

A TRAINING METHOD, A READABLE STORAGE MEDIUM AND A VOICE CLONING METHOD FOR A VOICE CLONING MODEL

TECHNICAL FIELD

[0001] One or more embodiments of the present application generally relate to the field of voice synthesis technology, and in particular to a training method, a readable storage medium and a voice cloning method for a voice cloning model.

BACKGROUND

[0002] Voice synthesis technology based on deep learning is applied to generate a corresponding piece of audio based on a given short text by using a neural network.

[0003] Currently available voice synthesis methods can be divided into such two categories as concatenative synthesis and parametric synthesis, of which the former refers to recording commonly used characters (words) by recorder and concatenating them based on the corresponding text, and the latter refers to constructing a speech generation model first which will generate a piece of audio (or spectrum) by itself after text inputting without any concatenation. And the voice synthesis technology based on deep learning is deemed to be applied in parametric synthesis.

[0004] At present, the difficulty in voice synthesis technology lies in naturalness of speech which is hard to be controlled in concatenative synthesis. However, parametric synthesis, especially synthesis based on deep learning, is likely to approximate the effect of human pronunciation recently. And for any speech involving more than one language, in particular any speech involving Chinese and English, it is widely feasible to generate Chinese and English parts of a sentence respectively by using a Chinese speech generation model and an English speech generation model and then concatenate such parts, which renders it almost impossible to meet the requirement of naturalness

of speech.

[0005] Based on the commonly used voice synthesis technology, voice synthesis of multiple speakers is further developed. The widely used deep-learning-based voice synthesis technology uses a dedicated model to generate a person's voice, which requires to build a large number of models for any speech involving multiple speakers which will occupy a large amount of storage and computing resources.

[0006] And based on voice synthesis involving multiple speakers, voice cloning is further developed. The voice synthesis involving multiple speakers generally only generate the voice of the speaker who has been involved during the model training period, and cannot generate corresponding voice for new speakers appearing during the use of the product. Voice cloning technology refers to clone the voice of any new speaker in a shorter time and with less data compared with training a new model (otherwise, a new model can be trained for the new speaker). Given that the model, based on any text input, can generate the corresponding audio that is akin to the new speaker's voice, it is necessary for voice cloning technology to ensure the similarity between the cloned voice and the new speaker's voice in addition to the naturalness.

[0007] The biggest challenge for voice cloning technology today lies in how to use less voice data of new speaker to obtain synthesized voice that is more akin to new speaker's voice in a shorter time (such as retraining time).

[0008] At present, adaptive retraining is widely used in the existing voice cloning technology in which adaptive retraining is conducted on the model based on new speaker's voice and fine-tuning the original model. However, this method generally requires dozens to hundreds of sentences for the sample data of new speaker which are hard to be obtained in a short time, and the adaptive method requires additional training time and requires a separate model for each new speaker.

SUMMARY

[0009] In order to solve the above problems, one or more embodiments of the present application provide a training method, a readable storage medium and a voice cloning

method for a voice cloning model.

[0010] Firstly, based on certain aspects of the present application, a training method for a voice cloning model is disclosed which includes:

[0011] Inputting the preprocessed corpus data into and applying the same in the neural network model of voice cloning, wherein the corpus data is used to train the neural network model and includes voice data of a plurality of speakers in a plurality of languages and text data corresponding to the voice data;

[0012] Processing the voice data of each speaker and the corresponding text data using speaker embedding module and encoder module of the neural network model, and thereby generating voice embedded data of the speaker and the corresponding text embedded data of the text data;

[0013] Combining the text embedded data with the voice embedded data as an input parameter of the attention mechanism module of the neural network model, and training the neural network model by using the voice data as a reference standard;

[0014] Obtaining the neural network model applied to the voice cloning if the loss function of the neural network model satisfies a predetermined value.

