TEXT PROCESSING METHOD, MODEL TRAINING METHOD AND RELATED APPARATUSES

TECHNICAL FIELD

[0001] The present disclosure relates to the field of audio processing technologies, and in particular, to speech processing methods, model training methods, apparatuses, devices, and a storage medium.

BACKGROUND

[0002] Text-to-Speech (TTS) aims to produce natural and intelligible speech from text input. With the advent of deep learning methods, neural network-based TTS systems have significantly enhanced the quality of synthesized speech. TTS has become a pivotal technology in conversational artificial intelligence and human-machine interaction.

[0003] This background information is provided as it may be relevant to the present disclosure. No admission is necessarily intended, nor should be construed, that any of the preceding information constitutes prior art against the present disclosure.

SUMMARY

[0004] In a first aspect, an embodiment of the present disclosure provides a text processing method, the method includes:

obtaining a first phonetic representation of reference acoustic information and a phoneme sequence of text information;

processing, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

processing, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to

obtain second acoustic features of the text information; and

obtaining speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0005] In a second aspect, an embodiment of the present disclosure provides a model training method, including:

acquiring a training sample set, where the training sample set includes a plurality of acoustic samples;

for an acoustic sample in the training sample set:

inputting sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

inputting a sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder to obtain a prediction result;

comparing the prediction result and the first phonetic representation of the sample reference acoustic information to obtain a first loss;

updating at least one parameter of the autoregressive LM decoder based on first losses corresponding to the plurality of acoustic samples.

[0006] In a third aspect, an embodiment of the present disclosure provides a model training method, including:

acquiring a training sample set, where the training sample set includes a plurality of acoustic samples;

for an acoustic sample in the training sample set:

inputting sample reference acoustic information corresponding to the acoustic sample into a pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

inputting a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample into a pre-trained autoregressive

language model (LM) decoder to obtain a second phonetic representation corresponding to the acoustic sample;

inputting the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample into a non-autoregressive LM decoder to obtain a prediction result;

comparing the prediction result and the first acoustic features of the sample reference acoustic information to obtain a second loss;

updating at least one parameter of the non-autoregressive LM decoder based on second losses corresponding to the plurality of acoustic samples.

[0007] In a fourth aspect, an embodiment of the present disclosure provides a model training method, including:

acquiring a training sample set, where the training sample set includes a plurality of acoustic samples;

for an acoustic sample in the training sample set:

inputting sample text information corresponding to the acoustic sample into a pre-trained phoneme generator of a speech synthesis model to obtain a sample phoneme sequence corresponding to the acoustic sample, where the pre-trained phoneme generator is used for phoneme generation based on text information;

inputting sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder of the speech synthesis model to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

inputting the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder of the speech synthesis model to obtain a second phonetic representation of the sample text information;

inputting the sample reference acoustic information into a pre-trained acoustic encoder of the speech synthesis model to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

inputting the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information into a non-autoregressive LM decoder of the speech synthesis model to obtain second acoustic features of the sample text information;

inputting the second acoustic features of the sample text information into a pre-trained acoustic decoder of the speech synthesis model to obtain a speech synthesis result, where the pre-trained acoustic decoder is used for audio decoding of acoustic features; and

comparing the speech synthesis result and the acoustic sample to obtain a third loss;

updating at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder based on third losses corresponding to the plurality of acoustic samples.

[0008] In a fifth aspect, an embodiment of the present disclosure provides a text processing method, including:

extracting, by a terminal device, text information from an application installed on the terminal device;

obtaining, by the terminal device, a first phonetic representation of reference acoustic information and a phoneme sequence of the text information;

processing, by the terminal device in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

processing, by the terminal device in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

obtaining, by the terminal device, speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0009] In a sixth aspect, an embodiment of the present disclosure provides a text processing method, including:

extracting, by a terminal device, text information from an application installed on the terminal device;

sending, by the terminal device, the text information to a server;

receiving, by the terminal device, speech information corresponding to the text information sent by the server, where the speech information is obtained by the server based on second acoustic features of the text information, the second acoustic features of the text information are obtained by processing a second phonetic representation of the text information, a phoneme sequence of the text information and first acoustic features of reference acoustic information in a non-autoregressive manner, and the second phonetic representation of the text information is obtained by processing a first phonetic representation of the reference acoustic information and the phoneme sequence in an autoregressive manner.

[0010] In a seventh aspect, an embodiment of the present disclosure provides a text processing apparatus, the apparatus includes:

a first obtaining model, configured to obtain a first phonetic representation of reference acoustic information and a phoneme sequence of text information;

a first processing module, configured to process, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

a second processing module, configured to process, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

a second obtaining model, configured to obtain speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0011] In an eighth aspect, an embodiment of the present disclosure provides a model training apparatus, the apparatus includes:

an acquiring module, configured to acquire a training sample set, where the training sample set includes a plurality of acoustic samples;

an inputting module, configured to:

for an acoustic sample in the training sample set, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder to obtain a first

phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

input a sample phoneme sequence corresponding to the acoustic sample into an autoregressive language mode (LM) decoder to obtain a prediction result;

a comparing module, configured to, for an acoustic sample in the training sample set, compare the prediction result and the first phonetic representation of the sample reference acoustic information to obtain a first loss;

an updating module, configured to update at least one parameter of the autoregressive LM decoder based on first losses corresponding to the plurality of acoustic samples.

[0012] In a ninth aspect, an embodiment of the present disclosure provides a model training apparatus, the apparatus includes:

an acquiring module, configured to acquire a training sample set, where the training sample set includes a plurality of acoustic samples;

an inputting module, configured to:

for an acoustic sample in the training sample set, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

input a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample into a pre-trained autoregressive language mode (LM) decoder to obtain a second phonetic representation corresponding to the acoustic sample;

input the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample into a non-autoregressive LM decoder to obtain a prediction result;

a comparing model, configured to, for an acoustic sample in the training sample set, compare the prediction result and the first acoustic features of the sample reference acoustic information to obtain a second loss;

an updating module, configured to update at least one parameter of the non-autoregressive

LM decoder based on second losses corresponding to the plurality of acoustic samples.

[0013] In a tenth aspect, an embodiment of the present disclosure provides a model training apparatus, the apparatus includes:

an acquiring module, configured to acquire a training sample set, where the training sample set includes a plurality of acoustic samples;

an inputting module, configured to:

for an acoustic sample in the training sample set, input sample text information corresponding to the acoustic sample into a pre-trained phoneme generator of a speech synthesis model to obtain a sample phoneme sequence corresponding to the acoustic sample, where the pre-trained phoneme generator is used for phoneme generation based on text information;

input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder of the speech synthesis model to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

input the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder of the speech synthesis model to obtain a second phonetic representation of the sample text information;

input the sample reference acoustic information into a pre-trained acoustic encoder of the speech synthesis model to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

input the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information into a non-autoregressive LM decoder of the speech synthesis model to obtain second acoustic features of the sample text information;

input the second acoustic features of the sample text information into a pre-trained acoustic decoder of the speech synthesis model to obtain a speech synthesis result, where the pre-trained acoustic decoder is used for audio decoding of acoustic features; and

a comparing model, configured to, for an acoustic sample in the training sample set, compare the speech synthesis result and the acoustic sample to obtain a third loss;

an updating module, configured to update at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder based on third losses corresponding to the plurality of acoustic samples.

[0014] In an eleventh aspect, an embodiment of the present disclosure provides a text processing apparatus, the apparatus includes:

an extracting module, configured to extract text information from an application installed on the terminal device;

an obtaining module, configured to obtain a first phonetic representation of reference acoustic information and a phoneme sequence of the text information;

a processing module, configured to process, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

the processing module is further configured to process, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

the obtaining module is further configured to obtain speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0015] In a twelfth aspect, an embodiment of the present disclosure provides a text processing apparatus, the apparatus includes:

an extracting module, configured to extract text information from an application installed on the terminal device;

a transceiving module, configured to send the text information to a server, and receive speech information corresponding to the text information sent by the server;

where the speech information is obtained by the server based on second acoustic features of the text information, the second acoustic features of the text information are obtained by processing a second phonetic representation of the text information, a phoneme sequence of the text information and

first acoustic features of reference acoustic information in a non-autoregressive manner, and the second phonetic representation of the text information is obtained by processing a first phonetic representation of the reference acoustic information and the phoneme sequence in an autoregressive manner.

[0016] In a thirteenth aspect, an embodiment of the present disclosure provides an electronic device including a processor coupled to a memory in a communicative way via an interface where the memory stores a computer executable instruction, the processor executes the computer executable instruction stored in the memory for executing the text processing method according to the first aspect or any possible implementation of the first aspect, the model training method according to the second aspect or any possible implementation of the second aspect, the model training method according to the third aspect or any possible implementation of the third aspect, or the model training method according to the fourth aspect or any possible implementation of the fourth aspect.

[0017] In a fourteenth aspect, an embodiment of the present disclosure provides a computer-readable storage medium storing computer execution instructions which, when executed by a processor, causes the processor to execute the text processing method according to the first aspect or any possible implementation of the first aspect, the text processing method according to the fifth aspect or any possible implementation of the fifth aspect, the text processing method according to the sixth aspect or any possible implementation of the sixth aspect, the model training method according to the second aspect or any possible implementation of the second aspect, the model training method according to the third aspect or any possible implementation of the third aspect, or the model training method according to the fourth aspect or any possible implementation of the fourth aspect.

[0018] In a fifteenth aspect, an embodiment of the present disclosure provides a computer program product including program code for performing the text processing method according to the first aspect or any possible implementation of the first aspect, the text processing method according to the fifth aspect or any possible implementation of the fifth aspect, the text processing method according to the sixth aspect or any possible implementation of the sixth aspect, the model training method according to the second aspect or any possible implementation of the second aspect, the model training method according to the third aspect or any possible implementation of the third aspect, or the model training method according to the fourth aspect or any possible implementation of the fourth aspect.

[0019] In a sixteenth aspect, an embodiment of the present disclosure provides a computer program,

including computer execution instructions which, when executed by a processor, cause the processor to execute the text processing method according to the first aspect or any possible implementation of the first aspect, the text processing method according to the fifth aspect or any possible implementation of the fifth aspect, the text processing method according to the sixth aspect or any possible implementation of the sixth aspect, the model training method according to the second aspect or any possible implementation of the second aspect, the model training method according to the third aspect or any possible implementation of the third aspect, or the model training method according to the fourth aspect or any possible implementation of the fourth aspect.

