Tailor, Test, Train: Novel Usage of Open Educational Content in ScribeAR for Front-end Testing and AI Training

Arman Michael Mehdipour, University of Illinois Urbana-Champaign

Before entering the Department of Computer Science as a graduate student, Arman studied organizational psychology. The incredible complexity of this transition and his previous research interests, including psychometric validity, have given him a deep appreciation for the power of effective engineering education.

Christian Classen, University of Illinois Urbana-Champaign

Christian Classen is a student at UIUC. His primary research interest is machine learning. He is interested in tackling problems in the intersection of AI and major social/ethical issues. By understanding the capabilities and flaws of current ML technology, we can ensure that emerging AI systems are used for the greater good while preventing harm.

Yuxuan Chen, University of Illinois Urbana-Champaign

Yuxuan Chen is an undergraduate Mathematics and Computer Science student at UIUC. His research interests include using HCI and AI technology to improve computer science education.

Colin P. Lualdi, University of Illinois Urbana-Champaign

Colin P. Lualdi is a Physics Ph.D. candidate in the group of Paul Kwiat at UIUC, where he studies quantum information science. As an experimentalist, Colin is particularly interested in using photons to explore fundamental physics and develop quantum technologies. Colin is also interested in exploring new approaches to communication, ranging from developing STEM sign language resources to making real-time captioning tools more versatile. Colin received his A.B. degree in Physics from Princeton University, along with certificates in linguistics and computer science.

Dr. Lawrence Angrave, University of Illinois Urbana-Champaign

Lawrence Angrave is an award-winning computer science Teaching Professor at UIUC. He creates and researches new opportunities for accessible and inclusive equitable education.

Tailor, Test, Train: Novel Usage of Open Educational Content in ScribeAR for Front-end Testing and AI Training

Abstract

ScribeAR is an open-source real-time captioning platform that employs automatic speech recognition (ASR) technology and human-centered design principles to advance equity and accessibility in engineering education [1] [2]. This service employs both online and offline models, operating on a variety of platforms, including augmented-reality headsets, mobile devices, and computers. Offline service is provided by the Web Speech API, and online service can be provided by either Web Speech or Azure [3]. The widespread availability of ScribeAR compounds existing challenges in captioning and front-end testing coverage; this includes designing end-to-end tests, homophone errors, and proper noun or jargon recognition. We present a novel front-end testing methodology that effectively mimics the user experience, with secondary benefits for AI training.

To advance transcription accuracy, a semi-automated front-end testing framework was developed. System-level driver emulation (optionally employed within virtual machine instances) feeds sanitized audio derived from open-access content to the ScribeAR platform. The resulting transcripts are downloaded, and a batch comparison between master texts and platform output is conducted using NLP tools. Variable speed and diverse speaking samples are employed. It was hypothesized that these attributes coincide with both improved AI fine-tuning and real-world platform demands. Students often speed up lecture videos, and educational platforms necessarily employ diverse speakers. Azure performance was comparable to existing literature findings, while Web Speech, though free and available offline, is poorly suited to long-form educational content as implemented. WhisperX fine-tuning was successful and presents a promising, free alternative to existing models [4].

Introduction

In its initial version, ScribeAR, as previously presented at ASEE [1], possessed a robust back-end testing system but no formalized front-end testing procedure. Natural, human-generated speech input was provided for brief periods to ensure the baseline functioning of deployed models on the live service. However, this standard did not formally assess the value of out-of-box recognition systems (especially in realistic scenarios) beyond face validity or facilitate the creation of alternative models that might be better suited to platform demands. Moreover, sole reliance on live human testers results in a bottleneck for testing capacity and input reproduction.

Assessing the performance of a captioning service designed to supplement education requires consideration and resources that enter the domain of human qualia, beyond the scope of simple unit testing. ScribeAR required a testing solution that balanced existing qualitative resources with the quantitative potential of emerging technology. ScribeAR employs Azure and Web Speech recognition models. A recently developed AI implementation, WhisperX, based on OpenAI's Whisper model, was designed specifically to address the difficulties of LLMs in accurately transcribing long form content [4]. A review of extant speech-to-text technologies confirms WhisperX offers significant potential improvements over both Web Speech and Azure. However, the benchmarked WhisperX model was not provided subject specific material [5]. Thus, it is possible the relative performance of WhisperX may be significantly improved with a fine-tuned implementation applied to a specific subject matter— engineering education. Furthermore, these same improvements can be constructed with the collected materials used for the bench-marking of ScribeAR. However, care must be taken to segregate test and train data to avoid over-fitting [6].

