Towards a Universal Speech Model

Prompting Speech Language Models for Diverse Speech Processing Tasks

張凱爲
Kai-Wei Chang

Advisor: Dr. Hung-yi Lee

Date: 2025/01/06

Outline

- Background
  - Pre-train, fine-tune paradigm vs.Prompting paradigm
  - Textless Speech Language Models
- SpeechPrompt: Prompting Speech LM for diverse tasks
- Exploring In-context Learning for Speech Language Model
- Conclusion
- Future Works

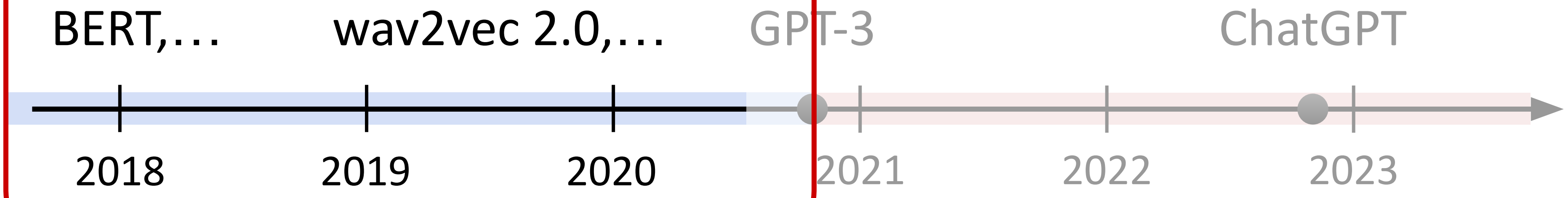
Background

Pre-train, Fine-tune Paradigm

Self-supervised Learning (SSL)

Prompting Paradigm

Prompt engineering

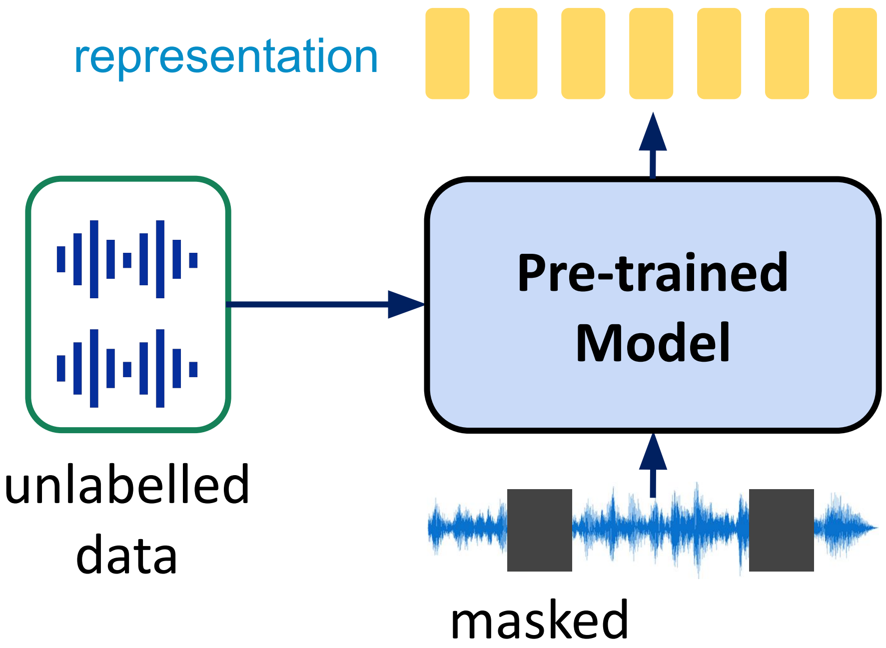


Representation Models

Pre-train, Fine-tune Paradigm

SSL objective
e.g. masked prediction

Pre-training stage

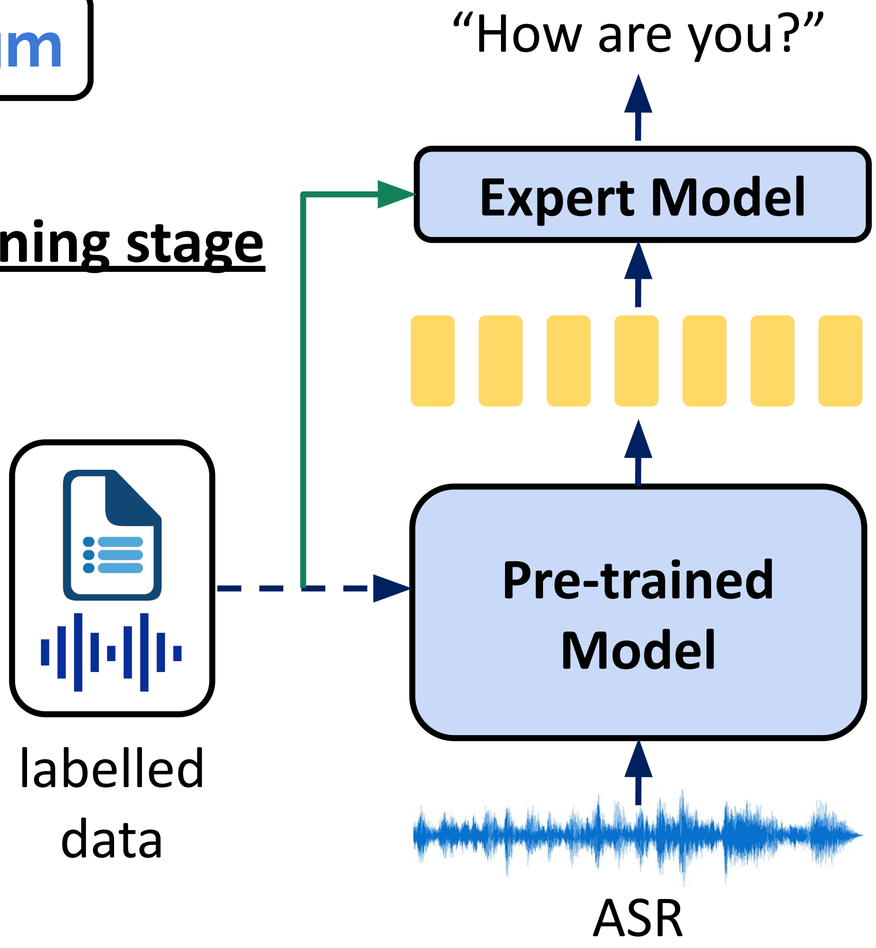


Pre-train, Fine-tune Paradigm

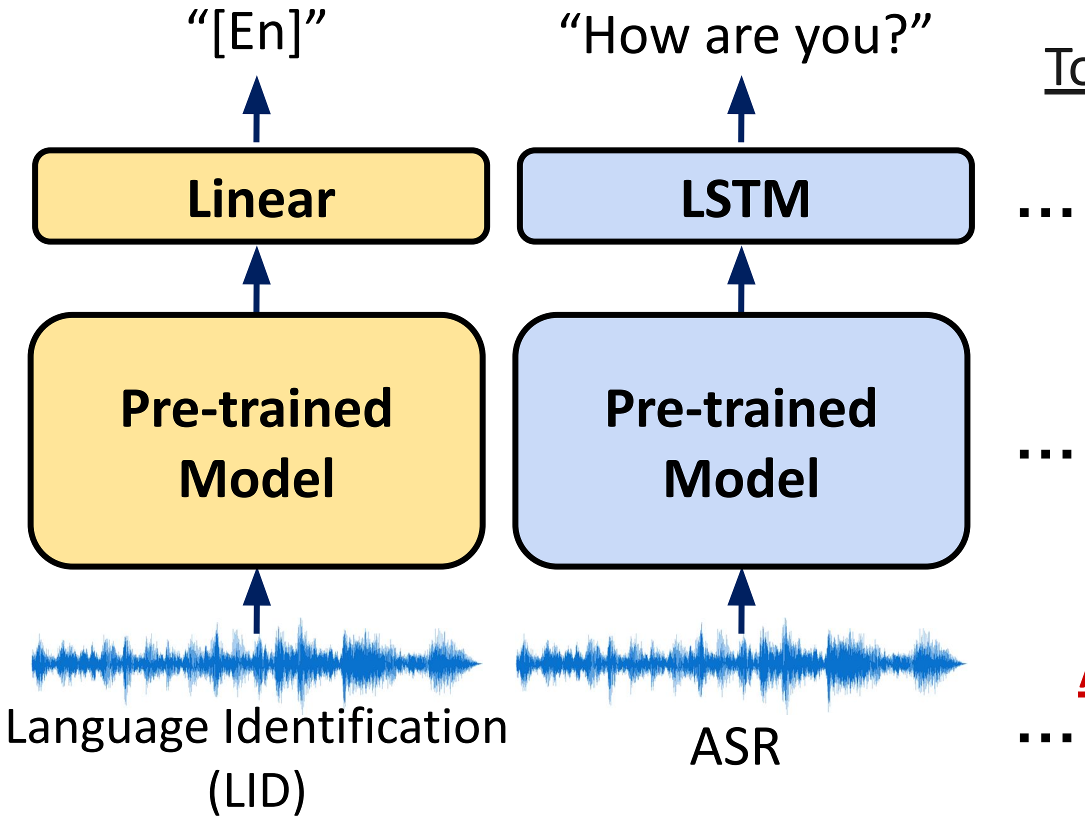
For a **downstream task:** **Fine-tuning stage**

1. Design a downstream expert model

2. **Fine-tune** the model with task-specific loss (e.g. CTC loss)



Pre-train, Fine-tune Paradigm



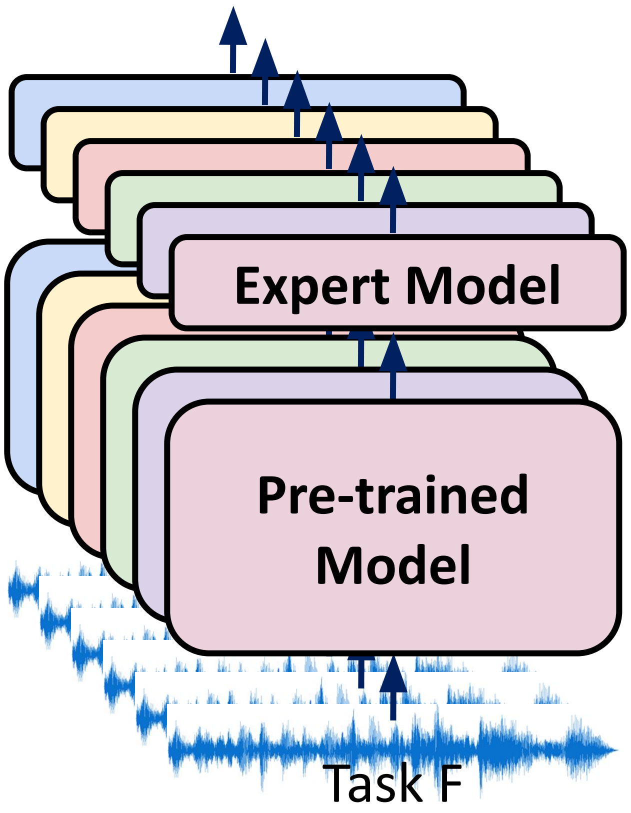
To perform a downstream task:

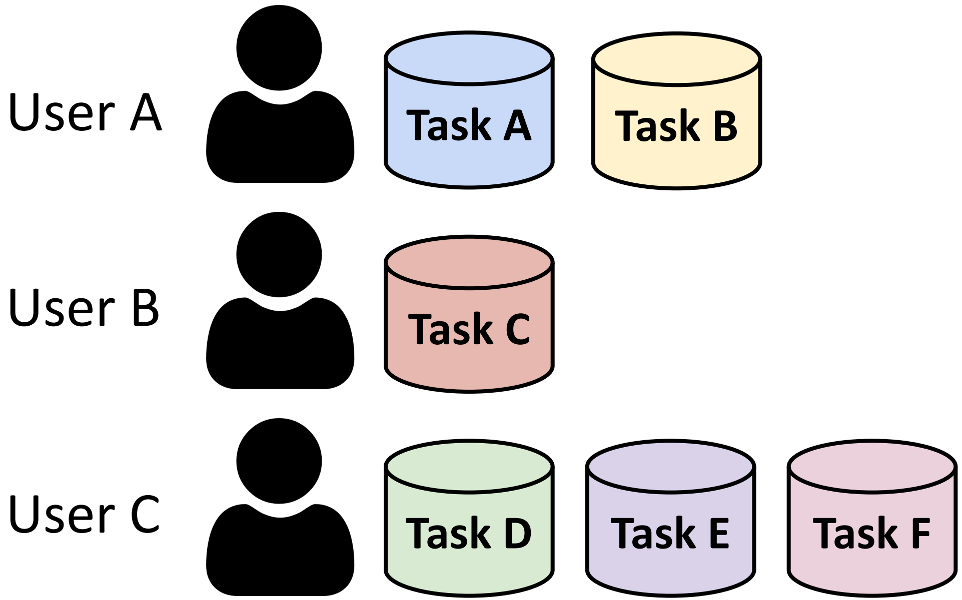
- Design an expert model
- Fine-tune the model
- Save the parameters

Achieve good performance

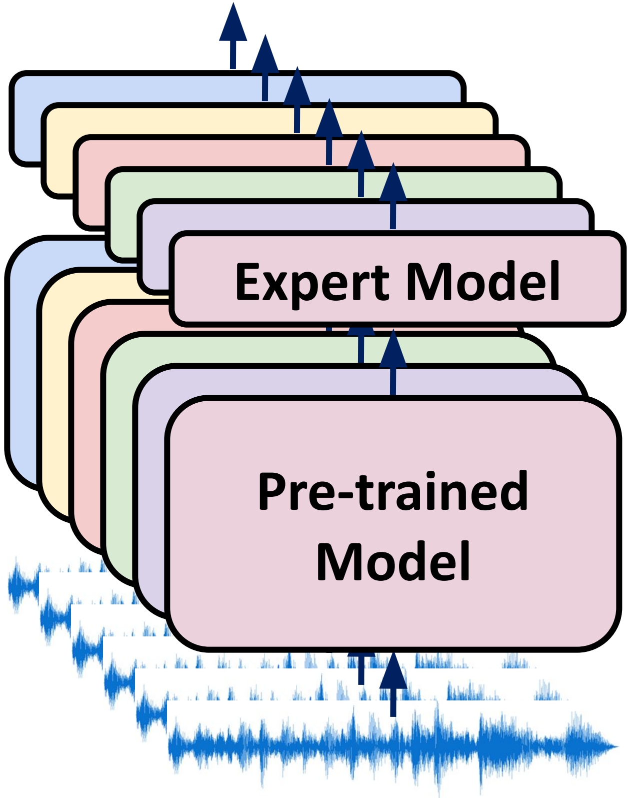
Pre-train, Fine-tune Paradigm

If you want to serve lots of users…





Pre-train, Fine-tune Paradigm



If there are lots of tasks to serve…

- Design an expert model
  human labor
- Fine-tune the model
  computational cost
- Save the parameters
  storage cost

Difficult to scale!

Research Question:

Is it possible to build a universal and efficient speech processing system?

Research Question:

Is it possible to build a universal and efficient speech processing system?

Solve diverse speech processing tasks in a unified manner

↓

No need to design expert models

Research Question:

Is it possible to build a universal and efficient speech processing system?

Trainable parameter efficiency

↓

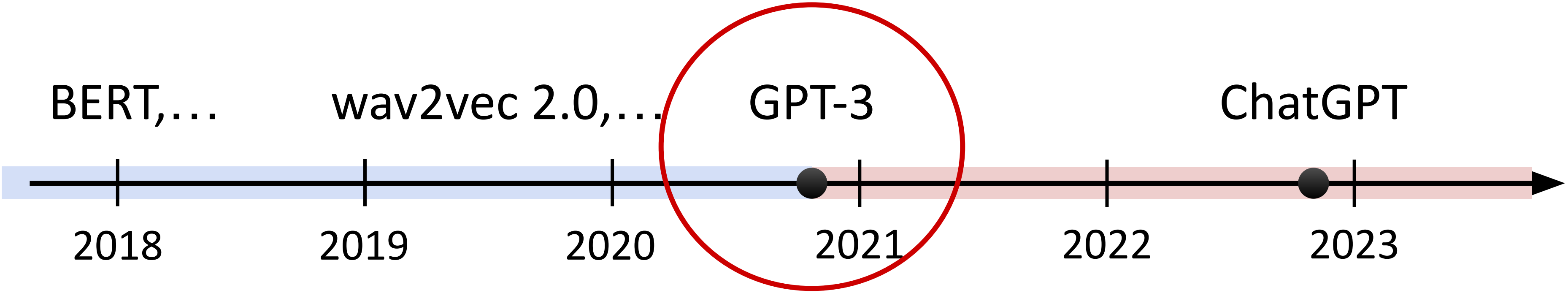
Computation and storage efficiency

Research Question:

Is it possible to build a universal and efficient speech processing system?

Inspiration: Prompting paradigm in NLP

Pre-train, Fine-tune Paradigm **Prompting Paradigm**



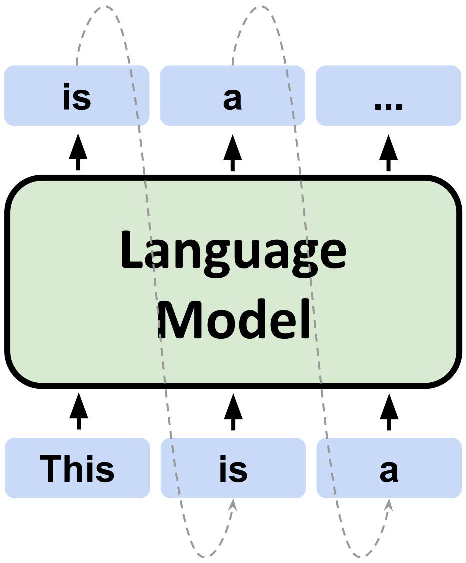
Prompting gained more and more attention

Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Prompting Paradigm

Decoder-only LM (e.g. GPT-3)

Pre-training: Next-token prediction



Prompting Paradigm

“prompt” an LM for various tasks

<u>Machine translation</u>

食物很美味

Language Model

Translate from English to Chinese

The food is delicious!

Instruction (prompt)

target data point

Prompting Paradigm

“prompt” an LM for various tasks

Sentiment classification

Positive

Language Model

Classify the sentiment of this sentence

The food is delicious!

Instruction (prompt)

target data point

Prompting Paradigm

Positive

Language Model

"**prompt**" an LM for various tasks

Sentiment classification

Classify the sentiment of this sentence | The food is delicious!

Instruction (prompt) | **target data point**

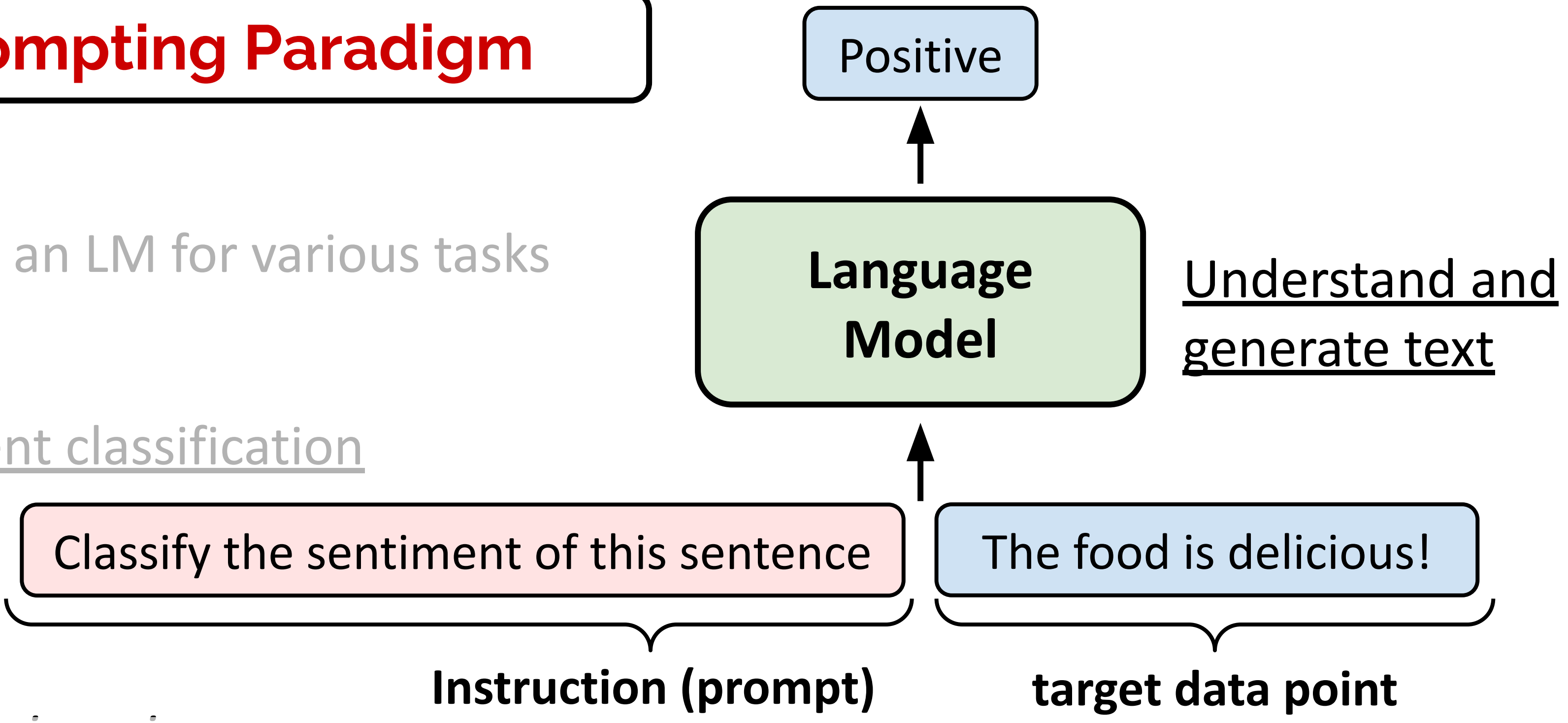
Prompt Engineering

- **Natural language**: Interpretable, manually design, difficult to optimize.
- **Continuous vectors**: Trainable and more capable, difficult to interpret.

Prompting Paradigm

"prompt" an LM for various tasks

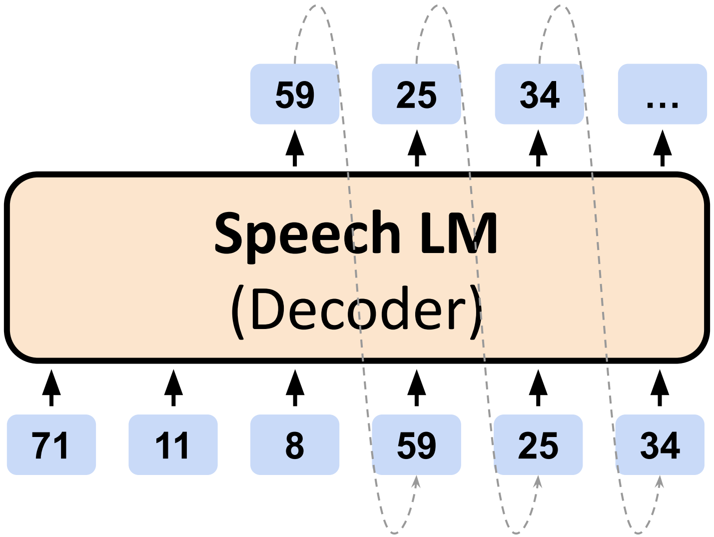
Sentiment classification

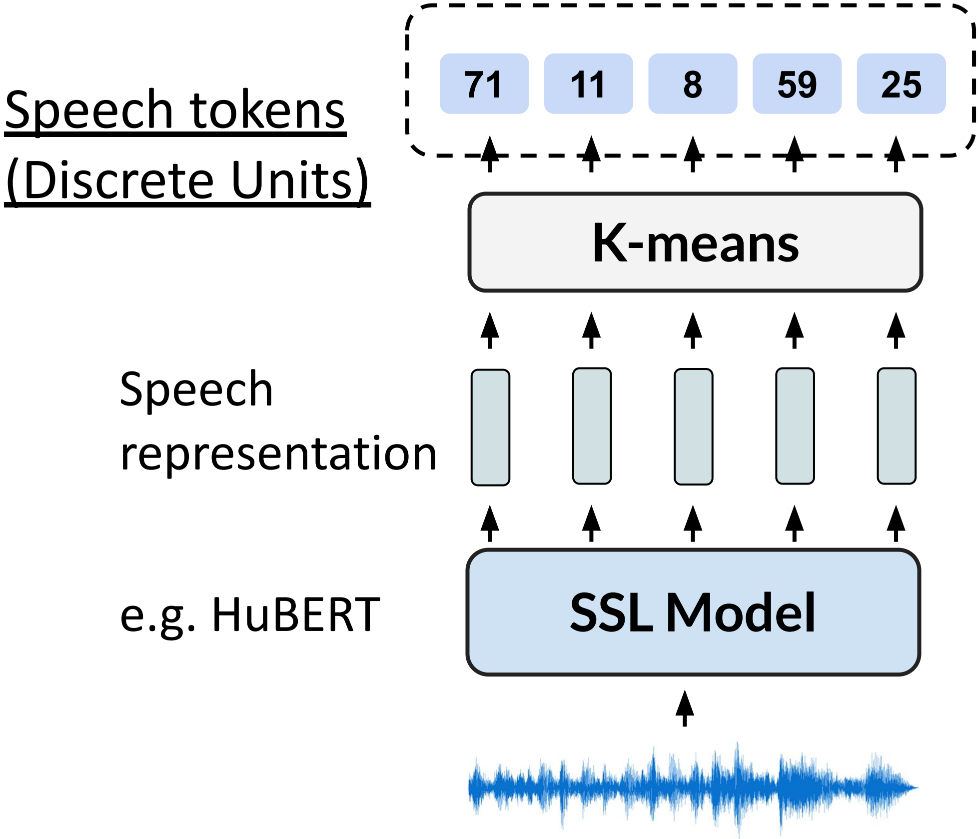


Prompt Engineering

- **Natural language**: Interpretable, manually design, difficult to optimize.
- **Continuous vectors**: Trainable and more capable, difficult to interpret.

