

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY

(19) World Intellectual Property
Organization International Bureau
(43) International Publication Date:
March 24, 2022 (24.03.2022)



(10) International Publication No.:
WO 2022/057759 A1

(51) International Patent Classification No.:

G10L 13/02 (2013.01) *G10L 15/26* (2006.01)
G10L 13/033 (2013.01) *G10L 25/18* (2013.01)
G10L 13/04 (2013.01) *G10L 25/30* (2013.01)
G10L 15/02 (2006.01) *H04M 1/72433* (2021.01)
G10L 15/06 (2013.01) *H04M 1/72403* (2021.01)
G10L 15/16 (2006.01)

(21) International Application No.: PCT/CN2021/117945

(22) International Filing Date:

September 13, 2021 (13.09.2021)

(25) Filing Language: Chinese

(26) Publication Language: Chinese

(30) Priority:

202010996501.5

September 21, 2020 (21.09.2020) CN

(71) Applicant: HUAWEI Technologies Co., LTD. [CN/CN];
Huawei Headquarters Office Building, Bantian,
Longgang District, Shenzhen, Guangdong Province,
China, Guangdong 518129 (CN).

(72) Inventors: FAN Fan; Huawei Headquarters Office
Building, Bantian, Longgang District, Shenzhen,
Guangdong Province, China, Guangdong 518129
(CN). LUO Jinghao; Huawei Headquarters Office
Building, Bantian, Longgang District, Shenzhen,
Guangdong Province, China, Guangdong 518129
(CN). LI Shuo; Huawei Headquarters Office
Building, Bantian, Longgang District, Shenzhen,
Guangdong Province, China, Guangdong 518129
(CN).

(74) Agent: SHENPAT INTELLECTUAL PROPERTY
AGENCY (General Partnership), 18C2, 18D, 18E,
18E2, Block B, Lushan Building, Chunfeng Road,
Nanhu Subdistrict, Luohu District, Shenzhen,
Guangdong Province, China, Guangdong 518001
(CN).

(54) Title: VOICE CONVERSION METHOD AND REALTED DEVICE

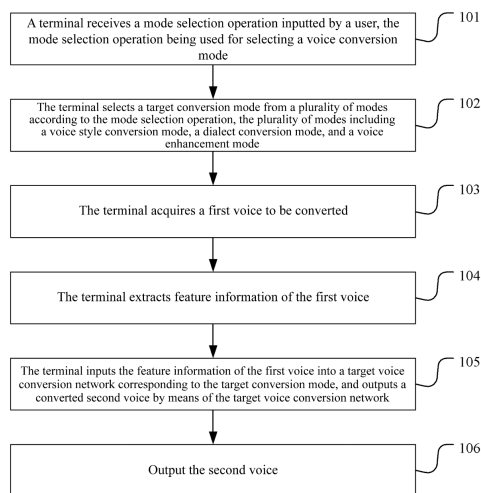


FIG. 1

- 101 A terminal receives a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode
- 102 The terminal selects a target conversion mode from multiple modes according to the mode selection operation, the multiple modes comprising a voice style conversion mode, a dialect conversion mode, and a speech enhancement mode
- 103 The terminal obtains a first voice to be converted
- 104 The terminal extracts feature information of said first voice
- 105 The terminal inputs the feature information of said first voice to a target voice conversion network corresponding to the target conversion mode, and outputs a second voice after conversion by means of the target voice conversion network
- 106 Output the second voice

(57) Abstract: A voice conversion method and a related device, used for realizing the diversification of sound beautification. The method in embodiments of the present application comprises: receiving a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode; providing multiple available modes, comprising: a style conversion mode used for converting the speaking style of a first voice to be converted, a dialect conversion mode used for adding an accent to or removing an accent from said first voice, and a speech enhancement mode used for performing speech enhancement on said first voice, and the three modes having corresponding voice conversion networks; selecting, according to a target conversion mode selected by the user, a target voice conversion network corresponding to the target conversion mode to convert said first voice; and outputting a second voice after conversion. In this way, the diversification of sound beautification is realized, and the requirements of the user under different application scenes are met.

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

This international Publication:

- Includes international search report (Article 21(3) of the Treaty).

(57) Abstract: Provided are a voice conversion method and a related device to implement diversified human voice beautification. The method in embodiments of the present application includes: receiving a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode. A plurality of selectable modes provided include: a style conversion mode for performing speaking style conversion on a first voice to be converted; a dialect conversion mode for implementing accent addition or accent removal on the first voice; and a voice enhancement mode for implementing voice enhancement on the first voice. The three modes have corresponding voice conversion networks. According to a target conversion mode selected by the user, a target voice conversion network corresponding to the target conversion mode is selected to perform conversion on the first voice, and a converted second voice is outputted, thereby implementing diversified human voice beautification and meeting the requirements of the user in different application scenarios.

VOICE CONVERSION METHOD AND RELATED DEVICE

The present application claims priority to Chinese Patent Application No. 202010996501.5, filed with the China National Intellectual Property Administration on September 21, 2020 and entitled “VOICE CONVERSION METHOD AND RELATED DEVICE”, which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

The present application relates to the technical field of voice processing, and in particular, to a voice conversion method and a related device.

BACKGROUND ART

Human voice beautification refers to modifying and beautifying sounds to produce a pleasant auditory effect. Currently, voice recording functions of a lot of audio recording software or social applications (apps) on the market each have a certain capability of human voice beautification. For example, denoising the voice, increasing the brightness or volume of the voice, and the like only modify the voice characteristics of the speaker himself, and human voice beautification modes are too limited.

SUMMARY OF THE INVENTION

Embodiments of the present application provide a voice conversion method and a related device to provide a plurality of human voice beautification modes, thereby implementing diversified human voice beautification.

According to a first aspect, an embodiment of the present application provides a voice conversion method. The method is applicable to a voice conversion apparatus. For example, the apparatus may be a terminal, and the method may comprise: receiving, by the terminal, a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode, and then selecting, by the terminal, a target conversion mode from a plurality of modes according to the mode selection operation, the plurality of modes comprising a style conversion mode, a dialect conversion mode, and a voice enhancement mode; the terminal being capable of selecting a target voice conversion network corresponding to the target conversion mode, and implementing voice conversion by means of the target voice conversion

network; acquiring, by the terminal device, a first voice to be converted, and further extracting feature information of the first voice, the feature information being used for retaining content information of the first voice; and then inputting, by the terminal, the feature information of the first voice into the target voice conversion network corresponding to the target conversion mode, outputting a converted second voice by means of the target voice conversion network, and finally outputting the second voice.

In this embodiment, a plurality of selectable modes are provided, for example, a style conversion mode for performing speaking style conversion on the first voice; a dialect conversion mode for implementing accent addition or accent removal; and a voice enhancement mode for implementing voice enhancement. The three modes have corresponding voice conversion networks, and the voice conversion networks corresponding to the modes can perform conversion on the first voice according to the user's requirements, thereby implementing diversified human voice beautification and meeting the requirements of the user in different application scenarios.

In an optional implementation, the extracting feature information of the first voice may specifically comprise: inputting, by the terminal, the first information into a voice feature extraction model; and extracting a phoneme posteriorgram (PPG) feature of the first voice by means of the voice feature extraction model, the PPG feature being used for retaining the content information of the first voice. The PPG feature describes the size of the probability of each phoneme in a phoneme set corresponding to a voice frame, which is equivalent to performing phoneme recognition. The PPG feature is used for retaining the content information of the first voice. In this example, the voice does not need to be converted into text to retain the content information of the voice. Instead, by directly inputting the voice, and retaining the content information of the first voice by means of the PPG feature, robustness can be further increased.

In an optional implementation, when the target conversion mode is the style conversion mode and the target voice conversion network is a style conversion network, the style conversion network comprises a style separation model and a voice fusion model. The method may further comprise: acquiring, by the terminal, a third voice for extracting a style feature, inputting the third voice into the style separation model, and separating a style feature of the third voice by means of the style separation model; and then, further inputting the style feature and the feature information of the first voice into the voice fusion model for fusion to obtain the second voice.

In this embodiment, the terminal receives the first voice to be converted and the third voice for extracting the style feature, then the first voice is inputted into the voice feature extraction model, and the PPG feature is extracted by means of the voice feature extraction model. The PPG feature is used for retaining the content information of the first voice, and the voice can be directly inputted by means of the PPG feature. The terminal inputs the third voice into the style separation model, separates the style feature of the third voice by means of the style separation model, and finally, inputs the style feature and the PPG feature into the voice fusion model for fusion to obtain the second voice in which the content of the first voice and the style of the third voice are fused. The third voice may be a voice of any person, so that the first voice is converted into a voice style of any person, thereby implementing diversified voice style conversion. In an optional implementation, the style feature comprises a first feature, and the first feature comprises a plurality of sub-features. The inputting the third voice into the style separation model, and separating a style feature of the third voice by means of the style separation model may specifically comprise: inputting, by the terminal, the third voice into the style separation model, and extracting a vector of the first feature in the third voice by means of the style separation model, where the first feature may be, for example, a timbre; then inputting the third voice into a sub-feature extraction model, and extracting vectors of the sub-features by means of the sub-feature extraction model; receiving a weight of each sub-feature among the plurality of sub-features inputted by the user; and determining the style feature of the third voice based on the vector of the first feature, the vector of each sub-feature, and the weight of each sub-feature.

In this embodiment, the similarity between the style of the voice to be converted and the third voice (that is, the voice corresponding to a target template) is adjusted by means of the weight corresponding to each sub-feature inputted by the user, and the degree of similarity is determined by the weight inputted by the user. The degree of similarity between the final outputted voice style and the third voice is determined by the weight, and the user can flexibly adjust the style of the voice to be converted by adjusting the inputted weight. The style of the converted voice may be completely the same as the style of the third voice, or may be changed on the basis of the style of the third voice, thereby implementing diversified styles of the voice to be converted.

In an optional implementation, determining the style feature of the third voice according to the vector of the first feature, the vector of each sub-feature, and the weight of each sub-feature may comprise: inputting, by the terminal, the vector of the first feature into a multi-head attention

structure; inputting a product of the vector of each sub-feature and the weight corresponding thereto into the multi-head attention structure; and outputting the style feature of the third voice by means of the multi-head attention structure.

In this embodiment, the multi-head attention structure enables the model to extract feature expressions from different subspaces, each head corresponding to one sub-feature space in a high-dimensional space, which is equivalent to decomposing the high-dimensional space, and each head being responsible for one sub-feature space. The multi-head attention mechanism is equivalent to a plurality of attention mechanisms having the same structure, so that the result outputted by the multi-head attention mechanism comprises a part of the timbre of the third voice.

In an optional implementation, acquiring the third voice for extracting the style feature may comprise:

receiving, by the terminal, a template selection operation inputted by the user, the template selection operation being used for selecting a target template, wherein the target template may be, for example, a style of a male announcer, a voice style of a certain “female announcer”, or the like; acquiring, by the terminal, a voice corresponding to the target template, using the voice corresponding to the target template as the third voice, and fusing the style feature of the third voice into the first voice, so that diversified voice styles can be implemented.

In an optional implementation, acquiring the third voice for extracting the style feature may specifically comprise: further being capable of receiving, by the terminal, the third voice inputted by a second speaker, the first voice being a voice of a first speaker, and the second speaker being any person different from the first speaker, so that diversified voice styles can be implemented.

In an optional implementation, when the target conversion mode is the dialect conversion mode and the target voice conversion network is a dialect conversion network, the inputting the feature information of the first voice into the target voice conversion network corresponding to the target conversion mode, and outputting a converted second voice by means of the target voice conversion network may further specifically comprise: inputting, by the terminal, the feature information of the first voice into the dialect conversion network, and outputting the second voice by means of the dialect conversion network, the first voice being a voice in a first dialect, and the second voice being a voice in a second dialect, thereby implementing dialect conversion,

enhancing the convenience of communication between users in different regions, and implementing diversified voice conversion.

In an optional implementation, the dialect conversion network comprises a plurality of dialect conversion models, each dialect conversion model being targeted at a different dialect to be converted, and the method further comprises: the terminal being capable of receiving a selection operation inputted by the user; and inputting the feature information of the first voice into a dialect conversion model corresponding to the selection operation, and outputting the second voice by means of the dialect conversion model corresponding to the selection operation. In this embodiment, the terminal may select a corresponding dialect conversion model according to a specific operation inputted by the user.

In an optional implementation, the method further comprises: inputting, by the terminal, the first voice into the style separation model, and separating a style feature of the first voice by means of the style separation model; and then inputting, by the terminal, the style feature of the first voice and the feature information of the first voice into the dialect conversion network, and outputting the second voice by means of the dialect conversion network, the second voice and the first voice having the same style.

In this embodiment, the content of the second voice is the same as that of the input voice (the first voice), and the speaking style of the input voice (the first voice) is retained.

In an optional implementation, the first voice is a far-field voice. When the target conversion mode is the voice enhancement mode and the target voice conversion network is a voice enhancement model, inputting the feature information of the first voice into the target voice conversion network corresponding to the target conversion mode, and outputting the converted second voice by means of the target voice conversion network may comprise: inputting, by the terminal, the feature information of the first voice into the voice enhancement model corresponding to the mode, and outputting the second voice by means of the voice enhancement model, the second voice being a near-field voice. In this example, the far-field voice is converted into the near-field voice to implement voice enhancement, so as to increase the clarity of the voice, increase application scenarios, and implement diversified voice conversion.

In an optional implementation, the method further comprises: inputting, by the terminal, the first voice into the style separation model, and separating the style feature of the first voice by

means of the style separation model; and then inputting the style feature of the first voice and the feature information of the first voice into the voice enhancement model, and outputting the second voice by means of the voice enhancement model, the second voice and the first voice having the same style. In this example, the converted voice is the same as the input voice (the first voice), and the speaking style of the input voice (the first voice) is retained.

In an optional implementation, acquiring the first voice to be converted may comprise: receiving the first voice inputted by the first speaker, or selecting the first voice from a locally stored file.

According to a second aspect, an embodiment of the present application provides a voice conversion apparatus, wherein the apparatus has a function performed by the terminal according to the first aspect described above, the function may be implemented by hardware or by hardware executing corresponding software, and the hardware or the software comprises one or more modules corresponding to the function.

According to a third aspect, an embodiment of the present application provides a terminal, comprising a processor, the processor being coupled to at least one memory, and the processor being configured to read a computer program stored in the at least one memory, so that the terminal performs any one of the methods according to the first aspect described above.

According to a fourth aspect, an embodiment of the present application provides a computer-readable storage medium, the computer-readable storage medium being configured to store a computer program, and when run on a computer, the computer program causing the computer to perform the method according to the first aspect described above.

According to a fifth aspect, the present application provides a chip system, the chip system comprising a processor, configured to support a terminal device in implementing functions related to the aspects described above. In a possible design, the chip system further comprises a memory, and the memory is configured to store program instructions and data that are necessary for the terminal device. The chip system may be formed of a chip, or may comprise a chip and other discrete devices.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a flowchart of steps of an embodiment of a voice conversion method in an embodiment of the present application;

FIG. 2 is a schematic diagram of a scenario of selecting a mode in an embodiment of the present application;

FIG. 3 is a schematic diagram of the correspondence between modes and voice conversion networks in an embodiment of the present application;

FIG. 4 is a flowchart of steps of another embodiment of a voice conversion method in an embodiment of the present application;

FIG. 5 is a schematic diagram of an example of performing a style conversion process on a voice in an embodiment of the present application;

FIG. 6A is a schematic diagram of a scenario of an interface of a style conversion mode in an embodiment of the present application;

FIG. 6B is a schematic diagram of another scenario of an interface of a style conversion mode in an embodiment of the present application;

FIG. 7 is a schematic diagram of training and updating a style conversion network in an embodiment of the present application;

FIG. 8 is a flowchart of steps of another embodiment of a voice conversion method in an embodiment of the present application;

FIG. 9 is a schematic diagram of an example of performing a dialect conversion process on a voice in an embodiment of the present application;

FIG. 10 is a schematic diagram of another example of performing a dialect conversion process on a voice in an embodiment of the present application;

FIG. 11A is a schematic diagram of a scenario of an interface of a dialect conversion mode and a voice enhancement mode in an embodiment of the present application;

FIG. 11B is a schematic diagram of another scenario of an interface of a dialect conversion mode and a voice enhancement mode in an embodiment of the present application;

FIG. 12 is a schematic diagram of training and updating a dialect conversion model in an embodiment of the present application;

FIG. 13 is a schematic structural diagram of an embodiment of a voice conversion apparatus in an embodiment of the present application;

FIG. 14 is a schematic structural diagram of an example of a chip in an embodiment of the present application; and

FIG. 15 is a schematic structural diagram of another embodiment of a voice conversion apparatus in an embodiment of the present application.

DETAILED DESCRIPTION OF THE EMBODIMENTS

The following describes the technical solutions in the embodiments of the present application with reference to the accompanying drawings in the embodiments of the present application. The term “and/or” appearing in the present application may be an association relationship describing associated objects, indicating that there may be three relationships, for example, A and/or B may indicate that A exists alone, A and B exist simultaneously, and B exists alone. In addition, the symbol “/” in the present application generally indicates an “or” relationship between the associated objects. The terms “first”, “second”, and the like in the specification, claims, and accompanying drawings of the present application are intended to distinguish between similar objects but do not necessarily indicate a specific order or sequence. It should be understood that the terms used in this way are interchangeable in appropriate circumstances, so that the embodiments described herein can be implemented in an order other than the order illustrated or described herein. Moreover, the terms “comprise”, “have” and any other variants are intended to cover non-exclusive inclusion, for example, a process, method, system, product, or device that includes a list of steps or modules is not necessarily limited to those steps or modules expressly listed, but may include other steps or modules not expressly listed or inherent to such a process, method, product, or device.

An embodiment of the present application provides a voice conversion method. The method is applicable to a terminal. The terminal may be an intelligent terminal having a voice function, such as a mobile phone, a tablet computer, a notebook computer, or a smartwatch. The terminal device may also be referred to as a terminal device, a user equipment (UE), or the like.

The terminal device may be described by using a mobile phone as an example. The terminal device is internally configured with a voice processing neural network, and the voice processing neural network mainly includes a style conversion network, a dialect conversion network, and a voice enhancement model. The voice processing neural network implements voice conversion (beautification). The present application provides a plurality of selectable modes, for example, a style conversion mode for performing speaking style conversion on the first voice; a dialect conversion mode for implementing accent addition or accent removal; and a voice enhancement mode for implementing voice enhancement. The three modes have corresponding voice conversion networks. According to a target conversion mode selected by a user, the terminal selects a target voice conversion network corresponding to the target conversion mode to perform conversion on the acquired first voice to be converted, and outputs a converted second voice, thereby implementing diversified human voice beautification and meeting the requirements of the user in different application scenarios.

To better understand the present application, words involved in the present application are first described.

Sequence to sequence (seq2seq) neural network: seq2seq is a type of encoder-decoder structure, in which an inputted sequence is compressed into a vector of a specified length by an encoder, and then the vector is inputted into a decoder to obtain an outputted sequence. The seq2seq neural network means that both the encoder and the decoder are formed of a neural network. The encoder is configured to encode the inputted sequence according to a specific rule to generate the vector. The decoder is configured to convert the generated vector into the outputted sequence.

Attention structure: A plurality of vectors of a specified length are generated in the encoder, each vector is obtained by weighting an input feature, and the values of weights are related to the output of the decoder. The purpose is for the output of the decoder to focus more on different key parts of an input instead of the entire input.

Neural network vocoder: The neural network vocoder in essence is a neural network, configured to convert an output feature of a voice processing neural network into a highly natural voice.

Style feature: The style feature includes, but is not limited to, a timbre. The timbre refers to the quality of a sound, and is also referred to as tone quality. The timbre can reflect the quality characteristic of a sound produced by a speaker. The timbre includes, but is not limited to, prosody, accent, and speech rate.

Phoneme: The phoneme is the smallest phonetic unit of an articulation. For example, hao (good) includes three phonemes in total, and wo (me) includes two phonemes in total.

Phoneme posteriorgram (PPG) feature: The phoneme posteriorgram is a posterior probability that each voice frame belongs to a group of predefined voice units (phonemes or triphones/auxiliary phonemes), and these voice units retain linguistic and voice information of a voice.

Mel feature: A voice signal being stored as a waveform diagram feature is represented as feature transformation in the time domain of the voice signal. However, the frequency distribution of a voice signal cannot be learned according to only the waveform diagram feature. Through the analysis of the frequency domain of the voice signal, the features of the voice can be extracted. As a feature representation of a voice signal, the Mel feature can clearly represent the formant characteristics of the voice.

Referring to FIG. 1, the present application provides an embodiment of a voice conversion method, and the method is performed by a terminal device. Alternatively, the method may be performed by a processor in the terminal device, or the method may be performed by a chip in the terminal device. In this embodiment of the present application, a description is provided using an example in which the method is performed by a terminal, and the terminal is a mobile phone. The terminal is internally configured with a voice processing neural network, and the voice processing neural network mainly includes a style conversion network, a dialect conversion network, and a voice enhancement model.

Step 101: A terminal receives a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode.

Referring to FIG. 2, the mode selection operation may be a tap operation, and a plurality of modes include a style conversion mode, a dialect conversion mode, and a voice enhancement mode. Each mode has a corresponding voice conversion network, and the voice is converted by

means of the corresponding voice conversion network. For example, the style conversion mode is used for converting a voice style. The dialect conversion mode is used for dialect conversion, such as “Accent addition”, “Accent removal”, or the like. The voice enhancement mode is used for converting a far-field voice into a near-field voice.

Step 102: The terminal selects a target conversion mode from a plurality of modes according to the mode selection operation, the plurality of modes including a style conversion mode, a dialect conversion mode, and a voice enhancement mode.

A list of the plurality of modes may be displayed on a display interface of the terminal, and the user selects a target conversion mode from the plurality of modes by means of a tap operation according to requirements, where the target conversion mode may be any one of the foregoing three modes.

The terminal selects a target voice conversion network corresponding to the target conversion mode.

Referring to FIG. 3, the style conversion mode corresponds to a style conversion network, the dialect conversion mode corresponds to a dialect conversion network, and the voice enhancement mode corresponds to a voice enhancement model. The terminal determines a target voice conversion network corresponding to the target conversion mode according to the correspondence between the modes and the voice conversion networks. For example, when the user selects the style conversion mode, the target voice conversion network is the style conversion network. When the user selects the dialect conversion mode, the target voice conversion network is the dialect conversion network. When the user selects the voice enhancement mode, the target voice conversion network is the voice enhancement model.

Step 103: The terminal acquires a first voice to be converted.

In a first implementation, the terminal receives the first voice inputted by a first speaker.

The terminal device receives, by means of a microphone, the first voice inputted by the first speaker. The terminal device receives a first operation inputted by the user (the user may be the same person as the first speaker, or may be a different person), and the first operation is an operation of recording a voice. For example, the first operation may be a tap operation. The user taps a “Start recording” button on the screen of the mobile phone, and after receiving the first

operation, the mobile phone starts to record the voice inputted by the user, the voice being a voice (i.e., the first voice) that the user desires to beautify (or process).

In a second implementation, the terminal selects the first voice from a locally stored file.

The terminal device may acquire the first voice from the locally stored file. The terminal device receives a second operation inputted by the user, and the second operation is an operation of selecting a voice file. For example, the user taps a “Select a file” button on the screen of the mobile phone, and according to the second operation, the mobile phone may display a list of voices to be selected. Then, the terminal device receives a third operation inputted by the user, the third operation is used for selecting a target voice, and the terminal device uses the target voice as the first voice.

Step 104: The terminal extracts feature information of the first voice.

The feature information may be a Mel frequency cepstrum coefficient (MFCC) feature. Alternatively, the feature information is a phoneme posteriorgram (PPG) feature. In this embodiment of the present application, the feature information of the first voice is described by using a PPG feature as an example. The PPG feature describes the size of the probability of each phoneme in a phoneme set corresponding to a voice frame, which is equivalent to performing phoneme recognition. The PPG feature is used for retaining content information of the first voice.

Step 105: The terminal inputs the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and outputs a converted second voice by means of the target voice conversion network.

In a first mode, the target voice conversion network is the style conversion network. Exemplarily, the terminal acquires a third voice for extracting a style feature, then inputs the third voice into a style separation model, and separates a style feature of the third voice by means of the style separation model; and finally inputs the style feature and the PPG feature of the first voice into the voice fusion model for fusion to acquire the second voice.

In a second mode, the feature information of the first voice is inputted into the dialect conversion network, and the second voice is outputted by means of the dialect conversion network, the first voice being a voice in a first dialect, and the second voice being a voice in a second dialect.

In a third mode, the PPG feature of the first voice is inputted into the voice enhancement model corresponding to the mode, and the second voice is outputted by means of the voice enhancement model, the second voice being a near-field voice.

Human voice beautification includes a plurality of modes, the plurality of modes include a style conversion mode, a dialect conversion mode and a voice enhancement mode, and the first voice may be beautified according to the mode selected by the user. For example, the style of the first voice is beautified, dialect conversion is performed on the first voice, or voice enhancement is performed on the first voice, and so on, implementing diversified human voice beautification.

Step 106: The terminal outputs the second voice.

The terminal outputs the converted second voice by means of a loudspeaker.

In the embodiments of the present application, there are provided a plurality of selectable modes, for example, a style conversion mode for performing speaking style conversion on the first voice; a dialect conversion mode for implementing accent addition or accent removal; and a voice enhancement mode for implementing voice enhancement. The three modes have corresponding voice conversion networks, and the voice conversion networks corresponding to the modes can perform conversion on the first voice according to the user's requirements, thereby implementing diversified human voice beautification and meeting the requirements of the user in different application scenarios.

Referring to FIG. 4, an embodiment of the present application provides another embodiment of voice conversion, and in this embodiment, a style conversion mode is described, that is, performing style conversion on a first voice is described as an example. The mode is a style conversion mode, and the target voice conversion network is a style conversion network.

Step 401: A terminal acquires a first voice to be converted and a third voice for extracting a style feature.

For a method for acquiring the first voice by the terminal in this step, refer to the description of step 103 in the embodiment corresponding to FIG. 1. Details are not described here again.

The terminal acquires the third voice in the following manner:

In a first implementation, the terminal receives a template selection operation inputted by a user, and the terminal selects, according to the template selection operation, a voice

corresponding to a target template, and then uses the voice corresponding to the target template as the third voice. For example, the target template includes, but is not limited to, “Male announcer”, “Female announcer”, “Actor's voice”, etc.

Optionally, the target template may also be a category. Further, the terminal device receives a sub-template selection operation, and the terminal device selects, according to the sub-template selection operation, a target voice corresponding to a sub-template. For example, a list of sub-templates under the category of “Male announcer” includes “Announcer A”, “Announcer B”, “Announcer C”, etc. A list of sub-templates under the category of “Female announcer” includes “Announcer D”, “Announcer C”, etc. A list of sub-templates under the category of “Actor's voice” includes “Actor D”, “Actor F”, etc. For example, the user selects “Announcer A”, and according to the sub-template selection operation inputted by the user, the terminal device selects a voice corresponding to “Announcer A” as the third voice. It should be noted that the examples of the target template and the sub-template in this embodiment are all examples for description, and are not intended to be limiting.

In a second implementation, the terminal may receive the third voice inputted by a second speaker. Exemplarily, the terminal device receives the template selection operation inputted by the user, the target template selected by the template selection operation is “Voice of any person”, and the terminal may use a voice corresponding to “Voice of any person” as the third voice. For example, the user taps a “voice of any person” option, and the terminal starts to record the third voice, the second speaker being any person different from the first speaker.

Optionally, the third voice may be a recorded voice that is locally pre-stored. For example, the third voice may be the voice of a favorite character in a movie (a cartoon character's voice, or a certain actor's voice). Optionally, the third voice may also be a voice that is downloaded from the Internet in advance and then stored in the terminal device.

Step 402: The terminal inputs the first voice into a voice feature extraction model, and extracts a PPG feature of the first voice by means of the voice feature extraction model, the PPG feature being used for retaining content information of the first voice.

Referring to FIG. 5, the terminal inputs the first voice into the voice feature extraction model, and extracts the PPG feature of the first voice by means of the voice feature extraction model. Exemplarily, the voice feature extraction model may be a deep neural network. The voice

feature extraction model includes a plurality of convolutional layers, two LSTM layers, and one fully connected layer. The voice feature extraction model outputs the PPG feature of the first voice. The PPG feature describes the size of the probability of each phoneme in a phoneme set corresponding to a voice frame, which is equivalent to performing phoneme recognition. The PPG feature is used for retaining the content information of the first voice. The voice feature extraction model is obtained by training in advance on a large number of linguistic materials.

In this embodiment, the PPG feature is inputted as the content information of the first voice, and these voice units (phonemes) retain linguistic and voice information of the voice. Compared with an automatic speech recognition (ASR) technology, robustness can be further improved by using the PPG feature, as the ASR technology needs to first convert a voice into a text, thereby increasing the probability of error in voice content recognition. In this embodiment, the PPG feature is inputted as the content information of the first voice, that is, a voice can be directly inputted without being converted into text content, thereby improving robustness of a system.

Step 403: The terminal inputs the third voice into a style separation model, and separates a style feature of the third voice by means of the style separation model.

The style feature includes a first feature, and the first feature includes a plurality of sub-features. For example, the first feature is a timbre feature, and the plurality of sub-features include prosody, accent, speech rate, etc.

The style separation model is used to separate the style feature of the third voice. The style separation model includes a timbre separation model. The timbre separation model is used to separate the timbre feature of the third voice, thereby obtaining a vector of the first feature (i.e. a timbre feature vector).

Optionally, the style conversion network further includes a plurality of sub-feature extraction models and a multi-head attention structure. The terminal inputs the third voice into the sub-feature extraction models, and extracts a vector of a sub-feature by means of the sub-feature extraction model. For example, the plurality of sub-feature extraction models include a prosody extraction model, an accent extraction model, and a speech rate extraction model. For example, the prosody extraction model is used to extract a prosody feature in the third voice to obtain a prosody vector. The accent extraction model is used to extract an accent feature in the third voice

to obtain an accent vector. The speech rate extraction model is used to extract a speech rate feature in the third voice to obtain a speech rate vector.

First, the terminal receives a weight of each of the plurality of sub-features inputted by the user. In an example, adjustment bars for prosody, speech rate, and accent are displayed on a screen of a mobile phone, and the user may input a weight corresponding to each sub-feature by adjusting the adjustment bar of each sub-feature. The weight corresponding to each sub-feature may be flexibly adjusted according to the user's own needs. For example, setting “Prosody” to 10% indicates that the final outputted voice has a similarity of 10% to the prosody of the target template, that is, a parameter having a value of 0.1 is transferred to a built-in sub-feature extraction model.

In another example, several levels are pre-configured. For example, there may be three levels, and the weight of each sub-feature at each level is pre-configured according to an empirical value. For example, at a first level, the weight of prosody is 0.1, the weight of speech rate is 0.2, and the weight of accent is 0.1. At a second level, the weight of prosody is 0.2, the weight of speech rate is 0.2, and the weight of accent is 0.1, and so on and so forth. The terminal determines the weight corresponding to each sub-feature by receiving a level inputted by the user. In this example, the user does not need to adjust the weight of each sub-feature separately, and the user only needs to select the level, which is convenient for the user to operate.

Then, the terminal determines the style feature of the third voice according to the vector of the first feature, the vector of each sub-feature, and the weight of each sub-feature.

