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

Assignment 3
CSL7770: Speech Understanding
AY 2024-25, Semester – II

REPORT

Under the guidance of
Prof. Richa Singh

Submitted by

Jyotishman Das
M24CSA013

Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur

April 2025

Paper Review

Title of the Paper:
SpeechGLUE: How Well Can Self-Supervised Speech Models Capture Linguistic Knowledge?

Summary of the Paper

This paper introduces **SpeechGLUE**, a speech-based version of the popular NLP benchmark GLUE. Evaluating the language ability of self-supervised learning (SSL) models trained on speech data is its main goal. The authors evaluate the ability of SSL models such as wav2vec 2.0, HuBERT, and WavLM to capture syntax, semantics, and other linguistic aspects without the use of text by converting GLUE inputs into voice using top-notch TTS (text-to-speech) systems. The findings indicate that SSL speech models have a great deal of promise for comprehending linguistic information since models like WaveLM Large outperform baselines and even come close to text-based models like BERT on several tasks.

Main Architecture (Figure)

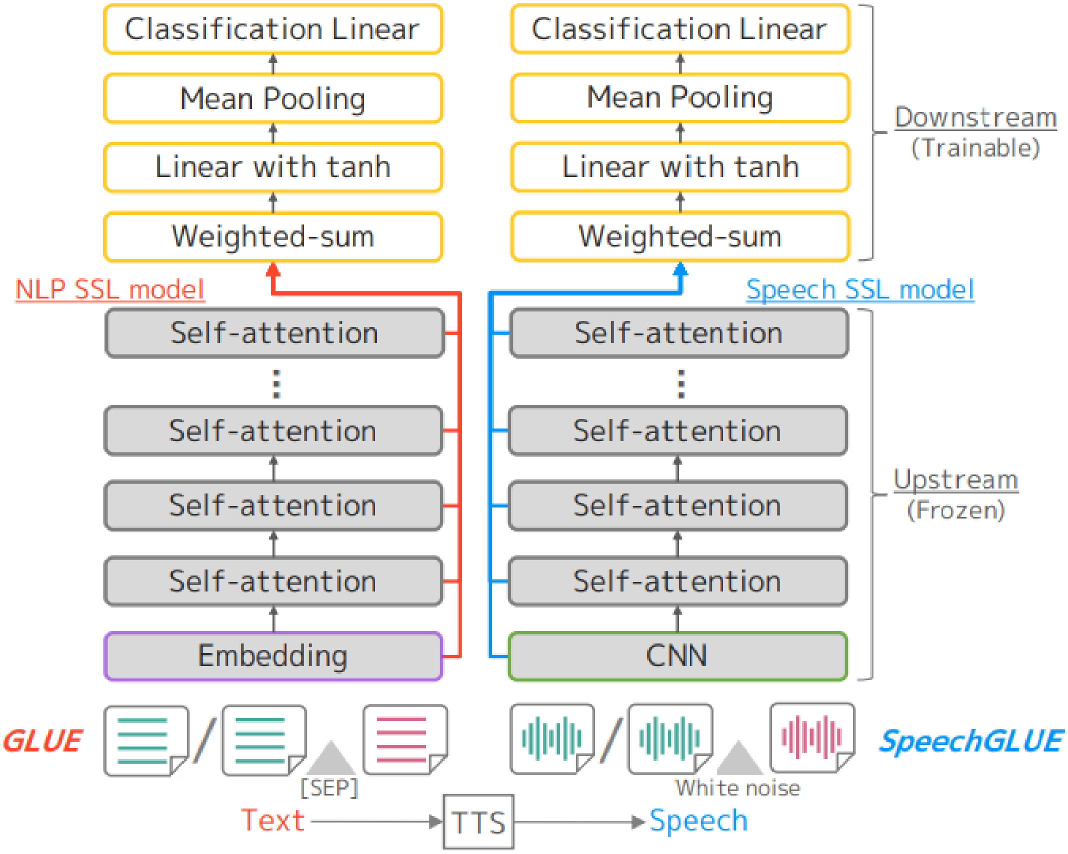

Figure 1: Overview of GLUE (text-based) vs SpeechGLUE (speech-based) pipeline.

Technical Strengths

- Presents a novel benchmark designed specifically for speech-based NLU assessment.
- Covers a broad range of tasks, such as inference, sentiment analysis, acceptability, and paraphrasing.
- By simply training the downstream classifier and freezing upstream SSL models, it guarantees fair comparison.
- Reveals the embedded locations of linguistic knowledge through layer-by-layer analysis.

Technical Weaknesses

- Top-performing NLP models, such as BERT, continue to outperform speech SSL models.
- Uses synthetic speech (TTS), which is devoid of the prosody and natural variation seen in natural speech.
- Because there are so few samples, the WNLI task exhibits instability.

Minor Questions / Minor Weaknesses

- What would happen if actual human speech recordings were used rather than sounds produced by TTS?
- Would SpeechGLUE's multilingual or cross-lingual versions provide more in-depth information?
- What impact do various TTS systems or noise augmentations have?

Reviewer Suggestions

Future research should think about using human-recorded datasets and broadening the scope of language tasks in order to improve SpeechGLUE's resilience and realism. The performance difference with NLP models might be closed by investigating speech-text combined SSL models or optimizing the SSL encoders. To improve generalization and relevance to actual SLU applications, the authors might potentially test code-switching scenarios or multilingual data.

Rating and Justification

Rating: 8/10

A current and practical standard for assessing linguistic understanding in speech SSL models is provided in this study. Although the results are encouraging, the reliance on artificial audio restricts realism and performance still falls short of text-based models. It does, however, lay a solid basis for further investigation.

Bonus Question: Reproduction and DoRA-Based Fine-Tuning

Part II(i): Reproduction on SST-2 and MRPC

We reproduced the evaluation pipeline for the paper on two SpeechGLUE benchmark datasets:

- **SST-2:** Sentiment classification
- **MRPC:** Paraphrase detection

We used the HuggingFace datasets and converted textual inputs into audio using `gTTS`. Audio features were extracted using the Wav2Vec2 Base model, and a shallow classifier was trained over the extracted features. The model achieved:

- **SST-2 Accuracy:** 54.6%
- **MRPC Accuracy:** 69.2%

Part II(ii): DoRA Fine-Tuning on SNIPS Dataset

We selected the SNIPS intent classification dataset as our third external dataset and trained a DoRA (Decomposed Rank-Adaptive Adapter)-based classifier. The training setup included:

- **Backbone:** Wav2Vec2 frozen encoder
- **DoRA Layer:** Rank 16 with 256-dimensional intermediate representation
- **Dropout:** 0.3, **Optimizer:** AdamW
- **Training:** 20 epochs on 80% of SNIPS

The final classification accuracy achieved on the SNIPS test set was **35.5%**.

We then reused the trained encoder and attached separate 2-class classifier heads to SST-2 and MRPC. These heads were trained for 10 epochs, achieving the following transfer results:

- **SST-2 Transfer Accuracy:** 54.6%
- **MRPC Transfer Accuracy:** 69.2%

These results demonstrate the ability of DoRA-trained models to generalize across speech understanding tasks using lightweight head adaptation. This reinforces the paper's findings on SSL model transferability in the speech domain.

Submission Links

Google Colab Notebook:

GitHub Repository: