November 2020

Speech Recognition Correction Based On Detected Topic In Speech

Victor Carbune

Matthew Sharifi

Speech Recognition Correction Based On Detected Topic In Speech

ABSTRACT

This disclosure describes techniques for adjusting speech recognition of a voice dictation based on topics that are detected from the speech during and after the dictation. A topic of the speech content is classified based on recognized speech. Speech recognition of further speech is adjusted to the detected topic, and text that was previously-recognized during the dictation is re-recognized and corrected using the adjustment. Described features can improve accuracy of text transcripts obtained automatically from dictation, reducing the amount of manual corrections required for recognized text.

KEYWORDS

- Voice dictation
- Automatic Speech Recognition (ASR)
- Speech biasing
- Topic classification
- Topic detection
- Speech-to-text
- Speech transcript

BACKGROUND

Voice dictation is a popular input method for mobile devices, for example, when users are on-the-go or in other hands-free situations. Automatic speech recognition features provided via such a device receive the user's speech, recognize the spoken words, and generate a text transcript of the spoken words. Dictation features, e.g., made available via a device operating system, a software keyboard application, a virtual assistant, word processor, or other application

allow users to conveniently compose text in various contexts, e.g., a message to a friend, reply to an email, composing an article, etc. Compared to typing, voice dictation can be faster for the user, and it can be particularly useful in situations where typing or swiping on a keyboard is inconvenient.

An important use case for voice dictation is long-form dictation, in which users compose long-form content including longer messages, emails, or documents. However, this form of dictation is particularly challenging for many speech recognition systems because such dictation often involves specialized terms related to the context of the long-form content, e.g., medical topics, financial topics, etc.) that a general-purpose speech recognition system is often unable to recognize accurately.

A high error rate of speech recognition systems is a hindrance to greater use of voice dictation using these systems, particularly for long-form content that relates to specialized topics. Furthermore, correcting mistakes and editing the dictated text is difficult since it requires the user to switch to keyboard input or tap through the words via a touchscreen.

DESCRIPTION

This disclosure describes features that enable correction of errors in text recognized from a user's speech. As a user continues to dictate speech content to a system, a topic is identified based on the received content and the ongoing speech recognition is adjusted to recognize further speech based on the topic. Furthermore, text previously-recognized during the same dictation is reanalyzed based on the topic and corrected in a displayed transcript.

The described techniques are implemented upon specific user permission to access a user's data, e.g., speech data, context information, etc. Users are provided with options to grant

permissions to and/or to disable features entirely. The user can enable or disable techniques discussed herein for particular locations, time periods, or for other conditions.

For long-form dictation, adjusting speech recognition techniques to take into account a current topic can help with speech recognition accuracy, particularly when the input content is more specialized to a particular topic such as medical topics in a medical report, financial topics in a financial report, etc. At the beginning of a dictation experience, there may be no indication that the user is about to provide input speech relating to a particular topic. However, partway through the dictation, it is possible to determine the topic (domain) from the received content. Therefore, corrections to recognized text (including retroactive corrections) can be triggered when the topic of the spoken message is detected. This disclosure describes automatic adjustment of speech recognition (referred to occasionally as speech biasing) that is provided simultaneously with processing of speech input which enables the accuracy of recognition of dictated content to be improved as more speech content is input by the user.

Reducing errors in speech recognition enables a user experience where the user can dictate long-form content, allowing initial errors from speech recognition and then retroactive correction of the errors after additional dictation content is received and processed. As a greater amount of content is received, the topic and intended use for the input can be better identified based on the structure of the content and based on correctly recognized terms (some terms may be still misrecognized or incorrect).

