WHAT IS CLAIMED IS:

1. An electronic device, comprising a memory and a processor connected with the memory, wherein the memory stores a system of identity verification which may be operated in the processor, and the system of identity verification is executed by the processor to implement the following steps:

S1, after current voice data of a target user to be subjected to identity verification are received, carrying out framing processing on the current voice data according to preset framing parameters to obtain multiple voice frames;

S2, extracting preset types of acoustic features in all the voice frames by using a predetermined filter, and generating multiple observed feature units corresponding to the current voice data according to the extracted acoustic features;

S3, pairwise coupling all the observed feature units with pre-stored observed feature units respectively to obtain multiple groups of coupled observed feature units;

S4, inputting the multiple groups of coupled observed feature units into a preset type of identity verification model generated by pre-training, and obtaining an output identity verification result to carry out the identity verification on the target user;

wherein the preset type of identity verification model generated by pre-training is a deep convolution neural network model which carries out the identity verification by adopting a recognition function, and the recognition function comprises:

$$Obj = -\sum_{x,y\in same} In(P(x,y)) - K\sum_{x,y\in diff} In(1-P(x,y))$$;

$$P(x,y) = \frac{1}{1+e^{-L(x,y)}}$$;

$$L(x,y) = x^T Uy - x^T Vx - y^T Vy + b$$,

where *Obj* is a target function of the deep convolution neural network model; x is a user feature obtained by one observed feature unit in one group of observed feature units on a normalization layer; y is a user feature obtained by the other observed feature unit in the group of observed feature units on the normalization layer; K is a constant; *P(x, y)* is a calculated probability that one group of observed feature units belong to the same user; *L(x, y)* is a calculated similarity L of one group of observed feature units; U represents an intra-class relation matrix of the user; V represents an inter-class relation matrix of the user; b represents an offset; and T represents matrix transposition.

2. The electronic device of claim 1, wherein the predetermined filter is a Mel filter, and the step of extracting the preset types of acoustic features in all the voice frames by using the predetermined filter comprises:

carrying out windowing processing on the voice frames;

carrying out Fourier transformation on each window to obtain a corresponding frequency spectrum;

inputting the frequency spectrums into the Mel filter so as to output Mel frequency spectrums;

carrying out cepstrum analysis on the Mel frequency spectrums to obtain Mel Frequency Cepstrum Coefficients (MFCCs) which serve as the acoustic features of the voice frames.

3. The electronic device of claim 2, wherein the step of generating the multiple observed feature units corresponding to the current voice data according to the extracted acoustic features comprises:

forming a voice frame set by all the voice frames in each recording datum in the current voice data, splicing 20 dimensions of MFCCs of each voice frame in the voice frame set according to the sequence of framing moments of the corresponding voice frames, and generating the observed feature units of a corresponding (20, N)-dimension matrix, wherein N is the total number of the frames in the voice frame set.

4. The electronic device of claim 1, wherein before the step S4, the method further comprises:

obtaining a first preset number of voice pairs of the same user, and obtaining a second preset number of voice pairs of different users, and carrying out the framing processing on voices in all the voice pairs respectively according to the preset framing parameters to obtain multiple voice frames corresponding to all the voice pairs;

extracting preset types of acoustic features in all the voice frames by using a predetermined filter, and generating multiple observed feature units of all the voice pairs according to the extracted acoustic features;

pairwise coupling the observed feature units corresponding to two voices belonging to the same user and two voices belonging to different users to obtain the multiple groups of coupled observed feature units;

dichotomizing each voice into a first percent of a training set and a second percent of a verification set, wherein the sum of the first percent and the second percent is less than or equal to 1;

training a deep convolution neural network model by using all the groups of observed feature units of all the voice pairs in the training set, and after the training is

completed, verifying the accuracy of the trained deep convolution neural network model by using the verification set;

if the accuracy is greater than a preset threshold value, after the training is completed, taking the trained deep convolution neural network model as the deep convolution neural network model in the step S4, or if the accuracy is less than or equal to the preset threshold value, increasing the number of the voice pairs for training for retraining.

