

VOICE INTERACTION SYSTEM AND METHOD

Technical Field

5 The present disclosure relates generally to voice interaction with a user, and in particular to systems and methods for providing an emotionally contextual response during voice interaction.

Background

10 With the development of artificial intelligence technology, intelligent voice interactive products are becoming widely used and regularly encountered by people in various products and processes. For example, exemplary voice interactive products may include outbound call products (such as customer service products), robot interaction products (such as companionship products, healing products), real-time digital human interaction products, and more.

15 Intelligent voice interactive products can recognize and process voice inputs from users, and output corresponding response voices based on the recognition and processing results, enabling voice interaction between humans and machines.

For example, a customer support system may be able to offer customer service to customers through telecommunication networks, including conducting multiple rounds of dialogue to
20 interact with customers. An intelligent voice interactive product forming part of this system may recognise and analyse a user's speech, process the content of the speech, formulate a suitable reply as a response text, and provide the response text to the customer as a reply. This may be accomplished rapidly and through multiple rounds of dialogue, allowing a seamless computer-human interaction.

25 However, typical intelligent voice interactive products, such as those providing speech recognition and formulated response text, will mechanically play the formulated response text as feedback to the customers voice data with a fixed tone, and without emotion. Because intelligent voice interactive products can only mechanically output response voices with fixed tones and lacking emotional context, this type of voice interaction is relatively rigid, and cannot
30 generate emotional interaction with users. This may result in a poor voice interaction experience for users.

Therefore, in order to address or alleviate at least one of the aforementioned problems and/or disadvantages, there is a need to provide an improved voice interaction system and method, particularly with an improved emotion recognition function.

5 Summary

In accordance with a first aspect of the present disclosure, a method for voice interaction with a user is provided. The method comprises acquiring user voice data forming part of a voice interaction, generating a response text for feedback to the user based on the acquired user voice data, performing emotion recognition on the user voice data to determine one or more
10 acoustic emotion parameters, and one or more semantic emotion parameters of the user voice data, combining the one or more acoustic emotion parameters and one or more semantic emotion parameters to determine a user's emotion state, determining a target emotion and a target tone from the user's emotion state for delivery of the generated response text, and providing the response text as an audio output using the target emotion and the target tone.

15 In an embodiment, the method steps are carried out sequentially in real time.

In an embodiment, the method further comprises converting the user voice data into corresponding user text via a speech recognition model for converting speech into text.

In an embodiment, the step of generating a response text comprises inputting the user text into a large language model for generating response texts.

20 In an embodiment, the step of generating a response text further comprises inputting the user text and one or more user attributes into a large language model for generating response texts.

In an embodiment, the step of generating a response text further comprises generating an initial response text based on speech-to-text results and/or one or more user attributes, analysing the acoustic and/or semantic emotional parameters from the user voice data, and
25 adjusting the initial response text based on at least a content and/or a tone of the determined acoustic emotion and/or semantic emotion of the user.

In an embodiment, the user attributes comprise one or more of: user intent, determined scenario, scenario context, habitual language patterns, personal attributes of the user, historical interaction data, and/or scenario attributes. Example embodiments of the user or
30 scenario attributes utilized by the method and system aspects as described herein may include:

User Personal Information Attributes, such as:

- Age: Users of different age groups may have different preferences and acceptance levels for language and emotionally contextual response as described herein.

- Gender: Gender differences may lead to different acceptance levels for emotional expression and emotionally contextual response as described herein.

- 5 - Cultural Background: Users from different cultural backgrounds may have different understandings and acceptance levels for emotional expression and emotionally contextual response as described herein.

User Historical Interaction Data or Attributes:

- 10 - Historical Emotional Records: For example, by analyzing a user's past emotional expressions, their current emotional state may be more accurately predicted.

- Interaction Habits: For example, a user's interaction habits, such as whether they prefer concise and clear responses, may also be factors considered when carrying out the methods and systems as described herein.

Scenario Attributes, such as:

- 15 - Interaction Scenario: Different interaction scenarios (such as work environments, leisure environments) may require different tones and emotions.

- Business Type: The type of business (such as finance, education, entertainment) associated with the user and/or voice interaction can influence the choice of tone and emotion in responses.

- 20 - Device Type: The type of device used by the user during or to access the voice interaction (such as a smartphone, tablet, smart speaker) may affect the presentation of the response.

In an embodiment, the determined one or more acoustic emotion parameters and one or more semantic emotion parameters correspond to an acoustic emotion state and semantic emotion state of the user.

- 25 In an embodiment, the determined target emotion and target tone correspond to the determined user emotion state.

In an embodiment, the step of performing emotion recognition comprises the use of a machine learning model trained to recognize acoustic features associated with emotional states from the user voice data, and/or wherein the step of performing emotion recognition comprises the

use of a machine learning model trained to recognize semantic features associated with emotional states from the user voice data.

5 In an embodiment, the step of determining the target emotion and target tone further comprises the identification and/or selection of an emotional feedback style that aligns with the user's determined true emotion (thereby providing an empathetic response).

In an embodiment, the target tone is determined based on at least one of a speech tempo, speech rate, pitch variation, speech timbre, speech rhythm, speech pauses, and/or speech volume of the user voice data and/or semantic structure, contextual analysis, topic modelling, and/or presence of words in the user voice data associated with a user's emotional state.

10 In an embodiment, the target emotion is determined based on a pre-defined emotional mapping that correlates the combined acoustic emotion and semantic emotion to one or more emotional states.

15 In an embodiment, the response text is generated by a natural language processing (NLP) algorithm, optionally wherein the NLP adapts to the recognized emotions in the user voice data and the determined target emotion and target tone.

In accordance with a second aspect of the present disclosure, a system for voice interaction is provided. The system comprises a voice data acquisition module configured to acquire user voice data from a user during a voice interaction, a response text generation module configured to generate a response text for feedback to the user based on the acquired user
20 voice data, an emotion recognition module configured to determine one or more acoustic emotion parameters and one or more semantic emotion parameters from the user voice data, an emotion combination module configured to combine the one or more acoustic emotion parameters and the one or more semantic emotion parameters to determine a user's emotion state, a target determination module configured to determine a target emotion and a target
25 tone for delivering the generated response text to the user based on the user's emotion state, and a response module configured to deliver the response text using the target emotion and target tone determined by the target determination module.

In an embodiment, the voice data acquisition module comprises a speech recognition module configured to convert the user voice data into corresponding user text.

30 In an embodiment, the response text generation module comprises a large language model configured to receive the acquired user voice data and generate the response text.

In an embodiment, the response text generation module is further configured to receive one or more user attributes for generating the response text.

In an embodiment, the response text generation module is further configured to generate an initial response text based on speech-to-text results and/or one or more user attributes, analyse the acoustic and/or semantic emotional parameters from the user voice data, and adjust the initial response text based on at least a content and/or a tone of the determined acoustic emotion and/or semantic emotion of the user.

In an embodiment, the user attributes comprise one or more of: user intent, determined scenario, scenario context, habitual language patterns, personal attributes of the user, historical interaction data, and/or scenario attributes. Example embodiments of the user or scenario attributes utilized by the method and system aspects as described herein may include:

User Personal Information Attributes, such as:

- Age: Users of different age groups may have different preferences and acceptance levels for language and emotionally contextual response as described herein.

- Gender: Gender differences may lead to different acceptance levels for emotional expression and emotionally contextual response as described herein.

- Cultural Background: Users from different cultural backgrounds may have different understandings and acceptance levels for emotional expression and emotionally contextual response as described herein.

User Historical Interaction Data or Attributes:

- Historical Emotional Records: For example, by analyzing a user's past emotional expressions, their current emotional state may be more accurately predicted.

- Interaction Habits: For example, a user's interaction habits, such as whether they prefer concise and clear responses, may also be factors considered when carrying out the methods and systems as described herein.

Scenario Attributes, such as:

- Interaction Scenario: Different interaction scenarios (such as work environments, leisure environments) may require different tones and emotions.

- Business Type: The type of business (such as finance, education, entertainment) associated with the user and/or voice interaction can influence the choice of tone and emotion in responses.

5 - Device Type: The type of device used by the user during or to access the voice interaction (such as a smartphone, tablet, smart speaker) may affect the presentation of the response.

