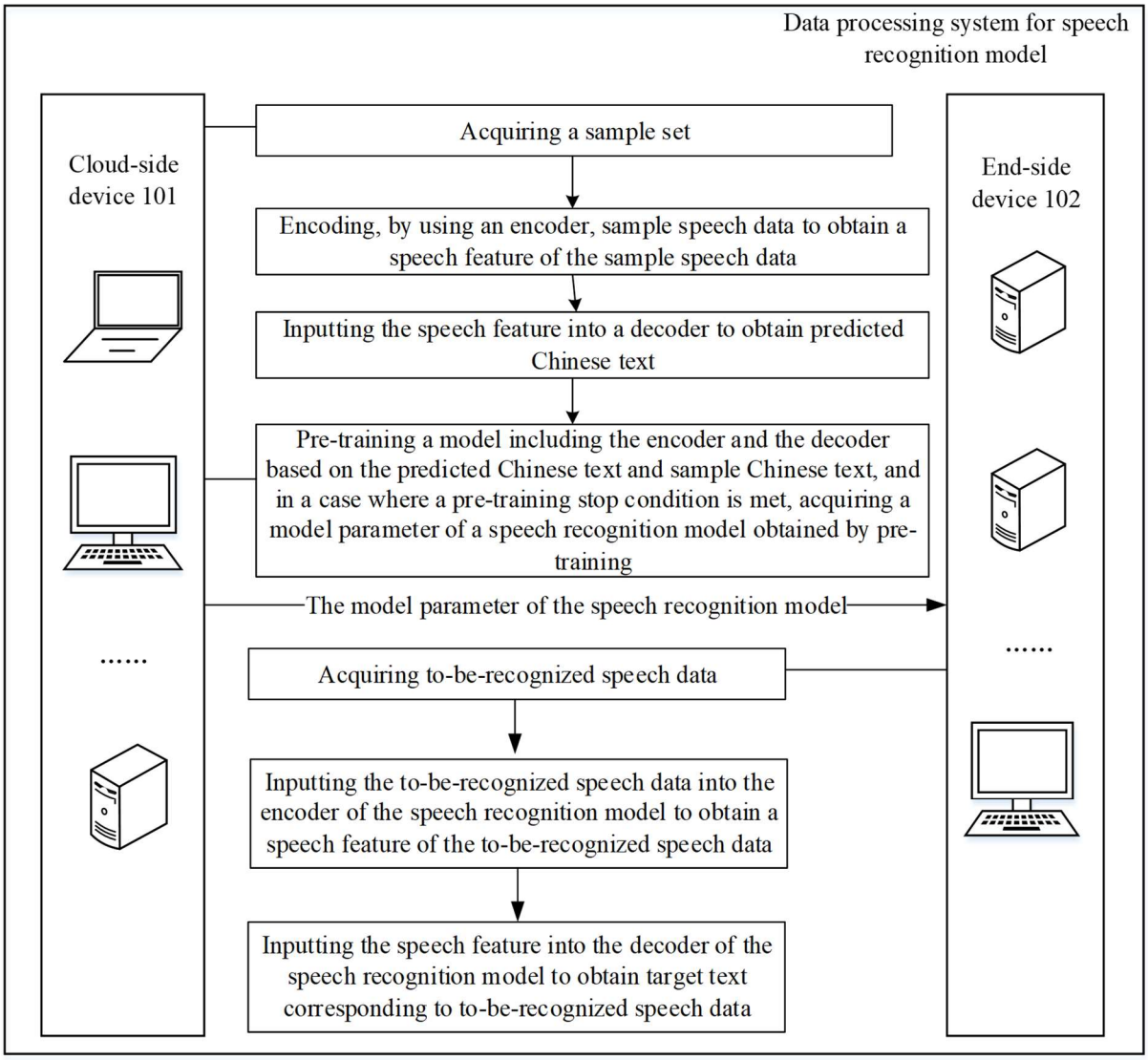

Data processing system for speech recognition model
Cloud-side device 101
End-side device 102
Acquiring a sample set
Encoding, by using an encoder, sample speech data to obtain a speech feature of the sample speech data
Inputting the speech feature into a decoder to obtain predicted Chinese text
Pre-training a model including the encoder and the decoder based on the predicted Chinese text and sample Chinese text, and in a case where a pre-training stop condition is met, acquiring a model parameter of a speech recognition model obtained by pre-training
The model parameter of the speech recognition model
Acquiring to-be-recognized speech data
Inputting the to-be-recognized speech data into the encoder of the speech recognition model to obtain a speech feature of the to-be-recognized speech data
Inputting the speech feature into the decoder of the speech recognition model to obtain target text corresponding to to-be-recognized speech data

