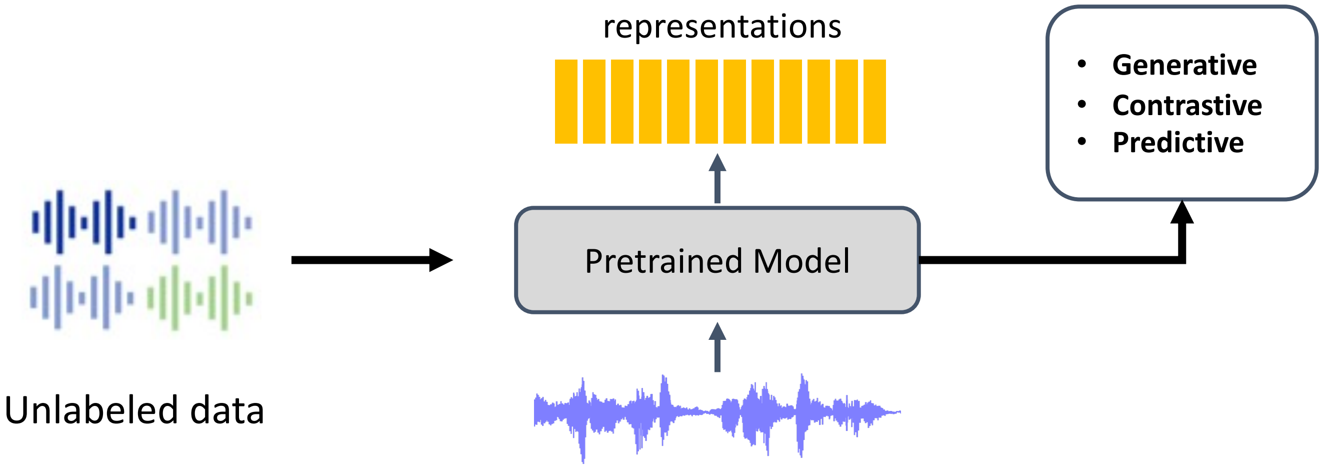
Pre-Training for Speech Processing

Chengyi Wang (王程一)

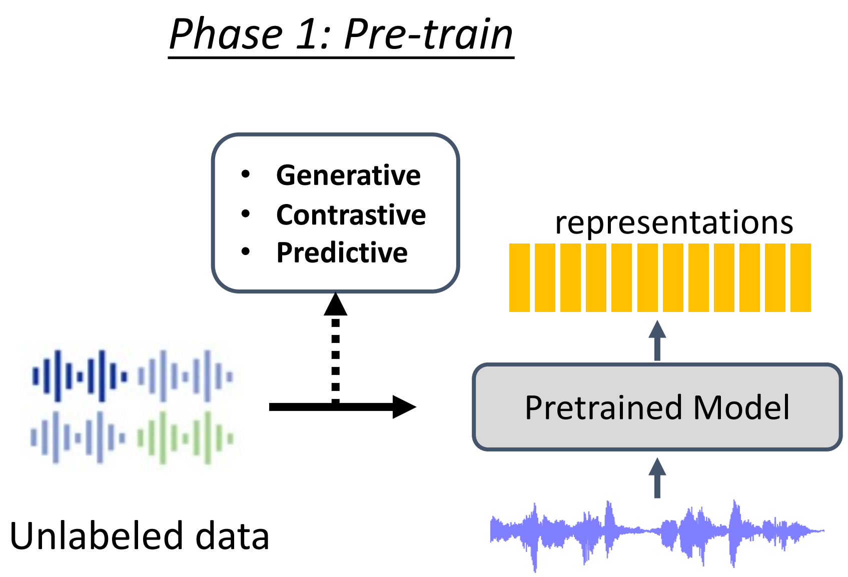
Joint Ph.D. student of Nankai University and Microsoft Research Asia

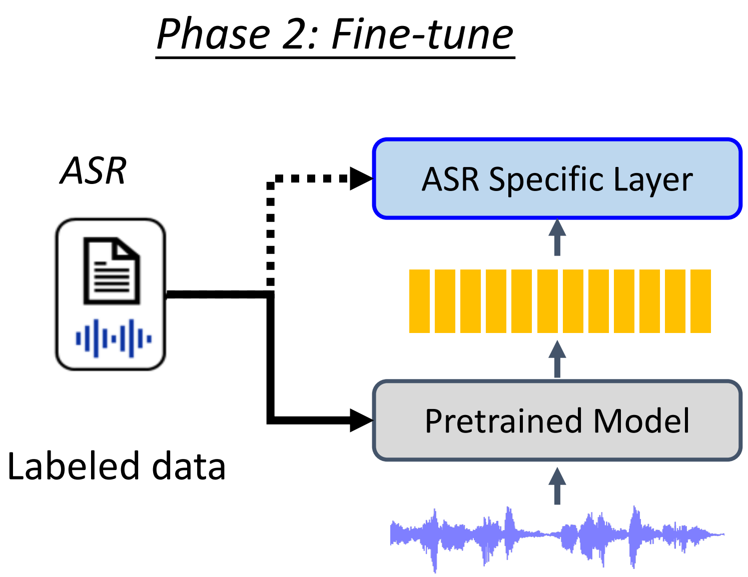
Three Key Elements

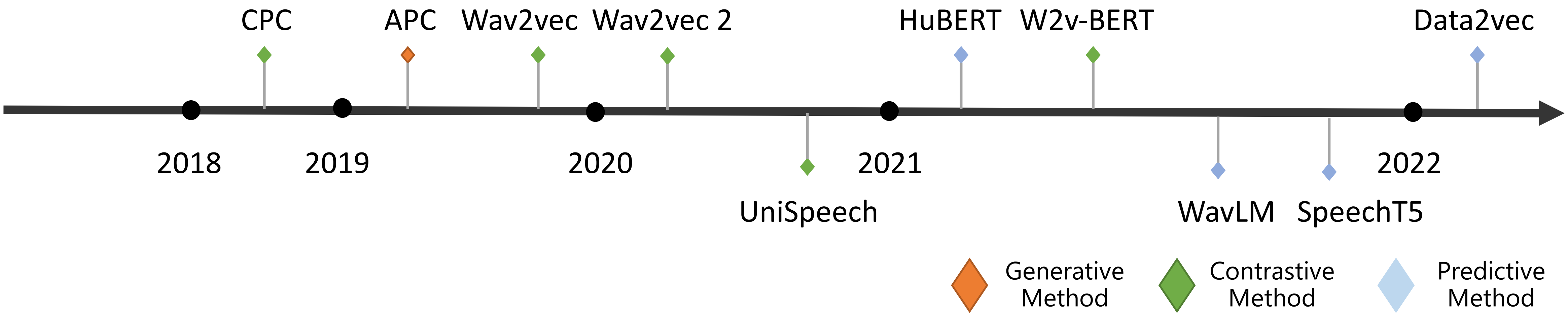
Data + Model + Task = Pre-training



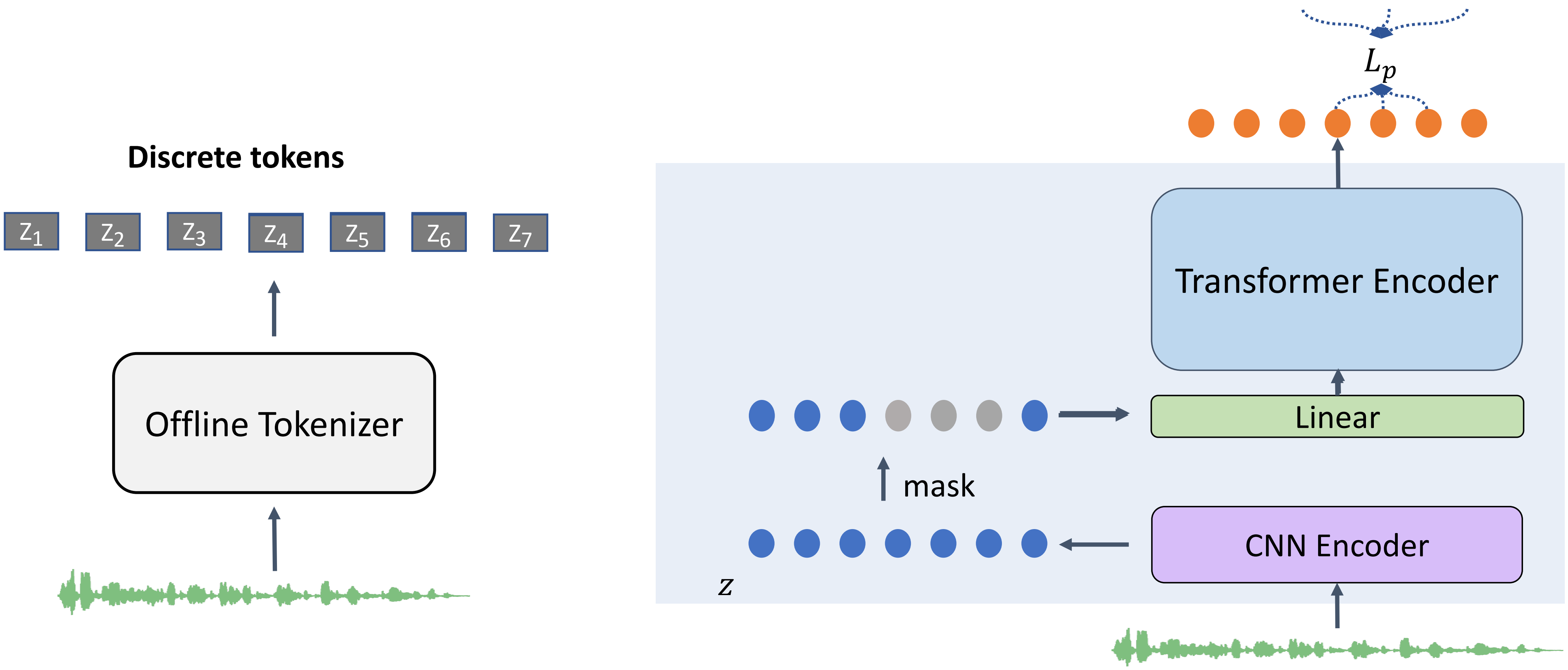
Background


Phase 1: Pre-train
Generative
Contrastive
Predictive
representations
Pretrained Model
Unlabeled data