In an embodiment, the emotion recognition module comprises a machine learning model configured to detect emotional features of the user voice data, optionally wherein the emotional features include acoustic emotion parameters and/or semantic emotion parameters.

10 In an embodiment, the response module includes an audio synthesis module configured to modulate an audio speech output based on the target emotion and tone.

A voice interaction system and method according to the present disclosure are thus disclosed herein. Various features, aspects, and advantages of the present disclosure will become more apparent from the following detailed description of the embodiments of the present disclosure, by way of non-limiting examples only, along with the accompanying drawings.

15

Brief Description of the Drawings

Fig. 1 is a flowchart illustrating a method of voice interaction according to embodiments of the present disclosure.

20 Fig. 2 is a flowchart illustrating a method of generating a response text according to embodiments of the present disclosure.

Fig. 3 is a flowchart illustrating a method of performing emotion recognition according to embodiments of the present disclosure.

Fig. 4 is a flowchart illustrating a process flow for determining an emotion state of a user according to embodiments of the present disclosure.

25 Fig. 5 is a flowchart illustrating a process flow for determining a target emotion and target tone according to embodiments of the present disclosure.

Fig. 6 is a flowchart illustrating a process flow for providing a response text according to embodiments of the present disclosure.

30 Fig. 7 is a schematic diagram illustrating an example of interactive relationships between modules of the systems and methods according to embodiments of the present disclosure.

Fig. 8 is a block diagram illustrating a system architecture of a voice interaction system according to embodiments of the present disclosure.

Fig. 9 is a block diagram illustrating a technical architecture of a computing server according to embodiments of the present disclosure.

5

Detailed Description

For purposes of brevity and clarity, descriptions of embodiments of the present disclosure are directed to a in accordance with the drawings. While aspects of the present disclosure will be described in conjunction with the embodiments provided herein, it will be understood that they are not intended to limit the present disclosure to these embodiments. On the contrary, the present disclosure is intended to cover alternatives, modifications and equivalents to the embodiments described herein, which are included within the scope of the present disclosure as defined by the appended claims. Furthermore, in the following detailed description, specific details are set forth in order to provide a thorough understanding of the present disclosure. However, it will be recognized by an individual having ordinary skill in the art, i.e. a skilled person, that the present disclosure may be practiced without specific details, and/or with multiple details arising from combinations of aspects of particular embodiments. In a number of instances, well-known systems, methods, procedures, and components have not been described in detail so as to not unnecessarily obscure aspects of the embodiments of the present disclosure.

In embodiments of the present disclosure, depiction of a given element or consideration or use of a particular element number in a particular figure or a reference thereto in corresponding descriptive material can encompass the same, an equivalent, or an analogous element or element number identified in another figure or descriptive material associated therewith.

References to “an embodiment / example”, “another embodiment / example”, “some embodiments / examples”, “some other embodiments / examples”, and so on, indicate that the embodiment(s) / example(s) so described may include a particular feature, structure, characteristic, property, element, or limitation, but that not every embodiment / example necessarily includes that particular feature, structure, characteristic, property, element or limitation. Furthermore, repeated use of the phrase “in an embodiment / example” or “in another embodiment / example” does not necessarily refer to the same embodiment / example.

The terms “comprising”, “including”, “having”, and the like do not exclude the presence of other features / elements / steps than those listed in an embodiment. Recitation of certain features / elements / steps in mutually different embodiments does not indicate that a combination of these features / elements / steps cannot be used in an embodiment.

5 As used herein, the terms “a” and “an” are defined as one or more than one. The use of “/” in a figure or associated text is understood to mean “and/or” unless otherwise indicated. The term “set” is defined as a non-empty finite organisation of elements that mathematically exhibits a cardinality of at least one (e.g. a set as defined herein can correspond to a unit, singlet, or single-element set, or a multiple-element set), in accordance with known
10 mathematical definitions. The terms “first”, “second”, etc. are used merely as labels or identifiers and are not intended to impose numerical requirements on their associated terms.

As used in the present disclosure, the terms “component”, “module”, “system”, “interface” and the like are generally intended to refer to a computer-related entity, either hardware, a combination of hardware and software, software, or software in execution. For example, a
15 module may be, but is not limited to being, a process running on a processor, a processor, an object, an executable, a thread of execution, a program, and/or a computer. By way of illustration, both an application running on a controller and the controller can be a component. One or more modules may reside within a process and/or thread of execution and a module may be localized on one computer and/or distributed between two or more computers. As
20 another example, an interface can include I/O components as well as associated processor, application, and/or API components. In the context of the present disclosure, the information processing and response generation system and its constituent parts may be implemented as hardware, software, or a combination thereof.

Furthermore, various embodiments of the present disclosure may be implemented as a
25 method, apparatus, or article of manufacture using standard programming and/or engineering techniques to produce software, firmware, hardware, or any combination thereof to control a computer to implement the disclosed subject matter. For instance, the claimed subject matter may be implemented as a computer-readable medium embedded with a computer executable program, which encompasses a computer program accessible from any computer-readable
30 storage device or storage media. For example, computer readable media can include but are not limited to magnetic storage devices (e.g., hard disk, floppy disk, magnetic strips, etc.), optical disks (e.g., compact disk (CD), digital versatile disk (DVD), etc.), smart cards, and flash memory devices (e.g., card, stick, key drive, etc.).

It will be understood that acoustic emotion features and semantic emotion features are
35 identified as two important dimensions for recognizing user emotion in voice interactions. In

example embodiments, these terms may refer to specific parameters and/or features that may allow more accurate recognition of acoustic and semantic emotion.

As used herein, the term acoustic emotion relates to the acoustic features of a user's speech, such as tone, volume, rhythm, speech tempo, pitch variation, and/or speech volume, which may reflect a user's emotional state.

The systems and methods as described herein may be configured to detect one or more Acoustic Emotion or Acoustic Emotional Features or Parameters. Acoustic emotion features or parameters may refer to one or more of the following:

- Pitch: Variations in pitch can reflect a user's emotional state, with high pitch possibly indicating excitement or tension, and low pitch possibly indicating calmness or depression.

- Volume: Changes in volume may be a way of detecting expression of emotion, with loudness possibly indicating excitement or anger, and softness possibly indicating shyness or unease.

- Speech Rate: The speed of speech may convey emotional information, with fast speech possibly indicating excitement or tension, and slow speech possibly indicating fatigue or contemplation.

- Timbre: Differences in timbre, such as hoarseness or clarity, may reflect a user's emotional state.

- Pauses and Rhythm: Changes in pauses and rhythm in speech may also be important components of emotional expression, with long pauses possibly indicating hesitation or thought.

As used herein, the term semantic emotion relates to the semantic features of a user's speech content or context, such as emotionally charged vocabulary, which may reflect a user's emotional state.

The systems and methods as described herein may be configured to detect one or more Semantic Emotion or Semantic Emotional Features or Parameters. Semantic emotion features or parameters may refer to one or more of the following:

- Emotional Words: The emotional words used in text may directly reflect a user's emotional state, such as "happy," "sad," "angry," etc.

- Contextual Analysis: In addition to individual words, a context and surrounding text of an entire sentence may be considered to accurately understand a user's emotion.

- Semantic Structure: The semantic structure of sentences, such as questions, statements, commands, etc., may also indirectly reflect a user's emotion.

- Topic Modelling: By analyzing one or more topics of text in a voice interaction using topic modelling, a user's emotional tendencies and areas of interest can be further understood.

5 Emotion in computer-human interaction is a widely researched field of technology, where it has been identified that computer-human interaction or conversation with contextually matched emotion can provide an enhanced user experience. For example, if after obtaining a user's voice data during a conversation, the user's real-time true emotions are recognized, and then a response is played back to the user through an emotion and tone that matches the
10 user's true emotions, real-time emotional interaction with the user can be achieved during a human-machine interaction or conversation. This allows users to perceive emotional feedback, enhancing their voice interaction experience.

This disclosure provides a technical solution for human-machine voice interaction with a user, and for providing an emotionally contextual response. Embodiments as disclosed herein may
15 comprise systems and methods for voice interaction with a user, including providing an emotionally contextual response during a voice interaction. Systems and methods as disclosed herein may comprise emotionally responsive, emotionally adaptive, and/or context-aware emotional response methods and/or systems.

