

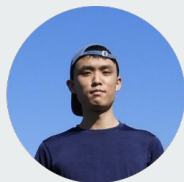
Section 1



Section 2



Section 3-a



Section 3-b
(now)



Section 3



11:20 to 11:50

12:00 to 12:20

opening

super-alignments

reprogramming

parameter-efficient nlp

prompting & speech LMs

close remarks

Single Modality
Parameter Efficient Learning

Cross-Modal
Parameter Efficient Learning

Multi-Task
Parameter Efficient Learning

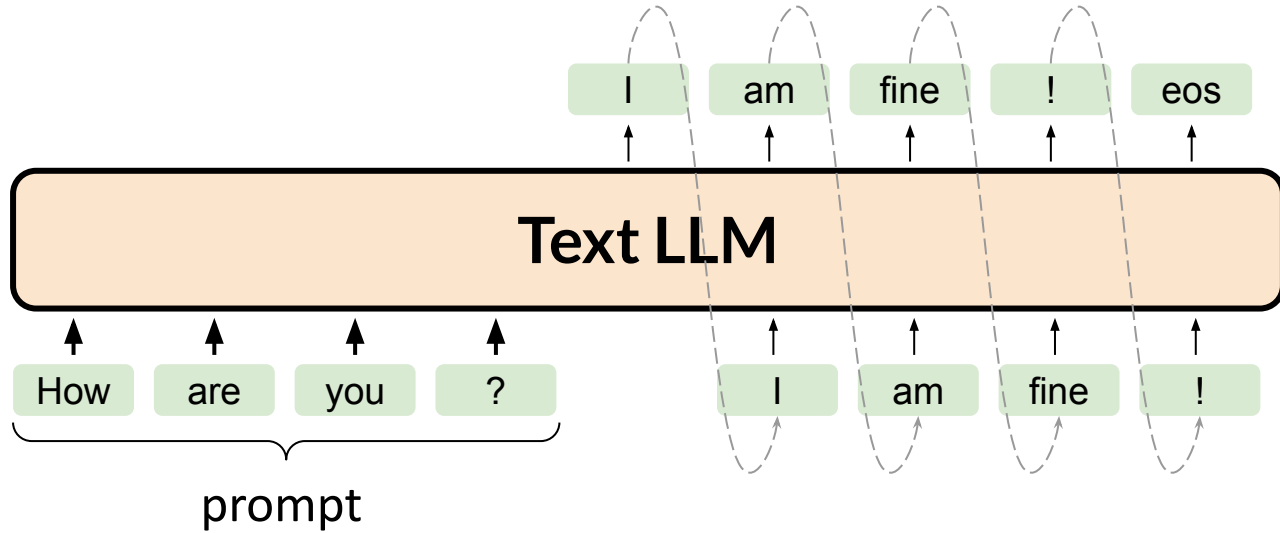
ICASSP 2024 Tutorial-17 part 3 (b)

Speech Language Models

Prompting and Parameter Efficient Learning

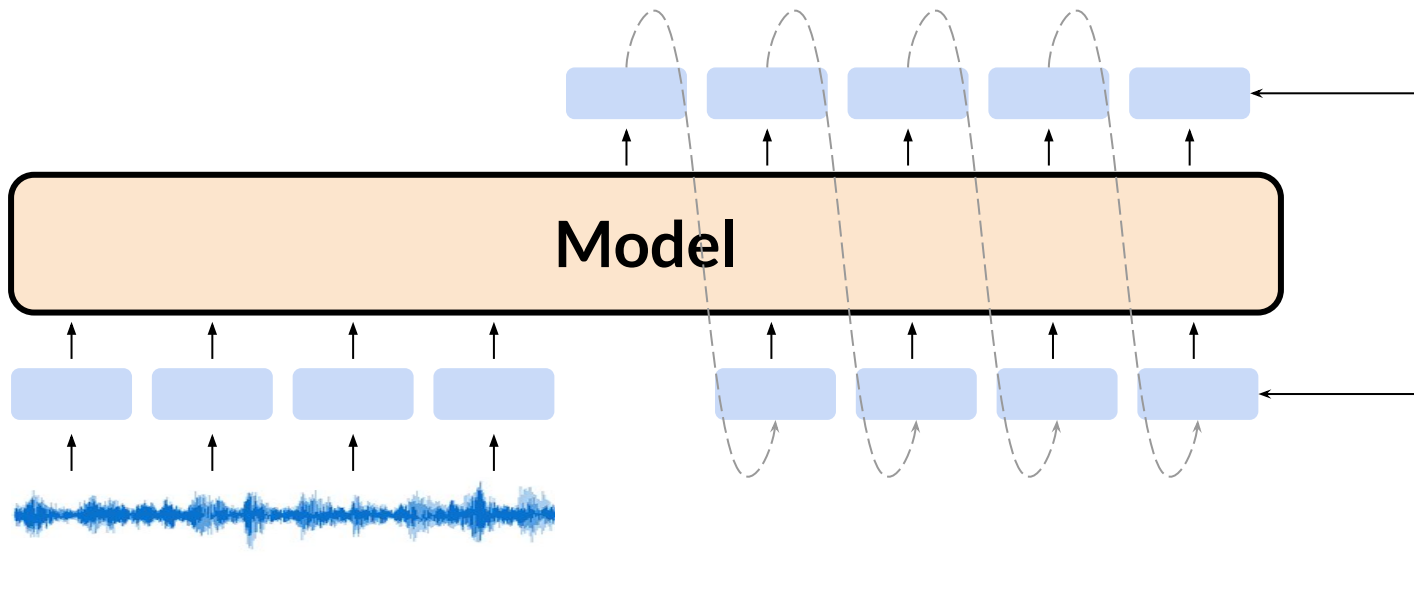
Presenter: Kai-Wei Chang (NTU)
kaiwei.chang.tw@gmail.com





Text LLMs performs
next-token-prediction.

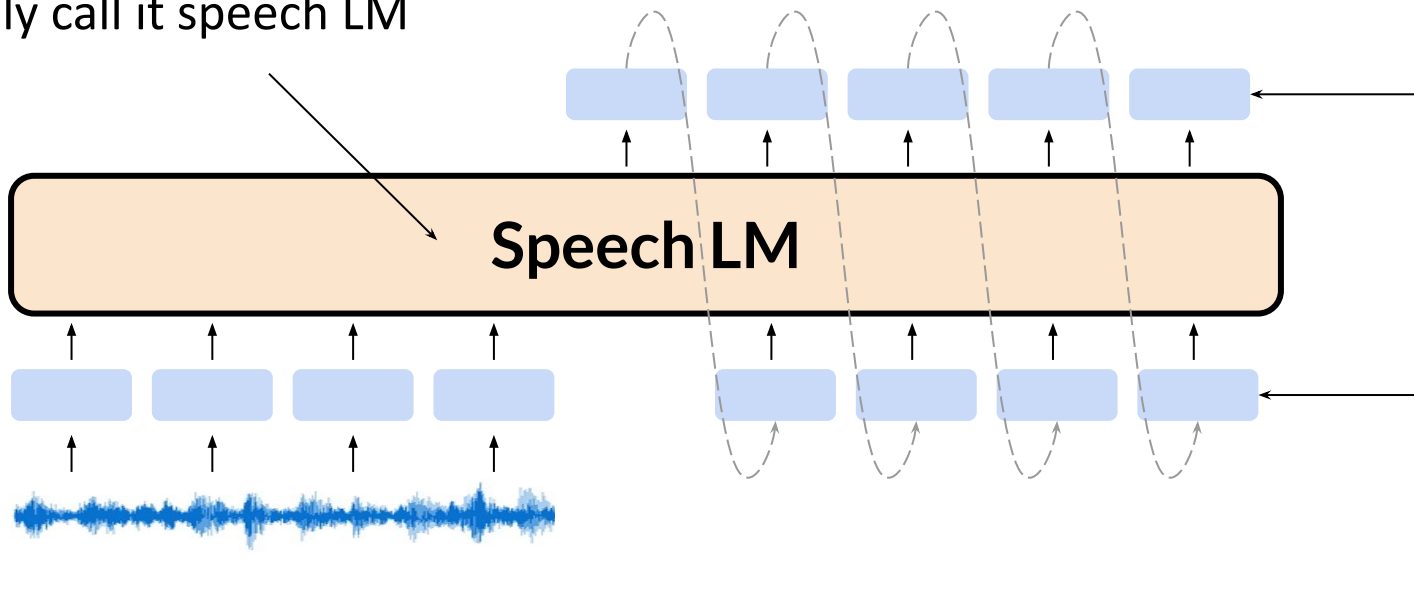
What about in speech processing?



Similar to text LM, there are also some models trained on speech tokens

Discrete speech tokens

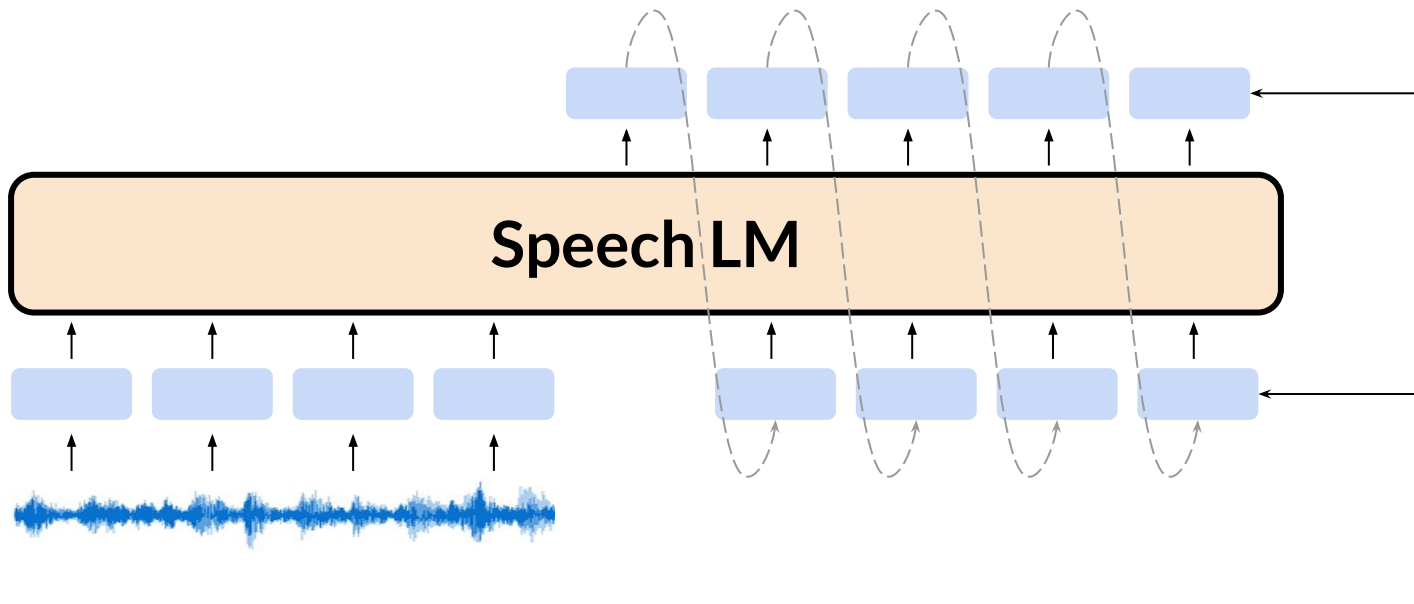
Temporarily call it speech LM



There are different terminologies...

There are various research on speech LM ...

discrete speech tokens

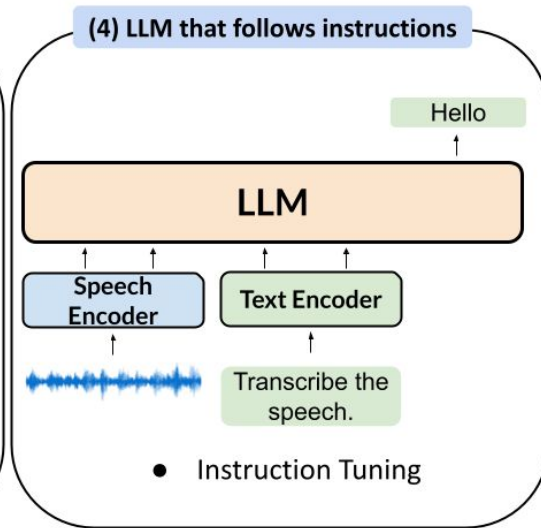
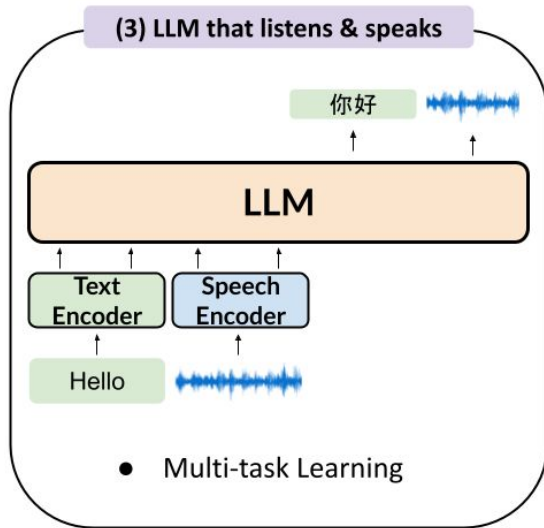
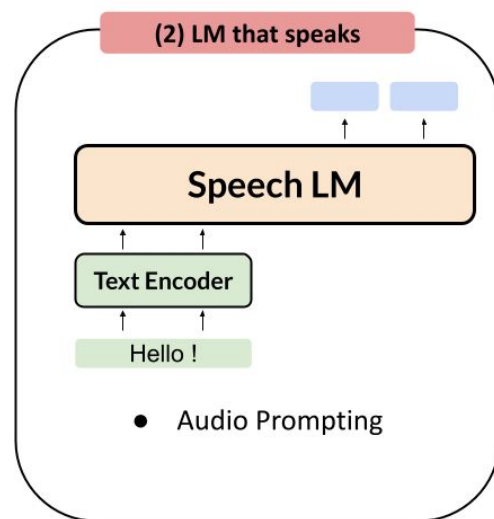
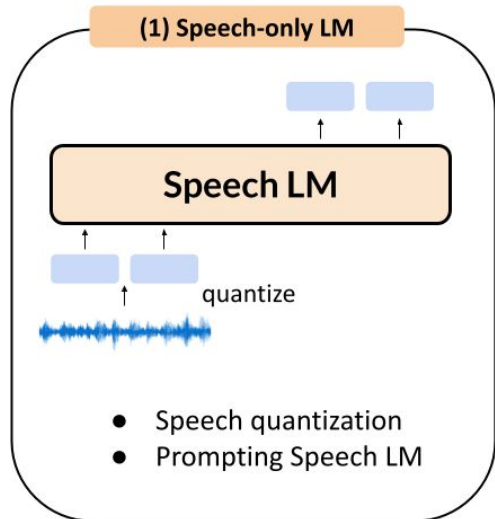


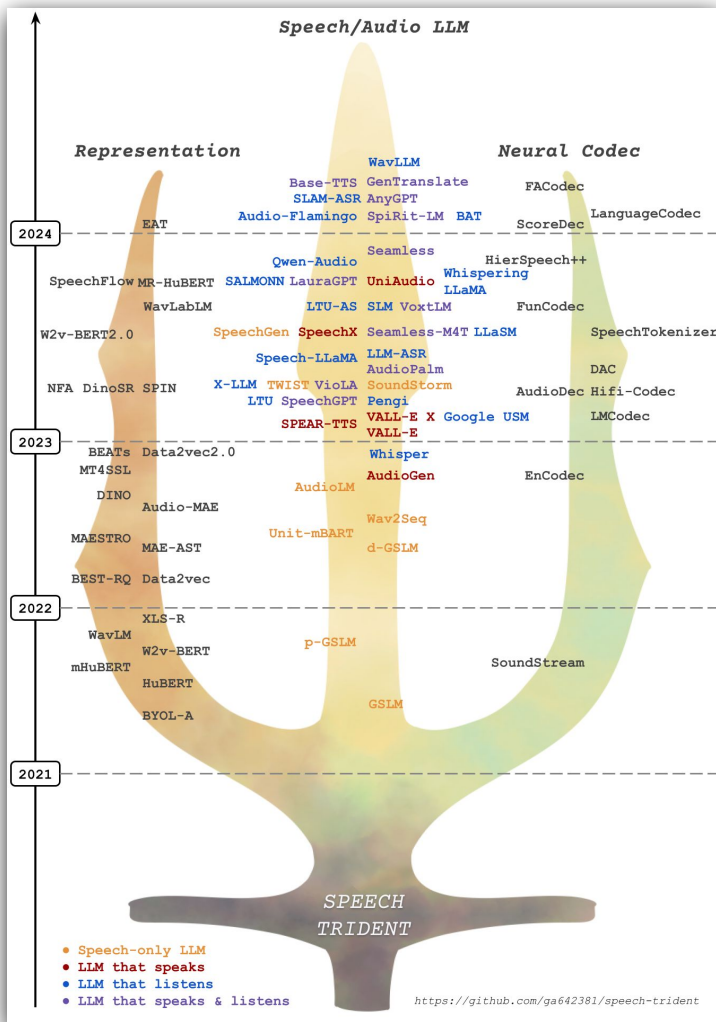
What are these tokens?

What do today's speech LM look like?

What can today's speech LM do?

Outline





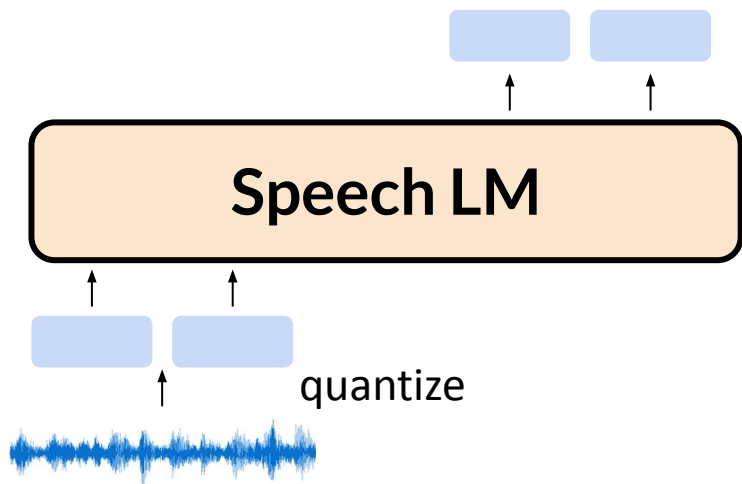
Speech Trident

- Speech / Audio LLMs
- Representation Learning Models
- Neural Codec Models



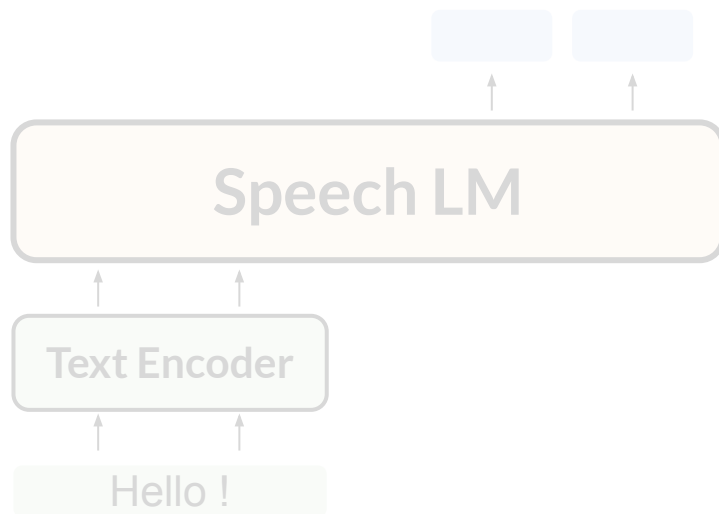
<https://github.com/ga642381/speech-trident>

(1) Speech-only LM



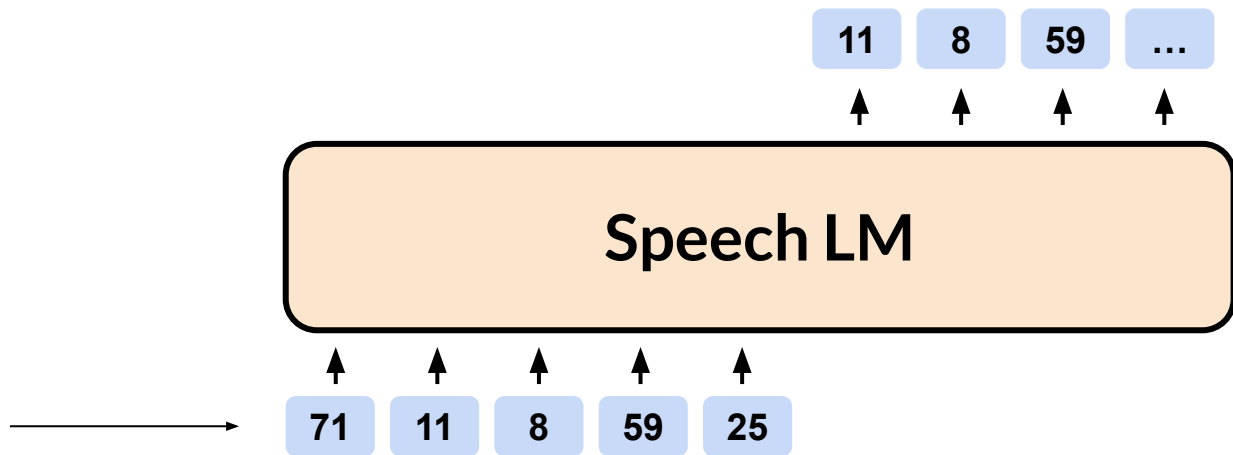
- Speech quantization
- Prompting Speech LM

(2) LM that speaks

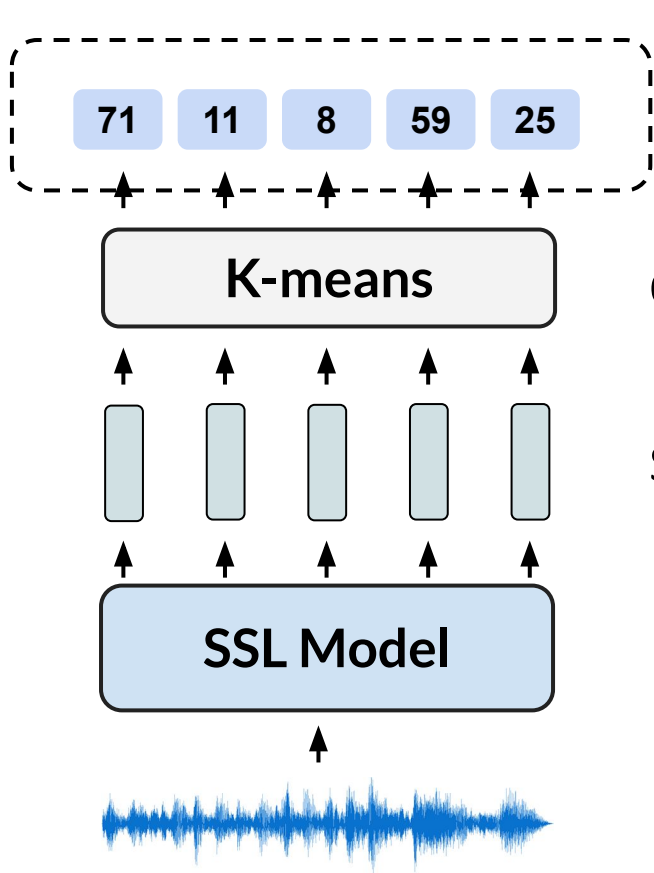


- Audio Prompting

Speech LM trained on
speech tokens



What are these speech tokens?