5. A method of identity verification, comprising:

S1, after current voice data of a target user to be subjected to identity verification are received, carrying out framing processing on the current voice data according to preset framing parameters to obtain multiple voice frames;

S2, extracting preset types of acoustic features in all the voice frames by using a predetermined filter, and generating multiple observed feature units corresponding to the current voice data according to the extracted acoustic features;

S3, pairwise coupling all the observed feature units with pre-stored observed feature units respectively to obtain multiple groups of coupled observed feature units;

S4, inputting the multiple groups of coupled observed feature units into a preset type of identity verification model generated by pre-training, and obtaining an output identity verification result to carry out the identity verification on the target user;

wherein the preset type of identity verification model generated by pre-training is a deep convolution neural network model which carries out the identity verification by adopting a recognition function, and the recognition function comprises:

$$Obj = -\sum_{x,y\in same} In(P(x,y)) - K\sum_{x,y\in diff} In(1-P(x,y))$$;

$$P(x,y)=\frac{1}{1+e^{-L(x,y)}};$$

$$L(x,y)=x^{T}Uy-x^{T}Vx-y^{T}Vy+b,$$

where *Obj* is a target function of the deep convolution neural network model; x is a user feature obtained by one observed feature unit in one group of observed feature units on a normalization layer; y is a user feature obtained by the other observed feature unit in the group of observed feature units on the normalization layer; K is a constant; *P(x, y)* is a calculated probability that one group of observed feature units belong to the same user; *L(x, y)* is a calculated similarity L of one group of observed feature units; U represents an intra-class relation matrix of the user; V represents an inter-class relation matrix of the user; b represents an offset; and T represents matrix transposition.

6. The method of identity verification of claim 5, wherein the predetermined filter is a Mel filter, and the step of extracting the preset types of acoustic features in all the voice frames by using the predetermined filter comprises:

carrying out windowing processing on the voice frames;

carrying out Fourier transformation on each window to obtain a corresponding frequency spectrum;

inputting the frequency spectrums into the Mel filter so as to output Mel frequency spectrums;

carrying out cepstrum analysis on the Mel frequency spectrums to obtain Mel Frequency Cepstrum Coefficients (MFCCs) which serve as the acoustic features of the voice frames.

7. The method of identity verification of claim 5, wherein the step of generating the multiple observed feature units corresponding to the current voice data according to the extracted acoustic features comprises:

forming a voice frame set by all the voice frames in each recording datum in the current voice data, splicing 20 dimensions of MFCCs of each voice frame in the voice frame set according to the sequence of framing moments of the corresponding voice frames, and generating the observed feature units of a corresponding (20, N)-dimension matrix, wherein N is the total number of the frames in the voice frame set.

8. The method of identity verification of claim 7, wherein before the step S4, the method further comprises:

obtaining a first preset number of voice pairs of the same user, and obtaining a second preset number of voice pairs of different users, and carrying out the framing processing on voices in all the voice pairs respectively according to the preset framing parameters to obtain multiple voice frames corresponding to all the voice pairs;

extracting preset types of acoustic features in all the voice frames by using a predetermined filter, and generating multiple observed feature units of all the voice pairs according to the extracted acoustic features;

pairwise coupling the observed feature units corresponding to two voices belonging to the same user and two voices belonging to different users to obtain the multiple groups of coupled observed feature units;

dichotomizing each voice into a first percent of a training set and a second percent of a verification set, wherein the sum of the first percent and the second percent is less than or equal to 1;

training a deep convolution neural network model by using all the groups of observed feature units of all the voice pairs in the training set, and after the training is completed, verifying the accuracy of the trained deep convolution neural network model by using the verification set;

if the accuracy is greater than a preset threshold value, after the training is completed, taking the trained deep convolution neural network model as the deep convolution neural network model in the step S4, or if the accuracy is less than or equal to the preset threshold value, increasing the number of the voice pairs for training for retraining.

