In the name of GOD

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

Maryam Afshari

Sharif Speech and Language Processing Lab

Introduction

- self-supervised learning (SSL) has achieved great success in the fields of (NLP)

- It leverages large amounts of text data to learn universal text representations, which can benefit almost all NLP downstream tasks by fine-tuning.

- Recently, SSL has also shown prominent results for speech processing, especially on phoneme classification (van den Oord et al., 2018) and (ASR) (Baevski et al., 2020b; Hsu et al., 2021a; Wang et al., 2021b).

- As speech signal contains multifaceted information including speaker identity, paralinguistics, spoken content, etc.

learning universal representations for all speech tasks is challenging.

Introduction

- Building a general pre-trained model can be essential to the further development of speech processing, because it can utilize large-scale unlabeled data to boost the performance in downstream tasks, reducing data labeling efforts.

- In the past, it has been infeasible to build such a general model, as different tasks focus on different aspects of speech signals.

• Speaker Verification	• Speech Recognition	• Speaker diarization & Speech separation
speaker characteristic	~~speaker characteristic~~	
~~Spoken content~~	Spoken content	Multiple Speaker

Introduction

- Building a general pre-trained model can be essential to the further development of speech processing, because it can utilize large-scale unlabeled data to boost the performance in downstream tasks, reducing data labeling efforts.

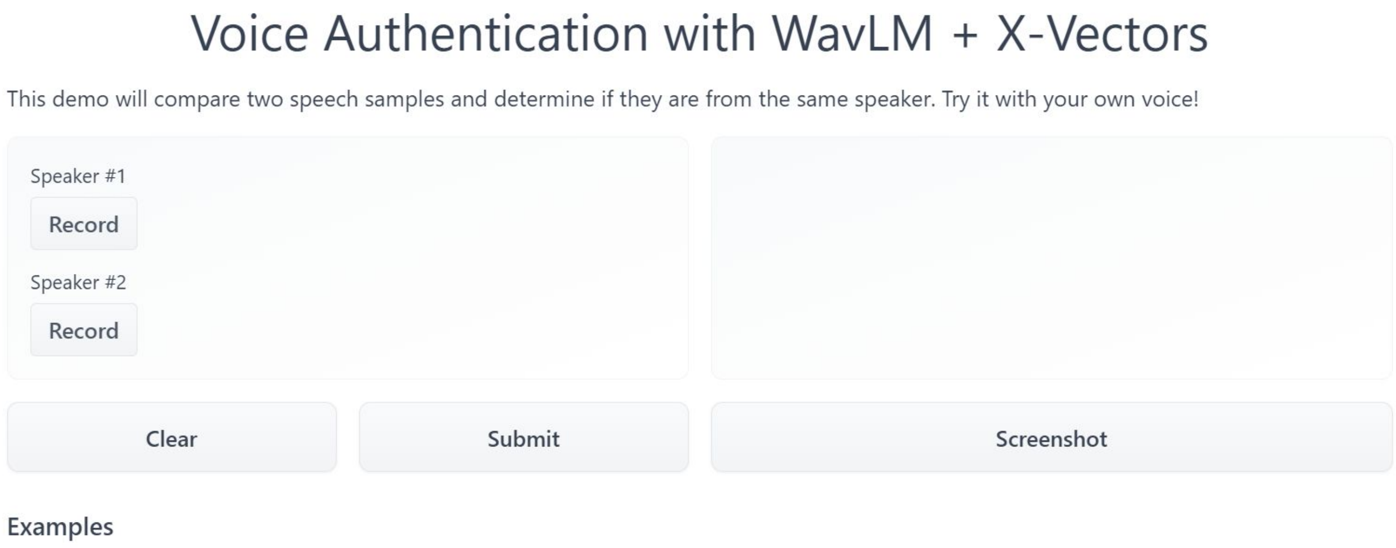
- In the past, it has been infeasible to build such a general model, as different tasks focus on different aspects of speech signals.

• Speaker Verification	• Speech Recognition	• Speaker (
		Speech
speaker characteristic	~~speaker characteristic~~	
~~Spoken content~~	Spoken content	Multiple Speaker

creates **additional obstacles** for learning general speech representations.

