SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities

Anonymous ACL submission

Abstract

Transfer learning has proven to be crucial in advancing the state of speech and natural language processing research in recent years. In speech, a model pre-trained by self-supervised learning transfers remarkably well on multiple tasks. However, the lack of a consistent evaluation methodology is limiting towards a holistic understanding of the efficacy of such models. SUPERB was a step towards introducing a common benchmark to evaluate pre-trained models across various speech tasks. In this paper, we introduce SUPERB-SG, a new benchmark focused on evaluating the semantic and generative capabilities of pre-trained models by increasing task diversity and difficulty over SUPERB. We use a lightweight methodology to test the robustness of representations learned by pre-trained models under shifts in data domain and quality across different types of tasks. It entails freezing pre-trained model parameters, only using simple task-specific trainable heads. The goal is to be inclusive of all researchers, and encourage efficient use of computational resources. We also show that the task diversity of SUPERB-SG coupled with limited task supervision is an effective recipe for evaluating the generalizability of model representation.

1 Introduction

Transfer learning is a paradigm in machine learning that has been very effective for natural language processing (NLP) (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2019; Dong et al., 2019; Yang et al., 2019; Raffel et al., 2020; Lewis et al., 2019; Conneau et al., 2020), and speech processing (van den Oord et al., 2018; Rivière et al., 2020; Chung et al., 2019; Schneider et al., 2019; Baevski et al., 2020; Hsu et al., 2021; Liu et al., 2020b,a; Ravanelli et al., 2020; Ling et al., 2020; Ling and Liu, 2020). Self-supervised learning (SSL) is the main driver of this paradigm, an effective and scalable way to learn high-level representation of language that transfers to a variety of tasks. SSL entails learning from the input or some perturbation of it without the need for labelled data. This has unlocked the usage of large amounts of cheaply available unlabelled data. It lends naturally to neural network models that have been shown to possess impressive scaling characteristics such that it is often enough to increase the model and data sizes to improve downstream performance (Hestness et al., 2017; Shazeer et al., 2017; Jozefowicz et al., 2016; Mahajan et al., 2018; Radford et al., 2019).

Speech signal consists of acoustic, linguistic, prosodic, and speaker characteristics. SSL algorithms in speech must be evaluated in their ability to produce representations that are useful for tasks that demand understanding of linguistic, speaker, and prosodic elements of spoken language as well as high-level semantics. Researchers have used auto-regressive, contrastive, discriminative and multi-task learning objectives to pre-train models, and have investigated their capabilities across tasks like phoneme recognition (van den Oord et al., 2018; Chung et al., 2019), automatic speech recognition (ASR) (Liu et al., 2020a; Schneider et al., 2019; Ling and Liu, 2020; Ravanelli et al., 2020; Hsu et al., 2021; Chang et al., 2021), speaker verification (Fan et al., 2020), speaker identification (Chung et al., 2019; Liu et al., 2020b), emotion recognition (Macary et al., 2021), speech translation (Chung et al., 2019), voice conversion (Lin et al., 2020; Huang et al., 2021a), spoken language understanding (Lai et al., 2021), and text-to-speech (Álvarez et al., 2019). However, the methodology in such studies vary in the use of datasets, fine-tuning strategy and task-specific model architecture. To bridge this gap, SUPERB (Yang et al., 2021) introduced a standardized benchmark of 10 speech tasks to compare 13 pre-trained models and a Log Mel-Filterbank baseline. It studied the models' performance in tasks focusing on linguistic

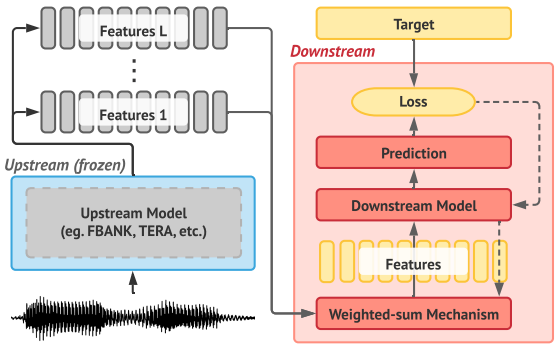

Figure 1: Illustration of the detailed training procedure. A trainable weighted-sum mechanism is used to summarize all layers' representations into a sequence of vectors and then taken by downstream model as input. Upstream is frozen through the whole process. Dashed arrow (- - →) is used to indicate the flow of gradient when back propagating.

(phoneme recognition and automatic speech recognition, keyword spotting and query by example), shallow semantic (intent classification and slot filling), speaker (speaker identification, speaker verification and speaker diarization), and prosodic (emotion recognition) characteristics.

In this paper, we introduce SUPERB-SG, a benchmark with 5 new tasks, which are speech translation, out-of-domain ASR, voice conversion, speech separation, and speech enhancement, with an emphasis on evaluating the semantic and generative capabilities of pre-trained models that require high-level representations to capture linguistic, semantic, and speaker characteristics. These tasks go beyond speech recognition by focusing on various other aspects that are essential to building intelligent speech interfaces. Further, we show that while SSL models achieve close to state-of-the-art performance on many tasks, there isn't one model that outperforms all others, and that a simple Log Mel-Filterbank can perform competitively on some tasks. We also demonstrate the robustness of our methodology with an ablation study over different task-specific model architectures and data sizes.