[0020] In a seventeenth aspect, an embodiment of the present disclosure provides a chip, including an input/output (I/O) interface and a processor, where the processor is configured to call and run a computer program stored in a memory, to enable a device installing with the chip to perform the text processing method according to the first aspect or any possible implementation of the first aspect, the text processing method according to the fifth aspect or any possible implementation of the fifth aspect, the text processing method according to the sixth aspect or any possible implementation of the sixth aspect, the model training method according to the second aspect or any possible implementation of the second aspect, the model training method according to the third aspect or any possible implementation of the third aspect, or the model training method according to the fourth aspect or any possible implementation of the fourth aspect.

BRIEF DESCRIPTION OF DRAWINGS

[0021] FIG. 1A is a schematic diagram of a text processing system for which a text processing method according to one or more embodiments of the present disclosure is applied.

[0022] FIG. 1B shows an exemplary scenario to which a text processing method according to one or more example embodiments of the present disclosure is applied.

[0023] FIG. 1C shows another exemplary scenario to which a text processing method according to one or more example embodiments of the present disclosure is applied.

[0024] FIG. 2 shows a schematic flowchart of a text processing method according to one or more embodiments of the present disclosure.

[0025] FIG. 3 shows a schematic flowchart of a model training method according to one or more

embodiments of the present disclosure.

[0026] FIG. 4 shows a schematic flowchart of a model training method according to one or more embodiments of the present disclosure.

[0027] FIG. 5 shows a schematic flowchart of an exemplary text processing process according to one or more embodiments of the present disclosure.

[0028] FIG. 6 shows a schematic flowchart of an exemplary text processing process according to one or more embodiments of the present disclosure.

[0029] FIG. 7 shows a schematic flowchart of an exemplary text processing process according to one or more embodiments of the present disclosure.

[0030] FIG. 8 shows a schematic flowchart of an exemplary text processing process according to one or more embodiments of the present disclosure.

[0031] FIG. 9 shows a schematic flowchart of a model training process according to one or more embodiments of the present disclosure.

[0032] FIG. 10 shows a schematic structural diagram of a text processing apparatus according to one or more embodiments of the present disclosure.

[0033] FIG. 11 shows a schematic structural diagram of a text processing apparatus according to one or more embodiments of the present disclosure.

[0034] FIG. 12 shows a schematic structural diagram of another text processing apparatus according to one or more embodiments of the present disclosure.

[0035] FIG. 13 shows a schematic structural diagram of a model training apparatus according to one or more embodiments of the present disclosure.

[0036] FIG. 14 shows a schematic structural diagram of another model training apparatus according to one or more embodiments of the present disclosure.

[0037] FIG. 15 shows a schematic structural diagram of yet another model training apparatus according to one or more embodiments of the present disclosure.

[0038] FIG. 16 shows a structural diagram of an electronic device according to one or more embodiments of the present disclosure.

DESCRIPTION OF EMBODIMENTS

[0039] To describe the technical solutions in embodiments of the present invention or in the prior art more clearly, the following briefly introduces the accompanying drawings needed for describing the embodiments or the prior art.

[0040] In the following description, reference is made to the accompanying figures, which form part of the present disclosure, and which show, by way of illustration, specific aspects of embodiments of the present disclosure or specific aspects in which embodiments of the present disclosure may be used. It is understood that embodiments of the present disclosure may be used in other aspects and include structural or logical changes not depicted in the figures. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present disclosure is defined by the appended claims.

[0041] Text-to-Speech (TTS) models facilitate translation between a text and a speech. TTS systems, such as Tacotron and Transformer-based TTS, utilize sequence-to-sequence autoregressive models, significantly improving the naturalness of synthesized speeches (voices). However, output sequences of TTS systems are continuous and often longer than text inputs, making it challenging to model abundant speech variations from discrete text inputs. Incorrect predictions of text-acoustic alignment often result in robustness issues for synthesized speeches. Non- autoregressive TTS models, such as FastSpeech and Glow-TTS, introduce a separate duration predictor, which improves TTS robustness but at the cost of reduced diversity and expressiveness.

[0042] The primary challenge for the text-to-speech (TTS) systems is computational modeling of the one-to-many mapping function between text and speech. A TTS model in the TTS system is expected to synthesize highly natural speech across various styles for multiple speakers, including unseen speakers or styles (zero-shot TTS). In the related art, a speaker or reference encoder is primarily focused on capturing speaker characteristics and incorporating them into a TTS framework (i.e., TTS system). Since the speaker characteristics are closely entangled with speech content and style, speaker similarity and speech naturalness often decline in zero-shot scenarios. Moreover, the TTS framework requires high-quality recorded speech data for training, which constrains a diversity of the synthesized speech.

[0043] Recent language models have achieved great success in natural language generation and then applied to audio generation and speech synthesis. Language model (LM) based TTS systems formulate text-to-speech as a conditional language modeling task, where a speech waveform is quantized into

discrete tokens and is then modeled with a language model. That is, neural audio codecs facilitate the integration of speech units into a language model by quantizing speech waveforms into highly compressed audio codecs. TTS is then formulated as a task of conditional language modeling rather than as traditional signal regression. Compared to conventional neural TTS systems, the language model-based TTS systems have demonstrated strong in-context learning capabilities on large and diverse speech data, and can synthesize highly natural speech. Despite their advancements, these systems often face issues with robustness and suffer from word deletion, repetition, and mispronunciation, which makes them less ideal for real-life applications.

[0044] For example, VALL-E, as one successful example of the language model-based TTS, has demonstrated scalability and the capability for in-context learning. It employs a two-stage training process: modeling a first layer of audio codecs autoregressively and predicting the following layers in parallel in a non-autoregressive manner. Specifically, in VALL-E, zero-shot TTS is regarded as a conditional language modeling task. Given phoneme sequences and an acoustic prompt, the TTS model is trained to generate discrete acoustic codecs. These codecs capture all the necessary acoustic details for synthesizing a speech waveform, which not only represent the linguistic content of the phoneme sequence but also reflect the speaker's characteristics as indicated by the acoustic prompt. Encodec is often used to quantize acoustic prompt into 8 layers of discrete acoustic codecs. The first layer of codecs is predicted autoregressively, while the remaining layers are generated in parallel in the non-autoregressive manner, based on the first layer of codecs.

[0045] However, given the paramount importance of the first quantizer in the reconstruction of the speech waveform, in the existing VALL-E, the autoregressive model is required to process an excessive amount of information. Consequently, this could lead to an accumulation of errors and result in the above robustness issues in synthesized speech. Although VALL-E can be trained on large and diverse multi-speaker datasets and clone the voices of speakers from just a short acoustic prompt, as described above, VALL-E inherits synthesis robustness issues, such as word mispronunciation, deletion, and repetition, from autoregressive language models.

[0046] In fact, one notable challenge in the language model-based TTS systems lies in their autoregressive approach. For instance, text tokens are mapped to speech tokens in an autoregressive manner. However, accurately modeling the speech tokens is difficult due to the presence of both verbal

and non-verbal information in the speech signal. Furthermore, additional challenges arise from confounding factors such as environmental noises, which are encoded into the speech tokens during quantization. False predictions of the speech tokens can accumulate, leading to robustness issues in synthesized speech.

[0047] Based on the above observation, the present disclosure introduces a text processing method, aiming to facilitate a more robust language modeling for speech units, to solve the above problems. As will be described later, in the embodiments of the present disclosure, the tasks to be completed in autoregressive and non-autoregressive manners are reassigned, allowing the phonetic prediction to be done in the autoregressive manner, and leaving the refinement of acoustic details to the subsequent non-autoregressive stage. In this way, the autoregressive model would be simply responsible for dealing with language modelling issue, thereby reducing the probability of error propagation from the autoregressive model to the subsequent non-autoregressive model.

[0048] FIG. 1A is a schematic diagram of a text processing system for which a text processing method according to one or more embodiments of the present disclosure is applied. The text processing system 100 may include a terminal device 101, and may further include a (cloud) server 102 as an optional choice. The text processing system 100 may correspond to various scenarios where text needs to be synthesized into speech. For example, the various scenarios can include at least one of the followings: an audiobook reading scenario, a voice dialogue system, an intelligent voice assistant such as Siri, a telephone information query system, auxiliary applications such as a car navigation and an audio e-book, a language learning system, real-time information broadcasting systems at an airport, a station, etc., an information acquisition and communication scenario for people with visual or voice impairments, etc. At the same time, different application scenarios also put forward higher requirements for various indicators of synthesized speech, including naturalness, intelligibility, sound quality, emotional style, and control.

[0049] In a possible implementation, the text processing system 100 may simply include the terminal device 101, and the terminal device 101 is deployed with a speech synthesis model that provides a text processing function for converting text information into speech information. The speech synthesis model can complete the text-to-speech task locally at the terminal device 101 side. In a possible implementation, the speech synthesis model may be deployed by the server 102 connected to the terminal device 101 in a communicative way (e.g., via a wired or wireless network). For example, the server 102 may construct

the speech synthesis model and deploy the speech synthesis model to the terminal device 101, and then the speech synthesis model deployed in the terminal device 101 may be used for converting text information into speech information. In this case, the text processing system 100 may be regarded as further including the server 102. For example, an application may be installed on the terminal device 101, and can provide a text processing function by calling the speech synthesis model to convert text information into speech information, the application providing the text processing function can be any application capable of converting text into speech, e.g., it could be an audiobook reading application, the application may also be revoked by another application which needs a text-to-speech service, the type and running manner of the application is not limited in the embodiments of the present disclosure. In a possible implementation, the server 102 is a background server of the application capable of deploying text processing functions in the terminal device 101. As shown in FIG. 1B, upon receiving from a user an instruction which requests a text-to-speech conversion, the terminal device 101 can run the application and, with the speech synthesis model deployed therein, complete the conversion from text to speech, so that the response is provided to the user in a form of voice.

[0050] In a possible implementation, the text processing system 100 also includes both the terminal device 101 and the server 102 connected to the terminal device 101 via a wired or wireless network, but different from the case shown in FIG. 1B, in this implementation, the speech synthesis model would be deployed in the server 102, so the server 102 may convert, in response to an instruction text information from the terminal device 101, into speech information, and feedback the converted speech information corresponding to the text information to the terminal device 101. This could be applied in some on-line communication scenarios. For example, a chatGPT application is installed on the terminal device 101, the chatGPT application receives text information input by a user and transmits the same to the server 102, then the server 102 generates speech information corresponding to the text by the speech synthesis model deployed within the server 102, and transmits the speech information to the terminal device 101, so the response would be presented to the user in a form of voice. As shown in FIG. 1C, which is an exemplary scenario to which a text processing method according to one or more example embodiments of the present disclosure is applied. The terminal device 101 receives from a user an instruction which requests converting a target text into a target speech, transmits the target text (which is also referred to as text information in context) to be converted into a target speech (which is also referred to as speech

information in context) to the server 102, and the server 102 converts the target text into the target speech by a speech synthesis model deployed therein, and transmits the target speech to the terminal device 101, so that the response is provided to the user a form of voice.

