Curtin University

Automatic Closed Captions and Immersive Learning in Higher Education

Katie Ellis, Kai-Ti Kao and Mike Kent

February 2020

Automatic Closed Captions and Immersive Learning in Higher Education

Authored by **Katie Ellis, Kai-Ti Kao and Mike Kent**

Published in **2020**

The Centre for Culture and Technology Curtin University

Website: **https://ccat.curtin.edu.au**

Email: **Katie.ellis@curtin.edu.au**

Telephone: **+61 8 9266 2509**

Table of contents

Figures and Tables

Acknowledgements

The research was funded by Echo360 via their 2019 Echo360 research grant. Preliminary research was funded by Curtin University's Teaching and Learning Centre and the School of Media, Creative Arts and Social Inquiry.

Colleagues at Curtin University in the Centre for Culture and Technology, School of Media, Creative Arts and Social Inquiry and Research Office have also provided invaluable support in administering this project.

An earlier version of the literature review addressing diverse students was authored by Natalie Latter. It was previously published in the report *Mainstreaming Captions for Online Lectures in Higher Education in Australia: Alternative Approaches to Engaging with Video Content at Curtin University*. We also acknowledge previous staff over the course of our captions research program, including Kathryn Locke, Natalie Latter, Gwyneth Peaty, Leanne McRae and Chris Mason.

Key researchers on this project were Kai-Ti Kao and Kate Corkish. The final report was formatted and proofread by Ceridwen Clocherty.

Executive summary

Captions are the text version of speech and sound in audio-visual media superimposed on the bottom of the screen typically to provide access to people who can't hear audio for whatever reason. While captions are most often associated with film and television entertainment, their relevance in an educational context is becoming increasingly apparent as a result of theories such as universal design for learning (UDL) and immersive learning.

In the higher education setting, captioned video lectures are a vital accessibility feature for students who are D/deaf or hard of hearing. In addition, students with diverse learning styles, older students, and students who experience difficulty accessing online videos for reasons related to issues with their environment (noise) or with technology (connectivity or equipment) also benefit from captions. As such, increasingly accurate automatic speech recognition (ASR) software employed within the Echo360 active learning system has the potential to revolutionise online learning via the widespread availability of captions.

This report details findings of the Echo360-funded project *Automatic Closed Captions and Immersive Learning in Higher Education*. The project sought to determine the usefulness of captioned lectures to the broader student population. In addition to reviewing developments regarding the creation of captions via speech recognition software, the international standards regarding caption accuracy, and the pedagogies of teaching and learning, the project also interviewed 53 online students enrolled in 11 Digital and Social Media, Screen Arts and Fine Art units at Curtin University during study periods 3 and 4 in 2018 in which Echo360 captions had been mainstreamed for the purpose of the project.

The research attempted to answer the following questions:

- What are the expectations regarding captioning accuracy of online lectures in an international context?
- What level of accuracy do we see in the transcripts generated by the Echo360 ASR functionality built with Amazon Transcribe?
- How does the mainstream student population perceive the experience of immersive learning via captioned lectures and the ASR transcript window?
- Can mainstream captions improve learning outcomes for students with and without disabilities?

The report has three parts. Part 1 is a literature review addressing two key areas of research. First, the potential benefits of captioned lectures for the broader student population and, second, the use of ASR in making these captions available, with a particular focus on their accuracy. Captions, while initially intended for people who are D/deaf and hard of hearing, were then seen to have benefits from students from other at risk groups, including those with English as an additional

language and other disability groups, and are now also embraced as a key resource for the mainstream student population via theories of UDL.

Part 2 presents a scoping study of international standards regarding caption accuracy. A significant number of international government, industry and advocacy organisations have articulated captioning standards and recommendations, both in general and pertaining to their automation. This includes the World Federation of the Deaf, the International Federation of Hard of Hearing People (IFHOH), 3Play Media, the Described and Captioned Media Program (DCMP), the Federal Communications Commission (FCC), Media Access Australia / Centre for Inclusive Design, Deafness Forum of Australia, AI Media, the UK government's Office of Communications (Ofcom), the World Wide Web Consortium (W3C), the Canadian Radio–Television and Telecommunications Commission (CRTC) standards, and Netflix. However, available industry standards regarding caption accuracy standards for online lectures remains unclear. While the preferred accuracy rate of 99% is often cited, this figure relates only to US legislation; in the Australian context, 95% is often cited as the international standard. This section covers some key definitions of what accuracy means in relation to captioning, as well as industry, advocacy group and educational interpretation of their use, including specific information regarding captioning on Echo360.

Part 3 of the report is concerned with the findings of, and offers discussion on, the results of the interview stage of this research with the 53 students who participated in the project. These students were enrolled in 11 Digital and Social Media, Screen Arts and Fine Art units which trialled mainstreaming captions in two study periods in 2018, resulting in data from 22 units in total. Insights were gained into both how students actually used the captions and how they anticipated using them in the future should Curtin University embrace them as a mainstream approach.

The report recommends:

- It is clear from this research that students both like and expect captions in a Higher Education setting. Automated captions provide a cost effective alternative to traditional captioning and should be turned on by default.

- Further study is needed into the impact of different error rates on the effectiveness of captions for student learning, and what can be considered effective.

- Further research is needed on the impact and use of captions by specific user groups, including the broader student population, those with English as a further language, students who are Deaf or hard of hearing, and those with learning disabilities.

- Captions need to be used in conjunction with training for presenters to make the best use of automated systems including appropriate use of audio recording systems, protocols for including questions and comments from people in presentations who are not captured by recording equipment and an understanding of the requirements of an audience that is not present in a lecture theatre or classroom.

Introduction

Background and historical context of captions

Captions are the text version of speech and sound in audio-visual media superimposed on the bottom of the screen typically to provide access to people who can't hear audio for whatever reason. While captions are most often associated with film and television entertainment, their relevance in an educational context, for students with and without hearing impairments, has been apparent since the earliest days of the technology. For example, in the 1950s educators in US schools for the D/deaf and hard of hearing captioned movies for their students (Downey, 2007). When these captioned films improved the educational outcomes for this group, educators began to hypothesise on how captions could benefit a broader range of students, such as students without any hearing impairment. As Malcolm Norwood reflected when opening the *First National Conference on Television for the Hearing-Impaired* in 1971, standards regarding caption quality change across time (Perkins, 1971, p. 3):

> *When the Office of Education began to caption motion pictures for deaf children and adults some 11 years ago, the subtitles were geared to a reading speed of 120 words per minute. Believe me, we had our share of complaints regarding the speed of the captions. Approximately one and a half years ago, we unilaterally increased the reading speed from 120 words per minute to 144 for all films aimed at adult audiences. We haven't received a single squawk. I mention this to you as a matter of interest for if captions have contributed to the advancement of our deaf population, what will they do for the general population?*

When captions were introduced on television during the 1970s there was no expectation that they would be verbatim. The average reading level of the D/deaf community was thought to be at a third-grade level. As a result, the audio content was reduced by one third and captions did not reflect idioms, puns, swearing, nor sentences in the passive voice. However, as technology improved, allowing more people access to captions, as legislative measures were introduced, and as audiences who were D/deaf and hard of hearing developed a caption literacy, the importance of verbatim captions began to be established in the 1980s and 1990s – see Figure 1 for a detailed timeline. Today, UK news readers speak and are captioned at a rate of 220 words per minute while in Spain this reportedly goes as high as 600 words per minute (Romero-Fresco & Perez, 2016).

