

DATA PROCESSING METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority to the Chinese Patent Application No. 201910745443.6, filed with the National Intellectual Property Administration, PRC on August 13, 2019 and entitled "DATA PROCESSING METHOD AND APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

This application relates to a data processing method and apparatus, a computer device, and a storage medium.

BACKGROUND

With the development of artificial intelligence, an intelligent interview system appears. A conventional intelligent interview system is mainly used to recognize facial micro-expressions to detect abnormal expressions of interviewees, and use the abnormal expressions as a basis for risk assessment. Micro-expression is a psychological noun. People express their feelings to others by making expressions. In different expressions, or in one expression, the face "leaks" other information. A "micro-expression" can last at least 1/25 second. Although a subconscious expression may last only a moment, it expresses an opposite emotion in some cases.

However, the inventor is aware that recognizing only the micro-expression features is insufficient to accurately and comprehensively capture psychological states of the interviewees, and is likely to cause a major difference between an interview result and an actual situation, thus lowering recognition accuracy.

SUMMARY

According to embodiments disclosed in this application, a data processing method and apparatus, a computer device, and a storage medium are provided.

A data processing method includes:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

A data processing apparatus includes:

an acquisition module, configured to obtain audio data and video data of an interviewee;

a first extraction module, configured to extract a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtain first speech emotion data based on the micro-speech feature;

a first processing module, configured to: convert the audio data of the interviewee into text data, split the text data into multiple sentences, perform word segmentation on the multiple sentences, search for and match a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determine a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain

second speech emotion data, where the emotion classification network is obtained through training based on first sample text data;

a second processing module, configured to: input the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculate an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data;

a second extraction module, configured to randomly capture a video frame from the video data of the interviewee, extract a micro-expression feature of the interviewee based on the video frame, and obtain a confidence level of the video data based on the micro-expression feature; and

an analysis module, configured to determine an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

A computer device includes a memory and one or more processors, where the memory stores computer readable instructions, and when the computer readable instructions are executed by the one or more processors, the one or more processors are enabled to perform the following steps:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

One or more non-volatile computer readable storage media storing computer readable instructions are provided, where when the computer readable instructions are executed by one or more processors, the one or more processors are enabled to perform the following steps:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

Details of one or more embodiments of this application are set forth in the following accompanying drawings and descriptions. Other features and advantages of this application will

become apparent from the specification, accompanying drawings, and claims.

BRIEF DESCRIPTION OF DRAWINGS

To describe the technical solutions in the embodiments of this application more clearly, the following briefly describes the accompanying drawings required for describing the embodiments. Clearly, the accompanying drawings in the following description show merely some embodiments of this application, and a person of ordinary skill in the art can still derive other accompanying drawings from these accompanying drawings without creative efforts.

FIG. 1 is a diagram of an application scenario of a data processing method according to one or more embodiments;

FIG. 2 is a schematic flowchart of a data processing method according to one or more embodiments;

FIG. 3 is a schematic sub-flowchart of step 204 in FIG. 2 according to one or more embodiments;

FIG. 4 is a schematic sub-flowchart of step 204 in FIG. 2 according to one or more embodiments;

FIG. 5 is a schematic sub-flowchart of step 204 in FIG. 2 according to one or more embodiments;

FIG. 6 is a schematic sub-flowchart of step 206 in FIG. 2 according to one or more embodiments;

FIG. 7 is a schematic sub-flowchart of step 212 in FIG. 2 according to one or more embodiments;

FIG. 8 is a block diagram of a data processing apparatus according to one or more embodiments; and

FIG. 9 is a block diagram of a computer device according to one or more embodiments.

DESCRIPTION OF EMBODIMENTS

To make the technical solutions and advantages of this application clearer, the following further describes this application in detail with reference to the accompanying drawings and embodiments. It should be understood that the specific embodiments described herein are merely

intended to explain this application and are not intended to limit this application.

A data processing method provided in this application may be applied to an application environment shown in FIG. 1. A terminal 102 communicates with a server 104 through a network. The server 104 obtains audio data and video data of an interviewee, extracts a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtains first speech emotion data based on the micro-speech feature; converts the audio data of the interviewee into text data, splits the text data into multiple sentences, and performs word segmentation on the multiple sentences; searches for and matches a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determines a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data; inputs the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculates an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data; randomly captures a video frame from the video data of the interviewee, extracts a micro-expression feature of the interviewee based on the video frame, and obtains a confidence level of the video data based on the micro-expression feature; and determines an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data, and pushes the interview result to the terminal 102. The terminal 102 may be but is not limited to various personal computers, laptop computers, smartphones, tablet computers, and portable wearable devices. The server 104 may be implemented by a standalone server or a server cluster including a plurality of servers.

