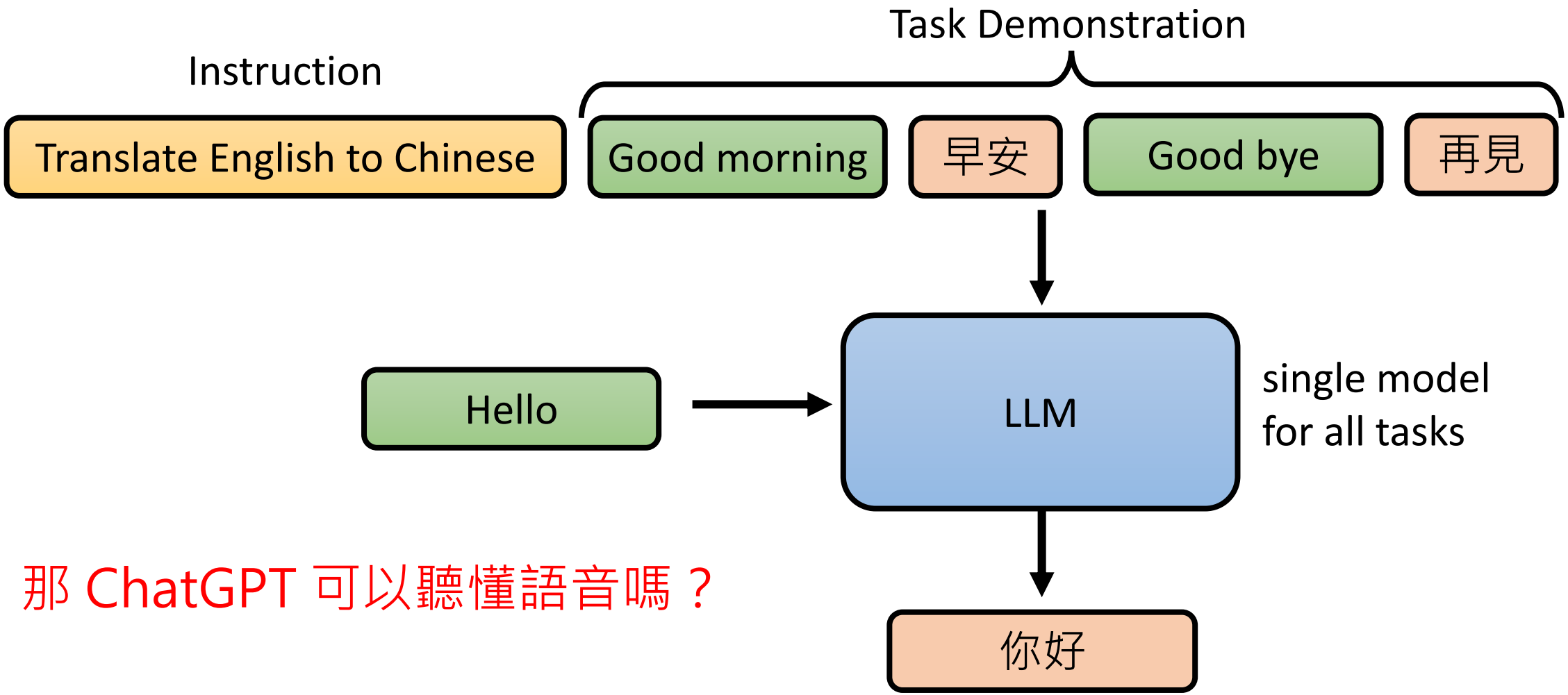
Towards a Speech Version of ChatGPT

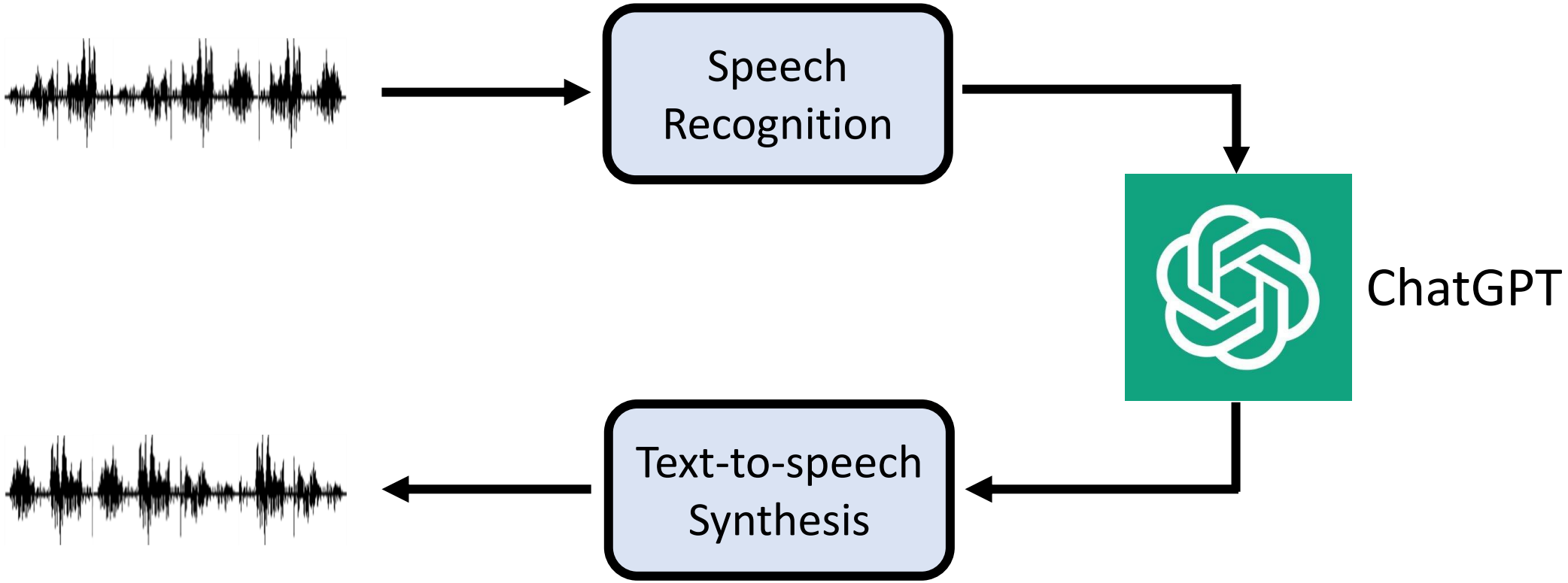
Hung-yi Lee

大家都已經見識到 ChatGPT 的能力



那 ChatGPT 可以聽懂語音嗎？

語音版 ChatGPT? 這不是已經有了嗎？

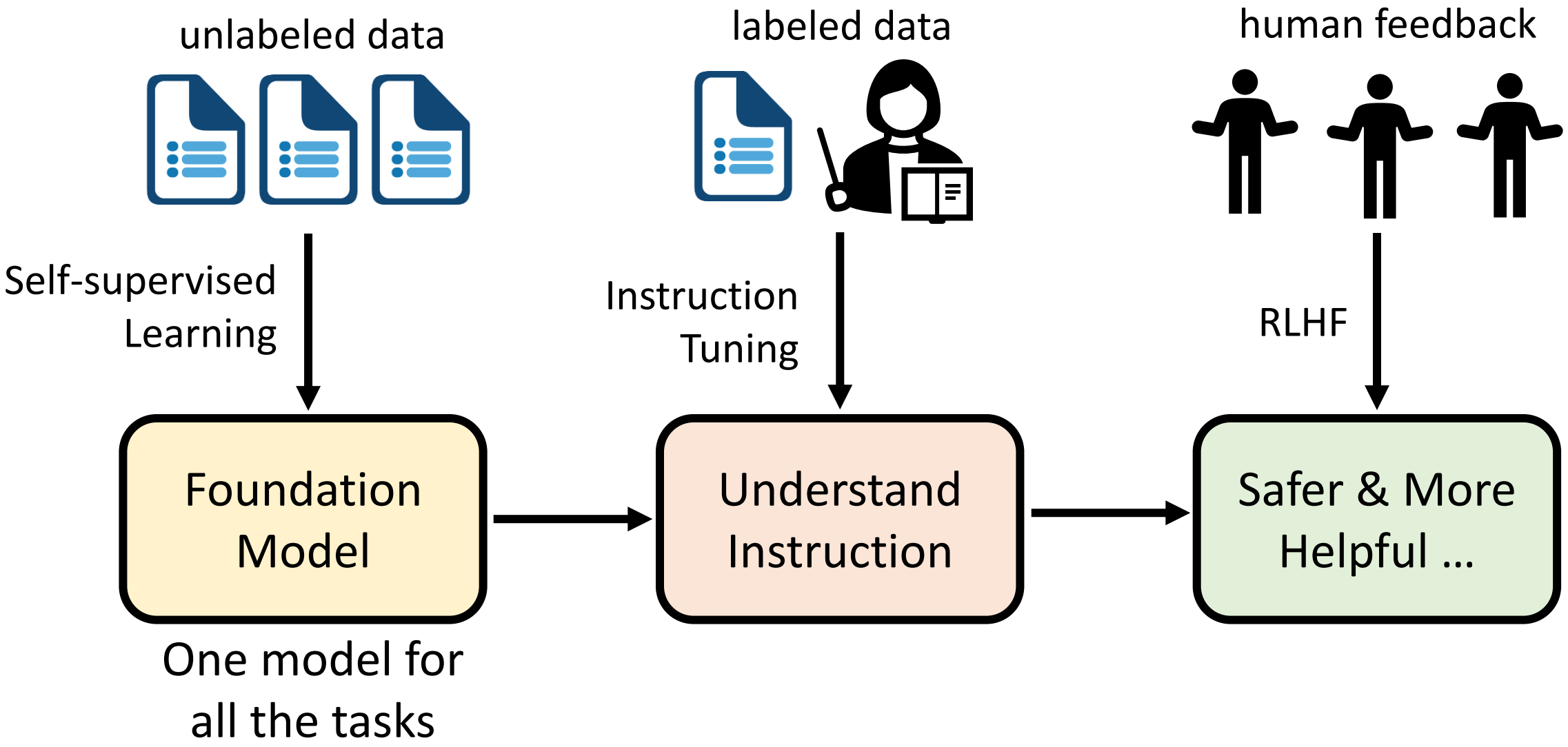


At the end of this talk, I will show you that this trivial solution is insufficient.

我們離語音版的 ChatGPT 其實還有距離

(獵人第八卷)

How did LLM get there?



How about Speech?

unlabeled data

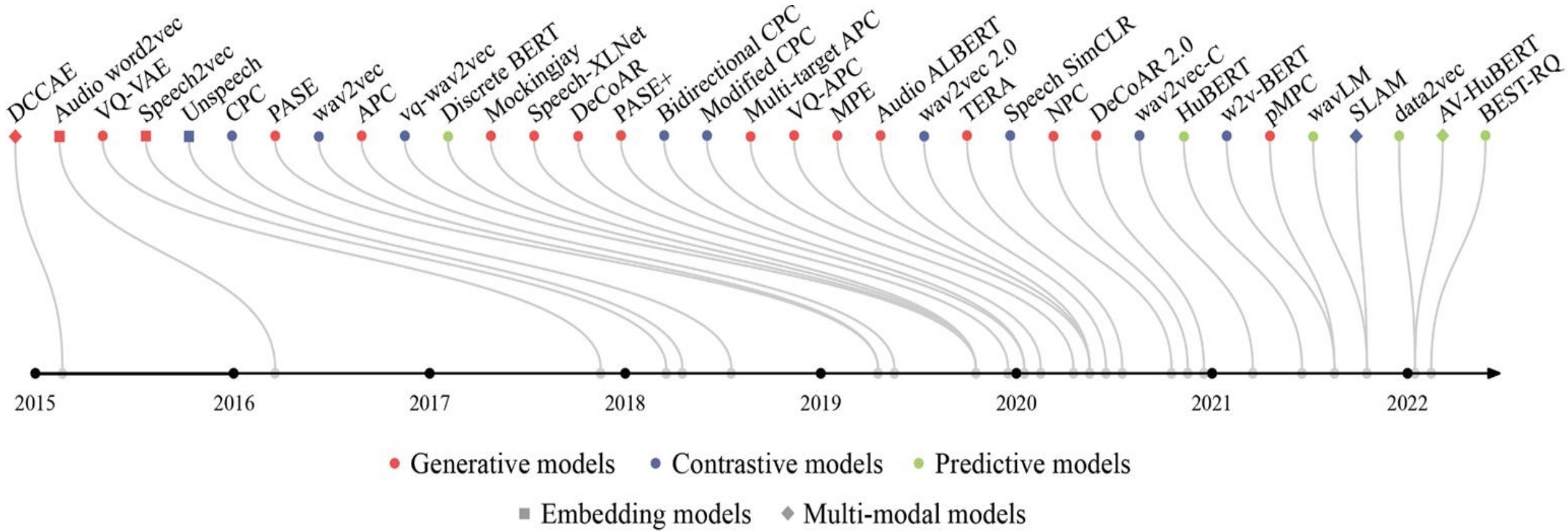
Self-supervised Learning

Speech Foundation Model

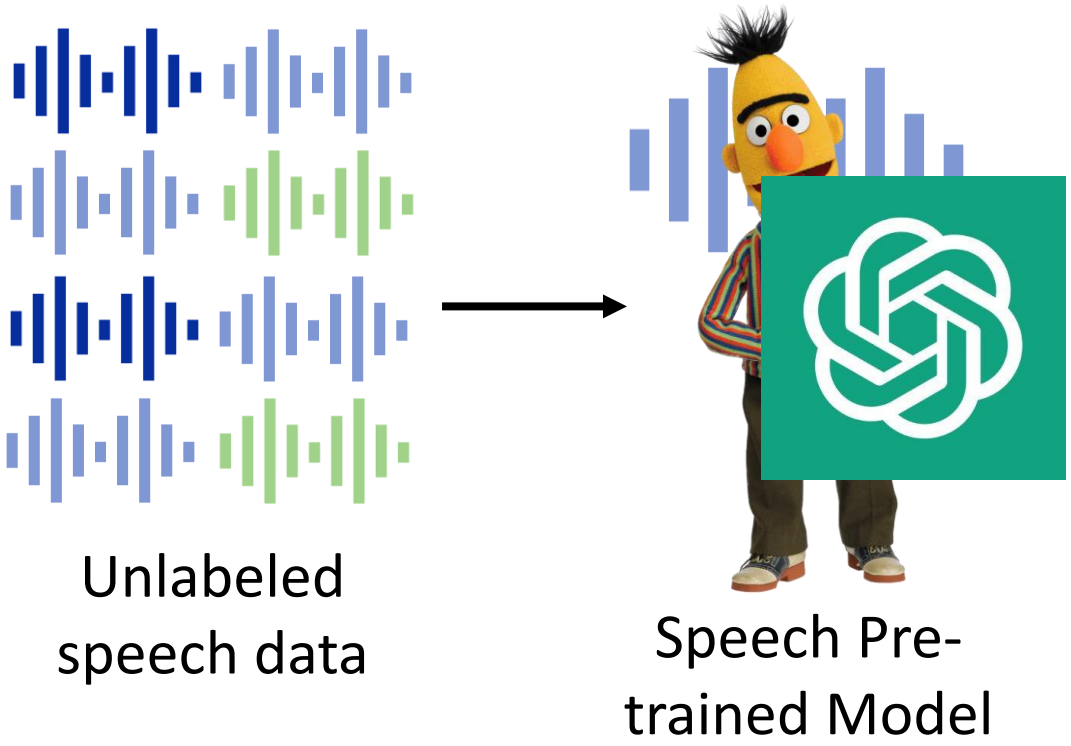
Self-Supervised Speech Representation Learning: A Review

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

https://arxiv.org/abs/2205.10643



Self-supervised Learning for Speech



Unlabeled speech data

Speech Pre-trained Model

- There are so many self-supervised models ……
- They have shown to achieve good performance on ASR.
- Are they specialist for ASR? Or are they universal?

I believe they are specialist.

