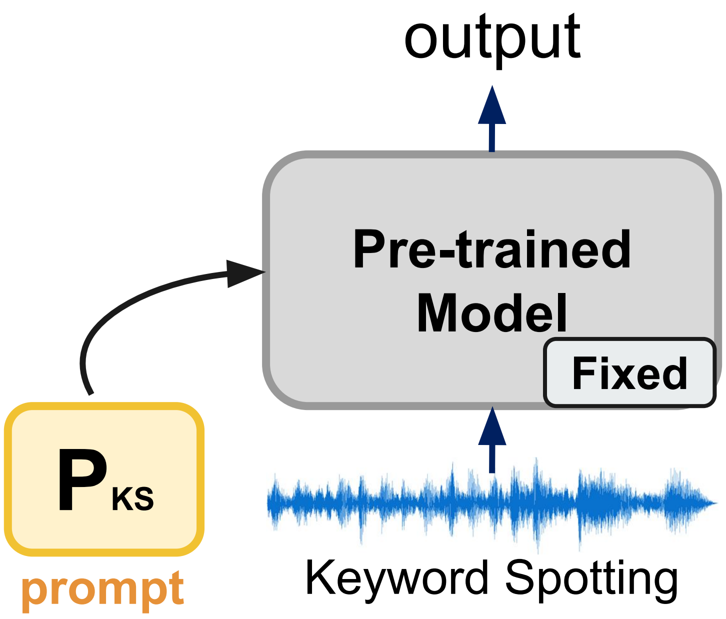
Prompting Paradigm



- Find a prompt for speech processing tasks
- Directly generate the output

Motivation

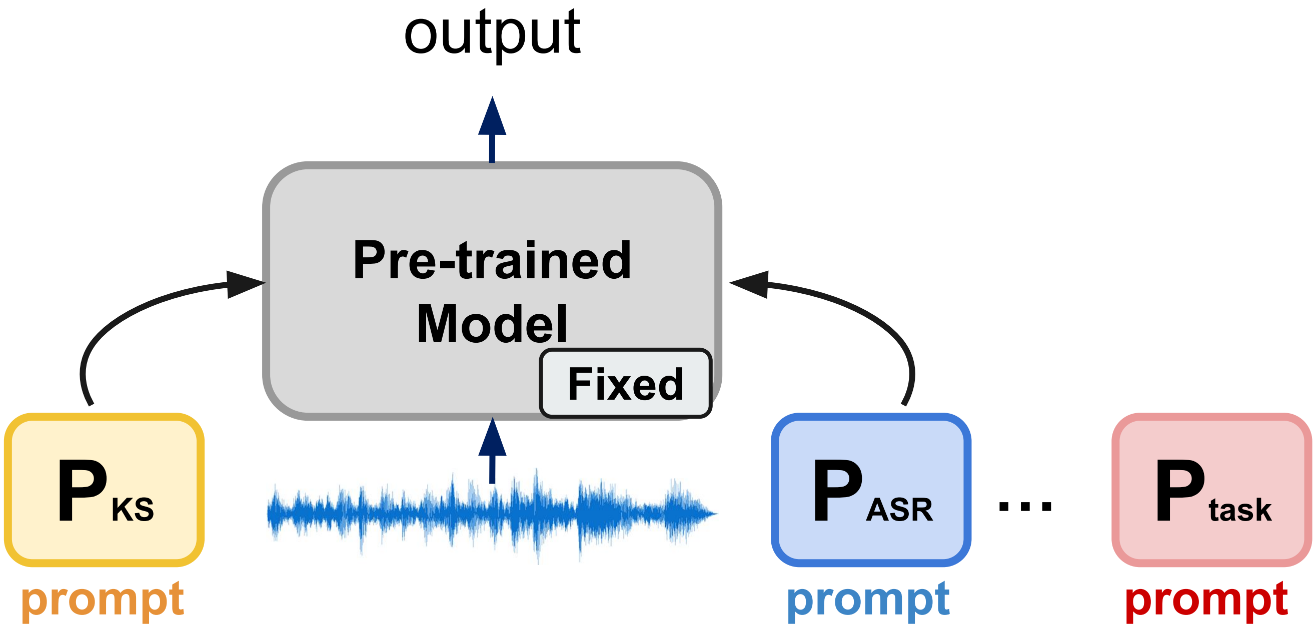
Prompting Paradigm



- Find a prompt for speech processing tasks
- Directly generate the output

Motivation

Prompting Paradigm



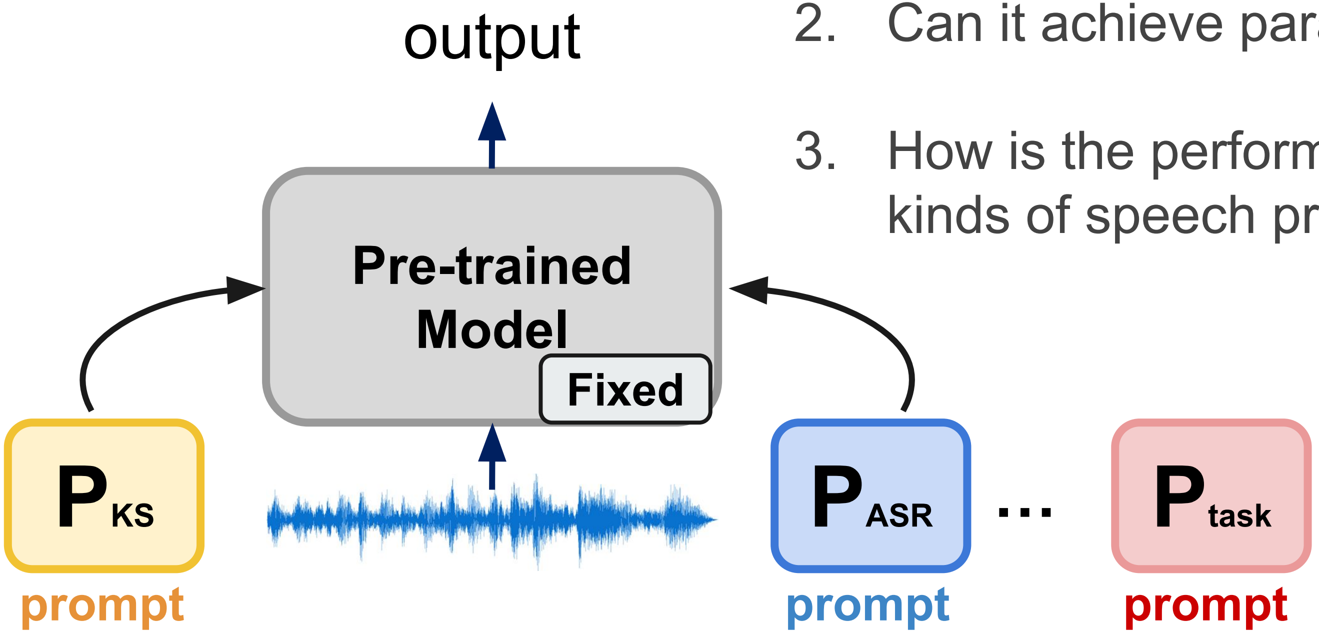
Motivation

Prompting Paradigm

1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency?

