Improving Speech-to-Text (STT) for African American Vernacular English (AAVE)

Karan Agarwal
University of Maryland, College Park
getkaran@terpmail.umd.edu

Aayush Gupta
University of Maryland, College Park
agupta47@terpmail.umd.edu

Tejas Iyer
University of Maryland, College Park
tiyer1@terpmail.umd.edu

Melvin Anand Rajendran
University of Maryland, College Park
mrajendr@terpmail.umd.edu

Vikas Thoti Reddy
University of Maryland, College Park
vikasred@terpmail.umd.edu

Abstract

This project aims to improve Speech-to-Text (STT) accuracy for African American Vernacular English (AAVE). This form of artificial intelligence modeling is known as Automatic Speech Recognition (ASR). Specifically, the team fine-tuned OpenAI's Whisper model on an AAVE audio dataset. Current ASR models underperform on underrepresented dialects like AAVE, perpetuating technological inequities. By selecting a labeled dataset of AAVE audio and evaluating model performance via Word Error Rate (WER), this project seeks to enhance ASR accuracy for AAVE speakers and pave the way for future research into addressing speech recognition inequalities.

1 Introduction

Speech-to-text (STT), also known as Automatic Speech Recognition (ASR), is a form of artificial intelligence modeling that transcribes spoken language into written text. Today, this type of modeling has widespread applications including real-time transcriptions, voice search, voice assistants, and language translation. Furthermore, sophisticated models such as OpenAI's Whisper, Deepgram's Voice AI, and AssemblyAI's Universal-1 have significantly enhanced Speech-To-Text performance, yet these systems still face critical challenges in handling diverse accents and dialects.

One of these models' crucial problems is that they underperform on uncommon varieties, accents, and dialects as mainstream English. One such variety is African American Vernacular English (AAVE), which is natively spoken, especially in urban communities, by many African Americans and Black Canadians. Hence, transcriptions of AAVE speakers across traditional media, social media, and other social settings often lag behind those of mainstream English speakers. This problem likely stems from training data bias, as ASR models are predominantly trained on datasets that contain mainstream English.

As such, the goal of this project is to improve OpenAI's Whisper model's performance on AAVE. This will have a widespread impact across the various applications of ASR modeling. By better capturing and transcribing this variety of English, our team will make meaningful progress towards equitable representation of AAVE speakers in modern settings. In the future, this work can be applied to speakers of other varieties, accents, and dialects to improve the representation of all speakers.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Related Works

Whisper by OpenAI is a large-scale model for automated speech recognition (ASR) [3]. It has been trained on 680,000 hours of diverse audio from the internet. This makes it good at understanding different languages and performing tasks like transcription and translation without additional training. The model is robust enough to handle diverse audio inputs and is designed for broad use cases. However, the model struggles with recognizing speech that is less represented in its training data. Thus, Whisper can be used as a foundational model to improve automated speech recognition for underrepresented accents and dialects.

Koenecke et al. (2020) explore racial disparities in the performance of ASR systems [2]. The study examines leading ASR technologies and finds that they consistently perform worse for African American speakers compared to white speakers. The authors suggest that these disparities are partly due to ASR systems being trained on datasets that underrepresent AAVE and regional dialects often used by Black speakers. This raises concerns about accessibility and equity in ASR systems that are used in virtual assistants and other technologies. The study recommends expanding training datasets to include diverse voices to improve ASR performance for all users.

The Corpus of Regional African American Language (CORAAL) is a collection of spoken language data from African American communities [1]. It was created to study regional variations in AAVE. The dataset includes recorded interviews, conversations, and narratives from different U.S. regions. CORAAL is an open-access project that aims to provide a resource for researchers studying AAVE and its applications.

3 Methods

3.1 Dataset Selection

We used the CORAAL dataset (Corpus of Regional African American Language) due to its extensive representation of AAVE. This dataset includes recorded interviews, conversations, and narratives from African American communities, making it a robust resource for studying linguistic variations specific to AAVE.

