(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 4 503 017 A1

(12) EUROPEAN PATENT APPLICATION
published in accordance with Art. 153(4) EPC

(43) Date of publication:
05.02.2025 Bulletin 2025/06

(21) Application number: **22933875.1**

(22) Date of filing: **09.09.2022**

(51) International Patent Classification (IPC):
G10L 13/02 *(2013.01)* ***G10L 13/04*** *(2013.01)*
G10L 13/10 *(2013.01)* ***G10L 13/08*** *(2013.01)*
G10L 13/06 *(2013.01)*

(52) Cooperative Patent Classification (CPC):
G10L 13/02; G10L 13/04; G10L 13/06; G10L 13/08; G10L 13/10

(86) International application number:
PCT/CN2022/118072

(87) International publication number:
WO 2023/184874 (05.10.2023 Gazette 2023/40)

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(30) Priority: **31.03.2022 CN 202210344448**
31.03.2022 CN 202210346114
31.03.2022 CN 202210346097
31.03.2022 CN 202210346094
31.03.2022 CN 202210344456

(71) Applicants:
- **Midea Group (Shanghai) Co., Ltd.**
  Shanghai 201702 (CN)
- **Midea Group Co., Ltd.**
  Foshan, Guangdong 528311 (CN)

(72) Inventors:
- **GAO, Yu**
  Shanghai 201702 (CN)
- **LIU, Xueling**
  Shanghai 201702 (CN)
- **TU, Jianhua**
  Shanghai 201702 (CN)

(74) Representative: **Haseltine Lake Kempner LLP**
Cheapside House
138 Cheapside
London EC2V 6BJ (GB)

(54) SPEECH SYNTHESIS METHOD AND APPARATUS

(57) The present application relates to the field of speech synthesis. Provided is a speech synthesis method, comprising: segmenting a rhythm phoneme sequence of target text, so as to generate a plurality of clause sequences, wherein the rhythm phoneme sequence comprises a plurality of phonemes corresponding to the target text, and rhythm identifiers, each of which is located between adjacent phonemes, with each clause sequence comprising at least one phoneme; performing speech synthesis on a first sub-rhythm phoneme sequence in the plurality of clause sequences, so as to obtain first speech information; and outputting the first speech information, and performing speech synthesis on a second sub-rhythm phoneme sequence in the plurality of clause sequences, so as to generate second speech information, wherein the second sub-rhythm phoneme sequence is at least one clause sequence, which is located after the first sub-rhythm phoneme sequence, in the rhythm phoneme sequence. By means of the speech synthesis method of the present application, the feedback from a system after same has received a network speech synthesis service request is effectively sped up, and the waiting time of users is shortened.

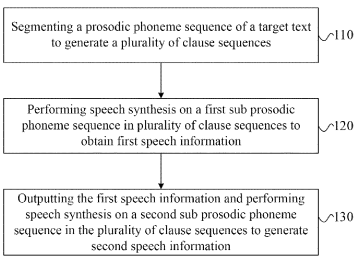

FIG. 1

EP 4 503 017 A1

Description

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present application claims priorities to Chinese Patent Application No. 202210344448X, filed on March 31, 2022, entitled "Speech Synthesis Method and Speech Synthesis Apparatus", Chinese Patent Application No. 2022103461146, filed on March 31, 2022, entitled "Speech Splicing Method and Speech Splicing Apparatus", Chinese Patent Application No. 2022103460976, filed on March 31, 2022, entitled "Method and Apparatus for Obtaining Audio and Video File Size", Chinese Patent Application No. 2022103460942, filed on March 31, 2022, entitled "Text Transliteration Method and Text Transliteration Apparatus", and Chinese Patent Application No. 2022103444564, filed on March 31, 2022, entitled "Text Segmentation Method and Text Segmentation Apparatus", which are hereby incorporated by reference in their entireties.

FIELD

[0002] The present application relates to the field of speech synthesis, and particularly to a speech synthesis method and apparatus, a speech splicing method and apparatus, a method and apparatus for obtaining audio and video file size, a text transliteration method and apparatus, and a text segmentation method and a text segmentation apparatus.

BACKGROUND

[0003] A text to speech (TTS) technology is widely used in the field of speech synthesis. In the related art, speech synthesis is usually performed directly on an entire to-be-synthesized text. For some longer to-be-synthesized texts, it takes longer to perform speech synthesis, which also means that a user needs to wait for a longer time for obtaining synthesized speech. The performance of speech synthesis is low, which not only wastes the user's time but also affects his experience.

SUMMARY

[0004] The present application aims to solve at least one of the problems in the related art. Therefore, the present application provides a speech synthesis method and a speech synthesis apparatus.

[0005] An embodiment provides a speech synthesis method, including: segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences, where the prosodic phoneme sequence includes a plurality of phonemes corresponding to the target text and a prosodic identifier located between adjacent phonemes, and each clause sequence includes at least one phoneme; performing speech synthesis on a first sub prosodic phoneme sequence in the plurality of clause sequences to obtain first speech information; and outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in the plurality of clause sequences to generate second speech information, where the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

[0006] An embodiment further provides a speech synthesis apparatus, including: a first processing module, used for segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences, where the prosodic phoneme sequence includes a plurality of phonemes corresponding to the target text and a prosodic identifier located between adjacent phonemes, and each clause sequence includes at least one phoneme; a second processing module, used for performing speech synthesis on a first sub prosodic phoneme sequence in the plurality of clause sequences to obtain first speech information; and a third processing module, used for outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in the plurality of clause sequences to generate second speech information, where the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

[0007] An embodiment further provides an electronic device, including a memory, a processor and a computer program stored in the memory and capable of running on the processor, where the processor, when executing the computer program, performs the speech synthesis method as described above.

[0008] An embodiment further provides a non-transient computer-readable storage medium, on which a computer program is stored, where the computer program is executed by a processor to perform the speech synthesis method as described above.

[0009] An embodiment further provides a computer program product, including a computer program, where the computer program is executed by a processor to perform the speech synthesis method as described above.

BRIEF DESCRIPTION OF DRAWINGS

[0010] To illustrate solutions according to the present application or the related art, the accompanying drawings used in the description of the embodiments of the present application or the related art are briefly described below. It should be noted that the drawings in the following description are of only part embodiments of the present application. For those of ordinary skill in the art, other drawings may also be obtained according to these drawings without creative efforts.

FIG. 1 is a first schematic flow chart of a speech

synthesis method according to an embodiment of the present application;

FIG. 2 is a second schematic flow chart of a speech synthesis method according to an embodiment of the present application;

FIG. 3 is a schematic structural diagram of a speech synthesis apparatus according to an embodiment of the present application;

FIG. 4 is a schematic structural diagram of an electronic device according to an embodiment of the present application;

FIG. 5 is a first schematic flow chart of a speech splicing method according to an embodiment of the present application;

FIG. 6 is a second schematic flow chart of a speech splicing method according to an embodiment of the present application;

FIG. 7 is a first schematic flow chart of a method for obtaining audio and video file size according to an embodiment of the present application;

FIG. 8 is a second schematic flow chart of a method for obtaining audio and video file size according to an embodiment of the present application;

FIG. 9 is a first schematic flow chart of a text transliteration method according to an embodiment of the present application;

FIG. 10 is a second schematic flow chart of a text transliteration method according to an embodiment of the present application;

FIG. 11 is a first schematic flow chart of a text segmentation method according to an embodiment of the present application;

FIG. 12 is a second schematic flow chart of a text segmentation method according to an embodiment of the present application; and

FIG. 13 is a third schematic flow chart of a text segmentation method according to an embodiment of the present application.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0011] The following will provide a further detailed description of implementation of the present application in conjunction with the accompanying drawings and embodiments. The following embodiments are used to illustrate the present application, but cannot be used to limit the scope of the present application.

[0012] In the description of the present specification, the reference terms "an embodiment", "some embodiments", "example", "specific example", or "some examples" refer to the specific features, structures, materials, or features described in conjunction with the embodiments or examples included in at least one embodiment or example of the present application. In the present specification, the illustrative expressions of the above terms do not necessarily refer to the same embodiments or examples. Moreover, the specific features, structures, materials, or features described may be combined in an appropriate manner in any one or more embodiments or examples. In addition, those skilled in the art may combining the different embodiments or examples described in the present specification, as well as the features of different embodiments or examples, without conflicting with each other.

[0013] A speech synthesis method according to the present embodiment is described below in conjunction with FIG. 1 and FIG. 2.

[0014] The speech synthesis method may be executed by a speech synthesis apparatus, a server, or a user terminal, including but not limited to a mobile phone, a tablet, an PC, a vehicle terminal and a smart appliance.

[0015] As shown in FIG. 1, the speech synthesis method includes step 110, step 120 and step 130.

[0016] Step 110: segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences.

[0017] In this step, the target text is a text currently used for speech synthesis.

[0018] The prosodic phoneme sequence is a sequence used to represent a prosodic feature and a phonemic feature of the target text.

[0019] The prosodic phoneme sequence includes prosodic identifiers located between adjacent phonemes and a plurality of phonemes corresponding to the target text.

[0020] The phoneme may be a combination of one or more phonetic units segmented based on natural properties of speech. The phonetic unit may be a Pinyin, an initial, or a final corresponding to a Chinese character, or an English word, an English phonetic symbol, or an English letter.

[0021] The prosodic identifier is an identifier used to represent the prosodic feature corresponding to each phoneme in the target text, where the prosodic feature includes, but not limited to: tones, syllables, prosodic words, prosodic phrases, intonational phrases, a silence, pauses, etc. corresponding to phonemes.

[0022] A fine grain of a prosodic identifier used to represent the pauses is greater than that of a prosodic identifier used to represent intonational phrases. A fine grain of a prosodic identifier used to represent intonational phrases is greater than that of a prosodic identifier used to represent prosodic phrases. A fine grain of a prosodic identifier used to represent prosodic phrases is greater than that of a prosodic identifier used to represent prosodic words. A fine grain of a prosodic identifier used to represent prosodic words is greater than that of a prosodic identifier used to represent the syllables.

[0023] In actual application process, different symbols may be used to represent prosodic features of different fine-grained levels.

[0024] For example, for a target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4", it may be converted into a prosodic phoneme sequence: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0

fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0025] It may be understood that for this prosodic phoneme sequence, prosodic identifiers may include: numbers, symbols, and English phonemes between adjacent Pinyin; phonemes may include Pinyin corresponding to each Chinese character.

[0026] Sil represents silence at the beginning and the end of a sentence in a prosodic phoneme sequence, #0 represents syllables, #1 represents prosodic words, #2 represents prosodic phrases, #3 represents intonational phrases, and #4 represents the end of a sentence. The number behind each phoneme represents the tone of that phoneme, such as 4 in shang4, representing a fourth tone in Pinyin "shang".

[0027] It may be understood that an entire prosodic phoneme sequence is sequentially connected by clause sequences.

[0028] In some embodiments, step 110 may include: converting the target text into a prosodic phoneme sequence; and segmenting the prosodic phoneme sequence based on at least partial of a plurality of prosodic identifiers to generate a plurality of clause sequences.

[0029] In this embodiment, the target text is a text currently used for speech synthesis.

[0030] The entire prosodic phoneme sequence includes a plurality of phonemes and a plurality of prosodic identifiers, and the plurality of prosodic identifiers include prosodic identifiers corresponding to different fine-grained levels.

[0031] In an actual application process, an appropriate fine-grained level may be selected as a segmentation standard based on actual situation, and a position of the corresponding prosodic identifier in the prosodic phoneme sequence may be used as a segmentation point to segment the prosodic phoneme sequence to obtain the plurality of clause sequences.

[0032] It should be noted that each clause sequence includes a prosodic identifier at the segmentation point and at least one phoneme.

[0033] It may be understood that, for each prosodic phoneme sequence at least two clause sequences corresponding to at least one segmentation point may be obtained.

[0034] For example, for prosodic phoneme sequence "sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil", based on actual needs, it is determined that the prosodic phoneme sequence is segmented at #3. Then, segmentation is performed at the positions containing #3 in the prosodic phoneme sequence, while retaining a prosodic separator #3 to the previous splicing unit, so that the prosodic phoneme sequence may be segmented into the following clause sequences:

clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3; and clause sequence 2: dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0035] Determination of the segmentation point will be explained in subsequent embodiments, and will not be elaborated here.

[0036] Step 120: performing speech synthesis on a first sub prosodic phoneme sequence in plurality of clause sequences to obtain first speech information.

[0037] In this step, the first sub prosodic phoneme sequence is a prosodic phoneme sequence before the first segmentation point of the prosodic phoneme sequence.

[0038] Continuing with the examples of clause sequence 1 and clause sequence 2 in the above embodiments, the first sub prosodic phoneme sequence is the clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3.

[0039] In an actual application process, a vocoder may be used to perform speech synthesis on the first sub prosodic phoneme sequence to generate the first speech information corresponding to the first sub prosodic phoneme sequence.

[0040] Step 130: outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in plurality of clause sequences to generate second speech information, where the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

[0041] In this step, after generating the first speech information, the first speech information is returned to a client for output, so that a user may play the first speech information.

[0042] During the first speech information is output, a background continues to perform speech synthesis on the second sub prosodic phoneme sequence to generate the second speech information corresponding to the second sub prosodic phoneme sequence.

[0043] The second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence. For the clause sequence 1 and clause sequence 2 in the above embodiments, the second sub prosodic phoneme sequence is the clause sequence 2: dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0044] In other embodiments, the method may further include: determining that any one of to-be-matched clause sequences in plurality of clause sequences matches a target clause sequence in a cache, obtaining a target clause speech corresponding to the target clause sequence from the cache, and determining speech corresponding to to-be-matched clause sequences among the plurality of clause sequences as the target clause speech; or determining that any one of to-be-matched clause sequences in plurality of clause sequences does not match a target clause sequence in a cache, and performing speech synthesis on the to-be-matched

clause sequences to generate a second clause speech.

[0045] For example, a clause sequence in a plurality of pre-cached clause sequences is matched with the first sub prosodic phoneme sequence or the second sub prosodic phoneme sequence, pre-generated and cached speech corresponding to the matched clause sequence is obtained, and synthesized speech corresponding to the first sub prosodic phoneme sequence or the second sub prosodic phoneme sequence is further obtained. The corresponding speech is first matched from the cache without real-time synthesis, which improves the efficiency of speech synthesis.

[0046] In some embodiments, the method may further include: segmenting the prosodic phoneme sequence of the target text to generate a plurality of candidate sequences; and combining the target candidate sequence from a plurality of candidate sequences with adjacent candidate sequences to generate the fine grain size corresponding to the clause sequence and a plurality of clause sequences.

[0047] For example, by segmenting the prosodic phoneme sequence "sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil" of sentence "Hope this song | may make you like it | play for you | XX", a plurality of candidate sequences may be obtained: "Xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3", "neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1", "wei4 #0 nin2 #1 bol #0 fang4 #1", and "EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4".

[0048] Any one candidate sequence from the candidate sequences is taken as the target candidate sequence, and combined with adjacent candidate sequences to obtain the following plurality of clause sequences (the one between two "|" is a clause sequence):

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 si1";

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 si1";

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil"; and

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil".

[0049] The plurality of clause sequences are ordered in descending order based on the corresponding fine grain size. The larger the fine grain, the higher the corresponding clause sequence is ordered. For example, "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 (Hope this song may make you like it play for you XX's X)" is greater thanordered before "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 (Hope this song may make you like it play for you)".

[0050] In some embodiments, the plurality of clause sequences may match with the target clause sequence in the cache in a descending order, and the speech of the successfully matched target clause sequence may be determined as the speech of the clause sequence.

[0051] In this embodiment, based on the generated descending order, the clause sequence precisely matches the target clause sequence from top to bottom. For example, the clause sequence "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bo1 #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 (Hope this song may make you like it play for you XX's X)" first precisely matches the target clause sequence. If it is matched successfully, the clause sequence is determined as a first clause sequence, and the target clause speech corresponding to the target sentence matched with the first clause sequence is determined as the speech corresponding to the first clause sequence, and the comparison ends.

[0052] If it is not matched, the clause sequence "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 (Hope this song may make you like it play for you)" is compared with the target clause. The above process is repeated until it is determined that a certain clause sequence may precisely match the target clause, and the comparison ends.

[0053] During research, the applicant found that in the related art, speech synthesis is usually performed directly on an entire to-be-synthesized text. Considering the temporal nature of speech, a conversion time of a system is usually proportional to a length of an input text, and the longer the sentence, the longer the time for synthesis is required.

[0054] For inputs ranging from tens to hundreds of characters, synthesis using current fastest deep learning model takes seconds to tens of seconds. For some longer to-be-synthesized texts, speech synthesis requires a longer time, which means that users need to wait for a

longer time for obtaining the synthesized speech, which not only wastes their time but also affects their experience.

[0055] During research, the applicant also found that to solve the above problem, there is a method that divides the to-be-synthesized text into a plurality of sub texts and performs speech synthesis on the plurality of sub texts through a system parallel synthesis method. However, this method is limited to be operated on GPU, and this method fails to improve synthesis performance and still requires a lot of time on CPU servers.

[0056] In the present application, by segmenting the target text into the plurality of clause sequences, speech synthesis is first performed on the first sub prosodic phoneme sequence and the first speech information synthesized from the first sub prosodic phoneme sequence is first output. During the first speech information is output, the clause sequence subsequent to the first sub prosodic phoneme sequence is synthesized, which effectively accelerates a feedback speed of the system after receiving a network speech synthesis service request, shortens waiting time of users, and conveniently improves the user experience.

[0057] In an actual application process, performing speech synthesis on the second sub prosodic phoneme sequence in the plurality of clause sequences may include sequentially performing speech synthesis on each clause sequence based on an order of segmenting the target text into each clause sequence.

[0058] For example, after the target text is sequentially segmented into a first clause sequence, a second clause sequence, and a third clause sequence, the first clause sequence is the first sub prosodic phoneme sequence, and speech synthesis is first performed on the first clause sequence to generate the first speech information, speech synthesis is performed on the second clause sequence while the first speech information is output, and speech synthesis is performed on the third clause sequence after second speech information corresponding to the second clause sequence is generated.

[0059] Performing speech synthesis on the second sub prosodic phoneme sequence in plurality of clause sequences may also include simultaneously performing speech synthesis on each clause sequence.

[0060] For example, after the target text is sequentially segmented into a first clause sequence, a second clause sequence, and a third clause sequence, the first clause sequence is the first sub prosodic phoneme sequence, and speech synthesis is first performed on the first clause sequence to generate the first speech information, speech synthesis is performed on the second clause sequence and the third clause sequence in parallel while the first speech information is output utilizing parallel synthesis capability of a system.

[0061] In the speech synthesis method provided by the embodiments of the present application, by segmenting the target text into the plurality of clause sequences, and speech synthesis is first performed on the first clause sequence to generate the first speech information. While the first speech information is output, speech synthesis is continuously performed on subsequent clause sequences, which effectively accelerates the feedback speed of the system after receiving network speech synthesis service request, shortens waiting time of users, and conveniently improves the user experience.

[0062] As shown in FIG. 2, according to some embodiments of the present application, before step 110, the method may further include: obtaining a to-be-synthesized text; if a size of the to-be-synthesized text exceeds a target threshold, segmenting the to-be-synthesized text and generating a target text of which a size does not exceed the target threshold.

[0063] In this embodiment, the to-be-synthesized text is an original text on which speech synthesis is to be performed.

[0064] The text level of the to-be-synthesized text may be tens to hundreds of levels of regular text, as well as thousands or tens of thousands of levels of super text.

