(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2022 年 12 月 8 日 (08.12.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/252904 A1

(51) 国际专利分类号:
G06F 40/126 (2020.01)

(21) 国际申请号: PCT/CN2022/090951

(22) 国际申请日: 2022 年 5 月 5 日 (05.05.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110620109.5 2021年6月3日 (03.06.2021) CN

(71) 申请人: 腾 讯 科 技 (深 圳) 有 限 公 司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(72) 发明人: 郑艺斌 (ZHENG, Yibin); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 李新辉 (LI, Xinhui); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 苏文超 (SU, Wenchao); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。 卢鲤 (LU, Li); 中国广东省深圳市南山区高新区科技中一路腾讯大厦35层, Guangdong 518057 (CN)。

(74) 代理人: 北京派特恩知识产权代理有限公司 (CHINA PAT INTELLECTUAL PROPERTY OFFICE); 中国北京市海淀区海淀南路21号中关村知识产权大厦B座2层, Beijing 100080 (CN)。

(54) **Title:** ARTIFICIAL INTELLIGENCE-BASED AUDIO PROCESSING METHOD AND APPARATUS, DEVICE, STORAGE MEDIUM, AND COMPUTER PROGRAM PRODUCT

(54) **发明名称**: 基于人工智能的音频处理方法、装置、设备、存储介质及计算机程序产品

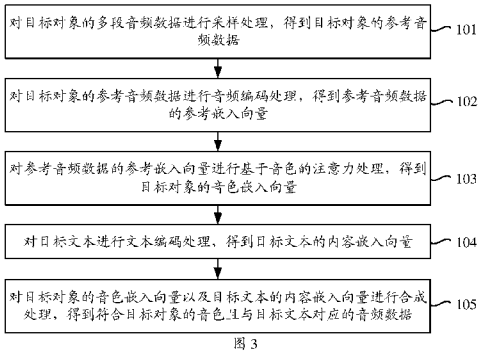

图 3

101 Perform sampling processing on a plurality of segments of audio data of a target object to obtain reference audio data of the target object
102 Perform audio coding processing on the reference audio data of the target object to obtain a reference embedding vector of the reference audio data
103 Perform timbre-based attention processing on the reference embedding vector of the reference audio data to obtain a timbre embedding vector of the target object
104 Perform text coding processing on target text to obtain a content embedding vector of the target text
105 Perform synthesis processing on the timbre embedding vector of the target object and the content embedding vector of the target text to obtain audio data which conforms to the timbre of the target object and corresponds to the target text

(57) **Abstract:** An artificial intelligence-based audio processing method and apparatus, an electronic device, a computer readable storage medium, and a computer program product, which relate to artificial intelligence technology. The method comprises: performing sampling processing on a plurality of segments of audio data of a target object to obtain reference audio data of the target object (101); performing audio coding processing on the reference audio data of the target object to obtain a reference embedding vector of the reference audio data (102); performing timbre-based attention processing on the reference embedding vector of the reference audio data to obtain a timbre embedding vector of the target object (103); performing text coding processing on target text to obtain a content embedding vector of the target text (104); and performing synthesis processing on the timbre embedding vector of the target object and the content embedding vector of the target text to obtain audio data which conforms to the timbre of the target object and corresponds to the target text (105).

(57) **摘要**: 一种基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品；涉及人工智能技术；方法包括：对目标对象的多段音频数据进行采样处理，得到目标对象的参考音频数据（101）；对目标对象的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量（102）；对参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到目标对象的音色嵌入向量（103）；对目标文本进行文本编码处理，得到目标文本的内容嵌入向量（104）；对目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据（105）。

WO 2022/252904 A1

[见续页]

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

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

本国际公布：

— 包括国际检索报告(条约第21条(3))。

基于人工智能的音频处理方法、装置、设备、存储介质及计算机程序产品

相关申请的交叉引用

本申请实施例基于申请号为 202110620109.5、申请日为 2021 年 06 月 03 日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请实施例作为参考。

技术领域

本申请涉及人工智能技术，尤其涉及一种基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品。

背景技术

人工智能（AI，Artificial Intelligence）是计算机科学的一个综合技术，通过研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科，涉及领域广泛，例如自然语言处理技术以及机器学习/深度学习等几大方向，随着技术的发展，人工智能技术将在更多的领域得到应用，并发挥越来越重要的价值。

相关技术中对于音频的合成方式比较粗糙，通常是直接对目标对象的音频数据进行特征提取，并基于提取到的目标对象的嵌入向量进行合成，以得到合成的音频数据，这种合成方式由于直接从音频数据中提取目标对象的嵌入向量，避免不了引入音频数据中与内容相关的信息（例如韵律、风格等），导致待合成的目标文本和音频数据的内容不一致时，音频合成效果不稳定，无法实现音频的精准合成，从而浪费大量的计算资源，并影响用户体验正常的音频合成。

发明内容

本申请实施例提供一种基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品，能够提高音频合成的准确性。

本申请实施例的技术方案是这样实现的：

本申请实施例提供一种基于人工智能的音频处理方法，包括：

对目标对象的多段音频数据进行采样处理，得到所述目标对象的参考音频数据；

对所述目标对象的参考音频数据进行音频编码处理，得到所述参考音频数据的参考嵌入向量；

对所述参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到所述目标对象的音色嵌入向量；

对目标文本进行文本编码处理，得到所述目标文本的内容嵌入向量；

基于所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行合成处理，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

本申请实施例提供一种基于人工智能的音频处理装置，包括：

采样模块，配置为对目标对象的多段音频数据进行采样处理，得到所述目标对象的参考音频数据；

音频编码模块，配置为对所述目标对象的参考音频数据进行音频编码处理，得到所

述参考音频数据的参考嵌入向量；

注意力模块，配置为对所述参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到所述目标对象的音色嵌入向量；

文本编码模块，配置为对目标文本进行文本编码处理，得到所述目标文本的内容嵌入向量；

合成模块，配置为基于所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行合成处理，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

本申请实施例提供一种用于音频处理的电子设备，所述电子设备包括：

存储器，用于存储可执行指令；

处理器，用于执行所述存储器中存储的可执行指令时，实现本申请实施例提供的基于人工智能的音频处理方法。

本申请实施例提供一种计算机可读存储介质，存储有可执行指令，用于引起处理器执行时，实现本申请实施例提供的基于人工智能的音频处理方法。

本申请实施例提供了一种计算机程序产品，包括计算机程序或指令，该计算机程序或指令使得计算机执行上述基于人工智能的音频处理方法。

通过对多段音频数据采样得到的目标对象的参考音频数据进行注意力处理，从而多样化表示音色嵌入向量，以提高音色嵌入向量提取的鲁棒性，并结合目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行音频合成处理，与相关技术直接从特定的音频数据中提取目标对象的嵌入向量进行音频合成不同，本申请实施例通过多样化的音色嵌入向量进行音频合成，能够避免引入音频数据中与内容相关的信息，从而提高音频合成的稳定性，进而基于准确的音色嵌入向量实现精准地个性化音频生成，以节约相关的计算资源。

附图说明

图 1 是本申请实施例提供的音频处理系统的应用场景示意图；

图 2 是本申请实施例提供的用于音频处理的电子设备的结构示意图；

图 3-图 5 是本申请实施例提供的基于人工智能的音频处理方法的流程示意图；

图 6 是本申请实施例提供的编码器的结构示意图；

图 7 是本申请实施例提供的级联卷积层的结构示意图；

图 8 是本申请实施例提供的嵌入空间的示意图；

图 9 是本申请实施例提供的训练流程图；

图 10 是本申请实施例提供的一种快速有效的语音合成定制模型结构的结构示意图；

图 11 是本申请实施例提供的编码器的具体结构框图。

具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，所描述的实施例不应视为对本申请的限制，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

在以下的描述中，所涉及的术语“第一\第二”仅仅是是区别类似的对象，不代表针对对象的特定排序，可以理解地，“第一\第二”在允许的情况下可以互换特定的顺序或先后次序，以使这里描述的本申请实施例能够以除了在这里图示或描述的以外的顺序实施。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技

术人员通常理解的含义相同。本文中所使用的术语只是为了描述本申请实施例的目的，不是旨在限制本申请。

对本申请实施例进行进一步详细说明之前，对本申请实施例中涉及的名词和术语进行说明，本申请实施例中涉及的名词和术语适用于如下的解释。

1）卷积神经网络（CNN，Convolutional Neural Networks）：一类包含卷积计算且具有深度结构的前馈神经网络（FNN，Feedforward Neural Networks），是深度学习（Deep Learning）的代表算法之一。卷积神经网络具有表征学习（Representation learning）能力，能够按其阶层结构对输入图像进行平移不变分类（Shift-invariant Classification）。

