

STREAMING SPEECH SYNTHESIS SYSTEM AND METHOD FOR STREAMING INPUT, TERMINAL, AND MEDIUM

FIELD OF THE INVENTION

[0001] The present disclosure relates to the field of speech synthesis technologies, and in particular to a streaming speech synthesis system and method for streaming input, a terminal, and a medium.

BACKGROUND OF THE INVENTION

[0002] With the popularity of speech interaction, speech synthesis technology is widely used. The conventional text-to-speech synthesis technology is typically based on a sentence level, and sentence-by-sentence processing of input text is used to generate corresponding speech. A text-to-speech (TTS) speech synthesis system has a structure as shown in Fig. 1. A front end may perform tasks including language determination, text standardization, linguistic feature extraction, prosody analysis, prediction, and the like, and a back end may perform tasks including concatenative synthesis, parameter synthesis, end-to-end model operation, vocoder operation, and the like.

[0003] A non-streaming speech synthesis technology usually requires a one-time input of text and involves a one-time output of speech. The non-streaming speech synthesis technology focuses on the overall computing speed of the speech synthesis system, and this TTS speech synthesis technology is not suitable for speech interaction. In actual applications, text information is often streaming, such as text generated by GPT, a chat conversation, text live broadcast, and the like. In such scenarios, the conventional speech synthesis technology based on sentence-level often results in high latency and poor experience, and cannot realize real conversion from streaming text to speech.

SUMMARY

[0004] An objective of the present disclosure is to provide a streaming speech synthesis system and method for streaming input, a terminal, and a medium, which have low latency and good experience, and implement real conversion from streaming text to speech.

[0005] A first aspect of the present disclosure provides 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, where 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; the streaming input encoding module is configured to perform speech encoding processing on received streaming text information according to the maximum context of 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.

[0006] In some embodiments of the first aspect of the present disclosure, the streaming text dynamic output module dynamically adjusts 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.

[0007] In some embodiments of the first aspect of the present disclosure, 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.

[0008] In some embodiments of the first aspect of the present disclosure, 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:

[0009] $t_{\max_waiting} = t_{audio} - 2t_{machine} = (1 - 2 * RTF)t_{audio}$; and

[0010] $RTF = \frac{t_{machine}}{t_{audio}}$,

[0011] where $t_{\max_waiting}$ 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.

[0012] In some embodiments of the first aspect of the present disclosure, 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.

[0013] In some embodiments of the first aspect of the present disclosure, 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.

[0014] In some embodiments of the first aspect of the present disclosure, 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.

[0015] A second aspect of the present disclosure provides 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; 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.

[0016] A third aspect of the present disclosure provides a computer-readable storage medium. The computer-readable storage medium stores a computer program, and the computer program, when executed by a processor, implements the streaming speech synthesis method for streaming input.

[0017] A fourth aspect of the present disclosure provides an electronic terminal, comprising: a processor and a memory, where 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.

[0018] As described above, the streaming speech synthesis system and method for streaming input, the terminal, and the medium in the present disclosure have the

following beneficial effects: the size and duration of output audio can be dynamically adjusted according to the overall real-time factor of streaming speech synthesis and the input rate of streaming text, and the best speech synthesis effect is obtained with that the streaming speech synthesis is not interrupted to the greatest extent.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] Fig. 1 is a schematic diagram of a structure of a TTS speech synthesis system according to an embodiment of the present disclosure.

[0020] Fig. 2 is a schematic diagram of a structure of a streaming speech synthesis system for streaming input according to an embodiment of the present disclosure.

[0021] Fig. 3 is a schematic flowchart of dynamically adjusting a maximum context by a streaming text dynamic output module according to an embodiment of the present disclosure.

[0022] Fig. 4 is a schematic flowchart of a streaming speech synthesis method for streaming input according to an embodiment of the present disclosure.

[0023] Fig. 5 is a schematic diagram of a structure of an electronic terminal according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0024] The following describes an implementation of the present disclosure through specific examples. A person skilled in the art can easily understand other advantages and effects of the present disclosure from the content disclosed in this specification. The present disclosure may be further implemented or applied through other different specific implementation modes. Various details in this specification may also be modified or changed in various ways based on different viewpoints and applications without departing from the spirit of the present disclosure. It should be noted that the following embodiments and features in the embodiments may be mutually combined if no conflict occurs.

