DESCRIPTION

TITLE OF THE INVENTION: LANGUAGE PROCESSING MODEL TRAINING METHOD, APPARATUS, ELECTRONIC DEVICE, AND READABLE STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to Chinese Patent Application No. 2025101964074, filed with the Chinese Patent Office on February 21, 2025, and titled "LANGUAGE PROCESSING MODEL TRAINING METHOD, APPARATUS, ELECTRONIC DEVICE, AND READABLE STORAGE MEDIUM", the entire contents of which are incorporated herein by reference.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of speech synthesis technology, and in particular, to a language processing model training method, an apparatus, an electronic device, and a computer-readable storage medium.

BACKGROUND ART

[0003] With the rapid development of information technology, speech synthesis technology (Text-To-Speech, TTS), as a technology for converting text into natural and smooth speech, has become an important component of human-computer interaction and directly affects the user experience. TTS is widely used in fields such as outbound broadcasting, voice navigation, and audiobooks.

[0004] In recent years, the training of neural network-based large speech synthesis models has become a research hotspot in the TTS field. However, prior large speech synthesis models perform inadequately in handling style information (such as prosody, and emotion).

SUMMARY

[0005] The present disclosure aims to provide a language processing model training method, an apparatus, an electronic device, and a readable storage medium to address the issue that prior large speech synthesis models perform inadequately in handling style information.

[0006] In a first aspect, the present disclosure provides a language processing model training method, comprising:

[0007] acquiring training text data and ground-truth speech token data;

[0008] inputting the training text data and the ground-truth speech token data into a prediction module of a language processing model to obtain predicted speech token data output by the prediction module;

[0009] acquiring ground-truth style features based on the ground-truth speech token data and acquiring predicted style features based on the predicted speech token data;

[0010] acquiring a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and

[0011] performing optimization training on the prediction module according to the total loss function.

[0012] In an optional embodiment, the language processing model comprises a style encoding module, and the step of acquiring ground-truth style features based on the ground-truth speech token data and acquiring predicted style features based on the predicted speech token data comprises:

[0013] performing style extraction on the ground-truth speech token data through the style encoding module to obtain the ground-truth style features; and

[0014] performing style extraction on the predicted speech token data through the style encoding module to obtain the predicted style features.

[0015] In an optional embodiment, the step of acquiring a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features comprises:

[0016] acquiring a prediction loss function based on the ground-truth speech token data and the predicted speech token data;

[0017] acquiring a style loss function based on the ground-truth style features and the predicted style features; and

[0018] acquiring the total loss function based on the prediction loss function and the style loss function.

[0019] In an optional embodiment, the step of acquiring a prediction loss function based on the ground-truth speech token data and the predicted speech token data comprises:

[0020] calculating a distribution distance between the ground-truth speech token data and the predicted speech token data to obtain the prediction loss function.

[0021] In an optional embodiment, the step of acquiring a style loss function based on the ground-truth style features and the predicted style features comprises:

[0022] calculating a similarity between the ground-truth style features and the predicted style features to obtain the style loss function.

[0023] In an optional embodiment, the step of acquiring the total loss function based on the prediction loss function and the style loss function comprises:

[0024] performing a weighted calculation on the prediction loss function and the style loss function to obtain the total loss function.

[0025] In a second aspect, the present disclosure provides a language processing model training apparatus, comprising:

[0026] a predicted speech token data acquisition module, configured to acquire training text data and ground-truth speech token data; input the training text data and the ground-truth speech token data into a prediction module of a language processing model to obtain predicted speech token data output by the prediction module;

[0027] a style feature acquisition module, configured to acquire ground-truth style features based on the ground-truth speech token data and acquire predicted style features based on the predicted speech token data;

[0028] a loss function acquisition module, configured to acquire a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and

[0029] an optimization training module, configured to perform optimization training on the prediction module according to the total loss function.

