Enhancing Learning Accessibility through Fully Automatic Captioning

Maria Federico
Servizio Accoglienza Studenti Disabili
Università di Modena e Reggio Emilia
Modena, Italy
maria.federico@unimore.it

Marco Furini
Dipartimento di Comunicazione ed Economia
Università di Modena e Reggio Emilia
Reggio Emilia, Italy
marco.furini@unimore.it

ABSTRACT

The simple act of listening or of taking notes while attending a lesson may represent an insuperable burden for millions of people with some form of disabilities (e.g., hearing impaired, dyslexic and ESL students). In this paper, we propose an architecture that aims at automatically creating captions for video lessons by exploiting advances in speech recognition technologies. Our approach couples the usage of off-the-shelf ASR (Automatic Speech Recognition) software with a novel caption alignment mechanism that smartly introduces unique audio markups into the audio stream before giving it to the ASR and transforms the plain transcript produced by the ASR into a timecoded transcript.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Science**]: Sociology; K.4.2 [**Computers and Society**]: Assistive technologies for persons with disabilities; I.7.2 [**Document and Text Processing**]: Multi/mixed Media

General Terms

Algorithms, Design, Experimentation

Keywords

Accessibility, Learning, Automatic Captioning

1. INTRODUCTION

Education is one of the most important factors that contributes to the growth of individual and society. A more educated society may translate into higher rates of innovation, higher overall productivity and faster introduction of new technology [1]. Unfortunately, while for many people is easy attend lessons in educational Institutes, for millions of people with some form of disabilities the simple act of listening or of taking notes may represent an insuperable burden. For instance, hearing impaired students have difficulties in just listening to a lesson; motion impaired students may have problems in taking notes; dyslexic students may have trouble with note-taking and with whiteboard reading; ESL (English as a Second Language) students may experience difficulties in understanding how a teacher speaks.

To ameliorate the gap, many educational institutes provide educational material in digital video format. The access to such material allows students to watch a lecture any time they want, to tune the speakers' volume according to their needs and to rewatch it when content is not understood. While many students benefit from this approach, many other students still have the problem of clearly listening to what is said in the video. Therefore, captioning techniques become of vital importance as they display a textual version of what is said in the video. Roughly, a caption technique integrates the timecoded video transcript with the video material so as to synchronously illustrate the scripts during the playing of videos [2]. Unfortunately, in many scenarios the timecoded video transcript is not always available (e.g., live TV talk-show) and therefore a transcript needs to be produced on-the-fly. In addition to fast typists, currently, the most used approach is so-called *shadow speaking*: a human being slowly repeats what is said by different people in the video so as to make the speech recognition software able to understand his/her speech. Again, the human presence may result in being too expensive for many educational Institutes. This is why, several recent studies are trying to exploit advances in speech recognition technologies to automatically create captions from video material.

Off-the-shelf speech recognition software promises an accuracy of 99% when correctly trained, when used for dictating purposes and while using good quality microphones in a good acoustic environment. One may think of using such tool to automatically produce caption of a video lesson, but a classroom is a scenario very different from the dictating one, as spontaneous speech that occurs in a lecture is acoustically, linguistically, and structurally different than the one used to create written documents: the speaker talks at different speeds and different volume to emphasize some part of the speech, he/she often uses fillers (e.g. uh, er, um, ah), sometimes he/she hesitates in the middle of a word and does not speak punctuation marks ('comma', 'dot', 'question mark', etc.). Therefore, the existing speech recognition technologies are far from being completely satisfactory. However, with no doubt speech recognition technologies represent an interesting and promising solution to increase content accessibility.

W4A2012 - Communication, April 16-17, 2011, Lyon, France. Co-Located with the 21st International World Wide Web Conference.

In this paper we propose an architecture that aims at automatically creating captions of video lessons by exploiting advances in speech recognition technologies. The main motivation behind our study is to provide equal access to learning material for students of all abilities with a cost-effective solution. Our approach couples the usage of an off-the-shelf ASR software with a novel caption alignment mechanism that smartly introduces unique audio markups into the audio stream before giving it to the ASR and transforms the plain transcript produced by the ASR into a timecoded transcript. A video player synchronizes the timecoded transcript with the video material so as to make learning contents accessible to students with all kind of abilities.

This paper is organized as follows: Section 2 presents approaches in the area of captioning; Section 3 describes details of our proposal, which is analyzed through an experimental evaluation in Section 4. Conclusions are drawn in Section 5.

2. RELATED WORKS

In the literature, several studies investigate different aspects of captioning: fields of application, timing display and generation.

The primary field of application of captioning techniques is the support of impaired audience. Different studies (e.g., [2], [3]) showed that captioned videos provide better comprehension of the content for students who are hearing impaired, suggesting that visual stimuli provide essential information for viewers who are hearing impaired.

