Microsoft

Unified Speech and Text Pre-Training

Long Zhou (周龙)

lozhou@microsoft.com

Natural Language Computing Group, MSRA

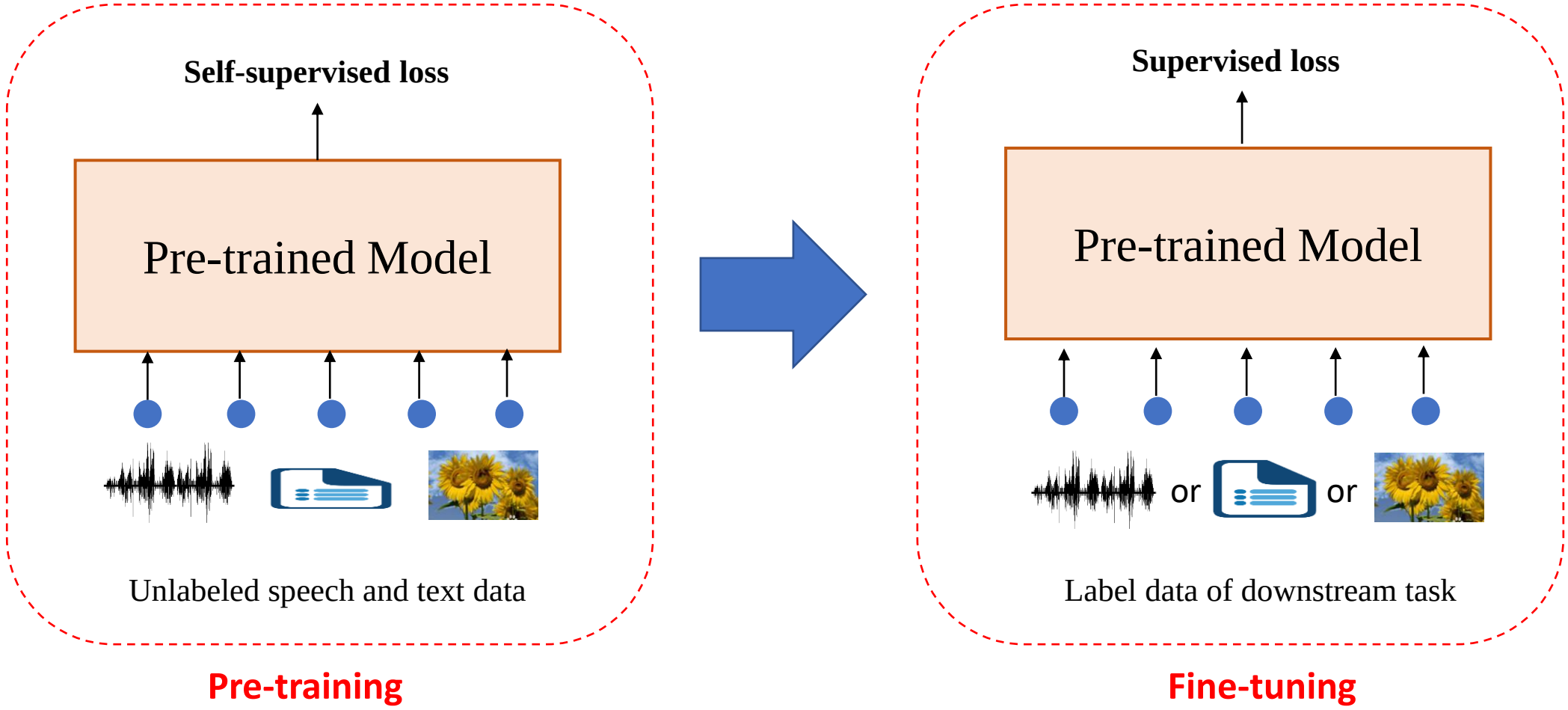
Outline

- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

Outline

- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

Pre-Training Method: A New Paradigm

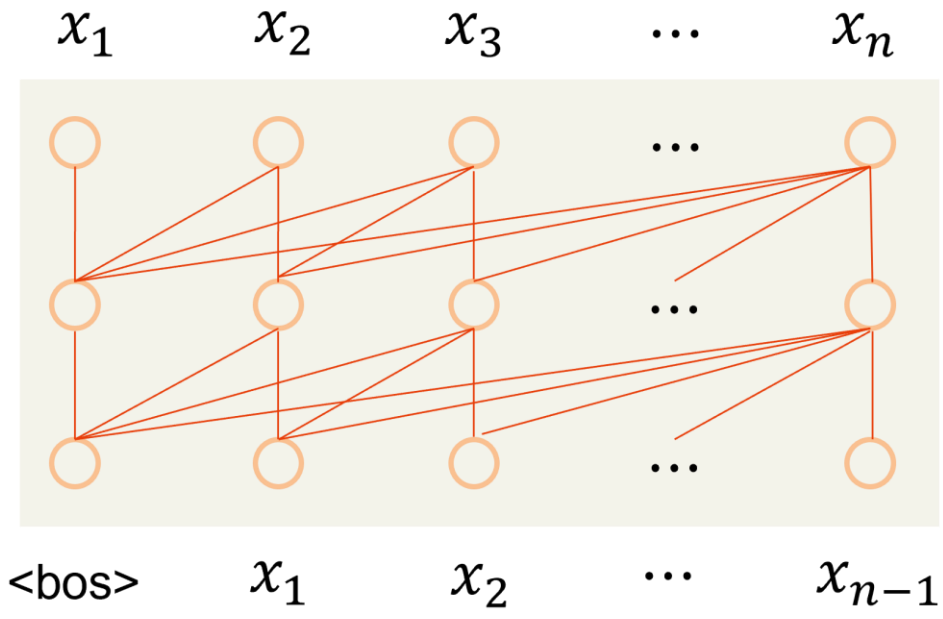


Why Pre-Trained Models?

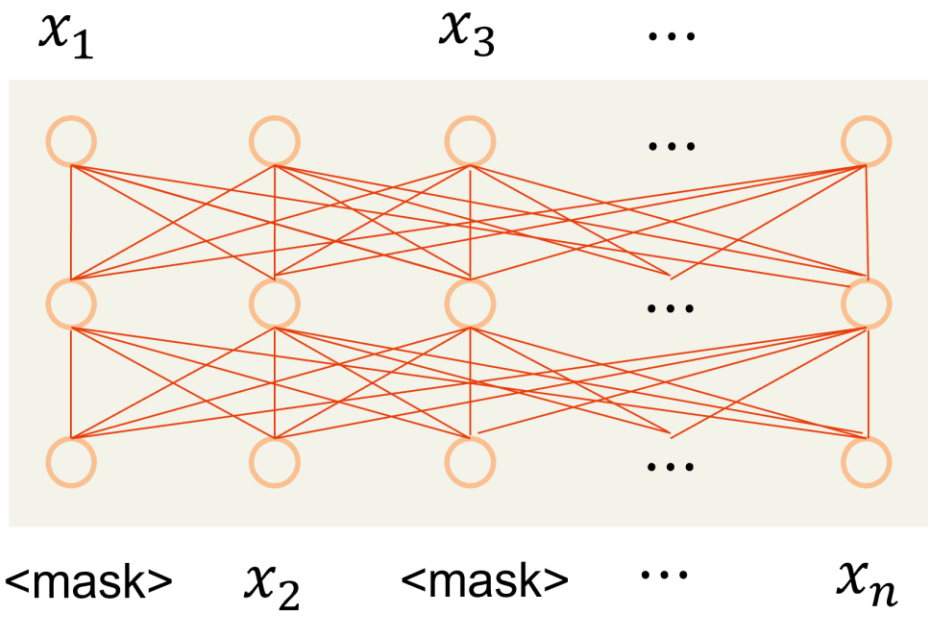
- Pre-trained models capture task-agnostic general knowledge from massive labeled and unlabeled data.
- Pre-trained models transfer learned knowledge to downstream tasks, and support almost all NLP/Speech tasks.
- Pre-trained models provide a scalable solution to various applications, which require less training and require less effort.

Text Pre-Training

- GPT-1/2/3
- BERT

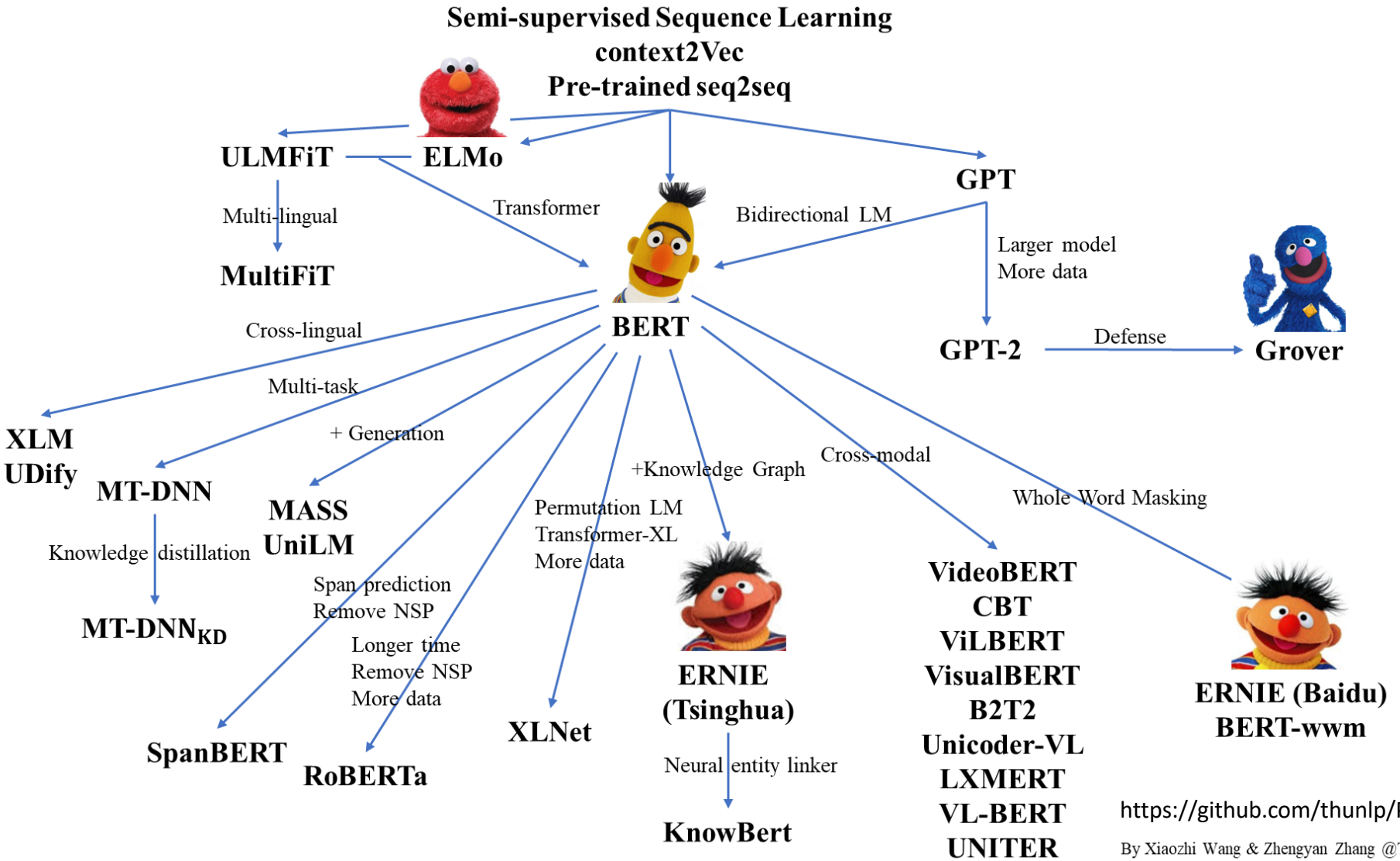


GPT



BERT

Text Pre-Training



https://github.com/thunlp/PLMpapers

By Xiaozhi Wang & Zhengyan Zhang @THUNLP

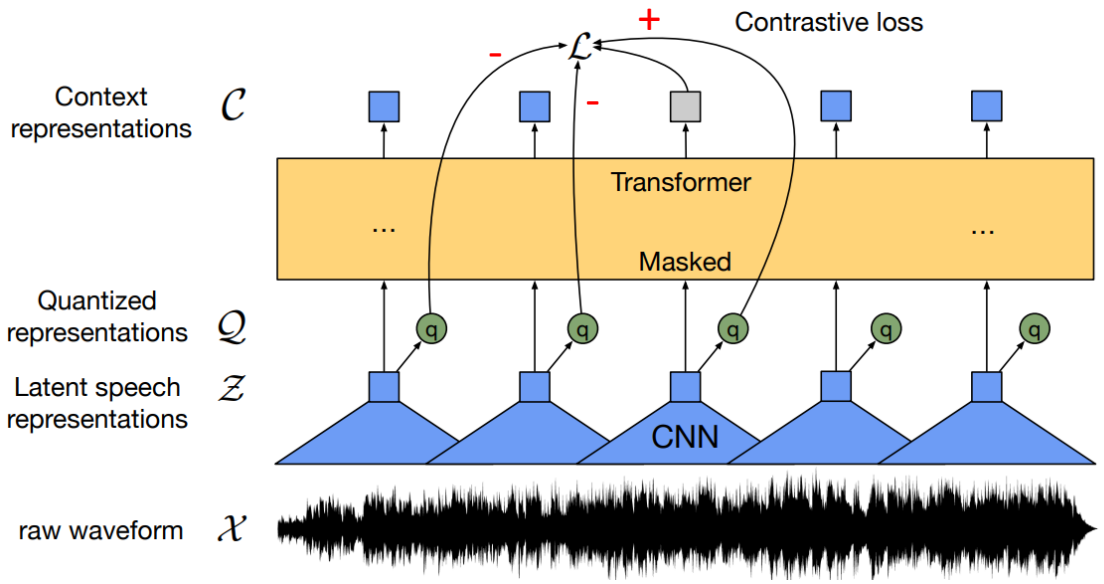
Speech Pre-Training

- Speech is unique and different from text
  - Speech inputs are much longer than text, e.g., 1s has 16000 frames for 16K HZ audio.
  - Speech inputs are continuous without a predefined dictionary (like the vocabulary of text data).
  - Speech inputs vary in the sentence length without segment boundaries.
  - Speech inputs contain more information than text, e.g., content and speaker information.

