

SPEECH SYNTHESIS

FIELD

[0001] The present disclosure generally relates to text-to-speech technology, and more specifically, to a method, apparatus, device and computer readable storage medium for speech synthesis.

BACKGROUND

[0002] Recently, there has been a progress in the development of text-to-speech (TTS) technologies. Among these technologies, adaptive TTS systems are capable of cloning personalized voices given a few minutes of speech data. However, the performance of these systems relies heavily on the quality and quantity of the data utilized during the fine-tuning phases. Insufficient data during the fine-tuning stages may lead to diminished audio naturalness or speech intelligibility. Moreover, the computational demands also constrain its application for cloning everyone's voice.

SUMMARY

[0003] In a first aspect of the present disclosure, there is provided a method for speech synthesis. The method comprises: obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech; extracting, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech; extracting, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech; extracting, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and generating, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

[0004] In a second aspect of the present disclosure, there is provided an apparatus for speech synthesis. The apparatus comprises: an input information obtaining module configured to obtain input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech; a timbre feature information extracting module configured to extract, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech; a prosody feature information extracting module configured to extract, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech; a content feature information extracting module configured to extract, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and a target speech generating module configured to generate, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

[0005] In a third aspect of the present disclosure, there is provided an electronic device. The electronic device comprises: at least one processing unit; and at least one memory coupled to the at least one processing unit and storing instructions executable by the at least one processing unit, the instructions, upon execution by the at least one processing unit, causing the electronic device to perform: obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech; extracting, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech; extracting, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech; extracting, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and generating, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content

feature information.

[0006] In a fourth aspect of the present disclosure, a computer-readable storage medium is provided. The computer-readable storage medium stores computer executable instructions which, when executed by an electronic device, causes the electronic device perform operations comprising: obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech; extracting, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech; extracting, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech; extracting, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and generating, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

[0007] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

[0008] The above and other features, advantages and aspects of the embodiments of the present disclosure will become more apparent in combination with the accompanying drawings and with reference to the following detailed description. In the drawings, the same or similar reference symbols refer to the same or similar elements, where:

[0009] FIG. 1 illustrates a schematic diagram of an example environment in which embodiments of the present disclosure may be implemented;

[0010] FIG. 2A illustrates a schematic diagram of a training process of a machine learning model utilized by a TTS system in accordance with some embodiments of the present disclosure;

[0011] FIG. 2B illustrates a schematic diagram of an inference process of a TTS system in accordance with some embodiments of the present disclosure;

[0012] FIG. 3 illustrates a schematic diagram of performing an interpolation operation on two prosodies in accordance with embodiments of the present disclosure;

[0013] FIG. 4 illustrates a flowchart of a process for speech synthesis in accordance with some embodiments of the present disclosure;

[0014] FIG. 5 illustrates a block diagram of an apparatus for speech synthesis in accordance with some embodiments of the present disclosure; and

[0015] FIG. 6 illustrates a block diagram of an electronic device in which one or more embodiments of the present disclosure can be implemented.

DETAILED DESCRIPTION

[0016] The embodiments of the present disclosure will be described in more detail below with reference to the accompanying drawings. Although some embodiments of the present disclosure are shown in the drawings, it would be appreciated that the present disclosure may be implemented in various forms and should not be interpreted as limited to the embodiments described herein. On the contrary, these embodiments are provided for a more thorough and complete understanding of the present disclosure. It would be appreciated that the drawings and embodiments of the present disclosure are only for the purpose of illustration and are not intended to limit the scope of protection of the present disclosure.

[0017] In the description of the embodiments of the present disclosure, the term "including" and similar terms would be appreciated as open inclusion, that is, "including but not limited to". The term "based on" would be appreciated as "at least partially based on". The term "one embodiment" or "the embodiment" would be appreciated as "at least one embodiment". The term "some embodiments" would be appreciated as "at least some embodiments". Other explicit and implicit definitions may also be included below. As used herein, the term "model" can represent the matching degree between various data. For example, the above matching degree can be obtained based on various technical solutions currently available and/or to be developed in the future.

[0018] It will be appreciated that the data involved in this technical proposal (including but not limited to the data itself, data acquisition or use) shall comply with the requirements of

corresponding laws, regulations and relevant provisions.

[0019] It will be appreciated that before using the technical solution disclosed in each embodiment of the present disclosure, users should be informed of the type, the scope of use, the use scenario, etc. of the personal information involved in the present disclosure in an appropriate manner in accordance with relevant laws and regulations, and the user's authorization should be obtained.

[0020] For example, in response to receiving an active request from a user, a prompt message is sent to the user to explicitly prompt the user that the operation requested operation by the user will need to obtain and use the user's personal information. Thus, users may select whether to provide personal information to the software or the hardware such as an electronic device, an application, a server or a storage medium that perform the operation of the technical solution of the present disclosure according to the prompt information.

[0021] As an optional but non-restrictive implementation, in response to receiving the user's active request, the method of sending prompt information to the user may be, for example, a pop-up window in which prompt information may be presented in text. In addition, pop-up windows may also contain selection controls for users to choose "agree" or "disagree" to provide personal information to electronic devices.

[0022] It will be appreciated that the above notification and acquisition of user authorization process are only schematic and do not limit the implementations of the present disclosure. Other methods that meet relevant laws and regulations may also be applied to the implementation of the present disclosure.

[0023] As used herein, the term "model" can learn a correlation between respective inputs and outputs from training data, so that a corresponding output can be generated for a given input after training is completed. The generation of the model can be based on machine learning techniques. Deep learning is a machine learning algorithm that processes inputs and provides corresponding outputs by using multiple layers of processing units. A neural networks model is an example of a deep learning-based model. As used herein, "model" may also be referred to as "machine learning model", "learning model", "machine learning network", or "learning network", and these terms are used interchangeably herein.

[0024] "Neural networks" are a type of machine learning network based on deep learning. Neural networks are capable of processing inputs and providing corresponding outputs, typically

comprising input and output layers and one or more hidden layers between the input and output layers. Neural networks used in deep learning applications typically comprise many hidden layers, thereby increasing the depth of the network. The layers of neural networks are sequentially connected so that the output of the previous layer is provided as input to the latter layer, where the input layer receives the input of the neural network and the output of the output layer serves as the final output of the neural network. Each layer of a neural network comprises one or more nodes (also known as processing nodes or neurons), each of which processes input from the previous layer.

