September 2020

# Using Visual Context to Improve Accuracy of Automated Speech Transcription

N/A



# Using Visual Context to Improve Accuracy of Automated Speech Transcription

ABSTRACT

Machine generated speech transcriptions are a feature of several products such as videoconferencing software, mobile operating systems, etc. However, automatic transcribers are poor at accurately understanding some types of real world user speech. Spoken terms that are phonetically similar but have different meanings can cause errors in machine generated transcription. Although automatic transcribers evaluate various probable phrases as the spoken phrase, the analysis of sound alone is not enough to accurately recognize speech.

Per the techniques of this disclosure, a machine transcription model evaluates probable options for spoken language and evaluates the options based in part on using user-permitted available visual context. Such visual content is analyzed to determine presence of text within the image. If text is detected, OCR techniques are applied to recognize the text and the recognized text is used to improve the accuracy of transcription.

KEYWORDS

- Speech recognition
- Speech transcription
- Visual context
- Optical character recognition (OCR)
- Text input
- Transcription accuracy
- Machine generated transcription
- Videoconference

## BACKGROUND

Machine generated speech transcriptions are a feature of several products such as videoconferencing software, mobile operating systems, etc. However, current transcription techniques sometimes perform poorly at recognizing real world user speech in certain contexts. For example, spoken terms that are phonetically similar (sound similar) but have different meanings can lead to errors in machine generated transcription. For example, when the audio includes the spoken phrase "*recognize speech*" it may be erroneously machine transcribed to "*wreck a nice beach*."

Although current automatic transcribers evaluate various probable interpretations of the spoken phrases as the detected phrase, the analysis of input sound alone is not enough to accurately recognize speech.

## DESCRIPTION

Per the techniques of this disclosure, an automatic transcriber evaluates probable transcription options for spoken language based on using user-permitted context, including visual context, as a weighting factor. For example, visual context can be text obtained from images. With user permission, captured or displayed images that occur in the context of audio that is to be transcribed are analyzed to detect presence of text within the image. If text is detected, OCR techniques are applied to recognize the text. This text is then used to evaluate and weight probable transcription options to improve transcription accuracy, e.g., by better distinguishing between phonetically similar spoken terms based on the recognized text.

For instance, consider a user that is using a feature of smart glasses to generate a machine generated transcription of a conversation with a colleague in a meeting room. The meeting room has a display with a shared notes document that includes the phrase "*using common sense*" in the

header. Using the described techniques the smart glasses use a visual sensor to capture the image of the shared notes document and obtain the text "*using common sense*."

During the discussion, the user's colleague utters the phrase "using common sense." To transcribe this audio, the automatic transcriber evaluates probable transcription options for the spoken phrase which may include, e.g., option 1: "*you sing calm incense*" and option 2: "*using common sense*." The automatic transcriber weights the options using the context of the obtained text from the analyzed image of the notes. Based on this analysis, it is determined in this example that option 2 is the more likely interpretation of the spoken phrase (based on the text of option 2 matching the detected text from the notes, which in turn assigns a higher weight to option 2). The resultant machine transcription of the uttered phrase "*using common sense*" is then provided to the user via her smart glasses.

