IEEE ICASSP 2022

UniSpeech-SAT: Universal Speech Representation Learning with Speaker Aware Pre-Training

Poster #2884

Sanyuan Chen[1], Yu Wu[2], Chengyi Wang[2], Zhengyang Chen[2], Zhuo Chen[2], Shujie Liu[2], Jian Wu[2], Yao Qian[2], Furu Wei[2], Jinyu Li[2], Xiangzhan Yu[1]

[1]Harbin Institute of Technology, [2]Microsoft Corporation

Demo: https://huggingface.co/spaces/microsoft/unispeech-speaker-verification

Code: https://github.com/microsoft/UniSpeech

Self-Supervised Learning

Self-supervised learning (SSL) has achieved great successes in NLP and CV, especially for limited resource tasks.

It is because SSL utilizes **a large amount of unlabeled data** to learn universal representation, and the **universal representation** with outstanding generalizability, re-usability, and effectiveness can significantly benefit various downstream tasks.

The common practice of SSL:

1. Optimize **the pre-train model** with SSL objective on the large-scale unlabeled data.
2. Optimize **the downstream model** on the various downstream annotated dataset, where the input feature is the universal representation extracted from the pre-trained model.

As for SSL in Speech, we have witnessed great success of SSL in content related task, e.g. wav2vec 2.0/HuBERT SSL methods achieve state-of-the-art results in ASR.

It is unknown if we can also boost the performance of SSL for speaker related task (e.g. speaker verification, diarization task).

Mask prediction loss

Mask prediction loss, proposed by HuBERT, is the **state-of-the-art** SSL method for content representation learning.

Main idea: conduct **iterative offline clustering** to provide target labels and perform BERT-like mask prediction loss.

Steps:

1. Conduct **k-means clustering** on the MFCC feature of input signals.
2. Set the clustering center of each input frame as the **pseudo target label**.
3. Train a Transformer-based model with the **mask prediction loss**, where the Transformer encoders are fed with the masked input features $\tilde{X}$, and predict the pseudo target label z_t in the masked region M:

$$\mathcal{L}_{\text{Content}} = -\sum_{t\in M} \log f(z_t|\tilde{X}, t)$$

4. Given the pre-trained model, we conduct k-means clustering on **the latent representations** generated by the pre-trained model, and start a new iteration from step 2.

Utterance-wise contrastive loss

Utterance-wise contrastive loss is proposed to enhance single-speaker representation learning.

Assume: 1) each training utterance contains one active speaker. 2) Each utterance in the training batch belongs to a different speaker,

Main idea: the pre-trained model is optimized to **discriminate** the representations from the same utterance or the different utterance

Methods: For each training batch, we extract and quantize the latent representations from the internal layer of Transformer encoders, then perform **contrastive loss** over the **quantized representations** in the **mask regions**, where the representations within the same utterance are considered as positive instances, the representations from other utterances are considered as negative instances.

The speaker information modeling loss:

$$\mathcal{L}_{\text{Contrastive}} = -\Big(\sum_{q_t^b \in \hat{Q}^b} \log \frac{\exp(\text{sim}(l_t^b, q_t^b)/\kappa)}{\exp(\text{sim}(l_t^b, q_t^b)/\kappa)+1} - \sum_{q_t^b \sim \hat{Q}\setminus\hat{Q}^b} \log \frac{1}{\exp(\text{sim}(l_t^b, q_t^b)/\kappa)+1}\Big)$$

$$\mathcal{L}_{\text{diversity}} = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

$$\mathcal{L}_{\text{Speaker}} = \mathcal{L}_{\text{Contrastive}} + \alpha \mathcal{L}_{\text{diversity}}$$

The final UniSpeech-SAT pre-training loss:

$$\mathcal{L}_{\text{UniSpeech-SAT}} = \mathcal{L}_{\text{Speaker}} + \beta \mathcal{L}_{\text{Content}}$$

