ASCILITE
SHAPING THE FUTURE OF TERTIARY EDUCATION
Show Me The Learning
Adelaide, Nov 27-30, 2016

Voice-to-Text Transcription of Lecture Recordings

Stuart Dinmore
Teaching Innovation Unit
University of South Australia

Jing Gao
School of Information Technology and Mathematical Sciences
University of South Australia

Educational institutions are increasingly recognising the potential value for students that same-language-subtitles can bring to lecture recordings and other digital content. During 2016 the University of South Australia's Teaching Innovation Unit and School of Information Technology and Mathematical Sciences collaborated on a project which aimed to test our ability transcribe every piece of digital video content hosted by the University in to same-language subtitles in a cost effective way. We believe this augmentation to our existing media content would have various benefits for our students. This paper discusses the benefits of same-language transcription of media content and goes on to outline the details of a technical feasibility study.

Subtitles, Transcription, Universal Design for Learning (UDL), Same Language Subtitles (SLS)

Introduction

During the past decade, use of multimedia for teaching, particularly digital video, has become extremely widespread in higher education. This is driven in large part by the cultural shift in education towards digital and blended learning models, exemplified particularly by the flipped classroom, but also by access to affordable digital technology, faster internet speeds and a rise in digital production skill sets. As part of this shift many universities are also refitting traditional lecture theatres and building collaborative teaching spaces to augment these types of pedagogical approaches. In short, pedagogical models in higher education are changing and digital video plays a significant part in this change.

Digital video can also play a key role in increasing access for students with varying abilities. For example, the flexible modes of consumption of digital video mean that students can access their lecture or course content at a time which best suits them, they can slow down and speed it up as they need, can review material for revision or access content specifically tied to an assessment. Students usually prefer video over audio only solutions as digital video can provide richer content for learning. Provision of these options for students conforms to one of the 3 key principles of Universal Design for Learning (UDL). Principle 1 states that course content must have 'multiple means of representation' (CAST, 2011). This means that students must be able to access similar material through multiple means, thus levelling the playing field for all students. Adding same-language subtitles (SLS) is an effective way of achieving this, with numerous studies, outlined in Gernsbacher (2015),demonstrating the benefits of adding captions or subtitles to video.

The process of creating fully-automated transcriptions represents a significant technological challenge for an institution, particularly a large university that creates 100's of hours of content each week. What follows is a brief outline of the benefits of subtitling for students and the details of a feasibility study conducted at the University of South Australia to assess the current capability to provide fully-automated SLS for all our video content.

Pedagogical benefits of subtitles and transcription

The benefits of incorporating digital video in to your course content are many, and have been documented elsewhere (Kuomi 2014, Woolfitt 2015). The addition of SLS with this digital video content potentially creates an environment for students which can greatly increase not only comprehension and engagement but equitable access for a range of students with varying abilities and needs. Some of the benefits include:

1. **Increased accessibility for deaf or hard of hearing viewers** – Perhaps the most obvious advantage of subtitles or captions is their use by those with hearing difficulties (Wald 2006, Burnham et al 2008, and Stinson 2009). The advantage of subtitles for those with hearing problems is clear, but it can also mean that video becomes more accessible for all students in sound sensitive situations.

2. **Improves comprehension for all students** – SLS can have a powerful impact on comprehension for all students, Steinfeld (1998), Kothari (2008), Brasel and Gips (2014). Providing this kind of access for students is an excellent example of UDL principle 1 (Provide multiple means of representation): it can enable the curriculum for all students, not just those with disabilities. 'Multiple studies have shown that the same options that allow students with physical and sensory disabilities to access materials, specifically captioning and video description, also provide educational benefits for students with other disabilities, English language learners, and general education students.' (Sapp, 2009, p. 496)
3. **Translation into foreign languages** – As higher education becomes increasingly globalised with many courses available internationally the need to provide means of comprehension for students from a variety of language backgrounds is crucial (Kruger, Hefer & Matthew 2014). For example, 25% of the University of South Australia's internal cohort are international students and the ability for those students to easily translate course content in to various languages can aid comprehension.
4. **Enhances foreign language Learning** – Multiple studies, such as Zanón (2005), Etamadi (2012), Vanderplank (2013), and Mohsen (2015) have outlined the effectiveness of SLS for students learning a new language. This is because they influence factors like pronunciation, context, speed, reading skill, understanding colloquialisms and aid with rapid word recognition.