[0030] In an optional embodiment, the language processing model comprises a style encoding module, and the style feature acquisition module is further configured to perform style extraction on the ground-truth speech token data through the style encoding module to obtain the ground-truth style features; and to perform style extraction on the predicted speech token data through the style encoding module to obtain the predicted style features.

[0031] In a third aspect, the present disclosure provides an electronic device, comprising a processor and a memory, wherein the memory stores a computer program executable by the processor to implement the language processing model training method described in any one of the foregoing embodiments.

[0032] In a fourth aspect, the present disclosure provides a computer-readable storage medium having a computer program stored thereon, wherein the computer program, when

executed by a processor, implements the language processing model training method described in any one of the foregoing embodiments.

[0033] The embodiments of the present disclosure provide a language processing model training method, an apparatus, an electronic device, and a readable storage medium. The method comprises: acquiring training text data and ground-truth speech token data and inputting them into a prediction module of a language processing model to obtain predicted speech token data; extracting style features from both the ground-truth and predicted speech token data respectively; constructing a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and then performing optimization training on the prediction module according to the total loss function. By incorporating the style feature loss into the total loss function, the language processing model is enabled to learn style information, thereby allowing the language processing model to more accurately reproduce speech styles. This improves the performance of the language processing model in handling speech style information. Since the speech processing model is a key component of a speech synthesis model, the performance of a large speech synthesis model in handling speech style information can thus be enhanced.

BRIEF DESCRIPTION OF DRAWINGS

[0034] In order to more clearly illustrate the technical solutions of the embodiments of the present disclosure, the drawings required for the embodiments are briefly introduced below. It should be understood that the following drawings only show some embodiments of the present disclosure and therefore should not be considered as limiting the scope. For those of ordinary skill in the art, other related drawings can be obtained based on these drawings without creative efforts.

[0035] Fig. 1 shows a schematic flowchart of a language processing model training method provided in an embodiment of the present disclosure;

[0036] Fig. 2 shows another schematic flowchart of a language processing model training method provided in an embodiment of the present disclosure;

[0037] Fig. 3 shows a framework schematic diagram of a large speech synthesis model provided in an embodiment of the present disclosure;

[0038] Fig. 4 shows another schematic flowchart of a language processing model training method provided in an embodiment of the present disclosure;

[0039] Fig. 5 shows a module block diagram of a language processing model training apparatus provided in an embodiment of the present disclosure; and

[0040] Fig. 6 shows a block schematic diagram of an electronic device provided in an embodiment of the present disclosure.

[0041] Reference numbers: 100 - electronic device; 110 - memory; 120 - processor; 130 - communication module; 200 - language processing model training apparatus; 210 - predicted speech token data acquisition module; 220 - style feature acquisition module; 230 - loss function acquisition module; 240 - optimization training module.

DETAILED DESCRIPTION OF EMBODIMENTS

[0042] To make the objectives, technical solutions, and advantages of the embodiments of the present disclosure clearer, the technical solutions in the embodiments of the present disclosure will be clearly and completely described below in conjunction with the drawings in the embodiments of the present disclosure. Obviously, the described embodiments are only some of the embodiments of the present disclosure, rather than all of them. Typically, the elements of the embodiments of the present disclosure described and shown in the drawings herein can be arranged and designed in various different configurations.

[0043] Therefore, the detailed descriptions of the embodiments of the present disclosure provided in the drawings are not intended to limit the scope of the claimed present disclosure, but merely represent selected embodiments of the present disclosure. For those skilled in the art, all other embodiments obtained without creative efforts based on the embodiments of the present disclosure shall fall within the protection scope of the present disclosure.

[0044] It should be noted that relative terms such as “first” and “second” are only used to distinguish one entity or operation from another entity or operation, and do not necessarily require or imply any actual relationship or sequence between these entities or operations. Additionally, terms such as “comprise”, “include”, or any other variations thereof are intended to cover non-exclusive inclusion, such that a process, method, article, or device comprising a series of elements not only comprises those elements but also may comprise other elements not explicitly listed, or elements inherent to such process, method, article, or device. Without further limitations, an element limited by the statement of “comprising a...” does not exclude the presence of additional identical elements in the process, method, article, or apparatus.

