

METHOD AND APPARATUS FOR LIP-SYNC GENERATION OF DIGITAL HUMANS

This application claims priority to Chinese Patent Application No. 2023108555002, filed with China National Intellectual Property Administration on July 12, 2023 and entitled
5 "METHOD AND APPARATUS FOR LIP-SYNC GENERATION OF DIGITAL HUMANS", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

One or more embodiments of this specification relate to the field of virtual digital humans, and in particular, to a method and an apparatus for lip-sync generation of digital humans.

10 BACKGROUND

In recent years, virtual digital humans have found extensive applications in fields such as virtual reality, digital marketing, metaverse, healthcare, and education. A crucial metric for evaluating digital human performance is the naturalness of experience across various dimensions, including visual texture, voices, motions, and lip shapes. The lip shapes refer to
15 whether a digital human accurately performs a "mouth-closing" action in silent intervals and forms correct "mouth shapes" when articulating different phonetic text during speech delivery. Manually configuring lip shapes for each segment of audio delivered by a digital human is prohibitively costly and lacks scalability. Therefore, it is essential to achieve automated orchestration in the process of configuring lip shapes for digital humans.

20 SUMMARY

One or more embodiments of this specification describe a method and an apparatus for lip-sync generation of digital humans, to achieve automated configuration of lip shapes for digital humans.

According to a first aspect, a method for lip-sync generation of digital humans is provided,
25 including:

- obtaining text data and corresponding audio data;

- aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set;

determining, for a target text unit over a target time period, target mouth shape information corresponding to the target text unit based on a first dictionary, where the first dictionary includes mouth shape information corresponding to any text unit;

5 determining target amplitude information corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit; and

determining a target lip shape over the target time period based on the target mouth shape information and the target amplitude information.

In a possible implementation, the audio data is pre-recorded audio, and the text unit is a character; and the aligning the text data with the audio data in the temporal dimension to obtain
10 a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set includes:

generating, based on the text data and the audio data, a character set and an audio segment and a time period corresponding to any character in the character set using a forced alignment model.

15 In a possible implementation, the audio data is pre-recorded audio, and the text unit is a phoneme; and the aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set includes:

20 determining, based on any character in the text data, a phoneme corresponding to the character to obtain a phoneme set; and

generating an audio segment and a time period corresponding to any phoneme in the phoneme set based on the phoneme set and the audio data using a forced alignment model.

In a possible implementation, the audio data is generated using a text-to-speech model based on the text data, and the text-to-speech model further generates the text unit set
25 corresponding to the text data and the audio segment and the time period corresponding to any text unit in the text unit set while generating the audio data; and

the aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set includes:

30 obtaining the text unit set and the audio segment and the time period corresponding to any text unit in the text unit set from a generation result of the text-to-speech model.

In a possible implementation, the determining target amplitude information corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit includes:

determining the target amplitude information based on an audio feature of the target audio segment, where the audio feature includes at least one of a volume feature, a fundamental frequency feature, and a formant feature.

In a possible implementation, the audio feature is the volume feature; and the determining
5 the target amplitude information based on an audio feature of the target audio segment includes:

determining the target amplitude information based on a maximum volume value and a minimum volume value within a specific text range in which the target text unit is located.

In a possible implementation, the specific text range in which the target text unit is located is a sentence or paragraph in which the target text unit is located.

10 In a possible implementation, the method further includes:

forming a lip shape set based on a plurality of target lip shapes, where the lip shape set includes a first lip shape and a second lip shape adjacent in the temporal dimension; and

performing, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip
15 shape sequence.

In a possible implementation, the interpolation operation is linear interpolation.

In a possible implementation, the performing, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip shape sequence includes:

20 determining a lip shape corresponding to any smallest time unit in the blank time period based on the first lip shape and the second lip shape to obtain the first lip shape sequence.

In a possible implementation, the method further includes:

performing a smoothing operation on a target lip shape sequence formed by the first lip shape, the second lip shape, and the first lip shape sequence in the temporal dimension to obtain
25 a target smoothed lip shape sequence.

In a possible implementation, the smoothing operation is exponential smoothing or one-dimensional Gaussian filter smoothing.

According to a second aspect, an apparatus for lip-sync generation of digital humans is provided, including:

30 an obtaining unit, configured to obtain text data and corresponding audio data;

an alignment unit, configured to align the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set;

a mouth shape determination unit, configured to determine, for a target text unit over a

target time period, target mouth shape information corresponding to the target text unit based on a first dictionary, where the first dictionary includes mouth shape information corresponding to any text unit;

an amplitude determination unit, configured to determine target amplitude information
5 corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit; and

a lip shape determination unit, configured to determine a target lip shape over the target time period based on the target mouth shape information and the target amplitude information.

