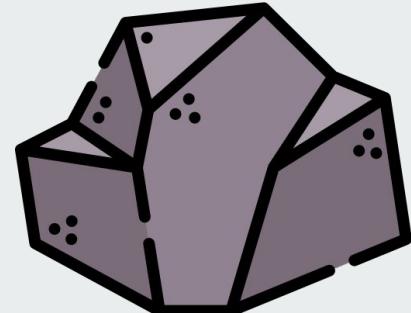


# Speech Foundation Models

## 語音基石模型

2023/05/12 張凱爲  
[kaiwei.chang.tw@gmail.com](mailto:kaiwei.chang.tw@gmail.com)



# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. VALL-E

### Part 3

#### Other Speech Foundation Models

1. OpenAI Whisper
2. Google USM

# Overview

## Part 1

### Speech

### Representation Learning

1. SSL Models  

2. Representation benchmarking

 **SUPERB**

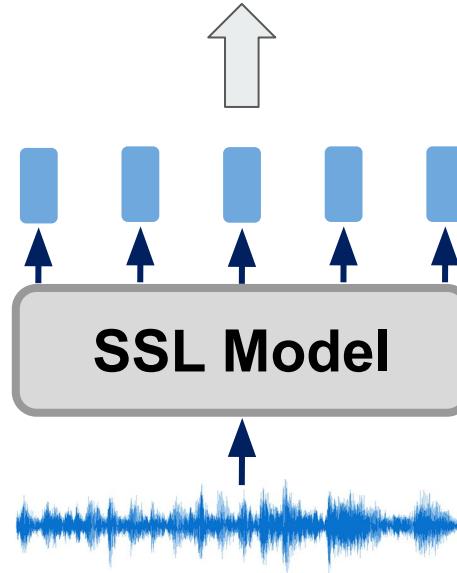
Speech Recognition  
Speaker identification

## Downstream Models

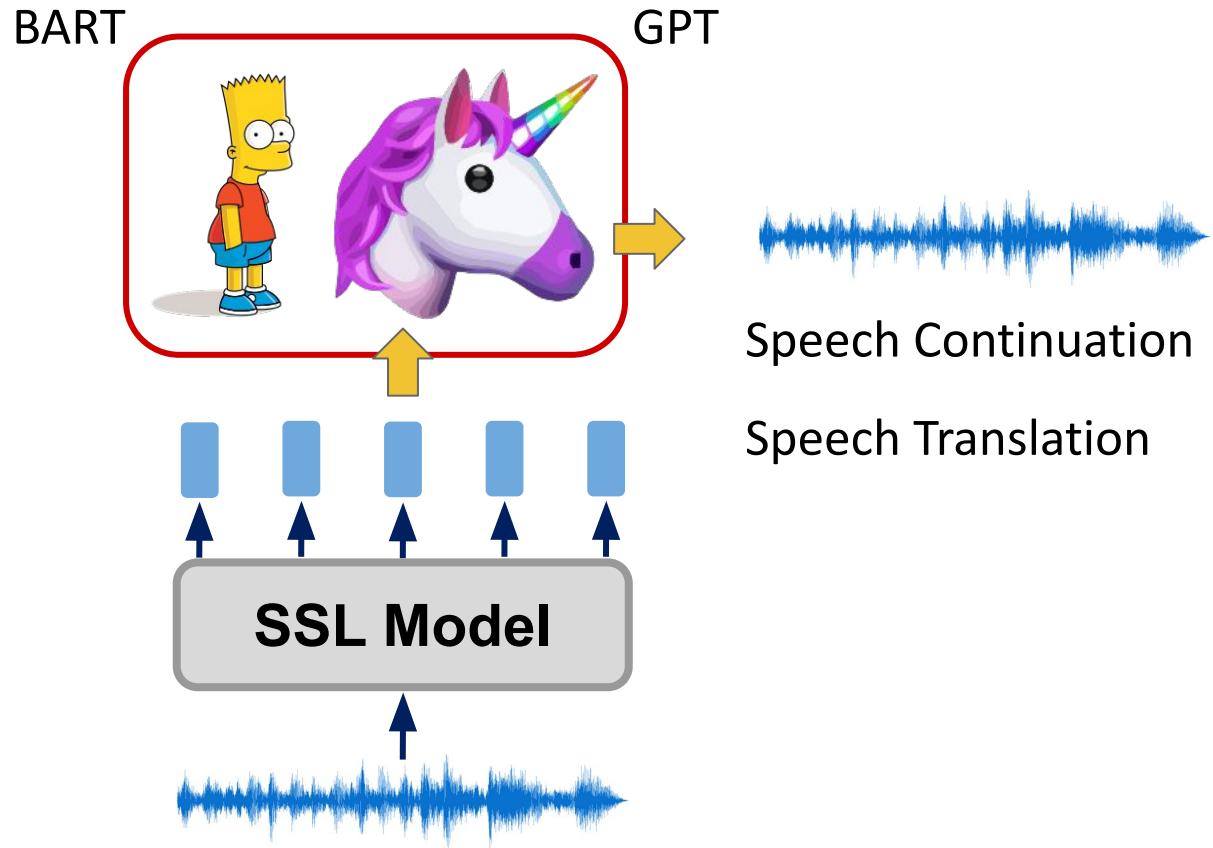
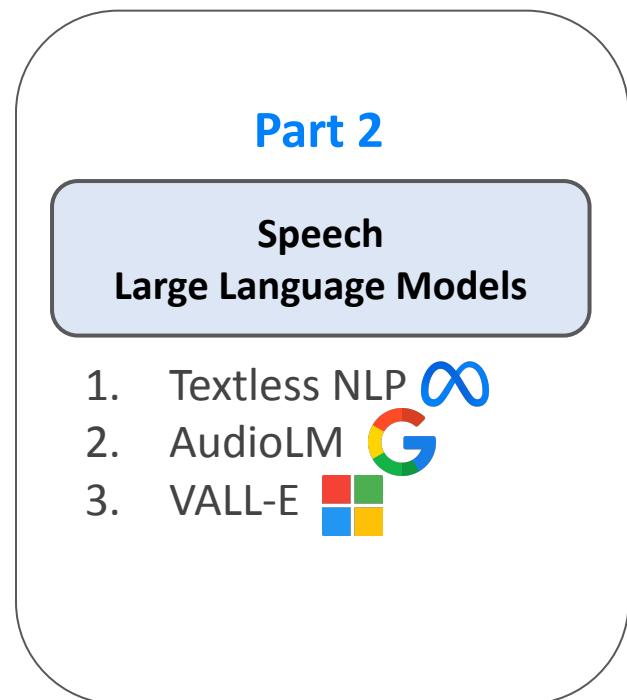
Speech  
Representation

Self-supervised  
Learning Model

Speech



# Overview

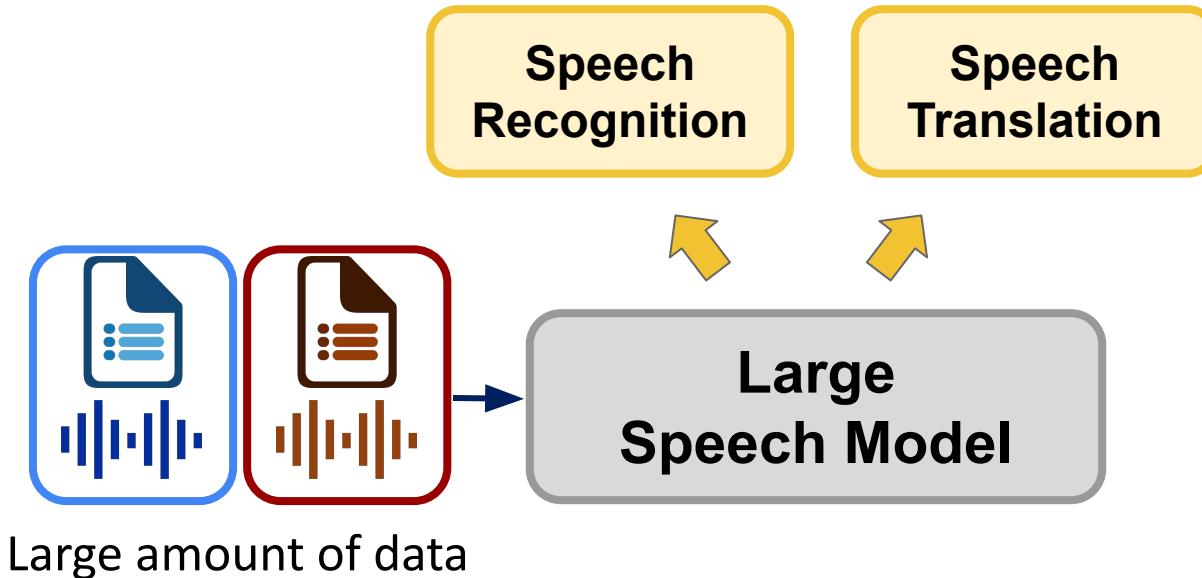


# Overview

## Part 3

### Other Speech Foundation Models

1. Whisper 
2. USM 



# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. Regeneration Framework

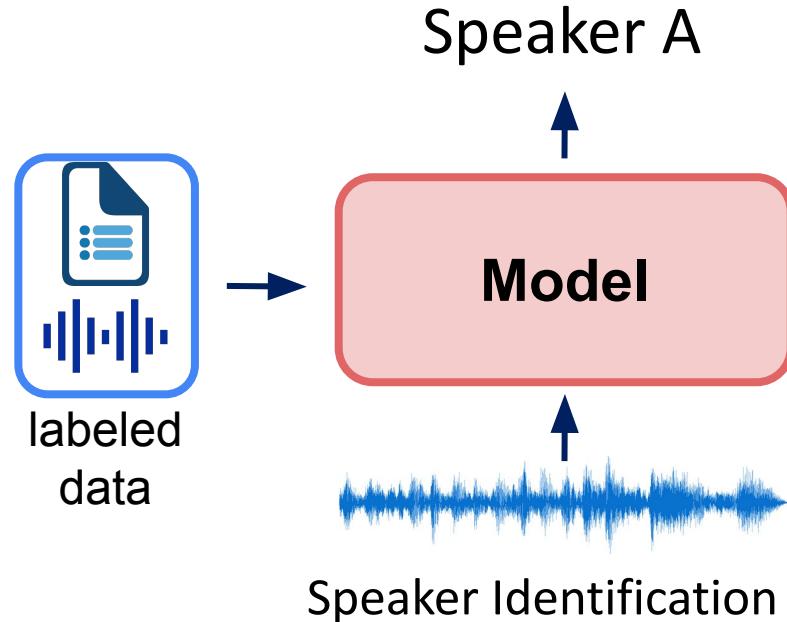
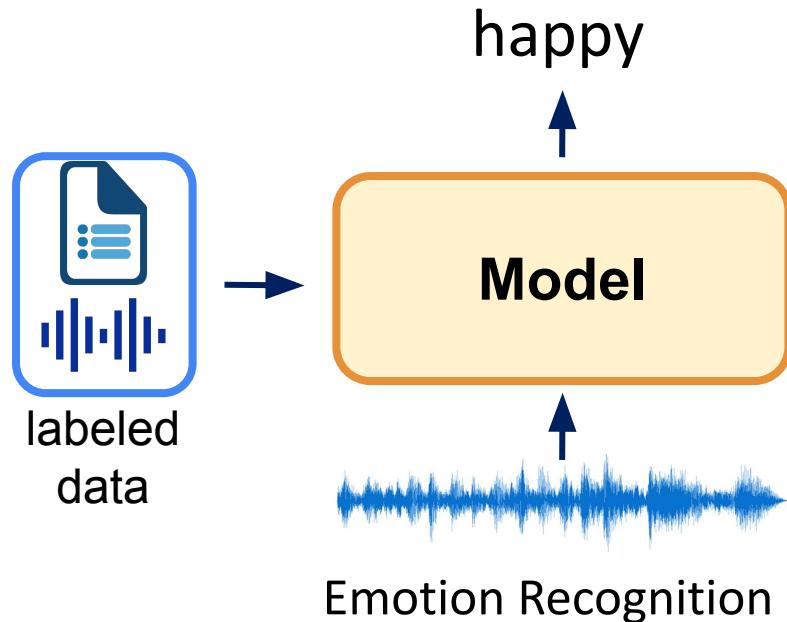
### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

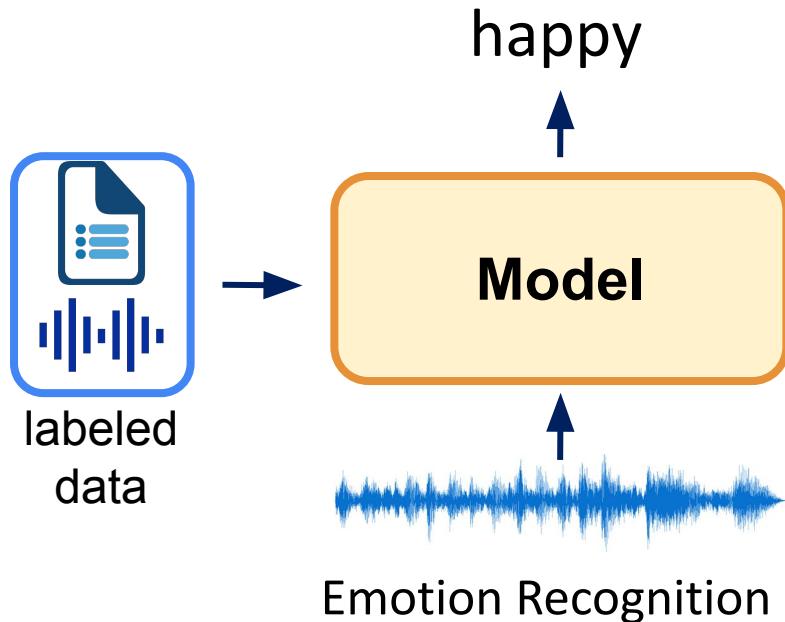
# Speech Representation Learning

Why speech representation learning?



# Speech Representation Learning

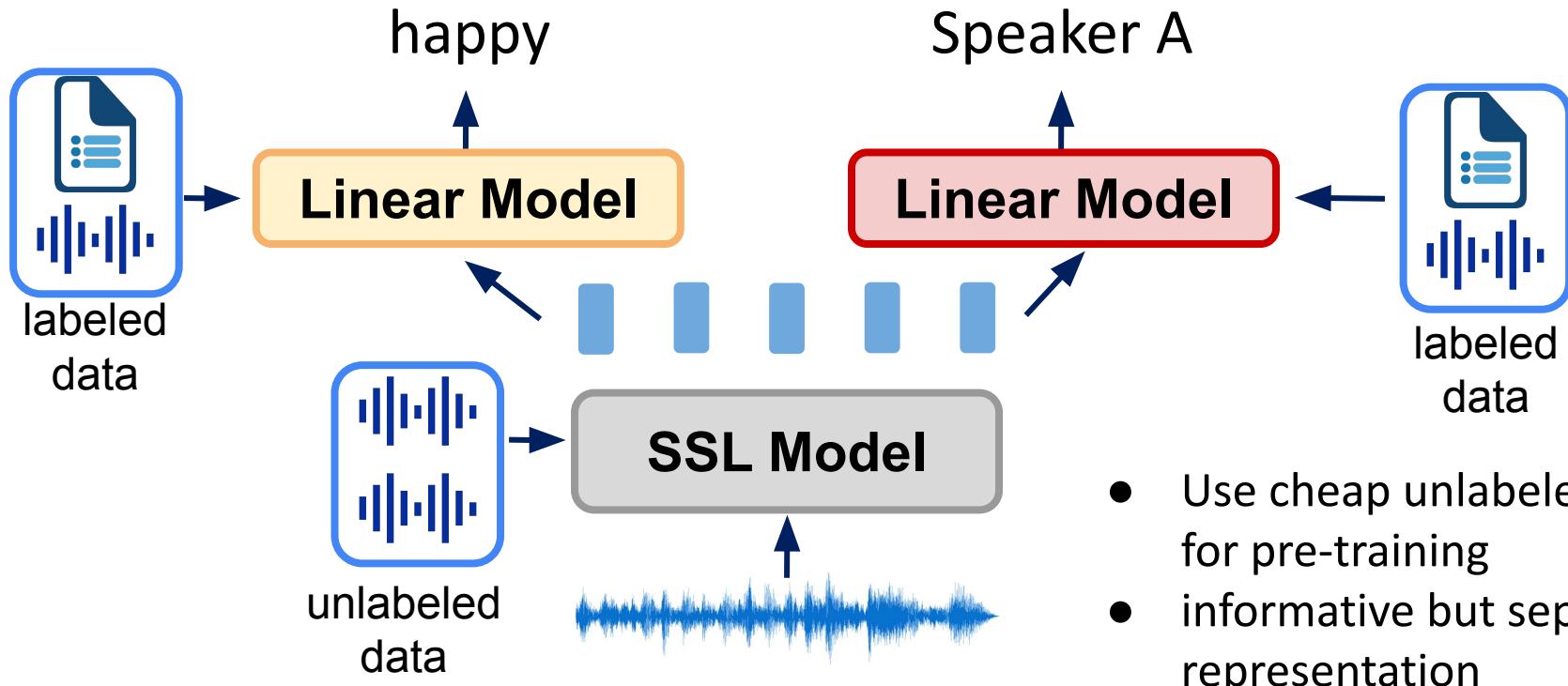
## Why speech representation learning?



- Fully supervised learning
- labeled data is expensive
  - Train a new model for each task

# Speech Representation Learning

Why speech representation learning?



# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

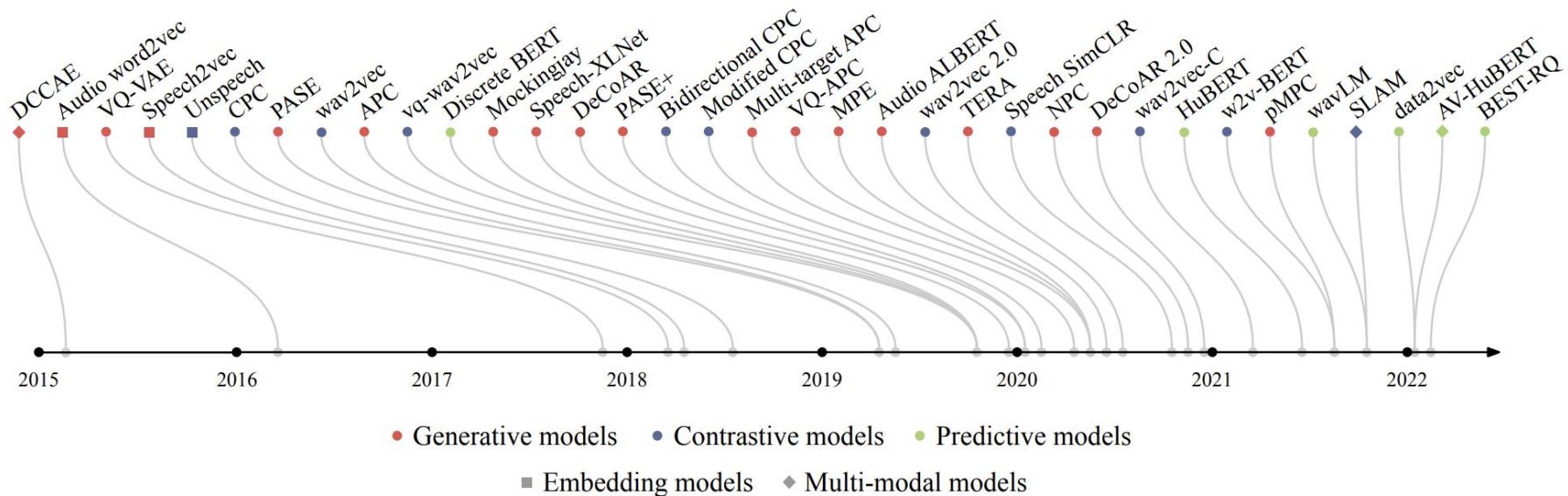
1. Textless NLP
2. AudioLM
3. Regeneration Framework

### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

# Self-Supervised Speech Representation Learning



# SSL Speech Representation Learning Models

CPC

HuBERT

wav2vec 2.0

WavLM

XLS-R

BEST-RQ

Contrastive Models

Predictive Models

# SSL Speech Representation Learning Models

CPC

wav2vec 2.0

XLS-R

HuBERT

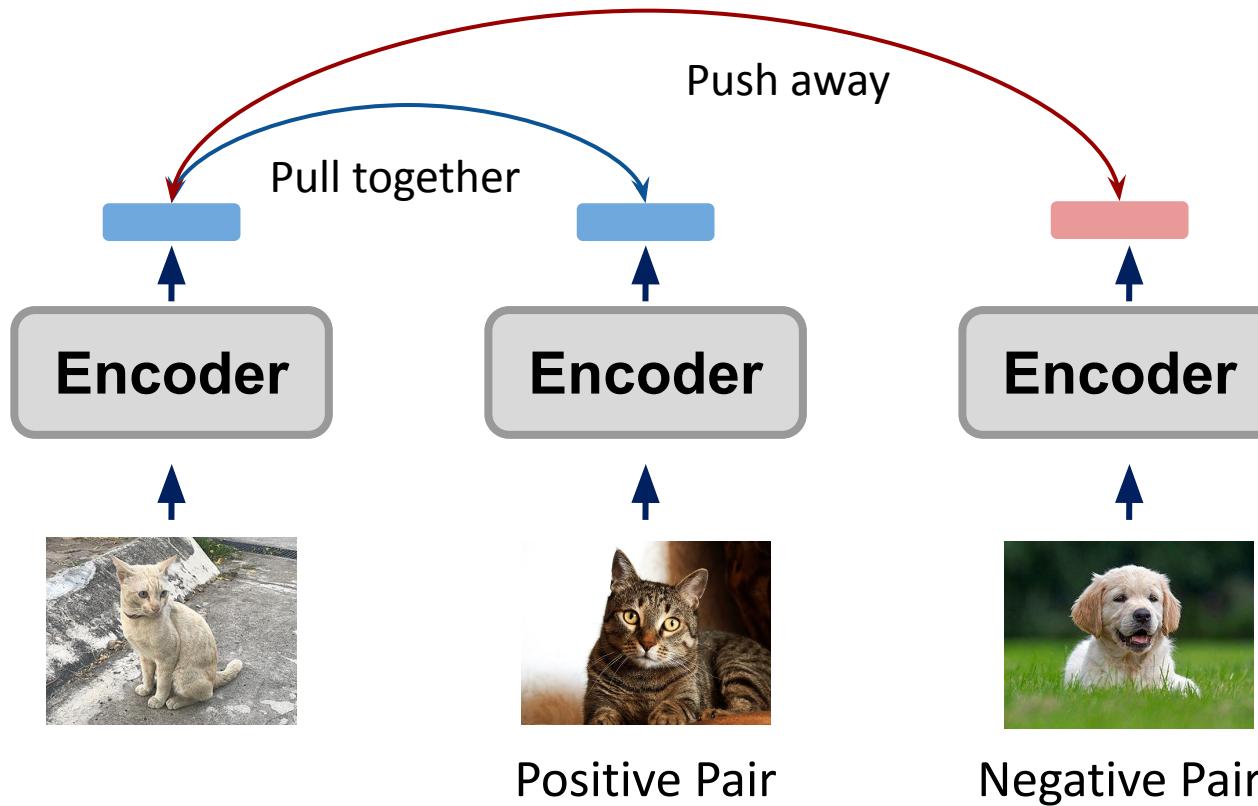
WavLM

BEST-RQ

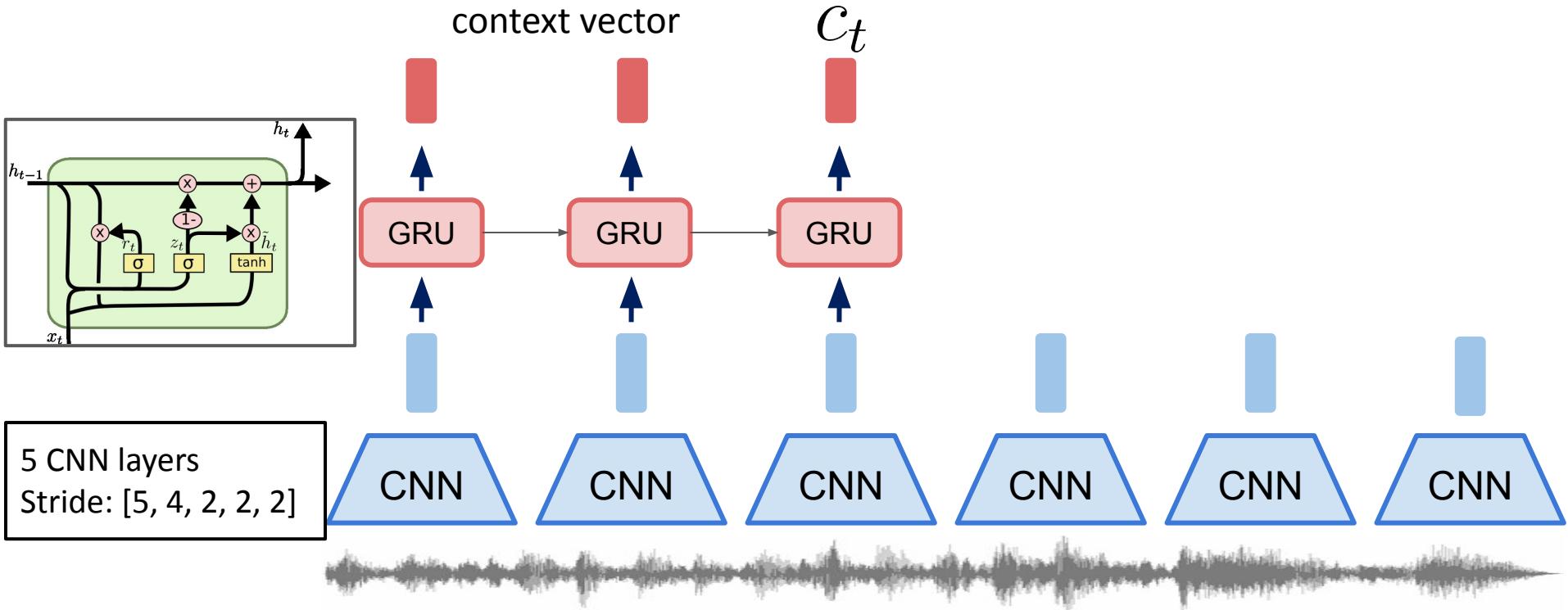
Contrastive Models

Predictive Models

# Contrastive Learning: Intuition



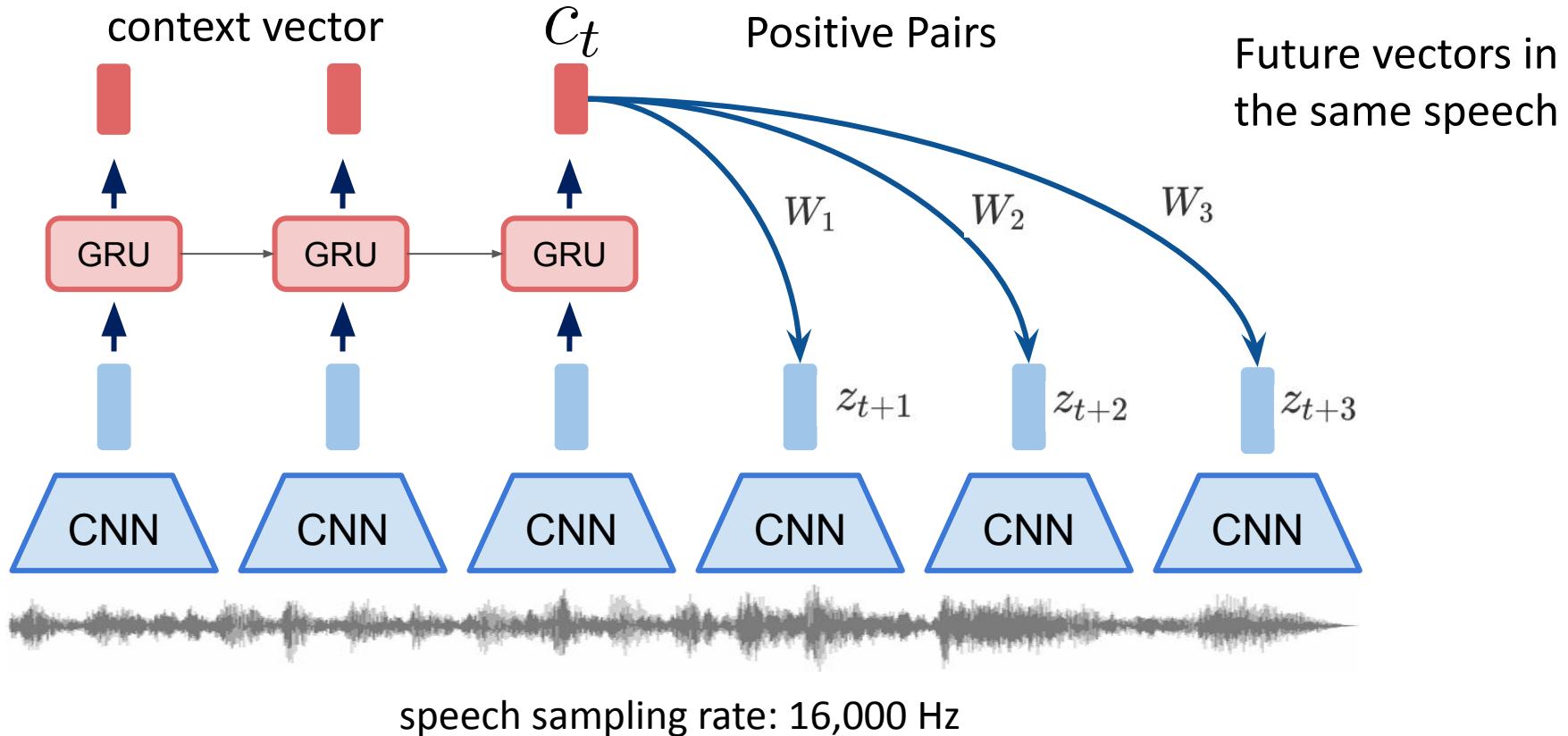
# Contrastive Predictive Coding (CPC)



