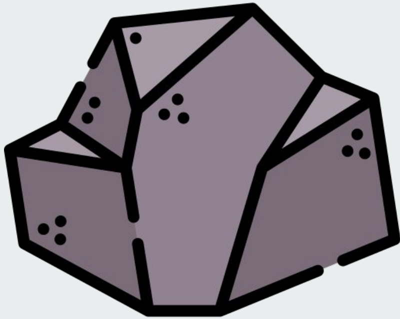
Speech Foundation Models

語音基石模型

2023/05/12 張凱爲
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

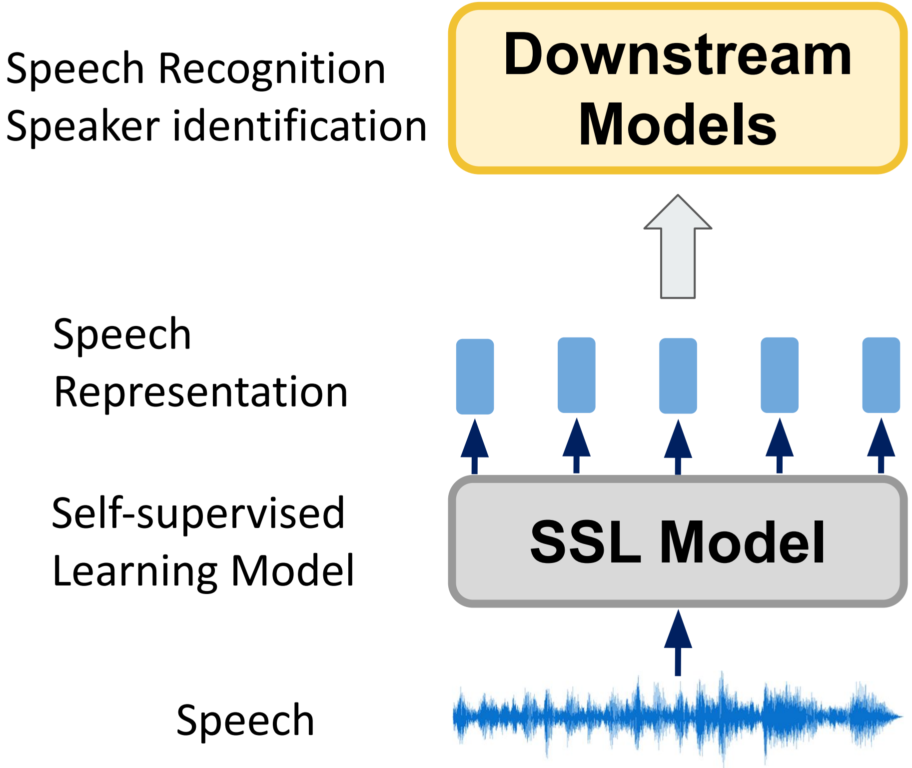
Overview

Part 1

Speech Representation Learning

1. SSL Models
2. Representation benchmarking

SUPERB

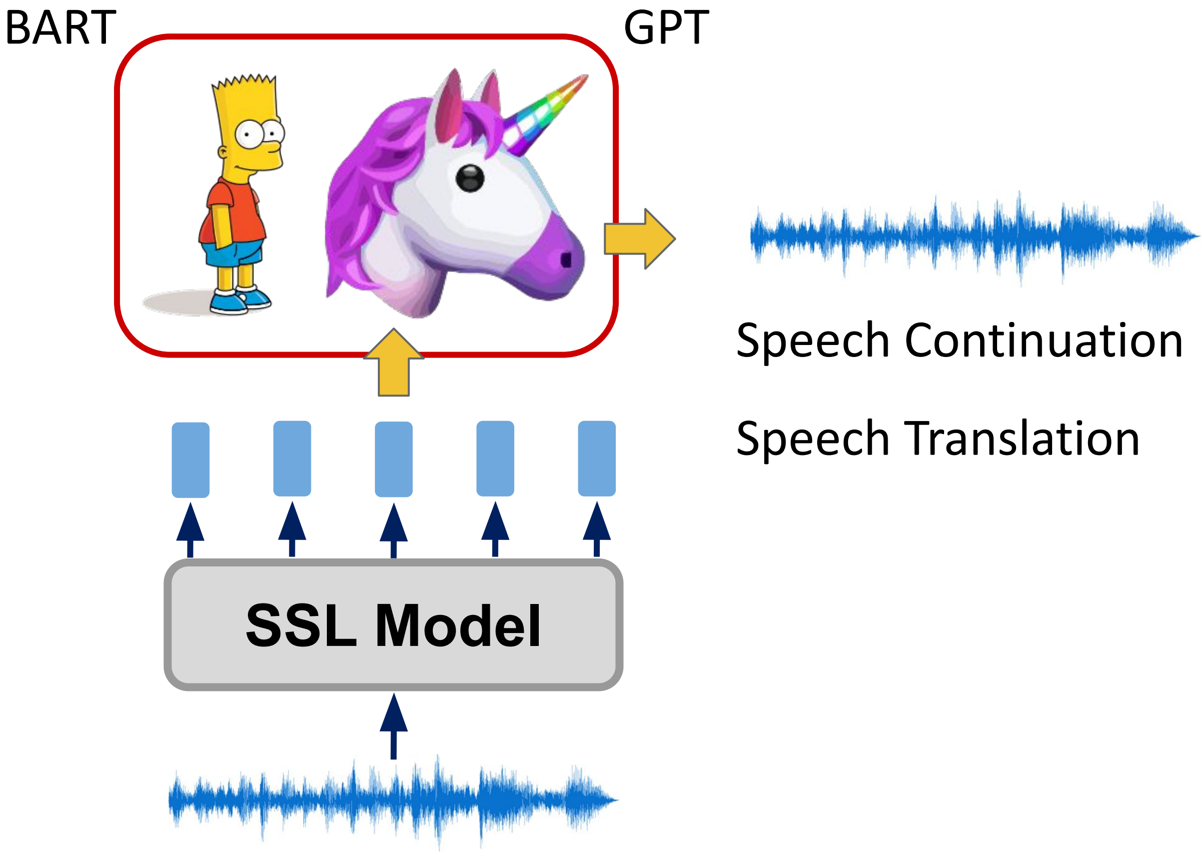


Overview

Part 2

Speech
Large Language Models

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

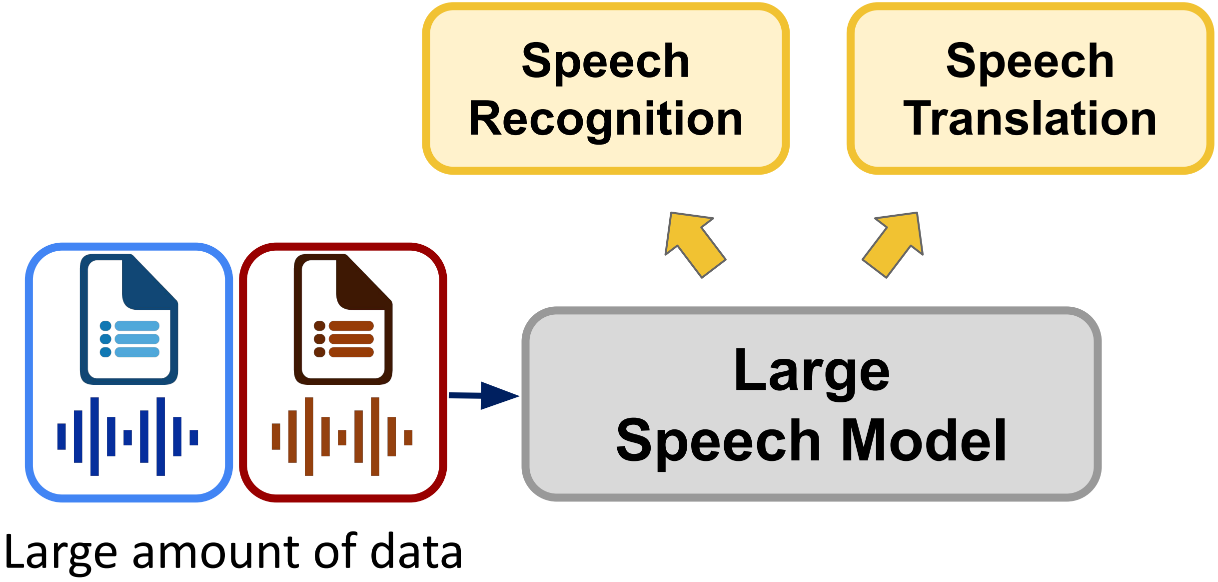


Overview

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

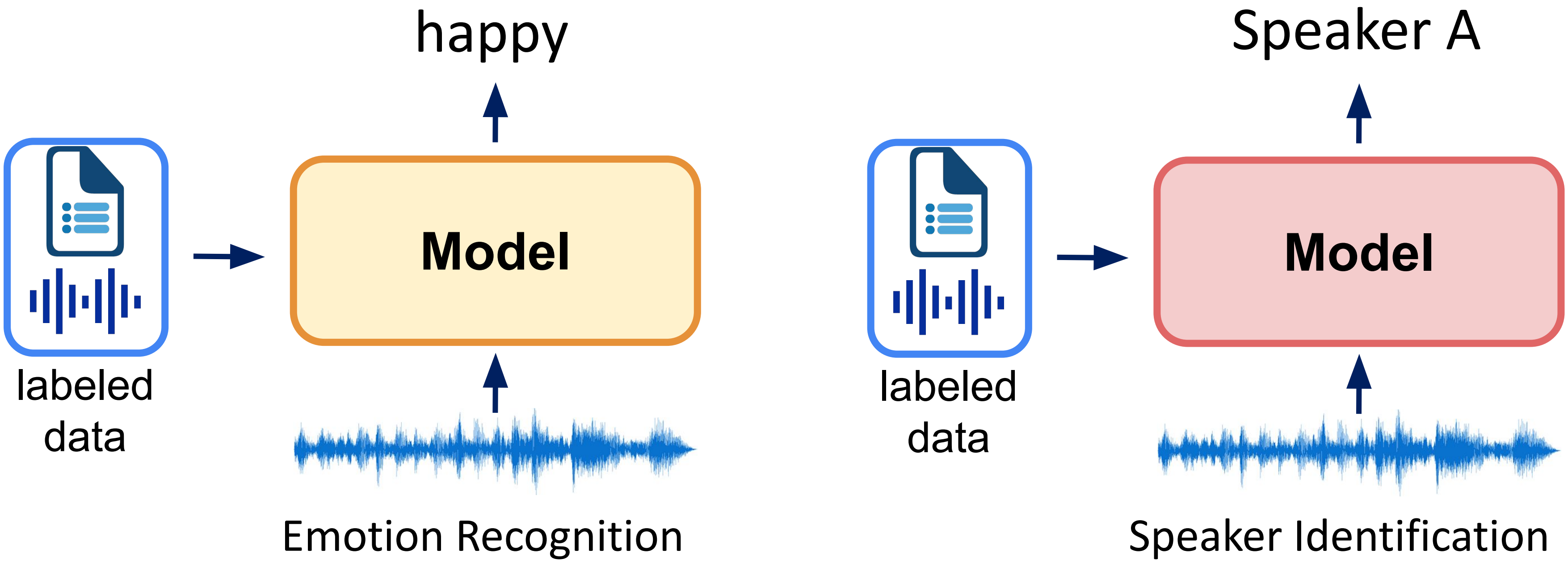
Part 3

Other Speech Foundation Models

1. Whisper
2. USM

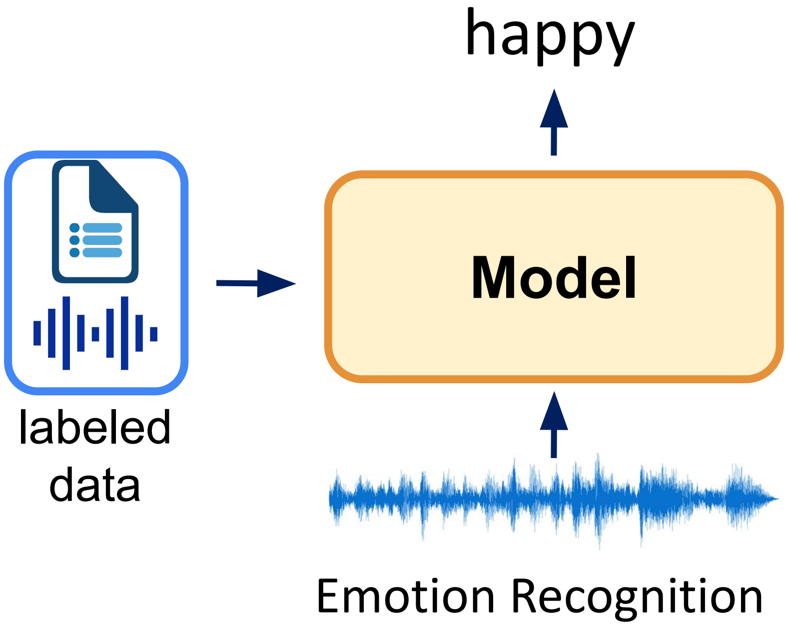
Speech Representation Learning

Why speech representation learning?



Speech Representation Learning

Why speech representation learning?

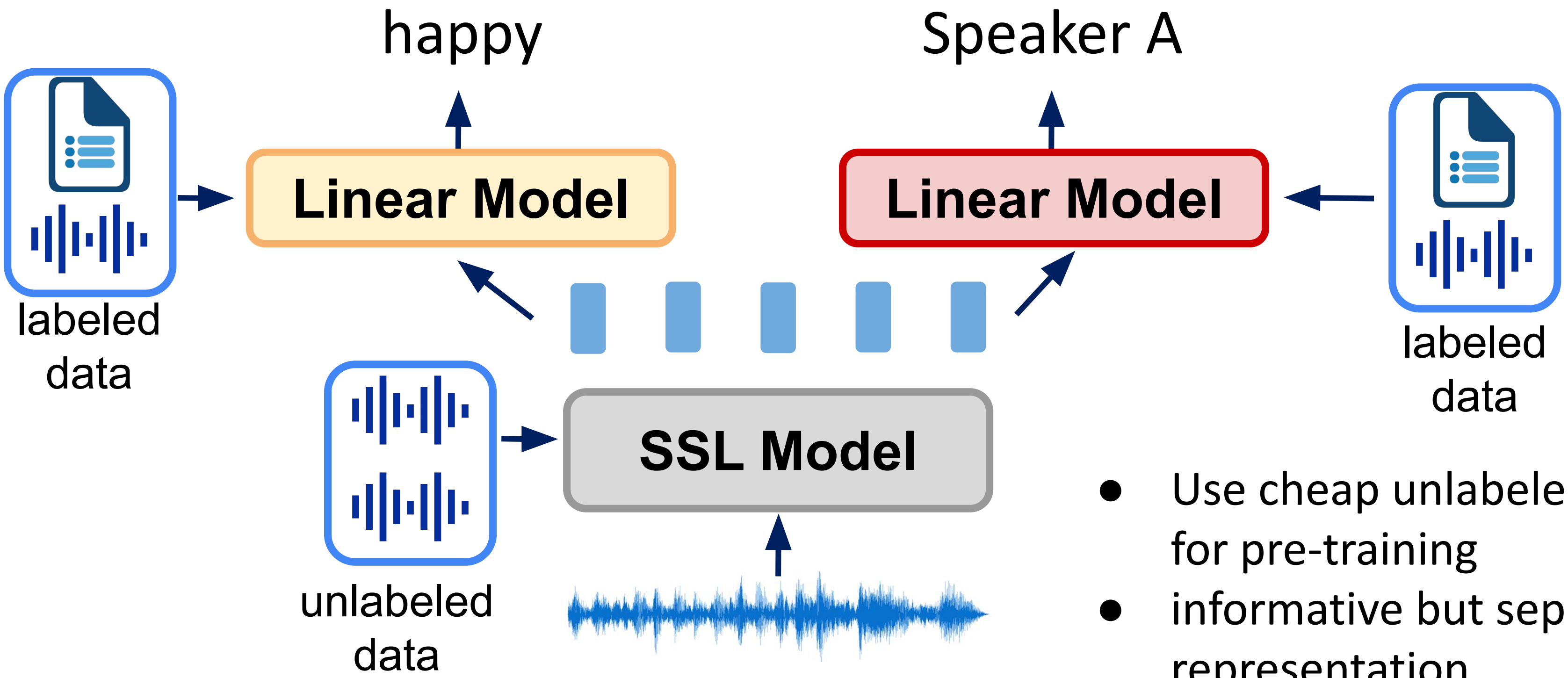


Fully supervised learning

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

Speech Representation Learning

Why speech representation learning?



- Use cheap unlabeled data for pre-training
- informative but separable representation

Overview

Speech Foundation Models

Part 1

Speech Representation Learning

1. SSL Models
2. Representation benchmarking

Part 2

Speech Large Language Models

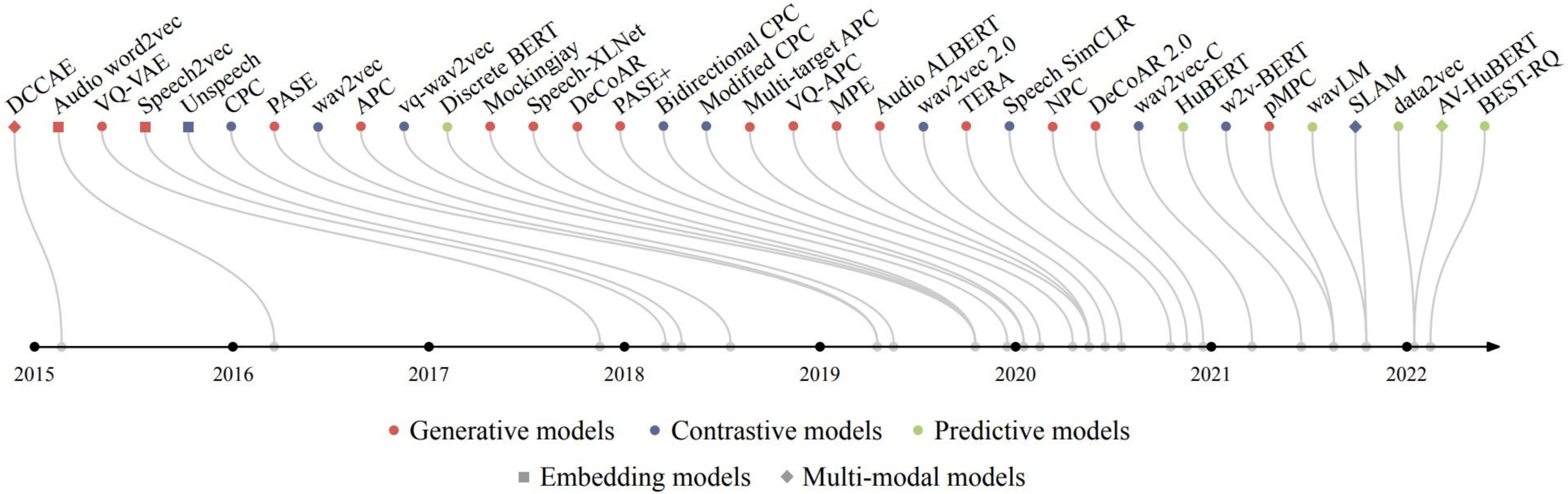
1. Textless NLP
2. AudioLM
3. Regeneration Framework

Part 3

Other Speech Foundation Models

1. Whisper
2. USM

Self-Supervised Speech Representation Learning



Mohamed, Abdelrahman, et al. "Self-supervised speech representation learning: A review." *IEEE Journal of Selected Topics in Signal Processing* (2022).

SSL Speech Representation Learning Models

Contrastive Models

- CPC
- wav2vec 2.0
- XLS-R

Predictive Models

- HuBERT
- WavLM
- BEST-RQ

SSL Speech Representation Learning Models

CPC

wav2vec 2.0

XLS-R

Contrastive Models

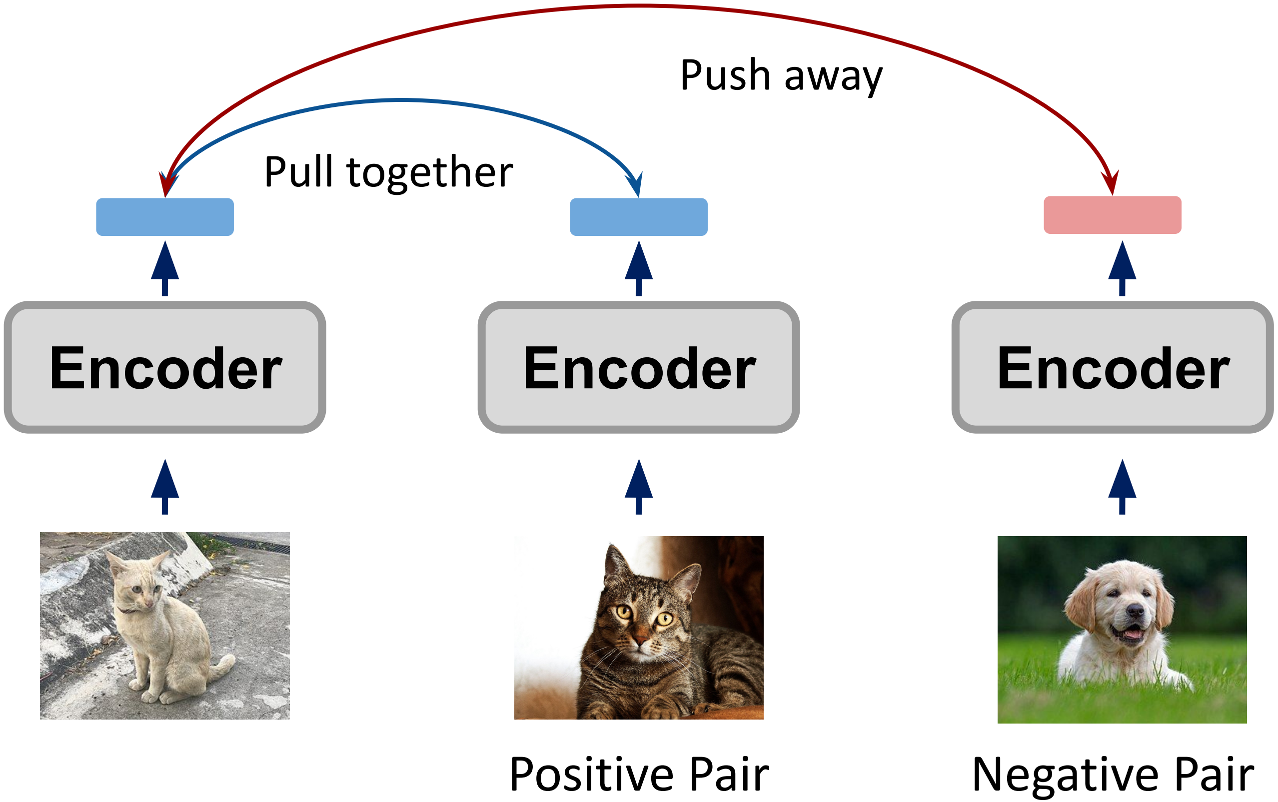
HuBERT

WavLM

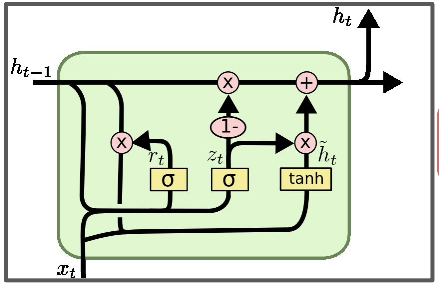
BEST-RQ

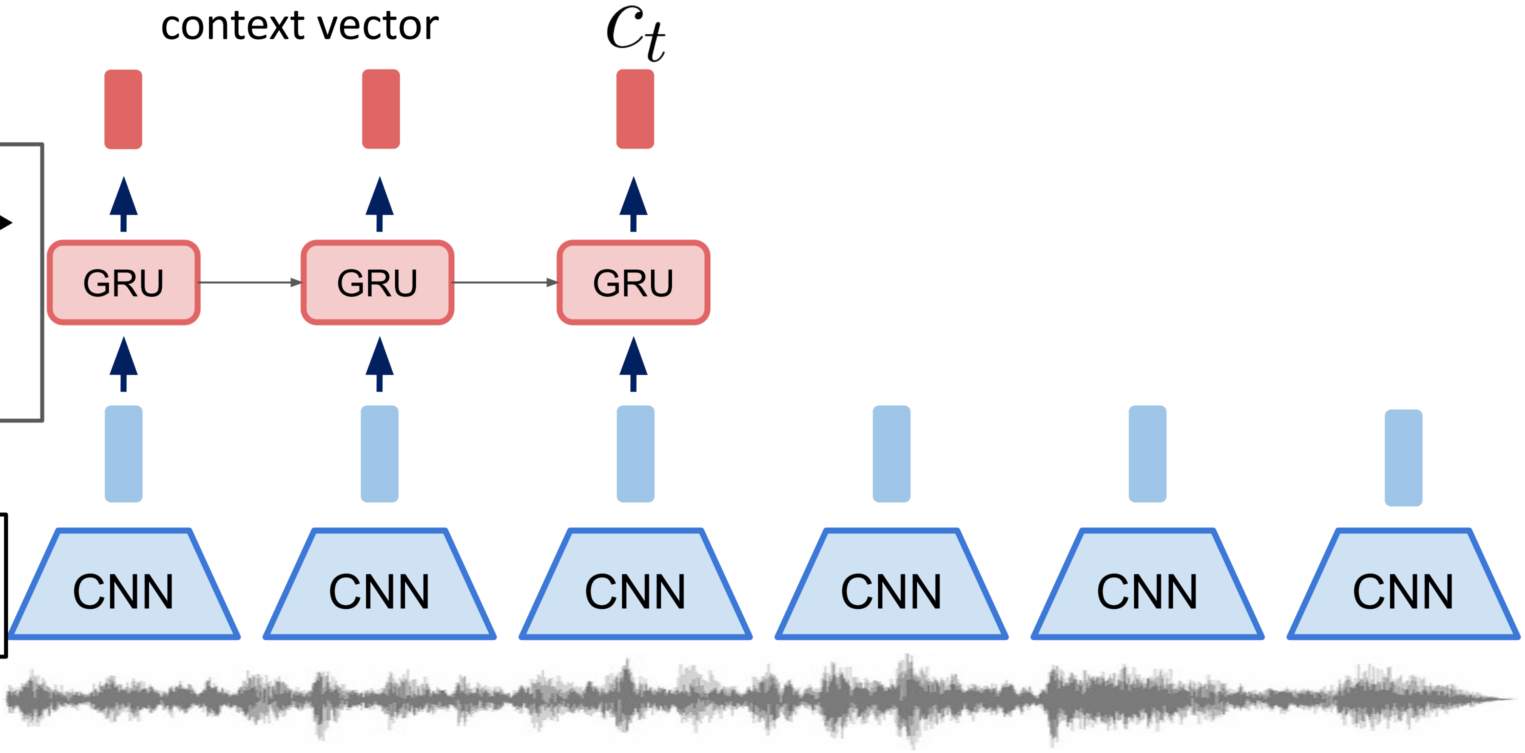
Predictive Models

Contrastive Learning: Intuition



Contrastive Predictive Coding (CPC)