In a possible implementation, the apparatus further includes:

10 a set determination unit, configured to form a lip shape set based on a plurality of target lip shapes, where the lip shape set includes a first lip shape and a second lip shape adjacent in the temporal dimension; and

an interpolation unit, configured to perform, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the
15 second lip shape to obtain a first lip shape sequence.

In a possible implementation, the apparatus further includes:

a smoothing unit, configured to perform a smoothing operation on a target lip shape sequence formed by the first lip shape, the second lip shape, and the first lip shape sequence in the temporal dimension to obtain a target smoothed lip shape sequence.

20 According to a third aspect, a computer-readable storage medium is provided, having a computer program stored thereon. The computer program, when executed by a computer, causes the computer to perform the method in the first aspect.

According to a fourth aspect, a computing device is provided, including a memory and a processor. The memory has executable code stored thereon, and the processor, when executing
25 the executable code, implements the method in the first aspect.

The method and apparatus for lip-sync generation of digital humans provided in the embodiments of this specification achieve fine-grained control over mouth opening and closing and various lip shapes without using annotated audio-text alignment information for two different types of application scenarios including offline video production and real-time speech
30 delivery of digital humans.

BRIEF DESCRIPTION OF THE DRAWINGS

To describe the technical solutions in multiple embodiments disclosed in this specification

more clearly, the following briefly describes the accompanying drawings required for describing the embodiments. It is clear that the accompanying drawings in the following description show merely multiple embodiments disclosed in this specification, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

FIG. 1 is a schematic diagram of an implementation scenario of a method for lip-sync generation of digital humans according to an embodiment;

FIG. 2 is a flowchart of a method for lip-sync generation of digital humans according to an embodiment;

FIG. 3 is a schematic diagram of lip shape interpolation according to an embodiment;

FIG. 4 is a schematic diagram of lip shape interpolation according to another embodiment; and

FIG. 5 is a schematic block diagram of an apparatus for lip-sync generation of digital humans according to an embodiment.

DETAILED DESCRIPTION

The solutions provided in this specification are described below with reference to the accompanying drawings.

A digital human, also known as a virtual avatar, is a digital character that is created using digital technology and that closely resembles a human appearance. Lip-sync driving is a critical capability of digital humans, referring to the process where mouth animations of a digital human change appropriately with audio during speech delivery, achieving synchronization between the audio and the mouth animations. In digital human scenarios such as customer service speech delivery and intelligent assistants, text and audio are both constantly changing and cannot be enumerated, making it impossible to manually replicate all lip shapes. Consequently, it is necessary to use certain methods to automatically generate lip shape sequences in real time based on audio and text content.

FIG. 1 is a schematic diagram of an implementation scenario of a method for lip-sync generation of digital humans according to an embodiment. In an example of FIG. 1, input data includes text data and audio data corresponding to the text data. The audio data may be pre-recorded, or may be generated by a machine based on the text data using a text-to-speech (TTS) model. After the text data and the audio data are input to an alignment module, the alignment module first determines whether the audio data is pre-recorded or generated in real time based

on text using TTS. If the audio data is pre-recorded, a current application scenario is an offline video production scenario. The alignment module aligns text and audio using a forced alignment model, divides the text data into a plurality of text units, and obtains an audio segment and a time period corresponding to each text unit. If the audio data is generated using the TTS model, a current application scenario is an online real-time speech delivery scenario. The TTS model can output a time period corresponding to any text unit while generating audio. The alignment module only needs to obtain the time period corresponding to any text unit from a generation result of the TTS model. The time period may be represented based on a start timestamp and an end timestamp of the time period. After the time periods corresponding to the text units are obtained, the audio may be divided into corresponding audio segments based on the time periods, so that the audio segment and the time period corresponding to any text unit can be obtained.

After the audio segment and the time period corresponding to any text unit are obtained, corresponding mouth shape information is obtained using a pre-configured text-mouth shape dictionary based on the text unit. Amplitude information corresponding to a mouth shape is obtained based on a specific audio feature of the audio segment, and a corresponding lip shape over one time period may be obtained by integrating the mouth shape information and the amplitude information. A preliminary lip shape sequence may be obtained by sorting lip shapes over different time periods in time order.

Interpolation and smoothing operations are performed on a blank time period between adjacent lip shapes in the lip shape sequence to implement transition between different lip shapes, so that a global lip shape animation is smoother and more natural, thereby obtaining a final smoothed lip shape sequence. After the final smoothed lip shape sequence is obtained, the smoothed lip shape sequence is sent to a digital-human lip shape animation generation module (not shown in the figure), so that a correct and smooth natural lip shape animation corresponding to a digital human at each time point may be generated based on the lip shape sequence.