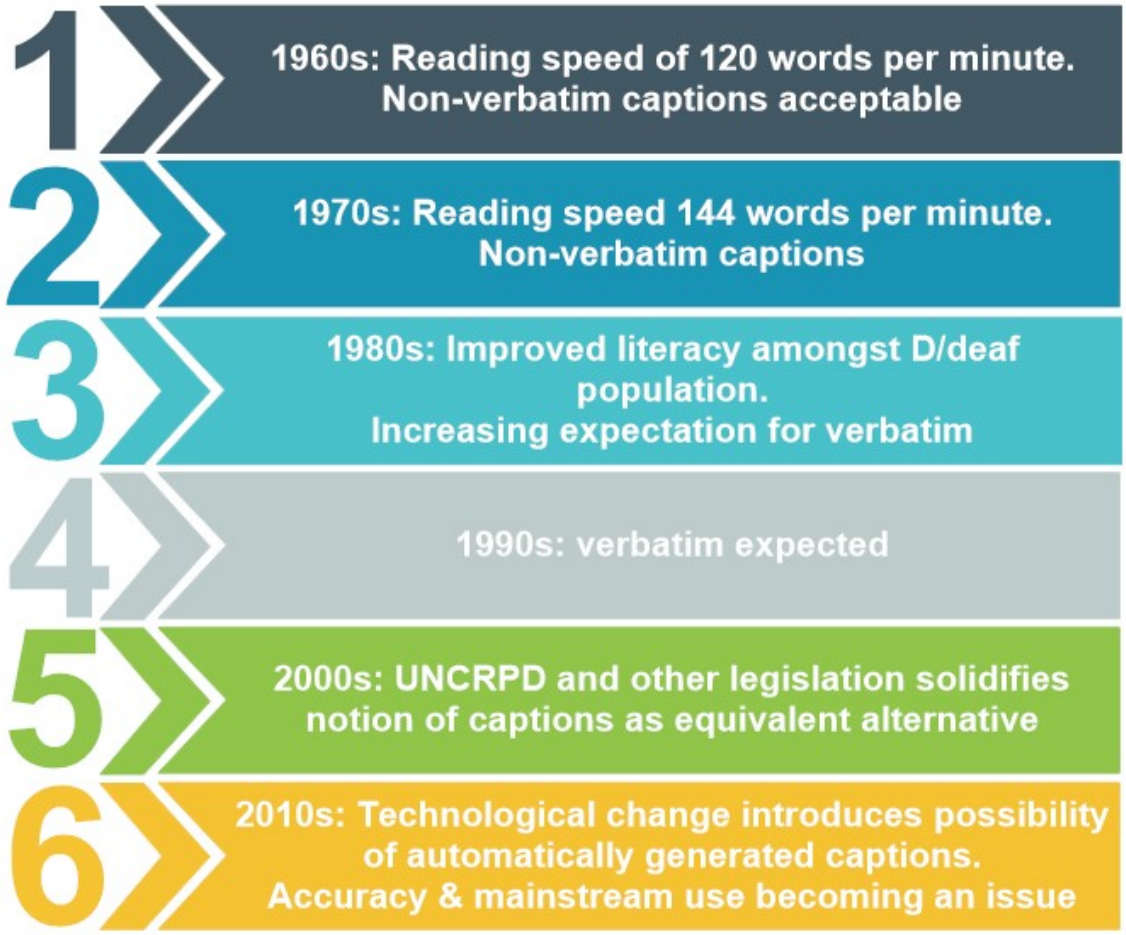

Figure 1. Changing expectations about captions since the 1960s

Additionally, from the beginning of this century, verbatim captions have also been more widely expected, in part due to treaties such as the United Nations Convention on the Rights of Persons with Disabilities (UNCRPD) and, in the Australian context, the Disability Standards for Education. As 3Play Media (3Play Media, 2018) explain:

> *Caption quality matters because captions are meant to be an equivalent alternative to video for individuals with hearing loss. When captions are inaccurate, they are inaccessible.*

Furthermore, advances in machine learning and artificial intelligence (AI) have both introduced and increased the possibility of automatically generated captions using automatic speech recognition (ASR). For example, the video sharing site YouTube began offering this feature to broadcasters and universities in 2009 and then to the general public the following year (Klie, 2010). However, a rate of 7.7 errors per minute established the feature as too inaccurate to be exclusively used in the context of online learning (Parton, 2016).

From speed, to literacy, to verbatim, with advances in AI and ASR, caption quality has now zeroed in on the notion of accuracy. Accuracy refers both to a lack of errors and the inclusion of punctuation and background sounds.

More recently, with the introduction of automatically generated captions using ASR on popular social media platforms such as Facebook and YouTube has increased, with diverse audiences now embracing this feature to watch videos in environments when sound would be inappropriate. In recognition of the increased use of – and demand for – ASR technology, the World Federation of the Deaf (WFD) and International Federation of Hard of Hearing people (IFHOH) released a joint statement encouraging further research in this area (WFD & IFHOH, 2019):

> *The field of Automatic Speech Recognition (ASR) has progressed significantly with the advancement of artificial intelligence technology. As a result, more applications that utilise ASR and AI technologies are developed and have shown a promising impact on communication and accessibility. However, this field is emerging and the World Federation of the Deaf (WFD) and International Federation of Hard of Hearing (IFHOH) are documenting a small number of user experiences and cases using ASR technology. There needs to be continuing effort in research and development with deaf and hard of hearing participation to improve the uses and applications of ASR technology.*

While their statement relates to replacing current communication methods such as telecommunication relay services (TRS), their recommendation to engage with all users' needs is relevant to the current study's focus on accuracy.

Benefits of captions for diverse student groups

Typically used in entertainment media, the pedagogical and accessibility benefits of captions in the educational arena are increasingly being recognised. As such, the educational video platform Echo360 has been seeking ways to improve access to video lectures via the provision of automatic captioning. According to Echo360:

> *Providing closed captions is the best way to ensure the highest level of accessibility for your video content.*

While research shows a clear benefit of captioned lectures to students with disabilities and other at risk groups, trends in online learning and personalised approaches to learning suggest a significant portion of the mainstream student population would also benefit from the delivery of captioned online lectures. Research by 3Play Media and the BBC (Griffin, 2015) demonstrates that captions benefit many people and are not only used by people with hearing loss. For example, captions also benefit students whose first language is not English, some older students, as well as those with learning disabilities, attention deficits or autism.

They help students, both with and without disabilities, to:

- Comprehend content that is spoken very quickly, with accents, or that which includes mumbling or background noise.
- Clarify full names and technical terminology.
- Watch videos in sound-sensitive environments, like offices and libraries.

Captions are therefore a key example of the ways universal design for learning (UDL) can help diverse student groups succeed in the university environment and, with the use of captions increasing in entertainment and social media, students are now beginning to also expect their availability on their lectures (Pitman, Ellis, Kent, & Mancini, 2020). In their book *Reach Everyone, Teach Everyone* (2018), Thomas J. Tobin and Kirsten T. Behling profile a number of students and instructors who benefit from captioned instructional lectures to illustrate the ways UDL has moved from the purview of disability access to an ease of use and general diversity framework. For example, students report using captions when studying while children are asleep, and evidence from both instructors and disability advocates describes the pedagogical benefits of captions for the entire student cohort. Users of captioned online videos report higher user engagement and better user experience (Griffin, 2015). They also offer the opportunity to increase search engine optimisation (MIT, 2016), and allow students to jump to the exact point in a lecture they are looking for during assignment revision.

Furthermore, more widely available captions are a key example of the importance of removing barriers and thereby promoting UDL – this is illustrated in Figure 2. The image displays three scenarios of children of varying heights trying to watch a baseball game from behind a fence. In the first image, they are given the same sized box to stand on. The two tallest children can see over the fence, while the shortest can't. By being given the same supports they are being treated equally. A lecture would be an example of this kind of support. In the next scenario, the tallest child is not given support but can still see over the fence, the second tallest is given one box and can see over the fence, while the shortest is given two boxes and can see over the fence. By being given different supports, they are being treated equitably. Allowing students with disability to request captions – but not offering the same to students with other diverse learning styles – would be an example of this kind of support. In the final picture, the wooden fence is replaced with one that can be seen through and all three can see the game without any support because the systemic barrier was removed. The mainstream availability of captioned lectures offers this kind of support by removing all barriers.

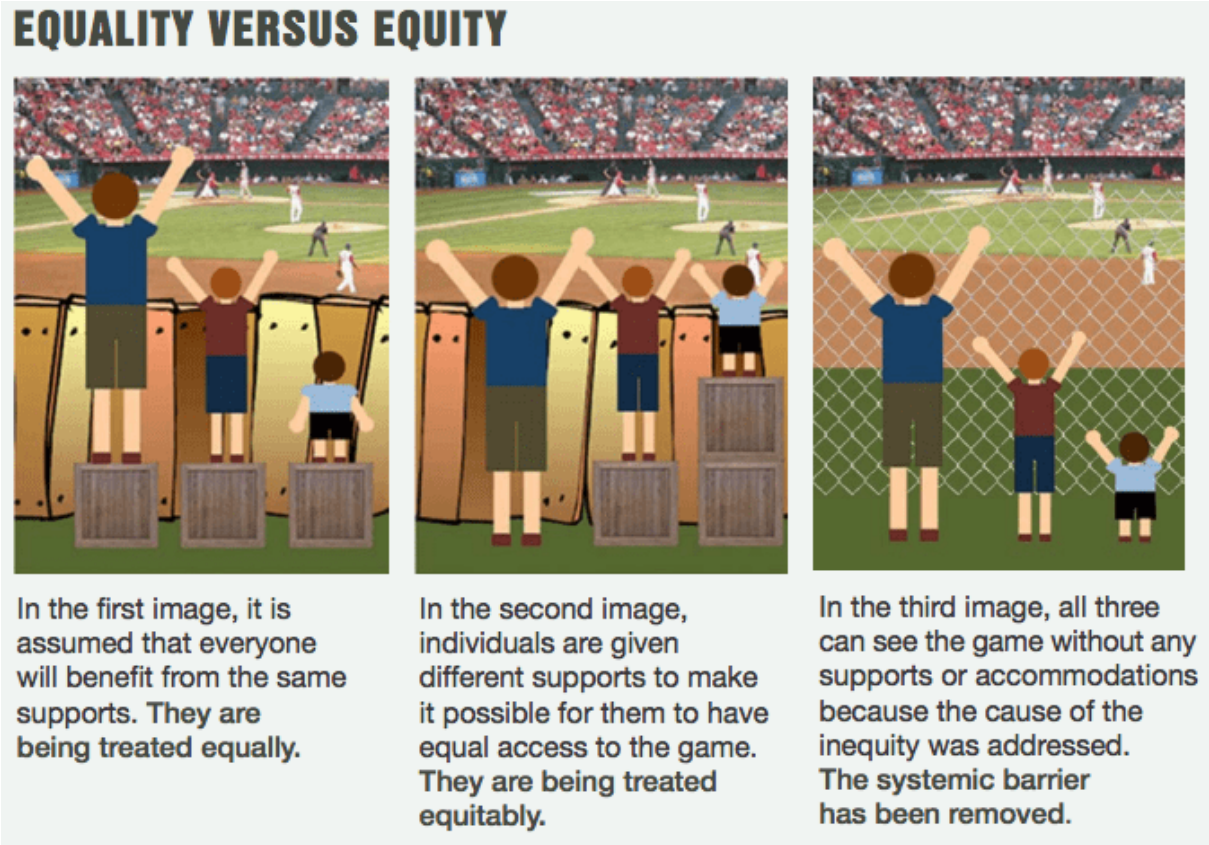

Figure 2. Equality versus equity

Provision of captions to promote immersive learning

Captions also offer an opportunity for immersive learning. The notion of immersive learning is described on the Echo360 blog as a simultaneous engagement with auditory and visual material that has the potential to make lectures more accessible to a significant range of students, from at risk groups to students whose specific learning styles lean towards both visual and auditory materials.