My two cents **(before 2021)**

SUPERB

Speech processing Universal PERformance Benchmark

Shu-wen Yang

SUPERB: Speech processing Universal PERformance Benchmark

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

https://arxiv.org/abs/2105.01051

INTERSPEECH 2021

National Taiwan University 國立臺灣大學

Johns Hopkins University

Carnegie Mellon University

Massachusetts Institute of Technology

Meta

SUPERB

Speech processing Universal PERformance Benchmark

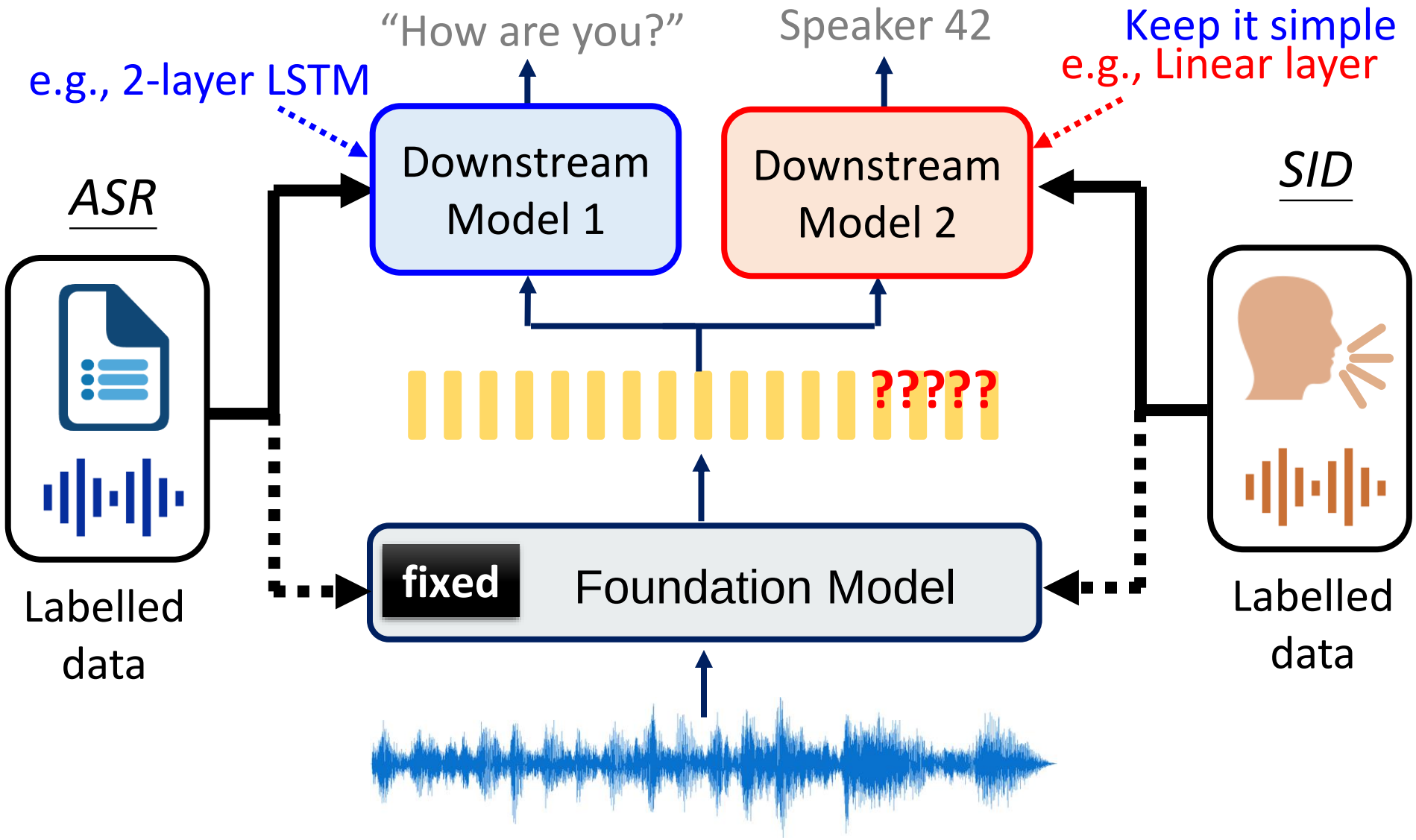
Evaluate a wide range of speech self-supervised models on many speech tasks

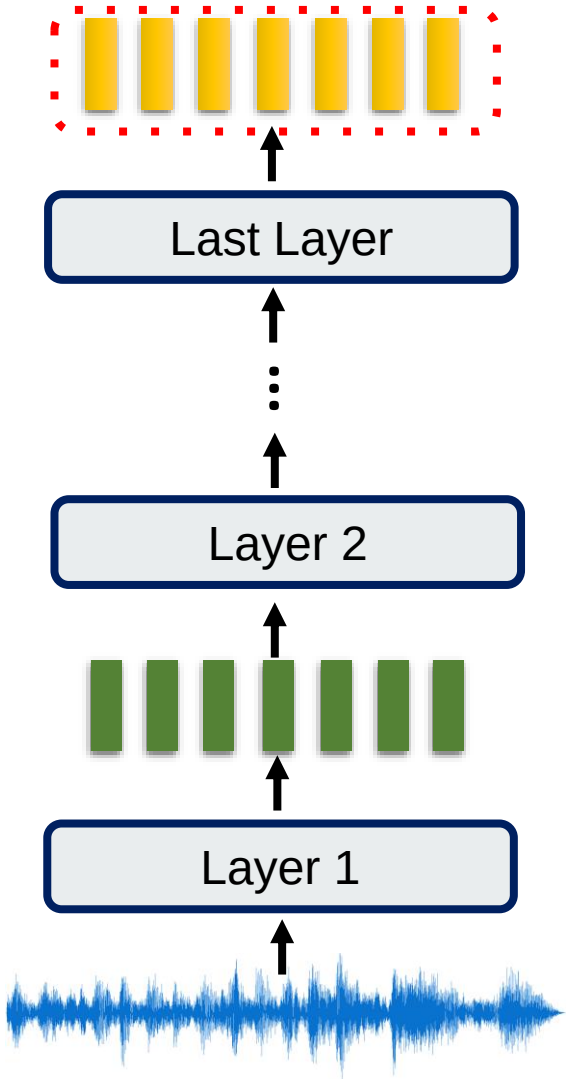
Phoneme Recognition (**PR**)	Speaker Identification (**SID**)	Emotion Recognition (**ER**)	Intent Classification (**IC**)	Voice Conversion (**VC**)
Keyword Spotting (**KS**)	Speaker Verification (**SV**)	Sentiment Analysis (**SA**)	Spoken Slot Filling (**SF**)	Speech Enhancement (**SE**)
Speech Recognition (**ASR**)	Speaker Diarization (**SD**)	Sarcasm Detection (**SarD**)	Speech Translation (**ST**)	Speaker Separation (**SS**)
Query-by-Example (**QbyE**)		Persuasiveness Prediction (**PP**)		

Content | Speaker | Prosody | Semantic | Synthesis

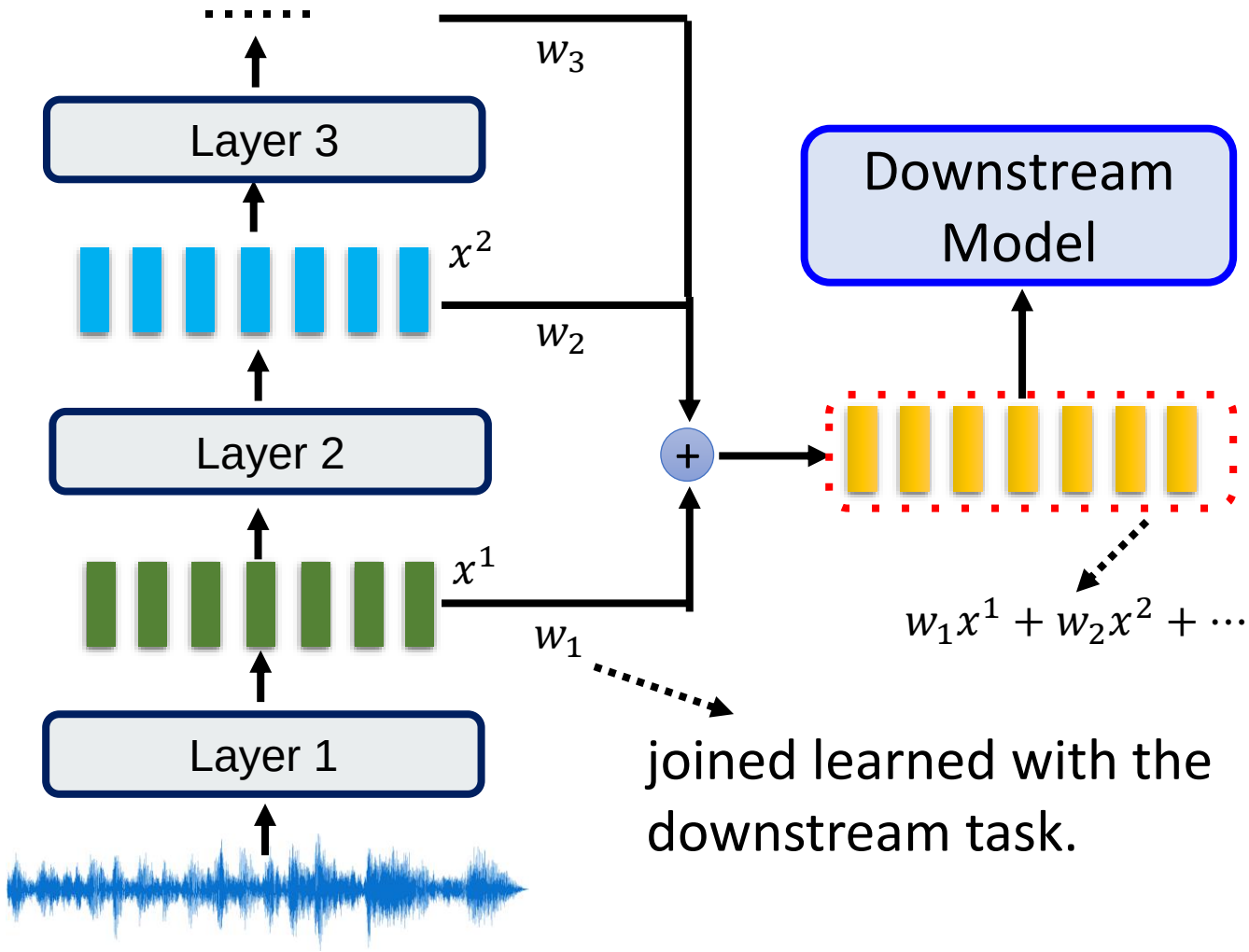
https://arxiv.org/abs/2105.01051
https://arxiv.org/abs/2203.06849
https://arxiv.org/abs/2210.08634
https://arxiv.org/abs/2210.07185
https://arxiv.org/abs/2110.06280

How to use Speech Foundation Model





Not always lead to decent performance.



joined learned with the downstream task.

"Weighted-sum" is Very effective!

Results of SUPERB

Emotion

	Content				Speaker			Semantic		
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.87	82.54	16.62	0.0072	37.99	11.61	8.68	29.82	62.14	57.86
APC	41.98	91.01	14.74	0.0310	60.42	8.56	10.53	74.69	70.46	59.33
VQ-APC	41.08	91.11	15.21	0.0251	60.15	8.72	10.45	74.48	68.53	59.66
NPC	43.81	88.96	13.91	0.0246	55.92	9.40	9.34	69.44	72.79	59.08
Mockingjay	70.19	83.67	15.48	6.60E-04	32.29	11.66	10.54	34.33	61.59	50.28
TERA	49.17	89.48	12.16	0.0013	57.57	15.89	9.96	58.42	67.50	56.27
DeCoAR 2.0	14.93	94.48	9.07	0.0406	74.42	7.16	6.59	90.80	83.28	62.47
modified CPC	42.54	91.88	13.53	0.0326	39.63	12.86	10.38	64.09	71.19	60.96
wav2vec	31.58	95.59	11.00	0.0485	56.56	7.99	9.90	84.92	76.37	59.79
vq-wav2vec	33.48	93.38	12.80	0.0410	38.80	10.38	9.93	85.68	77.68	58.24
wav2vec 2.0 base	5.74	96.23	4.79	0.0233	75.18	6.02	6.08	92.35	88.30	63.43
wav2vec 2.0 large	4.75	96.66	3.10	0.0489	86.14	5.65	5.62	95.28	87.11	65.64
HuBERT base	5.41	96.30	4.79	0.0736	81.42	5.11	5.88	98.34	88.53	64.92
HuBERT large	3.53	95.29	2.94	0.0353	90.33	5.98	5.75	98.76	89.81	67.62

https://arxiv.org/abs/2105.01051

Results of SUPERB

Emotion

	Content				Speaker			Semantic		
	PR	KS	ASR	QbE	SID	ASV	SD	IC	SF	ER
fbank	82.01	8.63	15.21	0.0058	8.50E-04	9.56	10.05	9.1	69.64	35.39
PASE+	58.87	82.54		0.0072	37.99		8.68	29.82		57.86
APC	41.98	91.01	14.74	0.0310	60.42	8.56		74.69	70.46	59.33
VQ-APC	41.08	91.11		0.0251	60.15	8.72		74.48		59.66
NPC	43.81	88.96	13.91	0.0246	55.92	9.40	9.34	69.44	72.79	59.08
Mockingjay	70.19	83.67			32.29			34.33		50.28
TERA	49.17	89.48	12.16		57.57		9.96	58.42		56.27
DeCoAR 2.0	14.93	94.48	9.07	0.0406	74.42	7.16	6.59	90.80	83.28	62.47
modified CPC	42.54	91.88	13.53	0.0326	39.63			64.09	71.19	60.96
wav2vec	31.58	95.59	11.00	0.0485	56.56	7.99	9.90	84.92	76.37	59.79
vq-wav2vec	33.48	93.38	12.80	0.0410	38.80		9.93	85.68	77.68	58.24
wav2vec 2.0 base	5.74	96.23	4.79	0.0233	75.18	6.02	6.08	92.35	88.30	63.43
wav2vec 2.0 large	4.75	96.66	3.10	0.0489	86.14	5.65	5.62	95.28	87.11	65.64
HuBERT base	5.41	96.30	4.79	0.0736	81.42	5.11	5.88	98.34	88.53	64.92
HuBERT large	3.53	95.29	2.94	0.0353	90.33	5.98				

They can be universal!

To learn more ……

Compression

Robust

Adapter/Prompt

Unsupervised ASR

Visual-enhanced

Prosody

Research Group @ JSALT 2022

https://youtu.be/_GbZRM18NxQ

Closing-day Presentation

More SUPERB families are coming .....

Multilingual SUPERB

Carnegie Mellon University

National Taiwan University 國立臺灣大學

Meta

REMBRAND

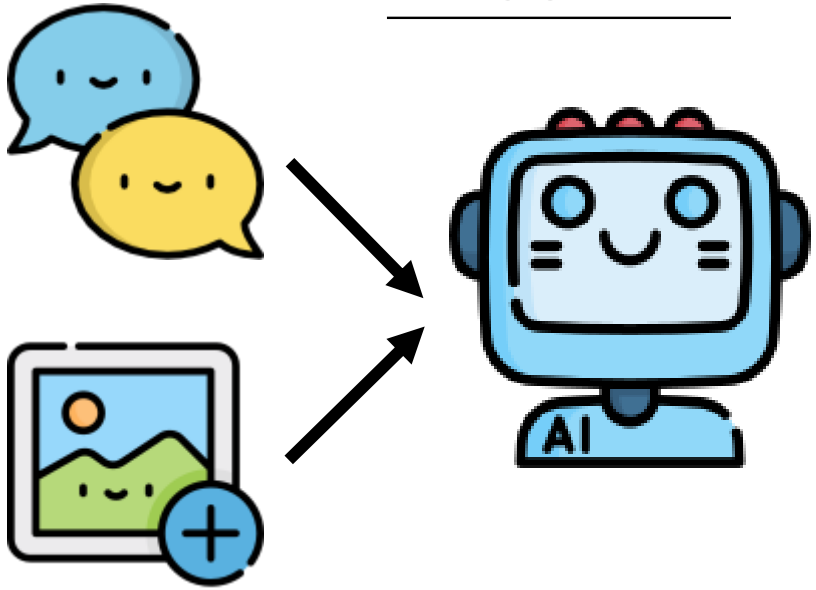
Challenge at ASRU 2023

https://multilingual.superbbenchmark.org/

- ***June 26, 2023***: New Language Track Submission Deadline
- ***July 07, 2023***: Model Submission Deadline

More SUPERB families are coming .....

AV-SUPERB



https://av.superbbenchmark.org/

National Taiwan University 國立臺灣大學

Carnegie Mellon University

Meta

REMBRAND

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

Yuan Tseng[1], *Layne Berry*[2*], *Yi-Ting Chen*[3*], *I-Hsiang Chiu*[1*], *Hsuan-Hao Lin*[1*], *Max Liu*[1*], *Puyuan Peng*[2*], *Yi-Jen Shih*[1*], *Hung-Yu Wang*[1*], *Haibin Wu*[1*], *Po-Yao Huang*[4], *Chun-Mao Lai*[1], *Shang-Wen Li*[4], *David Harwath*[2], *Yu Tsao*[3], *Shinji Watanabe*[5], *Abdelrahman Mohamed*[6], *Chi-Luen Feng*[1], *Hung-yi Lee*[1]

Yuan Tseng

https://arxiv.org/abs/2309.10787

Spoken Language Understanding Evaluation (SLUE) v2

https://arxiv.org/abs/2212.10525

SLUE Phase-2: A Benchmark Suite of Diverse Spoken Language Understanding Tasks

Suwon Shon[1], Siddhant Arora[2*], Chyi-Jiunn Lin[*3], Ankita Pasad[4*],
Felix Wu[1], Roshan Sharma[2], Wei-Lun Wu[3],
Hung-Yi Lee[3], Karen Livescu[4], Shinji Watanabe[2]

[1]ASAPP [2]Carnegie Mellon University [3]National Taiwan University
[4]Toyota Technological Institute at Chicago

New spoken QA corpus – SLUE-SQA-5: 47k spoken question-context pairs

Real speech!

Useful Toolkit!

