

METHOD AND DEVICE FOR PROCESSING CHORUS AUDIO, AND STORAGE MEDIUM

5 **[0001]** This application claims priority to Chinese Patent Application No. 202110460280.4 titled "METHOD AND DEVICE FOR PROCESSING CHORUS AUDIO, AND STORAGE MEDIUM", filed on April 27, 2021 with the China National Intellectual Property Administration (CNIPA), which is incorporated herein by reference in its entirety.

FIELD

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

BACKGROUND

15 **[0003]** Audio software, video software, office software, and other types of software are constantly becoming more prevalent thanks to the rapid development of computer technology, which results in a much more convenient life. With the audio type software, users can experience listening to and singing songs.

20 **[0004]** In the conventional technology, singing data from multiple users are directly superimposed to provide the users with an auditory experience of a concert chorus. However, the audio obtained through such simple superimposition provides a sound field located at the center of the human head in the sense of hearing, which causes an in-head effect. Since such sound field is not wide enough, it compromises the auditory experience.

SUMMARY

25 **[0005]** An objective of the present disclosure is to provide a method and device for processing a chorus audio, and a storage medium, which can eliminate the in-head effect due to the sound field is located at the center of the human head, to provide a wider sound field and an improved auditory experience.

30 **[0006]** Technical solutions provided through embodiments of the present disclosure for solving the above technical problems are described below.

[0007] A method for processing a chorus audio is provided. The method includes: obtaining multiple dry audios respectively from multiple performers singing a same target song; performing time alignment on the multiple dry audios; performing virtual sound image localization on the multiple dry audios after the time alignment, to locate the dry audios to multiple virtual sound images; where the multiple virtual sound images are located in a pre-configured virtual sound image coordinate system; where, the virtual sound image coordinate system is configured as: a human head is a center of the virtual sound image coordinate system; a midpoint of a line segment where a left ear and a right ear are located is a coordinate origin; a positive direction of a first axis represents a straight front direction of the human head; a positive direction of a second axis represents a lateral direction of the human head from the left ear to the right ear, and a positive direction of a third axis represents a straight above direction of the human head; a distance between each virtual sound image and the coordinate origin is within a set distance range, and an elevation angle of each virtual sound image relative to a plane formed by the first axis and the second axis is within a set angle range; generating a chorus audio based on the multiple dry audios after the virtual sound image localization; and synthesizing, in a case that a lead audio based on the target song is obtained, the lead audio, the chorus audio and an accompaniment, to output a chorus-effect audio.

[0008] In an embodiment of the present disclosure, the performing time alignment on the obtained multiple dry audios includes: determining a reference audio corresponding to the target song; for each of the obtained multiple dry audios, extracting an audio feature of the dry audio and an audio feature of the reference audio, where the audio feature is a fingerprint feature or fundamental frequency feature; determining a time instant corresponding to a maximum similarity between the audio feature of the dry audio and the audio feature of the reference audio as an audio alignment time; and performing the time alignment on the dry audio based on the audio alignment time.

[0009] In an embodiment of the present disclosure, the method further includes: performing band-pass filtering on the each of the multiple dry audios respectively, to obtain bass data; and the generating the chorus audio based on the multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after the virtual sound image localization and the bass data.

[0010] In an embodiment of the present disclosure, the method further includes: performing reverberation simulation on the each of the obtained multiple dry audios respectively; and the generating the chorus audio based on the multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after the
5 virtual sound image localization and the multiple dry audios after the reverberation simulation.

[0011] In an embodiment of the present disclosure, the performing the reverberation simulation on each of the obtained multiple dry audios respectively includes: performing the reverberation simulation on each of the obtained multiple dry audios respectively by using a
10 cascade of a comb filter and an all-pass filter.

[0012] In an embodiment of the present disclosure, after the performing virtual sound image localization on the multiple dry audios after the time alignment, the method further includes: performing reverberation simulation respectively on each of the multiple dry audios after the virtual sound image localization; and the generating the chorus audio based on the
15 multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after both the virtual sound image localization and the reverberation simulation.

[0013] In an embodiment of the present disclosure, the method further includes: performing dual-channel simulation on each of the obtained multiple dry audios respectively; and the
20 generating the chorus audio based on the multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation.

[0014] In an embodiment of the present disclosure, the method further includes: after the performing the dual-channel simulation on each of the obtained multiple dry audios
25 respectively, the method further includes: performing reverberation simulation on the multiple dry audios after the dual-channel simulation; and the generating the chorus audio based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation includes: generating the chorus audio based on the multiple dry
30 audios after the virtual sound image localization and the multiple dry audios after both the dual-channel simulation and the reverberation simulation.

[0015] In an embodiment of the present disclosure, the performing virtual sound image

localization on the multiple dry audios after the time alignment includes: grouping the multiple dry audios after the time alignment based on a quantity of virtual sound images, where a quantity of groups is equal to the quantity of the virtual sound images; and locating each group of dry audios respectively to one of the virtual sound images, so that each group of dry audios correspond to a different one of the virtual sound images.

[0016] In an embodiment of the present disclosure, among the virtual sound images, an elevation angle of a virtual sound image located behind the human head relative to the plane formed by the first axis and the second axis is greater than an elevation angle of a virtual sound image located in front of the human head relative to the plane formed by the first axis and the second axis; or the virtual sound images are evenly distributed in a full circle on the plane formed by the first axis and the second axis.

[0017] In an embodiment of the present disclosure, the synthesizing the lead audio, the chorus audio and the accompaniment includes: performing volume adjustment on the lead audio and the chorus audio respectively, and/or performing reverberation simulation on the lead audio and the chorus audio; and synthesizing the accompaniment with the lead audio and the chorus audio after the volume adjustment and/or the reverberation simulation.

[0018] An apparatus for processing a chorus audio is provided. The apparatus includes: a dry audio obtaining module, configured to obtain multiple dry audios respectively from multiple performers singing a same target song; an alignment module, configured to perform time alignment on the obtained multiple dry audios; a virtual sound image localization module, configured to perform virtual sound image localization on the multiple dry audios after the time alignment, to locate the dry audios to multiple virtual sound images; where the multiple virtual sound images are located in a pre-configured virtual sound image coordinate system; where, the virtual sound image coordinate system is configured as: a human head is a center of the virtual sound image coordinate system; a midpoint of a line segment where a left ear and a right ear are located is a coordinate origin; a positive direction of a first axis represents a straight front direction of the human head; a positive direction of a second axis represents a lateral direction of the human head from the left ear to the right ear, and a positive direction of a third axis represents a straight above direction of the human head; a distance between each virtual sound image and the coordinate origin is within a set distance range, and an elevation angle of each virtual sound image relative to a plane formed by the

first axis and the second axis is within a set angle range; a chorus audio generating module, configured to generate a chorus audio based on the multiple dry audios after the virtual sound image localization; and a chorus-effect audio output module, configured to synthesize, in a case that a lead audio based on the target song is obtained, the lead audio, the chorus audio and an accompaniment, to output a chorus-effect audio.

