Business model for Captioning University Lecture Recordings

Mike Wald [1]

Abstract

Deaf and Hard of Hearing people can benet from captioning of video recordings and transcription of audio recordings and guidelines (e.g. Web Content Accessibility Guidelines (WCAG 2.0)) or legislation in many countries has been drafted to encourage or require organisations to caption and transcribe their recordings. Captioning and transcription by qualied and trained professionals is expensive and not affordable in many countries and situations (e.g. University Lectures). Captions and transcriptions can also be of great benet to non-native speakers or when there is a high level of background noise or there is no audio available (e.g. TV screens in public places such as airports) and also allow video and audio recordings to be searched. Professional captioning is expensive and crowdsourcing may reduce the costs but requires comparisons to verify the accuracy of the work and so as the number of crowdsourced captioners is increased to increase accuracy, the cost is also increased. Automatic Speech Recognition can be used to automatically caption and transcribe recordings but requires human correction of errors. This paper compares captioning methods and costs and suggests an affordable model might be to use students to correct automatic speech recognition errors of their lecture recordings.

Keywords: Captioning, Transcription speech recognition, Costs, Business model

[1] Professor in University of Southampton, UK. E-mail: m.wald@soton.ac.uk

โมเดลธุรกิจสำหรับคำบรรยายการบันทึกการสอนในมหาวิทยาลัย

Mike Wald [1]

บทคัดย่อ

คนหูหนวกและผู้ที่มีความบกพร่องทางการได้ยินได้รับผลประโยชน์จากคำบรรยายเสียงพูดและคำอธิบายเสียงพูดของการบันทึกเสียง (เช่น แนวทางในการพัฒนาเนื้อหาบนเว็บไซต์ที่เข้าถึงได้ (WCAG 2.0)) หรือกฎ ระเบียบในหลายประเทศ ซึ่งได้จัดทำขึ้นเพื่อกระตุ้นหรือต้องการให้หน่วยงานต่าง ๆ จัดทำคำบรรยายและคำอธิบายการบันทึกเสียงของตน การจัดทำคำบรรยายและคำอธิบายเสียงพูดโดยผู้ที่มีชำนาญและมีประสบการณ์ในการให้คำบรรยายเสียงพูดมีราคาแพงมาก และหน่วยงานต่าง ๆ ไม่สามารถที่จะรับผิดชอบค่าใช้จ่ายเหล่านี้ได้ (เช่น อาจารย์มหาวิทยาลัย) คำบรรยายและคำอธิบายเสียงพูดยังเกิดประโยชน์กับกลุ่มผู้ใช้ที่ไม่ใช่เจ้าของภาษา หรือเมื่ออยู่ในเหตุการณ์ที่มีเสียงรบกวน หรือไม่มีเสียงประกอบ (เช่น จอโทรทัศน์ที่สนามบินที่มีเฉพาะภาพแต่ไม่มีเสียงประกอบ) คำบรรยายเสียงพูดที่ผู้เขียนพัฒนาขึ้นยังสามารถที่จะค้นหาเนื้อหาของเสียงหรือวีดีโอที่บันทึกไว้แล้ว การจ้างผู้ที่มีความเชี่ยวชาญในการจัดทำคำบรรยายนั้นมีราคาแพง ถึงแม้ว่ามีการใช้เทคนิคการร่วมกันแก้ไขคำบรรยายออนไลน์ ซึ่งอาจช่วยลดค่าใช้จ่ายลงได้ แต่เมื่อเปรียบเทียบความถูกต้องของการทำงานกับจำนวนของผู้ที่เข้ามาทำงานร่วมในการแก้ไขคำบรรยายให้มีความถูกต้องยังคงส่งผลให้ค่าใช้จ่ายเพิ่มขึ้น คำบรรยายแบบอัตโนมัติสามารถใช้เทคโนโลยีการให้คำบรรยายและการให้คำอธิบายแบบอัตโนมัติ แต่อย่างไรก็ตาม ยังคงต้องการมนุษย์ในการตรวจสอบความผิดพลาดที่เกิดขึ้น บทความนี้จะเปรียบเทียบวิธีการในการให้คำบรรยาย ค่าใช้จ่าย และเสนอแนะโมเดลธุรกิจในการแก้ไขความผิดพลาดที่เกิดขึ้นจากการใช้เทคโนโลยีให้คำบรรยายแบบอัตโนมัติ โดยนักศึกษาเข้ามาร่วมแก้ไขข้อผิดพลาดจากการฟังวีดีโอบรรยายการเรียนของพวกเขา