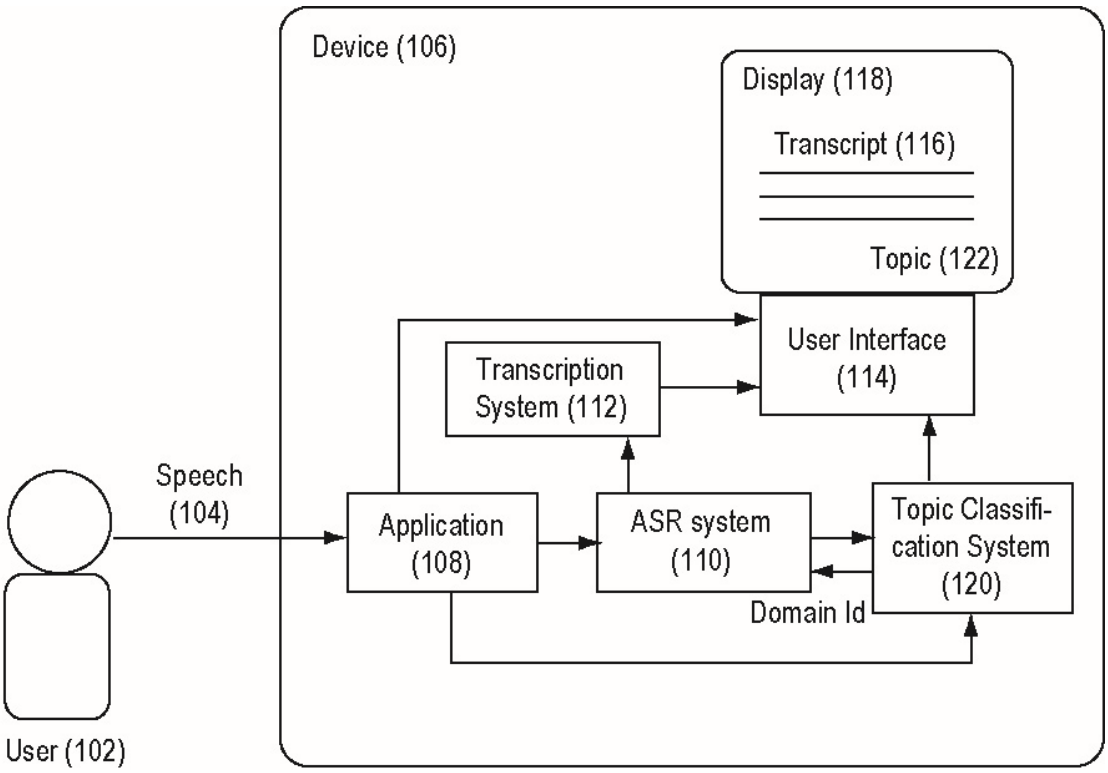

Fig. 1: Speech recognition system using biasing based on detected speech topic

Fig. 1 shows an operational implementation of techniques described in this disclosure. A user (102) starts a voice dictation session and issues speech (104) that is detected by a microphone of a device (106). The speech is provided as input to an application (108), e.g., a messaging application, document application, a software assistant, etc. For example, the speech can be a reply to a received message, a dictation of a new email message, a dictation of a document, etc.

With user permission, the speech is processed by trained machine learning models included in an Automatic Speech Recognition (ASR) system (110) which may be part of the application that receives the input speech or can be implemented separately, e.g., as part of a device operating system. The ASR system performs speech recognition, including determining hypotheses for words, detecting when there is voice activity from the user, determining

endpoints indicating when the user has finished speaking, etc. The ASR system provides recognized text to a transcription system (112) that outputs the recognized text via a user interface (114) as a transcript (116) provided on a display (118) while the user is speaking.

A topic classification system (120) receives information from the ASR system, including top-transcribed hypotheses for detected words and phrases, one or more alternate hypotheses, detected voice activity indicators, endpointing indicators, etc. If the user permits, the topic classification system also obtains current context information from the application related to the dictation, including the identity of the application in which the speech is being received, time of day, calendar events of the user, currently-running applications on the device, or other user-permitted information.

The topic classification system runs in parallel with the ASR system and determines the likely topic of the user's speech, where the topic indicates the type of information being recognized and transcribed. Some examples of topics include types of reports, documents, emails, messages, or other types of information. A list of candidate topics, categories, or types can be produced, such as "personal message," "work email," "medical report," "finance report," "wedding invitation," "location description," "event description," etc.

The topic classification system can be implemented using a text-based neural network that processes the information from the ASR system and the context information. For example, the topic classification system can receive top alternate recognitions from the ASR system in parallel. The topic classification system can include, for example, a standard classification layer that returns to the ASR system, as its output, the most likely topic as a domain identifier and an associated confidence score indicating the confidence for that topic.