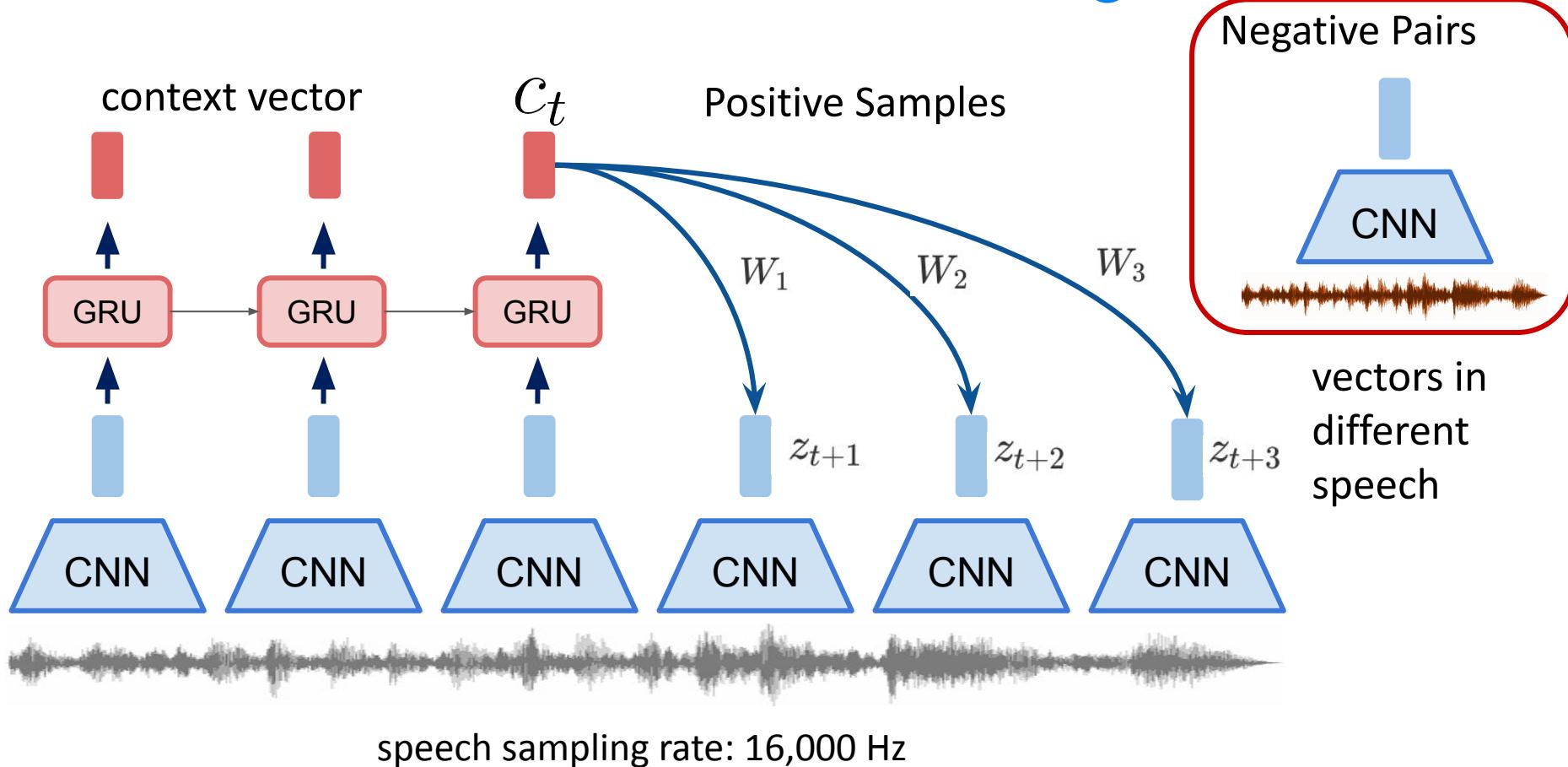
down sampling  
rate: 160

speech sampling rate: 16,000 Hz

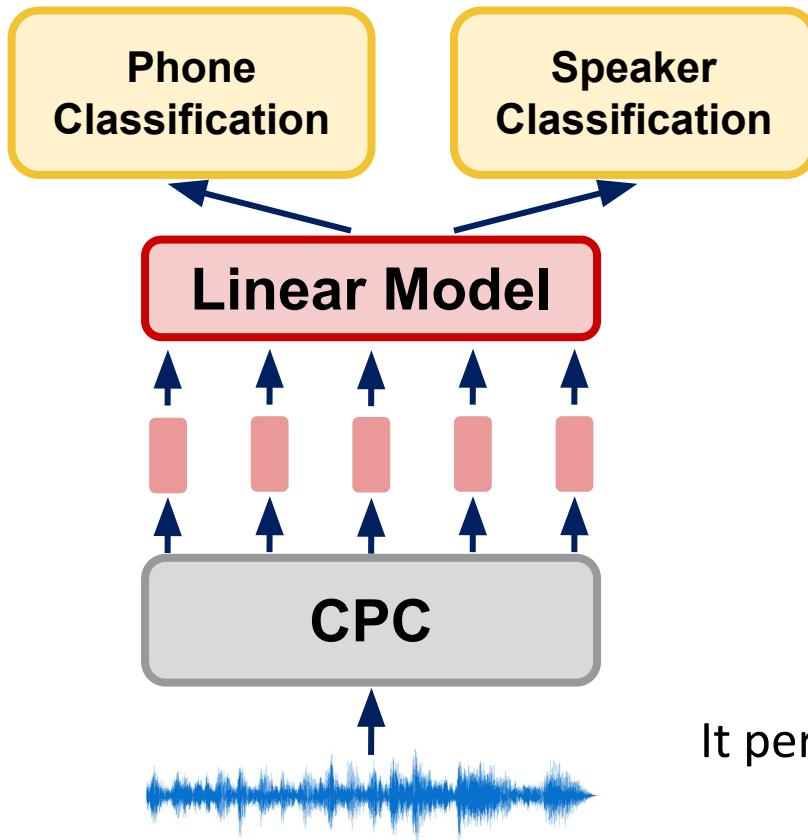
# Contrastive Predictive Coding (CPC)



# Contrastive Predictive Coding (CPC)



# Contrastive Predictive Coding (CPC)



Method	ACC
<b>Phone classification</b>	
Random initialization	27.6
MFCC features	39.7
<b>CPC</b>	64.6
Supervised	74.6
<b>Speaker classification</b>	
Random initialization	1.87
MFCC features	17.6
<b>CPC</b>	97.4
Supervised	98.5

It performs well on both content and speaker tasks!

# SSL Speech Representation Learning Models

CPC

HuBERT

**wav2vec 2.0**

WavLM

XLS-R

BEST-RQ

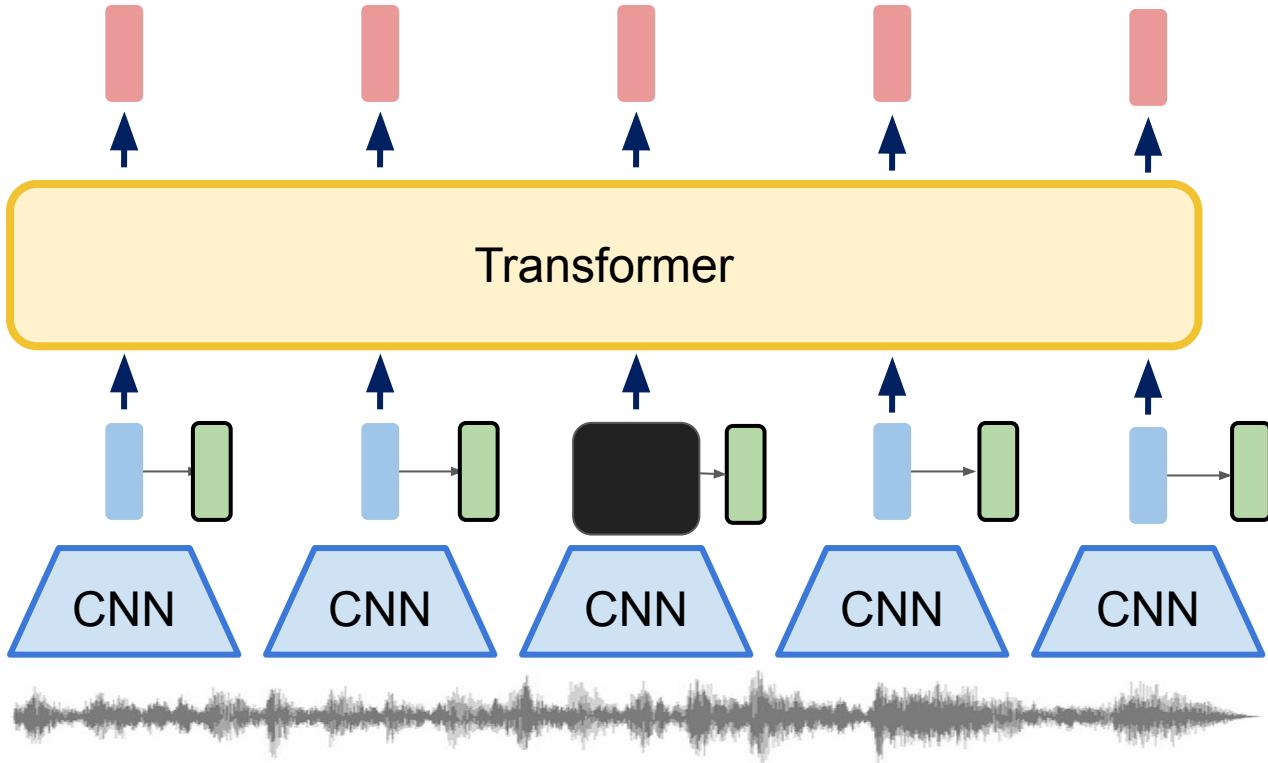
Contrastive Models

Predictive Models

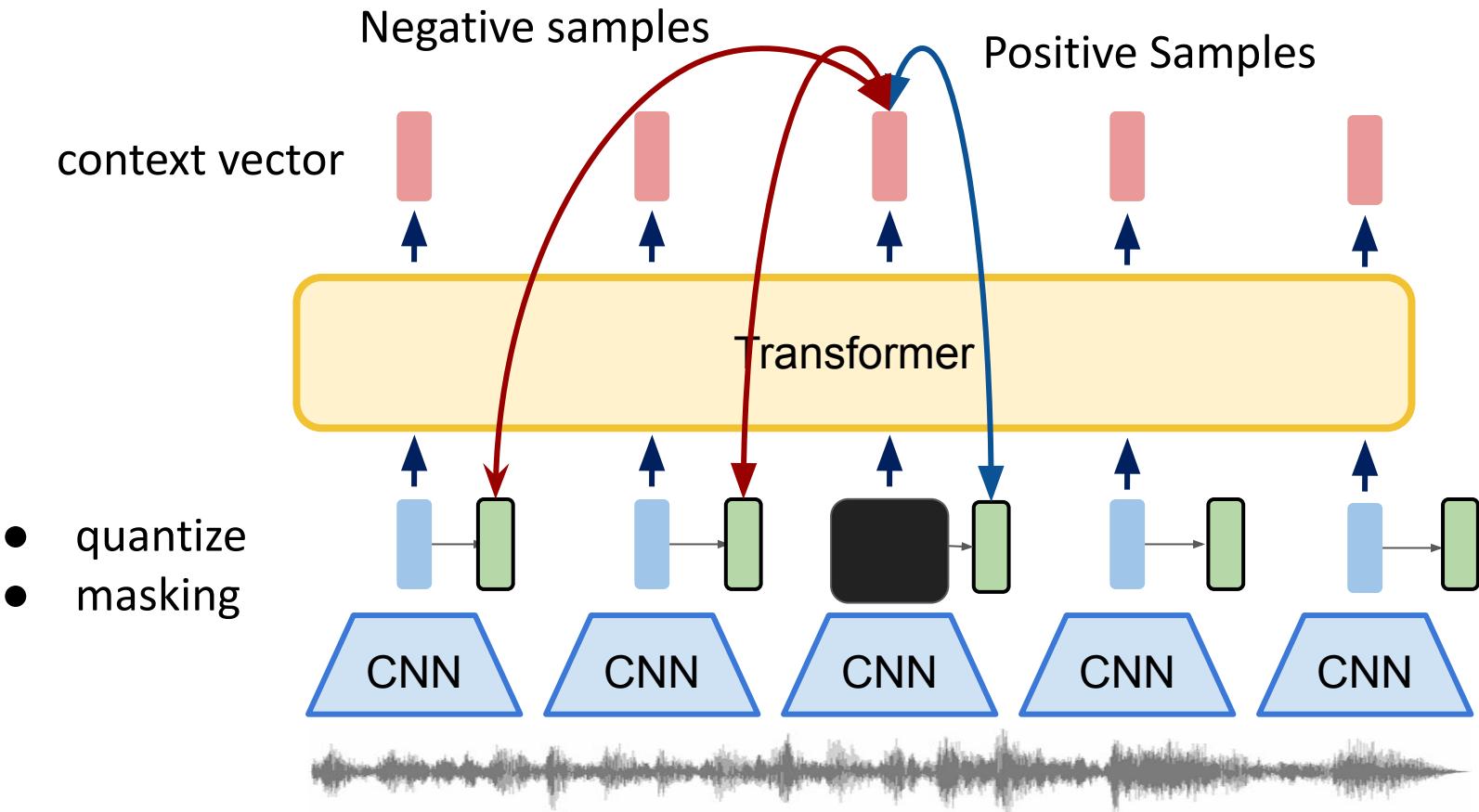
# Wav2vec 2.0

context vector

- quantize
- masking



# Wav2vec 2.0



# SSL Speech Representation Learning Models

CPC

HuBERT

wav2vec 2.0

WavLM

XLS-R

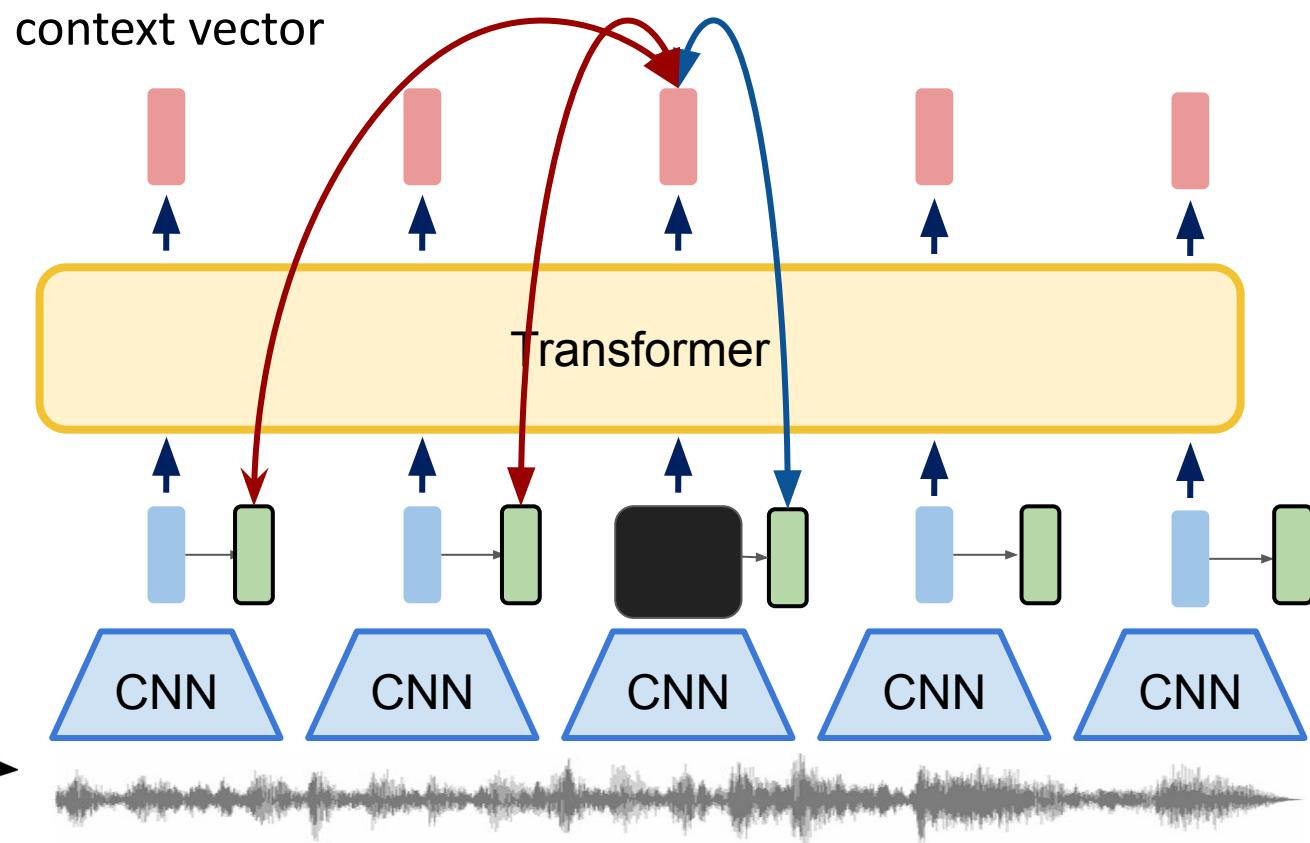
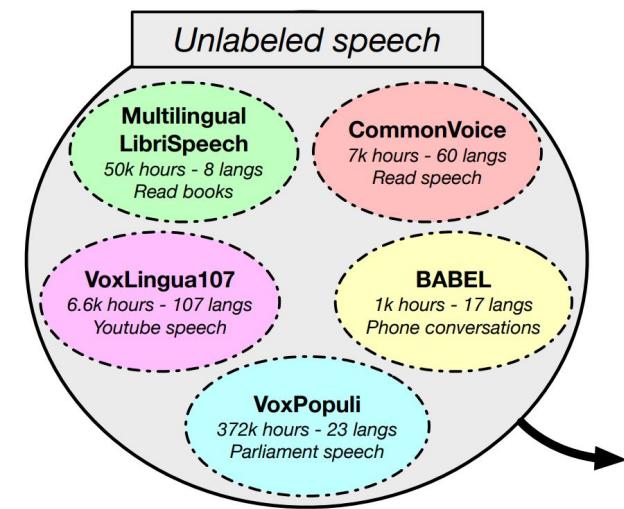
BEST-RQ

Contrastive Models

Predictive Models

# XLS-R

~ 400,000 hours  
107 languages



- Performance drop when the size is the same as wav2vec 2.0

LibriSpeech ASR results

Model	dev		test	
	clean	other	clean	other
<b>10 min labeled</b>				
wav2vec 2.0 LV-60K (0.3B)	31.7	35.0	32.1	34.5
XLS-R (0.3B)	33.3	39.8	34.1	39.6
XLS-R (1B)	<b>28.4</b>	<b>32.5</b>	<b>29.1</b>	<b>32.5</b>
<b>1h labeled</b>				
wav2vec 2.0 LV-60K (0.3B)	13.7	<b>16.9</b>	13.7	<b>17.1</b>
XLS-R (0.3B)	17.1	23.7	16.8	24.0
XLS-R (1B)	<b>13.2</b>	17.0	<b>13.1</b>	17.2
<b>10h labeled</b>				
wav2vec 2.0 LV-60K (0.3B)	<b>5.7</b>	<b>9.2</b>	<b>5.6</b>	<b>9.4</b>
XLS-R (0.3B)	8.3	15.1	8.3	15.4
XLS-R (1B)	5.9	10.5	5.9	10.6

## LibriSpeech ASR results

Model	dev		test	
	clean	other	clean	other
<b>10 min labeled</b>				
wav2vec 2.0 LV-60K (0.3B)	31.7	35.0	32.1	34.5
XLS-R (0.3B)	33.3	39.8	34.1	39.6
XLS-R (1B)	<b>28.4</b>	<b>32.5</b>	<b>29.1</b>	<b>32.5</b>
<b>1h labeled</b>				
wav2vec 2.0 LV-60K (0.3B)	13.7	<b>16.9</b>	13.7	<b>17.1</b>
XLS-R (0.3B)	17.1	23.7	16.8	24.0
XLS-R (1B)	<b>13.2</b>	17.0	<b>13.1</b>	17.2
<b>10h labeled</b>				
wav2vec 2.0 LV-60K (0.3B)	<b>5.7</b>	<b>9.2</b>	<b>5.6</b>	<b>9.4</b>
XLS-R (0.3B)	8.3	15.1	8.3	15.4
XLS-R (1B)	5.9	10.5	5.9	10.6

- Achieve competitive performance when model size is bigger

## Multilingual LibriSpeech

	#ft	en	de	nl	fr	es	it	pt	pl*	Avg.
Full labeled data (h)		44.7K	2K	1.6K	1.1K	918	247	161	104	
<i>Previous work</i>										
Pratap et al. (2020)	full	<b>5.9</b>	<b>6.5</b>	12.0	<b>5.6</b>	<b>6.1</b>	<b>10.5</b>	19.5	19.4	<b>10.7</b>
XLSR-53	10h	14.6	8.4	12.8	12.5	8.9	13.4	18.2	17.8	13.8
<i>This work</i>										
XLS-R (0.3B)	10h	15.9	9.0	13.5	12.4	8.1	13.1	17.0	13.9	12.8
XLS-R (1B)	10h	12.9	7.4	<b>11.6</b>	10.2	7.1	12.0	15.8	10.5	<b>10.9</b>
XLS-R (2B)	10h	14.0	7.6	11.8	10.0	6.9	12.1	<b>15.6</b>	<b>9.8</b>	11.0

