SPONTANEOUS SPEECH QUESTION-ANSWERING EXTRACTION WITH UNDER-REPRESENTED DIALECTS AND SPEAKING STYLES

Kavan Mehrizi[1*] *Nadia Pelaez*[2*] *Annie Villalta*[3*] *Alexander Johnson*[4] *Abeer Alwan*[4]

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[2] Department of Computer Science and Engineering, University at Buffalo
[3] Department of Computer Science, Stanford University
[4] Department of Electrical and Computer Engineering, University of California, Los Angeles
kavanmehrizi@berkeley.edu, nadiapel@buffalo.edu, anniev18@stanford.edu

ABSTRACT

Automatic Question Answering (QA) from spontaneous speech performs worse when the speech is from an underrepresented dialect such as African-American Vernacular English (AAVE). In this work, we experiment with various methods of increasing QA performance on spontaneous AAVE speech including prompt engineering to fine-tuning. However, significant challenges arose with prompt engineering such as generating irrelevant or false information and ignoring given prompts. Therefore, we adopted fine-tuning as a more effective and reliable method. To decrease transcription errors that negatively impact information retrieval by QA models, we use the Corpus of Regional African American Language (CORAAL) database to train and evaluate the Whisper Automatic Speech Recognition (ASR) system. We then input the transcript generated by Whisper to the QA model DeBERTa and compare its performance with large language models (LLMs) such as LLaMA-2. By utilizing Meta's LLaMA-2 and fine-tuning OpenAI's Whisper ASR system, preliminary results show improved QA performance compared to conventional question answering models.

Index Terms— Neural Networks, Question-Answering, Automatic Speech Recognition Systems, Large-Language Models, African-American Vernacular English

1. INTRODUCTION

Automatic Spontaneous Speech (ASR) systems have been incorporated into various facets of daily life, from virtual assistants like Amazon Alexa to speech recognition technology such as Youtube's automatic captioning. However, like many other machine learning tools, ASR systems hold biases that degrade performance for some users [1]. To evaluate ASR performance, researchers often use the word error rate (WER), which measures the number of substitutions, deletions, and insertions made by the system. A Stanford University Policy Lab study found that racial disparities in these systems led to high WERs for certain ethnic groups [2]. In the end, researchers found a positive correlation between the usage of African American Vernacular English (AAVE) in speech and high WERs. These insights shed light on the biases that continue to exist within ASR technology and demonstrate the importance of pursuing research in this field.

Within this context, our research aims to improve question-and-answer (QA) models, particularly within the realm of spontaneous speech and underrepresented dialects. The central focus of our research lies in the comparative analysis of two prominent models: LLaMA-2 and DeBERTa. Transcription errors further compound the challenges faced by QA models, affecting the system's overall accuracy. To address this, we use the CORAAL database to train and evaluate the Whisper ASR system. The transcripts generated by Whisper subsequently serve as the foundation for input into the DeBERTa QA model and LLaMA-2. Through the comparison of these models and the finetuning of Whisper, we aim to advance the accurate interpretation of AAVE by ASR systems, contributing to the ongoing endeavor of refining language processing technologies to be more inclusive.

2. BACKGROUND

Whisper is a speech transcription API trained on the common voice corpus [3], an open-source dataset with various languages spoken by the general public in order to aid automatic speech recognition in a diverse setting. In this work, we utilize Whisper for transcription with audio files from CORAAL and fine-tune to better understand AAVE and improve accessibility in both ASR and QA models. Whisper makes on average 55.2% less errors with the LibriSpeech test-clean dataset compared to the LibriSpeech test-other dataset [5, 6]. Although the LibriSpeech test-clean contains carefully chosen clear audio with no stutters, background noise, or other distracting factors, this quality is not reflective of spontaneous speech heard in the real world. LibriSpeech test-other con-

*Equal contribution. Work performed while at the University of California, Los Angeles Science Hub for Humanity and Artificial Intelligence

