Fusion of Speech and Gesture Inputs for Improved Speech Recognition

Tyler Thornblade

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

Recommended Citation
Thornblade, Tyler, "Fusion of Speech and Gesture Inputs for Improved Speech Recognition", Technical Disclosure Commons, (April 27, 2020)
https://www.tdcommons.org/dpubs_series/3193

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.
Fusion of Speech and Gesture Inputs for Improved Speech Recognition

ABSTRACT

While speech is an important input mechanism used in many products, interpretation of user speech is challenging when the user input is ambiguous, e.g., due to the presence of punctuation or commands, as opposed to only verbatim text. This disclosure describes the utilization of a combination of speech analysis and gesture recognition to automatically disambiguate between verbatim text input (dictation) and commands. User provided speech and gestures are analyzed and used for interpretation of the spoken input, without the user having to switch between text entry and command entry modes.

KEYWORDS

- Speech recognition
- Gesture recognition
- Dictation
- Speech input
- Voice input
- Word processor
- Multimodal
- Smart speaker

BACKGROUND

Speech is an important input mechanism for users to provide dictation input or to give a command to a computing device such as a smartphone, smart speaker, computer, etc. However, interpretation of user speech is challenging for these devices when the user input is ambiguous, e.g., due to the presence of punctuation or commands, as opposed to only verbatim text. For
example, the spoken phrase "The presentation is due Friday, period" can be transcribed as either “The presentation is due Friday, period” or “The presentation is due Friday.” because the user intent for the term "period" is ambiguous. Further, current speech recognition systems struggle for correct interpretation when the user input mixes verbatim text with command instructions. For example, in the phrase "The dog is wet <italicize wet>,” the words “italicize wet” can be incorrectly interpreted as verbatim text, rather than as a command.

Still further, it is difficult for users to make corrections to transcribed text when using current speech recognition techniques. In human conversation, it is typical for speakers to correct themselves as they speak e.g., "We should go to Phoenix, excuse me, Pittsburgh." The transcribed text that corresponds to this spoken input is “We should go to Pittsburgh.” However, current speech recognition solutions require the user to perform a mode switch from a verbatim text entry or dictation mode to a command entry mode to delete the text that was misinterpreted (e.g., “Phoenix, excuse me,”), prior to resuming dictation. Such mode switching is both unnatural and time consuming.

DESCRIPTION

This disclosure describes the utilization of a combination of speech analysis and gesture recognition to automatically disambiguate between verbatim text input (dictation) and commands. Body language, finger movements, hand gestures, etc. can all be used to provide gesture input.

The recognition of speech and gestures is performed with specific user permission. Users are provided with options to turn off gesture recognition, in which case speech recognition is performed without obtaining user data for gesture recognition.
Fig. 1: Use of speech and gesture input jointly for speech disambiguation

Fig. 1 illustrates an example of use of the described techniques for speech disambiguation that make use of speech and gesture input provided by the user. A fusion of gestural data with speech data is performed when performing speech recognition. In the example illustrated in Fig. 1, the user provides speech input (“We should go to Phoenix, excuse me, Pittsburgh.”) At the time of providing the inadvertent input (“Phoenix”) and subsequent correction (“excuse me, Pittsburgh”) the user performs a gesture, e.g., a shake of the head or other gesture, that indicates that there was an error in the speech input. A device (100), e.g., a smartphone, tablet, computer, or other device that performs speech recognition jointly processes the speech and gesture input and produces as output the correct phrase “We should go to Pittsburgh.”

The device (100) includes a speech recognition module (102) that detects user speech, e.g., obtained via a microphone. The device also includes a gesture recognition module (104) that detects user gestures. For example, the user may perform simple gestures and/or provide input via body language.

The gestures performed by the user can be identified from images obtained by a camera and/or a depth sensor. For example, gesture data can be obtained via one or more of monocular...
or binocular vision technologies, depth sensing technologies (e.g., structured light, time-of-flight, LIDAR, RADAR, etc.). Further, hand gestures can be recognized via an accelerometer or other sensor, or an explicit input device such as a joystick on a ring, or an input pad on a wearable device such as a smartwatch. User permission is obtained to activate such image or gesture capture hardware.

The gestures can include, e.g., specific hand gestures such as those used in American Sign Language, to indicate special intent. Some examples of special intent are:

- treat the speech as punctuation (i.e., transliterate "period" to ".");
- treat the speech as a word processing command (i.e., transliterate "excuse me, Pittsburgh" as deleting the previous phrase and replacing it by "Pittsburgh");
- treat the speech as a non-word processing command (e.g., deliver a command to "increase screen brightness" or "send call to voicemail" without exiting dictation); and
- let the device know that the user is pausing to collect their thoughts and that it should not exit speech transcription mode due to the silence.

While the foregoing discussion refers to gestures, facial expressions can also be used. For example, winking with the left eye, gazing in a specific direction (e.g., up, down, left, or right), or head shaking or nodding, can be configured to be associated with specific special intents. Similarly, a certain number of finger taps can be used to convey a specific intent, such as tapping a specific finger twice to enter a command intent. This type of input may be perceived as less natural, but can offer higher input rates and enable more complex commands for advanced users.

The use of speech and gesture input jointly allows users to perform complex text input tasks by simply speaking to their device, without having to resort to other input devices such as a keyboard or a mouse/pointing device. It can enable higher data input rates by avoiding
cumbersome switches between speech processing and command entry modes. The resultant user experience of speech input is robust and powerful. The techniques are also valuable for users that use devices that lack traditional input hardware and for users with a physical impairment that precludes them from using typing-based input devices.

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 or gestures, documents, 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

This disclosure describes the utilization of a combination of speech analysis and gesture recognition to automatically disambiguate between verbatim text input (dictation) and commands. User provided speech and gestures are analyzed and used for interpretation of the spoken input, without the user having to switch between text entry and command entry modes.