The introduction of these new tasks of varying difficulty takes us closer to a more comprehensive unified standard speech benchmark. We hope that this will motivate the development of more powerful, generalizable, and reusable pre-trained models to democratize the advancement of speech research. To facilitate this, we will release the code and integrate the tasks with the SUPERB benchmark.

2 Related Work

As more powerful SSL models are proposed with promising performance on various tasks, researchers continually try to find extensive evaluation methods to assess model performance, and wish to holistically understand the capability of the learned representations in these models.

SUPERB (Yang et al., 2021) is a framework to benchmark the SSL models on 10 speech tasks by learning task-specific prediction heads on top of the frozen shared SSL models. Although the tasks in SUPERB span across different domains, most of them are simple classification problems, or only require utilization of shallow semantics. In contrast, we focus on harder semantic and generative tasks.

Another recently proposed benchmark is the LeBenchmark (Evain et al., 2021), investigating the performance of SSL models trained on French data with three semantic tasks. However, they only consider wav2vec 2.0 (Baevski et al., 2020) with different architectures as their upstream models (i.e., networks pre-trained with SSL). Here, we evaluate a diverse set of SSL models, and offer a more comprehensive analysis.

The Zero Resource Speech Benchmark 2021 (Nguyen et al., 2020) introduces unsupervised speech processing tasks, particularly the spoken language modeling problem. They evaluate the SSL models via zero-shot probings at four linguistic levels. While their benchmark task is specific for certain domain, we use various tasks to evaluate different aspects of SSL models.

The HEAR 2021 Challenge[1] aims to develop general-purpose audio representation by focusing on audio tasks beyond speech that include sound event detection, speech commands and pitch chroma classification. We specifically focus on various aspects of speech processing, thus providing a wide variety of spoken language tasks.

3 SUPERB-SG

3.1 Tasks and Datasets

This section introduces the tasks in SUPERB-SG, including why we choose these tasks and how we design the task-specific heads for fine-tuning. Following SUPERB's methodology, we use a lightweight fine-tuning approach wherein we freeze the pre-trained model parameters and only keep

[1] https://neuralaudio.ai/hear2021-holistic-evaluation-of-audio-representations.html

Upstream	Network	#Params	Stride	Input	Corpus	Pretraining	Official Github
FBANK	-	0	10ms	waveform	-	-	-
PASE+	SincNet, 7-Conv, 1-QRNN	7.83M	10ms	waveform	LS 50 hr	multi-task	santi-pdp / pase
APC	3-GRU	4.11M	10ms	FBANK	LS 360 hr	F-G	iamyuanchung / APC
VQ-APC	3-GRU	4.63M	10ms	FBANK	LS 360 hr	F-G + VQ	iamyuanchung / VQ-APC
NPC	4-Conv, 4-Masked Conv	19.38M	10ms	FBANK	LS 360 hr	M-G + VQ	Alexander-H-Liu / NPC
Mockingjay	12-Trans	85.12M	10ms	FBANK	LS 360 hr	time M-G	s3prl / s3prl
TERA	3-Trans	21.33M	10ms	FBANK	LS 960 hr	time/freq M-G	s3prl / s3prl
DeCoAR 2.0	12-Trans	89.84M	10ms	FBANK	LS 960 hr	time M-G + VQ	awslabs / speech-representations
Modified CPC	5-Conv, 1-LSTM	1.84M	10ms	waveform	LL 60k hr	F-C	facebookresearch / CPC_audio
wav2vec	19-Conv	32.54M	10ms	waveform	LS 960 hr	F-C	pytorch / fairseq
vq-wav2vec	20-Conv	34.15M	10ms	waveform	LS 960 hr	F-C + VQ	pytorch / fairseq
wav2vec 2.0 Base	7-Conv 12-Trans	95.04M	20ms	waveform	LS 960 hr	M-C + VQ	pytorch / fairseq
wav2vec 2.0 Large	7-Conv 24-Trans	317.38M	20ms	waveform	LL 60k hr	M-C + VQ	pytorch / fairseq
HuBERT Base	7-Conv 12-Trans	94.68M	20ms	waveform	LS 960 hr	M-P + VQ	pytorch / fairseq
HuBERT Large	7-Conv 24-Trans	316.61M	20ms	waveform	LL 60k hr	M-P + VQ	pytorch / fairseq

Table 1: Details of the investigated SSL representations. LibriSpeech and LibriLight are denoted as LS and LL, respectively. For the pretraining methods, we abbreviate "vector quantization" as VQ, "future" as F, "masked" as M, "generation" as G, "contrastive discrimination" as C, and "token prediction/classification" as P. Parameters for both pretraining and inference are counted.

the task-specific head's parameters trainable. This setting serves the dual purpose of evaluating the robustness as well as the generalizability of the speech representations, and provides a resource-efficient way of fine-tuning the models that is inclusive of participants with constrained compute resources. We call the pre-trained model as upstream model and the task-specific heads as downstream model. We now discuss the newly added tasks in `SUPERB-SG` in the following sub-sections.