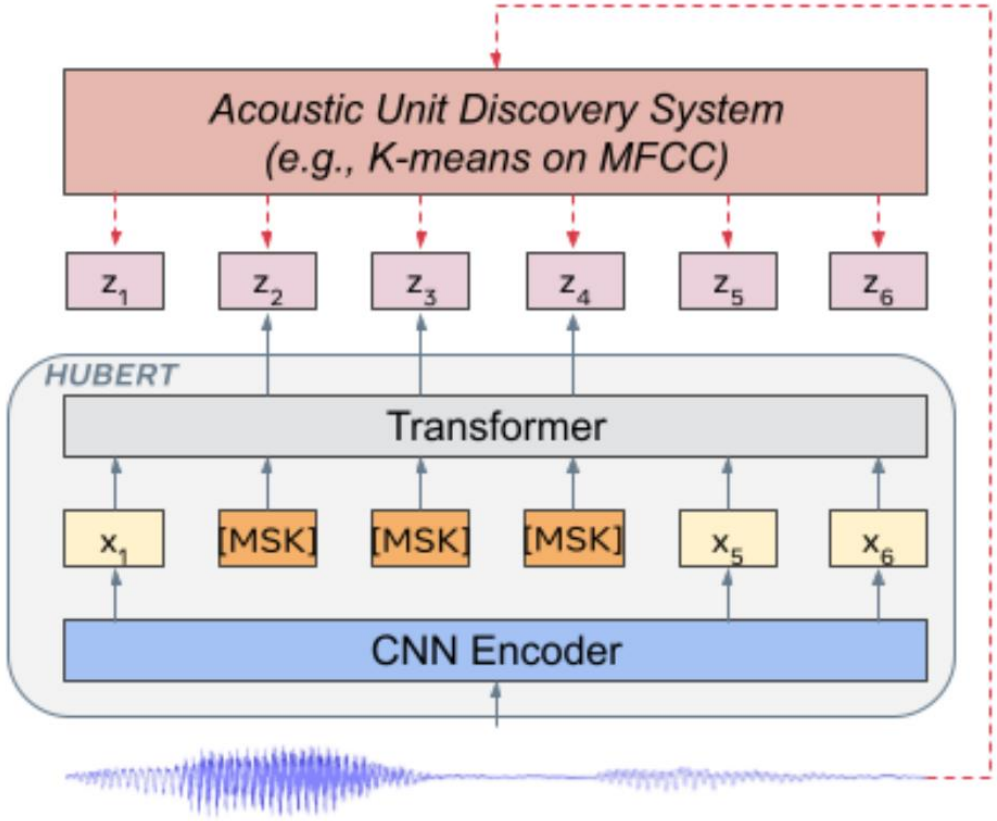
Speech is different from text.

Speech Pre-Training

- wav2vec 2.0
- HuBERT



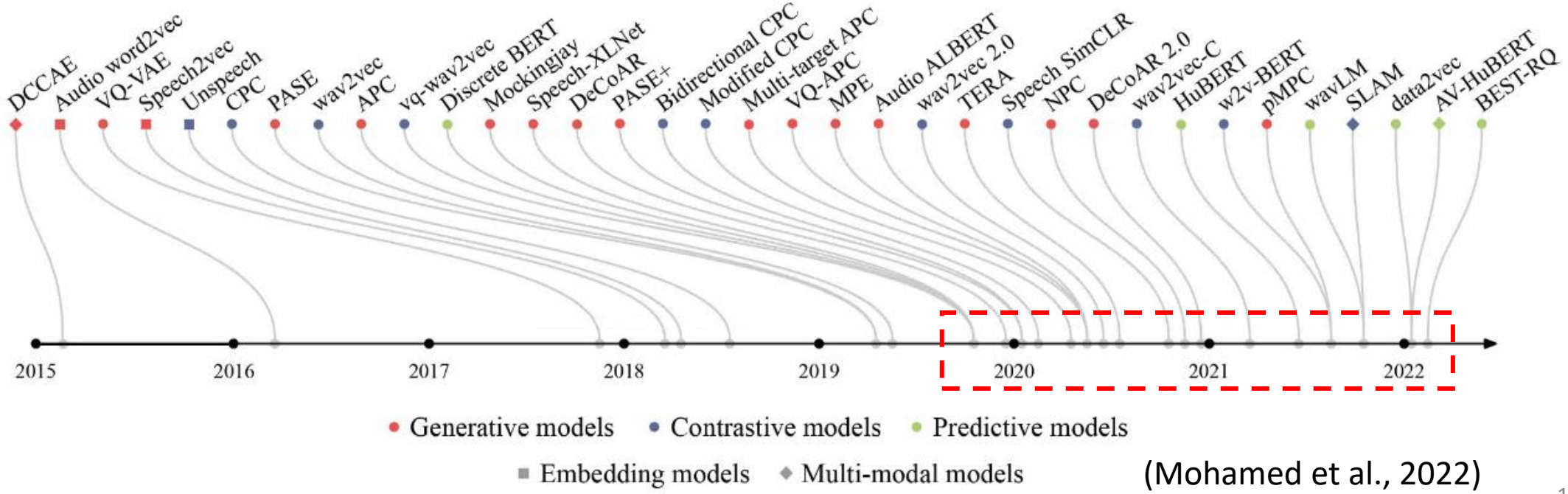
wav2vec 2.0



HuBERT

Speech Pre-Training

- Contrastive methods (e.g., CPC, wav2vec 2.0)
- Predictive methods (e.g., HuBERT, WavLM)
- Generative methods (e.g., APC, Audio-MAE)

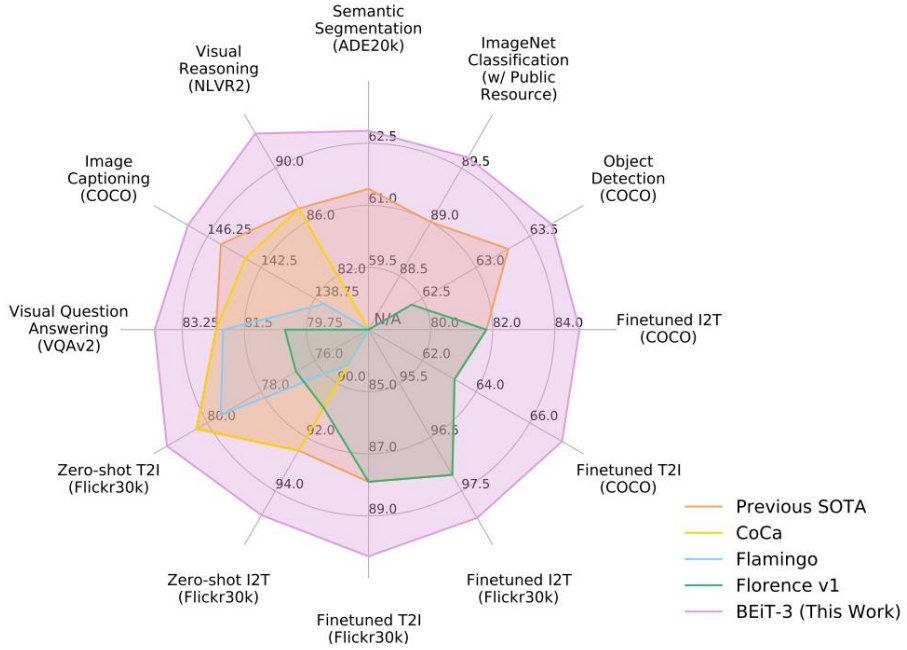


(Mohamed et al., 2022)

The Big Convergence

- Model architecture
  - Transformers becomes the de facto backbone networks across AI areas like NLP, CV, and Speech.
- Training paradigm
  - Mainstream: pre-training then fine-tuning
  - Others: prompt learning
- Pre-training tasks
  - Self-supervised pre-training tasks converge across different modalities
  - Generative learning: language to vision and audio
  - Contrastive learning: vision to language and audio

The Big Convergence



BEiT-3: [Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks]

TorchScale - A Library for Transformers at (Any) Scale

license MIT | pypi package 0.1.1

TorchScale is a PyTorch library that allows researchers and developers to scale up Transformers efficiently and effectively. It has the implementation of fundamental research to improve modeling generality and capability as well as training stability and efficiency of scaling Transformers.

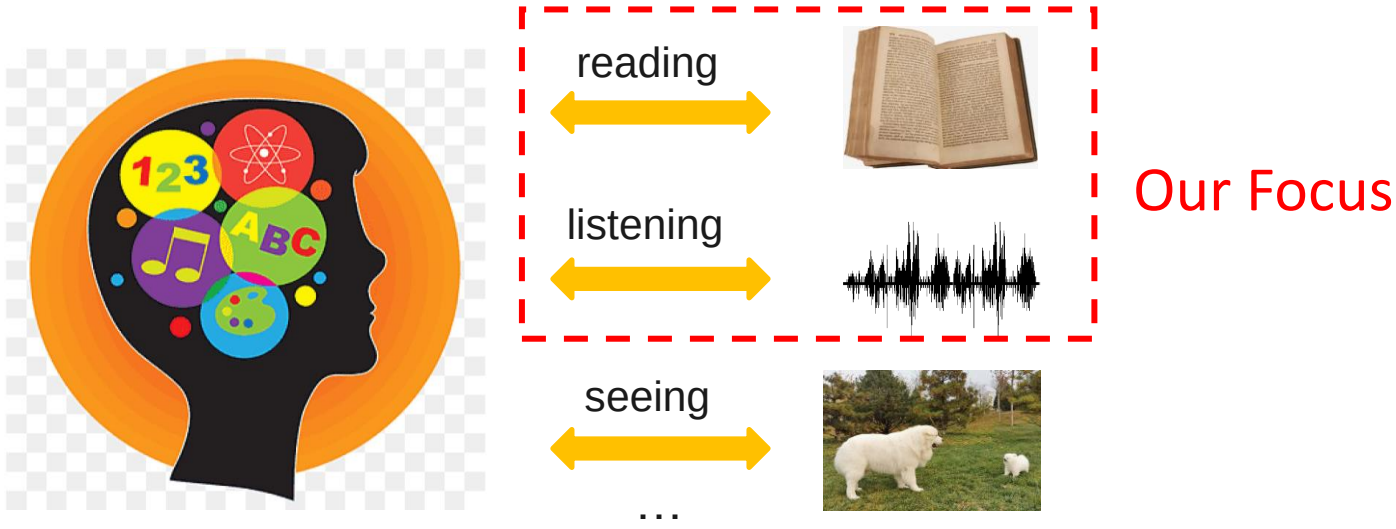
- Stability - DeepNet: scaling Transformers to 1,000 Layers and beyond
- Generality - Foundation Transformers (Magneto): towards true general-purpose modeling across tasks and modalities (including language, vision, speech, and multimodal)
- Efficiency - X-MoE: scalable & finetunable sparse Mixture-of-Experts (MoE)

Github: https://github.com/microsoft/torchscale

TORCHSCALE: [Transformers at Scale]

Motivation of Our Work

- The convergence is a big trend across different modalities, e.g., text, speech, image, and video.
- Speech and text are two important carriers of human communication. They are two similar modalities with natural alignment relationship.
- Push the convergence of speech and text.

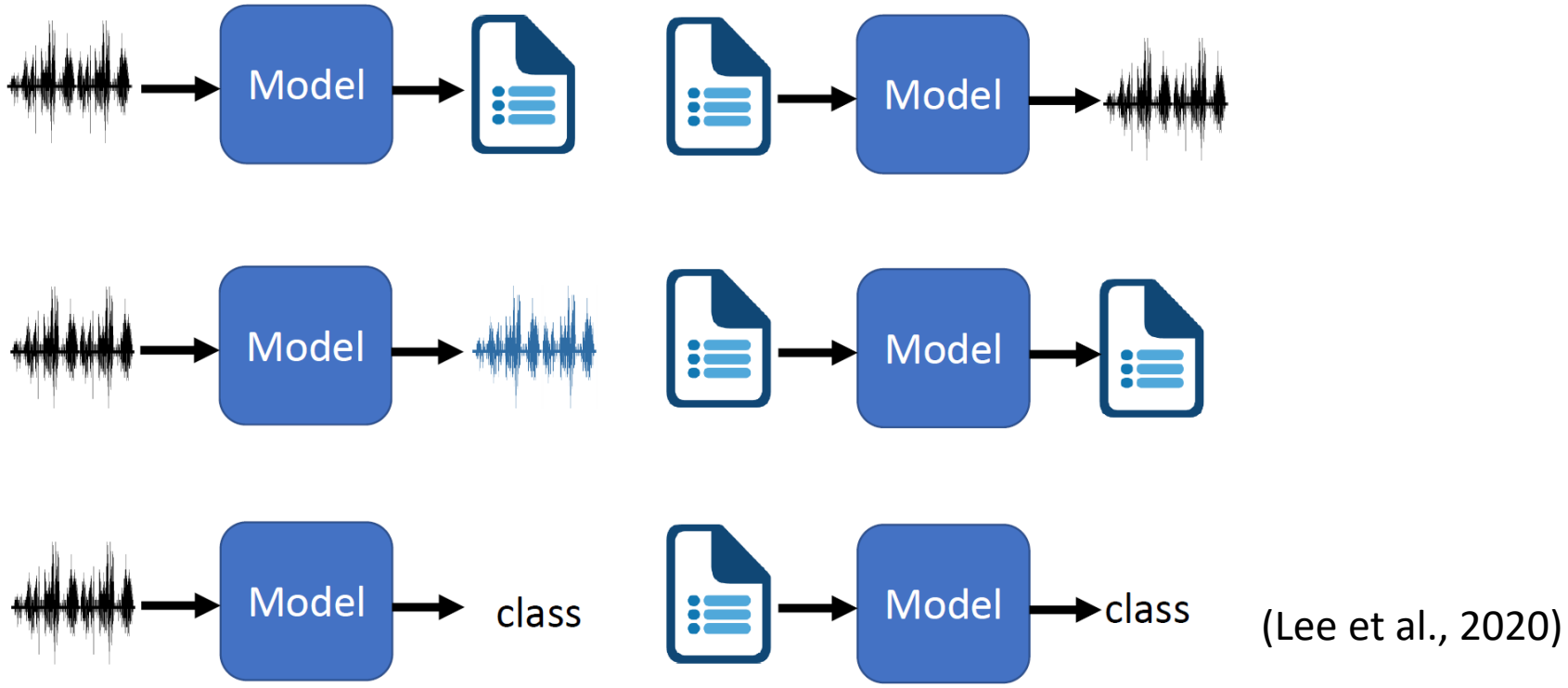


Outline

- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

Background: Spoken Language Processing

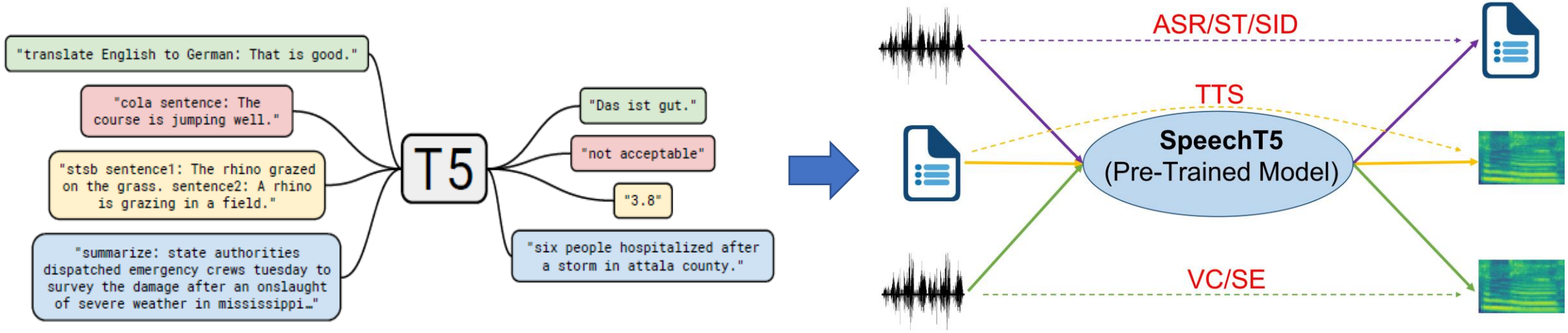
- Speech/Text → Speech/Text



(Lee et al., 2020)

