INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

SpeechGLUE: How Well Can Self-Supervised Speech Models Capture Linguistic Knowledge?

Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, Tomohiro Tanaka
Yusuke Ijima, Taichi Asami, Marc Delcroix, Yukinori Honma

NTT Corporation, Japan

takanori.ashihara@ntt.com

Abstract

Self-supervised learning (SSL) for speech representation has been successfully applied in various downstream tasks, such as speech and speaker recognition. More recently, speech SSL models have also been shown to be beneficial in advancing spoken language understanding tasks, implying that the SSL models have the potential to learn not only acoustic but also linguistic information. In this paper, we aim to clarify if speech SSL techniques can well capture linguistic knowledge. For this purpose, we introduce SpeechGLUE, a speech version of the General Language Understanding Evaluation (GLUE) benchmark. Since GLUE comprises a variety of natural language understanding tasks, SpeechGLUE can elucidate the degree of linguistic ability of speech SSL models. Experiments demonstrate that speech SSL models, although inferior to text-based SSL models, perform better than baselines, suggesting that they can acquire a certain amount of general linguistic knowledge from just unlabeled speech data.

Index Terms: self-supervised learning, speech representation, linguistic knowledge, natural language processing

1. Introduction

Self-supervised learning (SSL) has become a prominent technique to leverage a large amount of unlabeled data in an unsupervised fashion. For the speech community, various SSL methods have been proposed [1–5] and accuracy has been dramatically improved, especially in automatic speech recognition (ASR) tasks under low-resource conditions. Subsequent studies have demonstrated success with task-generalizability, i.e., performance improvement in a wide range of tasks such as speaker recognition, emotion recognition, and speech enhancement [6, 7]. These positive results likely reflect the ability of the SSL model to learn a wide range of speech information (e.g., phonemes and speaker characteristics) from only speech data without any labels [1, 8–10]. Actually, a previous SSL study has demonstrated a clear relationship between latent representations and phonetic units [3, 9].

More recently, speech SSL models have also been utilized in spoken language understanding (SLU) tasks [11–13], e.g., named entity recognition and sentiment analysis, and these universal models have demonstrated superiority over conventional approaches. Their success can be naturally attributed to the speech information captured by SSL as described above. However, since performing SLU tasks requires natural language processing (NLP) ability, the benefit of SSL in these tasks may also imply that speech SSL models can latently capture linguistic characteristics, such as semantics and syntax, from speech signals in addition to acoustic characteristics.

The above implications are supported by other studies. Previous studies [14–16] have presented and utilized the zero-resource benchmark to evaluate the spoken language models, which are language models trained using the discrete acoustic units obtained by clustering the speech SSL output. Benchmark results have shown that the unit-based language models are feasible, indicating that self-supervised representation seemingly retains some multi-level information such as phonetics, lexicon, syntax, and semantics. The other work of [17, 18] comprehensively analyzed the speech representation layer-wise and demonstrated that SSL models capture some word and semantic information in the middle layers. While the aforementioned papers shed light on the language properties acquired by the speech SSL, further investigation is needed because, for example, it is important to align the representations across speech and text modalities for a unified multimodal SSL model [19, 20]. In particular, motivated by the above efforts, we aim to elucidate if linguistic information captured by speech SSL models is enough to solve practical and diverse natural language understanding (NLU) tasks. Moreover, we would like to compare the linguistic capabilities of not only speech SSL models but also text-based ones such as BERT [21] and identify their main differences to confirm if text data is still required to represent the linguistic information.