9. A computer readable storage medium, which stores a system of identity verification, wherein the system of identity verification may be executed by at least one processor to implement the following steps:

S1, after current voice data of a target user to be subjected to identity verification are received, carrying out framing processing on the current voice data according to preset framing parameters to obtain multiple voice frames;

S2, extracting preset types of acoustic features in all the voice frames by using a predetermined filter, and generating multiple observed feature units corresponding to the current voice data according to the extracted acoustic features;

S3, pairwise coupling all the observed feature units with pre-stored observed feature units respectively to obtain multiple groups of coupled observed feature units;

S4, inputting the multiple groups of coupled observed feature units into a preset type of identity verification model generated by pre-training, and obtaining an output identity verification result to carry out the identity verification on the target user;

wherein the preset type of identity verification model generated by pre-training is a deep convolution neural network model which carries out the identity verification by adopting a recognition function, and the recognition function comprises:

$$Obj = -\sum_{x,y\in same} In(P(x,y)) - K\sum_{x,y\in diff} In(1-P(x,y))$$;

$$P(x,y) = \frac{1}{1+e^{-L(x,y)}}$$;

$$L(x,y) = x^T Uy - x^T Vx - y^T Vy + b$$,

where *Obj* is a target function of the deep convolution neural network model; x is a user feature obtained by one observed feature unit in one group of observed feature units on a normalization layer; y is a user feature obtained by the other observed feature unit in the group of observed feature units on the normalization layer; K is a constant; *P(x, y)* is a calculated probability that one group of observed feature units belong to the same user; *L(x, y)* is a calculated similarity L of one group of observed feature units; U represents an intra-class relation matrix of the user; V represents an inter-class relation matrix of the user; b represents an offset; and T represents matrix transposition.

10. The computer readable storage medium of claim 9, wherein the predetermined filter is a Mel filter, and the step of extracting the preset types of acoustic features in all the voice frames by using the predetermined filter comprises:

carrying out windowing processing on the voice frames;

carrying out Fourier transformation on each window to obtain a corresponding frequency spectrum;

inputting the frequency spectrums into the Mel filter so as to output Mel frequency spectrums;

carrying out cepstrum analysis on the Mel frequency spectrums to obtain Mel Frequency Cepstrum Coefficients (MFCCs) which serve as the acoustic features of the voice frames.

11. The computer readable storage medium of claim 10, wherein the step of generating the multiple observed feature units corresponding to the current voice data according to the extracted acoustic features comprises:

forming a voice frame set by all the voice frames in each recording datum in the current voice data, splicing 20 dimensions of MFCCs of each voice frame in the voice frame set according to the sequence of framing moments of the corresponding voice frames, and generating the observed feature units of a corresponding (20, N)-dimension matrix, wherein N is the total number of the frames in the voice frame set.

12. The computer readable storage medium of claim 9, wherein before the step S4, the method further comprises:

obtaining a first preset number of voice pairs of the same user, and obtaining a second preset number of voice pairs of different users, and carrying out the framing processing on voices in all the voice pairs respectively according to the preset framing parameters to obtain multiple voice frames corresponding to all the voice pairs;

extracting preset types of acoustic features in all the voice frames by using the predetermined filter, and generating multiple observed feature units of all the voice pairs according to the extracted acoustic features;

pairwise coupling the observed feature units corresponding to two voices belonging to the same user and two voices belonging to different users to obtain the multiple groups of coupled observed feature units;

dichotomizing each voice into a first percent of a training set and a second percent of a verification set, wherein the sum of the first percent and the second percent is less than or equal to 1;

training a deep convolution neural network model by using all the groups of observed feature units of all the voice pairs in the training set, and after the training is completed, verifying the accuracy of the trained deep convolution neural network model by using the verification set;

if the accuracy is greater than a preset threshold value, after the training is completed, taking the trained deep convolution neural network model as the deep convolution neural network model in the step S4, or if the accuracy is less than or equal to the preset threshold value, increasing the number of the voice pairs for training for retraining.