The terminal multiplies the weight corresponding to each sub-feature inputted by the user by the vector of each sub-feature. The result of the multiplication and the vector of the first feature (such as the timbre vector) are simultaneously inputted into the multi-head attention structure for attention alignment, so that the multi-head attention structure outputs a style vector, which is the style feature of the third voice. The multi-head attention structure enables the model to extract feature expressions from different subspaces, each head corresponding to one sub-feature space in a high-dimensional space, which is equivalent to decomposing the high-dimensional space, each head being responsible for one sub-feature space. The multi-head attention mechanism is equivalent to a plurality of attention mechanisms having the same structure, so that the result outputted by the multi-head attention mechanism includes a part of the timbre of the third voice.

For example, the structure outputted by the multi-head attention mechanism may be a 256-dimension style embedding vector (style feature).

In this embodiment, the similarity between the style of the voice to be converted and the third voice is adjusted by means of the weight corresponding to each sub-feature inputted by the user, and the degree of similarity is determined by the weight inputted by the user. The user may select whether to input the weight, and if the user selects to input the weight, the degree of similarity between styles of the final outputted voice and the third voice is determined by the input weight. If the user selects not to input the weight, the style of the final outputted voice will be exactly the same as the style (or timbre, in which timbre is used as an example of the style) of the third voice. The user can flexibly adjust the style of the voice to be converted by adjusting the input weight. The style of the converted voice may be completely the same as the style of the third voice, or may be changed on the basis of the style of the third voice, thereby implementing diversified styles of the voice to be converted. For example, the style of the converted voice may be a voice style of a certain “Male announcer”, or a voice style of a certain “Female announcer”, or a voice style of any person, and may be changed on the basis of the voice styles of the “Male announcer”, the “Female announcer”, or any person, so that diversified voice styles can be implemented.

Step 404: The terminal inputs the style feature and the PPG feature into a voice fusion model for fusion to obtain a second voice.

Exemplarily, the voice fusion model is a seq2seq neural network, and an attention mechanism is introduced into the seq2seq neural network. The seq2seq neural network includes an encoder, a decoder, and an attention structure. Any combination of neural network models may be used for the models of the encoder and the decoder, including, for example, any one or a combination of any two of a convolutional neural network (CNN), a recurrent neural network (RNN), and a long short-term memory (LSTM). For example, in the present application, the encoder may include three convolutional layers and one bidirectional LSTM layer. The PPG feature is first inputted into the encoder, and the encoder encodes the inputted PPG sequence into a vector of a fixed dimension. Since the length of the input sequence may be long, it is difficult to express rich information in one vector during decoding, and thus an attention mechanism is introduced. Then, the vector of the style feature and the PPG vector outputted by the encoder are

tilled in width to obtain an attention input matrix. Then, the attention input matrix is sent to the attention structure frame by frame, and the attention structure cooperates with the encoder to output a Mel (which may also be referred to as a Mel frequency spectrum) feature of the second voice. The Mel frequency spectrum feature is a feature representation form of the second voice. After encoding is completed, an output vector of an encoding module is weighted differently to obtain several vectors, each vector corresponding to one output, thereby ensuring that an output sequence is focused on an input sequence instead of depending on all input sequences.

Step 405: The terminal outputs the second voice.

The Mel frequency spectrum feature is an intermediate representation, and although the Mel frequency spectrum contains information of the output voice, the Mel frequency spectrum cannot be directly played, and needs to go through the inverse process of converting a voice signal into a Mel spectrogram. In the present application, a playable audio file may be generated by a vocoder by means of an acoustic feature of the Mel frequency spectrum. The vocoder may be a neural network vocoder, and the network is responsible for converting a Mel feature into a highly natural voice signal. The network is formed of a plurality of convolutional layers and deconvolutional layers, and the final output is a voice that can be played. Content in the voice is the same as that in the first voice, and the speaking style is the same as that in the third voice. The terminal may directly play the second voice. Optionally, the user may select whether to retain the second voice. If the user selects to retain the second voice, the terminal stores the second voice.

In this embodiment, the terminal receives the first voice to be converted and the third voice for extracting the style feature, then the first voice is inputted into the voice feature extraction model, and the PPG feature is extracted by means of the voice feature extraction model. The PPG feature is used for retaining the content information of the first voice, and the voice can be directly inputted by means of the PPG feature. The terminal inputs the third voice into the style separation model, separates the style feature of the third voice by means of the style separation model, and finally, inputs the style feature and the PPG feature into the voice fusion model for fusion to obtain the second voice in which the content of the first voice and the style of the third voice are fused. The third voice may be a voice of any person, so that the first voice is converted into a voice style of any person, thereby implementing diversified voice style conversion.

In an application scenario, the voice processing neural network in the present application is built in an application. For example, the app may be an instant messaging app or a human voice beautification app, and the human voice beautification app may be applied to an audio recording or video recording scenario. For example, the user may beautify a voice to be sent by means of the app. In the audio recording or video recording scenario, after recording a voice or a video with a voice by using the terminal, the user may beautify the voice by means of an app formed in the present invention.

Referring to FIGS. 6A and 6B, a method for performing human voice beautification by a user by means of the app may include:

1). After the user taps a “Start recording” button on a screen, the user may record a voice (a first voice) that the user desires to process. Alternatively, the user may tap “Select a file”, and then select a voice from a local file as the first voice. Moreover, the first voice that has been read may be displayed by means of an audio waveform.

2). The user selects a target template in an interface, that is, the user selects a third voice. When the user selects “Simulating Actor F”, “Male announcer” or “Female announcer”, a voice already stored in the terminal is used by default. When the user selects “Voice of any person”, the user needs to tap a “Start recording” or “Select a file” button to input the voice of any person into the app.

If the user taps the “Start recording” button, the terminal device starts to record voice of any person, that is, to receive third voice inputted by a second speaker. If the user taps “Select a file”, the terminal device selects an already stored voice of any person from a local file, and uses the voice as the third voice.

3). The user selects an input style feature weight (also referred to as a “sub-feature weight”) under “Style similarity adjustment”. For example, setting “Prosody” to 20% indicates that the final outputted voice has a similarity of 20% to the prosody of the target template, and the app transfers a parameter having a value of 0.2 to a built-in neural network model. Similarly, setting “Speech rate” to 30% indicates that the speech rate of the final outputted voice has a similarity of 30% to the speech rate of the target template, and setting “Accent” to 40% indicates that the accent of the final outputted voice has a similarity of 40% to the accent of the target template. It should be noted

that the adjustable style features include, but is not limited to, pitch, intensity, duration, timbre, etc. The timbre includes, but is not limited to, prosody, speech rate, and accent.

4). After the user taps a “Start beautification” button, the inputs in steps 1, 2, and 3 are simultaneously inputted into a trained voice processing neural network built in the app, and a processed voice (a second voice) is outputted after being processed by the voice processing neural network, where content of the voice is the same as content of the first voice, the style of the voice is similar to the target template, and the degree of similarity is determined by the style feature weight used for input.

5). The user can see a processed voice waveform on the interface of the app, and there will also be three buttons, “Play”, “Save file” and “Re-record”. When the user taps “Play”, the processed voice (the second voice) is played by means of a loudspeaker of the mobile phone. When the user taps “Save file”, the terminal saves the processed voice (the second voice) locally. When the user taps “Re-record”, the terminal starts to perform re-processing, that is, returns to step 4.

The style conversion network and the application scenario are described above. The training and updating processes of the style conversion network will be described below.

The style conversion network is a neural network including three parts, namely, a voice feature extraction part (i.e., a voice feature extraction model), a style separation part (i.e., including a style separation model, a multi-head attention structure, and a plurality of sub-feature extraction models), and a voice fusion model.

Referring to FIG. 7, in FIG. 7, a black solid line indicates a flow direction of input data, and a thick dashed line indicates a direction of parameter updating. In the style conversion network, the style separation model and the voice feature extraction model do not participate in network updating.

The style separation model is trained in advance using a large amount of voice data of different speakers. That is, a training dataset of the style separation model includes a large number of linguistic materials of different speakers (voices of different speakers have different timbre features), and the style separation model is trained on the training set to obtain the style separation model. The style separation model does not participate in parameter updating of the entire network.

The voice feature extraction model is trained in advance on a large number of linguistic materials. For example, an input of sample data included in the training dataset is a voice, and a label is a PPG feature of the inputted voice. The voice feature extraction model is used to extract the PPG feature, and the voice feature extraction model does not participate in parameter updating of the entire network.

Except for the style separation model and the voice feature extraction model, other models in the entire network participate in the training and updating of the entire network.

Please refer to the models through which the thick dashed lines pass in FIG. 7. Among the multi-head attention structure, each sub-feature extraction model, and the Seq2seq neural network (including the encoder, the decoder, and the attention structure), each structure may be a neural network, and each neural network may include a plurality of convolutional layers, fully connected layers, LSTM layers, and other structural layers, while the weight parameters of each layer need to be trained on a large amount of data to be obtained.

The updating of the style conversion network in the present application is different from the conventional method in that, since the style separation model and the voice feature extraction model in the present application have been pre-trained and do not participate in the network updating, the training of the style separation model does not need to be considered. Accordingly, an input and a label do not need to be two pieces of voice having the same content (two pieces of voice spoken by two different people), which greatly reduces the amount of sample data. A first Mel feature is obtained from the label through a short-time fourier transform (STFT) algorithm. After the voice is inputted into the style conversion network, the output through this network is a second Mel feature. A loss value and a gradient are obtained by comparing the first Mel feature and the second Mel feature. The output of this network is compared with the Mel feature obtained from the input through the STFT algorithm to obtain the loss value and the gradient thereof. In FIG. 7, the thick dashed lines indicate flow directions of gradients. Only when a gradient flows through a corresponding structure will the parameters in the structure be updated. Further, the style separation model and the voice feature extraction model are obtained by training in advance and do not need to participate in the parameter updating of the network. The loss value is an indicator for determining when the network stops training. When the loss value drops to a certain value and

there is no obvious continuing decreasing trend, it indicates that the network has converged, and the training can be stopped.

The entire style conversion network does not use a style label (that is, the input and the label are in the same voice sequence) to obtain the capability of voice beautification. This network learning method is an unsupervised learning method, and the style conversion network trained by the unsupervised learning method can support the user to manually adjust the degree of beautification for the recorded voice and can combine the speaking styles of different speakers.

Referring to FIG. 8, an embodiment of the present application provides another embodiment of voice conversion, and the dialect conversion mode and the voice enhancement mode are described in this embodiment.

The dialect conversion network includes a plurality of dialect conversion models, and the dialect conversion models are used to perform dialect conversion on the inputted first voice. The voice enhancement model is used for converting a far-field voice into a near-field voice to implement voice enhancement.

The plurality of dialect conversion models include at least two types. A first type is accent removal, namely, converting a dialect into Mandarin. For example, Sichuanese is converted into Mandarin. A second type is accent addition, that is, converting Mandarin into a dialect, for example, converting Mandarin into Sichuanese, or the like. Each dialect conversion model is targeted at a different dialect to be converted. Dialect conversion can enhance the convenience of communication between users in different regions, thereby implementing diversified voice conversion. It should be noted that only two types are exemplified herein for illustration; certainly, in an optional solution, two dialect conversion models may also be used in conjunction to perform conversion between two dialects. For example, to convert Sichuanese into Cantonese, it is possible to first convert Sichuanese into Mandarin, and then convert Mandarin into Cantonese.

Step 801: A terminal receives a mode selection operation inputted by a user.

For example, different processing modes for a voice are displayed on a user interface of the terminal. For example, the processing modes may include a “dialect switching mode” and a “voice enhancement mode”. The model selection operation may be a tap operation, and when the user taps the “dialect switching mode”, the user selects a dialect switching processing mode.

Step 802: The terminal receives a selection operation inputted by the user.

Alternatively, when the mode selection operation is used for selecting the dialect switching mode, the mode selection operation (a first-level operation) may further include a plurality of selection operations of a next level (second-level operations), for example, a first selection operation and a second selection operation. The first selection operation is used for selecting “Accent removal”, and the second selection operation is used for selecting “Accent addition”. Each second-level operation may include a plurality of third-level operations. In an application scenario, when the user selects “Accent dialect switching”, second-level classification options of “Accent removal” and “Accent addition” will appear. When the user selects “Accent addition”, the user interface of the terminal displays third-level classification labels of accent options of different places, for example, “Sichuan accent”, “Guangdong accent”, etc. The terminal receives a third-level selection operation inputted by the user, such as a first operation, a second operation, and the like. The first operation is used for selecting to convert Sichuanese into Mandarin, the second operation is used for selecting to convert Mandarin into Sichuanese, and so on. The terminal may select a corresponding dialect conversion model according to a specific operation inputted by the user.

Step 803: The terminal inputs feature information of a first voice into a dialect conversion model corresponding to the selection operation, and outputs a second voice by means of the dialect conversion model corresponding to the selection operation.

Referring to FIG. 9, when the mode selection operation is used for selecting the dialect switching mode, the terminal inputs a PPG feature into a dialect switching model (a seq2seq neural network) according to the mode selection operation, outputs a Mel feature form of the second voice by means of the dialect switching model, and then converts the Mel feature into a playable voice form by means of a neural network vocoder. The first voice is a voice in a first dialect, and the second voice is a voice in a second dialect.

Optionally, the terminal inputs third voice into a style separation model, and separates the style feature of the first voice by means of the style separation model to obtain a timbre feature vector of the first voice. Then, the PPG feature of the first voice and the timbre feature vector of the first voice are inputted into the dialect conversion model corresponding to the selection operation, and the second voice is outputted by means of the dialect conversion model

corresponding to the selection operation. Content of the second voice is the same as that of the input voice (the first voice), and the speaking style of the input voice (the first voice) is retained.

It should be noted that, in the embodiments of the present application, the dialect is a “language of a place”, for example, Sichuanese, Cantonese, etc. Mandarin is also a dialect. In Chinese, Mandarin uses the Beijing voice as the standard pronunciation and can be understood as a dialect based on Northern Chinese. That is, in the present application, Mandarin is also a dialect.

Exemplarily, the first dialect is Sichuanese, and the second dialect is Mandarin. Alternatively, the first dialect is Mandarin, and the second dialect is Sichuanese, or the second dialect is a Northeastern dialect, and so on. Specifically, the first dialect and the second dialect are not limited.

Referring to FIG. 10, the dialect conversion network includes N dialect conversion models, each dialect conversion model is targeted at a different dialect to be converted, and conversion of each dialect has a corresponding model. For example, among the plurality of dialect conversion models, a first dialect conversion model is used to convert Sichuanese into Mandarin, a second dialect conversion model is used to convert Mandarin into Cantonese, an integrated model of the first dialect conversion model and the second dialect conversion model may convert Sichuanese into Cantonese, and so on. In this example, the first dialect and the second dialect are illustrated, which is not intended to limit the present application.

Optionally, when a mode corresponding to the mode selection operation is the voice enhancement mode, the terminal inputs the PPG feature of the first voice into the voice enhancement model corresponding to the mode, and outputs the second voice by means of the voice enhancement model, where the first voice is a far-field voice, and the second voice is a near-field voice. The far-field voice is converted into the near-field voice, thereby implementing voice enhancement, increasing application scenarios, and implementing diversified voice conversion. The voice enhancement model is obtained by learning performed based on sample data in a training dataset, the sample data includes an input and a label, the input is a far-field voice, and the label is a near-field voice.

Optionally, the first voice is inputted into the style separation model, and the style feature of the first voice is separated by the style separation model. Then, the style feature of the first voice and the feature information of the first voice are inputted into the voice enhancement model, and

the second voice is outputted by means of the voice enhancement model, the second voice and the first voice having the same style. In this example, the converted voice is the same as the input voice (the first voice), and the speaking style of the input voice (the first voice) is retained.

For example, in an application scenario, refer to FIGS. 11A and 11B.

a. After the user taps the “Start recording” button, the terminal records a voice (the first voice) that the user desires to process, or the user may tap the “Select a file” button, and the terminal selects a voice from a local file as the first voice. Optionally, the terminal may display a waveform of an audio that has been read.

b. The user selects a voice processing mode in the interface. The interface displays a first-level classification label. For example, the first-level classification label includes “Dialect switching” and “Voice enhancement”. When the user selects “Dialect switching”, the interface displays second-level classification labels. For example, the second-level classification labels include “Accent removal” and “Accent addition”. When the user selects “Accent addition”, the interface displays third-level classification labels of accent options of different places, for example, “Guangdong accent”, “Sichuan accent”, “Fujian accent”, etc. Alternatively, the user may select a “Voice enhancement” mode.

c. The terminal selects a corresponding model according to the mode selected by the user.

d. After the user taps the “Start beautification” button, the first voice is inputted into the selected model (for example, a dialect conversion model for converting Mandarin to a Guangdong accent), and after being processed by the dialect conversion model for a period of time, a processed second voice is outputted, where the content of the second voice is the same as that of the input voice (the first voice), and the speaking style of the input voice (the first voice) is retained. Alternatively, the first voice is inputted into the selected voice enhancement model, and after being processed by the voice enhancement model for a period of time, a processed second voice is outputted, where the content of the second voice is the same as that of the input voice (the first voice), and the speaking style of the input voice (the first voice) is retained. The first voice is a far-field voice, and the second voice is a near-field voice.

e. Finally, the user can see a processed voice waveform on the display interface. The display interface further displays three buttons, “Play”, “Save file”, and “Re-record”. When the

user taps the “Play” button, the processed voice will be played by means of the loudspeaker of the mobile phone. When the user taps the “Save file” button, the terminal saves the processed voice locally. When the user taps the “Re-record” button, the terminal re-processes the first voice, and selects a corresponding model according to the selection operation of the user on the interface at this time, while returning to step d.

The training and updating processes of the dialect conversion model will be described below.

Referring to FIG. 12, each dialect conversion model has the same structure, and each dialect conversion model is a seq2seq neural network. The seq2seq neural network includes an encoder, a decoder, and an attention structure. Different dialect conversion models correspond to different training datasets, and parameters of each layer are different. For example, a first dialect conversion model is to convert Sichuanese to Mandarin, and each piece of sample data in the training dataset corresponding to the first dialect conversion model includes an input and a label. The input is a Sichuanese voice, and the label is a Mandarin voice having the same content as the Sichuanese voice. A second dialect conversion model is to convert Mandarin to Sichuanese, and each piece of sample data in the training dataset corresponding to the first dialect conversion model includes an input and a label. The input is a Mandarin voice, and the label is a Sichuanese voice having the same content as the Mandarin voice. So on and so forth. Different dialect conversion models perform learning on different training datasets. In FIG. 12, parameter updates of modules in a dialect conversion model are used as an example.

A training manner of this training solution is substantially the same as that of a style conversion network. A style separation model and a voice feature extraction model are already trained in advance, and do not participate in network updating, that is, training of the style separation model and the voice feature extraction model does not need to be considered. Accordingly, a third Mel feature is obtained from a label through an STFT algorithm, and the output through this network is a fourth Mel feature. A loss value and a gradient are obtained by comparing the third Mel feature and the fourth Mel feature. The fourth Mel feature outputted by this network is compared with the third Mel feature obtained from the input through the STFT algorithm to obtain the loss value and the gradient. In FIG. 12, the thick dashed lines indicate flow directions of gradients. Only when a gradient flows through a corresponding structure (for

example, an encoder, an attention structure, and a decoder) will the parameter in the module be updated. Further, the style separation model and the voice feature extraction model are obtained by training in advance and do not need to participate in the parameter updating of the network. The loss value is an indicator for determining when the network stops training. When the loss value drops to a certain value and there is no obvious continuing decreasing trend, it indicates that the network has converged, and the training can be stopped.

In the embodiment of the present application, style features such as timbre, prosody, and speech rate in the voice are extracted by means of unsupervised learning, so as to implement controllable beautification of the style of the voice. For a voice beautification scenario in which the user desires the style of the processed voice to be the same as that before processing, for example, an application scenario of dialect conversion and voice enhancement, in the present application, the style of the processed voice can remain unchanged while implementing dialect conversion or voice enhancement. In the present application, artificial intelligence technology is used to provide a more convenient and abundant method for implementing human voice beautification, which can cover more scenarios. The effect of accent removal, accent addition or voice enhancement of any person is achieved, while the input and output voice styles can remain unchanged.

In the embodiments of the present application, training and updating of the models may be performed by a server, and the server downloads the trained voice processing neural network from the terminal to a local end. In practical applications, the voice processing neural network is built in an app, for example, the app is an instant messaging app, and the like.

A voice conversion method has been described above. Referring to FIG. 13, a voice conversion apparatus 1300 to which the method is applied will be described below. The voice conversion apparatus includes an input module 1320, a processing module 1310, an acquisition module 1330, and an output module 1340.

The input module 1320 is configured to receive a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode.

The processing module 1310 is configured to select a target conversion mode from a plurality of modes according to the mode selection operation received by the input module 1320,

the plurality of modes including a style conversion mode, a dialect conversion mode, and a voice enhancement mode.

The acquisition module 1330 is configured to acquire a first voice to be converted.

The processing module 1310 is further configured to: extract feature information of the first voice acquired by the acquisition module 1330; input the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and output a converted second voice by means of the target voice conversion network.

The output module 1340 is configured to output the second voice.

In an optional implementation, the processing module 1310 is further configured to: input the first information into a voice feature extraction model, and extract a phoneme posteriogram (PPG) feature of the first voice by means of the voice feature extraction model, the PPG feature being used for retaining content information of the first voice.

In an optional implementation, when the target conversion mode is the style conversion mode and the target voice conversion network is the style conversion network, the style conversion network includes a style separation model and a voice fusion model.

The acquisition module 1330 is configured to acquire a third voice for extracting a style feature.

The processing module 1310 is further configured to: input the third voice into the style separation model, and separate a style feature of the third voice by means of the style separation model.

The processing module 1310 is further configured to input the style feature and the feature information of the first voice into the voice fusion model for fusion to obtain the second voice.

In an optional implementation, the style feature includes a first feature, and the first feature includes a plurality of sub-features.

The processing module 1310 is further configured to: input the third voice into the style separation model, and extract a vector of the first feature of the third voice by means of the style separation model; and input the third voice into a sub-feature extraction model, and extract vectors of the sub-features by means of the sub-feature extraction model.

The input module 1320 is further configured to receive a weight of each sub-feature among the plurality of sub-features inputted by the user.

The processing module 1310 is further configured to determine the style feature of the third voice based on the vector of the first feature, the vector of each sub-feature received by the input module 1320, and the weight of each sub-feature received by the input module 1320.

In an optional implementation, the processing module 1310 is further configured to: input the vector of the first feature into a multi-head attention structure, input a product of the vector of each sub-feature and the weight corresponding to the sub-feature into the multi-head attention structure, and output the style feature of the third voice by means of the multi-head attention structure.

In an optional implementation, the acquisition module 1330 is further configured to: receive a template selection operation inputted by the user, the template selection operation being used for selecting a target template; and acquire a voice corresponding to the target template, and use the voice corresponding to the target template as the third voice.

In an optional implementation, the acquisition module 1330 is further configured to receive the third voice inputted by a second speaker, the first voice being a voice of a first speaker, and the second speaker being any person different from the first speaker.

In an optional implementation, the target conversion mode is the dialect conversion mode and the target voice conversion network is a dialect conversion network. The processing module 1310 is further configured to input the feature information of the first voice into the dialect conversion network, and output the second voice by means of the dialect conversion network, the first voice being a voice in a first dialect, and the second voice being a voice in a second dialect.

In an optional implementation, the dialect conversion network includes a plurality of dialect conversion models, and each dialect conversion model is targeted at a different dialect to be converted.

The input module 1320 is further configured to receive a selection operation inputted by the user.

The processing module 1310 is further configured to: input the feature information of the first voice into a dialect conversion model corresponding to the selection operation, and output the second voice by means of the dialect conversion model corresponding to the selection operation.

In an optional implementation, the processing module 1310 is further configured to: input the first voice into the style separation model, and separate a style feature of the first voice by means of the style separation model.

The processing module 1310 is further configured to: input the style feature of the first voice and the feature information of the first voice into the dialect conversion network, and output the second voice by means of the dialect conversion network, the second voice and the first voice having the same style.

In an optional implementation, the first voice is a far-field voice. When the target conversion mode is a voice enhancement mode and the target voice conversion network is a voice enhancement model,

the processing module 1310 is further configured to: input the feature information of the first voice into the voice enhancement model corresponding to the mode, and output the second voice by means of the voice enhancement model, the second voice being a near-field voice.

In an optional implementation, the processing module 1310 is further configured to: input the first voice into the style separation model, and separate a style feature of the first voice by means of the style separation model.

The processing module 1310 is further configured to: input the style feature of the first voice and the feature information of the first voice into the voice enhancement model, and output the second voice by means of the voice enhancement model, the second voice and the first voice having the same style.

In an optional implementation, the acquisition module 1330 is further configured to receive the first voice inputted by the first speaker, or select the first voice from a locally stored file.

In a possible design, the processing module 1310 may be a processing apparatus, and some or all functions of the processing apparatus may be implemented by software.

Alternatively, the functions of the processing apparatus may be implemented partially or entirely by software. In this case, the processing apparatus may include a memory and a processor,

the memory is configured to store a computer program, and the processor reads and executes the computer program stored in the memory, to perform corresponding processing and/or steps in any one of the method embodiments.

Alternatively, the processing apparatus may include only a processor. The memory for storing the computer program is located outside the processing apparatus, and the processor is connected to the memory via a circuit/wire to read and execute the computer program stored in the memory.

Optionally, the processing apparatus may be one or more chips, or one or more integrated circuits.

Exemplarily, an embodiment of the present application provides a chip structure. Referring to FIG. 14, the chip includes:

The chip may be represented as a neural-network processing unit (NPU) 140. The NPU is mounted to a host CPU as a coprocessor, and the host CPU allocates a task. A core part of the NPU is an operation circuit 1403, and a controller 1404 controls the operation circuit 1403 to extract matrix data in the memory and perform a multiplication operation.

In some implementations, the operation circuit 1403 internally includes a plurality of process engines (PE). In some implementations, the operation circuit 1403 is a two-dimensional systolic array. The operation circuit 1403 may also be a one-dimensional systolic array or other electronic circuits capable of performing mathematical operations such as multiplication and addition. In some implementations, the operation circuit 1403 is a general matrix processor.

For example, it is assumed that there are an input matrix A, a weight matrix B, and an output matrix C. The operation circuit obtains data corresponding to the matrix B from a weight memory 1402, and buffers the data on each PE in the operation circuit. The operation circuit obtains data of the matrix A from an input memory 1401 to perform a matrix operation with the matrix B, and stores an obtained partial result or final result of the matrix in an accumulator 1408.

A unified memory 1406 is configured to store input data and output data. Weight data is directly transferred to the weight memory 1402 through a direct memory access controller (DMAC) 1405. The input data is also transferred to the unified memory 1406 through the DMAC.

A BIU is a bus interface unit, i.e., a bus interface unit 1410, and is configured to interact between an AXI bus and the DMAC and an instruction fetch buffer 1409.

The bus interface unit (BIU for short) 1410 is used by the instruction fetch buffer 1409 to acquire an instruction from an external memory, and is further used by the direct memory access controller 1405 to acquire original data of the input matrix A or the weight matrix B from the external memory.

The DMAC is mainly configured to transfer the input data in the external memory DDR to the unified memory 1406, or transfer the weight data to the weight memory 1402, or transfer the input data to the input memory 1401.

A vector computation unit 1407 includes a plurality of operation processing units, and if necessary, performs further processing of the output of the operation circuit, such as vector multiplication, vector addition, exponential operation, logarithmic operation, size comparison, etc. The vector computation unit is mainly used for non-convolution/FC layer network calculation in a neural network, such as pooling, batch normalization, local response normalization, etc.

In some implementations, the vector computation unit 1407 can store the vector of processed outputs to the unified memory 1406. For example, the vector computation unit 1407 may apply a non-linear function to the output of the operation circuit 1403, such as a vector of accumulated values, to generate an activation value. In some implementations, the vector computation unit 1407 generates a normalized value, a merged value, or both. In some implementations, the vector of processed outputs can be used as an activation input to the operation circuit 1403, for example, for use in a subsequent layer in the neural network.

The instruction fetch buffer 1409 connected to the controller 1404 is configured to store an instruction used by the controller 1404.

The unified memory 1406, the input memory 1401, the weight memory 1402, and the instruction fetch buffer 1409 are all on-chip memories. The external memory is private to the hardware architecture of this NPU.

In the present application, the operations of the layers in each module of the voice feature extraction model, the style separation model, the multi-head attention structure, the sub-feature extraction models, the encoder, the decoder and the attention structure in the Seq2seq neural

network, the dialect conversion models, and the voice enhancement model may be performed by the operation circuit 1403 or the vector computation unit 1407.

The operation circuit 1403 or the vector computation unit 1407 calculates a parameter value (for example, a first parameter value), and the main CPU is configured to read a computer program stored in the at least one memory, so that the terminal performs the method performed by the terminal in the foregoing method embodiments.

Referring to FIG. 15, an embodiment of the present invention further provides another voice conversion apparatus. As shown in FIG. 15, only parts related to the embodiment of the present application are shown, and for specific technical details that are not disclosed, reference may be made to the method part of the embodiments of the present application. The voice conversion apparatus may be a terminal, and the terminal may be a mobile phone, a tablet computer, a notebook computer, a smart watch, or the like. An example in which the terminal is a mobile phone is used.

FIG. 15 is a block diagram of a partial structure of a mobile phone related to a terminal provided by an embodiment of the present invention. Referring to FIG. 15, the mobile phone includes components such as a radio frequency (RF) circuit 1510, a memory 1520, an input unit 1530, a display unit 1540, an audio circuit 1560, a processor 1580, and a power supply 1590. It can be understood by those skilled in the art that the structure of the mobile phone shown in FIG. 15 does not constitute a limitation on the mobile phone, and the mobile phone may include more or fewer components than those shown in the figure, or some components may be combined, or a different component arrangement may be used.

Each component of the mobile phone will be described in detail below with reference to FIG. 15.

The RF circuit 1510 may be configured to receive and send a signal in an information receiving and sending process or a call process, and particularly, receive downlink information of a base station and send the downlink information to the processor 1580 for processing.

The memory 1520 may be configured to store a software program and a module, and the processor 1580 runs the software program and module stored in the memory 1520, to implement various functional applications and data processing of the mobile phone. The memory 1520 may

mainly include a storage program area and a storage data area, and the storage program area may store an operating system, an application required by at least one function (such as a sound playing function), etc. In addition, the memory 1520 may include a high-speed random access memory, and may also include a non-volatile memory, such as at least one magnetic disk storage device, a flash memory device, or other volatile solid-state storage devices.

The input unit 1530 may be configured to receive input digit or character information, and generate key signal input related to user setting and function control of the mobile phone. Specifically, the input unit 1530 may include a touch panel 1531 and other input devices 1532. The touch panel 1531 is also referred to as a touchscreen and may collect a touch operation by a user on or near the touch panel (such as an operation performed by the user on the touch panel 1531 or near the touch panel 1531 by using any proper object or accessory, such as a finger or a stylus), and drive a corresponding connection apparatus according to a preset program. Optionally, the touch panel 1531 may include two parts: a touch detection apparatus and a touch controller. The touch detection apparatus detects a touch orientation of a user, detects a signal brought about by the touch operation, and sends the signal to the touch controller. The touch controller receives touch information from the touch detection apparatus, converts the touch information into contact coordinates, and sends the contact coordinates to the processor 1580, and can receive and execute a command sent by the processor 1580. The input unit 1530 may further include other input devices 1532 in addition to the touch panel 1531. Specifically, other input devices 1532 may include, but are not limited to, one or more of a physical keyboard, function keys (such as volume control keys, switch keys, etc.), a trackball, a mouse, a joystick, etc.

In the present application, the input unit 1530 is configured to receive various operations inputted by a user, for example, a mode selection operation. A function of the input module 1320 in FIG. 13 may be performed by the input unit 1530, or a function of the acquisition module 1330 in FIG. 13 may be performed by the input unit 1530.

