Connecting through Educational Technology
Proceedings of the European Distance and E-Learning Network 2019 Annual Conference
Bruges, 16-19 June, 2019
ISBN 978-615-5511-27-1
ISSN: 2707-2819
doi: https://doi.org/10.38069/edenconf-2019-ac-0044

AUTOMATIC TRANSCRIPTION SOFTWARE: GOOD ENOUGH FOR ACCESSIBILITY? A CASE STUDY FROM BUILT ENVIRONMENT EDUCATION

Tharindu R. Liyanagunawardena, University College of Estate Management, United Kingdom

Abstract

The increasing use of multimedia in learning resources in higher education poses a challenge for learners with hearing disabilities, unless these are accompanied by transcripts or captions. This paper reports on a small study where six Automatic Transcription Software (ATS) were analysed for their accuracy. Although economical and timesaving, at present, it seems an automatically generated transcript is not yet accurate enough to be an accessibility aid for the subjects relating to built environment sector.

Accessibility

The use of multimedia has allowed the creation of rich learning experiences for students, especially for distance learners. This poses a huge challenge in making these resources accessible. Oxford Living Dictionary (n.d.) defines accessibility as "the quality of being able to be reached or entered". More specifically, the term "Web accessibility" is used to denote that "websites, tools, and technologies are designed and developed so that people with disabilities can use them". That is, people with disabilities can "perceive, understand, navigate, and interact with the Web and contribute to the Web" (W3C, 2018a).

Web Content Accessibility Guidelines (WCAG) is a set of guidelines produced with the goal of creating one international shared standard for web content accessibility (W3C, 2018b). With the recent European Parliament directive on accessibility (Europa, 2016) there will be an increasing need for transcripts/captions to meet the legal requirement of time-based media accessibility.

Accuracy, Cost and Automatic Speech Recognition (ASR)

Transcription services that employ human transcribers can be expensive. In December 2017, the author's institution paid an average rate of £1.30 per minute for instructional videos to be professionally transcribed, amounting to £78 per hour (subjected to a minimum fee and additional cost for more than one speaker). Manchester University guidelines for research students indicate that an hour of recording is likely to take four to seven hours to transcribe (Burke, Jenkins, & Higham, 2010). University of California Berkeley removed over 20,000 videos and podcasts from public access in March 2017, in response to an order by the

Department of Justice to make these contents accessible to people with disabilities because it was deemed "extremely expensive" (Straumsheim, 2017). Prohibitive cost and extensive time requirement are major disadvantages of manual transcription.

Automatic Speech Recognition (ASR) technologies have improved rapidly with both IBM and Microsoft reaching 5.5% (Fogel, 2017) and 5.1% (Lant, 2017) Word Error Rate (WER) respectively, almost on a par with a professional human transcriber. Today, off-the-shelf Automatic Transcription Software (ATS), for example Nuance Dragon Professional Individual v15, promises 99% accuracy once the software is trained for the speaker's voice (Dragon Speech Recognition, 2016). However, automatically generated captions/transcripts, can sometimes fail to convey the meaning accurately; it can be humorous or intelligible, but for people with hearing-impairments this can result in inaccessible content. Bokhove and Downey (2018) argue that given the quality of ATS is sufficient, these tools could provide a useful first draft for transcription creation.

In this context, this paper reports an experiment conducted at University College of Estate Management (UCEM) to assess the suitability of using ATS in the built environment sector subject disciplines.

University College of Estate Management

UCEM, previously known as the College of Estate Management, is a leading supported online learning provider for the built environment. Celebrating its centenary in 2019, UCEM was originally a postal distance education provider until 2014 when it offered fully online modules to students. Since then UCEM has been offering almost all courses fully online. The core purpose of UCEM, as described in its vision statement, is to "provide truly accessible, relevant and cost-effective education". UCEM caters for students in various circumstances; due to their circumstances some of them would not have had an opportunity to gain professional recognition in built environment in a more traditional university setting. The institution also has a large international student population.