[0051] The terminal device mentioned above can include, but is not limited to, a tablet computer, a mobile phone (such as a folding screen mobile phone, a large screen mobile phone, etc.), a wearable device, a vehicle-mounted device, an augmented reality (AR)/virtual reality (VR) equipment, a laptop computer, a ultra-mobile personal computer (UMPC), a netbook, a personal digital assistant (PDA), a smart TV, a smart screen, a HD TV, a 4K TV, a smart speaker, a smart projector and other Internet of Things (IOT) devices. The server can be an independent physical server, a server cluster or a distributed system composed of multiple physical servers, or can be a cloud server that provides basic cloud computing services such as cloud services, cloud databases, cloud computing, cloud functions, cloud storage, network services, cloud communication, middleware services, domain name services, security services, content distribution networks, and big data and artificial intelligence platforms. The specific types of the terminal device and the server are not limited in the present disclosure.

[0052] According to the embodiments of the present disclosure, the operation process of the above speech synthesis model includes: obtaining a first phonetic representation of reference acoustic information and a phoneme sequence of text information; processing, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information; processing, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and obtaining speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0053] The technical solution of the present disclosure will be described in detail below through specific embodiments. In the following, the inference and training processing of the speech synthesis model for implementing the text processing method according to the embodiments of the present disclosure will be illustrated with reference to FIG. 2 and FIG. 5 to FIG. 8 (inference processes) and FIG. 9 (training process).

[0054] FIG. 2 is a schematic flowchart of a text processing method according to one or more

embodiments of the present disclosure. As shown in FIG. 2, the method includes the following steps. The method may be applied in the scenarios shown in FIG. 1B and 1C, and may also be applied in other scenarios a text-speech conversion is necessary. Illustratively, the text processing method of the present disclosure can be implemented by a text processing apparatus provided by the present disclosure, and the text processing apparatus can be implemented by any software and/or hardware. For example, the text processing apparatus can be the speech synthesis model deployed in the terminal device or the server mentioned above with reference to FIG. 1A to FIG. 1C.

[0055] S201, obtain a first phonetic representation of reference acoustic information and a phoneme sequence of text information.

[0056] The reference acoustic information, which can also be referred to as acoustic prompt, is used for providing acoustic characteristic for synthesized speech information of the (to-be-converted) text information. In a possible implementation, the reference acoustic information is a piece of reference speech of a speaker. The identity of the speaker would be known or unknown, which is not limited in the embodiments of the present disclosure. The duration of the reference speech would be several seconds, for example, a 3-second speech, and the reference acoustic information can be pre-stored in the text processing apparatus or loaded from the network, which is not limited in the embodiments of the present disclosure.

[0057] In a possible implementation of the present disclosure, the step S201 of obtaining the first phonetic representation of the reference acoustic information includes: inputting the reference acoustic information into a pre-trained phonetic encoder to extract the first phonetic representation, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information. In a possible implementation, the first phonetic representation of the reference acoustic information can be phonetic information of the reference acoustic information, for example, self-supervised learning (SSL) tokens or phonetic posterior grams (PPG) features. When the first phonetic representation is the SSL tokens of the reference acoustic information, the pre-trained phonetic encoder can be a SSL encoder, when the first phonetic representation is the PPG features, the pre-trained phonetic encoder can be a PPG encoder, and the pre-trained phonetic encoder can be trained based on a set of sample reference acoustic information.

[0058] Since the pre-trained phonetic encoder is trained on large and diverse datasets of speech,

exposed to a wide range of phonetic variations, including different accents and speaking styles, so that the pre-trained phonetic encoder can learn robust phonetic representations across various speakers and context, and can extract the first phonetic representation effectively.

[0059] In a possible implementation of the present disclosure, the step S201 of obtaining the phoneme sequence of text information includes: inputting the text information into a grapheme-to-phoneme (G2P) model (which is also referred to as pre-trained phoneme generator below) to obtain the phoneme sequence corresponding the text information. In a possible implementation, the G2P model is used for phoneme generation based on text information and is pre-trained based on a set of sample text information.

[0060] In a possible implementation of the present disclosure, after obtaining the first phonetic representation of the reference acoustic information and the phoneme sequence, the method further includes the following steps.

[0061] S202, process, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information.

[0062] After obtaining the first phonetic representation of the reference acoustic information, the first phonetic representation of the reference acoustic information can be processed in an autoregressive manner, e.g., using an autoregressive decoder. In this way, it is possible to extract phonetic information from the to-be-converted text information by virtue of autoregressive processing and leave the acoustic information for later non-autoregressive processing.

[0063] In a possible implementation, the first phonetic representation of the reference acoustic information includes self-supervised learning (SSL) tokens of the reference acoustic information, and the second phonetic representation of the text information includes SSL tokens of the text information. In this case, the pre-trained phonetic encoder for extracting the first phonetic representation from the reference acoustic information can be a SSL encoder.

[0064] In a possible implementation, the first phonetic representation of the reference acoustic information includes phonetic posterior grams (PPG) features of the reference acoustic information, and the second phonetic representation of the text information includes PPG features of the text information. In this case, the pre-trained phonetic encoder for extracting the first phonetic representation from the

reference acoustic information can be a PPG encoder.

[0065] Based on the above, there may be different pre-trained phonetic encoders for generating different first phonetic representations of the reference acoustic information, a pre-trained phonetic encoder can be selected according to actual applications, and the flexibility of the speech synthesis model can be improved.

[0066] S203, process, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information.

[0067] After obtaining the second phonetic representation of the text information, the second phonetic representation of the text information can be processed in a non-autoregressive manner, e.g., via a non-autoregressive decoder. In a possible implementation of the present disclosure, the first acoustic features (which may also be referred to as first acoustic encoding vectors) are first acoustic tokens of the reference acoustic information, and the first acoustic tokens include fine-grained acoustic details of the reference acoustic information; where the second acoustic features (which may also be referred to as second acoustic encoding vectors) are second acoustic tokens of the text information, and the second acoustic tokens include fine-grained acoustic details for reconstructing the speech information corresponding to the text information. In a possible implementation, a dimension of a vector/embedding representing the first acoustic features is as same as that of a vector/embedding representing the second acoustic features. The fine-grained acoustic details may be non-verbal information such as a speaker's identity, a rhythm of the speaker, a tone/pitch of the speaker, etc. Here the first acoustic tokens correspond to the reference acoustic information, so the verbal information contained in the first acoustic tokens may be text information corresponding to the reference acoustic information, but the verbal information contained in the second acoustic tokens may be the text information obtained in step S201. The first and second acoustic tokens may share same non-verbal information.

[0068] From the perspective of the input, a size/length of the first phonetic representation (which would be a 1-dimension vector) and a size/length of the phoneme sequence (which would be a 1-dimension vector) would be the same or different, in the case where they are different, alignment would be possible in the autoregressive processing of the first phonetic representation of the reference acoustic information and the phoneme sequence.

[0069] From the perspective of the output, the first and second acoustic features are normally represented as 2-dimension matrices, e.g., each being represented as 5 or 8 layer codecs, so a size/length of the second phonetic representation (which would be a 1-dimension vector) would be as same as a size/length of each codec representing the first/second acoustic features, and the size/length of each codec representing the first/second acoustic features would be dependent on the size of the acoustic prompt.

[0070] In a possible implementation, the first acoustic tokens of the reference acoustic information can be obtained by inputting the reference acoustic information into a pre-trained acoustic encoder, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information. The pre-trained acoustic encoder can be trained based on sample acoustic tokens of a set of sample acoustic information. In this regard, the pre-trained acoustic encoder would be the encoder used in VALL-E.

[0071] Since the performance of the proposed speech synthesis is limited by the effectiveness of the pre-trained acoustic encoder, in terms of its reconstruction distortion, noise robustness, codebook size, and compression bitrate, so the performance of the proposed speech synthesis can be improved with the enhanced pre-trained acoustic encoder.

[0072] In a possible implementation of the present disclosure, the step S202 of processing the first phonetic representation of the reference acoustic information and the phoneme sequence in the autoregressive manner includes: processing, with a pre-trained autoregressive language model (LM) decoder, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain the second phonetic representation of the text information, where the pre-trained autoregressive LM decoder is trained based on a plurality of phonetic representations of sample reference acoustic information corresponding to a plurality of acoustic samples one-by-one and a plurality of sample phoneme sequences corresponding to the plurality of acoustic samples one-by-one. In a possible implementation, the multiple sample phoneme sequences are extracted from multiple pieces of sample text information corresponding to the multiple acoustic samples by the G2P model.

[0073] The sample reference acoustic information may also be referred to as a reference acoustic prompt, for which a phonetic representation could be obtained by using, e.g., a pre-trained phonetic encoder. The multiple acoustic samples could be different speeches of a speaker, and different acoustic samples correspond to different pieces of sample text information or same pieces of sample text information, and different acoustic samples correspond to different pieces of reference acoustic

information (an acoustic sample has its corresponding sample text information and corresponding reference acoustic information), so each acoustic sample has corresponding sample text information and corresponding reference acoustic information, and the sample text information may be processed, e.g., with a G2P model, to obtain a corresponding sample phoneme sequence.

[0074] With reference to FIG. 3, for training the autoregressive LM decoder, the process would be as follows:

S301, acquire a training sample set, where the training sample set includes a plurality of acoustic samples.

[0075] For an acoustic sample in the training sample set, the method further includes:

S302, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

S303, input a sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder to obtain a prediction result;

S304, compare the prediction result and the first phonetic representation of the sample reference acoustic information to obtain a first loss.

[0076] After steps S302 to S304 are performed on the plurality of acoustic samples, the method further includes:

S305, update at least one parameter of the autoregressive LM decoder based on first losses corresponding to the plurality of acoustic samples.

[0077] In the embodiment, a training sample set including multiple acoustic samples is acquired, and for an acoustic sample in the training sample set, sample reference acoustic information corresponding to the acoustic sample is input into a pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, a sample phoneme sequence corresponding to the acoustic sample is input into an autoregressive LM decoder to obtain a prediction result; the prediction result is compared with the first phonetic representation of the sample reference acoustic information to obtain a first loss; and then at least one parameter of the autoregressive LM decoder is updated based on the first loss corresponding to the acoustic sample. The acquiring of the training sample set would be performed

once for one or more training epochs, and in each training iteration of a training epoch, S302 to S305 would be performed. For each training iteration, there would one or more acoustic samples used for training, which is not limited herein. In this way, the trained autoregressive model would be simply responsible for dealing with language modelling issue, thereby reducing the probability of error propagation from the trained autoregressive model to the subsequent non-autoregressive model.

[0078] The pre-trained phonetic encoder may be, e.g., an SSL encoder or a PPG encoder, which is used to process sample reference acoustic information corresponding to the acoustic sample to output SSL tokens or PPG features as the first phonetic representation of the sample reference acoustic information.

