Technical Disclosure Commons

Defensive Publications Series

February 23, 2018

Simultaneous multimodal user interface

Sandro Feuz

Thomas Deselaers

Simultaneous multimodal user interface

ABSTRACT

This disclosure describes techniques to use a combination of voice based input with other input mechanism in human-computer interfaces. The techniques enable users to utilize in real-time, the suitable input mode(s) in a given context, without having to switch between input modes. With user permission, speech analysis techniques are utilized to analyze user speech and detect when speech includes user instructions, and to determine corresponding actions to be performed. By enabling simultaneous user input via multiple modes, the techniques facilitate effective navigation of complex tasks performed using a computing device.

KEYWORDS

- Multimodal interaction
- Voice input
- Tone analysis
- Speech recognition
- Natural language processing
- Word processor

BACKGROUND

Users interact with computing devices through different input mechanisms or modes. For example, some common input modes include interactions via a keyboard, mouse, touchscreen keyboards, voice, etc. Each of the above modes of interaction has different advantages and are utilized independent of the other modes.

In some contexts, using combinations of these input mechanisms or modes simultaneously can be advantageous. For example, in word processing, while some users input text using a keyboard and choose formatting options with a mouse, other users utilize keyboard shortcuts instead of a mouse, thereby avoiding switching between the keyboard and mouse for text entry and formatting respectively.

In current computer systems, the commonly used modes, e.g., keyboard, mouse, touchscreen, and voice, are provided such that the users are provided with limited ability to employ multiple input modes simultaneously.

DESCRIPTION

Techniques described herein enable users to provide voice inputs simultaneously along with input via input devices such as keyboard, mouse, or touchscreen. For example, the techniques enable users to interact with text and word processing applications by utilizing a combination of keyboard and voice inputs. By simultaneously using voice with another input mechanism, users can seamlessly utilize multiple suitable modes for a given context without having to switch between input modes. By enabling users to provide input via multiple input mechanisms simultaneously, the techniques facilitate effective navigation of complex tasks. The described techniques can be incorporated within an operating system or in a software application.

Users provide voice inputs to a software application (e.g., a text application such as a word processor) using natural language via multiple input mechanisms that include voice and one or more of keyboard, mouse, touchscreen, or other mechanisms simultaneously. For example, implementing these techniques enables users to correct typographical errors that occur while typing using the voice interface, instead of retracing to fix such errors using a keyboard.

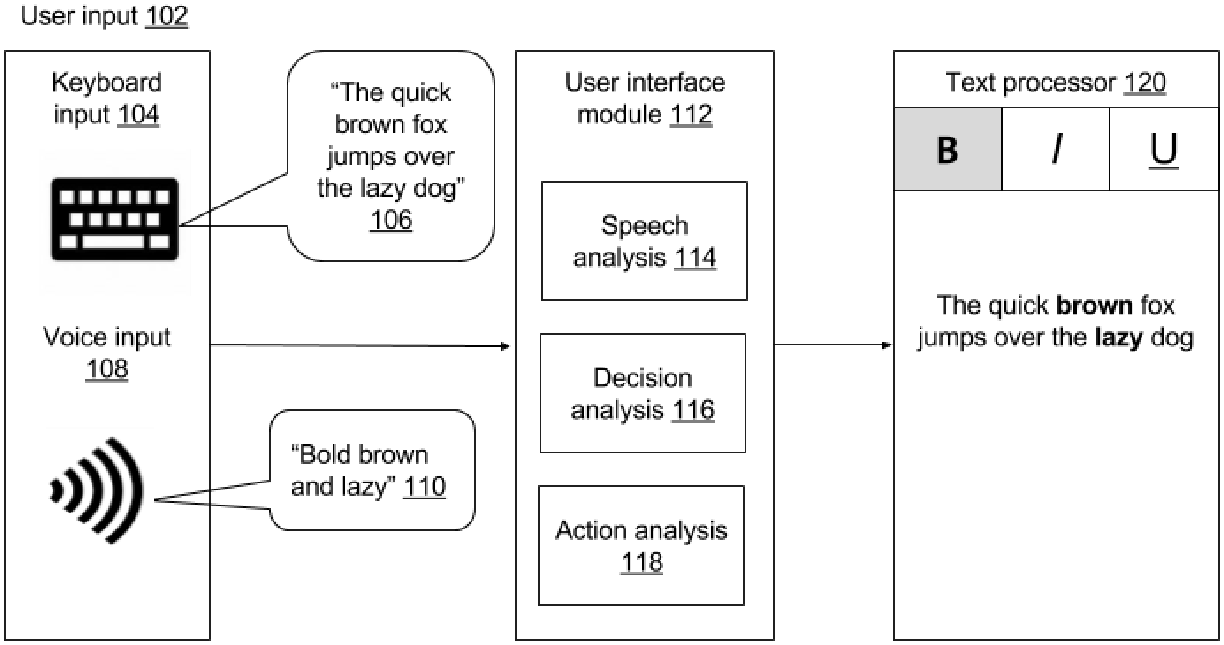
User input 102
Keyboard input 104
"The quick brown fox jumps over the lazy dog" 106
Voice input 108
"Bold brown and lazy" 110
User interface module 112
Speech analysis 114
Decision analysis 116
Action analysis 118
Text processor 120
B
I
U
The quick **brown** fox jumps over the **lazy** dog

