Exploring practical Metrics to support Automatic Speech Recognition Evaluations

E.A. DRAFFAN[a1][1], Mike WALD[a], Chaohai DING[a] and Yunjia LI[b]
[a] *ECS, University of Southampton, UK*
[b] *Yunjia Li, Habitat Learn, UK*
ORCiD ID: E.A. Draffan https://orcid.org/0000-0003-1590-7556

Abstract. Recent studies into the evaluation of automatic speech recognition for its quality of output in the form of text have shown that using word error rate to see how many mistakes exist in English does not necessarily help the developer of automatic transcriptions or captions. Confidence levels as to the type of errors being made remain low because mistranslations from speech to text are not always captured with a note that details the reason for the error. There have been situations in higher education where students requiring captions and transcriptions have found that some academic lecture results are littered with word errors which means that comprehension levels drop and those with cognitive, physical and sensory disabilities are particularly affected. Despite the incredible improvements in general understanding of conversational automatic speech recognition, academic situations tend to include numerous domain specific terms and the lecturers may be non-native speakers, coping with recording technology in noisy situations. This paper aims to discuss the way additional metrics are used to capture issues and feedback into the machine learning process to enable enhanced quality of output and more inclusive practices for those using virtual conferencing systems. The process goes beyond what is expressed and examines paralinguistic aspects such as timing, intonation, voice quality and speech understanding.

Keywords. automatic speech recognition, error correction, word error rate, captions, transcriptions, disability,

1. Introduction

During the last few years Higher Educational Institutions (HEIs) across the world increased their use of elearning and video conferencing, in part due to the COVID-19 pandemic [1]. Speaker independent automatic speech recognition (ASR) systems were often used in lecture recording situations for the provision of captions and transcriptions in order to support those students with cognitive, physical and sensory disabilities such as dyslexia, visual and hearing impairments as well as dexterity issues making note taking difficult. Research into presentations in English have also shown that many individuals benefit from these alternative formats such as when the language spoken is not a person’s first language and those who may find text-based content easier to work with when concentrating on audio or video output is impossible [2]. However, the audio to text output needs to correctly represent what has been said so users gain maximum benefit and in some countries, such as the United States of America, 99% accuracy rates are necessary for captions as a legal requirement [3]. Ideally the output in an academic situation, for example a lecture or seminar, also needs to be transcribed in a timely fashion. This is becoming more achievable with ASR rates of accuracy reaching between

[1] Corresponding Author: E.A. Draffan, ead@ecs.soton.ac.uk

90-95% in ideal classroom settings with native English speakers [4]. But where optimal environments are not possible it is important to discover which aspects are causing an increase in word error rates, whether they are substitutions, insertions or deletions.

In this paper we explore the practical aspects of a series of metrics that are particularly relevant to academic lecture situations with a series of checks that depend on confidence scores as well as values related to comments about the recording situation.

2. Background

Over the years several types of metrics have been used alongside Word Error Rates (WER) such as the Levenshtein distance, Match Error Rates (MER), Word Recognition Rate (WRR), Number of phrase-level insertions, deletions, and mismatches and Concept Error Rates (CER) in order to evaluate the accuracy of ASR[2]. However, none seem to have been used to the same extent as WER despite criticism from many researchers such as those mentioned by Kuhn, Kersken and Zimmermann who add that "ASR output should be validated for real-world use-cases" [5]. So, although WER may provide insights into the accuracy levels of automatic captions and transcriptions [6] there remain other factors that can be explored to provide additional feedback for ASR training models.

Ulasik et al in their work on a Corpus for Evaluating the quality of Automatic Speech Recognition (CEASR) highlighted issues that arise from "disfluencies, speaker and non-speaker noise as well as non-native speech" [7] which may well be representative of the type of problems experienced in an academic setting, but automatically calculated error rates do not necessarily provide reasons why these errors occur [8]. That is why it is important to ensure more information about the speech recordings can be provided to enhance training data.

3. Methodology

As a pilot project 30-minute recordings were taken from four lectures on anatomy, accounting, statistics and counselling, each given by 4 college tutors, two Canadian females and two males, one Chinese and one Canadian in three different lecture theatre settings. The language was English and the automated captions were developed using YouTube[3] with the transcription being copied into a text file. This was compared to the 'ground truth' provided by Otter.ai[4] in the form of speaker notes in order to find the Word Error Rate. Otter.ai was used as it proved remarkably accurate despite the complexity of the content and could easily be manually corrected in situ with the recording running. It also allowed the researchers to clearly see the type of errors that were developing. There followed a review using additional metrics under the main headings of Speaker, Environment, Content and Technology (Table 1). These were listed in an Excel spreadsheet against a 1-3 series where 1 = not confident, 2 = neutral and 3 = confident. The idea of using confidence levels was based on the need to understand how three evaluators, who were experienced in working with disabled

[2] https://symbl.ai/blog/key-metrics-and-data-for-evaluating-speech-recognition-software/

[3] https://www.youtube.com/

[4] https://otter.ai/home

students and creating alternative formats, marked the various metrics as to why each one affected certain types of word error.