2）循环神经网络（RNN，Recurrent Neural Network）：一类以序列（Sequence）数据为输入，在序列的演进方向进行递归（Recursion）且所有节点（循环单元）按链式连接的递归神经网络（Recursive Neural Network）。循环神经网络具有记忆性、参数共享并且图灵完备（Turing Completeness），因此在对序列的非线性特征进行学习时具有一定优势。

3）音素：语音中最小的基本单位，音素是人类能区别一个单词和另一个单词的基础。音素构成音节，音节又构成不同的词和短语。

4）音色：不同音频表现在波形方面总是有与众不同的特性，不同的物体振动都有不同的特点。不同的发声体由于其材料、结构不同，则发出音频的音色也不同。例如钢琴、小提琴和人发出的声音不一样，每一个人发出的音频也不一样，即音色可以理解为音频的特征。

5）目标对象：真实世界中的真实对象或者虚拟场景中的虚拟对象，例如真实用户、虚拟人物、虚拟动物、动漫人物等。

近几年来，随着计算能力的大规模提升，深度学习技术得到了大规模的研究与运用，进一步推动了语音合成技术的发展。开始涌现出端到端的语音合成声学建模方法。该方法直接从输入的字符或者音素序列上预测对应的声学特征序列，在学术界和工业界上都获得了较为广泛的运用。然而训练一个这样的商用语音合成系统一般都需要数十个小时的数据量，这样的数据要求在许多运用场景上都不太可能。因此，基于少量数据的语音合成定制技术的需求日益迫切。

语音合成定制声学建模方法可以分为两大类：第一种方法，首先在多个说话人（对象）的语料上预训练多说话人模型（也称为平均模型），然后在平均模型的基础上利用说话人少量的数据进行自适应训练；第二种方法，直接从目标说话人的音频中预测说话人嵌入向量，然后将该嵌入向量直接输入到平均模型，没有经过任何模型的微调训练。

然而，相关技术存在以下问题：在说话人嵌入空间建模的采用独热编码对说话人进行表示，该表示信息仅仅能将不同的说话人区分开，但并不含有说话人音色相关的信息，另外一种直接从音频中提取嵌入空间表示的方法，由于是直接从与文本相配对的音频中提取说话人嵌入信息，避免不了引入音频内容相关的信息（比如韵律、风格等），导致待合成的文本和参考音频的内容不一致时合成效果不稳定。

为了解决上述问题，本申请实施例提供了一种基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品，能够提高音频合成的稳定性。

本申请实施例所提供的基于人工智能的音频处理方法，可以由终端/服务器独自实现；也可以由终端和服务器协同实现，例如终端独自承担下文所述的基于人工智能的音频处理方法，或者，终端向服务器发送针对音频的生成请求（包括目标对象以及目标文本），服务器根据接收的针对音频的生成请求执行基于人工智能的音频处理方法，响应于针对音频的生成请求，基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据，从而实现音频的智能化

地精准生成。

本申请实施例提供的用于音频处理的电子设备可以是各种类型的终端设备或服务器，其中，服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云计算服务的云服务器；终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。

以服务器为例，例如可以是部署在云端的服务器集群，向用户开放人工智能云服务（AIaaS，AI as a Service），AIaaS 平台会把几类常见的 AI 服务进行拆分，并在云端提供独立或者打包的服务，这种服务模式类似于一个 AI 主题商城，所有的用户都可以通过应用程序编程接口的方式来接入使用 AIaaS 平台提供的一种或者多种人工智能服务。

例如，其中的一种人工智能云服务可以为音频处理服务，即云端的服务器封装有本申请实施例提供的音频处理的程序。用户通过终端（运行有客户端，例如音响客户端、车载客户端等）调用云服务中的音频处理服务，以使部署在云端的服务器调用封装的音频处理的程序，对目标对象的音频数据进行采样、注意力处理，得到目标对象的音色嵌入向量，并基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据，从而实现音频的智能化地精准生成。

作为一个应用示例，对于音响客户端，目标对象可以是某广播平台的广播员，需要向社区的住户定期广播注意事项、生活小知识等。例如，广播员在音响客户端输入一段目标文本，该文本需要转化为音频，以向社区的住户广播，基于广播员的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合广播员的音色且与目标文本对应的音频数据，以向住户广播生成的音频。

作为另一个应用示例，对于车载客户端，当用户在开车时，不方便通过文本的形式了解信息，但是可以通过读取音频的方式了解信息，避免遗漏重要的信息。例如，用户在开车时，会议主持者向参会用户发送一段重要会议的文本，需要参会用户及时读取并处理该文本，则车载客户端接收到该文本后，需要将该文本转化为音频，以向该参会用户播放该音频，基于会议主持者的音色嵌入向量以及文本的内容嵌入向量进行合成处理，得到符合会议主持者的音色且与文本对应的音频数据，以向参会用户播放生成的音频，使得参会用户可以及时读取到会议主持者的音频。

参见图 1，图 1 是本申请实施例提供的音频处理系统 10 的应用场景示意图，终端 200 通过网络 300 连接服务器 100，网络 300 可以是广域网或者局域网，又或者是二者的组合。

终端 200（运行有客户端，例如音响客户端、车载客户端等）可以被用来获取针对音频的生成请求，例如，用户通过终端 200 输入目标对象以及目标文本，则终端 200 自动获取目标对象的多段音频数据以及目标文本，并自动生成针对音频的生成请求。

在一些实施例中，终端中运行的客户端中可以植入有音频处理插件，用以在客户端本地实现基于人工智能的音频处理方法。例如，终端 200 获取针对音频的生成请求（包括目标对象以及目标文本）后，调用音频处理插件，以实现基于人工智能的音频处理方法，对目标对象的音频数据进行采样、注意力处理，得到目标对象的音色嵌入向量，并基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据，从而实现音频的智能化地精准生成，例如，对于录音应用，用户在非录音室场景下，无法进行高质量的个性化声音定制，则在录音客户端中输入一段需要录制的文本，该文本需要转化为个性化的音频，基于个性化的音色嵌入向量以及文本的内容嵌入向量进行合成处理，从而基于准确的音色嵌入向量生成

精准的个性化音频，以实现非录音室场景下的个性化声音定制。

在一些实施例中，终端 200 获取针对音频的生成请求后，调用服务器 100 的音频处理接口（可以提供为云服务的形式，即音频处理服务），服务器 100 对目标对象的音频数据进行采样、注意力处理，得到目标对象的音色嵌入向量，并基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据，并将音频数据发送至终端 200，例如，对于录音应用，用户在非录音室场景下，无法进行高质量的个性化声音定制，则在终端 200 中输入一段需要录制的文本，并自动生成针对音频的生成请求，并将针对音频的生成请求发送至服务器 100，服务器 100 基于个性化的音色嵌入向量以及文本的内容嵌入向量进行合成处理，从而基于准确的音色嵌入向量生成精准的个性化音频，并将生成的个性化音频发送至终端 200，以响应针对音频的生成请求，实现非录音室场景下的个性化声音定制。

下面说明本申请实施例提供的用于音频处理的电子设备的结构，参见图 2，图 2 是本申请实施例提供的用于音频处理的电子设备 500 的结构示意图，以电子设备 500 是服务器为例说明，图 2 所示的用于音频处理的电子设备 500 包括：至少一个处理器 510、存储器 550、至少一个网络接口 520 和用户接口 530。电子设备 500 中的各个组件通过总线系统 540 耦合在一起。可理解，总线系统 540 用于实现这些组件之间的连接通信。总线系统 540 除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见，在图 2 中将各种总线都标为总线系统 540。

处理器 510 可以是一种集成电路芯片，具有信号的处理能力，例如通用处理器、数字信号处理器（DSP，Digital Signal Processor），或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等，其中，通用处理器可以是微处理器或者任何常规的处理器等。

存储器 550 包括易失性存储器或非易失性存储器，也可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（ROM，Read Only Memory），易失性存储器可以是随机存取存储器（RAM，Random Access Memory）。本申请实施例描述的存储器 550 旨在包括任意适合类型的存储器。存储器 550 可选地包括在物理位置上远离处理器 510 的一个或多个存储设备。

在一些实施例中，存储器 550 能够存储数据以支持各种操作，这些数据的示例包括程序、模块和数据结构或者其子集或超集，下面示例性说明。

操作系统 551，包括用于处理各种基本系统服务和执行硬件相关任务的系统程序，例如框架层、核心库层、驱动层等，用于实现各种基础业务以及处理基于硬件的任务；

网络通信模块 552，用于经由一个或多个（有线或无线）网络接口 520 到达其他计算设备，示例性的网络接口 520 包括：蓝牙、无线相容性认证（WiFi）、和通用串行总线（USB，Universal Serial Bus）等；