The embodiments described herein may comprise various steps to provide for emotionally
20 contextual responses during voice interaction. For example, the systems and methods as disclosed herein may comprise the steps of real-time acquisition of a user's voice data during a current conversation process to generate a response text for feedback to the voice data. Emotion recognition performed on the voice data to determine the user's real-time acoustic emotion and semantic emotion. Determination of the user's real-time true emotion by
25 combining the acoustic emotion and semantic emotion. Determination of a target emotion and target tone suitable for feedback based on the true emotion of the user, and providing the response text to the user using the determined target emotion and target tone.

Advantageously, the solution provided by one or more embodiments of this disclosure may not mechanically play the response text for feedback to the voice data with a fixed tone and
30 without emotion. Instead, after real-time acquisition of the user's voice data during a conversation process, a user's real-time true emotion may be identified through the user's real-time acoustic emotion and semantic emotion as described herein. Following this, a response text for feedback to the user's voice data may be played or conveyed to the user with an emotion and tone suitable for feedback of the true emotion. In this way, real-time
35 emotional interaction with the user can be achieved when playing a response text, allowing

the user to perceive emotional feedback and thereby enhancing the user's voice interaction experience.

Whilst example embodiments have been described herein, it will be understood that the order of response text generation is not necessarily always before the emotion recognition step. In alternative embodiments, the response text can be generated based on the user's voice data, scenario information, user attributes, and/or recognized user intent.

The voice interaction method as described herein may involve real-time acquisition of a user's voice data during a current conversation, and generation of a response text for feedback to the voice data. Next, emotion recognition may be performed on the voice data to determine a user's real-time acoustic emotion and semantic emotion, and the user's real-time or genuine emotion may be determined by combining the acoustic emotion and semantic emotion components. Finally, a target emotion and target voice suitable for feedback of the genuine emotion may be determined, and the response text can be played to the user using the target emotion and target voice.

As outlined above, the solution provided in various embodiment of the present application do not involve mechanical playback of the response text for feedback to the voice data with a fixed voice and lack of emotion. Instead, after real-time acquisition of the user's voice data during the conversation, the user's real-time genuine emotion may be accurately identified through their real-time acoustic emotion and semantic emotion. Then, a response text for feedback to the voice data may be played to the user through emotions and voices or tones suited to feedback of the user's genuine emotion. In this way, real-time emotional interaction with the user can be achieved when playing or conveying the response text, allowing the user to perceive emotional feedback, thereby enhancing their voice interaction experience.

The technical solution for voice interaction as described herein can be applied to any type of voice interactive product to enable the product to achieve real-time emotional interaction with users, and enhance a user's voice interaction experience. The embodiments described herein do not limit the type of voice interactive product, which can be flexibly selected based on specific situational needs. Exemplary voice interactive products may include, but are not limited to, outbound call products (such as customer service products), robot interaction products (such as companionship products, healing products), real-time digital human interaction products, and more.

Fig. 1 is a flow chart illustrating an example embodiment of a method of voice interaction as disclosed herein. The method 100 comprises acquiring, at step 110, user voice data

forming part of a voice interaction. In an example embodiment, the user voice data may comprise a user's voice data acquired during or as part of a current conversation process of a voice interaction between a user and a computer system.

The method 100 further comprises generating, at step 120, a response text based on the acquired user voice data for feedback to the user. In an example embodiment, the response text may comprise a suitable or relevant response or set of responses to the user voice data. For example, response text that may comprise a suitable response to a request identified or forming part of the user voice data.

Following this, the method 100 further comprises performing, at step 130, emotion recognition on the user voice data. In an example embodiment, the emotion recognition may comprise determining at step 132 an acoustic emotion or set of emotion. In an example embodiment, the emotion recognition may additionally or alternatively comprise determining at step 134 a semantic emotion or set of emotions. In an example embodiment, the determined one or more acoustic emotion parameters and/or semantic emotion parameters may correspond to an acoustic emotion state and/or semantic emotion state of a user.

Next, the method 100 further comprises combining, at step 140, the one or more acoustic emotion parameters and semantic emotion parameters to determine a user's emotion state. The combination of the acoustic emotion parameters determined at step 132, with the semantic emotion parameters determined at step 134, may allow the method 100 to determine or calculate an emotion state or true emotion of a user.

Next, the method 100 further comprises determining, at step 150, a target emotion and/or a target tone for delivery of the generated response text. In an example embodiment, the determined target emotion and target tone may correspond to the determined user emotion state from step 140. For example, the determined target emotion and target tone may match a determined emotion state or true emotion of the user.

Finally, the method 100 comprises providing, at step 160, the response text determined at step 120 as an audio output using the target emotion and target tone as determined at step 150. In an example embodiment, the step 160 of providing the response text may comprise playing the response text as audio output in reply to the user as part of the voice interaction, where the providing a reply may comprise broadcasting the reply text as audio with a tone and emotion corresponding to an acoustic and semantic emotion in response to the determined user emotion state at step 140, and/or via a tone and emotion determined at step 150.

As outlined herein, in an example embodiment, the method 100 of Fig. 1 may comprise step 110, where the method may acquire or obtain user voice data. In an example embodiment,

the acquired voice data may comprise real-time voice data of the user during the current conversation.

Further, as outlined above in an example embodiment the method 100 may comprise step 120, of generating a response text for feedback to the user voice data.

- 5 In an example embodiment, the timeliness of voice interaction may be an important performance aspects of voice interaction products, such as those described herein. Therefore, in order to improve the timeliness of voice interaction, an example embodiment may obtain real-time voice data of the user during a current conversation with the user, so as to provide timely responses to the voice data.
- 10 In an example embodiment as described herein, after obtaining at step 110 the real-time voice data of the user during a current conversation, the system may generate at step 120 a response text suitable to act as feedback or reply to a content of the user voice data, in order to respond to the real-time obtained voice data of the user.

15 In some example embodiments, methods for generating a response text for feedback to the user voice data can include converting the acquired user voice data into corresponding user text based on a speech recognition model used for converting speech into text.

The speech recognition model may be trained based on multiple sets of data and may be used for converting voice data into corresponding text. Each set of data among multiple sets can include voice data and corresponding text. It will be understood that various types of speech
20 recognition or text to speech model could be utilised in the embodiments disclosed herein, and the specific type of speech recognition model can be flexibly selected based on particular requirements of the systems and methods for voice interaction as disclosed herein. Exemplary speech recognition models can include, but are not limited to a multi-label speech large model, an Automatic Speech Recognition (ASR) model, where the multi-label speech large model
25 can be a multi-label speech large model after Supervised Fine-Tuning (SFT) in a vertical domain (for example related to the scenario corresponding to the current conversation process of the voice interaction).

After selecting the suitable speech recognition model, the voice data may be converted into corresponding user text based on the speech recognition model. The user text may be text
30 that expresses the semantics of the user voice data.

Next, the method 100 may comprise generating at step 120 a response text to act as feedback or a reply to the voice data based on the user text. The generating may utilise a large language model for generating response texts. The response texts may be target texts.

The large language model may be trained based on multiple sets of data and may be used for generating response texts. Each set of data among multiple sets can include user text and a corresponding response text. It will be understood that various types of large language model could be utilised in the embodiments disclosed herein, and the specific type of large language model can be flexibly selected based on particular requirements of the systems and methods for voice interaction as disclosed herein. Exemplary large language models can include, but are not limited to, the Llama 3.1 Large Language Model (LLM). In some embodiments, both the speech recognition model utilised in step 110 and the large language model utilised in step 120 can be existing products that are pre-trained, which can save model training time and costs, therefore reducing the cost of voice interaction.

After determining the large language model, a response text or target text for feedback or reply to the voice data may be generated based on the user text using the large language model. The response text may be specifically used to respond to the user voice data. Since the response text may be generated using only the large language model to process the user text, it may enable the large language model to quickly generate the response text, which may advantageously enable timely response during voice interaction. Put another way, in an example embodiment the systems and methods disclosed herein may comprise a dedicated large language model to process the user text as described herein to generate the response text.