3. How is the performance for different kinds of speech processing tasks?

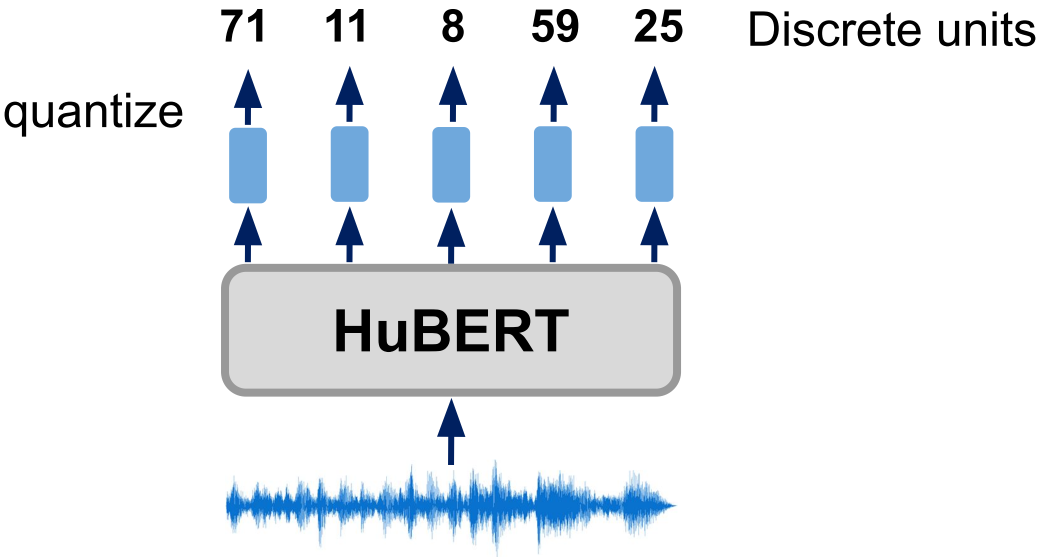


1. Motivation
2. **Method**
3. Experiment & Analysis
4. Discussions

- Background: Generative Spoken Language Model (GSLM)
- Prompt tuning on GSLM

Background - GSLM

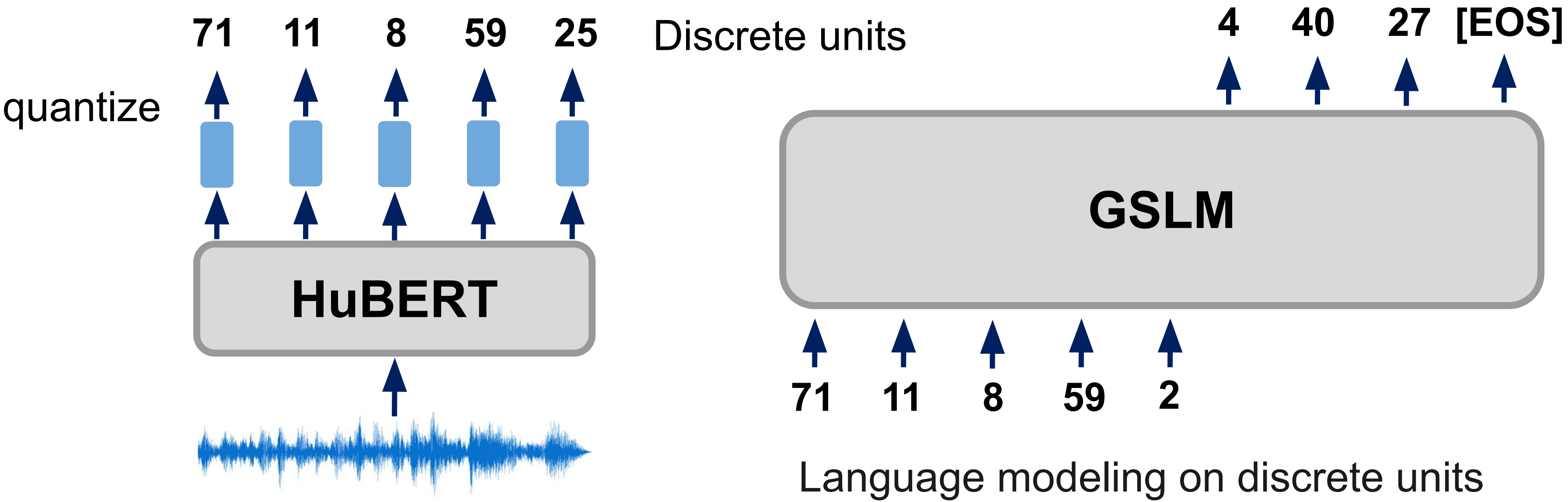
Generative Spoken Language Model



Lakhotia et.al., arXiv 21’

Background - GSLM

Generative Spoken Language Model

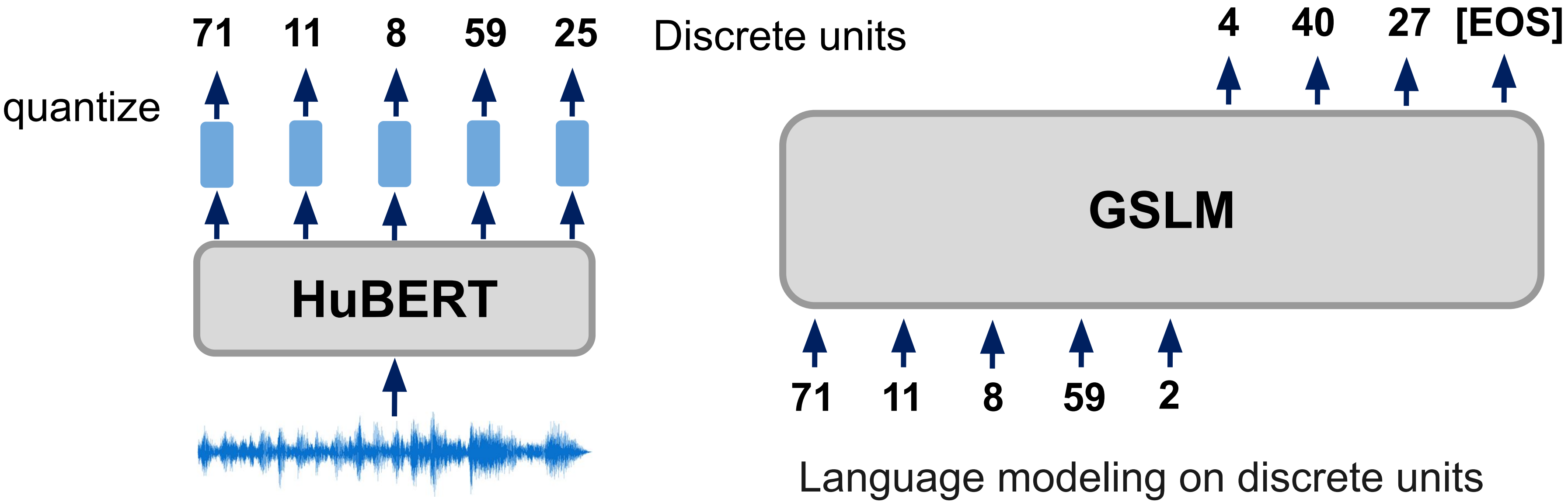


Lakhotia et.al., arXiv 21’

Background - GSLM

Generative Spoken Language Model

- **speech LM trained on a large corpus**
- **speech version of GPT-3**

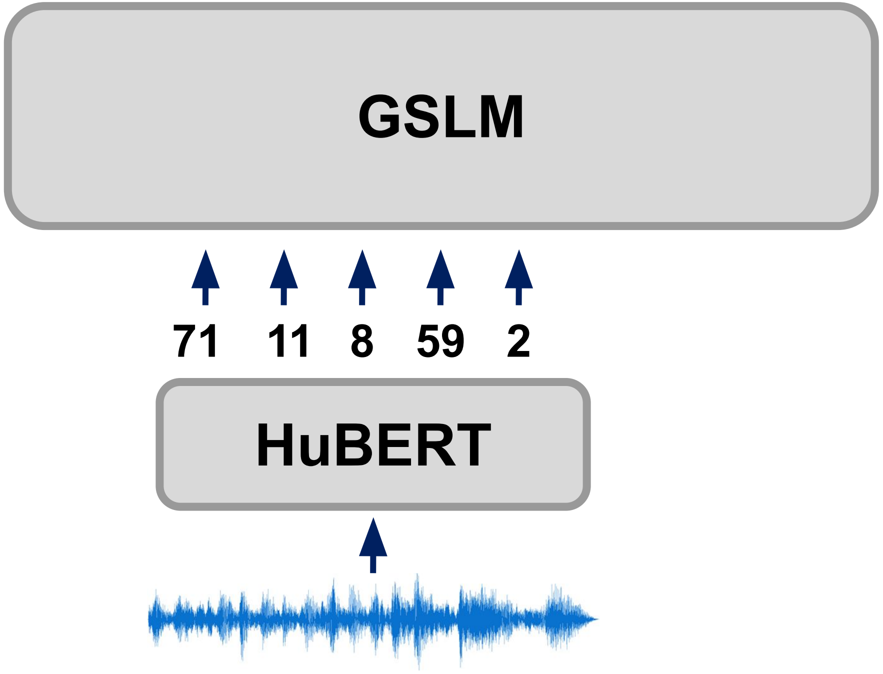


Lakhotia et.al., arXiv 21'

