Technical Disclosure Commons

Defensive Publications Series

August 17, 2018

Assistance Tailored to User Mood

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, "Assistance Tailored to User Mood", Technical Disclosure Commons, (August 17, 2018)
https://www.tdcommons.org/dpubs_series/1416

This work is licensed under a Creative Commons Attribution 4.0 License.

This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

Assistance tailored to user mood

ABSTRACT

User interaction with assistive devices and applications often includes use of speech-based interfaces where contents of the user's speech serve as input. However, in real-life human conversations, people often use additional information beyond the content of speech, such as tone, pace, etc., to infer attributes of the speaker's emotional state and intention, and make nuanced adjustments to the style and content of their own speech. With user permission, the techniques of this disclosure build and utilize a machine learning model to infer mood based on an analysis of user speech. The inferred mood information is then applied to tailor the content and style of the speech output of a voice-based assistive device or application. The techniques enable user interaction that more closely resembles real-world human conversations.

KEYWORDS

- Speech content
- Mood inference
- Machine learning
- Voice interface
- Voice interaction
- Speech variation
- Smart speaker
- Virtual assistant
- Speech assistant
- Natural interaction

BACKGROUND

In real-life human conversations, people often use additional information beyond the content of speech, such as tone, pace, word choices, etc., to infer attributes of the speaker's emotional state and intention, and make nuanced adjustments to the style and content of their own speech. For example, from the tone of a conversation partner's voice, one can typically gauge the mood of the person and accordingly adjust one's own speech by choosing appropriate words and tone that fit the partner's likely conversational expectations. User interaction with assistive devices and applications often includes use of speech-based interfaces where the contents of the user's speech serve as the input. However, aspects of the user's speech other than the natural language content are currently not taken into account when the speech is processed and are not used to determine the conversational speech response of the assistive device or application. As a result, the voice output used to interact with the user of the assistive device or application lacks some of the conversational nuance of real-life human conversations.

DESCRIPTION

With user permission, the techniques of this disclosure build and utilize a machine learning model to infer a person's mood based on an analysis of their speech. Training of the machine learning model is based on training data that is obtained specifically for this purpose, and with consent, from users that utilize the assistive application or device. Such training data includes the user's speech collected as input, and serves as unlabeled data.

Subsequently, the training data is labeled, by human coders who listen to the audio and annotate the data to indicate the speaker's mood. The annotations can include categorical labels that indicate common moods, such as "sad," "happy," "serious," "funny," etc. or numeric scale indicators that denote the intensity of the mood, such as "0.0" for extremely sad to "1.0" for "extremely cheerful." The annotated speech serves as ground truth training data for a machine learning model suitable for speech data, such as a multi-layer recurrent neural network based on Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). In addition to the annotations, the training input includes the raw audio of the speech, audio processed via feature extraction or pre-trained convolutional or recurrent networks, and additional attributes inferred with permission from the speaker of the audio, such as time of day, country, user characteristics, etc.

Once the model is sufficiently trained using the training data (e.g., stored as a database of annotated ground truth audio), audio of new speech samples and associated metadata, obtained with user consent, serve as input to the model. Based on analyzing the input, the model provides as output the inferred mood of the speaker, e.g., in the form of a categorical or numeric label along with an indication of the confidence in the correctness of the prediction. The model can be utilized by an assistive device or application that incorporates speech-based user interaction by passing it the user's speech input.

The interaction of assistive device or application with the user is adjusted appropriately based on the user's mood inferred by the model. Such adjustments can involve one or more of a variety of customizations to the interaction of the assistive device or application with the user. For instance, the tone of the voice output of the voice-based interface is altered to match the mood of the user. In another example, the content provided in response to user commands or queries is customized based on the user's mood. For example, in response to a user command to "play music," the playlist determined by the assistive device or application may include songs that fit the user's inferred mood. Similarly, the inferred mood may be utilized for mood-relevant content recommendations. For example, a query such as "I'm hungry" may trigger different responses based on mood, e.g., if the mood is detected as happy or relaxed, restaurant recommendations for eating out may be provided, e.g., “Restaurant A recently opened and has great reviews; Restaurant B and C near you are highly rated, would you like to book a table?” while if the mood is detected as upset or hurried, a more specific recommendation such as “would you like me to order a pizza from Pizza Man?” that saves time and navigating through further menus may be provided. Additionally, the content of the output of the assistive device or the application may be determined based on the inferred mood of the user. For example, users detected to be in a rush may be provided broad, high-level answers while other users detected to be curious of interested may be provided in-depth information in response to similar questions.