[0015] In some embodiments, generating the voice embedded data of the speaker further includes:

[0016] Calculating voiceprint information of the speaker in each piece of voice data included in voice data of the speaker by means of the speaker embedding module;

[0017] Calculating the average of all the voiceprint information, and generating the voice embedded data of the speaker.

[0018] In some embodiments, processing the corresponding text data further includes expanding an array for expressing the characters accordingly based on the language of each character in the text data.

[0019] In some embodiments, expanding an array for expressing the characters accordingly based on the language of each character in the text data further includes:

[0020] Distinguishing the language of each character in the text data;

[0021] For Chinese characters, merging null arrays having the same dimension as the array expressing the Chinese characters in the back of the array expressing the Chinese

characters; and

[0022] For English characters, merging null arrays having the same dimension as the array expressing the English characters in front of the array expressing the English characters.

[0023] In some embodiments, combining the text embedded data with the voice embedded data further includes merging the array expressing the voice embedded data with the array of each time stamp expressing the text embedded data.

[0024] In some embodiments, the method further includes pre-processing the corpus data and the pre-processing includes:

[0025] Converting Chinese characters in the text data into pinyin with four tones of pinyin being represented by numbers; and

[0026] Converting English characters in the text data into Unicode non-English letters.

[0027] In some embodiments, the method further includes pre-processing the corpus data and the pre-processing includes converting the voice data into a Mel spectrum after removing silent tones in the voice data.

[0028] Secondly, the embodiments of the present application provide a voice cloning method for cloning voice of target speaker which includes:

[0029] Processing voice data of the target speaker and the corresponding text data to generate the processed voice data and text data;

[0030] Inputting the processed voice data and text data into the trained voice cloning model; and

[0031] Obtaining cloned voice of the target speaker by using the voice cloning model;

[0032] Wherein, the voice cloning model is generated using the training method as described in the first part above.

[0033] Thirdly, the embodiments of the present application provide a computer readable medium which stores an instruction which, when executed on a computer, triggers the computer to perform the training method of the voice cloning model as described in the first part above.

[0034] Fourthly, the embodiments of the present application provide a system which includes:

[0035] A memory which is configured to store the instructions executed by one or more processors of the system; and

[0036] A processor which is configured to execute the instructions stored in the memory to further perform the training method of the voice cloning model as described in the first part above.

[0037] The present application, based on certain aspects herein, sets out the effects including but not limited to:

[0038] Only one voice cloning model is needed to support the generation of voice involving multiple speakers, and only one such model is needed to support the generation of voice involving more than one language. The voice cloning model can directly clone voice of the speaker after obtaining the voice and the voice embedded data of the target speaker without any additional training. And it is found after actual tests that the voice cloning model of the present application can produce acceptable voice results by using only 15 to 30 sentences of the target speaker. In contrast, existing models that requires adaptive training requires tens of minutes to hours of preparation time.

BRIEF DESCRIPTION OF ATTACHED DRAWINGS

[0039] FIG. 1 is a structural schematic diagram of the computing device used in the training method of the voice cloning model as described in embodiments of the present invention.

[0040] FIG. 2 is a module schematic diagram of the voice cloning model as described in embodiments of the present application.

[0041] FIG. 3 is a schematic diagram of an example of generating a combined tensor in the voice cloning model as described in embodiments of the present application.

[0042] FIG. 4 is a schematic flow chart of the training method of the voice cloning model as described in embodiments of the present application.

[0043] FIG. 5 is a schematic flow chart of the voice cloning model as described in embodiments of the present application.

DESCRIPTION OF THE EMBODIMENTS

[0044] In order to have a better understanding of the purposes of and the technical solutions applied in the embodiments of the present application, the technical solutions applied in the embodiments of the present application will be clearly and completely described as below with reference to the accompanying drawings. It is apparent that the described embodiments are only a part of but not all the embodiments of the present application. All other embodiments obtained by those of ordinary skill in the art based on the embodiments of the present application without creative efforts fall within the scope of protection of the present application.