The production of captions and the alignment with the video (i.e., the synchronization) are usually considered the two most difficult challenges in captioning. Since the manual approach is time consuming and expensive, many studies (e.g., [6], [7]) are trying to find a way to automatically produce a textual transcript of a video content. Some approaches relies on speech recognition systems that automatically produces a transcript with timing information (e.g., CMUSphinx), whereas others use speech recognition technologies and a plain transcript to find when a word is pronounced (e.g., the AutoCap project [8] runs CMUSphinx to obtain a transcription and then the transcript is used to create a language model which is used to obtain a more accurate transcript with timing information).

Finally, in the field of learning accessibility, it is worth mentioning the Liberated Learning Consortium (liberatedlearning.com) and Net4Voice projects: these are international research networks dedicated to advancing speech recognition technology and techniques to create and foster barrier-free learning environments to improve accessibility.

3. OUR PROPOSAL

In most educational scenarios (e.g., Schools, Colleges and Universities) students with different abilities are experiencing problems that make their life very different from the one of other students as they have troubles in accessing to educational material. The motivations behind our proposal is to create a cost-effective scenario where technologies can be used to enhance learning accessibility of students with different abilities like hearing impaired, dyslexic or ESL students. To this aim, we propose an architecture able to assist in the automatic production of video lesson captions by exploiting advances in speech recognition technologies. We design a novel caption alignment mechanism that smartly introduces unique audio markups into the audio stream before giving it to an ASR and transforms the plain transcript produced by the ASR into a timecoded transcript.

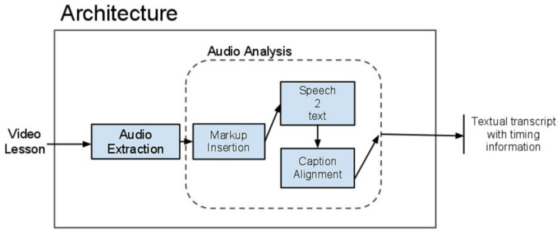

Figure 1: Architecture. Audio/Video lessons are analyzed and automatically transcribed in textual format with timing information.

From the system point of view, our proposal aims at providing a solution with the following characteristics: i) Automatic: no manual transcription or caption alignment; ii) Efficient: ASR runs just one time and iii) Technology transparent: no binding to specific technologies and methodologies (i.e., any speech recognition, multimedia and networking technology can be used without any major change to the overall architecture).

3.1 Architecture

The architecture we propose is depicted in Figure 1: it is designed to analyze and transform a video lesson into a video lesson with synchronized textual transcript. In doing this, the audio is extracted from the video lesson file and is then processed by an audio markup insertion, by a speech2text module and finally by the caption alignment module. The architecture produces a timecoded transcript that will be used by a video player (as later explained) to display captions synchronized with audio/video contents.

3.1.1 Markup Insertion

To produce a timecoded transcript it is necessary to know the time a word is spoken. Off-the-shelf speech recognition products do not provide timing information within the produced textual transcript and therefore it is necessary to design an efficient solution that inserts timing information within the transcript. In this paper we propose a novel caption alignment mechanism that works as follows: the original audio stream is coupled with unique audio markups that are automatically introduced in it (see Figure 2). The modified audio stream is then passed to the ASR and the produced transcript will also contain the transcription of the audio markup. By exactly knowing when the markups were inserted, we can produce a timecoded transcript.

As a unique markup we consider a word that is unlikely to be pronounced during a lesson: 'Goofy'. The audio format of this word is frequently inserted in the audio stream. In particular, the markup is inserted in *silence periods* (i.e., when the speaker does not speak), otherwise our approach could truncate words, resulting in the impossibility for the ASR to recognize those words. Therefore, we need to identify silences within the audio stream. In a speech stream, silence lengths can span from few ms to secs: a short silence happens very frequently as it is present even between

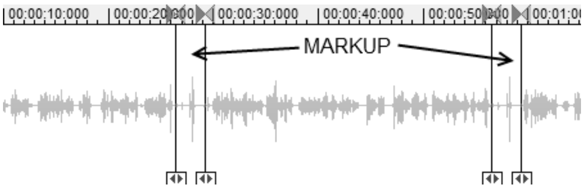

Figure 2: Automatic Caption Alignment. A unique markup is periodically inserted to retrieve timing information about the automatic produced transcript.

two consecutive syllables of a single word (in this case, if we insert the markup within the syllables, the word will be truncated); a long silence may rarely be present and therefore few markups would be inserted. Therefore, finding a reasonable silence length is very important.

After the markup insertion, the audio file is passed to the speech2text module for textual transcript generation.

