

AUDIO SYNTHESIS METHOD, ELECTRONIC DEVICE AND READABLE STORAGE MEDIUM

[0001] This application claims priority to Chinese Patent Application No. 202210460941.8,
5 titled "AUDIO SYNTHESIS METHOD, ELECTRONIC DEVICE AND READABLE
STORAGE MEDIUM", filed on April 28, 2022 with the China National Intellectual Property
Administration, which is incorporated herein by reference in its entirety.

FIELD

10 **[0002]** The present disclosure relates to the technical field of audio processing, and in
particular to a method for synthesizing an audio, an electronic device and a computer readable
storage medium.

BACKGROUND

15 **[0003]** With the development of artificial intelligence, it is possible to synthesize text into
highly realistic speech audio using speech synthesis technology. Synthesized audio is
commonly used for news broadcasting, book readings, human-computer dialogue, and other
scenarios. In addition, it is accustomed to listening to various types of music. However, how
to combine music and speech to create a new audio format, allowing audiences to hear
20 comforting or uplifting messages while enjoying music, is a challenge. The current solution is
in the form of radio broadcasts, where a host speaks intermittently while playing a
predetermined song for the audiences. However, such solution requires human intervention,
and the effect is limited by the broadcasting skills of the host.

SUMMARY

25 **[0004]** In views of this, an objective of the present disclosure is to provide a method for
synthesizing an audio, an electronic device and a computer readable storage medium, which
are not limited by human broadcasting skills and achieve better playing effect.

[0005] To this end, in a first aspect, a method for synthesizing an audio is provided
30 according to the present disclosure, the method includes:

obtaining an audio of a to-be-processed song and song information corresponding to the to-be-processed song;

performing a vocal separation processing on the audio of the to-be-processed song to obtain a vocal audio;

5 determining target timbre information from multiple pieces of candidate timbre information based on the vocal audio;

obtaining a text template;

generating a to-be-processed text based on the text template and the song information;

10 performing a speech synthesis processing on the to-be-processed text based on the target timbre information, to obtain a to-be-synthesized audio; and

synthesizing the to-be-synthesized audio and the audio of the to-be-processed song to obtain a synthesized audio.

[0006] In an embodiment, the process of determining the target timbre information from the
15 multiple pieces of the candidate timbre information based on the vocal audio includes:

extracting a voiceprint feature vector corresponding to the vocal audio;

generating a cosine similarity between the voiceprint feature vector and each of candidate voiceprint feature vectors, wherein each of the candidate voiceprint feature vectors corresponds to each of the pieces of the candidate timbre information; and

20 determining the candidate timbre information corresponding to a maximum cosine similarity as the target timbre information.

[0007] In an embodiment, the process of determining the target timbre information from the multiple pieces of candidate timbre information based on the vocal audio includes:

25 inputting the vocal audio into a timbre matching model obtained by training based on candidate audios, to obtain a similarity probability between the vocal audio and each of the candidate audios; and

determining the candidate timbre information of the candidate audio corresponding to a maximum similarity probability as the target timbre information.

[0008] In an embodiment, the song information includes multiple pieces of song sub-information, and the generating the to-be-processed text based on the text template and the song information includes:

5 inserting, based on information type of the song sub-information, each of the pieces of the song sub-information into a corresponding position in the text template, to obtain the to-be-processed text.

[0009] In an embodiment, the process of performing the speech synthesis processing on the to-be-processed text based on the target timbre information, to obtain the to-be-synthesized audio includes:

10 inputting the target timbre information and the to-be-processed text into a feature encoding model to obtain a text code; and

inputting the text code into a vocoder model to obtain the to-be-synthesized audio.

[0010] In an embodiment, a training process of the feature encoding model and the vocoder model includes:

15 obtaining training data corresponding to the pieces of candidate timbre information, wherein the training data comprises training texts and candidate audios corresponding to the training texts;

inputting the training texts into an initial feature encoding model to obtain training codes;

20 inputting the training codes into an initial vocoder model to obtain training output audios;

extracting a label code, which has a same type as each of the training codes, from each of the candidate audios, and generating a first loss value based on the label codes and the training codes;

25 generating a second loss value based the candidate audios and the training output audios;

performing parameter correction on the initial feature encoding model based on the first loss value, and performing feature correction on the initial vocoder model based on the second loss value, and

determining the initial feature encoding model after the parameter correction as the feature encoding model, and determining the initial vocoder model after parameter correction as the vocoder model, on detecting that training completion conditions are satisfied.

[0011] In an embodiment, the process of synthesizing the to-be-synthesized audio and the audio of the to-be-processed song audio includes:

determining an accompaniment time interval corresponding to the audio of the to-be-processed song;

determining a synthesis position type based on the template information, and filtering the accompaniment time interval based on the synthesis position type to obtain a target accompaniment time interval;

synthesizing the to-be-synthesized audio and a part of the audio of the to-be-processed song corresponding to the target accompaniment time interval, in a case that a length of the target accompaniment time interval is no less than a time length of the to-be-synthesized audio; and/or

splicing the to-be-synthesized audio and the audio of the to-be-processed song based on a type of the target accompaniment time interval, in a case that the length of the target accompaniment time interval is less than the time length of the to-be-synthesized audio.

