Johnson, A., Chance, C., Stiemke, K., Veeramani, H., Shankar, N. B., & Alwan, A. (2023). An Analysis of Large Language Models for African American English Speaking Children's Oral Language Assessment. *Journal of Black Excellence in Engineering, Science, & Technology, 1.*

BLACK EXCELLENCE IN ENGINEERING, SCIENCE & TECHNOLOGY MANUSCRIPTS

An Analysis of Large Language Models for African American English Speaking Children's Oral Language Assessment

Alexander Johnson[1], Christina Chance[1], Kaycee Stiemke[1], Hariram Veeramani[1], Natarajan Balaji Shankar[1], Abeer Alwan[1]

[1] Electrical and Computer Engineering, University of California, Los Angeles

Keywords: Speech Recognition, African American English, Education, Language Model, Assessment Scoring

Journal of Black Excellence in Engineering, Science, & Technology

Vol. 1, 2023

This paper evaluates the performance of widely-used open-source automatic speech recognition systems in transcribing primarily African American English-speaking children's speech for educational applications. We investigate the performance of the Whisper, HuBERT, and Wav2Vec2 ASR systems as well as the capability of the transformer-based language model, BERT, for automatically grading the student's oral responses to assessment prompts through use of the generated ASR transcripts. We achieve a 95% oral response scoring accuracy through the methods described. We also show a thorough analysis of ASR system performance over a diverse set of metrics going beyond the standard word error rate.

Introduction

Artificial intelligence has the potential to greatly improve outcomes in education. Researchers have long been interested in creating systems that automatically score educational assessments (Shibata & Uto, 2022), provide realtime feedback on reading and speaking practice (Williams et al., 2000), and interact with children to stimulate play and growth (Breazeal, 2003). For example, the "Read Along App" by Google uses automatic speech recognition (ASR) technology to listen to children as they practice reading and offer feedback when the student appears to have difficulty (Google, n.d.). These systems can save time for teachers and enable parents to give their children more speaking and literacy practice at home. With large language models and artificial intelligence, non-experts could provide children with expert-level language instruction or cheaply test children for language/reading impairments in order to offer early intervention. However, further work is needed with such devices before they can be considered both high-performing and equitable across diverse demographics of speakers. ASR for children's speech is a difficult task because of the high variability in oral language that children display as they develop (Dutta et al., 2022; Lee et al., 1999). To train ASR systems, researchers use a large amount of audio recordings containing speech samples together with corresponding hand-written transcripts for the words said in the recordings. Researchers then program the machine to calculate an optimal mapping function from the numerical representations of the audio to the characters in the transcripts (Rabiner & Juang, 1993). That mapping function can then be used to estimate the transcript of new audio files during inference. The success of these systems is heavily dependent on the similarity between the training data and the inference audio. As such, ASR systems trained only on adult speech data often struggle to capture the pronunciations and syntax unique to children. The performance of several

popular ASR systems have also been shown to degrade for underrepresented dialects such as African American English (Koenecke et al., 2020). They also show worse performance for speakers with a higher dialect density, or more frequent use of language characteristics that are not found in the mainstream dialect. When these systems are trained only on Mainstream American English, they do not learn to infer dialect or accent-specific speech patterns that were not present in the training set, often resulting in erroneous transcriptions. Several transformer-based ASR frameworks have been developed in recent years to improve system performance even for low-resource or underrepresented language. For example, Meta's Wav2Vec2 uses unsupervised pre-training, which means that it is first conditioned on a large amount of unlabeled audio data before starting the process of learning from audio with transcripts (Baevski et al., 2020). Hubert improves on this process by using unsupervised clustering to assign pseudo-labels to the audio data and then train with those artificially generated labels (Hsu et al., 2021). Recently, OpenAI's Whisper has also achieved great improvements in ASR through supervised training on 680k hours of mined audio data (Radford et al., 2022). In particular, Whisper claims improved performance over Wav2Vec2 for the AAE speech test set from the Corpus of Regional African American Language (Kendall & Farrington, 2021). Improvements in ASR systems are most commonly shown by demonstrating the newer system gives a lower word error rate (WER), which is calculated as the number of words incorrectly transcribed (substitutions), missed (deletions), or added (insertions) in the ASR transcript divided by the number of total words in the human-labeled transcript. While Whisper shows lower WER than the other systems for several adult speech datasets, few such audits have been conducted on large speech systems for AAE children's speech. Furthermore, for educational applications, it is important to also use metrics which reflect an ASR system's performance for downstream learning and assessment tasks. For example, if a teacher wanted to use ASR systems to transcribe and provide feedback on a student's grammar usage during an oral assessment, they would likely want assurance that the ASR system both has a low WER and accurately captures the student's grammar patterns. As WER alone gives no indication as to which parts of speech or speech patterns were incorrectly transcribed, metrics that provide additional language information for educators are necessary for this application. This paper 1) conducts an audit of commonly-used open-source ASR systems on a children's speech dataset containing recordings of AAE and non-AAE speaking 3rd-8th graders from the Atlanta, Georgia area, and 2) evaluates the performance of education-centered metrics for each of systems in addition to the standard WER.