When UCEM achieved the University College status in 2015, the rebranding exercise was taken as an opportunity to improve accessibility of UCEM learning material templates (Liyanagunawardena & Hussain, 2017). UCEM provides transcripts/captions for all pre-recorded audio and video learning materials. Captions for webinars were only provided where this was required for accessibility (Liyanagunawardena, Forster, & Nadan, 2017). However, it is shown that all students benefit from closed captions (Linder, 2016). Therefore, UCEM is keen to explore capability of ATS to improve webinar recording accessibility to all students. UCEM webinars are generally an hour long and conducted using Blackboard Collaborate. For UCEM students, studying at a distance, webinars are important for real-time interaction with tutors and peers. Due to the volume of webinars conducted it is not feasible to source manual transcription – both due to time and cost implications.

ISBN 978-615-5511-27-1

The Study

With this backdrop, this study investigated whether off-the-shelf automatic transcription software (ATS) is good enough to be used as an accessibility tool in the built environment sector?

Many commercial ATS are available that accept a variety of file formats and offer additional features such as human involvement as add-on services. In identifying a small set of software to be trialled, first the Association for Learning Technology community mailing list was consulted for recommendations. The cost and the availability of a free trial were also considered and the following software were selected: Descript (www.descript.com), IBM Watson Speech to Text (Watson) (www.ibm.com/watson/services/speech-to-text/), Sonix (sonix.ai), Synote (www.synote.com), Trint (trint.com) and Zoom (zoom.us). Synote, Zoom and IBM Watson were selected for the trial as UCEM was already using one or more of their services.

The research obtained ethical approval from the UCEM Research Ethics Committee prior to the data collection. A built environment specific text, containing 1000 words, was created for the experiment. In sourcing this text, expert tutors in were consulted and a section of text (200-500 words) containing subject specific terminology was requested. Four submissions were selected, covering the subject areas of property management, construction management, property and contract law, and building pathology were selected. These submissions were compiled to create the experiment text.

A purposely selected sample of 14 participants, all UCEM employees, with a range of native and non-native English accents, were invited to participate. This small sample was selected to maximise the project's benefits, within the available time, by representing the diversity of UCEM staff and students in webinars. A course, "Transcription testing" was set up on the institutional VLE, Moodle 3.4. This course contained an information sheet, consent form, instructions, the experiment text to be read and a Blackboard Collaborate webinar to create the voice recording. Out of the 14 invited participants, ten who expressed an interest in the study were enrolled on the course. After receiving consent, they were directed to the webinar setup in the course to create a recording of their voice. Seven participants created recordings; four had to be re-recorded due to recording errors. All recordings were created between 17th July and 30th August 2018. Participant demographics is provided in Table 1.

The recordings were downloaded in MP4 format. Whenever MP4 files were supported these were used, otherwise the recordings were converted into a compatible format using Camtasia Studio 8 software.

All transcriptions, except Zoom and Synote were created using each provider's free web service. With Zoom, a trial Zoom Webinar license provided to UCEM was used. The existing Synote workflow at UCEM was followed with Synote. All automatic transcriptions were generated during the period July to October 2018. The transcripts were then extracted into

ISBN 978-615-5511-27-1

Microsoft Word 2016 documents and any additional information such as timings for captions were removed manually. Altogether there were 42 transcripts containing 42,000 words to be analysed

Table 3: Participant demographics

Participant	Gender	Native English Speaker	English Accent (as identified by participant)	Length of Recording (min)
1	Male	Yes	Generic British	6:02
2	Female	No	South American	7:49
3	Male	Yes	Generic Scottish	6:32
4	Male	Yes	Generic British	6:36
5	Female	No	South Asian	6:59
6	Male	No	Greek	5:57
7	Male	No	African	8:21

Analysis

The voice recordings were first compared with the experiment text and variations were noted. For example, instead of "bond yields by contrast" a participant read "both yields by contrast". Each automatic transcription was then compared against the transcript of the voice recording (what the participant read) using Microsoft Word 2016 "Compare" function and a new document was generated where comparison was shown as track changes. These documents were then manually checked. Contractions, punctuations and capitalisation differences were ignored. Transcription of numbers and dates either using word or digits were accepted. Spelling mistakes, tense or singular/plural mismatches were considered as errors. However, where a word such as "webcam" was being transcribed as "web cam" this was ignored (not considered an error).