Alternatively, or in addition, the embodiments as disclosed herein may take into consideration that the voice data of users can be related to a particular scenario to which the voice interaction is occurring, and/or user attributes. Therefore, in order to improve the adaptation of the response text to the scenario and/or user attributes, the voice interaction methods and systems described herein may further include additional steps before generating a response text for feedback or reply to the voice data. For example, embodiments may comprise determining a scenario related to the current conversation process occurring as part of the voice interaction, and/or determine user attributes related to the user and/or scenario.

It will be understood that, when generating response text or determining the target tone or emotion as described herein, example embodiments may consider and incorporate additional user or scenario attributes to ensure the accuracy and appropriateness of the response.

In example embodiments, the user attributes may comprise one or more of a user intent, a determined scenario, scenario context, and/or habitual language patterns. Example embodiments of the user or scenario attributes utilized by the methods and systems as described herein may include:

User Personal Information Attributes, such as:

- Age: Users of different age groups may have different preferences and acceptance levels for language and emotionally contextual response as described herein.

- Gender: Gender differences may lead to different acceptance levels for emotional expression and emotionally contextual response as described herein.

- 5 - Cultural Background: Users from different cultural backgrounds may have different understandings and acceptance levels for emotional expression and emotionally contextual response as described herein.

User Historical Interaction Data or Attributes:

- 10 - Historical Emotional Records: For example, by analyzing a user's past emotional expressions, their current emotional state may be more accurately predicted.

- Interaction Habits: For example, a user's interaction habits, such as whether they prefer concise and clear responses, may also be factors considered when carrying out the methods and systems as described herein.

Scenario Attributes, such as:

- 15 - Interaction Scenario: Different interaction scenarios (such as work environments, leisure environments) may require different tones and emotions.

- Business Type: The type of business (such as finance, education, entertainment) associated with the user and/or voice interaction can influence the choice of tone and emotion in responses.

- 20 - Device Type: The type of device used by the user during or to access the voice interaction (such as a smartphone, tablet, smart speaker) may affect the presentation of the response.

25 It will be understood that scenario information and/or user attributes, such as those set out above, can be utilised by the voice interaction systems and methods as described herein, for example to generate, adjust, or modify any of a response text or response text generation process, emotion recognition process, emotion parameter combination process, target emotion or tone determination process and/or delivery of a response to a user.

30 In an example embodiment, the conversation process with the user may be initiated for a specific scenario or purpose of the voice interaction. Additionally, or alternatively, the voice data of the user during a conversation process may also be related to the scenario or purpose of the voice interaction. Therefore, in order to accurately respond to the voice data of the user,

the embodiments disclosed herein may determine a scenario related to the current conversation process, so that the scenario can be used as a reference factor for determining the response text. This may allow improvement of the accuracy of the response text to the voice data. For example, if the current conversation process is initiated for credit card repayment reminders, then the scenario related to the current conversation process is determined to be credit card repayment reminders.

Alternatively, or in addition, in an example embodiment the response to the user voice data may be related to the user's attributes associated with the scenario. Therefore, in order to accurately respond to the voice data of the user, some embodiments may determine a user's attributes related to the scenario, such that these attributes can be used as a reference factor for determining the response text. This may allow improvement of the accuracy of the response text to the voice data. For example, if the scenario is credit card repayment reminders, the user's attributes related to the scenario can include, but are not limited to: credit card limit, monthly salary of the user, balance of the user's related savings account, etc.

Accordingly, in an example embodiment the process, such as method step 120 in Fig. 1, for generating a response text for feedback or reply to the voice data can be based on the determination of the scenario related to the current conversation process, and the user's attributes related to the scenario. In example embodiments, this process may comprise converting the voice data into corresponding user text based on a speech recognition model used for converting speech into text. In an example embodiment, the speech recognition model may be trained based on multiple sets of data and used for converting voice data into corresponding text. In order to improve the accuracy of the user text, each set of data among the multiple sets may include voice data and corresponding text under the scenario related to the current conversation process, and/or user attributes related to the scenario as described above.

Alternatively, or in addition, in order to improve the accuracy of the user text, each set of data among the multiple sets may include user's habitual language patterns. Attributes related to the user's habitual language patterns can include, but are not limited to: gender, education level, age, occupation, personality, etc.

In this way, after converting the voice data into corresponding user text based on the speech recognition model used for converting speech into text, the obtained user text can more accurately express the semantics of the voice data.

Fig. 2 illustrates an example embodiment of a method 200 corresponding to the step of generating a response text as described herein, for example at step 120 of Fig. 1. The step of generating a response text may include generating an initial response text 220 based on speech-to-text results and/or one or more user attributes 222 as described herein. The user attributes may be predetermined or provided to the system based on a context or scenario to which the voice interaction is occurring. Alternatively, or in addition, the user attributes 222 may be derived from any of the steps of the systems and methods disclosed herein, including steps 130 to 160, and/or previous interactions with the user.

Further, the method 200 may include analysing the acoustic and/or semantic emotional parameters generated from the user voice data, as described at step 130, 132 and 134 in Fig. 1, allowing determination of acoustic emotion parameters 232 and semantic emotion parameters 234. Based on this analysis, the method 200 of generating a response text may further include refining or adjusting the initial response text at step 250 based on at least a content and/or a tone of the determined acoustic emotion and/or semantic emotion of the user, such as those acoustic and semantic emotion parameters determined at steps 232 and 234.

Alternatively, or in addition, the method 200 of generating a response text may include determining at step 240 a user's emotional state from the determined acoustic and semantic emotion parameters at steps 232 and 234, in line with step 140 as described in Fig. 1. Further, the method 200 may comprise analysis of the determined emotion state 240 to provide feedback and adjustment to the generated initial response text, thereby generating at step 250 an updated or revised response text that may better take into account an acoustic or semantic emotional state of a user for delivery of the response text.

In an example embodiment, such as that described in relation to step 120 of Fig. 1, or method 200 of Fig. 2, the systems and methods as described herein may generate a response text for feedback or reply to the voice data based on the user text, scenario, and/or user attributes using a large language model for generating response texts.

The large language model may be trained based on multiple sets of data and used for generating response texts. In order to improve the accuracy of the response text, each set of data among the multiple sets may include user text and corresponding response text under the scenario related to the current conversation process and the user's attributes related to the scenario. In this way, after generating a response text for feedback or reply to the voice data based on the user text, scenario, and user attributes using the large language model, the obtained response text may suitably or correctly respond to the voice data of the user, which may reduce a likelihood of response errors.

It will be understood that the above embodiments for generating a response text for feedback or reply to the voice data can be flexibly selected based on needs of the present scenario to which the voice interaction is related. Furthermore, it is noted that regardless of which method is used, the type of speech recognition model outlined above can be one or more, or a combination. When multiple speech recognition models are used, the process for the step of converting the voice data into corresponding user text based on a speech recognition model used for converting speech into text in the method for generating a response text for feedback or reply to the voice data can include the following steps:

- convert the voice data into corresponding text based on each speech recognition model;

- integrate the texts converted by each speech recognition model based on the corresponding model weights of each speech recognition model to obtain the user text,

- where the model weights indicate the credibility of the text converted by the speech recognition model for the scenario and user attributes.

Considering that each speech recognition model has a certain error probability when converting voice data into text, in order to ensure that the converted user text can express the true semantics of the voice data as accurately as possible, the embodiments as disclosed herein may utilise more than one speech recognition model.

Each speech recognition model may comprise different credibility for the text converted from the scenario related to the conversation process and the user's attributes related to the scenario. Higher credibility may mean that the text converted by the speech recognition model can more accurately express the semantics of the voice data. Therefore, the model weight of each speech recognition model can be set based on the credibility of the text converted by each speech recognition model for the scenario and user attributes.

In example embodiments, after converting the voice data into corresponding user text based on each speech recognition model, the texts converted by each speech recognition model may be integrated based on the corresponding model weights of each speech recognition model to obtain the user text, so that the user text can express the true semantics of the voice data. The process for integrating the texts converted by each speech recognition model can include:

- determine an associated text vector corresponding to the text converted by each speech recognition model,

- determine the product between the text vector and the model weight corresponding to each speech recognition model,

- sum the products corresponding to each speech recognition model to determine the user text vector, and

5 - determine the text corresponding to the user text vector as the user text.