[0012] In an embodiment, the process of splicing the to-be-synthesized audio and the audio of the to-be-processed song based on the type of the target accompaniment time interval includes:

splicing the to-be-synthesized audio before the audio of the to-be-processed song, in a case that the type of the target accompaniment time interval is an intro; and

splicing the to-be-synthesized audio after the audio of the to-be-processed song, in a case that the type of the target accompaniment time interval is an outro.

[0013] In an embodiment, the determining the accompaniment time interval corresponding to the audio of the to-be-processed song includes:

performing vocal detection on the audio of the to-be-processed song based on a vocal detection model to obtain a plurality of start time points and end time points;

removing an end time point and a start time point between which a distance is less

than a preset threshold, to obtain a target start time point and a target end time point; wherein the end time point is adjacent to the start time point; and

determining the accompaniment time interval based on the target start time point and the target end time point.

5 **[0014]** In a second aspect, an electronic device is provided according to the present disclosure, the electronic device includes a memory and a processor, where the memory is configured to store a computer program; and the processor is configured to, when executing the computer program, implement the above method for synthesizing the audio.

10 **[0015]** In a third aspect, a computer readable storage medium is provided according to the present disclosure, the computer readable storage medium is configured to store a computer program. The computer program, when executed by a processor, implements the above method for synthesizing the audio.

15 **[0016]** According to the method for synthesizing an audio provided in the present disclosure, the audio of the to-be-processed song and the corresponding song information are obtained; the vocal audio is obtained by performing vocal separation processing on the audio of the to-be-processed song; the target timbre information is determined from multiple pieces of candidate timbre information based on the vocal audio; the text template is obtained, and the to-be-processed text is generated based on the text template and the song information; speech synthesis processing is performed on the to-be-processed text to obtain the
20 to-be-synthesized audio, based on the target timbre information; and the to-be-synthesized audio and the audio of the to-be-processed song are synthesized to obtain the synthesized audio.

25 **[0017]** With the method for synthesizing an audio provided according to the embodiments of the present disclosure, the target timbre information is determined from multiple pieces of candidate timbre information using the vocal audio obtained by the vocal separation processing, where the target timbre information is the timbre information similar to the vocal timbre in the audio of the to-be-processed song. After a to-be-processed text is generated, the content of the vocal voice in the to-be-synthesized audio generated based on the target timbre information is the same as the to-be-processed text, and the timbre is similar to the timbre of
30 the singer of the audio of the to-be-processed song. In the synthesized audio obtained by synthesizing the to-be-synthesized audio and the audio of the to-be processed-song, the timbre

of the vocal part remains similar in the former part and later part, which has an excellent listening experience. In addition, the quality of the to-be-synthesized audio is high, which is not subjected to human broadcasting skills. As a result, the quality of the synthesized audio is high and the playing effect is excellent, thereby solving the problem that human intervention is required and the effect is limited by the broadcasting skills of the host.

[0018] Furthermore, an electronic device and a computer-readable storage medium provided according to the present disclosure, which also have the above beneficial effects.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] For clearer illustration of the technical solutions according to embodiments of the present disclosure or conventional techniques, hereinafter are briefly described the drawings to be applied in embodiments of the present disclosure or conventional techniques. Apparently, the drawings in the following descriptions are only merely embodiments of the present disclosure, and those skilled in the art can obtain other drawings from these drawings without any creative effort.

[0020] Figure 1 is a schematic diagram of a hardware composition framework applicable to a method for synthesizing an audio according to an embodiment of the present disclosure;

[0021] Figure 2 is a schematic diagram of a hardware composition framework applicable to a method for synthesizing an audio according to another embodiment of the present disclosure;

[0022] Figure 3 is a schematic flowchart of a method for synthesizing an audio according to an embodiment of the present disclosure;

[0023] Figure 4 is a schematic diagram of positions of synthesizing an audio according to an embodiment of the present disclosure; and

[0024] Figure 5 is a schematic flowchart of a whole processing of synthesizing an audio according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0025] For clearer illustration of the objective, technical solutions and advantages of the

embodiments of the present disclosure, technical solutions in the embodiments of the present disclosure hereinafter are described clearly and thoroughly with reference to the drawings in the embodiments of the present disclosure. Apparently, the described embodiments are only some, rather than all of the embodiments of the present disclosure. Any other embodiments
5 obtained by those skilled in the art based on the embodiments in the present disclosure without any creative effort fall within the protection scope of the present disclosure.

[0026] In order to facilitate understanding, firstly, the hardware composition framework applied in a solution corresponding to the method for synthesizing an audio according to the embodiments of the present disclosure is provided. Reference is made to Figure 1. Figure 1
10 illustrates a schematic diagram of a hardware composition framework applicable to a method for synthesizing an audio according to an embodiment of the present disclosure. An electronic device 100 may include a processor 101 and a memory 102, which may further include one or more of a multi-media component 103, an information input/output (I/O) interface 104 and a communication component 105.

