


# 語音語言模型的發展史

**Speaker:**

**Hung-yi Lee 李宏毅**

---

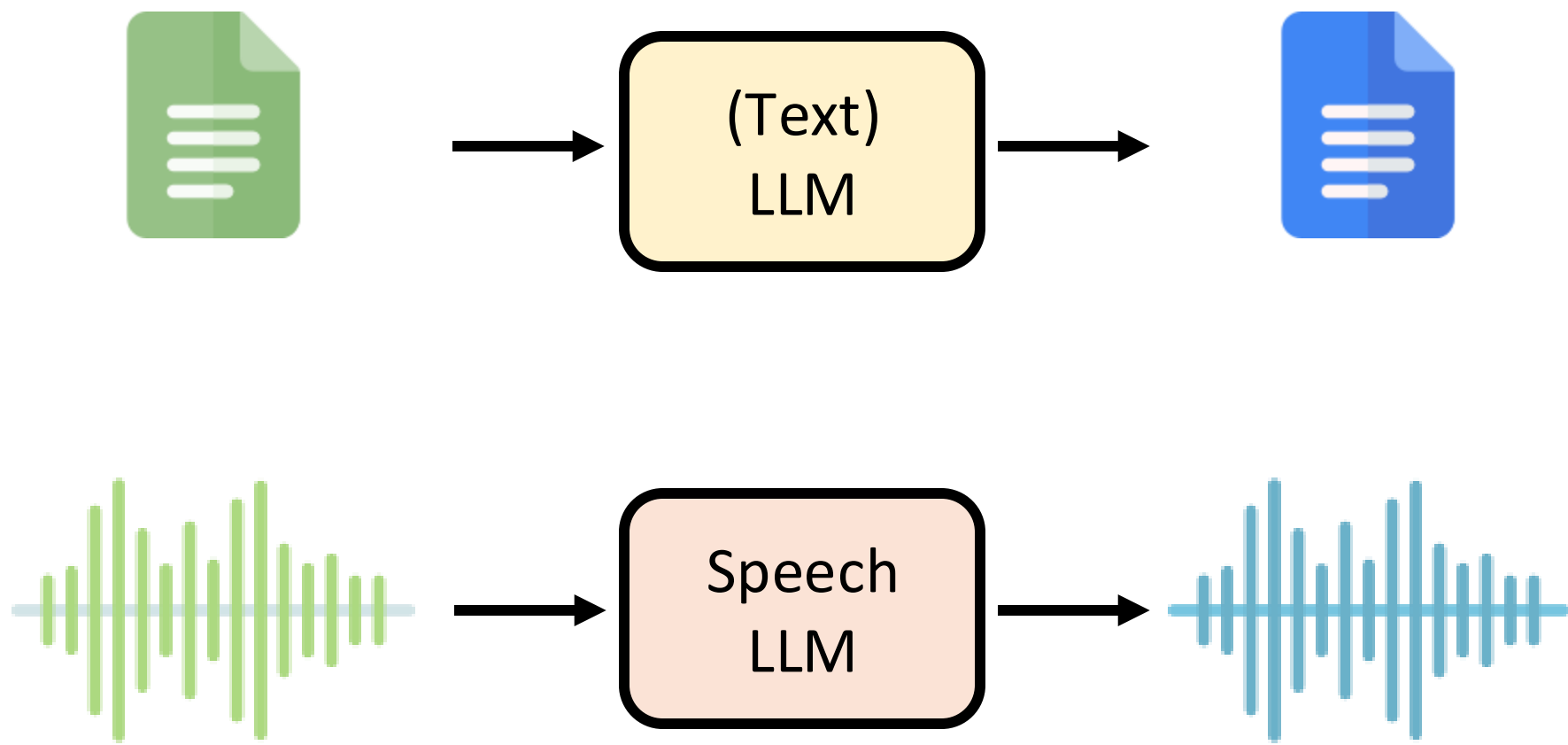


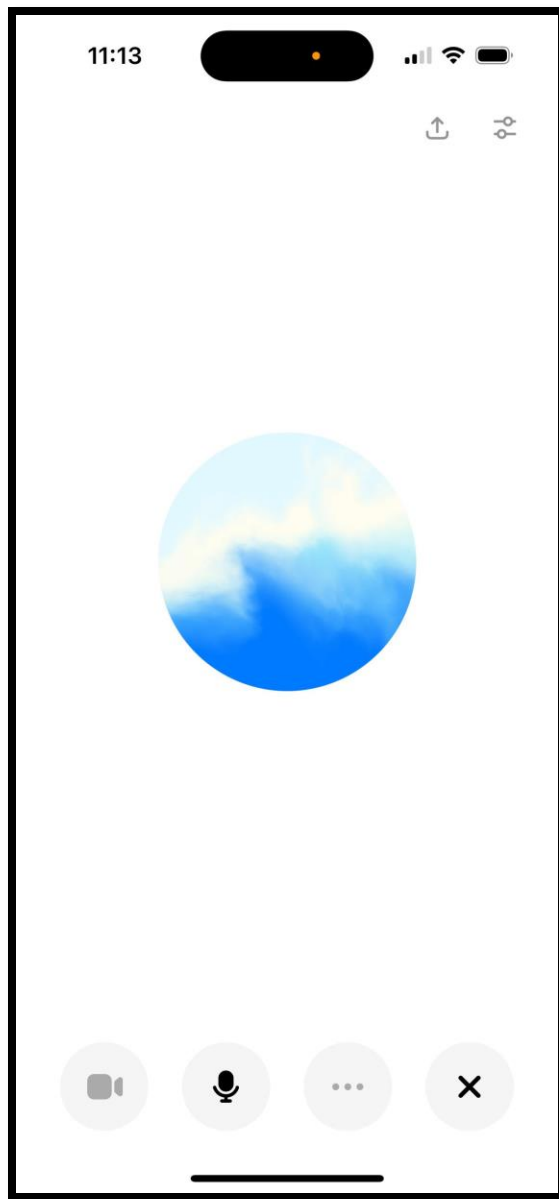
# 免責聲明

本課程並非完整介紹語音語言模型，而是以講者經驗講述語音語言模型的發展

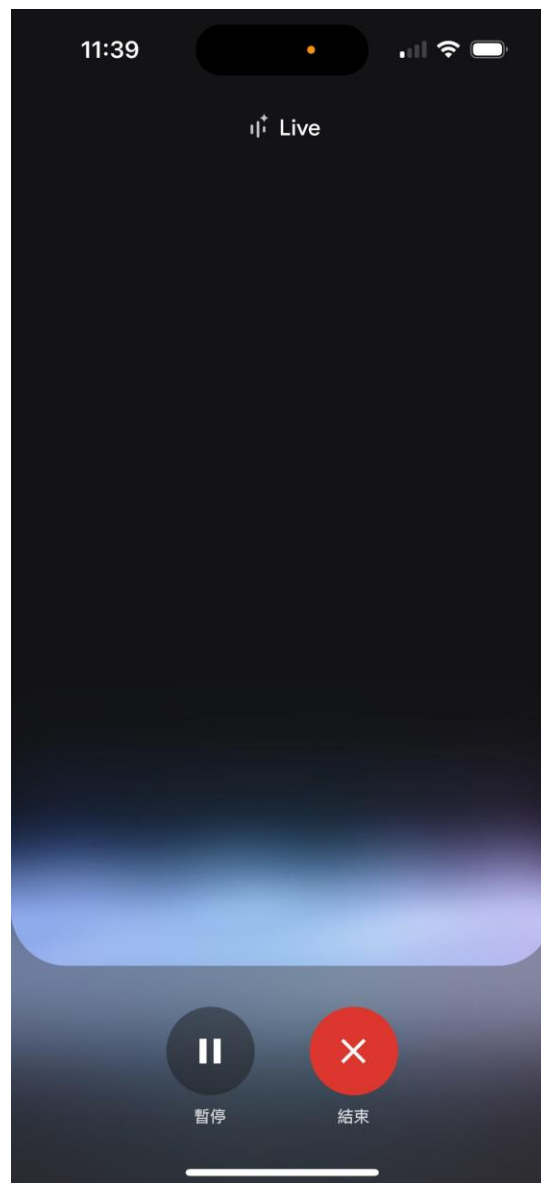
---

# 語音語言模型 (Speech Language Model)





ChatGPT  
voicemode

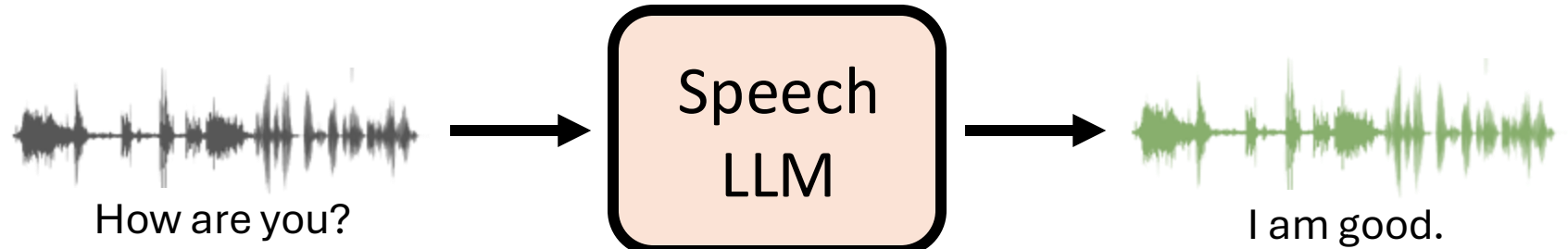


Gemini  
Live

- Moshi
    - <https://arxiv.org/abs/2410.00037>
  - Sesame
    - [https://www.sesame.com/research/crossing\\_the\\_uncanny\\_valley\\_of\\_voice](https://www.sesame.com/research/crossing_the_uncanny_valley_of_voice)
  - GLM-4-Voice
    - <https://arxiv.org/abs/2412.02612>
  - Step-Audio
    - <https://arxiv.org/abs/2502.11946>
  - Qwen2.5-Omni
    - <https://arxiv.org/abs/2503.20215>
  - Kimi-Audio
    - <https://arxiv.org/abs/2504.18425>
  - SpeechGPT
    - <https://github.com/OpenMOSS/SpeechGPT-2.0-preview>
  - Doubao Realtime Voice Model
    - [https://seed.bytedance.com/en/realtime\\_voice](https://seed.bytedance.com/en/realtime_voice)
- ..... just to name a few

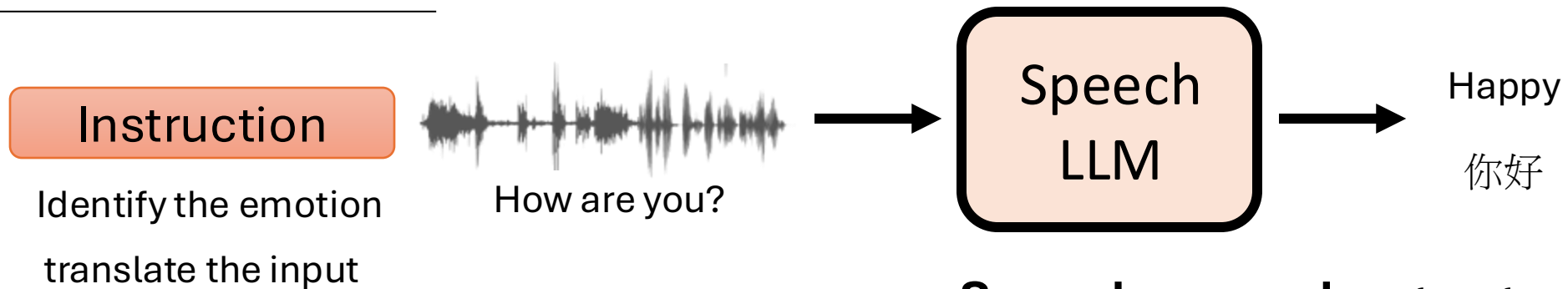
# 語音語言模型 (Speech Language Model)

## Dialogue Mode



the focus of today

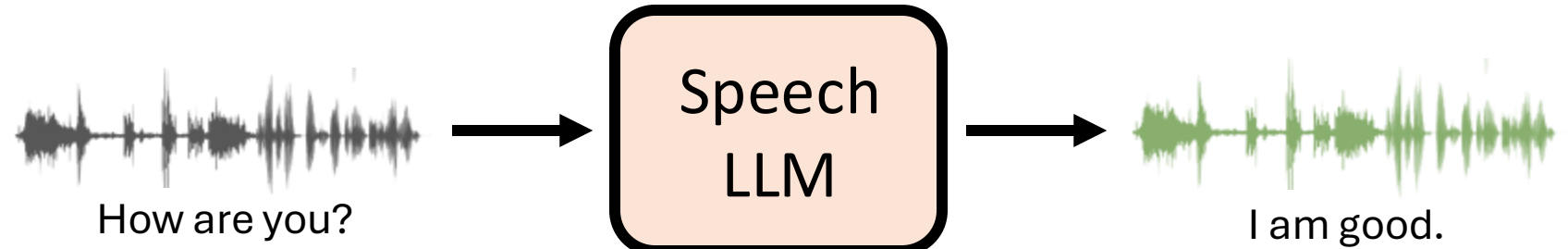
## Command Mode



**Speech-aware language model**

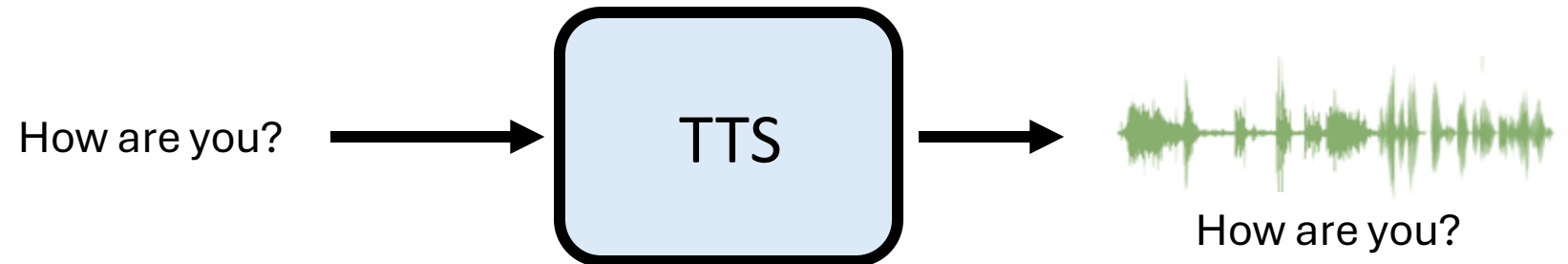
# 語音語言模型 (Speech Language Model)

## Dialogue Mode



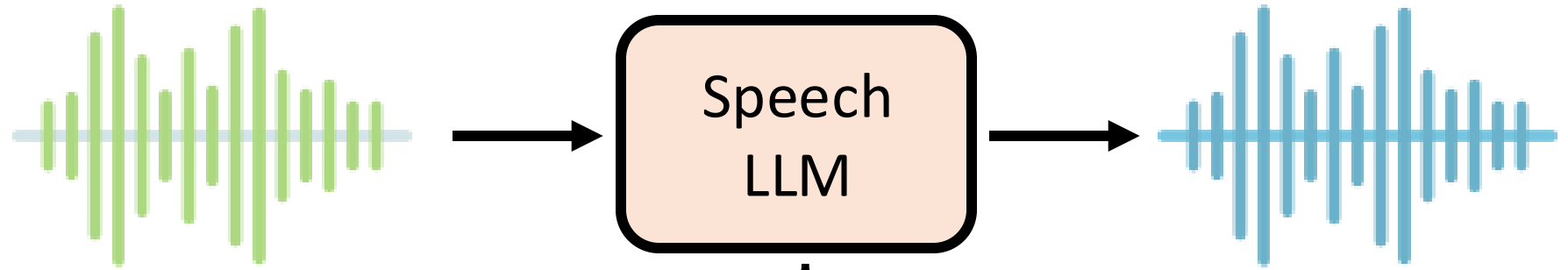
the focus of today

**Not TTS**



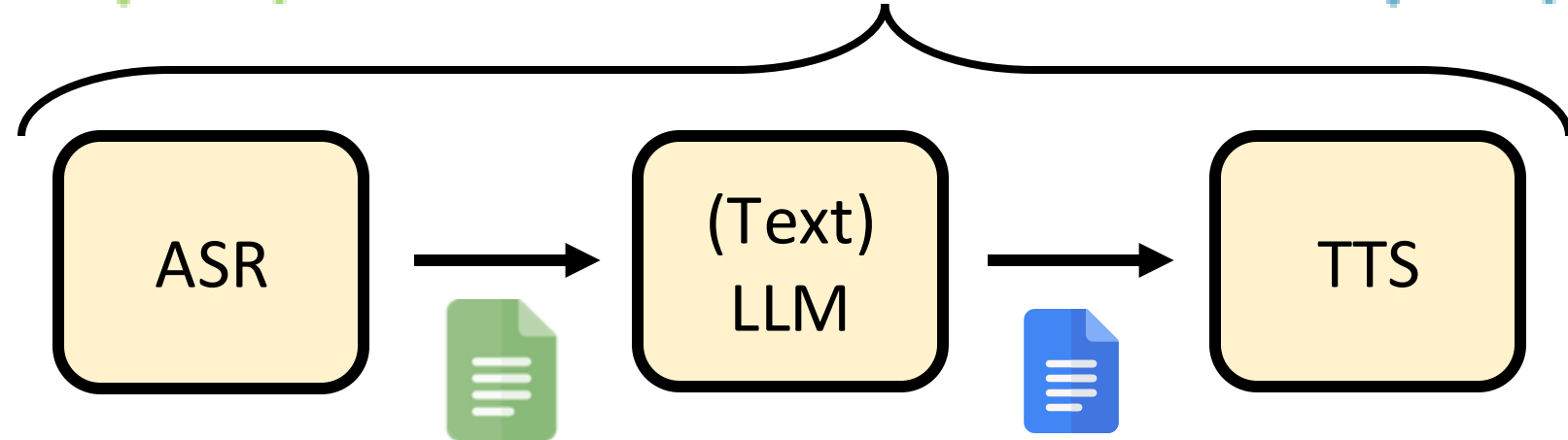
Some TTS models also claim themselves to be speech LMs.

# Non end-to-end solution

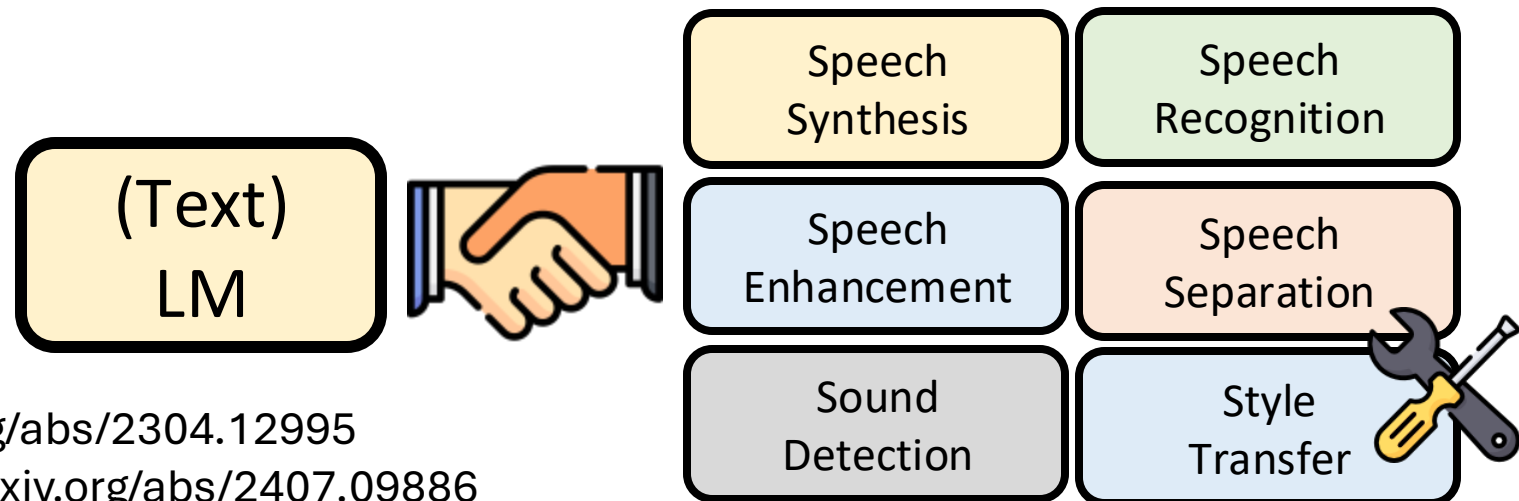


## Cascade

Most speech-to-speech systems use this solution.

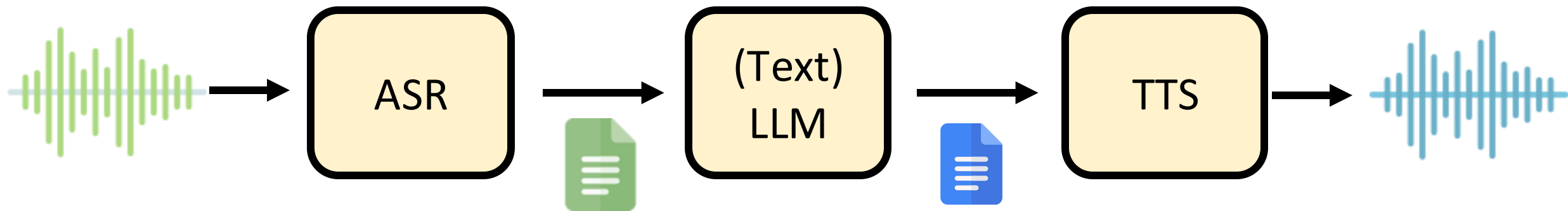


## Agentic Solution



AudioGPT:<https://arxiv.org/abs/2304.12995>

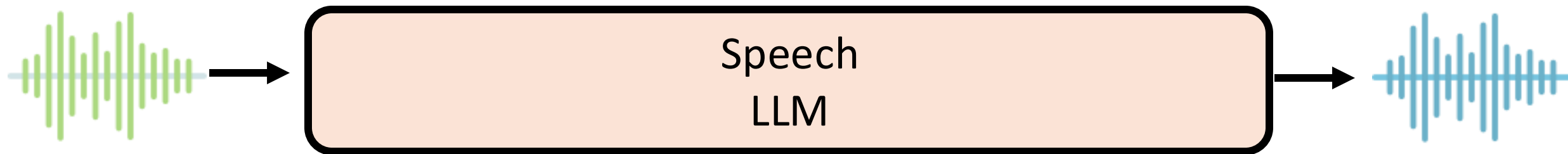
Speech-Copilot:<https://arxiv.org/abs/2407.09886>



Strength: Easy to construct

Weakness: Information loss (e.g., emotion), latency

在 2025 年會有  
比較好的結果



Strength: Capture all speech information, probably low latency

Weakness: Still under investigation

較高的能力  
上限

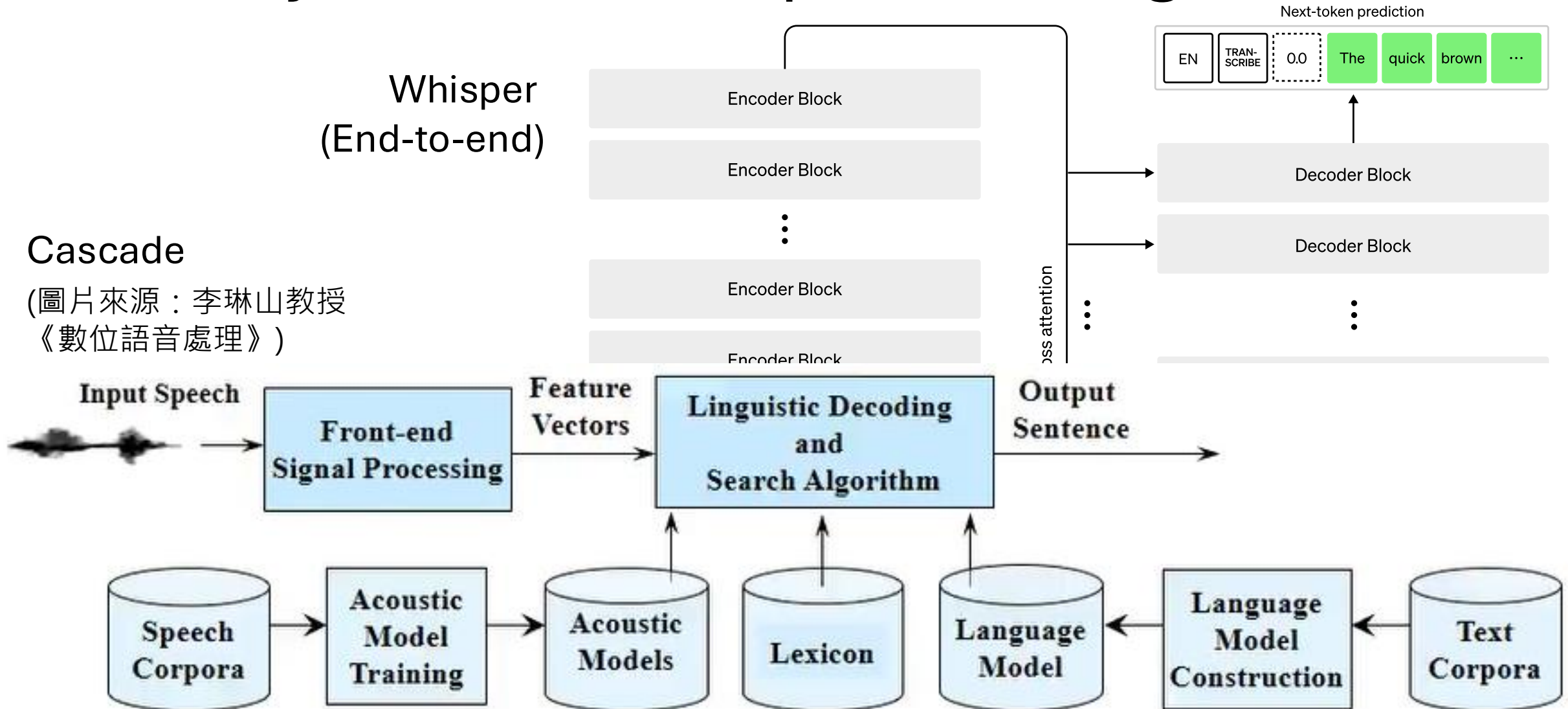


# History of Automatic Speech Recognition ...

Whisper  
(End-to-end)

Cascade

(圖片來源：李琳山教授  
《數位語音處理》)



# History of Automatic Speech Recognition ...

[https://www.isca-archive.org/interspeech\\_2015/lu15e\\_interspeech.html](https://www.isca-archive.org/interspeech_2015/lu15e_interspeech.html)

10.4% on SWB ...

<https://ieeexplore.ieee.org/abstract/document/6854669>  
(ICASSP'14)

Step	Splicing	Space	CHM	SWB	Avg
1	$\pm 5$	feature	62.7	47.6	55.2
2	$\pm 5$	feature	61.3	40.8	51.1
3	$\pm 5$	feature	59.9	<b>38.8</b>	<b>49.4</b>
4	$\pm 5$	feature	60.2	41.7	51.0
1	$\pm 7$	feature	65.5	47.6	56.6
2	$\pm 7$	feature	59.9	41.7	50.9
3	$\pm 7$	feature	59.8	40.3	50.1
4	$\pm 7$	feature	60.0	43.0	51.6
2	$\pm 5$	hidden	60.7	42.3	51.5
3	$\pm 5$	hidden	<b>58.9</b>	41.7	50.3

# History of Automatic Speech Recognition ...

## PHONE RECOGNITION USING RESTRICTED BOLTZMANN MACHINES

*Abdel-rahman*

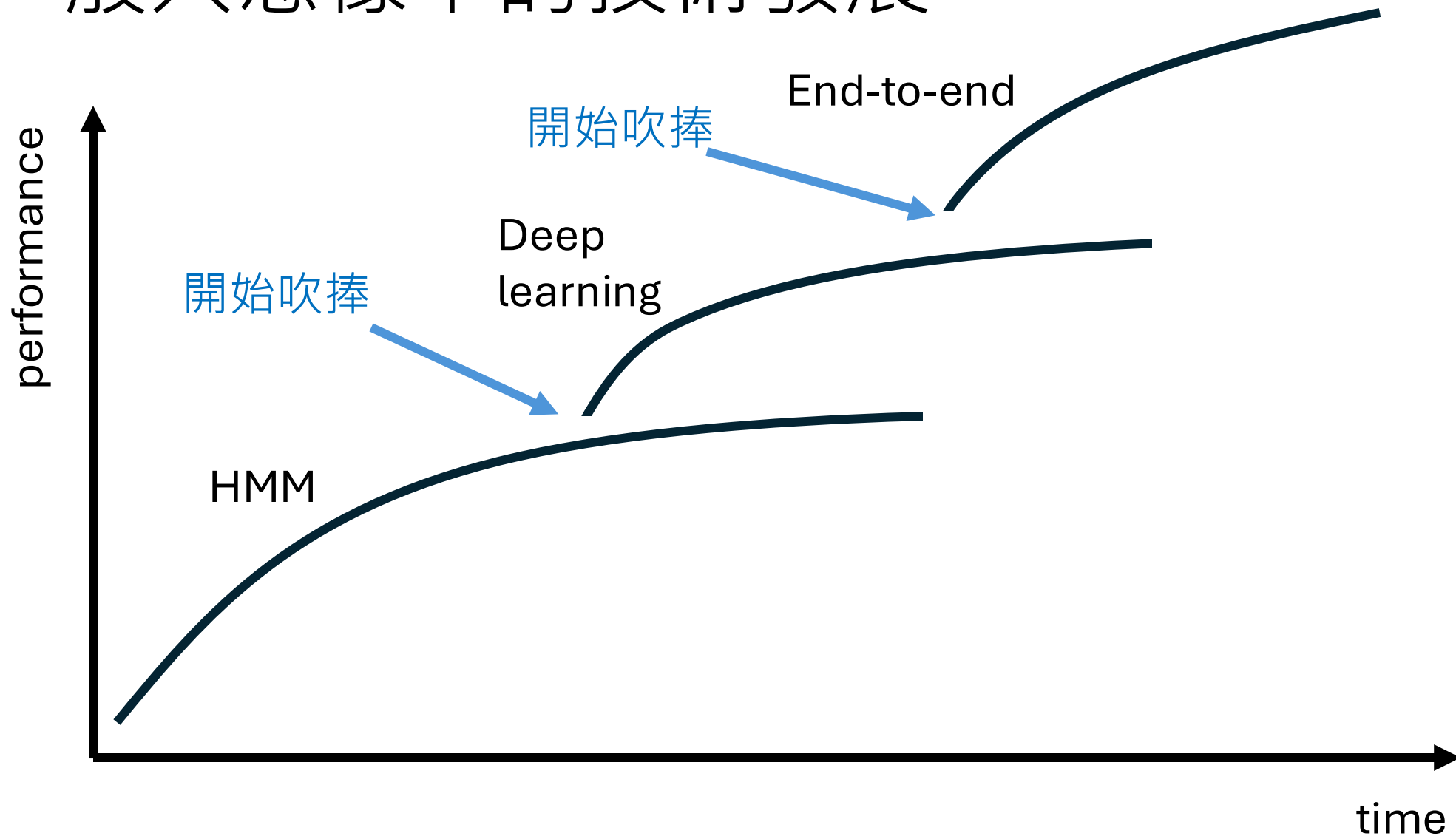
Department of Co

<https://ieeexplore.ieee.org/abstract/>

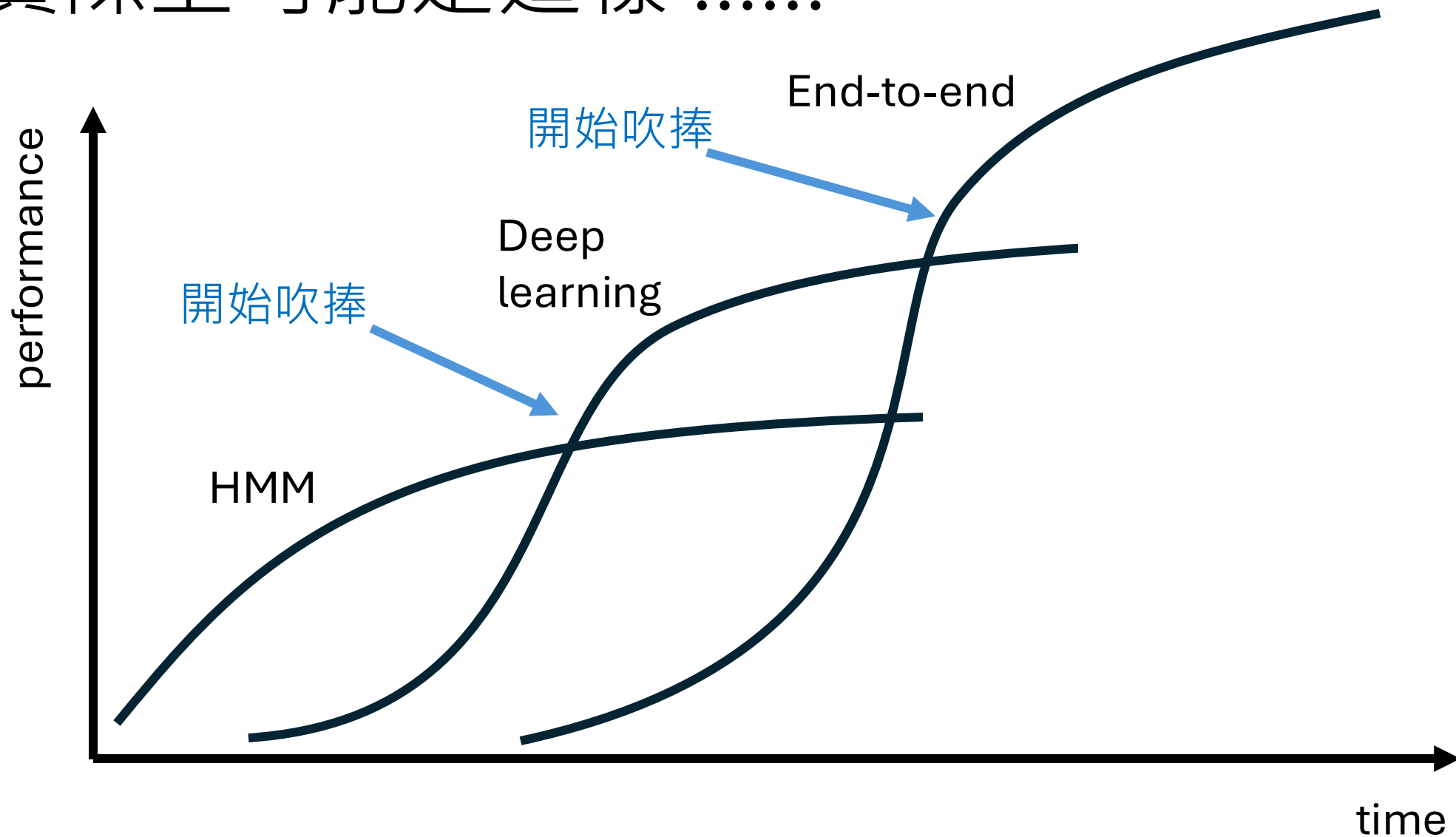
**Table 3.** *Reported results on TIMIT core test set*

Method	PER
Conditional Random Field [11]	34.8%
Large-Margin GMM [12]	28.2%
CD-HMM [2]	27.3%
ICRBM (this paper)	<b>26.7%</b>
Augmented conditional Random Fields [2]	26.6%
Recurrent Neural Nets [13]	26.1%
Monophone HTMs [1]	24.8%
Heterogeneous Classifiers [14]	24.4%

# 一般人想像中的技術發展

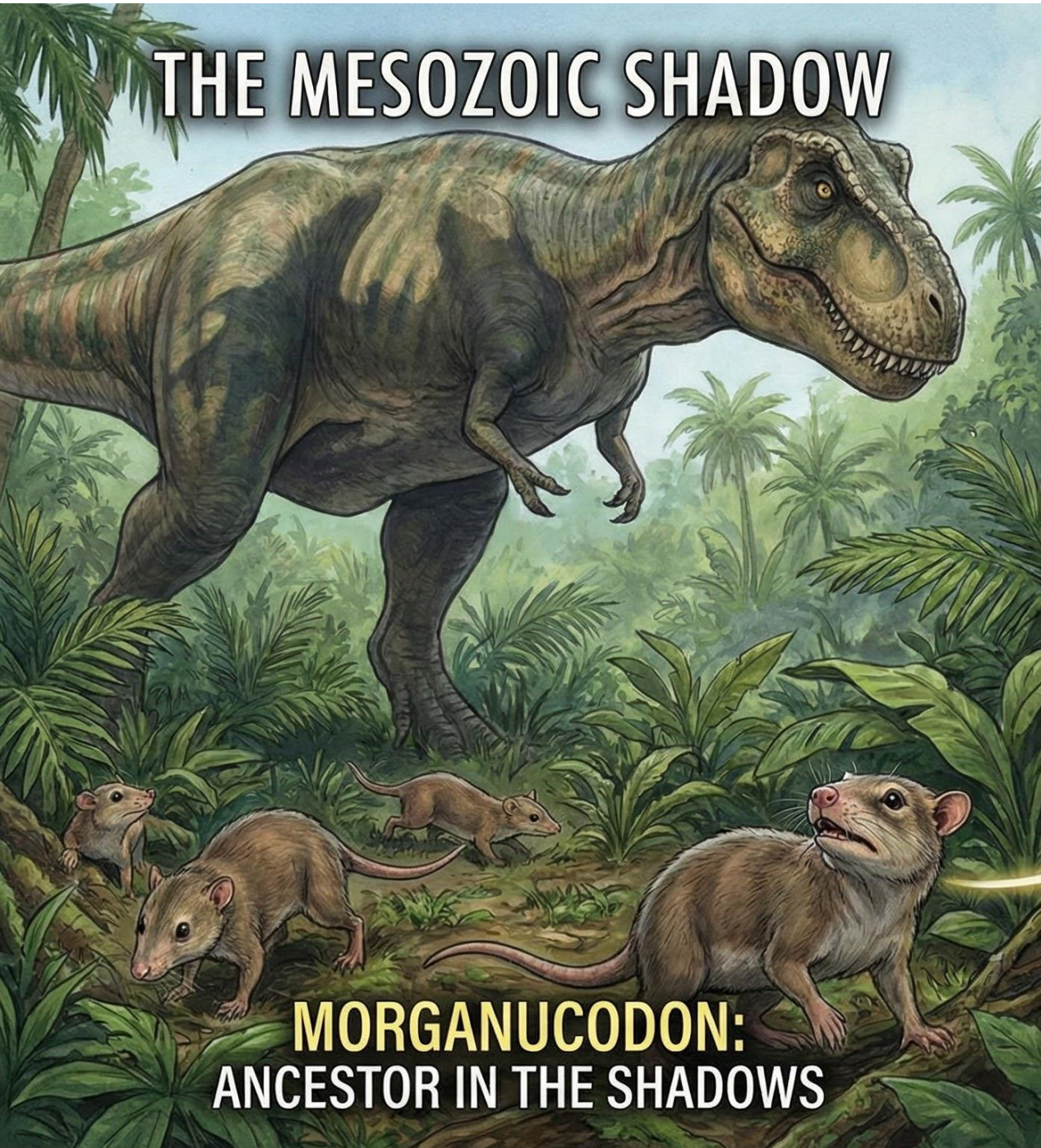


實際上可能是這樣 .....