[0065] The target threshold may be determined based on at least one of computing power of the system and the upper limit of capability of a speech synthesis model, for example, the target threshold may be determined within a range of several hundred characters.

[0066] In an actual application process, for the obtained to-be-synthesized text, the size of the to-be-synthesized text is first determined and compared with the target threshold. If the size of the to-be-synthesized text does not exceed the target threshold, the entire synthesized text is directly determined as the target text.

[0067] If the size of the to-be-synthesized text exceeds the target threshold, based on at least one of the obtained computing power of the system and capability information of the speech synthesis model, the synthesized text is first segmented to obtain a plurality of segments of first texts, so that a size of each segment of first texts does not exceed the target threshold, and the first segment of the plurality of segments of first texts is determined as the target text.

[0068] In the speech synthesis method provided by the embodiments of the present application, the synthesized text is segmented based on the target threshold to generate the target text, which may fully consider the actual capability of a server to provide speech synthesis of the target text within the processing capacity range of the server, thereby improving the performance of speech synthesis.

[0069] In some embodiments, step 110 may include: obtaining end-of-sentence information, an intonational phrase, a prosodic phrase, a prosodic word and a syllable of the target text; and

labeling the target text based on at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word, and the syllable to generate the prosodic phoneme sequence.

[0070] In this embodiment, the syllable is a phonetic unit in a speech flow and is also the most easily recogniz-

able phonetic unit in auditory perception by human. For example, a syllable may be each Chinese character in the target text.

[0071] The prosodic word is a group of syllables that are closely associated and pronounced together in actual speech flow.

[0072] The prosodic phrase is a medium rhythmic block that lies between the prosodic word and the intonational phrase. The prosodic phrase may include a plurality of prosodic words and mood particles, and the plurality of prosodic words that make up the prosodic phrase sound like they share a common rhythm group.

[0073] The intonational phrase is a sentence formed by connecting a plurality of prosodic phrases based on a certain intonation pattern, and is used to represent longer pauses.

[0074] The end-of-sentence information is used to represent the end of each long sentence.

[0075] For example, for a target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4 ", each Chinese character such as "shang", "hai", and "shi" is the corresponding syllable of the target text; words or phrases composed of words such as "Shanghai", "today", and "overcast to cloudy " are the corresponding intonational phrase of the target text; the sentence "overcast to cloudy today in Shanghai" composed of the intonational phrases "Shanghai", "today", and "overcast to cloudy" is the intonational phrase corresponding to the target text.

[0076] The target text is labeled based on at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word and the syllable of the target text after the end-of-sentence information, the intonational phrase, the prosodic phrase the prosodic word of the target text are obtained, to generate the prosodic phoneme sequence.

[0077] During research, the applicant found that in the related art, punctuations, such as commas or periods, are often used to represent the prosody of sentences, such as the sentences are segmented at the position of commas or periods to obtain a plurality of clauses. On the one hand, this mode cannot meet the segmentation requirements for text without punctuation, and on the other hand, it may lead to uneven segmentation of a sentence, resulting in poor segmentation performance.

[0078] In the present application, at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word and the syllable are used to represent the prosody of the sentence, and the target text is segmented based on this, without the situation of being segmented in the middle of a word, making the pauses of the segmented sentence and prosody sounds more natural.

[0079] In some embodiments, the target text is labeled based on at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word and the syllable, and generating the prosodic phoneme sequence includes: converting the target text into a phoneme sequence; generating a plurality of prosodic identifiers based on at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word, and the syllable; and labeling the phoneme sequence based on the plurality of prosodic identifiers to generate the prosodic phoneme sequence.

[0080] In this embodiment, the prosodic phoneme sequence is a sequence connecting articulation mark corresponding to each Chinese character or English word in the target text, including Pinyin, tone, or English phonetic notation.

[0081] For example, for a target text "overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4", it may be converted into a phoneme sequence: shang4 hai3 shi4 jin1 tian1 yin1 zhuan3 duo1 yun2 dong1 nan2 fengl san1 dao4 si4 ji2.

[0082] The prosodic identifier is an identifier used to represent the prosodic feature corresponding to each phoneme in the target text, that is, the prosodic identifier is a symbol used to represent the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word and the syllable.

[0083] In an actual application process, a combination of special symbol and number or a combination of specific letters may be used to represent prosodic identifiers. For example, "#0", "#1", "#2", "#3", and "#4" are used to represent the prosodic identifiers respectively. Different combinations represent different fine-grained levels.

[0084] For example: #0 represents a syllable, #1 represents a prosodic word, #2 represents a prosodic phrase, #3 represents an intonational phrase, and #4 represents an ending of the sentence. In this embodiment, the fine grain in an ascending order is: #0 < #1 < #2 < #3 < #4.

[0085] The prosodic identifiers are inserted into the corresponding position in the phoneme sequence after the corresponding phoneme sequence and prosodic identifier of the target text are obtained. For example, the prosodic identifier #0 used to represent the syllable is inserted into the Pinyin corresponding to each syllable in the phoneme sequence, and the prosodic identifier #2 used to represent the prosodic phrase is inserted into each phoneme sequence, and the phoneme sequence is converted into the prosodic phoneme sequence.

[0086] For example, #0, #1, #2, #3, and #4 are respectively used to label the phoneme sequence "shang4 hai3 shi4 jin1 tian1 yin1 zhuan3 duo1 yun2 dong1 nan2 fengl san1 dao4 si4 ji2" to generate a prosodic phoneme sequence: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0087] Sil represents the silence at the beginning and end of the sentence.

[0088] In the speech synthesis method provided by the embodiments of the present application, by converting the target text into the phoneme sequence and labeling the phoneme sequence based on prosodic identifiers

corresponding to at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word and the syllable to generate the prosodic phoneme sequence, a more refined prosodic representation is provided, which helps to improve the segmentation delicacy and the accuracy of the subsequent segmentation.

[0089] Continuing with reference to FIG. 2, according to some embodiments of the present application, the method may further include: generating a target file size for third speech information based on the prosodic phoneme sequence.

[0090] Step 130: generating second speech information based on the target file size.

[0091] In this embodiment, the third speech information is speech information generated by synthesizing speech information from at least two clause sequences in the plurality of clause sequences corresponding to the target text, where one of the at least two clause sequences is the first sub prosodic phoneme sequence.

[0092] The target file size is the predicted file size of the third speech information.

[0093] The target file size may be file volume information or length information of the third speech information, and no limitation is made in the present application.

[0094] After the target file size of the third speech information is obtained, the speech data generated based on the second sub prosodic phoneme sequence is supplemented based on the target file size, to generate the second speech information.

[0095] In some embodiments, generating the target file size for the third speech information based on the prosodic phoneme sequence may include: generating a predicted file size for third speech information based on the prosodic phoneme sequence; and correcting the predicted file size based on a target residual value to generate the target file size.

[0096] In this embodiment, the predicted file size is an uncorrected initial file size of the speech synthesized from the target text, which is predicted based on the prosodic phoneme sequence.

[0097] The target residual value is used to correct the predicted file size to improve the accuracy of the final generated target file size.

[0098] The target residual value is determined based on a sample file size and a predicted sample audio file size corresponding to the sample text. The sample file size is the actual sample audio file size corresponding to the sample text. The target file size is a file size of the speech synthesized from the target text, which is predicted based on the prosodic phoneme sequence and corrected. It may be understood that the accuracy of the target file size is greater than that of the predicted file size.

[0099] The target residual value is a predetermined value, for example, the target residual value may be a maximum residual value.

[0100] In this embodiment, residual is supplemented to the predicted file size to correct it, thereby improving the accuracy of the final generated target file size.

[0101] In some embodiments, the target residual value may be determined by the following steps:

obtaining the sample text, the sample file size corresponding to the sample audio file, and the sample audio file corresponding to the sample text, where the sample audio file is generated by performing speech synthesis on the sample text;

converting the sample text into the sample prosodic phoneme sequence;

predicting the sample audio file size based on the sample prosodic phoneme sequence, and generating a sample predicted file size of the sample audio file; and

determining an absolute value of a difference between the sample file size and the sample predicted file size as the target residual value.

[0102] In this embodiment, the text level of the sample text may be tens to hundreds of levels of regular text, as well as thousands or tens of thousands of levels of super text.

[0103] The sample audio file is the audio file generated by performing speech synthesis on the sample text.

[0104] The sample file size is an actual size or an actual audio duration of the sample audio file.

[0105] For example, a speech synthesis system may be used to calculate an actual WAV file size or audio duration of the sample audio file corresponding to the sample text.

[0106] The sample predicted file size is an uncorrected and predicted size or an audio duration of the sample predicted audio file .

[0107] It should be noted that the generation mode of the sample predicted file size should be consistent with the generation mode of the predicted file size.

[0108] The absolute value of the difference between the sample predicted file size and the sample file size is calculated and taken as the target residual value.

[0109] It may be understood that during performing, a plurality of predictions may be made on the sample prosodic phoneme sequence to obtain a plurality of sample predicted file sizes. Differences between each predicted file size and the sample file size are calculated separately, and a plurality of candidate differences are obtained; then, the absolute value of the minimum value is selected from the plurality of candidate differences to determine the target residual value, which improves the accuracy of the target residual value.

[0110] In the speech synthesis method provided by the embodiments of the present application, the size information of the target audio file synthesized from the target text is predicted based on the prosodic phoneme sequence, and the predicted value is corrected based on the target residual value, which may predict the size of the target audio file before it is generated, and has a prediction result of high accuracy and high precision.

[0111] Continuing with reference to FIG. 2, in some embodiments, step 110 may include:

> determining a first segmentation position based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence,
> determining a second segmentation position from positions of the prosodic identifier located after the first segmentation position in the prosodic phoneme sequence; and
> segmenting the prosodic phoneme sequence based on the first segmentation position and the second segmentation position to generate the first sub prosodic phoneme sequence and at least two second sub prosodic phoneme sequences, where the first sub prosodic phoneme sequence is the prosodic phoneme sequence located before the first segmentation position in the prosodic phoneme sequence, the at least two second sub prosodic phoneme sequences are the prosodic phoneme sequences located after the first segmentation position in the prosodic phoneme sequence, the adjacent second sub prosodic phoneme sequences are determined based on the second segmentation position, and a speech synthesis duration corresponding to the first sub prosodic phoneme sequence is within a target duration.

[0112] In this embodiment, the first segmentation position is the segmentation point used for the first segmentation.

[0113] The second segmentation position refers to the position of the segmentation points corresponding to all segmentations other than the first segmentation.

[0114] The prosodic phoneme sequence may be segmented into two sub sequences based on the first segmentation position, and the sub sequence located before the first segmentation position is determined as the first sub prosodic phoneme sequence.

[0115] It should be noted that the speech synthesis duration corresponding to the first sub prosodic phoneme sequence generated based on the first segmentation position is within the target duration.

[0116] The speech synthesis duration corresponding to the first sub prosodic phoneme sequence is time that it takes to synthesize the first sub prosodic phoneme sequence into speech.

[0117] The speech synthesis duration is related to a computing power of the speech synthesis system.

[0118] The target duration is a relatively short duration, and the value of the target duration may be customized by a user or a system default value may be used, such as setting the target duration to 0.2 s or 0.3 s.

[0119] After the first segmentation position is determined, at least a portion of the prosodic identifiers after the first segmentation position are searched from the prosodic phoneme sequence as candidate sets for determining the second segmentation position, and the position of the prosodic identifiers in the candidate set is determined as the second segmentation position.

[0120] It may be understood that in other embodiments, in absence of second segmentation position, the second sub prosodic phoneme sequence is the entire prosodic phoneme sequence located after the first segmentation position in the prosodic phoneme sequence. For example, if the second segmentation position is a position corresponding to #3, but #3 cannot be found in the second prosodic phoneme sequence, it may be understood that there is no second segmentation position.

[0121] In this embodiment, the first segmentation position is determined based on the prosodic identifier in the prosodic phoneme sequence, so that the speech synthesis duration corresponding to the first sub prosodic phoneme sequence obtained based on the first segmentation position may be within a reasonable duration range, thereby shortening first sentence response time and delay time of the synthesis system. In addition, the first segmentation position determined based on this method is the position with a longer pause time, which makes the pause and prosody of the first sub prosodic phoneme sequence obtained by segmentation more natural, thereby making the subsequent output of speech synthesized based on the first sub prosodic phoneme sequence more natural and smooth.

[0122] Continuing with reference to FIG. 2, according to some embodiments of the present application, after step 130, the method may further include: combining the first speech information and the second speech information to generate third speech information.

[0123] In this embodiment, the second speech information is speech information obtained by performing speech synthesis on the second sub prosodic phoneme sequence, where the second sub prosodic phoneme sequence may be one or more clause sequences, and all the second sub prosodic phoneme sequences are located after the first sub prosodic phoneme sequence in the target text.

[0124] For example, in case of sequentially synthesizing the second speech information, while the first speech information is output, speech synthesis may be performed on the second clause sequence adjacent to the first sub prosodic phoneme sequence located after the first sub prosodic phoneme sequence, to generate the second speech information corresponding to the second clause sequence. While the second speech information is output, the first speech information and the second speech information are combined to generate the third speech information.

[0125] For example, in case of sequentially synthesizing the second speech information, while the first speech information is output, speech synthesis may be performed on the second sub prosodic phoneme sequence adjacent to the first sub prosodic phoneme sequence located after the first sub prosodic phoneme sequence, to generate the second speech information. While the second speech information is output, speech synthesis may

be performed on the third clause sequence adjacent to the second clause sequence located after the second clause sequence, to generate the second speech information corresponding to the third clause sequence. While the second speech information is output, speech synthesis is performed on the subsequent clause sequences until the second speech information corresponding to all clauses is generated. The second speech information corresponding to all clauses and the first speech information are synthesized to generate the third speech information.

[0126] For the parallel synthesis of the second speech information, while outputting the first speech information, the plurality of clause sequences located after the first prosodic phoneme sequence may be synthesized in parallel, and the corresponding second speech information for each clause sequence may be generated. Then, the first speech information and the plurality of pieces of second speech information obtained are synthesized to generate the third speech information.

[0127] In some embodiments, combining the first speech information and the second speech information may include: combining the first speech information and the second speech information based on a phoneme duration corresponding to the first speech information and a phoneme duration corresponding to the second speech information.

[0128] In this embodiment, the phoneme duration is a corresponding articulation duration of the phoneme.

[0129] For example, for clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3, "shang" may be segmented into two phonemes: "sh" and "ang", each phoneme corresponds to a articulation duration.

[0130] During splicing, it is necessary to first truncate the speech information corresponding to each clause sequence based on a redundant phoneme duration at the beginning or end of each clause sequence, to remove the redundant phoneme duration at the beginning or end of the speech information.

[0131] Taking the first speech information for an example, in case where the first speech information is speech, after the speech is synthesized, the duration corresponding to the redundant phonemes at the beginning or end of the first sub prosodic phoneme sequence in the speech is truncated to generate truncated first speech information.

[0132] In case where the first speech information is an advanced acoustic feature, after the first speech information is synthesized, the duration corresponding to the redundant phonemes at the beginning or end of the first sub prosodic phoneme sequence in the advanced acoustic feature corresponding to the first speech information is truncated, to generate a truncated advanced acoustic feature; then, a vocoder is used to perform speech synthesis on the truncated advanced acoustic feature to generate truncated first speech information.

[0133] The truncation mode of the second speech information is the same as that of the first speech information, and will not be elaborated here.

[0134] Based on an order of segmenting the prosodic phoneme sequence into the first sub prosodic phoneme sequence corresponding to the first speech information, and an order of segmenting the prosodic phoneme sequence into the second sub prosodic phoneme sequence corresponding to the second speech information, starting from the first clause sequence, the truncated speech information corresponding to the adjacent clause sequence is sequentially spliced until all the speech information corresponding to the clause sequence is spliced.

[0135] In the speech synthesis method provided by the embodiments of the present application, by splicing the first speech information and the second speech information based on the phoneme duration, the naturalness and fluency of adjacent speech information splicing may be improved without a preset speech splicing unit library.

[0136] As shown in FIG. 3, a speech synthesis apparatus includes: a first processing module 310, a second processing module 320 and a third processing module 330.

[0137] The first processing module 310 is used for segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences, where the prosodic phoneme sequence includes a plurality of phonemes corresponding to the target text and prosodic identifiers located between adjacent phonemes, and each clause sequence includes at least one phoneme. The second processing module 320 is used for performing speech synthesis on a first sub prosodic phoneme sequence in the plurality of clause sequences to obtain first speech information. The third processing module 330 is used for outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in the plurality of clause sequences to generate second speech information, where the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

[0138] In the speech synthesis apparatus provided by the embodiments of the present application, by segmenting the target text into the plurality of clause sequences, and speech synthesis is first performed on the first clause sequence to generate the first speech information. While the first speech information is output, speech synthesis is continuously performed on subsequent clause sequences, which effectively accelerates the feedback speed of the system after receiving network speech synthesis service request, shortens waiting time of users, and conveniently improves the user experience.

[0139] In some embodiments, the first processing module 310 is further used for: converting the target text into the prosodic phoneme sequence, which includes prosodic identifiers located between adjacent phonemes and a plurality of phonemes corresponding to the target text; and segmenting the prosodic phoneme sequence based on at least a portion of a plurality of prosodic

identifiers to generate a plurality of clause sequences, and each clause sequence includes at least one phoneme.

[0140] In some embodiments, the apparatus may further include: a fifth processing module, used for combining the first speech information and the second speech information to generate third speech information after generating the second speech information.

[0141] In some embodiments, the apparatus may further include: a sixth processing module, used to generate a target file size of the third speech information based on the prosodic phoneme sequence after converting the target text into the prosodic phoneme sequence; a fourth processing module, used to generate second speech information based on the target file size.

[0142] In some embodiments, the sixth processing module is further used for: generating predicted file size for third speech information based on the prosodic phoneme sequence; and correcting the predicted file size based on a target residual value and generating the target file size. The target residual value is determined based on a sample file size and a predicted sample audio file size corresponding to the sample text. The sample file size is the actual sample audio file size corresponding to the sample text.

[0143] In some embodiments, the fifth processing module is further used for combining the first speech information and the second speech information based on a phoneme duration corresponding to the first speech information and a phoneme duration corresponding to the second speech information.

[0144] In some embodiments, the apparatus may further include a seventh processing module, used for obtaining a to-be-synthesized text before converting the target text into the prosodic phoneme sequence; segmenting the to-be-synthesized text and generating the target text if a size of the to-be-synthesized text exceeds a target threshold, where the target text size does not exceed the target threshold.

[0145] In some embodiments, the first processing module 310 is further used for: determining a first segmentation position based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence; determining a second segmentation position from positions of the prosodic identifier located after the first segmentation position in the prosodic phoneme sequence; and segmenting the prosodic phoneme sequence based on the first segmentation position and the second segmentation position to generate the first sub prosodic phoneme sequence and at least two second sub prosodic phoneme sequences. The first sub prosodic phoneme sequence is the prosodic phoneme sequence located before the first segmentation position in the prosodic phoneme sequence, and at least two second sub prosodic phoneme sequences are the prosodic phoneme sequences located after the first segmentation position in the prosodic phoneme sequence. The adjacent second sub prosodic phoneme sequences are determined based on the second segmentation position, and a speech synthesis duration corresponding to the first sub prosodic phoneme sequence is within a target duration.

[0146] In some embodiments, the first processing module 310 is further used for: obtaining a prosodic word, a syllable, a prosodic phrase, end-of-sentence information and an intonational phrase of the target text; and labeling the target text based on at least two of the prosodic word, the syllable, the prosodic phrase, the end-of-sentence information and the intonational phrase to generate the prosodic phoneme sequence.

