Technical Disclosure Commons

Defensive Publications Series

April 2020

Audiovisual Transcriptions for Input Verification of Voice Queries

Matthew Sharifi
Victor Carbune

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
https://www.tdcommons.org/dpubs_series/3122

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.
Audiovisual Transcriptions for Input Verification of Voice Queries

ABSTRACT

Speech interfaces provide a natural and accessible mechanism allowing users that cannot read to interact with computing devices to perform information-centric tasks. To verify that a spoken query has been understood correctly, devices such as smart displays, smartphones, etc. show users a streaming transcription of the spoken query, obtained using a speech recognition engine. However, text transcription based input verification is not usable by those who cannot read. The techniques described in this disclosure generate and provide an audiovisual transcription of the user’s speech input during and/or after the input, enabling users to verify that the speech input was understood correctly.

KEYWORDS

- Automatic speech recognition (ASR)
- Virtual assistant
- Voice assistant
- Spoken input
- Spoken query
- Entity recognition
- Smart display
- Smart speaker
- Low literacy

BACKGROUND

A significant proportion of the world’s adult population cannot read because of low levels of literacy. Moreover, children learn to speak much before they can read with reasonable
proficiency. Taken together, low-literacy adults and pre-reading-age children comprise more than a billion individuals who are unable to consume written information and take advantage of services that rely on the user’s ability to read.

Speech interfaces provide a natural and accessible mechanism allowing users that cannot read to interact with computing devices to perform information-centric tasks. For instance, such users can issue spoken queries to a virtual assistant, e.g., provided via devices such as a smart display, smart speaker, smartphone, or other appliance, and get an audio response to their request.

In the case of voice input, it can be challenging to determine whether a spoken query was processed accurately. To verify that a spoken query has been understood correctly, devices such as smart displays, smartphones, etc. show users a streaming transcription of the spoken query, obtained using a speech recognition engine. The text transcription can help users determine if the voice is processed inaccurately, partly or completely, thus requiring the user to reissue or modify the request. The text transcription can also indicate whether the query cannot be handled by the voice assistant. However, text transcription based input verification is unusable by those who cannot read.

DESCRIPTION

The techniques described in this disclosure generate and provide an audiovisual transcription of the user’s speech input during and/or after the input, enabling users to verify that the speech input was understood correctly. The audiovisual transcriptions are accessible to those who cannot read. If the user permits, the audiovisual transcription is streamed as real time feedback while the user is speaking and/or presented as the final output of an automatic speech
recognition (ASR) system that interprets the spoken query after the user has completed speaking the entire query.

Specifically, the techniques involve the use of image(s) and audio/video to provide feedback that conveys the results of interpreting the user’s voice input. The streaming text output from the ASR engine used for initial processing of the user’s voice input is further processed via a semantic annotator. Portions of the unstructured streaming text input provided to the semantic annotator are mapped to entities present in a structured knowledge base, each of which has an associated image or set of images. Typically, a single default image is selected for each mapped entity within the input. However, in some cases, other images can be selected to adapt the image selection based on factors such as other entities within the input, screen parameters of the display device, etc. When multiple entities are detected within the query, each with a corresponding image, the multiple images can be shown in sequence or composited into a single composite image. For example, compositing multiple images can be performed using machine-learning based style transfer, with the use of Generative Adversarial Networks (GANs), etc.

Entities that are commonly encountered in queries, e.g., pizza, dog, etc., are often associated with a corresponding emoji. If an emoji is available for an entity detected in the user’s voice input, the emoji can be used as the corresponding visual transcription rather than other types of images. Owing to their familiarity, users are likely to process and understand the meaning of emojis more effectively. Additionally, younger users are likely to relate to emojis more readily.

The above-described visual transcription converts the entities and other phrases contained in the user’s voice query with corresponding visuals such as emojis, photos, icons, etc. For example, the visual transcription for a user query “what is a good pizza place?” can show
emojis for a pizza and a building. Similarly, if a user asks, “how tall is the Eiffel Tower?,” the query is depicted visually with a photo (or sketch) of the Eiffel Tower along with the icon of a measuring tape, composed into a single image.

Voice queries can sometimes include information that is difficult (or impossible) to convey in visual form. With user permission, the visual transcription can be augmented by echoing back the audio of the query fully or in part. The word stability confidence scores associated with the ASR generated text output are examined for this purpose. Parts of the speech with confidence scores lower than a threshold value are deemed as potential sources of misunderstanding and selected for echoing back to the user. As such, the high confidence parts of the query are rephrased or skipped, thus shortening the amount of audio that is played for user verification.

Parts of the query can be echoed back while the user is in the process of speaking. With permission, the echoed playback can mimic the user’s style with a cadence that avoids interfering with the user’s train of thought during voice input. Alternatively, or in addition, the entire query can be played back after the user is done speaking so that the user can inspect if the query as a whole was understood correctly.
Fig. 1: Audiovisual transcription for voice query verification

Fig. 1 shows an operational implementation of the techniques described in this disclosure. A user (102) issues a voice query (104) regarding a pizza establishment to a device (106) such as a smart speaker, smart display, smartphone, or other appliance that includes a voice query interface, e.g., provided by a virtual assistant. With user permission, the text of the query is extracted by automatic speech recognition (108) and passed to a semantic annotator (110).

The entities (112) detected within the text by the semantic annotator are associated with corresponding images (116) obtained from a structured knowledge base (114). In case an entity is associated with multiple images, an image selector and compositor module (118) is used to select a single fitting image or combine multiple images into a single composite image as appropriate for the user’s query. While Fig. 1 shows the module on-device, it can also be
implemented on the system that provides the structured knowledge base, or a combination of various systems.

The images are provided to a virtual assistant (120) along with the text output and corresponding confidence scores (122) of the ASR process. The images are shown to the user as a visual transcription (124) on a transcription user interface (120). Additionally, parts of the query with low transcription confidence can be played back to the user via audio (126). The user can optionally request audio playback of the entire voice query.

The threshold values used to determine transcription confidence can be set by the developers, specified by the users, or determined dynamically at runtime. Users can enable the audiovisual transcription as described in this disclosure via a setting, e.g., controlled by voice input. Implementation of the techniques can improve the user experience and accessibility of online content and information-driven services for users that are unable to read. With user permission, the techniques can be extended to support audiovisual delivery of information from web pages and other information sources relevant to the user’s voice query.

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 voice queries, a user’s ability to read, a user’s preferences), 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

The techniques described in this disclosure generate and provide an audiovisual transcription of the user’s speech input during and/or after the input, enabling users to verify that the speech input was understood correctly. With user permission, streaming text obtained via automatic speech recognition is further processed via a semantic annotator to map entities within the text to corresponding visual representations, such as emojis, photos, icons, etc. The visual transcription is augmented by echoing back the audio of the query fully or in part. Such audiovisual feedback can improve the user experience and accessibility of spoken query interfaces for online content and information-driven services for users that are unable to read.

REFERENCES