Thank you to Amazon Science and the Amazon Summer Undergraduate Research Experience at the University of California, Los Angeles (SURE @ UCLA) for funding this work.

tains both clear and unfiltered audio, making it more representative of real-world spontaneous speech. Often automatic speech recognition technologies are developed and trained with the lack of labeled data for different speaking styles, accents, or dialects. Therefore, ASR systems trained only on clean speech often struggle with noisy or spontaneous speech [7, 8]. To improve speech recognition with Whisper, the system must be fine-tuned on a dataset with underrepresented dialects like CORAAL which can later be used in question-answering models without the inaccuracy of the transcription issues [9, 10, 11].

However, the transformer-based architecture of Whisper shows promising capabilities in fine-tuning. Transformers are a recent neural net architecture that excels in performance and runtime due to its ability to work in parallel. Transformers utilize self-attention mechanisms which have proven to be highly effective in identifying connections among tokens within a sequence, and cross-attention mechanisms, which identify relationships between tokens in an input sequence to those in an output sequence. Therefore, these systems are particularly well-suited for addressing sequence-to-sequence tasks [5, 12]. For example, transformers have been used effectively in mapping audio frame sequences into corresponding sequences of textual tokens. In addition, they are effective in natural language processing tasks such as question answering [12].

The language model, DeBERTa, also utilizes transformers that use attention mechanisms to better understand text-based tasks. The model is able to calculate attention weights and understand dependencies and significance between words within the text [13]. DeBERTa surpasses other language models in text comprehension, making it the chosen QA model for this research [14].

With the growing popularity of large language models (LLMS) such as ChatGPT, we will also be utilizing the LLM, LLaMA-2, for question answer extraction and comparing it with DeBERTa. LLaMA-2 has recently been updated in July 2023 and outperforms most open-source chat models on the benchmarks that it was tested on [15]. LLaMA-2 exhibits promising capabilities in situations requiring information retrieval and is available for research purposes at no cost [16]. In this investigation, we will utilize the large LLaMA-2 model which has 70 billion parameters for our analyses. The methods section will demonstrate a thorough comparison between LLaMA-2 and DeBERTa on extractive QA with questions from the CORAAL dataset. For our analysis, we use Whisper for automatic speech recognition and transcription as shown in Figure 1. The aim is to develop a model proficient in understanding and performing tasks involving African-American Vernacular English (AAVE).

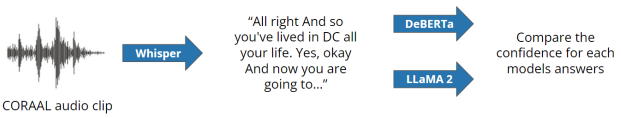

Fig. 1. General process of how Whisper, DeBERTa, and LLaMA-2 will be used. Audio clips from CORAAL are taken and passed through Whisper. Whisper then transcribes the audio into text and is given to DeBERTa and LLaMA-2 for question answer extraction. We will be analyzing the confidence of each model's answers based on the question given to determine which model is best to move forward with.

3. METHODOLOGY

3.1. Datasets

Developed by the University of Oregon, the Corpus of Regional African American Language (CORAAL) is the first public corpus of African American Language (AAL) speech data. CORAAL features recorded speech from regional varieties of AAL and includes audio recordings along with ground truth transcriptions from over 220 sociolinguistic interviews with speakers born between 1888 and 2005.

We utilize the CORAAL:DCB subset [16] as the finetuning dataset for Whisper. This dataset consists of 63 audio files recorded from 48 different speakers in the Washington D.C. area. The recordings were collected between July 2015 and December 2017 where speakers were selected through a friend-of-a-friend network to best represent four age groups and three social class groups.

We utilize the CORAAL:PRV subset [17] to evaluate fine-tuned Whisper model. The CORAAL: PRV dataset consists of 16 primary speakers across 32 audio files in Princeville, NC. The speakers were recorded between August 2003 and June 2004 and are between 19 and 83 years old.