SpeechT5: Motivation

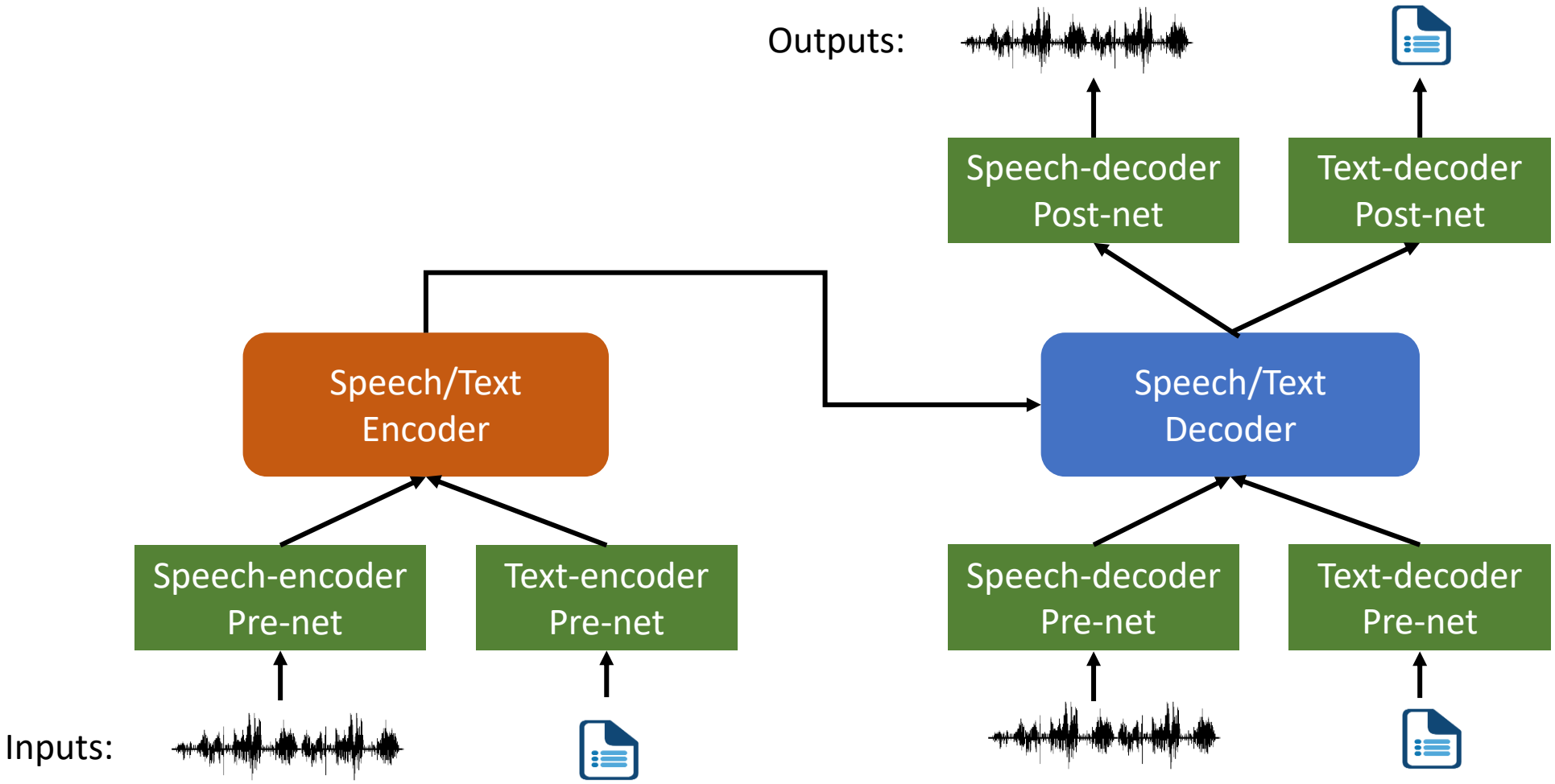
- How to pre-train a model for all spoken language tasks?
  - Convert all spoken language tasks as a speech/text to speech/text problem
  - Pre-train a single encoder-decoder model with unlabeled speech and text data



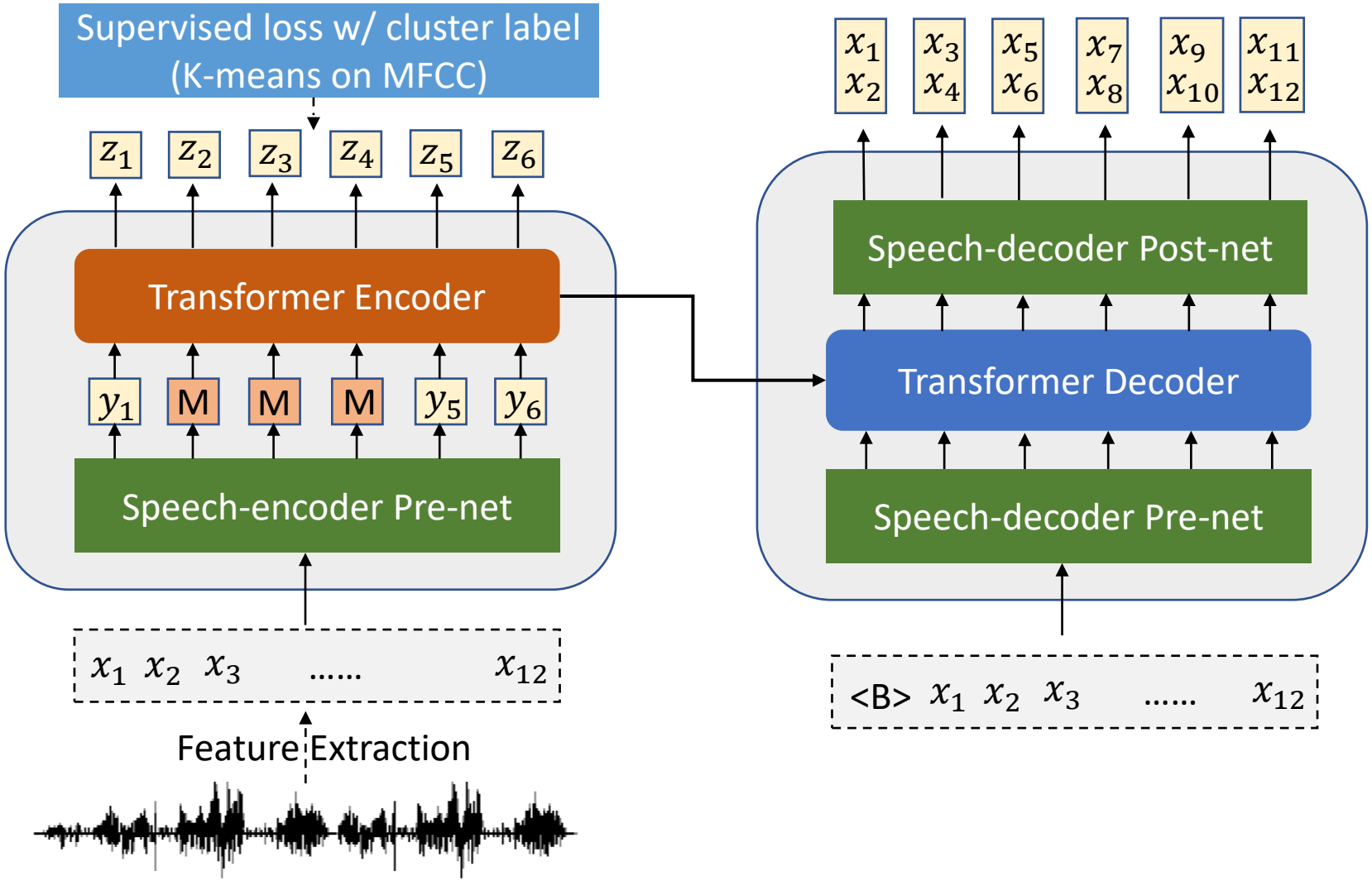
T5 (Raffel et al., 2019)