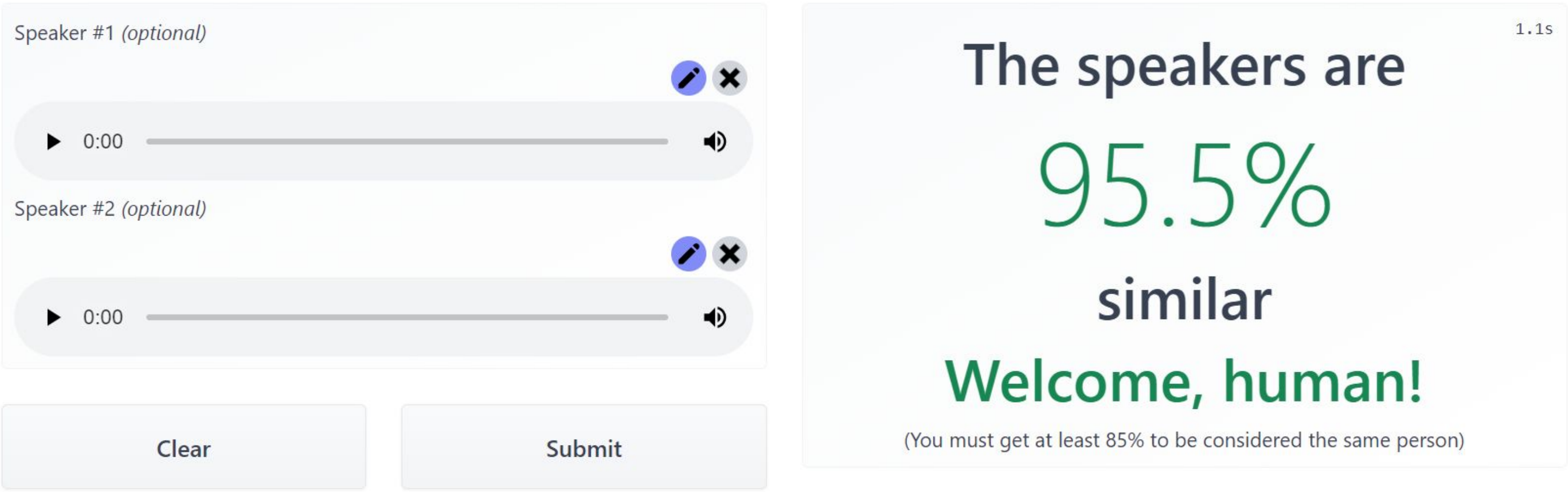
Speaker Verification

https://huggingface.co/spaces/microsoft/wavlm-speaker-verification



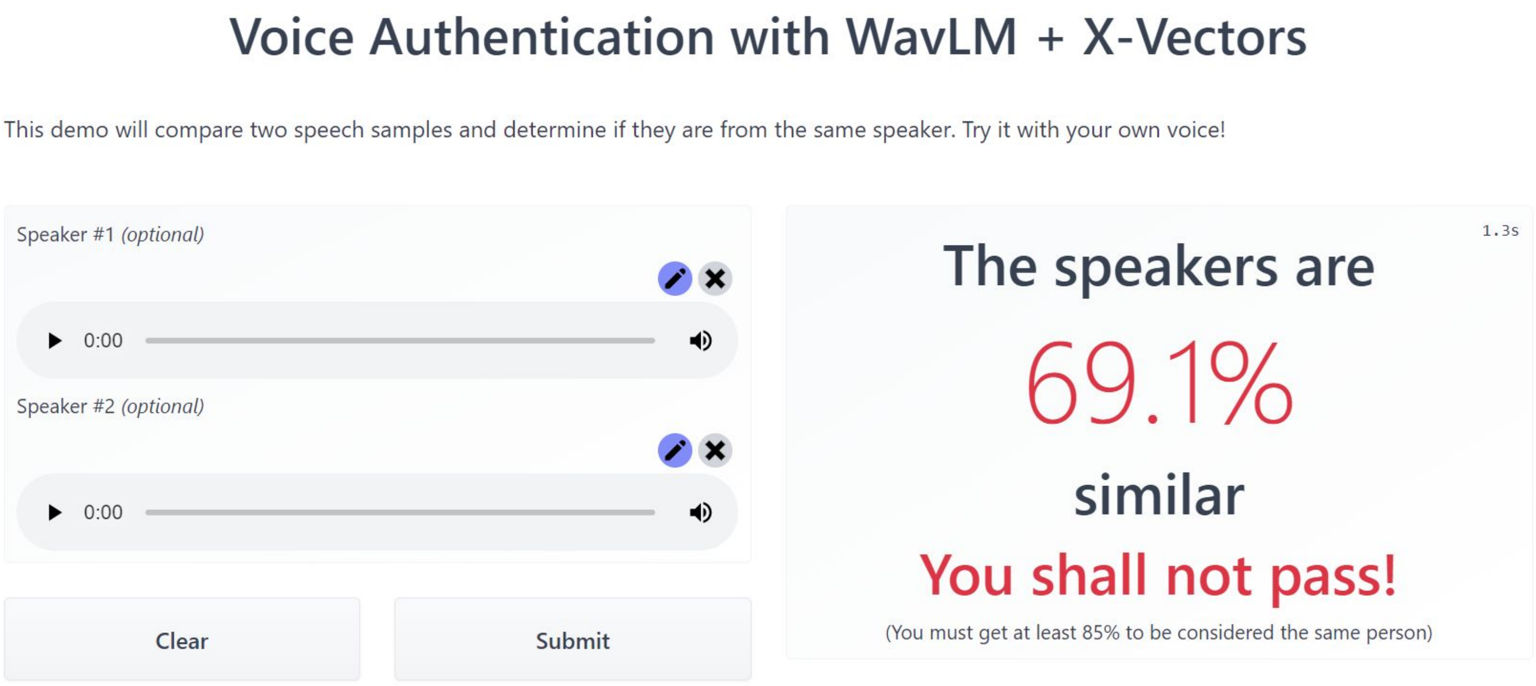
Speaker Verification

https://huggingface.co/spaces/microsoft/wavlm-speaker-verification



Speaker Verification

https://huggingface.co/spaces/microsoft/wavlm-speaker-verification



Speech Separation

Let's **say we want to write a program to generate the lyrics of a song**. As we know this process includes the usage of Automatic Speech Recognition(ASR). But will it be able to recognize the speech properly?
While some of the state-of-the-art methods can, it still won't be able to recognize the lyrics because of the background music.

Speech separation is also called the cocktail party problem.

The audio can contain **background noise, music, speech by other speakers, or even a combination of these.**

the task of extracting the target speech signal from a mixture of sounds as *speech enhancement*.

Example:
https://www.analyticsvidhya.com/blog/2021/08/speech-separation-by-facebook-ai-research/#:~:text=What%20is%20speech%20separation%3F,-Let's%20say%20we&text=Speech%20separation%20is%20also%20called,of%20sounds%20as%20speech%20enhancement.

Speaker Diarization

In most **real-world scenarios** speech does not come in well defined audio segments with only one speaker. In most of the conversations that our algorithms will need to work with, **people will interrupt each other** and cutting the audio between sentences won't be a trivial task.

In many applications we will want to **identify multiple speakers in a conversation**, for example when writing a protocol of a meeting. For such occasions, identifying the different speakers and connect different sentences under the same speaker is a critical task.
It can be described as the question **"who spoke when?"** in an audio segment.

```
SPEAKER rec1 0 86.200 16.400 <NA> <NA> 1 <NA> <NA>`
SPEAKER rec1 0 103.050 5.830 <NA> <NA> 1 <NA> <NA>`
SPEAKER rec1 0 109.230 4.270 <NA> <NA> 1 <NA> <NA>`
SPEAKER rec1 0 113.760 8.625 <NA> <NA> 1 <NA> <NA>`
SPEAKER rec2 0 122.385 4.525 <NA> <NA> 2 <NA> <NA>`
SPEAKER rec2 0 127.230 6.230 <NA> <NA> 2 <NA> <NA>`
SPEAKER rec2 0 133.820 0.850 <NA> <NA> 2 <NA> <NA>`
```

Introduction

- **WavLM** proves the potential of pre-trained models on full-stack speech tasks by using the weighted sum of embeddings from different layers.

- They find different layers contain information useful for different tasks.

For instance, the hidden states of the top layers are useful for ASR, while the bottom layers are more effective for speaker verification.

Drawbacks in existing pre-trained models

1. Current pre-trained models are <u>unsatisfactory for multi-speaker tasks</u>

- **speech separation models** trained on top of **HuBERT**, a top performed speech pre-trained model, achieve only <u>marginal improvement</u> compared with the models <u>trained from scratch</u>.

- because the pre-training methods **do not sufficiently enforce the speaker discrimination**, and the **training data** contain only **single-speaker audios**.

2. Speech pretraining crucially relies on high quality and large quantities of unlabeled audios.

- audiobook **data mismatches** the data in a **real scenario** and using it exclusively hurts the model performance <u>when the acoustic characteristics of the downstream tasks are different</u> from those of the audiobook.

- To **eliminate** the audiobook data bias, we try to **gather data from different sources** as much as possible in our experiments.

In this paper:

◦The contribution of the paper can be summarized as follows:

1)WavLM sheds light on a **general pre-trained model** for **full stack speech processing tasks**, in contrast to the previous SSL works focusing on a group of similar tasks.

2)We propose simple but effective modifications to the existing pre-trained models, which show general and consistent improvements across downstream tasks.

3)We scale-up self-supervised speech pre-training with more **unlabeled data and longer training steps.**

4)We achieve **state-of-the- art results on the SUPERB benchmark**, and significantly boost the performance for various speech processing tasks on their representative benchmarks, including speech separation, speaker verification, and speaker diarization. The models and code are released to facilitate future research.

Related Work

Based on the training objective, SSL methods can be categorized into:

•Generative learning

traced back to the **auto-encoding model**, which reconstructs the whole speech from **latent variables**, either continuous or discrete.

predict future frames from the history with an autoregressive model or

recover the masked frames from the corrupted speech with a non-autoregressive model

•Discriminative learning

well-known examples : **CPC , wav2vec , vq-wav2vec , wav2vec 2.0 , DiscreteBERT , HuBERT**

CPC and the wav2vec series models use the contrastive InfoNCE loss to discriminate the correlated positive samples from negative samples

•Multitask learning

is adopted in **PASE** and **PASE+**

They employ lots of pre-training objectives such as waveform generation, prosody regression and contrastive objectives.

Related Work

UniSpeech combines self-supervised learning and supervised learning for ASR, and shows impressive results on multi-lingual testsets.

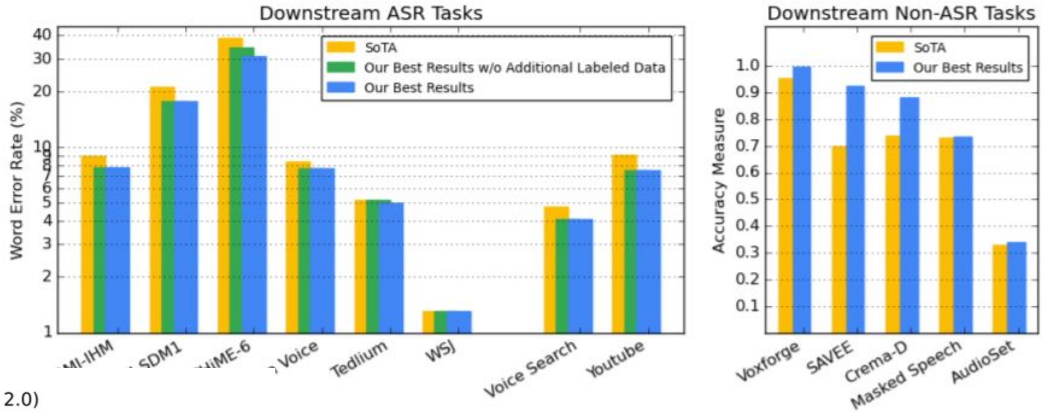
Unlike SSL in computer vision (CV) and NLP fields, where one pre-trained model is adapted to various downstream tasks, most speech SSL methods **focus on phoneme classification and ASR**.