3.2 Data Cleaning and Audio Processing

To prepare the dataset for model training, the following steps were undertaken:

1. **Audio Resampling:** Audio files in CORAAL were resampled from their original frequency (48kHz) to 16kHz to align with Whisper's input specifications.
2. **Segmentation:** Long audio files were split into smaller segments of manageable duration, ensuring compatibility with the model's input size limits.
3. **Normalization:** Audio levels were normalized to avoid biases introduced by varying input amplitudes.

3.3 Feature Extraction and Tokenization

1. **Feature Extraction:** Used *WhisperProcessor* to convert audio segments into log-Mel spectrograms, a suitable representation for the Whisper model.
2. **Tokenization:** Transcriptions were tokenized using *WhisperTokenizer*, converting them into input token sequences for the model. The tokenization process included handling special characters and ensuring consistency across the dataset.

3.4 Dataset Preparation

- **Splitting the Dataset:** The CORAAL dataset was divided into training (80%), validation (10%), and test (10%) subsets.
- Evaluating the model's performance by comparing baseline and fine-tuned WER on the CORAAL validation set.

- Conducting subgroup-specific evaluations based on speaker age, gender, and region to identify areas of improvement.

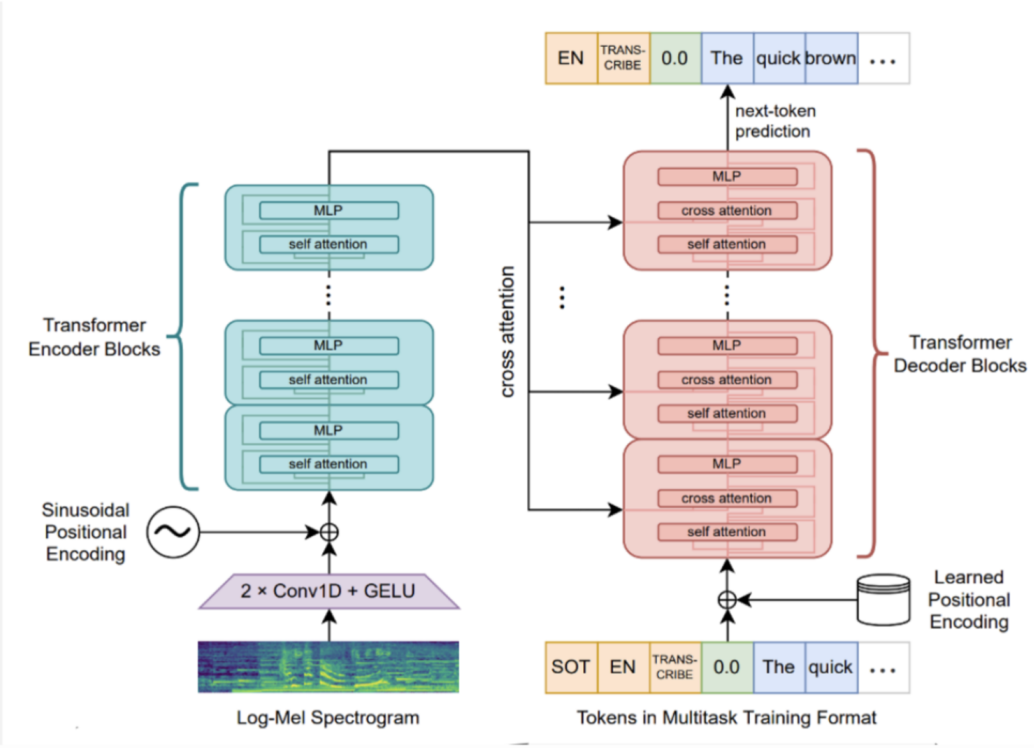
3.5 Model Architecture

Whisper ASR Model

Whisper is a Transformer-based encoder-decoder architecture, trained on 680,000 hours of multilingual and multitask data. We used the tiny variant of Whisper. It has 39M parameters and is trained on both English-only and multilingual data.