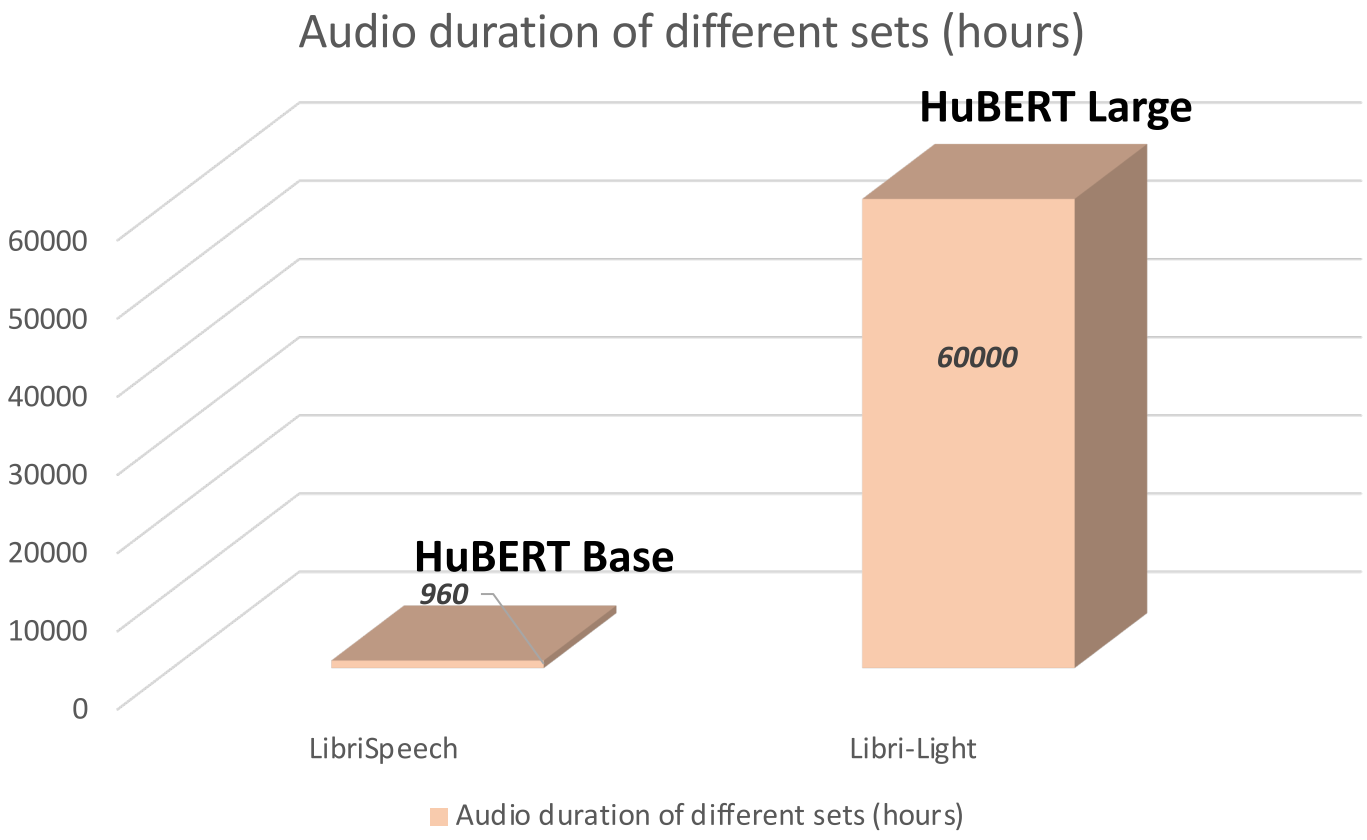

Phase 2: Fine-tune
ASR
ASR Specific Layer
Pretrained Model
Labeled data

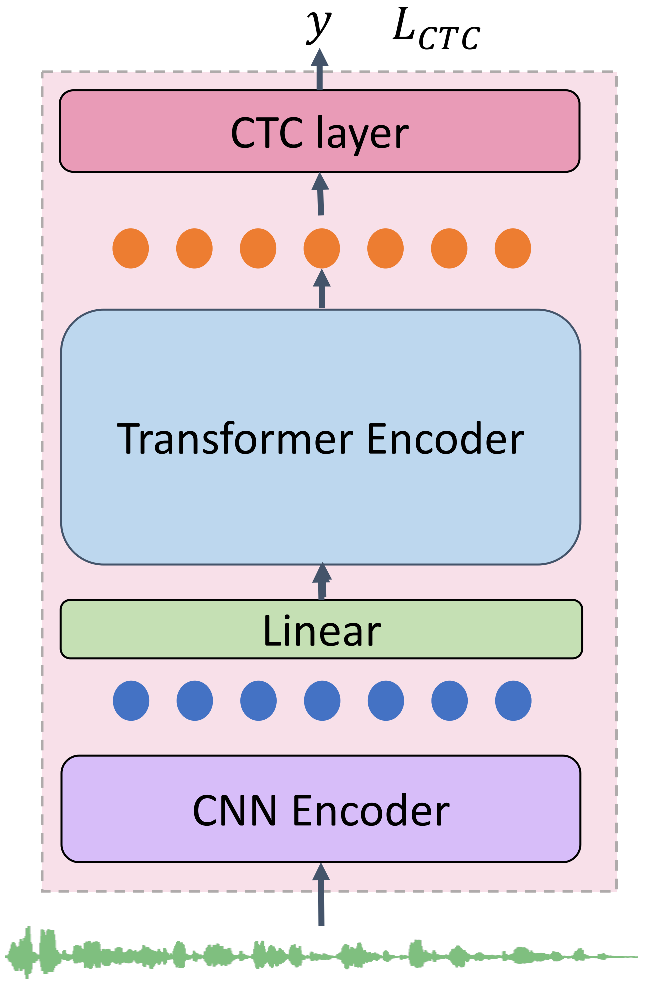

CPC
APC
Wav2vec
Wav2vec 2
HuBERT
W2v-BERT
Data2vec
2018
2019
2020
2021
2022
UniSpeech
WavLM
SpeechT5
Generative Method
Contrastive Method
Predictive Method

Background: HuBERT



HuBERT Setup



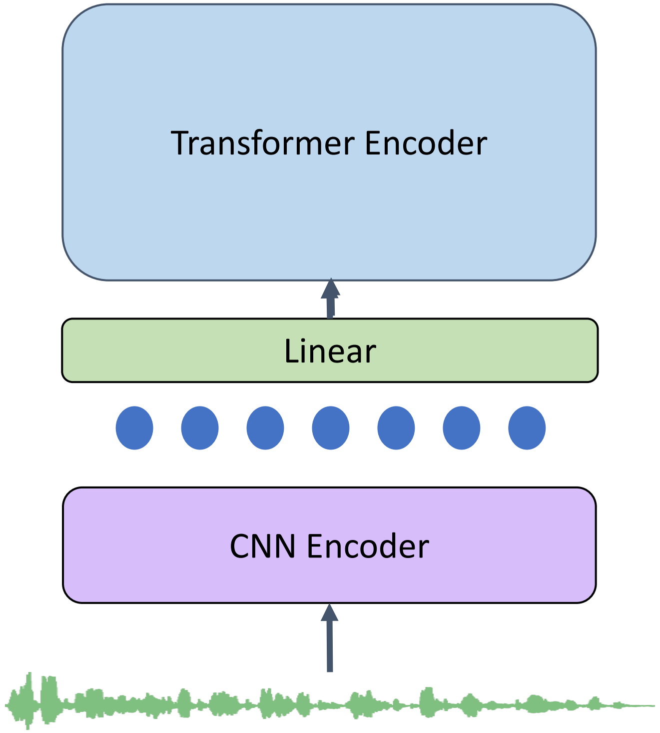


Three Key Elements

Data + Model + Task = Pre-training

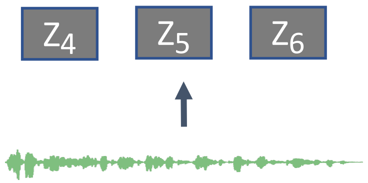
Librilight

+



Speech Transformer

+

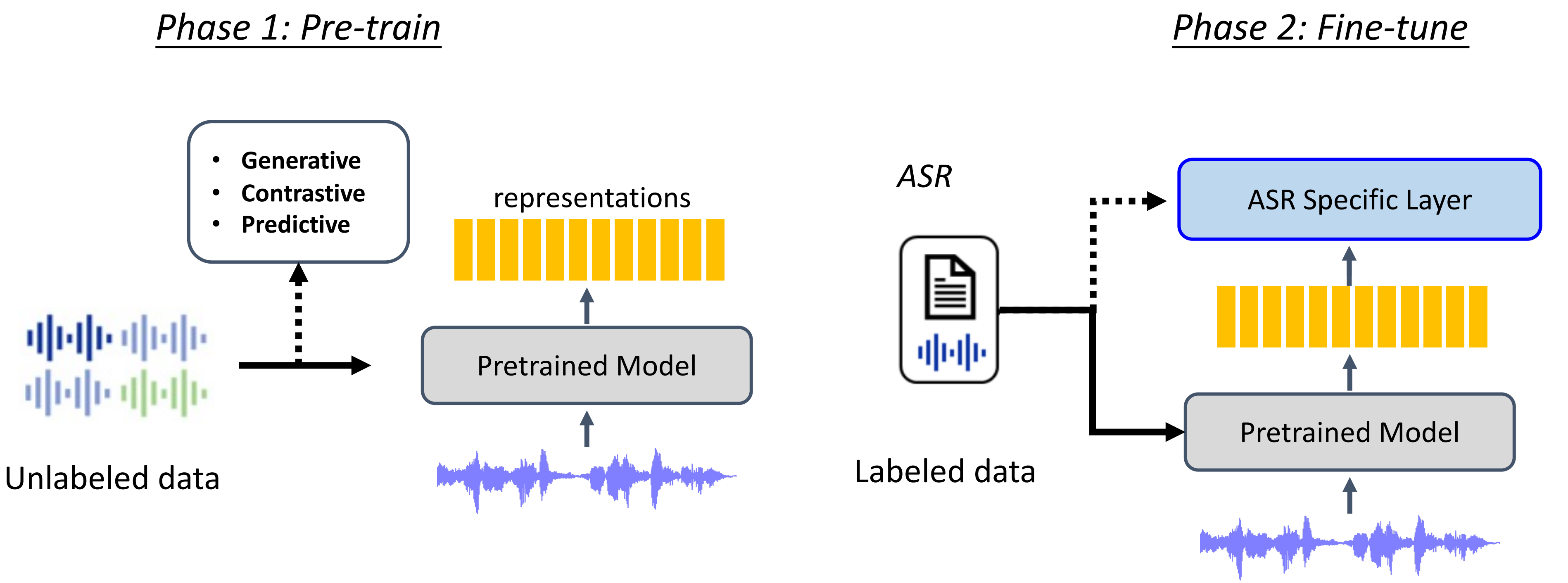


Mask Prediction

= HuBERT

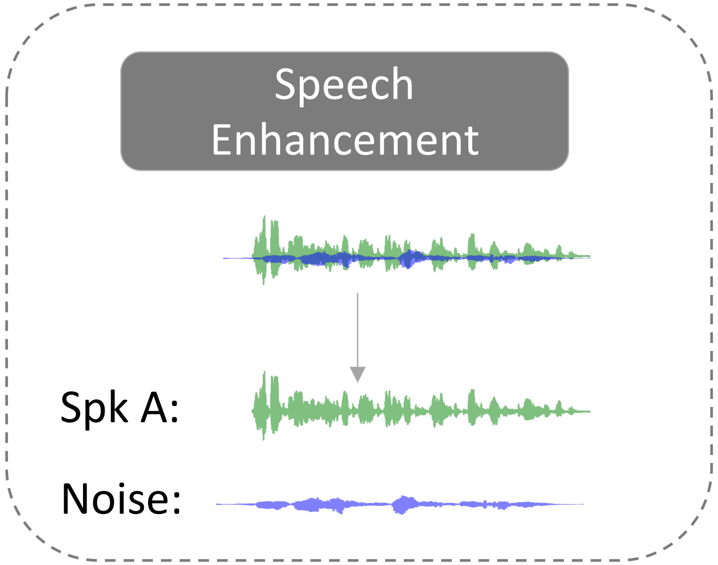
Key Question

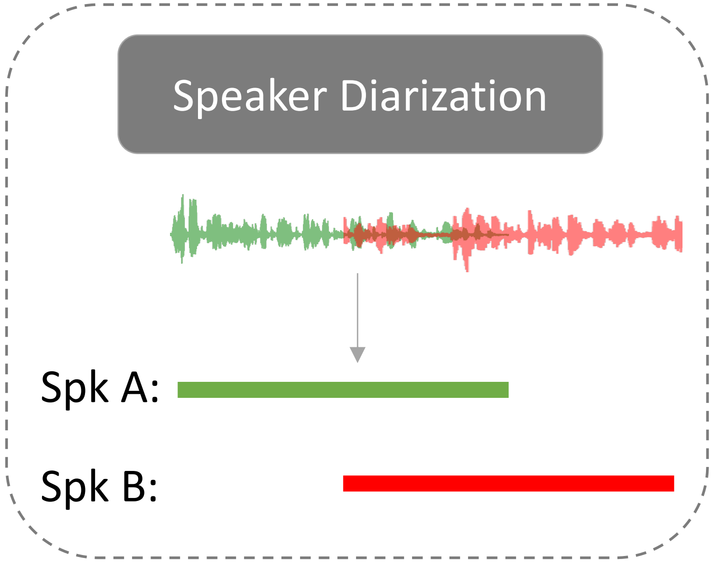
Can a single universal model benefit various speech tasks?



Key Question

Can a single universal model benefit various speech tasks?

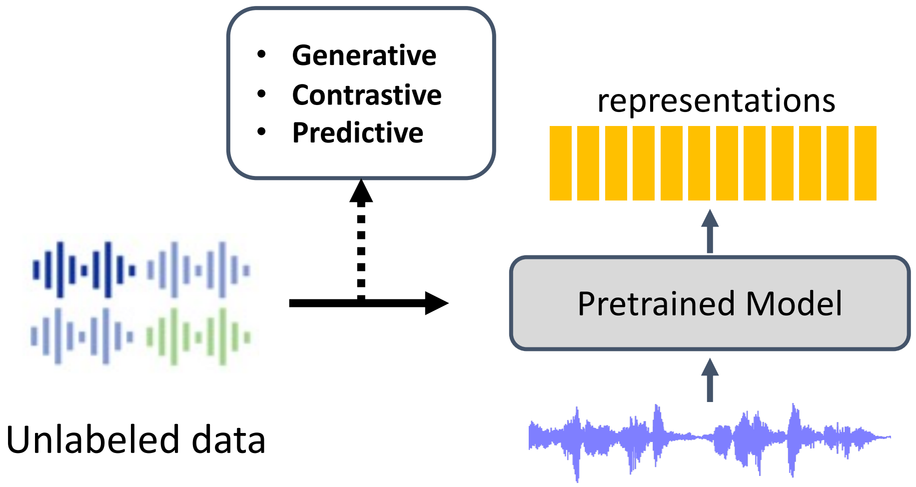




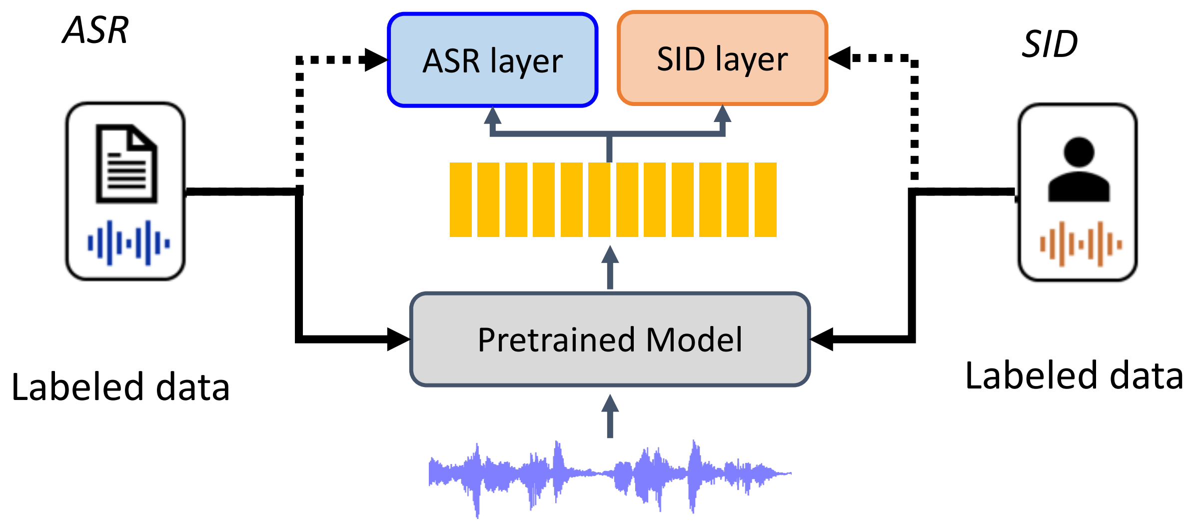
WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing

Translate Success from ASR to Full Stack Speech Processing Tasks

Phase 1: Pre-train



Phase 2: Fine-tune



Our Solution

WavLM

One model for full-stack downstream tasks

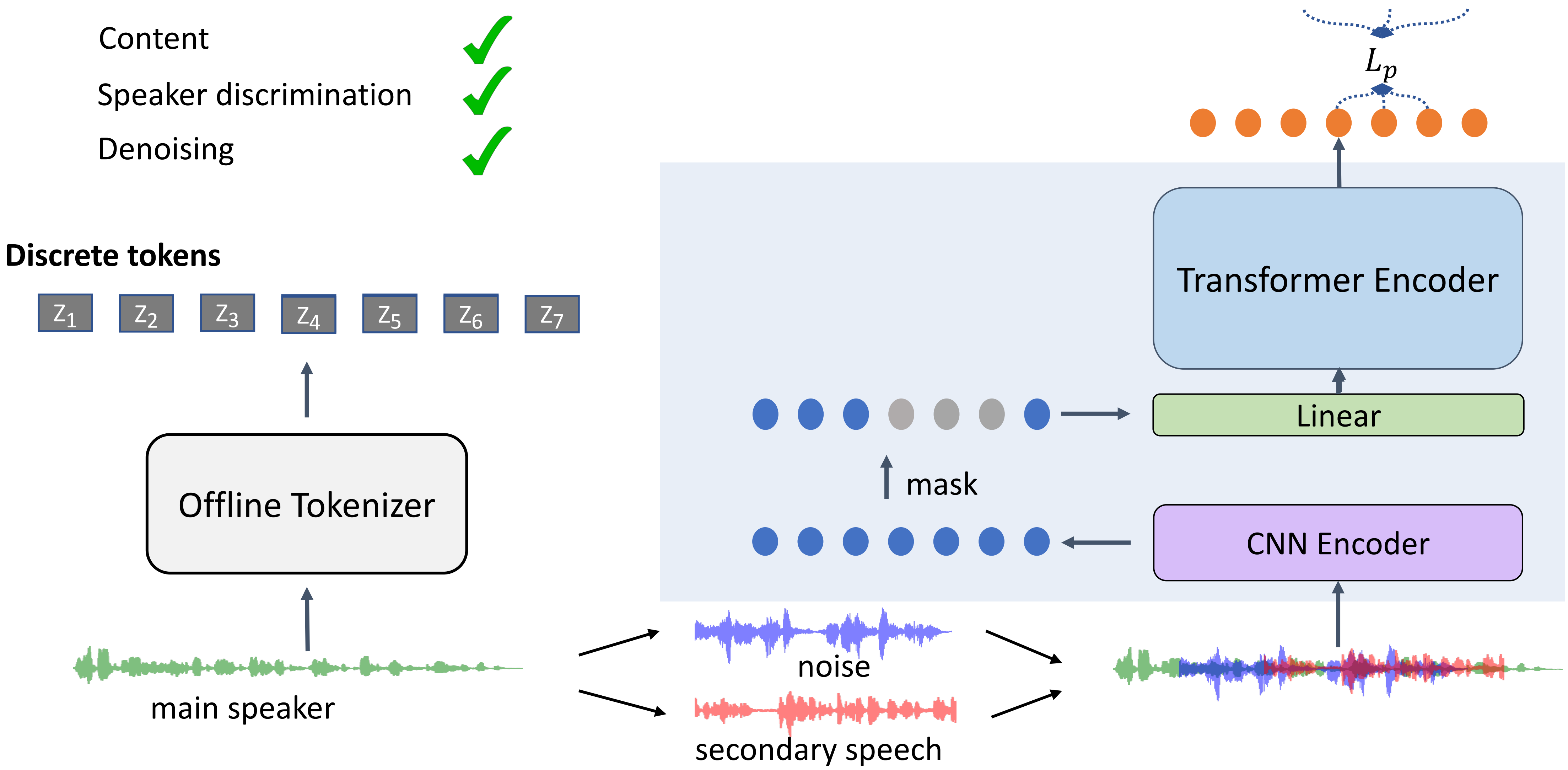
- Content Modeling
  - Speech Recognition, Speech Translation
- Denoising Modeling
  - Speech Enhancement, Speech Separation
- Speaker Modeling
  - Speaker Diarization

WavLM: Masked Speech **Prediction** and **Denoising**

Content ✓

Speaker discrimination ✓

Denoising ✓

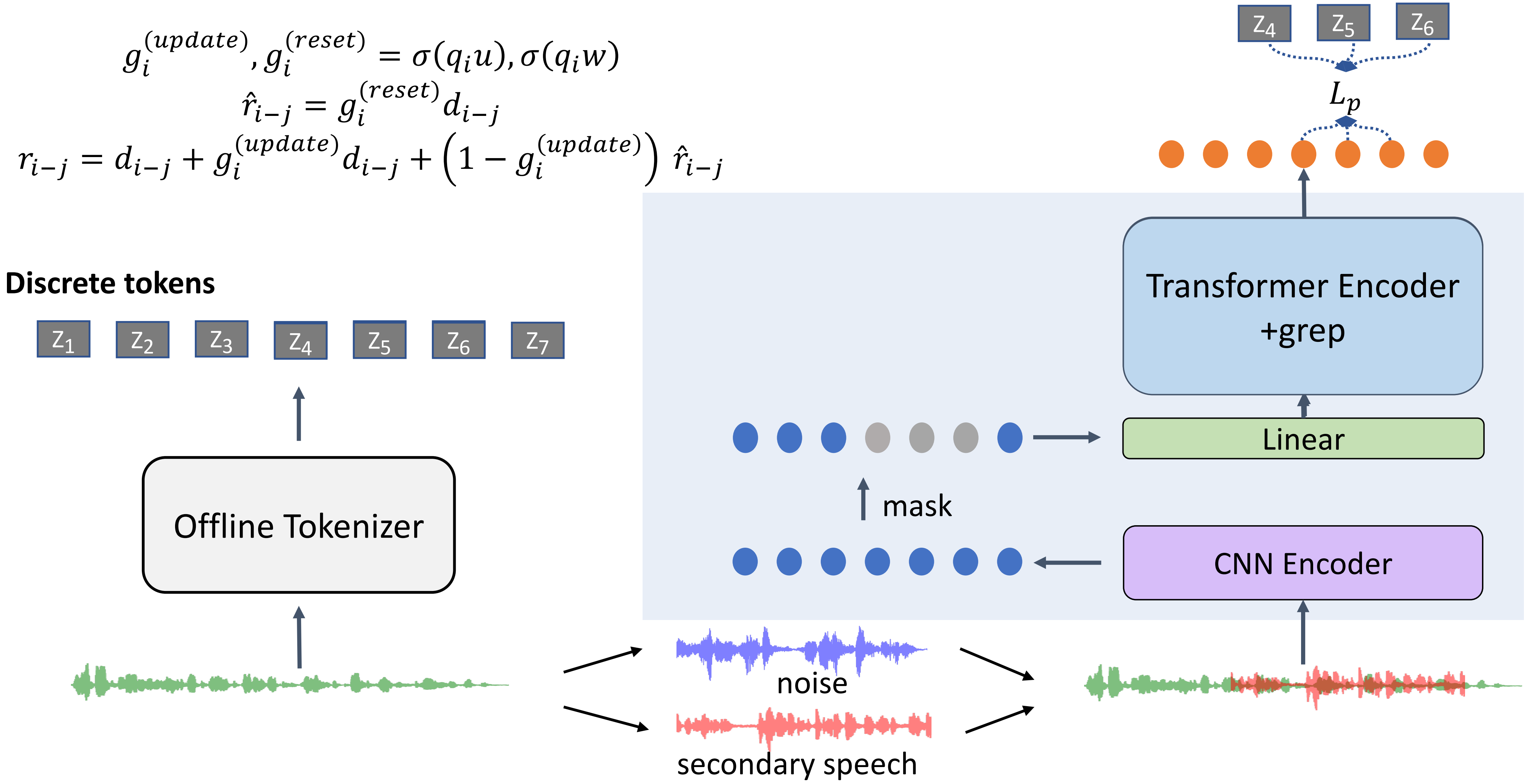


WavLM: Masked Speech **Prediction** and **Denoising**

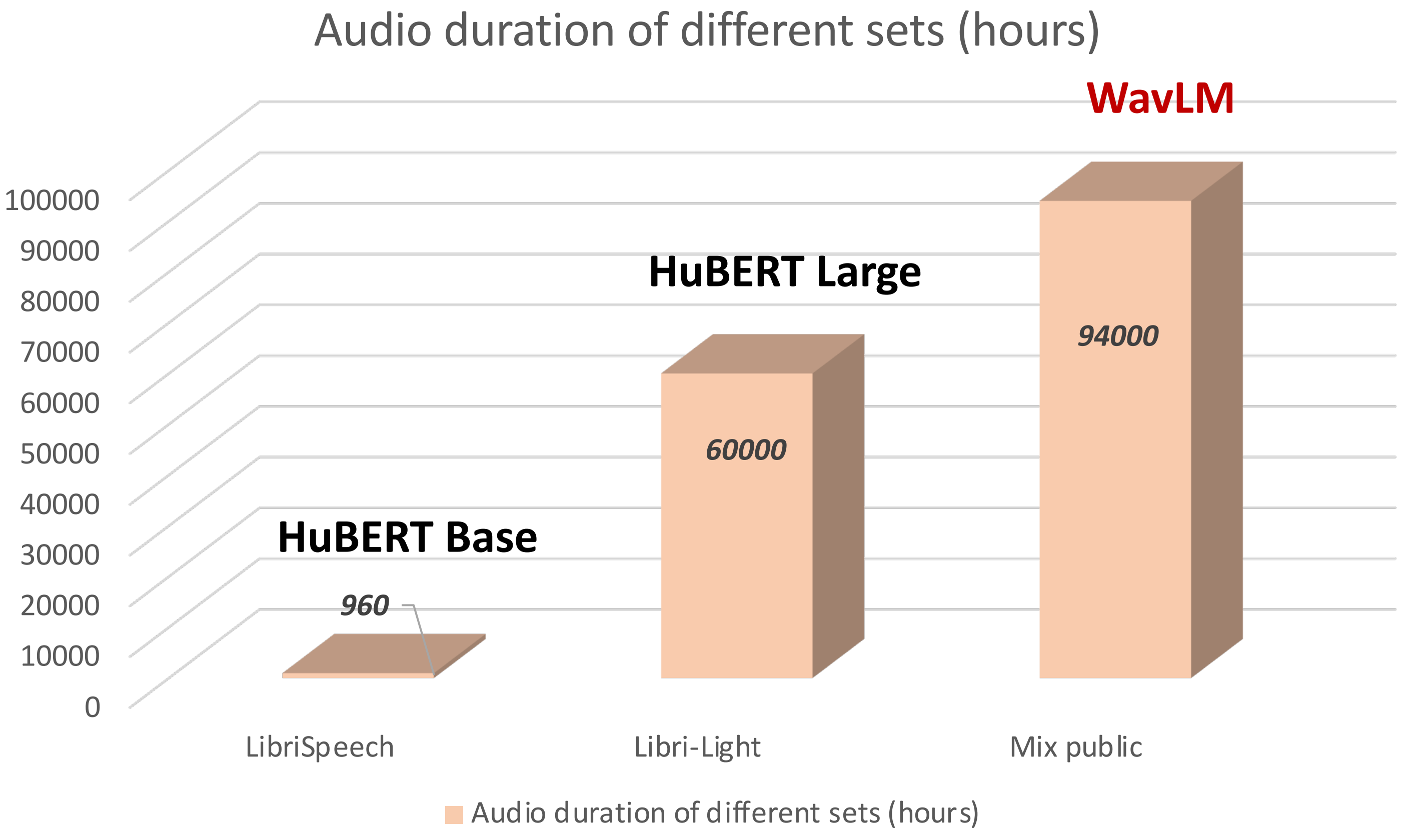
$$g_i^{(update)}, g_i^{(reset)} = \sigma(q_i u), \sigma(q_i w)$$

$$\hat{r}_{i-j} = g_i^{(reset)} d_{i-j}$$

$$r_{i-j} = d_{i-j} + g_i^{(update)} d_{i-j} + \left(1 - g_i^{(update)}\right) \hat{r}_{i-j}$$