[0019] A device for processing a chorus audio is provided. The device includes: a memory, configured to store a computer program; and a processor, configured to implement, when executing the computer program, the method for processing a chorus audio according to any of the above aspects.

[0020] A computer-readable storage medium storing a computer program is provided. The computer program, when executed by a processor, implements the method for processing a chorus audio according to any of the above aspects.

[0021] With the technical solution provided in the embodiments of the present disclosure, after the dry audios from multiple performers singing the same target song are obtained respectively, the time alignment is performed on the obtained multiple dry audios. The virtual sound image localization is performed on the multiple dry audios after the time alignment to locate the multiple dry audios to multiple virtual sound images. The virtual sound images are located in the virtual sound image coordinate system where the center is the human head. A distance from each virtual sound image to the coordinate origin is within the set distance range. The virtual sound images surround the human ears. The chorus audio is generated based on the multiple dry audios after the virtual sound image localization. In a case that the lead audio based on the target song is obtained, the lead audio, the chorus audio and the accompaniment are synthesized to output the chorus-effect audio. Since the multiple dry audios are located to the multiple virtual sound images surrounding the human ears, the generated chorus audio has a sound field with the surround sound effect. In a sense of hearing, it can effectively prevent the sound field of the final output chorus-effect audio from locating on the center of the head causing the in-head effect. As a result, the sound field is wider.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] In order to more clearly illustrate the technical solutions in the embodiments of the present disclosure or in the conventional technology, the drawings used in the description of

the embodiments or the conventional technology are briefly described hereinafter. It is apparent that the drawings in the following description illustrate only some embodiments of the present disclosure. Other drawings may be obtained by those ordinary skilled in the art without creative efforts based on the provided drawings.

5 **[0023]** Figure 1 is a flow chart of implementing a method for processing a chorus audio according to an embodiment of the present disclosure;

[0024] Figure 2 is a schematic diagram of orientations of virtual sound images in a virtual sound image localization coordinate system according to an embodiment of the present disclosure;

10 **[0025]** Figure 3 is a schematic diagram of virtual sound image localization according to an embodiment of the present disclosure;

[0026] Figure 4 is a schematic diagram of virtual sound images after localization according to an embodiment of the present disclosure;

15 **[0027]** Figure 5 is a schematic diagram of a composition of a spatial sound field process according to an embodiment of the present disclosure;

[0028] Figure 6 is a schematic diagram of a cascade of comb filters and all-pass filters according to an embodiment of the present disclosure;

[0029] Figure 7 is a schematic diagram of a reverberation impulse response according to an embodiment of the present disclosure;

20 **[0030]** Figure 8 is a schematic diagram of dual-channel simulation according to an embodiment of the present disclosure;

[0031] Figure 9 is a schematic block diagram of a system for processing a chorus audio according to an embodiment of the present disclosure;

25 **[0032]** Figure 10 is a specific schematic structural diagram of a system for processing a chorus audio according to an embodiment of the present disclosure;

[0033] Figure 11 is a schematic structural diagram of an apparatus for processing a chorus audio according to an embodiment of the present disclosure; and

[0034] Figure 12 is a schematic structural diagram of a device for processing a chorus audio according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0035] A method for processing a chorus audio is provided in the present disclosure. After the multiple dry audios from performers singing a same target song are obtained, the time alignment is performed on the multiple dry audios. The virtual sound image localization is performed on the multiple dry audios after the time alignment to locate the multiple dry audios to multiple virtual sound images. The virtual sound images are located in the virtual sound image coordinate system centered on the human head. A distance from each virtual sound image to the coordinate origin is within the set distance range. The virtual sound images surround the human ears. The chorus audio is generated based on the multiple dry audios after the virtual sound image localization. In a case that a lead audio based on the target song is obtained, the lead audio, the chorus audio and a corresponding accompaniment are synthesized to output a chorus-effect audio. Since multiple dry audios are located to the virtual sound images surrounding the human ears, the generated chorus audio has a sound field with surround audio effect. In a sense of hearing, it can effectively prevent the sound field of the final output chorus-effect audio from locating on the center of the head causing the in-head effect. As a result, the sound field is wider.

[0036] In practical application, the method provided in the embodiments of the present disclosure is applicable in various scenarios where a chorus effect is desired. The solutions may be specifically implemented through an interaction between a server and a client.

[0037] Examples are provided below. In a first scenario, the server may pre-obtain multiple dry audios from multiple performers (e.g., performers 1, 2, 3, 4, and the like) singing a same target song; perform time alignment on the multiple dry audios; perform virtual sound image localization on the multiple dry audios after the time alignment, to locate the multiple dry audios to multiple virtual sound images which may surround human ears; and generate a chorus audio based on the multiple dry audios after the virtual sound image localization. In a case that a user X desires to achieve a chorus effect for the song sung by the user X, the user X may sing the target song through the client. The server obtains a lead audio sung by the user X through the client, synthesizes the lead audio, the chorus audio, and an accompaniment, to obtain a chorus-effect audio. The chorus-effect audio is outputted through the client, to enable the user X to experience the chorus effect.

[0038] In a second scenario, several friends (e.g., users 1, 2, 3, 4, and 5) sing a target song in a same time period but in different spaces, expecting a chorus effect. From the perspective of any of the users, the current user may be a lead singer. For example, from the perspective of user 1, the server may obtain dry audios from user 2, user 3, user 4 and user 5 singing the target song, perform the time alignment on the dry audios, locate the dry audios after the time alignment to multiple virtual sound images which surround human ears, and generate a chorus audio based on the dry audios after the virtual sound image localization. In a case that a lead audio from the user 1 singing the target song through the client is obtained by the server, the server synthesizes the lead audio, the chorus audio and the accompaniment to obtain a chorus-effect audio. The chorus-effect audio is outputted to the user 1 through the client. In this way, the user 1 can experience the chorus effect.

[0039] Described above are only exemplary application scenarios. In practical application, the technical solution of the present disclosure may also be applicable to, for example, multi-user chorus, multi-user band, and other sound effect processing scenarios.

[0040] In order to enable those skilled in the art to better understand the solutions in the present disclosure, the present disclosure is described in further detail below in conjunction with the accompanying drawings and specific embodiments. Apparently, the embodiments described below are only some embodiments, rather than all the embodiments of the present disclosure. Any other embodiments obtained by those ordinary skilled in the art based on the embodiments in the present disclosure without any creative effort shall fall within the protection scope of the present disclosure.

[0041] Reference is made to Figure 1, Figure 1 is a flow chart of implementing a method for processing a chorus audio according to an embodiment of the present disclosure. The method includes the following steps S110 to S150.

