

# **METHODS AND APPARATUSES FOR DRIVING AN INTERACTIVE OBJECT, DEVICES AND STORAGE MEDIA**

## **CROSS REFERENCE TO RELATED APPLICATIONS**

The present disclosure is based on Chinese Patent Application No. 2020102458024, filed on March 31, 2020, and claims priority to the Chinese Patent Application, which is hereby incorporated herein by reference in its entirety.

## **TECHNICAL FIELD**

**[0001]** The present disclosure relates to the field of computer technology, and in particular, to a method and apparatus for driving an interactive object, a device and a storage medium.

## **BACKGROUND**

**[0002]** Human-computer interaction is mostly based on keystrokes, touch events and voice input, which is responded by presenting images, text, or a virtual human on a display screen. Currently, the virtual human is mostly obtained by improving on a voice assistant.

## **SUMMARY**

**[0003]** Embodiments of the present disclosure provide solutions for driving an interactive object.

**[0004]** According to an aspect of the present disclosure, provided is a method for driving an interactive object, which includes: obtaining a phoneme sequence corresponding to text data; obtaining values of control parameters for at least one local region of the interactive object matching the phoneme sequence; and controlling a posture of the interactive object based on the values of the control parameters.

**[0005]** In combination with any embodiment provided by the present disclosure, the method further includes at least one of the following: controlling a display device displaying the interactive object to display text based on the text data; or controlling the display device to output a voice based on the phoneme sequence corresponding to the text data.

**[0006]** In combination with any embodiment provided by the present disclosure, the control parameters for the at least one local region of the interactive object include a posture control vector of the at least one local region; wherein obtaining values of control parameters for at least one local region of the interactive object matching the phoneme sequence includes: obtaining a first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence; obtaining a feature encoding corresponding to at least one phoneme based on the first encoding sequence; and obtaining

a posture control vector of at least one local region of the interactive object corresponding to the feature encoding.

**[0007]** In combination with any embodiment provided by the present disclosure, obtaining the first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence includes: generating, for each of multiple phonemes contained in the phoneme sequence, a encoding sub-sequence corresponding to the phoneme; and obtaining a first encoding sequence corresponding to the phoneme sequence based on the encoding sub-sequence corresponding to each of the multiple phonemes.

**[0008]** In combination with any embodiment provided by the present disclosure, generating, for each of the multiple phonemes contained in the phoneme sequence, the encoding sub-sequence corresponding to the phoneme includes: detecting whether there is the phoneme corresponding to each of time points; and obtaining the encoding sub-sequence corresponding to the phoneme by setting an encoding value at a time point with the phoneme to a first value and setting an encoding value at a time point without the phoneme to a second value.

**[0009]** In combination with any embodiment provided by the present disclosure, the method further includes: for the encoding sub-sequence corresponding to each of the multiple phonemes, performing a Gaussian convolution operation on successive values of the phoneme in time by using a Gaussian filter.

**[0010]** In combination with any embodiment provided by the present disclosure, controlling the posture of the interactive object based on the values of the control parameters includes: obtaining a posture control vector sequence corresponding to the second encoding sequence; and controlling a posture of the interactive object based on the posture control vector sequence.

**[0011]** In combination with any embodiment provided by the present disclosure, the method further includes: in response to determining that a time interval between the phonemes in the phoneme sequence is greater than a preset threshold, controlling the posture of the interactive object based on values of preset control parameters of the at least one local region.

**[0012]** In combination with any embodiment provided by the present disclosure, obtaining the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding includes: obtaining the posture control vector of the

at least one local region of the interactive object corresponding to the feature encoding by inputting the feature encoding to a pre-trained recurrent neural network.

**[0013]** In combination with any embodiment provided by the present disclosure, the recurrent neural network is obtained by training with a feature encoding sample; wherein the method further includes: obtaining a video in which a character utters a voice, and obtaining a plurality of first image frames containing the character based on the video; extracting a piece of voice from the video, and obtaining a phoneme sequence sample based on the piece of voice, and performing feature encoding on the phoneme sequence sample; obtaining a feature encoding of at least one phoneme corresponding to the plurality of first image frames; converting the plurality of first image frames into a plurality of second image frames containing the interactive object, and obtaining a value of a posture control vector of at least one local region corresponding to the plurality of second image frames; and labeling the feature encoding corresponding to the plurality of first image frames based on the value of the posture control vector to obtain the feature encoding sample.

**[0014]** In combination with any embodiment provided by the present disclosure, the method further includes: training an initial recurrent neural network based on the feature encoding sample, to obtain the recurrent neural network when a change of network loss satisfies a convergence condition, wherein the network loss includes a difference between a value of the posture control vector of the at least one local region predicted by the recurrent neural network and the labeled value of the posture control vector.

**[0015]** According to an aspect of the present disclosure, provided is an apparatus for driving an interactive object , which includes: a first obtaining unit configured to obtain a phoneme sequence corresponding to text data; a second obtaining unit configured to obtain values of control parameters for at least one local region of the interactive object matching the phoneme sequence; and a driving unit configured to control a posture of the interactive object based on the values of the control parameters.

**[0016]** According to an aspect of the present disclosure, provided is an electronic device, which includes a processor and a memory for storing computer instructions executable by the processor, wherein the processor is configured to execute the computer instructions to implement the method according to any one of the above.

**[0017]** According to an aspect of the present disclosure, provided is a computer-readable storage medium storing a computer program, when being executed by a processor, the

computer program causes the processor to implement the method according to any one of the above.

