CLAIMS

1. An acoustic model training method, comprising:

(i) performing feature extraction on a training speech signal to obtain an audio feature sequence;

(ii) training the audio feature sequence by using a phoneme mixed Gaussian Model-Hidden Markov Model to obtain a phoneme feature sequence, wherein training the audio feature sequence includes:

(a) obtaining an original monophone truth corresponding to each audio feature in the audio feature sequence based on an original monophone Mixture Gaussian Model-Hidden Markov Model;

(b) iteratively training the original monophone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original monophone truth to obtain a target monophone Mixture Gaussian Model-Hidden Markov Model;

(c) aligning each original monophone truth based on the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain a target monophone feature; and

(d) obtaining the phoneme feature sequence based on the target monophone feature; and

(iii) training the obtained phoneme feature sequence by using a Deep Neural Net-Hidden Markov Model-sequence training model to obtain a target acoustic model.

2. The acoustic model training method according to claim 1, wherein performing feature extraction on the training speech signal in (i) comprises:

performing pre-emphasis, framing, windowing, endpoint detection, fast Fourier transform, and Mel filter bank and discrete cosine transform on the training speech signal.

3. The acoustic model training method according to claim 1, wherein prior to obtaining the original monophone truth in (ii)(a), the method further comprises:

calculating mean and variance of some of the audio feature sequences to obtain the original monophone Mixture Gaussian Model-Hidden Markov Model.

4. The acoustic model training method according to claim 1, wherein training of the obtained phoneme feature sequence by the Deep Neural Net-Hidden Markov Model-sequence in (iii) produces an output phoneme feature sequence, and the method further comprises:

training the output phoneme feature sequence by using a triphone Mixture Gaussian Model-Hidden Markov Model to obtain an updated phoneme feature sequence;

wherein training the phoneme feature sequence by using a triphone Mixture Gaussian Model-Hidden Markov Model to obtain the updated phoneme feature sequence comprises:

obtaining an original triphone truth based on the target monophone feature in the output phoneme feature sequence;

inputting the original triphone truth into the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain an original triphone Mixture Gaussian Model-Hidden Markov Model;

clustering the original triphone Mixture Gaussian Model-Hidden Markov Model to obtain a clustered triphone Mixture Gaussian Model-Hidden Markov Model;

iteratively training the clustered triphone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original triphone truth to obtain a target triphone Mixture Gaussian Model-Hidden Markov Model;

aligning each original triphone truth based on the target triphone Mixture Gaussian Model-Hidden Markov Model to obtain a target triphone feature; and

obtaining the updated phoneme feature sequence based on the target triphone feature.

5. The acoustic model training method according to claim 1, wherein training the obtained phoneme feature sequence by using the Deep Neural Net-Hidden Markov Model-sequence training model to obtain the target acoustic model in (iii) comprises:

training an original Deep Neural Net-Hidden Markov Model-sequence training model with the obtained phoneme feature sequence;

obtaining an original state level truth corresponding to each target monophone feature or each target triphone feature, based on the original Deep Neural Net-Hidden Markov Model-Sequence Training Model;

training the original Deep Neural Net-Hidden Markov Model-Sequence Training Model based on the audio feature sequence and the original state level truth to obtain an intermediate Deep Neural Net-Hidden Markov Model-Sequence Training Model; and

performing a combined training on the intermediate Deep Neural Net-Hidden Markov Model-sequence training model to obtain the target acoustic model by using a cross-entropy training criterion, a L2-norm training criterion and a Leaky HMM training criterion.

6. A speech recognition method, comprising:

recognizing an original voice signal by using the target acoustic model obtained by the acoustic model training method according to claim 1, and acquiring a current voiceprint feature signal, the original voice signal comprises a user ID;

obtaining a pre-stored voiceprint feature signal corresponding to the user ID; and

determining whether the current voiceprint feature signal and the pre-stored voiceprint feature signal correspond to a same user, and obtaining a determination result.

7. An acoustic model training apparatus, comprising:

(i) an audio feature sequence obtaining module configured to perform feature extraction from a training voice signal to obtain an audio feature sequence;

(ii) a phoneme feature sequence obtaining module configured to perform training on the audio feature sequence by using a phoneme mixed Gaussian Model-Hidden Markov Model to obtain a phoneme feature sequence, wherein the phoneme feature sequence obtaining module is further configured to:

(a) obtain an original monophone truth corresponding to each audio feature in the audio feature sequence based on an original monophone Mixture Gaussian Model-Hidden Markov Model;

(b) iteratively train the original monophone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original monophone truth to obtain a target monophone Mixture Gaussian Model-Hidden Markov Model; and

(c) aligning each original monophone truth based on the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain a target monophone feature; and

(d) obtaining the phoneme feature sequence based on the target monophone feature; and

(iii) a target acoustic model obtaining module configured to train the obtained phoneme feature sequence by using a Deep Neural Net-Hidden Markov Model-sequence training model to obtain a target acoustic model.