Prompt tuning on GSLM

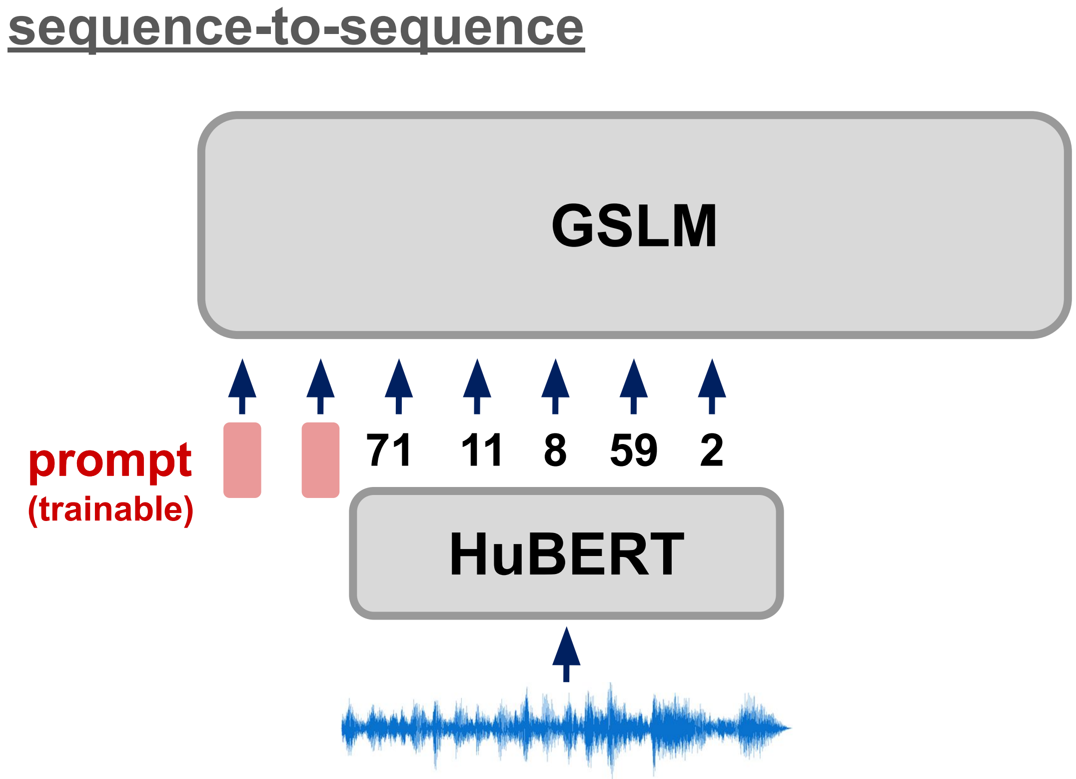
Sequence Generation
(e.g. ASR)

<u>sequence-to-sequence</u>



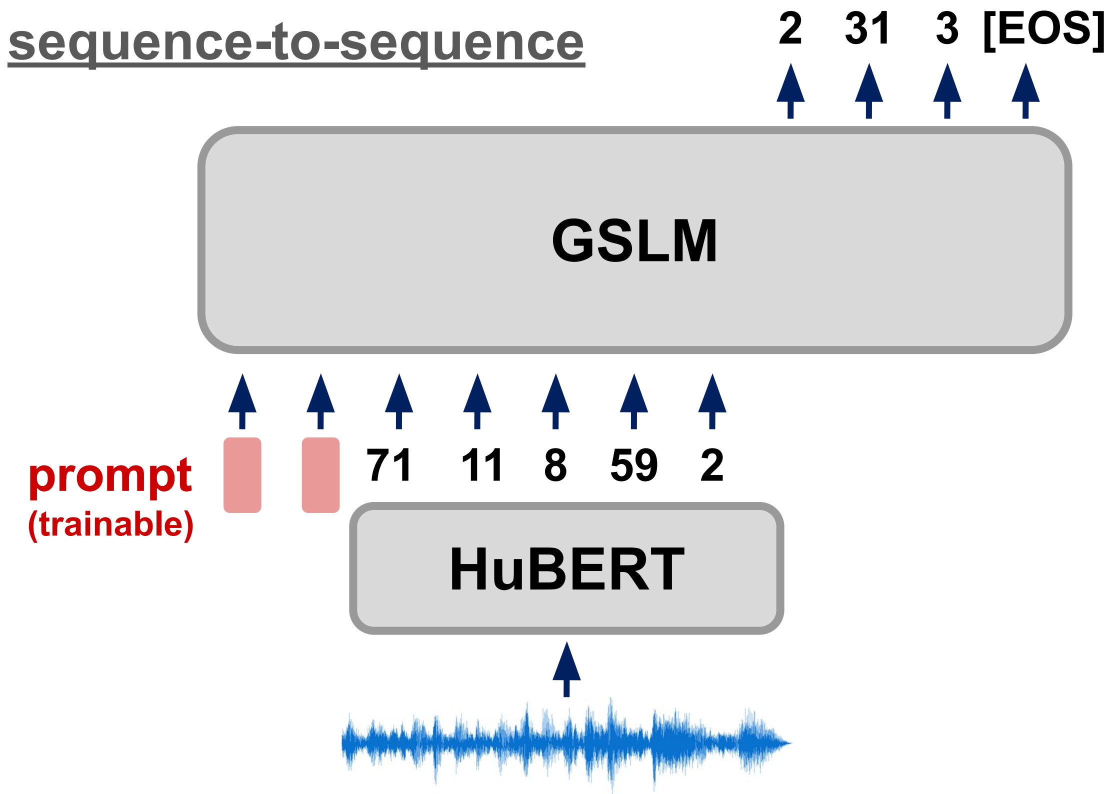
Prompt tuning on GSLM

Sequence Generation
(e.g. ASR)



Prompt tuning on GSLM

Sequence Generation
(e.g. ASR)