Immersive learning is linked to the cognitive theory of multimedia learning. This theory draws on three key assumptions from theories of learning processes, that:

- People have separate channels for processing visual and verbal information.
- People have limited capacity or cognitive load in their working memory for each channel (visual and verbal) at any given time.
- Students must actively process information for meaningful learning to occur.

Lectures with captions enable dual channel processing of the spoken lecture (verbal) and the written captions (visual). To provide a comprehensive overview of the potential and actual use of captioned lectures by the broader student population and improve understanding of these services, we need to investigate how students use captions and how they could be improved.

Methodology

This project proceeds from the position that captions are a vital pedagogical and accessibility tool that benefit a wide range of students, including people with and without disability. This has been firmly established in our prior research, including the Curtin University-funded 2017 pilot study *Alternative approaches to engaging with video content* in which every student enrolled in three Digital and Social Media units were provided captions by default. A key recommendation of that study was that Australian universities provide this feature to all enrolled students. Further research into students' use of digital technologies in 2018 reconfirmed a widespread student demand for captions. Echo360's ASR functionality now provides this option.

In 2018 we made this service available to students enrolled in 11 Digital and Social Media, Screen Arts and Fine Art units during study periods 3 and 4 in 2018 as part of Curtin's wider review of the technology (see Appendix 1). Students responded to a second survey, again indicating a preference for this facility. A total of 53 students from the 22 units offered over the two study periods participated in these interviews. The current project returns to those students, as well as to international standards and technologies, to answer the following research questions:

- What are the expectations regarding captioning accuracy of online lectures in an international context?
- What level of accuracy do we see in the transcripts generated by the Echo360 ASR functionality built with Amazon Transcribe?
- How does the mainstream student population perceive the experience of immersive learning via captioned lectures and the ASR transcript window?
- Can mainstream captions improve learning outcomes for students with and without disabilities?

The project sought to assess the accuracy and usefulness of Echo360's captions to the entire student population participating in the eleven units across the two study periods. This cohort exist within a media landscape in which the availability of captions is increasingly expected, available and accurate.

The study was funded by Echo360. The program through which the project is funded was established to assist in evaluating the impact of video and active learning strategies on student outcomes and processes for students.

The project adopted a multi-modal methodology across three parts:

Part 1: A literature review. The literature review focuses on two key areas of research. First, the potential benefits of captioned lectures for the broader student population and, second, the accuracy of ASR in this format.

Part 2: A scoping study of international standards regarding caption accuracy. A significant number of international government, industry and advocacy organisations have articulated captioning standards and recommendations, both in general and pertaining to their automation. This includes the World Federation of the Deaf, IFHOH, 3Play Media, the Described and Captioned Media Program (DCMP), the Federal Communications Commission (FCC), Media Access Australia / Centre for Inclusive Design, Deafness Forum of Australia, AI Media, the UK government's Office of Communications (Ofcom), the World Wide Web Consortium (W3C), the Canadian Radio–Television and Telecommunications Commission (CRTC) standards, and Netflix. However, available industry standards regarding caption accuracy standards for online lectures remains unclear. While the preferred accuracy rate of 99% is often cited, this figure relates only to US legislation; in the Australian context, 95% is often cited as the international standard. The scoping study will investigate regulations in other countries to establish a current baseline of existing standards, including specific information regarding captioning on Echo360.

Part 3: The findings of, and discussion on, the results of the interview stage of this research. Online interviews were conducted with 53 students enrolled in 11 Digital and Social Media, Screen Arts and Fine Art units (see Appendix 1 for unit details). Echo360 captions were made available to all students, regardless of dis/ability, enrolled in these units during study periods 3 and 4 in 2018. Evaluation following this exercise discovered students valued the availability of accurate captions as a mechanism through which to more deeply engage with lecture content. The online interview re-introduced students to the feature, asked questions about their possible uses, and obtained insights regarding the potential of immersive learning, for example using both sound and captions when accessing lectures.

It should be noted that there were some limitations which were encountered. While captions were made available to the 11 units each study period, some students were not made aware of how to access them.

Literature review

Benefits of captioned online lectures for the broader student population

Trends in online learning as well as recognition of personalised approaches to learning suggest a significant portion of the student population benefit from captioned online lectures. Adoption of frameworks of UDL in higher education since the 1990s has emphasised the importance of creating learning experiences that offer multiple ways of engaging with content, and of accessing and representing information. Students with and without disability learn and comprehend information in diverse ways depending on the accessibility of information (Rose, Harbour, Johnston, Daley, & Abarbanell, 2006, p. 3).

Captions benefit at risk students such as students who are D/deaf, have learning difficulties, older students, as well as students from non-English speaking backgrounds – (for a comprehensive review see Kent, Ellis, Peaty, Latter, & Locke, 2017). It is argued that the same improvements in comprehension amongst these groups could be achieved within the mainstream student population as many students encounter similar barriers to learning (Shadiev, Hwang, Chen, & Huang, 2014). For example, mainstream students also report difficulties hearing lectures (Fuller, Bradley, & Healey, 2004; Fuller, Healey, Bradley, & Hall, 2004; Healey, Bradley, Fuller, & Hall, 2006; Madriaga et al., 2010). Captions may also be an important tool to help tackle the under-representation of at risk groups in particular subject areas such as science, technology, engineering and maths (Wheatly, Flach, Shingledecker, & Golshani, 2010). Captions have also demonstrated a positive effect on vocabulary acquisition, which may help students learn subject-specific language and vocabulary of these and other academic disciplines.

In addition, the overall retention of course content has been seen to significantly improve for students – both with and without disabilities – when using captions (Steinfeld, 1998). Captions offer accessibility to all, as well as the ability to adapt content to different context, constraints and audiences. Learners can be broadly grouped into three categories – visual, auditory and kinaesthetic (VAK). The so-called VAK learning styles refer to the human observation channels of seeing, hearing and feeling. Echo360's notion of immersive learning recognises these different approaches to learning. That is, many students better comprehend content when it is presented in particular media combinations such as auditory and visually (Moreno & Mayer, 2002). Alty, Al-Sharrah and Beacham (2006) build on this earlier research, confirming that particular media combinations can impact on learner motivation, comprehension – particularly of complex information or large data sets – and cognitive load, as well as improving accessibility for those with different needs, such as students who are D/deaf or hard of hearing. Representing content in multiple ways such as through captioned and transcribed

lectures can therefore enable more students to learn in their preferred way (Schweppe & Rummer, 2016).

There are also proven benefits purely regarding the technology that captions offer. Captions and transcripts can enable more consistent access to content even when it is not possible or suitable to listen to a lecture (Elliot, Foster, & Stinson, 2002; Stinson, Elliot, Kelly, & Liu, 2009) – students can therefore access the content in public spaces or noisy environments, and can avoid having to replay video content whenever background noise interferes with their ability to hear clearly. In addition, the provision of captions can also be an advantage in a purely online learning environment. A much more diverse group of students enrol in online learning as it can be more accessible for many. However, students will not always have access to the ideal technology environment through which to do so – slow or intermittent internet connections, poor quality speakers or headphones, and computers that struggle to stream large files can all negatively impact a student's ability to make use of recorded lectures. The provision of a lecture transcript or, better still, captions, may help to bridge this often invisible technology gap faced by students with diverse learning styles accessing content in diverse learning environments.

Further, creating captions or transcripts for video media can revolutionise the way students index, search and retrieve information (Tuna et al., 2011; Wactlar, Kanade, Smith, & Stevens, 1996). For example, captions uploaded with YouTube videos can be indexed for increased discoverability through search engines (Bond, 2014; Griffin, 2016). In addition, students can more easily search for particular terms to find relevant lecture content when revising (Gernsbacher, 2015).

Accuracy of automatic speech recognition in online lectures

Automatic speech recognition has been seen as a way of overcoming the challenges associated with traditional captioning, particularly those associated with the production and synchronisation of captioned videos (Federico & Furini, 2012, p. 2). Traditionally, real-time captioning is performed by human captioners whose highly specialised skills and intensive training mean that they are often highly in demand and need to be booked well in advance. They are also a costly option which, when combined with their limited availability, mean that they are not well suited to any impromptu or casual interactions which may benefit from captions (Kawas, Karalis, Wen, & Ladner, 2016).

However, while ASR has been viewed as a cheaper, more accessible option for generating captions, there have been concerns regarding issues with accuracy, latency, and ASR's limited capacity to discern context (Kawas et al., 2016). In addition, it has been noted that ASR struggles to manage variations in speakers' accents and speaking rates, background noise, and multiple or synchronous speakers, all of which are common occurrences in the higher education context (Kawas et al., 2016).

