Improving ASR Performance on Accented English using LLMs

Rebekah Kim, Arthur Zhao, Divyansh Jain, Jason Lee, Kunling Tong, and Luke Fu

Abstract

Automatic Speech Recognition (ASR) systems translate spoken language into text. However, they are not completely representative of the large variety in the English Language, including accents. Integrating large language models (LLMs) can help improve these systems. Through supplementing BERT with Whisper, we investigated investigated how we can utilize LLMs to disambiguate tokens that ASR models are uncertain about when transcribing accented English using the Edinburgh International Accents of English Corpus. Through a baseline evaulation of the Whisper model, we compared metrics to 3 methods depending on BERT at varying levels to determine the effects of LLMs and to gain insights on ASR in accented speech.

1 Problem Statement

Building a representative and accurate automatic speech recognition (ASR) system that encompasses accented speech during training is often difficult to achieve. The variability of accents is due to geographic location, socio-economic status, and other socio-linguistic factors. Through bias auditing, several areas of high error have been identified in current English Language ASR, such as for Black Americans compared to White Americans, speakers whose first language is a tone language compared to those whose first language is not, and speakers with Indian accents compared to speakers who with "American" accents (DiChristofano et al., 2023).

Currently, many studies have attempted to use approaches to lower the high error rates that have been identified in ASR models. These methods include multi-accent training, accent-aware training using accent embeddings or adversarial learning, and accent adaptation using supervised or unsupervised techniques (Prabhu et al., 2023). However, these methods have only achieved partial success, and it remains challenging to accurately transcribe accented English.

Our research question is: How can we utilize LLMs to disambiguate tokens that ASR models are uncertain about during transcription of accented English? Our focus is to integrate Whisper and BERT for accent recognition and correction through exploring LLMs to disambiguate tokens that ASR models are uncertain about during transcription. We aim to find an approach in efforts to improve the word-error rate of current ASR models.

2 Background

Automatic Speech Recognition (ASR) systems are computational technologies designed to translate spoken language into text. They analyze acoustic signals of speech and break them down into components that can be matched against a database of known linguistic patterns to produce textual transcriptions. Over the years, ASRs have evolved significantly, from rule-based systems that relied heavily on phonetic algorithms to modern deep learning models that utilize neural networks to process and understand complex speech patterns. This evolution has enabled ASR systems to achieve remarkable levels of accuracy and efficiency, making them indispensable tools in various applications, such as voice-activated assistants and automated transcription services. Despite these advancements, ASR systems continue to face challenges, particularly in handling diverse accents and dialects, underscoring the ongoing need for innovation in the field (Graham, 2024).

Integrating LLMs to speech recognition presents a promising approach to address the limitations of

conventional ASR systems (Min Wang, 2023). Unlike traditional ASR systems, which primarily focus on acoustic features to transcribe speech, LLMs offer the ability to incorporate linguistic context into the transcription process from past conversation. By leveraging the contextual information encoded in LLMs, ASR systems can potentially enhance their transcription accuracy, particularly in scenarios where speech may be ambiguous.

More specifically, LLMs offer flexibility and adaptability in handling diverse linguistic contexts. Through pre-training on large-scale text corpora spanning various accents, dialects, and languages, LLMs can effectively capture the linguistic nuances inherent in different speech patterns. This pre-training enables ASR systems to generalize better across diverse speaking styles and linguistic backgrounds, thereby improving transcription accuracy and robustness.

Incorporating LLMs into ASR systems also opens avenues for continual learning and adaptation. By fine-tuning LLMs on domain-specific speech data or incorporating feedback mechanisms, ASR systems can dynamically adjust their transcription strategies to better align with evolving language use patterns and user preferences.

3 Methods

3.1 Model

In order to work with speech recognition as well as language modeling, we have chosen Whisper and BERT as our primary models to implement in our methods.

Whisper has emerged as a standard in ASR. Created by OpenAI, this model is trained on large sets of audio, both English and other languages, and is effective and efficient in transcribing speech. However, there may be shortcomings when it comes to accented English (Graham, 2024).