[0025] It should be noted that the following description referred to the accompanying drawings illustrates several embodiments of the present disclosure. It should be understood that other embodiments may be further used, and mechanical, structural, electrical, and operational changes may be made without departing from the spirit and scope of the present disclosure. The following detailed description should not be considered as limitative, and a scope of embodiments of the present disclosure is limited only by claims of the published patent. The terms used herein are merely used for describing particular embodiments rather than limiting the present disclosure. Space-related terms such as "up", "down", "left", "right", "under", "below", "lower portion", "above", "upper portion", and the like may be used to describe a relationship between one component or feature and another component or feature as shown in the figures.

[0026] In the present disclosure, unless otherwise explicitly specified or defined, the terms such as "mount/install", "connect", "connection", "fix", and "retain" should be understood in a broad sense. For example, the connection may be a fixed connection, a detachable connection, or an integral connection; or the connection may be a mechanical connection or an electrical connection; or the connection may be a direct connection, an indirect connection through an intermediary, or an internal communication between two components. A person of ordinary skill in the art may understand the specific meanings of the foregoing terms in the present disclosure according to specific situations.

[0027] In addition, as used herein, singular forms "a", "an" and "the" are intended to also comprise plural forms, unless the context indicates otherwise. It should be further understood that the terms "include" and/or "comprise" indicate the presence of the stated features, operations, elements, components, items, species, and/or groups, but do not preclude the presence, occurrence, or addition of one or more other features, operations, elements, components, items, species, and/or groups. The terms "or" and "and/or" as used herein are to be construed as inclusive or to mean any one or any combination. Therefore, "A, B, or C" or "A, B, and/or C" means "any of

the following: A; B; C; A and B; A and C; B and C; A, B, and C". Exceptions to the definition occur only when a combination of elements, functions, or operations is inherently mutually exclusive in some manners.

[0028] The present disclosure provides a streaming speech synthesis system and method for streaming input, a terminal, and a medium, aiming for a dual-stream scenario in which both input and output are streaming, to realize a natural, smooth, low-latency, and continuous speech synthesis for a streaming input scenario. The present disclosure can be widely used in speech interaction, multi-modal dialog, and the like, especially in an interaction scenario involving a large model (such as GPT, and the like).

[0029] To make the objectives, technical solutions, and advantages of the present disclosure clearer, the technical solutions in the embodiments of the present disclosure will be further described in detail through the following embodiments and with reference to the accompanying drawings. It should be understood that the embodiments herein are provided for describing the present disclosure and are not intended to limit the present disclosure.

[0030] Before the present disclosure is further described in detail, nouns and terms related to the embodiments of the present disclosure are described, and the nouns and the terms related to the embodiments of the present disclosure are appropriate for the following explanations.

[0031] <1> Speech synthesis is a technology that simulates human speech by using a computer device and some specialized equipment to produce speech. Speech synthesis, also referred to as Text to Speech technology, may convert any text information into standard and smooth speech in real time, thus allowing a machine to speak like a human.

[0032] <2> Non-streaming speech synthesis usually requires a one-time input of text and involves a one-time output of speech. The non-streaming speech synthesis technology focuses on the overall computing speed of the speech synthesis system.

[0033] <3> Streaming speech synthesis may segment words and sentences based on text prediction, and send the synthesized audio back in segments. This speech synthesis technology mainly focuses on a first packet response time. The shorter the first packet response time, the faster a user gets a response. Due to the user's waiting time and blank time being reduced, the user will not lose patience after waiting for a response, therefore, a better experience is provided, and streaming speech synthesis is a TTS scheme more suitable for speech interaction.

[0034] <4> Phoneme is the most basic element in speech. There are about 128 to 256 phonemes. An average of 10 phonemes are uttered per second at a normal speaking rate, and an information rate is $I = [\log_2(256)10] \text{bps} = 80 \text{bps}$.

