January 2022

Improved Dictation Using Virtual Assistant

Eric Stavarache

Henry Dlhopolsky

Improved Dictation Using Virtual Assistant

ABSTRACT

Writing on small form factor devices such as smartphones is inefficient due to the small dimensions of such devices. This disclosure describes techniques that support natural-language, context-aware voice commands that refer to individual words, sentences, paragraphs, sections, chapters, etc. of text content. Various features are supported, such as: automatically splitting text into paragraphs; creating lists; using different fonts for subheadings, chapter titles, etc.; inserting footnotes; adding citations and references; etc. To assist the user in navigating the document, the transcribed text is displayed overlaid with annotations that indicate available commands; numbered textual components; etc. Machine learning and natural language processing techniques are used to automatically differentiate between text and commands.

KEYWORDS

- Speech input
- Dictation
- Automatic speech recognition (ASR)
- Speech-to-text
- Word processing
- Command recognition
- Punctuation
- Virtual assistant
- Voice assistant
- Machine learning
- Natural language processing (NLP)

BACKGROUND

Writing on small form factor devices such as smartphones is inefficient due to the small dimensions of such devices. Virtual keyboards attempt to make smartphone typing easier. For example, typing via swipe gestures enables a user to drag a finger across the virtual keyboard, with text prediction techniques being deployed to complete the word intended by the user. Several input methods are available to make typing in Japanese more efficient [2]. Yet, these cannot compare to typing on a keyboard or to using voice input. For example, the average typing speed on mobile devices is 38 words per minute (WPM) compared to 52 WPM on a standard PC keyboard and 150 WPM while talking. Adding punctuation marks further slows down phone typing, since a change in virtual keyboard layout is often needed to enter punctuation marks.

The fastest technique to write a document is voice dictation. However, voice dictation doesn't lend itself to easy content organization or to fixing mistakes. At best, smart dictation enables the creation of paragraphs or fixing errors at the current cursor location. In particular, smart dictation does not support arbitrary navigation within a document, e.g., commands such as 'go back to the third sentence.' Smart dictation currently can control apps, but only with basic commands, and does not support an end-to-end writing experience suitable for long form writing such as books.

DESCRIPTION

This disclosure describes techniques that support natural-language, context-aware voice commands such as 'next paragraph,' 'next chapter,' 'next section,' 'move this paragraph after the next one,' etc. Within a given paragraph, sentence-based commands such as 'start writing after the second sentence,' etc. are supported. Various features are supported, such as: automatically

splitting text into paragraphs; creating lists; using different fonts for subheadings, chapter titles, etc.; inserting footnotes; adding citations and references; etc.