3.1.1 Speech Translation

Speech translation (ST) involves translating the acoustic speech signals in the source language into the words in the target language. We use it to evaluate the semantic capability of SSL models, and how they benefit the translation task. We use the CoVoST2 En→De (Wang et al., 2020) dataset with their official train, validation, and test splits while removing all the samples containing "REMOVE", resulting in 425.8, 25.9 and 24.5 hours respectively. For text, we keep original case, normalize punctuation, and build character vocabulary with 100% train-set coverage. We report case-sensitive detokenized BLEU using sacreBLEU (Post, 2018). Our downstream model has an encoder-decoder architecture with 3 layers of Transformers (Vaswani et al., 2017) each. A convolutional sub-sampler is used to reduce the sequence length of the input before feeding it to the encoder. We train our model with label-smoothing using a probability of 0.1. A beam size of 20 is used for inference.

3.1.2 Out-of-domain ASR

Although an ASR is included in `SUPERB`, it only examines SSL models on read English corpus LibriSpeech (Panayotov et al., 2015). Therefore, we introduce out-of-domain ASR (OOD-ASR), which aims to evaluate the models' capabilities across languages, and out-of-domain scenarios. The OOD-ASR tasks are categorized into cross-lingual and spontaneous speech tasks. For the cross-lingual tasks, we choose the Mexican Spanish (es), Mandarin (zh), and Arabic (ar) subsets from Common Voice 7.0 (Ardila et al., 2020) containing 21.5, 31.2, and 30.7 hours of training data respectively. The validation set sizes are 1.2 hours, 14.4 hours and 12.24 hours, and the test set sizes are 0.6 hour, 15.3 hours and 12.5 hours for es, zh and ar respectively. For the spontaneous speech task (spon), we use the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al., 2000 – 2005), consisting of 60 conversations over different topics spanning 16.7 hours of data. The validation and test set sizes are 1.6 hours and 2.2 hours respectively. For evaluation, we use word error rate (WER) as the metric except for Mandarin which character error rate (CER) is used. The error rates are averaged across all sub-tasks to offer an overall score. The ASR model is a 3-layer BLSTM (Hochreiter and Schmidhuber, 1997) with hidden states of 1024 dimension. The training objective is to minimize the Connectionist Temporal Classification (CTC) loss (Graves et al., 2006). During inference, we use CTC greedy decoding without language model

re-scoring to simplify the process and to highlight the impact of the learned acoustic representations.

3.1.3 Voice Conversion

For voice conversion (VC), we consider the intra-lingual VC task in VCC2020 (Zhao et al., 2020) under the any-to-one (A2O) setting. A2O VC aims to convert speech from any arbitrary speaker into that of a predefined target speaker. We use the task to evaluate the speaker transferability as well as the generalizability of the SSL models. We use 60 utterances from the target speaker that spans 5 minutes for training, and 25 utterances for testing that span 2 minutes. No validation set was used. We use the commonly used mel-cepstrum distortion (MCD), word error rate (WER) and automatic speaker verification (ASV) accept rate from off-the-shelf ASR and ASV models as evaluation metrics. The downstream model is trained to reconstruct the acoustic feature from the upstream representations in a target-speaker-dependent manner. In the conversion phase, given the representations extracted by the upstream, the model generates the converted acoustic features, which are then sent to a neural vocoder to synthesize the converted waveform. We adopted Tacotron2 (Shen et al., 2018) as the downstream model, which is an autoregressive network consisting of convolutional and LSTM layers. For the neural vocoder, we used the Hifi-GAN (Kong et al., 2020). We follow an implementation described in (Huang et al., 2021b).

3.1.4 Speech Separation

Speech separation (SS) is the task of separating target speech from background interference (Wang and Chen, 2018). It is an important step in speech processing, especially for noisy and multi-speaker scenarios. We investigate the speech separation problem on a dataset simulated from LibriSpeech (Cosentino et al., 2020) and WHAM! (Wichern et al., 2019) noise. We use 16kHz version of the dataset containing 2 speakers, and focus on the *mix_clean* condition. The train and evaluation sets contain 43.3 and 4.2 hours of speech simulated from LibriSpeech's *train-clean-100* and *test-clean*. This task is used to evaluate the generative capability of SSL models when input is a mixture of acoustic signals. We use the scale-invariant signal-to-distortion ratio improvement (SI-SDRi) as the evaluation metric. For the downstream model, we use a 3-layer BLSTM model with dimension of 896 for each direction to predict the short-time Fourier transform (STFT) masks for each speaker, and the predictions are transformed back to the time domain using inverse short-time Fourier transform (iSTFT). Permutation invariant training (PIT) (Yu et al., 2017) is performed to optimize the mean square error between the predicted mask and Ideal Non-negative Phase Sensitive Mask (INPSM) (Erdogan et al., 2015; Kolbæk et al., 2017). We choose frequency domain method instead of a time domain based method because of the stride size constraint and computational cost.

3.1.5 Speech Enhancement

Speech enhancement (SE) is the task of removing background noise from a degraded speech signal, and it aims to improve the perceived quality and intelligibility of the signal. We include this task to evaluate the generative capability under noisy conditions. In `SUPERB-SG`, we discuss the speech enhancement problem on the Voicebank-DEMAND (Veaux et al., 2013) corpus. The train, validation, and test sets contain 8.8, 0.6 and 0.6 hours of speech respectively. Our evaluation metrics are Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI). For the downstream model, we follow the mask-based speech enhancement pipeline in (Kolbæk et al., 2017). A 3-layer BLSTM model similar to the speech separation task is trained to predict the spectral mask for the clean signal. The mean square error between the predicted mask and INPSM is used as the objective.

