Final Project Report: Asante Twi Speech Recognition Using Whisper and BERT

Group 12

Isaac Kwame Acquah Baah

Emmanuel Nhyira Freduah-Agyemang

Paa Kwesi Jnr. Thompson

Deep Learning

Mr Dennis A. Owusu, Ms. Leanne Annor-Adjaye

11th December 2024

1.0 Introduction

This project aimed to develop a speech recognition system for Asante Twi, a native Ghanaian language, using OpenAI Whisper fine-tuned on Asante Twi datasets. Subsequently, we attempted to use a BERT-based language model to refine the transcriptions through grammar error correction. The main objectives were to train an automatic speech recognition (ASR) model, evaluate its performance using held-out and newly compiled test datasets, and deploy the model with an accessible API.

We chose Whisper as our baseline model because initial testing with Twi audio recordings showed it could capture phonemes accurately but transcribe them in different languages. This observation led us to believe that fine-tuning Whisper with appropriate Twi datasets could help the model learn better representations of the Twi language and yield more accurate transcriptions.

2.0 Methodology

2.1 Data Preparation

We utilized two datasets:

1. Financial Inclusion Speech Dataset: Focused on financial contexts.

2. Asante Twi TTS Dataset: A general-purpose Twi dataset containing 28,000 samples[1].

The financial dataset was split into 80% training and 20% validation. The general dataset was similarly split, with transcription columns aligned for compatibility. Combined datasets were shuffled to mitigate bias and prepared using the `Audio` class from the Hugging Face `datasets` library. All audio files were resampled to 16 kHz to standardize inputs (Whisper only accepts 16kHz audio for training).

2.2 Model Training

We fine-tuned OpenAI Whisper (medium) using Hugging Face's `transformers` library. Key components included:

- Feature Extraction: Log-Mel spectrograms were computed from audio inputs. We chose Log-Mel because it provides a more detailed representation of audio data.

- Tokenization: Transcriptions were tokenized, leveraging Whisper's multilingual support to handle code-mixed English and Twi sentences.

- Data Collation: A custom data collator ensured proper padding for ASR tasks.

The training process employed a batch size of 16, gradient accumulation, and mixed precision using `fp16`. Evaluation used Word Error Rate (WER) and Character Error Rate (CER) metrics, computed via Hugging Face's `evaluate` library. Below is a screenshot of our training process and metrics.

Step	Training Loss	Validation Loss	Wer	Cer
1000	0.263000	0.980465	70.970942	34.670390
2000	0.169500	0.987131	57.553838	25.829134
3000	0.129000	1.071878	63.534907	33.332198

Figure 1: Evidence of training

2.3 Refinement with BERT

To refine ASR predictions, we attempted to use a pre-trained BERT model. Predicted transcriptions from Whisper were fed into BERT for contextual corrections, particularly addressing grammatical errors and code-mixed ambiguities. This approach did not fully materialize due to time and resource constraints. However, in Fig. 2 below, you can see how the Bert model faired on a very small scale data set.

```
Input Sentence >>> Good morning Kojo, eti s3n?
Best Suggestion >>> Good morning Kojo, eti s3n?
Probabilities(avg. logits) >>> [-10.385024070739746, -10.383777618408203, -10.384968757629395, -10.38626766204834]
```

Figure 2: Bert Model Output

2.4 Deployment

The final model was deployed on Hugging Face, providing an API endpoint for real-time transcription. This ensured accessibility for practical applications.

Our Api Endpoint:

```
pip install gradio_client
from gradio_client import Client, handle_file

client = Client("Ibaahjnr/ASR_TWI_MODEL")
result = client.predict(
    audio=handle_file(file_path),
    api_name="/predict"
)
print(result)
```

To use this API endpoint, change the filepath in the handle_file function call to the file path of the actual audio you want to transcribe.

3.0 Results

3.1 Evaluation on Held-Out Test Data

The model was evaluated on 10 randomly selected samples from the held-out test dataset, with predictions shown in Figure 2 (find evidence in Figure 3 of the appendix):

- Word Error Rate (WER): 8% - Character Error Rate (CER): 4%

3.2 Evaluation on Compiled Test Data

A second evaluation was conducted on 10 newly compiled test sentences, including mixed English phrases (find evidence in Figure 4 of the appendix):

- WER: 58% - CER: 24%

Discussion

The Whisper model demonstrated robust performance, achieving high transcription accuracy on Twi audio, even with code-mixed English. Integrating BERT would have further improved contextual understanding and corrected grammatical errors. Some challenges we faced included handling numerals and disambiguating similar-sounding words. Ultimately, with more time and computing units, we could use the predictions from our training set to fine-tune the BERT model for grammatical corrections and include this model in the pipeline.

References

[1]

G. Birikorang, “akuapem-twi-tts,” Huggingface.co, 2024. https://huggingface.co/datasets/kojo-george/akuapem-twi-tts (accessed Dec. 12, 2024).

[2]

Link to model: “Ibaahjnr/Asanti_Twi_Model_V2.1 · Hugging Face.” *Huggingface.co*, 2024, huggingface.co/Ibaahjnr/Asanti_Twi_Model_V2.1. Accessed 12 Dec. 2024.

Appendix

	predictions	references
0	Adwoa Aboagye	Adwoa Aboagye
1	Hwɛ soro	Hwɛ soro
2	Te so ma me	Te so ma me
3	Efiri sɛ	Esian sɛ
4	Medaase nyame nhyira wo	Medaase Nyame nhyira wo
5	Yɛ wɔ	Yɛ wɔ
6	To so ma me	To so ma me
7	Mayera	Mayera
8	Na biribiara dwoi	Na biribiara dwoi
9	Aane	Aane

Figure 3 3: Predictions for given held-out test set

	predictions	references
0	Sika mpɛ, didi	Sika mpɛ dede
1	Wode ka 6,000 cedis	Wode ka 6000 cedis
2	Sika mpɛ di	Sika mpɛ dede
3	Me pɛsɛ me bue bank akɔnt	Mepɛ sɛ mebue bank account
4	Wobɔ bosea saa ade	Woabɔ bosea saa adɛn
5	Wo abɔ busua saa adeɛ	Woabɔ bosea saa adɛn
6	Woabɔ bosea saa adɛn	Woabɔ bosea saa adɛn
7	Nti Ntres ketewaa wei na menni aya	ɛnti interest ketwa wei na menyaeɛ
8	Wode ka 6,000	Wode ka 6000 cedis
9	Afenhyia so a aban tua wo ka	Afe no so ah aban betua wo ka

Figure 44: Predictions for compiled test set

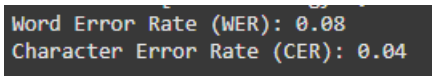

Figure 5 5: WER and CER for held-out test set

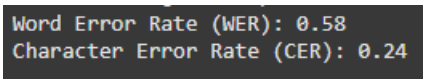

Figure 6 6: WER and CER for compiled test set