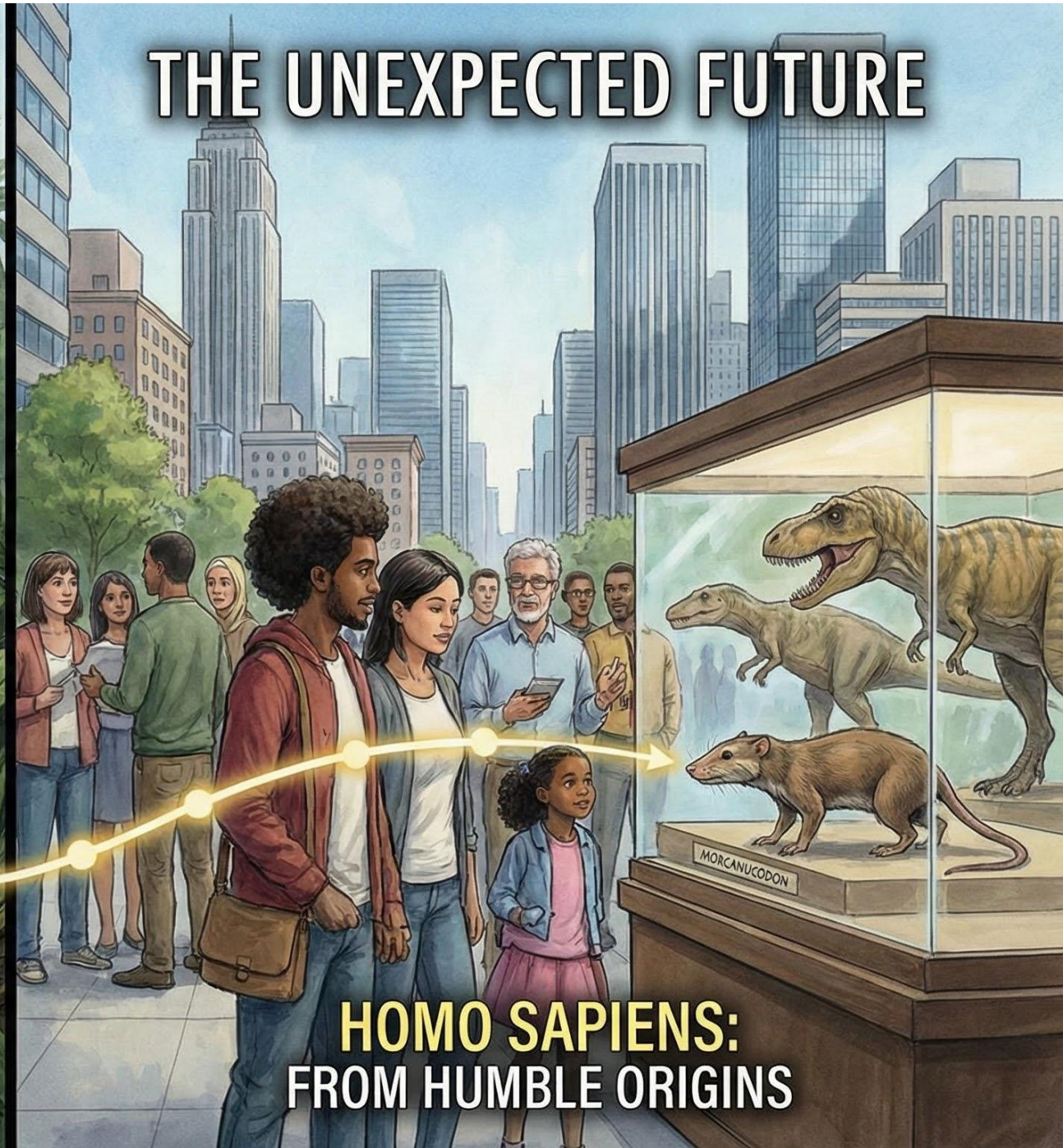


# THE MESOZOIC SHADOW



**MORGANUCODON:**  
ANCESTOR IN THE SHADOWS

# THE UNEXPECTED FUTURE



**HOMO SAPIENS:**  
FROM HUMBLE ORIGINS

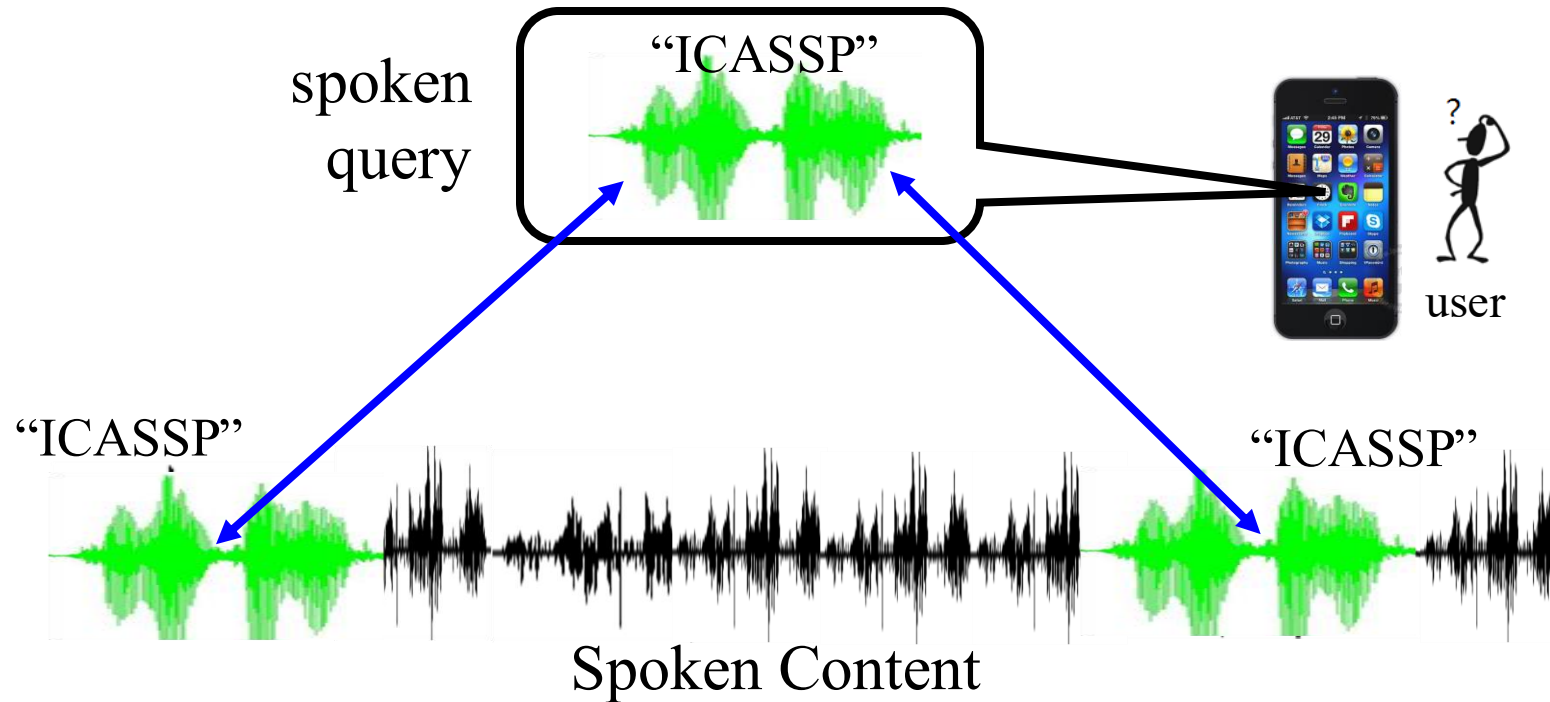




# 1. 序章



# Spoken Content Retrieval

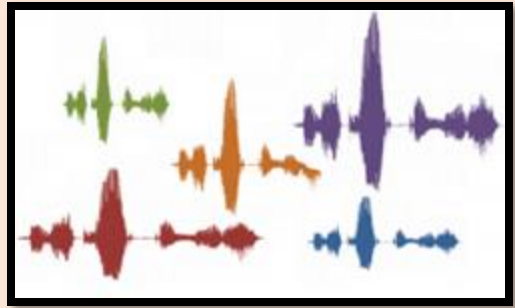


Compute similarity between spoken queries and audio files on acoustic level, and find the query term

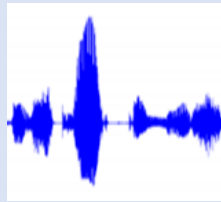
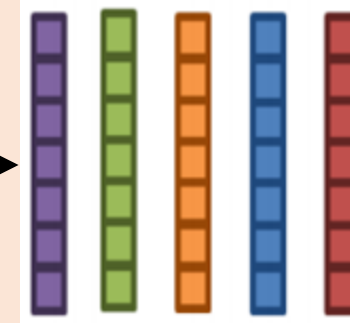


Audio archive divided into variable-length audio segments

*Off-line*



Audio Word  
to Vector



Spoken  
Query

Audio Word  
to Vector



Similarity

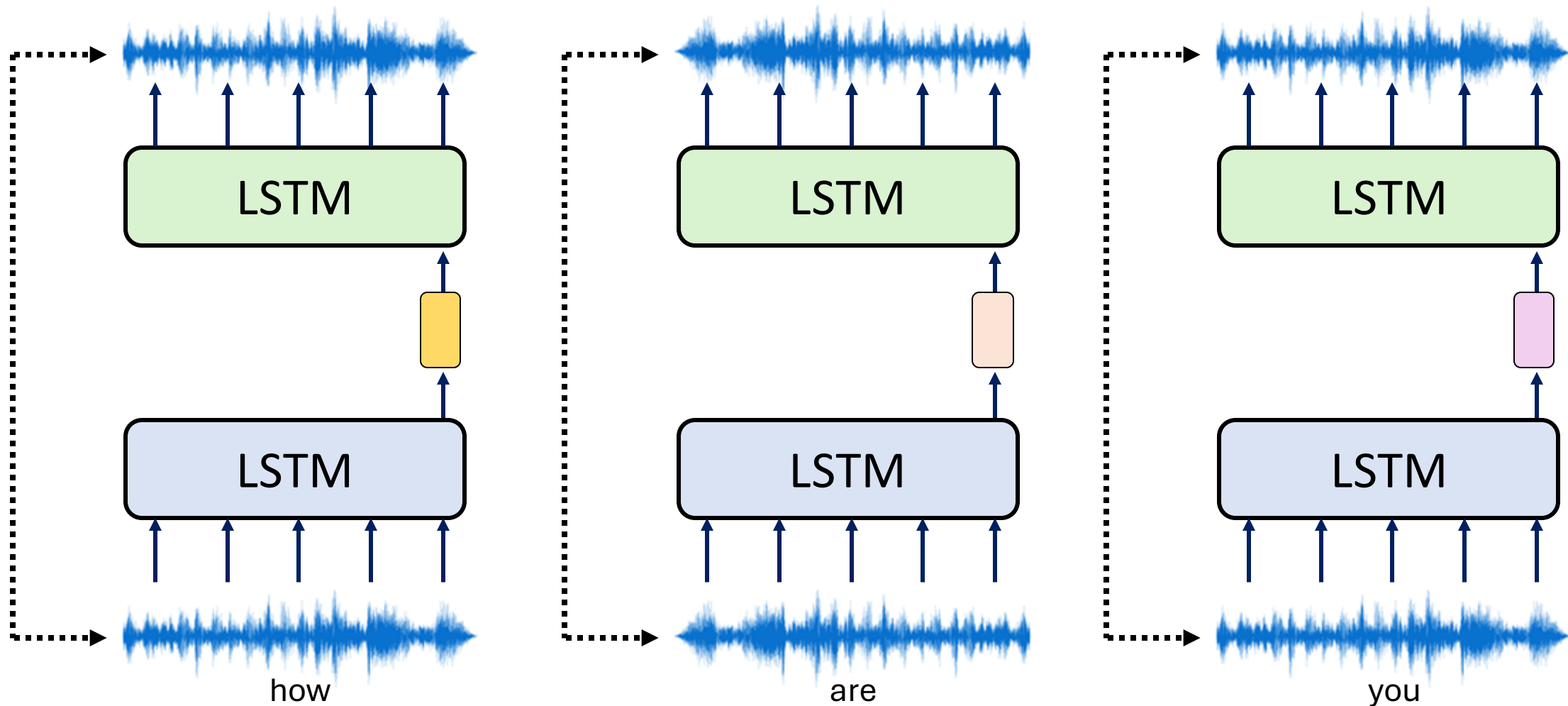
Search Result

*On-line*

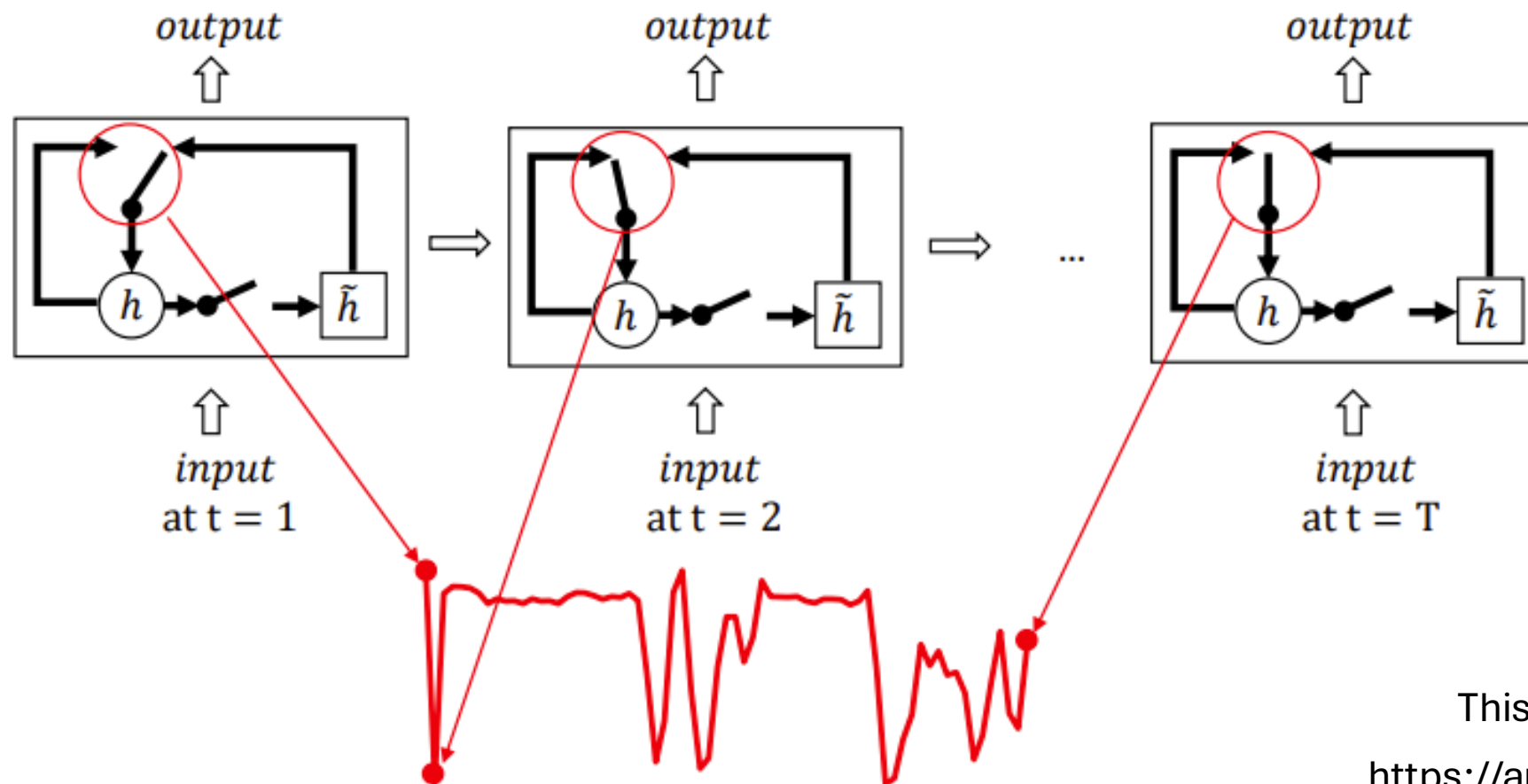
# Audio Word2Vector

<https://arxiv.org/abs/1603.00982>

This work is done by 鍾毓安.



# How can we automatically find certain types of boundaries in speech?



This work is done by 王育軒  
<https://arxiv.org/abs/1703.07588>

# Segmental Audio Word2Vector

3 1 4 | 1 5 9 | 2 7 1 | 8 3 1 | 4 1 5 | 9 2 7 | 1 8 3 | 1 4 1 | 5 9

3 1 4 1 5 9 | 2 7 1 8 | 3 1 4 1 5 9 | 2 7 1 8 | 3 1 4 1 5 9

A B A B A

# Joint Learning of Segmentation and Seq2seq Auto-encoder

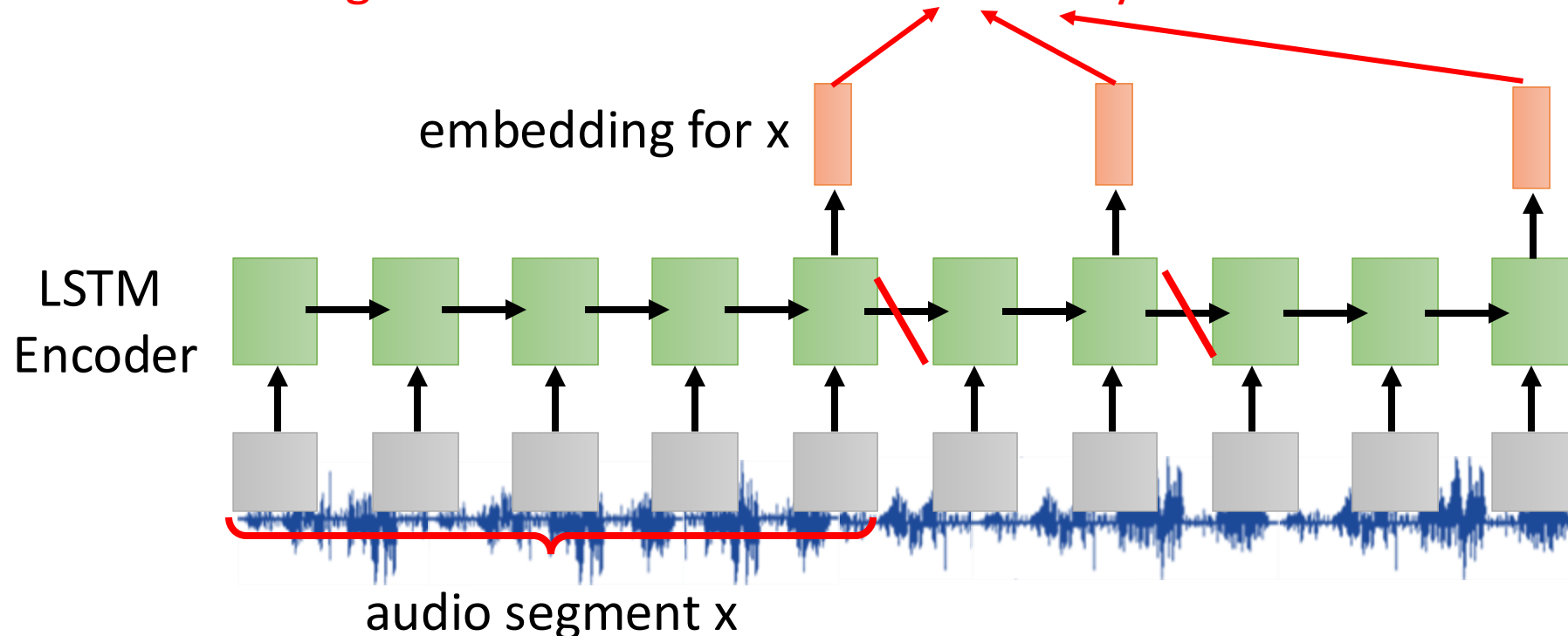
Segmental Audio Word2Vector

<https://arxiv.org/abs/1808.02228>

This work is done by 王育軒

- At each time step, LSTM encoder determines whether it is right before a boundary.
- If it is determined as right before a boundary, a vector (an embedding for an audio segment) is outputted.

Where to segment is determined automatically.



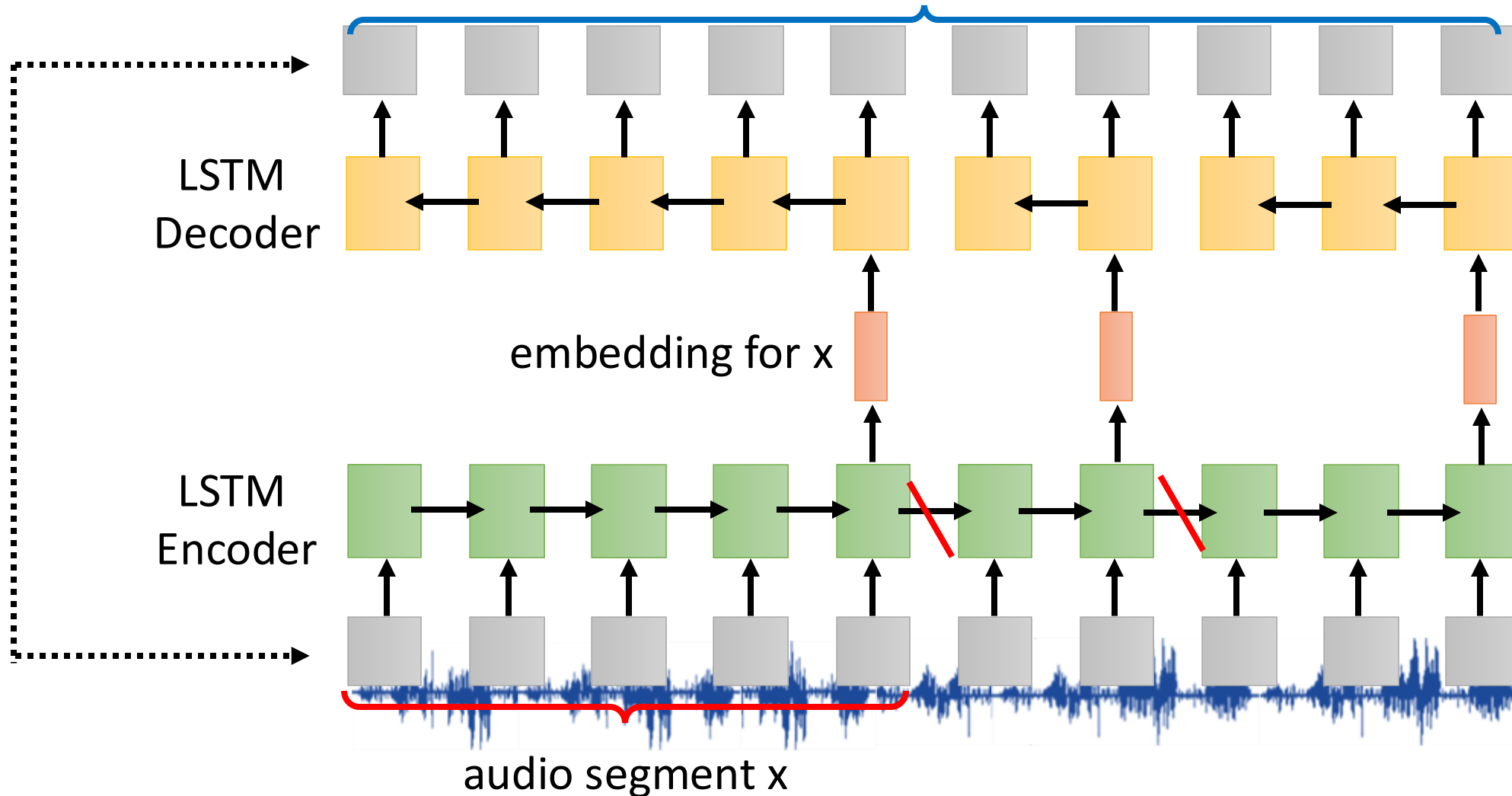
# Joint Learning of Segmentation and Seq2seq Auto-encoder

Segmental Audio Word2Vector

<https://arxiv.org/abs/1808.02228>

This work is done by 王育軒

LSTM decoder reconstructs the input utterance



# Joint Learning of Segmentation and Seq2seq Auto-encoder

Segmental Audio Word2Vector

<https://arxiv.org/abs/1808.02228>

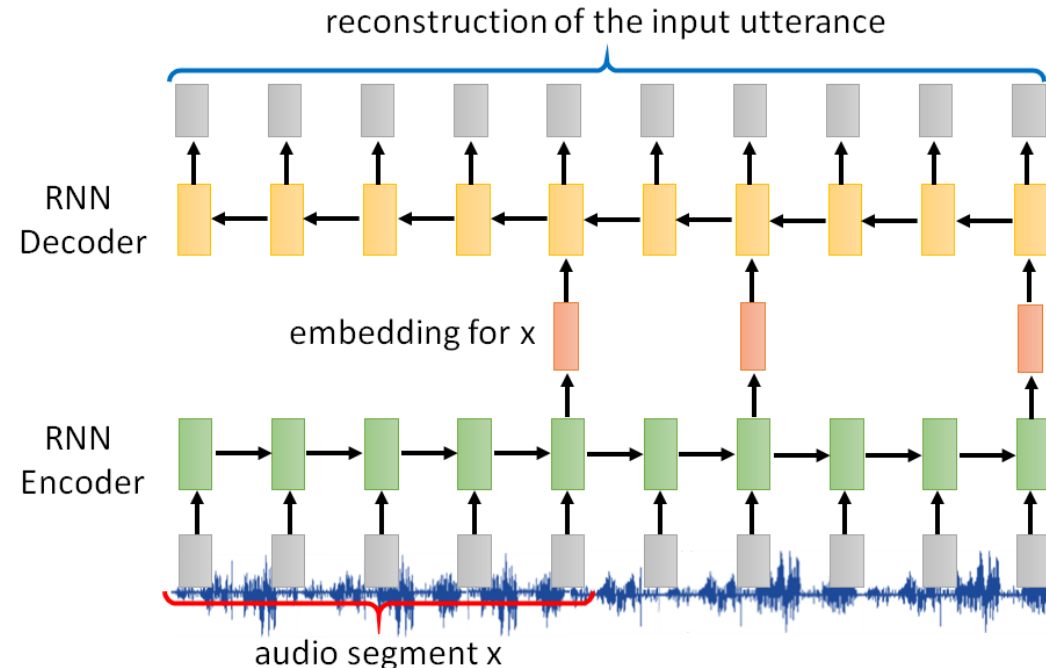
This work is done by 王育軒

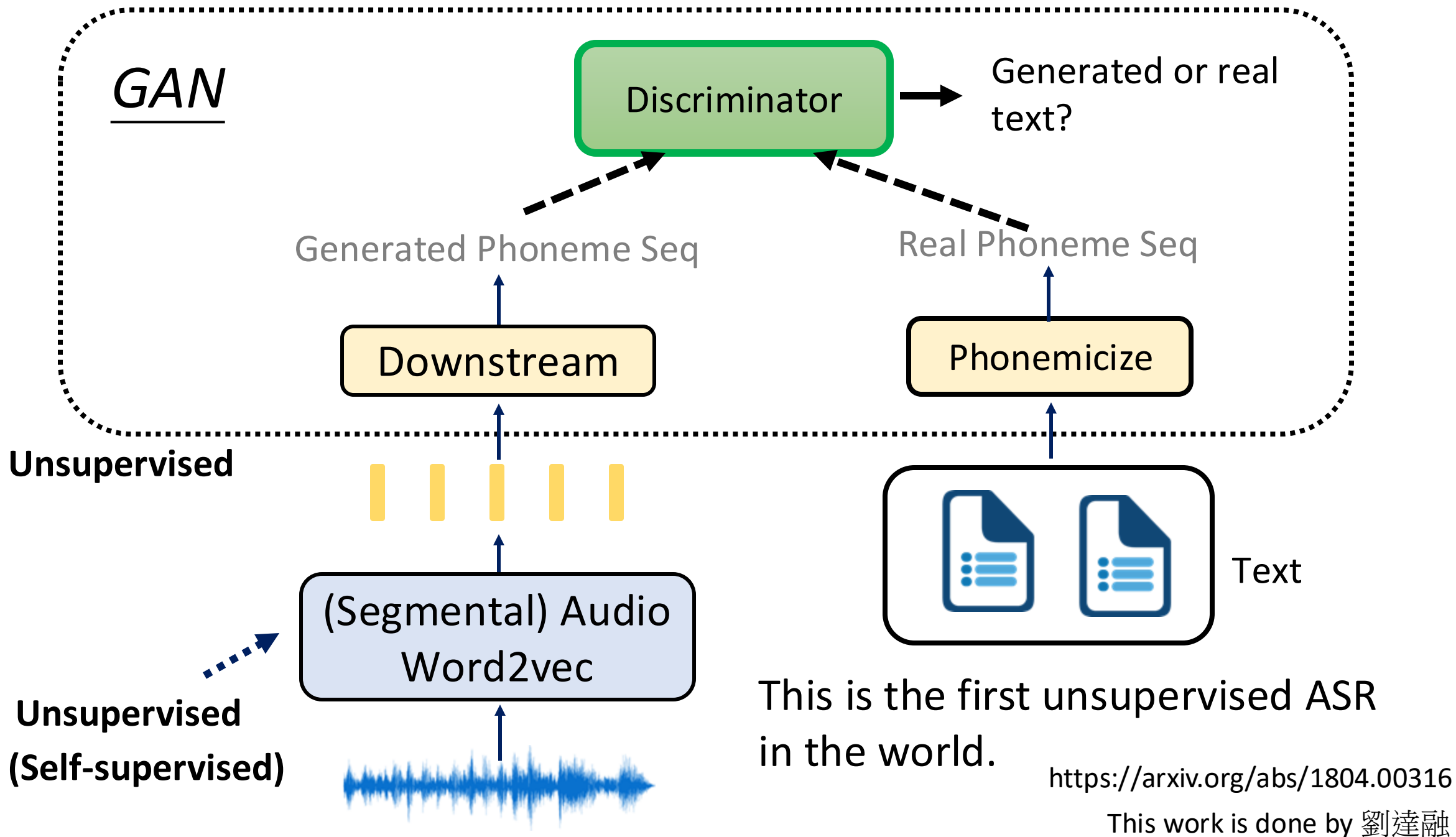
- The learning criterion of LSTM encoder and decoder is the weighted sum of the following two terms.
  - 1. Minimizing Reconstruction error
  - 2. Minimizing the number of segments, that is the number of output embedding

The **second term** is necessary.

If we only **minimize reconstruction error**,

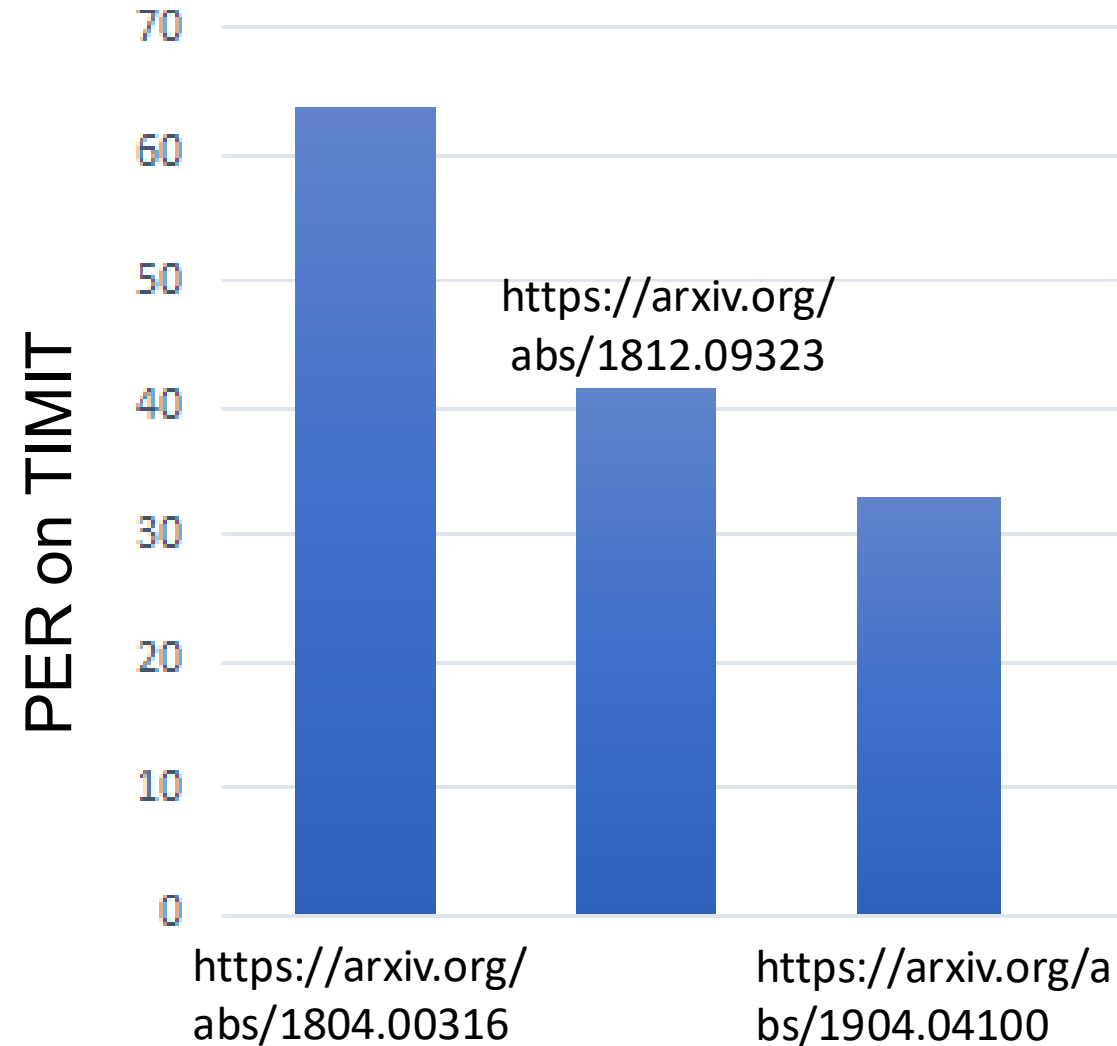
LSTM encoder would output embedding at all of the time steps.

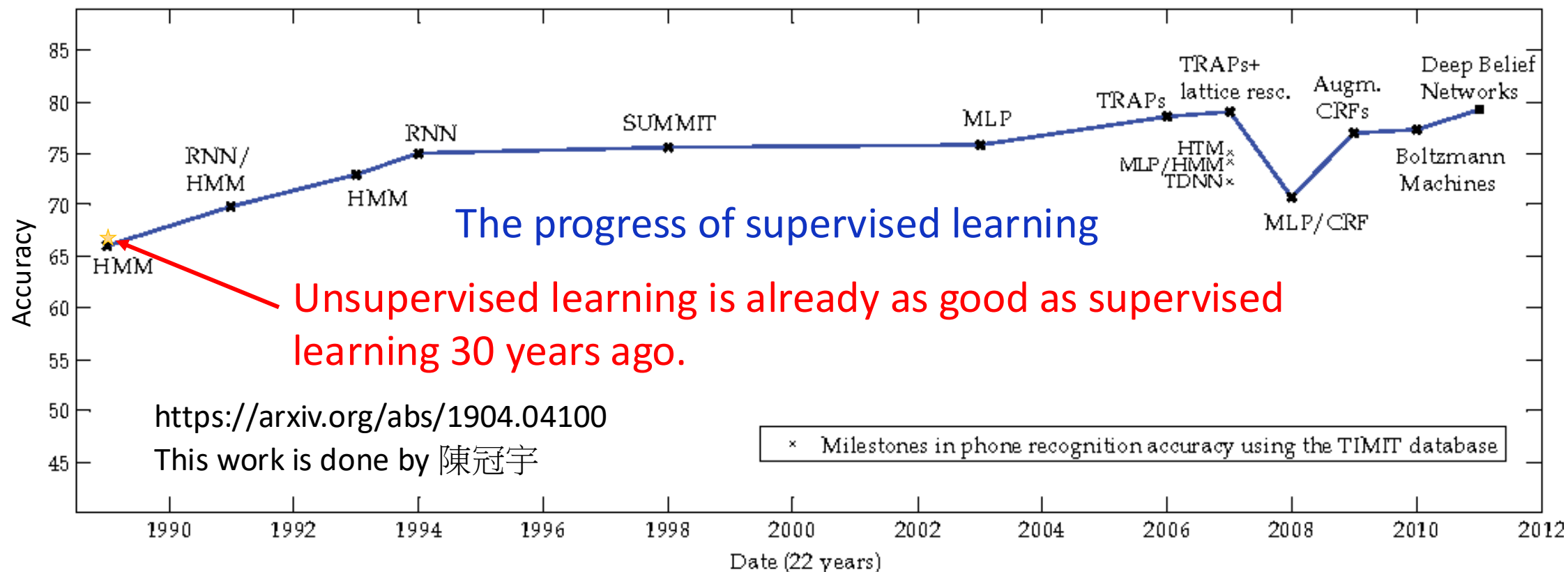






# Is Unsupervised Speech Recognition possible?





The image is modified from: Phone recognition on the TIMIT database Lopes, C. and Perdigão, F., 2011. Speech Technologies, Vol 1, pp. 285--302.

# Speech BERT

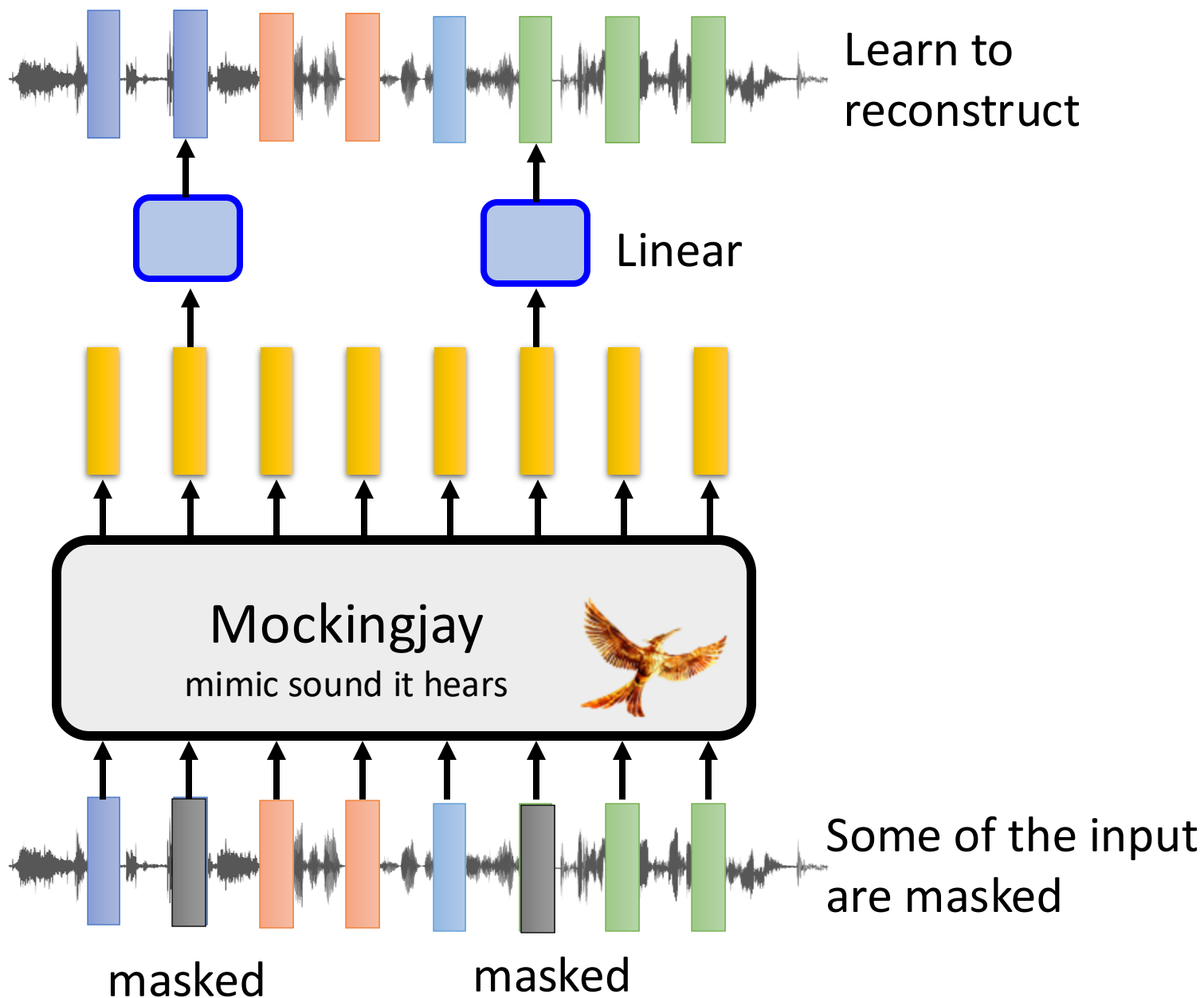
<https://arxiv.org/abs/1910.12638>



Andy T. Liu



Like BERT  
(also, like MaskGIT)

