Research Report

Comparison of accuracy of machine-generated or human-generated captions of Zoom live lectures in a comparative theriogenology course

Margaret Root Kustritz,[a] Ryan Rupprecht,[b] Perle Zhitnitskiy[c]
[a]Department of Veterinary Clinical Sciences, University of Minnesota College of Veterinary Medicine, St. Paul, MN, USA
[b]Office of Academic and Student Affairs, University of Minnesota College of Veterinary Medicine, St. Paul, MN, USA
[c]Department of Veterinary Population Medicine, University of Minnesota College of Veterinary Medicine, St. Paul, MN, USA

Abstract

Captions are available with captured lectures for student review. In this study, automatically generated captions from Zoom, Kaltura, and YouTube were compared for accuracy with captions generated by a human being. Also investigated was the effect of speaker on accuracy of captioning – does the speed with which someone speaks or their accent alter accuracy of captioning? YouTube was by far the most accurate of the automatic captioning systems. There were numerous mistakes made by Zoom and Kaltura and some significantly altered meaning. Mistakes were due to the transcribing systems, not to things specific to the presenters. Instructors should get in the habit of reviewing tran scripts of their lectures to ensure students are not misled.

Keywords: Accommodations, accessibility, public speaking

Introduction

Lecture capture is a widely accepted tool in health sciences education. Lecture capture systems vary from simple cameras on a tripod with a microphone to pick up the speaker's voice; to a webcam and computer microphone to capture lectures in a digital conference system; to extensive systems that capture audio, a video following the speaker as they move or demonstrate things, and the visual displayed through the lecture hall projector. Students use captured lectures to permit them to listen to lectures that they were unable to attend in-person or at presentation, to ensure their own notes are accurate, to review complex or confusing subjects as they study, and to prepare for examinations.[1–4] Availability of captions permits students to 'watch' the video in spaces where they cannot play the audio aloud (e.g. library or other quiet study space).[5] Students appreciate the flexibility this tool affords and feel that it supports their wellbeing.[6,7]

Many of these lecture capture systems also create captions, either in real-time or after the video is saved. Captions generally are created as transcribed text of spoken information, not inclusive of audio tracks from videos or other audible content not provided by the presenter or participants and may be time-stamped. Captions are beneficial for student learning in many ways. Historically, captions were provided for students with specific learning needs (e.g. hearing loss) but it has been demonstrated that provision of captions can enhance learning in students with a variety of other disabilities including autism spectrum disorder, attention deficit hyperactivity disorder, or dyslexia.[5] A written transcript is also helpful for those students for whom English is not their first language.[3] This may be increasingly necessary as we strive to create a more diverse student body, which may be associated with increasing responsiveness to accessibility concerns.[4]

Students without a diagnosed disability also benefit from captioning. Acoustic and visual stimuli are processed in different areas of the brain and taking in information through both channels can strengthen learning as higher level neurologic processes must be employed to integrate and encode these different inputs into a coherent piece of knowledge.[1,8,9] Joint presentation of auditory and visual data increases comprehension compared to presentation of auditory data alone.[9]

Captioning is accomplished by a human or through an automated system. Use of any system may help to overcome certain limitations of the speakers; some literature suggests that

CONTACT Margaret V. Root Kustritz rootk001@umn.edu

Citation line: Clinical Theriogenology 2023, 15, 9596, http://dx.doi.org/10.58292/CT.v15.9596

students may benefit from having a written transcript if the speaker has an accent or speaks very quickly or if audio quality is poor.[5,10–12]

Human captioners are very accurate, with reported accuracy of over 95% for those with some level of familiarity with the content they are captioning and only a slightly lower accuracy (93%) for those who are less familiar with the content.[13,14] A human being can readily adjust to a speaker's accent, speed of presentation, and volume, and can work around background noise and other distractions.[12] A human captioner could also ask for things known to be more difficult to catch in a lecture and to be provided with these ahead of time, for example, technical terms, proper nouns, and acronyms or other terms that contain letters or numbers.[15] Human captioners can be very accurate if they caption after production, where they are watching a saved video and thus can slow it down or speed it up as needed and repeat sections where the speech is unclear. Human captioners also ensure that the spelling, grammar, and punctuation are accurately transcribed so as to create an easily readable document.[11] Human captioning is expensive and may require scheduling such that there is substantial lapse of time between presentation of the lecture and availability of captions, limiting timely review of lectures by students.

