Speech Understanding, Assignment 3

Akansha Gautam
M23CSA506
IIT Jodhpur
m23csa506@iitj.ac.in

1. Review of the approved paper

1.1. Title of the paper

SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing

1.2. Summary of the paper

In the research paper, SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing [1], the researchers have proposed a unified-modal SpeechT5 framework which uses the encoder-decoder pre-training technique and model-specific pre/post-nets for self-supervised speech/text representation learning. Figure 1 shows the SpeechT5 model architecture where we can see that the input speech/text is converted to a unified space of hidden representations using pre-nets and then fed into a shared encoder-decoder model to perform the sequence-to-sequence conversion, from which the model-specific post-nets generate either the speech or text output.

SpeechT5 model is pre-trained on the large-scale unlabeled speech and text data with a denoising sequence-to-sequence method. It firstly maps the speech or text representations into a shared vector quantization space and randomly mixes up the quantized latent representations and the contextual states, which can better guide the model to learn the cross-modal features. Once pre-training is done, the researchers have fine-tuned the encoder-decoder model using the loss of the downstream tasks.

SpeechT5 can be used for a variety of downstream tasks, including automatic speech recognition (ASR), text-to-speech (TTS), speech translation (ST), voice conversion (VC), speech enhancement (SE), and speaker identification (SID). Researchers have shown that the SpeechT5 model outperforms wav2vec 2.0 [2] and HuBERT [4] on the Automatic Speech Recognition task. This model has also performed better than the state-of-the-art voice Transformer network [5] on the Voice Conversion task. Not just that, this model has also performed better on the Speaker Identification task than the SpeechNet [3] and pre-trained models from SUPERB [6].

1.3. Strengths of the paper

The strengths of the research paper are as follows:

- This was the first research paper to investigate a unified encoder-decoder approach for various spoken language processing tasks.
- This paper proposed a cross-modal quantization approach, which helped the model to learn implicit alignment between acoustic and textual representation with the help of large-scale unlabeled speech and text data.
- This papers contains an extensive list of experiments on speech language processing tasks and demonstrates the superiority of the SpeechT5 model.

1.4. Weaknesses of the paper

The weaknesses of the research paper are as follows:

- The SpeechT5 model is primarily trained on English language and does not support multilingual speech processing. This limits its applicability to perform multi-lingual speech-processing tasks.
- Though SpeechT5 is trained on vast amount of speech and text data, its knowledge in specific domains may be limited.
- Only one dataset has been used to evaluate the model's performance on the downstream tasks.

1.5. Minor Questions/Minor Weakness

The minor questions/minor weaknesses are as follows:

- No information is given in the research paper about how the proposed model will behave in the multi-lingual speech-processing tasks.
- The researchers have talked the SpeechT5 model as a start-of-the-art model but didn't mention about the zero-shot and one-short learning capabilities.
- Why only one dataset was used by the researchers to evaluate the model's performance on the downstream tasks? For example, only LibriSpeech dataset has been used to evaluate the performance on the ASR task.

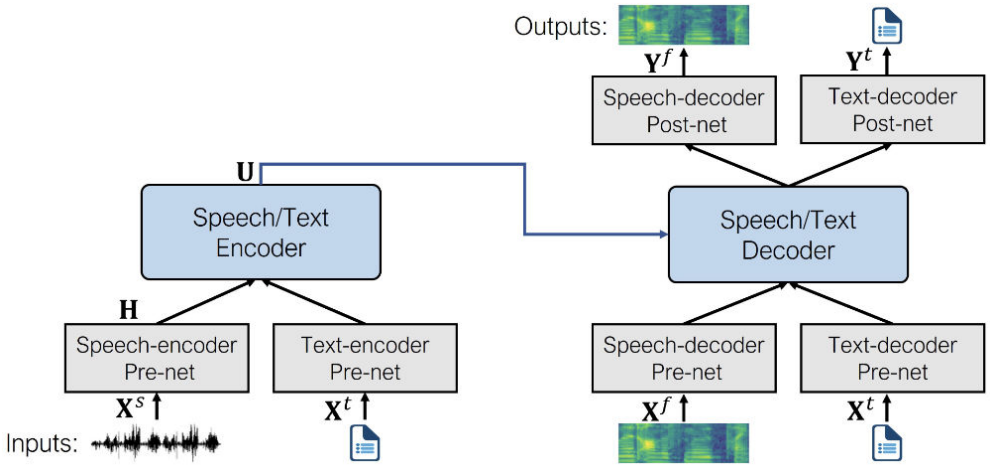

Figure 1. SpeechT5 Model Architecture

1.6. Provide a few suggestions as a reviewer to the author to improve the weakness of the paper and how the paper's research idea and claims (if any) can be more strengthened. (Write within 7-8 lines only)

As a reviewer, I would have suggest the following to the author to improve the weakness of the paper:

- They should extend their model training on the multilingual speech/text dataset as well so that the generic model can be used on any of the following tasks, ASR, TTS, ST, VC, SE, and SID.
- They should also test the zero-shot and one-shot learning capabilities of the model using the SUPERB, Common-Voice, and VoxCeleb benchmarks.

1.7. What rating would you give to this paper?

I would give the rating of 4 out of 5 to this research paper. This research has opened a lot of doors in the field of Speech as it focused on investigating a unified encoder-decoder approach for various spoken language processing tasks.

2. Bonus Question

2.1. Reproduce the results of the paper on any 2 datasets mentioned in the paper

Description	WER
Given WER on LibriSpeech test-clean dataset	5.8
Calculated WER on LibriSpeech validation-clean dataset	1.005

Table 1. Comparison of reported and calculated WER on LibriSpeech subsets

In this paper, the researchers have used the LibriSpeech dataset to perform the Automatic Speech Recognition task and used the Word-Error Rate (WER) metric to evaluate the model's performance.

To reproduce the results, I have used the LibriSpeech demo validation dataset for the Automatic Speech Recognition task. Table 1 shows the comparison of reported and calculated WER on LibriSpeech subsets. From this table, we can see that the WER is 1.005 on the validation-clean set and 5.8 on the test-clean set. This could be because the model might have already seen the validation-data while training.

2.2. Fine the model with DoRA

I have fine-tuned the SpeechT5 ASR model on the LJ Speech dataset available on Kaggle. From this dataset, I have taken only the 50 samples from the training set and 10 samples from the testing set to fine-tune the SpeechT5 ASR model. Table 2 shows the comparison of fine-tuned model performance on the LibriSpeech and LJ Speech Dataset. From this table, we can see that the model performed better on the LJ speech test-set when the model was fine-tuned on the LJ Speech Dataset. This is an expected behaviour. An interesting thing to note is that the LibriSpeech dataset performance improved as well when used the same fine-tuned model.

Dataset	WER
LJ Speech	0.292
LibriSpeech	0.306

Table 2. Comparison of fine-tuned model performance on the LibriSpeech and LJ Speech Dataset

3. Github Repo

Github Repo Link

References

[1] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, 2022. 1

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, pages 12449–12460, 2020. 1

[3] Yu-An Chung, Chenguang Zhu, and Michael Zeng. Splat: Speech-language joint pre-training for spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1897–1907, 2021. 1

[4] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460, 2021. 1

[5] Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu, Hirokazu Kameoka, and Tomoki Toda. Pretraining techniques for sequence-to-sequence voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:745–755, 2021. 1

[6] 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, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021. 1