In an embodiment, as shown in FIG. 2, a data processing method is provided. In an example for description, the method is applied to the server in FIG. 1, and includes the following steps.

Step S202: Obtain audio data and video data of an interviewee.

The video data of the interviewee is video data recorded when the interviewee is being interviewed. The audio data of the interviewee is audio data recorded when the interviewee is being

interviewed. The audio data of the interviewee can be extracted from the video data of the interviewee.

Step S204: Extract a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtain first speech emotion data based on the micro-speech feature.

The server can invoke a speech feature extraction tool to extract the micro-speech feature of the interviewee from the audio data of the interviewee. The micro-speech feature includes a speech rate feature, a pitch feature, and a Mel-frequency cepstral coefficient. The speech rate refers to the number of words per second in speech data. The words may be in Chinese or in English. The pitch refers to a frequency of speech. A Mel-frequency cepstrum is linear transformation of a logarithmic energy spectrum based on nonlinear mel scale of a sound frequency. The Mel-frequency cepstral coefficient is a coefficient forming the Mel-frequency cepstrum. The server inputs the micro-speech feature into a speech emotion classification model that matches gender information of the interviewee in a trained speech emotion classification model set, to obtain the first speech emotion data corresponding to the micro-speech feature. The first speech emotion data refers to a confidence level of the micro-speech feature belonging to each preset emotion category.

The trained speech emotion classification model set includes speech emotion classification models that are obtained through training based on sample data for interviewees of different genders, that is, an emotion classification model for analyzing male speech data and an emotion classification model for analyzing female speech data. The server obtains the gender information of the interviewee, matches the trained speech emotion classification model set based on the gender information of the interviewee, and obtains the speech emotion classification model that matches the gender information of the interviewee from the trained speech emotion classification model set. The speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information includes emotion category information and gender information. The server divides the sample speech data based on the gender information, and performs model training based on the sample speech data obtained after division, to obtain the speech emotion classification model set.

Step S206: Convert the audio data of the interviewee into text data, split the text data into

multiple sentences, perform word segmentation on the multiple sentences, search for and match a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determine a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data.

The emotion classification network may be a network that is based on Bidirectional Encoder Representation from Transformer (BERT) and includes an overlaid classification layer of N neurons (assuming that N emotions are preset). The server splits the text data into multiple sentences, performs word segmentation on each sentence, searches for and matches a dictionary corresponding to BERT based on each word in each sentence, converts each word into a sequence number corresponding to the word in the dictionary corresponding to BERT, inputs a sequence number of an entire sentence into BERT to obtain a confidence level of each sentence belonging to each preset emotion category, and then determines a confidence level of the text data belonging to each preset emotion category based on the confidence level of each sentence belonging to each preset emotion category, to obtain the second speech emotion data. The emotion classification network may be obtained through training based on first sample text data. Each sample sentence in the first sample text data carries label information, and the label information is emotion category information of each sample sentence.

Because the text data requires less cache space than the audio data and the video data, during data processing, converting the audio data of the interviewee into text data and then processing the text data can reduce cache space of the server during processing, thereby optimizing the cache space of the server.

Step S208: Input the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculate an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data.

The syntax analysis network can be trained by using Corpus of Linguistic Acceptability (CoLA) as the second sample text data. The dataset includes multiple single sentences carrying a

label, and the label indicates whether the syntax is correct (0: incorrect; 1: correct). The trained syntax analysis network can be used to determine syntax accuracy of a sentence. The syntax score ranges from 0 to 1, where 0 represents incorrect syntax, 1 represents correct syntax, and a confidence level between 0 and 1 can be understood as syntax accuracy. After obtaining a syntax score of each sentence in the text data, the server calculates an average value of the syntax scores of the sentences to obtain a syntax score of the text data. The syntax analysis network automatically learns based on the text data, and does not need to split each sentence in the text data or match a syntax structure.

Step S210: Randomly capture a video frame from the video data of the interviewee, extract a micro-expression feature of the interviewee based on the video frame, and obtain a confidence level of the video data based on the micro-expression feature.

The server randomly captures a video frame from the video data of the interviewee at a preset time interval, obtains the micro-expression feature of the interviewee based on the video frame, inputs the micro-expression feature into the trained micro-expression model to obtain a confidence level of the micro-expression feature belonging to each preset emotion category, sorts the confidence levels of the micro-expression feature belonging to the preset emotion categories, and obtains a maximum confidence level as the confidence level of the video data. The micro-expression model is obtained through training based on sample micro-expression data.