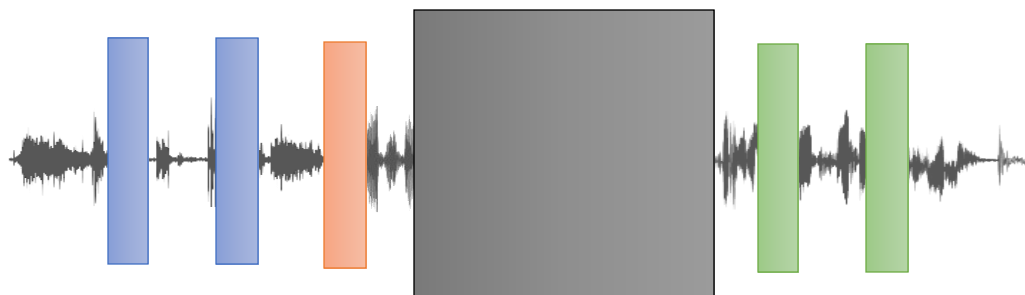


# Masking

- Smoothness of acoustic features

<https://arxiv.org/abs/1910.12638>

Masking consecutive features

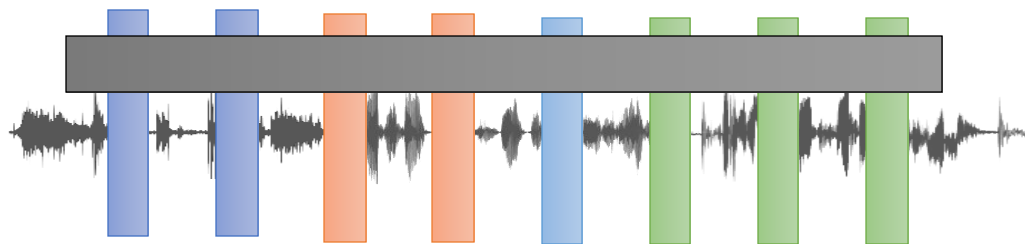


- Masking strategies for speech

Learn more speaker  
information in this way

TERA: <https://arxiv.org/abs/2007.06028>

Masking specific dimensions

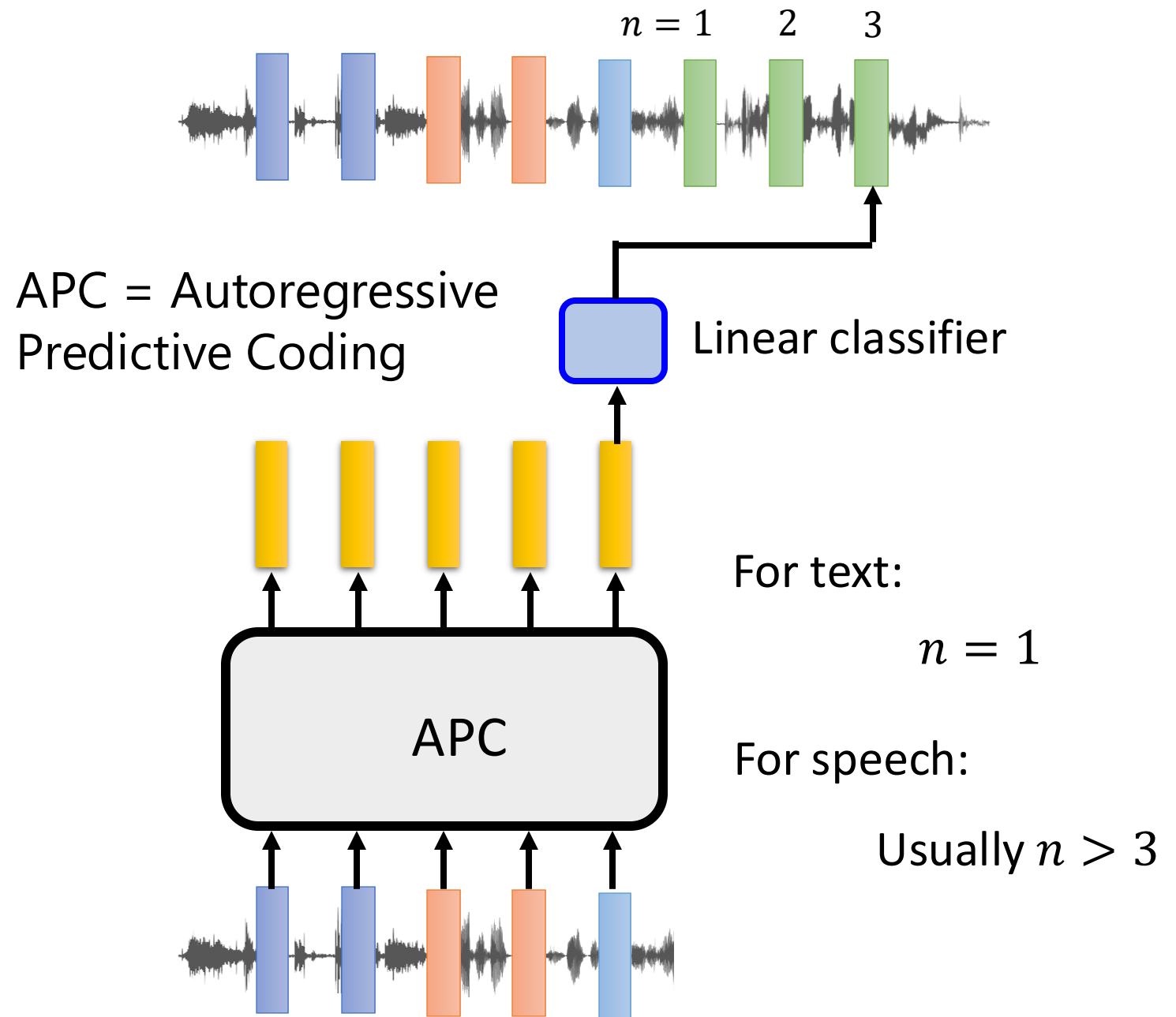


# Speech GPT

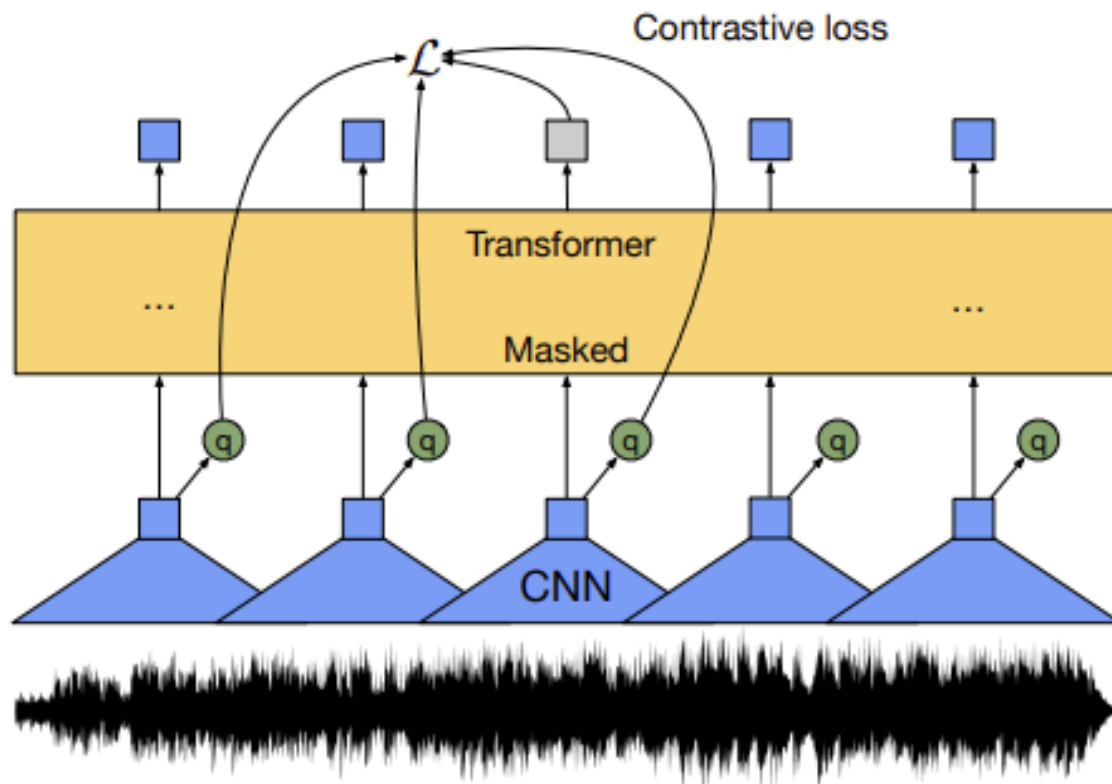
<https://arxiv.org/abs/1910.12607>



Similar to GPT



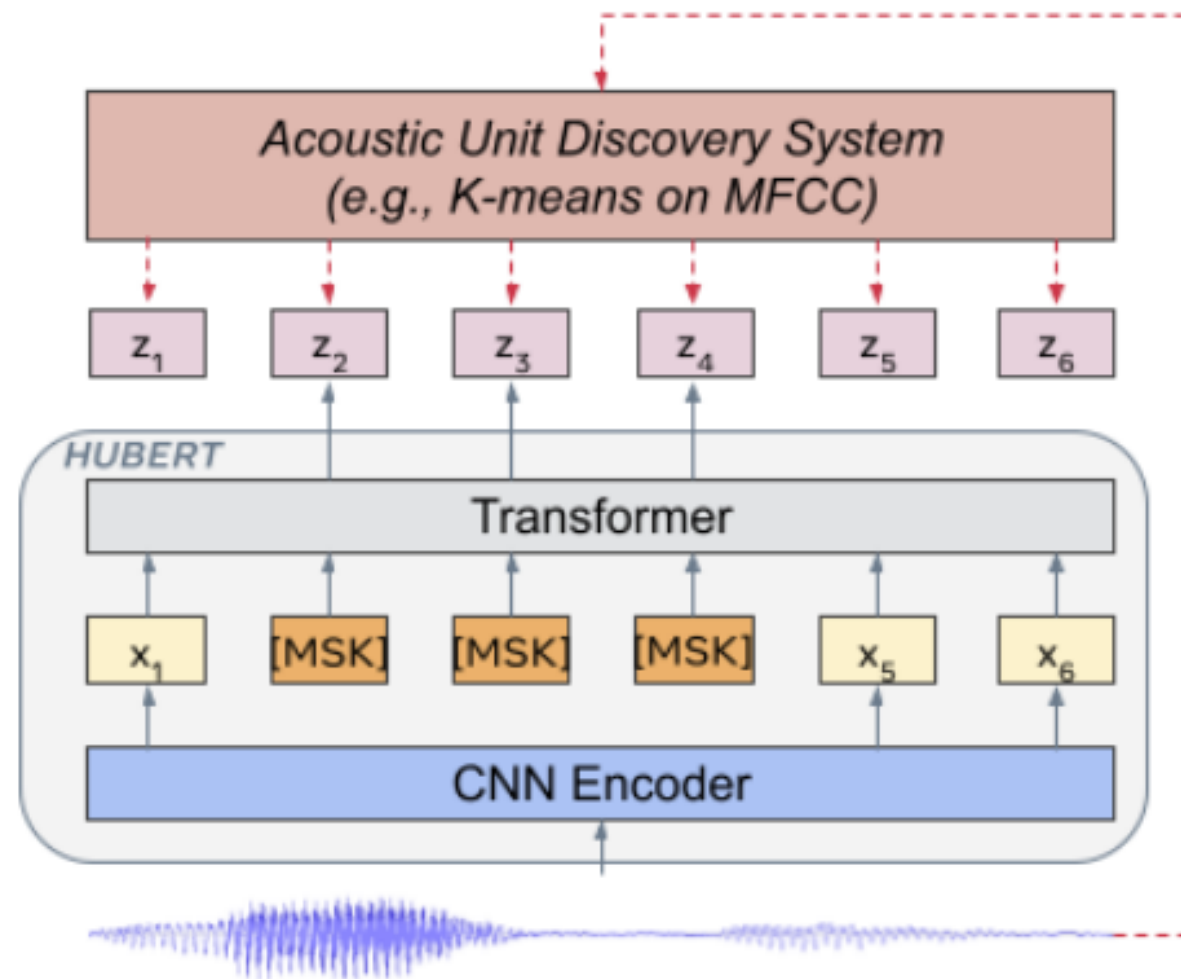
# Wav2vec / HuBERT



**Wav2vec**

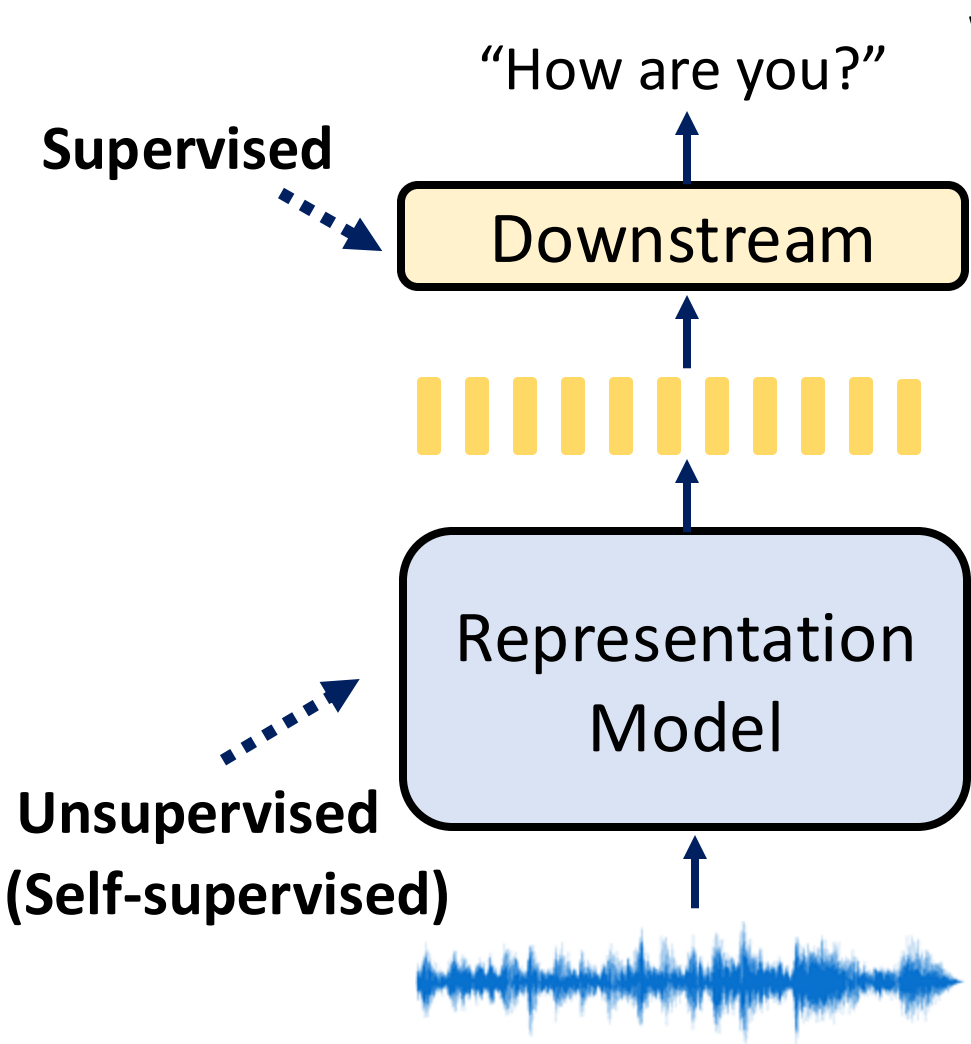
<https://arxiv.org/abs/1904.05862>

<https://arxiv.org/abs/2006.11477>

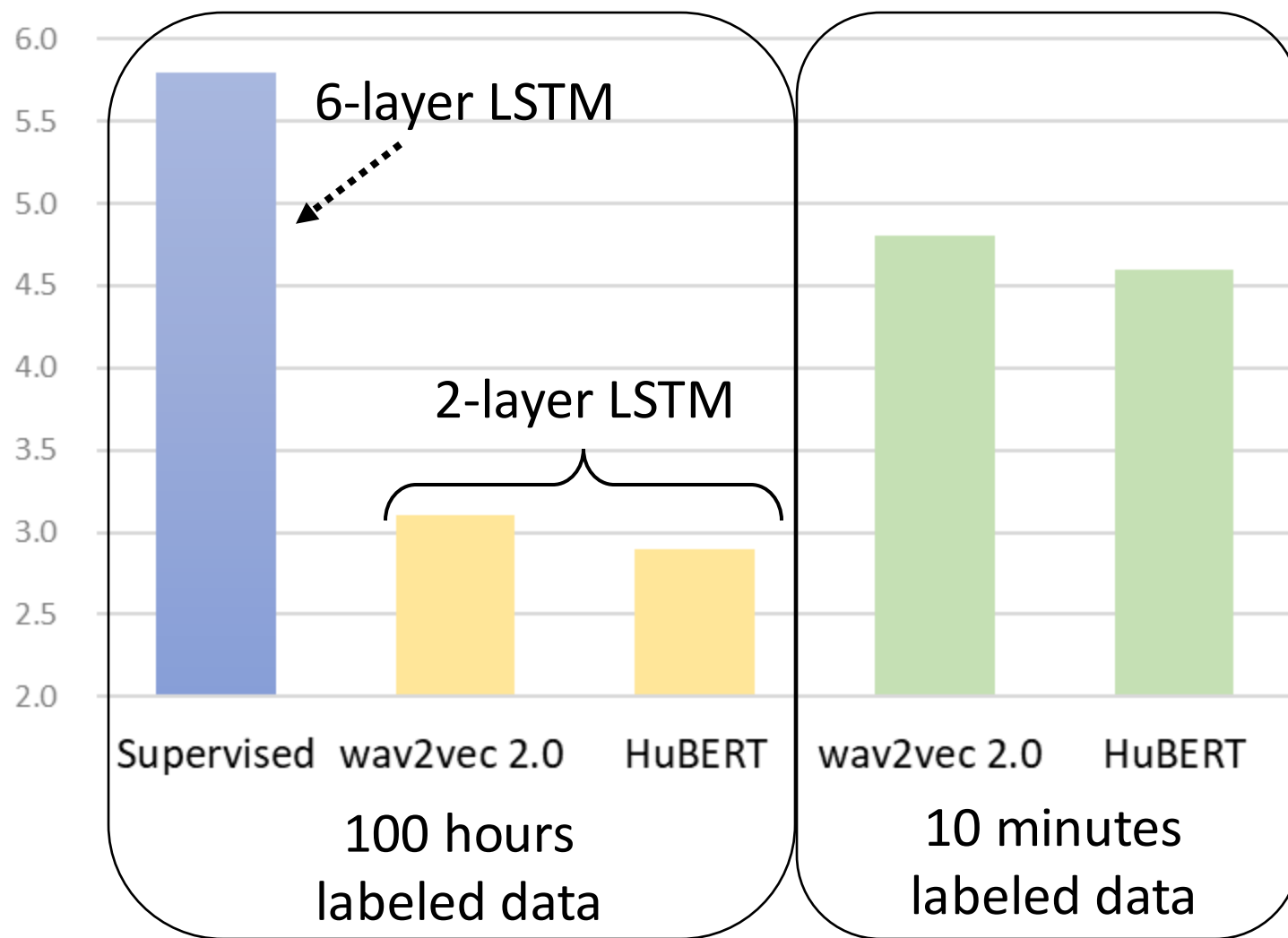


**HuBERT**

<https://arxiv.org/abs/2106.07447>



WER (Librispeech)



<https://arxiv.org/abs/1905.03072>

<https://arxiv.org/abs/2006.11477>

<https://arxiv.org/abs/2106.07447>

<https://arxiv.org/abs/2105.01051>

GAN

Discriminator

Generated or real  
text?

Generated Phoneme Seq

Real Phoneme Seq

Downstream

Phonemicize

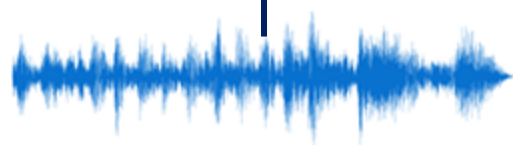
~~Segmental Audio  
Word2vec~~



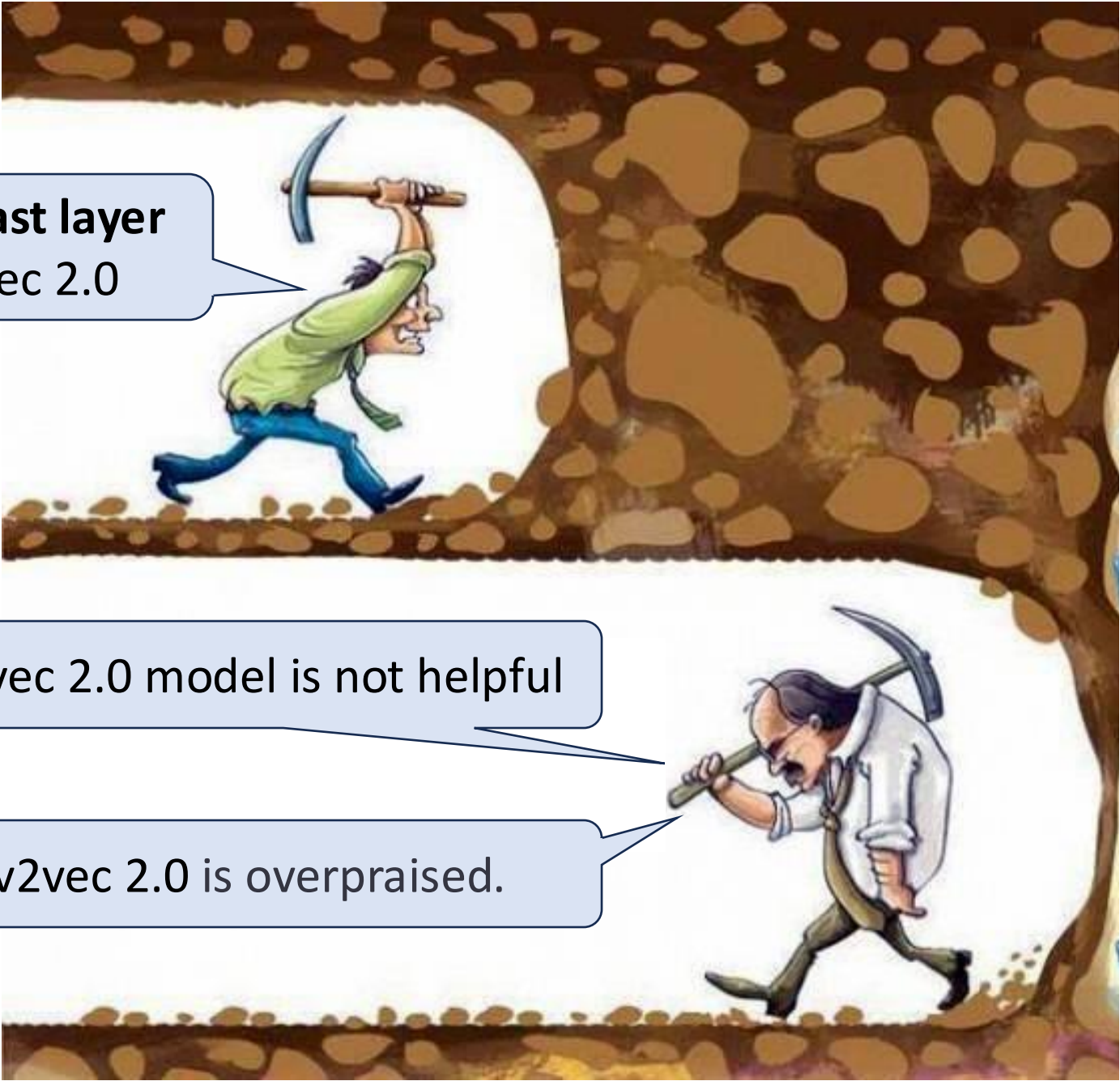
Text

Wav2vec 2.0

Unsupervised  
(Self-supervised)





A cartoon illustration of a man in a green shirt and blue pants running through a cave. He is holding a pickaxe over his shoulder. The cave walls are brown and rocky. A speech bubble points to him.

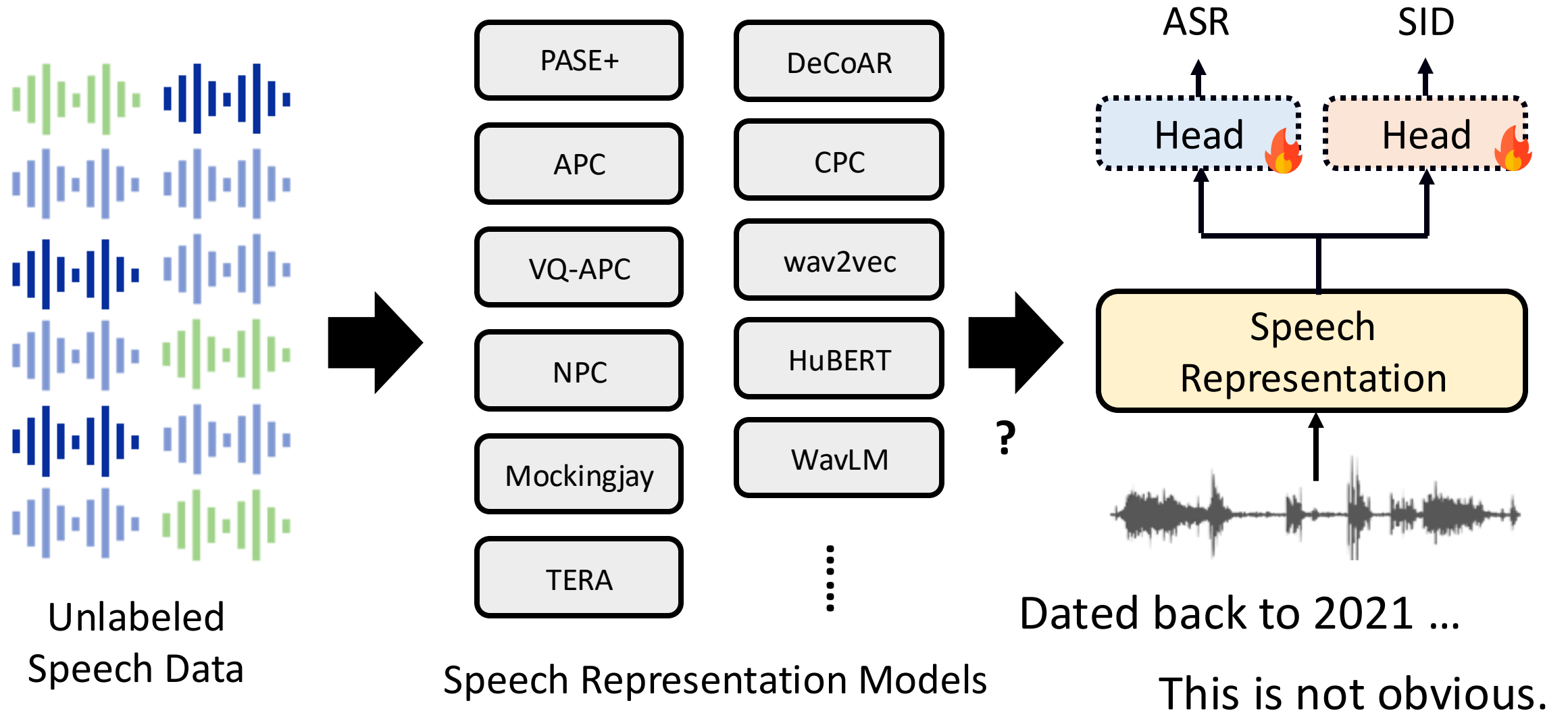
Using the **last layer**  
of wav2vec 2.0

wav2vec 2.0 model is not helpful

wav2vec 2.0 is overpraised.

Groundbreaking  
performance of  
Unsupervised ASR

# Are speech representation models universal?





# SUPERB

## Speech processing Universal PERformance Benchmark



Shu-wen  
(Leo) Yang

### **SUPERB: Speech processing Universal PERformance Benchmark**

*Shu-wen Yang<sup>1</sup>, Po-Han Chi<sup>1\*</sup>, Yung-Sung Chuang<sup>1\*</sup>, Cheng-I Jeff Lai<sup>2\*</sup>, Kushal Lakhota<sup>3\*</sup>,  
Yist Y. Lin<sup>1\*</sup>, Andy T. Liu<sup>1\*</sup>, Jiatong Shi<sup>4\*</sup>, Xuankai Chang<sup>6</sup>, Guan-Ting Lin<sup>1</sup>,  
Tzu-Hsien Huang<sup>1</sup>, Wei-Cheng Tseng<sup>1</sup>, Ko-tik Lee<sup>1</sup>, Da-Rong Liu<sup>1</sup>, Zili Huang<sup>4</sup>, Shuyan Dong<sup>5†</sup>,  
Shang-Wen Li<sup>5†</sup>, Shinji Watanabe<sup>6</sup>, Abdelrahman Mohamed<sup>3</sup>, Hung-yi Lee<sup>1</sup>*



Shang-Wen  
Li



Abdelrahman  
Mohamed



Shinji  
Watanabe



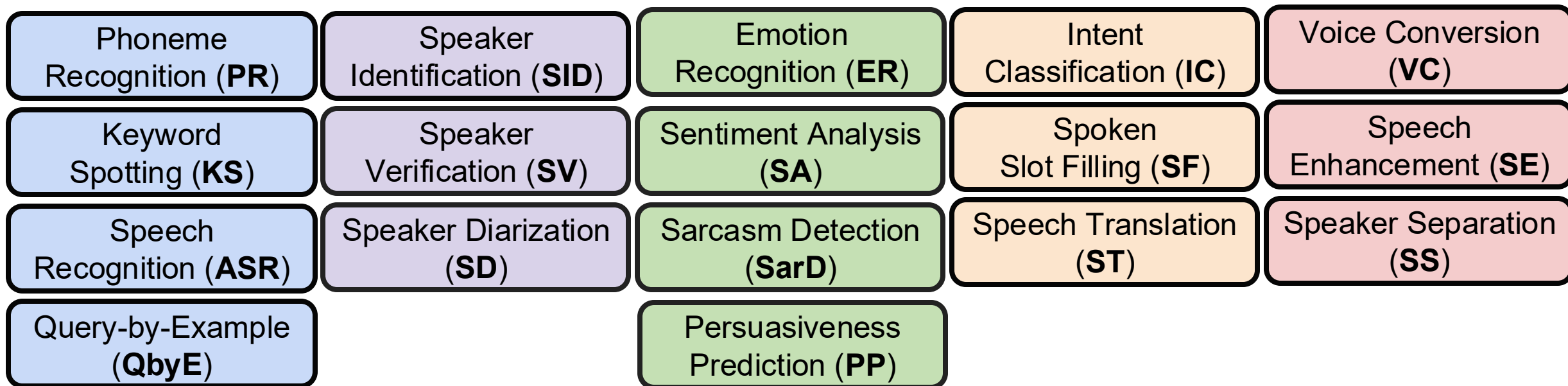
Hung-yi  
Lee



# SUPERB

## Speech processing Universal PERformance Benchmark

Evaluate a wide range of speech representation models on many speech tasks



Content



Speaker



Prosody



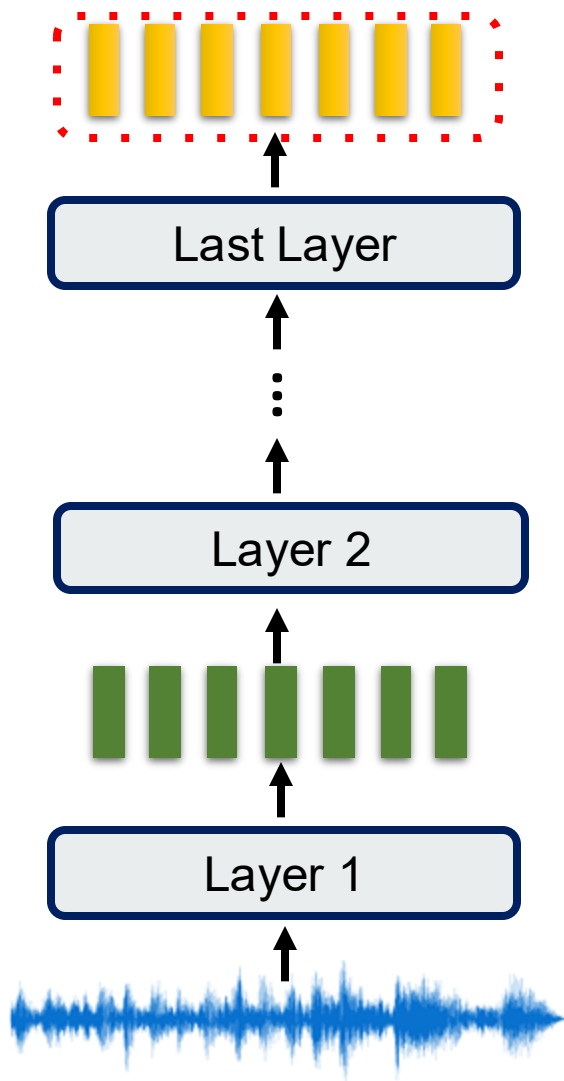
Semantic



Synthesis

<https://arxiv.org/abs/2105.01051>

<https://arxiv.org/abs/2203.06849>



Does not always lead to decent performance.

Most speech representation models are worse than acoustic features (fbank) for speaker verification.

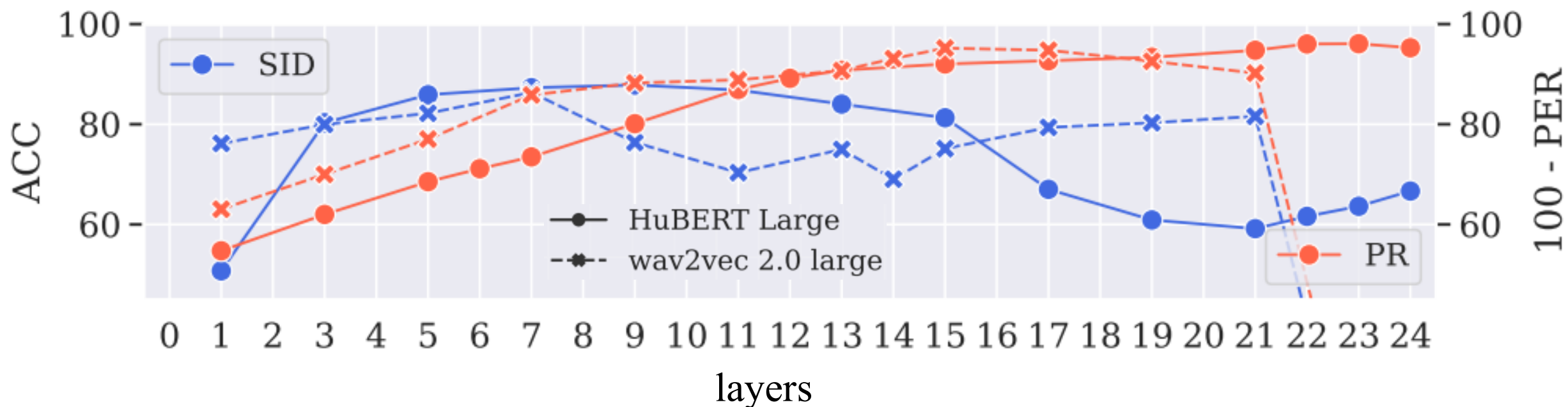


Hung-yi Lee

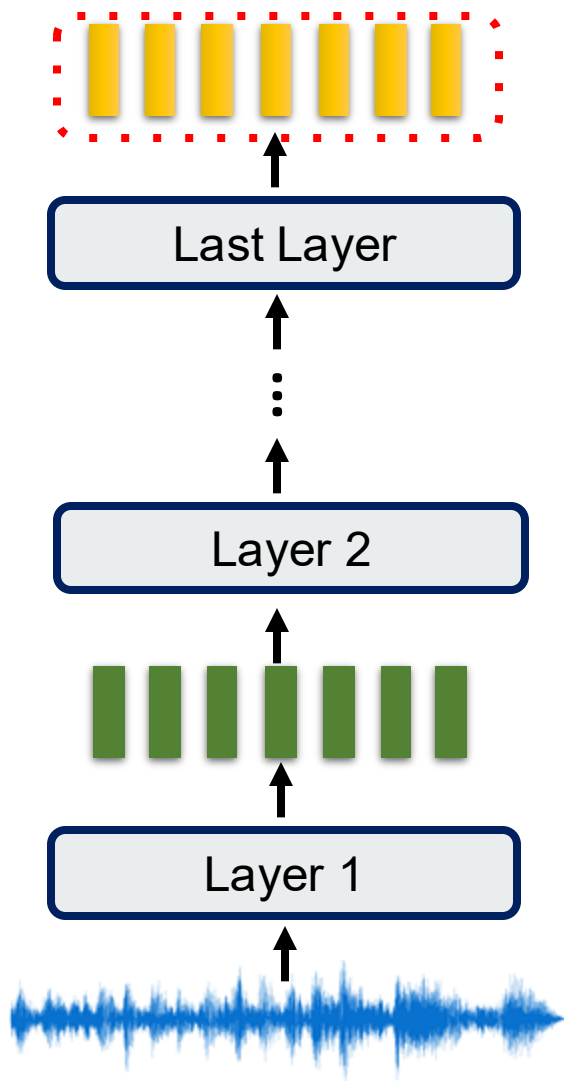
I knew it. Speech representation models cannot be universal.

Large models such as **wav2vec-large** and **HuBERT-large** perform as poorly as fbank on some tasks.

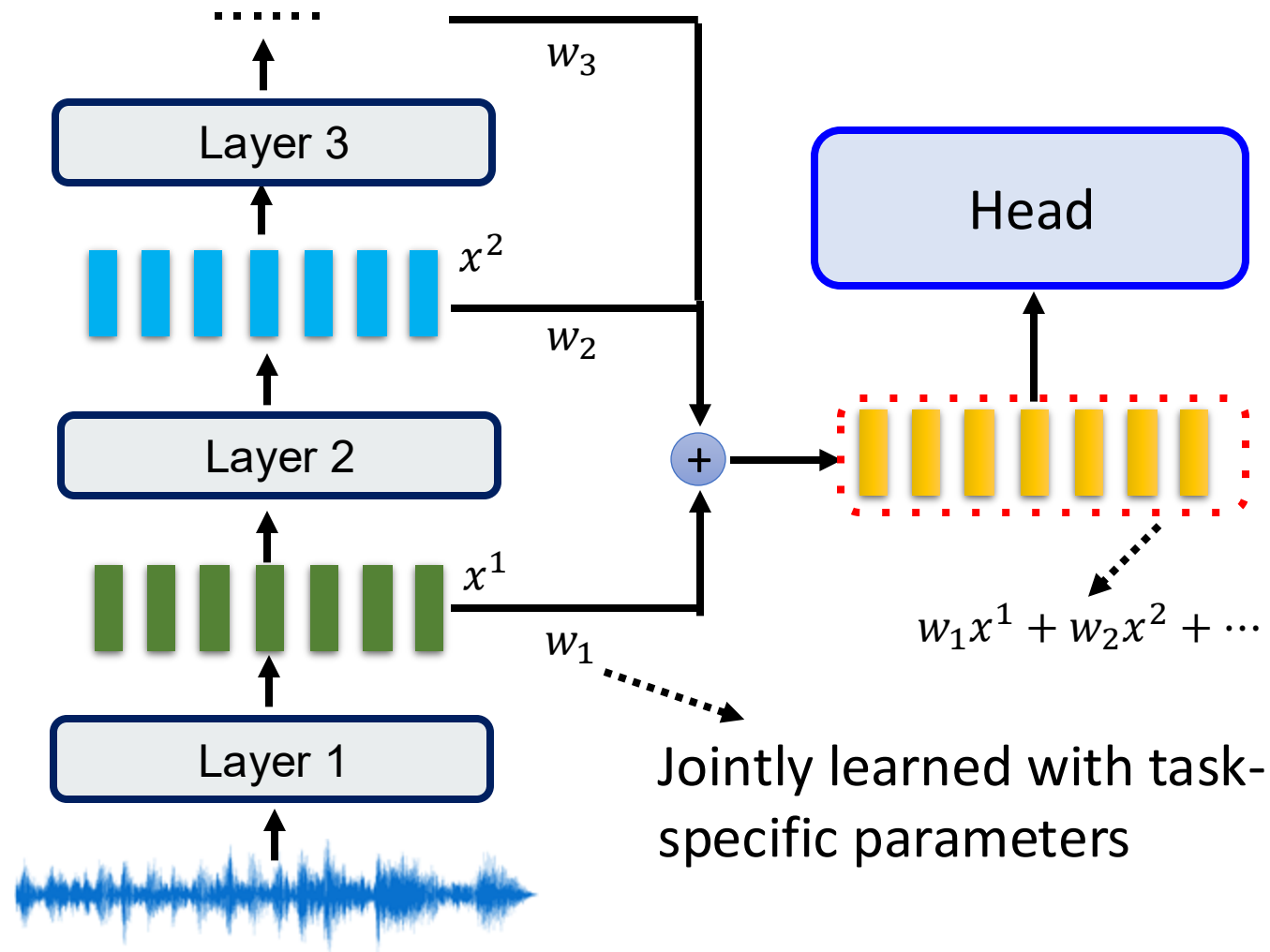
# Different Layers Encode Different Information



- **PR** (phoneme recognition) - **phonetic** information in the last few layers
- **SID** (speaker identification) - **speaker** information in the first few layers
- The last layer of Wav2vec 2.0 large encodes either phonetic and speaker information.



Does not always lead to decent performance.



“Weighted-sum” is very effective!



	PASE+	58.87	82.54	29.82	37.99	57.86	25.11	61.12	51.41	34.66	70.18	88.25	62.14	60.18	0.72	11.61	8.68	3.16	8.66	30.60	63.20	2.56	93.90	9.87
	APC	41.98	91.01	74.69	60.42	59.33	21.28	62.34	54.03	35.05	69.95	90.32	70.46	50.90	3.10	8.57	10.53	5.95	8.05	27.20	87.25	2.56	93.40	8.92
	VQ-APC	41.08	91.11	74.48	60.15	59.66	21.20	62.85	54.14	35.80	71.04	90.42	68.53	52.91	2.51	8.72	10.45	4.23	7.84	22.40	94.25	2.56	93.40	8.44
	NPC	43.81	88.96	69.44	55.92	58.21	20.20	60.75	49.68	34.43											52	93.10	8.04	
	Mockingjay	70.19	83.67	34.33	32.29	50.28	22.82	65.01	58.28	36.87											53	93.40	9.29	
	TERA	49.17	89.48	58.42	66.62	56.27	18.17	57.89	48.56	30.80											54	93.60	10.19	
	Audio Albert	49.71	86.01	53.10	56.51	56.52	20.43	61.30	52.14	34.82											61	93.80	9.93	
	DeCoAR	22.27	91.82	85.34	69.76	60.66	16.57	57.25	47.11	32.12											61	93.70	9.69	
	DeCoAR 2.0	14.93	94.48	90.80	74.42	62.48	13.02	52.56	41.98	27.27	60.01	80.99	83.28	34.73	4.06	7.16	6.59	9.94	7.83	17.10	90.75	2.47	93.20	8.54
	Modified CPC	42.54	91.89	64.09	39.63	60.96	20.18	61.82	53.45	34.92	68.13	90.78	71.19	49.91	3.26	12.86	10.38	4.82	8.41	26.20	71.00	2.57	93.70	10.40
	wav2vec	31.58	95.59	84.92	56.56	59.79	15.86	54.95	45.33	30.25	59.74	84.49	76.37	43.72	4.85	8.00	9.90	6.61	7.45	10.10	98.25	2.53	93.80	9.30
	vq-wav2vec	33.48	93.38	85.68	38.80	58.24	17.71	60.02	51.44	35.26	65.21	88.19	77.68	41.55	4.10	10.38	9.93	5.66	7.08	13.40	100.00	2.48	93.60	8.16
	FaST-VGS+	7.76	97.27	98.97	41.34	62.71	8.83	46.48	35.53	25.32	54.19	70.89	88.15	27.12	5.62	5.87	6.05	14.45	7.73	10.85	92.75	2.57	93.94	9.76
LightHuBERT	Stage1	4.15	96.82	98.50	80.01	66.25	5.71	43.72	34.69	24.32	52.82	63.07	88.44	25.92	7.37	5.14	5.51	16.41	7.74	13.12	94.50	2.59	94.11	9.93
	LightHuBERT Small	6.60	96.07	98.23	69.70	64.12	8.34	47.26	37.48	26.34	54.89	70.33	87.58	26.90	7.64	5.42	5.85	13.83	7.34	9.38	98.25	2.54	93.85	9.45
	DPHuBERT	9.67	96.36	97.92	76.83	63.16	10.47	48.98	38.46	26.09	56.12	75.25	86.86	28.26	6.93	5.84	5.92	12.00	7.63	10.28	96.25	2.56	93.97	9.65
	DPWavLM	8.22	96.27	98.58	82.11	65.24	10.19	46.29	33.99	24.71												7.47		
	CoBERT Base	3.08	96.36	98.87	72.66	65.32	4.74	44.36	34.27	24.51													8.25	
Table credit:	Shuwen Yang																							
deeplearning4j	Shuwen Yang	4.14	96.55	98.92	59.87	67.59	5.39	42.39	32.60	23.62	52.44	60.92	89.39	22.88	6.65	5.82	4.84	18.74	7.23	9.68	99.25	2.59	94.13	10.59
	CCC-wav2vec 2.0	5.95	96.72	96.47	72.84	64.17	6.30	41.79	31.96	22.71	50.65	61.84	88.08	24.34	6.73	5.61	4.27	16.20	7.71	11.22	94.50	2.62	94.16	10.86

- Row - SSL model
- Column – Task
- Value – The darker, the better

Speech SSL models can be universal!



## Most cited Interspeech papers in last 5 years

1. Conformer: Convolution-augmented Transformer for Speech Recognition.  
Gulati et al. (2020)
2. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification.  
Desplanques et al. (2020)
3. AST: Audio Spectrogram Transformer.  
Gong et al. (2021)
4. SUPERB: Speech Processing Universal PERFORMANCE Benchmark  
Yang et al. (2021)
5. Unsupervised Cross-Lingual Representation Learning for Speech Recognition.  
Conneau et al. (2021)
6. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.  
Babu et al. (2022)
7. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement.  
Hu et al. (2020)
8. Real Time Speech Enhancement in the Waveform Domain  
Défossez et al. (2020)
9. MLS: A Large-Scale Multilingual Dataset for Speech Research  
Pratap et al (2020)
10. In Defence of Metric Learning for Speaker Recognition  
Chung et al (2020)

INTERSPEECH is one of the top three internationally recognized conferences in the field of speech processing.

mid

# The SUPERB Universe

## AV-SUPERB: A MULTI-TASK EVALUATION BENCHMARK FOR AUDIO-VISUAL REPRESENTATION MODELS

*Yuan Tseng<sup>1</sup>, Layne Berry<sup>2\*</sup>, Yi-Ting Chen<sup>3\*</sup>, I-Hsiang Chiu<sup>1\*</sup>, Hsuan-Hao Lin<sup>1\*</sup>, Max Liu<sup>1\*</sup>,  
Dongdong Chen<sup>2\*</sup>, Wei Fan<sup>3,4,1\*</sup>, Huihui Wu<sup>1\*</sup>, Huihui Wu<sup>1\*</sup>, Deyu Huang<sup>4</sup>, Chun-Mao Lai<sup>1</sup>,  
Shih-Luen Feng<sup>1</sup>, Hung-yi Lee<sup>1</sup>*

## ML-SUPERB: Multilingual Speech Universal PERFORMANCE Benchmark

*Jiatong Shi<sup>1</sup>, Dan Berrebbi<sup>1\*</sup>, William Chen<sup>1\*</sup>, Ho-Lam Chung<sup>2\*</sup>, En-Pei Hu<sup>2\*</sup>, Wei Ping Huang<sup>2\*</sup>,  
Xuankai Chang<sup>1</sup>, Shang-Wen Li<sup>3</sup>, Abdelrahman Mohamed<sup>1</sup>, Hung-yi Lee<sup>2</sup>, Shinji Watanabe<sup>1</sup>* <sup>1</sup>at Austin, USA

<sup>1</sup>Carnegie Mellon University, {jiatongs, dberrebbi, wchen, hchung, ehu, wphuang, xuankai, shangwen, amohamed, hylee, shwatanabe}@cmu.edu

## DYNAMIC-SUPERB: TOWARDS A DYNAMIC, COLLABORATIVE, AND COMPREHENSIVE INSTRUCTION-TUNING BENCHMARK FOR SPEECH

*Chien-yu Huang<sup>1</sup>, Ke-Han Lu<sup>\*1</sup>, Shih-Heng Wang<sup>\*1</sup>, Chi-Yuan Hsiao<sup>†1</sup>, Chun-Yi Kuan<sup>†1</sup>, Haibin Wu<sup>†1</sup>,  
Roshan Sharma<sup>2</sup>, Shinji Watanabe<sup>2</sup>,  
Junyi Peng<sup>3</sup>, Hung-yi Lee<sup>1</sup>*

## IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian languages

Conferences > ICASSP 2025 - 2025 IEEE Inter...

Tahir Javed<sup>1,2</sup> Kaushal Sa  
Anoop Kunchukuttan<sup>2,3</sup> Pr

## TS-SUPERB: A Target Speech Processing Benchmark for Speech Self-Supervised Learning Models

<sup>1</sup>Indian Institute of Technology  
<sup>2</sup>AI4BH

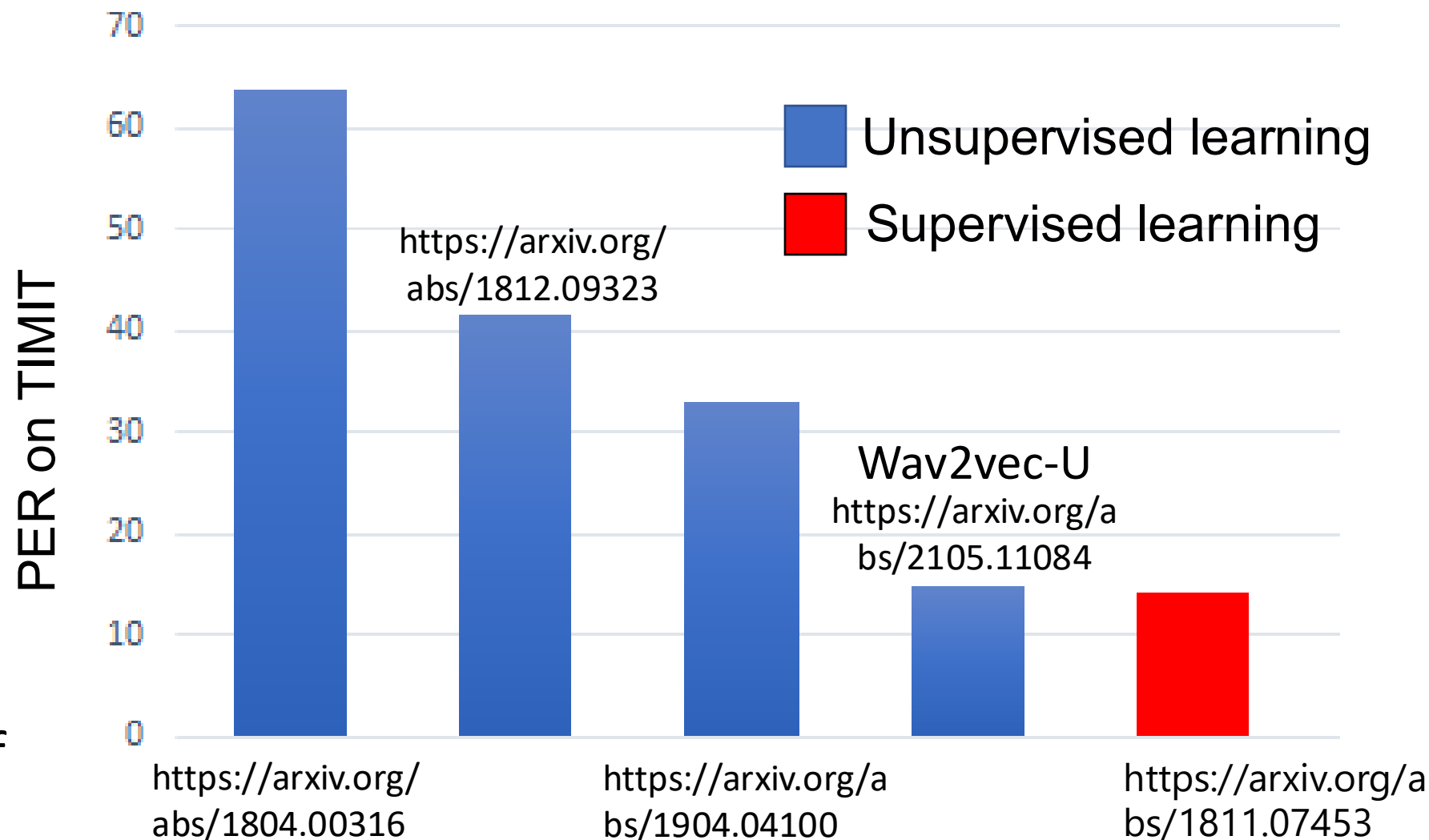
Publisher: IEEE

Cite This

PDF

Junyi Peng ; Takanori Ashihara ; Marc Delcroix ; Tsubasa Ochiai ; Oldrich Plchot ; Shoko Araki All Authors

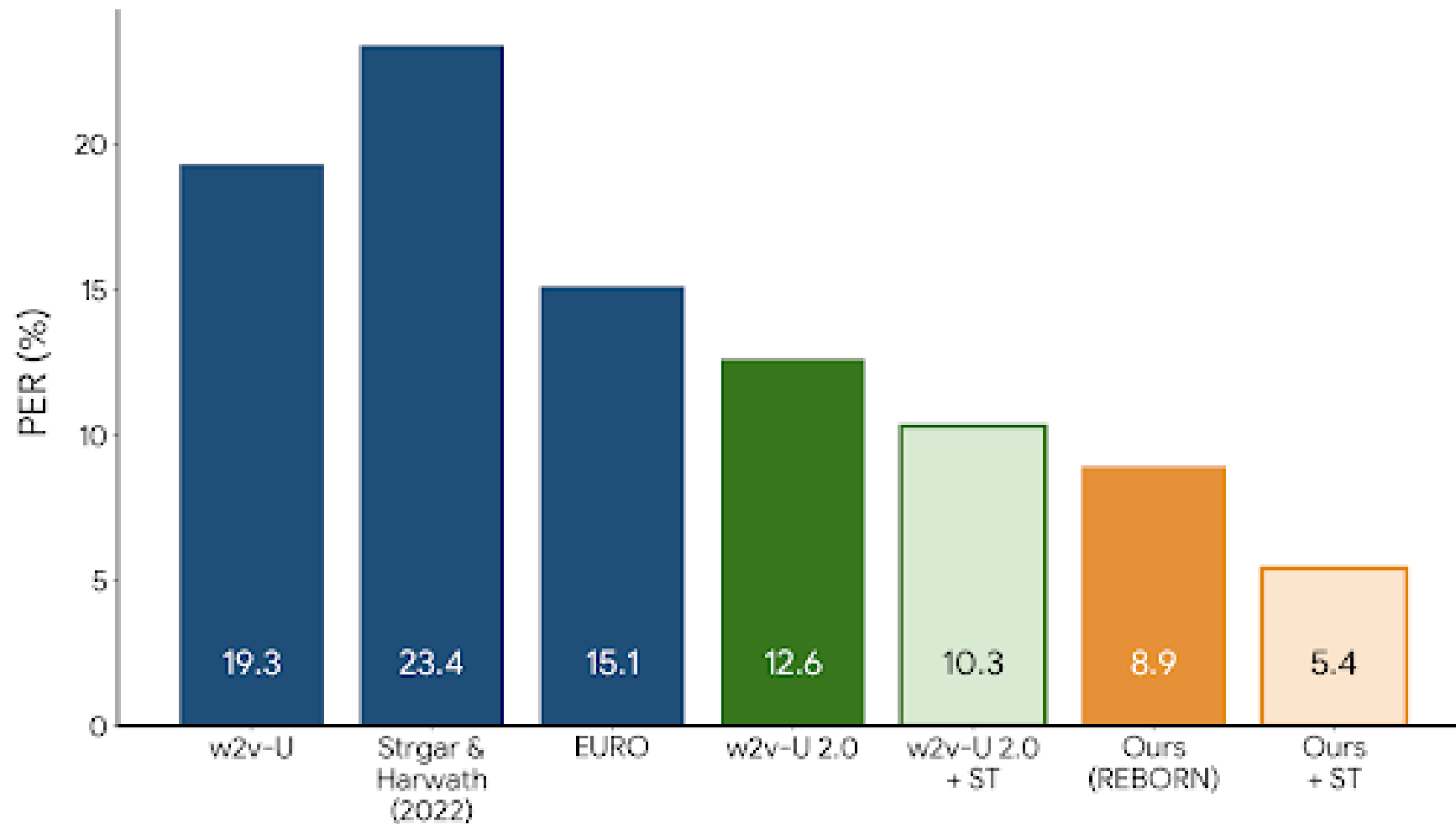
# Wav2vec-U: Unsupervised ASR with Wev2vec



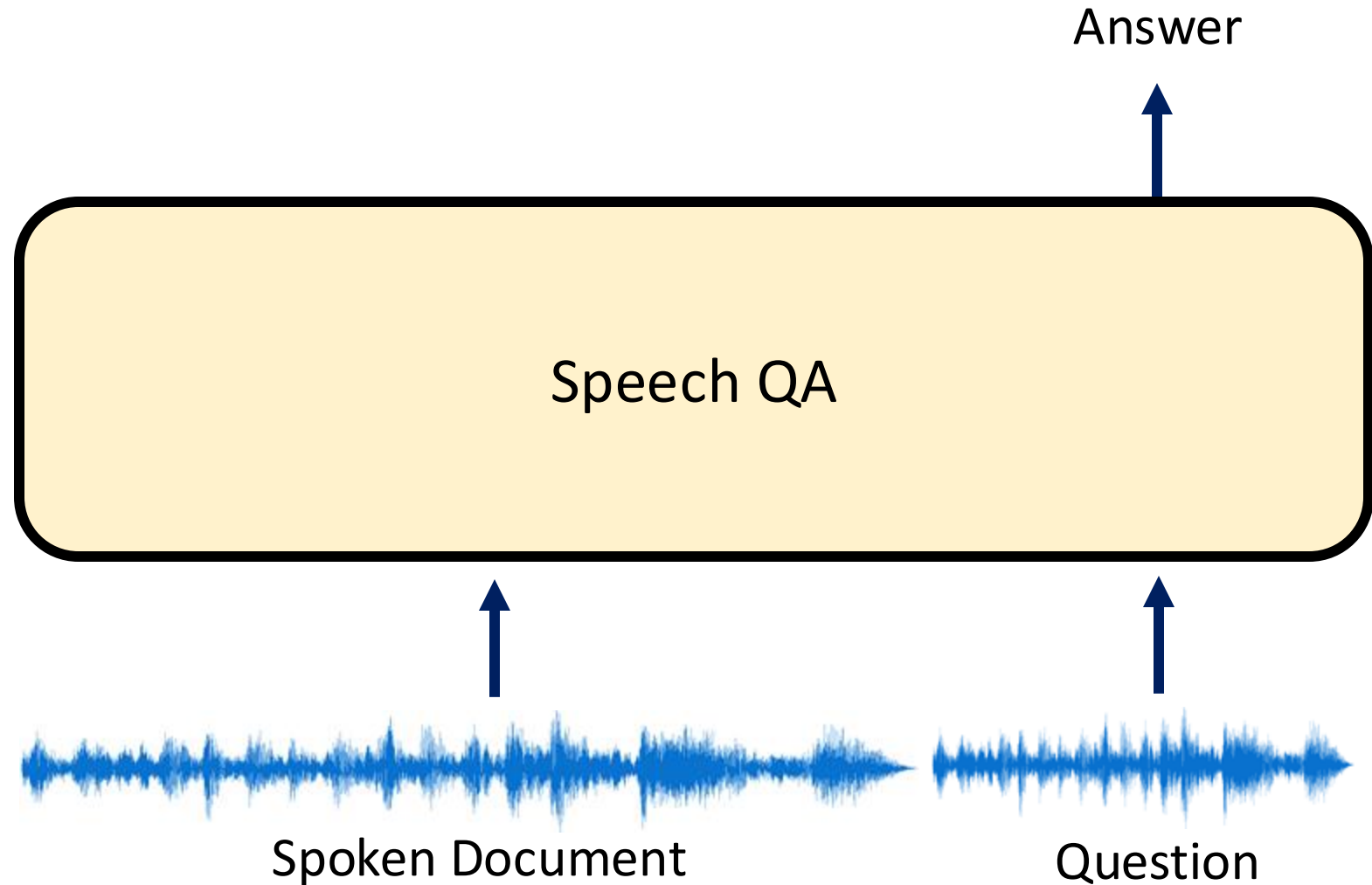
Wav2vec-U uses the **15th layer** instead of the **last layer**.

