August 22, 2016

COLLABORATIVE DISTRIBUTED SPEECH RECOGNITION

Sara Basson

Dimitri Kanevsky

COLLABORATIVE DISTRIBUTED SPEECH RECOGNITION

ABSTRACT

The disclosure includes a captioning system configured to caption a video. A video may be identified for captioning. The video may be submitted to automated speech recognition engines. A transcription of audio from the video may be received from the automated speech recognition engines. It may be determined whether to accept or create a final transcription. If not, the video may be submitted to one or more manual speech recognition engines. If the final transcription is accepted or created, at least one of the automated speech recognition engines or one of the manual speech recognition engines may be rewarded based on the transcriptions. The video may be captioned with the final transcription.

KEYWORDS

- video captioning
- transcription
- hearing impaired
- crowdsourced speech recognition

BACKGROUND

Videos are often published on websites without high quality captioning included. This may present a problem for disabled users, such as users that have experienced hearing loss, and for users that may benefit from reading the captions, such as non-native speakers. Captioning systems exist; however, the quality of the speech recognition used by the captioning systems is unpredictable and depends on acoustic characteristics of the video. Low quality

captioning may not be useful, particularly if the user depends on the captioning to gain understanding of the content of the video.

DESCRIPTION

Figure 1 illustrates a diagram of an example captioning system 100 that includes a video server 101, an arbitration server 104, automated speech recognition engines 107, manual speech recognition engines 108, user devices 115a, 115n, and a network 105.