[0042] In S110, multiple dry audios are obtained respectively from multiple performers singing a same target song.

[0043] In an embodiment of the present disclosure, multiple dry audios may be obtained based on actual requirements. The multiple dry audios may be audio data obtained from different performers singing a same target song. The different performers may be in a same or different environment.

[0044] In S120, the time alignment is performed on the obtained multiple dry audios.

[0045] The multiple dry audios are obtained respectively from multiple performers singing the same target song. Since the multiple dry audios may be produced by different performers in different time periods, there may be a misalignment, such as a delay. In order to achieve a better chorus effect, the multiple dry audios, first, may be subjected to the time alignment processing. By doing so, the multiple dry audios after the time alignment are not serious ahead or behind the beat, such as being ahead or behind the beat more than one second. In an implementation, the obtained multiple dry audios may be time aligned at a same start position by using an alignment tool.

[0046] In a specific embodiment of the present disclosure, before performing time alignment on the obtained multiple dry audios, a preliminary sorting may be performed on the obtained multiple dry audios. For example, audio(s) with a poor sound quality, such as audio(s) with a noise, accompaniment backbeat, short duration, small audio energy, or distortion are eliminated by sorting via tools such as sound quality detection sorted. Then, the time alignment and subsequent processing are performed on the dry audios pass the preliminary sorting.

[0047] In S130, virtual sound image localization is performed on the multiple dry audios after the time alignment, to locate the multiple dry audios to multiple virtual sound images.

[0048] Where, the virtual sound images are in a preconfigured virtual sound image coordinate system. The human head is configured as a center of the virtual sound image coordinate system. A midpoint of a line segment where a left ear and a right ear are located is configured as a coordinate origin. A positive direction of a first axis represents a straight front direction of the human head. A positive direction of a second axis represents a lateral direction of the human head from the left ear to the right ear. A positive direction of a third axis represents a straight above direction of the human head. A distance between each virtual sound image and the coordinate origin is within a set distance range. An elevation angle of each virtual sound image relative to a plane formed by the first axis and the second axis is within a set angle range.

[0049] In an embodiment of the present disclosure, the virtual sound image coordinate system may be configured in advance for demonstrating orientations of sound images. In an implementation, the virtual sound image coordinate system may be a Cartesian coordinate

system. Reference is made to Figure 2. Where, the human head may be configured as a center of the virtual sound image coordinate system. A midpoint of a line segment where a left ear and a right ear are located is configured as a coordinate origin. A positive direction of a first axis (i.e., x-axis) represents a straight front direction of the human head. A positive direction of a second axis (i.e., y-axis) represents a lateral direction of the human head from the left ear to the right ear. A positive direction of a third axis (i.e., z-axis) represents a straight above direction of the human head, that is, a direction towards the top of the human head. Each sound image in the space has an azimuth and an elevation, which may be represented by (θ, ϕ) , where rad represents a distance between the current sound image and the coordinate origin.

[0050] Generally, an acoustic signal is a single-channel signal, which may be regarded as having a sound image at a position $(\theta, \phi)=(0^\circ, 0^\circ)$. In order to obtain a virtual sound image, a head related transfer function (HRTF) may be applied to perform data convolution to perform image localization. Reference is made to Figure 3, which is a schematic diagram of virtual sound image localization. As shown in Figure 3, X represents a real sound source (a single-channel signal), Y_L represents an acoustic signal received by the left ear, Y_R represents an acoustic signal received by the right ear, and HRTF represents a transfer function of a transmission path for an acoustic signal from a position of a sound source to the ears. Based on the HRTF technology, the real sound source (i.e., single-channel signal) may be filtered through the HRTF of the left ear and the right ear at a position (θ, ϕ) , to obtain a dual-channel acoustic signal.

[0051] Frequency domain characteristics of the acoustic signals received by the left ear and the right ear may be expressed as:

$$\begin{cases} Y_L = X \cdot H_L \\ Y_R = X \cdot H_R \end{cases}$$

[0052] It can be simply considered that the acoustic signal heard by the human ears is a result of the sound source X filtered through the HRTF. Therefore, when performing the virtual sound image localization, the acoustic signals are filtered through the HRTF at a corresponding position. Multiple virtual sound images may be set in the virtual sound image coordinate system. The distance between each virtual sound image and the coordinate origin

may be within a set distance range, such as within a range of one meter. An elevation angle of each virtual sound image relative to the plane formed by the first axis and the second axis of the virtual sound-image coordinate system may be within a set angle range, such as within a range of 10° , to enable the multiple virtual sound images to surround the human ears.

5 **[0053]** In an implementation, the virtual sound images may be evenly distributed in a full circle on the plane formed by the first axis and the second axis. In other words, the virtual sound images circle a horizontal plane where the human ears are located at a same interval angle. The interval angle may be set based on an actual situation or analysis of history data, for example, set to 30° . In a case that the interval angle is set to 30° , twelve virtual sound
10 images may be located, at intervals of 30° , around the horizontal plane where the human ears are located. The twelve virtual sound images each have an elevation angle of 0° , and azimuth angles of 0° , 30° , 60° , ..., 330° , respectively. The interval angle may be set to other values, such as 15° or 60° .

[0054] In another embodiment, among the virtual sound images, an elevation angle of a
15 virtual sound image located behind the human head relative to the plane formed by the first axis and the second axis may be greater than an elevation angle of a virtual sound image located in front of the head relative to the plane formed by the first axis and the second axis. In other words, among the multiple virtual sound images, the elevation angle of a virtual sound image located behind the head may be greater than the elevation angle of a virtual
20 sound image located in front of the head. In this way, it can enhance a localization effect and reduce a front-rear mirroring problem of virtual sound images. For example, the elevation angle of a virtual sound image located behind the head may be increased by 10° . In other words, the virtual sound image located in front of the head has an elevation angle $\theta=0^\circ$, and the virtual sound image located behind the head has an elevation angle $\theta=10^\circ$.

25 **[0055]** As shown in Figure 4, multiple virtual sound images circle the horizontal plane where the ears are located by intervals of 30° . The virtual sound image located in front of the head has an elevation angle $\theta=0^\circ$, and the virtual sound image located behind the head has an elevation angle $\theta=10^\circ$.

[0056] It should be noted that locations of the virtual sound images set in the virtual sound
30 image coordinate system are not limited to what mentioned above, and may be set based on an actual need, as long as the distance between each virtual sound image and the coordinate

origin is within the set distance range and the elevation angle of each virtual sound image relative to the plane formed by the first axis and the second axis is within the set angle range. For example, among the virtual sound images, a first part of the virtual sound images circle the plane where the human ears are located at intervals of 30° , and each has an elevation angle of 0° ; and a second part of the virtual sound images circle the plane where the human ears are located with intervals of 60° , and each has an elevation angle of 10° . The distance between the first part of the virtual sound images and the coordinate origin and the distance between the second part of the virtual sound images and the coordinate origin may be the same or different, yet those distances are all within the set distance range. In this way, a surround effect of the generated chorus audio is enhanced.