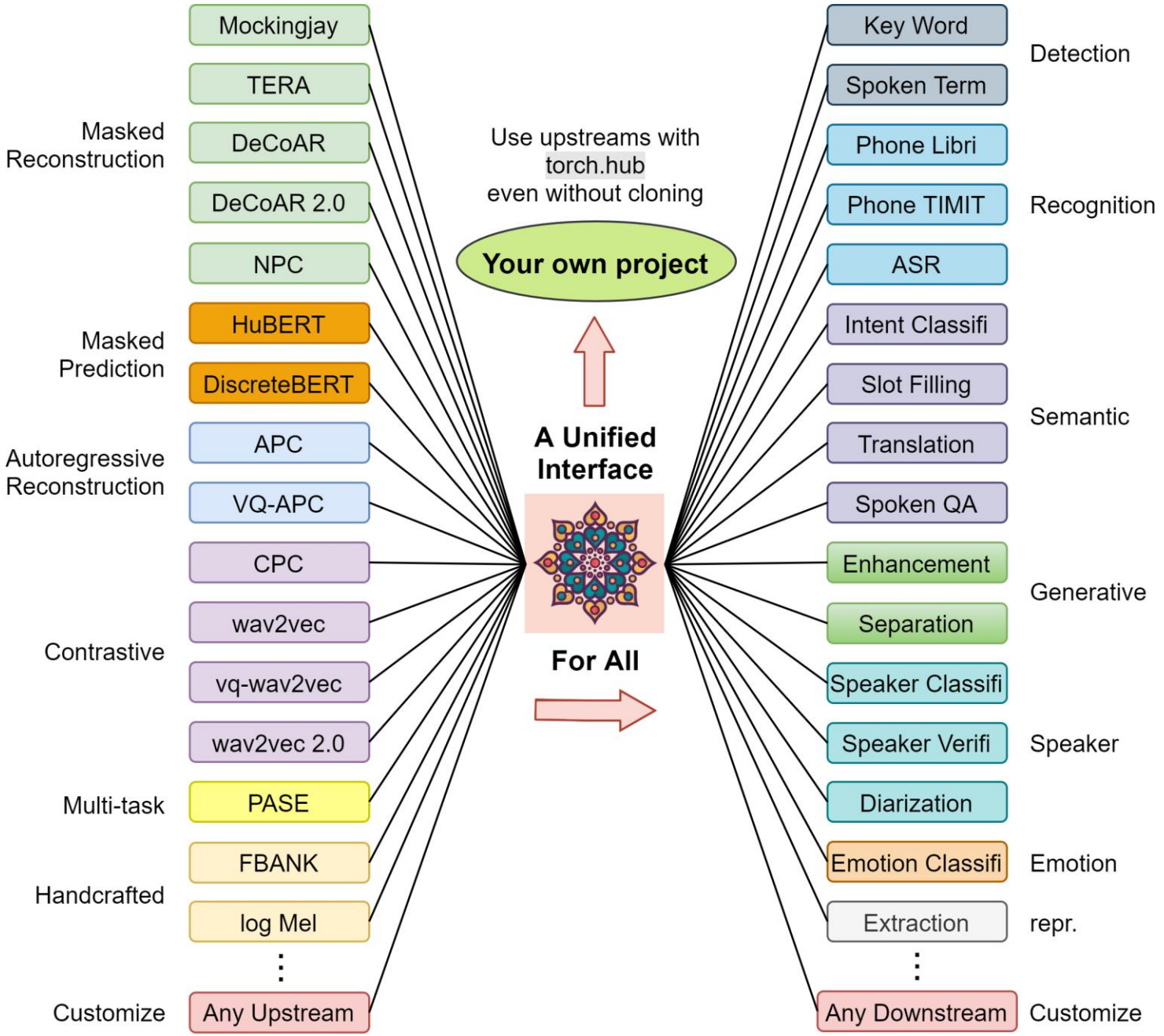
The S3PRL toolkit

https://github.com/s3prl/s3prl

Creator

Shu-wen Yang

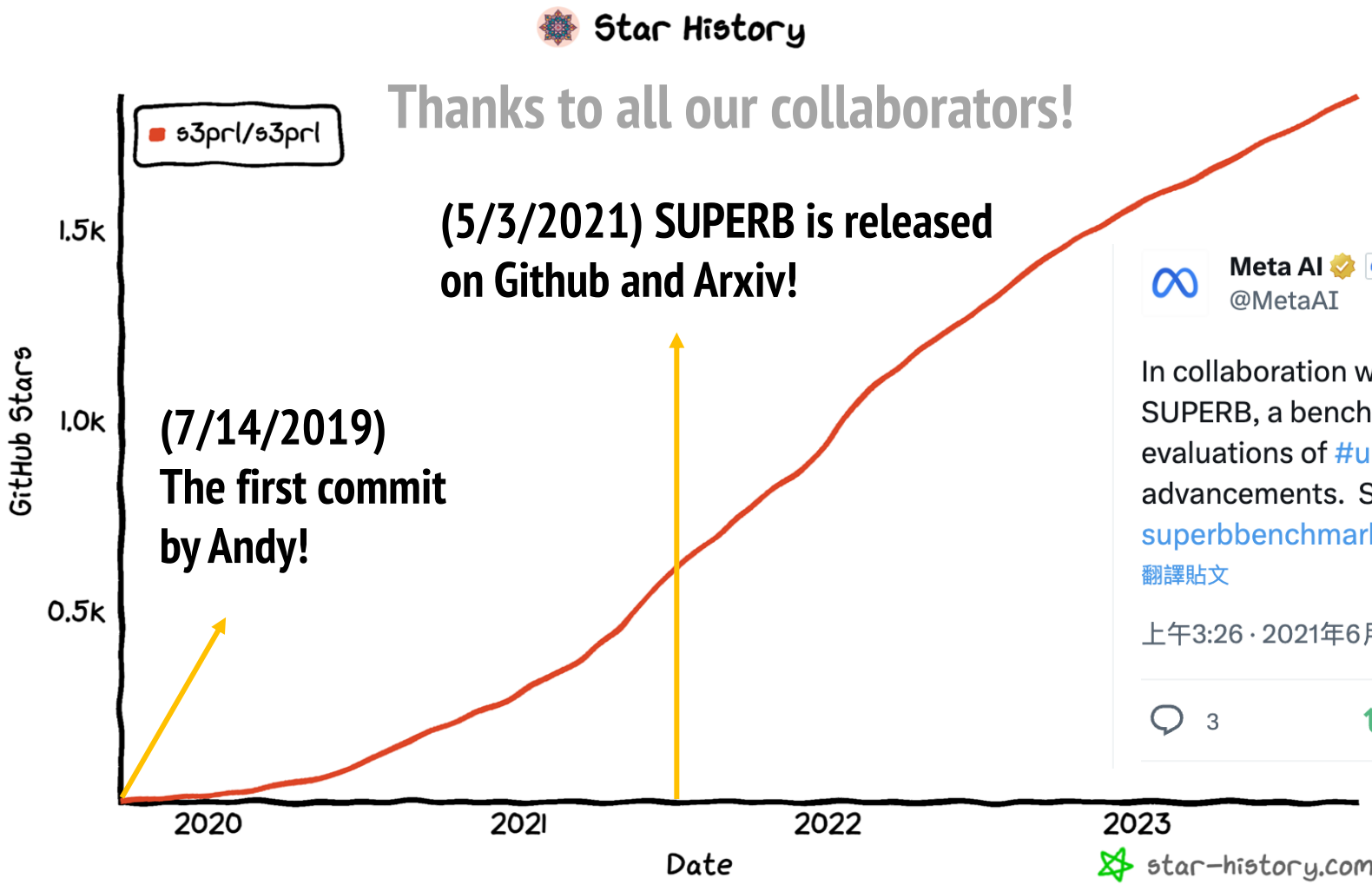
Andy T. Liu



投影片來源: 楊書文

Over 1.9k stars & used by over 48 repos

Shu-wen Yang



(6/19/2021) Meta helped promote SUPERB!

Meta AI
@MetaAI

In collaboration with @ntu_spml, @LTIatCMU, & @jhuclsp we introduce SUPERB, a benchmark using 10 speech processing tasks to standardize evaluations of #unsupervised models used in speech processing advancements. Submit & evaluate your models here: superbbenchmark.org

翻譯貼文

上午3:26 · 2021年6月19日

3 60 169 18

data from https://star-history.com/

How about Speech?

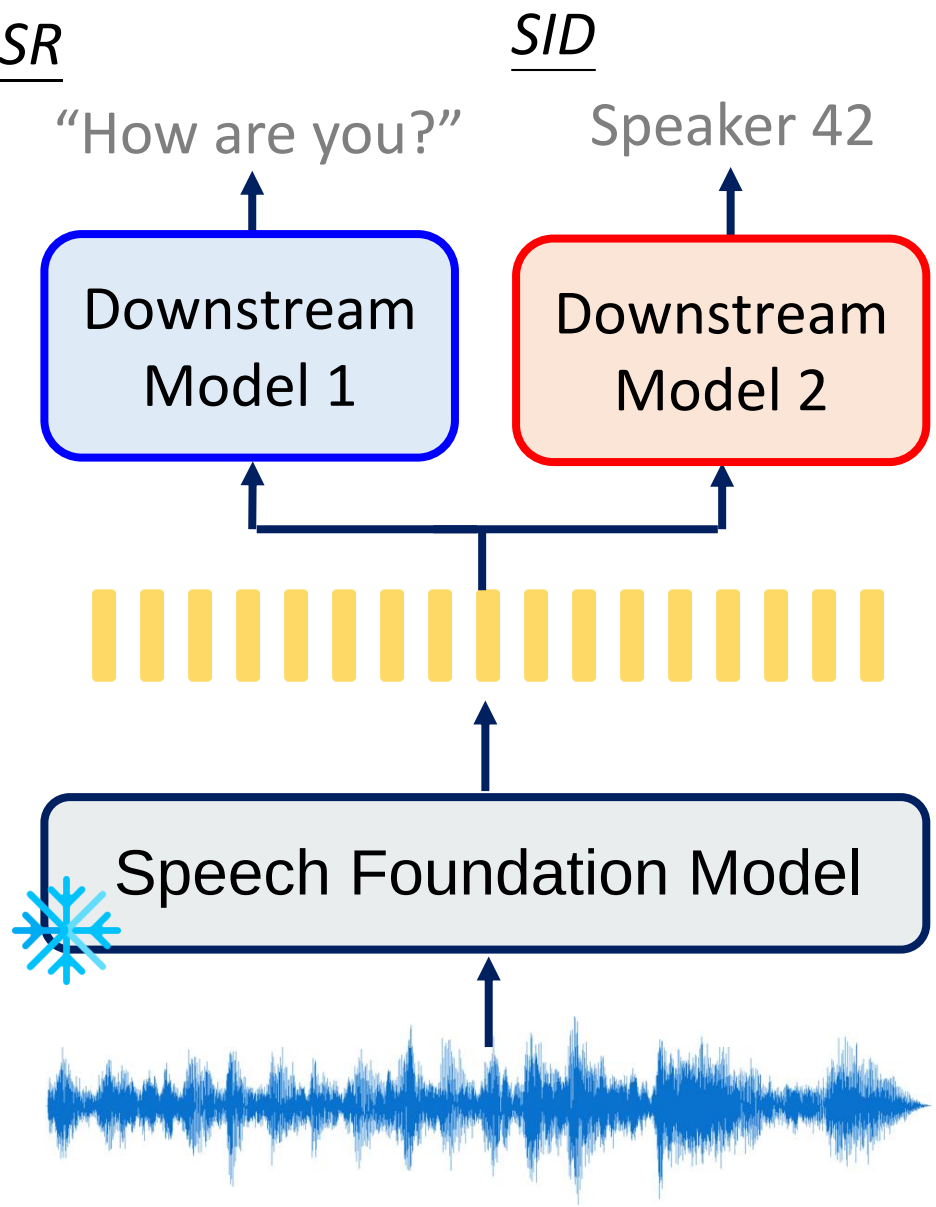
unlabeled data

Self-supervised Learning

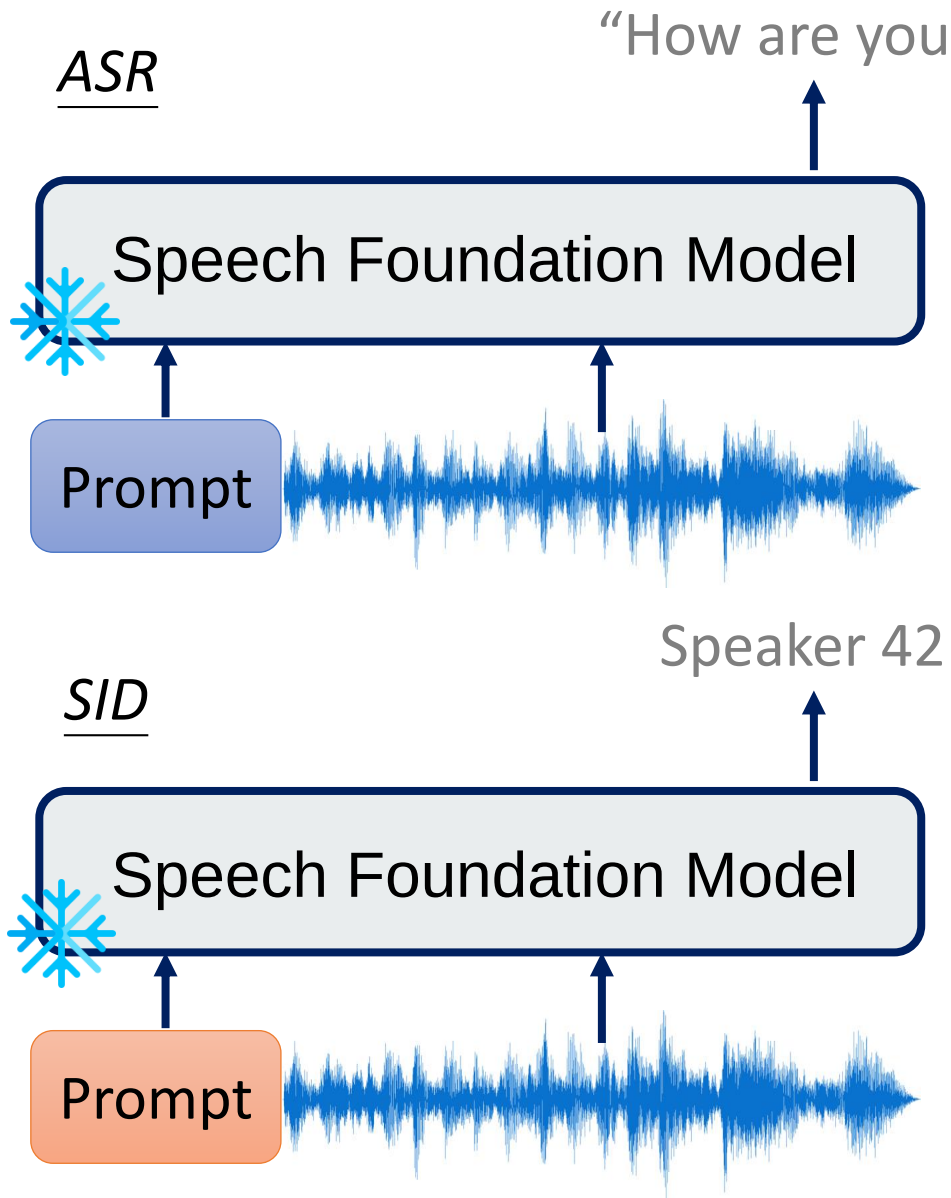
Speech Foundation Model

One model for all the tasks ?

SUPERB



Prompting



How about Speech?

unlabeled data

Self-supervised Learning

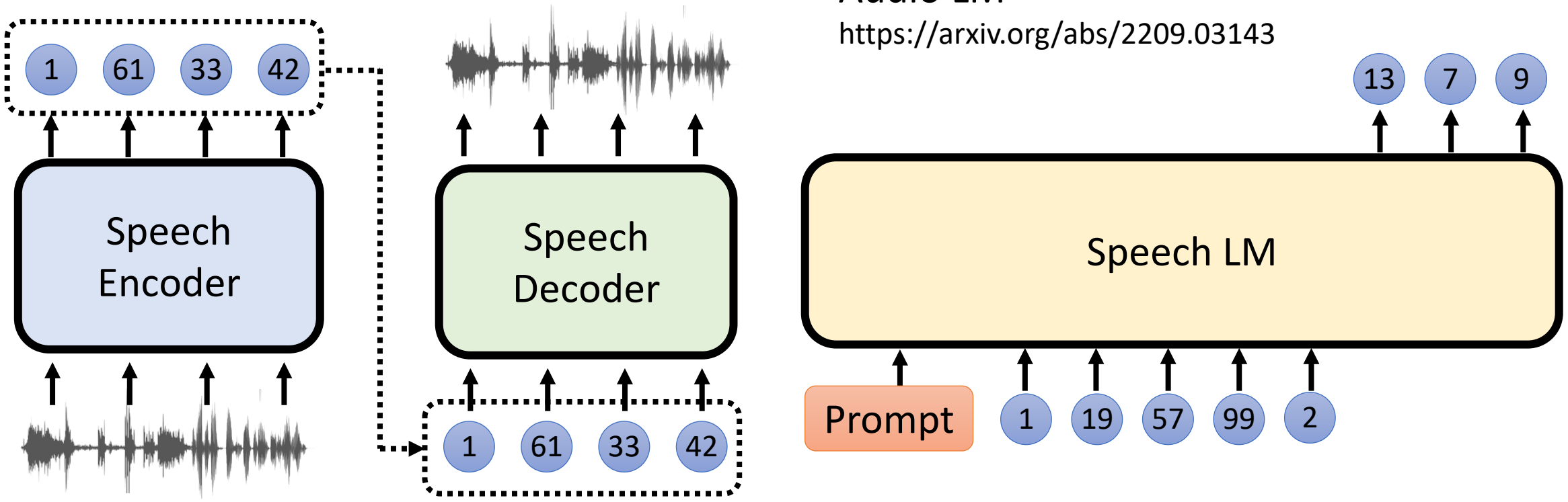
Speech Foundation Model

One model for all the tasks ?