The ASR system uses the identified topic to adjust the recognition of the user's speech. The adjustment can include directing a particular list of items (e.g., words or phrases) to the recognition process, the items being much more likely to be present in speech related to the detected topic. The list of items can be provided by the ASR system or by the application. For example, an email application may adjust items for an email type of topic based on words that the user has used in previous emails (if such access is permitted by the user). Or, biasing can be implicitly provided in the neural network of the models of the ASR system that have an input for the topic (domain ID), which can be an embedding in the neural network corresponding to that topic. For example, if the topic is detected as relating to a particular sport, adjustment to ASR can include interpreting user input as terms associated with the sport, e.g., team or player names, terminology associated with the particular sport, e.g., for basketball, "assist," "three-pointer," "steal," "block," "double-double," etc. such that these terms are preferred by the ASR system over alternate interpretations of the spoken term. For example, an ASR system adjusted for the topic basketball may interpret "double-double" rather than "double trouble" which may be the default interpretation.

When a topic is identified, e.g., in response to the ASR system receiving a domain ID from the topic classification system, the ASR system may, using adjustments for the topic, re-recognize parts or the entirety of the speech that was previously recognized and transcribed as part of the same user dictation without use of that topic adjustment. The re-recognized content is provided to the transcription system to be displayed to the user as corrections to previously-transcribed words, sentences, or paragraphs. For example, if a sports topic is detected, certain words are more likely transcribed as terms associated with the sport and the displayed transcript shows the corrections as suggestions in the user interface. In various cases, based on user

settings, and confidence scores associated with the suggestions, the suggestions can be accepted into the transcript based on user input, or can be made automatically without user input.

As the topic classification system receives further information from the ASR system, it may change the confidence level of an identified topic such that a different topic becomes the most likely topic of the user's speech. For example, the topic may change from "work email" to "finance report." If such a topic change occurs, the topic classification system sends the new topic identification (domain ID) to the ASR system. The ASR system then recognizes the speech that follows based on a new list of items associated with the new topic.

In addition, similarly as described above, based on the changed topic, the ASR can re-recognize parts of or the entirety of the speech that was previously recognized and transcribed as part of the same user dictation. The re-recognized content is provided to the transcription system to be displayed to the user as corrections to previously-transcribed words.

An identified topic (122) can be displayed as the user's dictation continues, or displayed when a topic change has been detected. For example, a currently-identified topic can be overlaid on the displayed transcript.

In some cases, if a topic is determined or if topic change occurs, the user interface can display a button or other interface element that allows the user to change the entire transcribed message based on the latest identified topic. For example, the entire recorded speech audio, which was stored in storage, can be re-accessed and reinterpreted by the ASR system using the detected topic as an adjustment input. Further, one or more of the derived features and/or classifier outputs may be reprocessed.

While Fig. 1 shows a speech recognition system interacting with an application on a device, described techniques can be implemented in other configurations as permitted by the

user. For example, the functionality of the speech recognition system and/or topic classifier may be incorporated within one or more applications or in a device operating system. In another example, if the user permits, one or more of the ASR system (or components thereof), topic classifier, transcription system, and/or other modules can be executed on a server that is remote from the user device.

According to described features, speech content dictated by a user is examined holistically to determine a topic as the speech continues to be input. Corrections are proposed for previously-recognized paragraphs, sentences, or words that better match the recognized topic, in view of a larger amount of received speech. This enables the user to fluently dictate an entire report and have high confidence that correction of recognized text is being performed, thus reducing end-to-end time of the user spent performing manual corrections of misrecognized speech.

The various features of the speech recognition system are implemented with specific user permission to access user data that serves as input to the system. 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 on-screen content, a user's speech input and commands, a user's preferences, a user's contextual items such as calendar items, etc.), and if the user is sent content or communications from a server. Certain techniques are not implemented if users deny permission. 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 for adjusting speech recognition of a voice dictation based on topics that are detected from the speech during and after the dictation. A topic of the speech content is classified based on recognized speech. Speech recognition of further speech is adjusted to the detected topic, and text that was previously-recognized during the dictation is re-recognized and corrected using the adjustment. Described features can improve accuracy of text transcripts obtained automatically from dictation, reducing the amount of manual corrections required for recognized text.