Discrete speech tokens:

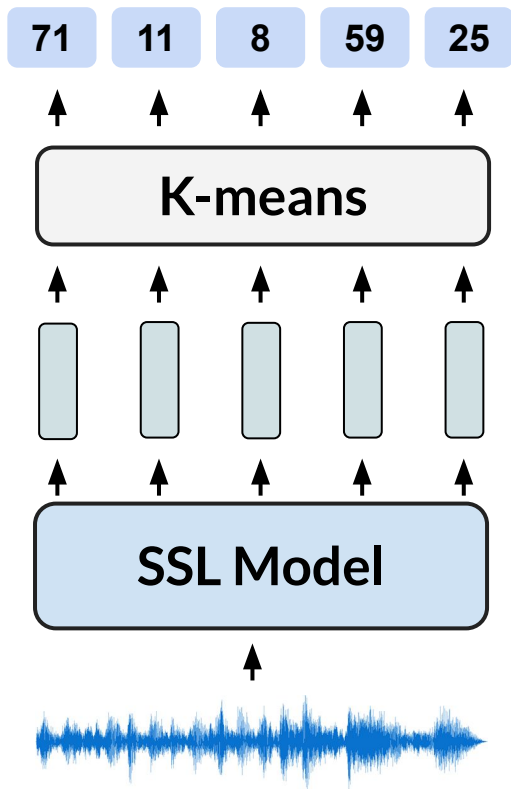
- pseudo text
- semantic tokens

Quantization

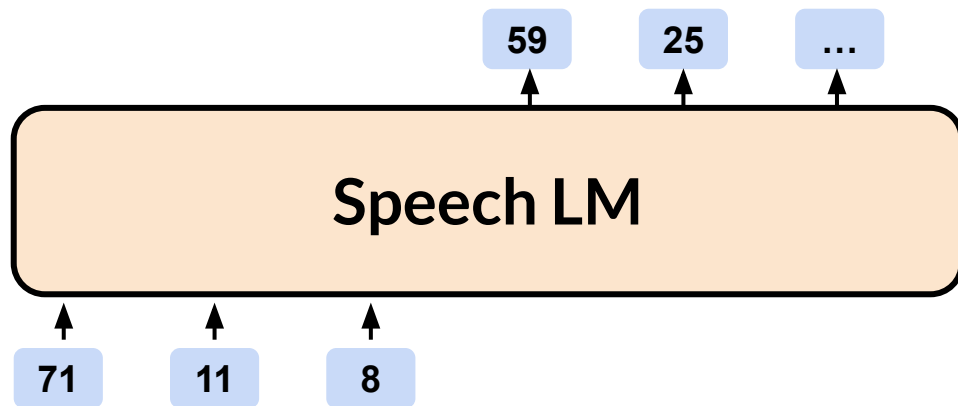
Speech representation

(e.g. HuBERT, w2v-BERT, ...)

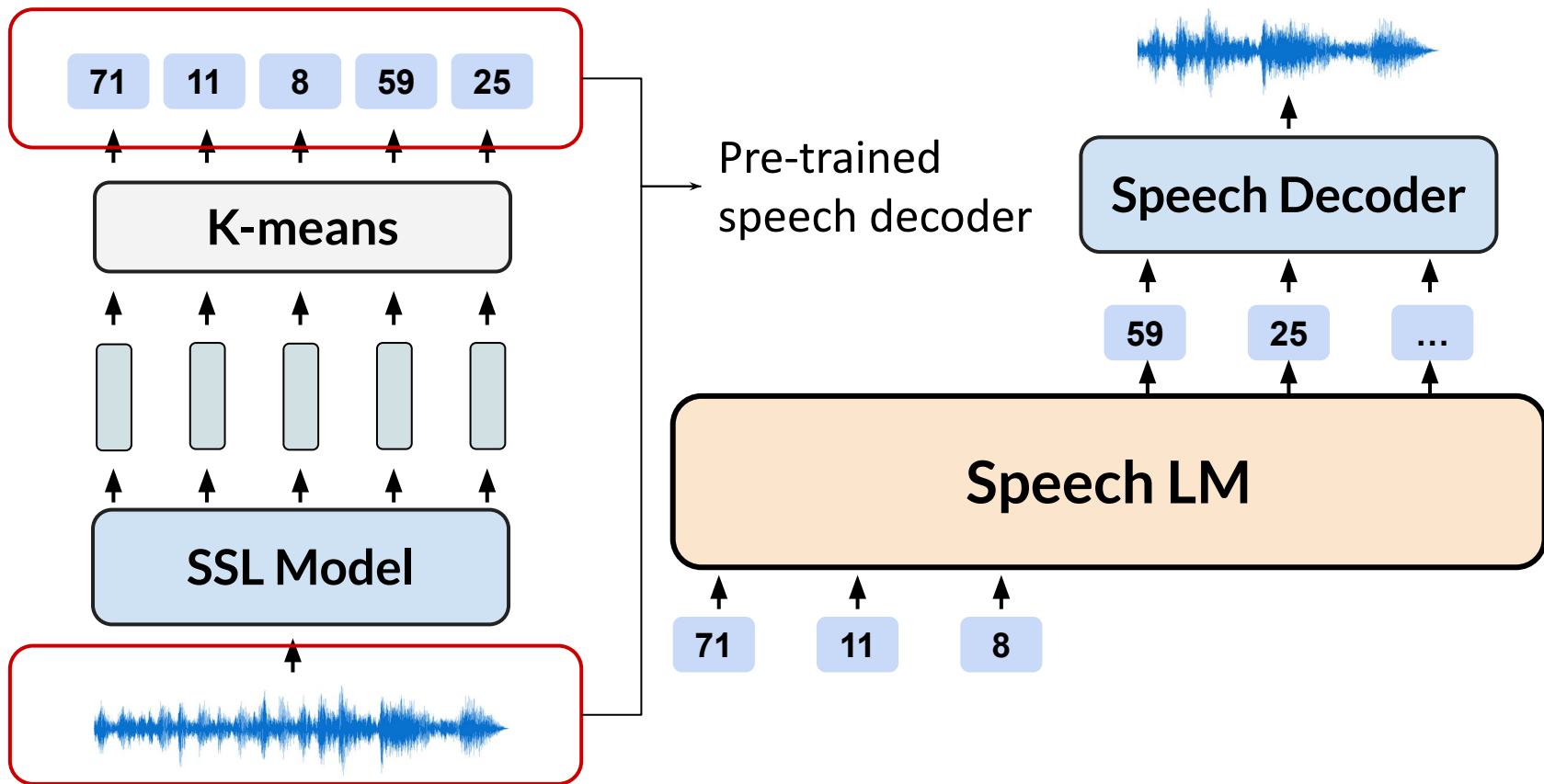
A common approach is to utilize pre-trained SSL representation models



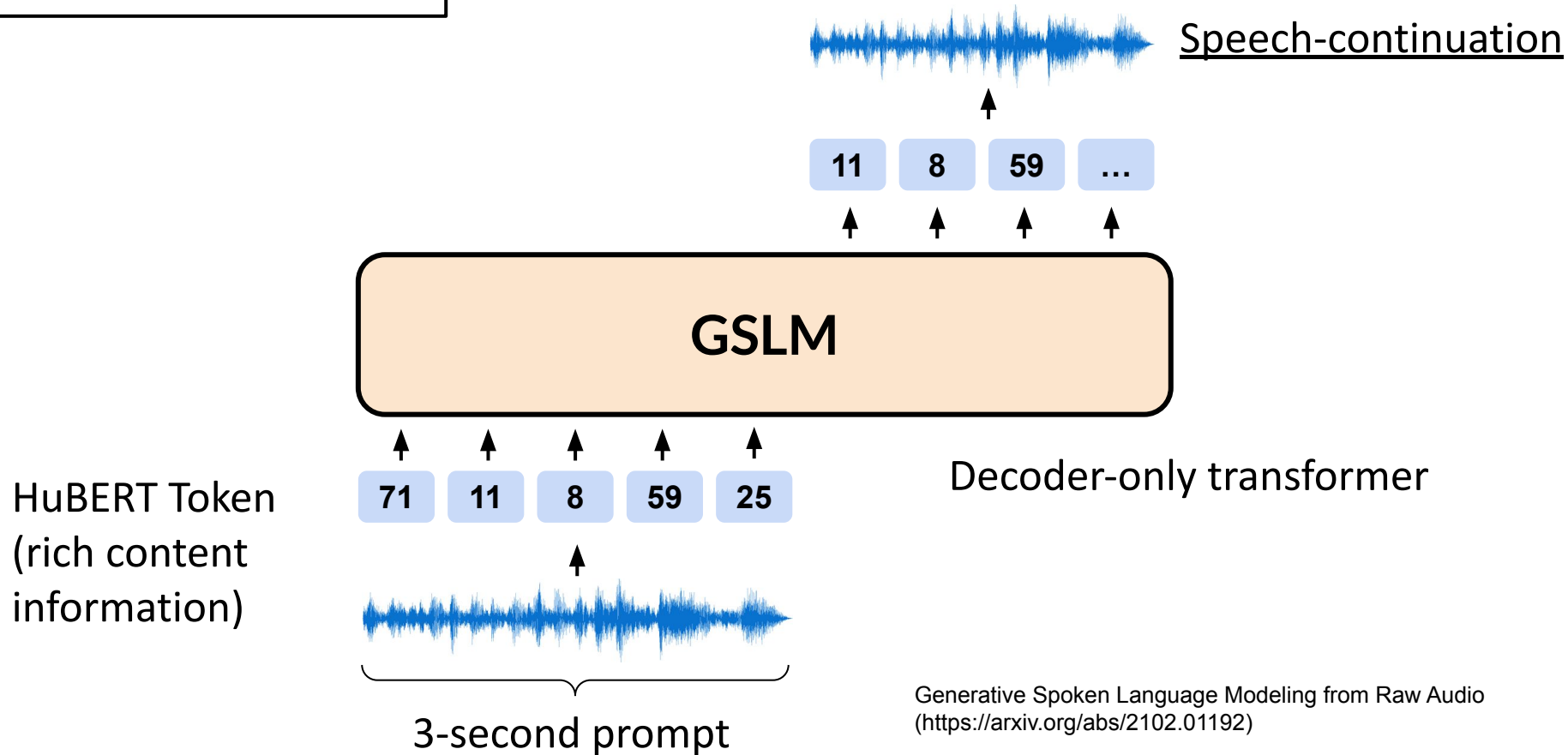
Speech LM performs next-token prediction on the speech tokens autoregressively



The speech tokens can be synthesized back to waveform

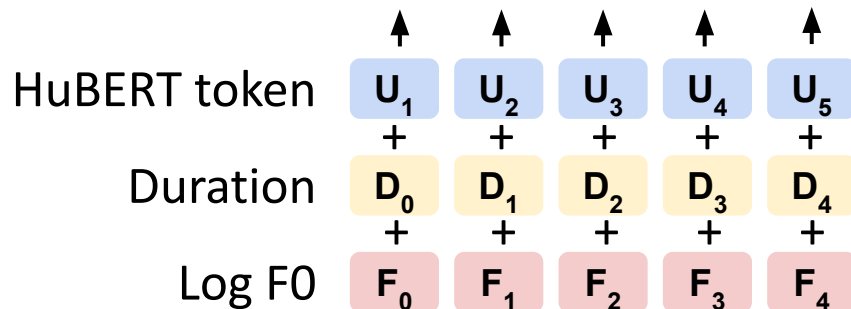
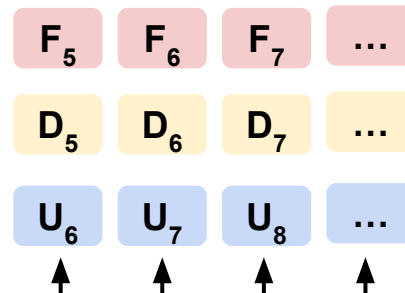


Textless NLP



Textless NLP

There are some variance works of GSLM
e.g. prosody aware GSLM



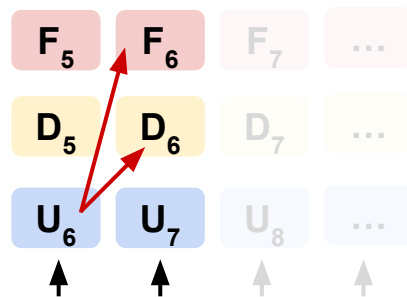
Multi-stream Transformer

Text-Free Prosody-Aware Generative Spoken Language Modeling
(<https://arxiv.org/abs/2109.03264>)

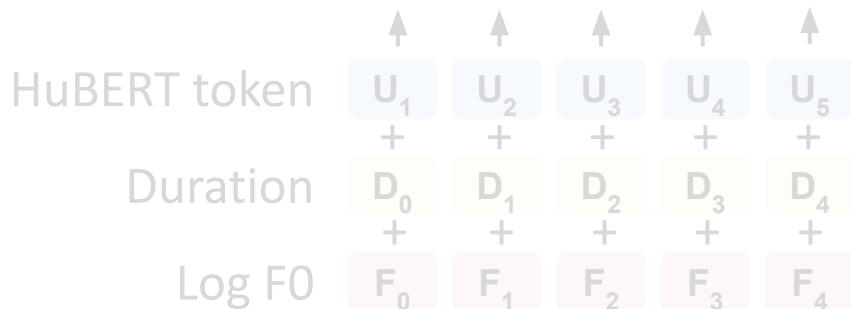
Textless NLP

Delayed prosody prediction

Predicting next-frame prosody conditioned on current HuBERT token



pGSLM

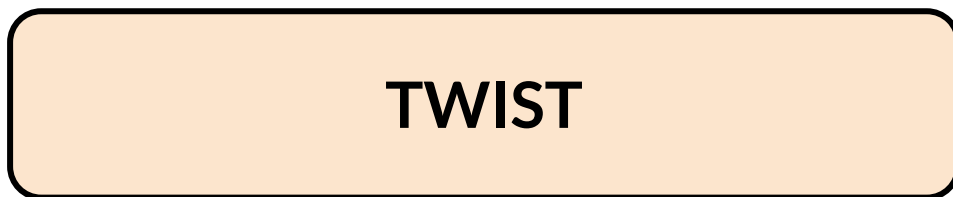
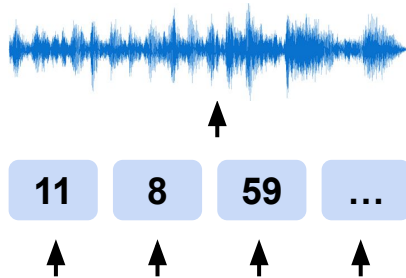


Text-Free Prosody-Aware Generative Spoken Language Modeling
(<https://arxiv.org/abs/2109.03264>)