在一些实施例中，本申请实施例提供的音频处理装置可以采用软件方式实现，例如，可以是上文所述的终端中的音频处理插件，可以是上文所述的服务器中音频处理服务。当然，不局限于此，本申请实施例提供的音频处理装置可以提供为各种软件实施例，包括应用程序、软件、软件模块、脚本或代码在内的各种形式。

图 2 示出了存储在存储器 550 中的音频处理装置 555，其可以是程序和插件等形式的软件，例如音频处理插件，并包括一系列的模块，包括采样模块 5551、音频编码模块 5552、注意力模块 5553、文本编码模块 5554、合成模块 5555 以及训练模块 5556；其中，采样模块 5551、音频编码模块 5552、注意力模块 5553、文本编码模块 5554、合成模块 5555 用于实现本申请实施例提供的音频处理功能，训练模块 5556 用于训练神经网络模型，其中，音频处理方法是通过调用神经网络模型实现的。

如前所述，本申请实施例提供的基于人工智能的音频处理方法可以由各种类型的电子设备实施。参见图 3，图 3 是本申请实施例提供的基于人工智能的音频处理方法的流程示意图，结合图 3 示出的步骤进行说明。

在步骤 101 中，对目标对象的多个音频数据进行采样处理，得到目标对象的参考音频数据。

作为获取目标对象的示例，用户通过终端输入目标对象以及目标文本，则终端自动生成针对音频的生成请求（其中包括目标对象的标识以及目标文本），并将针对音频的生成请求发送至服务器，服务器解析针对音频的生成请求，得到目标对象的标识，并基于目标对象的标识从数据库中获取目标对象的多段音频数据（即音频片段），对多段音频数据进行随机采样，将采样得到的音频数据作为目标对象的参考音频数据（参考音频数据为从多段长度不同的音频数据中采样得到的一段不定长的音频数据，用于辅助表征目标对象的多段音频数据），通过采样处理提高目标对象的参考音频数据的多样性，避免局限于一个特定的音频数据，从而通过随机采样保证后续得到的参考嵌入向量与音频数据的内容无关，从而避免待合成的目标文本和音频数据的内容不一致时，导致音频合成效果不稳定的问题。

例如，任一目标对象（即说话人 m），从目标对象对应的语料（包括多段不同长度的音频数据）中随机采样出一段音频数据作为参考音频数据$\mathrm{y}^{r}=\mathrm{y}^{Random(N)}$，其中，Random(N)表示[1, N]中的任意正整数，N 表示目标对象的音频数据对应的最大文本数，例如限定语料中音频数据对应的文本序列的最大长度为 256，即 N 为 256。

在步骤 102 中，对目标对象的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量。

例如，将目标对象的参考音频数据作为编码器的输入，通过编码器对随机采样得到的参考音频数据（变长的音频序列）进行音频编码处理，得到参考音频数据的参考嵌入向量（即参考音频数据的嵌入向量），以便后续基于参考嵌入向量进行注意力处理，以构建一个更加鲁棒、准确的对象嵌入空间，以便后续提高生成音频的自然度（相似度主观平均意见值（MOS，Mean Opinion Score）中通过音频的连贯性、韵律感等进行自然度评测，即连贯性韵、律感越好，则自然度越好）和生成的音频的音色与目标对象的音色的适配度。

需要说明的是，音频编码处理是通过神经网络中的编码器对音频进行压缩实现的，以将参考音频数据（模拟信号）压缩转化为参考嵌入向量（数字信号）。其中，本申请实施例并不局限编码器的模型结构，例如编码器可以是卷积神经网络、循环神经网络、深度神经网络等。

在一些实施例中，对目标对象的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量，包括：对目标对象的参考音频数据进行卷积处理，得到参考音频数据的卷积嵌入向量；对参考音频数据的卷积嵌入向量进行编码处理，得到参考音频数据的参考嵌入向量。

如图 6 所示，通过编码器中级联的卷积层对目标对象的参考音频数据进行卷积处理，得到参考音频数据的卷积嵌入向量，然后通过编码器中的循环神经网络对参考音频数据的卷积嵌入向量进行编码处理，得到参考音频数据的参考嵌入向量。从而，相对于仅通过卷积处理，通过卷积处理以及编码处理这两种处理方式，能够更加精确地提取参考音频数据的嵌入向量，以便后续基于精确的嵌入向量进行音频合成，提高音频合成的准确性。

其中，对参考音频数据的卷积嵌入向量进行向量空间转换，以实现对卷积嵌入向量的编码处理，即将卷积嵌入向量（一种向量空间的 K 维向量，K 为大于 1 的正整数）转

换为参考嵌入向量（另一种向量空间的H维向量，H为大于1的正整数）。

在一些实施例中，音频编码是通过编码器实现的，编码器包括多个级联的卷积层；对目标对象的参考音频数据进行卷积处理，得到参考音频数据的卷积嵌入向量，包括：通过多个级联的卷积层中的第一个卷积层，对目标对象的参考音频数据进行卷积处理；将第一个卷积层的卷积结果输出到后续级联的卷积层，通过后续级联的卷积层继续进行卷积处理以及卷积结果输出，直至输出到最后一个卷积层，并将最后一个卷积层输出的卷积结果作为参考音频数据的卷积嵌入向量。

如图7所示，编码器包括J个级联的卷积层，第1个卷积层对参考音频数据进行卷积编码，并将卷积结果输出到第2个卷积层，第2个卷积层继续进行卷积编码和卷积结果输出，直至输出到第J个卷积层，通过第J个卷积层对第J-1个卷积层输出的卷积结果进行卷积编码，得到参考音频数据的卷积嵌入向量，其中，J为多个级联的卷积层的总数，J为大于1的正整数。从而，相对于单个卷积层，通过级联的卷积处理，提高卷积结果的精度，以便基于精确的卷积嵌入向量进行后续编码处理，提高参考音频数据的嵌入向量的精度。

在一些实施例中，音频编码是通过编码器实现的，编码器包括循环神经网络；对参考音频数据的卷积嵌入向量进行编码处理，得到参考音频数据的参考嵌入向量，包括：对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行向量更新处理，得到参考音频数据的更新信息；对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行向量重置处理，得到参考音频数据的重置信息；基于参考音频数据的更新信息以及参考音频数据的重置信息，对参考音频数据的卷积嵌入向量进行上下文编码处理，得到参考音频数据的参考嵌入向量。

例如，循环神经网络能够解决长依赖的问题，循环神经网络包括两个门：分别是更新门和重置门。更新门用于控制前一状态的隐藏向量被带入到当前状态中的程度，更新门的值越大说明前一状态的带入的隐藏向量越多；重置门控制前一状态有多少信息被写入到当前状态的候选集上，重置门的值越小，前一状态的信息被写入的越少。通过循环神经网络中的更新门，结合循环神经网络的隐藏向量（表示循环神经网络所自带的可学习向量）以及参考音频数据的卷积嵌入向量进行向量更新处理，得到参考音频数据的更新信息z_t，通过循环神经网络中的重置门，对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行向量重置处理，得到参考音频数据的重置信息r_t；基于参考音频数据的更新信息以及参考音频数据的重置信息，对参考音频数据的卷积嵌入向量进行上下文编码处理，得到参考音频数据的参考嵌入向量y_t。从而，通过循环神经网络解决参考音频数据的长依赖问题，提高考音频数据的嵌入向量的鲁棒性。

在一些实施例中，对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行向量更新处理，得到参考音频数据的更新信息，包括：对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行拼接处理，得到第一拼接向量；对第一拼接向量进行基于更新门的映射处理，得到参考音频数据的更新信息。

例如，向量更新处理的过程如下公式所示：$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]$，其中，σ表示激活函数，$W_z$表示更新门的可学习参数，$h_{t-1}$表示循环神经网络的隐藏向量，$x_t$表示参考音频数据的卷积嵌入向量，[.]表示拼接处理，z_t表示参考音频数据的更新信息，$[h_{t-1}, x_t]$表示第一拼接向量，通过激活函数对W_z点乘$[h_{t-1}, x_t]$的结果进行激活处理，以实现基于更新门的映射处理。

在一些实施例中，对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行向量重置处理，得到参考音频数据的重置信息，包括：对循环神经网络的隐藏向量以及参考音频数据的卷积嵌入向量进行拼接处理，得到第二拼接向量；对第二拼接向量进

行基于重置门的映射处理，得到参考音频数据的重置信息。

例如，向量重置处理的过程如下公式所示：$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]$，其中，σ表示激活函数，$W_r$表示重置门的可学习参数，$h_{t-1}$表示循环神经网络的隐藏向量，$x_t$表示参考音频数据的卷积嵌入向量，[.]表示拼接处理，r_t表示参考音频数据的重置信息，$[h_{t-1}, x_t]$表示第二拼接向量，通过激活函数对W_r乘$[h_{t-1}, x_t]$的结果进行激活处理，以实现基于重置门的映射处理。