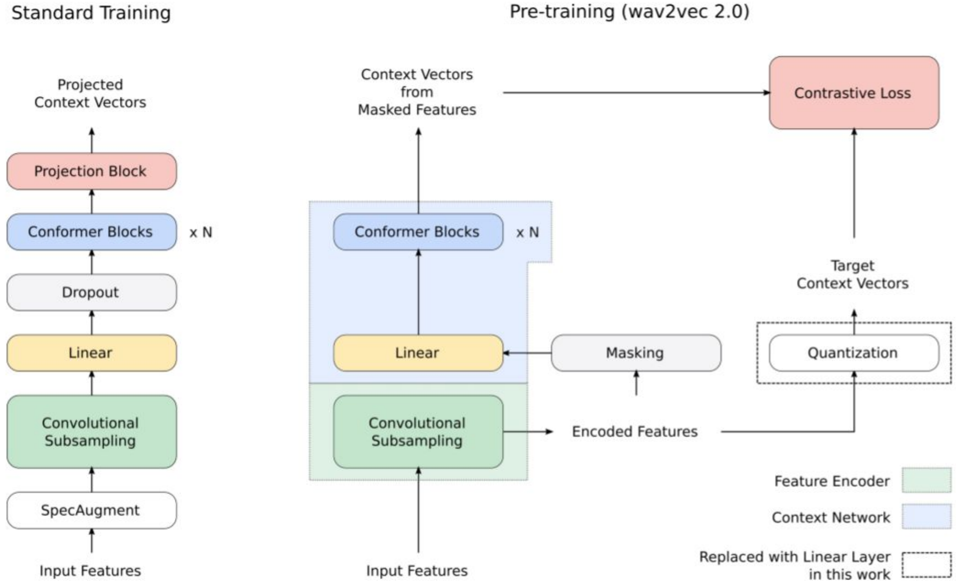
According to the results, HuBERT enjoys the best generalization ability in the overall evaluation. To better learn speaker characteristics, proposed UniSpeech-SAT, which extends the HuBERT framework with speaker aware pre-training. It significantly outperforms other pre-trained models on the speaker-relate tasks with a slight degradation on the ASR.

a concurrent work BigSSL also mentions **large self-supervised learning model** could handle various speech tasks.

The **difference** is that our work demonstrates that the full stack tasks can be handled by the careful pre-training and fine-tuning strategy design, even without scaling-up the model size to **8 billion parameters.**

Related Work (BigSSL)





Background : HUBERT

- HuBERT is an SSL method which benefits from an offline clustering step to provide target labels for a BERT-like prediction loss (Devlin et al., 2019).
- The backbone is a Transformer encoder (Vaswani et al., 2017) with **L** blocks.
- During pre-training, the Transformer consumes masked acoustic features $\tilde{\mathbf{x}}$ and output hidden states h^L.
- The network is optimized to predict the discrete target sequence **z**, where each z_t $\in$ [C] is a C-class categorical variable.

Background : HUBERT

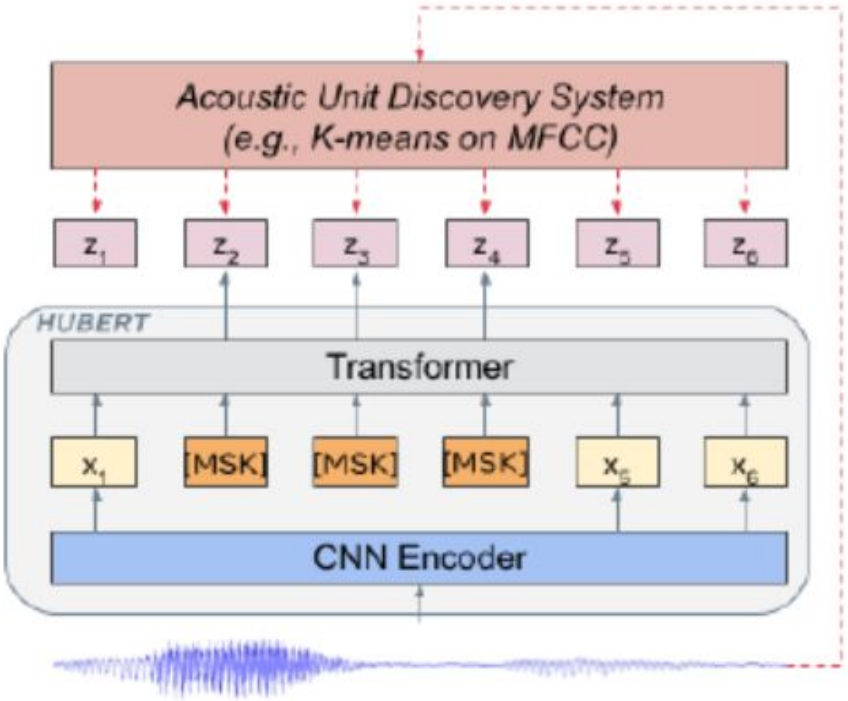


Fig. 1: The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.

Background : HUBERT

- The distribution over codewords is parameterized with

$$p(c|\mathbf{h}_t) = \frac{\exp\left(sim(\mathbf{h}_t^L \mathbf{W}^P, \mathbf{e}_c)/\tau\right)}{\sum_{c'=1}^{C} \exp\left(sim(\mathbf{h}_t^L \mathbf{W}^P, \mathbf{e}_{c'})/\tau\right)}$$

- where $\mathbf{W}^P$ is a projection matrix, $\mathbf{h}_t^L$ is the output hidden state for step t, $\mathbf{e}_c$ is the embedding for codeword c, $\mathrm{sim}(a, b)$ computes the cosine similarity and $\tau = 0.1$ scales the logit.
- A key ingredient of HuBERT is that the prediction loss is only applied over the masked regions, forcing the model to learn a combined acoustic and language model over the continuous inputs.

Background : HUBERT

HuBERT adopts an iterative re-clustering and re-training process:

- For the first iteration, the targets are assigned by clustering the MFCC features of the training data
- For the second iteration, a new generation of training targets are created by clustering the latent representations generated by the first iteration trained model.

WavLM : Model Structure

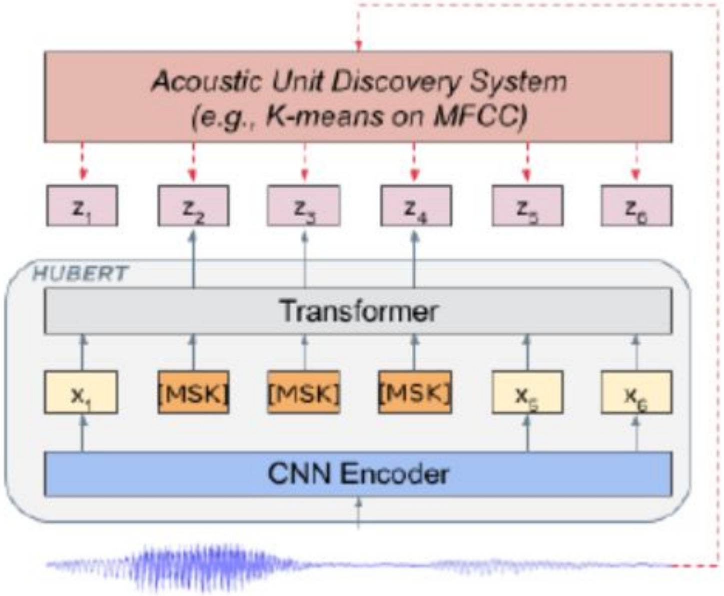
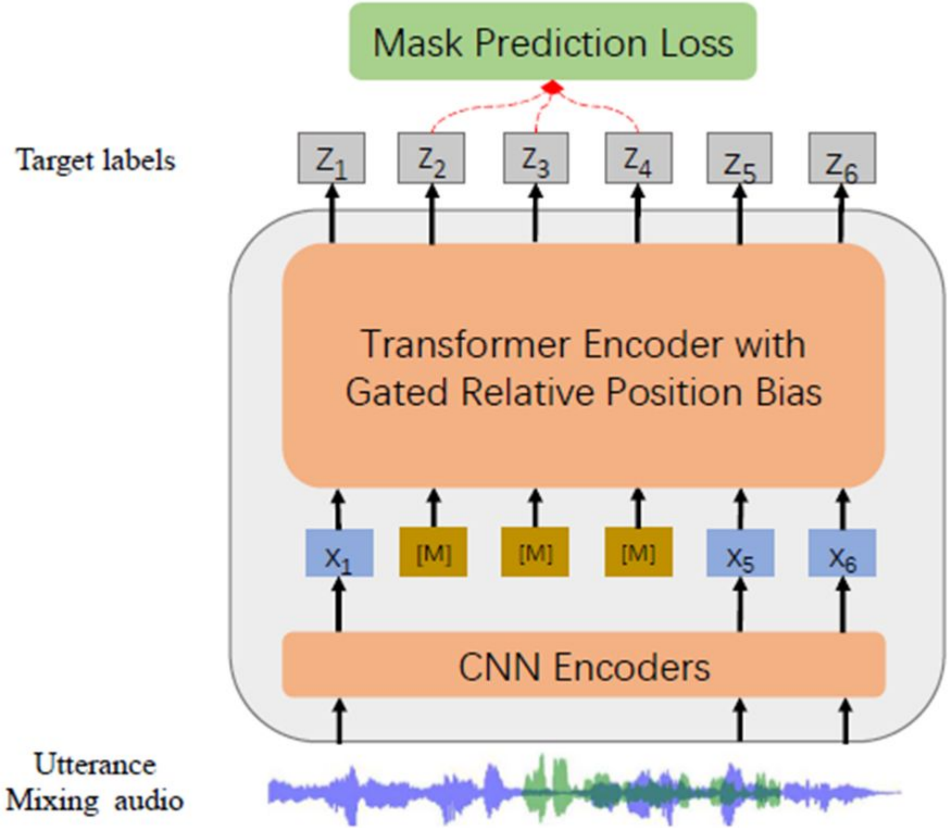


Fig. 1: The HuBERT approach predicts hidden cluster assignments of the masked frames (y_2, y_3, y_4 in the figure) generated by one or more iterations of k-means clustering.