[0045] With the rapid development of information technology, speech synthesis technology (Text-To-Speech, TTS), as a technology for converting text into natural and smooth speech, has become an important component of human-computer interaction and directly affects the user experience. TTS is widely used in fields such as outbound broadcasting, voice navigation, and audiobooks.

[0046] Prior large speech synthesis models mainly consist of three components: a language processing model (LLM), an acoustic feature prediction model, and a vocoder model. The predicted speech token data output by the language processing model in the first stage mainly comprises semantic information and partial style information. The loss function of the prior language processing model mainly focuses on semantic information but pays less attention to style information. Therefore, when the large speech synthesis model performs speech synthesis, it cannot well reproduce the speech style, resulting in poor performance of the large speech synthesis model in handling style information (such as prosody, and emotion).

[0047] To address the above problems, the inventors propose the language processing model training method, apparatus, server, and storage medium provided in the embodiments of the present disclosure. By introducing style feature loss and performing optimization training on the language processing model according to the total loss function, the performance of the language processing model in handling speech style information is improved, thereby improving the performance of the large speech synthesis model in handling speech style information. The specific language processing model training method is described in detail in the subsequent embodiments.

[0048] Please refer to Fig. 1, Fig. 1 shows a schematic flowchart of a language processing model training method provided in an embodiment of the present disclosure. The language processing model training method may be applied to an electronic device. Taking the electronic device as an example below, the specific process of this embodiment is described. The following provides a detailed explanation of the process shown in Fig. 1, and the language processing model training method may specifically include the following steps, including steps 110 to 150.

[0049] Step 110: acquiring training text data and ground-truth speech token data.

[0050] Here, the training text data refers to text sentences that have been sorted and annotated. By inputting the training text data into the language processing model, the language processing model can learn how to generate corresponding predicted speech token data based on the training text data and the ground-truth speech token data. The ground-truth speech token data are discrete tokens extracted from reference speech. The ground truth speech token data includes the actual semantic information and partial actual style information of the reference speech. The actual semantic information refers to the real semantic content of the reference speech, and the actual style information refers to the real stylistic characteristics of the reference speech. Examples of style information may include speaking rate, prosody, pitch, emotion, etc. The content of the reference speech may be the same as that of the training text data, that is, the training text data is the text sentence corresponding to the reference speech.

[0051] In some embodiments, the electronic device may acquire reference speech, preprocess the reference speech (e.g., cleaning, denoising, and normalization), extract features from the preprocessed reference speech to obtain speech features, and tokenize each frame of speech features to obtain ground-truth speech token data.

[0052] Step 120: inputting the training text data and the ground-truth speech token data into a prediction module of a language processing model to obtain predicted speech token data output by the prediction module.

[0053] Here, the prediction module may be an LLM model.

[0054] In some embodiments, the electronic device may obtain the predicted speech token data of the training text data output by the prediction module in the language processing model through the following formula. The predicted speech token data comprises the predicted semantic information and predicted style information of the reference speech, where the predicted semantic information refers to the predicted semantic content of the reference speech, and the predicted style information refers to the predicted stylistic characteristics of the reference speech,

[0055] $$speech_token_{predict} = LLM(speech_token_{truth}, text),$$

[0056] where $speech_token_{predict}$ represents predicted speech token data, LLM represents the LLM model, $speech_token_{truth}$ represents ground-truth speech token data, and text represents training text data.

[0057] Step 130: acquiring ground-truth style features based on the ground-truth speech token data and acquiring predicted style features based on the predicted speech token data.

[0058] In some embodiments, the electronic device may perform style extraction on the ground-truth speech token data to obtain ground-truth style features, and may also perform style extraction on the predicted speech token data to obtain predicted style features.

[0059] Step 140: acquiring a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features.

[0060] In some embodiments, the electronic device may construct a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features.