In this paper, for the purpose of exploiting the linguistic knowledge learned via SSL, we apply a probing task, which is a popular assessment method, to self-supervised speech representations. Specifically, we introduce the speech version of the General Language Understanding Evaluation (GLUE) benchmark [22], called SpeechGLUE, and evaluate speech SSL models extensively, in a fair comparison to NLP SSL models. While there is a large body of NLU probing tasks and benchmarks [22–24], we adopt GLUE which is relatively basic among the existing NLU benchmarks.[1] Since GLUE is designed to cover a diverse range of NLU tasks, SpeechGLUE, a collection of NLU tasks that convert input text to speech, is intended to evaluate the general-purpose NLU knowledge within the speech SSL models. For the conversion, we adopt text-to-speech (TTS) systems, which allow for the realization of tasks that assess purely linguistic knowledge by constraining acoustic conditions such as variations in speakers, speaking styles and recording settings. We base our implementation of SpeechGLUE on the S3PRL toolkit developed for SUPERB [6], which facilitates comparisons with various speech SSL models.[2]

From published experiments, speech SSL models lag behind NLP SSL models in performance, especially in the task of judging whether a sentence is linguistically acceptable. However, strong speech SSL models, such as WavLM Large, perform substantially better than chance level or baselines

[1] Note that our approach can be applied to any probing task of NLP.

[2] We release SpeechGLUE for reproducibility and comparison with successive SSL techniques at https://github.com/ashi-ta/speechGLUE.

10.21437/Interspeech.2023-1823

Table 1: *Brief summary of GLUE. MCC, PCC, and SCC denote Matthews, Pearson, and Spearman correlation coefficients, respectively. For details on this benchmark, see the original paper of [22].*

Corpus	Task	Metrics	Labels
	Single-sentence tasks		
CoLA	acceptability (grammaticality)	MCC	unacceptable / acceptable
SST2	sentiment analysis	accuracy	positive / negative
	Similarity and paraphrase tasks using sentence pairs		
MRPC	semantic equivalence (paraphrase)	accuracy & F1	equivalent / not equivalent
QQP	semantic equivalence (paraphrase)	accuracy & F1	duplicate / not duplicate
STS-B	sentence similarity	PCC & SCC	similarity score (1–5)
	Natural language inference (NLI) tasks using sentence pairs		
MNLI-m	NLI (in-domain)	accuracy	entailment / contradiction / neutral
MNLI-mm	NLI (cross-domain)		
QNLI	NLI (question-answering)	accuracy	entailment / not entailment
RTE	NLI	accuracy	entailment / not entailment
WNLI	NLI (coreference)	accuracy	entailment / not entailment

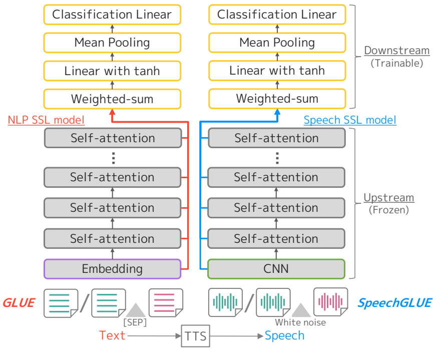

Figure 1: *Schematic diagrams of GLUE and SpeechGLUE.*

(e.g., log-mel filterbank output) and achieve close to the performance of the NLP SSL models, especially in sentence similarity and natural language inference (NLI) tasks. The experiments confirm that SSL models can capture enough linguistic information to tackle purely NLU tasks. By releasing the SpeechGLUE task, we hope to not only clarify the linguistic capabilities of current SSL models, but also to allow future models to be assessed in terms of their improvements in these tasks. Indeed, we believe that SLU tasks, which require capturing fine linguistic information, will be more and more important in future speech processing research.

2. Related work

There are several speech benchmarks related to the current work. ASR-GLUE [25] is a collection of human speech recordings based on selected sentences intended for some of the GLUE tasks. This benchmark includes only development and test sets to evaluate the negative impact of ASR error propagating to the backend NLU system. Therefore, we cannot train downstream models on this dataset, making it unsuitable for our purpose. Other datasets such as [6, 26] have been helpful in benchmarking speech SSL models. While they are designed to evaluate the generalizability through diverse speech processing tasks, SpeechGLUE, a collection of purely NLU tasks based on GLUE, aims to delve into the linguistic properties.