Achieve competitive performance when only using 10 hour labeled data

# SSL Speech Representation Learning Models

CPC

HuBERT

wav2vec 2.0

WavLM

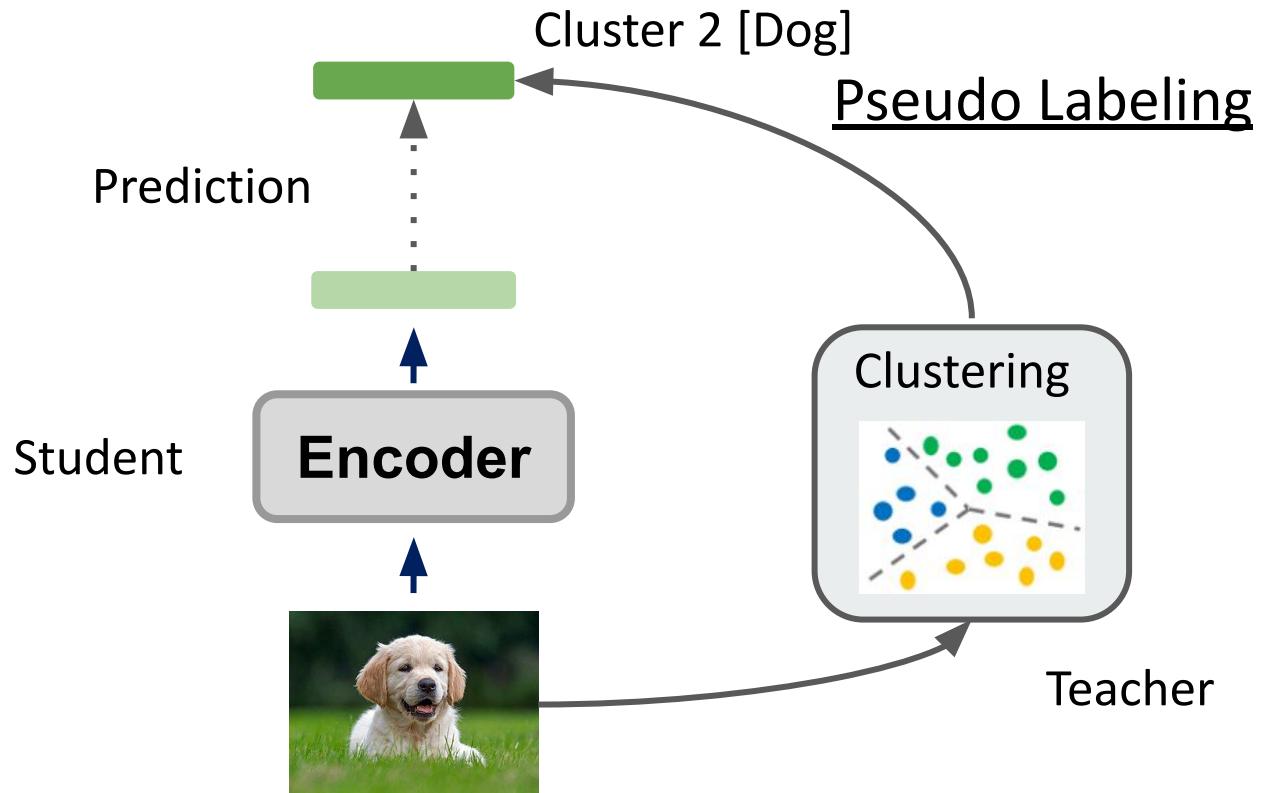
XLS-R

BEST-RQ

Contrastive Models

Predictive Models

# Predictive



# SSL Speech Representation Learning Models

CPC

**HuBERT**

wav2vec 2.0

WavLM

XLS-R

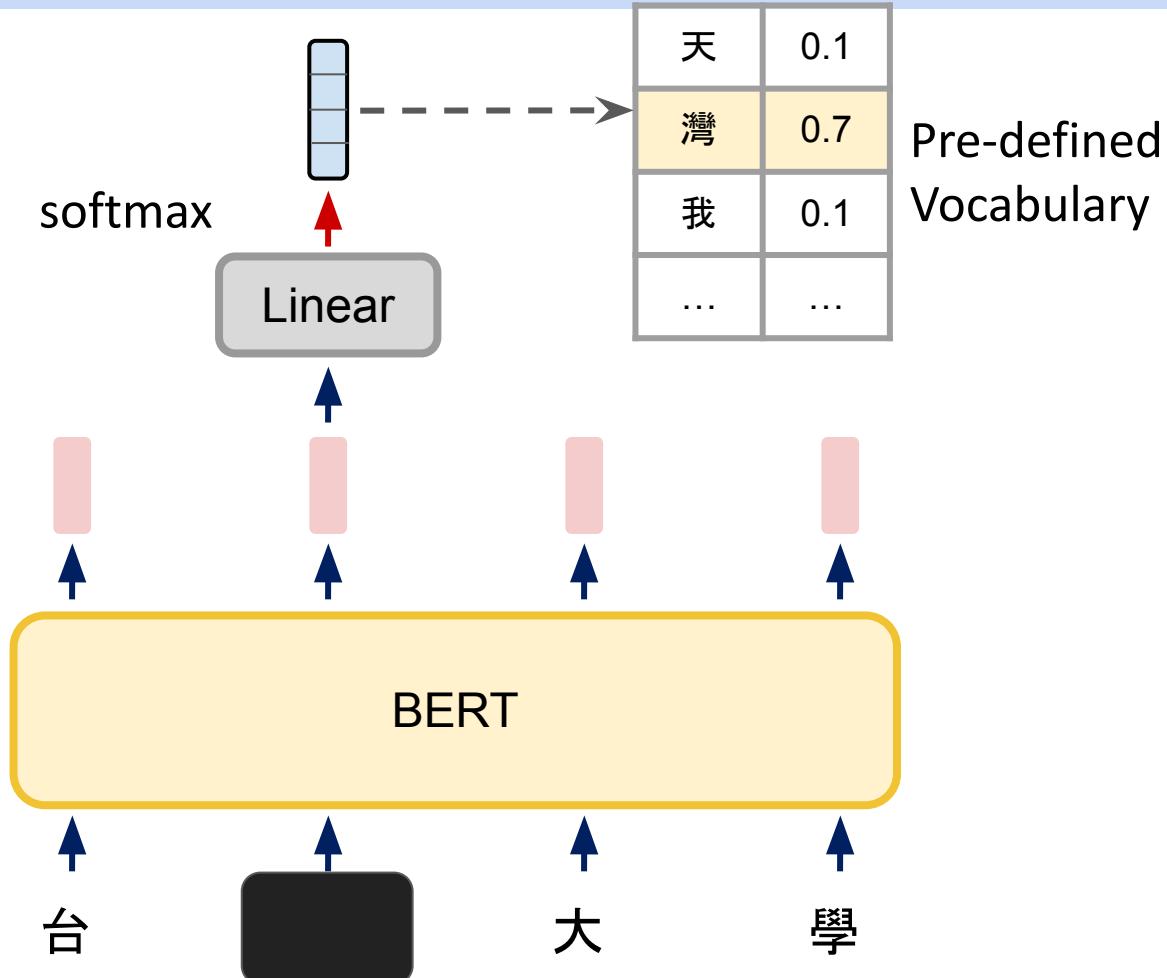
BEST-RQ

Contrastive Models

Predictive Models

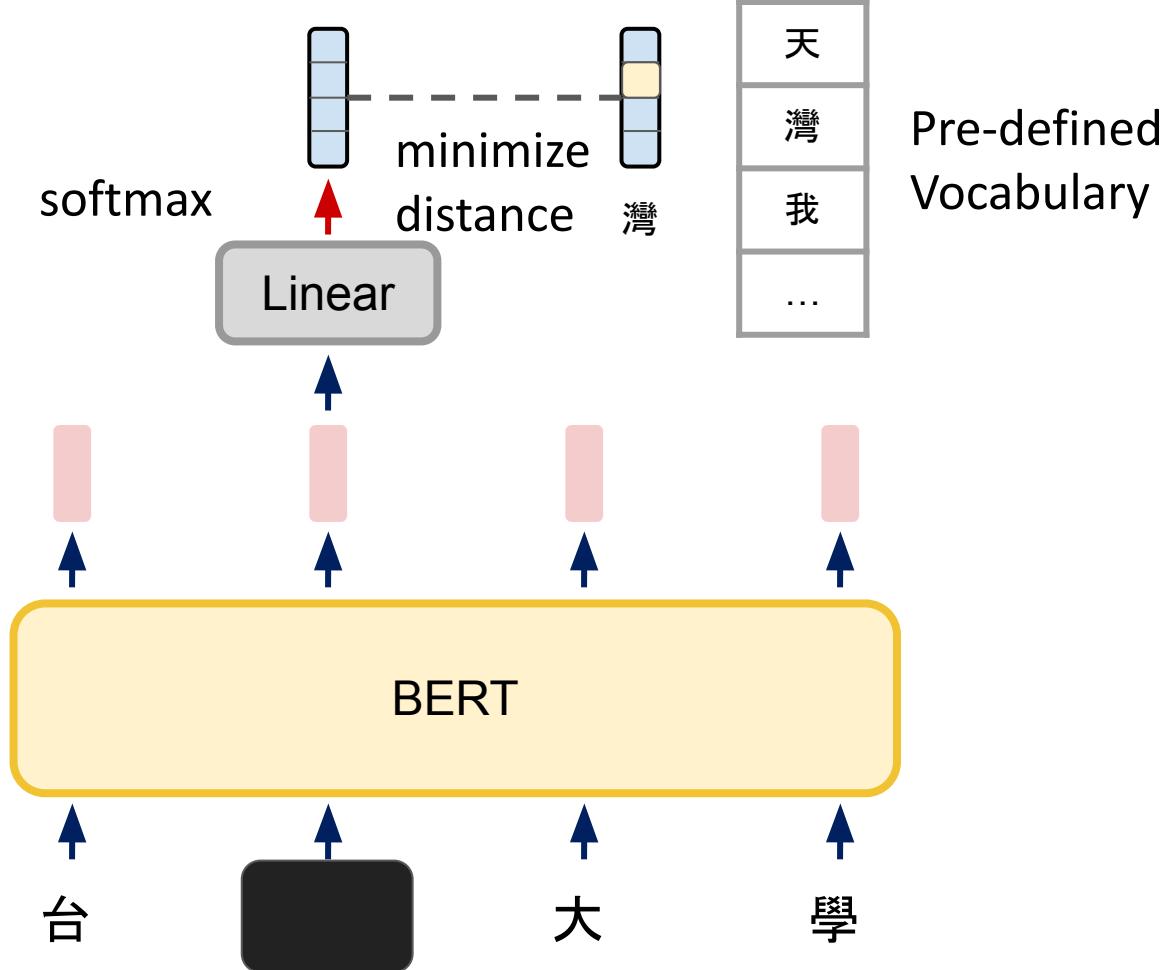
# BERT

Maksed Language  
Modeling

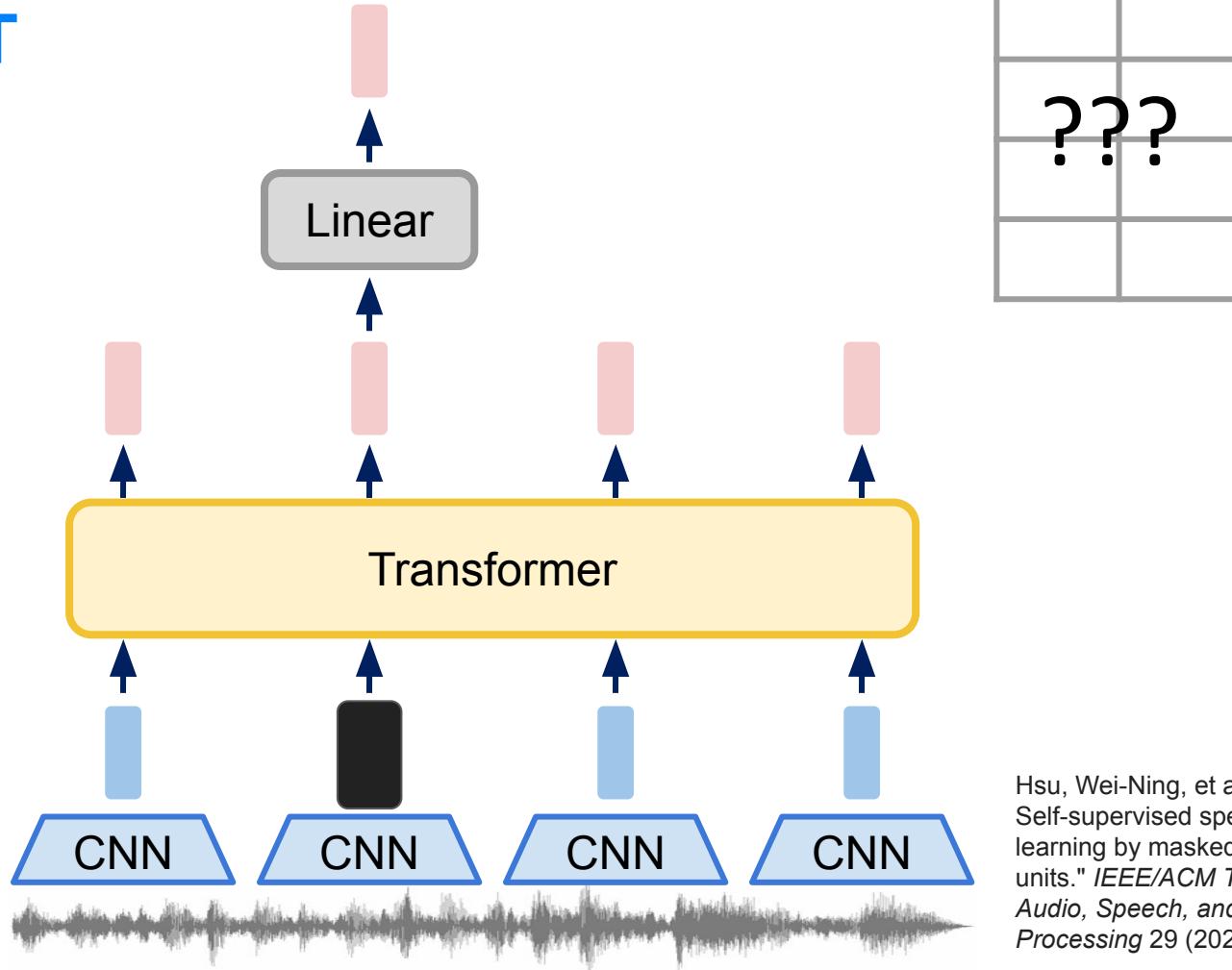


# BERT

Maksed Language  
Modeling

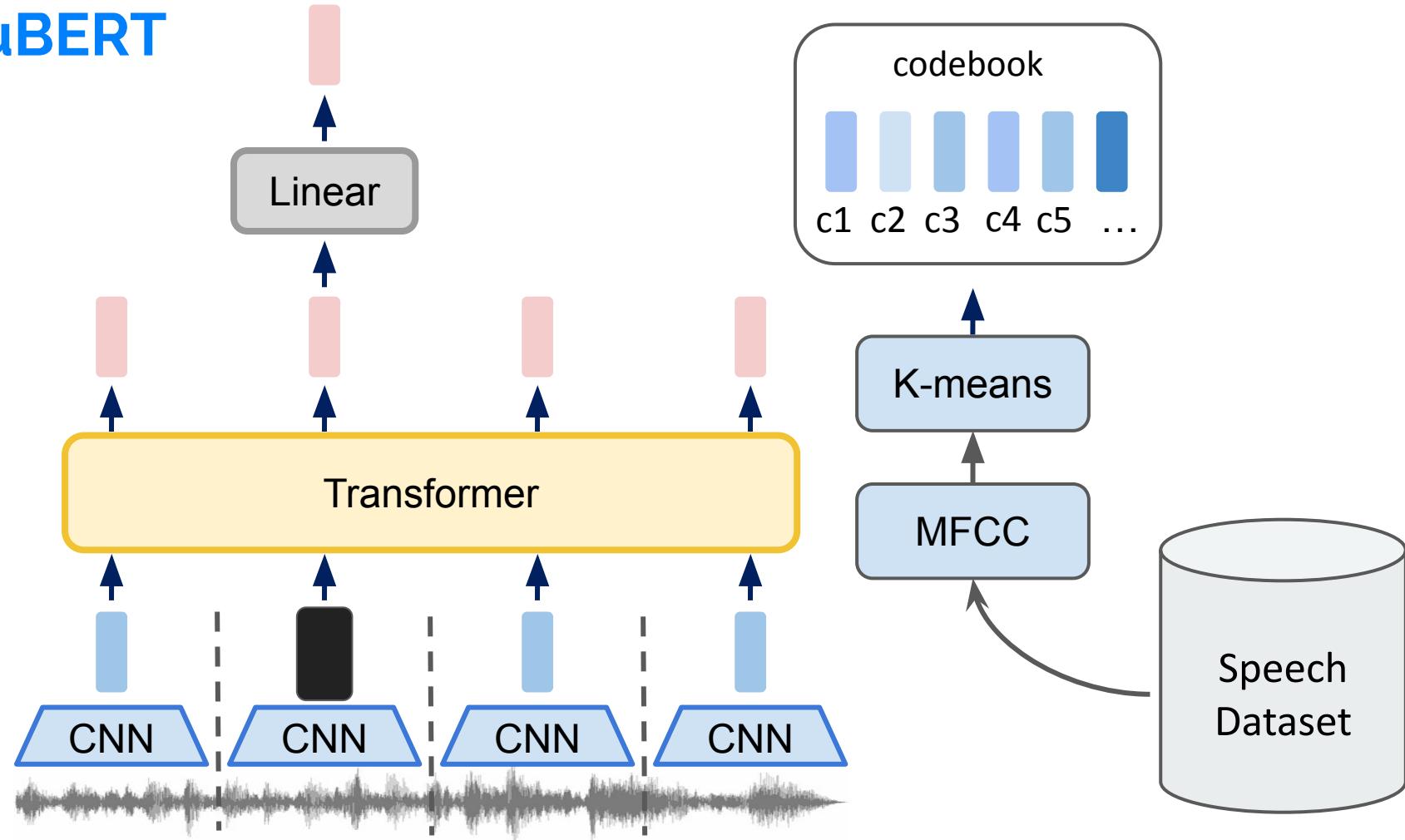


# HuBERT

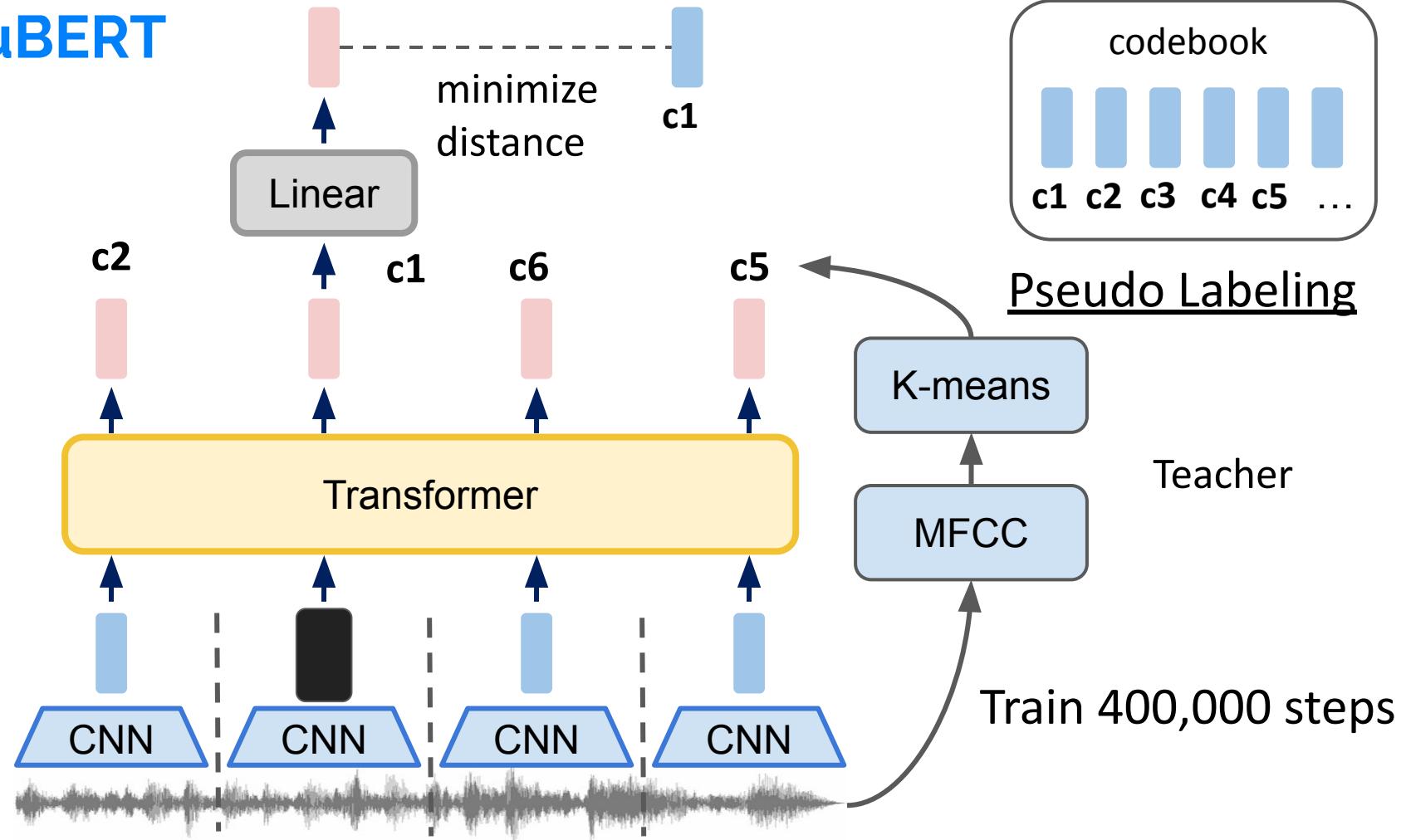


Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 3451-3460.

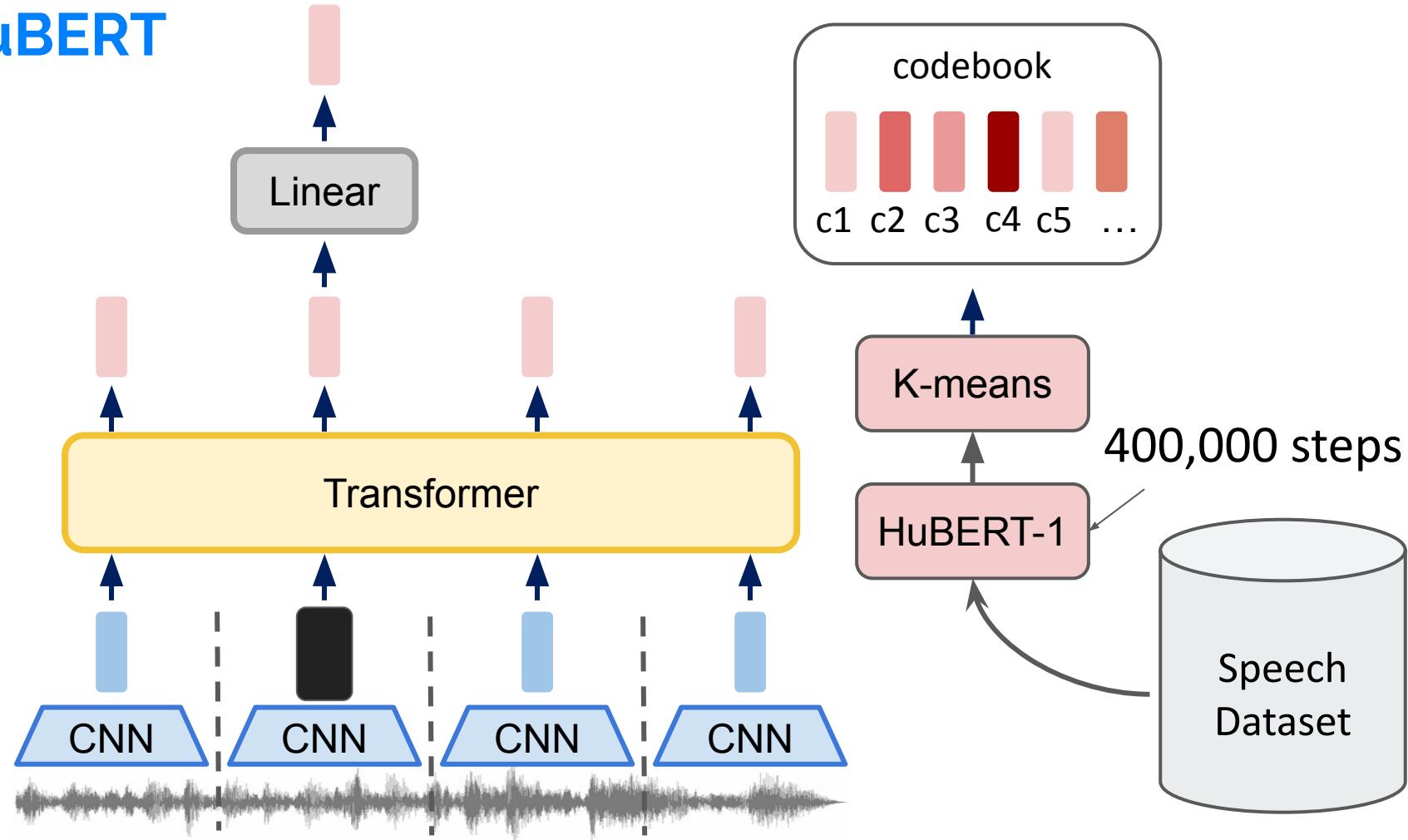
# HuBERT



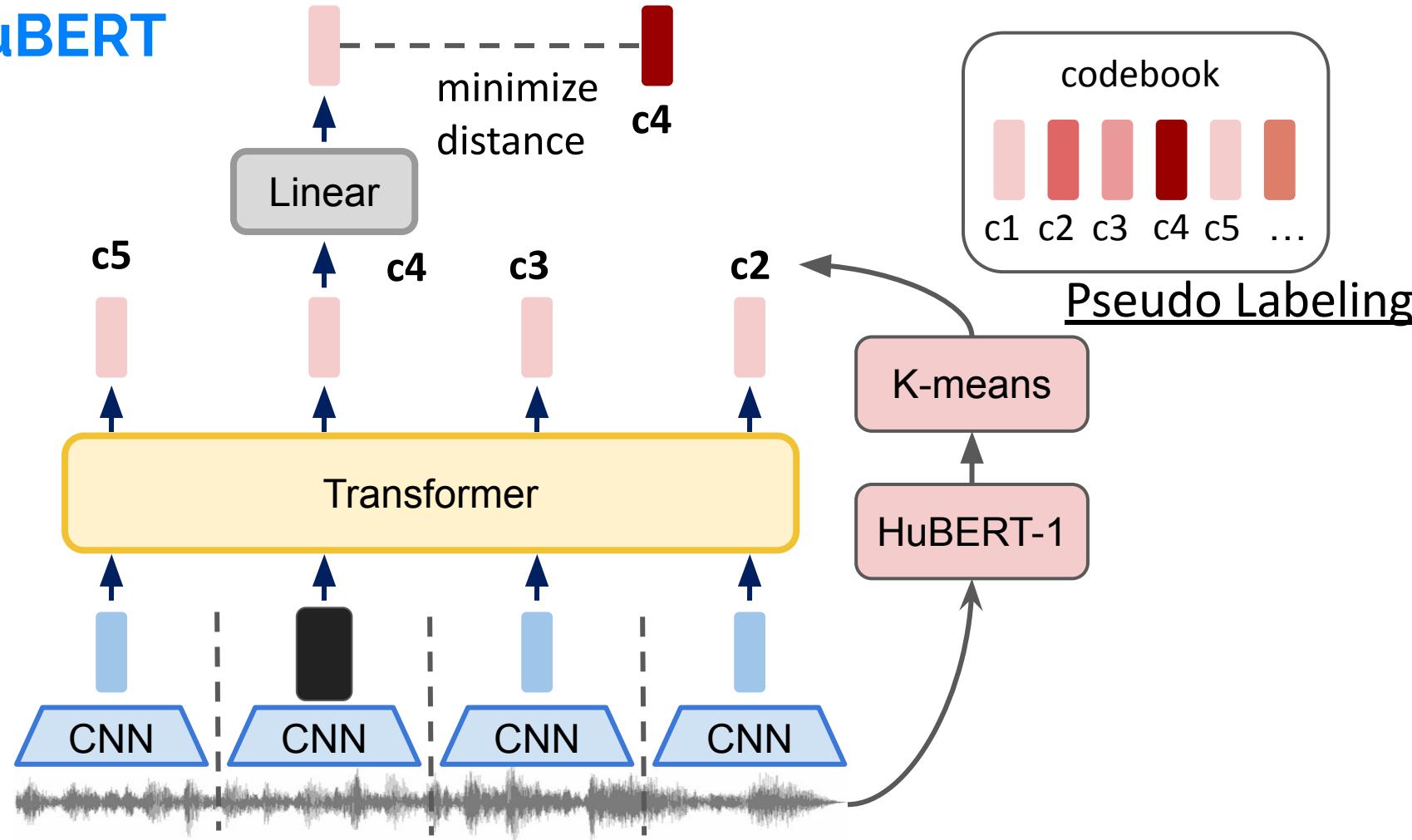
# HuBERT



# HuBERT



# HuBERT



# SSL Speech Representation Learning Models

CPC

HuBERT

wav2vec 2.0

**WavLM**

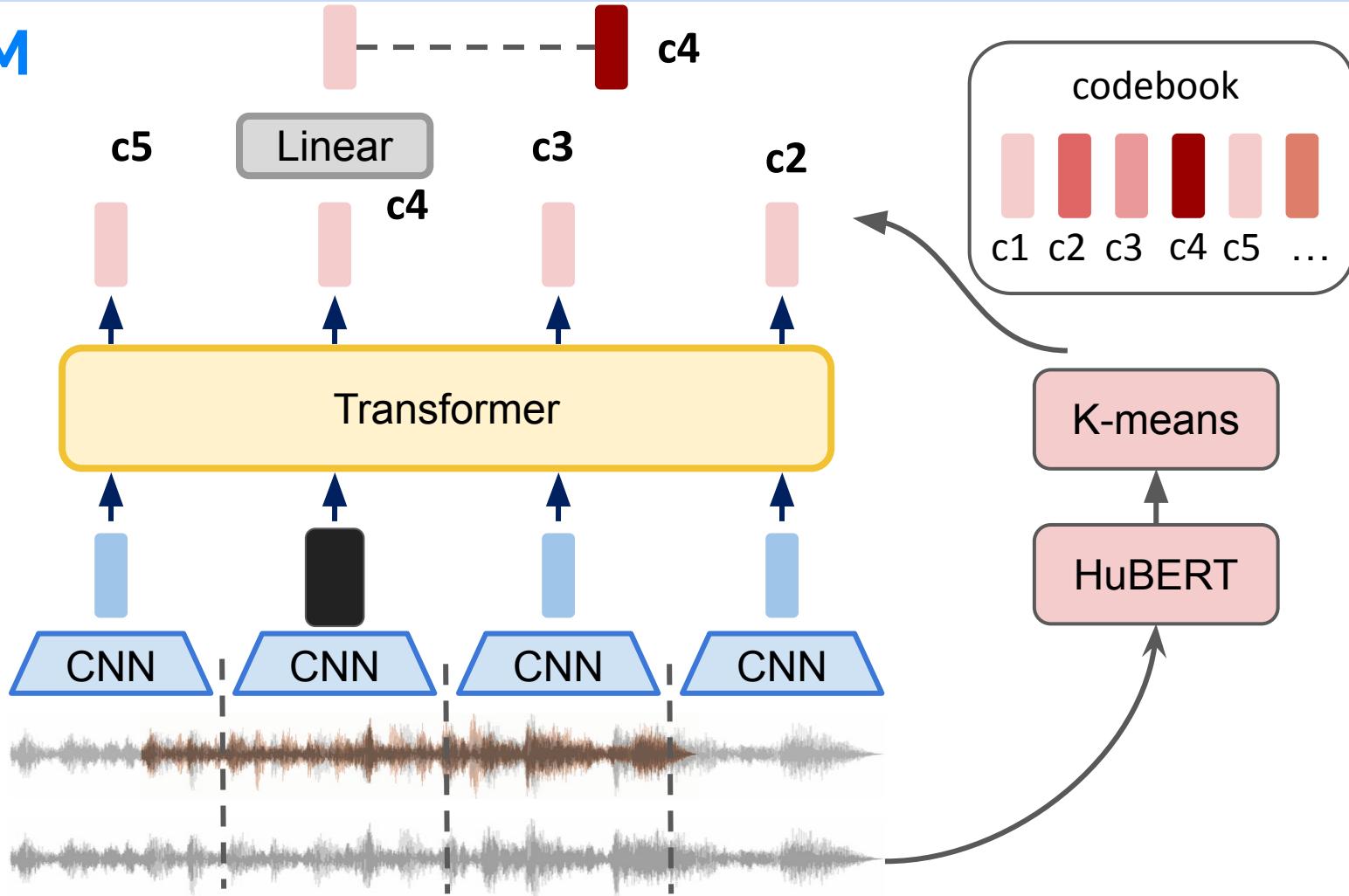
XLS-R

BEST-RQ

Contrastive Models

Predictive Models

# WavLM



## Speaker Verification

Feature	EER (%)		
	Vox1-O	Vox1-E	Vox1-H
ECAPA-TDNN [19]	1.010	1.240	2.320
ECAPA-TDNN (Ours)	1.080	1.200	2.127
HuBERT Base	0.989	1.068	2.216
HuBERT Large	0.808	0.822	1.678
WavLM Base+	0.84	0.928	1.758
WavLM Large	0.617	0.662	1.318
HuBERT Large*	0.585	0.654	1.342
WavLM Large*	<b>0.383</b>	<b>0.480</b>	<b>0.986</b>

EER: equal error rate, the lower the better

With speech denoising,  
WavLM performs better  
than HuBERT

# SSL Speech Representation Learning Models

CPC

HuBERT

wav2vec 2.0

WavLM

XLS-R

**BEST-RQ**

Contrastive Models

Predictive Models

# BERT-based Speech Pre-training with Random Projection Quantizer

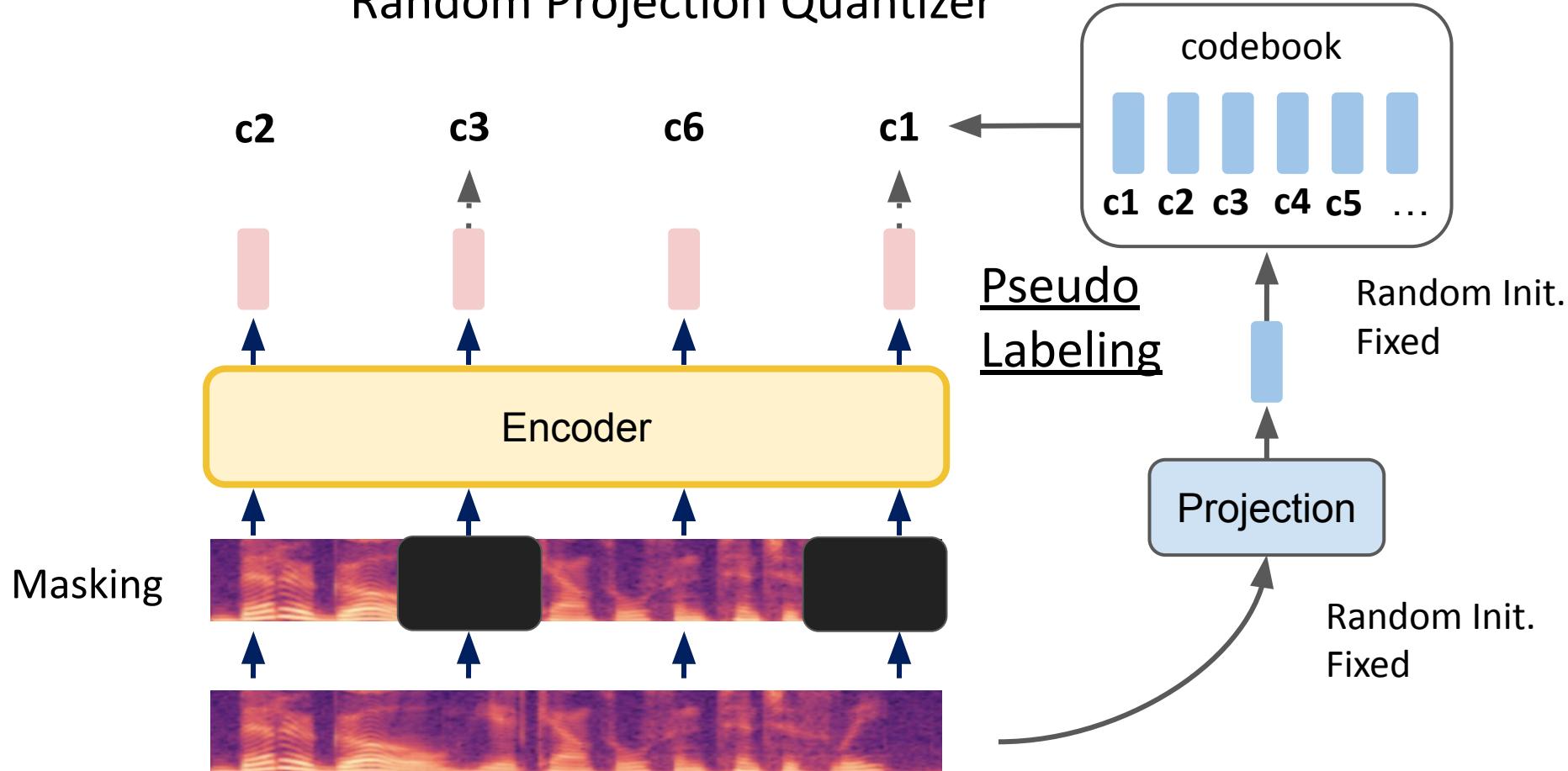


Table 1. LibriSpeech results with non-streaming models. The LM used in our experiment is a Transformer LM with model size 0.1B.