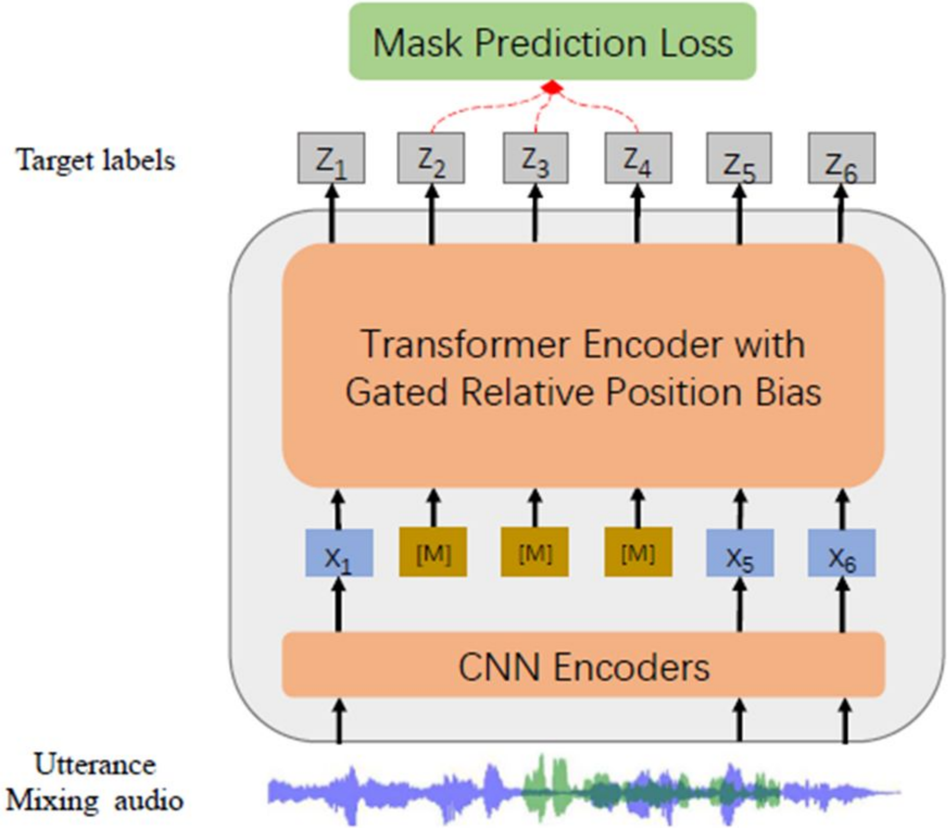
WavLM : Model Structure

- ## Convolutional feature encoder

The convolutional encoder is composed of **seven blocks** of **temporal convolution** followed by **layer normalization** and a **GELU activation layer**. The temporal convolutions have **512 channels** with **strides** (5,2,2,2,2,2,2) and **kernel widths** (10,3,3,3,3,2,2), resulting in each output representing about **25ms of audio** strided by 20ms.

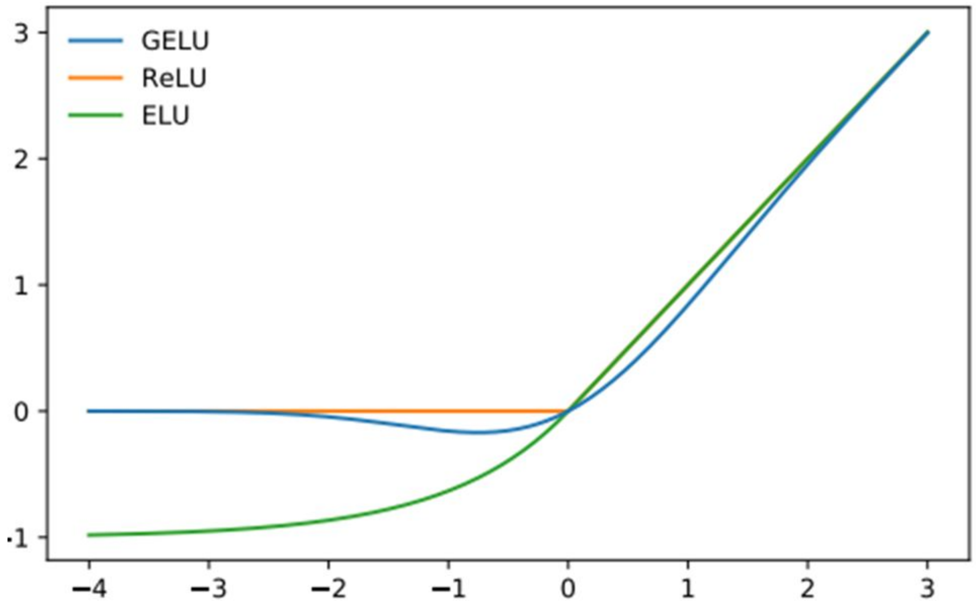
- ## Transformer encoder

The convolutional output representations **x** is **masked** as the Transformer input. The Transformer is equipped with a **convolution based relative position embedding layer** with kernel size 128 and 16 groups at the bottom.



WavLM GELU activation

Gaussian Error Linear Unit



$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x)$$
$$\approx 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right)$$

WavLM : Model Structure

- Following HuBERT, we also use the mask prediction loss to optimize our network as:

$$\mathcal{L} = -\sum_{l \in K} \sum_{t \in M} \log p\,(z_t | \mathbf{h}_t)$$

- where M denotes the set of masked indices in time domain and $\mathbf{h}_t^L$ is the Transformer output for step t.
- The model is trained for two or three iterations depending on the data size.
- For the first iteration, we run k-means clustering on the MFCC features of the training data to obtain the training targets.
- For the second and third iterations, we run k-means on the latent representations generated by the previous iteration model to get the new pre-training targets.

WavLM : gated relative position bias

gated relative position bias vs **convolutional relative position embedding**

WavLM **vs** **Hubert , Wav2vec2**

- To improve the model, we employ **gated relative position bias** (Chi et al., 2021) which is encoded based on the offset between the "**key" and "query" in the Transformer self-attention** mechanism.
- Let $\{h_i\}_{i=1}^{T}$ denote the input hidden states for the self-attention module, each h_i is linearly projected to a triple of query, key and value (q_i, k_i, v_i) as:

$$q_i, k_i, v_i = h_i W^Q, h_i W^K, h_i W^y$$

WavLM : gated relative position bias

- The self-attention $\{\tilde{h}\}_{i=1}^{T}$ outputs are computed via:

$$a_{ij} \propto \exp\{\frac{q_i \cdot k_j}{\sqrt{d_k}} + r_{i-j}\}$$

$$\tilde{h} = \sum_{j=1}^{T} a_{ij} v_j$$

- where r_{i-j} is the gated relative position bias added to the attention logits.
- It is computed by:

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

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

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

the same distance offset between two frames tends to play different roles if one frame is silence while the other belongs to a speech segment.

WavLM : Masked Speech Denoising and Prediction

- We propose masked **speech denoising and prediction framework** to improve model robustness for complex acoustic environments and the preservation of speaker identity

- propose a **masked speech denoising and prediction framework**, where some inputs are simulated noisy/overlapped with masks and the target is to predict pseudo-labels of the origin speech on masked region.

- Unlike existing masked speech modeling (HuBERT), which just focuses on the ASR task, the masked speech denoising allows us to extend pre-trained speech models to non-ASR tasks, since it implicitly models information we need in the speaker identification, separation, and diarization tasks.

- further optimize the Transformer backbone and extend pre-training data to 94k publish English data.