在一些实施例中，基于参考音频数据的更新信息以及参考音频数据的重置信息，对参考音频数据的卷积嵌入向量进行上下文编码处理，得到参考音频数据的参考嵌入向量，包括：基于目标对象的重置信息、循环神经网络的隐藏向量以及目标对象的卷积嵌入向量进行基于候选向量的映射处理，得到参考音频数据的候选嵌入向量；对参考音频数据的更新信息、循环神经网络的隐藏向量以及参考音频数据的候选嵌入向量进行向量映射处理，得到参考音频数据的参考嵌入向量。

例如，基于候选向量的映射处理的过程如下公式所示：$\widetilde{h_t} = tanh\left(W_{\widetilde{h_t}} \cdot [r_t * h_{t-1}, x_t]\right)$，其中，$W_{\widetilde{h_t}}$表示可学习参数，$h_{t-1}$表示循环神经网络的隐藏向量，$x_t$表示参考音频数据的卷积嵌入向量，[.]表示拼接处理，*表示矩阵的乘积，r_t表示参考音频数据的重置信息，$\widetilde{h_t}$表示参考音频数据的候选嵌入向量。

例如，向量映射处理的过程如下公式所示：$y_t = \sigma\left(W_o \cdot \left((1 - z_t) * h_{t-1} + z_t * \widetilde{h_t}\right)\right)$，其中，$W_o$表示可学习参数，σ表示激活函数，$h_{t-1}$表示循环神经网络的隐藏向量，$z_t$表示参考音频数据的更新信息，$y_t$表示参考音频数据的参考嵌入向量。

在步骤 103 中，对参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到目标对象的音色嵌入向量。

例如，在得到参考音频数据的参考嵌入向量后，基于参考嵌入向量进行注意力处理，得到目标对象的音色嵌入向量，其注意力处理能够构建一个更加鲁棒、准确的对象嵌入空间，使得音色嵌入向量与音频的内容分离开，音色嵌入向量可以准确表征目标对象的音色，从而提高了生成音频的自然度和与目标对象的相似度。

需要说明的是，注意力处理是通过注意力机制实现的。在认知科学中，注意力机制（Attention Mechanism）用于选择性地关注所有信息的一部分，同时忽略其他信息。注意力机制可以使得神经网络具备专注于部分输入的能力，即选择特定的输入。在计算能力有限情况下，注意力机制是解决信息超载问题的主要手段的一种资源分配方案，将计算资源分配给更重要的任务。其中，本申请实施例并不局限于注意力机制的形式，例如注意力机制可以是多头注意力、键值对注意力、结构化注意力等。

参见图 4，图 4 是本申请实施例提供的基于人工智能的音频处理方法的流程示意图，图 4 示出图 3 的步骤 103 可以通过步骤 1031-步骤 1032 实现：在步骤 1031 中，对参考音频数据的参考嵌入向量进行基于多个对象音色的嵌入空间的映射处理，得到多个对象音色的权重；在步骤 1032 中，基于多个对象音色的权重，对多个对象音色的嵌入向量进行加权求和处理，得到目标对象的音色嵌入向量。

如图 8 所示，当基于多个对象音色的嵌入空间存在 4 个对象音色，分别为 A、B、C、D，将参考音频数据的参考嵌入向量映射到包括多个对象音色的嵌入空间中（即对参考音频数据的参考嵌入向量进行基于多个对象音色的嵌入空间的映射处理），得到对象音色 A 的权重（0.3）、对象音色 B 的权重（0.4）、对象音色 C 的权重（0.1）、对象音色 D 的权重（0.2），基于 4 个对象音色的权重，对 4 个对象音色的嵌入向量进行加权求和处理，得到目标对象的音色嵌入向量（即表征目标对象的音色的嵌入向量）。从而，通过将参考音频数据的参考嵌入向量映射到包括多个对象音色的嵌入空间中，准确标识目标对象的音色，以便后续合成符合目标对象的音色的音频数据，提高合成音频的准确

性。

在步骤 104 中，对目标文本进行文本编码处理，得到目标文本的内容嵌入向量。

作为获取目标文本的示例，用户通过终端输入目标对象以及目标文本，则终端自动生成针对音频的生成请求，并将针对音频的生成请求发送至服务器，服务器解析针对音频的生成请求，得到目标文本，并通过文本编码器对目标文本进行文本编码，得到目标文本的内容嵌入向量（即表征目标文本的文本内容的嵌入向量），以便后续结合音色嵌入向量进行音频合成，以实现个性化音频定制。

需要说明的是，文本编码处理是通过神经网络中的文本编码器对文本进行压缩实现的，以将目标文本（模拟信号）压缩转化为内容嵌入向量（数字信号）。其中，本申请实施例并不局限文本编码器的模型结构，例如编码器可以是卷积神经网络、循环神经网络、深度神经网络等。

在步骤 105 中，对目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据。

例如，由于音色嵌入向量能够精准地表征目标对象的音色，因此基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行音频合成，能够得到符合目标对象的音色且与目标文本对应的音频数据，该合成的音频数据与目标对象真实的音频数据相似，使得合成的音频数据更加逼真，相对于相关技术中无法实现音频的精准合成，能够提高音频合成的效率，进而节约相关的计算资源。

在一些实施例中，对目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据，包括：对目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行对齐处理，得到目标文本的对齐向量；对目标文本的对齐向量进行解码处理，得到目标文本的声学特征序列；对声学特征序列进行声学特征转换，得到符合目标对象的音色且与目标文本对应的音频数据。

例如，对齐处理用于结合目标对象的音色嵌入向量，计算目标文本中的每一个字符被选择的概率，对齐向量的值为输入的目标文本中的每一个字符被选择的概率（或关注的程度），表征了输入的目标文本和输出音频的对齐关系。对目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行对齐处理，得到目标文本的对齐向量，并对目标文本的对齐向量进行解码处理，得到目标文本的声学特征序列，通过声码器（Vocoder）对声学特征序列进行声学特征转换，得到符合目标对象的音色且与目标文本对应的音频数据，即将声学特征序列转换为与输入的目标文本序列相匹配的合成语音数据。其中，声学特征序列具体可以为梅尔频谱图（Mel-Spectrogram）序列，声学特征序列中每个音素对应的声学特征均为目标对象的声学特征，例如，目标文本的长度为 100（即包括 100 个字符），那么可以将每个字符对应的声学特征都确定为目标对象的声学特征，那么可以将 100 个字符对应的声学特征组成声学特征序列，当声学特征为 1*5 维的向量，那么声学特征序列包括 100 个 1*5 维的向量，可以组成 100*5 维的向量。

其中，声码器具体可以为一种依靠流的从梅尔频谱图合成高质量语音的网络（WaveGlow 网络），可以实现并行化的语音合成，或者可以为一种可用于移动端语音合成的轻量级的流模型（SqueezeWave 网络），可以有效提升语音合成的速度，或者还可以使用诸如 Griffin-Lim、WaveNet、Parallel 的声码器将声学特征序列合成语音，可以根据实际需要选取合适的声码器，本申请实施例对此不做限制。

承接上述示例，声学特征转换的过程如下：对声学特征序列进行平滑处理，得到声学特征序列对应的频谱数据；对声学特征序列对应的频谱数据进行傅里叶变换，得到符合目标对象的音色且与目标文本对应的音频数据。

参见图 5，图 5 是本申请实施例提供的基于人工智能的音频处理方法的流程示意图，

音频处理方法是通过调用神经网络模型实现的，图 5 示出神经网络模型的训练过程，即步骤 106-步骤 109：在步骤 106 中，通过初始化的神经网络模型对对象样本的参考音频数据进行基于音色的注意力处理，得到对象样本的音色嵌入向量；在步骤 107 中，基于对象样本的音色嵌入向量进行对象预测处理，得到参考音频数据的预测对象；在步骤 108 中，基于参考音频数据的预测对象以及对象样本的对象标签，构建神经网络模型的第一损失函数；在步骤 109 中，基于第一损失函数更新神经网络模型的参数，将神经网络模型的更新的参数作为训练后的神经网络模型的参数。

例如，神经网络模型包括编码器、第一分类器，为了能够得到与音频的内容和句子独立的音色嵌入信息以及加大不同对象的区分性，进一步在音色嵌入向量的基础上增加了分类器。对对象样本的多段音频数据进行采样处理，得到对象样本的参考音频数据，通过编码器对对象样本的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量，对参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到对象样本的音色嵌入向量，通过第一分类器对对象样本的音色嵌入向量进行对象预测处理，得到参考音频数据的预测对象（即参考音频数据对应的预测对象（说话人）的概率），当基于参考音频数据的预测对象以及对象样本的对象标签（即真实的说话人标签），确定神经网络模型的第一损失函数的值后，可以判断第一损失函数的值是否超出预设阈值，当第一损失函数的值超出预设阈值时，基于第一损失函数确定神经网络模的误差信号，将误差信号在神经网络模中反向传播，并在传播的过程中更新各个层的模型参数。