The display unit 1540 may be configured to display information inputted by the user or information provided for the user, and various menus of the mobile phone. The display unit 1540 may include a display panel 1541, and optionally, the display panel 1541 may be configured in a form of a liquid crystal display (LCD), an organic light-emitting diode (OLED), etc. Further, the touch panel 1531 may cover the display panel 1541, and when the touch panel 1531 detects a touch

operation on or near the touch panel, the touch operation is transmitted to the processor 1580 to determine a type of a touch event, and then the processor 1580 provides a corresponding visual output on the display panel 1541 according to the type of the touch event. Although in FIG. 15, the touch panel 1531 and the display panel 1541 are used as two separate components to implement input and output functions of the mobile phone, in some embodiments, the touch panel 1531 and the display panel 1541 may be integrated to implement the input and output functions of the mobile phone.

In the present application, the display unit 1540 is configured to display the app interfaces shown in FIGS. 6A, 6B, 11A, and 11B corresponding to the method embodiments.

The audio circuit 1560, a loudspeaker 1561, and a microphone 1562 may provide an audio interface between the user and the mobile phone. The audio circuit 1560 may transmit an electrical signal converted from received audio data to the loudspeaker 1561, and the loudspeaker 1561 may convert the electrical signal into a sound signal for output; on the other hand, the microphone 1562 may convert a collected sound signal into an electrical signal, and the audio circuit 1560 may receive the electrical signal, convert the electrical signal into audio data, and then output the audio data to the processor 1580 for processing.

In the present application, the audio circuit 1560 receives a first voice of a first speaker or a third voice of a second speaker by means of the microphone 1562. The loudspeaker 1561 is configured to output a processed second voice. For example, the second voice is a voice that has undergone style conversion, or the second voice is a voice that has undergone dialect conversion, or the second voice is a voice that has undergone voice enhancement. The loudspeaker 1561 is configured to output the second voice.

In a possible design, a function of the output module 1340 in FIG. 13 may be performed by the loudspeaker 1561.

The processor 1580 is a control center of the mobile phone, connects various parts of the entire mobile phone by using various interfaces and lines, and performs various functions of the mobile phone and data processing by running or executing a software program and/or a module stored in the memory 1520 and invoking data stored in the memory 1520, to perform overall monitoring on the mobile phone. Optionally, the processor 1580 may include one or more processing units.

The mobile phone further includes a power supply 1590 (such as a battery) for supplying power to each component. Preferably, the power supply may be logically connected to the processor 1580 via a power management system, so as to implement functions such as charging, discharging, and power consumption management through the power management system.

Although not shown in the figure, the mobile phone may further include a camera, a Bluetooth module, etc. Details are not described herein.

In the present application, when the process instruction stored in the memory is executed by the processor, the method performed by the terminal in the foregoing method embodiments is implemented. For details, refer to the descriptions in the foregoing method embodiments. Details are not described herein again.

The embodiments of the present application further provide a computer-readable storage medium, where the computer-readable storage medium stores a program, and when run on a computer, the program causes the computer to perform the steps performed by the terminal device in the method described in the foregoing method embodiments.

The embodiments of the present application further provide a computer program product, and when run on a computer, the computer program product causes the computer to perform the steps performed by the terminal in the method described in the foregoing method embodiments.

The embodiments of the present application further provide a circuit system, the circuit system includes a processing circuit, and the processing circuit is configured to perform the steps performed by the terminal device in the method described in the foregoing method embodiments.

In another possible design, when the apparatus is a chip in a terminal, the chip includes a processing unit and a communications unit, the processing unit may be, for example, a processor, and the communications unit may be, for example, an input/output interface, a pin, a circuit, etc. The processing unit may execute a computer-executable instruction stored in a storage unit, so that the chip in the terminal performs any one of the wireless communication methods according to the first aspect described above. Optionally, the storage unit is a storage unit in the chip, for example, a register or a cache, or the storage unit may be a storage unit in the terminal but outside the chip, for example, a read-only memory (ROM), another type of static storage device that can store static information and an instruction, a random access memory (RAM), etc.

The processor mentioned anywhere above may be a general-purpose central processing unit (CPU), a microprocessor, an application-specific integrated circuit (ASIC), or one or more integrated circuits configured to control program execution of the wireless communication method according to the first aspect described above.

Those skilled in the art can clearly understand that, for the convenience and conciseness of description, reference may be made to the corresponding processes in the foregoing method embodiments for specific working processes of the system, apparatus and unit described above, and details are not described herein again.

In several embodiments provided by the present application, it should be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the apparatus embodiments described above are merely illustrative, for example, the division of the units is merely a logical function division, and there may be other division manners in actual implementation. For example, multiple units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual coupling or direct coupling or communication connection may be indirect coupling or communication connection through some interfaces, devices or units, and may be in electrical, mechanical or other forms.

The units described as separate parts may or may not be physically separate, and parts displayed as units may or may not be physical units, may be located in one position, or may be distributed on a plurality of network units. Some or all of the units may be selected according to actual needs to achieve the objectives of the solutions of the embodiments.

In addition, functional units in the embodiments of the present application may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit. The integrated unit may be implemented in the form of hardware, or may be implemented in the form of a software functional unit.

When the integrated unit is implemented in the form of a software functional unit and sold or used as an independent product, the integrated unit may be stored in a computer-readable storage medium. On the basis of such an understanding, the technical solutions of the present application essentially, or the part of the technical solutions that contributes to the related art, or all or part of the technical solutions, may be embodied in the form of a software product, and the computer

software product is stored in a storage medium and includes several instructions for causing a computer device (which may be a personal computer, a server, a network device, or the like) to perform all or part of the steps of the methods described in the various embodiments of the present application. The foregoing storage medium includes various media capable of storing 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.

The foregoing embodiments are merely intended for describing the technical solutions of the present application other than limiting the present application. Although the present application has been described in detail with reference to the foregoing embodiments, those of ordinary skill in the art should understand that they may still make modifications to the technical solutions described in the foregoing embodiments or make equivalent replacements to some technical features thereof, and these modifications or replacements do not make the essence of the corresponding technical solutions depart from the spirit and scope of the technical solutions of the embodiments of the present application.

CLAIMS

1. A voice conversion method, characterized by comprising:

receiving a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode;

selecting a target conversion mode from a plurality of modes according to the mode selection operation, the plurality of modes comprising a style conversion mode, a dialect conversion mode, and a voice enhancement mode;

acquiring a first voice to be converted;

extracting feature information of the first voice;

inputting the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and outputting a converted second voice by means of the target voice conversion network; and

outputting the second voice.

2. The method according to claim 1, wherein the extracting feature information of the first voice comprises:

inputting the first information into a voice feature extraction model, and extracting a phoneme posterior probability (PPG) feature of the first voice by means of the voice feature extraction model, the PPG feature being used for retaining content information of the first voice.

3. The method according to claim 1 or 2, wherein when the target conversion mode is the style conversion mode, and the target voice conversion network is a style conversion network, the style conversion network comprises a style separation model and a voice fusion model, and the method further comprises:

acquiring a third voice for extracting a style feature; and

inputting the third voice into the style separation model, and separating a style feature of the third voice by means of the style separation model; and

the inputting the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and outputting a converted second voice by means of the target voice conversion network comprises:

inputting the style feature and the feature information of the first voice into the voice fusion model for fusion to obtain the second voice.

4. The method according to claim 3, wherein the style feature comprises a first feature, and the first feature comprises a plurality of sub-features; and

the inputting the third voice into the style separation model, and separating a style feature of the third voice by means of the style separation model comprises:

inputting the third voice into the style separation model, and extracting a vector of the first feature of the third voice by means of the style separation model;

inputting the third voice into a sub-feature extraction model, and extracting vectors of the sub-features by means of the sub-feature extraction model;

receiving a weight of each sub-feature among the plurality of sub-features inputted by the user; and

determining the style feature of the third voice based on the vector of the first feature, the vector of each sub-feature, and the weight of each sub-feature.

5. The method according to claim 4, wherein the determining the style feature of the third voice based on the vector of the first feature, the vector of each sub-feature, and the weight of each sub-feature comprises:

inputting the vector of the first feature into a multi-head attention structure, inputting a product of the vector of each sub-feature and the weight corresponding to the sub-feature into the multi-head attention structure, and outputting the style feature of the third voice by means of the multi-head attention structure.

6. The method according to claim 3, wherein the acquiring a third voice for extracting a style feature comprises:

receiving a template selection operation inputted by the user, the template selection operation being used for selecting a target template; and

acquiring a voice corresponding to the target template, and using the voice corresponding to the target template as the third voice.

7. The method according to claim 3, wherein the acquiring a third voice for extracting a style feature comprises:

receiving the third voice inputted by a second speaker, the first voice being a voice of a first speaker, and the second speaker being any person different from the first speaker.

8. The method according to claim 1 or 2, wherein when the target conversion mode is the dialect conversion mode and the target voice conversion network is a dialect conversion network, the inputting the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and outputting a converted second voice by means of the target voice conversion network comprises:

inputting the feature information of the first voice into the dialect conversion network, and outputting the second voice by means of the dialect conversion network, the first voice being a voice in a first dialect, and the second voice being a voice in a second dialect.

9. The method according to claim 8, wherein the dialect conversion network comprises a plurality of dialect conversion models, each dialect conversion model is targeted at a different dialect to be converted, and the method further comprises:

receiving a selection operation inputted by the user; and

inputting the feature information of the first voice into a dialect conversion model corresponding to the selection operation, and outputting the second voice by means of the dialect conversion model corresponding to the selection operation.

10. The method according to claim 8, wherein the method further comprises:

inputting the first voice into a style separation model, and separating a style feature of the first voice by means of the style separation model; and

the inputting the feature information of the first voice into the dialect conversion network, and outputting the second voice by means of the dialect conversion network comprises:

inputting the style feature of the first voice and the feature information of the first voice into the dialect conversion network, and outputting the second voice by means of the dialect conversion network, the second voice and the first voice having the same style.

11. The method according to claim 1 or 2, wherein the first voice is a far-field voice, and when the target conversion mode is the voice enhancement mode and the target voice conversion network is a voice enhancement model, the inputting the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and outputting a converted second voice by means of the target voice conversion network comprises:

inputting the feature information of the first voice into the voice enhancement model corresponding to the mode, and outputting the second voice by means of the voice enhancement model, the second voice being a near-field voice.

12. The method according to claim 11, wherein the method further comprises:

inputting the first voice into a style separation model, and separating a style feature of the first voice by means of the style separation model; and

the inputting the feature information of the first voice into the voice enhancement model corresponding to the mode, and outputting the second voice by means of the voice enhancement model comprises:

inputting the style feature of the first voice and the feature information of the first voice into the voice enhancement model, and outputting the second voice by means of the voice enhancement model, the second voice and the first voice having the same style.

13. The method according to any one of claims 1 to 12, wherein the acquiring a first voice to be converted comprises:

receiving the first voice inputted by the first speaker;

or,

selecting the first voice from a locally stored file.

14. A voice conversion apparatus, characterized by comprising:

an input module, configured to receive a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode;

a processing module, configured to select a target conversion mode from a plurality of modes according to the mode selection operation received by the input module, the plurality of modes comprising a style conversion mode, a dialect conversion mode, and a voice enhancement mode;

an acquisition module, configured to acquire a first voice to be converted;

the processing module further being configured to: extract feature information of the first voice acquired by the acquisition module; input the feature information of the first voice into a target voice conversion network corresponding to the target conversion mode, and outputting a converted second voice by means of the target voice conversion network; and

an output module, configured to output the second voice.

15. The apparatus according to claim 14, wherein

the processing module is further configured to input the first information into a voice feature extraction model, and extract a phoneme posterior probability (PPG) feature of the first voice by means of the voice feature extraction model, the PPG feature being used for retaining content information of the first voice.

16. The apparatus according to claim 14 or 15, wherein when the target conversion mode is the style conversion mode and the target voice conversion network is a style conversion network, the style conversion network comprises a style separation model and a voice fusion model;

the acquisition module is configured to acquire a third voice for extracting a style feature;

the processing module is further configured to input the third voice into the style separation model, and separate a style feature of the third voice by means of the style separation model; and

the processing module is further configured to input the style feature and the feature information of the first voice into the voice fusion model for fusion to obtain the second voice.

17. The apparatus according to claim 16, wherein the style feature comprises a first feature, and the first feature comprises a plurality of sub-features;

the processing module is further configured to: input the third voice into the style separation model, and extract a vector of the first feature of the third voice by means of the style separation model; and input the third voice into a sub-feature extraction model, and extract vectors of the sub-features by means of the sub-feature extraction model;

the input module is further configured to receive a weight of each sub-feature among the plurality of sub-features inputted by the user; and

the processing module is further configured to determine the style feature of the third voice based on the vector of the first feature, the vector of each sub-feature received by the input module, and the weight of each sub-feature.

18. The apparatus according to claim 16, wherein

the acquisition module is further configured to: receive a template selection operation inputted by the user, the template selection operation being used for selecting a target template; and acquire a voice corresponding to the target template, and use the voice corresponding to the target template as the third voice.

19. The apparatus according to claim 14 or 15, wherein the target conversion mode is the dialect conversion mode and the target voice conversion network is a dialect conversion network; and

the processing module is further configured to input the feature information of the first voice into the dialect conversion network, and output the second voice by means of the dialect conversion network, the first voice being a voice in a first dialect, and the second voice being a voice in a second dialect.

20. The apparatus according to claim 19, wherein the processing module is further configured to input the first voice into a style separation model, and separate a style feature of the first voice by means of the style separation model; and

the processing module is further configured to input the style feature of the first voice and the feature information of the first voice into the dialect conversion network, and output the second voice by means of the dialect conversion network, the second voice and the first voice having the same style.

21. The apparatus according to claim 14 or 15, wherein the first voice is a far-field voice, and when the target conversion mode is the voice enhancement mode and the target voice conversion network is a voice enhancement model,

the processing module is further configured to input the feature information of the first voice into the voice enhancement model corresponding to the mode, and output the second voice by means of the voice enhancement model, the second voice being a near-field voice.

22. A terminal device, characterized by comprising a processor, wherein the processor is coupled to a memory, the memory stores program instructions, and when the program instructions stored in the memory are executed by the processor, the method according to any one of claims 1 to 13 is implemented.

23. A computer-readable storage medium, characterized by comprising a program, wherein when run on a computer, the program causes the computer to perform the method according to any one of claims 1 to 13.

ABSTRACT

Provided are a voice conversion method and a related device to implement diversified human voice beautification. The method in embodiments of the present application includes: receiving a mode selection operation inputted by a user, the mode selection operation being used for selecting a voice conversion mode. A plurality of provided selectable modes include: a style conversion mode for performing speaking style conversion on a first voice to be converted; a dialect conversion mode for implementing accent addition or accent removal on the first voice; and a voice enhancement mode for implementing voice enhancement on the first voice. The three modes have corresponding voice conversion networks. According to a target conversion mode selected by the user, a target voice conversion network corresponding to the target conversion mode is selected to perform conversion on the first voice, and a converted second voice is outputted, thereby implementing diversified human voice beautification and meeting the requirements of the user in different application scenarios.