3. Method

In this section, we briefly explain GLUE [22] tasks in Section 3.1 and then, introduce SpeechGLUE in Section 3.2.

3.1. GLUE

The original GLUE benchmark contains 9 tasks divided into 3 categories: 1) the Corpus of Linguistic Acceptability (CoLA) [27] and the Stanford Sentiment Treebank (SST-2) [28] for single-sentence tasks, 2) the Microsoft Research Paraphrase Corpus (MRPC) [29], Quora Question Pairs (QQP) [30] and the Semantic Textual Similarity Benchmark (STS-B) [31] for similarity and paraphrase tasks, and 3) Multi-Genre NLI (MNLI) [32], Question-answering NLI (QNLI) [33], Recognizing Textual Entailment (RTE) [34–37] and Winograd NLI (WNLI) [38] for NLI tasks. An overview of each task is given in Table 1.

3.2. SpeechGLUE

As described in Section 3.1, the GLUE benchmark was originally designed to assess NLU systems. Since our objective is to evaluate speech SSL models in terms of NLP capability, the text sentences must be converted into corresponding speech utterances. In this work, we applied a single-speaker TTS system to investigate purely linguistic knowledge of speech SSL by suppressing acoustic variabilities such as speaker, speaking style, and recording noise. Recent neural-based TTS systems can generate high-quality speech in terms of naturalness and intelligibility [39]. Moreover, the conversion cost via TTS is lower than rerecording the text by humans.

The overall system is summarized in Figure 1. Since SpeechGLUE is essentially a counterpart of GLUE, speech and NLP SSL models are fairly comparable except for the `[SEP]` token utilized in NLP SSL models. `[SEP]` token is a special separation token that indicates where to split pairs in the examples (e.g., pairs of question-answer). For speech SSL models, this study simply employs a white noise signal of 50 ms as an alternative.

To incorporate SSL upstream models, we utilize the pre-trained model as a feature extractor and do not update the parameters during the training on GLUE/SpeechGLUE tasks in order to verify just the linguistic representation captured only by SSL.

4. Experimental Setup

4.1. SpeechGLUE dataset

For the GLUE benchmark itself, we utilized the publicly available dataset[3] provided by Hugging Face. To generate the speech, we adopted the VITS [40] model[4] trained by LJSpeech[5] using the ESPnet toolkit [41]. Because the VITS model was trained with a sampling frequency of 22050 Hz, we resampled the output data to 16000 Hz to match the sampling frequency assumed by the SSL models. The dataset after applying TTS is summarized in Table 2. Note that the original number of examples in the test set of QQP and in the training set of SST2 were 390965 and 67349, but the sizes were reduced to 390963 and 67347 through the execution of TTS. This is because some samples were deemed impractical as they included a huge number of digits or only null text, which could not be synthesized into speech. In addition, the original text samples were altered with the ESPnet-based text normalization such as by removing symbols (e.g., quotation marks) and by translating Latin abbreviations (e.g., "*i.e.*") into English (e.g., *"that is"*). Thus, the word sequence of the GLUE benchmark was also transformed for a fair comparison; nevertheless, no significant degradation was noted in our preliminary GLUE experiment.

4.2. Upstream

To explore the linguistic ability of speech SSL models, we utilized the multiple baselines and SSL methods with speech and text modalities summarized in Table 3. In this table, the upstream components were divided into three sections: the baselines, the speech SSL models, and the NLP SSL models. For the

[3] `https://huggingface.co/datasets/glue`
[4] `https://huggingface.co/espnet/kan-bayashi_ljspeech_vits`
[5] `https://keithito.com/LJ-Speech-Dataset`

Table 2: *Summary of the data size in SpeechGLUE. The tasks with underline indicate the relatively high-resource tasks. Note that the number of examples in some tasks is decreased from the original number as noted in Section 4.1.*