Our SpeechT5

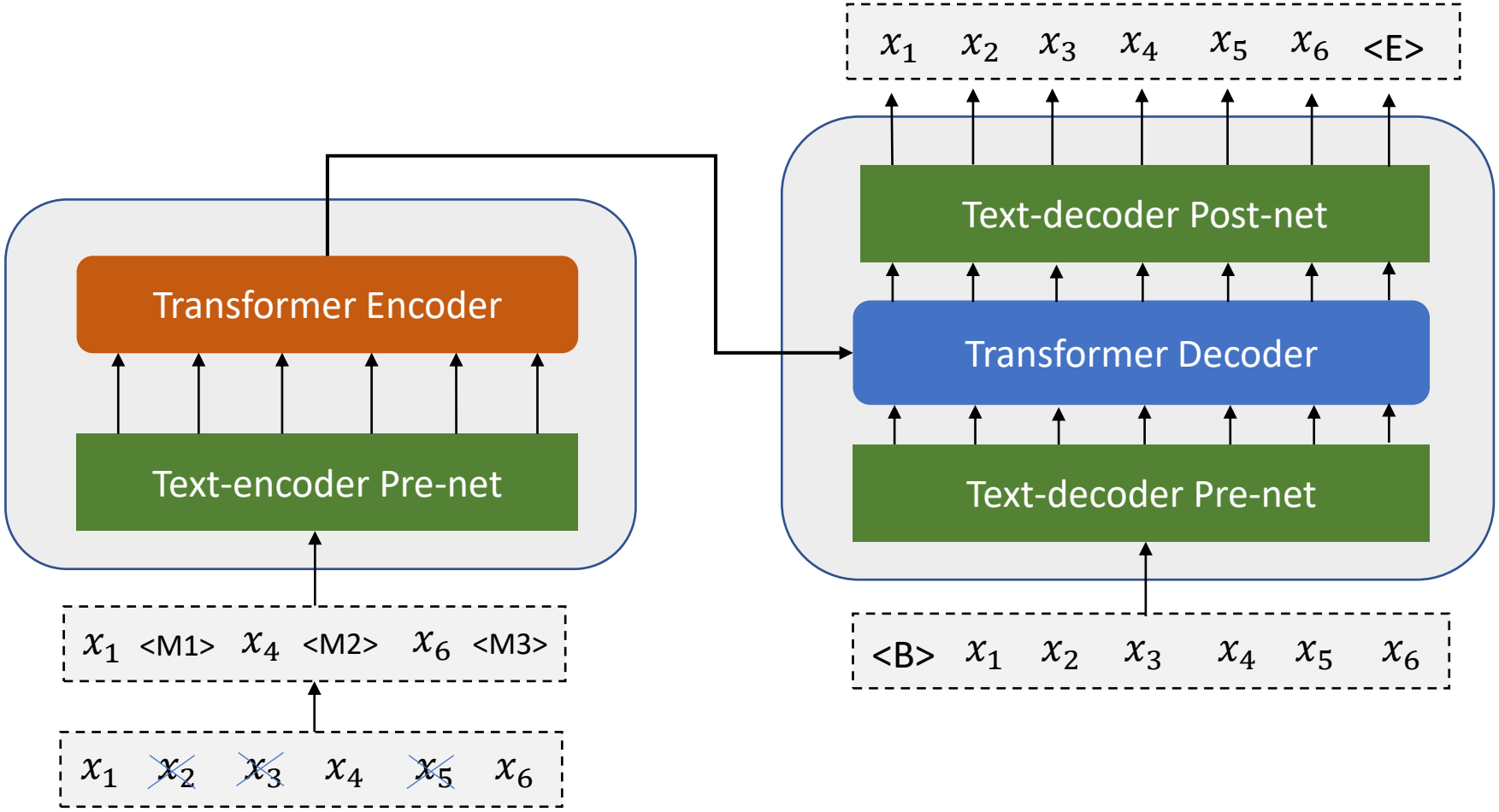
SpeechT5: Model Structure



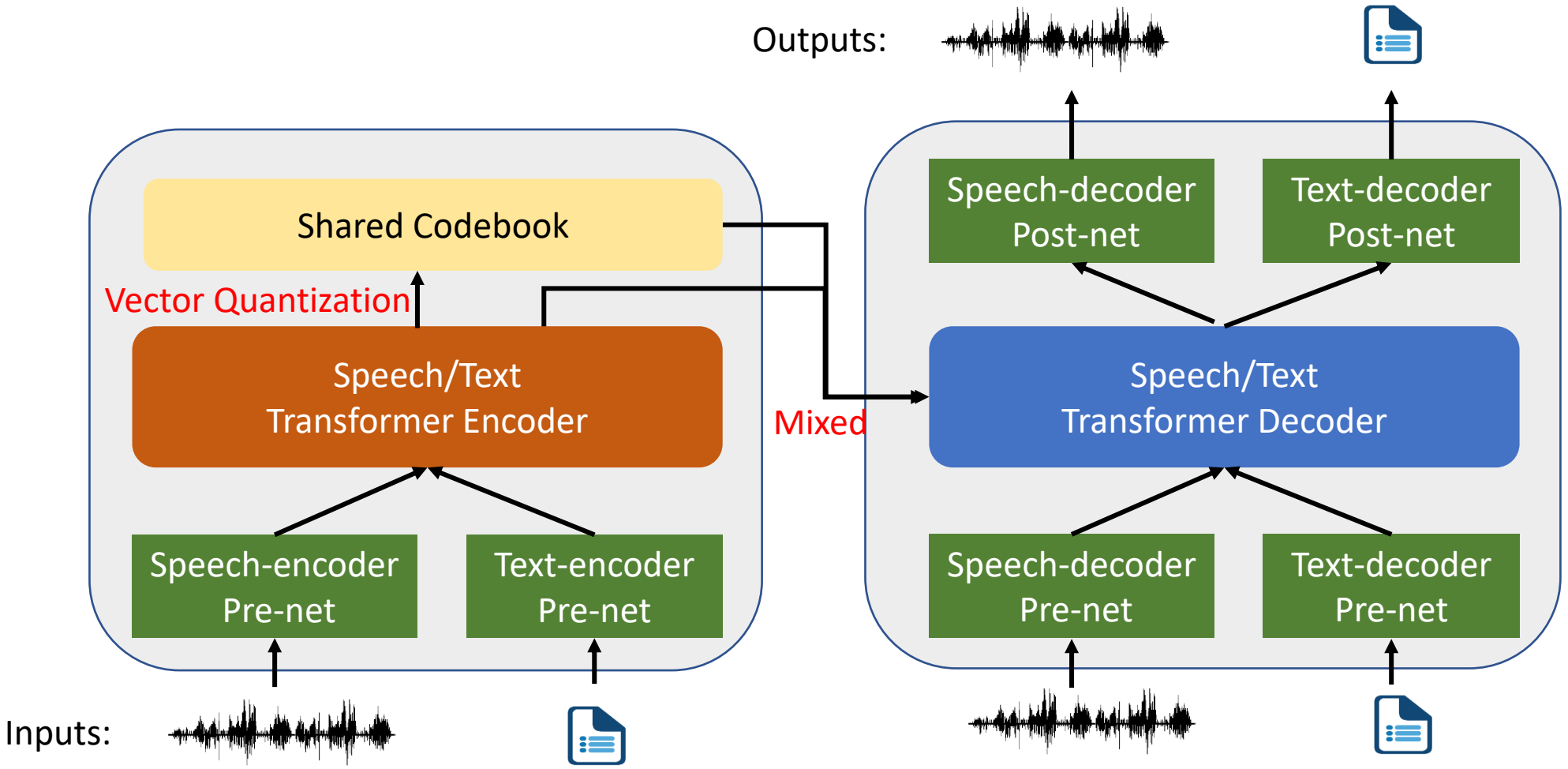
Pre-Train with Unlabeled Speech



Pre-Train with Unlabeled Text



Joint Pre-Train with Speech and Text



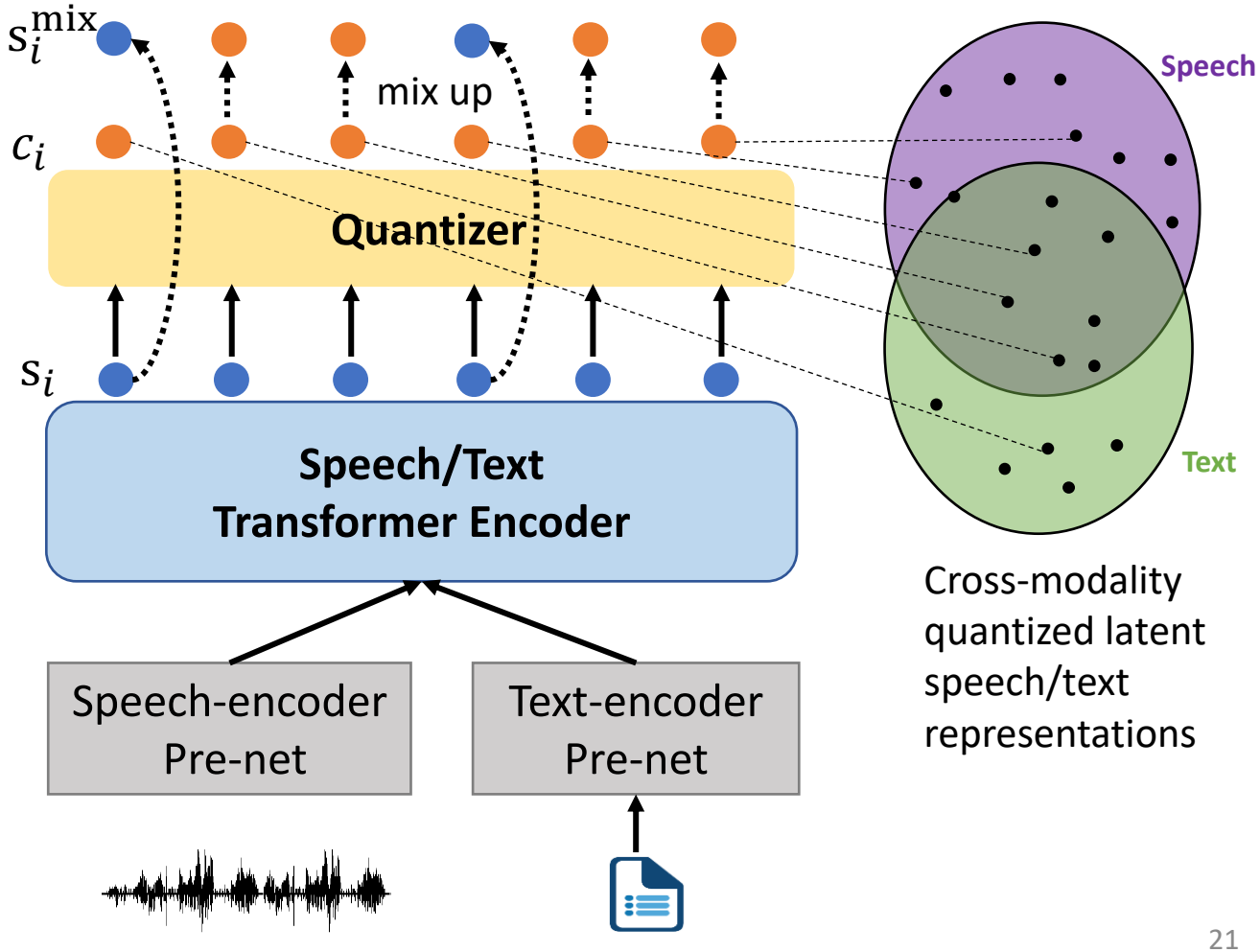
Joint Pre-Train with Speech and Text

- Step 1: vector quantization

$$\mathbf{c}_i = \arg\min_{j \in [K]} \|\mathbf{s}_i - \mathbf{c}_j\|_2$$

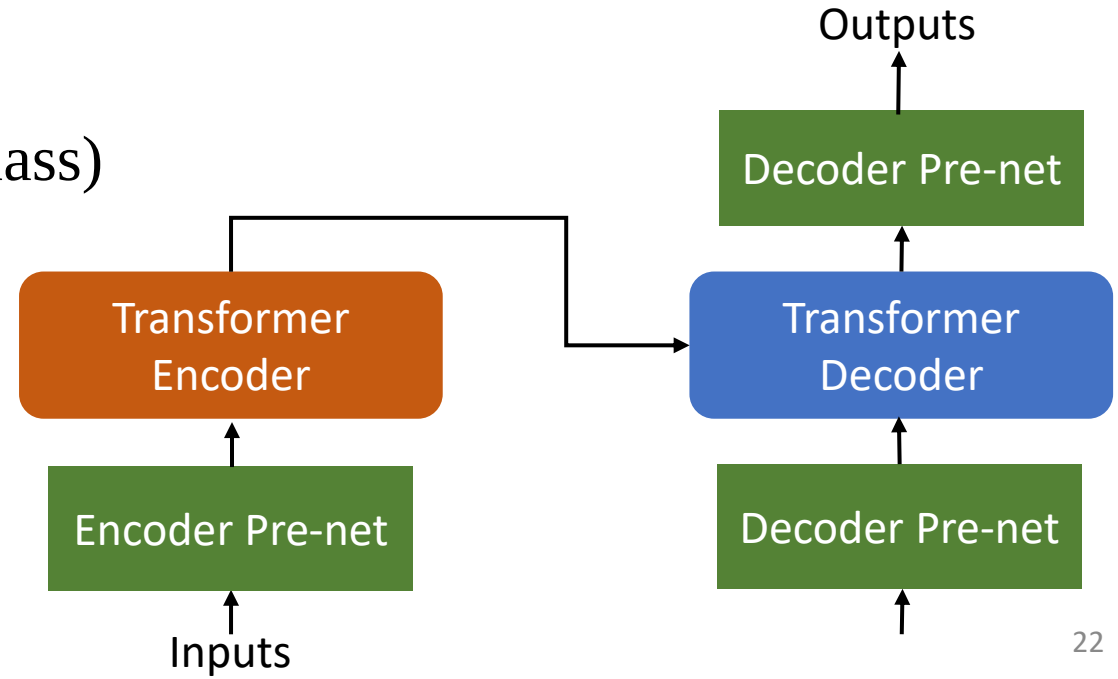
- Step 2: randomly mix up

$$\mathbf{s}_i^{mix} = \begin{cases} \mathbf{c}_i, & \text{if } \theta \leq 0.1 \\ \mathbf{s}_i, & \text{others} \end{cases}$$

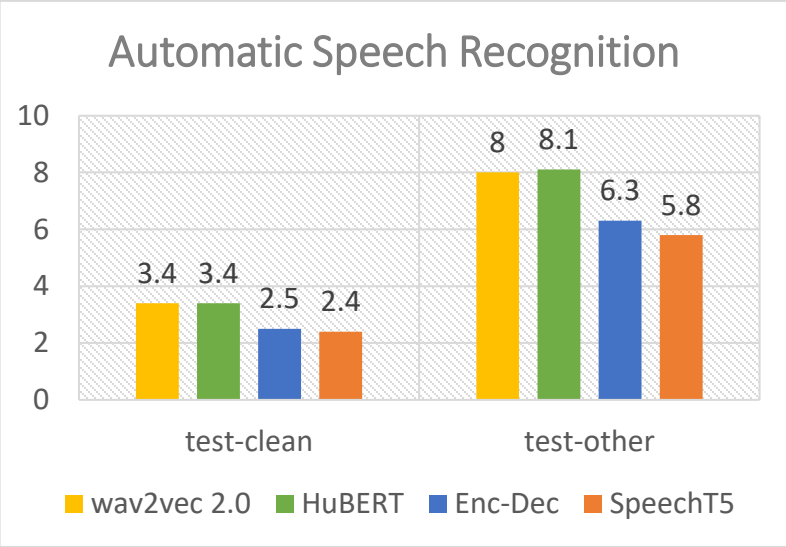


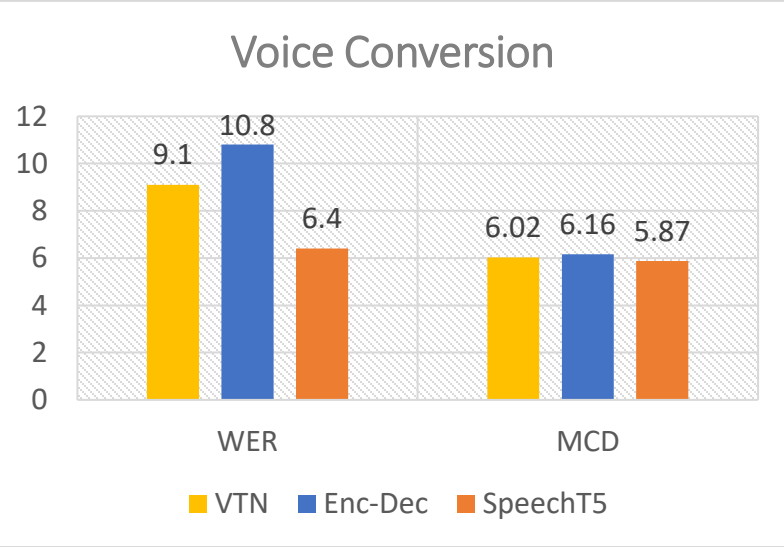
SpeechT5: Fine-Tune

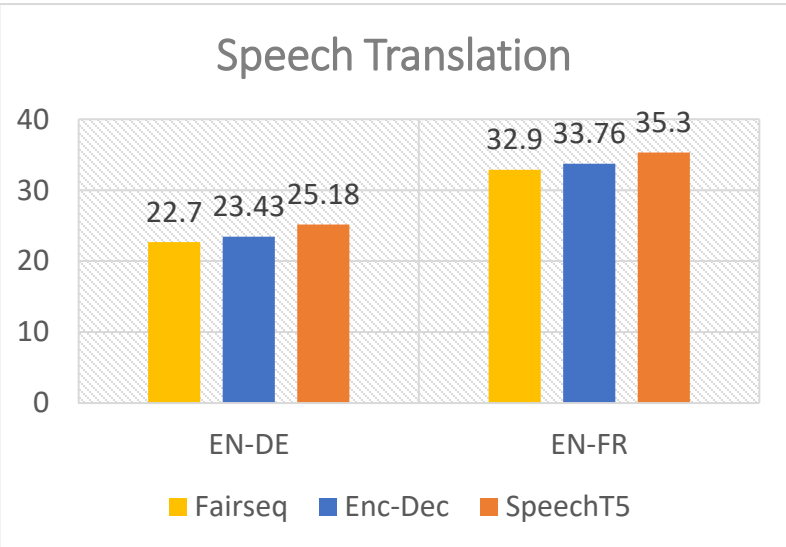
- Fine-tuning tasks
  - Speech conversion (speech to speech)
  - Automatic speech recognition (speech to text)
  - Speech synthesis (text to speech)
  - Speech Identification (Speech to class)
  - ……

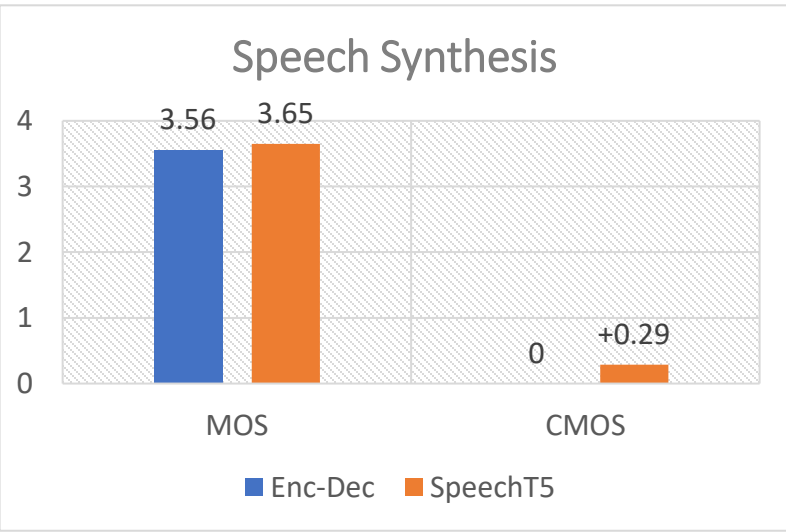


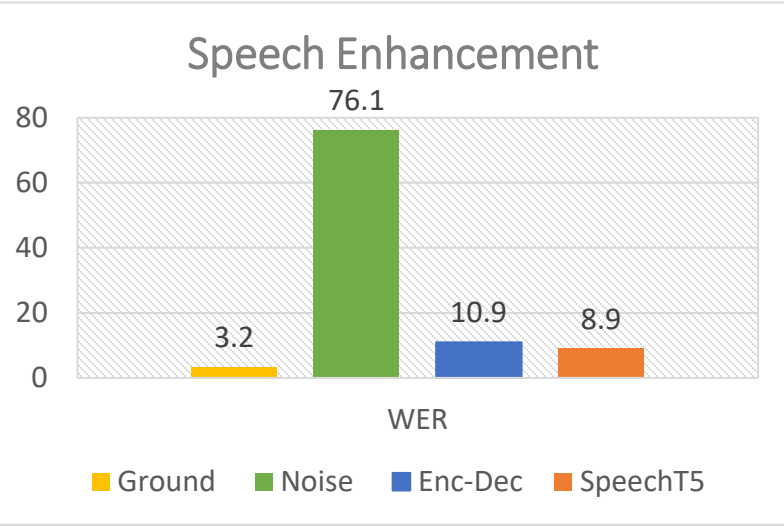
SpeechT5: Evaluation

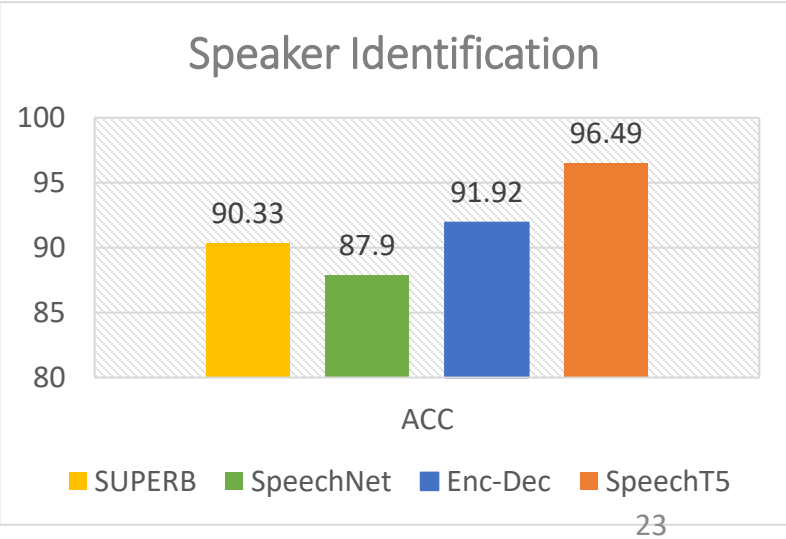












SpeechT5: Ablation Study

- Verify the effectiveness of different pre-training losses
  - Speech, text, and joint pre-training methods are important to SpeechT5.
  - Speech pre-training is more critical than text pre-training on these tasks that need to encode speech.