[0061] Step 150: performing optimization training on the prediction module according to the total loss function.

[0062] In some embodiments, the prediction module may gradually adjust internal parameters based on the feedback of the total loss function to improve its capability in generating predicted speech token data and capturing style information. The prediction module may continuously iterate according to the total loss function until the total loss function reaches a threshold or tends to stabilize.

[0063] The language processing model training method provided in the embodiments of the present disclosure comprises: acquiring training text data and ground-truth speech token data and inputting them into a prediction module of a language processing model to obtain predicted speech token data; extracting style features from both the ground-truth and predicted speech token data respectively; constructing a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and performing optimization training on the prediction module according to the total loss function. By incorporating the style feature loss into the total loss function, the language processing model is enabled to learn style information, thereby allowing the language processing model to more accurately reproduce speech styles. This improves the performance

of the language processing model in handling speech style information. Since the speech processing model is a key component of a speech synthesis model, the performance of a large speech synthesis model in handling speech style information can thus be enhanced.

[0064] To ensure that the language processing model can learn style information, a style encoding module is introduced into the language processing model. Therefore, the language processing model comprises the style encoding module. As shown in Fig. 2, step 130 specifically comprises the following steps, including steps 131 and 132.

[0065] Step 131: performing style extraction on the ground-truth speech token data through the style encoding module to obtain ground-truth style features.

[0066] The architecture of the large speech synthesis model with the introduced style encoding module is shown in Fig. 3. The training text data (text) and the ground-truth speech token data ($speech_token_{truth}$) are input into the language processing model (LLM) of the first stage of the large speech synthesis model to obtain the predicted speech token data ($speech_token_{predict}$) output by the language processing model (LLM). The style encoding module (style encoder) performs style extraction on the ground-truth speech token data ($speech_token_{truth}$) to obtain ground-truth style features (style truth), and performs style extraction on the predicted speech token data ($speech_token_{predict}$) to obtain predicted style features (style predict). The predicted speech token data ($speech_token_{predict}$) output by the language processing model is input into the acoustic feature prediction model (Token-to-mel) of the second stage of the large speech synthesis model, and then the data output by the acoustic feature prediction model (Token-to-mel) is input into the vocoder model (Mel-to-wav).

[0067] In some embodiments, the style encoding module may utilize Linear Predictive Coding (LPC) technology to perform style extraction on speech tokens and obtain style features.

[0068] It should be noted that the LPC model characterizes the spectral characteristics of speech token data by extracting linear prediction coefficients (LPC coefficients). The LPC coefficients can reflect the formant positions and shapes of the speech token data, thereby indirectly representing style features of speech, such as intonation, speech rate, and timbre.

[0069] In some embodiments, the electronic device may obtain the ground-truth style features through the following formula:

[0070] $style_hidden_{truth}=Style_encoder(speech_token_{truth})$,

[0071] where $style_hidden_{truth}$ represents the ground-truth style features, Style_encoder represents the style encoding module, and $speech_token_{truth}$ represents the ground-truth speech token data.

[0072] Please continue to refer to Fig. 2. Step 132: performing style extraction on the predicted speech token data through the style encoding module to obtain predicted style features.

[0073] In some embodiments, the electronic device may obtain the predicted style features through the following formula:

[0074] $style_hidden_{predict}=Style_encoder(speech_token_{predict})$,

[0075] where $style_hidden_{predict}$ represents the predicted style features, Style_encoder represents the style encoding module, and $speech_token_{predict}$ represents the predicted speech token data.

[0076] In some embodiments, the style encoding module comprised in the language processing model is a pre-trained model. During the training process of the prediction module, the style encoding module serves as a component for supervising the training of the language processing model.

[0077] In other embodiments, the style encoding module comprised in the language processing model is untrained, and its parameters can be adjusted simultaneously during the training process of the prediction module.

