

SPEECH EMOTION RECOGNITION METHOD AND APPARATUS BASED ON TARGET SCENARIO, DEVICE, AND MEDIUM

[0001] The present application claims priority to Chinese Patent Application No.
5 2025109181231, filed with the China National Intellectual Property Administration on
July 3, 2025 and entitled “SPEECH EMOTION RECOGNITION METHOD AND
APPARATUS BASED ON TARGET SCENARIO, DEVICE, AND MEDIUM”, which is
incorporated herein by reference in its entirety.

TECHNICAL FIELD

10 [0002] The present application relates to the technical field of speech semantics, and in
particular, to a speech emotion recognition method and apparatus based on a target
scenario, a device, and a medium.

BACKGROUND

[0003] The inventor has realized that, in intelligent speech quality inspection systems,
15 emotion recognition capabilities directly affect service quality assessment and customer
satisfaction monitoring. However, traditional technologies generally rely on fixed time
windows for speech quality analysis, which cannot capture fine-grained emotional
fluctuations and overlook the dynamic changes in customer emotions throughout the
service process, resulting in the inadequate recognition of critical turning points.

20 [0004] For example, in the financial technology field, intelligent speech emotion
recognition is widely applied to scenarios such as customer service hotlines and financial
advisory services. When customers inquire about financial products or conduct
transactions, the emotional states change in response to factors such as the satisfaction
with the service and the level of understanding regarding the product terms. Traditional
25 speech emotion analysis methods, which rely on fixed time windows, are difficult to
capture the subtle emotional fluctuations exhibited by customers during consultations,
such as a momentary flash of dissatisfaction upon learning of high service fees, or a brief
moment of pleasant surprise after discovering product returns. These critical emotional
turning points cannot be accurately identified, thereby affecting the accuracy of service
30 quality assessment and customer satisfaction monitoring.

[0005] For example, in the healthcare field, intelligent speech emotion recognition is used in scenarios such as hospital customer service hotlines and online medical consultations. When patients consult about illnesses or schedule appointments, their emotional states are influenced by factors such as the severity of illness and waiting time.

- 5 Traditional analysis methods relying on fixed time windows cannot capture fine-grained emotional fluctuations of patients, such as a sudden surge in anxiety of patients while describing their illnesses.

[0006] Therefore, how to improve the accuracy and recognition efficiency of speech emotion recognition has become an urgent problem to be solved.

10 SUMMARY

[0007] The present application provides a speech emotion recognition method and apparatus based on a target scenario, a device, and a medium, which mainly aims to solve problems of inaccuracy and low efficiency in speech emotion recognition.

- [0008] According to a first aspect, to achieve the above purpose, the present application
15 provides a speech emotion recognition method based on a target scenario, including:

[0009] acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature;

[0010] performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature;

- 20 [0011] performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature;

[0012] performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and

- 25 [0013] acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion trajectory in the target speech scenario.

[0014] According to a second aspect, the present application further provides a computer device, including:

- 30 [0015] at least one processor; and

[0016] a memory communicatively connected to the at least one processor, where

[0017] the memory stores a computer program that can be executed by the at least one

processor, and the computer program is executed by the at least one processor, so that the at least one processor can execute the above speech emotion recognition method based on a target scenario, and the method includes:

[0018] acquiring an initial speech signal, and performing signal encoding on the initial
5 speech signal, to obtain a speech encoding feature;

[0019] performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature;

[0020] performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the
10 plurality of speech level features, to obtain a target speech feature;

[0021] performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and

[0022] acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion
15 trajectory in the target speech scenario.

[0023] According to a third aspect, the present application further provides a non-volatile computer-readable storage medium, storing at least one computer program, where the at least one computer program is executed by a processor in a computer device to implement the above speech emotion recognition method based on a target scenario, and
20 the method includes:

[0024] acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature;

[0025] performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature;

[0026] performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the
25 plurality of speech level features, to obtain a target speech feature;

[0027] performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and

[0028] acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion
30 trajectory in the target speech scenario.

[0029] According to a fourth aspect, the present application further provides a speech emotion recognition apparatus based on a target scenario, including:

[0030] a speech signal encoding module, configured to: acquire an initial speech signal, and perform signal encoding on the initial speech signal, to obtain a speech encoding feature;

[0031] an acoustic feature extraction module, configured to: perform acoustic feature
5 extraction on the speech encoding feature, to obtain an initial acoustic feature;

[0032] a speech feature generation module, configured to: perform feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and perform hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature;

10 [0033] an emotion feature extraction module, configured to: perform emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and

[0034] an emotion trajectory recognition module, configured to: acquire a target speech scenario, map the plurality of speech emotion features to the target speech scenario, and
15 recognize a mapped target speech emotion trajectory in the target speech scenario.

[0035] In embodiments of the present application, performing signal enhancement on an initial speech signal can effectively eliminate noise interference and improve speech quality, and framing windowing and feature extraction can accurately capture time-domain and frequency-domain characteristics of a speech signal, thereby enhancing
20 feature comprehensiveness and representativeness and improving accuracy and efficiency of speech processing. Acquiring a spectrum envelope through cepstral analysis reflects sound channel characteristics and enhances feature discriminability, to obtain a more robust initial acoustic feature. An attention value reflects real-time importance of different levels of speech while suppressing irrelevant characteristics, thereby generating a target
25 speech feature having local details and global consistency, and improving an expression capability of complex speech data. Richness of emotion expression is enhanced through rhythm analysis and emotion clustering, thereby improving accuracy and reliability of emotion analysis. Speech emotion features are mapped to a target speech scenario, so that an association between an emotion and a scenario can be accurately positioned, and an
30 expression of the emotion in a specific scenario can be understood more deeply. A mapped speech emotion trajectory is recognized, so that changes of the emotion over time can be dynamically captured, helping implement a more intelligent emotion interaction.

BRIEF DESCRIPTION OF THE DRAWINGS

[0036] To describe the technical solutions in the embodiments of the present application more clearly, the following briefly describes the accompanying drawings required for describing the embodiments. Apparently, the accompanying drawings in the following description show merely some embodiments of the present application, and those of
5 ordinary skill in the art may still derive other drawings from these accompanying drawings without making creative efforts.

[0037] FIG. 1 is a schematic diagram of an application environment of a speech emotion recognition method based on a target scenario according to an embodiment of the present application;

10 [0038] FIG. 2 is a schematic flowchart of a speech emotion recognition method based on a target scenario according to an embodiment of the present application;

[0039] FIG. 3 is a schematic flowchart of recognizing a mapped target speech emotion trajectory in a target speech scenario according to an embodiment of the present application;

15 [0040] FIG. 4 is a schematic diagram of modules of a speech emotion recognition apparatus based on a target scenario according to an embodiment of the present application;

[0041] FIG. 5 is a schematic structural diagram of a computer device of a speech emotion recognition method based on a target scenario according to an embodiment of
20 the present application; and

[0042] FIG. 6 is another schematic structural diagram of a computer device of a speech emotion recognition method based on a target scenario according to an embodiment of the present application.