Corpus	#hours (#examples)		
	Training	Development	Test
Single-sentence tasks			
CoLA	6.3 (8551)	0.8 (1043)	0.8 (1063)
SST2	66.5 (67347)	1.6 (872)	3.4 (1821)
Similarity and paraphrase tasks (first / second sentence)			
MRPC	8.4 / 8.4 (3668)	0.9 / 0.9 (408)	3.9 / 3.9 (1725)
QQP	399.3 / 404.0 (363846)	44.4 / 44.9 (40430)	438.8 / 437.4 (390963)
STS-B	6.6 / 6.6 (5749)	1.9 / 1.8 (1500)	1.5 / 1.5 (1379)
Natural language inference (NLI) tasks (first / second sentence)			
MNLI-m	811.2 / 399.7 (392702)	19.9 / 10.0 (9815)	20.1 / 9.9 (9796)
MNLI-mm		21.0 / 11.1 (9832)	21.1 / 11.1 (9847)
QNLI	111.0 / 332.9 (104743)	5.8 / 17.8 (5463)	5.8 / 18.0 (5463)
RTE	12.7 / 2.6 (2490)	1.4 / 0.3 (277)	14.4 / 3.1 (3000)
WNLI	1.2 / 0.5 (635)	0.1 / 0.1 (71)	0.5 / 0.2 (146)

Table 3: *Overview of upstream models.* `LS`*,* `LL`*,* `GS`*,* `VP`*,* `MLS`*,* `CV`*,* `VL`*,* `BBL`*,* `BC` *and* `EW` *denote LibriSpeech, Libri-Light, GigaSpeech, VoxPopuli, Multilingual LibriSpeech, CommonVoice, VoxLingua107, BABEL, BookCorpus, and English Wikipedia, respectively. Note that the* `VP` *utilized in WavLMs is the subset of only English data.*

Upstreams	#Params	Input	Unlabeled data (#hours or #words)
FBANK	-	waveform	-
w/o SSL Large	315M	waveform	-
Phoneme	0.01M	text	-
wav2vec2.0 Base [2]	94M	waveform	LS (960 hours)
wav2vec2.0 Large [2]	315M	waveform	LL (60k hours)
HuBERT Base [3]	94M	waveform	LS (960 hours)
HuBERT Large [3]	315M	waveform	LL (60k hours)
data2vec-s Base [4]	94M	waveform	LS (960 hours)
data2vec-s Large [4]	315M	waveform	LL (60k hours)
WavLM Base [5]	94M	waveform	LS (960 hours)
WavLM Base+ [5]	94M	waveform	LL + GS + VP (94k hours)
WavLM Large [5]	315M	waveform	LL + GS + VP (94k hours)
XLS-R (0.3B) [42]	315M	waveform	VP + MLS + CV + VL + BBL (436k hours)
data2vec-t Base [4]	125M	text	BC + EW (3300M words)
BERT Base [21]	110M	text	BC + EW (3300M words)
BERT Large [21]	340M	text	BC + EW (3300M words)

model architecture, the encoder of all SSL models with Base (Large) structure consisted of 12 (24) Transformer blocks with 768-dim (1024-dim) embeddings, 3072-dim (4096-dim) feed-forward networks, and attention heads of 12 (16). As explained in Section 4.1, the parameters of all SSL models were frozen during training.

For the baseline models listed in the first section in Table 3, we adopted three types of upstream feature extractors. FBANK was the 80-dimensional log-mel filterbank output combined with delta and delta-delta features. We also evaluated a randomly initialized model with Large architecture (i.e., w/o SSL Large; see the second row of Table 3). We adopted this model to test the extent to which SpeechGLUE could be accurately handled by the structure itself since the Transformer architecture can inherently access long-range context. This model had input of raw waveforms, and hence, a feature encoder with subsampling was added before the Transformer blocks. The architecture of the feature encoder was identical to that of an existing SSL study [2] and comprised a 7-layer convolutional neural network (CNN). The third baseline was grapheme-to-phoneme (G2P) conversion followed by a 128-dim embedding layer to investigate the performance of ideal speech units without higher-level context. In this paper, since we utilized ESPnet for the TTS system as described in Section 4.1, the converter was in accordance with the G2P[6] utilized inside ESPnet.