WavLM : Utterance Mixing

Algorithm 1 Utterance Mixing

given a batch of speech utterances $\mathbf{U} = \{\mathbf{u}^i\}_{i=1}^{B}$ with batch size B and length L, mixing probability p
Choose S utterances $\mathbf{U}^S \subset \mathbf{U}$ by Bernoulli sampling with probability p
for each primary utterance $\mathbf{u}^{\text{pri}} \in \mathbf{U}^S$ **do**
    Sample a secondary utterance $\mathbf{u}^{\text{sec}}$ from discrete uniform distribution with probability $P(\mathbf{u}^{\text{sec}} = \mathbf{x}) = \frac{1}{B}, \mathbf{x} \in \mathbf{U}$
    Sample the mix length l from discrete uniform distribution with probability $P(l = x) = \frac{2}{L}, x \in \{1, \cdots, \frac{L}{2}\}$
    Sample a start position s^{pri} of $\mathbf{u}^{\text{pri}}$ from discrete uniform distribution with probability $P(s^{\text{pri}} = x) = \frac{1}{L-l}, x \in \{1, \cdots, L-l\}$
    Sample a start position s^{sec} of $\mathbf{u}^{\text{sec}}$ from discrete uniform distribution with probability $P(s^{\text{sec}} = x) = \frac{1}{L-l}, x \in \{1, \cdots, L-l\}$
    Sample the mixing energy ratio r from the continuous uniform distribution $\mathcal{U}(-5, 5)$
    Calculate the energy of the primary utterance $E^{\text{pri}} \leftarrow \frac{\sum \mathbf{u}^{\text{pri}} \cdot \mathbf{u}^{\text{pri}}}{L}$
    Calculate the energy of the secondary utterance $E^{\text{sec}} \leftarrow \frac{\sum \mathbf{u}^{\text{sec}} \cdot \mathbf{u}^{\text{sec}}}{L}$
    Calculate the mixing scale $scl \leftarrow \sqrt{\frac{E^{\text{pri}}}{10^{\frac{r}{10}} E^{\text{sec}}}}$
    $\mathbf{u}^{\text{pri}}[s^{\text{pri}} : s^{\text{pri}} + l] \leftarrow \mathbf{u}^{\text{pri}}[s^{\text{pri}} : s^{\text{pri}} + l] + scl \cdot \mathbf{u}^{\text{sec}}[s^{\text{sec}} : s^{\text{sec}} + l]$
return U

WavLM : Utterance Mixing

- Introduce utterance mixing to improve the multi-speaker information modeling in pre-training. The utterance mixing method aims to simulate the multi-speaker speech for self-supervised pre-training when only single-speaker pre training data are available.

- To generate the overlapped speech for pre-training, we randomly select multiple utterances from each training batch, and mix each of them with another secondary utterance at a random region.

- The secondary utterance is randomly selected from the same batch, randomly cropped and scaled by a random source energy ratio. We ensure that the overlap region is less than 50% and refer the speaker from the first utterance as main speaker.

- With the utterance mixing method, the model is trained to predict the content information corresponding to the **main speaker** with the **mask prediction loss** .

WavLM : Pre-Training Data

- We leverage large-scale unsupervised data from diverse domains to improve the robustness of our model.
- Previous works use LibriSpeech (Panayotov et al., 2015) or Libri-Light (Kahn et al., 2020) datasets for pre-training, which limits the generalization capability of the pre-trained model since the input data are all extracted from the audiobook.
- The background acoustics of the speech obtained from the audiobook is different from what is observed in real scenarios, since the real captured sounds are usually accompanied by various types of noise.

Experiment: Setup

- We first evaluate our models on **SUPERB**, which is designed to provide a standard and comprehensive testbed for pre-trained models on various speech tasks.

- It covers 10 tasks

- These tasks can be grouped into four aspects of speech:
    - **Content**
    - **Speaker**
    - **Semantics**
    - **paralinguistics**

Experiment: Universal Representation Evaluation

WavLM Base

parameters : 94.70M

corpus: LS 960 hr

WavLM Base+

parameters: 94.70M

corpus: Mix 94k hr

larger and **more diverse pre-training data.**

WavLM Large

parameters: 316.62M

corpus: Mix 94k hr

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

1) We use the same downstream models as the SUPERB implementations for each downstream task

2) Pre-trained models are *frozen* to limit the space of the fine-tuning hyperparameter search

3) The downstream models consume the weighted sum results of the hidden states extracted from each layer of the pre-trained model.

The **overall score** is computed by ourselves: we multiply the QbE score with 100, replace each error rate score with (1 - error rate), and average the scores of all tasks.

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

10 task

TABLE I

UNIVERSAL SPEECH REPRESENTATION EVALUATION ON SUPERB BENCHMARK. PARAL DENOTE PARALINGUISTICS ASPECT OF SPEECH.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ [37]	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC [25]	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC [24]	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC [28]	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay [30]	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA [29]	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC [44]	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec [33]	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec [34]	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base [5]	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base [6]	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.55	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o denoising task	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	89.42	4.07	3.50	3.92	5.59	97.37	**0.0988**	99.00	90.58	21.20	68.65	83.3
wav2vec 2.0 Large [5]	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large [6]	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.49**	**3.77**	**3.24**	**3.06**	**3.44**	**97.86**	0.0886	**99.31**	**92.21**	**18.36**	**70.62**	**84.8**

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

It is a **fair comparison** as the **three models** use the **same amount of pre-training data** and **the number of parameters.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			[illegible]	[illegible]	[illegible]	[illegible]	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 202[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	[illegible]	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**	**84.6**

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	[illegible]	[illegible]	[illegible]	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	81.7
WavLM Base+	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	80.8
HuBERT Large (Hsu et al., 2021a)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	82.2
WavLM Large	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	84.6

WavLM Base performs better than **wav2vec 2.0 Base** and **HuBERT** Base on all downstream tasks.

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	[illegible]	[illegible]	[illegible]	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification														
WavLM Base+														
wav2vec 2.0 Large (Baevski et al., 2020b)														
HuBERT Large (Hsu et al., 2021a)														
WavLM Large														

WavLM Base performs better than **wav2vec 2.0 Base** and **HuBERT** Base on all downstream tasks.

indicate the effectiveness of our **structure** and the **masked speech denoising modeling**

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	[illegible]	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Base+	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
wav2vec 2.0 Large (Baevski et al., 2020b)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HuBERT Large (Hsu et al., 2021a)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Large	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

the most <u>impressive result</u> is **speaker diarization**, where the **WavLM Base** outperforms **HuBERT Base** by **22.6% relatively**

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	[illegible]	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Base+	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
wav2vec 2.0 Large (Baevski et al., 2020b)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HuBERT Large (Hsu et al., 2021a)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Large	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

explanation is that the **additional overlapped speech** forces the model to deal with **multi-speaker signals** during pre-training.

Ablation study to remove utterance mixing

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61		5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84				5.64	96.69	0.0990	99.16	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38													
HuBERT Large (Hsu et al., 2021a)	316.61													
WavLM Large	316.62													

The performance of “**w/o utterance mixing**” drops significantly for the **speaker diarization task**.

Ablation study to remove utterance mixing

Phoneme Recognition (PR)

Automatic Speech Recognition (ASR)

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	[illegible]	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HuBERT Large (Hsu et al., 2021a)	316.61M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Large	316.62M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

in the "**w/o structure modification**" setting, performance degradation can be witnessed especially for **PR** and **ASR** tasks.

Ablation study to remove <u>utterance mixing</u>

Phoneme Recognition (PR)

Automatic Speech Recognition (ASR)

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker SID Acc ↑	Speaker ASV EER ↓	Speaker SD DER ↓	Content PR PER ↓	Content ASR WER ↓	Content KS Acc ↑	Content QbE MTWV ↑	Semantics IC Acc ↑	Semantics SF F1 ↑	Semantics SF CER ↓	ParaL ER Acc ↑	Overall Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	[illegible]	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HuBERT Large (Hsu et al., 2021a)	316.61M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Large	316.62M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

It indicates that the ***gated relative position bias*** contributes to the **performance improvement** of the **content related tasks**

Ablation study to remove structure modification

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	[illegible]	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	[illegible]	[illegible]	[illegible]	67.62	82.2
WavLM Large	316.62M	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

WavLM performs very well on **semantic** and **paralinguistics** tasks as well

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75			0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT													67.62	82.2
WavLM													**70.03**	**84.6**

demonstrating our **model** is **general for the full stack speech processing tasks.**

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	[illegible]	[illegible]	[illegible]
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	[illegible]	[illegible]
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	[illegible]	[illegible]
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	[illegible]
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**	**84.6**

WavLM Base+ outperforms the **wav2vec 2.0 Large** and **HuBERT Large** in the overall score.

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	[illegible]	[illegible]	[illegible]
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	[illegible]	[illegible]
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	[illegible]	[illegible]
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	[illegible]
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**	**84.6**

indicates that the **960h data** are **insufficient** to fulfill the capacity of the Base model.

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
WavLM Base	94.70M	LS 960 hr	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	[illegible]	[illegible]	[illegible]	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	[illegible]	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**	**84.6**

Most tasks benefit from the **larger model size**, especially for the **ASR.**
We obtain **38% word error rate** reduction on the ASR by model scaling-up.