Step S212: Determine an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

The server can input the first speech emotion data, the second speech emotion data, and the syntax score into a trained audio classification model to obtain the confidence level of the audio data, and then determine the interview result for the interviewee based on the confidence level of the audio data, the confidence level of the video data, and a confidence level parameter. Specifically, parameters of the audio classification model include the confidence level of the audio data belonging to each preset emotion category in the first speech emotion data, the confidence level of the text data belonging to each preset emotion category in the second speech emotion data, and the syntax score. The audio classification model can be trained by using the sample speech data

and the sample text data that carry the label information as a training set. The label information is used to mark whether the interviewee corresponding to the sample speech data and the sample text data tells a lie. The confidence level parameter is an adjustable parameter, and can be set as required.

In the previously described data processing method, the micro-speech feature is extracted based on the audio data of the interviewee, the first speech emotion data is obtained based on the micro-speech feature, the audio data of the interviewee is converted into the text data, the text data is analyzed to obtain the second speech emotion data and the syntax score, the micro-expression feature is extracted based on the video data of the interviewee, the confidence level of the video data is obtained based on the micro-expression feature, and the interview result for the interviewee is determined based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data. Multiple features of the interviewee are recognized in multiple ways, and the interview result for the interviewee is determined based on multiple recognition results, so that a psychological state of the interviewee can be accurately and comprehensively captured, recognition accuracy is improved, and the interview result is closer to the actual situation.

In an embodiment, as shown in FIG. 3, step S204 includes the following:

Step S302: Invoke a speech feature extraction tool to extract the micro-speech feature of the interviewee based on the audio data of the interviewee, where the micro-speech feature includes a speech rate feature, a Mel-frequency cepstral coefficient, and a pitch feature.

Step S304: Input the micro-speech feature into the matched speech emotion classification model to obtain the first speech emotion data corresponding to the micro-speech feature.

The method for invoking the speech feature extraction tool to extract the Mel-frequency cepstral coefficient is as follows: performing fast Fourier transform on the audio data of the interviewee to obtain a spectrum, mapping the spectrum to the mel scale, taking the logarithm, and then performing discrete cosine transform to obtain the Mel-frequency cepstral coefficient. The pitch feature includes a current segment pitch average, a current segment pitch standard deviation, a historical pitch average, and a historical pitch standard deviation. The method for extracting the

current segment pitch average is as follows: performing fast Fourier transform on the audio data of the interviewee to obtain a spectrum diagram of the audio data, calculating a variance between each frequency band and a center value of the spectrum, performing summation on the variances, and then taking the square root. The historical pitch average and the historical pitch standard deviation are an average and a standard deviation calculated from the beginning of the current interview to the current segment for the interviewee. The data is memorized and stored in the server at the beginning of the interview. For ease of calculation, approximate calculation can be performed by using an exponential moving average value. Updated formulas are as follows:

$$\text{Historical pitch average} = \alpha * \text{Historical pitch average} + (1-\alpha) * \text{Current pitch average}$$

$$\text{Historical pitch standard deviation} = \alpha * \text{Historical pitch standard deviation} + (1-\alpha) * \text{Current pitch standard deviation}$$

In the formulas, α represents a weight parameter between 0 and 1, which can be set as required. The default value is 0.9 herein.

The speech rate feature includes a current speech rate, a historical speech rate average, and a historical speech rate standard deviation. The historical speech rate average and the historical speech rate standard deviation are calculated, memorized, and stored by the server at the beginning of the interview. Similarly, for ease of calculation, approximate calculation can be performed by using an exponential moving average value. Updated formulas are as follows:

$$\text{Historical speech rate average} = \alpha * \text{Historical speech rate average} + (1-\alpha) * \text{Current speech rate}$$

$$\text{Historical speech rate mean square error} = \alpha * \text{Historical speech rate mean square error} + (1-\alpha) * (\text{Current speech rate} - \text{Historical speech rate average})^2$$

$$\text{Historical speech rate standard deviation} = \text{Square root of historical speech rate mean square error}$$

In the previously described embodiment, the speech feature extraction tool is invoked to extract the micro-speech feature of the interviewee based on the audio data of the interviewee.

In an embodiment, as shown in FIG. 4, step S204 includes the following:

Step S402: Obtain gender information of the interviewee, and obtain a speech emotion

classification model that matches the gender information of the interviewee from a trained speech emotion classification model set, where the speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information includes emotion category information and gender information.

Step S404: Obtain a pitch feature, a Mel-frequency cepstral coefficient, and a speech rate feature from the micro-speech feature.

Step S406: Input the pitch feature, the Mel-frequency cepstral coefficient, and the speech rate feature into the matched speech emotion classification model, and obtain a confidence level of the micro-speech feature belonging to each preset emotion category, to obtain the first speech emotion data of the micro-speech feature.

