Claims

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

extracting (S301) and encoding speech acoustic features of a received voice stream in units of frames;

performing (S302) 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 (S303) 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 same block includes an encoding result corresponding to a frame of the voice stream; and

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

3. The method of any one of the preceding claims, wherein 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).

4. The method of any one of the preceding claims, 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.

5. The method of any of the preceding claims, wherein:

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

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.

6. The method of any of the preceding claims, wherein 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.

7. The method of any of the preceding claims, further comprising:

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

8. The method of any one of the preceding claims, wherein:

performing the block processing on the encoding frames includes:

buffering the encoding 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.

9. The method of claim 8, 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.

10. A streaming end-to-end speech recognition system, comprising:

one or more processors (1520); and

memory (1504) storing executable instructions that, when executed by the one or more processors (1520), cause the one or more processors (1520) to perform acts comprising:

extracting (S301) and encoding speech acoustic features of a received voice stream in units of frames;

performing (S302) 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 (S303) 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.

11. The system of claim 10, wherein the same block includes an encoding result corresponding to a frame of the voice stream; and

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

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).

12. The system of any one of preceding claims 10-11, the acts 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.

13. The system of any of preceding claims 10-12, wherein:

the same block includes encoding results corresponding to a multi-frame voice stream, and

the acts further include: determine 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.

14. The system of any of the preceding claims 10-13, the acts further comprising:

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

15. A computer-readable storage medium on which a computer program is stored, wherein: when the program, when executed by a processor (1520), implements the steps of the method of any one of claims 1 to 9.