# REBORN <https://arxiv.org/abs/2402.03988>

- Librispeech
- training data:
  - 100 hours



# Listening Comprehension (Speech QA)



# Listening Comprehension

- **TOEFL Listening Comprehension Test by Machine**

Audio Story:  (The original story is 5 min long.)

Question: “ What is a possible origin of Venus’ clouds? ”

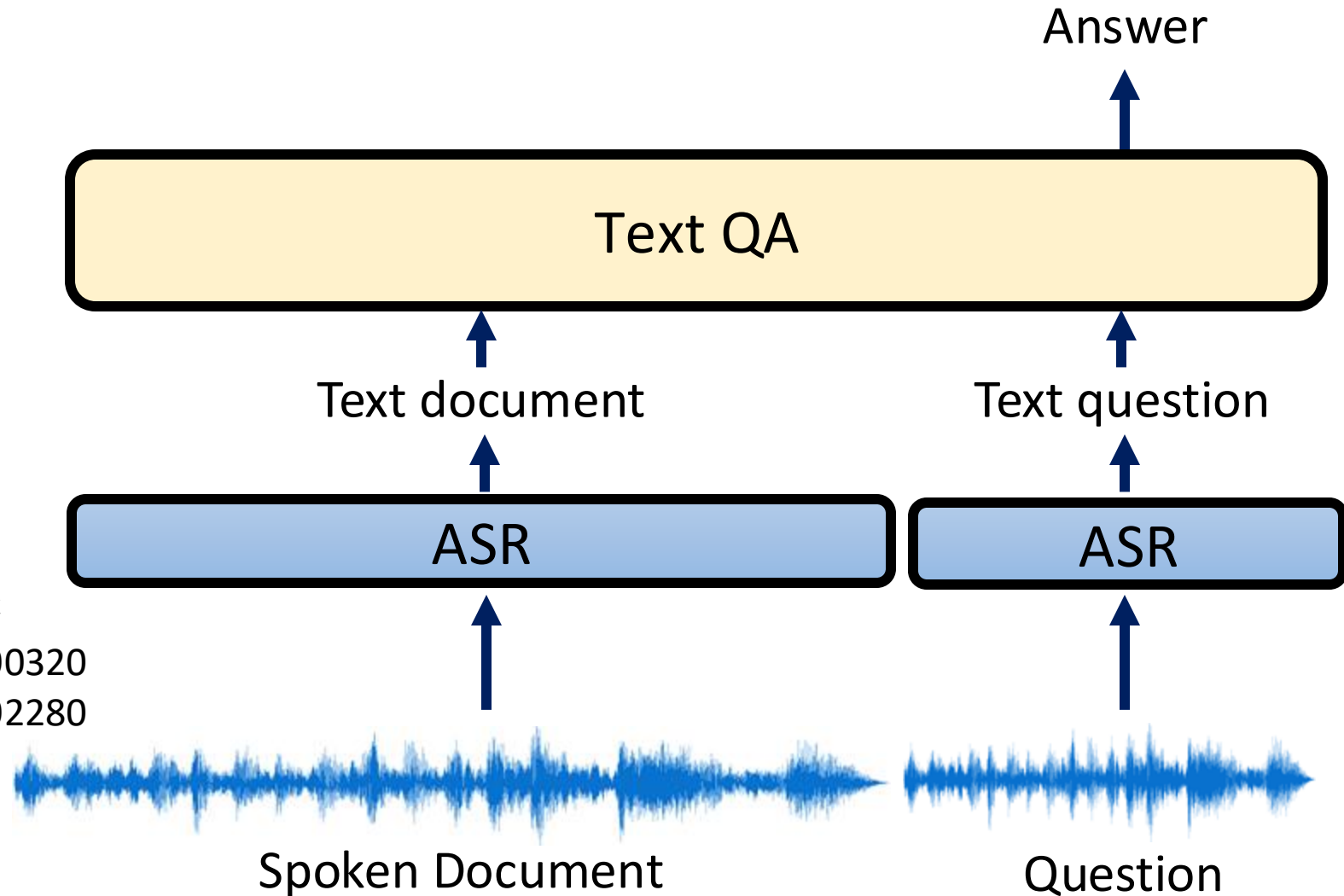
Choices:

- (A) gases released as a result of volcanic activity
- (B) chemical reactions caused by high surface temperatures
- (C) bursts of radio energy from the plane's surface
- (D) strong winds that blow dust into the atmosphere

This work is done by 曾柏翔

<https://arxiv.org/abs/1608.06378>

# Listening Comprehension (Speech QA)

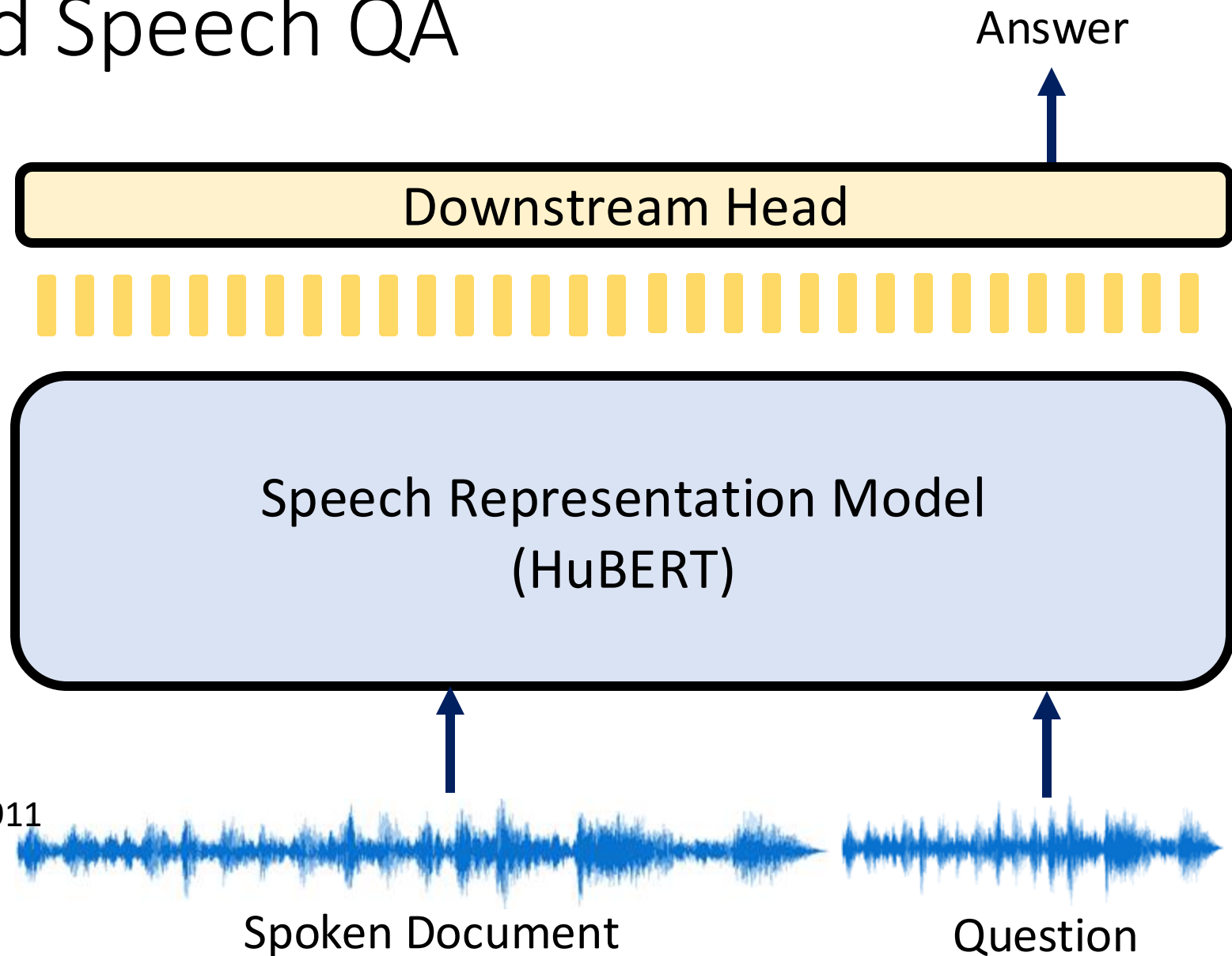


This work is done by 李佳軒

<https://arxiv.org/abs/1804.00320>

<https://arxiv.org/abs/1808.02280>

# End-to-End Speech QA



This work is done by 林冠廷  
<https://arxiv.org/abs/2203.04911>

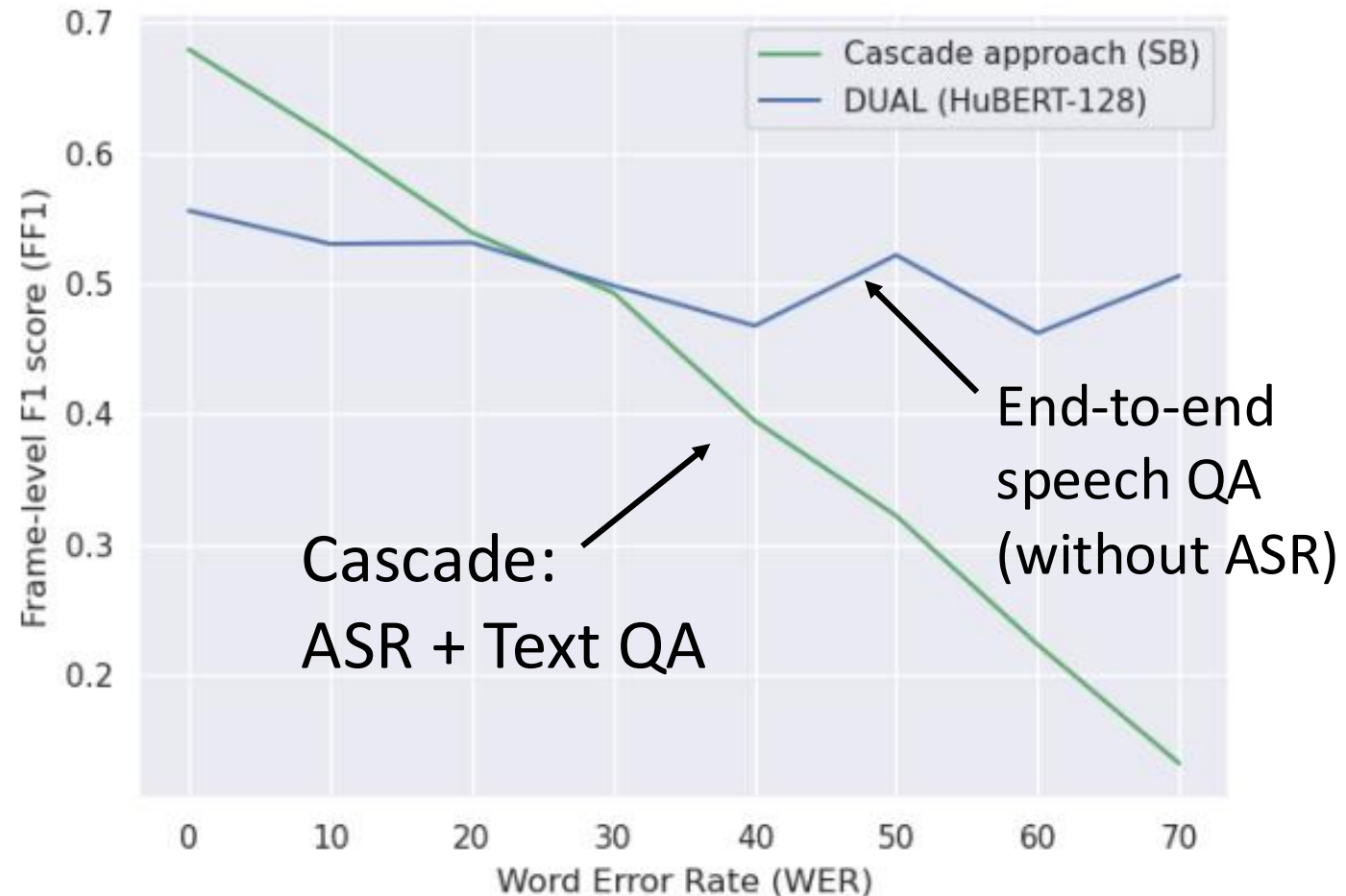


# End-to-End Speech QA

<https://arxiv.org/abs/2203.04911>



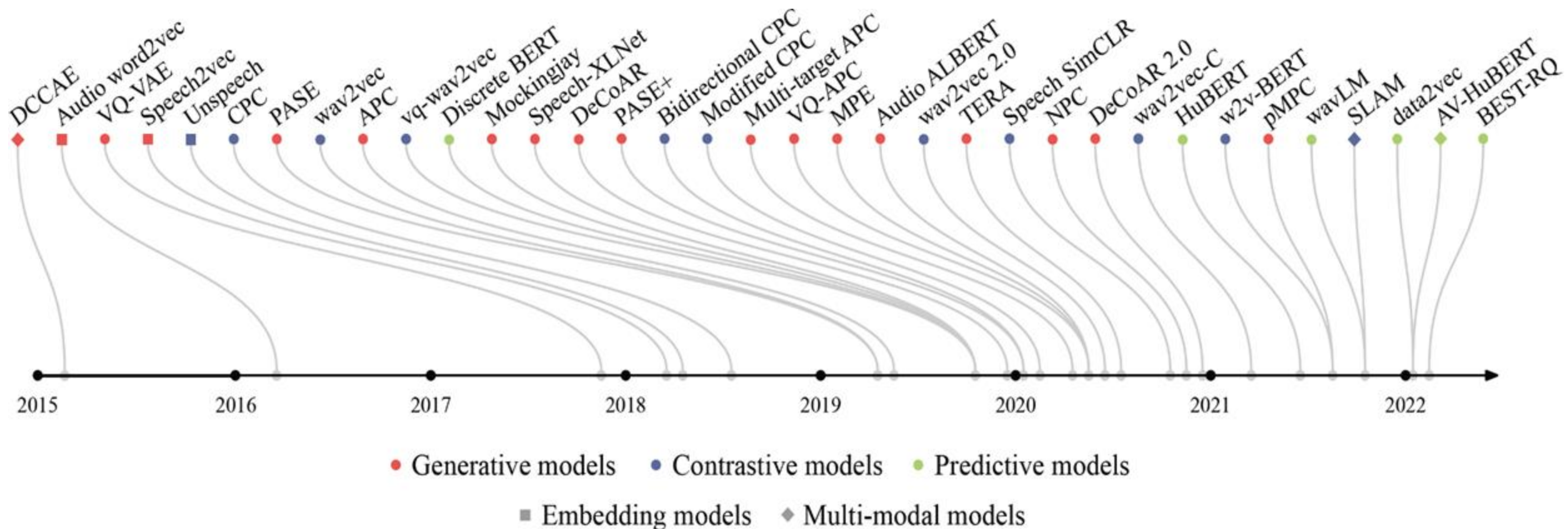
<https://rc.signalprocessingsociety.org/conferences/icassp-2022/spsicassp22vid1971>



# Self-Supervised Speech Representation Learning: A Review

<https://arxiv.org/abs/2205.10643>

Abdelrahman Mohamed\*, Hung-yi Lee\*, Lasse Borgholt\*, Jakob D. Havtorn\*, Joakim Edin, Christian Igel  
Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, Shinji Watanabe



# Useful Toolkit! The S3PRL toolkit



<https://github.com/s3prl/s3prl>

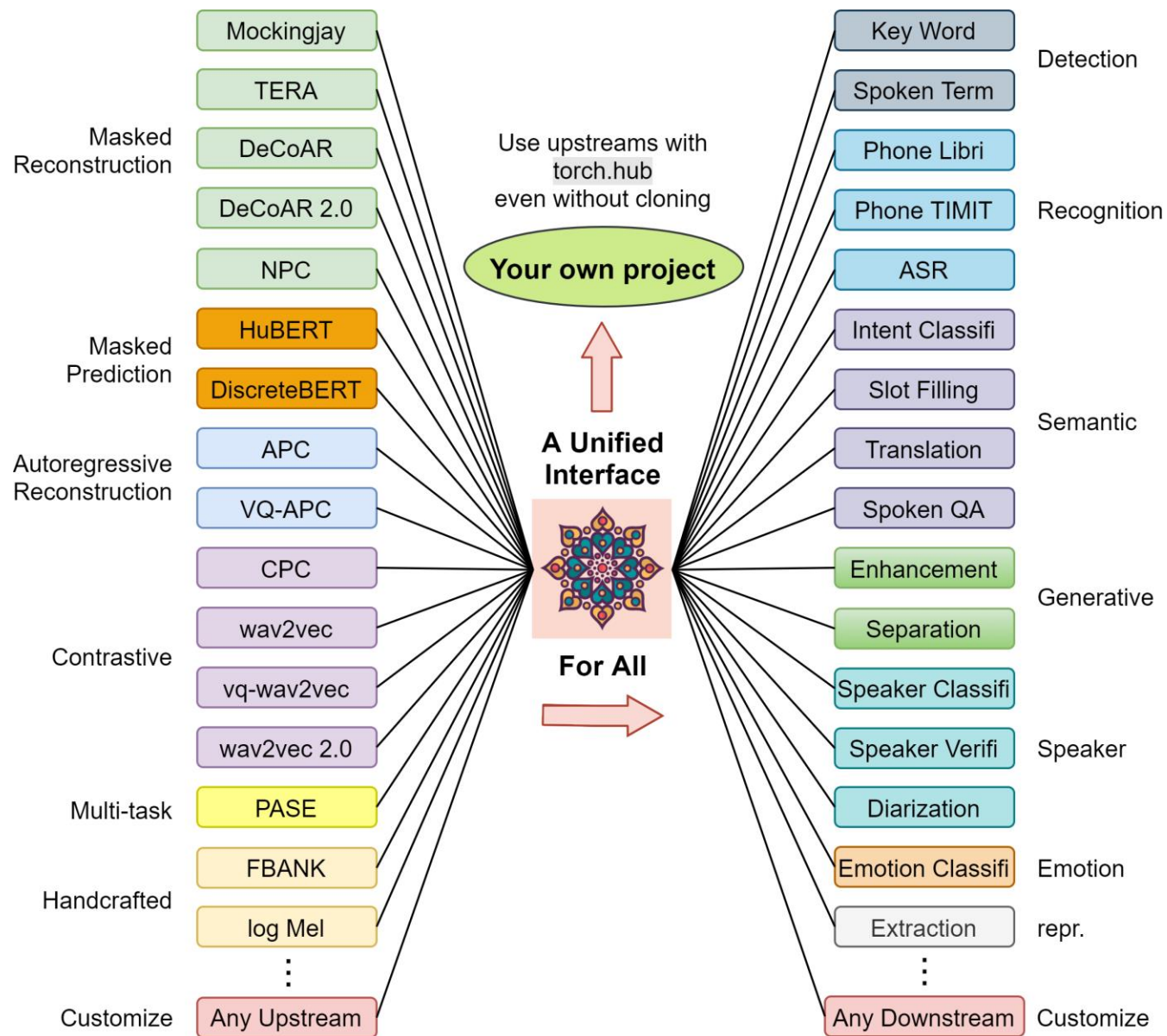
Creator



Shu-wen  
(Leo) Yang



Andy T. Liu



Over 2.5k stars & used by over 180 repos

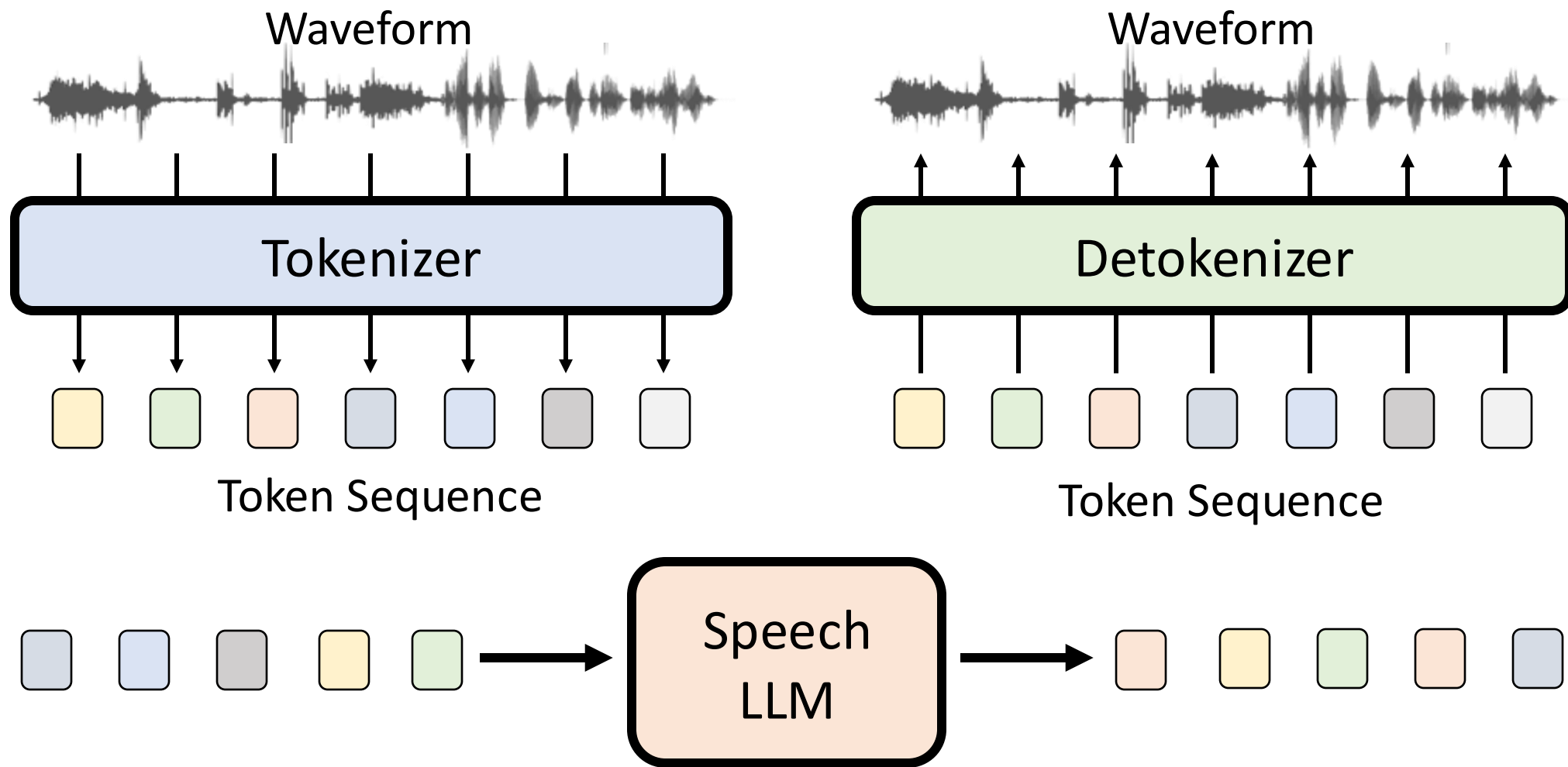


## 2. 初代語音語言模型

# Speech LLM

Generative Spoken Language Modeling  
(GSLM) from Raw Audio

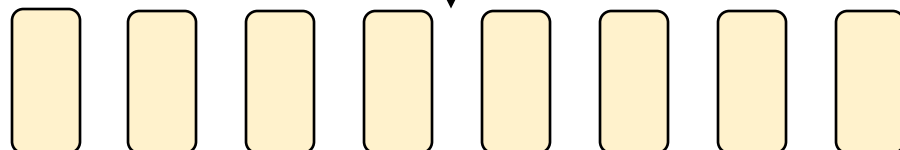
<https://arxiv.org/abs/2102.01192>



Tokenization



Representation Model 



K-means or VQ-layer




Deduplicate



BPE (Byte Pair Encoding)



Detokenization Model 

?????

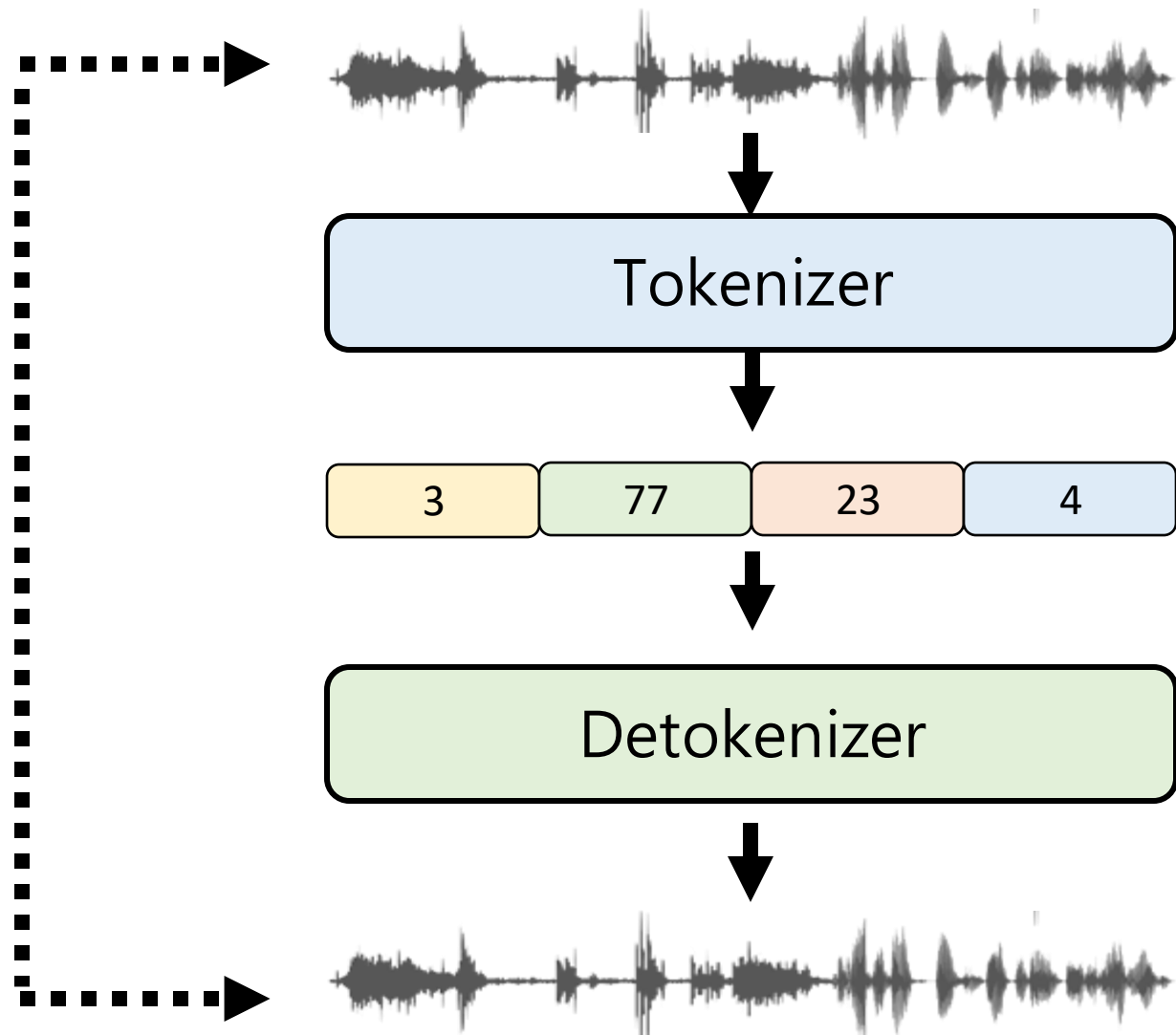
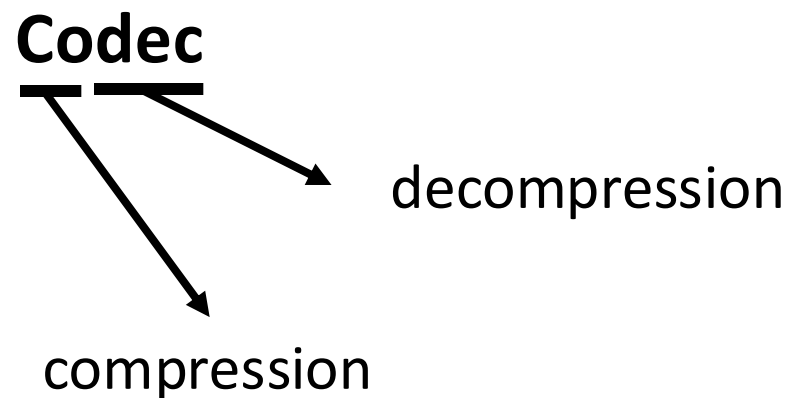




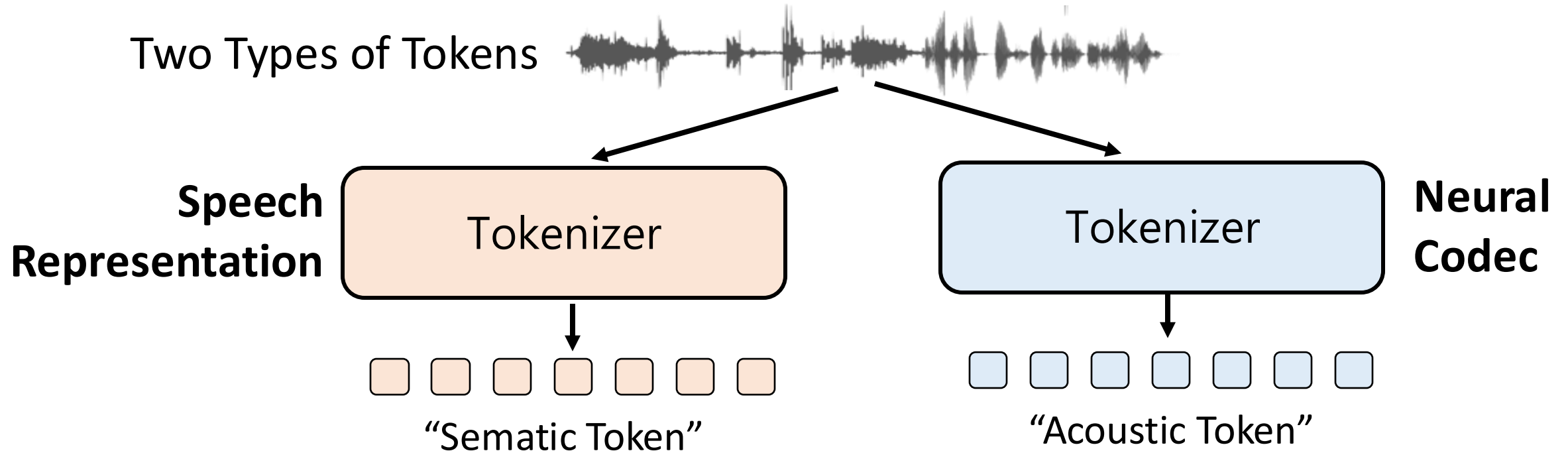
# Another possible pipeline of speech tokenization

## Neural Speech Codec

The tokenizer and detokenizer are learned jointly.



# Various Types of Speech Tokenizers

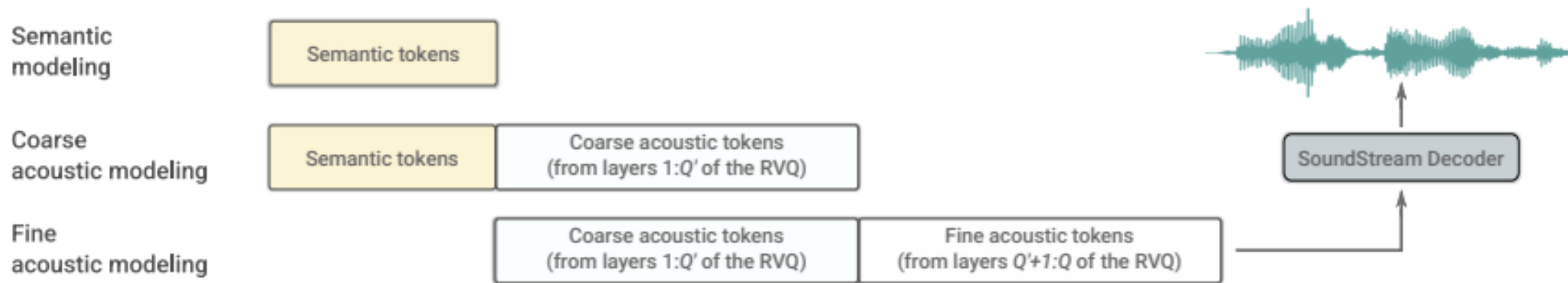


**Which one is the best choice?**



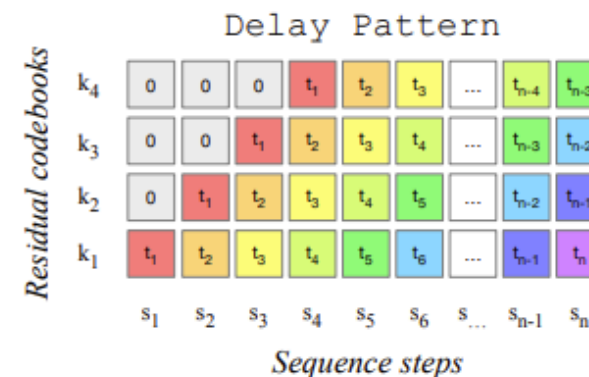
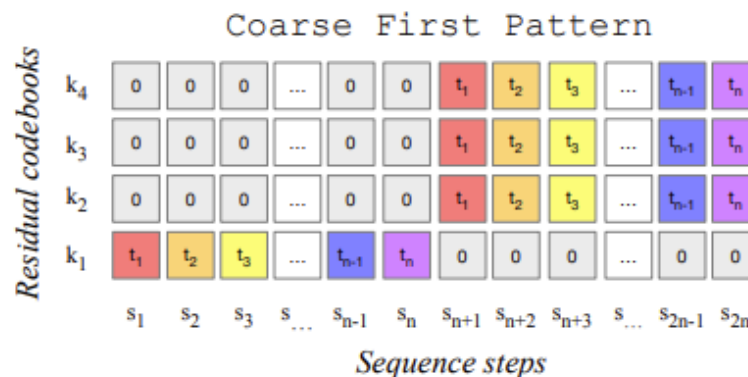
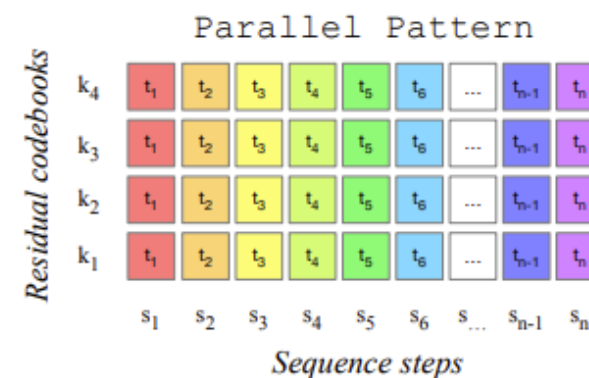
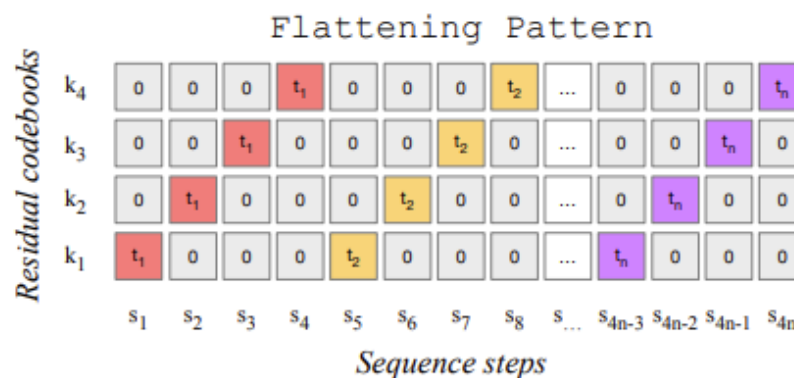
# Just use all of them!

Audio LM <https://arxiv.org/abs/2209.03143>



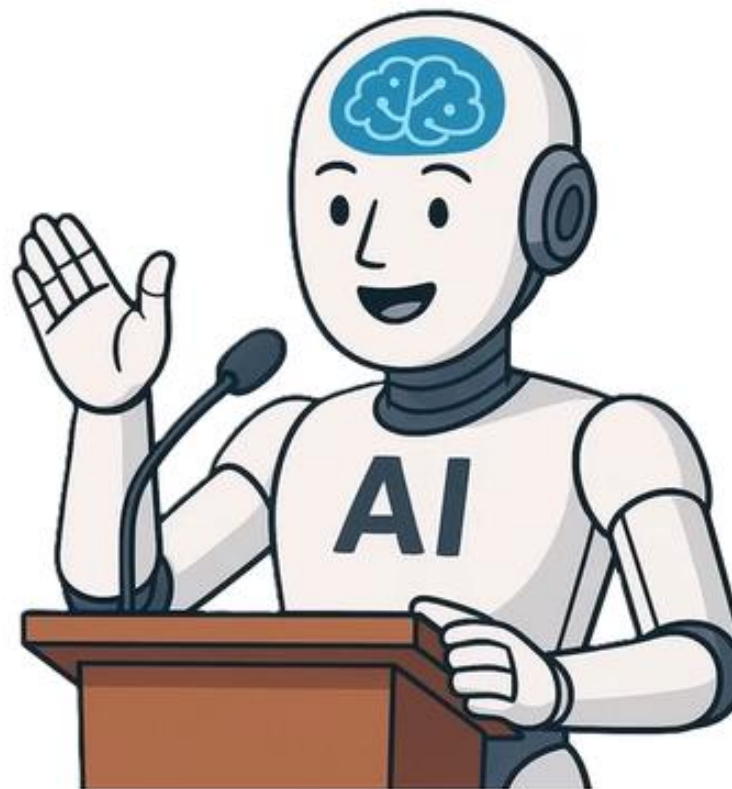
## How to arrange the order of their generation

<https://arxiv.org/abs/2306.05284>





## 可以聽和說的 語音語言模型

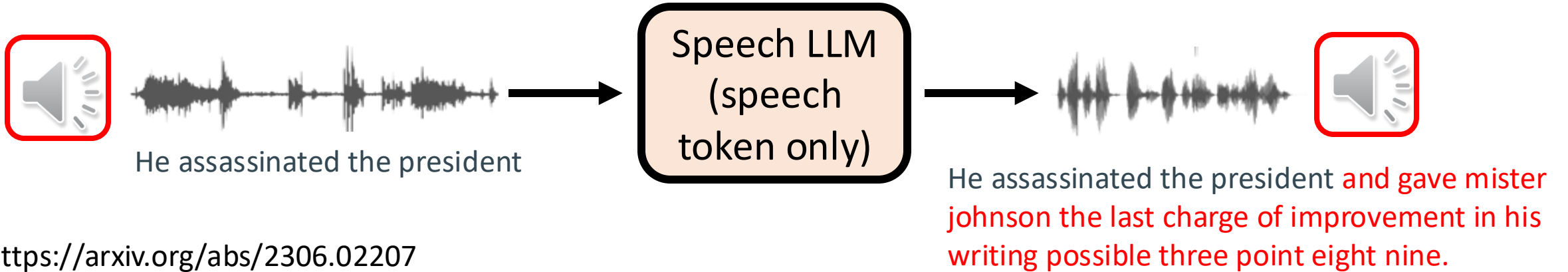


【生成式AI時代下的機器學習(2025)】第十二講：語言模型如何學會說話－概述語音語言模型發展歷程

<https://www.youtube.com/watch?v=gkAyqoQkOSk>

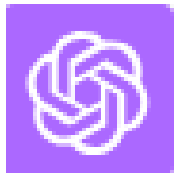
27:00 – 39:00

# Very challenging!



<https://arxiv.org/abs/2306.02207>

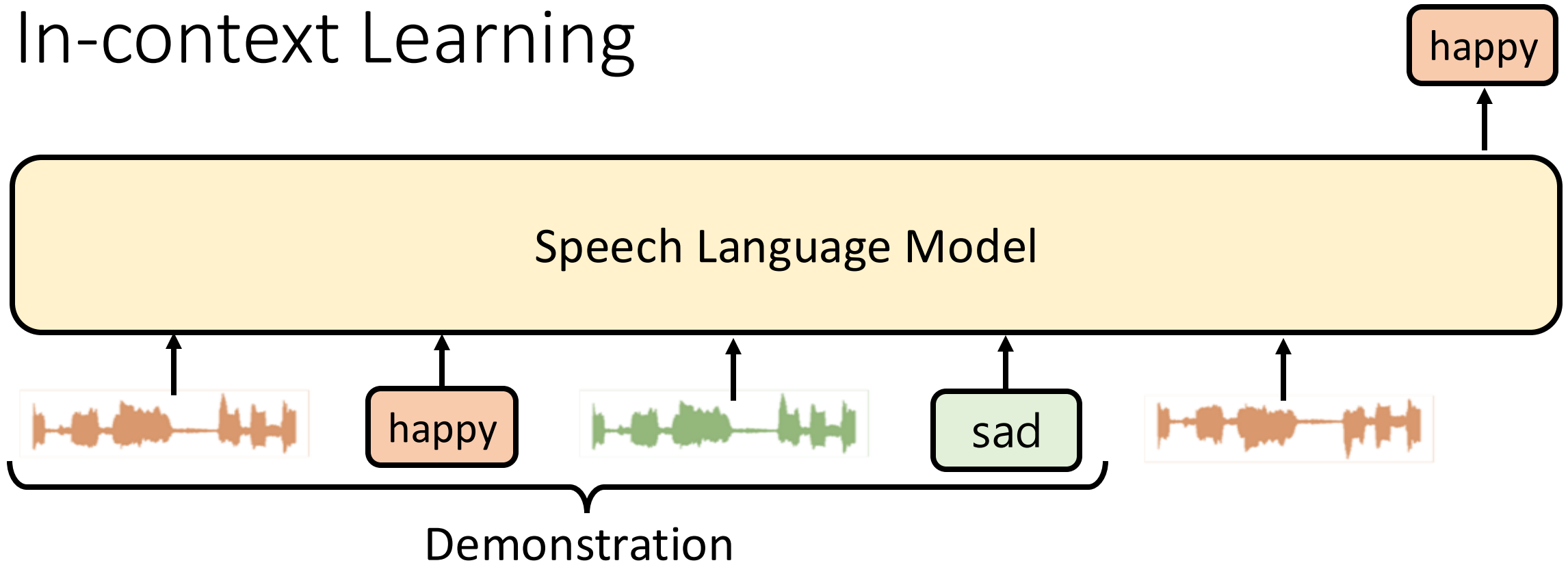
Does this sentence make sense?



