April 2023

Enabling Inline Correction of Speech Transcript via Audio Cues

Gaetano Ling

Amelia Schladow

Michael Colville

Matthew Sibigtroth

George Joseph Rickerby

See next page for additional authors

Inventor(s)
Gaetano Ling, Amelia Schladow, Michael Colville, Matthew Sibigtroth, George Joseph Rickerby, and Benjamin Guy Alexander Pawle

Enabling Inline Correction of Speech Transcript via Audio Cues

ABSTRACT

While voice input has become a popular way of interacting with devices, user frustration due to incorrect transcription is common. Speech-to-text (STT) conversion errors can require users to provide the spoken input again, manually issue a correction command, or use a non-voice modality to make corrections. This disclosure describes techniques to automatically play audio cues to indicate when the confidence in the accuracy of speech transcription is low. The cues enable timely, inline correction of the transcript as the user speaks, in a manner akin to human conversation. The cues can include a discernible tone/ beep or spoken phrases that indicate that particular spoken phrases were not transcribed with sufficient confidence.

KEYWORDS

- Voice-based interaction
- Voice UI
- Voice assistant
- Speech transcription
- Speech-To-Text (STT)
- Speech conversion
- Transcription error
- Audio cue
- Hands-free interaction
- Inline correction
- Contextual text correction

BACKGROUND

Many users take advantage of the voice-based interactive capabilities of their devices. Such capabilities include the ability to issue commands and queries via voice, composing text content with voice dictation, etc. The voice-based interactions rely on converting the user's spoken words to corresponding text via speech-to-text (STT) techniques, e.g., STT models. Although converting speech to text via such techniques is reasonably accurate, there may sometimes be portions of the text where the conversion is inaccurate. Such portions correspond to portions of the converted speech where the confidence score of the underlying STT model is low. Low-confidence in conversion can result from one or more of a number of reasons such as similar sounding words, accent differences, slang terms, interference from background noise, etc. Such issues in speech conversion are usually limited to small portions, such as individual words or phrases. Portions of the text for which model confidence is low are more likely to be erroneous than other portions.

Detecting errors in speech conversion, such as those that result from low confidence of the model, currently requires users to examine the converted text on the device screen. However, in many situations in which a user engages in voice-based interaction, it is difficult or infeasible for the user to look at the device screen. In such cases, detecting errors requires the user to listen to the entire converted text of their speech by having it read back to them in audio format by running it through text-to-speech (TTS) conversion. Either way, if there are speech conversion errors, correcting them requires that the user engage in non-voice interaction to edit the erroneous text on the device screen. If the user does not wish to switch away from voice-based interaction mode, the user needs to dictate the entire text in which there are erroneous portions,

even when the error is limited to a single word. As a result, detecting and correcting speech conversion errors when using voice-based interaction is cumbersome and inefficient.

DESCRIPTION

This disclosure describes techniques to play appropriate audio cues to indicate when the confidence in the accuracy of speech conversion is low. Such cues can include a discernible tone or beep or spoken phrases, such as “huh?” that mimic typical human reactions upon encountering spoken words that are confusing.

The audio cues can be played while the user is engaging in voice based interaction with the device, e.g., immediately upon encountering a portion of the user’s speech that is associated with low confidence during conversion to text. Placing the cues immediately after low confidence conversion occurs alerts the user to potential inaccuracies in speech conversion at the earliest opportunity. The user can then choose to correct the speech conversion errors before proceeding with the rest of the voice interaction. Such an approach mimics how listeners convey confusion in human conversations, enabling speakers to react and correct any erroneous understanding.

Alternatively, or in addition, the cues can be played in the audio stream generated by applying TTS to play back the converted text. Additionally, visual indicators can be displayed within the converted text of the speech displayed on the device screen. The latter approach is analogous to the visual indicators, such as colored squiggly underlines, used to flag likely spelling or grammar errors in text content.

When the user encounters an audio cue indicating that the portion of the user’s speech that immediately preceded the cue may not have been converted accurately, the user can pause and issue a correction or speak the content again. For instance, if the user hears the audio cue

after saying “message Jenny,” the user can infer that the name Jenny might not have been heard accurately. Instead of continuing with the message to Jenny, the user can then say: “I meant Jenny” and/or spell out the name Jenny and/or speak the name again slower and/or louder. When the speech of the user’s correction attempt is converted accurately with high confidence, the initial erroneous text conversion is replaced with the correct text. Users can employ a similar approach to correct errors when listening to a playback of the converted text of large pieces of voice dictation, such as a paragraph, an email message, etc.