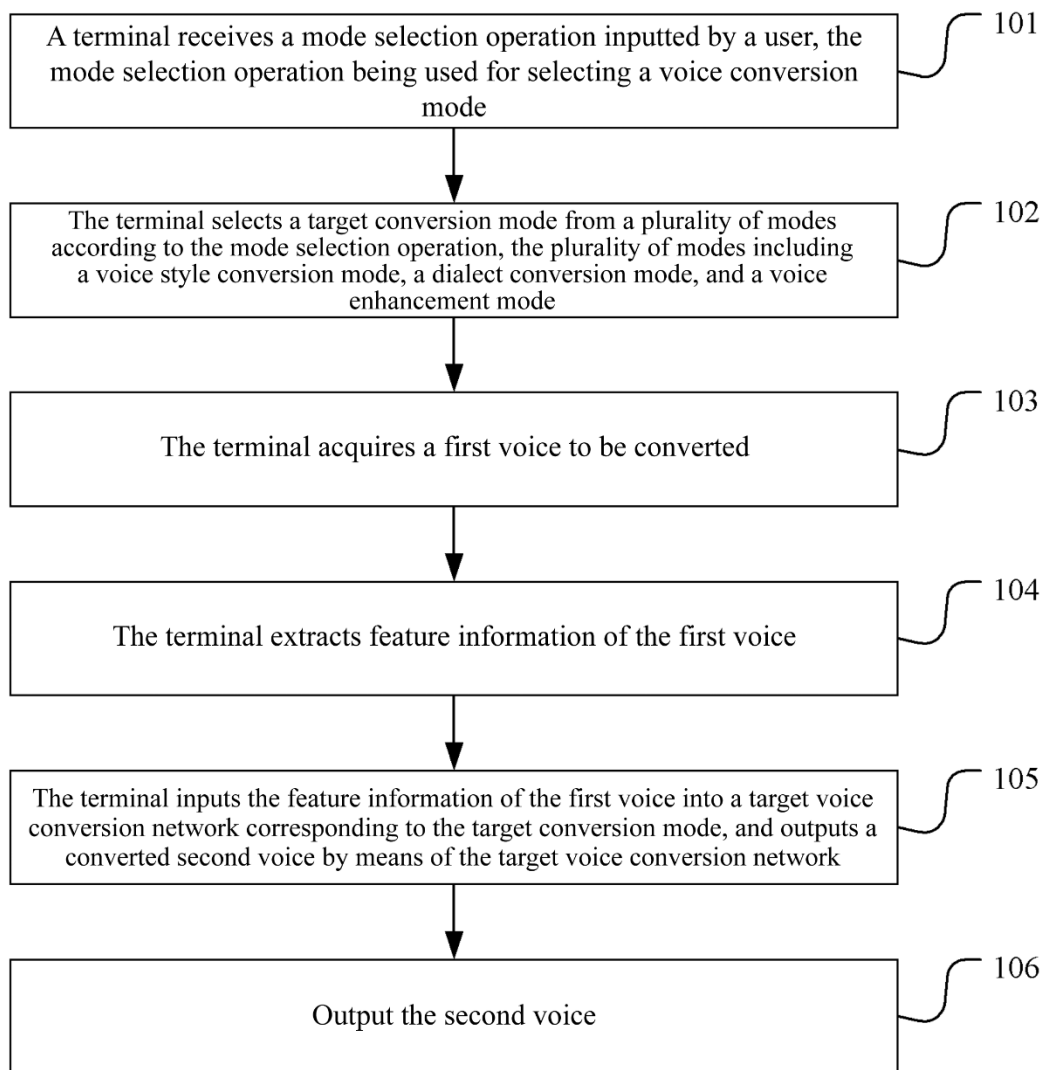


FIG. 1

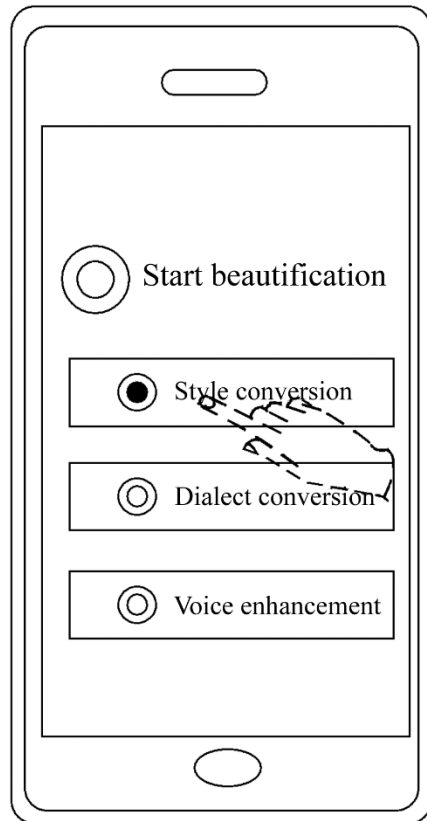


FIG. 2

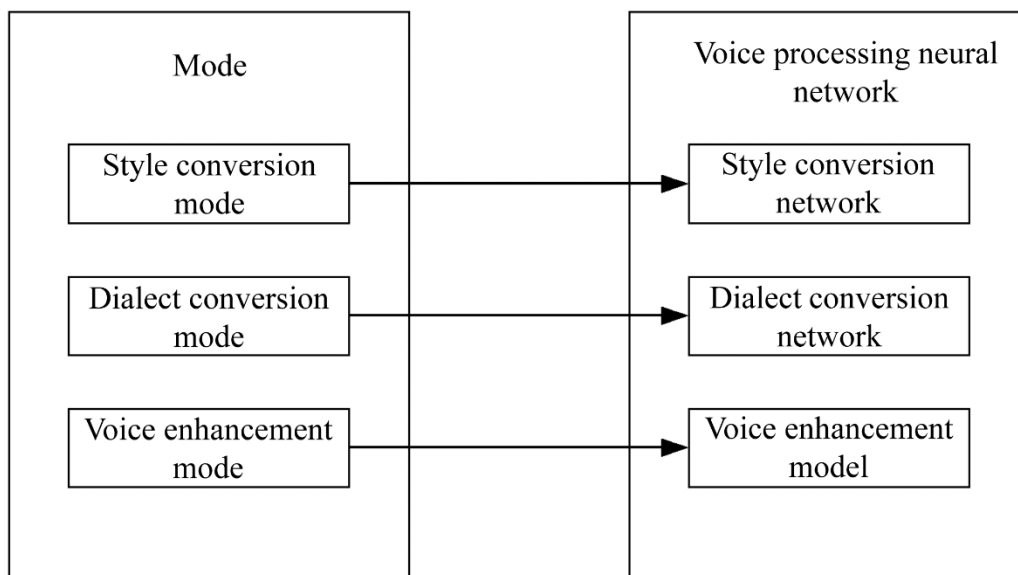


FIG. 3

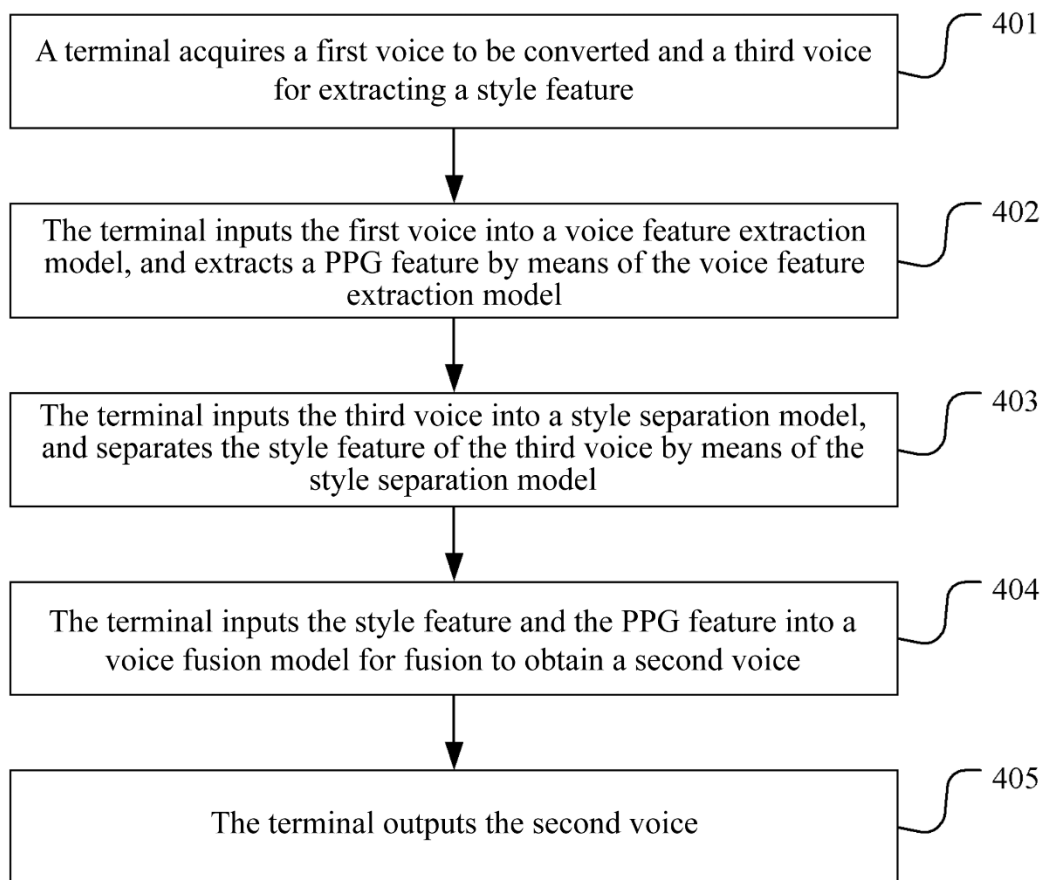


FIG. 4

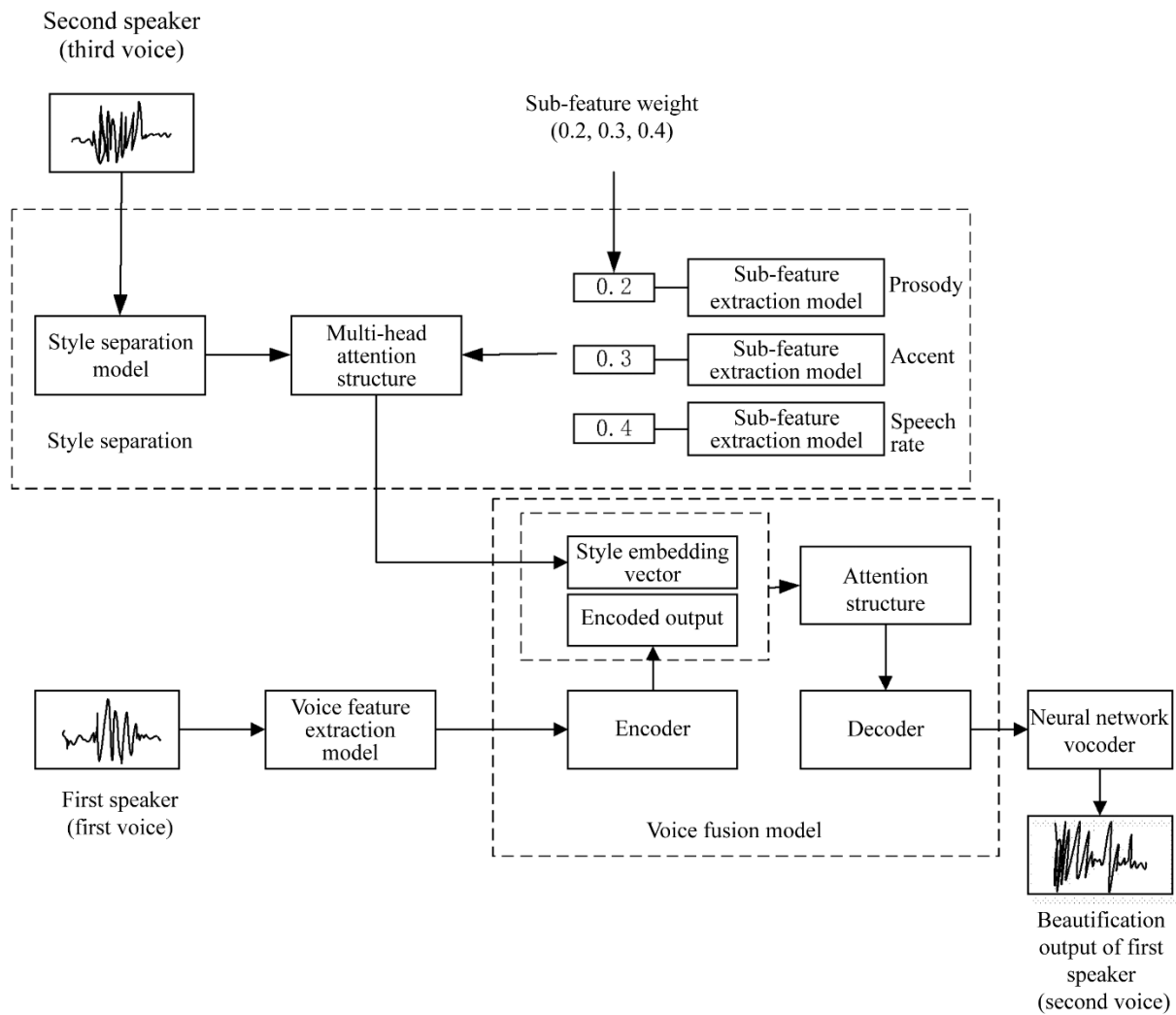


FIG. 5

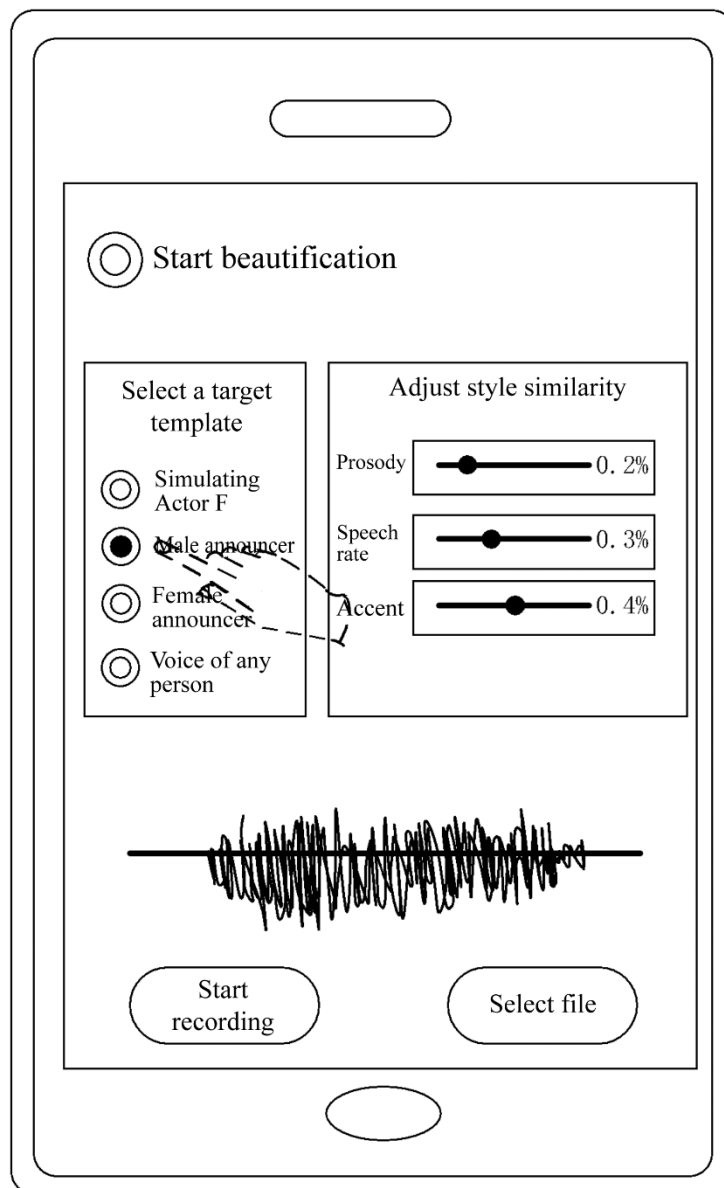


FIG. 6A

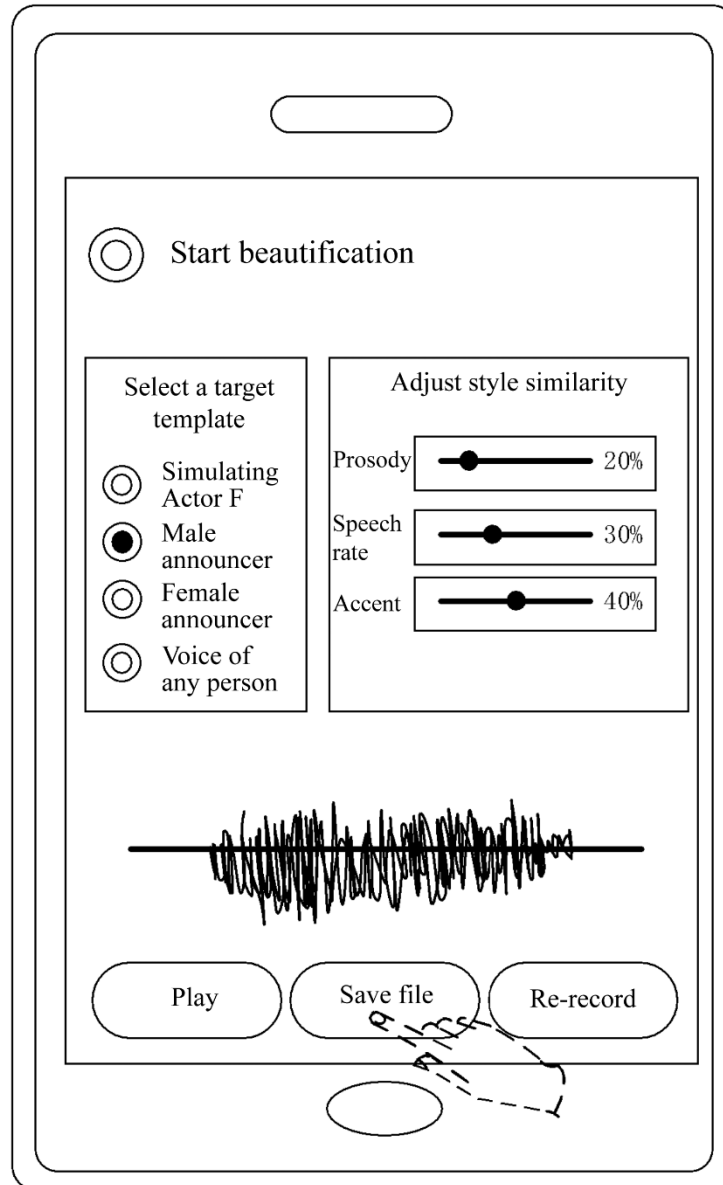


FIG. 6B

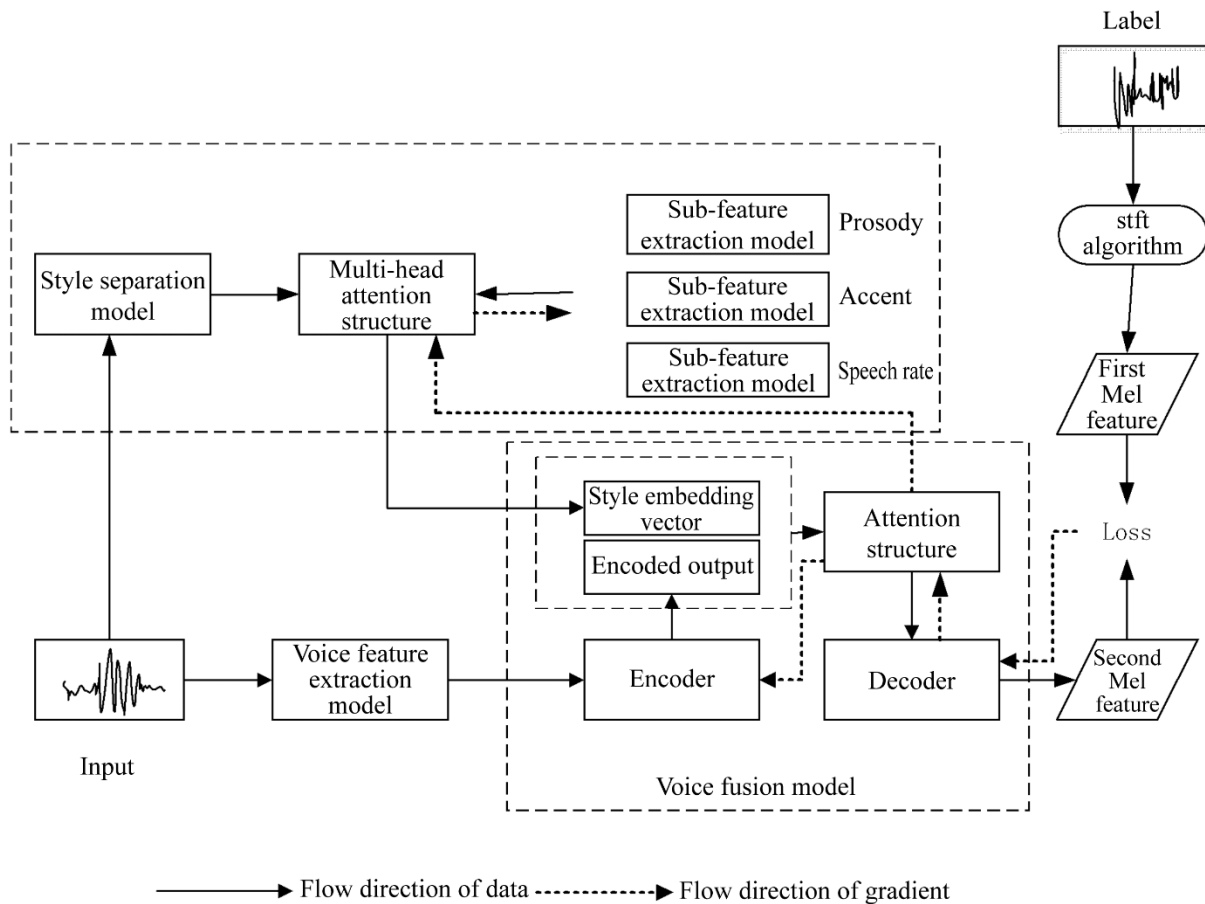


FIG. 7

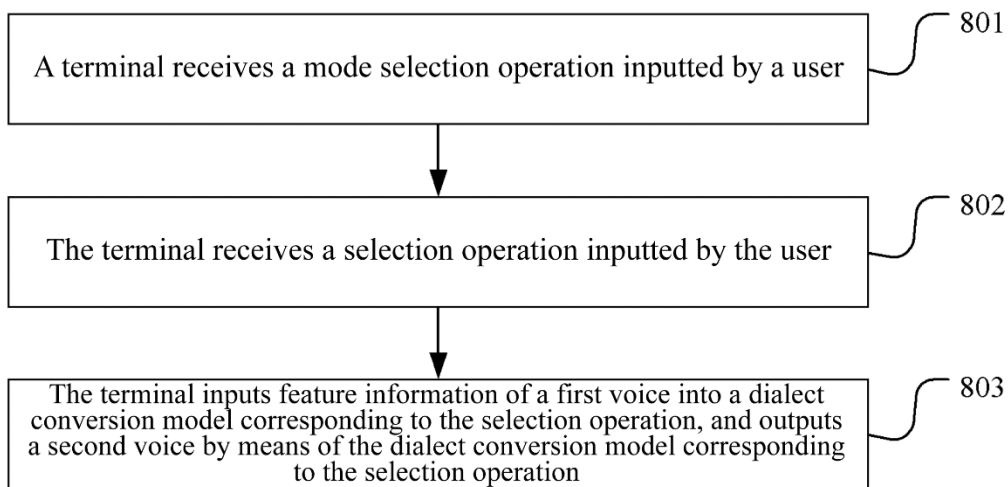


FIG. 8

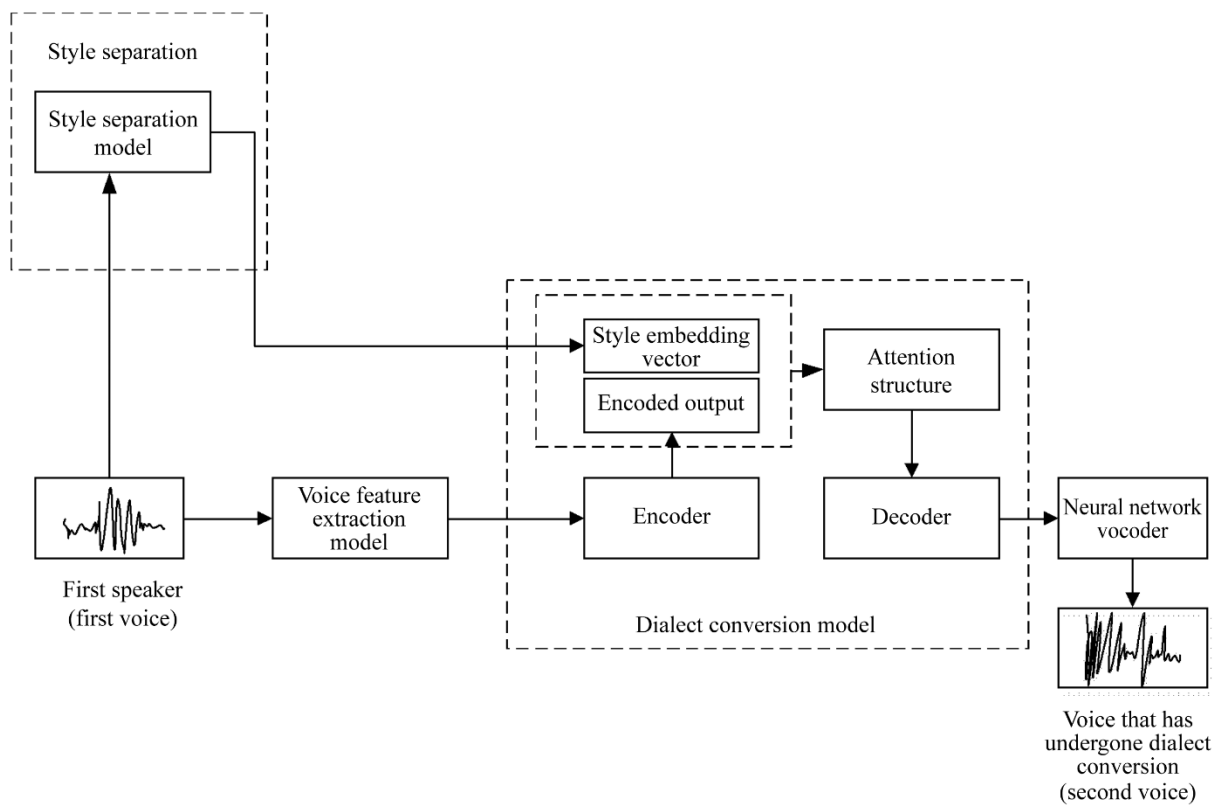


FIG. 9

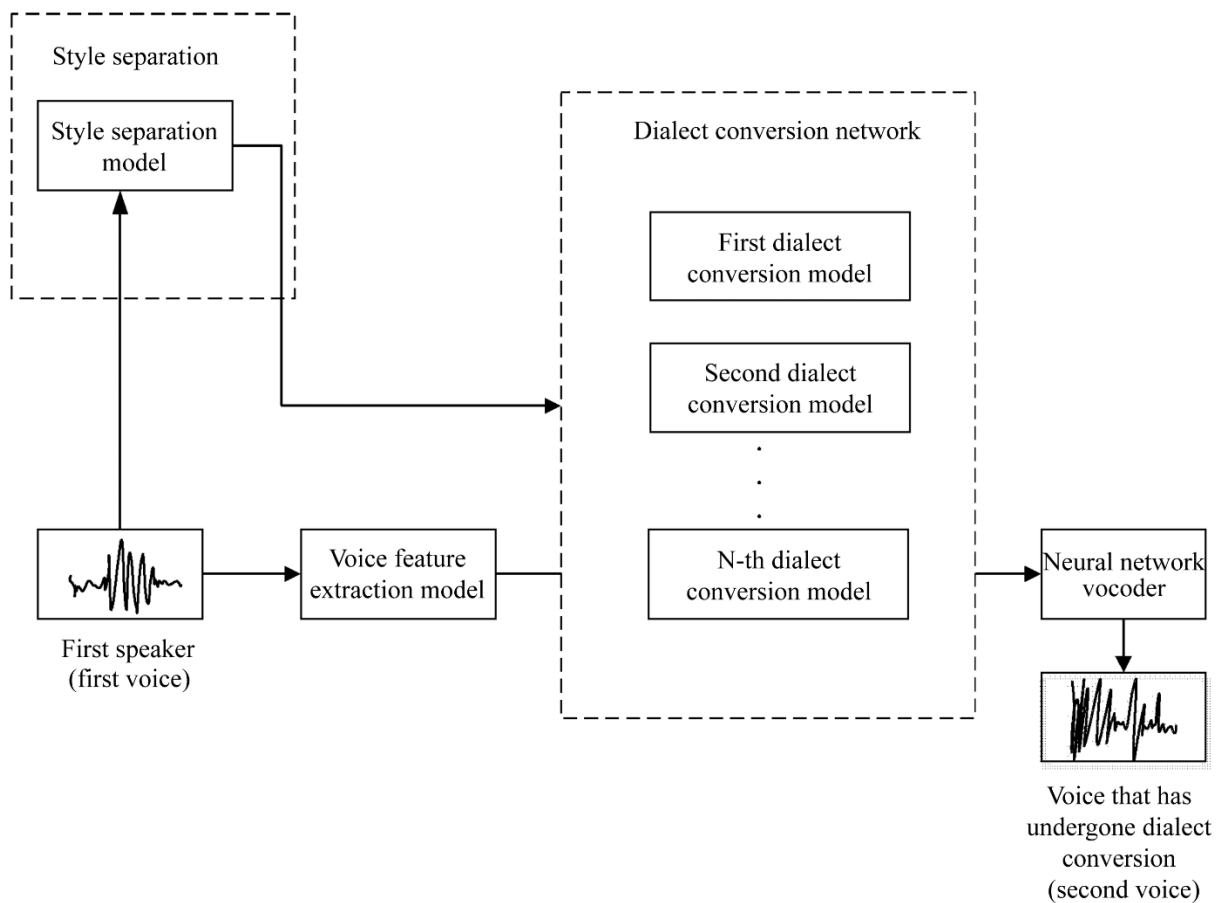


FIG. 10

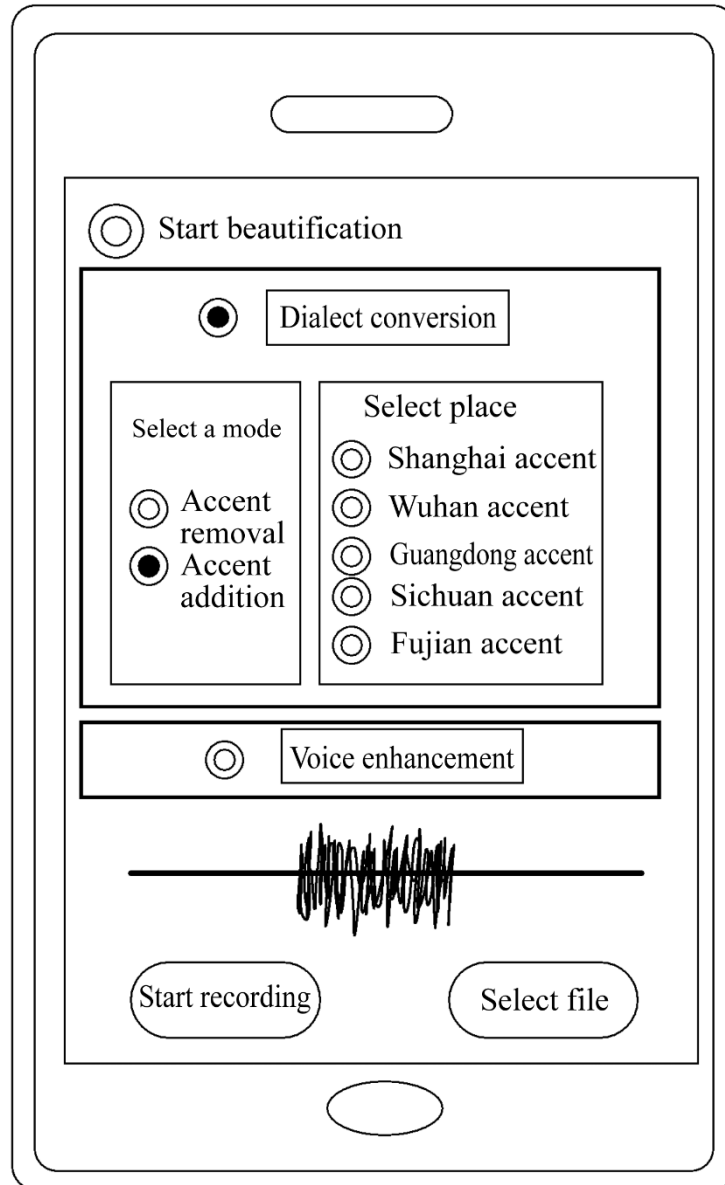


FIG. 11A

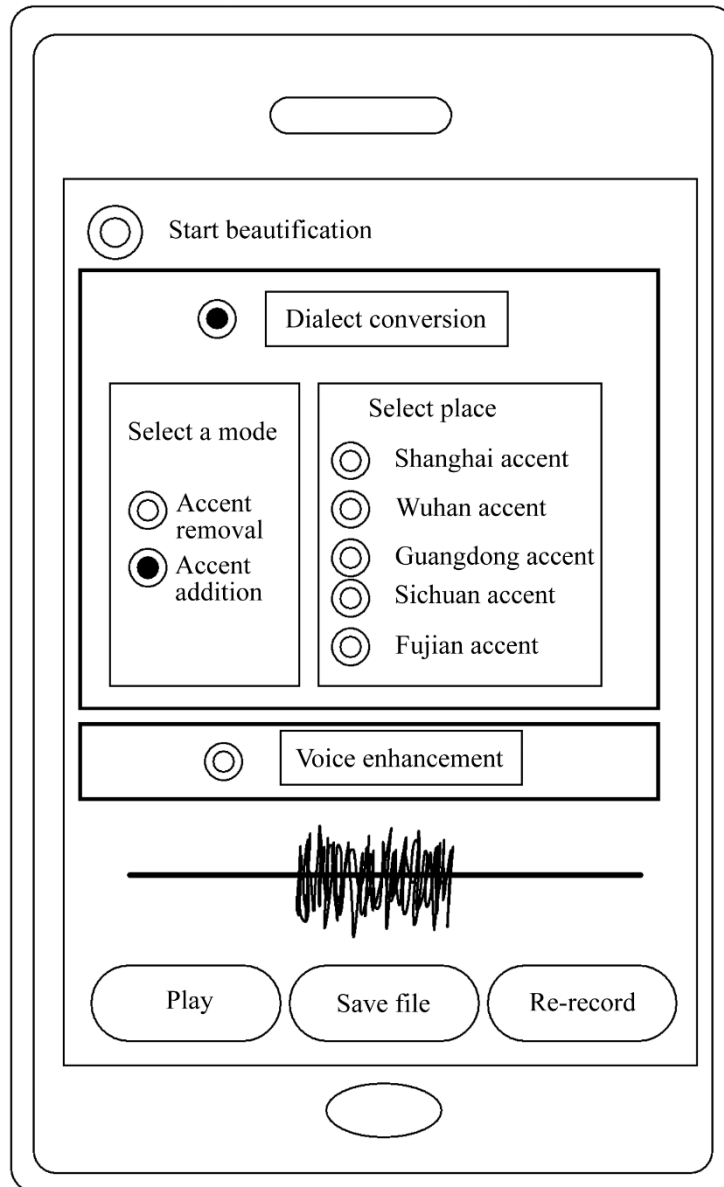


FIG. 11B

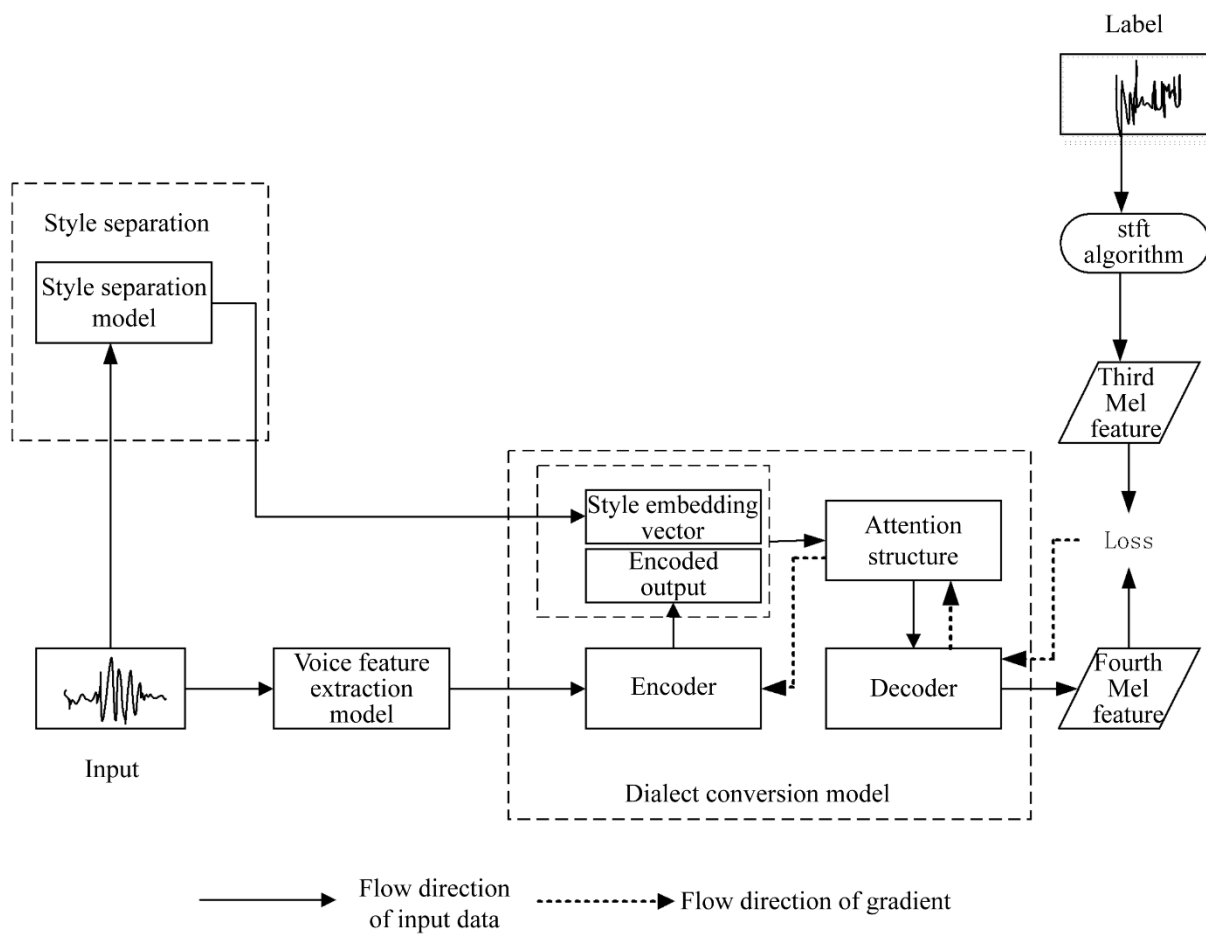


FIG. 12

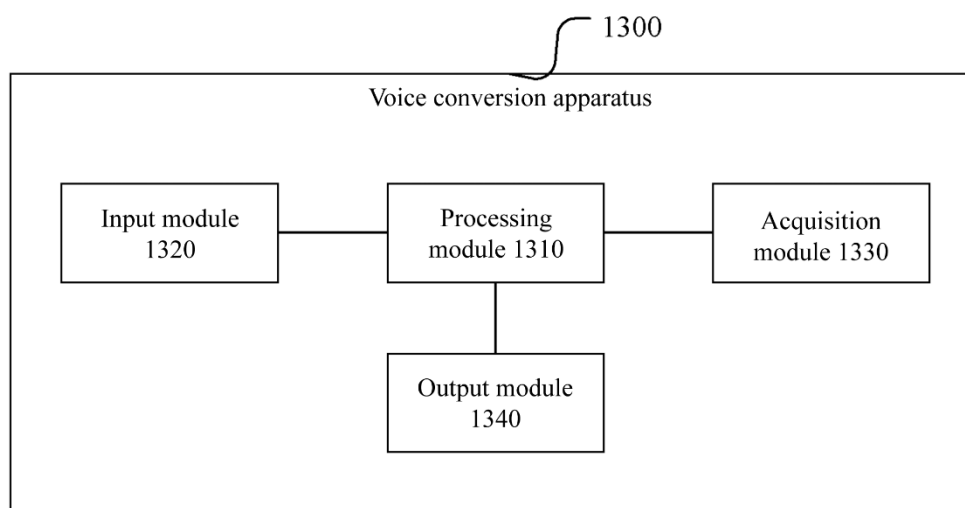


FIG. 13

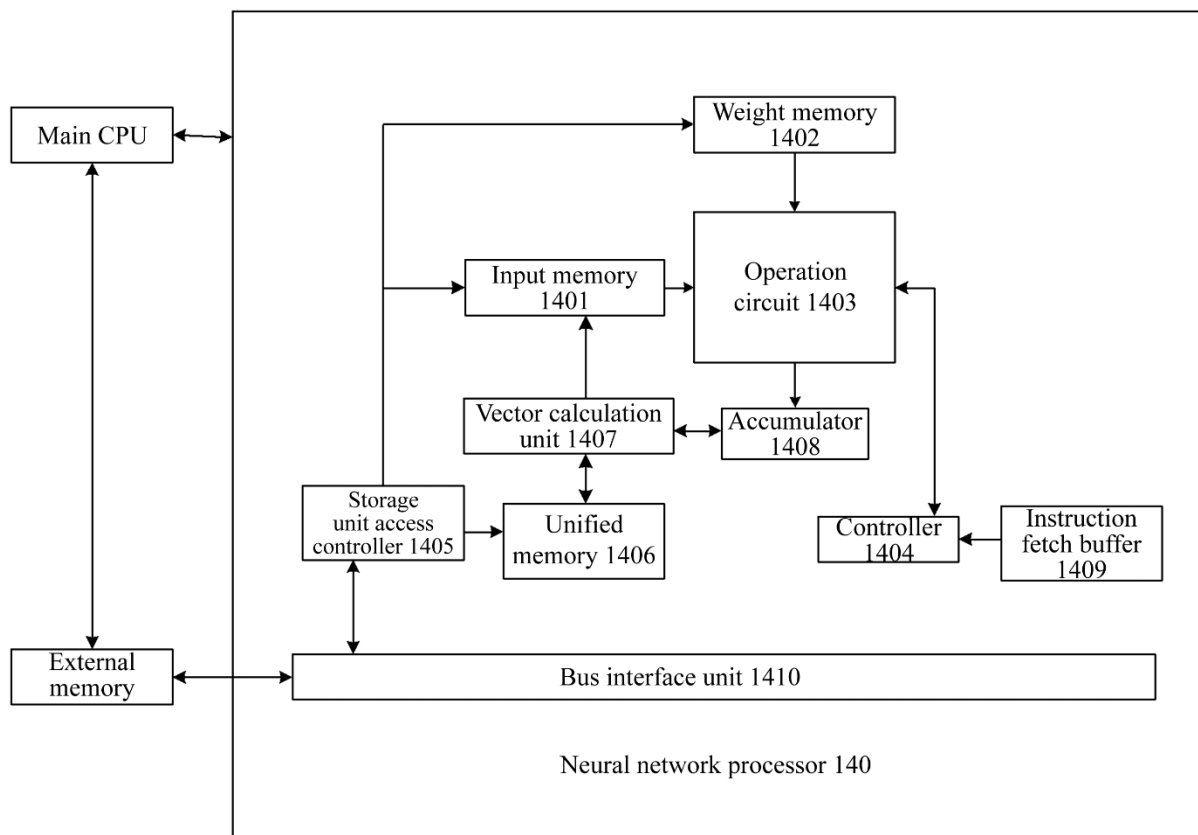


FIG. 14

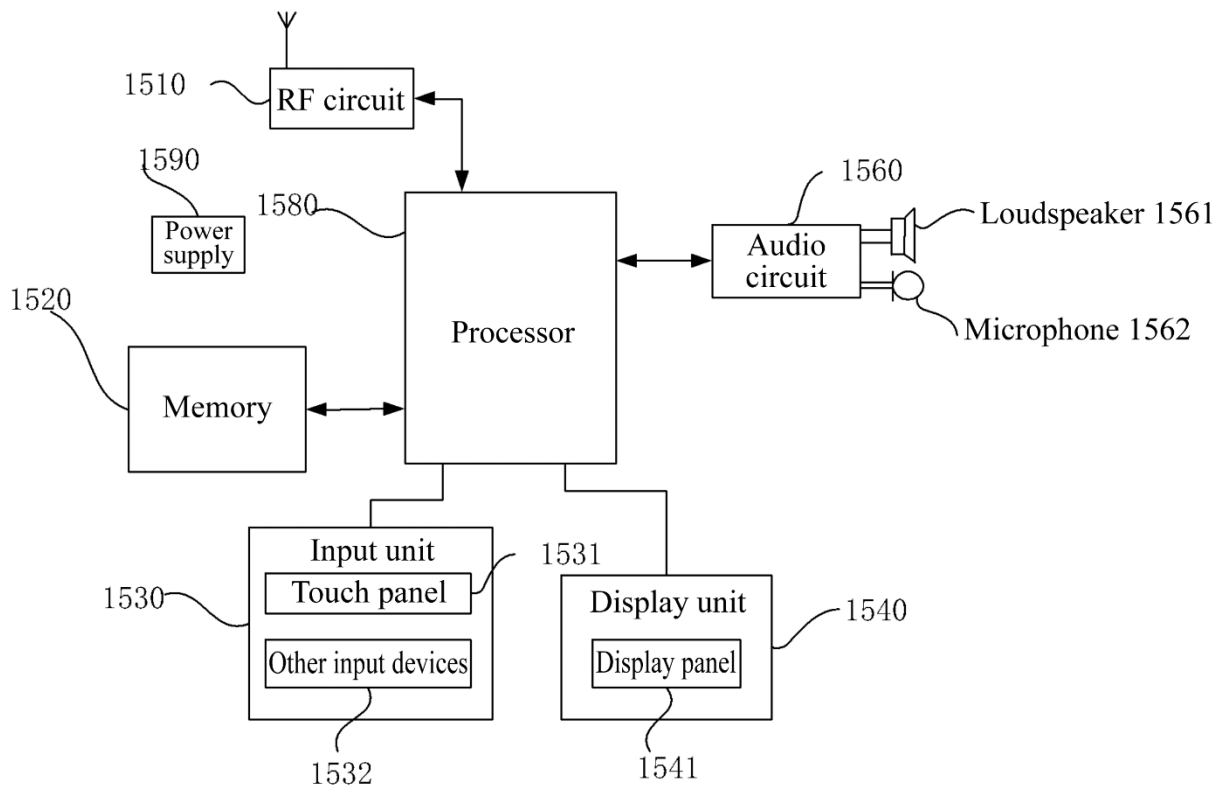


FIG. 15

Date: February 12, 2026

To whom it may concern:

This is to certify that the attached translation is an accurate representation of the documents received by this office. The translation was completed from:

- Chinese

To:

- English (USA)

The documents are designated as:

- [WO2022057759- Fan](#)

I, Zoe Brunell, Translator in this company, attest to the following:

I am a professional translator and I am knowledgeable of and well acquainted with the Chinese language and the English language.

The document in the English language attached hereto is a translation of [WO2022057759 - Fan](#) and is to the best of my ability, knowledge, and expertise, a true and correct English translation of the original document written in the Chinese language.

I declare under penalty of perjury that all statements made herein of my own knowledge are true and all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.