Developed by Google, BERT (Bidirectional Encoder Representations from Transformers) is a widespread model in the field of natural language processing. It is pre-trained on vast data and utilizes transformers. As a result, it can effectively analyze the context of words, and thus BERT has become a baseline in NLP (Koroteev, 2021).

The integration of Whisper and BERT aims to leverage Whisper's transcription capabilities with BERT's contextual understanding to address inaccuracies in recognizing accented English. We begin with Whisper transcribing audio inputs. In cases where Whisper's confidence level in a transcription is low, we use BERT to predict the most probable alternatives.

3.2 Corpus

Because of the diverse range of English spoken around the world, including accents and dialects, many public datasets lack diversity in representing English and its speakers. Even with the advancements in English ASR, these limitations occur and are usually reported as word error rates (WER) on established benchmark datasets.

The Edinburgh International Accents of English Corpus (EdAcc) is a dataset that aims to address the varieties of English by providing almost 40 hours of Zoom video conversations in various English accents. Each conversation lasted between 20 and 60 minutes and there were over 40 self-reported English accents from speakers of 51 different first languages. The participants received a questionnaire to develop a more well-documented dataset. Additionally, they gathered information on their linguistic background, including languages spoken, years of speaking English, and places of extended residence.

Professional transcribers manually transcribed all conversations and the transcriptions were then post-processed for evaluation. A linguist then standardized accents descriptions to compare performance across speaker groups.

Qualitative findings indicate that EdAcc encompasses greater linguistic variation compared to traditional datasets. The best-performing model achieved a 19.7% WER, which is significantly higher than the 2.7% WER on US English clean read speech. Performance dropped across models when evaluating Jamaican, Indonesian, Nigerian, and Kenyan English speakers. The results also suggest that further research is needed for English ASR systems to work effectively with different English accents.

3.3 Implementation Details

We will implement various combinations of our proposed Whisper + BERT architecture that vary its dependence on BERT. Let x denote our input audio sequence and $\hat{y}_i = argmax[P(y_i|x)]$, or the most likely token for time i. For all $\hat{y}_i$ with $P_{Whisper}(\hat{y}_i|x) < \tau$, we will add i to a set $\tau_{rescore}$. Then, for each $j \in \tau_{rescore}$, we will compute y'_j, a re-scored prediction for time j, using the following 3 methods:

1. Hybrid Re-Scoring of ASR and BERT (section 3.4)
2. N-Best of ASR (section 3.5)
3. BERT only (section 3.6)

We will compare these 3 rescoring methods to the baseline Whisper model on accented English with reported performance on clean English using our evaluation metric (more in section 3.8). Specifically, we aim to answer the following two questions:

1. Is there truly a notable dropoff in the performance of Whisper when evaluating traditional vs. accented English?
2. Do LLMs help improve Whisper's performance on accented English? If so, which of our 3 proposed methods works best?

We ran the baseline Whisper model on 200 and 1000 samples and plotted the distribution of the maximum predicted probabilities for each token in Figures 1 and 2. From these figures, we set our threshold τ = 0.5, as there is a drop in frequency of uncertain tokens past this point.

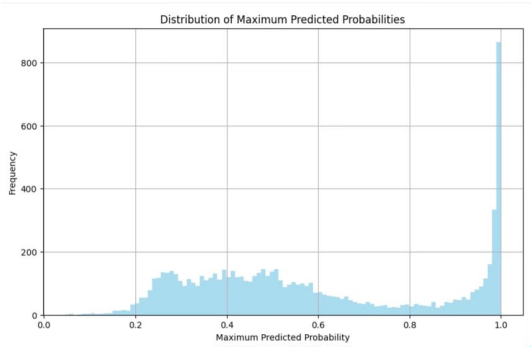

Figure 1: Distribution of 200 Samples

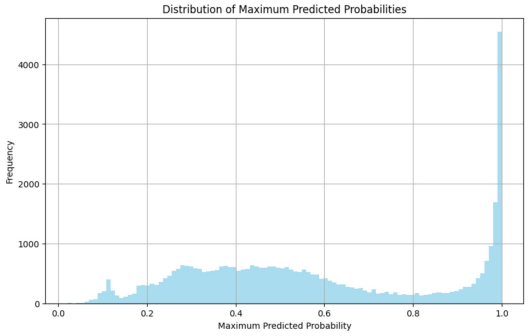

Figure 2: Distribution of 1000 Samples

3.4 Approach 1: hybrid Re-Scoring of ASR and BERT

Following the rescoring method proposed by Ortiz and Burud (2022), we use:

