Context-dependent multi-mode chat

Sandro Feuz
Yannick Stucki
Jorim Jaggi
Adrian Roos

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation
Feuz, Sandro; Stucki, Yannick; Jaggi, Jorim; and Roos, Adrian, "Context-dependent multi-mode chat", Technical Disclosure Commons, (July 03, 2018)
https://www.tdcommons.org/dpubs_series/1290
Context-dependent multi-mode chat

ABSTRACT

Chat applications often support several communication modes, e.g., texting, voice messaging, voice/video calling, sending memes/GIFs/images, sending video messages, etc. In a group chat, a given participant may not be disposed to use every chat modality. For example, one participant may be in a quiet space and constrained to use only text messaging, another participant may be in a position to receive video but not turn on their own camera, and so on.

With user permission, this disclosure applies machine-learning techniques to detect the user’s context. Based on the context, the user-preferred input and output modes for a group chat are determined. For example, such determination of user-preferred input and output modes is based upon features such as sensor data, historical usage, etc.

KEYWORDS

- Chat application
- Messaging application
- Multi-mode chat
- Machine learning
- Context-dependent mode

BACKGROUND

A multi-participant chat session taking place in a multi-mode messaging application where different participants have different contexts is illustrated by an example.
Fig. 1 illustrates a chat session with four participants, e.g., that are chatting about planning a party. A first participant (102a) is in charge of shopping, e.g., for beverages, and starts a video call (102b) to show to the others the different options available in the store. This user is able to use several modes of the chat service, e.g., camera, voice, text, etc. A second participant (104a) is driving, and is able to use, e.g., via a hands-free mode, only the voice mode (104b) of the chat application. This user is in a position to listen and respond via voice but is unable to view the video stream, other media content (e.g., images), or text messages in the chat session.

A third participant (106a) is in a situation where he can see the video stream (106b) and also respond via voice or text, but is unable to turn on their own camera. A fourth user (108a) is at a meeting, and the only usable mode available is text messaging (108b). In this manner, each participant uses the chat application differently.
DESCRIPTION

This disclosure describes the use of machine-learning techniques that detect user context by use of user-permitted contextual features. Based on the user context, a determination is made of preferred input and output modes for users of a chat application. Further, the incoming and outgoing streams in the chat are converted, if necessary, to the user-preferred mode. For example, for a chat participant who is disposed to only use text, an incoming voice message is automatically transcribed to text. An indication of such mode conversion is provided.

Predicting user-preferred input and output modes

Fig. 2: Detecting user context and deducing user-preferred modes

Fig. 2 illustrates detection of user context and deducing user-preferred modes, per techniques of this disclosure. Different factors are provided to a machine-learning model (202) as input for the purpose of mode detection. The machine learning model is trained for such prediction and can be implemented, e.g., as an on-device model on a participant device. The factors include, for example,

- contextual features relating to the user (204), e.g., data from applications such as calendars, maps, etc. for which the user has provided consent.
- permitted sensor data (206), e.g., GPS, accelerometer data, etc.
historical group chat settings (208)

The machine-learning model predicts as output a prediction of the mode(s) (210) that are likely to be preferred by the user for input and output of the chat session. For example, a prediction from the model may be that the preferred input mode includes one or more of voice, video, text, etc. Such determination is based on using the contextual factors to deduce, for example, whether a participant user is in a public setting (e.g., meeting, cinema, restaurant, etc.) where disturbance is forbidden, whether the user is driving, whether the user is wearing a headset with a phone device that is in the chat session being inaccessible, e.g., in pocket, whether the user can view incoming video in the chat session or provide video to other participants, etc.

Additionally, the model produces a certainty score for each input and output mode. A high score indicates that the prediction about the preferred mode is likely to be accurate while a low score indicates the prediction may not be reliable. Regardless of the certainty score, e.g., even if the score is high, the user is provided with options to manually override the predicted mode, e.g., via prominently displayed switches, on-screen icons, etc. If the certainty score is below a threshold, the chat mode is not switched automatically, e.g., user input is sought regarding the preferred mode of chat input/output. Alternately, when the certainty score is low, the chat mode is set to a default mode set by the user.

The machine-learning model can be implemented as a multi-layer neural network, e.g., a long short-term memory (LSTM) neural network. Other types recurrent neural networks, as well as other types of models such as convolutional neural networks) can also be used. Further, techniques such as support vector machines, random forests, boosted decision trees, etc., can also be used.
Converting to and from preferred input and output modes

Once the preferred input and output modes are determined, incoming or outgoing streams at the user’s device are converted accordingly. This is explained with a few examples.

Example 1: The output mode of the chat application for a certain user is text only. Conversion of the incoming stream is performed per the below rules:

- a text message: do nothing;
- a voice message: transcribe the speech content using speech recognition techniques, e.g., a machine-learned speech recognition model that takes voice as input and provides text as output;
- an audio stream other than speech: classify the type of input (e.g., “classical music,” “applause,” “airplane starting,” etc.) and describe it using text;
- an image or video message: describe the content of the image or video, e.g., using image recognition or video summarization techniques.

Example 2: The output mode of the chat application for a certain user is voice only. Conversion of the incoming stream is performed per the below rules:

- a text message: use a text-to-speech synthesizer to voice out the messages;
- an audio message: do nothing;
- an image: use a machine-learning model to obtain content description and use text-to-speech to provide an audio of the description;
- a video: combine the audio from the video and a summary of the video, using, e.g., video summarization techniques, and provide the combined audio.

The incoming and/or outgoing streams are converted upon specific user permission. The user is provided with indications that mode conversion is on and is provided with options to
turn the conversion off and participate in the chat session using the original, unconverted media streams.

In addition, or alternatively to machine-learning based chat-mode determination and mode-conversion, heuristics, e.g., hand-crafted heuristics, can be used to determine and convert modes if necessary.

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

Chat applications often support several communication modes, e.g., texting, voice messaging, voice/video calling, sending memes/GIFs/images, sending video messages, etc. In a group chat, a given participant may not be disposed to use every chat modality. For example, one participant may be in a quiet space and constrained to use only text messaging, another participant may be in a position to receive video but not turn on their own camera, and so on.
With user permission, this disclosure applies machine-learning techniques to detect the user’s context. Based on the context, the user-preferred input and output modes for a group chat are determined. For example, such determination of user-preferred input and output modes is based upon features such as sensor data, historical usage, etc.