[0147] In some embodiments, the first processing module 310 is further used for: converting the target text into the phoneme sequence; generating a plurality of prosodic identifiers based on at least two of the prosodic word, the syllable, the prosodic phrase, the end-of-sentence information, and the intonational phrase; and labeling the phoneme sequence based on the plurality of prosodic identifiers to generate the prosodic phoneme sequence.

[0148] FIG. 4 illustrates a schematic physical structural diagram of an electronic device. As shown in FIG. 4, the electronic device may include: a processor 410, a communications interface 420, a memory 430, and a communication bus 440, among which the processor 410, the communication interface 420, and the memory 430 communicate with each other through communication bus 440. The processor 410 may call logical an instruction in the memory 430 to execute a speech synthesis method, which includes: segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences, where the prosodic phoneme sequence includes a plurality of phonemes corresponding to the target text and prosodic identifiers located between adjacent phonemes, and each clause sequence includes at least one phoneme; performing speech synthesis on a first sub prosodic phoneme sequence in plurality of clause sequences to obtain first speech information; and outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in plurality of clause sequences to generate second speech information, where the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

[0149] If the integrated unit is implemented in the form of a software functional unit and sold or used as an independent product, the above-mentioned memory 430 may be stored in a computer readable storage medium. Based on such understanding, the solutions of the present application in essence or a part of the solutions that contributes to the related art, or all or part of the solutions, may be embodied in the form of a software product, which is stored in a storage medium, including several instructions to cause a computer device (which may be a personal computer, server, or network device, etc.) or a processor to perform all or part of the steps of the

methods described in the respective embodiments of the present application. The storage medium described above includes various media that may store program codes such as flash disk, mobile hard disk, read-only memory (ROM), random access memory (RAM), magnetic disk, or optical disk.

[0150] The present application further provides a computer program product, including a computer program that may be stored in a non-transient computer-readable storage medium. When the computer program is executed by a processor, the computer may implement the speech synthesis method provided by the aforementioned embodiments.

[0151] The present embodiment further provides a non-transient computer-readable storage medium on which a computer program is stored, and the computer program is executed by a processor to perform the speech synthesis method provided by the aforementioned embodiments.

[0152] The following describes a speech splicing method of the present embodiment in conjunction with FIG. 5 to FIG. 6.

[0153] As shown in FIG. 5, the speech splicing method includes step 510, step 520, and step 530.

[0154] Step 510: segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences, where the prosodic phoneme sequence includes a prosodic identifier located between adjacent phonemes and a plurality of phonemes corresponding to the target text, and each clause sequence includes at least one phoneme.

[0155] In some embodiments, step 510 may include: segmenting the prosodic phoneme sequence based on at least partial of the plurality of prosodic identifiers to generate a plurality of clause sequences.

[0156] In this embodiment, an entire prosodic phoneme sequence includes a plurality of phonemes and a plurality of prosodic identifiers, where the plurality of prosodic identifiers include prosodic identifiers corresponding to different fine-grained levels.

[0157] In an actual application process, an appropriate fine-grained level may be selected as a segmentation standard based on actual situation, and a position of the corresponding prosodic identifier in the prosodic phoneme sequence may be used as a segmentation point for segmenting the prosodic phoneme sequence to obtain the plurality of clause sequences.

[0158] It should be noted that each clause sequence includes a prosodic identifier at the segmentation point and at least one phoneme.

[0159] It may be understood that, for each prosodic phoneme sequence, at least one segmentation point may obtain at least two clause sequences.

[0160] For example, for a prosodic phoneme sequence "sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil", in case that segmentation is determined at #3 based on actual needs, segmentation is performed at positions containing #3 in the prosodic phoneme sequence, and the prosodic separator #3 is retained to the previous splicing unit, so that the prosodic phoneme sequence may be segmented into the following plurality of clause sequences:

clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3; and clause sequence 2: dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0161] In other embodiments, a first segmentation position and a second segmentation position may be determined based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence; the prosodic phoneme sequence is segmenting based on the first segmentation position and the second segmentation position to generate a clause sequence corresponding to a first sub prosodic phoneme sequence and a clause sequence corresponding to a second sub prosodic phoneme sequence.

[0162] A speech synthesis duration corresponding to the first sub prosodic phoneme sequence is within a target duration. For example, the target duration may be determined based on at least one of computing power of the system and an upper limit of capability of a speech synthesis model. For example, the target duration is a shorter duration, a value of the target duration may be customized by a user or a system default value may be used, for example the target duration is set to 0.2 s or 0.3 s.

[0163] It is possible to continue segmenting the second sub prosodic phoneme sequence based on the prosodic identifier to obtain a clause sequence, for example, segmenting the second sub prosodic phoneme sequence at prosodic separator #3. It may be understood that when there is no prosodic separator #3 in the second prosodic phoneme sequence, the segmentation of the second prosodic phoneme sequence will not continue.

[0164] In the above embodiment, the speech synthesis duration corresponding to the first sub prosodic phoneme sequence obtained based on the first segmentation position may be within a reasonable time range, thereby shortening first sentence response time and delay time of the synthesis system. In addition, the first segmentation position and the second segmentation position determined based on this method are the position with a longer pause time, which makes the pause and prosody of the clause sequence obtained by segmentation more natural, thereby making the subsequent output of speech synthesized based on the clause sequence more natural and smooth.

[0165] Step 520: performing speech synthesis on each clause sequence to generate a plurality of pieces of first clause speech information.

[0166] In this step, the first sentence speech information is speech information generated by performing speech synthesis on clause sequences, and each clause

sequence corresponds to first sentence speech information.

[0167] The first clause speech information includes a first duration corresponding to each prosodic identifier and phoneme.

[0168] It should be noted that the first clause speech information may be either speech or an advanced acoustic feature.

[0169] The advanced acoustic feature is physical quantities used to represent speech acoustic characteristics and may be used to reconstruct speech, including but not limited to: a linear spectrum, a Mel spectrum, a Mel cepstral, an energy concentration zone, a resonance peak frequency, a resonance peak intensity and bandwidth of timbre, as well as a duration, a fundamental frequency, an average speech power representing prosodic characteristics of speech, etc.

[0170] The phoneme may be a combination of one or more phonetic units segmented based on natural properties of speech. The phonetic unit may be a Pinyin, an initial, or a final corresponding to a Chinese character, or an English word, an English phonetic symbol, or an English letter.

[0171] The first duration refers to an articulation duration corresponding to the prosodic identifier or the phoneme.

[0172] For example, for clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3, "shang4" may be used as one phoneme or split into two phonemes, "sh" and "ang4", with each prosodic identifier or phoneme corresponding to a articulation duration.

[0173] In an actual application process, all prosodic identifiers and phonemes in the first clause sequence may be obtained first, and then the first duration corresponding to each prosodic identifier and phoneme may be obtained based on each prosodic identifier and phoneme.

[0174] In some embodiments, step 520 may include: inputting the clause sequence into a target speech synthesis model to obtain the first clause speech information output from the target speech synthesis model, where the target speech synthesis model is trained with a sample prosodic phoneme sequence as a sample and a sample clause speech corresponding to the sample prosodic phoneme sequence as a sample label.

[0175] In this embodiment, the target speech synthesis model may be an end-to-end speech synthesis model.

[0176] The clause sequence is obtained by segmenting an entire prosodic phoneme sequence into a plurality of segments.

[0177] An input value of the target speech synthesis model is a clause sequence, and an output value thereof is first clause speech corresponding to the clause sequence, or an advanced acoustic feature corresponding to the first clause speech.

[0178] The target speech synthesis model is trained by taking a sample clause sequence as a sample and taking corresponding sample clause speech as a sample label.

[0179] Training of the target speech synthesis model is similar to the training mode of a neural network model, and will not be elaborated here.

[0180] As shown in FIG. 6, in an actual application process, each clause sequence may be converted into a prosodic phoneme sequence that may be received by the end-to-end speech synthesis model, and first durations corresponding to each phoneme and each prosodic identifier in the prosodic phoneme sequence may be obtained based on the prosodic phoneme sequence.

[0181] For example, the clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3, is converted into a prosodic phoneme sequence 1 that may be received by the end-to-end speech synthesis model: sil sh ang4 #0 h ai3 #0 sh i4 #2 j in1 #0 t ian1 #2 y in1 #0 zh uan3 #1 d uo1 #0 y vn2 #3 sil eos.

[0182] Speech synthesis is performed on each prosody and phoneme in the prosodic phoneme sequence 1, speech or the advanced acoustic feature corresponding to each prosody and phoneme are synthesized to generating the first clause speech or the advanced acoustic feature corresponding to the first clause speech. The first durations corresponding to each prosodic identifier and phoneme are calculated.

[0183] In some embodiments of the present application, after step 520 and before step 530, the method may further include: outputting first clause speech information.

[0184] In this embodiment, the first sentence speech information may be output after the first sentence speech information corresponding to the clause sequence is generated.

[0185] In this embodiment, by segmenting the target text into the plurality of clause sequences and performing speech synthesis on each clause sequence, the first speech information corresponding to each clause sequence is generated, and the first speech information corresponding to the first clause sequence in the target text is first output, which effectively accelerates a feedback speed of the system after receiving a network speech synthesis service request, shortens waiting time of users, and conveniently improves the user experience.

[0186] Step 530: splicing a plurality of pieces of first clause speech information based on the first duration and an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the first clause speech information to generate target speech.

[0187] In this step, the target speech is speech obtained by performing speech synthesis on the target text.

[0188] During splicing, it is necessary to first truncate the clause speech information corresponding to each clause sequence based on a redundant first duration at the beginning or end of each clause sequence.

[0189] In case where the first clause speech information is first clause speech, after the first clause speech is synthesized, the speech corresponding to the duration of the redundant phonemes at the beginning or end of the

clause sequence in the first clause is truncated, and the truncated first clause speech is generated. The adjacent truncated first clause speech is sequentially spliced until all the truncated first clause speech is spliced to generate the target speech.

[0190] In case where the first clause speech information is an advanced acoustic feature corresponding to the first clause speech, after the advanced acoustic feature corresponding to the first clause speech is synthesized, the advanced acoustic features corresponding to the duration of the redundant phonemes at the beginning or end of the clause sequence among the advanced acoustic features corresponding to the first clause speech are truncated, and the truncated advanced acoustic feature corresponding to the first clause speech is generated. Speech synthesis is performed on the advanced acoustic feature corresponding to the truncated first clause speech using a vocoder, to generate the truncated first clause speech.

[0191] Based on the order of segmenting the target text into each clause sequence, the adjacent truncated first clause speech is sequentially spliced until all the truncated first clause speech is spliced to generate the target speech.

[0192] In some embodiments, step 530 may include: truncating the speech corresponding to the target phoneme in the first clause speech information based on the first duration corresponding to the target phoneme in a plurality of phonemes to generate second clause speech information; and splicing the second clause speech information based on the order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the second clause speech information to generate the target speech.

[0193] In this embodiment, the speech corresponding to the target phoneme is the redundant phoneme in the clause sequence, including but not limited to the silent phoneme corresponding to the beginning or end of the clause sequence.

[0194] The second clause speech information is speech information generated by truncating the redundant pause or silence duration from the first clause speech information.

[0195] The second speech information may include speech or advanced acoustic feature.

[0196] The content of the second speech information corresponds to the content of the first speech information.

[0197] For example, the first speech information generated by performing speech synthesis on prosodic phoneme sequence 1: sil sh ang4 #0 h ai3 #0 sh i4 #2 j in1 #0 t ian1 #2 y in1 #0 zhuan3 #1 d uo1 #0 y vn2 #3 sil eos includes speech information corresponding to duration such as sil and eos. This speech information is redundant speech information such as silence or pause. The redundant duration corresponding to the sil and eos of the first clause speech information may be truncated based on the first duration corresponding to the target phonemes "sil" and "eos" at the end of the sentence to generate the second clause speech information.

[0198] The second clause speech information corresponding to adjacent clause sequences are sequentially spliced starting from the first clause sequence based on the order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the second clause speech information until all the clause sequences corresponding to the second clause speech information are spliced.

[0199] It should be noted that for different forms of second speech information, the splicing processes corresponding to different forms of second speech information also are different. Specific explanations will be provided in subsequent embodiments, and will not be elaborated here.

[0200] In the speech splicing method provided by the embodiments of the present application, after the target text is segmented into the plurality of clause sequences and the first clause speech information corresponding to each clause sequence is synthesized, speech corresponding to the redundant phonemes in the first clause speech information are truncated based on the duration of the phonemes corresponding to each clause sequence, thereby improving the naturalness and fluency of adjacent first clause speech information splicing without a preset speech splicing unit library and smooth processing of the spliced speech units.

[0201] Below are specific explanations of the implementation of step 530 from two perspectives.

1. The clause sequence corresponding to the first sentence speech information is not the first clause sequence in the target text.

[0202] Referring to FIG. 2, in some embodiments, the target phoneme includes at least one of a beginning-of-sentence redundant phoneme and an end-of-sentence redundant phoneme. Truncating the speech corresponding to the target phoneme in the first clause speech information based on the first duration corresponding to the target phoneme among the plurality of phonemes may include: determining that the clause sequence corresponding to the first sentence speech information is not the first clause sequence in the target text, then truncating the speech corresponding to the end-of-sentence redundant phoneme and the speech corresponding to the beginning-of-sentence redundant phoneme from the first sentence speech information.

[0203] This embodiment is explained by continuously taking the target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4" as an example.

[0204] The target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4" is converted into:

prosodic phoneme sequences 1 that may be re-

ceived by the end-to-end speech synthesis model: sil sh ang4 #0 h ai3 #0 sh i4 #2 j in1 #0 t ian1 #2 y in1 #0 zh uan3 #1 d uo1 #0 y vn2 #3 sil eos; and

prosodic phoneme sequences 2 that may be received by the end-to-end speech synthesis model: sil d ong1 #0 n an2 #0 f engl #2 s an1 #0 d ao4 #1 s i4 #0 j i2 #4 sil eos.

[0205] The prosodic phoneme sequences 1 that may be received by the end-to-end speech synthesis model is the first clause sequence in the target text, and the prosodic phoneme sequences 2 that may be received by the end-to-end speech synthesis model is not the first clause sequence in the target text.

[0206] After the prosodic phoneme sequences 2 that may be received by the end-to-end speech synthesis model is synthesized into the first speech information, speech or advanced acoustic feature the corresponding to the duration is truncated at the beginning based on the duration corresponding to the redundant phonemes at the beginning and end of the sentence in prosodic phoneme sequences 2 that may be received by the end-to-end speech synthesis model, that is, based on the first duration of "sil" at the beginning of the sentence, and the speech or advanced acoustic feature corresponding to the duration is truncated at the end based on the first duration of "sil" and "eos" at the end of the sentence, to generate the second speech information.

[0207] 2. The clause sequence corresponding to the first sentence speech information is the first clause sequence in the target text.

[0208] Referring to FIG. 6, in other embodiments, the target phoneme includes at least one of the beginning-of-sentence redundant phoneme and the end-of-sentence redundant phoneme. Truncating the speech corresponding to the target phoneme in the first clause speech information based on the first duration corresponding to the target phoneme among the plurality of phonemes may further include: determining the clause sequence corresponding to the first sentence speech information to be the first clause sequence in the target text, and truncating the speech corresponding to the end-of-sentence redundant phoneme from the first sentence speech information.

[0209] This embodiment is explained by continuously taking the target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4" as an example.

[0210] The target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4" is converted into:

prosodic phoneme sequences 1 that may be received by the end-to-end speech synthesis model: sil sh ang4 #0 h ai3 #0 sh i4 #2 j in1 #0 t ian1 #2 y in1 #0 zh uan3 #1 d uo1 #0 y vn2 #3 sil eos; and

prosodic phoneme sequences 2 that may be received by the end-to-end speech synthesis model: sil d ong1 #0 n an2 #0 f eng1 #2 s an1 #0 d ao4 #1 s i4 #0 j i2 #4 sil eos.

[0211] The prosodic phoneme sequences 1 that may be received by the end-to-end speech synthesis model is the first clause sequence in the target text, and the prosodic phoneme sequences 2 that may be received by the end-to-end speech synthesis model is not the first clause sequence in the target text.

[0212] After the prosodic phoneme sequence 1 that may be received by the end-to-end speech synthesis model is synthesized into the first speech information, the second speech information may be generated by truncating the corresponding duration speech or advanced acoustic feature at the end based on the duration of the end-of-sentence redundant phoneme in the remaining phonemes at the end of the sentence, that is, based on the first duration of "sil" and "eos".

[0213] In the speech splicing method provided by the embodiments of the present application, after the target text is segmented into the plurality of clause sequences and the first clause speech information corresponding to each clause sequence is synthesized, speech corresponding to the redundant phonemes in the first clause speech information are truncated based on the duration of the phonemes corresponding to each clause sequence, thereby improving the naturalness and fluency of adjacent first clause speech information splicing without a preset speech splicing unit library and smooth processing of the spliced speech units.

[0214] A method for obtaining audio and video file size in the embodiments of the present application is described below in conjunction with FIG. 7 to FIG. 8.

[0215] As shown in FIG. 7, the method for obtaining the audio and video file size includes step 710, step720, and step 730.

[0216] Step 710: obtaining a target text.

[0217] In this step, the target text is a text currently used for speech synthesis.

[0218] A text level of the target text may be tens to hundreds of levels of regular text, as well as thousands or tens of thousands of levels of super text.

[0219] The target text may be a local file stored in a database or a file downloaded from a network, and is not limited in the present application.

[0220] Step 720: extracting a feature from the target text to generate a target prosodic feature and a target phoneme feature.

[0221] In this step, the target prosodic feature is used to represent the prosodic feature of the target text, and the target phoneme feature is used to represent the phoneme feature of the target text.

[0222] The target prosodic feature and the target phoneme feature include but not limited to: phonemes and their corresponding tones, syllables, prosodic words, prosodic phrases, intonational phrases, silence, and pauses.

[0223] In some embodiments, step 720 may include:

converting the target text into a prosodic phoneme sequence, where the prosodic phoneme sequence includes a prosodic identifier located between adjacent phonemes and a plurality of phonemes corresponding to the target text; and extracting a feature of the prosodic phoneme sequence to generate the target prosodic feature and the target phoneme feature.

[0224] In some embodiments, converting the target text into the prosodic phoneme sequence may include: converting the target text into a phoneme sequence; obtaining end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word, and the syllable of the phoneme sequence; and labeling the phoneme sequence based on at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word, and the syllable, to generate the prosodic phoneme sequence.

[0225] After the prosodic phoneme sequence is obtained, the prosodic feature and the phoneme feature are extracted from the prosodic phoneme sequence, then the target prosodic feature and the target phoneme feature may be generated.

[0226] In some embodiments, the target prosodic feature and the target phoneme feature may include at least one of: a length of the prosodic phoneme sequence, the number of Chinese Pinyin in the prosodic phoneme sequence, the number of pauses in the prosodic phoneme sequence, the number of English phonemes in the prosodic phoneme sequence, the number of Chinese initials in the prosodic phoneme sequence, the number of Chinese finals in the prosodic phoneme sequence, and English phonemes of any category in the prosodic phoneme sequence.

[0227] The length of the prosodic phoneme sequence may be the number of phonemes in the prosodic phoneme sequence.

[0228] Step 730: obtaining a target file size of a target audio file based on the target prosodic feature and the target phoneme feature.

[0229] In this step, the target audio file is an audio file generated by performing speech synthesis on the entire target text.

[0230] It may be understood that for audio files, the target audio file is that audio file; for video files, the target audio file is the audio file included in the video file.