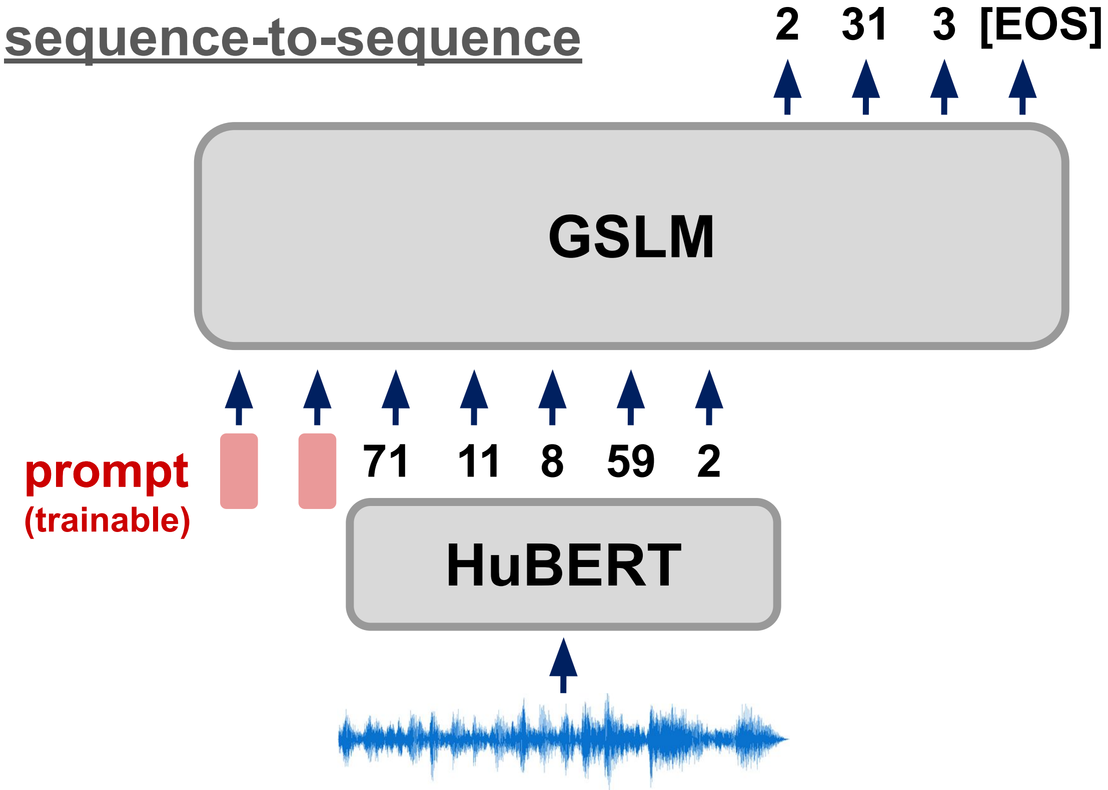


Prompt tuning on GSLM

Sequence Generation (e.g. ASR)

Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)



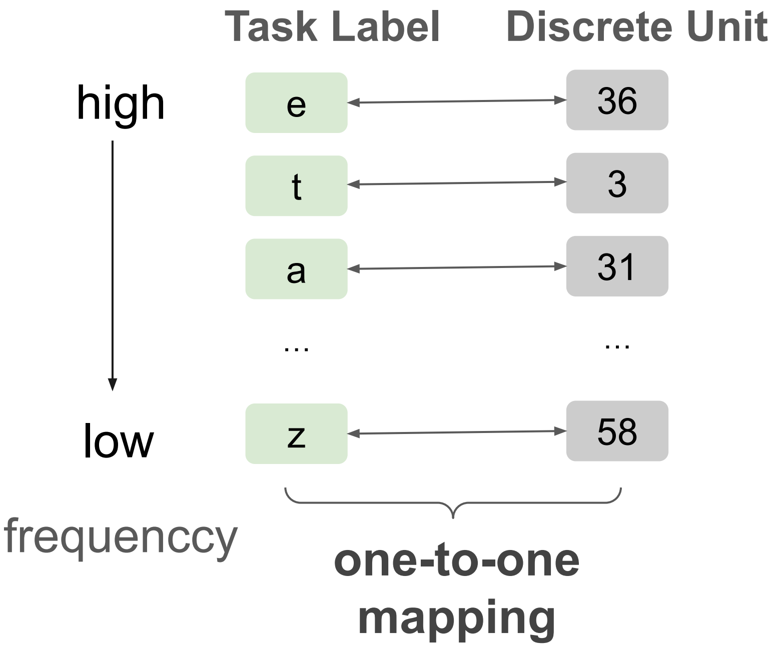
Prompt tuning on GSLM

Sequence Generation (e.g. ASR)

Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

Find and sort the top frequent task labels and discrete units from the training data and map them in order

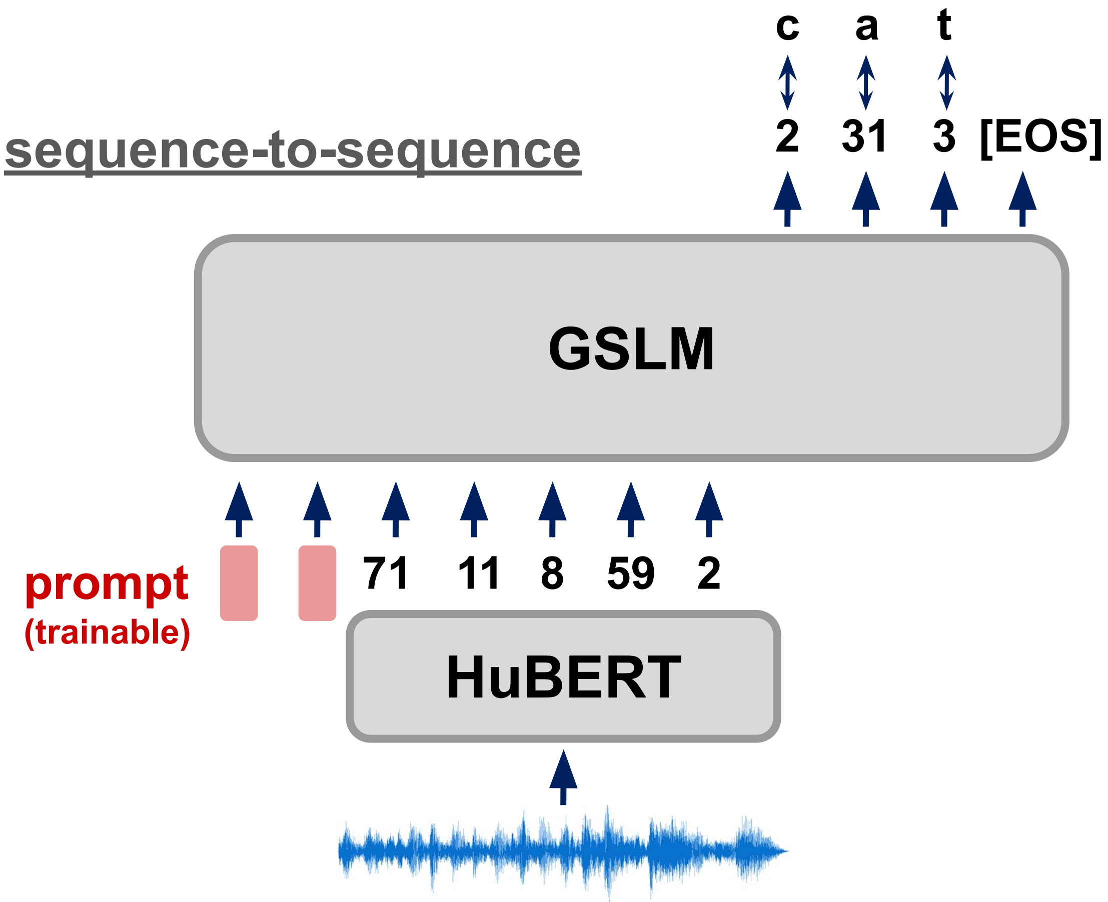


Prompt tuning on GSLM

Sequence Generation (e.g. ASR)

Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

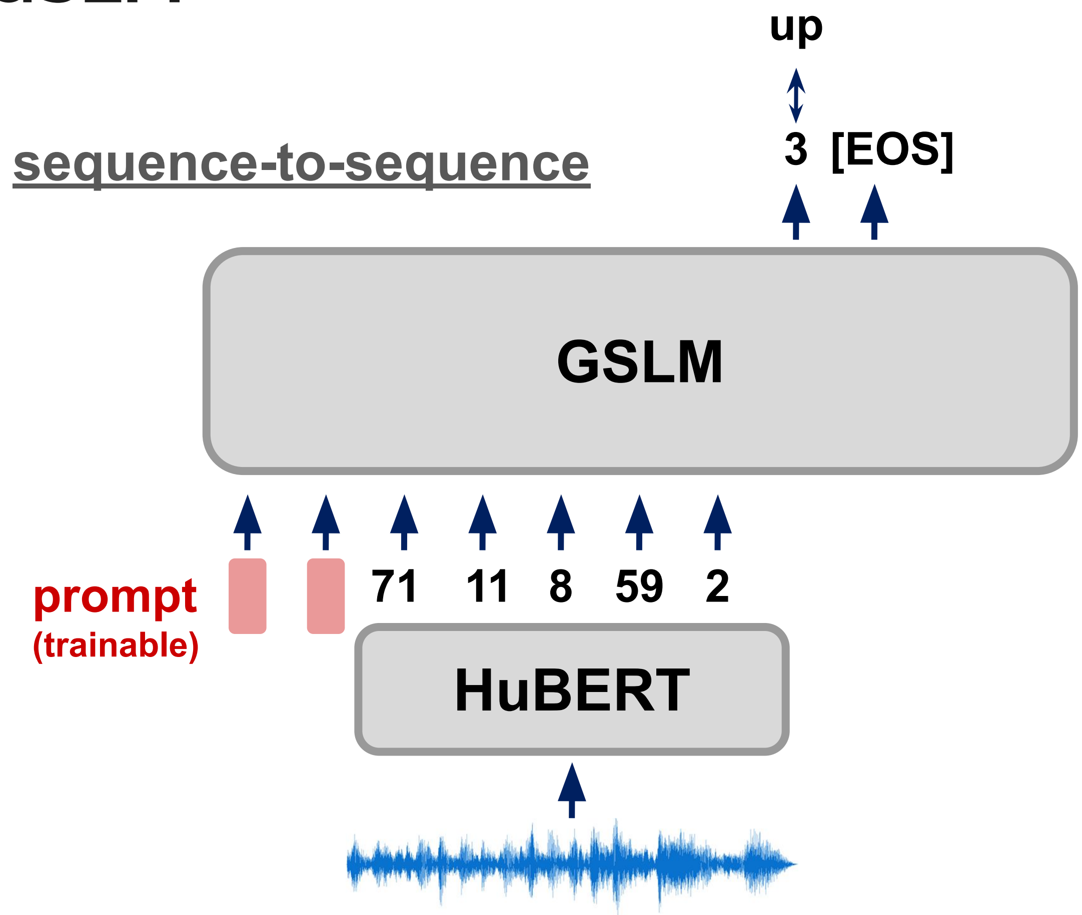


Prompt tuning on GSLM

Speech Classification
(e.g. Keyword Spotting)

Keyword	Unit ID
yes	31
no	68
up	3
down	25
…	…

Mapping table
(Verbalizer)

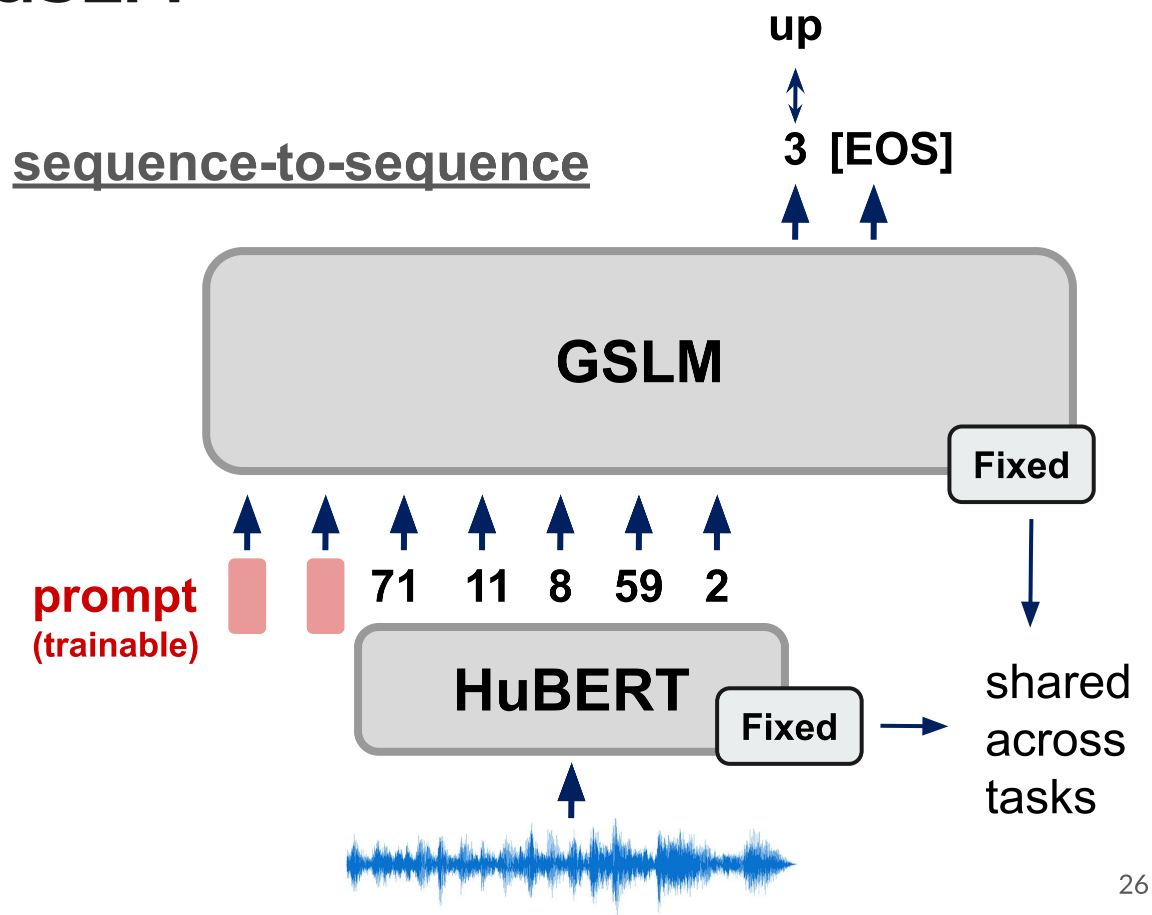


Prompt tuning on GSLM

Speech Classification (e.g. Keyword Spotting)

Keyword	Unit ID
yes	31
no	68
up	3
down	25
…	…

Mapping table (Verbalizer)