The trained speech emotion classification model set includes speech emotion classification models that are obtained through training based on sample data for interviewees of different genders, that is, an emotion classification model for analyzing male speech data and an emotion classification model for analyzing female speech data. The server obtains the gender information of the interviewee, matches the trained speech emotion classification model set based on the gender information of the interviewee, and obtains the speech emotion classification model that matches the gender information of the interviewee from the trained speech emotion classification model set. The speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information includes emotion category information and gender information. The server divides the sample speech data based on the gender information, and performs model training based on the sample speech data obtained after division, to obtain the speech emotion classification model set.

The pitch feature includes a current segment pitch average, a current segment pitch standard deviation, a historical pitch average, and a historical pitch standard deviation. The speech rate feature includes a current speech rate, a historical speech rate average, and a historical speech rate standard deviation. The server inputs all features included in the three features as parameters into the matched speech emotion classification model. A convolutional neural network in the speech emotion classification model provides a confidence level of the micro-speech feature belonging to

each preset emotion category based on all the features.

In the previously described embodiment, the matched speech emotion classification model is obtained based on the gender information of the interviewee; the pitch feature, the Mel-frequency cepstral coefficient, and the speech rate feature are input into the matched speech emotion classification model; the confidence level of the micro-speech feature belonging to each preset emotion category is obtained, so as to obtain the first speech emotion data of the micro-speech feature.

In an embodiment, as shown in FIG. 5, before step S402, the method further includes the following:

Step S502: Obtain the sample speech data that carries the label information.

Step S504: Divide the sample speech data into a training set and a verification set.

Step S506: Perform model training based on the training set and an initial speech emotion classification model, to obtain the speech emotion classification model set.

Step S508: Perform model verification based on the verification set, and adjust each speech emotion classification model in the speech emotion classification model set.

After obtaining the sample speech data that carries the label information, the server first divides the sample speech data into a first sample speech dataset and a second sample speech dataset based on the gender information in the label information; divides each of the first sample speech dataset and the second sample speech dataset into a training set and a verification set; performs model training based on the training sets in the first sample speech dataset and the second sample speech dataset to obtain the first speech emotion classification model and the second speech emotion classification model; performs model verification based on the verification sets in the first sample speech dataset and the second sample speech dataset; and adjusts the first speech emotion classification model and the second speech emotion classification model. Each of the first sample speech dataset and the second sample speech dataset includes only sample speech data of interviewees of the same gender.

In the previously described embodiment, the sample speech data that carries the label information is obtained; the sample speech data is divided into the training set and the verification

set; model training is performed based on the training set, and model verification is performed based on the verification set, to obtain each speech emotion classification model in the speech emotion classification model set.

In an embodiment, as shown in FIG. 6, step S206 includes the following:

Step S602: Search for and match the preset dictionary corresponding to the trained emotion classification network based on each word in each sentence, and determine a sequence number corresponding to each word in each sentence in the dictionary.

Step S604: Input the sequence number corresponding to each word in each sentence in the dictionary into the emotion classification network, to obtain a confidence level of each sentence in the text data belonging to each preset emotion category.

Step S606: Obtain an average value of the confidence levels of the sentences in the text data belonging to the preset emotion categories, and obtain the confidence level of the text data belonging to each preset emotion category based on the average value of the confidence levels.

The emotion classification network may be a network that is based on BERT and includes an overlaid classification layer of N neurons (assuming that N emotions are preset). The server splits the text data into multiple sentences, performs word segmentation on each sentence, searches for and matches a dictionary corresponding to BERT based on each word in each sentence, converts each word into a sequence number corresponding to the word in the dictionary corresponding to BERT, inputs a sequence number of an entire sentence into BERT to obtain a confidence level of each sentence belonging to each preset emotion category, and then determines a confidence level of the text data belonging to each preset emotion category based on the confidence level of each sentence belonging to each preset emotion category, to obtain the second speech emotion data. The emotion classification network may be obtained through training based on first sample text data. Each sample sentence in the first sample text data carries label information, and the label information is emotion category information of each sample sentence.

In the previously described embodiment, the sequence number corresponding to each word in each sentence in the dictionary is input into the emotion classification network, to obtain the confidence level of each sentence in the text data belonging to each preset emotion category; and

then the confidence level of the text data belonging to each preset emotion category is obtained based on the confidence level of each sentence in the text data belonging to each preset emotion category.

In an embodiment, as shown in FIG. 7, step S212 includes the following:

Step S702: Obtain a confidence level of the audio data based on the first speech emotion data, the second speech emotion data, and the syntax score.

Step S704: Determine the interview result for the interviewee based on the confidence level of the audio data, the confidence level of the video data, and a preset confidence level parameter.

