ASR - Fine-Tuning the Whisper Model

Alex Adams Jake Walper

April 30, 2024

Abstract

We studied the effect of fine-tuning the Whisper-Small ASR model with a multi-ethnic accent dataset. The resulting model was compared against the small-Whisper model using word error rate (WER) as a measurement tool to determine which model performed better. The fine-tuned model performed significantly better when comparing the WER for the same testing dataset, but had limitations that necessitate further research into whether supervised accented speech could be a viable training option when trying to enhance ASR models.

1 Introduction

Automatic Speech Recognition (ASR) was first introduced in 1952 when Bell labs created "Audrey," the first audio recognition device that could recognize spoken single digits from a single voice (1). Since then, ASR has become significantly more mainstream, with prominent companies like Apple and Amazon implementing models for their respective speech bots, Siri and Alexa. By 2030, it is reported that 90% of customer service interactions among Fortune 500 companies will include some sort of automatic speech recognition.

ASR recognizes speech patterns in a wide variety of individuals using two main methods; Traditional methods, which utilize Hidden Markov models and Gaussian Matrix models combined with CNN, and end-to-end deep learning, which makes use of CTC, LAS, and RNNT architectures (2),(3). The most state-of-the-art architecture is Transformers, which were developed by a group at google and have since become the staple architecture for ASR models, in addition to other applications such as DNA analysis, protein structure analysis, and Large language processing (4). The transformer architecture is ultimately what we used in our fine-tuned Whisper model.

2 Literature Review

The Whisper model was developed by a group at OpenAI with the intention of developing an ASR model trained only on supervised data (5). The group compiled 680,000 hours of trained data from public sources such as LibriSpeech, Kaggle etc. The motivation behind the development of a new model came from the analyzed drawbacks of the Wav2Vec models that had gained popularity due to their unsupervised approach, which enabled them to train millions of hours of random speech without the need for human labeling (6). Their results were significant, suggesting that Whisper not only outperforms all other commercially available ASR models, but also that the Whisper model is approaching human-level accuracy. These findings are graphically displayed in Figure 1.

Despite the success of Whisper, the authors did

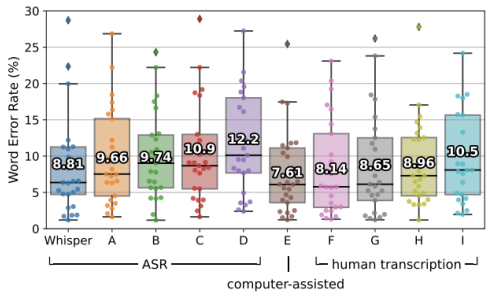

Figure 1: This plot shows the WER distributions of 25 recordings from the Kincaid46 dataset transcribed by Whisper, 4 commercial ASR systems, one computer-assisted human transcription service (E) and 4 human transcription services (F-I).

make note of a few limitations to their model. The speech the model was trained on was padded or cut to be only 30 seconds in length at a maximum, which is largely due to the datasets that were used containing mostly small speech segments. The authors make a note that long supervised speech datasets would be beneficial so the model could be trained on longer speech, potentially increasing the robustness of the model on longer segments of speech. Additionally, the model was trained on less than 1000 hours of language that was not English, which leads to the model performing significantly worse on languages such as Mandarin, Hindu, and Spanish. The authors suggest a targeted effort should be undertaken to increase the amount of supervised data available from languages other than English if the model is to become more robust for languages other than English.

3 Problem Formulation

The limitations of the model discussed in the paper suggest that another possible limitation could be accented speech, given that many of the datasets they used for training Whisper are known to be a single, non-accented English speaker. This theory supported by an article published by the Acoustical Society of America, in which they tested Whisper on American-accented English, Canadian-accented English, British-accented English, and Australian-accented English (7). The results showed that the match error rate (MER) of the American-accented English was significantly better than the other 3 accents tested, and is shown in Figure 2.

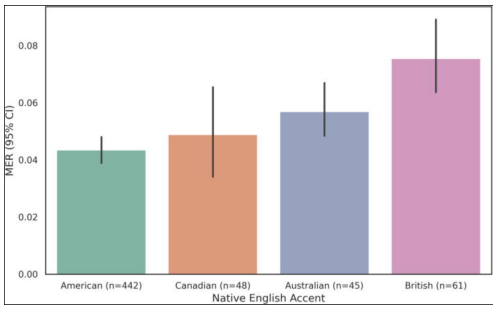

Figure 2: Match Error rates for American-accented English, Canadian-accented English, British-accented English, and Australian-accented English.