Signature of Zoe Brunell

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 3 月 24 日 (24.03.2022)



(10) 国际公布号
WO 2022/057759 A1

(51) 国际专利分类号:

G10L 13/02 (2013.01) *G10L 15/26* (2006.01)
G10L 13/033 (2013.01) *G10L 25/18* (2013.01)
G10L 13/04 (2013.01) *G10L 25/30* (2013.01)
G10L 15/02 (2006.01) *H04M 1/72433* (2021.01)
G10L 15/06 (2013.01) *H04M 1/72403* (2021.01)
G10L 15/16 (2006.01)

(21) 国际申请号: PCT/CN2021/117945

(22) 国际申请日: 2021 年 9 月 13 日 (13.09.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010996501.5 2020 年 9 月 21 日 (21.09.2020) CN

(71) 申请人: 华为技术有限公司 (HUAWEI TECHNOLOGIES CO., LTD.) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 范泛 (FAN, Fan); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 罗敬昊 (LUO, Jinghao); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。 李硕 (LI, Shuo); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦 B 座 18C2、18D、18E、18E2, Guangdong 518001 (CN)。

(54) Title: VOICE CONVERSION METHOD AND RELATED DEVICE

(54) 发明名称: 一种语音转换的方法及相关设备

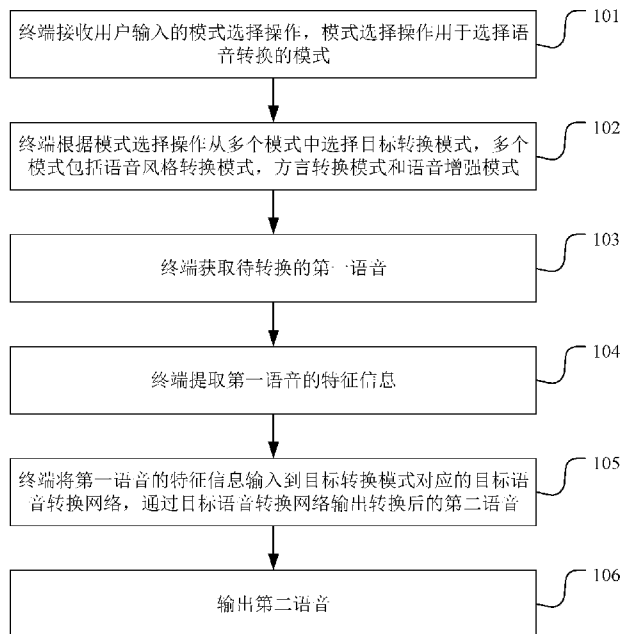


图 1

- 101 A terminal receives a mode selection operation input by a user, the mode selection operation being used for selecting a voice conversion mode
- 102 The terminal selects a target conversion mode from multiple modes according to the mode selection operation, the multiple modes comprising a voice style conversion mode, a dialect conversion mode, and a speech enhancement mode
- 103 The terminal obtains a first voice to be converted
- 104 The terminal extracts feature information of said first voice
- 105 The terminal inputs the feature information of said first voice to a target voice conversion network corresponding to the target conversion mode, and outputs a second voice after conversion by means of the target voice conversion network
- 106 Output the second voice

(57) Abstract: A voice conversion method and a related device, used for realizing the diversification of sound beautification. The method in embodiments of the present application comprises: receiving a mode selection operation input by a user, the mode selection operation being used for selecting a voice conversion mode; providing multiple available modes, comprising: a style conversion mode used for converting the speaking style of a first voice to be converted, a dialect conversion mode used for adding an accent to or removing an accent from said first voice, and a speech enhancement mode used for performing speech enhancement on said first voice, and the three modes having corresponding voice conversion networks; selecting, according to a target conversion mode selected by

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

the user, a target voice conversion network corresponding to the target conversion mode to convert said first voice; and outputting a second voice after conversion. In this way, the diversification of sound beautification is realized, and the requirements of the user under different application scenes are met.

(57) 摘要: 一种语音转换方法及相关设备, 用于实现人声美化的多样化。本申请实施例方法包括: 接收用户输入的模式选择操作, 所述模式选择操作用于选择语音转换的模式; 提供的多种可以选择的模式包括: 风格转换模式, 用于对待转换的第一语音进行说话风格转换; 方言转换模式, 用于对第一语音实现加口音或去口音; 语音增强模式, 用于对第一语音实现语音增强; 三种模式具有对应的语音转换网络, 根据用户选择的目标转换模式, 选择目标转换模式对应的目标语音转换网络对第一语音进行转换, 输出转换之后的第二语音, 从而实现人声美化的多样化, 满足用户在不同应用场景下的需求。

一种语音转换的方法及相关设备

本申请要求于 2020 年 9 月 21 日提交中国专利局、申请号为“202010996501.5”、申请名称为“一种语音转换的方法及相关设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及语音处理技术领域，尤其涉及一种语音转换的方法及相关设备。

背景技术

- 10 人声美化是指对声音进行修饰和美化，产生悦耳的听觉效果。目前市面上很多的录音软件或者社交应用（application，APP）的语音录制功能都会带有一定的人声美化的能力。例如，对语音进行去噪、提高语音的亮度或音量等，仅是对说话人自身的语音特点进行修饰，对于人声美化的模式过于单一。

发明内容

- 15 本申请实施例提供了一种语音转换的方法及相关设备，用于提供多种人声美化模式，实现人声美化的多样性。
- 第一方面，本申请实施例提供了一种语音转换方法，该方法应用于语音转换装置，例如，该装置可以为终端，该方法可以包括：终端接收用户输入的模式选择操作，模式选择操作用于选择语音转换的模式，然后，终端，根据模式选择操作从多个模式中选择目标转换模式，多个模式包括风格转换模式，方言转换模式和语音增强模式；终端可以选择目标转换模式对应的目标语音转换网络，通过目标语音转换网络实现语音转换；终端设备获取待转换的第一语音，进一步提取第一语音的特征信息，该特征信息用于保留第一语音的内容信息，然后，终端将第一语音的特征信息输入到目标转换模式对应的目标语音转换网络，通过目标语音转换网络输出转换后的第二语音，最后输出第二语音。
- 20 本实施例中，提供了多种可以选择的模式，如风格转换模式，用于对第一语音进行说话风格转换；方言转换模式，用于实现加口音或去口音；语音增强模式，用于实现语音增强；三种模式具有对应的语音转换网络，根据用户的需求，模式对应的语音转换网络可以对第一语音进行转换，从而实现人声美化的多样化，满足用户不同应用场景的需求。
- 在一种可选的实现方式中，提取第一语音的特征信息可以具体包括：终端将第一信息输入到语音特征提取模型，通过语音特征提取模型提取第一语音的音素后验概率 PPG 特征；PPG 特征用于保留第一语音的内容信息。PPG 特征描述的是语音帧所对应音素集中每个音素概率的大小，相当于对音素进行了识别，该 PPG 特征用于保留第一语音的内容信息，本示例中，不需要将语音转换成文本来保留语音的内容信息，而是直接通过语音输入，且通过 PPG 特征保留第一语音的内容信息，更能增加鲁棒性。
- 30 在一种可选的实现方式中，当目标转换模式为风格转换模式，目标语音转换网络为风格转换网络时，风格转换网络包括风格分离模型和语音融合模型，该方法还可以包括：终端获取用于提取风格特征的第三语音，并将第三语音输入到风格分离模型，通过风格分离
- 35

—2—

模型分离第三语音的风格特征；然后进一步的将风格特征和第一语音的特征信息输入到语音融合模型进行融合，得到第二语音。

本实施例中，终端接收待转换的第一语音和用于提取风格特征的第三语音，然后，第一语音输入到语音特征提取模型，通过语音特征提取模型提取 PPG 特征；PPG 特征用于保留第一语音的内容信息，通过 PPG 特征实现语音的直接输入。终端将第三语音输入到风格分离模型，通过风格分离模型分离第三语音的风格特征，最后，将风格特征和 PPG 特征输入到语音融合模型进行融合，得到融合了第一语音内容和第三语音风格的第二语音。第三语音可以是任意人的语音，从而实现将第一语音转换成任意人的语音风格，实现语音风格转换的多样性。在一种可选的实现方式中，风格特征包括第一特征，第一特征包括多个子特征；将第三语音输入到风格分离模型，通过风格分离模型分离第三语音的风格特征可以具体包括：终端将第三语音输入到风格分离模型，通过风格分离模型提取第三语音中的第一特征的向量，例如，该第一特征可以为音色，然后，将第三语音输入到子特征提取模型，通过子特征提取模型提取子特征的向量；接收用户输入的多个子特征中每个子特征的权重；根据第一特征的向量，每个子特征的向量及每个子特征的权重确定第三语音的风格特征。

本实施例中，通过用户输入的各子特征对应的权重，调整待转换语音的风格与第三语音（即目标模板对应的语音）的相似度，该相似程度由用户输入的权重决定。最终输出的语音风格与第三语音的相似程度由该权重决定，用户可以通过调整输入的权重，灵活调整待转换语音的风格，转换语音的风格既可以与第三语音的风格完全相同，也可以在第三语音的风格的基础上进行改变，从而实现待转换语音的风格多样化。

在一种可选的实现方式中，根据第一特征的向量，每个子特征的向量及每个子特征的权重确定第三语音的风格特征可以包括：终端将第一特征的向量输入多头注意力结构，且将每个子特征的向量及与其对应权重的乘积输入到多头注意力结构，通过多头注意力结构输出第三语音的风格特征。

本实施例中，多头注意力结构使得模型能够从不同的子空间提取特征表达，每个头对应高维空间中的一个子特征空间，相当于将高维空间分解，每个头负责一个子特征空间。多头注意力机制相当于多个结构相同的注意力机制，从而使得多头注意力机制输出的结果包含第三语音的部分音色。

在一种可选的实现方式中，获取用于提取风格特征的第三语音可以包括：

终端接收用户输入的模板选择操作，模板选择操作用于选择目标模板，例如，该目标模板可以是男播音员的风格，或者，可以为某一位“女播音员”的语音风格等，终端获取目标模板对应的语音，将目标模板对应的语音作为第三语音，将第三语音的风格特征融合到第一语音中，从而可以实现语音风格的多样化。

在一种可选的实现方式中，获取用于提取风格特征的第三语音可以具体包括：终端还可以接收第二说话人输入的第三语音，第一语音为第一说话人的语音，第二说话人为与第一说话人不同的任意人，从而可以实现语音风格的多样化。

在一种可选的实现方式中，当目标转换模式为方言转换模式，目标语音转换网络为方言转换网络时，将第一语音的特征信息输入到目标转换模式对应的目标语音转换网络，通

— 3 —

过目标语音转换网络输出转换后的第二语音，还可以具体包括：终端将第一语音的特征信息输入到方言转换网络，通过方言转换网络输出第二语音，第一语音为第一方言的语音，第二语音为第二方言的语音，从而实现方言转换，增强不同地域用户的交流便捷性，实现语音转换的多样化。

5 在一种可选的实现方式中，方言转换网络包括多个方言转换模型，每个方言转换模型分别针对不同的待转换方言，方法还包括：终端可以接收用户输入的选择操作；并且将第一语音的特征信息输入到选择操作对应的方言转换模型，通过选择操作对应的方言转换模型输出第二语音。本实施例中，终端可以根据用户输入的具体操作选择对应的方言转换模型。

10 在一种可选的实现方式中，方法还包括：终端将第一语音输入到风格分离模型，通过风格分离模型分离第一语音的风格特征；然后，终端将第一语音的风格特征和第一语音的特征信息输入到方言转换网络，通过方言转换网络输出第二语音，第二语音和第一语音的风格相同。

15 本实施例中，该第二语音内容与输入语音（第一语音）相同，且保留了输入语音（第一语音）的说话风格。

 在一种可选的实现方式中，第一语音为远场语音，当目标转换模式为语音增强模式，目标语音转换网络为语音增强模型时，将第一语音的特征信息输入到目标转换模式对应的目标语音转换网络，通过目标语音转换网络输出转换后的第二语音可以包括：终端将第一语音的特征信息输入到模式对应的语音增强模型，通过语音增强模型输出第二语音，第二
20 语音为近场语音。本示例中，将远场语音转换成近场语音，实现语音增强，以增加语音的清晰度，增加了应用场景，实现语音转换的多样化。

 在一种可选的实现方式中，方法还包括：终端将第一语音输入到风格分离模型，通过风格分离模型分离第一语音的风格特征；然后，将第一语音的风格特征和第一语音的特征信息输入到语音增强模型，通过语音增强模型输出第二语音，第二语音和第一语音的风格
25 相同。本示例中，转换之后的语音与输入语音（第一语音）相同，且保留了输入语音（第一语音）的说话风格。

 在一种可选的实现方式中，获取待转换的第一语音可以包括：接收第一说话人输入的第一语音；或者，从本地存储文件中选择第一语音。

 第二方面，本申请实施例提供了一种语音转换装置，该装置具有实现上述第一方面终端所执行的功能，该功能可以通过硬件实现，也可以通过硬件执行相应的软件实现；该硬件或软件包括一个或多个与上述功能相对应的模块。
30 第三方面，本申请实施例提供了一种终端，包括处理器，处理器和至少一个存储器耦合，处理器用于读取至少一个存储器所存储的计算机程序，使得终端执行上述第一方面中任一项的方法。

 第四方面，本申请实施例提供了一种计算机可读介质，所述计算机可读存储介质用于存储计算机程序，当所述计算机程序在计算机上运行时，使得所述计算机执行上述第一方面的方法。
35 第五方面，本申请实施例提供了一种计算机可读介质，所述计算机可读存储介质用于存储计算机程序，当所述计算机程序在计算机上运行时，使得所述计算机执行上述第一方面的方法。

—4—

第五方面，本申请提供了一种芯片系统，该芯片系统包括处理器，用于支持终端设备实现上述方面中所涉及的功能。在一种可能的设计中，所述芯片系统还包括存储器，所述存储器，用于保存终端设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包括芯片和其他分立器件。

5 附图说明

图1为本申请实施例中一种语音转换方法的一个实施例的步骤流程图；

图2为本申请实施例中选择模式的场景示意图；

图3为本申请实施例中各模式与各语音转换网络之间的对应关系的示意图；

图4为本申请实施例中一种语音转换方法的另一个实施例的步骤流程图；

10 图5为本申请实施例中语音进行风格转换过程的一个示例的示意图；

图6A为本申请实施例中风格转换模式的界面的一个场景示意图；

图6B为本申请实施例中风格转换模式的界面的另一个场景示意图；

图7为本申请实施例中风格转换网络训练并更新的示意图；

图8为本申请实施例中一种语音转换方法的另一个实施例的步骤流程图；

15 图9为本申请实施例中语音进行方言转换过程的一个示例的示意图；

图10为本申请实施例中语音进行方言转换过程的另一个示例的示意图；

图11A为本申请实施例中方言转换模式及语音增强模式的界面的一个场景示意图；

图11B为本申请实施例中方言转换模式及语音增强模式的界面的另一个场景示意图；

图12为本申请实施例中方言转换模型训练并更新的示意图；

20 图13为本申请实施例中一种语音转换装置的一个实施例的结构示意图；

图14为本申请实施例中一种芯片的一个示例的结构示意图；

图15为本申请实施例中一种语音转换装置的另一个实施例的结构示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行描述。本申请
25 中出现的术语“和/或”，可以是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本申请中字符“/”，一般表示前后关联对象是一种“或”的关系。本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换，以便这里描述的实施例
30 能够以除了在这里图示或描述的内容以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或模块的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或模块，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或模块。

本申请实施例提供了一种声音转换方法，该方法应用于终端，该终端可以为手机、平
35 板电脑，笔记本电脑，智能手表等具有语音功能的智能终端，该终端设备也可以称为终端设备，用户设备（user equipment，UE）等。该终端设备可以以手机为例进行说明。该终端设备内配置有语音处理神经网络，语音处理神经网络主要包括风格转换网络，方言转换网

— 5 —

络和语音增强模型。语音处理神经网络实现对语音的转换（美化）。本申请中提供了多种可以选择的模式，如风格转换模式，用于对第一语音进行说话风格转换；方言转换模式，用于实现加口音或去口音；语音增强模式，用于实现语音增强；三种模式具有对应的语音转换网络，终端根据用户选择的目标转换模式，选择目标转换模式对应的目标语音转换网络对获取到的待转换的第一语音进行转换，输出转换之后的第二语音，从而实现人声美化的多样化，满足用户不同应用场景的需求。

为了更好的理解本申请，首先对本申请中涉及的词语进行说明。

序列到序列（sequence to sequence, seq2seq）神经网络：seq2seq属于编码（encoder）-解码（decoder）结构的一种，输入的序列（sequence）通过编码层压缩成指定长度的向量，然后将该向量输入到解码层中得到输出的序列（sequence）。seq2seq神经网络指编码层和解码层都由神经网络构成。其中，编码层（encoder）用于通过一定规则对输入的序列进行编码生成向量。解码层（decoder）用于将生成的向量转化成输出的序列。

注意力（attention）结构：编码层中生成多个指定长度的向量，每个向量都由输入特征加权得到，权值与解码层的输出相关，其目的是让解码层的输出更关注输入的不同关键部分，而不是关注整段输入。

神经网络声码器（vocoder）：本质上是一种神经网络，用于将语音处理神经网络的输出特征转换为自然度高的语音。

风格特征：包括但不限于音色，音色是指声音的品质，又叫音品，音色可以反映说话人发出的声音特有的品质。音色包括但不限于韵律，重音和语速。

音素：是一个发音动作的最小语音单位，如 hao（好），一共有 3 个音素，wo（我），一种有 2 个音素。

音素后验概率（phoneme posteriorgram, PPG）特征：每个语音帧属于一组预先定义的语音单元（音素或三音素/助音素）的后验概率，这些语音单元保留了语音的语言和语音信息。

梅尔（mel）特征：语音信号存储为波形图特征是表示为其时域上的特征变换，但是仅根据波形图特征是不能看出语音信号的频率分布的。通过对语音信号频率域的分析，可以提取语音的特征。梅尔特征作为语音信号的特征表示，可以清晰表现语音共振峰特性。

请参阅图 1 所示，本申请提供了一种语音转换方法的一个实施例，该方法的执行主体为终端设备。或者，该方法的执行主体也可以为终端设备中的处理器，或者，该执行主体也可以是终端设备中的芯片。本申请实施例中，该方法的执行主体以终端为例，该终端以手机为例进行说明。终端内配置有语音处理神经网络，语音处理神经网络主要包括风格转换网络、方言转换网络和语音增强模型。

步骤 101、终端接收用户输入的模式选择操作，模式选择操作用于选择语音转换的模式。

请参阅图 2 所示，模式选择操作可以为点击操作，多个模式包括风格转换模式，方言

—6—

转换模式和语音增强模式。其中，每种模式具有对应的语音转换网络，通过对应的语音转换网络对语音进行转换。例如，风格转换模式用于转换语音风格。方言转换模式用于方言转换，例如“加口音”或者“去口音”等。语音增强模式用于将远场语音转换成近场语音。

5 步骤 102、终端根据模式选择操作从多个模式中选择目标转换模式，多个模式包括风格转换模式，方言转换模式和语音增强模式。

终端的显示界面上可以显示多个模式的列表，用户根据需求通过点击操作选择多个模式中的目标转换模式，该目标转换模式可以是上述三种模式中的任一种模式。

终端选择目标转换模式对应的目标语音转换网络。

10 请参阅图 3 所示，风格转换模式对应风格转换网络，方言转换模式对应方言转换网络，语音增强模式对应语音增强模型。终端根据模式与语音转换网络之间的对应关系，确定目标转换模式对应的目标语音转换网络。例如，当用户选择风格转换模式时，目标语音转换网络为风格转换网络。当用户选择言转换模式时，目标语音转换网络为方言转换网络。当用户选择语音增强模式时，目标语音转换网络为语音增强模型。

步骤 103、终端获取待转换的第一语音。

15 第一种实现方式，终端接收第一说话人输入的第一语音。

终端设备通过麦克风接收第一说话人输入的第一语音。终端设备接收用户（该用户可以和第一说话人为同一人，也可以为不同人）输入的第一操作，该第一操作为录制语音的操作。例如，该第一操作可以为点击操作。用户在手机屏幕上点击“开始录制”按钮，手机接收到该第一操作后，开始录制用户输入的语音，该语音作为用户希望被美化（或处理）
20 的语音（即第一语音）。

第二种实现方式，终端从本地存储文件中选择第一语音。

终端设备可以从本地存储文件中获取第一语音。终端设备接收用户输入的第二操作，该第二操作为选择语音文件的操作。例如，用户在手机屏幕上点击“选择文件”按钮，手机可以根据该第二操作显示待选择的语音列表，然后，终端设备接收用户输入的第三操作，
25 该第三操作用于选择目标语音，终端设备将该目标语音作为第一语音。

步骤 104、终端提取第一语音的特征信息。

该特征信息可以为梅尔频率倒谱系数（mel frequency cepstrum coefficient, MFCC）特征。或者，该特征信息为音素后验概率 PPG 特征。本申请实施例中，该第一语音的特征信息以 PPG 特征为例进行说明。PPG 特征描述的是语音帧所对应音素集中每个音素概率的大小，
30 相当于对音素进行了识别，该 PPG 特征用于保留第一语音的内容信息。

步骤 105、终端将第一语音的特征信息输入到目标转换模式对应的目标语音转换网络，通过目标语音转换网络输出转换后的第二语音。

第一种模式，目标语音转换网络为风格转换网络。示例性的，终端获取用于提取风格特征的第三语音，然后，再将第三语音输入到风格分离模型，通过风格分离模型分离第三
35 语音的风格特征；最后，将风格特征和第一语音的 PPG 特征输入到语音融合模型进行融合，得到第二语音。

第二种模式，将第一语音的特征信息输入到方言转换网络，通过方言转换网络输出第

—7—

二语音，第一语音为第一方言的语音，第二语音为第二方言的语音。

第三种模式，将第一语音的 PPG 特征输入到模式对应的语音增强模型，通过语音增强模型输出第二语音，第二语音为近场语音。

对于人声美化包括多个模式，多个模式包括风格转换模式，方言转换模式和语音增强模式，可以根据用户选择的模式对第一语音进行美化。例如，对第一语音的风格进行美化，对第一语音进行方言转换，或对第一语音进行语音增强等，实现人声美化的多样性。

步骤 106、终端输出第二语音。

终端通过扬声器输出转换后的第二语音。

本申请实施例中，提供了多种可以选择的模式，如风格转换模式，用于对第一语音进行说话风格转换；方言转换模式，用于实现加口音或去口音；语音增强模式，用于实现语音增强；三种模式具有对应的语音转换网络，根据用户的需求，模式对应的语音转换网络可以对第一语音进行转换，从而实现人声美化的多样化，满足用户不同应用场景的需求。

请参阅图 4 所示，本申请实施例提供了一种语音转换的另一个实施例，本实施例中针对风格转换模式进行说明，即对第一语音进行风格转换进行示例性说明。模式为风格转换模式，目标语音转换网络为风格转换网络。

步骤 401、终端获取待转换的第一语音及用于提取风格特征的第三语音。

本步骤中终端获取第一语音的方法请参阅图 1 对应的实施例中的步骤 103 的说明，此处不赘述。

终端获取第三语音的方式：

第一种实现方式中，终端接收用户输入的模板选择操作，终端根据模板选择操作选择目标模板对应的语音，然后将该目标模板对应的语音作为第三语音。例如，目标模板包括但不限于“男播音员”，“女播音员”，“演员声音”等等。

可选地，该目标模板也可以为一个类别。进一步的，终端设备接收子模板选择操作，终端设备根据子模板选择操作选择子模板对应的目标语音。例如，“男播音员”类别下的子模板列表包括“播音员 A”、“播音员 B”和“播音员 C”等。“女播音员”类别下的子模板列表包括“播音员 D”、“播音员 C”等。“演员声音”类别下的子模板列表包括“演员 D”、“演员 F”等。例如，用户选择“播音员 A”，终端设备根据用户输入的子模板选择操作选择“播音员 A”对应的语音作为第三语音。需要说明的是，本实施例中目标模板及子模板的示例，均是举例说明，并不造成限定。

第二种实现方式中，终端可以接收第二说话人输入第三语音。示例性的，终端设备接收用户输入的模板选择操作，该模板选择操作选择的目标模板是“任意人语音”，终端可以将该“任意人语音”对应的语音作为第三语音。例如，用户点击“任意人语音”选项，终端开始录制第三语音，该第二说话人是与第一说话人不同的任意人。

可选地，该第三语音可以是预先存储在本地的录制好的语音。例如，第三语音可以是电影中喜欢的角色的声音（卡通人物声音，或者某个演员的声音）。可选地，该第三语音也可以是预先从网上下载后，保存到终端设备的语音。

—8—

步骤 402、终端将第一语音输入到语音特征提取模型，通过语音特征提取模型提取第一语音的 PPG 特征，PPG 特征用于保留第一语音的内容信息。

请参阅图 5 所示，终端将第一语音输入到语音特征提取模型，通过语音特征提取模型提取第一语音的 PPG 特征。示例性的，语音特征提取模型可以为深度神经网络，语音特征提取模型包含多层的卷积层、两层 LSTM 层和一层全连接层，语音特征提取模型输出第一语音的 PPG 特征，PPG 特征描述的是语音帧所对应音素集中每个音素概率的大小，相当于对音素进行了识别，该 PPG 特征用于保留第一语音的内容信息。该语音特征提取模型预先根据大量的语料进行训练得到的。

本实施例中，使用 PPG 特征作为第一语音的内容信息输入，这些语音单元（音素）保留了语音的语言和语音信息。使用 PPG 特征相比自动语音识别技术（automatic speech recognition, ASR）更能增加鲁棒性，ASR 技术需要先将语音转换成文本，增加了语音内容识别错误的概率，本实施例中，通过使用 PPG 特征作为第一语音的内容信息输入，即直接可以实现语音输入，不需要转换成文本内容，增加系统的鲁棒性。

步骤 403、终端将第三语音输入到风格分离模型，通过风格分离模型分离第三语音的风格特征。

风格特征包括第一特征，第一特征包括多个子特征。例如，第一特征为音色特征，多个子特征包括韵律，重音和语速等。

风格分离模型用于分离第三语音的风格特征。该风格分离模型包括音色分离模型。该音色分离模型用于分离第三语音的音色特征，从而得到第一特征的向量（即音色特征向量）。

可选地，风格转换网络还包括多个子特征提取模型及多头注意力结构。终端将第三语音输入到子特征提取模型，通过子特征提取模型提取子特征的向量。例如，多个子特征提取模型包括韵律提取模型，重音提取模型和语速提取模型。例如，韵律提取模型用于提取第三语音中的韵律特征，得到韵律向量。重音提取模型用于提取第三语音中的重音特征，得到重音向量。语速提取模型用于提取第三语音中的语速特征，得到语速向量。

首先，终端接收用户输入的多个子特征中每个子特征的权重。一个示例中，手机的屏幕上显示韵律、语速和重音的调节条，用户可以通过对每个子特征的调节条的调节来输入每个子特征对应的权重。每个子特征对应的权重可以根据用户自己的需求灵活调整。例如，将“韵律”设为 10% 表示最终输出的语音跟目标模板的韵律有 10% 的相似度，即将一个值为 0.1 的参数传给了内置的子特征提取模型。

另一个示例中，预先配置几个档位。例如，可以分为三个档位，每个档位中各子特征的权重是根据经验值预先配置好的。例如，一档中，韵律的权重为 0.1，语速为 0.2，重音为 0.1。二档中，韵律的权重为 0.2，语速为 0.2，重音为 0.1 等等。终端通过接收用户输入的档位从而确定各子特征对应的权重。本示例中，不需要用户单独调整每个子特征的权重，用户只需要选择档位即可，方便用户操作。

然后，终端根据第一特征的向量，每个子特征的向量及每个子特征的权重确定第三语音的风格特征。

终端将用户输入的各子特征对应的权重与各子特征的向量相乘。将相乘之后的结果和

—9—

第一特征的向量（如音色向量）同时输入多头注意结构中进行注意对齐，使得多头注意结构输出风格向量，该风格向量为第三语音的风格特征。多头注意力结构使得模型能够从不同的子空间提取特征表达，每个头对应高维空间中的一个子特征空间，相当于将高维空间分解，每个头负责一个子特征空间。多头注意力机制相当于多个结构相同的注意力机制，从而使得多头注意力机制输出的结果包含第三语音的部分音色。例如，多头注意力机制输出的结构可以为一个 256 维的风格嵌入向量（风格特征）。

本实施例中，通过用户输入的各子特征对应的权重，调整待转换语音的风格与第三语音的相似度，该相似程度由用户输入的权重决定。用户可以选择是否输入该权重，若用户选择输入该权重，则最终输出的语音风格与第三语音的相似程度由输入的权重决定。若用户选择不输入该权重，则最终输出的语音风格将和第三语音的风格（或音色，风格以音色为例）完全相同。用户可以通过调整输入的权重，灵活调整待转换语音的风格，转换语音的风格既可以与第三语音的风格完全相同，也可以在第三语音的风格的基础上进行改变，从而实现待转换语音的风格多样化。例如，该转换语音的风格可以为某一位“男播音员”的语音风格；或者，可以为某一位“女播音员”的语音风格；或者，可以为任意人的语音风格，而且还可以在“男播音员”、“女播音员”或任意人的语音风格的基础上进行改变，从而实现语音风格的多样化。

步骤 404、终端将风格特征和 PPG 特征输入到语音融合模型进行融合，得到第二语音。

示例性的，该语音融合模型为seq2seq神经网络，在seq2seq神经网络中引入注意力机制。其中，该seq2seq神经网络包括编码层、解码层和注意力结构。编码层和解码层的模型可以采用任意神经网络模型的组合。例如，包括卷积神经网络（convolutional neural networks, CNN）、循环神经网络（recurrent neural network, RNN）、长短期记忆网络（long short-term memory, LSTM）中的任意种或任意两种的组合等等。例如，本申请中，该编码层可以包含三个卷积层和一个双向LSTM层。PPG特征首先会输入编码层，编码层将输入的PPG序列编码成一个固定维度的向量，由于输入序列长度可能较长，在解码时一个向量难以表达丰富的信息，因而引入了注意力机制。然后，风格特征的向量与编码层输出的PPG向量在宽度上拼接，得到注意力输入矩阵。然后，将这个注意力输入矩阵一帧一帧地送入注意力结构中，并注意力结构与编码层配合输出第二语音的梅尔（或也可以称为梅尔频谱）特征。梅尔频谱特征为第二语音的特征表示形式。在编码结束后，对编码模块的输出向量进行不同的加权得到若干向量，每个向量对应一个输出，从而保证了输出序列不再依据全部的输入序列，而是对输入序列有所聚焦。

步骤 405、终端输出第二语音。

该梅尔频谱特征为中间表示，梅尔频谱虽然蕴含了输出语音的信息，却不能直接播放，需要经过语音信号转化为梅尔谱图的逆过程。本申请中，可以通过声码器通过梅尔频谱的声学特征生成可播放的音频文件。该声码器可以为神经网络声码器，该网络负责将梅尔特征转换为高自然度的语音信号。该网络由多个卷积层和反卷积层构成，最终输出为可以播放的语音。语音中的内容与第一语音相同，说话风格与第三语音相同。该终端可以直接播放第二语音。可选的，用户可以选择是否保留第二语音。若用户选择保留该第二语音，终

端将该第二语音进行存储。

本实施例中，终端接收待转换的第一语音和用于提取风格特征的第三语音，然后，第一语音输入到语音特征提取模型，通过语音特征提取模型提取 PPG 特征，PPG 特征用于保留第一语音的内容信息，通过 PPG 特征实现语音的直接输入。终端将第三语音输入到风格分离模型，通过风格分离模型分离第三语音的风格特征，最后，将风格特征和 PPG 特征输入到语音融合模型进行融合，得到融合了第一语音内容和第三语音风格的第二语音。第三语音可以是任意人的语音，从而实现将第一语音转换成任意人的语音风格，实现语音风格转换的多样性。