[0027] Where, the processor 101 is configured to control an overall operation of the electronic device 100, to implement all or part of steps of the method for synthesizing an audio. The memory 102 is configured to store various types of data to support the operation of the electronic device 100. These data may include, for example, any application programs or instructions of the method operating on the electronic device 100, and related data of the
20 application programs. The memory 102 may be implemented by any type of volatile or non-volatile storage device or a combination thereof, such as one or more of a static random access memory (SRAM), an electrically erasable programmable read-only memory (EEPROM), an erasable programmable read-only memory (EPROM), a programmable read-only memory (PROM), a read-only memory (ROM), a magnetic memory, a flash
25 memory, a magnetic disk or an optical disk. In this embodiment, the memory 102 stores at least programs and/or data used to implement functions as follows:

obtaining an audio of a to-be-processed song and song information corresponding to the to-be-processed song;

performing a vocal separation processing on the audio of the to-be-processed song to
30 obtain a vocal audio;

determining target timbre information from multiple pieces of candidate timbre

information, based on the vocal audio;

obtaining a text template;

generating a to-be-processed text based on the text template and the song information;

5 performing a speech synthesis processing on the to-be-processed text to obtain a to-be-synthesized audio based on the target timbre information,; and

synthesizing the to-be-synthesized audio and the audio of the to-be-processed song, to obtain a synthesized audio.

[0028] The multi-media component 103 may include a screen and an audio component. The
10 screen may be, for example, a touch screen, and the audio component is used for outputting and/or inputting audio signals. For example, the audio component may include a microphone for receiving external audio signals. The received audio signal may be further stored in the memory 102 or sent through the communication component 105. The audio component further includes at least one speaker for outputting audio signals. The I/O interface 104
15 provides interfaces between the processor 101 and other interface modules. Other interface modules described above may be a keyboard, a mouse, buttons, and the like. The buttons may be virtual buttons or physical buttons. The communication component 105 is configured to perform wired or wireless communication between the electronic device 100 and other devices. The wireless communication may be, such as Wi-Fi, Bluetooth, near field
20 communication (NFC), 2G, 3G or 4G, or a combination of one or more thereof, therefore the corresponding communication component 105 may include a Wi-Fi component, a Bluetooth component, and an NFC component.

[0029] The electronic device 100 may be implemented by one or more of an application specific integrated circuit (ASIC), a digital signal processor (DSP), a digital signal processing
25 device (DSPD), a programmable logic device (PLD), a field programmable gate array (FPGA), a controller, a microcontroller, a microprocessor or other electronic components, which is used to implement the method for synthesizing an audio.

[0030] Apparently, the structure of the electronic device 100 shown in Figure 1 does not constitute a limitation on the electronic device provided in the embodiments of the present
30 disclosure. In practical applications, the electronic device 100 may include more or fewer

components than components shown in Figure 1, or a combination of some components.

[0031] It should be understood that the quantity of electronic devices in the embodiment of the present is not limited, and the method for synthesizing an audio may be implemented by cooperation of multiple electronic devices. In an embodiment, reference is made to Figure 2.

5 Figure 2 illustrates a schematic diagram of a hardware composition framework applicable to a method for synthesizing an audio according to another embodiment of the present disclosure. As shown in Figure 2, the hardware composition framework may include a first electronic device 11 and a second electronic device 12, where the first electronic device 11 and the second electronic device 12 are connected through a network 13.

10 **[0032]** In an embodiment, hardware structures of the first electronic device 11 and the second electronic device 12 may refer to the electronic device 100 in Figure 1. It can be understood that there are two electronic devices 100 in this embodiment, where the two electronic devices perform data interaction with each other. Further, the form of the network 13 is not limited in the embodiment of the present disclosure, for example, the network 13
15 may be a wireless network (such as WIFI, Bluetooth), or a wired network.

[0033] Where, the first electronic device 11 and the second electronic device 12 may be the same type of electronic device. For example, the first electronic device 11 and the second electronic device 12 are both servers. The first electronic device 11 and the second electronic device 12 may be different types of electronic devices. For example, the first electronic device
20 11 may be a smart phone or other intelligent terminal, and the second electronic device 12 may be a server. In an embodiment, a server with high computing ability may be the second electronic device 12 to improve the efficiency and reliability of data processing, and thereby improving the processing efficiency of audio synthesis. In addition, a smart phone with low cost and wide application range is the first electronic device 11 to implement interactions
25 between the second electronic device 12 and users. It can be understood that the interaction process may be as follows. The smartphone obtains a song information and sends the song information to the server, and the server generates a synthesized audio. The server then sends the synthesized audio to the smartphone, and the smartphone plays the synthesized audio.

[0034] In an embodiment, reference is made to Figure 3. Figure 3 is a schematic flowchart of a method for synthesizing an audio according to an embodiment of the present disclosure.
30 The method in the embodiment includes steps S101 to S106 as follows.

[0035] In step S101, audio of a to-be-processed song and song information corresponding to the to-be-processed song are obtained.