3.2 Self-supervised Models

We evaluate the tasks on 15 upstream models that span across different architectures, sizes, and learning objectives. Some of these models also use vector quantization which has an added benefit of signal compression. For grounding, we use Log Mel Filterbank as our baseline. The detailed properties of upstream models are shown in Table 1.

4 Experimental Setup

Following `SUPERB`, we fix upstream models parameters for all downstream tasks' training. We extract the frame-level representations for each hidden layer of the upstream models from raw waveform, and use a trainable task-specific weighted-sum mechanism to summarize all layers' representations into a sequence of vectors. The summarized representations are then used as the downstream

Upstream	ST	OOD-ASR	VC			SS	SE	
	BLEU↑	WER↓	MCD↓	WER↓	ASV↑	SI-SDRi↑	PESQ↑	STOI↑
FBANK	2.32	63.58	8.47	38.3	77.25	9.23	2.55	93.6
PASE+	3.16	61.56	8.66	30.6	63.20	9.87	2.56	93.9
APC	5.95	63.12	8.05	27.2	87.25	8.92	2.56	93.4
VQ-APC	4.23	63.56	7.84	22.4	94.25	8.44	2.56	93.4
NPC	4.32	61.66	7.86	30.4	94.75	8.04	2.52	93.1
Mockingjay	4.45	65.27	8.29	35.1	79.75	9.29	2.53	93.4
TERA	5.66	58.49	8.21	25.1	83.75	10.19	2.54	93.6
DeCoAR 2.0	9.94	53.62	7.83	17.1	90.75	8.54	2.47	93.2
Modified CPC	4.82	62.54	8.41	26.2	71.00	10.40	2.57	93.7
wav2vec	6.61	55.86	7.45	10.1	98.25	9.30	2.53	93.8
vq-wav2vec	5.66	60.66	**7.08**	13.4	**100.00**	8.16	2.48	93.6
wav2vec 2.0 Base	14.81	46.95	7.50	10.5	98.00	9.77	2.55	93.9
wav2vec 2.0 Large	12.48	44.69	7.63	15.8	97.25	10.02	2.52	94.0
HuBERT Base	15.53	46.69	7.47	**8.0**	98.50	9.36	2.58	93.9
HuBERT Large	**20.01**	**44.08**	7.22	9.0	99.25	**10.45**	**2.64**	**94.2**

Table 2: Evaluating various SSL representations on new semantic and generative downstream tasks. ↑ indicates the higher the better and ↓ indicates the lower the better. The complete results of OOD-ASR are in Appendix A.

model's input. An overview of the training procedure is demonstrated in Figure 1. This procedure is consistent for all experiments, offering a fair and simple evaluation strategy for all upstream models.

5 Results and Discussion

5.1 Main result

The results of the upstream models evaluated on SUPERB-SG are shown in Table 2. We only report the averaged WER for OOD-ASR. Full results can be found in Appendix A. For speech-to-text tasks (ST and OOD-ASR), wav2vec 2.0 and HuBERT offer competitive results, while DeCoAR 2.0 shows some improvements. In speech generation tasks (VC, SS, and SE), FBANK yields comparable or superior performance than some SSL models, especially for those metrics that take the quality of the output signal into account. For VC, the 3 reported metrics have the same trend for respective models. Here, vq-wav2vec achieves the best performance on MCD and ASV, while HuBERT performs the best on WER. For SS, Hubert-Large achieves the best performance, followed by Modified CPC. PASE+, which is pre-trained with denoising tasks, performs better than half the SSL models, but this observation doesn't transfer to the other tasks. For SE, all upstream models perform comparably. The largest gap is only 0.17 in PESQ and 1.1 in STOI.

Overall, no model outperforms all others on all tasks. However, HuBERT-Large performs most competitively on all downstream tasks, especially those requiring linguistic and semantic signals.

5.2 Correlation between tasks

We analyze the correlations between tasks in SUPERB-SG to understand the similarity between tasks, and verify if the experimental results agree with the common understanding of related tasks based on shared representation they require.

To compute the correlation, we first change all metrics into a higher-better manner. Then, we compute the Spearman's rank correlation coefficients (Spearman's ρ) between all pairs of tasks. For multiple metrics contained in a single task, such as MCD/WER/ASV in VC as well as PESQ/STOI in SE, we compute each of them separately.

To make our analysis more representative and generalized to all speech domains, we bring back the six tasks from SUPERB (Yang et al., 2021) that are considered representative of the following four domains: (i) Content recognition tasks containing Phoneme Recognition (PR), Automatic Speech Recognition (ASR) (ii) Speaker identity tasks including Identification (SID), Automatic Speaker Verification (ASV) (iii) Semantics task which is Intent Classification (IC) and (iv) Prosodic task which is Emotion Recognition (ER). Together with the 5 tasks introduced in this paper, we show the results of total 11 downstream tasks with the 14 corresponding metrics in Figure 2.