13. A system of identity verification, which is stored in a memory, and may be executed by at least one processor to implement the following steps:

S1, after current voice data of a target user to be subjected to identity verification are received, carrying out framing processing on the current voice data according to preset framing parameters to obtain multiple voice frames;

S2, extracting preset types of acoustic features in all the voice frames by using a predetermined filter, and generating multiple observed feature units corresponding to the current voice data according to the extracted acoustic features;

S3, pairwise coupling all the observed feature units with pre-stored observed feature units respectively to obtain multiple groups of coupled observed feature units;

S4, inputting the multiple groups of coupled observed feature units into a preset type of identity verification model generated by pre-training, and obtaining an output identity verification result to carry out the identity verification on the target user;

wherein the preset type of identity verification model generated by pre-training is a deep convolution neural network model which carries out the identity verification by adopting a recognition function, and the recognition function comprises:

$$Obj = -\sum_{x,y\in same} In(P(x,y)) - K\sum_{x,y\in diff} In(1-P(x,y))\ ;$$

$$P(x,y) = \frac{1}{1+e^{-L(x,y)}}\ ;$$

$$L(x,y) = x^T Uy - x^T Vx - y^T Vy + b\ ,$$

where *Obj* is a target function of the deep convolution neural network model; x is a user feature obtained by one observed feature unit in one group of observed feature units on a normalization layer; y is a user feature obtained by the other observed feature unit in the group of observed feature units on the normalization layer; K is a constant; *P(x, y)* is a calculated probability that one group of observed feature units belong to the same user; *L(x, y)* is a calculated similarity L of one group of observed feature units; U represents an intra-class relation matrix of the user; V represents an inter-class relation matrix of the user; b represents an offset; and T represents matrix transposition.

14. The system of identity verification of claim 13, wherein the predetermined filter is a Mel filter, and the step of extracting the preset types of acoustic features in all the voice frames by using the predetermined filter comprises:

carrying out windowing processing on the voice frames;

carrying out Fourier transformation on each window to obtain a corresponding frequency spectrum;

inputting the frequency spectrums into the Mel filter so as to output Mel frequency spectrums;

carrying out cepstrum analysis on the Mel frequency spectrums to obtain Mel Frequency Cepstrum Coefficients (MFCCs) which serve as the acoustic features of the voice frames.

15. The system of identity verification of claim 14, wherein the step of generating the multiple observed feature units corresponding to the current voice data according to the extracted acoustic features comprises:

forming a voice frame set by all the voice frames in each recording datum in the current voice data, splicing 20 dimensions of MFCCs of each voice frame in the voice frame set according to the sequence of framing moments of the corresponding voice frames, and generating the observed feature units of a corresponding (20, N)-dimension matrix, wherein N is the total number of the frames in the voice frame set.

16. The system of identity verification of claim 13, wherein before the step S4, the method further comprises:

obtaining a first preset number of voice pairs of the same user, and obtaining a second preset number of voice pairs of different users, and carrying out the framing processing on voices in all the voice pairs respectively according to the preset framing parameters to obtain multiple voice frames corresponding to all the voice pairs;

extracting preset types of acoustic features in all the voice frames by using the predetermined filter, and generating multiple observed feature units of all the voice pairs according to the extracted acoustic features;

pairwise coupling the observed feature units corresponding to two voices belonging to the same user and two voices belonging to different users to obtain the multiple groups of coupled observed feature units;

dichotomizing each voice into a first percent of a training set and a second percent of a verification set, wherein the sum of the first percent and the second percent is less than or equal to 1;

training a deep convolution neural network model by using all the groups of observed feature units of all the voice pairs in the training set, and after the training is completed, verifying the accuracy of the trained deep convolution neural network model by using the verification set;

if the accuracy is greater than a preset threshold value, after the training is completed, taking the trained deep convolution neural network model as the deep convolution neural network model in the step S4, or if the accuracy is less than or equal to the preset threshold value, increasing the number of the voice pairs for training for retraining.