[0079] The sample phoneme sequence may be obtained by inputting sample text information corresponding to the acoustic sample into a pre-trained phoneme generator, the pre-trained phoneme generator is used for phoneme generation based on text information, for example, it may be a G2P model.

[0080] The prediction result may be a predicted phonetic representation for the sample phoneme sequence, and the parameter(s) of the autoregressive LM decoder may be adjusted based on the first loss, which is obtained by comparing the prediction result and the first phonetic representation of the sample reference acoustic information.

[0081] Take the SSL token as an example, the autoregressive language model (i.e., the above autoregressive (AR) LM decoder) is trained to predict the SSL tokens from input phoneme sequences. Given a dataset D = $\{x_i, y_i\}$, where y_i is the i-th audio sample (a specific example of the above acoustic sample) and x_i is the corresponding phoneme sequence (a specific example of the above sample phoneme sequence), an autoregressive (AR) decoder-only language model is trained to predict SSL tokens $\tilde{s}$(a specific example of the above first phonetic representation) extracted from the acoustic prompt (sample reference acoustic information above):

$$\mathrm{p}(\mathrm{s}|\mathrm{x}, \tilde{s}; \theta_{AR}) = \prod_{t=0}^{T} p(s_t|s < t, \tilde{s}, X; \theta_{AR}) \quad (1)$$

where T is the sequence length of the SSL tokens $\tilde{s}$.

[0082] The AR model aims to predict SSL tokens from the input phoneme sequence and SSL tokens of the acoustic prompt. This stage can be referred as “Phoneme→ Phonetic”. Since the SSL tokens not only encode contextual cues but also prosodic variations, the AR model learns to map the phoneme sequence to a wide range of phonetic variations (i.e., predicted SSL tokens).

[0083] Since the pre-trained autoregressive language model (LM) decoder focuses solely on linguistic modeling to reduce error propagation that occurs in training pre-trained non-autoregressive LM decoder, thus improving the performance of the proposed speech synthesis.

[0084] In a possible implementation of the present disclosure, the step S203 of processing the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information in the non-autoregressive manner includes: processing, with a pre-trained non-autoregressive LM decoder, the second phonetic representation of the text information, the phoneme sequence and the first acoustic features of the reference acoustic information to obtain the second acoustic features of the text information, where the pre-trained non-autoregressive LM decoder is trained based on first acoustic features of plural pieces of sample reference acoustic information, a plurality of second phonetic representations corresponding to a plurality of acoustic samples one-by-one and a plurality of sample phoneme sequences corresponding to the plurality of acoustic samples one-by-one, and the plural pieces of sample reference acoustic information correspond to the plurality of acoustic samples one-by-one.

[0085] The sample reference acoustic information may also be referred to as a reference acoustic prompt, for which a phonetic representation could be obtained by using, e.g., a pre-trained phonetic encoder. The multiple acoustic samples could be different speeches of a speaker, and different acoustic samples correspond to different pieces of sample text information or same pieces of sample text information, and different acoustic samples correspond to different pieces of reference acoustic information (an acoustic sample has its corresponding sample text information and corresponding reference acoustic information), so each acoustic sample has corresponding sample text information and corresponding reference acoustic information, and the sample text information may be processed, e.g., with a G2P model, to obtain a corresponding sample phoneme sequence. Further, the corresponding sample phoneme sequence of each acoustic sample may also be processed, e.g., with the above pre-trained AR LM decoder, to obtain a second phonetic representation of the acoustic sample. Besides, the corresponding sample reference acoustic information may also be processed, e.g., with an acoustic encoder, to obtain its corresponding first acoustic features (e.g., codecs) which contain not only verbal information but also non-verbal information of the sample reference acoustic information.

[0086] With reference to FIG. 4, for training the non-autoregressive LM decoder, the process would

be as follows:

S401, acquire a training sample set, where the training sample set includes a plurality of acoustic samples.

[0087] For an acoustic sample in the training sample set, the method further includes:

S402, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

S403, input a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample into a pre-trained autoregressive language model (LM) decoder to obtain a second phonetic representation corresponding to the acoustic sample;

S404, input the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample into a non-autoregressive LM decoder to obtain a prediction result;

S405, compare the prediction result and the first acoustic features of the sample reference acoustic information to obtain a second loss.

[0088] After steps S402 to S405 are performed on each of the multiple acoustic samples, the method further includes:

S406, update at least one parameter of the non-autoregressive LM decoder based on second losses corresponding to the plurality of acoustic samples.

[0089] In the embodiment, a training sample set including multiple acoustic samples is acquired, and for an acoustic sample in the training sample set, sample reference acoustic information corresponding to the acoustic sample is input into a pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample are input into a pre-trained autoregressive LM decoder to obtain a second phonetic representation corresponding to the acoustic sample; the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample are input into a non-autoregressive LM decoder

to obtain a prediction result; the prediction result is compared with the first acoustic features of the sample reference acoustic information to obtain a second loss; and then at least one parameter of the non-autoregressive LM decoder is updated based on the second loss corresponding to the acoustic sample. The acquiring of the training sample set would be performed once for one or more training epochs, and in each training iteration of a training epoch, S402 to S406 would be performed. For each training iteration, there would one or more acoustic samples used for training, which is not limited herein. In this way, the trained non-autoregressive LM decoder is employed to recover fine-grained acoustic details of the sample reference acoustic information, thereby achieving better speech quality in terms of naturalness and zero-shot speaker similarity compared to the related art. It should be noted that there is no restriction with respect to the execution order of the steps, as long as the steps are executed logically. For example, step S402 and S403 may be executed simultaneously, or in sequence, which is not limited in the embodiments of the present disclosure.

[0090] The pre-trained acoustic encoder may be any encoding device capable of generating codecs for an input acoustic prompt, e.g., it could be the encoder used in VALL-E, which may be used to obtain corresponding first acoustic features (e.g., codecs) which contain not only verbal information but also non-verbal information of the sample reference acoustic information.

[0091] The first phonetic representation of the sample reference acoustic information may be obtained by inputting the sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder, the pre-trained phonetic encoder may be, e.g., an SSL encoder or a PPG encoder, which is used to output SSL tokens or PPG features as the first phonetic representation of the sample reference acoustic information.

[0092] The sample phoneme sequence may be obtained by inputting sample text information corresponding to the acoustic sample into a pre-trained phoneme generator, the pre-trained phoneme generator is used for phoneme generation based on text information, for example, it may be a G2P model.

[0093] The prediction result may be second acoustic features (SSL tokens or PPG features) for the sample phoneme sequence and the first phonetic representation of the sample reference acoustic information, and the parameter(s) of the non-autoregressive LM decoder may be adjusted based on the second loss, which is obtained by comparing the prediction result and the first acoustic features of the sample reference acoustic information.

[0094] For example, a non-autoregressive (NAR) language model is trained to recover fine-grained acoustic details from the phonetic variations, this stage is referred to as "Phonetic→Acoustic". The NAR model is conditioned on the phoneme sequence x (a specific example of the above sample phoneme sequence), the predicted SSL tokens *s* (a specific example of the above second phonetic representation), and the acoustic codecs $\tilde{C}$ (specific examples of first acoustic features above) extracted from the acoustic prompt (sample reference acoustic information above) by a pre-trained acoustic encode:

$$\mathrm{p}(\mathrm{C}|\mathrm{x}, \tilde{C}; \theta_{NAR}) = \prod_{j=1}^{8} p(C_{:,j}|C_{:,<j}, s, \tilde{C}, X; \theta_{NAR}) \quad (2)$$

[0095] The acoustic codecs C consist of 8 layers of discrete codec sequences, each containing fine-grained acoustic details that could be used to reconstruct a speech waveform. The amount of information decreases with each layer.

[0096] The pre-trained non-autoregressive LM decoder is employed to recover fine-grained acoustic details of the reference acoustic information, thereby achieving better speech quality in terms of naturalness and zero-shot speaker similarity compared to the related art.

[0097] In a possible implementation of the present disclosure, after obtaining the acoustic features of the text information, the method further includes step S204.

[0098] S204, obtain speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0099] After determining the second acoustic features of the text information, the speech information corresponding to the text information can be recovered. In a possible implementation, the second acoustic features of the text information is input into a pre-trained acoustic decoder to obtain the speech information corresponding to the text information and the reference acoustic information. The pre-trained acoustic decoder can be any decoding device capable of reconstructing speech information based on input codecs, e.g., it could be the decoder used in VALL-E.

[0100] In the above text processing method, a first phonetic representation of reference acoustic information and a phoneme sequence of text information are obtained, and processed in an autoregressive manner to obtain a second phonetic representation of the text information. In this way, it is possible to extract phonetic information from the to-be-converted text information by virtue of autoregressive processing and leave the acoustic information for later non-autoregressive processing. The second phonetic representation of the text information, the phoneme sequence and first acoustic features of the

reference acoustic information are processed in a non-autoregressive manner, to obtain second acoustic features of the text information. That is, the acoustic information for speech synthesis of the text information is obtained from the reference acoustic information, and processed in the non-autoregressive manner together with the second phonetic representation of the text information and the phoneme sequence. Based on the above, by processing verbal information related to the to-be-converted text information into the speech information in the autoregressive manner, and processing non-verbal information related to the text information in the non-autoregressive manner to recover fine-grained acoustic details of the reference acoustic information, the present disclosure focuses solely on linguistic modeling in the process of autoregressive manner, thereby reducing error propagation that occurs in the process of non-autoregressive manner, and effectively improving robustness of speech synthesis system and achieving better speech quality in terms of naturalness and zero-shot speaker similarity compared to the related art.

[0101] FIG. 5 shows a schematic flowchart of an exemplary text processing method according to one or more embodiments of the present disclosure, the method includes the following steps. As shown in FIG. 5, the speech synthesis model includes a pre-trained phoneme generator, a pre-trained phonetic encoder, a pre-trained autoregressive LM decoder, a pre-trained acoustic encoder, a pre-trained non-autoregressive LM decoder and a pre-trained acoustic decoder.

[0102] S501, input (to-be-converted) text information into the pre-trained phoneme generator to obtain a phoneme sequence.

[0103] S502, input reference acoustic information into the pre-trained phonetic encoder to obtain a first phonetic representation of the reference acoustic information.

[0104] S503, input the first phonetic representation and the phoneme sequence into the pre-trained autoregressive LM decoder to obtain a second phonetic representation of the text information.

[0105] S504, input the reference acoustic information into the pre-trained acoustic encoder to obtain first acoustic features of the reference acoustic information.

[0106] S505, input the second phonetic representation of the text information, the phoneme sequence and the first acoustic features of the reference acoustic information into the pre-trained non-autoregressive LM decoder to obtain second acoustic features of the text information.