Fig. 3 illustrates an example embodiment of a module 300 for performing emotion recognition as described herein, for example as part of step 130 in Fig. 1. Turning to Fig. 1, as described herein, the methods and systems disclosed perform emotion recognition on the user voice data. In an example embodiment, the emotion recognition may comprise determining at step 10 132 an acoustic emotion or set of emotion. In an example embodiment, the emotion recognition may additionally or alternatively comprise determining at step 134 a semantic emotion or set of emotions. In an example embodiment, the determined one or more acoustic emotion parameters and/or semantic emotion parameters may correspond to an acoustic 15 emotion state and/or semantic emotion state of a user.

A user's emotional state or true emotions may be expressed at both acoustic and semantic levels. Accordingly, emotion recognition may be carried out on the user voice data acquired at step 110, and/or the user text that may be generated as part of step 120 as outlined herein. In this way, the systems and methods disclosed herein may determine a user's real-time 20 acoustic and semantic emotions, in order to enable emotional interaction with the user and allow a user to perceive emotional feedback when providing a response text to the user in response to the user voice data. The following provides example embodiments of how the systems and methods as disclosed herein may perform emotion recognition on user speech, such as user voice data, to determine a user's real-time acoustic and semantic emotion.

25 In an example embodiment, the process of performing emotion recognition on user speech data to determine a user's real-time acoustic emotion, such as that disclosed at step 132 in Fig. 1, may include using a multi-class emotion model adapted to a user's acoustic emotion expression habits, and performing emotion recognition on the user's speech, such as user voice data, to obtain the user's real-time acoustic emotions.

30 In an example embodiment, the multi-class emotion model may be a model trained based on multiple sets of data for recognizing acoustic emotions. In order to accurately recognize a user's acoustic emotions, the multi-class emotion model may be adapted to the user's acoustic emotion expression habits. Therefore, each type of data in the multiple sets of data for the

multi-class emotion model may include a user's speech or user voice data and the corresponding acoustic emotions of the speech or users voice data.

Turning back to Fig. 3, in an example embodiment the process for performing emotion recognition 300 may comprise feature extraction 320, where acoustic features like pitch, volume, and speech rate may be extracted from the user speech or user voice data 310. Further, the process 300 may include acoustic model analysis 322, for example using a trained multi-class emotion model to determine a user's emotional state (e.g., happy, sad, angry) based on the extracted acoustic features. From this, the process 300 may comprise determining acoustic emotion parameters at 324, based on the acoustic feature extraction at 320, and acoustic model analysis at 322.

In an example embodiment, the process 300 of performing emotion recognition on speech data or user voice data 310 to determine a user's real-time semantic emotions may comprise converting the speech to text at 340, for example based on a speech recognition model used for converting speech into text. The resultant user text may then be analysed at 342 using a large language model adapted to the user's semantic emotion expression habits, to provide semantic model analysis at 342. From this, the process 300 may perform emotion recognition on the user text to obtain the user's real-time semantic emotions, by determining semantic emotion parameters at 344, based on the converted speech to user text at 340, and the semantic model analysis at 342.

In some example embodiments, text may be able to better express a user's semantic emotions. For example, when recognizing semantic emotions, it may be necessary to convert the speech data into corresponding user text 340 based on a speech recognition model used for converting speech into text, and then use a large language model adapted to the user's semantic emotion expression habits to perform emotion recognition 342 on the user text, thereby obtaining the user's real-time semantic emotions at 344.

In an example embodiment, the large language model may be a model trained based on multiple sets of data for recognizing semantic emotions. Considering that different users have different semantic expression habits, in order to accurately recognize a user's semantic emotions, the large language model may be a model adapted to the user's semantic emotion expression habits. Based on this, each type of data in the multiple sets of data for the large language model may include text corresponding to the user's speech data and the semantic emotions corresponding to the text.

The determined acoustic emotion parameters 324 and semantic emotion parameters 344 may be conveyed to the next step of determining a user's emotion state, corresponding to step 140 of Fig. 1.

After determining the user's real-time acoustic emotions and semantic emotions, as set out in Fig. 3, in order to more accurately determine the user's real-time true emotions, the acoustic emotions and semantic emotions may be combined to determine a user's real-time true emotions.

5

Fig. 4 illustrates an example embodiment of a process 400 to determine an emotion state of a user, corresponding to step 140 of Fig. 1. The process 400 of combining acoustic emotions and semantic emotions to determine the user's real-time true emotions may include the following steps: At steps 420 and 440, the process may receive acoustic emotion parameters and semantic emotion parameters, respectively. From this, at steps 422 and 442, the process may determine emotion weights corresponding to the acoustic emotions and semantic emotions, respectively, from the acoustic and semantic emotion parameters.

Finally, based on the emotion weights corresponding to acoustic emotions 422 and semantic emotions 442, the process 400 may perform a weighted calculation at 450 on the acoustic emotions and semantic emotions to combine the acoustic and semantic emotion parameters and weights to obtain the user's real-time true emotions.

In an example embodiment, the determined emotion state of a user at step 450 may comprise a numerical value, or may be a state or value selected from a list or set of quants.

Emotion weights, such as those determined at 422 and 442, may be used to indicate a degree of influence of corresponding acoustic and semantic emotions on a user's true emotions. For example, the emotion weight of acoustic emotions 422 may be used to indicate the degree of influence of acoustic emotions on the user's true emotions, and the emotion weight of semantic emotions 442 may be used to indicate the degree of influence of semantic emotions on the user's true emotions.

Generally, for a user, if their emotions are usually expressed through sound, then the emotion weight of their acoustic emotions 422 may higher than that of their semantic emotions 442. Conversely, for a user, if their emotions are usually expressed through semantics, then the emotion weight of their semantic emotions 442 may be higher than that of their acoustic emotions 422. That is to say, for any user, the magnitude of the respective emotion weights corresponding to acoustic emotions and semantic emotions may be related to the user's emotional expression habits.

In an example embodiment, the emotional weights determined at 422 and 442 may indicate the influence of acoustic and semantic emotional features on the user's true emotional

characteristics. The emotional weights determined at 422 and 442 may be determined based on:

- User Research: Understanding users' sensitivity to acoustic and semantic emotions through surveys, interviews, etc.

5 - Data Analysis: Using historical data to analyze which emotional expression method (acoustic or semantic) has a greater impact on user emotional feedback.

- Model Training: Adjusting weights during model training to optimize performance.

After determining the emotion weights corresponding to acoustic emotions 422 and semantic emotions 442, respectively, the process 400 may carry out step 450 to combine the respective
10 emotion weights to obtain a user's true emotion. This process may include determining at 452 a product of the feature vector corresponding to the acoustic emotions 420 and the emotion weight 422, determining at 454 a product of the feature vector corresponding to the semantic emotions 440 and the emotion weight 442, determining a sum of the two products at 460 as a target vector, and determining an emotion at 470 corresponding to the target vector as the
15 user's real-time true emotion or emotion state.

After determining the user's real-time true emotion or emotion state, as set out in Fig. 4 and corresponding to step 140 of Fig. 1, systems and methods as disclosed herein may determine a target emotion and/or target tone for responding to the user's voice data, as set out in step 150 of Fig. 1.

20

Fig. 5 illustrates an example embodiment of a process 500 for determining a target emotion and target tone that may be suitable for response based on the emotion state of the user as described herein. The determined target emotion and target tone or voice may be considered suitable for feedback of true emotions in response to a user's determined emotion state, such
25 as that determined in Fig. 4.

After determining the user's real-time true emotions or emotion state, in order to achieve emotional interaction with the user when playing back the response text, it may be necessary to determine a target emotion and target voice or tone suitable for feedback of true emotions, such that a response text (for example that determined at step 120 in Fig. 1, or in Fig. 2) can
30 be played back to the user based on the target emotion and target voice or tone, allowing the user to perceive emotional feedback.