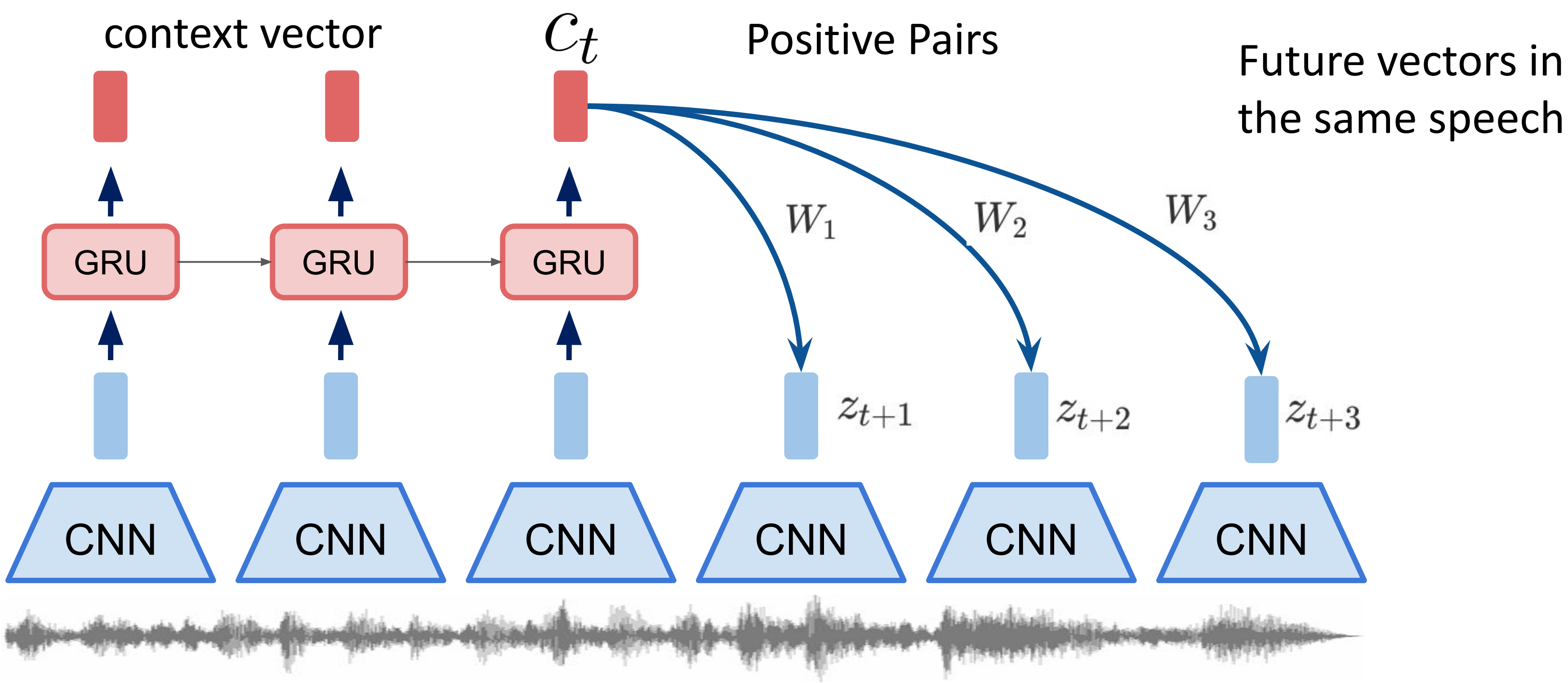


5 CNN layers
Stride: [5, 4, 2, 2, 2]

down sampling rate: 160

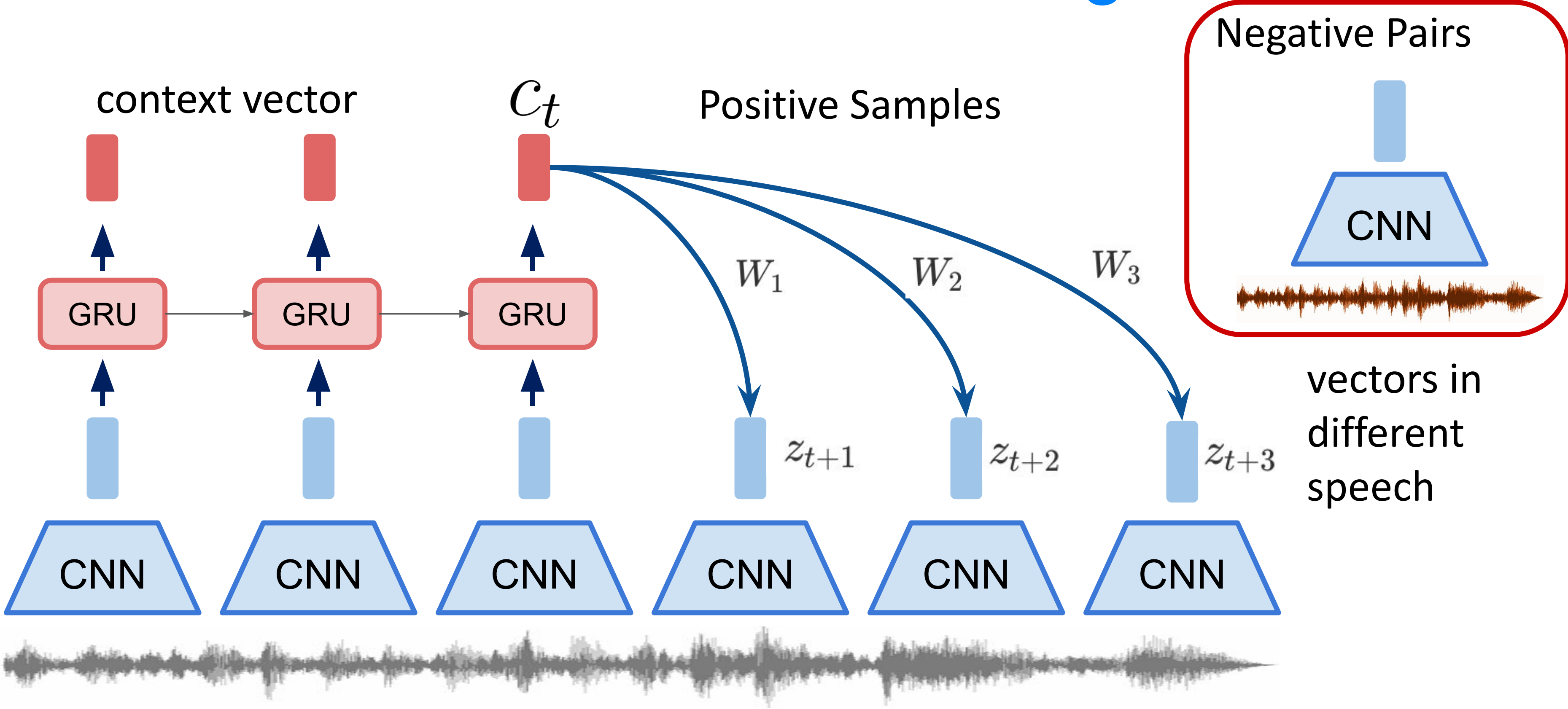
speech sampling rate: 16,000 Hz

Contrastive Predictive Coding (CPC)



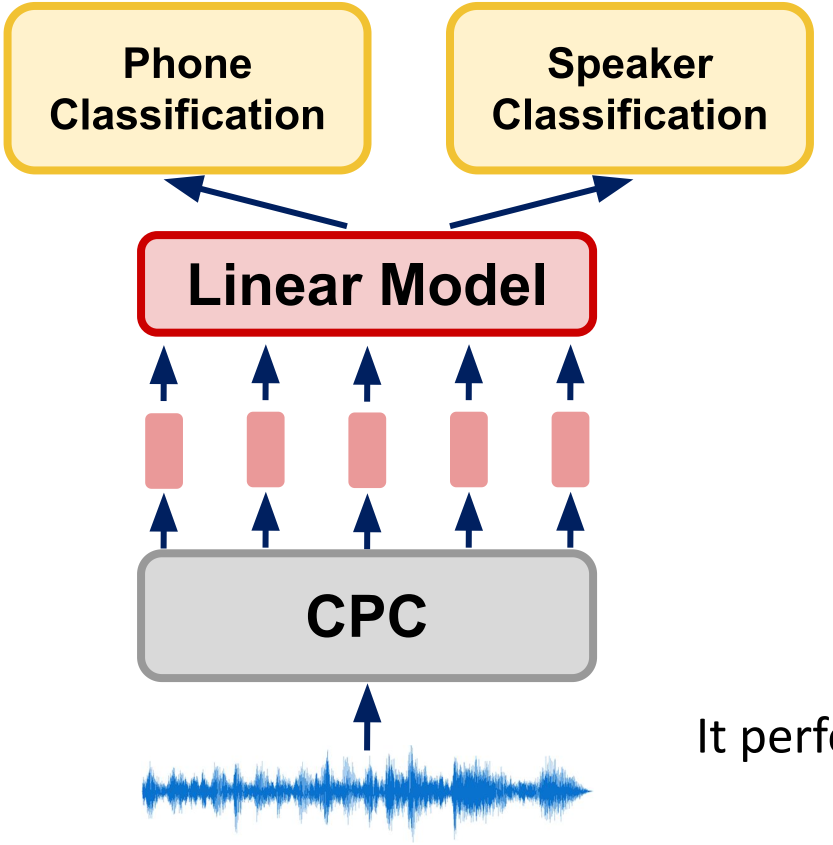
speech sampling rate: 16,000 Hz

Contrastive Predictive Coding (CPC)



speech sampling rate: 16,000 Hz

Contrastive Predictive Coding (CPC)



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

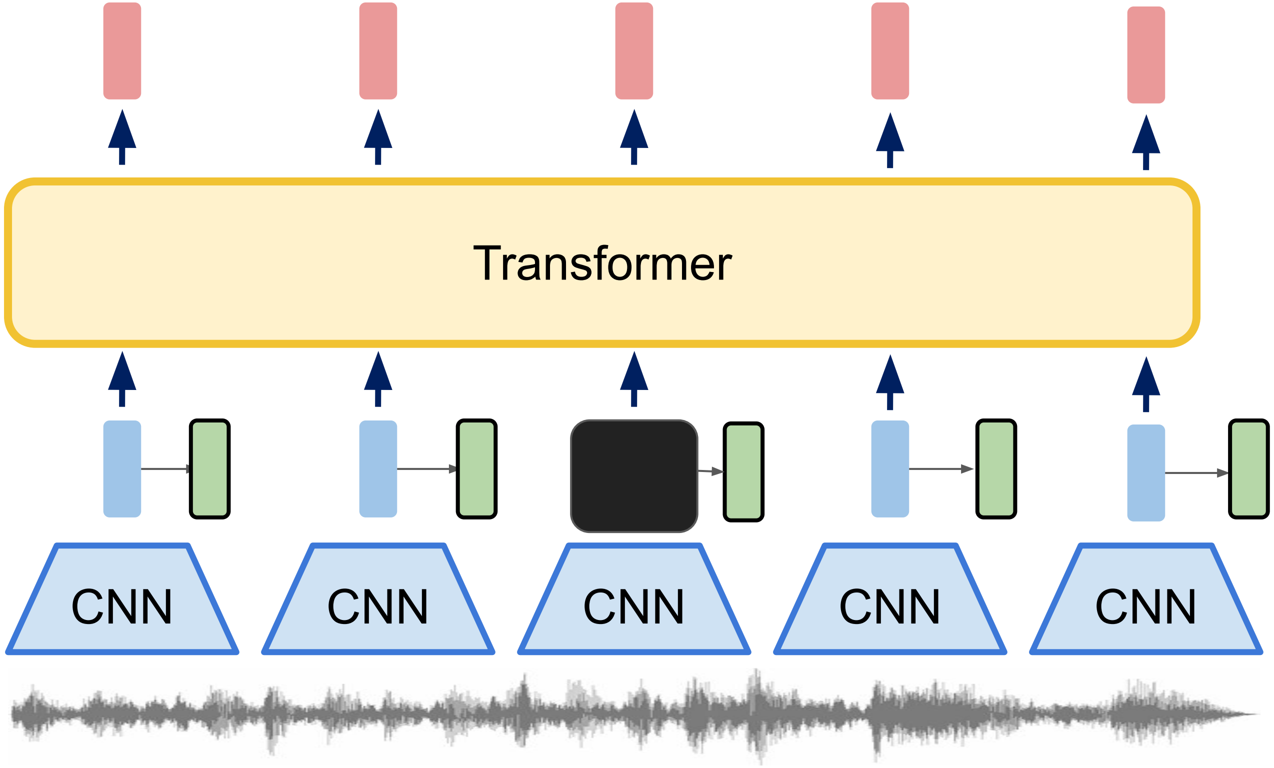
It performs well on both content and speaker tasks!

SSL Speech Representation Learning Models

Contrastive Models	Predictive Models
CPC	HuBERT
wav2vec 2.0	WavLM
XLS-R	BEST-RQ

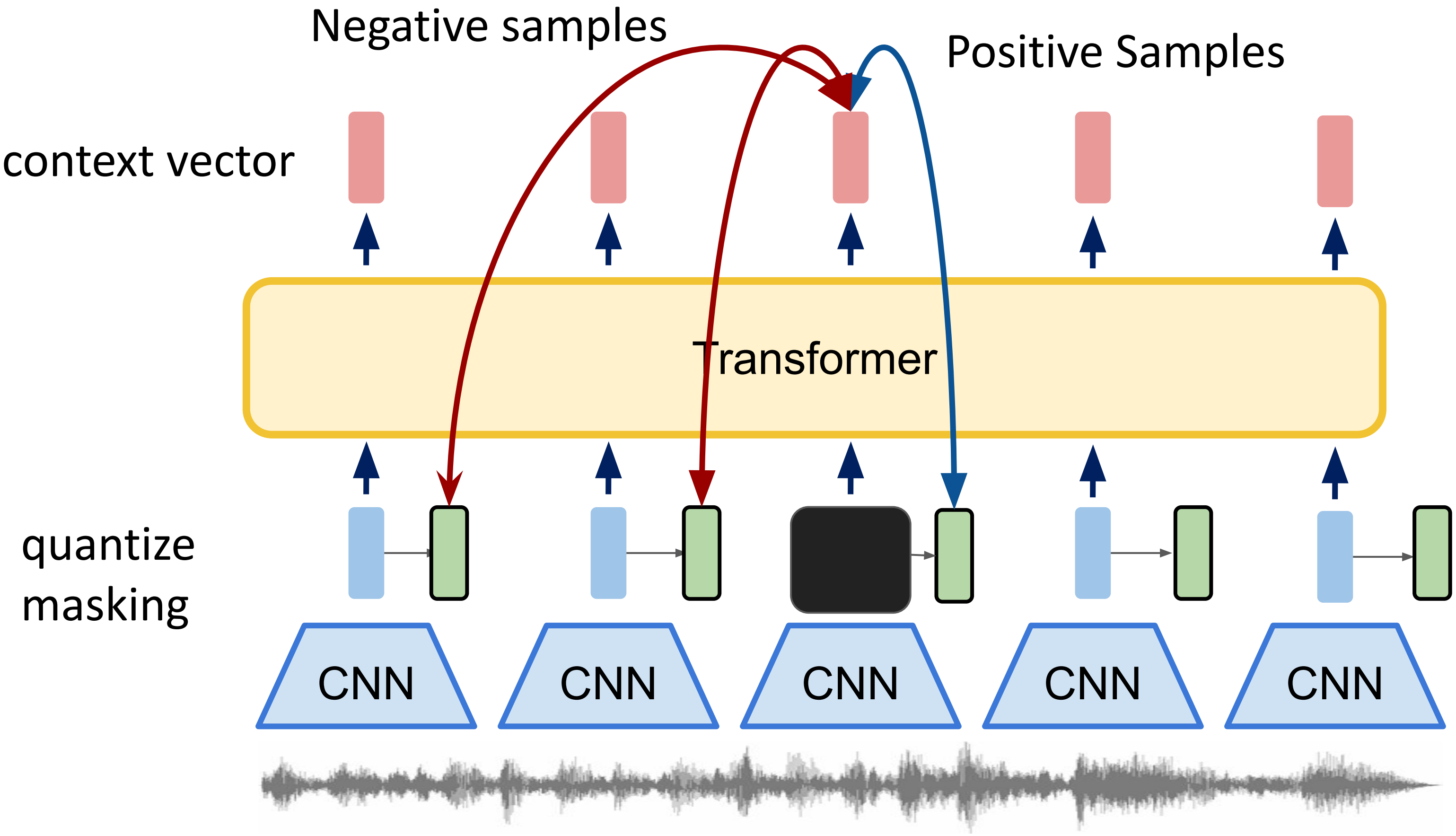
Wav2vec 2.0

context vector



- quantize
- masking

Wav2vec 2.0



- quantize
- masking

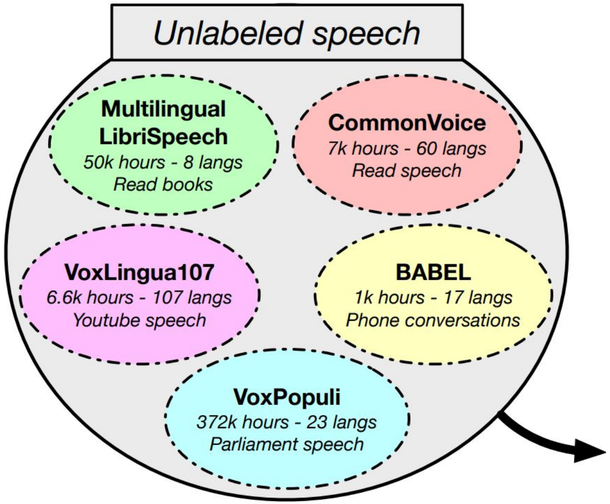
SSL Speech Representation Learning Models

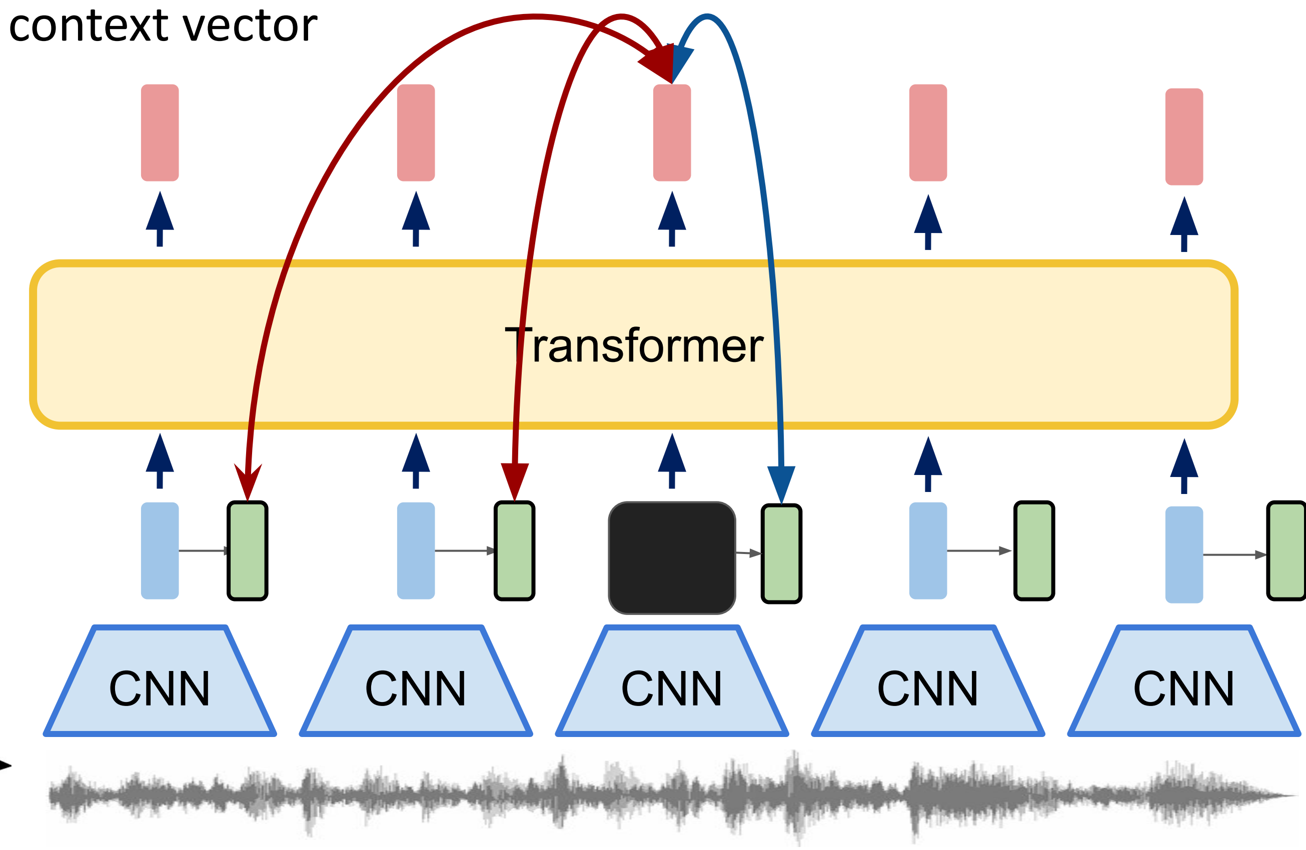
Contrastive Models
- CPC
- wav2vec 2.0
- **XLS-R**

Predictive Models
- HuBERT
- WavLM
- BEST-RQ

XLS-R

~ 400,000 hours
107 languages





LibriSpeech ASR results

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

Model	dev clean	dev other	test clean	test other
10 min labeled				
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)	**28.4**	**32.5**	**29.1**	**32.5**
1h labeled				
wav2vec 2.0 LV-60K (0.3B)	13.7	**16.9**	13.7	**17.1**
XLS-R (0.3B)	17.1	23.7	16.8	24.0
XLS-R (1B)	**13.2**	17.0	**13.1**	17.2
10h labeled				
wav2vec 2.0 LV-60K (0.3B)	**5.7**	**9.2**	**5.6**	**9.4**
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

- Achieve competitive performance when model size is bigger

Model	dev clean	dev other	test clean	test other
10 min labeled				
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)	**28.4**	**32.5**	**29.1**	**32.5**
1h labeled				
wav2vec 2.0 LV-60K (0.3B)	13.7	**16.9**	13.7	**17.1**
XLS-R (0.3B)	17.1	23.7	16.8	24.0
XLS-R (1B)	**13.2**	17.0	**13.1**	17.2
10h labeled				
wav2vec 2.0 LV-60K (0.3B)	**5.7**	**9.2**	**5.6**	**9.4**
XLS-R (0.3B)	8.3	15.1	8.3	15.4
XLS-R (1B)	5.9	10.5	5.9	10.6

Multilingual LibriSpeech

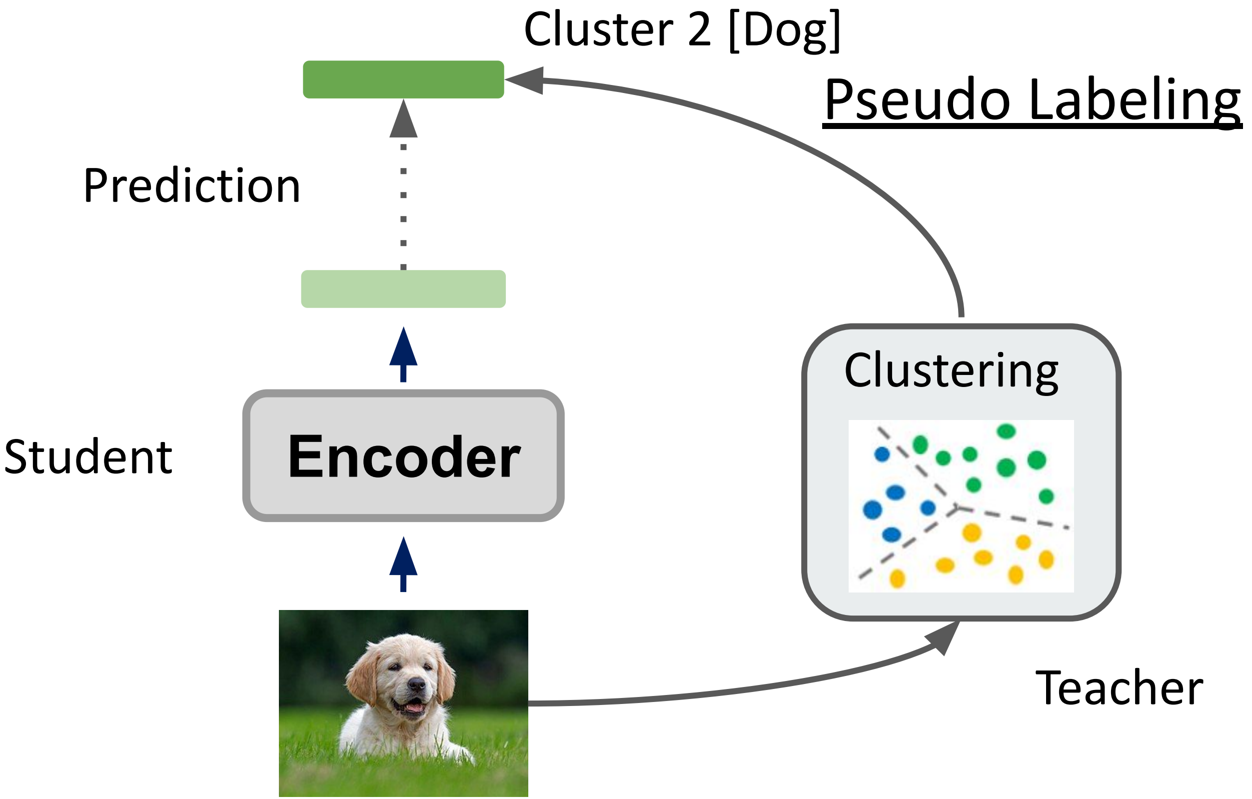
	#ft	en	de	nl	fr	es	it	pt	pl*	Avg.
Full labeled data (h)		44.7K	2K	1.6K	1.1K	918	247	161	104	
Previous work										
Pratap et al. (2020)	full	**5.9**	**6.5**	12.0	**5.6**	**6.1**	**10.5**	19.5	19.4	**10.7**
XLSR-53	10h	14.6	8.4	12.8	12.5	8.9	13.4	18.2	17.8	13.8
This work										
XLS-R (0.3B)	10h	15.9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

Achieve competitive performance when only using 10 hour labeled data

SSL Speech Representation Learning Models

Contrastive Models	Predictive Models
CPC	HuBERT
wav2vec 2.0	WavLM
XLS-R	BEST-RQ

Predictive

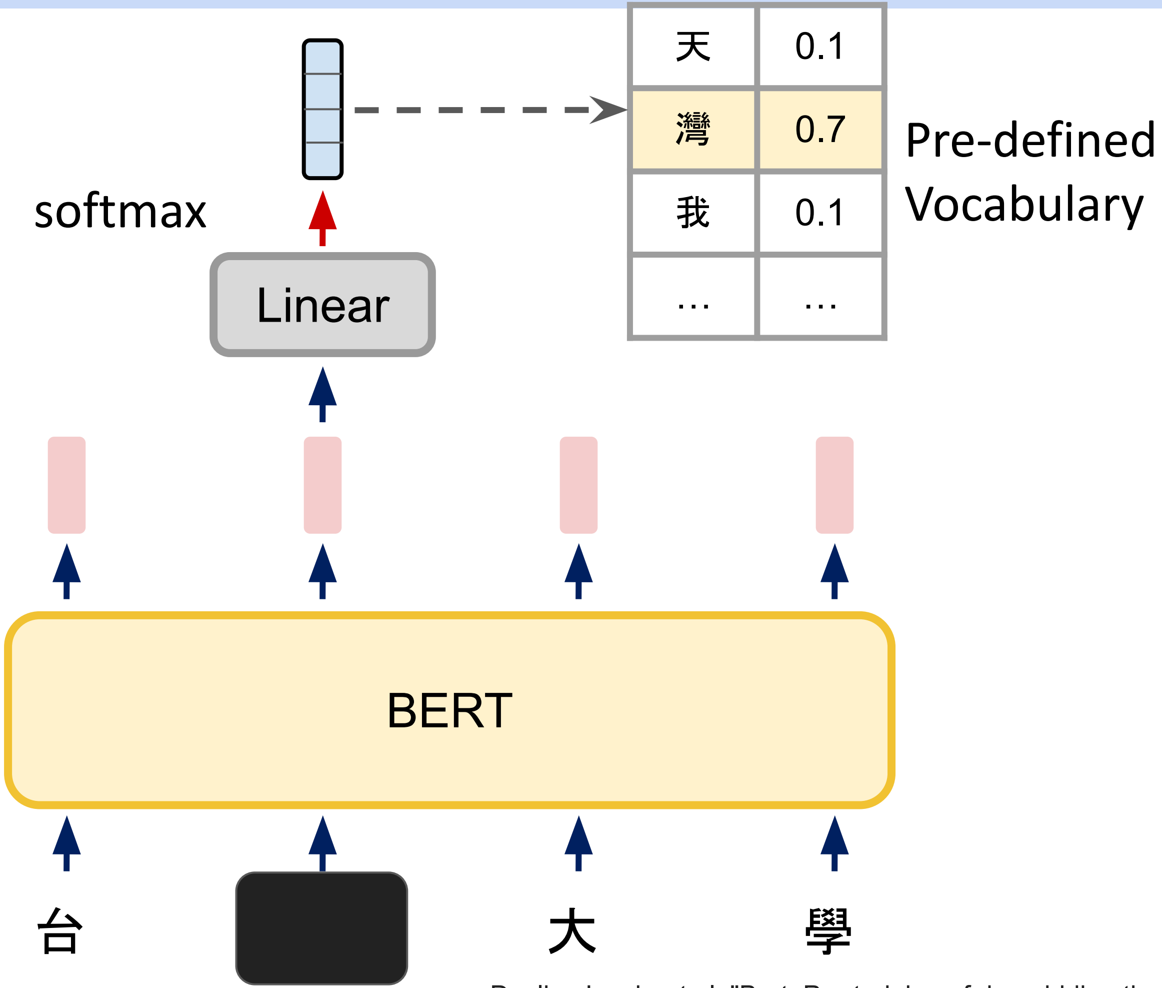

Cluster 2 [Dog]
Pseudo Labeling
Prediction
Student
Encoder
Clustering
Teacher

SSL Speech Representation Learning Models

Contrastive Models	Predictive Models
CPC	**HuBERT**
wav2vec 2.0	WavLM
XLS-R	BEST-RQ

BERT

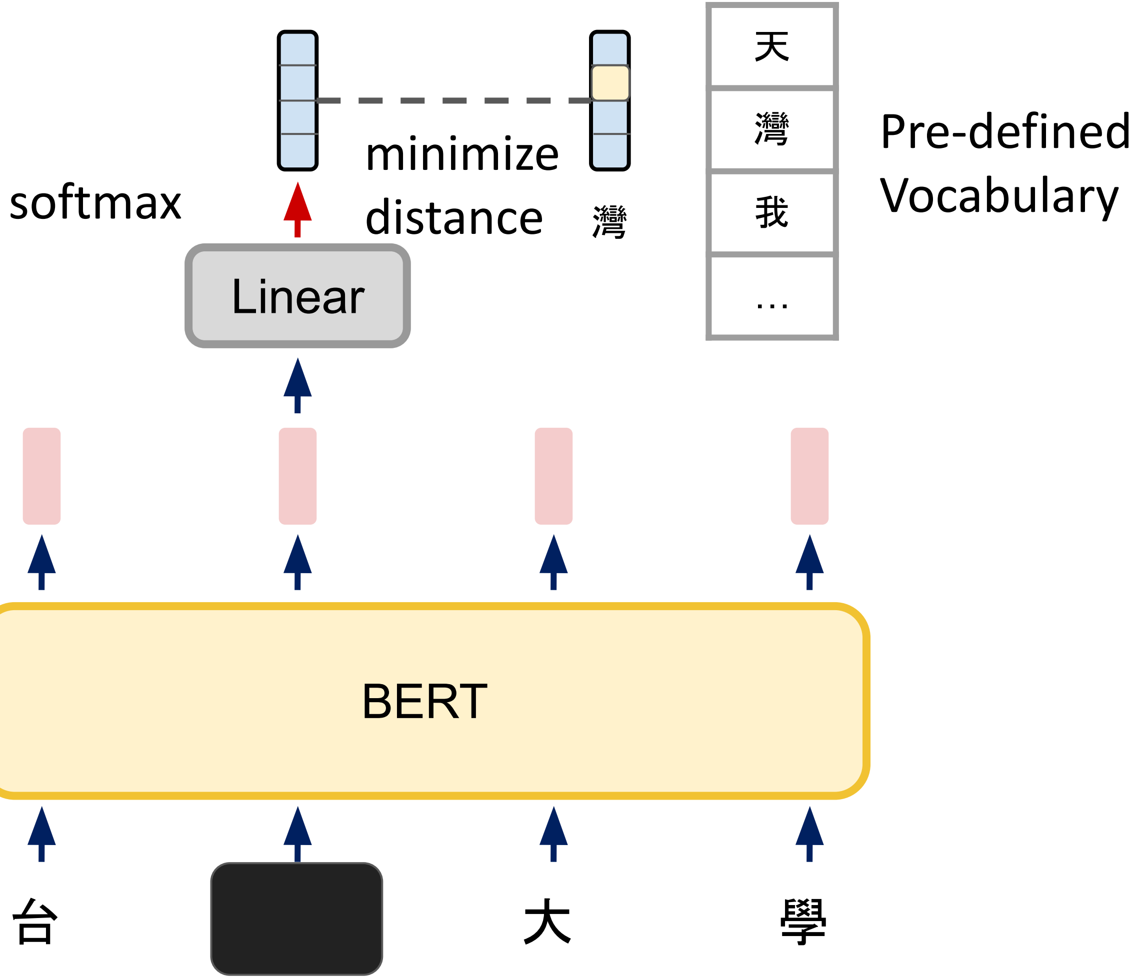
Maksed Language Modeling



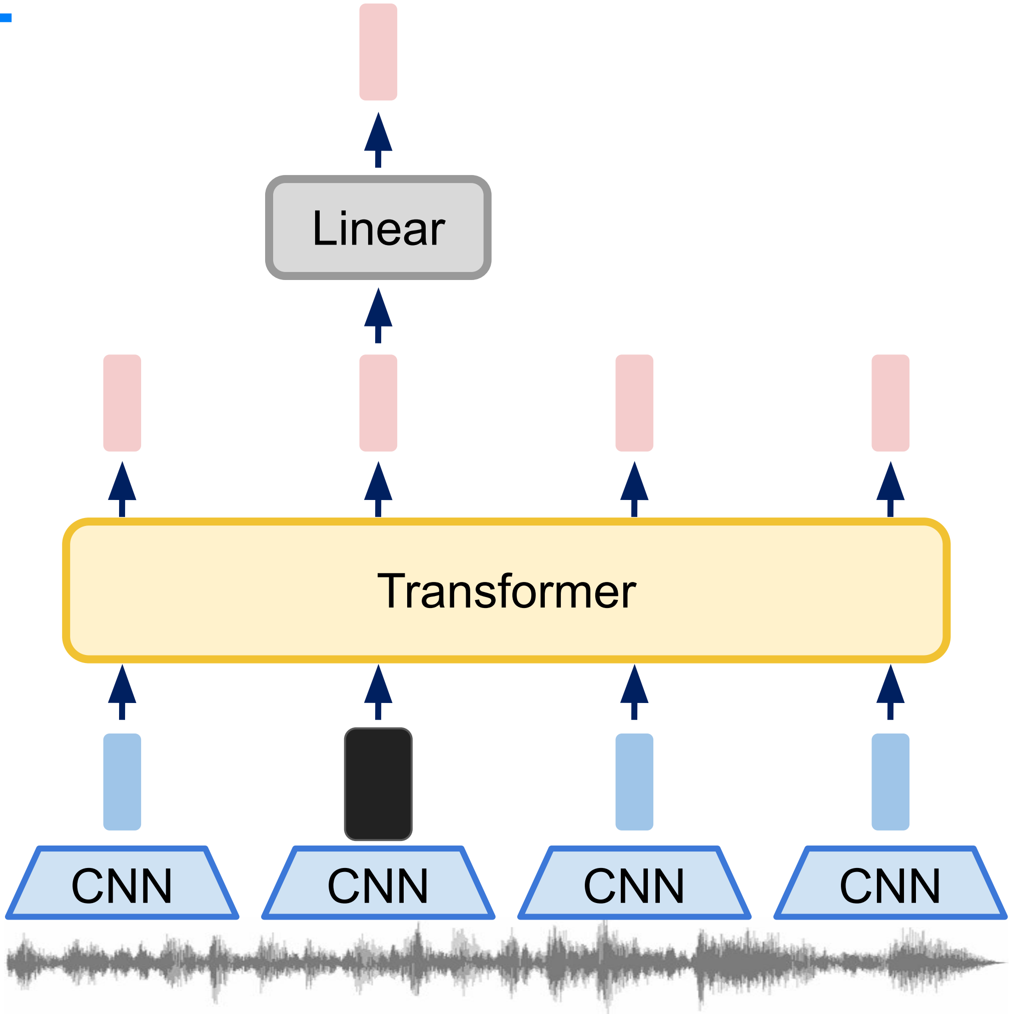
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