Indeed, accuracy is frequently raised as a key issue in relation to ASR in general; however, this is often measured in ideal contexts, that being a single, trained speaker in a quiet environment. There have not been many studies that have specifically sought to establish a viable standard of caption accuracy within the context of higher education. Those that do exist are now dated and do not clearly reflect our social and cultural relationships with recent advances in communication technology. For example, Hede's (2002) study into students' reactions to speech recognition technology in lectures reported accuracy rates between 56–85% and a lukewarm reception from students themselves. He observed that "cognitive overload" from audio, visual and screen text – captions – may be a contributing factor to the lack of student enthusiasm; however, this is a far cry from today's average student who is not just familiar with but also deeply entrenched in a multimodal communication environment.

Stuckless (1999) was one of the earliest to consider the use of ASR within an educational context, suggesting that an accuracy rate of 97.6% or higher may be needed for full comprehension of the material due to the additional challenges and complexity of higher education. He also draws an important distinction between "word accuracy" and "readability", pointing out that speech recognition technology can produce text with 100% accuracy but very low readability if it cannot indicate sentence markers nor change in speakers. Given that today's higher education environments are both multicultural and transnational in nature, it would also be important to consider how speakers' accents and the additional needs of students with English as a further language to factor into accuracy measurements.

A more recent study of captions in higher education by Papadopoulos and Pearson (2012) determined that an accuracy of 88.5% or higher is required for automatically generated transcripts of lectures to be considered usable. Although their study focused on the production of such transcripts as post-lecture material rather than live captions, their findings are consistent with other studies that indicate a high accuracy rate of 90–98% is necessary in a higher education setting (Kheir & Way, 2007; Kushalnagar, Lasecki, & Bigham, 2014; Pan et al., 2010; Stuckless, 1999).

Although accuracy is widely acknowledged as a key concern when it comes to ASR, the way that accuracy is assessed does come under some debate. A commonly used measure of accuracy (Kafle & Huenerfauth, 2017, p. 166) is word error rate (WER), where WER is:

> *… based on S (number of erroneous substitutions of one word for another), D (number of deletions, i.e. erroneous omissions of words that were spoken), I (number of insertions of spurious words in the ASR output), and N (number of words actually spoken).*

The difficulty here, however, as Kafle and Huenerfauth point out, is that the WER measure assumes that all words are of equal value and does not consider the relative importance and predictability of words in context (2017, p. 166). In other

words, subject keywords and technical terms would be given the same weighting as prepositions, despite the greater contextual importance of the former. Kafle and Huenerfauth (2017) argue that, as the D/deaf or hard of hearing users employ a more keyword-driven reading strategy compared to the hearing population, it is not useful to consider every caption error or omission equally. Kushalnagar et al. (2014) similarly point out that the errors made during the captioning process may be more consequential in higher education than in other contexts. Simple errors in captioning – for example, numbers being misinterpreted – or slight omissions can result in changing the meaning of the content being delivered, which then impacts students' learning experience.

Others have noted that ASR accuracy tends to be measured in ideal environments which consist of single speakers with no background noise, and ideally using equipment that has been specifically chosen for its ability to recognise the speaker's voice and cadence (Lasecki, Kushalnagar, & Bigham, 2014). This, however, is quite different to a live classroom or lecture environment which can sometimes include multiple speakers, noisy backgrounds, poor acoustics, varying volumes and accents, complex and specific technical vocabularies, and multiple information sources (Kushalnagar et al., 2014; Lasecki et al., 2014). Kushalnagar et al. (2014) also point out that difficulties can even arise if the speaker's speech is altered due to a cold. In addition, it is nearly impossible for ASR to identify when a speaker may use different volumes or speeds for emphasis (Federico & Furini, 2012), or when the speaker may be incorporating visual references, such as pointing at key information on slides without actually vocalising it (Lasecki et al., 2014).

Finally, Kawas et al. (2016) point out that accuracy is only one aspect of the overall experience of using captions in higher education. They point out that it is also important to consider the overall experience of using the system interface and technology. They highlight features such as easy set-up and use, the ability to control caption displays – for example adjusting the size or position of the font –, the availability of transcripts, the ability to easily identify and troubleshoot issues, as well as device agnosticism as playing an important role in students' general experience and satisfaction with the technology. Overall, the more familiar and intuitive the interface and technology are to the student, the more likely they are to report a satisfactory experience.

Scoping study of international standards

The industry standard for caption quality is 99% or "a 1% chance of error or a leniency of 15 errors total per 1,500 words" (3Play Media, 2018). Captioning standards have been articulated by several international government, industry and advocacy organisations, including the World Federation of the Deaf, IFHOH, 3Play Media, the DCMP, the FCC, Media Access Australia / Centre for Inclusive Design, Deafness Forum of Australia, AI Media, Ofcom, W3C, the CRTC standards, and Netflix. These captioning standards and recommendations identify both general standards and those pertaining to the automation of captions.

This scoping study will firstly outline the specific definitions of caption quality and accuracy as outlined by the FCC, the DCMP and the W3C before briefly considering how these have been interpreted by industry, advocacy groups and educational providers in a variety of contents, including specific information regarding Echo360.

Definitions of caption accuracy

The FCC (2019) define quality captions as accurate, synchronous, complete and properly placed. These are expanded on further:

- Accurate: Captions must match the spoken words in the dialogue and convey background noises and other sounds to the fullest extent possible.
- Synchronous: Captions must coincide with their corresponding spoken words and sounds to the greatest extent possible and must be displayed on the screen at a speed that can be read by viewers.
- Complete: Captions must run from the beginning to the end of the program to the fullest extent possible.
- Properly placed: Captions should not block other important visual content on the screen, overlap one another or run off the edge of the video screen.

Building on these standards, the DCMP (2020) define caption quality as follows:

- Accurate: Errorless captions are the goal for each production.
- Consistent: Uniformity in style and presentation of all captioning features is crucial for viewer understanding.
- Clear: A complete textual representation of the audio, including speaker identification and non-speech information, provides clarity.
- Readable: Captions are displayed with enough time to be read completely, are in synchronisation with the audio, and are not obscured by (nor do they obscure) the visual content.
- Equal: Equal access requires that the meaning and intention of the material is completely preserved.

The W3C develops protocols and guidelines to ensure long-term growth for the web. It is the main international standards organisation for the world wide web. Their Web Content Accessibility Guidelines (WCAG) 2.0 offer guidance on how to make the web accessible for people with disabilities. WCAG 2.0 is guided by four principles, namely that it is (W3C, 2016):

- Perceivable: Information and user interface components must be presentable to users in ways they can perceive. This means that users must be able to perceive the information being presented; that is, it can't be invisible to all of their senses.
- Operable: User interface components and navigation must be operable. This means that users must be able to operate the interface; that is, the interface cannot require interaction that a user cannot perform.
- Understandable: Information and the operation of user interface must be understandable. This means that users must be able to understand the information as well as the operation of the user interface; that is, the content or operation cannot be beyond their understanding.
- Robust: Content must be robust enough that it can be interpreted reliably by a wide variety of user agents, including assistive technologies. This means that users must be able to access the content as technologies advance; that is, as technologies and user agents evolve, the content should remain accessible.

Captions are listed within the highest priority in WCAG 2.0 (Hollier, Ellis, & Kent, 2017). There are guidelines related to both live and prerecorded captions (W3C, 2008):

> *Captions (prerecorded): Captions are provided for all prerecorded audio content in synchronized media, except when the media is a media alternative for text and is clearly labeled as such.*
>
> *Captions (live): Captions are provided for all live audio content in synchronized media.*

Industry, advocacy group and educational interpretation of caption accuracy

In the UK, Ofcom offer the following recommendations around caption accuracy (OfCom, 2017):

> *…subtitle users need to be able both to watch what is going on, and to read the subtitles, so it is important that these are as accurate as possible, so that viewers do not need to guess what is meant by an inaccurate subtitle. Broadcasters should ensure that subtitles for pre-recorded programmes are reviewed for accuracy before transmission. Where live subtitling is to be provided, advance preparation is vital – where possible, any scripted material should be obtained, and special vocabulary should be prepared. The subtitling*

for repeated programmes first broadcast live should be reviewed and edited if necessary.

In the Australian context, captions must be made available on all content screened on the primary digital channels between 6am and midnight. The Australian Communications and Media Authority (n.d.) draws on The Broadcasting Services (Television Captioning) Standard 2013 to advise that captions must be:

- Easy to read.
- Easy to understand.
- Accurate.

However, no specific information regarding accuracy is mentioned. Yet businesses offering captioning services in Australia variously claim 98% (Red Bee Media) or 99% (AI Media). While Red Bee do not offer ASR, AI Media use re-speaking technology – where the captioner repeats what is heard into voice recognition software – and stenography. They also use CART – a live speech-to-text platform via audio platforms, for example phone calls, webinars, Skype etc – in which live captioners provide captions in real time to the user's device.

Advocacy group Media Access Australia (2012) have also been active in this space offering some general principles:

- All dialogue and important audio elements in a video need to be captioned.
- Captions should always be synchronised with the audio.
- Captions should be correctly spelled and punctuated.

They also have specific guidance about font size, colouring and reading speed.

However, while captioning availability – and accuracy – is mandated in the broadcast industry arena via legislations such as *The Broadcasting Services Act* 1992 and the *Disability Discrimination Act* 1992, in the Australian educational context it is not as well articulated. In the US, captioned online lectures are required under the *21st Century Video Accessibility Act*; however, in Australia, they are typically made available only by request. Following the ruling in the US that video on demand site Netflix caption its entire catalogue (Ellis, 2015), advocates for the Deaf launched legal action against American educational institutions Harvard and MIT for not offering captioned video lectures (The Associated Press, 2015).