Given how similar the analyzed accents are to the American accent, it is reasonable to conclude that the model would perform worse on less similar accents, including both native English accents and non-native English accents.
We also compiled a graph from data collected in the original paper, which is displayed as Figure 3. It shows the amount of supervised data the Whisper model was trained on, and the resulting WER on a given testing dataset. It is evident observing the graph that significant decreases in WER are observed for large dataset increases initially, but then exponentially decays as dataset size grows very large. This is indicative that large increases in dataset sizes moving forward would have close to negligible effects on ASR accuracy, and new methods are needed if ASR's are to one day reach and possible exceed human level speech recognition.

Using the above analysis, we created a project that would attempt to address the issue of robust-

ness on non-American accented English. We did this by training the Whisper-small model on additional data that is comprised of supervised, accented speech. Then, we compared the WER for the testing dataset of the Whisper-small model to the new, fine-tuned model to determine their was any difference in the evaluation measurement.

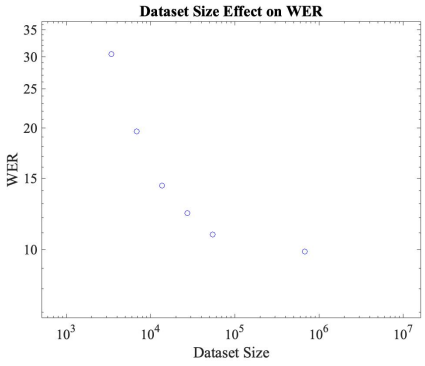

Figure 3: Match Error rates for American-accented English, Canadian-accented English, British-accented English, and Australian-accented English.

4 Dataset

The dataset chosen for the project comes directly from Hugging Face, an open source website with a wide variety of datasets for training a wide range of different models. The dataset is titled "common-accent" and is broken up into a training and testing section. The training section had 20 hours of supervised speech ranging from 5 to 10 seconds in length from 173 different accents. Likewise, the testing data set also had 173 accents represented in 5 to 10 second supervised speech segments, but only 2 hours of total speech. A link to the dataset can be found here.

5 Model Architecture

The overall fine-tuning process was guided by a paper called "Fine-Tune Whisper For Multilingual ASR with Transformers" by Sanchit Gandhi (8). The fine-tuned model architecture is the exact same transformer architecture as that described in the paper "Attention is all you need," which happens to also be the same architecture used in the original Whisper model (9). An image of this transformer model is shown in Figure 4., and was taken directly from the original Whisper paper. The source code for the transformer can be found on OpenAI's GitHub (10).

In preprocessing, the raw audio data was first converted into Log-Mel spectograms and padded to 30 seconds utilizing the "feature_extractor()" function taken from the transformers library, which is available on Hugging Face and is a PyTorch based library specific to the Whisper model. Next, the human labels were tokenized via the hugging face function called "tokenizer()" which was pre-trained with weights based on the original Whisper model. As an aside, both the transformation of raw audio data into Mel-spectograms and the tokenization of the human speech labeling can be combined into a single step using the Whisper processor.

After preprocessing, the pre-trained Whisper model was called from Hugging Face. The model employed in this project is Whisper-Small, which is one member from a family of Whisper models. The model has 244M parameters, 12 layers, and 12 heads. In the encoder block, the model first feeds the preprocessed Mel-spectograms through 2 CNN layers, before applying the GELU non-linearity (11). Following this, the tokens are positionally encoded using sinusoidal positional encoding. Sinusoidal positional encodings are constant and do not update with the model. An example of their implementation and mathematical theory is given in Figure 5.

Next, the tokens enter into an encoder block, which contains a multihead attention and MLP sub-layer, followed by layer normalization. Each