Methods

In this work, we take a quantitative approach to assess the usability of a variety of metrics and assessment mechanisms to support more dialectically dense speech and text. We use various metrics to capture speech complexity and narrative-building skills and assess alignment with our metrics and standard

metrics used for educational assessment. This process will determine the correlation between metrics used by language education experts to assess diverse students and those used in large language models to assess performance which will in turn describe the aptness of large language models for use in common classroom tasks such as oral assessments or reading practice. A high agreement between human-scored metrics and machine-scored metrics would suggest that these systems could be readily applied to educational tasks, and a low agreement would show in what ways language models and artificial intelligence must be improved before it can be applied to early language instruction. For our metrics, we choose approaches that align with many of the factors that educators look for during their assessments that are not related to the dialect the children speak. These metrics allow for a more robust and transferable assessment by not coupling the performance scores with a specific dialect. We additionally use a language model, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), to study whether automated assessment is plausible for this task. BERT is a transformer-based model, meaning it uses an attention mechanism to understand and learn the relationships between the words. Its bidirectionality allows the model to gain a deeper understanding of the text and allows it to perform masking. Masking occurs when the model masks or hides a word and looks at the surrounding context, both before and after the masked word, to predict the word. BERT is more powerful because it allows multiple representations of the same word based on the varying context. Dataset This work uses the Georgia State University Kids' Speech Dataset (GSU Kids), a speech dataset of 191 children aged 8-13 from the Atlanta, GA area (Fisher et al., 2019; Johnson et al., 2022). The children were recorded in a standard classroom environment as they performed educational assessments in story-retelling and picture-description tasks. Each child was recorded for approximately 5-10min resulting in about 20 hours of total speech. The educational assessments included three tasks from the Test of Narrative Language (Gillam & Pearson, 2017): one story-retelling task, in which each child was asked to repeat a story told to them, and two picture description tasks, in which each child was asked to tell a story about an image shown to them. The story retelling task was graded based on the child's correct repetition of designated keywords that are essential to the narrative (eg. character names, setting, action verbs, etc.). The child must have said these keywords in the correct order and tense to receive credit for them. For example, suppose that the child were read the sentence, "Sam and Jordan played basketball in the park," where the bolded words are the graded keywords. If the child repeated the sentence as "Jordan plays basketball in the park" then they would receive credit for two of the four keywords since the others are missing or incorrectly conjugated. The picture description tasks were graded based on the child's inclusion of key story elements (eg. character names, rising action, resolution, etc.) as well as their correct use of grammar, coherence of the narrative, and completeness of the description of the image shown. The audio was recorded at 44.1kHz and later downsampled

to 16kHz for experimentation. The children's assessments were transcribed and graded by experts in children's language and subsequently annotated for aspects of AAE dialect by the authors of this paper according to the protocol in (Koenecke et al., 2020). 124 of the children spoke with characteristics of AAE.

