Technical Disclosure Commons

Defensive Publications Series

May 2021

MULTIMODAL CONTENT EDITING

Matt Sharifi

MULTIMODAL CONTENT EDITING

ABSTRACT

This publication describes systems and techniques for multimodal content editing that enables a user to edit or extend content in a given modality using a variety of different input modes while matching the underlying format of the content and while preserving the original input. When the user uses an input mode different from the underlying format of a piece of content to provide input at a computing device to edit or extend the piece of content, the computing device may convert the input provided by the user from the input mode to the underlying format and may edit the piece of content based on the converted input. For example, if the user uses voice input to edit a text document, the computing device may convert the voice input into text and may include the converted text in the text document. In another example, if the user uses text input to edit an audio recording, the computing device may convert the text input into audio using a text-to-speech technique and may include the converted audio in the audio recording.

The computing device may also preserve the provided input in its original form, such as by storing the provided input at the computing device, and may associate the stored provided input with the converted input so that the user may be able to refer back to the originally provided input. For example, if the computing device converts text input into audio for inclusion in an audio recording, the computing device may store the text input and may link the audio recording to the text input so that a user may be able to view the text input while listening to the audio recording.

DESCRIPTION

Mobile computing devices, such as smartphones, may enable a user to input text in a variety of modalities, such as by typing text at a physical or virtual keyboard, providing voice input, providing handwritten input, and the like. Different pieces of content may have different underlying formats. For example, a text document may have an underlying format of text while a voice recording may have an underlying format of audio. Typically, a user may use an input mode that matches the underlying format of a content to edit or extend the content. For example, a user may type text to edit or extend a text document or may provide voice input to edit or extend a voice recording.

However, in certain situations, it may be inconvenient for the user to input text to edit or extend a document using an input mode that matches the underlying format of the document. For example, excessive ambient noise may make it difficult for the user to provide voice input to edit or extend a voice recording. In another example, if the user is performing other tasks with his or her hands, the user may be unable to input text by typing text to edit or extend a text document such as an e-mail message, a word processing document, and the like.

This publication describes systems and techniques for multimodal content editing that enables a user to edit or extend content having an underlying format at a computing device using input modes that do not match the underlying format of the content. When the user uses an input mode different from the underlying format of a content to provide input at a computing device to edit or extend the content, the computing device may convert the input provided from the input mode to the underlying format and may edit the content based on the converted input. The computing device may also preserve the provided input in its original form and may associate the provided input with the converted input.

For example, when the user provides voice input to edit a text document at a computing device, the computing device may perform speech recognition on the voice input to convert the voice input into text and may add the text converted from the voice input into the text document. The computing device may also preserve the voice input provided by the user and may associate the voice input with the text converted from the voice input in the document, so that the computing device may be able to play back the voice input, such as when the user selects the text converted from the voice input in the document. The computing device may also include, in the text document, an indicator that the text converted from voice input was synthetically generated (i.e., not originally inputted as text by the user).

In another example, a user may provide text input, such as by providing input at a keyboard (e.g., a physical keyboard or a virtual keyboard) of a computing device to edit an audio recording, such as a voice memo. When the user provides the text input, the computing device may perform text-to-speech conversion of the text input to convert the text input into audio and may add the audio converted from the text input into the audio recording. The computing device may also preserve the text input provided by the user and may associate the text input with the audio converted from the text input in the audio recording. The computing device may also include, in the audio recording, an indicator that the audio converted from text input was synthetically generated (i.e., not originally inputted as audio by the user).

A user may create a piece of content having an underlying format at a computing device. For example, a user may create a draft e-mail message having an underlying format of text, a word processing document having an underlying format of text, a notes document having an underlying format of handwriting, a voice memo having an underlying format of audio, and the like. The user may edit the piece of content by providing input at the computing device using an

input mode that matches the underlying format of the document, and may save the edited content. For example, the user may edit the word processing document by typing text on a virtual keyboard outputted by the computing device, edit the notes document by providing handwriting gestures using a stylus, edit the voice memo by providing voice input at the computing device, and the like.