Methods

In the interest of reproducibility, a collection of open educational materials was assembled from Project Gutenberg [7], LibriVox [8], MIT OpenCourseWare [9], and the University of Illinois Urbana-Champaign's ClassTranscribe service [10]. These educational resources are in English, but the speakers are of varying races, genders, and backgrounds. Some presenters possess an accent and are not native English speakers. The overwhelming majority of materials are related to engineering education, but two out of the 12 total are recitations of public domain texts in unrelated fields. Every audio segment, over an hour long each, has a corresponding master transcript or caption file in .VTT format. The use of pre-recorded audio enables exact input over multiple trials, and the deployment of open educational content ensures these inputs are widely and ethically accessible. Not just for the purpose of platform testing, but for additional model training or even study reproduction. Following the acquisition of educational materials, pre-processing was conducted. Master transcripts were created if none existed, and manually audited for accuracy. In some instances, master transcripts were further converted to .VTT format to facilitate AI training. Audio files were trimmed and cleaned as needed. In accordance with the license of Project Gutenberg, stripping all organizational references from texts not protected by U.S. copyright law returns texts that are fully in the public domain [11].

After a collection of reference materials was assembled, testing infrastructure was programmed. ScribeAR requires live microphone input, which limits testing potential. Therefore, a system to emulate a microphone at the driver level was deployed. The virtualized microphone input can be broadcast across multiple virtual machine instances simultaneously. Subsequently, a Python script utilizing the spaCy NLP library was programmed to enable the comparison of the text output of ScribeAR instances with master texts. The use of spaCy enables exploratory analysis of word categories during instances of error and the ability to ignore subtle punctuation differences relative to simple difference checkers. 25 trials were conducted per out-of-box service (Azure and Web Speech), wherein at least one hour of audio was fed to a distinct runtime for each trial.

After the finalization of data pre-processing and testing infrastructure, WhisperX fine-tuning was conducted. The 8-bit version of the ADAM optimizer and the large-sized model were selected for fine-tuning on approximately 5.5 hours of audio, which were segmented into 656 chunks of approximately 30 seconds in length [12]. Each chunk was added to either the training set, validation set, or test set with a probability of 70 percent for training and 15 percent for both validation and testing. Subsequently, there were 452 chunks in the training set, 97 in the testing set, and 94 in the validation set. During hyperparameter selection, batch sizes were set to 1 for training and 16 for validation. A prompt was provided to the model only 50 percent of the time. While WhisperX does not require timestamp training, timestamps were provided 50 percent of the time. Gradient accumulation was utilized, and model parameters were updated every 64 steps. To determine the learning rate, 400 warm up steps were conducted. Then, the model was fine-tuned for 500 steps. There was a maximum gradient norm of 1.0.

Results

Testing on Azure and Web Speech is shown in Fig. 1. For Azure, results were comparable to recent literature review results [5]. This review does not directly mention the Web Speech API, which seems poorly suited to previous applications. According to analysis in spaCy, the most commonly misinterpreted part of speech token (when accounting for omitted text in Web Speech) was the Proper Noun. Web Speech as currently deployed appears to be unsuited to long-form, complex, single-speaker content such as university mathematics lectures. While sentence- or paragraph- level performance appears somewhat similar to Azure over very short periods, drifting and hallucinations occurred across every trial over extended periods of time. Web Speech appeared to be overwhelmed by lectures. Subsequent sentences could be accurately transcribed, only to disappear in front of the current sentence. Web Speech would frequently revise previous sentences as though they were being corrected by the ongoing lecture, respond sporadically, or simply stop responding altogether. This may be a function of imposed service limitations on continuous speech, which warrants further investigation.

Model	Accuracy	Most Missed Token
Web Speech	12.23	Proper Noun
Azure	88.21	Proper Noun

Figure 1: Accuracy Rates Across Out-Of-Box Services

It was observed that temporarily pausing and resuming input could lead to a resumed transcript. However, in the interest of accuracy, transcription during all recorded trials was allowed to occur without human interference. Out of an abundance of caution and given the unusually poor performance over multiple trials, results were briefly re-checked and subsequently replicated (n=3) using Chrome's official Web Speech demo applet and lecture content [13].