$$y'_j = argmax_{y_j}[(1-\gamma)P_{Whisper}(y_j|x) + \gamma P_{BERT}(y_j|x)] \quad (1)$$

where we use γ to weigh the predicted probability distributions of our Whisper and BERT model, then identify the token y that maximizes the weighted sum of these distributions.

3.4.1 Model Initialization

Initializing each models pre-trained configurations, we used Whisper's open-ai/small and BERT's based-uncased mdoels.

3.4.2 Processing and Prediction

The audio inputs are first processed by Whisper, which converts them into textual transcriptions and calculates confidence scores for each token in the transcription. Then, tokens will be re-evaluated by BERT. For low-confidence scores, BERT's language modeling capabilities are utilized.

The integration strategy specifically involves a hybrid rescoring mechanism where the confidence scores from Whisper and the contextual understanding from BERT are combined using a weighted average controlled by a parameter γ = 0.5 by arbitrary default. This approach allows us to dynamically balance the influence of acoustic accuracy and contextual relevance on the final transcription.

3.4.3 Confidence and Correction

Our method not only predicts transcription but also assigns confidence levels to each token to identify which parts of the transcription are uncertain and thus need further processing from BERT:

- Tokens with confidence below a threshold τ = 0.5 were flagged for re-scoring.
- For these tokens, a new confidence score was calculated by blending Whisper's original confidence with the new score derived from BERT's analysis, using the gamma parameter to control the blend ratio.

This hybrid approach aimed that each word in the transcription is contextually appropriate, aiming to significantly reduce misinterpretations caused by accent variations or complex linguistic contexts.

3.5 Approach 2: N-Best of ASR

We first identify the N-Best predictions of our Whisper model as candidate tokens

$$\hat{Y}_j = NBest[P_{Whisper}(y_j|x)] \tag{2}$$

then, we extract BERT's predicted probability for the N candidates and select the most likely candidate:

$$y'_j = argmax_{y_j}[P_{BERT}(y_j|x)] \mid y_j \in \hat{Y}_j \tag{3}$$

3.6 Approach 3: BERT Only

We disregard the predicted probability distribution of our Whisper model and rely solely on BERT's prediction for time j.

$$y'_j = argmax_{y_j}[P_{BERT}(y_j|x)] \tag{4}$$

3.7 Data Pre-Processing

We pre-processed the candidate and reference texts using the packages $werpy$ and $jiwer$. The normalization of texts included: removing punctuation, duplicated or white spaces, leading/trailing blanks, converting all words to lowercase. In addition, we expanded common English contractions, removed Kaldi non-words (ex. <laugh>) and specific words such as "uh", "um", "mm." Before testing, we also removed any candidate samples predicted numbers, as our reference text contained words (ex. "eleven") while the transcription would output numbers ("11").

3.8 Evaluation Metrics

We will be using word error rate (WER), word-information lost (WIL), and match-error rate (MER) to evaluate our results.

Our main evaluation metric will be WER, which is a standard evaluation metric used in ASR studies.

Word Error Rate:

$$WER = \frac{S + D + I}{N} \tag{5}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of the words in the reference (Gurevych & Miyao, 2018).

WIL and MER are also common evaluation metrics that provide more detailed comparison between the reference and candidate texts (Morris et al., 2004).

Word Information Lost:

$$WIL = 1 - \frac{H^2}{(H + S + D)(H + S + I)} \tag{6}$$

Match Error Rate:

$$MER = \frac{S + D + I}{S + D + I + H} \tag{7}$$

where H is the number of hits, D is the number of deletions, I is the number of insertions, and S is the number of substitutions between the reference and candidate.