คำสำคัญ: คำบรรยายเสียงพูด คำอธิบายเสียงพูด ค่าใช้จ่าย โมเดลธุรกิจ

[1] Professor in University of Southampton, UK. E-mail: m.wald@soton.ac.uk

Introduction

Deaf and Hard of Hearing people can benefit from captioning of video recordings and transcription of audio recordings and guidelines (e.g. Web Content Accessibility Guidelines (WCAG 2.0)) or legislation in many countries has been drafted to encourage or require organisations to caption and transcribe their recordings resulting in law suits (NYTimes 2015). Captioning and transcription by qualified and trained professionals is expensive and not affordable in many situations (e.g. University Lectures) and countries.

Captions and transcriptions can also be of great benefit to non-native speakers or when there is a high level of background noise or there is no audio available (e.g. TV screens in public places such as airports) and also allow video and audio recordings to be searched.

Crowdsourcing has been used to try and reduce the costs but requires comparisons to verify the accuracy of the work and so as the number of crowdsourced captioners is increased to increase accuracy, the cost is also increased.

Speech Recognition can be used either to automatically caption and transcribe recordings or by captioners and transcribers to respeak what was said instead of typing on a keyboard. Live captioning of up to 240 words per minute in courtrooms or for television programmes was originally undertaken using a phonetic keyboard by stenographers with 5 years training but nowadays is often undertaken using speech recognition respeaking.

While there has been much research into the use of speech recognition based captioning and transcription there has been little published research into the costs and benefits of using speech recognition technology for lecture captioning and transcription in comparison to commercial charges for manual captioning. This article addresses this gap.

Background

The use of video including lecture recording is increasing in many countries (Sonic Foundry 2016) but is rarely captioned or transcribed because of the high costs involved. Professional manual captioning is time consuming and therefore expensive, typically $180/hr (3Playmedia 2016). Automatic captioning is possible using speech recognition technologies but this results in many recognition errors requiring manual correction (Bain et al 2002). With training of the software and experience some speakers can sometimes achieve less than 10% word error rates with current speech recognition technologies for conversational speech using good quality microphones in a good acoustic environment. With conversational speech however the accuracy can drop as the speaker speeds up and begins to run the ends of words into the beginnings of the next word. Speakers also use fillers (e.g. ums and ahhs) and sometimes hesitate in the middle of a word. People do not speak punctuation marks aloud when conversing normally but speech recognition technologies designed for dictation use dictated punctuation to indicate the end of one phrase or sentence and the beginning of another to assist the statistical recognition processing of which words are likely to follow other words. However, often it is not possible to train the speaker or the software and in these situations, depending on the speaker and acoustic environment, word error rates can increase to over 30% (Fiscus et. al. 2005) even using the best speaker independent systems and therefore extensive manual corrections may be required. If close to 100% accuracy is required then a human editor will be required and even if the Word Error Rate is very low, as unless a human editor checks it nobody can be certain of the accuracy.

Comparative Study

The only extensive comparative study of captioning and transcription quality and costs appears to have been undertaken in 2012 by members of the international Liberated Learning Consortium (LLC) with the support of the author, a LLC founding member.

Recordings were organized into four 'sets' based on 3 topics, 4 lengths, 3 recording qualities, and 3 speaker accents as follows:

Topic: easy - general topic and terminology; medium - some specialized terminology, first-year college course; difficult - specialized terminology, second or third year University course

Length: Short (<10 min), medium (10 to 25 min), long (>25 min)

Audio quality: Low , medium or high

Accent: Native North American speaker or Australian or Indian English

Researchers measured transcription accuracy with a 3-5% confidence interval using Word Error Rate (number of incorrect (misrecognized) words divided by the total number of words spoken * 100). The average WER among providers was 2.65% and there were significant differences in costs between providers. The average cost for transcribing one hour of media was $164.20 (Figure 1).