Specific implementation steps of the foregoing method for lip-sync generation are described below with reference to specific embodiments. FIG. 2 is a flowchart of a method for lip-sync generation of digital humans according to an embodiment. The method may be performed by any platform, server, device cluster, or the like having computing and processing capabilities.

Step 202: Obtain text data and corresponding audio data.

The audio data may be pre-recorded or may be generated by a machine in real time based

on the text data using a TTS model.

During the generation of the audio data using the TTS model based on the text data in real time, the text data is first divided into a plurality of phonemes. The phonemes are fundamental speech units divided based on natural phonetic properties, refer to a set of distinctive units of speech, and are the smallest distinguishable audio segments in a language's phonetic system. After the text data is divided into the plurality of phonemes, the TTS model predicts a time period corresponding to any phoneme in the entire audio, and generates corresponding audio data using digital signal processing and speech synthesis-related technologies. A time period of a phoneme may be represented by a start timestamp and an end timestamp. An audio segment corresponding to the phoneme in the audio data may be obtained based on the start timestamp and the end timestamp. In an embodiment, a frame count, a start frame, and an end frame of any phoneme are predicted using a duration predictor in a non-autoregressive TTS model. The duration of the phoneme can be obtained by multiplying the frame count by the duration of each frame.

When the audio data is pre-recorded, time periods corresponding to phonemes cannot be simultaneously obtained. For the pre-recorded audio, text may be manually read out, and then an audio acquisition device records the voice to obtain corresponding audio data. Alternatively, the pre-recorded audio may be generated using a constrained TTS model based on text data. During use of the constrained TTS model, a specific module related to duration in the TTS model cannot be used. For example, the TTS model is an interface provided by an upstream service provider, or the TTS model has been exported as an executable file. In this case, the time periods corresponding to the phonemes cannot be obtained based on the TTS model.

Step 204: Align the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set.

The text unit is the smallest unit of text that is manually set and that cannot be divided again in the embodiments. For example, in Chinese text, the text unit may be a single Chinese character or may be a Chinese phoneme. In English text, the text unit may be a word or may be an English phoneme.

In an embodiment, the audio data is pre-recorded audio, and the text unit is a character. The character refers to a linguistic unit that is in any language included in the text data and that is one level above a phoneme of the language. For example, the character is a single Chinese character in Chinese, and is a single English word in English. A character set and an audio segment and a time period corresponding to any character in the character set are generated

based on the text data and the audio data using a forced alignment model.

The forced alignment model may be any existing model, for example, may be a Montreal Forced Aligner (MFA), or a module related to forced alignment in a pre-trained acoustic model Wav2Vec2. This is not limited herein. The forced alignment model aligns phrases, characters, or phonemes in text with corresponding speech segments in audio by statistically analyzing co-occurrence and correlation between features of audio data and text sequences, thereby achieving more accurate and precise correspondence between text and speech.

After the text data and the audio data are input to the forced alignment model, a start timestamp and an end timestamp of any character in the text data are obtained, and an audio segment corresponding to the character may be segmented from the audio data based on the start timestamp and the end timestamp of any character.

In another embodiment, the audio data is pre-recorded audio, and the text unit is a phoneme.

First, a phoneme corresponding to the character is determined based on any character in the text data to obtain a phoneme set.

Based on text in different languages in the text data, characters are divided into phonemes using corresponding methods.

In a specific embodiment, when a text language is Chinese, because polyphonic characters exist in Chinese, for any character, it is first determined whether the character is a monophonic character or a polyphonic character using a polyphonic character set. Because the number of polyphonic characters in Chinese characters is limited, the polyphonic character set may be built in advance.

When the character is a monophonic character, a monophonic character is converted into a corresponding Hanyu pinyin syllable using a pinyin library, and then the pinyin syllable is divided into corresponding phonemes based on a preset division rule. For example, a monophonic character "你" (Pinyin: ni, literal meaning: "you") is first converted using the pinyin library into "ni", which is then divided according to a preset division rule into phonemes "n" and "i". The used pinyin library may be an existing open-source library, for example, pypinyin, or may be specially customized according to specific requirements. This is not limited herein.

It should be noted that because the number of Hanyu pinyin syllables is limited and the number of phonemes in Hanyu is also limited, a division rule may be formulated in advance. The correspondence between any Hanyu pinyin syllable and corresponding phonemes of the Hanyu pinyin syllable is established to form the division rule. In addition, because tones have

very limited influence on the lip shapes of digital humans, different tones are not distinguished in this process.