Method	Size (B)	No LM				With LM			
		dev	dev-other	test	test-other	dev	dev-other	test	test-other
wav2vec 2.0 (Baevski et al., 2020b)	0.3	2.1	4.5	2.2	4.5	1.6	3.0	1.8	3.3
HuBERT Large (Hsu et al., 2021)	0.3	—	—	—	—	1.5	3.0	1.9	3.3
HuBERT X-Large (Hsu et al., 2021)	1.0	—	—	—	—	1.5	<b>2.5</b>	1.8	2.9
w2v-Conformer XL (Zhang et al., 2020)	0.6	1.7	3.5	1.7	3.5	1.6	3.2	<b>1.5</b>	3.2
w2v-BERT XL (Chung et al., 2021)	0.6	<b>1.5</b>	2.9	<b>1.5</b>	<b>2.9</b>	<b>1.4</b>	2.8	<b>1.5</b>	2.8
BEST-RQ (Ours)	0.6	<b>1.5</b>	<b>2.8</b>	1.6	<b>2.9</b>	<b>1.4</b>	2.6	<b>1.5</b>	<b>2.7</b>

- Comparable to other Speech SSL models
- Teacher (clusters) is not necessary to be good

# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. Regeneration Framework

### Part 3

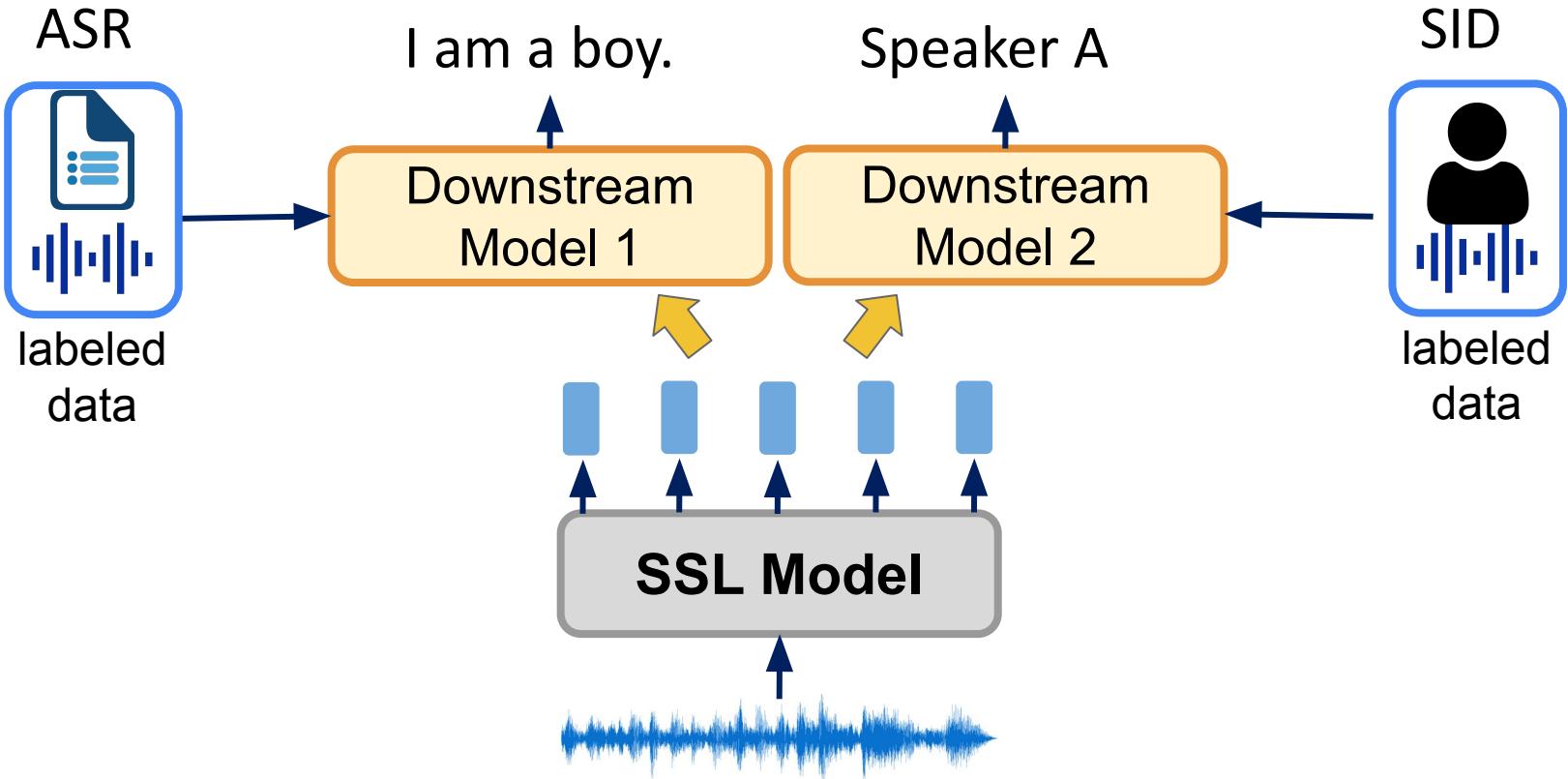
#### Other Speech Foundation Models

1. Whisper
2. USM

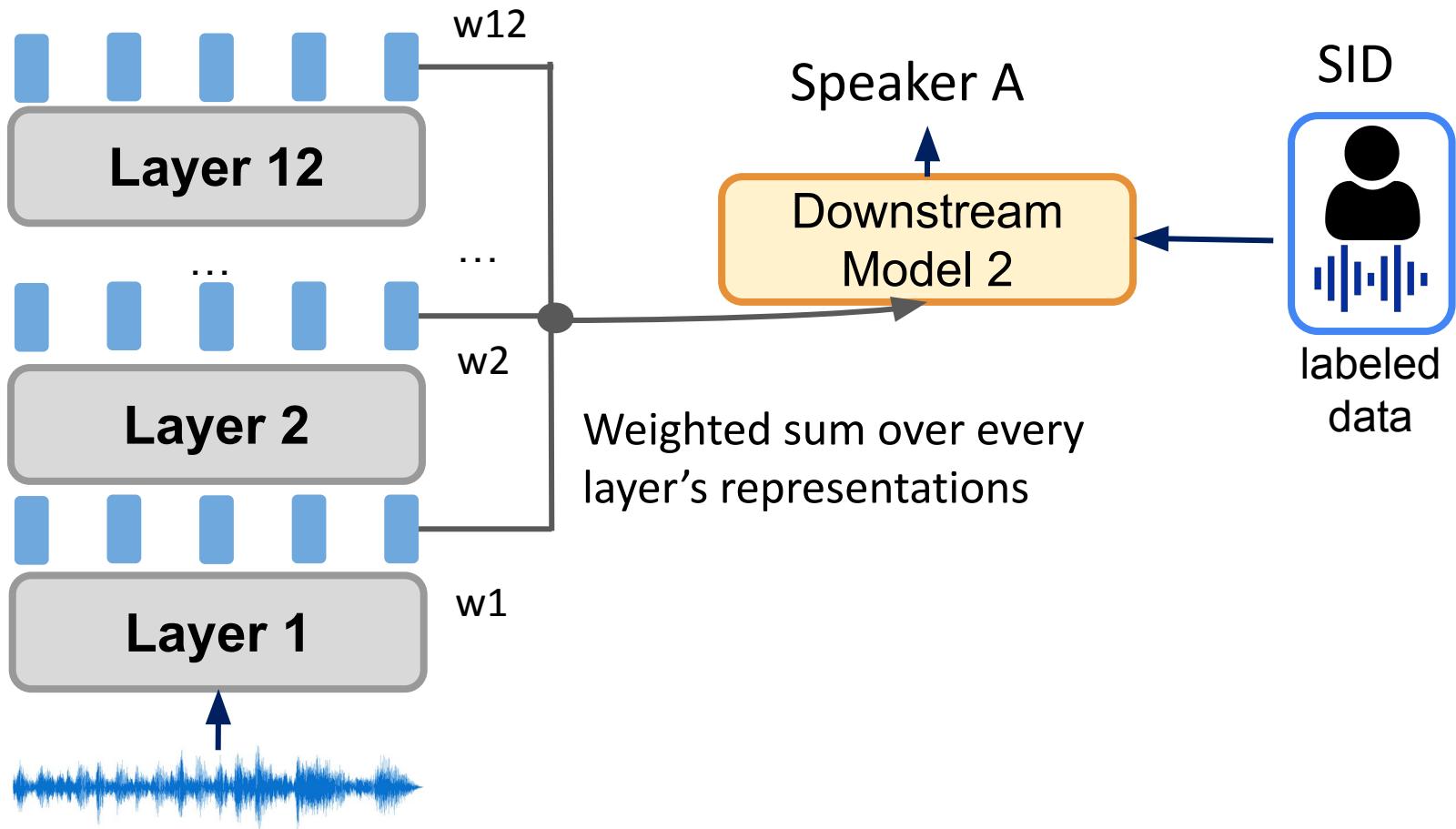
# SUPERB



# SUPERB



# SUPERB





Query By Example

Speaker  
Diarization

Speech  
Recognition

Speaker  
Identification

Phoneme  
Recognition

Intent  
Classification

Speaker  
Verification

Keyword  
Spotting

Slot Filling

Emotion  
Recognition

**Speaker**

**Content**

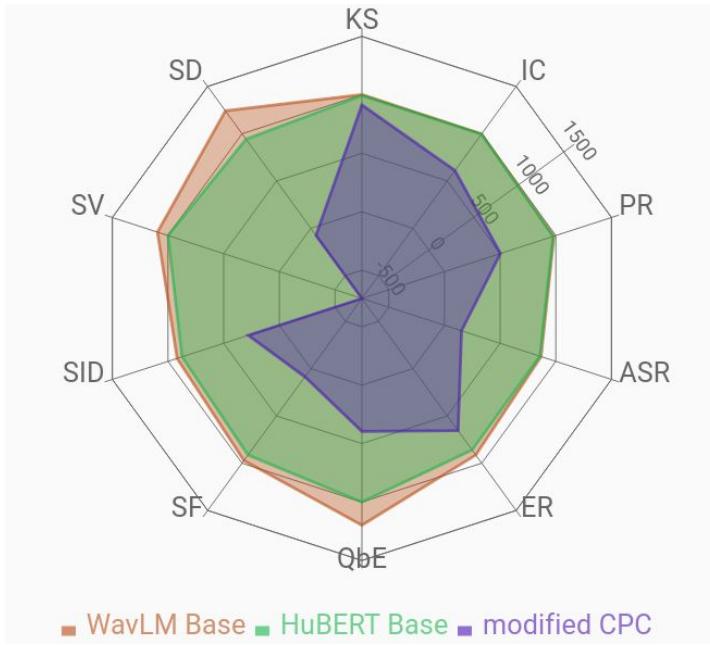
**Semantics**

**Paralinguistics**

Method	Name	Description	URL	Params ↓	MACs ↓	(1) ↓	(2) ↓	(3) ↓	(4) ↓	Rank ↑	Score ↑	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑
WavLM Large	Microsoft	M-P + VQ +...	🔗	3.166e+8	4.326e+12	3.8...	6.7...	1.0...	2.1...	25.8	1145	97.86	99.31	3.06	3.44	70.62
WavLM Base+	Microsoft	M-P + VQ +...	🔗	9.470e+7	1.670e+12	1.4...	2.6...	4.2...	8.3...	24.05	1106	97.37	99	3.92	5.59	68.65
WavLM Base	Microsoft	M-P + VQ +...	🔗	9.470e+7	1.670e+12	1.4...	2.6...	4.2...	8.3...	20.95	1019	96.79	98.63	4.84	6.21	65.94
data2vec Large	CI Tang	Masked Ge...	🔗	3.143e+8	4.306e+12	3.8...	6.7...	1.0...	2.1...	20.8	949	96.75	98.31	3.6	3.36	66.31
LightHuBERT Stag...	LightHuBE...	Once-for-AI...	🔗	9.500e+7	-	-	-	-	-	20.1	959	96.82	98.5	4.15	5.71	66.25
HuBERT Large	paper	M-P + VQ	🔗	3.166e+8	4.324e+12	3.8...	6.7...	1.0...	2.1...	19.15	919	95.29	98.76	3.53	3.62	67.62
data2vec-aqc Base	Speech La...	Masked Ge...	🔗	9.384e+7	1.657e+12	1.4...	2.5...	4.1...	8.3...	19.05	935	96.36	98.92	4.11	5.39	67.59
CoBERT Base	ByteDance ...	Code Repr...	🔗	9.435e+7	1.660e+12	1.4...	2.5...	4.1...	8.3...	18	894	96.36	98.87	3.08	4.74	65.32
HuBERT Base	paper	M-P + VQ	🔗	9.470e+7	1.669e+12	1.4...	2.6...	4.2...	8.3...	17.75	941	96.3	98.34	5.41	6.42	64.92
wav2vec 2.0 Large	paper	M-C + VQ	🔗	3.174e+8	4.326e+12	3.8...	6.7...	1.0...	2.1...	17.7	914	96.66	95.28	4.75	3.75	65.64
ccc-wav2vec 2.0 B...	Speech La...	M-C + VQ	🔗	9.504e+7	1.670e+12	1.4...	2.6...	4.2...	8.3...	17.45	940	96.72	96.47	5.95	6.3	64.17
data2vec base	CI Tang	Masked Ge...	🔗	9.375e+7	1.657e+12	1.4...	2.5...	4.1...	8.3...	16.85	884	96.56	97.63	4.69	4.94	66.27
LightHuBERT Small	LightHuBE...	Once-for-AI...	🔗	2.700e+7	8.607e+11	7.7...	1.3...	2.1...	4.3...	15.45	901	96.07	98.23	6.6	8.34	64.12
FaST-VGS+	Puyuan Pe...	FaST-VGS I...	-	2.172e+8	-	-	-	-	-	14.7	809	97.27	98.97	7.76	8.83	62.71
wav2vec 2.0 Base	paper	M-C + VQ	🔗	9.504e+7	1.669e+12	1.4...	2.6...	4.2...	8.3...	13.2	818	96.23	92.35	5.74	6.43	63.43
DistilHuBERT	Heng-Jui C...	multi-task I...	-	2.349e+7	7.859e+11	7.2...	1.2...	2.0...	3.8...	11.8	717	95.98	94.99	16.27	13.37	63.02
DeCoAR 2.0	paper	M-G + VQ	🔗	8.984e+7	1.114e+12	9.7...	1.7...	2.7...	5.6...	11.1	722	94.48	90.8	14.93	13.02	62.47
wav2vec	paper	F-C	🔗	3.254e+7	1.086e+12	1.0...	1.7...	2.7...	5.2...	8.9	529	95.59	84.92	31.58	15.86	59.79

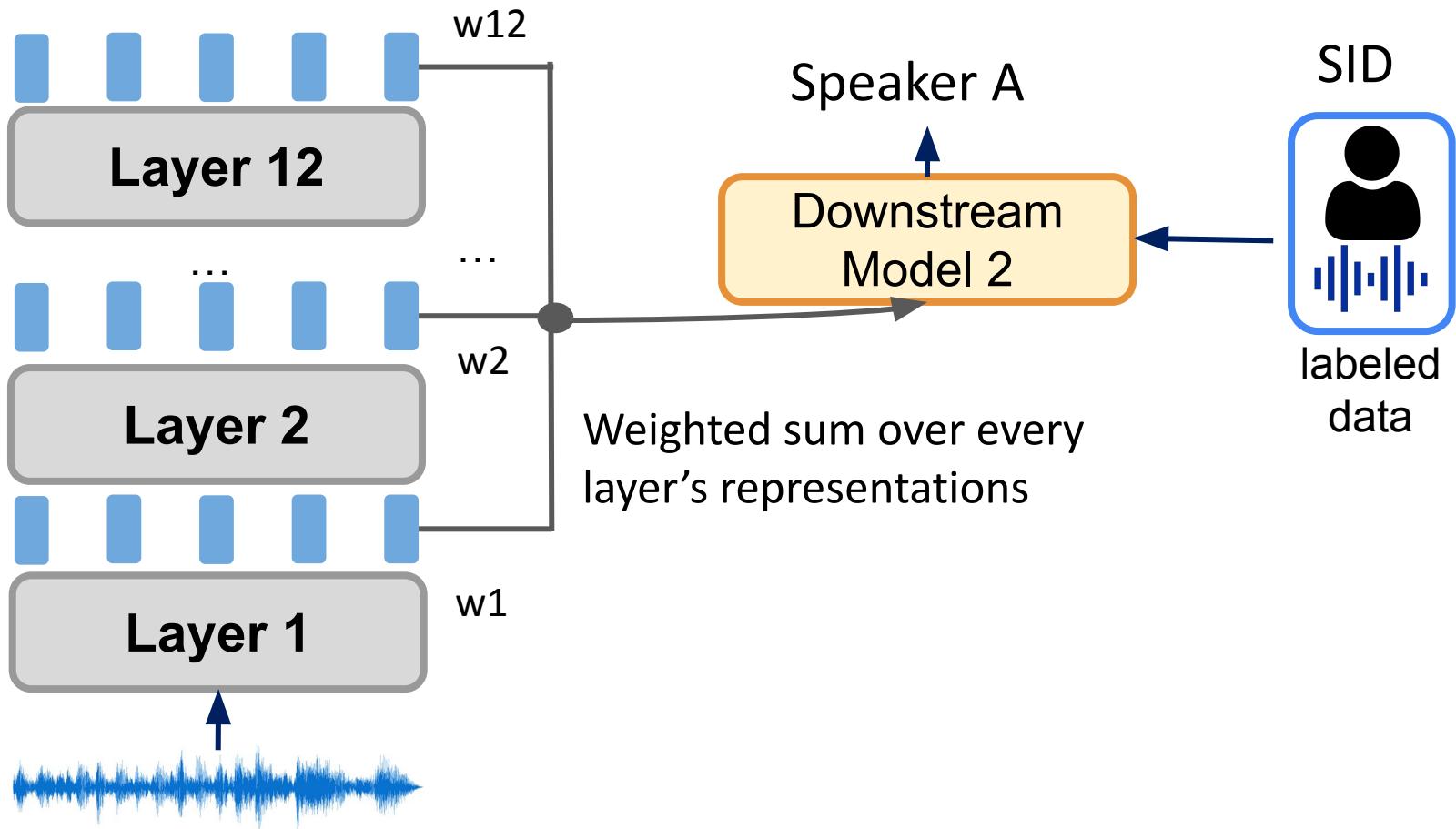


The bigger, the better



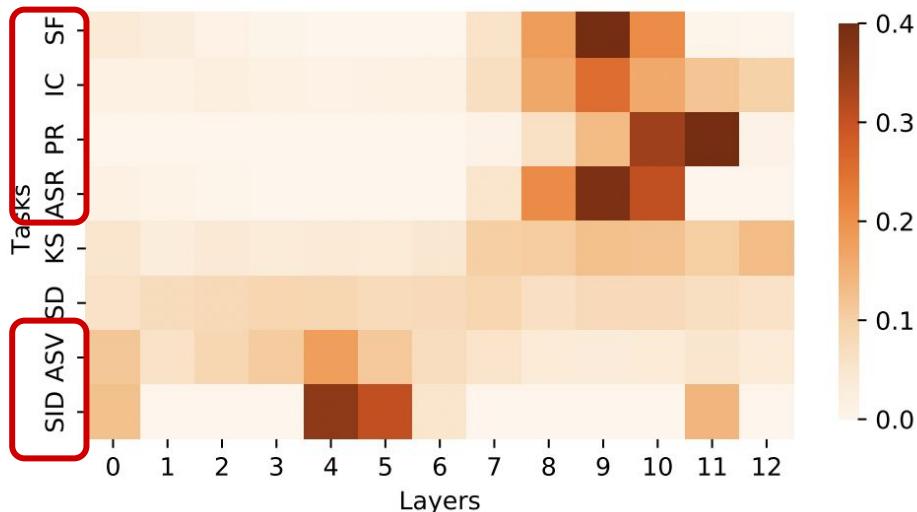
Strong model performs better on all kinds of tasks

# SUPERB



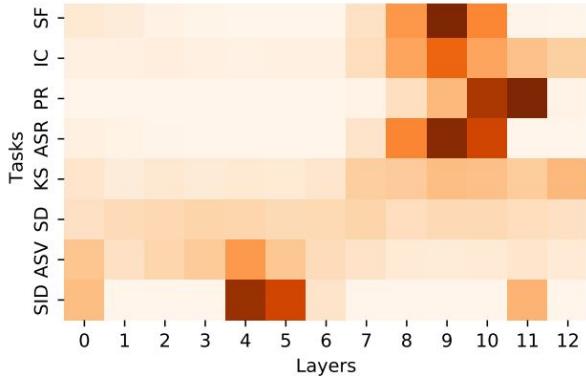
# Representation Weight Analysis

- **Speaker tasks:**
  - ASV (Speaker Verification)
  - SID (Speaker Identification)
  - ...
- **Content tasks:**
  - ASR (Speech Recognition)
  - IC (Intent Classification)

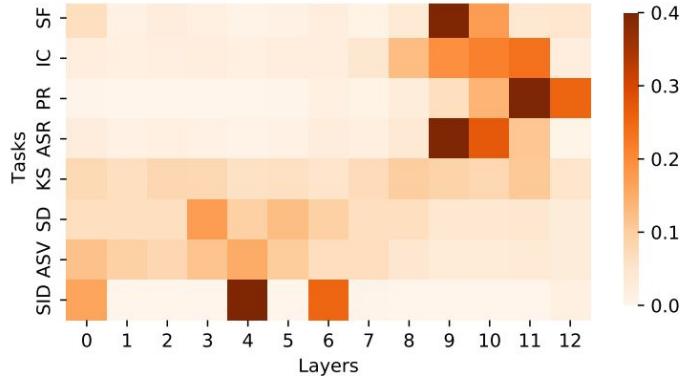


(a) HuBERT Base

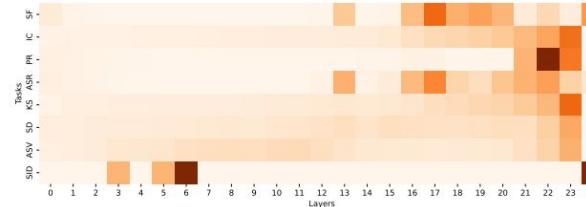
# Representation Weight Analysis



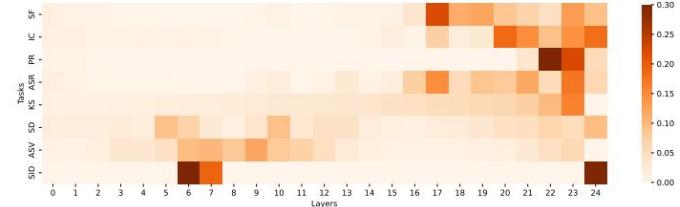
(a) HuBERT Base



(b) WavLM Base+



(c) HuBERT Large



(d) WavLM Large

# Part 1 Summary: Speech Representation Learning

Predict future vectors	CPC	HuBERT	Masked prediction (BERT)
Quantize, masking	wav2vec 2.0	WavLM	Denoising
Multilingual	XLS-R	BEST-RQ	Random Projection during pseudo labeling

**Contrastive Models      Predictive Models**

# Part 1 Summary: Speech Representation Learning



Contrastive Models      Predictive Models



Speaker  
Content  
Semantics  
Paralinguistics  
...

- SSL Speech Models are versatile!
- Different information is encoded in different layer's representation

# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. VALL-E

### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. VALL-E

### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

[Research](#)[Blog](#)[Resources](#)[About](#)

RESEARCH | NLP

# Textless NLP: Generating expressive speech from raw audio

September 9, 2021

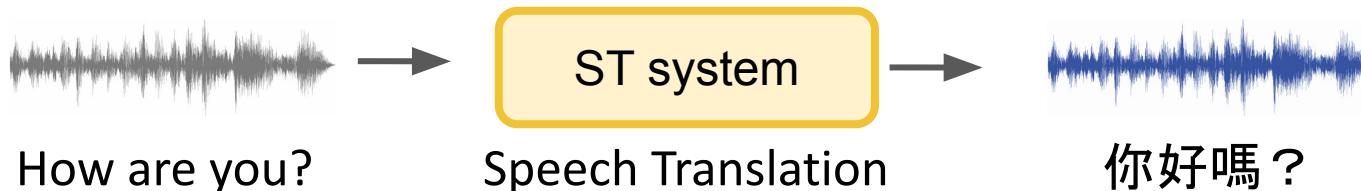
→ Share on Facebook

→ Share on Twitter

## Our Work



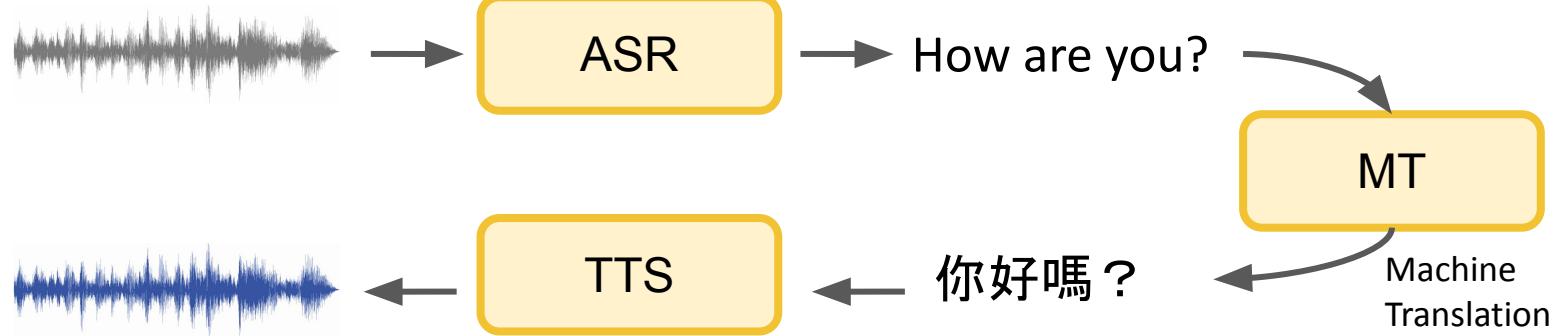
# Textless NLP Project



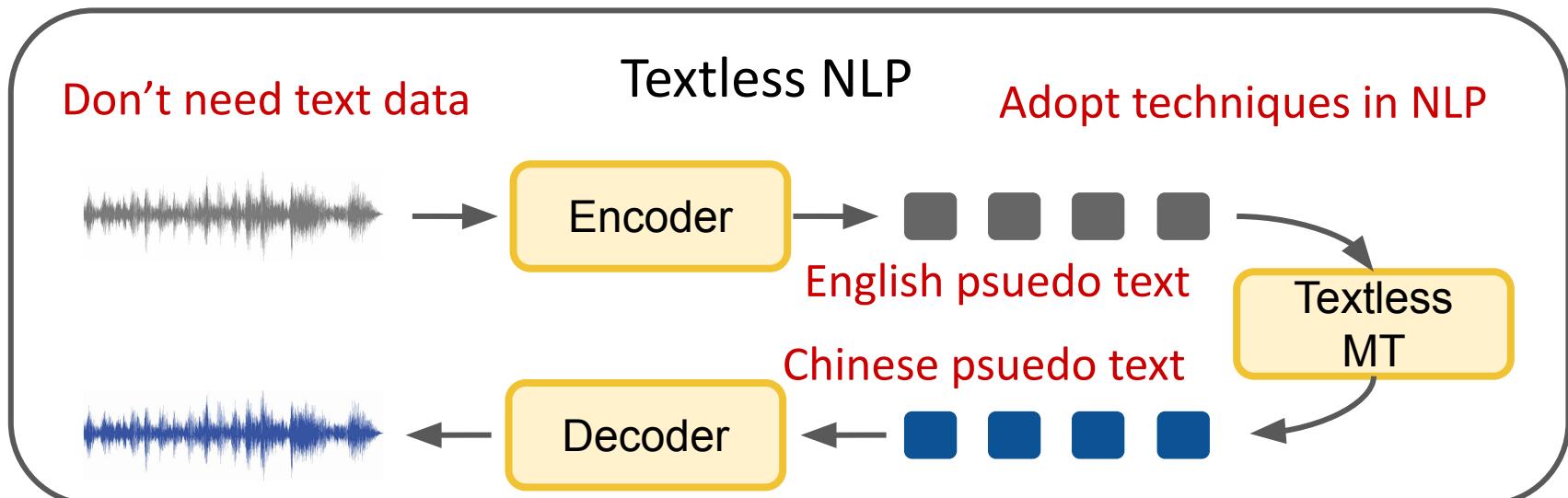
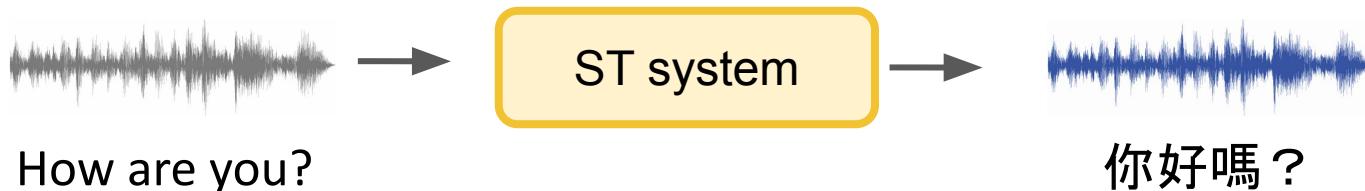
Need Text Data!