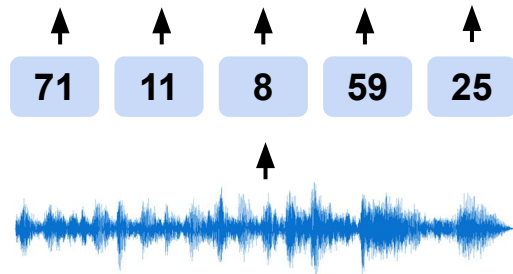
Textless NLP

An improved work of GSLM

Textually **W**arm Initialized **S**peech **T**ransformer

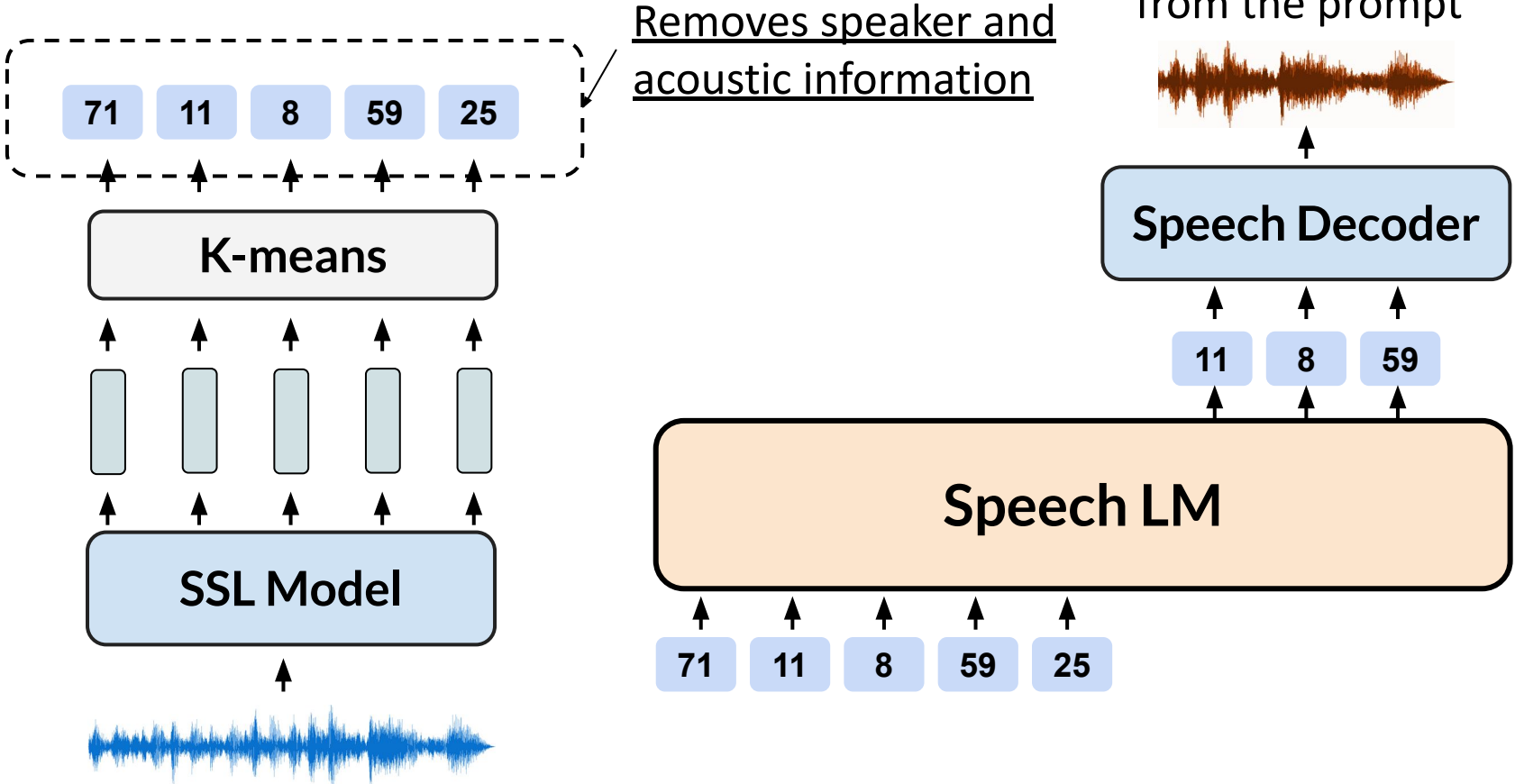


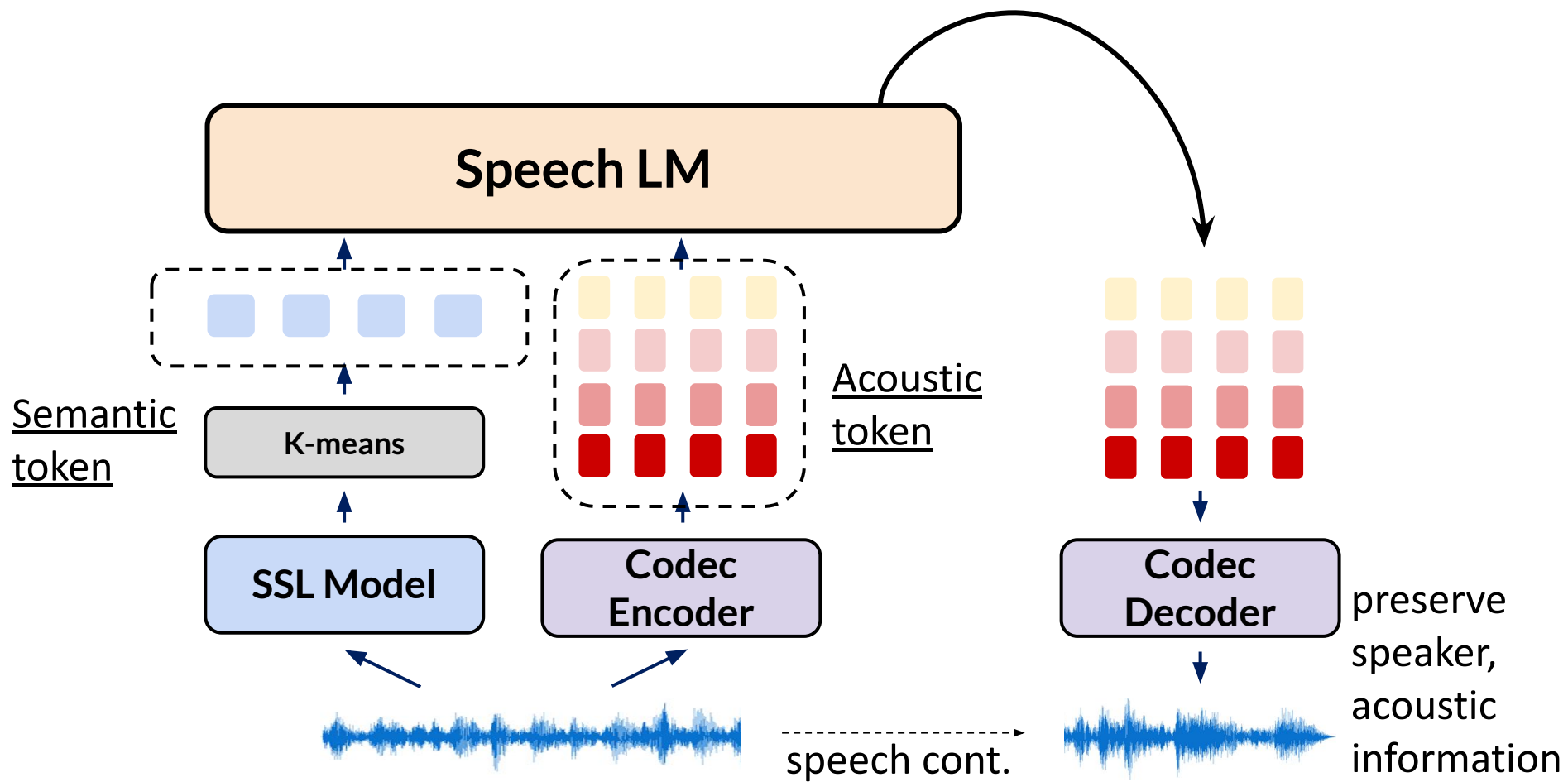
HuBERT Token

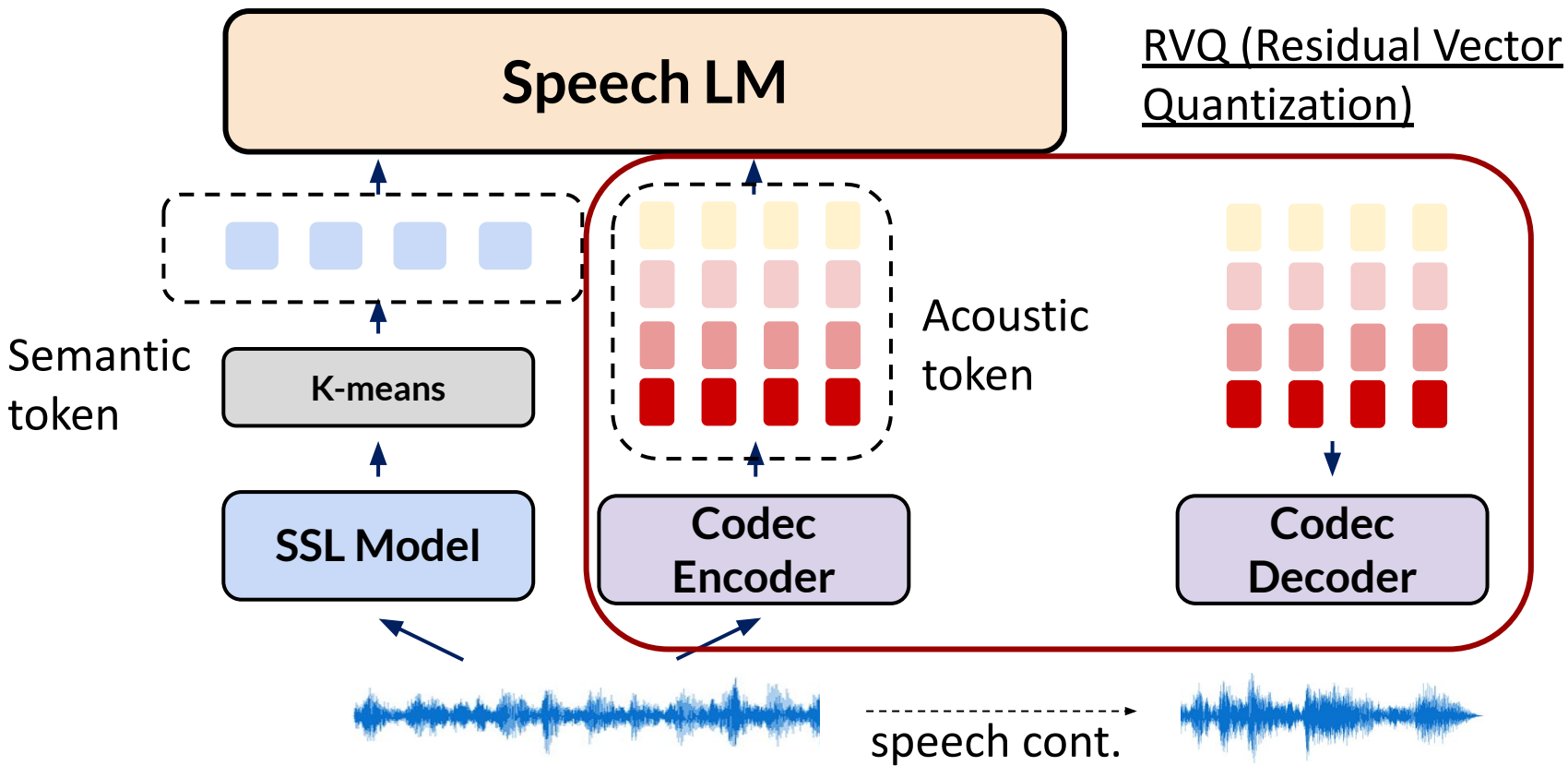


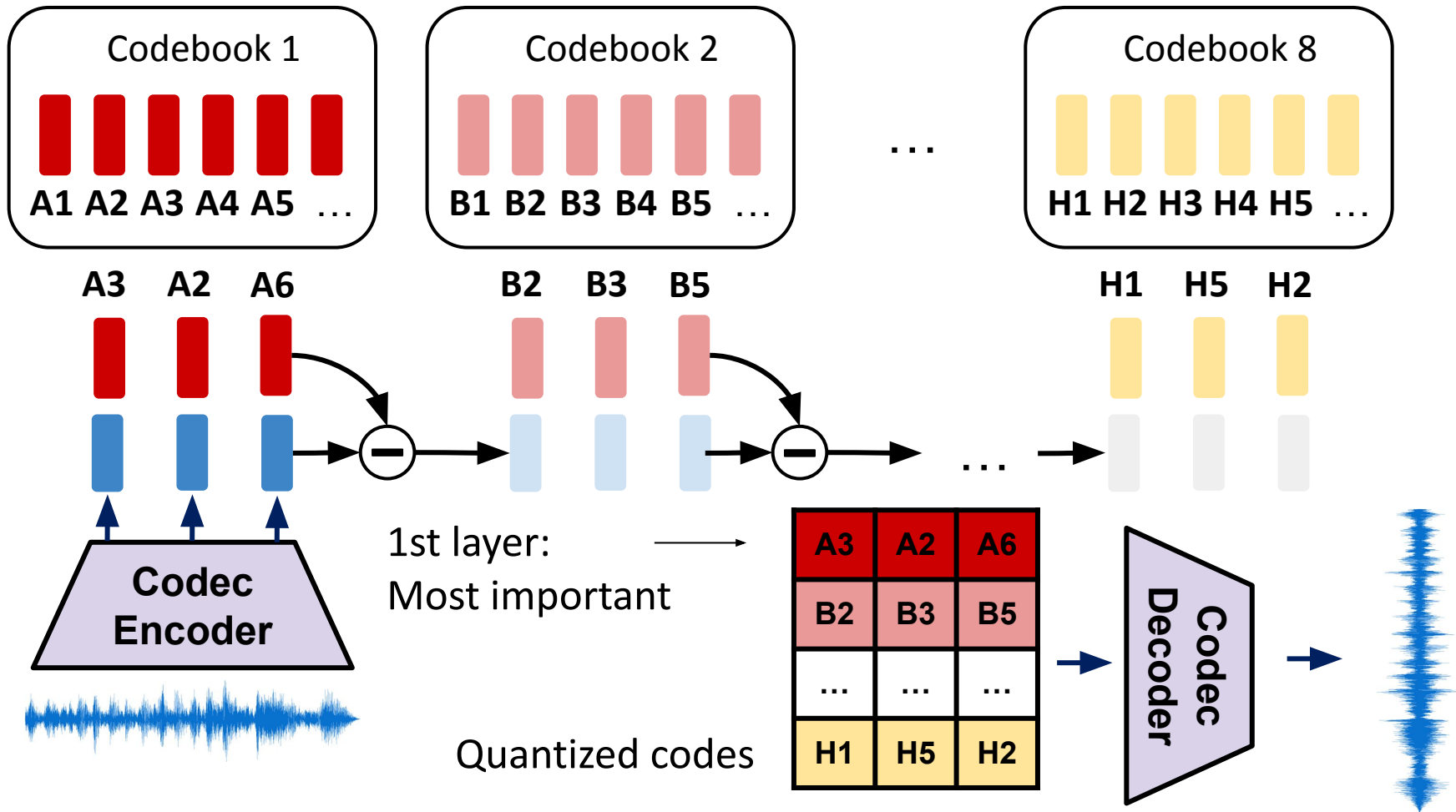
- Decoder-only transformer initialized with pre-trained text LLM (e.g. LLaMA)
- Replace text vocabulary with speech tokens

There's a problem/feature of this framework

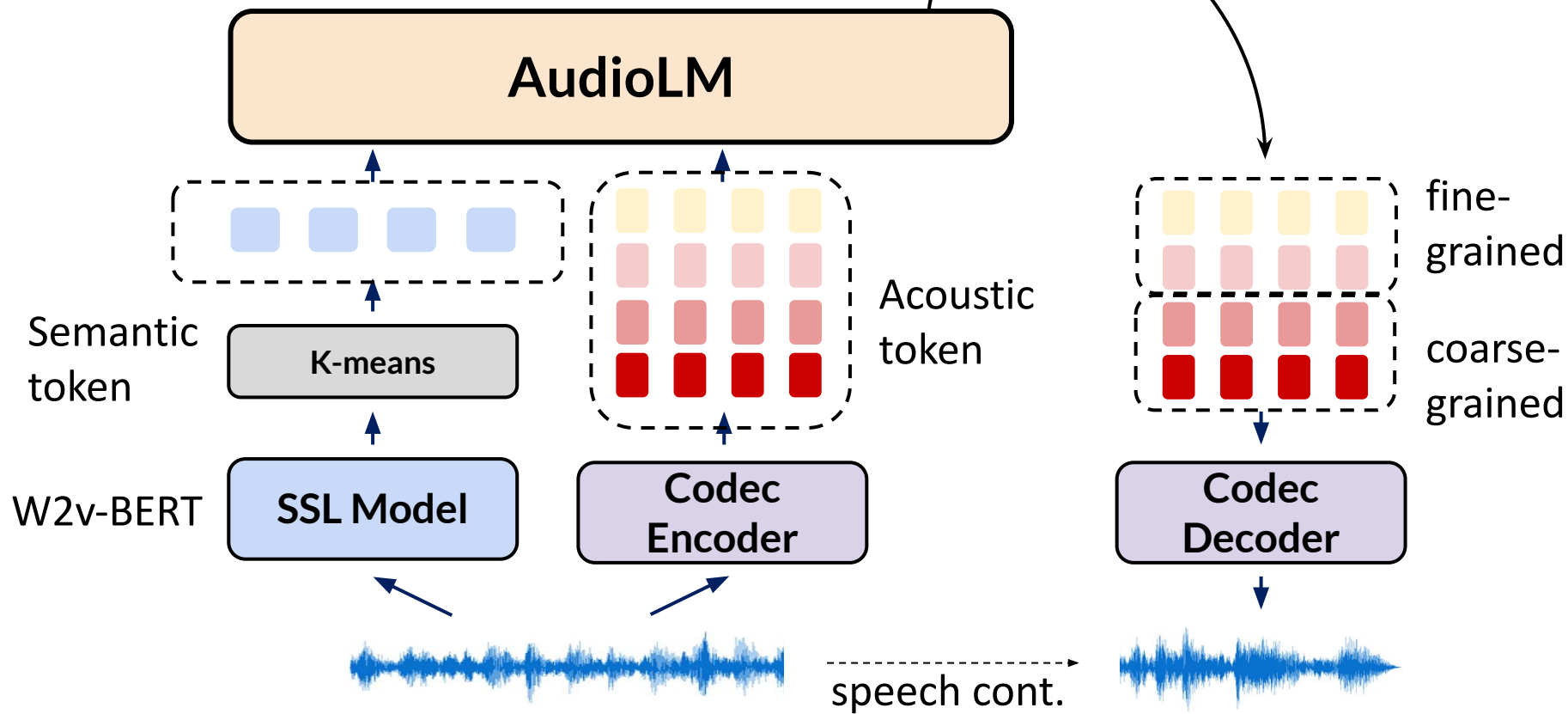




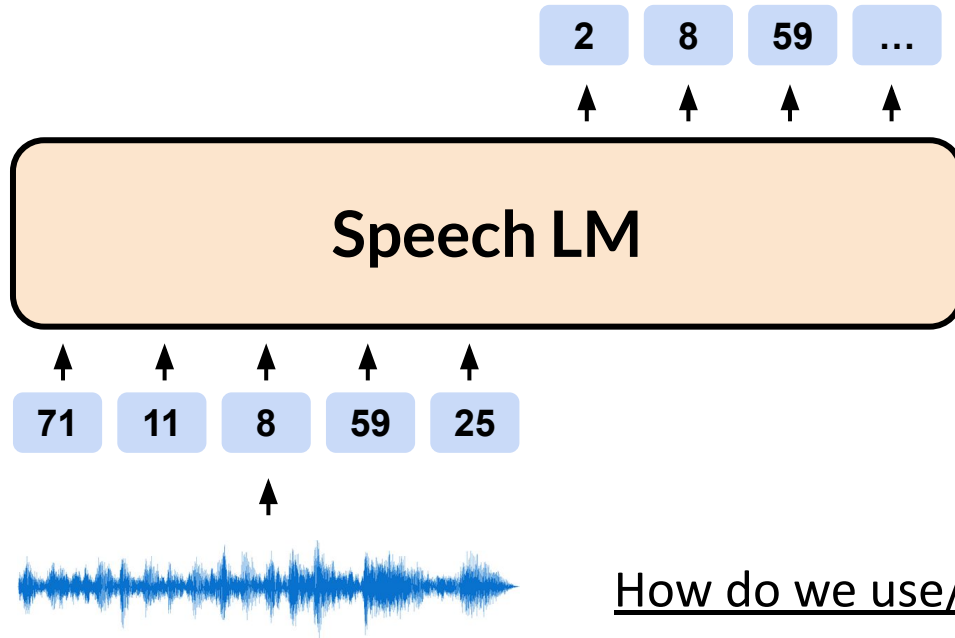




A notable example is Audio LM



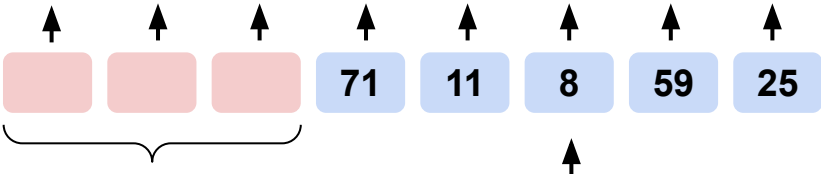
Speech LM can generate tokens with rich linguistic information



How do we use/prompt this model?

SpeechPrompt

2 8 59 ... rich linguistic information



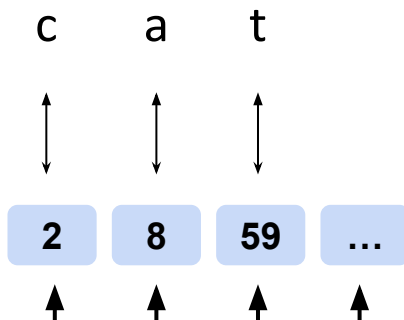
Task Prompts
(trainable vectors)



How do we use/prompt this model?

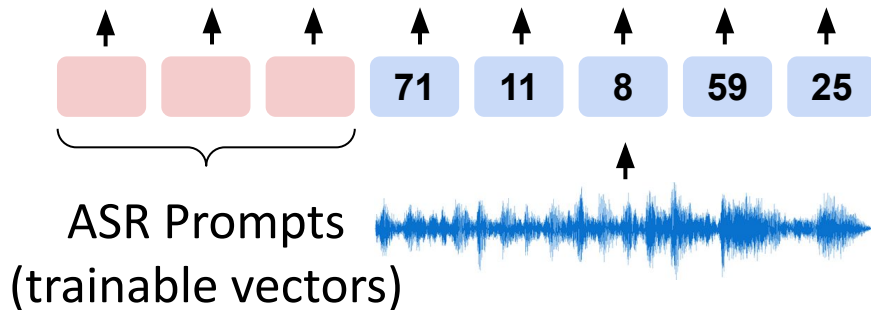
SpeechPrompt

GSLM



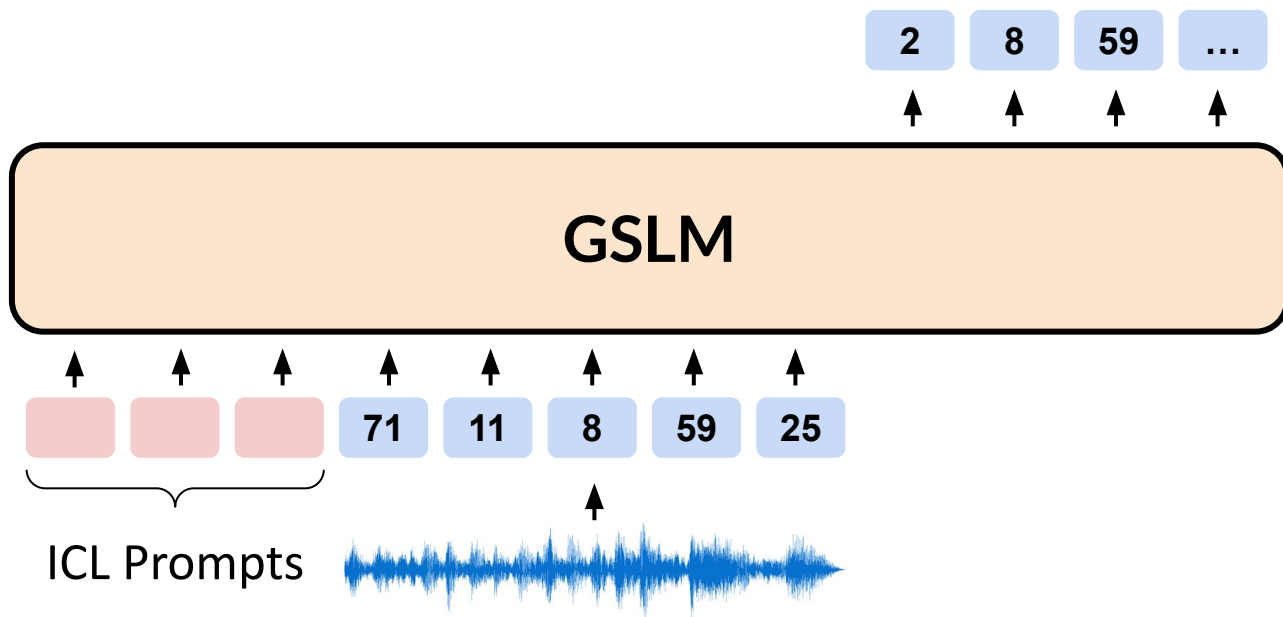
Mapping table

a	8
b	68
c	2
...	...
t	59
...	...