FIG. 1

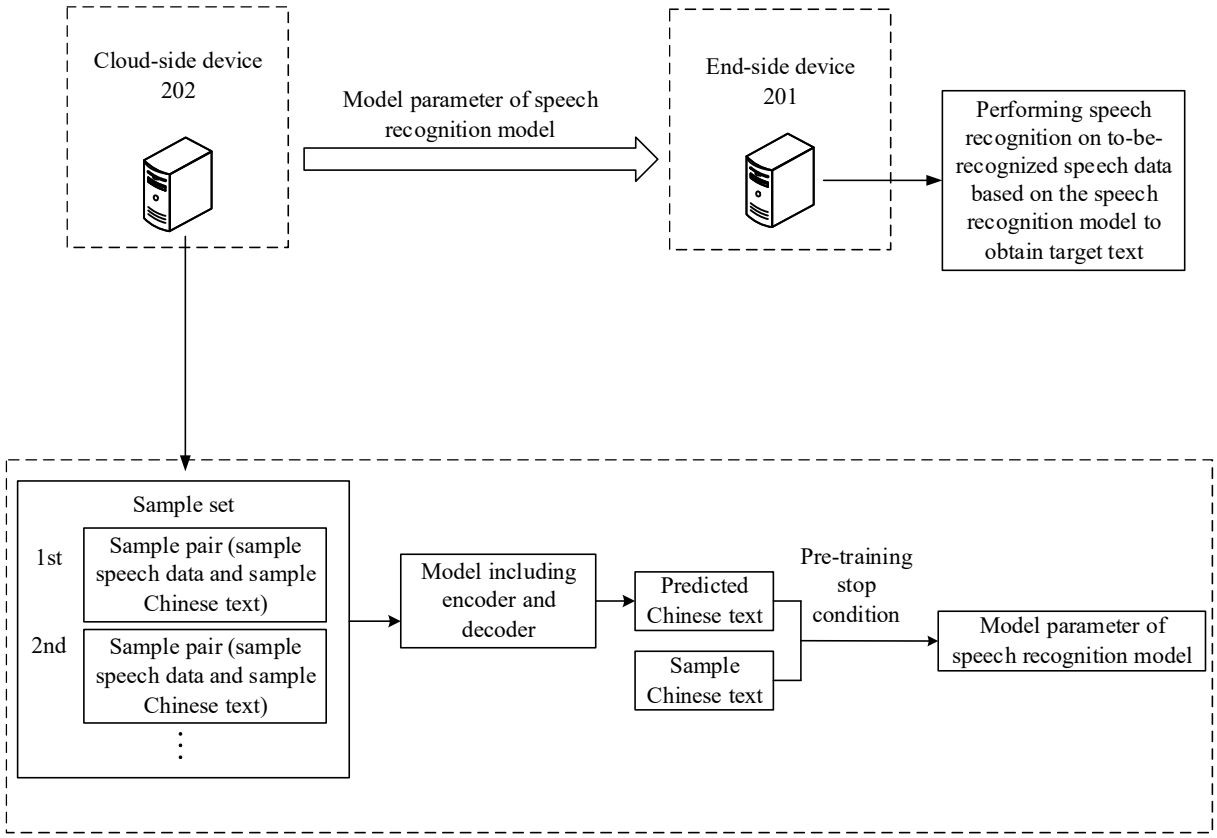

Cloud-side device
202
Model parameter of speech recognition model
End-side device
201
Performing speech recognition on to-be-recognized speech data based on the speech recognition model to obtain target text
Sample set
1st
Sample pair (sample speech data and sample Chinese text)
2nd
Sample pair (sample speech data and sample Chinese text)
Model including encoder and decoder
Predicted Chinese text
Sample Chinese text
Pre-training stop condition
Model parameter of speech recognition model

FIG. 2

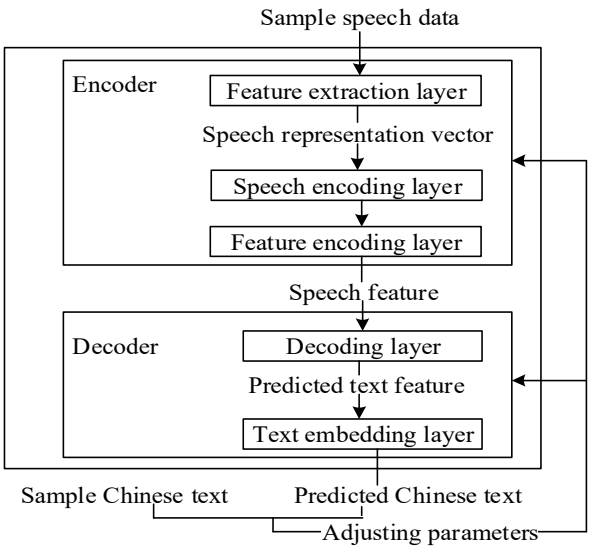

Sample speech data
Encoder
Feature extraction layer
Speech representation vector
Speech encoding layer
Feature encoding layer
Speech feature
Decoder
Decoding layer
Predicted text feature
Text embedding layer
Sample Chinese text
Predicted Chinese text
Adjusting parameters

FIG. 3

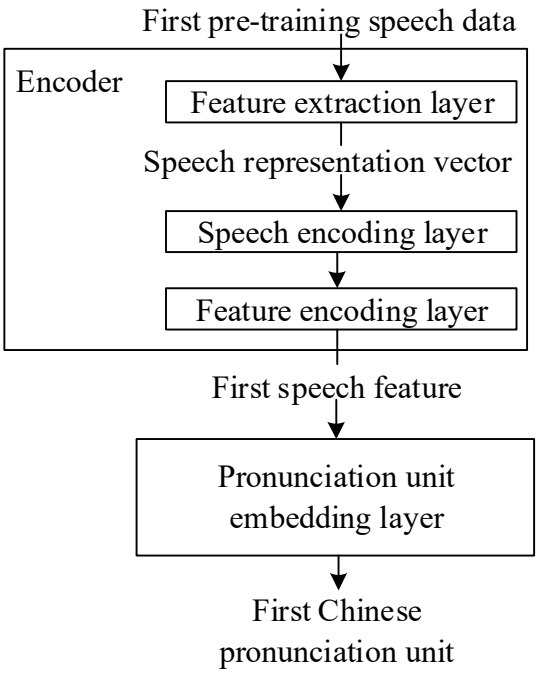

First pre-training speech data
Encoder
Feature extraction layer
Speech representation vector
Speech encoding layer
Feature encoding layer
First speech feature
Pronunciation unit embedding layer
First Chinese pronunciation unit

FIG. 4a

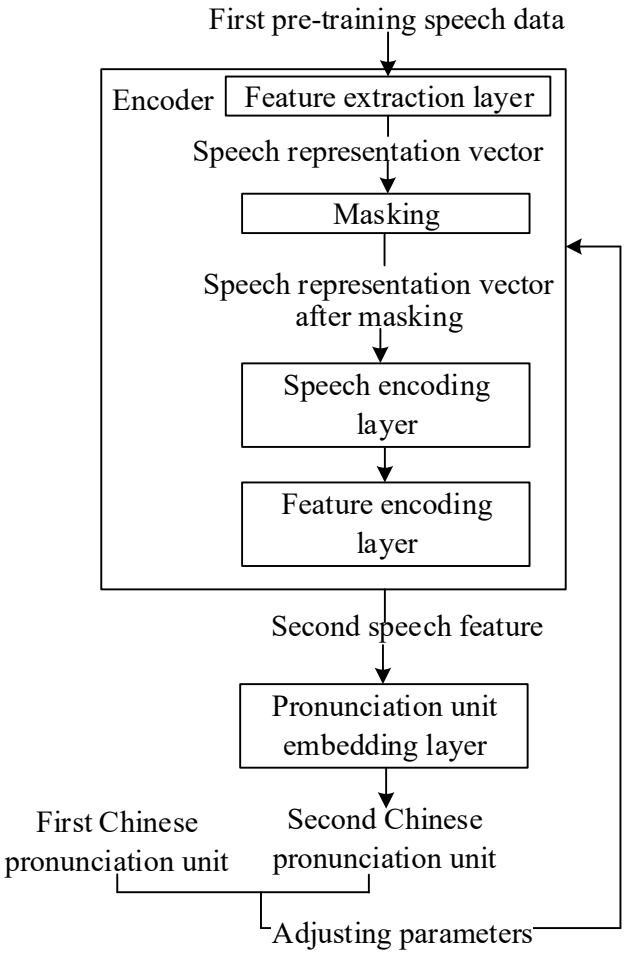

First pre-training speech data
Encoder
Feature extraction layer
Speech representation vector
Masking
Speech representation vector after masking
Speech encoding layer
Feature encoding layer
Second speech feature
Pronunciation unit embedding layer
First Chinese pronunciation unit
Second Chinese pronunciation unit
Adjusting parameters

