CLAIMS

What is Claimed is:

1. A streaming speech synthesis system for streaming input, comprising: a streaming text dynamic output module, a streaming input encoding module, and a streaming speech decoding and generation module, wherein

the streaming text dynamic output module is configured to obtain streaming text information, and dynamically adjust a maximum context of released text information according to an overall real-time factor of streaming speech synthesis, to cause streaming synthesis not to be interrupted;

wherein dynamically adjusting the maximum context of released text information according to the overall real-time factor of streaming speech synthesis, to cause streaming synthesis not to be interrupted comprises:

obtaining a corresponding synthesis real-time factor through calculation according to synthesis time and audio duration of streaming output audio generated last time;

calculating, according to the synthesis real-time factor, a maximum waiting time without interruption in which a length of the streaming output audio remains unchanged;

obtaining a first text length of streaming text received during the maximum waiting time, and obtaining a second text length of text required for the streaming output audio generated last time; and calculating the maximum context of released text information this time according to the first text length and the second text length, wherein the streaming synthesis is in an interruption state when the maximum context of released text information this time is less than 0;

the streaming input encoding module is configured to perform speech encoding processing on received streaming text information according to the maximum context of the released text information, to output a text information feature required for streaming synthetic audio; and

the streaming speech decoding and generation module is configured to perform speech decoding processing on the text information feature, to output a time domain signal or a spectral feature of a streaming speech.

2. The streaming speech synthesis system for streaming input according to claim 1, wherein in a process of calculating the synthesis real-time factor by the streaming text dynamic output module, a result of dividing the synthesis time by the audio duration serves as the synthesis real-time factor.

3. The streaming speech synthesis system for streaming input according to claim 1, wherein the calculating of, according to the synthesis real-time factor, the maximum waiting time without interruption in which the length of the streaming output audio remains unchanged comprises:

$t_{\max_waiting} = t_{audio} - 2t_{machine} = (1 - 2 * RTF)t_{audio}$; and

$\mathrm{RTF} = \frac{t_{machine}}{t_{audio}}$, wherein

$t_{max_waiting}$ represents the maximum waiting time without interruption; t_{audio}represents the audio duration; $t_{machine}$represents the synthesis time; and RTF represents the synthesis real-time factor.

4. The streaming speech synthesis system for streaming input according to claim 1, wherein the calculating of the maximum context of released text information this time according to the first text length and the second text length comprises:

using a difference between the first text length and the second text length as the maximum context of released text information this time.

5. The streaming speech synthesis system for streaming input according to claim 1, wherein the streaming input encoding module comprises a decoder module in a Transformer model; and the Transformer model captures long-range dependency information of an input sequence based on a self-attention mechanism, and then

performs real-time encoding on streaming text according to the long-range dependency information.

6. The streaming speech synthesis system for streaming input according to claim 1, wherein the streaming speech decoding and generation module performs speech synthesis based on a waveform concatenative speech synthesis, a statistical parametric speech synthesis, or a neural network-based end-to-end speech synthesis.

7. A streaming speech synthesis method for streaming input, comprising:

obtaining streaming text information, and dynamically adjusting a maximum context of released text information according to an overall real-time factor of streaming speech synthesis, to cause streaming synthesis not to be interrupted;

wherein dynamically adjusting the maximum context of released text information according to the overall real-time factor of streaming speech synthesis, to cause streaming synthesis not to be interrupted comprises:

obtaining a corresponding synthesis real-time factor through calculation according to synthesis time and audio duration of streaming output audio generated last time;

calculating, according to the synthesis real-time factor, a maximum waiting time without interruption in which a length of the streaming output audio remains unchanged;

obtaining a first text length of streaming text received during the maximum waiting time, and obtaining a second text length of text required for the streaming output audio generated last time; and calculating the maximum context of released text information this time according to the first text length and the second text length, wherein the streaming synthesis is in an interruption state when the maximum context of released text information this time is less than 0;

performing speech encoding processing on received streaming text information according to the maximum context of released text information this time,

to output a text information feature required for streaming synthetic audio; and

performing speech decoding processing on the text information feature, to output a time domain signal or a spectral feature of a streaming speech.

8. A computer-readable storage medium, with a computer program stored thereon, wherein the computer program, when executed by a processor, implements the streaming speech synthesis method for streaming input according to claim 7.

9. An electronic terminal, comprising: a processor and a memory, wherein

the memory is configured to store a computer program; and

the processor is configured to execute the computer program stored in the memory, to enable the terminal to perform the streaming speech synthesis method for streaming input according to claim 7.