Overall, results show that all tasks except SS and SE have strong positive correlation among them. One possible explanation for SS and SE not showing strong correlation is that the low-level information closely related to audio signals is more criti-

Tasks	Top 2		Last 2	
ST	ASR (0.92)	IC (0.92)	ASV (0.75)	VC (0.76)
OOD-ASR	ASR (0.92)	PR (0.86)	ASV (0.70)	VC (0.72)
VC	PR (0.84)	ASR (0.77)	SID (0.64)	ER (0.66)
SS	SE (0.65)	OOD-ASR (0.46)	VC (0.01)	ASV (0.04)
SE	SS (0.65)	ER (0.39)	VC (0.17)	IC (0.25)

Table 3: Top 2 and last 2 tasks correlated with the five SUPERB-SG tasks ranked by Spearman's ρ.

Cluster	Metrics
A	ST, OOD-ASR, PR VC (WER), ASR, IC
B	SID, ASV, ER
C	VC (MCD), VC (ASV)
D	SS
E	SE (PESQ)
F	SE (STOI)

Table 4: K-means clustering result based on the correlation between each downstream tasks.

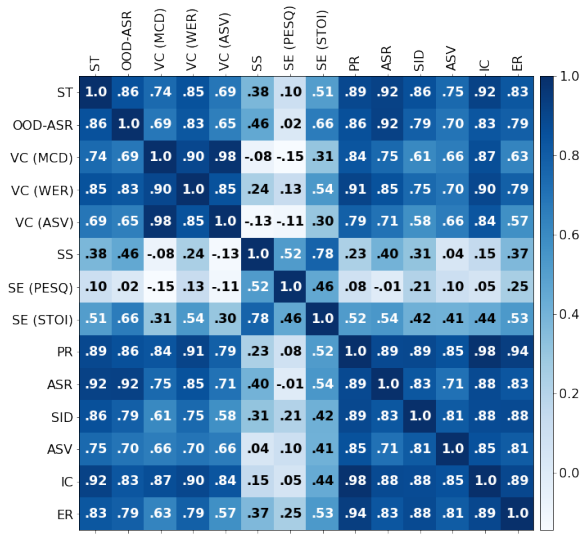

Figure 2: Spearman's ρ between tasks.

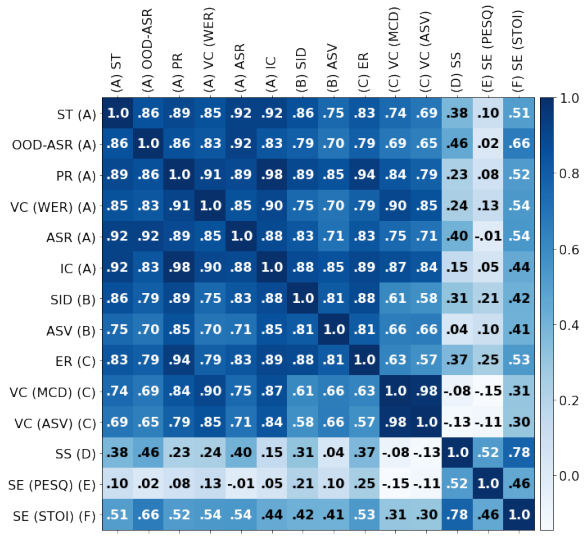

Figure 3: Spearman's ρ between tasks rearranged by clustering result.

cal as they need to reconstruct clean speech from interfering speakers and background noise by estimating the STFT masks. As a result, high-level information extracted from SSL models has little benefit for these tasks but is helpful for other tasks. As noted earlier, there is only a small gap in performance between FBANK and SSL models. If we leave SS and SE out, all correlation coefficients are greater than 0.58, showing that the SSL model representations are useful for multiple domains.

Although the Spearman's ρ are large in general in Figure 2, differences between tasks are observable. Here, we focus on the relation between correlation and similarity of tasks. We list the most and the least two correlated tasks comparing with ST, OOD-ASR, VC, SS, and SE. SS and SE are skipped as candidates for for the least correlated tasks since they dominate the results. For VC, we average the correlation coefficients across the three metrics. The results are shown in Table 3. ST and OOD-ASR are highly correlated with ASR since they both transform speech signals into discrete text tokens. IC is also correlated with ST since semantic information is required to perform both tasks. Moreover, ASV and VC are the least correlated tasks since they primarily focus on the speaker information with lesser regard to the semantic content. However, the absolute correlation values are still larger than 0.7. For VC, the speaker information needs to be removed while the content has to be kept, similar to PR and ASR but different from SID. SS and SE are correlated with each other and have a much lower correlation with speaker identity and semantics tasks, supporting our assumption. Overall, we find that empirically highly-correlated tasks require similar knowledge or understanding ability.

To give a broader view of our correlation results, we further cluster the downstream tasks by their

Architecture	ST		OOD-ASR		SS	
	architecture	#params	architecture	#params	architecture	#params
default	3-layer encoder 3-layer decoder Transformer	28.8M	2-layer BLSTM	53.4M	3-layer BLSTM dim 896	51.4M
small	no encoder 1-layer decoder Transformer	10.9M (× 0.38)	1-layer BLSTM	24.1M (× 0.45)	2-layer BLSTM dim 768	24.4M (× 0.47)
large	12-layer encoder 6-layer decoder Transformer	69.8M (× 2.42)	4-layer BLSTM	112.2M (× 2.10)	4-layer BLSTM dim 1152	114.50M (× 2.23)

Table 5: A detailed comparison of dowsntream model architectures. We report the number of trainable parameters when using TERA as upstream model while minor difference (< 10%) exists due to different upstream dimensions. For OOD-ASR, we average values across all sub-tasks since sub-tasks have different vocabulary sizes.