FIG. 4b

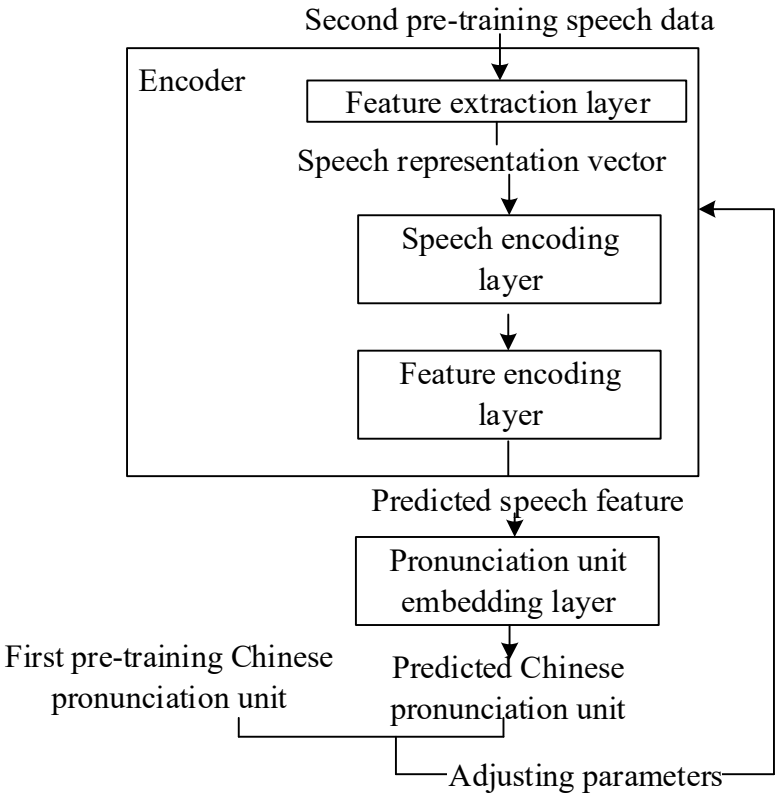

Second pre-training speech data
Encoder
Feature extraction layer
Speech representation vector
Speech encoding layer
Feature encoding layer
Predicted speech feature
Pronunciation unit embedding layer
First pre-training Chinese pronunciation unit
Predicted Chinese pronunciation unit
Adjusting parameters

FIG. 5

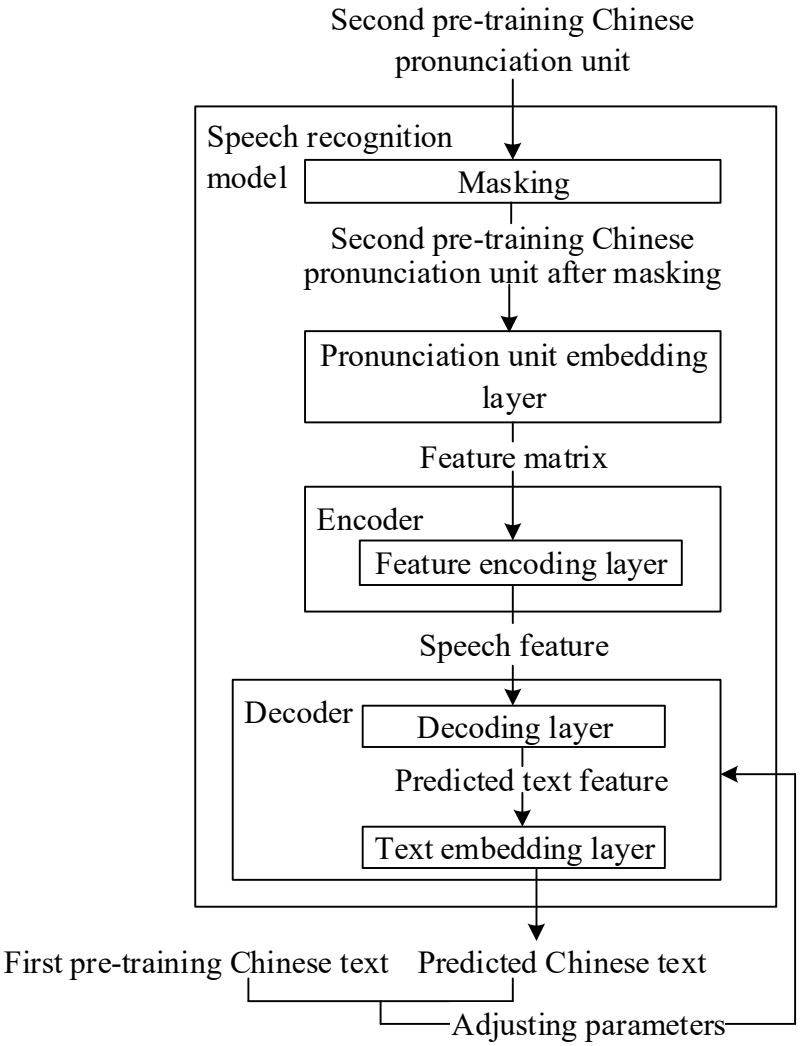

Second pre-training Chinese pronunciation unit
Speech recognition model
Masking
Second pre-training Chinese pronunciation unit after masking
Pronunciation unit embedding layer
Feature matrix
Encoder
Feature encoding layer
Speech feature
Decoder
Decoding layer
Predicted text feature
Text embedding layer
First pre-training Chinese text
Predicted Chinese text
Adjusting parameters