3.2. Automatic Speech Recognition System

We utilize OpenAI's neural network ASR model, Whisper, [5] for our speech-to-text engine. Whisper is trained in a supervised manner on a large and diverse dataset of 680,000 hours of multilingual speech collected from the web. The input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then is passed into the encoder. A decoder is trained to predict the corresponding text caption from the input audio sequence. Whisper uses the transformer architecture [4] which has an encoder-decoder structure. The encoder and decoder each consist of a stack of identical layers, where each layer has two modules: a multi-head self-attention mechanism that allows the model to learn the dependencies between different parts of the input and output sequences, and a fully connected feed-forward network that applies two linear transformations.

3.3. Background Setup

In the initial phase of our study, the audio files and accompanying ground truth transcripts sourced from the CORRAL database are uploaded to our designated repository within Google Drive. Subsequently, mounted our Google Drive to Google Colab where all the programming took place. In Google Colab, we performed a meticulous cleaning process for the ground truth transcripts. Redacted statements, laughter, pauses, and any other non-contextual information in the ground-truth transcripts were removed. Word error rate is the ratio of errors in a transcript to the total number of words spoken. This allows us to analyze the performance difference of Whisper with and without the help of outside factors explored in section 3.4. Cleaning the ground truth transcript enables a precise word error rate calculation, as Whisper does not transcribe these excised statements.

3.4. Selecting an Approach for Speech Processing

We explored two primary approaches for enhancing transcription accuracy using Whisper: prompt engineering and fine-tuning. Prompt engineering is an emerging research area with promising potential for generating specific pieces of information. This approach emphasizes the utilization of prompts, such as lists of location names or entities, to rectify transcription errors in the Whisper system. For large language models, this can include inquiries, instructions, or statements provided by the user that are then used to generate a response.

Upon transcribing the CORAAL:DCB dataset with Whisper, we conducted a comparative analysis between the generated transcripts and the ground truth transcripts to identify the types of errors made by Whisper. In instances where Whisper encountered challenges in recognizing location names, we experimented with providing a prompt in the form of a list of locations. The intention was to assist Whisper in considering this additional information when generating the transcript for the audio. However, this approach did not address the fundamental issues associated with Whisper's ability to transcribe spontaneous speech with African-American Vernacular English (AAVE). Whisper would continue to incorrectly transcribe locations, names, and make grammatical errors since AAVE contains patterns that were not seen in the original training data of Whisper. Consequently, we observed an increase in the word error rate when testing the CORAAL:PRV subset, potentially impacting the performance of the question-answer extraction process. Moreover, during the use of prompt engineering, we observed instances of hallucinations in Whisper's output, where the system generated non-existent words or combined words from the provided prompt. This raised concerns about the reliability of the approach, prompting us to explore alternative methods of improvement.

Subsequently, the fine-tuning approach emerged as a compelling direction for our investigation. Fine-tuning proved particularly suitable for our research objective of mitigating racial bias within Whisper's transcription capabilities. In this method, we built upon the pre-existing knowledge of the Whisper system by providing it with the CORAAL:DCB subset as supplementary data for training. This fine-tuning procedure yielded substantial performance enhancements in Whisper's transcription of the testing audio files from the CORAAL:PRV dataset. From these outcomes, we determined fine-tuning to be the optimal and conclusive approach for audio transcription with Whisper.

3.5. Comparing LLaMA-2 vs DeBerta

Having established fine-tuning as the preferred approach for audio transcription, we then proceeded to compare two prominent language models, LLaMA-2 and DeBERTa, in their roles as question-answer models. We assessed the performance of these models by analyzing the precision, recall, and F1 scores of correctly detecting the answer span. These metrics are widely used in machine learning for evaluating a classifier's ability to detect the target.