Nevertheless, the importance of captions – as well as the use of ASR to offer these and at a suitable level of accuracy – is increasingly being recognised within the Australian higher education sector. Several apps embracing this technology, as well as services seeking to differentiate themselves within this field, have emerged, each claiming various degrees of accuracy – or accusing their

competition of being inaccurate. Table 1 outlines a number of key organisations in this space.

Table 1. Advertised accuracy of automatic captioning

Name	Description	Advertised accuracy
YouTube	These automatic captions are generated by machine learning algorithms, so the quality of the captions may vary. We encourage creators to provide professional captions first. YouTube is constantly improving its speech recognition technology. However, automatic captions might misrepresent the spoken content due to mispronunciations, accents, dialects, or background noise. You should always review automatic captions and edit any parts that haven't been properly transcribed.	Not advertised, but claims to have increased accuracy by 50% since their introduction in 2009 (Sprangler, 2017)
Otter	Generate rich notes for meetings, interviews, lectures, and other important voice conversations with Otter, your AI-powered assistant.	Does not advertise an accuracy level
Ava	Ava is an app designed to empower people who are deaf or hard of hearing by allowing them to follow conversations in real time. The app provides 24/7 real-time captioning (with up to 95% accuracy, based on artificial intelligence), on your smartphone.	95%
Transcribe – Speech-to-Text	Transcribe is your own Personal Assistant for transcribing videos and voice memos into text. Leveraging almost-instant Artificial Intelligence technologies, Transcribe provides quality, readable transcriptions with just a tap of a button.	90%
PowerPoint	PowerPoint for Office 365 can transcribe your words as you present and display them on-screen as captions in the same language you are speaking, or as subtitles translated to another language. This can help accommodate individuals in the audience who may be D/deaf or hard of hearing, or more familiar with another language, respectively.	Does not advertise an accuracy level
Rev	Rev provides a web-based captioning editor you use to capture all audible English speech, sound effects, music, and lyrics in a video file. Customers receive an easy to edit version of the caption file that can be downloaded in many forms.	99%
Scribie	AI-powered automated transcripts.	99%
Echo360	Uses Amazon Transcribe to offer speech-to-text translation	Close to 98%

Captioning online lectures using Echo360

Echo360's automatic transcription service is made possible using Amazon Transcribe, one of the many products offered by Amazon Web Services (AWS). First announced in late 2017, Amazon Transcribe is a speech recognition engine aimed at converting audio files into text (Dillet, 2017). It offers a scalable service capable of recognising multiple speakers, inserting timestamps, developing custom vocabularies, vocabulary filtering, channel identification, and automatic content redaction (AWS, 2020a). Echo360's partnership with AWS and use of the Amazon Transcribe service was announced in April 2018 (Kelly, 2018).

While Amazon Transcribe offers speech-to-text translation in up to 31 languages (Eigenbrode, 2019), the real-time streaming transcription aspect of the service (such as that used by Echo360) is currently only available in the following languages – Australian English, British English, US English, French, Canadian French and US Spanish (AWS, 2020b).

A significant advantage of Amazon Transcribe is that the audio-to-text service automatically recognises and inserts natural punctuation and formatting, meaning that speakers do not have to actively voice punctuation such as 'comma' or 'full stop' for it to be inserted into the text (Dillet, 2017; Perez, 2019). This makes it much easier to capture natural spoken language such as that commonly used in classroom and lecture settings. The transcripts are automatically generated from the audio track and made readily available for the instructor or student directly within the Echo360 player, a significant advantage over previous manual captioning services (Lynch, 2019).

In addition, the text files produced in this process are both indexable and searchable across various digital devices, providing greater flexibility in student use (AWS, 2020a; Dillet, 2017; Lynch, 2019). This also results in the creation of transcripts which students can use to further assist their studies by navigating, searching and referencing via the use of key words or phrases (AWS Public Sector Blog Team, 2018). According to Fred Singer, CEO and founder of Echo360, "It's not about the video capture itself, but the ability to use that dynamic technology to capture very complex, live, real-time interactions—moments in the classroom, you could say—and extend them" (quoted in Waters, 2019). These transcripts can also be used as study guides for the students to reference when reviewing course content for assignments or exams.

Amazon Transcribe's machine learning technology also means that the service is continually learning and improving its transcription process. While they do not claim to offer 99% accuracy in their transcripts, the transcripts are able to be further manually edited for greater accuracy. Instructors have the ability to download the video transcripts, edit them in a common text editor, and upload them back to the caption track of the associated video (Holding, 2018).

Apart from its benefits as an assistive technology, Echo360's use of Amazon Transcribe's ASR service enables a standard lecture or classroom delivery to become a more dynamic and interactive experience. Students can replay sections using the search function, which in turn provides instructors with learning analytics and data that they can use to review their course content (Lynch, 2019).

Furthermore, the ASR transcription service can be toggled on as the default at the institution, organisation or department level (Echo360, 2020b). It is also possible for the ASR transcription service to be toggled off at institution level, and overrides allowed for lower levels to switch it on as required (Echo360, 2020b). Step by step instructions on how to enable ASR transcriptions at various levels are available at https://admin.echo360.com/hc/en-us/articles/360035035512

In addition, the transcription service is triggered once the video is published, which means that an instructor can record or create a video, edit it and upload it, all before generating any transcriptions. It is only once they hit 'publish' that the transcription will be created. Transcriptions take approximately 30 minutes to complete, and are available for videos less than 4 hours long (Winfrey, 2019). Then, once the transcript is generated, instructors can download it and edit it for greater accuracy. This editing can be completed using a common word processing software such as Microsoft Word (Echo360, 2020a). Instructions on how to download transcripts are available at https://admin.echo360.com/hc/en-us/articles/360038310372 and instructions on how to edit transcripts are available at https://admin.echo360.com/hc/en-us/articles/360038310392

Once the transcript is available, the transcription panel can be toggled on and off in the viewing pane via the transcription button in the classroom toolbar (Winfrey, 2019). The transcription panel highlights segments of the text as it corresponds to the audio track, which also makes it easy to navigate to different sections of the video as needed. A search bar is available at the top of the transcription panel to enable viewers to search for specific keywords or phrases. Search results are identified as underlined text in the transcription panel, and users are able to skip forward and backward between search results (Winfrey, 2019).

Instructors also have the ability to disable the automatic transcription of videos on publication by toggling off the ASR option in the settings for their course (Winfrey, 2019). By toggling the ASR option off, the video will not have transcriptions automatically generated; this is a useful option for when manual transcripts are used.

Results of interviews

> *… every single time! I ALWAYS use captions where available, and get actively disappointed when they aren't there.*

In order to better evaluate the potential uses of captions amongst the entire student population, all Curtin University students enrolled in 11 Digital and Social Media, Screen Arts and Fine Art units offered in study periods 3 and 4 in 2018 were invited, via email, to participate in a short online interview to discuss their understandings of captions and reflect on the ways they could potentially be used in their future teaching and learning. A total of 53 students from the 22 units offered over the two study periods participated in these interviews. Some students participating in the interviews self-identified as being from at risk groups, including being hard of hearing, English as an additional language, and having sensory processing difficulties.

The interview was designed to identify current and anticipated expectations regarding captions as a pedagogical tool. Questions were grouped into four main categories – students' current usage of captions in online lectures, the potential benefits – and therefore likelihood of using captions – if they were made available in other units, their expectations regarding caption accuracy, and the impacts of in/accurate captions. For a full list of questions see Appendix 2.

Several prevalent themes and experiences regarding participants' views on the educational benefits of captioned lectures became apparent across the interviews. From the interviews, it can be seen that online students with and without disability at Curtin University state that they:

- Have diverse learning styles and that captions can be used alongside a variety of other learning tools in ways that suit their visual, auditory and kinaesthetic approaches to learning.
- Expect captions to be accurate.
- Expect captions in online lectures because they are widely available in other media.
- Multitask while accessing lectures, and therefore see captions as a way to retain focus and improve clarity.
- Consider captions as a way to provide the correct technical spelling.
- Consider that the provision of captions can assist with improving the quality of teaching in an online environment. For example, they state that lecturers continue to be unaware of the way their practices may impact on people watching lectures at a distance, and that captions could provide a way to compensate for lecturers moving around the room or not repeating questions asked by students attending in person.
- Assume that automatically generated captions are constantly improving.

- Expect the University to provide tools that support their learning.
- Hope that accurate captions will become a tool that facilitates improved comprehension of lectures – and therefore higher grades – to better manage access to educational materials.
- Are aware that they are dealing with complex visual and audio material in these online lectures. They state that they not only need to know what the lecturer is saying, but are also simultaneously trying to read the slides, make the connection between the content on the slides and what the lecturer is saying, interpret the lecturer's body language and movement, and decipher all of this in the context of the course itself.

These will be discussed below, grouped under four main categories to reflect the interview questions.