When the user subsequently edits a previously-created piece of content having an underlying format at a computing device, the user may continue to edit the piece of content using an input mode that matches the underlying format of the document, or may switch to using a different input mode that does not match the underlying format of the piece of content. The user may switch to using a different input mode that does not match the underlying format of the piece of content to, for example, edit the word processing document by providing voice input, edit the notes document by typing text on a virtual keyboard outputted by the computing device, edit the voice memo by typing text on a virtual keyboard outputted by the computing device, and the like. In some examples, the user may explicitly indicate to the computing device that the user is switching to using a different input mode to edit the piece of content. In other examples, the computing device may automatically detect that the user is editing the piece of content using a different input mode.

If the computing device determines that the user has switched to using a different input mode that does not match the underlying format of the piece of content to edit the piece of content, the computing device may, in response to the user providing input using the input mode that does not match the underlying format of the piece of content to edit the piece of content, convert the provided input into text if the provided input is not already text. For example, if the user is providing voice input to edit a word processing document, the computing device may use

speech recognition to convert the voice input provided by the user to text. In another example, if the user is providing handwritten input to edit an e-mail message, the computing device may perform handwriting recognition to convert the handwritten input provided by the user to text. In another example, if the user is providing text input by typing text on a virtual keyboard to edit a voice memo, the computing device may not need to convert the inputted text into text.

The computing device may convert the input provided by the user to edit the piece of content to the underlying format. Because the computing device converts the provided input into text if the provided input is not already text, the computing device may convert the text associated with the provided input into the underlying format. For example, if the piece of content being edited is a word processing document, the computing device may add the text associated with the provided input in the word processing document.

In another example, if the piece of content being edited is a notes document containing handwritten notes, the computing device may convert the text associated with the provided input into handwriting and may include the handwriting in the handwritten notes. The computing device may convert the text associated with the provided input into handwriting (i.e., stroke form) using any suitable technique, such as using a standard generative model trained on the reverse of a handwriting recognition task. The computing device may condition such a technique to convert text to handwriting on a vector which encodes the user's handwriting style to generate handwriting having attributes that match the handwriting in the rest of the handwritten notes document.

In another example, if the piece of content being edited is a voice memo, the computing device may convert the text associated with the provided input into audio, such as by using a text-to-speech engine, and may include the audio in the voice memo. In some examples, the text-

to-speech engine may use a speaker embedding to condition the text-to-speech process to generate spoken audio that sounds similar to the voice of the user. In some examples, the text-to-speech engine may produce audio that matches some characteristics of the user's speech (e.g., pitch, speed, etc.) while sounding different from the user in other ways to clearly indicate that the generated audio was not originally spoken by the user.

Besides converting the input provided by the user to edit the piece of content to the underlying format, the computing device may also preserve the original input provided by the user in the original input mode used by the user, such as preserving the handwritten input provided by the user, the voice input provided by the user, the text inputted by the user, and the like. The computing device may, after receiving explicit permission by the user, store the original input provided by the user at the computing device, at an external storage device coupled to the computer, at an external system (e.g., the cloud), and the like.

In some examples, the computing device may perform filtering of the original input provided by the user before storing the originally provided input. For example, if the original input provided by the user is in the form of voice input, the computing device may perform filtering of the voice input to remove background noises, to crop silences, and the like prior to storing the filtered input.

The computing device can associate the stored original input provided by the user with the piece of content being edited. For example, the computing device may store, in the piece of content being edited, an identifier that links to the stored original input provided by the user. For example, if the original input provided by the user is voice input, and if the underlying format of the piece of content is text, the computing device may store an identifier in the piece of context

that links to the original input, such that when the user selects text in the piece of content that is converted from the voice input, the computing device may audibly output the stored voice input.

It is noted that the techniques of this disclosure may be combined with any other suitable technique or combination of techniques. As one example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2019/0056909 A1. As another example the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2004/0243415 A1. As another example the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2005/0177369 A1. As another example the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2003/0185444 A1. As another example the techniques of this disclosure may be combined with the techniques described in International Patent Publication No. WO 2000041166 A2.