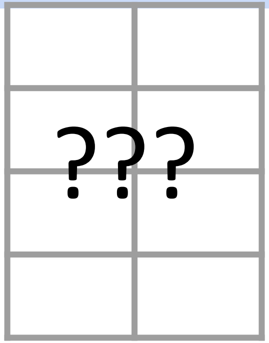
BERT

Maksed Language Modeling



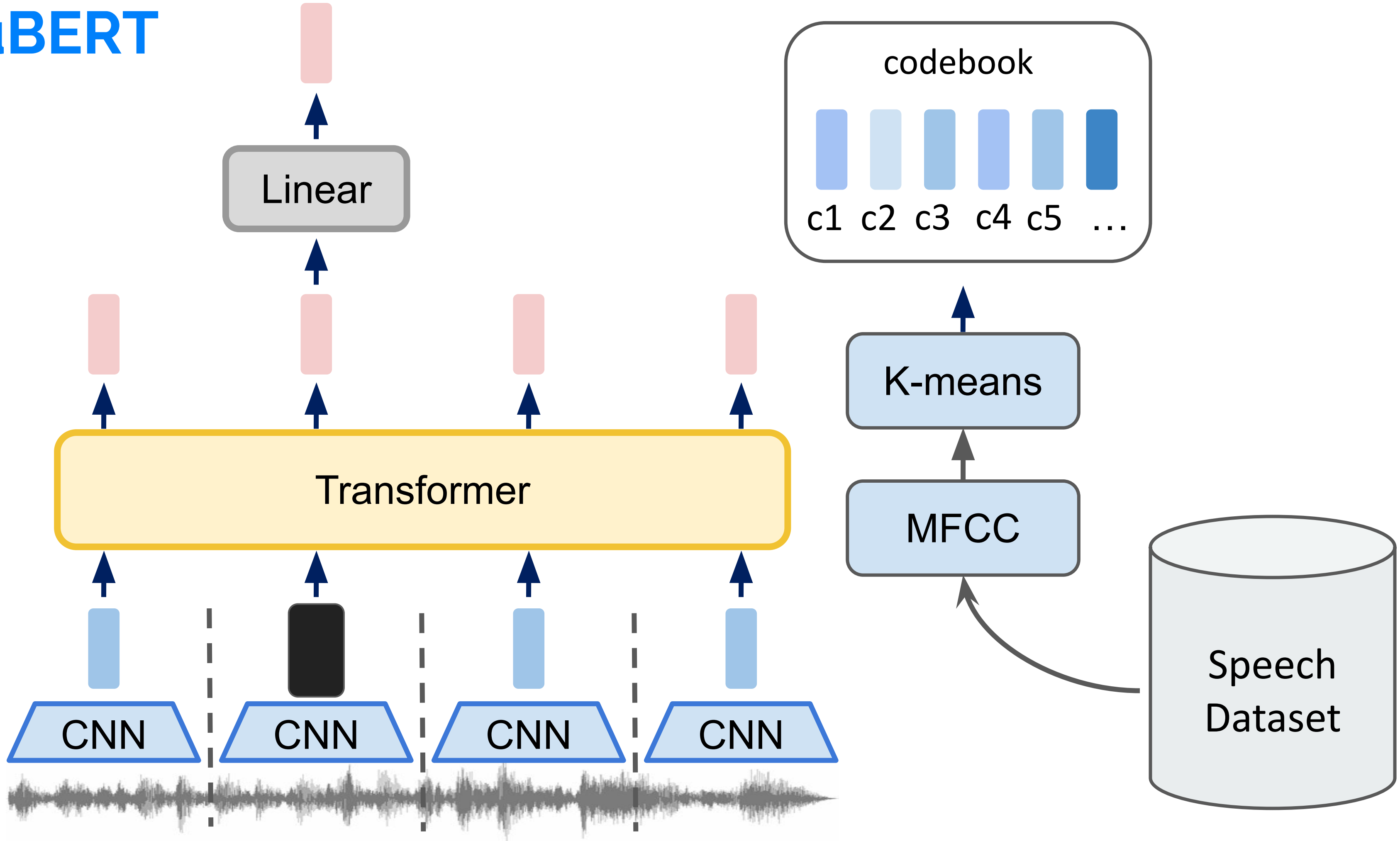
HuBERT





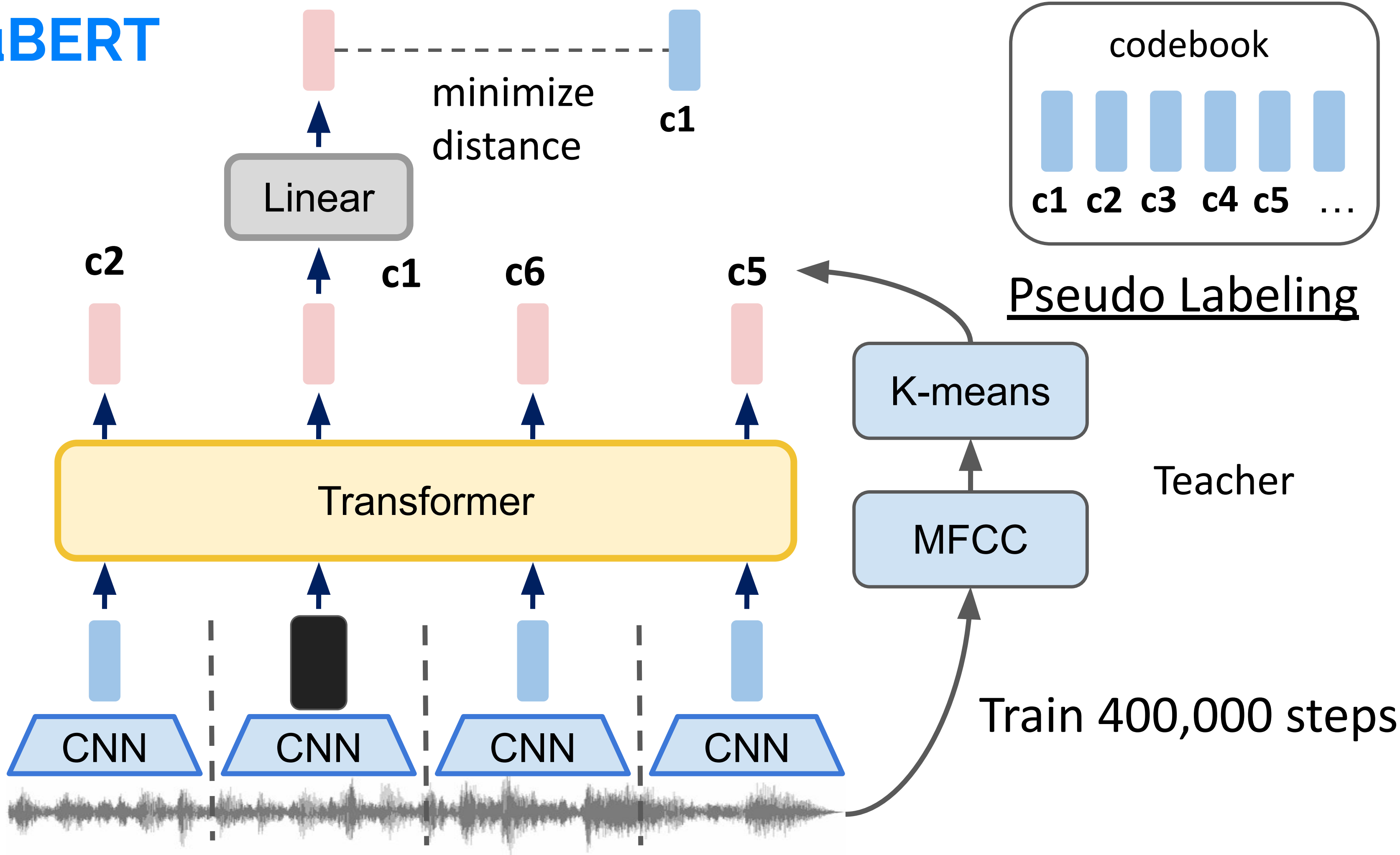
Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 3451-3460.

HuBERT



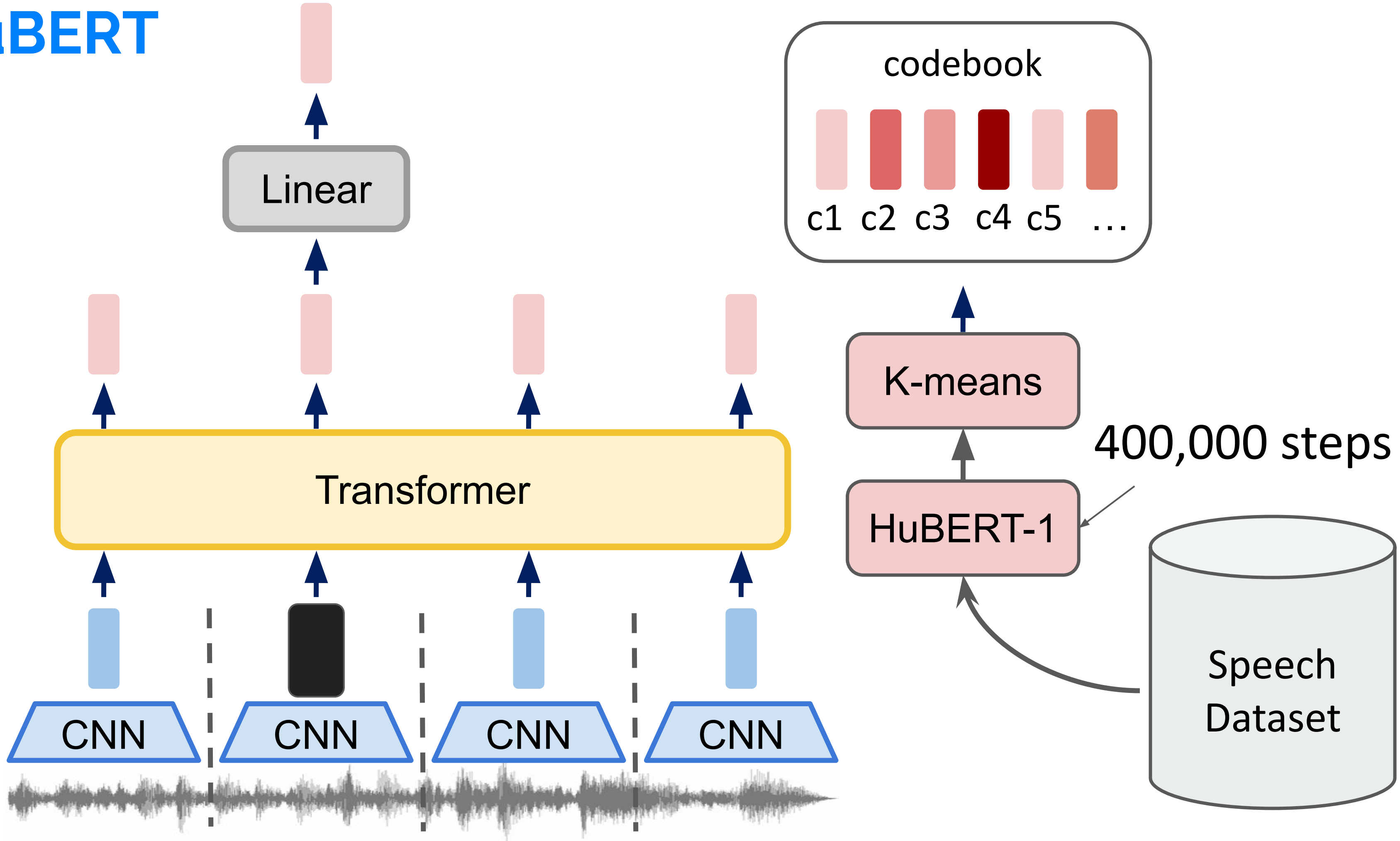
codebook
c1 c2 c3 c4 c5 ...
Linear
K-means
Transformer
MFCC
Speech Dataset
CNN
CNN
CNN
CNN

HuBERT

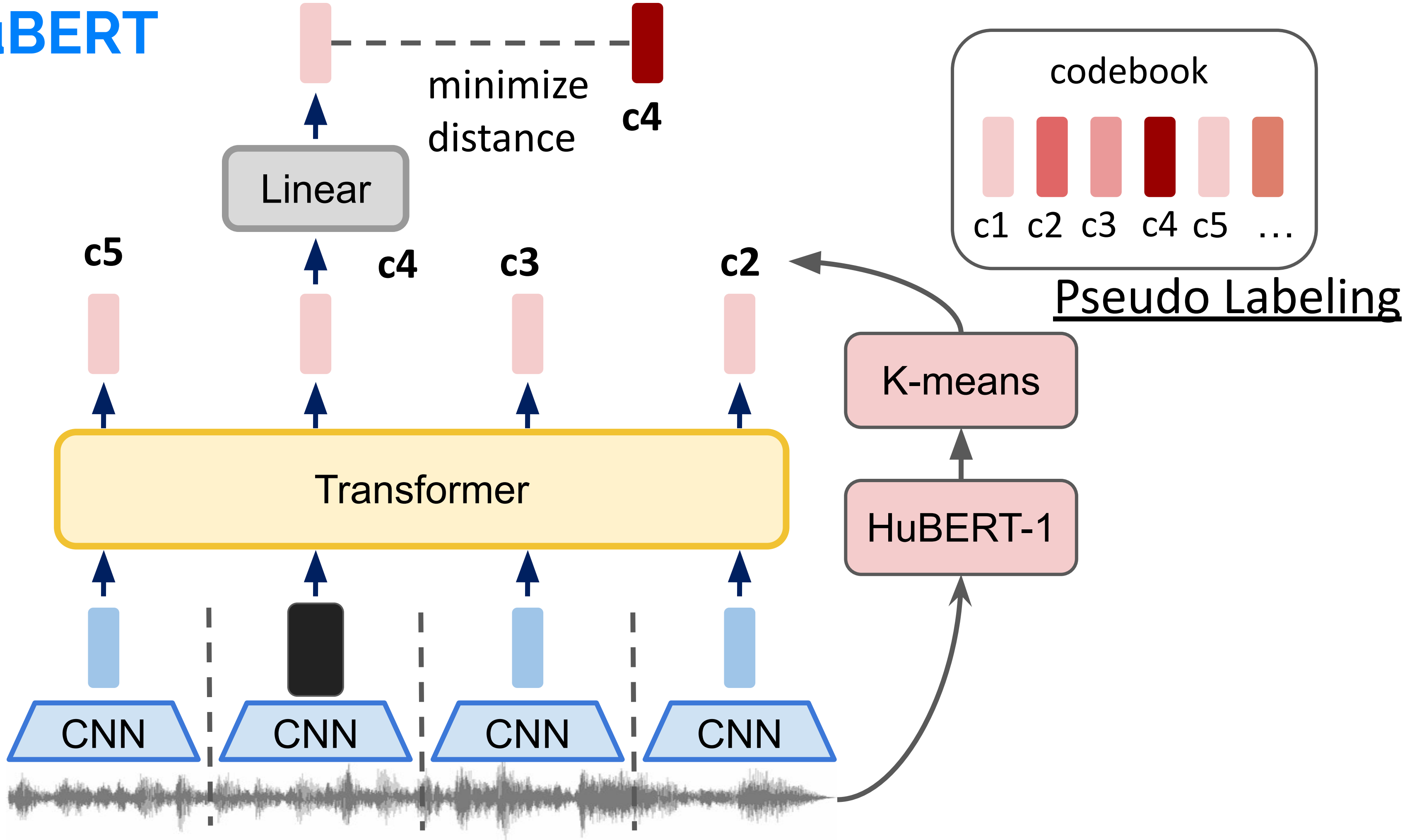


minimize
distance
c1
codebook
c1 c2 c3 c4 c5 …
Linear
c2
c1
c6
c5
Pseudo Labeling
K-means
Teacher
Transformer
MFCC
Train 400,000 steps
CNN
CNN
CNN
CNN

HuBERT


codebook
c1 c2 c3 c4 c5 …
Linear
K-means
400,000 steps
HuBERT-1
Transformer
Speech
Dataset
CNN
CNN
CNN
CNN

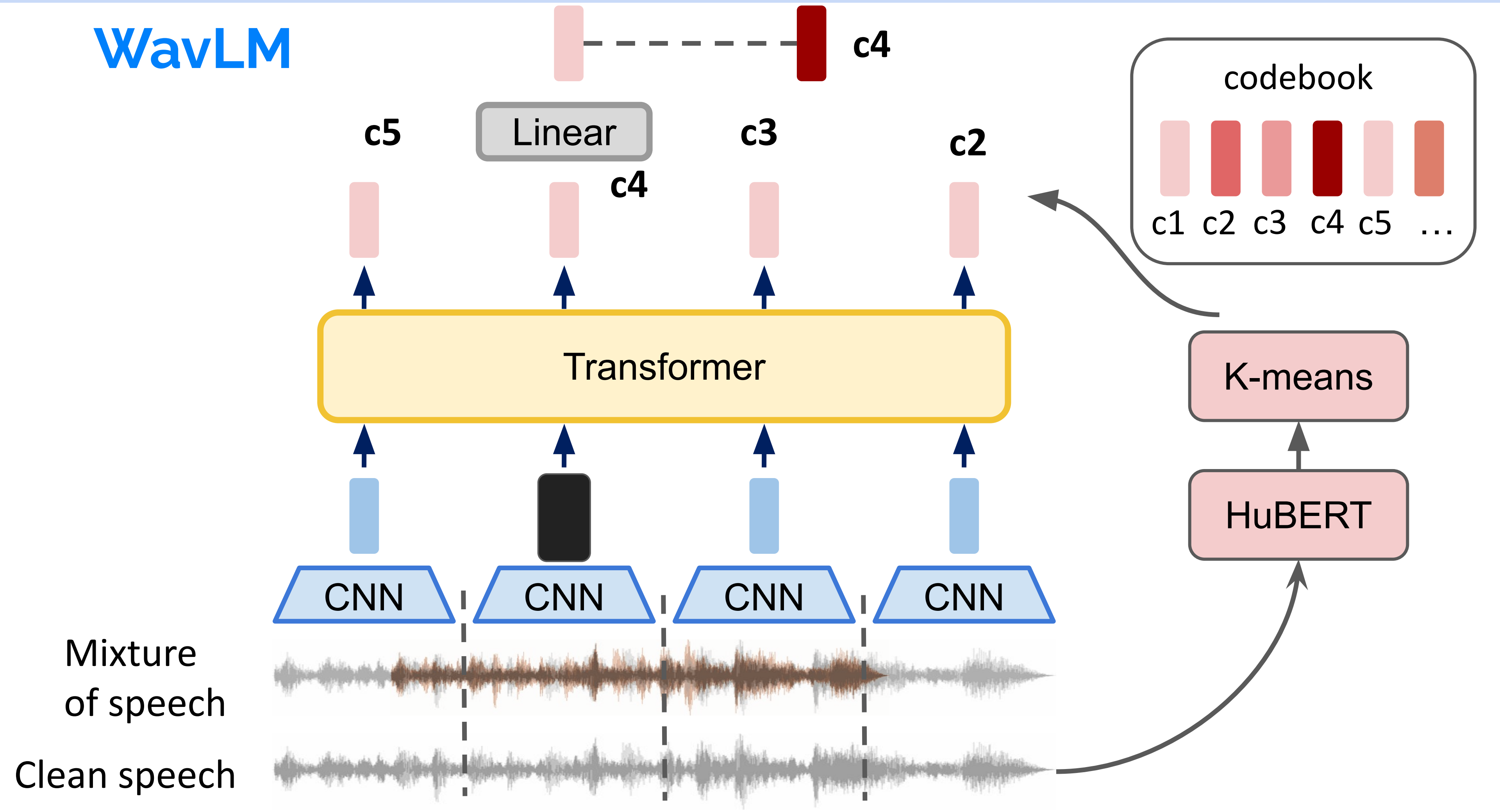
HuBERT



minimize
distance
c4
codebook
c1 c2 c3 c4 c5 …
Pseudo Labeling
Linear
c5
c4
c3
c2
K-means
Transformer
HuBERT-1
CNN
CNN
CNN
CNN

SSL Speech Representation Learning Models

Contrastive Models	Predictive Models
CPC	HuBERT
wav2vec 2.0	**WavLM**
XLS-R	BEST-RQ



WavLM
c4
codebook
c5
Linear
c3
c2
c4
c1 c2 c3 c4 c5 …
Transformer
K-means
HuBERT
CNN
CNN
CNN
CNN
Mixture of speech
Clean speech

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*	**0.383**	**0.480**	**0.986**

EER: equal error rate, the lower the better

With speech denoising, WavLM performs better than HuBERT

SSL Speech Representation Learning Models

CPC

wav2vec 2.0

XLS-R

Contrastive Models

HuBERT

WavLM

BEST-RQ

Predictive Models

BEST-RQ BERT-based Speech Pre-training with Random Projection Quantizer

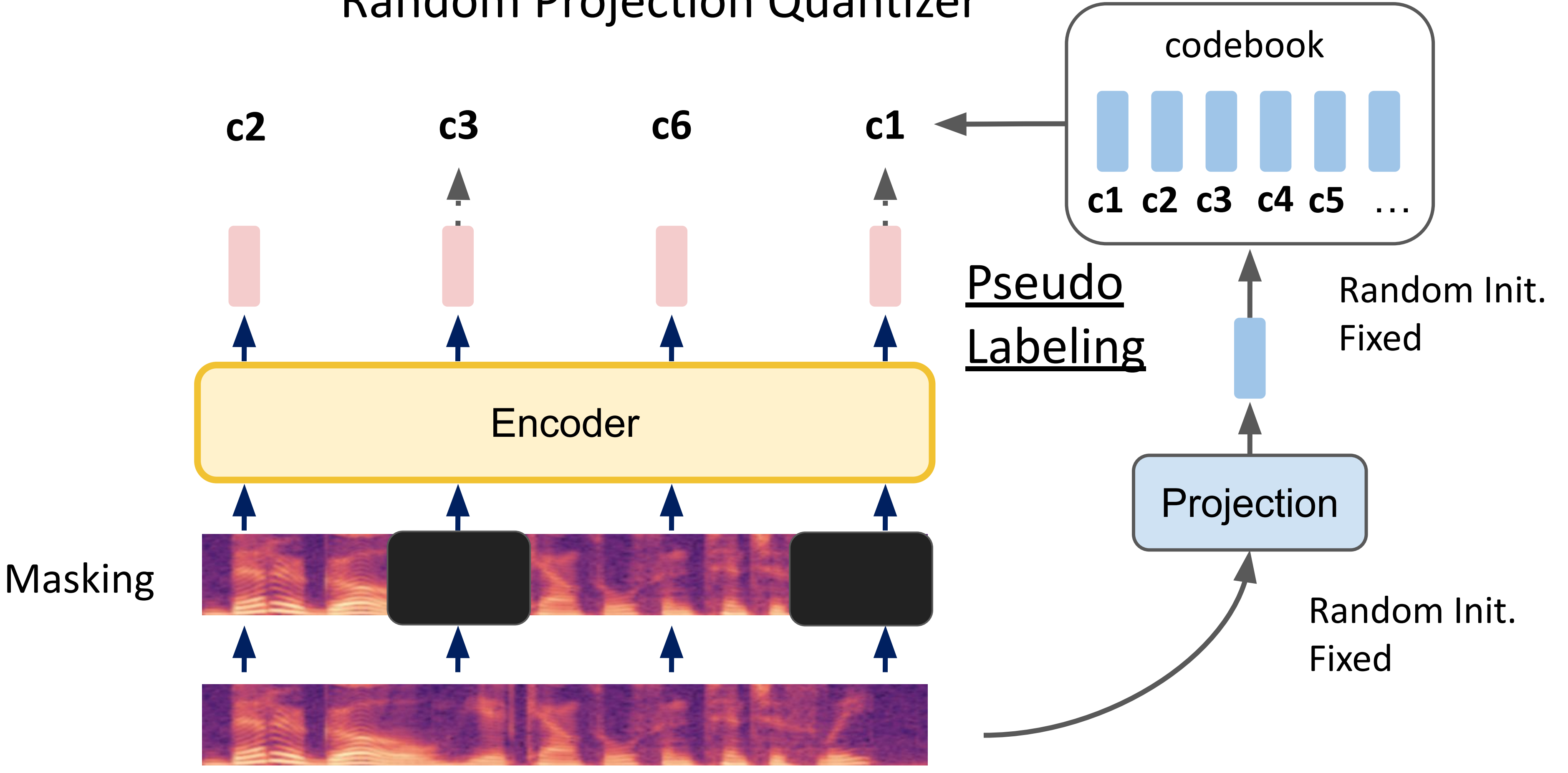


Table 1. LibriSpeech results with non-streaming models. The LM used in our experiment is a Transfomer LM with model size 0.1B.

