Concurrent Collaborative Captioning

SERP'13
Mike Wald
University of Southampton, UK
+442380593667
M.Wald@soton.ac.uk

ABSTRACT
Captioned text transcriptions of the spoken word can benefit hearing impaired people, non native speakers, anyone if no audio is available (e.g. watching TV at an airport) and also anyone who needs to review recordings of what has been said (e.g. at lectures, presentations, meetings etc.) In this paper, a tool is described that facilitates concurrent collaborative captioning by correction of speech recognition errors to provide a sustainable method of making videos accessible to people who find it difficult to understand speech through hearing alone. The tool stores all the edits of all the users and uses a matching algorithm to compare users' edits to check if they are in agreement.

Keywords
Accessibility, Speech recognition, Captioning, Collaborative editing

1. INTRODUCTION
As more videos are becoming available on the web these require captioning/(subtitling) if they are to benefit hearing impaired people, non-native speakers, anyone if no audio is available (e.g. watching TV at an airport) and also anyone who needs to search, review recordings of what has been said (e.g. at lectures, presentations, meetings etc.) or translate the recording.

The provision of synchronized text captions with video also enables all their different communication qualities and strengths to be available as appropriate for different contexts, content, tasks, learning styles, learning preferences and learning differences. For example, text can reduce the memory demands of spoken language; speech can better express subtle emotions; while images can communicate moods, relationships and complex information holistically.

Professional manual captioning is time consuming and therefore expensive[i] (e.g.180$/hour). 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, unless a human editor checks it nobody can be certain of the accuracy.

In this paper, further details of the development of a tool is described that facilitates collaborative correction of speech recognition captioning errors to provide a sustainable method of making audio or video recordings accessible to people who find it difficult to understand speech through hearing alone (Wald 2011). 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. The new approach described in this paper therefore is to allow many people to edit the captions at the same time and automatically compare their edits to verify they are correct. The term 'Social Machines'[ii] has been used to describe such large scale collaborative problem solving by humans and computers using the web.

Section 2 reviews other approaches, section 3 describes Synote and its captioning method, section 4 describes the new collaborative caption creation tool while section 5 summarises the conclusions and future planned work.

2. Review of Other Approaches
There are many web based captioning tools some which only allow captioning of videos they host (e.g. YouTube[iii], overstream[iv], dotsub[v]) while others allow manual captioning of web based videos hosted elsewhere (e.g.

Amara[vi], originally a Mozilla Drumbeat project called Universal Subtitles; CaptionTube[vii]; Subtitle Horse[viii]; Easy YouTube Caption Creator[ix]).

There are also many examples of desktop captioning/ subtitling software (e.g. magpie[x], MovieCaptioner[xi], Subtitle Workshop[xii] etc.) but these cannot normally be used with web hosted video and would involve transferring files between captioners if more than one person was involved in captioning.

None of the captioning systems are designed to allow more than one person at a time to create the captions or edit the captions.

Transcription is not only used for hearing impaired and non native speakers. Speech recognition scientists need transcribed speech to build and improve their acoustic speech models but the accuracy of the transcriptions is less important as Novotney and Chris Callison-Burch (2010) showed that the accuracy of speech recognition models could be improved more cheaply using more lower accuracy transcriptions by Amazon Mechanical Turk to transcribe speech for 3% of the cost of more accurate professional transcription. Lee & Glass (2011) used workers on the Amazon Mechanical Turk[xiii] with two stages of transcription each using ASR to filter out poor quality. The first stage presented each worker with five, five to six second clips created automatically by silence detection. A 15% word error rate (WER) was achieved by proving feedback using an automatic quality detector measuring both the range of words used (e.g. to detect lots of 'ums') and how closely it matched the n best words and phoneme sequences (e.g. to detect random corrections) rejecting poor quality transcripts with a WER greater than 65%. The second stage joined together clips to make 75 seconds of audio synchronised with the first stage transcripts to provide more audio context. Feedback on performance quality (with 80% being the acceptance threshold) was given by comparing the number of corrections made with the number of corrections needed estimated by using ASR word confidence scores. Their trained support vector machine classifier was able to judge 96.6% of the submitted transcripts correctly, reducing poor quality transcripts by over 85% and WERs to 10%.