[0045] Certain embodiments included in all embodiments of the present invention provide a training method for a voice cloning model. It should be noted that the steps illustrated in the accompanying flow chart may be performed in a computer system such as a set of computer executable instructions. Furthermore, the steps shown or described may be performed in an order different from that shown in some cases even though logical sequences are shown in the accompanying flow chart.

[0046] The embodiments of the present application can be performed in a computing device. FIG. 1 is a structural schematic diagram of the computing device used in the training method of the voice cloning model as described in embodiments of the present invention, in which the computing device 100 includes but is not limited to laptop, desktop, handheld PC, personal digital assistant, engineering workstation, service station, network device, network hub, switch, embedded processor, Digital Signal Processor (DSP), graphics device, video game device, set top box, microcontroller, cellular telephone, portable media player, handheld device, wearable device (such as display glasses or goggles, Head-Mounted Display (HMD), watch, headset, arm band, jewelry, etc.), Virtual Reality (VR) and/or Augment Reality (AR) device, Internet of Things (IoT) device, industrial control device, on-board infotainment device, streaming media client device, e-book reading device, POS machine, control system of electric vehicle and various other electronic devices. In general, a plurality of devices and

electronic devices that can house the processors and/or other execution logics disclosed herein are generally suitable.

[0047] As shown in FIG. 1, the computing device 100 may include one or more (only one shown in FIG. 1) processors 101 (the processor 101 may include but is not limited to a central processing unit CPU, an image processor GPU, a digital signal processor DSP, a microprocessor MCU or a programmable logic device FPGA and other processing devices), buses 102, memories 103 for storing data and communication interfaces 104 for communication. It could be understood by those skilled in the art that the structure shown in FIG. 1 is merely illustrative and does not impose any limitations on the structure of the above electronic device. For example, the computing device 100 may also include more or fewer components than those shown in FIG. 1, or have a configuration different from that shown in FIG. 1.

[0048] The memory 103 can be not only used to store databases as well as various corpus data, including corpus data for model training and corpus data of target speakers for model adjustment, but also used to store software programs and modules of neural network models as well as software programs and modules of other application software, including the program instructions/modules corresponding to the training method of the voice cloning model described in the embodiments of the present invention. The processor 101 executes various functional applications and conducts data processing by running the software programs and modules stored in the memory 103, which is deemed as a training method and a voice cloning method for implementing a voice cloning model.

[0049] The memory 103 may comprise of a high-speed random access memory such as SRAM and DRAM, and may also comprise of a nonvolatile memory such as one or more nonvolatile random access memories (NVRAM), various flash memories like SPI flash and NAND flash, or other nonvolatile solid state memories.

[0050] As shown in some embodiments of the present application, the memory 103 may further comprise of a memory remotely configured relative to the processor 101 which may be connected to the computing device 100 through the communication interface 104 via a network. Examples of the aforesaid network include, but are not limited to,

the Internet, intranets, local area networks, mobile communication networks and any combination thereof.

[0051] The communication interface 104 is configured to receive or send data via a network which may include various connection types such as wired connection, connection via wireless communication link or connection using fiber optic cables. And the specific network examples above may include the Internet provided by the communication provider of the computing device 100.

[0052] FIG. 2 is a module schematic diagram of the voice cloning model as described in one or more embodiments of the present application. As shown in FIG. 2, the voice cloning model 200 includes an encoder 210, a speaker embedding module 220, a merging module 230, an attention mechanism module 240, a decoder 250, etc.. Optionally, the merging module 230, although not shown in FIG. 2, may also be placed inside the encoder 210 as one or more units of the encoder 210.

[0053] Although the voice cloning model 200 can be implemented based on an existing neural network model, and therefore can be implemented in the embodiments of the present application based on the Google Tacotron2 model, the existing Tacotron2 model is only suitable for single-person single-language model training and does not support voice cloning. The present application, based on improvement of the Tacotron2 model, implements the multi-person multi-language voice cloning model 200 of the present application. In the following various embodiments, an explicit description is only made on the improved part of the voice cloning model 200 of the present application but not on the existing architecture of the Tacotron2 model.

