October 12, 2018

Crowdsourcing Training Data For Real-Time Transcription Models

Sara Basson

Dimitri Kanevsky

CROWDSOURCING TRAINING DATA FOR REAL-TIME TRANSCRIPTION MODELS

ABSTRACT

A system and method are disclosed to train speech transcription models via crowdsourcing. Users of a media sharing platform may view real-time transcriptions associated with media on the user devices and identify the transcriptions as correct or incorrect. Users may determine with high accuracy correct and incorrect parts of transcribed text, using a general context of a conversation that is being transcribed. The users may select or mark blocks of transcription text and assign the selected text as correct transcription or incorrect transcription on the input user device. The system may aggregate a large amount of marked transcriptions from multiple user devices and store the marked transcriptions. The stored marked transcriptions may be used as training data for transcription and captioning models. The disclosed concept may also be extended to machine translation. The accurate training data enables development of better transcription models.

BACKGROUND

Over the years, speech transcription and captioning have significantly improved and are finding various real-time applications. For instance, speech transcription is a viable substitute for real-time stenography. Captioning tools configured to run on mobile computing devices for captioning spontaneous or live conversations have also been developed. High quality speech recognition may be achieved using acoustic and language models, which are trained on large data samples of already-transcribed speech. Existing speech recognition models use data training sets that include already-captioned audio to provide high quality captioning. However, engaging human transcribers to provide the correctly transcribed audio may be both time consuming and extremely expensive.

A crowd-sourcing version of the transcriptions may be developed, whereby media-sharing platform users add captions to a video or correct the existing captions in poorly transcribed videos. Such crowd-sourced transcriptions may not be of sufficient high quality as the training data may not match the conversational speech styles of actual users. Conversion of large samples of actual conversational speech into an accurate, vetted transcribed form remains a challenge. Additionally, editing captions on smaller devices, such as mobile phones, can be complex.

DESCRIPTION

A system and a method are disclosed to train speech transcription models by crowdsourcing transcription. The system and method provide accurate transcription of conversational speech as it occurs on the mobile device for training the speech transcription model. The system, as shown in FIG. 1, may include a server **102** to which inputs could be provided via multiple input user devices **104** to mark transcribed text as accurate or inaccurate for crowdsourcing training data for transcription models. The device **104** may be a mobile device or any smart device with a user interface **106**, such as a touch screen. The device **102** may play media **108**, such as audio or video clips, and display text **110** representing transcriptions or captions associated with the media in real-time. The device **102** is configured to access media through a media sharing platform accessed via a web browser or an application. Further, the media sharing platform provides a graphical user interface with marking options **116**, which may include “correct” or “incorrect”, for marking the captions. The server **102** may include data stores **112** for aggregating the training data and a transcription engine **114**.

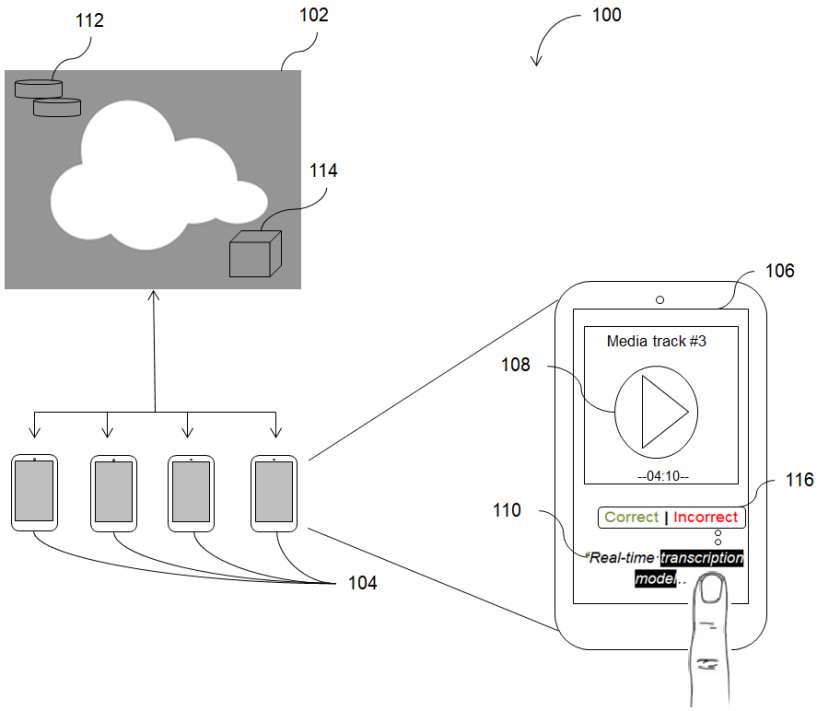

FIG. 1: A system for marking transcription text as “accurate” or “inaccurate” on a device