[0057] After the virtual sound image localization is performed on the multiple dry audios after the time alignment, the multiple dry audios are located to multiple virtual sound images. The method may proceed to the following steps.

[0058] In S140, a chorus audio is generated based on the multiple dry audios after the virtual sound image localization.

[0059] Subsequent to the steps of obtaining multiple dry audios respectively from a multiple performer singing a same target song, performing the time alignment on the dry audios, and performing the virtual sound image localization on the dry audios after the time alignment to locate the multiple dry audios to multiple virtual sound images, an HRTF filtering may be performed on each of the multiple dry audios at a corresponding location of a virtual sound image, to obtain corresponding audio data at each of the virtual sound images. The chorus audio may be generated based on the multiple dry audios after the virtual sound image localization. In an implementation, the audio data obtained after the HRTF filtering of locations of the multiple virtual sound images may be superimposed or weighted superimposed to obtain the chorus audio. The chorus audio has a sound effect of a three-dimensional sound field.

[0060] In S150, in a case that a lead audio based on the target song is obtained, the lead audio, the chorus audio and an accompaniment are synthesized to output a chorus-effect audio.

[0061] In an application scenario according to an embodiment of the present disclosure, the chorus audio, after generated, may be stored in a database and used as required. For example,

in a case that a user expects a chorus effect of the song sung by the user, the chorus audio may be used to achieve such effect.

[0062] An audio from a current user singing based on the target song may be obtained as the lead audio. Then, the lead audio, the chorus audio and the accompaniment are synthesized to obtain the chorus-effect audio. The chorus-effect audio is outputted so that the current user can enjoy the chorus effect.

[0063] Synthesis of the lead audio, the chorus audio and the accompaniment may be achieved in a variety of manners. For example, the lead audio and the corresponding accompaniment may be firstly synthesized with each other, and then synthesized with the chorus audio. Alternatively, the chorus audio and the accompaniment may be firstly synthesized with each other, and then synthesized with the lead audio. Alternatively, the lead audio and the chorus audio may be firstly synthesized with each other, and then synthesized with the accompaniment. For example, the lead audio and the chorus audio may be balanced and then superimposed with the accompaniment based on a set vocal-to-accompaniment ratio. The chorus effects obtained through different implementations may be different, and a specific implementation thereof may be determined based on an actual situation.

[0064] With the method provided in the embodiments of the present disclosure, after the dry audios from multiple performers singing the same target song are obtained respectively, the time alignment is performed on the multiple dry audios. The virtual sound image localization is performed on the multiple dry audios after the time alignment to locate the multiple dry audios to multiple virtual sound images. The virtual sound images are located in the virtual sound image coordinate system where the center is the human head. A distance from each virtual sound image to the coordinate origin is within the set distance range. The virtual sound images surround the human ears. The chorus audio is generated based on the multiple dry audios after the virtual sound image localization. In a case that the lead audio based on the target song is obtained, the lead audio, the chorus audio and the accompaniment are synthesized to output the chorus-effect audio. Since the multiple dry audios are located to the virtual sound images surrounding the human ears, it can effectively prevent the sound field of the final output chorus-effect audio from locating on the center of the human head causing the in-head effect. As a result, the sound field is wider.

[0065] In an embodiment of the present disclosure, the S120 of performing time alignment

on the multiple dry audios includes the following steps.

[0066] In a first step, a reference audio corresponding to the target song is determined.

[0067] In a second step, for each obtained dry audio, an audio feature of the current dry audio and an audio feature of the reference audio are extracted, where the audio feature is a fingerprint feature or fundamental frequency feature.

[0068] In a third step, a time instant corresponding to a maximum similarity between the audio feature of the current dry audio and the audio feature of the reference audio is determined as an audio alignment time.

[0069] In a fourth step, the time alignment is performed on the current dry audio based on the audio alignment time.

[0070] For ease of description, the above steps are described in a combined manner.

[0071] In an embodiment of the present disclosure, after the dry audios from multiple performers singing a same target song are respectively obtained, the reference audio corresponding to the target song may be determined first in the process of performing time alignment on the multiple dry audios. In an implementation, a dry audio with a good sound quality may be determined from the multiple dry audios as the reference audio. Alternatively, an original dry audio of the target song may be determined as the reference audio.

[0072] For each of the dry audios, an audio feature of the current dry audio and an audio feature of the reference audio may be extracted. The audio feature is a fingerprint feature or fundamental frequency feature. For example, Mel band information, Bark band information, and Erb band power may be extracted through a multi-band filtering. Then, the fingerprint feature may be obtained through half-wave rectification, binary determination, and the like. For another example, a fundamental frequency feature may be extracted through a fundamental frequency extraction tool, such as pyin, crepe, and harvest and the like. The audio feature of the reference audio may be stored once extracted and invoked directly as required.

[0073] The audio feature of the current dry audio is compared with the audio feature of the reference audio. A time instant corresponding to a maximum similarity may be determined as an audio alignment time, based on a similarity curve and other characteristics. Subsequently, the time alignment is performed on the dry audio based on the audio alignment time.

[0074] Each of the dry audios is compared with the reference audio in terms of the audio feature to obtain the audio alignment time corresponding to the dry audio, and then the time alignment is performed on the dry audio. Thereby, multiple dry audios after the time alignment can be obtained.

5 **[0075]** In an embodiment of the present disclosure, the method may further include: each of the obtained dry audios is respectively subjected to the band-pass filtering to obtain bass data.

[0076] In this case, the generating a chorus audio based on the multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after the virtual sound image localization and the bass data.

10 **[0077]** In an embodiment of the present disclosure, after the multiple dry audios from performers singing the same target song are respectively obtained, the band-pass filtering may be performed on each of the obtained multiple dry audios to obtain the bass data respectively. For example, the band-pass with a cutoff frequency in a range of [33Hz, 523Hz] is performed on the multiple dry audios to obtain the multiple bass data.

15 **[0078]** The chorus audio may be generated based on the multiple dry audios after the virtual sound image localization and the bass data. In an implementation, a superimposition or weighted superimposition may be performed on the bass data and the dry audios after the virtual sound image localization, to generate the chorus audio. The superimposition of the bass signal may result in an enhanced heaviness of the sound signal.

20 **[0079]** In an embodiment of the present disclosure, the method may further include: performing reverberation simulation on the each of the obtained multiple dry audios respectively.

[0080] In this case, the generating a chorus audio based on the dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry
25 audios after the virtual sound image localization and the multiple dry audios after the reverberation simulation.

