November 2022

Visual Depiction of Voice Message Content

Fuxiao Xin

Shumin Zhai

Visual Depiction of Voice Message Content

ABSTRACT

Voice messages, while faster to input, are inconvenient if the message recipient is in a situation where they are not able to listen to the message. This disclosure describes techniques that, with permission, automatically analyze incoming voice messages and provide visual information within a messaging application to enable the user to understand the contents of the incoming voice messages at a glance. The glanceable visual information is derived by generating concise text of the message content, determining the message sentiment, analyzing the prosody and emotion of the voice, and adding emojis that correspond to the content, prosody, and emotion. The visual information is provided attached to the voice message in a user interface, which enables recipients to know about message contents immediately, without having to listen to the voice message. The techniques can thus improve the convenience and user experience of interacting with incoming voice messages within any application or platform.

KEYWORDS

- Voice message
- Audio message
- Speech prosody
- Speech disfluency
- Sentiment analysis
- Emoji prediction
- Automatic speech recognition (ASR)
- Natural language processing (NLP)

BACKGROUND

Many users of messaging and social media and chat applications communicate with others by sending messages by recording their voice rather than typing content in the form of a text message. Sending voice messages can be substantially faster because people typically speak at the rate of about 150 words per minute. In contrast, even professional typists can type only about 60-80 words of text per minute. Moreover, recording and sending voice messages can be more intuitive and easier to use, especially for those who are not used to typing text on their devices on a regular basis. Further, voice messages are more convenient in situations such as driving or walking when it is not readily possible to type. The recipients may also find voice messages more personal and intimate than simply reading them in text form.

Despite the various advantages, voice messages have a few important limitations. Recipients of voice messages may not be able to listen to the messages right away if they are not in the right setting to do so. For instance, recipients might be in settings that require silence such as libraries, classrooms, meetings, etc. or in a public place where others may be able hear the message such as on public transportation, at an airport, etc. In such cases, recipients need to wait to access the content of the voice message until they are in a setting where it is feasible to listen to the message or wear earphones. Until then, recipients cannot obtain any information about the contents of a received message since the applications show only that an audio file with a voice message was received, with no indication of the message contents. When recipients are in situations in which they cannot access the contents of voice messages immediately, delayed access can impact the communication.

DESCRIPTION

This disclosure describes techniques that, with user permission, analyze received voice messages within applications and derive visual information that can be shown below the messages to signal the nature of their contents at a glance. The glanceable visual information is obtained by generating concise text of the message content, determining the sentiment contained in the text, analyzing the prosody and emotion of the speaker's voice, and adding emojis that correspond to the content, prosody, and emotion. Delivering voice messages with the visual information attached makes it possible for a recipient to learn about message contents immediately upon receiving a message, even when the recipient is in a situation that prohibits them from listening to the voice message.