[0043] The purpose realization, functional features, and advantages of the present
25 application will be further described with reference to the embodiments and accompanying drawings.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0044] The technical solutions in the embodiments of the present application are clearly and completely described below with reference to the accompanying drawings in the
30 embodiments of the present application. Apparently, the described embodiments are only some rather than all of the embodiments of the present application. All other embodiments obtained by those of ordinary skill in the art based on the embodiments of the present

application without making creative efforts shall fall within the protection scope of the present application.

[0045] An embodiment of the present application provides a speech emotion recognition method based on a target scenario. An execution body of the speech emotion recognition method based on a target scenario includes but is not limited to at least one of a service end, a terminal, and the like that can be configured to execute a computer device of an apparatus provided in the embodiment of the present application. In other words, the speech emotion recognition method based on a target scenario may be executed by software or hardware mounted on a terminal device or a service end device. The service end includes but is not limited to: a single server, a server cluster, a cloud server, a cloud server cluster, or the like. The server may be an independent server or a cloud server providing fundamental cloud computing services such as cloud services, cloud databases, cloud computing, cloud functions, cloud storage, network services, cloud communication, middleware services, domain name services, security services, content delivery networks (CDNs), and big data and artificial intelligence platforms.

[0046] The speech emotion recognition method based on a target scenario of the present application may be applied to an application environment shown in FIG. 1. A client communicates with the service end through a network. The service end can acquire an initial speech signal through the client, performing signal enhancement on the initial speech signal can effectively eliminate noise interference and improve speech quality, and framing windowing and feature extraction can accurately capture time-domain and frequency-domain characteristics of a speech signal, thereby enhancing feature comprehensiveness and representativeness and improving accuracy and efficiency of speech processing. Acquiring a spectrum envelope through cepstral analysis reflects sound channel characteristics and enhances feature discriminability, to obtain a more robust initial acoustic feature. An attention value reflects real-time importance of different levels of speech while suppressing irrelevant characteristics, thereby generating a target speech feature having local details and global consistency, and improving an expression capability of complex speech data. Richness of emotion expression is enhanced through rhythm analysis and emotion clustering, thereby improving accuracy and reliability of emotion analysis. Speech emotion features are mapped to a target speech scenario, so that an association between an emotion and a scenario can be accurately positioned, and an expression of the emotion in a specific scenario can be understood more deeply. A mapped speech emotion trajectory is recognized, so that changes of the emotion over time can be

dynamically captured, helping implement a more intelligent emotion interaction. Finally, an emotion trajectory result is output and fed back to the client. The client may be but is not limited to various personal computers, laptops, smart phones, tablet computers, and portable wearable devices. The service end may be implemented by using an independent
5 server or a server cluster composed of a plurality of servers. The present application is described in detail below through specific embodiments.

[0047] FIG. 2 is a schematic flowchart of a speech emotion recognition method based on a target scenario according to an embodiment of the present application. In this embodiment, the speech emotion recognition method based on a target scenario includes:

10 [0048] S1: acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature.

[0049] In this embodiment of the present application, the initial speech signal refers to an unprocessed original audio electrical signal directly collected by acoustic-electro conversion devices such as microphones, which exists in the form of a continuous analog
15 signal, includes feature information such as speech content, pitch, timbre, and speaking speed of a speaker, and serves as basic data of subsequent speech processing (for example speech recognition, speech synthesis, and speech enhancement).

[0050] Detailedly, the initial speech signal may be collected through a hardware device, for example, a microphone, to convert an acoustic wave into an electrical signal, or a
20 sound card, to convert an analog signal output by the microphone into a digital signal (ADC). The analog signal is discretized at a fixed frequency (for example, 8 kHz, 16 kHz, and 44.1 kHz), that is, continuous voltage amplitudes are mapped to discrete digital values. The digital signal is transmitted to a processing device through Bluetooth, Wi-Fi, or a wired interface.

25 [0051] For example, in a healthcare scenario, the initial speech signal refers to speech information of a patient or a doctor collected by devices such as the microphone. The speech information includes content such as illness descriptions, diagnosis suggestions, and medical records. Upon arriving at a hospital, the patient describes his symptoms, for example, "I've been coughing constantly lately, and I also have a fever", via a microphone
30 at an intelligent reception desk. A system collects this speech signal, then processes and analyzes the speech signal, and subsequently recommends a corresponding medical department to the patient.

[0052] For example, in a financial technology scenario, the initial speech signal refers to speech information generated by a customer during an interaction with a financial

institution, for example, inquiring about services, conducting transactions, and undergoing identity verification. The customer calls a customer service hotline of a bank to inquire about credit card repayments, stating, “I would like to check my credit card repayment date”. A customer service system collects this speech signal, then converts the

5 speech signal into text through a speech recognition technology, and then provides the customer with corresponding information according to a preset procedure.

[0053] In this embodiment of the present application, the performing signal encoding on the initial speech signal, to obtain a speech encoding feature includes:

[0054] performing signal enhancement processing on the initial speech signal, to obtain

10 an enhanced speech signal;

[0055] extracting a time-domain feature and a frequency-domain feature of the enhanced speech signal;

[0056] performing feature fusion on the time-domain feature and the frequency-domain feature, to obtain a fused spatial-temporal feature; and

15 [0057] performing nonlinear transformation on the fused spatial-temporal feature, to obtain the speech encoding feature.

[0058] Detailedly, the initial speech signal generally includes various types of noise, for example, environmental noise and device noise. The signal enhancement processing aims to improve quality of the initial speech signal and suppress noise interference, to enable

20 subsequent processing to be more accurate and reliable.

[0059] Specifically, assuming that the noise is stable, a noise power spectrum is calculated, then the noise power spectrum is subtracted from a power spectrum of noisy speech, to obtain an enhanced speech power spectrum, and finally, an enhanced speech signal in a time domain is obtained through inverse Fourier transform. For example, in a

25 silent office environment, when sudden noise of a working printer intrudes upon the speech signal, a spectral subtraction method may eliminate the noise by analyzing spectrum characteristics of the noise.

[0060] The present application may model the noisy speech by using a deep neural network such as a convolutional neural network (CNN) and learn noise and speech

30 characteristics, thereby enhancing the speech signal. For example, a neural network model is trained by using a large amount of noisy speech and corresponding pure speech data, to enable the model to learn how to recover pure speech from the noisy speech.

[0061] Detailedly, the time-domain feature and the frequency-domain feature describe characteristics of a speech signal from different perspectives respectively. The time-

domain feature extraction includes: calculating energy of each frame of the speech signal, which reflects an amplitude change of the speech signal, and counting the number of times of a signal value in each frame of the speech signal crossing a zero point, which reflects a frequency characteristic of the speech signal. The frequency-domain feature
5 extraction is to transform each frame of the speech signal from a time domain to a frequency domain, to obtain a spectrum of the speech signal, and features such as frequency components and formants of the speech signal may be extracted by analyzing the spectrum.