When the character is a polyphonic character, the character cannot be directly converted into a corresponding Hanyu pinyin syllable using the pinyin library. In this case, the pronunciation of the character is predicted based on context information of the character using a polyphonic character prediction model. For example, a pre-trained BERT-based text classification model is used. A sentence or paragraph in which the character is located is input to the text classification model to predict the pronunciation of the character. After the correct pronunciation of the polyphonic character is obtained, the polyphonic character is divided into corresponding phonemes using a processing solution similar to that for a monophonic character above.

In a specific embodiment, when a text language is English, English text is first divided into a set of English words. For any English word, the English word is divided into corresponding phonemes using a grapheme-to-phoneme model. The grapheme-to-phoneme model may be an existing model, for example, a CMU Pronouncing Dictionary, Arpabet, or the like, or may be specially customized according to specific requirements. This is not limited herein.

When text data contains another language, a character may also be divided into phonemes using an existing technology corresponding to the language. Details are not described here again.

After the phoneme set is obtained, an audio segment and a time period corresponding to any phoneme in the phoneme set are generated based on the phoneme set and the audio data using a forced alignment model. The case is similar to the solution in the foregoing embodiment in which the text unit is a character. Refer to partial descriptions in the foregoing embodiment. Details are not described here again.

In yet another embodiment, the audio data is generated in real time based on the text data using the TTS model. While generating the audio data in Step 202, the TTS model has generated the time period corresponding to any phoneme in the audio data and the audio segment corresponding to the time period. Therefore, it is only necessary to obtain corresponding data from a generation result of the TTS model. If the text unit is a character, a time period of the character may be obtained by concatenating time periods of a plurality of phonemes corresponding to any character. An audio segment corresponding to any character may be obtained by concatenating a plurality of audio segments corresponding to the character or by segmenting the audio data again based on time periods.

Step 206: Determine, for a target text unit over a target time period, target mouth shape

information corresponding to the target text unit based on a first dictionary, where the first dictionary includes mouth shape information corresponding to any text unit.

The first dictionary may be an existing text-mouth shape dictionary, or may be specially customized according to specific requirements. This is not limited herein. During customization, a person puts on a motion capture device, and the motion capture device captures and analyzes corresponding mouth shape information when the person reads out any text unit to establish the first dictionary. The first dictionary may be established based on blend shape, or may be established based on any other animation technology. This is not limited herein.

Step 208: Determine target amplitude information corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit.

Specifically, the target amplitude information is determined based on an audio feature of the target audio segment, where the audio feature includes at least one of a volume feature, a fundamental frequency feature, and a formant feature.

In an embodiment, the target amplitude information is determined by using the volume feature corresponding to the text unit separately. For any text unit, a maximum volume value X_{max} and a minimum volume value X_{min} within a specific text range in which the text unit is located are determined first. The specific text range may be a sentence or a paragraph in which the text unit is located or another preset range. In a long piece of audio, an audio feature of a segment may be significantly different from an audio feature of another segment. For example, for a volume feature, a segment has large overall volume, and another segment has small overall volume. Therefore, if only the maximum volume value X_{max} and the minimum volume value X_{min} of the entire audio are taken into consideration, a target amplitude may be excessively small in a part with small volume, resulting in an indistinct mouth motion in an eventually generated animation. The specific text range is set to determine maximum values and minimum values of audio features by segments, so that the foregoing problems can be avoided.

After X_{max} and X_{min} within the specific text range are determined, for a text unit with a volume value being X within the text range, amplitude information of the text unit may be represented by an amplitude value P . A method for calculating P is shown in Formula (1):

$$P = \alpha + \alpha \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

α is a hyperparameter configured to control a minimum value of the amplitude value P . In Formula (1), a value range of P is $[\alpha, 2\alpha]$. Preferably, the value range of α is 0.5 to 0.8.

In another embodiment, the target amplitude information corresponding to the target mouth shape information may alternatively be determined using the fundamental frequency feature, the formant feature, or any feature combination thereof with the volume feature with reference to the method in the foregoing embodiment.

5 Step 210: Determine a target lip shape over the target time period based on the target mouth shape information and the target amplitude information.

For the target text unit, after the target mouth shape information and the target amplitude information corresponding to the target text unit are determined, the target lip shape corresponding to the target text unit may be determined.

10 In an embodiment, the mouth shape information is encoded in a manner similar to one-hot encoding to obtain a one-dimensional vector. The vector has a value of 1 in one dimension and 0 in all other dimensions. The dimensionality of the vector is equal to the total number of mouth shapes. For example, a mouth shape A corresponds to [1, 0, 0, ..., 0], a mouth shape B corresponds to [0, 1, 0, ..., 0], and a mouth shape H corresponds to [0, 0, ..., 1, 0].