Chat GPT

... while the sentence has recognizable English words and phrases, as it is currently constructed, **it doesn't coherently communicate a clear, singular idea or sequence of connected ideas.** ...

# In-context Learning



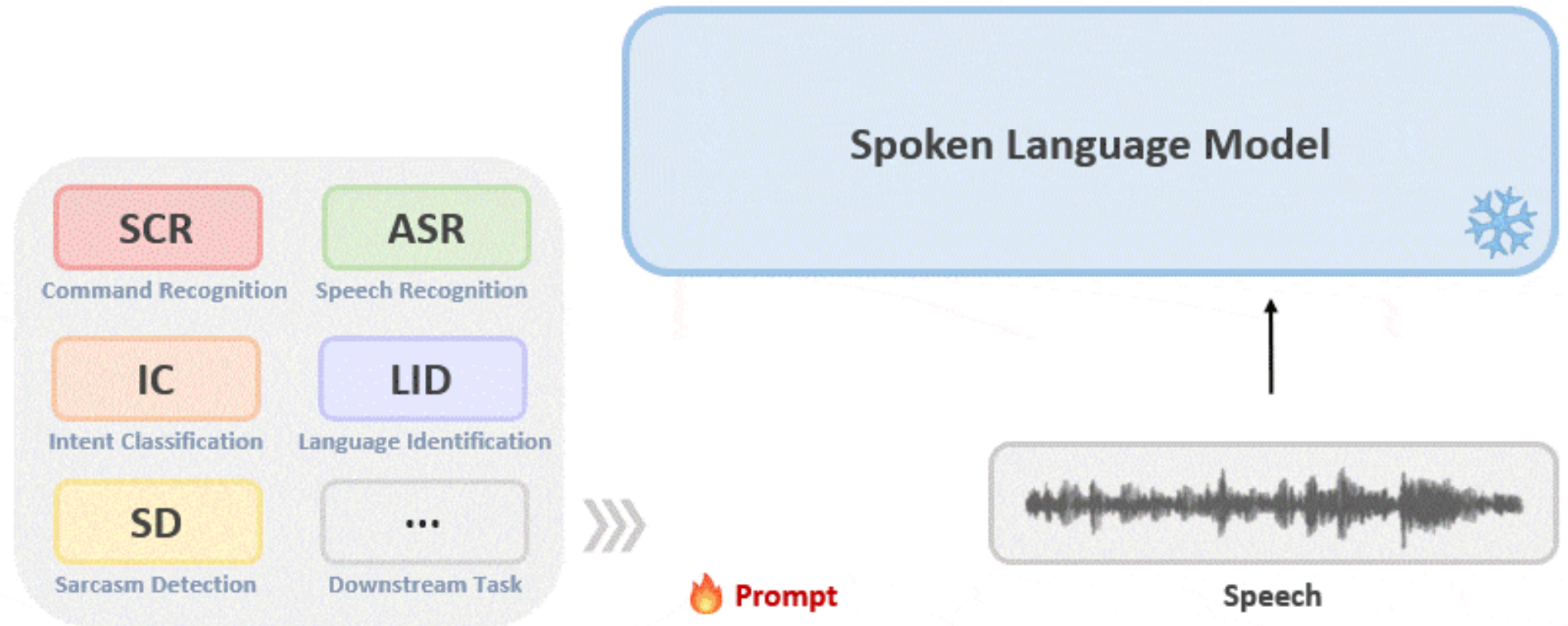
- Text LLM usually has the in-context learning capability, but speech LM (GSLM) has not shown similar capability.

# Prompting Speech LLM for Different Tasks

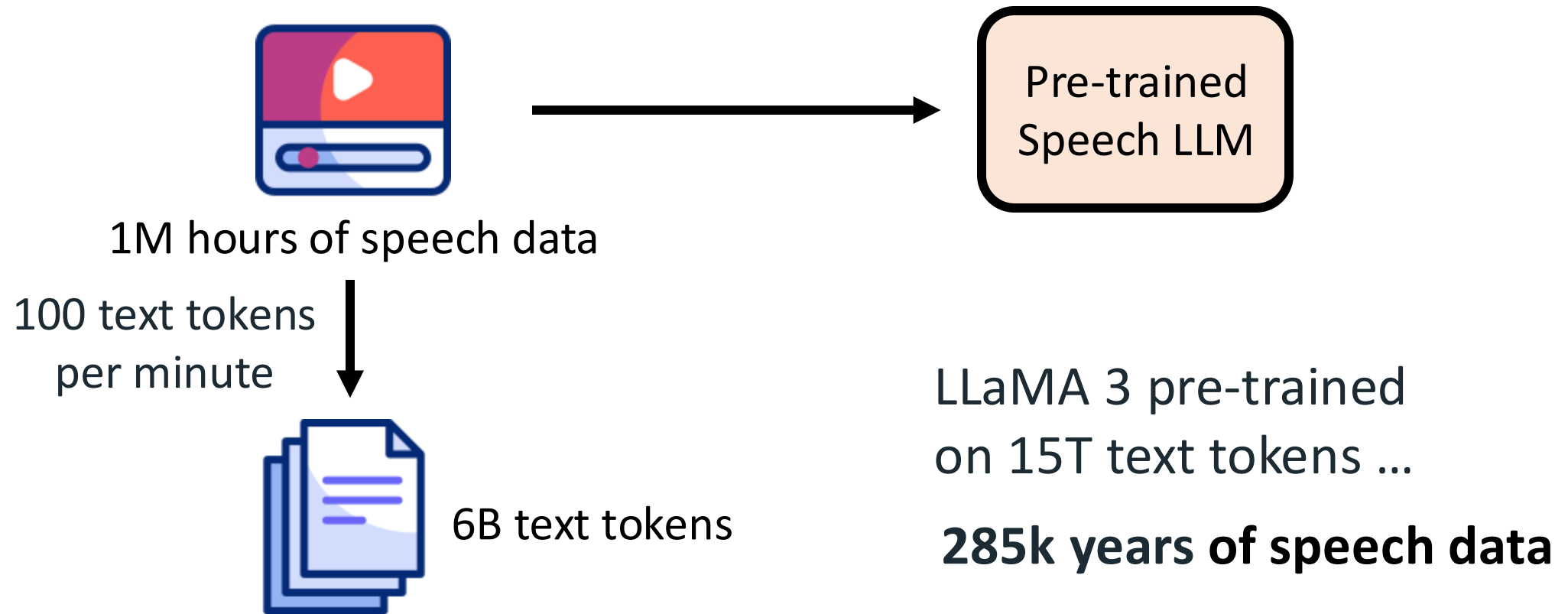
<https://ga642381.github.io/SpeechPrompt/>



Kai-Wei Chang

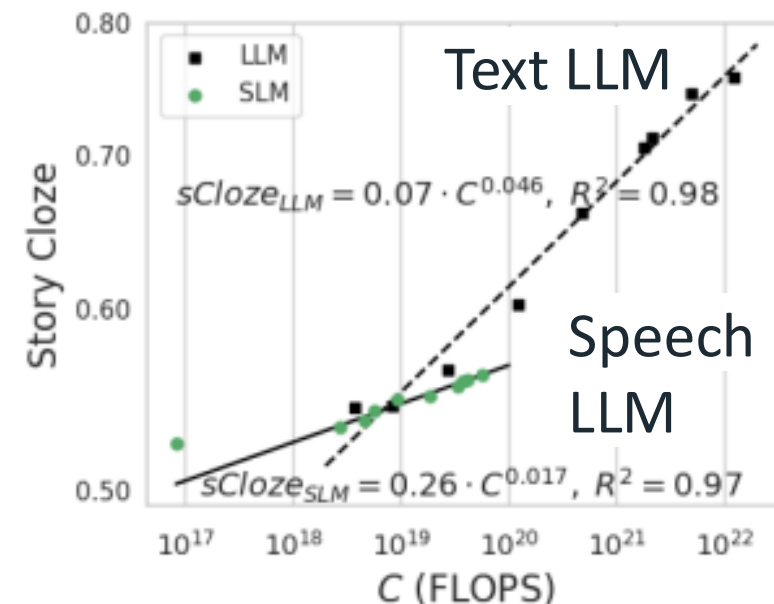
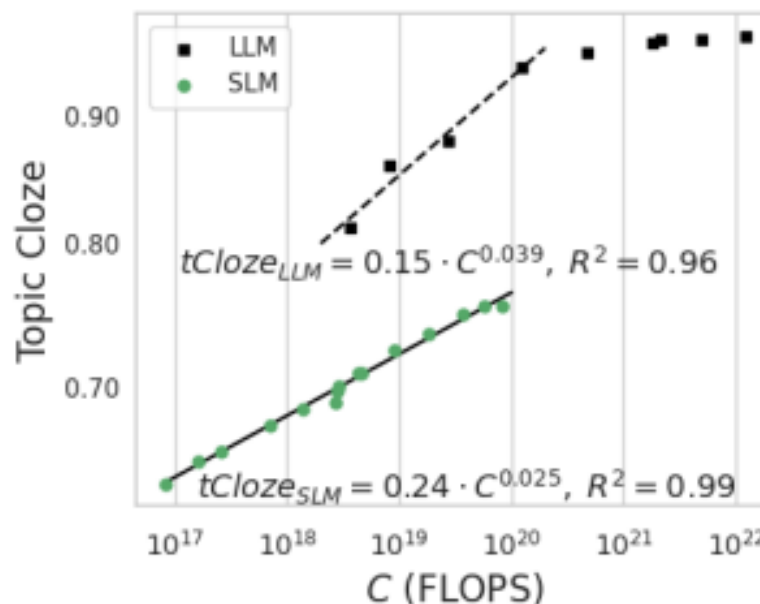
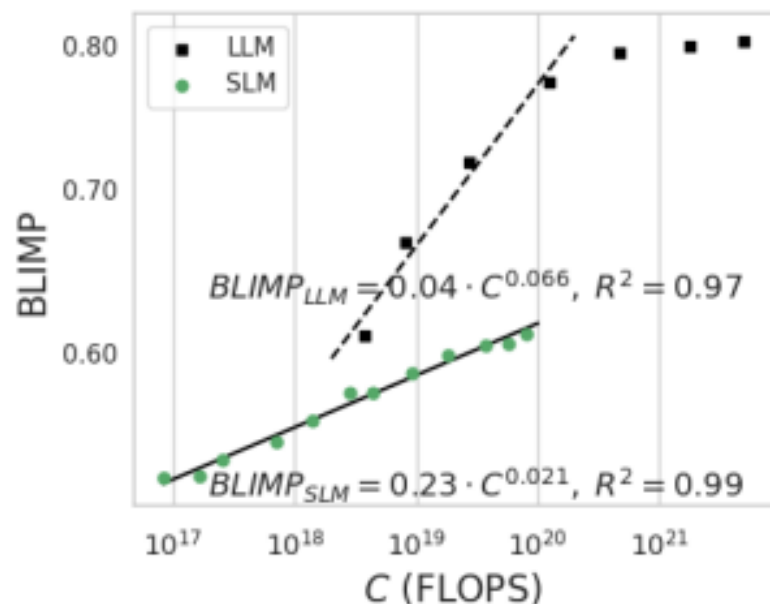


# Why is training speech LM so challenging?



# Why is training speech LM so challenging?

<https://arxiv.org/abs/2404.00685>



Besides content, speech LLMs also have to learn to understand other information (such as speaker identity, emotion, etc.) that text LLMs do not have to.



Until .....

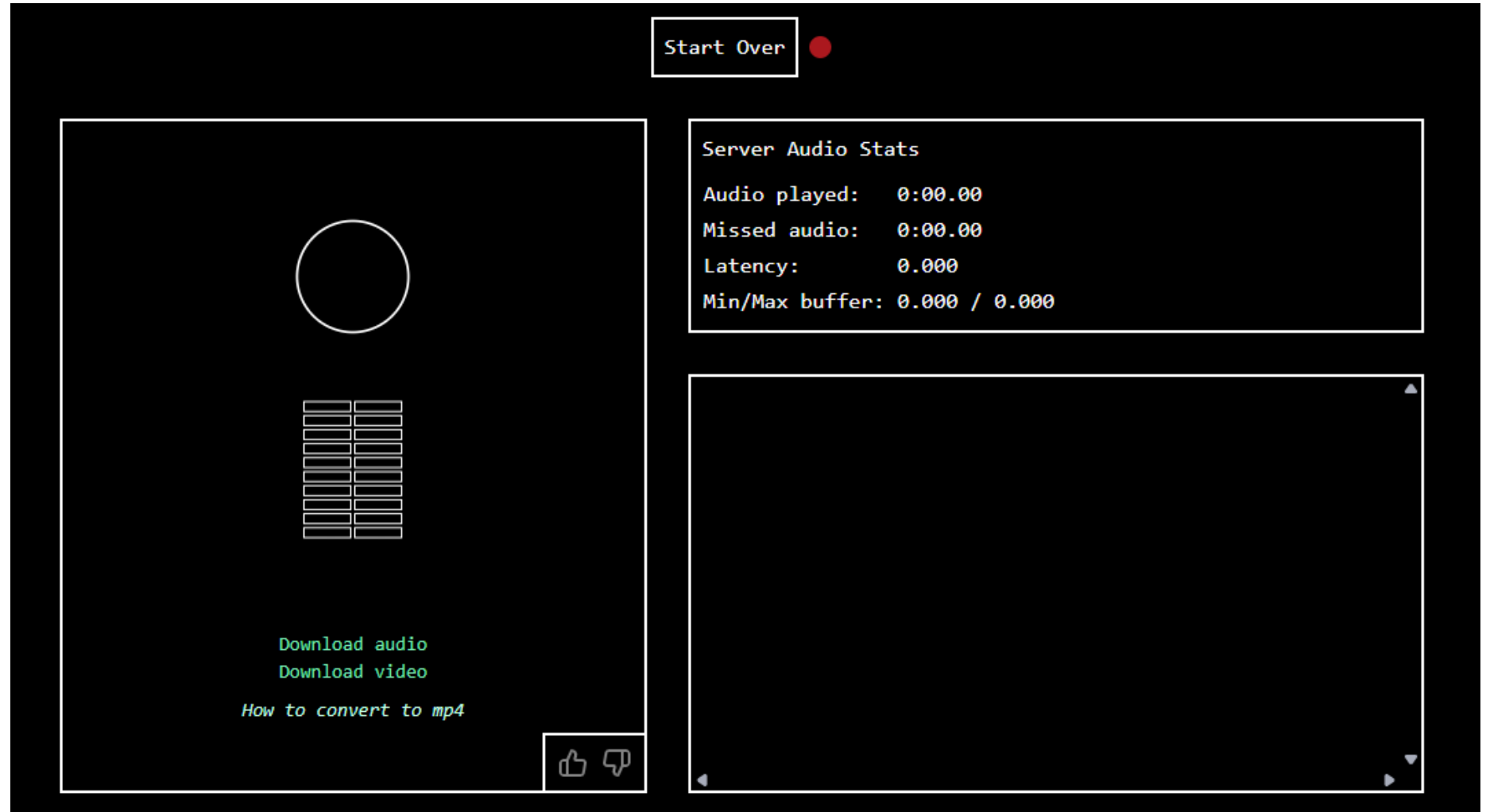


GPT-4o Voice Mode

Source of image: <https://www.youtube.com/watch?v=DQacCB9tDaw>



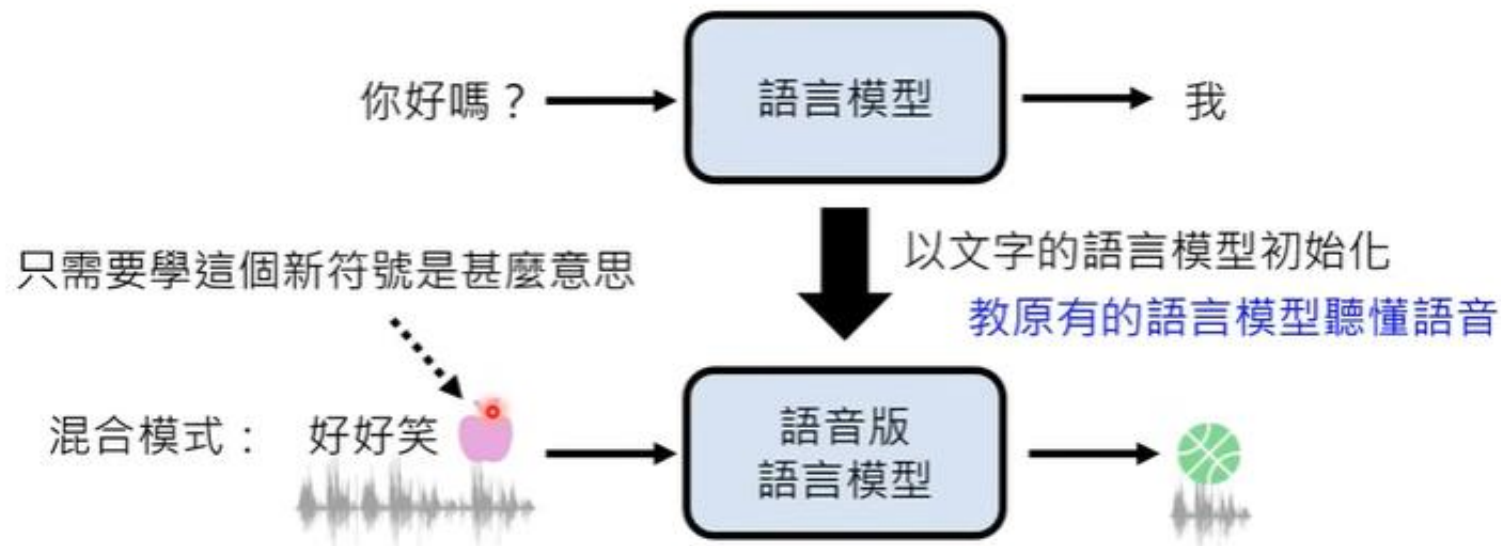
# Moshi



<https://moshi.chat/>

# 3. 如何利用強大的文字模型

## 模型訓練：Pre-train (利用文字資訊)



其他利用文字資訊的方式

<https://arxiv.org/abs/2310.08715>

<https://arxiv.org/abs/2402.05755>

27:47 / 38:12

GPT-4o 背後可能的語音技術猜測

<https://youtu.be/CgQ3lUOpXgc?si=LhzquKSJTdOTjZCw>

# Starting from Text LLM

Initialized by text  
LLM (BERT)

Downstream Head

Answer

Speech Representation Model  
(HuBERT)

DUAL

<https://arxiv.org/abs/2203.04911>

GSQA

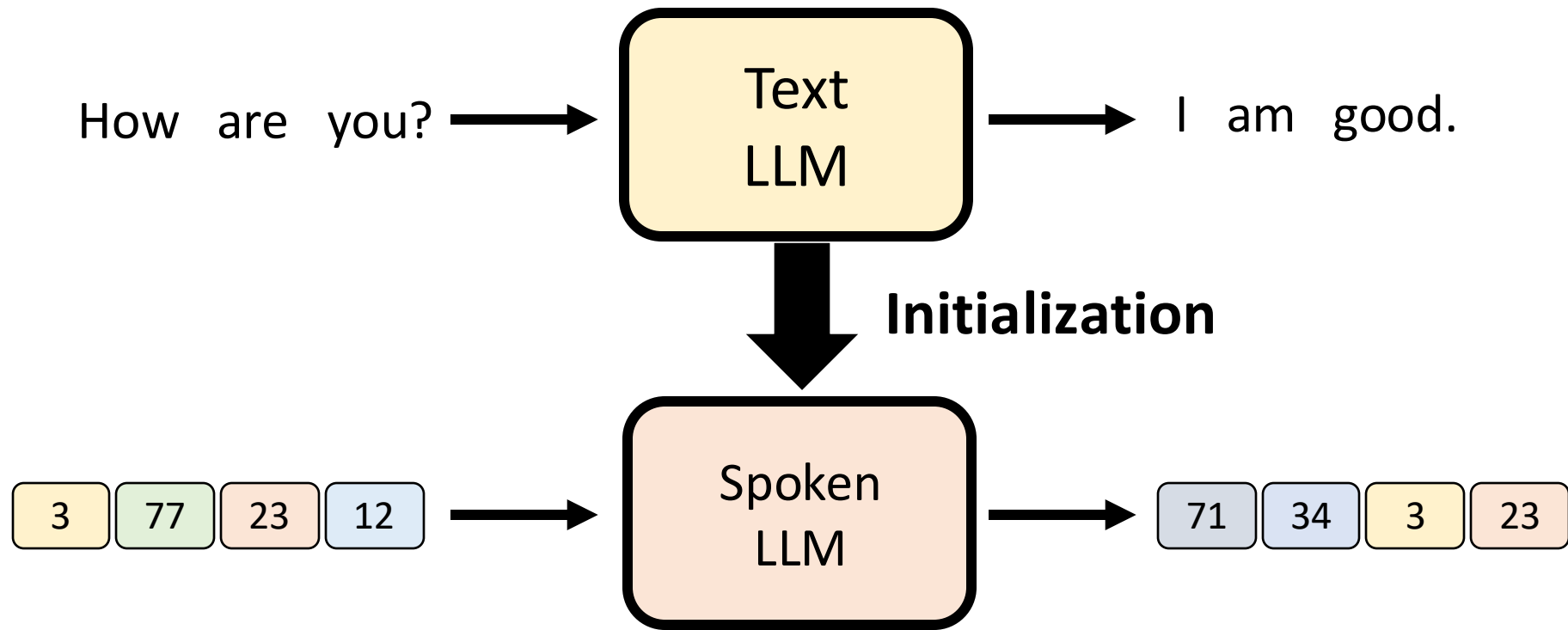
<https://arxiv.org/abs/2312.09781>



Spoken Document

Question

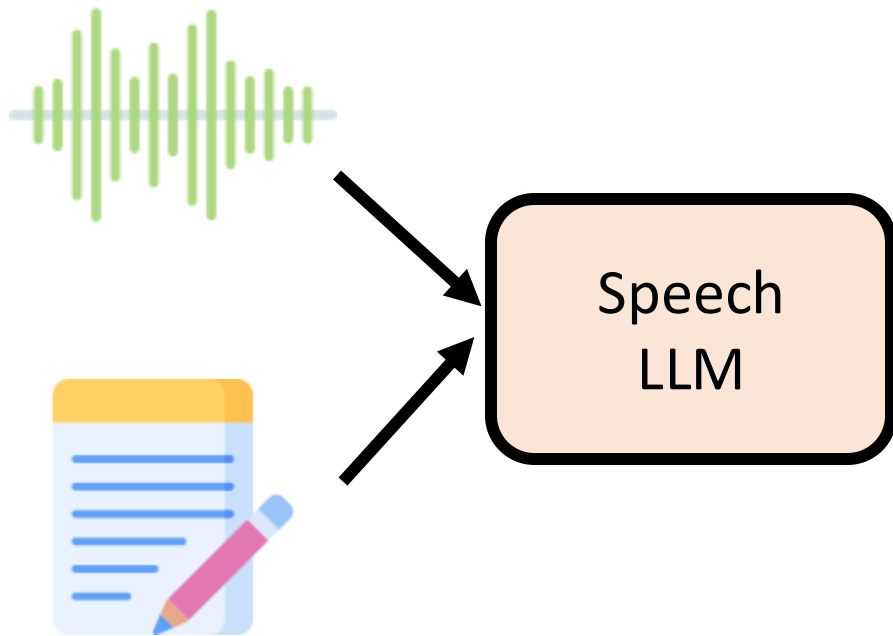
# Starting from Text LLM



TWIST

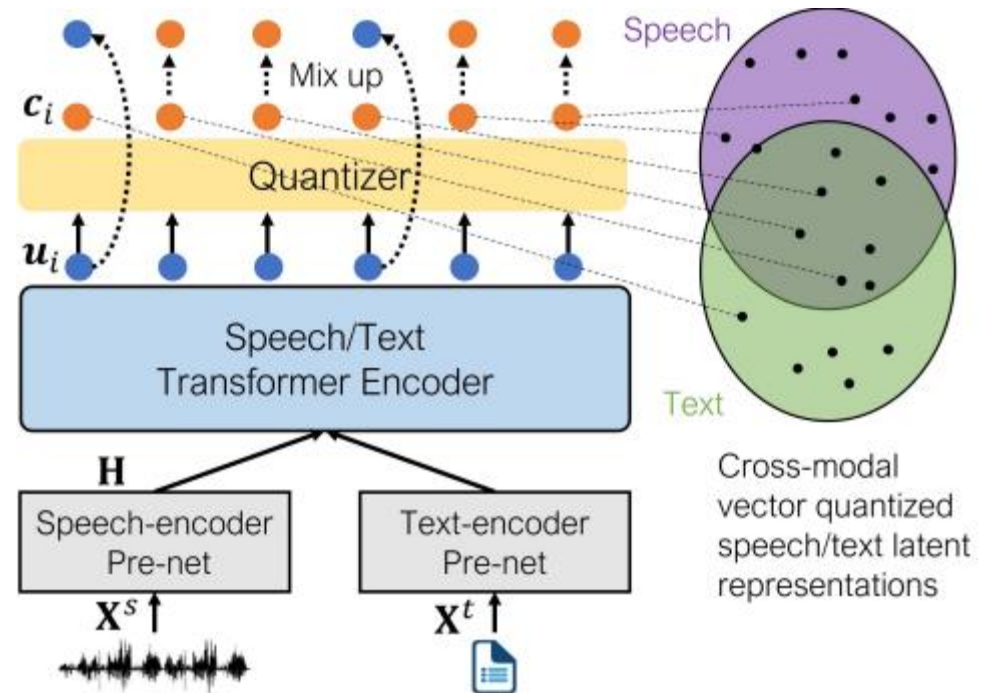
<https://arxiv.org/abs/2305.13009>

# Jointly learned from the speech and text



How to better alignment  
between speech and text

Widely studied in the era of shared encoder

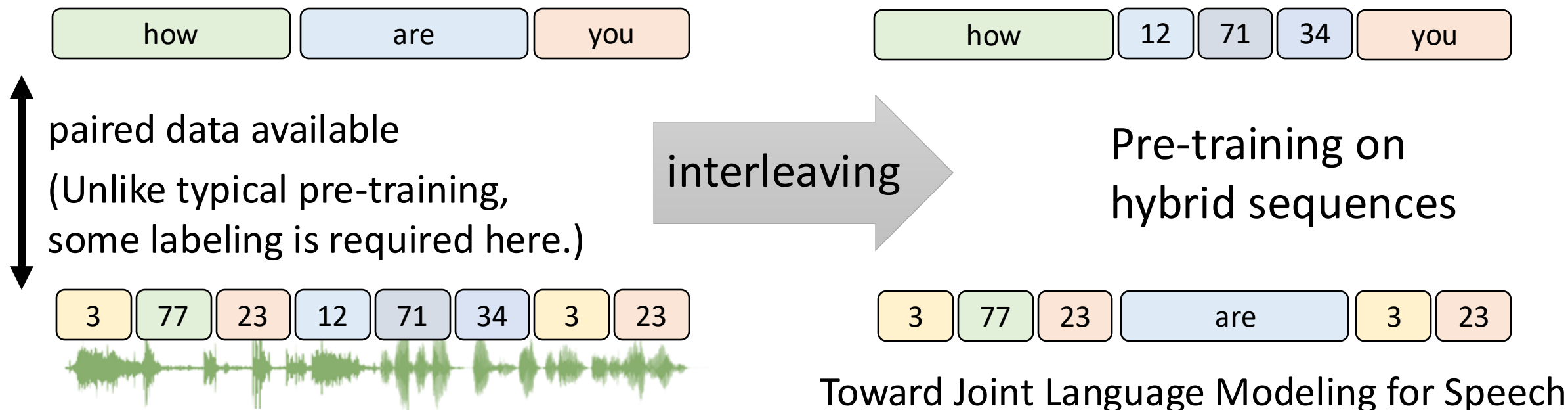


Source:  
speechT5

<https://arxiv.org/abs/2110.10329>  
<https://arxiv.org/abs/2110.07205>  
<https://arxiv.org/abs/2204.03409>

# Jointly learned from the speech and text

- Methods for spoken LM



Toward Joint Language Modeling for Speech Units and Text

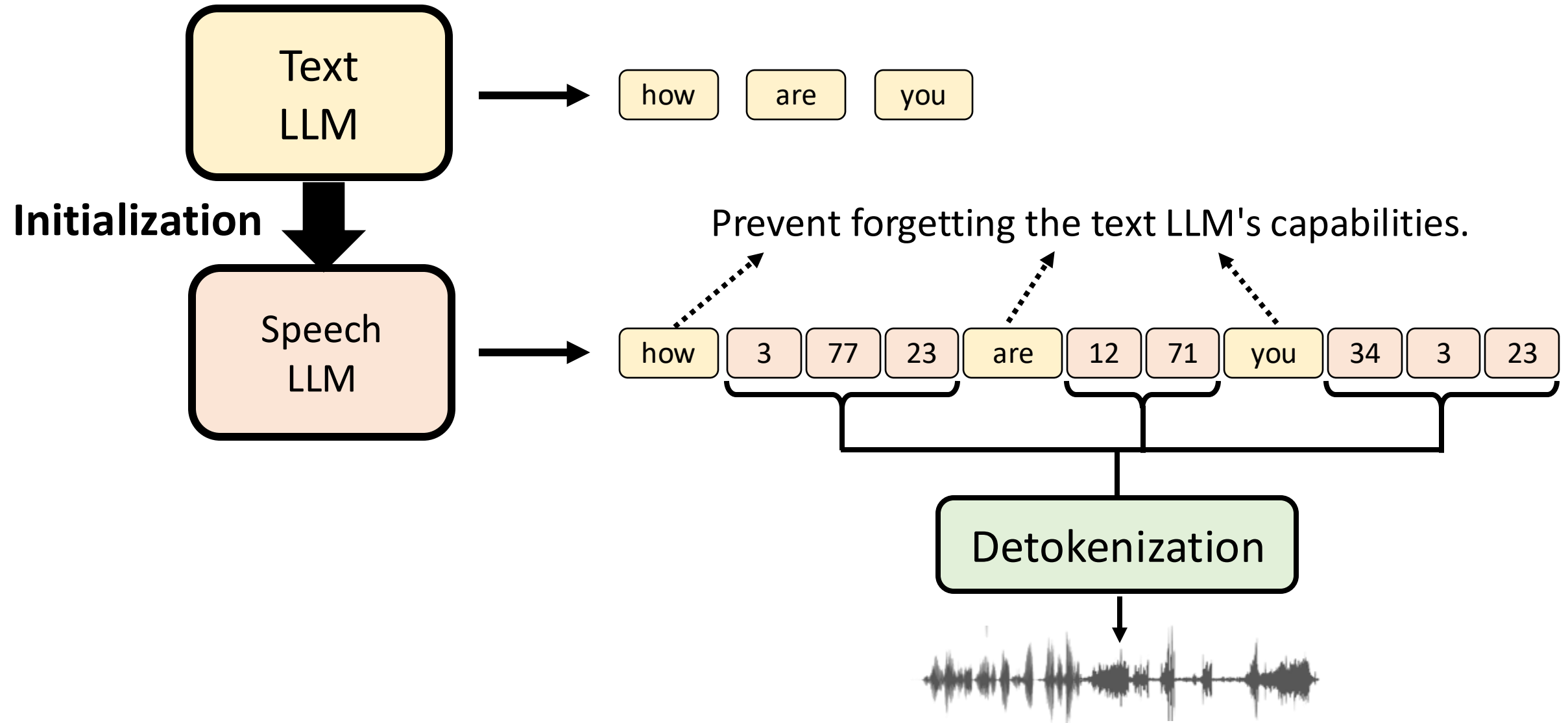
<https://arxiv.org/abs/2310.08715>

Spirit LM

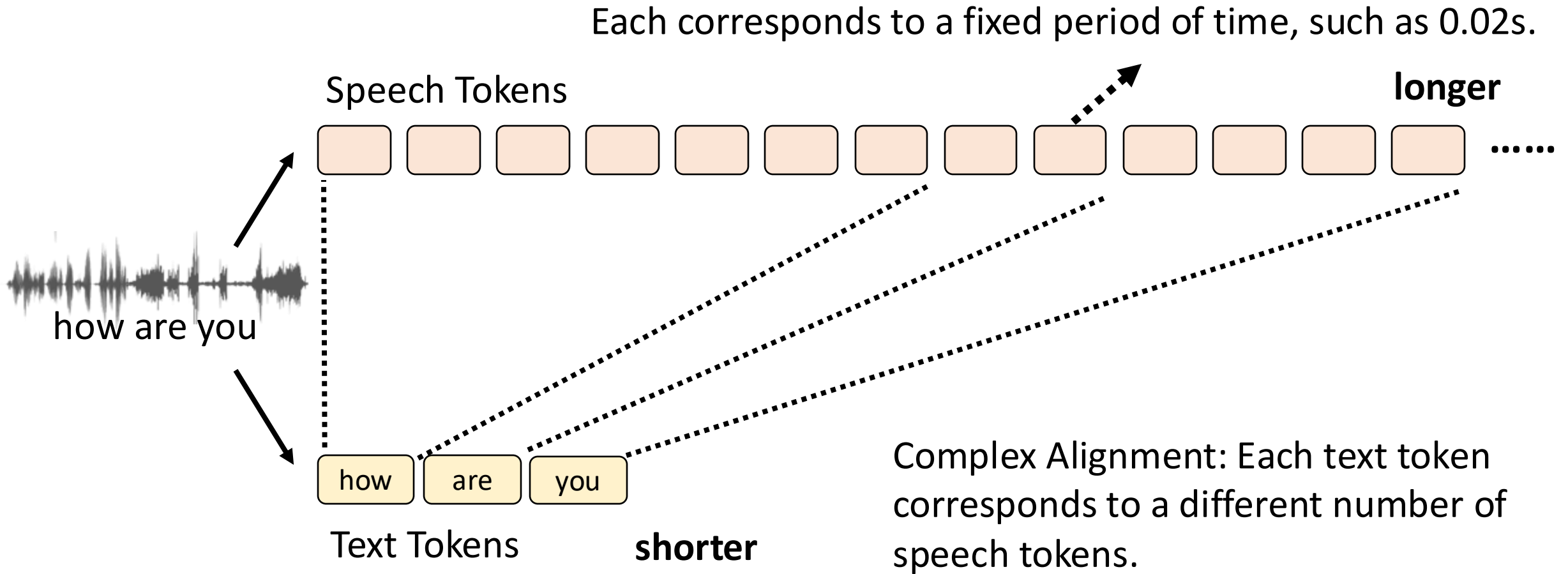
<https://arxiv.org/abs/2402.05755>



# Text-Speech Hybrid Generation

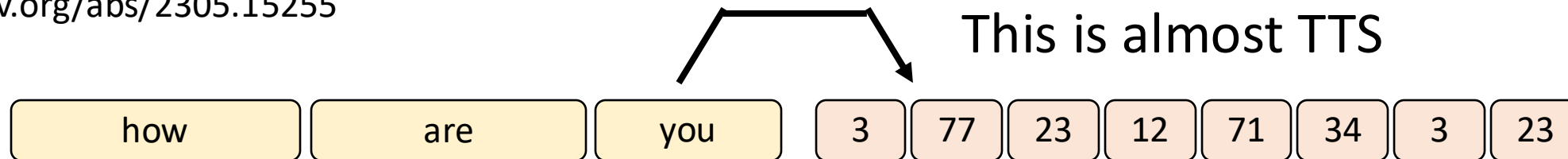


# Text-Speech Hybrid Generation



## Text then speech (sentence-level):

<https://arxiv.org/abs/2305.15255>

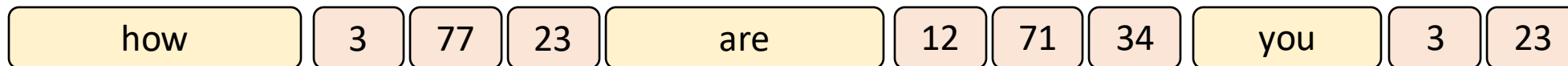


Drawback: cannot streaming

## Text then speech (word-level):

<https://arxiv.org/abs/2411.07111>

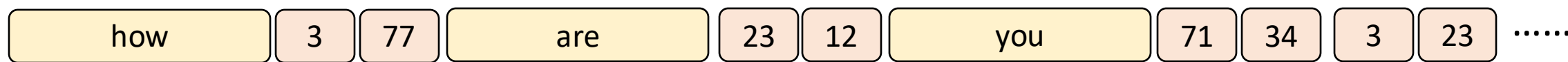
<https://arxiv.org/abs/2505.17496>



We need alignment between text and speech during training.

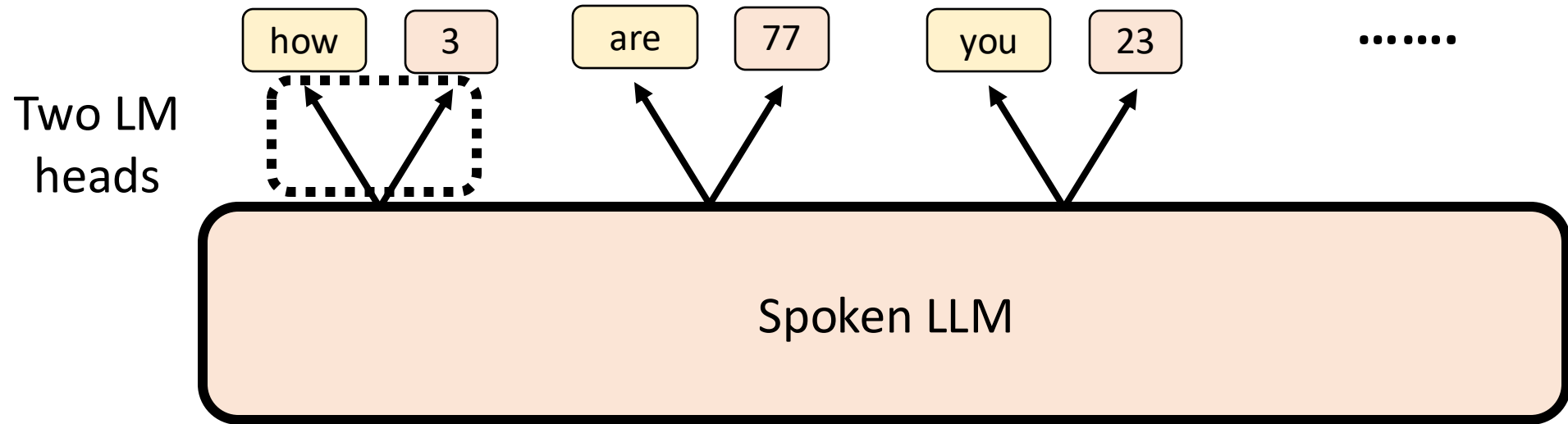
## Text then speech (chunk-level):

<https://arxiv.org/abs/2412.02612>

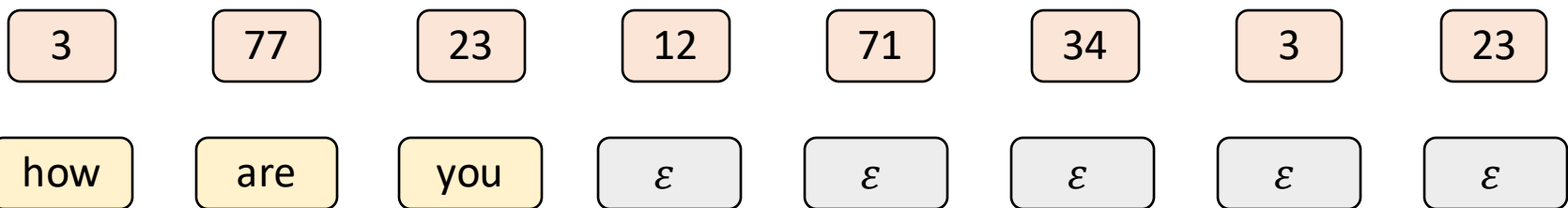


# Text-Speech Hybrid Generation

Text and speech at the same time



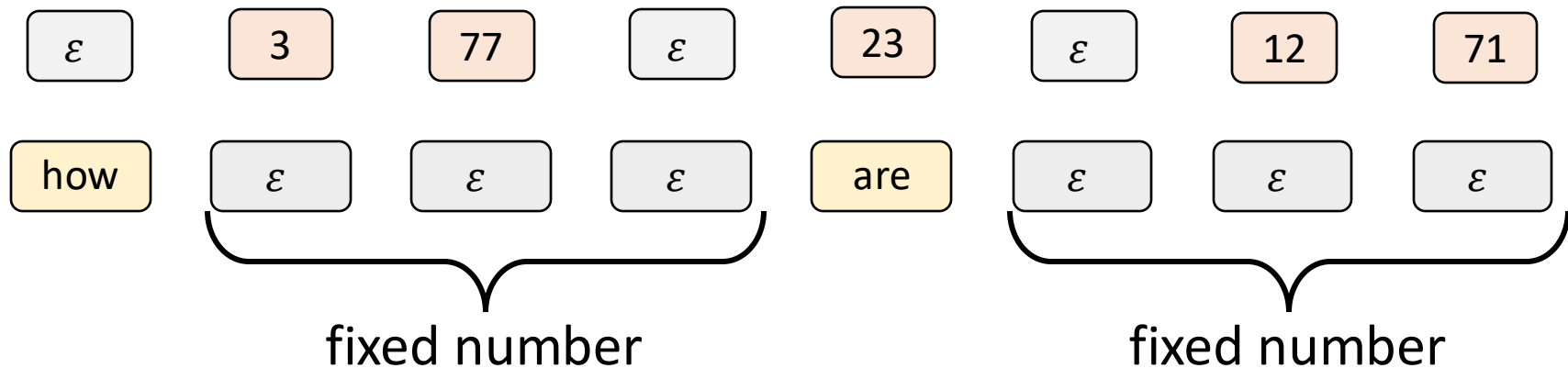
The text token and speech token do not have the same scale (their lengths differ significantly).



