Technical Disclosure Commons

Defensive Publications Series

April 2020

Accuracy Evaluation Of Automated Speech Transcription Via User Feedback

Omar Abdelaziz

Justice Ogbonna

Ágoston Weisz

Ragnar Groot Koerkamp

Xinran Lu

See next page for additional authors

Inventor(s)
Omar Abdelaziz, Justice Ogbonna, Ágoston Weisz, Ragnar Groot Koerkamp, Xinran Lu, Saisuresh Krishnakumaran, Tâmara Baía, Vladimir Vuskovic, and Tony Yu

Accuracy Evaluation Of Automated Speech Transcription Via User Feedback

ABSTRACT

Effective operation of a voice-activated virtual assistant requires accurate speech recognition. Manual determination of the accuracy of machine-generated speech transcriptions requires involvement of third parties to evaluate transcriptions of a user's speech. Automated accuracy evaluation approaches that use machine-generated speech as input and determine quality of transcription have limited effectiveness since the machine-generated speech is a poor proxy for real-world user speech, e.g., volume of input, microphone and room characteristics, pronunciations, etc. This disclosure describes obtaining user confirmation of the transcription of a small subset user queries as performed by a virtual assistant or other applications that accept speech input. With user permission, the obtained data, e.g., the user verifying that the transcription was accurate or indicating that it was wrong, are used to rate and/or update the speech recognition technology, e.g., train speech recognition machine learning models.

KEYWORDS

- Speech recognition
- Speech transcription
- Spoken input
- Spoken query
- Virtual assistant
- Smart speaker
- Smart display
- Transcription accuracy

BACKGROUND

Effective operation of a voice-activated virtual assistant requires accurate speech recognition. One technique for determining the accuracy of machine-generated speech transcriptions is to compare them with human transcriptions of the same speech. Discrepancies between machine-generated and human-generated transcriptions can be utilized to identify and fix the shortcomings of the corresponding speech recognition techniques (e.g., machine learning models) used for machine-generated transcriptions. However, such an approach typically requires allowing third parties to listen to a user's speech interactions with a voice based assistant.

Automated accuracy evaluation approaches that use machine-generated speech as input and determine quality of transcription have limited effectiveness. The machine-generated speech can be based on prior transcriptions of user queries. The percentage of words transcribed incorrectly is indicative of the error rate of the speech transcription. However, such an approach suffers some important shortcomings:

- If the original machine-generated transcription is inaccurate, it makes the benchmark not representative of the actual user speech.
- Since the machine-generated speech is a poor proxy for real-world user speech, e.g., volume of input, microphone and room characteristics, capture quality, pronunciations, etc.

DESCRIPTION

This disclosure describes obtaining user confirmation of the transcription of a small subset user queries as performed by a virtual assistant or other applications that accept speech input. With user permission, the obtained data, e.g., the user verifying that the transcription was

accurate or indicating that it was wrong, are used to rate and/or update the speech recognition technology, e.g., train speech recognition machine learning models.

For instance, if the user issues a speech command to set a wake-up alarm for the next morning, the corresponding machine-transcribed text of the command can be echoed back with user permission: "You said: 'Set an alarm for 7:15am for tomorrow morning.' Is that correct?" The user can then respond to indicate whether the presented machine transcription accurately captured the user's speech. Enabling users to provide such an indication of accuracy can be achieved via any suitable user interface (UI) options such as a dialog box with "Yes" and "No" buttons, input via voice, etc. After the user indicates whether the machine-generated transcription is accurate, the rest of the interaction proceeds based on the user confirmation.

The user confirmation operation described above can be sought for a small subset of the user's speech input, thus keeping user burden to a minimum. The user speech subset, used with permission for seeking user confirmation, can be selected randomly or can be based on other criteria, e.g., low confidence level for the transcribed speech for a particular query.

User responses when prompted for accuracy confirmation of speech transcription are aggregated over a large set of transcription samples. The percentage of the transcription sample for which the user confirms that the transcription was accurate is the user-perceived transcription accuracy.

The quality of such user-perceived transcription accuracy measurement can be determined by comparing it with current machine and/or human based approaches for evaluating the accuracy of speech recognition models. When the quality of user-perceived transcription accuracy meets an acceptable threshold, the measure can replace other manual and/or automated approaches to evaluate the accuracy of speech transcription models.