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	**2.5**	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	**1.5**	3.2
w2v-BERT XL (Chung et al., 2021)	0.6	**1.5**	2.9	**1.5**	**2.9**	**1.4**	2.8	**1.5**	2.8
BEST-RQ (Ours)	0.6	**1.5**	**2.8**	1.6	**2.9**	**1.4**	2.6	**1.5**	**2.7**

- 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

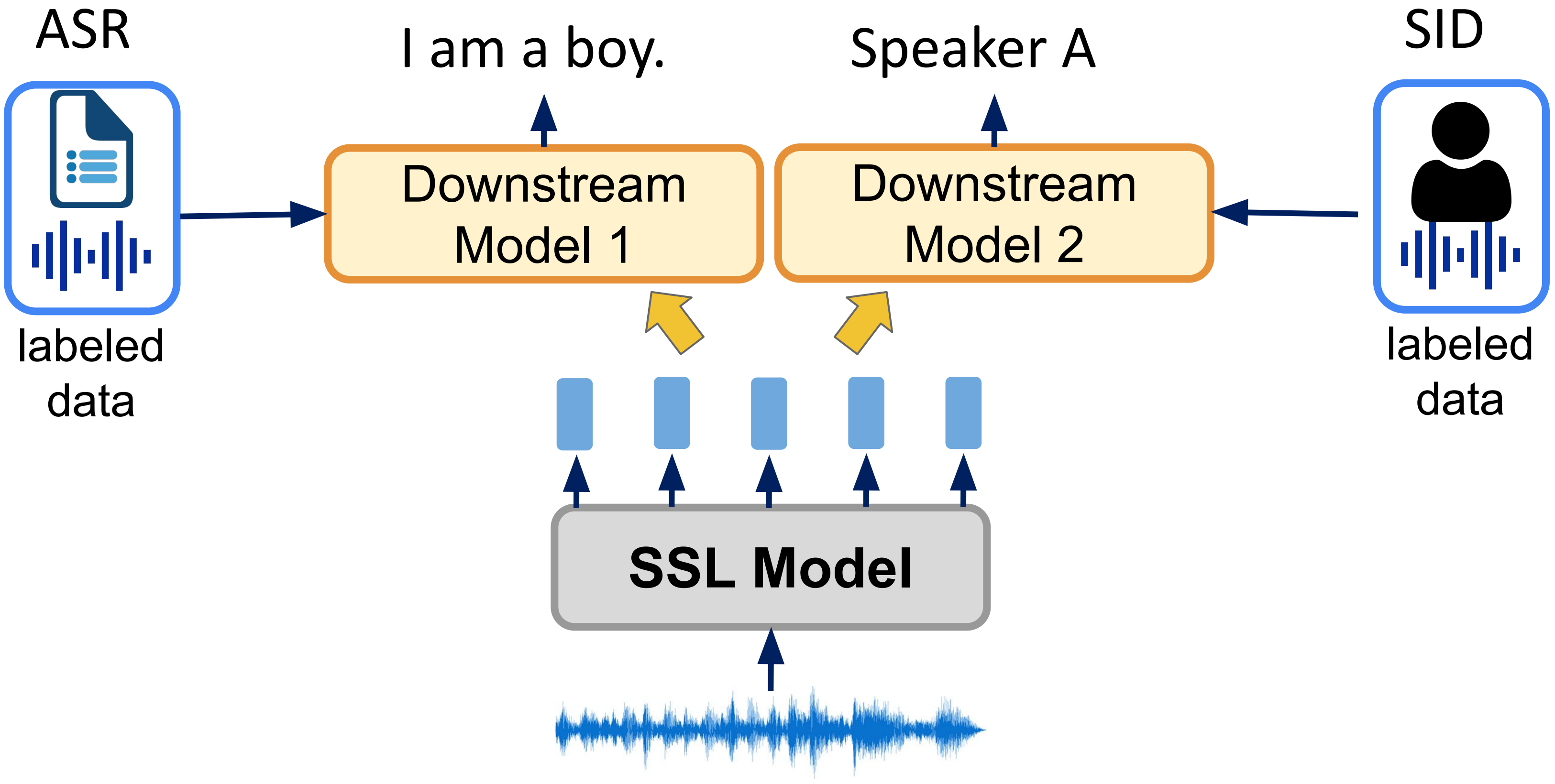
1. Textless NLP
2. AudioLM
3. Regeneration Framework

Part 3

Other Speech Foundation Models

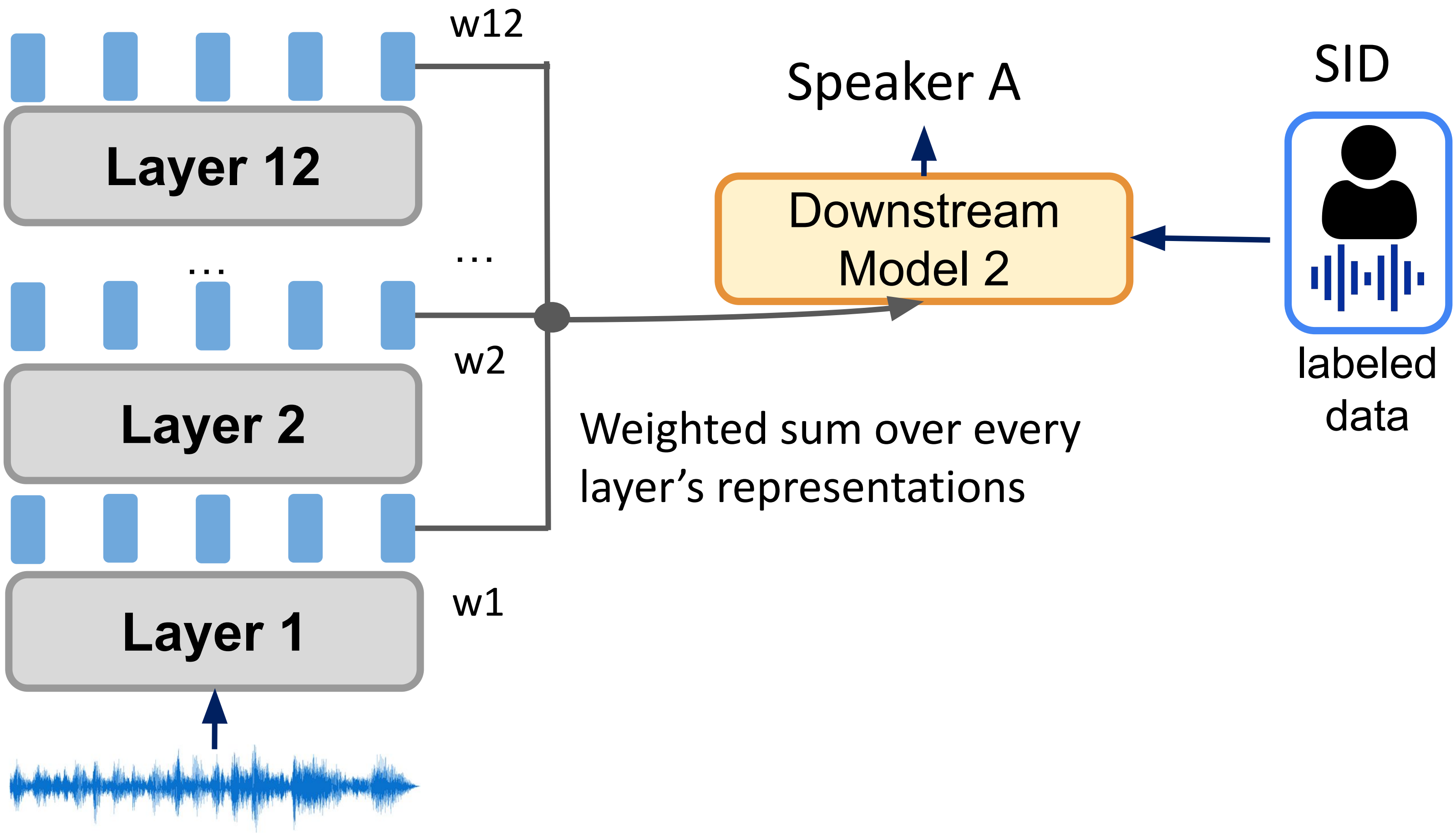
1. Whisper
2. USM

SUPERB



ASR
I am a boy.
Speaker A
SID
Downstream Model 1
Downstream Model 2
labeled data
labeled data
SSL Model

SUPERB

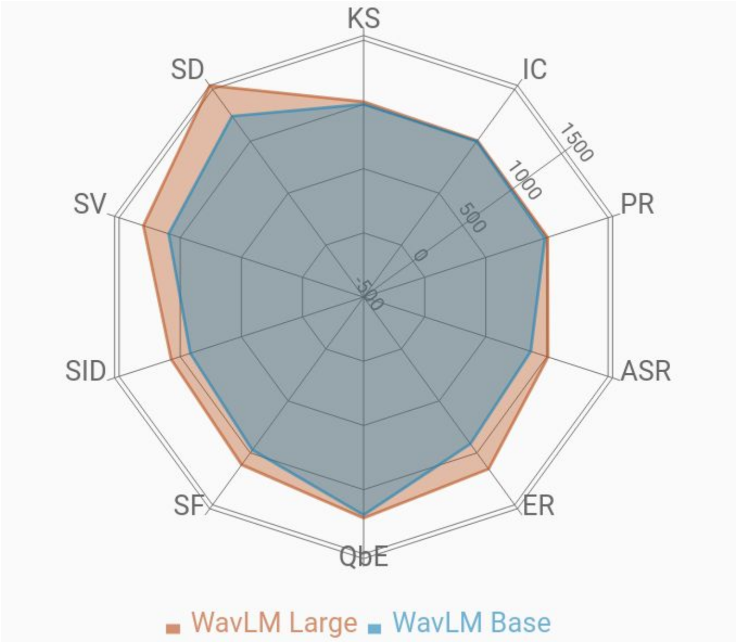

w12
Layer 12
...
...
w2
Layer 2
w1
Layer 1
Speaker A
Downstream Model 2
SID
labeled data
Weighted sum over every layer's representations

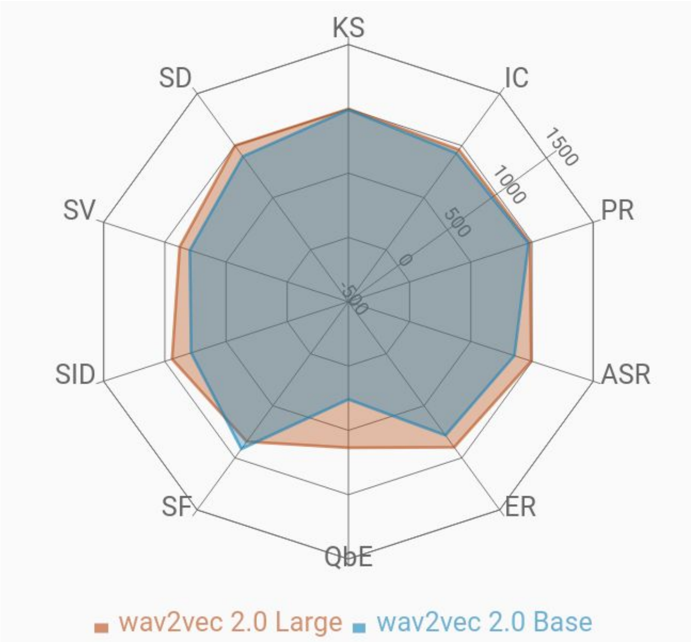
SUPERB

Speaker	Content	Semantics	Paralinguistics
	Query By Example		
Speaker Diarization	Speech Recognition		
Speaker Identification	Phoneme Recognition	Intent Classification	
Speaker Verification	Keyword Spotting	Slot Filling	Emotion Recognition

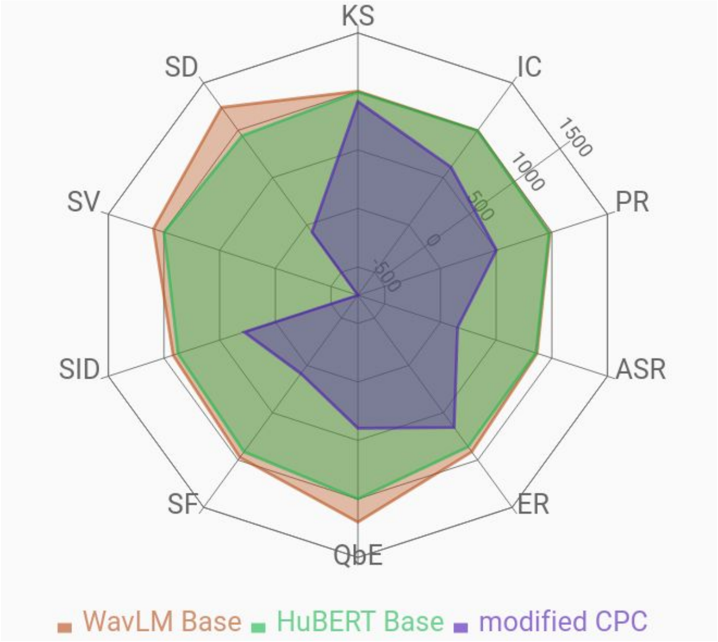
Method	Name	Description	URL	Params ↓	MACs ↓	(1) ↓	(2) ↓	(3) ↓	(4) ↓	Rank ↑	Score ↑	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑
WavLM Large	Microsoft	M-P + VQ +...		3.166e+8	4.326e+12	3.8...	6.7...	1.0...	2.1...	25.8	1145	97.86	99.31	3.06	3.44	70.62
WavLM Base+	Microsoft	M-P + VQ +...		9.470e+7	1.670e+12	1.4...	2.6...	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	Cl 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-Al...		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	Cl 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-Al...		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 l...	-	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 l...	-	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

https://superbbenchmark.org/leaderboard



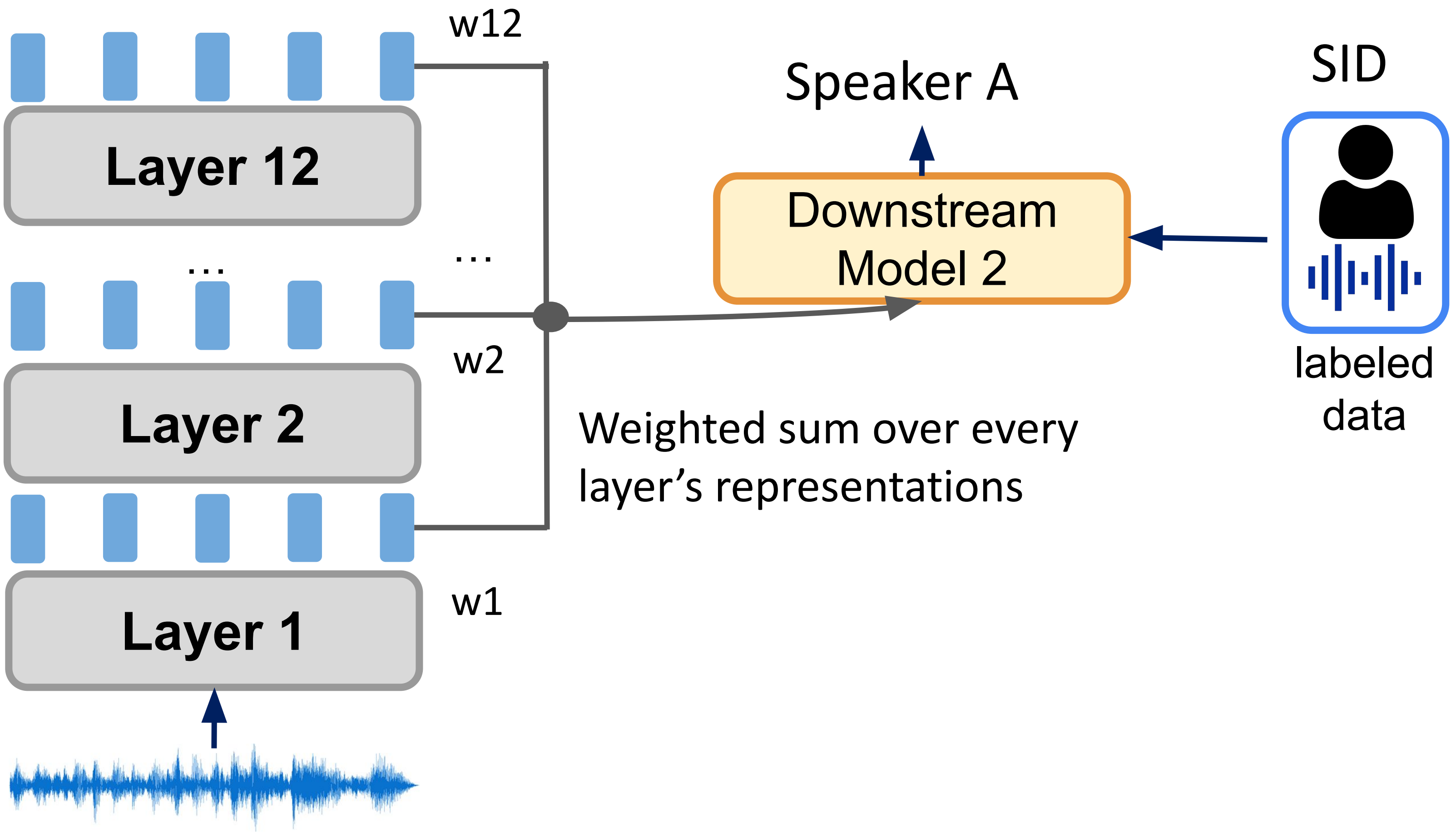


The bigger, the better



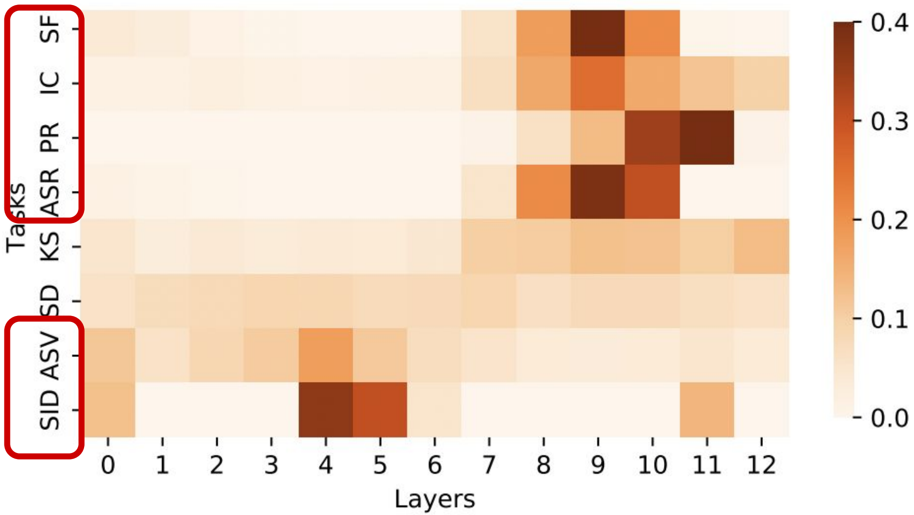
Strong model performs better on all kinds of tasks

SUPERB


w12
Layer 12
...
...
w2
Layer 2
w1
Layer 1
Speaker A
Downstream Model 2
SID
labeled data
Weighted sum over every layer's representations

Representation Weight Analysis

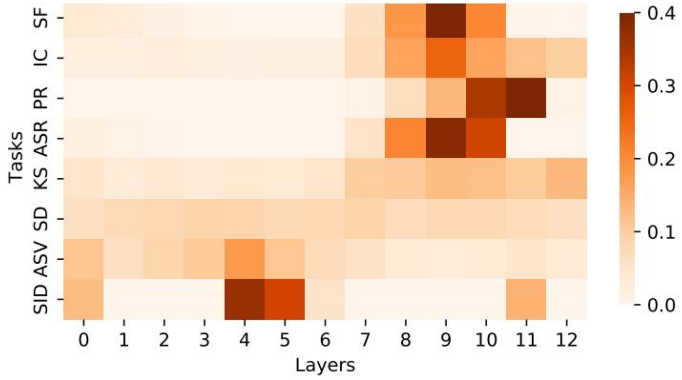
- **Speaker tasks:**
  - ASV (Speaker Verification)
  - SID (Speaker Identification)
  - …
- **Content tasks:**
  - ASR (Speech Recognition)
  - IC (Intent Classification)



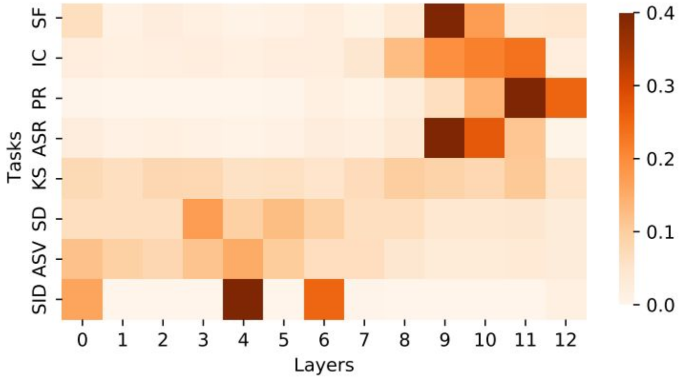
(a) HuBERT Base

WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing

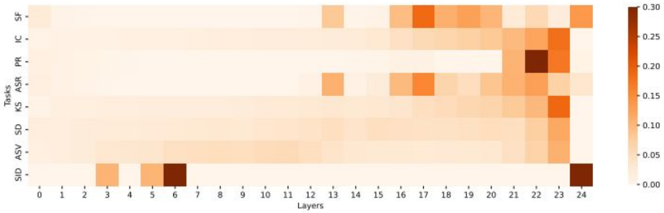
Representation Weight Analysis



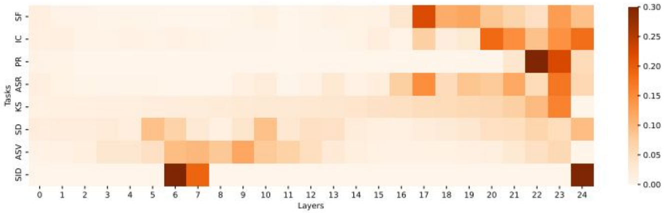
(a) HuBERT Base



(b) WavLM Base+



(c) HuBERT Large



(d) WavLM Large

Part 1 Summary: Speech Representation Learning

	Contrastive Models	Predictive Models	
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

Part 1 Summary: Speech Representation Learning

Contrastive Models	Predictive Models
CPC	HuBERT
wav2vec 2.0	WavLM
XLS-R	BEST-RQ

SUPERB

Speaker
Content
Semantics
Paralingustics
…

- 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 | NLP

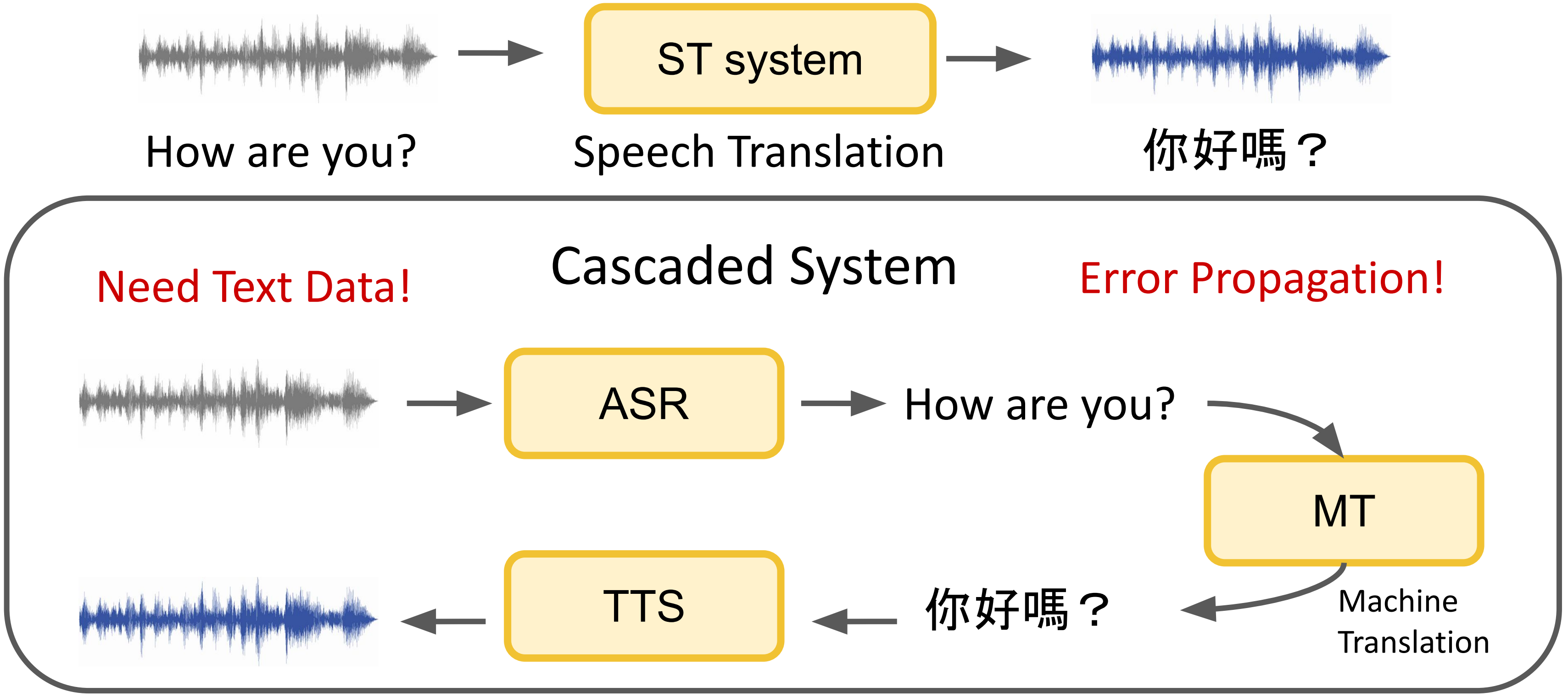
Textless NLP: Generating expressive speech from raw audio