The method of crowdsourcing transcription and training transcription models, as illustrated in FIG. 2, includes presenting audio or video with real-time transcription text **110** on user devices **102**, in block **201**. The users may select or mark blocks of text and assign the selected text **111** as correct transcription or incorrect transcription, in block **203**. The user may select blocks of text by touching the display screen using fingers or by using a mouse pointer. In some aspects, the touching operation may be similar to copy and paste blocks of text. In some cases, the transcribed text may scroll by quickly when the audio or video is playing. The viewers may then use a single touch to mark the onset and offset of the block as “correct” when the user detects that the particular block of text is “correct”. The selected or marked text **111** may then be aggregated in one or more data stores, in block **205**. The aggregated training data is used to train transcription and captioning models accurately for conversational speech, in block **207**.

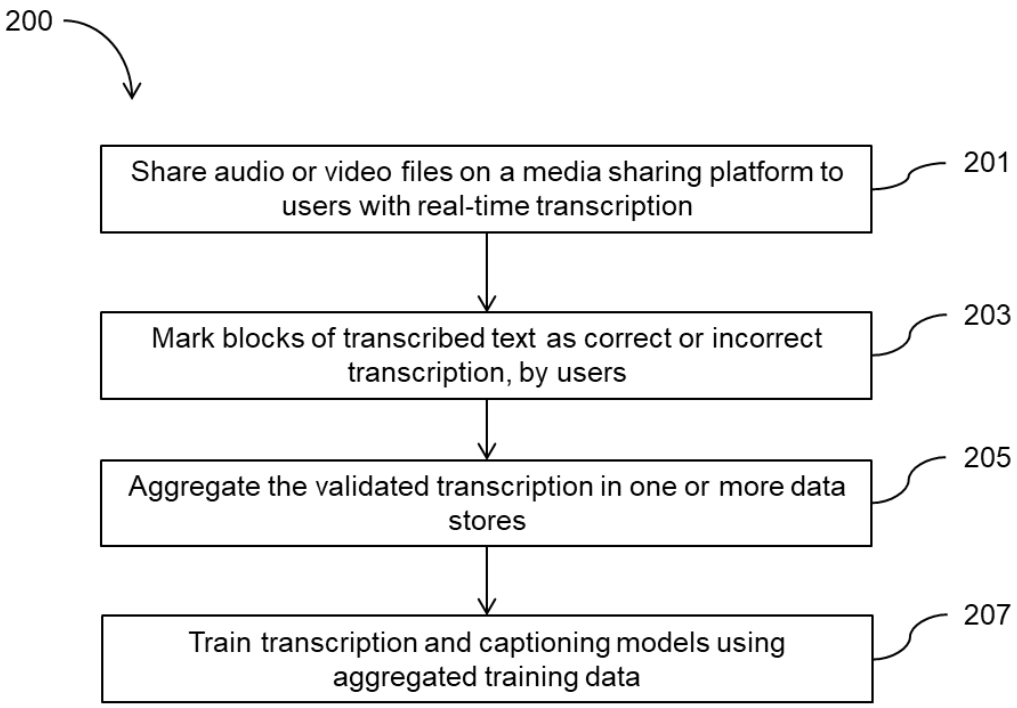

FIG. 2: A method of crowdsourcing transcription of audio or video

Alternatively, the user may opt to select blocks of text that may be inaccurate. For blocks of text marked inaccurate, the viewer may be presented an option to provide their hypothesis about the source of the problem. The options may include “multiple speakers talking”, “noisy environment”, or “speaker with strong accent”, etc. The audio associated with that block of text will be saved for training or improving the transcription and captioning models. Based on this the correct transcription is associated with the corresponding audio or video. Further, the device may also provide a menu to the user or viewer to provide scores ranging from 1 to 5. The scores may indicate how confident the user is about the texts selected as “correct” or “incorrect”.

Users may determine with high accuracy correct and incorrect parts of transcribed text, using a general context of a conversation that is being transcribed. This allows obtaining labeled data for additional training of speech recognition even when users do not know a "ground truth" - what was actually spoken.

The method creates a large corpus of audio or video with aligned decoded text based on the user's feedback. In some aspects, the audio-text segments with high confidence scores may be automatically filtered from the large corpus and used for training the transcription and captioning models. Therefore, the models may be trained on high quality transcripts.

In some aspects, the method may also be extended to machine translation, for example, users may mark blocks of translated data that they think are translated correctly. In some aspects, several high quality offline speech recognition systems to decode live streamed audio may be used. The decoded data may be used for training conversational live stream models.

The system and method may be implemented by transcription providers, interpretation and translation services and speech recognition software developers. The system and method of crowdsourcing transcriptions of conversational speech in videos and audios via mobile devices allows collection of large amount of training data. The training data is more accurate as any user can easily mark texts as "correct" or "incorrect" via mobile devices. The accurate training data enables development of better transcription models.