As the speech SSL models, we evaluated publicly-available models with a combination of four SSL approaches and two model sizes as shown in the second section in Table 3. Specifically, we employed wav2vec2.0 [2], HuBERT [3], data2vec-s [4] and WavLM [5] for the SSL method, and Base and Large for the model size. These models passed raw waveforms to a 7-layer CNN before the Transformer encoder as the baseline model of w/o SSL Large. Note that data2vec-s and data2vec-t, described below, were the speech and NLP versions of data2vec, respectively.

We evaluated not only speech SSL models but also NLP SSL models in this paper. Since the NLP SSL models were specialized to obtain language representation from large unlabeled texts, we treated the results as the performance upper bound in SpeechGLUE tasks. The vocabulary size of data2vec-t and BERT were 50265 and 30522 subwords, resulting in the number of parameters being different. For BERT models, segment embedding, which encodes which of the two sentences contains the subword, was always set to zero; because the embedding was not used by the speech SSL models and data2vec-t, and was disabled for a fair comparison. In addition, recent research [43] and our preliminary experiment reported no significant difference in performance with and without the embedding. When evaluating the NLP SSL models, we utilized the modified version of the GLUE benchmark due to text normalization for TTS as explained in Section 4.1. Moreover, the normalized text was composed of lowercase, and hence, we utilized the uncased BERT models.[7]

4.3. Downstream

To benchmark the upstream models on the SpeechGLUE tasks, the downstream model was straightforwardly connected to the backend of the upstream models. The parameters of downstream models were updated during training unlike upstream models, which acted as feature extractors. The architecture of the downstream model, as motivated by an existing study [6], consisted of the weighted-sum of all hidden layers of upstream models followed by a 256-dim linear layer with tanh function, mean-pooling across whole sequences, and a final linear layer for classification or regression task. With respect to the last linear layer, the number of classes was 2 except for MNLI and STS-B, 3 for MNLI, and 1 for STS-B to perform the regression task. We applied a dropout with probability of 0.1 to the output of tanh function. The downstream model structure is basically identical regardless of the upstream model and tasks, while the weighted-sum was not utilized for upstream models without multiple layers (i.e., FBANK and phoneme).

For optimization, we adopted Adam with a learning rate of 3×10^{-4} and a batch size of 32. Two types of total training steps were used depending on the amount of training data: 50k steps for low-resource tasks and 150k steps for high-resource tasks, i.e., tasks underlined in Table 2. For the loss function, the models were updated using cross-entropy loss for all tasks except STS-B which used a mean squared error loss.

The entire system was evaluated on the development set of low-resource (high-resource) tasks for every 1k (12.5k) steps, and only the highest performances are reported here. Note that the evaluations were conducted only on the development set and not on the private test set proceeding on the GLUE server since the goal of this study was to investigate whether or not linguistic information was acquired rather than to achieve state-of-the-art performance on the GLUE benchmark. Additionally, there were seemingly no clear differences in performance tendency across

[6] `https://github.com/Kyubyong/g2p`

[7] `https://huggingface.co/bert-base-uncased` and `https://huggingface.co/bert-large-uncased` for Base and Large architecture. Note that, for data2vec-t, we used the only publicly available case-sensitive model at `https://huggingface.co/facebook/data2vec-text-base`.

Table 4: *Evaluation result for each model and each task on the development set of SpeechGLUE and GLUE. Acc, MCC, PCC and SCC denote accuracy, Matthews, Pearson and Spearman correlation coefficients, respectively. The resulting score with bold font (underline) indicates the highest score among the speech (NLP) SSL models.*