[0062] Specifically, the time-domain feature and the frequency-domain feature are
10 concatenated into a longer feature vector according to a certain order. For example, time-domain features such as short-time energy and short-time zero-crossing rate are concatenated with frequency-domain features such as MFCC to form a comprehensive feature vector, or different attention weights may be allocated to different features according to feature importance through an attention mechanism network, thereby
15 improving quality of the fused spatial-temporal feature.

[0063] Detailedly, the fused spatial-temporal feature is mapped to a more compact and discriminative feature space through the nonlinear transformation. The nonlinear transformation may be performed on the fused spatial-temporal feature by using a multi-layer fully connected neural network. A DNN may be trained through a large amount of
20 data to learn a complex nonlinear relationship between an input feature and an output encoding feature. For example, in a speech recognition task, the DNN may map the fused spatial-temporal feature to a low-dimensional encoding space while retaining key information of the speech signal.

[0064] In this embodiment of the present application, performing signal enhancement
25 on the initial speech signal can effectively eliminate noise interference and improve speech quality, and framing windowing and feature extraction can accurately capture time-domain and frequency-domain characteristics of the speech signal, thereby enhancing feature comprehensiveness and representativeness, improving accuracy and efficiency of speech processing, and enhancing a capability of a computer to process the
30 speech signal.

[0065] S2: performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature.

[0066] In this embodiment of the present application, by performing a progressive processing procedure of framing processing, energy extraction, spectrum analysis,

cepstral transform, and multi-level feature extraction and fusion on the speech encoding feature, the speech encoding feature is gradually converted into the initial acoustic feature having rich acoustic information.

[0067] In this embodiment of the present application, the performing acoustic feature
5 extraction on the speech encoding feature, to obtain an initial acoustic feature includes:

[0068] performing framing processing on the speech encoding feature, to obtain a framing encoding feature;

[0069] performing feature energy extraction on the framing encoding feature, to obtain a framing energy feature;

10 [0070] performing cepstral analysis on the framing energy feature, to obtain a spectrum envelope feature;

[0071] performing first acoustic feature extraction on the spectrum envelope feature, to obtain a first acoustic feature;

[0072] performing second acoustic feature extraction on the first acoustic feature, to
15 obtain a second acoustic feature; and

[0073] performing weighted fusion on the first acoustic feature and the second acoustic feature, to obtain the initial acoustic feature.

[0074] Detailedly, the framing processing is to segment a continuous speech encoding feature sequence into a plurality of short-time frames, facilitating subsequent feature
20 analysis on each frame. A fixed frame length (for example, the number of sampling points corresponding to 25 ms: assuming that a sampling rate is 16 kHz, the frame length is 400 sampling points) and a frame shift (for example, 160 sampling points corresponding to 10 ms) are determined, and fixed-length data is successively intercepted from the speech encoding feature sequence according to the frame shift, to form the framing encoding
25 feature.

[0075] Specifically, the framing energy feature reflects energy of each frame of the speech signal and may be used for distinguishing between an unvoiced sound and a voiced sound. The energy of the voiced sound is relatively large, and the energy of the unvoiced sound is relatively small, which is one of the important features for analysis of the speech
30 signal. For each frame of the framing encoding feature, an energy value of each frame is calculated according to an energy calculation formula, to obtain the framing energy feature.

[0076] Detailedly, the cepstral analysis may decompose a spectrum of the speech signal into two parts: a spectrum envelope and a spectrum detail. The spectrum envelope feature

reflects the sound channel characteristics and is an important feature in tasks such as speech recognition and speaker recognition. Discrete cosine transform is performed on each line (that is, a framing spectrum feature of each frame) of a target spectrum feature matrix, and the signal is transformed from the frequency domain to a cepstral domain, to
5 obtain a cepstral coefficient. The spectrum envelope feature generally corresponds to a low-order part in the cepstral coefficient.

[0077] For example, after the discrete cosine transform is performed on a frame of a 26-dimensional framing spectrum feature, the first 12 to 13 cepstral coefficients are taken as a spectrum envelope feature of this frame.

10 [0078] In the present application, a more representative and discriminative acoustic feature is extracted from the spectrum envelope feature, for example, a first-order difference and a second-order difference. The first-order difference reflects a change rate of the feature over time, and the second-order difference reflects a change rate of a feature change rate. The static spectrum envelope feature, the first-order difference, and the
15 second-order difference are combined together to be used as the first acoustic feature.

[0079] Specifically, feature extraction is performed on the first acoustic feature by using deep neural networks (for example, a convolutional neural network (CNN), a recurrent neural network (RNN), and variants thereof such as an LSTM and a GRU). A convolutional layer performs a sliding convolutional operation on the first acoustic
20 feature through a convolution kernel to extract a local feature. A pooling layer performs downsampling on an output of the convolutional layer to reduce feature dimensions. A fully connected layer integrates and classifies an output of the pooling layer to output the second acoustic feature. Different weights are allocated for feature fusion according to importance of the first acoustic feature and the second acoustic feature to obtain the initial
25 acoustic feature.

[0080] In this embodiment of the present application, the framing processing enables a non-stationary speech signal to be analyzed frame by frame. The energy extraction can distinguish between the unvoiced and voiced sounds to provide key information for speech analysis. Filtering processing highlights an effective frequency component and
30 improves accuracy of the spectrum feature. The spectrum envelope is acquired through the cepstral analysis, which reflects the sound channel characteristics and enhances the feature discriminability, to obtain a more robust initial acoustic feature.

[0081] S3: performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the

plurality of speech level features, to obtain a target speech feature.

[0082] In this embodiment of the present application, the speech level features are constructed through principal component dimension reduction, multi-granularity unit division, and time series modeling, and feature fusion is implemented by using a multi-head attention mechanism, to obtain the target speech feature having rich semantic information.

[0083] In this embodiment of the present application, the performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features includes:

10 [0084] performing principal component dimension reduction processing on the initial acoustic feature, to obtain a target acoustic feature;

[0085] performing unit division on the target acoustic feature, to obtain a speech phoneme unit;

15 [0086] performing phoneme recognition on the speech phoneme unit, to obtain a speech phoneme feature;

[0087] performing sliding window average processing on the target acoustic feature, to obtain a speech segment-level feature;

[0088] performing time series modeling analysis on the target acoustic feature, to obtain a time series feature convolutional layer;

20 [0089] performing lexical analysis on the target acoustic feature according to the time series feature convolutional layer, to obtain a speech lexical feature; and

[0090] generating the speech level features according to the speech phoneme feature, the speech segment-level feature, and the speech lexical feature.

[0091] Detailedly, the principal component dimension reduction processing aims to map a high-dimensional acoustic feature to a low-dimensional space through linear transformation while retaining principal information of data. Principal component analysis is a common dimension reduction method. A covariance matrix of the initial acoustic feature is calculated, and then feature values and feature vectors of the covariance matrix are solved. The first few feature vectors are selected in a descending order according to the feature values, to constitute a projection matrix. The initial acoustic feature is projected onto the projection matrix, to obtain a dimension-reduced target acoustic feature.