[0078] It can be understood that by using a unified style encoding module to perform style extraction on both the ground-truth speech token data and the predicted speech token data respectively, the ground-truth style features and predicted style features can be compared under the same standard, thereby improving the accuracy of the style loss function calculation during the training process of the language processing model and further enhancing the performance of the language processing model in handling language styles.

[0079] To improve the performance of the language processing model in handling language styles, a style loss function is introduced. As shown in Fig. 4, step 140 specifically comprises the following steps, including steps 141 to 143.

[0080] Step 141: acquiring a prediction loss function based on the ground-truth speech token data and the predicted speech token data.

[0081] Here, the prediction loss function is configured to measure the difference between the ground-truth speech token data and the predicted speech token data.

[0082] In some embodiments, the electronic device may calculate the distribution distance between the ground-truth speech token data and the predicted speech token data to obtain the prediction loss function.

[0083] In some embodiments, the electronic device may calculate the cross-entropy between the ground-truth speech token data and the predicted speech token data through the following formula to calculate the distribution distance between the ground-truth speech token data and the predicted speech token data, thereby acquiring the prediction loss function:

[0084] $$Loss_{ce}=Cross_Entropy(speech_token_{truth}, speech_token_{predict}),$$

[0085] where $Loss_{ce}$ represents the prediction loss function, Cross_Entropy ($speech_token_{truth}$, $speech_token_{predict}$) represents the cross-entropy loss between the ground-truth speech token data and the predicted speech token data, $speech_token_{truth}$ represents the ground-truth speech token data, and $speech_token_{predict}$ represents the predicted speech token data.

[0086] Step 142: acquiring a style loss function based on the ground-truth style features and the predicted style features.

[0087] In some embodiments, the similarity between the ground-truth style features and the predicted style features is calculated to obtain the style loss function.

[0088] Here, the electronic device may calculate the similarity between the ground-truth style features and the predicted style features through techniques such as cosine similarity, KL divergence, and Dynamic Time Warping (DTW).

[0089] In some embodiments, the electronic device may calculate the cosine similarity between the ground-truth style features and the predicted style features through the following formula to obtain the prediction loss function:

[0090] $$Loss_{style}=Cosine_Embedding(style_hidden_{truth}, style_hidden_{predict}),$$

[0091] where $Loss_{style}$ represents the style loss function, Cosine_Embedding ($style_hidden_{truth}$, $style_hidden_{predict}$) represents the cosine similarity between the ground-truth

style features and the predicted style features, $style_hidden_{truth}$ represents the ground-truth style features, and $style_hidden_{predict}$ represents the predicted style features.

[0092] Step 143: acquiring the total loss function based on the prediction loss function and the style loss function.

[0093] In some embodiments, the prediction loss function and the style loss function are performed a weighted calculation to obtain the total loss function.

[0094] In some embodiments, a preset weighting value is preset and stored in the electronic device. The electronic device may perform weighted calculation on the prediction loss function and the style loss function through the following formula to obtain the total loss function:

[0095] $$Loss_{llm} = \alpha \times Loss_{ce} + (1-\alpha) \times Loss_{style},$$

[0096] where $Loss_{llm}$ represents the total loss function, α represents a preset weighting value, $Loss_{ce}$ represents the prediction loss function, and $Loss_{style}$ represents the style loss function.

[0097] It should be noted that the prediction loss function optimizes the hard matching between the ground-truth speech token data and the predicted speech token data. When the difference between the ground-truth speech token data and the predicted speech token data is smaller, the style information contained in the speech token data may not be accurately predicted. Therefore, a trade-off is made between the differences in semantic information and style information through weighted calculation.

[0098] The language processing model training method provided in the embodiments of the present disclosure ensures that the speech generated by the large speech synthesis model is consistent with the style of the training text data and the ground-truth speech token data (speech token data corresponding to the reference speech) in both semantic and stylistic aspects by introducing an additional style encoding module and a style loss function, thereby achieving more precise speech style control and enhancing the user experience.