**[0018]** The method and apparatus for driving an interactive object, and devices, and computer-readable storage medium of one or more embodiments of the present disclosure may cause an interactive object to make a posture that matches the phoneme corresponding to the text data by obtaining a phoneme sequence corresponding to the text data and obtaining the values of the control parameter of at least one local region of the interactive object that matches the phoneme sequence to control the posture of the interactive object. The posture includes a facial posture and a body posture, thereby causing the target object to produce the feeling that the interactive object is speaking the text content, and the interactive experience between the target object and the interactive object is improved.

### **BRIEF DESCRIPTION OF THE DRAWINGS**

**[0019]** In order to more clearly illustrate one or more embodiments of the present specification, the following is a brief description of the drawings required for the embodiments, and it will be obvious that the drawings in the following description are only some of the embodiments recorded in the present specification, and that other drawings may be obtained on the basis of these drawings without inventive work to those skilled in the art.

**[0020]** FIG. 1 illustrates a schematic diagram of a display device in a method for driving an interactive object in accordance with at least one embodiment of the present disclosure.

**[0021]** FIG. 2 illustrates a schematic flowchart of a method for driving an interactive object in accordance with at least one embodiment of the present disclosure.

**[0022]** FIG. 3 illustrates a schematic diagram of a process for feature encoding of a phoneme sequence in accordance with at least one embodiment of the present disclosure.

**[0023]** FIG. 4 illustrates a schematic structural diagram of an apparatus for driving an interactive object in accordance with at least one embodiment of the present disclosure.

**[0024]** FIG. 5 illustrates a schematic structural diagram of an electronic device in accordance with at least one embodiment of the present disclosure.

### **DETAILED DESCRIPTION OF THE EMBODIMENTS**

**[0025]** Examples will be described in detail herein, with the illustrations thereof represented in the drawings. When the following descriptions involve the drawings, like numerals in different drawings refer to like or similar elements unless otherwise indicated. The embodiments described in the following examples do not represent all embodiments consistent with the present disclosure. Rather, they are merely examples of apparatuses and

methods consistent with some aspects of the present disclosure as detailed in the appended claims.

**[0026]** The term “and/or” herein is merely an association relationship for describing associated objects, and indicates that there may be three relationships, for example, A and/or B may indicate that there are three cases: A alone, both A and B, and B alone. In addition, the term “at least one” herein means any one of a plurality or any combination of at least two of a plurality, for example, including at least one of A, B and C, and may mean including any one or more elements selected from the set of A, B, and C.

**[0027]** At least one embodiment of the present disclosure provides a method for driving an interactive object, the method may be performed by an electronic device such as a terminal device or a server, the terminal device may be a fixed terminal or a mobile terminal, such as a cell phone, a tablet computer, a game console, a desktop computer, an advertising machine, a vehicle terminal, and the like, the server includes a local server or a cloud server, and the method may also be implemented by a processor executing computer readable instructions stored in a memory.

**[0028]** In embodiments of the present disclosure, the interactive object may be any kind of virtual images capable of interacting with the target object. In an embodiment, the interactive object may be a virtual character, and may also be a virtual animal, a virtual object, a cartoon image, and other virtual images capable of implementing interaction functions. The presentation of the interactive object may be in 2D or 3D form, without being limited by the present disclosure. The target object may be a user, a robot, or other intelligent device. The interaction between the interactive object and the target object may be in an active or passive interaction mode. In an example, the target object may send a demand by making a posture or body movement to trigger the interaction with the interactive object by means of active interaction. In another example, the interactive object may actively greet the target object, prompt the target object to make a movement, etc., so that the target object interacts with the interactive object in a passive manner.

**[0029]** The interactive object may be displayed by a terminal device, the terminal device may be a television set, an all-in-one machine with a display function, a projector, a Virtual Reality (VR) device, an Augmented Reality (AR) device, etc., without being limited by the present disclosure.

**[0030]** FIG. 1 illustrates a display device in accordance with at least one embodiment of the present disclosure. As shown in FIG. 1, the display device has a transparent display on

which a 3D picture may be displayed to present a virtual scene with a 3D effect as well as an interactive object. For example, the interactive object displayed on the transparent display in FIG. 1 include a virtual cartoon character. In some embodiments, the terminal device described in the present disclosure may also be a display device having a transparent display as described above, the display device being configured with a memory and a processor, the memory is used to store computer instructions that may be executed on the processor, the processor is used to implement the method for driving the interactive object provided in the present disclosure when executing the computer instructions, to drive the interactive object displayed in the transparent display to communicate with or respond to the target object.

**[0031]** In some embodiments, in response to voice drive data used to drive the interactive object to output a voice, the interactive object may utter a specified voice to the target object. The terminal device may generate sound-driven data based on the movements, expressions, identities, preferences, etc. of the target object around the terminal device to drive the interactive object to respond by uttering the specified voice, thereby providing anthropomorphic services to the target object. It should be noted that the sound-driven data may also be generated by other means, e.g., generated by a server and sent to the terminal device.

**[0032]** During the interaction between the interactive object and the target object, when driving the interactive object to utter the specified voice based on the sound-driven data, it may not be possible to drive the interactive object to make facial movement synchronized with the specified voice, thereby causing the interactive object dull and unnatural when uttering the voice, which affects the interactive experience between the target object and the interactive object. Based on this, at least one embodiment of the present disclosure proposes a method for driving an interactive object to improve the experience of the target object interacting with the interactive object.

**[0033]** FIG. 2 illustrates a schematic flowchart of a method for driving an interactive object in accordance with at least one embodiment of the present disclosure, and as shown in FIG. 2, the method includes steps 201 to 203.

**[0034]** At step 201, a phoneme sequence corresponding to text data is obtained.

**[0035]** The text data may be drive data for driving the interactive object. The drive data may be drive data generated by the server or terminal device based on the movements, expressions, identities, preferences, etc. of the target object with which the interactive object

is interacted, or may be drive data called by the terminal device from the internal memory. The present disclosure does not limit the manner for obtaining the text data.