25
30

[0092] Specifically, phoneme units are divided by analyzing energy and a zero-crossing change of the target acoustic feature. A position where the energy suddenly increases or

the zero-crossing rate changes obviously may correspond to a boundary of a phoneme, thereby dividing the target acoustic feature into the plurality of speech phoneme units. A phoneme category corresponding to each speech phoneme unit is recognized, and a representative phoneme feature is extracted, thereby providing a basis for subsequent

5 speech analysis and processing.

[0093] In the present application, based on phoneme recognition of a hidden Markov model, a state transition model is established for each phoneme. By calculating matching probabilities between observation sequences (that is, features of the speech phoneme units) and various phoneme models, the most probable phoneme category is determined.

10 [0094] Specifically, the sliding window average processing may smooth the target acoustic feature and reduce the impact of noise. A fixed sliding window is defined and sequentially slides on target acoustic feature sequences. For feature values in each window, an average value thereof is calculated to be used as a speech segment-level feature of a position corresponding to the window.

15 [0095] Detailedly, the speech signal has an obvious time series characteristic. The time series modeling analysis may capture a change law of the target acoustic feature in a time dimension and extract a feature with time series information. Local feature extraction is performed on the target acoustic feature through the convolutional neural network, and the convolution kernel slides on the time dimension, to extract a local time series feature

20 and capture a time series dependence relationship between features.

[0096] For example, a convolutional operation is performed on the target acoustic feature by using a one-dimensional convolutional layer, to extract the local time series feature, and then time series modeling is performed on an output of the convolutional layer by using the fully connected layer, to obtain the time series feature convolutional

25 layer.

[0097] In the present application, by using the time series information captured by the time series feature convolutional layer, analysis at a lexical level is performed on the target acoustic feature, to recognize the vocabulary in the speech. Classifiers such as a fully connected neural network or a support vector machine (SVM) may be used to

30 perform lexical classification on the target acoustic feature. Different levels of speech features are integrated, to obtain the speech level features.

[0098] In this embodiment of the present application, the performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature includes:

[0099] performing product calculation on the speech level features and a preset first query weight, second key weight, and third value weight respectively, to obtain a first query matrix, a second key matrix, and a third value matrix;

[0100] aggregating the first query matrix, the second key matrix, and the third value
5 matrix into an attention matrix;

[0101] by using a pre-constructed multi-head attention mechanism, calculating attention values corresponding to different attention heads according to the attention matrix; and

[0102] performing feature fusion on the speech level features according to the attention values, to obtain the target speech feature.

10 [0103] In this embodiment of the present application, the speech level features are multiplied by preset query, key, and value weight matrices respectively, to generate a query matrix, a key matrix, and a value matrix, which are combined to form an attention matrix. The attention matrix is divided into a plurality of sub-spaces, that is, attention heads. An attention weight is determined through a matching degree between the query
15 and the key, and then weighted summing is performed on the value matrix, to obtain attention values of various heads.

[0104] Detailedly, the first query weight is to map the speech level features to a “query” space and represents information that needs to be focused on currently. The second key weight is used for mapping the input feature to a “key” space and represents referable
20 information. The third value weight is used for mapping the input feature to a “value” space and represents actually carried content information.

[0105] Specifically, the query matrix, the key matrix, and the value matrix are concatenated or combined into a three-dimensional attention matrix according to a specific dimension which generally is a feature dimension, thereby providing a unified
25 input for subsequent multi-head attention calculation. The attention matrix is divided into a plurality of sub-spaces along a certain dimension (for example, a channel dimension). Each sub-space is referred to as an “attention head”. For example, if an 8-head attention is used, the matrix is divided into 8 parts.

[0106] In the present application, for each attention head, similarity (for example, dot
30 product) between a query vector and a key vector is calculated, to obtain an attention score, which represents a matching degree between the query and the key. In addition, the attention score is normalized and converted into a probability distribution, which represents attention weights for different positions. Weighted summing is performed on the value matrix according to the normalized weights, to obtain an attention value of a

current head. Outputs of all attention heads are concatenated or averaged, to obtain a final multi-head attention value, thereby enhancing an expression capability.

[0107] Detailedly, the attention values reflect real-time importance of different levels of speech, and differentiated weighting is performed on various scale features according to magnitudes of the importance. By combining the attention values with the speech level features, key information in current context can be highlighted while suppressing irrelevant characteristics, thereby generating the target speech feature having local details and global consistency, and improving the expression capability of complex speech data.

[0108] S4: performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features.

[0109] In this embodiment of the present application, through multi-level processing of dynamic and static spectrum feature analysis, rhythm and emotion feature fusion, and high-dimensional emotion space projection, accurate mapping and extraction from speech features to multi-dimensional emotion features are implemented.

[0110] In this embodiment of the present application, the performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features includes:

[0111] extracting a dynamic spectrum feature and a static spectrum feature of the target speech feature respectively;

[0112] performing rhythm analysis on the dynamic spectrum feature, to obtain a rhythm change feature;

[0113] performing emotion clustering analysis on the static spectrum feature, to obtain an emotion category feature;

[0114] performing feature fusion on the rhythm change feature and the emotion category feature, to obtain a combined emotion feature;

[0115] performing dimension expansion processing on the combined emotion feature, to obtain a high-dimensional emotion feature space; and

[0116] performing multi-emotion dimension projection on the combined emotion feature in the high-dimensional emotion feature space, to obtain the plurality of speech emotion features.

[0117] Detailedly, the dynamic spectrum feature reflects a change situation of the speech signal over time and can capture dynamic information in the speech, for example, tone rising and falling and speaking speed changes. The static spectrum feature reflects a spectrum characteristic of the speech at a certain moment. The dynamic and static

spectrum features can be combined to comprehensively describe spectrum information of the speech, providing a basis for emotion analysis.

[0118] Specifically, the framing processing is performed on the target speech feature, to obtain a spectrum representation. A difference of a spectrum between adjacent frames is
5 calculated, for example, a first-order difference, and a second-order difference, to obtain the dynamic spectrum feature. Features such as an amplitude and power of the spectrum are directly extracted to be used as the static spectrum feature.

[0119] In this embodiment of the present application, rhythm is a factor closely related to the emotion expression. Rhythm features, such as a change law of a fundamental
10 frequency, energy, and duration, that reflect speech emotion changes may be extracted through the rhythm analysis. A fundamental frequency of each frame is extracted by using a self-correlation function, a cepstrum method, and the like, and a statistical feature of the fundamental frequency is calculated, for example, a fundamental frequency average value, a fundamental frequency range, and a fundamental frequency change rate. Duration of
15 different phonemes, syllables, or words in the speech is counted, and a distribution feature of the duration is calculated, for example, average duration and a duration variance. Statistical results of the features such as the fundamental frequency, the energy, and the duration are combined to be used as the rhythm change features.

[0120] Detailedly, speech samples having similar emotion features are classified into a
20 category through clustering analysis, thereby obtaining a feature reflecting a speech emotion category and helping recognize an emotion expressed by the speech. A proper clustering quantity is selected, which is generally determined according to the emotion category (for example, happiness, sadness, anger, and calmness). A plurality of clustering centers are initialized randomly, each static spectrum feature vector is allocated to a
25 category where the closest clustering center is located, and a clustering center of each category is recalculated. A process of allocating and updating the clustering center is repeated until the clustering center does not change or reaches the maximum number of times of iterations, and a clustering category to which each speech sample belongs is used as the emotion category feature.