15 The vector corresponding to the mouth shape information is then multiplied by the target amplitude to obtain a vector corresponding to the target lip shape. For example, for a text unit, a target mouth shape is a mouth shape A, and a target amplitude is 0.8. A vector corresponding to a target lip shape of the text unit is [0.8, 0, 0, ..., 0]. For another text unit, a target mouth shape is a mouth shape B, and a target amplitude is 0.5. A vector corresponding to a target lip
20 shape of the text unit is [0, 0.5, 0, ..., 0].

For a vector corresponding to any lip shape, a value of the vector in any dimension represents an amplitude of the lip shape in a mouth shape corresponding to the dimension. For example, a vector [0, 0.5, 0, ..., 0] of a lip shape represents that an amplitude of a mouth shape B is 0.5, and amplitudes of all other mouth shapes are 0. A vector [0.2, 0.4, 0, ..., 0] of a lip
25 shape represents that an amplitude of a mouth shape A is 0.2, an amplitude of a mouth shape B is 0.4, and amplitudes of all other mouth shapes are 0.

In other embodiments, another manner of representing a lip shape may alternatively be designed, provided that different lip shapes can have different representations. For example, another encoding manner of mouth shape information is defined, and another manner of
30 combining mouth shape information with amplitude information is defined. For example, a two-dimensional vector is defined. The first row is an encoded vector of mouth shape information, and the second row is a value or encoded vector corresponding to amplitude information.

Step 202 to Step 210 may determine a lip shape corresponding to a single text unit in text.

A lip shape sequence corresponding to a piece of text may be obtained by sorting lip shapes corresponding to a plurality of text units in the piece of text in time order.

In some possible implementations, the method further includes the following steps.

Step 212: Form a lip shape set based on a plurality of target lip shapes, where the lip shape set includes a first lip shape and a second lip shape adjacent in the temporal dimension.

Step 214: Perform, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip shape sequence.

During speaking, a silent blank time exists between adjacent pronunciations. If the lip shape is directly set to a closed-mouth state within the blank time, the eventually generated lip shape animation seems extremely rigid and abrupt. That is, the mouth is opened suddenly when there is sound, and the mouth is closed suddenly within the blank time without sound. Therefore, a transitional lip shape needs to be set for the blank time between adjacent text units, making the eventually generated lip shape animation appear more natural.

Specifically, the first lip shape corresponds to a first text unit, the second lip shape corresponds to a second text unit, and the blank time is segmented into a plurality of smallest time units using a preset smallest time unit of time. The smallest time unit may be set according to actual requirements. For example, duration of one frame is used as the smallest time unit, or the smallest time unit is set to 0.05 seconds, 0.02 seconds, or the like. Interpolation is then performed on the blank time between the first lip shape and the second lip shape based on the first lip shape and the second lip shape to determine the lip shape corresponding to any smallest time unit within the blank time period to obtain the first lip shape sequence.

In an embodiment, the first text unit and the second text unit are located in the same sentence. It may be assumed that the first lip shape precedes the second lip shape in time order. The blank time is segmented into a plurality of smallest time units. Interpolation is performed on the blank time through linear interpolation based on the first lip shape and the second lip shape to obtain the lip shape corresponding to any smallest time unit, and the lip shapes are combined to obtain the first lip shape sequence.

For example, in a specific example, it is assumed that a total of 5 types of lip shapes are provided. The first lip shape is [0, 1, 0, 0, 0], and the second lip shape is [0, 0, 0, 0, 0.8]. The smallest time unit is the duration of a frame. A blank time of 3 frames exists between the first lip shape and the second lip shape, that is, [0, 0, 0, 0, 0]. The lip shape sequence before interpolation is:

[0, 1, 0, 0, 0],

[0, 0, 0, 0, 0],

[0, 0, 0, 0, 0],

[0, 0, 0, 0, 0],

[0, 0, 0, 0, 0.8].

5 Interpolation is performed on the blank time through linear interpolation based on the first lip shape and the second lip shape to obtain the interpolated first lip shape sequence.

[0, 1, 0, 0, 0],

[0, 0.75, 0, 0, 0.2],

[0, 0.5, 0, 0, 0.4],

10 [0, 0.25, 0, 0, 0.6],

[0, 0, 0, 0, 0.8].

A schematic diagram of the obtained first lip shape sequence is shown in FIG. 3.