The server can input the first speech emotion data, the second speech emotion data, and the syntax score into a trained audio classification model to obtain the confidence level of the audio data, and then determine the interview result for the interviewee based on the confidence level of the audio data, the confidence level of the video data, and a confidence level parameter. Specifically, parameters of the audio classification model include the confidence level of the audio data belonging to each preset emotion category in the first speech emotion data, the confidence level of the text data belonging to each preset emotion category in the second speech emotion data, and the syntax score. The audio classification model can be trained by using the sample speech data and the sample text data that carry the label information as a training set. The label information is used to mark whether the interviewee corresponding to the sample speech data and the sample text data tells a lie. The confidence level parameter is an adjustable parameter, and can be set as required. The interview result can be obtained from the interview score. The formula for calculating the interview score may be as follows: Interview score = $A * \text{Confidence level of audio data} + B * \text{Confidence level of video data}$, where A and B represent confidence level parameters.

In an embodiment, before step S206, the method further includes the following:

obtaining the first sample text data, where each sample sentence in the first sample text data carries emotion category information; and

performing model training by using the first sample text data as a training set, to obtain the emotion classification network.

It should be understood that although the steps in the flowcharts of FIG. 2 to FIG. 7 are shown

in sequence indicated by arrows, these steps are not necessarily performed in such sequences. Unless expressly stated herein, these steps are not limited to a strict execution sequence, and may be performed in another sequence. In addition, at least some of the steps in FIG. 2 to FIG. 7 may include a plurality of substeps or stages. These substeps or stages are not necessarily performed and completed at the same moment, but may be performed at different moments. These substeps or stages are not necessarily performed in sequence, either, but may be performed in turn or alternately with other steps or at least some substeps or stages of the other steps.

In an embodiment, as shown in FIG. 8, a data processing apparatus is provided, including an acquisition module 802, a first extraction module 804, a first processing module 806, a second processing module 808, a second extraction module 810, and an analysis module 812.

The acquisition module 802 is configured to obtain audio data and video data of an interviewee.

The first extraction module 804 is configured to extract a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtain first speech emotion data based on the micro-speech feature.

The first processing module 806 is configured to: convert the audio data of the interviewee into text data, split the text data into multiple sentences, perform word segmentation on the multiple sentences, search for and match a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determine a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data.

The second processing module 808 is configured to: input the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculate an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data.

The second extraction module 810 is configured to randomly capture a video frame from the video data of the interviewee, extract a micro-expression feature of the interviewee based on the

video frame, and obtain a confidence level of the video data based on the micro-expression feature.

The analysis module 812 is configured to determine an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

In the previously described data processing apparatus, the micro-speech feature is extracted based on the audio data of the interviewee, the first speech emotion data is obtained based on the micro-speech feature, the audio data of the interviewee is converted into the text data, the text data is analyzed to obtain the second speech emotion data and the syntax score, the micro-expression feature is extracted based on the video data of the interviewee, the confidence level of the video data is obtained based on the micro-expression feature, and the interview result for the interviewee is determined based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data. Multiple features of the interviewee are recognized in multiple ways, and the interview result for the interviewee is determined based on multiple recognition results, so that a psychological state of the interviewee can be accurately and comprehensively captured, recognition accuracy is improved, and the interview result is closer to the actual situation.

In an embodiment, the first extraction module is further configured to invoke a speech feature extraction tool to extract the micro-speech feature of the interviewee based on the audio data of the interviewee, where the micro-speech feature includes a speech rate feature, a Mel-frequency cepstral coefficient, and a pitch feature.

In an embodiment, the first extraction module is further configured to: obtain gender information of the interviewee; obtain a speech emotion classification model that matches the gender information of the interviewee from a trained speech emotion classification model set, where the speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information includes emotion category information and gender information; obtain a pitch feature, a Mel-frequency cepstral coefficient, and a speech rate feature from the micro-speech feature; input the pitch feature, the Mel-frequency cepstral coefficient, and the speech rate feature into the matched speech emotion classification model; and

obtain a confidence level of the micro-speech feature belonging to each preset emotion category, to obtain the first speech emotion data of the micro-speech feature.

In an embodiment, the first extraction module is further configured to: obtain the sample speech data that carries the label information; divide the sample speech data into a training set and a verification set; perform model training based on the training set and an initial speech emotion classification model, to obtain the speech emotion classification model set; and perform model verification based on the verification set, and adjust each speech emotion classification model in the speech emotion classification model set.

In an embodiment, the first processing module is further configured to: search for and match the preset dictionary corresponding to the trained emotion classification network based on each word in each sentence, and determine a sequence number corresponding to each word in each sentence in the dictionary; input the sequence number corresponding to each word in each sentence in the dictionary into the emotion classification network, to obtain a confidence level of each sentence in the text data belonging to each preset emotion category; and obtain an average value of the confidence levels of the sentences in the text data belonging to the preset emotion categories, and obtain the confidence level of the text data belonging to each preset emotion category based on the average value of the confidence levels.