Cascaded System

Error Propagation!

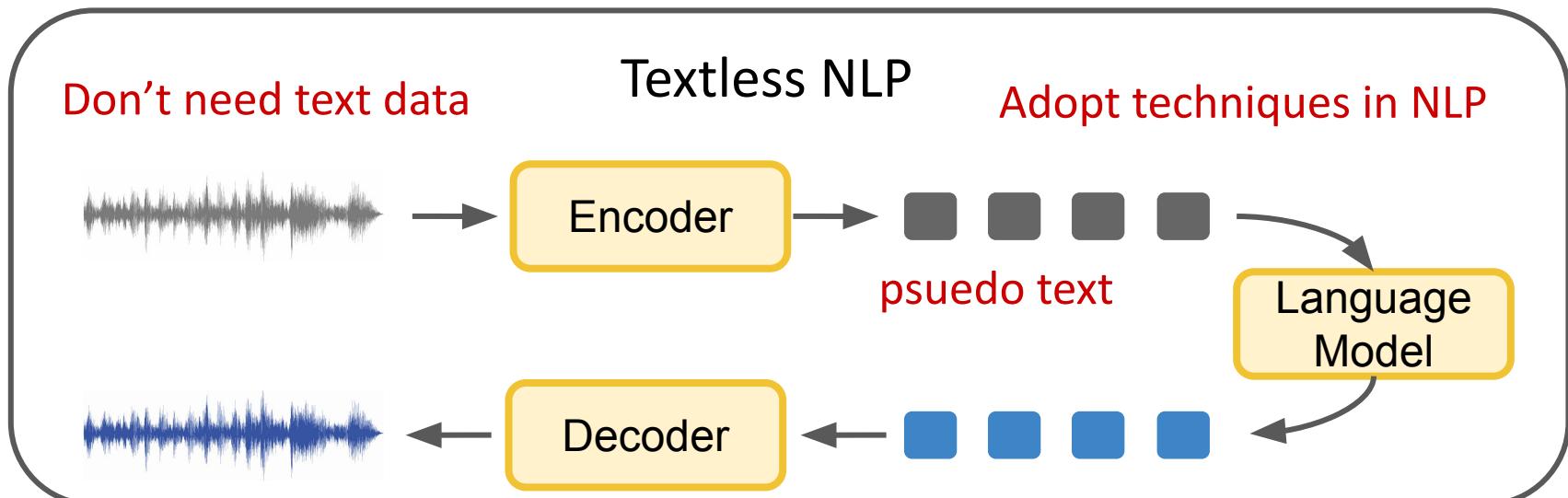


# Textless NLP Project

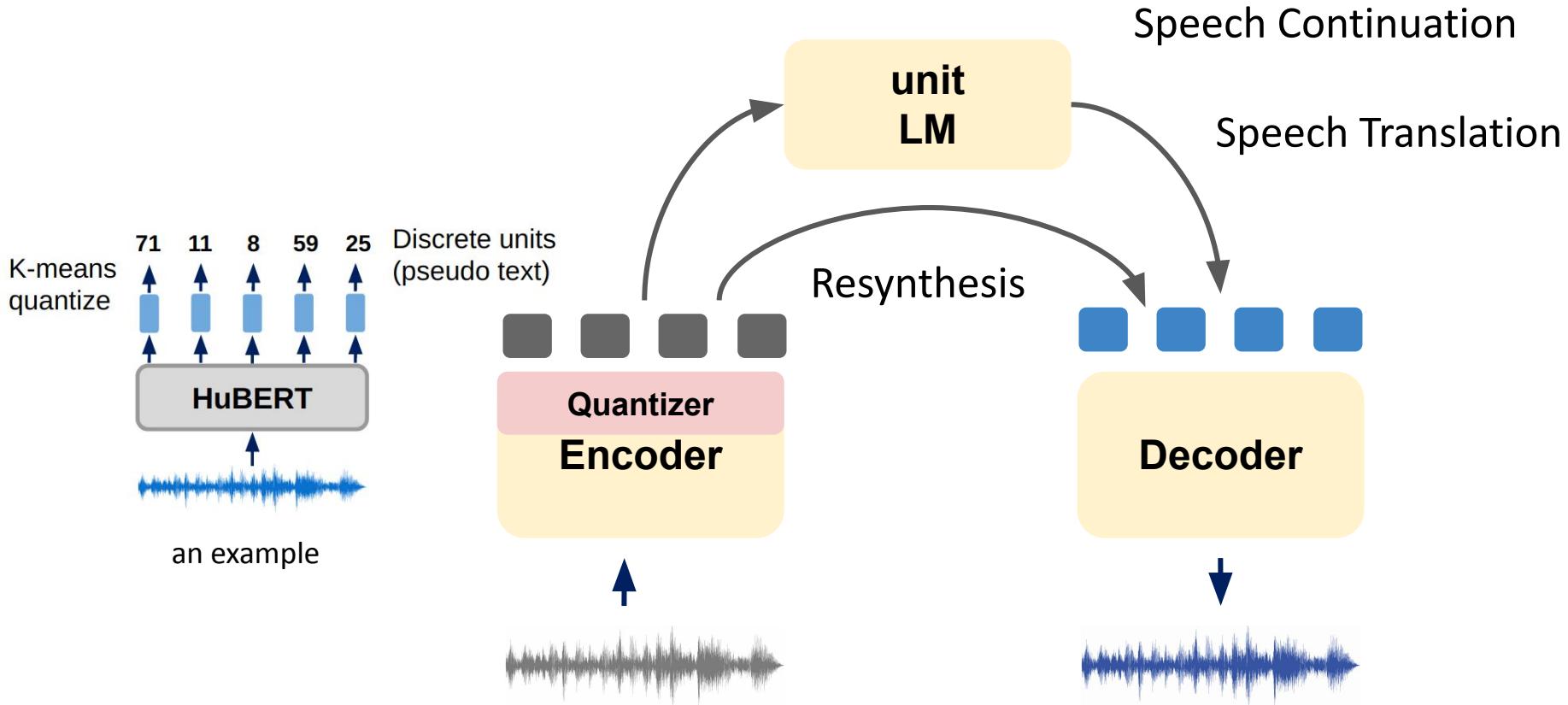


# Textless NLP Project

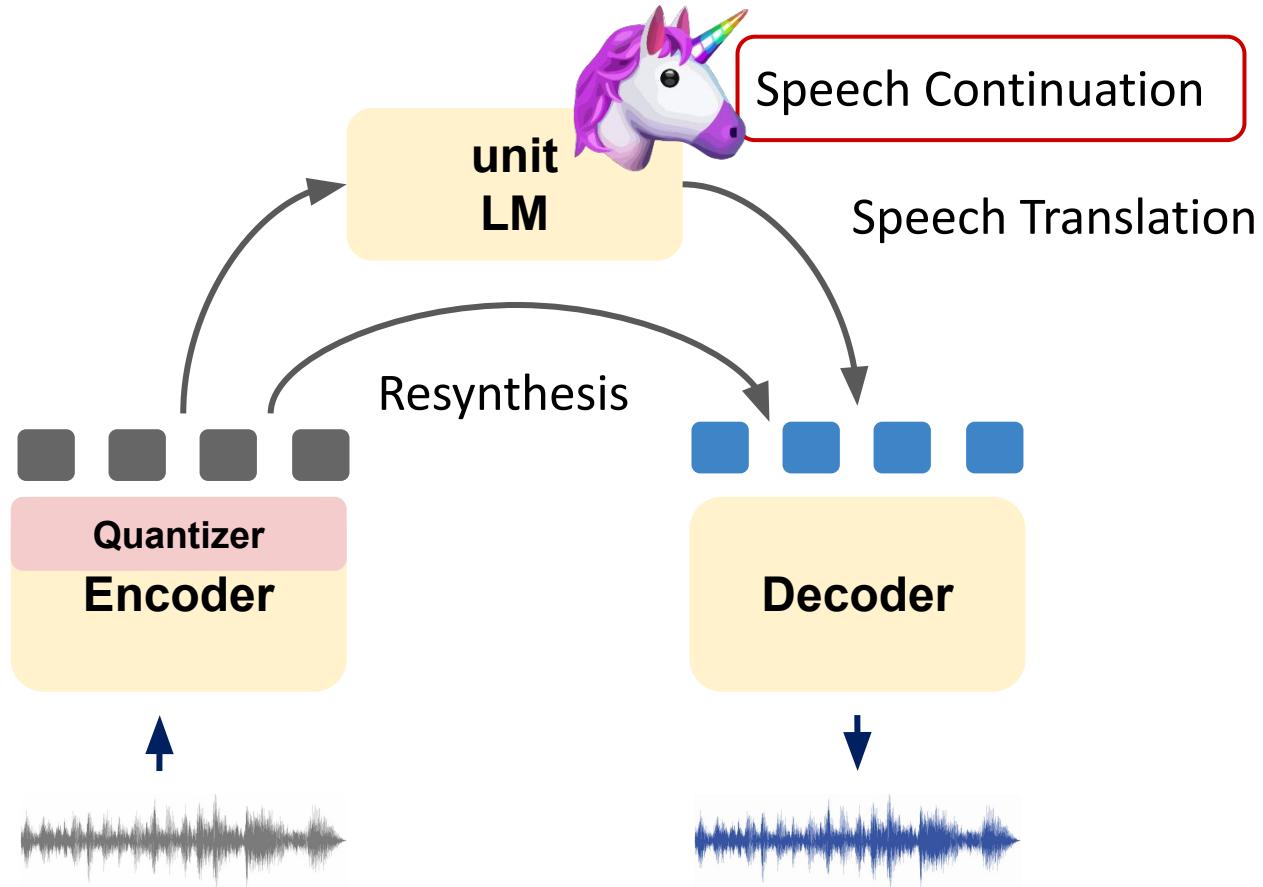
- GPT: Speech Continuation
- BART: Speech Translation
- no LM: Speech Resynthesis



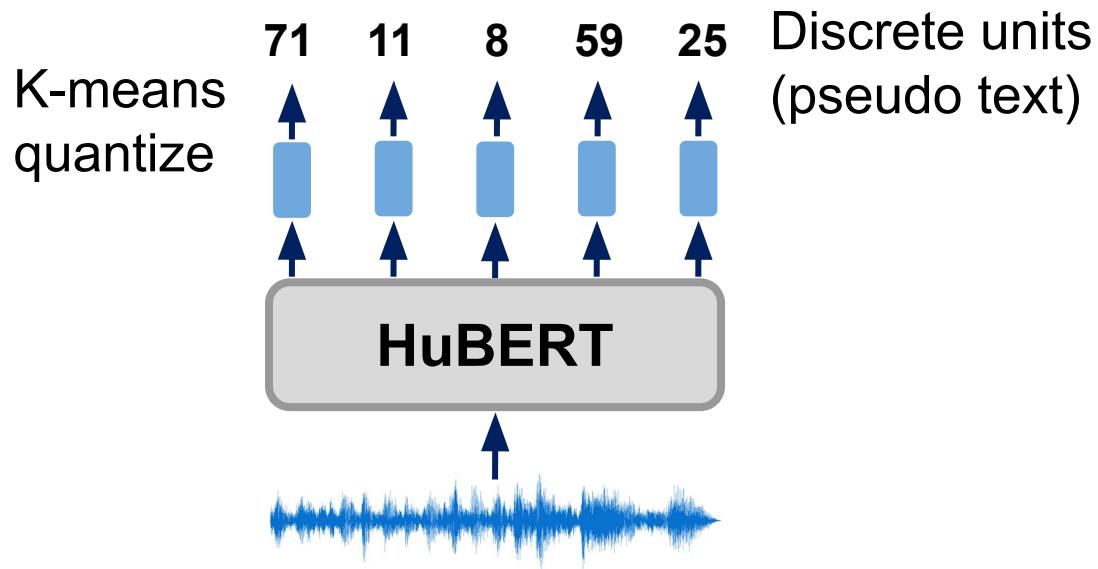
# Textless NLP Project



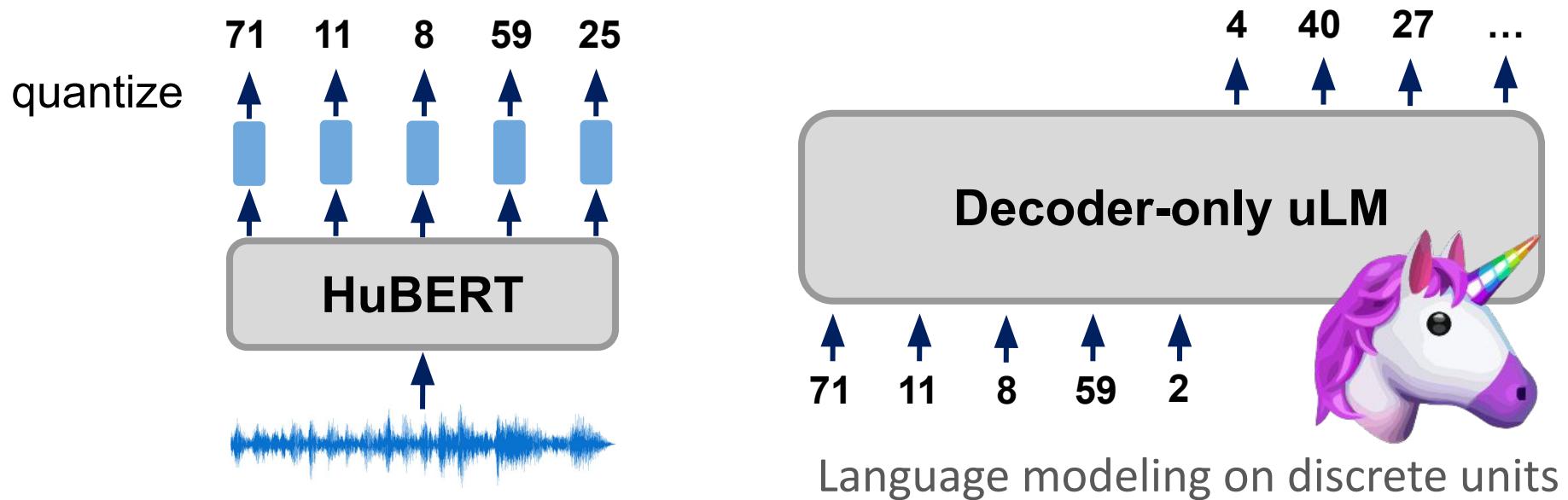
# Textless NLP Project



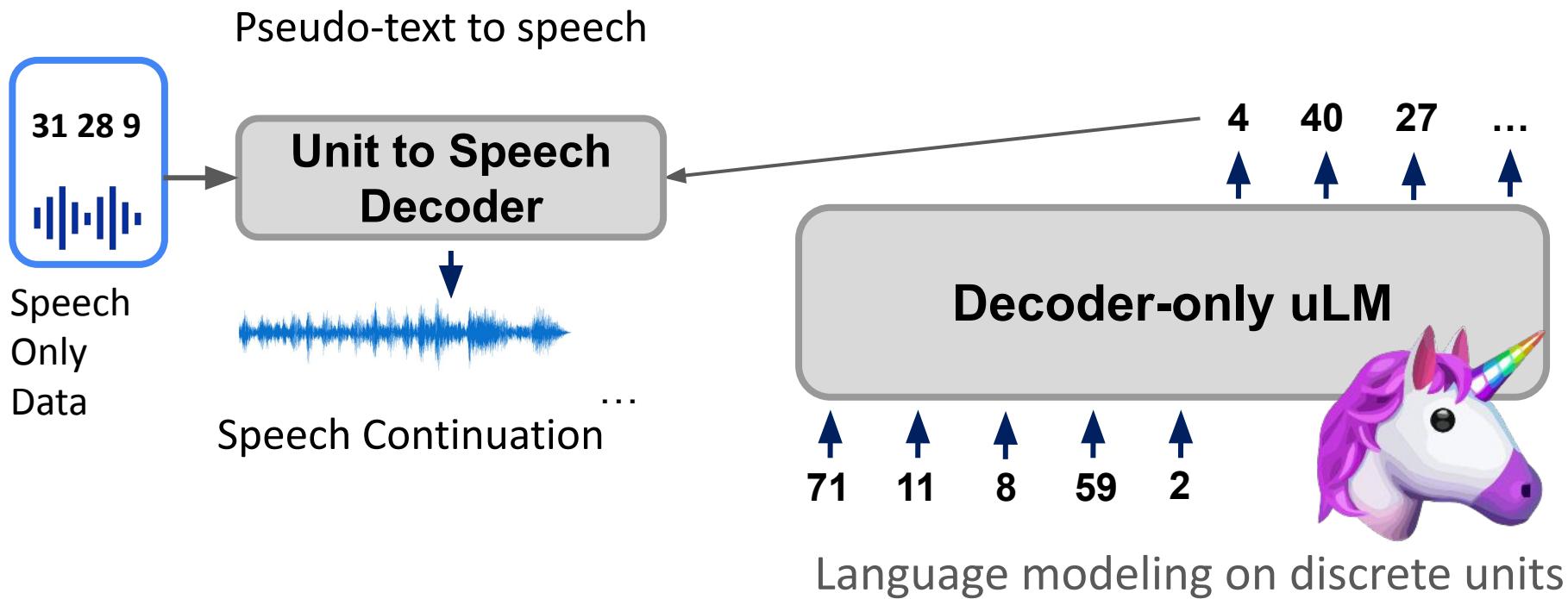
# Generative Spoken Language Modeling



# Generative Spoken Language Modeling

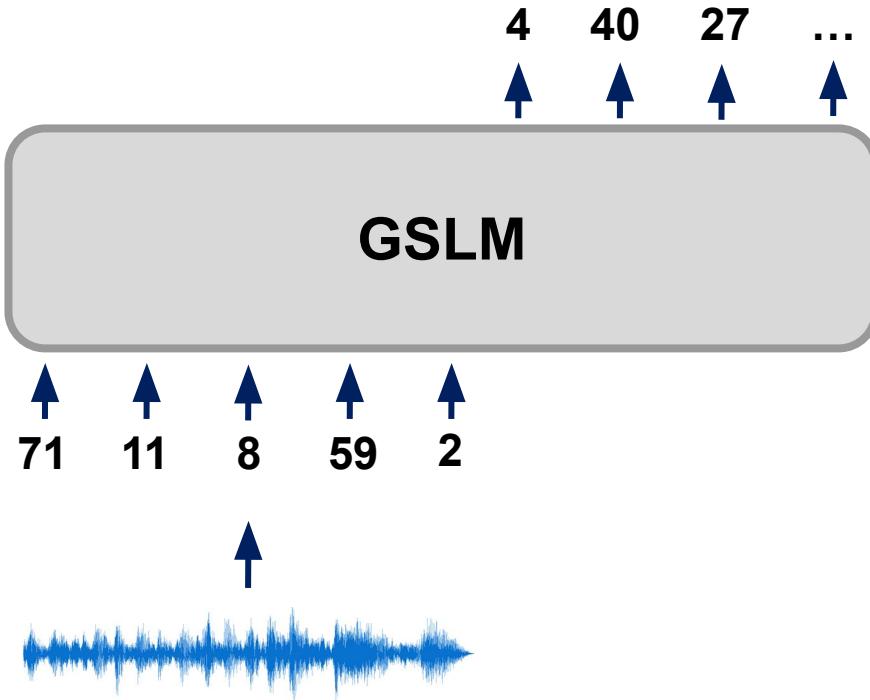


# Generative Spoken Language Modeling

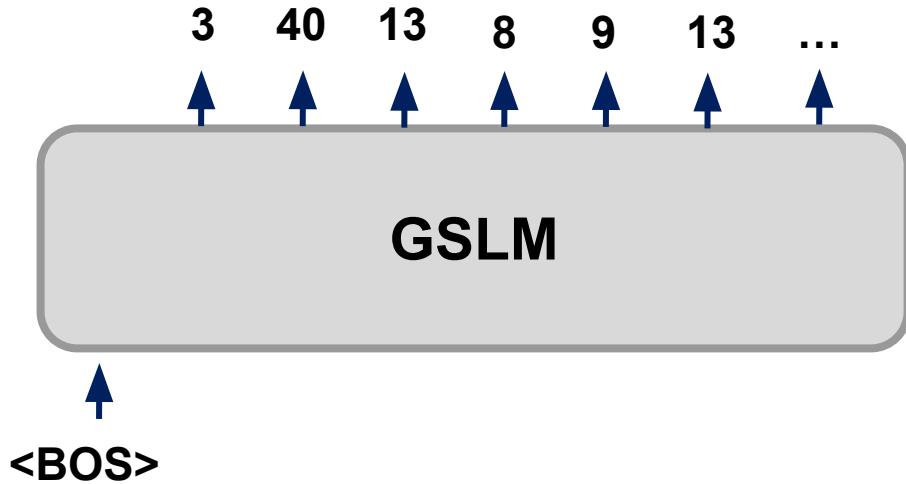


# Generative Spoken Language Modeling

Conditional Generation



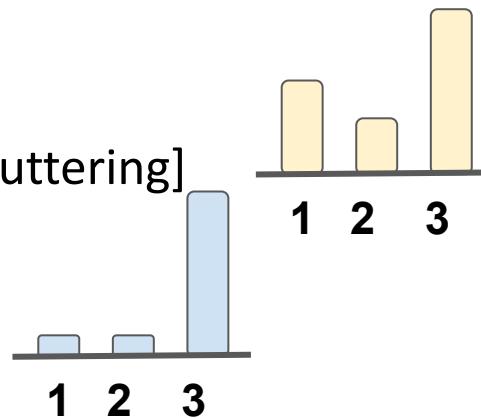
Unconditional Generation



# Generative Spoken Language Modeling

temperature = 0.3 [stuttering]

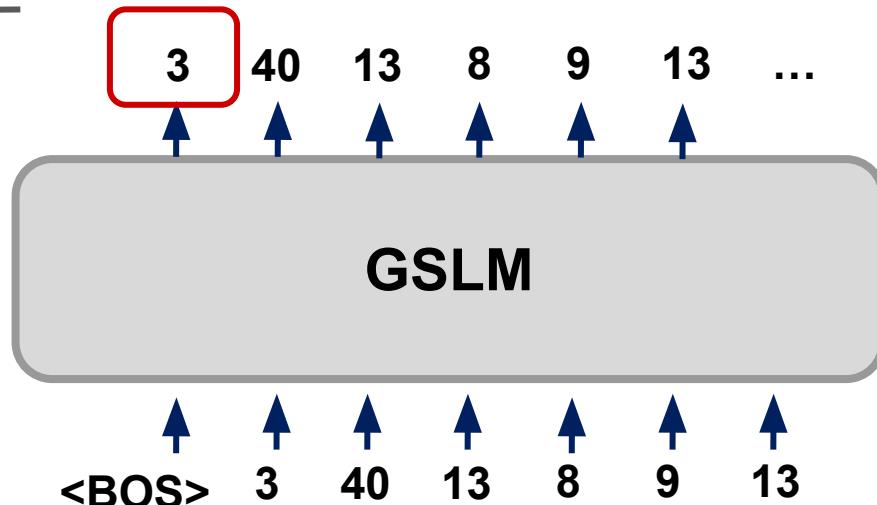
and to take in another path  
and to take in another path  
and to take in another path  
and to take in another path ...



temperature = 1.5 [babble]

at the swing here as to motions out of the events not  
time and abe he was any stump headed and flow any  
he's the kiln are tama why do ye take the floor ...

Unconditional Generation

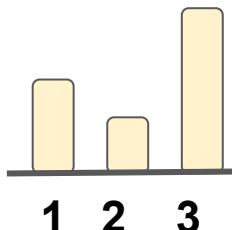


# Generative Spoken Language Modeling

temperature = 1

but it is attendant from the people to defend himself from  
this information pride of the potential in criminal activity a  
curiosity and impetuosity of the world a war soon acquired

“Locally” coherent



Unconditional Generation

4 40 13 8 9 13 ...



<BOS> 4 40 13 8 9 13

# Generative Spoken Language Modeling

temperature = 1

but it is attendant from the people to defend himself from  
this information pride of the potential in criminal activity a  
curiosity and impetuosity of the world a war soon acquired

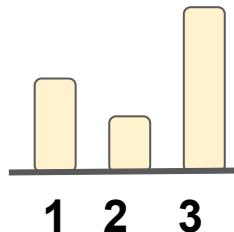


請翻譯 but it is attendant from the people to defend himself from this information pride of  
the potential in criminal activity a curiosity and impetuosity of the world a war soon  
acquired



這句話的意思可能有些混亂，但我會盡力翻譯出最合理的意思：

"但是，人們必須保護自己，免受罪犯活動、世界好奇心和冒失的影響，這些情況可能很快引  
發戰爭。"



Unconditional Generation

1 2 3 4 40 13 8 9 13 ...

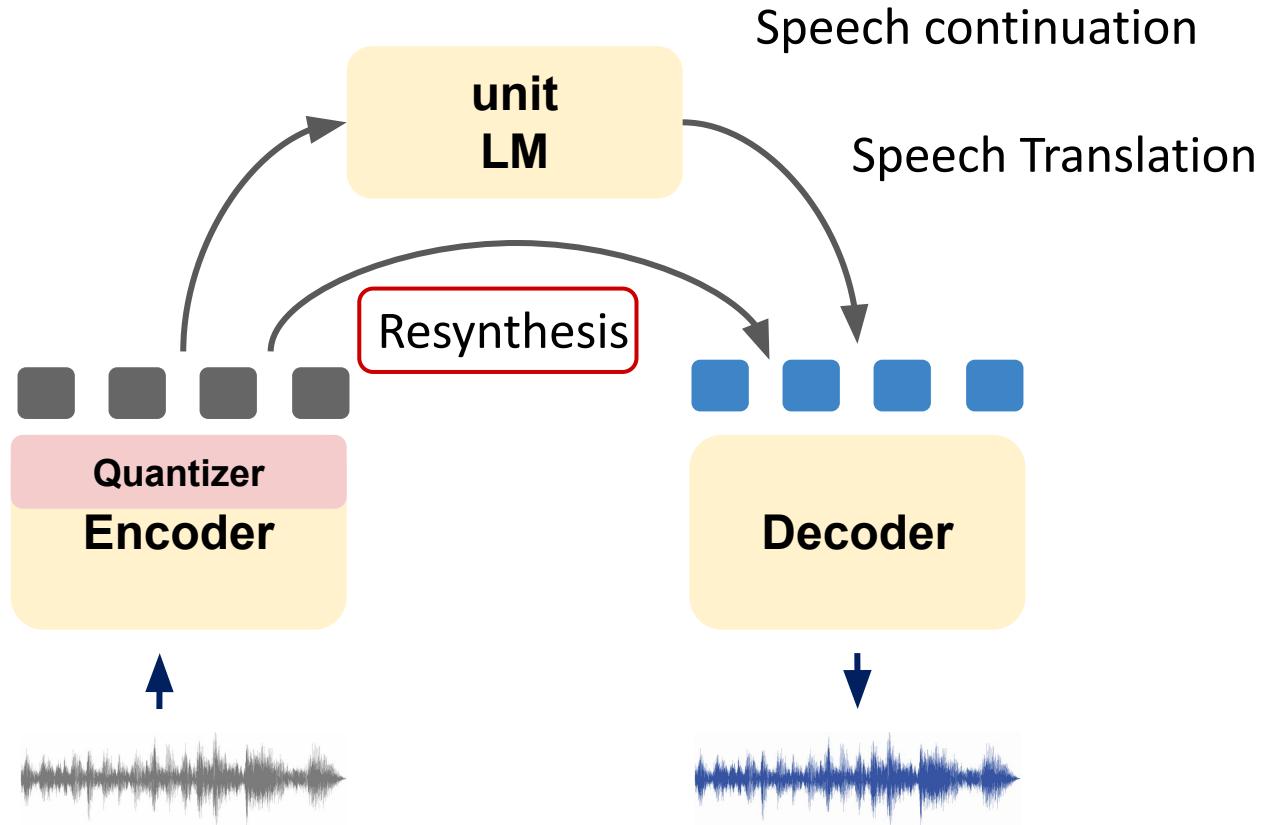
↑ ↑ ↑ ↑ ↑ ↑

GSLM

↑ ↑ ↑ ↑ ↑ ↑

40 13 8 9 13

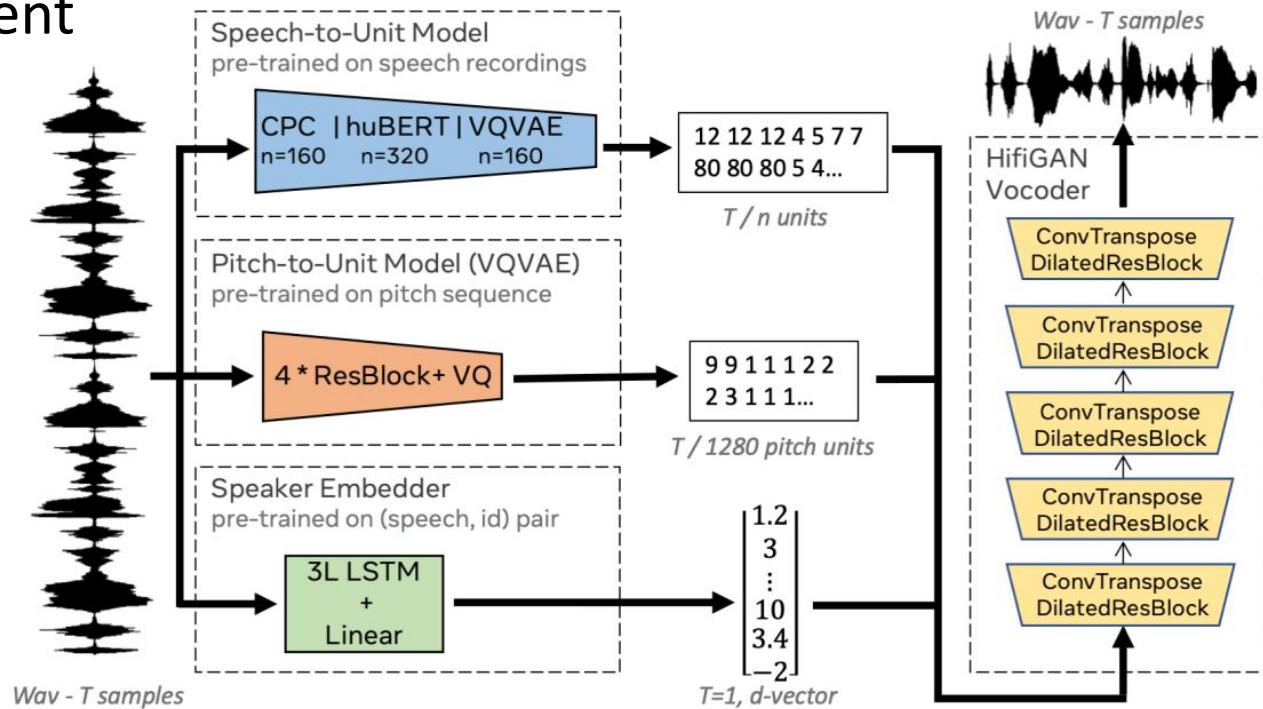
# Textless NLP Project



# Speech Resynthesis

## Feature disentanglement

- Content
- Pitch
- Speaker



Polyak, Adam, et al. "Speech resynthesis from discrete disentangled self-supervised representations." *arXiv preprint arXiv:2104.00355* (2021).