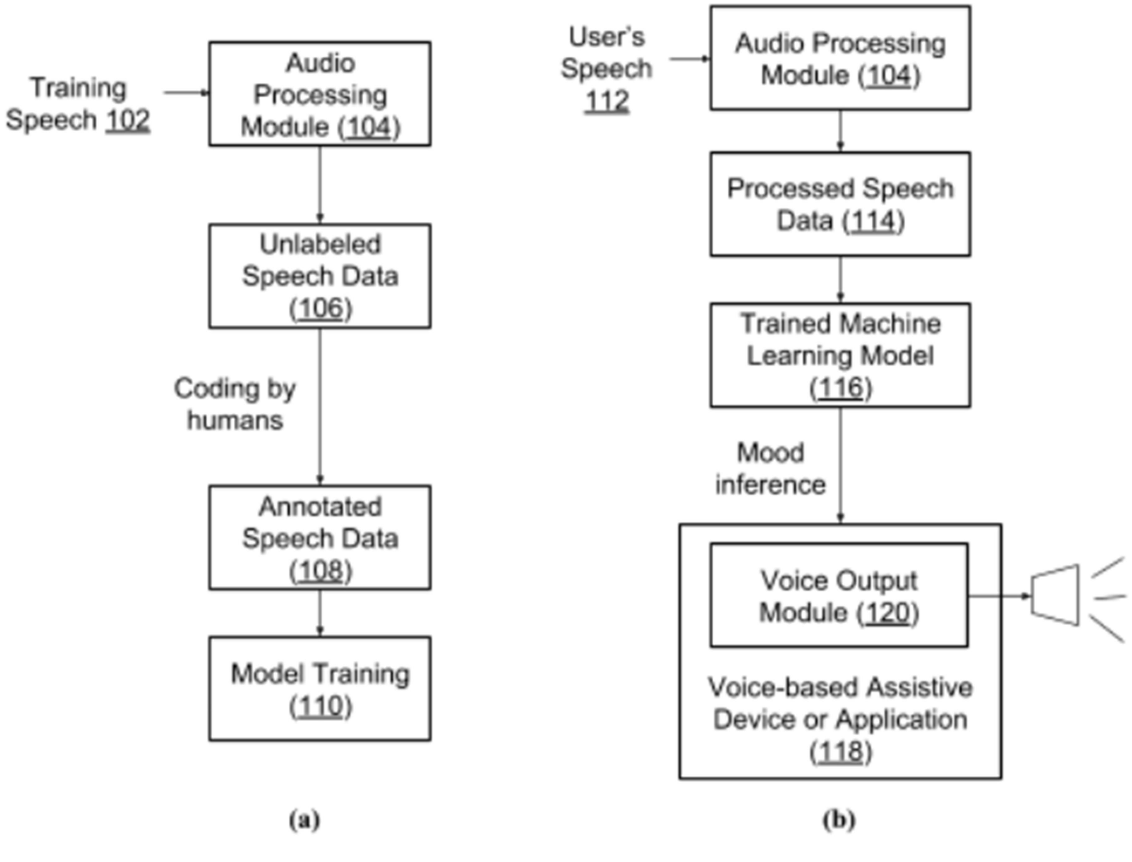
Training Speech 102
Audio Processing Module (104)
Unlabeled Speech Data (106)
Coding by humans
Annotated Speech Data (108)
Model Training (110)
(a)
User's Speech 112
Audio Processing Module (104)
Processed Speech Data (114)
Trained Machine Learning Model (116)
Mood inference
Voice Output Module (120)
Voice-based Assistive Device or Application (118)
(b)

Fig. 1: (a) Model training for mood inference; and (b) Using mood inference to adjust voice output of an assistive application

Fig. 1(a) shows training of a machine learning model for mood inference. Training speech (102) is obtained from users that consent to use of such speech data for training a mood inference model. An audio processing module (104) processes the training speech to obtain unlabeled speech data (106). With the user's permission, the audio processing module is used to apply feature extraction or pre-trained convolutional or recurrent networks to determine various audio and speaker attributes, such as time of day, country, user characteristics, etc.

Human coders provide annotations for the unlabeled speech data, as explained above, which provides annotated speech data (108). For example, human coders listen to training speech and provide categorical or numeric labels that correspond to the moods expressed in the speech and their intensities, respectively. A machine learning model is trained, e.g., by providing as input, unlabeled speech data (106) (or training speech 102) and the annotated speech data (108) which serves as ground truth, such that the model learns to predict the labels. Output of the model training phase is a trained machine learning model (116), as illustrated in Fig. 1(b)

Fig. 1(b) illustrates use of mood inference from a trained model to adjust voice output of an assistive application or device. If the user permits, the user's speech interaction (112) with the assistive device or application (118) along with the metadata associated with the speech is provided to audio processing module (104) to obtain processed data (114). The processed data and metadata are provided as input to the trained machine learning model (116) that provides a mood inference for the speaking user, e.g., in categorical or numeric form along with an indication of the confidence in the correctness of the prediction.

The mood inference and corresponding confidence values are received by a voice output module (120) that compares the confidence values to corresponding threshold values. If the confidence value does not meet a threshold value, the mood inference is considered insufficiently reliable to be utilized. In this instance, the voice output may be a default response that is not based on the mood inference. If the confidence value meets the threshold value, the voice output module (120) generates the speech response to be delivered to the user by the assistive device or application by taking into account the mood inference. The generated speech response is derived by making appropriate style and tone adjustments based on the mood inference to the default response. The generated speech output may also include other content, such as recommendations or queries, appropriate for the inferred mood.

Applying the inferred mood information to tailor the content and style of the speech output enables a voice-based assistive device or application to achieve user interaction that more closely resembles real-world human conversations, thus increasing its utility and appeal to the users. The threshold values for the confidence levels used to determine whether to use the inferred mood information may vary, e.g., based on the type of mood or the strength of the mood. Further, the threshold values may be set by the voice output module or may be configurable by the developer or the user.

The described techniques can be incorporated in a voice-based assistive application or device, and can also be provided as a web service or an Application Programming Interface (API) that enables developers to share and embed the features across applications and websites while providing fast performance and user experience comparable to native applications. The mood inference capabilities are turned off for users who do not provide permission to use their speech input for these purposes. In such cases, the user will receive the default voice output that would have been delivered in the absence of mood inference information.

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

In real-life human conversations, people often use additional information beyond the speech content, such as tone, pace, word choices, etc., to infer attributes of a speaker's emotional state and intention, and make nuanced adjustments to the style and contents of their own speech. With user permission, the techniques of this disclosure build and utilize a machine learning model to infer mood based on an analysis of user speech. A trained model is used to analyze speech input and provides a mood inference along with an indication of the confidence in the prediction. The inferred mood information is selectively applied to tailor the content or style of the speech output of a voice-based assistive device or application.