[0054] In the illustrated embodiments, the voice cloning model 200 can include a multi-language hybrid input. In order to prevent the characters of different languages from interfering with each other during training, the encoder 210 processes the characters in the text data corresponding to the voice data in a specific manner and based on the involved language, so as to prevent characters of different languages from being trained in the same pronunciation space.

[0055] Specifically, taking a piece of voice data involving both Chinese and English as an example, the encoder 210 determines whether each character in the text data input

into the encoder 210 is a Chinese character or an English character, with the specific determination method detailed as below. Although any array output from the encoder to represent each character may generally be in 128-dimension, 256-dimension or 512-dimension, in the embodiments of the present application, the dimension of any array output from the encoder 210 is doubled for the purpose of completely separating the space between Chinese and English characters and ensuring as far as possible that the voice cloning model 200 shows the same expressiveness on Chinese and English characters. For example, in the case where each character is represented by a 512-dimension array, the encoder 210 expands the dimension of the output array to 1024 dimension. Specifically, if the character involved is in Chinese, a 512-dimension null array will be merged at the back of the 512-dimension array of the character; and if the character involved is in English, a 512-dimension null array will be merged in the front of the 512-dimension array of the character. In this way, Chinese and English characters are allocated to the two subspaces in the space (array) of the output of the encode 210, under which circumstance the dot product of any two Chinese and English embeddings is zero and no interference exists among Chinese and English.

[0056] In different embodiments, multi-language hybrid input is not limited to Chinese and English input but may also include hybrid input of any two different languages of the four major languages; no such limitation is imposed on the present application.

[0057] In some embodiments, the voice cloning model 200 can also include voice input involving a plurality of speakers, and the speaker embedding module 220 can be configured to process voice data of a plurality of speakers and generate voice embedded data of the speakers.

[0058] The speaker embedding module 220 calculates voiceprint information of each speaker and stores the information in a 128-dimension array. Specifically, in the speaker embedding module 220, each piece of voice data of each speaker is inputted into a 3-layer LSTM (Long Short-Term Memory) with a dimension of 768 of each layer, and the output of the LSTM is mapped in a 128-dimension array and is normalized to obtain a piece of voice embedded data of voice data. For each speaker, the final voice embedded data of the speaker is obtained by calculating the average of the voice

embedded data obtained from the speaker.

[0059] After the speaker's voice data and the corresponding text data are respectively processed by the speaker embedding module 220 and the encoder 210, the merging module 230 is configured to merge the array of expressed characters output by the encoder 210 with the voice embedded data of the speaker.

[0060] In some embodiments, the final embedded data of each speaker expressed by the 128-dimension array and the array of expressed characters output by the encoder 210 are merged based on each time stamp. For example, since the shape of the output array of the encoder 210 and the shape of the array of the voice embedded data of the speaker can be expressed as (batch_size, length, encoder_dimension) and (batch_size, 128) respectively, the dimension of the array of the voice embedded data is increased by one dimension to (batch_size, 1, 128), the array is then copied to (batch_size, length, 128) in the increased dimension and finally a new tensor is formed by merging the output of the encoder 210 with the array of the speaker's voice embedded data in the last dimension into (batch_size, length, encoder_dimension + 128).

[0061] FIG. 3 is a schematic diagram of an example of generating a combined tensor in the voice cloning model. According to FIG. 3, one of the speaker Xiaowang's several pieces of voice data corresponds to the text saying "this is apple" which is preprocessed and then output by the encoder 210 as a text embedding array as shown in FIG. 3. In FIG. 3, the part in (.) represents the Chinese embedding of each character (it will contain context information as a result of convolution) and the part in [.] represents the English embedding of each character (it will contain context information as a result of convolution), and the voice information of the speaker Xiaowang, after being pre-processed, generates a 128-dimension array { } which represents the speaker embedding of the speaker Xiaowang through the speaker embedding module 220. And after processing by the merging module 230, the final text embedding tensor in FIG. 3 is formed.

