CLAIMS

1. A voice dialogue processing method based on a multi-modal feature, comprising:

acquiring, in a process of performing dialogue interaction with a user, first voice information that the user currently inputs, wherein the first voice information comprises a silent segment;

determining, according to text information of the first voice information and historical context information of the first voice information, semantic feature information of the text information;

determining, according to a voice fragment, which is before the silent segment, in the first voice information, phonetic feature information of the first voice information;

acquiring temporal feature information of the first voice information; and

determining, according to the semantic feature information, the phonetic feature information and the temporal feature information, whether the user ends voice input.

2. The method according to claim 1, wherein the determining, according to text information of the first voice information and historical context information of the first voice information, semantic feature information of the text information comprises:

performing voice recognition on the first voice information to obtain text information of the first voice information;

acquiring historical context information of the first voice information; and

inputting the text information and the historical context information into a semantic representation model to obtain semantic feature information of the text information.

3. The method according to claim 1, wherein the determining, according to a voice fragment, which is before the silent segment, in the first voice information, phonetic feature information of the first voice information comprises:

acquiring a voice fragment of a first preset time length, which is before the silent segment, in the first voice information;

segmenting, according to a second preset time length, the voice fragment to obtain multiple

voice fragments;

extracting respective acoustic feature information of the multiple voice fragments, and splicing the respective acoustic feature information of the multiple voice fragments, respectively, to obtain respective splicing features of the multiple voice fragments;

inputting the splicing features into a deep residual network to obtain phonetic feature information of the first voice information.

4. The method according to claim 1, wherein the acquiring temporal feature information of the first voice information comprises:

acquiring a voice duration, a speaking speed and a text length of the first voice information;

inputting the voice duration, the speaking speed and the text length into a pre-trained multi-layer perceptron MLP model to obtain temporal feature information of the first voice information.

5. The method according to claim 1, wherein the determining, according to the semantic feature information, the phonetic feature information and the temporal feature information, whether the user ends voice input comprises:

inputting the semantic feature information, the phonetic feature information and the temporal feature information into a multi-modal fusion model;

determining, according to an output result of the multi-modal fusion model, whether the user ends voice input.

6. The method according to claim 1, further comprising:

determining, in the case of determining that the user ends the voice input, first reply voice information corresponding to the first voice information, and outputting the first reply voice information.

7. The method according to claim 1, further comprising:

acquiring, in the case of determining that the user does not end the voice input, second voice information input again by the user; and

determining, according to the first voice information and the second voice information,

corresponding second reply voice information, and outputting the second reply voice information.

8. An electronic device, comprising: a memory, and a processor, wherein the memory stores computer instructions that, when executed by the processor, implement the voice dialogue processing method based on a multi-modal feature according to any one of claims 1 to 7.

9. A non-transitory computer-readable storage medium storing computer instructions, wherein the computer instructions are used to cause the computer to perform the voice dialogue processing method based on a multi-modal feature according to any one of claims 1 to 7.

10. A computer program product, wherein instructions in the computer program product, when executed by a processor, implement the voice dialogue processing method based on a multi-modal feature according to any one of claims 1 to 7.