In an embodiment, the analysis module is further configured to: obtain a confidence level of the audio data based on the first speech emotion data, the second speech emotion data, and the syntax score; and determine the interview result for the interviewee based on the confidence level of the audio data, the confidence level of the video data, and a preset confidence level parameter.

In an embodiment, the first processing module is further configured to: obtain the first sample text data, where each sample sentence in the first sample text data carries emotion category information; and perform model training by using the first sample text data as a training set, to obtain the emotion classification network.

For a specific limitation on the data processing apparatus, refer to the limitation on the data processing method described above. Details are not described herein again. All or some of the modules in the data processing apparatus may be implemented by using software, hardware, or a

combination thereof. The foregoing modules may be embedded in or independent of a processor of a computer device in a hardware form, or may be stored in a memory of a computer device in a software form, so that the processor invokes the modules and performs operations corresponding to the foregoing modules.

In an embodiment, a computer device is provided. The computer device may be a server. An internal structural diagram of the computer device may be shown in FIG. 9. The computer device includes a processor, a memory, and a network interface that are connected to each other by using a system bus. The processor of the computer device is configured to provide computing and control capabilities. The memory of the computer device includes a non-volatile storage medium and an internal memory. The non-volatile storage medium stores an operating system and computer readable instructions. The internal memory provides an environment for running the operating system and the computer readable instructions in the non-volatile storage medium. The network interface of the computer device is configured to communicate with an external terminal through a network connection. The computer readable instructions are executed by the processor to implement a data processing method.

A person skilled in the art can understand that the structure shown in FIG. 9 is merely a block diagram of a partial structure related to the solution of this application, and does not constitute a limitation on the computer device to which the solution of this application is applied. Specifically, the computer device may include more or fewer components than those shown in the figure, or combine some components, or have different component arrangements.

In an embodiment, a computer device includes a memory and one or more processors, where the memory stores computer readable instructions, and when the computer readable instructions are executed by the one or more processors, the one or more processors are enabled to perform the following steps:

- obtaining audio data and video data of an interviewee;

- extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

In an embodiment, one or more non-volatile computer readable storage media storing computer readable instructions are provided, where when the computer readable instructions are executed by one or more processors, the one or more processors are enabled to perform the following steps:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, where the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, where the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

A person of ordinary skill in the art can understand that all or some of the processes of the methods in the embodiments may be implemented by the computer readable instructions instructing related hardware. The computer readable instructions may be stored in a non-volatile computer readable storage medium. When the computer readable instructions are executed, the processes of the method embodiments are performed. Any reference to a memory, a storage, a database, or another medium used in the embodiments provided in this application may include a non-volatile memory and/or volatile memory. The non-volatile memory may include a read-only memory (ROM), a programmable ROM (PROM), an electrically programmable ROM (EPROM), an electrically erasable programmable ROM (EEPROM), or a flash memory. The volatile memory may include a random access memory (RAM) or an external cache memory. By way of illustration and not limitation, the RAM is available in a variety of formats, such as a static RAM (SRAM), a dynamic RAM (DRAM), a synchronous DRAM (SDRAM), a double data rate SDRAM (DDR SDRAM), an enhanced SDRAM (ESDRAM), a synchlink DRAM (SLDRAM), a Rambus direct RAM (RDRAM), a direct Rambus dynamic RAM (DRDRAM), and a Rambus dynamic RAM (RDRAM).

The technical features of the foregoing embodiments may be combined arbitrarily. For brevity of description, not all possible combinations of the technical features of the foregoing embodiments are described. However, the combinations of these technical features should be considered as falling within the scope of the present specification provided that there is no contradiction between the

combinations.

The foregoing embodiments merely describe several implementations of this application, and description of the implementations is relatively specific and detailed, but shall not be understood as a limitation on the scope of the present invention. It should be noted that a person of ordinary skill in the art may make several variations and improvements without departing from the idea of this application, and the variations and improvements shall fall within the protection scope of this application. Therefore, the protection scope of this application shall be subject to the protection scope of the claims.

CLAIMS

1. A data processing method, comprising:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, wherein the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, wherein the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

2. The method according to claim 1, wherein the extracting a micro-speech feature of the interviewee based on the audio data of the interviewee comprises:

invoking a speech feature extraction tool to extract the micro-speech feature of the interviewee based on the audio data of the interviewee, wherein the micro-speech feature comprises a speech rate feature, a Mel-frequency cepstral coefficient, and a pitch feature.