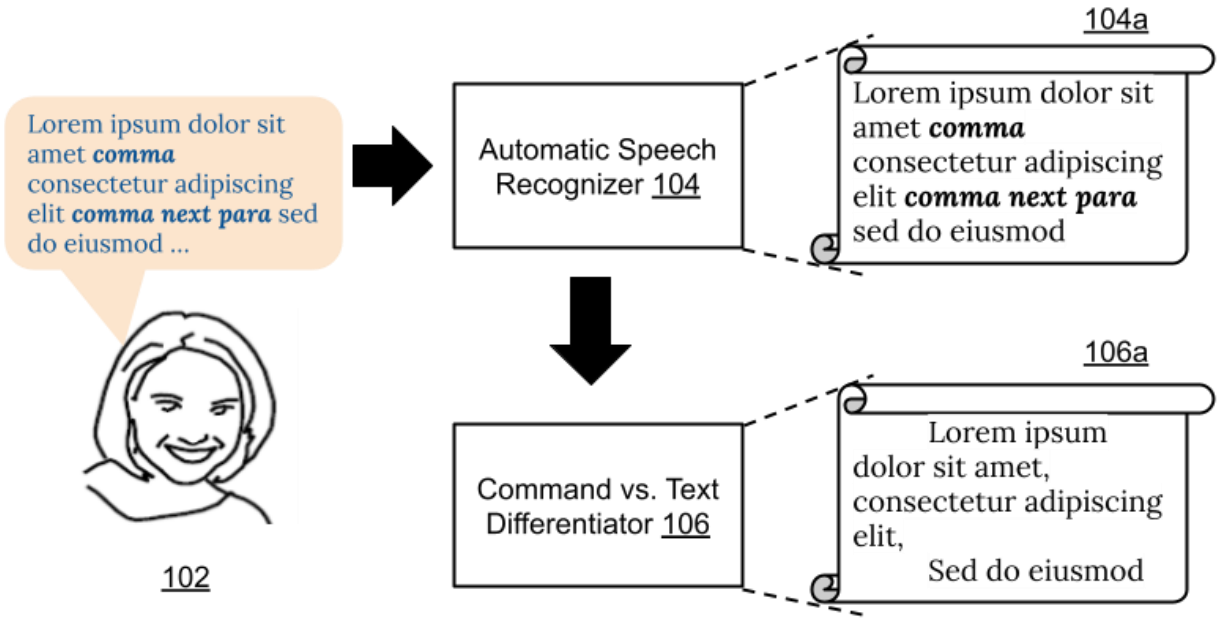

Fig. 1: Writing text using a virtual assistant

Fig. 1 illustrates writing text using a virtual assistant that implements techniques described herein. A user (102) utters aloud sentences that include text (‘lorem ipsum …’), punctuations (‘comma’), and commands (‘next para’). An automatic speech recognizer (104) transcribes the user’s speech, producing a transcription (104a) that treats punctuations and commands as equivalent to text. A command-vs-text differentiator module (106) differentiates between commands, text, and punctuation, and produces a modified transcription (106a) that reflects the text structure (paragraphs, punctuations, font, etc.) intended by the user.

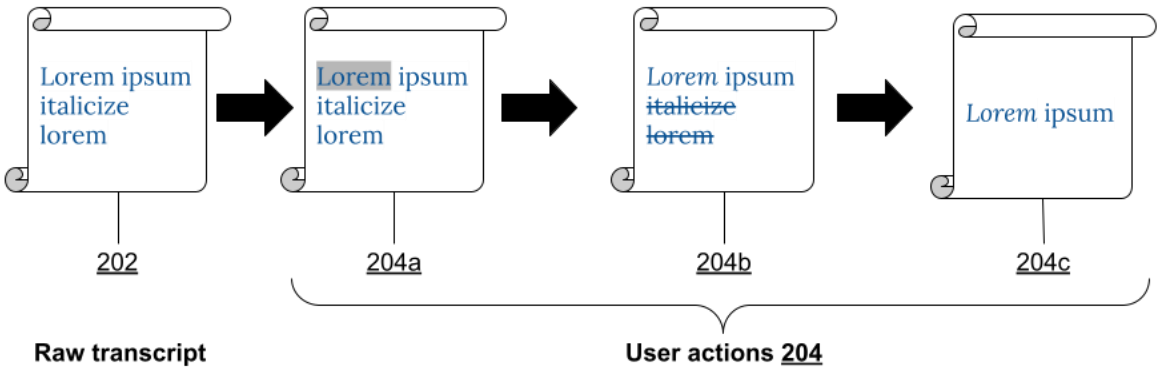

Fig. 2(a)

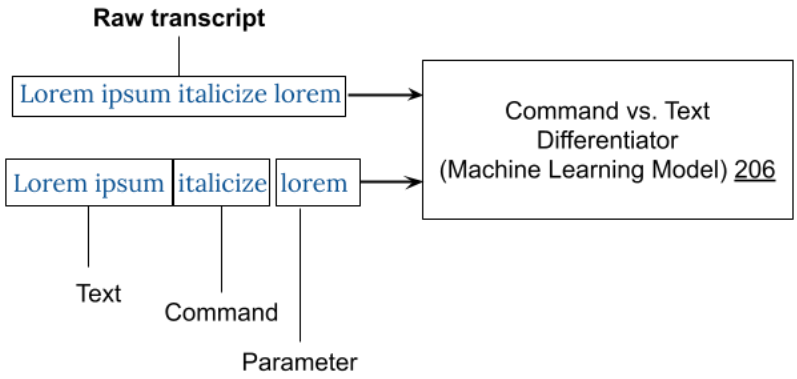

Fig. 2(b)

The command-vs-text differentiator module can implement a machine-learning model that is initialized with a set of commands, e.g., ‘next para,’ ‘delete the last word,’ ‘change font to Arial,’ etc. The model can be trained using raw transcripts with components labeled as text, command, punctuation, command-parameters, etc.

For example, as illustrated in Fig. 2(a), a raw transcript (202) comprises the phrase ‘lorem ipsum italicize lorem,’ and the user action (204) that follows transcription includes highlighting ‘lorem’ (204a); italicizing ‘lorem’ (204b); and deleting ‘italicize lorem’ (204c). This indicates that the command-vs-text differentiator ML model (206, Fig. 2b) can be trained with the raw transcript ‘lorem ipsum italicize lorem’ labeled as ‘lorem ipsum (text)’; ‘italicize (command)’; and ‘lorem (parameter)’. The labeling can be manual (e.g., human-labeled) or

automatic (e.g., based on clustering). Natural language processing (NLP) techniques can also be leveraged to support natural language commands, such as ‘change everything to font Arial’. Heuristics can also be leveraged to support command recognition. For example, a phrase ‘next chapter’ (found using regular expression searching) followed by a pause can indicate that the phrase is a command.