[0231] The target file size is a predicted file size of the target audio file.

[0232] The target file size may be file volume information or length information of the third speech information, and no limitation is made in the present application.

[0233] In some embodiments, step 730 may include: obtaining a first predicted file size of the target audio file based on the target prosodic feature and the target phoneme feature; summing a target residual value and the first predicted file size to generate the target file size.

[0234] In this embodiment, the first predicted file size is an uncorrected initial file size of the speech synthesized from the target text, which is predicted based on the target prosodic feature and the target phoneme feature.

[0235] The target residual value is used to correct the first predicted file size, therefore improving the accuracy of the final generated target file size.

[0236] The target residual value is determined based on a sample file size and a predicted sample audio file size corresponding to the sample text. The sample file size is the actual sample audio file size corresponding to the sample text.

[0237] The target file size is a file size of the speech synthesized from the target text, which is predicted based on the target prosodic feature and the target phoneme feature, and corrected. It may be understood that the accuracy of the target file size is greater than that of the first predicted file size.

[0238] The target residual value is a predetermined numerical value, for example, the target residual value may be a maximum absolute value of the residual value.

[0239] In this embodiment, residual is supplemented to the first predicted file size to correct the first predicted file size, thereby improving the accuracy of the final generated target file size.

[0240] In an actual application process, a neural network model may be used to predict the first predicted file size.

[0241] By taking the neural network model being a file size prediction model as an example, the generation mode of the first predicted file size in this embodiment will be explained.

[0242] In some embodiments, step 730 may include: inputting the target prosodic feature and the target phoneme feature into a file size prediction model to obtain a first predicted file size output from the file size prediction model.

[0243] In this embodiment, the file size prediction model may be a pre-trained neural network model.

[0244] The file size prediction model is used to predict the file size of the synthesized speech based on the prosodic feature and the phoneme feature of the text.

[0245] Training of the file size prediction model is as follows. The file size prediction model is trained by using a sample prosodic feature and a sample phoneme feature as a sample, and using a sample file size corresponding to a sample prosodic feature and a sample phoneme feature as a sample label.

[0246] The sample prosodic feature and the sample phoneme feature are generated by extracting the prosodic feature and the phoneme feature from the sample text. The mode of extracting the sample prosodic feature and the sample phoneme feature is similar to the mode of extracting the target prosodic feature and the target phoneme feature mentioned above, and will not be elaborated here.

[0247] The sample file size corresponding to the sample prosodic feature and the sample phonemic feature is an actual size of the sample audio file generated by performing speech synthesis on the sample text.

[0248] In an actual application process, the target pro-

sodic feature and the target phoneme feature are input into the trained file size prediction model, and the file size prediction model may output the initial file size, i.e., the first predicted file size, corresponding to the speech generated by performing speech synthesis on the target text corresponding to the target prosodic feature and the target phoneme feature.

[0249] After the first predicted file size is obtained, the sum of the first predicted file size and the target residual value is calculated to generate the target file size.

[0250] In this embodiment, computational efficiency in actual performing may be improved using a pre-trained model for obtaining the first predicted file size.

[0251] In addition, the target prosodic feature and the target phoneme feature corresponding to each target text in the actual application may be used as training samples for subsequently training the file size prediction model. As the training sample volume increases, the intelligence level of the file size prediction model will also continue to improve, and the final predicted results will be more accurate.

[0252] The mode of determining the target residual value will be explained below through specific embodiments.

[0253] In some embodiments, the target residual value is determined by the following steps:

obtaining the sample text, the sample file size corresponding to the sample audio file, and the sample audio file corresponding to the sample text, where the sample audio file is generated by performing speech synthesis on the sample text;

extracting a feature from the sample text to generate a sample prosodic feature and a sample phonemic feature;

obtaining a second predicted file size of the sample audio file based on the sample prosodic feature and the sample phoneme feature; and

determining a maximum absolute value of a difference between the second predicted file size and the sample file size as the target residual value.

[0254] In this embodiment, the text level of the sample text may be tens to hundreds of levels of regular text, as well as thousands or tens of thousands of levels of super text.

[0255] The sample audio file is the audio file generated by performing speech synthesis on the sample text.

[0256] The sample file size is an actual size or an actual audio duration of the sample audio file.

[0257] For example, a speech synthesis system may be used to calculate a true wav file size or audio duration of the sample audio file corresponding to the sample text.

[0258] The second predicted file size is a size or audio duration of the uncorrected sample audio file obtained through prediction.

[0259] It should be noted that the generation mode of the second predicted file size should be consistent with the generation mode of the first predicted file size.

[0260] In an actual application process, feature may be extracted from the sample text to generate the sample prosodic feature and the sample phoneme feature, and the sample prosodic feature and the sample phoneme feature may be input into the file size prediction model to obtain the second predicted file size output from the file size prediction model.

[0261] The maximum absolute value of the difference between the second predicted file size and the sample file size is then calculated and taken as the target residual value.

[0262] It may be understood that in an actual application process, the sample prosodic feature and the sample phoneme feature may be predicted for many times to obtain a plurality of second predicted file sizes. The difference between each second predicted file size and the sample file size is then calculated to obtain a plurality of candidate differences; then, the absolute value of the minimum non-positive value is selected from the plurality of candidate differences as the target residual value, which improve the accuracy of the target residual value.

[0263] In the method for obtaining audio and video file size provided by the embodiments of the present application, by extracting the prosodic feature and the phoneme feature from the target text, and predicting the size of the target audio file synthesized from the target text based on the extracted target prosodic feature and target phoneme feature, the target file size may be predicted before the target audio file is generated, which has a certain timeliness, and the accuracy and precision of the prediction results are relatively high.

[0264] As shown in FIG. 8, according to some embodiments of the present application, after step 730, the method may further include: segmenting the target text based on target prosodic feature and the target phoneme feature to generate a plurality of clause sequences; performing speech synthesis on the clause sequence to generate clause speech; and outputting clause speech and the target file size, and splicing the clause speech to generate the target audio file.

[0265] In this embodiment, each clause sequence includes at least one phoneme, which may be either a Chinese phoneme or an English phoneme.

[0266] The target text is segmented based on at least one feature from the target prosodic features, including the syllable, the prosodic word, the prosodic phrase, and the intonational phrase, to obtain at least two clause sequences.

[0267] For example, for the target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4", it may be first converted into a prosodic phoneme sequence: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil;

the prosodic phoneme sequence is then segmented at #3 into the following plurality of clause sequences:

clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3; and clause sequence 2: dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0268] Speech synthesis is performed on the clause sequence with the highest segmentation order among the plurality of clause sequences to generate the clause speech corresponding to that clause sequence; and the corresponding clause speech and target file size of the clause sequence are output, and subsequent clause sequences are synthesized.

[0269] For example, for sample text: please search for detailed information on the APP, the sample text may be converted into sample prosodic phoneme sequence: sil xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #2 zai4 #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil;
then features are extracted from the sample prosodic phoneme sequence, the extracted target prosodic feature and target phoneme feature includes but not limited to: a length of the sample prosodic phoneme sequence, the number of Chinese Pinyin in the sample prosodic phoneme sequence, the number of pause symbols (#0 #1 #2 #3 sil) in the sample prosodic phoneme sequence, the number of English phonemes in the sample prosodic phoneme sequence, the number of Chinese phonemes in the sample prosodic phoneme sequence, the number of Chinese initials in the sample prosodic phoneme sequence, the number of Chinese finals in the sample prosodic phoneme sequence, and the English phonemes of each category in the sample prosodic phoneme sequence (vowels, diphthongs, R colored finals, stops, affricates, fricatives, nasals, liquids, semivowels).

[0270] After training data is prepared, a wav file size prediction model based on ElasticNet regression model may be trained.

[0271] The obtained sample prosodic feature and sample phoneme feature are input into the WAV file size prediction model, and a target output during the training is the actual number of WAV file bytes of the sample audio file.

[0272] Cross validation may be used to select the best model parameter, and then the ElasticNet regression model may be trained with the selected parameter.

[0273] The target residual value is then calculated, for example, by using sample prosodic feature and sample phoneme feature as inputs of the model, to obtain the second predicted file size.

[0274] The absolute value of the minimum non-positive value of the second predicted file size minus the sample file size is calculated and taken as the maximum residual value.

[0275] In an actual application process, a client initiates a request. For example, the target text: Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4 is obtained.

[0276] The system responds to the request by extracting the target prosodic feature and the target phoneme feature from the target text requested by the client.

[0277] The extracted target prosodic feature and the target phoneme feature are input into the model described above for obtaining the first predicted file size.

[0278] Residual is supplemented to the first predicted file size, and the generated target file size is the sum of the first predicted file size and the target residual value.

[0279] The generated target file size is used as the predicted WAV file size.

[0280] The predicted WAV file size is written into a header of the WAV file.

[0281] The target text requested by the client is then segmented to generate a plurality of clause sequences, such as:

a first clause sequence: overcast to cloudy today in Shanghai; and
a second clause sequence: with southeast wind of a scale from 3 to 4.

[0282] The audio of the first clause sequence "overcast to cloudy today in Shanghai" is synthesized, the first clause speech is generated and written into the wav file, and it is returned to the client.

[0283] The audio after the first clause sequence is then synthesized in order and written into the wav file until all requests are synthesized. For example, the audio of "with southeast wind of a scale from 3 to 4" is synthesized and written into the wav file, then the process ends.

[0284] For example, in case where the file size is represented by duration, and the sample text is "may control", which may be converted into a prosodic phoneme sequence: sil k e2 #0 y i3 #1 k ong4 #0 zh i4 #3 sil eos, and a duration of prosody and phonemes (Mel spectrum frame number) are predicted: 3 1 3 1 1 6 2 2 7 4 5 11 4 12, and the total duration of the phonemes are taken as the sample file size.

[0285] In subsequent model training, the model may be set as a 1-layer 256-dimensional embedding layer, connected to 4 layers of a 1-dimensional convolutional neural network with 256 channels, connected to layer norm, connected to dropout, and connected to a fully connected layer with an output dimension of 1.

[0286] A phoneme duration sequence d is then converted to log domain, where $d' = \log(d + 1)$.

[0287] A loss function may include an MSE loss of the phoneme duration sequence and an average total duration MAE loss of each phoneme.

[0288] An Adam optimizer is then used to iteratively optimize the model.

[0289] When the target residual value is calculated, the predicted total Mel spectrum frame number obtained based on the above model may be taken as the second predicted file size, and then the maximum residual value is calculated.

[0290] In an actual application process, a client initiates a request. For example, the target text: Overcast to cloudy today in Shanghai with southeast wind of a scale

from 3 to 4 is obtained.

[0291] The system responds to the request by extracting the target prosodic feature and the target phoneme feature from the target text requested by the client.

[0292] The extracted target prosodic feature and the target phoneme feature are input into the model described above for obtaining the first predicted file size.

[0293] Residuals is supplemented to the first predicted file size, and the generated target file size is the sum of the first predicted file size and the target residual value.

[0294] It should be noted that in this embodiment, the Mel spectrum frame number is calculated, and the audio duration (total Mel spectrum number) is converted to the WAV file size based on a Mel spectrum frame shift and a sampling frequency of the WAV file (16000), the number of sampling bits (16), and the number of channels (1), where the wav file size is (Mel spectrum frame number × Mel spectrum frame shift / 16000) * 16000 * 16 * 1 / 8 + 44) bytes.

[0295] In to the method for obtaining audio and video file size provided by the embodiments of the present application, by extracting the prosodic feature and the phoneme feature from the target text, and predicting the size information of the target audio file synthesized from the target text based on the extracted target prosodic feature and the target phoneme feature, the target file size may be predicted before the generation of the target audio file, which has a certain timeliness; and the accuracy and precision of the prediction results are relatively high.

[0296] A text transliteration method provided by embodiments of the present application is described below in conjunction with FIG. 9 to FIG. 10.

[0297] As shown in FIG. 9, the text transliteration method includes step 910 and step920.

[0298] Step 910: segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences.

[0299] In this step, the target text is a text currently used for speech synthesis.

[0300] In some embodiments, a prosodic identifier may include: at least one of identifiers used to represent syllables, prosodic words, prosodic phrases, end-of-sentence information and intonational phrases.

[0301] It may be understood that different prosodic identifiers correspond to different fine-grained levels. The fine grain of a prosodic identifier used to represent pauses is greater than that of a prosodic identifier used to represent intonational phrases, the fine grain of a prosodic identifier used to represent intonational phrases is greater than that of a prosodic identifier used to represent prosodic phrases, the fine grain of a prosodic phrase used to represent prosodic words is greater than that of a prosodic identifier used to represent syllables.

[0302] In an actual application process, different symbols may be used to represent prosodic features of different fine-grained levels.

[0303] For example, for a target text "Overcast to cloudy today in Shanghai with southeast wind of a scale from 3 to 4", it may be converted into a prosodic phoneme sequence: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0304] It may be understood that for the prosodic phoneme sequence, the prosodic identifier may include: # and numbers between adjacent Pinyin, and phonemes may include the corresponding Pinyin and tone of each Chinese character, or English phonetic symbols.

[0305] Sil represents silence at the beginning and the end of a sentence in the prosodic phoneme sequence, #0 represents syllables, #1 represents prosodic words, #2 represents prosodic phrases, #3 represents intonational phrases, and #4 represents the end of a sentence. The number after each phoneme represents the tone of that phoneme. For example, in shang4, 4 represents the fourth tone of the Pinyin "shang". An order of the fine grain in an ascending order is: #0 < #1 < #2 < #3 < #4.

[0306] For each prosodic phoneme sequence, at least one segmentation point may obtain at least two clause sequences.

[0307] In some embodiments, step 910 may include: converting the target text into a prosodic phoneme sequence; segmenting the prosodic phoneme sequence based on at least partial of a plurality of prosodic identifiers to generate a plurality of clause sequences.

[0308] In this embodiment, an entire prosodic phoneme sequence includes a plurality of phonemes and a plurality of prosodic identifiers, where the plurality of prosodic identifiers include prosodic identifiers corresponding to different fine-grained levels.

[0309] In an actual application process, an appropriate fine-grained level may be selected as a segmentation standard based on actual situation, and a position of the corresponding prosodic identifier in the prosodic phoneme sequence may be used as a segmentation point to segment the prosodic phoneme sequence to obtain the plurality of clause sequences.

[0310] It should be noted that each clause sequence includes a prosodic identifier at the segmentation point and at least one phoneme.

[0311] It may be understood that, for each prosodic phoneme sequence, at least two clause sequences corresponding to at least one segmentation point may be obtained.

[0312] For example, for a prosodic phoneme sequence "sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3 dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil", based on actual needs, it is determined that the prosodic phoneme sequence is segmented at #3. Then, segmentation is performed at positions containing #3 in the prosodic phoneme sequence, while retaining a prosodic separator #3 to the previous splicing unit, so that the prosodic phoneme sequence may be segmented into the following clause sequences:

clause sequence 1: sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3; and clause sequence 2: dong1 #0 nan2 #0 fengl #2 san1 #0 dao4 #1 si4 #0 ji2 #4 sil.

[0313] In some embodiments, step 910 may further include:

segmenting the prosodic phoneme sequence based on a target identifier from a plurality of prosodic identifiers to generate a plurality of candidate sequences;
combining the target candidate sequence from a plurality of candidate sequences with adjacent candidate sequences to generate the fine grain size corresponding to the clause sequence and generate a plurality of clause sequences; and
ordering the plurality of clause sequences in descending order based on the corresponding fine grain size.

[0314] In this embodiment, the prosodic phoneme sequence may be segmented into the plurality of candidate sequences based on a position corresponding to the target identifier. A speech synthesis duration corresponding to the candidate sequence located before the first segmentation point position is within the target duration.

[0315] The speech synthesis duration is the time that it takes to synthesize the candidate sequence into speech.

[0316] The target duration is a shorter duration, a value of the target duration may be customized by a user or a system default value may be used, such as setting the target duration to 0.2 s or 0.3 s.

[0317] For example, for a prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3 tiao2 #0 jie2 #1 wen1 #0 du4 #3 ding4 #0 shi2 #1 kail #0 guan1 #3 xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil", the prosodic identifier corresponding to "yi3" and all prosodic identifiers corresponding to "#3" after "yi3" may be determined as the target identifiers, and the following segmentation sequence is generated:

sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1| kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3| tiao2 #0 jie2 #1 wen1 #0 du4 #3 |ding4 #0 shi2 #1 kail #0 guan1 #3 |xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil,
where "|" is the segmentation point corresponding to the target identifier.

[0318] The target candidate sequence may be any one of the candidate sequences among multiple candidate sequences. The target candidate sequence is combined with adjacent other candidate sequences to generate a plurality of combined clause sequences, where the plurality of clause sequence includes the original candidate sequences and the original prosodic phoneme sequences corresponding to the target text.

[0319] It may be understood that, the fine-grained level of the candidate sequence is greater than that of the target candidate sequence.

[0320] For example, in the prosodic phoneme sequence "sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil" of sentence "Hope this song | may make you like it | play for you | XX", "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3", "neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1", "wei4 #0 nin2 #1 bol #0 fang4 #1", and "EH1 K S: #10 EH1 K S #0 de5 #1 EH1 K S #4" are all candidate sequences.

[0321] Any one candidate sequence from the candidate sequences is taken as the target candidate sequence, and combined with adjacent other candidate sequences to obtain the following plurality of clause sequences (the one between two "|" is a clause sequence):

"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";
"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";
"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";
"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 | EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 si1";
"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 | wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil";
"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 | neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil"; and
"sil xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 sil".

[0322] It may be understood that each clause sequence corresponds to a fine-grained level, and the more candidate sequences included in the clause sequence, the greater the corresponding fine-grained level. For example, the fine-grained level corresponding to "Hope this song may make you like it" is greater than the fine-

grained level corresponding to "Hope this song".

[0323] The plurality of clause sequences are ordered based on the corresponding fine grain size. The higher the fine grain size, the higher the corresponding clause sequence ordered. For example, "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 (Hope this song may make you like it play for you XX's X)" is ordered ahead of "xi1 1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1" (Hope this song may make you like it play for you).

[0324] In this embodiment, the prosodic phoneme sequence is segmented based on the predicted prosodic results of semantics and human speaking habits and separating at longer pauses, rather than simply segmented based on punctuations, which helps to improve the naturalness of the target speech generated by the subsequent splicing of multiple clause speech.

[0325] Step 920: determining that any one of to-be-matched clause sequences among the plurality of clause sequences matches a target clause sequence in a cache, obtaining the target clause speech corresponding to the target clause sequence from the cache, and determining the speech corresponding to the target clause sequence as the target clause speech.

[0326] In this step, the target clause sequence is a pre-generated and stored clause sequence in the system.

[0327] The target clause sequence may be any one of the stored clause sequences cached in the system.

[0328] The target clause speech is the speech generated by performing speech synthesis on the target clause sequence, and the target clause speech is stored in the system and has a correspondence with the target clause sequence.

[0329] In an actual application process, the cached target clause sequence may be precisely matched with multiple clause sequences to determine that the target clause sequence matches any one of the multiple to-be-matched clause sequences. Then, the target clause sequence is determined as the first clause sequence, and the target clause speech corresponding to the target clause sequence is obtained from the cache.

[0330] After the clause sequence is determined as the first clause sequence, the target clause speech corresponding to the target clause that matches the first clause sequence may be directly determined as the speech corresponding to the first clause sequence.

[0331] In some embodiments, the to-be-matched clause sequence includes any one of multiple clause sequences and combinations between different clause sequences.