Experiments

This paper seeks to provide a comprehensive analysis of the performance of state-of-the-art ASR and NLP systems in transcribing and understanding AAE-speaking children's spoken language. First, we look at the performance of Whisper-Large, Wav2Vec2-Large, and Hubert-Large in transcribing children's AAE speech. In addition to word error rate, we also evaluate: • Character Error Rate (CER): WER as a metric gives no partial credit for narrow misspellings. However, CER, a similar measure that gives the percentage of characters mis-transcribed by an ASR system, would show smaller error for a close misspelling. For example, if a person said the phrase "Call her," and an ASR system transcribed it as "Caller," the WER would be (1 substitution + 1 deletion)/(2 words) = 100%. However, the CER would be (2 deleted characters)/(8 total characters) = 25%. We investigate the CER of these systems to determine if close misspellings may be unfairly penalized by looking at WER alone. • Semantic Similarity: Another shortcoming in WER is that it weights all words in the sentence equally. For example, suppose that a person uttered the phrase "I am a tall person." If one ASR system interpreted this as "I am tall," and another system interpreted this as "I am a small purse," both of these outputs would receive the same WER, as they have both transcribed two words incorrectly. However, the first ASR system has preserved the sentence's meaning much more faithfully. Proposed for use in ASR in (Kim et al., 2021), semantic similarity is a metric that aims to capture this. We use the pre-trained transformer-based architecture from (Reimers & Gurevych, 2019) in this task. Here, Sentence-BERT first calculates a compressed vector representation of the text designed to numerically represent the important features of the input text. The semantic similarity of two texts is then calculated as the cosine distance between the BERT-representations of the texts. We use this method to automatically encode and measure the interpreted similarity between the ground truth transcripts and ASR transcripts for the children's speech. • Mean Length Communication Unit (MLCU): The mean length communication unit is a unit that assesses the complexity of a sentence (Nutter, 1981). This is calculated from the average count of verb phrases used in modifying clauses in a sample per each sentence. For example, the utterance, "The boy who wore yellow likes chocolate. The girl likes vanilla" has 2 communication units in the first sentence (1 modifying clause and one main clause) and one communication unit in the second sentence, and so the MLCU is 1.5. The MLCU has been proposed as a dialect-invariant measure of grammatical abilities, as grammatical differences between AAE and MAE often lead AAE-speaking children to be under-rated in language abilities (Craig & Washington, 2000). Utilizing part of speech tagging (POS), we cluster verb clauses or verb phrases and count the frequency of unique phrases per sentence. • Vocabulary

Size: The vocabulary size count is a metric to assess the variability of words used in a text and word count captures the overall number of words in a text (Milton & Treffers-Daller, 2013). These metrics are used to assess the complexity and depth of vocabulary that the child utilizes. The vocabulary count captures the number of unique words present in the text while the word count captures the overall number or words. Here, we assess similarity between the vocabulary counts of the children's ground truth and ASR transcripts. • Word Count: In addition to vocabulary size, we also examine the total number of words stated by each child and compare that to the number of words captured by the ASR system. Last, an ultimate goal of large language models may be to automatically assess children's narrative language abilities. The presence of automatic assessments would alleviate the burden on teachers and enable people without access to language specialists to conduct battery assessments for reading or language disorders at home. Following (Johnson et al., 2023) we use the language model, BERT, appended with a fully connected classification layer, to automatically classify the ASR transcripts into 1 of 5 classes (corresponding to students who scored 0% - 20%, 20% - 40% , 40% - 60% , 60% - 80%, and 80%-100% respectively). The BERT model itself takes text as input and outputs a compressed numerical representation of the input. This compressed representation ideally only preserves crucial information about the input text such as the general topic or information needed to create a compact summary. We then appended a fully connected neural network layer for classification layer to the BERT model which learns to classify the score of the input transcript from the BERT representation. Given the small sample size of the data, we additionally use text data from the WeeBit corpus (Vajjala & Meurers, 2012) to jointly train the system on ASR transcripts and text data. We then train the BERT model to learn a representation of the input text that can easily be mapped to the grammatical complexity and amount of detail in the text, correlating to an assigned grade.

Results

Table 1 shows the comparison of metrics across the Whisper, Hubert, and Wav2Vec2 ASR Systems for the children's speech. The table first shows the word error rate and character error rate between the ground truth transcripts and ASR transcripts (for each of the three systems) for each dialect demographic. The table then shows the semantic similarity between the ground truth text and ASR transcripts as well as the root mean square error between the average number of mean length communication units, vocabulary size, and number of words in the ground truth transcripts vs the ASR transcripts. We additionally show the effects of higher dialect density on WER in Figure 1. Figure 2 shows the average number of character-level substitutions, deletions, and insertions for the three ASR systems in both assessment tasks. For the story-retelling task, Table 2 shows 1) the ASR systems' precision (the percentage of time the system detected a keyword and there actually was a keyword in the human-labeled transcripts at that time), recall (the percentage of time that the hand labeled transcripts showed that the child used a keyword