3.1.2 *Speech2text*

The speech2text module is in charge of transcribing the audio stream into text. Our proposal is not bound to any specific speech technology and therefore is possible to use any given speech recognition technology.

3.1.3 *Caption Alignment*

The goal of this module is to insert timing information into the textual transcript produced by the speech2text module. As mentioned, caption alignment is usually done manually, but our goal is to automatically insert timing information. We recall here that the markup insertion module introduced in the audio stream several unique markups and we are aware of where these markups have been inserted. The transcript produced by the speech2text module does not have timing information, but it has the textual form of the audio markup (e.g., in our case the 'Goofy' word). When the 'goofy' word appears in the transcript, we substitute it with the time it was inserted. As a result, the module produces a timecoded transcript.

4. EXPERIMENTAL STUDY

To evaluate our proposal we set up an experimental scenario involving different *Computer Science* and *Linguistics* Professors of the Communication Sciences degree of the University of Modena and Reggio Emilia. The goal of the experimental study is to find the most appropriate hardware and software products to build the recording scenario, to investigate the accuracy achieved by our proposal and to tune the parameters that are used to locate the positions where to insert the audio markups.

4.1 Testbed scenario

To create a testbed scenario it is necessary to select both the speech technology for the speech2text module and the microphone to record the speech. We consider Dragon NaturallySpeaking version 11 as the ASR to be used in the speech2text module of the architecture for three main reasons: i) support for Italian language, ii) availability of speech-to-text transcription from digital audio file, and iii) easy access to the product. The professional wireless clip-on microphone (SENNHEISER FreePort FP12) is device selected for the classroom scenario.

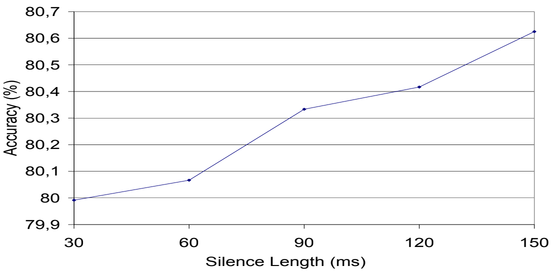

Figure 3: Accuracy achieved while varying the silence length used to insert the audio markup.

In addition, since the selected ASR is speaker dependent, the software is trained with all the professors who agree to collaborate so as to have a voice profile for each of them. In the following, we present results that are an average of the single results obtained while analyzing lessons recorded in classroom environment where Professors teach in front of a live audience.

4.2 Experimental Evaluation

The first performance parameter investigated during the experimental evaluation is accuracy. One may think that such parameter depends only on the characteristics of the ASR, but it is worth mentioning that many ASRs use context-sensitive algorithms in recognizing continuous speech; hence, a non natural speech, as the one produced by the caption alignment mechanism, may affect the achieved accuracy. In fact, the caption alignment mechanism periodically inserts audio markups into the original audio stream, transforming the natural speech of the speaker into a non natural one. Since the behavior of the ASR depends on the acoustic and language models, the audio markup insertion is likely to affect the performance of the speech analyzer. This is why, it is necessary to measure the achieved accuracy: if it drops too much, our approach is not worth using.

Figure 3 presents the accuracy results obtained from varying the length of silence where to insert the markup. The first set of experiments has been carried out by inserting the markup in speech silences that lasts a minimum of 30 ms. Then, we performed additional sets of experiments by considering longer silence (e.g., 60, 90, 120 and 150 ms). Results obtained show that there is no much difference in the measured accuracy (less than 1% of difference between a silence length of 30 ms and a silence length of 150 ms).

The goal of Figure 4 is to investigate if the presence of too many markups could influence the accuracy. Results show that the longer the minimum distance is, the better is, but it is worth highlighting that the accuracy difference between the shorter period (10 secs) and the longer one (40 secs) is within 1%. Therefore, we can state that the minimum markup distance does not greatly affect the achieved accuracy.

According to the previous investigations, it seems that the higher the values (both the silence length and the minimum distance between two consecutive markups) the better is. However, it is to note that such parameters affect the length

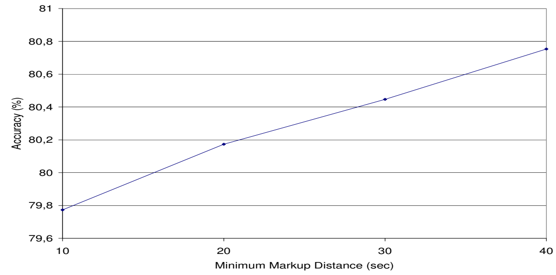

Figure 4: Accuracy while varying the minimum distance between two consecutive markups.

of the produced caption. In particular, the higher the values are, the longer the subtitles are. In fact, a caption is the textual transcript of a part of the speech located between two consecutive markups. Since our proposal is designed to support students with different abilities, captions must be displayed through few lines and with reasonable font size. In our prototype, we considered a 1024x80 area for subtitles, which leads to a maximum of 375 characters (ARIAL family font and font size of 16) per single caption.

