Technical Disclosure Commons

November 15, 2017

Voice activated spell-check

Pedro Gonnet

Victor Carbune

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

Voice activated spell-check

ABSTRACT

Voice based user queries sometimes include words in multiple languages that could stymie speech recognition systems. Alternatively, different words in the same language may have similar pronunciation. A device, e.g., smart home speaker, smartphone, etc. that implements a voice-based user interface receives a user command, where portions of the user speech are parsed with low confidence, e.g., due to similar pronunciation, or use of a different language. The described techniques enable improved voice-based interaction by permitting a user to provide corrections for such portions. For example, the techniques enable users to provide a correction via speech, via a displayed interface, etc.

KEYWORDS

- Speech recognition

- Virtual assistant

- Multilingual

- Query correction

- Smart home

- Smart speaker

- Smart appliance

- Internet of Things

- Audio UI

BACKGROUND

Voice-based user interfaces are popular, e.g., in appliances such as smart home speakers, and in voice-activated assistant software provided on smartphones. The quality of the response

from such appliances and software depends on accurate recognition of the user's speech. However, portions of the user's speech, e.g., spellings of names of people or places, are sometimes incorrectly parsed. This is often the case with multi-language voice queries, e.g., where the names in a query are in a different language than the rest of the query.

For example, the recognition of a voice command "Take me to Bahnhof Enge" can result in a transcript "Take me to Bonifay." Further, different words in the same language may sometimes have similar pronunciation. Incorrect transcription can lead to incorrect responses and a sub-optimal user experience. There is currently no voice-activated way for the user to correct the transcription.

DESCRIPTION

Speech recognition is used to transcribe voice commands into text for the system to process and provide a response. In certain situations, portions of the recognized speech may be associated with confidence scores that are below accuracy thresholds. In such situations, the techniques of this disclosure provide a voice-based user interface for the user to provide corrections or clarification.

A device that implements the voice-based user interface, requests the user to spell out the words or phrases that are associated with low confidence scores during the recognition. For example, the device may ask "can you spell that?" or use a more specific question such as "can you spell the first letter of the last word?" The device provides a response to the user query only after the user's speech has been parsed correctly.

This eliminates presentation of inaccurate results, and improves the user experience. In some instances, other user interfaces are used in addition to, or alternative to a voice UI. For

example, when the user query is regarding a location, the device presents a map on a display

screen and requests the user to point to the location.

Voice-based user interaction

User: "Take me to Bahnhof Enge"

Virtual Assistant: "I didn't quite understand the last part, could you spell that?"

User: "B-A-H-N-H-O-F E-N-G-E"

Virtual Assistant: "Thank you! The fastest route to Bahnhof Enge is ..."

Fig. 1: Voice interaction to correct speech transcription errors

Fig. 1 illustrates an interaction between a voice-based virtual assistant and a user. A

speech recognition system, e.g., that utilizes beam search, determines that the possible results are

"Take to me to Bahnhof Onga," "Take me to Bahnhof Inga," and "Take me to Bahnhof Enge."

When these results are associated with low confidence scores, e.g., for the last two words, the

virtual assistant seeks user confirmation of the spelling of the two words, e.g., "could you spell

that?" By enabling the user to provide corrections using voice, the techniques can interpret user

commands correctly before providing a response. Further, the user does not need to switch

between different interaction modes, e.g., from speaking to typing in this case.

When user provide permission for use of user data, user queries and corresponding

corrections can be provided as feedback to improve speech recognition. Specifically, such data

points can be characterized as weakly-labeled data to improve future iterations of speech

recognition.

Additionally, the present techniques may identify a category of objects that are a subject of the user voice command and suggest app actions for the identified objects. For example, if a virtual assistant identifies two locations that match the parsed user voice command, a maps user interface is presented for the user to select the intended location.

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 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

A device, e.g., smart home speaker, smartphone, etc. that implements a voice-based user interface receives a user command, where portions of the user speech are parsed with low confidence, e.g., due to similar pronunciation, or use of a different language. The described techniques enable improved voice-based interaction by permitting a user to provide corrections for such portions. For example, the techniques enable users to provide a correction via speech, via a displayed interface, etc.