Upstream group	Upstream model	**Single sentence**		**Similarity and paraphrase**			**Natural language inference (NLI)**				Avg.
		CoLA	SST-2	MRPC	QQP	STS-B	MNLI-m / -mm	QNLI	RTE	WNLI	w/o
		MCC	Acc	Acc (F1)	Acc (F1)	PCC (SCC)	Acc	Acc	Acc	Acc	WNLI
Baselines	Chance rate	-	50.9	68.4 (81.2)	63.2 (0.0)	-	35.4 / 35.2	50.5	52.7	56.3	-
	FBANK	3.3	64.6	70.8 (81.8)	70.8 (53.8)	12.4 (10.0)	39.9 / 40.4	57.9	52.7	43.7	45.9
	w/o SSL LARGE	6.6	55.0	68.4 (81.2)	64.1 (20.4)	8.5 (7.7)	35.1 / 35.0	54.9	53.4	62.0	42.3
	Phoneme	0.0	62.0	71.8 (82.4)	65.0 (35.5)	15.4 (14.8)	37.6 / 37.1	58.3	57.8	42.3	45.0
Speech SSL	wav2vec2.0 BASE [2]	5.5	65.8	71.6 (81.2)	71.2 (58.7)	45.8 (45.5)	42.9 / 43.7	63.7	56.7	36.6	51.9
	wav2vec2.0 LARGE [2]	0.0	73.3	72.5 (81.8)	76.1 (67.1)	58.1 (58.1)	47.4 / 49.3	73.8	56.0	**57.7**	56.3
	HuBERT BASE [3]	3.1	73.3	70.3 (81.4)	72.3 (60.5)	50.4 (50.9)	44.8 / 46.1	64.3	**57.4**	36.6	53.6
	HuBERT LARGE [3]	24.8	81.0	73.0 (82.1)	81.0 (72.8)	70.0 (70.5)	60.7 / 62.7	76.3	54.9	35.2	64.9
	data2vec-s BASE [4]	13.1	72.8	71.8 (81.7)	73.8 (62.2)	56.4 (56.8)	46.9 / 48.5	67.9	54.9	28.2	56.2
	data2vec-s LARGE [4]	20.4	77.8	73.8 (83.3)	74.6 (65.1)	59.1 (59.2)	53.5 / 55.2	71.6	54.2	47.9	60.0
	WavLM BASE [5]	5.8	71.6	71.8 (81.6)	73.3 (63.3)	69.4 (69.8)	47.6 / 48.6	71.0	**57.4**	38.0	57.4
	WavLM BASE+ [5]	6.9	74.5	72.8 (79.9)	75.3 (66.1)	74.3 (74.5)	49.2 / 50.0	73.8	54.5	40.8	59.0
	WavLM LARGE [5]	**29.6**	**82.7**	**75.7 (83.0)**	**83.3 (76.8)**	**79.5 (79.7)**	**63.8 / 65.5**	**80.6**	52.0	35.2	**68.1**
	XLS-R (0.3B) [42]	7.2	74.5	71.1 (81.2)	78.6 (69.8)	69.1 (69.2)	54.5 / 55.9	74.3	56.0	54.9	60.1
NLP SSL	data2vec-t BASE [4]	41.0	86.9	79.2 (84.8)	82.2 (76.7)	80.0 (80.2)	68.4 / 69.8	83.7	58.5	23.9	72.2
	BERT BASE [21]	49.0	90.5	77.2 (84.2)	85.4 (80.3)	82.8 (82.9)	69.1 / 70.2	84.4	53.1	15.5	73.5
	BERT LARGE [21]	51.4	90.3	76.7 (83.6)	85.2 (80.5)	82.8 (83.1)	70.4 / 71.0	85.0	53.4	14.1	74.0

tasks between the development and test sets from the previous NLP studies such as [44].