There are other positive aspects of SLS which may apply to the general student cohort. For example, this more flexible style of delivery aid to personalised learning – students are able work at their own pace and blend the time and place of their learning. Subtitles and transcripts can also help make content searchable, so students can locate the relevant information among an enormous amount of material.

Outline of Research Project – Design and Methodology

In order to test our capability to create fully-automated subtitles for all our digital video content we conducted a feasibility study. We used an automated process to transcribe sample videos housed on our dedicated media server. What we aimed to test was the accuracy vs. cost of using automated voice-to-text generators, given that a very high level of accuracy is essential in higher education due to the use of technical and discipline specific language. A number of experiments were designed to answer the following research question: *Can automated speech-recognition provide acceptable results for lecture recordings?*
In total, 30 recordings from the university media library were used, ranging from 2 minutes (a short welcome message) to 2 hours (a standard university lecture). Four key areas were considered during data collection:

1. **Discipline area:** covering IT, law, management, engineering and health science.
2. **Single voice vs multiple voices:** Covering sole speaker and multiple speakers (seminars and workshops).
3. **English speakers from different native language background:** covering British, Chinese and Indian.
4. **With and without background noise:** Covering recordings from the lecture theatres, individual offices and classrooms.

Three engines were utilised to perform speech recognition on sample videos (used with default settings, no training required).
1. **Google Speech-to-Text:** Industry leading, available as beta-testing to selected users only.
2. **IBM Bluemix Speech-to-Text:** Industry leading, available commercially to the public (enhanced and cloud version or Dragon Naturally Speaking).
3. **CMUSphinx:** leading open source solution, developed at Carnegie Mellon, free to the public.

Unlike CMUSphinx (an offline solution), both Google and IBM engines are cloud-based and require audio data to be sent as chunks (e.g. 60 seconds per chunk). This project has also considered the potential recognition results differences between short and large chunks (large chunks contain more context so the accuracy is potentially improved).

Due to a very high volume of recordings (approximately 250,000 hours recording per annum at the authors' university) and the varying background of lecturers there are other requirements, outlined below, when adopting a speech-recognition system and preparing the audio for transcription:

1. **Speaker-independent:** It is time-consuming and nearly impossible to create training voice data sets for individual speakers. Although the training model of the Dragon Naturally Speaking software can be exported and re-used on a different machine, due to the recording hardware (different microphones) and background noise, the applicability of the training model significantly degrades.
2. **Context-specific:** University lectures often have discipline-specific terminologies. The speech-recognition engine should be capable of identifying terminology based on the content discipline.
3. **Big Data friendly:** University recordings are managed by a central server. In a large scale deployment (to take a large volume of recordings), recognition cannot be done on individual lecturer's computers, a fully automated server environment is essential.

4. **Usable results:** In addition to the expected level of accuracy, the results need to be provided to the end-users (both staff and students) in a way that makes viewing recordings more effective and efficient.
5. **Minimum human intervention:** Human transcribing and editing is expensive and time-consuming.

Further to the considerations outlined above, a typical speech recognition process includes four elements:

1. **Core engine:** Process the input audio files and match the dictionary words base on the statistical models specified in the language model and acoustic model;
2. **Pronouncing Dictionary:** It can map from words to their pronunciations (e.g. en-US, in the ARPAbet phoneme set, a standard for English pronunciation).
3. **Language model:** A simple one may contain a small set of keywords (e.g. used in automated phone answering machine) and the grammar of the language. The other variant, statistical language models, describe more complex language. They contain probabilities of the words and word combinations. Those probabilities are estimated from a sample data; and
4. **Acoustic model:** This is a statistical model as a result of a large set of training data which are carefully optimised to achieve best recognition performance (e.g. adapter to a certain accent and recording environment).