In another embodiment, the first text unit and the second text unit are located in different sentences. The first text unit is located at the end of a former sentence, and the second text unit is located in the beginning of a latter sentence. In this case, because a long pause exists between the sentences, a closed-mouth lip shape should be set for a partial time within the pause. If a lip shape is configured for any smallest time unit of the blank time period with reference to the previous embodiment, the transition from the first lip shape to the second lip shape may become excessively slow, and no closed-mouth lip shape may exist.

20 In this case, the blank time between the first lip shape and the second lip shape is first divided into two segments from the middle. The front half segment belongs to the trailing blank time of the sentence in which the first text unit is located, and the rear half segment belongs to the beginning blank time of the sentence in which the second text unit is located. For the trailing blank time after the first text unit, interpolation is performed in forward-time order according to a fixed rate. For the second text unit located in the beginning of the sentence, interpolation is performed in reverse-time order according to the fixed rate. The rate is a preset value, representing an amount of a lip shape amplitude that decreases within each time unit. The forward-time order interpolation is sequentially decreasing lip shapes within time units following a lip shape in time order. The reverse-time order interpolation is sequentially decreasing lip shapes of time units preceding a lip shape in reverse time order.

For example, in a specific example, in an initial case, as shown in FIG. 4(a), a blank time exists between the first lip shape and the second lip shape. The blank time includes a plurality of time units. The blank time is divided into a trailing blank time and a beginning blank time from the middle. It is assumed that the first lip shape is [0, 0.6, 0, 0, 0], the trailing blank time

following the first lip shape has 4 frames, and the rate is 0.2/frame. The lip shape sequence after the forward-time order interpolation is:

[0, 0.6, 0, 0, 0],
[0, 0.4, 0, 0, 0],
5 [0, 0.2, 0, 0, 0],
[0, 0, 0, 0, 0],
[0, 0, 0, 0, 0].

It is assumed that the second lip shape is [0, 0.4, 0, 0, 0], and the beginning blank time preceding the second lip shape has 4 frames. The lip shape sequence following the reverse-time order interpolation is:

[0, 0, 0, 0, 0],
[0, 0, 0, 0, 0],
[0, 0, 0, 0, 0],
[0, 0.2, 0, 0, 0],
15 [0, 0.4, 0, 0, 0].

The lip shape sequences of the trailing blank time and the beginning blank time are concatenated in time order to obtain the first lip shape sequence. A schematic diagram of the obtained first lip shape sequence is shown in FIG. 4(b).

In other embodiments, another interpolation method, for example, quadratic interpolation, cubic interpolation, spline interpolation, or the like, may alternatively be used, provided that a transitional lip shape for a blank time can be reasonably set.

The transitional lip shape may be reasonably set for the blank time between the lip shapes using Step 212 to Step 214, so that the lip shape animation generated based on the overall lip shape sequence is more natural compared with that without a transitional lip shape.

25 In some possible implementations, the method further includes the following step.

Step 216: Perform a smoothing operation on a target lip shape sequence formed by the first lip shape, the second lip shape, and the first lip shape sequence in the temporal dimension to obtain a target smoothed lip shape sequence.

After the transitional lip shape is set through the interpolation operation, a change between lip shapes within adjacent time units may still be not natural enough, and a change trajectory of lip shapes is adjusted through the smoothing operation, making the lip shapes smoother, thereby avoiding abrupt changes.

Specifically, the target lip shape sequence is a two-dimensional vector. The vertical direction is the temporal dimension, and the horizontal direction is the amplitude dimension. A

value of the vector at any position represents an amplitude of a mouth shape corresponding to the dimension of the lip shape sequence at the time point. A smoothing operation is performed on a one-dimensional amplitude sequence formed in any amplitude dimension of the target lip shape sequence at a plurality of time points, and smoothing results of all amplitude dimensions are integrated to obtain the target smoothed lip shape sequence. The used smoothing operation is exponential smoothing, one-dimensional Gaussian filter smoothing or any other smoothing method.

For example, in a specific example, exponential smoothing is performed on an amplitude sequence [10, 12, 13, 15, 17, 20, 19, 18, 16, 14, 12, 11] of the second column of the target lip shape sequence in 12 consecutive time units to obtain a smoothed sequence [10.00, 10.30, 10.91, 12.13, 13.29, 15.01, 16.20, 16.94, 16.46, 15.12, 13.88, 12.82]. Smoothed sequences of other columns are calculated using similar methods. The sequences are then restored in an original order into a two-dimensional vector, so that the target smoothed lip shape sequence is obtained.