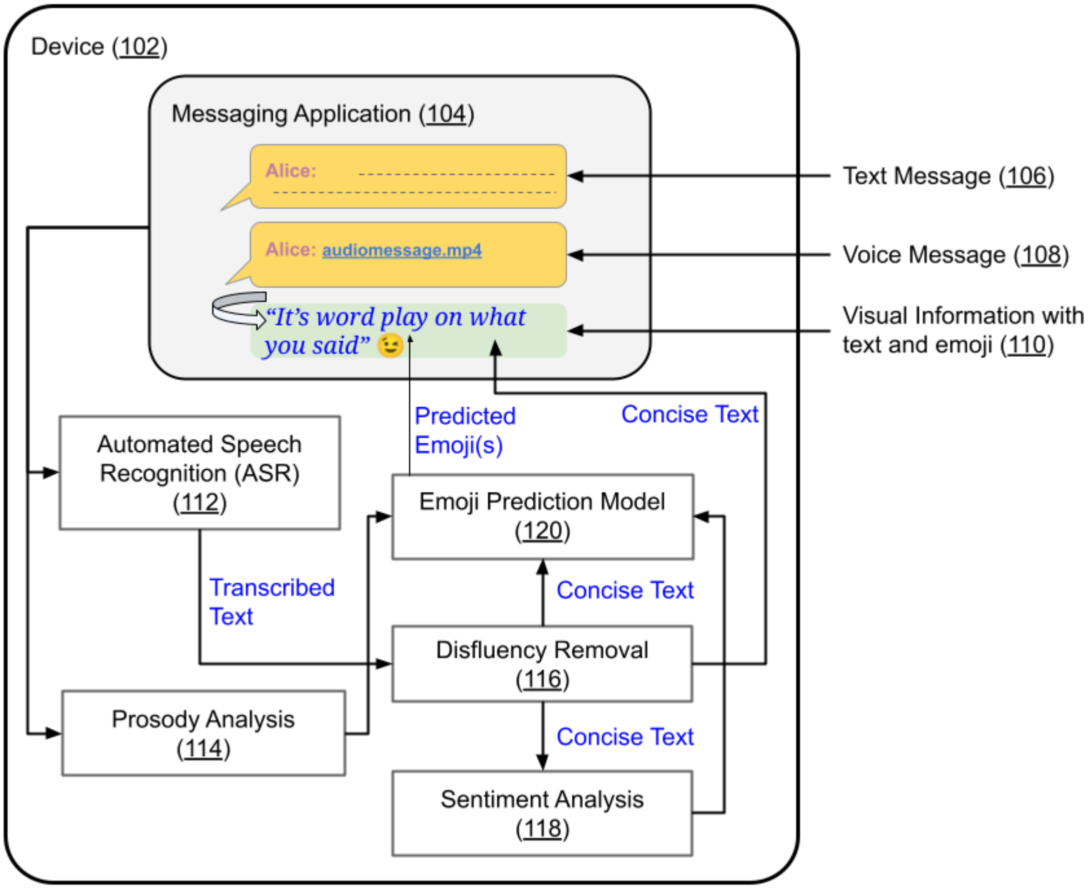

Fig. 1: Deriving text and emojis corresponding to voice messages

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. A user communicates with a contact using a messaging application (104) on a device (102). The contact can send messages by typing them as text (106) or recording the voice and sending it as an audio file (108). With user permission, the audio of the voice message is processed by automated speech recognition (ASR) techniques (112) to transcribe the audio and by prosody analysis tools (114) to analyze non-linguistic affect.

To enhance readability, the text transcription of the audio generated by ASR can be further processed via disfluency identification and removal tools (116) to make it concise by removing non-essential content, such as hesitations, repetitions, filler words, etc. For example, a verbatim transcription of the voice message can yield the text “But that's it's not…it's not…it's, uh, it's a word play on what you just said.” Removing disfluency results in converting the text to the concise form “But it’s word play on what you just said.” Sentiment analysis (118) of the concise text can help identify the emotional valence of the text.

Prosody analysis of the audio can be based on relevant non-linguistic emotional components, such as tone, pitch, volume, etc. Such affective aspects of prosody can indicate the intended emotional perception and can help avoid miscommunication. For instance, the same sentence can convey different meanings depending on whether the speaker’s tone is regular or sarcastic.

The results of the prosody and sentiment analyses can be combined for better detection of the speaker’s intent since the analyses pertain to non-linguistic and linguistic components of the message, respectively. Moreover, the combination can help deal with cases in which one of analyses is less reliable. For instance, if the speaker has a monotone voice, prosody cannot be determined with high confidence. In such cases, the results of the sentiment analysis can still serve as a source for determining the speaker’s intended emotion.

The concise text generated upon disfluency removal along with the results of the prosody and sentiment analyses can be input to a trained emoji prediction model (120). A generic model that predicts emojis based on the text and emotion information can be fine-tuned with permission for a specific user’s communication style and/or personal preferences.

The concise text and the predicted emoji(s) can be shown visually (110) below the received voice message. The text and the emoji(s) can be displayed in a suitable format that makes it visually obvious that these are automatically derived and connected to the received voice message. Glancing at the visual information can help the recipient quickly glean the content and emotional intent of the sender without playing the audio of the voice message. The recipient can subsequently listen to the voice message whenever convenient.

The techniques described herein can employ state-of-the-art ASR tools that can handle variations in accents and can operate locally on the user device with low latency, high reliability, and high accuracy. Moreover, the occasional errors in ASR can be tolerable since users are made aware that the text and emoji(s) attached to the voice message are automatically generated. Further, if the confidence level of ASR transcription and/or sentiment/prosody analysis falls below certain thresholds, no visual information is shown to the user, thereby reducing the likelihood of erroneous transcription being shown.

The techniques described in this disclosure can be implemented with user permission within any application or platform that supports interpersonal voice messages. The various audio and text processing components mentioned above can be implemented using any suitable state-of-the-art techniques. If users permit, one or more of the components can be implemented external to the user device, such as on a remote server. Implementation of the described techniques can help users understand the content and the intent of voice messages with a quick glance, without having to play the audio. The techniques can thus improve the convenience and user experience (UX) of interacting with incoming voice messages.

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, voice messages, chat messages, 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. The information obtained with user permission can be leveraged to create a personalized voice message experience for the user.

CONCLUSION

Voice messages, while faster to input, are inconvenient if the message recipient is in a situation where they are not able to listen to the message. This disclosure describes techniques that, with permission, automatically analyze incoming voice messages and provide visual information within a messaging application to enable the user to understand the contents of the incoming voice messages at a glance. The glanceable visual information is derived by generating concise text of the message content, determining the message sentiment, analyzing the prosody and emotion of the voice, and adding emojis that correspond to the content, prosody, and emotion. The visual information is provided attached to the voice message in a user interface, which enables recipients to know about message contents immediately, without having to listen to the voice message. The techniques can thus improve the convenience and user experience of interacting with incoming voice messages within any application or platform.

REFERENCES

1. “Google's new messaging app translates your voice into emoji” available online at https://mashable.com/article/google-supersonic-voice-emoji accessed Oct 9, 2022.
2. Hu, Jiaxiong, Qianyao Xu, Limin Paul Fu, and Yingqing Xu. "Emojilization: An automated method for speech to emoji-labeled text." In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-6. 2019.
3. Seroussi, Joseph and Doron Taboh. “Systems and methods for generating emotionally-enhanced transcription and data visualization of text” PCT Patent Application Number 2021260611 published Dec 30, 2021.