4 Results

4.1 Baseline Whisper Model Evaluation

We ran the baseline for 15-30 word transcripts (1635), 15-100 word transcripts (2835), and 50-100 word transcripts (496) out of 9848, and got the following WER, WIL, and MER as shown in Table 1:

Text Length	WER	WIL	MER
15-30 words	27.09%	30.90%	23.39%
50-100 words	24.68%	26.64%	21.14%
15-100 words	24.42%	27.68%	21.24%

Table 1: WER, WIL, MER for Various Reference Text Lengths

We found the WIL to be relatively high because the Whisper transcription does not account for repeat words well. For instance, the Whisper transcription turns repeats of the word "I" into a singular "I". In addition, the WER was slightly higher

than previous 19.7% determined by the best model (Ramon et al., 2023).

To further analyze our baseline Whisper model, we determined the WER for various accents characterized in the dataset. We selected a few representative accents and plotted them below.

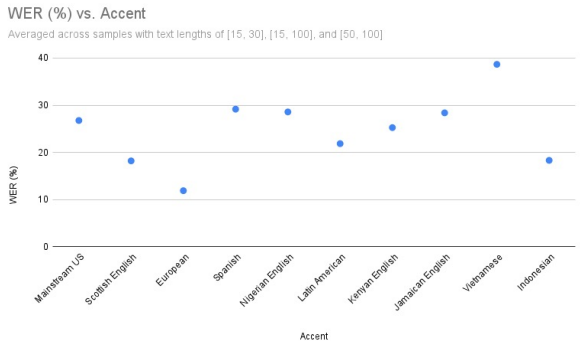

Figure 3: Graph of WER for Various Accents

4.2 Approach 1: Hybrid Re-Scoring of ASR and BERT

Text Length	WER	WIL	MER
15-30 words	33.03%	41.39%	30.66%
50-100 words	36.33%	43.65%	33.75%
15-100 words	43.33%	51.80%	38.24%

Table 2: WER, WIL, MER for Various Reference Text Lengths

4.3 Approach 2: N-Best of ASR

Text Length	WER	WIL	MER
15-30 words	24.56%	30.50%	24.03%
50-100 words	21.12%	26.21%	20.86%
15-100 words	22.69%	28.29%	22.21%

Table 3: WER, WIL, MER for Various Reference Text Lengths

4.4 Approach 3: BERT Only

Text Length	WER	WIL	MER
15-30 words	30.21%	41.08%	29.40%
50-100 words	25.58%	33.69%	25.11%
15-100 words	26.94%	36.53%	26.36%

Table 4: WER, WIL, MER for Various Reference Text Lengths

4.5 Conclusion

We explored three different approaches to improve Automatic Speech Recognition (ASR) performance and evaluated their effectiveness using various evaluation metrics.

Firstly, we employed a baseline evaluation of the Whisper model, which demonstrated reasonable performance with Word Error Rate (WER) ranging from 24.42% to 27.09%, depending on the text length. However, the Word Information Lost (WIL) and Match Error Rate (MER) were relatively high due to limitations in handling repeated words and linguistic nuances.

The findings suggest that while the hybrid approach has potential, careful calibration and enhancements are necessary to truly harness the strengths of both Whisper and BERT in improving ASR performance on accented English.

The increase in confidence probabilities when compared to the baseline Whisper model, despite higher error rates highlights the complexities of integrating different models with distinct strengths. This suggests that the integration of BERT did not effectively enhance ASR accuracy, possibly due to challenges in aligning ASR outputs with BERT's tokenization. By addressing these integration challenges, there is potential to combining Whisper and BERT to improve ASR accuracy, particularly in handling diverse and accented English, but that process would take several extra fine-tuning steps, as discussed below.

In our implementation of the N-Best of ASR approach, where uncertain tokens were replaced using BERT's predictions, we observed notable improvements, with WER reduced to 22.69% for the range of 15-100 words. This is despite problems with data cleaning, and other limitations The performance enhancements suggest that incorporating BERT's predictions for uncertain tokens can effectively mitigate ASR errors and improve transcription accuracy.

Finally, in the BERT only approach, we obtained metrics deemed worse than the baseline model. However, we cannot necessarily conclude that this approach is invalid. One, this could be a testament to the limitations of LLMs in this use case. Additionally, calibrating tokenization with BERT could have led to inconsistencies.