[0025] Usually, machine learning can roughly comprise three stages, namely training stage, test stage, and application stage (also known as inference stage). During the training stage, a given model can be trained using a large scale of training data, iteratively updating parameter values until the model can obtain consistent inference from the training data that meets the expected objective. Through the training, the model can be considered to learn the correlation between input and output (also known as input-to-output mapping) from the training data. The parameter values of the trained model are determined. In the test stage, test inputs are applied to the trained model to test whether the model can provide correct outputs, thereby determining the performance of the model. In the application stage, the model can be used to process actual inputs and determine corresponding outputs based on the parameter values obtained from training.

[0026] FIG. 1 illustrates a block diagram of an example environment 100 in which various embodiments of the present disclosure may be implemented. In the environment 100 of FIG. 1, a text-to-speech (TTS) system 110 applies a machine learning model 105 to perform speech generation. The machine learning model 105 is configured to generate a speech 114. The machine learning model 105 is configured to process a text 112 input by a user to generate the speech 114.

[0027] In FIG. 1, the TTS system 110 may be implemented at any computing system with computing capability, such as various computing devices/systems, terminal devices, servers, etc. Terminal devices may include any type of mobile terminals, fixed terminals, or portable terminals, including mobile phones, desktop computers, laptops, netbooks, tablets, media computers, multimedia tablets, or any combination of the aforementioned, including accessories and peripherals of these devices or any combination thereof. Servers include but are not limited to mainframe, edge computing nodes, computing devices in cloud environment, etc.

[0028] It should be understood that the structure and function of each element in the

environment 100 is described for illustrative purposes only and does not imply any limitations on the scope of the present disclosure.

[0029] As mentioned above, adaptive TTS systems rely heavily on the data utilized during the fine-tuning phases. To reduce such a reliance, related works leverage generative models to perform zero-shot TTS. These powerful models can effectively synthesize speech given only a single speech prompt, eliminating the need for data preparation and the computational requirements for fine-tuning methods. However, the prompting mechanisms of current solutions still face two primary challenges.

[0030] One challenge is a lack of multi-sentence prompting strategies. Related works of zero-shot TTS typically employ single-sentence speech prompts during training. In inference, the information in the single-sentence speech prompt is insufficient to guide the zero-shot TTS systems to imitate the voice variability of a natural person perfectly. From another perspective, the performance of fine-tuning methods can be further improved by increasing the amount of data, while zero-shot TTS systems lack an appropriate strategy to extract useful information from multi-sentence speech prompts.

[0031] The other challenge is a lack of specialized prompting mechanism for prosodic information. Current solutions for zero-shot TTS primarily concentrate on improving the similarity of timbre and prosody between the generated speech and the prompts. However, they neglect to express various unseen prosodic styles in a controlled manner while also preserving the unique timbre of the given one-sentence prompt. In order to control the prosodic styles, it is necessary to disentangle the prosody information from speech prompts.

[0032] Adaptive TTS focuses on synthesizing personalized voice for any user with few data. During the adaptation process, a TTS system pre-trained on a multi-speaker speech dataset is typically fine-tuned with few adaptation data for the target voice. Some related works design independent learned embeddings for each speaker, which requires few data at deployment time to adapt to new speakers rapidly. Some related works propose an acoustic-condition modeling method for high-quality and efficient customization of new voices. There are also some works leveraging the meta-learning approach and data augmentation for speaker adaptation. However, although some works are data-efficient and parameter-efficient, these systems still suffer from audio quality issues when data size is small, as well as computational cost issues due to hundreds of fine-tuning steps.

[0033] Zero-shot TTS aims to synthesize unseen voices with a speaker encoder that extracts speaker embeddings from the reference audio. This scenario is highly attractive because it does

not require any adaptation data or parameters. The attention-based adaptation method utilizes attention mechanisms to extract fine-grained speech features from reference audios. Some related works propose to extract useful style information from arbitrary number of reference audios. However, they do not separately model the timbre and prosody information, lacking controllability over timbre and prosody. Some recent works propose to use in-context learning methods to efficiently extract speaker information from acoustic prompts and have achieved progress in zero-shot TTS. Some related works propose the neural codec language model that exhibits strong in-context learning capability for zero-shot speech generation. Some related works introduce in-context learning to latent diffusion model, which is achieved by partitioning a speech clip into the prompt and target regions. However, these methods are trained with single-sentence prompts, lacking an appropriate strategy to extract fine-grained information from multi-sentence speech prompts.

[0034] Prosody transfer aims to transfer the prosody from a reference utterance to the synthesized target speech, which is essential for producing natural and expressive speech in a controlled manner. Some related works propose to integrate a prosody reference encoder into a TTS system, which is capable of performing similar-text prosody transfer. Some recent works try to transfer prosody in different-text and different-speaker settings with the bottleneck of the prosody encoder. Some related works propose to use a gradient reversal layer to penalize the prosody encoder if its output contains information about the speaker identity from the reference utterance, which enhances the target speaker fidelity for cross-speaker prosody transfer. However, current solutions do not learn a transferable representation of prosody, but rather an utterance-level representation that is relatively dependent on both the reference speaker and reference text.

[0035] To address at least some of the above issues, embodiments of the present disclosure propose an improved solution for speech synthesis. In this solution, input information for a trained acoustic autoencoder model is obtained, where the input information comprises a target text, at least one first reference speech, and at least one second reference speech. Timbre feature information is extracted from the at least one first reference speech using a timbre encoder in the acoustic autoencoder model. The timbre feature information characterizes timbre represented in the at least one first reference speech. Prosody feature information is extracted from the at least one second reference speech using a prosody encoder in the acoustic autoencoder model. The prosody feature information characterizes a prosodic pattern represented in the at least one second reference speech. Content feature information is extracted from the target text using a content

encoder in the acoustic autoencoder model. A target speech corresponding to the target text is generated based on the timbre feature information, the prosody feature information, and the content feature information using a decoder in the acoustic autoencoder model.

[0036] With these embodiments of the present disclosure, timbre feature information and prosody feature information are separately extracted using different encoders in the acoustic autoencoder model, thereby achieving decomposition of the timbre and prosody information from speech. In this way, multi-sentence reference speech may be processed effectively by the acoustic autoencoder model to extract more feature information and thus the quality of the generated speech is improved.

