SyncWords: A Platform for Semi-Automated Closed Captioning and Subtitles

Aleksandr Dubinsky
SyncWords LLC., USA
alex@syncwords.com

Abstract
As progressing social mores and new legislation mandate greater levels of accessibility across broadcast television, online video, education, and the workplace, it becomes ever more important to deliver captioned content at lower cost, with higher quality, and more quickly. To reach those goals, attention is shifting toward automation. While many attempts have been made to apply large vocabulary ASR transcription for this task, the results on noisy, post-processed, multi-speaker recordings are insufficient for most real-world applications, to say the least. More viable is the automated synchronization of a human-prepared transcript with the media, and automating the breaking up of the text into well-structured, meaningful captions. SyncWords is a platform that incorporates innovative, originally developed algorithms for both of these tasks, and presents an easy-to-use commercial Web interface for submitting work and receiving captions. In addition, an interactive review tool is available in which user corrections are fed back and cause a re-evaluation of the results in real time.

Index Terms: long alignment, closed captioning

1. Description
SyncWords [1] is an online service that lowers the cost of closed captioning and subtitling. Users are able to upload a pre-prepared transcript along with a media file and receive a caption file in one of several popular formats. (Human transcription services are also offered.) Under the hood, a proprietary long alignment algorithm synchronizes the text to the audio. Audio files up to 5 hours in duration are supported, and may include sound effects, distortions, and non-speech regions of any length. Following alignment, chunking is performed to produce caption or subtitle files that attempt to optimize a number of criteria recognized to be important for readable captions [2], such as placing line and caption breaks at grammatically opportune times, respecting heuristics regarding minimum and maximum on-screen duration, and obeying line length limits that are enforced by certain closed caption formats. Many of the parameters, such as maximum line length, the visibility of speaker labels, and the closing of gaps between closely-spaced captions are user-configurable. Finally, the captions are encoded in formats that are used in broadcast television production and on the web.

Of particular note is a feature that allows the per-word timings to be previewed and manipulated in an interactive web-based application. The user is presented with the transcript, a display of the sound waveform, and controls for playback and editing. As the audio plays, words are highlighted in sync. Crucially, the user can edit the start or end time of a mistimed word or its text. When this happens, a request is sent back to SyncWords specifying the new constraint. The alignment algorithm uses this constraint to re-align surrounding text. Updated timing information is sent back and displayed to the user within a few seconds.

Compared to some earlier approaches to long alignment, our approach is highly robust and does not suffer from the problem of error accumulation. The acoustic model uses deep neural networks with rectifier neurons, as recently demonstrated to perform well on this task [3]. Rectifier networks do not require pre-training, and train well with ordinary back-propagation through many layers. We use a 5-hidden-layer network of 896 units in each layer that takes as input 31 frames of mel-filtered speech features to predict the posterior probability of 6144 context-dependent tied-state senones. Training was done on a 550-hour corpus representative of material found in broadcast television as well as lectures and conferences. The alignment algorithm is proprietary. The alignment is performed approximately at twice real-time on a single CPU core. The implementation is written in Java and CUDA C. The service is hosted in Amazon’s “cloud” virtual machine rental service.

2. Evaluation
We evaluate the accuracy of alignment of SyncWords, along with YouTube’s alignment service [4], another commercial competitor, and professionally-created (presumably human-generated) subtitles. For the evaluation we chose four full-length recordings, totaling 3.7 hours. These recordings consisted of an episode of Cheers [5] which first aired in 1985, to give an example of medium-difficulty archival footage; an episode of Teen Wolf [6] which aired in 2011, to give an example of a modern high-production-value TV program containing pervasive adverse audio conditions such as reverberation and background music; the feature film The Lost World: Jurassic Park [7]; and a recording of a presentation.

The following methodology was used for evaluation: professionally-created subtitles were taken of the recordings, and stripped of non-spoken text such as speaker identification labels and sound effect cues. These transcripts, which retained their line-breaks, were submitted for alignment to the three services. To eliminate remaining differences in chunking among the returned subtitle files, only the captions which started on the same word for all subtitle files were kept and the
others discarded. Among the kept subtitles, a random set was chosen and the start of the first word was timed by an operator using an internally-developed software that featured playback controls and a spectrogram display. The initially presented start time of these words was randomized so that the operator had to make some adjustment in every case. The start time of the captions in which these words appeared were compared against those timed by the operator. For each recording, between 123 and 200 data points were used. The results are presented as the percent of captions that had an error less than one of several thresholds.

We evaluated timing at the caption level, because word-level timing was not available for YouTube’s output, and because it is more representative of real-world requirements. The line-breaks were retained in the transcripts to “hint” the various chunking algorithms and increase the similarity of chunking of the returned subtitle files and the number of captions they had in common. We do not believe the extra line-breaks affected the alignment results of any of the services, although this was not verified. The transcripts were stripped of non-spoken textual elements in order to eliminate differences in how the different services might handle (or fail to handle) them. At least one of the evaluated services (that of the competitor) claims to use speaker labels to perform speaker-dependent adaptation (if these speaker labels conform to a strict typographical style). However, speaker labels were sparse and inconsistent in the two recordings that had them (e.g., often represented by a dash rather than the actual name of the speaker), and we felt it was reasonable to omit them. The start times (as opposed to the end times) of the captions were evaluated because the end times are more often purposefully modified. For example, captioners extended the end times of captions to remove gaps between closely-spaced successive captions. The lowest precision threshold is 250 ms because it is more representative of real-world requirements. This is echoed by new, more-stringent regulation from the FCC and other government bodies, which mandates access. This is more imperative to offer the deaf and hard-of-hearing greater access. As more educational material is being delivered in video form, such as on YouTube, it becomes ever more imperative to offer the deaf and hard-of-hearing greater access. This is echoed by new, more-stringent regulation from the FCC and other government bodies, which mandates captioning not just on broadcast television but also online. SyncWords is an effective solution to reduce the workload of creating captioning.

SyncWords offers superior alignment performance for challenging content. For the full-length feature film, major alignment errors were reduced by 3- to 5-fold compared to other automated alignment services. However, performance was slightly worse for less-challenging material when evaluated at high thresholds of precision. Possible reasons include training on a corpus that was over-representative of challenging audio, an inherent limitation of the algorithm, or an undetected bug in the implementation. We will work to address this.

Of note is the accuracy of human captioning. As can be expected, large errors are infrequent. However, small errors are prevalent for some of the programs. This may be due to the prohibitive expense of meticulously aligning each caption. Alternatively, some of the errors may be purposeful, such as to allow for more time for a caption to be read. The timings may also reflect sound effect cues which were subsequently deleted and not timed during the evaluation. Nevertheless, it appears that the more recent program had significantly better captions, pointing to increasing attention to quality.

4. Conclusions
SyncWords isn’t perfect, but it has large potential for reducing the cost of captioning and subtitling and represents a solid step in the state of the art. As more educational material is being delivered in video form, such as on YouTube, it becomes ever more imperative to offer the deaf and hard-of-hearing greater access. This is echoed by new, more-stringent regulation from the FCC and other government bodies, which mandates captioning not just on broadcast television but also online. SyncWords is an effective solution to reduce the workload of creating captioning.

5. References
[1] “SyncWords” Internet: www.syncwords.com