在一个应用场景中，本申请中的语音处理神经网络内置于应用内。例如，该 APP 可以为即时通讯类的 APP，或者，人声美化 APP，人声美化 APP 可以应用于录音或录像的场景中。例如，用户可以通过该 APP 将待发送的语音进行美化。在录音录像场景中，用户利用终端录制完语音或带有语音的视频后，可以通过本发明组成的 APP 将语音进行美化。

请参阅图 6A 和 6B 所示，用户通过该 APP 进行人声美化的方法可以包括：

1)、用户点击屏幕上的“开始录制”按钮后，即可录制自己希望被处理的语音（第一语音）。或者，用户也可以点击“选择文件”，然后从本地文件中选取语音作为第一语音。并且可以将已读取到第一语音通过音频波形显示出来。

2)、用户在界面中选择目标模板，即选择第三语音。当用户选择“仿演员 F”、“男播音员”或“女播音员”时，则默认使用终端内已经保存的语音。当用户选择“任意人语音”时，则需要通过用户点击“开始录制”或“选择文件”按钮，将任意人的语音输入到 APP 中。

若用户点击“开始录制”按钮，则终端设备开始录制任意人的语音，即接收第二说话人输入的第三语音。若用户点击“选择文件”，则终端设备从本地文件中选择已经保存的任意人的语音，将该语音作为第三语音。

3)、用户在“风格相似度调节”下选择输入风格特征权重（也称为“子特征权重”）。例如，将“韵律”设为 20% 表示最终输出的语音跟目标模板的韵律有 20% 的相似度，app 实现将一个值为 0.2 的参数传给了内置的神经网络模型。同理，将“语速”设为 30% 表示最终输出的语音跟目标模板的语速有 30% 的相似度，将“重音”设为 40% 表示最终输出的语音跟目标模板的重音有 40% 的相似度。需要说明的是，可调节的风格特征包括但不限于音高、音强、音长、音色等。其中，音色包括但不限于韵律、语速和重音。

4)、用户点击“开始美化”按钮后，步骤 1、2 和 3 的输入将同时输进 APP 内置的已训练好的语音处理神经网络，经过语音处理神经网络处理之后输出一段处理完的语音（第二语音），该语音内容与第一语音的内容相同，该语音的风格与目标模板相似，相似程度由用于输入的风格特征权重决定。

5)、用户在 APP 界面可以看到处理完的语音波形，同时会有“播放”、“保存文件”和“重新录制”三个按钮。当用户点击“播放”时，该处理完的语音（第二语音）将通过手机的扬声器播放出来。当用户点击“保存文件”时，终端将该处理后的语音（第二语音）保存到本地。当用户点击“重新录制”时，终端开始进行重新处理，即回到步骤 4。

以上对风格转换网络及应用场景进行了说明。下面对该风格转换网络的训练和更新过程进行说明。

风格转换网络包括三个部分的神经网络，语音特征提取部分（即语音特征提取模型），
5 风格分离部分（即包括风格分离模型、多头注意力结构和多个子特征提取模型），语音融合模型。

请参阅图 7 所示，图 7 中黑色实线表示输入数据的流向，黑色虚线表示参数更新的方向。风格转换网络中，风格分离模型和语音特征提取模型不参与网络更新：

10 风格分离模型是预先使用大量不同说话人的语音数据训练好的。即风格分离模型的训练数据集包括大量不同说话人的语料（不同的人的语音具有不同的音色特征），通过该训练集对该风格分离模型进行训练，得到风格分离模型。该风格分离模型不参与整个网络的参数更新。

语音特征提取模型是预先通过大量的语料训练好的，例如，训练数据集中包括的样本
15 数据的输入为语音，标签为该输入语音的 PPG 特征。语音特征提取模型用于提取 PPG 特征，该语音特征提取模型不参与整个网络的参数更新。

整个网络中除了风格分离模型和语音特征提取模型之外，其他的模型参与整个网络的训练和更新。

请参阅图 7 中黑色虚线所经过的模型。其中，多头注意结构、各子特征提取模型、Seq2seq
20 神经网络（包括编码层、解码层及注意力结构）中，每个结构都可以是一个神经网络，每个神经网络可以包括多个卷积层、全连接层、LSTM 层等结构层，而各层的权重参数需要通过大量的数据训练得到。

本申请中对于风格转换网络的更新与传统方法不同之处在于：由于本申请中风格分离模型和语音特征提取模型已经是预先训练好的，不参与网络更新，即不需要考虑风格分离
25 模型的训练，由此，输入与标签不需要是同样内容的两条语音（两个不同的人说的两条语音），极大的减少了样本数据的数量。标签通过短时傅里叶变换（short-time fourier transform, STFT）算法得到第一梅尔特征，将语音输入到风格转换网络后，通过本网络的输出为第二梅尔特征，通过比对第一梅尔特征和第二梅尔特征得到 loss 值和梯度，本网络的输出与输入通过 STFT 算法得到的梅尔特征进行比对得到 loss 值及其梯度，图 7 中黑色虚线为梯度
30 流动方向，当梯度流过相应结构时，结构中的参数才会进行更新，而风格分离模型和语音特征提取模型是预先训练得到，不需要参与本网络的参数更新。Loss 值作为判断网络何时停止训练的指标，当 loss 值下降到一定值，且无明显继续下降趋势时，说明网络已收敛，即可停止训练。

整个风格转换网络获得语音美化的能力并没有通过使用风格标签（即输入与标签为同
35 一条语音序列），这种网络学习方式属于无监督学习方式，通过无监督学习方式训练的风格转换网络，可以支持用户手动调节对所录语音美化的程度并可组合不同说话人的说话风格。

请参阅图 8 所示, 本申请实施例提供了一种语音转换的另一个实施例, 本实施例中对方言转换模式和语音增强模式进行说明。

方言转换网络包括多个方言转换模型, 其中, 方言转换模型用于对输入的第一语音进行方言转换。语音增强模型用于将远场语音转换成近场语音, 实现语音增强。

5 多个方言转换模型至少包括两类, 第一类是去口音, 即将方言转换成普通话。例如, 四川话转换成普通话。第二类是加口音, 即将普通话转换成方言, 例如, 普通话转换成四川话等。每个方言转换模型分别针对不同的待转换方言。方言转换可以增强不同地域用户的交流便捷性, 实现语音转换的多样化。需要说明的是, 这里仅是举了两类进行示例性说明, 当然在可选的方案中, 也可以将两个方言转换模型进行联合使用, 从而进行两种方言之间的转换, 例如, 将四川话转换成广东话, 可以先将四川话转换成普通话, 再将普通话转换成广东话。

步骤 801、终端接收用户输入的模式选择操作。

例如, 终端的用户界面上显示对语音的不同的处理模式。例如, 该处理模式可以包括“方言切换模式”和“语音增强模式”。该模型选择操作可以为点击操作, 当用户点击“方言切换模式”时, 则用户选择方言切换处理模式。

步骤 802、终端接收用户输入的选择操作。

可选地, 当模式选择操作用于选择方言切换模式时, 模式选择操作(第一个层级的操作)还可以包括下一层级(第二个层级的操作)的多个选择操作。例如, 第一选择操作和第二选择操作。该第一选择操作用于选择“去口音”, 而第二选择操作用于选择“加口音”。
20 每一个第二层级的操作又会包括多个第三层级的操作。在一个应用场景中, 当用户选择“口音方言切换”时, 会出现二级分类“去口音”, “加口音”选项。当用户选择“加口音”时, 终端的用户界面又会显示不同地方口音选项的三级分类标签。例如, “四川口音”、“广东口音”等。终端会接收用户输入的第三个层级的选择操作, 如第一操作, 第二操作等。其中, 第一操作用于选择将四川话转换成普通话, 第二操作用于选择将普通话转换四川话等。终端
25 可以根据用户输入的具体操作选择对应的方言转换模型。

步骤 803、终端将第一语音的特征信息输入到选择操作对应的方言转换模型, 通过选择操作对应的方言转换模型输出第二语音。

请参阅图 9 所示, 当模式选择操作用于选择方言切换模式时, 终端根据模式选择操作, 将 PPG 特征输入到方言切换模型(seq2seq 神经网络), 通过方言切换模型输出第二语音的梅尔特征形式, 然后, 通过神经网络声码器将梅尔特征转换成可播放的语音形式。第一语音为第一方言的语音, 第二语音为第二方言的语音。

可选地, 终端将第三语音输入到风格分离模型, 通过风格分离模型分离第一语音的风格特征, 得到第一语音的音色特征向量。然后, 将第一语音的 PPG 特征和第一语音的音色特征向量输入到选择操作对应的方言转换模型, 通过选择操作对应的方言转换模型输出第二语音。该第二语音内容与输入语音(第一语音)相同, 且保留了输入语音(第一语音)的说话风格。

需要说明的是, 本申请实施例中, 方言为“地方语言”, 例如, 四川话, 广东话等等。

- 13 -

而普通话也为方言的一种，汉语中，普通话是以北京语音为标准音，可以理解为以北方话为基础的方言，即在本申请中，普通话为也方言的一种。

示例性的，第一方言为四川话，第二方言为普通话。或者，第一方言为普通话，第二方言为四川话，或者，第二方言为东北话等等，具体的，并不限定第一方言和第二方言。

5 请参阅图 10 所示，方言转换网络包括 N 个方言转换模型，每个方言转换模型分别针对不同的待转换方言，每一种方言转换都具有对应的模型。例如，多个方言转换模型中的第一方言转换模型用于四川话转换为普通话，而第二方言转换模型用于将普通话转换为广东话，而第一方言转换模型和第二方言转换模型的整合模型可以将四川话转换为广东话等等。本示例中，对于第一方言和第二方言进行举例说明，并不造成对本申请的限定。

10 可选地，当模式选择操作对应的模式为语音增强模式时，终端将第一语音的 PPG 特征输入到模式对应的语音增强模型，通过语音增强模型输出第二语音，第一语音为远场语音，第二语音为近场语音。将远场语音转换成近场语音，从而实现语音增强，增加了应用场景，实现语音转换的多样化。语音增强模型是通过训练数据集中的样本数据进行学习得到的，样本数据包括输入和标签，输入为远场语音，标签为近场语音。

15 可选地，将所述第一语音输入到风格分离模型，通过所述风格分离模型分离所述第一语音的风格特征。然后，将第一语音的风格特征和第一语音的特征信息输入到所述语音增强模型，通过语音增强模型输出所述第二语音，所述第二语音和所述第一语音的风格相同。本示例中，转换之后的语音与输入语音（第一语音）相同，且保留了输入语音（第一语音）的说话风格。

20

示例性的，在一个应用场景中，请参阅图 11A 和 11B 所示。

a、用户点击“开始录制”按钮后，终端录制自己希望被处理的语音（第一语音），用户也可以点击“选择文件”按钮，终端从本地文件中选取一段语音作为第一语音。可选的，终端可以显示已读取到音频的波形。

25 b、用户在界面中选择语音处理模式。界面显示一级分类标签。例如，该一级分类标签包括“方言切换”和“语音增强”。当用户选择“方言切换”时，界面显示二级分类标签，如二级分类标签包括：“去口音”，“加口音”。当用户选择“加口音”时，界面会显示不同地方口音选项的三级分类标签，例如：“广东口音”、“四川口音”、“福建口音”等。或者，用户也可以选择“语音增强”模式。

30 c、终端会根据用户选择的模式选取对应的模型。

d、用户点击“开始美化”按钮后，将第一语音输入已选择好的模型中（例如，普通话转广东口音的方言转换模型），经过方言转换模型处理一段时间后，输出一段处理完的第二语音，该第二语音内容与输入语音（第一语音）相同，且保留了输入语音（第一语音）的说话风格。或者，将第一语音输入已选择好的语音增强模型中，经过语音增强模型处理一段
35 时间后，输出一段处理完的第二语音，该第二语音内容与输入语音（第一语音）相同，且保留了输入语音（第一语音）的说话风格。第一语音为远场语音，第二语音为近场语音。

e、最后用户在显示界面可以看到处理完的语音波形。显示界面还会显示“播放”、“保

存文件”和“重新录制”三个按钮。当用户点击“播放”按钮时，处理完的语音将通过手机的扬声器播放出来。当用户点击“保存文件”按钮时，终端将处理后的语音保存到本地。当用户点击“重新录制”按钮时，终端重新对该第一语音进行处理，根据此时界面上用户的选择操作选取对应的模型，同时回到步骤 d。

5

下面对方言转换模型训练和更新过程进行说明。

请参阅图 12 所示，每个方言转换模型的结构相同，每个方言转换模型都为 seq2seq 神经网络，该 seq2seq 神经网络包括编码层、解码层及注意力结构。不同的方言转换模型对应的训练数据集不同，各层的参数不同。例如，第一方言转换模型为四川话转普通话，该

10 第一方言转换模型对应的训练数据集中的每个样本数据包括输入和标签。其中，输入为四川话，标签为与该四川话相同内容的普通话。第二方言转换模型为普通话转四川话，该第一方言转换模型对应的训练数据集中的每个样本数据包括输入和标签，输入为普通话，标签为与该普通话相同内容的四川话等等。不同的方言转换模型对不同的训练数据集进行学习。图 12 中以一个方言转换模型中各模块的参数更新进行示例。

15 本训练方案与风格转换网络的训练方式基本相同，风格分离模型和语音特征提取模型已经是预先训练好的，不参与网络更新，即不需要考虑风格分离模型和语音特征提取模型的训练，由此，标签通过 STFT 算法得到第三梅尔特征，通过本网络的输出为第四梅尔特征，通过比对第三梅尔特征和第四梅尔特征得到 loss 值和梯度，本网络的输出的第四梅尔特征与输入通过 STFT 算法得到的第三梅尔特征进行比对得到 loss 值及其梯度，图 12 中黑色虚线为梯度流动方向，当梯度流过相应结构（如编码层、注意结构和解码层）时，模块

20 中的参数才会进行更新，而风格分离模型和语音特征提取模型是预先训练得到，不需要参与本网络的参数更新。loss 值作为判断网络何时停止训练的指标，当 loss 值下降到一定值，且无明显继续下降趋势时，说明网络已收敛，即可停止训练。

本申请实施例中，利用无监督学习的方式提取语音中的音色、韵律和语速等风格特征，

25 实现对语音的风格的可控美化。对于用户希望处理后的语音的风格跟处理前是相同的人声美化场景，比如，方言转换和语音增强的应用场景，本申请能保持处理后的语音的风格不变，同时实现方言的转换或语音增强。本申请利用人工智能技术为实现人声美化提供了一种更加便捷、丰富和覆盖场景更多的方法。实现任意人的去口音、加口音或语音增强的效果，同时可以保持输入和输出的语音风格不发生改变。

30 本申请实施例中，对于各模型的训练及更新的执行主体可以为服务器，服务器将该语音处理神经网络训练好后，由终端下载到本地端。实际应用中，该语音处理神经网络内置于 APP 中，如该 APP 为即时通讯类的 APP 等。

上面对一种语音转换方法进行了说明，请参阅图 13 所示，下面对该方法应用的语音转换装置 1300 进行说明，语音转换装置包括：输入模块 1320，处理模块 1310，获取模块 1330，

35 输出模块 1340。

输入模块 1320，用于接收用户输入的模式选择操作，所述模式选择操作用于选择语音转换的模式；

— 15 —

处理模块 1310，用于根据所述输入模块 1320 接收的所述模式选择操作从多个模式中选择目标转换模式，所述多个模式包括风格转换模式，方言转换模式和语音增强模式；

获取模块 1330，用于获取待转换的第一语音；

处理模块 1310，还用于提取所述获取模块 1330 获取到的所述第一语音的特征信息；

5 将所述第一语音的特征信息输入所述目标转换模式对应的目标语音转换网络，通过所述目标语音转换网络输出转换后的第二语音；

输出模块 1340，用于输出第二语音。

在一个可选的实现方式中，所述处理模块 1310，还用于将所述第一信息输入到语音特征提取模型，通过所述语音特征提取模型提取所述第一语音的音素后验概率 PPG 特征；所述 PPG 特征用于保留所述第一语音的内容信息。

在一个可选的实现方式中，当所述目标转换模式为风格转换模式，所述目标语音转换网络为风格转换网络时，风格转换网络包括风格分离模型和语音融合模型；

所述获取模块 1330，用于获取用于提取风格特征的第三语音；

15 所述处理模块 1310，还用于将所述第三语音输入到风格分离模型，通过所述风格分离模型分离所述第三语音的风格特征；

所述处理模块 1310，还用于将所述风格特征和第一语音的特征信息输入到语音融合模型进行融合，得到所述第二语音。

在一个可选的实现方式中，所述风格特征包括第一特征，所述第一特征包括多个子特征；

20 所述处理模块 1310，还用于将所述第三语音输入到风格分离模型，通过所述风格分离模型提取所述第三语音中的第一特征的向量；将所述第三语音输入到子特征提取模型，通过所述子特征提取模型提取所述子特征的向量；

所述输入模块 1320，还用于接收用户输入的所述多个子特征中每个所述子特征的权重；

25 所述处理模块 1310，还用于根据所述第一特征的向量，所述输入模块 1320 接收的每个所述子特征的向量及每个所述子特征的权重确定所述第三语音的风格特征。

在一个可选的实现方式中，所述处理模块 1310，还用于将所述第一特征的向量输入多头注意力结构，且将每个子特征的向量及与其对应权重的乘积输入到所述多头注意力结构，通过所述多头注意力结构输出所述第三语音的风格特征。

30 在一个可选的实现方式中，所述获取模块 1330，还用于接收用户输入的模板选择操作，所述模板选择操作用于选择目标模板；获取所述目标模板对应的语音，将所述目标模板对应的语音作为所述第三语音。

在一个可选的实现方式中，所述获取模块 1330，还用于接收第二说话人输入的第三语音，所述第一语音为第一说话人的语音，所述第二说话人为与所述第一说话人不同的任意人。

35 在一个可选的实现方式中，所述目标转换模式为方言转换模式，所述目标语音转换网络为方言转换网络；所述处理模块 1310，还用于将所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，所述第一语音为第一方言的语音，所

述第二语音为第二方言的语音。

在一个可选的实现方式中，所述方言转换网络包括多个方言转换模型，每个方言转换模型分别针对不同的待转换方言；

所述输入模块 1320，还用于接收用户输入的选择操作；

5 所述处理模块 1310，还用于将第一语音的特征信息输入到所述选择操作对应的方言转换模型，通过所述选择操作对应的方言转换模型输出所述第二语音。

在一个可选的实现方式中，所述处理模块 1310，还用于将所述第一语音输入到风格分离模型，通过所述风格分离模型分离所述第一语音的风格特征；

10 所述处理模块 1310，还用于将所述第一语音的风格特征和所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，所述第二语音和所述第一语音的风格相同。

在一个可选的实现方式中，所述第一语音为远场语音，当所述目标转换模式为语音增强模式，所述目标语音转换网络为语音增强模型时，

15 所述处理模块 1310，还用于将所述第一语音的特征信息输入到所述模式对应的语音增强模型，通过所述语音增强模型输出第二语音，所述第二语音为近场语音。

在一个可选的实现方式中，所述处理模块 1310，还用于将所述第一语音输入到风格分离模型，通过所述风格分离模型分离所述第一语音的风格特征；

20 所述处理模块 1310，还用于将所述第一语音的风格特征和所述第一语音的特征信息输入到所述语音增强模型，通过所述语音增强模型输出所述第二语音，所述第二语音和所述第一语音的风格相同。

在一个可选的实现方式中，所述获取模块 1330，还用于接收第一说话人输入的所述第一语音；或者，从本地存储文件中选择所述第一语音。

在一种可能的设计中，处理模块 1310 可以是一个处理装置，处理装置的功能可以部分或全部通过软件实现。

25 可选地，处理装置的功能可以部分或全部通过软件实现。此时，处理装置可以包括存储器和处理器，其中，存储器用于存储计算机程序，处理器读取并执行存储器中存储的计算机程序，以执行任意一个方法实施例中的相应处理和/或步骤。

可选地，处理装置可以仅包括处理器。用于存储计算机程序的存储器位于处理装置之外，处理器通过电路/电线与存储器连接，以读取并执行存储器中存储的计算机程序。

30 可选地，所述处理装置可以是一个或多个芯片，或一个或多个集成电路。

示例性的，本申请实施例提供了一种芯片结构，请参阅图 14 所示，该芯片包括：

35 所述芯片可以表现为神经网络处理器 140 (neural-network processing unit, NPU)，NPU 作为协处理器挂载到主 CPU (host CPU) 上，由 Host CPU 分配任务。NPU 的核心部分为运算电路 1403，通过控制器 1404 控制运算电路 1403 提取存储器中的矩阵数据并进行乘法运算。

在一些实现中，运算电路 1403 内部包括多个处理单元 (process engine, PE)。在一些实现中，运算电路 1403 是二维脉动阵列。运算电路 1403 还可以是一维脉动阵列或者能够

- 17 -

执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中，运算电路 1403 是通用的矩阵处理器。

举例来说，假设有输入矩阵 A，权重矩阵 B，输出矩阵 C。运算电路从权重存储器 1402 中取矩阵 B 相应的数据，并缓存在运算电路中每一个 PE 上。运算电路从输入存储器 1401 中取矩阵 A 数据与矩阵 B 进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加器 1408 accumulator 中。

统一存储器 1406 用于存放输入数据以及输出数据。权重数据直接通过存储单元访问控制器 1405 (direct memory access controller, DMAC) 被搬运到权重存储器 1402 中。输入数据也通过 DMAC 被搬运到统一存储器 1406 中。

BIU 为 Bus Interface Unit 即，总线接口单元 1410，用于 AXI 总线与 DMAC 和取指存储器 1409 Instruction Fetch Buffer 的交互。

总线接口单元 1410 (Bus Interface Unit, 简称 BIU)，用于取指存储器 1409 从外部存储器获取指令，还用于存储单元访问控制器 1405 从外部存储器获取输入矩阵 A 或者权重矩阵 B 的原数据。

DMAC 主要用于将外部存储器 DDR 中的输入数据搬运到统一存储器 1406 或将权重数据搬运到权重存储器 1402 中或将输入数据数据搬运到输入存储器 1401 中。

向量计算单元 1407 包括多个运算处理单元，在需要的情况下，对运算电路的输出做进一步处理，如向量乘，向量加，指数运算，对数运算，大小比较等等。主要用于神经网络中非卷积/FC 层网络计算，如池化 (pooling)，批归一化 (batch normalization)，局部响应归一化 (local response normalization) 等。

在一些实现种，向量计算单元 1407 能将经处理的输出的向量存储到统一缓存器 1406。例如，向量计算单元 1407 可以将非线性函数应用到运算电路 1403 的输出，例如累加值的向量，用以生成激活值。在一些实现中，向量计算单元 1407 生成归一化的值、合并值，或二者均有。在一些实现中，处理过的输出的向量能够用作到运算电路 1403 的激活输入，例如用于在神经网络中的后续层中的使用。

控制器 1404 连接的取指存储器 (instruction fetch buffer) 1409，用于存储控制器 1404 使用的指令；

统一存储器 1406，输入存储器 1401，权重存储器 1402 以及取指存储器 1409 均为 On-Chip 存储器。外部存储器私有于该 NPU 硬件架构。

其中，本申请中，语音特征提取模型、风格分离模型、多头注意结构、各子特征提取模型、Seq2seq 神经网络中的编码层、解码层和注意力结构、方言转换模型、语音增强模型中每个模块中各层的运算可以由运算电路 1403 或向量计算单元 1407 执行。

运算电路 1403 或向量计算单元 1407 计算得到参数值 (如第一参数值)，主 CPU 用于读取所述至少一个存储器所存储的计算机程序，使得终端执行上述方法实施例中终端所执行的方法。

请参阅图 15 所示，本发明实施例还提供了另一种语音转换装置，如图 15 所示，仅示

出了与本申请实施例相关的部分，具体技术细节未揭示的，请参照本申请实施例方法部分。语音转换装置可以为终端，该终端可以为手机、平板电脑，笔记本电脑，智能手表等，终端以为手机为例：

图 15 示出的是与本发明实施例提供的终端相关的手机的部分结构的框图。参考图 15，手机包括：射频（radio frequency，RF）电路 1510、存储器 1520、输入单元 1530、显示单元 1540、音频电路 1560、处理器 1580、以及电源 1590 等部件。本领域技术人员可以理解，图 15 中示出的手机结构并不构成对手机的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。

下面结合图 15 对手机的各个构成部件进行具体的介绍：

RF 电路 1510 可用于收发信息或通话过程中，信号的接收和发送，特别地，将基站的下行信息接收后，给处理器 1580 处理。

存储器 1520 可用于存储软件程序以及模块，处理器 1580 通过运行存储在存储器 1520 的软件程序以及模块，从而执行手机的各种功能应用以及数据处理。存储器 1520 可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序（比如声音播放功能等）等。此外，存储器 1520 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。

输入单元 1530 可用于接收输入的数字或字符信息，以及产生与手机的用户设置以及功能控制有关的键信号输入。具体地，输入单元 1530 可包括触控面板 1531 以及其他输入设备 1532。触控面板 1531，也称为触摸屏，可收集用户在其上或附近的触摸操作（比如用户使用手指、触笔等任何适合的物体或附件在触控面板 1531 上或在触控面板 1531 附近的操作），并根据预先设定的程式驱动相应的连接装置。可选的，触控面板 1531 可包括触摸检测装置和触摸控制器两个部分。其中，触摸检测装置检测用户的触摸方位，并检测触摸操作带来的信号，将信号传送给触摸控制器；触摸控制器从触摸检测装置上接收触摸信息，并将它转换成触点坐标，再送给处理器 1580，并能接收处理器 1580 发来的命令并加以执行。除了触控面板 1531，输入单元 1530 还可以包括其他输入设备 1532。具体地，其他输入设备 1532 可以包括但不限于物理键盘、功能键（比如音量控制按键、开关按键等）、轨迹球、鼠标、操作杆等中的一种或多种。

本申请中，输入单元 1530 用于接收用户输入的各种操作，例如，模式选择操作等。图 13 中输入模块 1320 的功能可以由该输入单元 1530 来执行，或者，图 13 中获取模块 1330 的功能可以由该输入单元 1530 来执行。

显示单元 1540 可用于显示由用户输入的信息或提供给用户的信息以及手机的各种菜单。显示单元 1540 可包括显示面板 1541，可选的，可以采用液晶显示器（Liquid Crystal Display，LCD）、有机发光二极管（Organic Light-Emitting Diode，OLED）等形式来配置显示面板 1541。进一步的，触控面板 1531 可覆盖显示面板 1541，当触控面板 1531 检测到在其上或附近的触摸操作后，传送给处理器 1580 以确定触摸事件的类型，随后处理器 1580 根据触摸事件的类型在显示面板 1541 上提供相应的视觉输出。虽然在图 15 中，触控面板

1531 与显示面板 1541 是作为两个独立的部件来实现手机的输入和输入功能，但是在某些实施例中，可以将触控面板 1531 与显示面板 1541 集成而实现手机的输入和输出功能。

本申请中，显示单元 1540 用于显示方法实施例中对应的图 6A、图 6B、图 11A 和图 11B 中示出的 APP 界面。

5 音频电路 1560、扬声器 1561，传声器 1562 可提供用户与手机之间的音频接口。音频电路 1560 可将接收到的音频数据转换后的电信号，传输到扬声器 1561，由扬声器 1561 转换为声音信号输出；另一方面，传声器 1562 将收集的声音信号转换为电信号，由音频电路 1560 接收后转换为音频数据，再将音频数据输出处理器 1580 处理后。

本申请中，音频电路 1560 通过传声器 1562 接收第一说话人的第一语音，或者，接收
10 第二说话人的第三语音。扬声器 1561 用于输出处理之后的第二语音，例如，第二语音为风格转换后的语音，或者，第二语音为方言转换的语音，或者，第二语音为语音增强后的语音。扬声器 1561 用于输出第二语音。

在一个可能的设计中，图 13 中输出模块 1340 的功能可以由扬声器 1561 来执行。

处理器 1580 是手机的控制中心，利用各种接口和线路连接整个手机的各个部分，通过
15 运行或执行存储在存储器 1520 内的软件程序和/或模块，以及调用存储在存储器 1520 内的数据，执行手机的各种功能和处理数据，从而对手机进行整体监控。可选的，处理器 1580 可包括一个或多个处理单元。

手机还包括给各个部件供电的电源 1590（比如电池），优选的，电源可以通过电源管理系统与处理器 1580 逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。
20

尽管未示出，手机还可以包括摄像头、蓝牙模块等，在此不再赘述。

本申请中，当所述存储器存储的程序指令被所述处理器执行时实现上述各方法实施例中终端执行的方法，具体的参阅上述各方法实施例中的描述，此处不赘述。

本申请实施例中还提供一种计算机可读存储介质，该计算机可读存储介质中存储有程序，当其在计算机上运行时，使得计算机执行如前述方法实施例描述的方法中终端设备所执行的步骤。
25

本申请实施例中还提供一种包括计算机程序产品，当其在计算机上行驶时，使得计算机执行如前述各方法实施例描述的方法中终端所执行的步骤。

本申请实施例中还提供一种电路系统，所述电路系统包括处理电路，所述处理电路配置为执行如前述方法实施例描述的方法中终端设备所执行的步骤。
30

在另一种可能的设计中，当该装置为终端内的芯片时，芯片包括：处理单元和通信单元，所述处理单元例如可以是处理器，所述通信单元例如可以是输入/输出接口、管脚或电路等。该处理单元可执行存储单元存储的计算机执行指令，以使该终端内的芯片执行上述第一方面任意一项的无线通信方法。可选地，所述存储单元为所述芯片内的存储单元，如
35 寄存器、缓存等，所述存储单元还可以是所述终端内的位于所述芯片外部的存储单元，如只读存储器（read-only memory, ROM）或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器（random access memory, RAM）等。

—20—

其中，上述任一处提到的处理器，可以是一个通用中央处理器（CPU），微处理器，特定应用集成电路（application-specific integrated circuit, ASIC），或一个或多个用于控制上述第一方面无线通信方法的程序执行的集成电路。

5 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统，装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

15 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

30

-21-

权 利 要 求

1. 一种语音转换方法，其特征在于，包括：

接收用户输入的模式选择操作，所述模式选择操作用于选择语音转换的模式；

根据模式选择操作从多个模式中选择目标转换模式，所述多个模式包括风格转换模式，

5 方言转换模式和语音增强模式；

获取待转换的第一语音；

提取所述第一语音的特征信息；

将所述第一语音的特征信息输入到所述目标转换模式对应的目标语音转换网络，通过
所述目标语音转换网络输出转换后的第二语音；

10 输出所述第二语音。

2. 根据权利要求1所述的方法，其特征在于，所述提取所述第一语音的特征信息，包
括：

将所述第一信息输入到语音特征提取模型，通过所述语音特征提取模型提取所述第一
语音的音素后验概率 PPG 特征；所述 PPG 特征用于保留所述第一语音的内容信息。

15 3. 根据权利要求1或2所述的方法，其特征在于，当所述目标转换模式为风格转换模
式，所述目标语音转换网络为风格转换网络时，所述风格转换网络包括风格分离模型和语
音融合模型，所述方法还包括：