[0332] As shown in FIG. 10, in some embodiments, step 920 may include: precisely matching the plurality of clause sequences with the target clause sequence from top to bottom in a descending order in step 910.

[0333] The plurality of clause sequences are matched with the target clause sequence in the cache from top to bottom in the descending order, and the speech corresponding to the matched target clause sequence is determined as the speech corresponding to the clause sequence.

[0334] In this embodiment, the clause sequence is precisely matched with the target clause sequence from top to bottom based on the descending ranking order generated in step 910. For example, the clause sequence "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 EH1 K S #10 EH1 K S #0 de5 #1 EH1 K S #4 (Hope this song may make you like it play for you XX's X)" is first precisely matched with the target clause, and in case of successful matching, the clause sequence is determined as the first clause sequence, and the target clause speech corresponding to the target clause matched with the first clause sequence is determined as the speech corresponding to the first clause sequence, then the comparison ends.

[0335] In case of unsuccessful matching, the clause sequence "xi1 #0 wang4 #1 zhe4 #0 shou3 #0 gel #3 neng2 #0 rang4 #1 ni2 #1 xi3 #0 huan1 #1 wei4 #0 nin2 #1 bol #0 fang4 #1 (Hope this song may make you like it play for you)" is precisely matched with the target clause, and the above process is repeated until it is determined that a certain clause sequence may be precisely matched with the target clause, then the comparison ends.

[0336] In some embodiments, if all clause sequences cannot be precisely matched with the target sentence, speech is generated based on the clause sequence. The specific implementation will be explained in subsequent embodiments and will not be elaborated here.

[0337] In some embodiments, after step 920, the method may further include: outputting the target clause speech.

[0338] In this embodiment, the target clause speech is the speech corresponding to the first clause sequence, and the target clause speech is pre-generated and stored in the cache.

[0339] In an actual application process, if it is determined that the clause sequence is precisely matched with the target clause sequence, the target clause speech corresponding to the target clause sequence that is similar to the first clause sequence is directly determined as the speech corresponding to the first clause sequence, and the target clause speech is output.

[0340] During research, the applicant found that due to the temporal nature of speech, it requires a lot of computing power to synthesize a text. If the request volume for speech synthesis is very large and the text is basically the same, a server will waste a lot of computing power and repeat the same work. A simple mode is to take the to-be-synthesized text as a key, and take the corresponding synthesized audio address as a value, and this set of key and value are stored in the cache. When there is a need for duplicate text synthesis, the corresponding audio may be directly found from the cache to avoid repeatedly using computing power to synthesize the same text.

[0341] However, this mode requires complete matching of the entire sentence, considering that in actual use, there are rarely requests for identical texts (such as only different punctuations, or only changes in certain parts of a sentence), which leads to a low hit rate and thus affects cache efficiency.

[0342] In the present application, by converting the target text into the prosodic phoneme sequence, the pause position and pause duration level of the target text are determined based on the prosodic feature, and the prosodic phoneme sequence is segmented into the plurality of clause sequences based on the prosodic feature. The clause sequence is compared with the cached target clause sequence respectively. Due to that the sequence as a keyword is shorter, it is easier to be hit in cache search, thereby effectively improving hit efficiency; in case where the clause sequence is the same as the target sequence, the target clause speech corresponding to the target clause sequence is directly determined as the speech corresponding to that clause sequence, without new speech synthesis, thereby effectively reducing the computational costs of the server.

[0343] As shown in FIG. 10, in an actual application process, a prosody prediction module and a segmentation module may be used respectively to perform the above steps.

[0344] In the text transliteration method provided by the embodiments of the present application, the prosodic phoneme sequence is segmented into the plurality of clause sequences based on the prosodic feature, and the clause sequence is compared with the cached target clause sequences respectively, which may effectively improve the hit efficiency. when the clause sequence is the same as the target sequence, the target clause speech corresponding to the target clause sequence is directly determined as the speech of the clause sequence, without speech synthesis, thereby improving the efficiency of speech synthesis.

[0345] Continuing with reference to FIG. 10, in some embodiments of the present application, the method may further include: determining that any one of the to-be-matched clause sequences in plurality of clause sequences does not match the target clause sequence; and performing speech synthesis on the to-be-matched clause sequence to generate the second clause speech.

[0346] In this embodiment, the target clause sequence is a pre-generated and stored clause sequence in the system.

[0347] The target clause sequence may be any one of the stored clause sequences cached in the system.

[0348] It is determined that any one of the to-be-matched clause sequences in plurality of clause sequences does not match the target clause sequence, then the to-be-matched clause sequence will be determined as the second clause sequence.

[0349] In an actual application process, the target clause sequence is precisely matched with any to-be-matched clause sequence among the plurality of clause sequences. If the target clause sequence does not be matched with none of them, the to-be-matched clause sequence is determined as the second clause sequence, and speech synthesis is performed on the second clause sequence to generate the second clause speech.

[0350] The second clause is a speech that does not exist in the cache.

[0351] In an actual application process, the clause sequence may be precisely matched with the target clause sequence from top to bottom in the descending order generated in step 910. In case where all clause sequences do not be matched with the target sentence, speech synthesis is performed on clause sequences that have not found similar sequences, to generate the second clause speech.

[0352] In some embodiments, performing speech synthesis on the second clause sequence to generate the second clause speech may include: converting the second clause sequence into a prosodic phoneme sequence that may be received by an end-to-end speech synthesis model, and performing speech synthesis on this prosodic phoneme sequence to generate the second clause speech.

[0353] In this embodiment, the prosodic phoneme sequence is used to represent prosodic information and phoneme information of the second clause sequence.

[0354] The phoneme is the smallest phonetic unit divided based on the natural properties of speech. The speech analysis is performed based on articulation actions in syllables, and each action constitutes a phoneme, which may be either a Chinese phoneme or an English phoneme.

[0355] For example, the second clause sequence may be expressed as sil shang4 #0 hai3 #0 shi4 #2 jin1 #0 tian1 #2 yin1 #0 zhuan3 #1 duo1 #0 yun2 #3, or as: sil sh ang4 #0 h ai3 #0 sh i4 #2 j in 1 #0 t ian1 #2 y in 1 #0 zh uan3 #1 d uo1 #0 y vn2 #3 sil eos and other phoneme sequences in different formats.

[0356] The second clause sequence is input into a speech synthesis system (such as the end-to-end speech synthesis model), and the speech synthesis system will synthesize the second clause speech.

[0357] In an actual application process, the text-to-phoneme module may be used to perform the above operations.

[0358] In this embodiment, phonemes are taken as keywords for caching, which overcomes the drawbacks of being cached as different sentences when the punctuation changes or number writing changes but the articulation is exactly the same. This may achieve standardized caching of the target text and improve caching efficiency.

[0359] Continuing with reference to FIG. 10, in some embodiments of the present application, after generating the second clause speech, the method may further include: segmenting the second clause speech based on the prosodic identifier to generate a plurality of pieces of sub second clause speech; and caching the sub clause

sequence corresponding to the second clause speech and a plurality of pieces of sub second clause speech.

[0360] In this embodiment, the second clause speech may be segmented based on the prosodic identifier in the second clause sequence corresponding to the second clause speech, and the plurality of pieces of sub second clause speech is generated.

[0361] The clause sequence corresponding to each sub second clause speech is the sub clause sequence.

[0362] After the second sub clause speech and its corresponding sub clause sequence are obtained, the sub clause sequence and the second clause speech may be cached in the system as the target clause sequence and its corresponding target clause speech in subsequent queries.

[0363] In some embodiments of the present application, after generating second sub clause speech, the method may further include: splicing the second clause speech and the target clause speech to generate the target speech corresponding to the target text.

[0364] In this embodiment, the target speech is the speech obtained by performing speech synthesis on the target text.

[0365] The target clause speech is the speech that exists in the cache.

[0366] The second clause is a speech that does not exist in the cache.

[0367] It may be understood that the target speech is generated based on at least one of the target clause speech in the cache and the newly generated second clause speech.

[0368] In some embodiments, splicing the second clause speech and the target clause speech may further include: splicing the second clause phoneme and the target clause phoneme based on an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the second clause phoneme, and an order of segmenting prosodic phoneme sequence the clause sequence corresponding to the target clause phoneme in the.

[0369] In this embodiment, starting from the first clause sequence, speech corresponding to the adjacent clause sequences is sequentially spliced based on the order of segmenting the prosodic phoneme sequence into the second clause sequence corresponding to the second clause speech, as well as the n order of segmenting the prosodic phoneme sequence into the first clause sequence corresponding to the target clause speech, until all speech corresponding to the clause sequence are spliced, to generate the target speech.

[0370] After generating the target speech, the target speech may also be output.

[0371] In the text transliteration method provided by the embodiments of the present application, the prosodic phoneme sequence is segmented into the plurality of clause sequences based on prosodic features, and the clause sequences are compared with the cached target clause sequences, which may effectively improve the hit efficiency; speech synthesis is only performed when the clause sequence is different from the target sequence, which effectively reduces the computational pressure on the server and improves the efficiency of speech synthesis.

[0372] A text segmentation method of the present embodiment is described below in conjunction with FIG. 11 to FIG. 13.

[0373] As shown in FIG. 11, the text segmentation method includes step 1110, step 1120, and step 1130.

[0374] Step 1110: converting a target text into a prosodic phoneme sequence.

[0375] In some embodiments, if a target text size exceeds a target threshold, the target text of which a size exceeds the target threshold is segmented. It may be understood that the target text exceeding the target threshold is a text of which speech synthesis duration exceeds the preset range. Therefore, the efficiency of speech synthesis may be improved by segmenting longer texts and then performing speech synthesis on the segmented texts.

[0376] Step 1120: determining a first segmentation position based on positions of a plurality of prosodic identifiers in a prosodic phoneme sequence.

[0377] In this step, the first segmentation position is a position of a segmentation point used for a first segmentation. The prosodic phoneme sequence may be segmented into two sub sequences based on the first segmentation position, and the sub sequence located before the first segmentation position is determined as the first sub prosodic phoneme sequence.

[0378] It should be noted that a speech synthesis duration corresponding to the first sub prosodic phoneme sequence generated based on the first segmentation position is within the target duration.

[0379] The speech synthesis duration is related to a computing power of the speech synthesis system.

[0380] The speech synthesis duration corresponding to the first sub prosodic phoneme sequence is time that it takes to synthesize the first sub prosodic phoneme sequence into speech.

[0381] The target duration is a relatively short duration, and the value of the target duration may be customized by a user or a system default value may be used, such as setting the target duration to 0.2 s or 0.3 s.

[0382] In some embodiments, the plurality of prosodic identifiers may include: at least one of the identifiers used to represent syllables, prosodic words, prosodic phrases, intonational phrases, and end-of-sentence information, where a fine grain of identifiers used to represent the end-of-sentence information is greater than that of identifiers used to represent intonational phrases, a fine grain of identifiers used to represent intonational phrases is greater than that of identifiers used to represent prosodic phrases, a fine grain of identifiers used to represent prosodic phrases is greater than that of identifiers used to represent prosodic words, and a fine grain of identifiers used to represent prosodic words is greater than that of

identifiers used to represent syllables.

[0383] In this embodiment, the prosodic identifier is a symbol used to represent at least one of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word, and the syllable.

[0384] In an actual application process, special symbols and number combinations or specific letter combinations may be used to represent prosodic identifiers, such as using "#0", "#1", "#2", "#3", and "#4" to represent the prosodic identifier respectively. Different combinations represent different fine-grained levels.

[0385] For example, #0 represents syllables, #1 represents prosodic words, #2 represents prosodic phrases, #3 represents intonational phrases, and #4 represents the end of a sentence.

[0386] For example, for target text "Currently, Xiaojia may control the water heater switch, adjust temperature and control the timing switch, please search for detailed information on Jiaju APP", the sample text may be converted into prosodic phoneme sequence: sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3 tiao2 #0 jie2 #1 wen1 #0 du4 #3 ding4 #0 shi2 #1 kail #0 guan1 #3 xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil.

[0387] The prosodic phoneme sequence includes the plurality of prosodic identifiers such as #0, #1, #3, and #4; the fine grain of prosodic identifiers from small to large is: #0 < #1 < #2 < #3 < #4.

[0388] Positions of each prosodic identifier in the prosodic phoneme sequence are obtained, and a position of the first target identifier is determined as the first segmentation position based on the relationship between the speech synthesis duration corresponding to the sub prosodic phoneme sequence before each position in the prosodic phoneme sequence and the target duration, which ensures that the speech synthesis duration corresponding to the first sub prosodic phoneme sequence generated based on the first segmentation position is within the target duration.

[0389] Below will provide a specific explanation of the implementation for this step based on FIG. 12 to FIG. 13.

[0390] In some embodiments, step 1120 may include: determining a prosodic identifier with largest fine grain from the plurality of prosodic identifiers based on a target threshold range; and determining a position of the prosodic identifier with largest fine grain in the prosodic phoneme sequence within the target threshold range as a first segmentation position.

[0391] In this embodiment, the target threshold range ranges from a maximum and a minimum of the element articulation length.

[0392] The element articulation length is the sum of the articulation length of all phonemes before a target position in the prosodic phoneme sequence.

[0393] The target position may be a location of any one prosodic identifier in the location of any prosodic identifier in a prosodic phoneme sequence.

[0394] The target threshold range may be represented by (n, n + m), where values of n and m may be determined based on user customization or algorithm, n and m are both positive integers, and the sum of n and m does not exceed the sum of the articulation length of all phonemes in the prosodic phoneme sequence.

[0395] For example, n may be set to 5 and m may be set to 5, thus, the target threshold range is determined to be 5-10 units of articulation length.

[0396] One Chinese character is one unit of articulation length, one English word that forms a word is two units of articulation length, and a string of English phonemes that do not form a word is four units of articulation length.

[0397] It should be noted that in some embodiments, if there are a plurality of prosodic identifiers with maximum fine grain, the position with the smallest position number is taken, that is, the position of the first prosodic identifier with maximum fine grain in the prosodic phoneme sequence is determined as the first segmentation position.

[0398] A target text "Currently, Xiaojia may control the water heater switch, adjust temperature, and control the timing switch, please search for detailed information on Jiaju APP" is taken as an example for explanation.

[0399] After the target text is converted into prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3 tiao2 #0 jie2 #1 wen1 #0 du4 #3 ding4 #0 shi2 #1 kail #0 guan1 #3 xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil", starting from the first character of the prosodic phoneme sequence, a sequence composed of syllables between the 5th and 10th units of articulation length: "ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0", is selected and determined as the candidate prosodic phoneme sequence within the target threshold range.

[0400] The fine grain corresponding to each prosodic identifier in the candidate prosodic phoneme sequence "ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0" is compared, the position of the prosodic identifier with the largest fine grain is selected and determining as the first segmentation position. For the candidate prosodic phoneme sequence mentioned above, #1 is the prosodic identifier with largest fine grain. Since there are a plurality of #1, the position where the first #1 is located is determined as the first segmentation position, that is, the position after the syllable "yi" is determined as the first segmentation position, as shown below:

> sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1 | kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3 tiao2 #0 jie2 #1 wen1 #0 du4 #3 ding4 #0 shi2 #1 kail #0 guan1 #3 xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil,
> where "|" represents the first segmentation position.

[0401] In some embodiments, determining the proso-

dic identifier with largest fine grain from the plurality of prosodic identifiers based on the target threshold range may include:

obtaining a first articulation length of all phonemes in the target sub prosodic phoneme sequence in the prosodic phoneme sequence,

when the first articulation length is within the target threshold range, determining the target position corresponding to the first articulation length as a candidate segmentation point position, and generating a plurality of candidate segmentation point positions; and

determining the prosodic identifier with largest fine grain from the prosodic identifiers corresponding to the plurality of candidate segmentation point positions.

[0402] In this embodiment, the target sub prosodic phoneme sequence is all prosodic phoneme sequences before a target position in the prosodic phoneme sequence, where the target position is the position of any one prosodic identifier in the prosodic phoneme sequence.

[0403] The first articulation length is the sum of the articulation length corresponding to each phoneme in the target sub prosodic phoneme sequence.

[0404] As shown in FIG. 13, in an actual application process, the first articulation length may be obtained through a first segmentation position searching module.

[0405] For example, the prosodic phoneme sequence is converted into a prosodic phoneme list, and the prosodic phoneme list is input into the first segmentation position searching module.

[0406] The first segmentation position searching module sets an initial value of the current articulation length and an initial value of the list index to 0, initializes an empty prosodic position dictionary dict, and starts a loop based on the following formula:

$$\text{element articulation length} = \text{get\_speech\_length (list index)};$$

$$\text{list index} = \text{List index} + 1;$$

and

first articulation length = first articulation length + element articulation length,

where the element articulation length is a articulation length corresponding to a syllable at the current target position, and the list index is used to represent the target position.

[0407] Function get_voice_length (list index) calculates the element articulation length specified in the index, using the following calculation method: one Chinese character is one unit of articulation length, one English word (an English string in the dictionary) is two units of articulation length, and one English string (not in the dictionary) is four units of articulation length.

[0408] By using the above method, N first articulation length with index values within a range of (1, N) may be obtained, where N is the number of syllables in the list.

[0409] The length of each generated first articulation is compared with the target threshold.

[0410] When the length of the first articulation is greater than the lower threshold in the target threshold range and less than the upper threshold in the target threshold range, the current prosodic identifier and position index are recorded in the dictionary dict.

[0411] In some embodiments, when the first articulation length is within the target threshold range, determining the target position corresponding to the first articulation length as the position of the candidate segmentation point includes:

when the first articulation length is within the target threshold range, determining that the prosodic identifier at the target position corresponding to the first articulation length appears for the first time, and determining the target position corresponding to the first articulation length as a candidate segmentation point position.

[0412] Continuing with the above example, if the first articulation length is greater than the lower threshold in the target threshold range and less than the upper threshold in the target threshold range, determining whether the corresponding prosodic identifier at the current list index is the first identifier. If it is determined that the corresponding prosodic identifier at the current list index is the first identifier, the current prosody and list index as a key value pair are recorded in the dictionary dict.

[0413] In other embodiments, if it is determined that the corresponding prosodic identifier at the current list index is an identifier that has already appeared before, the current list index is skipped and the next loop proceeds.

[0414] For example, for prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3 tiao2 #0 jie2 #1 wen1 #0 du4 #3 ding4 #0 shi2 #1 kail #0 guan1 #3 xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil", the first articulation length corresponding to each list index value is calculated starting from the list index value of 1.

[0415] After the calculated first articulation length corresponding to the position of "ke2 #0" is within the target threshold range, the list index value and the prosodic identifier "#0" at this position are recorded.

[0416] The list index is added by one, and the first articulation length corresponding to the position of "yi3 #1" is calculated. When it is determined that the first articulation length corresponding to the position of "yi3 #1" is within the target threshold range, the list index value and the prosodic identifier "#1" at this position are recorded.

[0417] The list index is added by one, and the first

articulation length corresponding to the position of "kong4 #0" is calculated. When it is determined that the first articulation length corresponding to the position of "kong4 #0" is within the target threshold range, if it is determined that the prosodic identifier "#1" at that position is not the first appearance, the list index is skipped and added by one. The above process is repeated until the first articulation length corresponding to the current list index exceeds the upper threshold of the target threshold range, and the loop ends.

[0418] The list index corresponding to all recorded prosodic identifiers may be used to determine the position of the candidate segmentation point for the first segmentation position.

[0419] The identifier with the largest fine grain prosodic is filtered out from all the recorded prosodic identifiers, such as determining the identifier "#1" with the largest fine grain prosodic from the "#0" and "#1" recorded above, and the target position (i.e., position index) corresponding to this prosodic identifier "#1" is determined as the all-around sub position.