FIG. 6

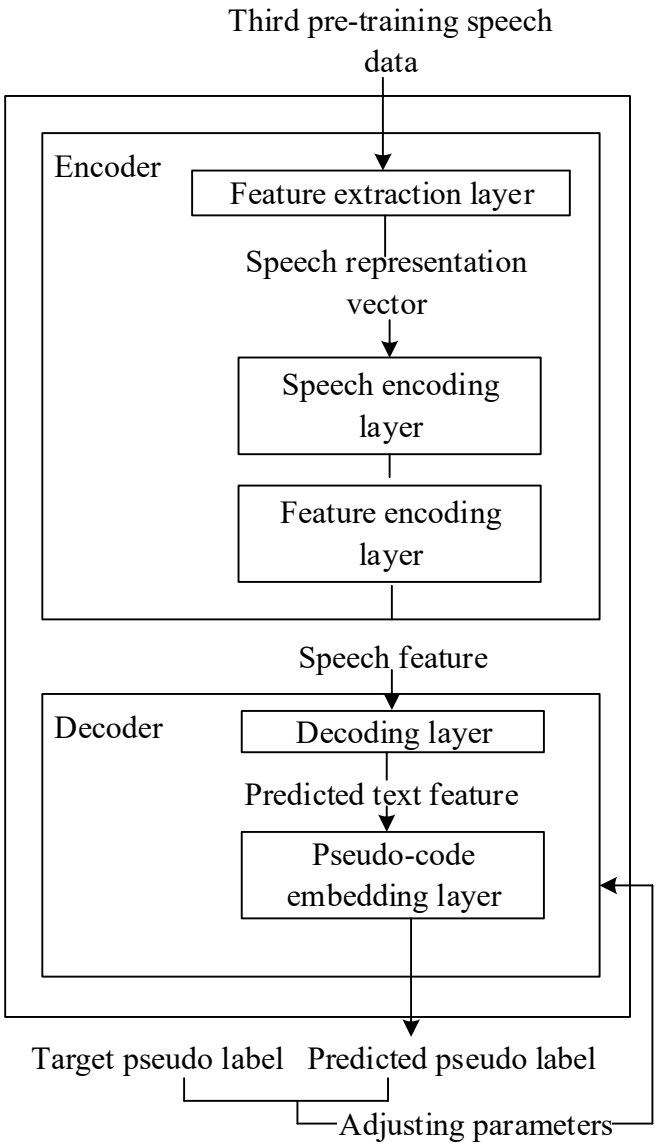

Third pre-training speech data
Encoder
Feature extraction layer
Speech representation vector
Speech encoding layer
Feature encoding layer
Speech feature
Decoder
Decoding layer
Predicted text feature
Pseudo-code embedding layer
Target pseudo label
Predicted pseudo label
Adjusting parameters

FIG. 7

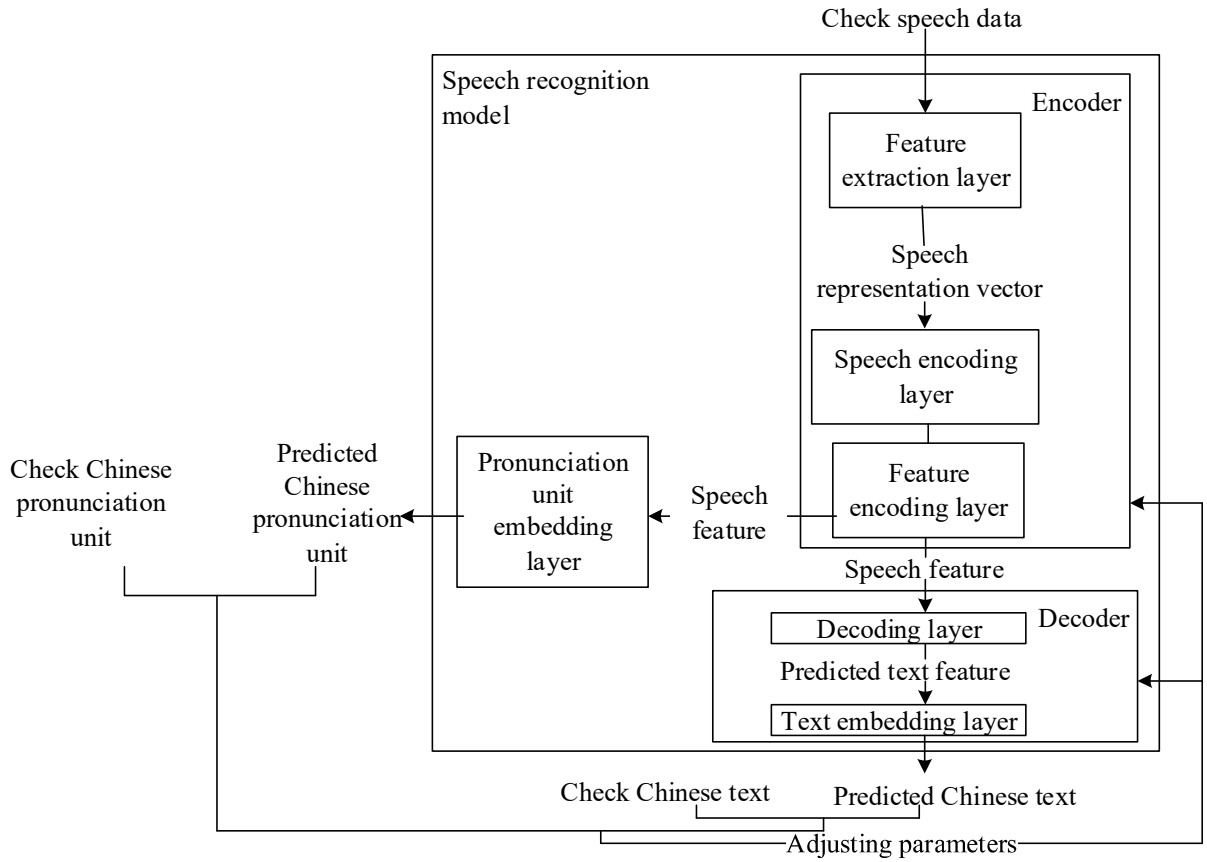

Check speech data
Speech recognition model
Encoder
Feature extraction layer
Speech representation vector
Speech encoding layer
Feature encoding layer
Speech feature
Pronunciation unit embedding layer
Check Chinese pronunciation unit
Predicted Chinese pronunciation unit
Speech feature
Decoder
Decoding layer
Predicted text feature
Text embedding layer
Check Chinese text
Predicted Chinese text
Adjusting parameters