需要说明的是，注意力处理是通过注意力机制实现的。在认知科学中，注意力机制（Attention Mechanism）用于选择性地关注所有信息的一部分，同时忽略其他信息。注意力机制可以使得神经网络具备专注于部分输入的能力，即选择特定的输入。在计算能力有限情况下，注意力机制是解决信息超载问题的主要手段的一种资源分配方案，将计算资源分配给更重要的任务。其中，本申请实施例并不局限于注意力机制的形式，例如注意力机制可以是多头注意力、键值对注意力、结构化注意力等。

作为示例，第一损失函数的计算公式为$L_{spk} = \sum_{r=1}^{N} CE(e_m^r, s_m^r)$，其中，N 表示对象样本 m 的参考音频数据的对应的文本数，e_m^r表示为通过第一分类器对对象样本的音色嵌入向量进行对象预测处理，得到参考音频数据的预测对象，s_m^r表示对象标签，CE 表示交叉熵损失，L_{spk}表示第一损失函数。本申请实施例并不局限于$L_{spk} = \sum_{r=1}^{N} CE(e_m^r, s_m^r)$这一种计算公式。

这里，对反向传播进行说明，将训练样本数据输入到神经网络模型的输入层，经过隐藏层，最后达到输出层并输出结果，这是神经网络模型的前向传播过程，由于神经网络模型的输出结果与实际结果有误差，则计算输出结果与实际值之间的误差，并将该误差从输出层向隐藏层反向传播，直至传播到输入层，在反向传播的过程中，根据误差调整模型参数的值，即根据输出结果与实际值之间的误差构建损失函数，并逐层求出损失函数对模型参数的偏导数，生成损失函数对各层模型参数的梯度，由于梯度的方向表明误差扩大的方向，因此对模型参数的梯度取反，与以各层模型的原始参数求和，将得到的求和结果作为更新后的各层模型参数，从而减小模型参数引起的误差；不断迭代上述过程，直至收敛。

在一些实施例中，基于第一损失函数更新神经网络模型的参数之前，通过初始化的神经网络模型执行以下处理至少之一：对文本样本进行文本编码处理，得到文本样本的内容嵌入向量，并基于文本样本的内容嵌入向量构建神经网络模型的第二损失函数；基于对象样本的音色嵌入向量以及文本样本的内容嵌入向量，构建神经网络模型的第三损失函数；其中，第二损失函数以及第三损失函数中至少之一用于与第一损失函数结合（结合的方式可以是加和，还可以是基于注意力机制的加权求和等），以更新神经网络模型

的参数。

例如，神经网络模型还包括文本编码器、第二分类器，为了最大化不同对象之间可以共享的信息，即文本编码器被所有对象共享，在文本编码器端引入了对抗训练机制，即在文本编码器后面加入带梯度反转层（Gradient Reversal Layer）的第二分类器，阻止文本编码捕捉对象信息。通过文本编码器对文本样本进行文本编码处理，得到文本样本的内容嵌入向量，通过第二分类器对文本样本的内容嵌入向量进行对象预测处理，得到文本样本的预测对象，基于文本样本的预测对象以及对象样本的对象标签，构建第四损失函数，对第四损失函数进行反转处理，得到神经网络模型的第二损失函数。从而，通过对抗训练机制阻止文本编码捕捉对象信息，以将文本与对象信息分离开，实现文本与对象信息的解耦，提高内容嵌入向量的准确性，避免与其他信息耦合。

作为示例，第四损失函数的计算公式为$L_1 = \mu \sum_{i=1}^{N} CE(t_m^i, s_m^i)$，其中，$\mu$表示缩放因子，N 表示对象样本 m 的参考音频数据的对应的文本数，t_m^i表示为通过第二分类器对文本样本的内容嵌入向量进行对象预测处理，得到文本样本的预测对象（即文本样本对应的预测对象（说话人）的概率），s_m^i表示对象标签，CE 表示交叉熵损失，第二损失函数的计算公式为$L_{txt} = -L_1$。本申请实施例并不局限于$L_1 = \mu \sum_{i=1}^{N} CE(t_m^i, s_m^i)$这一种计算公式。

例如，为了帮助语音合成在少量数据下更好地学习得到文本和音频之间的对齐关系，利用了文本和音频之间存在的一些关系来训练神经网络模型，即引入了多种不同的文本与音频之间的对齐信息。基于对象样本的音色嵌入向量以及文本样本的内容嵌入向量进行合成处理，得到符合对象样本的音色且与文本样本对应的音频数据样本，首先，对对象样本的音色嵌入向量以及文本样本的内容嵌入向量进行对齐预测处理，得到文本样本的预测对齐信息$\alpha_{t't}$，然后，对对象样本的音色嵌入向量以及文本样本的内容嵌入向量进行基于语音识别的强制对齐处理，得到文本样本的强制对齐信息$A^*_{t't}$；对文本样本的字符数以及音频数据样本的音频帧数进行线性映射处理，得到文本样本的线性对齐信息$A'_{t't}$；基于文本样本的预测对齐信息、文本样本的强制对齐信息以及文本样本的线性对齐信息，构建神经网络模型的第三损失函数。从而，通过对齐信息（即预测对齐信息、强制对齐信息以及线性对齐信息），在少量数据下更好地学习得到文本和音频之间的对齐关系，利用文本和音频之间存在的一些关系来训练神经网络模型，使得训练后的神经网络模型能够进行准确的对齐操作，以提高语音合成的准确性。

需要说明的是，通过语音合成中的声学模型（Acoustic Model）学习文本与音频之间的对齐关系，在学习到文本与音频之间的对齐关系后，通过声学模型对对象样本的音色嵌入向量以及文本样本的内容嵌入向量进行对齐预测处理，得到文本样本的预测对齐信息$\alpha_{t't}$。其中，声学模型是语音识别系统中最为重要的部分之一，多采用隐马尔科夫模型(一种离散时域有限状态自动机)进行建模，声学模型用于提取音频中的声学特征。

通过自动语音识别（ASR，Automatic Speech Recognition）算法学习文本与音频之间的对齐关系，在学习到文本与音频之间的对齐关系后，通过 ASR 算法对对象样本的音色嵌入向量以及文本样本的内容嵌入向量进行基于语音识别的强制对齐（又称维特比对齐），得到文本样本的强制对齐信息$A^*_{t't}$。强制对齐（Forced Alignment)）是指给定音频和文本的情况下，确定每个音素在音频中所处的位置。例如通过 Viterbi 解码技术科实现强制对齐，Viterbi 解码是一个动态规划的算法。

文本样本的线性对齐信息$A'_{t't}$是文本与音频之间存在接近于线性单调的对应关系（即对齐图，输入的文本样本和输出的音频数据样本之间的对角关系），线性对齐信息的计算公式为$A'_{t't} = 1 - \exp\{-\frac{\left(\frac{t'}{T'} - \frac{t}{T}\right)^2}{2g^2}\}$，其中，$T'$表示输入的文本样本的最大字符数，$t'$表

示输入的第t'个字符，T表示输出的最大的音频数，t 表示输出的第 t 个音频帧，g表示缩放因子。

承接上述示例，基于文本样本的预测对齐信息、文本样本的强制对齐信息以及文本样本的线性对齐信息，构建神经网络模型的第三损失函数，包括：基于文本样本的线性对齐信息与基于文本样本的预测对齐信息的差值，构建第一差异信息；基于文本样本的强制对齐信息与基于文本样本的预测对齐信息的差值，构建第二差异信息；对第一差异信息以及第二差异信息进行加权求和处理，得到神经网络模型的第三损失函数。

例如，第一差异信息为$A'_{t't}-\alpha_{t't}$，第二差异信息为$A^*_{t't}-\alpha_{t't}$，第三损失函数的计算公式为$L_{ali}=\frac{1}{T'}\sum_{t'=1}^{T'}\sum_{t=1}^{T}\{(A'_{t't}-\alpha_{t't})^2+(A^*_{t't}-\alpha_{t't})^2$，其中，$\alpha_{t't}$表示预测对齐信息，$A^*_{t't}$表示通过自动语音识别（ASR，Automatic Speech Recognition）得到的强制对齐信息。如果$\alpha_{t't}$与$A'_{t't}$，或者$\alpha_{t't}$与$A^*_{t't}$差距越大，将会得到越大的惩罚。这种基于多对齐信息的训练策略一方面能够通过引入 ASR 中得到的矩阵$A^*_{t't}$来避免缩放因子g的敏感性，另外一方面能够通过$A'_{t't}$来避免强制对齐中存在的对齐偏差的影响。本申请实施例并不局限于$L_{ali}=\frac{1}{T'}\sum_{t'=1}^{T'}\sum_{t=1}^{T}\{(A'_{t't}-\alpha_{t't})^2+(A^*_{t't}-\alpha_{t't})^2$这一种计算公式。