Mini-Omni

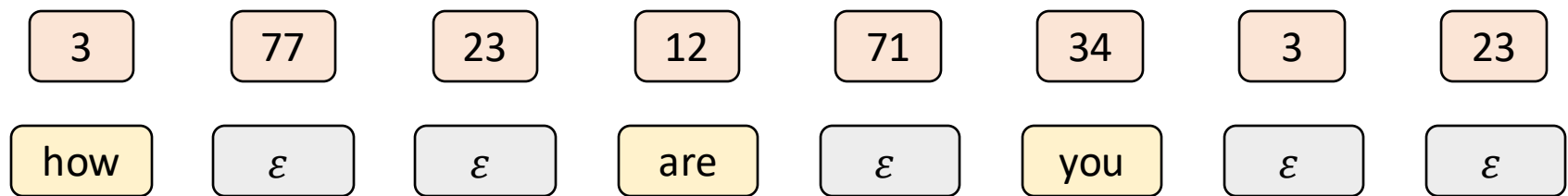
<https://arxiv.org/abs/2408.16725>

CTC loss



LLaMA-Omni

<https://arxiv.org/abs/2409.06666>

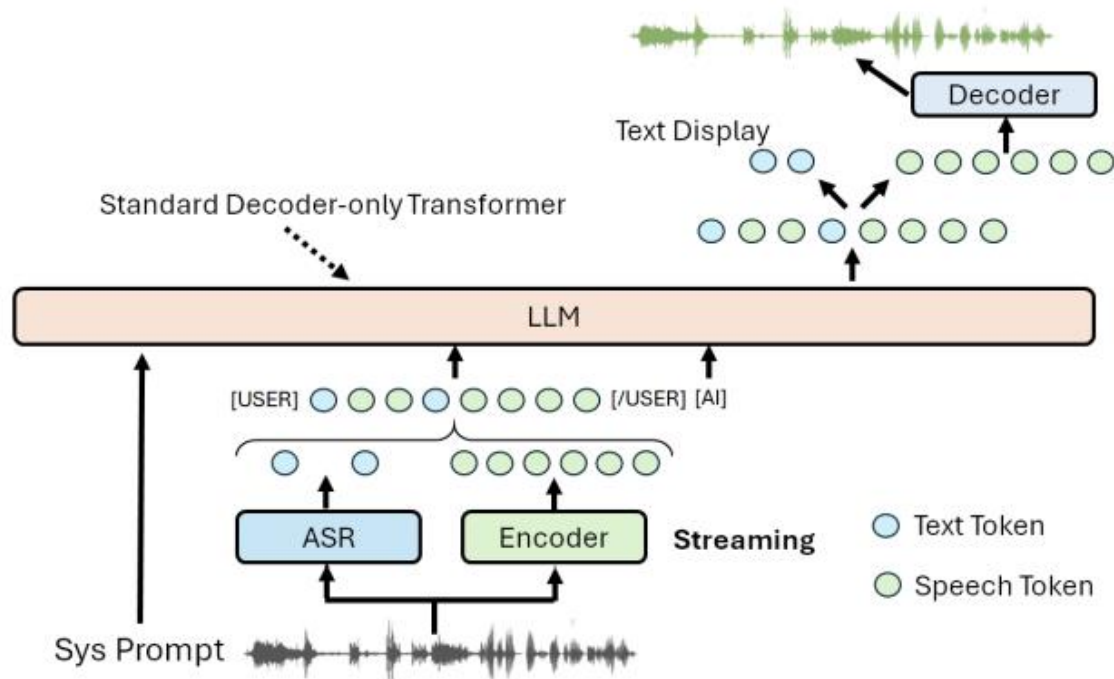


Moshi

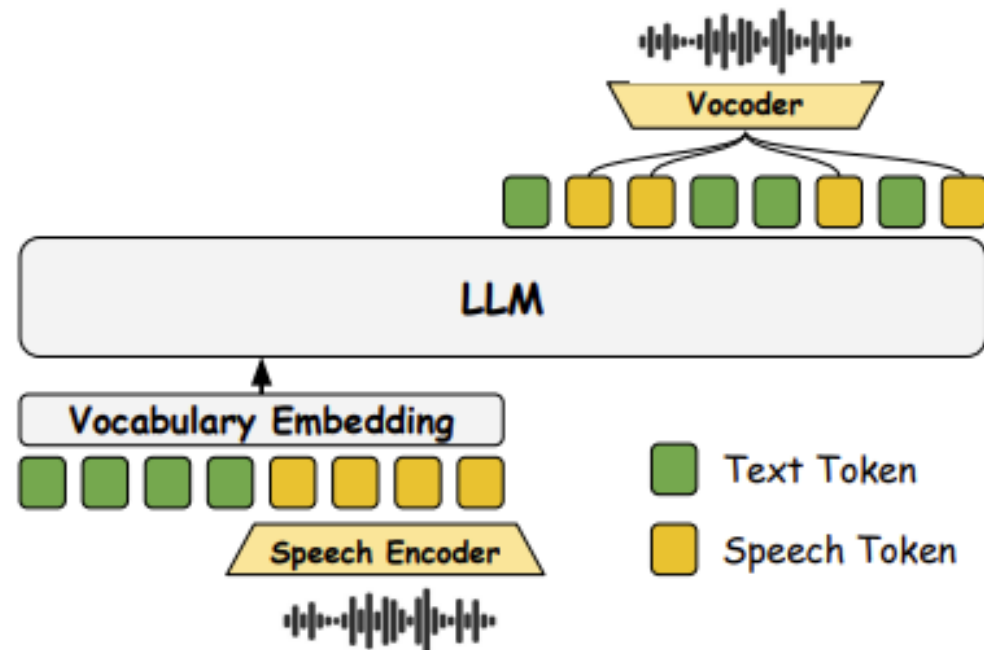
<https://arxiv.org/abs/2410.00037>

This is similar to a duration model.



# Text then speech (word-level)

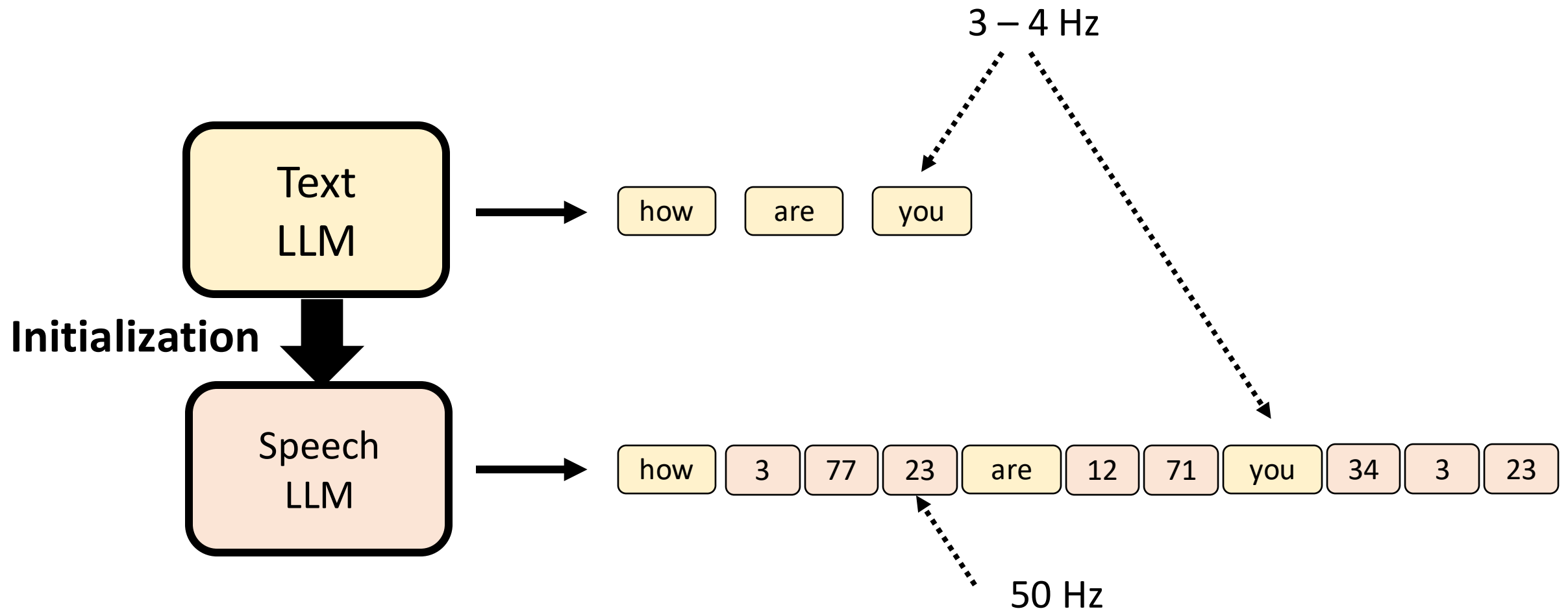


<https://arxiv.org/abs/2411.07111>



<https://arxiv.org/abs/2505.17496>

	<p>I am hungry now. Can you give me some ideas?</p> <p>The training data is approximately 1,000 hours.</p>		<p>I'd be happy to help you with some ideas for something to eat. What kind of food are you in the mood for? Are you craving something sweet, savory, or something else? Do you have any dietary restrictions or preferences? Let me know and I can suggest some options for you.</p>
---	--	---	---



**Ten** times the length and **one hundred** times the computational cost for attention...



## 4. 尋找更合適的語音表示方式

---

# TASTE: Text-Aligned Speech Tokenization and Embedding for Spoken Language Modeling

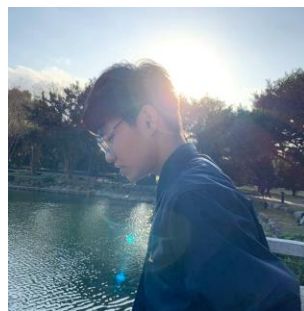
---

Liang-Hsuan Tseng<sup>\*23</sup> Yi-Chang Chen<sup>\*1</sup> Kuan-Yi Lee<sup>23</sup> Da-Shan Shiu<sup>1</sup> Hung-yi Lee<sup>3</sup>

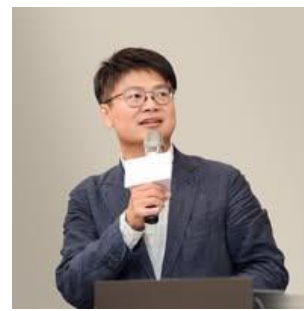
<sup>\*</sup>Equal contribution    <sup>1</sup>MediaTek Research

<sup>2</sup>Internship at MediaTek Research    <sup>3</sup>National Taiwan University

<https://arxiv.org/abs/2504.07053>



Liang-Hsuan Tseng  
(NTU)



Yi-Chang Chen  
(MediaTek)

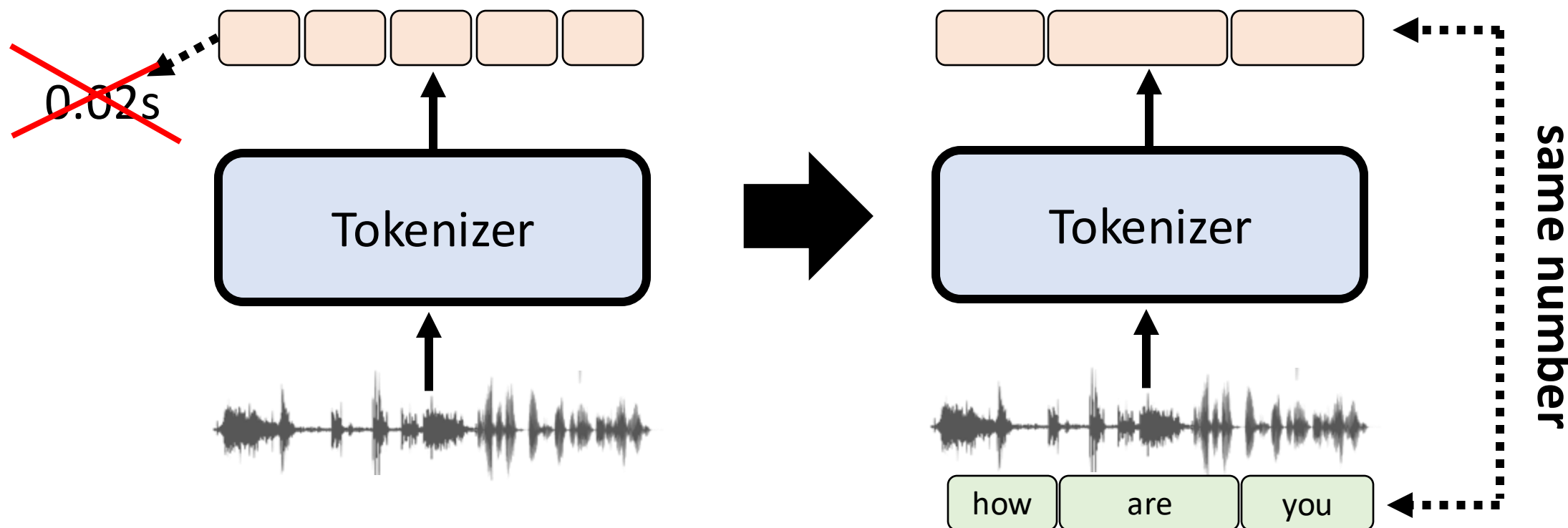


Kuan-Yi Lee  
(NTU)

# Can we have speech tokens designed for text-speech hybrid generation?

## 1. Straightforward relationship with text (cannot be fixed duration)

e.g., one speech token corresponds to one text token

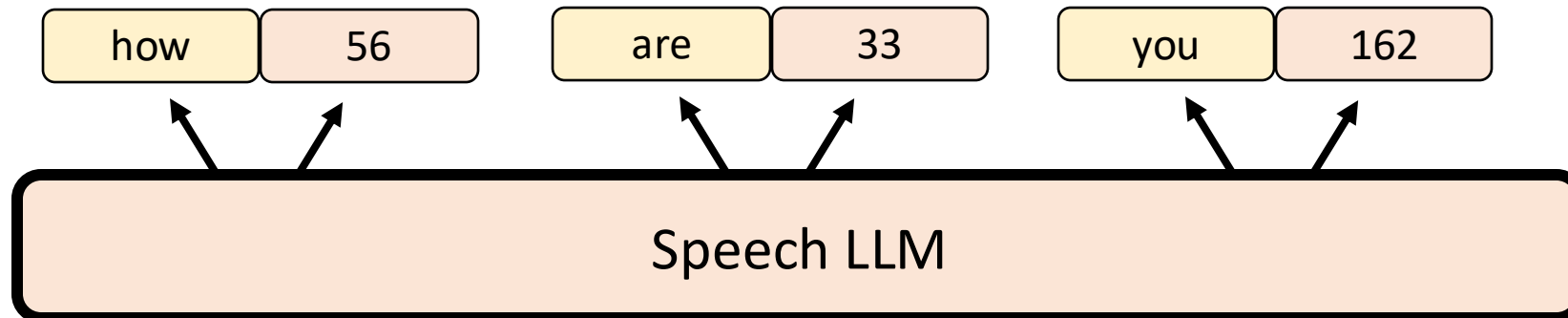


# Can we have speech tokens designed for text-speech hybrid generation?

## 1. Straightforward relationship with text (cannot be fixed duration)

e.g., one speech token corresponds to one text token

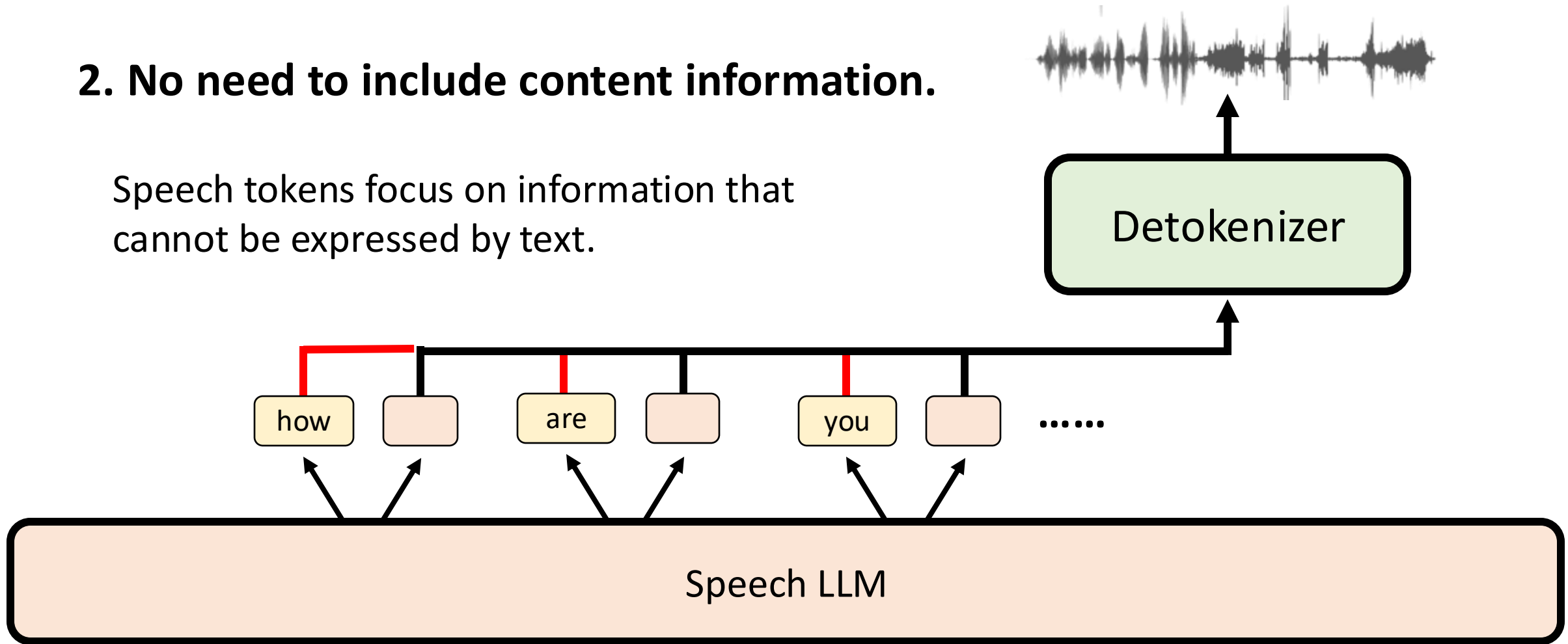
Text-speech hybrid generation can be so simple.

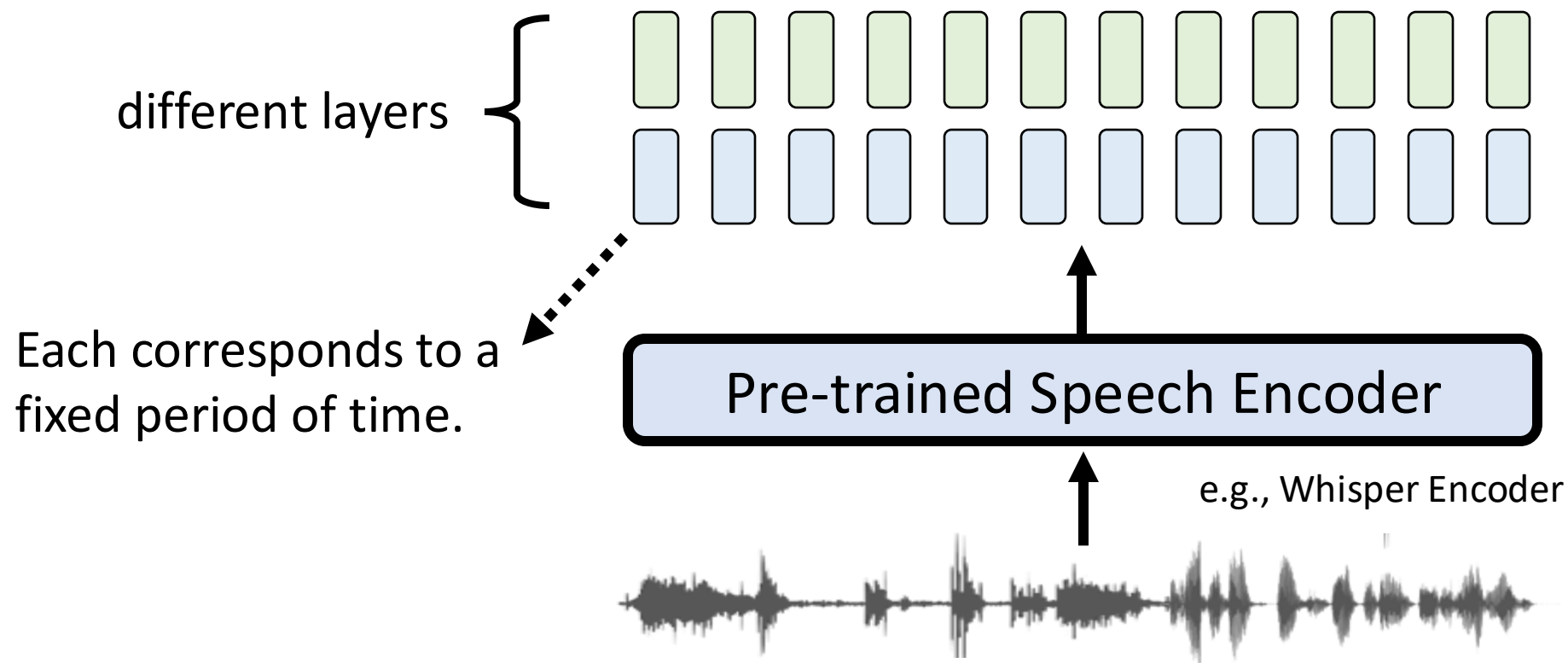


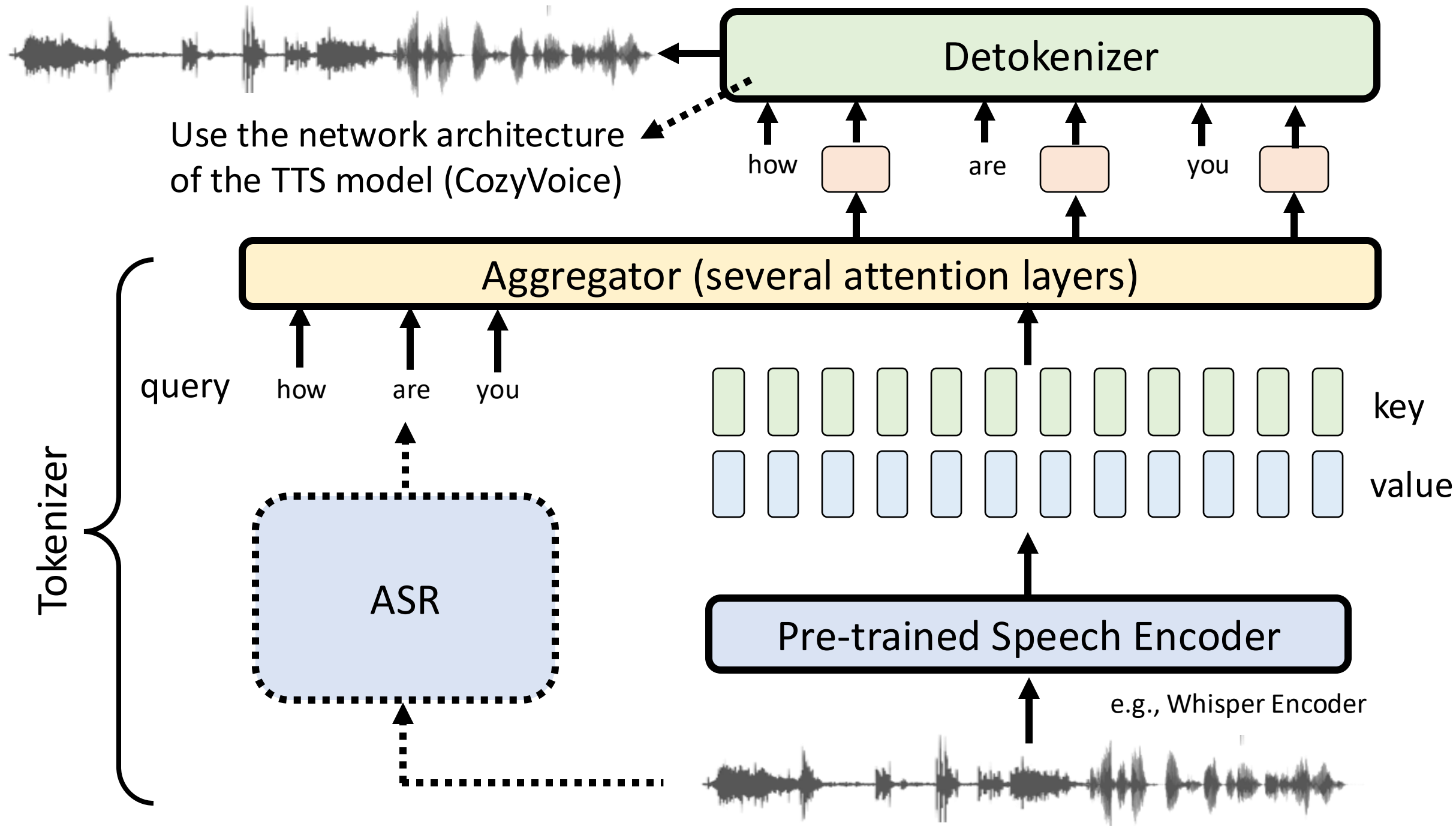
# Can we have speech tokens designed for text-speech hybrid generation?

## 2. No need to include content information.

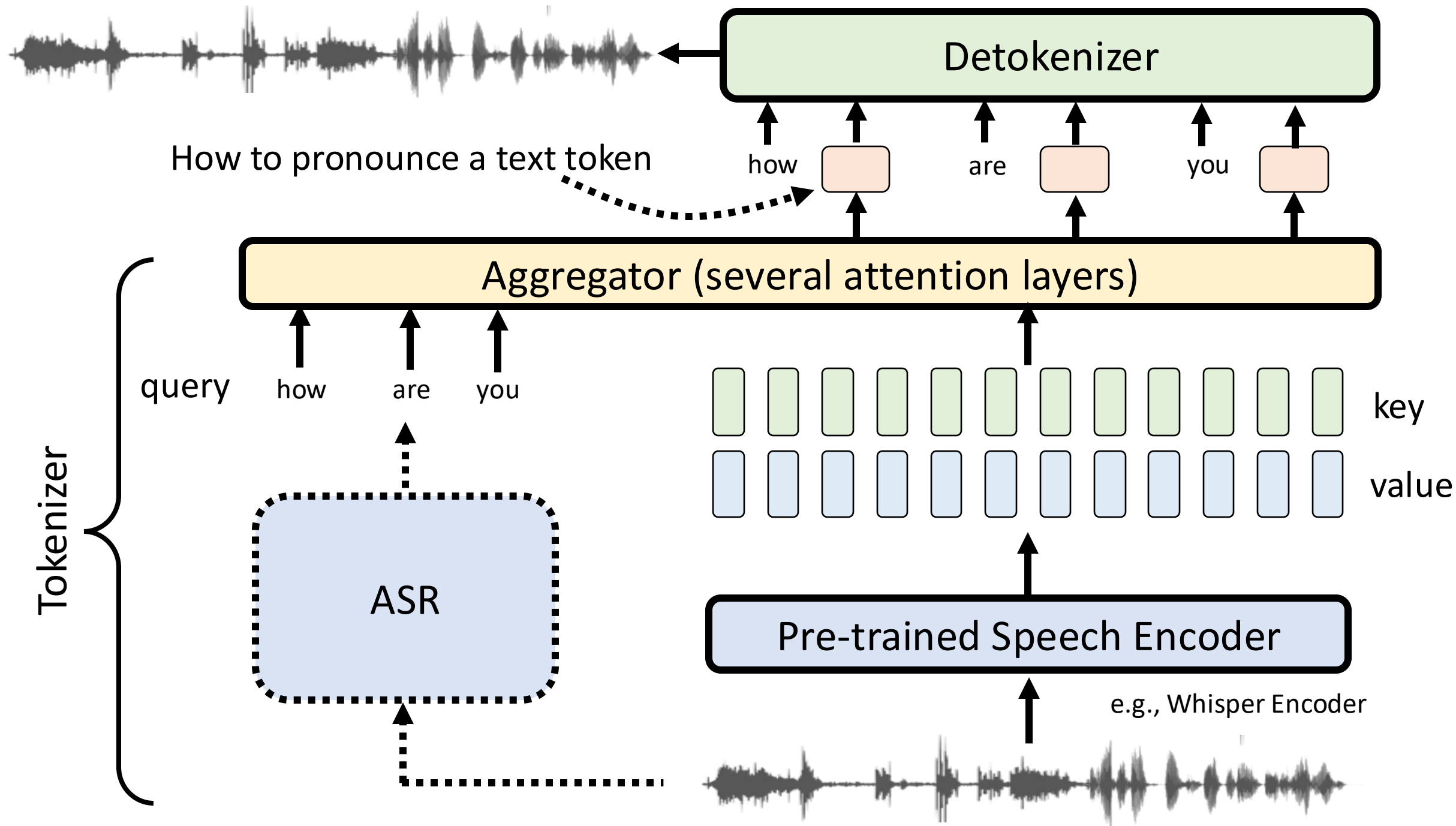
Speech tokens focus on information that cannot be expressed by text.

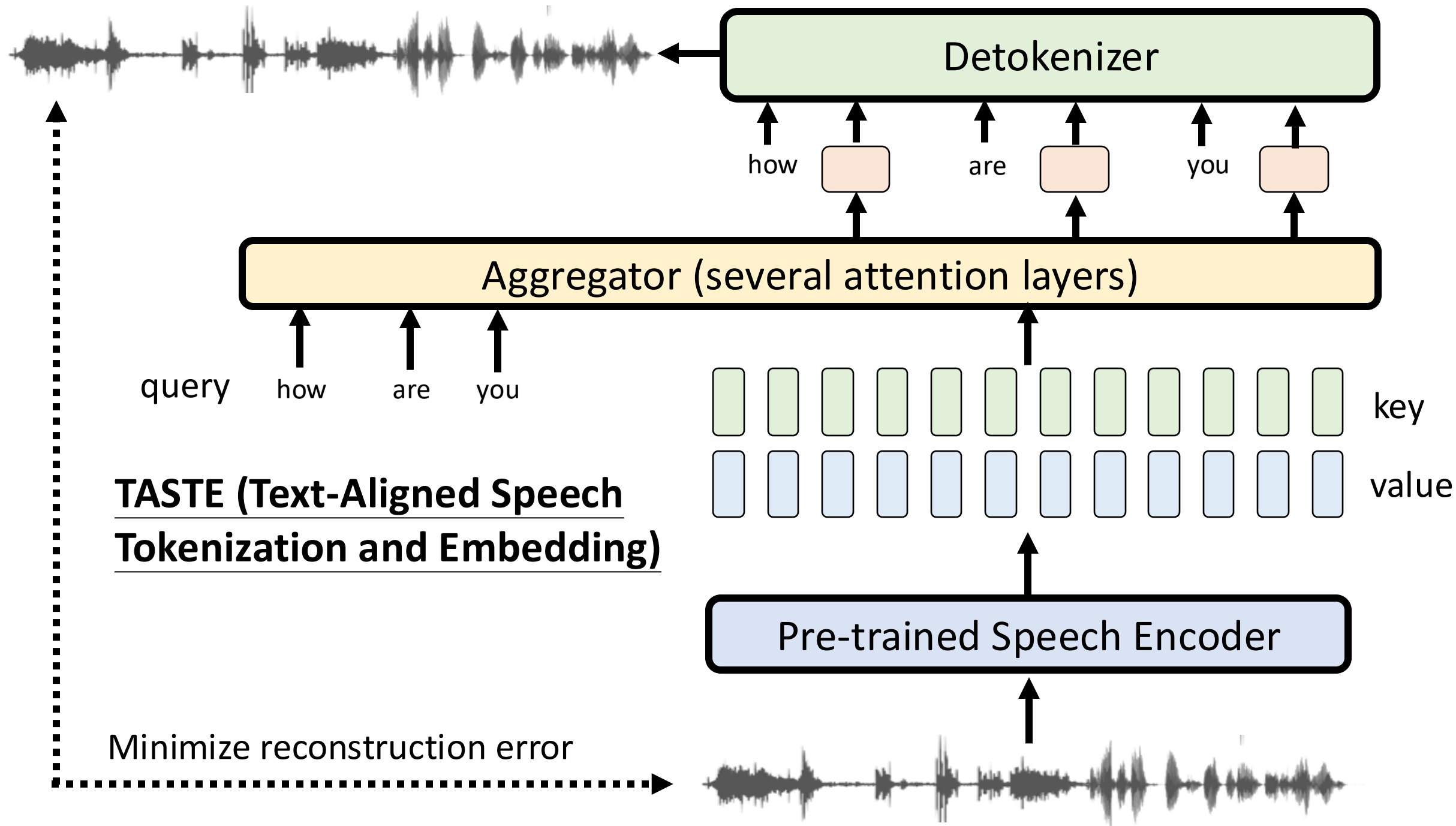


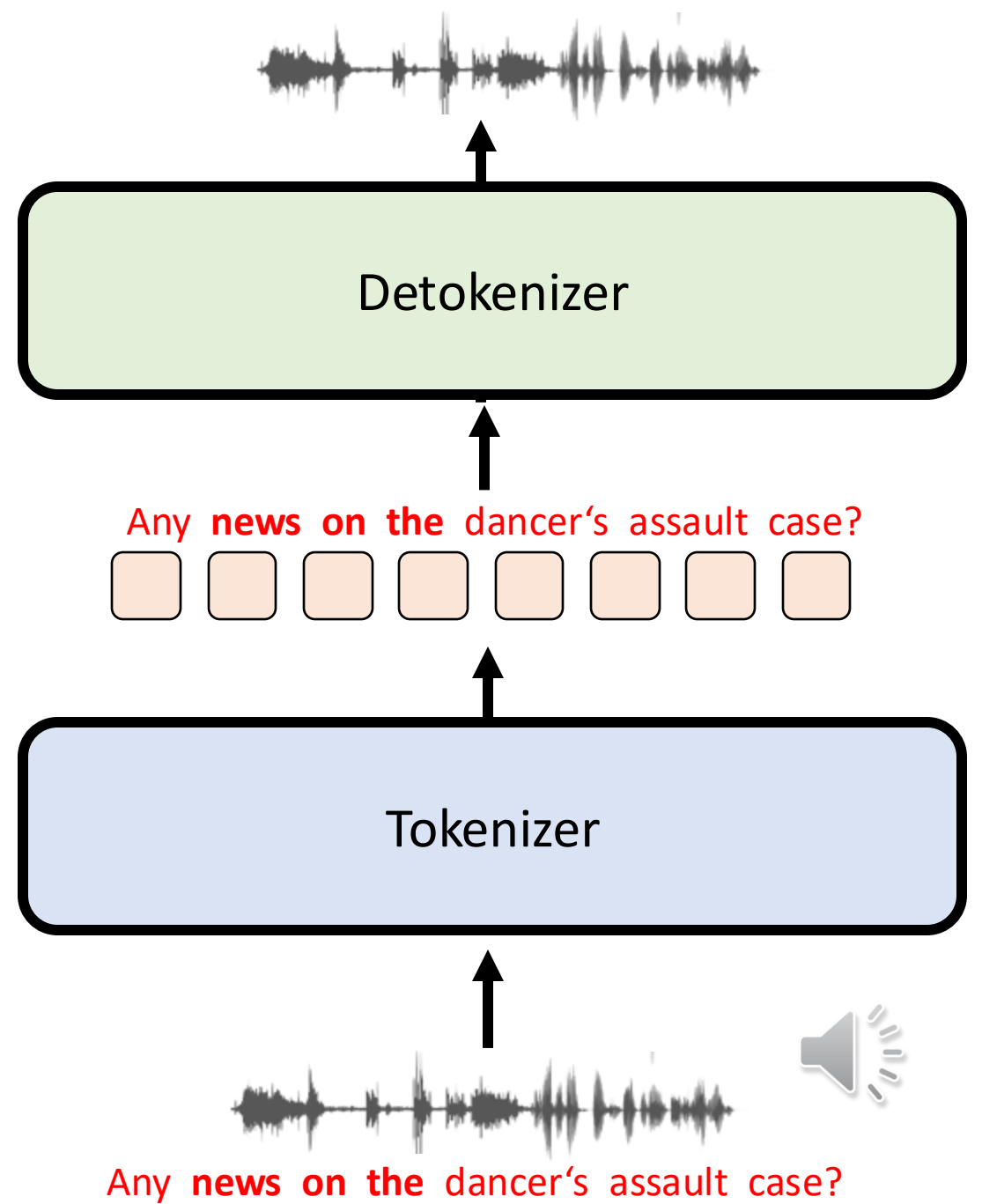
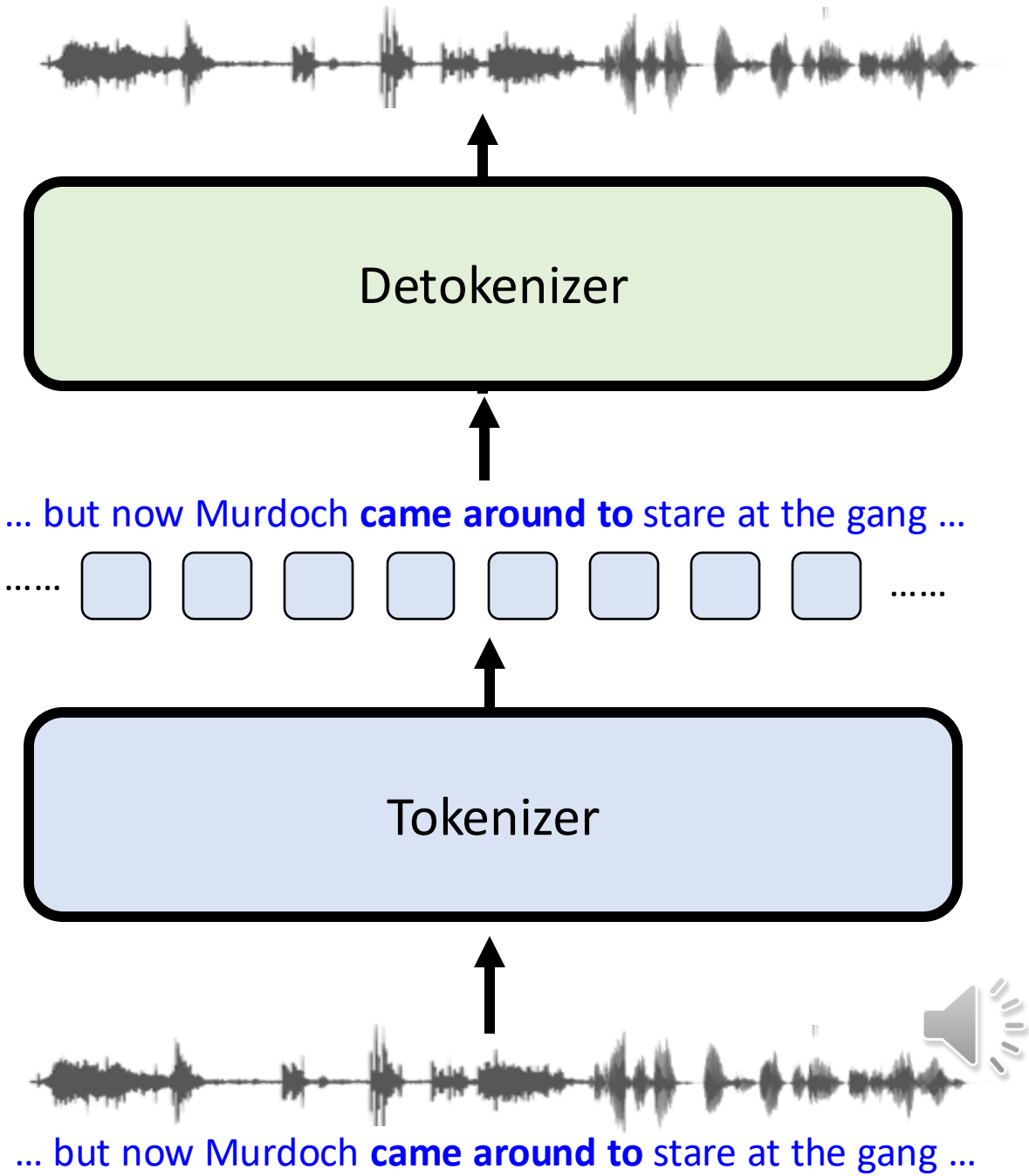


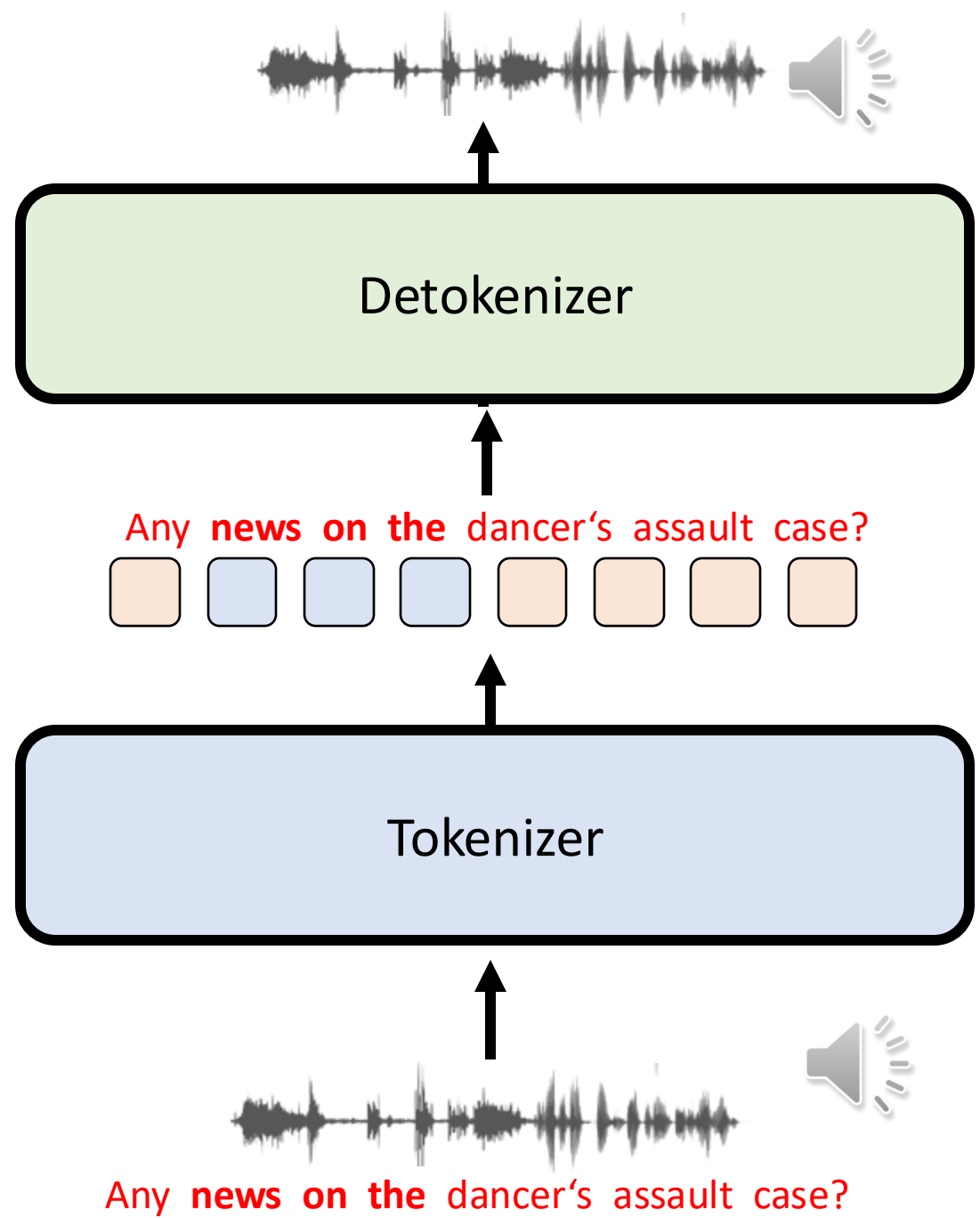
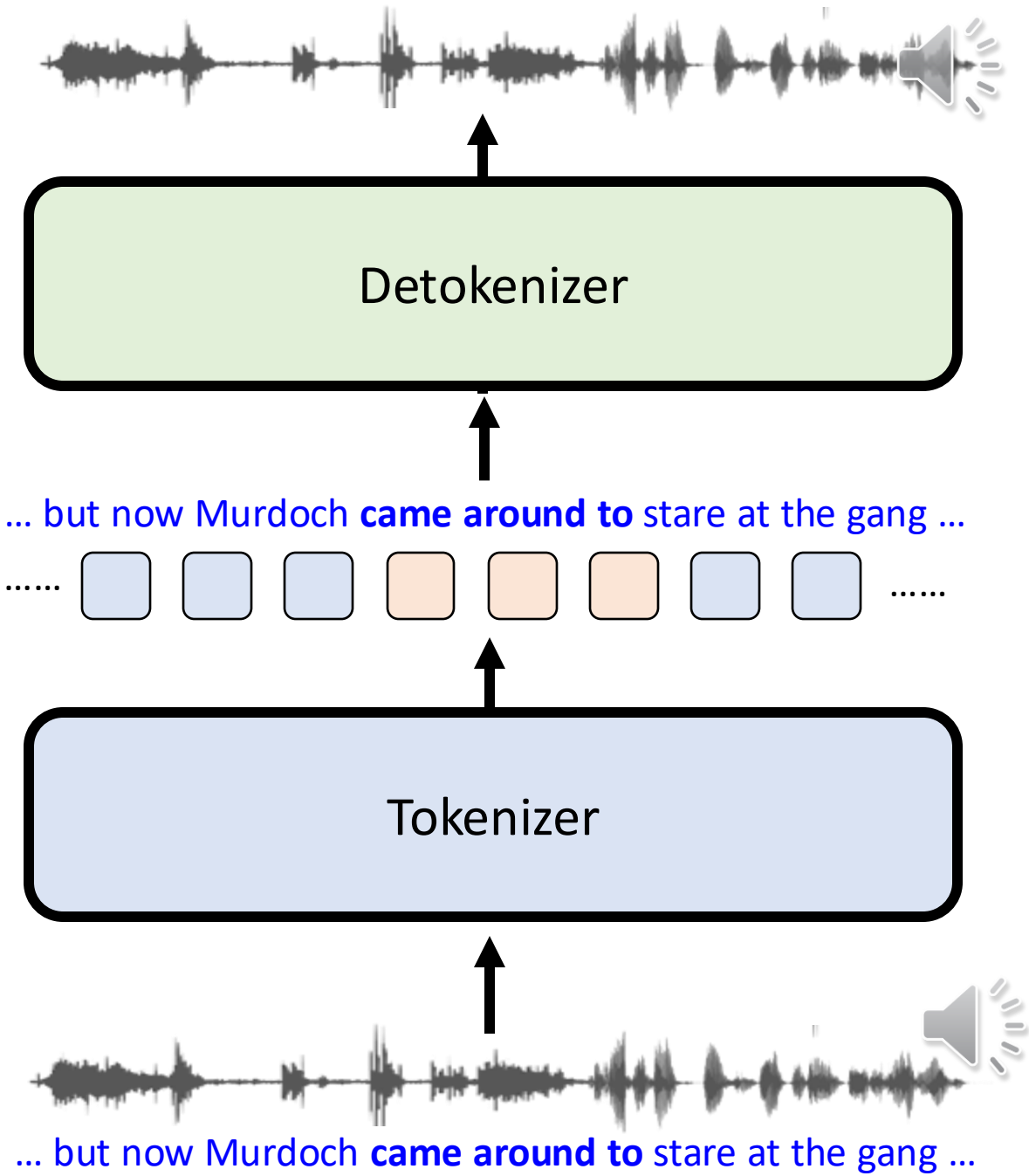