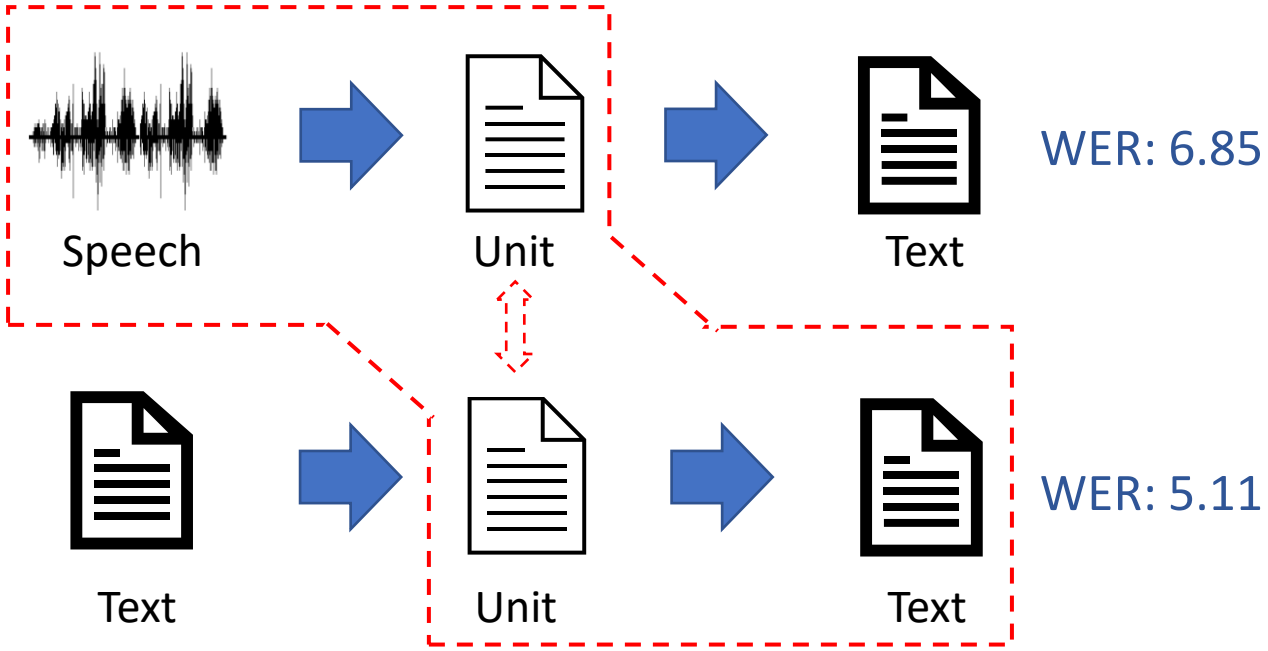
Model	ASR		VC	SID
	clean	other		
SpeechT5	4.4	10.7	5.93	96.49%
w/o Speech PT	-	-	6.49	38.61%
w/o Text PT	5.4	12.8	6.03	95.60%
w/o Joint PT	4.6	11.3	6.18	95.54%
w/o $\mathcal{L}^s_{mlm}$	7.6	22.4	6.29	90.91%

Outline

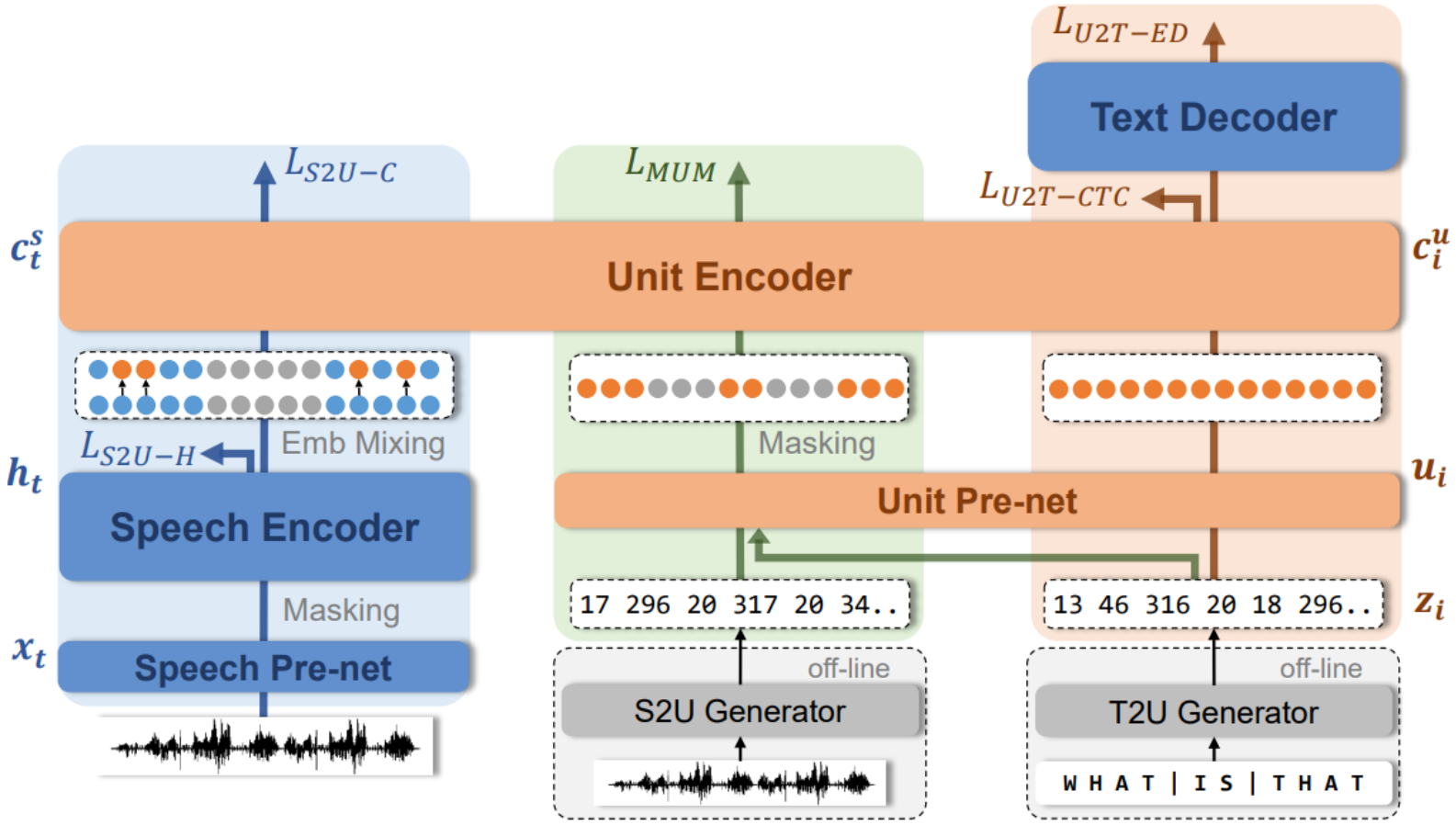
- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

SpeechUT: Motivation

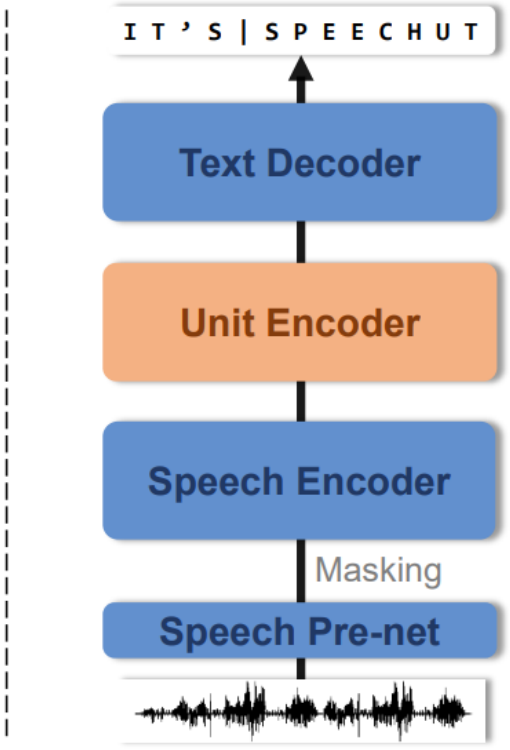
- Design for cross-modal tasks, e.g., ASR and ST
  - Our preliminary observation shows hidden units have strong mapping relationship with both speech and text.
  - This inspires us to decompose the speech-to-text model into a speech-to-unit model and unit-to-text model.



SpeechUT: Model Architecture



(a) Pre-training pipeline of SpeechUT



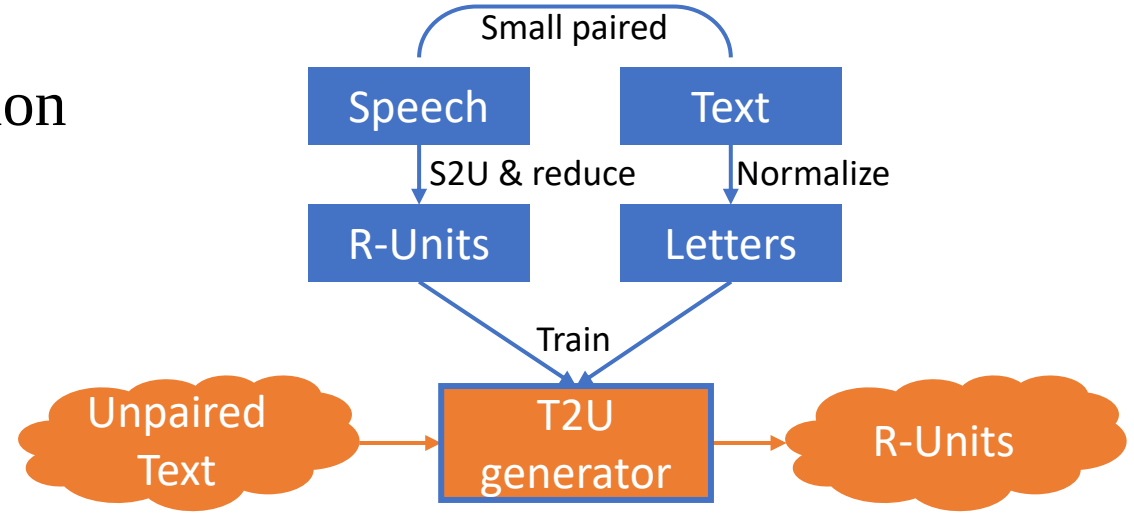
(b) Fine-tuning for ASR/ST

SpeechUT: Pre-Training Tasks

- Speech-to-unit task
  - Use speech encoder and unit encoder
  - Predict the unit categories of the masked positions of a speech sequence based on the non-mask region.
- Unit-to-text task
  - Use unit encoder and text decoder
  - Perform the unit-to-text transformation as a regular sequence-to-sequence task like standard Transformer.
- Masked unit modeling task
  - Use unit encoder
  - Recover the hidden units of masked positions of a unit sequence.

SpeechUT: Data Acquisition

- Speech-unit data (S2U generator)
  - Unsupervised clustering model: e.g., HuBERT
- Unit-text data (T2U generator)
  - A sequence-to-sequence model
  - R(educed)-Units: 17 17 17 17 296 296 20 20 20 34 34,… → 17 296 20 34,…
- Unit data
  - Combination



Training and **inference** of the T2U generator

SpeechUT: Embedding Mixing

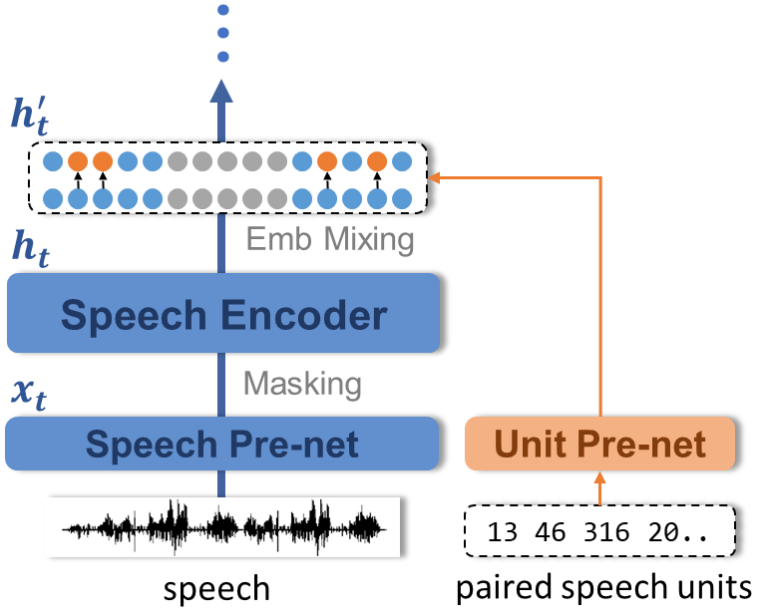
- Embedding mixing mechanism
  - Mix the embeddings of two modalities

$$h_t' = \begin{cases} u_t & t \in \mathcal{R} - \mathcal{M} \\ h_t & \text{otherwise} \end{cases}$$

$\mathcal{R}$ is the selected positions for mixing,

$\mathcal{M}$ is the masked positions,

$\mathcal{R} - \mathcal{M}$ is for preventing information leakage



  - Perform in S2U pre-training task
  - SpeechUT can employ the mechanism on all unlabeled data