[0062] The new is then passed to the attention mechanism module 240 as an input parameter, and finally generates the target speaker's audio corresponding to the target text through the modified parallel WaveNet after being processed by the attention

mechanism module 240 and the decoder 250.

[0063] According to the embodiments of the present application, only one voice cloning model is needed to support the generation of voice involving multiple speakers as well as voice involving multiple languages, and this voice cloning model can directly clone the target speaker's voice after obtaining the speaker's voice and voice embedded data without requiring additional adaptive training. And it is found after actual tests that the voice cloning model of the present application can produce acceptable voice results by using only 15 to 30 sentences of the target speaker. In contrast, existing models that require adaptive training requires tens of minutes to hours of preparation time.

[0064] One or more embodiments of the training method of the voice cloning model in the present application are described as below. And this method is used to train the previously described voice cloning model.

[0065] FIG. 4 shows an embodiment of the voice cloning model of the present application. In some embodiments, the method 400 is implemented on an electronic device, such as the computing device 100 shown in FIG. 1.

[0066] For the content that is not described in the embodiments of the foregoing computing device and voice cloning model, reference may be made to the embodiments of the following method; similarly, for the content that is not described in the embodiments of the method, reference may be made to the embodiments of the foregoing computing device and voice cloning model.

[0067] In the training method 400 of the voice cloning model shown in FIG. 4, the preprocessed corpus data input is applied to the neural network model of voice cloning at block 401.

[0068] At block 402, the speaker embedding module and the encoder module of the neural network model are used to process voice data and corresponding text data of each speaker respectively, so as to generate each speaker's voice embedded data and the text embedded data of the corresponding text data.

[0069] At block 403, the text embedded data and the voice embedded data are merged as an input parameter of the attention mechanism module of the neural network model, and voice data is used as a reference standard to train the neural network model.

[0070] Finally, at block 404, the neural network model applied to voice cloning is obtained under the circumstance where the loss function of the neural network model satisfies a predetermined value.

[0071] Furthermore, generating the speaker's voice embedded data also includes: calculating voiceprint information of the speaker for each piece of voice data in the voice data of the speaker by using the speaker embedding module; and calculating the average of all voiceprint information to generate voice embedded voice of the speaker.

[0072] Furthermore, processing the corresponding text data also includes expanding the array for expressing characters according to the language of each character included in the text data.

[0073] Furthermore, expanding the array for expressing characters according to the language of each character included in the text data also includes: distinguishing the language of each character in the text data; for Chinese characters, merging null arrays with the same dimension as the arrays expressing Chinese characters at the back of the array expressing Chinese characters, and for English characters, merging null arrays with the same dimension as the arrays expressing English characters in the front of the arrays expressing English characters.

[0074] Furthermore, merging the text embedded data with the voice embedded data also includes merging the array expressing the voice embedded data with the array expressing the text embedded data at each time stamp.

[0075] In some implementations, the corpus data, including voice data involving multiple languages and multiple speakers and the corresponding text data thereof, is used to train the aforesaid neural network model, and should be preprocessed prior to training the voice cloning model.

[0076] For the text data in multiple languages and take Chinese-English corpus data as an example, Chinese characters may be converted into Pinyin to reduce character space to 26 English letters and 4 digits representing tones since a large number of Chinese characters are included and too large character space is required if Chinese characters are used directly without conversion; and 26 English letters can be converted into 26 non-English Unicode special letters to avoid any conflict with Chinese Pinyin. Besides,

expressing Chinese and English characters in different ways as stated above may allow the above-mentioned encoder 210 to easily distinguish Chinese and English characters in the preprocessed text data.