Pre-Training Data

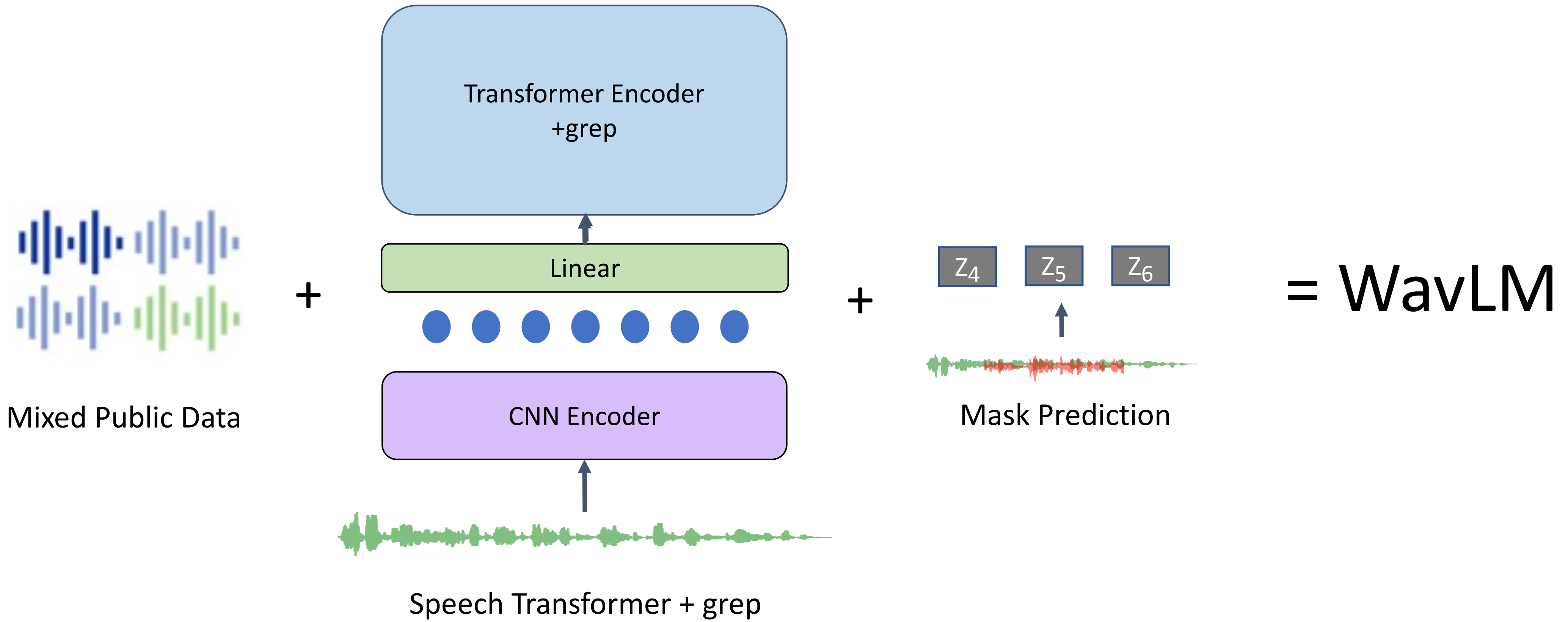


Public unlabeled data

- Libri-Light (60kh)
- VoxPopuli (24kh)
- GigaSpeech (10kh)

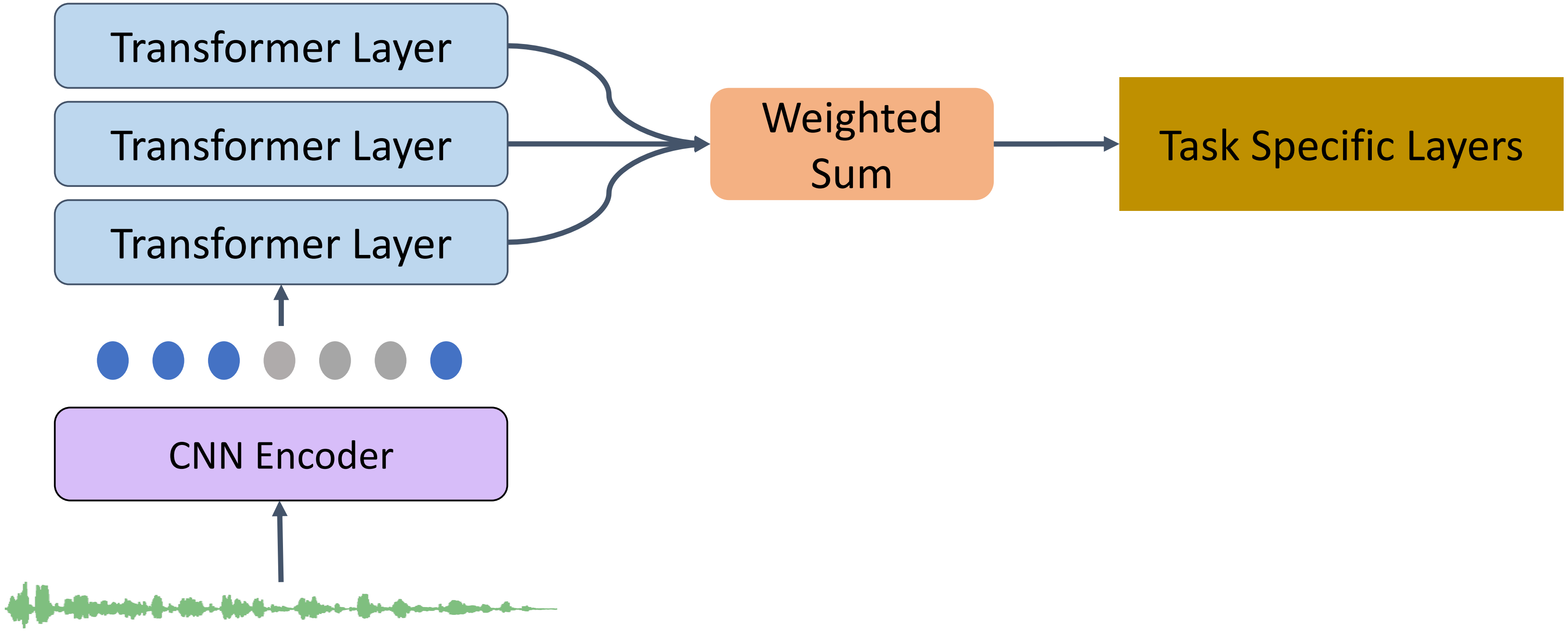
Three Key Elements

Data + Model + Task = Pre-training



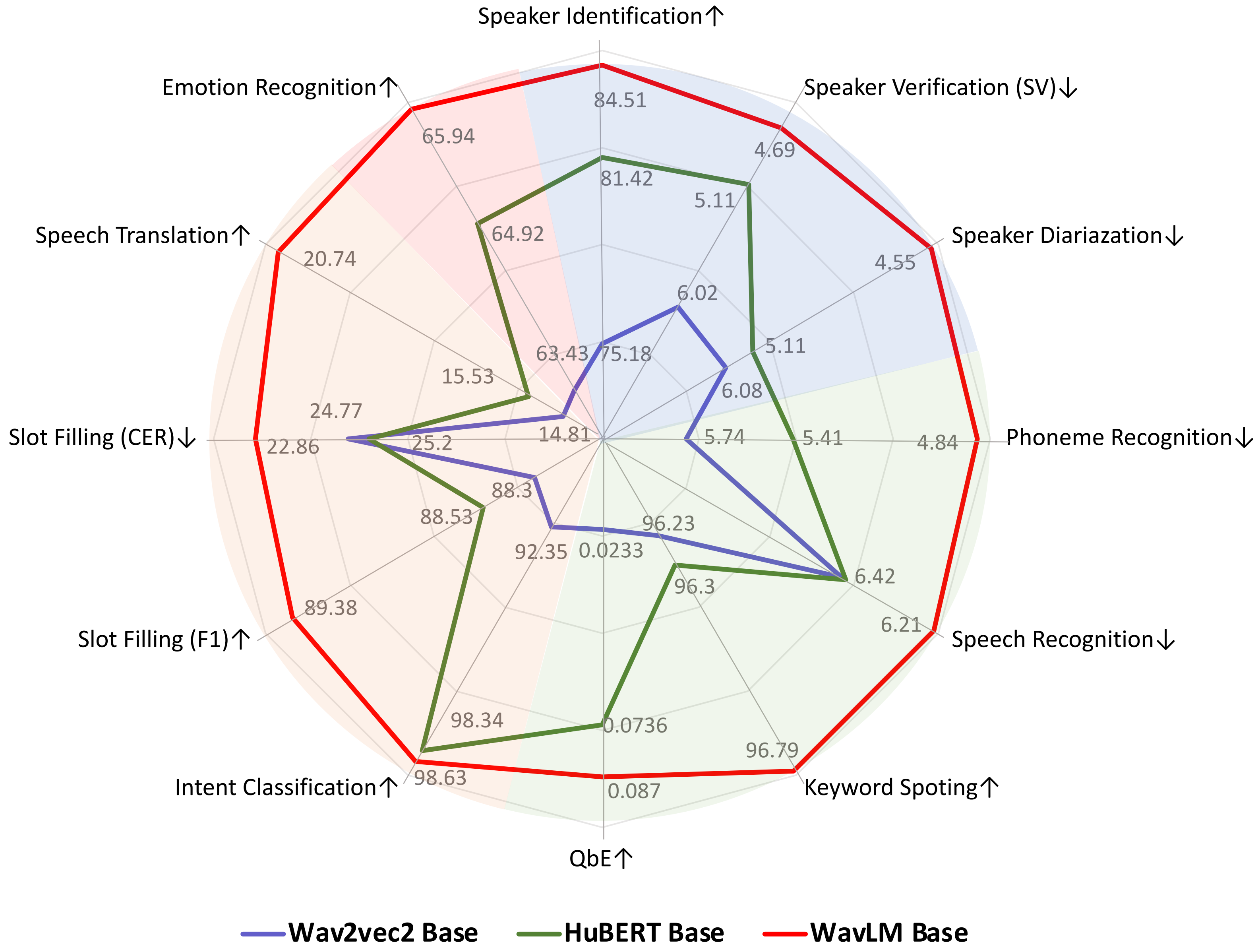
SuperB Finetuning Setup

Representations are weighted sum.



Yang et al. SUPERB: Speech processing Universal PERformance Benchmark Interspeech 2021

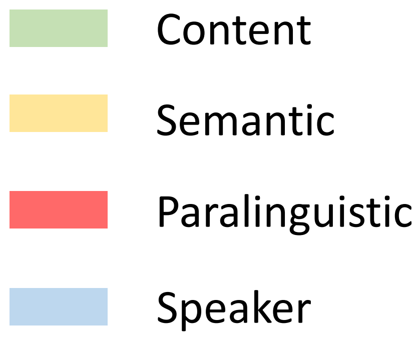
The **Best** Universal Speech Pre-trained Model