Once this pre-processing was completed, a manual colour coding was applied to identify substitutions, deletions and insertions. To verify consistency, another member of the Learning Technology Research team was asked to perform the coding on a sample of transcripts and these were compared. Manual comparison was time consuming; however, this was chosen over calculating similarity with a software as used in Bokhove and Downey (2018) due to the high level of differences shown by the Microsoft Word "Compare" function when in fact they were presentation differences. In this study only the meaning of the text was important and presentation differences were not relevant.

Transcript accuracy was checked using the measure Word Error Rate (WER), which was calculated using the formula: WER = (Substitution + Deletion + Insertions) / N; where N is the total number of words in the reference transcript (Apone, Botkin, Brooks, & Goldberg, 2011).

WER calculation for the first few ATS showed that Participant 1's recording was an outlier. In fact, Participant 1 had recorded the voice using laptop microphone while all other participants had used headset microphones mimicking the UCEM setup for webinars. This difference was

ISBN 978-615-5511-27-1

evident in the quality of the recording and thus Participant 1's recording was eliminated from the experiment to manage the like for like comparison.

Subject matter experts were contacted to get their opinion of the automated transcript's quality. Only the transcripts created by the best performing ATS on WER measure was presented to them. Separate documents were created for each subject area consisting only of the text and transcriptions relevant to the subject area. At the top of the document was the original text followed by each of the six (Participants 2-7) numbered transcriptions. Experts' opinion whether each was "good enough" for a student to understand the meaning of the original text was recorded.

Results

Table 2 shows the calculated WER for each software. Some recordings created by non-native speakers performed better than native English speakers' recordings on the WER. For example, Participant 5 (South Asian accent) recorded lower WER than Participant 3 and Participant 4 in all except one. Furthermore, Participant 7 (African accent) recorded lower WER than Participant 3 in three software: Sonix, Descript and Zoom.

Some ATS seemed to have worked better than others in the experiment. For example, Synote has a low WER for Participant 4 (Generic British English accent) but relatively higher WER for all other participants while IBM Watson recorded the highest WER in the experiment for all except one participant. Descript have performed consistently for all participants in the experiment. Table 3 provides details of the total errors and average errors per each software considered and calculated average accuracy. Descripts has the highest average accuracy in this experiment followed by Trint, Sonix, Zoom, Synote and Watson.

Table 2: Word Error Rate for automatic transcription services

Participant	Trint	Sonix	Descript	Watson	Zoom	Synote
2	0.277	0.275	0.181	0.490	0.292	0.471
3	0.168	0.301	0.176	0.412	0.274	0.277
4	0.120	0.140	0.110	0.434	0.245	0.139
5	0.115	0.108	0.101	0.428	0.211	0.359
6	0.392	0.383	0.254	0.555	0.293	0.689
7	0.226	0.147	0.159	0.560	0.263	0.473

Table 3: Average errors per each software

Software	Trint	Sonix	Descript	Watson	Zoom	Synote
Total errors	1098	1354	981	2879	1578	2408
Average errors per transcript (1000 words)	216.33	225.67	163.50	479.83	263.00	401.33
Average accuracy (%)	78.37	77.43	83.65	52.02	73.70	59.87

Expert opinion of the automatic transcripts is provided in Table 4.

ISBN 978-615-5511-27-1

Table 4: Expert opinion of Descript transcripts

Subject area/ Participant	Property Management	Construction Management		Building Pathology		Property and Contract Law	
	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7
2	x	x	x	x	x	x	x
3	x	x	x	x	x	x	x
4	x	Good enough	Good enough	x	Good enough	Good enough	x
5	x	Good enough	x	Almost good enough	x	x	x
6	x	x	x	x	x	x	x
7	x	x	x	x	x	x	x

Despite receiving the lowest WER for Descript software, Participant 5's transcript was "not good enough" as an accessibility aid in its original form except for the Construction Management section and the Building Pathology section where it received an "almost good enough" rating. On the other hand, Participant 4's transcripts received "good enough" rating from four experts. It was also interesting to see that in some instances two experts in the same discipline having conflicting views on acceptability of a transcript. None of the transcripts received a "good enough" rating in the subject discipline Property Management. The selected section was on property and bonds. However, the word "bonds" was transcribed as "bumps", "bones", "blondes", "buttons" deeming them not good enough from the outset. ATS also struggled with a range of other technical terms in the valuation section such as inflation, liquidity, yield, hedge and redemption.