[0107] S506, input the second acoustic features of the text information into the pre-trained acoustic

decoder to obtain the speech information corresponding to the text information.

[0108] It should be noted that there is no restriction with respect to the execution order of the steps, as long as the steps are executed logically. For example, step S501, S502, and S504 may be executed simultaneously, or in sequence, which is not limited in the embodiments of the present disclosure.

[0109] FIG. 6 shows a schematic flowchart of an exemplary text processing method according to one or more embodiments of the present disclosure, the method includes the following steps. The method may be applied in the scenario shown in FIG. 1B.

[0110] S601, a terminal device extracts text information from an application installed on the terminal device.

[0111] S602, the terminal device obtains a first phonetic representation of reference acoustic information and a phoneme sequence of the text information.

[0112] S603, the terminal device processes, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information.

[0113] S604, the terminal device processes, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information.

[0114] S605, the terminal device obtains speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0115] Here the application can be any application which needs a text-to-speech service, and the terminal device can extract, e.g., in response to an instruction from a user, text information and complete the text-to-speech conversion locally, e.g., by using a speech synthesis model deployed therein. For detailed description with reference to the terminal device and the server, reference may be made to the above description with reference to FIG. 1A, which is not repeated here for brevity.

[0116] By dealing with the text information extracted from the application locally, the text-to-speech synthesis can be done efficiently, thereby improving user experience, which is particularly useful for applications where bandwidth is limited.

[0117] FIG. 7 shows a schematic flowchart of an exemplary text processing method according to one or more embodiments of the present disclosure, the method includes the following steps. The method

may be applied in the scenario shown in FIG. 1C.

[0118] S701, a terminal device extracts text information from an application installed on the terminal device.

[0119] S702, the terminal device sends the text information to a server.

[0120] S703, the terminal device receives speech information corresponding to the text information sent by the server, where the speech information is obtained by the server based on second acoustic features of the text information, the second acoustic features of the text information are obtained by processing a second phonetic representation of the text information, a phoneme sequence of the text information and first acoustic features of reference acoustic information in a non-autoregressive manner, and the second phonetic representation of the text information is obtained by processing a first phonetic representation of the reference acoustic information and the phoneme sequence in an autoregressive manner.

[0121] Here the application can be any application which needs a text-to-speech service, and the terminal device can extract, e.g., in response to an instruction from a user, text information, send the text information to the server for text-to-speech conversion, and the server can perform such conversion, e.g., by using a speech synthesis model deployed therein. For detailed description with reference to the terminal device and the server, reference may be made to the above description with reference to FIG. 1A, which is not repeated here for brevity.

[0122] By interacting with the server and make the server fulfill the task of text-to-speech synthesis, power consumption can be realized at the terminal device side, which is particularly useful for applications where the terminal device has limited processing capabilities.

[0123] FIG. 8 shows a schematic flowchart of an exemplary text processing process according to one or more embodiments of the present disclosure.

[0124] As shown in FIG. 8, the overall diagram follows the pipeline of "Phoneme→Phonetic→Acoustic". Given phoneme sequence of a text (which is equivalent to phoneme sequence of text information above) and an acoustic prompt (which is also referred to as reference acoustic information above) as inputs, an autoregressive decoder (which is also referred to as a pre-trained autoregressive LM decoder above) first predicts SSL tokens (which is a specific example of the second phonetic representation of the text information above) that contain phonetic information. A non-

autoregressive decoder (which is also referred to as a pre-trained non-autoregressive LM decoder above) then predicts 8 layers of acoustic codecs (which is a specific example of the second acoustic features of the text information) that represent fine-grained acoustic details. Finally, from these predicted acoustic codecs, a pre-trained acoustic decoder synthesizes speech waveform corresponding to the text.

[0125] Rather than directly mapping phoneme sequence to acoustic tokens, the proposed approach maps phoneme sequence to phonetic tokens in an autoregressive manner, which is referred to as "phonetic enhanced". Drawing inspiration from the success of self-supervised learning (SSL) representations in automatic speech recognition, the present disclosure leverages SSL representations as phonetically rich targets for training of the autoregressive model, the SSL representations are obtained by processing the acoustic prompt by an SSL model. The phonetic richness of SSL features (SSL representations) arises from the models' training objective to capture the underlying structure of the audio signal. For instance, to accurately predict masked parts of speech, the SSL model requires a deep understanding of phonetics and the combinations of sounds in natural language. Moreover, these models (for example, SSL encoder, Grapheme-to-Phoneme model, autoregressive and non-autoregressive LM decoder shown in FIG. 8) are trained on large and diverse datasets of speech, exposing them to a wide range of phonetic variations, including different accents and speaking styles. This exposure allows them to learn robust phonetic representations across various speakers and context.

[0126] During inference shown in FIG. 8, an acoustic prompt is provided, which triggers the SSL encoder (which is a specific example of pre-trained phonetic encoder above) and the pre-trained acoustic encoder to generate SSL tokens and 8 layers of acoustic codecs, respectively. Subsequently, the AR LM decoder predicts SSL tokens of the text from the input phoneme sequence and the SSL tokens of the acoustic prompt. Then, the NAR LM decoder predicts acoustic codecs of the text using the predicted SSL tokens, the phoneme sequence and the acoustic codecs of the acoustic prompt. Finally, an acoustic decoder (which is also referred to as pre-trained acoustic decoder above) synthesizes the speech waveform based on the predicted acoustic codecs.

[0127] FIG. 9 shows a schematic flowchart of a model training method for training a speech synthesis model of the text processing method. As shown in FIG. 9, the speech synthesis model includes a pre-trained phoneme generator, a pre-trained phonetic encoder, an autoregressive language model (LM) decoder, a pre-trained acoustic encoder, a non-autoregressive LM decoder and a pre-trained acoustic

decoder. The model training method includes the following steps.

[0128] S901, acquire a training sample set, where the training sample set includes a plurality of acoustic samples.

[0129] For an acoustic sample in the training sample set, the method further includes:

[0130] S902, input sample text information corresponding to the acoustic sample into the pre-trained phoneme generator to obtain a sample phoneme sequence corresponding to the acoustic sample, where the pre-trained phoneme generator is used for phoneme generation based on text information.

[0131] S903, input sample reference acoustic information corresponding to the acoustic sample into the pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information.

[0132] S904, input the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample into the autoregressive LM decoder to obtain a second phonetic representation of the sample text information.

[0133] S905, input the sample reference acoustic information into the pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information.

[0134] S906, input the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information into the non-autoregressive LM decoder to obtain second acoustic features of the sample text information.

[0135] S907, input the second acoustic features of the sample text information into the pre-trained acoustic decoder to obtain a speech synthesis result, where the pre-trained acoustic decoder is used for audio decoding of acoustic features.

[0136] S908, compare the speech synthesis result and the acoustic sample to obtain a third loss;

[0137] S909, update at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder based on third losses corresponding to the plurality of acoustic samples.

[0138] In the embodiment, a training sample set including multiple acoustic samples is acquired, and

for an acoustic sample in the training sample set, sample text information corresponding to the acoustic sample is input into the pre-trained phoneme generator to obtain a sample phoneme sequence corresponding to the acoustic sample; sample reference acoustic information corresponding to the acoustic sample is input into the pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information; the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample are input into the autoregressive LM decoder to obtain a second phonetic representation of the sample text information; the sample reference acoustic information is input into the pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information; the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information are input into the non-autoregressive LM decoder to obtain second acoustic features of the sample text information; the second acoustic features of the sample text information are input into the pre-trained acoustic decoder to obtain a speech synthesis result; the speech synthesis result is compared with the acoustic sample to obtain a third loss; and then at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder is/are updated based on the third loss corresponding to the acoustic sample. The acquiring of the training sample set would be performed once for one or more training epochs, and in each training iteration of a training epoch, S902 to S909 would be performed. For each training iteration, there would one or more acoustic samples used for training, which is not limited herein. It should be noted that there is no restriction with respect to the execution order of the steps, as long as the steps are executed logically. For example, step S902, S903, and S905 may be executed simultaneously, or in sequence, which is not limited in the embodiments of the present disclosure.

[0139] Here sample text information used in different training iterations may be texts corresponding to different acoustic samples. The speech synthesis result would be a reconstructed speech signal corresponding to the sample text information used in each iteration.

[0140] In the training process, the autoregressive LM decoder focuses solely on linguistic modeling, so that reduce error propagation that occurs during training non-autoregressive LM decoder, and the non-autoregressive LM decoder recovers fine-grained acoustic details of the reference acoustic information based on the output of the autoregressive LM decoder, thereby effectively improving robustness of the

speech synthesis model and achieving better speech quality in terms of naturalness and zero-shot speaker similarity compared to the related art.

[0141] After obtaining both the trained autoregressive LM decoder and the trained non-autoregressive LM decoder, the trained speech synthesis model with improved robustness is obtained.

[0142] A set of experiments based on the text processing method will be described below, the experiments include the following stages.

[0143] (1) Experimental Setup

[0144] All experiments are performed on LibriTTS dataset, which consists of around 600 hours of speech data with transcriptions at a 24kHz sampling rate from 2,456 speakers. 'train-clean' and 'train-other' subsets in the LibriTTS dataset are merged to train the TTS framework (which refers to the speech synthesis model above) and evaluate its performance on 'test-clean' and 'test-other' subsets, respectively. Note that the 'test-other' subset contains more challenging recording conditions compared to the 'test-clean'. The speakers from both 'test-other' and 'test-clean' subsets are unseen during training.

[0145] The zero-shot performance of the TTS framework is evaluated on the test sets of the LibriTTS dataset, where the TTS framework is expected to synthesize speech mimicking the voice of a given speaker not seen during training. To set up the experiments, for each speaker in the test set, one speech sample is randomly chose from the same speaker to serve as an acoustic prompt, which is distinct from speech to be synthesized, and then the speech is synthesized based on transcriptions. All synthesized speech is within 3 to 10 seconds. All demos are publicly available online (see https://demos46.github.io/IS24/).

[0146] For comparison, VALL-E is chosen as a baseline TTS framework. VALL-E utilizes a similar decoder-only language model (LM) and predicts the first layer of acoustic codecs autoregressively, followed by predicting the other layers non-autoregressively. Additionally, the experiments compare different configurations using self-supervised learning (SSL) tokens. Specifically, the experiments compare Hubert andWavLM features, followed by K-means clustering with the cluster sizes of 500 and 1024 for both Hubert and WavLM.