Current use of captions in online lectures

Students recruited to participate in the interviews were studying online via Open Universities Australia. Respondents reported a diversity of approaches to accessing online lectures. The majority reported accessing the lectures on a weekly basis and engaged in a stop/start approach. That is, they would pause the lecture to take in-depth notes or attend to other things, sometimes rewinding if they needed clarity of what the lecturer was saying. Many respondents reported multitasking such as simultaneously caring for children or working while accessing their lectures. The below response is indicative of the overall sentiment regarding the use of lectures amongst this cohort:

> *As I am an online student, all of my lectures are online. I sometimes view them multiple times. I will stop the lectures if required while I am taking notes, and sometimes replay sections if I have lost my focus, missed the main points or have difficulty understanding what the lecturer is saying. I usually watch/listen to the lectures in our home office which is separated from the rest of the family, however, I can be interrupted by the teenagers living in the house if they want attention.*

Nevertheless, some students did prefer to listen to lectures in a quiet place without distraction and were protective of their time spent studying. Another popular strategy was listening to the lecture at a faster speed.

A key issue identified in accessing online lectures was note taking. Students felt they were potentially disadvantaged by their distance, especially when lecturers responded to questions in the room or moved away from the podium microphone:

> *… when the lecturer turns or walks away from the microphone, the sound either drops or becomes unintelligible... or the tutor gets the class to answer a question but doesn't repeat what they say which usually means the mic can't pick it up. For lectures without captions this means you're not getting all the available information and it's infuriating! additionally, sometimes the*

caption doesn't understand a field specific term and just captions the closest approximation which is all well and good if you can make out what the tutor is actually saying, but not great for everyone else.

All my studies are online. It is frustrating when I cannot download the slides used and there is no transcription instead. I am mostly a kinaesthetic learner and to be forced to rely on audio-visual alone is disappointing.

Taking this further, there was also a keen sense amongst the group that, as online students, they were not valued in the same way as people in the room and that technology did not always compensate:

Some camera work is bad and distracting; some lecturers move away from the microphone making them difficult to understand; some lectures recorded from live sessions include group discussion portions that are not valuable for online study.

Despite this, there was a recognition too of the advantages of being able to manipulate the delivery of the lecture according to your own requirements:

Sometimes there are bits of speech which are unclear. However, I think this happens in a standard lecture room too. I love online lectures because one can stop and replay parts or the whole as often as desired.

As such, captioned lectures were recognised as a potentially powerful tool to mitigate many of the issues identified:

I would use them to help remember the material covered and take more comprehensive notes and to search for specific information.

Students were aware of their learning styles, with some describing themselves as visual learners or referring to the interaction of seeing and hearing to assist absorption of information:

I don't required captions but I do prefer captions… for some reason they help me to absorb and process the information better.

They could help a bit, as while I am able to hear, I think I process information more quickly when reading.

100% likely [to use captions], I absorb more information when reading.

Students that we might describe as both visual or kinaesthetic described the way they would print the transcript and highlight key points or make additional notes:

Online captions could help me with note taking as I am more of a visible learner.

I would use them to help remember the material covered and take more comprehensive notes and to search for specific information.

The potential for this search functionality was mentioned several times with reference to revision, note taking and clarifying information. Captions were seen as potentially providing greater clarity when lecturers had accents or moved around the theatre:

> *I don't need captions but they are good to have since lecturers have different accents.*

Whereas students who are D/deaf or hard of hearing, the original audience for captions, might only use the caption track, for this cohort everyone indicated that would use both tracks. For some this was to compensate for the technical difficulties experienced in lecture recordings:

> *If captions were available I may use both. I might only switch the captions on when bits of the audio had poor sound for some reason.*
>
> *Yes, I can read whilst the buffering is working itself out.*

While others again referred to their learning style:

> *Both – I am fortunate to have a choice, in that I do not have a hearing or sight impairment, but the captions offer some clarity, but I learn better with a combination of visual and audial cues.*

Accuracy was highlighted as a key issue:

> *If the captions accurately state what the lecturer says, then I would just use the captions.*
>
> *If captions are quite inaccurate I am unlikely to use them when watching a lecture presented by someone whose first language is English, but I would probably still use them when watching a lecture presenting an especially complex idea or presented by someone with a thick accent.*

Students again referred to being diverse or non-traditional students in their responses, having to study whenever, wherever they could:

> *Captions over sounds. With sound/audio, some accents can be hard to distinguish words. Some lecturers have monotone voices and can make a subject quite uninspiring. Audio is harder to use at night while husband is sleeping and I want to study.*

While all believed they should have access to captions as an option, there was variety in how these would be used. For some, the caption track alone was the preferred method of access, particularly if a transcript could be printed out while, for others, captions and sound together was the preferred option.

Potential benefits of captions in online lectures

A number of potential benefits were identified, from accessibility for students who are D/deaf and hard of hearing, to students accessing the lecture in a noisy environment, to increasing clarity and reducing the need to stop/start the lecture:

> *The captions may pick up words that I'm unable to decipher. If the captions were available as a downloadable document after the lecture, that would be fantastic. It means I could watch and listen through the entire lecture, without having to stop and start.*

While this comment was made by a student who identified as having a mild hearing impairment, others again emphasised the functionality of captions for students with diverse learning styles:

> *… beyond the standard inclusivity and not making people feel like an inconvenience for requesting what they need? It will provide people with different learning methods with more options.*

In addition to lecturers' accents, which was raised several times throughout the 53 interviews, students recognised the potential for revision:

> *Improved quality of assessment work and a better understanding of requirements.*

Captions also improved comprehension by communicating the correct technical spelling of some words, a valuable resource for distance students:

> *… easier to understand information and improved spelling and knowledge of unit/subject specific terms. I don't always get the pronunciation right when I read them.*

Expectations regarding caption accuracy in online lectures

Accurate captions were expected amongst the cohort. The issue of grammatical or spelling errors yielded some disagreement. Some students were staunch in their belief that captions should be 100% accurate:

> *I think that the accuracy is the most important thing; including things like 'their' and 'they're'. This means that the lecturer has to be very clear in the way they speak and probably not use contractions and other speech shortcuts, to maintain clarity. Good pronunciation is vital. It is very frustrating listening to educated people saying words like 'imporDant' instead of 'imporTant'. This is very imporTant for ESL students.*

At the same time, other students felt they could manage with some minor errors:

> *I expect captions to be fairly accurate. Grammatical or spelling errors are ok, but using the incorrect word when it sounds similar isn't.*

During this part of the interview, students reflected on the use and accuracy of captions in other media and the potential for speech-to-text technology to constantly improve. However, the service was expected to be of equal quality to a professional stenographer:

> *I would expect 98% to 100% accuracy – what you would expect from any stenography/caption service.*

The role of the lecturer and their delivery was again mentioned as a contributing factor to caption accuracy:

> *I think the quality may be affected by the lecturer's accent and clarity of speech as well as the software use to capture the words.*

This cohort of students were remarkably aware of the variety of learning tools at their disposal and viewed captions as a tool within this available arsenal:

> *I will utilise whatever tools are in place; I have a choice.*

There was a firm belief amongst some that caption availability improved their marks and reduced the need to stop and start the lecture to get clarity regarding what the lecture had said. This in turn had a positive time management effect with less perceived wasted time.

While a number of students participating in the interviews were keen to point out that they did not have any type of hearing impairment but that they would nevertheless use captions as a learning tool suggests that captions are a key feature of UDL. Similarly, these students believed their disabled peers had a right to accessibility:

> *Accessibility should be the default – we should offer captions whether the majority of students are hearing or deaf.*

Impacts of in/accurate captions in online lectures

In the final part of the online interview, students were presented with two short sample clips of online lectures. They were asked to view a clip, give it a rating out of five stars, and then explain why they provided that rating. A total of 27 students responded in this part of the interview.

The sample clips were selected on the basis of demonstrating two different levels of captioning quality. Sample A was drawn from the unit NET102/NETS1002 Digital Culture and Everyday Life, and Sample B was from the unit WEB207/NETS2007 Web Media. Sample A was delivered by a female lecturer and Sample B by a male lecturer. Both clips were similar in length, 1:38 minutes and 1:41 minutes respectively.

Sample A was selected as an example of good quality captioning with few inaccuracies, errors or lag between audio and text. Sample B was selected as a

sample of poorer quality captioning as it contained a number of errors, including some lag between captions and audio, misinterpreted words, a portion labelled "inaudible" in the captions, and a few omissions. Students were not made aware of the reasons behind the choice of sample clips, nor were the differences in quality communicated to them beforehand.

Despite the difference in the quality of the clips, they received a mix of ratings, as summarised in Table 2 below.

Table 2. Rating of the two sample online lecture clips

Rating	**Sample A**		**Sample B**	
	No. of respondents	**% of total**	**No. of respondents**	**% of total**
1 star	2	7.41%	1	3.70%
2 stars	0	0.00%	5	18.52%
3 stars	4	14.81%	7	25.93%
4 stars	13	48.15%	9	33.33%
5 stars	8	29.63%	5	18.52%

It can be seen that Sample A largely received good ratings, with just over three quarters of respondents rating the clip 4 stars or above. Sample B received a greater mix of ratings, but still had just over half the respondents rate the clip 4 stars or above. Respondents that rated Sample A highly generally gave Sample B an equal or lower rating, further reflecting the difference in the quality of the clips. However, this wasn't always the case and there were a minority who ranked Sample B higher than Sample A.