Speech recognition systems are used for automatic captioning and vary greatly in accuracy depending on the system used and the level of specialized terminology in the discipline of the presentation. Automatic captioners are not generally as accurate as humans and do not produce documentation as an easily readable document, with some using no punctuation at all or otherwise providing a 'stream of consciousness' of information that may be difficult for novices to understand out of context.[16,17] Speech recognition systems may or may not include fillers ('um' or 'er') and transcribe words that sound the closest to what was said whether or not their transcription makes any sense. Discipline-specific terminology is not well transcribed and may be presented in a manner that changes the intended meaning.[13] Automatic captioning is less expensive than human captioning and can be done in real-time or after production without need for scheduling so the lecture and associated captions can be made available soon after the presentation.

At our college, lecture capture has been used for over 10 years. Over that time, faculty have become more accepting of the value of this tool and students have come to expect that all lectures will be captured. Technology used also has varied over time. Evaluation of new technology options was underway at the onset of the COVID-19 pandemic when our college had to make an abrupt switch from in-person to virtual lectures. The digital conference system that was supported by the university was Zoom[a] and our college has continued to use Zoom as the pandemic had its course. Currently, lectures are offered either in-person and on Zoom concurrently or completely on Zoom. Zoom lectures are available to students and faculty with captions generated by Kaltura[b]; lectures are uploaded within 24 hours but captions may take longer to load dependent on usage of the system, which is shared by colleges across the university. Kaltura captions are automatically appended to these videos using automatic speech recognition, which is described by Kaltura as up to 90% accurate, and are posted in Canvas,[c] the learning management system of our college. This is the default captioning method at our college. Faculty are recommended by Kaltura to review and edit their captions; at present, our college does not require this, and it is not known how many faculty review their captions for accuracy.

Zoom also creates captions if requested by the presenter, and these are visible in real-time and are sent to the presenter or host of the meeting as a text file after the session is ended. Videos also can be uploaded to YouTube[d] and captions generated through that website; however, this is not commonly done at our college.

Captions created using Kaltura, Zoom, and YouTube are generated by machine learning algorithms and details of these systems are proprietary. Machine learning algorithms are subject to limitations in quality due to how their algorithm is built, vocabulary available within the captioning system, effect of background noise, the clarity and volume of the presenter's voice, presence of multiple overlapping voices, mispronunciations by the speakers, and accents including dialects.

Comparative Theriogenology is a 3-credit course offered in third year spring. A variety of instructors participate in the course. Lectures and review sessions are offered live via Zoom and those sessions are recorded. The lectures and associated captions generated by Kaltura are posted in the course Canvas site for student access. The goal of this study was to evaluate what factors may make automatic captioning more or less accurate for a cohort of instructors in a course.

Materials and methods

Data were collected from 6 instructors in the course. For each, a 25-minute segment from the beginning of their presentation (after initial greetings and assurances that the technology was working) was evaluated for number of homonym errors, number of other errors, and number of medically significant errors, and speaking rate (words per minute [wpm]) with and without repeaters (e.g. 'um' and 'er'). An accurate transcript was created for each presentation by the first author after listening to each presentation at 0.5 speed and at full speed. Speaking rate was calculated from the corrected transcript. Machine-generated transcripts were generated in Kaltura, Zoom, and YouTube. A trained captioner who was not a medical expert created the human-generated transcripts.[e] YouTube and Zoom generated transcripts with time stamps and no punctuation. Kaltura and the human captioner generated documents that looked more like course notes, with punctuation. The first author compared the transcripts from Kaltura, Zoom, YouTube, and the human captioner to the accurate transcripts. Homonyms were not counted as errors in any system. Errors included missed words, additional words, and incorrect words. A given error was counted either as an overall error or a medically significant error; the latter changed the meaning in ways that would have taught the reader something that was medically incorrect. Comparison of characteristics of speakers by captioning system was analyzed by either the paired or unpaired Student's t-test. Correlation between speaking rate and number of errors was evaluated by calculation of the Pearson correlation coefficient. Overall errors and medically significant errors by system for all speakers were analyzed using ANOVA and the paired t-test. Significance was set at $p < 0.05$.