FIG. 8

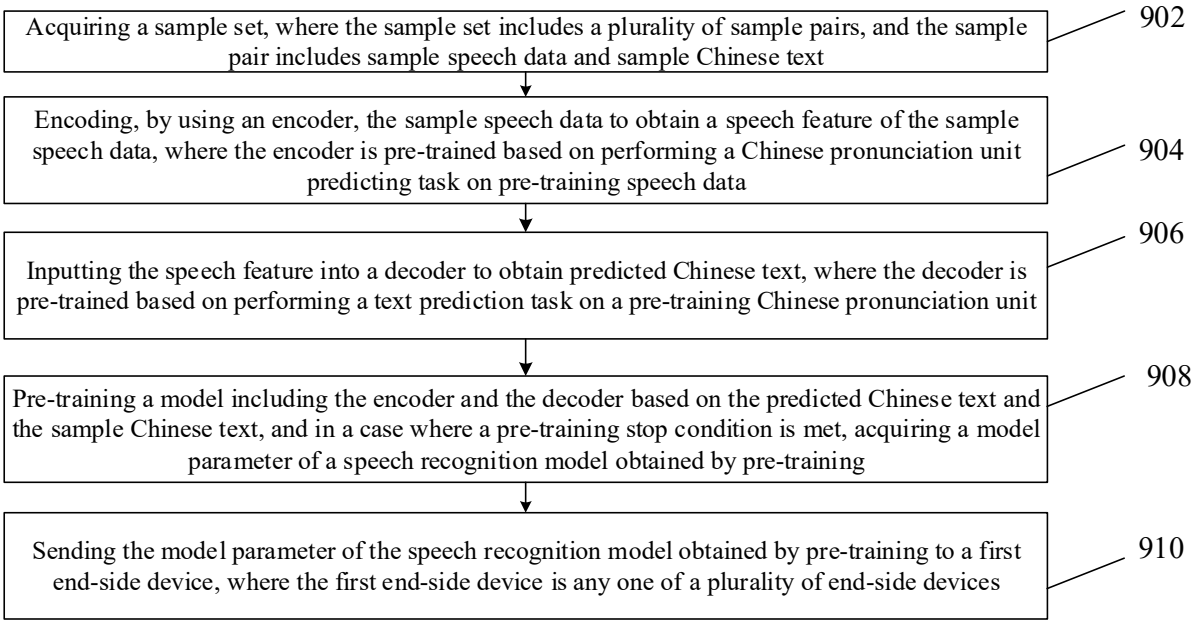

Acquiring a sample set, where the sample set includes a plurality of sample pairs, and the sample pair includes sample speech data and sample Chinese text
902
Encoding, by using an encoder, the sample speech data to obtain a speech feature of the sample speech data, where the encoder is pre-trained based on performing a Chinese pronunciation unit predicting task on pre-training speech data
904
Inputting the speech feature into a decoder to obtain predicted Chinese text, where the decoder is pre-trained based on performing a text prediction task on a pre-training Chinese pronunciation unit
906
Pre-training a model including the encoder and the decoder based on the predicted Chinese text and the sample Chinese text, and in a case where a pre-training stop condition is met, acquiring a model parameter of a speech recognition model obtained by pre-training
908
Sending the model parameter of the speech recognition model obtained by pre-training to a first end-side device, where the first end-side device is any one of a plurality of end-side devices
910

FIG. 9

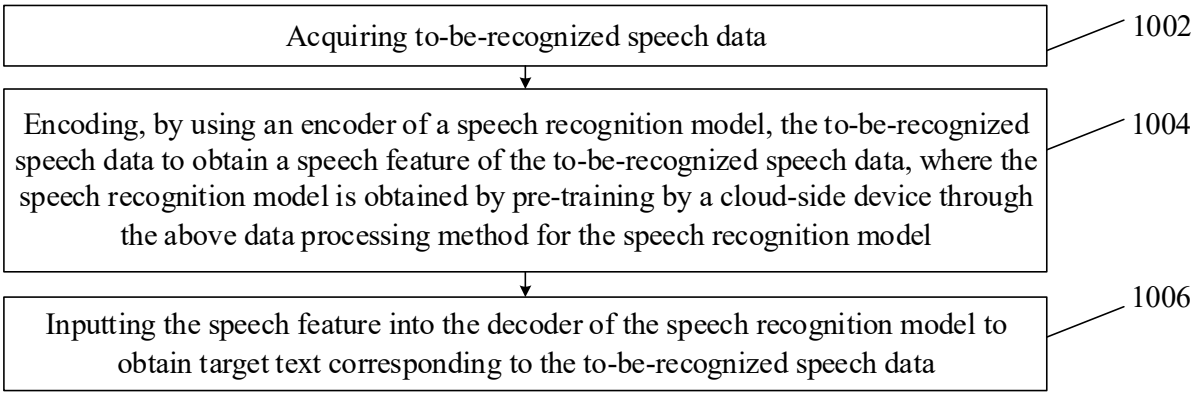

Acquiring to-be-recognized speech data
1002
Encoding, by using an encoder of a speech recognition model, the to-be-recognized speech data to obtain a speech feature of the to-be-recognized speech data, where the speech recognition model is obtained by pre-training by a cloud-side device through the above data processing method for the speech recognition model
1004
Inputting the speech feature into the decoder of the speech recognition model to obtain target text corresponding to the to-be-recognized speech data
1006

FIG. 10

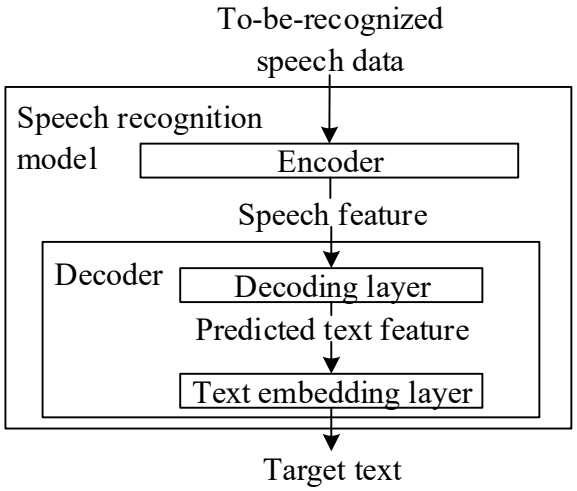

To-be-recognized speech data
Speech recognition model
Encoder
Speech feature
Decoder
Decoding layer
Predicted text feature
Text embedding layer
Target text

