What is claimed is:

1. A text-based speech synthesis method, comprising:

obtaining a target text to be recognized;

discretely characterizing each character in the target text to generate a feature vector corresponding to each character;

obtaining a preset number of training texts and matching speech corresponding to the training texts;

discretely characterizing the training text to obtain a feature vector corresponding to each character in the training text;

inputting the feature vector corresponding to each character in the training text into a spectrum conversion model to be trained to obtain a Mel-spectrum output by the spectrum conversion model to be trained;

when an error between the Mel-spectrum output by the spectrum conversion model to be trained and a Mel-spectrum corresponding to the matching speech is less than or equal to a preset threshold, obtaining a trained spectrum conversion model;

inputting the feature vector into a pre-trained spectrum conversion model to obtain a Mel-spectrum corresponding to each character in the target text output by the spectrum conversion model; and

converting the Mel-spectrum into speech to obtain speech corresponding to the target text.

2. The method as claimed in claim 1, wherein inputting the feature vector corresponding to each character in the training text into the spectrum conversion model to be trained to obtain the Mel-spectrum output by the spectrum conversion model to be trained comprises:

coding the training text through the spectrum conversion model to be trained to obtain a hidden state sequence corresponding to the training text, the hidden state sequence comprising at least two hidden nodes;

according to a weight of a hidden node corresponding to each character, weighting the hidden node to obtain a semantic vector corresponding to each character in a training sample; and

decoding the semantic vector corresponding to each character, and outputting the Mel-spectrum corresponding to each character.

3. The method as claimed in claim 1, wherein after inputting the feature vector corresponding to each character in the training text into the spectrum conversion model to be trained to obtain the Mel-spectrum output by the spectrum conversion model to be trained, the

method further comprises:

when the error between the Mel-spectrum output by the spectrum conversion model to be trained and the Mel-spectrum corresponding to the matching speech is greater than the preset threshold, updating the weight of each hidden node;

weighting the hidden node whose weight is updated to obtain a semantic vector corresponding to each character in a training sample;

decoding the semantic vector corresponding to each character, and outputting the Mel-spectrum corresponding to each character; and

when the error between the Mel-spectrum corresponding to each character and the Mel-spectrum corresponding to the matching speech is less than or equal to the preset threshold, stopping the process of updating the weight of each hidden node, and obtaining the trained spectrum conversion mode.

4. The method as claimed in any one of claims 1 to 3, wherein converting the Mel-spectrum into speech to obtain the speech corresponding to the target text comprises:

performing inverse Fourier transform on the Mel-spectrum through a vocoder to convert the Mel-spectrum into a speech waveform signal in the time domain to obtain the speech.

5. A text-based speech synthesis device, comprising:

an obtaining module, configured to obtain a target text to be recognized and a feature vector corresponding to each character in the target text that is discretely characterized by a processing module, obtain a preset number of training texts and matching speech corresponding to the training texts, discretely characterize the training text to obtain a feature vector corresponding to each character in the training text, input the feature vector corresponding to each character in the training text into a spectrum conversion model to be trained to obtain a Mel-spectrum output by the spectrum conversion model to be trained, and when an error between the Mel-spectrum output by the spectrum conversion model to be trained and a Mel-spectrum corresponding to the matching speech is less than or equal to a preset threshold, obtain a trained spectrum conversion model, and input the feature vector corresponding to each character in the target text into a pre-trained spectrum conversion model to obtain a Mel-spectrum corresponding to each character in the target text output by the spectrum conversion model; a processing module, configured to discretely characterize the target text to obtain the feature vector corresponding to each character in the target text;

a converting module, configured to convert the Mel-spectrum obtained by the obtaining module into speech to obtain speech corresponding to the target text.

6. The device as claimed in claim 5, wherein the obtaining module is configured to code

the training text through the spectrum conversion model to be trained to obtain a hidden state sequence corresponding to the training text, the hidden state sequence comprising at least two hidden nodes, according to a weight of a hidden node corresponding to each character, weight the hidden node to obtain a semantic vector corresponding to each character in a training sample, and decode the semantic vector corresponding to each character and output the Mel-spectrum corresponding to each character.

7. The device as claimed in claim 5, wherein the obtaining module is configured to, when the error between the Mel-spectrum output by the spectrum conversion model to be trained and the Mel-spectrum corresponding to the matching speech is greater than the preset threshold, update the weight of each hidden node, weight the hidden node whose weight is updated to obtain a semantic vector corresponding to each character in a training sample, decode the semantic vector corresponding to each character and output the Mel-spectrum corresponding to each character, and

when the error between the Mel-spectrum corresponding to each character and the Mel-spectrum corresponding to the matching speech is less than or equal to the preset threshold, stop the process of updating the weight of each hidden node and obtain the trained spectrum conversion mode.

8. The device as claimed in any one of claims 5 to 7, wherein the converting module is configured to perform inverse Fourier transform on the Mel-spectrum through a vocoder to convert the Mel-spectrum into a speech waveform signal in the time domain to obtain the speech.

