CLAIMS

1. A voice conversion system, comprising:

a speaker-independent automatic speech recognition model, comprising at least a bottleneck layer, configured to: convert a mel-scale frequency cepstral coefficients feature of an inputted source speech into a bottleneck feature of the source speech through the bottleneck layer, and output the bottleneck feature of the source speech to an Attention voice conversion network through the bottleneck layer;

where a training method for the speaker-independent automatic speech recognition model comprises:

inputting a number of a character encoding to which a word in a multi-speaker speech recognition training corpus is converted, together with a mel-scale frequency cepstral coefficients feature of the multi-speaker speech recognition training corpus, to the speaker-independent automatic speech recognition model; executing a backward propagation algorithm; and performing iterative optimization until the speaker-independent automatic speech recognition model is converged;

the Attention voice conversion network configured to convert the bottleneck feature of the source speech into a mel-scale frequency cepstral coefficients feature in conformity with a target speech; and

a neural network vocoder configured to convert the mel-scale frequency cepstral coefficients feature in conformity with the target speech into and output audio.

2. A voice conversion method, characterized by comprising steps of:

by a speaker-independent automatic speech recognition model, converting a mel-scale frequency cepstral coefficients feature of an inputted source speech into a bottleneck feature of the source speech through the bottleneck layer, and outputting the bottleneck feature of the source speech to an Attention voice conversion network through the bottleneck layer;

where a training method for the speaker-independent automatic speech recognition model

comprises:

inputting a number of a character encoding to which a word in a multi-speaker speech recognition training corpus is converted, together with a mel-scale frequency cepstral coefficients feature of the multi-speaker speech recognition training corpus, to the speaker-independent automatic speech recognition model; executing a backward propagation algorithm; and performing iterative optimization until the speaker-independent automatic speech recognition model is converged;

converting, by the Attention voice conversion network, the bottleneck feature of the source speech into a mel-scale frequency cepstral coefficients feature in conformity with a target speech; and

converting, by a neural network vocoder, the mel-scale frequency cepstral coefficients feature in conformity with the target speech into and output audio.

3. A training method for a voice conversion system, applicable to the voice conversion system according to claim 1, characterized by that, a training method for the speaker-independent automatic speech recognition model comprises :

inputting a number of a character encoding to which a word in a multi-speaker speech recognition training corpus is converted, together with a mel-scale frequency cepstral coefficients feature of the multi-speaker speech recognition training corpus, to the speaker-independent automatic speech recognition model; executing a backward propagation algorithm; and performing iterative optimization until the speaker-independent automatic speech recognition model is converged.

4. The training method for the voice conversion system according to claim 3, characterized by that, the training method for the speaker-independent automatic speech recognition model further comprises preprocessing the multi-speaker speech recognition training corpus, wherein the preprocessing comprises de-blanking and normalizing.

5. A training method for a voice conversion system, applicable to the voice conversion system according to claim 1, characterized by that, a training method for the Attention voice conversion network comprises :

converting a mel-scale frequency cepstral coefficients feature of a target speech into a bottleneck feature of the target speech;

inputting the bottleneck feature of the target speech to a basic Attention voice conversion network, with the mel-scale frequency cepstral coefficients feature corresponding to the target speaker as a ground truth, and training the Attention voice conversion network by a method of Deep Transfer Learning.

6. The training method for the voice conversion system according to claim 5, characterized by that, the converting of the mel-scale frequency cepstral coefficients feature of the target speech into the bottleneck feature of the target speech is performed by using the speaker-independent automatic speech recognition model that is pre-trained.

7. A training method for a voice conversion system, applicable to the voice conversion system according to claim 1, characterized by that, a training method for the neural network vocoder comprises :

inputting a mel-scale frequency cepstral coefficients feature of a target speech and the target speech to the neural network vocoder that is pre-trained; and training the neural network vocoder by a method of Deep Transfer Learning.

8. The training method for the voice conversion system according to any one of claims 5 to 7, characterized by further comprising preprocessing the target speech, wherein the preprocessing comprises de-blanking and normalizing.

9. The training method for the voice conversion system according to any one of claims 3 to

7, characterized by further comprising extracting a parameter to obtain the mel-scale frequency cepstral coefficients feature of the multi-speaker speech recognition training corpus, a mel-scale frequency cepstral coefficients feature of a target speech, and a mel-scale frequency cepstral coefficients feature of a source speech.

10. A terminal, characterized by, comprising the voice conversion system according to claim 1.

11. A computer-readable storage medium, characterized by that, a computer program is stored in the computer-readable storage medium, wherein the computer program is configured to, when executed, perform the method according to any one of claims 3 to 7.