Chapter 12
Fast Caption Alignment for Automatic Indexing of Audio

Allan Knight
University of California, Santa Barbara USA

Kevin Almeroth
University of California, Santa Barbara, USA

ABSTRACT

For large archives of audio media, just as with text archives, indexing is important for allowing quick and accurate searches. Similar to text archives, audio archives can use text for indexing. Generating this text requires using transcripts of the spoken portions of the audio. From them, an alignment can be made that allows users to search for specific content and immediately view the content at the position where the search terms were spoken. Although previous research has addressed this issue, the solutions align the transcripts only in real-time or greater. In this paper, the authors propose AutoCap. It is capable of producing accurate audio indexes in faster than real-time for archived audio and in real-time for live audio. In most cases it takes less than one quarter the original duration for archived audio. This paper discusses the architecture and evaluation of the AutoCap project as well as two of its applications.

INTRODUCTION

Over the past 10 years, automatic speech recognition has become faster, more accurate, and speaker independent. One tool that these systems rely on is *forced alignment*, the alignment of text with speech. This application is especially useful in automated captioning systems for video play out. Traditionally, forced alignment's main application was training for automatic speech recognition. By using the text of recognized speech ahead of time, the Speech Recognition System (SRS) can learn how phonemes map to text. However, there exist other uses for forced alignment.

Caption alignment is another application of forced alignment. It is the process of finding the

DOI: 10.4018/978-1-4666-1791-9.ch012

exact time all words in a video are spoken and matching them with the textual captions in a media file. For example, closed captioning systems use aligned text transcripts of audio/video. The result is that when the audio of the media plays, the text of the spoken words is displayed on the screen at the same time. Finding such alignments manually is very time consuming and requires more than the duration of the media itself, i.e., it cannot be performed in real-time. Automatic alignment of captions is possible using the new generation of SRS, which are fast and accurate.

There are several applications that benefit from these aligned captions. Foremost, and quite obviously, are captions for media. Providing consumers of audio and video with textual representations of the spoken parts of the media has many benefits. Other uses are also possible. For example, indexing the audio portion of the media is a useful option. By aligning media with the spoken components, users can find the exact place where text occurs within the audio content. This functionality makes the media searchable.

The technical challenge is how to align the transcript of the spoken words with the media itself. As stated before, manual alignment is possible, but requires a great deal of time. A better solution would be to find algorithms to automatically align captions with the media. There are, however, several challenges to overcome in order to obtain accurate caption timestamps. The first is aligning unrecognized utterances. No modern SRS is 100% perfect, and therefore, any system for caption alignment must deal with this problem. The second challenge is determining what techniques to apply if the text does not exactly match the spoken words of the media. This problem arises if the media creators edit transcripts to remove grammatical errors or other types of extraneous words spoken during the course of the recorded media (e.g., frequent use of the non-word "uh"). The third challenge is to align the caption efficiently. For indexing large archives of media, time is important. Therefore, any solution should balance how much time it takes with the greatest possible accuracy.

The work discussed in this paper is part of a project called AutoCap. The goal of this project is to automatically align captured speech with their transcripts while directly addressing the questions above. AutoCap includes of two previously available components: a language model toolkit and a speech recognitions system. By combining these components with an alignment algorithm and caption estimator, developed as part of this research, we are able to achieve accurate timestamps in a timely manner. Then, using the longest common subsequence algorithm and local speaking rate, AutoCap can quickly and accurately align long media files that include audio (and video) with a written transcript that contains many edits, and therefore, does not exactly match the spoken words in the media file.

While other researchers have previously addressed a similar problem (Hazen, 2006; Moreno & Jeorg, 1998; Placeway & Lafferty, 1996; Robert-Ribes & Mukhtar, 1997), they use different techniques and do not accomplish the task as fast as AutoCap can. The cited projects either do more work than is needed, such as a recursive approach (Moreno & Joerg, 1998), or add more features than are needed (Hazen, 2006), for example, correcting the transcripts. In either case, both approaches, while very accurate, take real-time or longer to align each piece of media. And as mentioned previously, for processing large archives of media, shorter processing times are critical. Finally, and most importantly, these works do not address the issue of edited transcripts.

Our research shows that AutoCap can accurately and efficiently align edited transcripts. AutoCap's accuracy, as measured by how closely aligned the spoken words are with when the text appears on the screen, is well within two seconds of the ground truth. This two second value is what other research cites as the minimum level of