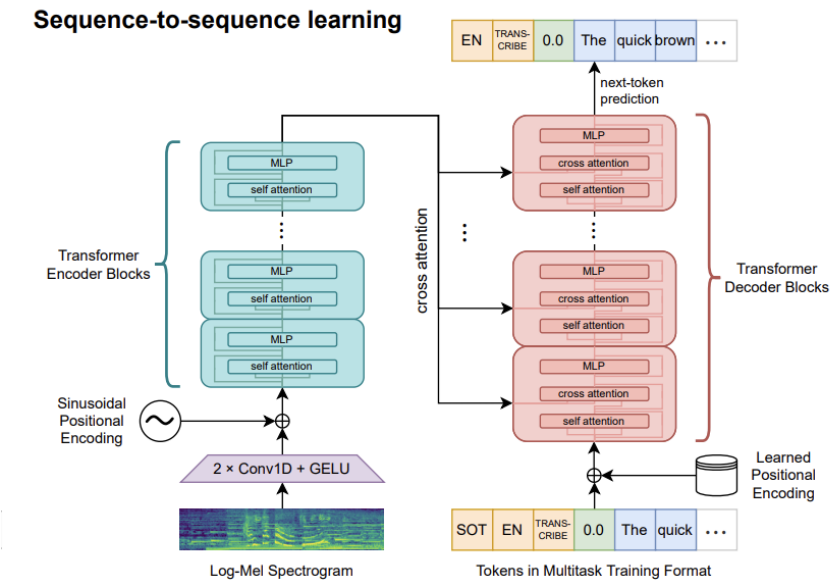

Figure 4: Transformer architecture used for fine-tuning the Whisper model. This is the same architecture employed in the original Whisper model.

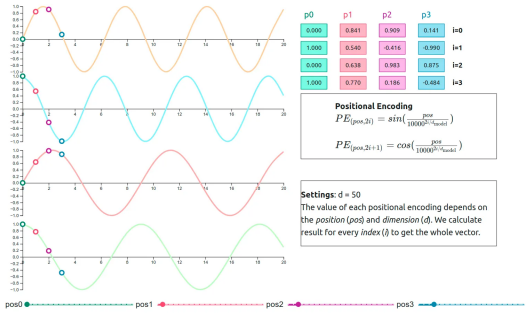

Figure 5: Example of how positional sinusoidal embedding can be implemented.

encoder block is identical. Simultaneously, in the decoder, the tokens from the human labeled text are put into learned positional encoding, which are not fixed like the sinusoidal encoding and will update with the model. They are then passed through decoder blocks, very similar to the encoder blocks but with an additional cross-attention sub-layer. And like the encoder blocks, each decoder block is the same. Despite the more complicated medium and large Whisper models, the small model was the largest model that we were confident we would be able to train given our limited access to high level GPUs. An image of the models in the Whisper family can be seen in Table 1.

Table 1: Whisper Model Sizes

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

A data collator was then employed to transform the data into pytorch tensors of equal dimensionality. We also needed to define an evaluation metric, and decided to use the WER metric. There has been significant research into developing more accurate evaluation metrics such as H_{eval}, but to be consistent with the original model, we decided to use WER (12). We finally employed a trainer from the Transformers library from Hugging Face titled ”Seq2SeqTrainer()” and defined various training parameters. The total training time was 2:24.57 using the Nvidia A100 GPU.

6 Results

The results from the training are shown in Table 2. and Figure 5. The average WER over the course of training was 0.21339, or 21.339%. The same testing data was ran using the normal Whisper-small model, and a WER of 0.25083 or 25.083%.

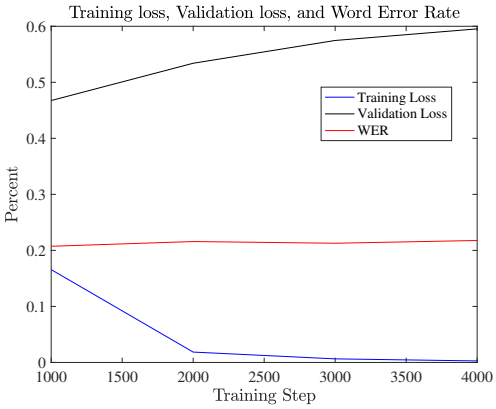

Figure 6: Graph of the training performance of the fine-tuned model over a wide range of training steps.

Table 2: Training and Validation Metrics

Step	Training Loss	Validation Loss	WER (%)
1000	0.165700	0.467473	20.74
2000	0.018500	0.534014	21.58
3000	0.006400	0.574676	21.27
4000	0.002700	0.595369	21.77
Whisper small	-	-	25.08

It is important to point out some of the limitations of the results. The only testing data set that was used was the accented speech testing dataset in "common-accent" from Hugging Face. Thus, although the WER for the fine-tuned model was smaller than that of the normal Whisper-small model, those results aren't able to be extrapolated beyond this dataset. It is a real possibility that by training the fine-tuned model only on additional data with accents that it might actually perform worse on American accented speech datasets. Further testing and validation is required to determine if this is in fact that case.