3. The method according to claim 1, wherein the obtaining first speech emotion data based on the micro-speech feature comprises:

obtaining gender information of the interviewee, and obtaining a speech emotion classification model that matches the gender information of the interviewee from a trained speech emotion classification model set, wherein the speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information comprises emotion category information and gender information;

obtaining a pitch feature, a Mel-frequency cepstral coefficient, and a speech rate feature from the micro-speech feature; and

inputting the pitch feature, the Mel-frequency cepstral coefficient, and the speech rate feature into the matched speech emotion classification model, and obtaining a confidence level of the micro-speech feature belonging to each preset emotion category, to obtain the first speech emotion data of the micro-speech feature.

4. The method according to claim 3, wherein before the obtaining a speech emotion classification model that matches the gender information of the interviewee from a trained speech emotion classification model set, the method further comprises:

obtaining the sample speech data that carries the label information;

dividing the sample speech data into a training set and a verification set;

performing model training based on the training set and an initial speech emotion classification model, to obtain the speech emotion classification model set; and

performing model verification based on the verification set, and adjusting each speech emotion classification model in the speech emotion classification model set.

5. The method according to claim 1, wherein the searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data comprises:

searching for and matching the preset dictionary corresponding to the trained emotion classification network based on each word in each sentence, and determining a sequence number corresponding to each word in each sentence in the dictionary;

inputting the sequence number corresponding to each word in each sentence in the dictionary into the emotion classification network, to obtain a confidence level of each sentence in the text data belonging to each preset emotion category; and

obtaining an average value of the confidence levels of the sentences in the text data belonging to the preset emotion categories, and obtaining the confidence level of the text data belonging to each preset emotion category based on the average value of the confidence levels.

6. The method according to claim 1, wherein the determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data comprises:

obtaining a confidence level of the audio data based on the first speech emotion data, the second speech emotion data, and the syntax score; and

determining the interview result for the interviewee based on the confidence level of the audio data, the confidence level of the video data, and a preset confidence level parameter.

7. The method according to claim 1, wherein before the searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, the method further comprises:

obtaining the first sample text data, wherein each sample sentence in the first sample text data carries emotion category information; and

performing model training by using the first sample text data as a training set, to obtain the emotion classification network.

8. A data processing apparatus, comprising:

an acquisition module, configured to obtain audio data and video data of an interviewee;

a first extraction module, configured to extract a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtain first speech emotion data based on the micro-speech feature;

a first processing module, configured to: convert the audio data of the interviewee into text data, split the text data into multiple sentences, perform word segmentation on the multiple sentences, search for and match a preset dictionary corresponding to a trained emotion classification

network based on each word in each sentence, and determine a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, wherein the emotion classification network is obtained through training based on first sample text data;

a second processing module, configured to: input the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculate an average value of the syntax scores of the sentences to obtain a syntax score of the text data, wherein the syntax analysis network is obtained through training based on second sample text data;

a second extraction module, configured to randomly capture a video frame from the video data of the interviewee, extract a micro-expression feature of the interviewee based on the video frame, and obtain a confidence level of the video data based on the micro-expression feature; and

an analysis module, configured to determine an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

9. The apparatus according to claim 8, wherein the first extraction module is further configured to invoke a speech feature extraction tool to extract the micro-speech feature of the interviewee based on the audio data of the interviewee, wherein the micro-speech feature comprises a speech rate feature, a Mel-frequency cepstral coefficient, and a pitch feature.

10. A computer device, comprising a memory and one or more processors, wherein the memory stores computer readable instructions, and when the computer readable instructions are executed by the one or more processors, the one or more processors are enabled to perform the following steps:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in

each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, wherein the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, wherein the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

11. The computer device according to claim 10, wherein when executing the computer readable instructions, the processor further performs the following steps:

invoking a speech feature extraction tool to extract the micro-speech feature of the interviewee based on the audio data of the interviewee, wherein the micro-speech feature comprises a speech rate feature, a Mel-frequency cepstral coefficient, and a pitch feature.

12. The computer device according to claim 10, wherein when executing the computer readable instructions, the processor further performs the following steps:

obtaining gender information of the interviewee, and obtaining a speech emotion classification model that matches the gender information of the interviewee from a trained speech emotion classification model set, wherein the speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information comprises emotion category information and gender information;

obtaining a pitch feature, a Mel-frequency cepstral coefficient, and a speech rate feature from the micro-speech feature; and

inputting the pitch feature, the Mel-frequency cepstral coefficient, and the speech rate feature into the matched speech emotion classification model, and obtaining a confidence level of the

micro-speech feature belonging to each preset emotion category, to obtain the first speech emotion data of the micro-speech feature.