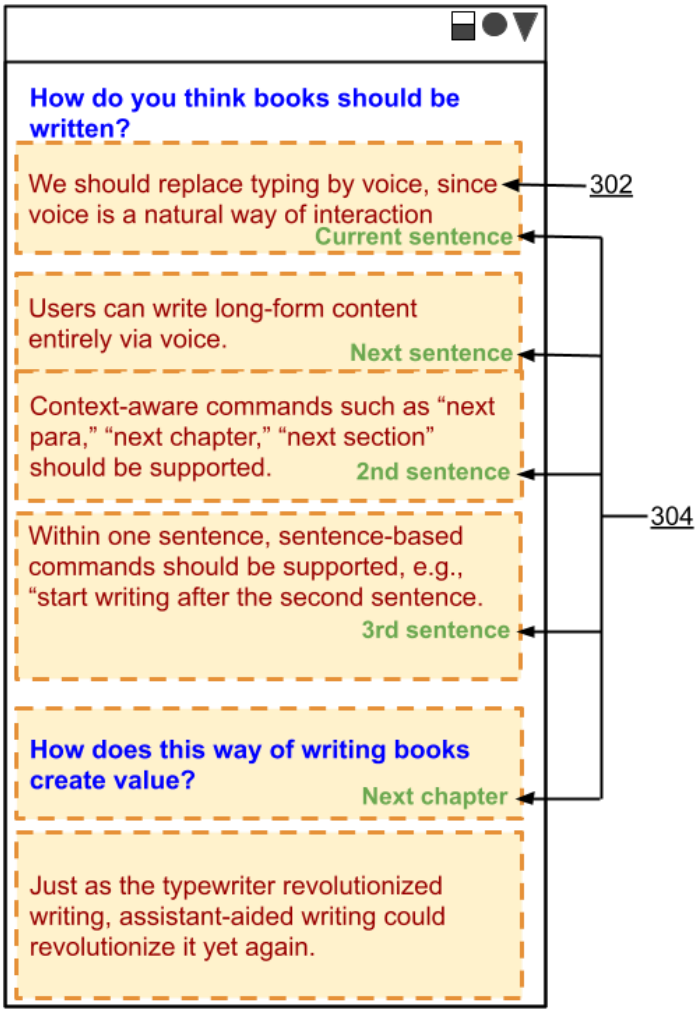

Fig. 3: User interface overlaid with in-text annotations

Furthermore, to assist the user in navigating the document, as illustrated in Fig. 3, the transcribed text (302) can be overlaid with annotations (304, shown in green). The annotations can indicate to the user, for example, the commands that are available at that point in the text,

thereby relieving the user of having to memorize commands; numbered components of the text (first sentence, tenth paragraph, etc.) so that the user can easily speak commands relating to numbered textual components (‘move the tenth para above the sixth’); etc.

In this manner, the techniques enable a dynamic writing style while providing voice input as compared to current techniques of touchscreen textual input. The techniques can enable a user to dictate to a virtual assistant at an efficiency similar to dictating to a person. Users can easily create and share long form writing such as reports and stories much faster. Given the substantially higher words-per-minute rate of speech input compared to typing (especially on a small screen device), document creation speed can be improved across multiple types of devices. The described in-context text annotations enable dynamic command discovery and can obviate the need for users to consult help pages. The in-context annotations also help users learn commands.

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 the collection of user information (e.g., information about a user’s spoken input, a user’s language, or 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 techniques that support natural-language, context-aware voice commands that refer to individual words, sentences, paragraphs, sections, chapters, etc. of text content. Various features are supported, such as: automatically splitting text into paragraphs; creating lists; using different fonts for subheadings, chapter titles, etc.; inserting footnotes; adding citations and references; etc. To assist the user in navigating the document, the transcribed text is displayed overlaid with annotations that indicate available commands; numbered textual components; etc. Machine learning and natural language processing techniques are used to automatically differentiate between text and commands.

REFERENCES

[1] "Type with your voice - Docs Editors Help" available online at https://support.google.com/docs/answer/4492226?hl=en accessed Jan. 9, 2022.

[2] "Japanese input method - Wikipedia" available online at https://en.wikipedia.org/wiki/Japanese_input_method accessed Jan. 9, 2022.

[3] "Nuance Dragon Dictate for Mac - User Manual" available online at https://www.nuance.com/content/dam/nuance/en_us/collateral/dragon/guide/gd-dragon-dictate-for-mac-user-manual-en-us.pdf accessed Jan. 9, 2022.

[4] "Zoho brings Zia AI assistant to its office apps | Computerworld" available online at https://www.computerworld.com/article/3340129/zoho-brings-zia-ai-assistant-to-its-office-apps.html accessed Jan. 9, 2022.

[5] "Use Voice Control on your Mac - Apple Support" available online at https://support.apple.com/en-us/HT210539 accessed Jan. 9, 2022.