**[0036]** In embodiments of the present disclosure, based on morphemes contained in the text, the phonemes corresponding to the morphemes may be obtained, so as to obtain the phoneme sequence corresponding to the text. Wherein a phoneme is the smallest phonetic unit divided according to the natural properties of speech, and a single articulatory action of a real person is capable of forming a phoneme.

**[0037]** In an embodiment, in response to the text being Chinese text, the phoneme sequence may be generated by converting characters of the Chinese text into pinyin, using pinyin, and generating a timestamp for each phoneme.

**[0038]** At step 202, values of control parameters for at least one local region of the interactive object matching the phoneme sequence are obtained.

**[0039]** The local region is obtained by dividing the interactive object as a whole, including the face and/or the body. The control of one or more local regions of the face may correspond to a series of facial expressions or movements of the interactive object, for example, the control of a eye region may correspond to the facial movements such as opening eyes, closing eyes, blinking eyes, perspective shifting, etc. of the interactive object; and for example, the control of a mouth region may correspond to the facial movements such as closing mouth, opening mouth with different degrees, etc. of the interactive object. The control of one or more local regions of the body may correspond to a series of body movements of the interactive object, for example, the control of a leg region may correspond to the movements of the interactive object such as walking, jumping, kicking, etc.

**[0040]** The control parameters of the local region of the interactive object include a posture control vector of the local region. The posture control vector for each local region is used to drive the local region of the interactive object to perform movements. Different values of the posture control vector correspond to different movements or different magnitudes of a movement. For example, one set of values of the posture control vector for the mouth region may cause the mouth of the interactive object to open slightly, while another set of values of the posture control vector may cause the mouth of the interactive object to open widely. By driving the interactive object with different values of the pose control vector, the corresponding local region may be caused to make different movements or different magnitudes of a movement.

**[0041]** The local region may be selected according to movements of the interactive object to be controlled. For example, when the face as well as the body of the interactive object are required to be controlled to make the movements at the same time, the values of the posture control vectors of all the local regions may be obtained; when the expression of the interactive object is required to be controlled, the values of the posture control vector of the local region corresponding to the face may be obtained.

**[0042]** In embodiments of the present disclosure, values of the control parameters corresponding to the feature encoding may be determined by performing feature encoding on the phoneme sequence, such that values of the control parameters corresponding to the phoneme sequence are determined. Different encoding manners may reflect different features of the phoneme sequence. The specific encoding manner is not limited by the present disclosure.

**[0043]** In embodiments of the present disclosure, the correspondence between the feature encodings of the phoneme sequence corresponding to the text data and the values of the control parameters of the interactive object may be established in advance, so that the corresponding values of the control parameters may be obtained through the text data. The specific method for obtaining the values the control parameters matching the feature encodings of the phoneme sequence of the text data is described in detail later.

**[0044]** At step 203, a posture of the interactive object is controlled based on the values of the control parameters.

**[0045]** Wherein the values of the control parameters, such as the value of the posture control vector, are matched with the phoneme sequence contained in the text data. For example, when controlling the display device displaying the interactive object to display text based on the text data, and/or controlling the display device to output a voice based on the phoneme sequence corresponding to the text data, the posture made by the interactive object is synchronized with the outputted voice and/or the displayed text, such that the target object feels that the interactive object is speaking.

**[0046]** In embodiments of the present disclosure, the posture of the interactive object is controlled by obtaining the phoneme sequence corresponding to the text data and obtaining the values of the control parameters of at least one local region of the interactive object matching the phoneme sequence, such that the interactive object may make a posture matching the phonemes corresponding to the text data, the posture includes a facial posture

and a body posture, such that the target object feels that the interactive object is saying the text content, and the interactive experience of the target object is improved.

**[0047]** In some embodiments, the method is applied to a server, which includes a local server, a cloud server, etc., wherein the server processes the text data, generates the values of the control parameters for the interactive object, and renders the interactive object by using a 3D rendering engine based on the values of the control parameters to obtain an animation of the interactive object. The server may send the animation to the terminal for display to communicate with or respond to the target object, and may also send the animation to the cloud side, so that the terminal may obtain the animation from the cloud side to communicate with or respond to the target object. After the server generates the values of the control parameters for the interactive object, the values of the control parameters may also be sent to the terminal to enable the terminal to complete the process of rendering, generating the animation, and performing the display.

**[0048]** In some embodiments, the method is applied to a terminal, the terminal processes for text data, generates the values of the control parameters for the interactive object, and renders the interactive object by using a 3D rendering engine based on the values of the control parameters to obtain an animation of the interactive object. The terminal may display the animation to communicate with or respond to the target object.

**[0049]** In some embodiments, a display device displaying the interactive object may be controlled to display text based on the text data, and/or the display device may be controlled to output speech based on a phoneme sequence corresponding to the text data.

**[0050]** In embodiments of the present disclosure, since the values of the control parameters match the phoneme sequence of the text data, the outputted voice and/or text according to the text data, and the posture of the interactive object according to the values of the control parameters are synchronized in cases where the posture made by the interactive object is synchronized with the outputted voice and/or the displayed text, such that the target object feels that the interactive object is speaking.

**[0051]** In some embodiments, the control parameters of at least one local region of the interactive object include a posture control vector, which may be obtained in the following method.

**[0052]** First, an encoding sequence corresponding to the phoneme sequence is obtained by performing feature encoding on the phoneme sequence. In order to distinguish the encoding sequence from the subsequently mentioned encoding sequence, the encoding sequence

corresponding to the phoneme sequence of the text data is referred to as the first encoding sequence, i.e., the first encoding sequence is obtained by performing feature encoding on the phoneme sequence.