To begin the comparison between LLaMA-2 and DeBERTa, we segmented the entire audio file into 2-minute intervals with an overlap of 30 seconds between adjacent segments. After transcribing each 2min segment of audio through the Whisper model, a set of questions associated with each audio file were inputted into both the LLaMA-2 and DeBERTa models. The purpose was to ascertain the models' ability to detect answers to posed questions upon looping through each chunk. For DeBERTa, this assessment compared the confidence intervals accompanying DeBERTa's responses against a pre-defined threshold. his threshold selection was motivated by the consistent observation that DeBERTa's responses exhibited a confidence level of 0.4 or above when identifying valid answers within the given audio segments. However, LLaMA-2 did not have a predefined confidence score. Recognizing that LLaMA-2 operates as a language model heavily reliant on the provided prompt, we devised the following prompt strategy in Figure 2.

The decision to utilize '-1' as the indicator was largely influenced by the nature of the questions in the audio files. The questions in the audio files included answers that weren't always numerical. If numerical responses did arise, they were always positive numbers. This helped prevent accidentally categorizing a correct answer as undetected by LLaMA-2. As a result, our evaluative criterion for whether LLaMA-2 detected an answer in a chunk revolved around whether the answer contained '-1' or not.

In order to evaluate how well both models performed, we compared them against a set of ground truth answers or a set of expected answers. For this evaluation, we constructed three matrices: the ground truth matrix, the DeBerta matrix, and the LLaMA-2 matrix. The ground truth matrix is built using human-extracted questions from the CORAAL audio files

```
result = client.predict(
"Instruction: Look at the following
story and in one sentence only,
if stated directly in the story
answer the following question
with only the answer do not make
predictions:" + llama_question +
"If the answer to this question is
not found in the story then say
'-1' else don't say anything.
transcript:" + llama_transcript,
api_name="/chat_1")
```

Fig. 2.

along with their corresponding answer timestamps. These questions are compiled in a document that serves as the reference for the correct answers. Next, we create the predicted labels for LLaMA-2 and DeBERTa. Each matrix represents the presence or absence of an answer to a specific question in each 2-minute audio chunk. If the answer to a question is found within a particular chunk, the corresponding entry in the matrix is marked as 1; otherwise, it is marked as 0. An example of this is depicted in Figure 3.

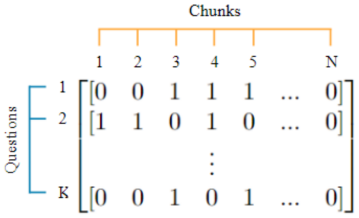

Fig. 3. A predictive matrix including N chunks and K questions contained in an audio file.

By aligning the generated answers from the models with the ground truth matrix, we can calculate the F1 scores for each model. We utilized the F1 score support function from the scikit-learn library to assess each row within our matrices, juxtaposed with their corresponding ground truth rows. This approach yielded individual F1 scores, as well as precision and recall values, for every row within the matrices. Subsequently, we computed the average scores for each matrix, effectively capturing the average performance across the rows in each file. To determine the overall performance across all 20 files, we calculated the mean of these average scores. This final aggregation provided us with comprehensive F1 scores, precision, and recall metrics, collectively characterizing the performance of our approach across the entire dataset of 20 files.

4. RESULTS

4.1. Prompt Engineering with Whisper

Our first approach was to use Whisper's ability to enter a prompt to mitigate word errors. We created these prompts around noticeable patterns we saw in Whisper's transcripts such as spelling city names wrong, or not understanding some words pronounced in AAVE such as 'imma'. We would input these prompts and compare the original and ground truth transcripts with the one created with prompts.

We found significant challenges and drawbacks to this approach. A notable concern was the presence of hallucinations, where Whisper would produce text outputs that were either garbled from or not present in the original speech. Because of these hallucinations, question-answering performance was degraded, where in some instances the answer given would be from the hallucinations.

As a consequence of these challenges, we concluded that prompt engineering, while a promising path, fell short of mitigating the word errors associated with AAVE. Because of this, we deemed it necessary to move on and explore alternative strategies to enhance the performance of QA from spontaneous speech, focusing on a more holistic approach that addresses both ASR and QA performance.