Discussion

Accuracy of transcriptions is an important factor to be considered especially when they are used as accessibility aids. Though WER is used as a common measure of transcription accuracy, there are issues with measuring transcription accuracy this way (Apone et al.,2011). Even though WER considers all words to be equally important, in a technical discipline to make sense of a paragraph an error in an article such as "the" or "a" may be insignificant compared to an error in a key technical term. To address this issue, instead of relying only on the WER, the study was designed to seek the opinion from subject matter experts so that the findings would be more robust. This in fact was reflected in the results of the study, where Participant 4's transcription with slightly higher WER than Participant 5's was judged to be better by the experts.

Providing accessible material is a legal requirement and even when automatic captioning is provided, if they are not "good enough" an institution could be found to be in breach of law. For example, the lawsuit by National Association for the Deaf against Massachusetts Institute of Technology (MIT) and Harvard University stated that "Much of Harvard's online content is either not captioned or is inaccurately or unintelligibly captioned, making it inaccessible for

ISBN 978-615-5511-27-1

individuals who are deaf or hard of hearing" (Lewin, 2015). As this study has shown, some automatic transcripts may not be "good enough" for the students who are hearing-impaired as technical terms have been transcribed incorrectly. Therefore, it is important that the generated automatic transcripts are checked for meaning before being made available to students.

The quality of the recording also plays a role in how well they can be transcribed automatically. In this study, despite Participant 1 being a native English speaker, all ATS struggled to transcribe the recording. The only difference between this recording and other recordings was the use of laptop's built-in microphone as opposed to a headset microphone. This shows the need to use suitable devices/tools to create good quality recordings when ATS is to be employed.

Recording of Participant 5 with a South Asian accent was more accurately transcribed (in terms of WER) than native English speakers. Furthermore, two of the experts rated passages from Participant 5's transcript as "good enough" while none of the passages from the native English speaker Participant 3's transcript was rated "good enough". This may be an indication that the ATS has improved over the years to consider accents other than standard British and standard American accents.

Experiments have shown that even when letters are transposed ("jumbled word effect") due to the way the human brain works, they are still able to recognise the word but at a slower pace (Davis, n.d.) and that the context plays an important role in understanding words. Many transcripts with lower WER were rated "not good enough" by the subject experts. But if these transcripts were presented to hearing-impaired students who have developed various skills in negotiating similar situations – for example, considering the transcript with lip reading or simply contextualising to correct errors – whether these transcripts would have been considered "good enough" is a question worth exploring. Due to the limitation of not having access to hearing-impaired learners as quality assessors this project was unable to explore this interesting avenue.

Limitations

The small sample of transcription software used for this research was selected based on the requirements of the institution. The research was conducted using freely available versions of selected transcription software and if the paid-for version was better than that of the trial version on the web, this would not be reflected in the results. Due to the small sample size, results of this experiment cannot be generalised. Participants were aware of the research and could have consciously read the text for the experiment, rather than the way they would normally communicate. Nevertheless, this study it is a useful reminder that WER on its own cannot be considered as a quality measure for transcription in technical subjects. Using subject experts to assess the quality of the transcripts rather than having a pool of hearing-impaired students was another limitation; as the later would have better reflected the quality of the transcript as an accessibility aid.

ISBN 978-615-5511-27-1

Conclusion

This study analysed six types of off-the-shelf automatic transcription software with a variety of native and non-native speakers' voice recordings to explore whether the transcripts created were "good enough" as accessibility aids in the built environment discipline. The quality of the recording affects the automatic transcription and it is important to take great care in selecting suitable recording devices. In a technical discipline, where key technical terms are of great importance to convey the meaning of a text, Word Error Rate (WER) on its own may not be a sufficient predictor of the quality of a transcript. Automatic transcriptions created using a software with a lower WER could be a great time saving first draft, that could then be checked for meaning and made available to students as an accessibility aid. At present, it seems off-the-shelf automatic transcription software does not produce a high enough level of accuracy for the creation of accessibility aids for the built environment sector even though they are economical and time saving.