Model Configuration

- **Encoder:** Extracted log-Mel spectrograms served as inputs to the encoder.
- **Decoder:** Used tokenized transcriptions to generate text outputs during the decoding stage. The Decoder:
  1. Converts tokenized outputs back to human-readable text.
  2. Removes special tokens for cleaner predictions.
  3. Applies case normalization by converting all text to lowercase.



3.6 Fine Tuning Pipeline

- Hugging Face's **transformers** library provided a streamlined pipeline for fine-tuning with the **Seq2SeqTrainer**
- **Evaluation Metric:** Word Error Rate (WER) was calculated using the **evaluate** library to assess transcription accuracy after each training epoch.
- **Hyperparameters:**
  1. Batch Size: 16
  2. Learning Rate: 1e-5

3. Epochs/Steps: 4000

- **Optimizer:** Adam's Optimizer
- **Linear Learning Rate Scheduler:** 500 warmup steps after which it starts decreasing
- **Loss Function:** Cross-Entropy Loss

4 Experiments and Results

Firstly, the team obtained the baseline WER of the Whisper (tiny) model on the CORAAL test dataset. The results are as follows:

```
Un-fine-tuned Whisper Model WER (uppercase): 51.73%

Reference: NOTHING LIKE THAT GO OVER MY HEAD CAUSE YOU KNOW WELL KILL OURSELVES TRYNA WORRY ABOUT
Prediction:  NOW LIKE THAT GOES MY HEAD CAUSE YOU KNOW WE'RE KILLER, I SAY I WAS TRYING TO WORRY ABOUT.

Reference: ID GO OUT BECAUSE ID BEEN SITTING DOWN ALL DAY AND I DIDNT FEEL LIKE DOING ANY HOMEWORK WHEN I
Prediction:  I'D GO OUT BECAUSE I'VE BEEN SITTING DOWN ALL DAY AND I DIDN'T FEEL LIKE DOING HOMEWORK WHENEVER.

Reference: I MEAN YOU YOU YOU YOURE DOING A GOOD JOB YOURE C CONSCIENCE CONSCIENCE CONSCIENTIOUS ABOUT IT
Prediction:  I MEAN, YOU DID A GOOD JOB OF CONTENT ABOUT IT.

Reference: YOU KNOW HOLLER AT ME TELL ME I SHOULD HAVE WENT TO BED EARLIER OR SOMETHING YOU KNOW
Prediction:  YOU KNOW, HAAL AT ME, TELL MY SHIT WITH THE BEER EARLIER, SO I'M, YOU KNOW.

Reference: BUT HE SCARED THOUGH LIKE YOU SAY OR YOU GO AND GET YOUR MOTHER OR FATHER
Prediction:  BUT HE'S SCARED, LIKE HE SAID, YOU'RE GOING TO GET YOUR MOTHER FATHER.
```