3. SYNOTE

Synote[xiv] (Wald 2010, 2011) is a cross browser web based application that can use speaker independent speech recognition[xv] for automatic captioning of recordings. Synote also allows synchronization of user's notes and slide images with recordings and has won national[xvi] and international awards[xvii] for its enhancement of education and over the past four years has been used in many countries[xviii]. Figure 1 shows the Synote interface with the video in the upper left panel, the synchronized transcript in the bottom left panel with the currently spoken words highlighted in yellow and the individually editable 'utterances' in the right panel. While Synote provides an editor to correct speech recognition errors in the synchronised transcript in the bottom left panel, the whole transcript rather than individual corrections are saved to the Synote server which can take a substantial time (many seconds). If two people edit the same transcript then the most recently saved version will overwrite the previously saved version. It is therefore only possible to use collaborative editing in this way by only permitting one person to edit at a time. While this approach can be used for professional editing, that is not an affordable solution for editing of lecture recordings in universities. The individual captions in the right hand panel are however saved individually and so it may be possible to motivate students to correct some of the errors while reading and listening to their lecture recordings by providing rewards, for example in the form of academic credits. Some short experiments using a few students have indicated that students who edit the transcript of a recorded lecture do better on tests on the content of that lecture than students who just listen to and watch the lecture. The top right hand 'Synmark' (**SYN**chronised book**MARKS**) panel was originally designed for creating synchronized notes rather than captions although it does also allow for multimedia captions as shown in Figure 2. where each caption has a picture of the speaker and a different colour for what they are saying which is very helpful to identify which speaker a caption refers to. The pictures of the speaker are not stored on Synote's server but can be stored anywhere on the web (e.g. imdb.com). A 'parser' was developed (Figure 3) to automatically split the transcript into utterances which could be uploaded as 'Synmarks'. This enables the best way of automatically splitting the synchronized transcript into editable utterances/captions to be investigated; including the number of words in an utterance, total time length of utterance, the length of silence between words or by the commas inserted by the default silence setting of the IBM speaker independent speech recognition system (Soltau et. al. 2010). The best way of automatically presenting the utterances for correction is also being investigated including separating utterances with commas or full stops or spaces and capitalizing the first word of each utterance. The system can produce both a standard text format SRT file for use with most captioning systems or an XML file for use with Synote.

Figure 4 shows some of a transcript created using speech recognition without splitting into utterances while Figure 5 shows the same transcript split into utterances using silences.

The transcript file format uploaded to the parser could be Synote XML (the native format of the IBM speech recognition used by Synote) , Synote Print preview (Synote's output format and so allowing uploading of Synote's manually edited and/or transcribed synchronized

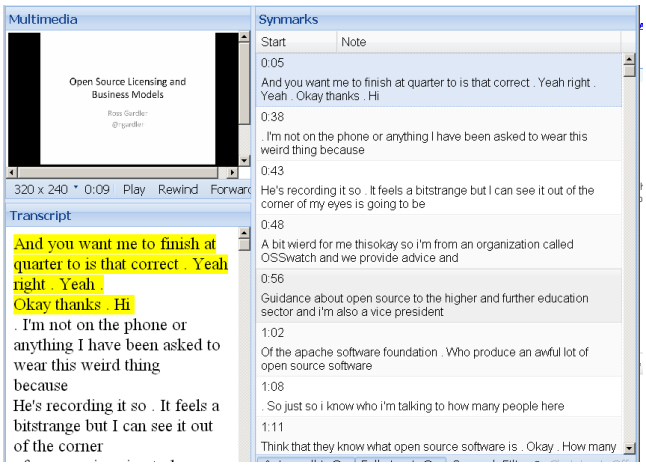

Figure 1. Synote Player Interface

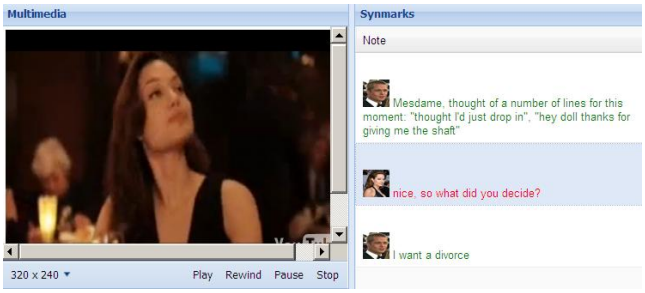

Figure 2. Captioning in Synmarks

transcripts) or SRT (A common video captioning format). Although Synmarks are saved to the server when they are created by a user any changes to Synmarks by users will only be updated in other users' Synmark panels when they choose to refresh the browser. This was a decision made when Synote was being designed as updating all the Synmarks whenever one Synmark was edited or created took a few seconds and so detracted from the user experience. This means that if users are editing the captions in the Synmark panel, they must regularly refresh the browser to check if any other users have edited or corrected any Synmarks. Synote only stores the most recent edit to a Synmark and keeps no record of previous edits. Synote also allows multiple users to concurrently manually caption or correct the errors in the speech recognition transcript. If two users concurrently select the same time period to caption (i.e. without realizing the other user is captioning Synmarks) this could create an unsatisfactory user experience of seeing multiple captions. If two users concurrently edit the same speech recognition utterance in a Synmark then the first person to save their correction will have their correction overwritten by the second person saving their corrections. A research tool was therefore developed to investigate what would be the best design for a collaborative editing tool.