The process 500 for determining the target emotion and target voice suitable for feedback of true emotions may include the following steps:

In a first example process embodiment, the process 500 for determining a target emotion and target tone suitable for feedback of true emotions may include the following steps:

The process 500 may use a first emotional voice model 510 to process the true emotions or emotion state 505, for example that determined in Fig. 4 at step 470, to obtain a target emotion 520 and target tone 530. The target emotion 520 and target tone 530 may be provided to the response delivery means as outlined at step 160 of Fig. 1, or may be combined into a step 570 comprising a determined target emotion 520 and target tone 530 as a combined data set to be provided to the response delivery means.

The first emotional voice model 510 may be a model trained based on multiple sets of first data to determine a target emotion and target tone suitable for feedback of true emotions or in response to the user's emotion state 505. Each set of first data among the multiple sets of first data may include true emotions or emotion states, as well as the emotions and voices or tones used to feedback the respective true emotions or emotion states. In this way, the model 510 may be trained to output a suitable emotion 520 and tone 530 based on learnings from the multiple sets of first data.

In an example embodiment, the first emotional voice model 510 may comprise a dedicated emotion voice model 510, and may primarily process the factor of true emotions or emotion state as outlined herein. Accordingly, the model 510 may have fewer factors to process, and may quickly obtain a target emotion and target voice suitable for feedback of true emotions, thus improving the efficiency of voice interaction.

Alternatively, or in addition, the process 500 for determining the target emotion and target voice suitable for feedback of true emotions may include the following steps:

In an example embodiment, the process 500 may include a module 540 to receive or determine a scenario, context, and/or user attributes related to the current dialogue process or voice interaction. This is described in relation to Fig. 2, and process step 120 in Fig. 1, where user attributes may be used to determine a response text. It will be understood that the scenario, context and/or attributes as described herein could be some or all of the same data received as part of earlier steps or processes of the disclosed systems and methods, and/or could be independently received and/or determined as part of the specific process of the method or system as required.

In an example embodiment, the process 500 may utilise a second voice model 550 adapted to the scenario, context and/or user attributes to process the true emotions or user emotion state 505 to obtain the target emotion 520 and target voice or tone 530, and/or assist or feed data into the emotion voice model 510 to determine the target emotion 520 and target tone

530. It will be understood that this process could be carried out entirely within or as part of emotion voice model 510, or entirely within second voice model 550, or a combination of both emotion voice model 510 and second voice model 550. For example, second voice model 550 may parse the scenario, context and/or user attributes and provide information relevant to determining the target emotion and tone to the emotion voice model 510. Alternatively, or in addition, the second voice model 550 could output the target emotion 520 and target tone 530 as a result of the scenario, context and/or user attributes.

As part of this example embodiment, the process 500 may consider a scenario and/or context associated with the voice interaction, and refine the identified target emotion 520 and/or identified target tone 530 to take into account this scenario or context information. Alternatively, or in addition, the process 500 may consider user attributes associated with a user taking part in the voice interaction, and refine the identified target emotion 520 and/or identified target tone 530 to take into account the user attributes.

The second voice model 550 may be a model trained based on multiple sets of second data to determine the target emotion 520 and target tone 530 suitable for feedback of true emotions or reply to the user emotion state determined previously. Each set of second data among the multiple sets of second data may include true emotions under specific scenario, context and attributes, as well as the emotions and tones used to feedback the true emotions, allowing the model to be trained.

Since the second voice model 550 may be trained based on second data that includes true emotions under specific scenarios, context and/or attributes, as well as the emotions and voices used to feedback the true emotions, the target emotion 520 and target tone 530 determined by this process may be more suitable for the voice interaction-related scenario or and context. Therefore, playing back the response text based on the target emotion 520 and target tone 530 may allow for improved emotional interaction with a user.

It will be understood that the process 500 for determining the target emotion 520 and target tone 530 suitable for feedback of true emotions can be flexibly selected based on requirements of the voice interaction.

Both of the above-mentioned models 510 and 550 for determining the target emotion 520 and target voice 530 suitable for feedback of true emotions may comprise emotion voice models. The training methods for emotion voice models such as models 510 and 550 may include the following two approaches:

One approach may be to obtain a large amount of emotional speech data, cluster the emotional speech data to obtain multiple sets of data for training, and then use an existing

emotional speech cloning base model (such as the prosody BERT model, or emotional VITS base model) to train the model 510, 550 based on the multiple sets of training data. From this, the systems and methods disclosed herein may obtain a corresponding emotional voice model 510, 550.

5 Alternatively, or in addition, an approach could be to obtain emotional speech data corresponding to the scenario or context related to the dialogue process, cluster the emotional speech data to obtain multiple sets of data for training, and train a personalized emotional voice model based on the multiple sets of data.

10 Fig. 6 illustrates an example embodiment of a process 600 for providing a response text 610 to a user as described herein, for example, corresponding to step 160 of method 100 in Fig. 1. The response text 610 may be that generated at step 250 as described in relation to Fig. 2, and the response text 610 may be provided with a target emotion and target tone 620, such as those determined at step 570 in Fig. 5.

15 After determining the target emotion and target tone 620 suitable for feedback of genuine emotions, the response text 610 may be played using the target emotion and target tone 620 to achieve real-time emotional interaction with the user as described herein.

The process 600 of using the target emotion and target tone 620 to play the response text 610 can include the following two approaches:

20 The first approach involves a process 600 of using the received target emotion and target tone 620 to play the response text 610, which can include the following steps:

At step 630, the process 600 may synthesize audio with the target emotion and target tone 620 based on the response text 610. The process 600 may modulate an audio speech output based on the target emotion and tone. At step 640, the process 600 may then play the
25 synthesized audio using the identified target emotion and target tone. This approach, which involves synthesizing the entire response text 610 into audio, may allow for the one-time playback of the response text 610, which may avoid interruptions during playback and therefore provide a seamless delivery of the response text 610.

Alternatively, or in addition, the process 600 may use the target emotion and target tone 620
30 to play the response text 610, which can include the following steps:

At step 650, the process 600 may intercept character sequences in order according to their arrangement in the response text 610 using a fixed window length. The window length may

be a pre-set window length, or may be determined based on one or more attributes of the response text, determined user emotion, or scenario/context as described herein.

At step 660, for each intercepted character sequence at step 650, the process 600 may synthesize a corresponding audio frame using the target emotion and target tone 620.

- 5 Finally, the process 600 may stream the audio frame for playback at step 670. It will be understood that the frames may therefore be streamed for audio playback as corresponding response text 610 and target emotion and tone 620, provided by the window length at 650.

10 In example embodiments, low latency may be one of the important performance features of voice interaction products. Therefore, in order to improve the low latency of responses, the method and system embodiments disclosed herein may adopt streaming playback when playing the response text, as set out at steps 650 to 670. A key point of streaming playback is to provide for real-time sentence segmentation of the response text (for example, based on punctuation marks), and then, after real-time sentence segmentation, intercepting character sequences, such as at step 650 in order according to their arrangement in the response text
15 610 using a fixed window length. From here, for each intercepted character sequence, synthesizing the corresponding audio frame at step 660 may be carried out using the target emotion and target tone 620, and then the process may stream the audio frame for playback at 670. In this way, due to the small amount of data in the intercepted character sequences, audio frames with the target emotion and target voice can be quickly synthesized based on
20 the character sequences and played back rapidly, thereby achieving low latency in voice interaction. Additionally, this low-latency processing method may quickly enable voice interaction and enhance a user's voice interaction experience even in cases of high concurrency of voice data.

25 In an example embodiment, the process 600 as described herein to provide for streaming playback to reduce latency for real-time responses could include one or more of:

- Text Segmentation: Dividing the response text into paragraphs or sentences based on punctuation or other rules.

- Window Processing: Setting a pre-set window length for each paragraph or sentence and processing them sequentially.

30 - Audio Synthesis: For each window of text, synthesizing corresponding audio frames using the target emotional and tone features.

- Streaming Playback: Continuously playing the synthesized audio frames to form a coherent speech response.

The above process 600 for using the target emotion and target tone 620 to play the response text 610 can be flexibly selected based on scenario needs, and this embodiment is not intended to impose any limitations in this regard. Regardless of which process is used, after playing the response text with the target emotion and target voice, real-time emotional interaction with the user can be achieved, allowing the user to perceive emotional feedback and thus enhancing their voice interaction experience.

