Towards Effective Speech-based AI in the Classroom: The Case of AAE-Speaking Children

Alexander Johnson[1], **Julie Washington**[2], **Robin Morris**[3], **Mari Ostendorf**[4], **Alison Bailey**[5], **Abeer Alwan**[1]
[1]University of California, Los Angeles, [2]University of California, Irvine,
[3]Georgia State University, [4]University of Washington
{ajohnson49@ucla.edu, alwan@ee.edu}

Abstract

This paper presents empirically driven recommendations for advancing the use of spoken language systems in children's language education. We propose shifts in the current paradigm of machine learning research to better fit the growing needs of educators as well as capture concerns expressed by those in the field of AI.

1 Introduction

AI has revolutionized practices in finance, defense, and entertainment. However, the education sector has significantly lagged behind in adopting machine learning-based technologies in teaching. One reason for educator's hesitancy to use AI with their students stems from AI researchers' failure to prove the efficacy and fairness of the technology. This paper presents recommendations on implementing AI systems in early education through the lens of employing spoken language technology for literacy education, especially for speakers of African American English (AAE).

2 Spoken Language Systems for Early Literacy Education

One of the most proposed uses of AI in education is Interactive Spoken Language Systems (ISLS) for literacy and language training. ISLS mainly involve both Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) Systems. These systems offer the possibility of a powerful tool for early childhood education, freeing up teachers' time and enabling more frequent classroom formative and diagnostic assessment. For example, these systems could be used to administer language assessments such as the Gray Oral Reading Test, 5th Edition (GORT-5) [1]. In this assessment, a child is asked to read a passage aloud and then scored by their fluency as well as their proficiency in answering reading comprehension questions on the passage. Here, TTS could be used to administer the instructions through an artificially synthesized voice while ASR is used to transcribe the response while detecting any disfluencies or inaccuracies in order to score the student's performance. Natural Language Processing (NLP) could also be applied to further capture discourse level understanding from the ASR outputs. Although the GORT-5 is typically an involved assessment that requires the full attention of a well-trained educator, ISLS could allow untrained caretakers with a smartphone application to administer it at home. This technology would be especially potent for schools in under-served communities, as students of minority communities have been shown to achieve lower literacy rates than their peers [2]. However, several challenges remain in employing these systems for under-served communities and in general.

3 Barriers to Fair and Inclusive Spoken Language Technology

State-of-the-art (SOTA) ISLS systems rely on training data-hungry models with large amounts of speech and text data. SOTA ASR systems typically use 1000 or more hours of speech data [3], and

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

SOTA TTS systems usually require a speaker to record their voice for 25 hours or more in order to provide a robust synthetic voice [4]. While corpora containing annotated data in these amounts exist for adult speech, they are not available for children's speech. Furthermore, the available data often lack a diversity of representation from different genders, dialects, accents, and speech-related disabilities, leaving machines trained on the standard corpora with poor performance for members of these groups. For example, it has been shown that ASR systems trained only on General American English perform less well for speakers of African American English [5]. This is because of dialectal differences in pronunciation, word usage, and grammar that do not appear in the Mainstream American English datasets on which most commercially available ISLS are trained. This disparity in performance must be fixed before these systems can be deployed equitably across schools.

In addition, ASR systems and TTS systems are often evaluated by word error rate and mean opinion score (found through having several listeners rate the audio quality and naturalness of a synthesized voice on a scale of 1-5) respectively. However, it has not been shown that these metrics correlate well to factors in a language assessment. For example, the ASR system's raw score of the percentage of words that a child repeated incorrectly may not be indicative of their actual understanding of a story. Further work is needed to develop training targets for ISLS that match the criterion needed in educational exercises.

4 Recommendations

4.1 Model Scaling

In the past few years, several of the innovations in ISLS such as Wav2Vec2.0-XLSR [6] and GPT-3 [7] have come from creating larger models with intractable numbers of parameters. Training these models has become inaccessible to many researchers and even less so to teachers who would benefit from their use in education settings. Given the lack of diverse sets of children's speech data, it is difficult to train a single large, data-hungry model that can be trained recognize a large range of children's voices without experiencing catastrophic forgetting or overfitting to the training set [8]. Therefore, the field should move towards creating smaller models, either through knowledge distillation or novel architectures that utilize linguistic knowledge, in order to create easily customizable, high-performing systems. Then teachers can rely on their experience to select smaller amounts of crowd-sourced data to train models for their specific use cases.

4.2 Data Diversity

Training data for ISLS systems is often measured in number of hours of speech or number of lines of text. However, these measures may be insufficient to explain the type and number of linguistic features that the machine will see during training. Instead, we should move towards also including metrics that show the diversity of training data such as the number of speakers from distinct linguistic backgrounds and the number of characteristics of speaker traits (eg. the number of markers of African American English or regional dialect, instances of codeswitching or translanguaging, the number of speech-pathology related disfluencies, etc.).

4.3 Community Feedback

Much of recent AI research has revolved around improving systems by objective metrics in order to climb leaderboards. Instead, the field should work to receive direct feedback from stakeholders when optimizing system performance. The gap between commonly used machine-learning metrics and the actual impact of the system on communities must be bridged for ISLS systems to become integral tools in education. Needs Assessments that collect feedback from educators and students will be vital in implementing these systems.

5 Conclusions

The issues and recommendations in this paper serve to highlight missing contributions in AI research that are needed to make widespread SLS systems in language education a reality. We focus on improving accessibility in model training as well as re-examining how we look at data diversity. These recommendations are intended to move AI research closer to realities in teaching practice.

References

[1] J. L. Wiederholt & B. R. Bryant. “Gray Oral Reading Tests—Fifth Edition (GORT-5).” Austin, TX: Pro-Ed. 2012.

[2] U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. “National Assessment of Educational Progress (NAEP), 2020 Long-Term Trend Reading and Mathematics Assessments.” Institute of Education Sciences, National Center for Education Statistics, 2021

[3] Alexei Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations", 2020.

[4] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," in Proc. ICLR 2021

[5] A. Koenecke, A Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, “Racial disparities in automated speech recognition.” Proceedings of the National Academy of Sciences Apr 2020, 117 (14) 7684-7689; DOI: 10.1073/pnas.1915768117

[6] Alexis Conneau et al., “Unsupervised cross-lingual representation learning for speech recognition", 2020.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. “Language models are few-shot learners." In NeurIPS, 2020.

[8] Vinay Venkatesh Ramasesh, Aitor Lewkowycz, Ethan Dyer, “Effect of scale on catastrophic forgetting in neural networks" in Proc. ICLR 2022