However, importantly, it should be noted that, from the qualitative feedback, it can be seen that the reasons for these ratings were often not related to the captions themselves, but rather the overall lecture experience. For example, the following ratings and comments are from the same student:

> *Not sure whether the captioning could be downloaded as a transcript [Sample A: 4 stars].*
>
> *Prefer option B over option A. Lecturer is easier to see, therefore can tell they are more engaged. I would be more inclined to watch a lecture as opposed to wanting to download a transcript [Sample B: 5 stars].*

While the qualitative feedback on the clips covered a range of issues, a number of respondents did comment on the captions, including their presentation and placement. The only respondent who rated both clips as 1 star criticised the size of the captions:

> *The caption text was way too small [Sample A].*
>
> *Again the caption text is way too small [Sample B].*

Others clearly referenced standards of expectations formed from familiarity with other captioned media, commenting on both the placement of the captions and the competition with other elements on screen:

> *I found it a bit distracting at first because I think the speech needs to be directly under the lecturer. I am still wanting to also observe the lecturer and I think the text needs to be larger so it is like subtitles in a movie.*
>
> *… the text was a little inaccurate and too small as well. The slide was really busy and this distracted me from the captions*
>
> *It was clearer throughout, no breaking up but it was still quite small. I also found that the caption fights with the video of the lecturer and the slides so making it larger would help. It also needs to keep up with the lecturer, it lagged a bit.*

Others commented on the inaccuracies they noticed in the captions, suggesting that the students themselves have a high standard of expectations when it comes to captioned lecture material:

> *The captions were well done and appeared to be accurate (it was good that the ums and repeated words were not included in the captions). There was a little of the volume drop when the lecturer turned away from the front of the room.*
>
> *Great not putting in the ums and ahs and you know, got all the information across well. There was an error where it read district but it was distinct that was spoken. The sentences seemed to run on, with some punctuation missed out. If this was the standard I could have, I would be happy.*
>
> *The captions were good, the only errors I could see were in punctuation and occasionally spelling but they were adequate in allowing the observer to understand what was being discussed.*
>
> *One obvious missed word "district" instead of "distinct" but mostly excellent, including avoiding the lecturer's stumbles, making it a more natural read.*
>
> *The captions were mostly accurate, however some words were incorrectly captioned. When what the lecturer said was inaudible, the captions said that, which I prefer to the captions trying to come up with something and making a mistake.*
>
> *The inaudible bit was actually audible for me, but since the lecturer couldn't remember the name, was kind of irrelevant. Some of the words were wrong, and some words were missed, that made what was on the screen not make sense. Overall not too bad, if this was what I had I would still use the captions.*
>
> *I gave this ranking because there were some areas not captioned – the lecturer is speaking very quickly so not everything could be captured, but there were also issues with punctuation and spelling.*

A harsh mark, but the quality here was not as good as the previous video, with many missed words, single word captions that should have been included in the previous caption, generally the quality was poorer.

Some students picked up on inaccurate translations of key terms that could impact the learning material:

Generally it was pretty accurate. However I noticed the word "distinct" was translated as "district" and in some contexts this kind of inaccuracy could lead to serious problems.

Many errors and omissions eg. "bespoke", "existing", "other spaces", "dramatised journalism", "such good quality".

Others noted the lack of sync between captioned text and audio:

… couple of errors in the text... was behind what the lecturer was saying and then in front so it was strange.

The timing is all off and there's parts where it doesn't even try.

Despite the criticisms of the samples as noted above, the general tone of the comments was largely positive, with students reflecting a wide range of reasons for why they found the captions useful:

Because my first language is not English … I think it is very important to have it.

Enforces the message. Helps with comprehending the lecture's message.

The caption allows me to follow the lecturer's words as they are spoken and this is much clearer to me. It also make the lecture available if I cannot hear the lecturer.

… it was easier to follow as he was moving around so was able to read parts that missed or double take and understand fully as using brain in two ways listing and reading.

The lecturer spoke fast and walked around a lot, but the captions picked up on most of what he said. This made them very useful as it would be hard to keep up with him without them.

As mentioned above, the range of comments and issues highlighted in the qualitative feedback on the sample clips demonstrates that students have a range of expectations when it comes to online lectures in general. The feedback on the captions not only referred to accuracy, but also their size, placement, and synchronisation with the audio. However, students were not commenting on the captions alone, but also the visibility of the lecturer and the slides, the presentation of the slides, and the lecturer's style of speaking:

Prefer option B over option A. Lecturer is easier to see, therefore can tell they are more engaged. I would be more inclined to watch a lecture as opposed to wanting to download a transcript.

I have 'attended' lecturers with this professor and he is very engaging. I did hear the 'inaudible' section, but I can understand that not everyone might be able to hear it. Training for the lecturers in how to get the best and consistent results for recording would probably be beneficial. Lapel mics and training would probably be best. You cannot ask a lecturer who is used to walking around to stay static – I teach adults in pre-accredited classes and I wander around (small classroom) and often lose my train of thought if I am static for too long, so I understand the need to 'move'.

… these captions did not match what was be said and it made them harder to follow. The lecturer will need to modify the way he speaks to fit in the captions as well.

These comments highlight that the students are aware that they are dealing with complex visual and audio material in these online lectures. They understand that they not only need to know what the lecturer is saying, but are also simultaneously trying to read the slides, make the connection between the content on the slides and what the lecturer is saying, interpret the lecturer's body language and movement, and decipher all of this in the context of the course itself. Greater accuracy in captions will go a long way towards smoothing the process of viewing online lectures, but will need to be recognised as but one element in the broader experience of online lectures overall.

Conclusions

This report has detailed student expectations regarding caption accuracy as a tool for teaching and learning focusing on Echo360's automatic captions. The report also considered the Global standards regarding caption accuracy in both entertainment and online learning as well as students' engagement with immersive learning strategies made available via online captions.

The report began by outlining the historical use of captions in both entertainment and education contexts focusing of the changing definition of quality. Advances in machine learning and increasing use of captions in a social media context have established an expectation that captions are readily available and accurate.

Part I of the report offered a literature review addressing two key areas of research. First, the potential benefits of captioned lectures for the broader student population and, second, the use of ASR in making these captions available, with a particular focus on their accuracy. Captions, while initially intended for people who are D/deaf and hard of hearing, were then seen to have benefits from students from other at risk groups, including those with English as an additional language and other disability groups, and are now also embraced as a key resource for the mainstream student population via theories of UDL.

Part 2 presents a scoping study of international standards regarding caption accuracy. A significant number of international government, industry and advocacy organisations have articulated captioning standards and recommendations, both in general and pertaining to their automation. This includes the World Federation of the Deaf, the International Federation of Hard of Hearing People (IFHOH), 3Play Media, the Described and Captioned Media Program (DCMP), the Federal Communications Commission (FCC), Media Access Australia / Centre for Inclusive Design, Deafness Forum of Australia, AI Media, the UK government's Office of Communications (Ofcom), the World Wide Web Consortium (W3C), the Canadian Radio–Television and Telecommunications Commission (CRTC) standards, and Netflix. However, available industry standards regarding caption accuracy standards for online lectures remains unclear. While the preferred accuracy rate of 99% is often cited, this figure relates only to US legislation; in the Australian context, 95% is often cited as the international standard. This section covers some key definitions of what accuracy means in relation to captioning, as well as industry, advocacy group and educational interpretation of their use, including specific information regarding captioning on Echo360.

Part 3 of the report is concerned with the findings of, and offers discussion on, the results of the interview stage of this research with the 53 students who participated in the project. These students were enrolled in 11 Digital and Digital and Social Media, Screen Arts and Fine Art units which trialled mainstreaming captions in two study periods in 2018, resulting in data from 22 units in total. Insights were gained into both how students actually used the captions and how they anticipated using

them in the future should Curtin University embrace them as a mainstream approach.

Captions are a key example of universal design for learning. While they are a vital accessibility feature for students with disability or from at risk groups, they offer flexibility and support for the entire student cohort. Increasingly students, particularly those studying online are aware of their learning needs and the importance of accessibility for students with disability. While the students interviewed for this research integrated captions into their distance education strategies they were adamant that students with disability be given the support they require. Echo360 automatic captioning was warmly received by the cohort of students interviewed.

Recommendations

- It is clear from this research that students both like and expect captions in a Higher Education setting. Automated captions provide a cost effective alternative to traditional captioning and should be turned on by default.

- Further study is needed into the impact of different error rates on the effectiveness of captions for student learning, and what can be considered effective.

- Further research is needed on the impact and use of captions by specific user groups, including the broader student population, those with English as a further language, students who are Deaf or hard of hearing, and those with learning disabilities.

- Captions need to be used in conjunction with training for presenters to make the best use of automated systems including appropriate use of audio recording systems, protocols for including questions and comments from people in presentations who are not captured by recording equipment and an understanding of the requirements of an audience that is not present in a lecture theatre or classroom.

Authors

Katie Ellis

Katie Ellis is Professor in Internet Studies and Director of the Centre for Culture and Technology at Curtin University. Her research is located at the intersection of media access and representation and engages with government, industry and community to ensure actual benefits for real people with disability. She has authored and edited 17 books and numerous articles on the topic of disability and the media, including most recently the monograph *Disability and Digital Television Cultures* (Routledge, 2019).

Kai-Ti Kao

Kai-Ti Kao is a Research Assistant and PhD candidate with the Centre for Culture and Technology at Curtin University. Her research interests lie in social engagement with digital media, particularly in relation to power, representation and inequality. She has previously published on a range of these topics including policy framing of Information and Communication Technology for Development (ICT4D), digital engagement and mental health, and issues of representation in the popular videogame *Overwatch.* Her current research focuses on investigating the collaborative learning experiences for students with disabilities in higher education.