An additional limitation to the study is the size of the model used. While it is fair to conclude that fine-tuning Whisper models on the accented data of model size Whisper-small or smaller will reduce the the WER for the an accented dataset, it is not possible to conclude that the same magnitude of the difference in WER will be seen in larger models. Larger models typically have more parameters, which means they have a higher capacity to learn complex patterns from data. With more parameters, can represent a wider variety of functions and capture more intricate relationships within the data. This increased capacity can make them more robust to variations in training data size. Additionally, larger models have a greater tendency to over-fit training data, which would result in worse testing performances than expected. These, in addition to other factors that affect larger models more heavily than smaller models, are the reasons the results of the project cannot be extended to the medium and large Whisper models.

A final, and critically important limitation to note is the content of the dataset. While 173 accents were represented, they were not all represented equally across the data, and 20 hours of speech is not enough time for significant segments of speech from all 173 accents. This leads to the analysis that the fine-tuned model may only have become prolific at recognizing a few types of accents far better than the normal Whisper-small model, which could have been more represented in the testing data, thus resulting in the large difference in WERs between the models. Future iterations of the study should utilize significantly more accented testing data specialized to certain accents to test if the theory outlined above is valid.

7 Conclusions

With thousands of hours of audio data used to train the whisper model without much improvement, this project's simple, small-scale training

on limited data has shown the potential to improve the whisper model to other variations in the human vocal range. However, the issues remain in finding a subsequent source of data. As is the case with many models that use human-centered data, the data is only available from those individuals who are interested in having themselves be recorded and labeled. There just isn't a lot of data of non-American accented speakers or heavily accented native English speakers.

It is possible that future criticism of ASR models like Whisper might come from its perceived "non-inclusive" tendencies for its decreased performance in non-American accents, as was the case with Apple's "face ID" being condemned for its propensity to recognize lighter-toned faces with greater accuracy than darker-toned faces. Similar labelling by some of society could hold back many AI models on this political front. Thus, the implementation of a wider range of data may not just be the next logical step, but add significant commercial appeal to a politically-conscious society.

Using the accented data could also potentially improve the model's robustness in decoding generic American speech as well, by picking up different speech patterns that may still be present in an American accent, but more pronounced in other accents. An example would be an American accented speaker who is sick. Additional testing is required in order for this conclusion to be made. Until a future improvement on the architecture of Whisper is constructed, such as improving the embedding process or contents of the encoding and decoding blocks, the foreseeable improvements on the current Whisper models stem entirely from the quality and variety of the data being used. While our project suggests using accented datasets is a viable option for improving the robustness of the Whisper model, other potential manipulations to data, such as increasing the length of the speech segments, should also be considered and could be viable options for improving robustness as well.

References

[1] Foster, Kelsey. (2023). *What Is ASR? An Overview of Automatic Speech Recognition.* Retrieved from https://www.assemblyai.com/blog/what-is-asr/

[2] Aguacil et al. (2023). *Predicting the Propagation of Acoustic Waves Using Deep Convolutional Neural Network.* Retrieved from www.sciencedirect.com/science/article/pii/S0022460X21003527.

[3] Panagiotis Antoniadis (2023). *What Is End-to-End Deep Learning?* Retrieved from www.baeldung.com/cs/end-to-end-deep-learning#:~:text=Definition,without%20any%20manual%20feature%20extraction.

[4] Amazon Web Services. (2024). *What are transformers in artificial intelligence?* Retrieved from https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/

[5] Radford et al. (2022). *Robust Speech Recognition Via Large-Scale Weak Supervision* Retrieved from https://cdn.openai.com/papers/whisper.pdf

[6] Zhang et al. (2021). *BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition* Retrieved from https://arxiv.org/pdf/2109.13226

[7] Graham et al. (2024). *Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits* Retrieved from https://pubs.aip.org/asa/jel/article/4/2/025206/3267247/Evaluating-OpenAI-s-Whisper-ASR-Performance

[8] Sanchit Gandhi (2022). *Fine-Tune Whisper For Multilingual ASR with Transformers*

Retrieved from https://huggingface.co/blog/fine-tune-whisper

[9] Vaswani et al. (2017). *Attention is all you Need* Retrieved from https://arxiv.org/pdf/1706.03762

[10] Whisper Model Code: https://github.com/openai/whisper

[11] Hendrycks et al. (2016). *GAUSSIAN ERROR LINEAR UNITS (GELUS)* Retrieved from https://arxiv.org/pdf/1606.08415

[12] Huang et al. (2023). *H_{eval}: A new hybrid evaluation metric for automatic speech recognition tasks* Retrieved from https://arxiv.org/html/2211.01722v3