Claims

1. A streaming end-to-end speech recognition method, comprising:

extracting and encoding speech acoustic features of a received voice stream in units of frames;

performing block processing on encoded frames, and predicting a number of activation points included in a same block that need to be encoded and outputted; and

determining position(s) of activation point(s) that need(s) to be decoded and outputted according to a prediction result, to allow a decoder to perform decoding at the position(s) of the activation point(s) and output a recognition result.

2. The method of claim 1, wherein:

the block includes an encoding result corresponding to a frame of the voice stream;

the prediction result includes: whether a current block includes an activation point that needs to be encoded and outputted;

determining the position(s) of the activation point(s) that need(s) to be decoded and outputted according to the prediction result, includes:

determining position(s) of block(s) including the activation point(s) as the position(s) of the activation point(s).

3. The method of claim 2, further comprising:

determining Attention coefficients of encoding results of each frame respectively, the Attention coefficients being used to describe probabilities that corresponding frames need to be decoded and outputted; and

verifying the prediction result based on the Attention coefficients.

4. The method of claim 1, wherein:

the block includes encoding results corresponding to a multi-frame voice stream;

the method further includes: determining Attention coefficients of encoding results of each frame respectively, the Attention coefficients being used to describe probabilities that corresponding frames need to be decoded and outputted;

determining the position(s) of the activation point(s) that need(s) to be decoded and outputted according to the prediction result, includes:

comparing the Attention coefficients of each frame in a same block and sort the Attention coefficients in order of magnitudes; and

determining positions of frames associated with a corresponding number of first few highest Attention coefficients among encoding results of each frame included in the block as the position(s) of the activation point(s) according to the number of activation points included in the block.

5. The method of claim 4, further comprising:

adaptively adjust a block size according to a frequency of occurrence of predicted activation points.

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

performing the block processing on the ~~encoding~~ encoded frames includes:

buffering the encoding results of the encoded frames; and

determining encoding results of various frames currently buffered as one block when a number of frames associated with the encoding results added to a buffer reaches a block size.

7. The method of claim 6, further comprising:

deleting the encoding results of the various frames of the one block from the buffer after prediction processing of the one block is completed.

8. A method for establishing a predictive model, comprising:

obtaining a training sample set, the training sample set including a plurality of block data and labeling information, wherein each block data frame includes an encoding result of respectively encoding a plurality of frames of a voice stream, and the labeling information includes a number of activation points included in each block that need to be decoded and outputted; and

inputting the training sample set into a predictive model for model training.

9. The method of claim 8, wherein:

the training sample set includes a situation in which multiple frames of the voice stream that correspond to a same modeling unit are divided into different blocks.

10. A method for providing speech recognition services, comprising:

receiving a voice stream provided by an application system after receiving a calling request from the application system;

extracting and encoding speech acoustic features of a received voice stream in units of frames;

performing block processing on encoded frames, and predicting a number of activation points included in a same block that need to be encoded and outputted;

determining a position of an activation point that needs to be decoded and outputted according to a prediction result, to allow a decoder to perform decoding at the

position of the activation point to obtain a speech recognition result; and

returning the speech recognition result to the application system.

11. A method for obtaining speech recognition information, comprising:

an application system submitting a call request and a voice stream to be recognized to a cloud service system by invoking an interface provided by the cloud service system, the cloud service system extracting speech acoustic features of the received voice stream in units of frames and performing encoding, performing block processing on encoded frames, predicting a number of activation points included in a same block that need to be encoded and outputted, and performing decoding at the position of the activation point to obtain a speech recognition result after determining a position of an activation point that needs to be decoded and outputted according to a prediction result; and

receiving the speech recognition result returned by the cloud service system.

12. A method for implementing court's self-service case filing, comprising:

a self-service case-filing all-in-one machine receiving case filing request information through voice input;

extracting and encoding speech acoustic features of a received voice stream in units of frames;

performing block processing on encoded frames, and predicting a number of activation points included in a same block that need to be encoded and outputted;

determining a position of an activation point that needs to be decoded and outputted according to a prediction result, to allow a decoder to perform decoding at the

position of the activation point and determine a recognition result; and

entering the recognition result into an associated case-filing information database.