30 [0121] Specifically, a rhythm change feature vector and an emotion category feature vector are directly concatenated into a longer feature vector according to a certain order, that is, the combined emotion feature.

[0122] In the present application, the combined emotion feature is mapped to a high-dimensional feature space by using a kernel function, and similarity between features is

calculated through the kernel function, which actually is that the features are mapped to the high-dimensional space. The multi-emotion dimension projection may project the features in the high-dimensional emotion feature space onto a plurality of different emotion dimensions, thereby obtaining the plurality of speech emotion features and more

5 comprehensively describing emotion information of the speech.

[0123] Detailedly, linear discriminant analysis projection enables projections of samples of different emotion categories in this direction to be separated as much as possible and projections of samples of the same emotion category to be aggregated as much as possible. A high-dimensional emotion feature vector is multiplied by a projection
10 matrix, to obtain a projected low-dimensional feature. These low-dimensional features may be regarded as representations on different emotion dimensions.

[0124] In this embodiment of the present application, the rhythm analysis captures the dynamic emotion change, and the emotion clustering clarifies the emotion category, thereby enhancing emotion expression richness. The dimension expansion improves a
15 feature expression capability, and the multi-emotion dimension projection describes the emotion from a plurality of perspectives, so that the computer can recognize different emotion states more accurately, thereby improving accuracy and reliability of emotion analysis and providing better quality of feature support for applications related to speech emotion recognition.

20 [0125] S5: acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion trajectory in the target speech scenario.

[0126] For example, in a healthcare field, an emotion state of the patient has an important impact on illness diagnosis and treatment. Through a speech emotion
25 recognition technology, the doctor may better learn about a psychological state of the patient, thereby providing more personalized medical services.

[0127] During communication between the patient and the doctor, speech data of the patient is collected in real time, an emotion feature is extracted and input into a trained emotion classification model to recognize the emotion state of the patient, and the doctor
30 adjusts a communication manner and a therapeutic scheme according to the recognized emotion state of the patient. For example, for an anxious patient, the doctor may give more comfort and explanations; and for a depressed patient, the doctor may recommend psychological therapy.

[0128] For example, in a financial technology field, an emotion state of the customer

has an important impact on sales and customer service quality of financial products. Through a speech emotion recognition technology, the financial institution may better learn about the requirements and emotions of the customer, thereby providing more personalized financial products and services.

5 [0129] During communication between the customer and customer service staff or sales personnel, speech data of the customer is collected in real time, and an emotion feature is extracted and input into the trained emotion classification model to recognize the emotion state of the customer. The customer service staff or sales personnel adjust a communication manner and a sales strategy according to the recognized emotion state of
10 the customer. For example, for a disgruntled customer, a problem is solved in time, and compensation is given; and for a hesitant customer, more product information and special offers are provided, thereby facilitating the completion of transactions.

[0130] In the present application, through the speech emotion recognition technology, the financial institution may better learn about the requirements and emotions of the
15 customer, thereby improving satisfaction and loyalty of the customer. At the same time, sales efficiency and customer service quality can also be improved, thereby increasing revenue of the financial institution.

[0131] In this embodiment of the present application, the plurality of speech emotion features are mapped to the target speech scenario, which is linear mapping based on
20 scenario adaptation. By using a powerful linear fitting capability of the deep neural network, a complex mapping relationship from the speech emotion features to the target scenarios is learned.

[0132] Specifically, a deep neural network model may be constructed, for example, a multi-layer perceptron (MLP) or a long short-term memory (LSTM) network. A large
25 amount of speech emotion features and corresponding scenario tags are used as training data to train the model, and after the training is completed, extracted speech emotion features are input into the model, and scenarized speech features mapped to the target scenario are output.

[0133] As shown in FIG. 3, in this embodiment of the present application, the
30 recognizing a mapped target speech emotion trajectory in the target speech scenario includes:

[0134] acquiring a scenarized speech feature in the target speech scenario, and performing time-domain segmentation on the scenarized speech feature, to obtain an emotion feature sequence;

[0135] extracting an emotion feature parameter in the emotion feature sequence;

[0136] performing multi-dimensional emotion prediction on the emotion feature parameter according to a preset emotion trajectory prediction model, to obtain an emotion state transition path; and

- 5 [0137] generating a continuous emotion state sequence according to the emotion state transition path; and drawing a corresponding initial speech emotion trajectory according to the emotion state sequence, and performing smoothing processing on the initial speech emotion trajectory, to obtain the target speech emotion trajectory.

[0138] Detailedly, semantic analysis is performed on the target speech scenario to
10 recognize a key element in the scenario, and scenario element information is combined to extract the scenarized speech feature. The scenarized speech feature includes acoustic characteristics of the speech itself and incorporates scenario-related semantic information.

[0139] Specifically, the speech emotion dynamically changes over time. The speech is segmented on the time domain, which can capture feature changes of the emotion in
15 different time periods, thereby forming an emotion feature sequence. A segmentation method based on acoustic features such as energy and fundamental frequency may be used, and an energy threshold is set. When the energy of the speech signal exceeds the threshold, it is considered that a new speech segment starts. A corresponding emotion feature of each segment is extracted, to form the emotion feature sequence.

20 [0140] Detailedly, statistical quantities of each feature in the emotion feature sequence are calculated, for example, an average value, a variance, a maximum value, and a minimum value. For a fundamental frequency feature sequence, calculating an average value thereof may reflect an average pitch of the speech, and a variance may reflect a change degree of the pitch.

25 [0141] In this embodiment of the present application, a mapping relationship between an emotion feature parameter and an emotion state is learned by using a machine learning or deep learning algorithm, thereby predicting a change path of the emotion state. The emotion state is regarded as a hidden state, the emotion feature parameter is regarded as an observation state, and a transition probability between the hidden states and an
30 emission probability from the hidden state to the observation state are learned. During prediction, the most probable hidden state sequence is found according to an input emotion feature parameter by using a Viterbi algorithm, that is, the emotion state transition path.

[0142] Detailedly, a possible transition relationship between different emotion states

may be described through an initial emotion state transition map, providing a reference framework for generating a continuous emotion state sequence. In combination with the emotion state transition path and the initial emotion state transition map, the emotion state is corrected and supplemented, generating a continuous emotion state sequence that better

5 conforms to an actual emotion change.

[0143] Specifically, the emotion state transition path is matched with the initial emotion state transition map, to check whether a state transition in the path conforms to a probability relationship in the transition map. For a non-conforming transition, it is corrected according to the probability in the transition map. At the same time, a possibly
10 missing state in the path is supplemented by using information in the transition map, generating the continuous emotion state sequence.

[0144] In the present application, each emotion state in the emotion state sequence is mapped to a corresponding emotion dimension, and various emotion state points are connected through a timeline, to form the initial speech emotion trajectory. According to
15 values of each state in the emotion state sequence on various dimensions, corresponding points in a three-dimensional coordinate system are drawn, and then these points are connected by using a line segment, to obtain the initial speech emotion trajectory. For each point on the initial speech emotion trajectory, an average value of a certain amount of points around this point is used as a new value of this point, thereby smoothing the
20 emotion trajectory and obtaining a more accurate target speech emotion trajectory.