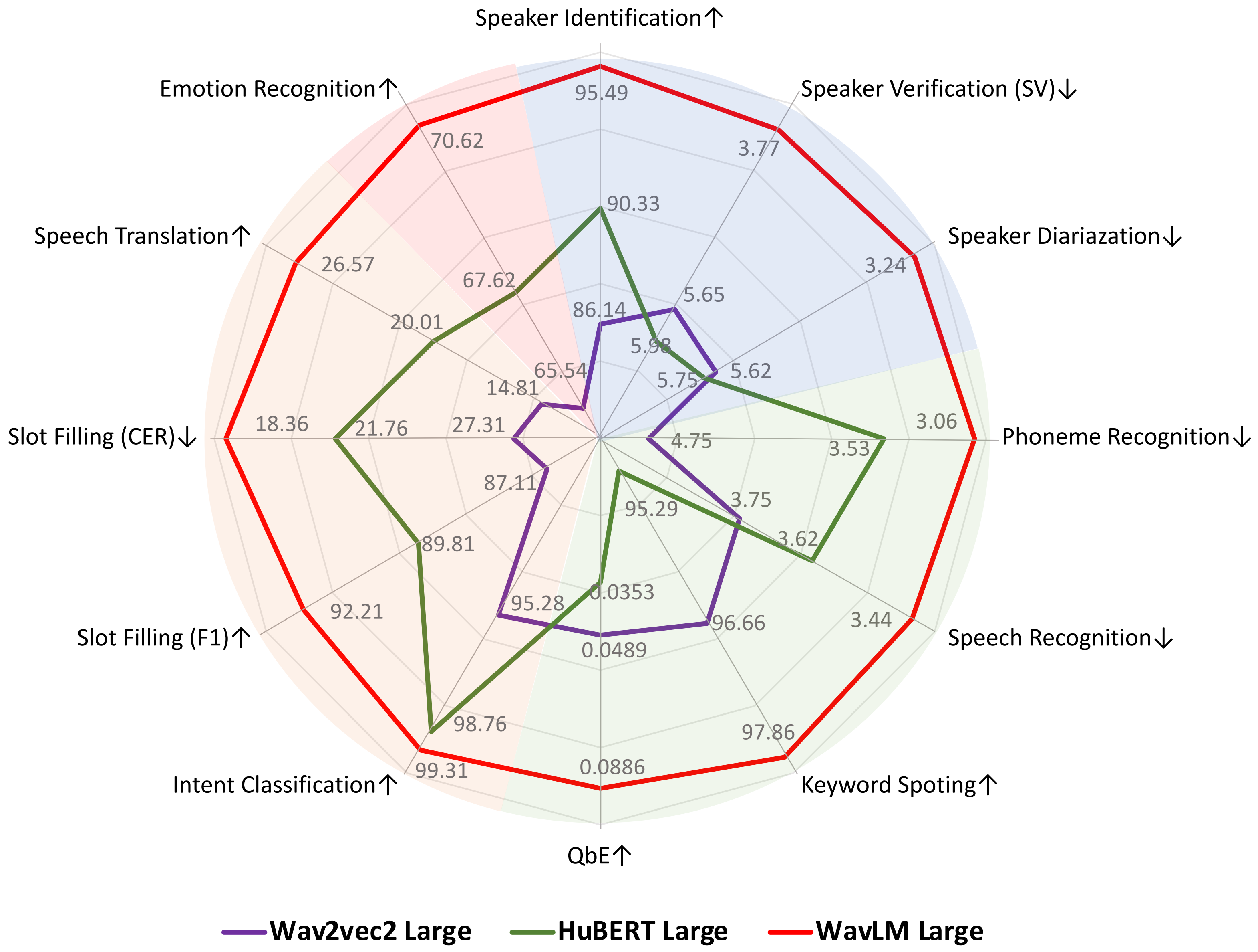
Base Model

Model size: 95M

Training data: 960hrs



The **Best** Universal Speech Pre-trained Model



Large Model

Model size: 316M

Training data:

Baseline(60k hrs)

WavLM(90k hrs)

Content

Semantic

Paralinguistic

Speaker

The **Best** Universal Speech Pre-trained Model

General rank and score

SUPERB

Speech processing **U**niversal **PER**formance **B**enchmark

Method	Rank ↑	Score ↑	PR ↓	KS ↑	IC ↑	SID ↑	ER ↑	ASR ↓	QbE ↑	SF-F1 ↑	SF-CER ↓	SV ↓	SD ↓
WavLM Large	22.8	1145	3.06	97.86	99.31	95.49	70.62	3.44	8.86	92.21	18.36	3.77	3.24
WavLM Base+	21.35	1106	3.92	97.37	99	89.42	68.65	5.59	9.88	90.58	21.2	4.07	3.5
WavLM Base	18.8	1019	4.84	96.79	98.63	84.51	65.94	6.21	8.7	89.38	22.86	4.69	4.55
data2vec Large	18.7	949	3.6	96.75	98.31	76.77	66.31	3.36	6.28	90.98	22.16	5.73	5.53
LightHuBERT Sta…	18.25	959	4.15	96.82	98.5	80.01	66.25	5.71	7.37	88.44	25.92	5.14	5.51
HuBERT Large	17.55	919	3.53	95.29	98.76	90.33	67.62	3.62	3.53	89.81	21.76	5.98	5.75
HuBERT Base	16.5	941	5.41	96.3	98.34	81.42	64.92	6.42	7.36	88.53	25.2	5.11	5.88
wav2vec 2.0 Large	16.3	914	4.75	96.66	95.28	86.14	65.64	3.75	4.89	87.11	27.31	5.65	5.62
data2vec base	15.6	884	4.69	96.56	97.63	70.21	66.27	4.94	5.76	88.59	25.27	5.77	6.67
LightHuBERT Small	14.65	901	6.6	96.07	98.23	69.7	64.12	8.34	7.64	87.58	26.9	5.42	5.85
FaST-VGS+	13.85	809	7.76	97.27	98.97	41.34	62.71	8.83	5.62	88.15	27.12	5.87	6.05
wav2vec 2.0 Base	12.65	818	5.74	96.23	92.35	75.18	63.43	6.43	2.33	88.3	24.77	6.02	6.08
DistilHuBERT	11.4	717	16.27	95.98	94.99	73.54	63.02	13.37	5.11	82.57	35.59	8.55	6.19
DeCoAR 2.0	10.8	722	14.93	94.48	90.8	74.42	62.47	13.02	4.06	83.28	34.73	7.16	6.59
wav2vec	8.9	529	31.58	95.59	84.92	56.56	59.79	15.86	4.85	76.37	43.71	7.99	9.9
vq-wav2vec	7	422	33.48	93.38	85.68	38.8	58.24	17.71	4.1	77.68	41.54	10.38	9.93
APC	5.8	392	41.98	91.01	74.69	60.42	59.33	21.28	3.1	70.46	50.89	8.56	10.53
VQ-APC	5.75	377	41.08	91.11	74.48	60.15	59.66	21.2	2.51	68.53	52.91	8.72	10.45

Phoneme Recognition (PR)
PER↓
Keyword Spotting (KS)
ACC↑
Intent Classification (IC)
ACC↑
Speaker Identification (SID)
ACC↑
Emotion Recognition (ER)
ACC↑
Automatic Speech Recognition (ASR)
WER↓
Query by Example Spoken Term Detection
MTWV↑
Slot Filling (Slot type)
F1↑
Slot Filling (Slot value)
CER↓
Speaker Diarization
DER↓
……

Ablation Study

	Speaker Identification ACC↑	Speaker Verification EER↓	Speaker Diarization DER↓	Query by Example MTWV↑	Phoneme Recognition PER↓	ASR WER↓	Keyword Spotting Acc↑	Intent Classification Acc↑	Slot Filling F1↑	Slot Filling CER↓	Emotion Recognition Acc↑
WavLM-Base	85.49	4.49	4.65	0.087	4.86	6.13	96.79	98.63	89.38	22.86	65.94
- denoising task	84.39	4.91	6.03	0.0799	4.85	6.08	96.79	98.42	88.69	23.43	65.55
- grep	84.74	4.61	4.72	0.0956	5.22	6.80	96.79	98.31	88.56	24.00	65.60
WavLM-Base+	89.42	4.07	3.50	0.0988	3.92	5.59	97.37	99.00	90.28	21.20	68.65

- Denoising task helps speaker related tasks.
- Gated relative position bias is effective on content modeling.

Academic Benchmarks

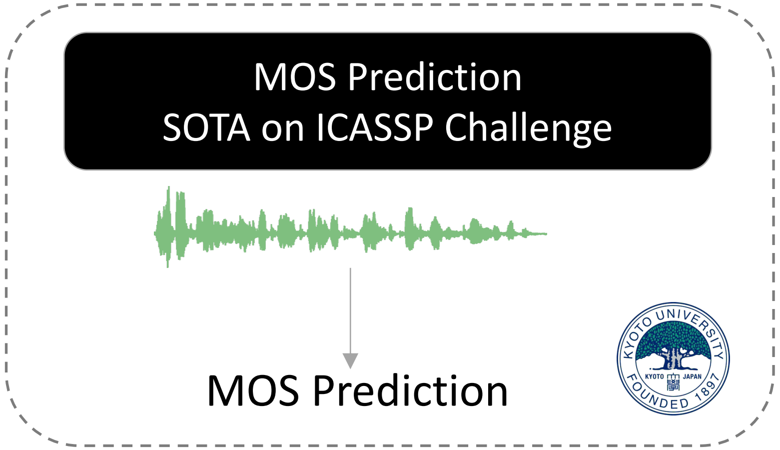
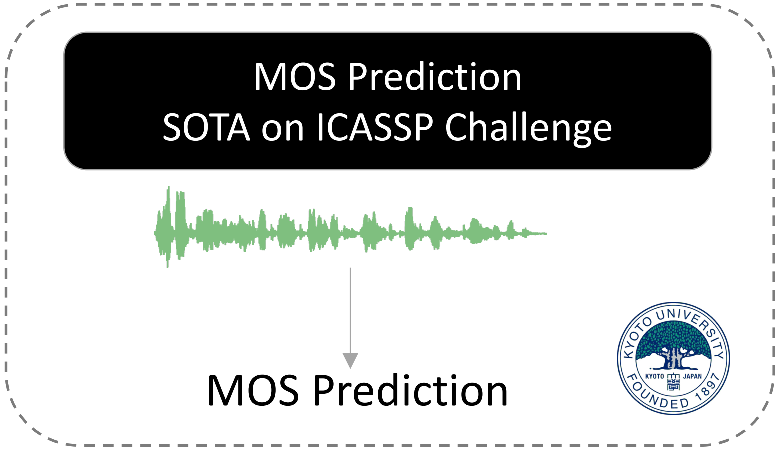


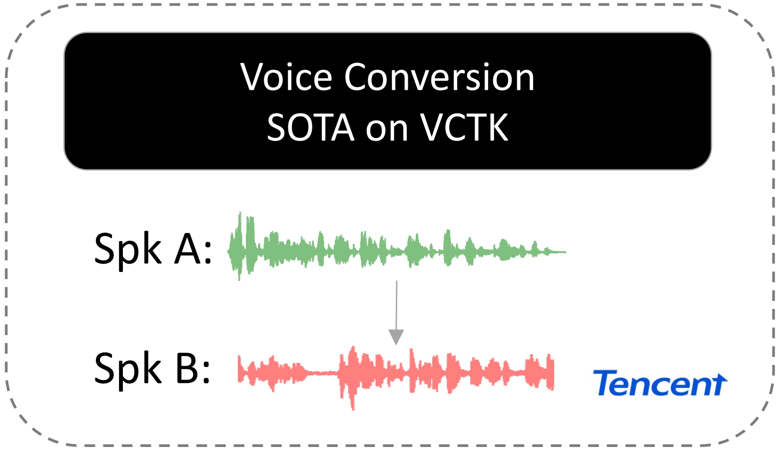
Table 1: *Results on the main track. Models finally selected for fusion are marked in bold.*

(a) Results of fine-tuning different pretrained SSL models individually for MOS prediction.

Pretrained SSL Model	Utterance level				System level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
W2V 2.0 Base	0.235	0.875	0.878	0.707	0.094	0.935	0.941	0.803
W2V 2.0 Large	0.197	0.875	0.873	0.697	0.068	0.948	0.953	0.820
W2V 2.0 Large (LV-60)	**0.191**	0.878	0.878	0.704	**0.060**	**0.950**	0.953	0.823
HuBERT Base	0.207	0.878	0.876	0.700	0.077	0.944	0.947	0.812
HuBERT Large	0.288	0.813	0.809	0.623	0.103	0.923	0.924	0.757
HuBERT Extra Large	0.229	0.852	0.849	0.666	0.082	0.930	0.931	0.777
WavLM Base	0.199	**0.891**	**0.891**	**0.722**	0.072	0.949	0.954	0.828
WavLM Base+	0.248	0.879	0.883	0.709	0.115	0.948	**0.958**	**0.832**
WavLM Large	0.192	0.876	0.872	0.695	0.063	**0.950**	0.952	0.827
Data2Vec	0.314	0.826	0.842	0.660	0.144	0.905	0.931	0.779

Academic Benchmarks



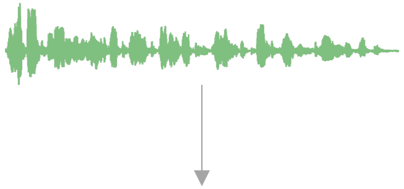


model	seen to seen naturalness	seen to seen similarity	unseen to unseen naturalness	unseen to unseen similarity
AUTOVC [17]	2.65±0.12	2.86±0.09	2.47±0.10	2.76±0.08
AdaIN-VC [17]	2.98±0.09	3.06±0.07	2.72±0.11	2.96±0.09
DSVAE [17]	3.40±0.07	3.56±0.06	3.22±0.09	3.54±0.07
DSVAE(HiFi-GAN)	3.76±0.07	3.83±0.06	3.65±0.07	3.89±0.05
C-DSVAE(BEST-RQ)	3.88±0.06	3.93±0.07	3.82±0.08	3.98±0.07
C-DSVAE(Mel)	3.86±0.10	3.65±0.07	3.78±0.05	3.58±0.08
C-DSVAE(Align)	4.03±0.04	4.12±0.07	3.93±0.06	4.06±0.07
C-DSVAE(WavLM)	4.08±0.06	4.17±0.06	3.98±0.07	4.12±0.05

Table 3: *The MOS (95% CI) test on different models.*

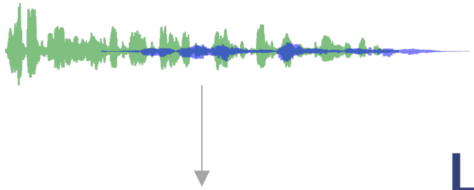
Academic Benchmarks

MOS Prediction
SOTA on ICASSP Challenge



MOS Prediction

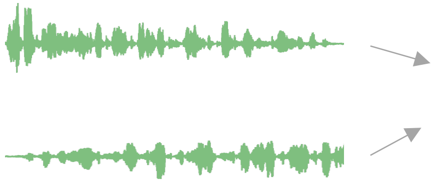
Overlapped Detection
SOTA on DIHARD3



Overlapped

LIUM
Laboratoire d'Informatique
Université du Mans

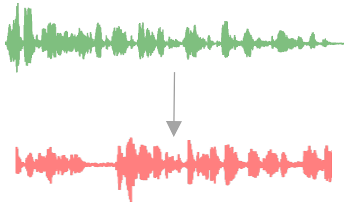
Speaker Verification
SOTA on VoxCeleb Challenge



Same Speaker

Voice Conversion
SOTA on VCTK

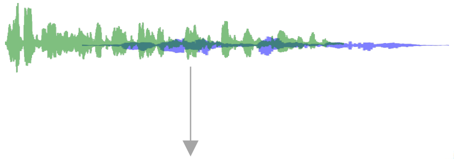
Spk A:



Spk B:

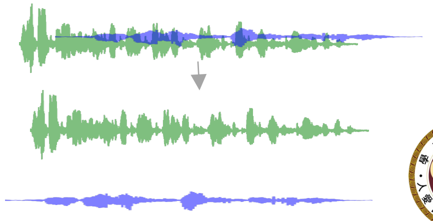
Tencent

Noisy Speech Recognition
SOTA on CHIME-4



Trans: What's the weather?