Textless Speech LM

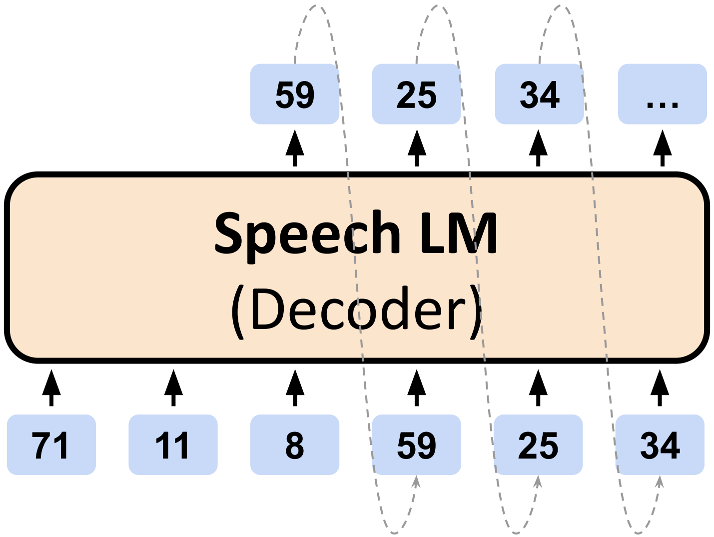




- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

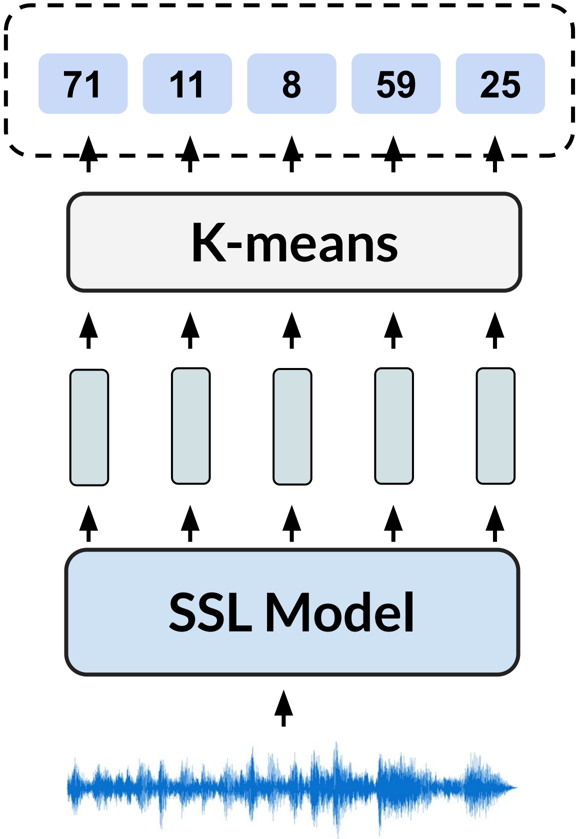
Textless Speech LM



- Task: Next-token prediction
- Example: GSLM

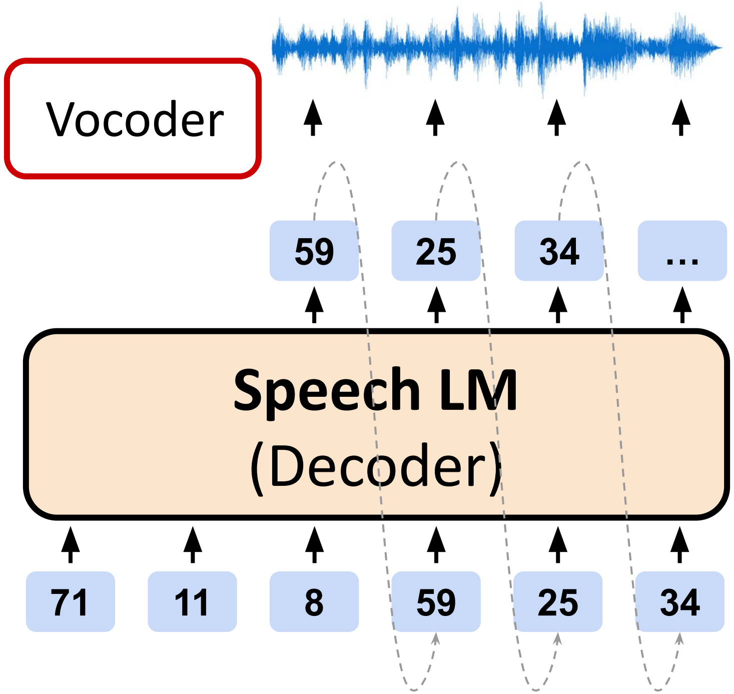
Speech tokens
(Discrete Units)

- Phonetic
- Semantic



Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

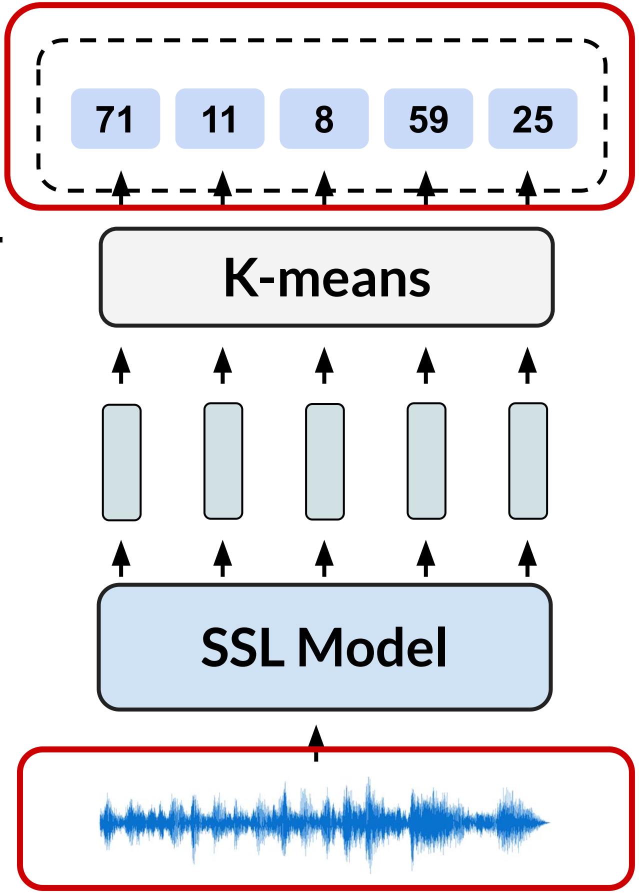
Textless Speech LM



- Task: Next-token prediction
- Example: GSLM

Speech tokens
(Discrete Units)

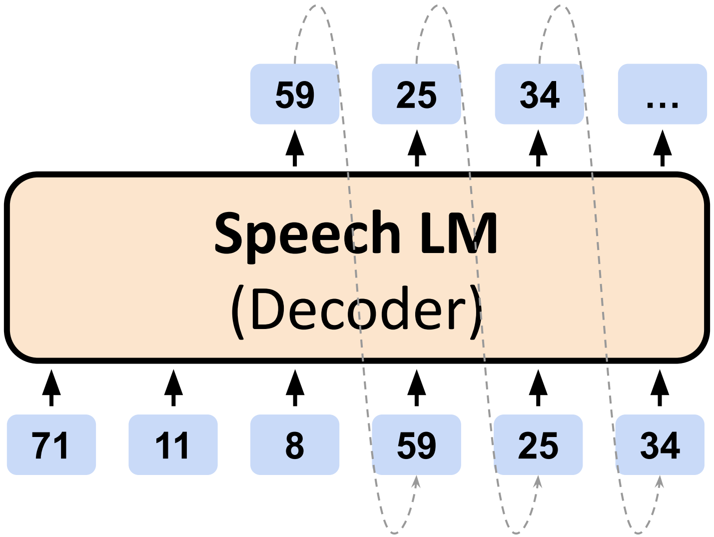
- Phonetic
- Semantic



Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Textless Speech LM

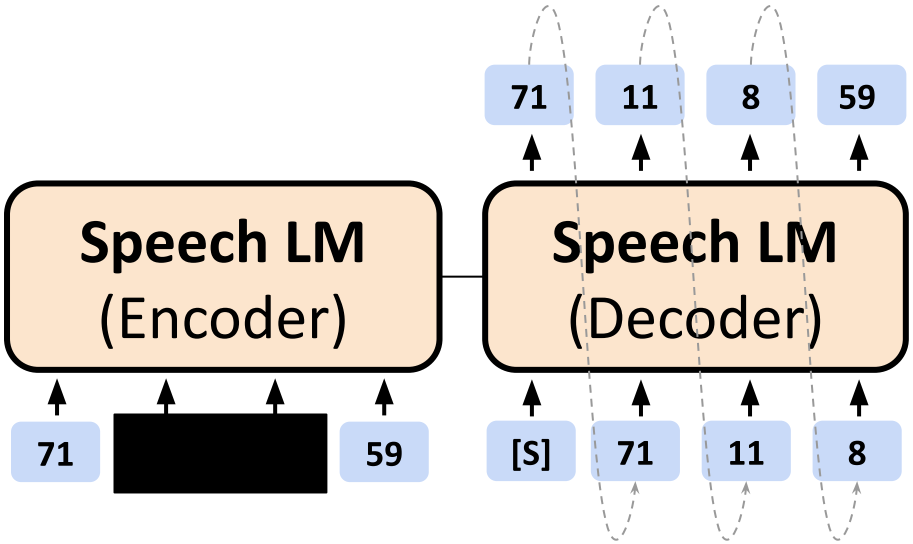
Decoder-only Speech LM



- Task: Next-token prediction
- Example: GSLM

Generative Spoken Language Modeling from Raw Audio (https://arxiv.org/abs/2102.01192)

Encoder-Decoder Speech LM

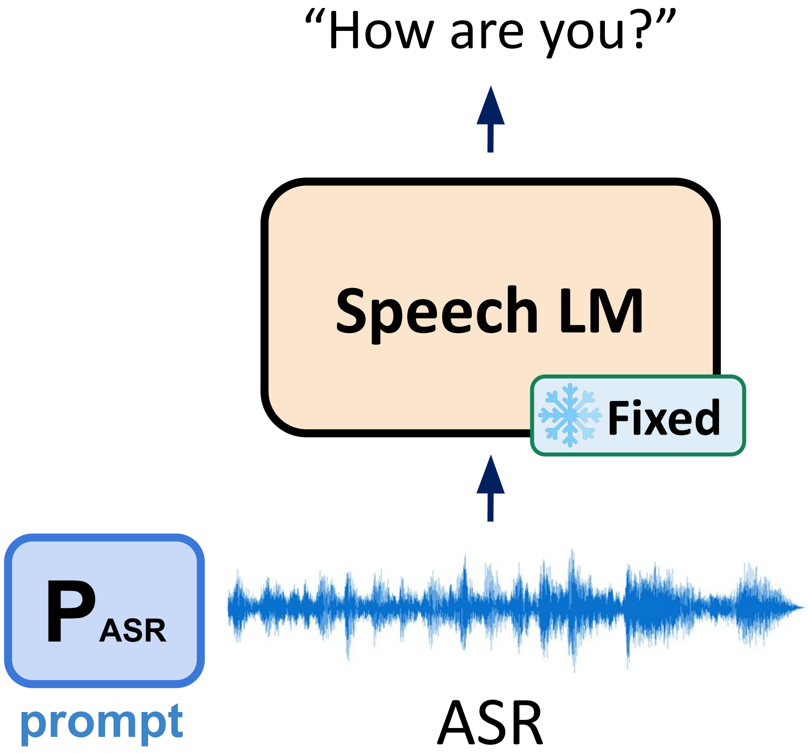


- Task: Reconstruction
- Example: Unit mBART

Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation (https://arxiv.org/abs/2204.02967)

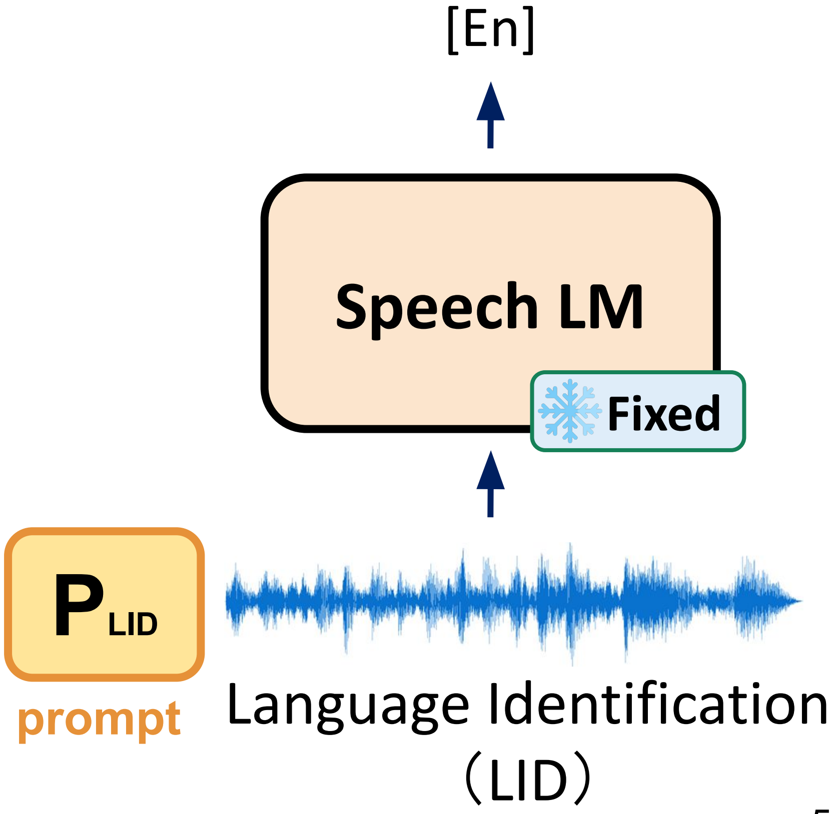
Prompting Paradigm

"prompt" a speech language model to perform various downstream tasks



Prompting Paradigm

"prompt" a speech language model to perform various downstream tasks

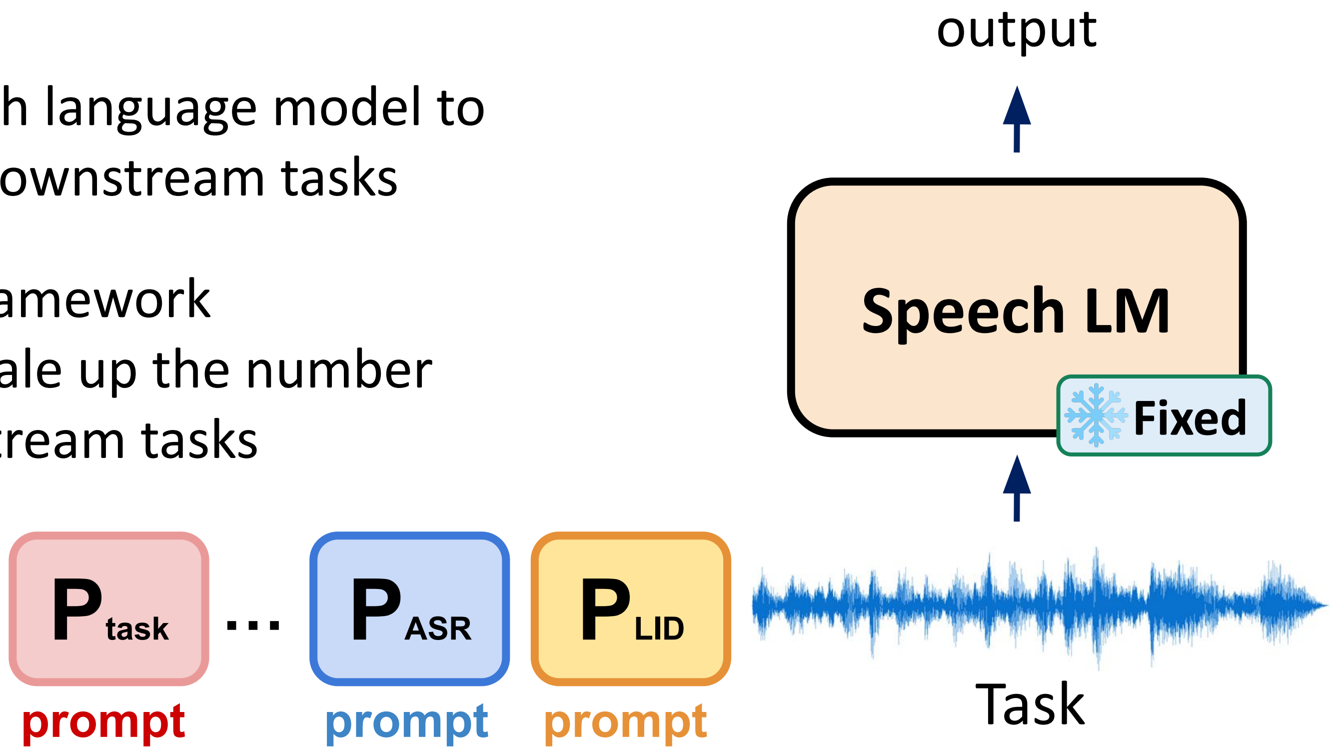


Prompting Paradigm

"prompt" a speech language model to perform various downstream tasks

- Unified framework
- Easy to scale up the number of downstream tasks

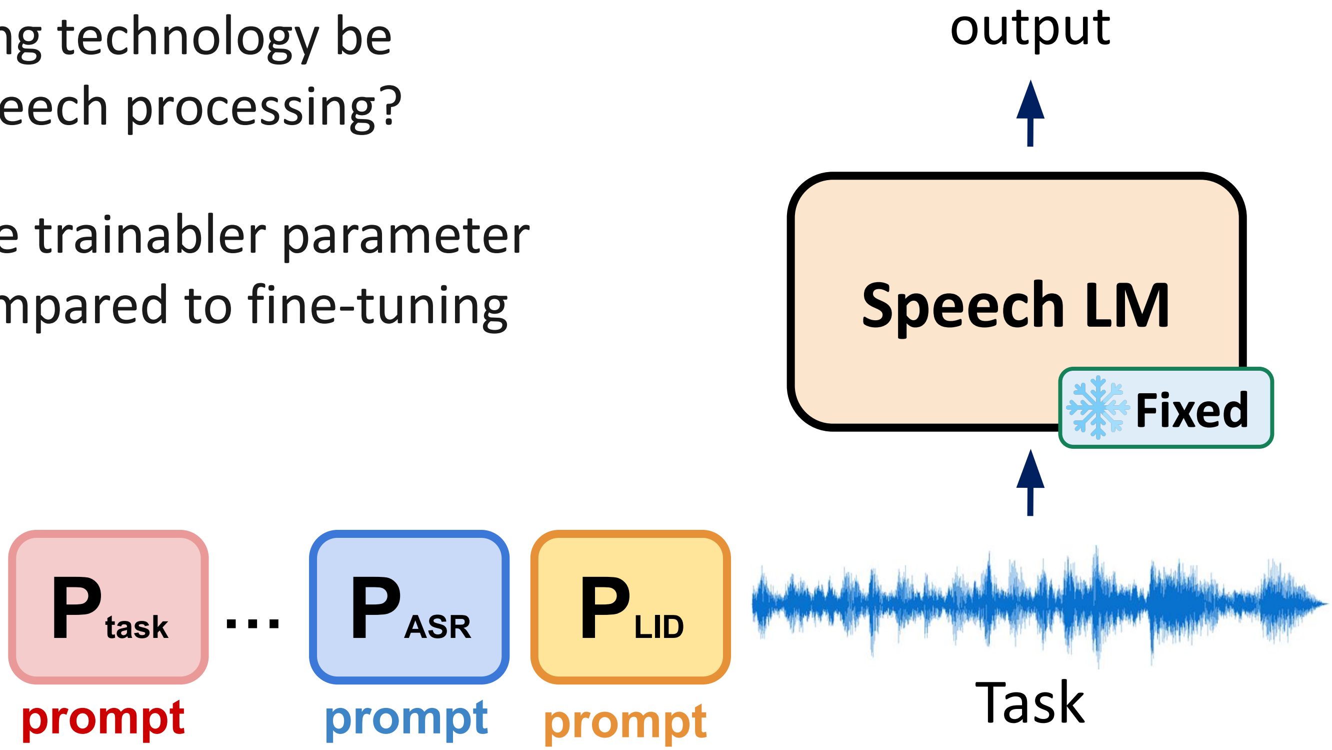
Contain few parameters



Prompting Paradigm

1. Can prompting technology be applied to speech processing?

2. Can it achieve trainabler parameter efficiency compared to fine-tuning paradigm?



SpeechPrompt

Outline

Diverse Speech Processing Tasks

Prompting Speech LM

Experiment Results

Further improvement

- Speech Classification Tasks
- Sequence Generation Tasks
- Speech Generation Tasks

3 kinds of speech processing tasks that take speech as input

1. Speech Classification

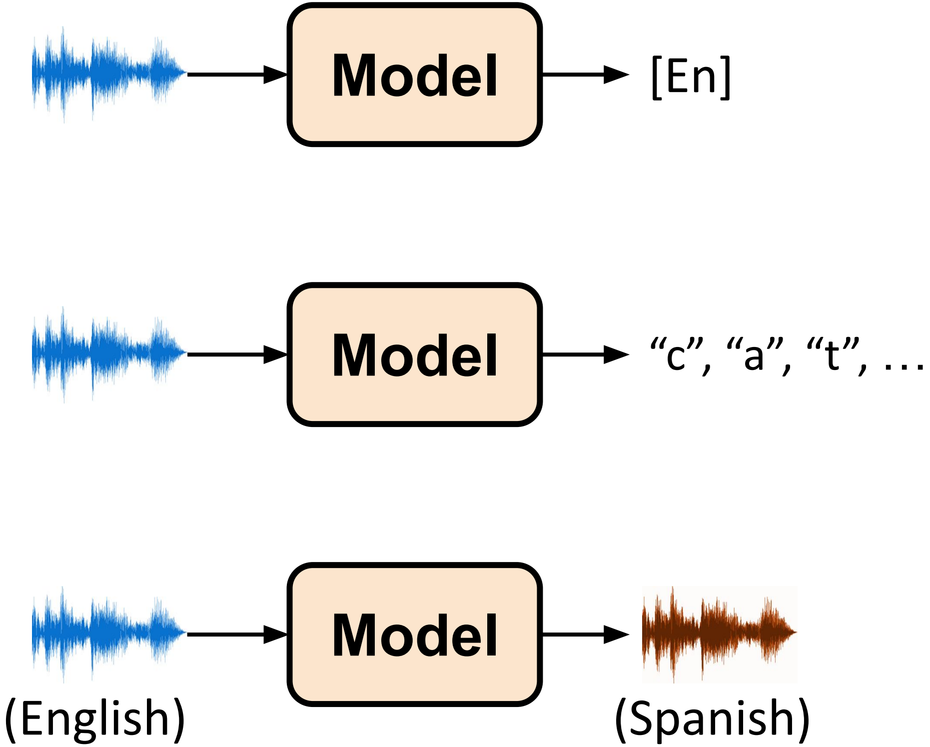
- ➔ Speech to class
- ➔ e.g. Langauge Identification