9. A computer device, comprising: a memory, a processor and a computer program stored in the memory and capable of running on the processor, wherein the computer program, when executed by the processor, causes the processor to implement:

obtaining a target text to be recognized;

discretely characterizing each character in the target text to generate a feature vector corresponding to each character;

obtaining a preset number of training texts and matching speech corresponding to the training texts;

discretely characterizing the training text to obtain a feature vector corresponding to each character in the training text;

inputting the feature vector corresponding to each character in the training text into a spectrum conversion model to be trained to obtain a Mel-spectrum output by the spectrum conversion model to be trained;

when an error between the Mel-spectrum output by the spectrum conversion model to be trained and a Mel-spectrum corresponding to the matching speech is less than or equal to a preset threshold, obtaining a trained spectrum conversion model;

inputting the feature vector into a pre-trained spectrum conversion model to obtain a Mel-spectrum corresponding to each character in the target text output by the spectrum conversion model; and

converting the Mel-spectrum into speech to obtain speech corresponding to the target text.

10. The computer device as claimed in claim 9, wherein the computer program, when executed by the processor, causing the processor to implement inputting the feature vector corresponding to each character in the training text into the spectrum conversion model to be trained to obtain the Mel-spectrum output by the spectrum conversion model to be trained causes the processor to implement:

coding the training text through the spectrum conversion model to be trained to obtain a hidden state sequence corresponding to the training text, the hidden state sequence comprising at least two hidden nodes;

according to a weight of a hidden node corresponding to each character, weighting the hidden node to obtain a semantic vector corresponding to each character in a training sample; and

decoding the semantic vector corresponding to each character, and outputting the Mel-spectrum corresponding to each character.

11. The computer device as claimed in claim 9, wherein after inputting the feature vector corresponding to each character in the training text into the spectrum conversion model to be trained to obtain the Mel-spectrum output by the spectrum conversion model to be trained, the computer program, when executed by the processor, further causes the processor to implement:

when the error between the Mel-spectrum output by the spectrum conversion model to be trained and the Mel-spectrum corresponding to the matching speech is greater than the preset threshold, updating the weight of each hidden node;

weighting the hidden node whose weight is updated to obtain a semantic vector corresponding to each character in a training sample;

decoding the semantic vector corresponding to each character, and outputting the Mel-spectrum corresponding to each character; and

when the error between the Mel-spectrum corresponding to each character and the

Mel-spectrum corresponding to the matching speech is less than or equal to the preset threshold, stopping the process of updating the weight of each hidden node, and obtaining the trained spectrum conversion mode.

12. The computer device as claimed in any one of claims 9 to 11, wherein the computer program, when executed by the processor, causing the processor to implement converting the Mel-spectrum into speech to obtain the speech corresponding to the target text causes the processor to implement:

performing inverse Fourier transform on the Mel-spectrum through a vocoder to convert the Mel-spectrum into a speech waveform signal in the time domain to obtain the speech.

13. A non-transitory computer-readable storage medium, storing a computer program, wherein the computer program, when executed by a processor, causes the processor to implement:

obtaining a target text to be recognized;

discretely characterizing each character in the target text to generate a feature vector corresponding to each character;

obtaining a preset number of training texts and matching speech corresponding to the training texts;

discretely characterizing the training text to obtain a feature vector corresponding to each character in the training text;

inputting the feature vector corresponding to each character in the training text into a spectrum conversion model to be trained to obtain a Mel-spectrum output by the spectrum conversion model to be trained;

when an error between the Mel-spectrum output by the spectrum conversion model to be trained and a Mel-spectrum corresponding to the matching speech is less than or equal to a preset threshold, obtaining a trained spectrum conversion model;

inputting the feature vector into a pre-trained spectrum conversion model to obtain a Mel-spectrum corresponding to each character in the target text output by the spectrum conversion model; and

converting the Mel-spectrum into speech to obtain speech corresponding to the target text.

14. The non-transitory computer-readable storage medium as claimed in claim 13, wherein the computer program, when executed by the processor, causing the processor to implement inputting the feature vector corresponding to each character in the training text into the spectrum conversion model to be trained to obtain the Mel-spectrum output by the

spectrum conversion model to be trained causes the processor to implement:

coding the training text through the spectrum conversion model to be trained to obtain a hidden state sequence corresponding to the training text, the hidden state sequence comprising at least two hidden nodes;

according to a weight of a hidden node corresponding to each character, weighting the hidden node to obtain a semantic vector corresponding to each character in a training sample; and

decoding the semantic vector corresponding to each character, and outputting the Mel-spectrum corresponding to each character.

15. The non-transitory computer-readable storage medium as claimed in claim 13, wherein after inputting the feature vector corresponding to each character in the training text into the spectrum conversion model to be trained to obtain the Mel-spectrum output by the spectrum conversion model to be trained, the computer program, when executed by the processor, further causes the processor to implement:

when the error between the Mel-spectrum output by the spectrum conversion model to be trained and the Mel-spectrum corresponding to the matching speech is greater than the preset threshold, updating the weight of each hidden node;

weighting the hidden node whose weight is updated to obtain a semantic vector corresponding to each character in a training sample;

decoding the semantic vector corresponding to each character, and outputting the Mel-spectrum corresponding to each character; and

when the error between the Mel-spectrum corresponding to each character and the Mel-spectrum corresponding to the matching speech is less than or equal to the preset threshold, stopping the process of updating the weight of each hidden node, and obtaining the trained spectrum conversion mode.

16. The non-transitory computer-readable storage medium as claimed in any one of claims 13 to 15, wherein the computer program, when executed by the processor, causing the processor to implement converting the Mel-spectrum into speech to obtain the speech corresponding to the target text causes the processor to implement:

performing inverse Fourier transform on the Mel-spectrum through a vocoder to convert the Mel-spectrum into a speech waveform signal in the time domain to obtain the speech.