This work was already covered in last year's tutorial

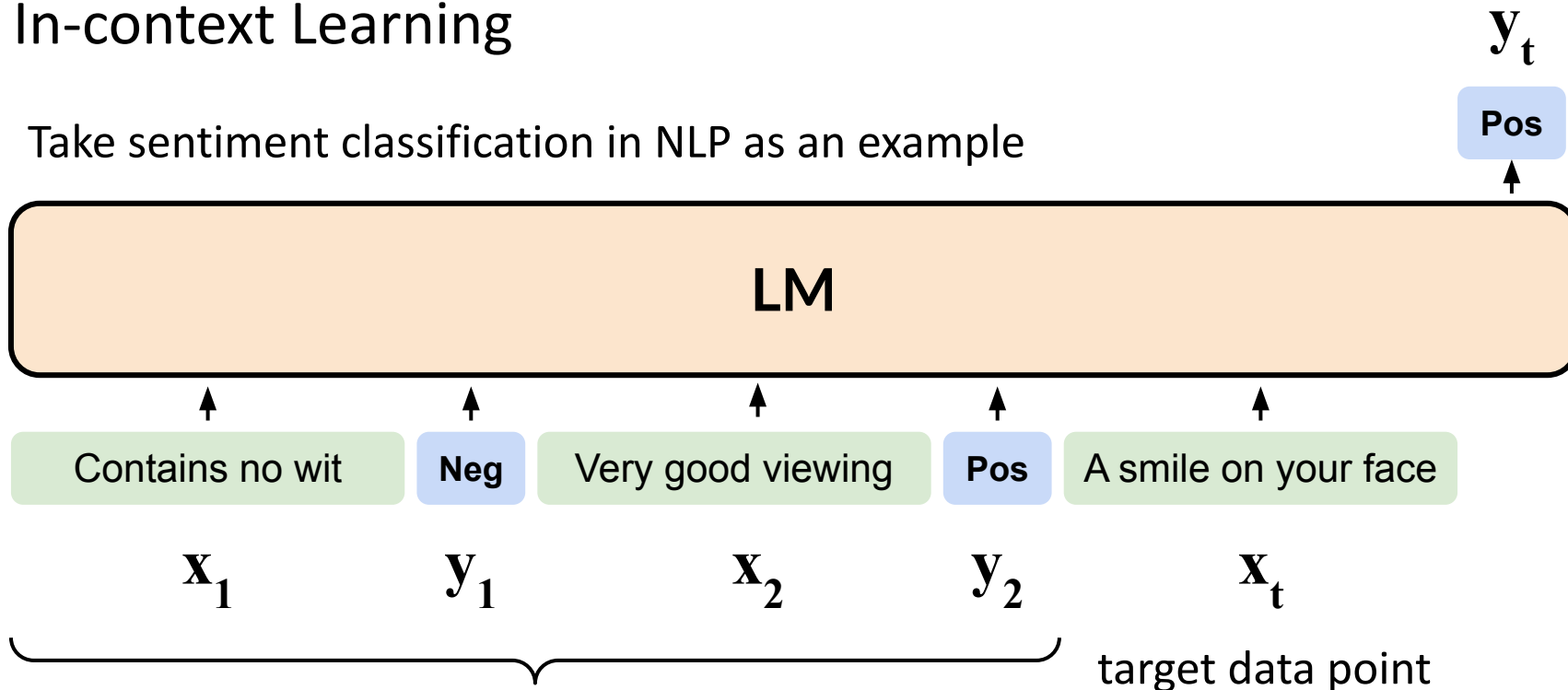
How far can we get with SpeechPrompt?



Prompts guiding speech LM perform In-context learning

In-context Learning

Take sentiment classification in NLP as an example

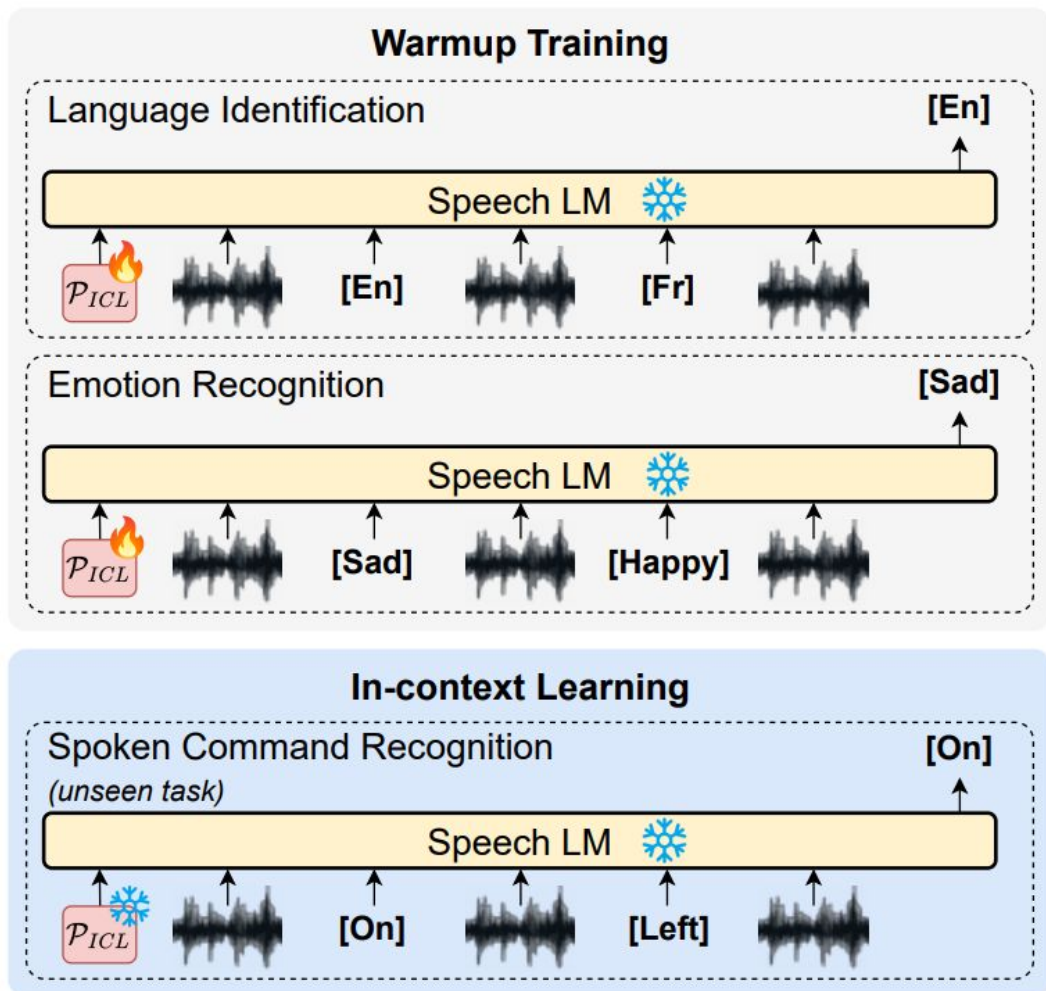
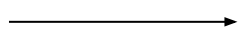


demonstrations concatenated at the beginning

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

The original GSLM does not have the ability to directly predict the label of \mathbf{x}_t

The LM is fixed
The prompt is fixed
The task is unseen



random guessing baseline

linear classifier

Speech classification task

Metric: accuracy

Task Type	Task	Dataset	ICL	Random	SVC
Unseen Task	SD	MUStARD	64.1	54.7	60.9
	SCR	Google SC	48.0	25.1	43.8
	SCR	Arabic SC	36.5	28.0	50.8

Speech classification task
Metric: accuracy

random guessing baseline

linear classifier

Task Type	Task	Dataset	ICL	Random	SVC
Unseen Task	SD	MUStARD	64.1	54.7	60.9
	SCR	Google SC	48.0	25.1	43.8
	SCR	Arabic SC	36.5	28.0	50.8

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

Speech classification task
Metric: accuracy

random guessing baseline

linear classifier

Task Type	Task	Dataset	ICL	Random	SVC
Unseen Task	SD	MUStARD	64.1	54.7	60.9
	SCR	Google SC	48.0	25.1	43.8
	SCR	Arabic SC	36.5	28.0	50.8

- Underperform SVC probably due to cross lingual setting

random guessing baseline

linear classifier

Speech classification task

Metric: accuracy

Task Type	Task	Dataset	ICL	Random	SVC
Unseen Task	SD	MUSARD	64.1	54.7	60.9
	SCR	Google SC	48.0	25.1	43.8
	SCR	Arabic SC	36.5	28.0	50.8

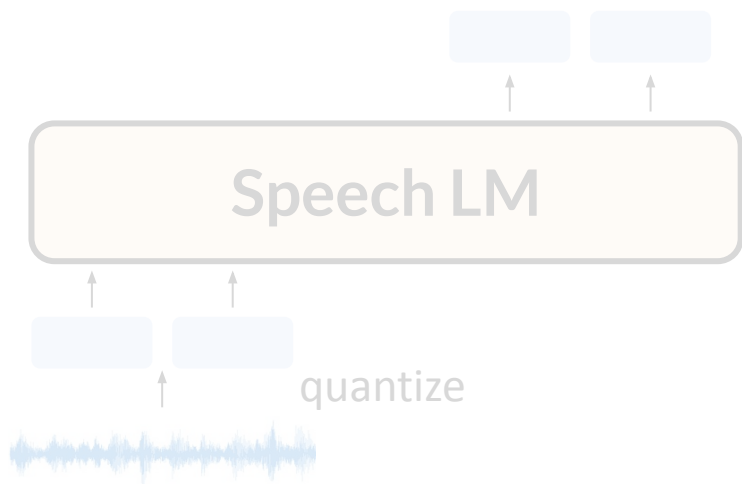
There's still a big performance gap between the simple supervised models.

Surprising to get a non trivial result. ICL as an emergent ability:

GPT-3 ~ 170 B parameters

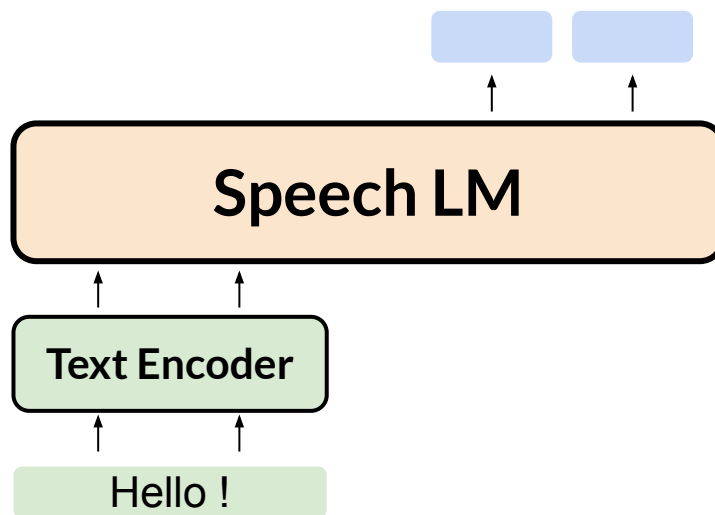
GSLM ~ 150 M parameters

(1) Speech-only LLM



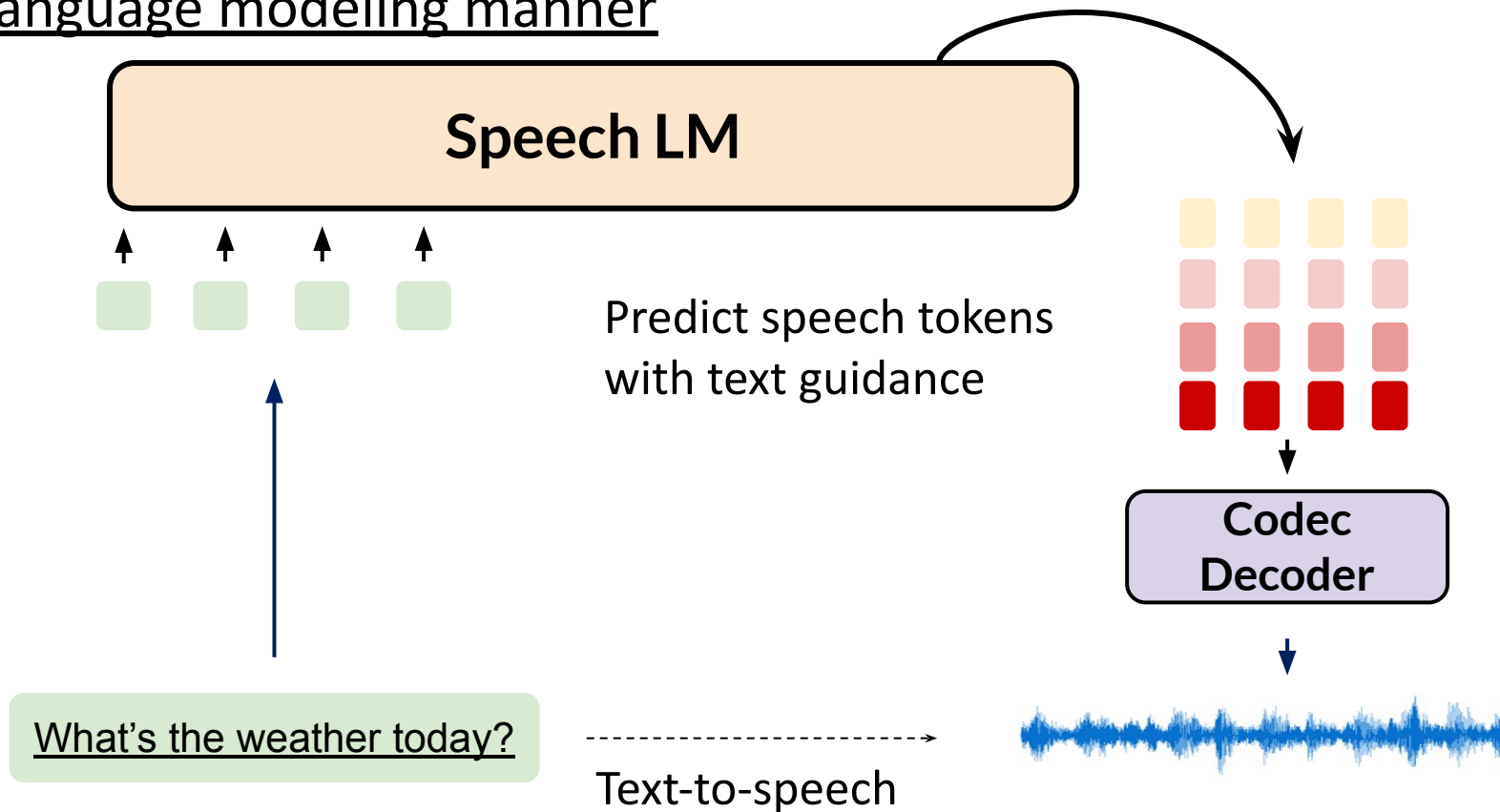
- Speech quantization
- Prompting Speech LM