Prompting Speech LM

Speech LM

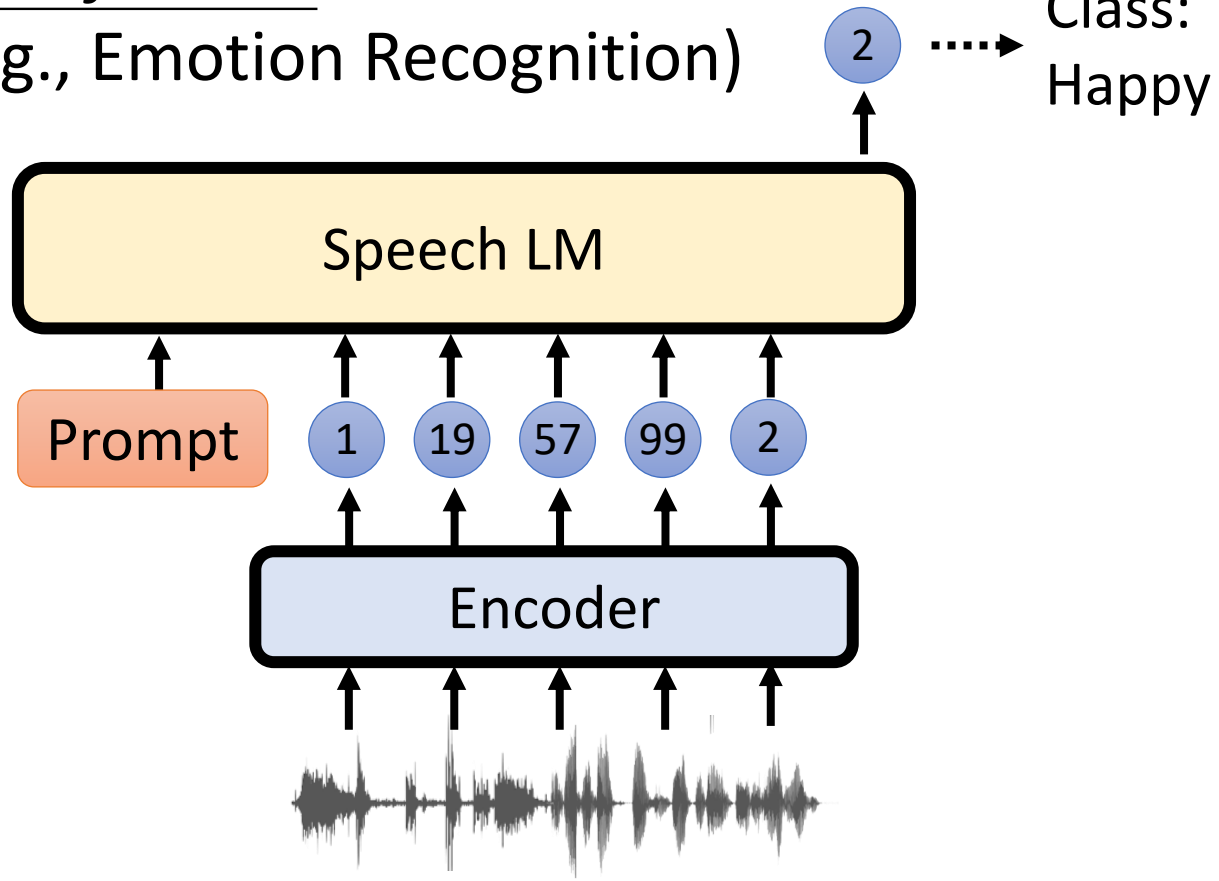
GSLM
https://arxiv.org/abs/2102.01192
Audio LM
https://arxiv.org/abs/2209.03143



Does a speech LM have potential to address for a wide range of speech tasks?

Classification

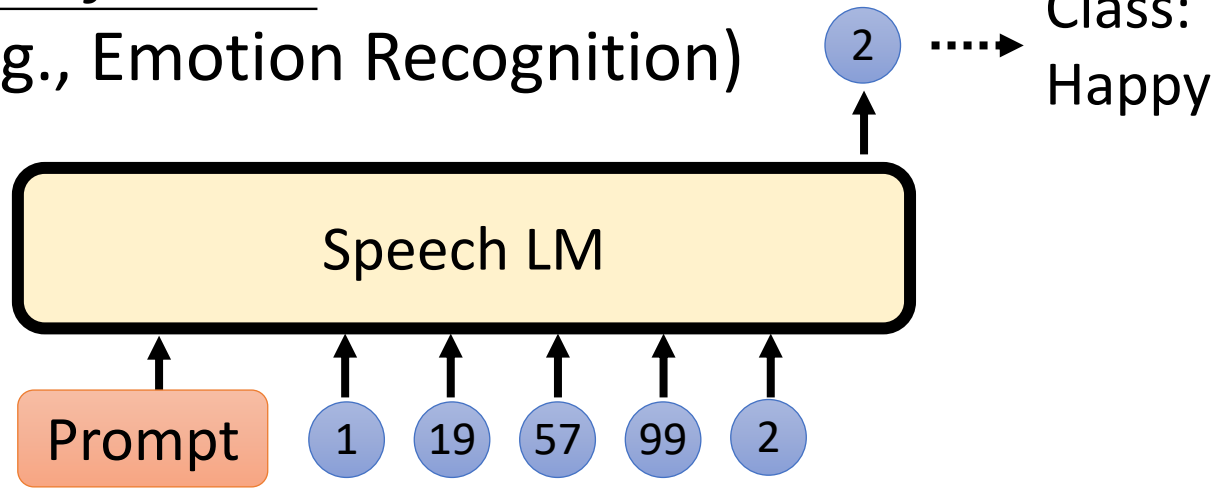
(e.g., Emotion Recognition)



Unit ID	Class
1	Angry
2	Happy
3	Sad
…	

Classification

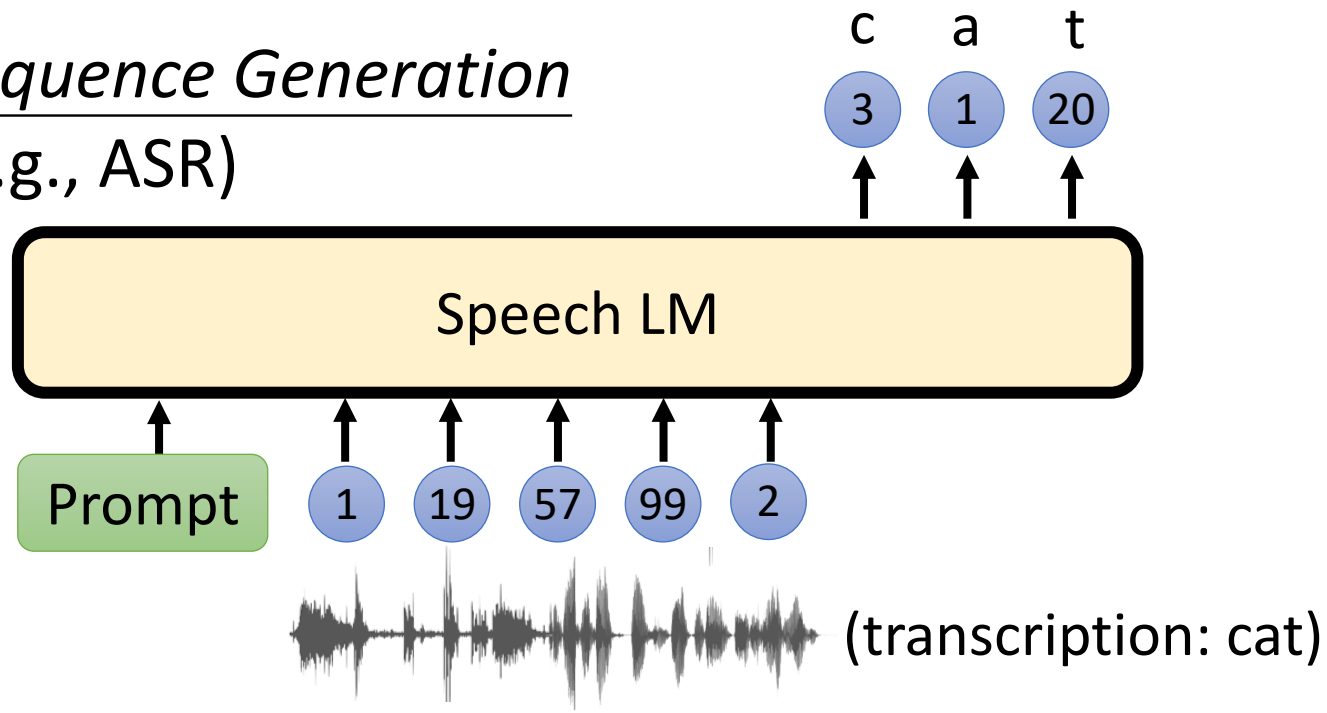
(e.g., Emotion Recognition)



Unit ID	Class
1	Angry
2	Happy
3	Sad
…	

Sequence Generation

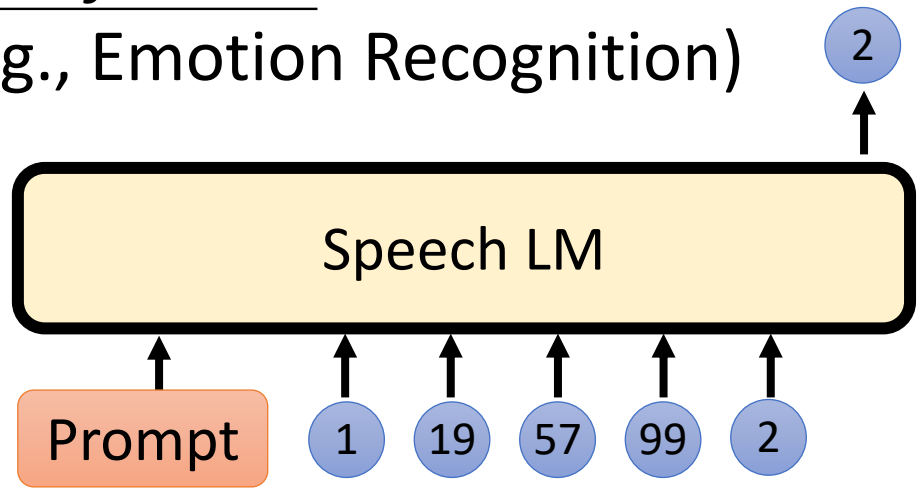
(e.g., ASR)



Unit ID	Class
1	a
2	b
3	c
…	

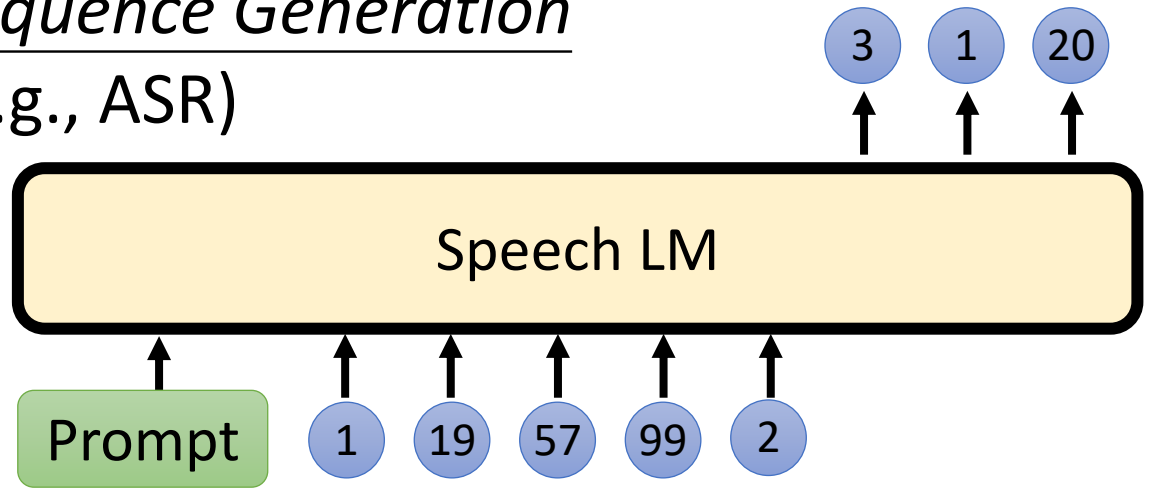
Classification

(e.g., Emotion Recognition)



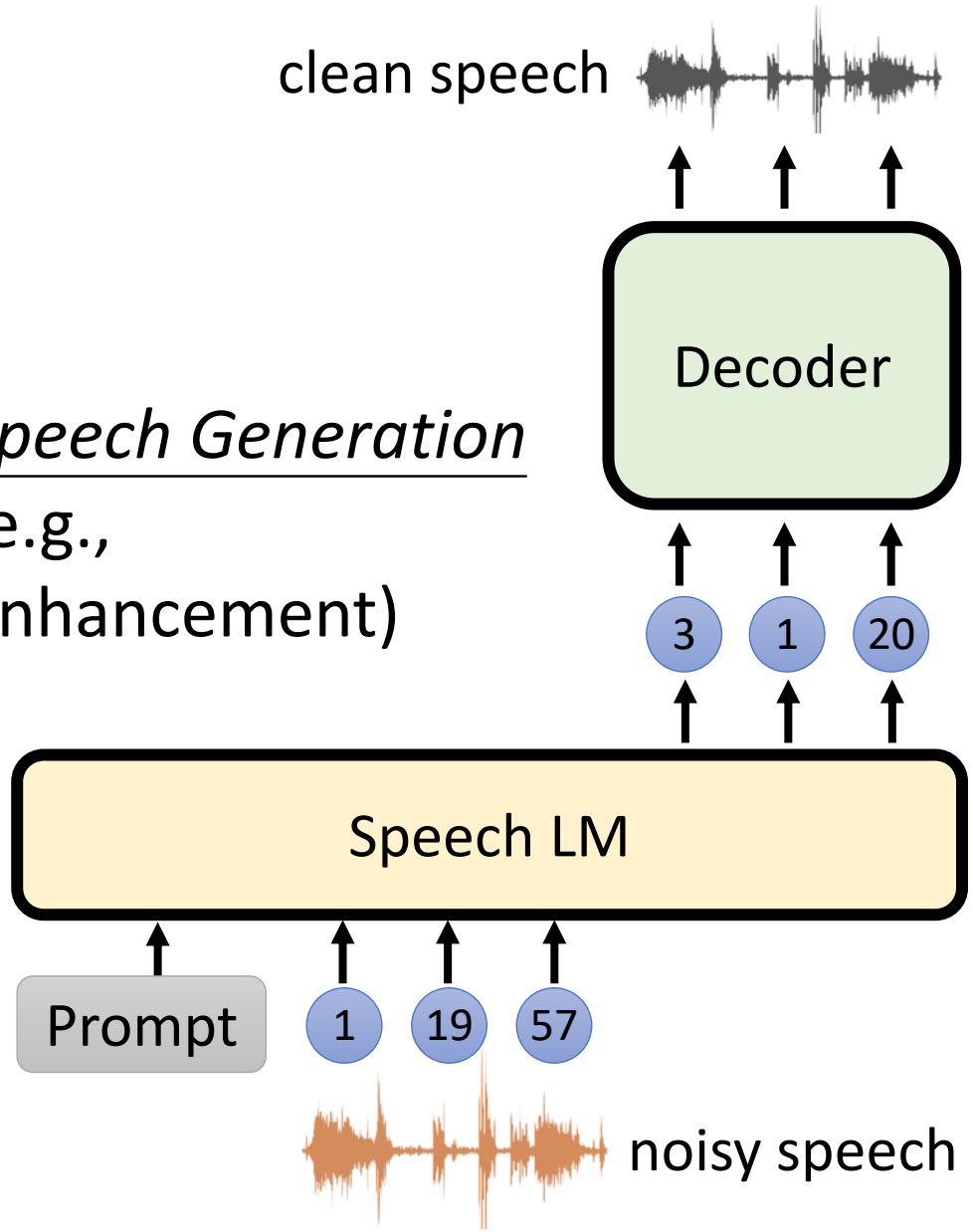
Sequence Generation

(e.g., ASR)



Speech Generation

(e.g., Enhancement)



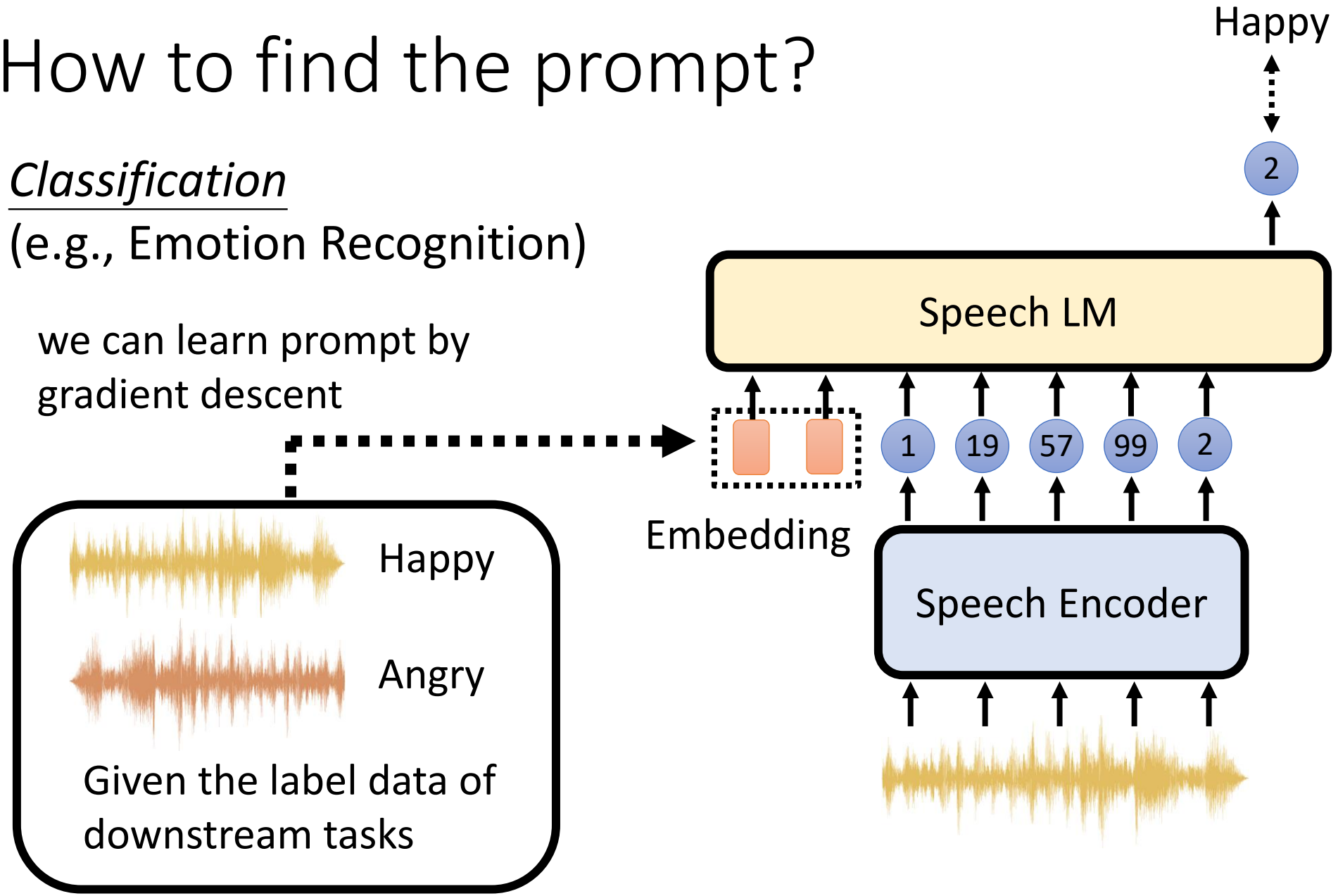
How to find the prompt?

