Training Report – Results, Evaluation and Future works

[Part I] Comparison of Model Performance (Pre-trained base model vs Fine-tuned models)

For our model evaluation on Automatic Speech Recognition (ASR) tasks, the selected evaluation metric is Word Error Rate (WER). Even though Character Error Rate (CER) penalizes minor spelling errors much less as compared to WER, WER evaluates the system's ability to learn more about the context of predictions (in English language). Thus, is more widely used to access the performance of speech systems.

We evaluated the fine-tuned "wav2vec2-large-960h" ASR model's inference capabilities against the performance of the pre-trained "wav2vec2-large-960h" baseline model development set (cv-valid-dev). Key dataset features and results are displayed in **Table 1**.

Model name	Type	Dataset	Dataset size	WER score
"wav2vec2-large-960h"	Pre-trained base model	cv-valid-dev	4,076	10.8%
"wav2vec2-large-960h"	Fine-tuned (6,300 files)	cv-valid-dev	4,076	7.7%
"wav2vec2-large-960h"	Fine-tuned (2,000 files)	cv-valid-test	3,995	12.0%
"wav2vec2-large-960h"	Fine-tuned (6,300 files)	cv-valid-test	3,995	7.3%

Table 1: Comparison on pre-trained base model vs fine-tuned model on development set

WER from using pre-trained "wav2vec2-large-960h" model (without fine-tuning) was approximately 10.8% while WER using fine-tuned "wav2vec2-large-960h" model was 3-percentage points lower at 7.7%. A better performance attributed to model fine-tuning can be attributed to better alignment to domain-specific data (common voice datasets), i.e. being able to capture the dataset's unique nuances like accent, gender, age and noise distribution. A key feature to speech variability is identified to be "accent". We found that "accent" distributions across training and test sets were consistent, possibly explaining an improved fine-tuned performance.

Following model inference on the development set, we observed the distribution of WER metrics across our key feature "accent" and compare our two models (refer to Figure 1 below).

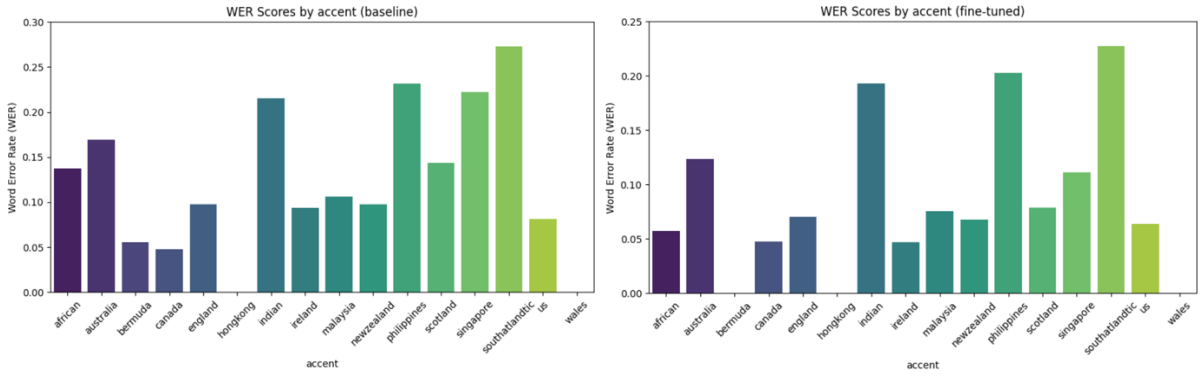

Figure 1: WER scores by "accent" – Baseline test (pre-trained wav2vec2-large-960h) vs Fine-tuned model (6,300 records)

From Figure 1, we observed that the baseline model performs well on speech/audio data from regions like the US, Canada and England. The key question now is: How does fine-tuning affect performance across regions? Our fine-tuned model shows improvements in WER scores across most other regions, indicating successful accent mapping. Notably, countries like Singapore and Africa recorded strong improvements while countries like Philippines and India shows less improvements. This could be due to unique speech nuances and pronunciations and more work needs to be done to explore potential solutions.