September 9, 2021

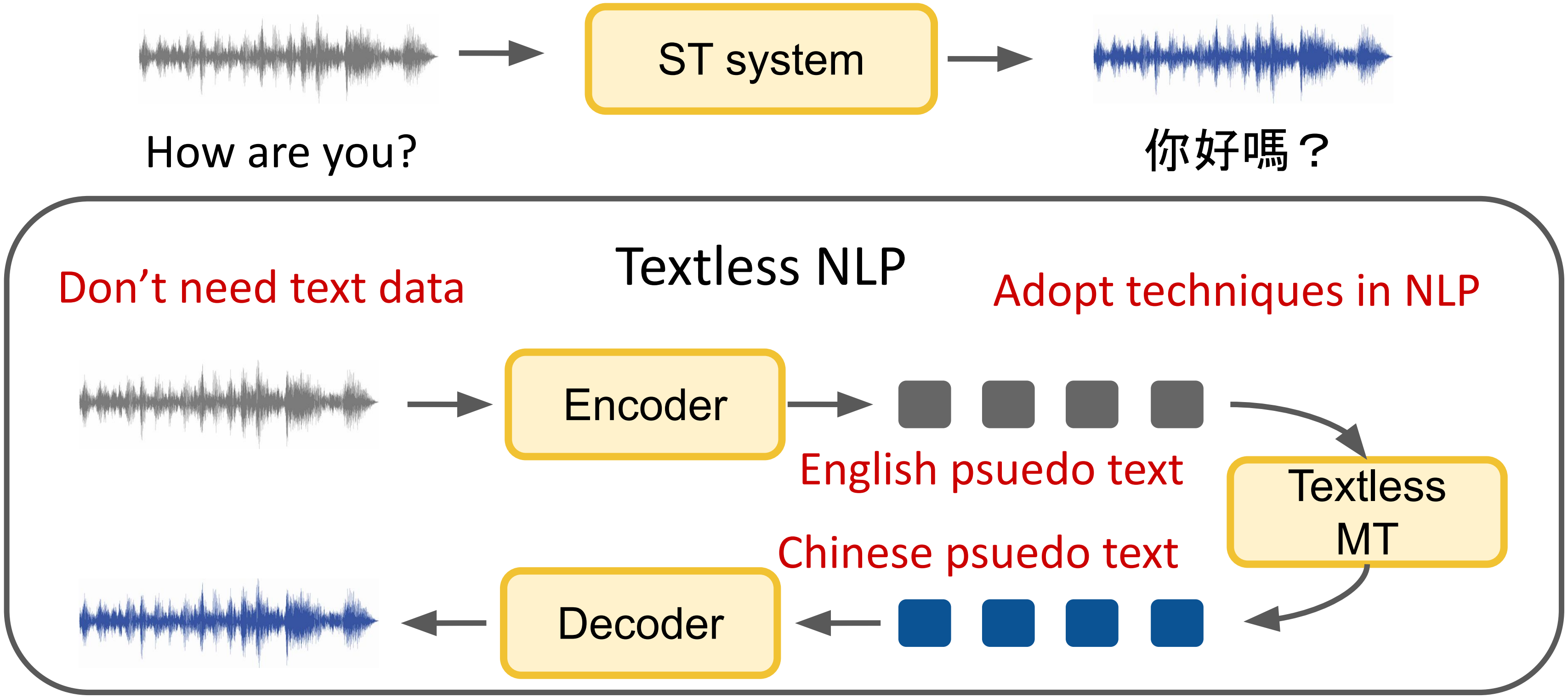
Share on Facebook
Share on Twitter

Our Work

Textless NLP Project

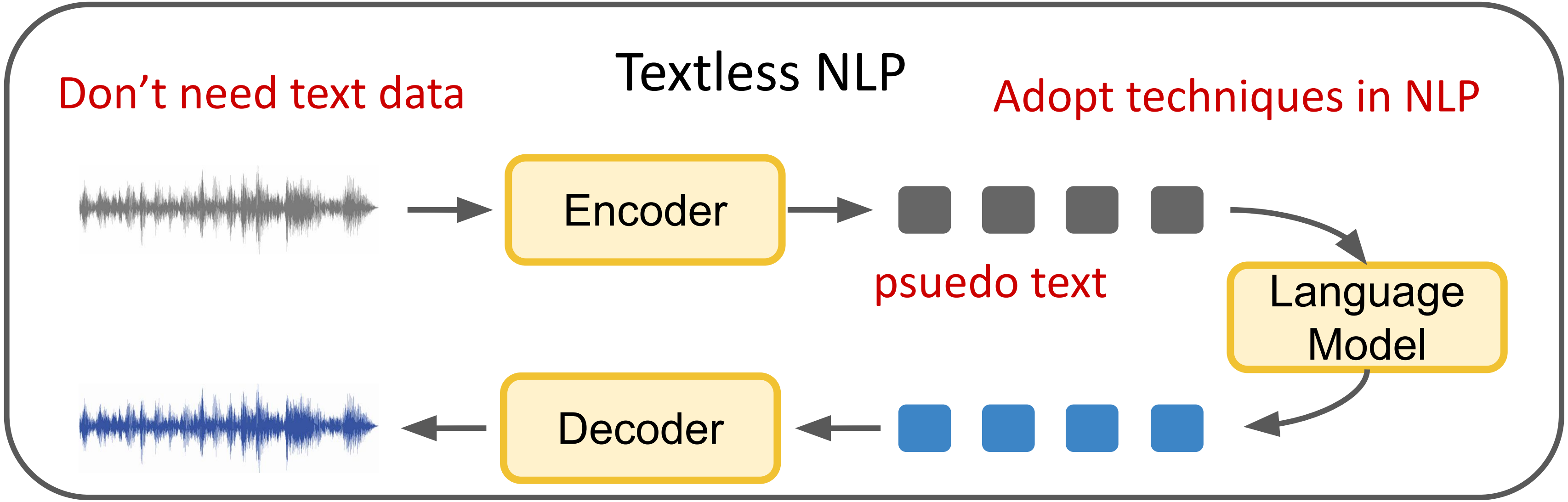


Textless NLP Project

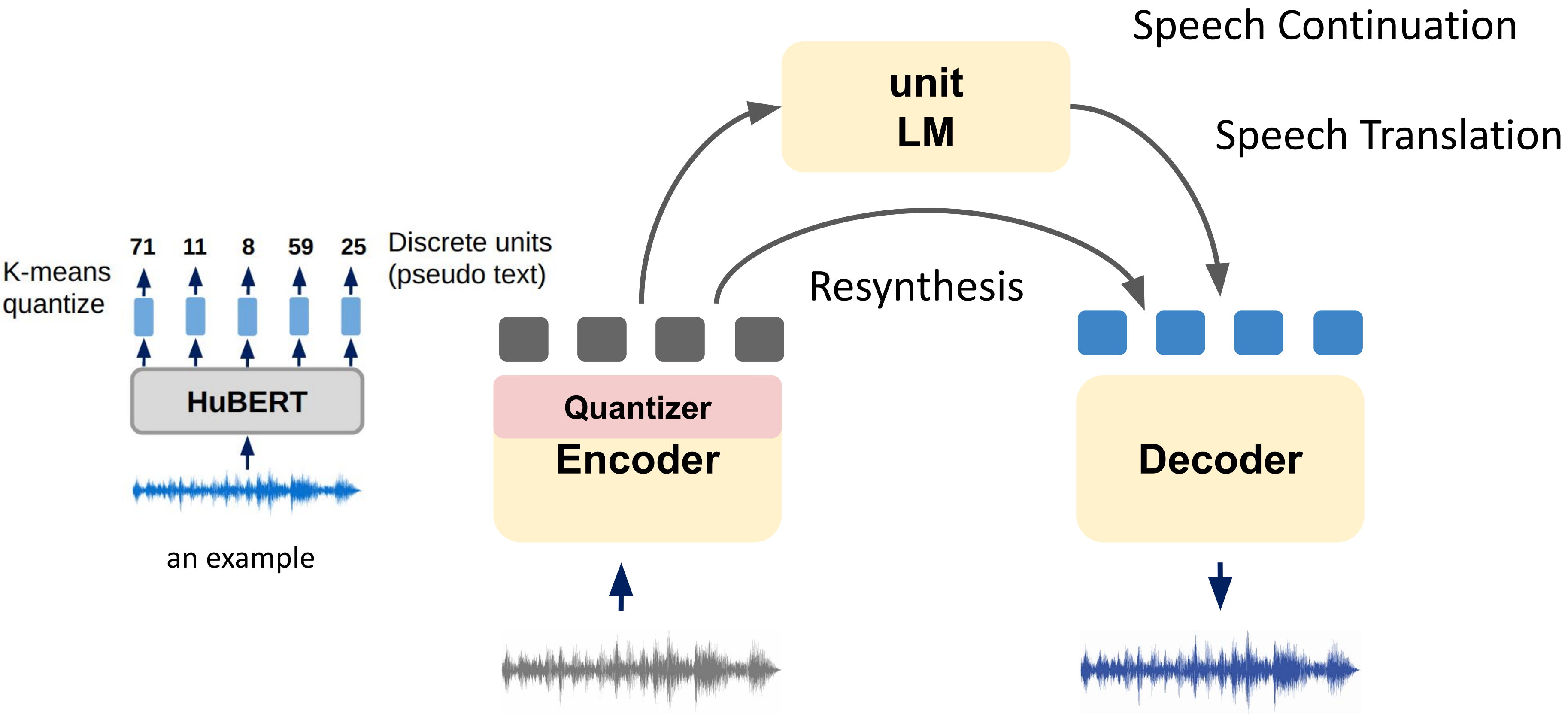


Textless NLP Project

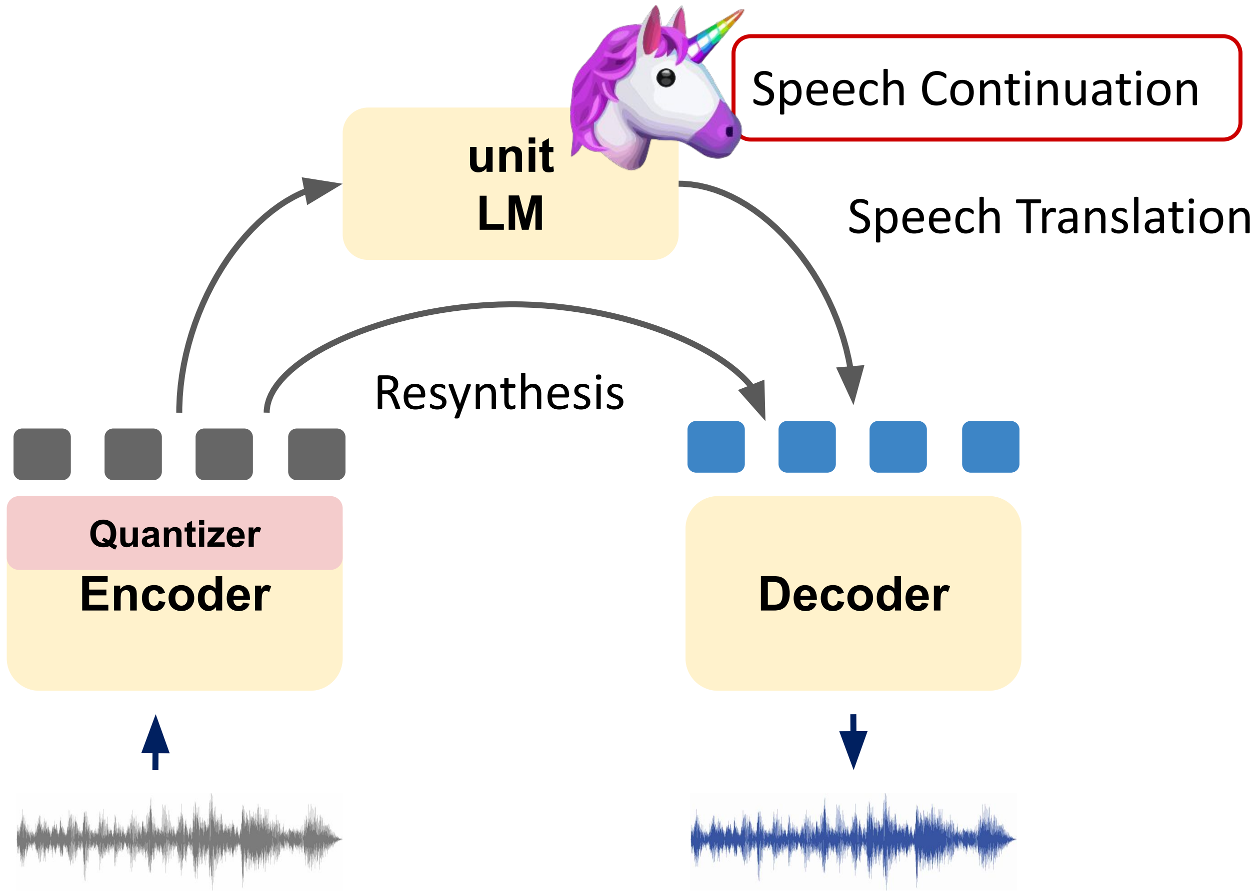
- GPT: Speech Continuation
- BART: Speech Translation
- no LM: Speech Resynthesis



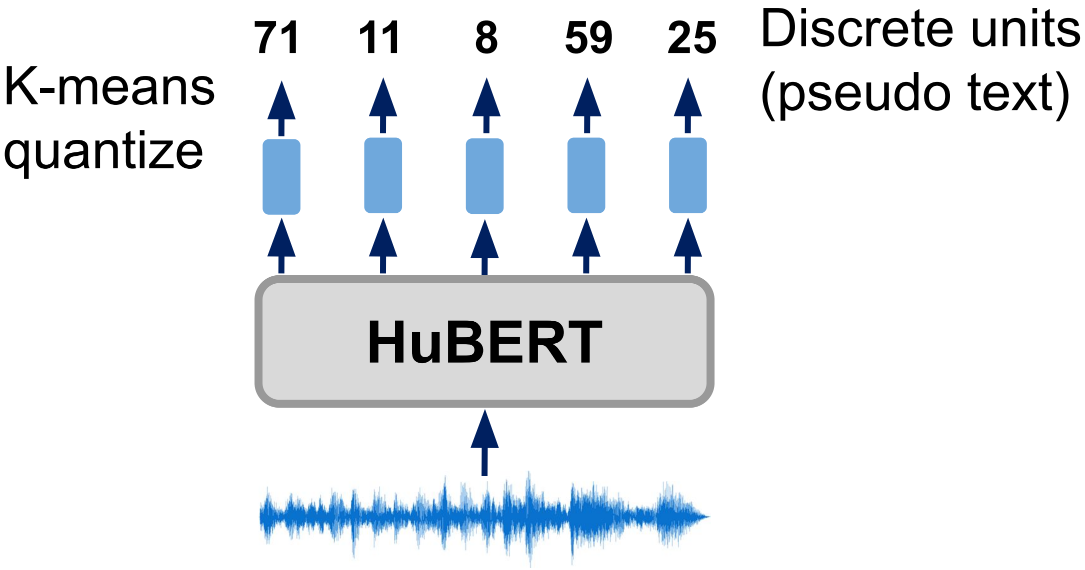
Textless NLP Project



Textless NLP Project

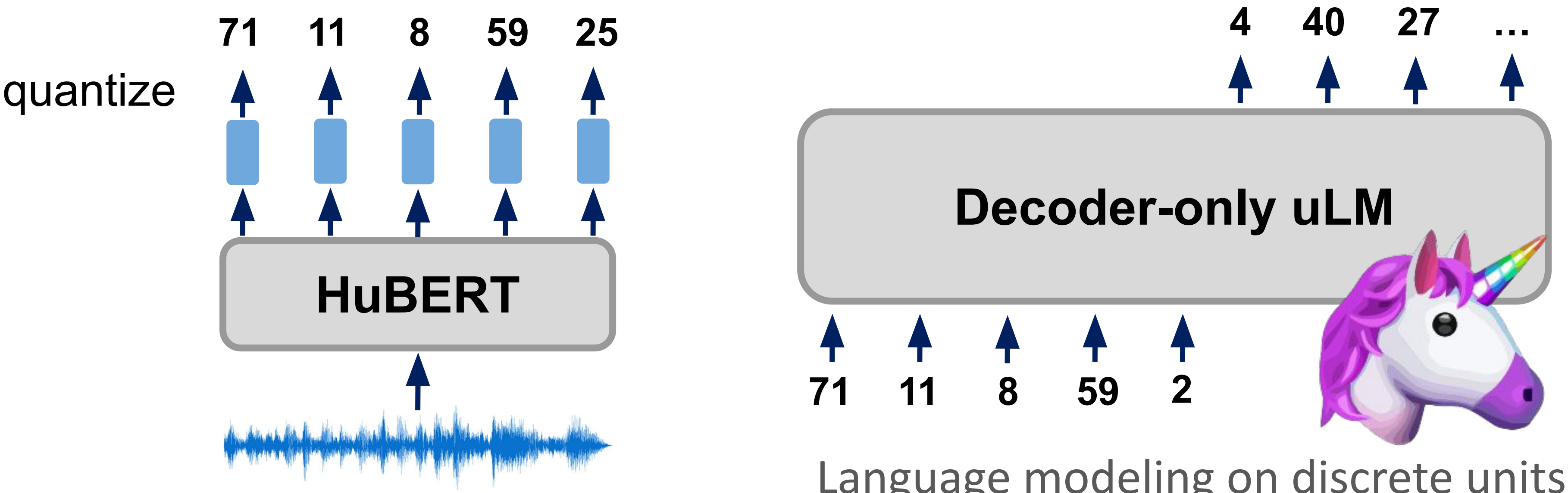


Generative Spoken Language Modeling



Lakhotia et.al., arXiv 21’

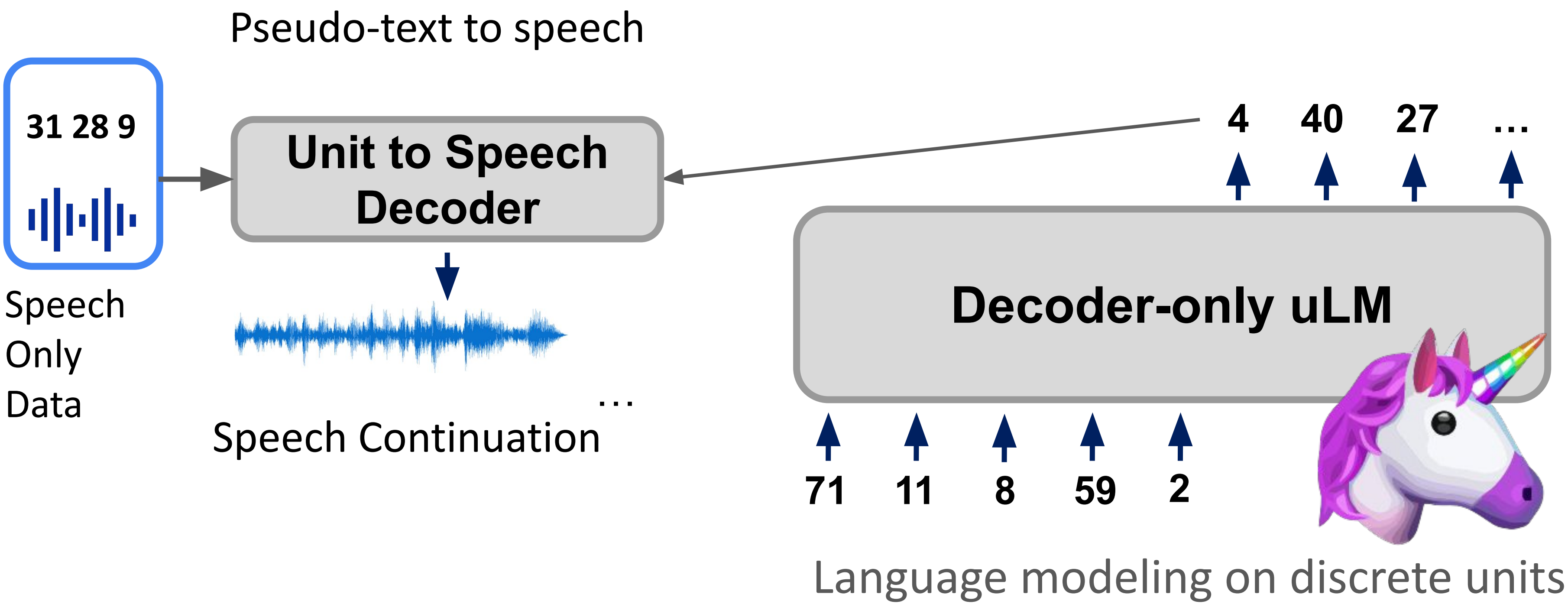
Generative Spoken Language Modeling



Language modeling on discrete units

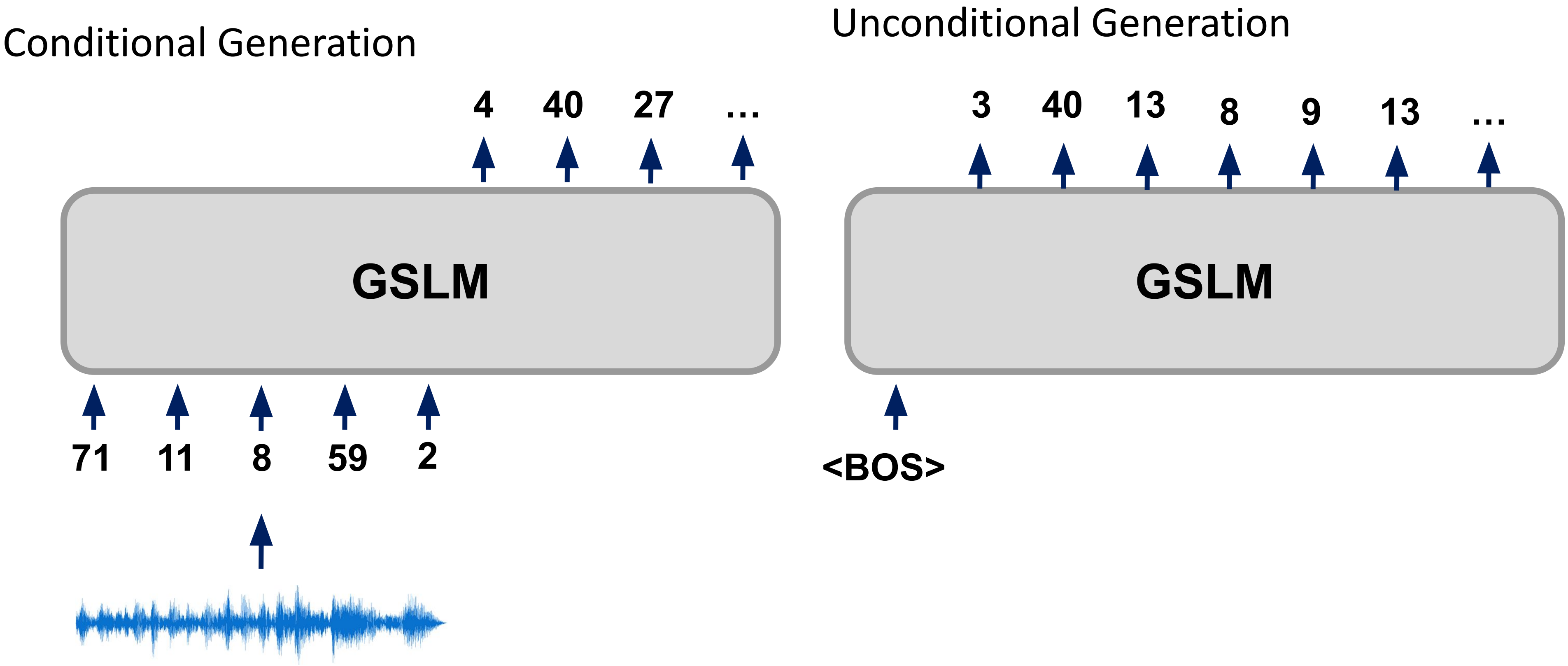
Lakhotia et.al., arXiv 21'

Generative Spoken Language Modeling



Lakhotia et.al., arXiv 21’

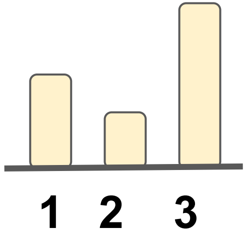
Generative Spoken Language Modeling



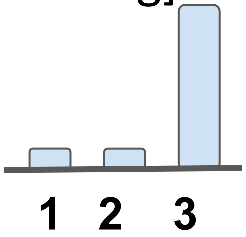
Lakhotia et.al., arXiv 21'

Generative Spoken Language Modeling

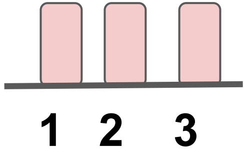
Unconditional Generation



temperature = 0.3 [stuttering]

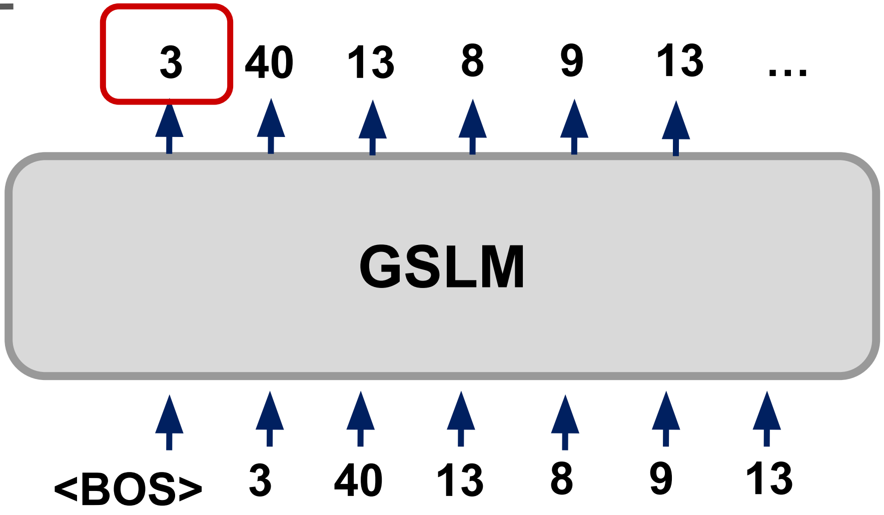


and to take in another path
and to take in another path
and to take in another path
and to take in another path …



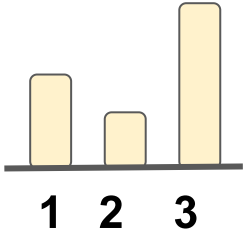
temperature = 1.5 [babble]

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 …



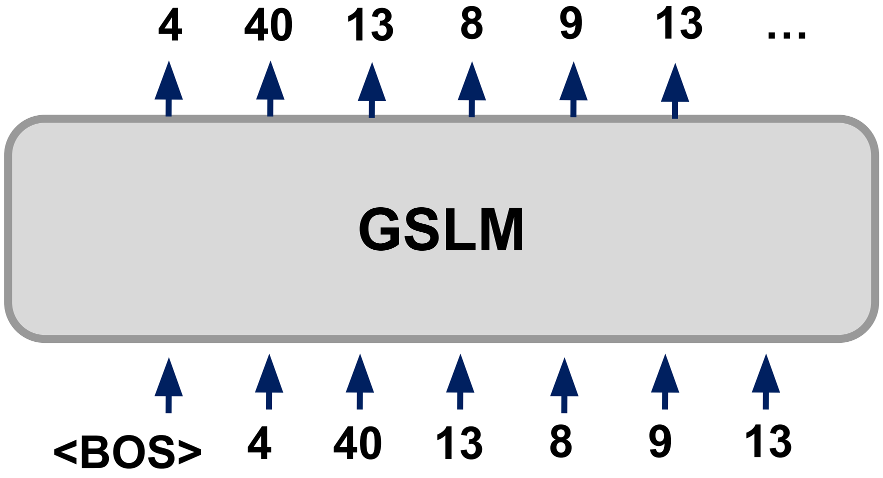
Lakhotia et.al., arXiv 21’

Generative Spoken Language Modeling



Unconditional Generation

temperature = 1

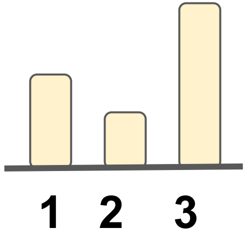


but it is attendant from the people to defend himself from this information pride of the potential in criminal activity a curiosity and impetuosity of the world a war soon acquired

“Locally” coherent

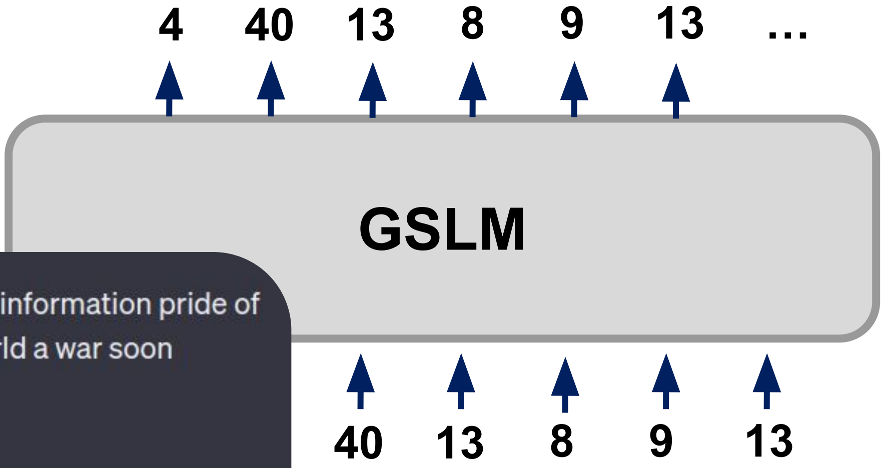
Lakhotia et.al., arXiv 21’

Generative Spoken Language Modeling



Unconditional Generation

temperature = 1



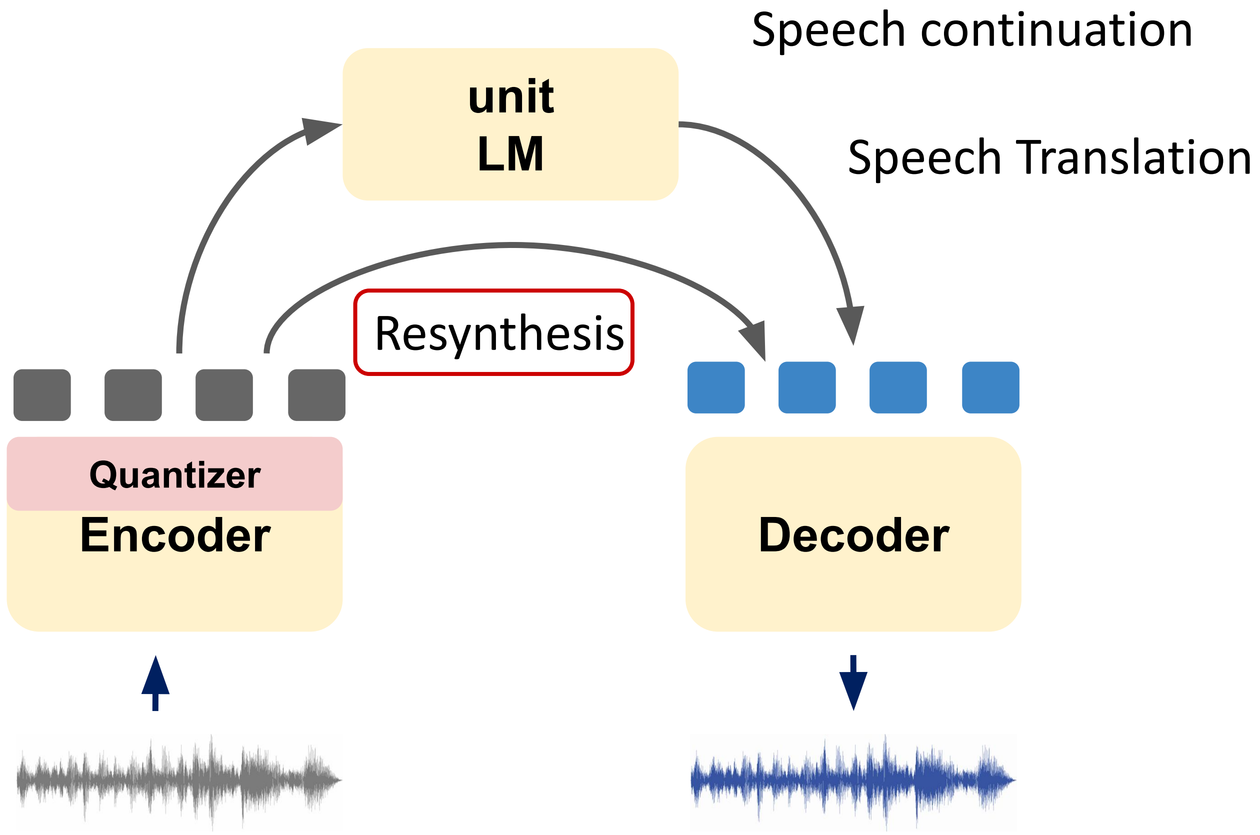
but it is attendant from the people to defend himself from this information pride of the potential in criminal activity a curiosity and impetuosity of the world a war soon acquired

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。"

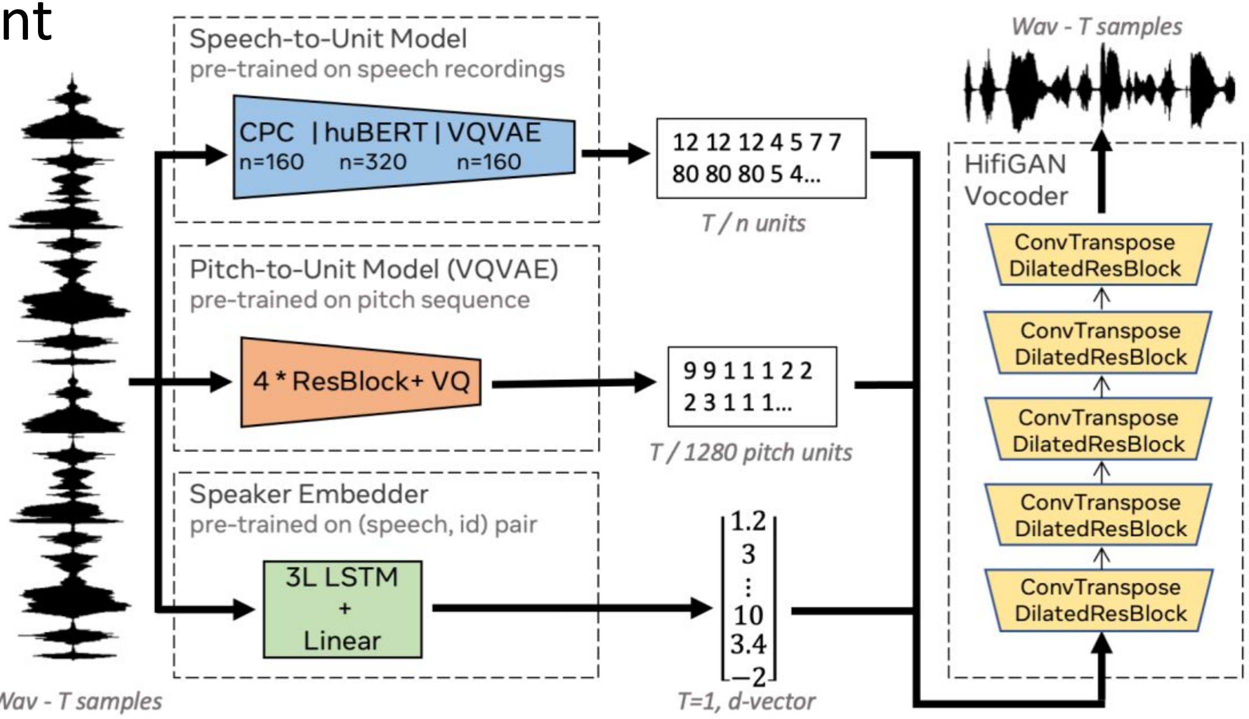
Textless NLP Project



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

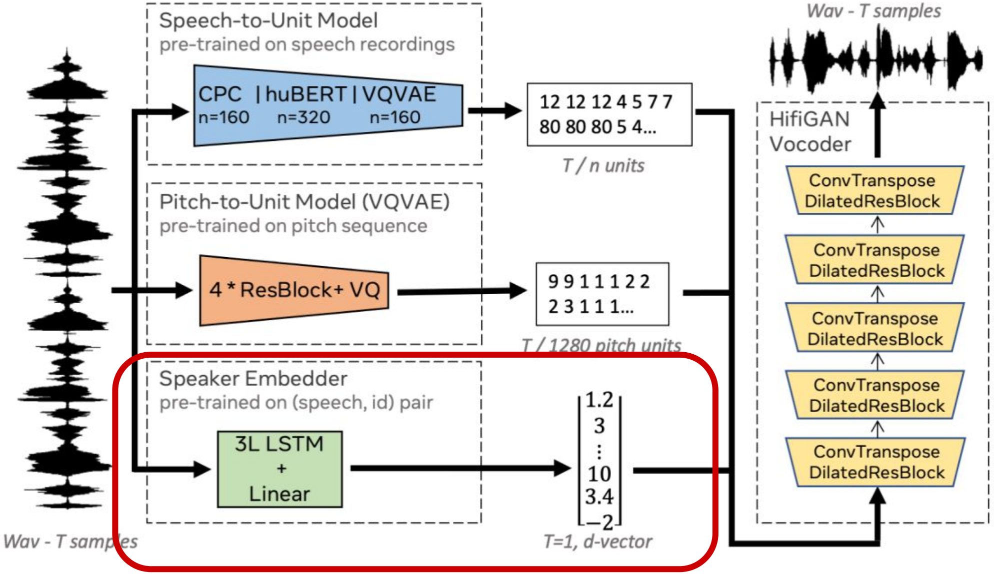
Multistream is required to perform resynthesis

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

Kharitonov, Eugene, et al. "textless-lib: A library for textless spoken language processing." *arXiv preprint arXiv:2202.07359* (2022).