For example, if the user's real-time genuine emotion is anger, the target emotion could be one that soothes anger, and the target tone could be one that is calming for anger. By using the target emotion and target tone to play the response text, not only can the response text be used to reply to the user's voice data, but the target emotion and target tone can also soothe the user's anger, providing the user with improved emotional feedback.

Fig. 7 illustrates a schematic diagram 700 of the interactive relationships among the models or modules that may be applied in the voice interaction methods and systems as disclosed herein.

The example embodiment of Fig. 7 may receive voice data 710 from a user as part of a voice interaction as disclosed herein. The example embodiment of Fig. 7 may include an Automatic Speech Recognition (ASR) model 720 for converting speech, such as the voice data 710, into text. The ASR model 720 may be configured to carry out the method steps 110 and/or 120 as described in relation to Fig. 1.

The example embodiment of Fig. 7 may further include a large language model 730 for generating response text and recognizing or determining semantic emotion parameters. The large language model 730 may be configured to carry out the method steps 130 and/or 134 as described in relation to Fig. 1.

The example embodiment of Fig. 7 may further include a multi-class emotion model 740 for recognizing or determining acoustic emotion parameters. The multi-class emotion model 740 may be configured to carry out the method steps 130 and/or 132 as described in relation to Fig. 1.

The example embodiment of Fig. 7 may further include a genuine emotion recognition model 750 for determining a user's emotion state. The emotion recognition model 750 may be configured to carry out the method steps 140 and/or 150 as described in relation to Fig. 1.

The example embodiment of Fig. 7 may further include an emotion-voice model 760 for determining target emotions and target voices. The emotion-voice model 760 may be configured to carry out the method step 160 as described in relation to Fig. 1.

In an example embodiment, when voice interaction is initiated with a user, the user's voice data may be acquired in real-time during a current conversation as part of the voice interaction. Following this, the systems and methods as disclosed herein may execute two processes:

Firstly, based on the ASR model 720, the voice data 710 may be converted into corresponding user text, which may then be provided to the large language model 730. The large language model 730 may then generate a response text corresponding to a suitable reply for feedback to the voice data 710 based on the user text, which may be provided to the emotion-voice model 760. As part of this, the large language model 730 may identify user intent within the voice data 710, to allow a response text to be generated based on a user query.

Simultaneously, the large language model 730 may perform emotion recognition on the user text to obtain semantic emotion parameters corresponding to a user's real-time semantic emotion, which may be provided to emotion recognition model 750.

Secondly, the multi-class emotion model 740 may be used to perform emotion recognition on the voice data 710 to obtain acoustic emotion parameters corresponding to a user's real-time acoustic emotion as described herein. The acoustic emotion parameters may be provided to the emotion recognition model 750.

Subsequently, the emotion recognition model 750 may combine the acoustic emotion parameters and the semantic emotion parameters to determine a user's real-time genuine emotion as described herein.

The user's genuine emotion may be then provided to the emotion-voice model 760. The emotion-voice model 760 may process the genuine emotion received from emotion model 750 to obtain a target emotion and target tone or voice as described herein, which may be used to play the response text generated by the large language model 730, thereby achieving real-time emotional interaction with the user, allowing the user to perceive emotional feedback, and enhancing a user's voice interaction experience.

In an alternative example embodiment, if a mature multi-label speech large model is available, which simultaneously possesses the functions of converting speech into text and recognizing semantic emotions, then the ASR model 720 and multi-class emotion model 740 in Fig.7 may be replaced by a single multi-label speech large model 770 as shown in Fig. 7.

Other module combinations may also be integrated in the systems and methods as described herein, including:

In an example embodiment, the ASR model 720 and the large language model 730 may be combined. For example, if the large language model 730 is sufficiently intelligent, it may be configured to provide both speech recognition, user intent recognition and response text generation simultaneously.

In an example embodiment, the emotion recognition model 750 and the emotion-voice model 760 may be combined. For example, because vocal features are also influenced by emotional states, the functions of these models could be integrated into a more comprehensive model.

FIG. 8 is a block diagram illustrating a system architecture of a voice interaction system 800 for providing an emotionally contextual response, according to various embodiments of the present disclosure.

As disclosed herein, the system 800 may comprise a voice data acquisition module 810 configured to acquire user voice data from a user during a voice interaction. The module 810 may comprise a speech recognition module 812 configured to convert the user voice data into corresponding user text.

The system 800 may further include a response text generation module 820 configured to generate a response text for feedback to the user based on the acquired user voice data from module 810. The module 820 may comprise a large language model (LLM) 822 configured to receive the acquired user voice data and generate the response text. The response text generation module 820 may also be configured to receive one or more user attributes 824 for generating the response text, where the response text generation module 820 may take into account both the user text and user voice data from module 810, and the user attributes 824 when generating the response text. In an example embodiment, the user attributes may comprise one or more of: user intent, determined scenario, scenario context, and/or habitual language patterns.

The system 800 may further include an emotion recognition module 830 configured to determine one or more acoustic emotion parameters and one or more semantic emotion parameters from the user voice data acquired at module 810. The emotion recognition module may comprise an acoustic emotion recognition sub-module 832 and a semantic emotion recognition module 834, which may be configured to analyse a respective acoustic and semantic portion of the user voice data and/or the user text.

In an example embodiment, the response text generation module 820 may be configured to generate an initial response text based on speech-to-text results from voice data acquisition module 810 and/or one or more user attributes 824 as described above. Further, the text generation module 820 may analyze the acoustic and/or semantic emotional parameters from the emotional recognition module 830, and adjust the initial response text based on at least a content and/or a tone of the determined acoustic emotion and/or semantic emotion of the user.

The system 800 may further include an emotion combination module 840 configured to combine the one or more acoustic emotion parameters and the one or more semantic emotion parameters from the emotion recognition module 830 to determine a user's emotion state.

The system 800 may further include a target determination module 850 configured to determine a target emotion and a target tone for delivering the generated response text generated by module 820 to the user based on the user's emotion state determined by module 840.

The system 800 may further include a response module 860 configured to deliver the response text generated by module 820 using the target emotion and target tone determined by the target determination module 850. The response module 860 may include an audio synthesis module 862 configured to modulate an audio speech output based on the target emotion and tone.

Embodiments of the present disclosure may be provided as a network of communicating devices (i.e. a "computerized network"). Embodiments of the invention may be also provided as a software application downloadable into a computer device to facilitate the method. The software application may be a computer program product, which may be stored on a non-transitory computer-readable medium on a tangible data-storage device (such as a storage device of a server, or one within a user device).

FIG. 9 is a block diagram illustrating a technical architecture of a computing server for implementing the processes of any one of Figs. 1 to 8 according to various embodiments of the present disclosure. The technical architecture 900 represents a computer server suitable for use as the voice interaction system 800 of Fig. 8, and/or for carrying out the voice interaction method 100 of Fig. 1, according to various embodiments of the present disclosure. While a single computing server is shown, the methods may be implemented across multiple computers in a distributed computing environment.

The technical architecture 900 includes a processor 902 (which may be referred to as a central processor unit or CPU) that is in communication with memory devices including secondary storage 904 (such as disk drives), read only memory (ROM) 906, random access memory (RAM) 908. The processor 902 may be implemented as one or more CPU chips. The RAM 908 may be partitioned to efficiently process different tasks, such as data pre-processing, feature engineering, model training, and evaluation as described herein. The partitioning of RAM 908 allows for efficient processing of large-scale data processing tasks, which is particularly beneficial for complex predictive modelling operations. The technical architecture may further comprise input/output (I/O) devices 910, and network connectivity devices 912.

The secondary storage 904 is typically comprised of one or more disk drives or tape drives and is used for non-volatile storage of data and as an over-flow data storage device if RAM 908 is not large enough to hold all working data. Secondary storage 904 may be used to store programs which are loaded into RAM 908 when such programs are selected for execution.

In an example embodiment, the secondary storage 904 includes a component 904a comprising non-transitory instructions operative by the processor 902 to perform various operations of the voice interaction method 100 of Fig. 1 and as described herein. For example, the component 904a may implement the functionality of the various components and modules illustrated in Fig. 8, and or implement the method steps as set out in Fig. 1.

