August 2020

# Automatic Transcription of Voice Notes Sent Via Messaging Applications

Anonymous



# Automatic Transcription of Voice Notes Sent Via Messaging Applications


## ABSTRACT

Sending a voice note over a messaging app is a convenient mechanism for the sender. However, the recipient of a voice note may perceive the voice note to be inconvenient and imposing on their time. This occurs since speaking is faster than typing and is more convenient for the sender, but listening is usually slower than reading and is less convenient for the recipient. This disclosure describes a messaging application that automatically converts a sender's voice note to text and provides an option for the sender to send the transcribed text as an attachment with the voice note. The techniques also enable the recipient of a voice note to independently convert it to text. At the recipient end, the recipient can choose to look at the transcribed text or hear the voice note or both, and search for words present in the text version of the voice note.



## KEYWORDS

- Speech-to-text
- Speech recognition
- Voice note
- Messaging app
- Chat application


## BACKGROUND

Sending a voice note over a messaging app is a convenient mechanism for the sender. In particular, if the sender's hands or eyes are busy, e.g., walking down a busy street, pushing a baby carriage, etc., it is easier for the user to provide spoken input than to type out a message. However, the recipient of a voice note may perceive the voice note to be inconvenient and

imposing on their time. This occurs since speaking is faster than typing and is more convenient for the sender, but listening is usually slower than reading and is less convenient for the recipient. Also, listening to a voice note may require using headphones due to various factors such as noise, requirements of privacy, etc. Further, voice messages are unsuitable when one or more recipients (e.g., in a group chat conversation) have hearing-related disabilities, or are in situations where audio playback is not feasible. However, a sender may not be aware of such circumstances.

It is possible to dictate text and have a device such as a smartphone that runs a chat application automatically convert the speech to text using speech-to-text techniques. However, such input may require a mode switch within the chat application, or the use of a separate speech-to-text application, and is therefore cumbersome. Also, speech recognition can be imperfect, causing a sender to correct the transcribed text (thereby defeating the purpose of dictating to avoid typing), or send as-is and hope that the recipient interprets things correctly based on the context of the conversation.

**DESCRIPTION**

When using speech-to-text transcription to send a message via a messaging application, the sender only sends text, and not a voice message, which needlessly limits the options of a receiving user. For example, sending just the speech-to-text transcription fails to leverage the emotive qualities of voice, which can convey nuance better than text. This disclosure describes techniques to convert a sender's voice note to text within a messaging application.