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Speaker Identification (SID)

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	64.92	80.8
WavLM Base	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	65.94	81.9
- w/o utterance mixing	94.70M	[illegible]	[illegible]	[illegible]	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	[illegible]	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**	**84.6**

there is **6.07% absolute improvement** on the **SID task**, indicating the ***large model size also impacts the speaker related tasks.***

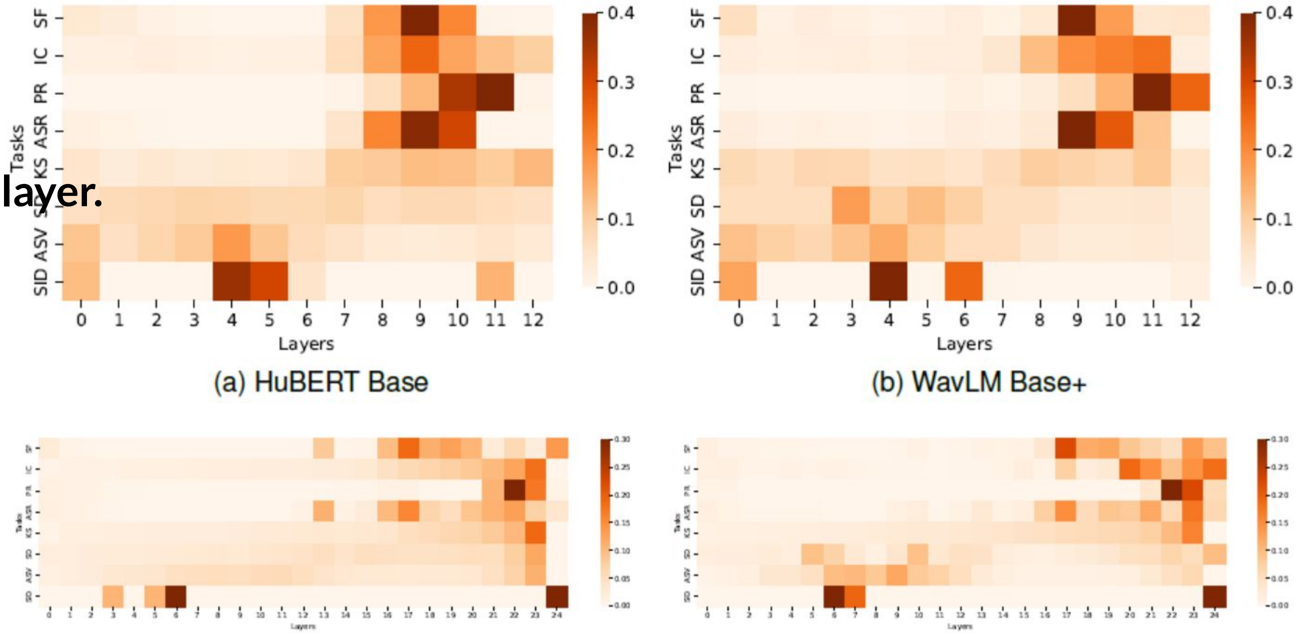
Weight analysis on the **SUPERB Benchmark**

weighted-sum the hidden states of different layers, and feed it to the **task specific layer**

the <u>weights of different layers </u>of HuBERT and WavLM models
<u> on the different downstream tasks</u>
of <u>SUPERB benchmark.</u>

A larger weight indicates

a larger contribution of the layer.



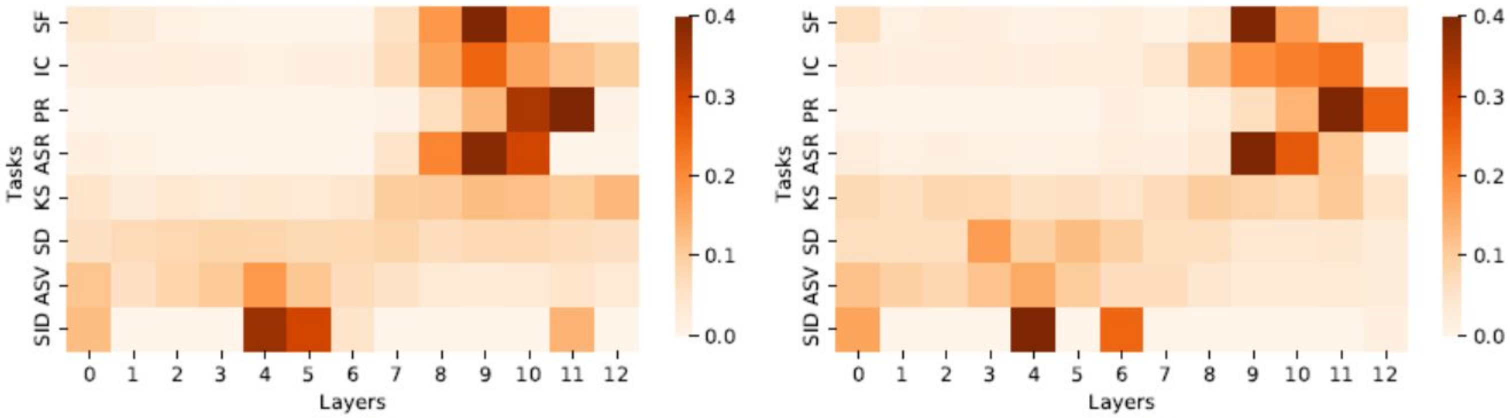
(a) HuBERT Base

(b) WavLM Base+

(c) HuBERT Large

(d) WavLM Large

Weight analysis on the **SUPERB Benchmark**



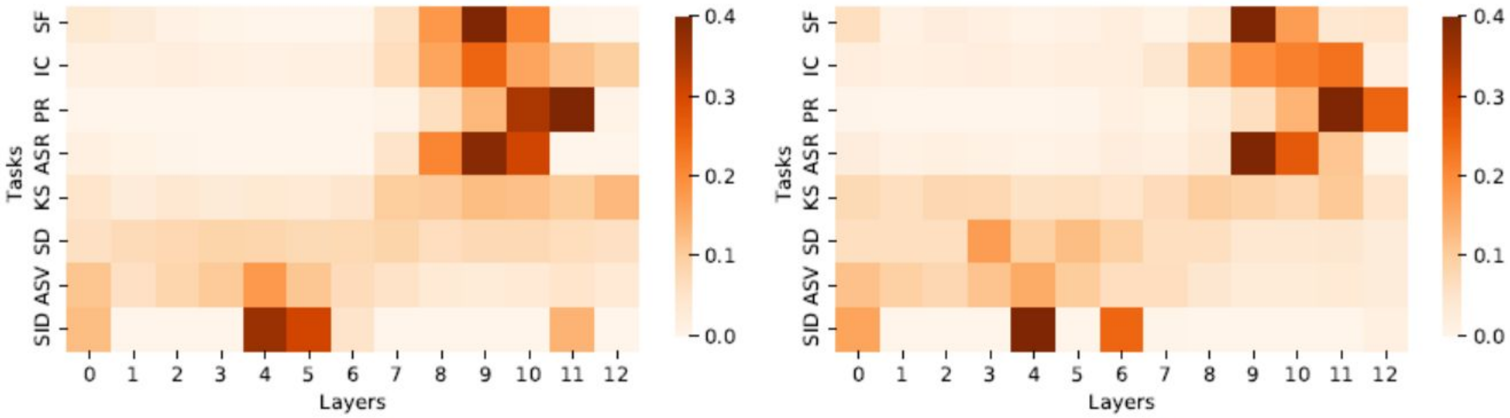
(a) HuBERT Base (b) WavLM Base+

for the Base models, the **contribution patterns of different layers** are **similar** between WavLM and HuBERT

observe that the **bottom layers** contribute **more to speaker related tasks**, such as speaker identification, automatic speaker verification and speaker diarization.

for the automatic speech recognition, phoneme recognition, intent classification and slot filling tasks, **the top layers are more important**