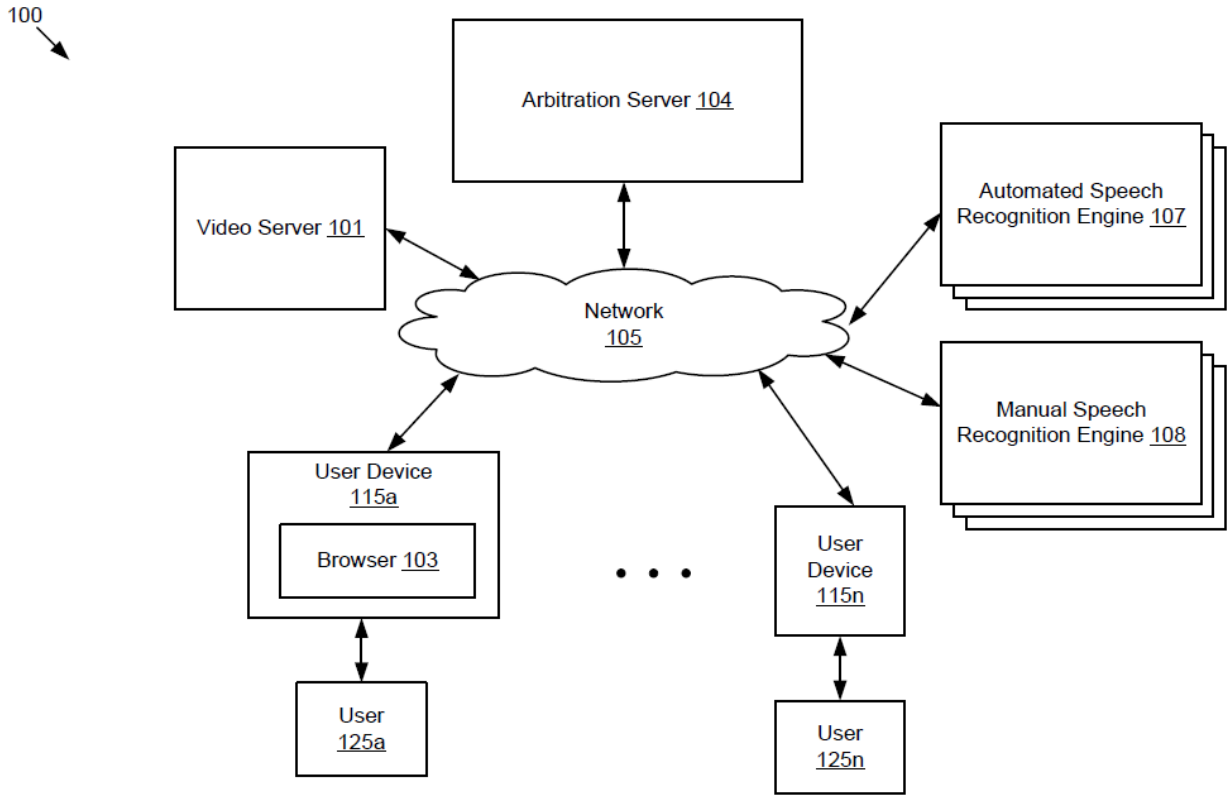

FIG 1

The video server 101 may be a hardware device that includes a processor, a memory, and network communication capabilities for accessing the network 105. The video server 101 may include software configured to publish videos, receive requests for videos from user devices 115, provide videos to the user devices 115 responsive to the request, etc. The video server 101 may store the videos. The videos may include content without captions. The videos may include

captions that may be accurate or inaccurate. The video server 101 may receive a captioned video from the arbitration server 104 and may replace the video with the captioned video.

The arbitration server 104 may be a hardware device that includes a processor, a memory, and network communication capabilities for accessing the network 105. Although the arbitration server 104 is illustrated as being a separate server from the video server 101, in some implementations they may be part of the same server. The arbitration server 104 may receive a request for captioning a video from the user 115 or a video to be captioned from the video server 101. The arbitration server 104 may also receive a video with captions that are a low quality transcription of the audio.

In some implementations, the arbitration server 104 may determine attributes of audio in the video including accents (e.g., southern accent, a French accent, etc.), gender (e.g., male or female), a level of environmental noise, speech over a telephone, channel characteristics, speech over broadband, a level of a person's speech (e.g., quiet or loud), etc.

The arbitration server 104 may submit a video to automated speech recognition engines 107. For ease of illustration, references to the video may include portions of the video. The arbitration server 104 may send the same portion of the video to the automated speech recognition engines 107 or different portions to different automated speech recognition engines 107.

In some implementations, the arbitration server 104 may determine which of the automated speech recognition engines 107 to transmit the video to, based on the attributes. For example, a first automated speech recognition engine 107 may be particularly adept at recognizing French accents and a second automated speech recognition engine 107 may excel at detecting voices in a noisy environment. The arbitration server 104 may divide the video into

different portions based on different attributes. For example, a speech recognition engine 107 that excels at detecting quiet voices may receive a portion of the video where the voices are quiet and not the whole video.

The ability of one of the speech recognition engines 107 to perform well in conjunction with a particular attribute may be referred to as voting power. A speech recognition engine may have a higher voting power for a particular type of attribute than another speech recognition engine. Voting power of a particular speech recognition engine may increase if it performs more accurate transcriptions than other speech recognition engines. In some implementations, the voting power may be expressed as weights.

The arbitration server 104 may receive a transcription of audio from the automated speech recognition engines 107. The arbitration server 104 may compare the transcriptions. The arbitration server 104 may receive different portions of the video from the automated speech recognition engines 107 and compare them accordingly. For example, a first speech recognition engine 107 may provide a full transcription, a second speech recognition engine 107 may provide a transcription of the first two minutes of the video, and a third speech recognition engine 108 may provide a transcription of the last five minutes of the video. The arbitration server 104 may determine where the three transcriptions overlap and compare the text in the overlapping areas to each other.

The arbitration server 104 may determine whether to accept or create a final transcription based on a level of agreement between the automated speech recognition engines 107. The final transcription could be one of the translations from an engine or an amalgam of multiple translations.

The arbitration server 104 may accept a transcription that is provided by a threshold number or percentage of automated speech recognition engines 107 or that has a threshold percentage of similarity between the automated speech recognition engines 107. In some implementations, the arbitration server 104 may determine whether a transcription of a portion of the video is consistent within the context of the entire video. For example, if the transcription includes geological terminology when the rest of the video is about rock music, the arbitration server 104 may not accept the transcription. In some implementations, the arbitration server 104 may accept transcriptions from certain automated speech recognition engines 107 based on voting power associated with the automated speech recognition engine 107.