Speech Resynthesis

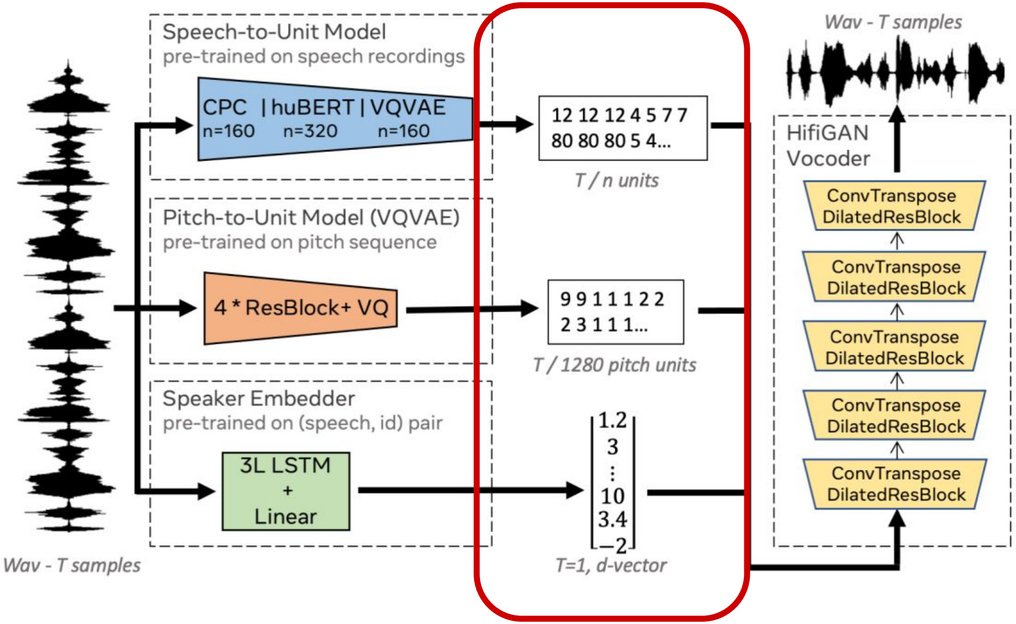


Replace the speaker embedding with other speaker's embedding

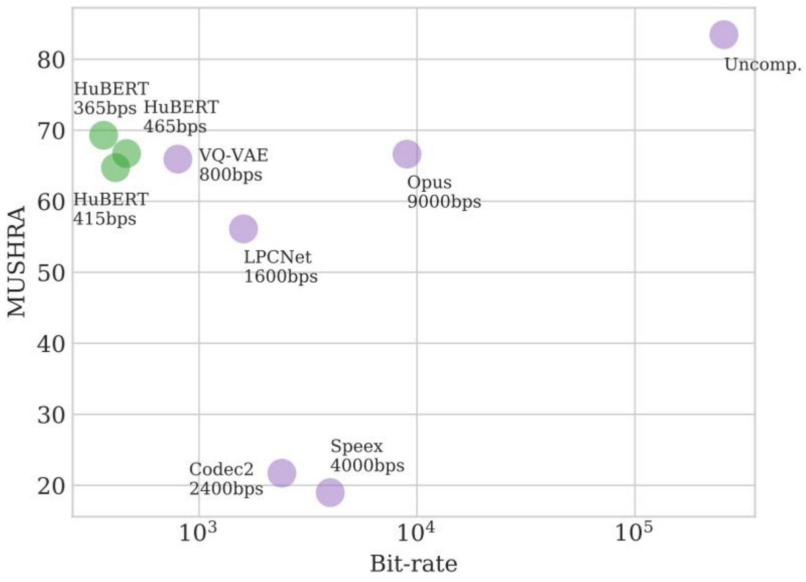
Voice Conversion

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

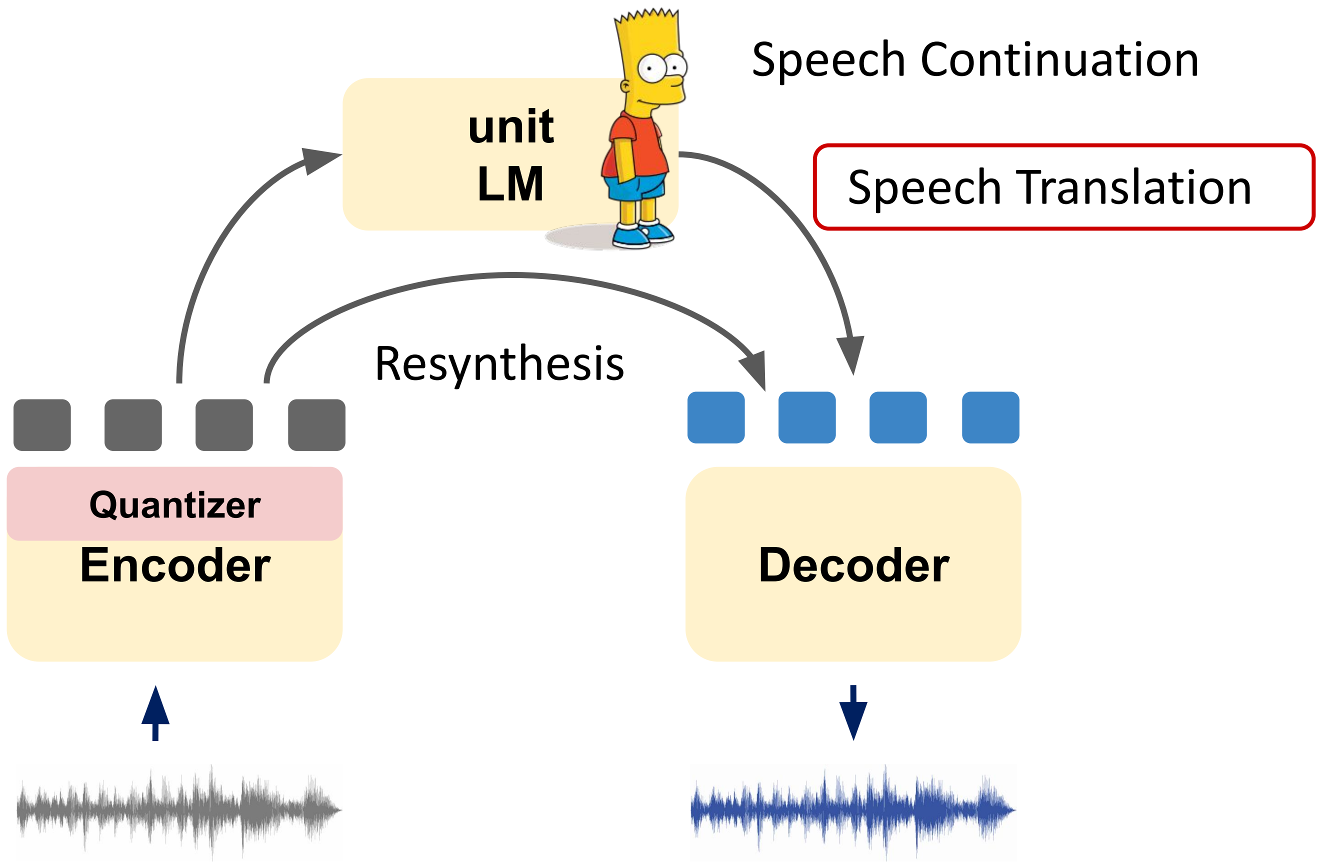
Speech Resynthesis



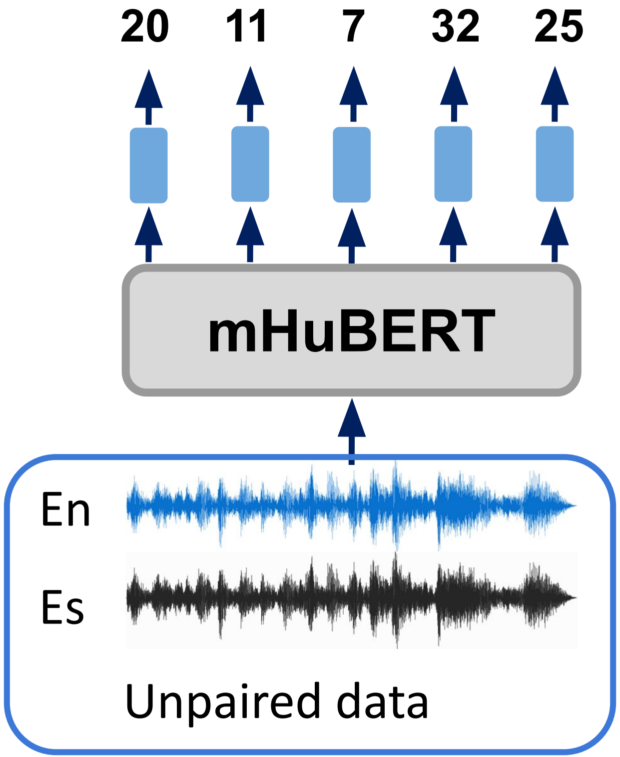
Speech Codec

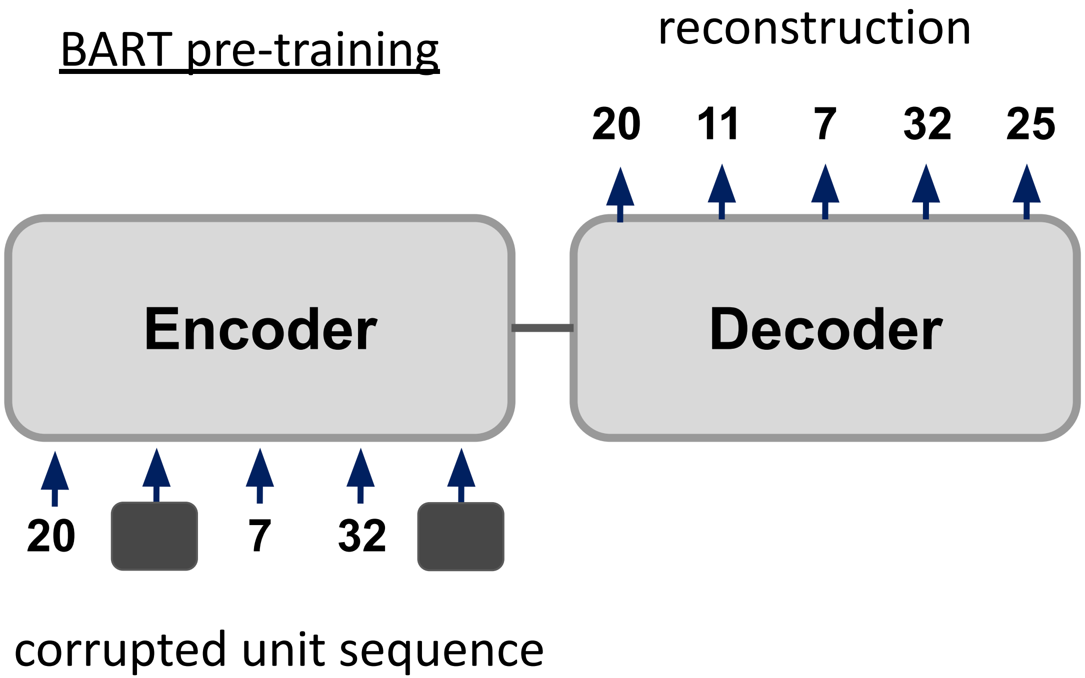


Textless NLP Project



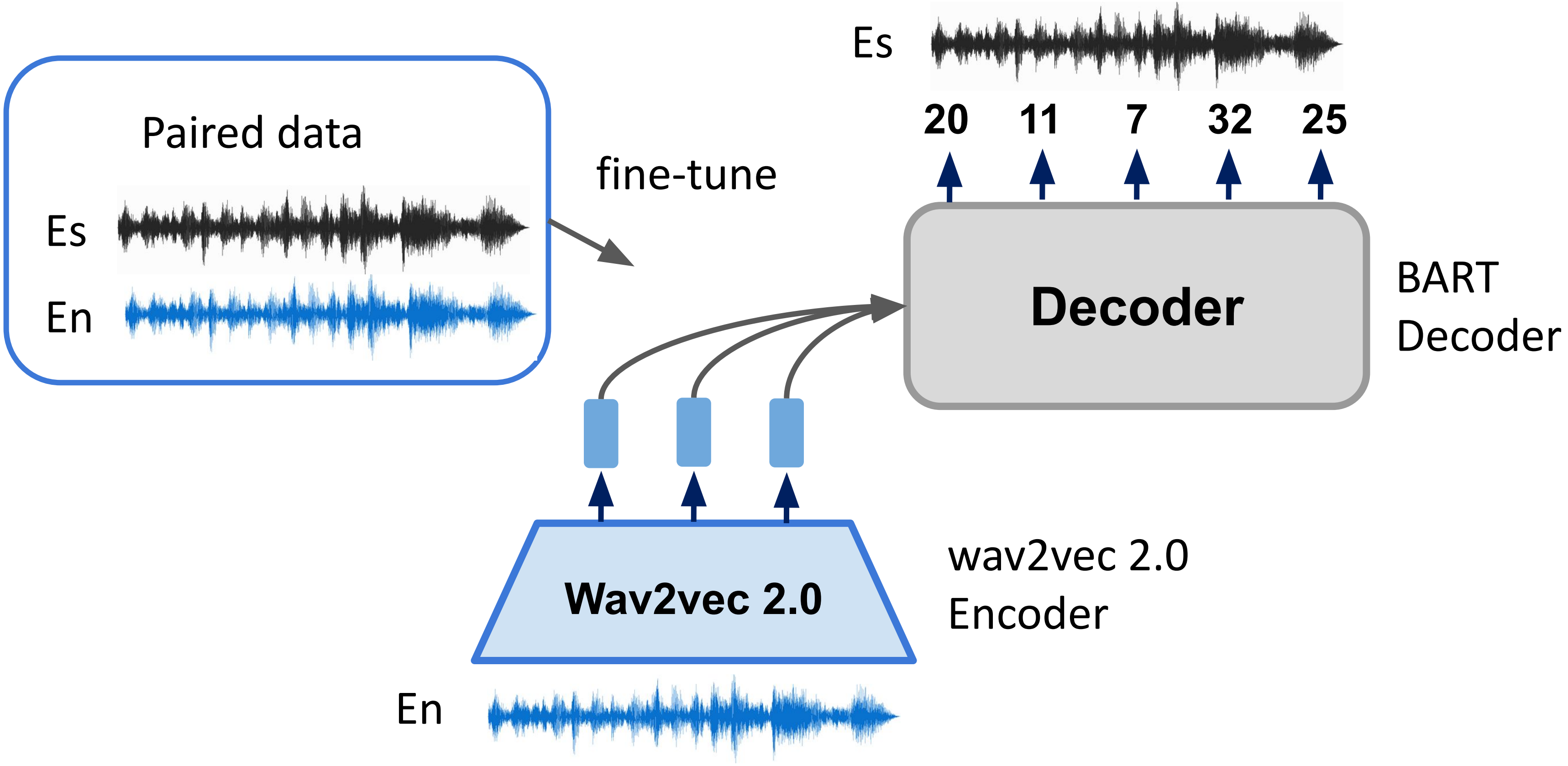
Speech Translation: Unit BART





Popuri, Sravya, et al. "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation." *arXiv preprint arXiv:2204.02967* (2022).

Speech Translation: Unit BART



Speech Translation: Unit BART

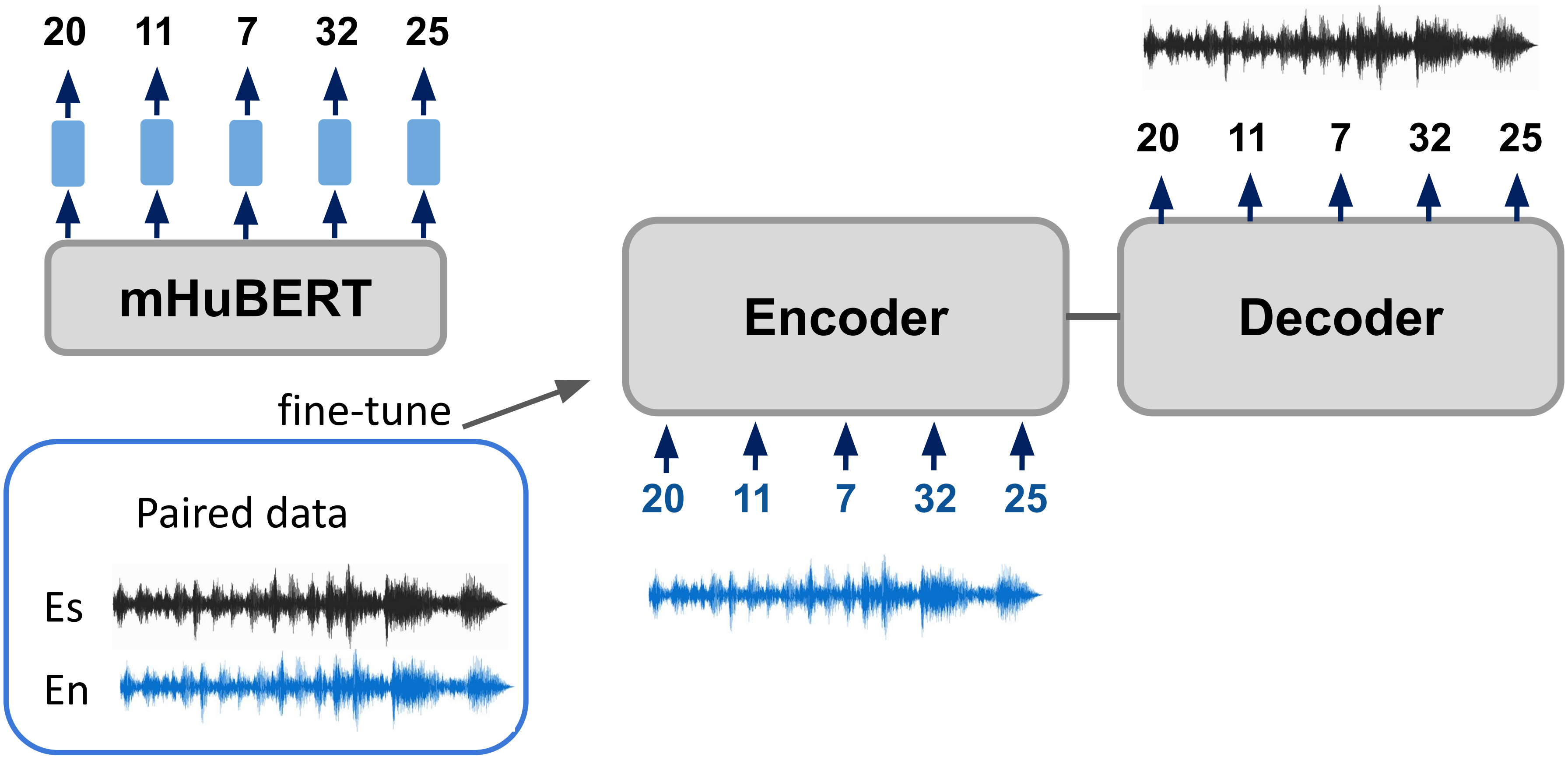


Table 2: *Dev / test BLEU on all the datasets included in the "S2ST-syn" data. All S2UT systems are decoded with beam size 10. MOS is reported with 95% confidence interval. (w2v2-L: wav2vec 2.0* LARGE*)*

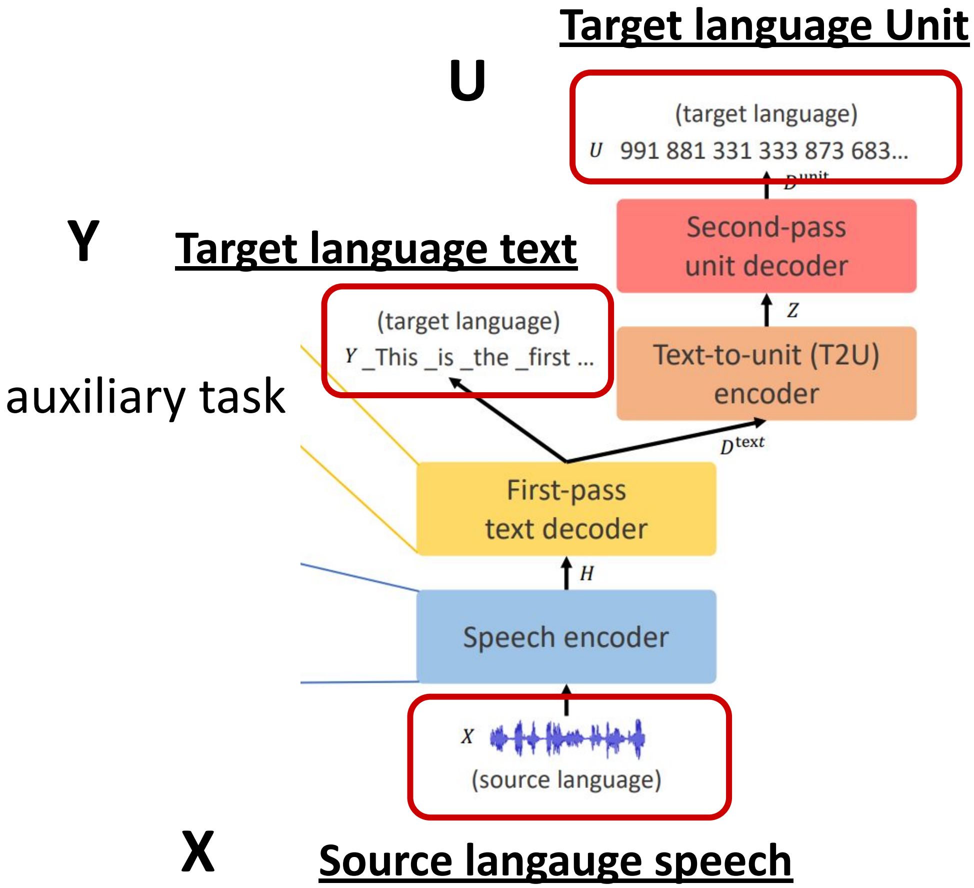
ID		En-Es BLEU Europarl-ST	En-Es BLEU MuST-C	En-Es MOS combined	Es-En BLEU CoVoST-2	Es-En BLEU Europarl-ST	Es-En BLEU mTEDx	Es-En MOS combined
Cascaded systems:								
1	S2T (w2v2-L)+TTS	33.0 / 32.6	30.3 / 30.1	3.80 ± 0.12	25.9 / 28.4	26.9 / 23.6	25.3 / 21.5	3.53 ± 0.14
2	ASR+MT+TTS	28.9 / 28.8	36.4 / 34.2	-	37.3 / 33.8	33.3 / 29.1	29.3 / 32.4	-
S2UT systems without pre-training:								
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
S2UT systems with model pre-training:								
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)	**32.2 / 32.5**	**32.6 / 30.8**	4.06 ± 0.10	**27.3 / 30.2**	**29.0 / 26.4**	**29.6 / 25.2**	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

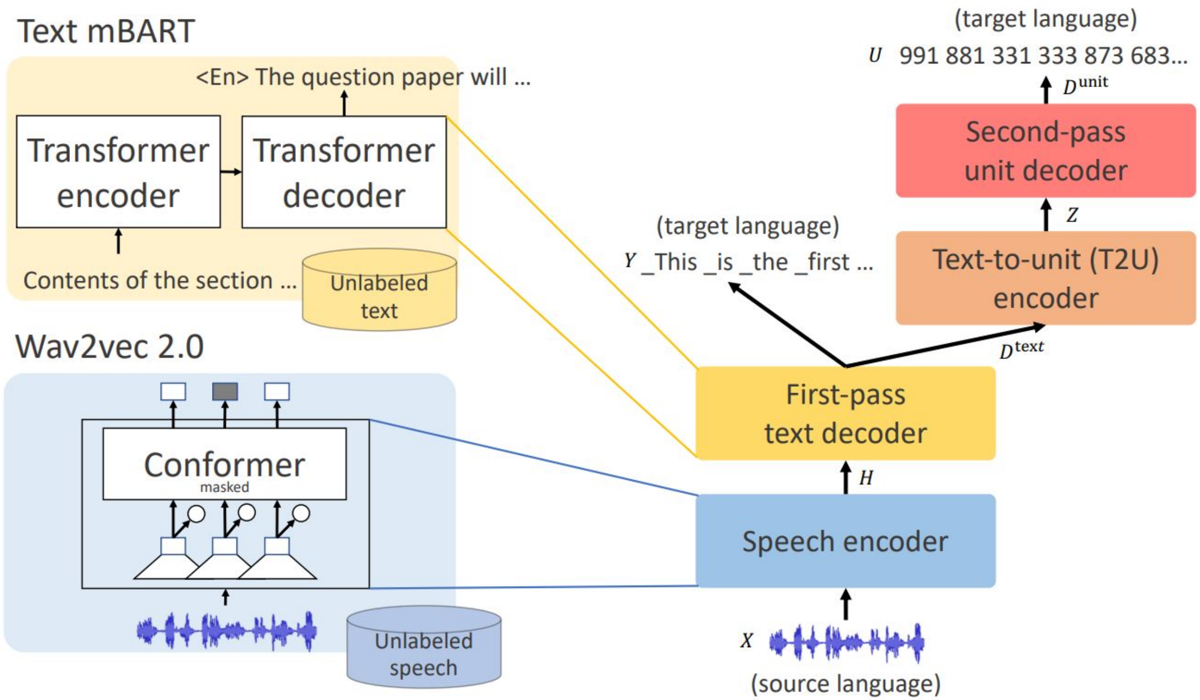
1. Speech Encoder
2. First-pass text decoder
3. Text-to-unit encoder
4. Second-pass unit decoder

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



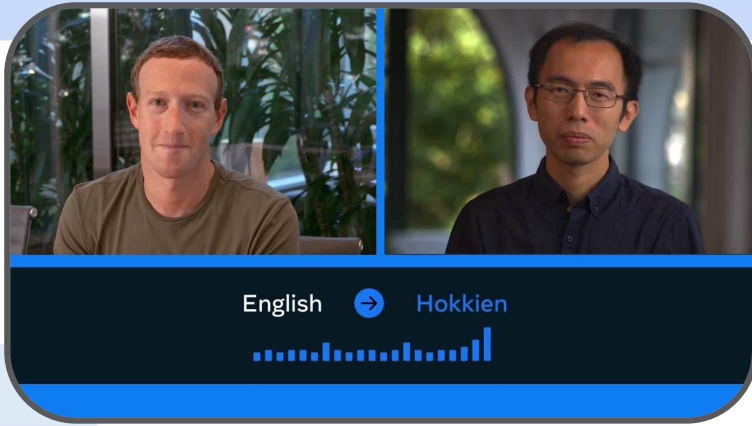
Speech Translation: UnitY

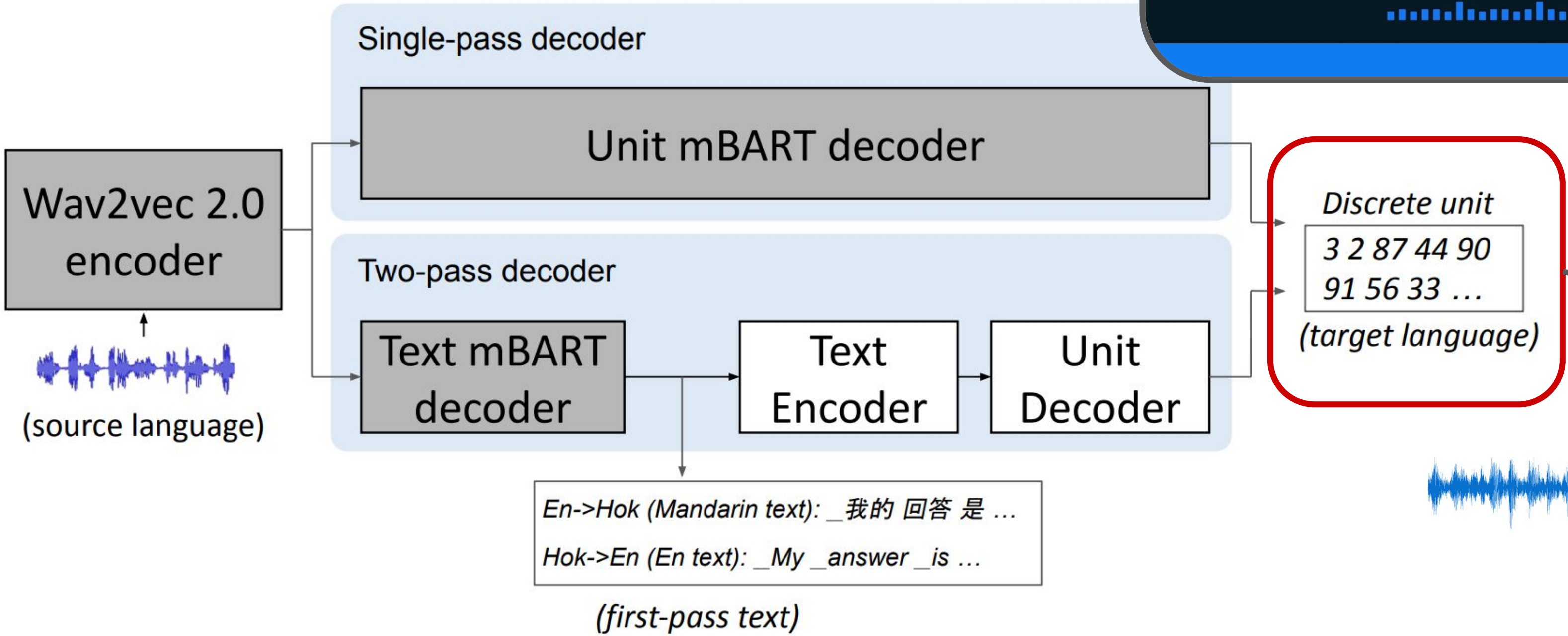
- Unlabeled text
- Unlabeled speech
- labeled speech
- paired speech



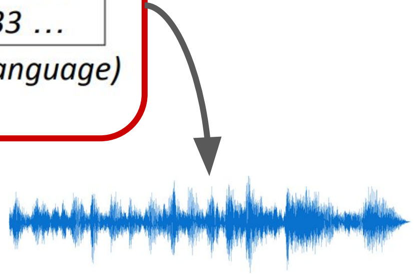
ID	Model	Encoder	ASR-BLEU (↑)		
			dev	dev2	test
A0	Synthetic target (Lee et al., 2022a)		88.5	89.4	90.5
Cascaded systems					
A1	ASR → MT → TTS	LSTM (Lee et al., 2022a)	42.1	43.5	43.9
A2	S2TT → TTS	LSTM (Jia et al., 2019b)	39.4	41.2	41.4
A3		LSTM (Jia et al., 2022b)	–	–	43.3
A4		LSTM (Lee et al., 2022a)	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
...					
Direct systems (speech-to-unit)					
A17	S2UT	Transformer (Lee et al., 2022a)	–	–	39.9
A18		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	**55.1**	**56.5**	**55.9**