[0036] Where, the song information refers to information used to identify the song, which may include a name of the song, a name of the singer, etc. In addition, the song information plays a role in the generation of the to-be-synthesized audio. In order to enrich the content of the to-be-synthesized audio, the song information may further include additional information, e.g., the background for creating the song, the purpose for creating the song and so on.

[0037] The audio of the to-be-processed song refers to an original audio of the song that is required to be synthesized to obtain a broadcasting effect. The broadcasting effect is the effect obtained after incorporating a speech related to the song into the song, which can improve the overall listening experience when the song is played.

[0038] According to different content of the song information, there are various approaches to obtain the audio of the to-be-processed song. For example, in a case that the song information itself includes the audio of the to-be-processed song, the audio of the to-be-processed song may be directly filtered from the song information. In a case that the song information does not include the audio of the to-be-processed song, the identity of the audio of the to-be processed song may be uniquely determined based on the information such as the name of the song and the name of the singer, and the audio of the to-be-processed song may be obtained from a preset library of the audio files of songs.

[0039] In step S102, a vocal separation processing is performed on the audio of the to-be-processed song to obtain a vocal audio.

[0040] To ensure the vocal timbre of the to-be-synthesized audio to be similar to the vocal timbre of the singer of the to-be-processed audio to maintain the vocal timbre of the overall synthesized audio relatively consistent and improve the playing effect, the following solution is provided according to the present disclosure. In the present disclosure, on the basis of the singer voice of the audio of the to-be-processed song, the timbre that is most similar to the singer voice of the audio of the to-be-processed song is selected from various timbres, to generate a to-be-synthesized audio. To achieve the above objective, it is firstly required to determine the timbre of the singer voice of the audio of the to-be-processed song. In an embodiment, the vocal separation processing is performed on the audio of the to-be-processed song to obtain a vocal audio, the vocal audio is the singing effect of the singer is singing

without accompanying instrumental music in listening experience. The specific manner of vocal separation processing is not limited in this embodiment, for example, a trained vocal separation model may be used for processing, and the vocal separation model may be obtained from training based on a pure vocal audio and a pure accompaniment audio. When
5 vocal separation processing is performed, the audio of the to-be-processed song is inputted into the vocal separation model, and the vocal audio and the accompaniment audio are separated and outputted by the vocal separation model. The specific types of vocal separation model are not limited, such as convolutional neural network (CNN), deep neural network (DNN), recurrent neural network (RNN), or UNET models in the complex domain.

10 **[0041]** In step S103, based on the vocal audio, target timbre information is determined from multiple pieces of candidate timbre information.

[0042] The multiple pieces of candidate timbre information refer to multiple pieces of identification information corresponding to multiple candidate vocal timbres respectively. And, the target timbre information is one of identification information that is most similar to
15 the singer timbre in the vocal audio. The target timbre information is the information of vocal timbre selected to be used for speech synthesis processing, so that the vocal timbre of the to-be-synthesized audio is similar to vocal timbre of the singer. The specific form of the timbre information is not limited, for example, it may be a serial number or a name of the candidate vocal.

20 **[0043]** In an embodiment, the specific determination method of the target timbre information is not limited. In an embodiment, the features that can characterize the timbre in the vocal audio may be extracted, and matched with features that characterize the timbre corresponding to each of multiple pieces of candidate timbre information, the candidate timbre information that is most similar to the singer timbre is determined as the target timbre
25 information. In an embodiment, a voiceprint feature vector corresponding to the vocal audio is extracted. The specific extraction manner of the voiceprint feature vector is not limited, as long as it can characterizes the timbre of the vocal audio. In addition, each of the candidate voiceprint feature vectors is generated in advance, and each of the candidate voiceprint feature vectors is corresponding to each pieces of candidate timbre information, which is
30 obtained by extracting the voiceprint features of the speaking audio of candidate corresponding to each of the multiple pieces of candidate timbre information. It should be

understood that all voiceprint feature vectors are extracted in the same manner, for example, the I-Vector model or the X-Vector model may be used for feature extraction. The cosine similarity between the voiceprint feature vector of the vocal audio and each of the candidate voiceprint feature vectors is generated, and the candidate timbre information corresponding to the maximum cosine similarity is determined as the target timbre information. It should be understood that the maximum cosine similarity can indicate that the two are the most similar.

[0044] In an embodiment, the vocal audio may be inputted into a timbre matching model obtained from training based on candidate audios, and similarity probability between the vocal audio and each of the candidate audios is obtained. In an embodiment, the timbre matching model is trained based on each of the speaking audios of candidates. After the vocal audio is inputted into the timbre matching model, the timbre matching model can output the similarity probability between the vocal audio and each of the candidate audios (i.e., the speaking audios of candidates). The similarity probability refers to the probability determined by the timbre matching model that the vocal audio and the candidate audio have the same timbre. The candidate timbre information of the candidate audio corresponding to the maximum similarity probability is determined as the target timbre information. The maximum similarity probability indicates that the vocal audio is most likely to have the same timbre with the candidate audio, i.e., the two are most similar.