[0099] To execute the corresponding steps in the above embodiments and various possible implementations, an implementation manner of a language processing model training apparatus is provided below. Optionally, please refer to Fig. 5, which is a functional module diagram of a language processing model training apparatus provided in an embodiment of the present disclosure. It should be noted that the basic principle and technical effect of the

language processing model training apparatus provided in this embodiment are the same as those of the above embodiments. For brief description, for parts not mentioned in this embodiment, reference may be made to the corresponding contents in the above embodiments. The language processing model training apparatus 200 comprises: a predicted speech token data acquisition module 210, a style feature acquisition module 220, a loss function acquisition module 230, and an optimization training module 240, wherein:

[00100] The predicted speech token data acquisition module 210 is configured to: acquire training text data and ground-truth speech token data; input the training text data and the ground-truth speech token data into a prediction module of a language processing model to obtain predicted speech token data output by the prediction module.

[00101] The style feature acquisition module 220 is configured to: acquire ground-truth style features based on the ground-truth speech token data and acquire predicted style features based on the predicted speech token data.

[00102] The loss function acquisition module 230 is configured to: acquire a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features.

[00103] The optimization training module 240 is configured to: perform optimization training on the language processing model according to the total loss function.

[00104] Optionally, the language processing model comprises a style encoding module. The style feature acquisition module is further configured to: perform style extraction on the ground-truth speech token data through the style encoding module to obtain ground-truth style features; and perform style extraction on the predicted speech token data through the style encoding module to obtain predicted style features.

[00105] Optionally, the loss function acquisition module 230 is specifically further configured to: acquire a prediction loss function based on the ground-truth speech token data and the predicted speech token data; acquire a style loss function based on the ground-truth style features and the predicted style features; acquire the total loss function based on the prediction loss function and the style loss function.

[00106] Optionally, the loss function acquisition module 230 is specifically further configured to: calculate the distribution distance between the ground-truth speech token data and the predicted speech token data to obtain the prediction loss function.

[00107] Optionally, the loss function acquisition module 230 is specifically further configured to: calculate the similarity between the ground-truth style features and the predicted style features to obtain the style loss function.

[00108] Optionally, the loss function acquisition module 230 is specifically further configured to: perform a weighted calculation on the prediction loss function and the style loss function to obtain the total loss function.

[00109] Those skilled in the art can clearly understand that for the convenience and conciseness of description, the specific working processes of the above-described apparatus and modules may refer to the corresponding processes in the foregoing method embodiments, and will not be repeated here.

[00110] In the several embodiments provided in this disclosure, the coupling between modules may be electrical, mechanical, or other forms of coupling.

[00111] Additionally, in various embodiments of this disclosure, the functional modules may be integrated into one processing module, exist individually as physical modules, or be integrated into one module with two or more modules. The above integrated modules may be implemented in the form of hardware or in the form of software functional modules.

[00112] Please refer to Fig. 6, which is a block diagram of an electronic device 100 provided in an embodiment of the present disclosure. The electronic device 100 comprises a memory 110, a processor 120, and a communication module 130. The memory 110, processor 120, and communication module 130 are directly or indirectly electrically connected to each other to enable data transmission or interaction. For example, these components may be electrically connected to each other through one or more communication buses or signal lines.

[00113] The memory 110 is configured to store programs or data. The memory 110 may be, but is not limited to, Random Access Memory (RAM), Read Only Memory (ROM), Programmable Read-Only Memory (PROM), Erasable Programmable Read-Only Memory (EPROM), Electrically Erasable Programmable Read-Only Memory (EEPROM), etc.

[00114] The processor 120 is configured to read/write data or programs stored in the memory and perform corresponding functions. For example, when the computer program stored in the memory 110 is executed by the processor 120, it can implement the language processing model training method disclosed in the above embodiments.

[00115] The communication module 130 is configured to establish a communication connection between the electronic device 100 and a cloud server through a network, and to send and receive data via the network.