[0081] In general, an acoustic signal emitted by a sound source in a sound field goes through processes such as the direct sound, reflection, and reverberation. Figure 5 shows a schematic diagram of a conventional composition of a spatial sound field process. In Figure 5,
30 an acoustic signal with the maximum amplitude is the direct sound. An acoustic signal

following the direct sound is a reflected acoustic signal obtained from a reflection of a sound wave on an object closest to a listener. The reflected signal has obvious directionality. A dense acoustic signal following the reflected acoustic signal is a reverberation signal obtained from superimposition of sound waves due to multiple reflections from surrounding objects. The
5 reverberation signal is a superimposition of a large number of reflected sounds in different directions, and has no directionality.

[0082] Based on a known characteristic of room impulse response, the reverberation is the superimposition of multiple reflected sounds, which is characterized by weak energy and no directionality. This is because that the reverberation is a superimposition of a large number of
10 reflected sounds from different directions and has a high echo density. Hence, the reverberation may be utilized to realize a surround sound effect with listener envelopment.

[0083] In an embodiment of the present disclosure, the reverberation simulation may be performed respectively on each of the multiple dry audios, after the multiple dry audios are obtained respectively from multiple performers singing a same target song. In an
15 implementation, the reverberation simulation may be performed on the multiple dry audios by using a cascade of a comb filter and an all-pass filter.

[0084] Figure 6 shows a cascade of comb filters and all-pass filters. In Figure 6, four comb filters are connected in parallel, and the parallel-connected comb filters are connected in series with two all-pass filters. An actual reverberation impulse response obtained in the
20 simulation is as shown in Figure 7.

[0085] It should be noted that Figure 6 shows only an implementation. In practical application, other forms may be provided. The quantity and cascade manner of comb filters and all-pass filters may be determined based on an actual requirement.

[0086] After performing the reverberation simulation on each of the multiple dry audios
25 respectively and performing the virtual sound image localization for locating the multiple dry audios to multiple virtual sound images, the chorus audio may be generated based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the reverberation simulation. In an implementation, the chorus audio may be generated by performing a superimposition or weighted superimposition on the multiple dry audios after
30 the virtual sound image localization and the multiple dry audios after the reverberation simulation. In this way, a spatial sound effect of the acoustic signal can be enhanced, the

in-head effect is further suppressed, and the sound field is expanded.

[0087] In an embodiment of the present disclosure, after the performing virtual sound image localization on the multiple dry audios after the time alignment, the method may further include: performing the reverberation simulation respectively on each of the multiple dry audios after the virtual sound image localization.

[0088] In this case, the generating the chorus audio based on the multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after both the virtual sound image localization and the reverberation simulation.

[0089] In the embodiment of the present disclosure, after the steps of obtaining the dry audios respectively from multiple performers singing the same target song, performing the time alignment on the obtained multiple dry audios, and performing the virtual sound image localization on the dry audios after the time alignment, the reverberation simulation may be further performed on the multiple dry audios after the virtual sound image localization. The process of the reverberation simulation may be referred to the reverberation simulation in the previous embodiment, which is not repeated here.

[0090] The chorus audio may be generated based on the multiple dry audios after both the virtual sound image localization and the reverberation simulation. In an implementation, the chorus audio may be generated by performing a superimposition or weighted superimposition on the multiple dry audios after both the virtual sound image localization and the reverberation simulation.

[0091] The reverberation simulation on the multiple dry audios after the virtual sound image processing can enhance the spatial sound effect of the acoustic signal, which can further suppress the in-head effect and expand the sound field.

[0092] In an embodiment of the present disclosure, the method may further include: performing dual-channel simulation on each of the obtained multiple dry audios respectively.

[0093] In this case, the generating the chorus audio based on the multiple dry audios after the virtual sound image localization includes: generating the chorus audio based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation.

[0094] In an embodiment of the present disclosure, after the steps of obtaining the dry audios respectively from multiple performers singing the same target song, performing the time alignment on the obtained multiple dry audios, the dual-channel simulation may be performed on the multiple dry audios. A delay may be applied to reduce correlation of two channel signals, to expand the sound field as much as possible to obtain a dual-channel output.

[0095] As shown in Figure 8, the dual-channel simulation on the multiple dry audios may be realized through eight different weighted delays on the left channel and eight different weighted delays on the right channel, where d represents a delay and g represents a weight.

Since a room impulse response generally takes 80ms as a reverberation time, a delay parameter may select sixteen different parameters ranging from 21ms to 79ms. An amplitude attenuation is for representing an energy loss caused by reflection of a sound wave, which can reduce correlation between environmental information from two channels. To put it another way, the dry audios may be copied to obtain two signals carrying the same information. The two signals are completely correlated. Then, different delays and amplitudes are applied to realize attenuation. Thereby, the correlation of the two signals is reduced and a pseudo stereo signal is obtained.

[0096] It should be noted that Figure 8 shows only a specific example. Fewer or more sets of different delays may be set based on an actual requirement for the dual-channel simulation.

[0097] The chorus audio may be generated based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation. In an implementation, the chorus audio may be generated by performing a superimposition or weighted superimposition on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation.

[0098] In an embodiment of the present disclosure, after the performing the dual-channel simulation on each of the obtained multiple dry audios respectively, the method further includes: performing reverberation simulation on the multiple dry audios after the dual-channel simulation.

[0099] In this case, the generating the chorus audio based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation includes: generating the chorus audio based on the multiple dry audios after the

virtual sound image localization and the multiple dry audios after both the dual-channel simulation and the reverberation simulation.

[0100] In the embodiment of the present disclosure, after the steps of obtaining the dry audios respectively from multiple performers singing the same target song, performing the time alignment on the multiple dry audios, and performing the dual-channel simulation on each of the multiple dry audios respectively, the reverberation simulation is further performed on the multiple dry audios after the dual-channel simulation. Thereby, a spatial effect of the acoustic signal is enhanced, the in-head effect is suppressed, and the sound field is expanded.

[0101] After performing virtual sound image localization on the multiple dry audios for locating the dry audios to multiple virtual sound images, the chorus audio may be generated based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after both the dual-channel simulation and the reverberation simulation. In an implementation, the chorus audio may be generated by performing a superimposition or weighted superimposition on the multiple dry audios after the virtual sound image localization and the multiple dry audios after both the dual-channel simulation and the reverberation simulation.

[0102] In practical application, after the dry audios from multiple performers singing the same target song are obtained, the time alignment may be first performed on the multiple dry audios, and then the virtual sound image localization, the bass enhancement, the reverberation simulation, the dual-channel simulation and other processing may be performed on multiple dry audios after the time alignment. The specific processing may be performed based on the above embodiments. Through the virtual sound image localization, the bass enhancement, the reverberation simulation and the dual-channel simulation, the eventually generated chorus audio is of a sound field surround effect, which presents a high robustness to misalignment of sounds in a large range. In a case that the chorus audio is to be superimposed with the lead audio, a harmonious auditory experience can be ensured for a user even if the lead audio has a large delay.

