December 2020

Identifying Speakers and Limiting Displayed Transcription to Select Speakers

N/A

Identifying Speakers and Limiting Displayed Transcription to Select Speakers

ABSTRACT

Machine-generated speech transcriptions are helpful for people that are hard of hearing or individuals who don't understand the spoken language to understand conversations that take place near them. Augmented reality glasses can display speech transcriptions to help users understand such conversations. However, the display of transcriptions can be distracting when the user wearing the AR glasses is herself speaking. This disclosure describes techniques to automatically identify speakers in a conversation and only display live transcriptions of speech from conversation participants other than the person to whom the transcription is being provided. Speakers may be identified using user-permitted factors, e.g., by use of a head-related transfer function (HRTF), biometric voice recognition, or facial feature (e.g., lip movement) recognition.

KEYWORDS

- Augmented reality
- Mixed reality
- Speech transcription
- Machine translation
- Live caption
- Speaking turn

BACKGROUND

Machine-generated speech transcriptions are a feature of several products such as videoconferencing software, mobile operating systems, electronic devices, etc. The transcriptions provided by such speech transcription systems are helpful for people that are hard of hearing to understand conversations that take place near them. The transcriptions can also help individuals comprehend speech that is in a language that they do not understand, e.g., while traveling to a place where a different language is spoken and interacting with locals. When using such

transcription systems on wearable or mobile devices, the user interface (UI) can only display a limited amount of transcribed text since the UI is relatively smaller in size.

DESCRIPTION

Techniques are described herein that, with express permission, automatically identify speakers in a conversation and limit the display of a live transcription of speech on a user interface, e.g., displayed on smart glasses or another wearable device, to speech from other speakers, and not the user's own speech. To differentiate between the user's own speech and that of other speakers, automatic speaker identification can be performed with user permission. Such identification can be based on applying a head-related transfer function (HRTF), voice recognition, or facial feature recognition. By limiting the display of a transcription to the speech from other conversation participants, the user interface is made more user-friendly and the limited available space is utilized to display transcriptions of others' speech.

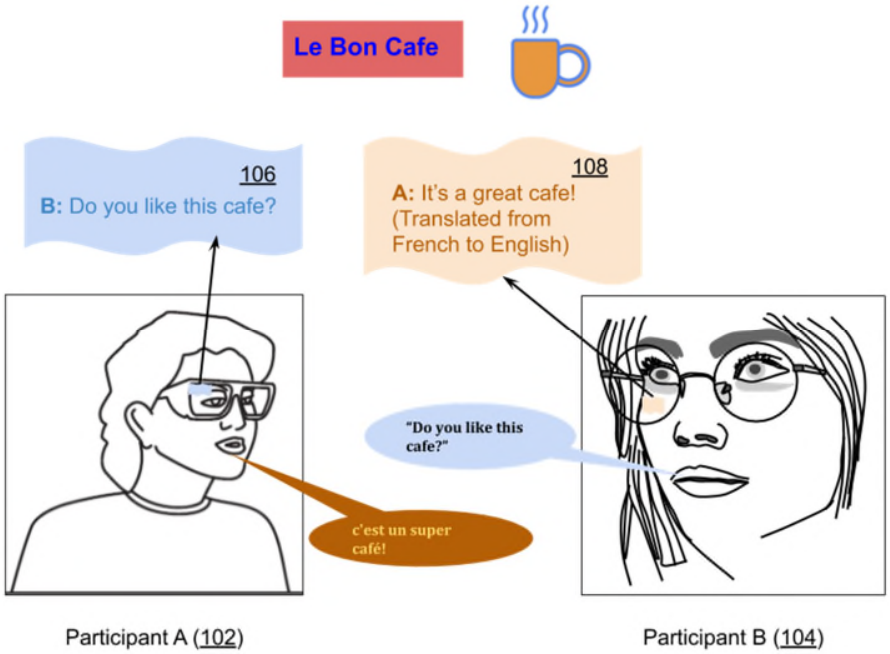

Fig. 1: Identifying speakers and limiting displayed transcription

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. Two participants (102, 104) are having a conversation outside of a cafe. A first participant A (102) and second participant B (104) are both wearing smart glasses that include the feature to display live captions via a machine-generated transcription of a conversation with other participants. In the example of Fig. 1, with express permission of all participants in the conversation, A and B, each wearing smart glasses, have enabled automatic transcription and provided appropriate permissions to access audio.

In the example of Fig. 1, during their discussion, participant B (104) utters the question, "Do you like this cafe?" The participant A (102) responds in their local language, which participant B does not understand, "c'est un super cafe!" Using the described techniques, the smart glasses provide UIs (106, 108) that display a transcription of only the respective speech (or translation) of the participant who is speaking in the listener's user interface. The user interface for each user excludes the transcription of the wearer's own speech. In the example of Fig. 1, participant B has also enabled a translation feature, which provides and displays in their UI a translation in English for the user, a language they understand.

The described techniques, using head-related transfer function (HRTF) or the camera on the smart glasses can also identify the general location of the person speaking for the user. For example, for participant B, the system can identify that the speech they are listening to is coming from participant A who's on their right. The speaker location information can be displayed along with the displayed transcription.

The described techniques can be implemented to identify speakers and to limit display of transcribed speech within any application via a device or system that accepts spoken input and can display transcription. For example, the techniques can be used in smart displays, mobile

devices, smartphones, smart glasses and head-mounted displays (HMDs), other wearable devices, 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, a user's identity, a user's image or facial features, a user's voice), 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

This disclosure describes techniques to automatically identify speakers in a conversation and only display live transcriptions of speech from conversation participants other than the person to whom the transcription is being provided. Speakers may be identified using user-permitted factors, e.g., by use of a head-related transfer function (HRTF), biometric voice recognition, or facial feature (e.g., lip movement) recognition.