2. Sequence Generation

- ➔ Speech to label sequence
- ➔ e.g. ASR

3. Speech Generation

- ➔ Speech to speech
- ➔ e.g. Speech translation



Outline

Diverse Speech Processing Tasks

↓

Prompting Speech LM

↓

Experiment Results

↓

Further improvement

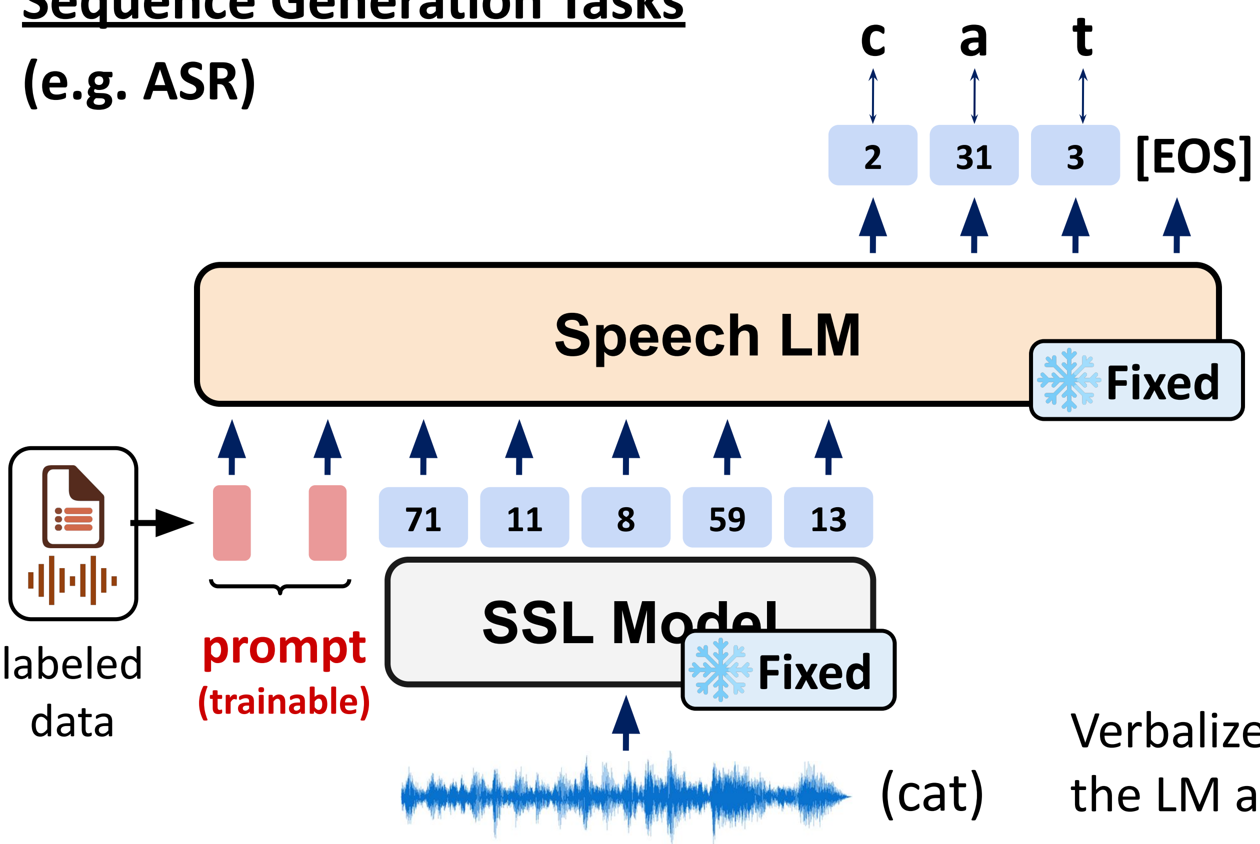
- Speech Classification Tasks
- Sequence Generation Tasks
- Speech Generation Tasks

Prompting Speech LM

downstream task's label

Sequence Generation Tasks

(e.g. ASR)



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

Mapping table (Verbalizer)

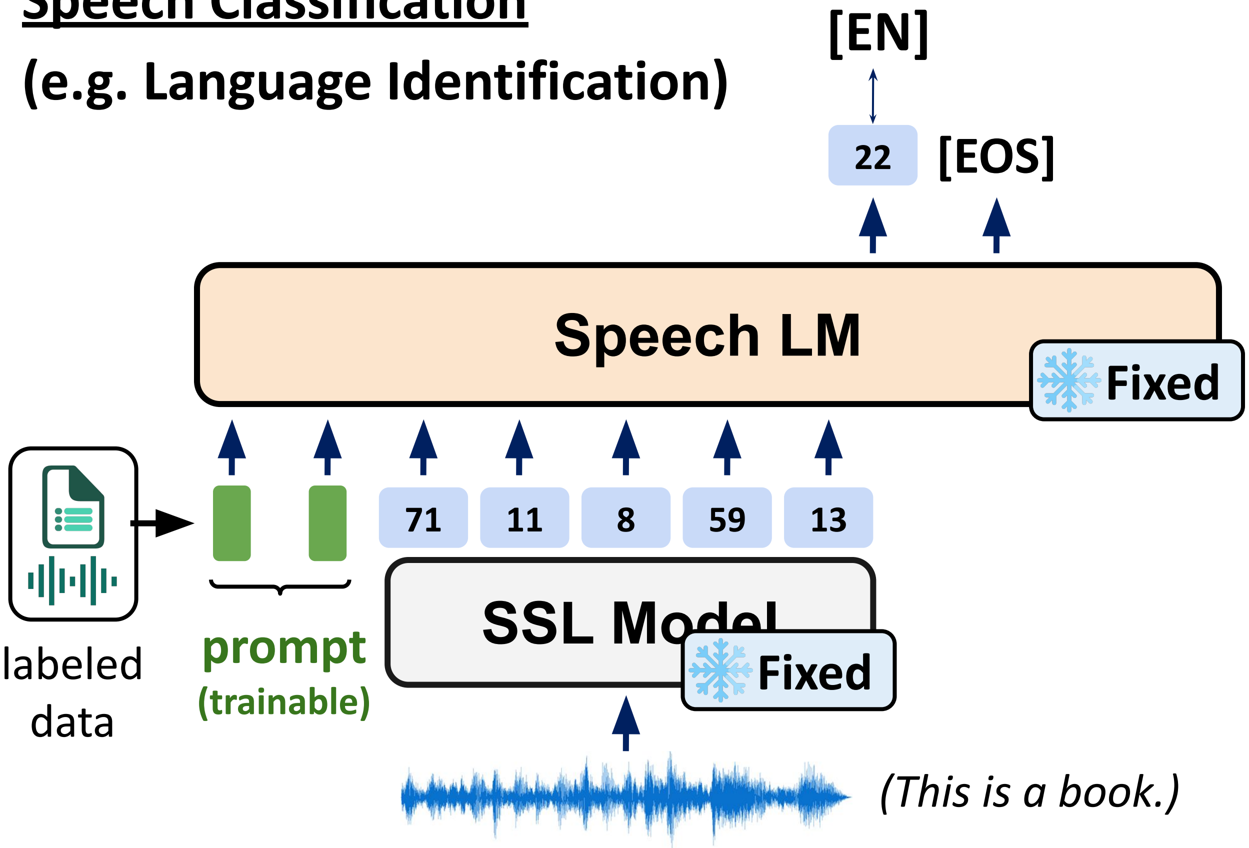
Verbalizer: Bridge the vocabulary of the LM and the task labels.

Prompting Speech LM

downstream task's label

<u>Speech Classification</u>

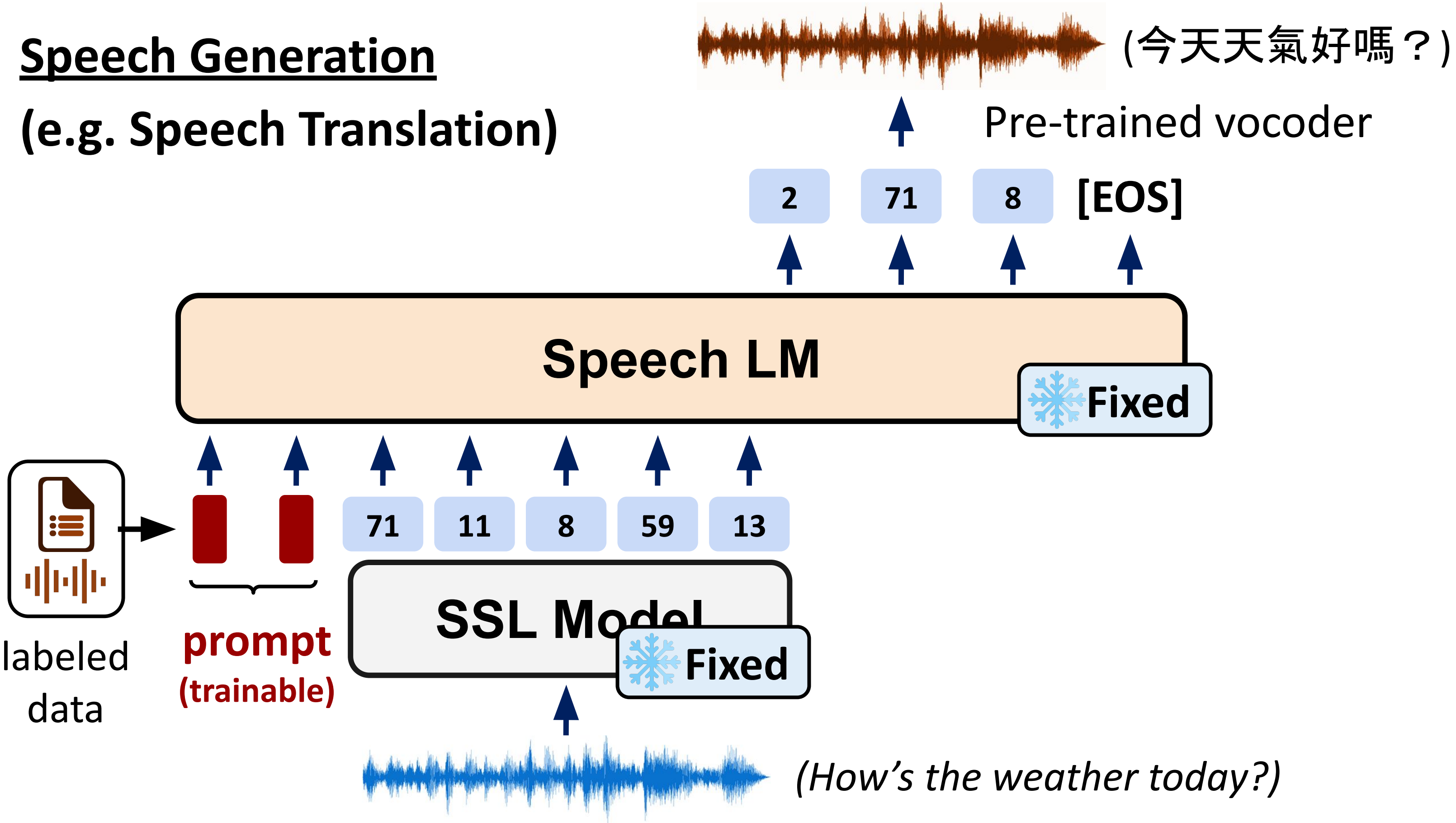
(e.g. Language Identification)



Labels	Unit ID
[EN]	22
[ES]	29
[CN]	17
…	…
[LT]	3
…	…

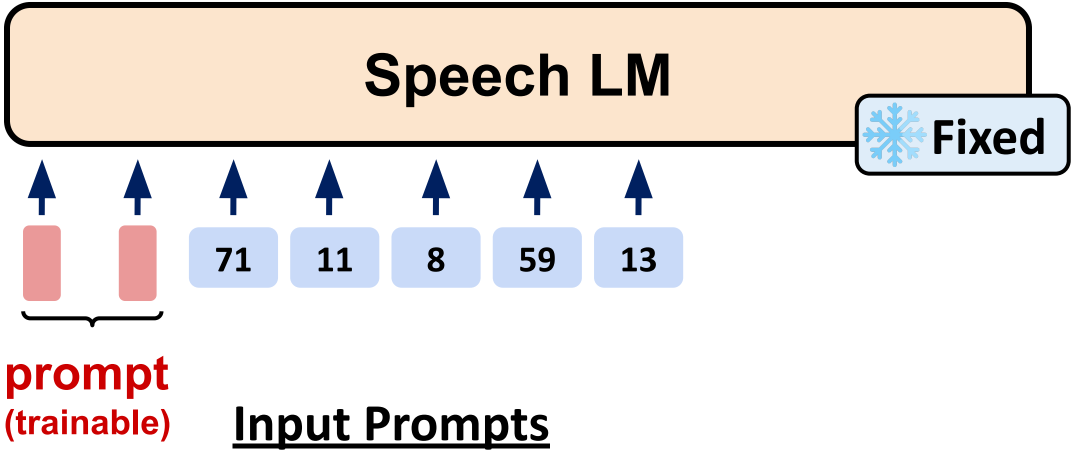
Mapping table (Verbalizer)

Prompting Speech LM



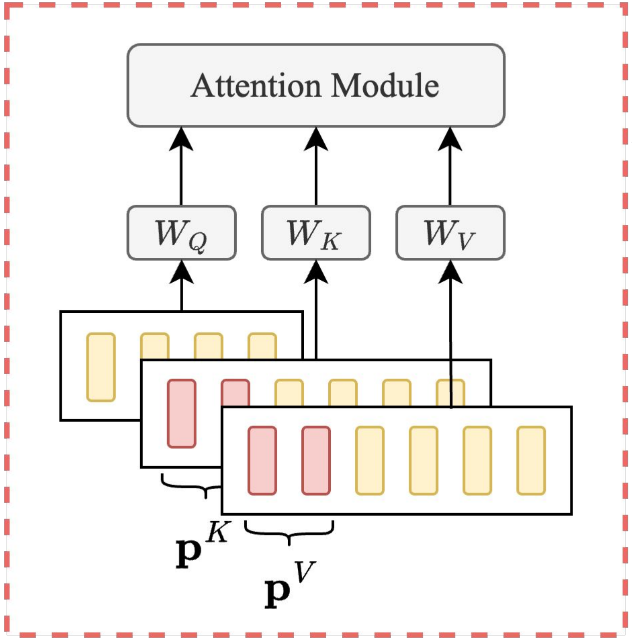
Prompting Speech LM

Prompting: Find the prompts and put them at the **input** without modifying the LM's architecture

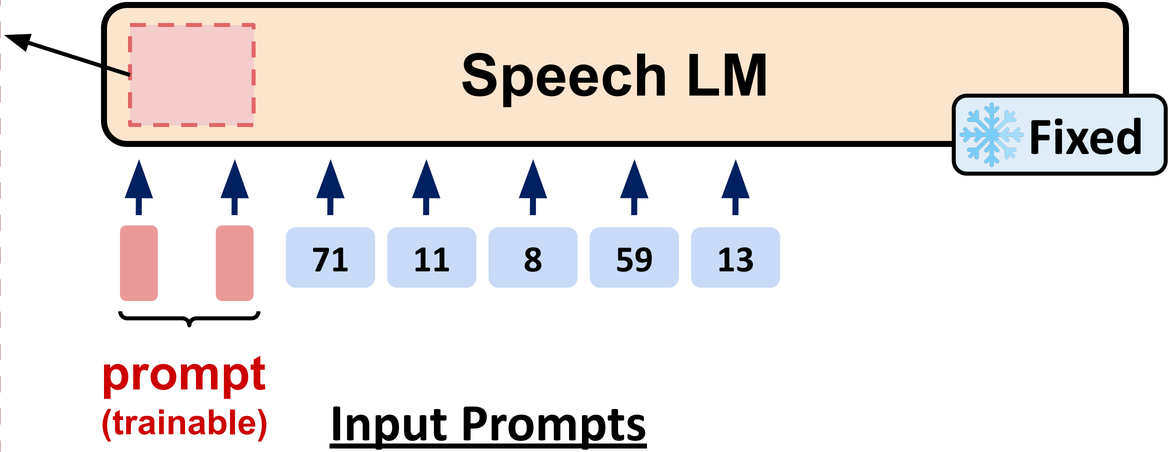


Prompting Speech LM

The prompts are prepended at the **input** of each transformer layer.



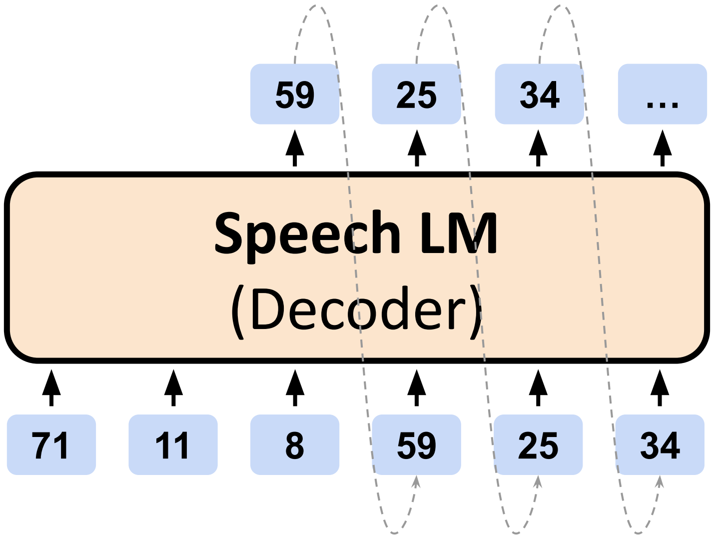
Deep Prompts guiding the attention mechnism



The following exps. show the performance of input prompts + deep prompts

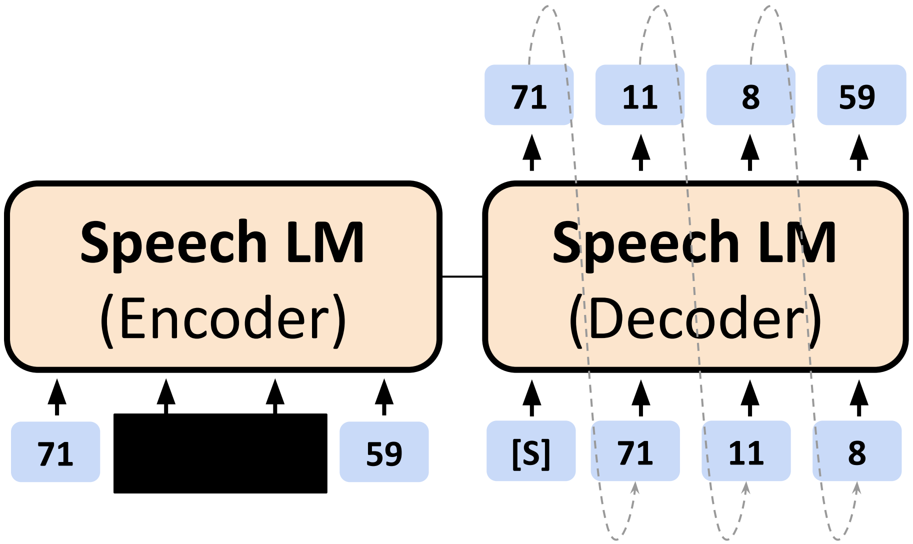
Textless Speech LM

Decoder-only Speech LM



- Task: Next-token prediction
- Example: GSLM

Encoder-Decoder Speech LM



- Task: Reconstruction
- Example: Unit mBART

Generative Spoken Language Modeling from Raw Audio
(https://arxiv.org/abs/2102.01192)

Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation (https://arxiv.org/abs/2204.02967)

Outline

Diverse Speech Processing Tasks

↓

Prompting Speech LM

↓

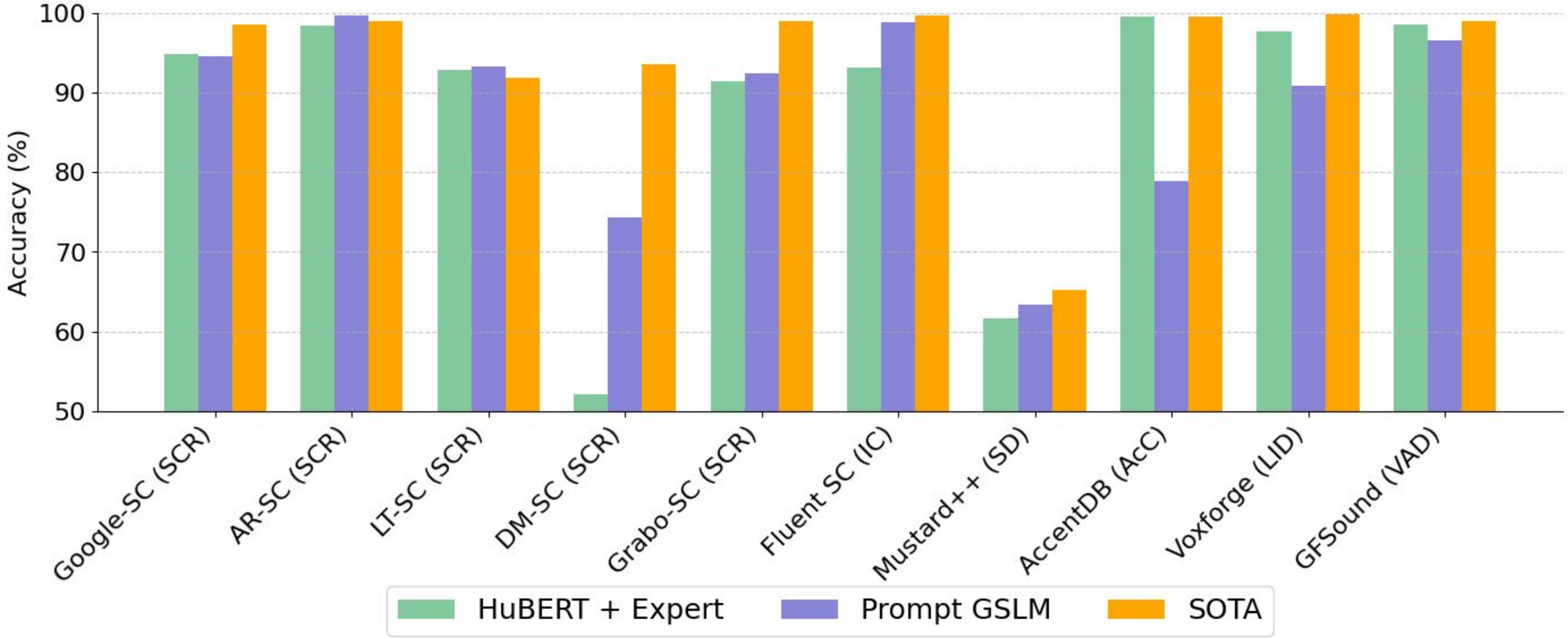
Experiment Results

↓

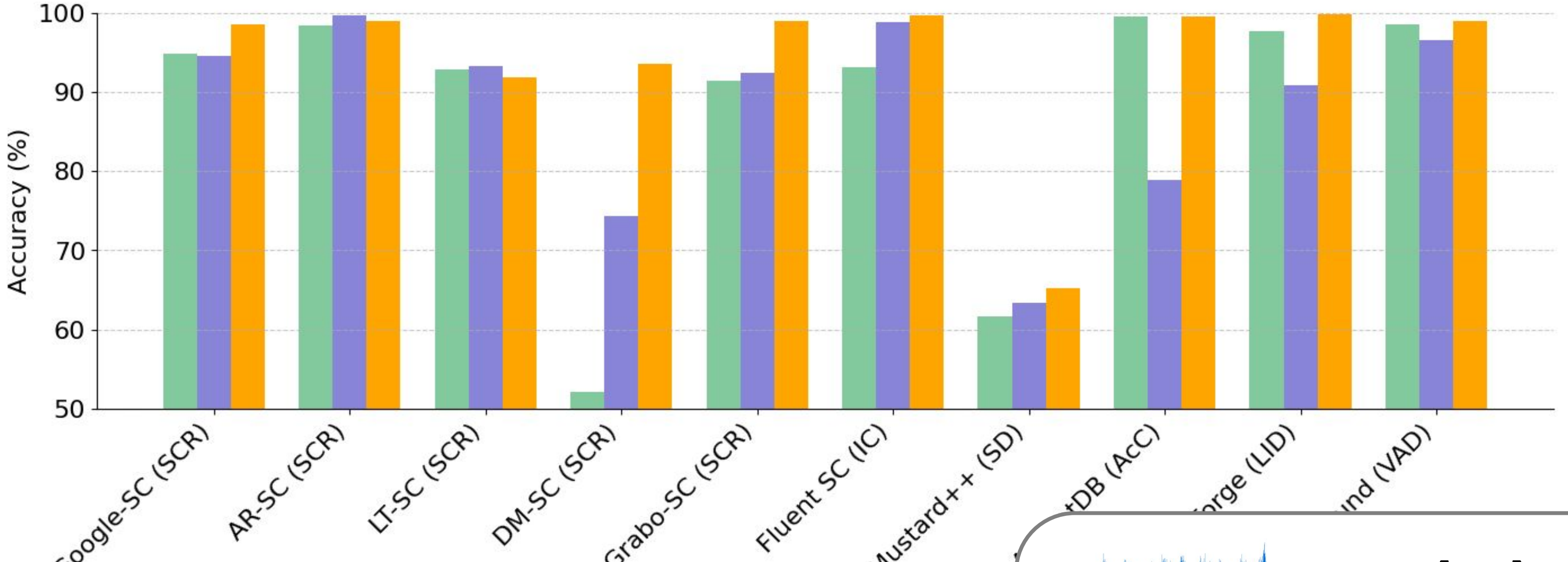
Further improvement

- **Speech Classification Tasks**
- Sequence Generation Tasks
- Speech Generation Tasks

Speech Classification - Prompt GSLM



Speech Classification - Prompt GSLM



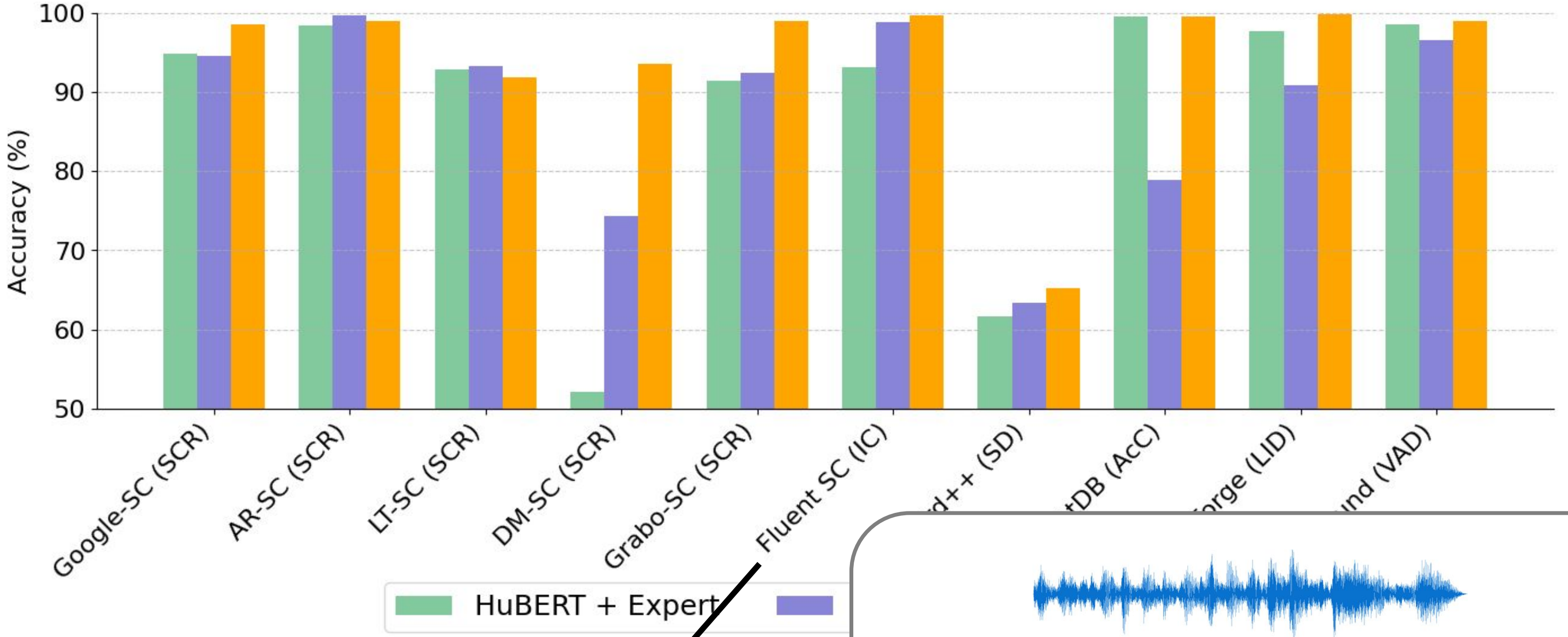
Spoken Command Recognition (SCR)
(English, Arabic, Lithuanian, Mandarin, German)

→ [UP]

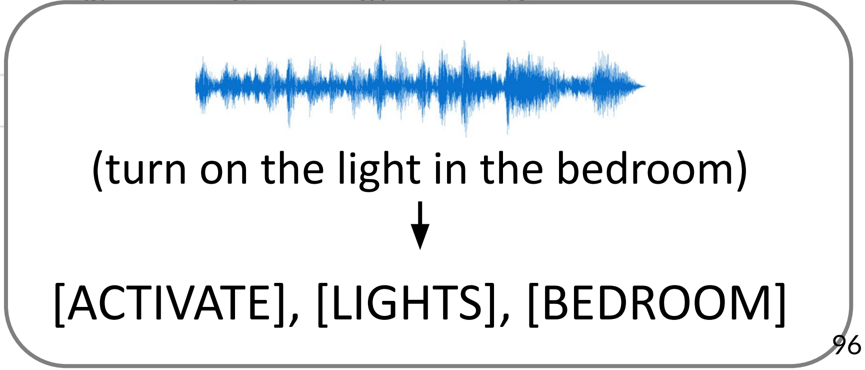
(UP)

Classify an utterance into a predefined keyword set.