[0103] Figure 9 shows a schematic block diagram of a system for processing multiple dry audios after the time alignment. The system includes a bass enhancement unit, a virtual sound image localization unit, a dual-channel simulation unit and a reverberation simulation unit. The bass enhancement unit is configured to perform band-pass filtering on the multiple dry

audios to obtain bass data. The virtual sound image localization unit is configured to perform virtual sound image localization on the multiple dry audios to locate the multiple dry audios to multiple virtual sound images. The dual-channel simulation unit is configured to perform dual-channel simulation on the multiple dry audios. The reverberation simulation unit is configured to perform reverberation simulation on the multiple dry audios. Both the virtual sound image localization unit and the dual-channel simulation unit may be connected to the reverberation simulation unit. After the virtual sound image localization on the multiple dry audios through the virtual sound image localization unit, the reverberation simulation may be further performed through the reverberation simulation unit. Similarly, after the dual-channel simulation on the multiple dry audios through the dual-channel simulation unit, the reverberation simulation may be further performed through the reverberation simulation unit. Finally, a weighted superimposition may be performed on audio data processed through these units to obtain the chorus audio.

[0104] Figure 10 shows a specific example of processing multiple dry audios. In Figure 10, H represents a transfer function for HRTF filtering. The virtual sound image localization may be performed on the multiple dry audios through the transfer function to locate the multiple dry audios to twelve virtual sound images surrounding a horizontal plane where the human ears are located. In addition, REV represents a reverberation simulation unit, BASS represents a bass enhancement unit, and REF represents dual-channel simulation unit. The reverberation simulation units here may use the same parameter. Alternatively, different parameters may be configured for different reverberation simulation units based on actual requirements, to obtain a flexible reverberation modulation.

[0105] According to the embodiment of the present disclosure, the chorus effect of the eventually generated chorus audio is closer to a listening feeling of a real concert chorus. In practical application, the accompaniment is added to the lead audio and then mixed with the chorus audio. In this way, the user is provided with an immersive sensation of a concert, realizing an experience of being surrounded in a more immersive sound field.

[0106] In an embodiment of the present disclosure, the performing virtual sound image localization on the multiple dry audios after the time alignment may include:

In step 1, grouping the multiple dry audios after the time alignment based on a quantity of virtual sound images, where a quantity of groups is equal to the quantity of the

virtual sound images; and

In step 2, locating each group of dry audios respectively to one of the virtual sound images, so that each group of dry audios correspond to a different one of the virtual sound images.

5 **[0107]** For ease of description, the above steps are described in a combined manner.

[0108] In an embodiment of the present disclosure, after the steps of obtaining the dry audios respectively from multiple performers singing the same target song and performing the time alignment on the obtained multiple dry audios, the multiple dry audios after the time alignment may be grouped based on a quantity of the virtual sound images. The quantity of groups is equal to the quantity of the virtual sound images. Each group includes several dry audio. In a case of a large quantity of the dry audios, a same dry audio may be in only one group. In a case of a small quantity of the dry audios, a same dry audio may be in multiple groups for a better chorus effect.

15 **[0109]** After the multiple dry audios are grouped, each group of dry audios is located to a corresponding one of the virtual sound images, to match groups of dry audios to different ones of the virtual sound images. The localization of the virtual sound images of the dry audios is realized to enhance the chorus effect.

[0110] In an embodiment of the present disclosure, the synthesizing the lead audio, the chorus audio and an accompaniment may include:

20 performing volume adjustment on the lead audio and the chorus audio respectively, and/or performing reverberation simulation on the lead audio and the chorus audio; and

synthesizing the accompaniment with the lead audio and the chorus audio after the volume adjustment and/or the reverberation simulation.

25 **[0111]** In an embodiment of the present disclosure, after the lead audio based on the target song is obtained, the volume adjustment may be performed on the lead audio and the chorus audio respectively. In this way, a volume of the lead audio is equal to or greater than a volume of the chorus audio. In addition, the reverberation simulation may be further performed on the lead audio and the chorus audio to obtain a surround sound effect with listener envelopment.

30 **[0112]** The lead audio and the chorus audio after the volume adjustment and/or the reverberation simulation are then synthesized with the accompaniment, so that the eventually

outputted chorus-effect audio provides a better experience to the user in terms of a sense of hearing.

[0113] Corresponding to the above method embodiments, an apparatus for processing a chorus audio is further provided according to the embodiments of the present disclosure. The apparatus for processing a chorus audio described below and the method for processing a chorus audio described above may be referenced to each other.

[0114] Reference is made to Figure 11. The apparatus may include a dry audio obtaining module 1110, a time alignment module 1120, a virtual sound image localization unit 1130, a chorus audio generating module 1140, and a chorus-effect audio obtaining module 1150.

[0115] The dry audio obtaining module 1110 is configured to obtain multiple dry audios respectively from multiple performers singing a same target song.

[0116] The time alignment module 1120 is configured to perform time alignment on the obtained multiple dry audios.

[0117] The virtual sound image localization module 1130 is configured to perform virtual sound image localization on the multiple dry audios after the time alignment, to locate the dry audios to multiple virtual sound images. The virtual sound images are in a pre-configured virtual sound-image coordinate system. In the virtual sound image coordinate system, a human head is configured as a center of the virtual sound image coordinate system; a midpoint of a line segment where a left ear and a right ear are located is a coordinate origin; a positive direction of a first axis represents a straight front direction of the human head; a positive direction of a second axis represents a lateral direction of the human head from the left ear to the right ear, and a positive direction of a third axis represents a straight above direction of the human head; a distance between each virtual sound image and the coordinate origin is within a set distance range, and an elevation angle of each virtual sound image relative to a plane formed by the first axis and the second axis is within a set angle range.

[0118] The chorus audio generating module 1140 is configured to generate a chorus audio based on the multiple dry audios after the virtual sound image localization.

[0119] The chorus-effect audio obtaining module 1150 is configured to synthesize, in a case that a lead audio based on the target song is obtained, the lead audio, the chorus audio and an accompaniment, to output a chorus-effect audio.

[0120] With the apparatus provided in the embodiments of the present disclosure, after the dry audios from multiple performers singing the same target song are obtained respectively, the time alignment is performed on the multiple dry audios. The virtual sound image localization is performed on the multiple dry audios after the time alignment to locate the multiple dry audios to multiple virtual sound images. The virtual sound images are located in the virtual sound image coordinate system where the center is the human head. A distance from each virtual sound image to the coordinate origin is within the set distance range. The virtual sound images surround the human ears. The chorus audio is generated based on the multiple dry audios after the virtual sound image localization. In a case that the lead audio based on the target song is obtained, the lead audio, the chorus audio and the accompaniment are synthesized to output the chorus-effect audio. Since the multiple dry audios are located to the virtual sound images surrounding the human ears, the generated chorus audio has a sound field with the surround sound effect. In a sense of hearing, it can effectively prevent the sound field of the final output chorus-effect audio from locating on the center of the head causing the in-head effect. As a result, the sound field is wider.