[0147] (2) Implementation Details

[0148] The proposed TTS framework shares a similar architecture with VALL-E for comparison purpose, but it should be noted that in actual applications, the proposed TTS framework would be

configured in other ways. Both autoregressive and non-autoregressive models consist of a transformer architecture consisting of 12 layers, 16 attention heads, an embedding dimension of 1024, a feed-forward layer dimension of 4096 and a dropout rate of 0.1. A 24kHz Encodec is used as the acoustic encoder to compress a speech waveform (which is a specific example of reference acoustic information above) into 8 layers of discrete acoustic codes (which is a specific example of first acoustic features of the reference acoustic information above), while a decoder of the Encodec (which is a specific example of pre-trained acoustic decoder above) is utilized to recover 8 layers of discrete acoustic codes (which is a specific example of second acoustic features of text information above) to a speech waveform (which is a specific example of speech information corresponding to text information above). To extract SSL tokens (which is a specific example of first phonetic representation of reference acoustic information above) from the speech waveform (which is a specific example of reference acoustic information above), official versions of the Hubert and WavLM models are used. Features are obtained from the 6th Hubert layer and the 24th WavLM layer respectively, and cluster them using the K-means algorithm. Upsample manually the SSL tokens from 16kHz to 24kHz to align with acoustic codecs. The TTS framework is trained on 4 NVIDIA TESLA V100 64GB GPUs for 1M steps. The other training configurations remain the same as those in the unofficial release of VALL-E.

[0149] (3) Objective Evaluation

[0150] The experiments conduct objective evaluation to assess a robustness performance of the proposed TTS framework. Whisper is used to generate transcriptions from the synthesized speech and calculate word error rate (WER) and character error rate (CER) with the ground-truth transcriptions.

[0151] Objective results are shown in Table 1. Table 1 shows objective evaluation results to compare the robustness of the baseline VALL-E and the proposed TTS systems (frameworks) with different configurations using SSL tokens. ‘WER’ and ‘CER’ denote as ‘Word Error Rate’ and ‘Character Error Rate’ respectively, where a smaller value indicates a better system robustness. ‘K’ represents the K-means cluster size of the SSL tokens.

Table 1

Test Dataset	LibriTTS (test-clean)		LibriTTS (test-other)	
Objective Metrics (%)	WER ↓	CER↓	WER↓	CER↓

Baseline (VALL-E)		21.05	18.19	25.56	23.51
Proposed Phonetic Enhanced LM-based TTS	w/ Hubert (K=500)	8.03	7.18	13.39	11.58
	w/ Hubert (K=1024)	7.66	5.84	12.14	10.05
	w/ WavLM (K=500)	12.72	11.96	14.34	12.73
	w/ WavLM (K=1024)	6.12	5.44	7.92	6.11

[0152] As shown in Table 1, the proposed phonetic enhanced LM-based TTS framework consistently outperforms the baseline VALL-E on both 'test-clean' and 'test-other' on WER and CER metrics. Regarding different SSL configurations, it is observed that the proposed TTS framework with WavLM tokens clustered with a K-means cluster size of 1024 ('with WavLM (K = 1024)') achieves the best performance on both 'test-clean' and 'test-other', while WavLM with a cluster size of 500 ('with WavLM (K = 500)') performs slightly worse than the others. Compared to WavLM, the cluster size of Hubert features does not have a significant impact on the final results. However, their performance declines on the 'test-other', indicating that Hubert tokens are less noise-robust than WavLM tokens.

[0153] (4) Subjective Evaluation

[0154] The experiments further conduct two listening tests to assess the performance of the proposed TTS framework. In each test, 10 native English speakers participated, with each participant listening to a total of 200 synthesized speech samples (200 samples = 10 samples×5 systems×2 test sets×2 tests). The first listening test (referred as 'MOS') evaluates the naturalness of the synthesized speech samples (which are specific examples of speech information corresponding to text information above). Participants are asked to score each sample on a 5-point scale (5: Excellent, 4: Good, 3: Satisfactory, 2: Unsatisfactory, 1: Bad). The second listening test (referred as 'S-MOS') evaluates the speaker similarity with the reference speech. Participants are asked to score each sample on a 5-point scale (5: Very Similar, 4: Similar, 3: Neutral, 2: Dissimilar, 1: Very Dissimilar). The results are reported in Table 2.

[0155] Table 2 shows subjective evaluation results with 95% confidence interval to assess the naturalness ('MOS') and the speaker similarity ('S-MOS') of the baseline VALL-E and the proposed TTS systems with different configurations using SSL tokens.

Table 2

Test Dataset	LibriTTS (test-clean)	LibriTTS (test-other)

Subjective Metrics		MOS	S-MOS	MOS	S-MOS
Reference Speech		4.81±0.07	_*	4.39±0.11	_*
Baseline (VALL-E)		2.51±0.19	3.12±0.17	2.49±0.16	3.09±0.13
Proposed Phonetic Enhanced LM-based TTS	w/ Hubert (K=500)	3.58±0.13	3.89±0.13	3.43±0.13	3.35±0.14
	w/ Hubert (K=1024)	3.29±0.14	3.59±0.14	3.19±0.10	3.33±0.15
	w/ WavLM (K=500)	3.38±0.13	3.61±0.16	3.33±0.14	3.53±0.13
	w/ WavLM (K=1024)	3.54±0.13	3.94±0.13	3.53±0.10	3.55±0.15

[0156] Note that * S-MOS results for 'Reference Speech' shown in Table 2 are not reported because it serves as the comparison target in speaker similarity evaluation.

[0157] It is first observed that all of the proposed TTS frameworks significantly outperform the Baseline (VALL-E) in naturalness and speaker similarity on both test sets ('test-clean' and 'test- other'). This demonstrates that the proposed phonetic enhanced language modeling approach improves the robustness of the TTS system and contributes to better zero-shot performance for synthesized speech.

[0158] The experiments further investigate the impacts of different configurations of SSL tokens. From Table 2, It is observed that the proposed framework with Hubert features clustered with a size of 500 ('w/ Hubert (K=500)') outperforms that with a size of 1024 ('w/ Hubert (K=1024)') on the test set of LibriTTS (test- clean). Conversely, the proposed TTS framework with WavLM features clustered with a size of 1024 ('w/ WavLM (K=1024)') outperforms that with a size of 500 ('w/ WavLM (K=500)') on the same test set. Additionally, it is noted that the proposed ('w/ WavLM (K=1024)') achieves slightly better speaker similarity than the proposed ('w/ Hubert (K=500)'), while the naturalness slightly declines.

[0159] It is observed that all TTS frameworks suffers from a performance decline on the test set of LibriTTS (test-other), which contains more challenging recording conditions. Notebly, the performance of the proposed TTS frameworks with WavLM tokens presents a more noise-robust performance by achieving less performance decline than those with Hubert tokens.

[0160] Based on the above experiments, the major difference between the proposed TTS framework and the baseline VALL-E is the utilization of self-supervised representations as phonetic targets for autoregressive language modeling. Both objective and subjective evaluations reveal two key findings:

- The proposed TTS framework utilizing WavLM tokens demonstrates greater noise

robustness compared to those using HuBERT tokens;

- A larger cluster size in K-means clustering of WavLM tokens contributes to improved performance.

[0161] As a result, configurations using WavLM features with a larger cluster size outperform other setups. This advantage is largely credited to WavLM's pre-training strategy, which includes de-noising and handling distorted audio signals, thereby enhancing its ability to capture phonetic details under challenging conditions. WavLM leverages more detailed phonetic information provided by a larger cluster and better model the nuances in complex audio environments. However, increasing the cluster size in Hubert could potentially add complexity to non-autoregressive acoustic modeling and might lead to diminished performance.

[0162] The experiments demonstrate that the proposed method could effectively reduce error propagation in speech unit modeling and improve the robustness and naturalness of the synthesized speech. The experiments also compare different configurations of self-supervised representations and analyze their impacts on the synthesized results.

[0163] FIG. 10 shows a schematic structural diagram of a text processing apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 10, the text processing apparatus 1000 may include:

a first obtaining model 1001, configured to obtain a first phonetic representation of reference acoustic information and a phoneme sequence of text information;

a first processing module 1002, configured to process, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

a second processing module 1003, configured to process, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

a second obtaining model 1004, configured to obtain speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0164] In a possible implementation, the text processing apparatus 1000 includes a first inputting

module, configured to input the reference acoustic information into a pre-trained phonetic encoder to extract the first phonetic representation, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information.

[0165] In a possible implementation, the first phonetic representation of the reference acoustic information includes self-supervised learning (SSL) tokens of the reference acoustic information, and the second phonetic representation of the text information includes SSL tokens of the text information.

[0166] In a possible implementation, the first phonetic representation of the reference acoustic information includes phonetic posterior grams (PPG) features of the reference acoustic information, and the second phonetic representation of the text information includes PPG features of the text information.

[0167] In a possible implementation, the first processing module 1002 is further configured to process, with a pre-trained autoregressive language model (LM) decoder, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain the second phonetic representation of the text information, where the pre-trained autoregressive LM decoder is trained based on a plurality of phonetic representations of sample reference acoustic information corresponding to a plurality of acoustic samples one-by-one and a plurality of sample phoneme sequences corresponding to the plurality of acoustic samples one-by-one.

[0168] In a possible implementation, the second processing module 1003 is further configured to process, with a pre-trained non-autoregressive LM decoder, the second phonetic representation of the text information, the phoneme sequence and the first acoustic features of the reference acoustic information to obtain the second acoustic features of the text information, where the pre-trained non-autoregressive LM decoder is trained based on first acoustic features of plural pieces of sample reference acoustic information, a plurality of second phonetic representations corresponding to a plurality of acoustic samples one-by-one and a plurality of sample phoneme sequences corresponding to the plurality of acoustic samples one-by-one, and the plural pieces of sample reference acoustic information correspond to the plurality of acoustic samples one-by-one.

[0169] In a possible implementation, the first acoustic features are first acoustic tokens of the reference acoustic information, and the first acoustic tokens include fine-grained acoustic details of the reference acoustic information; where the second acoustic features are second acoustic tokens of the text information, and the second acoustic tokens include fine-grained acoustic details for reconstructing the

speech information corresponding to the text information.

[0170] In a possible implementation, the text processing apparatus 1000 includes a second inputting module, configured to input the reference acoustic information into a pre-trained acoustic encoder to obtain the first acoustic tokens of the reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information.

[0171] FIG. 11 shows a schematic structural diagram of a text processing apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 11, the text processing apparatus 1100 may include:

an extracting module 1101, configured to extract text information from an application installed on the terminal device;

an obtaining module 1102, configured to obtain a first phonetic representation of reference acoustic information and a phoneme sequence of the text information;

a processing module 1103, configured to process, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

the processing module 1103 is further configured to process, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

the obtaining module 1102 is further configured to obtain speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

[0172] FIG. 12 shows a schematic structural diagram of a text processing apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 12, the text processing apparatus 1200 may include:

an extracting module 1201, configured to extract text information from an application installed on the terminal device;

a transceiving module 1202, configured to send the text information to a server, and receive speech information corresponding to the text information sent by the server;

where the speech information is obtained by the server based on second acoustic features of the text information, the second acoustic features of the text information are obtained by processing a second phonetic representation of the text information, a phoneme sequence of the text information and first acoustic features of reference acoustic information in a non-autoregressive manner, and the second phonetic representation of the text information is obtained by processing a first phonetic representation of the reference acoustic information and the phoneme sequence in an autoregressive manner.