[0035] <5> GPT (Generative Pre-Trained Transformer) is a generative pre-trained Transformer model and is an internet-based deep learning model for text generation that may be trained with data. The GPT model is different from a small model that focuses on a specific task such as playing Go or machine translation. The GPT model is an AI large model with attributes comprising both large-scale and pre-training, and may be pre-trained on massive general data, thus greatly improving the generalization, versatility, and practicality of AI.

[0036] Embodiments of the present disclosure provide a streaming speech synthesis method for streaming input, a system for the streaming speech synthesis method for streaming input, and a storage medium that stores an executable program for implementing the streaming speech synthesis method for streaming input. Regarding the implementation of the streaming speech synthesis method for streaming input, embodiments of the present disclosure describe an exemplary implementation scenario of streaming speech synthesis for streaming input.

[0037] Fig. 2 is a schematic diagram of a structure of a streaming speech synthesis system for streaming input according to an embodiment of the present disclosure. A streaming speech synthesis system for streaming input in this embodiment mainly comprises the following modules: a streaming text dynamic output module 201, a streaming input encoding module 202, and a streaming speech decoding and generation module 203.

[0038] The streaming text dynamic output module 201 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.

[0039] In the embodiments of the present disclosure, an essential function of the streaming text dynamic output module 201 is to accumulate text information and release text information. Specifically, how to accumulate and release the text information is determined by the overall real-time factor of the streaming speech synthesis, so as to dynamically adjust the maximum context of the released text information for the streaming input encoding module 202 each time.

[0040] In some examples, a process in which the streaming text dynamic output module 101 dynamically adjusts the maximum context for the streaming input encoding module 202 according to the overall real-time factor of the streaming speech synthesis of the streaming text information is shown in Fig. 3 and mainly comprises the following steps:

[0041] Step S3A: Obtaining a corresponding synthesis real-time factor through calculation according to synthesis time and audio duration of streaming output audio generated last time.

[0042] In the embodiments of the present disclosure, the corresponding synthesis real-time factor is obtained through calculation according to synthesis time and audio

duration of streaming output audio generated last time, where a result of dividing the synthesis time by the audio duration serves as the synthesis real-time factor.

[0043] Specifically, the synthesis real-time factor is represented as RTF, the synthesis time is represented as t_{machine} , and the audio duration is represented as t_{audio} , then a calculation equation of the synthesis real-time factor is as follows:

[0044] $RTF = \frac{t_{\text{machine}}}{t_{\text{audio}}}$; equation (1).

[0045] It should be understood that in a scenario of streaming synthetic audio, broadcasting has already started when audio starts to be synthesized, and the broadcasting is finished when the audio synthesis is finished, so there is no waiting process. The above process is mainly used in a speech interaction scenario. For example, an intelligent robot can give a reply immediately after receiving a speech signal, without leaving the user waiting for too long. To ensure the best user experience, it is required that the synthesis time $t_{\text{machine}} \leq$ the audio duration t_{audio} , i.e., the real-time factor $RTF \leq 1$.

[0046] Step S3B: 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.

[0047] Specifically, when the audio duration t_{audio} is determined, i.e., the length of the streaming output audio remains unchanged, a calculation equation of the maximum waiting time without interruption $t_{\text{max_waiting}}$ is as follows:

[0048] $t_{\text{max_waiting}} = t_{\text{audio}} - 2t_{\text{machine}}$;

[0049] $= t_{\text{audio}} - 2 * RTF * t_{\text{audio}} = (1 - 2 * RTF)t_{\text{audio}}$; equation (2).

[0050] It should be noted that in $2 t_{\text{machine}}$ involved in equation (2) in the embodiments of the present disclosure, the multiplier 2 is determined by the calculation of streaming, specifically, a play time (t_{audio}) of the synthesized audio

minus a machine running time ($1 \cdot t_{\text{machine}}$) required to generate a streaming audio block this time to obtain a remaining waiting time, and then subtracts a machine running time ($1 \cdot t_{\text{machine}}$) required to generate a next streaming audio block to obtain a theoretical limit waiting time. Therefore, the value 2 is used as a multiplier of t_{machine} .