SpeechUT: ASR Evaluation

Encoder-based

Encoder-Decoder

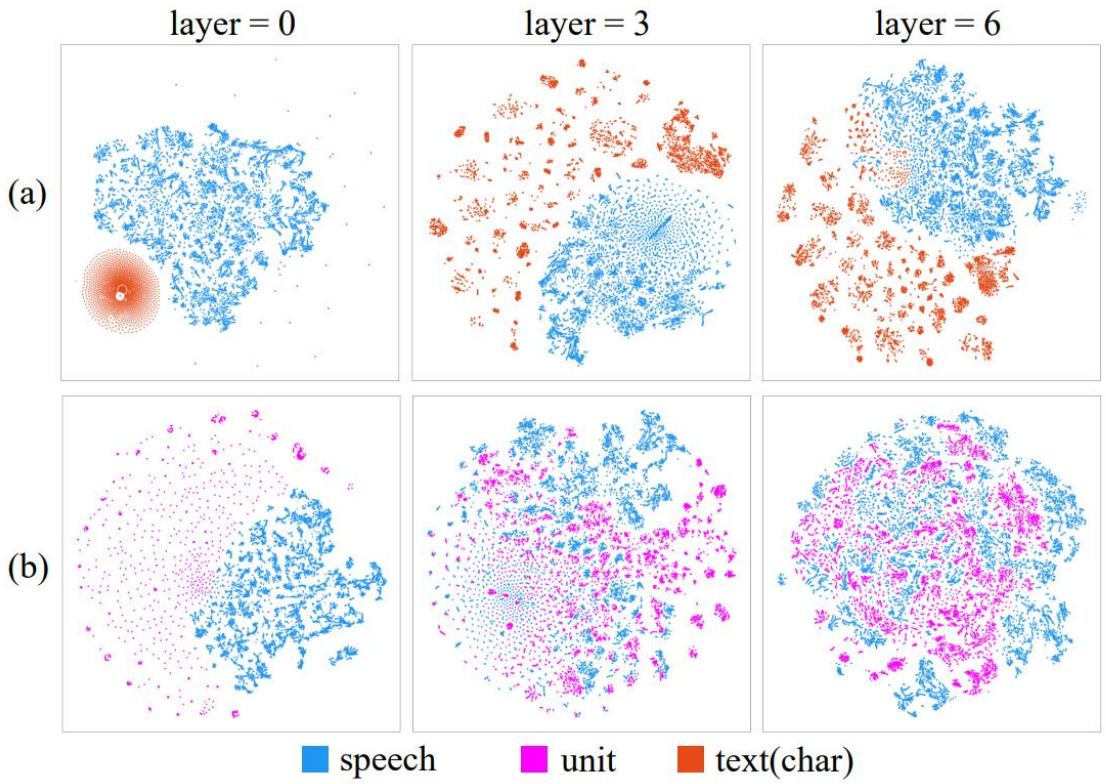
Model	Size	Pre-training Data			WER (↓) Without LM		WER (↓) With LM		
		Speech	Paired	Text	test-clean	test-other	LM	test-clean	test-other
960h hours pre-trained									
wav2vec 2.0 (Baevski et al., 2020)	Base (0.1B)	960h	-	-	6.1	13.3	4-gram	3.4	8.0
HuBERT (Hsu et al., 2021)	Base (0.1B)	960h	-	-	6.3	13.2	4-gram	3.4	8.1
WavLM (Chen et al., 2021)	Base (0.1B)	960h	-	-	5.7	12.0	4-gram	3.4	7.7
ILS-SSL (Wang et al., 2022b)	Base (0.1B)	960h	-	-	4.7	10.1	4-gram	3.0	6.9
data2vec (Baevski et al., 2022)	Base (0.1B)	960h	-	-	4.2*	9.7*	4-gram	2.8	6.8
PBERT (Wang et al., 2022a)	Base (0.15B)	960h	100h†	-	4.7	10.7	4-gram	3.1	7.3
SpeechT5 (Ao et al., 2022a)	Base (0.15B)	960h	-	40M	4.4	10.4	Transf.	2.4	5.8
Speech2C (Ao et al., 2022b)	Base (0.15B)	960h	-	-	4.3	9.0	Transf.	2.4	5.2
Wav2seq (Wu et al., 2022)	Base (0.15B)	960h	-	-	-	11.2	-	-	-
wav2vec 2.0 (Baevski et al., 2020)	**Large** (0.3B)	960h	-	-	4.7	9.0	Transf.	2.3	5.0
Baseline (Ours)	Base (0.15B)	960h	-	40M	3.8	8.0	Transf.	2.3	5.1
SpeechUT (Ours)	Base (0.15B)	960h	100h†	40M	**2.7**	**6.8**	Transf.	**2.0**	**4.5**
60kh hours pre-trained									
wav2vec 2.0 (Baevski et al., 2020)	**Large** (0.3B)	**60kh**	-	-	3.1	6.3	Transf.	2.0	4.0
HuBERT (Hsu et al., 2021)	**Large** (0.3B)	**60kh**	-	-	-	-	Transf.	2.1	3.9
WavLM (Chen et al., 2021)	**Large** (0.3B)	**94kh**	-	-	-	-	Transf.	2.1	4.0
ILS-SSL (Wang et al., 2022b)	**Large** (0.3B)	**60kh**	-	-	2.9	5.8	Transf.	2.0	4.0
STPT (Tang et al., 2022)	Base (0.16B)	**60kh**	100h	40M	3.5	7.2	-	-	-
SpeechUT (Ours)	**Large** (0.38B)	**60kh**	100h†	40M	**2.2**	**4.5**	Transf.	**1.9**	**3.6**

SpeechUT: ST Evaluation

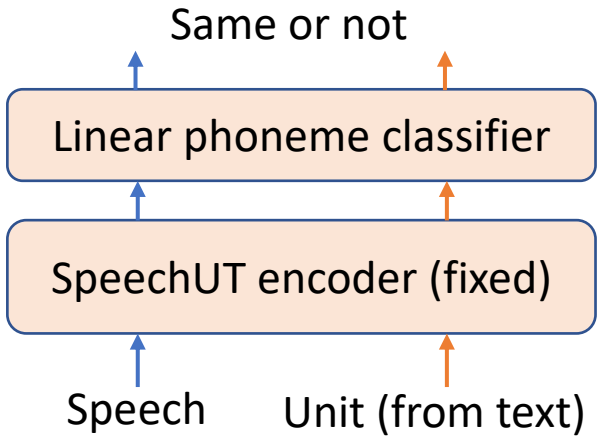
Models	Sizes	Pre-training Data			Fine-tuning BLEU (↑)		
		Speech (h)	ASR (h)	MT (#utt)	En-De	En-Es	En-Fr
FAT-ST (Zheng et al., 2021)	-	3.7k	1.4∼1.5k	1.9∼2.0M	25.5	30.8	-
SATE (Xu et al., 2021a)	-	-	1.4k	18M	28.1	-	-
STEMM (Fang et al., 2022)	-	960	408∼504	4.6∼40M	28.7	31.0	37.4
ConST (Ye et al., 2022)	0.15B	960	408∼504	4.6∼40M	28.3	32.0	38.3
STPT (Tang et al., 2022)	0.16B	60k	408∼504	4.6∼40M	29.2[6]	33.1	39.7
SpeechUT (Ours)	0.16B	1.4∼1.5k	100†	4.6∼40M	**30.1**	**33.6**	**41.4**

- Our SpeechUT achieves the performance of 30.1, 33.6, and 41.4 BLEU scores on En-De, En-Es, and En-Fr, respectively, demonstrating the superiority of SpeechUT over previous work.

Are the Speech and Unit Aligned?



Visualization analysis



Total	Vowels	Consonants	Silence
85.4%	79.6%	85.5%	96.7%

(Proportion where the paired speech and unit representations agree to the same phonemes.)

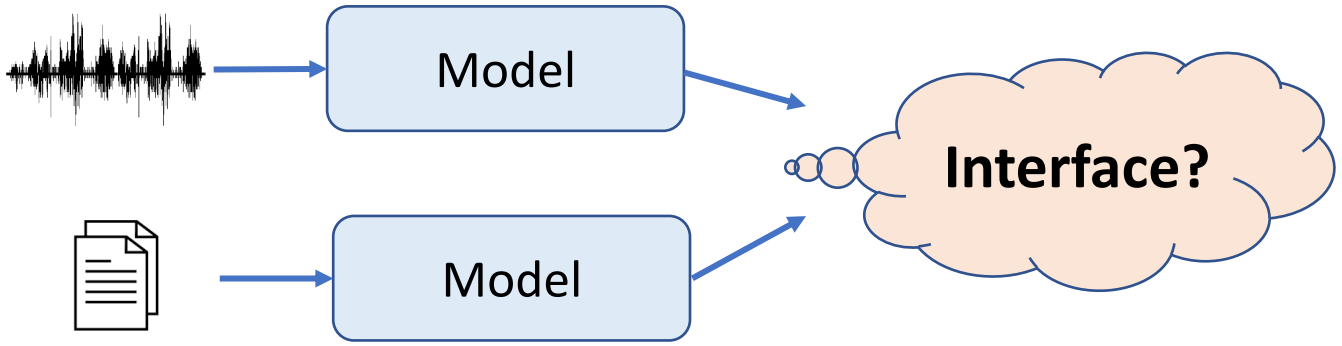
Quantitative analysis

Outline

- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

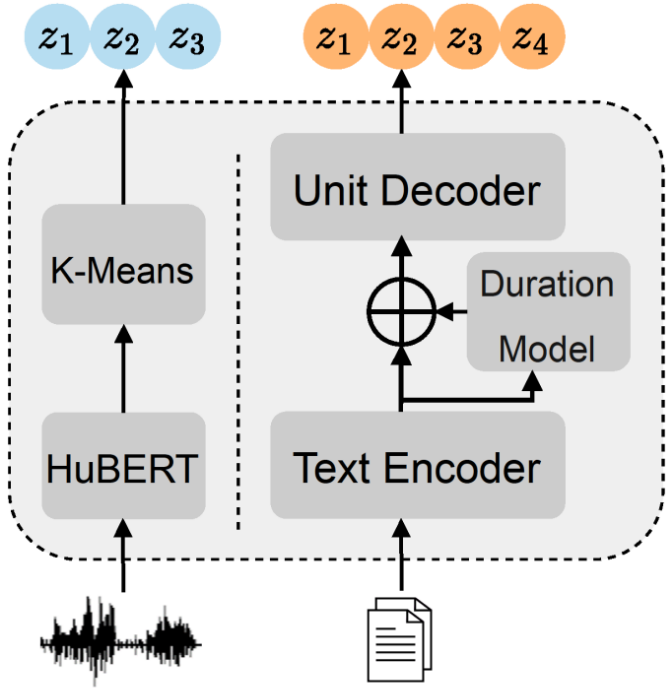
SpeechLM: Motivation

- How to boost speech pre-training with textual data is an unsolved problem.
- Almost all previous work follows the same structure with a speech/text encoder and a shared encoder, however,
  - the interface between the speech encoder and the text encoder is not well studied
  - probably leads to the outputs of the two encoders in different spaces
  - suffers from transfer interference and capacity dilution for the shared encoder



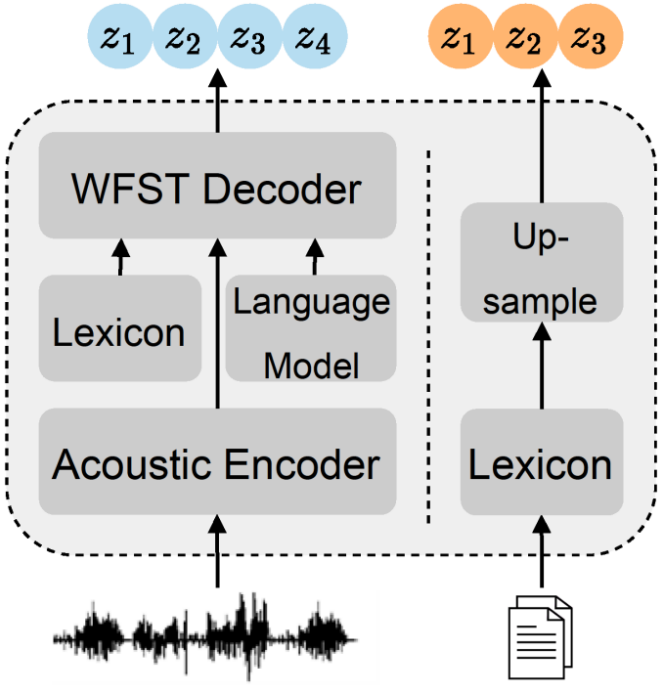
SpeechLM: Tokenizer

- Convert speech/text into the same space



(b) $\mathcal{T}_S^H$ and $\mathcal{T}_T^H$

Hidden-unit tokenizer

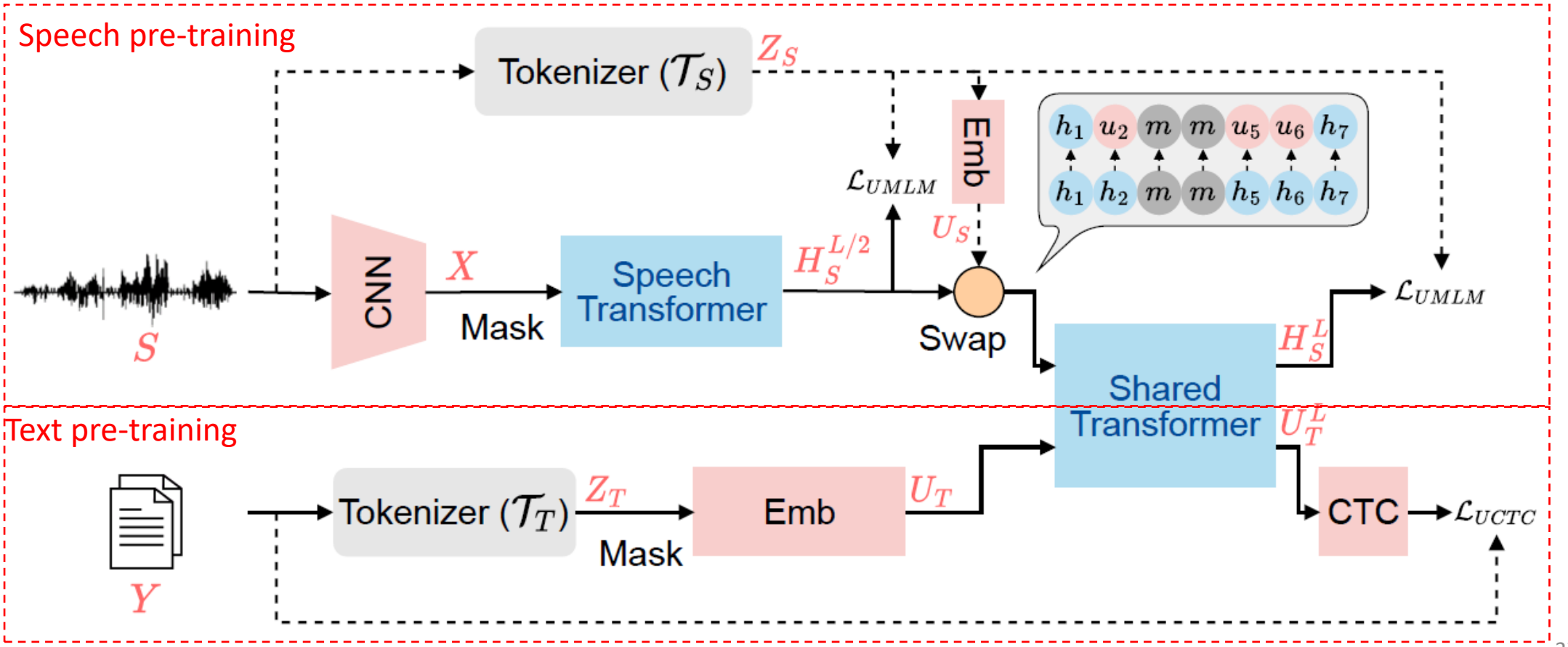


(a) $\mathcal{T}_S^P$ and $\mathcal{T}_T^P$

Phoneme-unit tokenizer

SpeechLM: Framework

- Equip with discrete tokenizers



SpeechLM: Pre-Training Tasks

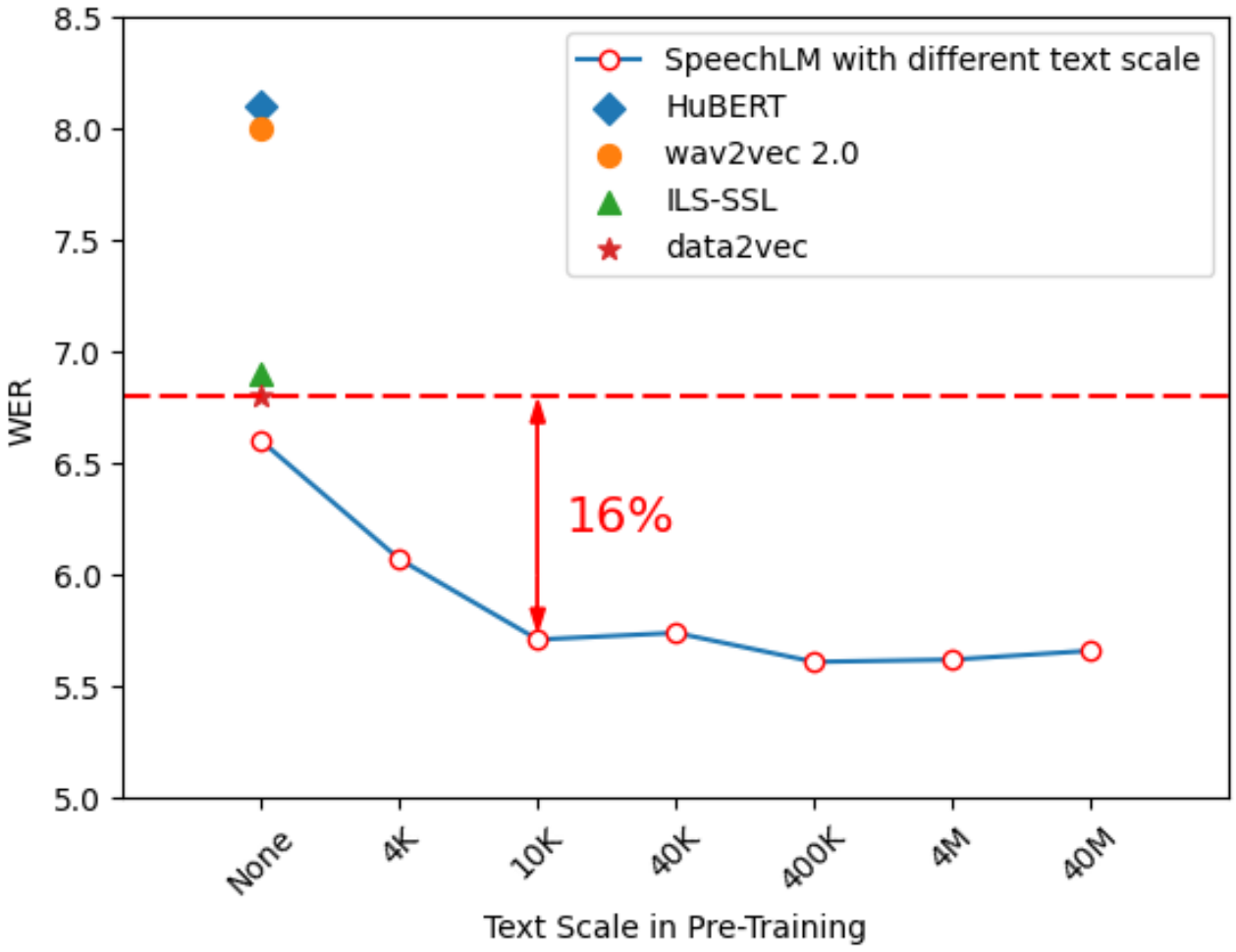
- Unit-based Masked Language Modeling (UMLM)
  - Predict the unit tokens from the masked speech

$$\mathcal{L}_{UMLM} = -\sum_{t \in \mathcal{M}} \left(\log p(z_t | h_t^{L/2}) + \log p(z_t | h_t^{L}) \right)$$

