On the use of automatic speech recognition for TV captioning

Jordi Robert-Ribes

CSIRO-MIS & ACSys
Locked Bag 17, North Ryde 1670, Australia
jordi.robert-ribes@cmis.csiro.au

ABSTRACT

This study analyses the possible use of automatic speech recognition (ASR) for the automatic captioning of TV programs. Captioning requires: (1) transcribing the spoken words and (2) determining the times at which the caption has to appear and disappear on the screen. These times have to match as closely as possible the corresponding times on the audio signal. Automatic speech recognition can be used to determine both aspects: the spoken words and their times. This paper focuses on the question: would perfect automatic speech recognition systems be able to automate the captioning process? We present quantitative data on the discrepancy between the audio signal and the manually generated captions. We show how ASR alone can even lower the efficiency of captioning. The techniques needed to automate the captioning process are presented.

1. INTRODUCTION

In this section we will introduce captioned TV. The next section will present the relevant aspect of the captioning process. Section 3 will present the analysis of four captioned TV programs. This sets the framework for the discussion on the use of ASR in the captioning process (section 4). Finally, Section 5 will present our conclusions.

1.1. Captioned TV

TV programs can have captions, which show the soundtrack of the program as text on the TV screen. Captions are coloured and positioned on the screen to show each character's speech. Sound effects, music and other audio cues are incorporated in the captions to ensure all relevant information is available to the viewer. This is the main difference between captions and subtitles, because subtitles just show a translation of the dialogue into another language. Figure 1 shows two examples of TV captions.

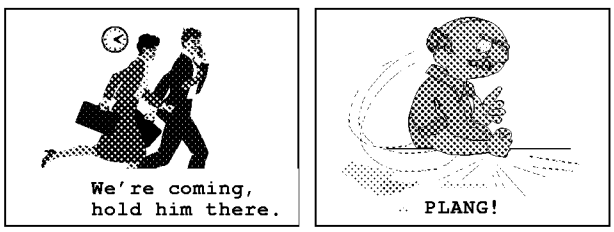

Figure 1: Examples of caption for: (1) speech signal and (2) some relevant noise (this text would not be subtitled)

What is captioned TV used for? Deaf people rely solely on captions to follow a TV program or commercial. Hearing impaired people with some degree of hearing use captions to help understand the soundtrack (see [3] for a recent study on this topic). However, not only people with hearing disorders use captions. Captions are also widely used to watch TV programs in noisy environments (sporting clubs, pubs, etc) and by people learning a foreign language.

1.2. Importance of captioned TV

Captioned TV is becoming an important part of the broadcasting industry and will grow in the coming years. The new systems of digital TV (DTV) have taken captions into consideration from the start of their development. Captions are not just an add-on into systems but they are built into the system and will evolve with it. The new standard EIA-708 for DTV captioning is a good example of this ([10], [11]).

The convergence between television sets and home computers is becoming a reality. It is possible nowadays to surf the Web from the TV set and to watch TV programs using the computer. This has made Microsoft propose the Synchronised Accessible Media Interchange standard [4]. This standard will allow any media or multimedia file to be captioned.

The importance of captioned TV is not only shown in the technical domain but also in the political one. In Australia, in July 1998 the Commonwealth Parliament passed the *Television Broadcasting Services (Digital Conversion) Act 1998* [12] which will require all networks to provide closed captions on all prime time programs and all news and current affairs programs by 1 January 2001.

All these factors will bring an increasing demand for captioned programs. This will generate a need for captioning tools that allow easy and fast captioning of TV programs.

1.3. Possible use of ASR

Automatic Speech Recognition (ASR) can be used in several ways to assist captioning. There are two main areas for the use of ASR: automatic time alignment of captions and automatic generation of captions.

We will consider a theoretical system with perfect ASR (that is 100% recognition rate). For automatic timing of captions, such a system would align the audio signal with some text manually entered by the captioner. For automatic captioning, the system would transcribe the audio track and automatically generate the captions. Both cases would need a second step for manual verification and refinement or adaptation of captions.

The efficiency of such systems would depend on the number of adaptations to be made during the manual step. It may even be the case that the number of manual corrections is so high that it is inefficient to use such systems.