and the ASR system correctly transcribed it), and the F1 score (macro average of precision of recall) using an 85% string similarity threshold for detection and 2) The classification accuracy and F1 score of the BERT system in predicting the student's oral assessment grade from each ASR transcript. Further analysis shows a 2% absolute drop in classification accuracy from the non-AAE speaking students to the AAE-speaking students. The Bert classification model for automatic response scoring achieves its highest accuracy of 95.6% with the Whisper ASR transcripts. This is in comparison to a 96.3% scoring accuracy achieved with the human-labeled transcripts. To analyze how well the ASR systems capture aspects of AAE for children's speech, we qualitatively examine how the most common phonological feature of AAE (ie. a difference in pronunciation) in the dataset and the most common morphosyntactic feature of AAE (ie. a difference in grammar) are represented by the ASR systems. The most common phonological feature of AAE in the dataset was a substitution of "ng" for "n" in word final position (eg. pronouncing "nothing" as "nothin"). 67 of the children displayed this pronunciation. While Whisper nearly always transcribed these pronunciations with the "g" (ie. always transcribing "nothin" as "nothing"), Hubert and Wav2Vec2 transcribed these words with no "g" (ie. as "nothin") over 50% of the time. The most common morphosyntactic feature of the dataset was use of a verb stem as past tense (eg. "He didn't know what he want to buy"). 101 of the children displayed this grammatical feature. Whisper nearly always changed this pattern to one with MAE subject-verb agreement (eg. "He didn't know what he wants to buy"). Hubert and Wav2Vec2 most often introduced spelling mistakes or incorrect words for this case but would be more likely to retain the original verb tense. (eg. "He didn't no what he want to buy").

Limitations

Current ASR technology has limited generality to out-of-domain data. For example, an ASR system not trained to recognize African American English will likely experience higher error rates than it would for a dialect that it was exposed to during training. Many ASR tools are not trained on and have very little exposure to children's speech which significantly differs linguistically from adult speech. Due to this, we expect sub-optimal performance for younger speakers. Additionally, as AAE speech data is low resource, ASR tools also have expected poor performance for that dialect. We further acknowledge that, as the participants in the study all come from one school district in the Atlanta, Georgia area, they represent a limited diversity of regional dialectal characteristics, socioeconomic statuses, and developmental influences. Further work is needed to both extrapolate the results presented to other populations and to disentangle the performance of large language models in educational tasks from other factors such as the users' socioeconomic status and experiences.

Discussion

In Table 1, we see that Whisper has the lowest WER as well as CER for all groups. There is a degradation in WER and CER for all three ASR system's performance for the AAE-speaking children as compared to the non-AAE speaking children. However, Figure 1 shows that this degradation is less steep for Whisper than for Wav2Vec2 or Hubert. All three systems show comparable performance in downstream counting of mean length communication units, suggesting that Whisper may capture several words better than the other systems but does not necessarily preserve children's use of verb tense and grammar structures more effectively. If a child uses an AAE grammar construction that is not found often in the ASR system training corpus, (eg. dropping of the auxiliary verb, as in, "This food good",) then this pattern may be equally unlikely under all three systems' language modeling. Performance in vocabulary size and word counting is more varied among the three systems. While whisper appears to apply language modeling that forces each output to correspond to a correctly spelled word, Hubert and Wav2Vec2 may output character sequences that seek to phonetically transcribe children's speech when they stutter or stop mid-word (eg. I went to Califor-), and so they may be more adept at capturing the number of unique pronunciations that the children made. For the story-retelling task, we see that Whisper had a significantly higher recall of the keywords used to score the test. As these keywords contain several proper nouns and names, Whisper may gain an advantage from having some of those names appear in its much larger training corpus. These findings are also consistent with findings in (Radford et al., 2022) that show an overall lower WER for Whisper than for competing ASR models. The higher recall of keywords also appears to translate to a much higher automatic scoring accuracy, as BERT is able to achieve a 95% classification accuracy in grading the student responses from these transcripts.

Implications

Our results suggest that language models, such as BERT, have the capacity to assess children's education speech data at an accuracy similar to human assessment. This model assessment is also less sensitive to dialectal bias compared to human assessment. While Table 1 does show that there are limitations in the overall accuracy of transcription of speakers, we see fairly similar performance between AAE and non-SAE speakers indicating that dialectal differences are not the main contributor to the rates of error. These results offer insight into how language models and automation can assist with educational assessment and provide a dialect-free and fair evaluation to support a more linguistically inclusive group of children. These results offer insights on key evaluation metrics for improving speech technology so that it can be advanced to more equitably assess oral response quality from diverse language learners and offer language feedback as a feature in educational technologies. As there is currently little work to assess the fairness of large language models

in real use cases such as the one presented here, this paper offers a foundational approach to improving artificial intelligence technology outcomes in education and other domains.