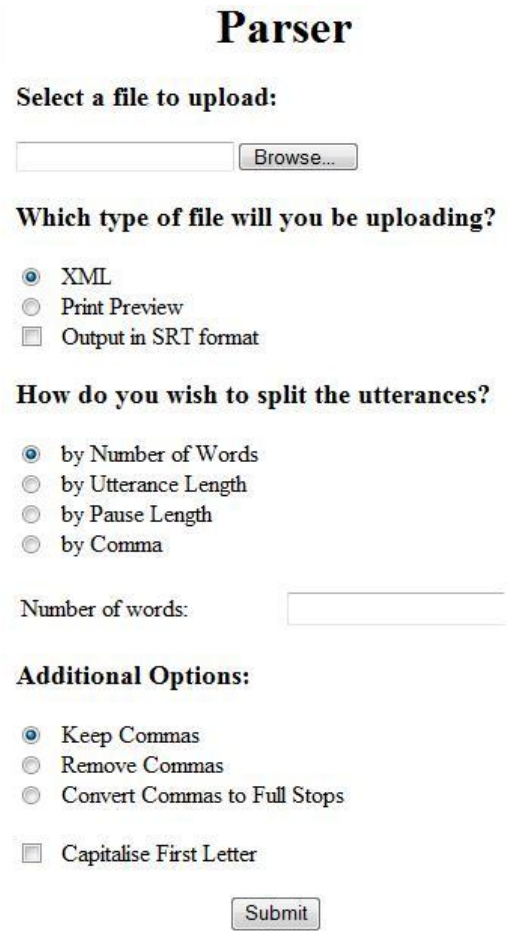

Figure 3. Transcript Parser

This is a demonstration of the problem of the readability of text created by commercial speech recognition software used in lectures they were designed for the speaker to dictate grammatically complete sentences using punctuation by saying comma period new paragraph to provide phrase sentence and paragraph markers when people speak spontaneously they do not speak in what would be regarded as grammatically correct sentences as you can see you just see a continuous stream of text with no obvious beginnings and ends of sentences normal written text would break up this text by the use of punctuation such as commas and periods or new lines by getting the software to insert breaks in the text automatically by measuring the length of the silence between words we can improve the readability greatly

Figure 4. Transcript without splitting into utterances

This is a demonstration of the problem of the readability of text created by commercial speech recognition software used in lectures

they were designed for the speaker to dictate grammatically complete sentences using punctuation by saying comma period new paragraph to provide phrase sentence and paragraph markers

when people speak spontaneously they do not speak in what would be regarded as grammatically correct sentences

as you can see you just see a continuous stream of text with no obvious beginnings and ends of sentences

normal written text would break up this text by the use of punctuation such as commas and period or new lines

by getting the software to insert breaks in the text automatically by measuring the length of the silence between words we can improve the readability greatly

Figure 5. Transcript split into utterances

4. COLLABORATIVE CAPTIONING TOOL

The collaborative correction tool shown in Figure 6 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 choose. Administrator settings (Figure 7) 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 (Figure 8). As shown in Figure 6, 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. 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 9). A report is available showing users' edits (Figure 10). Investigations are currently underway using this research tool in order to determine the most sustainable approach to adopt for collaborative editing. The tool has been designed to be scalable for wide scale 'crowdsourcing' of captioning.