To validate that the synthesized speech was generated properly, speech SSL models further addressed ASR tasks by using the SpeechGLUE dataset. The training setup for the ASR task was the same as the settings of SUPERB [6] except for low-resource tasks. For low-resource tasks, i.e., tasks with no underline in Table 2, the total training steps were reduced to 50k and the evaluation by the development set was performed every 500 steps, in addition to changing the learning rate to 2×10^{-4}. With respect to high-resource tasks, we randomly selected a maximum of 100 hours of data from the training set as in SUPERB, where only *train-clean-100* from LibriSpeech was utilized. In line with this, we also randomly selected the development set to be a maximum of 5 hours and reported the best word error rates (WERs) on the development set only as well as for the setting noted above. From the ASR experiments, we confirmed that the average WER for all tasks except for WNLI[8] ranged from 13.1% to 18.2%. For example, WavLM LARGE, which yielded the best averaged score, was able to achieve WERs of less than 10% on most tasks.

5. Results

The experimental results of SpeechGLUE and GLUE are shown in Table 4. Note that, in calculating the average score, we excluded the score of the WNLI task due to its performance instability[9] caused by the extremely small number of examples as was also pointed out in a previous NLP paper [21]. From the results, we can find the overall performance tendency that the NLP SSL models attain the highest performance, followed by the speech SSL models, and finally the baselines, especially in the CoLA task, which requires grammatical knowledge mainly. Among speech SSL, as with the ASR task, WavLM LARGE demonstrated the best performance. Especially in some sentence similarity and NLI tasks, the performance was comparable to that of the NLP SSL models, suggesting that the speech SSL models can learn enough linguistic information to handle those tasks. By comparing WavLM BASE, BASE+ and LARGE, we observe that model capacity is more critical than the size of unlabeled data. However, XLS-R (0.3B), which is a multilingual model, almost matched the accuracy of WavLM BASE+ despite its larger data volume and size, indicating the language dependency of the SSL model [45]. The performance degradation of NLP SSL models in GLUE compared to the scores presented in the previous papers [4, 21] may be due to the limitation that the entire model was not fine-tuned to ensure a fair comparison.

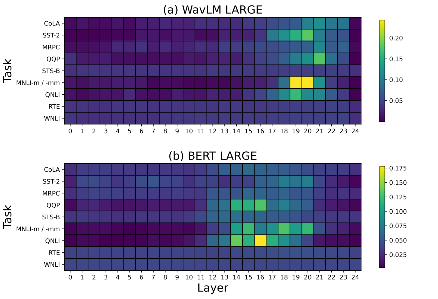

Figure 2: *The weights of weighted-sum. The 0th layer corresponds to the input to the 1st layer of the encoder.*

To investigate the contribution of features in each layer, Figure 2 depicts the weights of weighted-sum for WavLM LARGE and BERT LARGE on each task. The weights of WavLM are concentrated on the layers between 18 and 24, and the features in those latter layers seem to be important in performing non-speaker-related tasks [5, 18]. Compared with BERT, those layers are seemingly shifted somewhat later. It is noteworthy that it seems difficult to learn clear weights for some low-resource tasks, even with an NLP SSL model. Moreover, some tasks exploit information captured in multiple layers (e.g., SST-2 for WavLM), while others focus more on some layers (e.g., MNLI for WavLM and QNLI for BERT). However, this tendency is not consistent for WavLM and BERT. This seems to indicate that the layers capture very different information.

6. Conclusions

In this paper, we endeavored to uncover to what extent SSL models could capture language information through our probing tasks called SpeechGLUE. The speech SSL models performed better than chance and baselines, indicating that the pre-trained models capture some general linguistic knowledge from speech alone. However, compared to the top-line NLP SSL models, the performance is somewhat poor in some tasks and there seems to be room for improvement through, for example, unified speech-text SSLs. Future works contain further probing by other NLP benchmarks to analyze linguistic properties in more detail.

[8] Since the training set in WNLI contains only 635 samples, the WER was unstable.

[9] In our preliminary experiment with wav2vec2 BASE, changing the random seed resulted in a standard deviation of 4.8% in accuracy over the five trials, and the training itself was also unstable.

7. References

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, 2021.

[4] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," *ICML*, 2022.

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *JSTSP*, 2022.

[6] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[7] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi, X. Chang, P. Hall, H.-J. Chen, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for semantic and generative capabilities," in *ACL*, 2022.

[8] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *JSTSP*, 2022.