[0121] In an embodiment of the present disclosure, the time alignment module 1120 is configured to: determine a reference audio corresponding to the target song; for each of the dry audios, extract an audio feature of the dry audio and an audio feature of the reference audio, where the audio feature is a fingerprint feature or fundamental frequency feature; determine a time instant corresponding to a maximum similarity between the audio feature of the dry audio and the audio feature of the reference audio as an audio alignment time; and perform the time alignment on the dry audio based on the audio alignment time.

[0122] In an embodiment of the present disclosure, the apparatus further includes a bass data obtaining module, which is configured to: perform band-pass filtering on the each of the multiple dry audios respectively to obtain bass data.

[0123] In this case, the chorus audio generating module 1140 is configured to: generate the chorus audio based on the multiple dry audios after the virtual sound image localization and the bass data.

[0124] In an embodiment of the present disclosure, the apparatus further includes a reverberation simulation module, which is configured to: perform reverberation simulation on the each of the multiple dry audios respectively.

[0125] In this case, the chorus audio generating module 1140 is configured to: generate the chorus audio based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the reverberation simulation.

[0126] In an embodiment of the present disclosure, the reverberation simulation module is configured to: perform the reverberation simulation respectively on each of the multiple dry audios by using a cascade of a comb filter and an all-pass filter.

[0127] In an embodiment of the present disclosure, the reverberation simulation module is further configured to: after the virtual sound image localization is performed on the multiple dry audios after the time alignment, perform reverberation simulation respectively on each of the multiple dry audios after the virtual sound image localization.

[0128] In this case, the chorus audio generating module 1140 is configured to: generate a chorus audio based on the multiple dry audios after both the virtual sound image localization and the reverberation simulation.

[0129] In an embodiment of the present disclosure, the apparatus further includes a dual-channel simulation module, which is configured to: perform dual-channel simulation on each of the obtained multiple dry audios respectively.

[0130] In this case, the chorus audio generating module 1140 is configured to: generate the chorus audio based on the multiple dry audios after the virtual sound image localization and the multiple dry audios after the dual-channel simulation.

[0131] In an embodiment of the present disclosure, the reverberation simulation module is further configured to: perform, after the performing the dual-channel simulation on each of the obtained multiple dry audios respectively, the reverberation simulation on the multiple dry audios after the dual-channel simulation.

[0132] In this case, the chorus audio generating module 1140 is configured to: generate the chorus audio based on the multiple dry audios after the virtual sound image localization and the dry audios after both the dual-channel simulation and the reverberation simulation.

[0133] In an embodiment of the present disclosure, the virtual sound image localization module 1130 is configured to: group the multiple dry audios after the time alignment based on a quantity of virtual sound images, where a quantity of groups is equal to the quantity of the virtual sound images; and locate each group of dry audios respectively to one of the virtual

sound images, so that each group of dry audios correspond to a different one of the virtual sound images.

[0134] In another embodiment of the present disclosure, among the virtual sound images, an elevation angle of a virtual sound image located behind the human head relative to the plane formed by the first axis and the second axis is greater than an elevation angle of a virtual sound image located in front of the human head relative to the plane formed by the first axis and the second axis; or the virtual sound images are evenly distributed in a full circle on the plane formed by the first axis and the second axis.

[0135] In an embodiment of the present disclosure, the chorus-effect audio obtaining module 1150 is configured to: perform volume adjustment on the lead audio and the chorus audio respectively, and/or perform reverberation simulation on the lead audio and the chorus audio; and synthesize the accompaniment with the lead audio and the chorus audio after the volume adjustment and/or the reverberation simulation.

[0136] Corresponding to the above method embodiments, a device for processing a chorus audio is further provided in an embodiment of the present disclosure. The device includes: a memory, configured to store a computer program; and a processor, configured to implement, when executing the computer program, the method for processing a chorus audio.

[0137] Reference is made to Figure 12, which is a schematic structural diagram of a device for processing a chorus audio. The device may include: a processor 10, a memory 11, a communication interface 12 and a communication bus 13. The processor 10, the memory 11, and the communication interface 12 communicate with each other through the communication bus 13.

[0138] In an embodiment of the present disclosure, the processor 10 may be a central processing unit (CPU), an application-specific integrated circuit, a digital signal processor, a field programmable gate array, or other programmable logic devices.

[0139] The processor 10 may be configured to invoke the program stored in the memory 11. Specifically, the processor 10 can perform operations in the method for processing a chorus audio according to the embodiments of the present disclosure.

[0140] The memory 11 is configured to store one or more programs. The program may include program codes. The program codes include computer operation instructions. In an

embodiment of the present disclosure, the memory 11 stores at least a program for implementing the following functions:

obtaining multiple dry audios respectively from multiple performers singing a same target song;

5 performing time alignment on the multiple dry audios;

performing virtual sound image localization on the multiple dry audios after the time alignment, to locate the dry audios to multiple virtual sound images; where the multiple virtual sound images are located in a pre-configured virtual sound image coordinate system; where the virtual sound image coordinate system is configured as: a human head is configured
10 as a center of the virtual sound image coordinate system; a midpoint of a line segment where a left ear and a right ear are located is a coordinate origin; a positive direction of a first axis represents a straight front direction of the human head; a positive direction of a second axis represents a lateral direction of the human head from the left ear to the right ear, and a positive direction of a third axis represents a straight above direction of the human head; a
15 distance between each virtual sound image and the coordinate origin is within a set distance range, and an elevation angle of each virtual sound image relative to a plane formed by the first axis and the second axis is within a set angle range;

generating a chorus audio based on the multiple dry audios after the virtual sound image localization; and

20 synthesizing, in a case that a lead audio based on the target song is obtained, the lead audio, the chorus audio and an accompaniment, to output a chorus-effect audio.

[0141] In a possible implementation, the memory 11 may include a program storage area and a data storage area. The program storage area may store an operating system and an application program required for at least one function (such as an audio playback function and
25 an audio synthesis function). The storage data area may store data created during use, such as sound image localization data, audio synthesis data, and the like.

[0142] In addition, the memory 11 may include a high-speed random-access memory, and may further include a non-volatile memory, such as at least one magnetic disk storage device, or other volatile solid storage device.

30 **[0143]** The communication interface 12 may be an interface of a communication module,

and is used to be connected to other devices or systems.

[0144] It should be noted that the structure shown in Figure 12 does not constitute a limitation on the device for processing a chorus audio in the embodiments of the present disclosure. In practical application, the device may include more or fewer components than
5 what is shown in Figure 12, or may include a combination of some components.