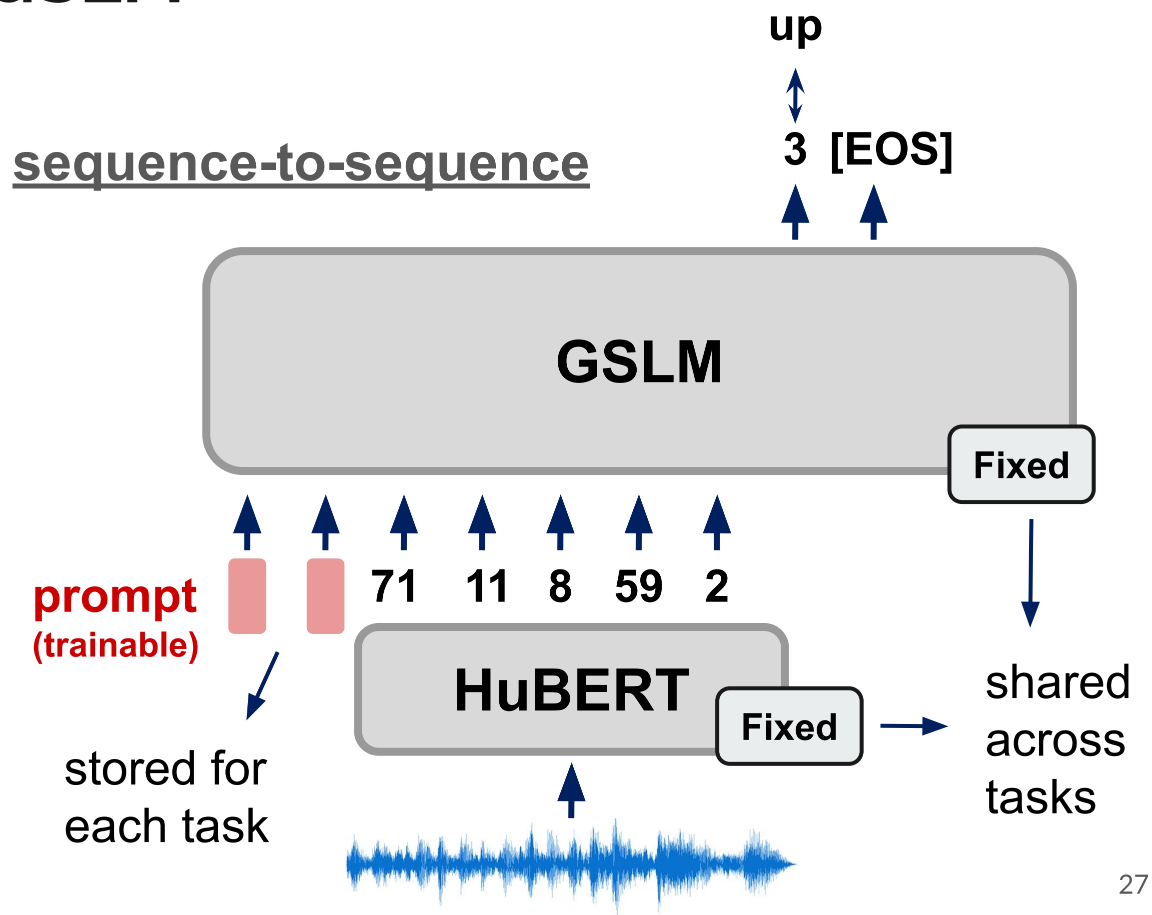


Prompt tuning on GSLM

Speech Classification
(e.g. Keyword Spotting)

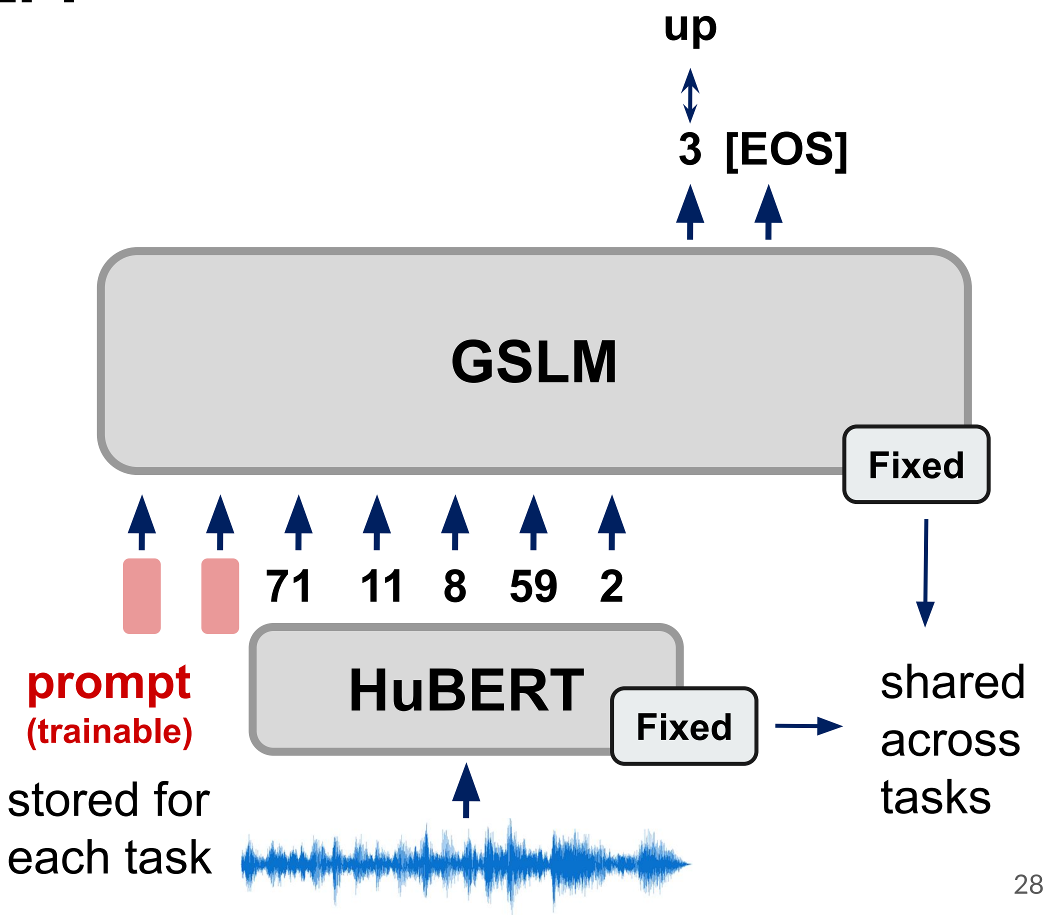
Keyword	Unit ID
yes	31
no	68
up	3
down	25
…	…

Mapping table (Verbalizer)



Prompt tuning on GSLM

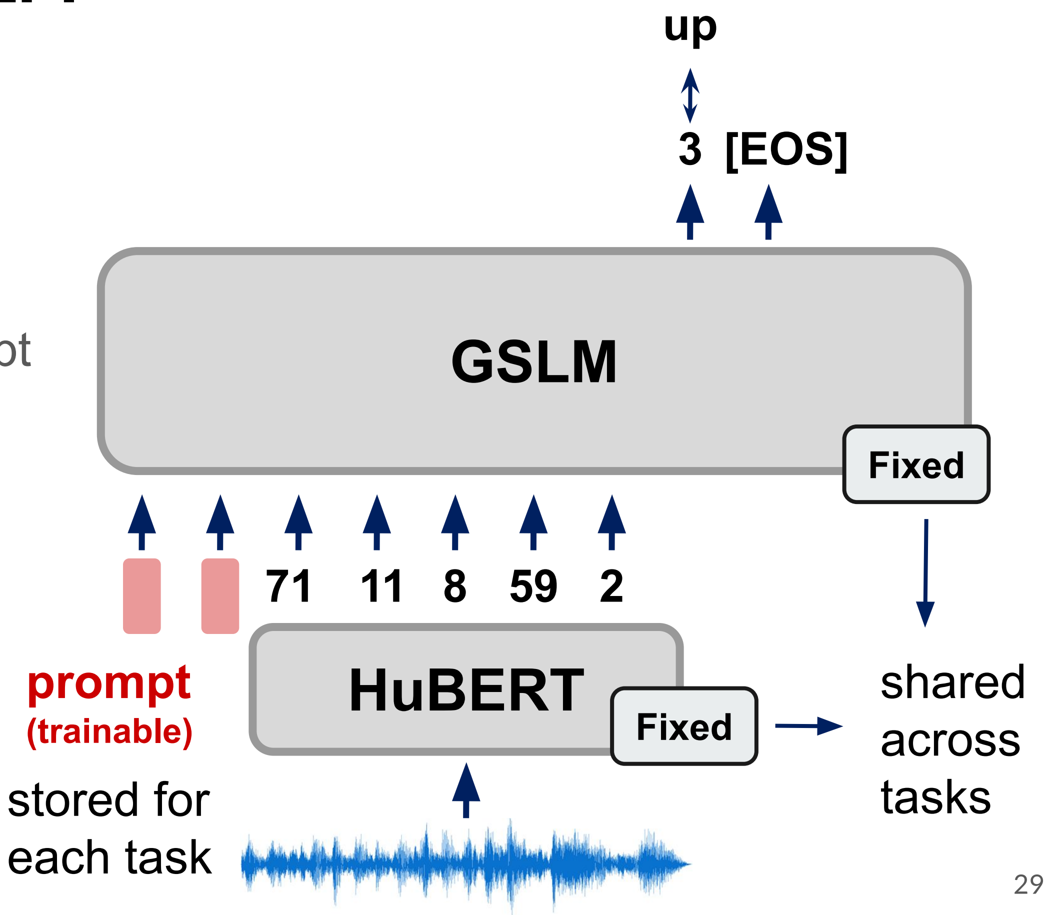
- Speech processing tasks are formulated into a seq2seq task →unified framework