[0420] In other embodiments, if the first articulation length is less than the lower threshold in the target threshold range, the current list index is added one to enter the next loop.

[0421] In some other embodiments, if the list index value exceeds a list element number, the loop ends.

[0422] After the first segmentation position is determined, the prosodic phoneme sequence may be segmented from that position to generate the first sub prosodic phoneme sequence.

[0423] During research, the applicant found that due to the temporal nature of speech, the conversion time of the system is usually proportional to the length of the input text. The longer the sentence, the longer the synthesis time required. Especially for some excessively long text inputs, it may also cause the system capacity to exceed the limit. A simple and direct idea to solve the above problems is to perform parallel synthesis utilizing the capabilities of computer systems, and the first problem faced in the segmentation is how to segment parallel tasks. In the related art, text segmentation is mainly based on punctuations, but this segmentation method cannot solve the segmentation of text without punctuation, nor may it solve the problem of imbalanced ends after segmentation.

[0424] In the present application, by converting the target text into the prosodic phoneme sequence and determining the first segmentation position based on the prosodic characteristics of the prosodic phoneme sequence, the speech synthesis duration corresponding to the first sub prosodic phoneme sequence obtained based on the first segmentation position may be within a reasonable time range, thereby shortening the first sentence response time and delay time of the synthesis system. In addition, the position determined based on this method is the position with a longer pause time, which makes the pause and prosody of the first sub prosodic phoneme sequence obtained by segmentation more natural, thereby making the subsequent output of speech synthesized based on the first sub prosodic phoneme sequence natural and smooth.

[0425] Step 1130: segmenting the prosodic phoneme sequence based on the first segmentation position and the second segmentation position to generate at least the first sub prosodic phoneme sequence and a second sub prosodic phoneme sequence, where the second sub prosodic phoneme sequences is the prosodic phoneme sequences located after the first segmentation position in the prosodic phoneme sequence.

[0426] In this step, the first sub prosodic phoneme sequence is generated based on the first segmentation position, and the first sub prosodic phoneme sequence is the prosodic phoneme sequence located before the first segmentation position in the prosodic phoneme sequence.

[0427] The second sub prosodic phoneme sequence is a prosodic phoneme sequence located after the first segmentation position.

[0428] In some embodiments, after step 1120 and before step 1130, the method may further include: determining the second segmentation position from the position of the prosodic identifier located after the first segmentation position in the prosodic phoneme sequence.

[0429] Step 1130 may include: segmenting the prosodic phoneme sequence based on the first segmentation position and the second segmentation position to generate the first sub prosodic phoneme sequence and at least two second sub prosodic phoneme sequences, at least two second sub prosodic phoneme sequences are the prosodic phoneme sequences located after the first segmentation position in the prosodic phoneme sequence, and the adjacent second sub prosodic phoneme sequences are determined based on the second segmentation position.

[0430] In this embodiment, the second segmentation position is the position of the segmentation points corresponding to all segmentations other than the first segmentation.

[0431] After the first segmentation position is determined, at least a portion of the prosodic identifiers after the first segmentation position are searched from the prosodic phoneme sequence as candidate sets for determining the second segmentation position, and the position of the prosodic identifiers in the candidate set are determined as the second segmentation position.

[0432] As shown in FIG. 12, in an actual application process, a second segmentation position searching module may be used to search for the second segmentation position.

[0433] For example, by inputting the first segmentation position and the prosodic phoneme sequence into the second segmentation position searching module, a segmentation point list output from the second segmentation position searching module is obtained, and the segmen-

tation point list includes the first segmentation position and the second segmentation position.

[0434] In some embodiments, determining the second segmentation position from the position of the prosodic identifier located after the first segmentation position in the prosodic phoneme sequence may include: determining the position corresponding to the identifier used to represent intonational phrases in the prosodic phoneme sequence that is located after the first segmentation position as the second segmentation position.

[0435] In this embodiment, the prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1 kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3 tiao2 #0 jie2 #1 wen1 #0 du4 #3 ding4 #0 shi2 #1 kail #0 guan1 #3 xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil" is further used as an example for explanation.

[0436] After determining the first segmentation position, starting from the position where the first segmentation is located, i.e., starting from "yi3 #1", positions of the prosodic identifier with the prosodic identifier "#3" are searched form subsequent positions in sequence, and these positions are determined as the second segmentation positions in sequence, to obtain the following segmentation sequence:

> sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1| kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3| tiao2 #0 jie2 #1 wen1 #0 du4 #3 |ding4 #0 shi2 #1 kail #0 guan1 #3 |xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil,
> where a first "|" is the first segmentation position, and subsequent "|" are all second segmentation positions.

[0437] In other embodiments, the position of other fine-grained level identifiers may also be determined as the second segmentation position, and is not limited in the present application.

[0438] The second sub prosodic phoneme sequence is generated by segmenting the entire prosodic phoneme sequence located after the first segmentation position.

[0439] In case where the second segmentation position is at least one, the second prosodic phoneme sequences are at least two.

[0440] In this embodiment, the second segmentation position is determined based on the first segmentation position and the prosodic feature to improve the naturalness of the second sub prosodic phoneme sequence generated by subsequent segmentation and the balance at both ends after segmentation, avoiding cutting in the middle of an entire word, which helps to improve the efficiency and quality of subsequent speech synthesis.

[0441] In some embodiments, in absence of the second segmentation position, the second sub prosodic phoneme sequence is the entire prosodic phoneme sequence located after the first segmentation position in the prosodic phoneme sequence. For example, if the second segmentation position is a position corresponding #3, and #3 cannot be found in the second prosodic phoneme sequence, it may be understood that there is no second segmentation position.

[0442] For example, the first segmentation position and the second segmentation position determined by step 1120 are as follows:

sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1| kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3| tiao2 #0 jie2 #1 wen1 #0 du4 #3 |ding4 #0 shi2 #1 kail #0 guan1 #3 |xiang2 #0 xi4 #1 nei4 #0 rong2 #2 ma2 #0 fan5 #1 zai4 #1 jia 3 #0 jul #1 AE1 P #0 shang4 #1 soul #0 xun2 #0 xia4 #4 sil.

[0443] In this step, starting from the position of the first "|", the prosodic phoneme sequence is sequentially segmented to generate the first sub prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1", as well as the second sub prosodic phoneme sequence "kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3", "tiao2 #0 jie2 #1 wen1 #0 du4 #3", and "ding4 #0 shi2 #1 kail #0 guan1 #3".

[0444] The speech synthesis duration of the first sub prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1" is about 0.2 s.

[0445] In the text segmentation method provided by the embodiments of the present application, the first segmentation position used for obtaining the first sub prosodic phoneme sequence is determined by the speech synthesis duration corresponding to the first sub prosodic phoneme sequence, so that the speech synthesis duration corresponding to the first sub prosodic phoneme sequence may be within a reasonable time range, thereby shortening the first sentence response time of the synthesis system.

[0446] In some embodiments, step 1110 may include: obtaining end-of-sentence information, an intonational phrase, a prosodic phrase, a prosodic word, and a syllable of the target text; converting the target text into the phoneme sequence, generating a plurality of prosodic identifiers based on at least two of the end-of-sentence information, the intonational phrase, the prosodic phrase, the prosodic word, and the syllable; and labeling the phoneme sequence based on the plurality of prosodic identifiers to generate prosodic phoneme sequence.

[0447] In some embodiments of the present application, after step 1130, the method may further include: performing speech synthesis on the first sub prosodic phoneme sequence to generate first speech, and outputting the first speech and performing speech synthesis on the second sub prosodic phoneme sequence to generate second speech.

[0448] In this embodiment, the first sub prosodic phoneme sequence is the sequence before the first segmentation point in the target text, which corresponds to the sequence of the first sentence in the synthesized speech of the target text.

[0449] After the first sub prosodic phoneme sequence is generated, speech synthesis may be performed on the first sub prosodic phoneme sequence to generate the first speech.

[0450] The first speech is then output for a client to play. While the client plays the first speech, the system may synthesize the subsequent second sub prosodic phoneme sequence to generate the second speech.

[0451] For example, after the first sub prosodic phoneme sequence "sil mu4 #0 qian2 #1 xiao2 #0 jia3 #3 ke2 #0 yi3 #1" is obtained, the first speech "Currently, Xiaojia may" may be synthesized based on the first sub prosodic phoneme sequence, and the first speech may be output; while playing the first speech on the client, the system performs speech synthesis on the second sub prosodic phoneme sequence: "kong4 #0 zhi4 #1 re4 #0 shui3 #0 qi4 #1 kail #0 guan1 #3".

[0452] In the text segmentation method provided by the embodiments of the present application, by first performing speech synthesis on the first speech corresponding to the first sub prosodic phoneme sequence, the subsequent second sub prosodic phoneme sequence is synthesized while the first speech is output, which may accelerate the feedback speed of the system after receiving the network speech synthesis service request and shorten the waiting time of the user.

[0453] The apparatus embodiments described above are only illustrative, where the units described as separate components may or may not be physically separated, and the components displayed as units may or may not be physical units, that is, they may be located in one place or distributed across a plurality of network units. Some or all modules may be selected according to actual needs to achieve the purpose of this embodiment. Those of ordinary skill in the art may understand and implement without creative labor.

[0454] Through the description of the above implementation, those of ordinary skill in the art may clearly understand that each implementation may be achieved through software and necessary universal hardware platforms, and of course, may also be achieved through hardware. Based on such understanding, the above-mentioned solutions or the parts that contribute to related art may be reflected in the form of software products. The computer software products may be stored in computer-readable storage media, such as ROM/RAM, disks, CDs, etc., including several instructions to enable a computer device (which may be a personal computer, a server, or a network device, etc.) to execute various embodiments or certain parts of the embodiments.

[0455] Finally, it should be noted that: the above embodiments are only used to illustrate the solution of the present application, not to limit it; although the present application has been described in detail with reference to the aforementioned embodiments, those of ordinary skill in the art should understand: they may still modify the solutions recorded in the aforementioned embodiments, or equivalently replace some of the features, and these modifications or replacements do not separate the corresponding solutions from the scope of the solutions of the various embodiments of the present application.

[0456] The above implementations are only intended to illustrate the present application but not to limit it. Although the present application has been described in detail with reference to the embodiments, those skilled in the art should understand that any combination, modification, or equivalent replacement of the solution of the present application does not deviate from the scope of the solution of the present application, and should be covered within the scope of the claims of the present application.

Claims

1. A speech synthesis method comprising:

   segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences, wherein the prosodic phoneme sequence comprises a plurality of phonemes corresponding to the target text and a prosodic identifier located between adjacent phonemes, and each clause sequence comprises at least one phoneme;
   performing speech synthesis on a first sub prosodic phoneme sequence in the plurality of clause sequences to obtain first speech information; and
   outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in the plurality of clause sequences to generate second speech information, wherein the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

2. The method of claim 1, wherein after generating the second speech information, the method further comprises:
   combining the second speech information and the first speech information to generate third speech information.

3. The method of claim 2, wherein before segmenting the prosodic phoneme sequence of the target text, the method further comprises:

   generating a target file size of the third speech information based on the prosodic phoneme sequence; and
   generating the second speech information comprises:
   generating the second speech information

based on the target file size.

4. The method of claim 3, wherein generating a target file size of the third speech information based on the prosodic phoneme sequence comprises:

generating a predicted file size of the third speech information based on the prosodic phoneme sequence; and
correcting the predicted file size based on a target residual value to generate the target file size, wherein the target residual value is determined based on a sample file size and a predicted sample audio file size corresponding to a sample text, and the sample file size is an actual sample audio file size corresponding to the sample text.

5. The method of claim 4, further comprising obtaining a sample audio file size, wherein obtaining the sample audio file size comprises:

extracting a feature from the target text to generate a target prosodic feature and a target phoneme feature; and
obtaining a target file size of a target audio file based on the target prosodic feature and the target phoneme feature of the target text, wherein the target audio file is generated by performing speech synthesis on the target text.

6. The method of claim 5, wherein obtaining the target file size of the target audio file based on the target prosodic feature and the target phoneme feature comprises:

obtaining a first predicted file size of the target audio file based on the target prosodic feature and the target phoneme feature; and
summing the first predicted file size and a target residual value to generate the target file size, wherein the target residual value is determined based on a sample file size and a predicted sample audio file size corresponding to the sample text, and the sample file size is the actual sample audio file size corresponding to the sample text.

7. The method of claim 6, wherein the target residual value is determined by following steps:

obtaining the sample text, a sample audio file corresponding to the sample text and the sample file size corresponding to the sample audio file, wherein the sample audio file is generated by performing speech synthesis on the sample text;
extracting a feature from the sample text to generate a sample prosodic feature and a sample phoneme feature;
obtaining a second predicted file size of the sample audio file based on the sample prosodic feature and the sample phoneme feature; and
determining an absolute value of a difference between the second predicted file size and the sample file size as the target residual value.

8. The method of claim 6, wherein obtaining the first predicted file size of the target audio file based on the target prosodic feature and the target phoneme feature comprises:

inputting the target prosodic feature and the target phoneme feature into a file size prediction model to obtain the first predicted file size output from the file size prediction model,
wherein the file size prediction model is trained using a sample prosodic feature and a sample phoneme feature as a sample, and a sample file size corresponding to the sample prosodic feature and the sample phoneme feature as a sample label.

9. The method of claim 5, wherein after obtaining the target file size of the target audio file, the method further comprises:

segmenting the target text based on the target prosodic feature and the target phoneme feature to generate a plurality of clause sequences;
perform speech synthesis on the clause sequenceto generate clause speech; and
outputting the clause speech and the target file size, and splicing the clause speech to generate the target audio file.

10. The method of any one of claims 5-9, wherein extracting the feature from the sample text to generate the target prosodic feature and the target phoneme feature comprises:

converting the target text into a prosodic phoneme sequence, wherein the prosodic phoneme sequence comprises a plurality of phonemes corresponding to the target text and prosodic identifiers located between adjacent phonemes; and
extracting a feature from the prosodic phoneme sequence to generate the target prosodic feature and the target phoneme feature.

11. The method of claim 2, wherein combining the second speech information and the first speech information comprises:
combining the second speech information and the first speech information based on a phoneme dura-

tion corresponding to the second speech information and a phoneme duration corresponding to the first speech information.

12. The method of claim 11, further comprising:

performing speech synthesis on the clause sequence respectively to generate a plurality of pieces of first clause speech information, wherein the first clause speech information comprises a first duration corresponding to each phoneme and the prosodic identifier; and
splicing a plurality of pieces of first clause speech information based on the first duration and an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the first clause speech information, to generate target speech.

13. The method of claim 12, wherein splicing the plurality of pieces of first clause speech information based on the first duration and an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the first clause speech information to generate the target speech comprises:

truncating speech corresponding to the target phoneme in the first clause speech information based on the first duration corresponding to the target phoneme in the plurality of phonemes, to generate second clause speech information, and
splicing the second clause speech information based on an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the second clause speech information, to generate the target speech.

14. The method of claim 13, wherein the target phoneme comprises at least one of a beginning-of-sentence redundant phoneme and an end-of-sentence redundant phoneme, and truncating the speech corresponding to the target phoneme in the first clause speech information based on the first duration corresponding to the target phoneme in the plurality of phonemes comprises:

determining that the clause sequence corresponding to the first clause speech information is not a first clause sequence in the target text, and truncating the speech corresponding to the beginning-of-sentence redundant phoneme and the speech corresponding to the end-of-sentence redundant phoneme in the first sentence phonetic information, respectively; or
determining that the clause sequence corresponding to the first clause speech information is a first clause sequence in the target text, and truncating the speech corresponding to the end-of-sentence redundant phoneme in the first sentence speech information.

15. The method of claim 11, wherein after generating the plurality of pieces of first clause speech information and before splicing the plurality of pieces of first clause speech information based on the first duration and the order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the first clause speech information, the method further comprises:
outputting the first clause speech information.

16. The method of any one of claims 12-15, wherein performing speech synthesis on the clause sequence respectively comprises:

inputting the clause sequence into a target speech synthesis model to obtain the first clause speech information output from the target speech synthesis model,
wherein the target speech synthesis model is trained using a sample prosodic phoneme sequence as a sample and sample clause speech corresponding to the sample prosodic phoneme sequence as a sample label.

17. The method of any one of claims 1-16, wherein before segmenting the prosodic phoneme sequence of the target text, the method further comprises:

obtaining a to-be-synthesized text; and
determining that a size of the to-be-synthesized text exceeds a target threshold, segmenting the to-be-synthesized text to generate the target text, wherein a size of the target text does not exceed the target threshold.

18. The method of any one of claims 1-16, wherein segmenting the prosodic phoneme sequence of the target text to generate a plurality of clause sequences comprises:

converting the target text into the prosodic phoneme sequence, wherein the prosodic phoneme sequence comprises a plurality of phonemes corresponding to the target text and the prosodic identifier located between adjacent phonemes; and
segmenting the prosodic phoneme sequence based on at least partial of the plurality of prosodic identifiers to generate the plurality of clause sequences.

19. The method of claim 18, wherein segmenting the prosodic phoneme sequence based on at least par-

tial of the plurality of prosodic identifiers to generate the plurality of clause sequences comprises:

determining a first segmentation position based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence; and
segmenting the prosodic phoneme sequence based on the first segmentation position to generate the plurality of clause sequences, wherein the clause sequence comprises the first sub prosodic phoneme sequence and the second sub prosodic phoneme sequence, the first sub prosodic phoneme sequence is the prosodic phoneme sequence located before the first segmentation position in the prosodic phoneme sequence, the second sub prosodic phoneme sequence is the prosodic phoneme sequence located after the first segmentation position in the prosodic phoneme sequence, and a speech synthesis duration corresponding to the first sub prosodic phoneme sequence is within a target duration.

20. The method of claim 19, wherein determining the first segmentation position based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence comprises:

determining a prosodic identifier with a largest fine grain from the plurality of prosodic identifiers based on a target threshold range; and
determining a position of the prosodic identifier with the largest fine grain in the prosodic phoneme sequence as the first segmentation position.

21. The method of claim 20, wherein determining the prosodic identifier with the largest fine grain from the plurality of prosodic identifiers based on the target threshold range comprises:

obtaining a first articulation length of all phonemes in a target sub prosodic phoneme sequence in the prosodic phoneme sequence, wherein the target sub prosodic phoneme sequence is all prosodic phoneme sequences in the prosodic phoneme sequence before a target position;
determining that the first articulation length is within the target threshold range, and determining that the prosodic identifier at the target position corresponding to the first articulation length appears for the first time, and determining the target position corresponding to the first articulation length as a candidate segmentation point position to generate a plurality of candidate segmentation point positions; and
determining the prosodic identifier with largest fine grain from prosodic identifiers corresponding to the plurality of candidate segmentation point positions.

22. The method of claim 18, wherein segmenting the prosodic phoneme sequence based on at least partial of the plurality of prosodic identifiers to generate the plurality of clause sequences comprises:

determining a first segmentation position based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence;
determining a second segmentation position from the position of the prosodic identifier located after the first segmentation position in the prosodic phoneme sequence; and
segmenting the prosodic phoneme sequence based on the first segmentation position and the second segmentation position to generate the first second sub prosodic phoneme sequence and at least two second sub prosodic phoneme sequences, wherein the first sub prosodic phoneme sequence is the prosodic phoneme sequence located before the first segmentation position in the prosodic phoneme sequence, the at least two second sub prosodic phoneme sequences are the prosodic phoneme sequence located after the first segmentation position in the prosodic phoneme sequence, adjacent second sub prosodic phoneme sequences are determined based on the second segmentation position, and the speech synthesis duration corresponding to the first sub prosodic phoneme sequence is within the target duration.