Table 1. Additional practical metrics to support the evaluation of ASR outcomes

Speaker	Environment	Content	Technology	Technology
Speech Pronunciation Clarity Speed Loudness Pitch Intonation Inflection Accent Age Gender Use of Technology Too far away / Near the microphone	**Noise** Ambient noise/continuous Reverberation Sudden noise Online/Offline User device Room system Conversation Presentation Single speaker Overlapping speakers Multi-speakers	Complexity Unusual names, locations, and other proper nouns Technical or industry-specific terms Out of Vocabulary / not in the dictionary Homonyms	**Hardware** Smart phone Tablet Laptop Desktop **Microphone** Array Headset Built-in Hand held **Camera** Specialist /Smart Computer Mobile Cont..	**Recording** Direct audio recording Synthetic speech recording Noise-network distorted speech **Connectivity** Live / Real-Time Recorded

The evaluators, whilst listening to the recording, scored their feelings of confidence against the various metrics, for example if they were confident that the speech was clear or sufficiently loud to be heard by students the score would be 3. A secondary value was added in the form of a comment to clarify the scores given, such as the reason pronunciation may be scored as neutral because the evaluator felt unsure about whether a few sounds were inaudible or mistranslated because English was not the speaker’s first language.

4. Findings and Discussion

A review of the findings has yet to be fully undertaken or checked against the developer’s decisions as to how many adaptations to the ASR are possible based on the comments. To date, the team have begun to label elements perceived as being the cause of bias based on the matrix and to categorize the training data. This has the potential to highlight the most frequent categories that cause a lack of confidence in the system. Under the speech category pronunciation, clarity and speed received neutral confidence levels for the accounting and counselling lectures and levels of confidence for inflection were neutral across all lectures. This could have been due to the way the evaluators understood this term, namely thinking about the sound of the voice as opposed to questionable changes to word forms[5]. There were differences related to both meanings, so this needs to be clarified in the next iteration of the matrix (Table 2).

One reason that became clear when reviewing the comments was that pronunciation with the aforementioned inflection and accent had an impact. However, developers may be able to preempt likely articulation errors that are typical for those speaking English as a foreign language. There may be specific errors related to some words for instance those that have “th”, “v” and “rl” sounds that do not appear in some Chinese dialects.

[5] https://www.oxfordreference.com/search?q=inflection&searchBtn=Search&isQuickSearch=true

Clarity of consonant cluster pronunciation, may also be affected by age which in turn has an impact on WER. The evaluation data highlighted tentative guesses at the age of the presenters and all four lectures received neutral scores for this category although one evaluator was more confident that they could judge the age of two presenters as being between 40-50 years old. Research into the effect of age on the voice tends to have been linked to an older population as described by Kim et al [9]. The authors discuss the concept of a voice conversion framework coupled with linguistic information that may help to reduce issues of bias where the voice files used to generate data sets are mainly from younger adults. It is felt that this framework might improve outcomes where there are multilingual speakers as well as older lecturers.

Table 2 Sample 10 categories with scores based on three evaluators using the 1-3 scale for confidence levels and percentages where all categories were completed.

	Total Scores	**Total Scores**	**Total Scores**			
Category	**1 Not Confident**	**2 Neutral**	**3 Confident**	**%Not Confident**	**% Neutral**	**% Confident**
Pronunciation		6	27		25%	75%
Clarity		4	30		17%	83%
Speed		2	33		8%	92%
Loudness			36			100%
Pitch			36			100%
Intonation			36			100%
Inflection	1	8	21	8%	33%	59%
accent		4	30		17%	83%
age		20	6		83%	17%
gender						100%

Not surprisingly sudden noises and overlapping speakers or multiple speakers affected transcription and caption accuracy and when reviewing the Word Error Rates (WER) these correlated with the 100% confidence levels that these problems affected outcomes. The WER scores for the four lectures were Anatomy WER 6.5%, Accounting WER 14.4%, Statistics WER 9.0% and Counselling WER 18.2%. The higher WERs were matched by lower confidence levels in the completed 10 categories for instance the Accounting lecture had a confidence score of 69 and Counselling 60 compared to 78 and 84 for Anatomy and Statistics consecutively.