Evaluation Set	Recording Length (Min)	Rate/Min	Actual Cost	Average Rate / Hour Audio
A	68	$4.29	$291.72	$257.40
	3	$3.86	$11.58	$231.60
	12	$3.86	$46.32	$231.60
	49	$2.87	$140.63	$172.20
	123	$2.87	$353.01	$172.20
	255	**avg $3.55**	**$843.26**	**$213.00**
B	74	$2.90	$214.60	$174.00
	7	$2.50	$17.50	$150.00
	12	$2.50	$30.00	$150.00
	56	$2.90	$162.40	$174.00
	106	$2.50	$265.00	$150.00
	255	**avg $2.66**	**$689.50**	**$159.60**
C	69	$2.57	$177.00	$153.91
	45	$2.53	$114.00	$152.00
	55	$2.18	$120.00	$130.91
	5	$2.40	$12.00	$144.00
	12	$2.75	$33.00	$165.00
	186	**avg $2.49**	**$456.00**	**$149.16**
D	49	$2.03	$99.21	$121.80
	102	$1.72	$175.23	$103.20
	69	$2.03	$140.74	$121.80
	220	**avg $1.93**	**$415.18**	**$123.99**

Figure 1: Average and total costs per minute and hour

While the cheapest company (4) gave the lowest accuracy, the most expensive company (1) did not give the highest accuracy transcripts (Figure 2).

Company	Rank Cost (Most Expensive)	Rank Quality (Lowest WER)
1	1	3
2	2	2
3	3	1
4	4	4

Figure 2: Cost and quality rankings

These costs were for a transcribed file and including time synchronization with the transcriptions for captioning purposes would add up to 58% to the total cost of service for companies. For example, transcribing a one hour media file and creating the timing data for publishing synchronized media would costs $259.44 (164.20 x 158%).

For comparison with speech recognition transcription researchers selected a sample of 299 transcriptions of educational media created using automatic speech recognition over a two year period from over 30 different post-secondary institutions with a variety of speaker accents and audio quality with an average WER of 32.08%, and with a large standard deviation of 30.70%. The speech recognition automatically time synchronized the transcript with the recordings allowing for search and captioning. The average editing effort required to produce corrected transcripts was 4.10 hours / media hour, with a standard deviation of 2.00 hours. For this range of expected accuracy levels and corresponding editing required, Figure 3 demonstrates potential variable costs that would be associated with this theoretical service. Given regional wage discrepancies, sample hourly wages were established from $15 to $30/hour.

Editing Effort (Hours)	Variable Cost / Editing Hour			
	$15	**$20**	**$25**	**$30**
2	$30	$40	$50	$60
2.5	$38	$50	$63	$75
3	$45	$60	$75	$90
3.5	$53	$70	$88	$105
4	**$60**	**$80**	**$100**	**$120**
4.5	$68	$90	$113	$135
5	$75	$100	$125	$150
5.5	$83	$110	$138	$165
5.8	$87	$116	$145	$174

Figure 3: Editing Costs and Level of Effort

Assuming a normal distribution, about 68% of future recordings could be transcribed and corrected with the currently available technology at an average variable cost between $60-$120/media hour, depending on wage levels.

Although this case study did not provide a direct comparison between traditional transcription services and state of the art speech recognition, it does provide baseline variable cost and quality indicators. The evidence suggests that if overheads and speech recognition costs were low, universities transcribing educational media using a combination of SR and human editing to correct recognition errors might be more economical than using commercial transcription services.

Collaborative Editing

The use of students in the class at little or no cost to edit the transcripts using a collaborative editor while they listened to or watched the recordings as part of their normal learning activity could reduce the editing costs greatly.

If there is no correct version of the transcript in existence there is no simple way of knowing whether the person creating or correcting the captions is making errors or not and so Wald (2011a, 2013) developed a collaborative editing tool that stores all the edits of all the users and uses a matching algorithm to compare users "edits to check if they are in agreement. A video demonstration of this tool is available for downloading (Wald 2011b) and is also available on Synote (Wald 2011c) captioned using Synote's speech recognition editing system. If users wish to annotate the recording on Synote they need to register before logging in with their registered user name and password, otherwise they can go to the "Read, Watch or Listen Only Version". The panels and size of the video can be adjusted up to full screen and the size of the text can also be enlarged.

The collaborative correction tool shown in Figure 3 stores all the edits of all the users and uses a matching algorithm to compare users' edits to check if they are in agreement before finalizing the 'correct' version of the caption. This improves the captioning accuracy and also reduces the chance of 'spam' captions. The tool allows contiguous utterances from sections of the transcript to be presented for editing to particular users or for users to be given the freedom to correct any utterance. The idea of the tool is that students could watch recordings of lectures that have captions created by automatic speech recognition and they could correct as many or as few of the recognition errors as they chose. Administrator settings allow for different matching algorithms based on the closeness of a match and the number of users whose corrections must agree before accepting the edit. Contractions

are accepted (e.g. I'm) as meaning the same as the full version (i.e. 'I am') and to enable these 'rules' to be easily extended a substitution rules XML file uploader is provided. As shown in Figure 4 the red bar on the left of the utterance and the tick on the right denote that a successful match has been achieved and so no further editing of the utterance is required while the green bar denotes that the required match for this utterance has yet to be achieved.