[a] https://zoom.us/

[b] https://corp.kaltura.com/

[c] https://www.instructure.com/canvas

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

[e] themaritimer at Fiverr

Table. Instructor characteristics

Characteristic	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5	Speaker 6
Male/Female	M	F	M	F	M	F
English as first language	N	Y	Y	Y	Y	N
Accent?	Y	N	Y	N	N	Y
Words per minute	146	165	179	199	158	124
Fillers/100 words	4.7	4.9	2.6	0.2	1.3	10.3

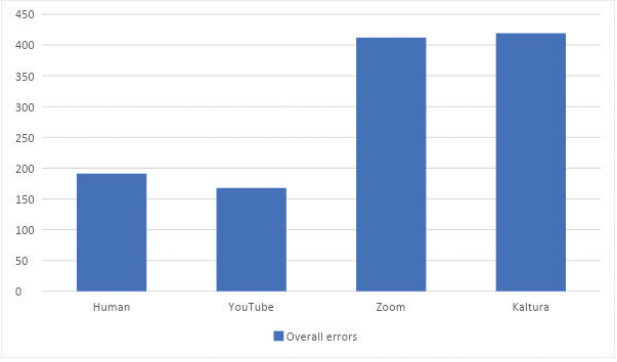

Figure 1. Mean number of overall errors per captioned lecture by captioning system

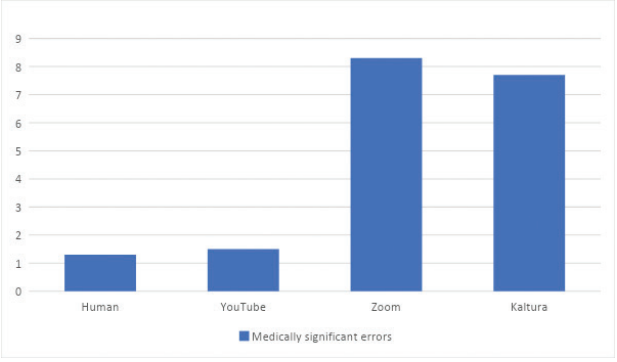

Figure 2. Mean number of medically significant errors per captioned lecture by captioning system

Results

Speaker characteristics are summarized (Table). There were no differences in number of overall errors in any system for men versus women, or for those for whom English was their first language versus those with a different first language. For those with accents versus those without accents, those with accents had fewer ($p = 0.03$) overall errors than those without only in Kaltura. There was no demonstrated correlation between words per minute and number of overall errors or number of medically significant errors.

Number of overall errors was much larger than the number of medically significant errors. The human captioner and YouTube were not different and both were superior to Zoom and Kaltura, both for number of overall errors ($p < 0.0001$) and number of medically significant errors ($p = 0.0003$) (Figures 1 and 2).

An example of a significant error was the following – Response when students were asked the 3 most common postpartum disorders in dogs – 'hypercalcemia, meteorities, and mastodons' OR 'hypoglycemia, detritus, and mastitis' – Should be 'hypocalcemia, metritis, and mastitis'.

Medical and scientific terms were commonly poorly transcribed but common words also were poorly transcribed including sow (south, cell, stylist) and milk (nuke, pill, melt). Favorites were 'evil looting' for involution and 'arctic eye shadow' for Dr. Caixeta.

Discussion

The default captioning system used by our college is much less accurate than one other commercial system or the human captioner. Knowing the extent of inaccuracy of the current system is valuable in helping the college plan for what technologies are best supported for lecture capture and captioning.

This study was limited in that only 6 individuals were evaluated and only 1 lecture was evaluated for each. More accurate data may have been generated by a wider data collection. Creating the gold-standard transcript and comparing all others to that gold-standard was very time-consuming; specific effort would have to be granted for a greater breadth of data collection. Another limitation is a practical one regarding the need for these data and future data collection. We do not have an easy way to determine how many students access captions or transcripts or how frequently a given student does so. If few students are using the captions because of their inaccuracy, we are missing a great opportunity to help support student learning but if many are accessing the captions and are managing despite the inaccuracies, we may be creating a solution for a problem that does not exist. Finally, our college is part of a larger unit and has somewhat limited capability to say what technologies it will and will not support, so any conversations about changing from the default system currently, will require extensive documentation and discussion.

Among the different transcript-generating software evaluated in this study, YouTube was the most accurate. This is in alignment with previously published literature that reported a 98% total accuracy when YouTube was used for the transcript of a college-level recorded lecture.[18] When comparing the accuracy of transcription, there were no significant differences based on the presenter's gender; this result was unexpected and at odds with a previous report of significant decrease of accuracy when the speaker was female.[19] This difference may be due to the limited sample size in this study.

Citation line: Clinical Theriogenology 2023, 15, 9596, http://dx.doi.org/10.58292/CT.v15.9596

This study did not evaluate whether speaking very quickly or having a strong accent alters student learning. A study evaluating speed of speech demonstrated that students were more engaged with presenters spoke moderately quickly (172 wpm) or very quickly (213 wpm) compared to when they spoke slowly (116 wpm).[20] Understanding the spoken word requires one to map acoustic sensory input to stored representations of sounds, grammar, and syntax, and listening to someone with a strong accent can require more cognitive effort and slow one's ability to understand.[21] Speakers who use more than 1.3 filler words per 100 content words are deemed by listeners as less credible, which may impair engagement and learning.[22] Only 1 of the 6 speakers in this study was below that limit, with most speakers well above that limit. In this study, characteristics of the speakers were not associated with changes in captioning. Speakers who know they speak slowly or extremely quickly, have a strong accent, or use many filler words may wish to strongly encourage students to use captions to ensure they are fully understanding what was presented.