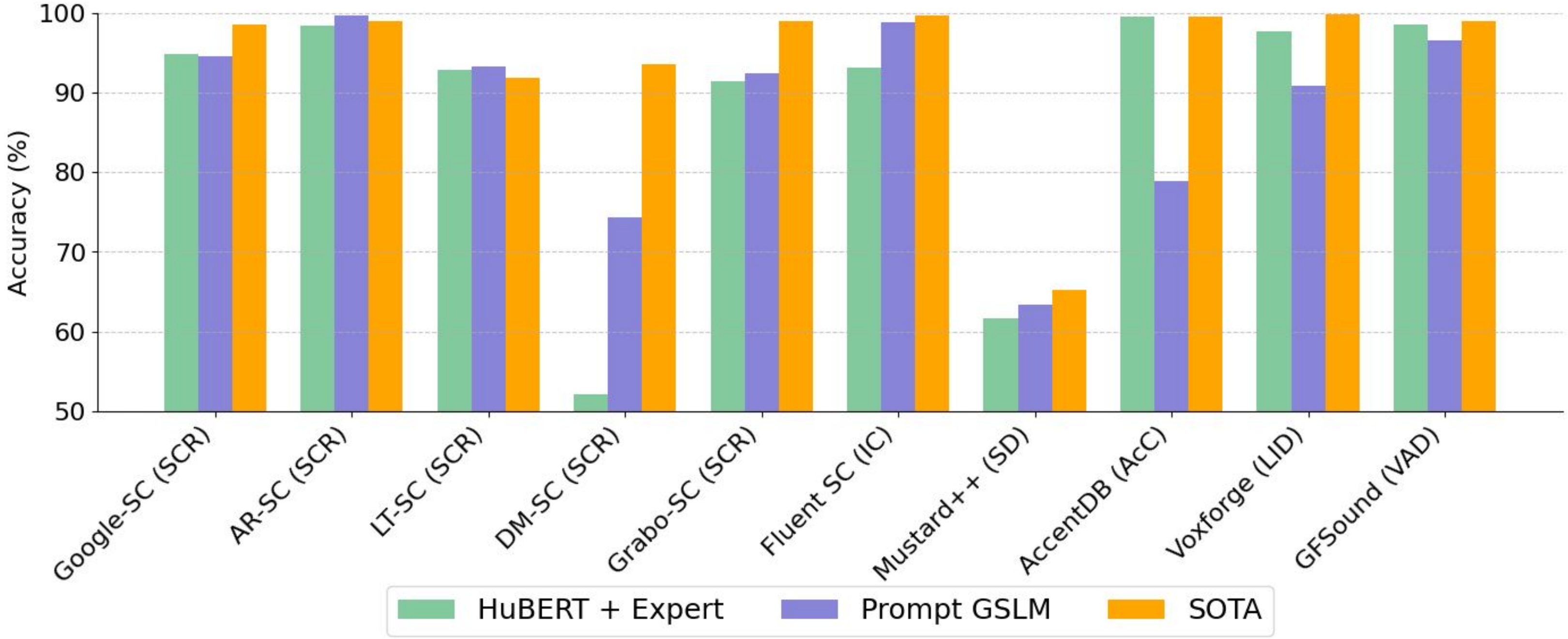
Speech Classification - Prompt GSLM



Intent Classification (IC)
Multi-label classification

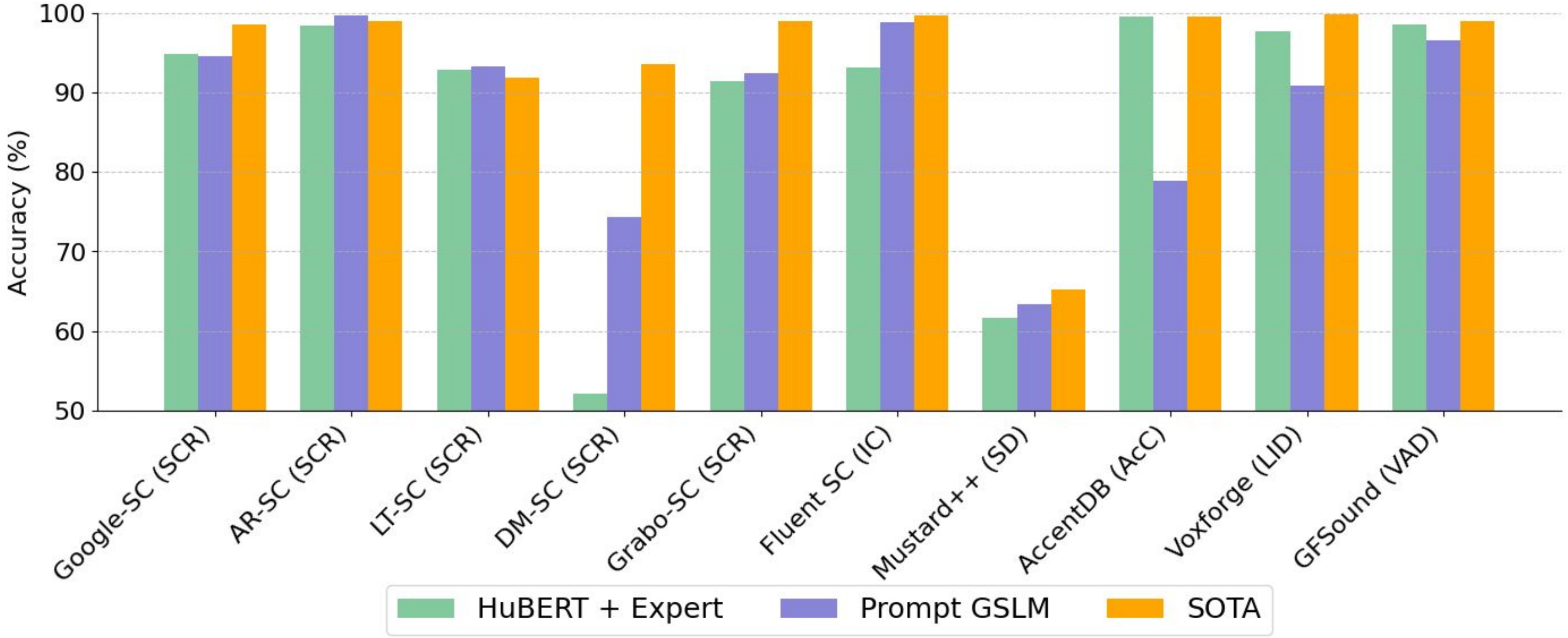


Speech Classification - Prompt GSLM



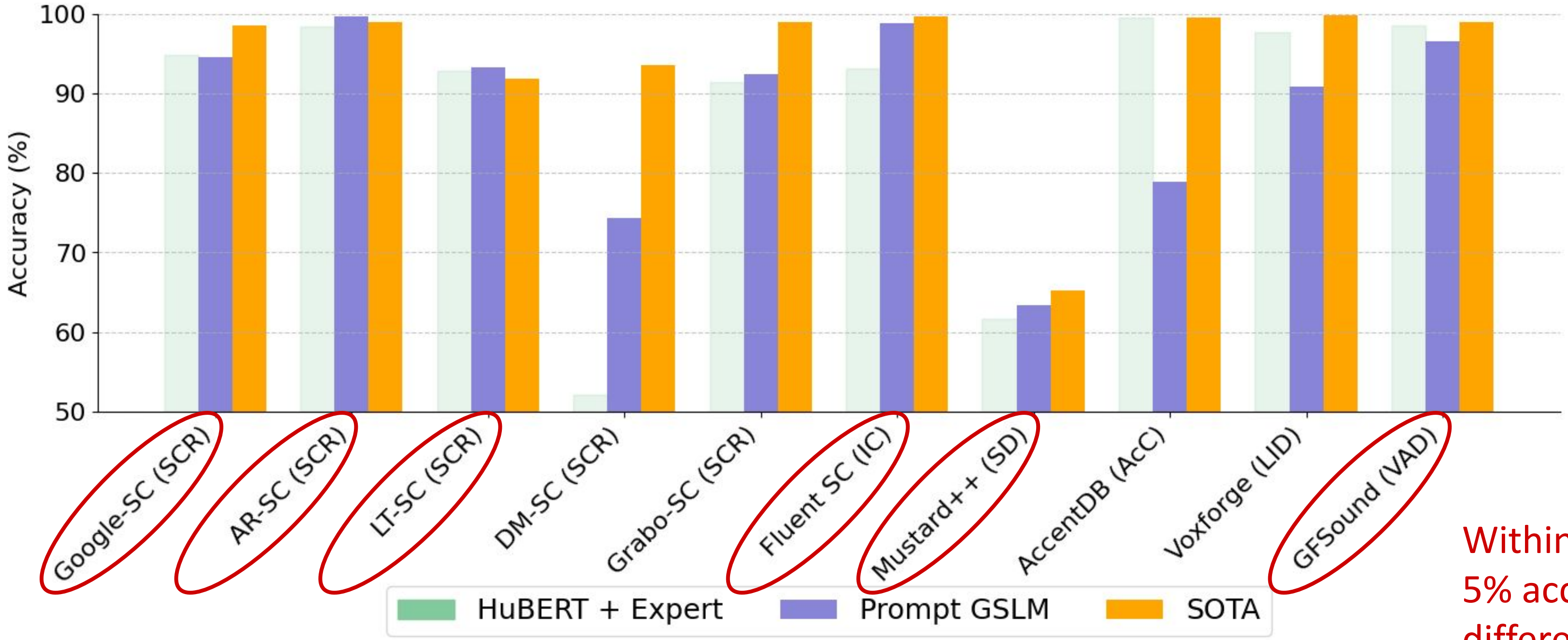
Sarcasm Detection (SD), Accent Classification(AcC)
Language Identification (LID), Voice Activity Detection (VAD)

Speech Classification - Prompt GSLM



- **HuBERT + Expert**: Fine-tuning paradigm - #Params.: **0.2M**
- **Prompt GSLM**: Prompting paradigm - #Params.: **0.15M**
- **SOTA**: Best model - dedicated trained

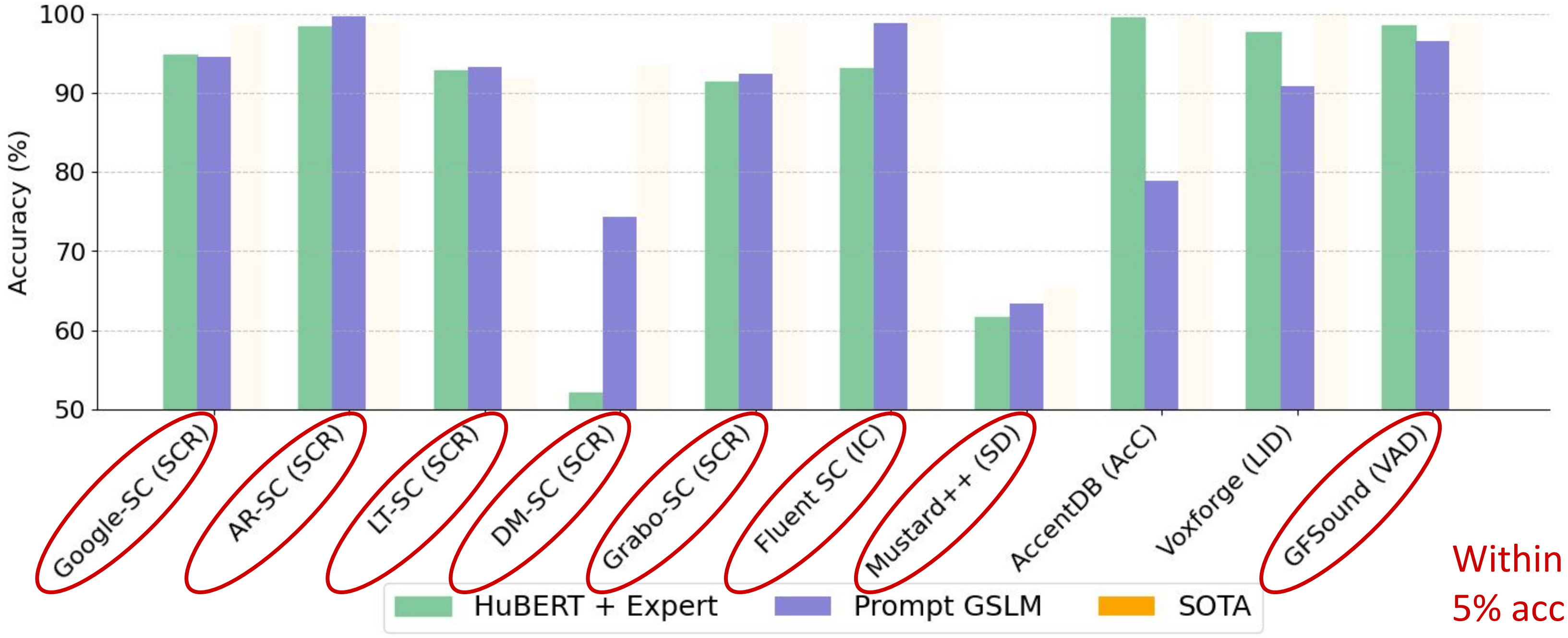
Speech Classification - Prompt GSLM



Within relative 5% accuracy difference

- **Prompt GSLM** can achieve comparable performance to **SOTA**
- **Prompting** is within a unified framework.

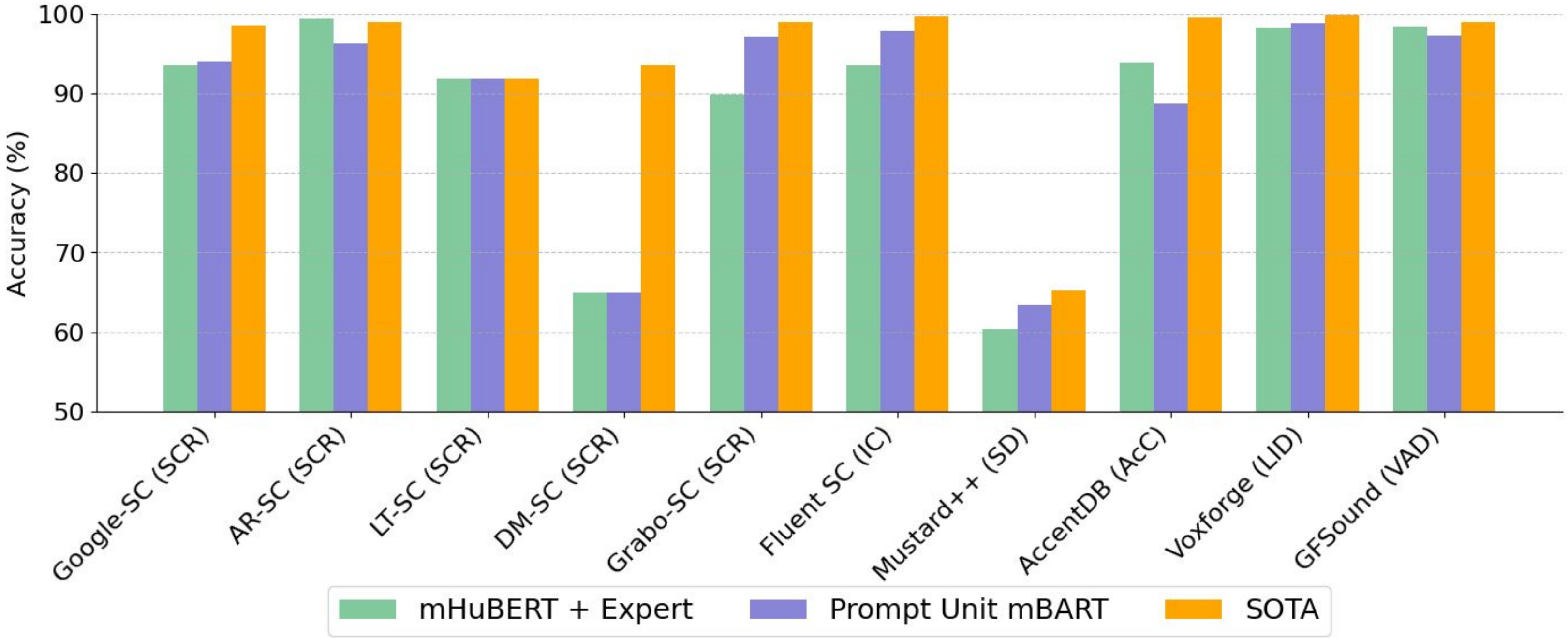
Speech Classification - Prompt GSLM



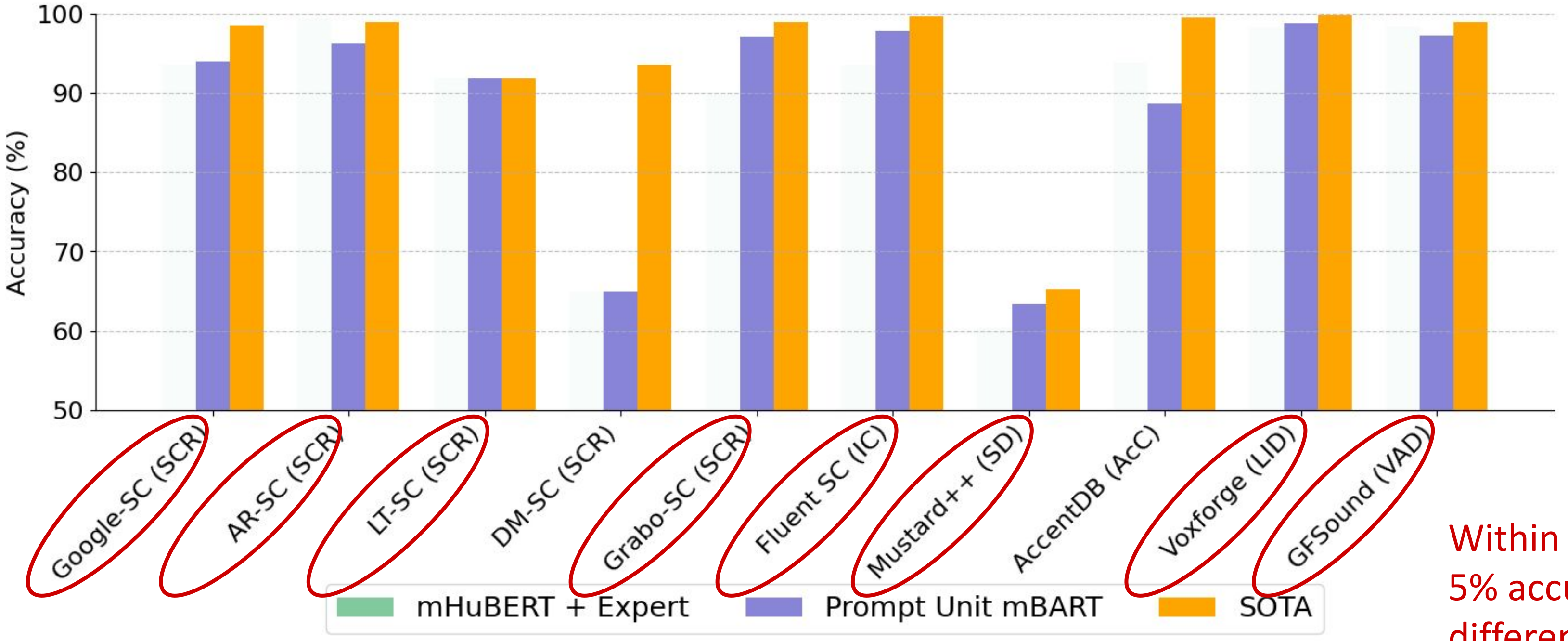
Within relative 5% accuracy difference

- Comparing **Prompt GSLM** and **HuBERT + Expert:** Prompting is competitive to pre-train, fine-tune paradigm in 8 out of 10 tasks.

Speech Classification - Prompt Unit mBART



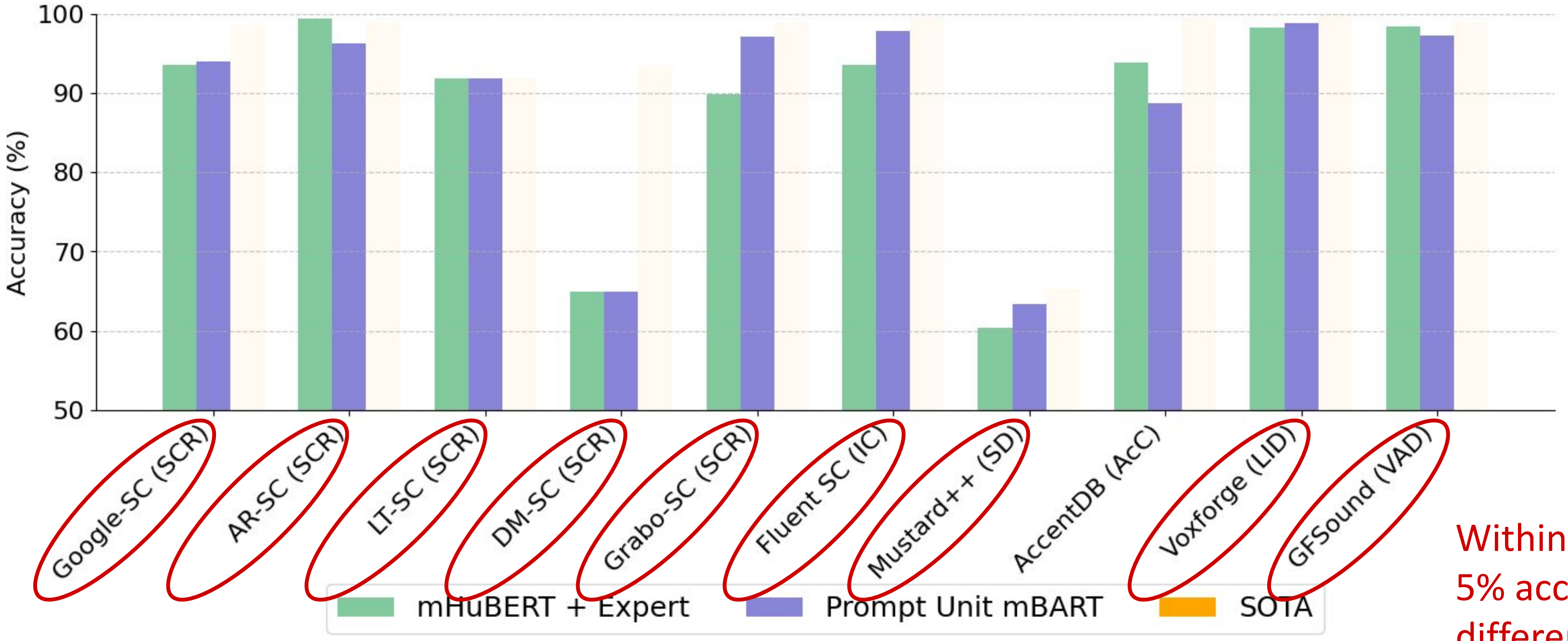
Speech Classification - Prompt Unit mBART



Within relative 5% accuracy difference

- **Prompt Unit mBART** is competitive to **SOTA** in 8 out of 10 tasks.