A comparison of the word error rates (weighted and unweighted) between Whisper's Large size model and the fine-tuned WhisperX version yields promising results. Shown in Fig. 2, Whisper's Large model yielded an unweighted Word Error Rate (WER) of 0.230 and a weighted WER of 0.149. The current study's fine-tuned WhisperX model yielded an unweighted WER of 0.183 and a weighted WER of 0.155. This suggests additional training data and longer training time could significantly improve the model, ultimately yielding results more comparable to ideal performance [4]. It is likely that an over-representation of both accented speech and symbol-heavy mathematics lectures in conjunction with a reduced sample pool relative to the testing pool for existing services resulted in an increased WER. The inconsistent representation of verbally described mathematical equations naturally results in inconsistent normalization. These factors could be better mitigated with additional training data and longer training time.

Model	WER (Unweighted)	WER (Weighted)	Most Missed Token
Whisper	0.230	0.149	Proper Noun
WhisperX	0.183	0.155	Proper Noun

Figure 2: Whisper and WhisperX Word Error Rate

Conclusion

Significant evidence, in the form of the current study's findings and existing literature, support the use of Azure as an automatic transcription tool to benefit engineering education. However, WhisperX, if effectively implemented, may hold more promise. A fine-tuned, custom WhisperX implementation with performance matching the expectations of existing literature could potentially outperform Azure while providing a free service [5].

Additionally, it is highly recommended to modify Web Speech usage in ScribeAR. It is possible that periodically refreshing the Web Speech service while maintaining continuity within the speech buffer may mitigate some of the discussed performance deficits. However, Web Speech as currently implemented is unsuited to continuous, long-form content in public environments. Ultimately, these results reinforce the importance of comprehensive testing coverage. During superficial usage, such as non-continuous speech within the span of a few minutes, no operational difficulties with Web Speech were observed. Collectively, it is likely the variance between the current study's results and existing literature has been contributed to by the usage of continuous long form content, differences in sampling procedures, the incorporation of accented speech, and the representation of mathematics and abstract computer science concepts. Future research will incorporate Google Speech-to-Text, allowing us to better compare the performance and suitability of different speech recognition systems [14].

Acknowledgements

We appreciate the Azure cloud credits provided by Microsoft Corporation for speech-to-text transcription. We also wish to acknowledge the work and dedication of all past and present ScribeAR team members.

References

[1] L. Angrave, C. P. Lualdi, M. Jawad, and T. Javid, "ScribeAR: A New Take on Augmented-Reality Captioning for Inclusive Education Access", *2021 ASEE Illinois-Indiana Regional Conference*, 2021. DOI: 10.18260/1-2--38276. [Online]. Available: https://peer.asee.org/38276.

[2] Y. Wang, C. P. Lualdi, L. Angrave, and G. N. Purushotam, "Using Deep Learning and Augmented Reality to Improve Accessibility: Inclusive Conversations Using Diarization, Captions, and Visualization", *2023 ASEE Annual Conference & Exposition*, 2023. DOI: 10.18260/1-2--44572. [Online]. Available: https://peer.asee.org/44572.

[3] MDN Web Docs, *Web Speech API*, 2023. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API.

[4] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio", 2023. arXiv: 2303.00747.

[5] A. Ferraro, A. Galli, V. La Gatta, and M. Postiglione, "Benchmarking open source and paid services for speech to text: An analysis of quality and input variety", *Frontiers in Big Data*, vol. 6, 2023, ISSN: 2624-909X. DOI: 10.3389/fdata.2023.1210559. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2023.1210559.

[6] P. Cohen and D. Jensen, "Overfitting Explained", *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, Apr. 2000.

[7] M. Hart, *Project Gutenberg*, 1971. [Online]. Available: https://www.gutenberg.org.

[8] LibriVox, *LibriVox Free Audiobook Collection*, 2005. [Online]. Available: https://librivox.org.

[9] Massachusetts Institute of Technology, *MIT OpenCourseWare*, 2001. [Online]. Available: https://ocw.mit.edu.

[10] C. Mahipal, L. Angrave, Y. Xie, B. Chatterjee, H. Wang, and Z. Qian, "'What did I just miss?!' Presenting ClassTranscribe, an Automated Live-captioning and Text-searchable Lecture Video System, and Related Pedagogical Best Practices", *2019 ASEE Annual Conference & Exposition*, 2019. DOI: 10.18260/1-2--31926. [Online]. Available: https://peer.asee.org/31926.

[11] Project Gutenberg. [Online]. Available: https://www.gutenberg.org/policy/license.html.

[12] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[13] Google Chrome, *Web Speech API Demonstration*. [Online]. Available: https://www.google.com/intl/en/chrome/demos/speech.html.

[14] Google, *Google Cloud Speech-to-Text*. [Online]. Available: https://cloud.google.com/speech-to-text/.