Multiple studies have identified that the availability of captions is associated with increased focus, increased understanding, better retention, and overall higher academic achievement.[4,23–25] Caption transcripts also are searchable that makes finding specific parts of lectures to be rewatched more efficient for learners.[1,23] Concerns have been expressed about cognitive load on students required to read captions while watching and listening to a video but students surveyed in a study did not report feeling undue pressure from this activity.[24]

Caption transcripts may be generated without punctuation or other attention to what makes for a readable document. In one study, transcripts were shown to be most valuable for learners if they were read while the students were hearing the video.[1] Some authors suggest that poor captions are worse than no written transcript at all as they may confuse the student regarding what is correct and trying to read the poorly transcribed wording may distract the student from hearing important information.[4,11] All transcripts should be evaluated for accuracy and edited as needed; faculty must be given time to complete this as part of their teaching effort. Some literature describes students, either individually or collectively, doing this editing as a part of the course in which they are enrolled.[17,26]

In this study, human captioning was not statistically different from YouTube but the mean number of overall errors was lower for YouTube (168) than for the human captioner (191). Technology is improving and at least one study has already demonstrated automatic captioning that is as accurate as human captioning.[27] Organizations should pay attention to advances in automatic captioning systems to take advantage of improved technology.

Conclusion

Projections in education are that more virtual training and more use of technology in all aspects of education will be the norm as we emerge from the COVID-19 pandemic.[28,29] This means that technologies such as captioning will need to be better understood and resources put toward making them as efficient and useful as possible. In this study, mistakes were due to the transcribing systems, not to specific characteristics of the presenters. However, speakers can improve student engagement and decrease cognitive load by watching how quickly they speak, by practicing their speech to minimize the number of filler words used, and to encourage students to use caption transcripts when reviewing captured lectures. All transcribing systems made some mistakes and instructors should get in the habit of reviewing transcripts of their lectures to ensure students are not misled. Students can best use these transcripts to search for specific areas of lectures to be rewatched, and by reading the transcript while hearing the presentation.

Conflict of interest

None.

Authors' contribution

Dr. Root Kustritz conceptualized the study, identified the resource for human evaluation of captions, gathered data from transcripts and analyzed, drafted the paper, and completed revisions requested by the reviewer. Mr. Rupprecht uploaded videos to YouTube and downloaded captions, investigated mechanisms used by mechanical captioning systems, and reviewed the paper before submission. Dr. Zhitnitskiy provided input on the study and provided an in-depth review of the paper as an initial draft. All authors have read and approved the final version of the manuscript and have agreed to the submission.

References

1. Dommett EJ, Dinu LM, Van Tilburg W, et al: Effects of captions, transcripts and reminders on learning and perceptions of lecture capture. Int J Educ Technol High Educ 2022;19:20 doi: 10.1186/s41239-022-00327-9
2. Dommett EJ, Gardner B, Van Tilburg W: Staff and students perception of lecture capture. Internet High Educ 2020;46:110732. doi: 10.1016/j.iheduc.2020.100732
3. Newton G, Tucker T, Dawson J, et al: Use of lecture capture in higher education – lessons from the trenches. Tech Trends 2014;58:32–44. doi: 10.1007/s11528-014-0735-8
4. Morris KK, Frechette C, Dukes L, et al: Closed captioning matters: Examining the value of closed captions for all students. J Postsec Educ Disability 2016;29:231–238.
5. Griffin E. Who uses closed captions? Not just the deaf or hard of hearing: 2015. Available from: http://www.3playmedia.com/2015/08/28/who-uses-closed-captions-not-just-the-deaf-or-hard-of-hearing/ [cited 15 July 2022].
6. Dommett EJ, Gardner B, Van Tilburg W: Staff and student views of lecture capture: a qualitative study. Int J Educ Technol High Educ 2019;16:1–2. doi: 10.1186/s41239-019-0153-2
7. Dommett EJ, Van Tilburg W, Gardner B: A case study: views on the practice of opting in and out of lecture capture. Educ Inform Technol 2019;24:3075–3090. doi: 10.1007/s10639-019-09918-y
8. Mayer RE: Cognitive theory of multimedia learning. In: Mayer RE: editor. The Cambridge handbook of multimedia learning. New York; Cambridge University Press: 2014:43–71. doi: 10.1017/CBO9781139547369.005
9. Moreno R, Mayer RE: Verbal redundancy in multimedia learning: when reading helps listening. J Educ Psychol 2002;94:156–163. doi: 10.1037/0022-0663.94.1.156