[0045] In step S104, a text template is obtained; and a to-be-processed text is generated based on the text template and the song information.

[0046] The to-be-processed text refers to a content text corresponding to the to-be-synthesized audio, which is generated based on the song information and the text template. The to-be-processed text generated based on the song information has a relatively close correlation with the audio of the to-be-processed song, thereby associating the to-be-synthesized audio with the audio of the to-be-processed song. Typically, the content of the song information is substantial, e.g., including multiple pieces of song sub-information. In this case, based on an information type of each of the multiple pieces of song sub-information, each song sub-information may be inserted into a corresponding position in the text template to obtain the to-be-processed text. For example, song information includes three pieces of song sub-information: a name of the song, a name of the singer and additional information. The text template is: "Welcome to listen to song B performed by A, this song wishes to C",

where A is the name of the singer, B is the name of the song, and C is the information of the purpose of the song. For example, C may be "encourage everyone to maintain a positive attitude in difficult situations", which belongs to the additional information. The text template may have different types. Through this text organization manner, the song information and the text content to be expressed are integrated to obtain a variety of different types of the to-be-processed text. The text template types include but is not limited to an encouragement type, straightforward narration type, a happy sharing type, and a playful type.

[0047] According to an embodiment, the specific manner for obtaining a text template is not limited. In an embodiment, the text template can be obtained based on the template information. The template information refers to the information used to describe the text template of the to-be-processed text for generating the to-be-synthesized audio. In an embodiment, the template information may be an identifier identifying the text template, such as a serial number or a name. According to this embodiment, the specific manner for obtaining the template information and the song information are not limited. For example, in an embodiment, interactions with the user can be performed through data interaction units, such as a keyboard, a mouse and a touch screen to obtain anyone of the template information and the song information based data transmitted by the user through the data interaction units. In an embodiment, data interaction with other electronic devices may be performed, or the song information and/or the template information may be obtained from a specified storage path.

[0048] It should be noted that, the quantity of the to-be-processed texts may be one or more, i.e., an audio of a to-be-processed song may correspond to multiple to-be-synthesized audios, and the content of the to-be-processed text corresponding to each of the to-be-synthesized audios may be different.

[0049] In step S105, based on the target timbre information, speech synthesis processing is performed on the to-be-processed text to obtain a to-be-synthesized audio.

[0050] By using the target timbre information, the vocal timbre of the to-be-synthesized audio may be specified, and a vocal content of the to-be-synthesized audio is determined by using the to-be-processed text. Therefore, based on the target timbre information, speech synthesis processing is performed on the to-be-processed text to obtain a to-be-synthesized audio. According to this embodiment, the specific manner of speech voice synthesis

processing is not limited. In an embodiment, the target timbre information and the to-be-processed text are inputted into a feature encoding model to obtain a text code. A feature encoding model is configured to determine the encoding method based on the target timbre information, and encode the to-be-processed text using the encoding method, where the encoding method is related to the vocal timbre. A vocoder model is configured to output a corresponding audio based on the text code, and the text code is inputted into the vocoder model to obtain the to-be-synthesized audio.

[0051] It should be understood that the identification manner of the vocoder model to the text code should match the feature encoding model. To ensure the above, the vocoder model and the feature encoding model are jointly trained. In a training process, the training data corresponding to each of the multiple pieces of the candidate timbre information is firstly obtained. The training data includes training texts and their corresponding candidate audios, and the timbre of each candidate corresponding to each one of candidate audios is consistent with the timbre of each candidate of the candidate audio in the step S102, which may be obtained by the same or different content reciting by several candidates. The specific content of the training text data is not limited, for example, including words, Chinese pinyin data, and pause identification data, etc. Where, the pause identification data is used to indicate where the words break sentences or pause. The training texts are inputted into the initial feature encoding model which extracts the corresponding code based on the texts. For example, it may be MEL spectrum encoding extraction to obtain the training codes.

[0052] The training codes are inputted into the initial vocoder model, and the initial vocoder model identifies the training codes to train output audios. For each candidate audio, a label code is extracted from the candidate audio, where the label code is of the same type as the corresponding training code. For example, when the initial feature encoding model extracts the MEL spectrum encoding, the candidate audio is subjected to the MEL spectrum encoding extraction to obtain the label code. And the label codes and the training codes are used to generate a first loss value. Correspondingly, each candidate audio is used as another label. The candidate audios and the training output audios are used to generate a second loss value. Parameter correction is performed on the initial feature encoding model based on the first loss value, and feature correction is performed on the initial vocoder model based on the second loss value, so that the initial feature encoding model and the initial vocoder model learn the related content. When it is detected that the training completion conditions are satisfied, the

initial feature encoding model after parameter correction is determined as the feature encoding model, and the initial vocoder model after parameter correction is determined as the vocoder model. The quantity and specific content of the training completion conditions are set according to requirements, such as an accuracy condition, a training time condition, or a training round condition.

[0053] It should be noted that the model types of vocoder model and feature encoding model are not limited. For example, the vocoder model may use model structures such as Tacotron, FastSpeech, etc.