[0145] In this embodiment of the present application, the speech emotion features are mapped to the target speech scenario, so that an association between an emotion and a scenario can be accurately positioned, thereby understanding an expression of the emotion in a specific scenario more deeply. The mapped speech emotion trajectory is
25 recognized, so that a change in the emotion over time can be dynamically captured, thereby helping implement a more intelligent emotion interaction, for example, an intelligent customer service can adjust a communication strategy according to an emotion trajectory of the customer in a financial consultation scenario.

[0146] It should be understood that the serial number of each step in the above
30 embodiments does not indicate the execution sequence, which should be determined by the function and internal logic of the step, and shall not limit the implementation of the embodiments of the present application.

[0147] FIG. 4 is a diagram of functional modules of a speech emotion recognition apparatus based on a target scenario according to an embodiment of the present

application.

[0148] In this embodiment of this disclosure, a speech emotion recognition apparatus based on a target scenario is provided, and the speech emotion recognition apparatus based on a target scenario is in one-to-one correspondence with the speech emotion
5 recognition method based on a target scenario in the above embodiments. As shown in FIG. 4, the speech emotion recognition apparatus 100 based on a target scenario may be mounted in a computer device. According to implemented functions, the speech emotion recognition apparatus 100 based on a target scenario includes a speech signal encoding
10 module 101, an acoustic feature extraction module 102, a speech feature generation module 103, an emotion feature extraction module 104, and an emotion trajectory recognition module 105. Various functional modules are described in detail below:

[0149] the speech signal encoding module is configured to: acquire an initial speech signal, and perform signal encoding on the initial speech signal, to obtain a speech encoding feature;

15 [0150] the acoustic feature extraction module is configured to: perform acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature;

[0151] the speech feature generation module is configured to: perform feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and perform hierarchical attention processing on the plurality of speech level features, to
20 obtain a target speech feature;

[0152] the emotion feature extraction module is configured to: perform emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features;
and

[0153] the emotion trajectory recognition module is configured to: acquire a target
25 speech scenario, map the plurality of speech emotion features to the target speech scenario, and recognize a mapped target speech emotion trajectory in the target speech scenario.

[0154] In the present application, specific limitations related to the speech emotion recognition apparatus based on a target scenario may refer to limitations to the speech emotion recognition method based on a target scenario in the above description, which
30 are not described herein again. Various modules in the above speech emotion recognition apparatus based on a target scenario may be implemented in whole or in part by using software, hardware, and a combination thereof. These modules may be embedded in or independent from a processor in a computer device in a hardware form or may be stored in a memory in the computer device in a software form so that the processor invokes and

executes operations corresponding to these modules.

[0155] In an embodiment, a computer device is provided. The computer device may be a service end, and an internal structure diagram thereof may be shown in FIG. 5. The computer device includes a processor, a memory, a network interface, and a database that
5 are connected through 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 and/or a volatile storage medium and an internal memory. The non-volatile storage medium stores an operating system, a computer program, and a database. The internal memory provides an environment for running the operating system
10 and the computer program stored in the non-volatile storage medium. The network interface of the computer device is configured to communicate with an external client via network connections. The computer program, when executed by the processor, implements functions or steps on a service end side of the speech emotion recognition method based on a target scenario.

15 [0156] In an embodiment, a computer device is provided. The computer device may be a client, and an internal structure diagram thereof may be shown in FIG. 6. The computer device includes a processor, a memory, a network interface, a display screen, and an input apparatus that are connected through a system bus. The processor of the computer device is configured to provide computing and control capabilities. The memory of the computer
20 device includes a non-volatile storage medium and an internal memory. The non-volatile storage medium stores an operating system and a computer program. The internal memory provides an environment for running the operating system and the computer program stored in the non-volatile storage medium. The network interface of the computer device is configured to communicate with an external server via network
25 connections. The computer program, when executed by the processor, implements functions or steps on a client side of the speech emotion recognition method based on a target scenario.

[0157] In an embodiment, a computer device is provided, including a memory, a processor, and a computer program stored on the memory and executable on the processor,
30 where the processor implements the following steps when executing the computer program:

[0158] acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature;

[0159] performing acoustic feature extraction on the speech encoding feature, to obtain

an initial acoustic feature;

[0160] performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature;

- 5 [0161] performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and

[0162] acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion trajectory in the target speech scenario.

- 10 [0163] In the several embodiments provided in the present application, it should be understood that the disclosed device and apparatus may be implemented in other ways. For example, the system embodiments described above are only examples. For example, the division of the modules is only a logical function division and may be another division in actual implementation.

- 15 [0164] In addition, various functional modules in various embodiments of the present application may be integrated into one processing unit, or each unit may physically exist separately, or two or more units may be integrated into one unit. The above integrated units may be implemented in the form of hardware or in the form of hardware with software functional modules.

- 20 [0165] In some implementations of the present application, a computer-readable storage medium is provided, having a computer program stored thereon, where the computer program, when executed by a processor, implements the steps in the method in the above embodiments.

- [0166] The readable storage medium of the present application stores a computer
25 program, and the computer program, when executed by a processor of a computer device, may implement:

[0167] acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature;

- [0168] performing acoustic feature extraction on the speech encoding feature, to obtain
30 an initial acoustic feature;

[0169] performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature;

[0170] performing emotion feature extraction on the target speech feature, to obtain a

plurality of speech emotion features; and

[0171] acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion trajectory in the target speech scenario.

5 [0172] It should be noted that the above functions or steps that can be implemented by the computer-readable storage medium or the computer device may correspondingly refer to related descriptions on a service end side and a client side in the above method embodiments, which are not described herein one by one for avoiding repetition.

[0173] The computer-readable storage medium may further store at least one computer-executable program/instruction, and the computer-executable program/instruction may be, 10 for example, a computer-readable instruction. The computer-readable storage medium may be non-volatile or volatile. A volatile memory may, for example, include a random access memory (RAM), a cache memory, and/or the like. The computer-readable storage medium may, for example, include a read-only memory (ROM), a hard disk drive, a flash 15 memory, and the like.

[0174] In an implementation, the at least one computer-executable instruction may also be compiled or combined into a software product/computer program product, where one or more computer-executable instructions, when run by a processor, execute various functions and/or steps of the method in the embodiments described in this technology.

20 [0175] Those of ordinary skill in the art may understand that all or part of the processes in the above method embodiments may be completed by instructing relevant hardware through a computer program. The computer program may be stored in a non-volatile computer-readable storage medium, and the computer program may include the processes of the above method embodiments when executed. Any reference to a memory, a storage, 25 a database, or any other medium used in the embodiments provided in the present application may include a non-volatile and/or a volatile memory.

[0176] Those skilled in the art may clearly understand that for the convenience and simplicity of description, the division of the above functional units and modules is taken for example only. In practical applications, the above function allocation may be 30 completed by different functional units and modules as needed, that is, the internal structure of the apparatus is divided into different functional units or modules to complete all or part of the above functions.