As shown above, the baseline Whisper (tiny) model has a WER of 51.73% on the CORAAL test dataset. This is reflected in the model outputs (i.e. predictions), which are often incorrect and fail to recognize slang phrases like "HOLLER AT ME". Furthermore, after training the model with initial hyperparameters, the team obtained the following results:

```
Fine-tuned Whisper Model WER (uppercase): 9.29%

Reference: NOTHING LIKE THAT GO OVER MY HEAD CAUSE YOU KNOW WELL KILL OURSELVES TRYNA WORRY ABOUT
Prediction: NOTHING LIKE THAT GO OVER MY HEAD CAUSE YOU KNOW WELL KILL OURSELVES TRYNA WORRY ABOUT

Reference: ID GO OUT BECAUSE ID BEEN SITTING DOWN ALL DAY AND I DIDNT FEEL LIKE DOING ANY HOMEWORK WHEN I
Prediction: ID GO OUT BECAUSE ID BEEN SITTING DOWN ALL DAY AND I DIDNT FEEL LIKE DOING ANY HOMEWORK WHEN I

Reference: I MEAN YOU YOU YOU YOURE DOING A GOOD JOB YOURE C CONSCIENCE CONSCIENCE CONSCIENTIOUS ABOUT IT
Prediction: I MEAN US YOU YOU YOU YOU YOU DID YOU DID A GOOD JOB YOUR CONTUREATION CONTENTION ABOUT IT

Reference: YOU KNOW HOLLER AT ME TELL ME I SHOULD HAVE WENT TO BED EARLIER OR SOMETHING YOU KNOW
Prediction: YOU KNOW HOLLER AT ME TELL ME I SHOULD HAVE WENT TO BED EARLIER OR SOMETHING YOU KNOW

Reference: BUT HE SCARED THOUGH LIKE YOU SAY OR YOU GO AND GET YOUR MOTHER OR FATHER
Prediction: BUT HE SCARED THOUGH LIKE YOU SAY OR YOU GO AND GET YOUR MOTHER OR FATHER
```

As shown above, our best finetuned Whisper (tiny) model has a WER of 9.29%, which is significantly better than the untrained model. Once again, this is reflected in the model outputs, as only one of them has incorrect words.

Next, the team moved on to identifying opportunities to further improve the model performance. First, the team leveraged the following table of the training losses and validation losses over the training period:

Step	Training Loss	Validation Loss
1000	0.008500	1.376329
2000	0.001700	1.508498
3000	0.000600	1.577645
4000	0.000500	1.602011

From the above chart, the team recognized that the total number of steps is likely too high, and thus, the model is not improving in performance after 1000 steps. The team re-ran the training to generate this new graph to determine when the Validation Loss tapered off. Since validation loss was the lowest at 350 steps, the team decided to only train up for 1000 steps.

The team had a warmup step hyperparameter set at 500 steps which does not fit into the typical $10 - 20\%$ of steps recommended to add a warmup function. Since our training was peaking so quickly, the team decided to lower the learning rate from **1e-5** to **8e-6**.

Step	Training Loss	Validation Loss	Model Preparation Time	Wer
25	2.089100	2.189947	0.002400	98.394348
50	2.035300	2.081369	0.002400	98.265896
75	1.858700	1.942532	0.002400	98.330122
100	1.739200	1.830470	0.002400	98.330122
125	1.645200	1.728229	0.002400	98.587026
150	1.474800	1.629255	0.002400	87.219011
175	1.394800	1.518430	0.002400	101.541426
200	1.221000	1.397553	0.002400	75.979448
225	1.114300	1.341171	0.002400	86.769428
250	1.041000	1.301116	0.002400	69.107258
275	0.952200	1.266917	0.002400	68.914579
300	0.859500	1.246908	0.002400	66.795119
325	0.872000	1.219921	0.002400	78.355812
350	0.758200	1.203079	0.002400	77.520873
375	0.689900	1.188292	0.002400	65.382145
400	0.664200	1.175120	0.002400	75.208735
425	0.559300	1.170618	0.002400	75.272961
450	0.544500	1.165461	0.002400	76.364804
475	0.464400	1.159150	0.002400	74.373796
500	0.421800	1.164427	0.002400	74.116891
525	0.374400	1.165352	0.002400	76.429030

The higher minimum validation loss was surprising to us because the team thought the reason for the quick training and overfitting was because of an overly aggressive learning rate. The model might have got stuck in a local minima, so to test this theory, the team increased the learning rate to 2e-5

and reduced the *warmup steps* to 100.