[0054] In step S106, the to-be-synthesized audio and the audio of the to-be-processed song is synthesized to obtain a synthesized audio.

[0055] By synthesizing the to-be-synthesized audio and the audio of the to-be-processed song, the synthesized audio with high quality, consistent timbre in an entire audio and listening experience is obtained. It should be understood that, in order to maintain integrity of the audio of the to-be-processed song, the section where there is no vocal in the audio of the to-be processed song may be synthesized to generate a to-be-synthesized audio. In an embodiment, the accompaniment time interval corresponding to the audio of the to-be-processed song may be determined. The accompaniment time interval refers to the time interval without vocal voice and only with accompaniment. The template information may indicate its fitting location. For example, the to-be-played song may usually be introduced before the song starts, the purpose for creating the song may be introduced after finishing the song playing, and the background for creating the song may be introduced during the song playing, specific settings may be performed. A synthesized position type is determined based on the template information, and the accompaniment time interval is filtered based on the synthesized position type to obtain a target accompaniment time interval. Each of the to-be-synthesized audios may correspond to one or more of the target accompaniment time intervals.

[0056] In a case that the length of the target accompaniment time interval is no less than the time length of the to-be-synthesized audio, it indicates that the to-be-synthesized audio may be played out within the target accompaniment time interval. In this case, the to-be-synthesized audio can be synthesized with the corresponding part of the target accompaniment time interval in the audio of the to-be-processed song, so that the

to-be-synthesized audio may be played simultaneously when playing the accompaniment. In a case that the length of the target accompaniment time interval is less than the time length of the to-be-synthesized audio, in order to maintain the integrity of the audio of the to-be-processed song, the to-be-synthesized audio and the audio of the to-be-processed song may be spliced based on the type of the target accompaniment time interval. In other words, the to-be-synthesized audio and the audio of the to-be-processed song are not overlapped in time, thereby extending the time length for the to-be-processed song. In an embodiment, in a case that the type of the target accompaniment time interval is an intro, the to-be-synthesized audio is spliced before the audio of the to-be-processed song; in a case that the type of the target accompaniment time interval is an outro, the to-be-synthesized audio is spliced after the audio of the to-be-processed song. Reference is made to Figure 4. Figure 4 is a schematic diagram of positions of synthesizing an audio according to an embodiment of the present disclosure. As shown in Figure 4, in a case that the type of the target accompaniment time interval is the intro, the to-be-synthesized audio (i.e., a voice audio or speech) is spliced to the front of the song audio; in a case that the type is the outro, the to-be-synthesized audio is spliced to the rear of the song audio. According to this embodiment, the processing manner of an interlude is not limited, for example, the interlude may be randomly spliced to the front or the rear of the song audio, or be spliced at a specified position, such as the rear.

[0057] In addition, when the to-be-synthesized audio is synthesized with the part corresponding to the target accompaniment time interval in the audio of the to-be-processed song, the volume of the to-be-synthesized audio and the part corresponding to the target accompaniment time interval may be adjusted, for a better listening experience. When the part corresponding to the target accompaniment time interval is an intro, the volume of the intro may be set to be increased from small to large, eventually returning to the normal volume, while the volume of the to-be-synthesized audio may be set to be decreased from large to small, creating a gradual introduction effect into the song. When the part corresponding to the target accompaniment time interval is an outro, the volume of the outro may be set to be decreased from large to small, while the volume of the to-be-synthesized audio may be set to be increased from small to large, creating a gradual outro effect. When the part corresponding to the target accompaniment time interval is an interlude, the volume of the interlude may be set to be kept consistently small, lower than the volume of the to-be-synthesized audio, or a similar setting to the intro may be applied.

[0058] According to an embodiment, the specific manner of obtaining the accompaniment time interval is not limited. In an embodiment, the lyrics corresponding to the audio of the to-be-processed song may be obtained, and the time interval of the accompaniment is determined based on the timestamp of the lyrics. For example, for a song "Common Jasmin Orange" of Jay Chou, through lyric analysis, the first line of the lyrics "The sparrow outside the window is talking on the telephone pole" starts at about 28 seconds, then the period from 0 to 28 seconds is a intro. For the lyric "It rained all night and my love is spilling over like the rain", this sentence starts at about 162 seconds, and the last sentence ends at about 135 seconds, during which there is a part about 27 seconds without vocal, i.e., the interlude. The last line of the lyrics ends at about 270 seconds, while the song ends at 299 seconds, where there is a part about 29 seconds without vocal, i.e., the outro. There are three accompaniment time intervals respectively from 0 to 28 seconds, from 135 to 162 seconds, and from 270 to 299 seconds, and the three accompaniment time intervals correspond to different types of synthesis positions, which are front, middle and rear respectively.