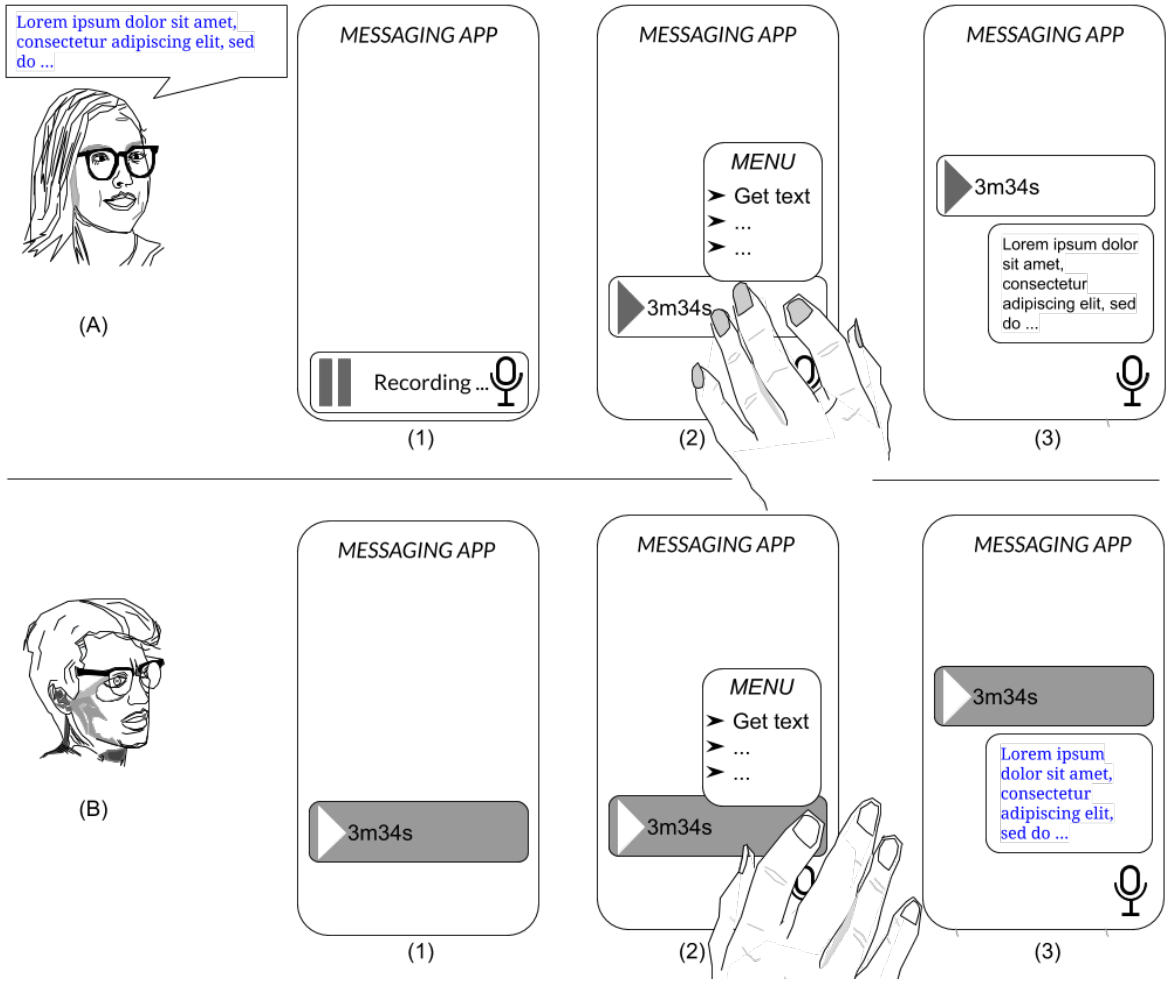


**Fig. 1: Speech-to-text for voice notes in messaging apps: (A) Actions at sender end; (B) Actions at recipient end**

As illustrated in Fig. 1, a sender first records a voice note within the messaging application (Fig. 1A-1). The sender then can activate the speech-to-text functionality provided within the messaging app. For example, such activation can be by long-pressing the voice note to activate a menu option (Fig. 1A-2) that is selectable to convert the voice note to text and display the text transcript alongside the voice note (Fig. 1A-3). At the recipient end, upon receiving the voice note (Fig. 1B-1), the recipient can select a menu option (Fig. 1B-2) within the messaging application to reveal the text attached to the voice note (Fig. 1B-3). Regardless of whether or not the sender attached transcribed text, the recipient can independently transcribe a received voice

note into text. The transcribed text can also be cached at the recipient device, so that subsequent requests for the transcribed text can be rapidly served.

The speech-to-text conversion can be performed by the messaging application at the client at either the sending or the receiving ends, or at a server. If performed by a client, the message can preserve end-to-end encryption. However, client-side speech-to-text conversion is sometimes less accurate and can increase battery consumption. Depending on the cost and performance of the speech-to-text conversion, e.g., based on accuracy, battery drain, and user preferences for privacy, the sender and the receiver can be provided a choice (with a default option) to use client-based or server-based speech-to-text conversion. Although both the voice note and the transcribed text can be sent to the recipient, it may be advantageous to perform the speech-to-text conversion at a server since the accuracy of transcription may be higher, e.g., if the server has more recent speech recognition models with greater accuracy. In such a case, the sender and/or the receiver can select to have speech-to-text conversion done at the server. Regardless of speech-to-text conversion being done by the (sending or receiving) client or at a server, the text content of the voice note is available for search.

Further, speech recognition models that recognize mood, emotion, intent, setting, etc., from the context of the voice note can be used to annotate these into the transcribed text, thereby improving the accuracy and the accessibility of the voice note. Mood, emotion, intent, and setting related to the voice-note serve a function similar to subtitles of a movie, where they describe and set the scene for differently-abled audiences or for audiences that don’t speak the language of the movie. These are also similar in function to descriptives added to images, which assist users with visual disabilities.

Speech-to-text conversion can be performed in the background, e.g., while the voice note

is being composed at the sender, or prior to hearing or viewing at the receiver. If the sender selects, the text can be sent in the background and attached to the voice note, such that a recipient can toggle the display of the text on or off using menu options provided in the user interface of the messaging application. Regardless of whether or not a sender attaches transcribed text, a voice note received by a receiver can be converted to text by the receiver, either at the receiving client, or by the receiving client invoking speech-to-text conversion at a server.

In this manner, this disclosure provides a mechanism and an example user interface to enable users of messaging applications to exchange voice notes transcribed to text. A recipient user can choose to listen to the voice note and/or read the transcribed text.

## CONCLUSION

This disclosure describes a messaging application that automatically converts a sender's voice note to text and provides an option for the sender to send the transcribed text as an attachment with the voice note. The techniques also enable the recipient of a voice note to independently convert it to text. At the recipient end, the recipient can choose to look at the transcribed text or hear the voice note or both, and search for words present in the text version of the voice note.