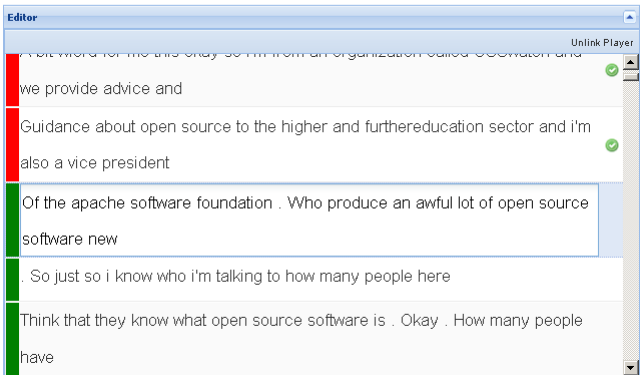

Figure 6. Collaborative correction tool

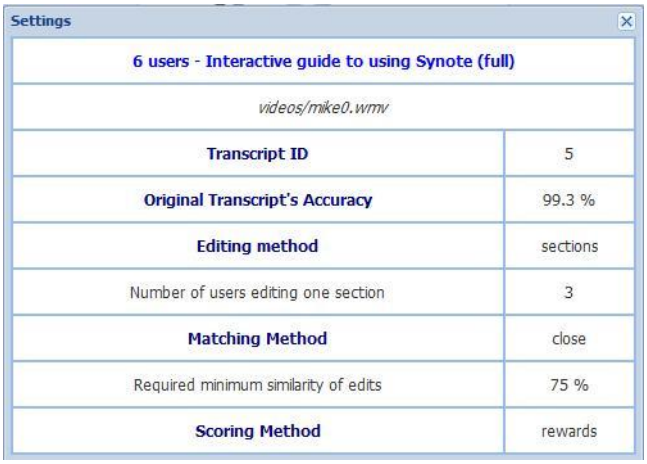

Figure 7. Collaborative Tool Settings

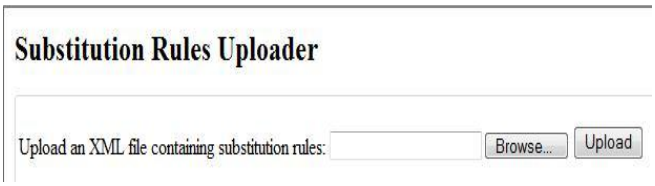

Figure 8. Substitution Rules Uploader

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 9. 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 10. Report showing users' edits

5. CONCLUSION

The use of collaborative correction of speech recognition errors offers a promising approach to providing sustainable captioning and Synote and its associated parser and collaborative correction tool provide the opportunity to investigate the best approach for both making it as easy as possible for users to correct the transcripts and also for providing the motivation for them to do so. Future work will involve further user trials of the system. A wmv format video demonstration of the systems tools described in this paper is available for downloading[xix] and is also available on Synote[xx] 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.

6. ACKNOWLEDGMENTS

Dawid Koprowski is the collaborative tool's lead developer and other ex ECS students Mike Kanani, Karolina Kaniewska, Stella Sharma were also involved in the tool's development and Alex Kilcoyne conducted the user trials

7. REFERENCES

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

Lee, C. Y., Glass, J. (2011) A transcription task for crowdsourcing with automatic quality control. *Proc. Interspeech2011, Florence*.

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

Novotney, S. Callison-Burch, C. (2010) "Cheap, fast and good enough: automatic speech recognition with non-expert transcription," in Proc. HLT-NAACL, pp. 207-215.

Soltau, Hagen; Saon, G.; Kingsbury, B. (2010) "The IBM Attila speech recognition toolkit," *Spoken Language Technology Workshop (SLT), 2010 IEEE* , pp.97-102

Wald, M. (2010) Synote: Designed for all Advanced Learning Technology for Disabled and Non-Disabled People. In, *Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies, Sousse, Tunisia*, pp 716-717.

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

[i] http://www.automaticsync.com/caption/

[ii] http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/J017728/1

[iii] www.youtube.com

[iv] http://www.overstream.net/
[v] http://dotsub.com/
[vi] http://www.amara.org
[vii] http://captiontube.appspot.com
[viii] http://www.subtitle-horse.com/
[ix] http://accessify.com/tools-and-wizards/accessibility-tools/easy-youtube-caption-creator/
[x] http://ncam.wgbh.org/invent_build/web_multimedia/tools-guidelines/magpie
[xi] http://www.synchrimedia.com/#movcaptioner
[xii] http://www.urusoft.net/products.php?cat=sw&lang=1
[xiii] https://www.mturk.com/
[xiv] http://www.synote.org
[xv] http://www.liberatedlearning.com/news/AGMSymposium2009.html
[xvi] http://www.ecs.soton.ac.uk/news/3874
[xvii] http://www.eunis.org/activities/tasks/doerup.html
[xviii] http://www.net4voice.eu
[xix] http://users.ecs.soton.ac.uk/mw/recordings/Mike%20Wald/webaccessibilitycompetitionsubmit/webaccessibilitycompetitionsubmit.wmv
[xx] http://www.synote.org/synote/recording/replay/55564