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





No speaker information



https://about.fb.com/news/2022/10/hokkien-ai-speech-translation/

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

Google Research   Philosophy   Research Areas   Publications   People   Resources

BLOG ›

AudioLM: a Language Modeling Approach to Audio Generation

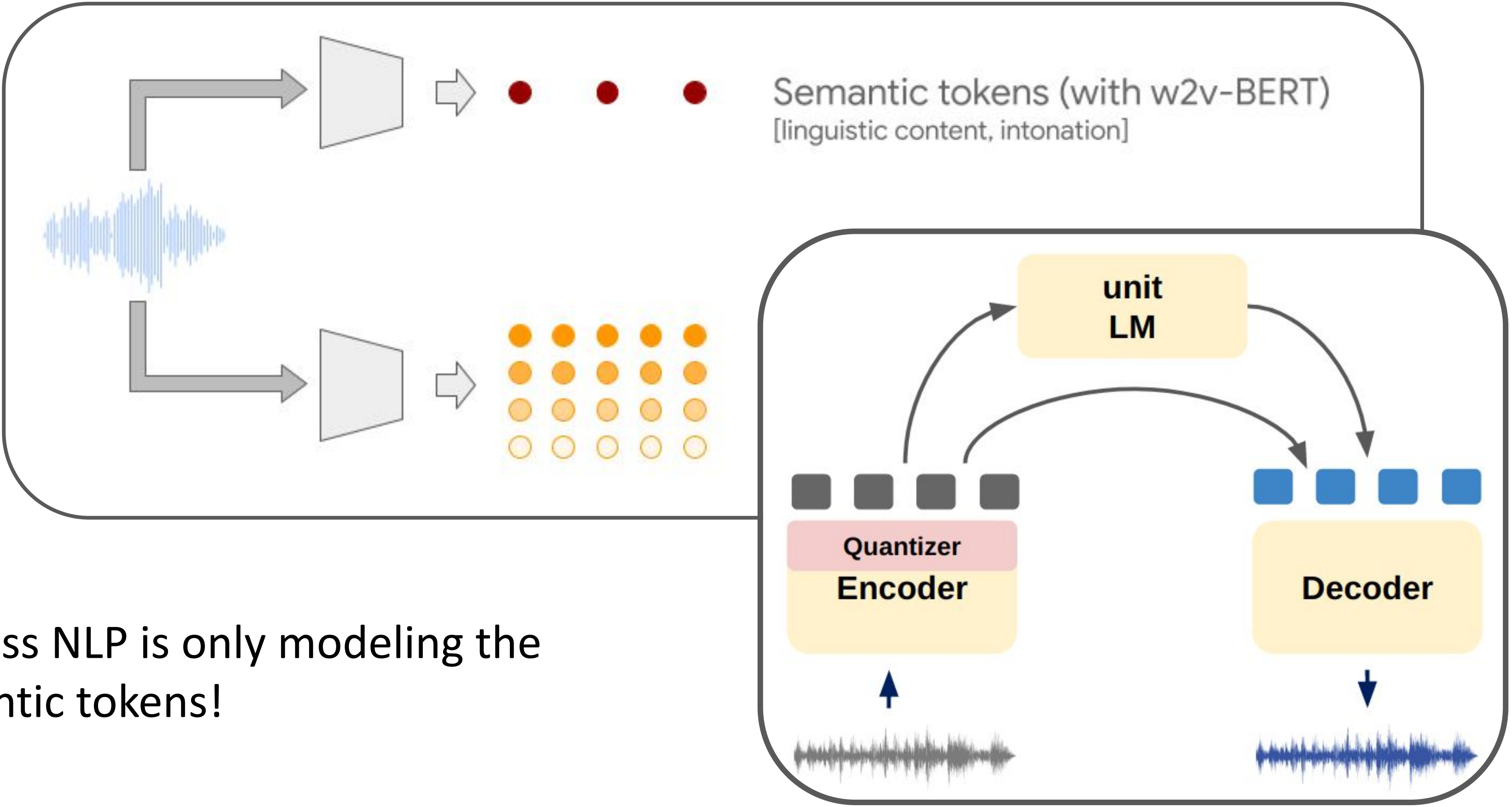
THURSDAY, OCTOBER 06, 2022

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

https://ai.googleblog.com/2022/10/audiolm-language-modeling-approach-to.html

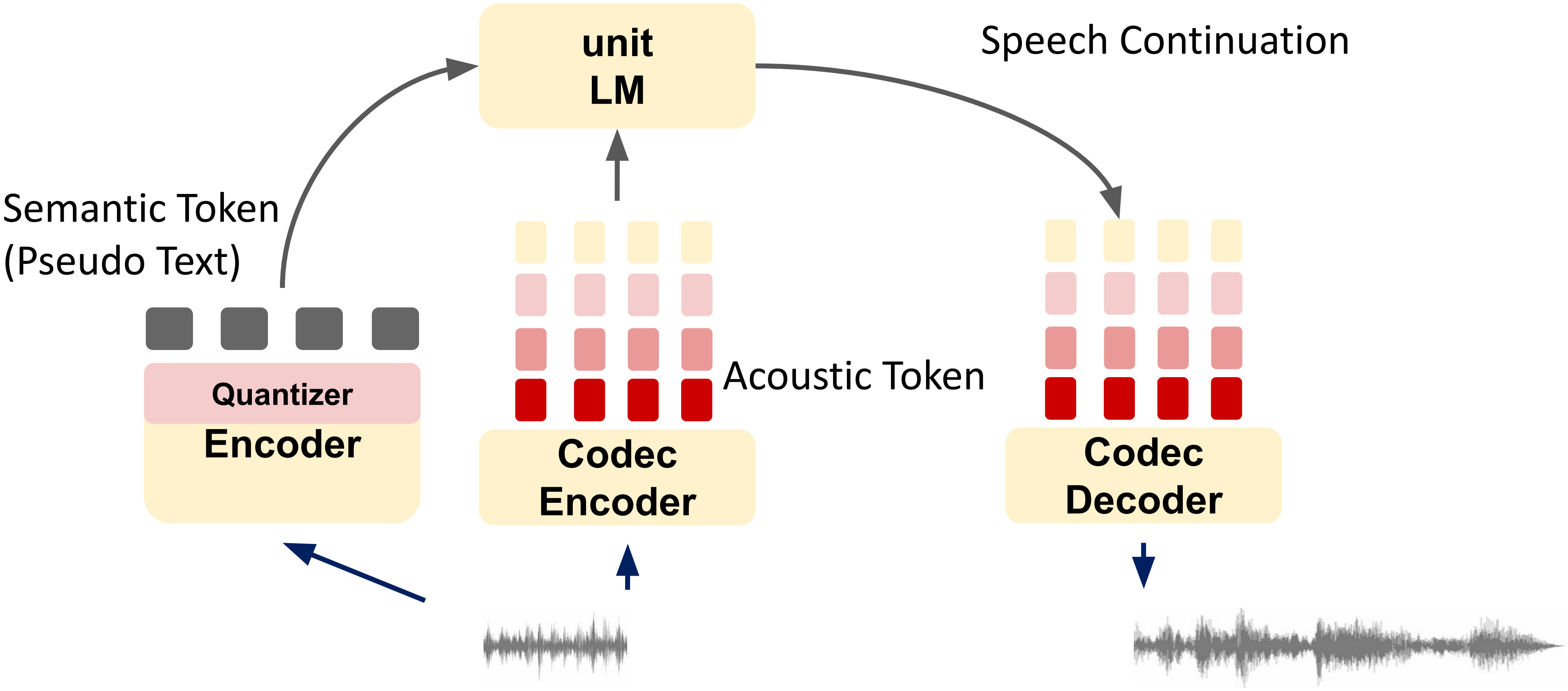
AudioLM

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

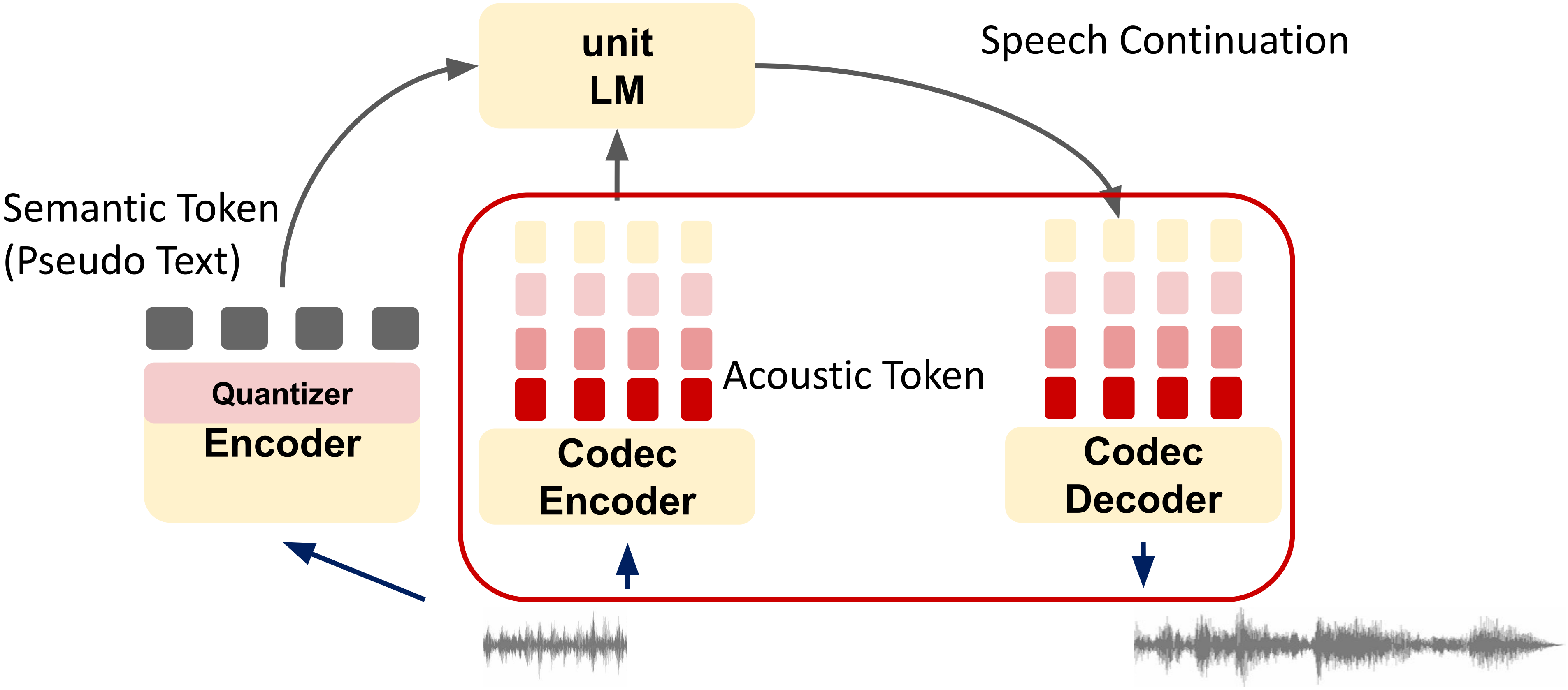


Textless NLP is only modeling the semantic tokens!

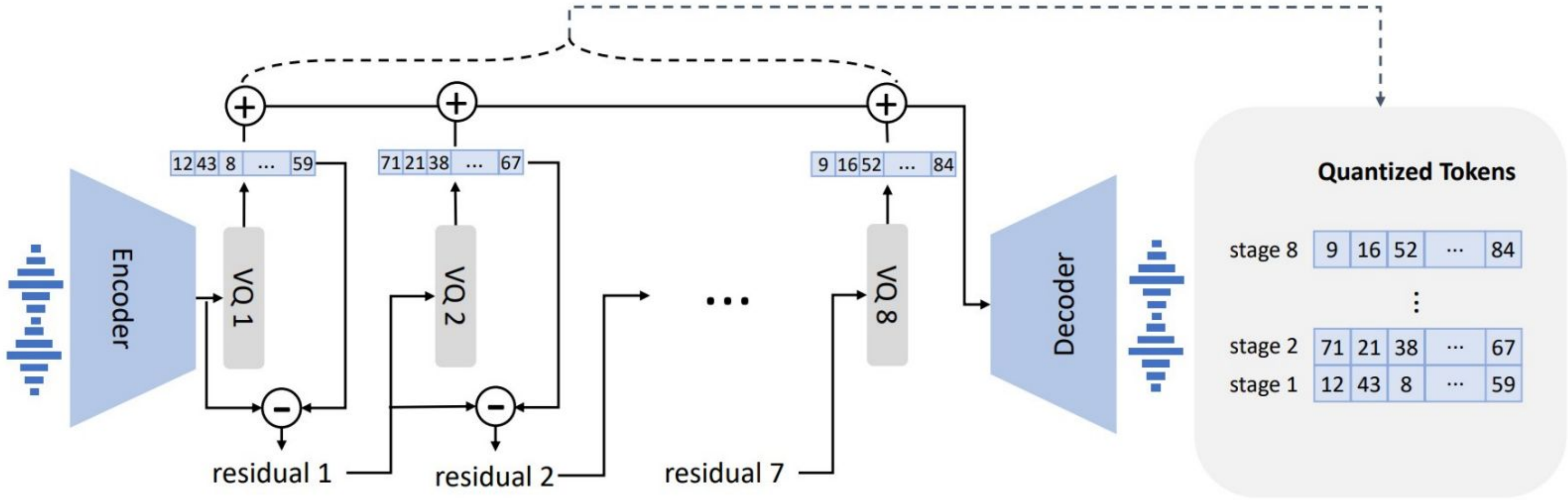
AudioLM


unit
LM
Speech Continuation
Semantic Token
(Pseudo Text)
Quantizer
Encoder
Acoustic Token
Codec
Encoder
Codec
Decoder

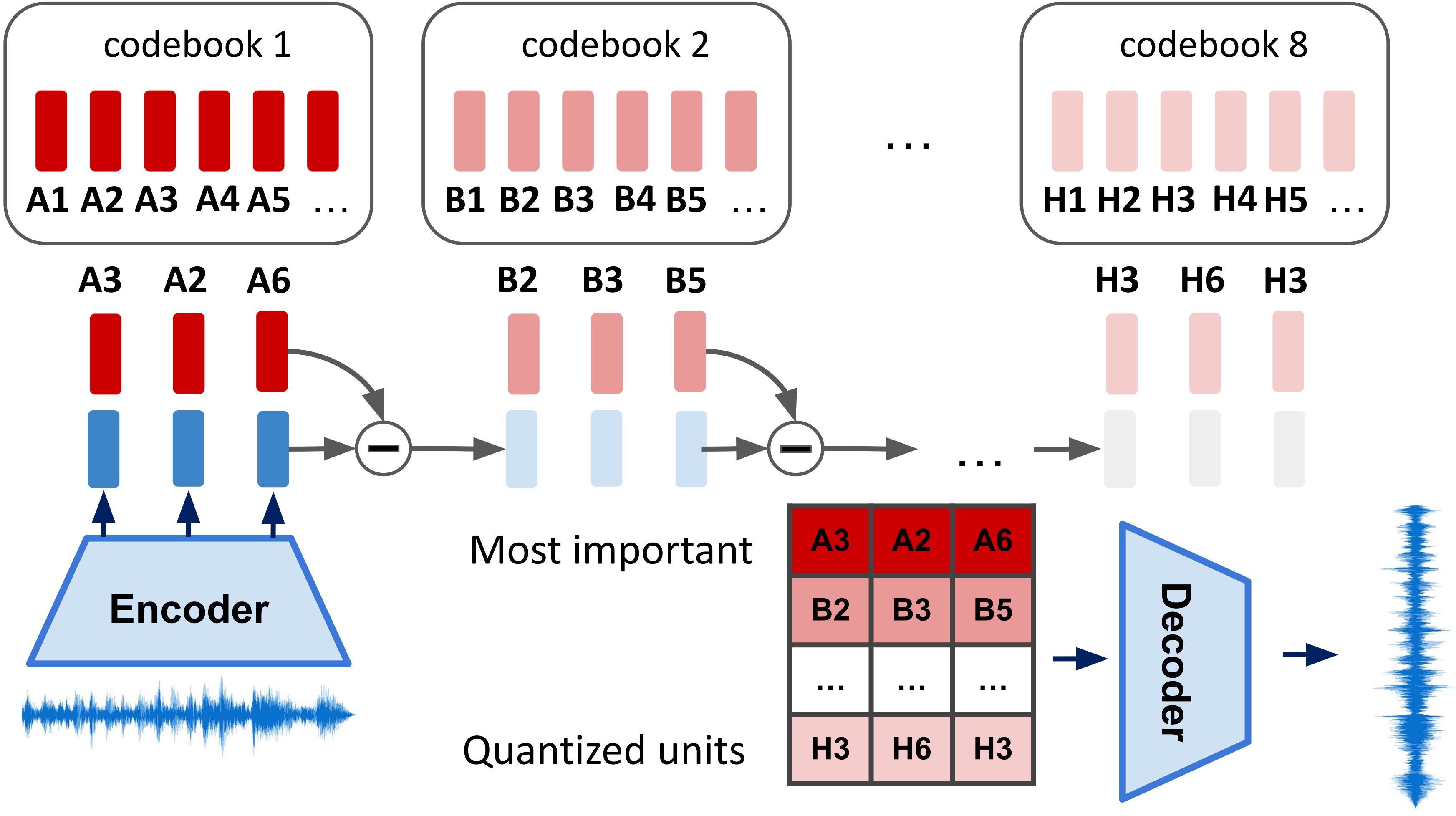
AudioLM


unit
LM
Speech Continuation
Semantic Token
(Pseudo Text)
Quantizer
Encoder
Acoustic Token
Codec
Encoder
Codec
Decoder

Codec Model

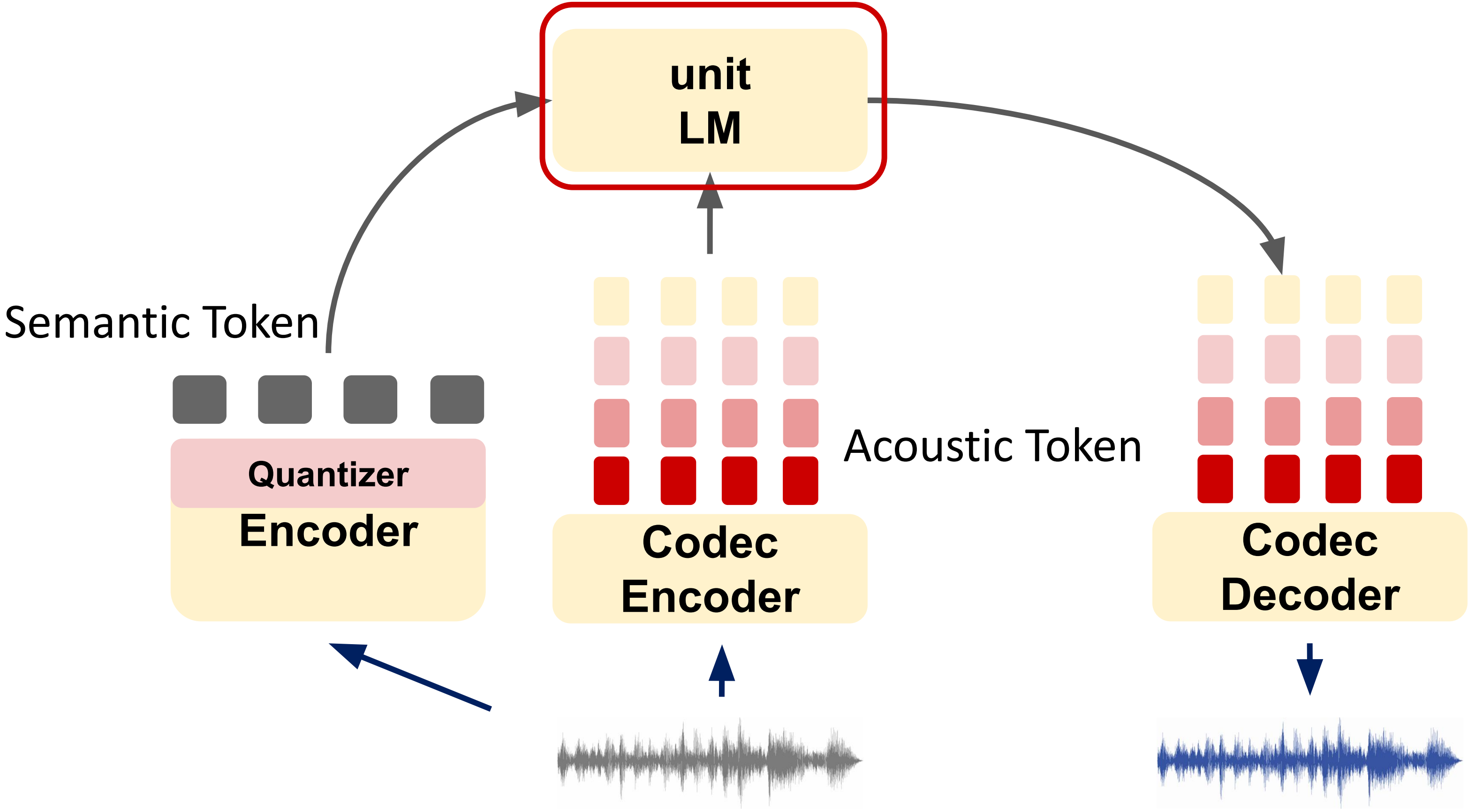


Residual Vector Quantization

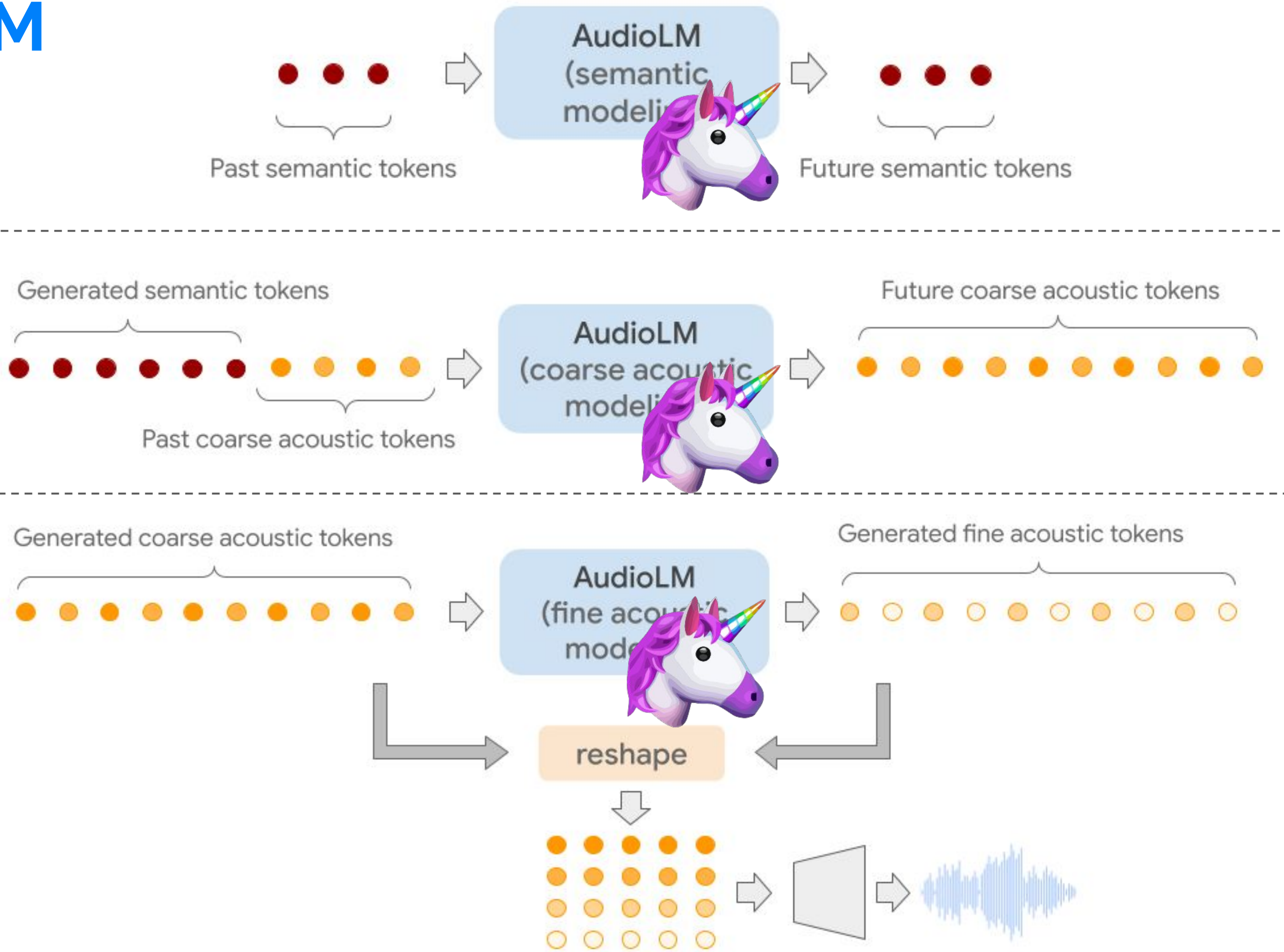


Défossez, Alexandre, et al. "High fidelity neural audio compression." *arXiv preprint arXiv:2210.13438* (2022).

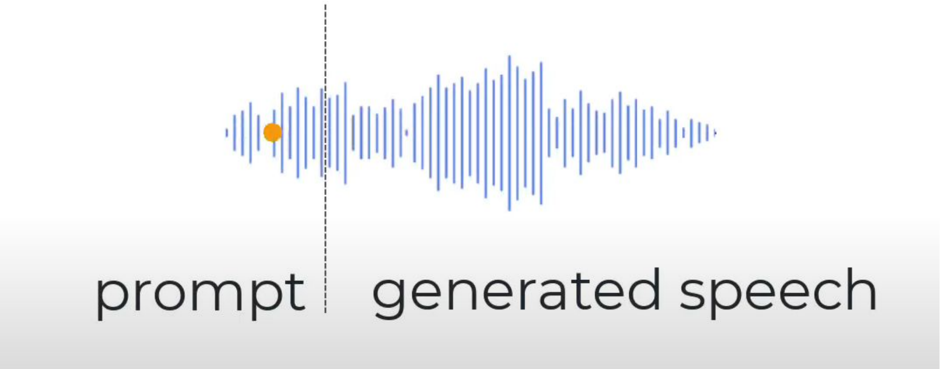
AudioLM


unit
LM
Semantic Token
Quantizer
Encoder
Codec
Encoder
Acoustic Token
Codec
Decoder

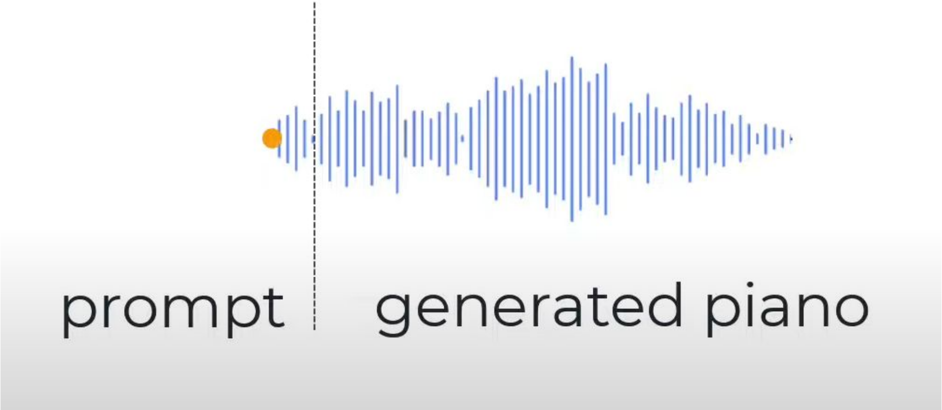
AudioLM



Past semantic tokens
Future semantic tokens
Generated semantic tokens
Future coarse acoustic tokens
Past coarse acoustic tokens
Generated coarse acoustic tokens
Generated fine acoustic tokens
reshape



Speech Continuation



Music Continuation

https://ai.googleblog.com/2022/10/audiolm-language-modeling-approach-to.html

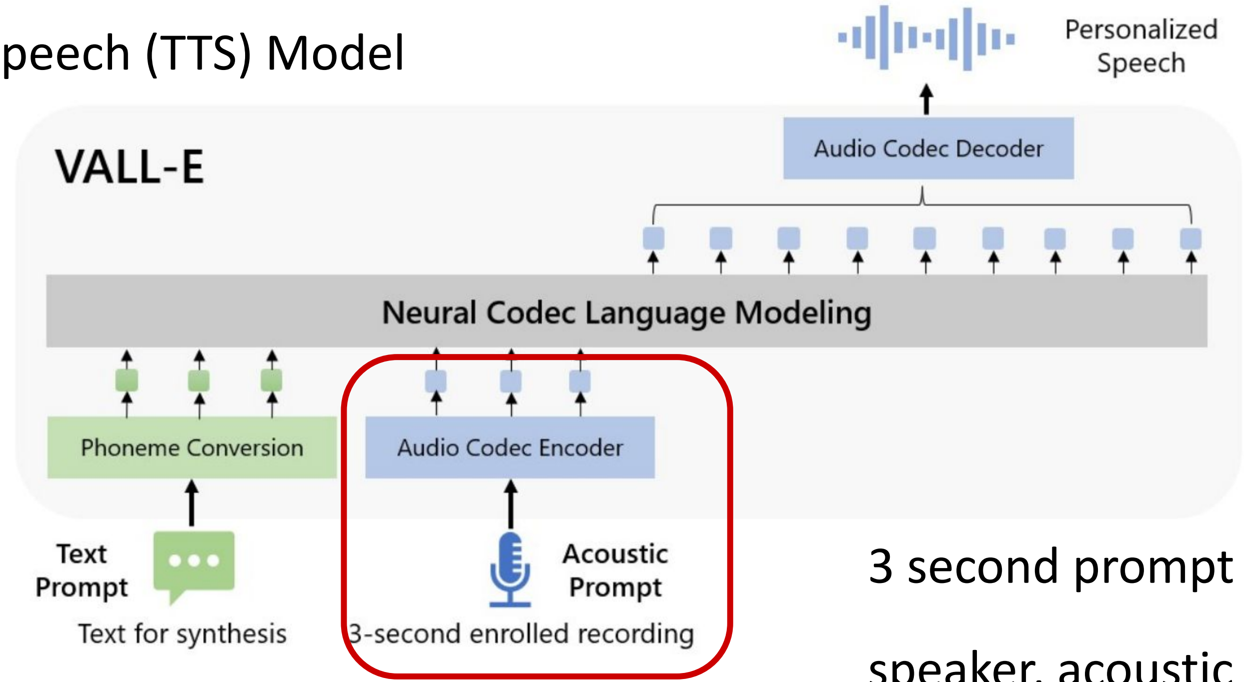
Overview

Speech Foundation Models

Part 1	Part 2	Part 3
Speech Representation Learning	**Speech Large Language Models**	Other Speech Foundation Models
1. SSL Models 2. Representation benchmarking	1. Textless NLP 2. AudioLM 3. VALL-E	1. Whisper 2. USM