8. A speech recognition apparatus, comprising:

a current voiceprint obtaining module configured to recognize an original voice signal by using the target acoustic model obtained by the acoustic model training method according to claim 1, and acquire a current voiceprint feature signal, the original voice signal comprises a user ID;

a pre-stored voiceprint obtaining module configured to obtain a pre-stored voiceprint feature signal corresponding to the user ID; and

a voiceprint feature comparison module configured to determine whether the current voiceprint feature signal and the pre-stored voiceprint feature signal correspond to a same user, and obtain a determination result.

9. A terminal device comprising a memory, a processor, and a computer program stored in the memory and operable by the processor, wherein the processor performs following steps when executing the computer program:

(i) performing feature extraction on a training speech signal to obtain an audio feature sequence;

(ii) training the audio feature sequence by using a phoneme mixed Gaussian Model-Hidden Markov Model to obtain a phoneme feature sequence, and wherein the processor is configured to further perform:

(a) obtaining an original monophone truth corresponding to each audio feature in the audio feature sequence based on an original monophone Mixture Gaussian Model-Hidden Markov Model;

(b) iteratively training the original monophone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original monophone truth to obtain a target monophone Mixture Gaussian Model-Hidden Markov Model; and

(c) aligning each original monophone truth based on the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain a target monophone feature; and

(d) obtaining the phoneme feature sequence based on the target monophone feature; and

(iii) training the phoneme feature sequence by using a Deep Neural Net-Hidden Markov Model-sequence training model to obtain a target acoustic model.

10. The terminal device according to claim 9, wherein performing feature extraction from a training speech signal in (i) comprises:

performing pre-emphasis, framing, windowing, endpoint detection, fast Fourier transform, and Mel filter bank and discrete cosine transform on the training speech signal.

11. The terminal device according to claim 9, wherein prior to obtaining the original monophone truth in (ii)(a), the processor further performs:

calculating mean and variance of some of the audio feature sequences to obtain the original monophone Mixture Gaussian Model-Hidden Markov Model.

12. The terminal device according to claim 9, wherein training of the obtained phoneme feature sequence by the Deep Neural Net-Hidden Markov Model-sequence in (iii) produces an output phoneme feature sequence, and the processor further performs:

training the output phoneme feature sequence by using a triphone Mixture Gaussian Model-Hidden Markov Model to obtain an updated phoneme feature sequence;

wherein training the phoneme feature sequence by using a triphone Mixture Gaussian Model-Hidden Markov Model to obtain the updated phoneme feature sequence by the processor further comprises:

obtaining an original triphone truth based on the target monophone feature in the output phoneme feature sequence;

inputting the original triphone truth into the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain an original triphone Mixture Gaussian Model-Hidden Markov Model;

clustering the original triphone Mixture Gaussian Model-Hidden Markov Model to obtain a clustered triphone Mixture Gaussian Model-Hidden Markov Model;

iteratively training the clustered triphone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original triphone truth to obtain a target triphone Mixture Gaussian Model-Hidden Markov Model;

aligning each original triphone truth based on the target triphone Mixture Gaussian Model-Hidden Markov Model to obtain a target triphone feature; and

obtaining the updated phoneme feature sequence based on the target triphone feature.

13. The terminal device according to claim 9, wherein training the obtained phoneme feature sequence by using a Deep Neural Net-Hidden Markov Model-sequence training model to obtain the target acoustic model in (iii) by the processor further comprises:

training an original Deep Neural Net-Hidden Markov Model-sequence training model with the obtained phoneme feature sequence;

obtaining an original state level truth corresponding to each target monophone feature or each target triphone feature, based on the original Deep Neural Net-Hidden Markov Model-Sequence Training Model;

training the original Deep Neural Net-Hidden Markov Model-Sequence Training Model based on the audio feature sequence and the original state level truth to obtain an intermediate Deep Neural Net-Hidden Markov Model-Sequence Training Model; and

performing a combined training on the intermediate Deep Neural Net-Hidden Markov Model-sequence training model to obtain the target acoustic model by using a cross-entropy training criterion, a L2-norm training criterion and a Leaky HMM training criterion.