4.2. Fine-Tuning Whisper

Our next approach showed more promising results. Testing on 40 CORAAL audio files, we found that regular Whisper performed with a WER of 41%, in line with the average WER of AAVE found by a Stanford University Policy Lab study. Our fine-tuned Whisper model performed with a WER of 37%, a 4% improvement as shown in Table 1. This is significant as the fine-tuned Whisper model was only trained on a small subset of less than 40 audio-transcript pairs.

We tested DeBERTa on both our fine-tuned Whisper model and the regular, non-fine-tuned Whisper. DeBERTa performed well using regular Whisper, with a precision score of 0.666, a recall score of 0.752, and an F1 score of 0.669. DeBERTa performed similarly using our fine-tuned Whisper model, with a precision score of 0.666, a recall score of 0.752, and an F1 score of 0.673 as shown in Table 2.

Whisper	Without fine-tuning	With fine-tuning
Word Error Rate (WER)	41%	37%

Table 1.

These results indicate that DeBERTa is robust to word errors in speech transcriptions and can effectively answer questions from spontaneous speech.

Model Used	Precision	Recall	F1-score
DeBERTa w/o fine-tuned	0.666	0.752	0.669
w/ fine-tuned whisper model	0.666	0.752	0.673

Table 2.

4.3. LLaMA-2

We attempted to utilize LLaMA-2, a large-language model (LLM), to understand if LLMs could perform better than models designed specifically for QA, such as DeBERTa. We planned to test LLaMA-2 on both our fine-tuned Whisper model and the regular, non-fine-tuned Whisper, using the same dataset and metrics as we did for DeBERTa.

However, we encountered several challenges and difficulties in using LLaMA-2 for our task. As previously stated, LLaMA-2 does not natively support confidence scores for its generated texts, which are essential for evaluating the quality and reliability of the answers. We tried to implement a custom confidence scoring function based on keywords in LLaMA-2's answers, as well as prompting LLaMA-2 before the context to return '-1' if it is not confident in its answer, but we found that it was not consistent or reliable across multiple questions and transcripts.

Second, we also faced issues with Google Colab and the LLaMA-2 API crashing frequently during our experiments. We suspect that this was due to the long execution times and memory consumption of LLaMA-2, which exceeded the limits of our free Google Colab account. We also experienced long delays and timeouts when sending requests to the LLaMA-2 API, which made it difficult to run multiple trials and compare results.

Due to these issues, we were not able to obtain any significant results for LLaMA-2 on our question-answering task from spontaneous speech.

5. FUTURE DIRECTIONS

Our research has shed light on several promising avenues for further exploration and improvement in the realm of question answering on spontaneous speech. As we move forward, the following directions present new opportunities for enhancing the accuracy of DeBERTa and LLaMA-2.

5.1. Confidence Score for LLaMA-2

One key area for future development involves the implementation of a confidence scoring mechanism for LLaMA-2. As LLaMA-2 does not inherently provide confidence scores for its generated responses, developing a reliable confidence estimation method becomes crucial. While DeBERTa already leverages a confidence score feature to evaluate the reliability of its responses, a similar tailored mechanism for LLaMA-2 holds great promise By introducing a quantifiable measure of confidence in LLaMA-2's answers, we aim to enhance the reliability of our model's predictions. This would facilitate a more comprehensive evaluation of LLaMA-2's performance and could bolster its integration into the broader landscape of question answering systems.

5.2. Fine-Tuning with Larger and Diverse Datasets

To further enhance the robustness and inclusivity of our models, exploring fine-tuning with larger and more diverse datasets holds significant promise. Expanding our training data to encompass a broader spectrum of dialects, accents, and spontaneous speech patterns could lead to improved accuracy in transcription and subsequently more accurate question answering. By capturing a more comprehensive range of linguistic variations, our models would become better equipped to address the inherent challenges posed by spontaneous speech.