Fig. 1: Example simultaneous multimodal interaction (voice + keyboard)

Fig. 1 illustrates simultaneous user input (102) with a combination of keyboard (104) and voice (108) inputs to a user interface (112) of a computing device. As illustrated in Fig. 1, at the same time as typing the text "the quick brown fox jumps over the lazy dog" (106) via a keyboard, the user also specifies a boldface font for the words "brown" and "lazy" (110) using voice input. In response to the user voice instruction "bold brown and lazy," a text processor application (120) formats the words "brown" and "lazy" in the sentence to utilize boldface. The user can specify a larger or smaller span of text by suitable instructions; for example, by specifying "more" or "less."

With user permission and express consent, the techniques employ a speech analysis module (114), a decision analysis module (116), and an action analysis module (118). Machine learned models, e.g., neural networks, are utilized to implement the modules. While Fig. 1 shows three distinct modules, the modules can be optionally combined or implemented as a single module.

The speech analysis module filters out background (ambient) noise from the voice input and separates user speech from extraneous sounds. Further, the speech analysis module recognizes user speech based on speaker identification (automatic speech recognition).

The decision analysis module determines, with user permission, whether the identified user speech is a user voice instruction. For example, if the decision analysis module detects that a user is speaking to another person, the corresponding portion of speech is excluded from interpretation for text processing. Further, the decision analysis module can also analyze the tone of the user's voice along with the context for the detection of instructions.

Based on a determination that the recognized speech is relevant (e.g., is a command), the action analysis module determines the actions to be taken in response to the user speech. In some instances, the module may determine that taking no action is an appropriate response to the recognized speech. In order to avoid unintended formatting or spelling changes, an explicit action is triggered specifically when user speech is identified as having a high probability (e.g., that meets a threshold) as being intended for interpretation and execution by the voice interface. In the context of applications that process text, the described techniques are an effective alternative to the use of keyboard shortcuts. For example, a user can issue voice commands to correct typographical errors even while entering additional text via a keyboard.

In some user interfaces, users are required to scroll through a list of options to find a specific option when using touchscreen, keyboard, mouse, or gesture inputs. In these contexts, a voice interface presents a simpler solution for a user to access the particular option, e.g., by saying aloud the name of the option. For example, to select German as an option from a list of languages, a user can simply say “German,” while a scrolling list interface is active. In response, the computing device can scroll to the corresponding entry in the list.

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

This disclosure describes techniques to use a combination of voice based input with other input mechanism in human-computer interfaces. The techniques enable users to utilize in real-time, the suitable input mode(s) in a given context, without having to switch between input modes. With user permission, speech analysis techniques are utilized to analyze user speech and detect when speech includes user instructions, and to determine corresponding actions to be performed. By enabling simultaneous user input via multiple modes, the techniques facilitate effective navigation of complex tasks performed using a computing device.