14. A terminal device comprising a memory, a processor, and a computer program stored in the memory and operable by the processor, the processor performs following steps when executing the computer program:

recognizing an original voice signal by using the target acoustic model obtained by the acoustic model training method according to claim 1, and acquiring a current voiceprint feature signal, the original voice signal comprises a user ID;

obtaining a pre-stored voiceprint feature signal corresponding to the user ID; and

determining whether the current voiceprint feature signal and the pre-stored voiceprint feature signal correspond to a same user, and obtaining a determination result.

15. A computer readable storage medium, the computer readable storage medium stores a computer program, wherein following steps are performed when the computer program is executed by the processor:

(i) performing feature extraction from a training speech signal to obtain an audio feature sequence;

(ii) training the audio feature sequence by using a phoneme mixed Gaussian Model-Hidden Markov Model to obtain a phoneme feature sequence, wherein training the audio feature sequence includes performing:

(a) obtaining an original monophone truth corresponding to each audio feature in the audio feature sequence based on an original monophone Mixture Gaussian Model-Hidden Markov Model;

(b) iteratively training the original monophone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original monophone truth to obtain a target monophone Mixture Gaussian Model-Hidden Markov Model; and

(c) aligning each original monophone truth based on the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain a target monophone feature; and

(d) obtaining the phoneme feature sequence based on the target monophone feature; and

(iii) training the obtained phoneme feature sequence by using a Deep Neural Net-Hidden Markov Model-sequence training model to obtain a target acoustic model.

16. The computer readable storage medium according to claim 15, wherein performing feature extraction from a training speech signal in (i) comprises:

performing pre-emphasis, framing, windowing, endpoint detection, fast Fourier transform, and Mel filter bank and discrete cosine transform on the training speech signal.

17. The computer readable storage medium according to claim 15, wherein prior to obtaining the original monophone truth in (ii)(a), the steps performed when the computer program is executed by the processor further comprises:

calculating mean and variance of some of the audio feature sequences to obtain the original monophone Mixture Gaussian Model-Hidden Markov Model.

18. The computer readable storage medium according to claim 15, wherein training of the obtained phoneme feature sequence by the Deep Neural Net-Hidden Markov Model-sequence in (iii) produces an output phoneme feature sequence, and the steps performed when the computer program is executed by the processor further comprises:

training the output phoneme feature sequence by using a triphone Mixture Gaussian Model-Hidden Markov Model to obtain an updated phoneme feature sequence;

wherein training the phoneme feature sequence by using a triphone Mixture Gaussian Model-Hidden Markov Model to obtain the updated phoneme feature sequence comprises:

obtaining an original triphone truth based on the target monophone feature in the output phoneme feature sequence;

inputting the original triphone truth into the target monophone Mixture Gaussian Model-Hidden Markov Model to obtain an original triphone Mixture Gaussian Model-Hidden Markov Model;

clustering the original triphone Mixture Gaussian Model-Hidden Markov Model to obtain a clustered triphone Mixture Gaussian Model-Hidden Markov Model;

iteratively training the clustered triphone Mixture Gaussian Model-Hidden Markov Model based on the audio feature sequence and the original triphone truth to obtain a target triphone Mixture Gaussian Model-Hidden Markov Model;

aligning each original triphone truth based on the target triphone Mixture Gaussian Model-Hidden Markov Model to obtain a target triphone feature; and

obtaining the updated phoneme feature sequence based on the target triphone feature.

19. The computer readable storage medium according to claim 15, wherein training the obtained phoneme feature sequence by using a Deep Neural Net-Hidden Markov Model-sequence training model to obtain a target acoustic model in (iii) comprises:

training an original Deep Neural Net-Hidden Markov Model-sequence training model with the obtained phoneme feature sequence;

obtaining an original state level truth corresponding to each target monophone feature or each target triphone feature, based on the original Deep Neural Net-Hidden Markov Model-Sequence Training Model;

training the original Deep Neural Net-Hidden Markov Model-Sequence Training Model based on the audio feature sequence and the original state level truth to obtain an intermediate Deep Neural Net-Hidden Markov Model-Sequence Training Model; and

performing a combined training on the intermediate Deep Neural Net-Hidden Markov Model-sequence training model to obtain the target acoustic model by using

a cross-entropy training criterion, a L2-norm training criterion and a Leaky HMM training criterion.

20. A computer readable storage medium, the computer readable storage medium stores a computer program, and wherein following steps are performed when the computer program is executed by the processor:

recognizing an original voice signal by using the target acoustic model obtained by the acoustic model training method according to claim 1, and obtaining a current voiceprint feature signal, the original voice signal comprises a user ID;

obtaining a pre-stored voiceprint feature signal corresponding to the user ID; and

determining whether the current voiceprint feature signal and the pre-stored voiceprint feature signal correspond to a same user, and obtaining a determination result.