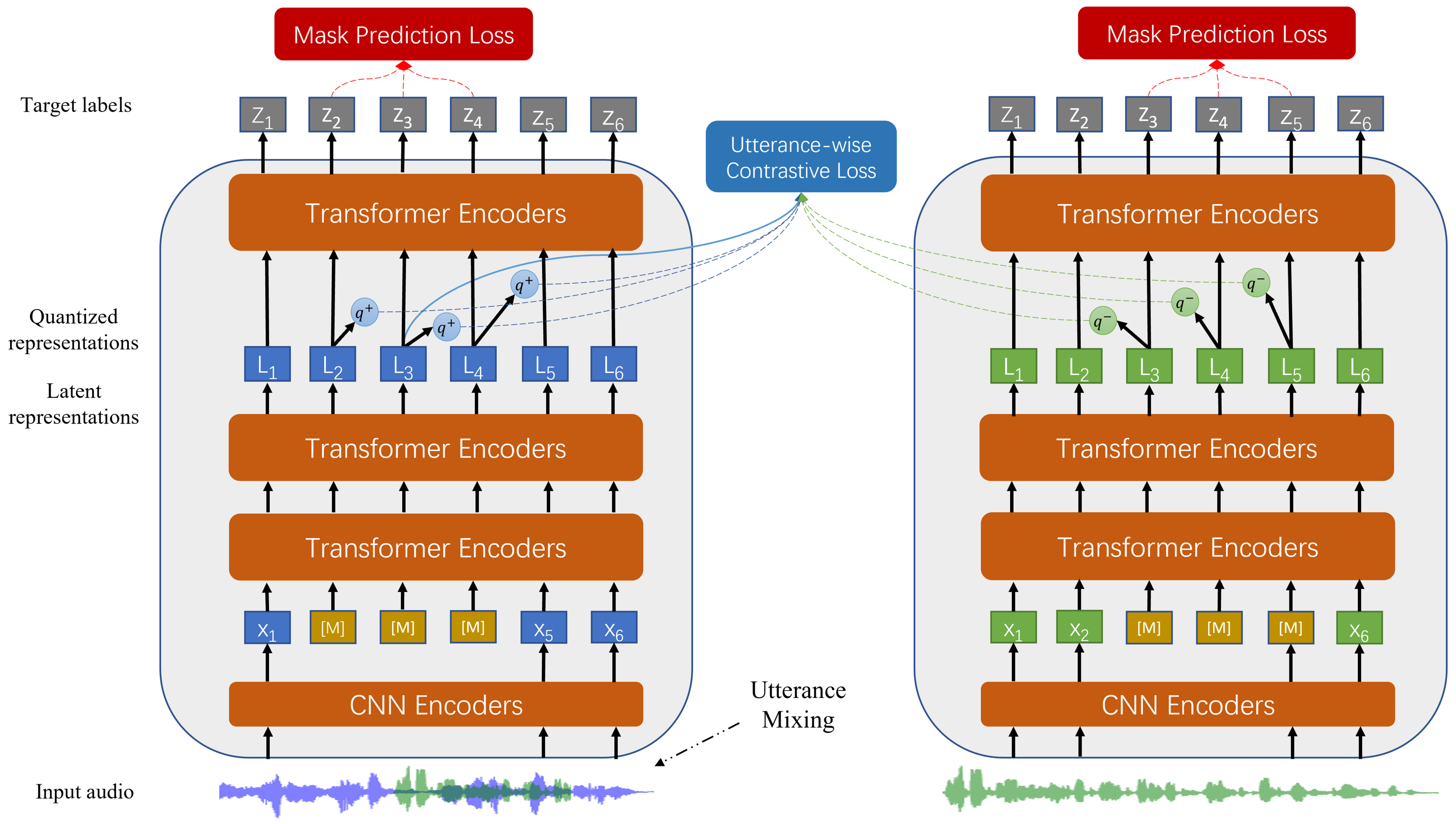
UniSpeech-SAT

Can we apply SSL for both content related task and speaker related task?

In this work, we propose Universal Speech representation learning with Speaker Aware pre-Training, to improve existing SSL framework for speaker representation learning.

Specifically, UniSpeech-SAT includes the following methods:

1. Mask prediction loss (from HuBERT) -> content representation learning
2. Utterance-wise contrastive loss -> single-speaker representation learning
3. Utterance mixing augmentation -> multi-speaker representation learning
4. Large and diverse pre-training data -> robust representation learning



Utterance mixing augmentation

Utterance mixing augmentation is proposed to further boost multi-speaker representation learning.

Main idea: simulate the **multi-speaker speech** for self-supervised pretraining when only single-speaker pretraining data is available.