Step	Training Loss	Validation Loss	Model Preparation Time	Wer
25	1.987800	1.878263	0.002400	98.330122
50	1.558100	1.519855	0.002400	109.184329
75	1.141500	1.275331	0.002400	75.401413
100	0.860900	1.190018	0.002400	61.271676
125	0.747900	1.139758	0.002400	75.465639
150	0.493200	1.130647	0.002400	86.962107
175	0.398500	1.136693	0.002400	101.156069

This yielded a much lower validation loss and reinforced the theory that the team was not being aggressive enough with the learning rate. The team ran more experiments by lowering the batch size hyperparameter from 16 to 8 and by adding a *weight decay*=.001, but it did not yield any significant changes to overall loss.

At this stage, the team noticed that the Word Error Rate was abnormally high, and printed a few examples to investigate. It turned out that the CORAAL dataset's transcription was all uppercase, and this was causing a high WER and possibly interfering with finetuning. The team decided to ignore the case when calculating the loss and the WER. The team reran finetuning and lowered the learning rate because of the proportionally lower loss. This led to yielding the model's lowest WER at 9.29% after 500 steps of training. The relative change of the WER between the baseline model and the final trained model is **82.03%**.

5 Conclusion

The Whisper Tiny model achieves a Word Error Rate (WER) of 5.6 on the LibriSpeech test-clean benchmark, which is a general English benchmark derived from read audiobooks—a medium that includes minimal African American Vernacular English (AAVE) [3] [4]. Notably, the team achieved a WER of 9.29 on the CORAAL AAVE benchmark, demonstrating performance that is competitive with Whisper Tiny's results on other English dialects. This significant improvement in recognizing AAVE suggests that speakers of this dialect will experience markedly better accuracy in speech-to-text applications. As a result, technologies like smart assistants will become more accessible and inclusive for AAVE speakers.

The improved recognition of African American Vernacular English (AAVE) by Whisper Tiny opens the door to numerous impactful applications in technology and beyond. Enhanced speech-to-text models can empower AAVE speakers to engage more effectively with voice-activated systems such as smart assistants, virtual customer service agents, and dictation tools. This advancement also has significant implications for accessibility, enabling more inclusive communication tools for individuals who rely on speech-to-text technologies, such as those with disabilities. Furthermore, better AAVE recognition could benefit industries such as education, where automated transcription tools are increasingly used in classrooms, and media, where accurate subtitles and captions for diverse dialects are essential for broader audience reach. By improving representation in ASR systems, this technology can contribute to bridging the digital divide and fostering equity in the use of AI-driven tools.

6 Future Work

The next steps would be to utilize larger models and expand the dataset to handle more complex audio scenarios. Within two weeks, the focus would be on fine-tuning the Whisper small and medium models using the existing AAVE dataset. This includes optimizing hyperparameters like learning rate and batch size for the larger models to maximize WER. A study would be conducted to evaluate the impact of different regularization techniques, such as dropout and weight decay, on the performance of the larger models. Additionally, efforts would focus on testing the fine-tuned models in real-world scenarios using Hugging Face pipelines for evaluation.

Over two months, the team would focus on expanding our dataset by collecting short audio snippets from publicly available sources like podcasts, interviews, and social media featuring AAVE speakers. These recordings would be processed into manageable segments, resampled to 16 kHz, and cleaned to reduce background noise. Initial transcriptions would be generated by the Whisper model and then manually labeled and corrected to ensure accuracy. Once the dataset is ready, the team would fine-tune the Whisper small and medium models to improve their performance in real-world scenarios. Finally, the team hopes to share our findings and dataset on open-source platforms or in academic publications to help advance speech-to-text technology.

7 References

[1] Kendall, T., & Farrington, C. (2023). The Corpus of Regional African American Language. Version 2023.06. Eugene, OR: The Online Resources for African American Language Project. Retrieved from https://doi.org/10.7264/1ad5-6t35.

[2] Koenecke, A., Namb, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences. Retrieved from https://doi.org/10.1073/pnas.1915768117.

[3] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. Retrieved from https://github.com/openai/whisper.

[4] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audiobooks. Retrieved from https://www.openslr.org/12.