In addition to these core components, there are other components which are designed to further improve recognition accuracy. For example, speaker dependent speech recognition software (e.g. Dragon Naturally Speaker) includes a software component to build the acoustic model from the speaker's voice (which is often referred as the 'training' process). Many cloud-based engines will use the recognised keywords to search for the possible context. Once the context is identified, a more relevant language model will be used instead of the generic one. Additionally, advanced engines such as Google speech-to-text API have built-in prediction algorithms (searching the database for similar results base on the recognised keywords). Taking in to consideration this wide range of factors and variables the researchers were confident of comprehensive and nuanced results in response to the research question.

Feasibility Study Results

In terms of the way the results are expressed it is worth noting the difference between user perceived accuracy and the machine confidence indicators. Both Google and IBM engines provide a confidence indicator (1 as the highest value) for the recognition results. The researchers read the text scripts while listening to the original audio for personal judgement. It was noted that over 30 recordings, the average accuracy confidence exceeded 0.70 (the highest one being 0.981). By listening to the audio, it can be determined that the machine confidence indicators are an underestimation of overall accuracy. The actual level of accuracy is significantly higher. For example:

> **[TRANSCRIPT]** seek out an activity or resource scroll to the bottom of the page and click the URL [resource] please add into the name of the URL resource and description open the video confirmation email highlight the link right click and copy the link close the email scroll down and paste the link into the external URL field. [confidence: 0.8921]
>
> **[TRANSCRIPT]** scroll to the bottom of the page and click save and return to course click the video resource link and hit the video. [confidence: 0.9311]

The above two transcripts actually matched every single word in the original audio, yet the confidence indicators do not reach 1. This finding is consistent across all sample data. Generally speaking, user perceived accuracy is higher than the machine generated confidence indicators.

Another area of difficulty from the transcribed results relates to the issue of readability. The transcripts may have a relatively high level of word recognition accuracy but their readability is low. For example:

> **[TRANSCRIPT]** if I wish to use the Today Show two of you to manage an appointment so I can click through on appointment time and that I showed you here I get a summary of all appointments for a particular guy I can talk with you today by using the left and right arrow or using the calendar drop down I can arrange to buy stuff. [confidence: 0.7954]
>
> **[TRANSCRIPT]** information about who the appointments with and if this unit but did you see here that older white coloured Apartments off of Mormons and the purple coloured appointments are booked appointments if I have the mass of it as appointments so I can get more details about the petition appointment down the bottom here we have a link for college and our country that gives a summary of all the symbols. [confidence: 0.7577]

The above two scripts represent the general accuracy level. Although it is possible to read through the scripts by themselves, it does not provide a pleasant reading experience for various reasons:

1. **Lack of punctuation:** Nearly all recognition engines skip punctuations if the speaker does not explicitly specify (e.g. say "Full Stop"). It seems that the Google and IBM engines will occasionally put one or two while still missing the majority. During the lectures, it is simply not practical for lecturers to say the punctuations. Without full stops, the scripts become difficult to read.
2. **Lack of grammar:** During speaking, speakers tend to focus less on grammar and the completeness of sentences. There is also tendency for speakers to repeat words that they think are important. However, while reading, without grammar and sentence structures, the reading experience is further reduced.
3. **Missing words:** Different speech recognition engines have different ways of dealing with mismatches or a complete miss. The Google engine seems to ignore the words if the quality of match is low. On the other hand, the IBM engine always tries to give some results even if it is not entirely accurate (as underlined below). Unfortunately, neither approach makes reading any easier. For example:

> **[Google]** recording is protected by copyright know. Maybe we produced without the prior permission of the University of South Australia. free moving into the final of the course which is all that digital communication is going to be looking at pricing for websites like designing and driving traffic to website was going to be looking at social media in the next few weeks sorry I'm sure you're very familiar with social media.
>
> **[IBM]** This recording is protected by copyright no part may be reproduced without the prior permission at the university of South Australia. I'm. Your. Hello everyone welcome. I have to excuse me tonight if I sound a bit nicely up quite a bit of a shocker called happening side has taken it's a bad bad I'm yeah obviously just if I sound a bit strange that's why tonight so you'll have to excuse me for that. So this week we.