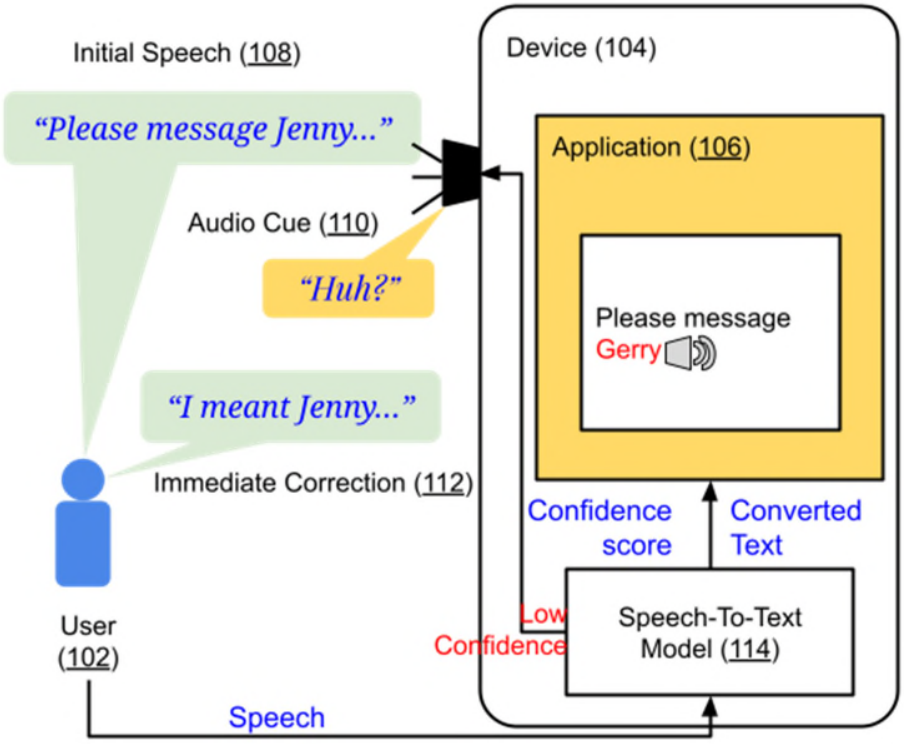

Fig. 1: Making corrections to speech transcription based on audio cues for low confidence

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. A user (102) is interacting with a device (104) via voice to compose a message via an application (106). The user’s speech is converted to text by a speech-to-text (STT) model (114).

The model also generates a score corresponding to the confidence in the accuracy of the converted text. When the confidence score is lower than a threshold, an audio cue (110) ("Huh?") is played to indicate that the conversion ("Gerry") immediately preceding the cue might be inaccurate. Upon hearing the cue, the user can issue an immediate correction (112) to change "Gerry" to "Jenny."

With appropriate user permissions, any voice-based corrections issued by the user can be applied in a contextually appropriate manner to relevant portions of the prior speech flagged with audio cues as having been converted to text with low confidence. For example, consider that a user uttered the following voice message:

> Please message Jenny to schedule an appointment with Bob from three to four next Monday to discuss accounting matters.

However, the conversion has errors, as indicated in red below:

> Please message Gerry to schedule an appointment with Bob from two to four next Monday to discuss accounting matters.

The audio cue is played after "message Gerry" and "two to four" to indicate that the confidence in the accuracy of the text conversion just prior to the cue is low. In this case, if the user says: "it should be three to four," the user's intention can be appropriately inferred as the instruction to correct the low-confidence snippet "two to four" to "three to four."

The voice-based interaction can be augmented to capture relevant aspects of the user's speech intonation, such as speed, tone, etc. The converted text and the captured intonation can be shared as a package of multimodal content. The parties receiving the multimodal content can

choose to engage with it in any manner of their choice, such as reading the text, listening to the audio of the text, etc.

The techniques described herein can be incorporated within any device, application, or platform that provides voice-based interaction based on speech-to-text conversion. Users can employ the techniques for any voice-based task, such as composing a document, dictating a message, creating notes, issuing commands or queries, etc. The threshold value of confidence for flagging converted speech with audio cues as well as the type of audio cues can be set by the developers and/or determined dynamically at runtime. Speech conversion can occur locally on the device and/or (if permitted by the user), external to the device, e.g., on a local or remote server.

By mimicking the detection and correction of communication errors in human conversations, the approach described herein can make it quicker, easier, and more intuitive for users to be notified of transcription errors when interacting with their devices via voice and to correct such errors. Moreover, the techniques enable users to use voice to issue local corrections to specific inaccurate portions within the converted speech, without requiring use of the device screen or having to repeat the entire spoken content. Implementation of the techniques can improve the user experience (UX) of voice-based interaction.

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 and attributes of spoken input, social network, social actions or activities, profession, a user's preferences, or a user's current location), 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, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. 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 play audio cues to indicate when the confidence in the accuracy of speech transcription is low. The cues enable timely, inline correction of the transcript as the user speaks, in a manner akin to human conversation. The cues can include a discernible tone/ beep or spoken phrases that indicate that particular spoken phrases were not transcribed with sufficient confidence.

REFERENCES

1. Qiu, David, Qiujia Li, Yanzhang He, Yu Zhang, Bo Li, Liangliang Cao, Rohit Prabhavalkar et al. "Learning Word-Level Confidence for Subword End-To-End Automatic Speech Recognition." U.S. Patent Application 17/182,592, filed February 23, 2021.