[0177] The above embodiments are provided only to illustrate the technical solutions of the present application and are not intended to limit the present application. Any software

tools or assemblies not belonging to the company that appear in the embodiments of the present application are used merely for illustrative purposes and do not represent actual usage.

What is claimed is:

1. A speech emotion recognition method based on a target scenario, comprising:
 - acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature;
 - 5 performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature;
 - performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features, and performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature;
 - 10 performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and
 - acquiring a target speech scenario, mapping the plurality of speech emotion features to the target speech scenario, and recognizing a mapped target speech emotion trajectory in the target speech scenario.
- 15 2. The speech emotion recognition method based on a target scenario according to claim 1, wherein the performing signal encoding on the initial speech signal, to obtain a speech encoding feature comprises:
 - performing signal enhancement processing on the initial speech signal, to obtain an enhanced speech signal;
 - 20 extracting a time-domain feature and a frequency-domain feature of the enhanced speech signal;
 - performing feature fusion on the time-domain feature and the frequency-domain feature, to obtain a fused spatial-temporal feature; and
 - performing nonlinear transformation on the fused spatial-temporal feature, to obtain
 - 25 the speech encoding feature.
3. The speech emotion recognition method based on a target scenario according to claim 1, wherein the performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature comprises:
 - performing framing processing on the speech encoding feature, to obtain a framing
 - 30 encoding feature;
 - performing feature energy extraction on the framing encoding feature, to obtain a framing energy feature;
 - performing cepstral analysis on the framing energy feature, to obtain a spectrum

envelope feature;

performing first acoustic feature extraction on the spectrum envelope feature, to obtain a first acoustic feature;

performing second acoustic feature extraction on the first acoustic feature, to obtain
5 a second acoustic feature; and

performing weighted fusion on the first acoustic feature and the second acoustic feature, to obtain the initial acoustic feature.

4. The speech emotion recognition method based on a target scenario according to claim 1, wherein the performing feature level division on the initial acoustic feature, to
10 obtain a plurality of speech level features comprises:

performing principal component dimension reduction processing on the initial acoustic feature, to obtain a target acoustic feature;

performing unit division on the target acoustic feature, to obtain a speech phoneme unit;

15 performing phoneme recognition on the speech phoneme unit, to obtain a speech phoneme feature;

performing sliding window average processing on the target acoustic feature, to obtain a speech segment-level feature;

performing time series modeling analysis on the target acoustic feature, to obtain a
20 time series feature convolutional layer;

performing lexical analysis on the target acoustic feature according to the time series feature convolutional layer, to obtain a speech lexical feature; and

generating the speech level features according to the speech phoneme feature, the speech segment-level feature, and the speech lexical feature.

25 5. The speech emotion recognition method based on a target scenario according to claim 1, wherein the performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature comprises:

performing product calculation on the speech level features and a preset first query weight, second key weight, and third value weight respectively, to obtain a first query
30 matrix, a second key matrix, and a third value matrix;

aggregating the first query matrix, the second key matrix, and the third value matrix into an attention matrix;

by using a pre-constructed multi-head attention mechanism, calculating attention values corresponding to different attention heads according to the attention matrix; and

performing feature fusion on the speech level features according to the attention values, to obtain the target speech feature.

6. The speech emotion recognition method based on a target scenario according to claim 1, wherein the performing emotion feature extraction on the target speech feature,
5 to obtain a plurality of speech emotion features comprises:

extracting a dynamic spectrum feature and a static spectrum feature of the target speech feature respectively;

performing rhythm analysis on the dynamic spectrum feature, to obtain a rhythm change feature;

10 performing emotion clustering analysis on the static spectrum feature, to obtain an emotion category feature;

performing feature fusion on the rhythm change feature and the emotion category feature, to obtain a combined emotion feature;

performing dimension expansion processing on the combined emotion feature, to
15 obtain a high-dimensional emotion feature space; and

performing multi-emotion dimension projection on the combined emotion feature in the high-dimensional emotion feature space, to obtain the plurality of speech emotion features.

7. The speech emotion recognition method based on a target scenario according to
20 claim 1, wherein the recognizing a mapped target speech emotion trajectory in the target speech scenario comprises:

acquiring a scenarized speech feature in the target speech scenario, and performing time-domain segmentation on the scenarized speech feature, to obtain an emotion feature sequence;

25 extracting an emotion feature parameter in the emotion feature sequence;

performing multi-dimensional emotion prediction on the emotion feature parameter according to a preset emotion trajectory prediction model, to obtain an emotion state transition path; and

generating a continuous emotion state sequence according to the emotion state
30 transition path; and

drawing a corresponding initial speech emotion trajectory according to the emotion state sequence, and performing smoothing processing on the initial speech emotion trajectory, to obtain the target speech emotion trajectory.

8. A computer device, comprising:

at least one processor; and
a memory communicatively connected to the at least one processor, wherein
the memory stores a computer program that can be executed by the at least one
processor, and the computer program is executed by the at least one processor, so that the
5 at least one processor can execute the speech emotion recognition method based on a
target scenario according to claim 1, and the method comprises:

acquiring an initial speech signal, and performing signal encoding on the initial
speech signal, to obtain a speech encoding feature;

performing acoustic feature extraction on the speech encoding feature, to obtain an
10 initial acoustic feature;

performing feature level division on the initial acoustic feature, to obtain a plurality
of speech level features, and performing hierarchical attention processing on the plurality
of speech level features, to obtain a target speech feature;

performing emotion feature extraction on the target speech feature, to obtain a
15 plurality of speech emotion features; and

acquiring a target speech scenario, mapping the plurality of speech emotion features
to the target speech scenario, and recognizing a mapped target speech emotion trajectory
in the target speech scenario.

9. The computer device according to claim 8, wherein the performing signal
20 encoding on the initial speech signal, to obtain a speech encoding feature comprises:

performing signal enhancement processing on the initial speech signal, to obtain an
enhanced speech signal;

extracting a time-domain feature and a frequency-domain feature of the enhanced
speech signal;

25 performing feature fusion on the time-domain feature and the frequency-domain
feature, to obtain a fused spatial-temporal feature; and

performing nonlinear transformation on the fused spatial-temporal feature, to obtain
the speech encoding feature.

10. The computer device according to claim 8, wherein the performing acoustic
30 feature extraction on the speech encoding feature, to obtain an initial acoustic feature
comprises:

performing framing processing on the speech encoding feature, to obtain a framing
encoding feature;

performing feature energy extraction on the framing encoding feature, to obtain a

framing energy feature;

performing cepstral analysis on the framing energy feature, to obtain a spectrum envelope feature;

performing first acoustic feature extraction on the spectrum envelope feature, to
5 obtain a first acoustic feature;

performing second acoustic feature extraction on the first acoustic feature, to obtain
a second acoustic feature; and

performing weighted fusion on the first acoustic feature and the second acoustic
feature, to obtain the initial acoustic feature.