Speech Enhancement
SOTA on Demand



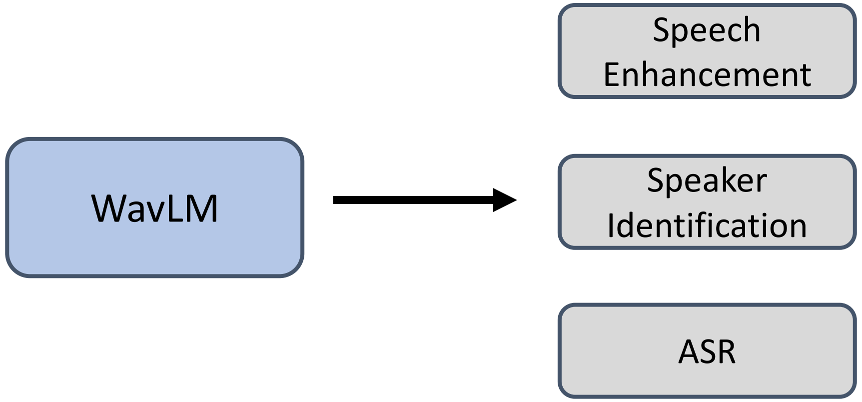
Spk A:

Noise:

Takeaway

Can a single universal model benefit various speech tasks?

Yes, WavLM is a pre-trained model for full-stack speech processing tasks.



SUPERB Achieve SOTA on 13 tasks

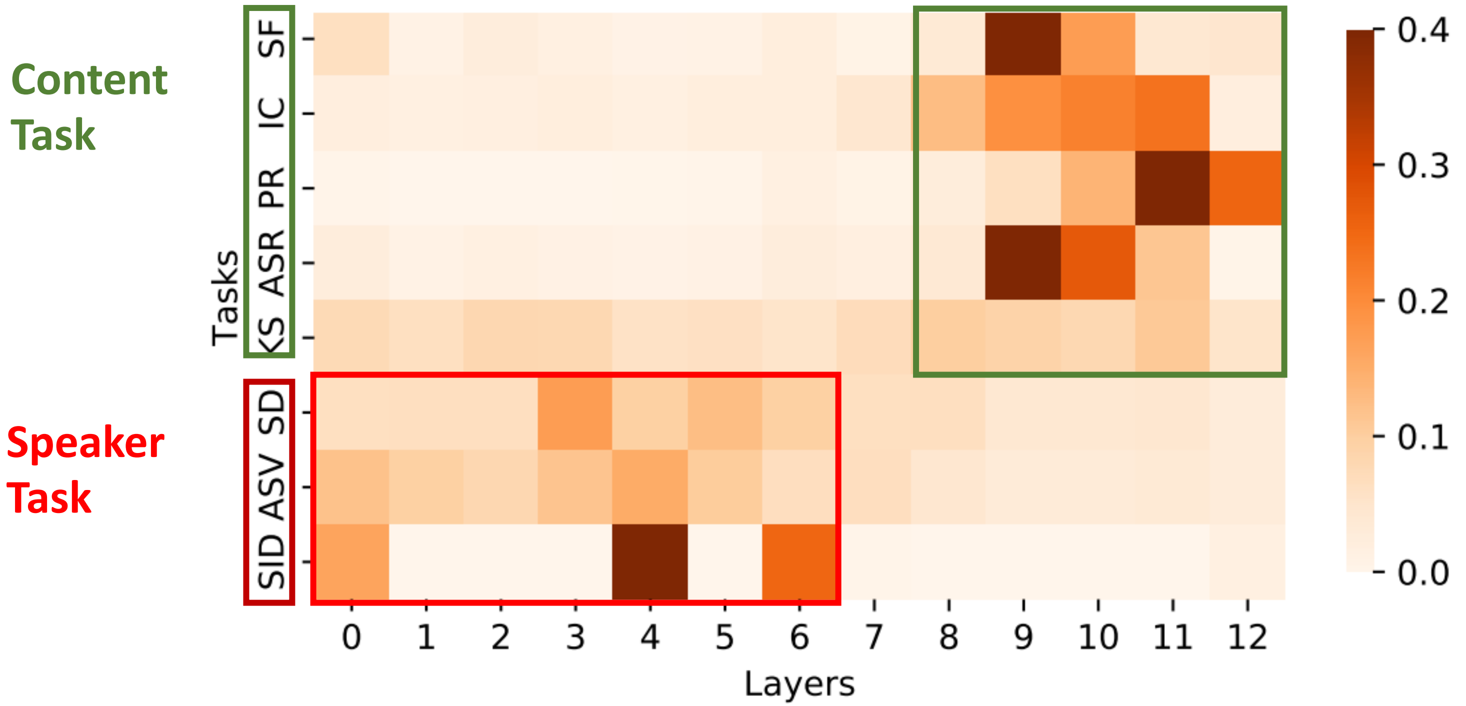
180 citations

Contribution for Products

Integrate to HuggingFace

Analysis

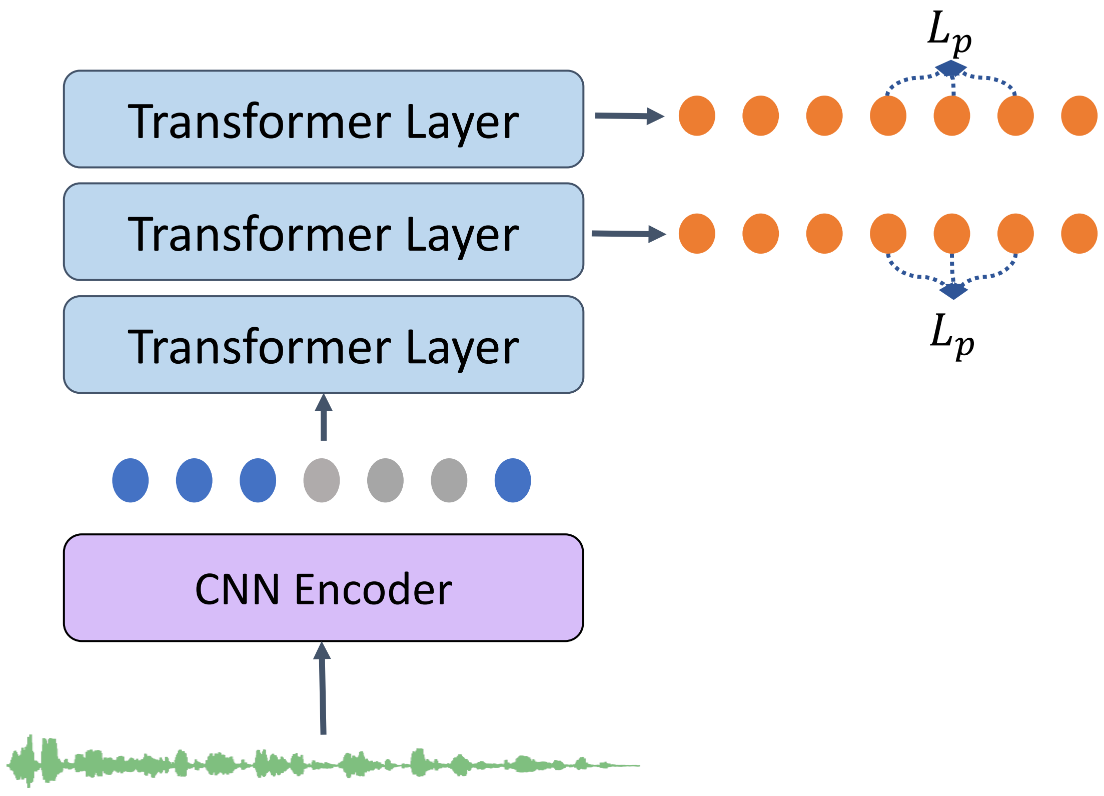
What's the relationship between representations and tasks?



Lower layers -> speaker
Higher layers -> content

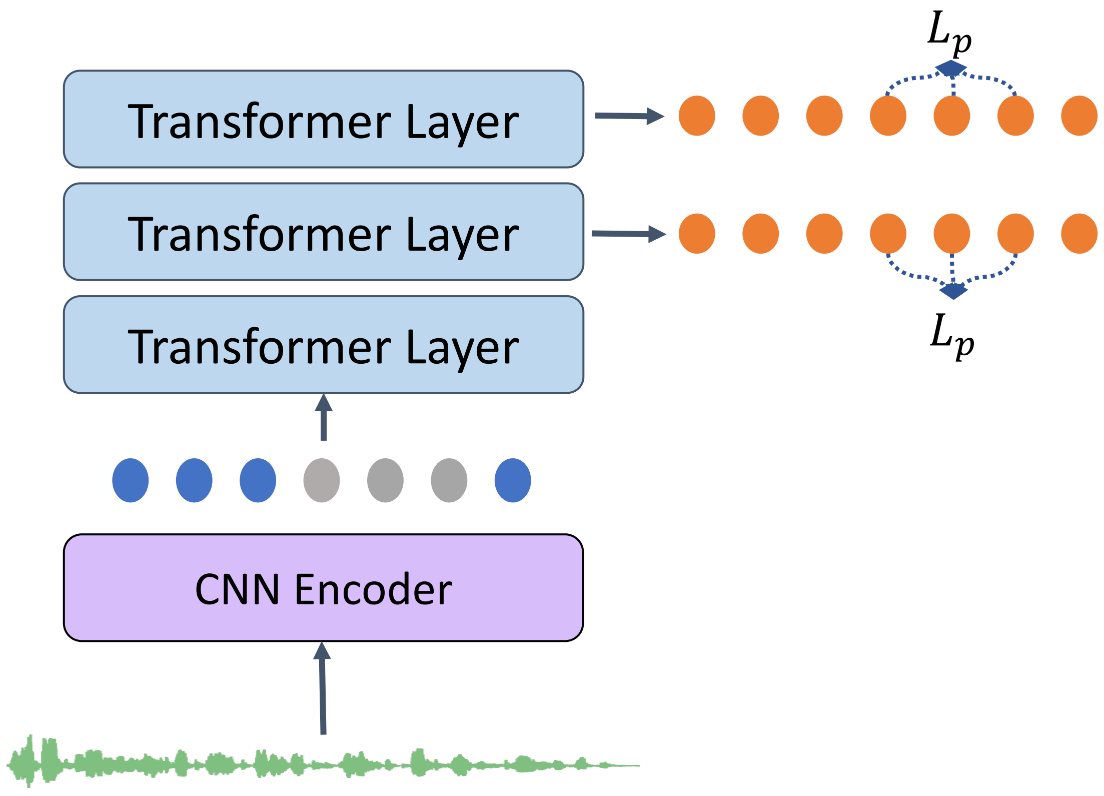
Self-Supervised Learning for Speech Recognition with Intermediate Layer Supervision

Intermediate Layer Supervision



	Clean	WERR	Other	WERR
HuBERT (1h)	20.9	-	27.5	-
WavLM (1h)	24.5	-	29.2	-
ILS-SSL (1h)	**17.9**	**14%**	**23.1**	**16%**
HuBERT (10h)	10.1	-	16.8	
WavLM (10h)	9.8	3%	16.0	5%
ILS-SSL (10h)	**8.3**	**18%**	**13.6**	**19%**
HuBERT (100h)	6.0	-	13.0	-
WavLM (100h)	5.7	5%	12.0	9%
ILS-SSL (100h)	**4.7**	**22%**	**10.1**	**24%**