Upstream	ST	OOD-ASR	SS
	BLEU↑	WER↓	SI-SDRi↑
default			
FBANK	2.32	63.58	9.23
TERA	5.66	58.49	10.19
Modified CPC	4.82	62.54	**10.40**
wav2vec 2.0 Base	14.81	46.95	9.77
HuBERT Base	**15.53**	**46.69**	9.36
small			
FBANK	0.58	70.86	8.19
TERA	1.84	64.80	9.20
Modified CPC	1.44	67.83	**9.56**
wav2vec 2.0 Base	8.55	50.75	8.83
HuBERT Base	**9.24**	**50.32**	8.73
large			
FBANK	3.02	60.49	9.77
TERA	6.64	57.95	(▲) **10.87**
Modified CPC	4.56	59.73	(▼) 10.61
wav2vec 2.0 Base	16.81	(▲) **45.61**	9.86
HuBERT Base	**17.59**	(▼) 45.78	9.83

Table 6: Results on SS, ST, OOD-ASR when using different architectures. ▲ and ▼ are used to denote the rank changing. The complete results of OOD-ASR are in Appendix A.

correlation with each other using K-means. In this way, all the tasks are considered simultaneously, and the grouping is driven automatically by the empirical correlation results. If more than one metric are used in a downstream task, we cluster them independently. The clustering results are shown in Table 4 and a rearranged correlation map is shown in Figure 3. The result shows that the clusters of the tasks align with our empirical knowledge. Cluster A includes tasks that require content information, while tasks in cluster B are more sensitive to speaker and prosodic features. Cluster C contains metrics MCD and ASV of VC, which are used to evaluate the signal quality and the rates of speaker transfer. It is worth noting that WER in VC belongs to cluster A, showing that it is more similar to content-related tasks. Furthermore, clusters D, E, and F each contain one of the metrics in SS and SE, aligning with our assumption that these tasks utilize different types of information compared to other tasks. Overall, the clustering results roughly categorize tasks with respect to the signals they need from the SSL representations.

With the analysis of the correlation between tasks, we empirically confirm the reliability of the results, and show that we increase the heterogeneity among speech tasks over SUPERB. We further discover shared properties between tasks with clustering, and the result is aligned with our common understanding of related tasks.

5.3 Robustness of SUPERB-SG

To study the impact of downstream model architecture and the data sizes used in SUPERB-SG we evaluate the robustness of SUPERB-SG with variations in downstream model as well as training data size, and show that our conclusions still hold true.

We choose ST, OOD-ASR and SS as the downstream tasks for evaluation with an aim to cover semantic, content recognition, and generative task types. For the upstream models, FBANK, TERA, CPC, wav2vec 2.0 Base and HuBERT Base are used to cover different SSL algorithms.

5.3.1 Downstream model

For each task, 2 additional downstream architectures are created by modifying the number of layers and the hidden dimensions compared to our default setting. We create *small* and *large* models that are roughly the half and twice of *default* in terms of the number of trainable parameters. A detailed com-

Partition	ST	OOD-ASR				SS
		es	zh	ar	spon	
Train						
100%	425.80	21.44	31.05	30.39	11.43	43.27
10%	42.58	2.15	3.11	3.04	1.14	4.34
5%	25.91	1.07	1.56	1.52	0.57	2.17
1%	4.26	0.22	0.31	0.31	0.12	0.43
Dev	25.91	1.19	14.41	12.24	1.59	1.52
Test	24.51	0.62	15.32	12.46	2.15	4.19

Table 7: Hours of data in pseudo datasets.

Upstream	ST	OOD-ASR	SS
	BLEU↑	WER↓	SI-SDRi↑
100%			
FBANK	2.32	63.58	9.23
TERA	5.66	58.49	10.19
Modified CPC	4.82	62.54	**10.40**
wav2vec 2.0 Base	14.81	46.95	9.77
HuBERT Base	**15.53**	**46.69**	9.36
10%			
FBANK	0.46	85.39	5.65
TERA	(▾) 0.88	80.32	(▴) **6.72**
Modified CPC	(▴) 1.30	85.32	(▾) 6.59
wav2vec 2.0 Base	5.04	63.85	6.45
HuBERT Base	**5.57**	**63.43**	6.13
5%			
FBANK	0.27	89.70	4.52
TERA	0.44	86.95	(▴ 1) **5.59**
Modified CPC	0.37	87.97	(▾ 3) 4.95
wav2vec 2.0 Base	2.91	69.88	(▴ 1) 5.36
HuBERT Base	**3.35**	**69.33**	(▴ 1) 5.03
1%			
FBANK	0.03	99.53	2.29
TERA	0.04	98.31	3.24
Modified CPC	0.03	98.37	(▾ 3) 2.87
wav2vec 2.0 Base	0.33	92.46	(▴ 2) **3.34**
HuBERT Base	**0.38**	**92.17**	(▴ 1) 3.01

Table 8: Results on ST, OOD-ASR and SS when using different amount of training data. ▴ and ▾ are used to denote the rank changing. The complete results of OOD-ASR are in Appendix A.

parison of the downstream architectures is shown in Table 5. The results are shown in Table 6.