Methods: For each training batch, we first randomly choose some utterances as the primary utterances. Then, for each primary utterance, we randomly choose an utterance from the same batch, crop a chunk of random length and mix it with the primary utterance in a random region.

With the utterance mixing method, the model is trained to extract **the information of the main speaker** from the mixed audio with the single-speaker information modeling loss and predict the **content information corresponding to the main speaker** with the content information modeling loss.

Note that we constrain the mixing portion in each utterance to be less than 50%, avoiding potential label permutation problem.

Conclusion

In this work, we propose a **speaker aware pre-training method** which is complementary to current ASR oriented pre-training.

The evaluation on the SUPERB benchmark shows our universal speech representation achieves **state-of-the-art overall performance** and outperforms other baselines by **a large margin**.

This work is extended to a journal paper **WavLM** (paper, code), where we sheds light on a general pre-trained model for **full stack speech processing task** and achieve 1) SOTA results on all the 10 tasks of SUPREB. 2) SOTA results on 4 typical speech tasks from different speech aspects: speaker verification, speech separation, speaker diarization and speech recognition.

Large and diverse pre-training data

Previous works only use the **audiobook speech** for pre-training, which limits the generalizability of the pre-trained speech representation in diverse scenarios.

Towards robust speech representation learning, we scale up unlabeled pre-training data to **94k hours** of public audios from **diverse domains**, including:

1. 10K hours Gigaspeech data, from audiobooks, podcasts and YouTube.
2. 24K hours VoxPopuli data, from European Parliament (EP) event recordings.
3. 60k hours LibriVox data, from audiobooks

Universal Representation Evaluation with SUPERB

We evaluate our models on SUPERB, which is designed to provide a standard and comprehensive testbed for pretrained models on various speech tasks.

SUPERB Policies:

1. The design of **task specific layers** follows the SUPERB official implementations for each downstream task.
2. Pre-trained models are **frozen** to limit the space of the fine-tuning hyperparameter search
3. The task specific layers consume the **weighted sum results** of the hidden states extracted from each layer of the pre-trained model

Universal Representation Evaluation Results

Table 1: Universal speech representation evaluation on SUPERB benchmark.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.

Method	#Params	Corpus	Speaker SID	Speaker ASV	Speaker SD	Content PR	Content ASR (WER)		Content KS	Content QbE	Semantics IC	Semantics SF		ParaL ER	Overall
			Acc ↑	EER ↓	DER ↓	PER ↓	w/o ↓	w/ LM ↓	Acc ↑	MTWV ↑	Acc ↑	F1 ↑	CER ↓	Acc ↑	Score ↑
FBANK	-	-	8.5E-4	9.56	10.05	82.01	23.18	15.21	8.63	0.0058	9.10	69.64	52.94	35.39	44.2
PASE+ [14]	7.83M	LS 50 hr	37.99	11.61	8.68	58.87	25.11	16.62	82.54	0.0072	29.82	62.14	60.17	57.86	57.5
APC [8]	4.11M	LS 360 hr	60.42	8.56	10.53	41.98	21.28	14.74	91.01	0.0310	74.69	70.46	50.89	59.33	67.6
VQ-APC [10]	4.63M	LS 360 hr	60.15	8.72	10.45	41.08	21.20	15.21	91.11	0.0251	74.48	68.53	52.91	59.66	67.2
NPC [11]	19.38M	LS 360 hr	55.92	9.40	9.34	43.81	20.20	13.91	88.96	0.0246	69.44	72.79	48.44	59.08	67.0
Mockingjay [12]	85.12M	LS 360 hr	32.29	11.66	10.54	70.19	22.82	15.48	83.67	6.6E-04	34.33	61.59	58.89	50.28	56.1
TERA [13]	21.33M	LS 360 hr	57.57	15.89	9.96	49.17	18.17	12.16	89.48	0.0013	58.42	67.50	54.17	56.27	64.2
modified CPC [2]	1.84M	LL 60k hr	39.63	12.86	10.38	42.54	20.18	13.53	91.88	0.0326	64.09	71.19	49.91	60.96	65.1
wav2vec [3]	32.54M	LS 960 hr	56.56	7.99	9.90	31.58	15.86	11.00	95.59	0.0485	84.92	76.37	43.71	59.79	71.5
vq-wav2vec [4]	34.15M	LS 960 hr	38.80	10.38	9.93	33.48	17.71	12.80	93.38	0.0410	85.68	77.68	41.54	58.24	69.3
wav2vec 2.0 Base [5]	95.04M	LS 960 hr	75.18	5.74	6.02	6.08	6.43	4.79	96.23	0.0233	92.35	88.30	24.77	63.43	80.3
HuBERT Base [6]	94.68M	LS 960 hr	81.42	5.11	5.88	5.41	**6.42**	**4.79**	96.30	0.0736	98.34	88.53	25.20	64.92	82.0
UniSpeech-SAT Base	94.68M	LS 960 hr	85.76	**4.31**	4.41	5.40	6.75	4.86	96.75	0.0927	98.58	88.98	23.56	66.04	83.0
– contrastive loss	94.68M	LS 960 hr	84.74	4.61	4.72	5.22	6.80	5.17	96.79	0.0956	98.31	88.56	24.00	65.60	82.8
– utterance mixing	94.68M	LS 960 hr	85.97	4.35	5.87	5.06	7.04	5.05	96.88	0.0866	98.10	88.50	24.52	65.97	82.7
UniSpeech-SAT Base+	94.68M	CD 94k hr	**87.59**	4.36	**3.80**	**4.44**	6.44	4.88	**97.40**	**0.1125**	**98.84**	**89.76**	**21.75**	**68.48**	**84.0**
wav2vec 2.0 Large [5]	317.38M	LL 60k hr	86.14	5.65	5.62	4.75	3.75	3.10	96.6	0.0489	95.28	87.11	27.31	65.64	82.1
HuBERT Large [6]	316.61M	LL 60k hr	90.33	5.98	5.75	3.53	**3.62**	**2.94**	95.29	0.0353	98.76	89.81	21.76	67.62	83.5
UniSpeech-SAT Large	316.61M	CD 94k hr	**95.16**	**3.84**	**3.85**	**3.38**	3.99	3.19	**97.89**	**0.0836**	**99.34**	**92.13**	**18.01**	**70.68**	**85.6**