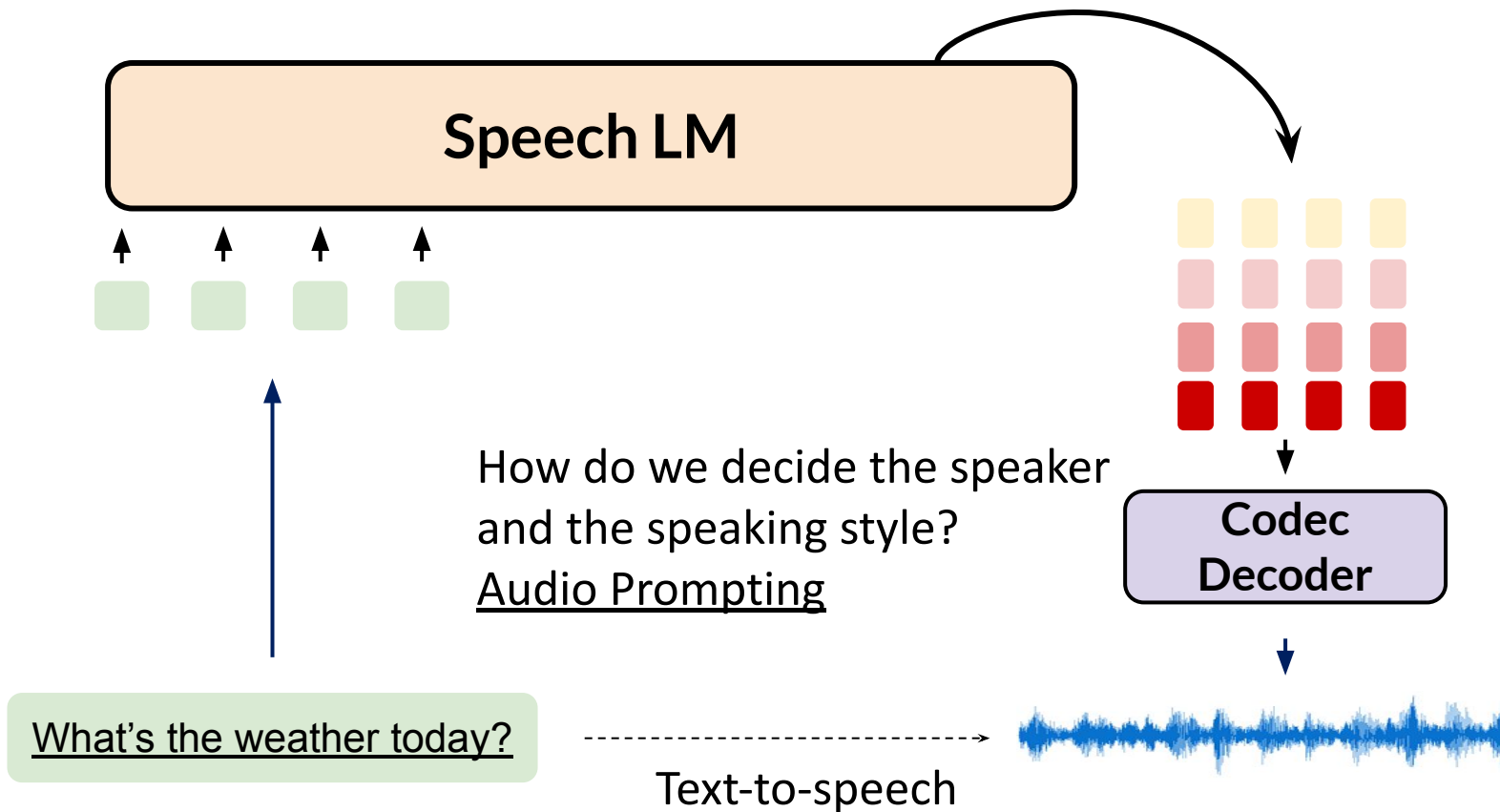
(2) LM that speaks

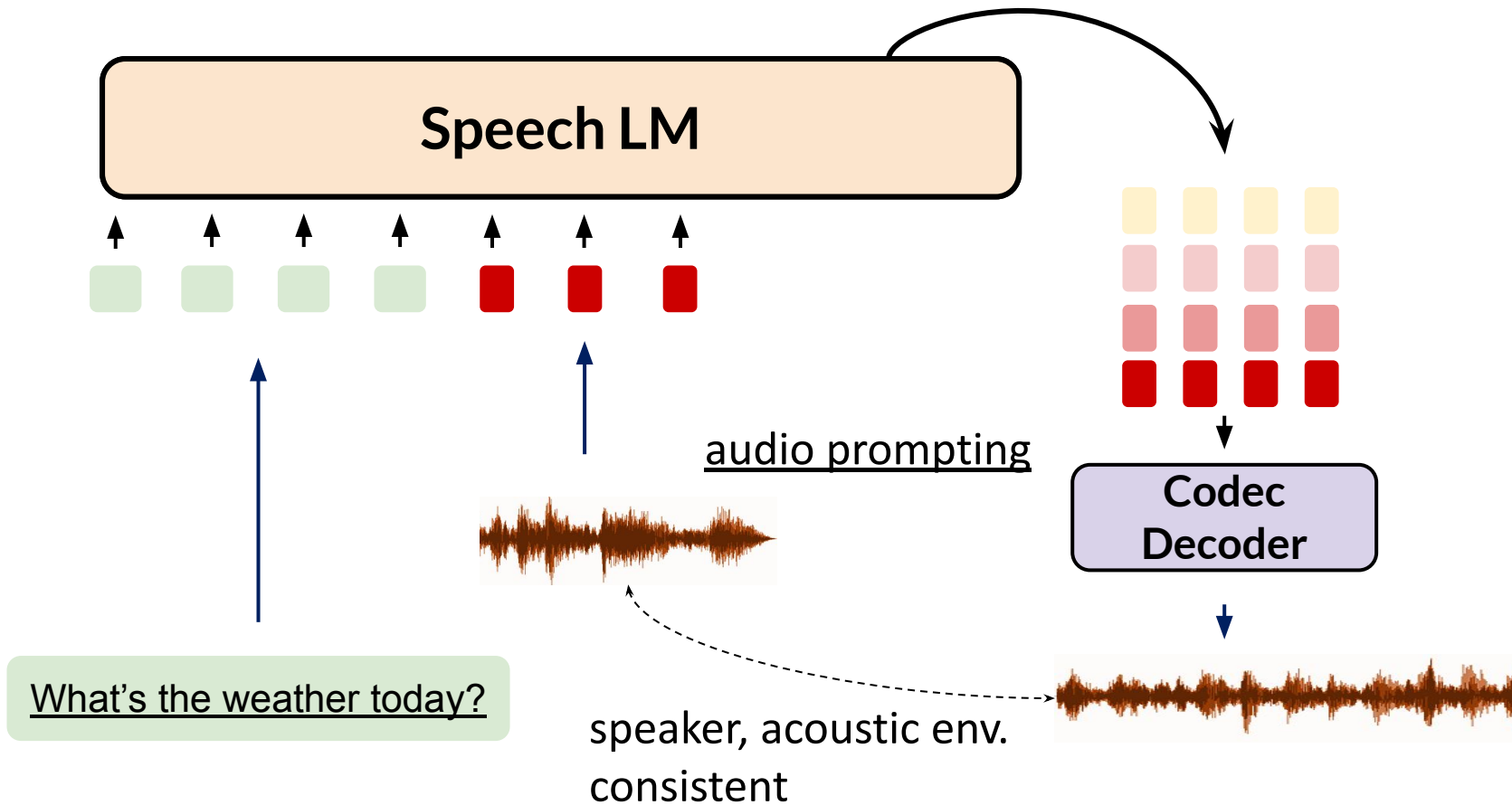


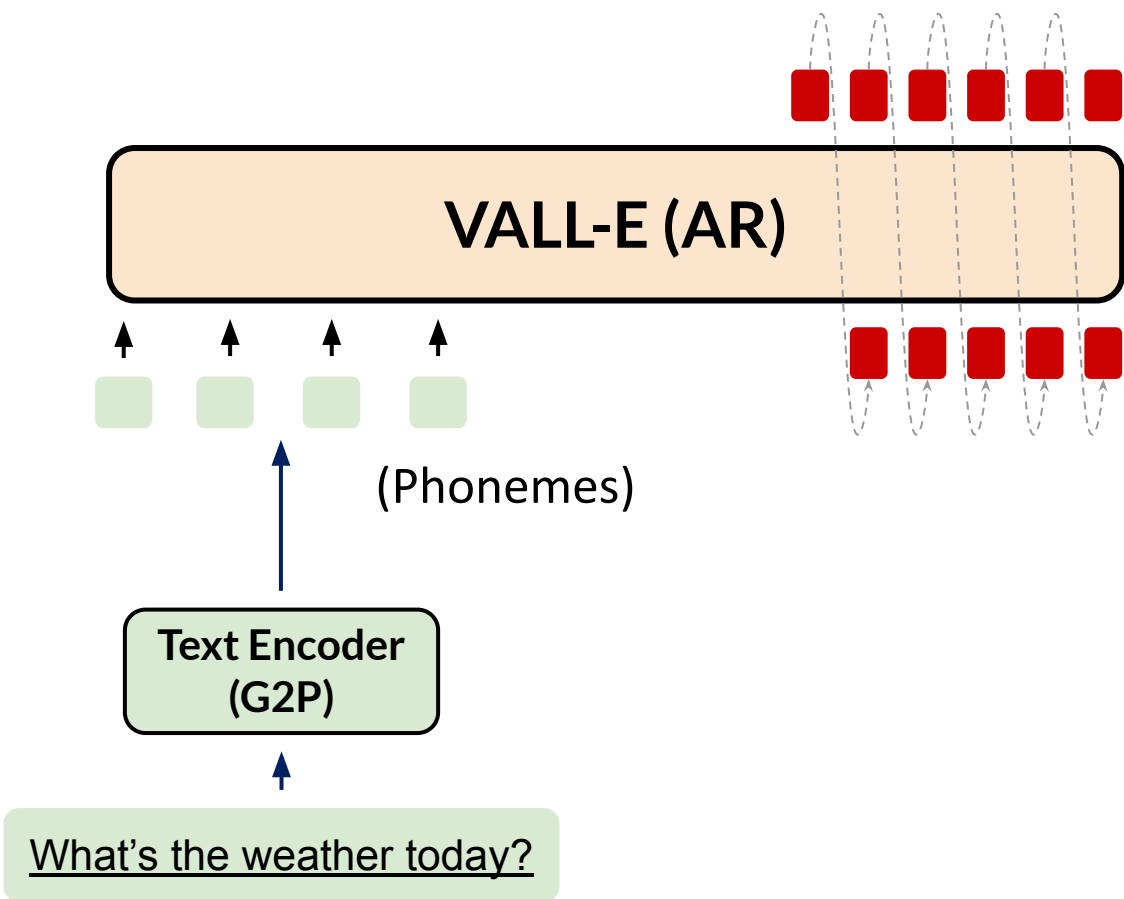
- Audio Prompting

Speech Synthesis in a language modeling manner



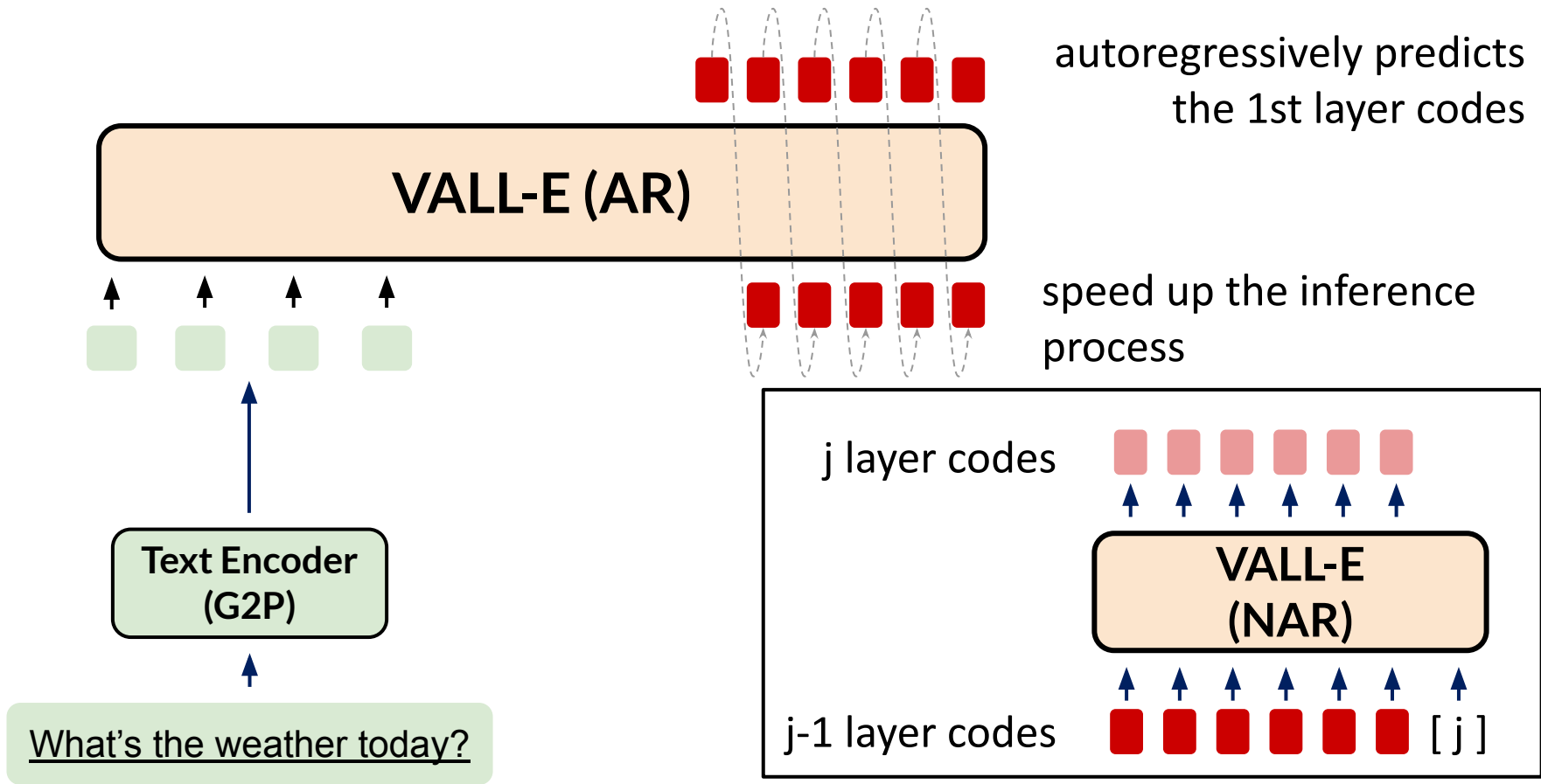


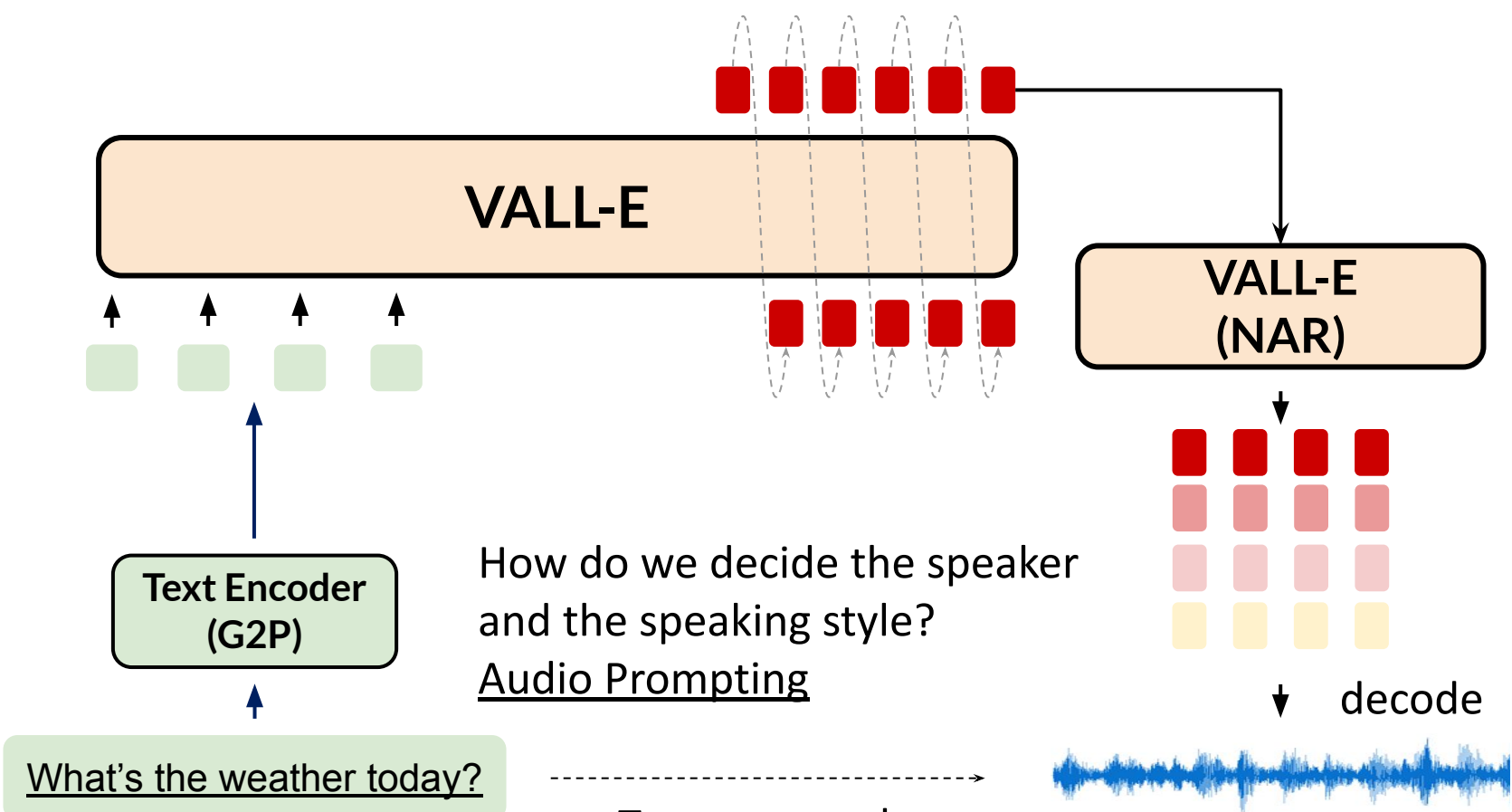




autoregressively predicts
the 1st layer codes

What's the weather today?



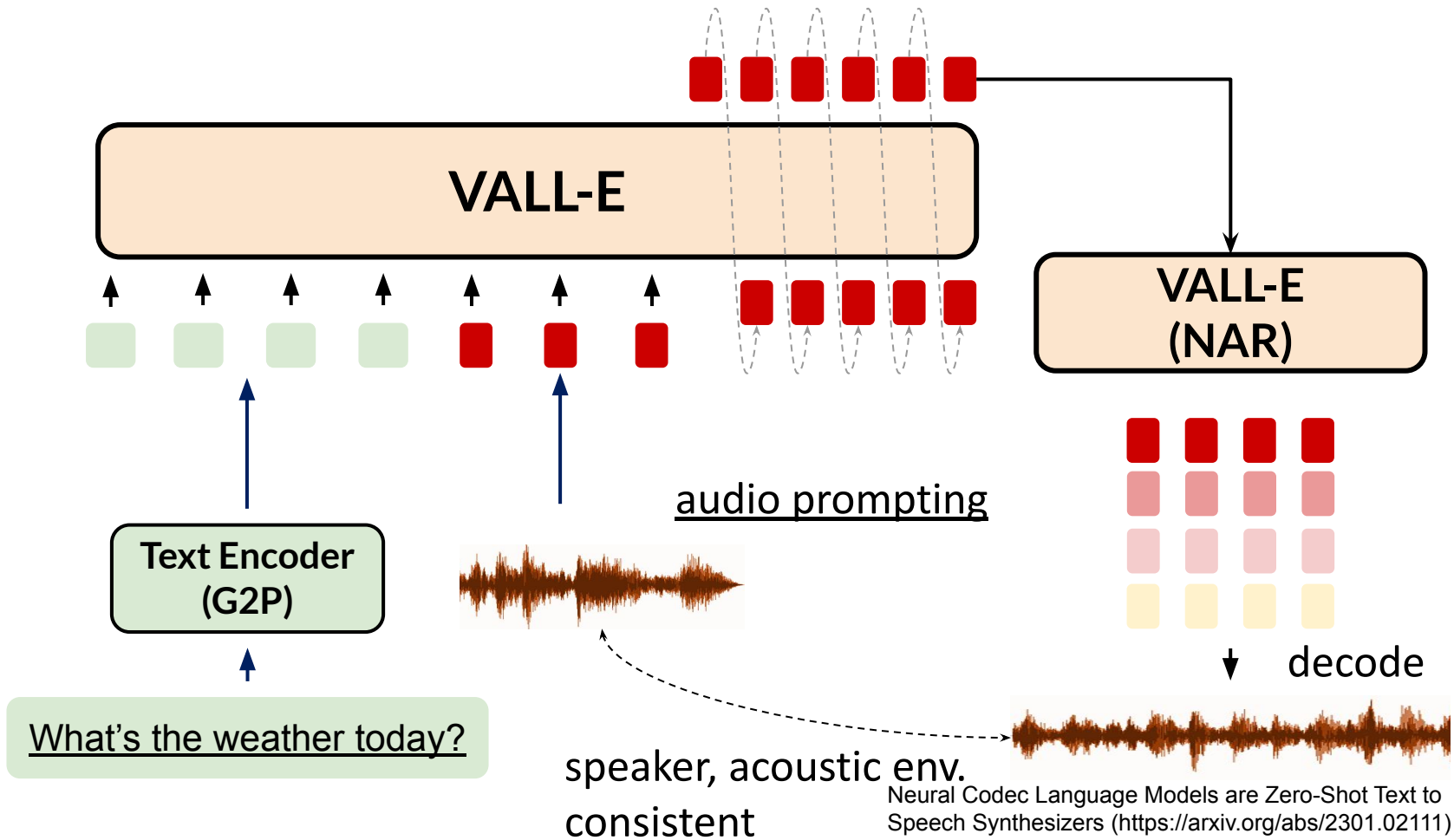


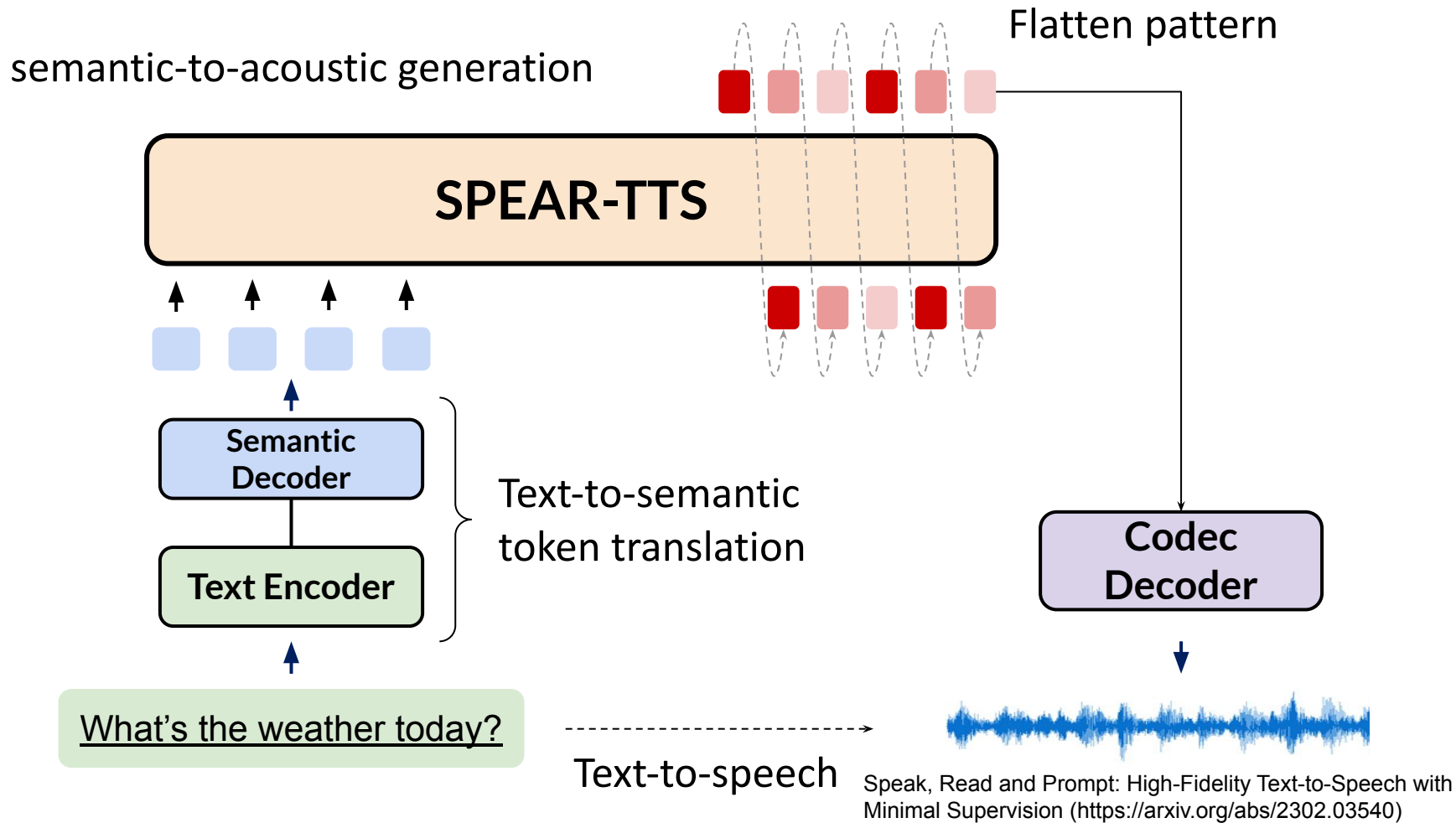
How do we decide the speaker and the speaking style?