[0051] Step S3C: 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 the released text information this time according to the first text length and the second text length. When the maximum context of released text information this time is less than 0, the streaming synthesis is in an interruption state.

[0052] The first text length of streaming text received during the maximum waiting time $t_{\text{max_waiting}}$ is represented as $L_{\text{accepted_text}}$, and the second text length of text required for the streaming output audio generated last time is represented as L_{text} , then the maximum context of the released text information this time is calculated as follows:

[0053] $L_{\text{context}} = L_{\text{accepted_text}} - L_{\text{text}}$; equation (3), and

[0054] when $L_{\text{context}} < 0$, streaming synthesis is interrupted.

[0055] The streaming input encoding module 202 is configured to perform speech encoding processing on received streaming text information according to the maximum context of the released text information, to output text information features required for streaming synthetic audio.

[0056] In an embodiment, the streaming input encoding module 202 comprises a Decoder module in a Transformer model, and the streaming input encoding module 202 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. When the streaming input encoding module 202 performs encoding on the streaming text, a maximum context of the streaming input encoding module 202 changes dynamically, and text information of the required maximum context may be long or short. An output sequence of the streaming input encoding module 202 is a description and encoding of time sequence input, which provides ideally conditional features for the subsequent streaming speech decoding and generation module 203.

[0057] The Transformer model is a neural network model based on an attention mechanism and is widely used in the natural language processing task. The self-attention mechanism of the Transformer model may simultaneously calculate the relationship among all words in one sentence. By using the self-attention mechanism, the model may learn the long-range dependency relationship among words, thereby achieving a better result when processing natural language.

[0058] The Transformer model usually comprises an encoder and a decoder. The encoder is configured to encode an input sequence into a series of hidden states, uses a word embedding vector at each position of the input sequence as initial input, and then encodes information at each position into one fixed-length hidden vector through a multi-layer self-attention mechanism and a fully connected layer; and an output result of the encoder is sent to the decoder to perform processing in the next step. The decoder is a module configured to generate an output sequence, and receives output of the encoder and a previously generated partial output sequence to serve as input. The main task of the decoder is to generate a word at a next position until the entire sequence is generated. The decoder also comprises a multi-layer self-attention mechanism and a fully connected layer, and further comprises one linear transformation layer which maps the output of the decoder into the probability distribution of an output word.

[0059] It should be noted that in the application of the embodiments of the present disclosure, such as applications similar to GPT, only the decoder of the Transformer

model needs to be employed, and the decoder of the Transformer model is streaming. When a token is input, the encoder does not need to be used. The token refers to the smallest unit of input text, and may be one word, one character, or one punctuation mark; and using Chinese as an example, one Chinese character is usually regarded as one token.

[0060] The streaming speech decoding and generation module 203 is configured to perform speech decoding processing on the text information features, to output a time domain signal or a spectral feature of a streaming speech.

[0061] In the embodiments of the present disclosure, the streaming speech decoding and generation module 203 comprises stacked convolution modules and is applied to end-to-end speech synthesis in the embodiments of the present disclosure. The input of the streaming speech decoding and generation module 203 is phoneme sequence features, and the output of the streaming speech decoding and generation module 203 are time domain signals or spectral features of a streaming speech.

[0062] In an embodiment, the streaming speech decoding and generation module 203 may implement final speech synthesis based on a waveform concatenative speech synthesis, a statistical parametric speech synthesis, or a neural network-based end-to-end speech synthesis. The waveform concatenative speech synthesis includes presetting one large-capacity real-person speech library to better fit the human voice, where the speech library is annotated according to phonemes and different feature parameters, and during synthesis, different phonemes in the speech library are found based on linguistic features and are spliced together to complete speech synthesis. If no corresponding phonemes in the speech library are found, basic phonemes are used to splice. The statistical parameter speech synthesis does not require the large-capacity real-person speech library, but is modeled based on a limited speech library. Through deep learning, a mapping relationship between text features and the speech library is constructed, to build a parameter synthesis model; and when one linguistic feature is input, an audio feature is provided based on

a neural network, and then a speech waveform is synthesized through a vocoder, so that adjustment and change in pronunciation may be achieved by simply modifying some basic parameters (such as a fundamental frequency, a fundamental frequency fluctuation range, a speaking speed, and the like) of the speech. The neural network-based end-to-end speech synthesis includes directly modeling through text or text features and speech, and skipping a vocoder stage. The entire engine is a black box to the user, and after the text is input, speech is output directly.