下面，将说明本申请实施例在一个实际的语音合成应用场景中的示例性应用。

本申请实施例可以应用于各种语音合成的应用场景（例如，智能音箱、有屏音箱、智能手表、智能手机、智能家居、智能地图、智能汽车等具有语音合成能力的智能设备等，XX 新闻、XX 听书、在线教育、智能机器人、人工智能客服、语音合成云服务等具有语音合成能力的应用等）中，例如对于车载应用，当用户在开车时，不方便通过文本的形式了解信息，但是可以通过读取语音的方式了解信息，避免遗漏重要的信息，当车载客户端接收到该文本后，需要将该文本转化为语音，以向该用户播放该语音，使得用户可以及时读取到文本对应的语音。

为了解决相关技术中语音合成定制声学建模方法所存在的问题，本申请实施例提出一种基于人工智能的音频处理方法，该方法能够在建模过程中对说话人（即对象）的信息和文本内容信息进行分离。该方法的说话人信息通过带有随机采样机制和说话人分类器的音频编码器（又称编码器）进行特征提取，以保证得到的说话人信息不带有韵律或者风格相关的信息；该方法提出一种对抗训练策略将说话人相关的信息从文本编码器进行剖离；为了让定制训练更加快速有效，充分利用了文本和音频之间的相关性，进一步提出基于多对齐信息的注意力损失来辅助模型进行学习。

为了快速、准确、稳定、高效地进行语音合成声学模型定制，本申请实施例提供了一种基于人工智能的音频处理方法。如图 9 所示，图 9 是本申请实施例提供的训练流程图，主要包括三个步骤：

步骤 11、训练数据准备，包括文本预处理、声学特征提取、音素信息提取。

步骤 12、利用给定的数据训练基于多说话人的语音合成声学模型（即神经网络模型），作为定制声学模型训练的初始模型。

步骤 13、利用给定目标说话人（即目标对象）的数据，利用本申请实施例提供的方法训练目标说话人的语音合成定制模型。

如图 10 所示，图 10 是本申请实施例提供的一种快速有效的语音合成定制模型结构的结构示意图。该结构主要包括带有随机采样机制、说话人注意力模块和说话人分类器的说话人编码器，带有说话人对抗训练的文本编码器和带有多对齐机制的基于序列到序列编解码的端到端声学模型。

如图 10 所示，第一部分是说话人编码器。为了让说话人编码器能够更专注于刻画说话人音色而忽略韵律、风格等音频内容相关的信息，本申请实施例提供了一种带有随

机采样机制的编码器。对于给定的文本和音频对$\{x_m^i, y_m^i\}, i \in [1, N]$，其中，N 表示说话人（即对象）的音频的文本数，首先基于随机采样机制，从说话人 m 对应的语料中随机采样出一个音频数据作为参考编码器（又称编码器）的输入，如公式（1）所示：

$$y^r = y^{Random(N)} \quad (1)$$

其中，Random(N)表示$[1, N]$中的任意正整数，y^r表示从说话人 m 对应的语料中随机采样出的一段音频数据作为参考音频数据，可以基于字符或者音素进行建模，限定最大序列长度为 256。

其中，采样得到的音频数据y^r直接送入到参考编码器（Reference Encoder）中进行编码，参考编码器对该变长的音频数据y^r进行编码，输出对应的参考嵌入向量（Reference Embedding）。其中，如图 11 所示，图 11 是本申请实施例提供的参考编码器的具体结构框图，输入的参考音频经过多层 CNN 编码（例如 6 层卷积层）后再送入单向门控循环单元（GRU，Gated Recurrent Unit）（例如，包括 128 个节点的门控循环网格）中，然后将单向 GRU 最后时刻的表示作为参考嵌入向量（Reference Embedding）。

上述参考嵌入向量（Reference Embedding）已经可以直接作为说话人嵌入向量对说话人音色进行控制。为了构建一个更加鲁棒、准确的说话人嵌入空间，本申请实施例进一步引入说话人注意力层，如图 10 所示，说话人嵌入空间包括 M 个不同类型的说话人令牌（token）（即不同对象），每个说话人 m 嵌入表示则可以由这 M 个说话人 token 的线性组合进行表示，这样将说话人嵌入空间进行进一步的多样化表示能够有利于提升说话人嵌入空间表示的鲁棒性，从而提高了生成音频的自然度、生成的音频的音色与说话人的音色的相似度。

为了能够得到与音频的内容独立的说话人嵌入信息以及加大不同说话人的区分性，进一步在说话人嵌入向量的基础上增加了说话人分类器（用于进行说话人分类，即第一分类器）。对于说话人$m \in [1, S]$中随机采样得到的音频数据y^r，其中，S 表示说话人的数量，说话人损失函数可以表示为基于说话人嵌入向量（即音色嵌入向量）预测的说话人概率e_m^r（即预测对象的概率）和目标说话人标签s_m^r之间的交叉熵（CE）损失（即第一损失函数），如公式（2）所示：

$$L_{spk} = \sum_{r=1}^{N} CE(e_m^r, s_m^r) \quad (2)$$

其中，N 表示说话人 m 的音频数据y^r对应的文本数，L_{spk}表示说话人损失函数，该说话人损失函数有助于从同一说话人的不同音频中获得一致的说话人嵌入向量。

如图 10 所示，第二部分是基于说话人对抗训练的文本编码器。为了最大化不同说话人之间可以共享的信息，即文本编码器被所有说话人共享，本申请实施例在文本编码器端引入了对抗训练机制，即在文本编码器后面加入带梯度反转层（Gradient Reversal Layer）的说话人分类器，阻止文本编码捕捉说话人信息，如公式（3）所示：

$$L_{txt} = -\mu \sum_{i=1}^{N} CE(t_m^i, s_m^i) \quad (3)$$

其中，μ表示缩放因子，N 表示说话人 m 的音频数据y^r对应的文本数，t_m^i表示为基于文本嵌入向量预测的说话人概率，s_m^i表示目标说话人标签，CE 表示交叉熵损失，L_{txt}表示第二损失函数，由于没有必要基于每个说话人学习文本编码器参数，因此对抗训练能够加快训练速度。

如图 10 所示，第三部分是基于多对齐信息指导的训练策略。为了帮助语音合成定制模型在少量数据下更好地学习得到文本和音频之间的对齐关系，本申请实施例利用了文本和音频之间存在的一些关系来训练模型，即引入了多种不同的文本与音频之间的预对齐信息，其中一种对齐信息来源于语音识别中的强制对齐（Force-Alignment）（即公式（5）中的矩阵$A_{t't}^*$），另外一种线性对齐信息是假定文本与音频之间存在接近于线性单调的对应关系（即对齐图，输入的文本和输出的音频之间的对角关系）。其中，线性

对齐信息如公式（4）所示：

$$A'_{t't} = 1 - \exp\{-\frac{\left(\frac{t'}{T'} - \frac{t}{T}\right)^2}{2g^2}\} \quad (4)$$

其中，T'表示输入的最大字符数，t'表示输入的第t'个字符，T表示输出的最大的声学特征帧数，t 表示输出的第 t 个声学特征帧，g表示缩放因子（例如 0.2）。

其中，多对齐信息的注意力损失函数（即第三损失函数）如公式（5）所示：

$$L_{ali} = \frac{1}{T'}\sum_{t'=1}^{T'}\sum_{t=1}^{T}\{(A'_{t't} - \alpha_{t't})^2 + (A^*_{t't} - \alpha_{t't})^2 \quad (5)$$

其中，$\alpha_{t't}$表示声学模型学习得到的文本与音频之间的对齐信息，$A^*_{t't}$表示通过自动语音识别（ASR，Automatic Speech Recognition）得到的文本与音频之间的对齐信息。如果$\alpha_{t't}$与$A'_{t't}$，或者$\alpha_{t't}$与$A^*_{t't}$差距越大，将会得到越大的惩罚。这种基于多对齐信息的训练策略一方面能够通过引入 ASR 中得到的矩阵$A^*_{t't}$来避免公式（4）中参数 g 的敏感性，另外一方面能够通过$A'_{t't}$来避免 ASR 强制对齐中存在的对齐偏差的影响。

本申请实施例提出的方法在录制的中文语音合成语料上进行测试，采用共 60 个说话人约 120 小时的中文普通话语料进行训练。本申请实施例采用自然度和相似度主观平均意见值（MOS，Mean Opinion Score）作为最终评价指标。