In some implementations, the arbitration server 104 may send the video to additional manual speech recognition engines 108. For example, arbitration server 104 may first send the video to automated speech recognition engines 107. If the transcriptions are of poor quality or if the arbitration server 104 is generating training data, the arbitration server 104 may transmit the video to one or more manual speech recognition engines 108. In another implementation, the arbitration server 104 may send the video to the manual speech recognition engines 108 at or around the same time as the arbitration server 104 sends the video to the automated speech recognition engines 107 and make the video available for transcription for a certain period of time, such as one hour.

The manual speech recognition engines 108 may take different forms. For example, the manual speech recognition engines 108 may include accessing human transcribers through a service where users are paid a fee for performing transcription services. In another example, the manual speech recognition engines 108 may include a crowdsourcing service where a crowd of members vote on different transcriptions. The crowdsourcing service may be particularly helpful

in instances where there is little or no consensus across the automated speech recognition engines 107, such as when names, terminology, or words spoken in a very noisy environment are included in the video. The services may include a user interface that provides users with editing tools for listening to the video and providing transcriptions or corrections to the output of automated speech recognition engines 107.

In some implementations, the arbitration server 104 may update the voting power for one or more automated speech recognition engines 107 based on a transcription from the manual speech recognition engine 108. For example, if the transcription from the manual speech recognition engine 108 contradicts the transcription generated by an automated speech recognition engine 107, the arbitration server 104 may decrease the voting power of the automated speech recognition engine 107.

Once the arbitration server 104 accepts or creates the final transcription, the arbitration server 104 may reward at least one of the automated speech recognition engines 107 and the manual speech recognition engines 108 based on the transcriptions. The reward may be access to a full transcription, a monetary reward, or higher voting power. The arbitration server 104 may reward the automated speech recognition engines 107 that provide a significant portion of the transcription with the full transcription, which may be used as training data for the automated speech recognition engines 107. The arbitration server 104 may also provide the automated speech recognition engines 107 and the manual speech recognition engines 108 with information about their accuracy and their contribution to the overall accuracy of a transcription. This information may be kept confidential and made available only to the corresponding automated speech recognition engine 107 or manual speech recognition engine 108.

The arbitration server 104 may caption at least the portion of the video with the final transcription. In implementations where the video included low-quality captions, the arbitration server 104 may replace the low-quality captions with captions from the final translation. The arbitration server 104 may transmit the captioned video to the video server 101 to replace the uncaptioned or poorly captioned video.

The automated speech recognition engines 107 may be code and routines for transcribing of the portion of the video. The automated speech recognition engines 107 may apply different recognition algorithms. In some implementations, one of the automated speech recognition engines 107 may apply a different recognition algorithm depending on attributes of the video. Automated speech recognition engines 107 may use different speech recognition techniques and may be implemented in hardware, software, or a combination.

The manual speech recognition engines 108 may be code and routines for generating user interfaces that include editing tools to provide to users. A user may enter a transcription of at least a portion of a video into the user interface and the manual speech recognition engine 108 may transmit the transcription to the arbitration system 104.

The user devices 115a, 115n may be computing devices that each include a memory and a processor. The user devices 115a, 115n are communicatively coupled to the network 105. Users 125a, 125n interact with the user devices 125a, 115n, respectively. The user devices 115a, 115n may access the video server 101 and the arbitration server 104 via the network 105.

A user 125 may select a video from the video server 101 or directly request captioning from the arbitration server 104. For example, the user 125 may access the video server 101 via a browser 103 stored on the user device 115. In another scenario, the user device 115 may include a thin-client application that communicates with the video server 101.

The network 105 can be a conventional type, wired or wireless, and may have numerous different configurations.