[0037] Example embodiments of the present disclosure will be described with the reference to the drawings.

[0038] Reference is now made to FIG. 2A, which illustrates a schematic diagram 200 of a training process of the machine learning model utilized by the TTS system 110 in accordance with some embodiments of the present disclosure. Such a machine learning model may be constructed as an acoustic autoencoder model 205. The training process of the acoustic autoencoder model 205 is first discussed. The acoustic autoencoder model 205 is configured to reconstruct a speech for a text with reference to one or more reference speeches.

[0039] As shown in FIG. 2A, the acoustic autoencoder model 205 comprises a timbre encoder 218 configured to extract timbre feature information from an input speech, a content encoder 222 configured to extract content feature information from an input text, and a prosody encoder 226 configured to extract prosody feature information from an input speech. The acoustic autoencoder model 205 further includes a decoder 230 configured to construct a speech based on the extracted timbre feature information, content feature information and prosody feature information. It is assumed that a speech (e.g., represented by mel-spectrogram) y may be reconstructed through the following generative process: $y = D(z_c, z_{pd}, z_t, g)$, where z_c and z_t denote the fine-grained content and timbre hidden states (also called content feature information and timbre feature information, respectively), g denotes the global style information that contains timbre and prosody, and D denotes a mel-spectrogram decoder (as an example of the decoder 230) in the acoustic autoencoder model 205. It is assumed that $z_{pd} = (z_p, z_d)$ contains the fine-grained prosodic style information of pitch and energy z_p and duration z_d . $z_d = \text{Aligner}(y)$ may be obtained by external alignment tools and disentangled from z_{pd} . The acoustic autoencoder model 205 may be trained based on the reconstruction of the speech y .

[0040] A sample speech corpus for a target speaker may be partitioned into a first sample speech 210 (denoted as $\tilde{y}$) and a second sample speech 212 (denoted as y_t). $Y = \{y_1, \dots, y_n\}$ is denoted as the speech corpus for a certain speaker S . In training, Y may be partitioned into the target mel-spectrogram (as an example of the second sample speech 210) and other mel-spectrograms (as an example of the first sample speech 212). A reconstructed speech 214 for a sample text 216 may be determined based on the sample text 216, the first sample speech 210, and the second sample speech 212 using the acoustic autoencoder model 205 under training. The sample text 216 may be the text corresponding to the second sample speech 212.

[0041] It is assumed that the mutual information between y_t and $\tilde{y}$ only contains timbre information $H(z_t)$ and global style information $H(g)$ of y_t . This assumption may be expressed as follows:

$$I(y_t; \tilde{y}) = H(z_t) + H(g) \quad (1)$$

[0042] In Eq. (1), $I(y_t; \tilde{y})$ represents the mutual information between y_t and $\tilde{y}$.

[0043] In embodiments of the present disclosure, sample timbre feature information 220 may be extracted from the first sample speech 210 using the timbre encoder 218 in the acoustic autoencoder model 205. Based on the above assumption, sample timbre feature information 220 (also referred to as fine-grained timbre information, denoted as z_t) may be extracted through $E_t(\tilde{y})$, where $E_t()$ represents the extraction process of the timbre encoder 218. In addition, global style information may be extracted through $E_t(\tilde{y})$.

[0044] In embodiments of the present disclosure, sample content feature information 224 may be extracted from the sample text 216 using a content encoder 222 in the acoustic autoencoder model 205. The sample text 216 may be represented by a phoneme sequence. By feeding the phoneme sequence to the content encoder 222 (denoted as E_c), sample content feature information 224 (also referred to as fine-grained content information, denoted as z_c) may be extracted.

[0045] In embodiments of the present disclosure, sample prosody feature information 228 may be extracted from the second sample speech 212 using a prosody encoder 226 in the acoustic autoencoder model 205. Because z_t and z_c are available, the prosody encoder 226 (denoted as E_p) may remove the sample timbre feature information 220 and the sample content feature information 224 if it is forced to lose some information by information bottleneck $B(\cdot)$. The bottleneck forces $E_p(y_t)$ to pass only the sample prosody feature information 228 (also referred to as fine-grained prosodic style, denoted as z_p) that other encoders cannot supply.

[0046] Then, the reconstructed speech 214 may be generated based on the sample prosody feature information 220, the sample timbre feature information 224 and the sample content feature information 228 using the decoder 230 (denoted as D) in the acoustic autoencoder model 205.

[0047] With embodiments of the present disclosure, decomposition of the prosody and timbre may be achieved by using different encoders in the acoustic autoencoder model 205, and thus the extracted sample prosody feature information 228 and sample timbre feature information 220 may be more accurate and the quality of the reconstructed speech 214 may be improved.

[0048] It is to be noted that in order to store timbre information for thousands of speakers, a large number of codebook entries are required. Because the prosody and timbre are decomposed, the sample prosody feature information may be compressed into a highly compact codebook and the sample timbre feature information may be extracted via a powerful speaker encoder. The decomposition strategy not only allows the acoustic autoencoder model 205 to accommodate extremely long prosody prompts (as an example of the second sample speech 212) but also enables the acoustic autoencoder model to control the prosodic styles of generated speeches (also referred to as the reconstructed speech).

[0049] In embodiments of the present disclosure, a vector quantized (VQ) encoder may be design as the prosody encoder 226 and a multi-reference timbre encoder which can process multi-sentence speech prompts may be design as the timbre encoder 218. Because the prosody encoder 226 mainly captures the prosodic variance information, a generative adversarial network (GAN)-based mel-spectrogram decoder is designed as the decoder 230, which models the high-frequency details in spectrograms and ensures perceptually high-quality reconstructions.

[0050] In embodiments of the present disclosure, the acoustic autoencoder model 205 may be trained based on a predetermined training objective. The training objective may be configured to reduce or minimize at least a difference 232 between the second sample speech 212 (the ground-truth speech) and the reconstructed speech 214 (denoted as $\hat{y}_t$). The difference may be measured via a loss function. In an example, the loss function used for the training may be formulated as $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{VQ} + \mathcal{L}_{Adv}$, where $\mathcal{L}_{rec} = \|y_t - \hat{y}_t\|$ represents the reconstruction loss between the second sample speech 212 and the reconstructed speech 214, $\mathcal{L}_{VQ}$ represents the VQ codebook loss, and $\mathcal{L}_{Adv}$ represents the LSGAN-styled adversarial loss whose objective is to minimize the distribution distance between the predicted mel-spectrograms (as an example of the reconstructed speech 214) and the ground truth mel-spectrograms (as an example of the second sample speech 212).