Audio Prompting

Text-to-speech

Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers (<https://arxiv.org/abs/2301.02111>)





Model	Parallel data	Prediction Process
SPEAR-TTS	15 min	(1) Text-to-semantic translation (2) Semantic-to-acoustic generation
VALL-E	60,000 hours	Direct acoustic tokens generation

Model	Parallel data	Prediction Process
SPEAR-TTS	15 min	(1) Text-to-semantic translation (2) Semantic-to-acoustic generation
VALL-E	60,000 hours	Direct acoustic tokens generation

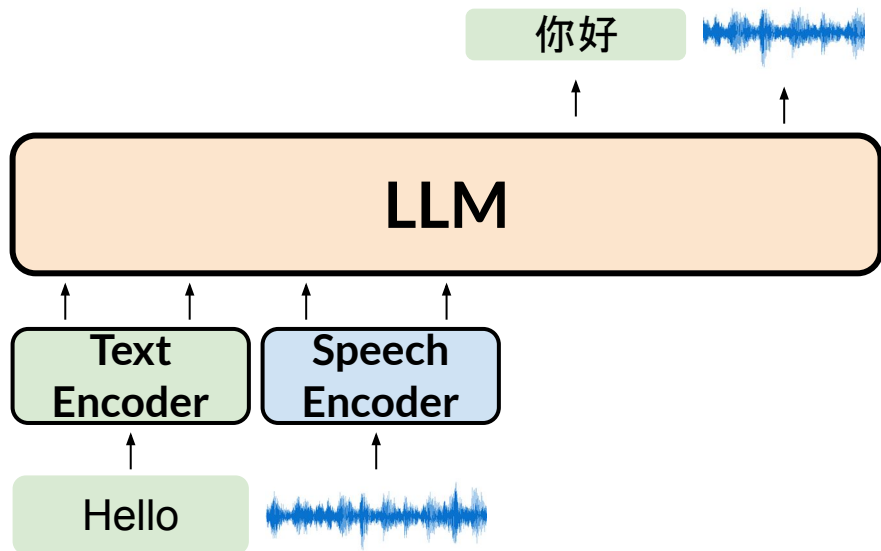
System	VALL-E	SPEAR-TTS (15 min)
MOS	3.35 \pm 0.12	4.75 \pm 0.06

With semantic tokens, SPEAR-TTS can achieve better quality with less parallel data

Model	Parallel training data	Cosine similarity
YourTTS	~ 600 h	0.34
VALL-E	60,000 h	0.58
SPEAR-TTS	15 min	0.56

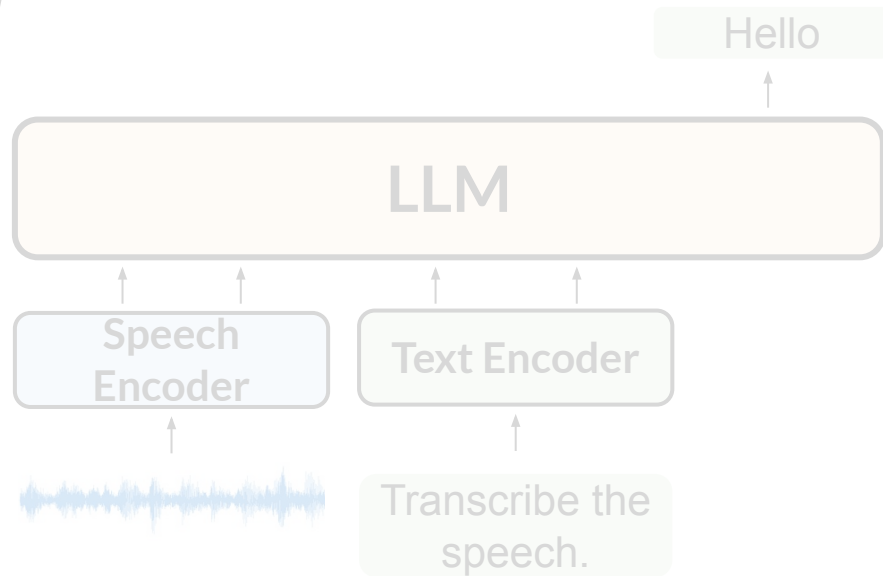
Speaker similarity between prompt and generated speech

(3) LLM that listens & speaks

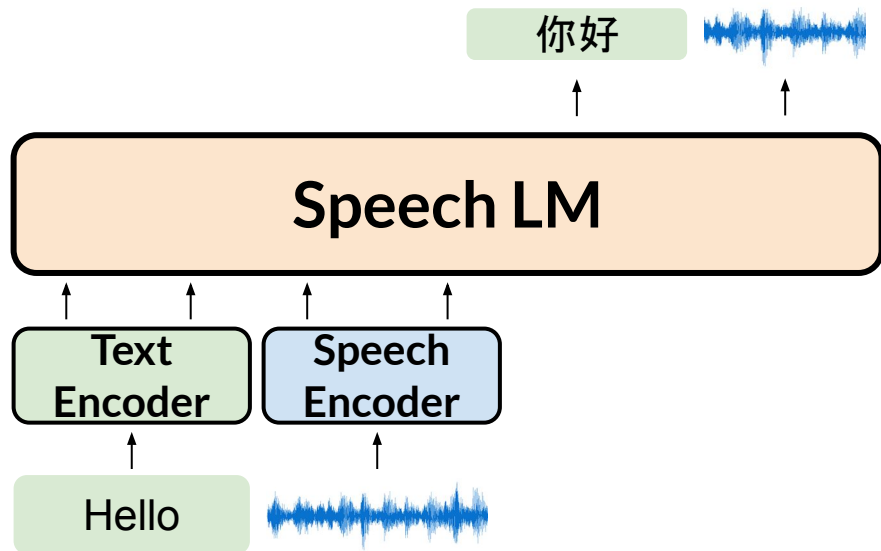


- Multi-task Learning

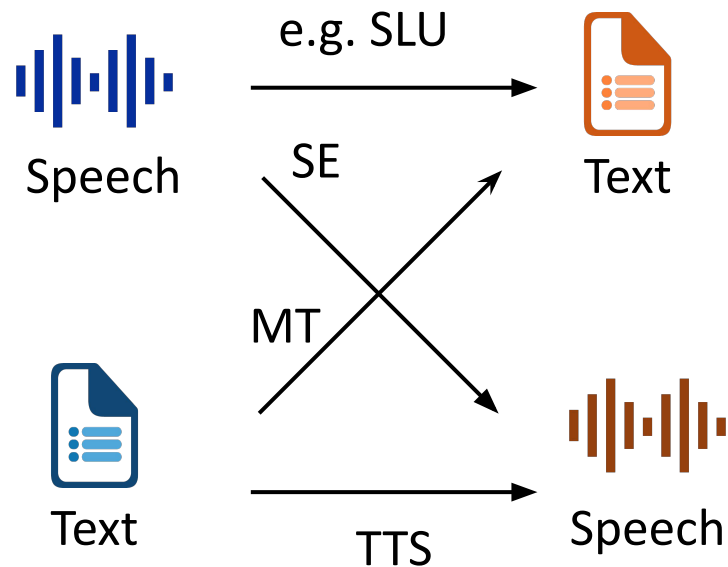
(4) LLM that follows instructions

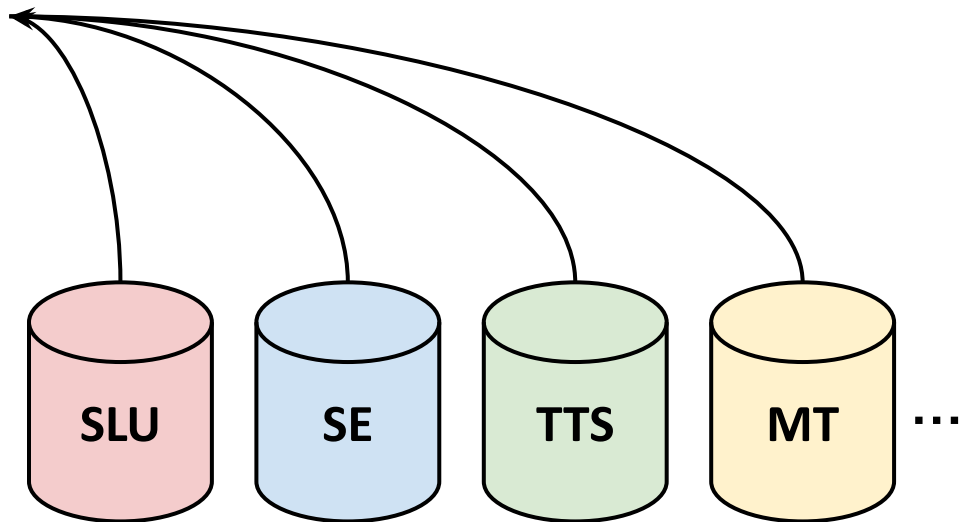
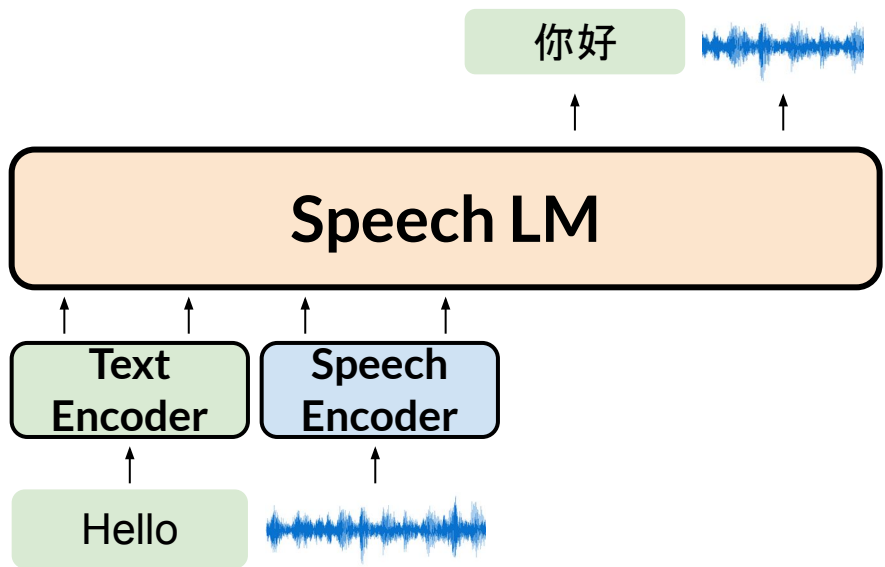


- Instruction Tuning



Multi-task learning

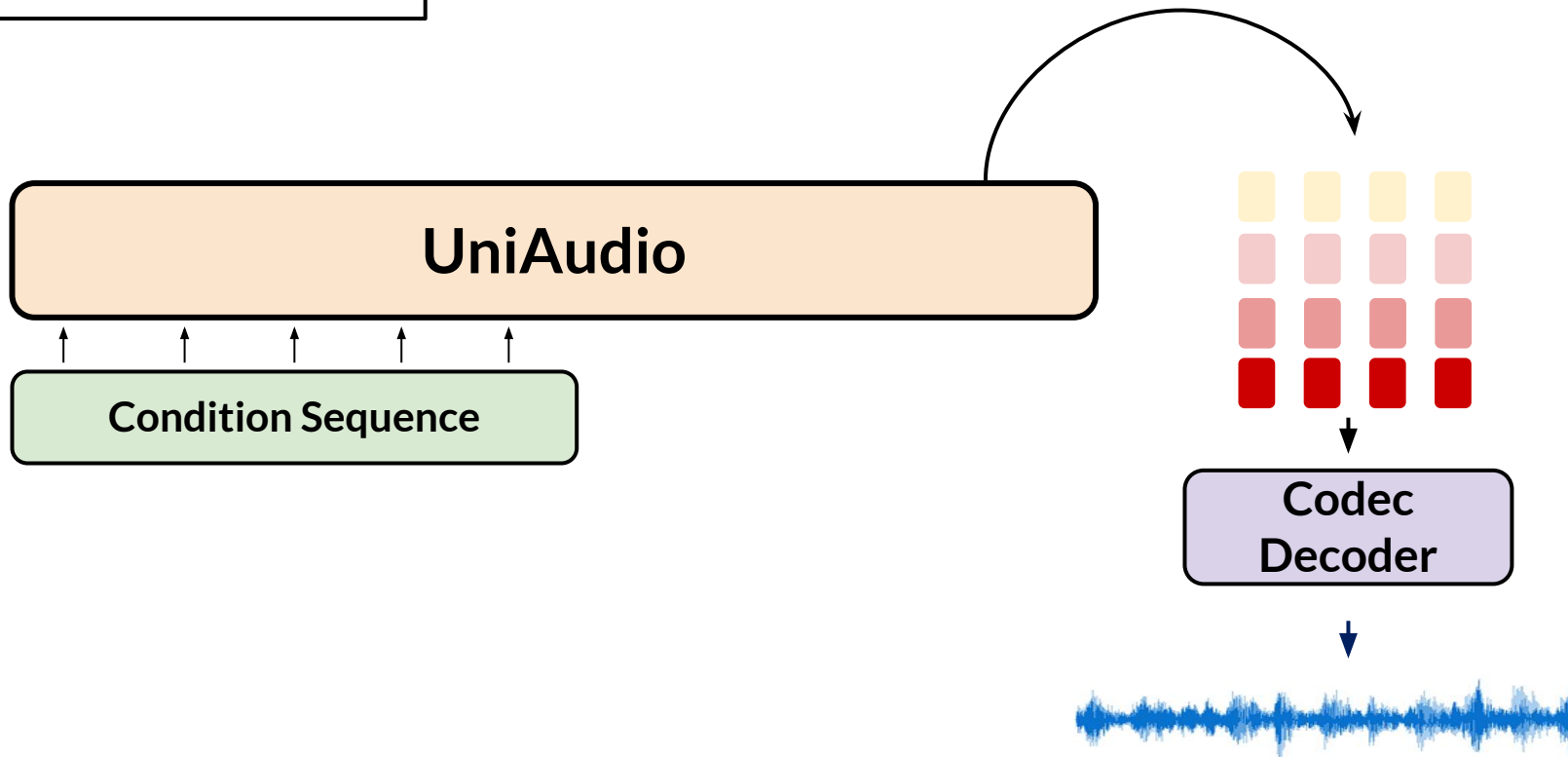




Collect dataset for various tasks

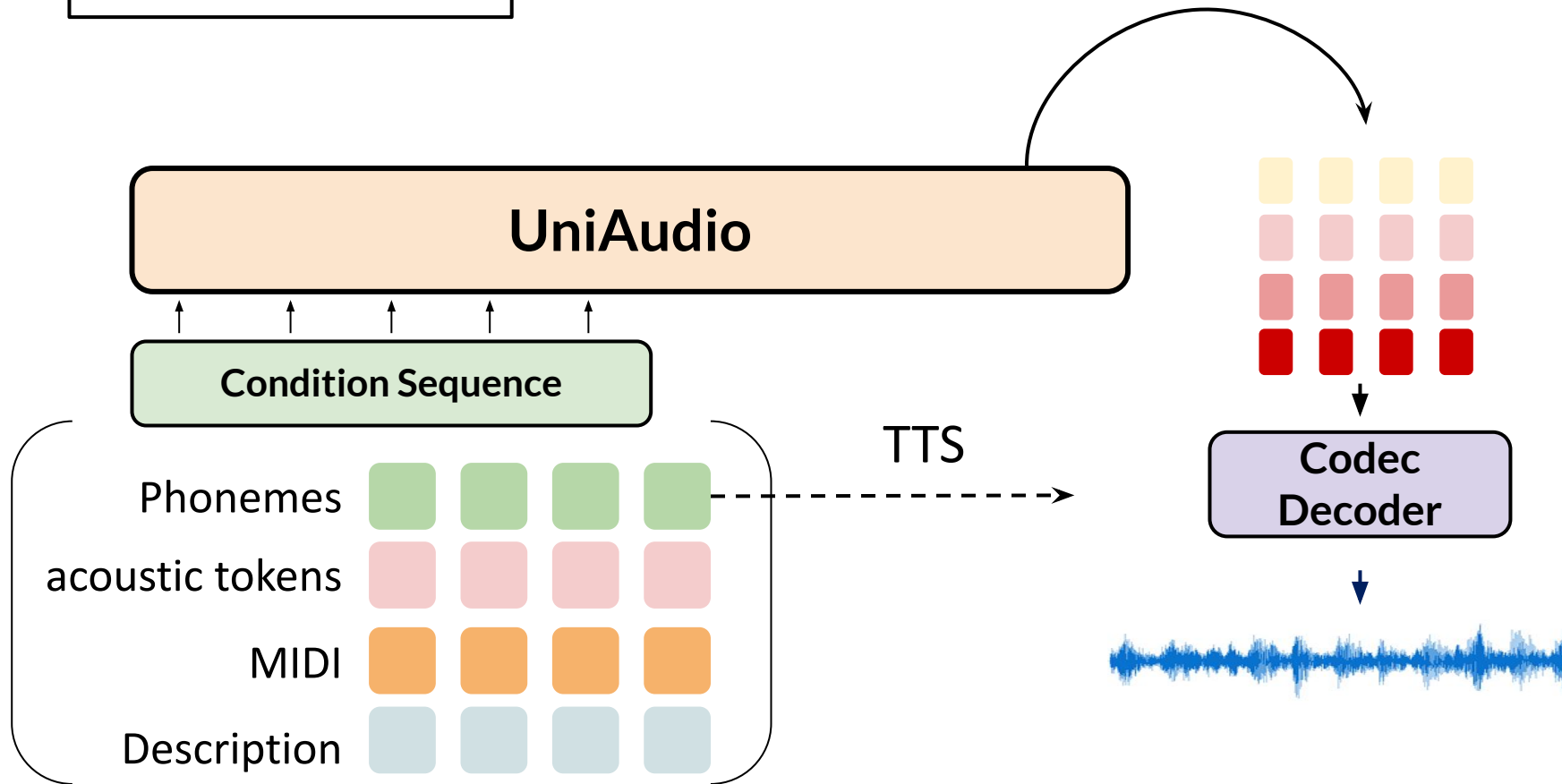
UniAudio

Unified audio generation



UniAudio

Unified audio generation

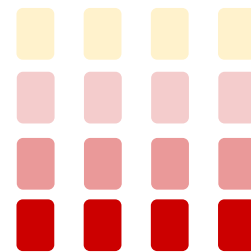
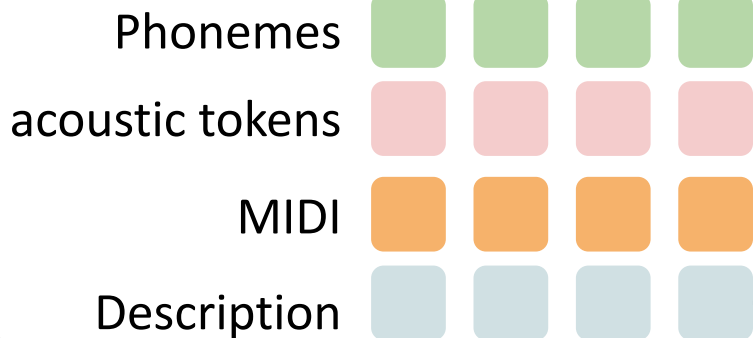