**[0053]** For multiple phonemes contained in the phoneme sequence, an encoding sub-sequence corresponding to each of the multiple phonemes is generated.

**[0054]** In an example, it is detected whether there is a first phoneme at various time points, and the first phoneme is any one of the multiple phonemes; an encoding value at a time point with the first phoneme is set to a first value, and an encoding value at a time point without the first phoneme is set to a second value. The encoding sub-sequence corresponding to the first phoneme is obtained after assigning the encoding value at various time points. For example, an encoding value at a time point with the first phoneme may be set to 1, and an encoding value at a time point without the first phoneme may be set to 0. That is, for each of the multiple phonemes in the phoneme sequence, it is detected whether there is a phoneme at various time points; an encoding value at a time point with the first phoneme is set to the first value, and an encoding value at a time point without the first phoneme is set to the second value, and the encoding sequence according to the phoneme is obtained after assigning values to the encoding values at various time points. It will be understood for those skilled in the art that the above setting of the encoding values is only an example, and the encoding values may also be set to other values, but the present disclosure is not limited thereto.

**[0055]** The first encoding sequence corresponding to the phoneme sequence is obtained based on the encoding sub-sequences respectively corresponding to the multiple phonemes.

**[0056]** In an example, for the encoding sub-sequence corresponding to the first phoneme, a Gaussian filter may be used to perform a Gaussian convolution operation on successive values of the first phoneme in time, to filter a matrix corresponding to the feature encoding, and to smooth the movement transition of the mouth region during phoneme conversion.

**[0057]** FIG. 3 illustrates a schematic flowchart of a method for driving an interactive object in accordance with at least one embodiment of the present disclosure. As shown in FIG. 3, a phoneme sequence 310 includes phonemes j, i1, j and ie4 (for brevity, only some of the phonemes are shown), and for each of the phonemes j, i1 and ie4, encoding sub-sequences 321, 322 and 323 respectively corresponding to the above phonemes are obtained. In each of the encoding sub-sequences, an encoding value at a time point with the phoneme is a first value (e.g., 1), and an encoding value at a time point without the phoneme is a second value

(e.g., 0). Illustrated by the example of encoding sub-sequence 321, a value of the encoding sub-sequence 321 at a time point with the phoneme *j* in the phoneme sequence 310 is the first value 1, and a value of encoding sub-sequence 321 at a time point without the phoneme *j* is the second value 0. All encoding sub-sequences form a first encoding sequence 320.

**[0058]** Then, a feature encoding corresponding to at least one phoneme is obtained based on the first encoding sequence.

**[0059]** Based on the encoding values of the encoding sub-sequences 321, 322 and 323 respectively corresponding to the phonemes *j*, *i1* and *ie4*, and the durations of the corresponding phonemes in the three encoding sub-sequences, i.e., the duration of the phoneme *j* in the encoding sub-sequence 321, the duration of the phoneme *i1* in the encoding sub-sequence 322, and the duration of the phoneme *ie4* in the encoding sub-sequence 323, feature information of the encoding sub-sequences 321, 322 and 323 is obtained.

**[0060]** In an example, a Gaussian filter may be used to respectively perform a Gaussian convolution operation on successive values in time of the phonemes *j*, *i1* and *ie4* in the encoding sub-sequences 321, 322 and 323, to smooth the feature encodings to obtain a smoothed first encoding sequence 330. That is, the Gaussian filter is used to perform the Gaussian convolution operation on successive values in time of the phonemes, so that stages of change of the encoding values from the second value to the first value or from the first value to the second value in the respective encoding sequence become smooth. For example, the values of the encoding sequence besides 0 and 1 also present intermediate state values, such as 0.2, 0.3, etc., and the posture control vectors obtained based on these intermediate state values make the movement transitions and expression changes of the interactive character more smooth and more natural, and the interactive experience of a target object is improved.

**[0061]** In some embodiments, a feature encoding corresponding to at least one phoneme may be obtained by performing a sliding window operation on the first encoding sequence, wherein the first encoding sequence may be an encoding sequence that has undergone a Gaussian convolution operation.

**[0062]** A sliding window operation is performed on the encoding sequence with a preset length of time window and a preset step size, and a feature encoding in the time window is used as a feature encoding of the corresponding at least one phoneme, after the sliding window operation is completed, a second encoding sequence is obtained based on multiple feature encodings. Since the durations of various phonemes are different and ratios of the

durations to the lengths of the time windows of various phonemes are different, the number of phonemes corresponding to the feature encodings within the time window may be 1, 2 or even more depending on the location of the time window. As shown in FIG. 3, a time window with a set length is slid over the first encoding sequence 320 or the smoothed first encoding sequence 330, to obtain feature encoding 1, feature encoding 2, feature encoding 3, and so on, respectively, i.e., the feature encodings 1, 2, 3, ..., and M are obtained after traversing the first encoding sequence, so as to obtain a second encoding sequence 340. Where M is a positive integer, and a value thereof is determined based on a length of the first encoding sequence, a length of the time window and a sliding step size of the time window.

**[0063]** Finally, a posture control vector of at least one local region of the interactive object corresponding to the feature encoding is obtained.

**[0064]** Based on the feature encoding 1, 2, 3, ..., and M, the corresponding posture control vectors 1, 2, 3, ..., and M are obtained, respectively, such that a posture control vector sequence 350 is obtained.

**[0065]** The posture control vector sequence 350 and the second encoding sequence 340 are aligned in time. Since each feature encoding in the second encoding sequence is obtained based on at least one phoneme in the phoneme sequence, each feature vector in the posture control vector sequence 350 is also obtained based on at least one phoneme in the phoneme sequence. While the phoneme sequence corresponding to the text data is played, the interactive object is driven to make a movement according to the posture control vector sequence, i.e., the interactive object can be driven to make voices according to the text content and make a movement synchronized with the voices at the same time, such that the target object feels that the interactive object is speaking, and the interactive experience of the target object is improved.