We show that the ranking of the upstream models is almost fixed when the model sizes are varied. As expected, the *small* architecture has worse performance than *default*, while *large* has better. Moreover, the scores causing the change in ranking are negligible, e.g., TERA/CPC in SS and wav2vec 2.0 Base/HuBERT Base in OOD-ASR with *large*. The results show that the relative performance achieved by different upstream models is agnostic to the downstream architecture, confirming the robustness of the framework used in SUPERB-SG.

5.3.2 Training data size

To study the effect of data size, we create 3 pseudo datasets per task by sub-sampling 10%, 5% and 1% from the original training set while fixing the validation and test sets. The statistics of the datasets are shown in Table 7, and the results are in Table 8.

The ranking of the upstream models remains almost the same for 10% of training data. When that is further reduced to 5%, there is a change in ranking in SS due to a performance drop in Modified CPC. Excluding Modified CPC, the ranking is still fixed showing that the relative performance of the upstream models is agnostic to data size.

Furthermore, when using only 1% of training data, most of the SSL models fail on the 3 downstream tasks. This phenomenon is caused by insufficient task-specific knowledge due to limited training data size. Although SSL models learn high-level representations from the unlabeled speech signal, acquisition of task-specific knowledge such as translingual ability in ST, text-level token mapping in OOD-ASR, and mask prediction in SS, requires non-trivial supervision.

We note that fewer training examples speeds training up but sacrifices the evaluation quality. Overall, we show the robustness of SUPERB-SG to variations in data size even when the training data is reduced to 5%, showing the reliability of the benchmark.

6 Conclusion

We introduce SUPERB-SG, a set of 5 new tasks that include speech translation, out-of-domain ASR, voice conversion, speech separation, and speech enhancement to evaluate the deep semantic and generative capabilities of SSL models. We evaluate 15 SSL models, and do a comprehensive analysis of the task correlations to demonstrate the reliability of our methodology. We test and confirm the robustness of SUPERB-SG in terms of the downstream model architecture as well as the training data size. The latest introduction of the semantic and generative tasks increases the diversity and difficulty of SUPERB, which can boost a more comprehensive understanding of the capability of various SSL models' representations, and help researchers discover the hidden properties of SSL techniques in development.

References

David Álvarez et al. 2019. Problem-agnostic speech embeddings for multi-speaker text-to-speech with samplernn. In *Proc. 10th ISCA Speech Synthesis Workshop*, pages 35–39.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*.

Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. 2021. DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit BERT. *arXiv preprint*.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Interspeech*, pages 146–150.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000 – 2005. Santa Barbara corpus of spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712. IEEE.

Solene Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Esteve, Benjamin Lecouteux, Francois Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech.

Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu. 2020. Exploring wav2vec 2.0 on speaker verification and language identification. *arXiv preprint arXiv:2012.06185*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

W.-C. Huang, Y.-C. Wu, T. Hayashi, and T. Toda. 2021a. Any-to-One Sequence-to-Sequence Voice Conversion using Self-Supervised Discrete Speech Representations. In *Proc. ICASSP*, pages 5944–5948.

Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, Hung-Yi Lee, Shinji Watanabe, and Tomoki Toda. 2021b. S3prl-vc: Open-source voice conversion framework with self-supervised speech representations. *arXiv preprint arXiv:2110.06280*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913.

J. Kong, J. Kim, and J. Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proc. NeurIPS*, volume 33, pages 17022–17033.

Cheng-I Lai, Yung-Sung Chuang, Hung-Yi Lee, Shang-Wen Li, and James Glass. 2021. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining. In *ICASSP*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ACL*.

Yist Y Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung-yi Lee, and Lin-shan Lee. 2020. Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention. *arXiv preprint arXiv:2010.14150*.

Shaoshi Ling and Yuzong Liu. 2020. DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*.

Shaoshi Ling, Yuzong Liu, Julian Salazar, and Katrin Kirchhoff. 2020. Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6429–6433. IEEE.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2020a. Tera: Self-supervised learning of transformer encoder representation for speech. *arXiv preprint arXiv:2007.06028*.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020b. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. *ICASSP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau. 2021. On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 373–380. IEEE.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tu Anh Nguyen et al. 2020. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*, pages 2227–2237.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv*.

Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. 2020. Multi-task self-supervised learning for robust speech recognition. In *ICASSP*, pages 6989–6993.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP*, pages 7414–7418.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions. In *Proc. ICASSP*, pages 4779–4783.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Christophe Veaux, Junichi Yamagishi, and Simon King. 2013. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE.

Changhan Wang, Anne Wu, and Juan Pino. 2020. CoVoST 2: A massively multilingual speech-to-text translation corpus.

DeLiang Wang and Jitong Chen. 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech processing universal performance benchmark. *Interspeech*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE.

Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda. 2020. Voice Conversion Challenge 2020 - Intra-lingual semi-parallel and cross-lingual voice conversion -. In *Proc. Joint Workshop for the BC and VCC 2020*, pages 80–98.