10. Tisdell C, Loch B: How useful are closed captions for learning mathematics via online video? Int J Math Educ Sci Technol 2017;48:229–243. doi: 10.1080/0020739X.2016.1238518
11. Millett P: Improving accessibility with captioning: an overview of the current state of technology. Can Audiol 2022;9. Available from: https://canadianaudiologist.ca/improving-accessibility-with-captioning-an-overview-of-the-current-state-of-technology [cited 15 July 2022].
12. Lasecki WS, Bigham JP: Real-time captioning with the crowd. ACM Digital Library: Interactions 2014;21:50–55. doi: 10.1145/2594459
13. Lasecki WS, Miller CD, Sadilek A, et al: Real-time captioning by groups of non-experts. Proceedings of the 25th annual ACM symposium on User interface software and technology, October 2012; 23–34. doi: 10.1145/2380116.2380122
14. Wald M: Creating accessible educational multimedia through editing automatic speech recognition captioning in real-time. Interact Tech Smart Educ 2006;3:131–141. doi: 10.1108/17415650680000058
15. Takeuchi Y, Kojima D, Sano S, et al: Detection of Input-Difficult Words by Automatic Speech Recognition for PC Captioning. In: Miesenberger, K., Kouroupetroglou, G. (eds) Computers Helping People with Special Needs. ICCHP 2018. Lecture Notes in Computer Science, vol 10896. Springer, Cham. doi: 10.1007/978-3-319-94277-3_32
16. Kent M, Ellis K, Latter N, et al: The case for captioned lectures in Australian higher education. Tech Trends 2018;62:158–165. doi: 10.1007/s11528-017-0225-x
17. Borgaonkar R: Captioning for classroom lecture video. Thesis: University of Houston: 2013.
18. Millett P: Accuracy of speech-to-text captioning for students who are deaf or hard of hearing. J Educ Ped Rehab Audiol 2021;25:1–13.
19. Tatman R: Gender and dialect bias in YouTube's automatic captions. Proc ACL Workshop on Ethics in Natural Language Processing, Valencia; Association for Computational Linguistics: 2017;53–59.
20. Simonds BK, Meyer KR. Quinlan MM, et al: Effects of instructor speech rate on student affective learning, recall, and perceptions of nonverbal immediacy, credibility, and clarity. Comm Res Report 2006;23;187–197.
21. Van Engen KJ, Peelle JE: Listening effort and accented speech. Front Hum Neurosci 2014;8:577. doi: 10.3389/fnhum.2014.00577
22. Duvall E, Robbins A, Graham T, et al: Exploring filler words and their impact. Schwa Lang Linguistics 2014;11:35–49.
23. Ellis K, Kent M, Peaty G: Captioned recorded lectures as a mainstream learning tool. M/C J 2017;20. doi: 10.5204/mcj.1262
24. Whitney M, Dallas B: Captioning online course videos: an investigation into knowledge retention and student perception. Minneapolis, MN: Proc ACM Technical Symposium on Computer Science Education; 2019.
25. Ranchal R, Taber-Doughty T, Guo Y, et al: Using speech-recognition for real-time captioning and lecture transcription in the classroom. IEEE Transactions Learn Technol 2013;6:299–311. doi: 10.1109/TLT.2013.21
26. Wald M: Using speech recognition transcription to enhance learning from lecture recordings. International Conference on Education and New Developments, Budapest, Hungary. 23–25 Jun 2018;111–115.
27. Xiong W, Droppo J, Huang X, et al: Achieving human parity in conversational speech recognition. Microsoft Technical Report. MSR-TR-2016-71, 2017; Available from: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/ms_parity.pdf [cited 10 May 2023].
28. Guppy N, Verpoorten D, Boud D, et al: The post-COVID-19 future of digital learning in higher education: views from educators, students, and other professionals in six countries. Br J Educ Technol 2022;53:1750–1765. doi: 10.1111/bjet.13212
29. Pew Research Center. Experts say the 'new normal' in 2025 will be far more tech-driven, presenting more big challenges, 2021. Pewresearch.org/internet/2021/02/18/experts-say-the-new-normal-in-2025-will-be-far-more-tech-driven-presenting-more-big-challenges/.