[0145] Corresponding to the method embodiments, a computer-readable storage medium storing a computer program is provided. The computer program, when executed by a processor, implements the steps of the method for processing a chorus audio above.

[0146] The embodiments in the present disclosure are described in a progressive manner,
10 and each of the embodiments focuses on its differences from the other embodiments. The same or similar parts among the embodiments may be referred to each other.

[0147] Those skilled in the art may further understand that, units and algorithm steps described in conjunction with the embodiments disclosed herein may be realized by electronic hardware, computer software or a combination thereof. In order to clearly describe
15 interchangeability of the hardware and the software, the composition and steps of each embodiment are generally described above based on functions. Whether a function is to be implemented by hardware or software depends on a particular application of the technical solutions and design constraints. Those skilled in the art may implement the described functions through different methods for each specific application. Such implementation
20 should not be considered beyond the scope of the present disclosure.

[0148] The steps of the method or algorithm described in conjunction with the embodiments disclosed herein may be directly implemented by hardware, a software module executed by a processor, or a combination thereof. The software module may be disposed in a random-access memory (RAM), a memory, a read only memory (ROM), an electrically
25 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 art.

[0149] The principles and implementations of the present disclosure are clarified with specific examples herein. The above description of the embodiments is only intended to assist understanding the method and the key concept of the present disclosure. It should be noted
30 that improvements and modifications can be made to the present disclosure by those of

ordinary skills in the art, without departing from the principles of the present disclosure. Such improvements and modifications fall within the protection scope of the claims in the present disclosure.

CLAIMS

1. A method for processing a chorus audio, comprising:

obtaining a plurality of dry audios respectively from a plurality of performers singing a
5 same target song;

performing time alignment on the obtained plurality of dry audios;

performing virtual sound image localization on the plurality of dry audios after the time
alignment, to locate the dry audios to a plurality of virtual sound images; wherein the plurality
of virtual sound images are located in a pre-configured virtual sound image coordinate system;

10 wherein, the virtual sound image coordinate system is configured as: a human head is a center
of the virtual sound image coordinate system; a midpoint of a line segment where a left ear
and a right ear are located is a coordinate origin; a positive direction of a first axis represents a
straight front direction of the human head; a positive direction of a second axis represents a
lateral direction of the human head from the left ear to the right ear, and a positive direction of
15 a third axis represents a straight above direction of the human head; a distance between each
virtual sound image and the coordinate origin is within a set distance range, and an elevation
angle of each virtual sound image relative to a plane formed by the first axis and the second
axis is within a set angle range;

generating a chorus audio based on the plurality of dry audios after the virtual sound
20 image localization; and

synthesizing, in a case that a lead audio based on the target song is obtained, the lead
audio, the chorus audio and an accompaniment, to output a chorus-effect audio.

2. The method according to claim 1, wherein the performing time alignment on the
25 obtained plurality of dry audios comprises:

determining a reference audio corresponding to the target song;

for each of the obtained plurality of dry audios, extracting an audio feature of the dry
audio and an audio feature of the reference audio, wherein the audio feature is a fingerprint
feature or fundamental frequency feature;

determining a time instant corresponding to a maximum similarity between the audio feature of the dry audio and the audio feature of the reference audio as an audio alignment time; and

performing the time alignment on the dry audio based on the audio alignment time.

5

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

performing band-pass filtering on the each of the plurality of dry audios respectively, to obtain bass data; and

the generating the chorus audio based on the plurality of dry audios after the virtual sound image localization comprises:

generating the chorus audio based on the plurality of dry audios after the virtual sound image localization and the bass data.

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

performing reverberation simulation on the each of the obtained plurality of dry audios respectively; and

the generating the chorus audio based on the plurality of dry audios after the virtual sound image localization comprises:

generating the chorus audio based on the plurality of dry audios after the virtual sound image localization and the plurality of dry audios after the reverberation simulation.

5. The method according to claim 4, wherein the performing the reverberation simulation on each of the obtained plurality of dry audios respectively comprises:

performing the reverberation simulation on each of the obtained plurality of dry audios respectively by using a cascade of a comb filter and an all-pass filter.

6. The method according to claim 1, wherein after the performing virtual sound image localization on the plurality of dry audios after the time alignment, the method further

comprises:

performing reverberation simulation respectively on each of the plurality of dry audios after the virtual sound image localization; and

the generating the chorus audio based on the plurality of dry audios after the virtual sound image localization comprises:

generating the chorus audio based on the plurality of dry audios after both the virtual sound image localization and the reverberation simulation.

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

performing dual-channel simulation on each of the obtained plurality of dry audios respectively; and

the generating the chorus audio based on the plurality of dry audios after the virtual sound image localization comprises:

generating the chorus audio based on the plurality of dry audios after the virtual sound image localization and the plurality of dry audios after the dual-channel simulation.

8. The method according to claim 7, wherein after the performing the dual-channel simulation on each of the obtained plurality of dry audios respectively, the method further comprises:

performing reverberation simulation on the plurality of dry audios after the dual-channel simulation; and

the generating the chorus audio based on the plurality of dry audios after the virtual sound image localization and the plurality of dry audios after the dual-channel simulation comprises:

generating the chorus audio based on the plurality of dry audios after the virtual sound image localization and the plurality of dry audios after both the dual-channel simulation and the reverberation simulation.

9. The method according to claim 1, wherein the performing virtual sound image localization on the plurality of dry audios after the time alignment comprises:

grouping the plurality of dry audios after the time alignment based on a quantity of virtual sound images, wherein a quantity of groups is equal to the quantity of the virtual sound images; and

locating each group of dry audios respectively to one of the virtual sound images, so that each group of dry audios correspond to a different one of the virtual sound images.

10. The method according to claim 1, wherein,

among the virtual sound images, an elevation angle of a virtual sound image located behind the human head relative to the plane formed by the first axis and the second axis is greater than an elevation angle of a virtual sound image located in front of the human head relative to the plane formed by the first axis and the second axis; or

the virtual sound images are evenly distributed in a full circle on the plane formed by the first axis and the second axis.

11. The method according to any one of claims 1 to 10, wherein the synthesizing the lead audio, the chorus audio and the accompaniment comprises:

performing volume adjustment on the lead audio and the chorus audio respectively, and/or performing reverberation simulation on the lead audio and the chorus audio; and

synthesizing the accompaniment with the lead audio and the chorus audio after the volume adjustment and/or the reverberation simulation.

12. A device for processing a chorus audio, comprising:

a memory, configured to store a computer program; and

a processor, configured to implement, when executing the computer program, the method for processing a chorus audio according to any of claims 1 to 11.

13. A computer-readable storage medium storing a computer program, wherein the computer program, when executed by a processor, implements the method for processing a chorus audio according to any of claims 1 to 11.