Weight analysis on the **SUPERB Benchmark**



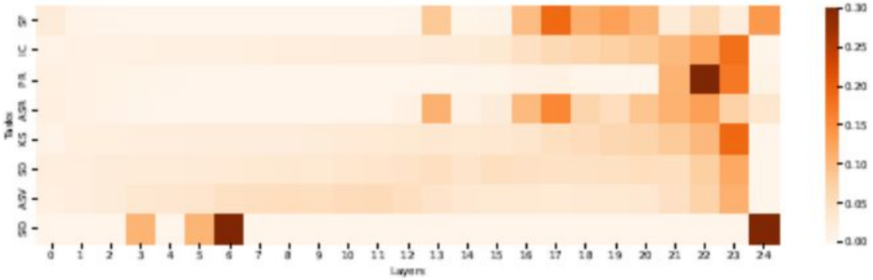
(a) HuBERT Base (b) WavLM Base+

Base models learn speaker information with the **bottom layers**

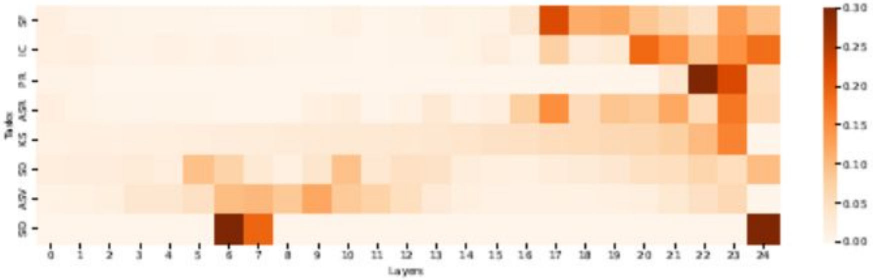
while

the content and semantic information are **encoded in the top layers.**

Weight analysis on the **SUPERB Benchmark**



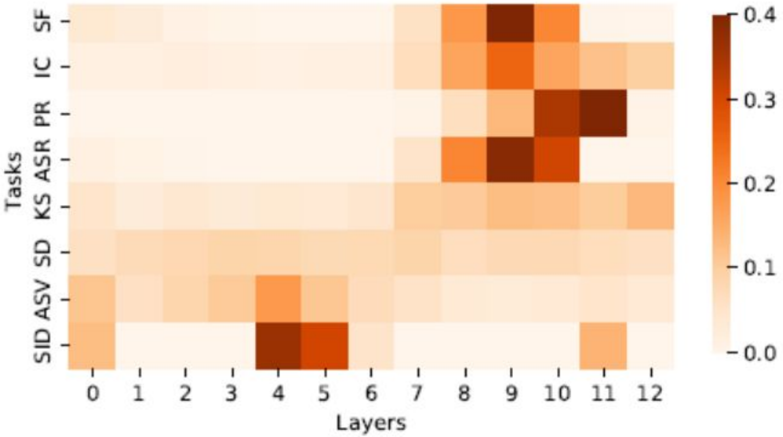
(c) HuBERT Large



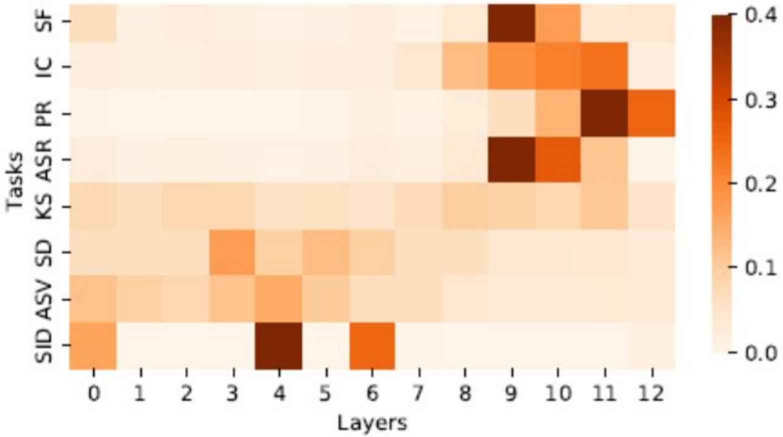
(d) WavLM Large

The model behaviour is similar to the Large models.

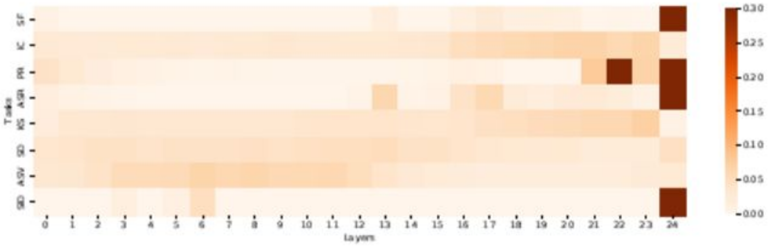
Weight analysis on the **SUPERB Benchmark**



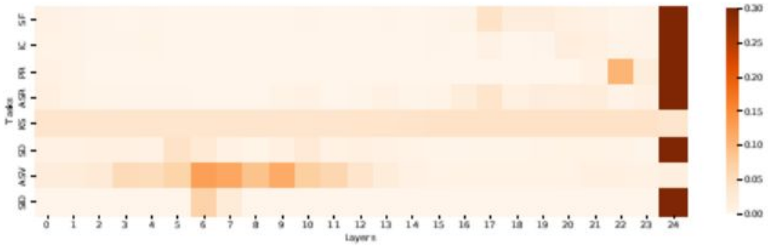
(a) HuBERT Base



(b) WavLM Base+



(c) HuBERT Large



(d) WavLM Large

END

Thanks for your attention …

UNIVERSAL SPEECH REPRESENTATION **EVALUATION** ON **SUPERB BENCHMARK.**

Table 1. Universal speech representation evaluation on SUPERB benchmark. ParaL denote Paralinguistics aspect of speech.

Method	#Params	Corpus	Speaker			Content				Semantics			ParaL	Overall
			SID	ASV	SD	PR	ASR	KS	QbE	IC	SF		ER	
			Acc ↑	EER ↓	DER ↓	PER ↓	WER ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	0	-	8.5E-4	9.56	10.05	82.01	23.18	8.63	0.0058	9.10	69.64	52.94	35.39	40.5
PASE+ (Ravanelli et al., 2020)	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	82.54	0.0072	29.82	62.14	60.17	57.86	55.1
APC (Chung et al., 2019)	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	91.01	0.0310	74.69	70.46	50.89	59.33	66.0
VQ-APC (Chung et al., 2020)	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	91.11	0.0251	74.48	68.53	52.91	59.66	65.6
NPC (Liu et al., 2020a)	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	88.96	0.0246	69.44	72.79	48.44	59.08	65.2
Mockingjay (Liu et al., 2020c)	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	83.67	6.6E-04	34.33	61.59	58.89	50.28	53.5
TERA (Liu et al., 2020b)	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	89.48	0.0013	58.42	67.50	54.17	56.27	62.0
modified CPC (Rivière et al., 2020)	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	91.88	0.0326	64.09	71.19	49.91	60.96	63.2
wav2vec (Schneider et al., 2019)	32.54M	LS 960 hr	56.56	7.99	9.9	31.58	15.86	95.59	0.0485	84.92	76.37	43.71	59.79	69.9
vq-wav2vec (Baevski et al., 2020a)	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	93.38	0.0410	85.68	77.68	41.54	58.24	67.7
wav2vec 2.0 Base (Baevski et al., 2020b)	95.04M	LS 960 hr	75.18	6.02	6.08	5.74	6.43	96.23	0.0233	92.35	88.30	24.77	63.43	79.0
HuBERT Base (Hsu et al., 2021a)	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	6.42	96.30	0.0736	98.34	88.53	25.20	64.92	80.8
WavLM Base	94.70M	LS 960 hr	84.51	4.69	4.83	4.84	6.21	96.79	0.0870	98.63	89.38	22.86	65.94	81.9
- w/o utterance mixing	94.70M	LS 960 hr	84.39	4.91	6.03	4.85	6.08	96.79	0.0799	98.42	88.69	23.43	65.55	81.5
- w/o structure modification	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	96.79	0.0956	98.31	88.56	24.00	65.60	81.7
WavLM Base+	94.70M	Mix 94k hr	86.84	4.26	4.07	4.07	5.64	96.69	**0.0990**	**99.16**	89.73	21.54	67.98	82.8
wav2vec 2.0 Large (Baevski et al., 2020b)	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	96.66	0.0489	95.28	87.11	27.31	65.64	80.8
HuBERT Large (Hsu et al., 2021a)	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	3.62	95.29	0.0353	98.76	89.81	21.76	67.62	82.2
WavLM Large	316.62M	Mix 94k hr	**95.25**	**4.04**	**3.47**	**3.09**	**3.51**	**97.40**	0.0827	99.10	**92.25**	**17.61**	**70.03**	**84.6**

Speaker Verification

Speaker Verification

Finetune the model with VoxCeleb2 dev data, and evaluate it on the VoxCeleb1

Model	Fix pre-train	Vox1-O	Vox1-E	Vox1-H
ECAPA-TDNN	-	0.87	1.12	2.12
HuBERT large	Yes	0.888	0.912	1.853
Wav2Vec2.0 (XLSR)	Yes	0.915	0.945	1.895
UniSpeech-SAT large	Yes	0.771	0.781	1.669
WavLM large	Yes	0.59	0.65	1.328
WavLM large	No	0.505	0.579	1.176
+Large Margin Finetune and Score Calibration				
HuBERT large	No	0.585	0.654	1.342
Wav2Vec2.0 (XLSR)	No	0.564	0.605	1.23
UniSpeech-SAT large	No	0.564	0.561	1.23
WavLM large (New)	No	**0.33**	**0.477**	**0.984**