In one or more embodiments of this specification, it is only necessary to obtain timestamps after the alignment of audio and text, and different lip shape driving solutions are built for two types of digital human animation scenarios: real-time speech delivery and offline production. One or more embodiments of this specification do not involve the collection of supervised data, and therefore almost zero data costs are incurred. In addition, a single-instance inference time is short, a required storage space is small, and the solution has higher concurrency, so that deployment costs can be greatly reduced. Through the orchestration of lip shapes within a specific time interval, it is ensured that a correct lip shape appears at every time point, so that an error rate can be greatly reduced. In addition, one or more embodiments of this specification can be migrated to different digital humans based on matching of timestamps and lip shape rules at almost zero costs, and adapt to different lip shape solutions, so that constraints caused by data and specific virtual persons are avoided, and a long and costly data collection phase is skipped, thereby achieving very high compatibility.

According to another aspect of embodiments, an apparatus for lip-sync generation of digital humans is further provided. FIG. 5 is a schematic block diagram of an apparatus for lip-sync generation of digital humans according to an embodiment. The apparatus may be deployed in any device, platform or device cluster having computing and processing capabilities. As shown in FIG. 5, the apparatus 500 includes:

- an obtaining unit 501, configured to obtain text data and corresponding audio data;
- an alignment unit 502, configured to align the text data with the audio data in the temporal

dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set;

a mouth shape determination unit 503, configured to determine, for a target text unit over a target time period, target mouth shape information corresponding to the target text unit based on a first dictionary, where the first dictionary includes mouth shape information corresponding to any text unit;

an amplitude determination unit 504, configured to determine target amplitude information corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit; and

a lip shape determination unit 505, configured to determine a target lip shape over the target time period based on the target mouth shape information and the target amplitude information.

In some possible implementations, the apparatus 500 further includes:

a set determination unit 506, configured to form a lip shape set based on a plurality of target lip shapes, where the lip shape set includes a first lip shape and a second lip shape adjacent in the temporal dimension; and

an interpolation unit 507, configured to perform, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip shape sequence.

In some possible implementations, the apparatus 500 further includes:

a smoothing unit 508, configured to perform a smoothing operation on a target lip shape sequence formed by the first lip shape, the second lip shape, and the first lip shape sequence in the temporal dimension to obtain a target smoothed lip shape sequence.

According to another aspect of embodiments, a computer-readable storage medium is further provided, having a computer program stored thereon. The computer program, when executed by a computer, causes the computer to perform the method described in any foregoing embodiment.

According to yet another aspect of embodiments, a computing device is further provided, including a memory and a processor. The memory has executable code stored thereon. The processor, when executing the executable code, implements the method described in any foregoing embodiment.

The embodiments of this specification are all described in a progressive manner, for same or similar parts in the embodiments, refer to these embodiments, and a description of each embodiment focuses on a difference from other embodiments. Especially, an apparatus

embodiment is basically similar to a method embodiment, and therefore is described briefly; for related parts, refer to partial descriptions in the method embodiment.

Particular embodiments of this specification are described above. Other embodiments fall within the scope of the appended claims. In some cases, the actions or steps recorded in the claims may be performed in sequences different from those in the embodiments and an expected result may still be achieved. In addition, the processes depicted in the accompanying drawings are not necessarily performed in a specific order or a successive order to achieve an expected result. In some implementations, multitasking and parallel processing may be feasible or beneficial.

It should be noted that the relational terms herein such as first and second are used only to differentiate an entity or operation from another entity or operation, and do not require or imply any actual relationship or sequence between these entities or operations. Moreover, the terms "include", "comprise", and any variation thereof are intended to cover a non-exclusive inclusion. Therefore, in the context of a process, a method, an object, or an apparatus that includes a series of elements, the process, method, object, or apparatus not only includes such elements, but also includes other elements not specified expressly, or may include inherent elements of the process, method, object, or apparatus. Without more limitations, elements defined by the sentence "including one ..." do not exclude that there are still other same elements in the processes, methods, objects, or apparatuses.

A person of ordinary skill in the art may understand that all or a part of the steps of the foregoing embodiments may be implemented by hardware or a program instructing relevant hardware. The program may be stored in a computer-readable storage medium. The storage medium may be a read-only memory, a magnetic disk, an optical disc or the like.

The foregoing specific implementations further describe the objectives, technical solutions, and beneficial effects of the present invention in detail. It should be appreciated that the foregoing descriptions are merely specific implementations of the present invention, but are not intended to limit the protection scope of the present invention. Any modification, equivalent replacement, improvement, and the like made within the spirit and principle of the present invention should fall within the protection scope of the present invention.