# Speech Resynthesis

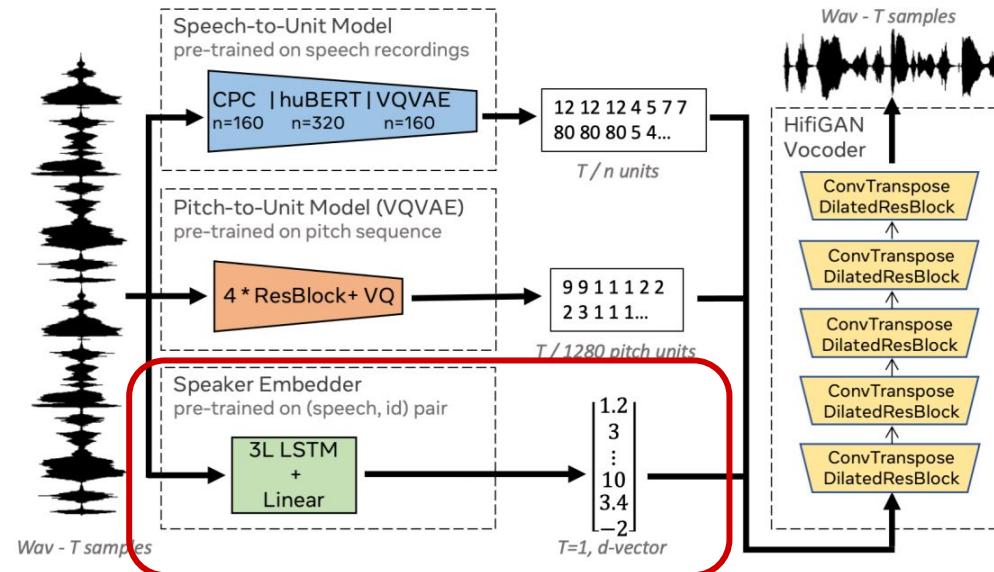
Speaker information is removed  
in the discrete units

Multistream is required to  
perform resynthesis

Model	Quantized?	Vocab. size	Accuracy
HuBERT	-	-	0.99
HuBERT	✓	50	0.11
HuBERT	✓	100	0.19
HuBERT	✓	200	0.29
HuBERT	✓	500	0.48
CPC	-	-	0.99
CPC	✓	50	0.19
CPC	✓	100	0.32
CPC	✓	200	0.34
CPC	✓	500	0.40

## Speaker Identification

# Speech Resynthesis

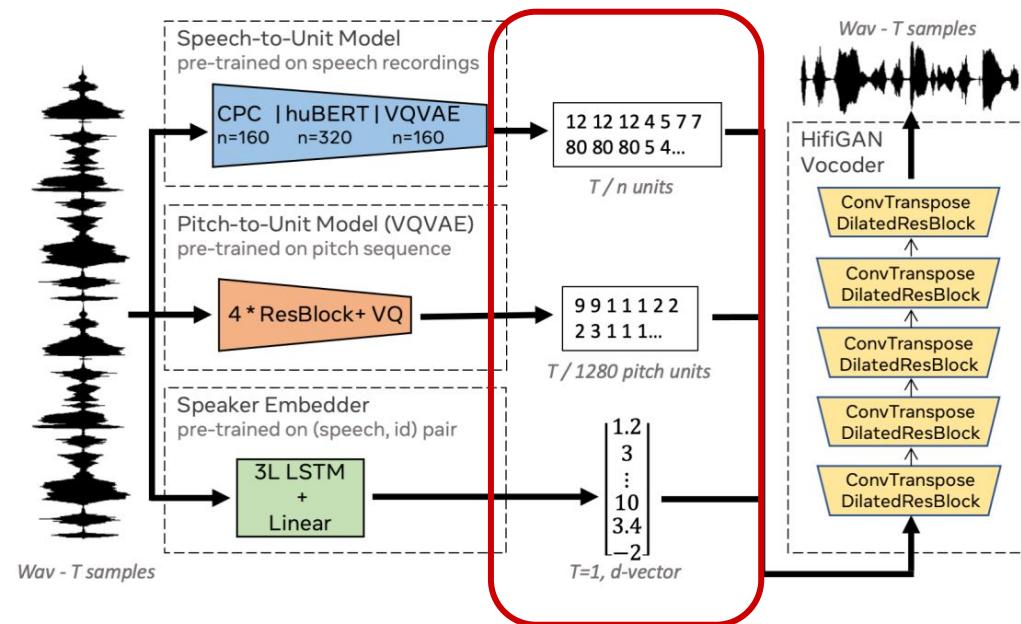


Replace the speaker embedding  
with other speaker's embedding

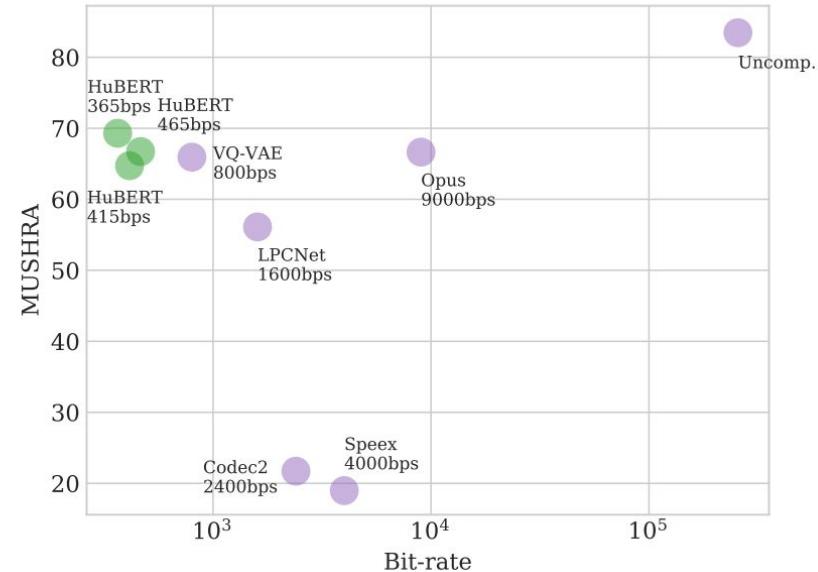
## Voice Conversion

Dataset	Method	Voice Conversion			
		PER ↓	WER ↓	EER ↓	MOS ↑
VCTK	GT	17.16	4.32	3.25	$4.11 \pm 0.29$
	CPC	22.22	16.11	0.46	$3.57 \pm 0.15$
	HuBERT	<b>19.09</b>	<b>12.23</b>	<b>0.31</b>	<b><math>3.71 \pm 0.24</math></b>
LJ	VQ-VAE	40.88	36.96	9.65	$2.90 \pm 0.17$
	CPC	23.58	15.98	<b>4.83</b>	$3.42 \pm 0.24$
	HuBERT	<b>20.85</b>	<b>12.72</b>	6.01	<b><math>3.58 \pm 0.28</math></b>
VCTK	VQ-VAE	36.88	29.44	11.56	$3.08 \pm 0.34$

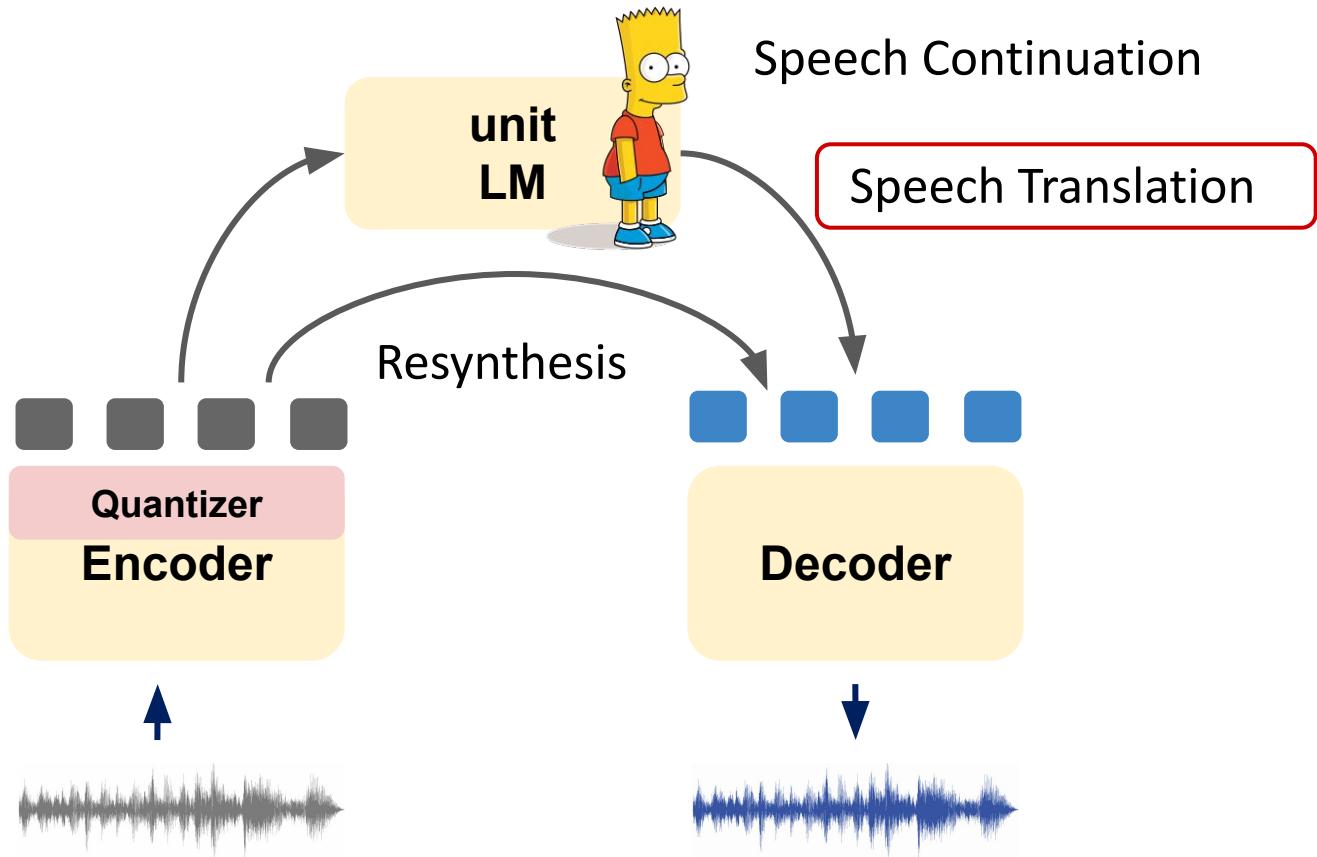
# Speech Resynthesis



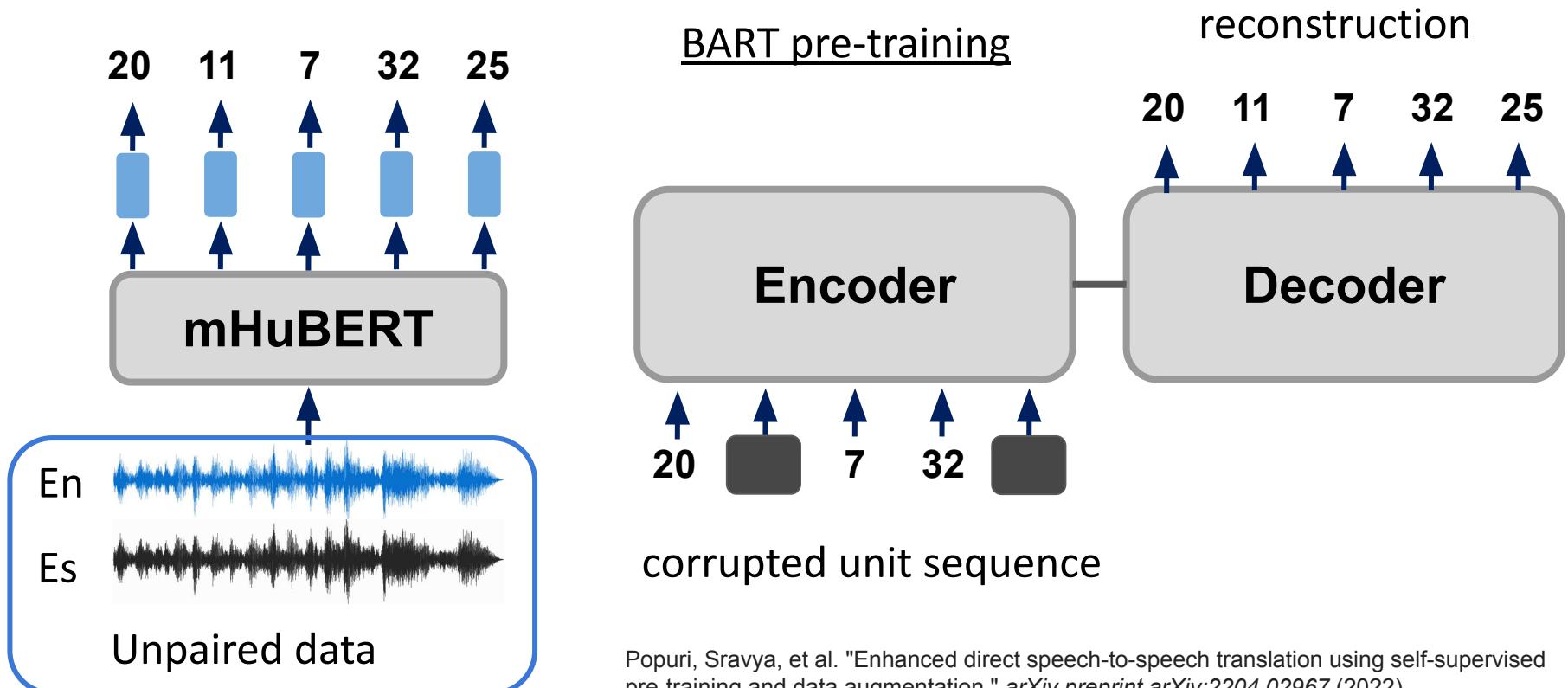
## Speech Codec



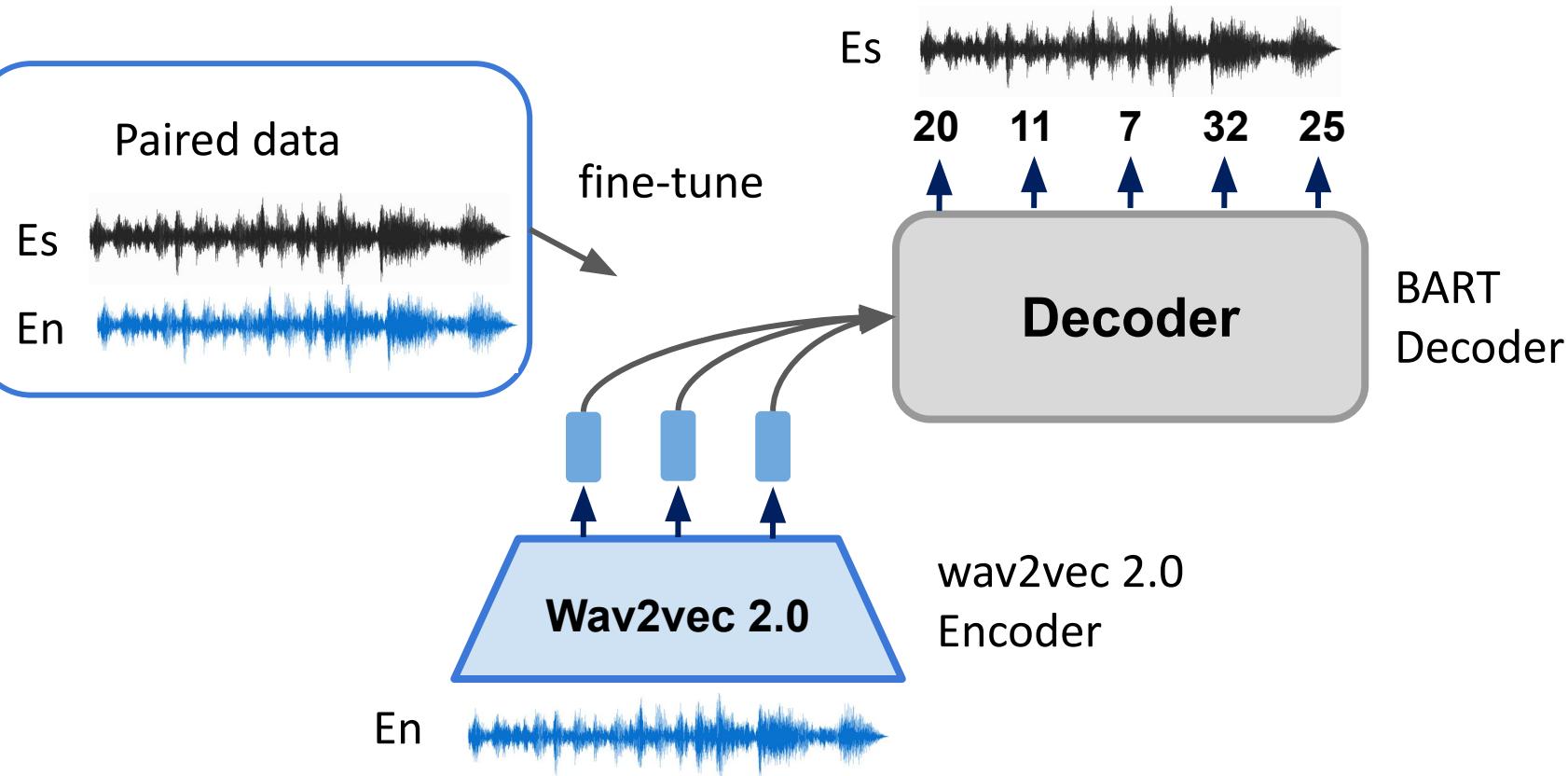
# Textless NLP Project



# Speech Translation: Unit BART



# Speech Translation: Unit BART



# Speech Translation: Unit BART

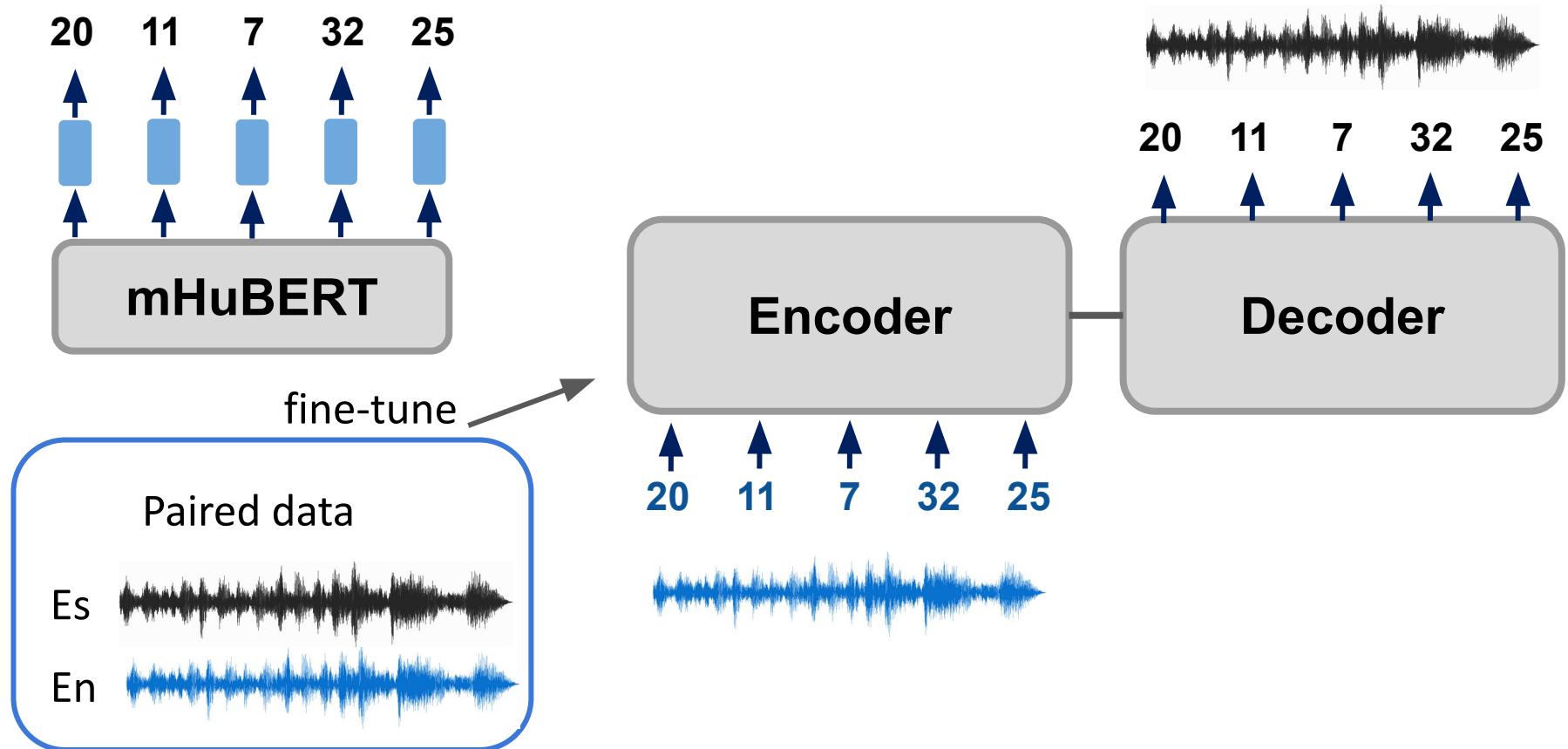


Table 2: Dev / test BLEU on all the datasets included in the “S2ST-syn” data. All S2UT systems are decoded with beam size 10. MOS is reported with 95% confidence interval. (w2v2-L: wav2vec 2.0 LARGE)

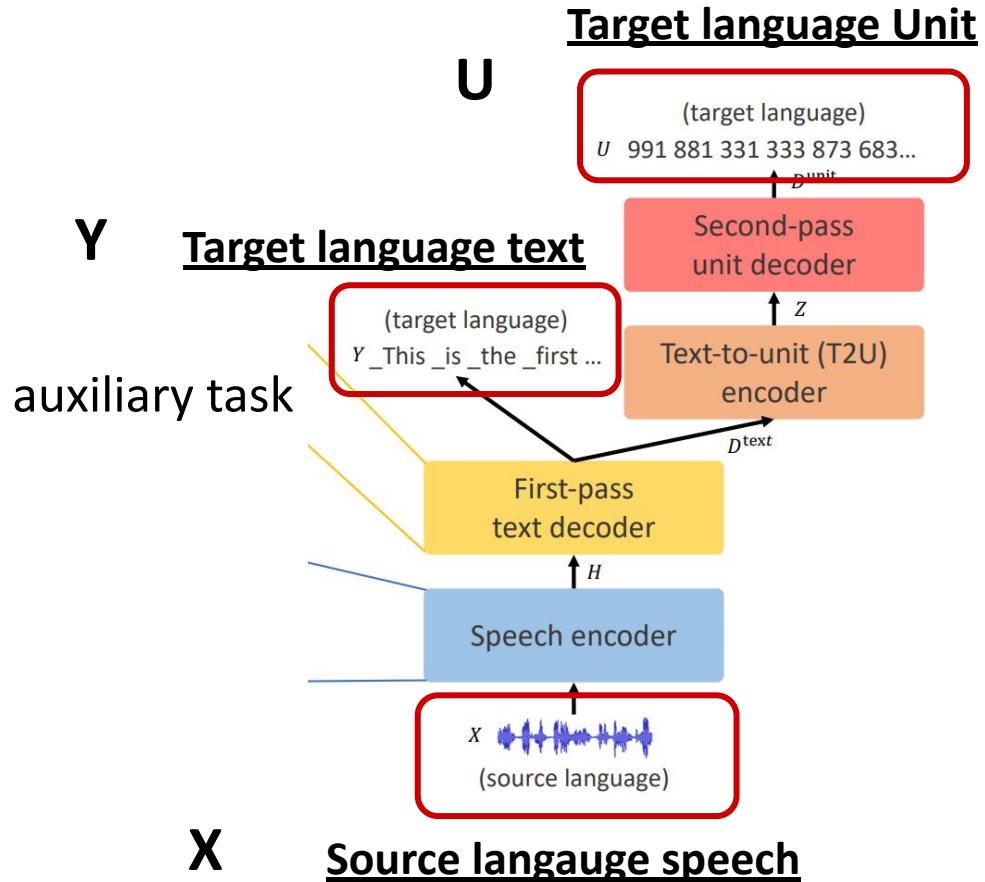
ID		En-Es			Es-En			MOS combined
		BLEU Europarl-ST	BLEU MuST-C	MOS combined	BLEU CoVoST-2	BLEU Europarl-ST	mTEDx	
<b>Cascaded systems:</b>								
1	S2T (w2v2-L)+TTS	33.0 / 32.6	30.3 / 30.1	3.80 ± 0.12	25.9 / 28.4	26.9 / 23.6	25.3 / 21.5	3.53 ± 0.14
2	ASR+MT+TTS	28.9 / 28.8	36.4 / 34.2	-	37.3 / 33.8	33.3 / 29.1	29.3 / 32.4	-
<b>S2UT systems without pre-training:</b>								
3	S2UT (w/o multitask) [4]	23.8 / 24.0	25.0 / 23.3	-	0.0 / 0.0	0.0 / 0.0	0.1 / 0.0	-
4	S2UT (w/ multitasks) [4]	25.5 / 25.8	26.3 / 24.3	3.97 ± 0.09	20.6 / 22.7	20.4 / 18.0	20.2 / 16.9	3.26 ± 0.09
<b>S2UT systems with model pre-training:</b>								
5	w2v2-L	30.8 / 31.0	31.1 / 30.3	3.35 ± 0.15	24.4 / 27.0	24.2 / 21.5	24.3 / 21.0	3.15 ± 0.14
6	w2v2-L + mBART (LNA-E)	30.1 / 30.4	31.0 / 28.2	-	24.4 / 27.1	24.0 / 21.4	23.6 / 21.1	-
7	w2v2-L + mBART (LNA-D)	<b>32.2 / 32.5</b>	<b>32.6 / 30.8</b>	4.06 ± 0.10	<b>27.3 / 30.2</b>	<b>29.0 / 26.4</b>	<b>29.6 / 25.2</b>	2.81 ± 0.16
8	w2v2-L + mBART (LNA-E,D)	30.6 / 31.0	31.3 / 29.3	-	26.8 / 29.6	27.6 / 25.2	24.7 / 22.3	-
9	w2v2-L + mBART (full)	31.4 / 30.8	31.2 / 30.5	-	27.3 / 30.1	27.0 / 24.4	26.6 / 24.2	-

Speech to speech translation is competitive to cascaded systems  
(Without any text supervision)