10 11. The computer device according to claim 8, wherein the performing feature level
division on the initial acoustic feature, to obtain a plurality of speech level features
comprises:

performing principal component dimension reduction processing on the initial
acoustic feature, to obtain a target acoustic feature;

15 performing unit division on the target acoustic feature, to obtain a speech phoneme
unit;

performing phoneme recognition on the speech phoneme unit, to obtain a speech
phoneme feature;

performing sliding window average processing on the target acoustic feature, to
20 obtain a speech segment-level feature;

performing time series modeling analysis on the target acoustic feature, to obtain a
time series feature convolutional layer;

performing lexical analysis on the target acoustic feature according to the time series
feature convolutional layer, to obtain a speech lexical feature; and

25 generating the speech level features according to the speech phoneme feature, the
speech segment-level feature, and the speech lexical feature.

12. The computer device according to claim 8, wherein the performing hierarchical
attention processing on the plurality of speech level features, to obtain a target speech
feature comprises:

30 performing product calculation on the speech level features and a preset first query
weight, second key weight, and third value weight respectively, to obtain a first query
matrix, a second key matrix, and a third value matrix;

aggregating the first query matrix, the second key matrix, and the third value matrix
into an attention matrix;

by using a pre-constructed multi-head attention mechanism, calculating attention values corresponding to different attention heads according to the attention matrix; and performing feature fusion on the speech level features according to the attention values, to obtain the target speech feature.

5 13. The computer device according to claim 8, wherein the performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features comprises:

extracting a dynamic spectrum feature and a static spectrum feature of the target speech feature respectively;

10 performing rhythm analysis on the dynamic spectrum feature, to obtain a rhythm change feature;

performing emotion clustering analysis on the static spectrum feature, to obtain an emotion category feature;

15 performing feature fusion on the rhythm change feature and the emotion category feature, to obtain a combined emotion feature;

performing dimension expansion processing on the combined emotion feature, to obtain a high-dimensional emotion feature space; and

20 performing multi-emotion dimension projection on the combined emotion feature in the high-dimensional emotion feature space, to obtain the plurality of speech emotion features.

14. The computer device according to claim 8, wherein the recognizing a mapped target speech emotion trajectory in the target speech scenario comprises:

25 acquiring a scenarized speech feature in the target speech scenario, and performing time-domain segmentation on the scenarized speech feature, to obtain an emotion feature sequence;

extracting an emotion feature parameter in the emotion feature sequence;

performing multi-dimensional emotion prediction on the emotion feature parameter according to a preset emotion trajectory prediction model, to obtain an emotion state transition path; and

30 generating a continuous emotion state sequence according to the emotion state transition path; and

drawing a corresponding initial speech emotion trajectory according to the emotion state sequence, and performing smoothing processing on the initial speech emotion trajectory, to obtain the target speech emotion trajectory.

15. A non-volatile computer-readable storage medium, storing a computer program, wherein the computer program, when executed by a processor, implements the speech emotion recognition method based on a target scenario according to claim 1, and the method comprises:

- 5 acquiring an initial speech signal, and performing signal encoding on the initial speech signal, to obtain a speech encoding feature;
 - performing acoustic feature extraction on the speech encoding feature, to obtain an initial acoustic feature;
 - performing feature level division on the initial acoustic feature, to obtain a plurality
 - 10 of speech level features, and performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature;
 - performing emotion feature extraction on the target speech feature, to obtain a plurality of speech emotion features; and
 - acquiring a target speech scenario, mapping the plurality of speech emotion features
 - 15 to the target speech scenario, and recognizing a mapped target speech emotion trajectory in the target speech scenario.

16. The non-volatile computer-readable storage medium according to claim 15, wherein the performing signal encoding on the initial speech signal, to obtain a speech encoding feature comprises:

- 20 performing signal enhancement processing on the initial speech signal, to obtain an enhanced speech signal;
 - extracting a time-domain feature and a frequency-domain feature of the enhanced speech signal;
 - performing feature fusion on the time-domain feature and the frequency-domain
 - 25 feature, to obtain a fused spatial-temporal feature; and
 - performing nonlinear transformation on the fused spatial-temporal feature, to obtain the speech encoding feature.

17. The non-volatile computer-readable storage medium according to claim 15, wherein the performing acoustic feature extraction on the speech encoding feature, to

30 obtain an initial acoustic feature comprises:

- performing framing processing on the speech encoding feature, to obtain a framing encoding feature;
- performing feature energy extraction on the framing encoding feature, to obtain a framing energy feature;

performing cepstral analysis on the framing energy feature, to obtain a spectrum envelope feature;

performing first acoustic feature extraction on the spectrum envelope feature, to obtain a first acoustic feature;

5 performing second acoustic feature extraction on the first acoustic feature, to obtain a second acoustic feature; and

performing weighted fusion on the first acoustic feature and the second acoustic feature, to obtain the initial acoustic feature.

18. The non-volatile computer-readable storage medium according to claim 15,
10 wherein the performing feature level division on the initial acoustic feature, to obtain a plurality of speech level features comprises:

performing principal component dimension reduction processing on the initial acoustic feature, to obtain a target acoustic feature;

15 performing unit division on the target acoustic feature, to obtain a speech phoneme unit;

performing phoneme recognition on the speech phoneme unit, to obtain a speech phoneme feature;

performing sliding window average processing on the target acoustic feature, to obtain a speech segment-level feature;

20 performing time series modeling analysis on the target acoustic feature, to obtain a time series feature convolutional layer;

performing lexical analysis on the target acoustic feature according to the time series feature convolutional layer, to obtain a speech lexical feature; and

25 generating the speech level features according to the speech phoneme feature, the speech segment-level feature, and the speech lexical feature.

19. The non-volatile computer-readable storage medium according to claim 15, wherein the performing hierarchical attention processing on the plurality of speech level features, to obtain a target speech feature comprises:

30 performing product calculation on the speech level features and a preset first query weight, second key weight, and third value weight respectively, to obtain a first query matrix, a second key matrix, and a third value matrix;

aggregating the first query matrix, the second key matrix, and the third value matrix into an attention matrix;

by using a pre-constructed multi-head attention mechanism, calculating attention

values corresponding to different attention heads according to the attention matrix; and
performing feature fusion on the speech level features according to the attention
values, to obtain the target speech feature.

20. A speech emotion recognition apparatus based on a target scenario, comprising:
- 5 a speech signal encoding module, configured to: acquire an initial speech signal, and
perform signal encoding on the initial speech signal, to obtain a speech encoding feature;
an acoustic feature extraction module, configured to: perform acoustic feature
extraction on the speech encoding feature, to obtain an initial acoustic feature;
a speech feature generation module, configured to: perform feature level division on
10 the initial acoustic feature, to obtain a plurality of speech level features, and perform
hierarchical attention processing on the plurality of speech level features, to obtain a target
speech feature;
an emotion feature extraction module, configured to: perform emotion feature
extraction on the target speech feature, to obtain a plurality of speech emotion features;
15 and
an emotion trajectory recognition module, configured to: acquire a target speech
scenario, map the plurality of speech emotion features to the target speech scenario, and
recognize a mapped target speech emotion trajectory in the target speech scenario.