[0077] For any voice data, the audio file contained is converted into a file in wav format, the extra silent tones contained are removed and the wave file is then converted into a Mel spectrum, which is usually stored in npy format, by short-time Fourier transform. The voice data converted into the Mel spectrum, on the one hand, will be inputted into the speaker embedding module 220 to generate language embedding data and on the other hand, will be used as a ground truth for training the neural network model.

[0078] The training is conducted by means of batch training by combining multiple random worked samples. And the model is basically available when the loss function of the voice cloning model satisfies a predetermined value. For example, in practical training, a voice cloning model that is basically available in the embodiments of the present application is obtained when the loss function is below 0.5. And optionally or alternatively, cloned voice outputted with ideal naturalness and similarity can be obtained based on the voice cloning model when the loss function is between 0.3 and 0.4.

[0079] In the embodiments of the present application, for the trained voice cloning model, only one such model is needed to support the generation of voice involving multiple speakers as well as voice involving multiple languages, and this voice cloning model can directly clone the target speaker's voice after obtaining the speaker's voice and voice embedded data without requiring additional adaptive training.

[0080] FIG. 5 shows an embodiment of the application method of the voice cloning model in the present application. In some embodiments, the method 500 is implemented on an electronic device, such as the computing device 100 shown in FIG. 1.

[0081] For the content that is not described in the foregoing embodiments, reference may be made to the embodiments of the following method; similarly, for the content that is not described in the embodiments of the following method, reference may be made to the foregoing embodiments.

[0082] The voice cloning method 500 shown in FIG. 5 is used to clone the voice of the

target speaker, and specifically at block 501, the voice data and corresponding text data of the target speaker are processed to generate processed voice data and text data;

[0083] At block 502, the processed voice data and text data are input into the trained voice cloning model; and

[0084] At block 503, the cloned voice of the target speaker is obtained through the voice cloning model.

[0085] And the voice cloning model includes the neural network model applied to voice cloning as described in the foregoing embodiments.

[0086] According to one or more embodiments of the present application, a computer-readable medium is also disclosed, which stores an instruction which, when executed on a computer, triggers the computer to perform the training method of any foregoing voice cloning model.

[0087] According to one or more embodiments of the present application, a system is also disclosed which comprises of:

[0088] A memory which is configured to store the instructions executed by one or more processors of the system; and

[0089] A processor which is configured to execute the instructions stored in the memory to further perform the training method of any foregoing voice cloning model. All embodiments of each method in the present application can be implemented in software, magnetic part, firmware, etc..

[0090] Program code may be applied to input instructions to perform the functions described herein and generate output information. And the output information may be applied to one or more output devices in a known manner. For the purpose of the present application, processing system includes any system configured with digital signal processor (DSP), microcontroller, application specific integrated circuit (ASIC), microprocessor or other processors.

[0091] Program code can be implemented in a high-level programming language or an object-oriented programming language to communicate with the processing system, and when necessary, may also be implemented in assembly or machine language. In fact, the mechanisms described herein are not limited to the scope of any particular

programming language which, in any case, can be a compiled or interpreted language.

[0092] One or more aspects of at least one embodiment may be implemented based on the representational instructions stored on machine-readable medium, which instruct various logics contained in processor and when read by a machine, trigger the machine to work out the logics for performing the techniques described herein. Those representations called “IP cores” can be stored on a tangible machine-readable medium and be provided to multiple customers or production facilities to be further loaded into a manufacturing machine that actually works out the logics or manufactures the processor.

[0093] The aforesaid machine-readable storage medium may include, but is not limited to, a non-transitory tangible arrangement of the items manufactured or formed by machine or equipment, such as storage medium like hard disks and any other types of disks, including floppy disks, compact disks, compact read-only memory (CD-ROM), compact rewritable disks (CD-RW) and magneto-optical disks; semiconductor devices like read-only memory (ROM), including dynamic random-access memory (DRAM), static random-access memory (SRAM) and other random-access memory (RAM), erasable programmable read-only memory (EPROM), flash memory and electrically erasable programmable read-only memory (EEPROM); phase change memory (PCM); magnetic or optical card; or any other type of medium suitable for storing electronic instructions.