# Speech Translation: UnitY

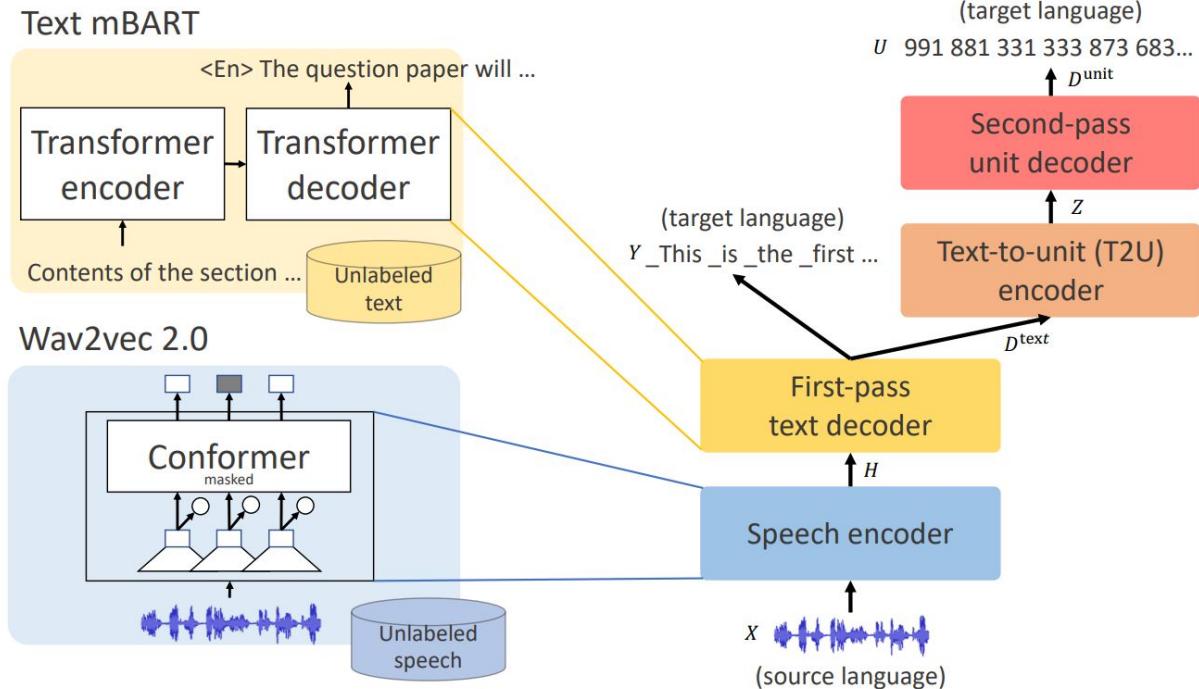
1. Speech Encoder
2. First-pass text decoder
3. Text-to-unit encoder
4. Second-pass unit decoder

X: Source language speech input  
Y: Target language text  
U: Target language discrete units



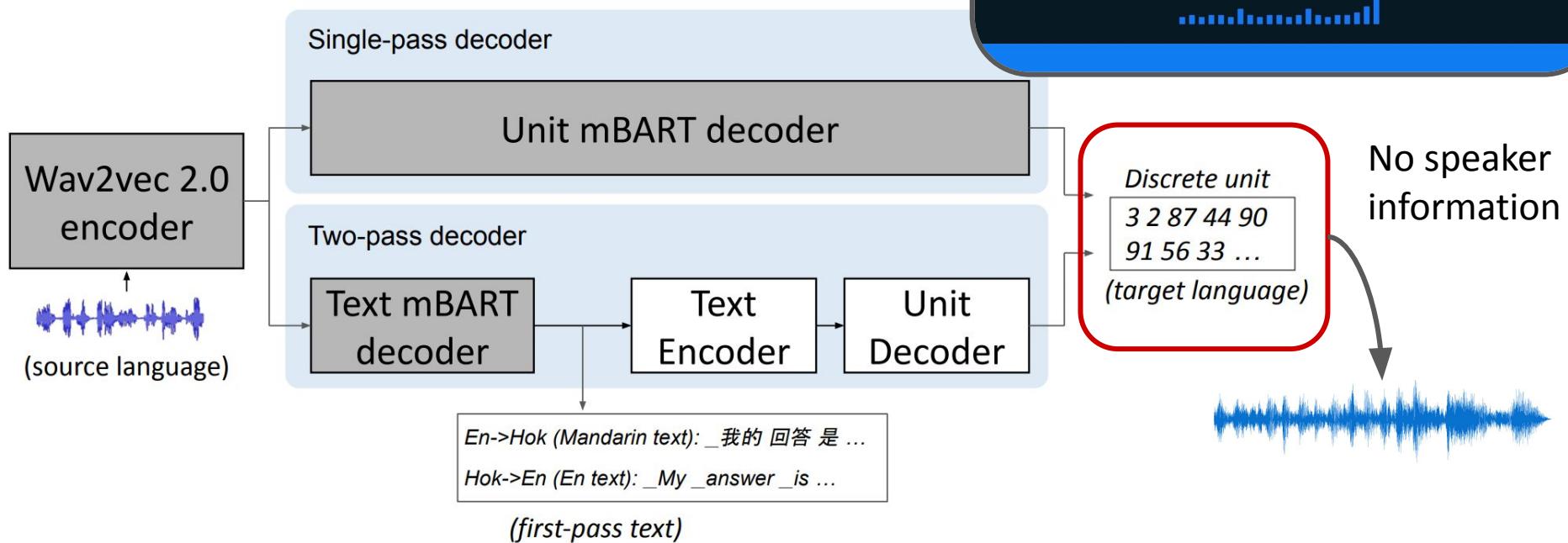
# Speech Translation: UnitY

- Unlabeled text
- Unlabeled speech
- labeled speech
- paired speech



ID	Model	Encoder	ASR-BLEU ( $\uparrow$ )				
			dev	dev2	test		
A0	Synthetic target (Lee et al., 2022a)			88.5	89.4	90.5	
<b>Cascaded systems</b>							
A1	ASR → MT → TTS	LSTM (Lee et al., 2022a)	42.1	43.5	43.9		
A2		LSTM (Jia et al., 2019b)	39.4	41.2	41.4		
A3		LSTM (Jia et al., 2022b)	–	–	43.3		
A4	S2TT → TTS	LSTM (Lee et al., 2022a)	38.5	39.9	40.2		
A5		Transformer (Dong et al., 2022)	44.3	45.4	45.1		
A6		Conformer	47.8	48.9	48.3		
A7		Conformer wav2vec2.0	51.0	52.2	52.1		
...							
<b>Direct systems (speech-to-unit)</b>							
A17		Transformer (Lee et al., 2022a)	–	–	39.9		
A18	S2UT	Conformer	46.2	47.6	47.4		
A19		Conformer wav2vec2.0	53.4	53.9	53.7		
A20	UnitY	Conformer	50.5	51.6	51.4		
A21		Conformer wav2vec2.0	<b>55.1</b>	<b>56.5</b>	<b>55.9</b>		

# Speech-to-Speech Translation For A Real-world Unwritten Language



# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. VALL-E

### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

## AudioLM: a Language Modeling Approach to Audio Generation

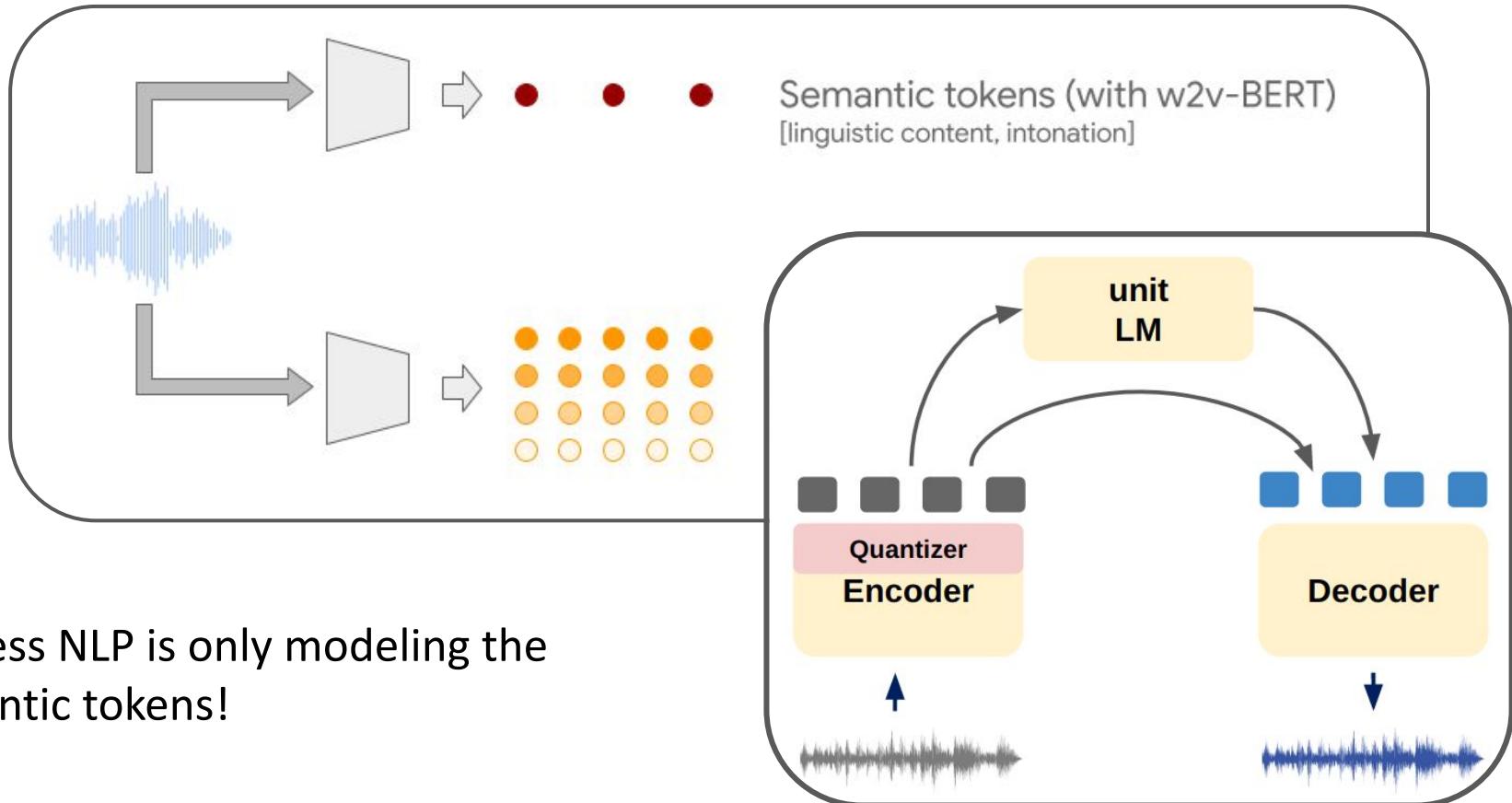
---

THURSDAY, OCTOBER 06, 2022

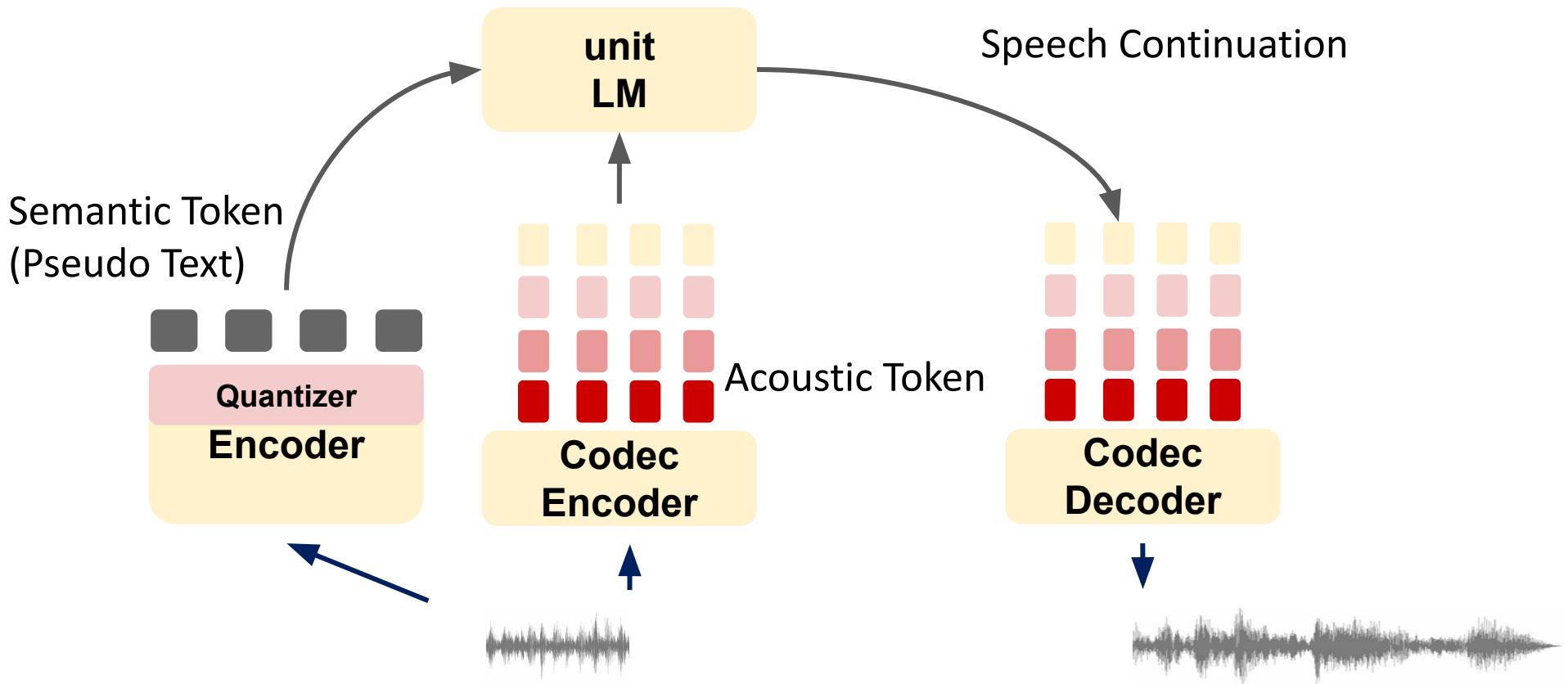
*Posted by Zalán Borsos, Research Software Engineer, and Neil Zeghidour, Research Scientist, Google Research*

# AudioLM

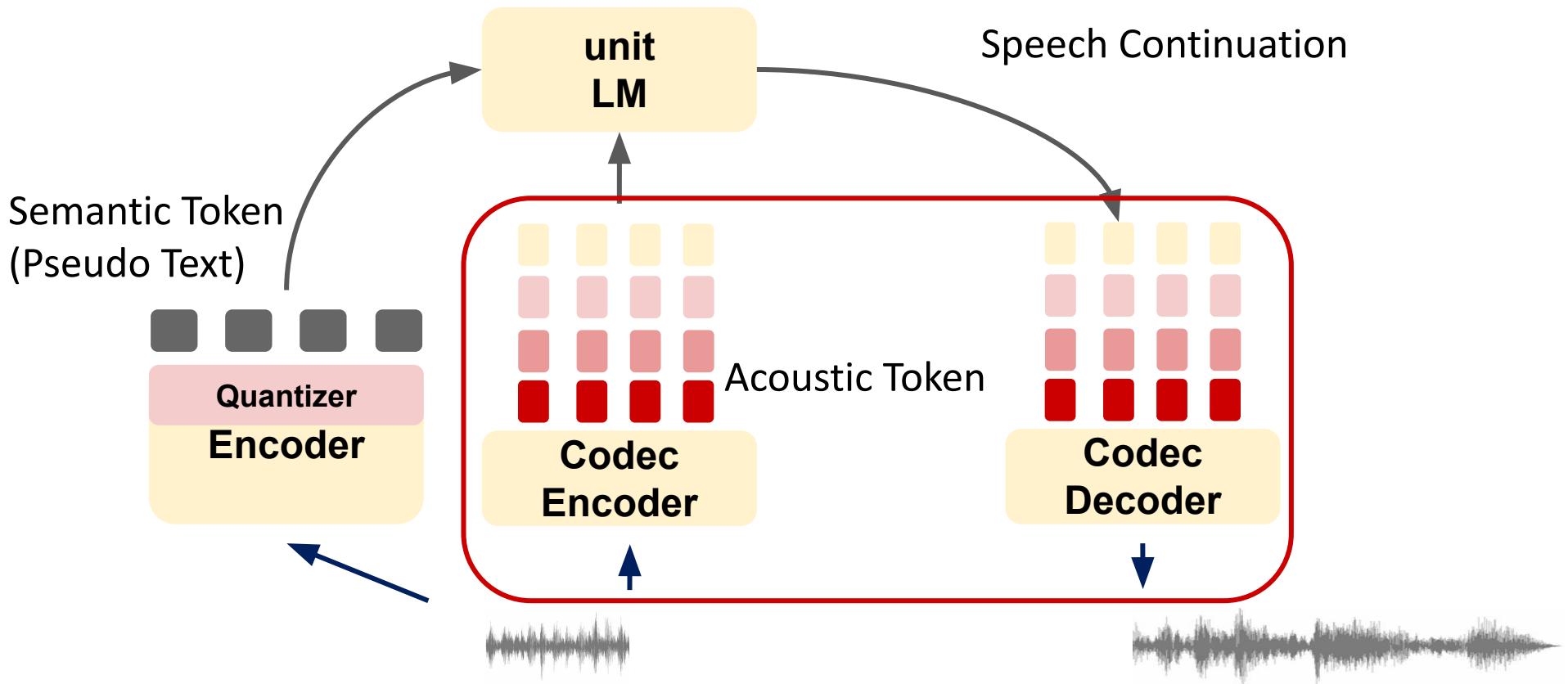
Borsos, Zalán, et al. "Audiolm: a language modeling approach to audio generation." *arXiv preprint arXiv:2209.03143* (2022).



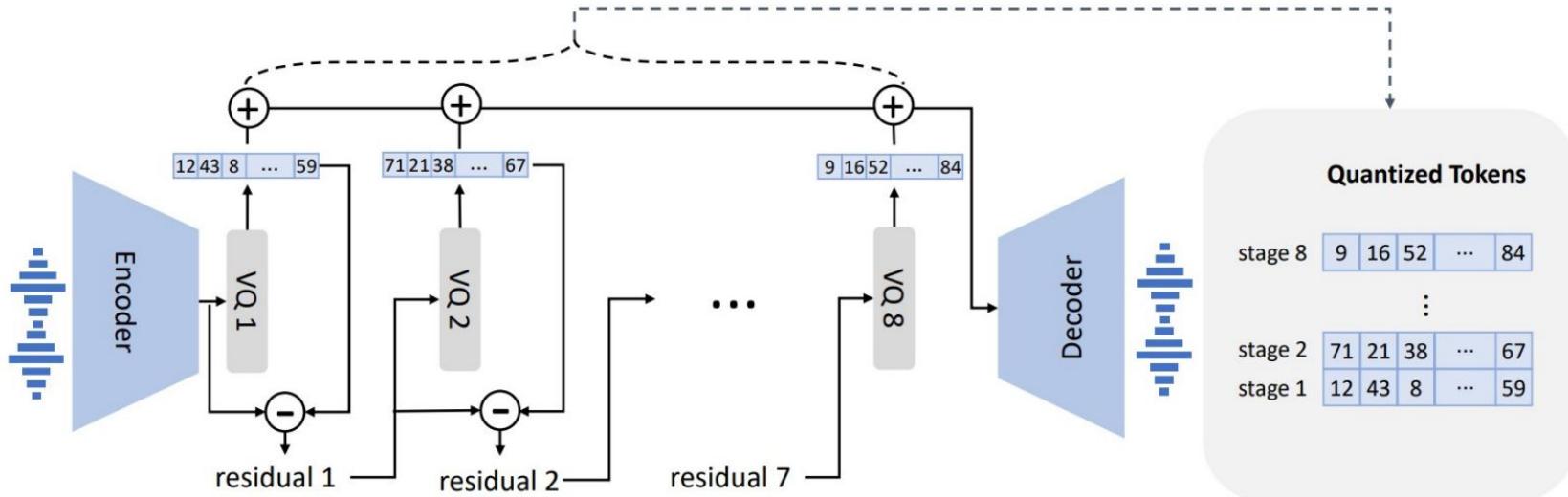
# AudioLM



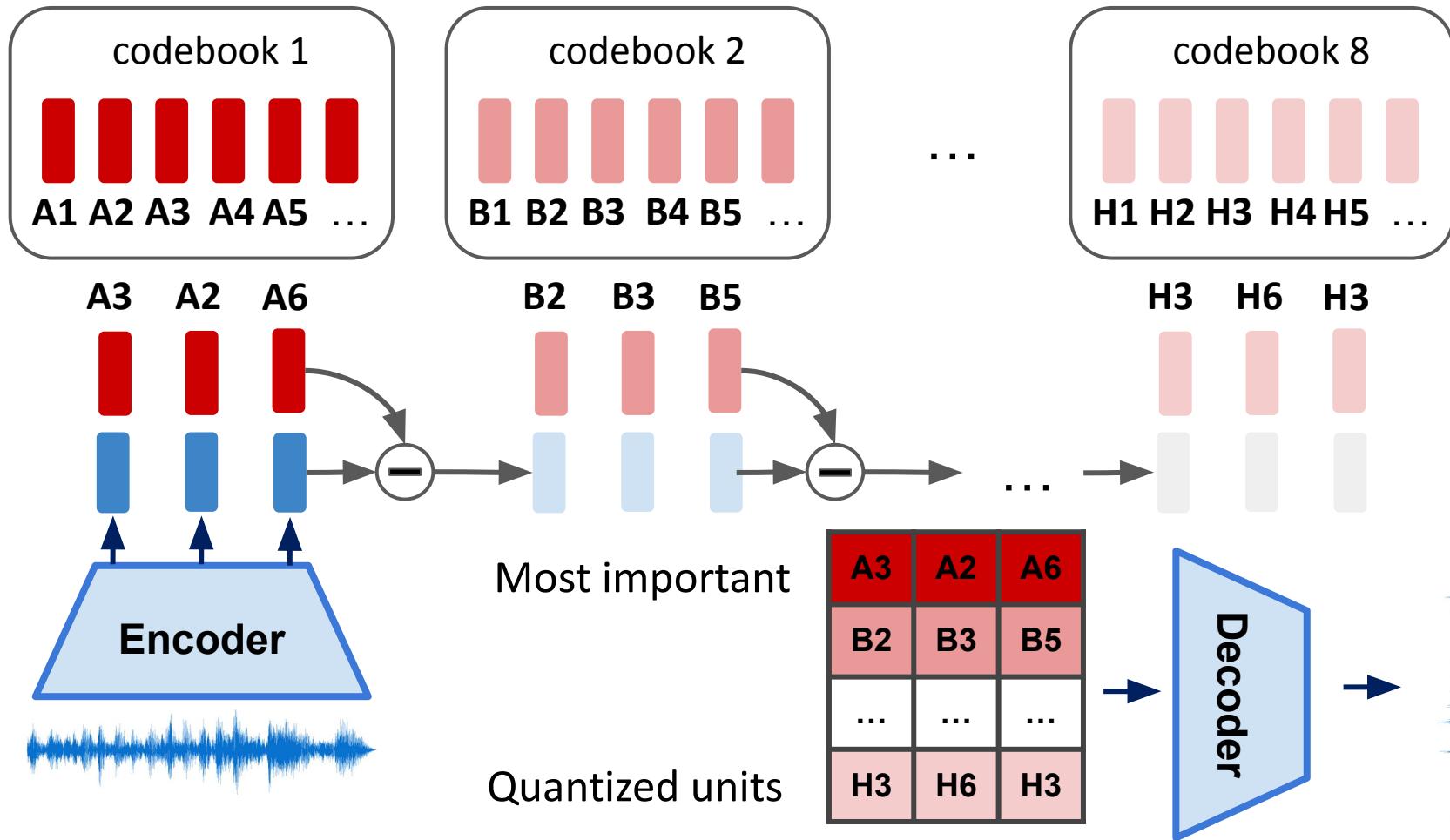
# AudioLM



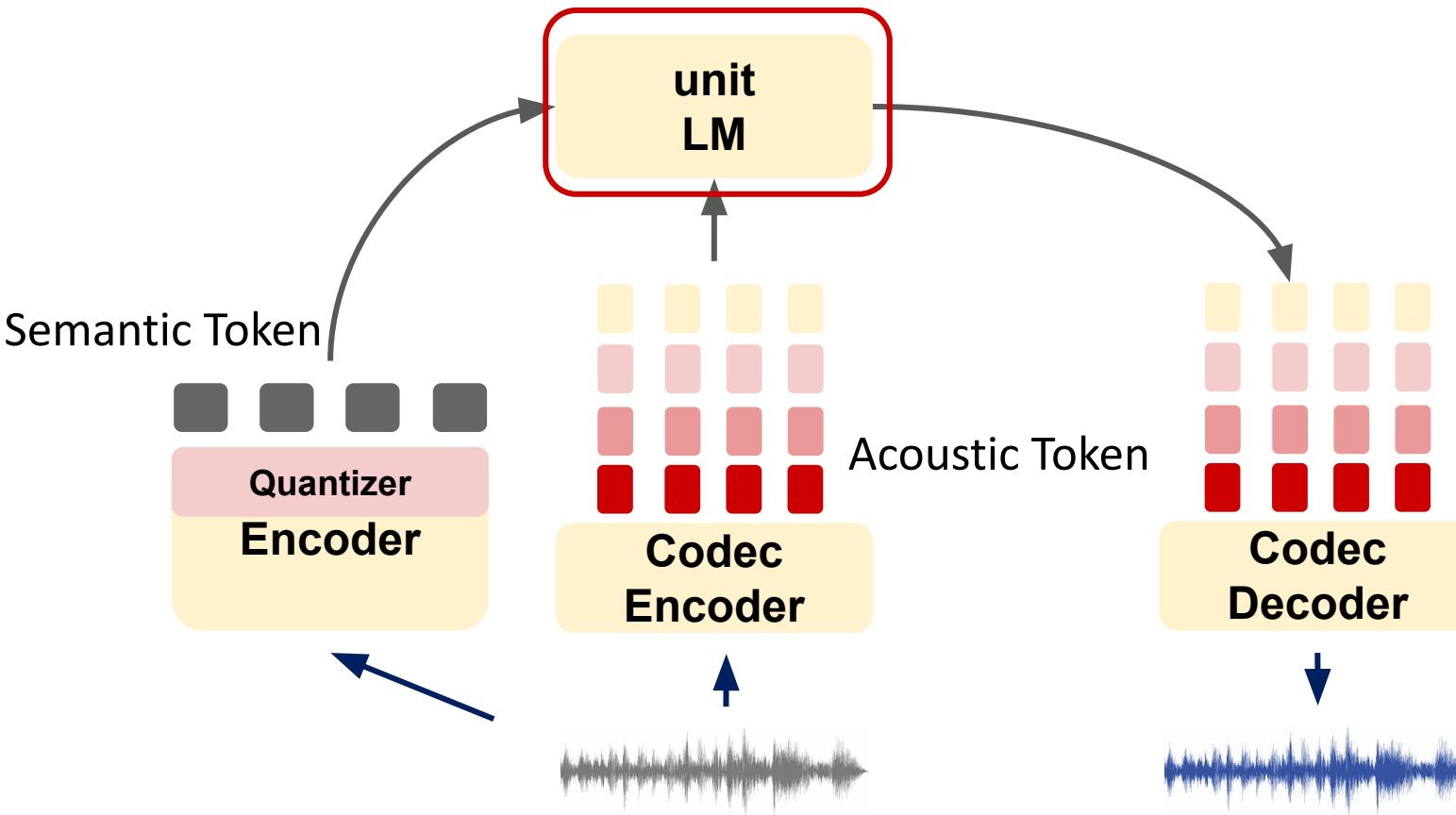
# Codec Model



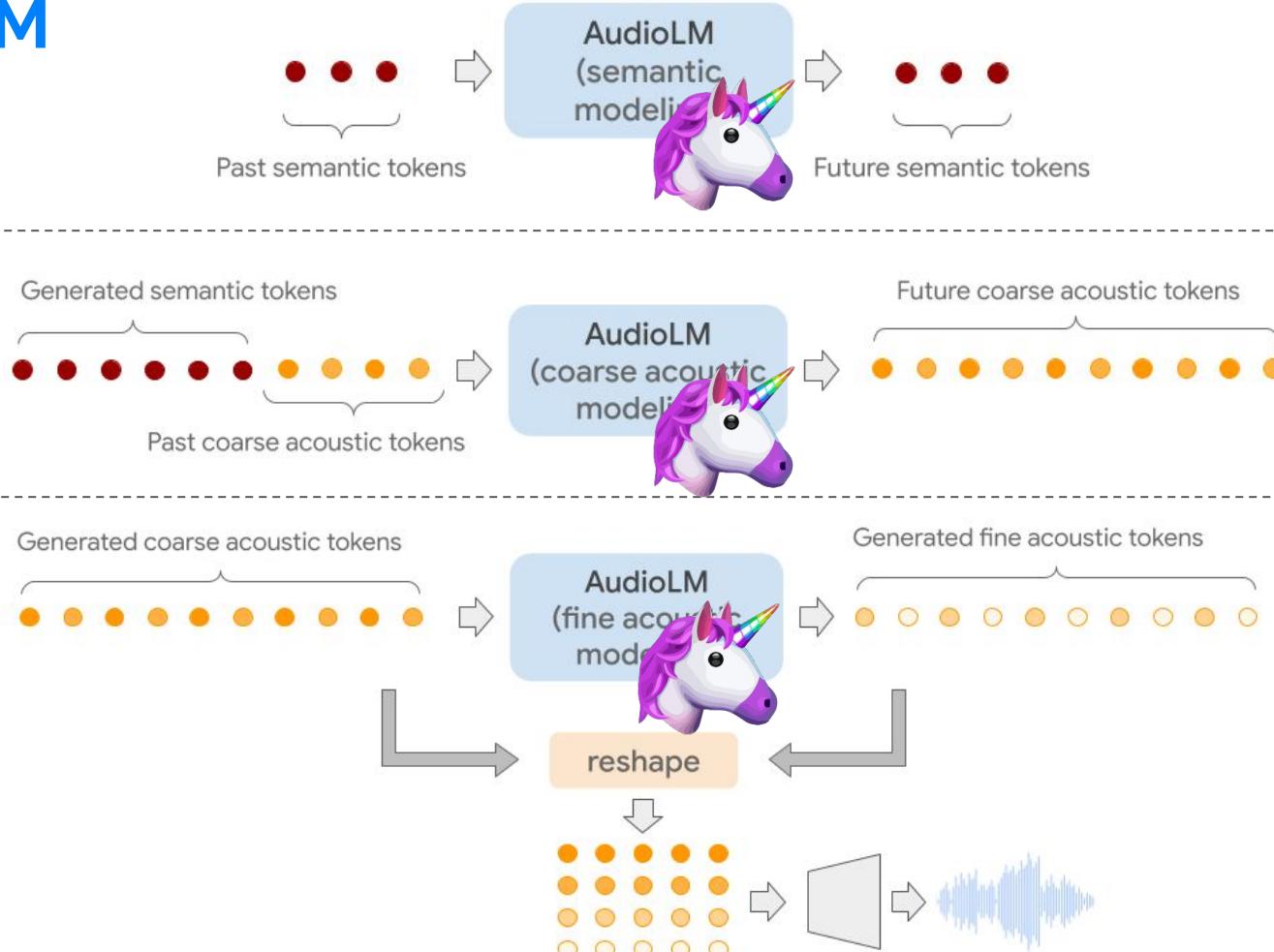
Residual Vector Quantization



# AudioLM



# AudioLM





prompt | generated speech

Speech Continuation



prompt | generated piano

Music Continuation

# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM
3. VALL-E

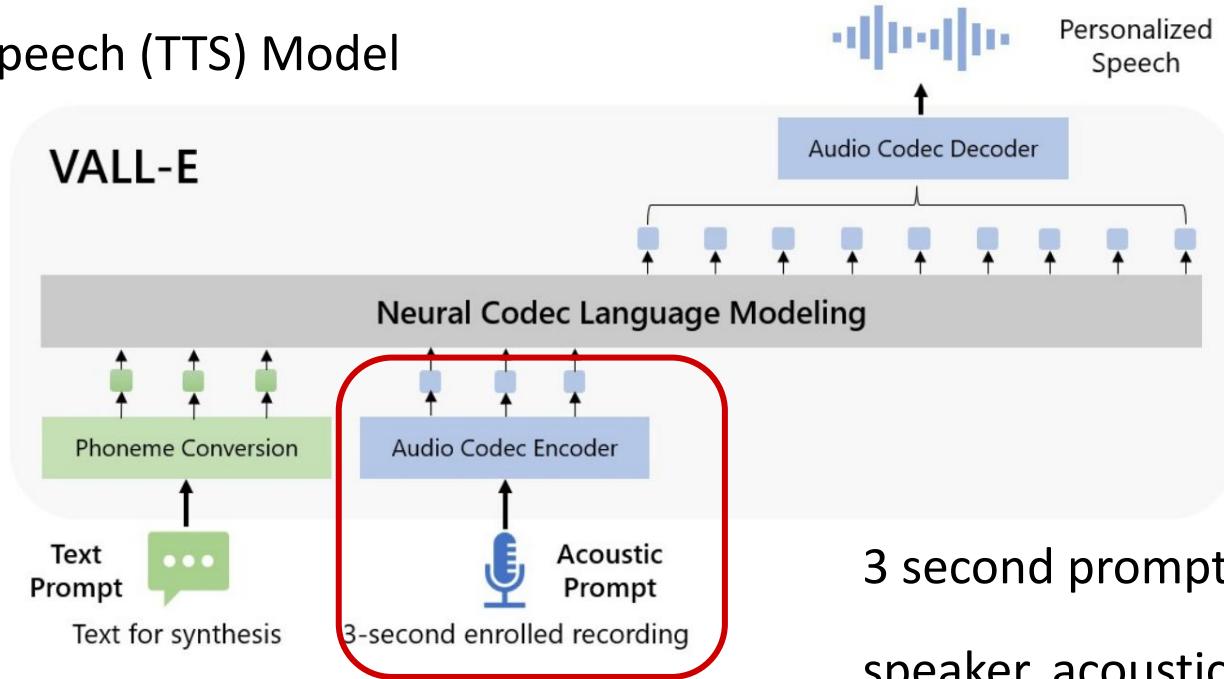
### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