The ROM 906 is used to store instructions and perhaps data which are read during program execution. The secondary storage 904, the RAM 908, and/or the ROM 906 may be referred to in some contexts as computer readable storage media and/or non-transitory computer readable media.

I/O devices 910 may include printers, video monitors, liquid crystal displays (LCDs), plasma displays, touch screen displays, keyboards, keypads, switches, dials, mice, track balls, voice recognizers, card readers, paper tape readers, or other well-known input devices.

The processor 902 executes instructions, codes, computer programs, scripts which it accesses from hard disk, floppy disk, optical disk (these various disk-based systems may all be considered secondary storage 904), flash drive, ROM 906, RAM 908, or the network connectivity devices 912. While only one processor 902 is shown, multiple processors may be present. Thus, while instructions may be discussed as executed by a processor, the instructions may be executed simultaneously, serially, or otherwise executed by one or multiple processors.

The network connectivity devices 912 enable the technical architecture 900 to communicate with external systems, which may include distributed computing resources or cloud-based services.

Although the technical architecture 900 is described with reference to a computer, it should be appreciated that the technical architecture 900 may be formed by two or more computers in communication with each other that collaborate to perform a task. For example, but not by way of limitation, an application may be partitioned in such a way as to permit concurrent and/or parallel processing of the instructions of the application. Alternatively, the data processed by the application may be partitioned in such a way as to permit concurrent and/or parallel processing of different portions of a data set by the two or more computers. In an embodiment, virtualization software may be employed by the technical architecture 900 to provide the functionality of a number of servers that is not directly bound to the number of computers in the technical architecture 900. In an embodiment, the functionality disclosed above may be provided by executing the application and/or applications in a cloud computing environment. Cloud computing may comprise providing computing services via a network connection using dynamically scalable computing resources. A cloud computing environment may be established by an enterprise and/or may be hired on an as-needed basis from a third-party provider.

By programming and/or loading executable instructions onto the technical architecture 900, at least one of the CPU 902, the RAM 908, and the ROM 906 are changed, transforming the technical architecture 900 in part into a specific purpose machine or apparatus having the novel functionality taught by the present disclosure. It is fundamental to the electrical engineering and software engineering arts that functionality that can be implemented by loading executable software into a computer can be converted to a hardware implementation by well-known design rules. Such hardware implementations may include field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs), which can further enhance the system's performance for specific information processing tasks.

In the foregoing detailed description, embodiments of the present disclosure in relation to a voice interaction system and method are described with reference to the provided figures. The description of the various embodiments herein is not intended to call out or be limited only to specific or particular representations of the present disclosure, but merely to illustrate non-limiting examples of the present disclosure. The present disclosure serves to address at least one of the mentioned problems and issues associated with the prior art. Although only some embodiments of the present disclosure are disclosed herein, it will be apparent to a person having ordinary skill in the art in view of the present disclosure that a variety of changes and/or

modifications can be made to the disclosed embodiments without departing from the scope of the present disclosure. It will be understood by those skilled in the art that many variations of the embodiments can be made within the scope of the following claims. Moreover, features of one or more embodiments may be mixed and matched with features of one or more other
5 embodiments. Therefore, the scope of the disclosure as well as the scope of the following claims is not limited to embodiments described herein.

Claims

1. A method for voice interaction with a user, the method comprising:
acquiring user voice data forming part of a voice interaction;
5 generating a response text for feedback to the user based on the acquired user voice data;
performing emotion recognition on the user voice data to determine one or more acoustic emotion parameters, and one or more semantic emotion parameters of the user voice data;
10 combining the one or more acoustic emotion parameters and one or more semantic emotion parameters to determine a user's emotion state;
determining a target emotion and a target tone from the user's emotion state for delivery of the generated response text; and
providing the response text as an audio output using the target emotion and the target
15 tone.
2. The method of claim 1, wherein the method steps are carried out sequentially in real time.
- 20 3. The method of claim 1 or claim 2, further comprising converting the user voice data into corresponding user text via a speech recognition model for converting speech into text.
4. The method of claim 3, wherein the step of generating a response text comprises inputting the user text into a large language model for generating response texts.
25
5. The method of claim 4, wherein the step of generating a response text further comprises inputting the user text and one or more user attributes into a large language model for generating response texts.
- 30 6. The method of any one of claims 1 to 5, wherein the step of generating a response text further comprises:
generating an initial response text based on speech-to-text results and/or one or more user attributes;
analyzing the acoustic and/or semantic emotional parameters from the user voice data;
35 and
adjusting the initial response text based on at least a content and/or a tone of the determined acoustic emotion and/or semantic emotion of the user.

7. The method of claim 6, wherein the user attributes comprise one or more of: user intent, determined scenario, scenario context, habitual language patterns, personal attributes of the user, historical interaction data, and/or scenario attributes.

5

8. The method of any one of claims 1 to 7, wherein the determined one or more acoustic emotion parameters and one or more semantic emotion parameters correspond to an acoustic emotion state and semantic emotion state of the user, optionally wherein the determined target emotion and target tone correspond to the determined user emotion state.

10

9. The method of any one of claims 1 to 8, wherein the step of performing emotion recognition comprises the use of a machine learning model trained to recognize acoustic features associated with emotional states from the user voice data, and/or wherein the step of performing emotion recognition comprises the use of a machine learning model trained to recognize semantic features associated with emotional states from the user voice data.

15

10. The method of any one of claims 1 to 9, wherein the step of determining the target emotion and target tone further comprises the identification and/or selection of an emotional feedback style that aligns with the user's determined true emotion.

20

11. The method of any one of claims 1 to 10, wherein the target tone is determined based on at least one of a speech tempo, speech rate, pitch variation, speech timbre, speech rhythm, speech pauses, and/or speech volume of the user voice data, and/or semantic structure, contextual analysis, topic modelling, and/or presence of words in the user voice data associated with a user's emotional state.

25

12. The method of any one of claims 1 to 11, wherein the target emotion is determined based on a pre-defined emotional mapping that correlates the combined acoustic emotion and semantic emotion to one or more emotional states.

30

13. The method of any one of claims 1 to 12, wherein the response text is generated by a natural language processing (NLP) algorithm, optionally wherein the NLP adapts to the recognized emotions in the user voice data and the determined target emotion and target tone.

35

14. A system for voice interaction, comprising:
a voice data acquisition module configured to acquire user voice data from a user during a voice interaction;

a response text generation module configured to generate a response text for feedback to the user based on the acquired user voice data;

an emotion recognition module configured to determine one or more acoustic emotion parameters and one or more semantic emotion parameters from the user voice data;

5 an emotion combination module configured to combine the one or more acoustic emotion parameters and the one or more semantic emotion parameters to determine a user's emotion state;

a target determination module configured to determine a target emotion and a target tone for delivering the generated response text to the user based on the user's emotion state;

10 and

a response module configured to deliver the response text using the target emotion and target tone determined by the target determination module.

15 15. The system of claim 14, wherein the voice data acquisition module comprises a speech recognition module configured to convert the user voice data into corresponding user text.

16. The system of claim 14 or claim 15, wherein the response text generation module comprises a large language model configured to receive the acquired user voice data and generate the response text.

20

17. The system of claim 16, wherein the response text generation module is further configured to receive one or more user attributes for generating the response text, optionally wherein the user attributes comprise one or more of: user intent, determined scenario, scenario context, habitual language patterns, personal attributes of the user, historical interaction data, and/or scenario attributes.

25

18. The system of any one of claims 14 to 17, wherein the response text generation module is further configured to:

30 generate an initial response text based on speech-to-text results and/or one or more user attributes;

analyze the acoustic and/or semantic emotional parameters from the user voice data; and

adjust the initial response text based on at least a content and/or a tone of the determined acoustic emotion and/or semantic emotion of the user.

35

19. The system of any one of claims 14 to 18, wherein the emotion recognition module
5 comprises a machine learning model configured to detect emotional features of the user voice
data, optionally wherein the emotional features include acoustic emotion parameters and/or
semantic emotion parameters.

20. The system of any one of claims 14 to 19, wherein the response module includes an
10 audio synthesis module configured to modulate an audio speech output based on the target
emotion and tone.

15