[9] S. Ren, S. Liu, Y. Wu, L. Zhou, and F. Wei, "Speech pre-training with acoustic piece," in *Interspeech*, 2022.

[10] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, "Deep versus wide: an analysis of student architectures for task-agnostic knowledge distillation of self-supervised speech models," in *Interspeech*, 2022.

[11] S. Arora, S. Dalmia, P. Denisov, X. Chang, Y. Ueda, Y. Peng, Y. Zhang, S. Kumar, K. Ganesan, B. Yan, N. Thang Vu, A. W. Black, and S. Watanabe, "ESPnet-SLU: Advancing spoken language understanding through ESPnet," in *ICASSP*, 2022.

[12] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech," in *ICASSP*, 2022.

[13] Y. Peng, S. Arora, Y. Higuchi, Y. Ueda, S. Kumar, K. Ganesan, S. Dalmia, X. Chang, and S. Watanabe, "A study on the integration of pre-trained SSL, ASR, LM and SLU models for spoken language understanding," in *SLT*, 2022.

[14] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeurIPS SAS*, 2020.

[15] E. Dunbar, M. Bernard, N. Hamilakis, T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, and E. Dupoux, "The Zero Resource Speech Challenge 2021: Spoken language modelling," in *Interspeech*, 2021.

[16] T. A. Nguyen, B. Sagot, and E. Dupoux, "Are discrete units necessary for spoken language modeling?" *JSTSP*, 2022.

[17] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *ASRU*, 2021.

[18] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP*, 2023.

[19] A. Bapna, Y.-a. Chung, N. Wu, A. Gulati, Y. Jia, J. H. Clark, M. Johnson, J. Riesa, A. Conneau, and Y. Zhang, "SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training," *arXiv preprint arXiv:2110.10329*, 2021.

[20] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "mSLAM: Massively multilingual joint pre-training for speech and text," *arXiv preprint arXiv:2202.01374*, 2022.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *NAACL*, 2019.

[22] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *ICLR*, 2019.

[23] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," in *NeurIPS*, 2019.

[24] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *arXiv preprint arXiv:2206.04615*, 2022.

[25] L. Feng, J. Yu, D. Cai, S. Liu, H. Zheng, and Y. Wang, "ASR-GLUE: A new multi-task benchmark for asr-robust natural language understanding," *arXiv preprint arXiv:2108.13048*, 2021.

[26] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," in *Interspeech*, 2020.

[27] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *TACL*, 2019.

[28] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *EMNLP*, 2013.

[29] W. B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *IWP*, 2005.

[30] S. Iyer, N. Dandekar, and K. Csernai, "First quora dataset release: Question pairs," https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs, 2017.

[31] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *SemEval-2017*, 2017.

[32] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in *NAACL*, 2018.

[33] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *EMNLP*, 2016.

[34] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *MLCW*, 2005.

[35] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, "The second PASCAL recognising textual entailment challenge," in *Proceedings of the second PASCAL challenges workshop on recognising textual entailment.*, 2006.

[36] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The third PASCAL recognizing textual entailment challenge," in *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007.

[37] L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo, "The fifth PASCAL recognizing textual entailment challenge," in *TAC*, 2009.

[38] H. J. Levesque, E. Davis, and L. Morgenstern, "The Winograd schema challenge," in *KR*, 2012.

[39] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[40] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, 2021.

[41] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 2018.

[42] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech*, 2022.

[43] R. van der Goot, M. Müller-Eberstein, and B. Plank, "Frustratingly easy performance improvements for low-resource setups: A tale on BERT and segment embeddings," in *LREC*, 2022.

[44] Y. Meng, C. Xiong, P. Bajaj, saurabh tiwary, P. N. Bennett, J. Han, and X. Song, "COCO-LM: Correcting and contrasting text sequences for language model pretraining," in *NeurIPS*, 2021.

[45] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, "Exploration of language dependency for japanese self-supervised speech representation models," in *ICASSP*, 2023.