VALL-E

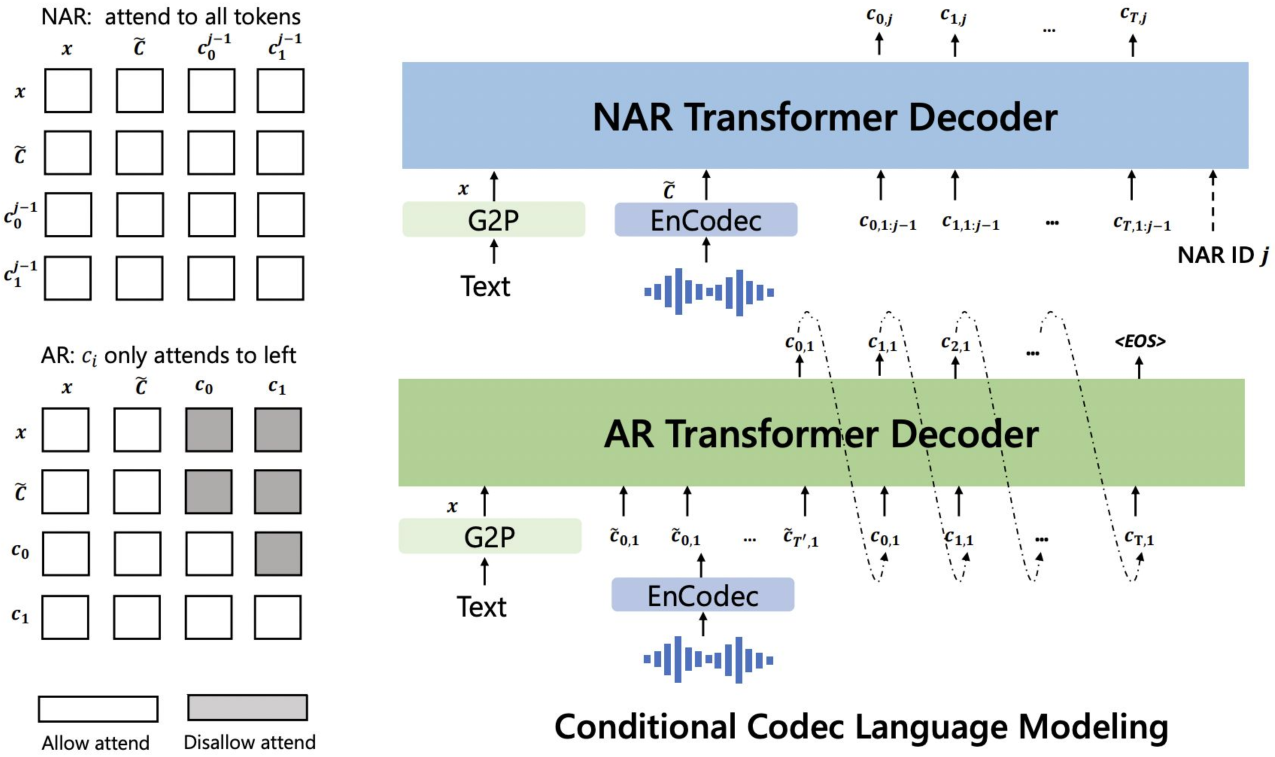
Text-to-speech (TTS) Model



3 second prompt

speaker, acoustic coherent

VALL-E



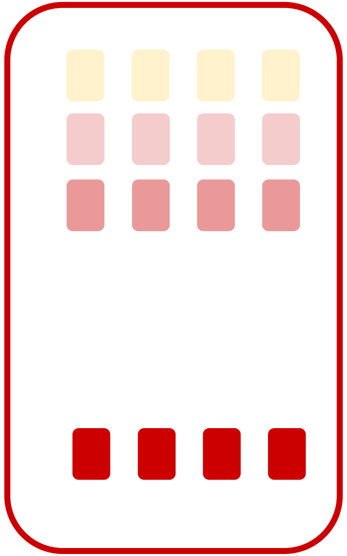
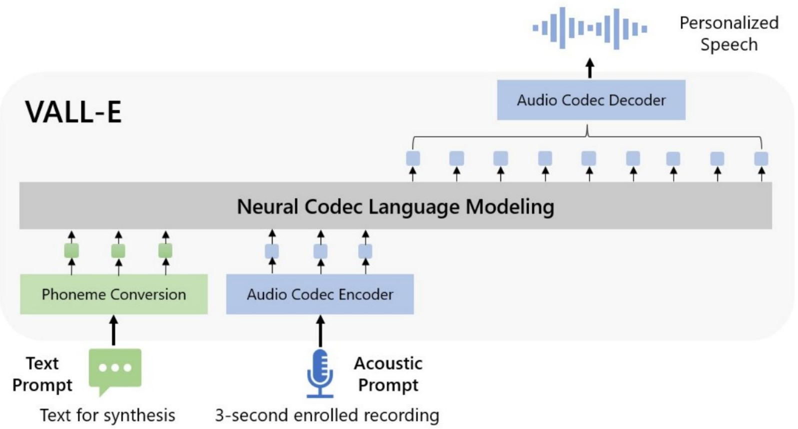


Figure 3: The structure of the conditional codec language modeling, which is built in a hierarchical manner. In practice, the NAR decoder will be called seven times to generate codes in seven quantizers.

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
Speech-to-Speech Systems		
GSLM	12.4	0.126
AudioLM*	6.0	-
TTS Systems		
YourTTS	7.7	0.337
VALL-E	5.9	**0.580**
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

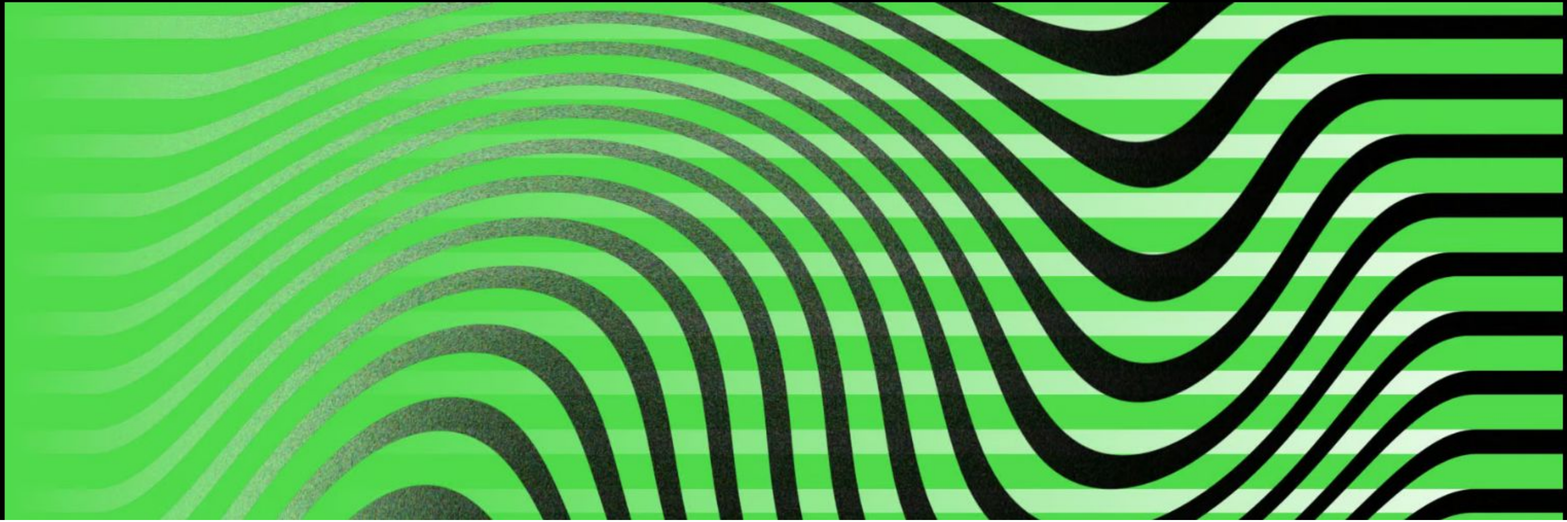
Part 3

Other Speech Foundation Models

1. Whisper
2. USM

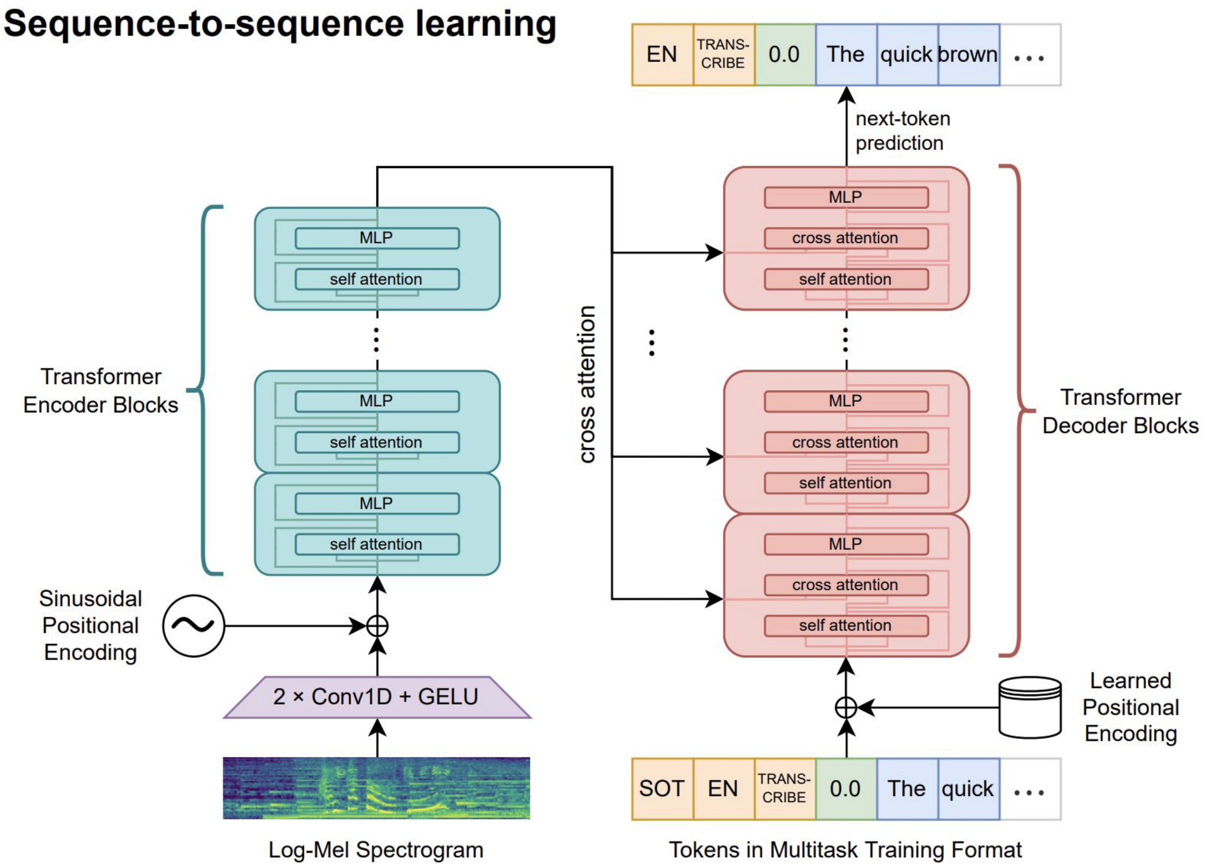
OpenAI Research Product Developers Safety Company Search

Introducing Whisper



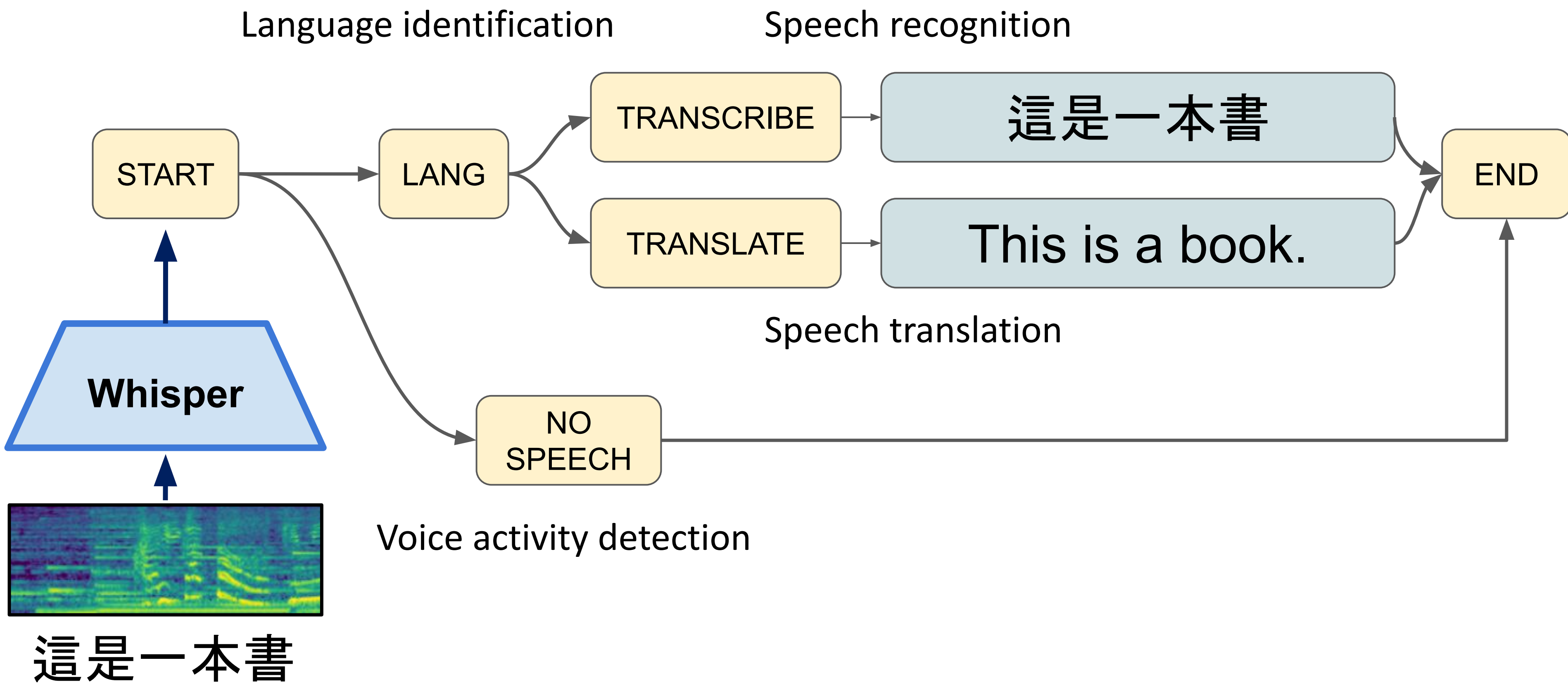
https://openai.com/research/whisper

Whisper



- 680,000 hours labeled data
- Multitask learning

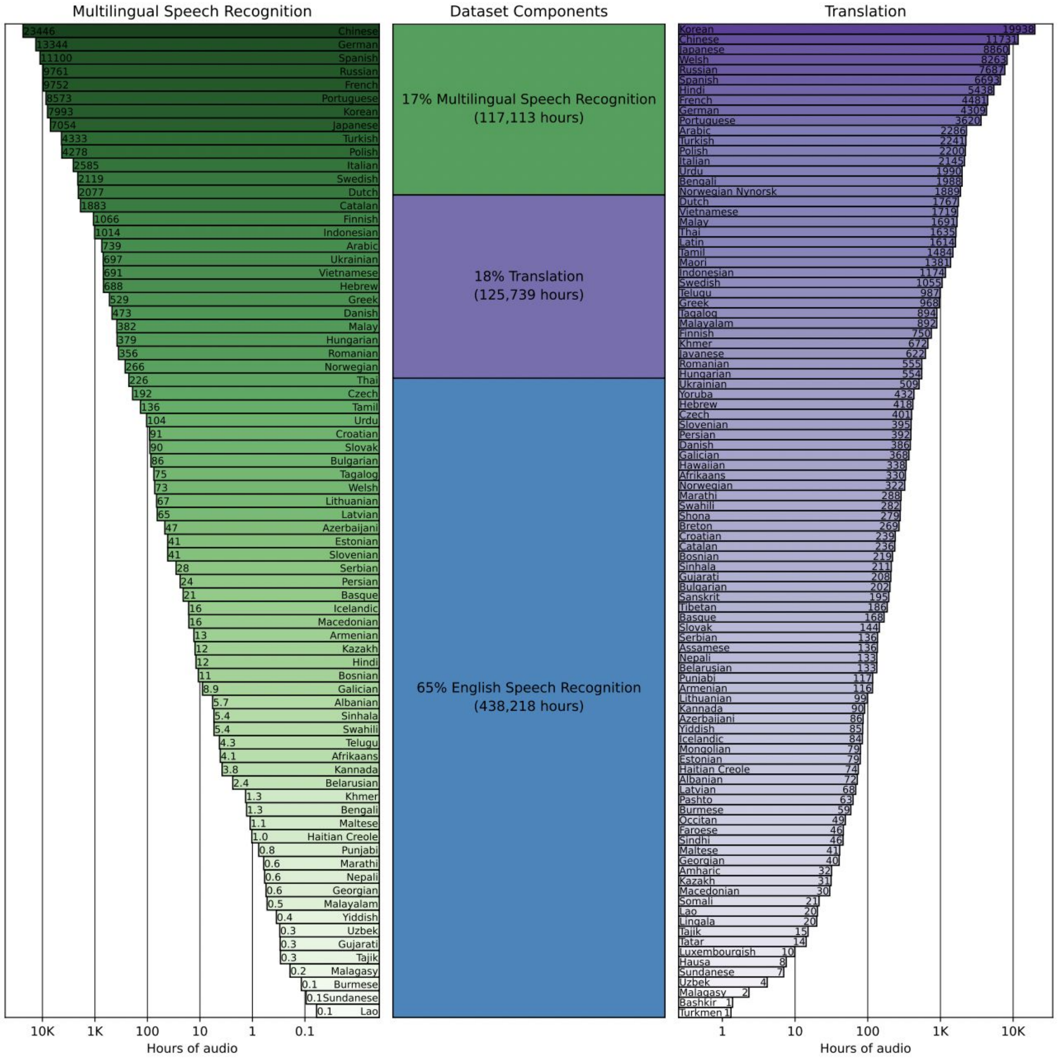
Whisper Multitasking



Multilingual Speech Recognition

(~120,000 hours)

Chinese
German
Spanish
…



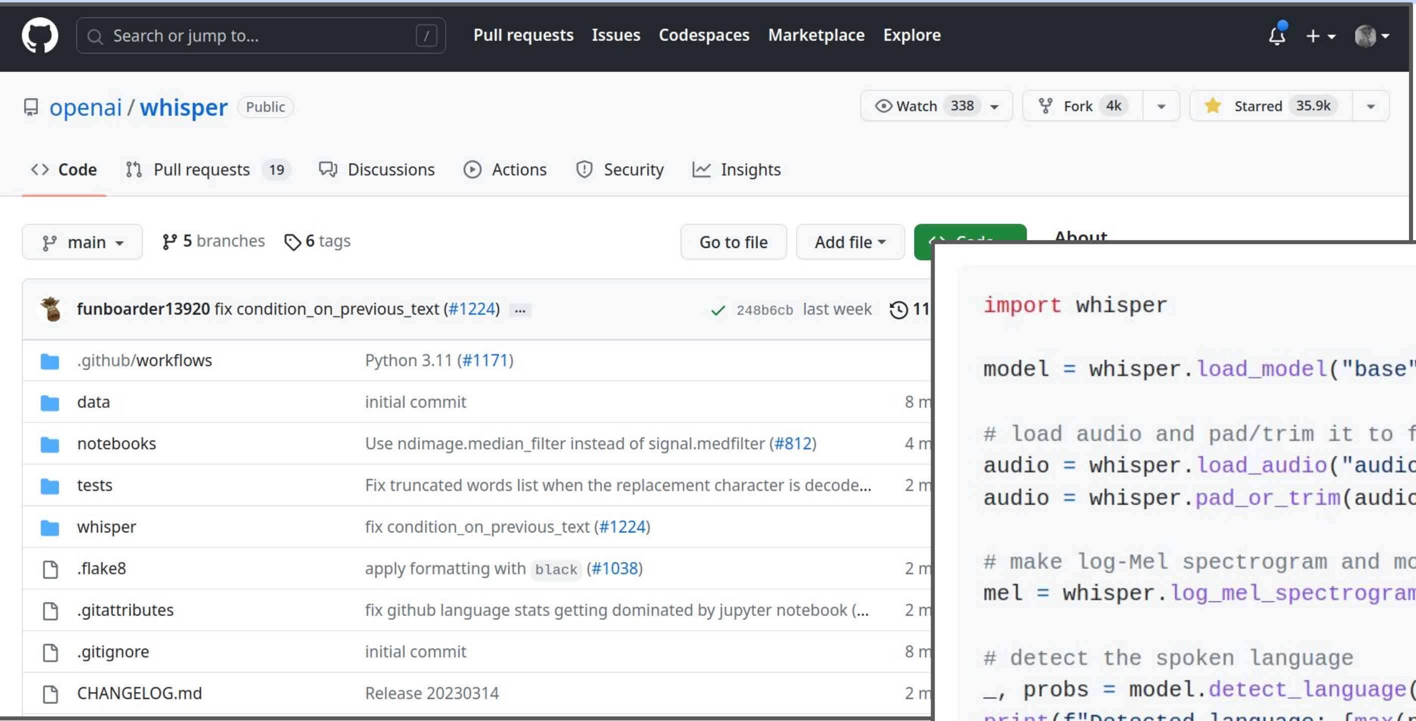
X to English Translation

(~120, 000 hours)

Korean
Chinese
Japanese
…

English Speech Recognition

(~440,000 hours)

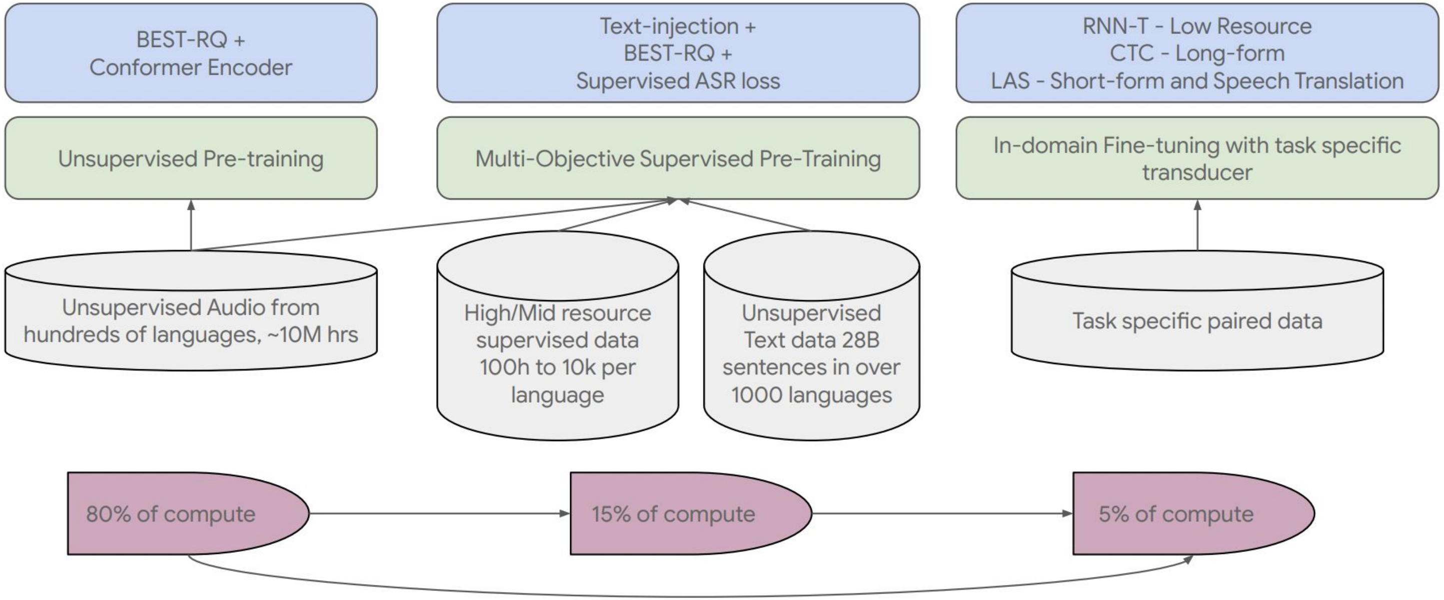


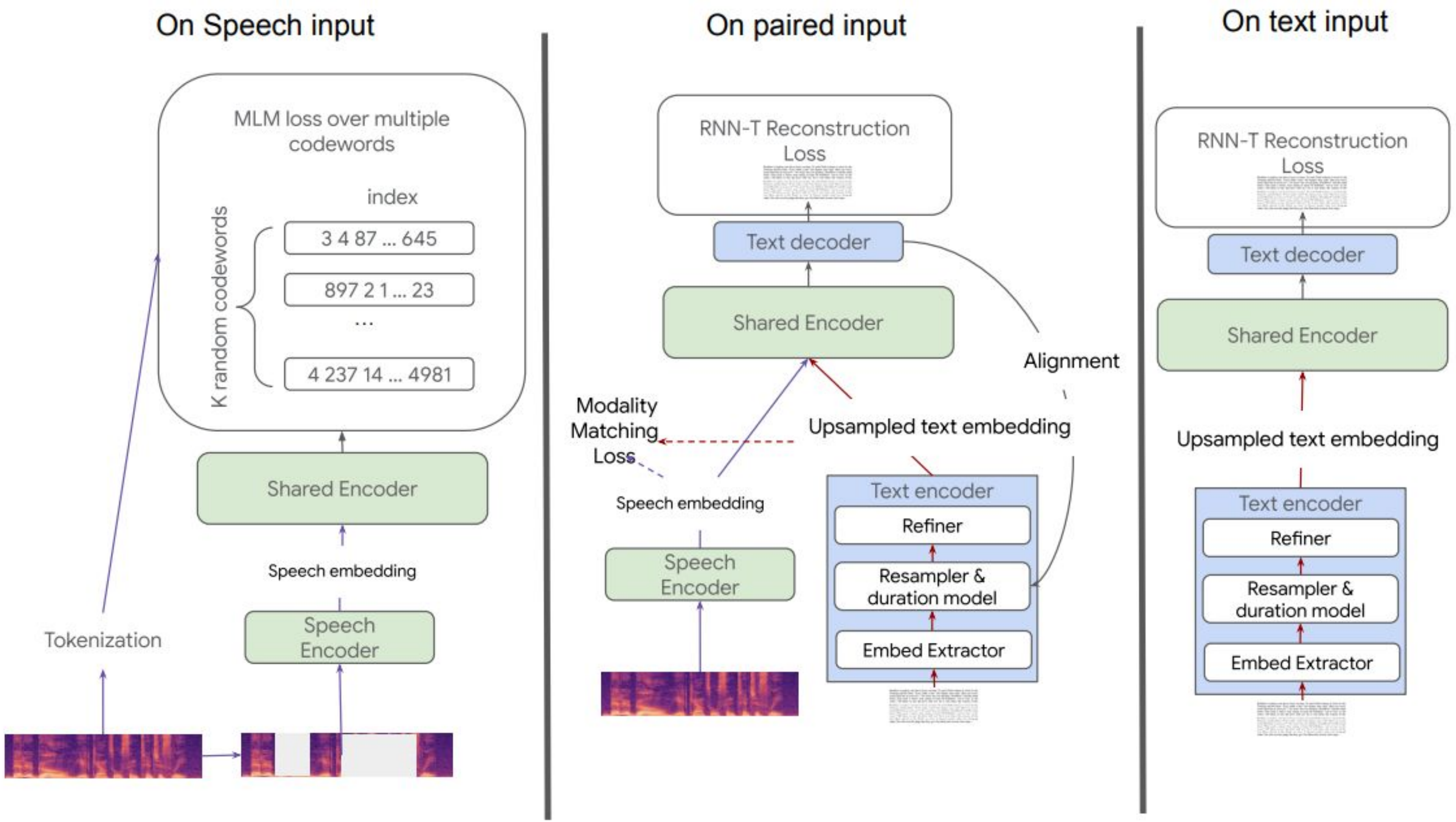

```python
import whisper

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)
```

Whisper is open sourced!

USM: Universal Speech Model

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





On Speech input
MLM loss over multiple codewords
index
K random codewords
3 4 87 ... 645
897 2 1 ... 23
...
4 237 14 ... 4981
Shared Encoder
Speech embedding
Speech Encoder
Tokenization
On paired input
RNN-T Reconstruction Loss
Text decoder
Shared Encoder
Alignment
Modality Matching Loss
Upsampled text embedding
Speech embedding
Speech Encoder
Text encoder
Refiner
Resampler & duration model
Embed Extractor
On text input
RNN-T Reconstruction Loss
Text decoder
Shared Encoder
Upsampled text embedding
Text encoder
Refiner
Resampler & duration model
Embed Extractor

Task	Multilingual Long-form ASR				Multidomain en-US	Multilingual ASR	
Dataset	YouTube			CORAAL	SpeechStew	FLEURS	
Langauges	en-US	18	73	en-US	en-US	62	102
Prior Work (single model)							
Whisper-longform	17.7	27.8	-	23.9	12.8		
Whisper-shortform[†]	-	-	-	13.2[‡]	11.5	36.6	-
Our Work (single model)							
USM-LAS	14.4	19.0	29.8	**11.2**	**10.5**	**12.5**	-
USM-CTC	**13.7**	**18.7**	**26.7**	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