**[0066]** Assuming that the feature encoding starts to be output at a set time of a first time window, a value of the posture control vector before the set time may be set to a default value, that is, when the phoneme sequence is just started to be played, the interactive object is driven to make a default movement, and after the set time, the interactive object is driven to make a movement by using the posture control vector sequence obtained according to the first encoding sequence. Illustrated by the example of FIG. 3, the feature coding 1 starts to be output at time  $t_0$ , the default posture control vector corresponds to the time before time  $t_0$ .

**[0067]** The length of the time window is associated with the amount of information contained in the feature encoding. In the case where the amount of information contained in the time window is relatively large, a more uniform result can be outputted by the processing of the recurrent neural network. If the length of the time window is too large, it may result in that the expression of the interactive object when speaking does not correspond to a part of the text; if the length of the time window is too small, it may result in that the expression of the interactive object when speaking appears stiff. Therefore, the length of the time window needs to be determined according to the minimum duration of the phonemes corresponding to the text data, the movements made by driving the interactive object are more relevant to the voice.

**[0068]** A sliding step size of the time window is associated with the time interval (frequency) of obtaining the posture control vector, i.e., with the frequency of driving the interactive object to make the movement. The length and the step size of the time window may be set according to the actual interaction scenario, such that the expressions and movements made by the interactive object are more relevant to the voice and more vivid and natural.

**[0069]** In some embodiments, when the time interval between phonemes in the phoneme sequence is greater than a set threshold, the interactive object is driven to make a movement according to the set posture control vector of the local region. That is, when the interactive character pauses for a long time, the interactive object is driven to make a set movement. For example, when there is a long pause in the output voice, the interactive character may be driven to make a smiling expression or make a slight body sway to avoid the interactive character standing upright with no expression when there is a long pause, so that the interactive object speaks naturally and smoothly, and the interactive feeling of the target object is improved.

**[0070]** In some embodiments, the feature encoding may be inputted to a pre-trained recurrent neural network, such that the recurrent neural network outputs a posture control vector of at least one local region of the interactive object corresponding to the feature encoding based on the first encoding sequence. Since the recurrent neural network is a time-recursive neural network, the recurrent neural network may learn the historical information of the input feature encoding and output the posture control vector of the at least one local region based on the feature encoding sequence. Wherein the feature encoding sequence

includes a first encoding sequence and a second encoding sequence. The recurrent neural network may be, for example, a Long Short-Term Memory (LSTM) network.

**[0071]** In embodiments of the present disclosure, the pre-trained recurrent neural network is used to obtain the posture control vector of at least one local region of the interactive object corresponding to the feature encoding, and the historical feature information of the feature encoding is integrated with the current feature information, so that the historical posture control vector influences the change of the current posture control vector, thereby the expression change and the body movement of the interactive object are more smooth and natural.

**[0072]** In some embodiments, the recurrent neural network may be trained by the following method.

**[0073]** First, a feature encoding sample is obtained, the feature encoding sample is labeled with an actual value, the actual value is a value of a posture control vector of at least one local region of the interactive object.

**[0074]** After obtaining the feature encoding sample, an initial recurrent neural network is trained based on the feature encoding sample, and the trained recurrent neural network is obtained after the change in network loss satisfies a convergence condition, wherein the network loss includes a difference between the value of the posture control vector of at least one local region predicted by the recurrent neural network and the actual value.

**[0075]** In some embodiments, the feature encoding sample may be obtained by the following method.

**[0076]** First, a video in which a character utters a voice is obtained, and a plurality of first image frames containing the character is obtained based on the video. For example, a video in which a real person is speaking may be obtained.

**[0077]** Then, a piece of voice is extracted from the video, and a phoneme sequence sample is obtained based on the piece of voice, and feature encoding of the phoneme sequence sample is performed. Wherein the phoneme sequence sample is encoded in the same manner as the phoneme sequence corresponding to the text data described above.

**[0078]** A feature encoding of at least one phoneme corresponding to the first image frame is obtained based on an encoding sequence sample obtained by performing feature encoding on the phoneme sequence sample. Wherein the at least one phoneme may be a phoneme within a preset range of the occurrence time of the first image frame.

**[0079]** Then, the first image frame is converted into a second image frame containing the interactive object, and a value of a posture control vector of at least one local region corresponding to the second image frame is obtained. Wherein the value of the posture control vector may include values of the posture control vectors of all local regions, or may include values of the posture control vectors of some local region.

**[0080]** Illustrated by the example of the first image frame being an image frame containing a real person, the image frame containing the real person may be converted into a second image frame containing an image represented by the interactive object, and the posture control vector of each local region of the real person corresponds to the posture control vector of each local region of the interactive object, so that the posture control vectors of various local regions of the interactive object in the second image frame may be obtained.

**[0081]** Finally, the feature encoding of at least one phoneme corresponding to the first image frame obtained above is labeled based on the value of the posture control vector, so as to obtain a feature encoding sample.

**[0082]** In embodiments of the present disclosure, the posture control vector corresponding to the feature encoding of the phoneme is obtained by splitting the video of the character into multiple first image frames and pieces of voice, and by converting the first image frame containing the real person into the second image frame containing the interactive object, such that the correspondence between the feature encoding and the posture control vector is better, thereby a high-quality feature encoding sample is obtained, and the movement of the interactive object is closer to the actual movement of the corresponding person.

**[0083]** FIG. 4 illustrates a schematic structural diagram of an apparatus for driving an interactive object in accordance with at least one embodiment of the present disclosure. As shown in FIG. 4, the apparatus includes: a first obtaining unit 401 configured to obtain a phoneme sequence corresponding to text data; a second obtaining unit 402 configured to obtain values of control parameters for at least one local region of the interactive object matching the phoneme sequence; and a driving unit 403 configured to control a posture of the interactive object based on the values of the control parameters.