13. The computer device according to claim 10, wherein when executing the computer readable instructions, the processor further performs the following steps:

obtaining the sample speech data that carries the label information;

dividing the sample speech data into a training set and a verification set;

performing model training based on the training set and an initial speech emotion classification model, to obtain the speech emotion classification model set; and

performing model verification based on the verification set, and adjusting each speech emotion classification model in the speech emotion classification model set.

14. The computer device according to claim 10, wherein when executing the computer readable instructions, the processor further performs the following steps:

searching for and matching the preset dictionary corresponding to the trained emotion classification network based on each word in each sentence, and determining a sequence number corresponding to each word in each sentence in the dictionary;

inputting the sequence number corresponding to each word in each sentence in the dictionary into the emotion classification network, to obtain a confidence level of each sentence in the text data belonging to each preset emotion category; and

obtaining an average value of the confidence levels of the sentences in the text data belonging to the preset emotion categories, and obtaining the confidence level of the text data belonging to each preset emotion category based on the average value of the confidence levels.

15. The computer device according to claim 10, wherein when executing the computer readable instructions, the processor further performs the following steps:

obtaining a confidence level of the audio data based on the first speech emotion data, the second speech emotion data, and the syntax score; and

determining the interview result for the interviewee based on the confidence level of the audio data, the confidence level of the video data, and a preset confidence level parameter.

16. One or more non-volatile computer readable storage media storing computer readable instructions, wherein when the computer readable instructions are executed by one or more processors, the one or more processors are enabled to perform the following steps:

obtaining audio data and video data of an interviewee;

extracting a micro-speech feature of the interviewee based on the audio data of the interviewee, and obtaining first speech emotion data based on the micro-speech feature;

converting the audio data of the interviewee into text data, splitting the text data into multiple sentences, performing word segmentation on the multiple sentences, searching for and matching a preset dictionary corresponding to a trained emotion classification network based on each word in each sentence, and determining a confidence level of the text data belonging to each preset emotion category based on a searching and matching result, to obtain second speech emotion data, wherein the emotion classification network is obtained through training based on first sample text data;

inputting the text data into a trained syntax analysis network to obtain a syntax score of each sentence in the text data, and calculating an average value of the syntax scores of the sentences to obtain a syntax score of the text data, wherein the syntax analysis network is obtained through training based on second sample text data;

randomly capturing a video frame from the video data of the interviewee, extracting a micro-expression feature of the interviewee based on the video frame, and obtaining a confidence level of the video data based on the micro-expression feature; and

determining an interview result for the interviewee based on the first speech emotion data, the second speech emotion data, the syntax score, and the confidence level of the video data.

17. The storage media according to claim 16, wherein when executing the computer readable instructions, the processor further performs the following steps:

invoking a speech feature extraction tool to extract the micro-speech feature of the interviewee based on the audio data of the interviewee, wherein the micro-speech feature comprises a speech rate feature, a Mel-frequency cepstral coefficient, and a pitch feature.

18. The storage media according to claim 16, wherein when executing the computer readable instructions, the processor further performs the following steps:

obtaining gender information of the interviewee, and obtaining a speech emotion classification model that matches the gender information of the interviewee from a trained speech emotion classification model set, wherein the speech emotion classification model is obtained through training based on sample speech data that carries label information, and the label information comprises emotion category information and gender information;

obtaining a pitch feature, a Mel-frequency cepstral coefficient, and a speech rate feature from the micro-speech feature; and

inputting the pitch feature, the Mel-frequency cepstral coefficient, and the speech rate feature into the matched speech emotion classification model, and obtaining a confidence level of the micro-speech feature belonging to each preset emotion category, to obtain the first speech emotion data of the micro-speech feature.

19. The storage media according to claim 16, wherein when executing the computer readable instructions, the processor further performs the following steps:

obtaining the sample speech data that carries the label information;

dividing the sample speech data into a training set and a verification set;

performing model training based on the training set and an initial speech emotion classification model, to obtain the speech emotion classification model set; and

performing model verification based on the verification set, and adjusting each speech emotion classification model in the speech emotion classification model set.

20. The storage media according to claim 16, wherein when executing the computer readable instructions, the processor further performs the following steps:

searching for and matching the preset dictionary corresponding to the trained emotion classification network based on each word in each sentence, and determining a sequence number corresponding to each word in each sentence in the dictionary;

inputting the sequence number corresponding to each word in each sentence in the dictionary into the emotion classification network, to obtain a confidence level of each sentence in the text data belonging to each preset emotion category; and

obtaining an average value of the confidence levels of the sentences in the text data belonging

to the preset emotion categories, and obtaining the confidence level of the text data belonging to each preset emotion category based on the average value of the confidence levels.