[0063] It should be noted that when the streaming speech synthesis system for streaming input provided in the foregoing embodiments performs streaming speech synthesis for streaming input, division of the foregoing program modules is only used as an example for description. During actual application, the foregoing processing may be allocated to and completed by different program modules as required, that is, an internal structure of the apparatus may be divided into different program modules, to complete all or some of the processing described above.

[0064] An implementation process and principles of the streaming speech synthesis system for streaming input provided in the embodiments of the present disclosure are explained in detail above. A streaming speech synthesis method for streaming input, a terminal, and a medium in the embodiments of the present disclosure will be further described below with reference to specific embodiments.

[0065] Fig. 4 is a schematic flowchart of a streaming speech synthesis method for streaming input according to an embodiment of the present disclosure. The streaming speech synthesis method for streaming input in the embodiments of the present disclosure mainly comprises the following steps:

[0066] Step S41: 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.

[0067] Step S42: Performing speech encoding processing on received streaming text information according to the maximum context of released text information, to output a text information feature required for streaming synthetic audio.

[0068] Step S43: Performing speech decoding processing on the text information feature, to output a time domain signal or a spectral feature of a streaming speech.

[0069] It should be noted that embodiments of the streaming speech synthesis method for streaming input and the streaming speech synthesis system for streaming input provided in the foregoing examples have the same concept. For a detailed implementation process, reference is made to the embodiments of the streaming speech synthesis system, and the details will not be described herein again.

[0070] The streaming speech synthesis method for streaming input provided in the embodiments of the present disclosure may be implemented on a terminal side or a server side. As for the hardware structure of a streaming speech synthesis terminal for streaming input, Fig. 5 shows a schematic diagram of an optional hardware structure of an electronic terminal 500 according to an embodiment of the present disclosure. The terminal 500 may be a mobile phone, a computer device, a tablet device, a personal digital processing device, a factory background processing device, and the like. The electronic device 500 comprises: at least one processor 501, a memory 502, at least one network interface 504, and a user interface 506. All the components in the apparatus are coupled together by using a bus system 505. It may be understood that the bus system 505 is configured to implement connection and communication between these components. In addition to a data bus, the bus system 505 further comprises a power bus, a control bus, and a status signal bus. However, for the sake of clarity, various buses are marked as a bus system in FIG. 5.

[0071] The user interface 506 may comprise a display, a keyboard, a mouse, a trackball, a click wheel, a key, a button, a touchpad, or a touch screen.

[0072] It may be understood that, the memory 502 may be a volatile memory or a non-volatile memory, or may comprise both the volatile memory and the non-volatile memory. The non-volatile memory may be a read-only memory (ROM), or a programmable read-only memory (PROM), which is used as an external cache. According to exemplary but not limited descriptions, many forms of RAMs are available, for example, a static random access memory (SRAM), a synchronous static random access memory (SSRAM), and the like. The memory according to the embodiments of the present disclosure is intended to comprise, but not limited to, these and any other suitable types of memories.

[0073] The memory 502 in the embodiments of the present disclosure is configured to store various types of data to support the operation of the electronic terminal 500. Examples of the data comprise: any executable program operating on the electronic terminal 500, for example, an operating system 5021 and an application 5022. The operating system 5021 comprises various system programs, for example, a framework layer, a core library layer, a driver layer, and the like, which are used for implementing various basic services and processing hardware-based tasks. The application 5022 may comprise various applications, for example, a media player and a browser, which are used for implementing various application services. The streaming speech synthesis method for streaming input provided in the embodiments of the present invention may be included in the application 5022.

