CLAIMS

1. A sample generation method, comprising:

acquiring a plurality of text-audio pairs, wherein each text-audio pair comprises a text segment and an audio segment;

calculating, for each text-audio pair among the plurality of text-audio pairs, an audio feature of the audio segment of the text-audio pair, and screening out from the plurality of text-audio pairs, according to the audio feature, a target text-audio pair and a splicing text-audio pair corresponding to the target text-audio pair;

splicing the target text-audio pair and the splicing text-audio pair into a to-be-detected text-audio pair, and detecting the to-be-detected text-audio pair; and

writing the to-be-detected text-audio pair into a training database in a case that the to-be-detected text-audio pair meets a preset detection condition.

2. The sample generation method according to claim 1, wherein the acquiring a plurality of text-audio pairs comprises:

acquiring a target text and an audio corresponding to the target text;

preprocessing the audio to obtain a target audio, and converting the target text into a phoneme sequence; and

aligning the phoneme sequence with the target audio, and generating the plurality of text-audio pairs according to an alignment processing result.

3. The sample generation method according to claim 2, wherein the generating the plurality of text-audio pairs according to an alignment processing result comprises:

acquiring a phoneme audio file according to the alignment processing result, and determining a segmentation position of the phoneme audio file;

segmenting the phoneme audio file according to the segmentation position to obtain a plurality of phoneme-audio pairs, wherein each phoneme-audio pair comprises a phoneme segment and an audio segment;

determining, based on the target text, a text segment corresponding to the phoneme segment of each phoneme-audio pair among the plurality of phoneme-audio pairs; and

generating the plurality of text-audio pairs according to the text segment corresponding to the phoneme segment of each phoneme-audio pair and the audio segment of each phoneme-audio pair.

4. The sample generation method according to claim 1, wherein the calculating, for each text-audio pair among the plurality of text-audio pairs, an audio feature of the audio segment of the text-audio pair comprises:

for each text-audio pair among the plurality of text-audio pairs,

extracting the audio segment of the text-audio pair, and performing a framing process on the audio segment of the text-audio pair, to obtain an audio frame set of the text-audio pair;

calculating, based on audio frames in the audio frame set of the text-audio pair, a pitch frequency feature and an audio frame feature of the audio segment of the text-audio pair; and

determining the audio feature of the audio segment of the text-audio pair according to the pitch frequency feature and the audio frame feature of the audio segment of the text-audio pair.

5. The sample generation method according to claim 1, wherein the screening out from the plurality of text-audio pairs, according to the audio feature, a target text-audio pair and a splicing text-audio pair corresponding to the target text-audio pair comprises:

integrating, for each text-audio pair among the plurality of text-audio pairs, the audio segment, the text segment and the audio feature of the text-audio pair to obtain a text-audio package corresponding to the text-audio pair, and writing the text-audio package into a segment database;

selecting a text-audio package in the segment database as a target text-audio package, and determining a text-audio pair in the target text-audio package as the target text-audio pair; and

determining a splicing text-audio package based on text-audio packages in the segment database other than the target text-audio package and the audio feature, and determining a text-audio pair in the splicing text-audio package as the splicing text-audio pair.

6. The sample generation method according to claim 5, wherein the determining a splicing text-audio package based on text-audio packages in the segment database other than

the target text-audio package and the audio feature comprises:

selecting the text-audio packages in the segment database other than the target text-audio package to form a to-be-screened text-audio package set;

determining text-audio pairs of to-be-screened text-audio packages in the to-be-screened text-audio package set as to-be-screened text-audio pairs; and

screening out the splicing text-audio package from the to-be-screened text-audio package set based on an audio feature of an audio segment of the target text-audio pair and audio features of audio segments of the to-be-screened text-audio pairs.

7. The sample generation method according to claim 6, wherein the screening out the splicing text-audio package from the to-be-screened text-audio package set based on an audio feature of an audio segment of the target text-audio pair and audio features of audio segments of the to-be-screened text-audio pairs comprises:

determining a first audio feature of the audio segment of the target text-audio pair according to the target text-audio package, and determining, for each of the to-be-screened text-audio pairs, a second audio feature of the audio segment of the to-be-screened text-audio pair according to the to-be-screened text-audio package;

calculating a feature distance between the first audio feature and the second audio feature; and

determining a to-be-screened text-audio package to which a to-be-screened text-audio pair whose feature distance is less than a preset distance threshold belongs, as the splicing text-audio package.

8. The sample generation method according to claim 1, wherein before the step of splicing the target text-audio pair and the splicing text-audio pair into the to-be-detected text-audio pair and detecting the to-be-detected text-audio pair, the method further comprises:

sampling the audio segment of the target text-audio pair to obtain target sampling information, and determining target text information of the text segment of the target text-audio pair; and

determining whether the target sampling information and the target text information meet the preset detection condition; wherein

in a case that the target sampling information and the target text information do not meet the preset detection condition, the step of splicing the target text-audio pair and the splicing text-audio pair into the to-be-detected text-audio pair and detecting the to-be-detected text-audio pair is performed.