# VALL-E

Text-to-speech (TTS) Model



3 second prompt  
speaker, acoustic coherent

# VALL-E

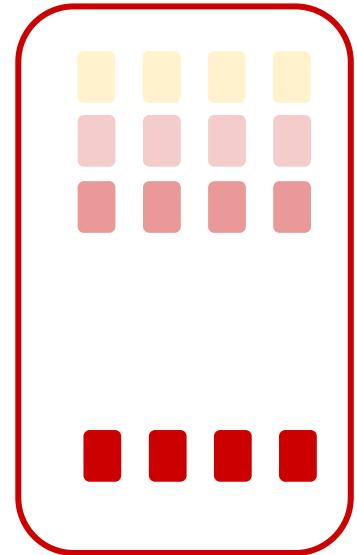
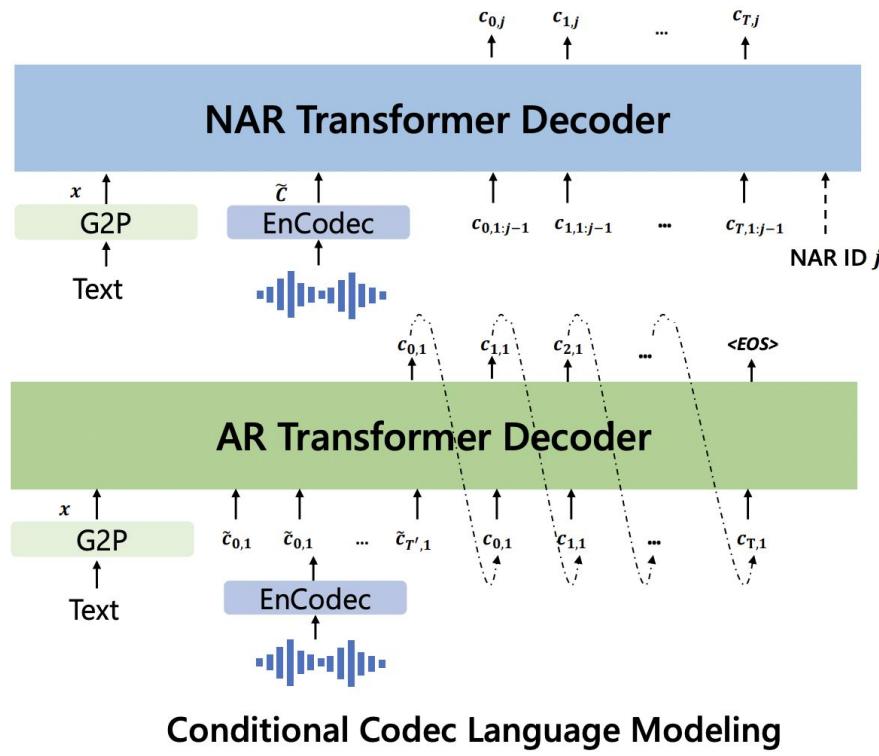
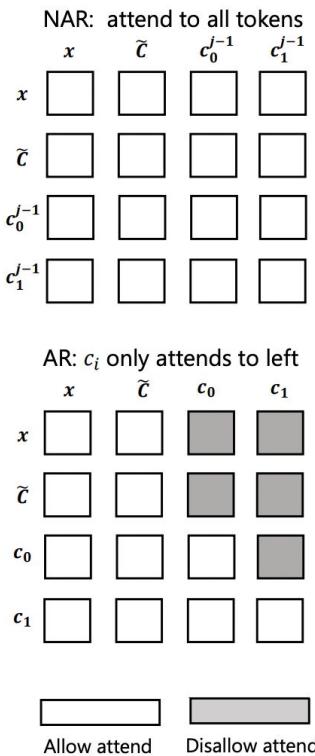
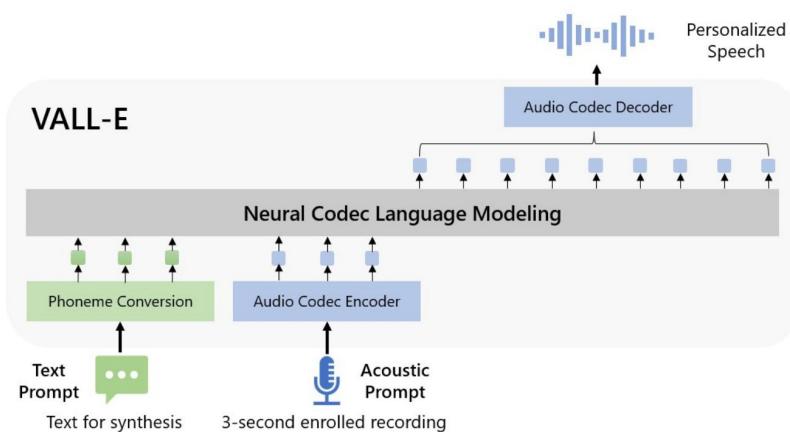


Figure 3: The structure of the conditional codec language modeling, which is built in a hierarchical manner. In practice, the NAR decoder will be called seven times to generate codes in seven quantizers.

# VALL-E

- Beat state-of-the-art TTS system
- Speaker similarity is high
- maintain emotion
- maintain acoustic environment



model	WER	SPK
GroundTruth	2.2	0.754
<b>Speech-to-Speech Systems</b>		
GSLM	12.4	0.126
AudioLM*	6.0	-
<b>TTS Systems</b>		
YourTTS	7.7	0.337
<b>VALL-E</b>	<b>5.9</b>	<b>0.580</b>
VALL-E-continual	3.8	0.508

<https://valle-demo.github.io/>

# Overview

## Speech Foundation Models

### Part 1

#### Speech Representation Learning

1. SSL Models
2. Representation benchmarking
3. Efficiently using these models

### Part 2

#### Speech Large Language Models

1. Textless NLP
2. AudioLM

### Part 3

#### Other Speech Foundation Models

1. Whisper
2. USM

# Introducing Whisper



<https://openai.com/research/whisper>

# Whisper

## Multitask training data (680k hours)

### English transcription

- 🗣 "Ask not what your country can do for ..."
- 📝 Ask not what your country can do for ...

### Any-to-English speech translation

- 🗣 "El rápido zorro marrón salta sobre ..."
- 📝 The quick brown fox jumps over ...

### Non-English transcription

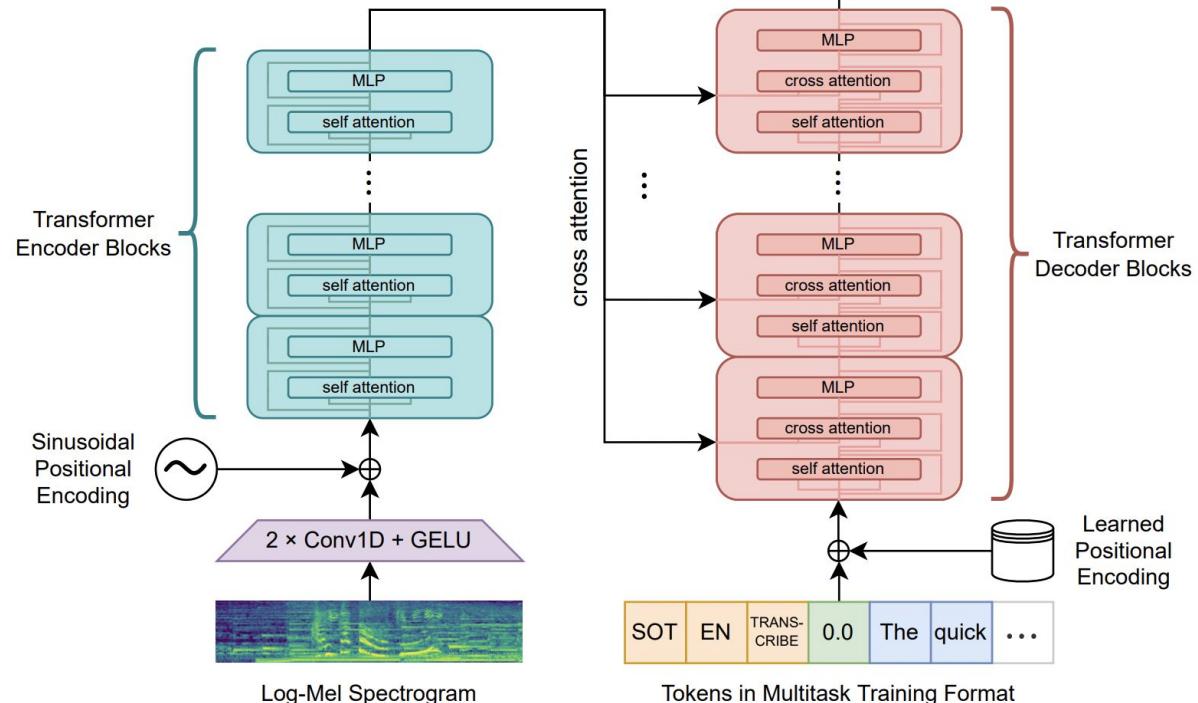
- 🗣 "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ..."
- 📝 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ...

### No speech

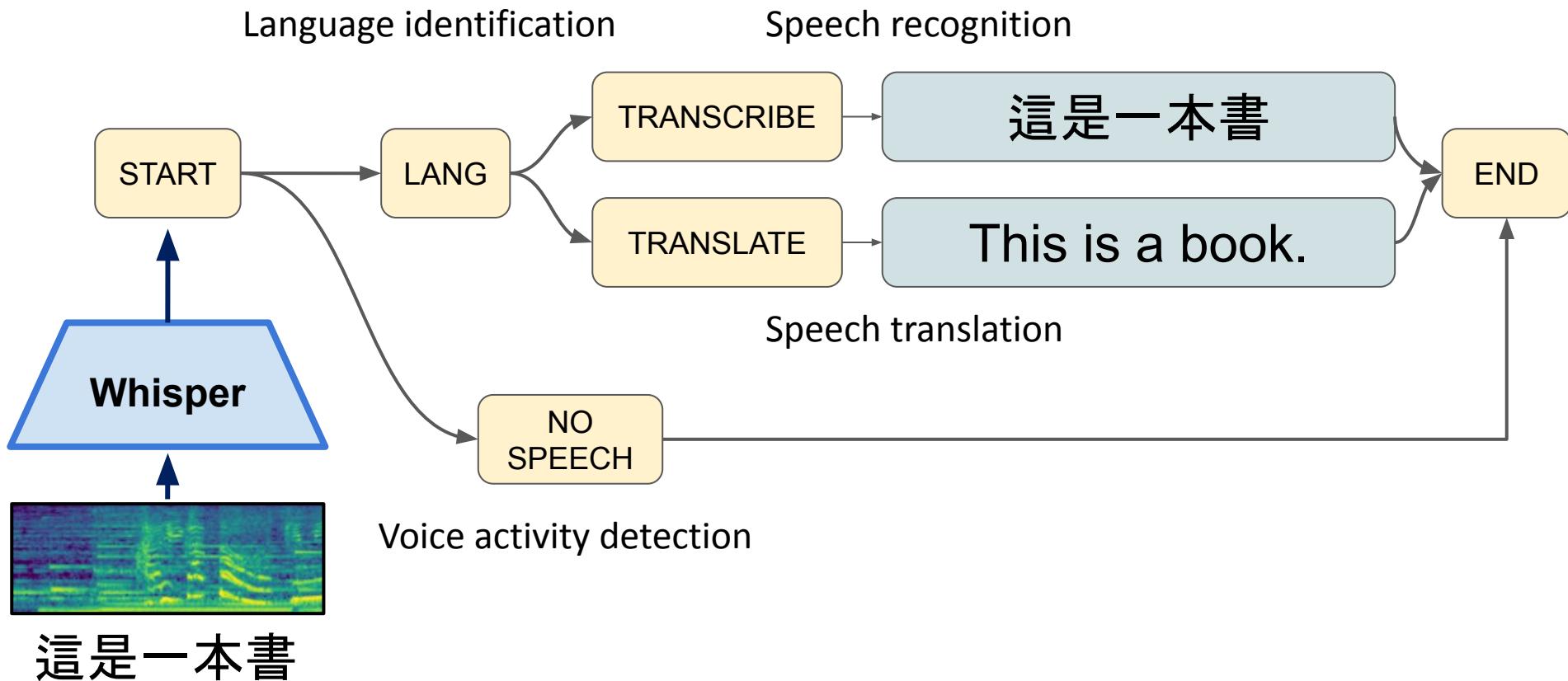
- 🔊 (background music playing)
- 📝 Ø

- 680,000 hours labeled data
- Multitask learning

## Sequence-to-sequence learning



# Whisper Multitasking



# Multilingual Speech Recognition

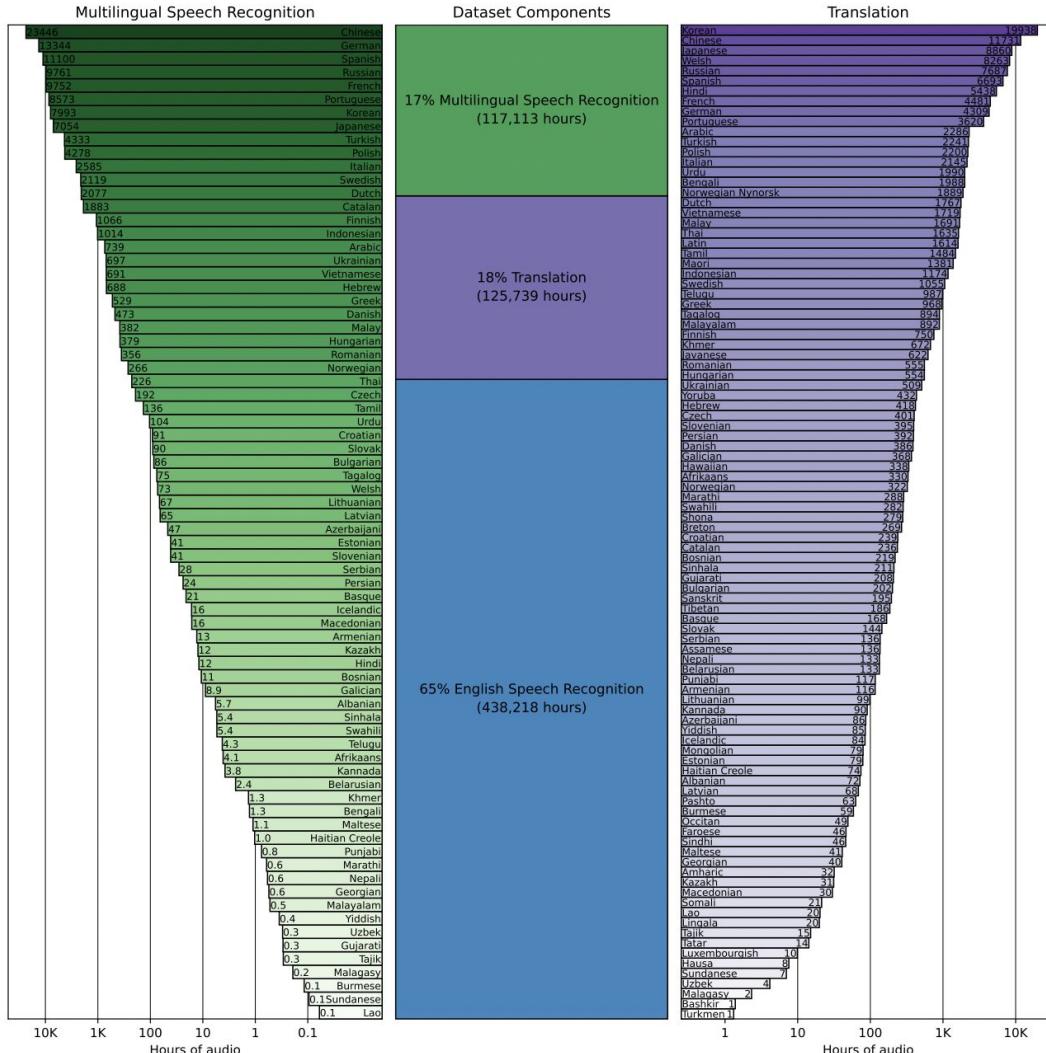
(~120,000 hours)

Chinese  
German  
Spanish

...

English  
Speech Recognition

(~440,000 hours)



X to English Translation

(~120,000 hours)

Korean  
Chinese  
Japanese

...



Search or jump to...

Pull requests Issues Codespaces Marketplace Explore



openai / whisper

Public

Watch 338

Fork 4k

Starred

35.9k

<> Code Pull requests 19 Discussions Actions Security Insights

main 5 branches 6 tags

Go to file

Add file

About

funboarder13920 fix condition\_on\_previous\_text (#1224) ...

248b6cb last week 11

.github/workflows Python 3.11 (#1171)

data initial commit 8 m

notebooks Use ndimage.median\_filter instead of signal.medfilter (#812) 4 m

tests Fix truncated words list when the replacement character is decode... 2 m

whisper fix condition\_on\_previous\_text (#1224)

.flake8 apply formatting with black (#1038) 2 m

.gitattributes fix github language stats getting dominated by jupyter notebook (...) 2 m

.gitignore initial commit 8 m

CHANGELOG.md Release 20230314 2 m

# Whisper is open sourced!

```
import whisper

model = whisper.load_model("base")

# load audio and pad/trim it to fit 30 seconds
audio = whisper.load_audio("audio.mp3")
audio = whisper.pad_or_trim(audio)

# make log-Mel spectrogram and move to the same device as the model
mel = whisper.log_mel_spectrogram(audio).to(model.device)

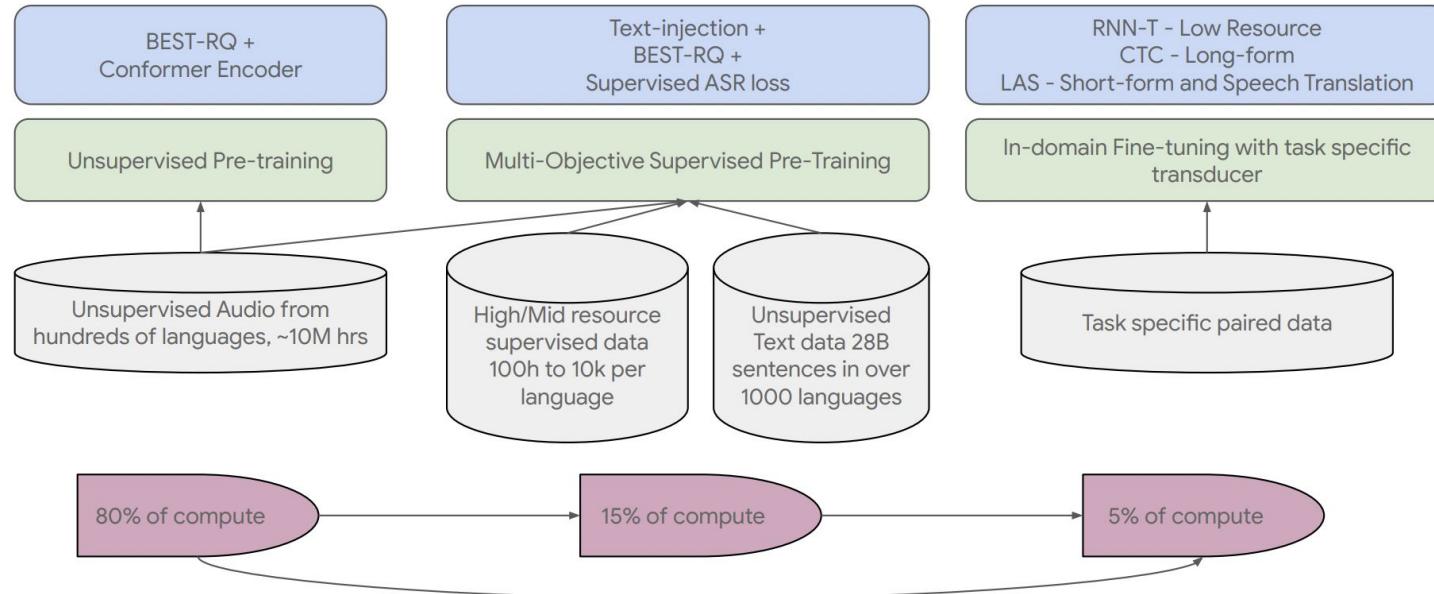
# detect the spoken language
_, probs = model.detect_language(mel)
print(f"Detected language: {max(probs, key=probs.get)}")

# decode the audio
options = whisper.DecodingOptions()
result = whisper.decode(model, mel, options)

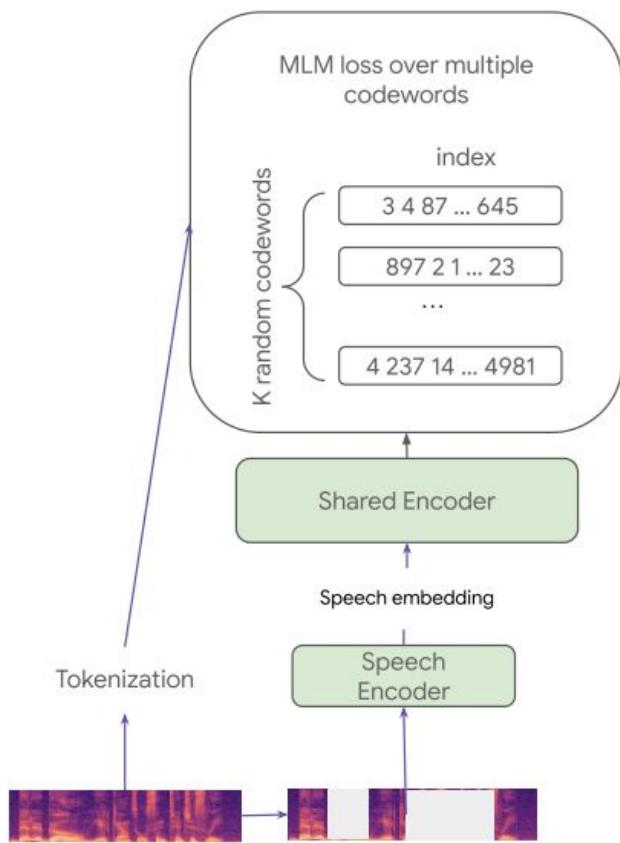
# print the recognized text
print(result.text)
```

# USM: Universal Speech Model

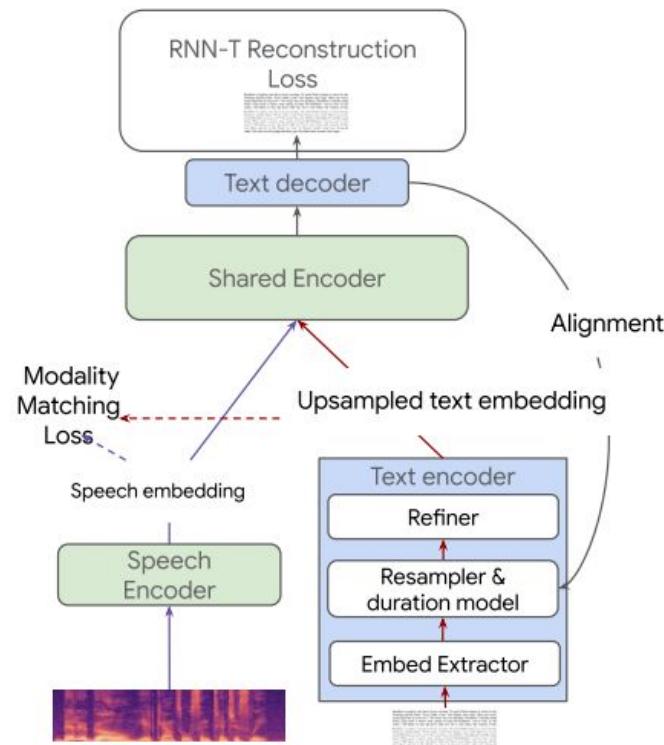
- Pre-train: 12M hours / 300 languages
- Fine-tune: 1/7 of the dataset used in Whisper



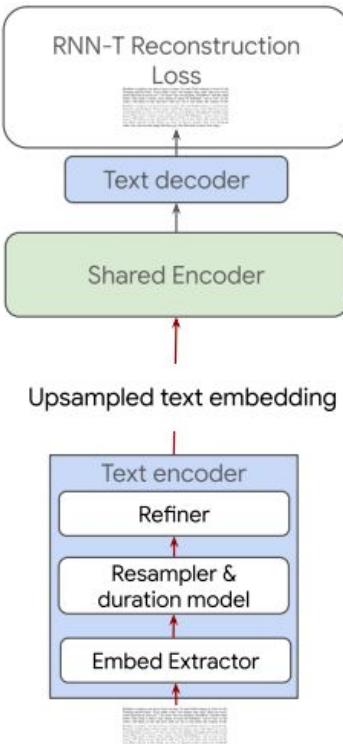
### On Speech input



### On paired input



### On text input



Task	Multilingual Long-form ASR			Multidomain en-US		Multilingual ASR	
	Dataset	YouTube	CORAAL	SpeechStew	FLEURS		
Langauges	en-US	18	en-US	en-US	62	102	
<b>Prior Work (single model)</b>							
Whisper-longform	17.7	27.8	-	23.9	12.8		
Whisper-shortform <sup>†</sup>	-	-	-	13.2 <sup>‡</sup>	11.5	36.6	-
<b>Our Work (single model)</b>							
USM-LAS	14.4	19.0	29.8	<b>11.2</b>	<b>10.5</b>	<b>12.5</b>	-
USM-CTC	<b>13.7</b>	<b>18.7</b>	<b>26.7</b>	12.1	10.8	15.5	-

- Fine-tune: 1/7 of the dataset used in Whisper

# Conclusion

1. Self-supervised Speech Models as feature extractor
2. Speech Large Language Model - Generative AI
3. Quantization is very important
4. How to efficiently use these speech foundation models?  
(not covered today)
  - a. prompting
  - b. adapters