[0059] In an embodiment, in a case that there is no lyrics, the vocal detection model may be used to detect the vocal of the audio of the to-be-processed song to obtain multiple start time points and end time points. When the singer sings, there is also an interval between two lines of adjacent lyrics, where the to-be-synthesized audio should not be inserted into. To prevent a short to-be-synthesized audio from being inserted in the middle of two lines of adjacent lyrics, the adjacent end time point and start time point, between which the distance is less than the preset threshold, may be removed, to obtain a target start time point and a target end time point. Where the distance refers to a distance in time intervals, and the specific value of the preset threshold is not limited. The time interval of the singing part may be determined based on the target start time point and the target end time point, and the accompaniment time interval may be determined similarly.

[0060] Referring to Figure 5, Figure 5 is a schematic flowchart of a whole processing of synthesizing an audio according to an embodiment of the present disclosure. The "word" indicates the above text template. The audio of the to-be-processed song is determined based on the song information, and then the vocal separation and timbre identification are performed on the audio of the to-be-processed song to obtain the target timbre information. After a to-be-processed text is generated using the words and song information, speech synthesis is performed on the to-be-processed text using the speech synthesis technology based on the

target timbre information, the synthesized speech is obtained, i.e., the above to-be-synthesized audio. The audio synthesis is performed on the synthesized speech audio and the audio of the to-be-processed song corresponding to the song information by using automatic crossfading technology, to obtain a long audio, which is the synthesized audio.

5 **[0061]** With the method for synthesizing an audio provided according to the embodiments of the present disclosure, the target timbre information is determined from multiple pieces of candidate timbre information using the vocal audio obtained by the vocal separation processing, where the target timbre information is the timbre information similar to the vocal timbre in the audio of the to-be-processed song. After a to-be-processed text is generated, the
10 content of the vocal voice in the to-be-synthesized audio generated based on the target timbre information is the same as the to-be-processed text, and the timbre is similar to the timbre of the singer of the audio of the to-be-processed song. In the synthesized audio obtained by synthesizing the to-be-synthesized audio and the audio of the to-be processed-song, the timbre of the vocal part remains similar in the former part and later part, which has an excellent
15 listening experience. In addition, the quality of the to-be-synthesized audio is high, which is not subjected to human broadcasting skills. As a result, the quality of the synthesized audio is high and the playing effect is excellent, thereby solving the problem that human intervention is required and the effect is limited by the broadcasting skills of the host.

[0062] Hereinafter, a computer readable storage medium is provided according to the
20 embodiment of the present disclosure. The computer readable storage medium described below and the method for synthesizing the audio described above may be referred to each other.

[0063] A computer readable storage medium is further provided in the present disclosure, where a computer program stores thereon. The computer program, when executed by a
25 processor, implements steps of the method for synthesizing the audio described above.

[0064] The computer readable storage medium may include various medium which can store program codes, such as a U disk, a mobile hard disk, a read-only memory (ROM), a random-access memory (RAM), a magnetic disk or an optical disk, etc.

[0065] The above embodiments of the present disclosure are described in a progressive
30 manner. Each embodiment is mainly focused on describing differences from other embodiments, and references may be made among these embodiments with respect to the

same or similar parts. Since the devices disclosed in the embodiment corresponds to the method disclosed in the embodiment, the description for the device is simple, and reference may be made to description of the method in the embodiment for the relevant parts.

[0066] Those skilled in the art may further appreciate that the units and algorithm steps of the examples described in conjunction with the embodiments disclosed herein may be implemented by electronic hardware, computer software, or a combination thereof. In order to clearly illustrate interchangeability between the hardware and the software, the components and steps of each example have been generally described according to their functions in the above description. Whether these functions are executed by means of hardware or software depends on the specific application and design constraints of the technical solution. Those skilled in the art may implement the described function by using different methods for each particular application, but such implementation should not be considered as beyond the scope of the present disclosure.

[0067] The steps of the method or algorithm described in conjunction with the embodiments of the present disclosure may be implemented by hardware, a software module executed by a processor, or a combination thereof. The software module may reside in a random access memory (RAM), an internal memory, a read-only memory (ROM), an electrically-programmable ROM, an electrically-erasable programmable ROM, a register, a hard disk, a removable disk, a CD-ROM, or any other forms of storage medium known in the technical field.

[0068] Finally, it should also be noted that in this specification, relationship terms such as "first" and "second" etc. are only used to distinguish one entity or operation from another entity or operation, and do not necessarily require or imply that there is any such actual relationship or order between these entities or operations. Moreover, the terms "comprises", "includes", or any other variation are intended to cover a non-exclusive inclusion such that a process, method, article, or device comprising a set of elements includes not only those elements but also other elements not expressly listed, and may further include elements inherent to such a process, method, article, or device.

[0069] In this specification, specific examples are used to illustrate the principles and implementations of the present disclosure. The above descriptions of embodiments are only used to help understand the methods and essential concept of the present disclosure.

Meanwhile, those skilled in the art may make changes in specific implementations and application scenarios based on concepts of the present disclosure. Based on the above description, the content of the specification shall not be understood as a limitation of the present disclosure.