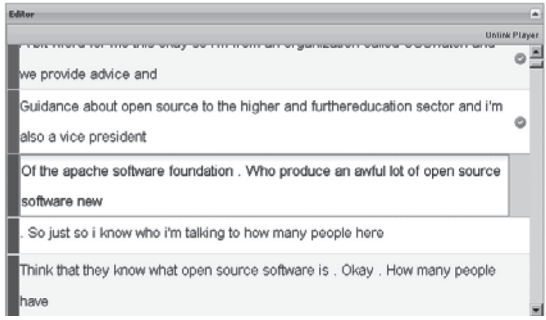

Figure 4: Collaborative correction tool

Scores

First Name	Last Name	Rewards	Penalties	Score
Alexander	Kilcoyne	0	0	0
dmk106	dmk106	0	0	0
Mike	Kanani	1	0	1
M	W	7	2	5
m	w	8	1	7
Alex	Kilcoyne	0	0	0
Stanley	Kubrick	0	0	0

Figure 5: Rewards and penalty scores

User	Final	Similarity	Word Changes	Utterance
				50001
				welcome to this brief interactive guide to using senate , what is senate ,
u1	✓	92	1	welcome to this brief interactive guide to using Synote , what is senate ,
u2	-	83	2	welcome to this brief interactive guide to using Synote , what is Synote ,
u3	✓	92	1	welcome to this brief interactive guide to using synote , what is senate ,
u4	-	-	-	NOT EDITABLE
u5	-	-	-	NOT EDITABLE
u6	-	-	-	NOT EDITABLE

Figure 6: Report showing users' edits

Various display and editing modes are provided for users. Users are awarded points for a matching edit and it is also possible to remove points for corrections that do not match other users' corrections (Figure 5). A report is available showing users' edits (Figure 6).

Conclusion

This paper compared the costs and benefits of manual and automatic transcription and captioning methods and suggested an affordable model for Universities might be to use students to collaboratively correct automatic speech recognition transcription errors of their lecture recordings. A new version of Synote has been developed and a company launched in 2016 to provide speech recognition captioning and a collaborative editing platform http://synote.com/ at 6/hr 'pay as you go' with a discount available for contracts for substantial number of hours of captioning.

Reference

3 playmedia. (2016). **Plans & Pricing** Retrieved from http://www.3playmedia.com /plans-pricing/.

Bain, K., Basson, S., Wald, M. (2002). **Speech recognition in university classrooms.** In: Proceedings of the Fifth International ACM SIGCAPH Conference on Assistive Technologies. ACM Press, 192-196.

Fiscus, J., Radde, N., Garofolo, J., Le, A., Ajot, J., Laprun, C., (2005). **The Rich Transcription 2005 Spring Meeting Recognition Evaluation,** National Institute of Standards and Technology.

NYTIMES. (2015). **Tamar Lewin.** Retrieved from http://www.nytimes.com/ 2015/02/13/education/harvard-and-mit-sued-over-failing-to-caption-online-courses.html?__r=0.

Sean Brown. (2016). **Sonic Foundry.** Retrieved from http://www.sonicfoundry. com/4-predictions-for-online-enterprise-video-in-2016/.

Wald, M. (2011). **Crowdsourcing Correction of Speech Recognition Captioning Errors.** In, W4A: 8th International Cross-Disciplinary Conference on Web Accessibility, Hyderabad, India, W4A.

Wald, M. (2011). **Crowdsource Captioning Demonstration Video.** Retrieved from http://users.ecs.soton.ac.uk/mw/recordings/Mike%20Wald/webaccessibilitycompetitionsubmit/webaccessibilitycompetitionsubmit.wmv

Wald, M. (2011). **Crowdsource Captioning Demonstration Video with Interactive Transcript.** Retrieved from http://www.synote.org/synote/recording/replay/55564.

Wald, M. (2013). **Concurrent Collaborative Captioning.** Proceedings of SERP'13 - The 2013 International Conference on Software Engineering Research and Practice Retrieved from http://eprints.soton.ac.uk/354312/. W3C 2018 https://www.w3.org/TR/WCAG20/.