获取用于提取风格特征的第三语音；

20 将所述第三语音输入到风格分离模型，通过所述风格分离模型分离所述第三语音的风
格特征；

所述将所述第一语音的特征信息输入到所述目标转换模式对应的目标语音转换网络，
通过所述目标语音转换网络输出转换后的第二语音，包括：

将所述风格特征和第一语音的特征信息输入到语音融合模型进行融合，得到所述第二
语音。

25 4. 根据权利要求3所述的方法，其特征在于，所述风格特征包括第一特征，所述第一
特征包括多个子特征；

所述将所述第三语音输入到风格分离模型，通过所述风格分离模型分离所述第三语音
的风格特征，包括：

30 将所述第三语音输入到风格分离模型，通过所述风格分离模型提取所述第三语音中的
第一特征的向量；

将所述第三语音输入到子特征提取模型，通过所述子特征提取模型提取所述子特征的
向量；

接收用户输入的所述多个子特征中每个所述子特征的权重；

35 根据所述第一特征的向量，每个所述子特征的向量及每个所述子特征的权重确定所述
第三语音的风格特征。

5. 根据权利要求4所述的方法，其特征在于，所述根据所述第一特征的向量，每个所
述子特征的向量及每个所述子特征的权重确定所述第三语音的风格特征，包括：

将所述第一特征的向量输入多头注意力结构，且将每个子特征的向量及与其对应权重的乘积输入到所述多头注意力结构，通过所述多头注意力结构输出所述第三语音的风格特征。

6. 根据权利要求 3 所述的方法，其特征在于，所述获取用于提取风格特征的第三语音，
5 包括：

接收用户输入的模板选择操作，所述模板选择操作用于选择目标模板；

获取所述目标模板对应的语音，将所述目标模板对应的语音作为所述第三语音。

7. 根据权利要求 3 所述的方法，其特征在于，所述获取用于提取风格特征的第三语音，
包括：

10 接收第二说话人输入的第三语音，所述第一语音为第一说话人的语音，所述第二说话人为与所述第一说话人不同的任意人。

8. 根据权利要求 1 或 2 所述的方法，其特征在于，当所述目标转换模式为方言转换模式，所述目标语音转换网络为方言转换网络时，所述将所述第一语音的特征信息输入到所述目标转换模式对应的目标语音转换网络，通过所述目标语音转换网络输出转换后的第二
15 语音，包括：

将所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，所述第一语音为第一方言的语音，所述第二语音为第二方言的语音。

9. 根据权利要求 8 所述的方法，其特征在于，所述方言转换网络包括多个方言转换模型，每个方言转换模型分别针对不同的待转换方言，所述方法还包括：

20 接收用户输入的选择操作；

将第一语音的特征信息输入到所述选择操作对应的方言转换模型，通过所述选择操作对应的方言转换模型输出所述第二语音。

10. 根据权利要求 8 所述的方法，其特征在于，所述方法还包括：

25 将所述第一语音输入到风格分离模型，通过所述风格分离模型分离所述第一语音的风格特征；

所述将所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，包括：

将所述第一语音的风格特征和所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，所述第二语音和所述第一语音的风格相同。

30 11. 根据权利要求 1 或 2 所述的方法，其特征在于，所述第一语音为远场语音，当所述目标转换模式为语音增强模式，所述目标语音转换网络为语音增强模型时，所述将所述第一语音的特征信息输入到所述目标转换模式对应的目标语音转换网络，通过所述目标语音转换网络输出转换后的第二语音，包括：

35 将所述第一语音的特征信息输入到所述模式对应的语音增强模型，通过所述语音增强模型输出第二语音，所述第二语音为近场语音。

12. 根据权利要求 11 所述的方法，其特征在于，所述方法还包括：

将所述第一语音输入到风格分离模型，通过所述风格分离模型分离所述第一语音的风

格特征;

所述将所述第一语音的特征信息输入到所述模式对应的语音增强模型, 通过所述语音增强模型输出第二语音, 包括:

5 将所述第一语音的风格特征和所述第一语音的特征信息输入到所述语音增强模型, 通过所述语音增强模型输出所述第二语音, 所述第二语音和所述第一语音的风格相同。

13. 根据权利要求 1-12 中任一项所述的方法, 其特征在于, 所述获取待转换的第一语音, 包括:

接收第一说话人输入的所述第一语音;

或者,

10 从本地存储文件中选择所述第一语音。

14. 一种语音转换装置, 其特征在于, 包括:

输入模块, 用于接收用户输入的模式选择操作, 所述模式选择操作用于选择语音转换的模式;

15 处理模块, 用于根据所述输入模块接收的所述模式选择操作从多个模式中选择目标转换模式, 所述多个模式包括风格转换模式, 方言转换模式和语音增强模式;

获取模块, 用于获取待转换的第一语音;

所述处理模块, 还用于提取所述获取模块获取到的所述第一语音的特征信息; 将所述第一语音的特征信息输入所述目标转换模式对应的目标语音转换网络, 通过所述目标语音转换网络输出转换后的第二语音;

20 输出模块, 用于输出所述第二语音。

15. 根据权利要求 14 所述的装置, 其特征在于,

所述处理模块, 还用于将所述第一信息输入到语音特征提取模型, 通过所述语音特征提取模型提取所述第一语音的音素后验概率 PPG 特征; 所述 PPG 特征用于保留所述第一语音的内容信息。

25 16. 根据权利要求 14 或 15 所述的装置, 其特征在于, 当所述目标转换模式为风格转换模式, 所述目标语音转换网络为风格转换网络时, 风格转换网络包括风格分离模型和语音融合模型;

所述获取模块, 用于获取用于提取风格特征的第三语音;

30 所述处理模块, 还用于将所述第三语音输入到风格分离模型, 通过所述风格分离模型分离所述第三语音的风格特征;

所述处理模块, 还用于将所述风格特征和第一语音的特征信息输入到语音融合模型进行融合, 得到所述第二语音。

17. 根据权利要求 16 所述的装置, 其特征在于, 所述风格特征包括第一特征, 所述第一特征包括多个子特征;

35 所述处理模块, 还用于将所述第三语音输入到风格分离模型, 通过所述风格分离模型提取所述第三语音中的第一特征的向量; 将所述第三语音输入到子特征提取模型, 通过所述子特征提取模型提取所述子特征的向量;

—24—

所述输入模块，还用于接收用户输入的所述多个子特征中每个所述子特征的权重；

所述处理模块，还用于根据所述第一特征的向量，所述输入模块接收的每个所述子特征的向量及每个所述子特征的权重确定所述第三语音的风格特征。

18. 根据权利要求 16 所述的装置，其特征在于，

5 所述获取模块，还用于接收用户输入的模板选择操作，所述模板选择操作用于选择目标模板；获取所述目标模板对应的语音，将所述目标模板对应的语音作为所述第三语音。

19. 根据权利要求 14 或 15 所述的装置，其特征在于，所述目标转换模式为方言转换模式，所述目标语音转换网络为方言转换网络；

10 所述处理模块，还用于将所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，所述第一语音为第一方言的语音，所述第二语音为第二方言的语音。

20. 根据权利要求 19 所述的装置，其特征在于，所述处理模块，还用于将所述第一语音输入到风格分离模型，通过所述风格分离模型分离所述第一语音的风格特征；

15 所述处理模块，还用于将所述第一语音的风格特征和所述第一语音的特征信息输入到方言转换网络，通过所述方言转换网络输出所述第二语音，所述第二语音和所述第一语音的风格相同。

21. 根据权利要求 14 或 15 所述的装置，其特征在于，所述第一语音为远场语音，当所述目标转换模式为语音增强模式，所述目标语音转换网络为语音增强模型时，

20 所述处理模块，还用于将所述第一语音的特征信息输入到所述模式对应的语音增强模型，通过所述语音增强模型输出第二语音，所述第二语音为近场语音。

22. 一种终端设备，其特征在于，包括处理器，所述处理器与存储器耦合，所述存储器存储有程序指令，当所述存储器存储的程序指令被所述处理器执行时实现权利要求 1 至 13 中任一项所述的方法。