References

Apone, T., Botkin, B., Brooks, M., & Goldberg, L. (2011). *Caption Accuracy Metrics Project: Research into Automated Error Ranking of Real-time Captions in Live Television News Programs.*

Bokhove, C., & Downey, C. (2018). Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data. *Methodological innovations, 11*(2), 1-14. doi: 10.1177/2059799118790743

Burke, H., Jenkins, L., & Higham, V. (2010). *Realities Toolkit #8: Transcribing your own qualitative data.* Retrieved from http://eprints.ncrm.ac.uk/973/2/08-toolkit-transcribing-your-qual-data.pdf

Davis, M. (n.d.). Psycholinguistic evidence on scrambled letters in reading. Retrieved from http://www.mrc-cbu.cam.ac.uk/people/matt-davis/cmabridge/

Dragon Speech Recognition (2016). Nuance Dragon Professional Individual: Data Sheet. Retrieved from https://www.nuance.com/content/dam/nuance/en_uk/collateral/dragon/data-sheet/ds-dragon-professional-individual-v15-en-uk.pdf

Europa (26 October 2016). Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016L2102&from=EN

Fogel, S. (2017, October 3). IBM inches toward human-like accuracy for speech recognition. Engadget [Blog post]. Retrieved from https://www.engadget.com/2017/03/10/ibm-speech-recognition-accuracy-record/

Lant, K. (2017, August 23). Microsoft's Speech Recognition is Now as Good as a Human Transcriber. Futurism the Byte [Blog post]. Retrieved from

ISBN 978-615-5511-27-1

https://futurism.com/microsofts-speech-recognition-is-now-as-good-as-a-human-transcriber

Lewin, T. (2015, February 12). Harvard and M.I.T. Are Sued Over Lack of Closed Captions. The New York Times [Blog post]. Retrieved from https://www.nytimes.com/2015/02/13/education/harvard-and-mit-sued-over-failing-to-caption-online-courses.html

Linder, K. (2016). *Student uses and perceptions of closed captions and transcripts: Results from a national study.* Corvallis, OR: Oregon State University. Retrieved from http://info.3playmedia.com/rs/744-UDO-697/images/Student-Survey-Report-10-25-16-Final.pdf

Liyanagunawrdena, T.R., Forster, S. & Nadan, T. (2017, September). *Captioning videos for accessibility: A case study of University College of Estate Management.* Paper presented at the 24th Annual Conference of the Association for Learning Technology, 5 – 7 September 2017, Liverpool. Retrieved from https://altc.alt.ac.uk/2017/sessions/captioning-videos-for-accessibility-a-case-study-1676/

Liyanagunawardena, T. R., & Hussain, A. (2017). Online Distance Education Materials and Accessibility: Case Study of University College of Estate Management. *Proceedings of the E-Learning, E-Education, and Online Training. Third International Conference, eLEOT 2016, Dublin, Ireland, August 31 – September 2, 2016*, 79-86. doi: 10.1007/978-3-319-49625-2_10

Oxford Living Dictionary. (n.d.). Accessibility. Retrieved from https://en.oxforddictionaries.com/definition/accessibility

Straumsheim, C. (2017, March 6). Berkeley Will Delete Online Content. Inside Higher ED [Blog post]. Retrieved from https://www.insidehighered.com/news/2017/03/06/u-california-berkeley-delete-publicly-available-educational-content#.W-QIwc1SSlI.link

W3C (24 March 2018a). What is Web accessibility. Retrieved from https://www.w3.org/WAI/fundamentals/accessibility-intro/#what

W3C (22 June 2018b). Web Content Accessibility Guidelines (WCAG) Overview. Retrieved from https://www.w3.org/WAI/standards-guidelines/wcag/#intro

Acknowledgement

I would like to thank Graham North, Dr Peter and Dr Adrian Shell for their support.

ISBN 978-615-5511-27-1