In the next section we will look into the relevant aspects of the captioning process. We will then analyse some captioned TV programs. This will set the framework for the discussion on the use of ASR in the two areas mentioned above.

2. THE CAPTIONING PROCESS

For some TV programs a transcript exists and can be used as the base for the captions [1]. However, this is not always the case and many times the captioners need to listen to the audio and type the corresponding text. In the following we will only consider the latter case; the former case can be treated as a sub-case of this general one.

There are three types of constraints when generating captions: timing, meaning and reading speed. These three constraints interact strongly. The timing constraints require the caption to appear and disappear as close as possible to when the corresponding audio starts or ends. The meaning constraints require that the text to be as verbatim (exact transcription) as possible. The reading speed constraints impose an upper limit to the number of text words per second to ensure the readability of the text displayed on the screen (normally 3 words/second).

For each caption, the captioner normally locates the start time, while typing the caption text and establishing the end time. These tasks are interdependent and the captioner approaches them in a holistic way. To ensure that readability constraints have been met: (1) the caption start and end times may be modified and (2) the text may be reduced. Time modification must follow certain rules (eg. a caption looks better if it starts and stops within the video shot boundaries) and text reduction must not change the meaning (but can use acronyms or change the wording, for instance). Furthermore, each caption text and its associated times are highly dependent on the preceding and following captions. The captioner often needs to change some of the captions already entered when it is realised that the particular characteristics of the current caption have a significant effect on previous captions (for instance, the existence of a video shot boundary).

Stress in the speech signal is also transcribed, for instance with capital letters (for example "`You DON'T love me`").

Timing of the captions with the corresponding audio is not always the most important factor to take into account. For instance, it is often preferred that captions do not go over scene boundaries and that they do not start or end too close to a scene boundary (to avoid a visual flickering effect).

Captions for noises or music (such as "`PLANG!`" or "`SOFT GUITAR MUSIC`") may not need an accurate timing. For instance a caption for music may appear on the screen when the music starts playing but may not stay on display for the whole length of the music.

3. ANALYSIS OF SAMPLE PROGRAMS

Any system that tries to automate the captioning process or the timing of captions will have to take into account the constraints detailed in section 2.

In this section we analyse four sample programs provided by the Australian Caption Centre (ACC) [2]. We quantify the difference between the captions generated manually by the ACC and the actual audio signal. We will quantify the differences in the wording and the timing. This will help us estimate the applicability of ASR technology to the captioning process.

Each of the sample programs is of a different type: news, magazine, documentary and drama. We had for each of these programs a manually generated captions file containing the captions and for each of them the starting and ending times.

3.1. Text reduction

In the four programs used in our testing, text reduction occurred in a considerable number of captions. Figure 2 shows the percentages of captions having a considerable reduction in the text. Small text changes are not taken into account. For instance, "Yeah" captioned as "`Yes`" would not be taken into account in Figure 2. An example of change that has been taken into account is "That story a little later" being captioned as "`That story later`".

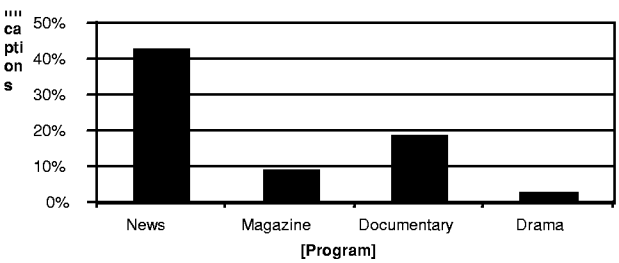

Figure 2: Percentage of non-verbatim captions

The high percentage (>40%) of reduced captions for the news program is because the intended audience was children. Therefore the reading constraints are much tighter (2 words/second) and consequently more captions need to be reduced. The drama program was captioned in the US where reading constraints are looser and less reduction is performed, explaining the low percentage (<5%) of captions with text reduction. In a similar Australian program, reduction occurs in about 20% of the captions if captioned by the Australian Caption Centre.

We also present the number of captions for non-speech sounds (such as noises or music) because it will be impossible to find the exact words of the captions in the audio signal. Figure 3 shows the percentage of captions for non-speech signals. These include noises (for instance "`PLANG!`"), music (for instance "`SOFT GUITAR MUSIC`") or names of the speaker (for instance "`JOHN: Here she comes`").