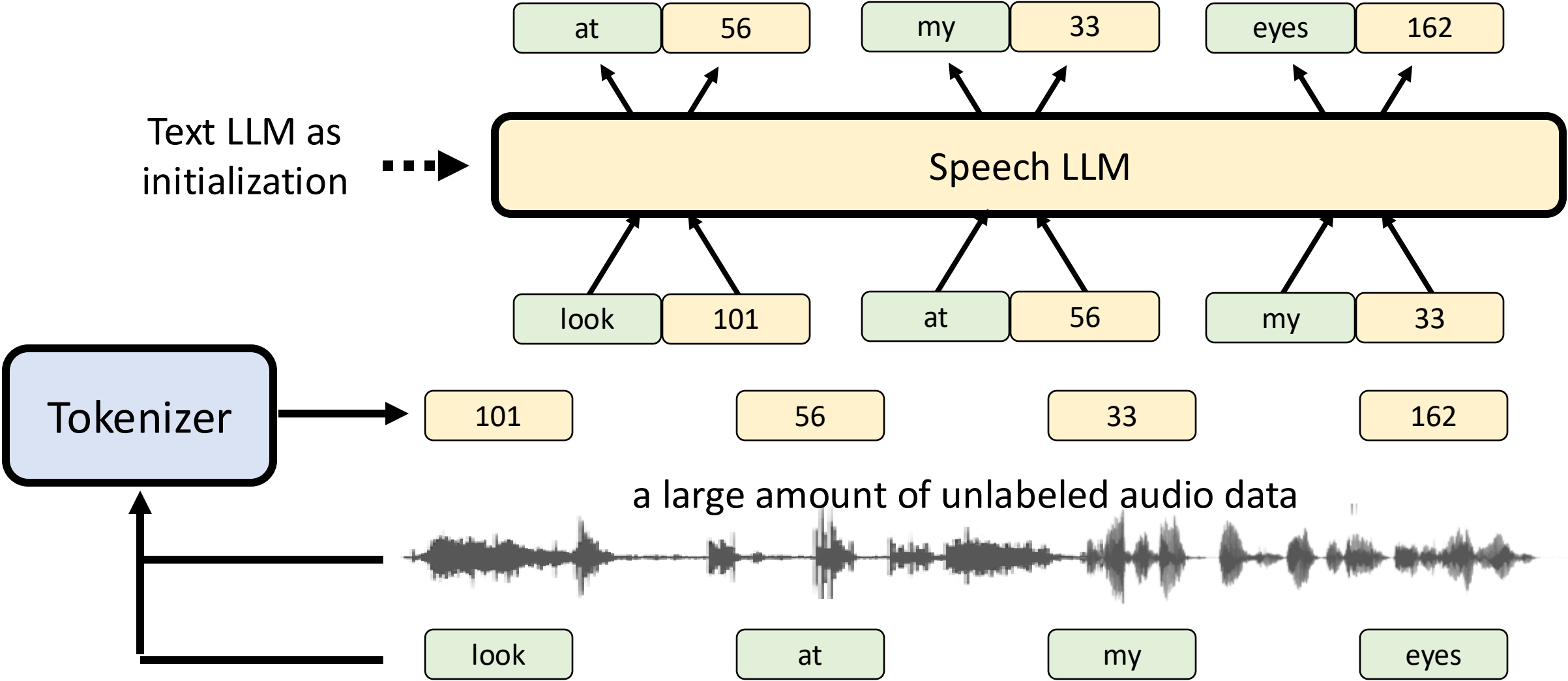




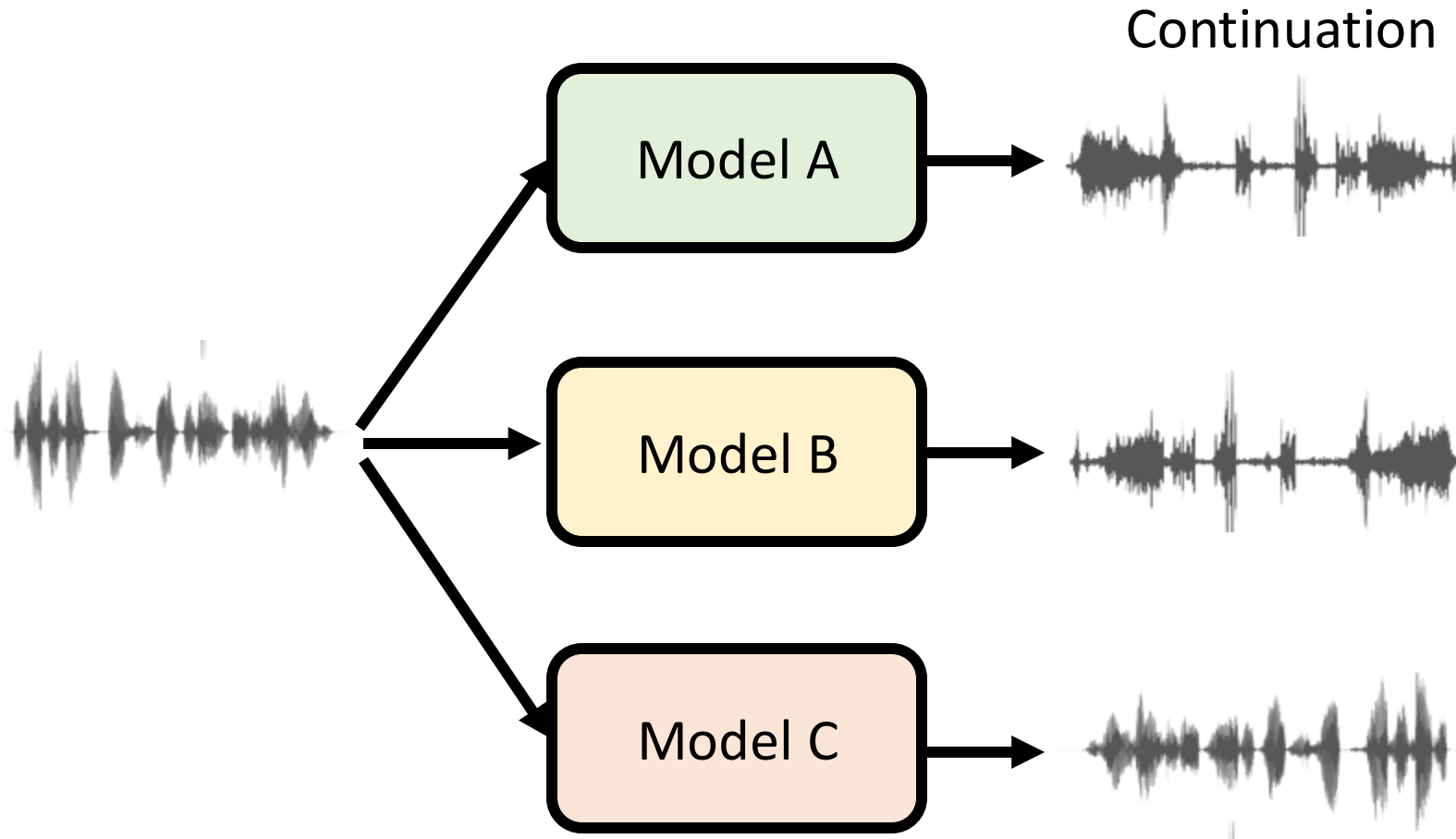




# Training Speech LLM



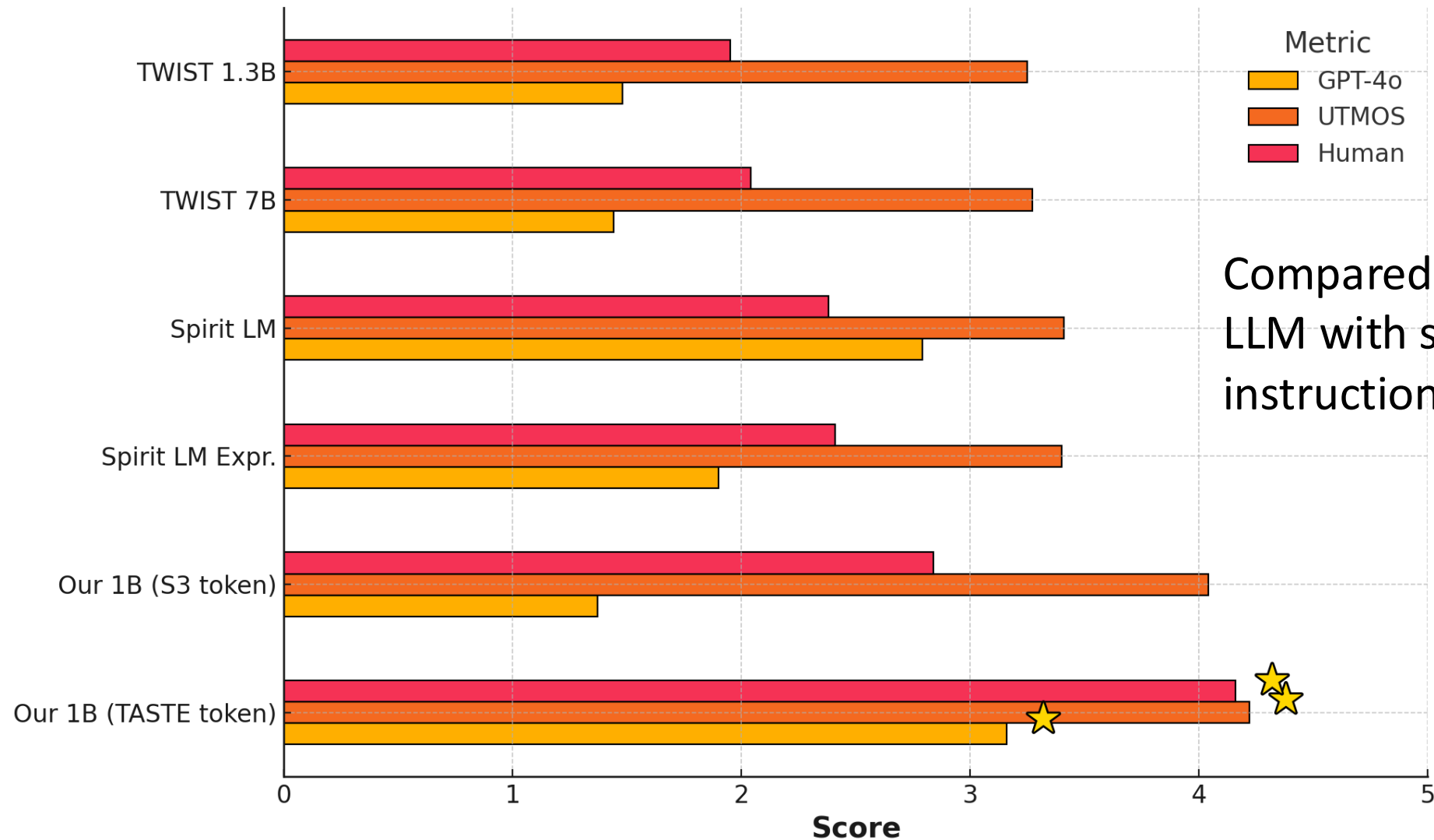
# Compared with Other **Pretrained** Models



## Evaluation

- ASR+GPT-4o: semantic coherence
- UTMOS: audio quality
- Human: how reasonable the utterances are

# Compared with Other Pretrained Models



Compared with the spoken  
LLM with supervised  
instruction fine-tuning

# English Demo



Source of video: <https://www.youtube.com/watch?v=Dc7gc7BEck0>

【生成式AI時代下的機器學習(2025)】第十二講：語言模型如何學會說話－概述語音語言模型發展歷程

<https://www.youtube.com/watch?v=gkAyqoQkOSk&t=4450s>



# Chinese Demo



大家好



你這個廢物



我是李宏毅



台灣最高的山



大家好我是李宏毅  
今天很高興來給這場演講



感謝陳竣瑋同學產生結果

# Chinese Demo



我信你個鬼



亮你好厲害  
又拿到了全學年的第一名



不要瞎掰好嗎



唉唷~我老爸得了 MVP



啊能能



感謝陳竣瑋同學產生結果

# More ....


## Dynamic Token Rate

- SyllableLM: Learning Coarse Semantic Units for Speech Language Models
  - <https://arxiv.org/abs/2410.04029>
- Sylber: Syllabic Embedding Representation of Speech from Raw Audio
  - <https://arxiv.org/abs/2410.07168>
- CodecSlime: Temporal Redundancy Compression of Neural Speech Codec via Dynamic Frame Rate
  - <https://arxiv.org/abs/2506.21074>

## Text

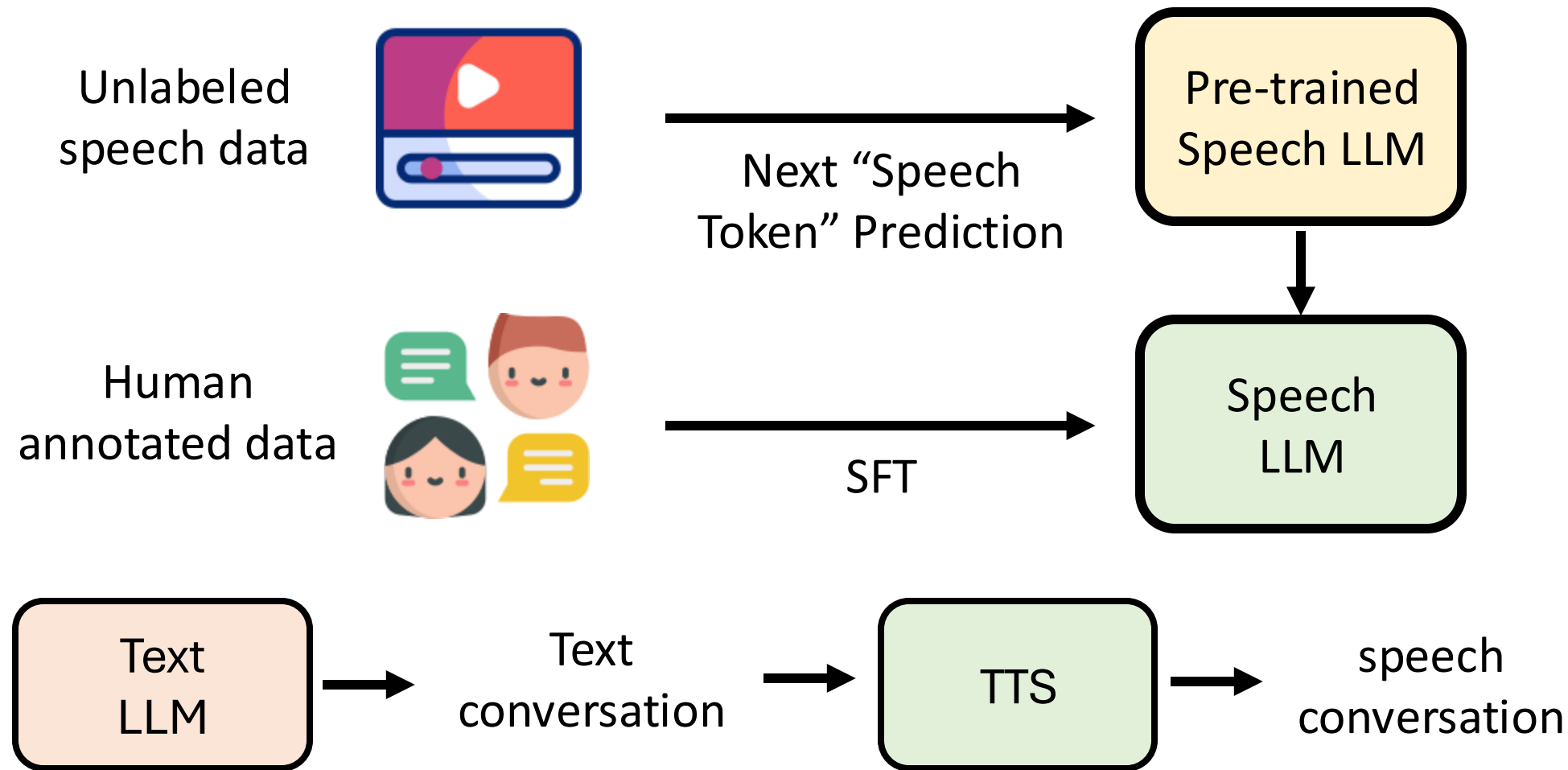
- TaDiCodec: Text-aware Diffusion Speech Tokenizer for Speech Language Modeling
  - <https://arxiv.org/abs/2508.16790>
- FlexiCodec: A Dynamic Neural Audio Codec for Low Frame Rates
  - <https://arxiv.org/abs/2510.00981>
- TASLA: Text-Aligned Speech Tokens with Multiple Layer-Aggregation
  - <https://arxiv.org/abs/2510.14934>

## Text + Dynamic Token Rate

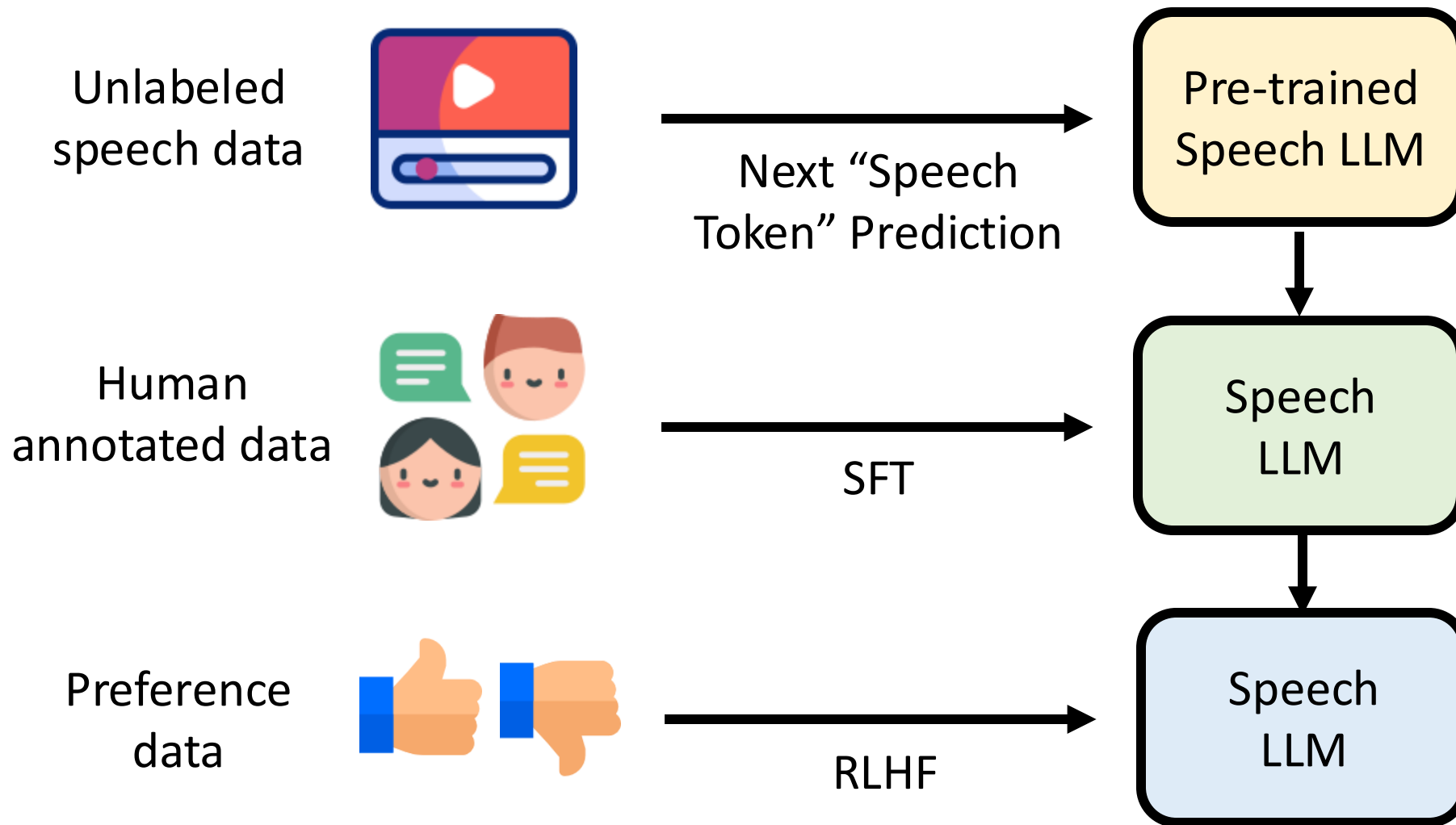
The background of the slide is a black grid with green ECG waveforms. The waveforms are arranged in four horizontal strips. The top strip is labeled 'I', 'aVR', and 'V1' in green. The second strip is labeled 'II', 'aVL', and 'V2' in green. The third strip is labeled 'III', 'aVF', and 'V3' in green. The bottom strip is labeled 'II' in green. The text '5. 語音語言模型的三部曲' is centered over the grid in white.

# 5. 語音語言模型的三部曲

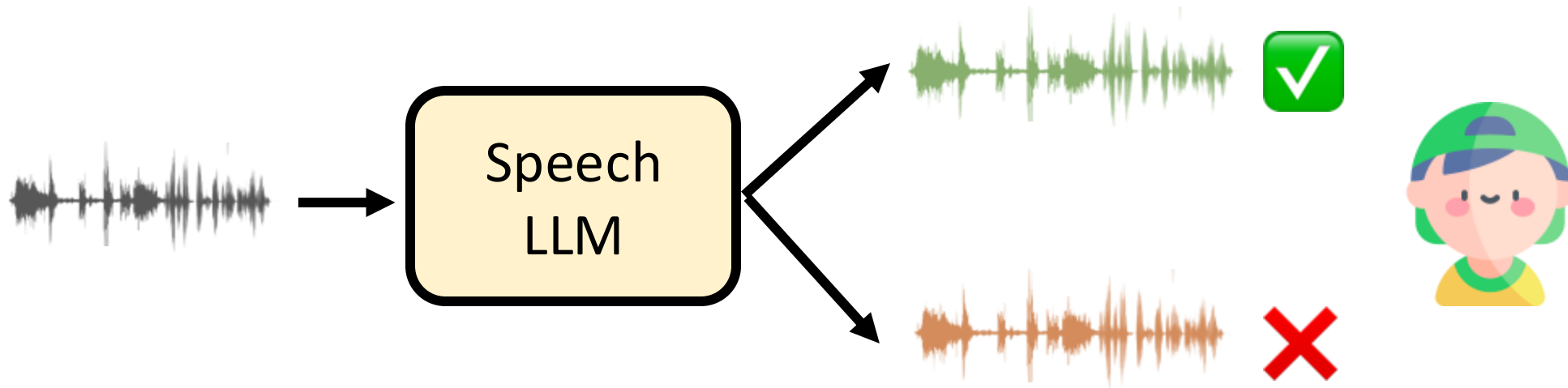
# How to Train Speech LLM



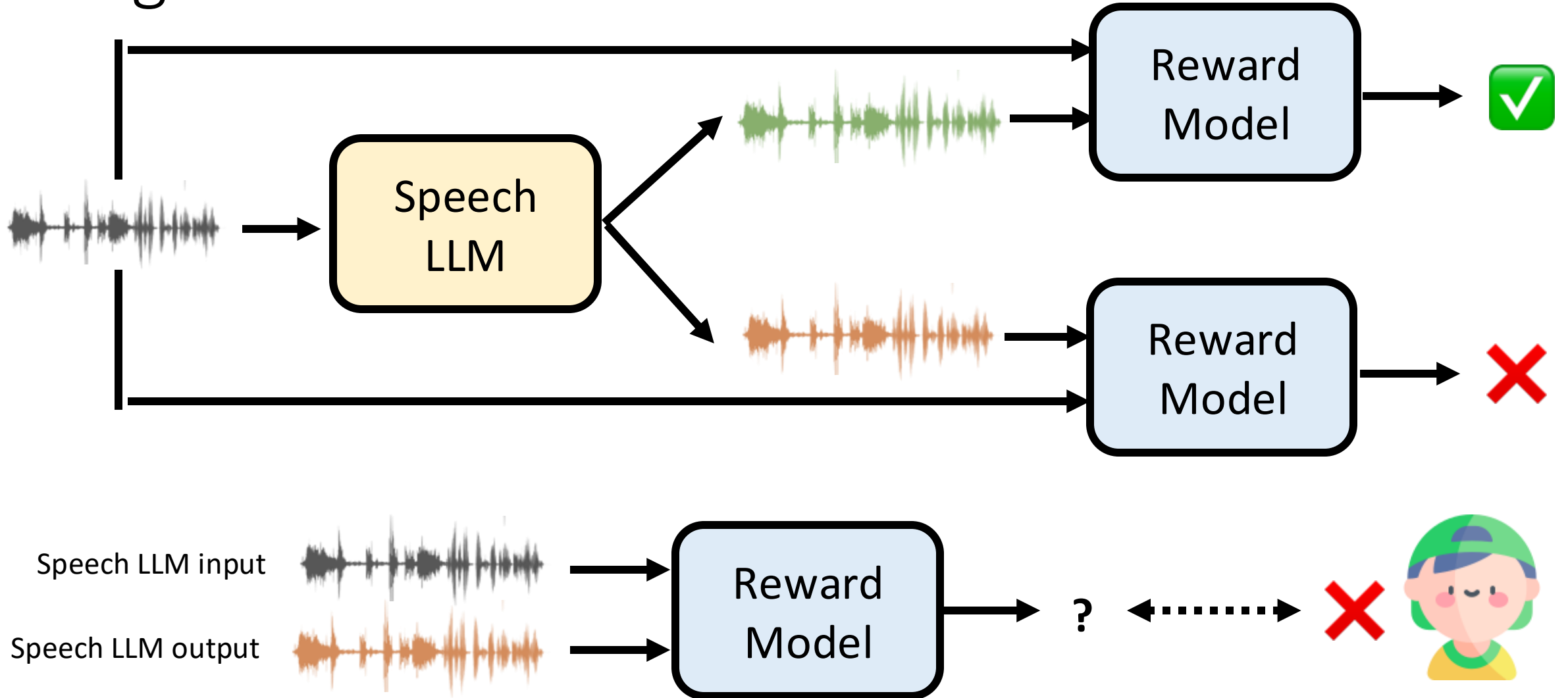
# How to Train Speech LLM



# Alignment with Feedback



# Alignment with Feedback



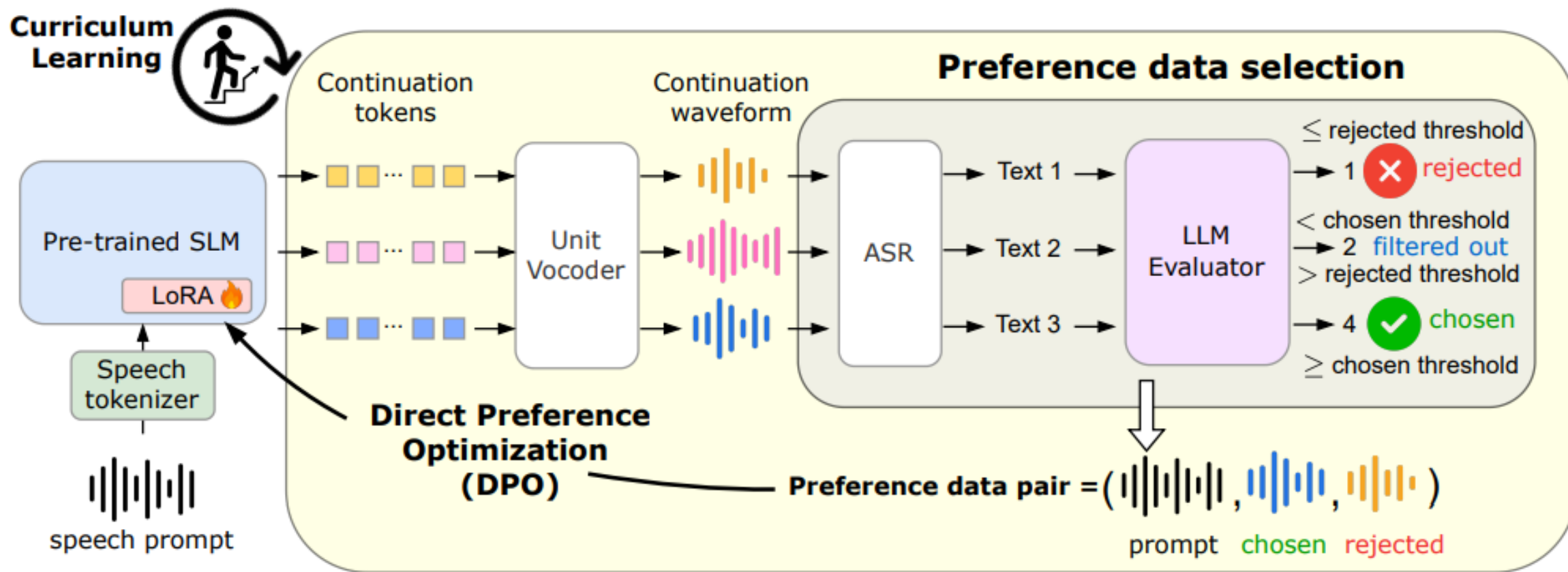


# Alignment with Feedback

Guan-Ting Lin  
(with researchers from  
the Amazon GAI team)



<https://arxiv.org/abs/2411.01834>



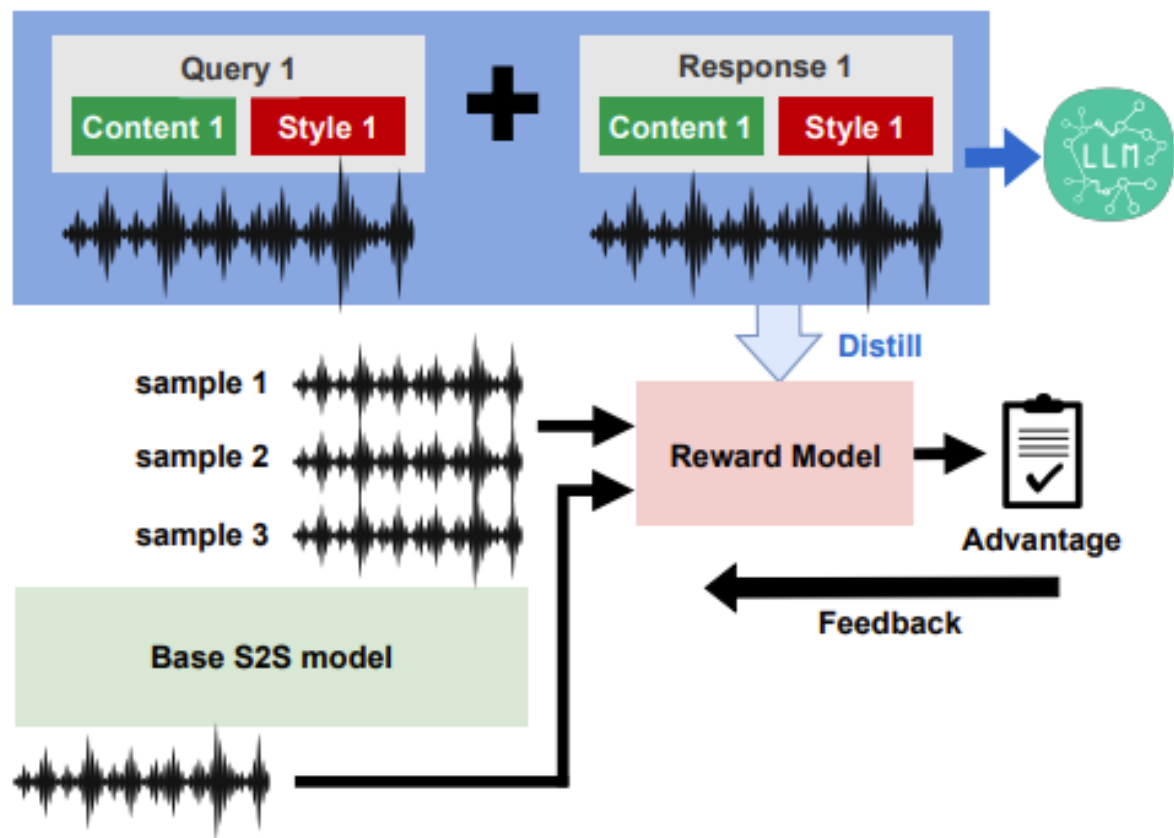
# Alignment with Feedback

Shu-wen  
(Leo) Yang

(with researchers from the  
ByteDance Seed team)



<https://arxiv.org/abs/2511.08723>



Same content



Emotion



Gender



Age

# Alignment with Feedback

- <https://paras2sbench.github.io/>



## 6. 一邊說一邊思考

# Reasoning (深度思考)

ChatGPT

告訴我人工智慧的未來

正在思考 >

立即回答

DeepSeek

告訴我人工智慧的未來



⊠ 已思考 ( 用时 4 秒 ) >

人工智慧 ( AI ) 的未來充滿無限可能，

告訴我人工智慧的未來

Claude

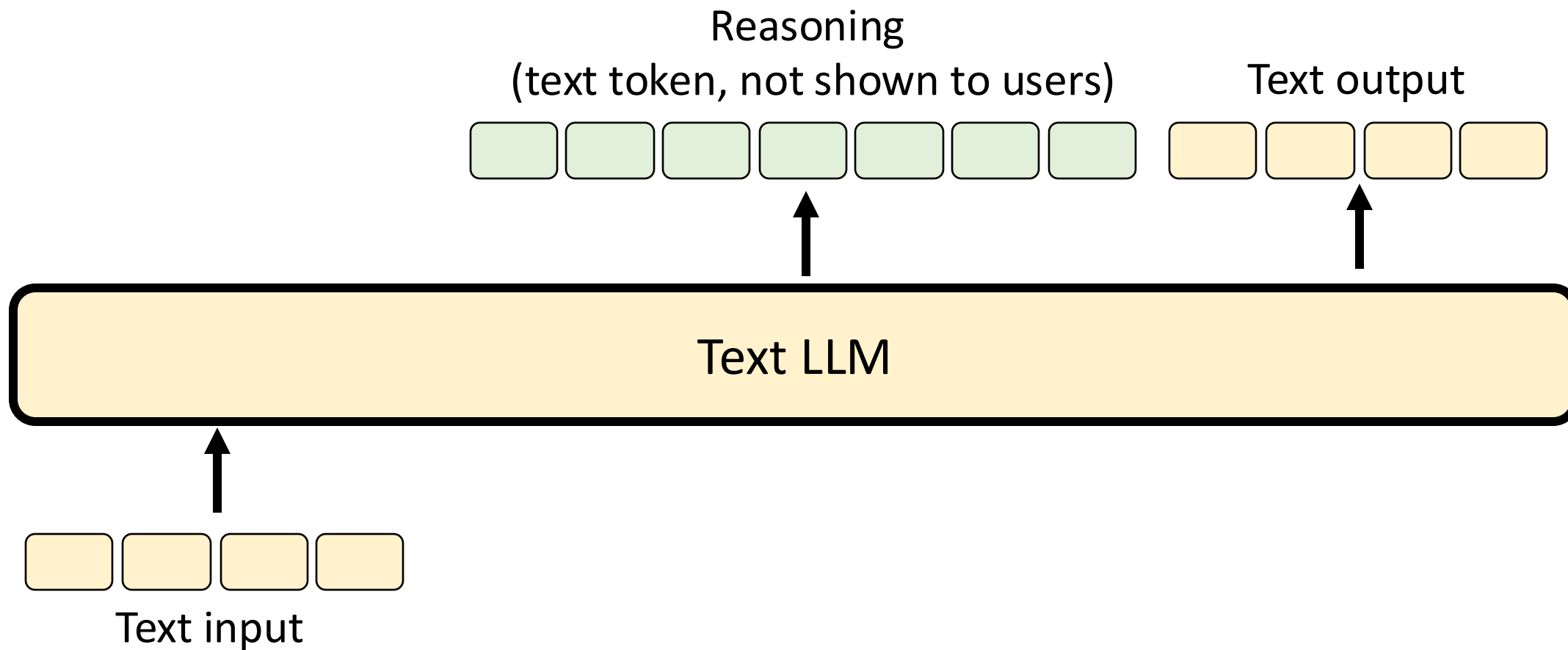
告訴我人工智慧的未來

籌劃全面回應涵蓋技術、應用與社會影響。

人工智慧的未來是個引人入勝的話題，我從幾個面向來談：

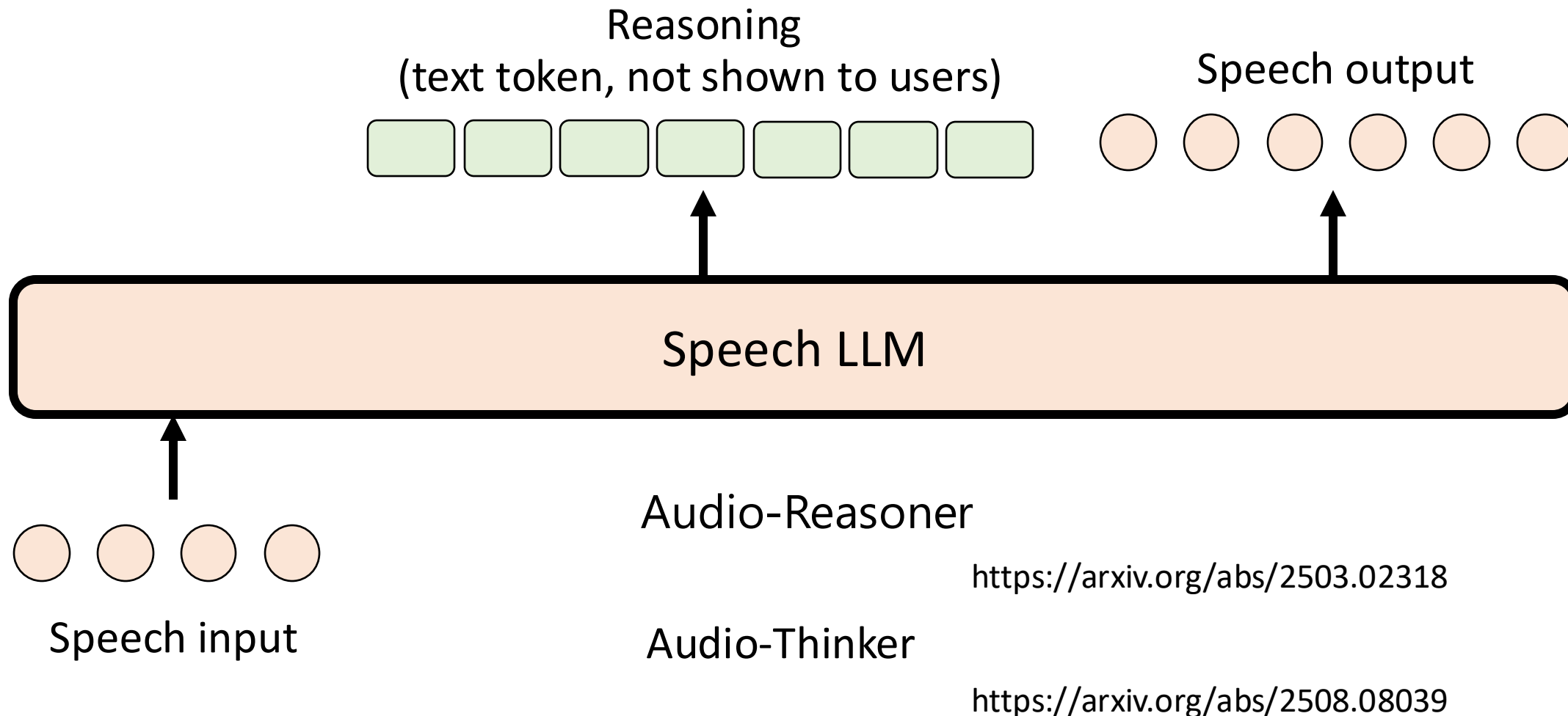
Gemini

# Text LLM + Reasoning

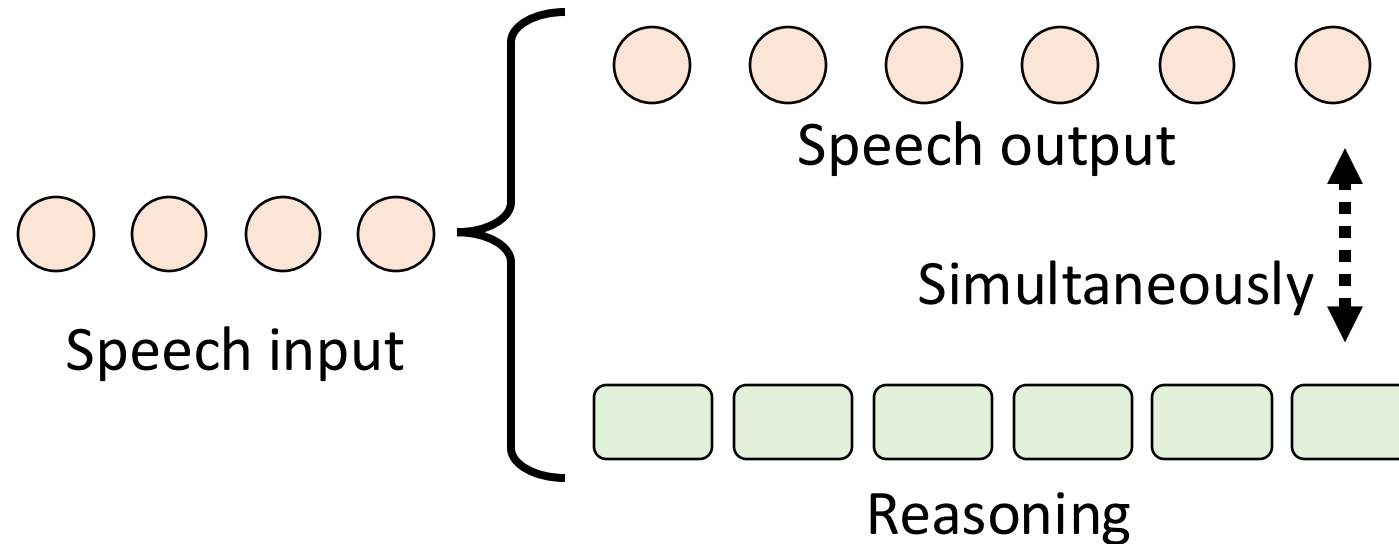
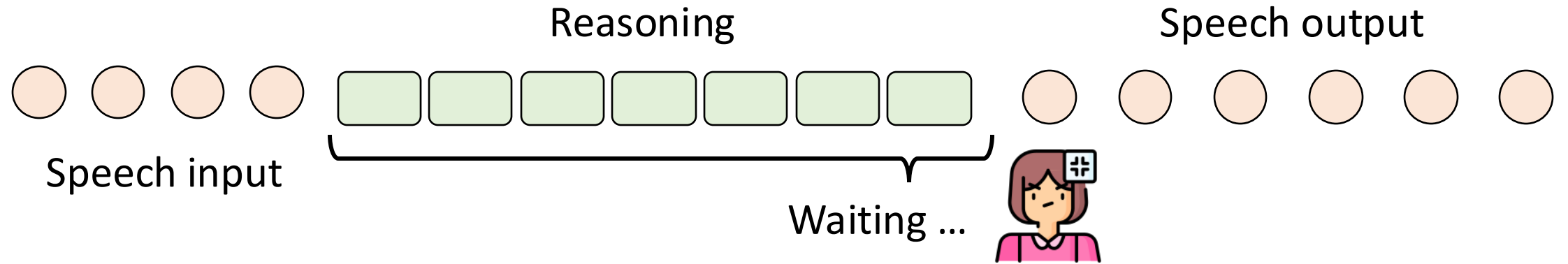


# Speech LLM + Reasoning

(Actually, text-speech hybrid generation is used, but do not show the text token here.)



# Simultaneous Thinking and Talking?



Do we need a new architecture to achieve it?

Two decoders with cross-attention?