CLAIMS

1. A method for synthesizing an audio, comprising:

obtaining an audio of a to-be-processed song and song information corresponding to the
5 to-be-processed song;

performing a vocal separation processing on the audio of the to-be-processed song to
obtain a vocal audio;

determining target timbre information from a plurality of pieces of candidate timbre
information based on the vocal audio;

10 obtaining a text template;

generating a to-be-processed text based on the text template and the song information;

performing a speech synthesis processing on the to-be-processed text based on the target
timbre information, to obtain a to-be-synthesized audio; and

synthesizing the to-be-synthesized audio and the audio of the to-be-processed song to
15 obtain a synthesized audio.

2. The method for synthesizing the audio according to claim 1, wherein the process of
determining the target timbre information from the plurality of pieces of candidate timbre
information based on the vocal audio comprises:

20 extracting a voiceprint feature vector corresponding to the vocal audio;

generating a cosine similarity between the voiceprint feature vector and each of
candidate voiceprint feature vectors, wherein each of the candidate voiceprint feature vectors
corresponds to each of the pieces of the candidate timbre information; and

determining the candidate timbre information corresponding to a maximum cosine
25 similarity as the target timbre information.

3. The method for synthesizing the audio according to claim 1, wherein the process of
the determining the target timbre information from the plurality of pieces of candidate timbre
information based on the vocal audio comprises:

inputting the vocal audio into a timbre matching model obtained by training based on candidate audios, to obtain a similarity probability between the vocal audio and each of the candidate audios; and

5 determining the candidate timbre information of the candidate audio corresponding to a maximum similarity probability as the target timbre information.

4. The method for synthesizing the audio according to claim 1, wherein the song information comprises a plurality of pieces of song sub-information, and the generating the to-be-processed text based on the text template and the song information comprises:

10 inserting, based on information type of the song sub-information, each of the pieces of the song sub-information into a corresponding position in the text template, to obtain the to-be-processed text.

5. The method for synthesizing the audio according to claim 1, wherein the process of performing the speech synthesis processing on the to-be-processed text based on the target timbre information, to obtain the to-be-synthesized audio comprises:

inputting the target timbre information and the to-be-processed text into a feature encoding model to obtain a text code; and

20 inputting the text code into a vocoder model to obtain the to-be-synthesized audio.

6. The method for synthesizing the audio according to claim 5, wherein a training process of the feature encoding model and the vocoder model comprises:

obtaining training data corresponding to the pieces of candidate timbre information, wherein the training data comprises training texts and candidate audios corresponding to the training texts;

25 inputting the training texts into an initial feature encoding model to obtain training codes;

inputting the training codes into an initial vocoder model to obtain training output audios;

extracting a label code, which has a same type as each of the training codes, from each of the candidate audios, and generating a first loss value based on the label codes and the training codes;

generating a second loss value based the candidate audios and the training output audios;

5 performing parameter correction on the initial feature encoding model based on the first loss value, and performing feature correction on the initial vocoder model based on the second loss value, and

determining the initial feature encoding model after the parameter correction as the feature encoding model, and determining the initial vocoder model after parameter correction
10 as the vocoder model, on detecting that training completion conditions are satisfied.

7. The method for synthesizing the audio according to claim 1, wherein the process of synthesizing the to-be-synthesized audio and the audio of the to-be-processed song audio comprises:

15 determining an accompaniment time interval corresponding to the audio of the to-be-processed song;

determining a synthesis position type based on the template information, and filtering the accompaniment time interval based on the synthesis position type to obtain a target accompaniment time interval;

20 synthesizing the to-be-synthesized audio and a part of the audio of the to-be-processed song corresponding to the target accompaniment time interval, in a case that a length of the target accompaniment time interval is no less than a time length of the to-be-synthesized audio; and/or

splicing the to-be-synthesized audio and the audio of the to-be-processed song based on
25 a type of the target accompaniment time interval, in a case that the length of the target accompaniment time interval is less than the time length of the to-be-synthesized audio.

8. The method for synthesizing the audio according to claim 7, wherein the process of splicing the to-be-synthesized audio and the audio of the to-be-processed song based on the

type of the target accompaniment time interval comprises:

splicing the to-be-synthesized audio before the audio of the to-be-processed song, in a case that the type of the target accompaniment time interval is an intro; and

5 splicing the to-be-synthesized audio after the audio of the to-be-processed song, in a case that the type of the target accompaniment time interval is an outro.

9. The method for synthesizing the audio according to claim 7, wherein the process of determining the accompaniment time interval corresponding to the audio of the to-be-processed song comprises:

10 performing vocal detection on the audio of the to-be-processed song based on a vocal detection model to obtain a plurality of start time points and end time points;

removing an end time point and a start time point between which a distance is less than a preset threshold, to obtain a target start time point and a target end time point; wherein the end time point is adjacent to the start time point; and

15 determining the accompaniment time interval based on the target start time point and the target end time point.

10. An electronic device, comprising a memory and a processor, wherein:

the memory is configured to store a computer program; and

20 the processor is configured to, when executing the computer program, implement the method for synthesizing the audio according to any one of claims 1 to 9.

11. A computer readable storage medium, configured to store a computer program, wherein the computer program, when executed by a processor, implements the method for
25 synthesizing the audio according to any one of claims 1 to 9.