A Complete Out-of-domain ASR Results

Here, we provide complete results of OOD-ASR tasks, as shown in Tables 9, 10, 11. All upstream models used in this paper are trained with English speech data, but we are also interested in multilingual pre-trained models in OOD-ASR. Therefore, we evaluate the wav2vec 2.0 XLSR model on the OOD-ASR tasks, as shown in the last row of Table 9. XLSR has identical architecture as wav2vec 2.0 Large, but is trained with 56k hours of speech including 53 different languages. The pre-training data of XLSR cover our cross-lingual tasks' training data. As expected, using multilingual data improves OOD-ASR tasks and achieves the best performance among all upstream models.

Upstream	es	zh	ar	spon	
	WER↓	CER↓	WER↓	WER↓	AVG
FBANK	54.03	35.44	72.07	92.78	63.58
PASE+	52.11	35.52	70.47	88.15	61.56
APC	55.23	36.38	70.79	90.07	63.12
VQ-APC	55.32	37.06	71.56	90.292	63.56
NPC	51.07	35.85	69.87	89.86	61.66
Mockingjay	58.11	38.13	73.57	91.27	65.27
TERA	48.67	32.21	66.18	86.89	58.49
Modified CPC	54.37	36.22	68.94	90.61	62.54
DeCoAR 2.0	43.18	28.77	61.00	81.53	53.62
wav2vec	46.16	31.69	60.85	84.72	55.86
vq-wav2vec	52.02	36.55	66.19	87.89	60.66
wav2vec 2.0 Base	37.85	26.44	55.95	67.55	46.95
wav2vec 2.0 Large	35.75†	25.07†	54.29†	**63.64**†	44.69
HuBERT Base	37.15	26.23	54.94	68.41	46.69
HuBERT Large	**30.90**	**23.73**†	**50.60**‡	71.09‡	**44.08**
wav2vec 2.0 XLSR	26.90†	22.97†	49.63†	63.05†	40.64†

Table 9: Results of OOD-ASR tasks, where spon denotes spontaneous speech. † Normalized across dimensionality of representation to stabilize training and ensure convergence. ‡ Uses linear warmup of learning rates in the first 8k steps to stabilize training and ensure convergence.

Upstream	es	zh	ar	spon	
	WER↓	CER↓	WER↓	WER↓	AVG
default					
FBANK	54.03	35.44	72.07	92.78	63.58
TERA	48.67	32.21	66.18	86.89	58.49
Modified CPC	54.37	36.22	68.94	90.61	62.54
wav2vec 2.0 Base	37.85	26.44	55.95	**67.55**	46.95
HuBERT Base	**37.15**	**26.23**	**54.94**	68.41	**46.69**
small					
FBANK	63.86	41.97	80.30	97.30	70.86
TERA	57.13	37.66	73.92	90.49	64.80
Modified CPC	60.81	41.47	76.45	92.59	67.83
wav2vec 2.0 Base	41.84	30.22	61.72	**69.23**	50.75
HuBERT Base	**41.45**	**29.68**	**59.93**	70.21	**50.32**
large					
FBANK	46.39	37.71	65.35	92.52	60.49
TERA	45.41	37.40	64.48	84.53	57.95
Modified CPC	48.70	35.16	69.15	85.93	59.73
wav2vec 2.0 Base	34.02	27.60	54.10	**66.73**	**45.61**
HuBERT Base	**33.91**	**27.22**	**53.43**	68.57	45.78

Table 10: Complete results of OOD-ASR tasks with different model sizes.

Upstream	es	zh	ar	spon	
	WER↓	CER↓	WER↓	WER↓	AVG
100%					
FBANK	54.03	35.44	72.07	92.78	63.58
TERA	48.67	32.21	66.18	86.89	58.49
Modified CPC	54.37	36.22	68.94	90.61	62.54
wav2vec 2.0 Base	37.85	26.44	55.95	**67.55**	46.95
HuBERT Base	**37.15**	**26.23**	**54.94**	68.41	**46.69**
10%					
FBANK	84.82	62.97	93.27	100.49	85.39
TERA	76.44	58.54	88.49	97.79	80.32
Modified CPC	83.84	64.78	91.20	101.44	85.32
wav2vec 2.0 Base	61.26	43.50	72.98	**77.65**	63.85
HuBERT Base	**58.08**	**42.94**	**72.78**	79.94	**63.43**
5%					
FBANK	89.48	71.99	96.69	100.65	89.70
TERA	83.98	71.04	93.15	99.62	86.95
Modified CPC	88.61	67.61	95.71	99.93	87.97
wav2vec 2.0 Base	67.09	**50.58**	78.53	**83.33**	69.88
HuBERT Base	**66.29**	50.72	**76.59**	83.74	**69.33**
1%					
FBANK	96.79	96.73	99.85	104.73	99.53
TERA	94.73	98.82	99.77	99.93	98.31
Modified CPC	95.93	97.94	99.80	99.84	98.37
wav2vec 2.0 Base	**82.00**	94.38	92.41	**101.06**	92.46
HuBERT Base	82.36	**94.34**	**90.37**	101.60	**92.17**

Table 11: Complete results of OOD-ASR tasks with different data sizes.