[Part II] Propose series of steps, including datasets and experiments to improve accuracy of fine-tuned wav2vec2 model

1. Dataset Diversification and Augmentations

Papers have shown that audio augmentation strategies has led to minor improvements in evaluation scores. In particular (Ko et.al., 2015) demonstrated the benefits of speech perturbations on model performance. Hence, exploring other strategies like speech perturbations, time masking, pitch shift and background noise injection might be beneficial in contributing to a more diverse training dataset, which could be crucial in improvements in model's generalisability to unique accents like those in India or the Philipines.

2. Integrating External Language Models for enhanced performance.

Leveraging Large Language Models (LLMs) for speech recognition is another feasible solution to improve fine-tuning evaluation accuracy. This post-processing strategy (after acoustic model decoding) involves integrating a transformer-based LLM decoder to perform speech recognition as next token prediction (Hono et.al., 2023). In the context of HuggingFace processors, we can implement one with a decoder that includes an Language Model such as "Facebook/wav2vec2-large-960h-lv60-self". As it was observed that there were some spelling mistakes contributing to error percentages in WER, these context-aware corrections and output re-ranking strategy could potentially improve WER accuracy in speech transcriptions after model fine-tuning.

3. Hyperparameter Tuning and Fine-tune Model over entire "cv-valid-train" Dataset (195,776 records)

Our current approach used open-source past projects as reference points for hyperparameter settings. Some sources include HuggingFace articles (with example colab notebooks), Medium and "readthedocs" articles. In future experiments, we could incorporate methodologies such as random search or Bayesian optimisation to determine optimal hyperparameters for fine-tuning our wav2vec2 model.

Another key limitation of this project is compute and memory limitations. We were only able to fine-tune our pre-trained "wav2vec2-large-960h" model on 6,300 audio files. Therefore, if resources permit, utilizing a large dataset for fine-tuning, coupled with hyperparameter tuning

to optimize model training might improve overall evaluation performance of the pre-trained model, leading to more accurate inferencing results.

4. Exploration of Other Methodologies to Enhance Training Data Quality for Model Fine-tuning

Conventional strategies like dataset augmentation and the integration of external language models have been shown to improve model fine-tuning performance in WER scores. Inspired by Guo et. al., 2024, we recommend experimenting with a semi-supervised learning strategy where we utilise self-transcribed, high confidence data to supplement the training data pool for model fine-tuning. These transcribed data can be selected based on model confidence levels (eg. WER <= 0.3).

[Part III] Conclusion

The fine-tuning of the wav2vec2-large-960h model on the Common Voice dataset resulted in a notable WER improvement over the baseline model, demonstrating the benefits of domain adaptation. Specifically, fine-tuning allowed the model to better align with accent variations and speech patterns, leading to improved transcription accuracy across diverse regions. However, performance discrepancies across certain accents indicate areas for further refinement.

To further enhance inferencing accuracy, we propose a multi-faceted approach involving dataset diversification, augmentation techniques, integration of external language models, and hyperparameter tuning. Additionally, semi-supervised learning strategies could leverage high-confidence transcriptions to expand training data, reducing WER even further. By implementing these enhancements, we aim to develop a more robust and generalizable ASR model, capable of accurately transcribing speech across diverse linguistic and acoustic conditions.

References

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. Interspeech 2015. Retrieved from https://www.isca-archive.org/interspeech_2015/ko15_interspeech.pdf

Hono, S., Kanda, N., Yoshioka, T., Wu, C., Li, X., & Xiao, X. (2023). Transformer-based language models for speech recognition post-processing. arXiv preprint arXiv:2312.03668. Retrieved from https://arxiv.org/pdf/2312.03668

Guo, J., Liu, Z., Zhang, T., & Chen, C. L. P. (2024). Incremental self-training for semi-supervised learning. arXiv preprint arXiv:2404.12398. Retrieved from https://arxiv.org/abs/2404.12398