From the above examples, both sourced from the same recording it is evident that currently, fully automated speech recognition, is not able to provide readable scripts from lecture recordings without extensive manual editing.

Subtitle creation

One the key aims of this feasibility study was to determine whether acceptably accurate subtitles could be automatically generated. Although it is possible to read the scripts while listening to the audio (or watching the videos) we concluded that this function is not feasible (without manual editing) for the following reasons:

1. **Timestamp is not accurate:** In order to link the scripts to actual play time, the audio has to be processed in small chunks – e.g. 5 seconds. Although it is technically possible to cut the audio into 5 seconds chunks, it is not possible to ensure that the speaking words are not chopped (e.g. start speaking at 4.9 seconds and finish speaking at 5.1 seconds). The smaller the chunks are, the more likely this will happen. As a result, the recognition result will be reduced. When cutting the audio in to tiny chunks, it appears that the recognition engines are not able to identify meaningful context from several words thus reducing the quality of recognition.
2. **Silence detection:** It is possible to cut the audio base on the pauses of speaking. This approach will not be able to guarantee consistent audio duration for each cut thus making the timestamp extremely complicated.
3. **Missing words or mismatching words**: Some audio chunks may not yield any results. For example, the following result below actually missed two sentences (that's also the reason why the confidence indicator is relatively low).

> **[TRANSCRIPT]** yeah I mean it's not the place to come and I'm happy to talk to you afterwards about it but I'll let you know around 6 I should go to professional internship. [confidence: 0.6304]

Results in relation to key research areas

1. **Discipline area:** Google and IBM engines perform very well in identifying specific words from different domains. For example:

> **[TRANSCRIPT]** products with heterosexual that that this culture of metrosexualality and they're more willing to be in the sea are submissive places such as exhausted as the epitome of metrosexuality the fall of David Beckham in the top shelf there but we do have the same maybe you changes in masculinity and in this regard let me to skip across a few here and there is change in relationships have to their bodies you actually changes.

[TRANSCRIPT] the old racism and whether you're right since it was erasing some of those I'm on my way to find racism in the consequences of that and use the case study of asylum-seekers is a case to think about racism but also to think about what might be a sociological approach to studying a highly controversial contested issue like this sociology as we talked about is collecting empirical information contrasting that testing social theories developing social theories but it's also you know I can't do it is.

2. **Single voice vs multiple voices:** The recognition results from single speakers are generally acceptable. However, it is not in the case in workshops where students will ask questions. The major issue is that the audiences are too far from the recording device thus not able to provide quality audio for recognition. For example, a 3 minute group discussion only produced the following results.

   [TRANSCRIPT] Belkin netcam out why do you want to share something about yourself.

   [TRANSCRIPT] emoticons greetings examples.

   [TRANSCRIPT] yeah, yeah, yeah, yeah.

3. **English speakers from different native language background:** Despite the speakers' background, the overall results are generally acceptable. However, for native English speakers who speak slowly and clearly, the recognition results are much better. For example, the example below almost had 100% accuracy.

   [TRANSCRIPT] seek out an activity or resource scroll to the bottom of the page and click the URL resource please add into the name of the URL resource and description open the video confirmation email highlight the link right click and copy the link close the email scroll down and paste the link into the external URL field.

4. **With and without background noise:** IBM and Google engines come with noise cancellation techniques. These techniques worked well for background music, but were less ideal for background human voices. For example, the example below wasn't effected by the loud background music.

   [TRANSCRIPT] become dishonest as adults lying to customers colleagues and even their Partners but all is not lost for the next 4 weeks I will be your lead educator guiding you through an exploration of several important questions such as what is academic Integrity why is it so important in Academia and how can you as a student at University but she with Integrity in this course will explore the answer to these and other questions each week.

Commercial vs Open Source
Open source engine CMUSphinx is able to produce some results, but not on par with Google and IBM, which both generated similar results exceeding researchers' expectations.

> **IBM:** This recording is protected by copyright no part may be reproduced without the prior permission at the university of South Australia.
>
> **Google:** This recording is protected by copyright know. Maybe we produced without the prior permission of the University of South Australia.
>
> **CMUSphinx:** this record is protected by copyright know what may be reproduced without our permission could the university.

Conclusion

For the creation of fully-automated, highly accurate subtitles in digital video it is recommended that a high quality audio recording is sourced, those related to podcasts rather than recorded lectures. The shorter average length of these types of recording mean manual editing would be more efficient. The level of accuracy currently available, however, is high enough to provide meaningful results for text analytics or topic modelling purposes and this is the direction in which this research now progresses. This is a potentially fruitful area of research and a function which may provide many benefits for students, though not to the extent that would fully support students in the ways mentioned above. After completing this feasibility study we concluded that, currently, none of the 3 transcription engines used are able to reach an acceptable level of accuracy for subtitle creation, without costly and time consuming human intervention. Given the amount of content produced by a university the level of manual editing would be far too costly to be of practical use.

References

Brasel, S. A., & Gips, J. (2014). Enhancing television advertising: same-language subtitles can improve brand recall, verbal memory, and behavioral intent. *Journal of the Academy of Marketing Science, 42*(3), 322-336.

Burnham, D., Leigh, G., Noble, W., Jones, C., Tyler, M., Grebennikov, L., & Varley, A. (2008). Parameters in television captioning for deaf and hard-of-hearing adults: Effects of caption rate versus text reduction on comprehension. *Journal of deaf studies and deaf education, 13*(3), 391-404.

CAST. (2011). Universal Design for Learning Guidelines Version 2.0. Wakefield MA.

Etemadi, A. (2012). Effects of bimodal subtitling of English movies on content comprehension and vocabulary recognition. *International journal of English linguistics, 2*(1), 239.

Gernsbacher, M. A. (2015). Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 195-202. https://doi.org/10.1177/2372732215602130

Janfaza, A., Jelyani, S. J., & Soori, A. (2014). Impacts of Captioned Movies on Listening Comprehension. *International Journal of Education & Literacy Studies, 2*(2), 80. https://doi.org/10.7575/aiac.ijels.v.2n.2p.80

Kothari, B. (2008). Let a billion readers bloom: Same language subtitling (SLS) on television for mass literacy. *International review of education, 54*(5-6), 773-780. https://doi.org/10.1007/s11159-008-9110-3

Koumi, J. (2014). Potent Pedagogic Roles for Video. *Media and learning association*.

Kruger, J.-L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research, 7*(5). https://doi.org/10.16910/jemr.7.5.4

Mohsen, M. A. (2015). The use of help options in multimedia listening environments to aid language learning: a review. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12305

Sapp, W. (2009). Universal Design: Online Educational Media for Students with Disabilities. *Journal of Visual Impairment & Blindness, 103*(8), 495-500. https://doi.org/10.1177/0145482X0910300807

Steinfeld, A. (1998). The Benefit of Real-Time Captioning in a Mainstream Classroom as Measured by Working Memory. *Volta review, 100*(1), 29-44.

Stinson, M. S., Elliot, L. B., Kelly, R. R., & Liu, Y. (2009). Deaf and hard-of-hearing students' memory of lectures with speech-to-text and interpreting/note taking services. *The Journal of Special Education, 43*(1), 52-64. https://doi.org/10.1177/0022466907313453

Vanderplank, R. (2016). 'Effects of' and 'effects with' captions: How exactly does watching a TV programme with same-language subtitles make a difference to language learners? *Language Teaching, 49*(02), 235-250. doi:10.1017/S0261444813000207

Wald, M. (2006). Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. *Interactive Technology and Smart Education, 3*(2), 131-141.

Woolfitt, Z. (2015). The effective use of video in higher education: The Hague. Retrieved from https://www. inholland. nl/media/10230/the-effective-use-ofvideo-in-higher-education-woolfitt-october-2015. pdf.

anón, N. T. (2006). Using subtitles to enhance foreign language learning. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*(6), 4.

Please cite as: Dinmore, S. & Gao, J. (2016). Voice-to-Text Transcription of Lecture Recordings. In S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds.), *Show Me The Learning. Proceedings ASCILITE 2016 Adelaide* (pp. 197-202). https://doi.org/10.14742/apubs.2016.864

Note: All published papers are refereed, having undergone a double-blind peer-review process.