[0094] Therefore, the embodiments of the present application, which are also called program products, also include non-transitory tangible computer-readable medium which contains instructions or design data, such as hardware description language (HDL) which defines the structures, circuits, devices, processors and/or system features described herein.

[0095] In some cases, an instruction converter may be used to convert instructions from a source instruction set to a target instruction set. For example, the instruction converter may convert instructions into one or more other instructions to be processed by cores by means of transformation (e.g. static binary transformation, including dynamic binary transformation or dynamic compilation), change, emulation or other ways. And the

instruction converter may be implemented in software, hardware, firmware or any combination thereof, and may be placed inside the processor, outside the processor, or partially inside the processor and partially outside the processor.

Claims

WHAT IS CLAIMED IS:

1. A training method of voice cloning model which is characterized in and further includes:

Inputting the preprocessed corpus data into and applying the same in the neural network model of voice cloning, wherein the corpus data is used to train the neural network model and includes voice data of a plurality of speakers in a plurality of languages and text data corresponding to the voice data;

Processing the voice data of each speaker and the corresponding text data using speaker embedding module and encoder module of the neural network model, and thereby generating voice embedded data of the speaker and the corresponding text embedded data of the text data;

Combining the text embedded data with the voice embedded data as an input parameter of the attention mechanism module of the neural network model, and training the neural network model by using the voice data as a reference standard;

Obtaining the neural network model applied to the voice cloning if the loss function of the neural network model satisfies a predetermined value.

2. The training method of voice cloning model described in Claim 1 is characterized in that generating the voice embedded data of the speaker further includes:

Calculating voiceprint information of the speaker in each piece of voice data included in voice data of the speaker by means of the speaker embedding module;

Calculating the average of all the voiceprint information, and generating the voice embedded data of the speaker.

3. The training method of voice cloning model described in Claim 1 is characterized in that processing the corresponding text data further includes expanding an array for expressing the characters accordingly based on the language of each character in the text data.

4. The training method of voice cloning model described in Claim 3 is characterized in that expanding an array for expressing the characters accordingly based on the language

of each character in the text data further includes:

Distinguishing the language of each character in the text data;

For Chinese characters, merging null arrays having the same dimension as the array expressing the Chinese characters in the back of the array expressing the Chinese characters; and

For English characters, merging null arrays having the same dimension as the array expressing the English characters in front of the array expressing the English characters.

5. The training method of voice cloning model described in Claim 1 is characterized in that combining the text embedded data with the voice embedded data further includes merging the array expressing the voice embedded data with the array of each time stamp expressing the text embedded data.

6. The training method of voice cloning model described in Claim 1 is characterized in that the method further includes pre-processing the corpus data and the pre-processing includes:

Converting Chinese characters in the text data into pinyin with four tones of pinyin being represented by numbers; and

Converting English characters in the text data into Unicode non-English letters.

7. The training method of voice cloning model described in Claim 1 is characterized in that the method further includes pre-processing the corpus data and the pre-processing includes converting the voice data into a Mel spectrum after removing silent tones in the voice data.

8. A voice cloning method for cloning voice of target speaker which is characterized in and includes:

Processing voice data of the target speaker and the corresponding text data to generate the processed voice data and text data;

Inputting the processed voice data and text data into the trained voice cloning model; and

Obtaining cloned voice of the target speaker by using the voice cloning model;

The voice cloning model is generated using the training method as described in any one of Claims 1 to 7.

9. A computer readable medium which is characterized in that the computer readable storage medium stores an instruction which, when executed on a computer, triggers the computer to perform the training method of the voice cloning model as described in any one of Claims 1 to 7.

10. A system which is characterized in and includes:

A memory which is configured to store the instructions executed by one or more processors of the system; and

A processor which is configured to execute the instructions stored in the memory to further perform the training method of the voice cloning model as described in any one of Claims 1 to 7.