Prompt tuning on GSLM

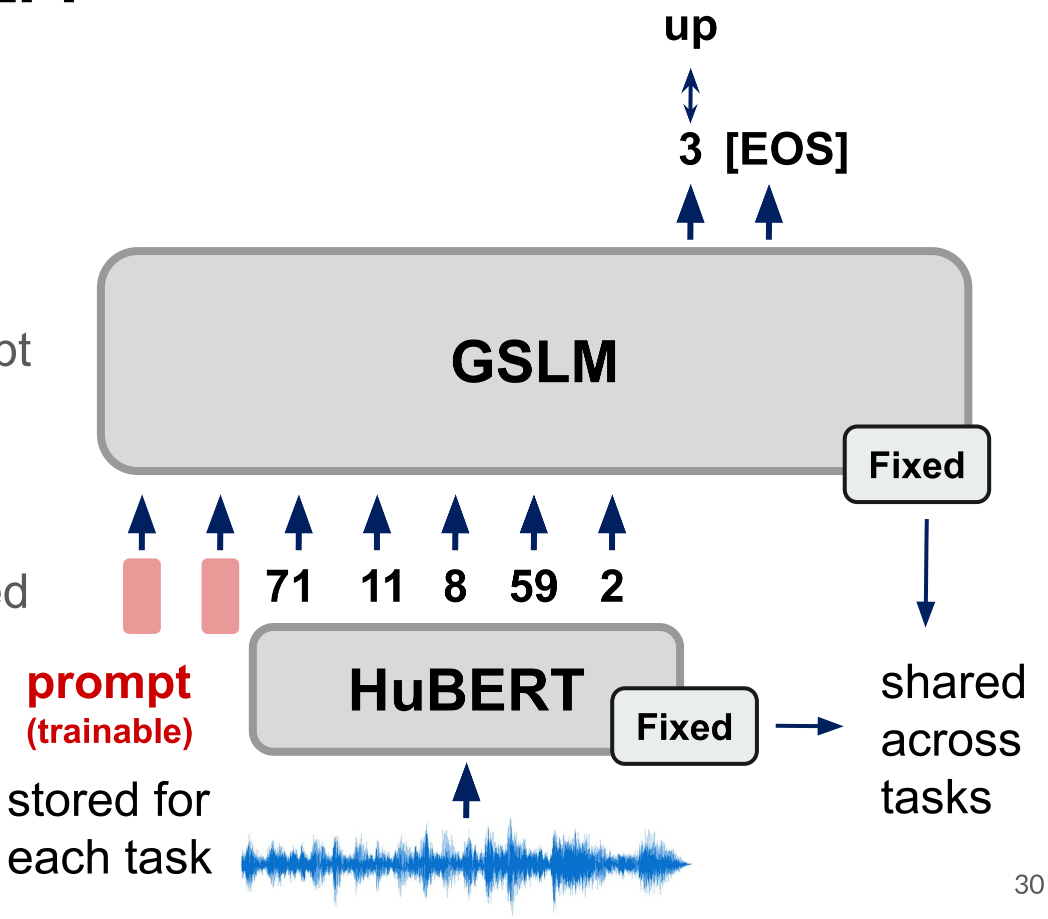
- Speech processing tasks are formulated into a seq2seq task
→unified framework

- We only need to train the prompt for each task
→computation efficient



Prompt tuning on GSLM

- Speech processing tasks are formulated into a seq2seq task →unified framework
- We only need to train the prompt for each task →computation efficient
- Only the prompt has to be saved for each task →parameter efficient (storage saving)



Prompt tuning on GSLM

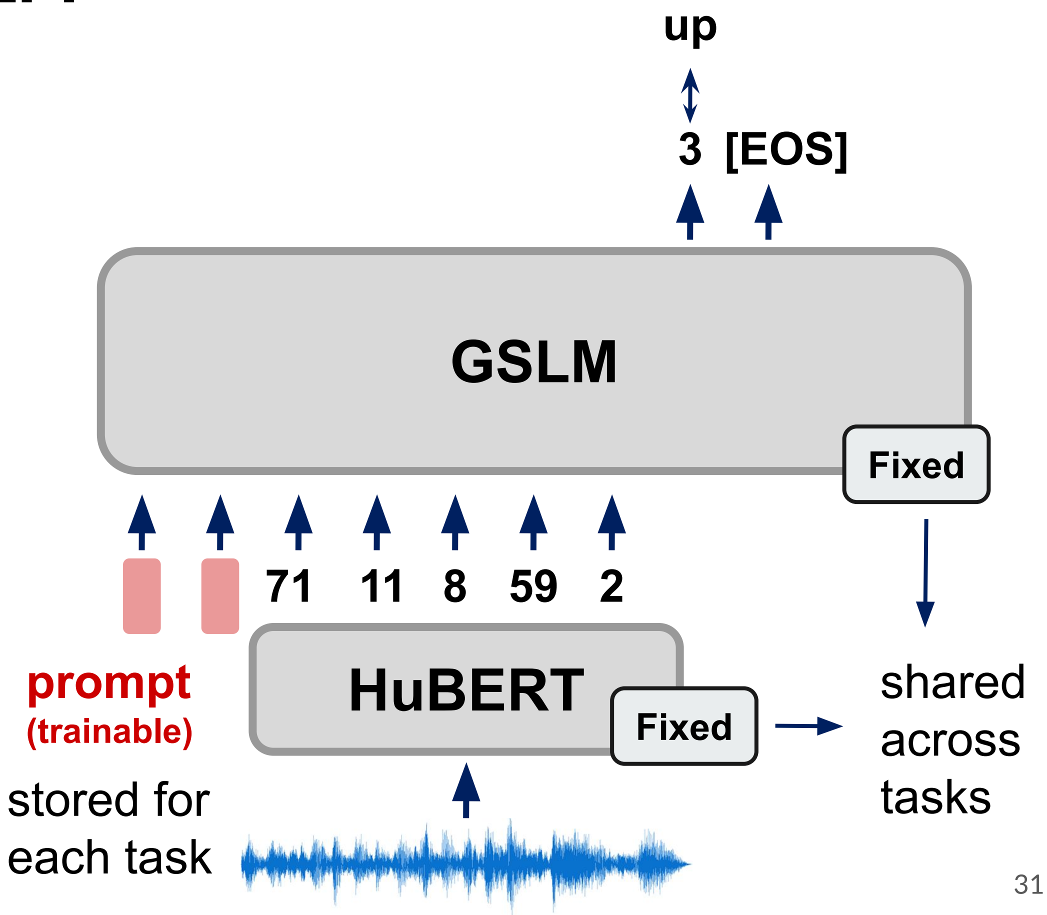
Prefix Tuning [Li and Liang ACL'21]

Prompts are prepended at:

1. Input embedding

2. Input of each Transformer layer

Prompts are at the input side. The pre-trained model is not modified



1. Motivation
2. Method
3. **Experiment & Analysis**
4. Discussions

- Speech classification tasks
- Sequence generation tasks
- Analysis

Experiment Setup

SUPERB

- **CLS**: Classification
- **SG**: Sequence Generation
- |y|: average label length

Task		Type	N_class	|y|
Keyword Spotting	KS	CLS	12	1
Intent Classification	IC	CLS	24	3
Speech Recognition	ASR	SG	29	173
Slot Filling	SF	SG	69	54

Yang et.al., INTERSPEECH 21’

Experiment Setup

SUPERB

- **CLS**: Classification
- **SG**: Sequence Generation
- |y|: average label length

Task		Type	N_class	|y|
Keyword Spotting	KS	CLS	12	1
Intent Classification	IC	CLS	24	3
Speech Recognition	ASR	SG	29	173
Slot Filling	SF	SG	69	54

Yang et.al., INTERSPEECH 21’

Experiment Setup

- Datasets:
  - Keyword Spotting: Speech Command
  - Intent Classification: Fluent Command
  - Speech Recognition: LibriSpeech-100
  - Slot Filling: Audio SNIPS

- Pre-trained models
  (SSL models and the corresponding GSLM)
  - HuBERT [Hsu et.al., TASLP Volume 29]
  - CPC [Oord et.al., arXiv 18']

Experiment Results - Speech Classification

- PT: Prompt Tuning
- FT: Fine-Tuning
- KS: Keyword Spotting - Single-label Cls.
- IC: Intent Classification - Multi-label Cls.