**[0084]** In some embodiments, the apparatus further includes an outputting unit configured to at least one of the following: control a display device displaying the interactive object to display text based on the text data; or control the display device to output a voice based on the phoneme sequence corresponding to the text data.

**[0085]** In some embodiments, the second obtaining unit is configured to: obtain a first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence; obtain a feature encoding corresponding to at least one phoneme based on the first encoding sequence; and obtain a posture control vector of at least one local region of the interactive object corresponding to the feature encoding.

**[0086]** In some embodiments, the second obtaining unit, when obtaining the first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence, is configured to: generating, for each of multiple phonemes contained in the phoneme sequence, an encoding sub-sequence corresponding to the phoneme; and obtaining a first encoding sequence corresponding to the phoneme sequence based on the encoding sub-sequence corresponding to each of the multiple phonemes.

**[0087]** In some embodiments, the second obtaining unit, when generating, for each of the multiple phonemes contained in the phoneme sequence, the encoding sub-sequence corresponding to the phoneme, is configured to: detect whether there is the phoneme corresponding to each of time points; and obtain the encoding sub-sequence corresponding to the phoneme by setting an encoding value at a time point with the phoneme to a first value and setting an encoding value at a time point without the phoneme to a second value.

**[0088]** In some embodiments, the apparatus further includes a filtering unit configured to: for the encoding sub-sequence corresponding to each of the multiple phonemes, perform a Gaussian convolution operation on successive values of the phoneme in time by using a Gaussian filter. In an embodiment, for the encoding sub-sequence corresponding to the first phoneme, a Gaussian convolution operation on successive values of the first phoneme in time is performed by using a Gaussian filter.

**[0089]** In some embodiments, the second obtaining unit, when obtaining the feature encoding corresponding to the at least one phoneme based on the first encoding sequence, is configured to: perform a sliding window operation on the first encoding sequence with a preset length of time window and a preset step size, take a feature encoding in the time window as the feature encoding corresponding to the at least one phoneme, and obtain a second encoding sequence based on multiple feature encodings obtained by completing the sliding window operation.

**[0090]** In some embodiments, the driving unit is configured to obtain a posture control vector sequence corresponding to the second encoding sequence; and control a posture of the interactive object based on the posture control vector sequence.

**[0091]** In some embodiments, the apparatus further includes a pause driving unit configured to, in response to determining that a time interval between the phonemes in the phoneme sequence is greater than a preset threshold, control the posture of the interactive object based on values of preset control parameters of the at least one local region.

**[0092]** In some embodiments, the second obtaining unit, when obtaining the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding, is configured to: obtain the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding by inputting the feature encoding to a pre-trained recurrent neural network.

**[0093]** In some embodiments, the recurrent neural network is obtained by training with a feature encoding sample; the apparatus further includes a sample obtaining unit configured to: obtain a video in which a character utters a voice, and obtain a plurality of first image frames containing the character based on the video; extract a piece of voice from the video, and obtain a phoneme sequence sample based on the piece of voice, and perform feature encoding on the phoneme sequence sample; obtain a feature encoding of at least one phoneme corresponding to the plurality of first image frames; convert the plurality of first image frames into a plurality of second image frames containing the interactive object, and obtain a value of a posture control vector of at least one local region corresponding to the plurality of second image frames; and label the feature encoding corresponding to the plurality of first image frames based on the value of the posture control vector to obtain the feature encoding sample.

**[0094]** In some embodiments, the apparatus further includes a training unit configured to train an initial recurrent neural network based on the feature encoding sample, to obtain the recurrent neural network when a change of network loss satisfies a convergence condition, wherein the network loss includes a difference between a value of the posture control vector of the at least one local region predicted by the recurrent neural network and the labeled value of the posture control vector.

**[0095]** At least one embodiment of the present disclosure provides an electronic device, as shown in FIG. 5, the electronic device includes a processor and a memory for storing computer instructions executable by the processor, wherein the processor is configured to execute the computer instructions to implement the method according to any one embodiment of the present disclosure.

**[0096]** At least one embodiment of the present disclosure provides a computer-readable storage medium storing a computer program, when being executed by a processor, the computer program causes the processor to implement the method according to any one embodiment of the present disclosure.

**[0097]** It will be understood for those skilled in the art that one or more embodiments of the present specification may be provide as a method, a system, or a computer program product. Therefore, one or more embodiments of the present specification may take a form of full hardware embodiment, full software embodiment, or embodiment combining software and hardware. Further, one or more embodiments of the present specification may take a form of computer program product implemented on one or more computer-readable storage mediums (including but not limited to disk memory, CD-ROM, optical memory, etc.) including computer available program codes.

**[0098]** Each embodiment in the present specification is described in a progressive manner, and the same and similar parts between each embodiment may be referred to each other, and each embodiment focuses on the differences from other embodiments. In particular, for the data processing device embodiment, since it substantially corresponds to the method embodiment, the description thereof is relatively simple, and relevant parts can be referred to the description of the method embodiment.

**[0099]** Specific embodiments of the present specification are described above. Other embodiments are within the scope of the claims. In some cases, the acts or steps in the claims may be performed in a different order than in the embodiments and still achieve the desired result. In addition, the processes depicted in the drawings do not necessarily require the particular order or the successive order shown in order to achieve the desired results. In some embodiments, multitasking and parallel processing are also possible or may be advantageous.