Mike Kent

Mike Kent is a Professor in the Centre for Culture and Technology at Curtin University. His research focuses on the intersecting areas of disability, eLearning and social media. His most recent publications include the two volume collection on the future of critical disability studies with Katie Ellis, Rosemarie Garland-Thomson and Rachel Robertson *Manifestos for the Future of Critical Disability Studies* (Routledge, 2019) and *Interdisciplinary Approaches to Disability: Looking Towards the Future* (Routledge, 2019).

Appendix 1

List of units that offered Echo360 captions in study periods 3 and 4 2018

APC 100	Academic and Professional Communications
COM 155	Culture to Cultures
HIS 513	Democracy and Dictatorship
NET 102	Digital Culture and Everyday Life
WEB 1010	Web Communications
WEB 207	Web Media
VAR 100	Art and Design Fundamentals
VAR 11	Visual Arts Research: Introduction to Drawing
VIS 330	Perspectives and Beauty in Art
SCA 210	Reading Screens
SCA 310	Thinking Screen Cultures

Appendix 2

1. Please describe how you currently use online lectures. Some things you could include are: frequency of use (all units, some units, weekly, fortnightly); viewing pattern (view once all the way through, repeated viewings, frequent stops or replaying); usual viewing environment (noisy/quiet, distracted/multitasking, interrupted/uninterrupted viewing).

2. What would you say is currently the greatest challenge of using online lectures?

3. How likely would you be to use captions if they were made available on online lectures? Would captions help address the challenge(s) identified previously?

4. Are you likely to use both sound and captions, or just one or the other? Why?

5. What do you think will be the potential benefits of captioned lectures?

6. What are your expectations of the quality of the captions (e.g. accuracy)?

7. How is meeting these expectations likely to impact on your study needs or requirements?

8. What is your likelihood of using captions if these expectations are not met?

9. Caption Lecture A: Please watch the following video and rate the quality of the captions by providing a ranking out of 5 stars (located below the video). You will then have the opportunity to elaborate on your rating.

10. Could you please elaborate on why you gave the Caption Lecture A video this ranking?

11. Caption Lecture B: Please watch the following video and rate the quality of the captions by providing a ranking out of 5 stars (located below the video). You will then have the opportunity to elaborate on your rating.

12. Could you please elaborate on why you gave the Caption Lecture B video this ranking?

References

3Play Media. (2018). What Is 99% Accuracy, Really? Why Caption Quality Matters. Retrieved from https://www.3playmedia.com/2018/06/06/caption-quality/

AWS. (2020a). *Amazon Transcribe – Automatic Speech Recognition*. Amazon Web Services. Retrieved from https://aws.amazon.com/transcribe/

AWS. (2020b). *Transcribing Streaming Audio—Amazon Transcribe*. Amazon Web Services. Retrieved from https://docs.aws.amazon.com/transcribe/latest/dg/how-streaming-transcription.html

AWS Public Sector Blog Team. (2018, June 28). *Echo360 Brings the Power of Machine Learning to the Classroom with Amazon Transcribe*. Amazon Web Services. Retrieved from https://aws.amazon.com/blogs/publicsector/echo360-brings-the-power-of-machine-learning-to-the-classroom-with-amazon-transcribe/

DCMP (2020). "Quality Captioning." Retrieved from http://www.captioningkey.org/quality_captioning.html.

Dillet, R. (2017, November 30). Amazon Transcribe is a sophisticated transcription service for AWS. *TechCrunch*. Retrieved from http://social.techcrunch.com/2017/11/29/amazon-transcribe-is-a-sophisticated-transcribing-service-for-aws/

Downey, G. (2007). Constructing closed-captioning in the public interest: from minority media accessibility to mainstream educational technology. *Info : the Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media, 9*(2/3), 69-82. doi:http://dx.doi.org/10.1108/14636690710734670

Ellis, K. (2015). Netflix Closed Captions Offer an Accessible Model for the Streaming Video Industry, But What about Audio Description? *Communication, Politics & Culture, 47*(3).

Eigenbrode, S. (2019, November 25). *Amazon Transcribe now supports speech-to-text in 31 languages*. Amazon Web Services. Retrieved from https://aws.amazon.com/blogs/machine-learning/amazon-transcribe-now-supports-speech-to-text-in-31-languages/

FCC (2019, 31 December). "Closed Captioning on Television." Retrieved from https://www.fcc.gov/consumers/guides/closed-captioning-television.

Griffin, E. (2015). Who uses Captions? Not just the deaf or hard of hearing. Retrieved from http://www.3playmedia.com/2015/08/28/who-uses-closed-captions-not-just-the-deaf-or-hard-of-hearing/

Hollier, S., Ellis, K., & Kent, M. (2017). User-Generated Captions: From Hackers, to the Disability Digerati, to Fansubbers. *Media Culture, 20*(3), http://journal.media-culture.org.au/index.php/mcjournal/article/view/1259.

Kelly, R. (2018, April 9). *Echo360 Integrates Amazon Transcribe Automated Speech Recognition*. Campus Technology. Retrieved from https://campustechnology.com/articles/2018/04/09/echo360-integrates-amazon-transcribe-automated-speech-recognition.aspx

Kent, M., Ellis, K., Peaty, G., Latter, N., & Locke, K. (2017). *Mainstreaming Captions for Online Lectures in Higher Education in Australia: Alternative approaches to engaging with video content at Curtin University*. Perth, Western Australia, National Centre for Student Equity. Retrieved from https://www.ncsehe.edu.au/publications/4074/?doing_wp_cron=1493183232.7519669532775878906250

Klie, L. (2010). YouTube Expands Video Transcription Option for All. In (Vol. 15, pp. 11). Medford.

Lynch, M. (2019, August 12). Automated Speech Recognition and the Future of Studying. *The Tech Edvocate*. Retrieved from https://www.thetechedvocate.org/automated-speech-recognition-and-the-future-of-studying/

Media Access Australia (2012). "Captioning Guidelines." Retrieved from https://mediaaccess.org.au/practical-web-accessibility/media/caption-guidelines.

OfCom. (2017). Ofcom's Code on Television Access Services. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0020/97040/Access-service-code-Jan-2017.pdf

Parton, B. S. (2016). Video Captions for Online Courses: Do YouTube's Auto-Generated Captions Meet Deaf Students' Needs? *Journal of Open, Flexible and Distance Learning, 20*(1), 8-18.

Perez, S. (2019, December 2). Amazon debuts automatic speech recognition service, Amazon Transcribe Medical. *TechCrunch*. Retrieved from http://social.techcrunch.com/2019/12/02/amazon-debuts-automatic-speech-recognition-service-amazon-transcribe-medical/

Perkins, R. (1971, December 14-16). Proceedings of the First National Conference on Television for the Hearing-Impaired. Retrieved from http://files.eric.ed.gov/fulltext/ED064828.pdf

Pitman, T., Ellis, K., Kent, M., & Mancini, V. (2020). *How Higher Education Students with Disability Engage with Digital Technologies*.

Romero-Fresco, P., & Perez, J. M. (2016). Accuracy rate in Live Subtitling: The NER model. In Jorge Díaz Cintas & R. B. Piñero (Eds.), *Audiovisual Translation in a Global Context: Mapping an Ever-changing Landscape* (pp. 13-28). London: Palgrave Macmillan.

Sprangler, T. (2017, 16 February). YouTube Has 1 Billion Videos With Closed-Captioning, but Not All of Them Are Accurate. Retrieved from https://variety.com/2017/digital/news/youtube-1-billion-videos-closed-captioning-accuracy-1201990083/

The Associated Press. (2015, February 12). Harvard, MIT sued over lack of closed captioning online. Retrieved from http://phys.org/news/2015-02-harvard-mit-sued-lack-captioning.html

Tobin, T. J., & Behling, K. T. (2018). *Reach Everyone, Teach Everyone: Universal Disign for Learning in Higher Educatio0n*. Morgantown: West Virginia University Press.

W3C. (2008, 11 December). Web Content Accessibility Guidelines (WCAG) 2.0. Retrieved from https://www.w3.org/TR/WCAG20/

W3C. (2016). Understanding WCAG 2.0. Retrieved from https://www.w3.org/TR/UNDERSTANDING-WCAG20/intro.html#introduction-fourprincs-head

Wactlar, H. D., Kanade, T., Smith, M. A., & Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *Computer*, *29*(5), 46–52. https://doi.org/10.1109/2.493456

Waters, J. (2019, January 7). *How Automated Speech Recognition Could Change Studying Forever*. EdSurge. Retrieved from https://www.edsurge.com/news/2019-01-07-how-automated-speech-recognition-could-change-studying-forever

Wheatly, M. G., Flach, J., Shingledecker, C., & Golshani, F. (2010). Delivering on the promise of Plato's academy: Educational accessibility for the 21st century. *Disability & Rehabilitation: Assistive Technology*, *5*(2), 79–82. https://doi.org/10.3109/17483100903387176

WFD & IFHOH (2019) WFD and IFHOH Joint Statement: Automatic Speech Recognition in Telephone Relay Services and in Captioning Services. Retrieved from https://wfdeaf.org/wp-content/uploads/2019/04/WFD-IFHOH-Joint-Statement-on-ARS-2019-Final.pdf

Winfrey, B. (2019). Enable or disable automatic transcriptioning for a section (ASR). Echo360: Teaching & Learning. Retrieved from http://learn.echo360.com/hc/en-us/articles/360035407651-Enable-or-Disable-Automatic-Transcriptioning-for-a-Section-ASR-