5.3. Transitioning to Longer Audio Files

As we continue to advance our research, the transition from 2-minute audio clips to longer audio files, perhaps spanning 1 hour or more, emerges as a logical progression. This shift reflects real-world scenarios more accurately, where extended conversations and dialogues are prevalent. Adapting our models to handle longer audio files will require addressing new challenges related to context preservation, memory efficiency, and computational resources.

6. CONCLUSION

Our exploration into question answering on spontaneous speech in underrepresented dialects has uncovered numerous pathways for further refinement and innovation. By embracing these future directions, we strive to contribute to the development of more accurate, reliable, and inclusive language processing technologies that effectively serve diverse linguistic backgrounds and contexts.

7. ACKNOWLEDGEMENTS

This work was supported and sponsored by the Amazon Summer Undergraduate Research Experience (SURE) program and UCLA Samueli School of Engineering Center for Excellence in Engineering and Diversity (CEED). We would like to thank our mentor, Alexander Johnson, and our Principal Investigator, Abeer Alwan, for giving us the opportunity to be a part of their research lab and for their amazing support throughout this experience. We would also like to thank Catherine Douglas and our cohort for their constant support.

8. REFERENCES

[1] J. Larson, J. Angwin, L. Kirchner, and S. Mattu, "How we analyzed the compas recidivism algorithm," *ProPublica*, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed Aug. 10, 2023).

[2] A. Koenecke et al., "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[3] R. Ardila et al., "Common Voice: A Massively-Multilingual Speech Corpus," *arXiv preprint*, arXiv:1912.06670, 2019.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," *arXiv preprint*, arXiv:2212.04356, 2022.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

[6] E. Shriberg, "Spontaneous Speech: How People Really Talk and Why Engineers Should Care," *CITESEERX*, 2005. doi: 10.1.1.73.3765

[7] M. Ostendorf, E. Shriberg, and A. Stolcke, "Human language technology: opportunities and challenges," *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 2005, vol. 5, p. v/949-v/952 Vol. 5. doi: 10.1109/ICASSP.2005.1416462.

[8] B. Talafha, A. Waheed, and M. Abdul-Mageed, "N-shot benchmarking of whisper on diverse Arabic speech recognition," *arXiv preprint*, arXiv:2306.02902, 2023.

[9] G.-T. Lin et al., "Dual: Discrete spoken unit adaptive learning for textless spoken question answering," *arXiv preprint*, arXiv:2203.04911, 2022.

[10] N. T. Vu, Y. Wang, M. Klose, Z. Mihaylova, and T. Schultz, "Improving ASR performance on non-native speech using multilingual and crosslingual information," *Proc. Interspeech 2014*, pp. 11–15. doi: 10.21437/Interspeech.2014-3.

[11] A. Vaswani et al., "Attention is all you need," *arXiv preprint*, arXiv:1706.03762, 2023.

[12] K. Nassiri and M. Akhloufi, "Transformer models used for text-based question answering systems," *Applied Intelligence*, vol. 53, no. 9, pp. 10602–10635, 2023.

[13] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced Bert with disentangled attention," *arXiv preprint*, arXiv:2006.03654, 2020.

[14] H. Touvron et al., "LLaMA-2: Open Foundation and fine-tuned chat models," *arXiv preprint*, arXiv:2307.09288, 2023.

[15] B. Bohnet et al., "Attributed question answering: Evaluation and modeling for attributed large language models," *arXiv preprint*, arXiv:2212.08037, 2022.

[16] Kendall, Tyler, Minnie Quartey, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler. *The Corpus of Regional African American Language: DCB (Washington DC 2016)*, Version 2018.10.06, 2018. Eugene, OR: The Online Resources for African American Language Project.

[17] Rowe, Ryan, Walt Wolfram, Tyler Kendall, Charlie Farrington, and Brooke Josler, *The Corpus of Regional African American Language: PRV (Princeville, NC 2004)*, Version 2018.10.06, 2018. Eugene, OR: The Online Resources for African American Language Project.