25 23. 一种计算机可读存储介质，其特征在于，包括程序，当其在计算机上运行时，使得计算机执行如权利要求 1 至 13 中任一项所述的方法。

— 1/14 —

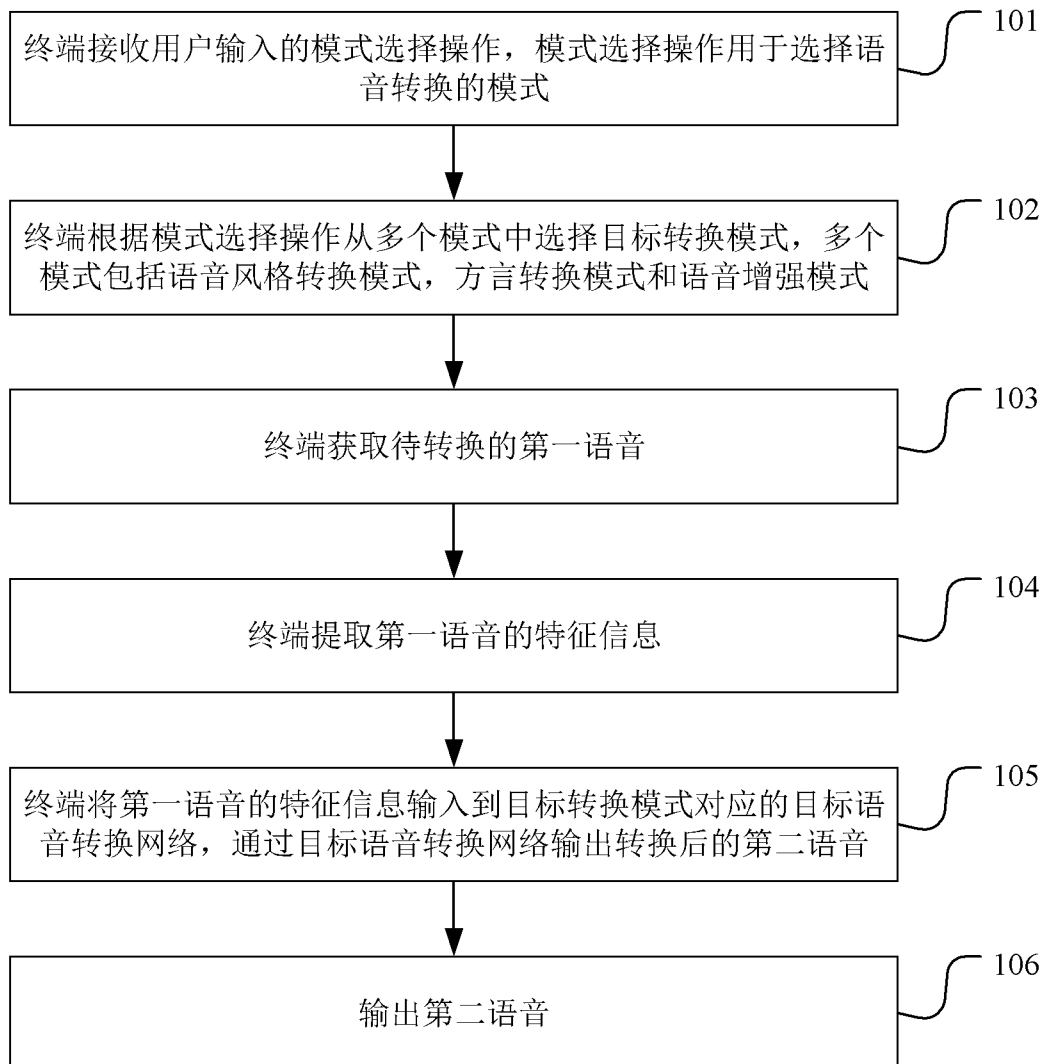


图 1

—2/14—

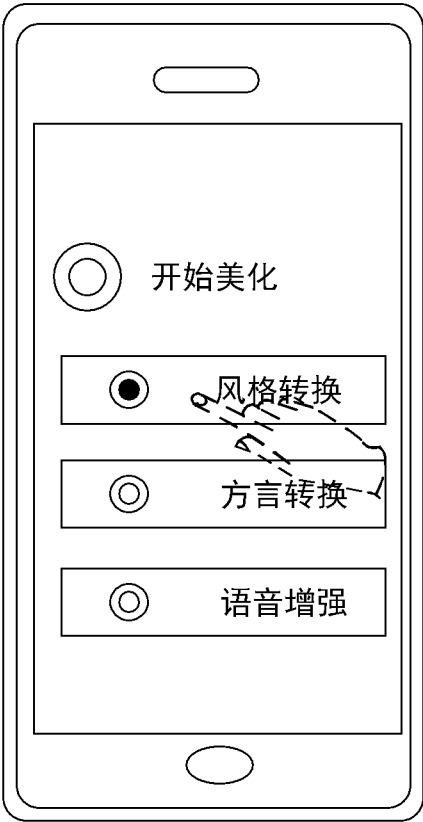


图 2

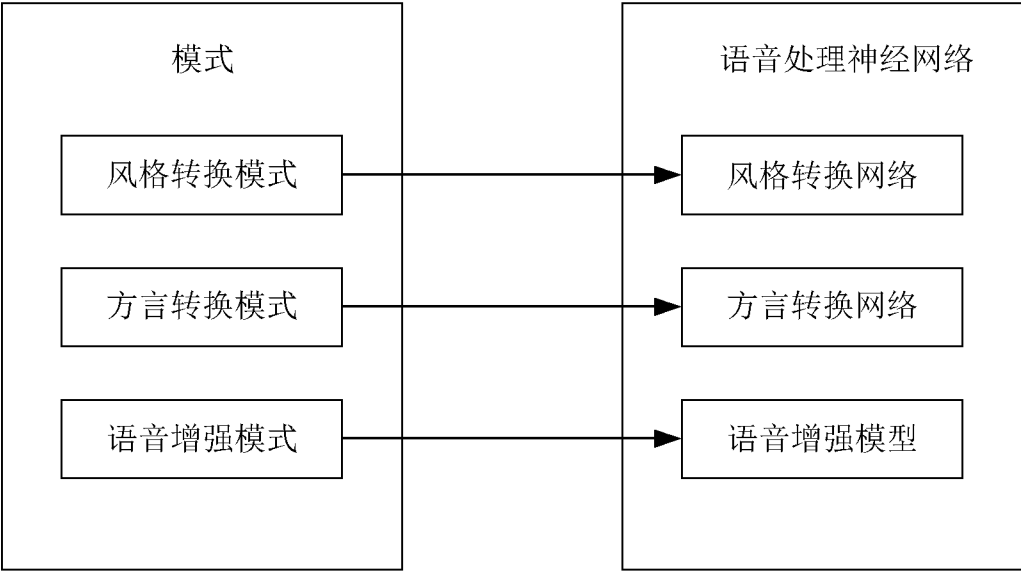


图 3

—3/14—

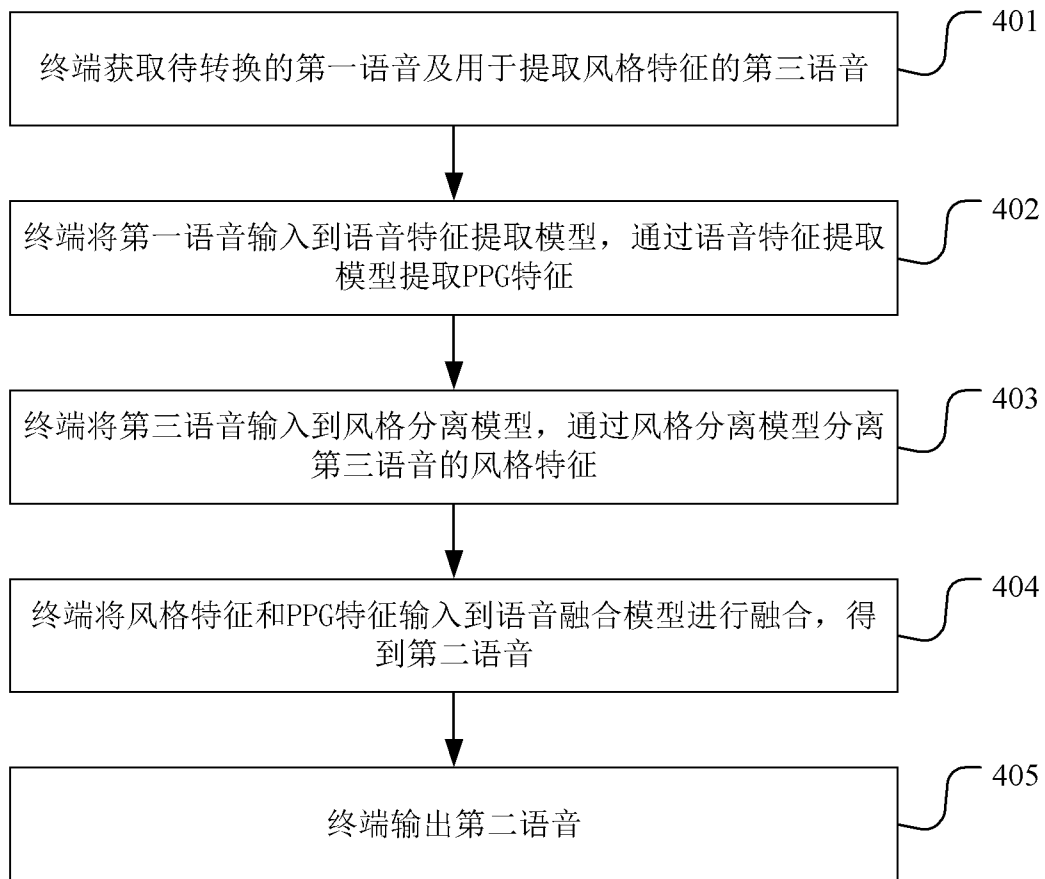


图 4

—4/14—

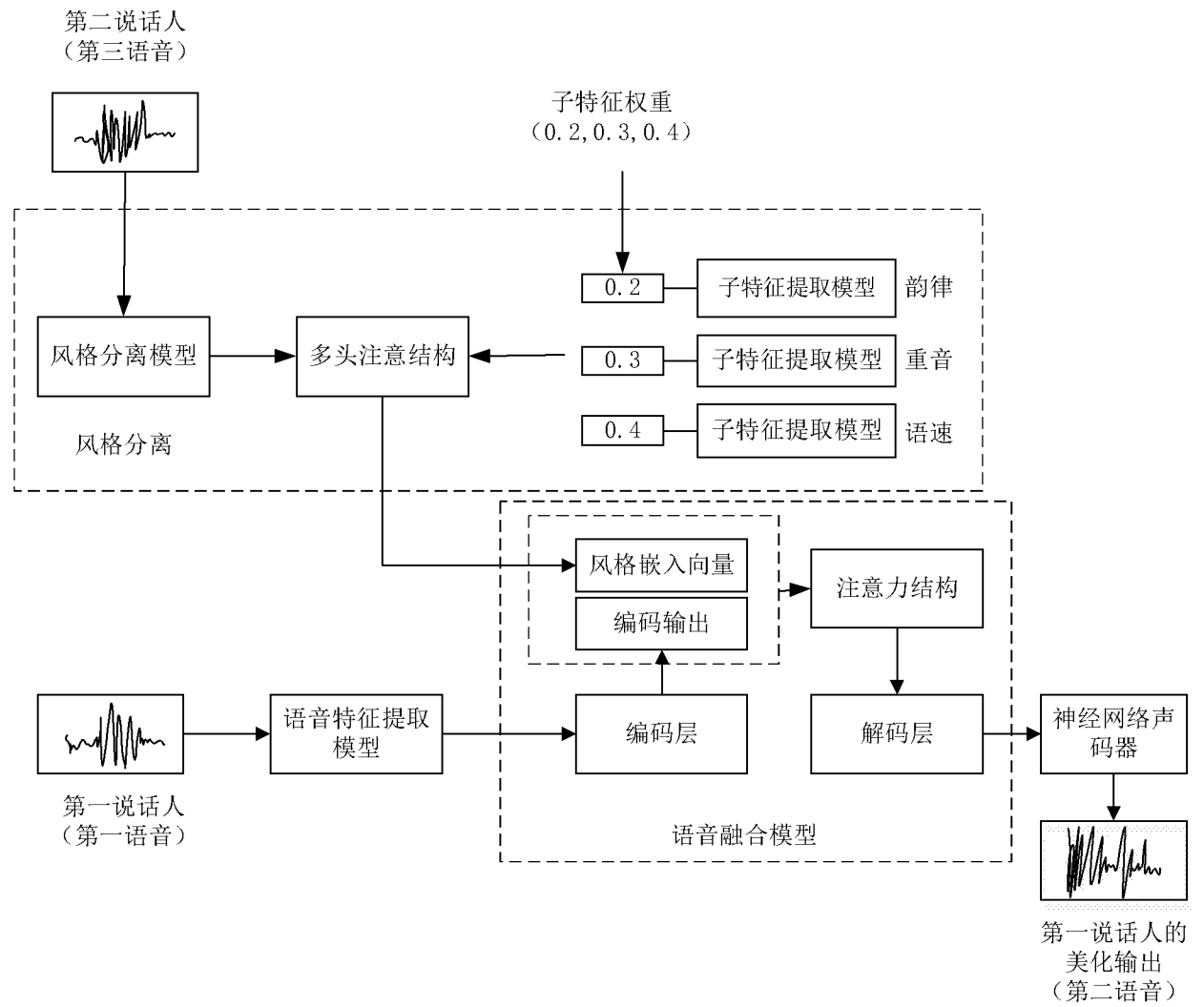


图 5

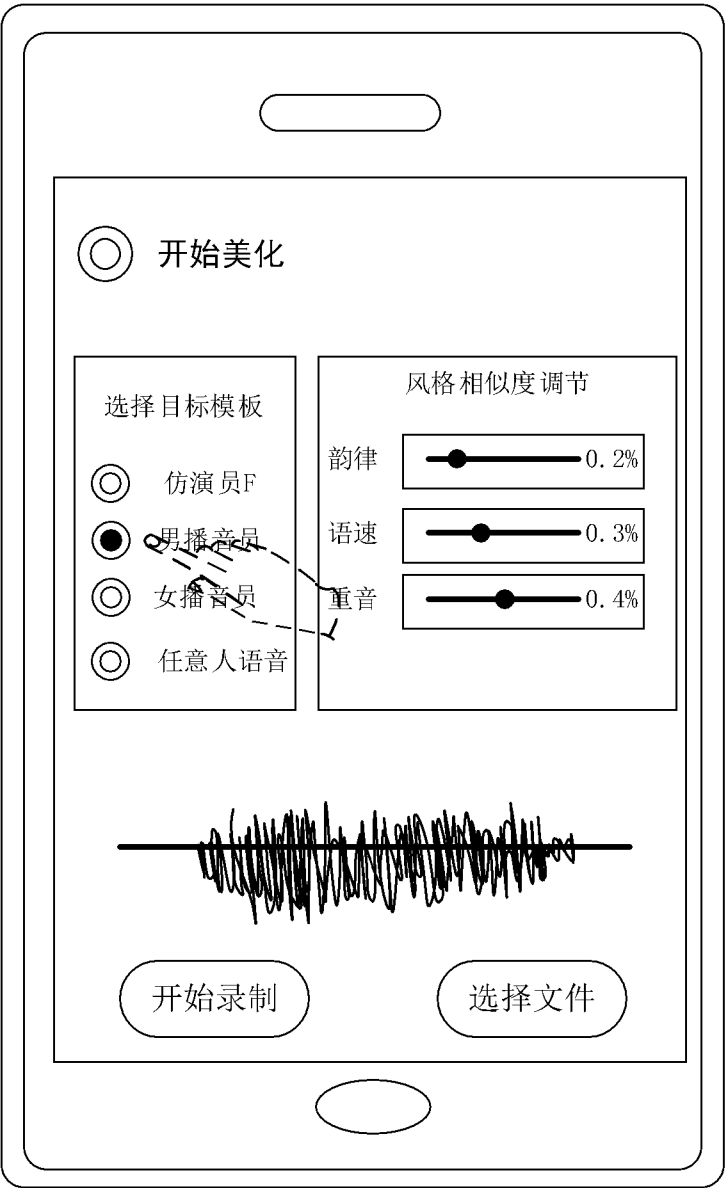


图 6A

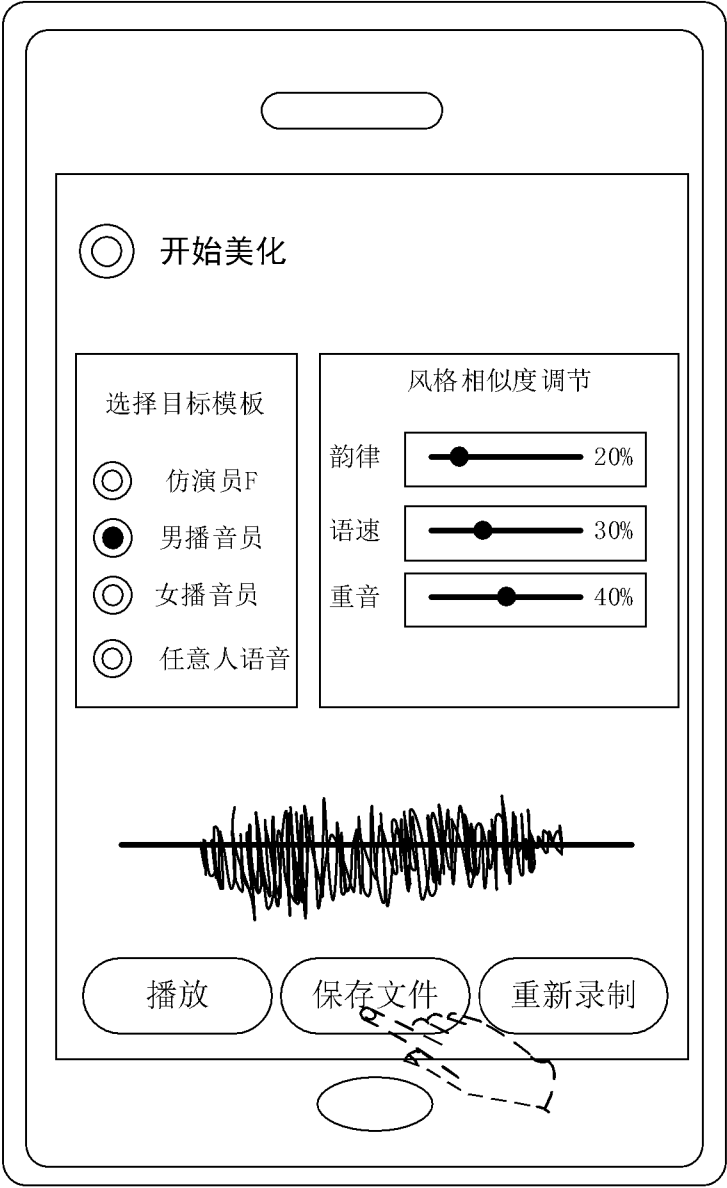


图 6B

—7/14—

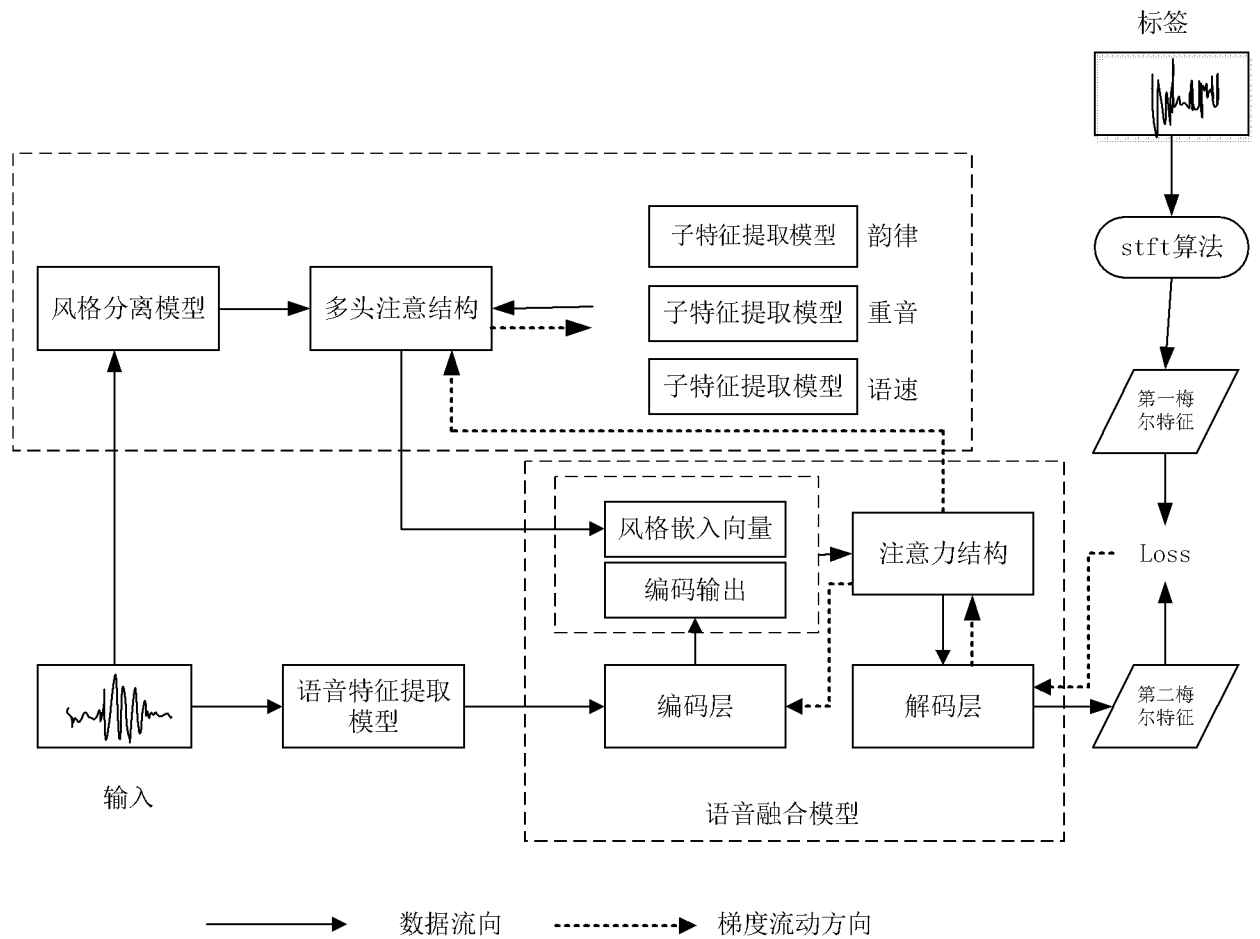


图 7

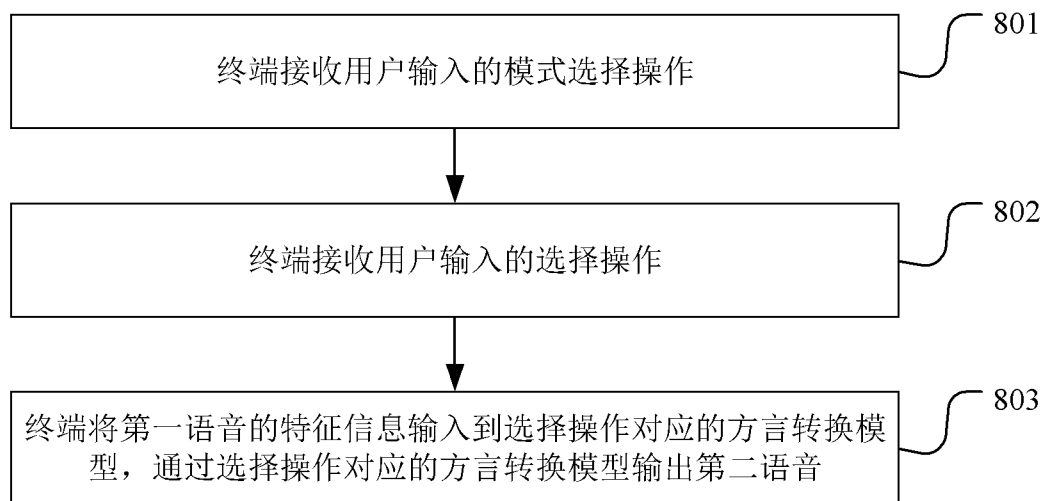


图 8

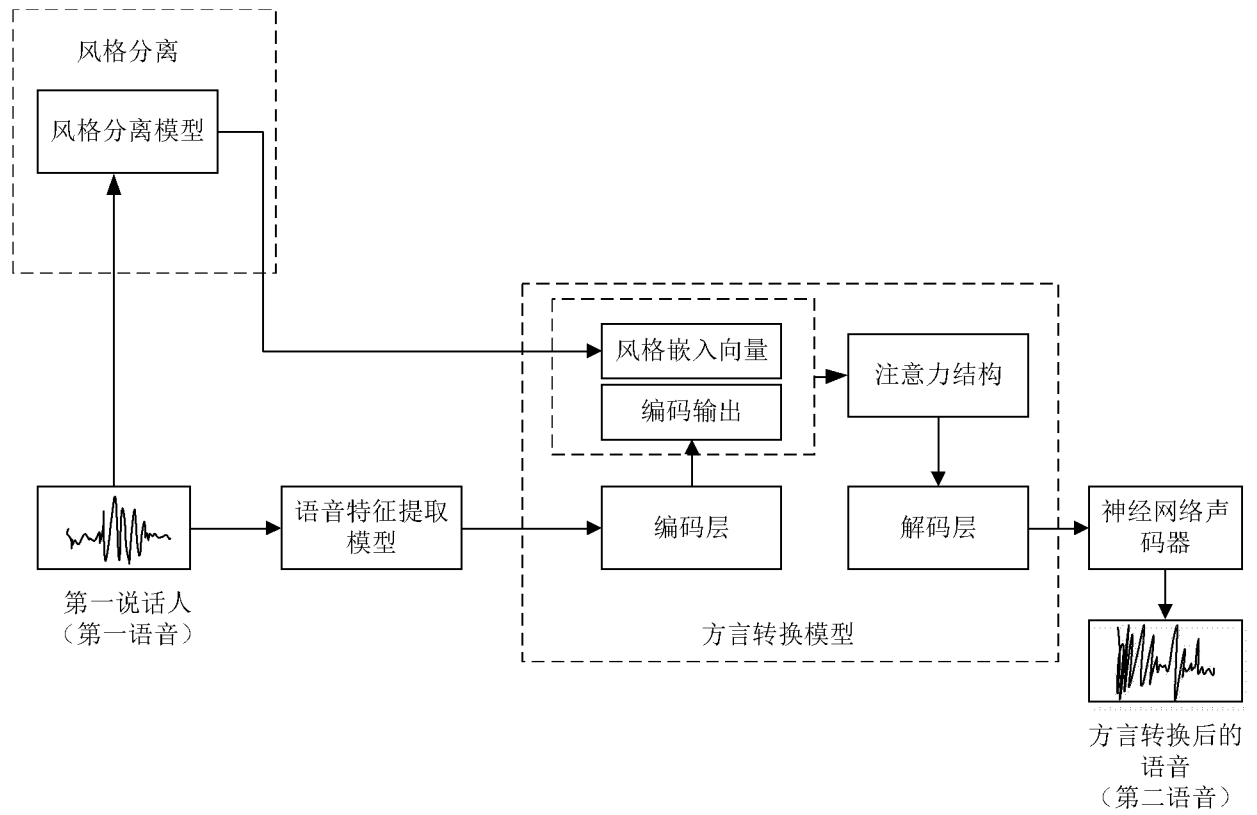


图 9

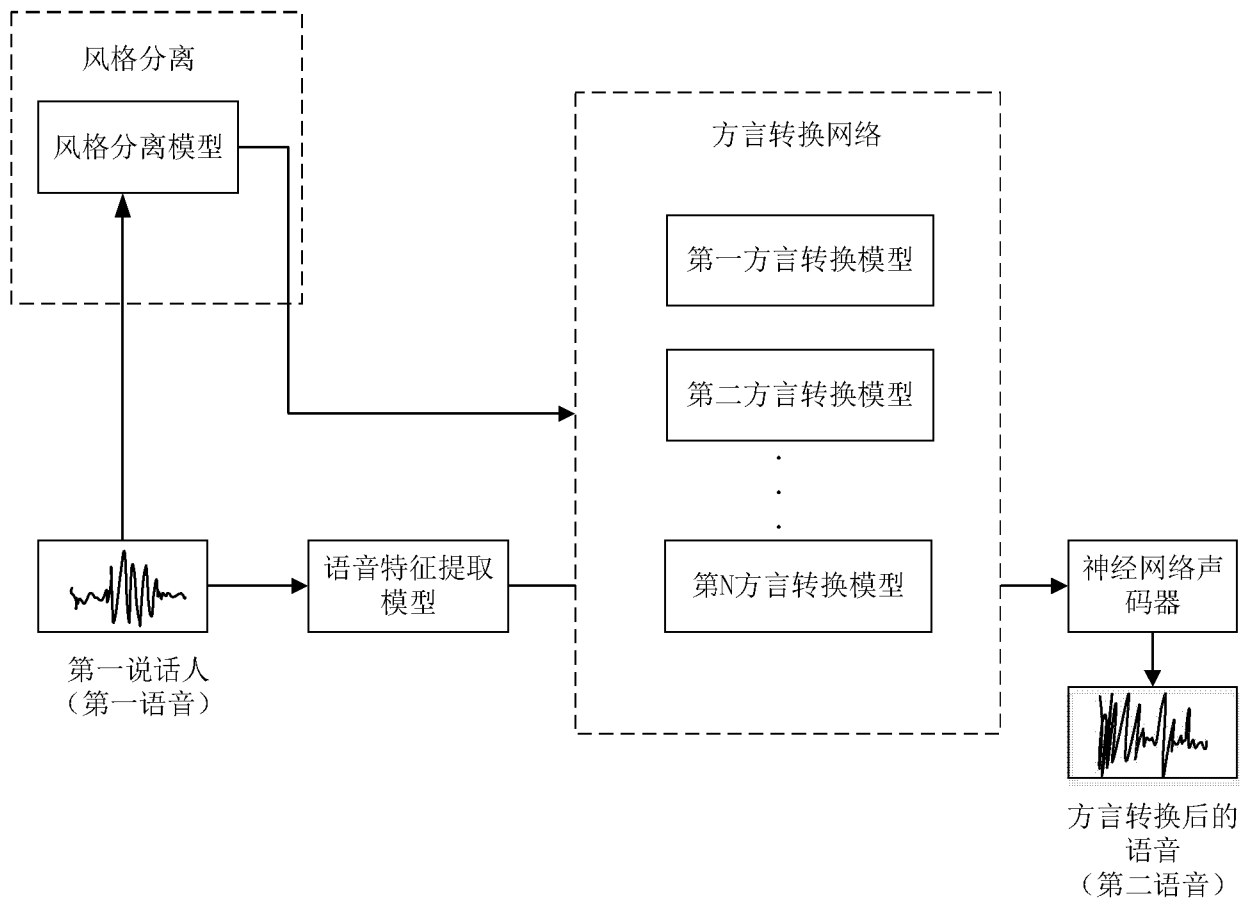


图 10

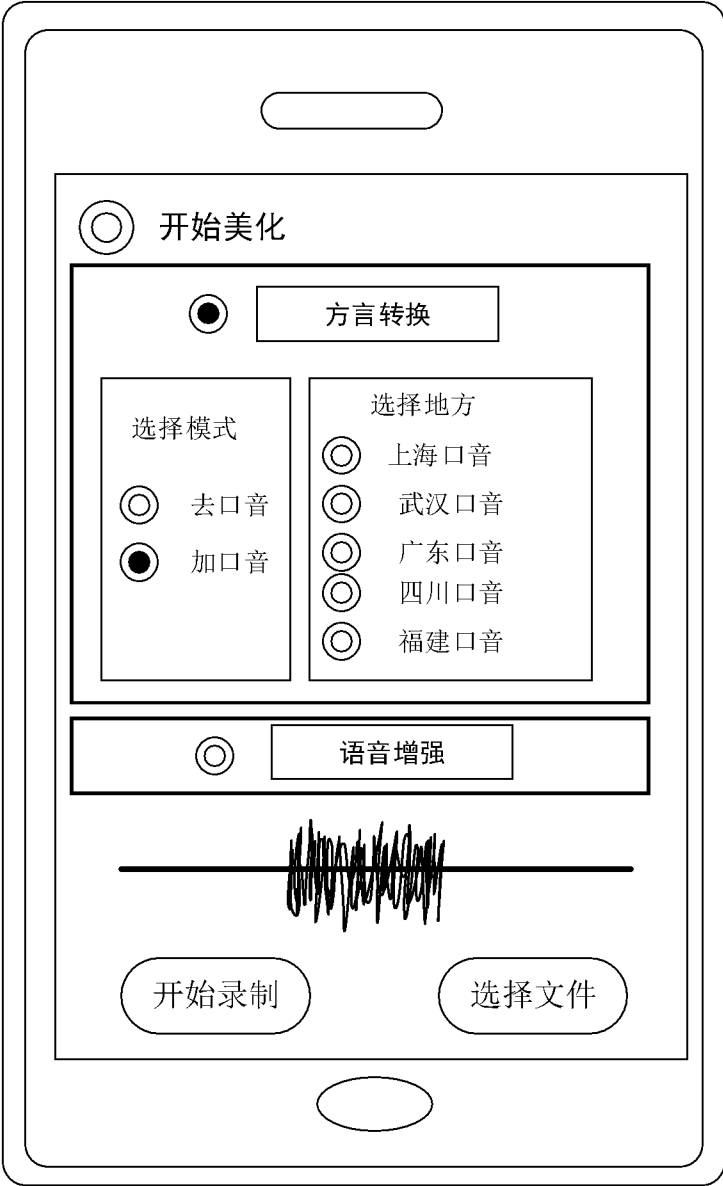


图 11A

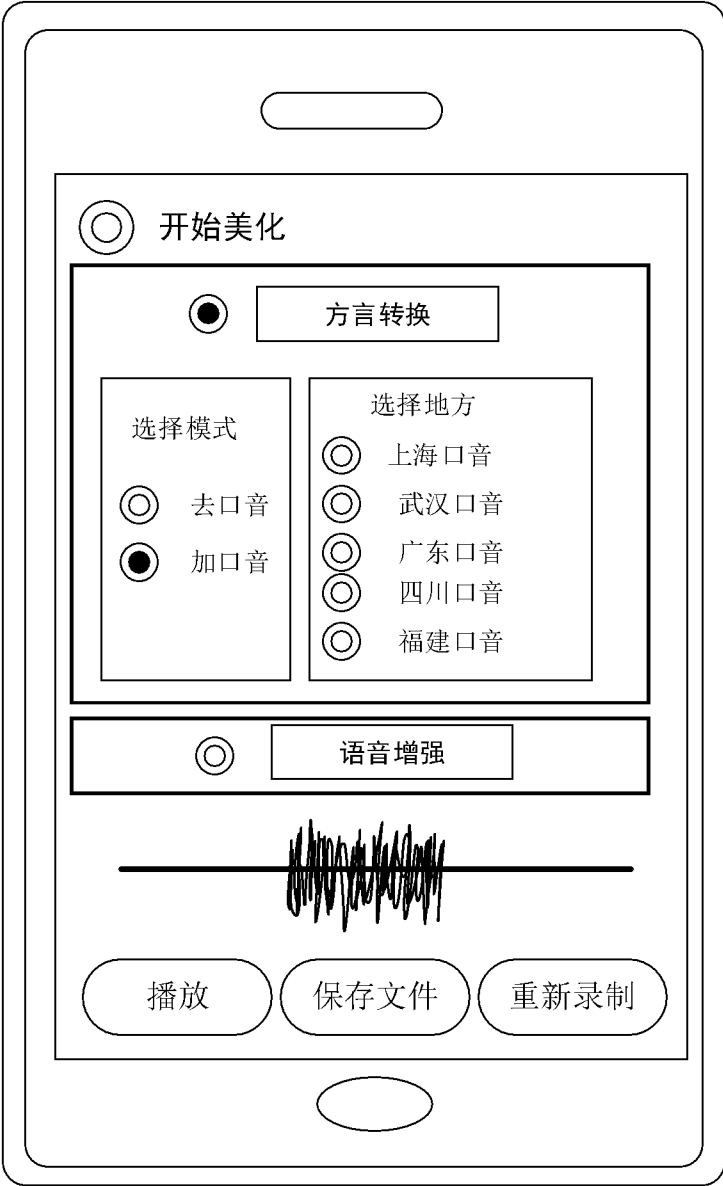


图 11B

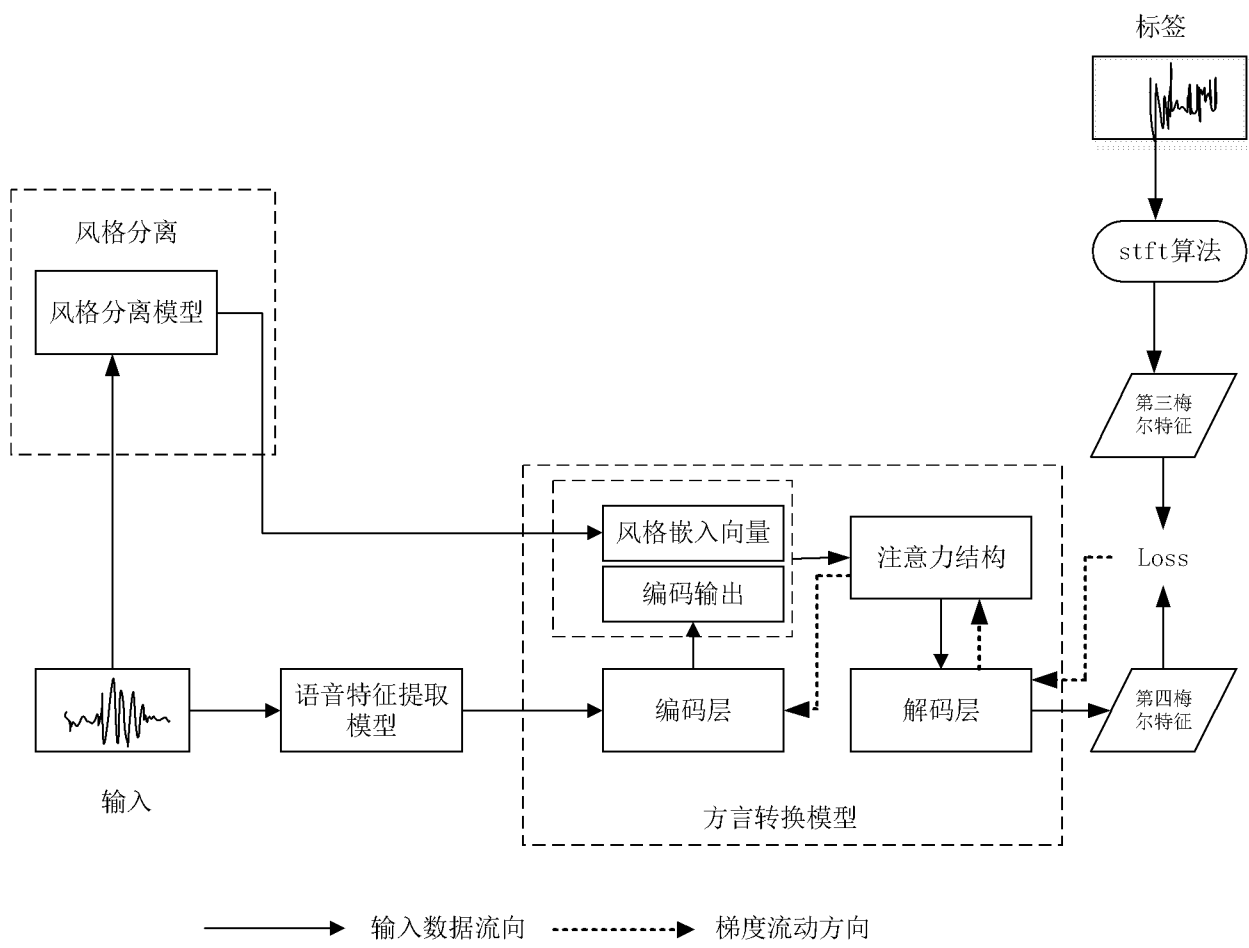


图 12

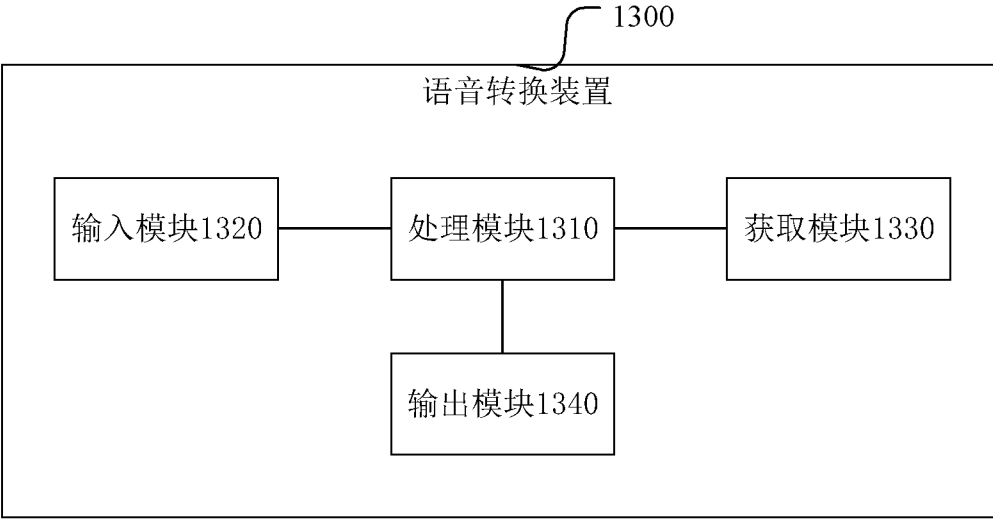


图 13

— 13/14 —

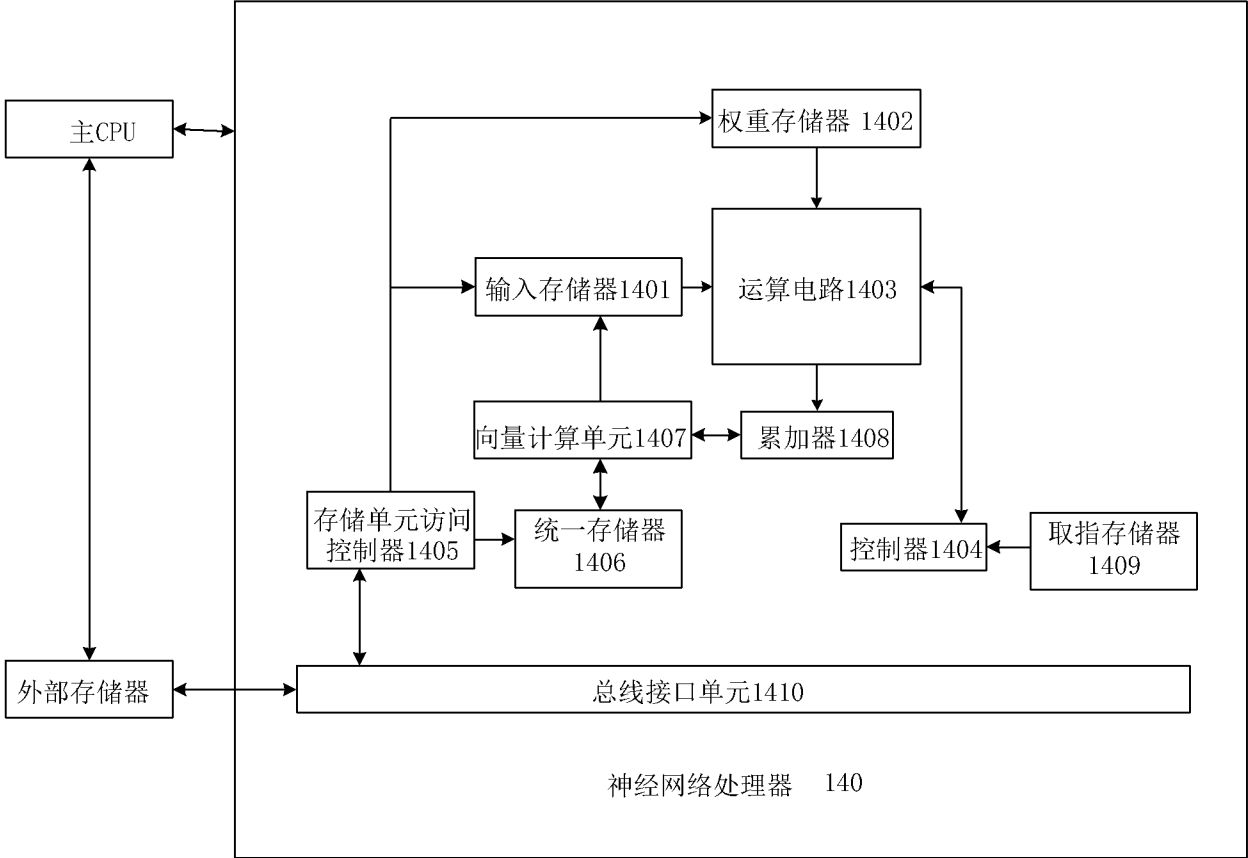


图 14

— 14/14 —

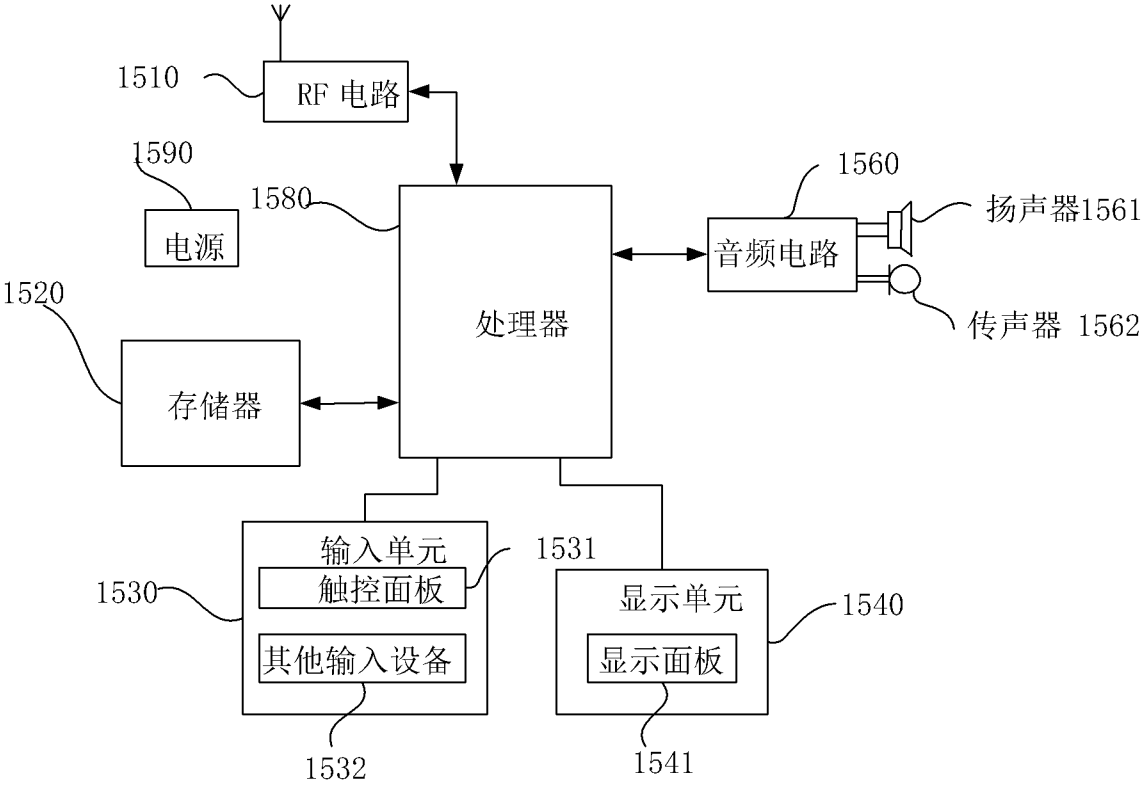


图 15

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/117945

A. CLASSIFICATION OF SUBJECT MATTER

G10L 13/02(2013.01)i; G10L 13/033(2013.01)i; G10L 13/04(2013.01)i; G10L 15/02(2006.01)i; G10L 15/06(2013.01)i; G10L 15/16(2006.01)i; G10L 15/26(2006.01)i; G10L 25/18(2013.01)i; G10L 25/30(2013.01)i; H04M 1/72433(2021.01)i; H04M 1/72403(2021.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L; H04M

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; VEN; EPTXT; USTXT; WOTXT; CNKI; IEEE: 模式, 转换, 方言, 风格, 选择, 语音, 增强, 特征, 音素, 后验, 概率, PPG, mode, pattern, transform, dial+, style, select, choose, sound, voice, speech, enhancement, character, phonetic, phoneme, posteriori, probability

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105551480 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 04 May 2016 (2016-05-04) description, paragraphs [0005]-[0109], and figures 1-8	1, 8-14, 19-23
Y	CN 105551480 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 04 May 2016 (2016-05-04) description, paragraphs [0005]-[0109], and figures 1-8	2-13, 15-23
Y	CN 111627457 A (GUANGZHOU GUOYIN INTELLIGENT TECHNOLOGY CO., LTD.) 04 September 2020 (2020-09-04) description, paragraphs [0002]-[0106], and figure 2	2-13, 15-23
A	CN 106356065 A (NUBIA TECHNOLOGY CO., LTD.) 25 January 2017 (2017-01-25) entire document	1-23
A	CN 111312267 A (GUANGZHOU CITY PAGODA INFORMATION TECHNOLOGY CO., LTD.) 19 June 2020 (2020-06-19) entire document	1-23

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance
 “E” earlier application or patent but published on or after the international filing date
 “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
 “O” document referring to an oral disclosure, use, exhibition or other means
 “P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
 “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
 “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
 “&” document member of the same patent family

Date of the actual completion of the international search

01 November 2021

Date of mailing of the international search report

18 November 2021

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China**

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/117945**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110895932 A (INSTITUTE OF ACOUSTICS, CHINESE ACADEMY OF SCIENCES et al.) 20 March 2020 (2020-03-20) entire document	1-23
A	JP 2013171196 A (TOSHIBA K. K.) 02 September 2013 (2013-09-02) entire document	1-23

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/117945

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	105551480	A	04 May 2016	CN	105551480	B	15 October 2019
CN	111627457	A	04 September 2020	None			
CN	106356065	A	25 January 2017	None			
CN	111312267	A	19 June 2020	None			
CN	110895932	A	20 March 2020	None			
JP	2013171196	A	02 September 2013	US	2013218568	A1	22 August 2013
				US	9135910	B2	15 September 2015
				JP	5665780	B2	04 February 2015

A. 主题的分类 G10L 13/02(2013.01)i; G10L 13/033(2013.01)i; G10L 13/04(2013.01)i; G10L 15/02(2006.01)i; G10L 15/06(2013.01)i; G10L 15/16(2006.01)i; G10L 15/26(2006.01)i; G10L 25/18(2013.01)i; G10L 25/30(2013.01)i; H04M 1/72433(2021.01)i; H04M 1/72403(2021.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																										
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G10L; H04M 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) CNABS; CNTXT; VEN; EPTXT; USTXT; WOTXT; CNKI; IEEE: 模式, 转换, 方言, 风格, 选择, 语音, 增强, 特征, 音素, 后 验, 概率, PPG, mode, pattern, transform, dial+, style, select, choose, sound, voice, speech, enhancement, character, phonetic, phoneme, posteriori, probability																										
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 105551480 A (百度在线网络技术北京有限公司) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0005]-[0109]段, 附图1-8</td> <td>1、8-14、19-23</td> </tr> <tr> <td>Y</td> <td>CN 105551480 A (百度在线网络技术北京有限公司) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0005]-[0109]段, 附图1-8</td> <td>2-13、15-23</td> </tr> <tr> <td>Y</td> <td>CN 111627457 A (广州国音智能科技有限公司) 2020年 9月 4日 (2020 - 09 - 04) 说明书第[0002]-[0106]段, 附图2</td> <td>2-13、15-23</td> </tr> <tr> <td>A</td> <td>CN 106356065 A (努比亚技术有限公司) 2017年 1月 25日 (2017 - 01 - 25) 全文</td> <td>1-23</td> </tr> <tr> <td>A</td> <td>CN 111312267 A (广州市百果园信息技术有限公司) 2020年 6月 19日 (2020 - 06 - 19) 全文</td> <td>1-23</td> </tr> <tr> <td>A</td> <td>CN 110895932 A (中国科学院声学研究所 等) 2020年 3月 20日 (2020 - 03 - 20) 全文</td> <td>1-23</td> </tr> <tr> <td>A</td> <td>JP 2013171196 A (TOSHIBA KK) 2013年 9月 2日 (2013 - 09 - 02) 全文</td> <td>1-23</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 105551480 A (百度在线网络技术北京有限公司) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0005]-[0109]段, 附图1-8	1、8-14、19-23	Y	CN 105551480 A (百度在线网络技术北京有限公司) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0005]-[0109]段, 附图1-8	2-13、15-23	Y	CN 111627457 A (广州国音智能科技有限公司) 2020年 9月 4日 (2020 - 09 - 04) 说明书第[0002]-[0106]段, 附图2	2-13、15-23	A	CN 106356065 A (努比亚技术有限公司) 2017年 1月 25日 (2017 - 01 - 25) 全文	1-23	A	CN 111312267 A (广州市百果园信息技术有限公司) 2020年 6月 19日 (2020 - 06 - 19) 全文	1-23	A	CN 110895932 A (中国科学院声学研究所 等) 2020年 3月 20日 (2020 - 03 - 20) 全文	1-23	A	JP 2013171196 A (TOSHIBA KK) 2013年 9月 2日 (2013 - 09 - 02) 全文	1-23
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																								
X	CN 105551480 A (百度在线网络技术北京有限公司) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0005]-[0109]段, 附图1-8	1、8-14、19-23																								
Y	CN 105551480 A (百度在线网络技术北京有限公司) 2016年 5月 4日 (2016 - 05 - 04) 说明书第[0005]-[0109]段, 附图1-8	2-13、15-23																								
Y	CN 111627457 A (广州国音智能科技有限公司) 2020年 9月 4日 (2020 - 09 - 04) 说明书第[0002]-[0106]段, 附图2	2-13、15-23																								
A	CN 106356065 A (努比亚技术有限公司) 2017年 1月 25日 (2017 - 01 - 25) 全文	1-23																								
A	CN 111312267 A (广州市百果园信息技术有限公司) 2020年 6月 19日 (2020 - 06 - 19) 全文	1-23																								
A	CN 110895932 A (中国科学院声学研究所 等) 2020年 3月 20日 (2020 - 03 - 20) 全文	1-23																								
A	JP 2013171196 A (TOSHIBA KK) 2013年 9月 2日 (2013 - 09 - 02) 全文	1-23																								
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																										
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的 公布日而引用的或者因其他特殊理由而引用的文件(如具体 说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解 发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是 新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并 且这种结合对于本领域技术人员为显而易见时, 要求保护的发 明不具有创造性 “&” 同族专利的文件																										
国际检索实际完成的日期 2021年 11月 1日		国际检索报告邮寄日期 2021年 11月 18日																								
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 郭忆 电话号码 (86-512) 88997073																								

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/117945

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	105551480	A	2016年 5月 4日	CN 105551480 B	2019年 10月 15日
CN	111627457	A	2020年 9月 4日	无	
CN	106356065	A	2017年 1月 25日	无	
CN	111312267	A	2020年 6月 19日	无	
CN	110895932	A	2020年 3月 20日	无	
JP	2013171196	A	2013年 9月 2日	US 2013218568 A1	2013年 8月 22日
				US 9135910 B2	2015年 9月 15日
				JP 5665780 B2	2015年 2月 4日