[00116] It should be understood that the structure shown in Fig. 6 is only a schematic diagram of the electronic device. The electronic device may comprise more or fewer components than those shown in Fig. 6. The components shown in Fig. 6 may be implemented by hardware, software, or a combination thereof.

[00117] The embodiments of the present disclosure also provide a computer-readable storage medium having a computer program stored thereon, wherein the computer program, when executed by a processor, implements the language processing model training method described in the above method embodiments.

[00118] The computer-readable storage medium may be an electronic memory such as flash memory, EEPROM (electrically erasable programmable read-only memory), EPROM, hard disk, or ROM. Optionally, the computer-readable storage medium comprises a non-transitory computer-readable storage medium. The computer-readable storage medium has storage space for program code that executes any of the method steps described above. The program code can be read from one or more computer program products or written into one or more computer program products, and the program code may be compressed in a suitable form.

[00119] In the several embodiments provided in this disclosure, it should be understood that the disclosed apparatus and method can also be implemented in other ways. The apparatus embodiments described above are merely illustrative. For example, the flowcharts and block diagrams in the drawings show possible implementation architectures, functions, and operations of apparatuses, methods, and computer program products according to multiple embodiments of the present disclosure. In this regard, each block in a flowchart or block diagram may represent a module, segment, or portion of code, which contains one or more executable instructions configured to implement the specified logical function(s). It should also

be noted that in some alternative implementations, the functions marked in the blocks may occur in an order different from that marked in the drawings. For example, two consecutive blocks may in fact be executed substantially in parallel, or they may sometimes be executed in the reverse order, depending on the functions involved. It should also be noted that each block of the block diagrams and/or flowcharts, and combinations of blocks in the block diagrams and/or flowcharts, can be implemented by a dedicated hardware-based system that performs the specified functions or actions, or by a combination of dedicated hardware and computer instructions.

[00120] If the functions are implemented in the form of software functional modules and sold or used as independent products, they may be stored in a computer-readable storage medium. Based on this understanding, the technical solution of the present disclosure—in essence, or the part that contributes to the related art, or a part of the technical solution—may be embodied in the form of a computer software product. The computer software product is stored in a storage medium and comprises several instructions for causing a computer device (which may be a personal computer, server, network device, etc.) to execute all or part of the steps of the methods in the various embodiments of the present disclosure. The aforementioned storage media comprise: various media that can store program code, such as USB flash drives, mobile hard disks, Read-Only Memory (ROM), Random Access Memory (RAM), magnetic disks, or optical discs.

[00121] The above are only optional embodiments of the present disclosure and are not intended to limit the present disclosure. For those skilled in the art, the present disclosure may have various modifications and changes. Any modifications, equivalent replacements, improvements, etc. made within the spirit and principles of the present disclosure shall be comprised in the protection scope of the present disclosure.

Industrial Applicability

[00122] In the above, the solutions comprise: acquiring training text data and ground-truth speech token data and inputting them into a prediction module of a language processing model to obtain predicted speech token data; extracting style features from both the ground-truth and predicted speech token data respectively; constructing a total loss function based on the ground-

truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and then performing optimization training on the prediction module according to the total loss function. By incorporating the style feature loss into the total loss function, the language processing model is enabled to learn style information, thereby allowing the language processing model to more accurately reproduce speech styles. This improves the performance of the language processing model in handling speech style information. Since the speech processing model is a key component of a speech synthesis model, the performance of a large speech synthesis model in handling speech style information can thus be enhanced.

CLAIMS

Claim 1: A language processing model training method, comprising:

acquiring training text data and ground-truth speech token data;

inputting the training text data and the ground-truth speech token data into a prediction module of a language processing model to obtain predicted speech token data output by the prediction module;

acquiring ground-truth style features based on the ground-truth speech token data and acquiring predicted style features based on the predicted speech token data;

acquiring a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and

performing optimization training on the prediction module according to the total loss function.