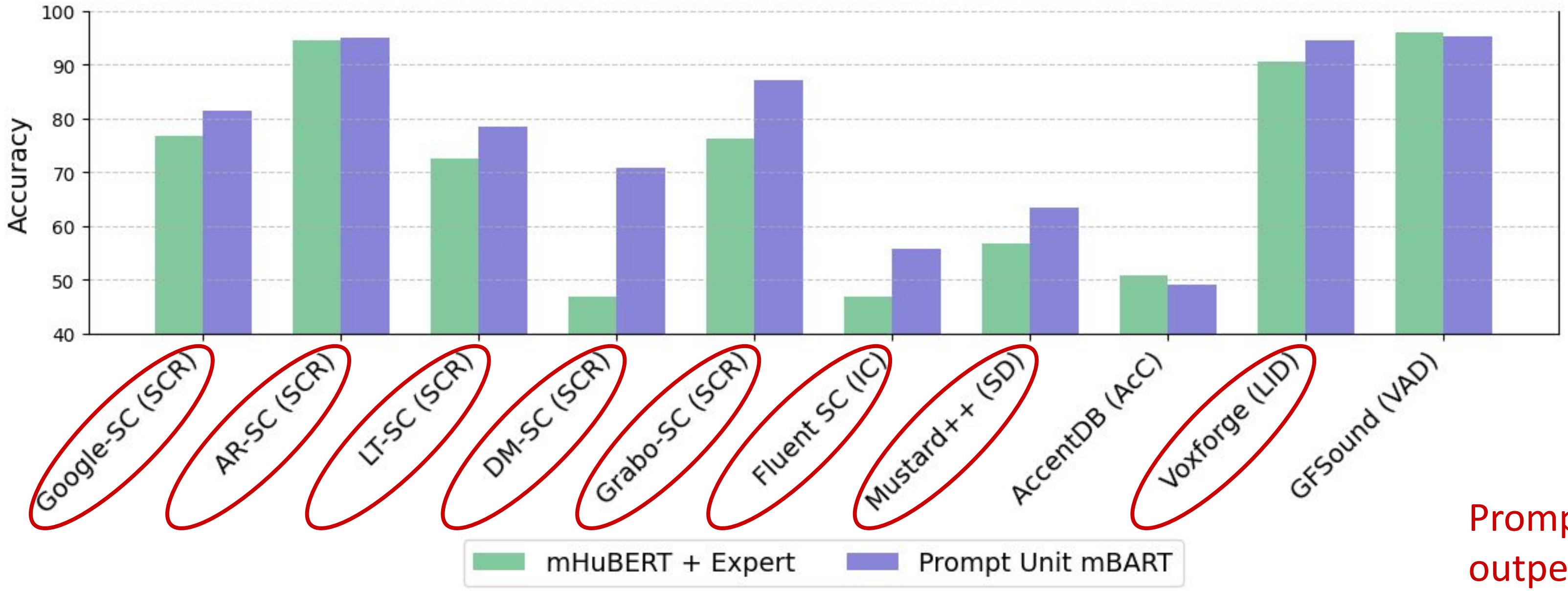
Speech Classification - Prompt Unit mBART



Within relative 5% accuracy difference

- **Prompt Unit mBART** is competitive to **mHuBERT + Expert** in 9 out of 10 tasks

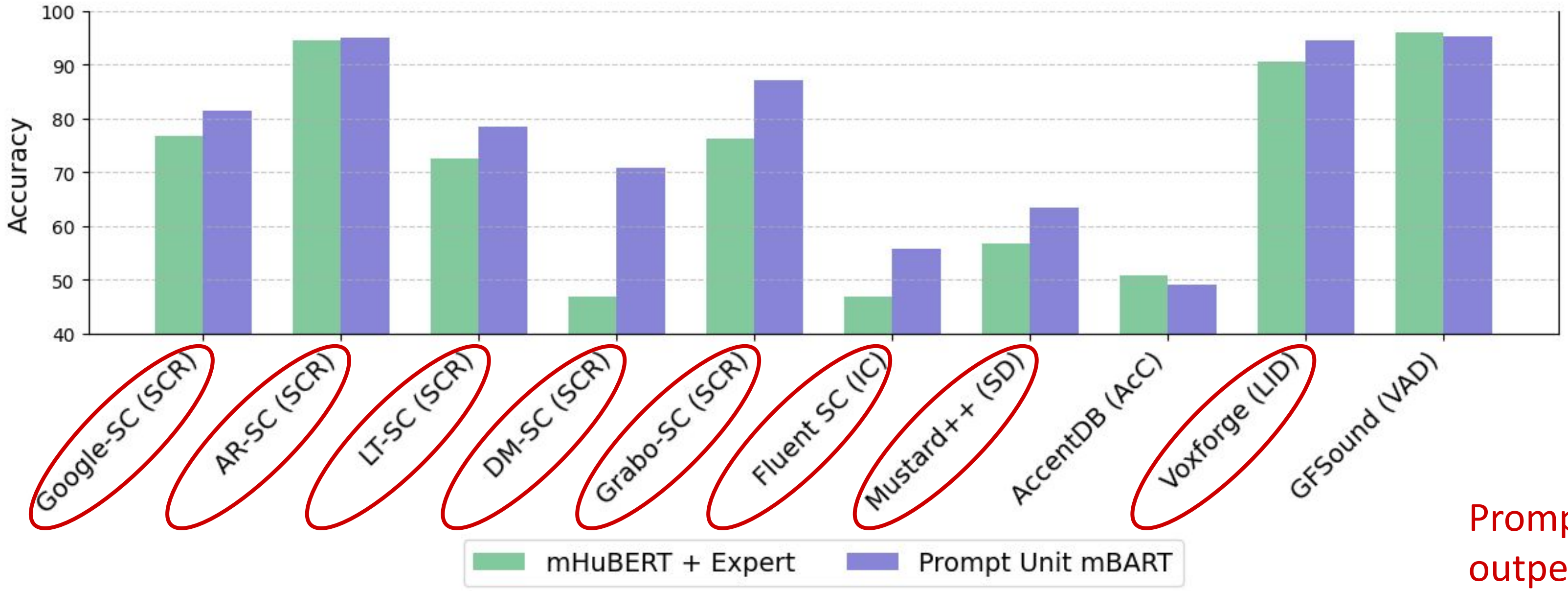
Speech Classification - Prompt Unit mBART



Prompting outperforms fine-tuning

- **10-shot Learning**. Each class contains only 10 training data.

Speech Classification - Prompt Unit mBART



Prompting outperforms fine-tuning

- **10-shot Learning**. Each class contains only 10 training data.
- **Prompt Unit mBART** outperforms **mHuBERT + Expert** in 8 out of 10 tasks.

Prompting for Speech Classification

1. **Prompting** is competitive to **fine-tuning**

2. **Prompting** can also be competitive to **SOTA**

3. **Prompting** has advantages in few-shot learning

Outline

Diverse Speech Processing Tasks

↓

Prompting Speech LM

↓

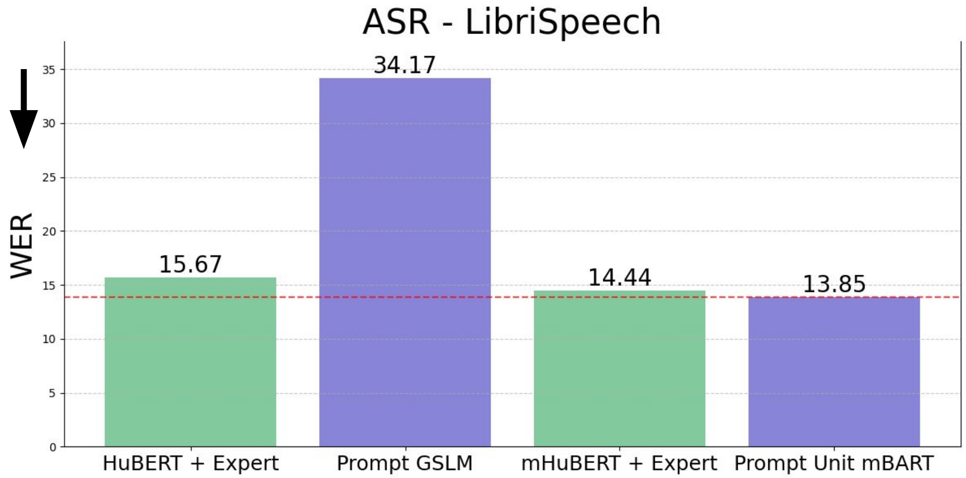
<u>Experiment Results</u>

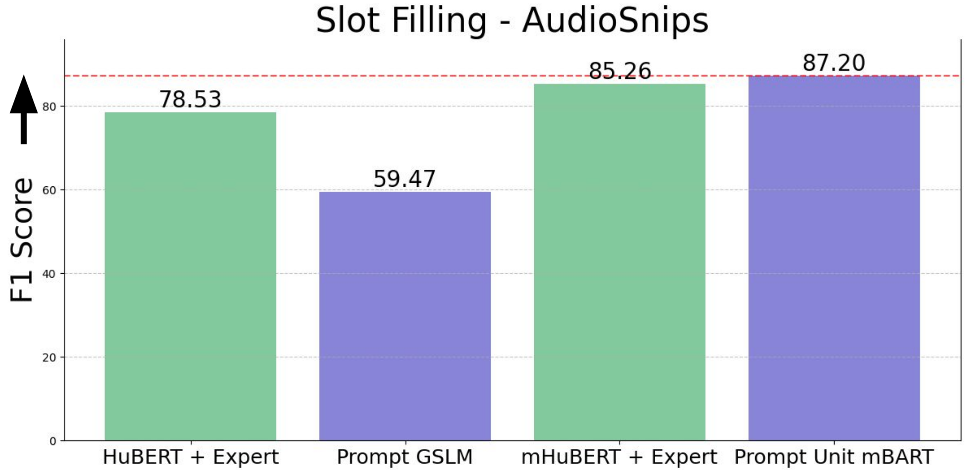
↓

Further improvement

- Speech Classification Tasks
- **<u>Sequence Generation Tasks</u>**
- Speech Generation Tasks

Sequence Generation Tasks



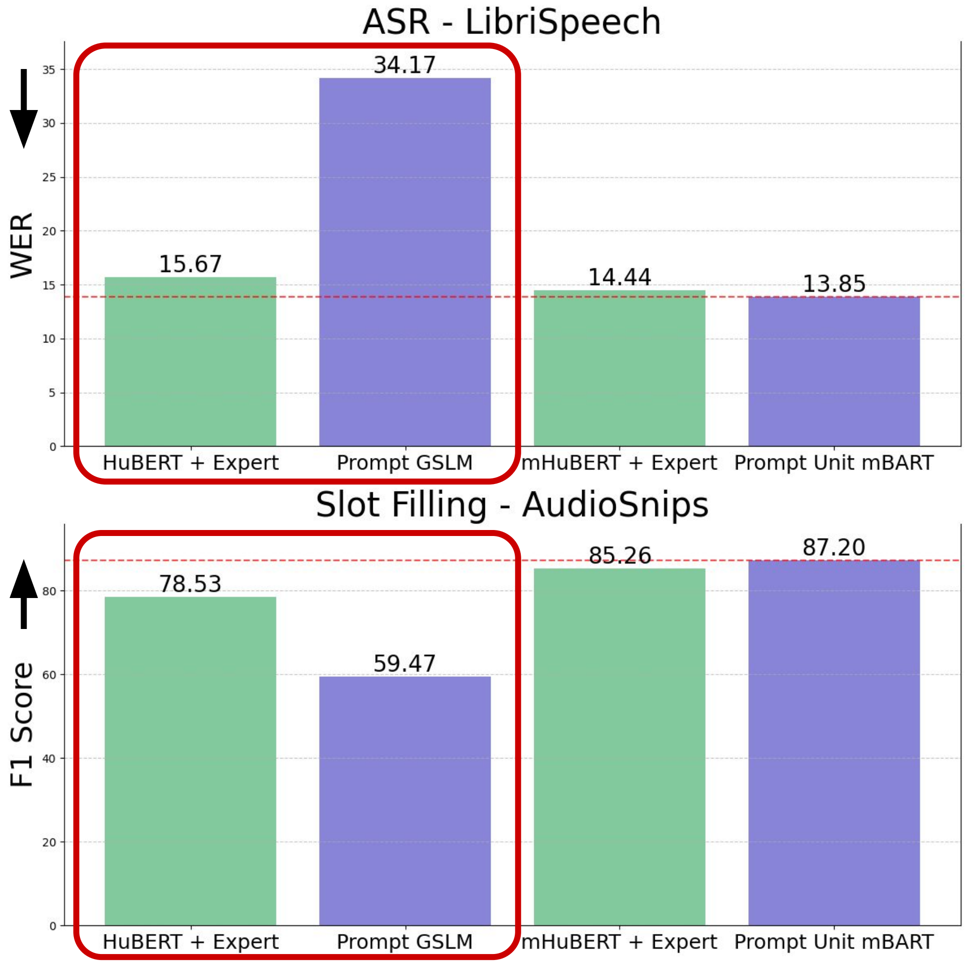


- **ASR**: transcribe an utterance into characters

- **Slot Filling**: conduct ASR and identify the slot types at the same time.

e.g. What's the weather like in <L> NewYork <L/> <T> tomorrow </T>?

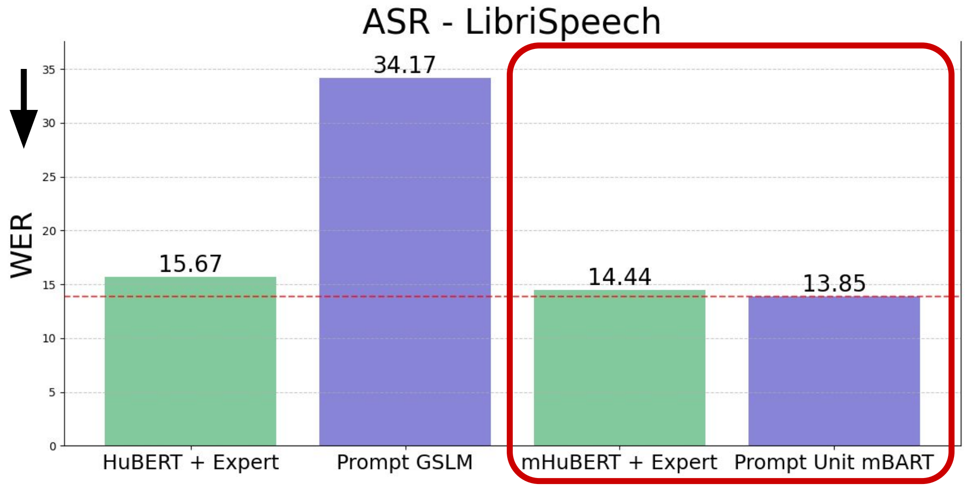
Sequence Generation Tasks

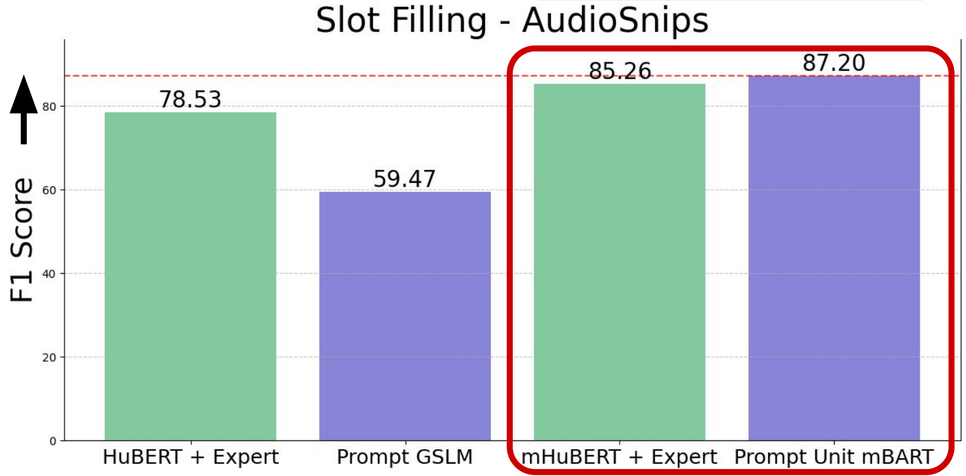


Scenario	Traniable Params.
HuBERT + Expert	2.9M
Prompt **GSLM**	4.5M

- Prompting GSLM underperforms the pre-train, fine-tune paradigm.

Sequence Generation Tasks

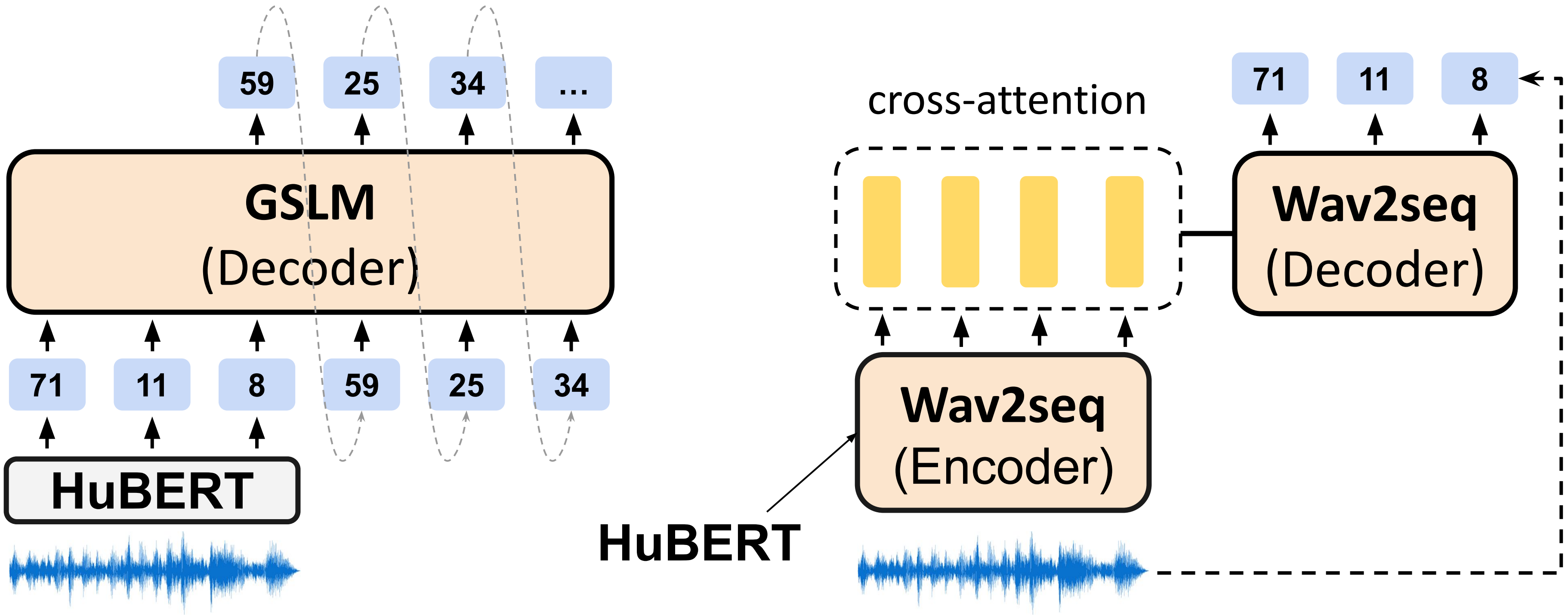




Scenario	Traniable Params.
mHuBERT + Expert	2.9M
Prompt **Unit mBART**	2.6M

- Prompting Unit mBART outperforms the pre-train, fine-tune paradigm.
- For prompting, model architecture and pre-training task matter.
- Encoder-decoder model is better than decoder-only model?

Decoder-only vs. Encoder-Decoder Speech LM



Next-token prediction

Pseudo speech recognition

Prompting and adapter tuning for self-supervised encoder-decoder speech model, ASRU2023 (https://arxiv.org/abs/2310.02971)

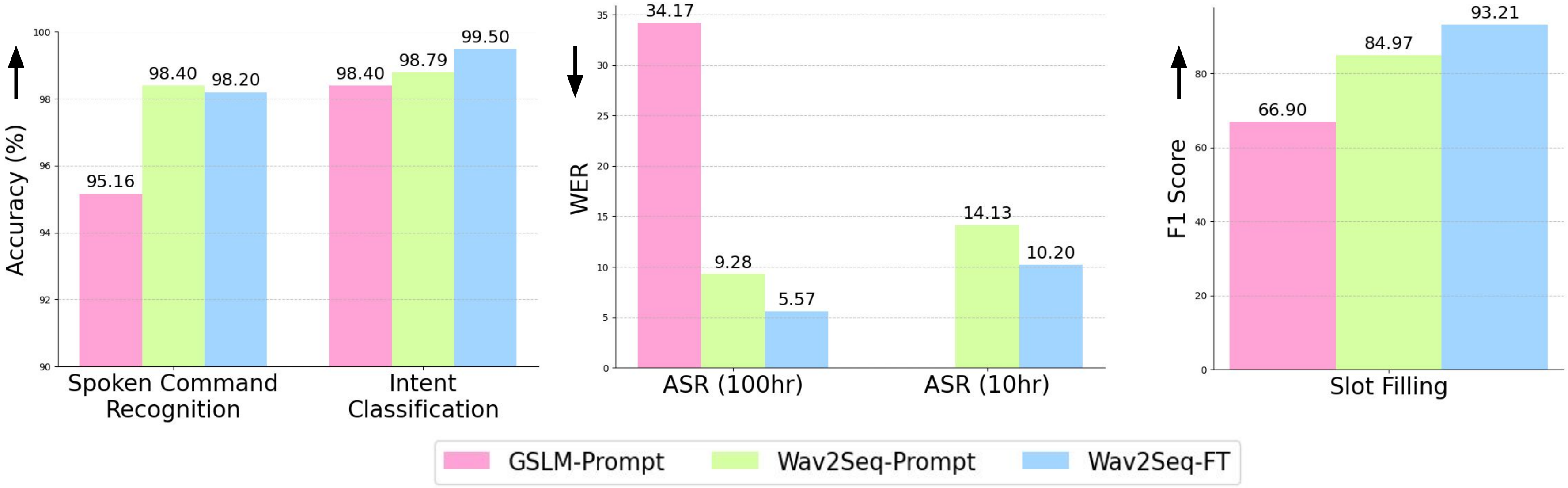
Decoder-only vs. Encoder-Decoder Speech LM

Model	Architecture	Params	Data	Pre-training Task
GSLM	HuBERT Unit (input) + 12-layer Transformer **(Decoder-only)**	~150M	LibriLight 60k hours	Next-token prediction
Wav2Seq	HuBERT Encoder + 6-layer Transformer **(Encoder-Decoder)**	~150M	LibriSpeech 960 hours	Pseudo speech recognition

GSLM has more training data

Similar model size

Decoder-only vs. Encoder-Decoder Speech LM



Speech Classification | **Sequence Generation**

Comparable performance | **Prompt Wav2Seq** is much better than **prompt GSLM**

Prompting and adapter tuning for self-supervised encoder-decoder speech model, ASRU2023 (https://arxiv.org/abs/2310.02971)