对于说话人嵌入空间的验证，比较了三种不同的说话人嵌入空间建模方法，包括相关技术的独热（One-hot）表示法、相关技术的说话人编码方法和本申请实施例提出的说话人编码器（SE）。随机选择一个目标说话人进行消融研究。从实验可知，使用相关技术的说话人编码方法生成的结果是不稳定的，而且合成的文本与参考音频的内容不匹配。

基于 20 个音频对（不包括在训练集中）对相关技术的独热（One-hot）表示法、以及本申请实施例提出的说话人编码器（SE）进行 AB 偏好测试，AB 偏好测试结果如表 1 所示：

表 1 AB 偏好测试

模型	SE	One-hot
AB 偏好测试	56.1%	20.5%

由表 1 可知，本申请实施例提出的基于人工智能的音频处理方法能够获得更多的偏好，这是因为提取的说话人嵌入向量包含有助于解码相应说话人的声学特征的说话人特征相关信息。同时，说话人编码器（SE）加入了一个说话人分类器，以保证说话人嵌入向量对不同说话人具有更好的区分性，从而可以更好地控制后续解码过程中的说话人特征，并且本申请实施例所提出的说话人嵌入向量可以预先离线计算，因此在推理过程中不会带来任何额外的计算成本。

在说话人编码器（SE）的基础上，引入一个对抗训练（SE+AT），以消除文本编码器中的说话人信息，其 AB 偏好测试结果如表 2 所示：

表 2 AB 偏好测试

模型	SE+AT	SE
AB 偏好测试	48.2%	20.3%

由表 2 可知，增加对抗训练可以进一步提高测试效果，这是因为对抗训练可以最大限度地增加说话者之间可以共享信息。

在极少量语料（每个说话人 20 句）的定制上，本申请实施例的测试结果如表 3 所示，基线模型（Baseline）采用相关技术的语音合成定制的方法，本申请实施例提出的语音合成定制模型（Proposed）：

表 3 极少量语料下语音合定制声学模型建模的 MOS 值

模型	说话人标识	Baseline	Proposed
自然度	1	4.08±0.07	4.22±0.09
	2	3.24±0.10	3.75±0.05
	3	3.65±0.05	3.85±0.08
	4	4.00±0.09	4.17±0.08
相似度	1	4.11±0.05	4.28±0.08
	2	3.59±0.13	3.97±0.03
	3	3.81±0.09	4.15±0.02
	4	4.04±0.06	4.25±0.05

由表 3 可知，本申请实施例提出的一种基于人工智能的音频处理方法取得了相比于 Baseline 有更好的性能。

同时，本申请实施例在不同说话人、不同语料规模下也都取得了明显的性能提升，收敛速度也更快，如表 4 所示：

表 4 不同说话人不同语料规模下语音合定制声学模型建模的 MOS 值

模型	语料规模	Baseline	Proposed
自然度	100	3.66±0.05	4.01±0.02
	1000	4.15±0.08	4.31±0.06
	2000	4.30±0.05	4.40±0.04
	10000	4.39±0.04	4.45±0.04
相似度	100	3.81±0.07	4.17±0.04
	1000	4.20±0.05	4.36±0.05
	2000	4.28±0.04	4.35±0.02
	10000	4.67±0.05	4.73±0.05

由上述表可知，本申请实施例提出的一种基于人工智能的音频处理方法在录制的语音合成语料中、在不同的说话人和不同的语料规模下均取得了比相关技术更高的自然度和相似度，合成语音的清晰度、自然度更好，合成语音的频谱细节上也更清晰，另外大大缩短了定制合成模型训练收敛的时间。

综上，本申请实施例提出的一种基于人工智能的音频处理方法具有以下有益效果：

1）从所有的音频数据中随机采样一个音频作为说话人编码器的信息，从而保证得到的说话人嵌入向量跟音频内容无关，另外，进一步引入说话人分类器，确保对不同的说话人有更好的区分性；

2）为了能够更好的学习到说话人无关的文本，引入说话人对抗训练机制，让文本编码器不能区分出文本属于哪个说话人，以最大化不同说话人之间可以共享的信息；

3）充分利用训练文本和音频之间的相关性，引入了多对齐机制的损失函数，从而有效提升了模型收敛的速度和模型稳定性。

至此已经结合本申请实施例提供的服务器的示例性应用和实施，说明本申请实施例提供的基于人工智能的音频处理方法。本申请实施例还提供音频处理装置，实际应用中，音频处理装置中的各功能模块可以由电子设备（如终端设备、服务器或服务器集群）的硬件资源，如处理器等计算资源、通信资源（如用于支持实现光缆、蜂窝等各种方式通信）、存储器协同实现。图 2 示出了存储在存储器 550 中的音频处理装置 555，其可以是程序和插件等形式的软件，例如，软件 C/C++、Java 等编程语言设计的软件模块、C/C++、Java 等编程语言设计的应用软件或大型软件系统中的专用软件模块、应用程序接口、插件、云服务等实现方式，下面对不同的实现方式举例说明。

示例一、音频处理装置是移动端应用程序及模块

本申请实施例中的音频处理装置555可提供为使用软件C/C++、Java等编程语言设计的软件模块，嵌入到基于Android或iOS等系统的各种移动端应用中（以可执行指令存储在移动端的存储介质中，由移动端的处理器执行），从而直接使用移动端自身的计算资源完成相关的音频合成任务，并且定期或不定期地通过各种网络通信方式将处理结果传送给远程的服务器，或者在移动端本地保存。

示例二、音频处理装置是服务器应用程序及平台

本申请实施例中的音频处理装置555可提供为使用C/C++、Java等编程语言设计的应用软件或大型软件系统中的专用软件模块，运行于服务器端（以可执行指令的方式在服务器端的存储介质中存储，并由服务器端的处理器运行），服务器使用自身的计算资源完成相关的音频合成任务。

本申请实施例还可以提供为在多台服务器构成的分布式、并行计算平台上，搭载定制的、易于交互的网络（Web）界面或其他各用户界面（UI，User Interface），形成供个人、群体或单位使用的音频合成平台（用于音频合成）等。

示例三、音频处理装置是服务器端应用程序接口（API，Application Program Interface）及插件

本申请实施例中的音频处理装置555可提供为服务器端的API或插件，以供用户调用，以执行本申请实施例的基于人工智能的音频处理方法，并嵌入到各类应用程序中。

示例四、音频处理装置是移动设备客户端API及插件

本申请实施例中的音频处理装置555可提供为移动设备端的API或插件，以供用户调用，以执行本申请实施例的基于人工智能的音频处理方法。

示例五、音频处理装置是云端开放服务

本申请实施例中的音频处理装置555可提供为向用户开发的信息推荐云服务，供个人、群体或单位获取音频。

其中，音频处理装置555包括一系列的模块，包括采样模块5551、音频编码模块5552、注意力模块5553、文本编码模块5554、合成模块5555以及训练模块5556。下面继续说明本申请实施例提供的音频处理装置555中各个模块配合实现音频处理方案。

采样模块5551，配置为对目标对象的多段音频数据进行采样处理，得到所述目标对象的参考音频数据；音频编码模块5552，配置为对所述目标对象的参考音频数据进行音频编码处理，得到所述参考音频数据的参考嵌入向量；注意力模块5553，配置为对所述参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到所述目标对象的音色嵌入向量；文本编码模块5554，配置为对目标文本进行文本编码处理，得到所述目标文本的内容嵌入向量；合成模块5555，配置为对所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行合成处理，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

在一些实施例中，所述注意力模块5553还配置为对所述参考音频数据的参考嵌入向量进行基于多个对象音色的嵌入空间的映射处理，得到所述多个对象音色的权重；基于所述多个对象音色的权重，对所述多个对象音色的嵌入向量进行加权求和处理，得到所述目标对象的音色嵌入向量。

在一些实施例中，所述音频编码模块5552还配置为对所述目标对象的参考音频数据进行卷积处理，得到所述参考音频数据的卷积嵌入向量；对所述参考音频数据的卷积嵌入向量进行编码处理，得到所述参考音频数据的参考嵌入向量。

在一些实施例中，所述音频编码是通过编码器实现的，所述编码器包括多个级联的卷积层；所述音频编码模块5552还配置为通过所述多个级联的卷积层中的第一个卷积

层，对所述目标对象的参考音频数据进行卷积处理；将所述第一个卷积层的卷积结果输出到后续级联的卷积层，通过所述后续级联的卷积层继续进行卷积处理以及卷积结果输出，直至输出到最后一个卷积层，并将所述最后一个卷积层输出的卷积结果作为所述参考音频数据的卷积嵌入向量。

