CLAIMS

1. A voiceprint recognition method, comprising:

extracting a first spectral feature from an audio signal;

independent from the extracting of the first spectral feature, inputting data corresponding to the audio signal to a memory deep neural network (DNN), the data corresponding to the audio signal being different from the first spectral feature, the memory DNN including a plurality of hidden layers that include at least one temporal recurrent layer and a bottleneck layer, an output of the at least one temporal recurrent layer being connected to the bottleneck layer, a feature extracted from the bottleneck layer indicating time sequence information, the bottleneck layer having a smallest number of dimensions among the plurality of hidden layers in the memory DNN;

extracting a bottleneck feature from the bottleneck layer of the memory DNN that received the data corresponding to the audio signal as input to the memory DNN;

forming an acoustic feature, the acoustic feature including both the first spectral feature and the bottleneck feature;

extracting an identity authentication vector based, at least in part, on the acoustic feature; and

performing speaker recognition using a classification model and based, at least in part, on the identity authentication vector.

2. The method according to claim 1, wherein the first spectral feature comprises a Mel frequency cepstral coefficient (MFCC) feature, and a first order difference feature and a second order difference feature of the MFCC feature.

3. The method according to claim 1, wherein the at least one temporal recurrent layer comprises a hidden layer based on a long-short term memory (LSTM) model or a hidden layer based on a long short-term memory projection (LSTMP) model.

4. The method according to claim 1, wherein the at least one temporal recurrent layer comprises a hidden layer based on a feedforward sequence memory (FSMN) model or a hidden layer based on a compact feedforward sequence memory (cFSMN) model.

5. The method according to claim 1, wherein:

the audio signal includes a plurality of consecutive speech frames; and

the inputting of the data corresponding to the audio signal to the memory deep neural network (DNN) comprises:

extracting a second spectral feature from the plurality of consecutive speech frames; and

inputting the second spectral feature to the memory DNN.

6. The method according to claim 5, wherein the second spectral feature is a Mel scale filter bank (FBank) feature.

7. The method according to claim 1, wherein the forming the acoustic feature of the speaker audio comprises: concatenating the first spectral feature and the bottleneck feature to form the acoustic feature.

8. A voiceprint recognition device, comprising:

a first extraction unit, configured to extract a first spectral feature from an audio signal;

a second extraction unit independent from the first extraction unit, configured to: input data corresponding to the audio signal to a memory deep neural network (DNN) and extract a bottleneck feature from a bottleneck layer of the memory DNN, wherein the data corresponding to the audio signal is different from the first spectral feature, the memory DNN includes at least one temporal recurrent layer and the bottleneck layer, an output of the at least one temporal recurrent layer is connected to the bottleneck layer, a feature extracted from the bottleneck layer indicates time sequence information, and a number of dimensions of the bottleneck layer is less than a number of dimensions of any other hidden layer in the memory DNN;

a feature combining unit, configured to form an acoustic feature, wherein the acoustic feature includes both the first spectral feature and the bottleneck feature;

a vector extraction unit, configured to extract an identity authentication vector based, at least in part, on the acoustic feature; and

a classification recognition unit, configured to perform speaker recognition using a classification model and based, at least in part, on the identity authentication vector.

9. The device according to claim 8, wherein the first spectral feature extracted by the first extraction unit comprises a Mel frequency cepstral coefficient (MFCC) feature, and a first order difference feature and a second order difference feature of the MFCC feature.

10. The device according to claim 8, wherein the at least one temporal recurrent layer comprises a hidden layer based on a long-short term memory (LSTM) model or a hidden layer based on a long short-term memory projection (LSTMP) model.

11. The device according to claim 8, wherein the at least one temporal recurrent layer comprises a hidden layer based on a feedforward sequence memory (FSMN) model or a hidden layer based on a compact feedforward sequence memory (cFSMN) model.

12. The device according to claim 8, wherein the second extraction unit is further configured to: extract a second spectral feature from a plurality of consecutive speech frames of the audio signal, and input the second spectral feature to the memory DNN.

13. The device according to claim 12, wherein the second spectral feature is a Mel scale filter bank (FBank) feature.

14. The device according to claim 8, wherein the feature combining unit is configured to concatenate the first spectral feature and the bottleneck feature to form the acoustic feature.

15. A computer-readable storage medium storing contents that, when executed by one or more processors, cause the one or more processors to perform actions comprising:

inputting data corresponding to an audio signal into a memory deep neural network (DNN), the DNN including at least a bottleneck layer and a temporal recurrent layer, wherein an output of the temporal recurrent layer is connected to an input of the bottleneck layer, wherein at least a feature extracted from the bottleneck layer conveys time sequence information;

obtaining a bottleneck feature from the bottleneck layer based, at least in part, on processing of the audio data using the memory DNN;

forming a combined feature, wherein the combined feature includes both (a) the bottleneck feature and (b) an other spectral feature derived from the audio signal independently from the inputting of the data corresponding to the audio signal into the DNN, the other spectral feature being different from the data corresponding to the audio signal that was input into the DNN; and

causing performing of speaker recognition based, at least in part, on the combined feature.

16. The computer-readable storage medium of claim 15, wherein the obtaining the bottleneck feature comprises accessing activation values of a plurality of nodes of the bottleneck layer.

17. The computer-readable storage medium of claim 16, wherein the bottleneck feature includes a vector formed by at least a subset of the activation values.

18. The computer-readable storage medium of claim 15, wherein the other spectral feature includes a single-frame spectral feature.

19. The computer-readable storage medium of claim 15, wherein the temporal recurrent layer is a first temporal recurrent layer and wherein the memory DNN

includes a second temporal recurrent that feeds directly or indirectly into the first temporal recurrent layer.

20. The computer-readable medium of claim 15, wherein the combined feature is larger in dimension than at least one of the bottleneck feature or the other feature.