Classification

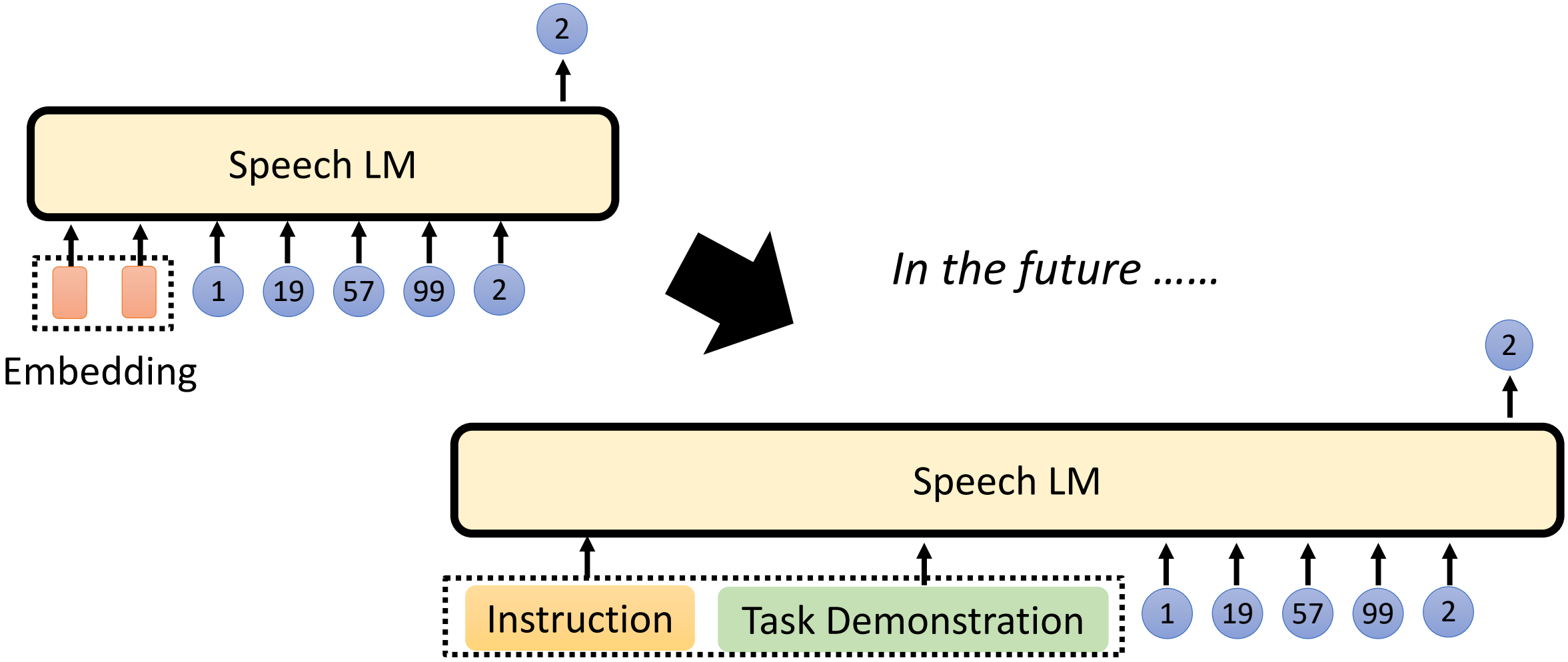
(e.g., Emotion Recognition)

we can learn prompt by gradient descent

Unit ID	Class
1	Angry
2	Happy
3	Sad
…	



How far are we from Universal Speech Model?



SPEECHPROMPT V2: PROMPT TUNING FOR SPEECH CLASSIFICATION TASKS

Kai-Wei Chang[1⋆] *Yu-Kai Wang*[2⋆] *Hua Shen*[3] *Iu-thing Kang*[4]
Wei-Cheng Tseng[1] *Shang-Wen Li*[5] *Hung-yi Lee*[1]

https://arxiv.org/abs/2303.00733

Kai-Wei Chang

Yu-Kai Wang

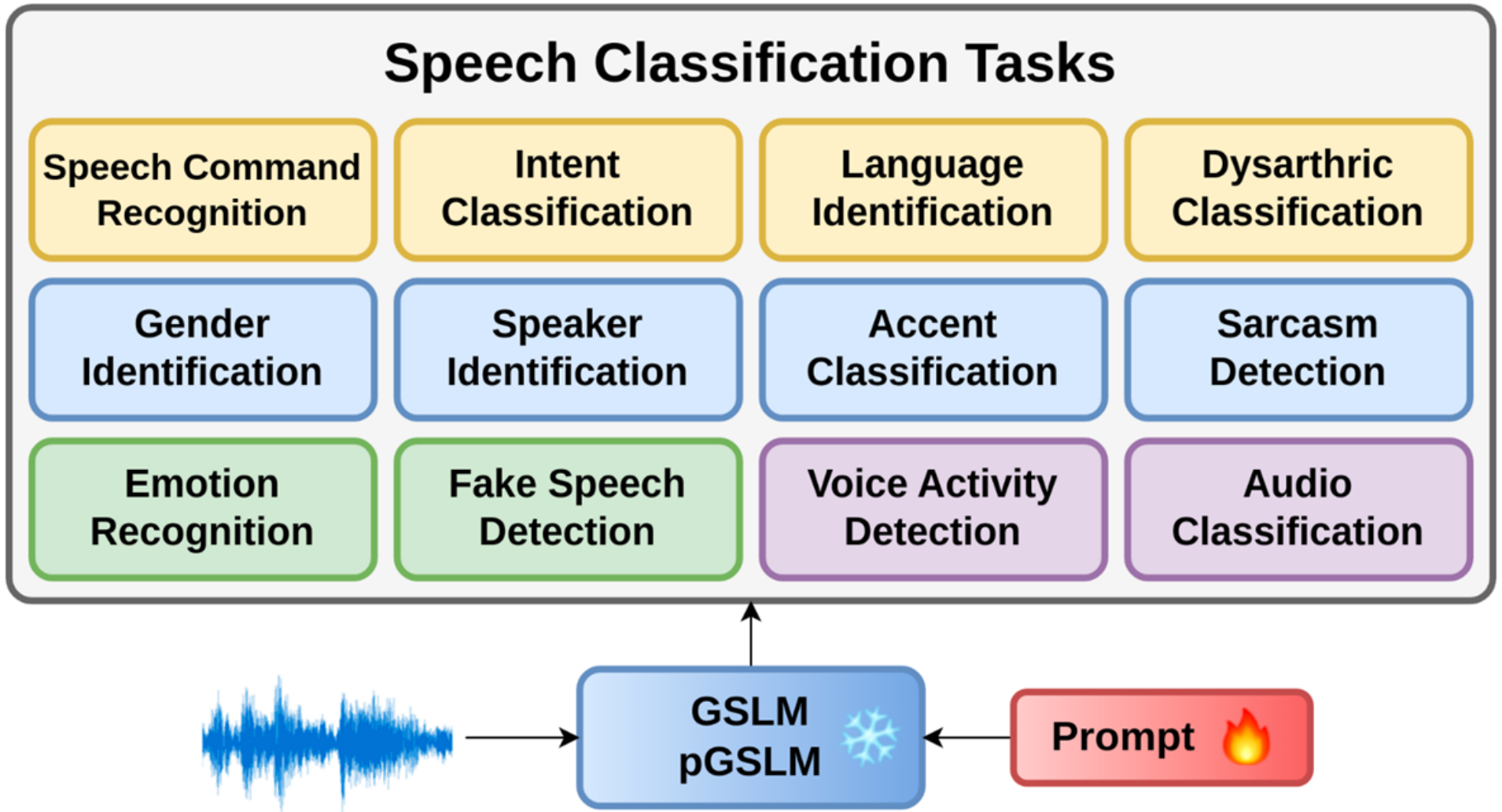
Hua Shen

Iu-thing Kang

Wei-Cheng Tseng

Shang-Wen Li

Hung-yi Lee


Speech Classification Tasks
Speech Command Recognition
Intent Classification
Language Identification
Dysarthric Classification
Gender Identification
Speaker Identification
Accent Classification
Sarcasm Detection
Emotion Recognition
Fake Speech Detection
Voice Activity Detection
Audio Classification
GSLM
pGSLM
Prompt

+:Trainable mapping

Task	Metric	Dataset	Language	#Class	SOTA (Topline)	GSLM	GSLM+	pGSLM	pGSLM+
SCR	ACC (↑)	Google SC v1	En	12	98.6 [10]	94.5	94.6	94.3	**94.7** (-3.9)
		Grabo SC	Du	36	98.9 [11]	92.4	**92.7** (-6.2)	17.5	19.6
		Lithuanian SC	Lt	15	91.8 [9]	93.2	**95.5** (+3.7)	90.9	79.5
		Arabic SC	Ar	16	98.9 [9]	99.7	**100.0** (+1.1)	85.6	92.6
IC	ACC (↑)	Fluent SC	En	24	99.7 [12]	97.2	97.3	98.1	**98.2** (-1.5)
LID	ACC (↑)	Voxforge	En, Es, Fr De, Ru, It	6	99.8 [13]	90.9	**94.2** (-5.6)	81.8	80.4
FSD	EER (↓)	ASVspoof	En	2	2.5 [13]	18.5	13.5	**13.1** (+10.6)	18.3
ER	ACC (↑)	IEMOCAP	En	4	79.2 [13]	42.1	44.3	49.9	**50.2** (-29)
AcC	ACC (↑)	AccentDB	En	9	99.5 [14]	78.9	83.4	86.5	**87.1** (-12.4)
SD	F1 (↑)	MUStARD	En	2	64.6 [15]	55.0	77.8	74.4	**78.7** (+13.1)
		MUStARD++	En	2	65.2 [16]	74.0	**75.2** (+10)	52.7	58.2
GID	F1 (↑)	VoxCeleb1	En	2	98.3 [17]	86.2	87.3	**91.6** (-6.7)	86.2
VAD	ACC (↑)	Google SC v2 & Freesound	En	2	98.8 [18]	96.6	96.9	**98.3** (-0.5)	98.1
AuC	ACC (↑)	ESC-50	✖	50	97.0 [19]	9.0	**37.5** (-59.5)	20.3	27.0

We can get reasonable performance on most classification tasks with prompting.

Experimental Results – Sequence Generation

Prompt length = 180

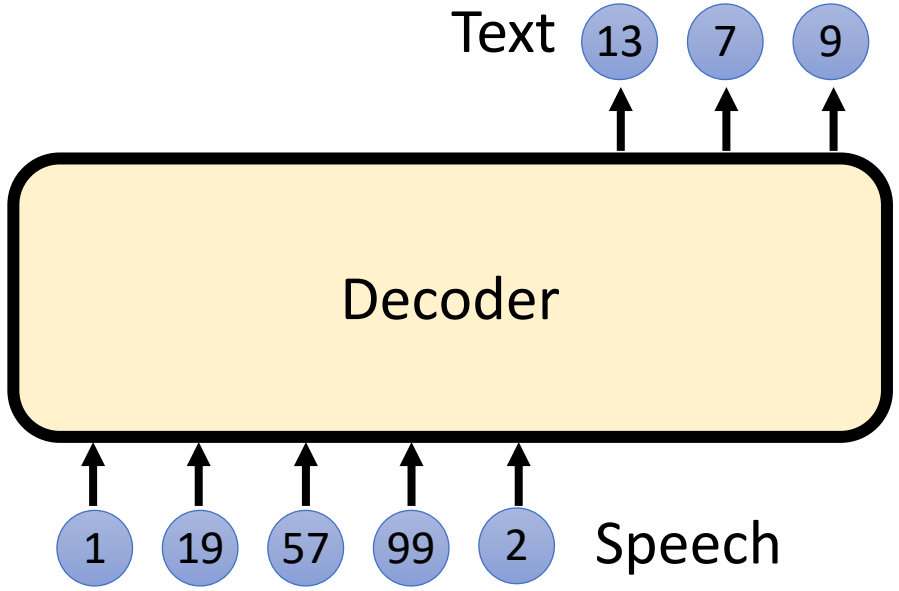
Slot Filling

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

Prompting speech LM is worse than other approaches .....

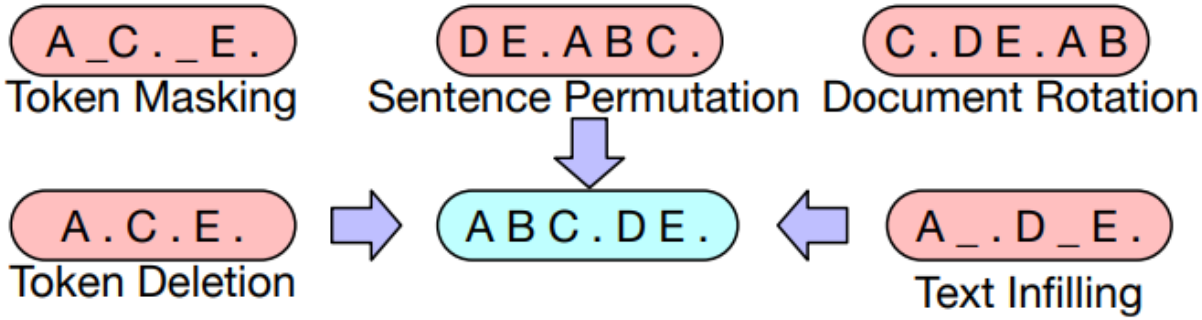
Other Speech LM

Poor performance for ASR

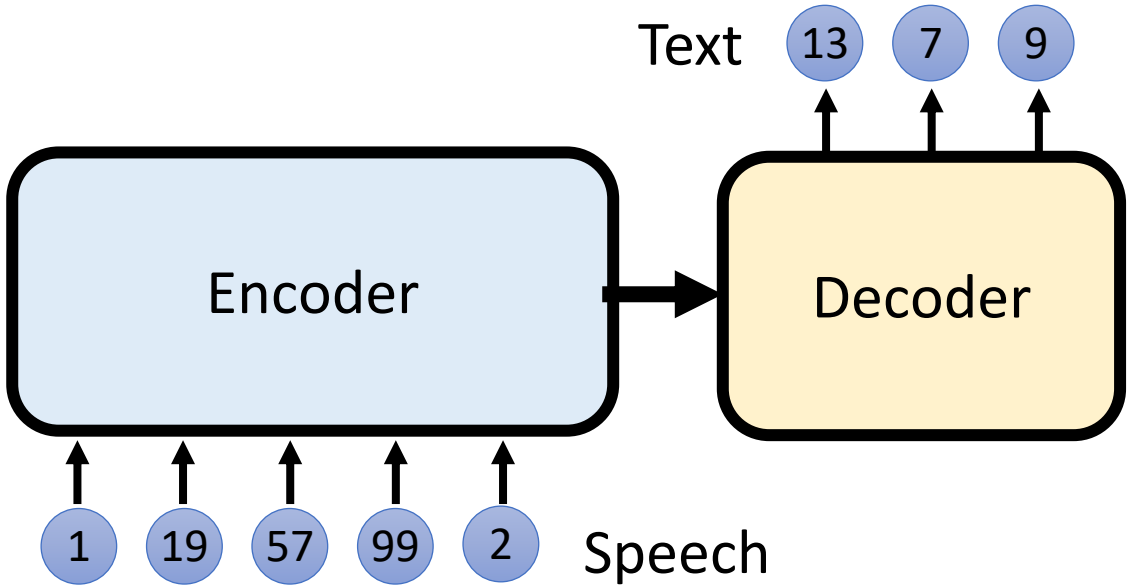


GSLM

https://arxiv.org/abs/2102.01192



https://arxiv.org/abs/1910.13461



Unit BART

https://arxiv.org/abs/2204.02967

Experimental Results – Sequence Generation

Slot Filling

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

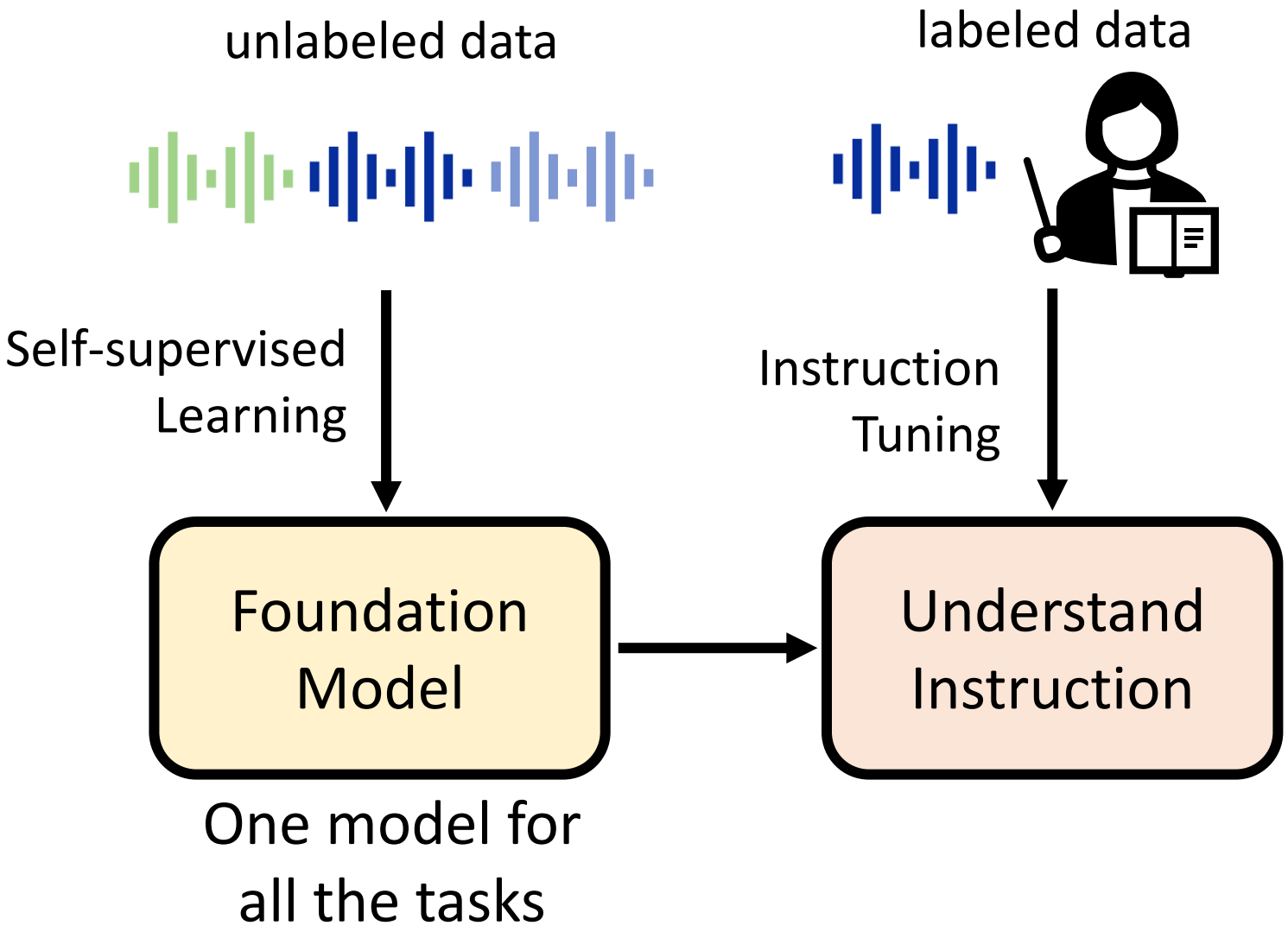
Prompting unit BART obtains 9.8% CER.