Sequence Generation Tasks

1. Prompting **Unit mBART** can achieve competitive performance

2. Prompting an **Encoder-Decoder** model is better than prompting a **Decoder-only** model

Outline

Diverse Speech Processing Tasks

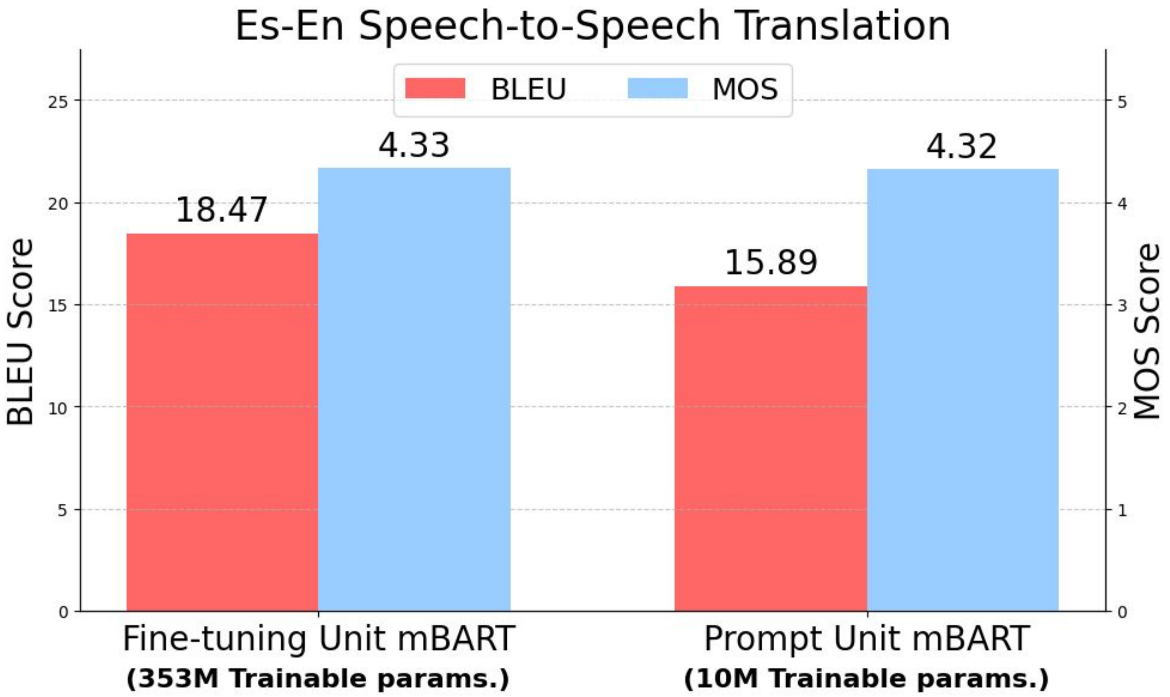
Prompting Speech LM

Experiment Results

Further improvement

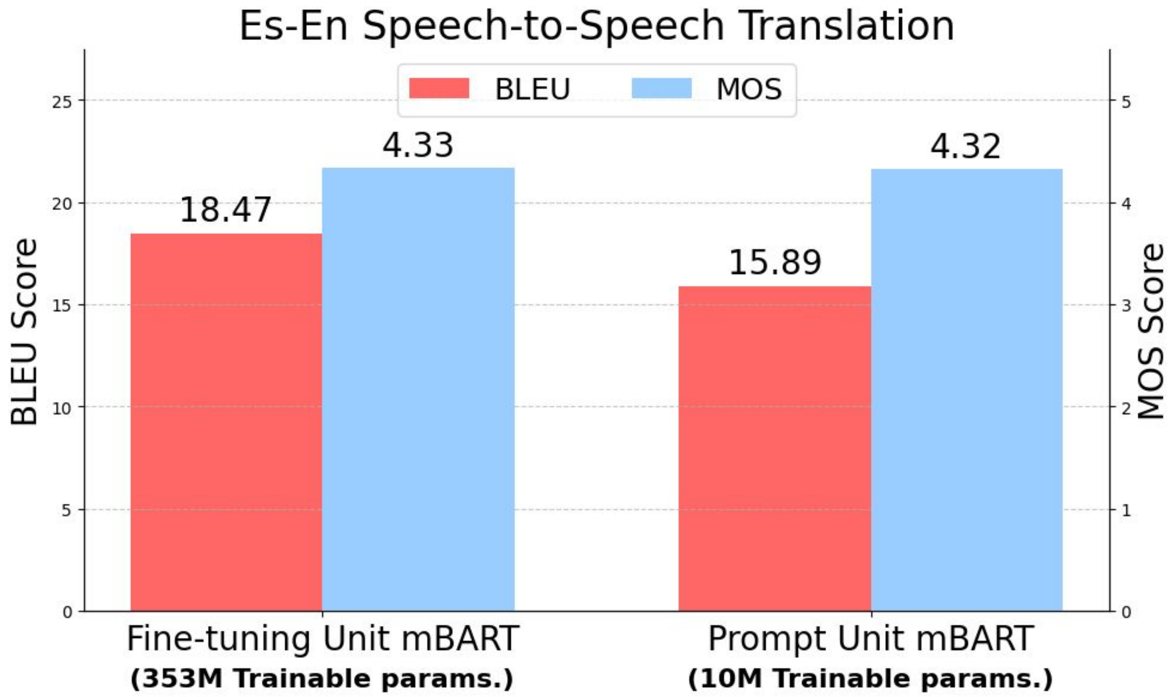
- Speech Classification Tasks
- Sequence Generation Tasks
- **Speech Generation Tasks**

Speech Generation



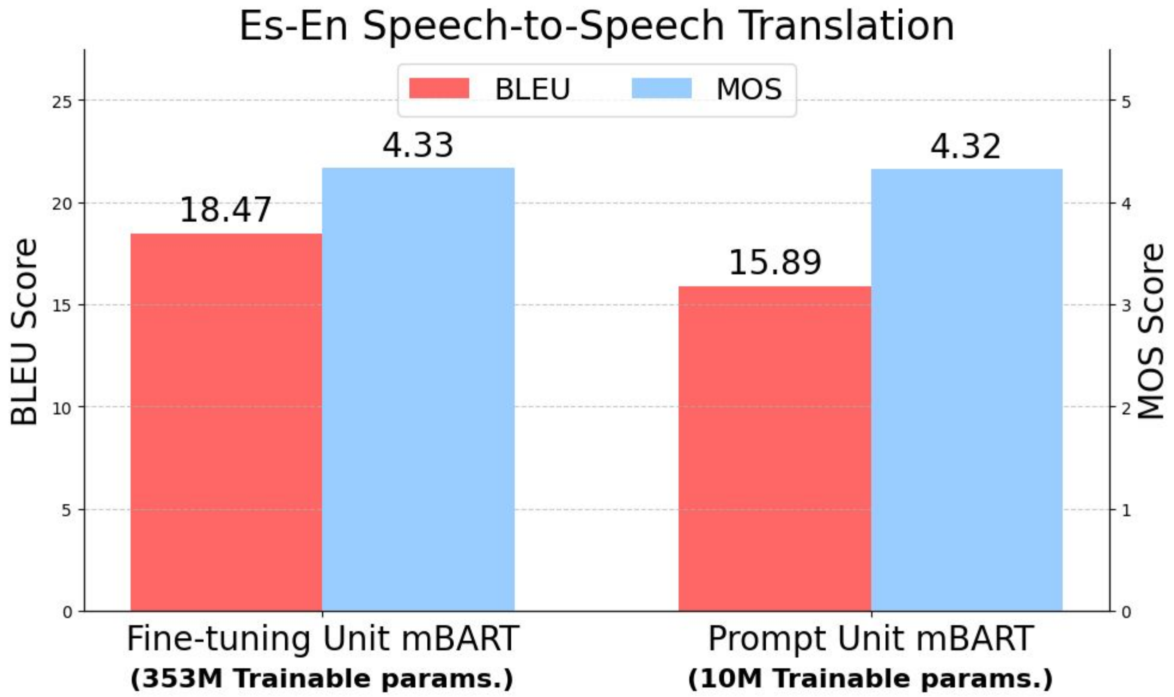
- BLEU score: Translation quality
- MOS score: Speech quality

Speech Generation



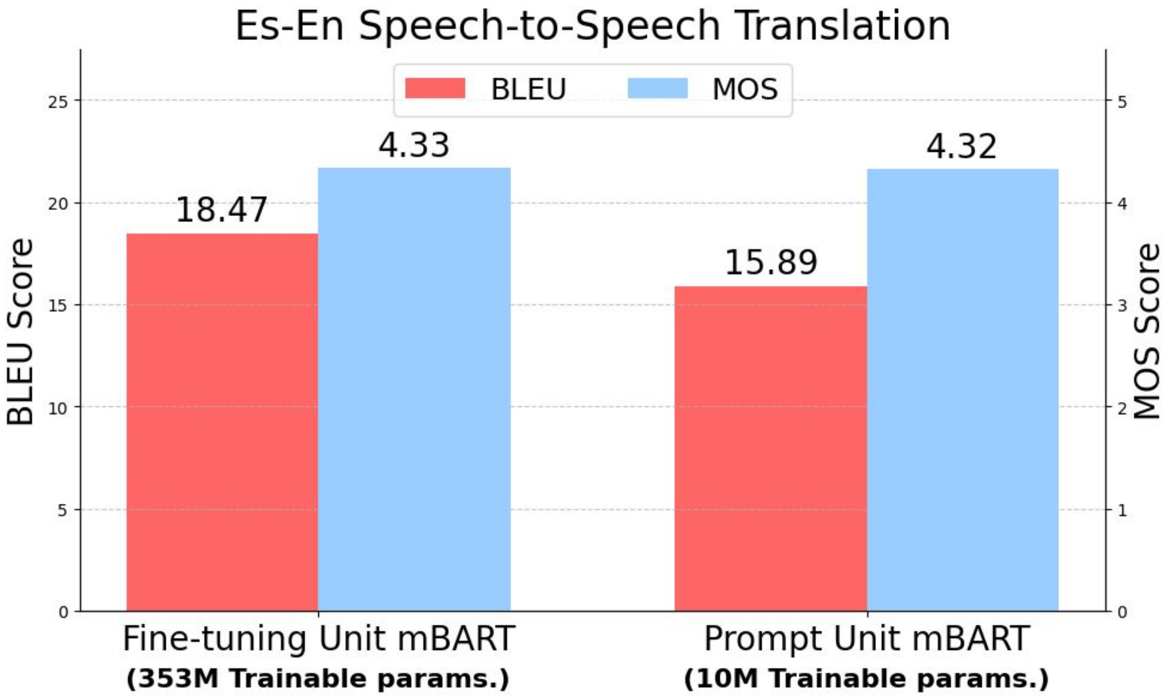
- Left: Fine-tuning the whole Unit mBART **(353M params.)**
- Right: Prompting Unit mBART **(10M params)**

Speech Generation



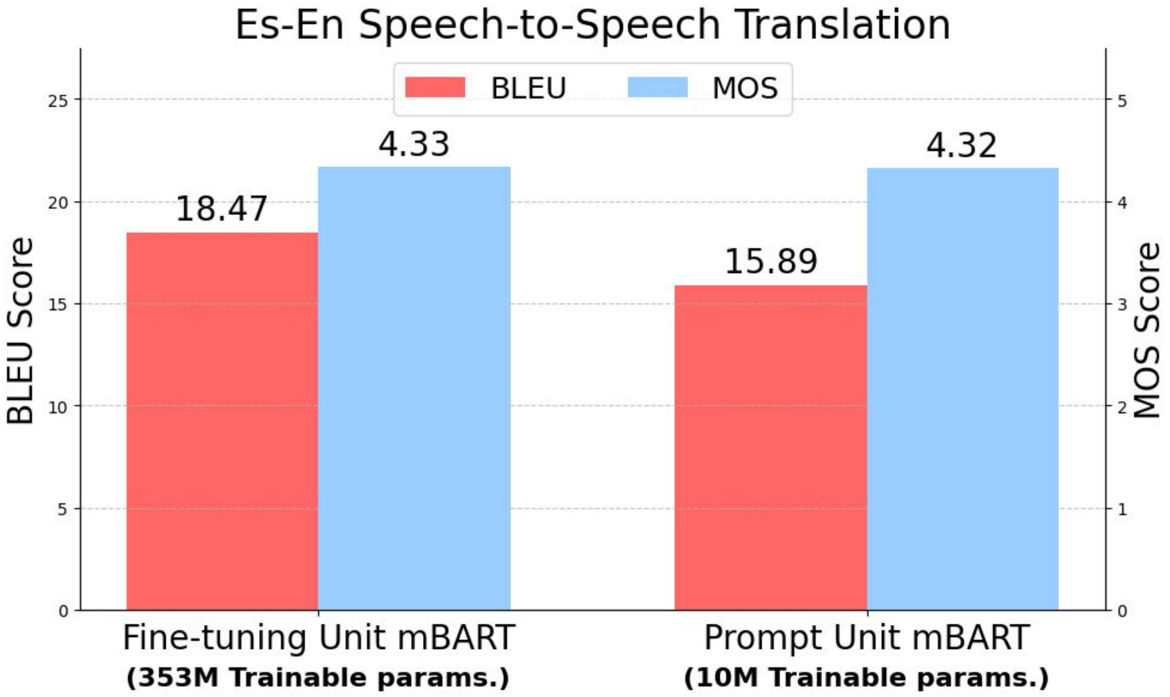
- Prompting: Performance drop but with much fewer trainable params.
- Both have similar MOS score.

Speech Generation



- Fine-tuning HuBERT/mHuBERT fails
- GSLM also fails

Speech Generation



- Speech-to-speech translation is challenging, often require auxiliary tasks

Direct speech-to-speech translation with a sequence-to-sequence model (https://arxiv.org/abs/1904.06037)

Summary

1. Prompting **GSLM** is feasible in speech classification tasks

2. Prompting **Wav2Seq** is competitive in speech classification and sequence generation

3. Prompting **Unit mBART** can achieve competitive performance in diverse tasks

<u>As more advanced Speech LM came out.</u>
<u>The performance is getting better</u>

Outline

Diverse Speech Processing Tasks

↓

Prompting Speech LM

↓

Experiment Results

↓

Further improvement

- Speech Classification Tasks
- Sequence Generation Tasks
- Speech Generation Tasks

Fully Utilize the information in Discrete Units

- Until now, we use <u>random mapping</u> to bridge the units and the labels.
  - Speech classification tasks
  - Sequence generation tasks

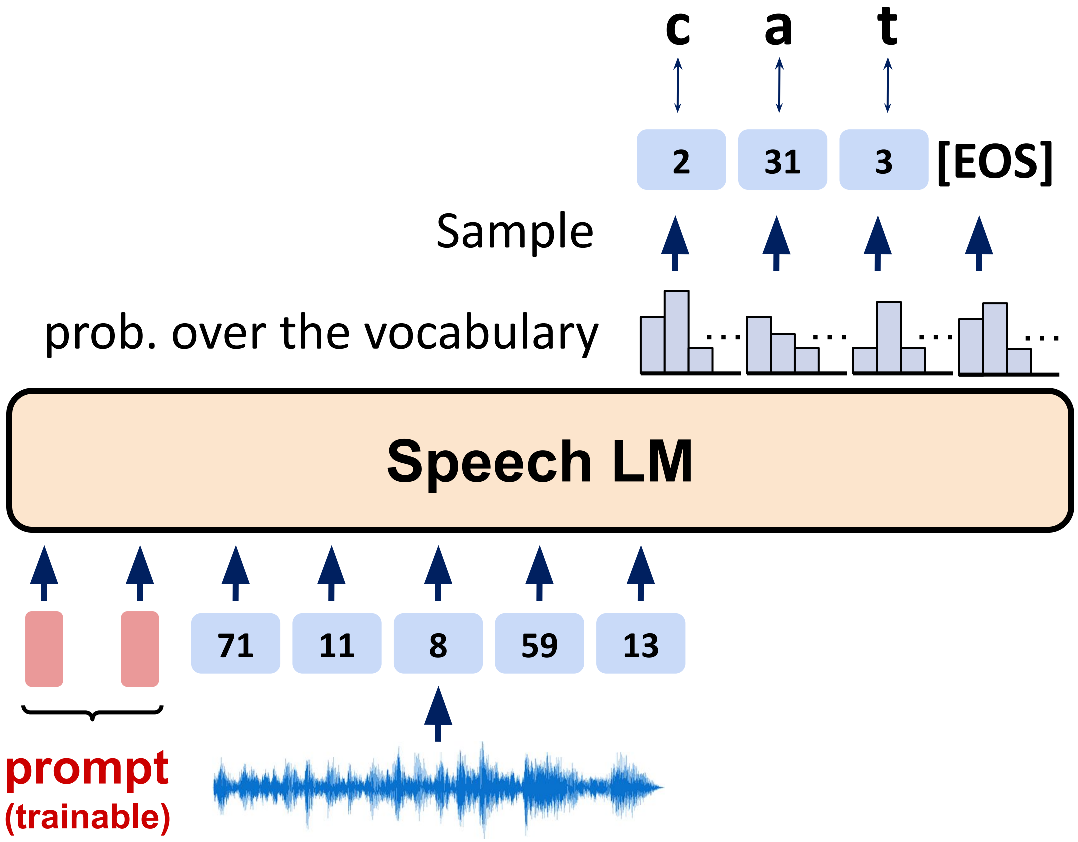
Character	Unit ID
a	31
b	7
c	2
…	…
t	3
…	…

Mapping table (Verbalizer)

Contains rich information.

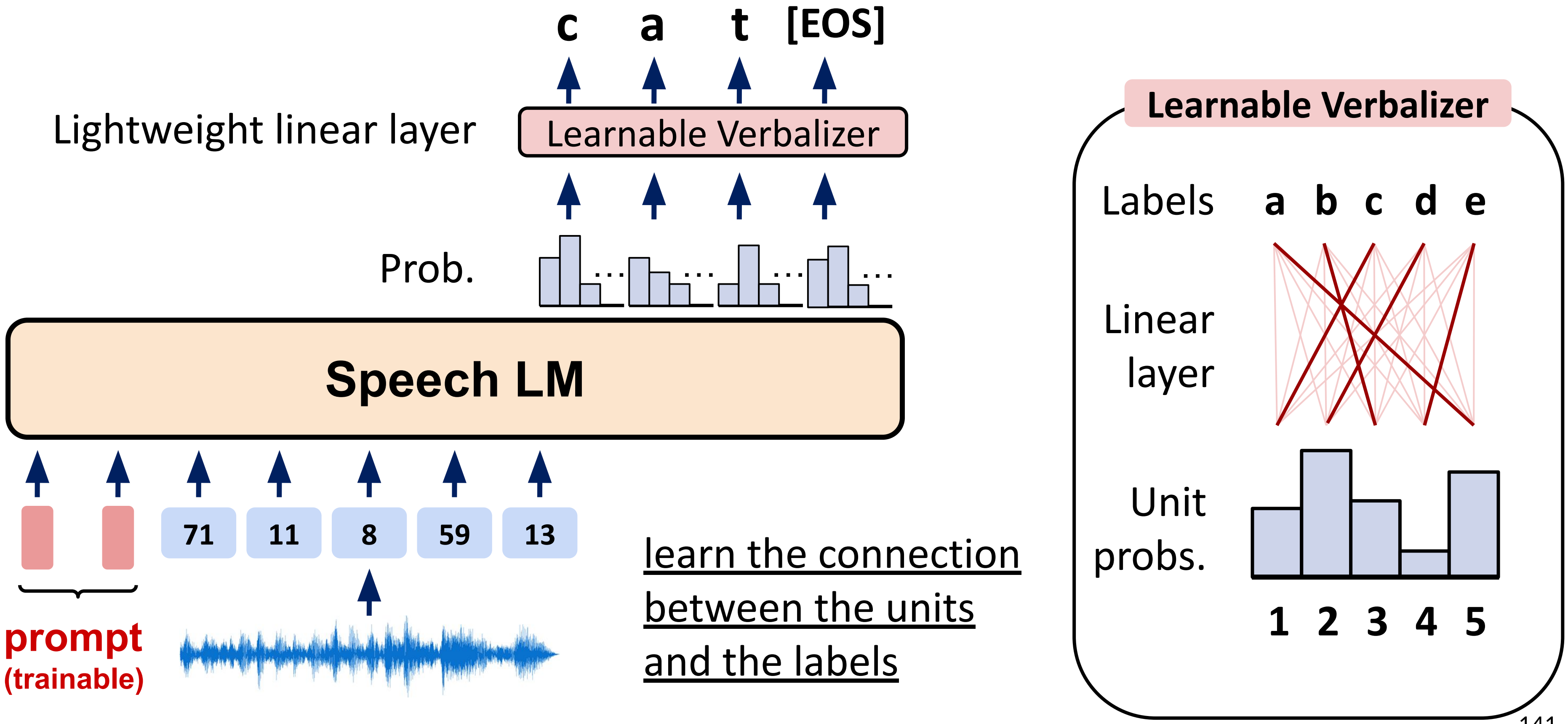
<u>Can we fully utilize the information in discrete units?</u>

Fully Utilize the information in Discrete Units



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

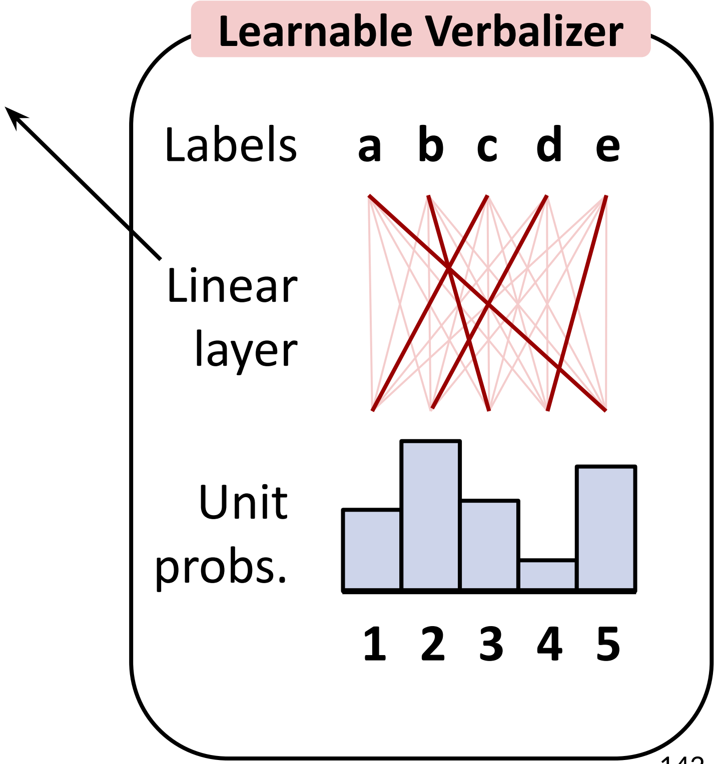
Fully Utilize the information in Discrete Units



Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	'B'	'H'	'V'



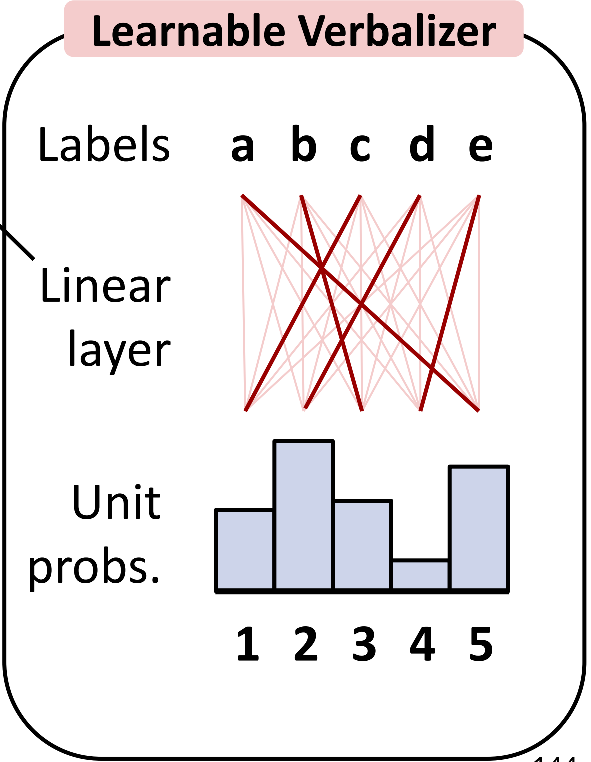
Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	'B'	'H'	'V'
Unit	290	470	577

Largest weight in the linear layer for a specific label.

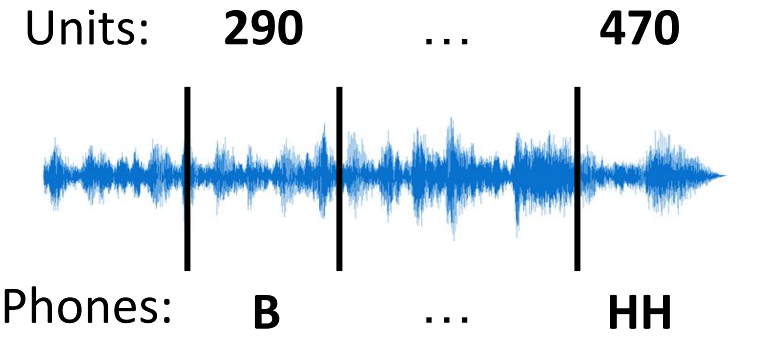
What is the meaning of these discrete units?