[0051] In some embodiments, the content encoder 222 may include a plurality of feed-forward Transformer layers following common practice in non-autoregressive TTS systems.

[0052] In some embodiments, the vector quantized prosody encoder 226 may include two convolution stacks and a vector quantization bottleneck. The first convolution stacks compress mel-spectrograms into hidden states by a factor of r in length, and the second stacks capture the correlation of features. After that, the vector quantization layer utilizes these hidden states to obtain prosody codes $u = \{u_1, u_2, \dots, u_n\}$ and hidden states z_p . The information bottleneck $B(\cdot)$ of the vector quantized prosody encoder may include temporal compression and the vector quantization layer. It is to be noted that the structures of the content encoder 222 and the prosody encoder 226 described above are merely examples, which are not limited by embodiments of the present disclosure. There may be other structures implemented in the content encoder 222 and the prosody encoder 226.

[0053] In some embodiments, the first sample speech 210 may comprise a plurality of speech clips. The plurality of speech clips may be concatenated to obtain a concatenated speech and the sample timbre feature information 220 may be extracted from the concatenated speech using the timbre encoder. The timbre encoder 218 may be used to extract the sample timbre feature information from multi-sentence speech prompts (e.g., including a plurality of speech clips). Since speakers can change their timbre by using different speaking techniques according to their speaking habits or desired semantic meanings, the timbre encoder 218 may need to extract the sample timbre feature information 220 from multiple prompts that can represent the speakers' habits. Firstly, a concatenation operation 234 may be performed on the plurality of speech clips contained in the first sample speech 210 that belong to the target speaker but are different from the second sample speech 212. Then, the timbre encoder 218 may compress the concatenated speech into the sample timbre feature information by a factor of d in length.

[0054] In some embodiments, a language model 268 may generate sample extended prosody feature information 240 based on the sample content feature information 224 and the sample prosody feature information 226. Then, the reconstructed speech may be generated based on the sample timbre feature information 220, the sample content feature information 224 and the sample extended prosody feature information 240 using the decoder 230.

[0055] In some embodiments, the timbre encoder 218, content encoder 222 and prosody encoder 226 may perform a down-sampling operation on respective input information to reduce the size of the input information. Correspondingly, up-sampling operation 236-1, up-sampling operation 236-2 and repeating operation 238 may be performed on the outputs of the timbre

encoder 218, content encoder 222 and prosody encoder 226 respectively before the decoding process of the decoder 230 to restore the intermediate data to its original dimensions or to a higher resolution.

[0056] After training the acoustic autoencoder model 205, in the inference stage, the acoustic autoencoder model 205 may generate a speech corresponding to a given text.

[0057] Reference is now made to FIG. 2B, which illustrates a schematic diagram 250 of an inference process of the TTS system 110 using the trained acoustic autoencoder model 205 in accordance with some embodiments of the present disclosure. As shown in FIG. 2B, input information for the trained acoustic autoencoder model 205 is obtained firstly. The input information includes at least one first reference speech 252, at least one second reference speech 254 and a target text 256. The at least one first reference speech 252 and the at least one second reference speech 254 may be speeches from different speakers or from the same speaker. In an example, the target text 256 may be represented by phonemes for processing by the content encoder 222.

[0058] In some embodiments, the at least one first reference speech 252 and the at least one second reference speech 254 may include one same speech from the same speaker. In this case, timbre feature information and prosody feature information are extracted from the same speech, to generate the target speech corresponding to the target text 256.

[0059] In some embodiments, the at least one first reference speech 252 and the at least one second reference speech 254 may include different speeches from the same speaker. In this case, timbre feature information is extracted from the at least one first reference speech 252 and prosody feature information is extracted from the at least one second reference speech 254, to generate the target speech corresponding to the target text 256.

[0060] In some embodiments, the at least one first reference speech 252 and the at least one second reference speech 254 may include different speeches from two or more speakers. In this case, timbre feature information is extracted from the at least one first reference speech 252 and different prosody feature information is extracted from the at least one second reference speech 254, to generate the target speech corresponding to the target text 256.

[0061] After the input information is obtained, timbre feature information 258 is extracted from the at least one first reference speech 252 using the timbre encoder 218 in the acoustic autoencoder model 205. The timbre feature information 258 characterizes timbre represented in the at least one first reference speech 252. The timbre encoder 218 may be used to extract fine-grained timbre information from the at least one first reference speech 252.

[0062] In some embodiments, the at least one first reference speech 252 may include a plurality of first speeches of a speaker. Different first speeches may include different timbres according to the speaker's habits or desired semantic meanings. The timbre encoder 218 may extract the timbre feature information from the plurality of first speeches to learn different timbre feature information. First, a concatenation operation 234 may be performed on the plurality of first speeches contained in the first reference speech. Then, the timbre encoder 218 may compress the concatenated speech into the timbre feature information 258.

[0063] In addition, prosody feature information 260 is extracted from the at least one second reference speech 254 using the prosody encoder 226 in the acoustic autoencoder model 205. The prosody feature information 260 characterizes a prosodic pattern represented in the at least one second reference speech 254. The prosody encoder 226 may be used to extract fine-grained prosodic style information from the at least one second reference speech 254.

[0064] Further, content feature information 262 is extracted from the target text 256 (or more specifically, the phonemes) using the content encoder 222 in the acoustic autoencoder model 205. The content encoder 222 may be used to extract fine-grained content information from the target text 256.

[0065] After these feature information are extracted, a target speech 264 corresponding to the target text 256 is generated based on the timbre feature information 258, the prosody feature information 260, and the content feature information 262 using the decoder 230 in the acoustic autoencoder model 205.

[0066] In embodiments of the present disclosure, an attention operation 266 may be performed on the timbre feature information 258 with the content feature information 262, to determine relevant timbre feature information for the target text. The timbre feature information 258 may include information of different timbres (e.g., happy, sad, angry and the like). Relevant timbre feature information may be determined based on the content feature information 262. To extract semantically relevant timbre feature information from speech prompts (e.g., at least one first reference speech 252), a timbre-to-content attention module may be introduced. This module takes the content feature information as the query and takes the timbre feature information as both the key and the value.