(unpublished results)

To learn more ……

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

How about Speech?



Speech Version of ChatGPT ……

Identify the emotion

Speech Version of ChatGPT

happy

Describe speaker characteristics

Speech Version of ChatGPT

deep male voice

Please denoise the utterance

Speech Version of ChatGPT

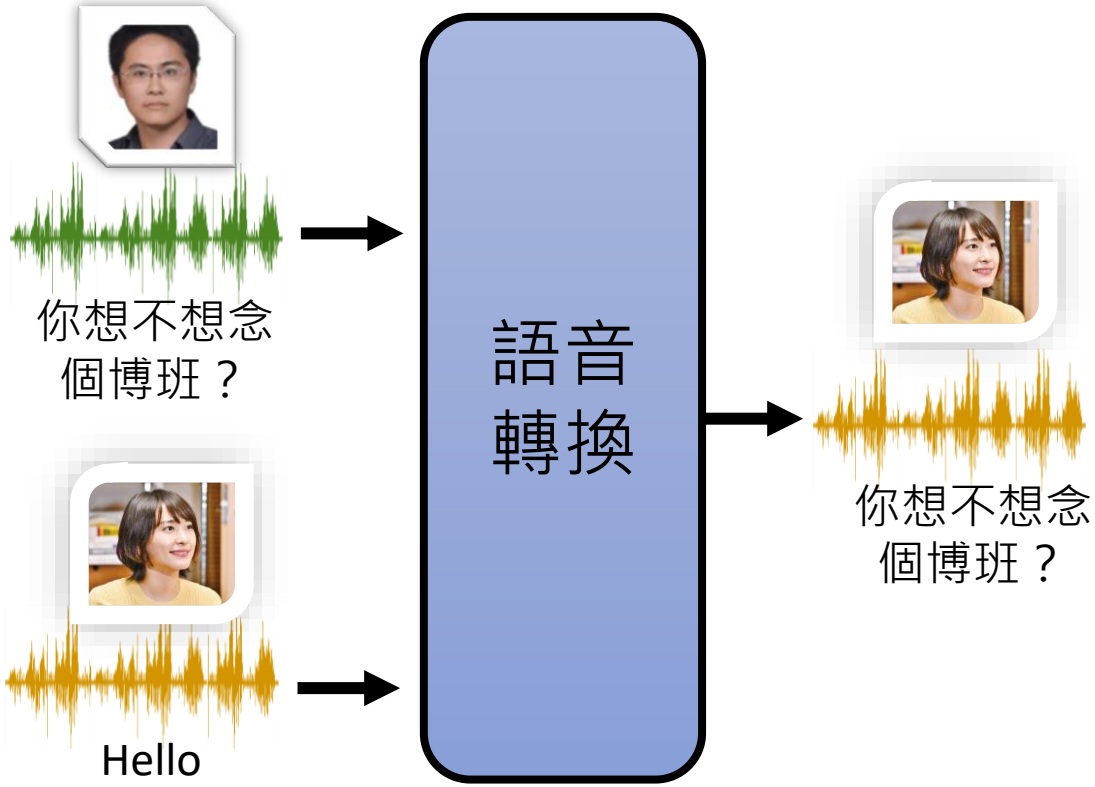
(clean speech)

https://www.kocpc.com.tw/archives/182613

Instruction Fine-tuning for Speech LM: Text-instruction-guided Voice Conversion

https://www.kocpc.com.tw/archives/182613

Instruction Fine-tuning for Speech LM: Text-instruction-guided Voice Conversion

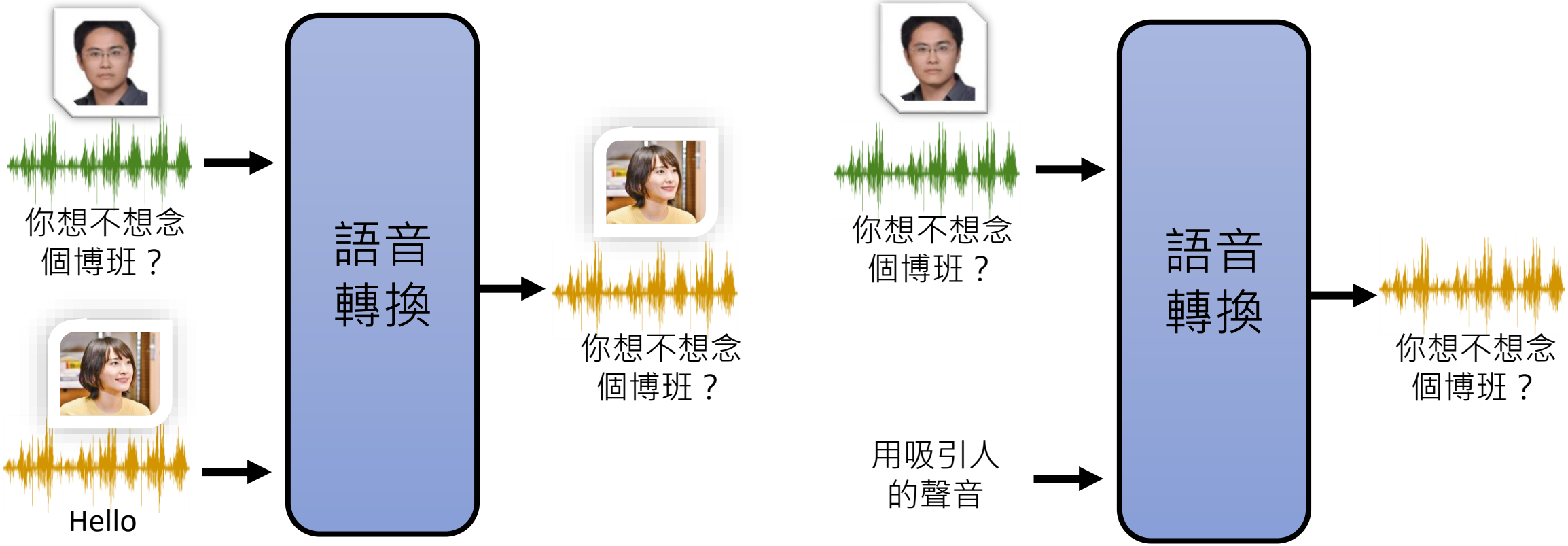


INTERSPEECH 2022 Tutorial
(Xu Tan and Hung-yi Lee)

https://tts-tutorial.github.io/interspeech2022/

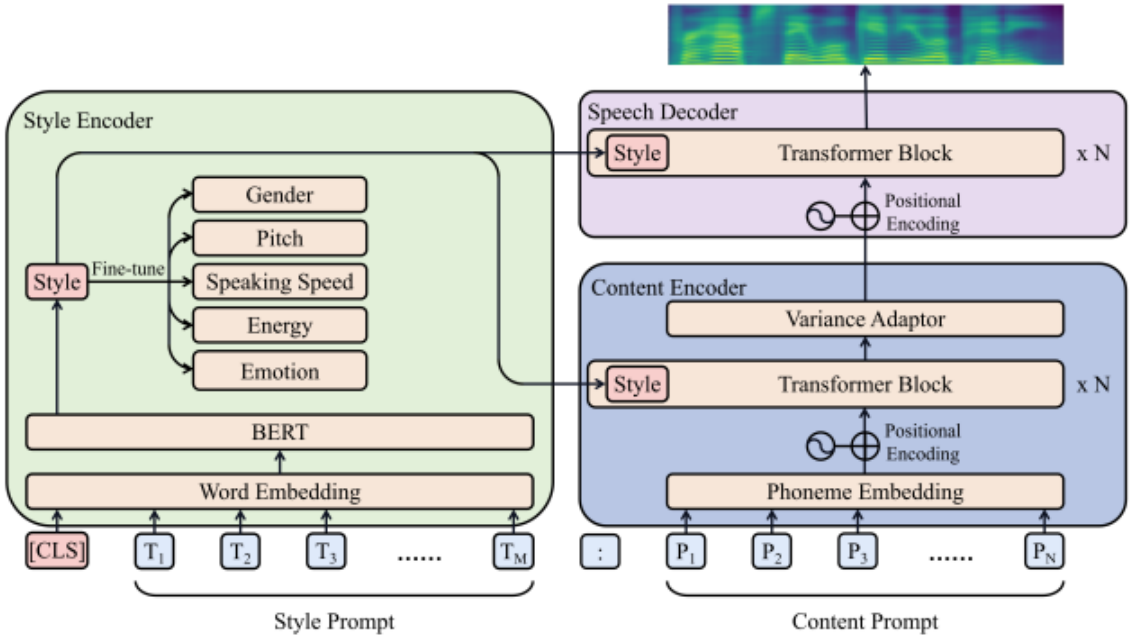
https://www.kocpc.com.tw/archives/182613

Instruction Fine-tuning for Speech LM: Text-instruction-guided Voice Conversion



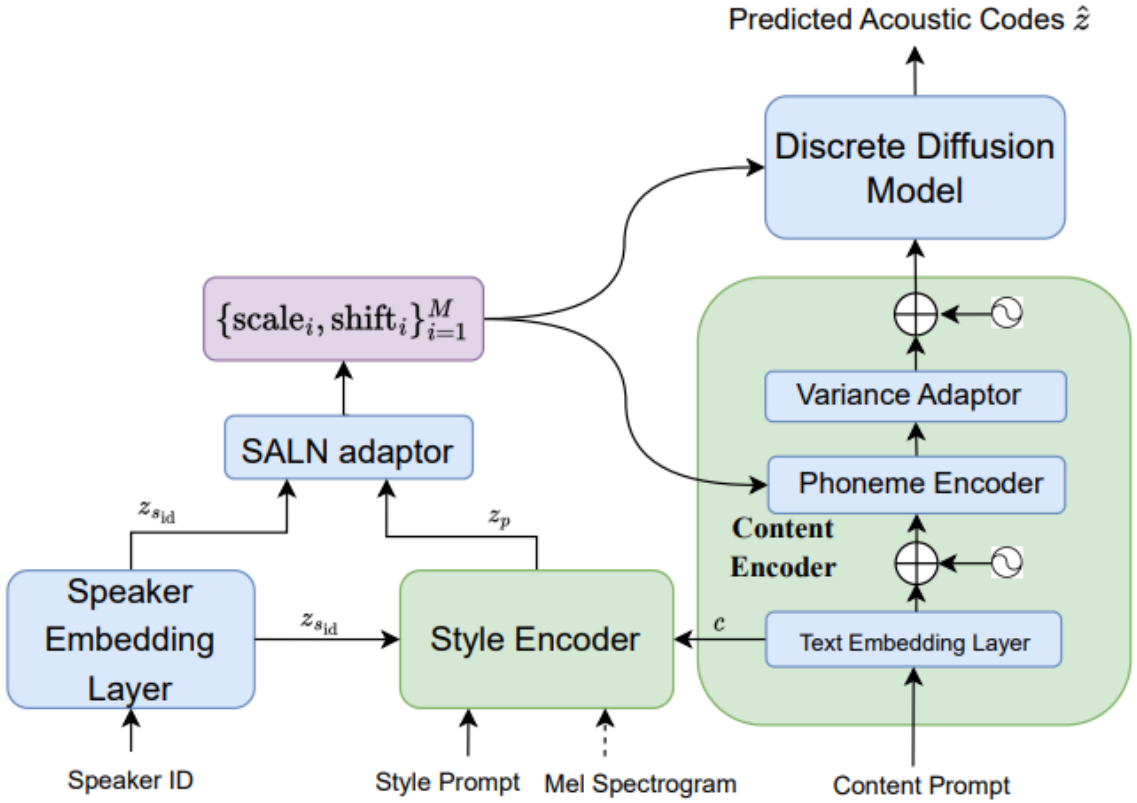
https://youtu.be/JZvEzb5PV3U

Related Work



PromptTTS

https://arxiv.org/abs/2211.12171



InstructTTS

https://arxiv.org/abs/2301.13662

Instruction Fine-tuning for Speech LM: Text-instruction-guided Voice Conversion

Chun-Yi Kuan
(NTU)

Chen-An Li
(NTU)

Tsu-Yuan Hsu
(NTU)

Tse-Yang Lin
(NTU)

Ho-Lam Chung
(NTU)

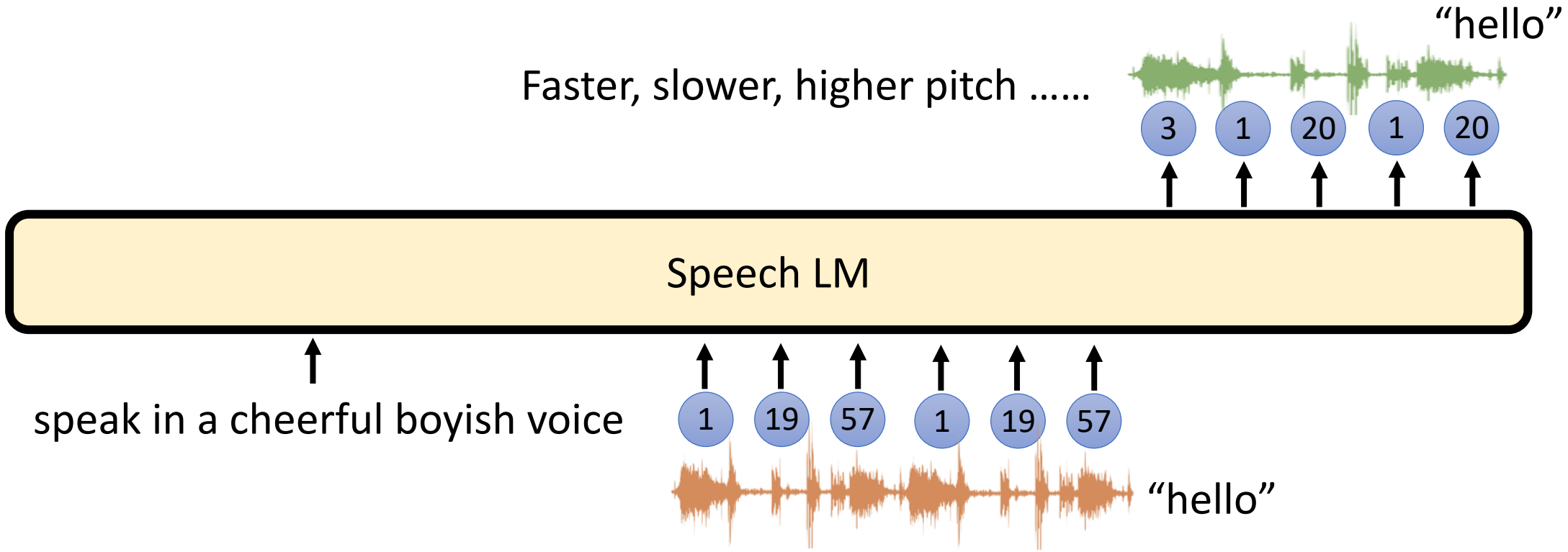
Kai-Wei Chang
(NTU)

Shuo-yiin Chang
(Google)

This research is based on work that is funded by a Googler-Initiated Grant from Shuo-Yiin Chang.