Intermediate Layer Supervision



960h	LM	Clean	Other
Transformer-CTC	Transf	2.5	5.5
Transformer-S2S	Transf	2.3	5.2
Transformer-Transducer	Transf	2.0	4.6
Conformer-Transducer	LSTM	1.9	3.9
HuBERT	None	2.1	4.3
HuBERT	4-gram	2.0	3.7
HuBERT	Transf	1.9	3.3
ILS-SSL	None	**1.9**	**3.8**
ILS-SSL	4-gram	**1.9**	**3.4**
ILS-SSL	Transf	**1.8**	**3.2**

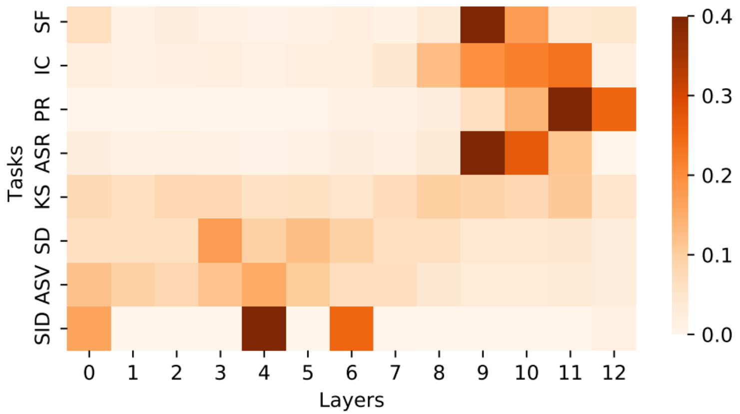
Analysis: Evaluation on SuperB Benchmark

	Speaker		
	SID (ACC↑)	ASV (EER↓)	SD (DER↓)
Fbank	0	9.56	10.05
HuBERT	**81.42**	**5.11**	**5.88**
ISL-SSL	79.29	5.24	6.31
	Content		
	PR (PER↓)	ASR (WER↓)	QbE (MTWV↑)
Fbank	82.01	23.18	0.0058
HuBERT	5.41	6.42	0.0736
ILS-SSL	**5.00**	**5.45**	**0.0789**
	Semantics		
	IC (ACC↑)	SF (F1↑)	SF (CER↓)
Fbank	9.10	69.64	52.94
HuBERT	98.34	88.53	25.20
ILS-SSL	**98.47**	**89.16**	**24.29**

Takeaway

1. Can a single universal model benefit various speech tasks?
2. What's the relationship between representations and tasks?

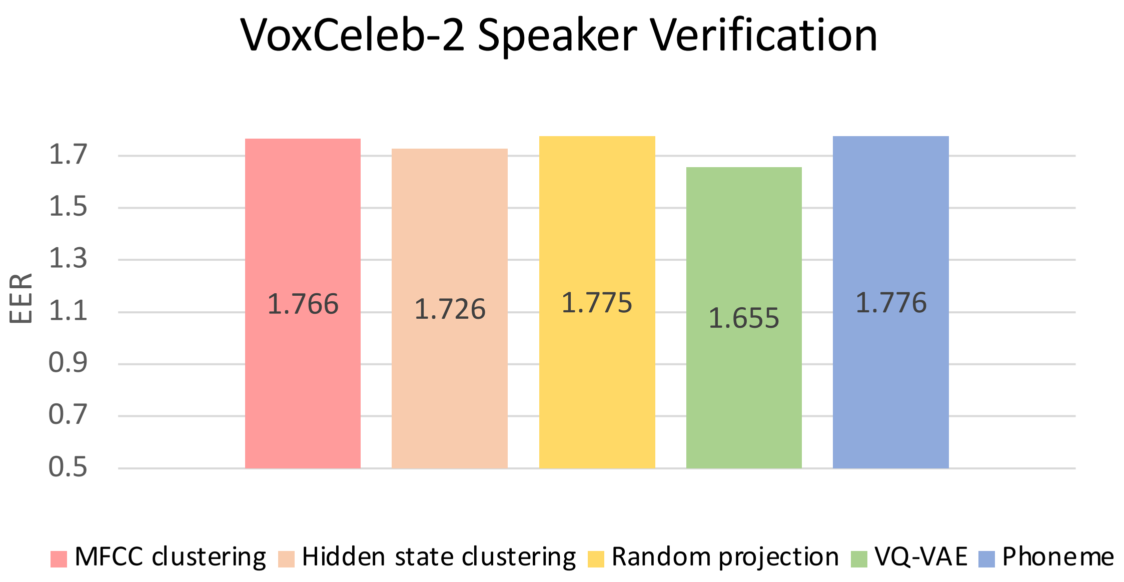
Bottom layers learn speaker information, and top layers learn content information.

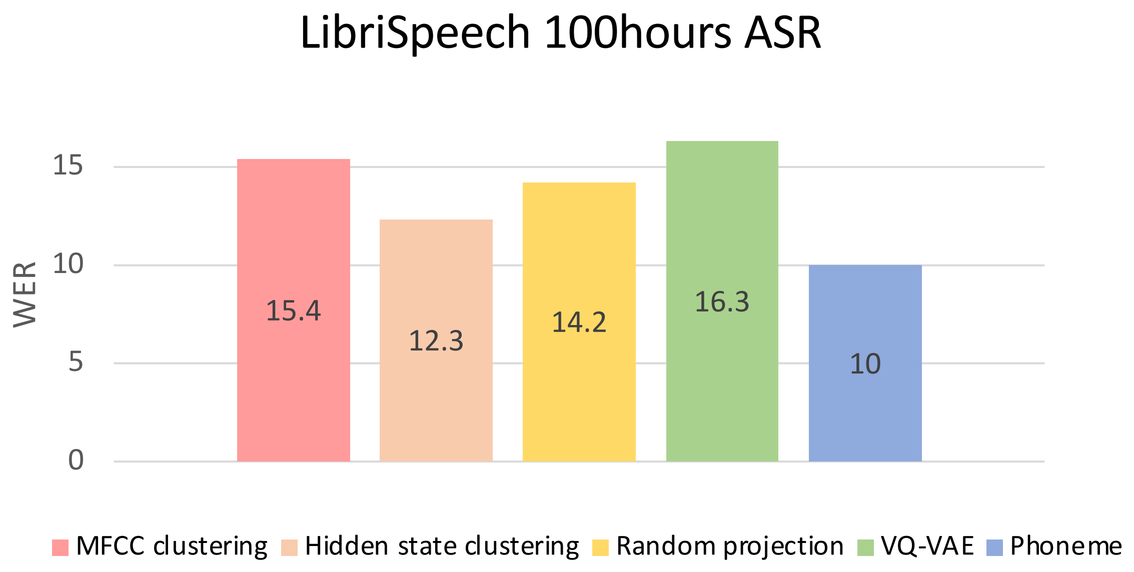
WavLM → 

→ ILS

Analysis

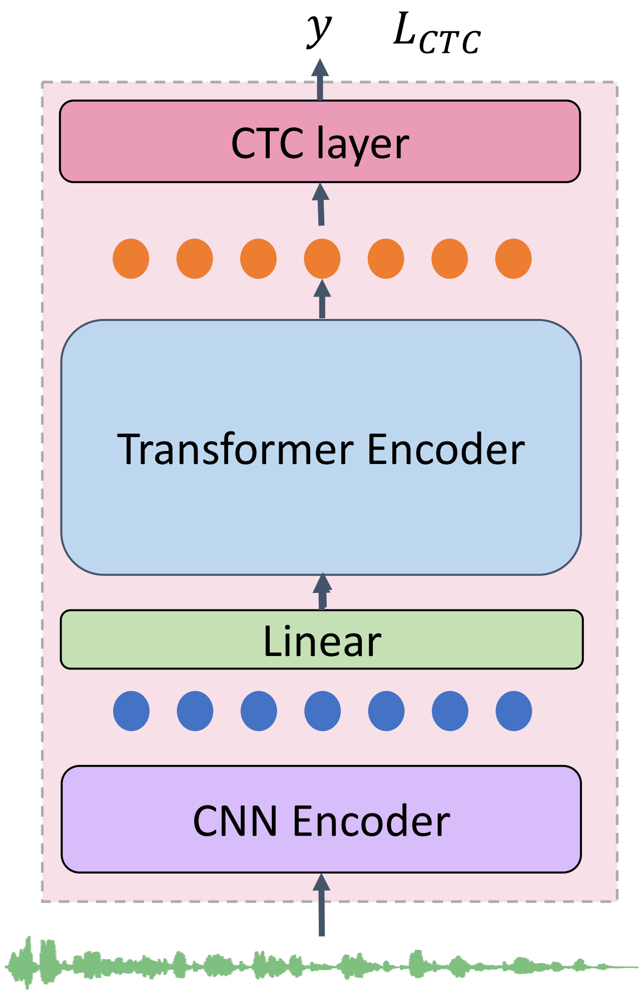
Does tokenizer matter for pre-trained model?

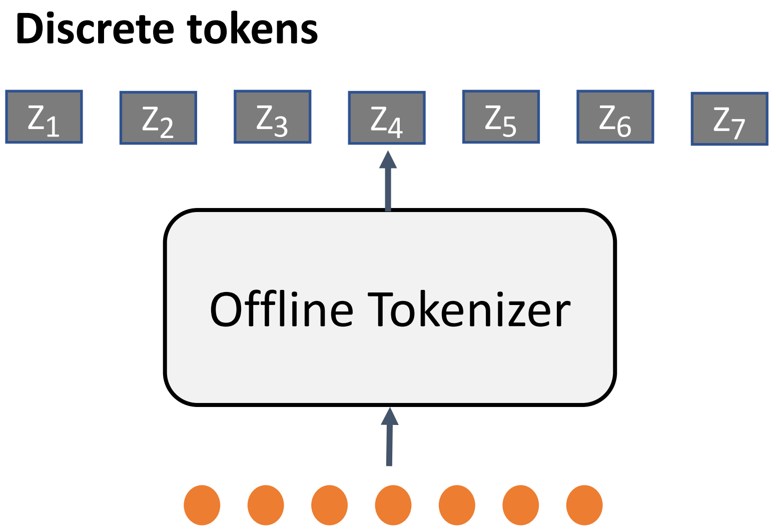




Supervision-Guided Codebooks for Masked Prediction in Speech Pre-training

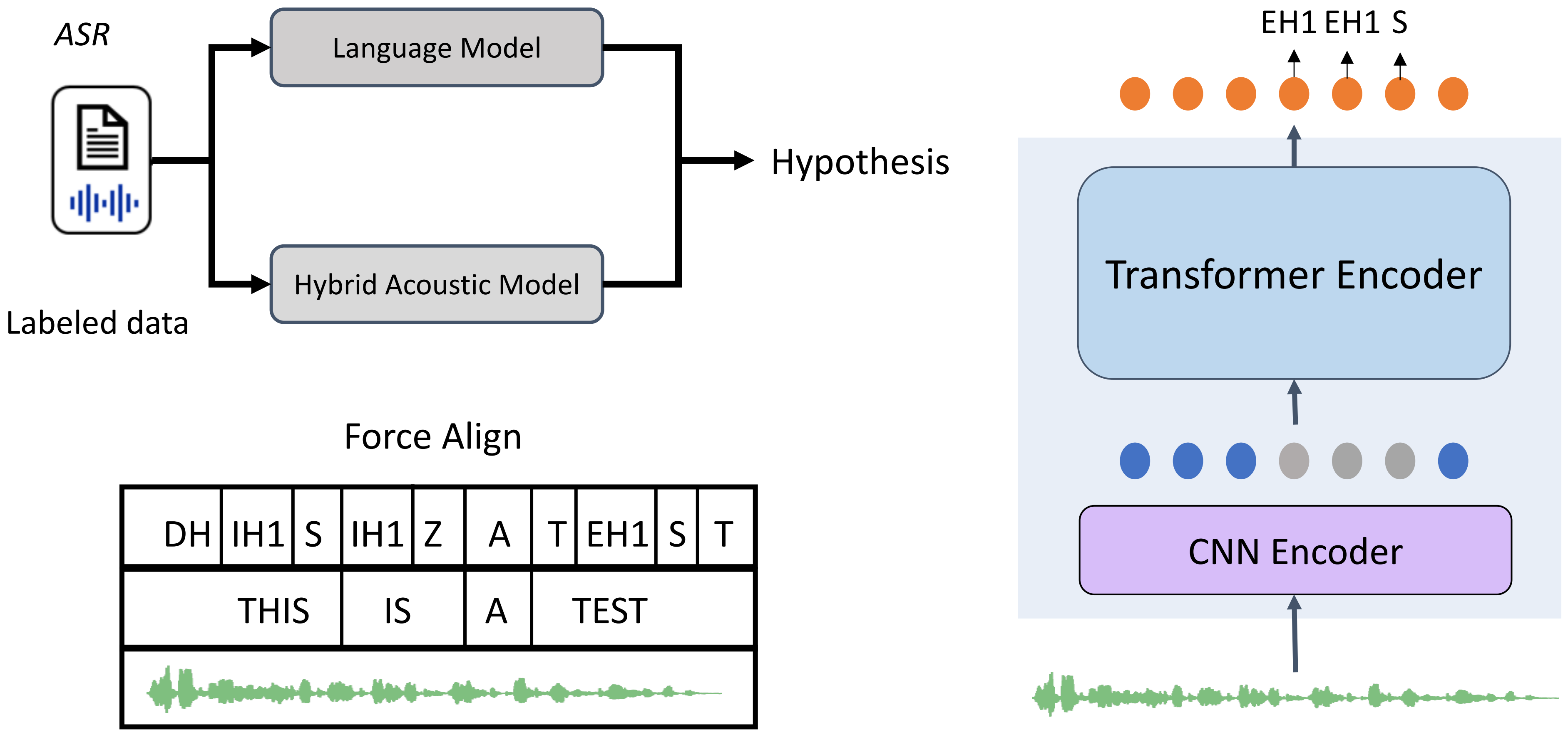
Clustering on Supervised Features





100h	Clean	WERR	Other	WERR
HuBERT Base	5.9	-	13.0	-
WavLM Base	5.7	3%	12.0	9%
CTC clustering	**5.2**	**12%**	**11.4**	**12%**