[0067] After the relevant timbre feature information is determined, the target speech 264 may be generated based on the relevant timbre feature information, the prosody feature information 260, and the content feature information 262. In some embodiments, the relevant timbre feature information output by the timbre-to-content attention module may be unsampled to match the

length of the target mel-spectrogram (also referred to as the at least one second reference speech) using a length regulator.

[0068] In embodiments of the present disclosure, extended prosody feature information 270 may be generated based on the extracted prosody feature information 260 and the content feature information 262 using a language model 268. The extended prosody feature information characterizes a prosodic pattern with an extended length. The language model 268 may be termed as prosody latent language model (P-LLM) which aims to capture the speaker's prosodic patterns from multi-sentence prompts effectively. In an example, the prosody feature information 260 $\{z_{p1}, z_{p2}, \dots, z_{pn}\}$ and the content feature information 262 $\{z_{c1}, z_{c2}, \dots, z_{cn}\}$ may be extracted from multiple speech clips $\{s_1, s_2, \dots, s_n\}$ of a target speaker using the acoustic autoencoder model 205. The language model 268 may be used to generate the extended prosody feature information 270 based on the extracted prosody feature information 260 and the content feature information 262.

[0069] In embodiments of the present disclosure, a speech duration for the target text may be determined. Then, a length of prosody feature information 260 may be aligned with a length of content feature information 262 according to the speech duration. In an example, the speech duration may include respective durations of respective phonemes in the target text. 256 The prosody feature information 260 and the content feature information 262 may be concatenated along the time axis to construct $z'_p = \text{Concat}(z_{p1}, z_{p2}, \dots, z_{pn})$ and $z'_c = \text{Concat}(z_{c1}, z_{c2}, \dots, z_{cn})$. In order to match the lengths of z'_p 270 and z'_c 272 in the temporal dimension, z'_c 272 is expanded to the frame level with duration information z_d (also referred to as speech duration) and is compressed r times with a max pooling layer.

[0070] After the length alignment is performed, the extended prosody feature information 270 may be generated based on the aligned prosody feature information and the aligned content feature information using the language model. After length alignment, z'_p 270 may be transformed to prosody code u' , and then u' and z'_c are fed into the language model 268, which predicts the extended prosody feature information 270 in an auto-regressive manner:

$$p(u' | z'_c; \theta) = \prod_{l=0}^L p(u'_l | u'_{<l}, z'_c; \theta), \quad (2)$$

[0071] In Eq. (2), θ represents the parameters of the parameters of the language model 268 and L represents the length of the concatenated prosody code u' .

[0072] In some embodiments, during the process of training the language model 268, batch size may be set as 1 to increase the maximum number m of prosody codes in each batch. If the total number of speech frames from a single speaker is less than $m \times r$, speech samples from other speakers may be included in this batch and speaker-level attention masks may be incorporated into the language model 268. The speech prompt (e.g., at least one first reference speech 252 and at least second first reference speech 254) may not be specifically defined and the language model 268 may be trained directly using the concatenated speech samples through the teacher-forcing technique with the cross-entropy loss. To avoid the transition area problems caused by directly concatenating the prompts, a start token and an end token may be assigned to each sentence, which guides the language model 268 to continue writing the current sentence and extract useful information from previous sentences (i.e., generating the extended prosody feature information). In this way, the useful prosody-level information contained in the multi-sentence prompts. Therefore, in the inference stage, users can flexibly improve the generation quality by extending the length of prompts by concatenating the reference speech clips.

[0073] In some embodiments, for duration modeling, a phoneme-level auto-regressive duration model is proposed. This model enhances the duration modeling by leveraging the powerful in-context learning capabilities of auto-regressive models. The overall architecture of the auto-regressive duration model remains the same as the language model, but a mean squared error (MSE) loss is used instead. It is to be noted that the structure of the duration model is merely an example, which is not limited by embodiments of the present disclosure.

[0074] In some embodiments, the timbre encoder 218, content encoder 222 and prosody encoder 226 may perform a down-sampling operation on respective input information to reduce the size of the input information. Correspondingly, up-sampling operation 236-1, up-sampling operation 236-2 and repeating operation 238 may be performed on the outputs of the timbre encoder 218, content encoder 222 and prosody encoder 226 respectively before the decoding process of decoder 230 to restore the intermediate data to its original dimensions or to a higher resolution.

[0075] In embodiments of the present disclosure, at least one second speech 254 may comprise a plurality of second speeches of a plurality of speakers. Respective probability distributions for the plurality of second speeches may be generated using the prosody encoder 226. The respective probability distributions indicate distributions of respective prosody tokens for the plurality of second speeches. In an example, the prosody encoder 226 may first extract

prosody feature information from the plurality of second speeches. Then, the language model 268 may output probability distributions based on the extracted prosody feature information 260.

[0076] After the respective probability distributions are generated, an interpolation operation may be performed on the respective probability distributions to obtain an interpolated probability distribution. In an example, a target speaker has a relatively sad speaking tone, speeches that sound happier for the target speaker may be generated while preserving the target speaker's tone. An interpolation operation may be performed on a probability distribution of a sad speech of the target speaker and a probability distribution of a happy speech of other speakers.

[0077] In embodiments of the present disclosure, the interpolation operation may be performed on the respective probability distributions based on a plurality of weights for the plurality of second speeches. A weight in the plurality of weights is in a range of 0 and 1. FIG. 3 illustrates a schematic diagram 300 of performing an interpolation operation on two prosodies in accordance with embodiments of the present disclosure. As shown in FIG. 3, prosody latent (also referred as prosody feature information) u_a 305 is extracted from speeches in a happy tone of other speakers and the sad prosody latent u_b 310 is extracted from a target speech prompt of the target speaker. Then, language models 268 are utilized to separately decode the target prosody code $\hat{u}$ 320 with the prosodic latent u_a 305 and u_b 310. The prosody latent u_a 305 and u_b 310 are output by the prosody encoder 226 and then are input to the language models 268. These language models 268 share the same parameters. In every step t of the decoding process, the probability distributions output by two language models 268 are interpolated with the weight γ , which can be formulated as follows:

$$p(\hat{u}) = \prod_{t=0}^T \left((1 - \gamma) \cdot p(\hat{u}_t \mid \hat{u}_{<t}, u_b, \text{Concat}(z_{cb}, \hat{z}_c); \theta) + \gamma \cdot p(\hat{u}_t \mid \hat{u}_{<t}, u_a, \text{Concat}(z_{ca}, \hat{z}_c); \theta) \right), \quad (3)$$

[0078] In Eq. (3), z_{cb} and z_{ca} represent the content feature information from speech clips s_b and s_a . $\hat{z}_c$ represents the content feature information of the target sentence. γ is in the range of 0 and 1.