Instruction Fine-tuning for Speech LM

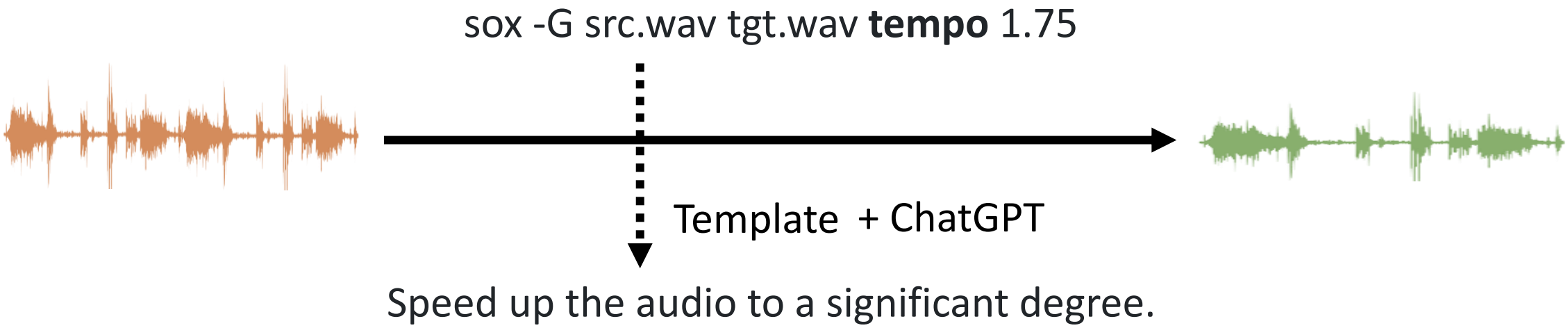
- Preliminary study: Text-instruction-guided Voice Conversion



Text-instruction-guided Voice Conversion

Training data format:

Signal Processing Effect Dataset

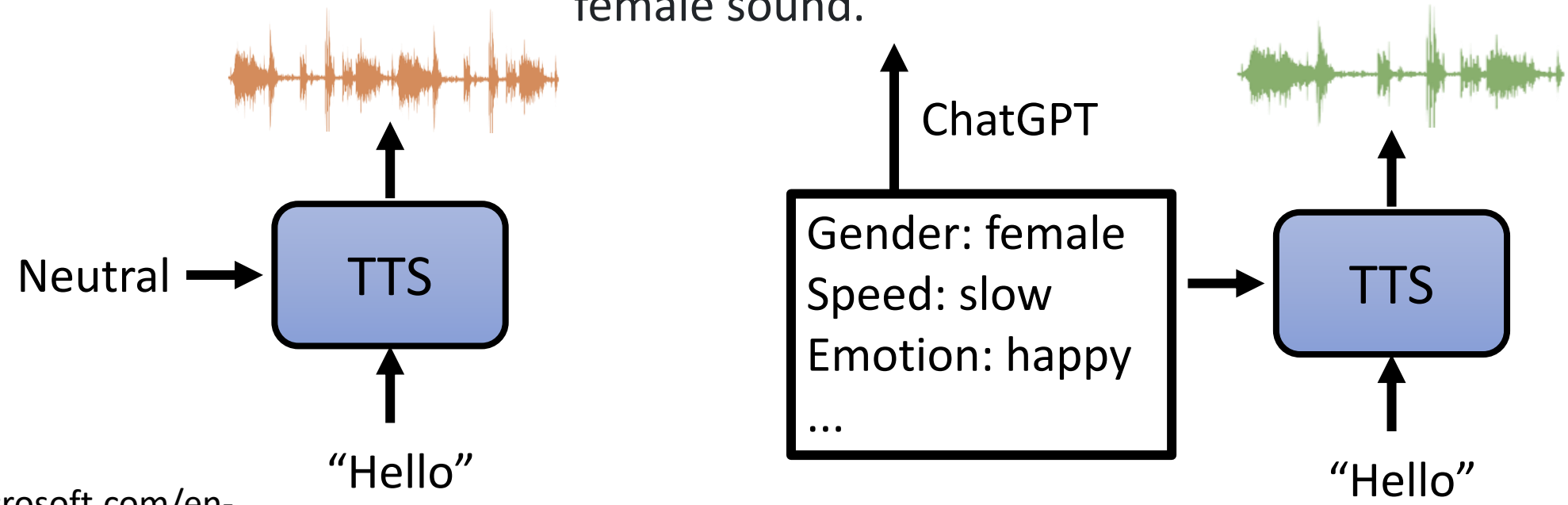


Text-instruction-guided Voice Conversion

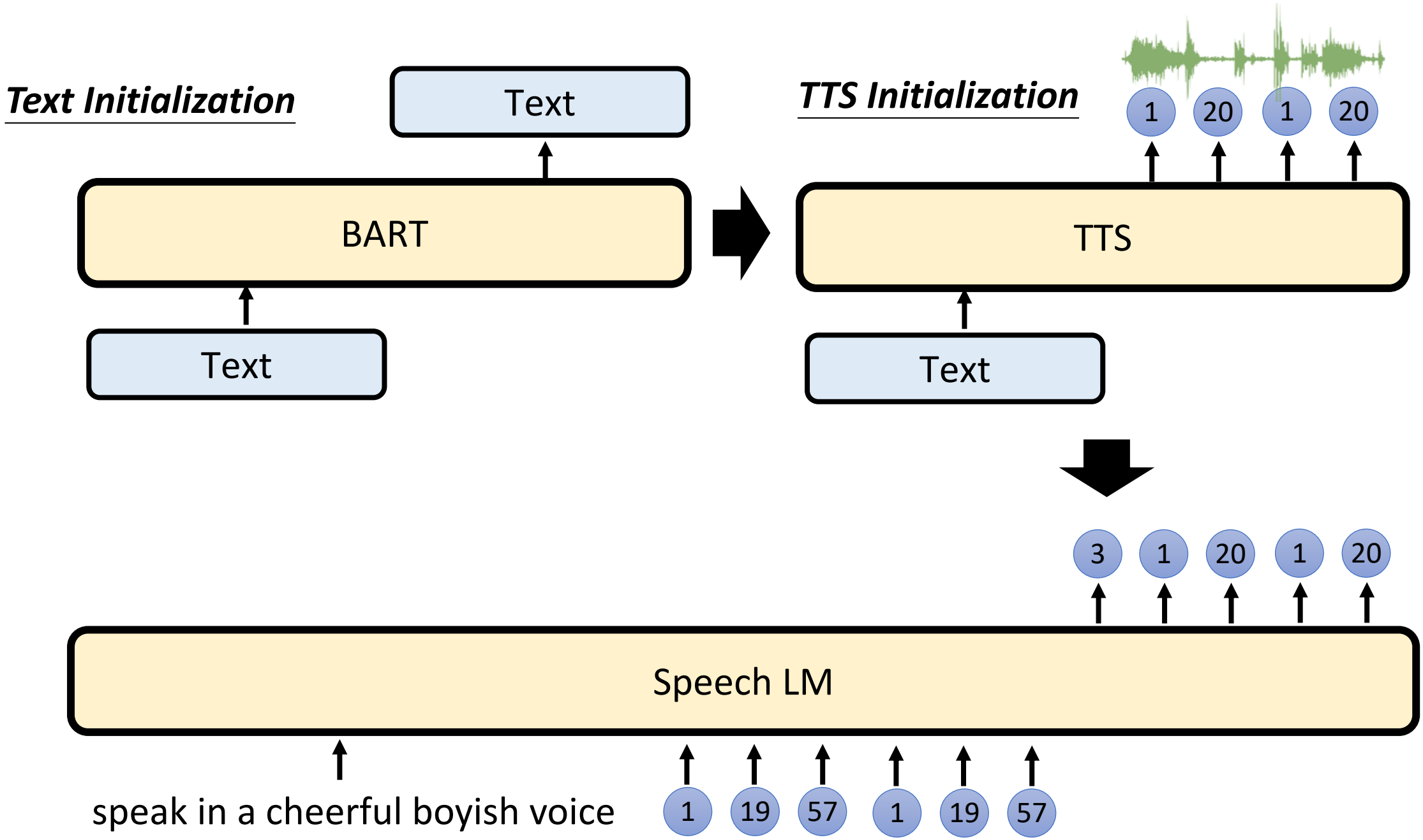
Training data format:

InstructSpeech Dataset

Please generate a slow and gaily female sound.



TTS: https://azure.microsoft.com/en-us/products/ai-services/text-to-speech


Text Initialization
Text
BART
Text
TTS Initialization
1
20
1
20
TTS
Text
3
1
20
1
20
Speech LM
speak in a cheerful boyish voice
1
19
57
1
19
57

Experimental Results

Human Evaluation (MOS Score)

	Quality	Instruction
Text-guided VC	2.81	4.19
Ground Truth	4.15	3.69

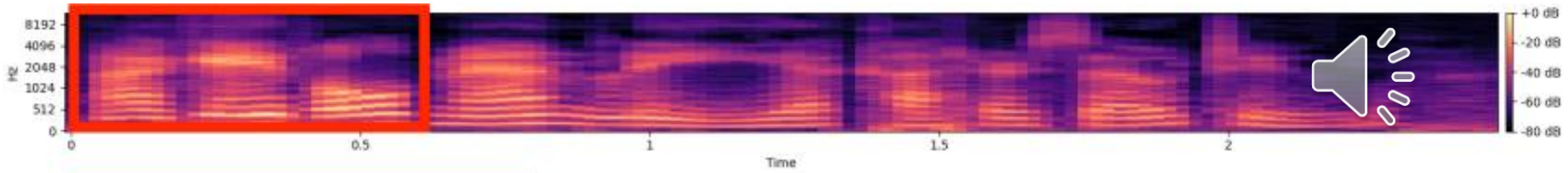
- Don’t worry. We have improved the quality by using a better EnCodec decoder.
- Even better than ground truth? The model generates more apparent conversion.

Experimental Results

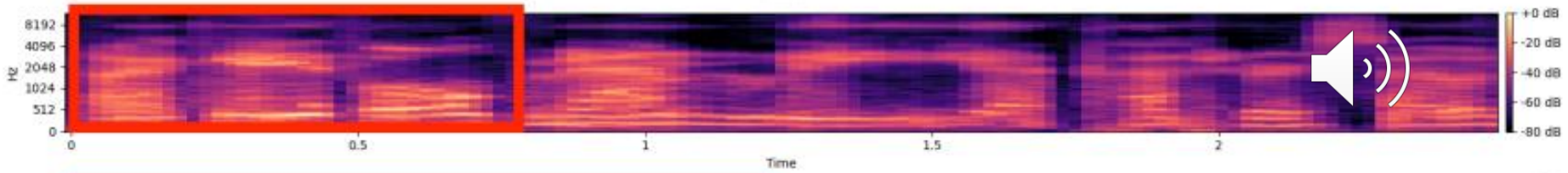
	Text Pre-training	No Pre-training
A slow sad moving female bass appeared in a low volume.		
Add a profound sense of spatial dimension for a more immersive audio experience.		
Speak as if you're telling a bedtime story to a child. **(no similar training instructions)**		
Adopt the tone of a news anchor delivering breaking news. **(no similar training instructions)**		

Experimental Results

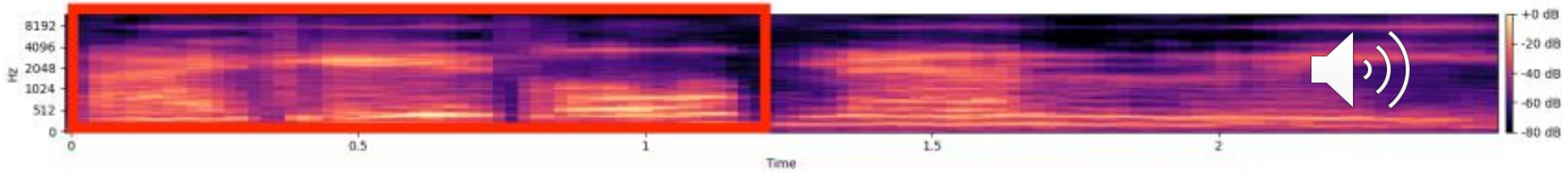
Source speech



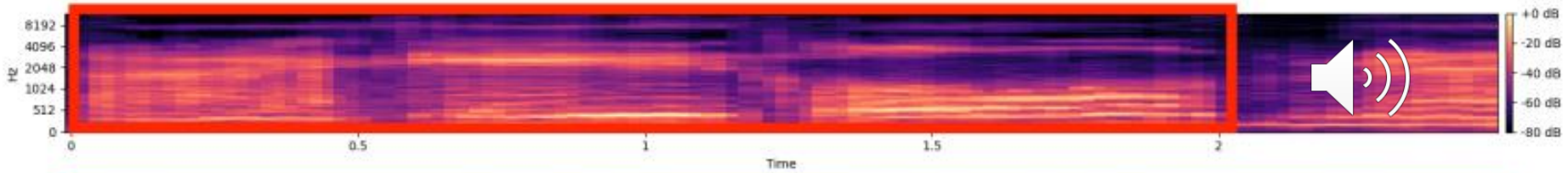
Decrease the speed of speech **slightly**.



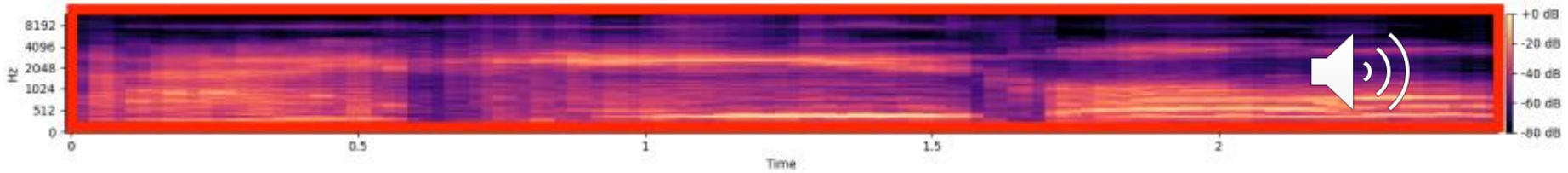
Decrease the speed of speech.



Decrease the speed of speech **notably**.



Decrease the speed of speech **extremely**.



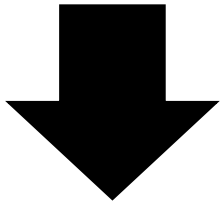
(This page is not a complete survey.)

What is still missing?

Speech

SUPERB series: 17 tasks

https://arxiv.org/abs/2105.01051
https://arxiv.org/abs/2203.06849
https://arxiv.org/abs/2210.08634
https://arxiv.org/abs/2210.07185
https://arxiv.org/abs/2110.06280



We need something bigger.

with instruction ☺

NLP

General Language Understanding Evaluation (GLUE): 9 tasks

https://arxiv.org/abs/1804.07461

Super GLUE: 8 tasks

https://arxiv.org/abs/1905.00537

FLAN: 62 tasks

https://arxiv.org/abs/2109.01652

CrossFit: 160 tasks

https://arxiv.org/abs/2104.08835

BIG-bench: 204 tasks

https://arxiv.org/abs/2206.04615

natural-instructions: 1616 tasks

https://arxiv.org/abs/2204.07705

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

Chien-yu Huang[1], *Ke-Han Lu*[*1], *Shih-Heng Wang*[*1], *Chi-Yuan Hsiao*[†1], *Chun-Yi Kuan*[†1], *Haibin Wu*[†1]
Siddhant Arora[§2], *Kai-Wei Chang*[§1], *Jiatong Shi*[2], *Yifan Peng*[2], *Roshan Sharma*[2], *Shinji Watanabe*[2]
Bhiksha Ramakrishnan[2,3], *Shady Shehata*[3], *Hung-yi Lee*[1]

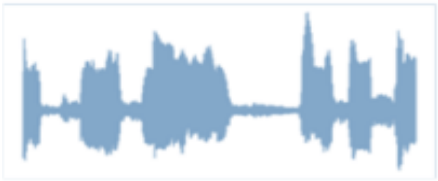
[1]National Taiwan University, Taiwan, [2]Carnegie Mellon University, USA
[3]Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates

https://arxiv.org/abs/2309.09510

Chien-yu Huang

Project page: https://github.com/dynamic-superb/dynamic-superb

Format

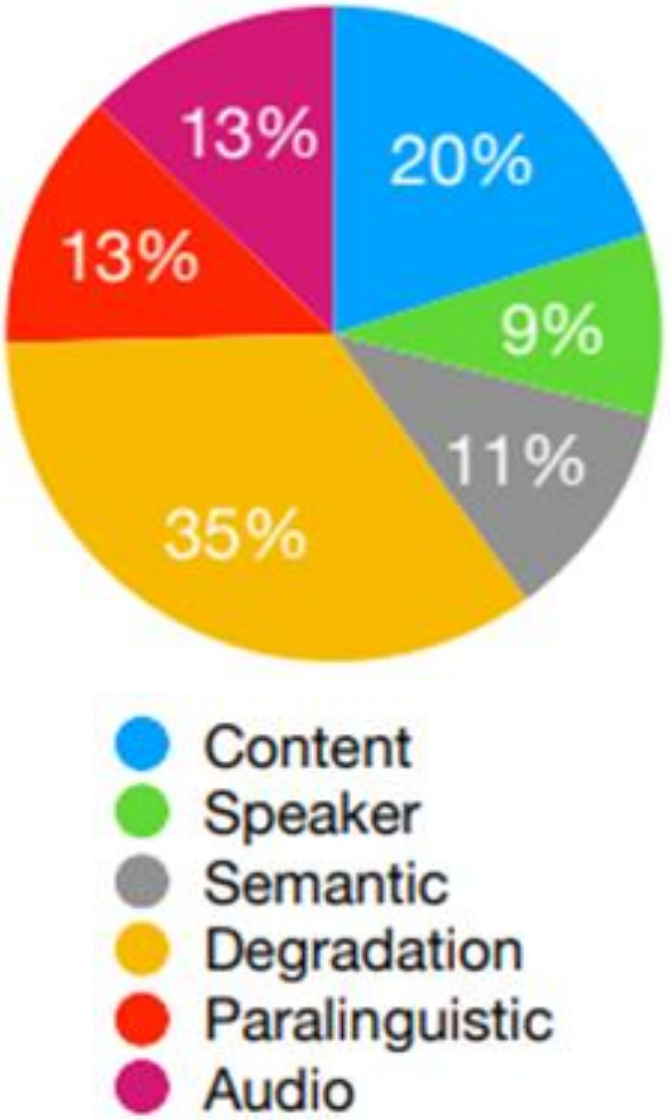
Instruction	Input	Output
Please identify the emotion in the audio. The answer could be ……		“Happiness”
Assess whether the provided speech is a spoofed voice. The answer could be ……		“Spoofed”
Recognize when sarcasm or irony is employed in the speech. The answer could be ……		“True”
Please transcribe the utterance.		“How are you”
Please speck slower.		