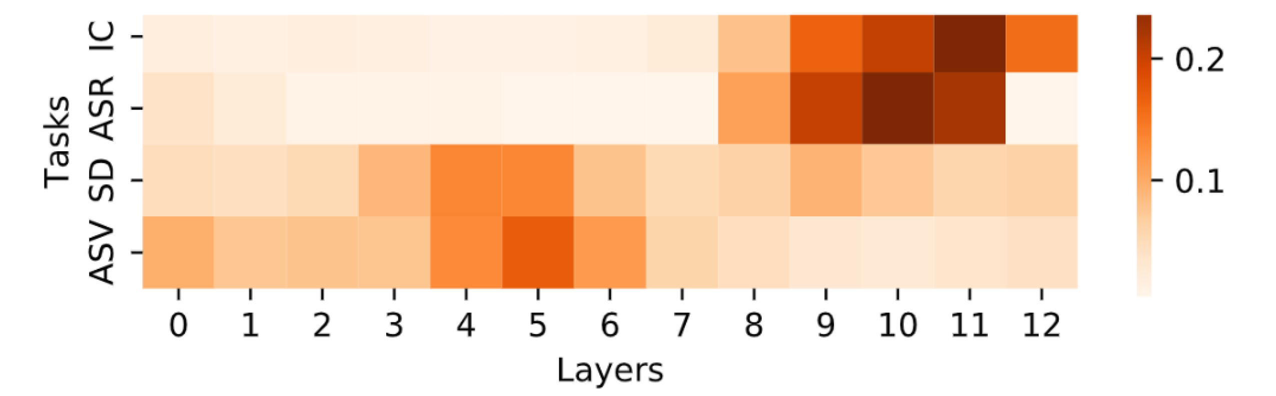


Fig. 2: Weight Analysis.

Table 2: Results of UniSpeech-SAT Base+ with various mixing ratios on 94k hours training data.

Method	Ratio	Speaker SD	Content ASR (WER)		Semantics IC	ParaL ER
		DER ↓	w/o ↓	w/ LM ↓	Acc ↑	Acc ↑
HuBERT Base [6]	-	5.88	6.42	4.79	98.34	64.92
UniSpeech-SAT Base+	0.0	5.04	**6.39**	**4.76**	99.24	66.32
	0.2	3.80	6.44	4.88	98.84	**68.48**
	0.5	**3.73**	6.65	5.18	**99.29**	67.36