P-BERT

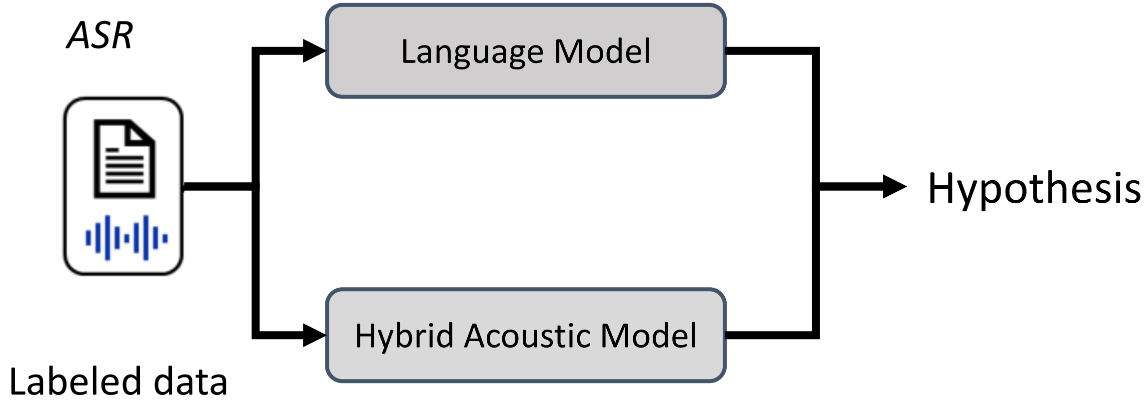

ASR
Language Model
Hybrid Acoustic Model
Labeled data
Hypothesis
Force Align
DH IH1 S IH1 Z A T EH1 S T
THIS IS A TEST
EH1 EH1 S
Transformer Encoder
CNN Encoder

Results

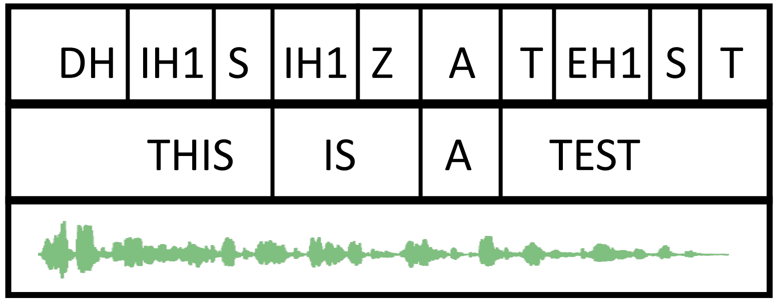
- Unlabeled data: 960h
- Labeled data: 100h

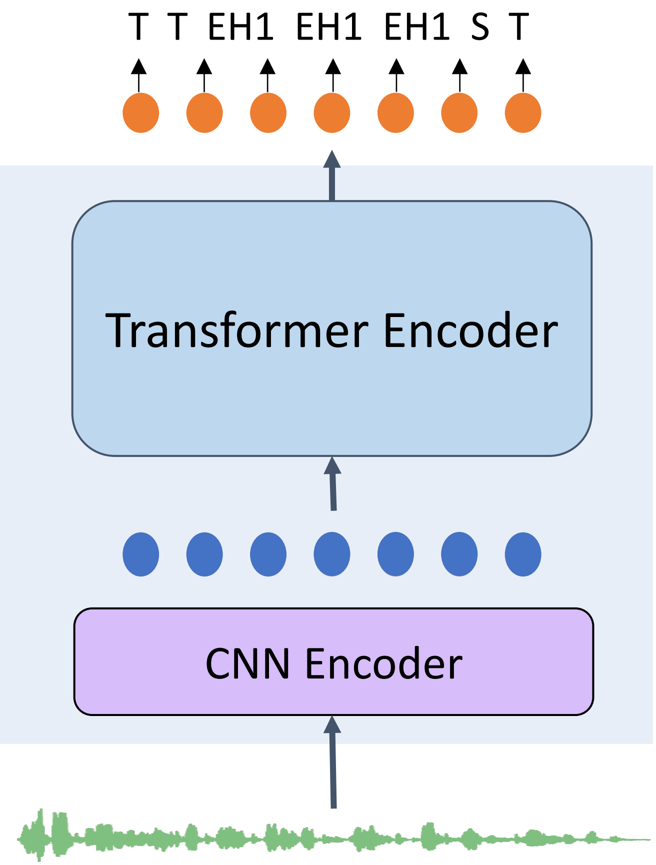
	Without LM		**With LM**	
	Clean	Other	Clean	Other
Hubert	6.0	13.0	3.4	8.1
WavLM	5.7	12.0	3.4	7.7
Noisy Student	4.9	14.4	3.5	9.7
Our Method	**4.7**	**11.2**	**3.1**	**7.5**

P-BERT



Forced Alignment





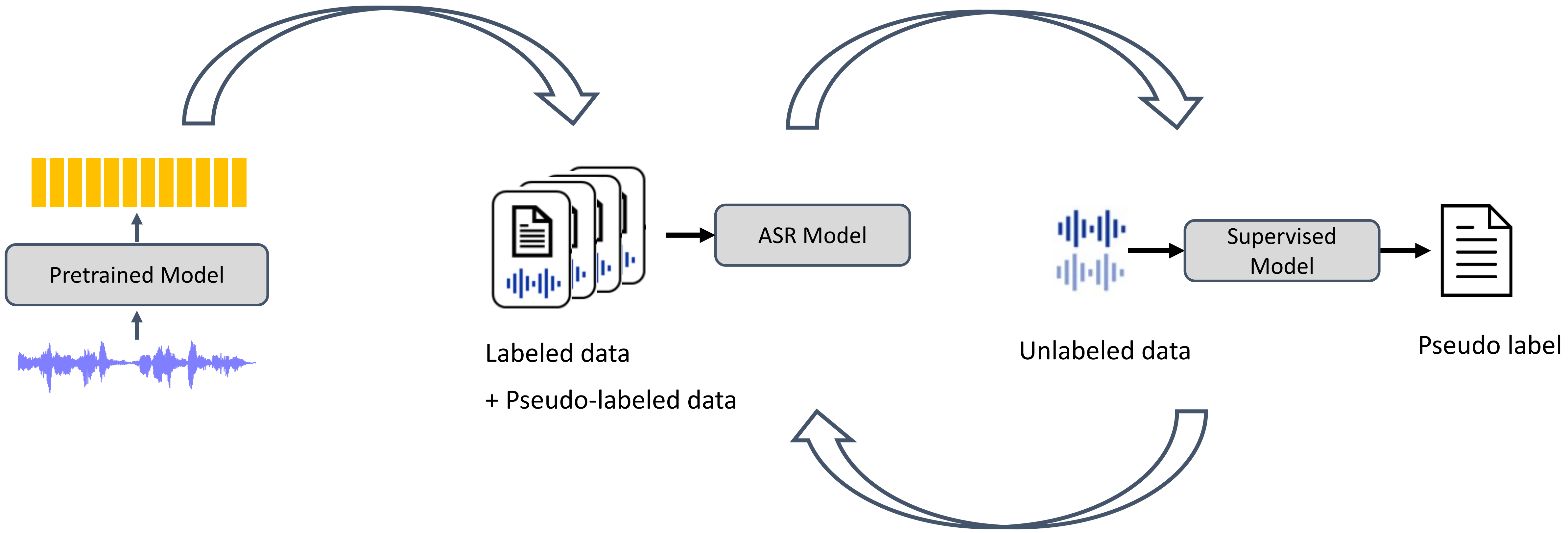
Results

- Unlabeled data: 960h
- Labeled data: 100h

	Without LM		With LM	
	Clean	Other	Clean	Other
Hubert	6.0	13.0	3.4	8.1
WavLM	5.7	12.0	3.4	7.7
Noisy Student	4.9	14.4	3.5	9.7
+2nd iter	**4.3**	11.0	3.3	8.4
Our Method	4.7	11.2	3.1	7.5
+2nd iter	4.7	**10.7**	**3.1**	**7.3**

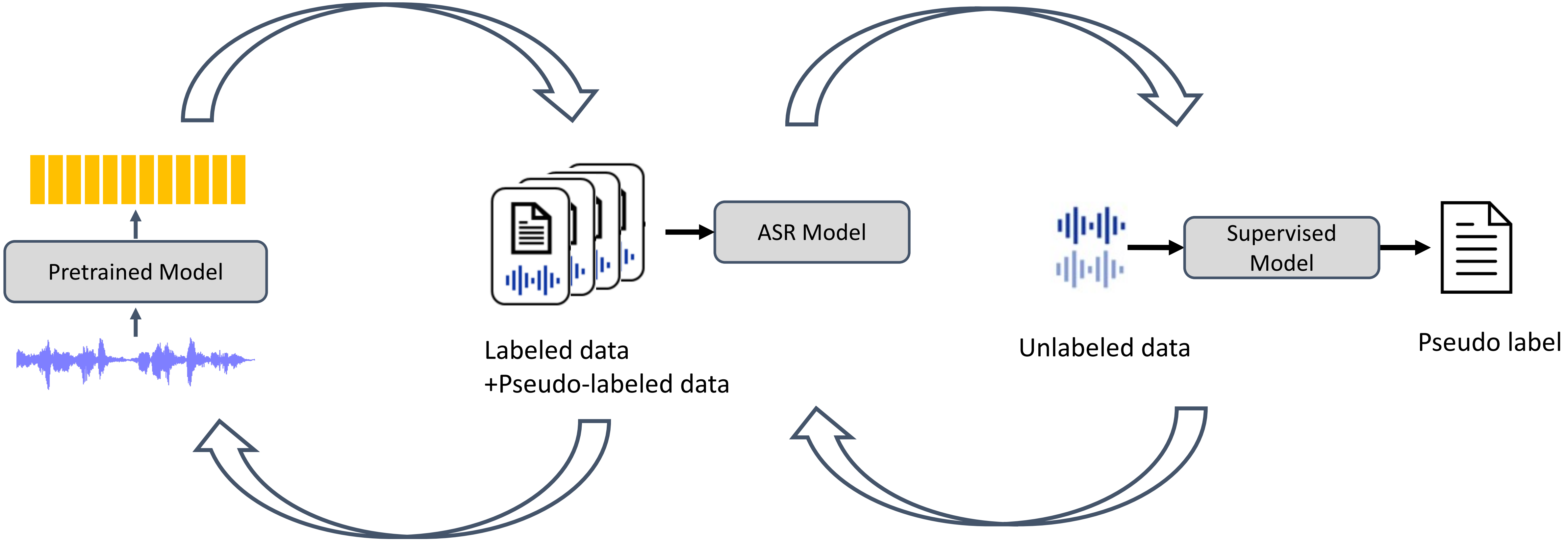
Combination with Noisy Student Learning

Previous Work



Combination with Noisy Student Learning

Our Work



Results

- Unlabeled data: 960h
- Labeled data: 100h

	Without LM		With LM	
	Clean	Other	Clean	Other
Hubert	6.0	13.0	3.4	8.1
WavLM	5.7	12.0	3.4	7.7
Noisy Student	4.9	14.4	3.5	9.7
+2nd iter	4.3	11.0	3.3	8.4
Our Method	4.7	11.2	3.1	7.5
+2nd iter	4.7	10.7	3.1	7.3
+ noisy student	**4.2**	**9.5**	**3.1**	**7.2**
Our Method + ILS	4.1	9.6	3.0	7.0
+noisy student	**3.2**	**7.0**	**2.8**	**6.1**

Takeaway

1. Can a single universal model benefit various speech tasks?
2. What's the relationship between representations and tasks?
3. Does tokenizer matter for pre-trained model?

   It depends! No for Speaker Verification, Yes for Speech Recognition

Thank you