- Unit-based Connectionist Temporal Classification (UCTC)
  - Reconstruct the whole text sequences from the masked unit sequences.

$$\mathcal{L}_{UCTC} = -\log p_{CTC}(\boldsymbol{Y} | \boldsymbol{U}^{L})$$

SpeechLM: ASR Evaluation



SpeechLM: ST Evaluation

Pre-trained Model	Size (Encoder)	en-de	en-ca	en-ar	en-tr	avg
Pre-ASR (Wang et al., 2020)	-	16.3	21.8	12.1	10.0	15.1
HuBERT (Hsu et al., 2021) *	Base (0.1B)	21.6	28.4	15.9	14.4	20.1
SpeechLM-H	Base (0.1B)	23.8	30.9	17.9	16.1	22.2
SpeechLM-P	Base (0.1B)	**24.2**	**31.2**	**18.3**	**16.2**	**22.5**
wav2vec 2.0 (Wang et al., 2021a)	Large (0.3B)	23.8	32.4	17.4	15.4	22.3
SLAM (Bapna et al., 2021)	X-Large (0.6B)	27.2	33.3	18.5	16.8	24.0
SLAM→w2v-bert (Bapna et al., 2021)	X-Large (0.6B)	27.1	34.2	21.2	17.5	25.0
SpeechLM-P	Large (0.3B)	**27.6**	**35.9**	**21.7**	**19.5**	**26.2**

- SpeechLM-H and SpeechLM-P achieve comparable results in the Base setting, and 2.4 BLEU improvement over HuBERT Base.
- Moreover, the proposed SpeechLM Large model significantly outperforms previous work.

SpeechLM: SUPERB Evaluation

Method	#Params	Corpus	Speaker			Content					Semantics			ParaL
			SID	ASV	SD	PR	ASR	OOD-ASR	KS	QbE	ST	IC	SF		ER
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	WER ↓	Acc ↑	MTWV ↑	BLEU ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	63.58	8.63	0.0058	2.32	9.10	69.64	52.94	35.39
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	61.56	82.54	0.0072	3.16	29.82	62.14	60.17	57.86
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	63.12	91.01	0.0310	5.95	74.69	70.46	50.89	59.33
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	63.56	91.11	0.0251	4.23	74.48	68.53	52.91	59.66
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	61.66	88.96	0.0246	4.32	69.44	72.79	48.44	59.08
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	65.27	83.67	6.6E-04	4.45	34.33	61.59	58.89	50.28
TERA (Liu et al., 2020b)	21.33M	LS 960 hr	57.57	15.89	9.96	49.17	18.17	58.49	89.48	0.0013	5.66	58.42	67.50	54.17	56.27
DeCoAR 2.0 (Ling & Liu, 2020)	89.84M	LS 960 hr	74.42	7.16	6.59	14.93	13.02	53.62	94.48	0.0406	9.94	90.80	83.28	34.73	62.47
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	62.54	91.88	0.0326	4.82	64.09	71.19	49.91	60.96
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	55.86	95.59	0.0485	6.61	84.92	76.37	43.71	59.79
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	60.66	93.38	0.0410	5.66	85.68	77.68	41.54	58.24
Wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	46.95	96.23	0.0233	14.81	92.35	88.30	24.77	63.43
HuBERT Base (Hsu et al., 2021)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	46.69	96.30	0.0736	15.53	98.34	88.53	25.20	64.92
WavLM Base (Chen et al., 2022a)	94.70M	LS 960 hr	84.51	4.69	**4.55**	4.84	6.21	**42.81**	**96.79**	**0.0870**	20.74	**98.63**	89.38	22.86	**65.94**
SpeechLM-H Base	94.70M	LS 960 hr	75.12	6.76	6.48	4.2	5.56	45.78	96.04	0.0526	20.72	97.6	88.76	23.49	63.31
SpeechLM-P Base	94.70M	LS 960 hr	75.11	6.13	6.56	**2.43**	**4.22**	47.22	94.61	0.0458	**22.79**	98.6	**89.4**	**22.36**	62.09

- Compared to the previous self-supervised learning methods, SpeechLM achieves better performance on several content-related and semantic-related tasks, such as PR, ASR, ST, and SF.

SpeechLM: Visualization Analysis

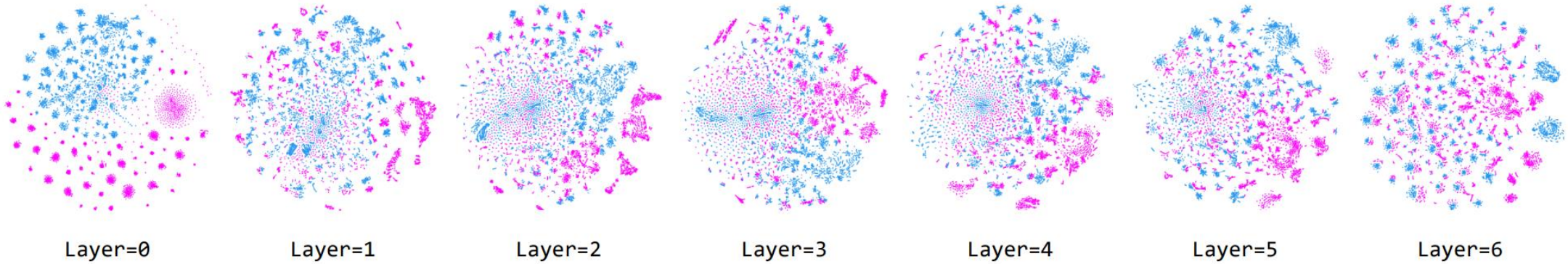


Figure: Layer-wise visualization of the Shared Transformer in SpeechLM-P Base. Frame-wise representations of unpaired speech (blue) and phonemes (red) are present.

Outline

- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

VATLM: Motivation

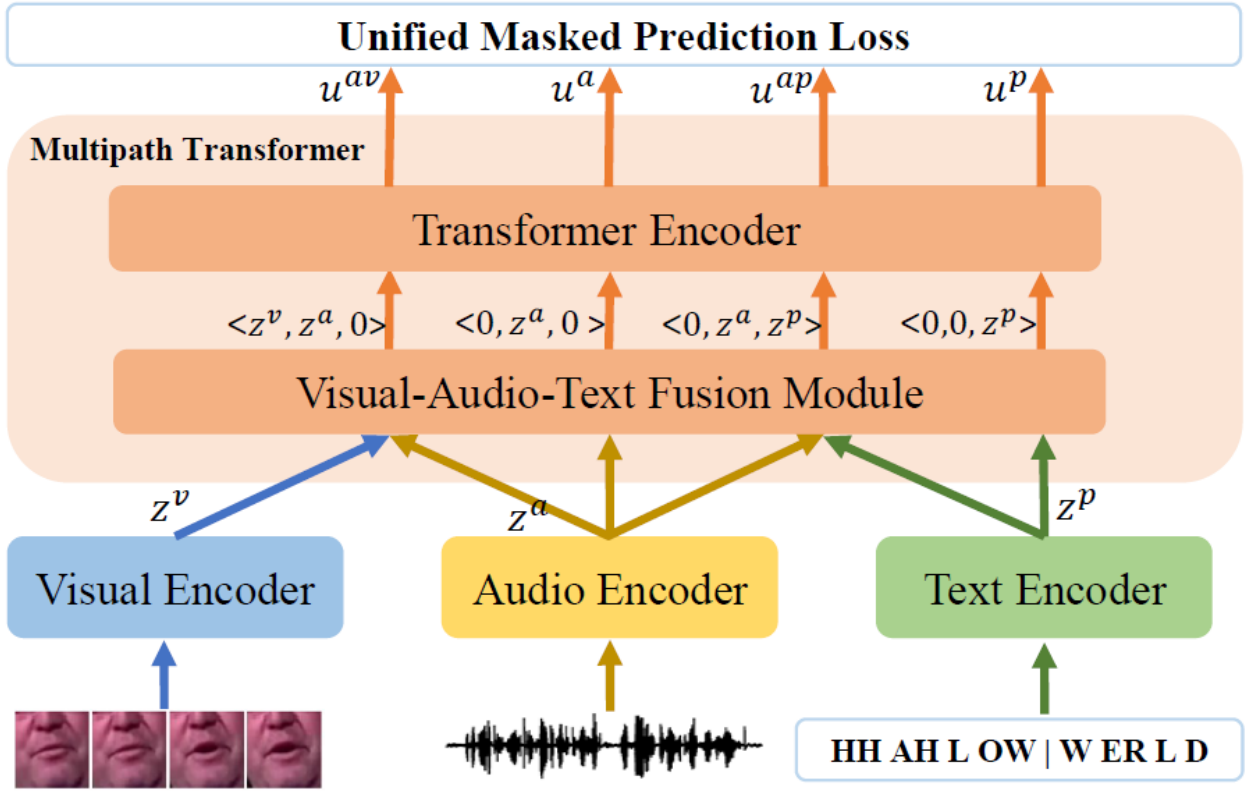
- Previous multi-modal (visual, speech, text) pre-training approaches mainly focus on visual-language tasks and cannot be extended to other spoken language processing tasks, such as AVSR.
- Previous speech representation learning methods can not make full use of diverse corpora, e.g., visual-audio pairs, audio-text pairs, and unlabeled speech and text, without considering both the visual and textual information.
- Previous methods mostly depend on a complicated model architecture and pre-training objects, **lacking a unified multi-modal framework for different modalities modeled in the same semantic space.**

AV-HuBERT →

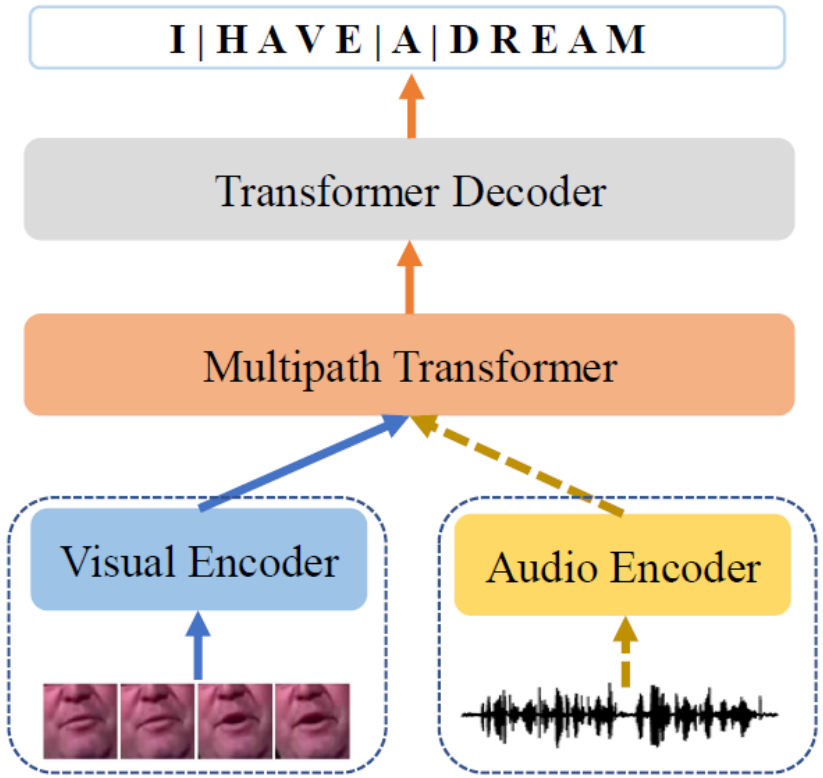
(Shi et al., 2022)

VATLM: Framework

- One shared model, one shared pre-training loss



(a) Pre-training structure of VATLM



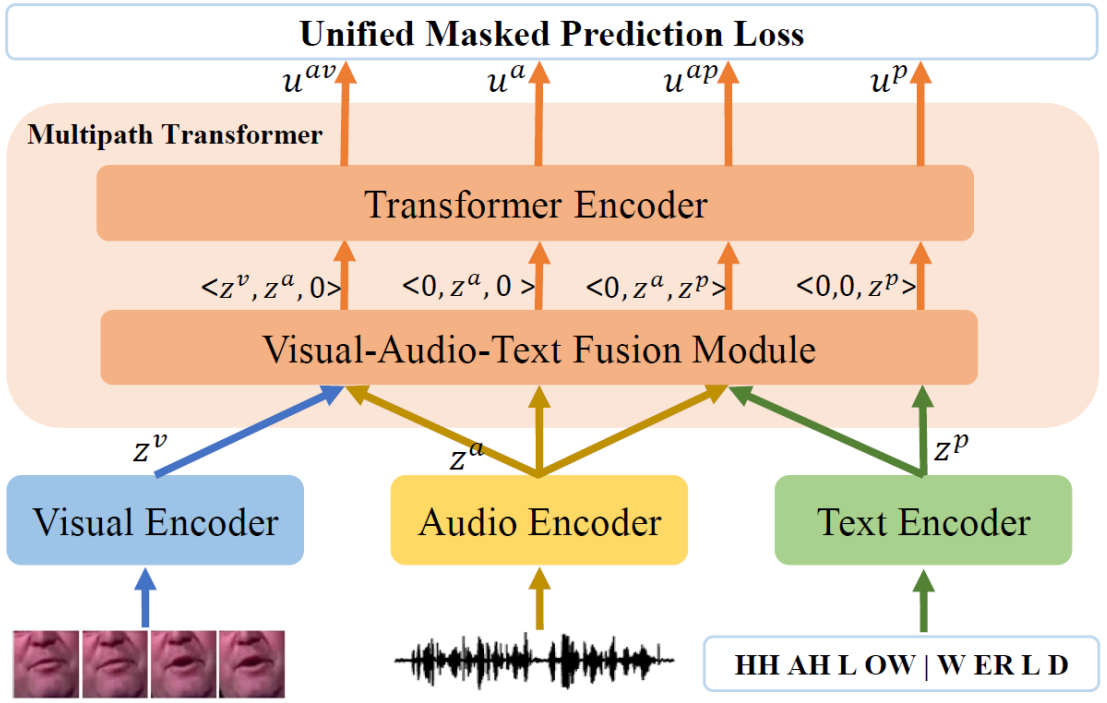
(a) Fine-tuning for AVSR/VSR

VATLM: Multipath Transformer

- Preprocessing modules for visual, audio, and text
  - Visual encoder (ResNet)
  - Audio encoder (linear layer)
  - Text encoder (embedding layer)
- Visual-audio-text fusion module

