Claims

1. A method for voice interaction with a user, the method comprising:

acquiring user voice data forming part of a voice interaction;

determining one or more scenario attributes associated with the voice interaction;

determining one or more user attributes associated with the user;

generating a response text for feedback to the user based on the acquired user voice data, the determined scenario attributes, and the determined user attributes, wherein the response text is generated by a large language model comprising a natural language processing (NLP) algorithm;

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, wherein the emotion recognition is performed by a voice model adapted to the determined scenario attributes and the determined user attributes;

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, the determined scenario attributes, and the determined user attributes for delivery of the generated response text; and

providing the response text as an audio output using the target emotion and the target tone;

wherein the large language model is configured to adapt to the determined user’s emotion state, and wherein the step of generating a response text further comprises:

generating an initial response text based on speech-to-text results of the acquired user voice data, the determined scenario attributes, and the determined user attributes;

analyzing the acoustic and semantic emotional parameters determined by the voice model, and the determined user’s emotion state; and

adjusting the initial response text based on at least a content and a tone of the determined acoustic and semantic emotional parameters, the determined user’s emotion state, the determined scenario attributes, and the determined user attributes.

2. The method of claim 1, wherein the method steps are carried out sequentially in real time.

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 any one of claims 1 to 3, 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. The method of any one of claims 1 to 4, wherein the scenario attributes comprise at least one of: an interaction scenario type, a business type associated with the voice interaction, or a device type used by the user.

6. The method of any one of claims 1 to 5, 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.

7. The method of any one of claims 1 to 6, 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.

8. The method of any one of claims 1 to 7, 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 determined user's emotion state.

9. The method of any one of claims 1 to 8, 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.

10. The method of any one of claims 1 to 9, 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.

11. The method of any one of claims 1 to 10, wherein the large language model comprising the NLP adapts to the recognized emotions in the user voice data and the determined target emotion and target tone.

12. A system for voice interaction, comprising:

a voice data acquisition module configured to acquire user voice data from a user during a voice interaction, and determine one or more scenario attributes associated with the voice interaction and one or more user attributes associated with the user;

a response text generation module configured to generate a response text for feedback to the user based on the acquired user voice data, the determined scenario attributes, and the determined user attributes, wherein the response text generation module comprises a large language model comprising a natural language processing (NLP) algorithm;

an emotion recognition module comprising a voice model configured to determine one or more acoustic emotion parameters and one or more semantic emotion parameters from the user voice data, wherein the voice model is adapted to the determined scenario attributes and the determined user attributes;

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, the determined scenario attributes, and the determined user attributes; and

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

wherein the large language model comprising the NLP algorithm adapts to the determined user’s emotion state, and wherein the response text generation module is further configured to:

generate an initial response text based on speech-to-text results of the acquired user voice data, the determined scenario attributes, and the determined user attributes;

analyze the acoustic and semantic emotional parameters determined by the voice model, and the determined user’s emotion state; and

adjust the initial response text based on at least a content and a tone of the determined acoustic and semantic emotional parameters, the determined user’s emotion state, the determined scenario attributes, and the determined user attributes.

13. The system of claim 12, wherein the voice data acquisition module comprises a speech recognition module configured to convert the user voice data into corresponding user text.

14. The system of claim 12 or claim 13, 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.

15. The system of any one of claims 12 to 14, wherein the scenario attributes comprise at least one of: an interaction scenario type, a business type associated with the voice interaction, or a device type used by the user.

16. The system of any one of claims 12 to 15, wherein 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.

17. The system of any one of claims 12 to 16, wherein the response module includes an audio synthesis module configured to modulate an audio speech output based on the target emotion and tone.