9. The sample generation method according to claim 8, wherein in a case that the target sampling information and the target text information meet the preset detection condition, the target text-audio pair is written into the training database.

10. The sample generation method according to claim 1, wherein the splicing the target text-audio pair and the splicing text-audio pair into a to-be-detected text-audio pair comprises:

extracting a target text segment and a target audio segment of the target text-audio pair, and extracting a splicing text segment and a splicing audio segment of the splicing text-audio pair;

splicing the target text segment and the splicing text segment into a to-be-detected text segment, and splicing the target audio segment and the splicing audio segment into a to-be-detected audio segment; and

forming the to-be-detected text-audio pair based on the to-be-detected text segment and the to-be-detected audio segment.

11. The sample generation method according to claim 10, wherein

the detecting the to-be-detected text-audio pair comprises:

sampling the to-be-detected audio segment to obtain to-be-detected sampling information, and determining to-be-detected text information of the to-be-detected text segment; and

detecting the to-be-detected sampling information and the to-be-detected text information based on the preset detection condition; and

the writing the to-be-detected text-audio pair into a training database in a case that the to-be-detected text-audio pair meets a preset detection condition comprises:

writing the to-be-detected text-audio pair into the training database in a case that both the to-be-detected sampling information and the to-be-detected text information meet the preset detection condition.

12. The sample generation method according to claim 1, wherein after the step of splicing the target text-audio pair and the splicing text-audio pair into the to-be-detected text-audio pair and detecting the to-be-detected text-audio pair, the method further comprises:

screening out a multi-degree splicing text-audio pair corresponding to the splicing text-audio pair from the plurality of text-audio pairs according to the audio feature in a case that the to-be-detected text-audio pair does not meet the preset detection condition; and

splicing the multi-degree splicing text-audio pair and the to-be-detected text-audio pair into a multi-degree to-be-detected text-audio pair, and determining whether the multi-degree to-be-detected text-audio pair meets the preset detection condition; wherein

in a case that the multi-degree to-be-detected text-audio pair meets the preset detection condition, the multi-degree to-be-detected text-audio pair is written into the training database; and

in a case that the multi-degree to-be-detected text-audio pair does not meet the preset detection condition, the multi-degree splicing text-audio pair is determined as the splicing text-audio pair, the multi-degree to-be-detected text-audio pair is determined as the to-be-detected text-audio pair, and the step of screening out the multi-degree splicing text-audio pair corresponding to the splicing text-audio pair from the plurality of text-audio pairs according to the audio feature is performed.

13. The sample generation method according to claim 1, wherein after the step of writing the to-be-detected text-audio pair into the training database, the method further comprises:

extracting a sample text-audio pair from the training database, wherein the sample text-audio pair comprises a sample text segment and a sample audio segment; and

training a speech synthesis model based on the sample text segment and the sample audio segment, to obtain a target speech synthesis model.

14. A sample generation device, comprising:

an acquisition module, configured to a plurality of text-audio pairs, wherein each text-audio pair comprises a text segment and an audio segment;

a calculation module, configured to calculate, for each text-audio pair among the plurality of text-audio pairs, an audio feature of the audio segment of the text-audio pair, and

screening out from the plurality of text-audio pairs, according to the audio feature, a target text-audio pair and a splicing text-audio pair corresponding to the target text-audio pair;

a splicing module, configured to splice the target text-audio pair and the splicing text-audio pair into a to-be-detected text-audio pair, and detect the to-be-detected text-audio pair; and

a writing module, configured to write the to-be-detected text-audio pair into a training database in a case that the to-be-detected text-audio pair meets a preset detection condition.

15. A computing device, comprising:

a memory configured to store a computer executable instruction; and

a processor configured to execute the computer executable instruction to perform a method comprising the following steps of:

acquiring a plurality of text-audio pairs, wherein each text-audio pair comprises a text segment and an audio segment;

calculating, for each text-audio pair among the plurality of text-audio pairs, an audio feature of the audio segment of the text-audio pair, and screening out from the plurality of text-audio pairs, according to the audio feature, a target text-audio pair and a splicing text-audio pair corresponding to the target text-audio pair;

splicing the target text-audio pair and the splicing text-audio pair into a to-be-detected text-audio pair, and detecting the to-be-detected text-audio pair; and

writing the to-be-detected text-audio pair into a training database in a case that the to-be-detected text-audio pair meets a preset detection condition.

16. A computer readable storage medium storing a computer instruction that is executed by a processor to perform the steps of the sample generation method according to any of claims 1 to 13.