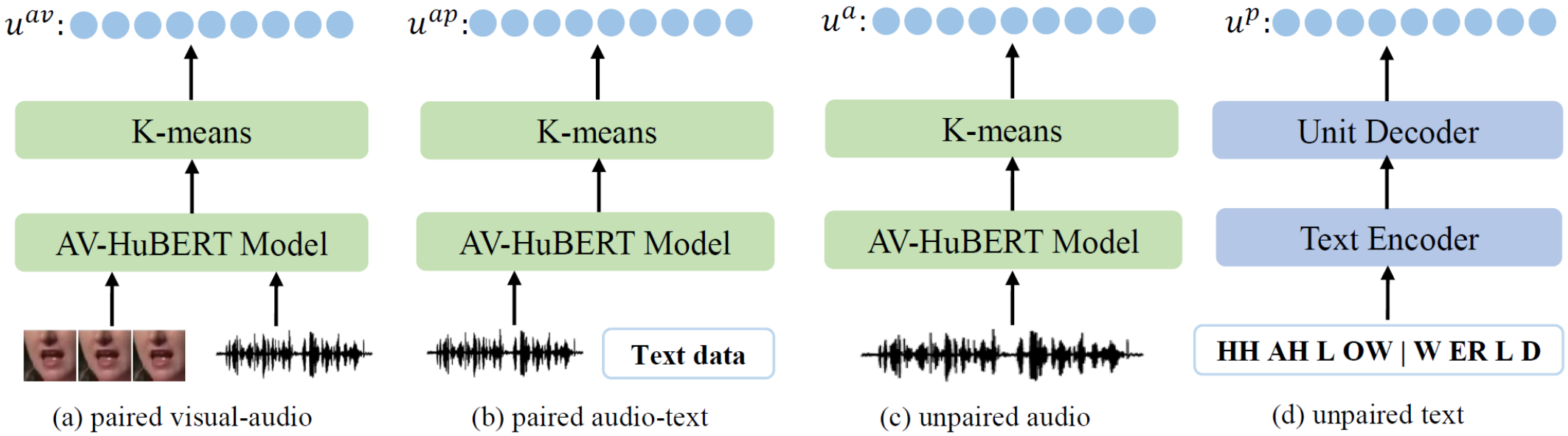
$$z^f = \text{concat}(z^v, z^a, z^p)$$

- Transformer encoder



VATLM: Unified Tokenizer

- The unified tokenizer generates the shared hidden units from different modalities and data resources.



VATLM: Masked Prediction Loss

- Existing multi-modal pre-training work usually employs multiple pre-training objectives, such as masked language modeling, contrastive learning, speech-text matching, image-text matching, and so on.
- **We pre-train VATLM via a unified masked prediction objective** on both mono-modal (i.e., audio and texts) and multi-modal data (i.e., audio-visual pairs and audio-text pairs).

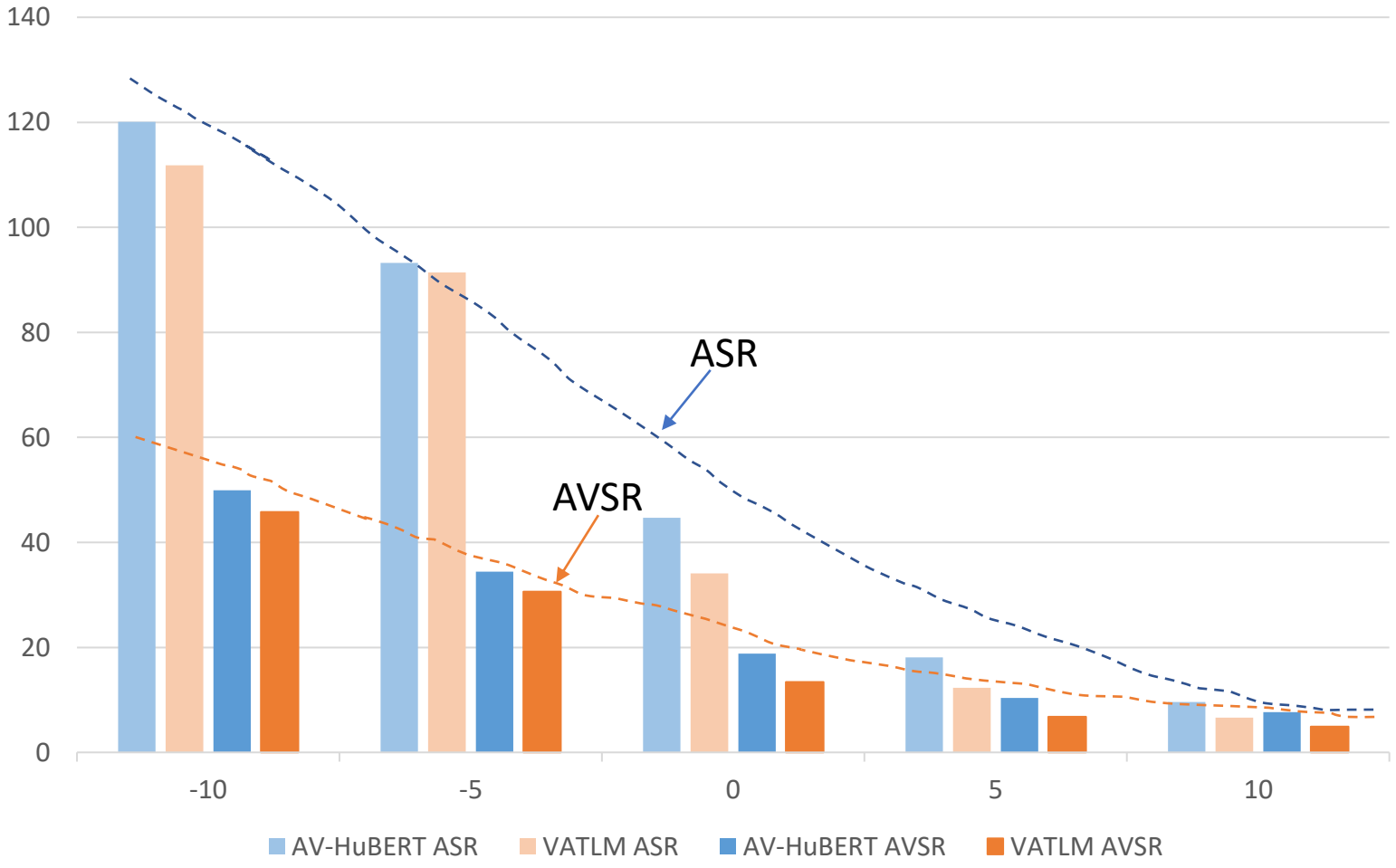
$$\mathcal{L} = -\sum_{t \in \mathcal{M}} \left(\log p(u_t | h_t^f) \right),$$

$$\mathcal{L}_{total} = \mathcal{L}^{av} + \lambda_1 \mathcal{L}^{a} + \lambda_2 \mathcal{L}^{ap} + \lambda_3 \mathcal{L}^{p},$$

VATLM: AVSR/VSR Evaluation

Method	Backbone	Criterion	Extra labeled (hrs)	Fine-tuned data (hrs)	Pre-trained AV data (hrs)	AVSR WER (%)	VSR WER (%)
Supervised							
Zhang et al. [59]	CNN	CE	157	698	-	-	60.1
Afouras et al. [24]	Transformer	CE	157	1362	-	7.2	58.9
Xu et al. [18]	RNN	CE	157	433	-	7.2	57.8
Shillingford et al. [60]	RNN	CTC	-	3886	-	-	55.1
Ma et al. [25]	Conformer	CTC+CE	-	433	-	-	46.9
Ma et al. [25]	Conformer	CTC+CE	157	433	-	2.3	43.3
Makino et al. [19]	RNN	Transducer	-	31000	-	-	33.6
Self-supervised & Semi-supervised							
Afouras et al. [41]	CNN	CTC	157	433	334		59.8
Zhang et al. [33]	Transformer-Base	CTC	-	30	433	9.1	67.8
Ma et al. [42]	Transformer-Base	CE	-	30	433	-	71.9
			-	433	1759	-	49.6
AV-HuBERT [30], [34]	Transformer-Base	CE	-	30	433	-	51.8
			-	30	1759	4.0	46.1
			-	433	1759	-	34.8
	Transformer-Large	CE	-	30	1759	3.3	32.5
			-	433	1759	1.4	28.6
AV-HuBERT (w/ self-training) [30]	Transformer-Large	CE	-	30	1759	-	28.6
			-	433	1759	-	26.9
VATLM (ours)	Transformer-Base	CE	-	30	433	**3.6**	**48.0**
			-	30	1759	**3.4**	**42.6**
			-	433	1759	**1.7**	**34.2**
	Transformer-Large	CE	-	30	1759	**2.7**	**31.6**
			-	433	1759	**1.2**	**28.4**
VATLM (w/ self-training)	Transformer-Large	CE	-	30	1759	**2.7**	**27.6**
			-	433	1759	**1.2**	**26.2**

VATLM: Robustness Evaluation



VATLM: Visualization Analysis

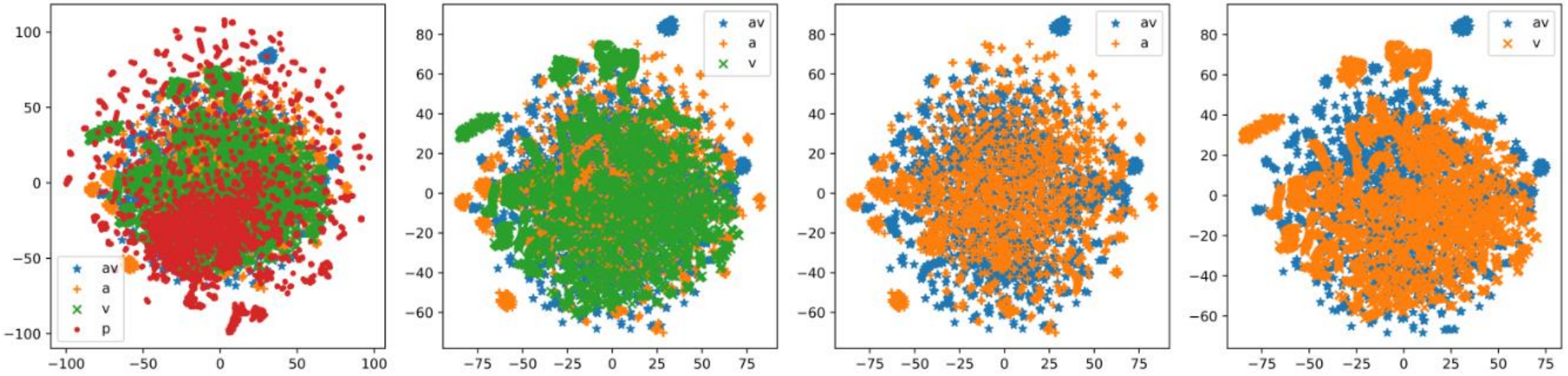


Figure: The 2D t-SNE visualization of representations obtained from data with different modalities, where ‘av’ in the figure denotes data of audio-visual modality, ‘a’ denotes data of audio modality, ‘v’ denotes data of visual modality, and ‘p’ denotes data of text modality.

Outline

- Background
  - Text/speech pre-training
  - The big convergence
- Our Work
  - SpeechT5: Unified-modal encoder-decoder pre-training for speech tasks
  - SpeechUT: Bridge speech and text with hidden unit for enc-dec pre-training
  - SpeechLM: Enhanced speech pre-training with unpaired textual data
  - VATLM: Visual-audio-text pre-training for speech representation learning
- Summary

Summary

- Comparison among different models

	SpeechT5	SpeechUT	SpeechLM	VATLM
Model architecture	Encoder-decoder	Encoder-decoder	Encoder	Encoder
Data sources	Speech/text	Speech/text	Speech/text	Speech/text/ visual
Need small paired data	No	Yes	Yes	Yes
Pre-training tasks	MLM/L1/MLE	MLM/CE	MLM/CTC	MLM
Fine-tuning tasks	Universal	ASR/ST	Universal	VSR/AVSR

Summary

- The big convergence across modalities (e.g., speech, text, and, image, and video) is emerging in recent years.

- We explored the unified-modal self-supervised representation learning, mainly for speech processing tasks, and proposed SpeechT5, SpeechUT, SpeechLM, and VATLM models.

- All code and models are available at:

https://github.com/microsoft/SpeechT5

https://github.com/microsoft/unilm

Challenges

- Joint speech and text modeling
  - Remove the need of small paired data
  - Deeply integrate the language model ability
  - Extend to natural language tasks
- Speech/audio generation
  - Text controllable speech/audio generation
  - Cross-lingual speech synthesis
  - Speech to speech translation with voice reservation
- Big convergence across modalities
  - Large-scale model training with more data
  - Integrate diverse video and audio data
  - Build universal large language model

Thank You!

Thank all co-authors, Furu Wei, Shujie Liu, Yu Wu, Shuo Ren, Junyi Ao, Rui Wang, Chengyi Wang, Ziqiang Zhang, Sanyuan Chen, Qiushi Zhu, et al.

Long Zhou (周龙)

Senior Researcher @ Microsoft Research Asia

Email: lozhou@microsoft.com

Homepage: long-zhou.github.io

Reference

[1] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., ... & Zhu, J. (2021). Pre-trained models: Past, present and future. AI Open, 2, 225-250.

[6] Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., ... & Watanabe, S. (2022). Self-Supervised Speech Representation Learning: A Review. arXiv preprint arXiv:2205.10643.

[7] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449-12460.

[8] Hsu, W. N., Tsai, Y. H. H., Bolte, B., Salakhutdinov, R., & Mohamed, A. (2021, June). HuBERT: How much can a bad teacher benefit ASR pre-training?. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6533-6537). IEEE.

Reference

[9] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505-1518.

[10] Hao, Y., Song, H., Dong, L., Huang, S., Chi, Z., Wang, W., ... & Wei, F. (2022). Language models are general-purpose interfaces. arXiv preprint arXiv:2206.06336.

[11] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., ... & Wei, F. (2022). Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442.

[12] Baevski, A., Hsu, W. N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555.

[13] Bapna, A., Chung, Y. A., Wu, N., Gulati, A., Jia, Y., Clark, J. H., ... & Zhang, Y. (2021). SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training. arXiv preprint arXiv:2110.10329.

[14] Bapna, A., Cherry, C., Zhang, Y., Jia, Y., Johnson, M., Cheng, Y., ... & Conneau, A. (2022). mSLAM: Massively multilingual joint pre-training for speech and text. arXiv preprint arXiv:2202.01374.

[15] Chen, Z., Zhang, Y., Rosenberg, A., Ramabhadran, B., Moreno, P., Bapna, A., & Zen, H. (2022). MAESTRO: Matched Speech Text Representations through Modality Matching. arXiv preprint arXiv:2204.03409.

[16] Shi, B., Hsu, W. N., Lakhotia, K., & Mohamed, A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184.