Crowdsourcing Correction of Speech Recognition Captioning Errors

M Wald,
University of Southampton, UK
+442380593667
M.Wald@soton.ac.uk

ABSTRACT
In this paper, we describe a tool that facilitates crowdsourcing correction of speech recognition captioning errors to provide a sustainable method of making videos accessible to people who find it difficult to understand speech through hearing alone.

Categories and Subject Descriptors
K.4.2 Social Issues - Assistive technologies for persons with disabilities

General Terms
Design, Human Factors

Keywords
accessibility

1. INTRODUCTION
Text transcriptions of the spoken word can benefit deaf people and also anyone who needs to review what has been said (e.g. at lectures, presentations, meetings etc.) The provision of synchronized text captions (subtitles) with video 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. As more videos are becoming available on the web these require captioning if they are to be accessible for those who find it difficult to understand speech through hearing alone. Captions also make it easier to search or translate the recording. Professional manual captioning is time consuming and therefore expensive[i]. Automatic captioning is possible using speech recognition technologies but this results in many recognition errors requiring manual correction[ii]. With training of the software some speakers can achieve less than 10% word error rates with current speech recognition technologies when dictating 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. With training and experience however some people can sometimes still achieve less than 10% word error rates for conversational speech. 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% even using the best speaker independent systems and therefore extensive manual corrections may be required. In this paper, we describe a tool that facilitates crowdsourcing 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.

W4A2011 – 'Microsoft Challenge', March 28-29, 2011, Hyderabad, India. Co-Located with the 20th
International World Wide Web Conference.

2. SYNOTE
Synote[iii] is a cross browser web based application that can use speaker independent speech recognition[iv] for automatic captioning. Synote also allows synchronization of user's notes and slide images with recordings and has won international awards[v] for its enhancement of education and over the past 3 years has been used in many countries[vi]. 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 it is possible to correct speech recognition errors in the synchronised transcript, the whole transcript rather than individual corrections are saved which can take some time. If two people edit the same transcript then the most recent saved version will overwrite the previously saved version. It is therefore only possible to use collaborative crowdsourcing 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. Previous studies 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 right hand 'Synmark' panel was originally designed for creating synchronized notes rather than captions and so only stores the most recent edit and keeps no

record of previous edits. A research tool was therefore developed to investigate what would be the best design for a crowdsourcing collaborative editing tool.

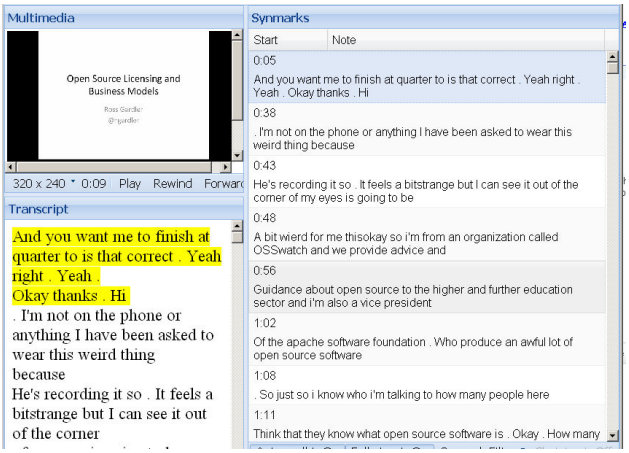

Figure 1. Synote Player Interface

3. CAPTION CREATION TOOL

The best way of automatically splitting the synchronized transcript into editable utterances/captions is being investigated; including the number of words in an utterance, total time length of utterance, or the length of silence between words. The best way of automatically presenting the utterances for correction is also being investigated including separating with commas or full stops and capitalizing the first word. The system can produce both a standard text format file for use with most captioning systems or an XML file for use with Synote.

4. CROWDSOURCING TOOL

The crowdsourcing correction tool shown in Figure 2 stores all the edits of all the users and uses a matching algorithm to compare users' edits to check if they are in agreement. 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. 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. 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. Users can be awarded points for a matching edit and it is also possible to remove points for corrections that do not match other users' corrections. Investigations are currently underway using this research tool in order to determine the most sustainable approach to adopt.

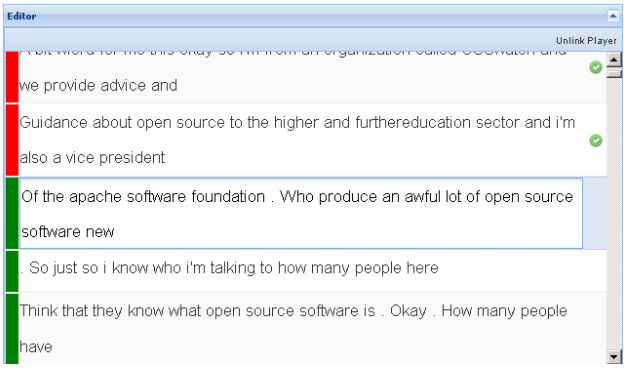

Figure 2. Crowdsourcing correction tool

5. CONCLUSION

The use of crowdsourcing of the correction of speech recognition errors offers a promising approach to providing sustainable captioning and Synote and its associated tools 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. A wmv format video demonstration of the systems described in this paper is available for downloading[vii] and is also available on Synote[viii] 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

Our thanks to ECS students Mike Kanani, Karolina Kaniewska, Dawid Koprowski, Stella Sharma for their help in developing the collaborative editing tools and to Alex Kilcoyne for help with user trials

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

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

[iii] http://www.synote.org

[iv] http://www.liberatedlearning.com/news/AGMSymposium2009.html

[v] http://www.eunis.org/activities/tasks/doerup.html

[vi] http://www.net4voice.eu

[vii] http://users.ecs.soton.ac.uk/mw/recordings/Mike%20Wald/webaccessibilitycompetitionsubmit/webaccessibilitycompetitionsubmit.wmv

[viii] http://www.synote.org/synote/recording/replay/55564