[0173] FIG. 13 shows a schematic structural diagram of a model training apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 13, the model training apparatus 1300 may include:

an acquiring module 1301, configured to acquire a training sample set, where the training sample set includes a plurality of acoustic samples;

an inputting module 1302, configured to:

for an acoustic sample in the training sample set, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

input a sample phoneme sequence corresponding to the acoustic sample into an autoregressive language mode (LM) decoder to obtain a prediction result;

a comparing module 1303, configured to, for an acoustic sample in the training sample set, compare the prediction result and the first phonetic representation of the sample reference acoustic information to obtain a first loss;

an updating module 1304, configured to update at least one parameter of the autoregressive LM decoder based on first losses corresponding to the plurality of acoustic samples.

[0174] FIG. 14 shows a schematic structural diagram of another model training apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 14, the model training apparatus 1400 may include:

an acquiring module 1401, configured to acquire a training sample set, where the training sample set includes a plurality of acoustic samples;

an inputting module 1402, configured to:

for an acoustic sample in the training sample set, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

input a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample into a pre-trained autoregressive language mode (LM) decoder to obtain a second phonetic representation corresponding to the acoustic sample;

input the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample into a non-autoregressive LM decoder to obtain a prediction result;

a comparing model 1403, configured to, for an acoustic sample in the training sample set, compare the prediction result and the first acoustic features of the sample reference acoustic information to obtain a second loss;

an updating module 1404, configured to update at least one parameter of the non-autoregressive LM decoder based on second losses corresponding to the plurality of acoustic samples.

[0175] FIG. 15 shows a schematic structural diagram of yet another model training apparatus according to one or more embodiments of the present disclosure. As shown in FIG. 15, the model training apparatus 1500 may include:

an acquiring module 1501, configured to acquire a training sample set, where the training sample set includes a plurality of acoustic samples;

an inputting module 1502, configured to:

for an acoustic sample in the training sample set, input sample text information corresponding to the acoustic sample into a pre-trained phoneme generator of a speech synthesis model to obtain a sample phoneme sequence corresponding to the acoustic sample, where the pre-trained phoneme generator is used for phoneme generation based on text information;

input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder of the speech synthesis model to obtain a first phonetic representation of the sample reference acoustic information, where the pre-trained phonetic

encoder is used for extracting a phonetic representation from acoustic information;

input the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder of the speech synthesis model to obtain a second phonetic representation of the sample text information;

input the sample reference acoustic information into a pre-trained acoustic encoder of the speech synthesis model to obtain first acoustic features of the sample reference acoustic information, where the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

input the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information into a non-autoregressive LM decoder of the speech synthesis model to obtain second acoustic features of the sample text information;

input the second acoustic features of the sample text information into a pre-trained acoustic decoder of the speech synthesis model to obtain a speech synthesis result, where the pre-trained acoustic decoder is used for audio decoding of acoustic features; and

a comparing model 1503, configured to, for an acoustic sample in the training sample set, compare the speech synthesis result and the acoustic sample to obtain a third loss;

an updating module 1504, configured to update at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder based on third losses corresponding to the plurality of acoustic samples.

[0176] It should be understood by a person skilled in the art that, the relevant description of the above modules in the possible implementations of the present disclosure may be understood with reference to the relevant description of the text processing methods and the model training methods in the possible implementations of the present disclosure. The technical effects achieved by the above apparatuses are similar as those achieved by the above corresponding method embodiments, which is not repeated herein.

[0177] FIG. 16 is a structural diagram of an electronic device according to one or more embodiments of the present disclosure, as shown in FIG. 16, the electronic device 1600 may include: a processor 1601 coupled to a memory 1602 in a communicative way via an interface 1603; where the memory 1602 stores

a computer executable instruction; the processor 1601 executes the computer executable instruction stored in the memory 1602 for executing any of the above text processing methods or any of the above model training methods. It should be noted that, the memory 1602 may be included or excluded from the electronic device, depending on actual needs.

[0178] In a possible implementation, the electronic device may include a transceiver, a processor, and a memory. The memory may be configured to store code, instructions, and the like executed by the processor.

[0179] It should be understood that the processor may be an integrated circuit chip and has a data processing capability. In an implementation process, steps of the foregoing method embodiments may be completed by using a hardware integrated logic circuit in the processor, or by using instructions in a form of software. The processor may be a general-purpose processor, a central processing unit (CPU), a graphics processing unit (GPU), a neural processing unit (NPU), a system on chip (SoC) or another programmable logic device, a discrete gate or transistor logic device, or a discrete hardware component. The processor may implement or perform the methods, the steps, and the logical block diagrams that are disclosed in the embodiments of the present application. The general-purpose processor may be a microprocessor, or the processor may be any conventional processor or the like. The steps of the methods disclosed with reference to the embodiments of the present application may be directly performed and completed by a hardware decoding processor, or may be performed and completed by using a combination of hardware in the decoding processor and a software module. The software module may be located in a mature storage medium in the art, such as a random-access memory, a flash memory, a read-only memory, a programmable read-only memory, an electrically erasable programmable memory, or a register. The storage medium is located in the memory, and the processor reads information in the memory and completes the steps of the foregoing methods in combination with hardware in the processor.

[0180] It may be understood that the memory in the embodiments of the present application may be a volatile memory or a non-volatile memory, or may include both a volatile memory and a non-volatile memory. The non-volatile memory may be a read-only memory (Read-Only Memory, ROM), a programmable read-only memory (Programmable ROM, PROM), an erasable programmable read-only memory (Erasable PROM, EPROM), an electrically erasable programmable read-only memory (Electrically EPROM, EEPROM), or a flash memory. The volatile memory may be a random-access

memory (Random Access Memory, RAM) and is used as an external cache. By way of example rather than limitation, many forms of RAMs may be used, and are, for example, a static random access memory (Static RAM, SRAM), a dynamic random access memory (Dynamic RAM, DRAM), a synchronous dynamic random access memory (Synchronous DRAM, SDRAM), a double data rate synchronous dynamic random access memory (Double Data Rate SDRAM, DDR SDRAM), an enhanced synchronous dynamic random access memory (Enhanced SDRAM, ESDRAM), a synchronous link dynamic random access memory (Synchronous link DRAM, SLDRAM), and a direct rambus random access memory (Direct Rambus RAM, DR RAM).

[0181] It should be noted that the memory described in this specification includes but is not limited to these memories and could be a memory of any other appropriate type.

[0182] An embodiment of the present disclosure provides a chip, including an input/output (I/O) interface and a processor, where the processor is configured to call and run a computer program stored in a memory, to enable a device installing with the chip to perform any of the above text processing methods or any of the above model training methods.

[0183] An embodiment of the present disclosure provides a computer-readable medium storing computer execution instructions which, when executed by a processor, causes the processor to execute any of the above text processing methods or any of the above model training methods. Optionally, the storage medium may be specifically a memory.

[0184] An embodiment of the present disclosure provides a computer program product including computer execution instructions which, when executed by a processor, causes the processor to execute any of the above text processing methods or any of the above model training methods.

[0185] An embodiment of the present disclosure provides a computer program including computer execution instructions which, when executed by a processor, causes the processor to execute any of the above text processing methods or any of the above model training methods.

[0186] A person of ordinary skill in the art may be aware that, in combination with the examples described in the embodiments disclosed in this specification, units and algorithm steps can be implemented by electronic hardware or a combination of computer software and electronic hardware. Whether the functions are performed by hardware or software depends on particular applications and design constraints of the technical solutions. A person skilled in the art may use different methods to

implement the described functions for each particular application, but it should not be considered that the implementation goes beyond the scope of this application.

[0187] In the several embodiments provided in this application, it should be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiment is merely an example. For example, the unit division is merely logical function division and may be other division in actual implementation. For example, multiple units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented through some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electronic, mechanical, or other forms.

[0188] The units described as separate parts may be or may not be physically separate, and parts displayed as units may be or may not be physical units, may be located in one position, or may be distributed on multiple network units. Some or all of the units may be selected based on actual requirements to achieve the objectives of the solutions of the embodiments.

[0189] In addition, functional units in the embodiments of this application may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit.

[0190] When the functions are implemented in a form of a software functional unit and sold or used as an independent product, the functions may be stored in a computer readable storage medium. Based on such an understanding, the technical solutions in this application essentially, or the part contributing to the prior art, or some of the technical solutions may be implemented in a form of a software product. The computer software product is stored in a storage medium, and includes several instructions for instructing a computer device (which may be a personal computer, a server, a network device, or the like) to perform all or some of the steps of the methods described in the embodiments of this application. The foregoing storage medium includes: any medium that can store program code, such as a USB flash drive, a removable hard disk, a read-only memory (Read-Only Memory, ROM), a random-access memory (Random Access Memory, RAM), a magnetic disk, or an optical disc.

[0191] Although the present disclosure describes methods and processes with steps in a certain order,

one or more steps of the methods and processes may be omitted or altered as appropriate. One or more steps may take place in an order other than that in which they are described, as appropriate.

[0192] Note that the expression “at least one of A or B”, as used herein, is interchangeable with the expression “A and/or B”. It refers to a list in which you may select A or B or both A and B. Similarly, “at least one of A, B, or C”, as used herein, is interchangeable with “A and/or B and/or C” or “A, B, and/or C”. It refers to a list in which you may select: A or B or C, or both A and B, or both A and C, or both B and C, or all of A, B and C. The same principle applies for longer lists having a same format.

[0193] Although the present disclosure is described, at least in part, in terms of methods, a person of ordinary skill in the art will understand that the present disclosure is also directed to the various components for performing at least some of the aspects and features of the described methods, be it by way of hardware components, software or any combination of the two. Accordingly, the technical solution of the present disclosure may be embodied in the form of a software product. A suitable software product may be stored in a pre-recorded storage device or other similar non-volatile or non-transitory computer readable medium, including DVDs, CD-ROMs, USB flash disk, a removable hard disk, or other storage media, for example. The software product includes instructions tangibly stored thereon that enable a processing device (e.g., a personal computer, a server, or a network device) to execute examples of the methods disclosed herein. The machine-executable instructions may be in the form of code sequences, configuration information, or other data, which, when executed, cause a machine (e.g., a processor or other processing device) to perform steps in a method according to examples of the present disclosure.