[0079] After the interpolation operation is performed, respective interpolated prosody tokens from the interpolated probability distribution may be determined, to form the prosody feature information. With embodiments of the present disclosure, users can freely control the prosodic style of the generated speech in the inference stage. Moreover, the proposed prosody interpolation algorithm utilizes the autoregressive probability distribution of the language model for prosody transfer. In this way, different prosodies may be mixed in a soft and fine-grained

manner in the autoregressive generation process, and thus the controllability of prosody may be ensured and various speaking styles may be transferred to the desired timbre.

[0080] FIG. 4 illustrates a flowchart of a process 400 for speech synthesis in accordance with some embodiments of the present disclosure. The process 400 may be implemented at the TTS system 110 of FIG. 1.

[0081] At block 410, the TTS system 110 obtains input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech.

[0082] At block 420, the TTS system 110 extracts, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech.

[0083] At block 430, the TTS system 110 extracts, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech.

[0084] At block 440, the TTS system 110 extracts, using a content encoder in the acoustic autoencoder model, content feature information from the target text.

[0085] At block 450, the TTS system 110 generates, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

[0086] In some embodiments, generating the target speech comprises: performing an attention operation on the timbre feature information with the content feature information, to determine relevant timbre feature information for the target text; and generating the target speech based on the relevant timbre feature information, the prosody feature information, and the content feature information.

[0087] In some embodiments, the at least one first reference speech comprises a plurality of first speeches of a speaker, and wherein extracting the timbre feature information from the at least one first reference speech comprises: concatenating the plurality of first speeches to obtain a concatenated speech; and extracting, using the timbre encoder, the timbre feature information from the concatenated speech.

[0088] In some embodiments, generating the target speech comprises: generating, using a language model, extended prosody feature information based on the extracted prosody feature information and the content feature information, the extended prosody feature information characterizing a prosodic pattern with an extended length; and determining the target speech based on the timbre feature information, the extended prosody feature information, and the content feature information.

[0089] In some embodiments, generating the extended prosody feature information comprises: determining a speech duration for the target text; aligning a length of prosody feature information with a length of content feature information according to the speech duration; and generating the extended prosody feature information based on the aligned prosody feature information and the aligned content feature information.

[0090] In some embodiments, the at least one second speech comprises a plurality of second speeches of a plurality of speakers. Extracting the prosody feature information comprises: generating, using the prosody encoder, respective probability distributions for the plurality of second speeches, the respective probability distributions indicating distributions of respective prosody tokens for the plurality of second speeches; performing an interpolation operation on the respective probability distributions to obtain an interpolated probability distribution; and determining respective interpolated prosody tokens from the interpolated probability distribution, to form the prosody feature information.

[0091] In some embodiments, performing the interpolation operation on the respective probability distributions to obtain an interpolated probability distribution comprises: performing the interpolation operation on the respective probability distributions based on a plurality of weights for the plurality of second speeches, a weight in the plurality of weights being in a range of 0 and 1.

[0092] In some embodiments, the acoustic autoencoder model is trained by: partitioning a sample speech corpus for a target speaker into a first sample speech and a second sample speech; determining, using the acoustic autoencoder model under training, a reconstructed speech for a sample text based on the sample text, the first sample speech, and the second sample speech; and training the acoustic autoencoder model based on a predetermined training objective, the training

objective being configured to reduce or minimize a difference between the second sample speech and the reconstructed speech.

[0093] In some embodiments, determining the reconstructed speech comprises: extracting, using the prosody encoder, sample prosody feature information from the second sample speech; extracting, using the timbre encoder, sample timbre feature information from the first sample speech; extracting, using the content encoder, sample content feature information from the sample text; and generating, using the decoder, the reconstructed speech based on the sample prosody feature information, the sample timbre feature information and the sample content feature information.

[0094] In some embodiments, the first sample speech comprises a plurality of speech clips, and wherein extracting the sample timbre feature information from the first sample speech comprises: concatenating the plurality of speech clips to obtain a concatenated speech; and extracting, using the timbre encoder, the sample timbre feature information from the concatenated speech.

[0095] FIG. 5 shows a block diagram of an apparatus 500 for speech synthesis in accordance with some embodiments of the present disclosure. The apparatus 500 may be implemented, for example, or included at the TTS system 110 of FIG. 1. Various modules/components in the apparatus 500 may be implemented by hardware, software, firmware, or any combination thereof.

[0096] As illustrated, the apparatus 500 includes an input information obtaining module 510 configured to obtain input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech.

[0097] The apparatus 500 includes a timbre feature information extracting module 520 configured to extract, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech.

[0098] The apparatus 500 includes a prosody feature information extracting module 530 configured to extract, using a prosody encoder in the acoustic autoencoder model, prosody feature

information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech.

[0099] The apparatus 500 further includes a content feature information extracting module 540 configured to extract, using a content encoder in the acoustic autoencoder model, content feature information from the target text.

[00100] The apparatus 500 further includes a target speech generating module 550 configured to generate, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

[00101] The apparatus 500 further includes corresponding modules that are configured to perform the operations of the process 400 and other embodiments as described herein.

[00102] FIG. 6 illustrates a block diagram of an electronic device 600 in which one or more embodiments of the present disclosure can be implemented. It would be appreciated that the electronic device 600 shown in FIG. 6 is only an example and should not constitute any restriction on the function and scope of the embodiments described herein. The electronic device 600 may be used, for example, to implement the TTS system 110 of FIG. 1. The electronic device 600 may also be used to implement the apparatus 500 of FIG. 5.