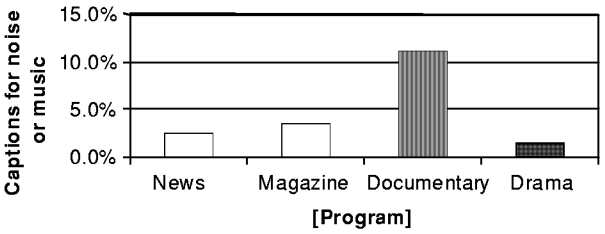

Figure 3: Percentage of captions for non-speech sounds

3.2. Difference in timing

For the four sample programs, we compared the start and end times of the captions with the start and end times of the corresponding segment in the audio track. Figure 4 shows the results of such comparisons for the start times. The horizontal axis shows the difference between the time start of the caption and the time start of the audio. The vertical axis is the percentage of captions with a difference corresponding to the value of the horizontal axis or better. For example, for the news program, 35% of the captions start at most 0.2 seconds apart from the speech signal. Figure 5 shows the data for the end times, and Figure 6 where both times are taken into account.

Overall, we can see that the start times are normally better aligned with the corresponding audio than the end times. In most of the cases this means that the caption is better synchronised when it appears on the screen than when it disappears.

Note, the child audience may explain the poor synchronisation of the news program.

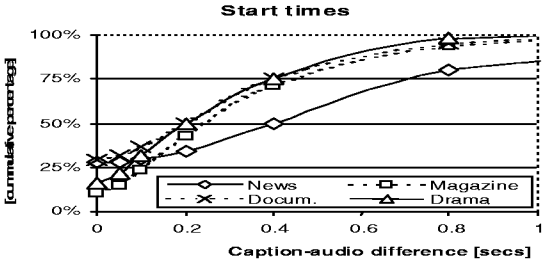

Figure 4: Start times, difference between captions and audio

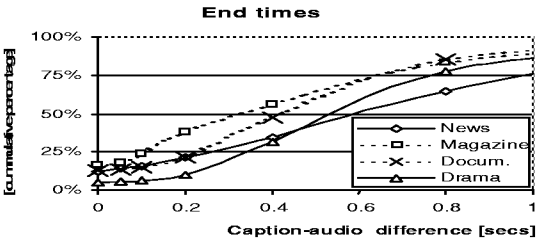

Figure 5: End times, difference between captions and audio

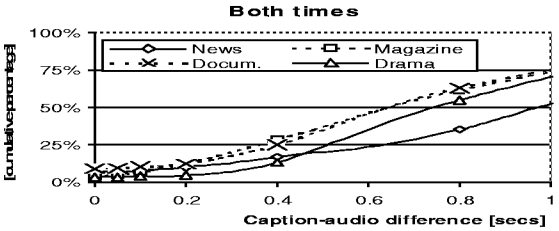

Figure 6: Both times, difference between captions and audio. This figure takes into account the maximum difference of start or end times

In some cases, the captioner does not intend there to be any difference. This happens mainly in slow segments of the program without extreme action. In such cases, the timing may not have the same importance as in more active segments. However, for most of the cases, the difference is intentional. For instance, this is the case for captions that would start or end too close to a video scene cut.

3.3. Background noise

Any background noise affects the automatic speech recognition systems currently available (see for instance [5], [6] and [8]). One common solution has been the use of close-microphones in dictation systems. Unfortunately this solution is not applicable to the case of TV programs. The research effort to develop systems that can be used for robust speech recognition with high accuracy even in the presence of background noise is considerable. However, a complete solution to this problem is not expected in the near future.

In our four sample programs, the number of captions for which there was considerable music or noise in the audio background was very high (see Figure 7 for details).

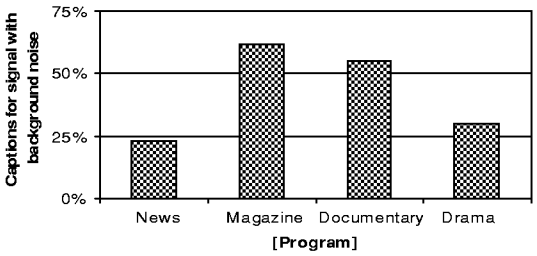

Figure 7: Percentage of captions with background noise

4. USE OF ASR FOR CAPTIONING

As stated above, we see two main areas where ASR could be used: automatic generation of captions and automatic alignment of captions. Such automatic systems would require some type of post-process manual verification and adaptation. From the data presented in section 3, it appears that the manual adaptations that would need to be made in such a second manual step would be considerable.