[0194] The present disclosure may be embodied in other specific forms without departing from the subject matter of the claims. The described example embodiments are to be considered in all respects as being only illustrative and not restrictive. Selected features from one or more of the above-described embodiments may be combined to create alternative embodiments not explicitly described, features suitable for such combinations being understood within the scope of this disclosure.

CLAIMS

1. A text processing method, comprising:

obtaining a first phonetic representation of reference acoustic information and a phoneme sequence of text information;

processing, in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

processing, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

obtaining speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

2. The method according to claim 1, wherein the obtaining a first phonetic representation of reference acoustic information comprises:

inputting the reference acoustic information into a pre-trained phonetic encoder to extract the first phonetic representation, wherein the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information.

3. The method according to claim 2, wherein the first phonetic representation of the reference acoustic information comprises self-supervised learning (SSL) tokens of the reference acoustic information, and the second phonetic representation of the text information comprises SSL tokens of the text information.

4. The method according to claim 2, wherein the first phonetic representation of the reference acoustic information comprises phonetic posterior grams (PPG) features of the reference acoustic information, and the second phonetic representation of the text information comprises PPG features of the text information.

5. The method according to any one of claims 1 to 4, wherein the processing, in an autoregressive manner, the phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information comprises:

processing, with a pre-trained autoregressive language model (LM) decoder, the first phonetic

representation of the reference acoustic information and the phoneme sequence to obtain the second phonetic representation of the text information, wherein the pre-trained autoregressive LM decoder is trained based on a plurality of phonetic representations of sample reference acoustic information corresponding to a plurality of acoustic samples one-by-one and a plurality of sample phoneme sequences corresponding to the plurality of acoustic samples one-by-one.

6. The method according to any one of claims 1 to 5, wherein the processing, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information comprises:

processing, with a pre-trained non-autoregressive LM decoder, the second phonetic representation of the text information, the phoneme sequence and the first acoustic features of the reference acoustic information to obtain the second acoustic features of the text information, wherein the pre-trained non-autoregressive LM decoder is trained based on first acoustic features of plural pieces of sample reference acoustic information, a plurality of second phonetic representations corresponding to a plurality of acoustic samples one-by-one and a plurality of sample phoneme sequences corresponding to the plurality of acoustic samples one-by-one, and the plural pieces of sample reference acoustic information correspond to the plurality of acoustic samples one-by-one.

7. The method according to any one of claims 1 to 6, wherein the first acoustic features are first acoustic tokens of the reference acoustic information, and the first acoustic tokens comprise fine-grained acoustic details of the reference acoustic information;

wherein the second acoustic features are second acoustic tokens of the text information, and the second acoustic tokens comprise fine-grained acoustic details for reconstructing the speech information corresponding to the text information.

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

inputting the reference acoustic information into a pre-trained acoustic encoder to obtain the first acoustic tokens of the reference acoustic information, wherein the pre-trained acoustic encoder is used for audio encoding of reference acoustic information.

9. A model training method, comprising:

acquiring a training sample set, wherein the training sample set comprises a plurality of acoustic

samples;

for an acoustic sample in the training sample set:

inputting sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, wherein the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

inputting a sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder to obtain a prediction result;

comparing the prediction result and the first phonetic representation of the sample reference acoustic information to obtain a first loss;

updating at least one parameter of the autoregressive LM decoder based on first losses corresponding to the plurality of acoustic samples.

10. A model training method, comprising:

acquiring a training sample set, wherein the training sample set comprises a plurality of acoustic samples;

for an acoustic sample in the training sample set:

inputting sample reference acoustic information corresponding to the acoustic sample into a pre-trained acoustic encoder to obtain first acoustic features of the sample reference acoustic information, wherein the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

inputting a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample into a pre-trained autoregressive language model (LM) decoder to obtain a second phonetic representation corresponding to the acoustic sample;

inputting the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample into a non-autoregressive LM decoder to obtain a prediction result;

comparing the prediction result and the first acoustic features of the sample reference acoustic information to obtain a second loss;

updating at least one parameter of the non-autoregressive LM decoder based on second losses corresponding to the plurality of acoustic samples.

11. A model training method, comprising:

acquiring a training sample set, wherein the training sample set comprises a plurality of acoustic samples;

for an acoustic sample in the training sample set:

inputting sample text information corresponding to the acoustic sample into a pre-trained phoneme generator of a speech synthesis model to obtain a sample phoneme sequence corresponding to the acoustic sample, wherein the pre-trained phoneme generator is used for phoneme generation based on text information;

inputting sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder of the speech synthesis model to obtain a first phonetic representation of the sample reference acoustic information, wherein the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

inputting the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder of the speech synthesis model to obtain a second phonetic representation of the sample text information;

inputting the sample reference acoustic information into a pre-trained acoustic encoder of the speech synthesis model to obtain first acoustic features of the sample reference acoustic information, wherein the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

inputting the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information into a non-autoregressive LM decoder of the speech synthesis model to obtain second acoustic features of the sample text information;

inputting the second acoustic features of the sample text information into a pre-trained acoustic decoder of the speech synthesis model to obtain a speech synthesis result, wherein the pre-trained acoustic decoder is used for audio decoding of acoustic features;

comparing the speech synthesis result and the acoustic sample to obtain a third loss; and

updating at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder based on third losses corresponding to the plurality of acoustic

samples.

12. A text processing method, comprising:

extracting, by a terminal device, text information from an application installed on the terminal device;

obtaining, by the terminal device, a first phonetic representation of reference acoustic information and a phoneme sequence of the text information;

processing, by the terminal device in an autoregressive manner, the first phonetic representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

processing, by the terminal device in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

obtaining, by the terminal device, speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

13. A text processing method, comprising:

extracting, by a terminal device, text information from an application installed on the terminal device;

sending, by the terminal device, the text information to a server;

receiving, by the terminal device, speech information corresponding to the text information sent by the server, wherein the speech information is obtained by the server based on second acoustic features of the text information, the second acoustic features of the text information are obtained by processing a second phonetic representation of the text information, a phoneme sequence of the text information and first acoustic features of reference acoustic information in a non-autoregressive manner, and the second phonetic representation of the text information is obtained by processing a first phonetic representation of the reference acoustic information and the phoneme sequence in an autoregressive manner.

14. A text processing apparatus, comprising:

a first obtaining model, configured to obtain a first phonetic representation of reference acoustic information and a phoneme sequence of text information;

a first processing module, configured to process, in an autoregressive manner, the first phonetic

representation of the reference acoustic information and the phoneme sequence to obtain a second phonetic representation of the text information;

a second processing module, configured to process, in a non-autoregressive manner, the second phonetic representation of the text information, the phoneme sequence and first acoustic features of the reference acoustic information to obtain second acoustic features of the text information; and

a second obtaining model, configured to obtain speech information corresponding to the text information and the reference acoustic information based on the second acoustic features of the text information.

15. A model training apparatus, comprising:

an acquiring module, configured to acquire a training sample set, wherein the training sample set comprises a plurality of acoustic samples;

an inputting module, configured to:

for an acoustic sample in the training sample set, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder to obtain a first phonetic representation of the sample reference acoustic information, wherein the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

input a sample phoneme sequence corresponding to the acoustic sample into an autoregressive language mode (LM) decoder to obtain a prediction result;

a comparing module, configured to, for an acoustic sample in the training sample set, compare the prediction result and the first phonetic representation of the sample reference acoustic information to obtain a first loss;

an updating module, configured to update at least one parameter of the autoregressive LM decoder based on first losses corresponding to the plurality of acoustic samples.

16. A model training apparatus, comprising:

an acquiring module, configured to acquire a training sample set, wherein the training sample set comprises a plurality of acoustic samples;

an inputting module, configured to:

for an acoustic sample in the training sample set, input sample reference acoustic information corresponding to the acoustic sample into a pre-trained acoustic encoder to obtain first acoustic

features of the sample reference acoustic information, wherein the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

input a first phonetic representation of the sample reference acoustic information and a sample phoneme sequence corresponding to the acoustic sample into a pre-trained autoregressive language model (LM) decoder to obtain a second phonetic representation corresponding to the acoustic sample;

input the second phonetic representation corresponding to the acoustic sample and the sample phoneme sequence corresponding to the acoustic sample into a non-autoregressive LM decoder to obtain a prediction result;

a comparing model, configured to, for an acoustic sample in the training sample set, compare the prediction result and the first acoustic features of the sample reference acoustic information to obtain a second loss;

an updating module, configured to update at least one parameter of the non-autoregressive LM decoder based on second losses corresponding to the plurality of acoustic samples.

17. A model training apparatus, comprising:

an acquiring module, configured to acquire a training sample set, wherein the training sample set comprises a plurality of acoustic samples;

an inputting module, configured to:

for an acoustic sample in the training sample set, input sample text information corresponding to the acoustic sample into a pre-trained phoneme generator of a speech synthesis model to obtain a sample phoneme sequence corresponding to the acoustic sample, wherein the pre-trained phoneme generator is used for phoneme generation based on text information;

input sample reference acoustic information corresponding to the acoustic sample into a pre-trained phonetic encoder of the speech synthesis model to obtain a first phonetic representation of the sample reference acoustic information, wherein the pre-trained phonetic encoder is used for extracting a phonetic representation from acoustic information;

input the first phonetic representation of the sample reference acoustic information and the sample phoneme sequence corresponding to the acoustic sample into an autoregressive language model (LM) decoder of the speech synthesis model to obtain a second phonetic representation of the sample text information;

input the sample reference acoustic information into a pre-trained acoustic encoder of the speech synthesis model to obtain first acoustic features of the sample reference acoustic information, wherein the pre-trained acoustic encoder is used for audio encoding of reference acoustic information;

input the second phonetic representation of the sample text information, the sample phoneme sequence corresponding to the acoustic sample and the first acoustic features of the sample reference acoustic information into a non-autoregressive LM decoder of the speech synthesis model to obtain second acoustic features of the sample text information;

input the second acoustic features of the sample text information into a pre-trained acoustic decoder of the speech synthesis model to obtain a speech synthesis result, wherein the pre-trained acoustic decoder is used for audio decoding of acoustic features;

a comparing model, configured to, for an acoustic sample in the training sample set, compare the speech synthesis result and the acoustic sample to obtain a third loss;

an updating module, configured to update at least one parameter of the autoregressive LM decoder and/or at least one parameter of the non-autoregressive LM decoder based on third losses corresponding to the plurality of acoustic samples.

18. An electronic device, comprising at least one processor coupled to a memory storing a set of instructions;

wherein the at least one processor is configured to read the set of instructions in the memory and execute the method according to any one of claims 1 to 8.

19. A computer-readable medium storing computer execution instructions which, when executed by a processor, cause the processor to execute the method according to any one of claims 1 to 8.