Speaker Diarization

Speaker Diarization

Evaluation on the CALLHOME

Model	spk_2	spk_3	spk_4	spk_5	spk_6	spk_all
EEND-vector clustering	7.96	11.93	16.38	21.21	23.1	12.49
EEND-EDA clustering (SOTA)	7.11	11.88	14.37	25.95	21.95	11.84
HuBERT base	7.93	12.07	15.21	19.59	23.32	12.63
HuBERT large	7.39	11.97	15.76	19.82	22.10	12.40
UniSpeech-SAT large	5.93	10.66	12.9	16.48	23.25	10.92
WavLM Base	6.99	11.12	15.20	16.48	21.61	11.75
WavLm large	6.46	10.69	11.84	12.89	20.70	10.35

Speech Recognition

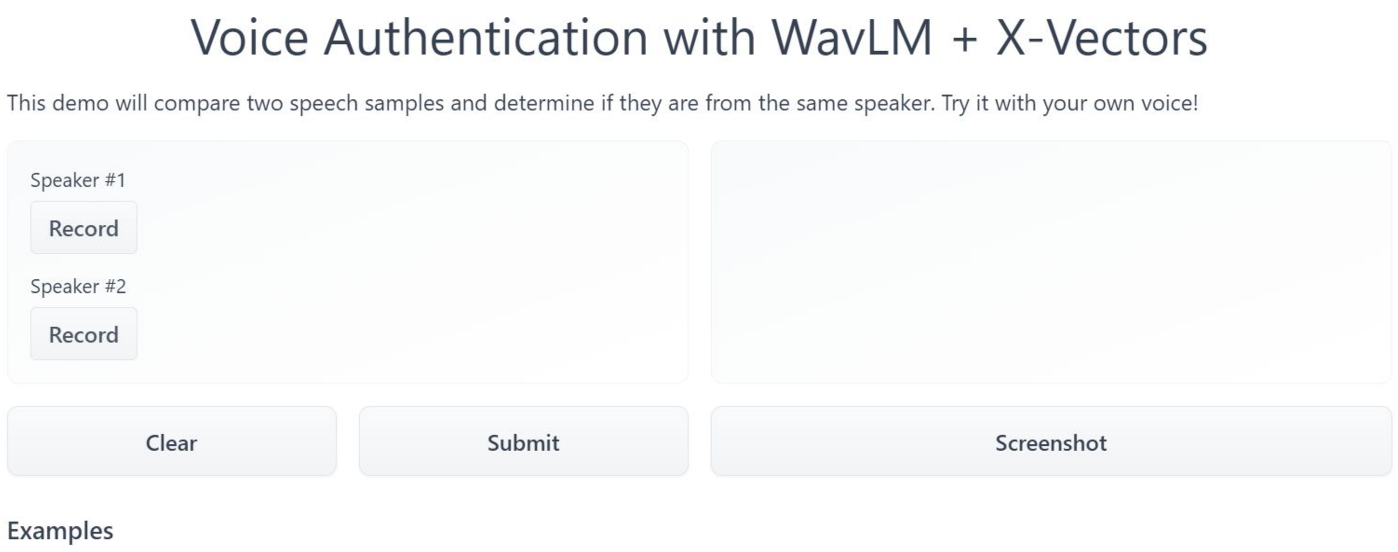
Speech Recogntion

Evaluate on the LibriSpeech

Model	Unlabled Data	LM	test-clean	test-other
1-hour labeled				
wav2vec 2.0 Base	LS-960	None	24.5	29.7
WavLM Base	LS-960	None	24.5	29.2
WavLM Base+	MIX-94K	None	22.8	26.7
DeCoAR 2.0	LS-960	4-gram	13.8	29.1
DiscreteBERT	LS-960	4-gram	9.0	17.6
wav2vec 2.0 Base	LS-960	4-gram	5.5	11.3
HuBERT Base	LS-960	4-gram	6.1	11.3
WavLM Base	LS-960	4-gram	5.7	10.8
WavLM Base+	MIX-94K	4-gram	5.4	9.8
wav2vec 2.0 Large	LL-60K	4-gram	3.8	7.1
WavLM Large	MIX-94K	4-gram	3.8	6.6
wav2vec2.0 Large	LL-60K	Transformer	2.9	5.8
HuBERT Large	LL-60K	Transformer	2.9	5.4
WavLM Large	MIX-94K	Transformer	2.9	5.1
10-hour labeled				
wav2vec 2.0	LS-960	None	11.1	17.6
WavLM Base	LS-960	None	9.8	16.0
WavLM Base+	MIX-94K	None	9.0	14.7
DeCoAR 2.0	LS-960	4-gram	5.4	13.3
DiscreteBERT	LS-960	4-gram	5.9	14.1
wav2vec 2.0	LS-960	4-gram	4.3	9.5
HuBERT Base	LS-960	4-gram	4.3	9.4
WavLM Base	LS-960	4-gram	4.3	9.2
WavLM Base+	MIX-94K	4-gram	4.2	8.8
wav2vec 2.0 Large	LL-60K	4-gram	3.0	5.8
WavLM Large	MIX-94K	4-gram	2.9	5.5
wav2vec 2.0 Large	LL-60K	Transformer	2.6	4.9
HuBERT Large	LL-60K	Transformer	2.4	4.6
WavLM Large	MIX-94K	Transformer	2.4	4.6
100-hour labeled				
wav2vec 2.0 Base	LS-960	None	6.1	13.3
WavLM Base	LS-960	None	5.7	12.0
WavLM Base+	MIX-94K	None	4.6	10.1
DeCoAR 2.0	LS-960	4-gram	5.0	12.1
DiscreteBERT	LS-960	4-gram	4.5	12.1
wav2vec 2.0 Base	LS-960	4-gram	3.4	8.0
HuBERT Base	LS-960	4-gram	3.4	8.1
WavLM Base	LS-960	4-gram	3.4	7.7
WavLM Base+	MIX-94K	4-gram	2.9	6.8
wav2vec 2.0 Large	LL-60K	4-gram	2.3	4.6
WavLM Large	MIX-94K	4-gram	2.3	4.6
wav2vec 2.0 Large	LL-60K	Transformer	2.0	4.0
HuBERT Large	LL-60K	Transformer	2.1	3.9
WavLM Large	MIX-94K	Transformer	2.1	4.0

Using

https://huggingface.co/spaces/microsoft/wavlm-speaker-verification



Pytorch code example

https://huggingface.co/docs/transformers/model_doc/wavlm

Implementation

https://colab.research.google.com/drive/1dHUzIHqh8vMBUt95bK-QzGn7025MZYOK?usp=sharing

https://colab.research.google.com/drive/1hz2oWHX0muHcTIM9p6O98ls0mmXwLNSz#scrollTo=HtDyCzD1Ez-b
this one is simpler

PreTrained models:

https://github.com/microsoft/unilm/tree/master/wavlm

Pre-Trained Models

Model	Pre-training Dataset	Fine-tuning Dataset	Model
WavLM Base	960 hrs LibriSpeech	-	Azure Storage Google Drive
WavLM Base+	60k hrs Libri-Light + 10k hrs GigaSpeech + 24k hrs VoxPopuli	-	Azure Storage Google Drive
WavLM Large	60k hrs Libri-Light + 10k hrs GigaSpeech + 24k hrs VoxPopuli	-	Azure Storage Google Drive

Thanks for your attention …

Background : HUBERT

- HuBERT is an SSL method which benefits from an offline clustering step to provide target labels for a BERT-like prediction loss (Devlin et al., 2019).
- The backbone is a Transformer encoder (Vaswani et al., 2017) with $\boldsymbol{L}$ blocks.
- During pre-training, the Transformer consumes masked acoustic features $\tilde{\mathbf{x}}$ and output hidden states h^L.
- The network is optimized to predict the discrete target sequence $\mathbf{z}$, where each $z_t \in [C]$ is a C-class categorical variable.

Introduction

- Building a general pre-trained model can be essential to the further development of speech processing, because it can utilize large-scale unlabeled data to boost the performance in downstream tasks, reducing data labeling efforts.

- In the past, it has been infeasible to build such a general model, as different tasks focus on different aspects of speech signals.

- Speaker Verification

speaker characteristic

~~Spoken content~~

- Speech Recognition

~~speaker characteristic~~

Spoken content