[0074] The method disclosed in the foregoing embodiment of the present disclosure may be applied to the processor 501, or implemented by the processor 501. The processor 501 may be an integrated circuit chip having a signal processing capability. During actual use, the steps of the method may be completed through an integrated logic circuit of hardware or an instruction in the form of software in the processor 501. The processor 501 may be a general-purpose processor, a digital signal processor (DSP), a programmable logic device, a discrete gate or transistor logic device, or discrete hardware component, or the like. The processor 501 can implement or perform the methods, steps, and logic block diagrams disclosed in the embodiments

of the present disclosure. The processor 501 may be a microprocessor, any conventional processor, or the like. Optimizing the steps in the method provided in the embodiments of the present disclosure in combination with accessory may be directly embodied as executing and completing the method by a hardware decoding processor or by a combination of hardware and software modules in the decoding processor. The software module may be located in a storage medium, wherein the storage medium is located in the memory. The processor reads the information in the memory and completes the steps of the foregoing method in combination with the hardware thereof.

[0075] In an exemplary embodiment, the electronic terminal 500 may be implemented by one or more application-specific integrated circuits (ASIC), a DSP, a programmable logic device (PLD), or a complex programmable logic device (CPLD) to perform the foregoing method.

[0076] A person of ordinary skill in the art may understand that all or some of the steps of the method in embodiments may be implemented by a computer program relevant hardware. The computer program may be stored in a computer-readable storage medium. When the program is executed, the steps of the foregoing method in embodiments are performed. The storage medium refers to a medium that can store program code, such as a ROM, a RAM, a magnetic disk, or an optical disc.

[0077] In the embodiments provided in the present disclosure, the computer readable and writable storage medium may comprise a read-only memory, a random access memory, EEPROM, CD-ROM or other optical disk storage apparatus, a magnetic disk storage apparatus or other magnetic storage devices, a flash memory, a U disk, or a mobile hard disk, or any other medium that stores desired program code in the form of instructions or data structures and may be accessed by a computer. In addition, any connection may be appropriately referred to as a computer-readable medium. For example, if the instructions are sent from a website, a server, or other remote sources by using a coaxial cable, a fiber optic cable, a twisted pair, a digital

subscriber line (DSL), or wireless technologies such as infrared, radio, and microwave, the coaxial cable, the fiber optic cable, the twisted pair, the DSL, or the wireless technologies such as the infrared, the radio, and the microwave are included in the definition of the medium. However, it should be understood that the computer readable and writable storage medium and data storage medium do not comprise a connection, a carrier wave, a signal, or other transitory mediums, and are instead intended for a non-transitory and tangible storage medium. The magnetic disk and the optical disc as used in the present disclosure comprise a compact disc (CD), a laser disc, an optical disc, a digital versatile disc (DVD), a floppy disk, and a Blu-ray disc, where the magnetic disk usually copy data magnetically, while the optical disc copies data optically by using a laser.

[0078] In summary, the present disclosure provides a streaming speech synthesis method and apparatus for streaming input, a terminal, and a medium. The present disclosure may dynamically adjust the size and duration of output audio according to the overall real-time factor of streaming speech synthesis and an input rate of streaming text, and obtain the best speech synthesis effect while ensuring that the streaming speech synthesis is not interrupted to the greatest extent. Therefore, this application effectively overcomes various defects in the related art, and has a high value in industrial use.

[0079] The foregoing embodiments only illustrate the principles and effects of the present disclosure, and are not used to limit the present disclosure. Anyone familiar with the technology may modify or change the foregoing embodiments without departing from the spirit and scope of the present disclosure. Therefore, all equivalent modifications or changes made by a person with ordinary knowledge of the art without departing from the spirit and technical ideas disclosed in the present disclosure shall still be covered by claims of the present disclosure.

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;

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 the streaming text dynamic output module dynamically adjusts 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.

3. The streaming speech synthesis system for streaming input according to claim 2, 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.

4. The streaming speech synthesis system for streaming input according to claim 2, 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:

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

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

$t_{\max_waiting}$ 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.

5. The streaming speech synthesis system for streaming input according to claim 2, 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.

6. 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.

7. 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.

8. 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;

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.

9. 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 8.

10. 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 8.