UniAudio

Unified audio generation

UniAudio

Condition Sequence

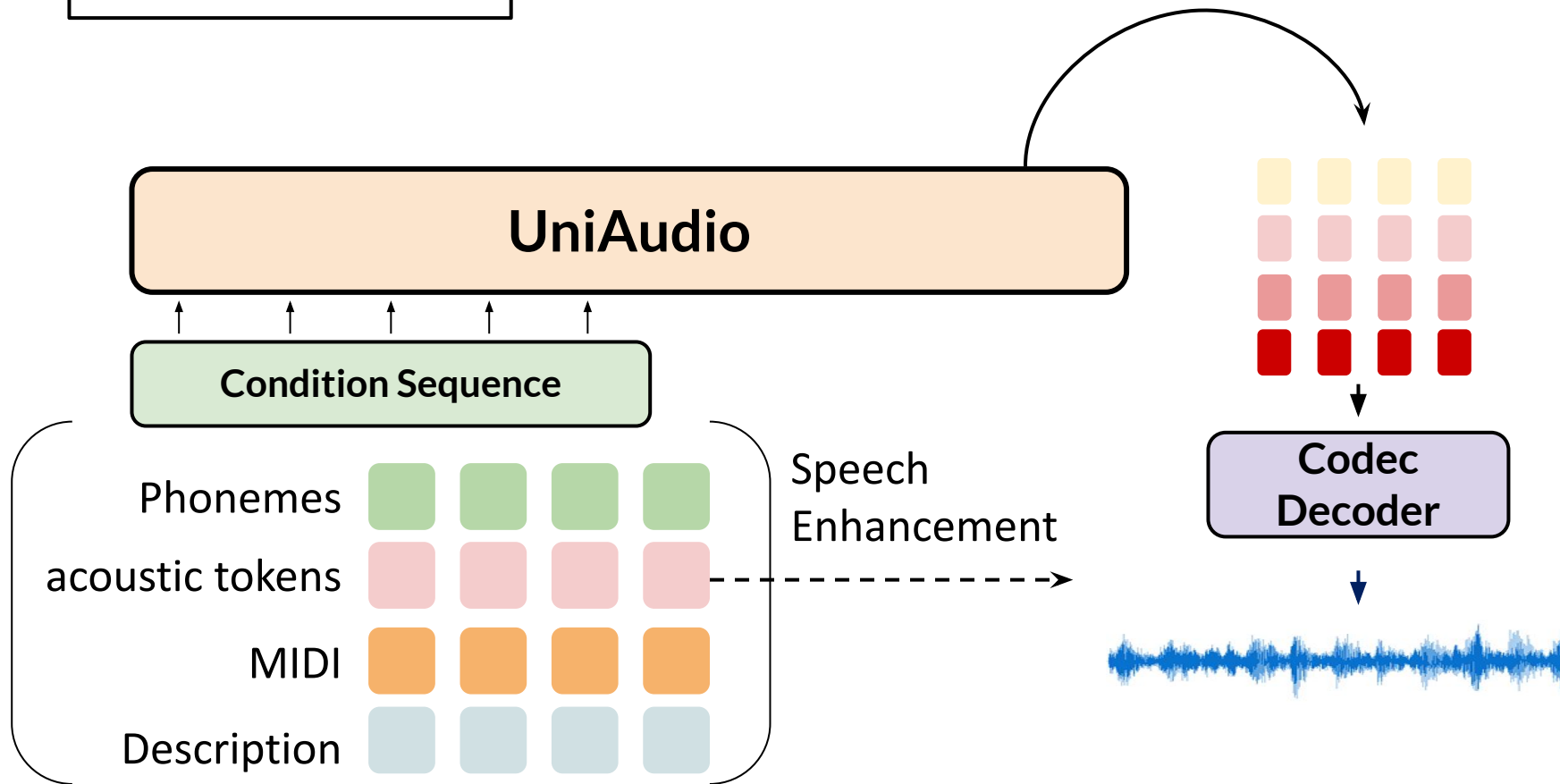


Codec Decoder



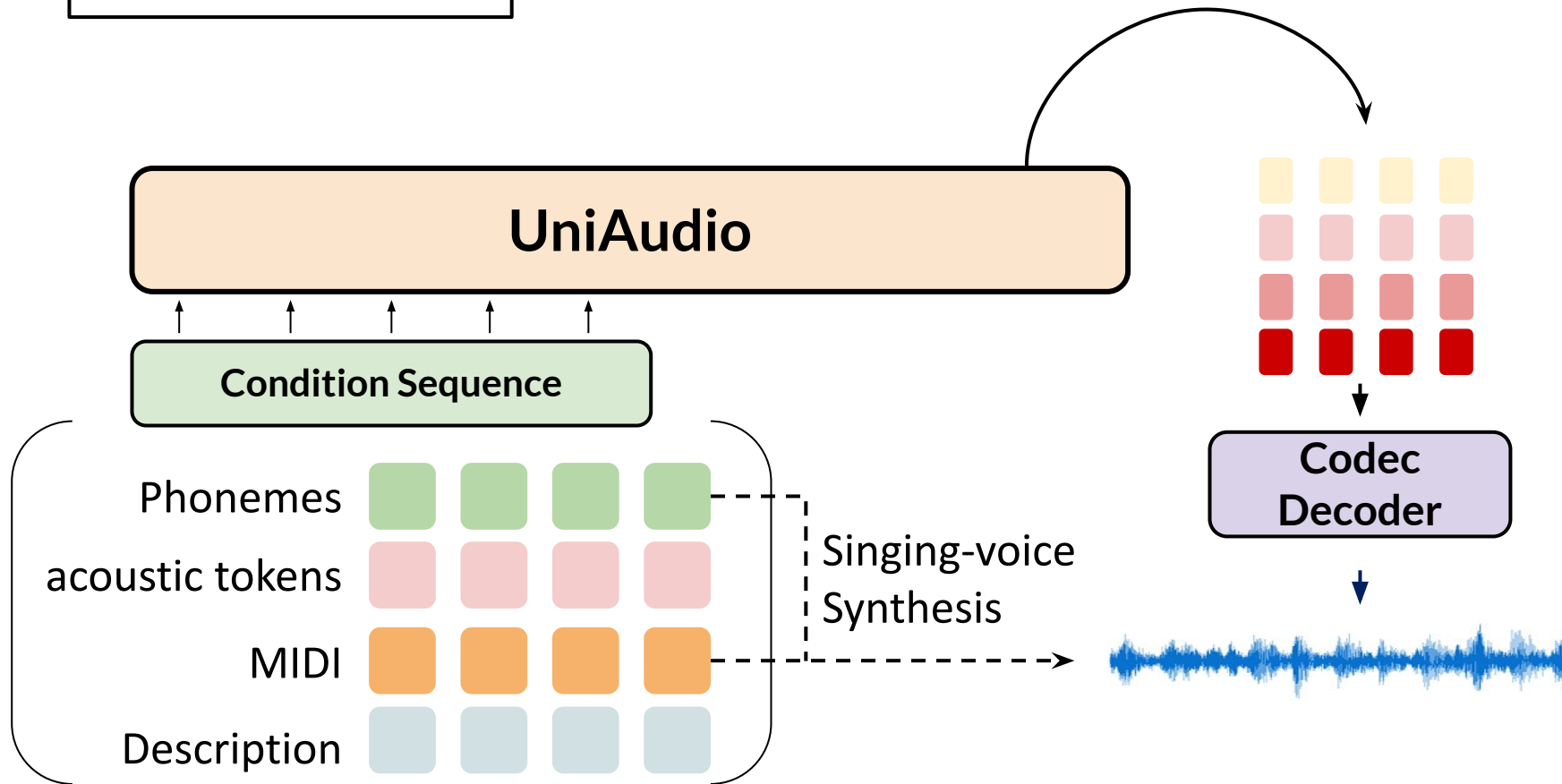
UniAudio

Unified audio generation



UniAudio

Unified audio generation



UniAudio

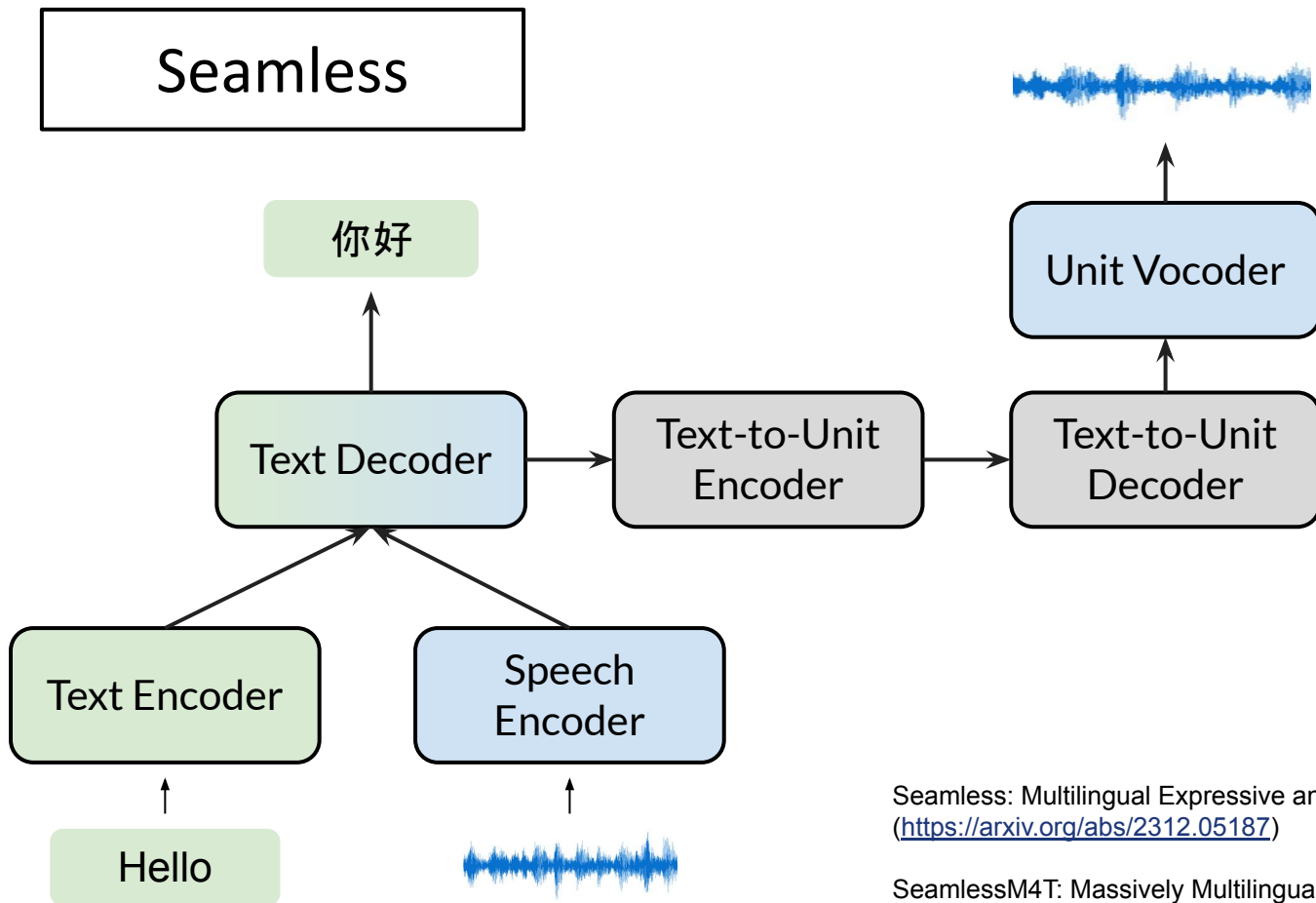
Task	Model	Subjective Evaluation	
		Metrics	Results
Text-to-Speech	Shen et al. (2023)	MOS(↑)	3.83±0.10 / 3.11±0.10
	UniAudio	/ SMOS(↑)	3.81±0.07 / 3.56±0.10
Voice Conversion	Wang et al. (2023e)	MOS(↑)	3.41±0.08 / 3.17±0.09
	UniAudio	/ SMOS(↑)	3.54±0.07 / 3.56±0.07
Speech Enhancement	Richter et al. (2023)		3.56±0.08
	UniAudio	MOS(↑)	3.68±0.07
Target Speaker Extraction	Wang et al. (2018)		3.43±0.09
	UniAudio	MOS(↑)	3.72±0.06
Singing Voice Synthesis	Liu et al. (2022)	MOS(↑)	3.94±0.02 / 4.05±0.06
	UniAudio	/ SMOS(↑)	4.08±0.04 / 4.04±0.05
Text-to-Sound	Liu et al. (2023a)	OVL (↑)	61.0±1.9 / 65.7±1.8
	UniAudio	/ REL (↑)	61.9±1.9 / 66.1±1.5
Text-to-Music	Copet et al. (2023)	OVL (↑)	73.3±1.5 / 71.3±1.7
	UniAudio	/ REL (↑)	67.9±1.7 / 70.0±1.5

Prior work

UniAudio

Achieve competitive results on various audio generation tasks

Seamless

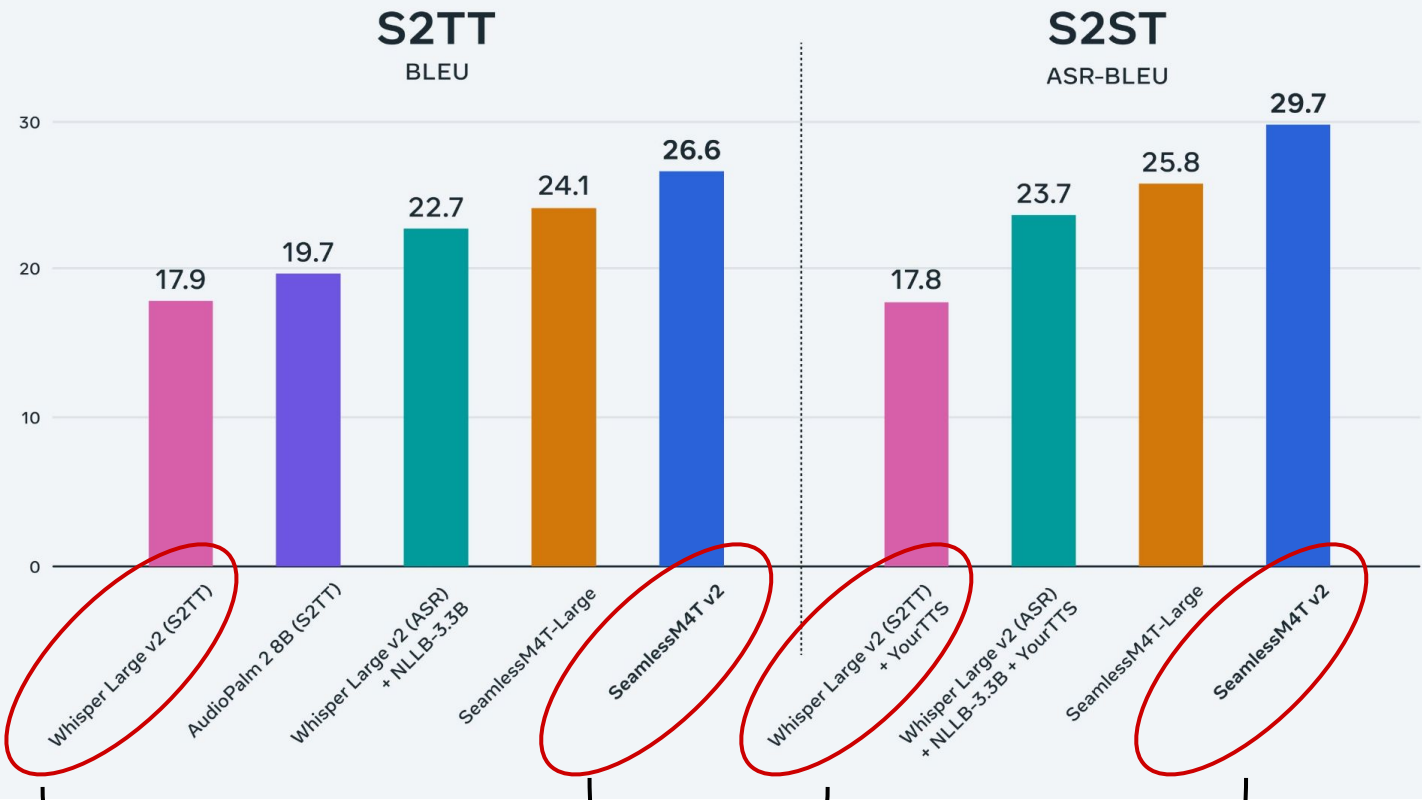


1. S2ST
2. S2TT
3. T2ST
4. T2TT
5. ASR

Seamless: Multilingual Expressive and Streaming Speech Translation
(<https://arxiv.org/abs/2312.05187>)

SeamlessM4T: Massively Multilingual & Multimodal Machine Translation
(<https://arxiv.org/abs/2308.11596>)

High-level competitive landscape for the SeamlessM4T v2 model



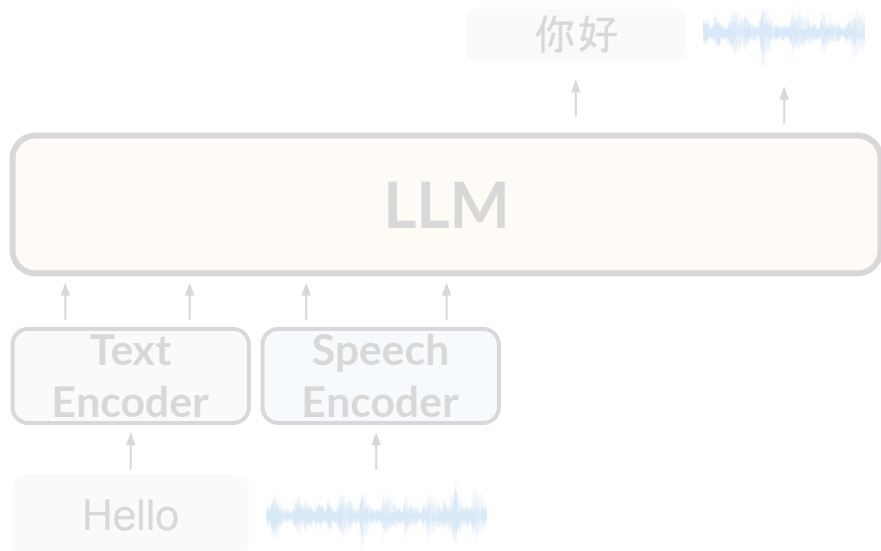
Whisper / Cascaded Model

Seamless

(Dec. 2023)

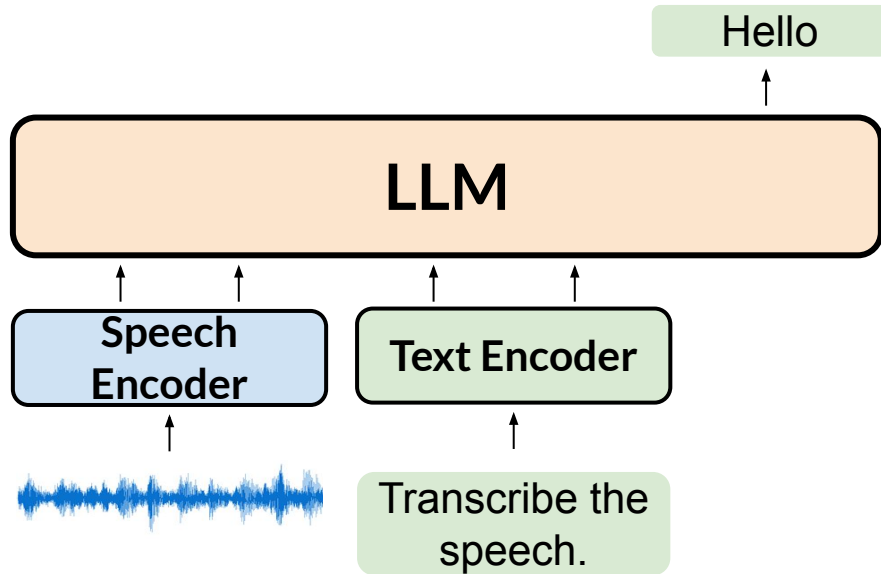
Source: <https://ai.meta.com/resources/models-and-libraries/seamless-communication-models/>

(3) LLM that listens & speaks



- Multi-task Learning

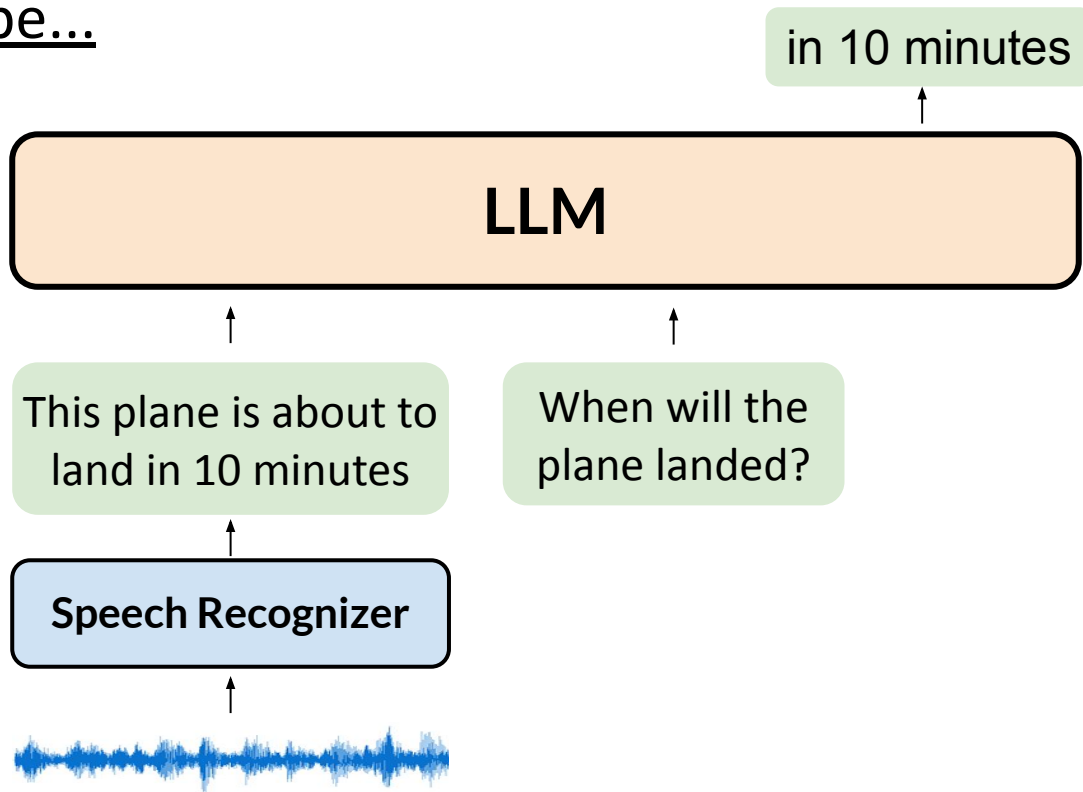
(4) LLM that follows instructions



- Instruction Tuning

We know that LLM can follow instructions.

A naive method would be...



Cascaded model has some problems...

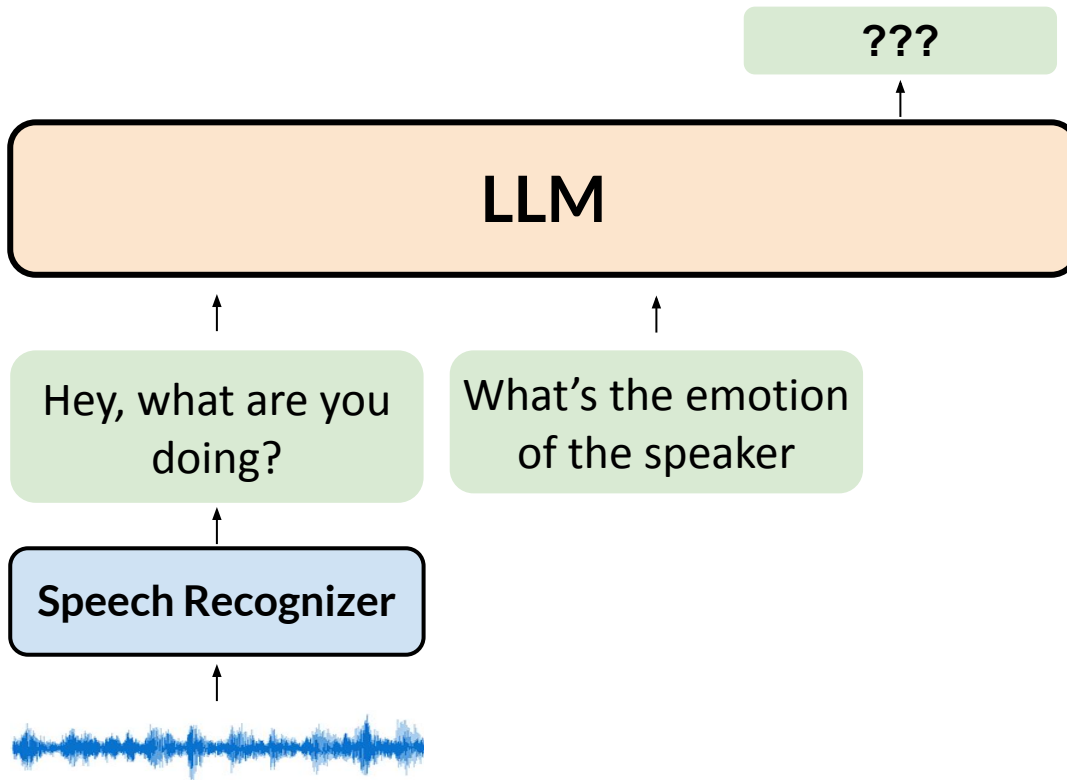
e.g. remove paralinguistic information



Hey, what are you
doing?