This manual process of checking all 48 categories in the matrix would eventually need to be set against a series of automated processes although in some cases, such as the way technology is used, this may not be possible. To date a limitation of the technology checks included the fact that data was collected from live and recorded audio or video output remotely, rather than during face to face lectures. Judgements about use of the microphone came from the sound of the voice and virtual views. For example, in one lecture, as the lecturer moved across the room the voice faded away or in another case the lecturer was wearing a headset and this could be seen on the video so the voice was judged to be very clear with full confidence.

The complexity of terms used in all the lectures was confidently noted with medical terms appearing in the anatomy lecture, but it was the business names and statistical terms that seemed to cause more accuracy problems, although those evaluators, who admitted they knew the terms, felt they would be transcribed accurately when clearly spoken as most were known in English at an academic level. The longer complex words were transcribed accurately in both the YouTube captions and the Otter.ai transcription. The Anatomy lecture having particularly low WER scores of around 6.5% despite the lengthy medical terminology.

The limited evaluation data, collected at the time of writing, highlighted potential matrix changes that would reduce the time taken to complete the checks. These changes would mainly happen in the technology hardware section which was relatively incomplete, with the type of computer used and recording techniques remaining blank. As there were many categories and some were not necessary in certain situations or evaluators did not feel they could fill in the scores, it may be necessary to make public more information surrounding the lecture settings and the way academics use systems.

5. Conclusion

The use of a wider range of practical metrics evaluated by a series of scores and value-added comments has the potential to improve rates of accuracy and tailor ASR for specific requirements. Further information gathered from the data collected will be presented at the conference. In particular it is hoped that the issue of selection bias in ASR [10], that in this case has meant that errors have occurred due to pronunciation differences affected by age, accents and English as a foreign language, can be addressed. ASR providers need to improve accuracy levels by using differently biased input data that is customized, instead of using one single accuracy percentage to denote the performance of the ASR services. External evaluators should be aware of these issues and suggest the need for more inclusive training data to enable corrections to automatically occur in a proactive manner. It is also important to keep raising awareness about best practices for recording settings and to improve the way technology is used by presenters to further enhance ASR caption and transcription outcomes.

Conflicts of Interest

Author Yunjia Li is employed by Habitat Learn UK. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The team from the University of Southampton would like to thank all those experts who took part in this review and the students, lecturers and colleagues who also helped by offering their opinions when using a newly developed e-learning platform with cloud-based video conferencing, summary and linked resources as part of UKRI Innovation UK funded projects (Technology Strategy Board Ref 10024466, 10013521, 103341).

References

[1] García-Morales, V.J., Garrido-Moreno, A., Martín-Rojas, R. The transformation of higher education after the COVID disruption: Emerging challenges in an online learning scenario, Frontiers in Psychology 12 (2021)

[2] Wilson, L. Making Online Conferencing Accessible: Platforms, Captions and Transcripts. The Journal of Inclusive Practice in Further and Higher Education. 12.1, 106 Available at: https://nadp-uk.org/wp-content/uploads/2020/12/JIPFHE.ISSUE-12.1-Winter-2020.docx (2020) Accessed 27 April 2023

[3] Klein, R., U.S. Laws for Video Accessibility: ADA, Section 508, CVAA, and FCC Mandates 3PlayMedia website https://www.3playmedia.com/blog/us-laws-video-accessibility/ (2021) Accessed 05 March, 2023

[4] Millett, P., Accuracy of Speech-to-Text Captioning for Students Who are Deaf or Hard of Hearing. Journal of Educational, Pediatric & (Re) Habilitative Audiology, 25. (2021)

[5] Kuhn, K., Kersken, V. and Zimmermann, G., Accuracy of AI-generated Captions With Collaborative Manual Corrections in Real-Time. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-7). (2023)

[6] Favre, B., Cheung, K., Kazemian, S., Lee, A., Liu, Y., Munteanu, C., Nenkova, A., Ochei, D., Penn, G., Tratz, S. and Voss, C.R., Automatic human utility evaluation of ASR systems: Does WER really predict performance? In INTERSPEECH (pp. 3463-3467). (2013)

[7] Ulasik, M.A., Hürlimann, M., Germann, F., Gedik, E., Benites, F. and Cieliebak, M., CEASR: a corpus for evaluating automatic speech recognition. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 6477-6485). (2020)

[8] Wald, M., Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. Interactive Technology and Smart Education. (2006)

[9] Kim, J.W., Yoon, H. and Jung, H.Y., Linguistic-coupled age-to-age voice translation to improve speech recognition performance in real environments. IEEE Access, 9, pp.136476-136486.(2021)

[10] Feng, S., Kudina, O., Halpern, B.M. and Scharenborg, O., Quantifying bias in automatic speech recognition. arXiv preprint arXiv:2103.15122. (2021)