Claims

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

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 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;

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;

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

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 claim 3, wherein the step of generating a response text comprises inputting the user text into a large language model for generating response texts.

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.

6. The method of any one of claims 1 to 5, wherein the step of generating an initial response text further comprises generating an initial response text based on the speech-to-text results and one or more user attributes.

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.

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.

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.

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 determined user's emotion state.

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.

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.

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.

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;

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; 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 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;

analyze the acoustic and/or semantic emotional parameters determined 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.

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.

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.

18. The system of any one of claims 14 to 17, wherein the response text generation module is further configured to generate an initial response text based on the speech-to-text results and one or more user attributes.

19. The system of any one of claims 14 to 18, 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.

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