**NO!**



# STITCH: SIMULTANEOUS THINKING AND TALKING WITH CHUNKED REASONING FOR SPOKEN LANGUAGE MODELS

**Cheng-Han Chiang<sup>1,2\*</sup> Xiaofei Wang<sup>2†</sup> Linjie Li<sup>2</sup> Chung-Ching Lin<sup>2</sup> Kevin Lin<sup>2</sup>**  
**Shujie Liu<sup>2</sup> Zhendong Wang<sup>2</sup> Zhengyuan Yang<sup>2</sup> Hung-yi Lee<sup>1</sup> Lijuan Wang<sup>2</sup>**

<sup>1</sup>National Taiwan University    <sup>2</sup>Microsoft

<https://arxiv.org/abs/2507.15375>

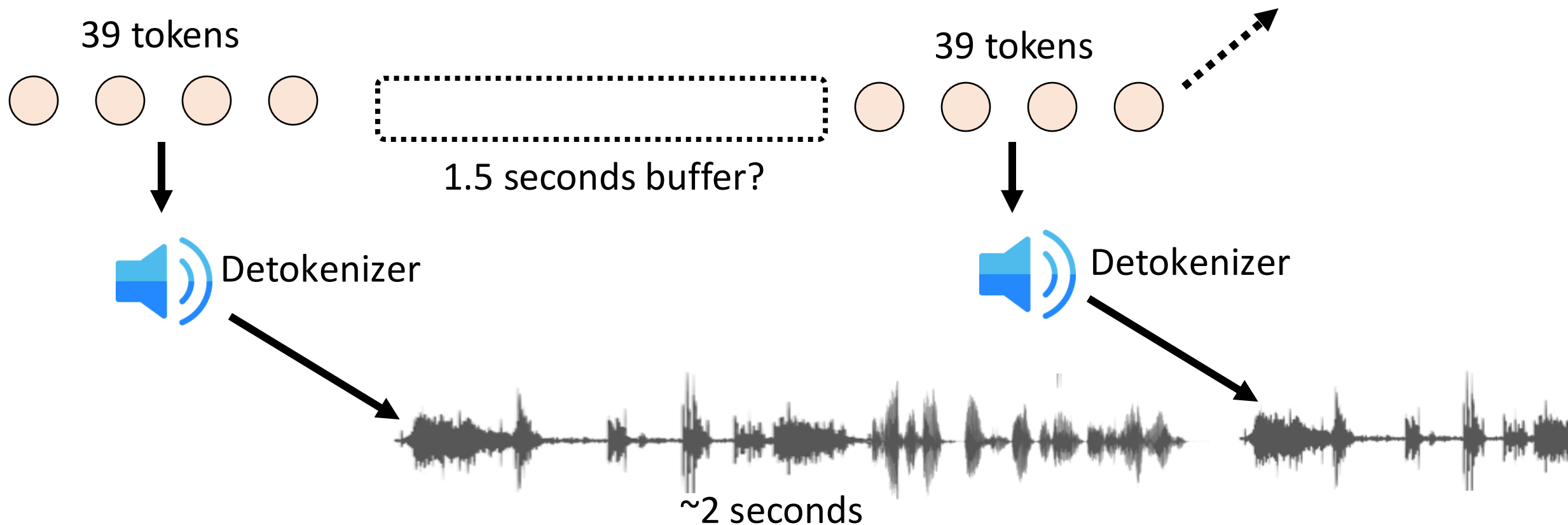
Cheng-Han Chiang and  
Microsoft Researchers



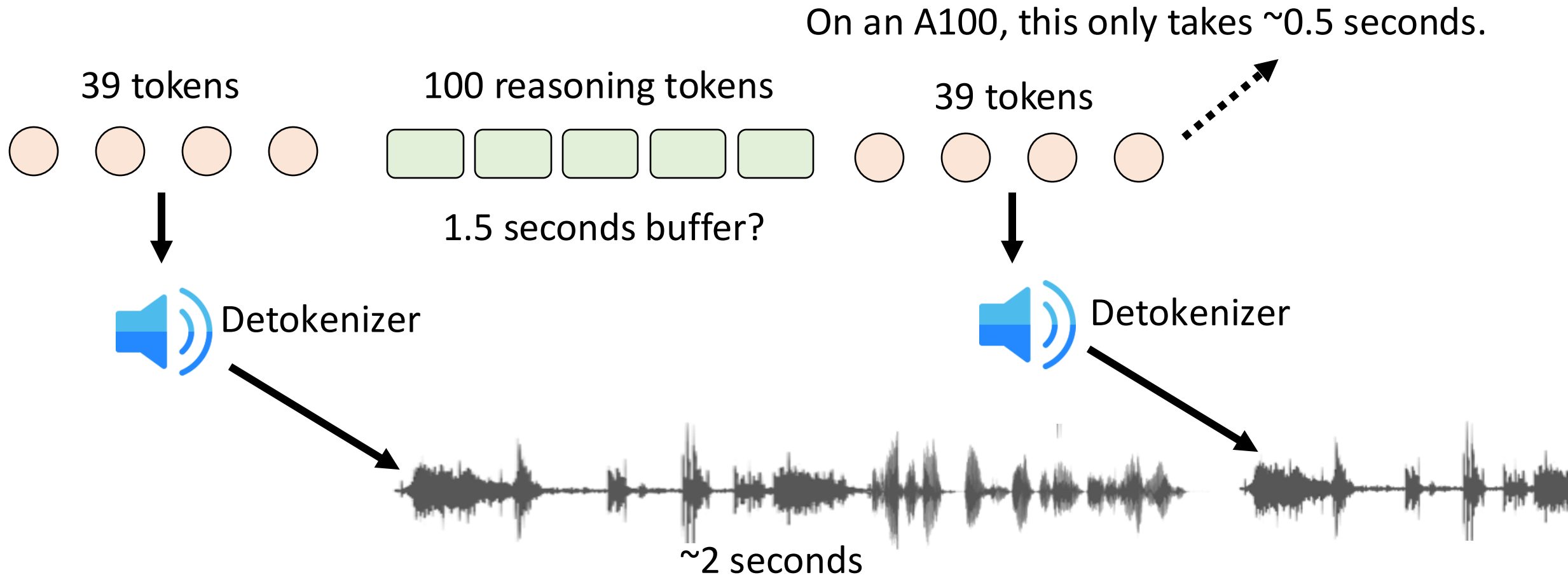
# STITCH: Simultaneous Thinking and Talking with Chunked Reasoning

Using GLM-4-Voice-9B as example

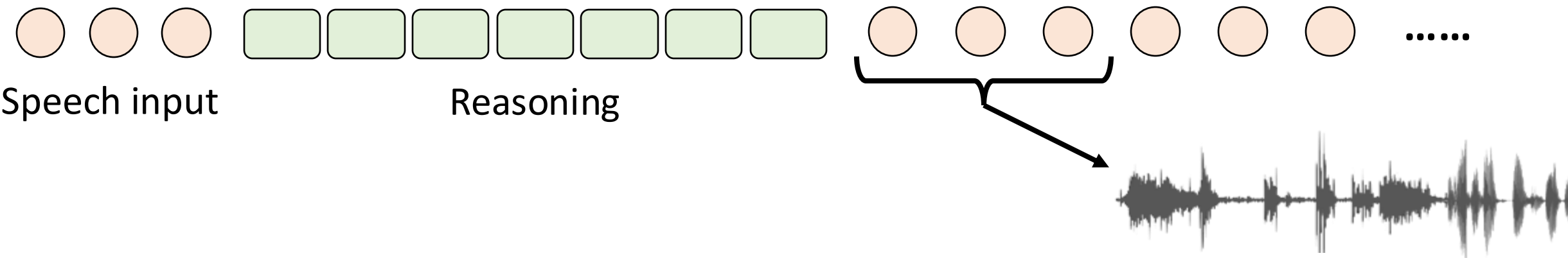
On an A100, this only takes ~0.5 seconds.



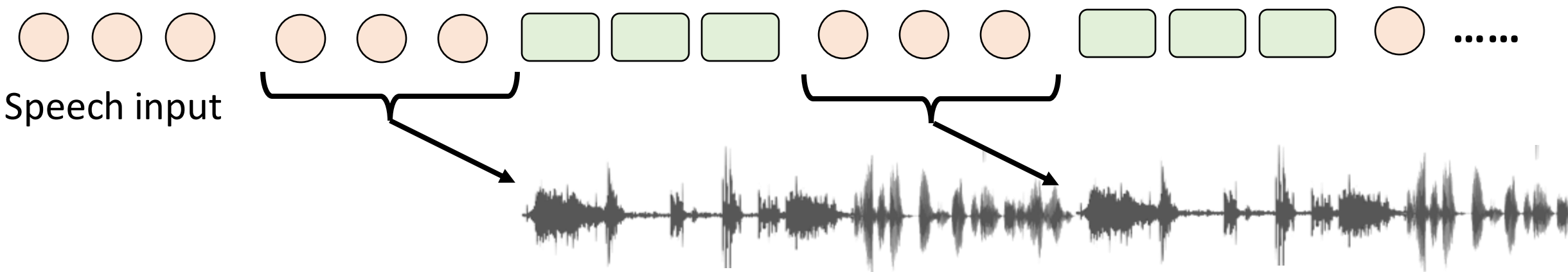
# STITCH: Simultaneous Thinking and Talking with Chunked Reasoning



# Typical Speech LLM + Reasoning

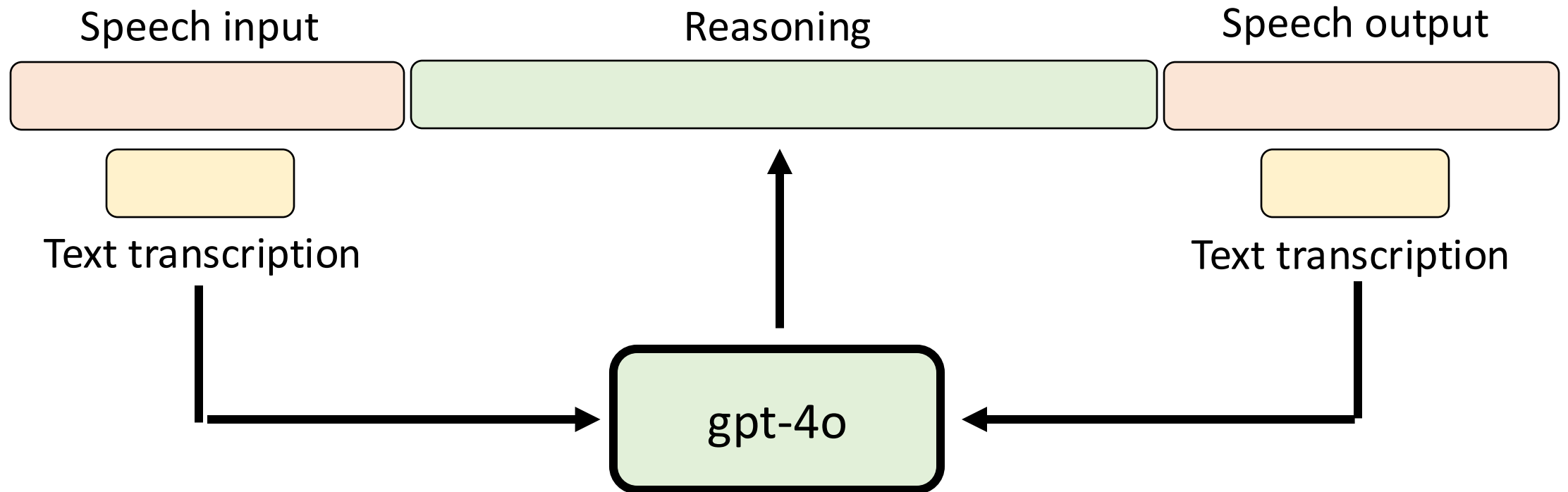


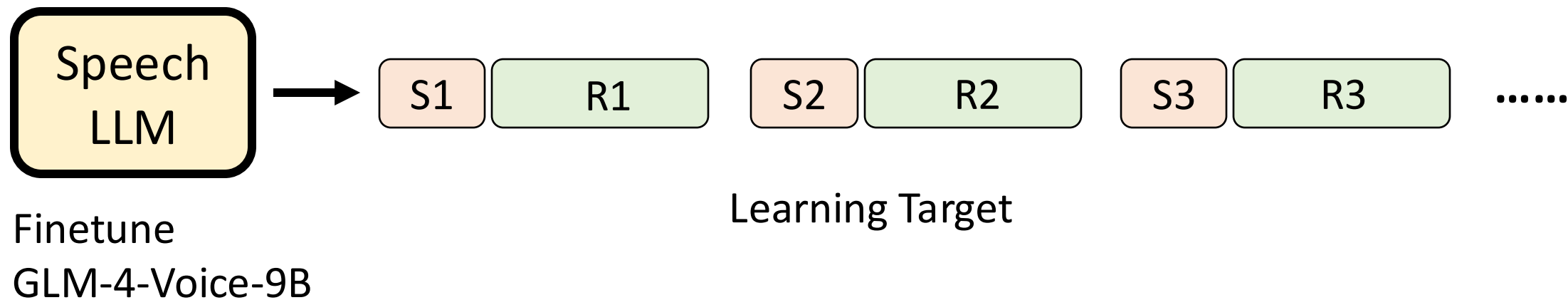
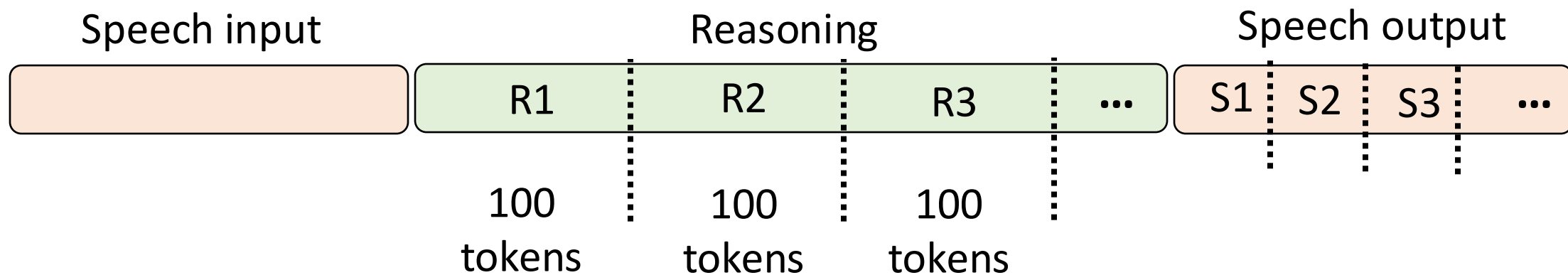
# STITCH

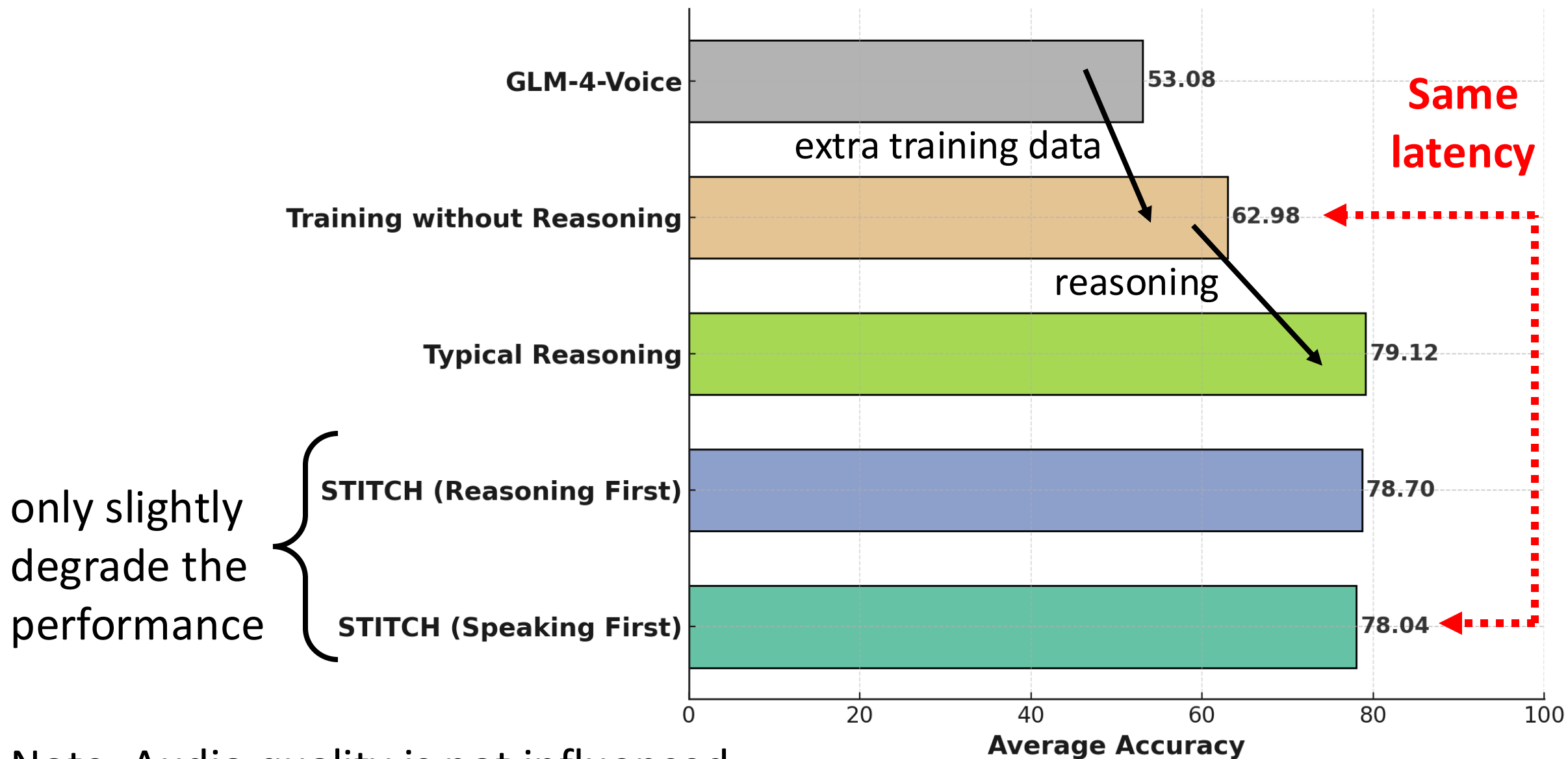


# How to prepare training data

- There is already some speech-to-speech dialogue datasets (e.g., VoiceAssistant400K)

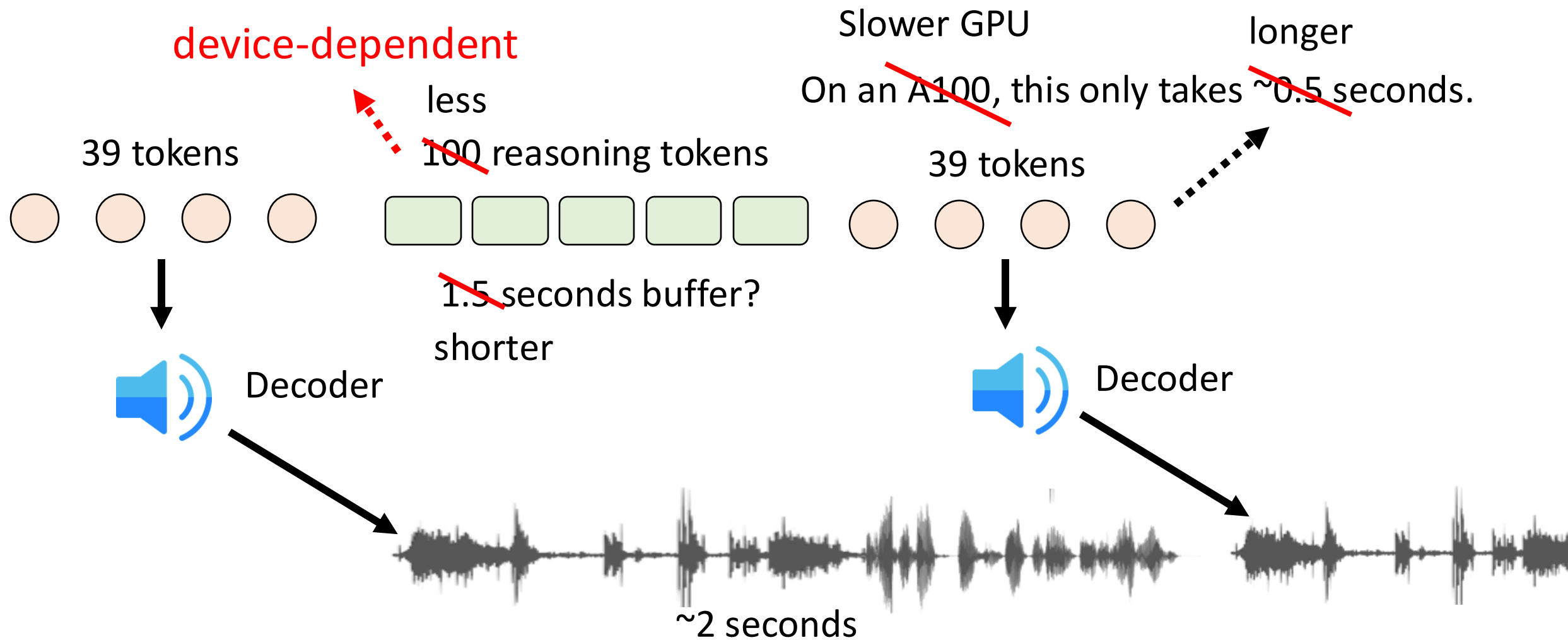






Note: Audio quality is not influenced.

# Dynamic Reasoning Chunk Sizes

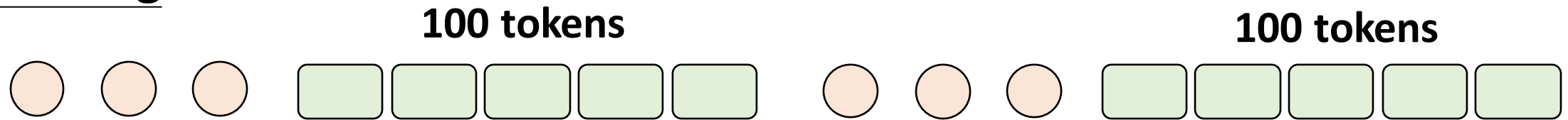




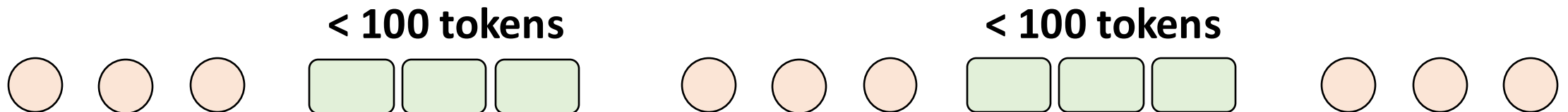
# Dynamic Reasoning Chunk Sizes

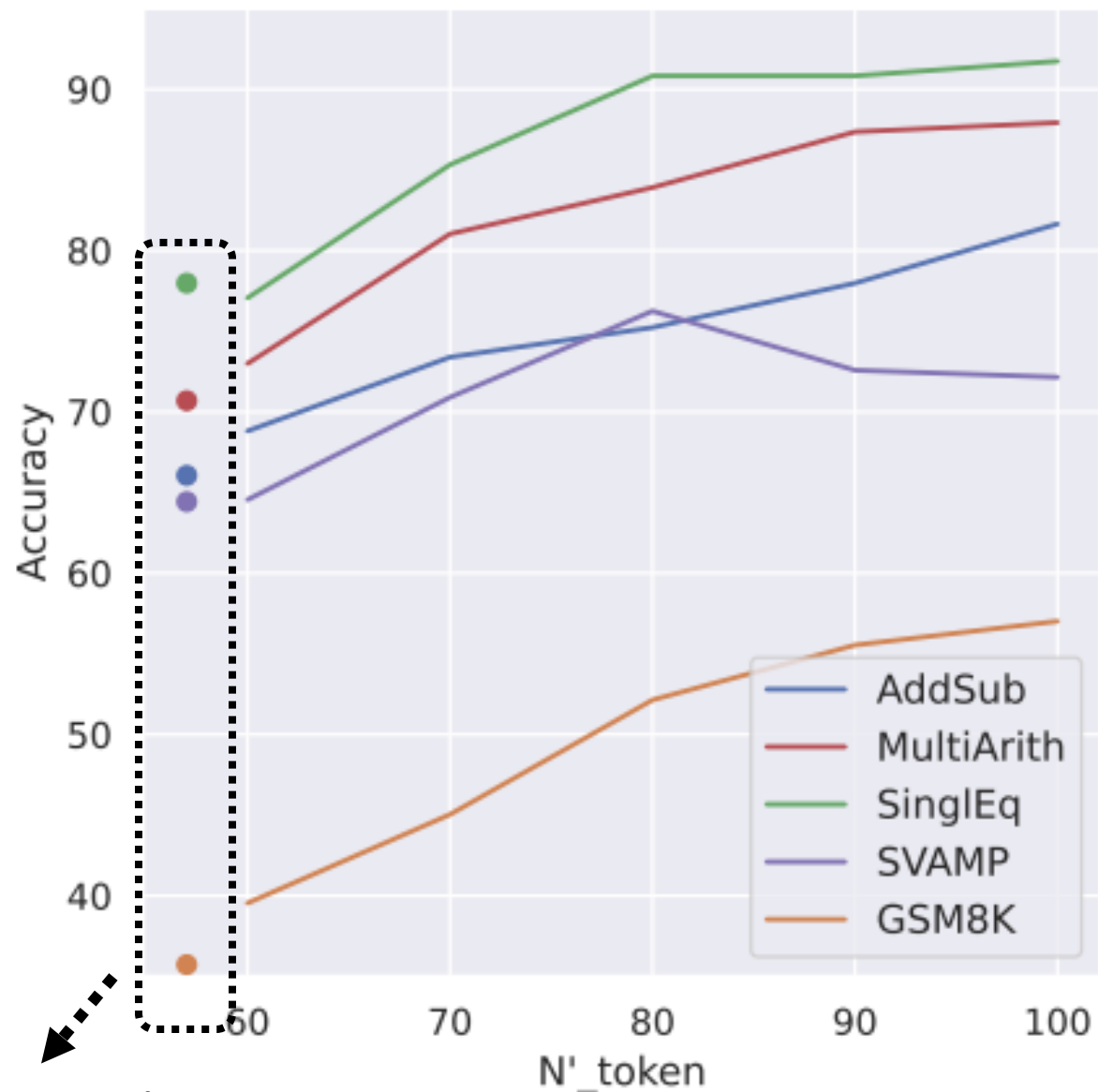
What would happen if the training and testing reasoning chunks have different numbers of tokens?

## Training



## Testing

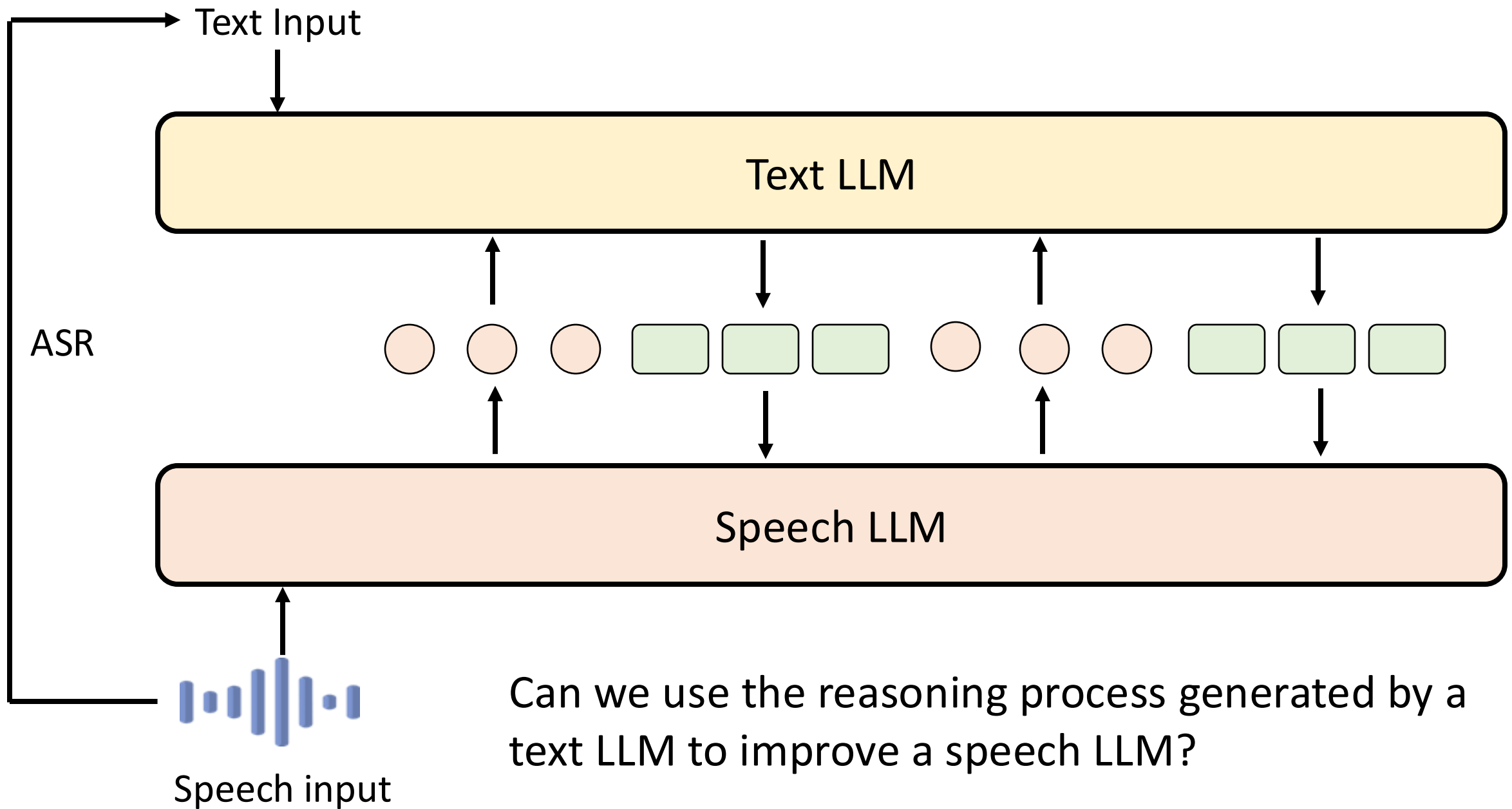


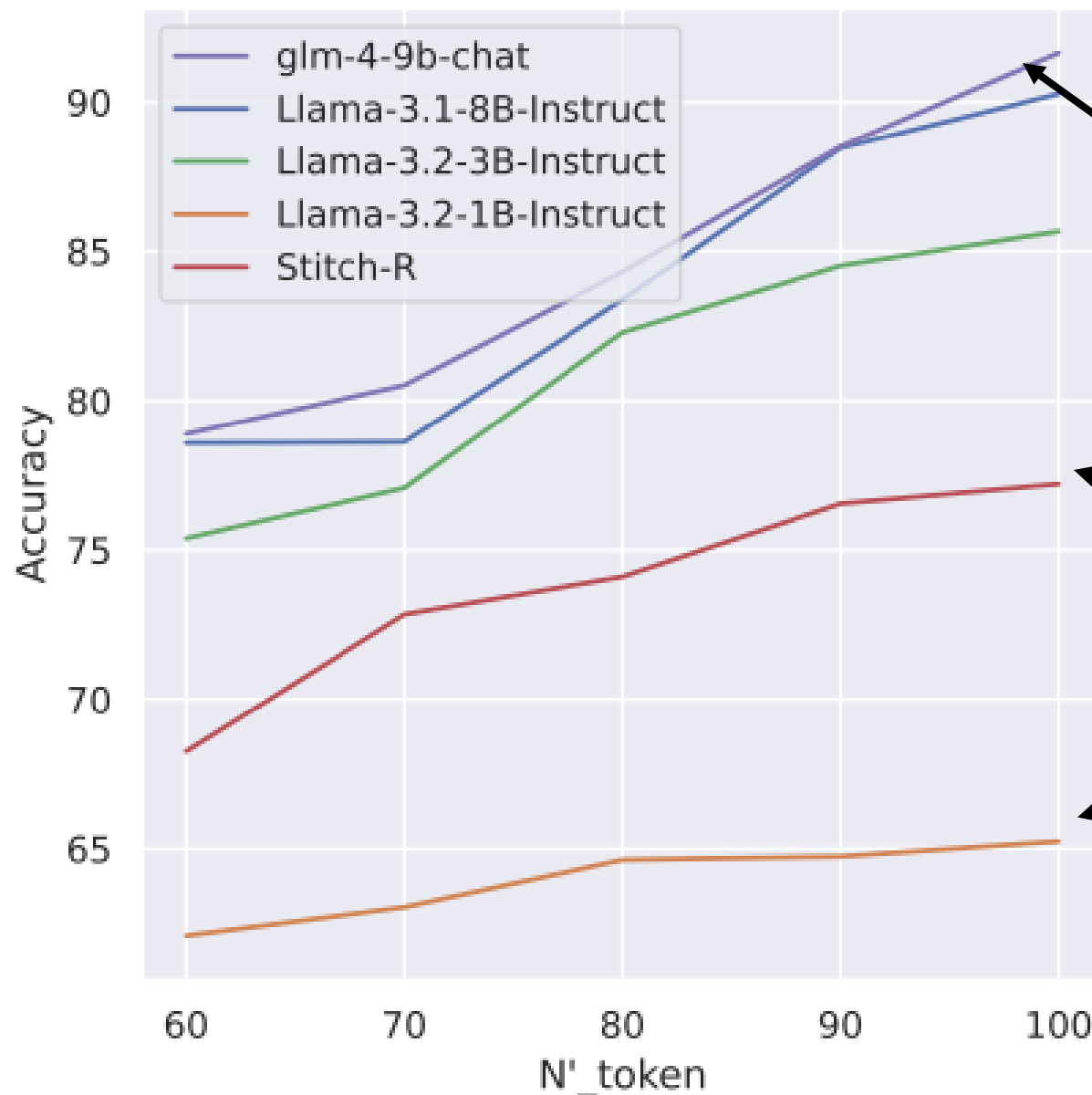


No reasoning

Tokens in each reasoning chunk during inference

- Shorter reasoning chunks yield to worse performances
- Still better than “no reasoning” in many cases





GLM-4-9B-Chat provided reasoning.

Reasoning by speech LLM itself  
(Finetune GLM-4-Voice-9B)

Llama-3.2-1B-instruct

**The speech LLM did leverage  
the reasoning tokens.**

Tokens in each reasoning chunk during inference

# Demo Video

```
User: Claire makes a 3 egg omelet every morning for breakfast. How many dozens of eggs will she eat in 4 weeks?
```

Source: <https://d223302.github.io/STITCH/>

# More .....

- Mini-Omni-Reasoner: Token-Level Thinking-in-Speaking in Large Speech Models
  - <https://arxiv.org/abs/2508.15827>
- Stream RAG: Instant and Accurate Spoken Dialogue Systems with Streaming Tool Usage
  - <https://arxiv.org/abs/2510.02044>
- SHANKS: Simultaneous Hearing and Thinking for Spoken Language Models
  - <https://arxiv.org/abs/2510.06917>
- Can SpeechLLMs Think while Listening?
  - <https://arxiv.org/abs/2510.07497>

# Concluding Remarks

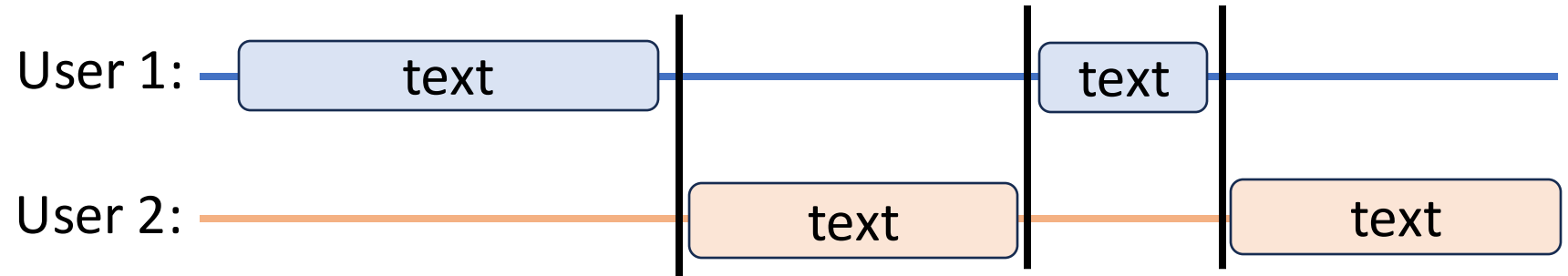
1. 序章：如何有效表示語音
2. 初代語音語言模型
3. 如何利用強大的文字模型
4. 尋找更合適的語音表示方式
5. 語音語言模型的三部曲
6. 一邊說一邊思考

其實還有很多沒有講到的技術 .....

# Beyond the Turn-based Game

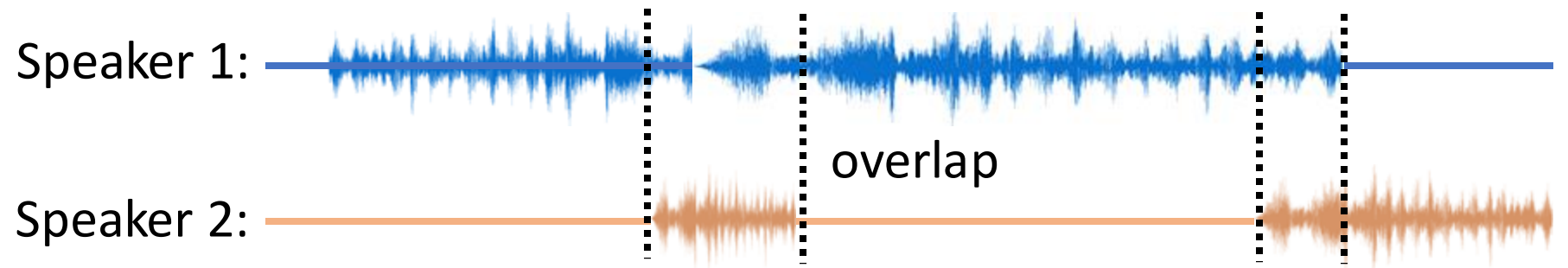
## Text Conversation

*Turn-based*

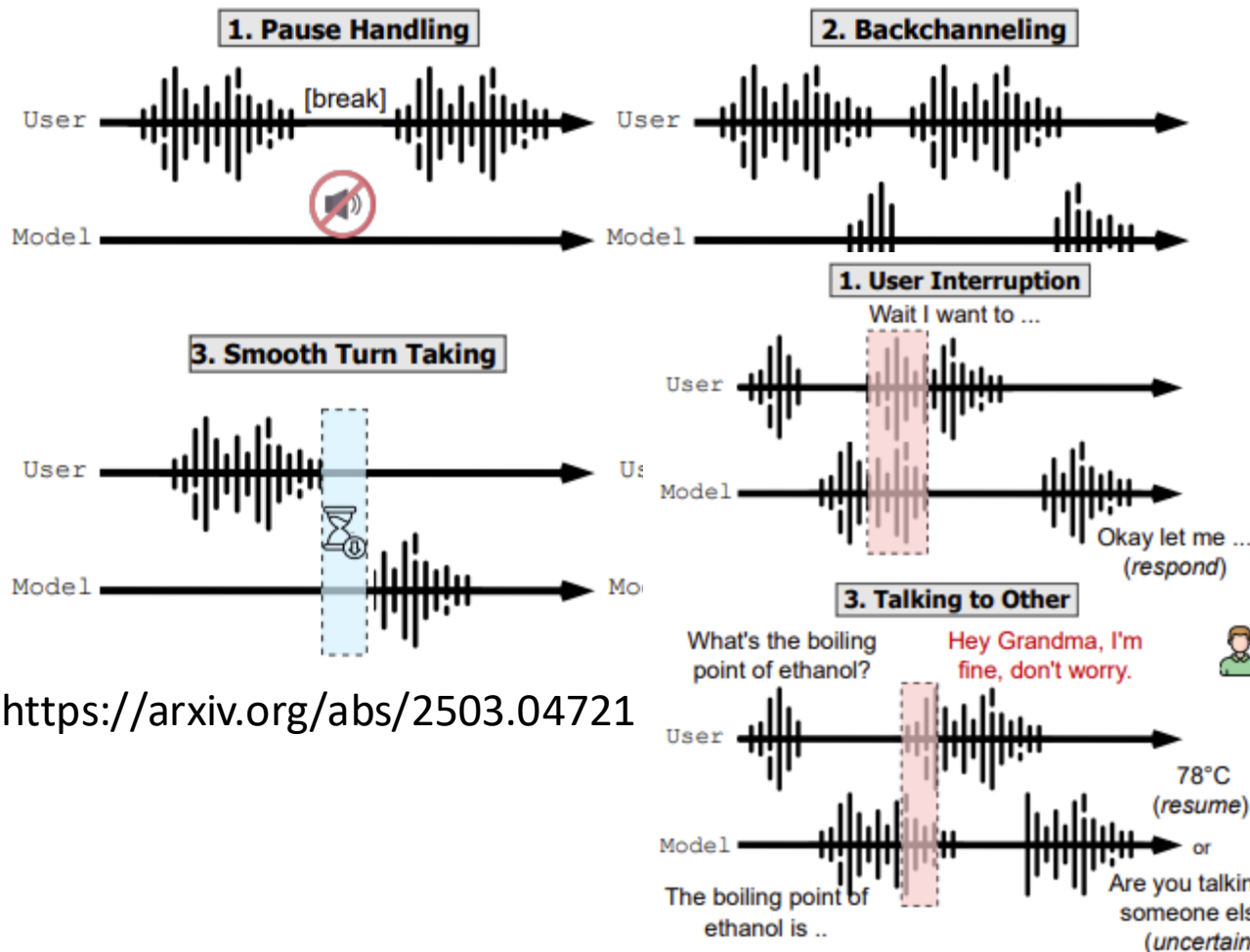


## Speech Conversation

*Full-duplex*







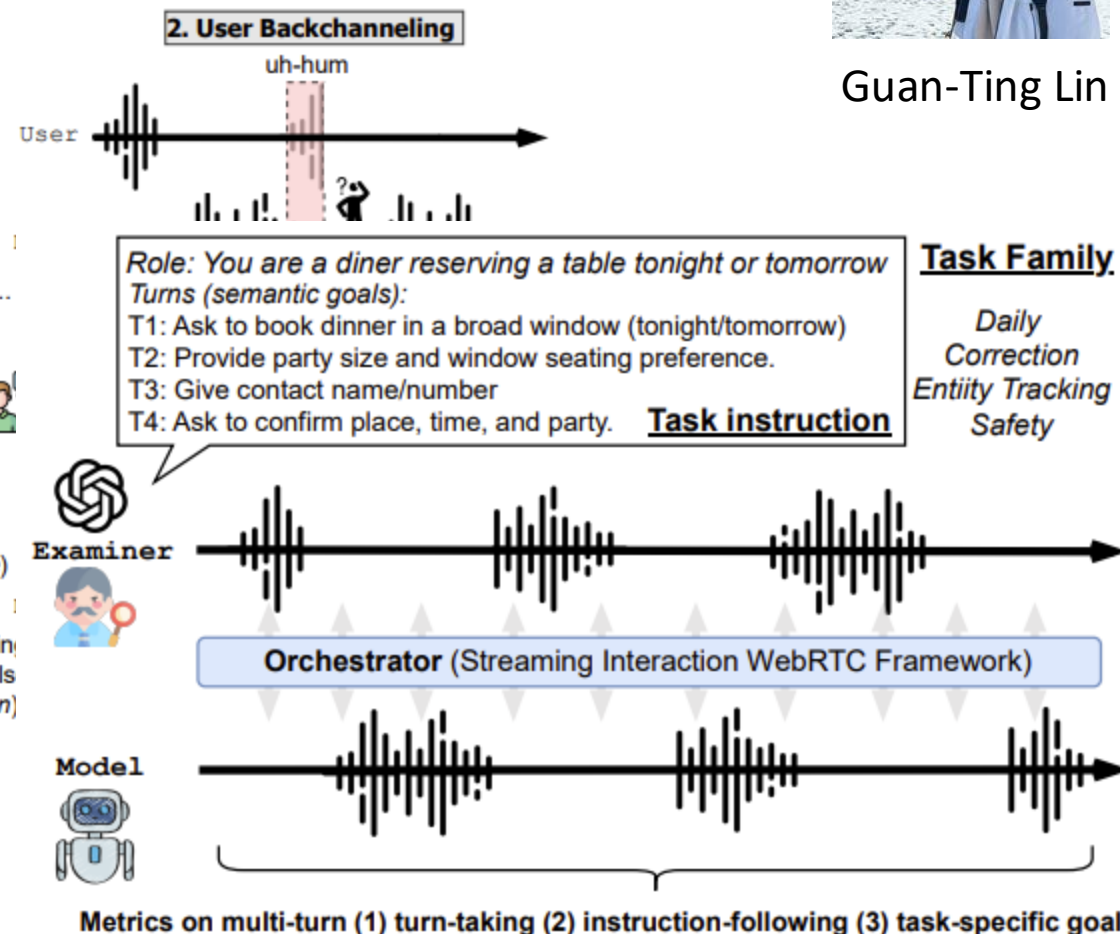
<https://arxiv.org/abs/2503.04721>

<https://arxiv.org/abs/2507.23159>

# Full-Duplex-Bench



Guan-Ting Lin



<https://arxiv.org/abs/2510.07838>

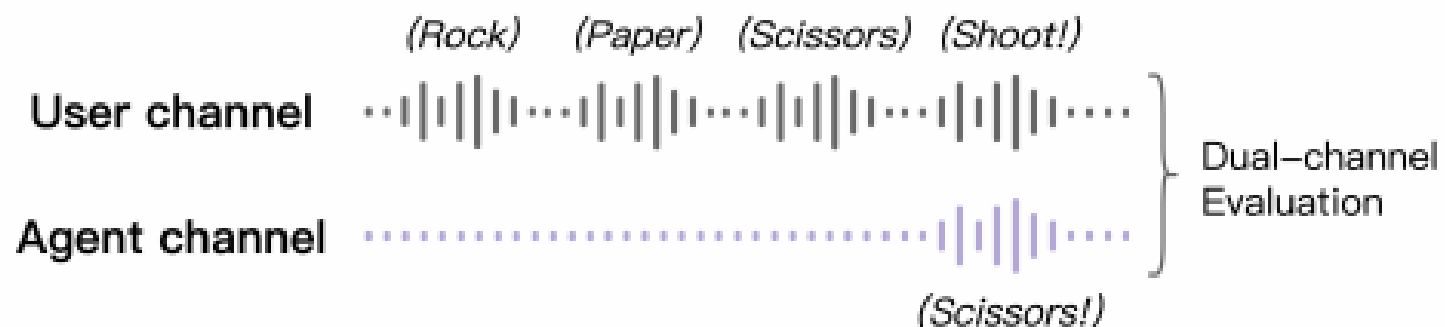
(With Shih-Yun Shan Kuan)

# Do speech LLMs know how long 10 seconds is?



Kai-Wei Chang

## Game-Time Benchmark



### Constraints



**Time**

### Examples

Remain silent for 10 seconds before giving your response.



**Tempo**

Count from one to ten with the tempo <bpm=60> dah-dah-dah.



**Simul.**

Say your choice on "shoot." Rock, Paper, Scissors... Shoot!

<https://arxiv.org/abs/2509.26388>

# On The Landscape of Spoken Language Models: A Comprehensive Survey

Siddhant Arora<sup>1\*</sup> Kai-Wei Chang<sup>2\*</sup> Chung-Ming Chien<sup>3\*</sup> Yifan Peng<sup>1\*</sup> Haibin Wu<sup>2\*#</sup>  
Yossi Adi<sup>4+</sup> Emmanuel Dupoux<sup>5+</sup> Hung-Yi Lee<sup>2+</sup> Karen Livescu<sup>3+</sup> Shinji Watanabe<sup>1+</sup>  
<sup>1</sup> Carnegie Mellon University, USA  
<sup>2</sup> National Taiwan University, Taiwan  
<sup>3</sup> Toyota Technological Institute at Chicago, USA  
<sup>4</sup> Hebrew University of Jerusalem, Israel  
<sup>5</sup> ENS - PSL, EHESS, CNRS, France

<https://arxiv.org/abs/2504.08528>