Current Status

- 55 tasks created from 33 datasets
- Covering 6 dimensions
  - Content: speech command recognition
  - Speaker: speaker verification
  - Semantics: sarcasm detection
  - Degradation: noise SNR prediction
  - Paralinguistic: emotion recognition
  - Audio: environmental sound classification

55 is not a big number …

They are all classification tasks ...



Let’s work together!

Role Model: BIG-bench

204 tasks

444 authors across 132 institutions

https://github.com/google/BIG-bench

Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Alphabetic author list:*

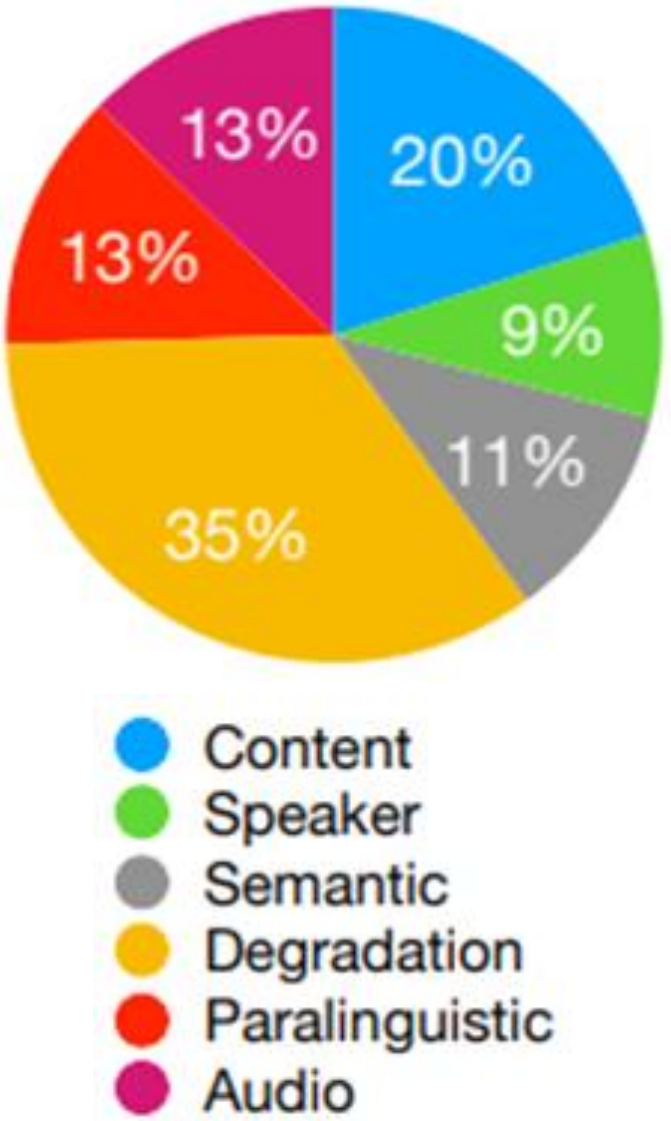
Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Mosegui Gonzalez, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jenny Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu

Current Status

- 55 tasks created from 33 datasets
- Covering 6 dimensions
  - Content: speech command recognition
  - Speaker: speaker verification
  - Semantics: sarcasm detection
  - Degradation: noise SNR prediction
  - Paralinguistic: emotion recognition
  - Audio: environmental sound classification

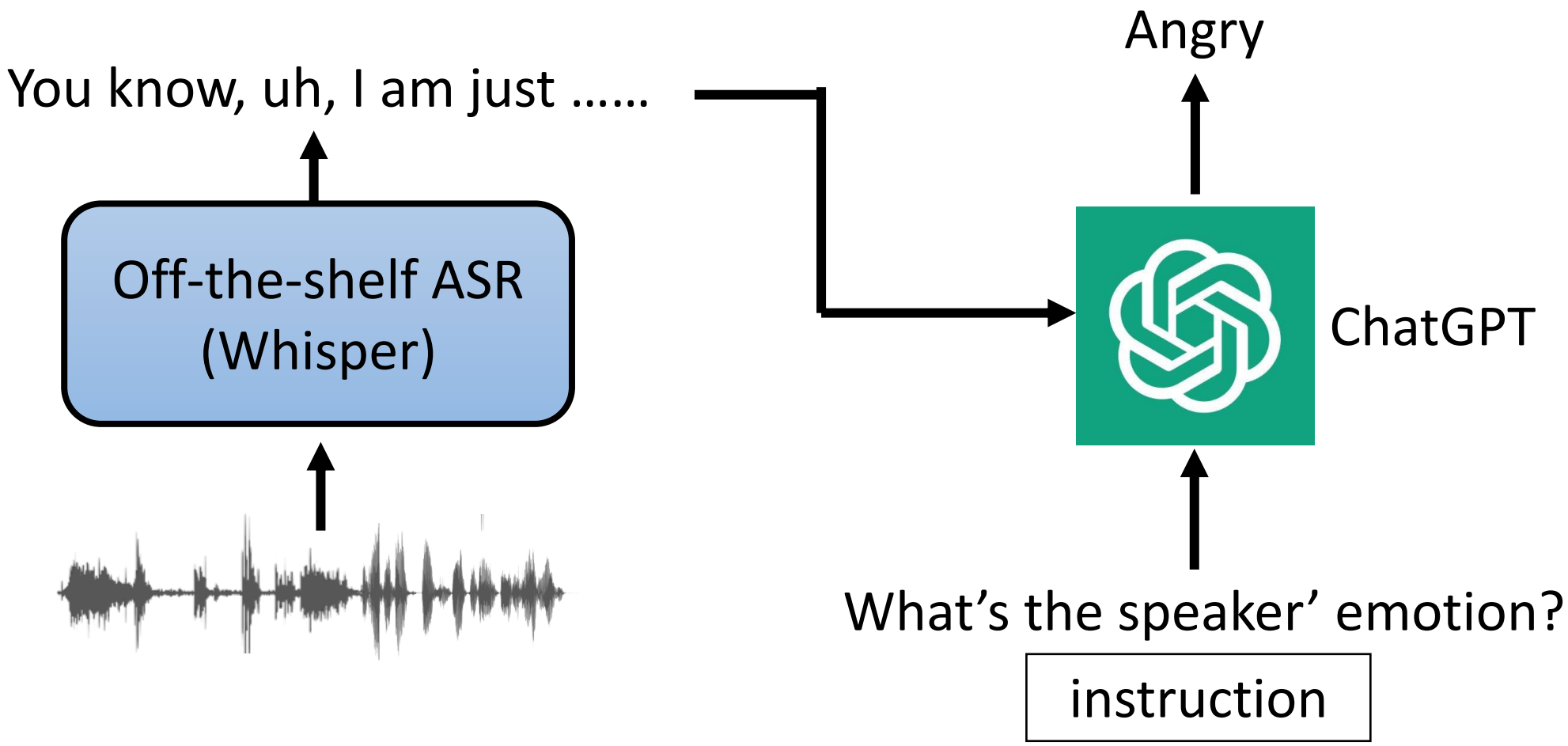
Everyone can add new tasks!

→ **Dynamic**

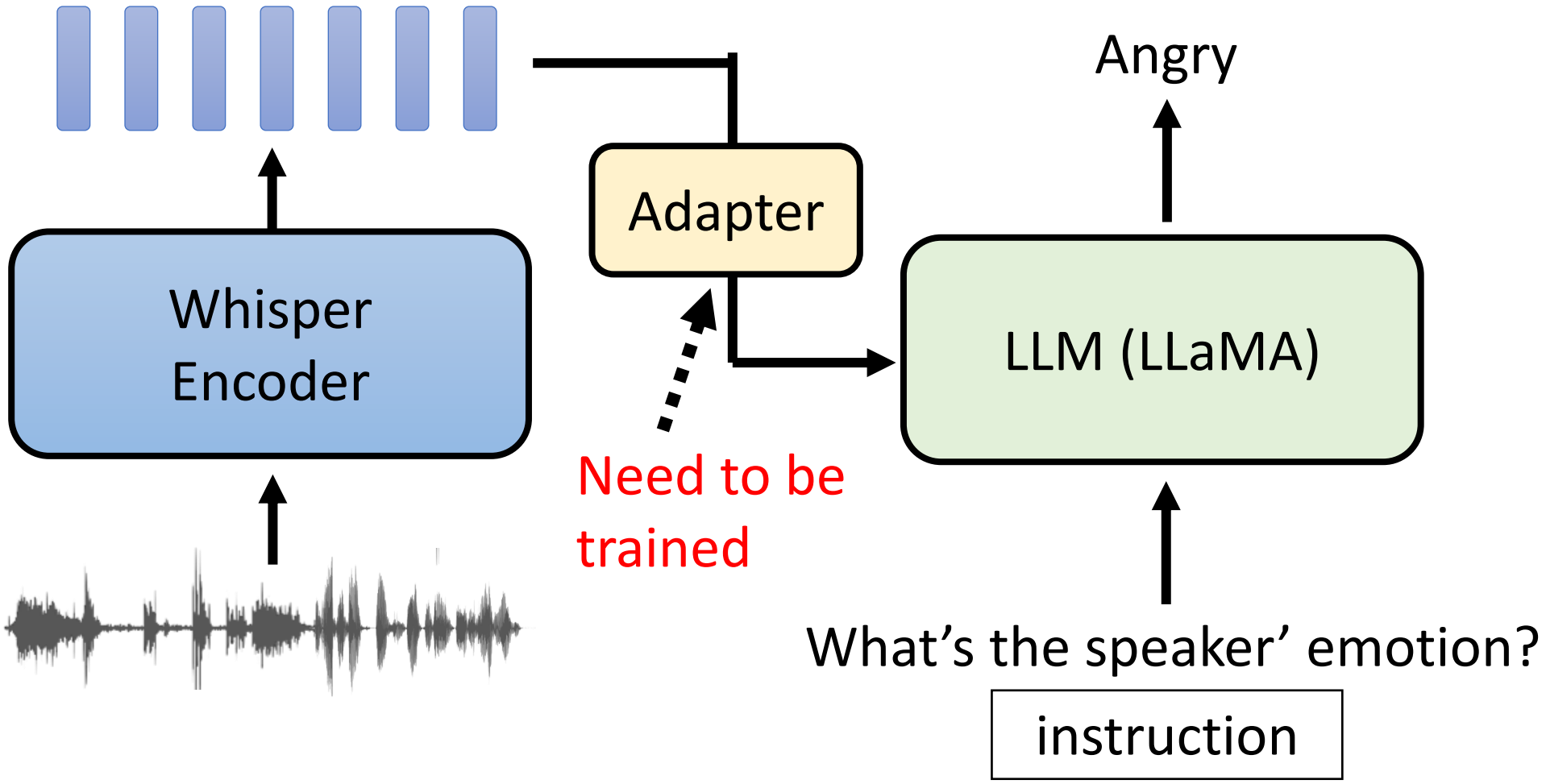


Let's work together!

Baseline: ASR + ChatGPT

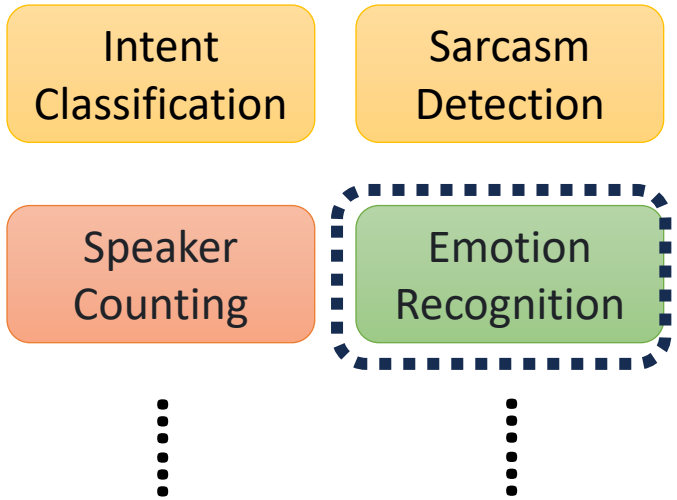


Baseline: Whisper Encoder + LLM



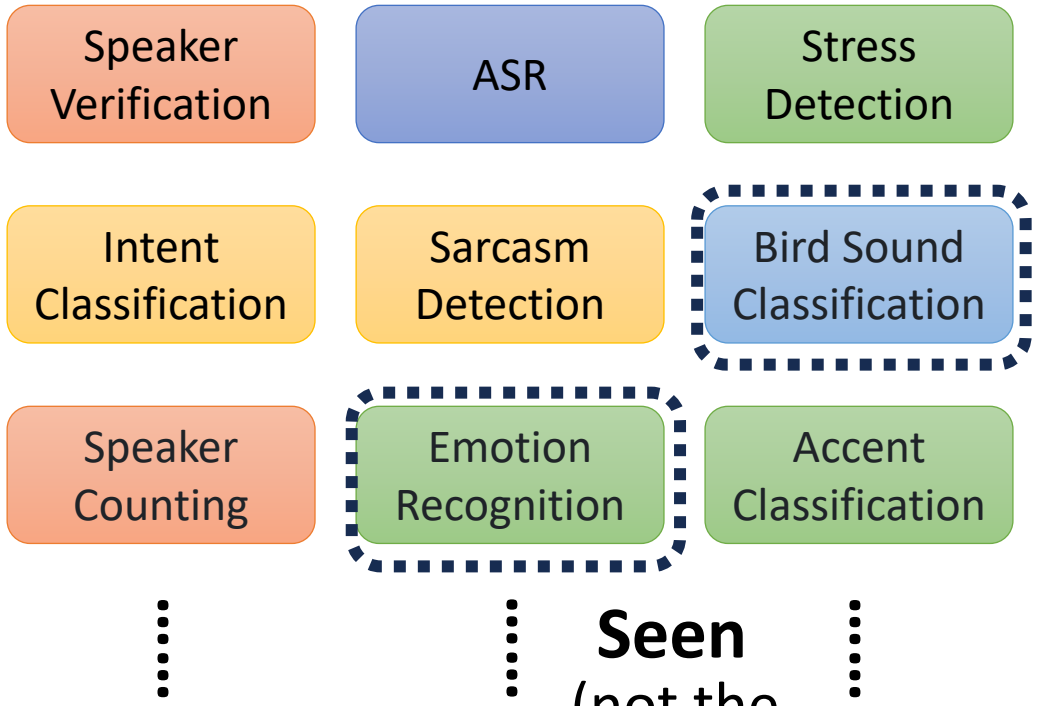
Training Tasks vs. Testing Tasks

Training Tasks (23 tasks)



Training model parameters

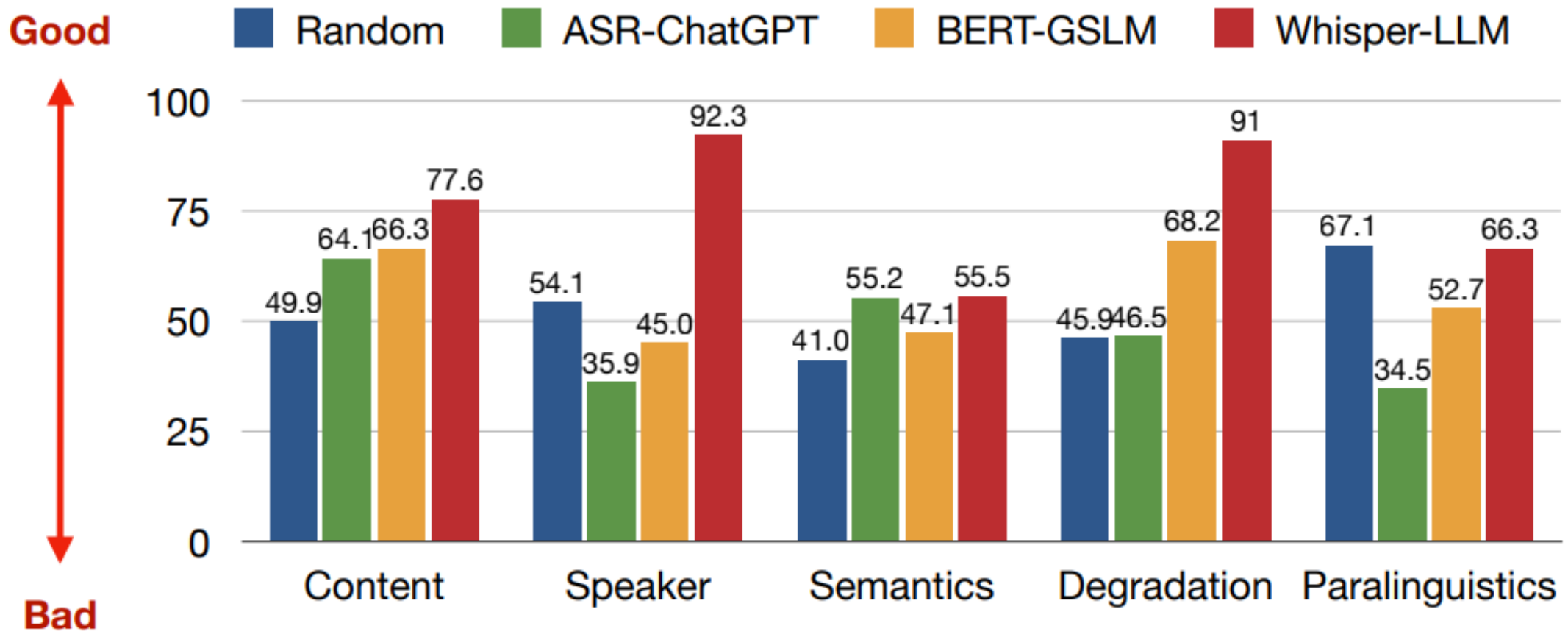
Dynamic-SUPERB (55 tasks)



Unseen

Seen (not the same data)

Overall Results



Dynamic-SUPERB