Conclusions

This paper demonstrates that large language models like Whisper and BERT have the potential to greatly benefit the field of education. When tested with speech from both AAE-speaking and non-AAE speaking children, Whisper achieves relatively good word and character error rate, semantic similarity with the ground truth transcripts, and precision and recall of keywords. We also show the ASR transcripts generated with Whisper to be useful in downstream educational tasks like automatic oral assessment response scoring, for which the BERT model achieves over 95% classification accuracy. Future work includes 1) fine-tuning these systems to better recognize AAE-specific grammar patterns and pronunciations 2) designing assessments metrics that are more invariant to dialectal patterns and can easily be interpreted by language models, and 3) integrating diverse user feedback into the process of incorporating these systems in educational practices.

Submitted: May 17, 2023 EST, Accepted: October 23, 2023 EST

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 12449–12460). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

Breazeal, C. (2003). Toward sociable robots [Socially Interactive Robots]. *Robotics and Autonomous Systems*, *42*(3–4), 167–175. https://doi.org/10.1016/s0921-8890(02)00373-1

Craig, H. K., & Washington, J. A. (2000). An assessment battery for identifying language impairments in african american children. *Journal of Speech, Language, and Hearing Research*, *43*(2), 366–379. https://doi.org/10.1044/jslhr.4302.366

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, minneapolis, mn, usa, june 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Dutta, S., Tao, S. A., Reyna, J. C., Hacker, R. E., Irvin, D. W., Buzhardt, J. F., & Hansen, J. H. L. (2022). Challenges remain in Building ASR for Spontaneous Preschool Children Speech in Naturalistic Educational Environments. *Proc Interspeech 2022*, 4322–4326. https://doi.org/10.21437/interspeech.2022-555

Fisher, E. L., Barton-Hulsey, A., Walters, C., Sevcik, R. A., & Morris, R. (2019). Executive functioning and narrative language in children with dyslexia. *American Journal of Speech-Language Pathology*, *28*(3), 1127–1138. https://doi.org/10.1044/2019_ajslp-18-0106

Gillam, R. B., & Pearson, N. A. (2017). *Test of narrative language*. Pro-ed.

Google. (n.d.). *Read along by google*. Retrieved May 2023, from https://play.google.com/store/apps/details?id=com.google.android.apps.seekhamp;hl=enUSamp;gl=US

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/taslp.2021.3122291

Johnson, A., Fan, R., Morris, R., & Alwan, A. (2022). LPC AUGMENT: An LPC-Based ASR Data Augmentation Algorithm for Low and Zero-Resource Children's Dialects. *ICASSP*. https://doi.org/10.1109/icassp43922.2022.9746281

Johnson, A., Veeramani, H., Natarajan, B., & Alwan, A. (2023). An equitable framework for automatically assessing children's oral narrative language abilities. *Proc. Interspeech*. https://doi.org/10.21437/interspeech.2023-1257

Kendall, T., & Farrington, C. (2021). The Corpus of Regional African American Language. *Version 2021.07*. http://oraal.uoregon.edu/coraal

Kim, S., Arora, A., Le, D., Yeh, C.-F., Fuegen, C., Kalinli, O., & Seltzer, M. L. (2021). Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *arXiv Preprint*.

Koenecke, A., Nam, 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*, *117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455–1468. https://doi.org/10.1121/1.426686

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4*(1), 151–172. https://doi.org/10.1515/applirev-2013-0007

Nutter, N. (1981). Relative merit of mean length of t-unit and sentence weight as indices of syntactic complexity in oral language. *English Education*, *13*(1), 17–19.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv Preprint*.

Reimers, N., & Gurevych, I. (2019). *Sentence-bert: Sentence embeddings using siamese bert-networks* (K. Inui, J. Jiang, V. Ng, & X. Wan, Eds.; pp. 3980–3990). Association for Computational Linguistics. http://dblp.uni-trier.de/db/conf/emnlp/emnlp2019-1.html#ReimersG19

Shibata, T., & Uto, M. (2022). Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. *Proceedings of the 29th International Conference on Computational Linguistics*, 2917–2926.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 163–173.

Williams, S. M., Nix, D., & Fairweather, P. G. (2000). *Using speech recognition technology to enhance literacy instruction for emerging readers*.

Supplementary Materials

Tables and Graphs

Download: https://nsbejournal.scholasticahq.com/article/92286-an-analysis-of-large-language-models-for-african-american-english-speaking-children-s-oral-language-assessment/attachment/192841.docx