FIG. 11

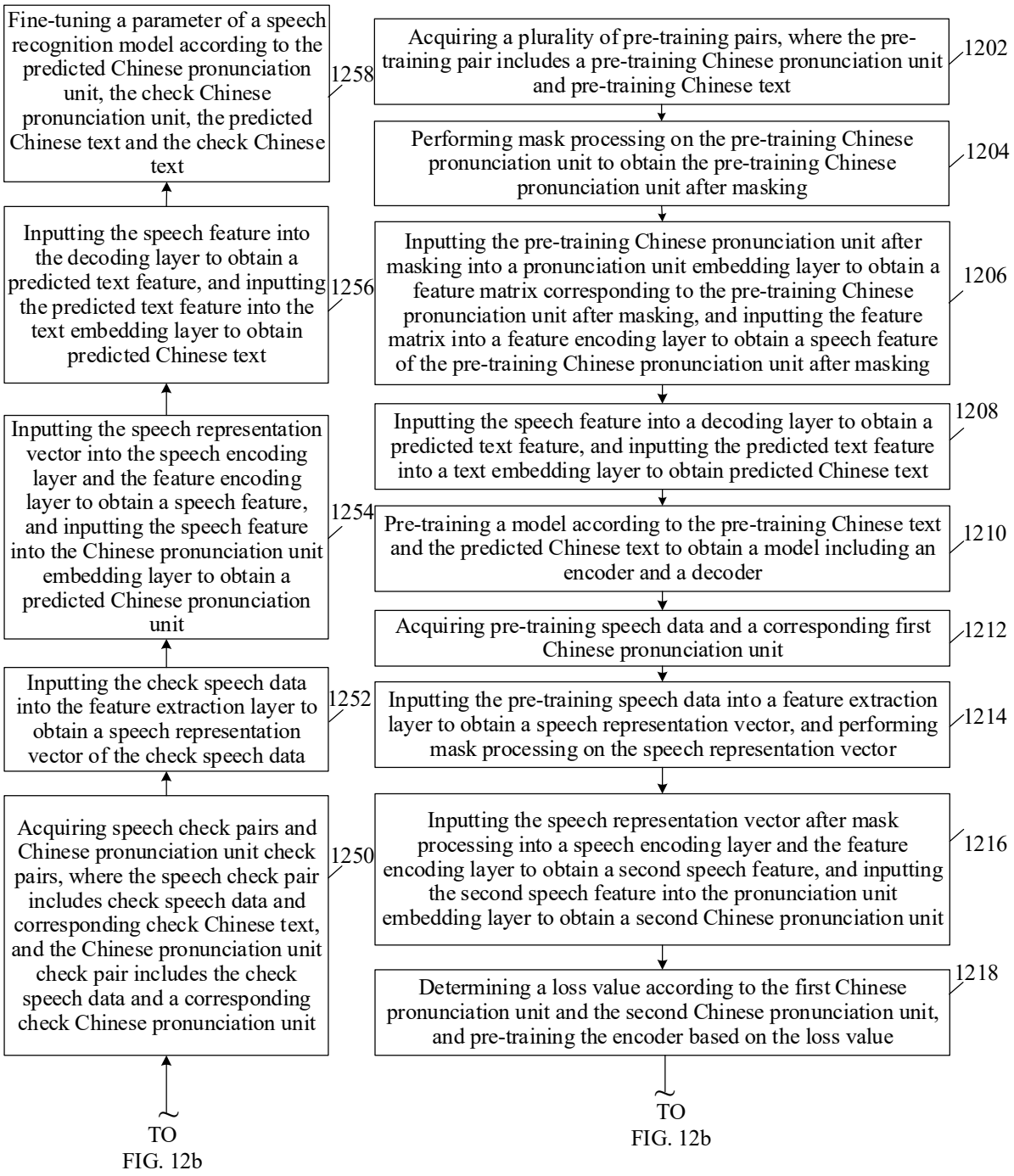

Acquiring a plurality of pre-training pairs, where the pre-training pair includes a pre-training Chinese pronunciation unit and pre-training Chinese text
1202
Performing mask processing on the pre-training Chinese pronunciation unit to obtain the pre-training Chinese pronunciation unit after masking
1204
Inputting the pre-training Chinese pronunciation unit after masking into a pronunciation unit embedding layer to obtain a feature matrix corresponding to the pre-training Chinese pronunciation unit after masking, and inputting the feature matrix into a feature encoding layer to obtain a speech feature of the pre-training Chinese pronunciation unit after masking
1206
Inputting the speech feature into a decoding layer to obtain a predicted text feature, and inputting the predicted text feature into a text embedding layer to obtain predicted Chinese text
1208
Pre-training a model according to the pre-training Chinese text and the predicted Chinese text to obtain a model including an encoder and a decoder
1210
Acquiring pre-training speech data and a corresponding first Chinese pronunciation unit
1212
Inputting the pre-training speech data into a feature extraction layer to obtain a speech representation vector, and performing mask processing on the speech representation vector
1214
Inputting the speech representation vector after mask processing into a speech encoding layer and the feature encoding layer to obtain a second speech feature, and inputting the second speech feature into the pronunciation unit embedding layer to obtain a second Chinese pronunciation unit
1216
Determining a loss value according to the first Chinese pronunciation unit and the second Chinese pronunciation unit, and pre-training the encoder based on the loss value
1218
TO FIG. 12b
Acquiring speech check pairs and Chinese pronunciation unit check pairs, where the speech check pair includes check speech data and corresponding check Chinese text, and the Chinese pronunciation unit check pair includes the check speech data and a corresponding check Chinese pronunciation unit
1250
Inputting the check speech data into the feature extraction layer to obtain a speech representation vector of the check speech data
1252
Inputting the speech representation vector into the speech encoding layer and the feature encoding layer to obtain a speech feature, and inputting the speech feature into the Chinese pronunciation unit embedding layer to obtain a predicted Chinese pronunciation unit
1254
Inputting the speech feature into the decoding layer to obtain a predicted text feature, and inputting the predicted text feature into the text embedding layer to obtain predicted Chinese text
1256
Fine-tuning a parameter of a speech recognition model according to the predicted Chinese pronunciation unit, the check Chinese pronunciation unit, the predicted Chinese text and the check Chinese text
1258
TO FIG. 12b