4.1. Automatic timing of captions

When automatically aligning the text manually entered by a captioner against the audio signal, several problems have to be considered.

The high number of captions for which there is a significant difference in the timing of the caption and the audio track (see Figure 4, Figure 5 and Figure 6) has been noted. As mentioned in section 3.2, in some cases the difference is unintentional. In these cases, the accurate timing of a perfect ASR would not pose any problem. On the contrary, it would help the captioner to produce highly accurate timing for the captions.

For most of the cases, the difference is intentional. In such cases, the accurate times given by a perfect ASR system would pose some problems. This suggests that a more complex system incorporating video processing techniques and artificial intelligence techniques would be needed. The video processing techniques would automatically detect the video cuts ([9]) and the artificial intelligence would be used to include knowledge of the video cut locations into the timing of the captions.

In contrast, captions representing noises or music would not be aligned by recognising the words in the audio track. A system for recognising and labelling noises and music would be needed for such task.

4.2. Automatic generation of captions

In addition to the considerations of section 4.1, there is a need for text reduction in a considerable number of captions. This would entail the use of natural language processing technology. Presently, some techniques have been developed to summarise text (see for instance [7]), but the reduction of short segments of text (such as one caption) has not received appropriate attention.

Transcribing the stress of the speech signal would need a powerful system of intonation recognition. Such a system is not available at present.

All the above considerations have assumed a perfect ASR system that performs at nearly 100% accuracy for all conditions (including noise or music in the background). However, current ASR systems are far away from this goal (see for instance [5], [6] and [8]). Figure 7 has shown that background noise and music are common in TV programs.

Even if the 100% recognition rate was achievable for signals with background noise or music, the total number of non-verbatim captions and captions for non-speech sounds may be up to 45% of the captions for some programs (see Figure 2 and Figure 3). This means that a system using perfect ASR, would still require manual adaptation in up to 45% of the captions.

5. CONCLUSIONS

Automatic Speech Recognition on its own is not enough to automate the timing of manually typed captions or to completely automate the captioning process. Such a system would need (among others):

- Improved ASR techniques to deal with background noises and music.
- Improvement and use of natural language processing techniques for text reduction.
- Use of video processing and artificial intelligence to incorporate knowledge of the video cut locations.
- Development and use of music and noise type recognition (for captioning relevant non-speech signals).

6. REFERENCES

1. 'Steps in the Captioning Process', http://www.gallaudet.edu/~tvweb/st.html
2. Australian Caption Centre, http://www.auscap.com.au/
3. Burnahm D. and Robert-Ribes J. (1998). "Why captions have to be on time". Proc of AVSP'98 (Terrigal, Australia)
4. http://microsoft.com/presspass/press/1997/jun97/SAMI.htm
5. Woodland, P. C et al. (1997). "Broadcast news transcription using HTK". ICASSP'97; Munich (Germany).
6. Cook, G. D. et al. (1997). "Transcription of broadcast television and radio news". ICASSP; Munich (Germany).
7. Dras, M. (1997). "Reluctant Paraphrase: Textual Restructuring under an Optimisation Model." PACLING, Ohme, Japan, pp. 98-104.
8. Bakis, R. et al. (1997). "Transcription of broadcast news". ICASSP'97; Munich (Germany).
9. Gu L., Tsui K., and Keightley D. (1997), "Dissolve detection in MPEG compressed video". IEEE Conference on Intelligent Processing Systems, pp 1692-1696.
10. "Closed Captioning Standards for Digital Television Established" http://www.cemacity.org/gazette/files/closecap.htm
11. Hutchins J. & Robson, G. (1998). "The Advantages and Pitfalls of Closed Captioning for Advanced Television". http://www.robson.org/gary/writing/icce98.html
12. http://SCALEplus.law.gov.au/html/pasteact/2/3156/top.htm

Acknowledgments:

The author wish to acknowledge the help and assistance provided by R. Ellison, B. Kim, G. Reynolds, T. Vu, K. Webb and all those not mentioned.