CLAIMS

1. A method for lip-sync generation of digital humans, comprising:
 - obtaining text data and corresponding audio data;
 - 5 aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set;
 - determining, for a target text unit over a target time period, target mouth shape information corresponding to the target text unit based on a first dictionary, wherein the first dictionary
 - 10 comprises mouth shape information corresponding to any text unit;
 - determining target amplitude information corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit; and
 - determining a target lip shape over the target time period based on the target mouth shape information and the target amplitude information.
- 15 2. The method according to claim 1, wherein the audio data is pre-recorded audio, and the text unit is a character; and the aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set comprises:
 - generating, based on the text data and the audio data, a character set and an audio segment
 - 20 and a time period corresponding to any character in the character set using a forced alignment model.
3. The method according to claim 1, wherein the audio data is pre-recorded audio, and the text unit is a phoneme; and the aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a
- 25 time period corresponding to any text unit in the text unit set comprises:
 - determining, based on any character in the text data, a phoneme corresponding to the character to obtain a phoneme set; and
 - generating an audio segment and a time period corresponding to any phoneme in the phoneme set based on the phoneme set and the audio data using a forced alignment model.
- 30 4. The method according to claim 1, wherein the audio data is generated using a text-to-speech model based on the text data, and the text-to-speech model further generates the text unit set corresponding to the text data and the audio segment and the time period corresponding to any text unit in the text unit set while generating the audio data; and

the aligning the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set comprises:

obtaining the text unit set and the audio segment and the time period corresponding to any text unit in the text unit set from a generation result of the text-to-speech model.

5 5. The method according to claim 1, wherein the determining target amplitude information corresponding to the target mouth shape information based on a target audio segment corresponding to the target text unit comprises:

determining the target amplitude information based on an audio feature of the target audio segment, wherein the audio feature comprises at least one of a volume feature, a fundamental frequency feature, and a formant feature.

6. The method according to claim 5, wherein the audio feature is the volume feature; and the determining the target amplitude information based on an audio feature of the target audio segment comprises:

15 determining the target amplitude information based on a maximum volume value and a minimum volume value within a specific text range in which the target text unit is located.

7. The method according to claim 6, wherein the specific text range in which the target text unit is located is a sentence or paragraph in which the target text unit is located.

8. The method according to claim 1, further comprising:

20 forming a lip shape set based on a plurality of target lip shapes, wherein the lip shape set comprises a first lip shape and a second lip shape adjacent in the temporal dimension; and performing, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip shape sequence.

25 9. The method according to claim 8, wherein the interpolation operation is linear interpolation.

10. The method according to claim 8, wherein the performing, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip shape sequence comprises:

30 determining a lip shape corresponding to any smallest time unit in the blank time period based on the first lip shape and the second lip shape to obtain the first lip shape sequence.

11. The method according to claim 8, further comprising:

performing a smoothing operation on a target lip shape sequence formed by the first lip shape, the second lip shape, and the first lip shape sequence in the temporal dimension to obtain

a target smoothed lip shape sequence.

12. The method according to claim 11, wherein the smoothing operation is exponential smoothing or one-dimensional Gaussian filter smoothing.

13. An apparatus for lip-sync generation of digital humans, comprising:

5 an obtaining unit, configured to obtain text data and corresponding audio data;

an alignment unit, configured to align the text data with the audio data in the temporal dimension to obtain a text unit set corresponding to the text data and an audio segment and a time period corresponding to any text unit in the text unit set;

10 a mouth shape determination unit, configured to determine, for a target text unit over a target time period, target mouth shape information corresponding to the target text unit based on a first dictionary, wherein the first dictionary comprises mouth shape information corresponding to any text unit;

an amplitude determination unit, configured to determine target amplitude information corresponding to the target mouth shape information based on a target audio segment
15 corresponding to the target text unit; and

a lip shape determination unit, configured to determine a target lip shape over the target time period based on the target mouth shape information and the target amplitude information.

14. The apparatus according to claim 13, further comprising:

20 a set determination unit, configured to form a lip shape set based on a plurality of target lip shapes, wherein the lip shape set comprises a first lip shape and a second lip shape adjacent in the temporal dimension; and

an interpolation unit, configured to perform, based on the first lip shape and the second lip shape, an interpolation operation on a blank time period between the first lip shape and the second lip shape to obtain a first lip shape sequence.

25 15. The apparatus according to claim 14, further comprising:

a smoothing unit, configured to perform a smoothing operation on a target lip shape sequence formed by the first lip shape, the second lip shape, and the first lip shape sequence in the temporal dimension to obtain a target smoothed lip shape sequence.

30 16. A computer-readable storage medium, having a computer program stored thereon, wherein the computer program, when executed by a computer, causes the computer to perform the method according to any one of claims 1 to 12.

17. A computing device, comprising a memory and a processor, wherein the memory has executable code stored thereon, and the processor, when executing the executable code, implements the method according to any one of claims 1 to 12.