Claim 2: The method according to claim 1, wherein the language processing model comprises a style encoding module, and the step of acquiring ground-truth style features based on the ground-truth speech token data and acquiring predicted style features based on the predicted speech token data comprises:

performing style extraction on the ground-truth speech token data through the style encoding module to obtain the ground-truth style features; and

performing style extraction on the predicted speech token data through the style encoding module to obtain the predicted style features.

Claim 3: The method according to claim 1 or 2, wherein the step of acquiring a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features comprises:

acquiring a prediction loss function based on the ground-truth speech token data and the predicted speech token data;

acquiring a style loss function based on the ground-truth style features and the predicted style features; and

acquiring the total loss function based on the prediction loss function and the style loss function.

Claim 4: The method according to claim 3, wherein the step of acquiring a prediction loss function based on the ground-truth speech token data and the predicted speech token data comprises:

calculating a distribution distance between the ground-truth speech token data and the predicted speech token data to obtain the prediction loss function.

Claim 5: The method according to claim 3 or 4, wherein the step of acquiring a style loss function based on the ground-truth style features and the predicted style features comprises:

calculating a similarity between the ground-truth style features and the predicted style features to obtain the style loss function.

Claim 6: The method according to any one of claims 3 to 5, wherein the step of acquiring the total loss function based on the prediction loss function and the style loss function comprises:

performing a weighted calculation on the prediction loss function and the style loss function to obtain the total loss function.

Claim 7: The method according to any one of claims 2 to 6, wherein the step of performing style extraction on the ground-truth speech token data through the style encoding module to obtain the ground-truth style features comprises:

performing style extraction on the ground-truth speech token data through the style encoding module to obtain linear prediction coefficients of the ground-truth speech token data, wherein the linear prediction coefficients of the ground-truth speech token data characterize spectral features of the ground-truth speech token data; and

acquiring the ground-truth style features based on the spectral features of the ground-truth speech token data.

Claim 8: The method according to any one of claims 2 to 7, wherein the step of performing style extraction on the predicted speech token data through the style encoding module to obtain the predicted style features comprises:

performing style extraction on the predicted speech token data through the style encoding module to obtain linear prediction coefficients of the predicted speech token data, wherein the linear prediction coefficients of the predicted speech token data characterize spectral features of the predicted speech token data; and

acquiring the predicted style features based on the spectral features of the predicted speech token data.

Claim 9: The method according to any one of claims 4 to 8, wherein the step of calculating a distribution distance between the ground-truth speech token data and the predicted speech token data to obtain the prediction loss function comprises:

calculating a cross-entropy between the ground-truth speech token data and the predicted speech token data; and

acquiring the prediction loss function based on the cross-entropy.

Claim 10: The method according to any one of claims 2 to 8, wherein the method further comprises:

performing optimization training on the style encoding module according to the total loss function.

Claim 11: A language processing model training apparatus, comprising:

a predicted speech token data acquisition module, configured to acquire training text data and ground-truth speech token data; input the training text data and the ground-truth speech token data into a prediction module of a language processing model to obtain predicted speech token data output by the prediction module;

a style feature acquisition module, configured to acquire ground-truth style features based on the ground-truth speech token data and acquire predicted style features based on the predicted speech token data;

a loss function acquisition module, configured to acquire a total loss function based on the ground-truth speech token data, the predicted speech token data, the ground-truth style features, and the predicted style features; and

an optimization training module, configured to perform optimization training on the prediction module according to the total loss function.

Claim 12: The apparatus according to claim 11, wherein the language processing model comprises a style encoding module, and the style feature acquisition module is further configured to perform style extraction on the ground-truth speech token data through the style encoding module to obtain the ground-truth style features; and to perform style extraction on the predicted speech token data through the style encoding module to obtain the predicted style features.

Claim 13: An electronic device, comprising a processor and a memory, wherein the memory stores a computer program executable by the processor to implement the language processing model training method according to any one of claims 1 to 10.

Claim 14: A computer-readable storage medium having a computer program stored thereon, wherein the computer program, when executed by a processor, implements the language processing model training method according to any one of claims 1 to 10.