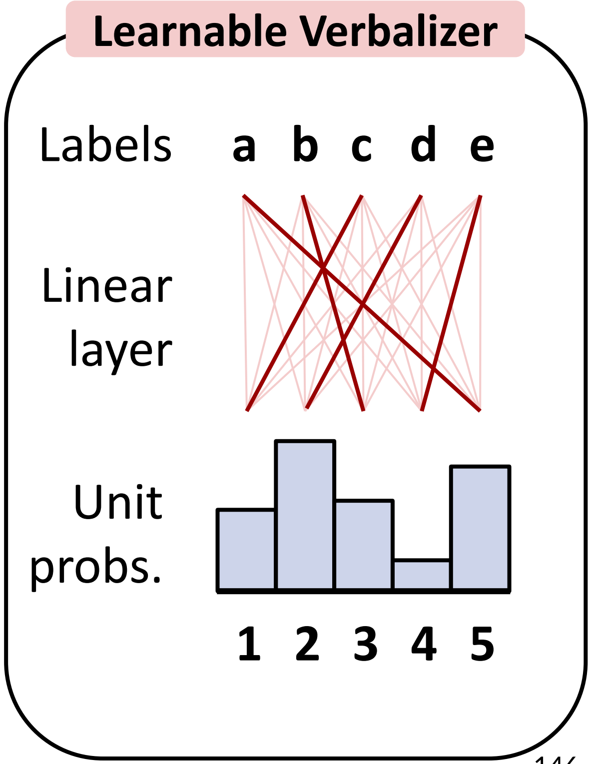
Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	‘B’	‘H’	‘V’
Unit	290	470	577
Phoneme	B	HH	V



Forced alignment on LibriSpeech

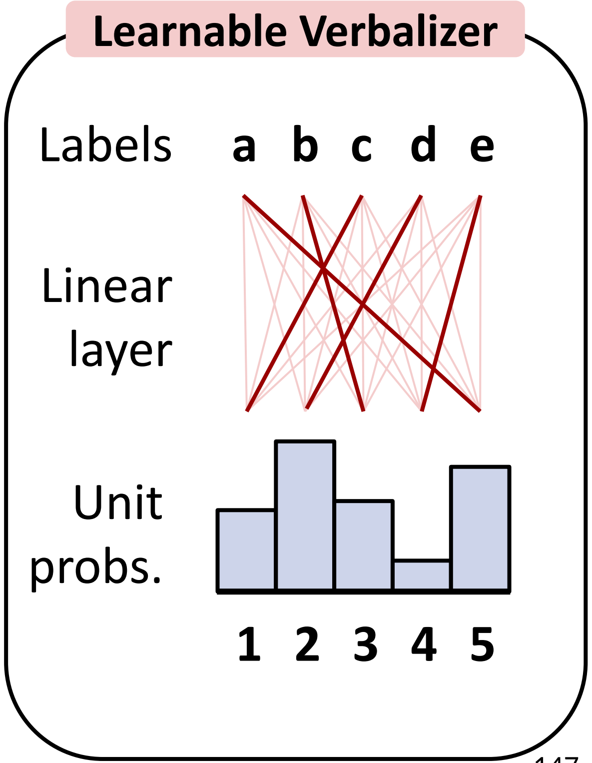


Learnable Verbalizer - A Case Study

For prompting Unit mBART in ASR

Label	'B'	'H'	'V'
Unit	290	470	577
Phoneme	B	HH	V

- The learnable verbalizer can automatically find the units for the labels

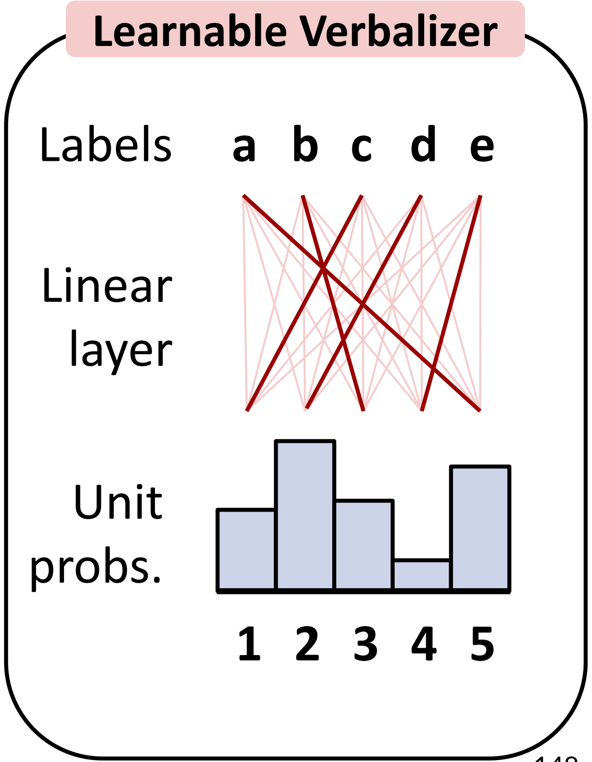


Learnable Verbalizer - A Case Study

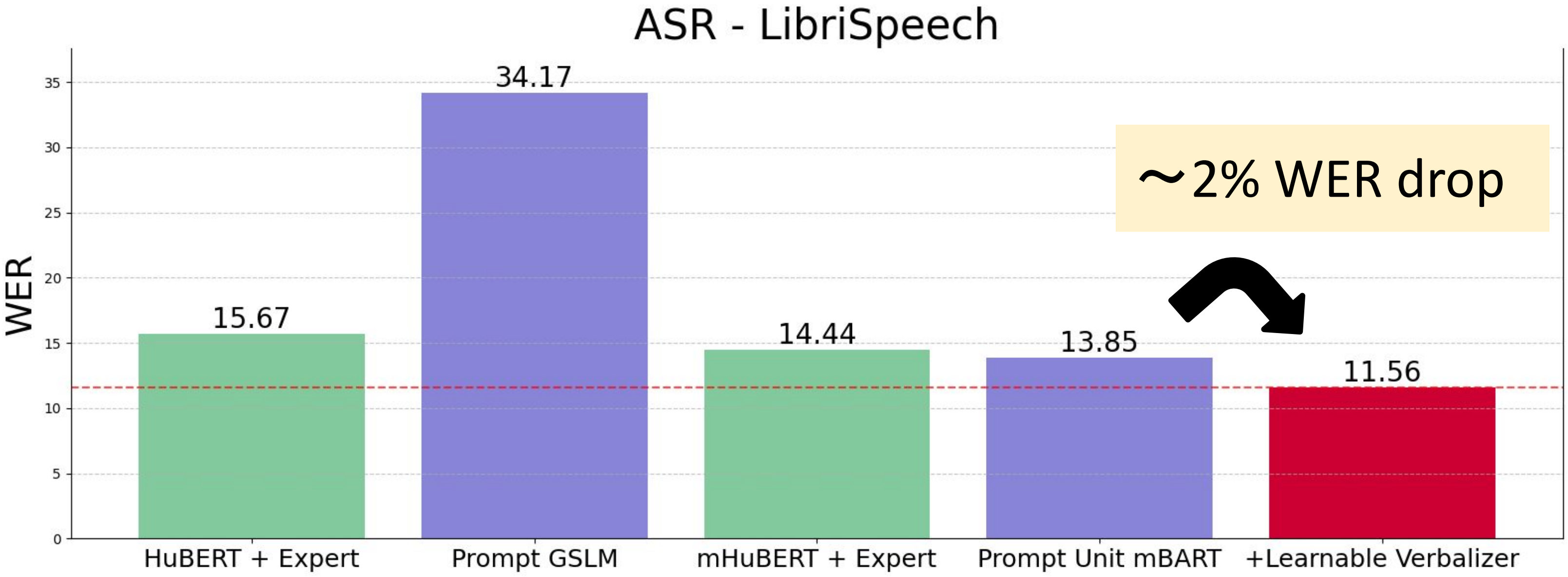
For prompting Unit mBART in Phoneme Recognition (PR)

Label	'F'	'K'	'TH'
Unit	958	487	918
Phoneme	F	K	TH

- The learnable verbalizer can automatically find the units for the labels



Learnable Verbalizer - A Case Study

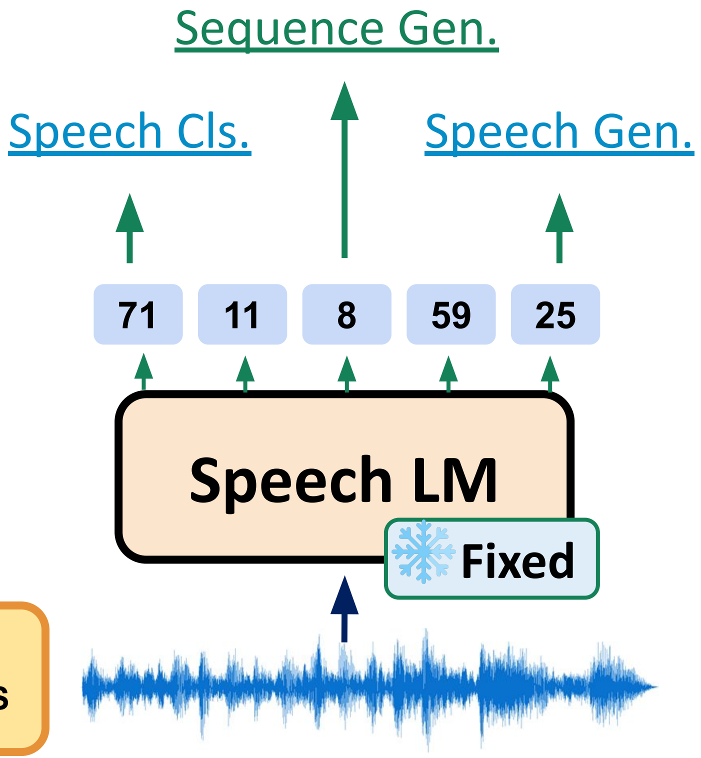


- Performance improvement with learnable verbalizer.
- With additional parameters less than **0.03M** (**~1%** of the prompt parameters)

Prompting Paradigm

1. Can prompting technology be applied to speech processing?

2. Can it achieve parameter efficiency compared to fine-tuning paradigm?

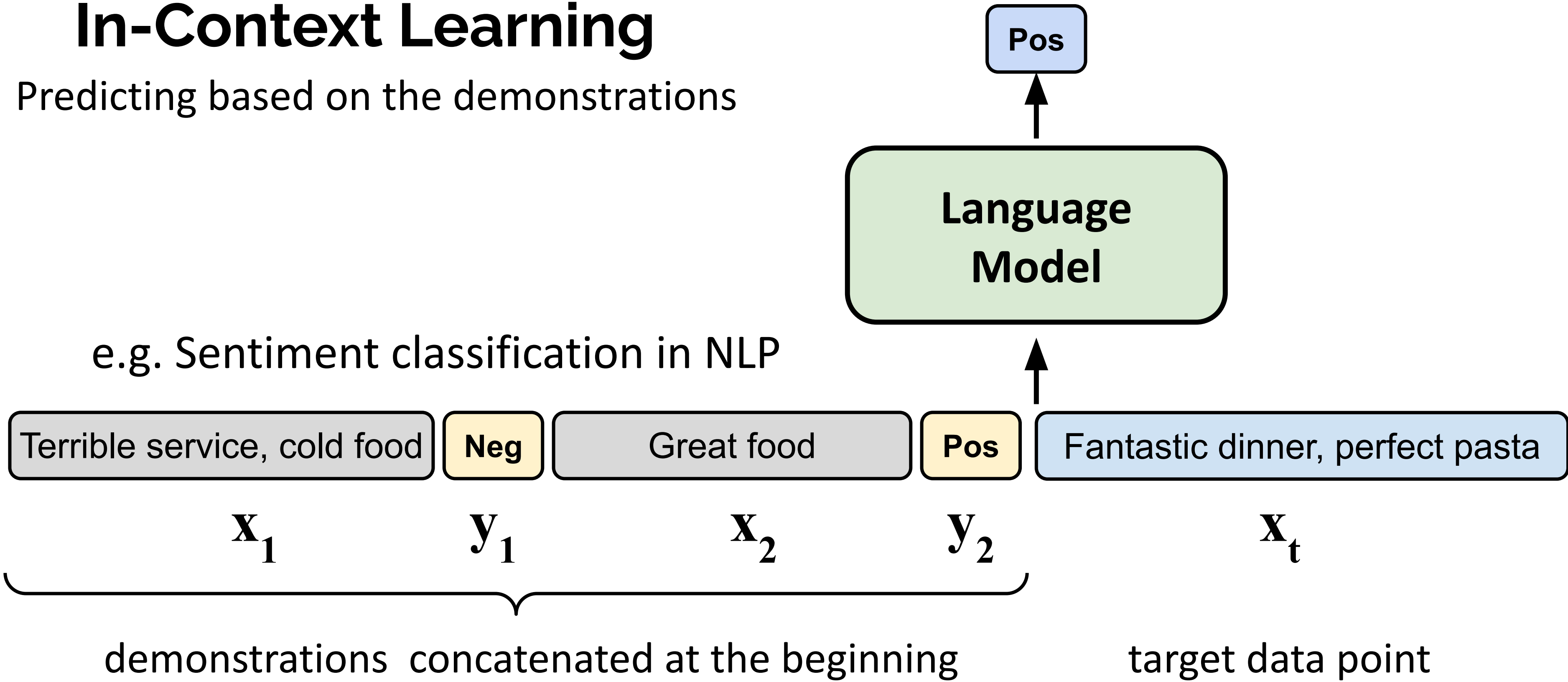


Limitation:
Still require training for a specifc task

In-Context Learning for Speech LM

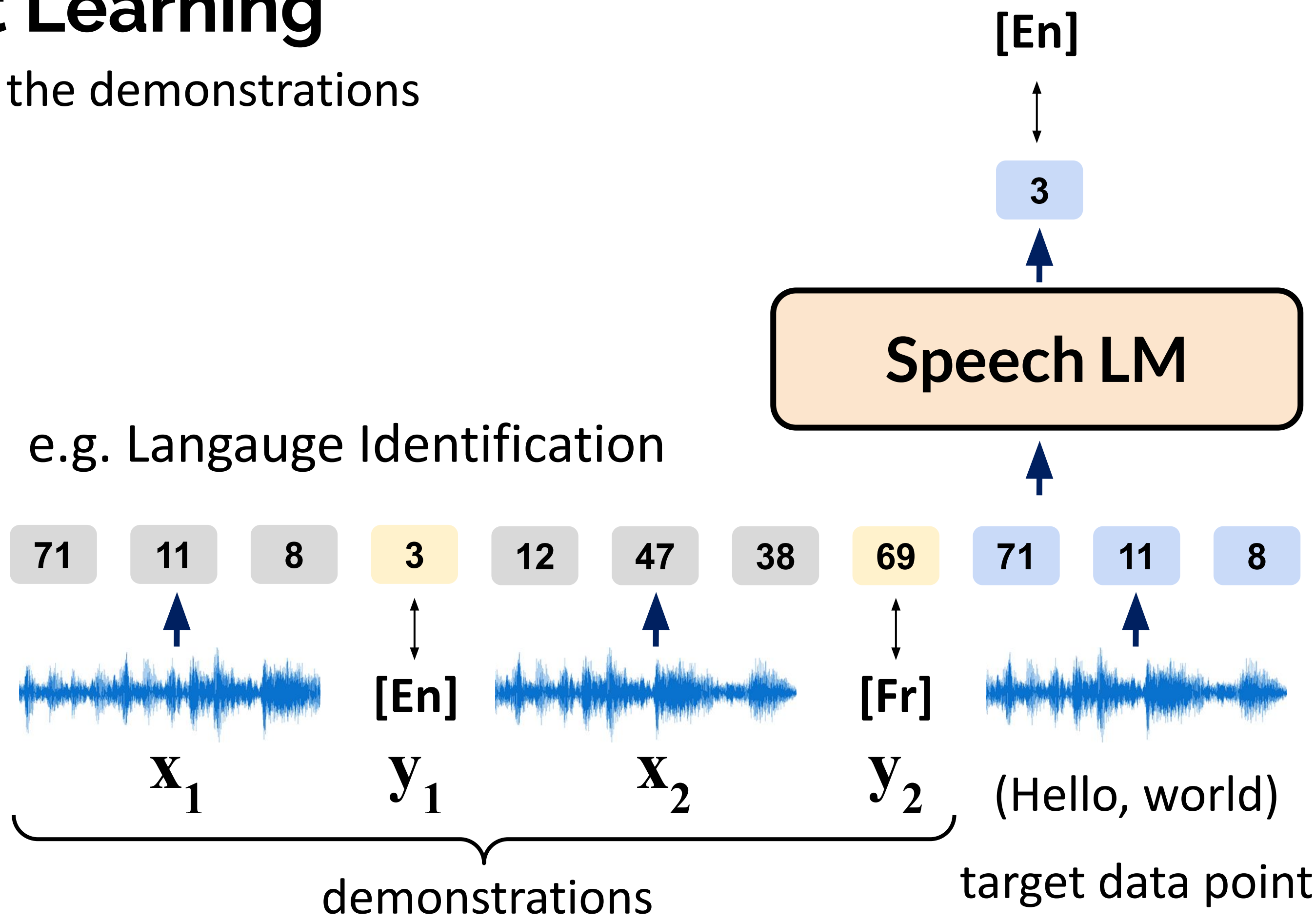
In-Context Learning

Predicting based on the demonstrations



In-Context Learning

Predicting based on the demonstrations

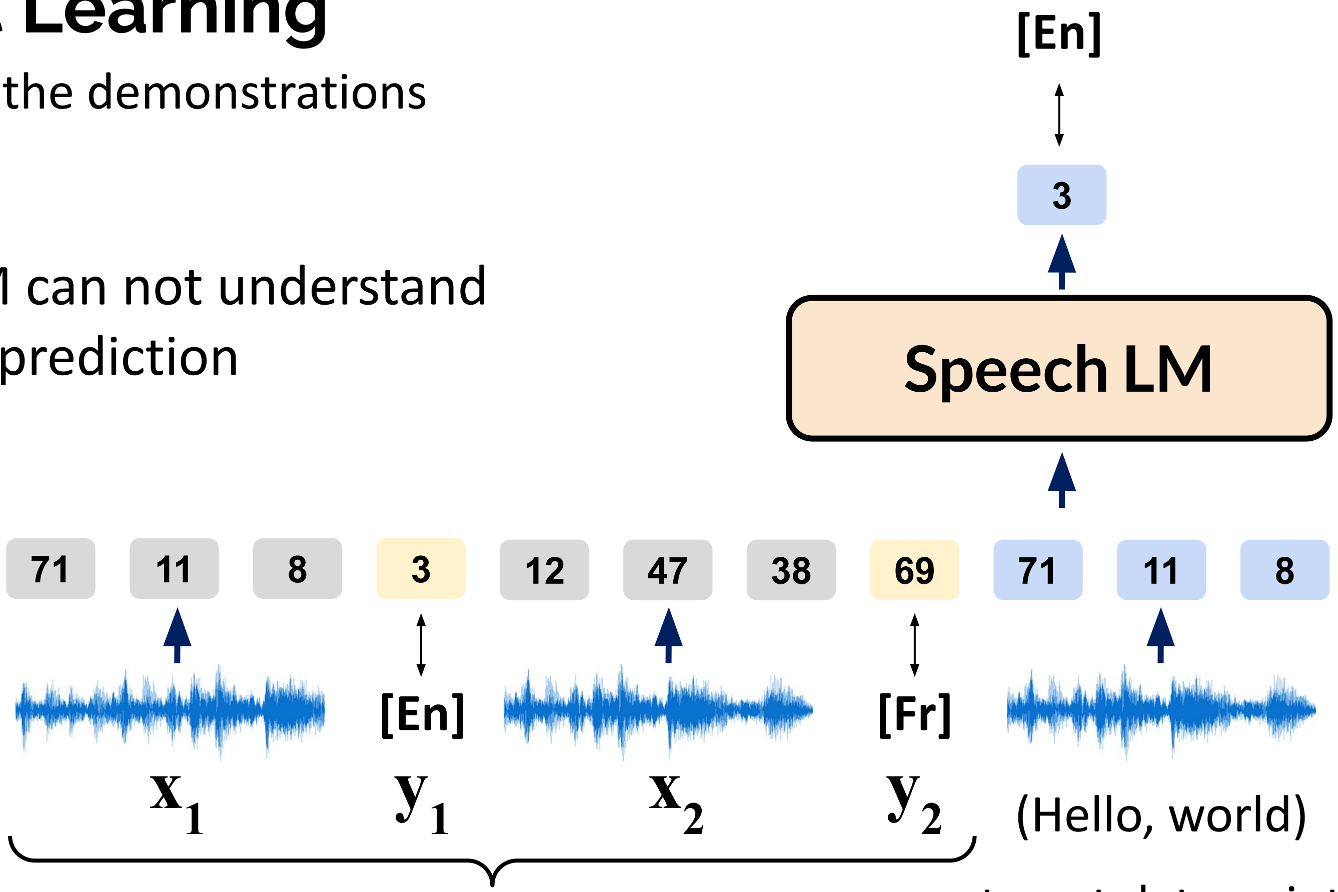


In-Context Learning

Predicting based on the demonstrations

The original GSLM can not understand and fails to make prediction

LLM can take care of random labels



demonstrations

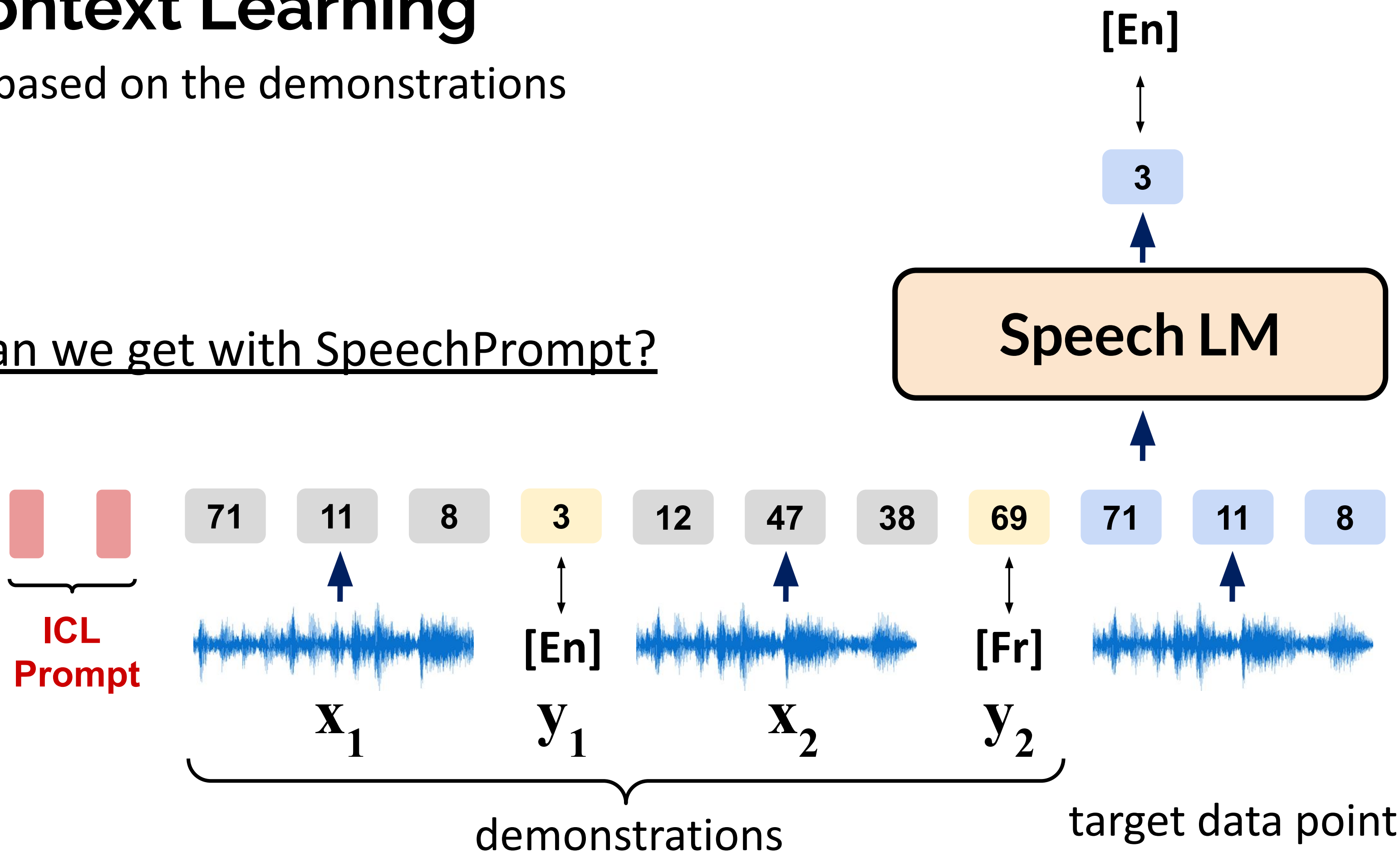
target data point

Larger language models do in-context learning differently (https://arxiv.org/abs/2303.03846)

In-Context Learning

Predicting based on the demonstrations

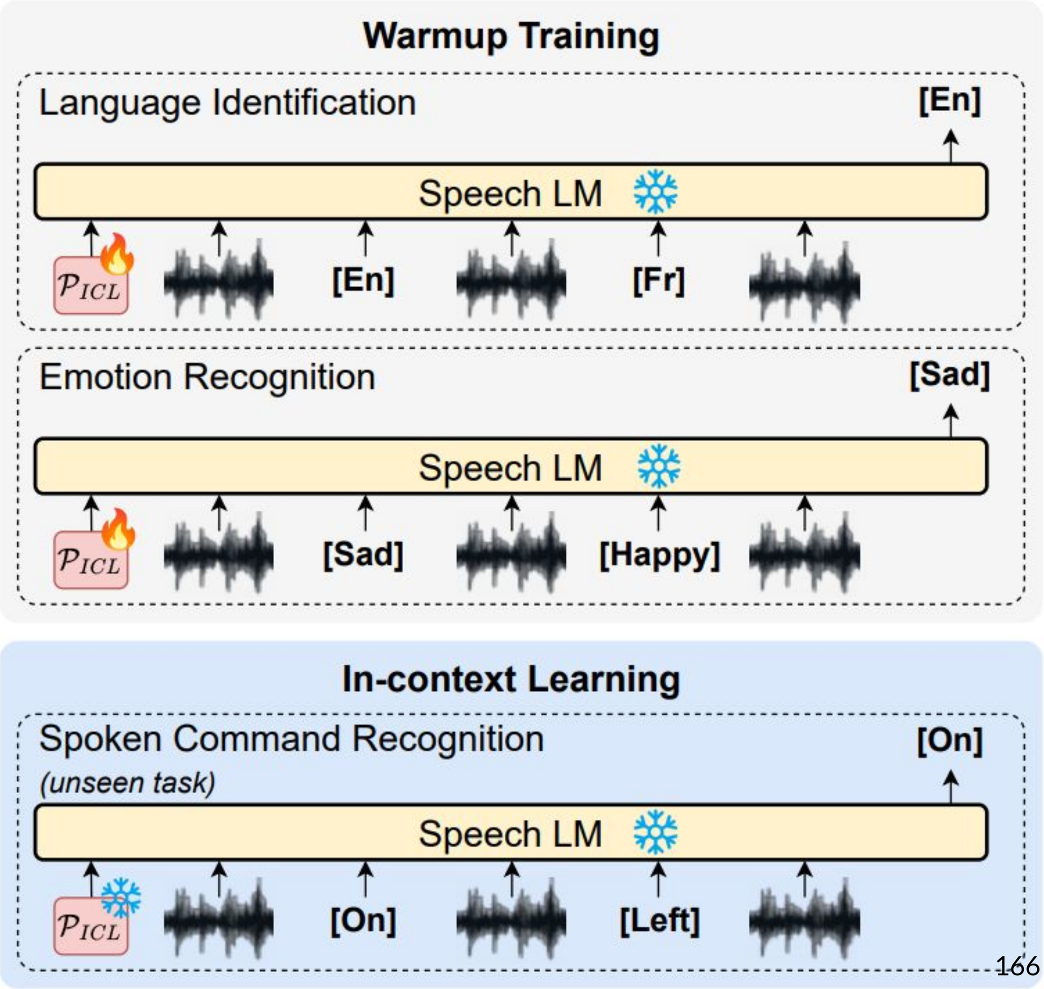
How far can we get with SpeechPrompt?



