SpeechGLUE: How Well Can Self-Supervised Speech Models Capture Linguistic Knowledge?
Takanori Ashihara, Takefumi 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. 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.¹

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.²

¹Note that our approach can be applied to any probing task of NLP.
²We release SpeechGLUE for reproducibility and comparison with successive SSL techniques at https://github.com/ashi-ta/speechGLUE.
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 pretrained 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 provided by Hugging Face. To generate the speech, we adopted the VITS [40] model trained by LJSpeech 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 baseline, the speech SSL models, and the NLP SSL models. For the
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.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>training #hours</th>
<th>development #hours</th>
<th>test #hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>6.3 (8551)</td>
<td>0.4 (1044)</td>
<td>0.8 (1060)</td>
</tr>
<tr>
<td>SST2</td>
<td>66.6 (5974)</td>
<td>1.6 (1822)</td>
<td>3.4 (1821)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Upstream</th>
<th>#Params</th>
<th>Input</th>
<th>Unlabeled data (hours or #words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBANK</td>
<td>315M</td>
<td>waveform</td>
<td>-</td>
</tr>
<tr>
<td>w/o SSL</td>
<td>315M</td>
<td>waveform</td>
<td>-</td>
</tr>
<tr>
<td>Phoneme</td>
<td>0.1M</td>
<td>text</td>
<td>-</td>
</tr>
<tr>
<td>wav2vec2.0 Base [2]</td>
<td>94M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>wav2vec2.0 Large [2]</td>
<td>315M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>HuBERT Base [3]</td>
<td>94M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>HuBERT Large [3]</td>
<td>315M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>data2vec-v [4]</td>
<td>94M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>data2vec-large [4]</td>
<td>315M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>wav2vec [5]</td>
<td>94M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>w/o SSL Base [5]</td>
<td>94M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>w/o SSL Large [5]</td>
<td>315M</td>
<td>waveform</td>
<td>12 (900 hours)</td>
</tr>
<tr>
<td>XLS-R 10B [42]</td>
<td>315M</td>
<td>waveform, Voiced + Unvoiced, C V C + V C L, 315M (35k hours)</td>
<td></td>
</tr>
<tr>
<td>data2vec-te [4]</td>
<td>125M</td>
<td>text</td>
<td>12 x (10M words)</td>
</tr>
<tr>
<td>BERT Base [21]</td>
<td>315M</td>
<td>text</td>
<td>12 x (300M words)</td>
</tr>
<tr>
<td>BERT Large [21]</td>
<td>340M</td>
<td>text</td>
<td>12 x (300M words)</td>
</tr>
</tbody>
</table>

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) with 768-dim embeddings. The third baseline was a grapheme-to-phone (G2P) conversion followed by a 128-dim embedding layer to investigate the performance of ideal speech units without output 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 utilized inside ESPnet.

The 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) with 768-dim embeddings. The third baseline was a grapheme-to-phone (G2P) conversion followed by a 128-dim embedding layer to investigate the performance of ideal speech units without output 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 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 [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 x 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

7https://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.

<table>
<thead>
<tr>
<th>Upstream group</th>
<th>Upstream model</th>
<th>Single sentence</th>
<th>Similarity and paraphrase</th>
<th>Natural language inference (NL)</th>
<th>Avg. w/o WNL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>UCL-A</td>
<td>SST-2</td>
<td>MRPC</td>
<td>QQP</td>
</tr>
<tr>
<td>Baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chance rate</td>
<td></td>
<td>50.9</td>
<td>66.4</td>
<td>81.2</td>
<td>83.2</td>
</tr>
<tr>
<td>FRANK</td>
<td></td>
<td>52.7</td>
<td>68.0</td>
<td>81.8</td>
<td>83.0</td>
</tr>
<tr>
<td>w/o SSL LARGE</td>
<td></td>
<td>66.4</td>
<td>68.0</td>
<td>81.8</td>
<td>83.0</td>
</tr>
<tr>
<td>Phoneme</td>
<td></td>
<td>55.0</td>
<td>68.4</td>
<td>81.2</td>
<td>84.0</td>
</tr>
</tbody>
</table>

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 caused by the extremely small number of examples as well as 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 multilin

6. Conclusions

In this paper, we endeavored to uncover to what extent SSL models could capture language information through their 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.
7. References