[00103] As shown in FIG. 6, the electronic device 600 is in the form of a general computing device. The components of the electronic device 600 may include, but are not limited to, one or more processors or processing units 610, a memory 620, a storage device 630, one or more communication units 640, one or more input devices 650, and one or more output devices 660. The processing unit 610 may be an actual or virtual processor and can execute various processes according to the programs stored in the memory 620. In a multiprocessor system, multiple processing units execute computer executable instructions in parallel to improve the parallel processing capability of the electronic device 600.

[00104] The electronic device 600 typically includes a variety of computer storage medium. Such medium may be any available medium that is accessible to the electronic device 600, including but not limited to volatile and non-volatile medium, removable and non-removable medium. The memory 620 may be volatile memory (for example, a register, cache, a random

access memory (RAM)), a non-volatile memory (for example, a read-only memory (ROM), an electrically erasable programmable read-only memory (EEPROM), a flash memory) or any combination thereof. The storage device 630 may be any removable or non-removable medium, and may include a machine-readable medium, such as a flash drive, a disk, or any other medium, which can be used to store information and/or data (such as training data for training) and can be accessed within the electronic device 600.

[00105] The electronic device 600 may further include additional removable/non-removable, volatile/non-volatile, transitory/non-transitory storage medium. Although not shown in FIG. 6, a disk driver for reading from or writing to a removable, non-volatile disk (such as a "floppy disk"), and an optical disk driver for reading from or writing to a removable, non-volatile optical disk can be provided. In these cases, each driver may be connected to the bus (not shown) by one or more data medium interfaces. The memory 620 may include a computer program product 625, which has one or more program modules configured to perform various methods or acts of various embodiments of the present disclosure.

[00106] The communication unit 640 communicates with a further computing device through the communication medium. In addition, functions of components in the electronic device 600 may be implemented by a single computing cluster or multiple computing machines, which can communicate through a communication connection. Therefore, the electronic device 600 may be operated in a networking environment using a logical connection with one or more other servers, a network personal computer (PC), or another network node.

[00107] The input device 650 may be one or more input devices, such as a mouse, a keyboard, a trackball, etc. The output device 660 may be one or more output devices, such as a display, a speaker, a printer, etc. The electronic device 600 may also communicate with one or more external devices (not shown) through the communication unit 640 as required. The external device, such as a storage device, a display device, etc., communicate with one or more devices that enable users to interact with the electronic device 600, or communicate with any device (for example, a network card, a modem, etc.) that makes the electronic device 600 communicate with one or more other computing devices. Such communication may be executed via an input/output (I/O) interface (not shown).

[00108] According to example implementation of the present disclosure, a computer-readable storage medium is provided, on which a computer-executable instruction or computer program is stored, where the computer-executable instructions or the computer program is executed by the processor to implement the method described above. According to example implementation of the present disclosure, a computer program product is also provided. The computer program product is physically stored on a non-transient computer-readable medium and includes computer-executable instructions, which are executed by the processor to implement the method described above.

[00109] Various aspects of the present disclosure are described herein with reference to the flow chart and/or the block diagram of the method, the device, the equipment and the computer program product implemented in accordance with the present disclosure. It would be appreciated that each block of the flowchart and/or the block diagram and the combination of each block in the flowchart and/or the block diagram may be implemented by computer-readable program instructions.

[00110] These computer-readable program instructions may be provided to the processing units of general-purpose computers, special computers or other programmable data processing devices to produce a machine that generates a device to implement the functions/acts specified in one or more blocks in the flow chart and/or the block diagram when these instructions are executed through the processing units of the computer or other programmable data processing devices. These computer-readable program instructions may also be stored in a computer-readable storage medium. These instructions enable a computer, a programmable data processing device and/or other devices to work in a specific way. Therefore, the computer-readable medium containing the instructions includes a product, which includes instructions to implement various aspects of the functions/acts specified in one or more blocks in the flowchart and/or the block diagram.

[00111] The computer-readable program instructions may be loaded onto a computer, other programmable data processing apparatus, or other devices, so that a series of operational steps can be performed on a computer, other programmable data processing apparatus, or other devices, to generate a computer-implemented process, such that the instructions which execute on a

computer, other programmable data processing apparatus, or other devices implement the functions/acts specified in one or more blocks in the flowchart and/or the block diagram.

[00112] The flowchart and the block diagram in the drawings show the possible architecture, functions and operations of the system, the method and the computer program product implemented in accordance with the present disclosure. In this regard, each block in the flowchart or the block diagram may represent a part of a module, a program segment or instructions, which contains one or more executable instructions for implementing the specified logic function. In some alternative implementations, the functions marked in the block may also occur in a different order from those marked in the drawings. For example, two consecutive blocks may actually be executed in parallel, and sometimes can also be executed in a reverse order, depending on the function involved. It should also be noted that each block in the block diagram and/or the flowchart, and combinations of blocks in the block diagram and/or the flowchart, may be implemented by a dedicated hardware-based system that performs the specified functions or acts, or by the combination of dedicated hardware and computer instructions.

[00113] Each implementation of the present disclosure has been described above. The above description is example, not exhaustive, and is not limited to the disclosed implementations. Without departing from the scope and spirit of the described implementations, many modifications and changes are obvious to ordinary skill in the art. The selection of terms used in this article aims to best explain the principles, practical application or improvement of technology in the market of each implementation, or to enable other ordinary skill in the art to understand the various embodiments disclosed herein.

CLAIMS

What is claimed is:

1. A method for speech synthesis, comprising:
 - obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech;
 - extracting, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech;
 - extracting, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech;
 - extracting, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and
 - generating, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.
2. The method of claim 1, wherein generating the target speech comprises:
 - performing an attention operation on the timbre feature information with the content feature information, to determine relevant timbre feature information for the target text; and
 - generating the target speech based on the relevant timbre feature information, the prosody feature information, and the content feature information.
3. The method of claim 1, wherein the at least one first reference speech comprises a plurality of first speeches of a speaker, and wherein extracting the timbre feature information from the at least one first reference speech comprises:
 - concatenating the plurality of first speeches to obtain a concatenated speech; and
 - extracting, using the timbre encoder, the timbre feature information from the concatenated speech.
4. The method of claim 1, wherein generating the target speech comprises:

generating, using a language model, extended prosody feature information based on the extracted prosody feature information and the content feature information, the extended prosody feature information characterizing a prosodic pattern with an extended length; and

determining the target speech based on the timbre feature information, the extended prosody feature information, and the content feature information.