Scenarios	KS		IC	
	ACC↑	# param.	ACC↑	# param.
HuBERT-PT	95.16	0.08M	**98.40**	0.15M
HuBERT-FT	**96.30**	0.2M	98.34	0.2M

Fine-tuning downstream linear model

Prompt tuning achieves competitive performance with fewer trainable parameters

Experiment Results - Speech Classification

- PT: Prompt Tuning
- FT: Fine-Tuning
- KS: Keyword Spotting - Single-label Cls.
- IC: Intent Classification - Multi-label Cls.

Scenarios	KS		IC	
	ACC↑	# param.	ACC↑	# param.
CPC-PT	**93.54**	0.05M	**97.57**	0.05M
CPC-FT	91.88	0.07M	64.09	0.07M

Fine-tuning downstream linear model

Prompt tuning achieves competitive performance with fewer trainable parameters

Experiment Results - Speech Classification

- PT: Prompt Tuning
- FT: Fine-Tuning
- KS: Keyword Spotting - Single-label Cls.
- IC: Intent Classification - Multi-label Cls.

Scenarios	KS		IC	
	ACC↑	# param.	ACC↑	# param.
CPC-PT	**93.54**	0.05M	**97.57**	0.05M
CPC-FT	91.88	0.07M	64.09	0.07M

The advantage of prompt tuning is even more obvious in Intent Classification for CPC

Experiment Results - Sequence Generation

- PT: Prompt Tuning
- FT: Fine-Tuning
- ASR: Automatic Speech Recognition
- SF: Slot Filling

Scenarios	ASR		SF	
	WER↓	# param.	F1↑	# param.
HuBERT-PT	34.17	4.5M	66.90	4.5M
HuBERT-FT	**6.42**	43M	**88.53**	43M

Fine-tuning downstream LSTM model

Prompt tuning is not competitive but with ~10 times fewer trainable parameters.

Experiment Results - Sequence Generation

- PT: Prompt Tuning
- FT: Fine-Tuning
- ASR: Automatic Speech Recognition
- SF: Slot Filling

Scenarios	ASR		SF	
	WER↓	# param.	F1↑	# param.
CPC-PT	59.41	4.5M	65.25	4.5M
CPC-FT	**20.18**	42.5M	**71.19**	42.5M

Fine-tuning downstream LSTM model

Prompt tuning is not competitive but with ~10 times fewer trainable parameters.

Analysis - The Curse of Long Sequences

- Analyze the performance and the data in ASR (LibriSpeech test-clean split)
- label length: #characters

Task	Type	Avg. label length
KS	CLS	1
IC	CLS	3
ASR	SG	173
SF	SG	54

Analysis - The Curse of Long Sequences

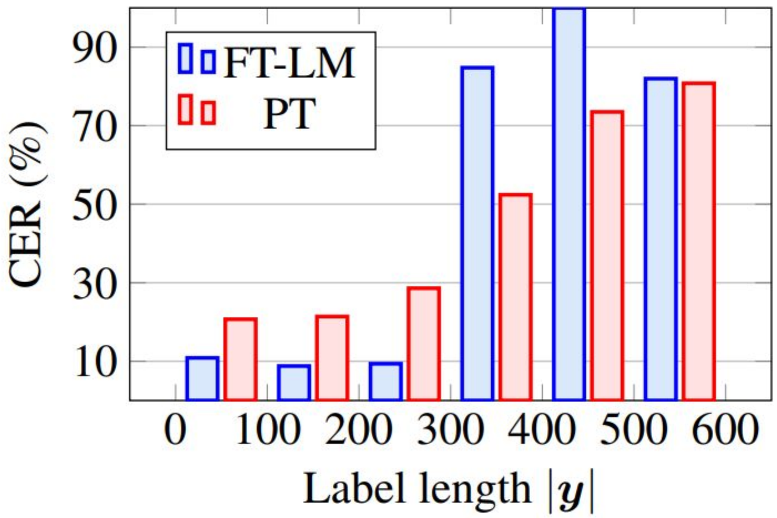
Divide the test dataset into several splits according to their label lengths

Plot their CER for

- PT: Prompt Tuning
- FT-LM: Fine-Tuning the whole GSLM

- The performance suffers from long sequences severely
- The performance might be restricted by the GSLM itself

ASR



1. Motivation
2. Method
3. Experiment & Analysis
4. **Discussions**

- Conclusions
- Future works

Conclusions

1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency?

3. How is the performance for different kinds of speech processing tasks?

Conclusions

1. Can prompting technology be applied to speech processing? **Yes**
2. Can it achieve parameter efficiency?
3. How is the performance for different kinds of speech processing tasks?

Conclusions

1. Can prompting technology be applied to speech processing? **Yes**
2. Can it achieve parameter efficiency? **Yes**
3. How is the performance for different kinds of speech processing tasks?

Conclusions

1. Can prompting technology be applied to speech processing? **Yes**

2. Can it achieve parameter efficiency? **Yes**

3. How is the performance for different kinds of speech processing tasks?
    - **Competitive for speech classification tasks**
    - **Underperform for sequence generation tasks**

Conclusions

1. Can prompting technology be applied to speech processing? **Yes**
2. Can it achieve parameter efficiency? **Yes**
3. How is the performance for different kinds of speech processing tasks?
    - **Competitive for speech classification tasks**
    - **Underperform for sequence generation tasks**

- The first exploration of prompt tuning for different kind of speech processing tasks.
- source code: https://github.com/ga642381/SpeechPrompt

Future Works

For sequence generation tasks, the performance suffers from “long sequences”

- Applying sequence compression/denoising techniques

Different from NLP, the discrete units are not meaningful

- Construct a better label mapping (e.g. learnable verbalizer)

Acknowledgement

Group's Website

https://jsalt-2022-ssl.github.io/

2022 Eighth Frederick Jelinek Memorial Summer Workshop

The Workshop June 27 to August 5, 2022

About the Eighth Frederick Jelinek Memorial Summer Workshop

The JSALT 2022 Program

JHU Summer School on Human Language Technology (June 13 June 24)

Opening Day Presentations Schedule (June 27)

Plenary Lectures by Invited Speakers (June 29, July 6, 13, 20, 27)

Closing Day Presentations (August 4 and 5)

Research Groups

- Speech Translation for Under-Resourced Languages
- Multilingual and Code-Switching Speech Recognition
- Leveraging Pre-Training Models for Speech Processing

References

- Yang et.al., INTERSPEECH 21’, SUPERB: Speech processing Universal PERformance Benchmark
- Hsu et.al., IEEE/ACM TASLP Volume 29, HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units
- Oord et.al., arXiv 18’, Representation Learning with Contrastive Predictive Coding
- Baevski et.al., NeurIPS’20, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations
- Brown et.al., NeurIPS’20, Language Models are Few-Shot Learners

Thanks for your listening!