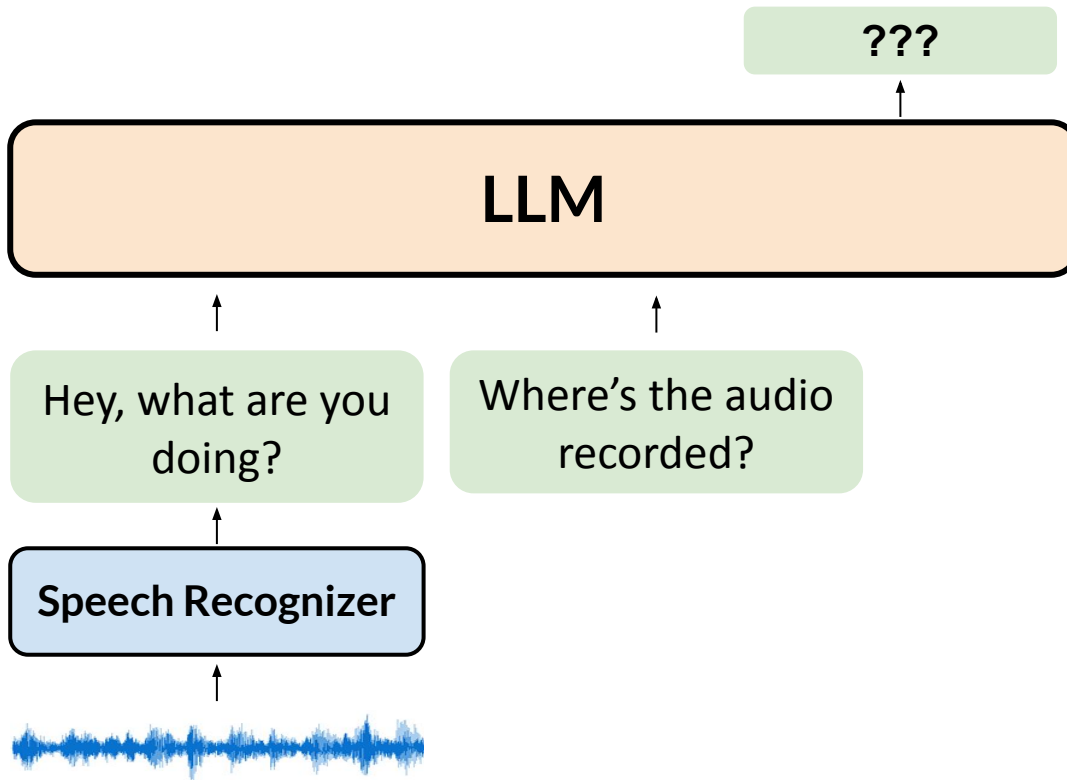
Hey, what are you
doing?



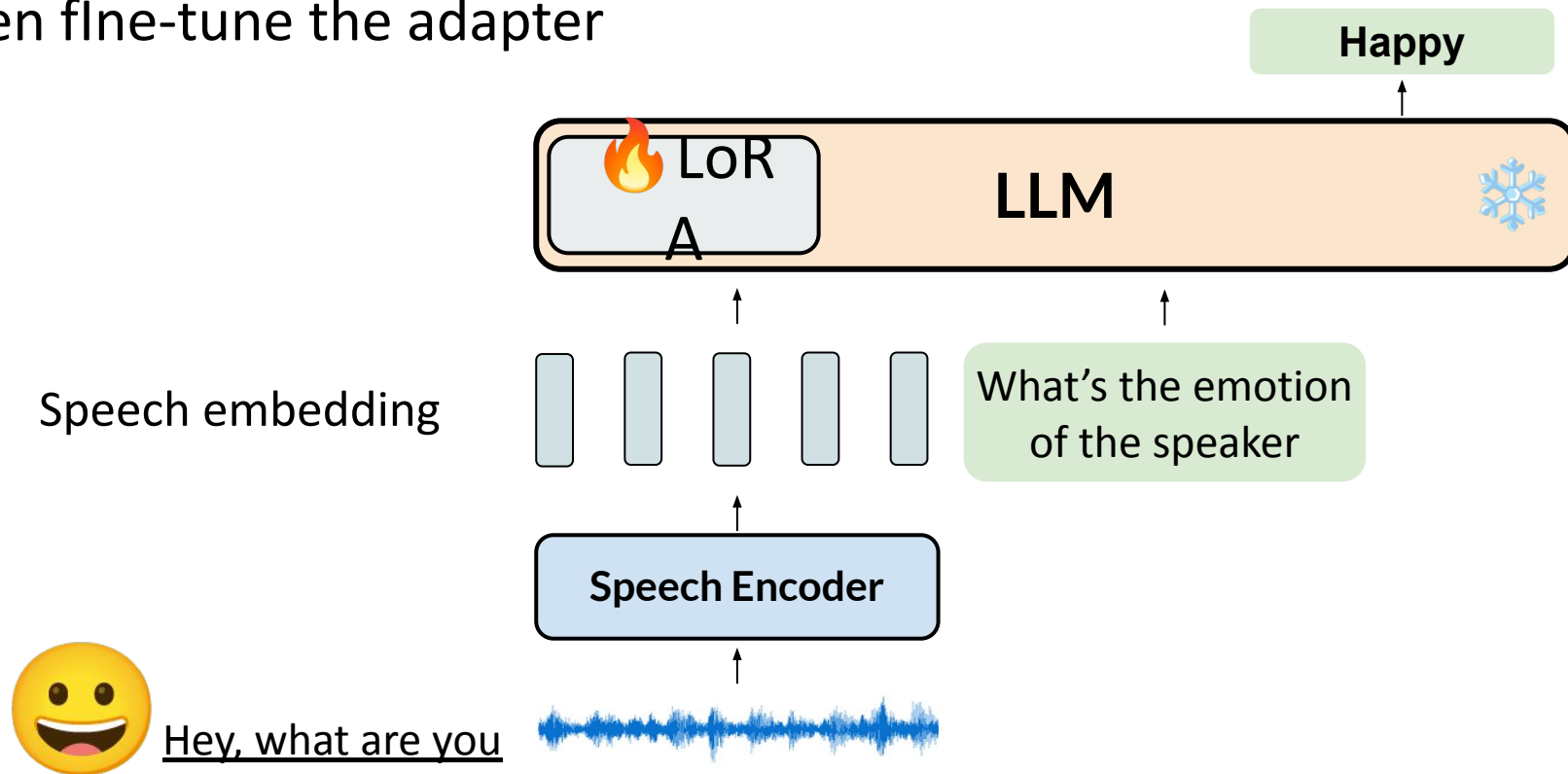
Cascaded model has some problems...

e.g. Remove acoustic information

Hey, what are you doing?
(Water sound and dishes clanking sound)



Use continuous speech features as input.
Then fine-tune the adapter

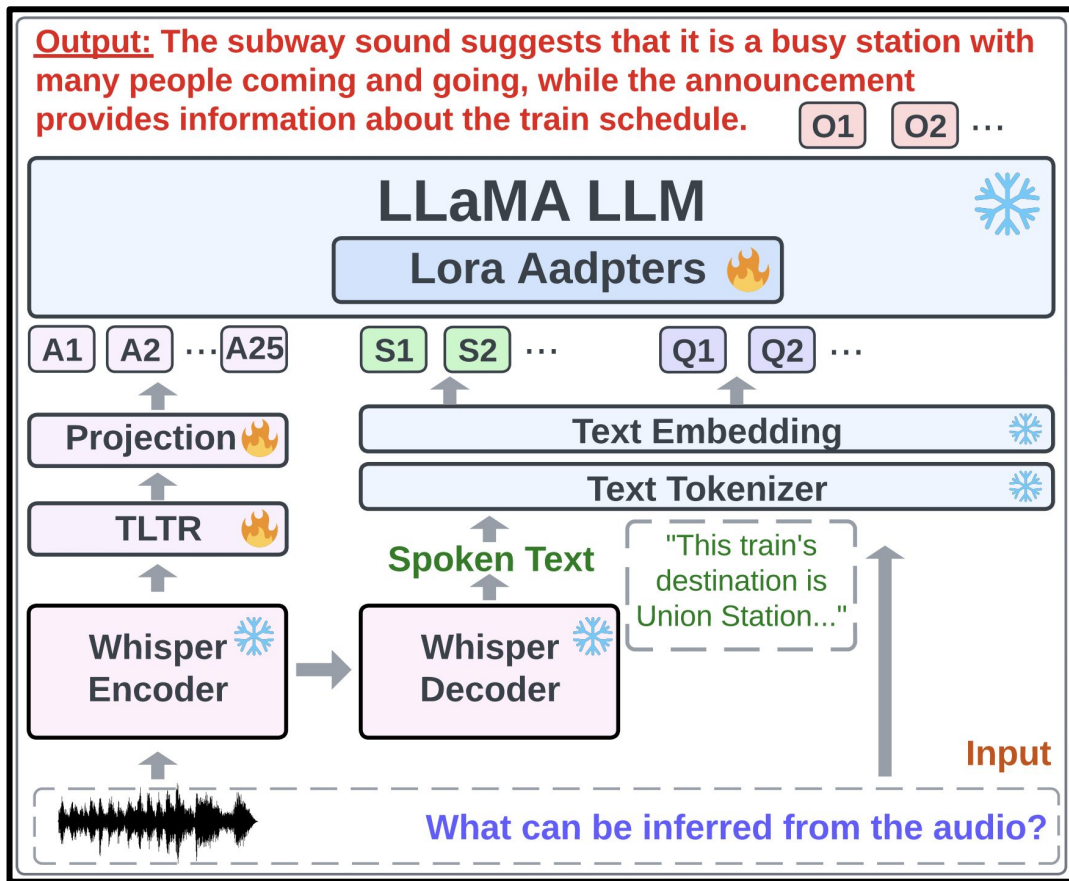


LTU-AS

LTU: Listen, Think and Understand

Whisper:

- Encode speech representation
- Transcribe spoken text



LTU-AS

input



(anxious speaker)

emotion
detection

Sample 2 Interview (VoxCeleb2-Eval)

Transcribed speech: ... I was so anxious about getting up and talking in front of people that I thought I still sort of lose...

Q: What is the emotion of the speaker? Why?

A: Based on the speech content, the speaker seems to express anxiety or nervousness about public speaking. The high pitch and fast speed suggest excitement or tension as well. Therefore, the emotion of the speaker can be inferred as nervousness or anxiety.

Q: Is the speaker a good public speaker?

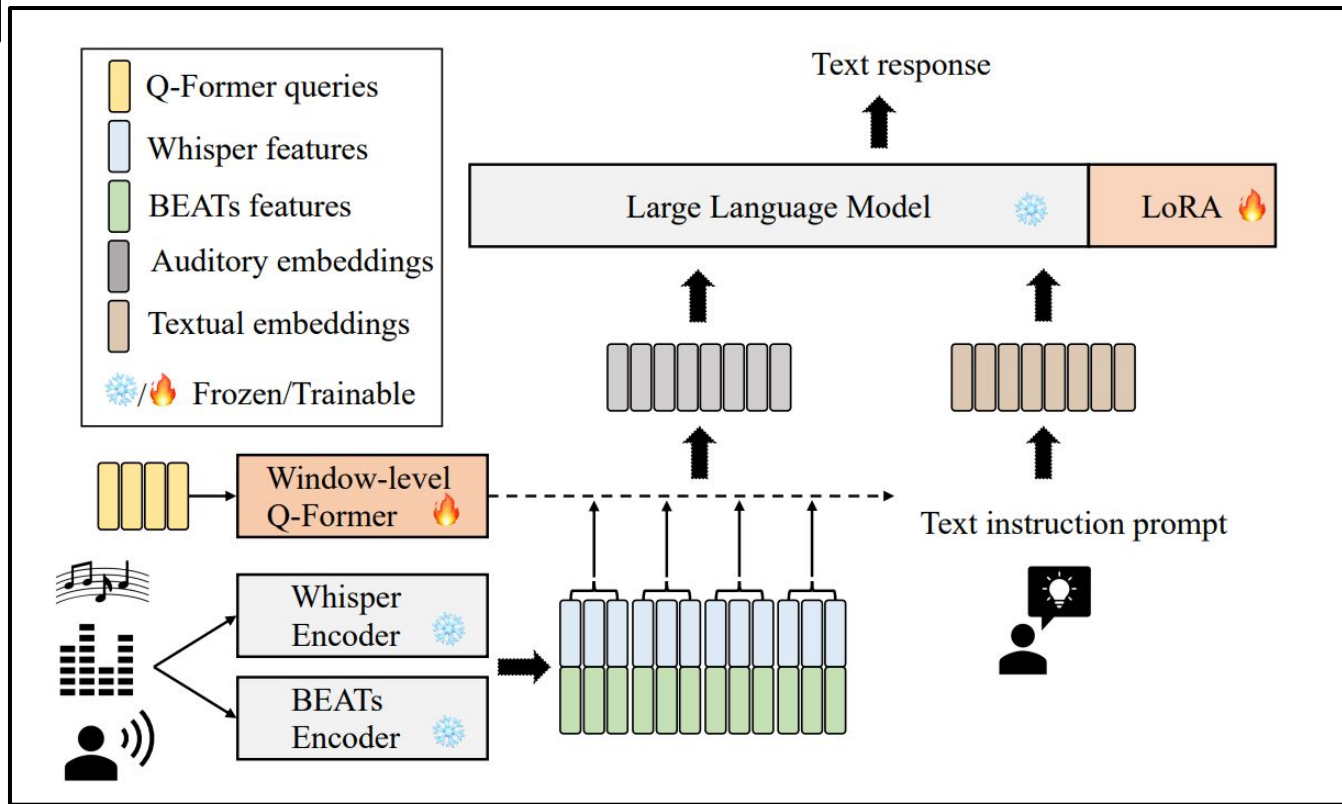
A: No, the speaker feels anxious about speaking in front of people and thinks they will still feel lose-lose about it.

Q: Provide the speaker 5 suggestions.

A: 1. Take a deep breath before speaking. 2. Practice public speaking regularly. 3. Use humor or lightheartedness to ease nerves. 4. Focus on the audience rather than yourself during speeches. 5. Seek feedback from others to improve your delivery.

SALMONN

- Whisper as speech encoder
- BEATs as audio encoder




SALMONN


Speech Recognition

Recognize the speech and give me the transcription.


Music Captioning


Please describe the music in detail.


 (A man asks "Can you guess where I am right now", and explosions and gunshots can be heard in the background.)

 USER: Recognize the speech and give me the transcription.


Can you guess where I am right now?

 SALMONN

 (It's a soothing, elegant piano piece. At the end of the audio a woman begins to sing.)

 USER: Please describe the music in detail.

This is a beautiful piano and vocal piece. The piano plays a simple melody while the female vocalist sings in a high register. The atmosphere is romantic and dreamy. The piece has a soft and mellow feel to it. This piece could be used in a romantic movie scene or as a lullaby.

 SALMONN

Instruct-VC

For speech generation tasks

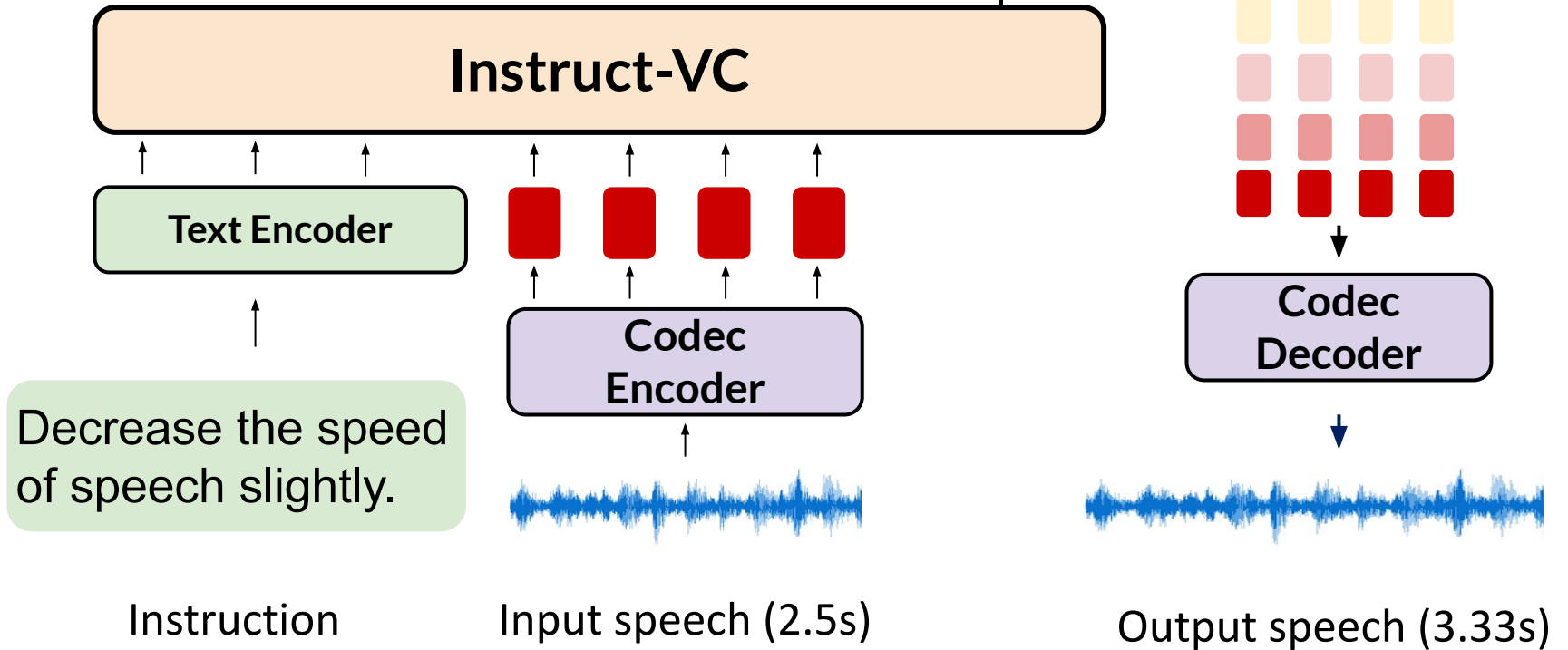
```
``sox input.wav output.wav speed 0.75``
```

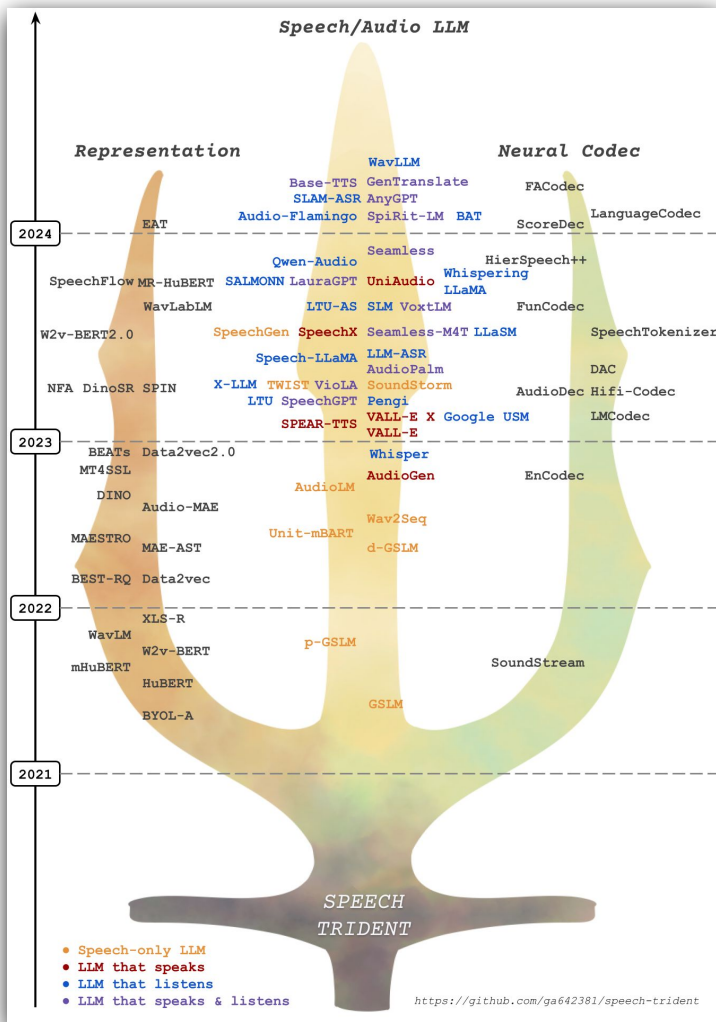


Input speech (2.5s)

Output speech (3.33s)

Instruct-VC





Speech Trident

- Speech / Audio LLMs
- Representation Learning Models
- Neural Codec Models



<https://github.com/ga642381/speech-trident>