**[0100]** The embodiments of the subject matter and functional operations described in the present specification may be implemented in the following: a digital electronic circuit, tangibly embodied computer software or firmware, computer hardware including the structure disclosed in the present specification and structural equivalent thereof, or a combination of one or more of them. The embodiments of the subject matter described in the present specification may be implemented as one or more computer programs, i.e., one or more modules of computer program instructions encoded on a tangible non-transitory program carrier to be executed by a data processing device or to control the operation of a

data processing device. Alternatively or additionally, the program instructions may be encoded on an artificially generated propagation signal, such as a machine-generated electrical, optical, or electromagnetic signal, which is generated to encode and transmit the information to a suitable receiver device for execution by the data processing device. The computer storage medium may be a machine-readable storage device, a machine-readable storage substrate, a random or serial access memory device, or a combination of one or more of them.

**[0101]** The processing and logic flow described in the present specification may be performed by one or more programmable computers executing one or more computer programs to perform corresponding functions by operating on input data and generating outputs. The processing and logic flow may also be performed by a dedicated logic circuit - such as FPGA (Field Programmable Gate Array) or ASIC (Application Specific Integrated Circuit) - and the device may also be implemented as a dedicated logic circuit.

**[0102]** A computer suitable for executing a computer program include, for example, a general-purpose and/or specialized microprocessor, or any other type of central processing unit. Typically, the central processing unit will receive instructions and data from read-only memory and/or random-access memory. The basic components of a computer include a central processing unit for implementing or executing instructions and one or more memory devices for storing instructions and data. Typically, the computer will also include one or more mass storage devices for storing data, such as a disk, a magnetic disk, or an optical disk, or the computer will be operable to couple to the mass storage device to receive data from or transmit data to it, or both. However, it is not necessary for a computer to have such a device. In addition, a computer can be embedded in another device, such as a cell phone, a personal digital assistant (PDA), a mobile audio or video player, a game console, a global positioning system (GPS) receiver, or a portable storage device such as a universal serial bus (USB) flash drive, just to name a few.

**[0103]** A computer-readable storage medium suitable for storing computer program instructions and data includes all forms of non-transitory memory, media, and memory devices, including, for example, a semiconductor memory device (e.g., EPROM, EEPROM, and a flash memory device), a disk (e.g., an internal hard disk or a removable disk), a magnetic disk, and a CD ROM and DVD-ROM disk. The processor and memory may be supplemented by or incorporated into a dedicated logic circuit.

**[0104]** While the specification includes many specific embodiment details, these should not be interpreted as limiting the scope of any invention or the scope of protection claimed but are primarily intended to describe the features of specific embodiments of a particular invention. Certain features described in multiple embodiments in the present specification may also be implemented in combination in a single embodiment. On the other hand, the various features described in a single embodiment may also be implemented separately in multiple embodiments or implemented in any suitable sub-combination. Further, although the features may function in some combinations as described above and even initially claimed protection, one or more features from the combination claimed protection may be removed from the combination in some cases, and the combination claimed protection may refer to a sub-combination or a variant of a sub-combination.

**[0105]** Similarly, although the operations are depicted in a particular order in the drawings, this should not be construed as requiring that the operations be performed in the particular order shown or sequentially, or that all of the illustrated operations be performed to achieve the desired result. In some embodiments, multitasking and parallel processing are also possible or may be advantageous. Furthermore, the separation of the various system modules and components in the above embodiments should not be understood to require such separation in all embodiments, and it should be understood that the program components and systems described may often be integrated together in a single software product or packaged into multiple software products.

**[0106]** Therefore, particular embodiments of the subject matter have been described. Other embodiments are within the scope of the claims. In some cases, the acts or steps in the claims may be performed in a different order than in the embodiments and still achieve the desired result. In addition, the processes depicted in the drawings do not necessarily require the particular order or successive order shown in order to achieve the desired results. In some embodiments, multitasking and parallel processing are also possible or may be advantageous.

**[0107]** The above are only some embodiments of the present disclosure and are not intended to limit the present disclosure, any modifications, equivalent substitutions, improvements, etc. made within the spirit and principle of the present disclosure shall be covered by the scope of the present disclosure.

## CLAIMS

1. A method for driving an interactive object, comprising:
  - obtaining a phoneme sequence corresponding to text data;
  - obtaining values of control parameters for at least one local region of the interactive object matching the phoneme sequence; and
  - controlling a posture of the interactive object based on the values of the control parameters.
2. The method according to claim 1, further comprising at least one of the following:
  - controlling a display device displaying the interactive object to display text based on the text data; or
  - controlling the display device to output a voice based on the phoneme sequence corresponding to the text data.
3. The method according to claim 1 or 2, wherein the control parameters for the at least one local region of the interactive object comprise a posture control vector of the at least one local region;
  - wherein obtaining values of control parameters for at least one local region of the interactive object matching the phoneme sequence comprises:
    - obtaining a first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence;
    - obtaining a feature encoding corresponding to at least one phoneme based on the first encoding sequence; and
    - obtaining a posture control vector of at least one local region of the interactive object corresponding to the feature encoding.
4. The method according to claim 3, wherein obtaining the first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence comprises:
  - generating, for each of multiple phonemes contained in the phoneme sequence, a encoding sub-sequence corresponding to the phoneme; and
  - obtaining a first encoding sequence corresponding to the phoneme sequence based on the encoding sub-sequence corresponding to each of the multiple phonemes.

5. The method according to claim 4, wherein generating, for each of the multiple phonemes contained in the phoneme sequence, the encoding sub-sequence corresponding to the phoneme comprises:

detecting whether there is the phoneme corresponding to each of time points; and

obtaining the encoding sub-sequence corresponding to the phoneme by setting an encoding value at a time point with the phoneme to a first value and setting an encoding value at a time point without the phoneme to a second value.

6. The method according to claim 5, further comprising:

for the encoding sub-sequence corresponding to each of the multiple phonemes, performing a Gaussian convolution operation on successive values of the phoneme in time by using a Gaussian filter.