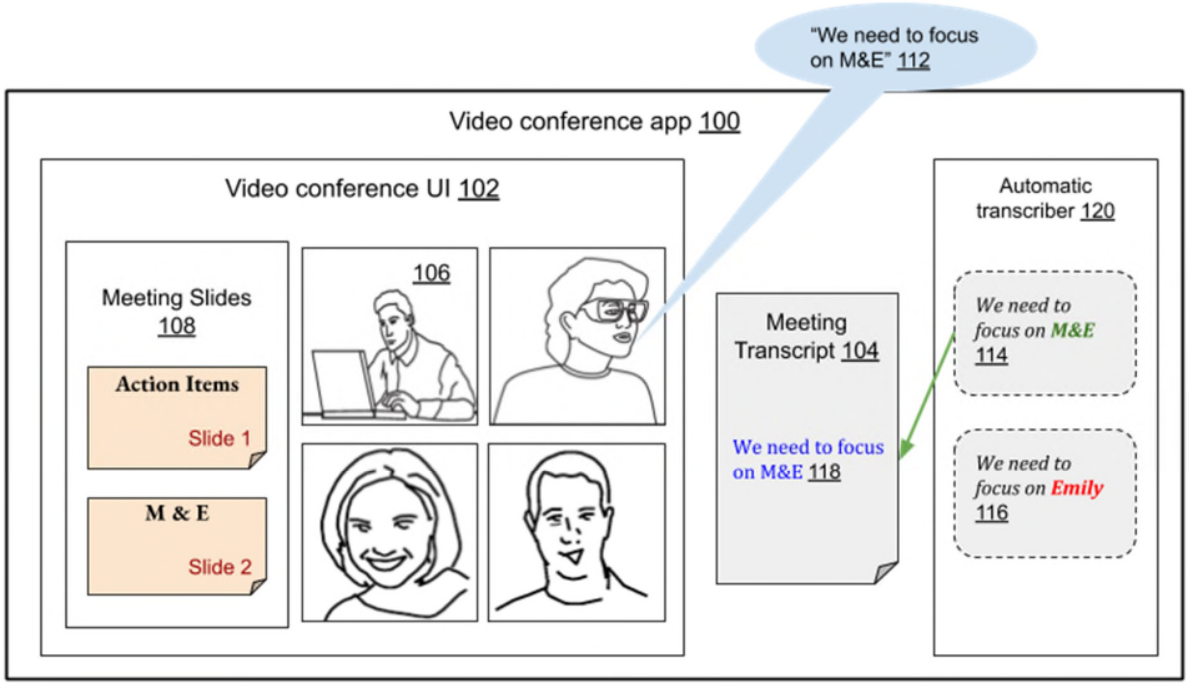


**Fig. 1: Using visual context to improve accuracy of automated speech transcription**

Fig. 1 shows an operational implementation of the techniques described in this disclosure. Four users (shown as thumbnail images) are conducting a meeting via a video conferencing application (100). The users have enabled automatic transcription and provided appropriate permission to access audio and other contextual information of the meeting, e.g., video feed, screen sharing content, etc. The automatic transcription may be provided by a module of the videoconferencing app, e.g., automatic transcriber (120) in the example of Fig. 1, or may be a separate app that can plug in to the video conference. A machine generated meeting transcript (104) of the meeting proceedings is automatically generated and displayed.

In the example of Fig. 1, a user (106) is using the video conference UI (102) to present meeting slides (108) to the video conference participants. The content of the meeting slides is With user permission, the content of the slides (e.g., slide 1 and slide 2) is accessed and OCR techniques are utilized to extract the text from the slides, e.g., “Action items” for slide 1 and “M&E” for slide 2.

While viewing the slides, another user as illustrated in Fig. 1, utters the phrase “We need to focus on M&E” (112). Per the described techniques, the multiple transcription options for the spoken phrase “We need to focus on M&E” are evaluated with the context of the text of the slides shared during the meeting. As seen in Fig. 1, option A includes “We need to focus on M&E” (114) and option B includes “we need to focus on Emily” (116). It will be understood that the options are shown in Fig. 1 for illustration and that only the correct option (118) is displayed in the user interface.

The described techniques can be implemented to improve the accuracy of machine transcription within any application via a device or system that accepts spoken input with additional contextual information, e.g., images or video. For example, the techniques can be used

in smart speakers, smart displays, mobile devices, smart glasses, head-mounted displays (HMDs), video conferencing systems, etc.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's spoken input, displayed images, video conference content), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

## CONCLUSION

Per the techniques of this disclosure, a machine transcription model evaluates probable options for spoken language and evaluates the options based in part on using user-permitted available visual context Such visual content is analyzed to determine presence of text within the image. If text is detected, OCR techniques are applied to recognize the text and the recognized text is used to improve the accuracy of transcription

## REFERENCES

1. Lieberman, Henry, Alexander Faaborg, Waseem Daher, and José Espinosa. "How to wreck a nice beach you sing calm incense." In Proceedings of the 10th international conference on Intelligent user interfaces, pp. 278-280. 2005.