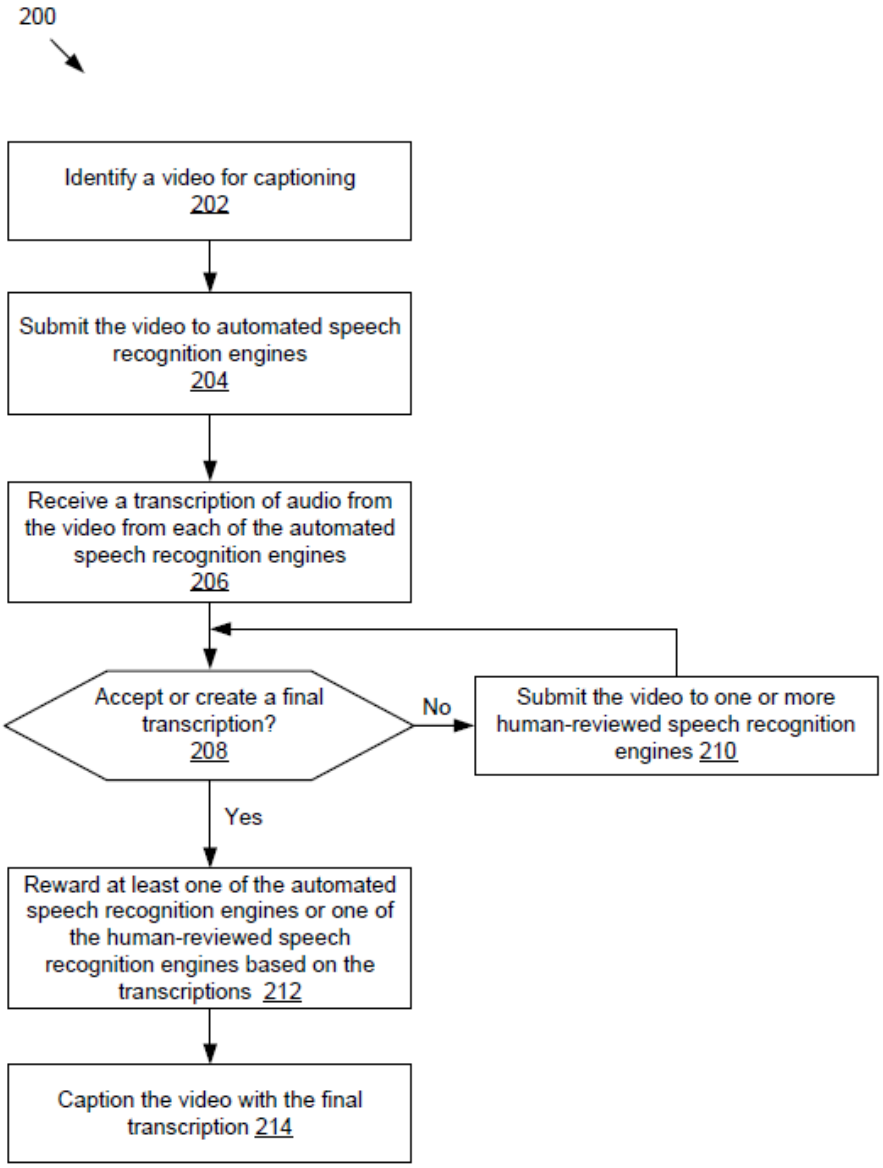

FIG 2

Figure 2 is a flowchart of an example method 200 of captioning a video. At block 202, a video is identified for captioning. For example, a user 125 may access a user device 115 to request that a video published by the video server 101 be captioned. The arbitration server

104 may identify at least the portion of the video to be transcribed. At block 204, the video may be submitted to automated speech recognition engines. For example, the arbitration server 104 may submit different portions of the video to different automated speech recognition engines 107. At block 206, a transcription of audio from the video from the automated speech recognition engines is received. The transcriptions may be the same, different, overlapping, completely different, etc.

At block 208, it may be determined whether to accept or create a final transcription. For example, the arbitration server 104 may compare different transcriptions and select one of the transcriptions that has the most agreement among the automated speech recognition engines 107. Responsive to the transcriptions from the automated speech recognition engines not being satisfactory, at block 210 the video may be submitted to one or more manual speech recognition engines. Block 208 may be repeated until the final transcription is accepted or created. Responsive to the final transcription being accepted or created, at block 212 at least one of the automated speech recognition engines 107 and the manual speech recognition engines 108 is rewarded based on the transcriptions. For example, the engine that provides all or a portion of the final transcription may be assigned a higher voting power. At block 214 the video is captioned with the accepted transcription.