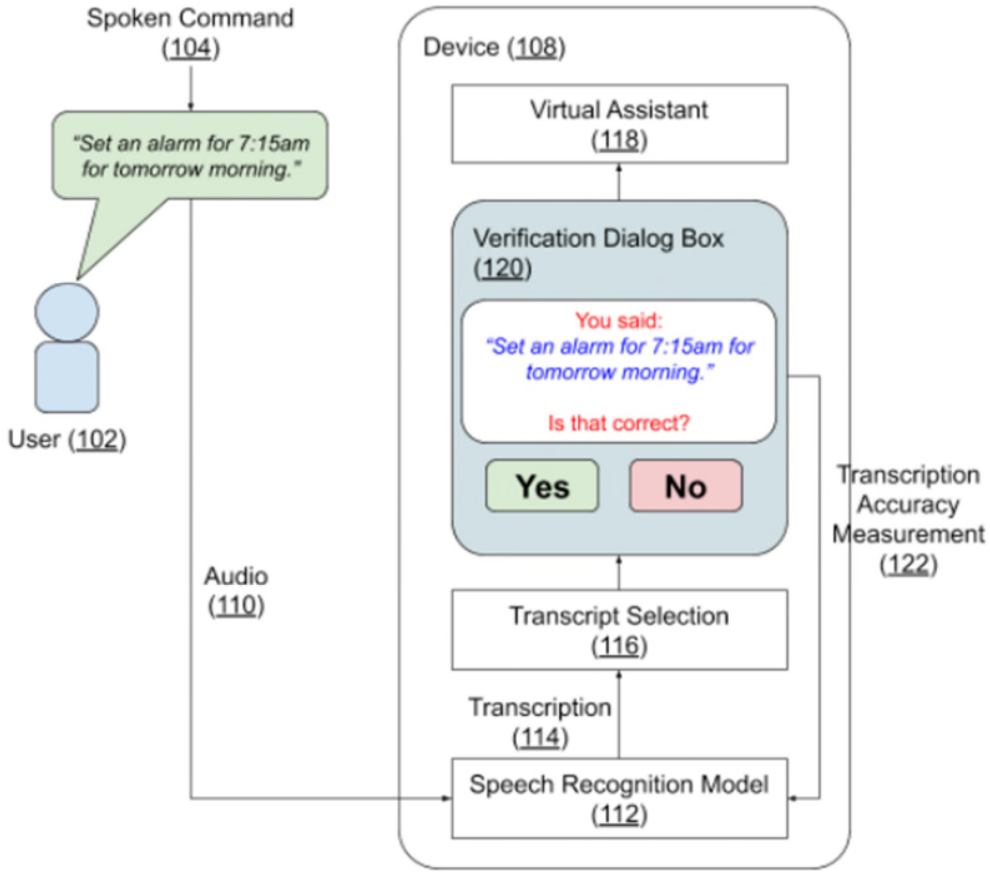

Fig. 1: Seeking user verification of machine transcription of a voice command

Fig. 1 shows an operational implementation of the techniques described in this disclosure. A user (102) issues a spoken command (104) for a voice assistant (118) application on the user's device (108). With user permission, the audio (110) of the voice command is analyzed by a speech recognition model (112) to generate a text transcript (114) of the user's speech.

A transcript selection module (116) is used to determine whether to seek user verification for the generated transcription. If selected, the user is presented with a dialog box (120) that requests the user to verify that the transcribed text matches the original voice command. The user's response serves as a measurement of transcription accuracy (122) that is provided to the speech recognition model for feedback and improvement. With user permission, model training

can be performed on-device or on a server, based on the transcription accuracy. The transcribed voice command is then provided to the virtual assistant for appropriate action based on the content of the command.

As described above, operational implementation of the techniques of this disclosure involves the potential use of one or more threshold values. For instance, threshold values can be used to determine the percentage of spoken commands for which user confirmation is sought or selecting a specific audio segment for transcription verification. The threshold values can be specified by the developers and/or determined dynamically at run time.

The described techniques can be implemented to evaluate the accuracy of machine transcription within any application (e.g., a virtual assistant) via a device that accepts spoken input, such as smart speakers, smart displays, mobile devices, virtual assistants, etc. The user confirmations obtained using the described techniques can help improve the accuracy of speech recognition models without requiring human transcribers to listen to the spoken command. These improvements can enhance the user experience of speech-based interactions.

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 commands, verification inputs), 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 obtaining user confirmation of the transcription of a small subset user queries as performed by a virtual assistant or other applications that accept speech input. With user permission, the obtained data, e.g., the user verifying that the transcription was accurate or indicating that it was wrong, are used to rate and/or update the speech recognition technology, e.g., train speech recognition machine learning models. The techniques enable determination of the accuracy of speech transcription models The user confirmation is sought for a small, random sample of user queries, keeping user burden to a minimum. The described techniques can be implemented for evaluating the accuracy of machine transcription within any application (e.g., virtual assistant) provided via any device such as smart speakers, smart displays, mobile devices, etc.