5. The method of claim 4, wherein generating the extended prosody feature information comprises:

determining a speech duration for the target text;

aligning a length of prosody feature information with a length of content feature information according to the speech duration; and

generating the extended prosody feature information based on the aligned prosody feature information and the aligned content feature information.

6. The method of claim 1, wherein the at least one second speech comprises a plurality of second speeches of a plurality of speakers, and wherein extracting the prosody feature information comprises:

generating, using the prosody encoder, respective probability distributions for the plurality of second speeches, the respective probability distributions indicating distributions of respective prosody tokens for the plurality of second speeches;

performing an interpolation operation on the respective probability distributions to obtain an interpolated probability distribution; and

determining respective interpolated prosody tokens from the interpolated probability distribution, to form the prosody feature information.

7. The method of claim 6, wherein performing the interpolation operation on the respective probability distributions to obtain an interpolated probability distribution comprises:

performing the interpolation operation on the respective probability distributions based on a plurality of weights for the plurality of second speeches, a weight in the plurality of weights being in a range of 0 and 1.

8. The method of claim 1, wherein the acoustic autoencoder model is trained by:

partitioning a sample speech corpus for a target speaker into a first sample speech and a second sample speech;

determining, using the acoustic autoencoder model under training, a reconstructed speech for a sample text based on the sample text, the first sample speech, and the second sample speech; and

training the acoustic autoencoder model based on a predetermined training objective, the training objective being configured to reduce or minimize a difference between the second sample speech and the reconstructed speech.

9. The method of claim 8, wherein determining the reconstructed speech comprises:

extracting, using the prosody encoder, sample prosody feature information from the second sample speech;

extracting, using the timbre encoder, sample timbre feature information from the first sample speech;

extracting, using the content encoder, sample content feature information from the sample text; and

generating, using the decoder, the reconstructed speech based on the sample prosody feature information, the sample timbre feature information and the sample content feature information.

10. The method of claim 9, wherein the first sample speech comprises a plurality of speech clips, and wherein extracting the sample timbre feature information from the first sample speech comprises:

concatenating the plurality of speech clips to obtain a concatenated speech; and

extracting, using the timbre encoder, the sample timbre feature information from the concatenated speech.

11. An electronic device, comprising:

at least one processing unit; and

at least one memory coupled to the at least one processing unit and storing instructions executable by the at least one processing unit, the instructions, upon execution by the at least one processing unit, causing the electronic device to perform:

obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech;

extracting, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech;

extracting, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech;

extracting, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and

generating, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.

12. The electronic device of claim 11, wherein generating the target speech comprises:

performing an attention operation on the timbre feature information with the content feature information, to determine relevant timbre feature information for the target text; and

generating the target speech based on the relevant timbre feature information, the prosody feature information, and the content feature information.

13. The electronic device of claim 11, wherein the at least one first reference speech comprises a plurality of first speeches of a speaker, and wherein extracting the timbre feature information from the at least one first reference speech comprises:

concatenating the plurality of first speeches to obtain a concatenated speech; and

extracting, using the timbre encoder, the timbre feature information from the concatenated speech.

14. The electronic device of claim 11, wherein generating the target speech comprises:

generating, using a language model, extended prosody feature information based on the extracted prosody feature information and the content feature information, the extended prosody feature information characterizing a prosodic pattern with an extended length; and

determining the target speech based on the timbre feature information, the extended prosody feature information, and the content feature information.

15. The electronic device of claim 14, wherein generating the extended prosody feature information comprises:

determining a speech duration for the target text;

aligning a length of prosody feature information with a length of content feature information according to the speech duration; and

generating the extended prosody feature information based on the aligned prosody feature information and the aligned content feature information.

16. The electronic device of claim 11, wherein the at least one second speech comprises a plurality of second speeches of a plurality of speakers, and wherein extracting the prosody feature information comprises:

generating, using the prosody encoder, respective probability distributions for the plurality of second speeches, the respective probability distributions indicating distributions of respective prosody tokens for the plurality of second speeches;

performing an interpolation operation on the respective probability distributions to obtain an interpolated probability distribution; and

determining respective interpolated prosody tokens from the interpolated probability distribution, to form the prosody feature information.

17. The electronic device of claim 16, wherein performing the interpolation operation on the respective probability distributions to obtain an interpolated probability distribution comprises:

performing the interpolation operation on the respective probability distributions based on a plurality of weights for the plurality of second speeches, a weight in the plurality of weights being in a range of 0 and 1.

18. The electronic device of claim 11, wherein the acoustic autoencoder model is trained by:

partitioning a sample speech corpus for a target speaker into a first sample speech and a second sample speech;

determining, using the acoustic autoencoder model under training, a reconstructed speech for a sample text based on the sample text, the first sample speech, and the second sample speech; and

training the acoustic autoencoder model based on a predetermined training objective, the training objective being configured to reduce or minimize a difference between the first sample speech and the reconstructed speech.

19. The electronic device of claim 18, wherein determining the reconstructed speech comprises:

extracting, using the prosody encoder, sample prosody feature information from the second sample speech;

extracting, using the timbre encoder, sample timbre feature information from the first sample speech;

extracting, using the content encoder, sample content feature information from the sample text;

generating, using the decoder, the reconstructed speech based on the sample prosody feature information, the relevant sample timbre feature information and the sample content feature information.

20. A computer readable storage medium having computer executable instructions stored thereon, the computer executable instructions, when executed by an electronic device, causing the electronic device perform operations comprising:

obtaining input information for a trained acoustic autoencoder model, the input information comprising a target text, at least one first reference speech, and at least one second reference speech;

extracting, using a timbre encoder in the acoustic autoencoder model, timbre feature information from the at least one first reference speech, the timbre feature information characterizing timbre represented in the at least one first reference speech;

extracting, using a prosody encoder in the acoustic autoencoder model, prosody feature information from the at least one second reference speech, the prosody feature information characterizing a prosodic pattern represented in the at least one second reference speech;

extracting, using a content encoder in the acoustic autoencoder model, content feature information from the target text; and

generating, using a decoder in the acoustic autoencoder model, a target speech corresponding to the target text based on the timbre feature information, the prosody feature information, and the content feature information.