23. The method of claim 18, wherein converting the target text into the prosodic phoneme sequence comprises:

obtaining a syllable, a prosodic word, a prosodic phrase, an intonational phrase, and end-of-sentence information of the target text; and
labeling the target text based on at least two of the syllable, the prosodic word, the prosodic phrase, the intonational phrase, and the end-of-sentence information to generate the prosodic phoneme sequence.

24. The method of claim 23, wherein labeling the target text based on at least two of the syllable, the prosodic word, the prosodic phrase, the intonational phrase, and the end-of-sentence information to generate the prosodic phoneme sequence comprises:

converting the target text into a phoneme sequence;
generating the plurality of prosodic identifiers

based on at least two of the syllable, the prosodic word, the prosodic phrase, the intonational phrase and the end-of-sentence information; and

labeling the phoneme sequence based on the plurality of prosodic identifiers to generate the prosodic phoneme sequence.

25. The method of any one of claims 1-24, further comprising:

segmenting the prosodic phoneme sequence of the target text to generate a plurality of clause sequences, wherein the prosodic phoneme sequence comprises a plurality of phonemes corresponding to the target text and the prosodic identifier located between adjacent phonemes, and each clause sequence comprises at least one phoneme; and

determining that any one of to-be-matched clause sequences among the plurality of clause sequences matches a cached target clause sequence, obtaining a target clause speech corresponding to the target clause sequence from a cache, and determining the speech corresponding to the to-be-matched clause sequences as the target clause speech.

26. The method of claim 25, further comprising:

determining that any one of the to-be-matched clause sequences of the plurality of clause sequences does not match the target clause sequence, and performing speech synthesis on the to-be-matched clause sequence to generate the second clause speech.

27. The method of claim 26, wherein after generating the second clause speech, the method further comprises:

segmenting the second clause speech based on the prosodic identifier to generate a plurality of pieces of sub second clause speech; and

caching the plurality of pieces of sub second clause speech and the sub clause sequence corresponding to the sub second clause speech.

28. The method of claim 26, wherein after generating the second clause speech, the method further comprises:

splicing the target clause speech and the second clause speech based on an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the target clause speech and an order of segmenting the prosodic phoneme sequence into the clause sequence corresponding to the second clause speech, to generate the target speech corresponding to the target text.

29. The method of claim 25, wherein segmenting the prosodic phoneme sequence based on at least partial of the plurality of prosodic identifiers to generate the plurality of clause sequences comprises:

segmenting the prosodic phoneme sequence based on a target identifier in the plurality of prosodic identifiers to generate a plurality of candidate sequences, wherein a speech synthesis duration corresponding to the candidate sequence before a first segmentation point is within the target duration;

combining a target candidate sequence with adjacent candidate sequences in the plurality of candidate sequences to generate a plurality of clause sequences, and determining a fine grain corresponding to the clause sequence; and

ranking the plurality of clause sequences in descending order based on the fine grain corresponding to the clause sequence.

30. The method of any one of claims 1-29, wherein the prosodic identifier comprises: at least one of the identifiers used to represent the syllable, the prosodic word, the prosodic phrase, the intonational phrase, and the end-of-sentence information; and

a fine grain of a prosodic identifier used to represent the end-of-sentence information is greater than a fine grain of a prosodic identifier used to represent the intonational phrase, a fine grain of a prosodic identifier used to represent the intonational phrase is greater than a fine grain of a prosodic identifier used to represent the prosodic phrase, a fine grain of a prosodic identifier used to represent the prosodic phrase is greater than a fine grain of a prosodic identifier used to represent the prosodic word, and a fine grain of a prosodic identifier used to represent the prosodic word is greater than a fine grain of a prosodic identifier used to represent the syllable.

31. The method of any one of claims 1-30, wherein the target prosodic feature and the target phoneme feature comprise at least one of: a length of the prosodic phoneme sequence, a number of Chinese Pinyin in the prosodic phoneme sequence, a number of pause symbols in the prosodic phoneme sequence, the number of English phonemes in the prosodic phoneme sequence, a number of Chinese phonemes in the prosodic phoneme sequence, a number of Chinese initials in the prosodic phoneme sequence, and English phonemes of any category in the prosodic phoneme sequence.

32. A speech synthesis apparatus, comprising:

a first processing module, configured to seg-

ment the prosodic phoneme sequence of the target text to generate a plurality of clause sequences, wherein the prosodic phoneme sequence comprises a plurality of phonemes corresponding to the target text and the prosodic identifier located between adjacent phonemes, and each clause sequence comprises at least one phoneme;

a second processing module, configured to perform speech synthesis on a first sub prosodic phoneme sequence in the plurality of clause sequences to obtain first speech information; and

a third processing module, configured to output the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in the plurality of clause sequences to generate second speech information, wherein the second sub prosodic phoneme sequence is at least one clause sequence located after the first sub prosodic phoneme sequence in the prosodic phoneme sequence.

33. An electronic device, comprising:

a processor; and

a memory storing a computer program capable of running on the processor, wherein the program, when executed by the processor, causes the electronic device to perform the speech synthesis method of any one of claims 1 to 31.

34. A non-transient computer-readable storage medium on which a computer program is stored, and the computer program, when executed by a processor, performs the speech synthesis method of any one of claims 1 to 31.

35. A computer program product comprising a computer program, wherein the computer program, when executed by a processor, performs the speech synthesis method of any one of claims 1 to 31.

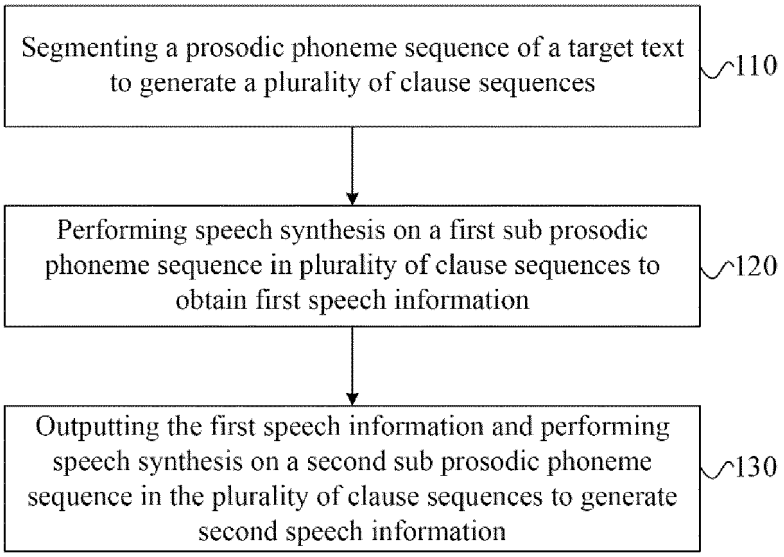
Segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences
110
Performing speech synthesis on a first sub prosodic phoneme sequence in plurality of clause sequences to obtain first speech information
120
Outputting the first speech information and performing speech synthesis on a second sub prosodic phoneme sequence in the plurality of clause sequences to generate second speech information
130

FIG. 1

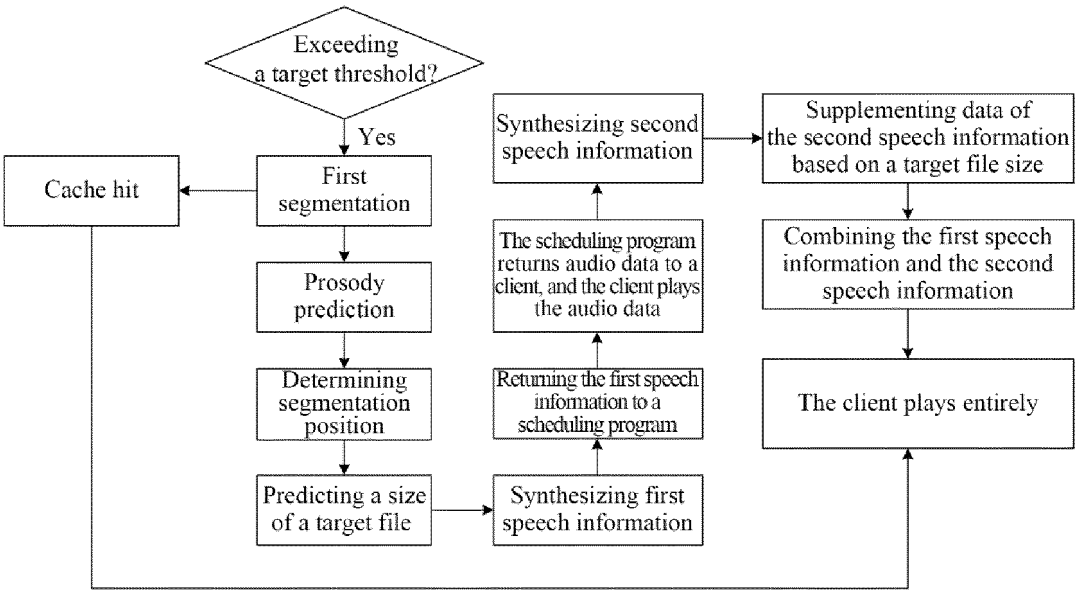
Exceeding a target threshold?
Yes
Cache hit
First segmentation
Prosody prediction
Determining segmentation position
Predicting a size of a target file
Synthesizing first speech information
Returning the first speech information to a scheduling program
The scheduling program returns audio data to a client, and the client plays the audio data
Synthesizing second speech information
Supplementing data of the second speech information based on a target file size
Combining the first speech information and the second speech information
The client plays entirely

FIG. 2

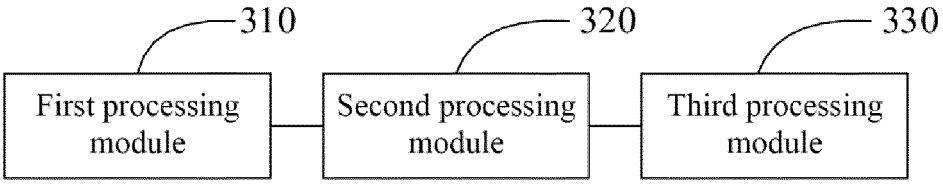

FIG. 3

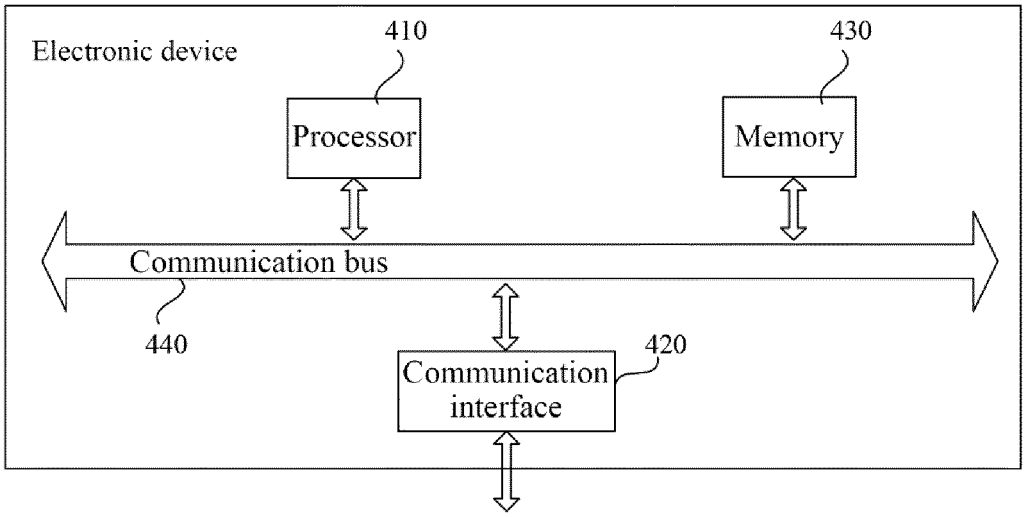

FIG. 4

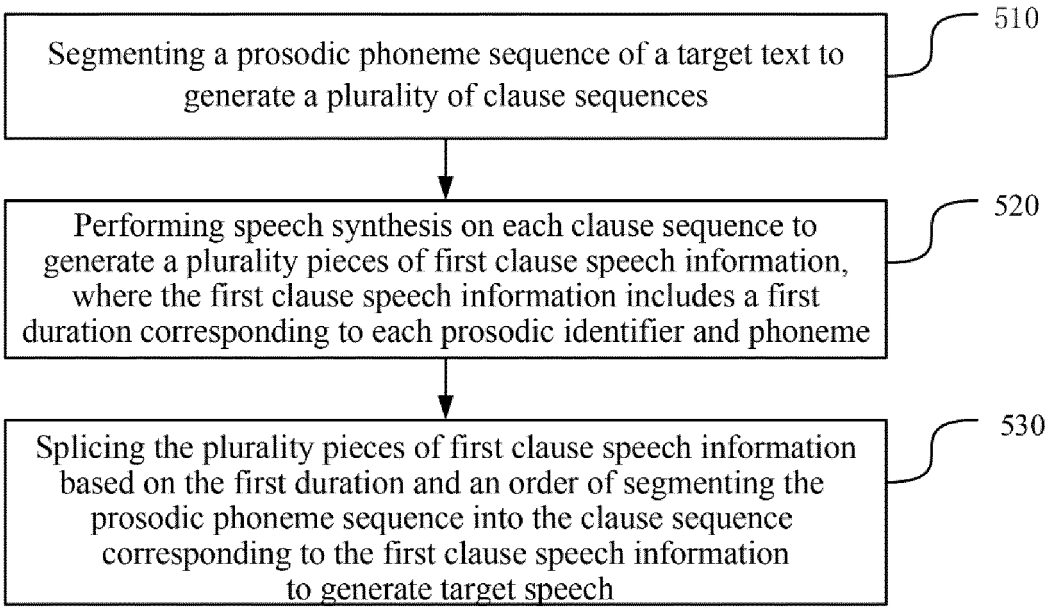

FIG. 5

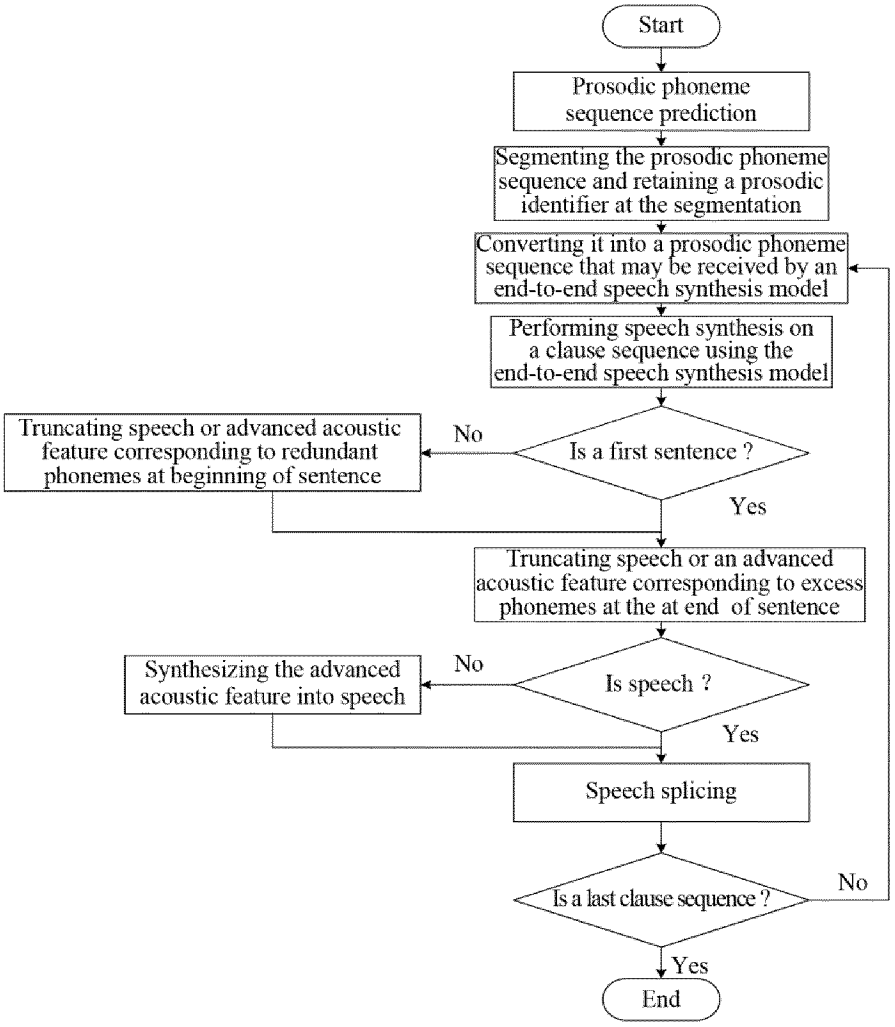
Start
Prosodic phoneme sequence prediction
Segmenting the prosodic phoneme sequence and retaining a prosodic identifier at the segmentation
Converting it into a prosodic phoneme sequence that may be received by an end-to-end speech synthesis model
Performing speech synthesis on a clause sequence using the end-to-end speech synthesis model
Is a first sentence ?
No
Truncating speech or advanced acoustic feature corresponding to redundant phonemes at beginning of sentence
Yes
Truncating speech or an advanced acoustic feature corresponding to excess phonemes at the at end of sentence
Is speech ?
No
Synthesizing the advanced acoustic feature into speech
Yes
Speech splicing
Is a last clause sequence ?
No
Yes
End

FIG. 6

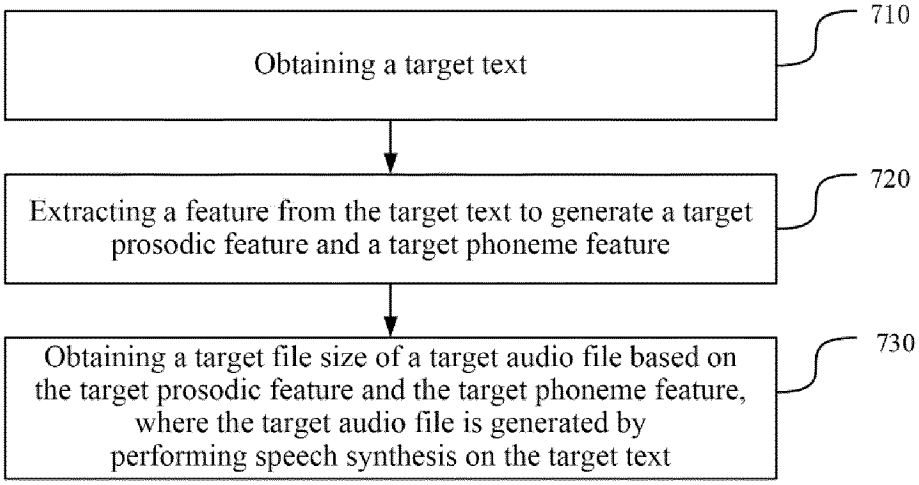
Obtaining a target text
710
Extracting a feature from the target text to generate a target prosodic feature and a target phoneme feature
720
Obtaining a target file size of a target audio file based on the target prosodic feature and the target phoneme feature, where the target audio file is generated by performing speech synthesis on the target text
730