FIG. 12a

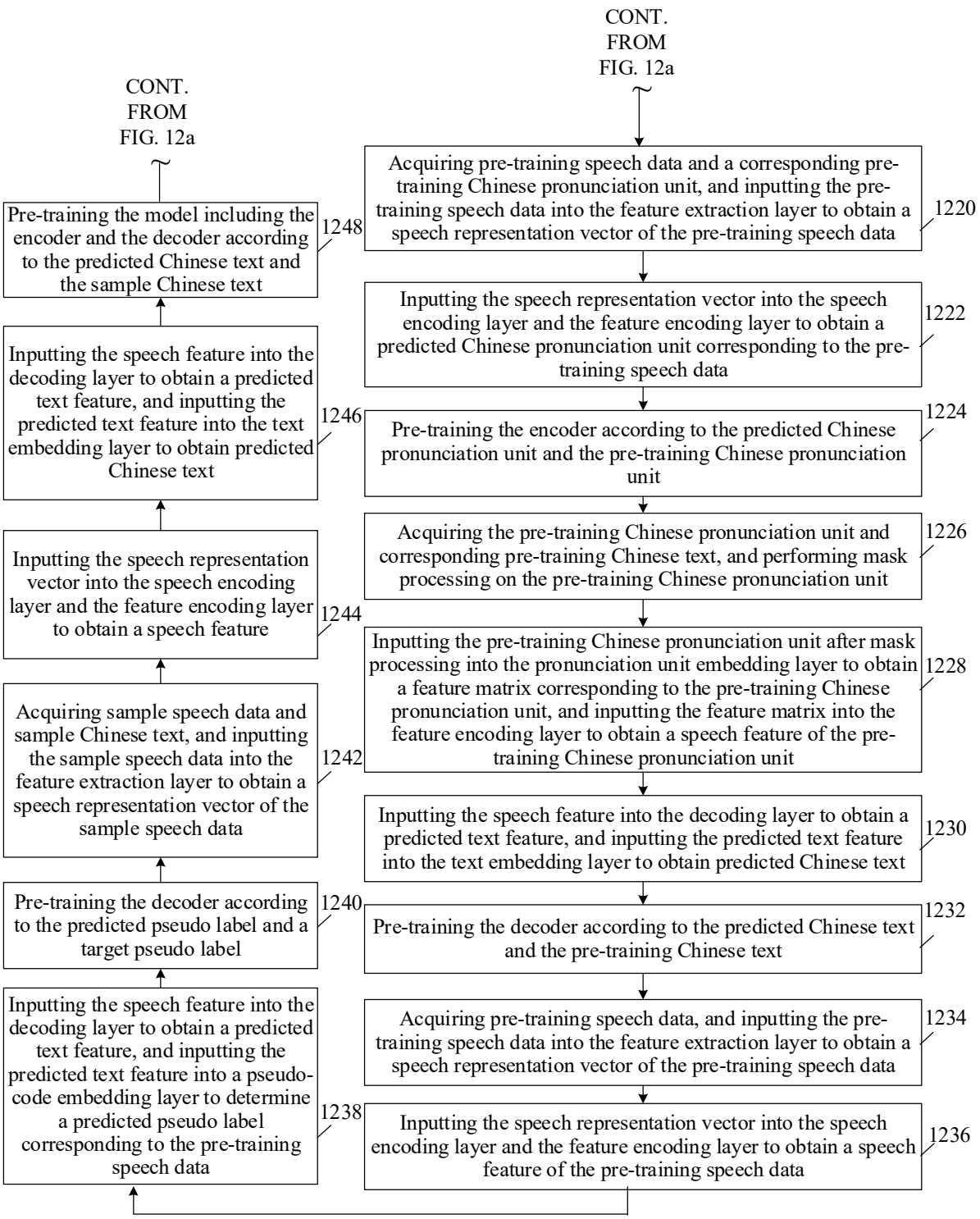

CONT. FROM FIG. 12a
Acquiring pre-training speech data and a corresponding pre-training Chinese pronunciation unit, and inputting the pre-training speech data into the feature extraction layer to obtain a speech representation vector of the pre-training speech data
1220
Inputting the speech representation vector into the speech encoding layer and the feature encoding layer to obtain a predicted Chinese pronunciation unit corresponding to the pre-training speech data
1222
Pre-training the encoder according to the predicted Chinese pronunciation unit and the pre-training Chinese pronunciation unit
1224
Acquiring the pre-training Chinese pronunciation unit and corresponding pre-training Chinese text, and performing mask processing on the pre-training Chinese pronunciation unit
1226
Inputting the pre-training Chinese pronunciation unit after mask processing into the pronunciation unit embedding layer to obtain a feature matrix corresponding to the pre-training Chinese pronunciation unit, and inputting the feature matrix into the feature encoding layer to obtain a speech feature of the pre-training Chinese pronunciation unit
1228
Inputting the speech feature into the decoding layer to obtain a predicted text feature, and inputting the predicted text feature into the text embedding layer to obtain predicted Chinese text
1230
Pre-training the decoder according to the predicted Chinese text and the pre-training Chinese text
1232
Acquiring pre-training speech data, and inputting the pre-training speech data into the feature extraction layer to obtain a speech representation vector of the pre-training speech data
1234
Inputting the speech representation vector into the speech encoding layer and the feature encoding layer to obtain a speech feature of the pre-training speech data
1236
Inputting the speech feature into the decoding layer to obtain a predicted text feature, and inputting the predicted text feature into a pseudo-code embedding layer to determine a predicted pseudo label corresponding to the pre-training speech data
1238
Pre-training the decoder according to the predicted pseudo label and a target pseudo label
1240
Acquiring sample speech data and sample Chinese text, and inputting the sample speech data into the feature extraction layer to obtain a speech representation vector of the sample speech data
1242
Inputting the speech representation vector into the speech encoding layer and the feature encoding layer to obtain a speech feature
1244
Inputting the speech feature into the decoding layer to obtain a predicted text feature, and inputting the predicted text feature into the text embedding layer to obtain predicted Chinese text
1246
Pre-training the model including the encoder and the decoder according to the predicted Chinese text and the sample Chinese text
1248
CONT. FROM FIG. 12a