7. The method according to any one of claims 3 to 6, wherein obtaining the feature encoding corresponding to the at least one phoneme based on the first encoding sequence comprises:

performing a sliding window operation on the first encoding sequence with a preset length of time window and a preset step size, taking a feature encoding in the time window as the feature encoding corresponding to the at least one phoneme, and obtaining a second encoding sequence based on multiple feature encodings obtained by completing the sliding window operation;

wherein controlling the posture of the interactive object based on the values of the control parameters comprises:

obtaining a posture control vector sequence corresponding to the second encoding sequence; and

controlling a posture of the interactive object based on the posture control vector sequence.

8. The method according to any one of claims 1 to 7, further comprising:

in response to determining that a time interval between the phonemes in the phoneme sequence is greater than a preset threshold, controlling the posture of the interactive object based on values of preset control parameters of the at least one local region.

9. The method according to claim 3, wherein obtaining the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding comprises:

obtaining the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding by inputting the feature encoding to a pre-trained recurrent neural network.

10. The method according to claim 9, wherein the recurrent neural network is obtained by training with a feature encoding sample;

wherein the method further comprising:

obtaining a video in which a character utters a voice, and obtaining a plurality of first image frames containing the character based on the video;

extracting a piece of voice from the video, and obtaining a phoneme sequence sample based on the piece of voice, and performing feature encoding on the phoneme sequence sample;

obtaining a feature encoding of at least one phoneme corresponding to the plurality of first image frames;

converting the plurality of first image frames into a plurality of second image frames containing the interactive object, and obtaining a value of a posture control vector of at least one local region corresponding to the plurality of second image frames; and

labeling the feature encoding corresponding to the plurality of first image frames based on the value of the posture control vector to obtain the feature encoding sample.

11. The method according to claim 10, further comprising:

training an initial recurrent neural network based on the feature encoding sample, to obtain the recurrent neural network when a change of network loss satisfies a convergence condition, wherein the network loss comprises a difference between a value of the posture control vector of the at least one local region predicted by the recurrent neural network and the labeled value of the posture control vector.

12. An apparatus for driving an interactive object, comprising:

a first obtaining unit configured to obtain a phoneme sequence corresponding to text data;

a second obtaining unit configured to obtain values of control parameters for at least one local region of the interactive object matching the phoneme sequence; and

a driving unit configured to control a posture of the interactive object based on the values of the control parameters.

13. The device according to claim 12, further comprising an outputting unit configured to at least one of the following:

control a display device displaying the interactive object to display text based on the text data; or

control the display device to output a voice based on the phoneme sequence corresponding to the text data.

14. The apparatus according to claim 12 or 13, wherein the second obtaining unit is configured to:

obtain a first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence;

obtain a feature encoding corresponding to at least one phoneme based on the first encoding sequence; and

obtain a posture control vector of at least one local region of the interactive object corresponding to the feature encoding;

wherein obtaining the first encoding sequence corresponding to the phoneme sequence by performing feature encoding on the phoneme sequence comprises:

generating, for each of multiple phonemes contained in the phoneme sequence, an encoding sub-sequence corresponding to the phoneme; and

obtaining a first encoding sequence corresponding to the phoneme sequence based on the encoding sub-sequence corresponding to each of the multiple phonemes.

15. The apparatus according to claim 14, wherein the second obtaining unit, when obtaining a feature encoding corresponding to at least one phoneme based on the first encoding sequence, is configured to:

perform a sliding window operation on the first encoding sequence with a preset length of time window and a preset step size, and take a feature encoding in the time window as the feature encoding corresponding to the at least one phoneme, and obtaining a second

encoding sequence based on multiple feature encodings obtained by completing the sliding window operation;

wherein the driving unit is configured to:

obtain a posture control vector sequence corresponding to the second encoding sequence; and

control a posture of the interactive object based on the posture control vector sequence.

16. The apparatus according to any one of claims 12 to 15, further comprising a pause driving unit configured to, in response to determining that a time interval between the phonemes in the phoneme sequence is greater than a preset threshold, control the posture of the interactive object based on values of preset control parameters of the at least one local region.

17. The apparatus according to claim 14, wherein the second obtaining unit, when obtaining the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding, is configured to:

obtain the posture control vector of the at least one local region of the interactive object corresponding to the feature encoding by inputting the feature encoding to a pre-trained recurrent neural network.

18. The apparatus according to claim 17, further comprising a sample obtaining unit configured to:

obtain a video in which a character utters a voice, and obtain a plurality of first image frames containing the character based on the video;

extract a piece of voice from the video, and obtain a phoneme sequence sample based on the piece of voice, and perform feature encoding on the phoneme sequence sample;

obtain a feature encoding of at least one phoneme corresponding to the plurality of first image frames;

convert the plurality of first image frames into a plurality of second image frames containing the interactive object, and obtain a value of a posture control vector of at least one local region corresponding to the plurality of second image frames; and

label the feature encoding corresponding to the plurality of first image frames based on the value of the posture control vector to obtain the feature encoding sample;

wherein the apparatus further comprising a training unit configured to train an initial recurrent neural network based on the feature encoding sample, to obtain the recurrent neural network when a change of network loss satisfies a convergence condition, wherein the network loss comprises a difference between a value of the posture control vector of the at least one local region predicted by the recurrent neural network and the labeled value of the posture control vector.

19. An electronic device, comprising a processor and a memory for storing computer instructions executable by the processor, wherein the processor is configured to execute the computer instructions to implement the method according to any one of claims 1 to 11.

20. A computer-readable storage medium storing a computer program, when being executed by a processor, the computer program causes the processor to implement the method according to any one of claims 1 to 11.