13. A method for upgrading a terminal device, comprising:

providing upgrade advice information to the terminal device; and

granting the terminal device a permission to perform streaming speech recognition in an upgraded manner after receiving an upgrade request submitted by the terminal device, performing the streaming speech recognition in the upgraded manner including: extracting and encoding speech acoustic features of a voice stream in units of frames, performing block processing on encoded frames, and predicting a number of activation points included in a same block that need to be encoded and outputted; and obtaining a speech recognition result by performing decoding at a position of an activation point through a decoder after determining the position of the activation point based on a prediction result.

14. The method of claim 13, wherein:

the terminal device includes a smart speaker device.

15. The method of claim 13, further comprising:

disabling the permission to perform the streaming speech recognition in the upgraded manner for the terminal device according to a downgrade request submitted by the terminal device.

16. A streaming end-to-end speech recognition apparatus, comprising:

an encoding unit configured to extract and encode speech acoustic features of a received voice stream in units of frames;

a prediction unit configured to perform block processing on encoded frames, and predict a number of activation points included in a same block that need to be encoded and outputted; and

an activation point position determination unit configured to determine position(s) of the activation point(s) that need(s) to be decoded and outputted according to a prediction result, to allow a decoder to perform decoding at the position(s) of the activation point(s) and output a recognition result.

17. An apparatus for building a predictive model, comprising:

a training sample set acquisition unit configured to obtain a training sample set, the training sample set including a plurality of block data and labeling information, wherein each block data frame includes an encoding result of respectively encoding a plurality of frames of a voice stream, and the labeling information includes a number of activation points included in each block that need to be decoded and outputted; and

an input unit configured to input the training sample set into a predictive model for model training.

18. An apparatus for providing speech recognition services, which is applied in a cloud service system, comprising:

a voice stream receiving unit configured to receive a voice stream provided by an application system after receiving a calling request from the application system;

an encoding unit configured to extract and encode speech acoustic features of a received voice stream in units of frames;

a prediction unit configured to perform block processing on encoded frames, and predict a number of activation points included in a same block that need to be encoded and outputted;

an activation point position determination unit configured to determine a position of an activation point that needs to be decoded and outputted according to a prediction result, to allow a decoder to perform decoding at the position of the activation point to obtain a speech recognition result; and

a recognition result returning unit configured to return the speech recognition result to the application system.

19. An apparatus for obtaining speech recognition information, which is applied in an application system, comprising:

a submission unit configured to submit a call request and a voice stream to be recognized to a cloud service system by invoking an interface provided by the cloud service system, the cloud service system extracting speech acoustic features of the received voice stream in units of frames and performing encoding, performing block processing on encoded frames, predicting a number of activation points included in a same block that need to be encoded and outputted, and performing decoding at the position of the activation point to obtain a speech recognition result after determining a position of an activation point that needs to be decoded and outputted according to a prediction result; and

a recognition result receiving unit configured to receive the speech recognition result returned by the cloud service system.

20. An apparatus for implementing court self-service case filing, which is is applied in a self-service case-filing all-in-one machine, comprising:

a request receiving unit configured to receive case filing request information through voice input;

an encoding unit configured is used to extract and encode speech acoustic features of a received voice stream in units of frames;

a prediction unit configured to perform block processing on encoded frames, and predict a number of activation points included in a same block that need to be encoded and outputted;

an activation point position determination unit configured to determine a position of an activation point that needs to be decoded and outputted according to a prediction result, to allow a decoder to perform decoding at the position of the activation point and determine a recognition result; and

an information entry unit configured to enter the recognition result into an associated case-filing information database.

21. An apparatus for upgrading a terminal device, comprising:

an upgrade suggestion providing unit configured to provide upgrade advice information to a terminal device; and

an authority granting unit configured to grant the terminal device a permission to perform streaming speech recognition in an upgraded manner after receiving an upgrade

request submitted by the terminal device, performing the streaming speech recognition in the upgraded manner including: extracting and encoding speech acoustic features of a voice stream in units of frames, performing block processing on encoded frames, and predicting a number of activation points included in a same block that need to be encoded and outputted; and obtaining a speech recognition result by performing decoding at a position of an activation point through a decoder after determining the position of the activation point based on a prediction result.

22. A computer-readable storage medium on which a computer program is stored, wherein: when the program, when executed by a processor, implements the steps of the method of any one of claims 1 to 15.

23. An electronic device comprising:

one or more processors; and

memory associated with the one or more processors, the memory being configured to store program instructions that, when read and executed by the one or more processors, execute the steps of the method of any one of claims 1 to 15.