CLAIMS

1. A speech wakeup method, comprising:

training a speech wakeup model with general speech data, wherein the training comprises iteratively optimizing parameters in the speech wakeup model with the general speech data by means of an asynchronous stochastic gradient descent method until the training converges;

acquiring keyword-specific speech data, and training the speech wakeup model with the keyword-specific speech data, wherein a learning rate used in the training with the keyword-specific speech data is less than that used in the training of the speech wakeup model with the general speech data; and

inputting speech data to the speech wakeup model trained with the general speech data and the keyword-specific speech data, and outputting, by the speech wakeup model, a result for determining whether to execute speech wakeup, wherein the speech wakeup model comprises a Deep Neural Network (DNN) and a Connectionist Temporal Classifier (CTC).

2. The method of claim 1, wherein the general speech data comprises a Large Vocabulary Continuous Speech Recognition (LVCSR) corpus.

3. The method of claim 1, further comprising:

cross-verifying the speech wakeup model with a verification data set in the training to determine whether the training converges.

4. The method of claim 1, wherein the outputting, by the speech wakeup model, a result for determining whether to execute speech wakeup specifically comprises:

extracting acoustic features from the input speech data;

inputting the acoustic features to the DNN comprised in the speech wakeup model for processing to obtain a class probability of the acoustic features respectively corresponding to each pronunciation phoneme;

inputting the class probability to the CTC comprised in the speech wakeup model for processing to obtain a confidence score of a speech wakeup term corresponding to a pronunciation phoneme sequence; and

determining whether to execute wakeup according to the confidence score, and outputting a determination result.

5. The method of claim 4, wherein the extracting acoustic features from the input speech data specifically comprises:

extracting acoustic feature frames of the input speech data from a window according to a specified time interval, wherein each of the acoustic feature frames is multi-dimension log filter bank energies;

stacking a plurality of adjacent acoustic feature frames respectively; and

taking the stacked acoustic feature frames respectively as acoustic features extracted from the monitored speech.

6. A speech wakeup apparatus, comprising a training module, an input module and a speech wakeup model;

wherein the training module trains a speech wakeup model with general speech data, wherein the training comprises iteratively optimizing parameters in the speech wakeup model with the general speech data by means of an asynchronous stochastic gradient descent method until the training converges;

the training module acquires keyword-specific speech data, and trains the speech wakeup model with the keyword-specific speech data, wherein a learning rate used in the training with the keyword-specific speech data is less than that used in the training of the speech wakeup model with the general speech data; and

speech data is input by the input module to the speech wakeup model trained with general speech data and the keyword-specific speech data, and the speech wakeup model outputs a result for determining whether to execute speech wakeup, wherein the speech wakeup model comprises a Deep Neural Network (DNN) and a Connectionist Temporal Classifier (CTC).

7. The apparatus of claim 6, wherein the general speech data comprises a large vocabulary continuous speech recognition (LVCSR) corpus.

8. The apparatus of claim 6, wherein the training module cross-verifies the speech wakeup model with a verification data set in the training to determine whether the training converges.

9. The apparatus of claim 6, wherein the outputting, by the speech wakeup model, a result for determining whether to execute speech wakeup specifically comprises:

extracting, by the speech wakeup model, acoustic features from the input speech data;

inputting the acoustic features to the DNN comprised in the speech wakeup model for processing to obtain a class probability of the acoustic features respectively corresponding to each pronunciation phoneme;

inputting the class probability to the CTC comprised in the speech wakeup model for processing to obtain a confidence score of a speech wakeup term corresponding to a pronunciation phoneme sequence; and

determining whether to execute wakeup according to the confidence score, and outputting a determination result.

10. The apparatus of claim 9, wherein the extracting, by the speech wakeup model, acoustic features from the input speech data specifically comprises:

extracting, by the speech wakeup model, acoustic feature frames of the input speech data from a window according to a specified time interval, wherein each of the acoustic feature frames is multi-dimension log filter bank energies;

stacking a plurality of adjacent acoustic feature frames respectively; and

taking the stacked acoustic feature frames respectively as acoustic features extracted from the monitored speech.

11. An electronic device, comprising:

at least one processor; and

a memory communicatively connected to the at least one processor; wherein

the memory stores an instruction executable by the at least one processor, and the instruction is executed by the at least one processor to enable the at least one processor to:

train a speech wakeup model with general speech data, wherein the training comprises iteratively optimizing parameters in the speech wakeup model with the general speech data by means of an asynchronous stochastic gradient descent method until the training converges;

acquire keyword-specific speech data, and train the speech wakeup model with the keyword-specific speech data, wherein a learning rate used in the training with the keyword-specific speech data is less than that used in the training of the speech wakeup model with the general speech data; and

input speech data to the speech wakeup model trained with general speech data and the keyword-specific speech data, and output, by the speech wakeup model, a result for determining whether to execute speech wakeup, wherein the speech wakeup model comprises a Deep Neural Network (DNN) and a Connectionist Temporal Classifier (CTC).