Figure 5 presents the maximum number of characters that compose a single subtitle obtained while varying the silence length. We recall here that the maximum number of characters in our settings is equal to 375: if a subtitle is longer, the subtitle will exceed the subtitle area and therefore it becomes unreadable. Therefore, looking at Figure 5 it is possible to observe that inserting markups in silences which are 30, 60 or 90 ms long produce readable subtitles, whereas silence lengths of 120 and 150 ms produce too long subtitles.

The length of subtitles is also affected by the distance between two consecutive markups. Therefore, Figure 6 presents a similar investigation, but here we varied the minimum distance between two consecutive markups. By observing the results, it is to note that only markups with a distance of 10 or 20 secs produce subtitles with less that 375 characters.

By combining results obtained in Figures 5 and 6, the readability of subtitles is ensured with silence lengths between 30, 60 and 90 ms and with markups distance between 10 and 20 secs. By observing Figures 3 and 4 it is possible to state that the best tuning to produce subtitles shorter than 375 characters is the one with 20 secs of minimum distance between two consecutive markups and with silence length of 90 ms, as these parameters produce higher accuracy with respect to the other possibilities.

5. CONCLUSIONS

In this paper we proposed an architecture to automatically create captions from audio/video lessons material. Our approach couples the usage of off-the-shelf ASR software with a novel caption alignment mechanism that smartly introduces unique audio markups into the audio stream before giving it to the ASR and transforms the plain transcript produced by the ASR into a timecoded transcript. An experimental assessment showed that our proposal does not negatively affect the achieved transcription accuracy and that it is possible to tune the parameters of the novel caption alignment mechanism so as to ensure captions readability in the video player.

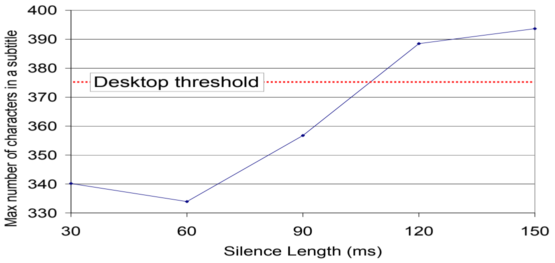

Figure 5: Maximum number of characters in a single subtitle by considering a subtitle area of 1024x80 pixel, a font size of 16 and ARIAL font-family.

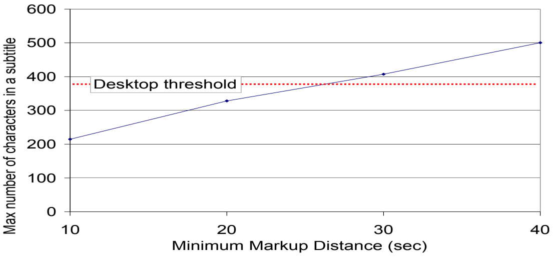

Figure 6: Maximum number of characters in a single subtitle by considering a subtitle area of 1024x80 pixel, a font size of 16 and ARIAL font-family.

6. REFERENCES

[1] Unesco Report 2005 - The quality imperative. *Global Monitoring Report*, 2005. [on-line] Available at http://www.unesco.org

[2] M. Xu, S. Yan, T-S. Chua, R. Hong, M. Wang. Dynamic captioning: video accessibility enhancement for hearing impairment. In *Proc of the ACM Multimedia Conference*, pp. 421–430, New York, NY, USA, 2010.

[3] L. Jelinek, D. Jackson. Television literacy: comprehension of program content using closed captions for the deaf. *Journal of Deaf Stud. Deaf Educ.*, Vol. 6, N. 1, pp. 43–53, 2001.

[4] T. Garza. Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals*, Vol. 24, N. 3, pp. 239-258, May 1991.

[5] S. Tsuboi, N. Shimogori, T. Ikeda. Automatically generated captions: will they help non-native speakers communicate in english? In *Proc of Intercultural collaboration conference*, ICIC '10, pp. 79–86, New York, NY, USA, 2010. ACM.

[6] G. Penn, E. Toms, D. James, C. Munteanu, R. Baecker. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pp. 493–502, New York, NY, USA, 2006. ACM.

[7] M. Wald. Crowdsourcing correction of speech recognition captioning errors. In *Proc of the W4A Conference*, New York, NY, USA, 2011. ACM.

[8] A. Knight, K.C. Almeroth. Fast caption alignment for automatic indexing of audio. *International Journal of Multimedia Data Engineering and Management*, Vol. 1, N. 2, pp. 1–17. June 2010.