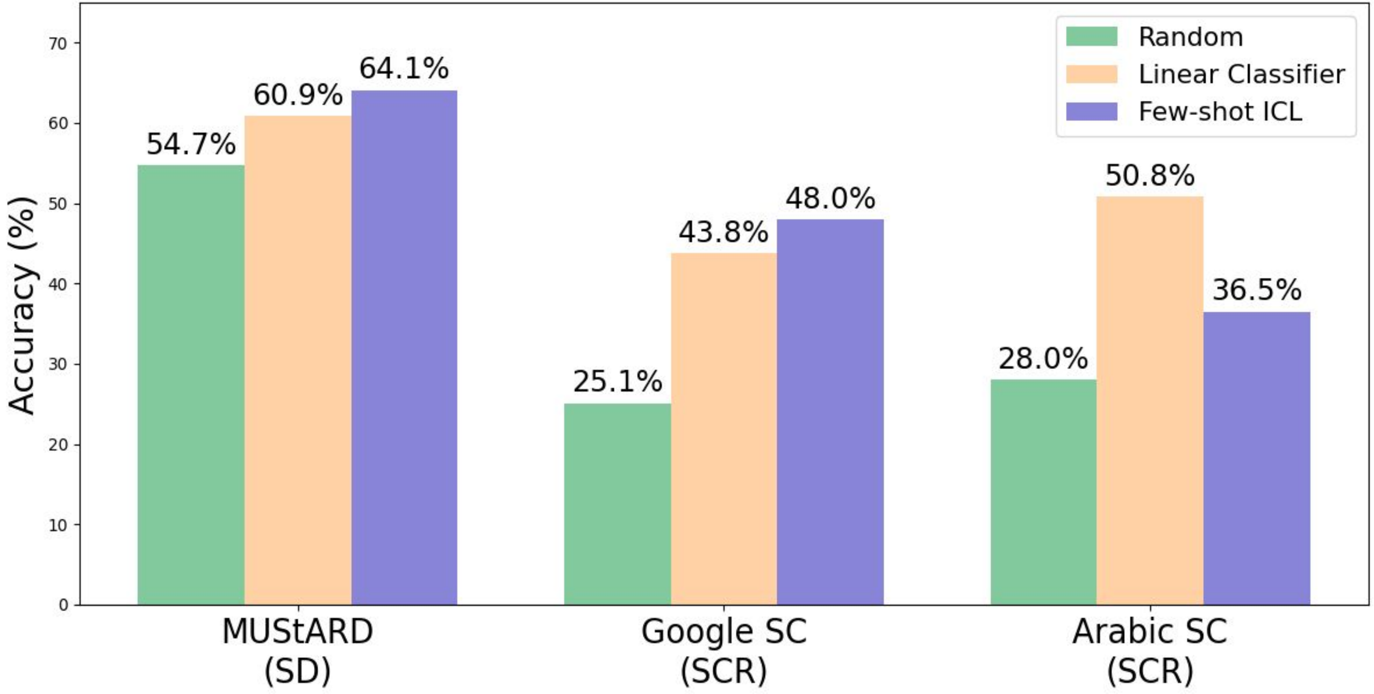
- **Warmup Training:**
  Learn ICL prompts to enable the speech LM with ICL capability.

ICL-Speech LM

- **In-context Learning**
  - The LM is fixed
  - The prompt is fixed
  - The task is unseen

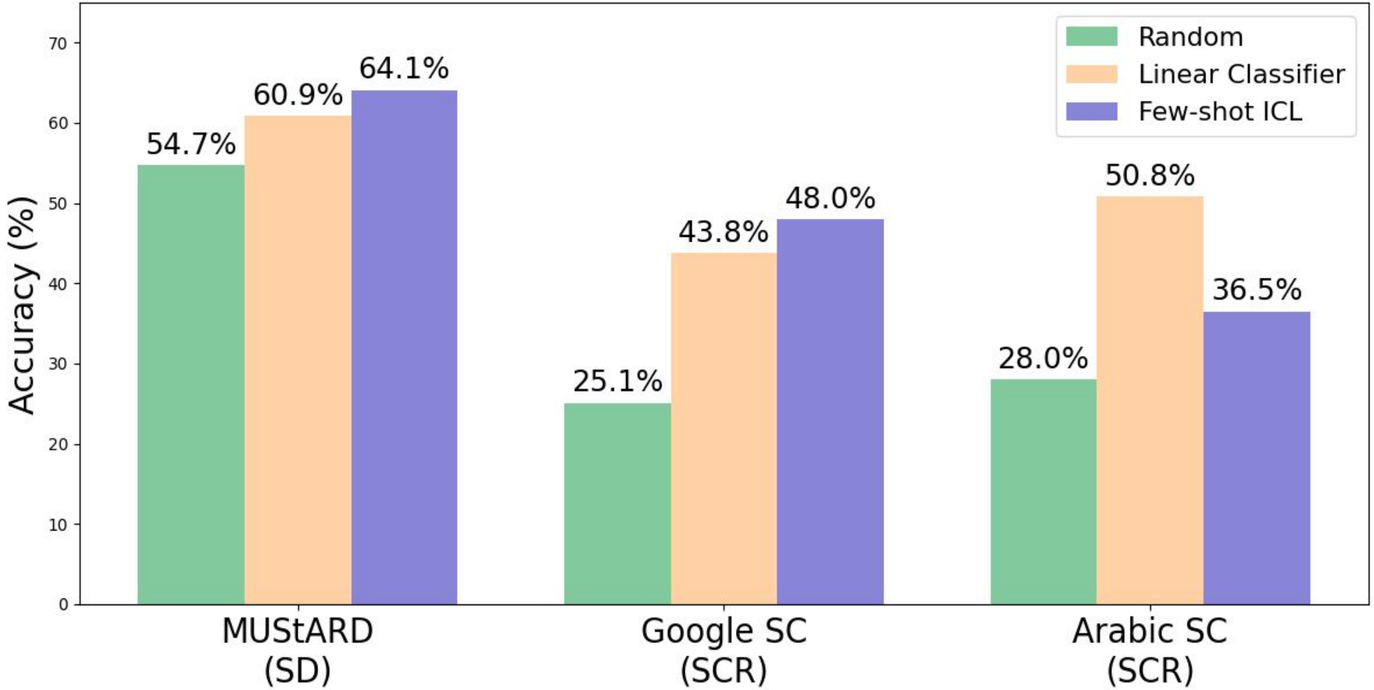


In-Context Learning on Unseen Tasks



Unseen Speech classification tasks

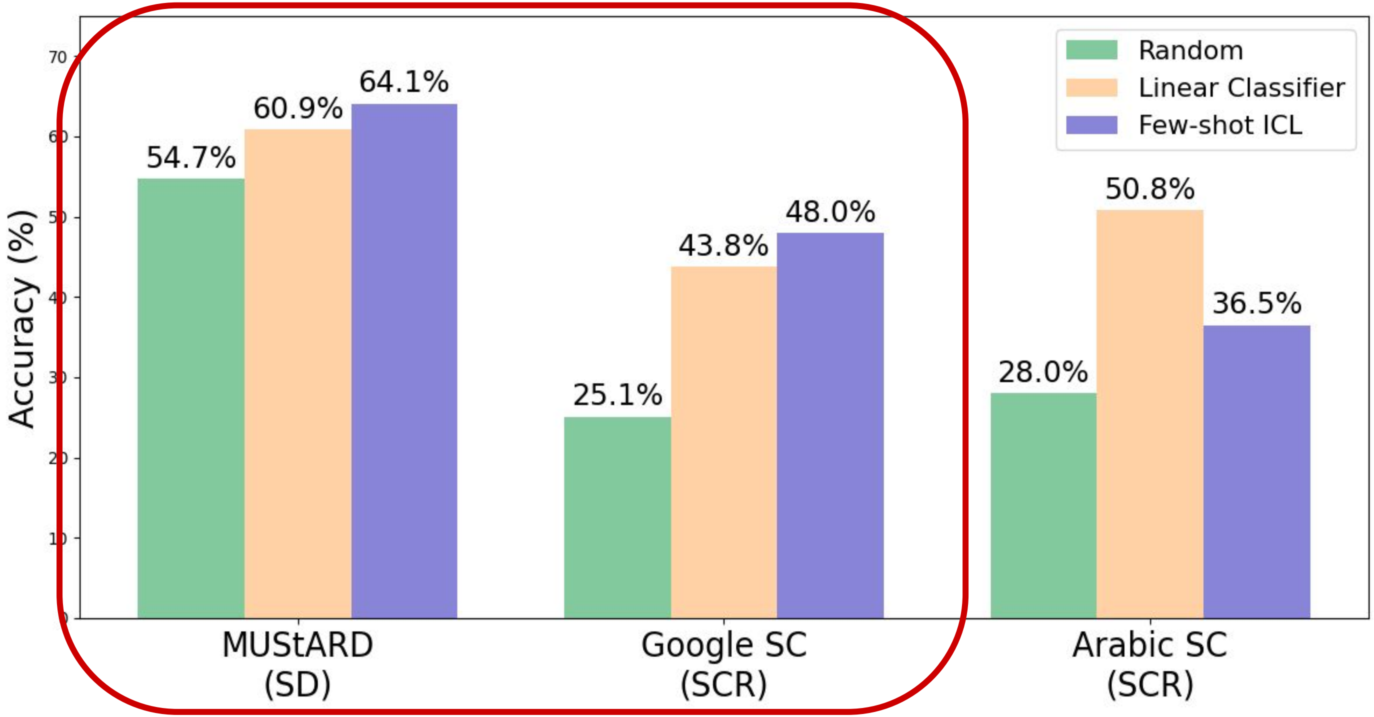
In-Context Learning on Unseen Tasks



Warmup training:

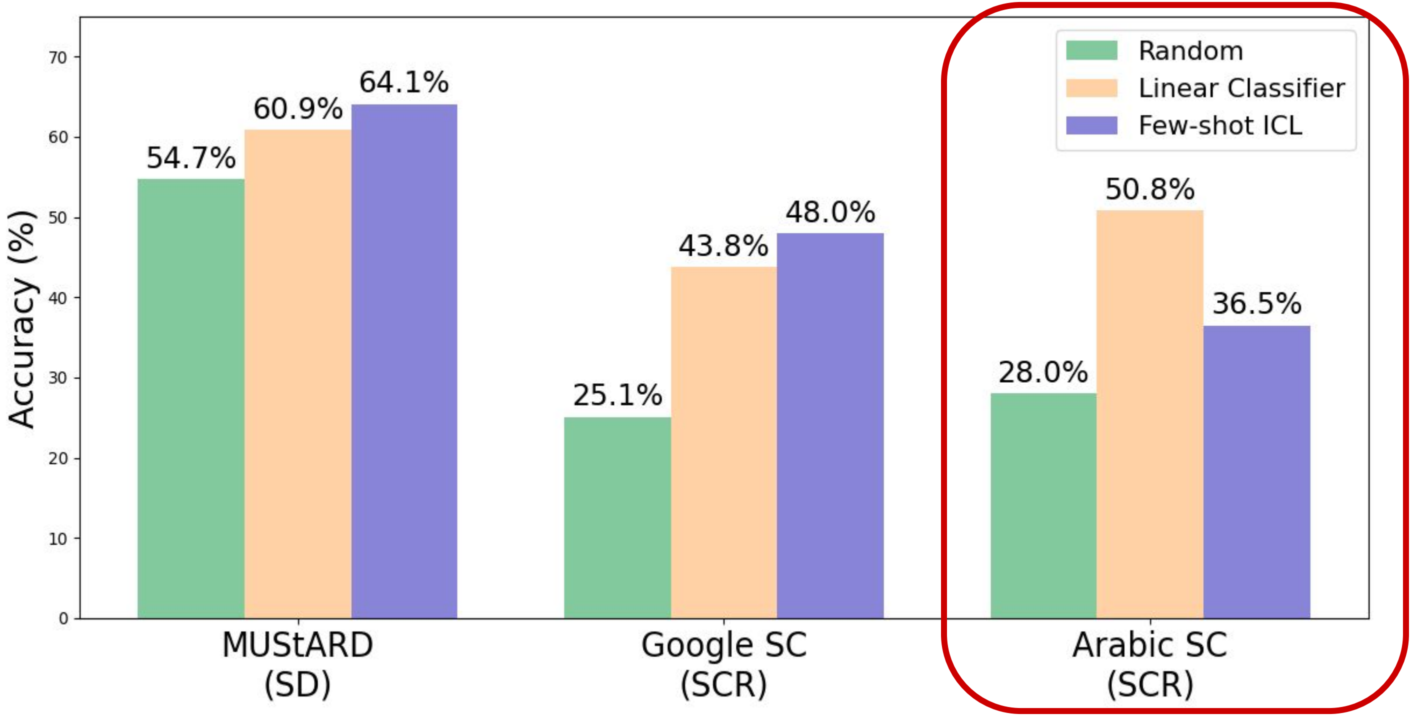
Mandarin SCR, Lithuanian SCR, Language ID, Emotion Recognition

In-Context Learning on Unseen Tasks



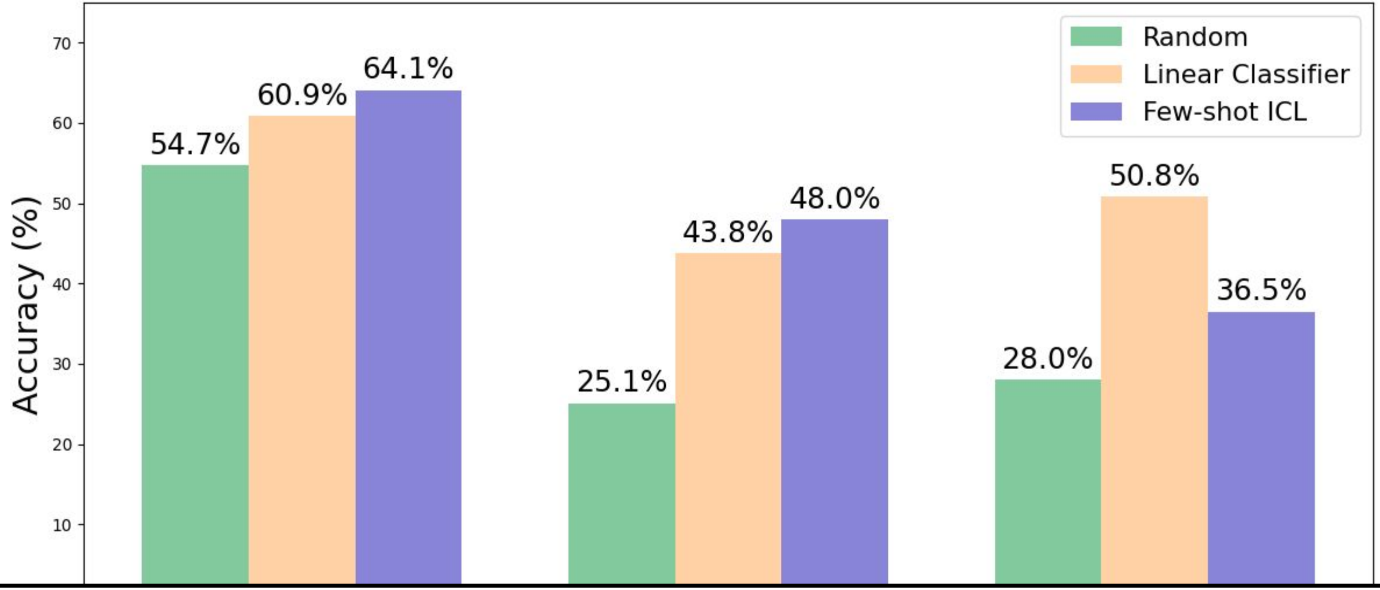
- GSLM can perform **In-context Learning** outperforming **random** guessing and **linear classifier**

In-Context Learning on Unseen Tasks



- In-context Learning underperform linear classifer probably due to cross-lingual setting

In-Context Learning on Unseen Tasks



There's still a big performance gap between the simple supervised models.
Surprising to get a non trivial result.

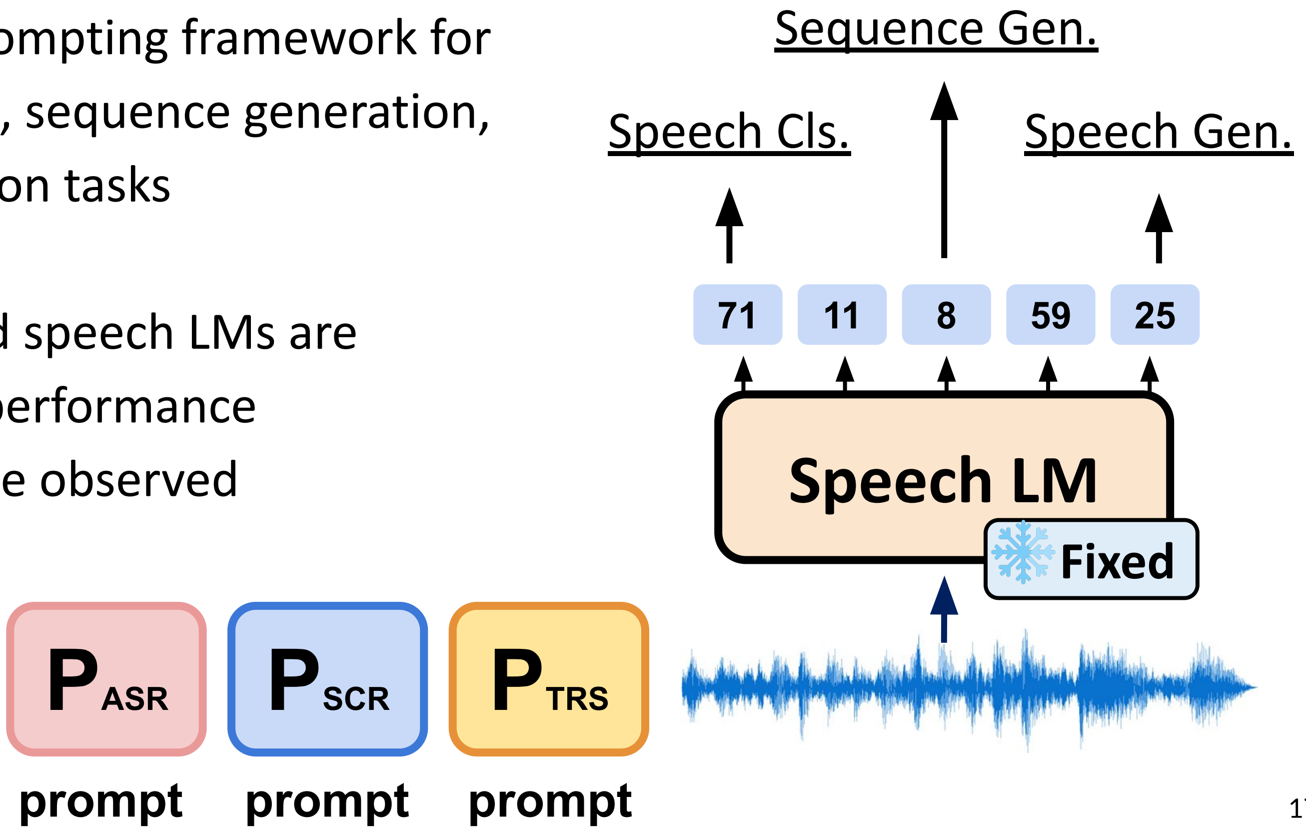
- GPT-3 ~ 170B parameters
- GSLM ~ 150M parameters + prompts (0.2M)

Conclusion

Conclusion

- Achieve a unified prompting framework for speech classification, sequence generation, and speech generation tasks

- With more advanced speech LMs are developed, further performance improvements can be observed



Future Works

Future Work:

Develop a more powerful and user friendly Speech service

↓

Nautral language prompts and good reasoning capability

Future Work:

Develop a more powerful and user friendly Speech service

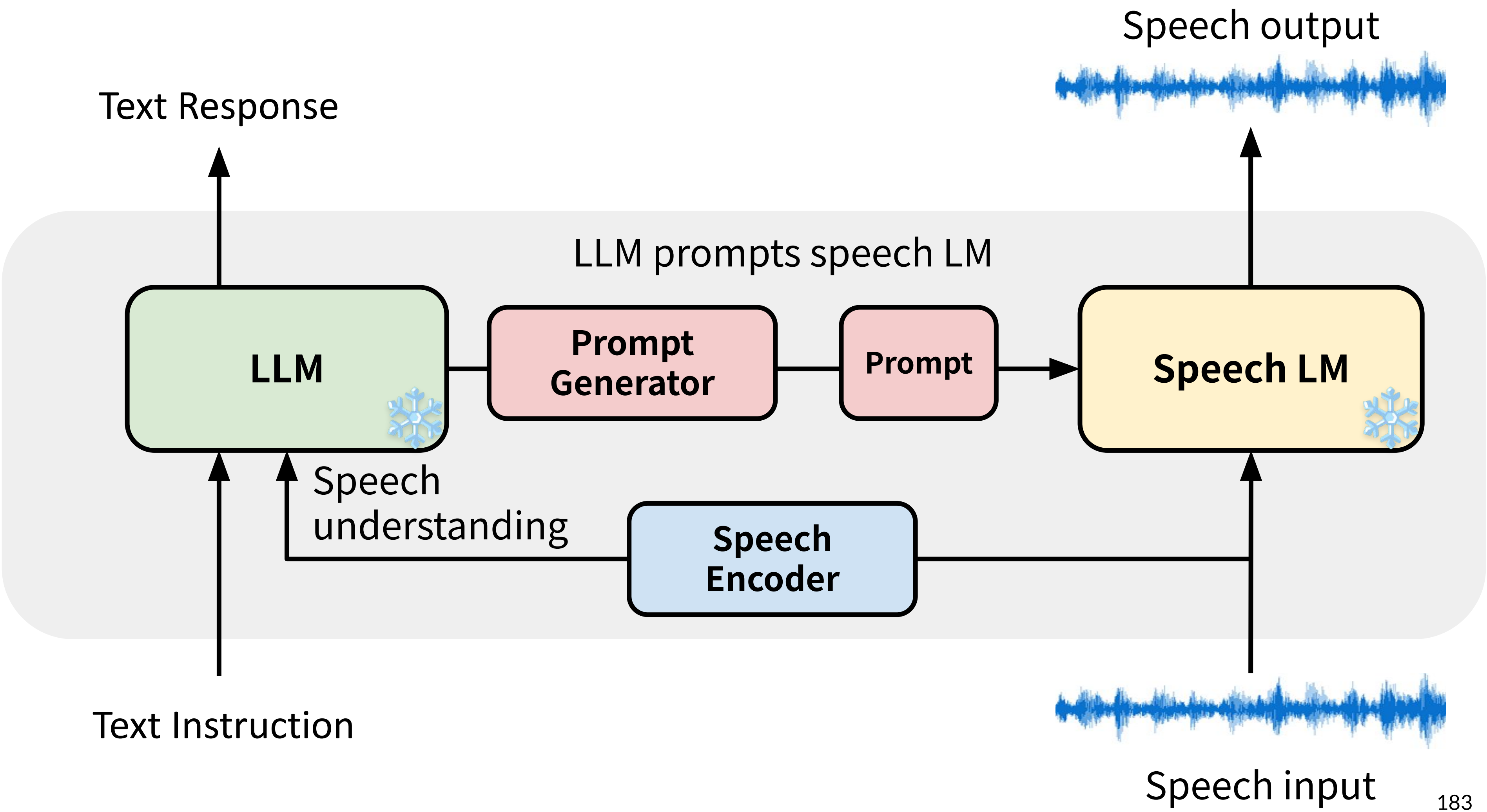
Idea: Develop a framework for combining the LLM and Speech LM

LLM :

- Generate good text response (V)
- Generate speech (X)

Speech LM :

- Generate speech (V)
- Reasoning capability (X)



Speech output
Text Response
LLM prompts speech LM
LLM
Prompt Generator
Prompt
Speech LM
Speech understanding
Speech Encoder
Text Instruction
Speech input

References

[1] **SpeechPrompt: Prompting Speech Language Models for Speech Processing Tasks**
(IEEE/ACM Transactions on Audio Speech and Language Processing, TASLP)
Kai-Wei Chang, *Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee*

[2] **SpeechPrompt: An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks**
(Interspeech 2022)
Kai-Wei Chang, *Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee*

[3] **SpeechPrompt v2: Prompt Tuning for Speech Classification Tasks**
(arXiv Preprint)
Kai-Wei Chang, *Yu-Kai Wang, Hua Shen, Iu-thing Kang, Wei-Cheng Tseng, Shang-Wen Li, Hung-yi Lee*

[4] **SpeechGen: Unlocking the Generative Power of Speech Language Models with Prompts**
(arXiv Preprint)
Kai-Wei Chang, *Haibin Wu, Yuan-Kuei Wu, Hung-yi Lee*

[5] **Prompting and Adapter Tuning for Self-supervised Encoder-Decoder Speech Model**
(ASRU 2023)
Kai-Wei Chang, *Ming-Hsin Chen, Yun-Ping Lin, Jing Neng Hsu, Chien-yu Huang, Shang-Wen Li, Hung-Yi Lee*

[6] **Exploring In-Context Learning of Textless Speech Language Model for Speech Classification Tasks**
(Interspeech 2024)
Kai-Wei Chang, *Ming-Hao Hsu, Shang-Wen Li, Hung-yi Lee*

Thanks for your listening