FIG. 7

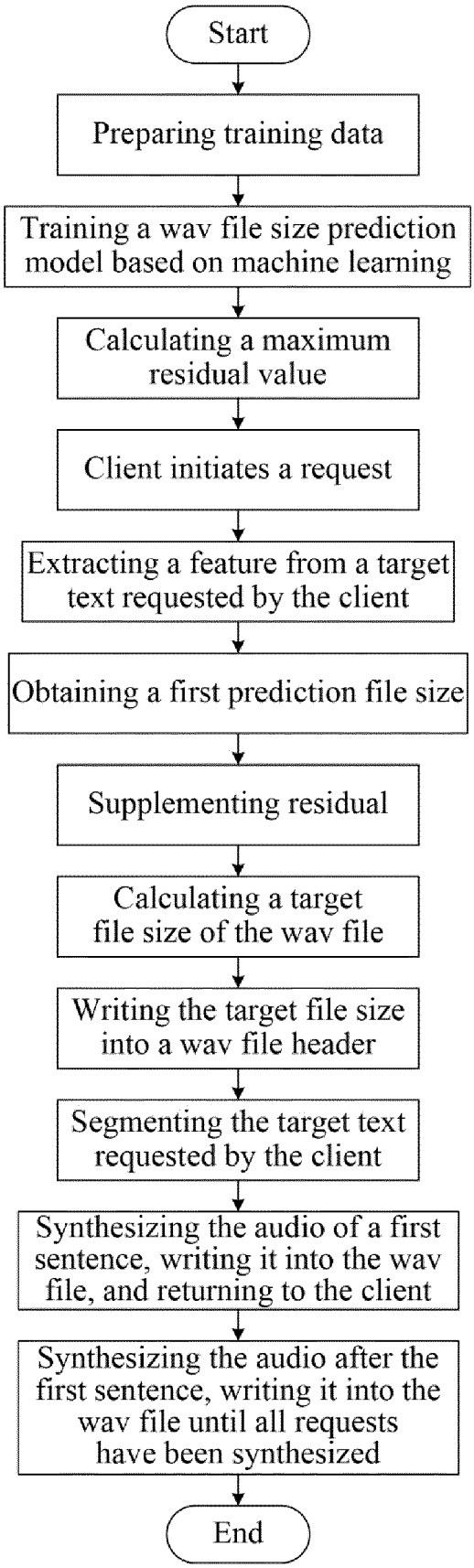
Start
Preparing training data
Training a wav file size prediction model based on machine learning
Calculating a maximum residual value
Client initiates a request
Extracting a feature from a target text requested by the client
Obtaining a first prediction file size
Supplementing residual
Calculating a target file size of the wav file
Writing the target file size into a wav file header
Segmenting the target text requested by the client
Synthesizing the audio of a first sentence, writing it into the wav file, and returning to the client
Synthesizing the audio after the first sentence, writing it into the wav file until all requests have been synthesized
End

FIG. 8

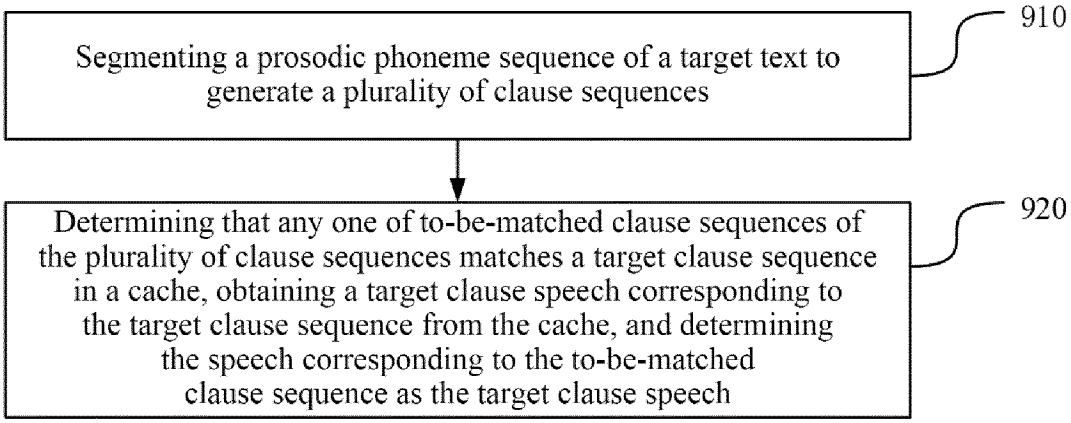
Segmenting a prosodic phoneme sequence of a target text to generate a plurality of clause sequences
910
Determining that any one of to-be-matched clause sequences of the plurality of clause sequences matches a target clause sequence in a cache, obtaining a target clause speech corresponding to the target clause sequence from the cache, and determining the speech corresponding to the to-be-matched clause sequence as the target clause speech
920

FIG. 9

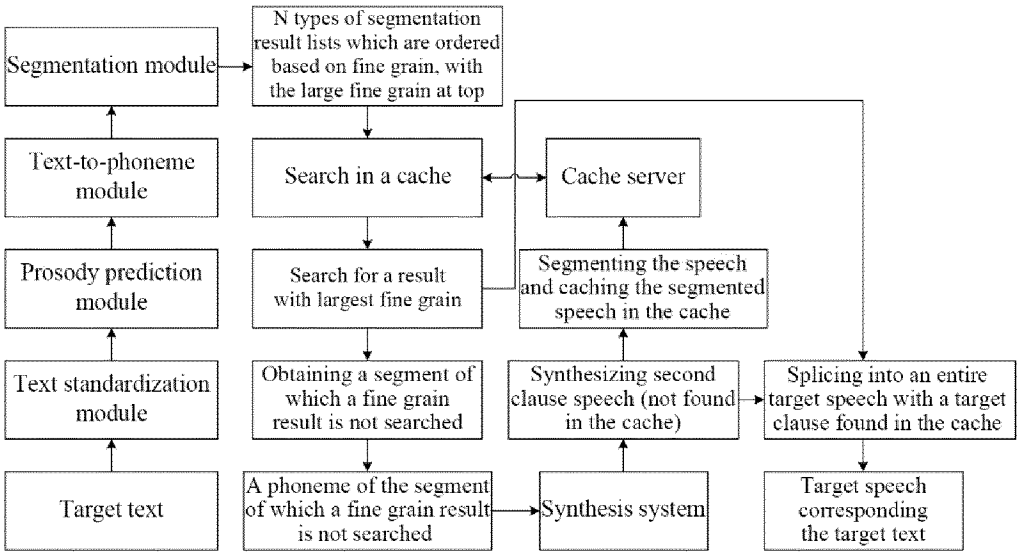
Segmentation module
N types of segmentation result lists which are ordered based on fine grain, with the large fine grain at top
Text-to-phoneme module
Search in a cache
Cache server
Prosody prediction module
Search for a result with largest fine grain
Segmenting the speech and caching the segmented speech in the cache
Text standardization module
Obtaining a segment of which a fine grain result is not searched
Synthesizing second clause speech (not found in the cache)
Splicing into an entire target speech with a target clause found in the cache
Target text
A phoneme of the segment of which a fine grain result is not searched
Synthesis system
Target speech corresponding the target text

FIG. 10

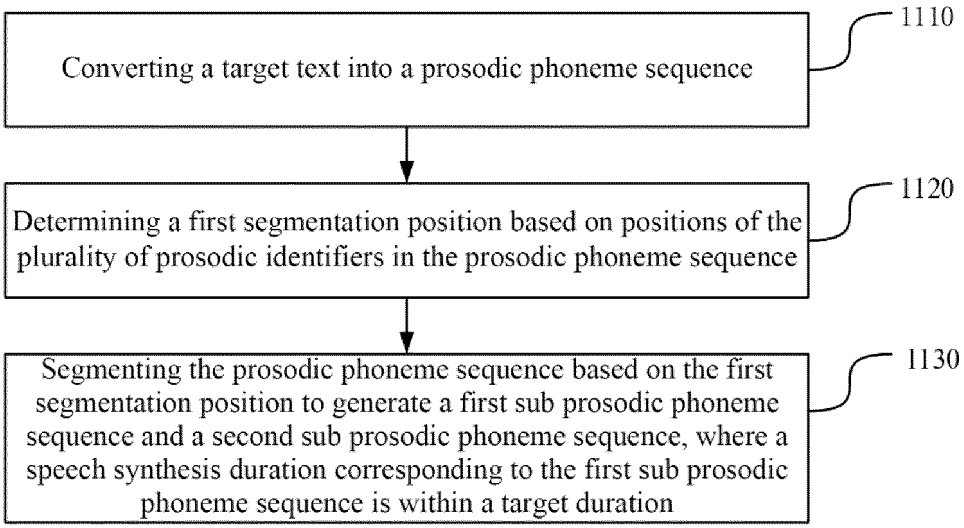
Converting a target text into a prosodic phoneme sequence
1110
Determining a first segmentation position based on positions of the plurality of prosodic identifiers in the prosodic phoneme sequence
1120
Segmenting the prosodic phoneme sequence based on the first segmentation position to generate a first sub prosodic phoneme sequence and a second sub prosodic phoneme sequence, where a speech synthesis duration corresponding to the first sub prosodic phoneme sequence is within a target duration
1130

FIG. 11

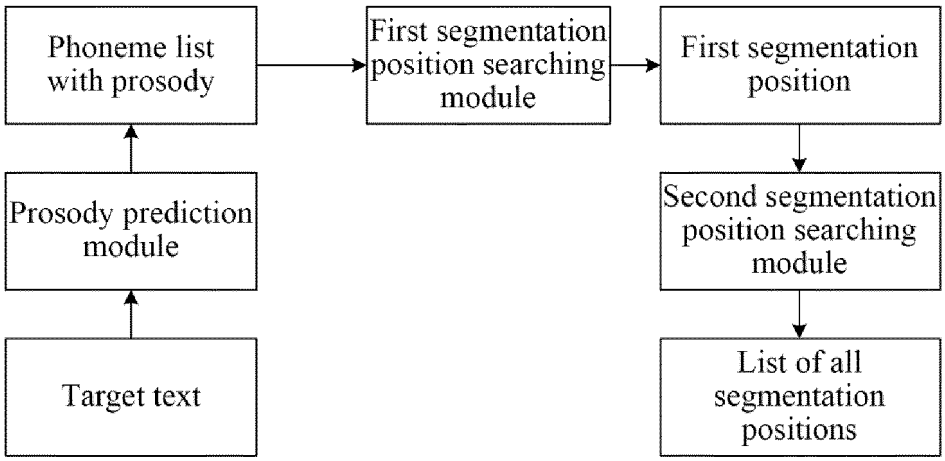
Phoneme list with prosody
First segmentation position searching module
First segmentation position
Prosody prediction module
Second segmentation position searching module
Target text
List of all segmentation positions

FIG. 12

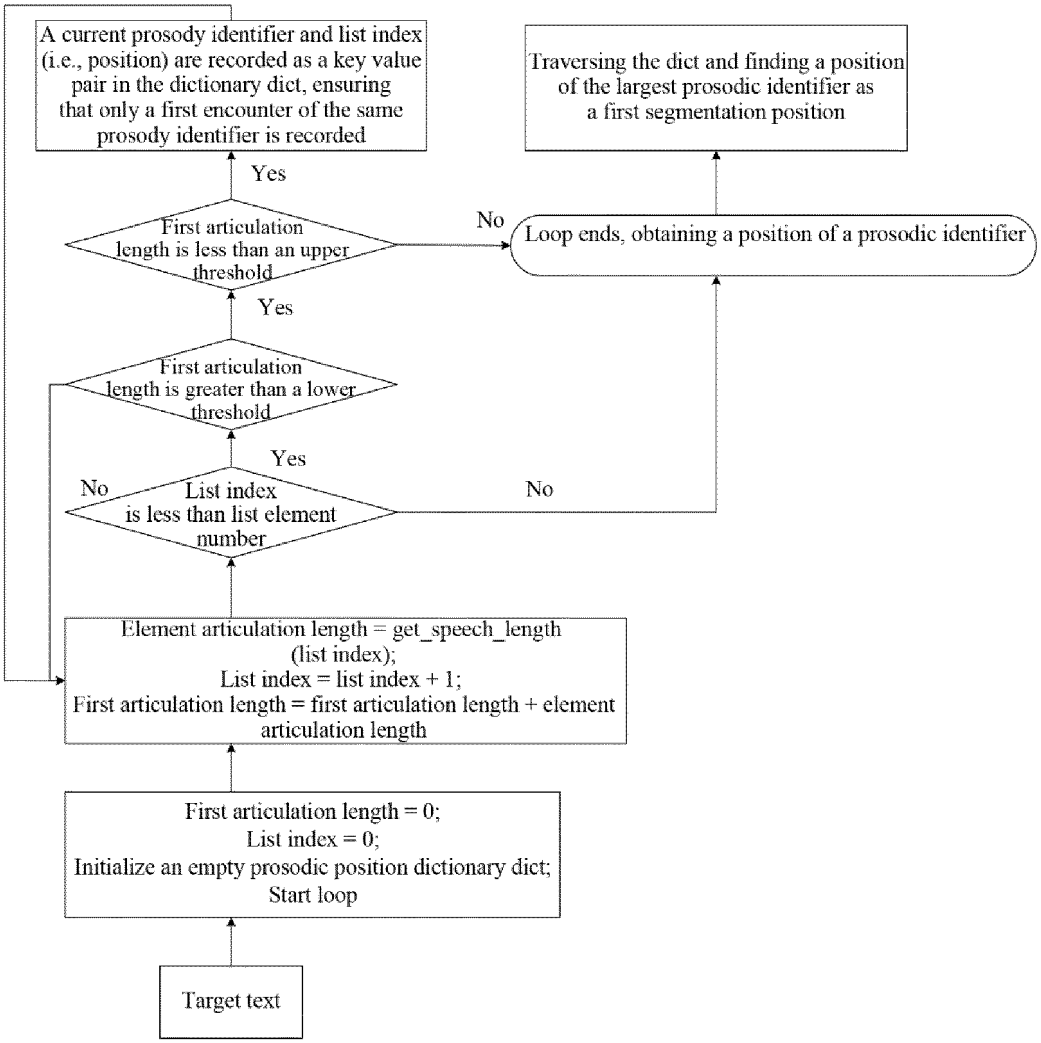
A current prosody identifier and list index (i.e., position) are recorded as a key value pair in the dictionary dict, ensuring that only a first encounter of the same prosody identifier is recorded
Traversing the dict and finding a position of the largest prosodic identifier as a first segmentation position
Yes
No
First articulation length is less than an upper threshold
Loop ends, obtaining a position of a prosodic identifier
Yes
First articulation length is greater than a lower threshold
Yes
No
No
List index is less than list element number
Element articulation length = get_speech_length (list index);
List index = list index + 1;
First articulation length = first articulation length + element articulation length
First articulation length = 0;
List index = 0;
Initialize an empty prosodic position dictionary dict;
Start loop
Target text

FIG. 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/118072

A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/02(2013.01)i; G10L 13/04(2013.01)i; G10L 13/10(2013.01)i; G10L 13/08(2013.01)i; G10L 13/06(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; CNKI; VEN; EPTXT; WOTXT; USTXT: 美的, 语音, 合成, 转换, 文本, 音素, 韵律, 切割, 切分, 划分, 分割, 处理, 速度, 速率, 等待, 时长, 时间, 反馈, voice, audio, speech, synthesi+, text, process+, phoneme, rhythm, conversion, convert, division, splic+, cut+, speed, efficient

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114678001 A (MIDEA GROUP (SHANGHAI) CO., LTD. et al.) 28 June 2022 (2022-06-28) description, paragraphs [0003]-[0257], and figures 1-4	1-4, 11, 17-19, 23, 24, 32-35
PX	CN 114678002 A (MIDEA GROUP (SHANGHAI) CO., LTD. et al.) 28 June 2022 (2022-06-28) description, paragraphs [0003]-[0220], and figures 1-5	1, 18-22, 30, 32-35
PY	CN 114708848 A (MIDEA GROUP (SHANGHAI) CO., LTD. et al.) 05 July 2022 (2022-07-05) description, paragraphs [0003]-[0222], and figures 1-4	5-10, 17-31, 33-35
PY	CN 114822490 A (MIDEA GROUP (SHANGHAI) CO., LTD. et al.) 29 July 2022 (2022-07-29) description, paragraphs [0003]-[0186], and figures 1-4	12-31, 33-35
PY	CN 114822489 A (MIDEA GROUP (SHANGHAI) CO., LTD. et al.) 29 July 2022 (2022-07-29) description, paragraphs [0003]-[0181], and figures 1-4	25-31, 33-35

☑ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 November 2022	**07 December 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/118072

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	CN 115223541 A (UBTECH ROBOTICS CORP.) 21 October 2022 (2022-10-21) description, paragraphs [0003]-[0081], and figures 1-8	1, 2, 32-35
Y	CN 112771607 A (SAMSUNG ELECTRONICS CO., LTD.) 07 May 2021 (2021-05-07) description, paragraphs [0057]-[0245], and figures 1-18	1-35
Y	CN 113516964 A (BEIJING FANGJIANGHU TECHNOLOGY CO., LTD.) 19 October 2021 (2021-10-19) description, paragraphs [0004]-[0026]	1-35
A	CN 111524500 A (ZHEJIANG TONGHUASHUN INTELLIGENT TECHNOLOGY CO., LTD.) 11 August 2020 (2020-08-11) entire document	1-35
A	CN 112037758 A (SICHUAN CHANGHONG ELECTRIC CO., LTD.) 04 December 2020 (2020-12-04) entire document	1-35
A	CN 110797006 A (BEIJING SPEECHOCEAN TECHNOLOGY CO., LTD.) 14 February 2020 (2020-02-14) entire document	1-35
A	CN 113053357 A (NETEASE (HANGZHOU) NETWORK CO., LTD.) 29 June 2021 (2021-06-29) entire document	1-35
A	CN 111226275 A (UBTECH ROBOTICS CORP.) 02 June 2020 (2020-06-02) entire document	1-35
A	CN 112885328 A (HUAWEI TECHNOLOGIES CO., LTD.) 01 June 2021 (2021-06-01) entire document	1-35
A	JP 2002304186 A (SHARP K. K.) 18 October 2002 (2002-10-18) entire document	1-35
A	CN 108073572 A (BEIJING SOGOU TECHNOLOGY DEVELOPMENT CO., LTD.) 25 May 2018 (2018-05-25) entire document	1-35

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/118072

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 114678001 A	28 June 2022	None	
CN 114678002 A	28 June 2022	None	
CN 114708848 A	05 July 2022	None	
CN 114822490 A	29 July 2022	None	
CN 114822489 A	29 July 2022	None	
CN 115223541 A	21 October 2022	None	
CN 112771607 A	07 May 2021	US 2020152194 A1	14 May 2020
		WO 2020101263 A1	22 May 2020
		KR 20200056261 A	22 May 2020
		EP 3818518 A1	12 May 2021
		EP 3818518 A4	11 August 2021
		US 11289083 B2	29 March 2022
		IN 202127018258 A	22 July 2022
CN 113516964 A	19 October 2021	CN 113516964 B	27 May 2022
CN 111524500 A	11 August 2020	None	
CN 112037758 A	04 December 2020	None	
CN 110797006 A	14 February 2020	CN 110797006 B	19 May 2020
CN 113053357 A	29 June 2021	None	
CN 111226275 A	02 June 2020	WO 2021134581 A1	08 July 2021
CN 112885328 A	01 June 2021	WO 2022156654 A1	28 July 2022
JP 2002304186 A	18 October 2002	JP 3681111 B2	10 August 2005
CN 108073572 A	25 May 2018	CN 108073572 B	11 January 2022

Form PCT/ISA/210 (patent family annex) (January 2015)

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- CN 202210344448X **[0001]**
- CN 2022103461146 **[0001]**
- CN 2022103460976 **[0001]**
- CN 2022103460942 **[0001]**
- CN 2022103444564 **[0001]**