FIG. 12b

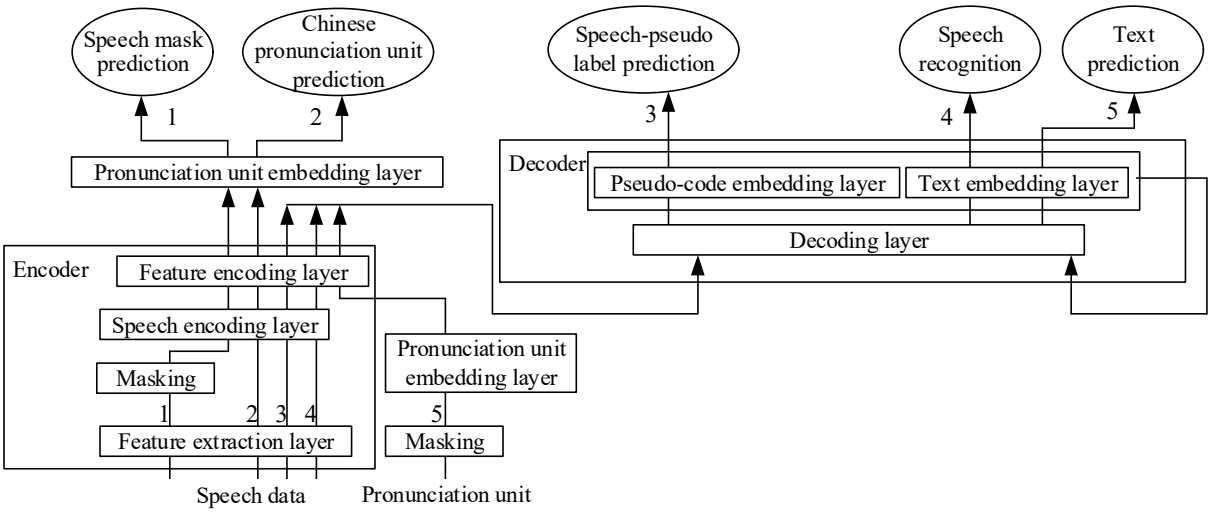


FIG. 13

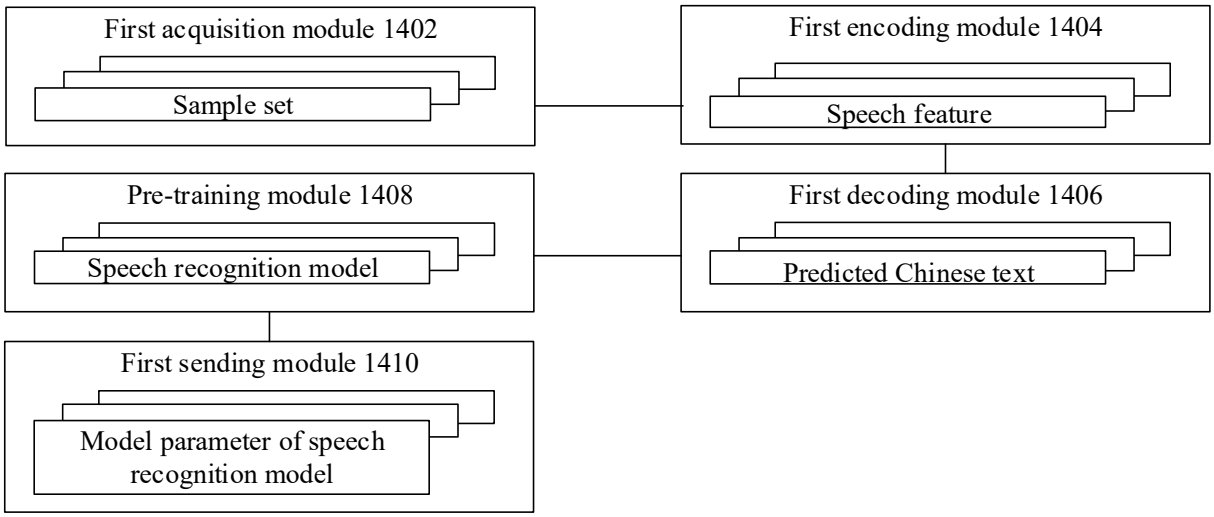


FIG. 14

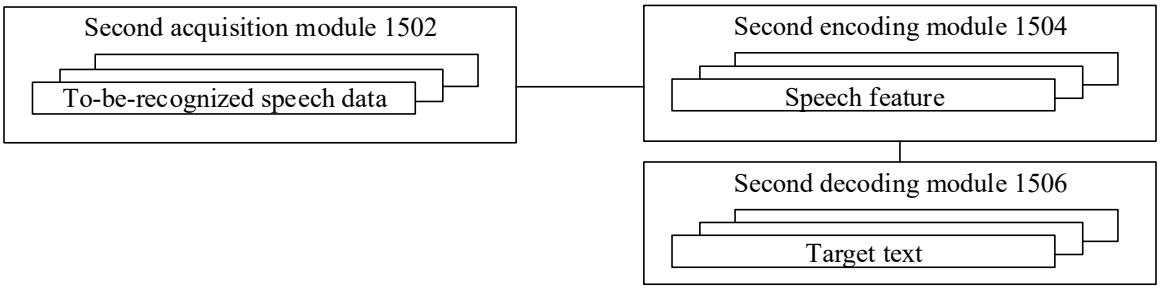


FIG. 15

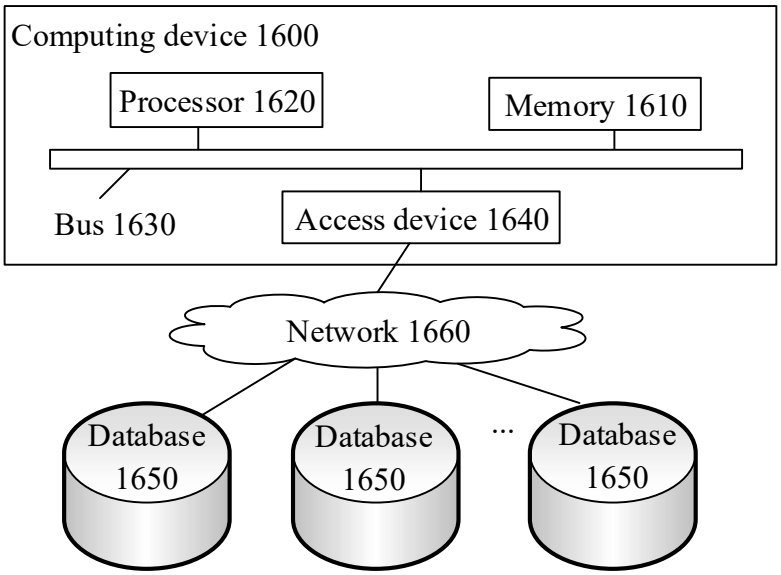

Computing device 1600
Processor 1620
Memory 1610
Bus 1630
Access device 1640
Network 1660
Database 1650
Database 1650
...
Database 1650

FIG. 16