4.6 Discussion

The integration of BERT in the Whisper-driven transcription process was intended to enhance the accuracy of transcriptions by providing contextual clarity where Whisper was uncertain. However, while preparing and comparing preliminary results,

although the error rates increased as shown in our results, it was noted that the confidence probabilities associated with the transcribed tokens also increased. There can be several interpretations behind these observations:

- **Higher Confidence Misalignments**: The increase in confidence probabilities suggests that the hybrid model was more certain about its outputs. However, this did not correlate with accuracy, indicating a possible misalignment between the confidence measures and the actual correctness of the outputs. This might be due to BERT reinforcing Whisper's mistakes rather than correcting them, especially when BERT's predictions were off-target for the nuances of spoken language captured in the audio.
- **Overfitting to Textual Context**: BERT's strong language modeling capabilities may have led to overconfidence in contexts where its training on written text did not align well with the spoken form. For example, BERT might generate high-confidence predictions based on common textual patterns which do not necessarily match the acoustic peculiarities or informal expressions present in spoken language.
- **Gamma Factor Implications**: The weighting factor gamma, used to balance the influence of Whisper and BERT, could have disproportionately favored the textual context provided by BERT, especially in ambiguous cases. This might have resulted in higher confidence scores but less accurate transcriptions, reflecting a false assurance in the contextual guesses made by BERT. In our code, a gamma value of 0.5 was used for equal weighting of Whisper and BERT, along with a confidence threshold of 0.5.

4.7 Limitations

Due to the length of the dataset, it took a long time to process when validating and testing our model. We decided to validate our threshold on only 200 and 1000 samples; but the distribution of the maximum probabilities looked similar, which helped us determine the threshold. In addition, we tested our dataset only a subset of word lengths, but varied these (1-30, 50-100, 15-100) these in order to validate our results. A large majority of the samples had less than 100 words.

While the dataset itself is large, it is also limiting in the way that there is only 40 hours of data so it may not encompass all of the challenges in ASR, such as additional accents or variations due to geographic region or socioeconomic status. During the pre-processing, we also took out samples with number since our ASR model had inconsistencies in transcription, which decreased our available dataset. While we normalized the text with several measures described in section 3.7, there still may have been inconsistencies in the texts that lead to a higher word-error rate.

4.8 Improvements and Future Work

Given our observations regarding the Hybrid approach, future work could aim to address the disconnect between confidence levels and transcription accuracy:

1. **Calibration of Confidence Measures**: Developing methods to calibrate the confidence scores so that they better reflect the true accuracy of the transcriptions. This could involve adjusting the model's confidence threshold or employing techniques from probabilistic modeling to assess confidence more judiciously.
2. **Refinement of Integration Strategy**: Refining the strategy for integrating BERT into the transcription process. This might include dynamic weighting mechanisms that adjust gamma based on the type of speech or the degree of ambiguity in Whisper's initial transcription.
3. **Enhanced Contextual Adaptation**: Enhancing BERT's training to include more spoken language data and scenarios with varied accents and informal styles. This adaptation could help BERT provide more accurate context understanding that aligns with the acoustic signals Whisper processes.

References

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24, Melbourne, Australia. Association for Computational Linguistics.

Alex DiChristofano, Henry Shuster, Shefali Chandra, and Neal Patwari. 2023. Global performance disparities between english-language accents in automatic speech recognition. arXiv:2208.01157. Version 2.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).

M. V. Koroteev. 2021. BERT: A review of applications in natural language processing and understanding. *CoRR*, abs/2103.11943.

Zeping Min and Jinbo Wang. 2023. Exploring the integration of large language models into automatic speech recognition systems: An empirical study.

Andrew C. Morris, Viktoria Maier, and Phil D. Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. *INTERSPEECH*.

Pablo Ortiz and Simen Burud. 2022. Bert attends the conversation: Improving low-resource conversational asr.

Darshan Prabhu, Preethi Jyothi, Sriram Ganapathy, and Vinit Unni. 2023. Accented speech recognition with accent-specific codebooks. arXiv:2310.15970v3. Version 3.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. 2022. Mitigating neural network overconfidence with logit normalization.