在一些实施例中，所述音频编码是通过编码器实现的，所述编码器包括循环神经网络；所述音频编码模块 5552 还配置为对所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行向量更新处理，得到所述参考音频数据的更新信息；对所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行向量重置处理，得到所述参考音频数据的重置信息；基于所述参考音频数据的更新信息以及所述参考音频数据的重置信息，对所述参考音频数据的卷积嵌入向量进行上下文编码处理，得到所述参考音频数据的参考嵌入向量。

在一些实施例中，所述音频编码模块 5552 还配置为对所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行拼接处理，得到拼接向量；对所述拼接向量进行基于更新门的映射处理，得到所述参考音频数据的更新信息。

在一些实施例中，所述音频编码模块5552还配置为基于所述目标对象的重置信息、所述循环神经网络的隐藏向量以及所述目标对象的卷积嵌入向量进行基于候选向量的映射处理，得到所述参考音频数据的候选嵌入向量；对所述参考音频数据的更新信息、所述循环神经网络的隐藏向量以及所述参考音频数据的候选嵌入向量进行向量映射处理，得到所述参考音频数据的参考嵌入向量。

在一些实施例中，所述合成模块 5555 还配置为对所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行对齐处理，得到所述目标文本的对齐向量；对所述目标文本的对齐向量进行解码处理，得到所述目标文本的声学特征序列；对所述声学特征序列进行声学特征转换，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

在一些实施例中，所述合成模块5555还配置为对所述声学特征序列进行平滑处理，得到所述声学特征序列对应的频谱数据；对所述声学特征序列对应的频谱数据进行傅里叶变换，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

在一些实施例中，所述音频处理方法是通过调用神经网络模型实现的；所述装置还包括：训练模块 5556，配置为通过初始化的所述神经网络模型对对象样本的参考音频数据进行基于音色的注意力处理，得到所述对象样本的音色嵌入向量；基于所述对象样本的音色嵌入向量进行对象预测处理，得到所述参考音频数据的预测对象；基于所述参考音频数据的预测对象以及所述对象样本的对象标签，构建所述神经网络模型的第一损失函数；基于所述第一损失函数更新所述神经网络模型的参数，将所述神经网络模型的更新的参数作为训练后的所述神经网络模型的参数。

在一些实施例中，所述训练模块 5556 还配置为通过初始化的所述神经网络模型执行以下处理至少之一：对文本样本进行文本编码处理，得到所述文本样本的内容嵌入向量，并基于所述文本样本的内容嵌入向量构建所述神经网络模型的第二损失函数；基于所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量，构建所述神经网络模型的第三损失函数；其中，所述第二损失函数以及所述第三损失函数中至少之一用于与所述第一损失函数结合，以更新所述神经网络模型的参数。

在一些实施例中，所述训练模块 5556 还配置为基于所述文本样本的内容嵌入向量进行对象预测处理，得到所述文本样本的预测对象；基于所述文本样本的预测对象以及所述对象样本的对象标签，构建第四损失函数；对所述第四损失函数进行反转处理，得到所述神经网络模型的第二损失函数。

在一些实施例中，所述训练模块 5556 还配置为对所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量进行合成处理，得到符合所述对象样本的音色且与所述文本样本对应的音频数据样本；对所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量进行对齐预测处理，得到所述文本样本的预测对齐信息；对所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量进行强制对齐处理，得到所述文本样本的强制对齐信息；对所述文本样本的字符数以及所述音频数据样本的音频帧数进行线性映射处理，得到所述文本样本的线性对齐信息；基于所述文本样本的预测对齐信息、所述文本样本的强制对齐信息以及所述文本样本的线性对齐信息，构建所述神经网络模型的第三损失函数。

在一些实施例中，所述训练模块 5556 还配置为基于所述文本样本的线性对齐信息与基于所述文本样本的预测对齐信息的差值，构建第一差异信息；基于所述文本样本的强制对齐信息与基于所述文本样本的预测对齐信息的差值，构建第二差异信息；对所述第一差异信息以及所述第二差异信息进行加权求和处理，得到所述神经网络模型的第三损失函数。

本申请实施例提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行本申请实施例上述的基于人工智能的音频处理方法。

本申请实施例提供一种存储有可执行指令的计算机可读存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本申请实施例提供的基于人工智能的音频处理方法，例如，如图 3-图 5 示出的基于人工智能的音频处理方法。

在一些实施例中，计算机可读存储介质可以是 FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或 CD-ROM 等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言（包括编译或解释语言，或者声明性或过程性语言）来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

作为示例，可执行指令可以但不一定对应于文件系统中的文件，可以可被存储在保存其它程序或数据的文件的一部分，例如，存储在超文本标记语言（HTML，Hyper Text Markup Language）文档中的一个或多个脚本中，存储在专用于所讨论的程序的单个文件中，或者，存储在多个协同文件（例如，存储一个或多个模块、子程序或代码部分的文件）中。

作为示例，可执行指令可被部署为在一个计算设备上执行，或者在位于一个地点的多个计算设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个计算设备上执行。

可以理解的是，在本申请实施例中，涉及到音频数据等用户相关的数据，当本申请实施例运用到具体产品或技术中时，需要获得用户许可或者同意，且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准。

以上所述，仅为本申请的实施例而已，并非用于限定本申请的保护范围。凡在本申请的精神和范围之内所作的任何修改、等同替换和改进等，均包含在本申请的保护范围之内。

权利要求书

1、一种基于人工智能的音频处理方法，由电子设备执行，所述方法包括:

对目标对象的多段音频数据进行采样处理，得到所述目标对象的参考音频数据;

对所述目标对象的参考音频数据进行音频编码处理，得到所述参考音频数据的参考嵌入向量;

对所述参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到所述目标对象的音色嵌入向量;

对目标文本进行文本编码处理，得到所述目标文本的内容嵌入向量;

对所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行合成处理，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

2、根据权利要求 1 所述的方法，其中，所述对所述参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到所述目标对象的音色嵌入向量，包括:

对所述参考音频数据的参考嵌入向量进行基于多个对象音色的嵌入空间的映射处理，得到所述多个对象音色的权重;

基于所述多个对象音色的权重，对所述多个对象音色的嵌入向量进行加权求和处理，得到所述目标对象的音色嵌入向量。

3、根据权利要求 1 所述的方法，其中，所述对所述目标对象的参考音频数据进行音频编码处理，得到所述参考音频数据的参考嵌入向量，包括:

对所述目标对象的参考音频数据进行卷积处理，得到所述参考音频数据的卷积嵌入向量;

对所述参考音频数据的卷积嵌入向量进行编码处理，得到所述参考音频数据的参考嵌入向量。

4、根据权利要求 3 所述的方法，其中，

所述音频编码是通过编码器实现的，所述编码器包括多个级联的卷积层;

所述对所述目标对象的参考音频数据进行卷积处理，得到所述参考音频数据的卷积嵌入向量，包括:

通过所述多个级联的卷积层中的第一个卷积层，对所述目标对象的参考音频数据进行卷积处理;

将所述第一个卷积层的卷积结果输出到后续级联的卷积层，通过所述后续级联的卷积层继续进行卷积处理以及卷积结果输出，直至输出到最后一个卷积层，并

将所述最后一个卷积层输出的卷积结果作为所述参考音频数据的卷积嵌入向量。

5、根据权利要求 3 所述的方法，其中，

所述音频编码是通过编码器实现的，所述编码器包括循环神经网络;

所述对所述参考音频数据的卷积嵌入向量进行编码处理，得到所述参考音频数据的参考嵌入向量，包括:

对所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行向量更新处理，得到所述参考音频数据的更新信息;

对所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行向量重置处理，得到所述参考音频数据的重置信息;

基于所述参考音频数据的更新信息以及所述参考音频数据的重置信息，对所述参考音频数据的卷积嵌入向量进行上下文编码处理，得到所述参考音频数据的参考嵌入向量。

6、根据权利要求 5 所述的方法，其中，对基于所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行向量更新处理，得到所述参考音频数据的更新信息，包括:

对所述循环神经网络的隐藏向量以及所述参考音频数据的卷积嵌入向量进行拼接处理，得到拼接向量；

对所述拼接向量进行基于更新门的映射处理，得到所述参考音频数据的更新信息。

7、根据权利要求 5 所述的方法，其中，所述基于所述参考音频数据的更新信息以及所述参考音频数据的重置信息，对所述参考音频数据的卷积嵌入向量进行上下文编码处理，得到所述参考音频数据的参考嵌入向量，包括：

基于所述目标对象的重置信息、所述循环神经网络的隐藏向量以及所述目标对象的卷积嵌入向量进行基于候选向量的映射处理，得到所述参考音频数据的候选嵌入向量；

对所述参考音频数据的更新信息、所述循环神经网络的隐藏向量以及所述参考音频数据的候选嵌入向量进行向量映射处理，得到所述参考音频数据的参考嵌入向量。

8、根据权利要求 1 所述的方法，其中，所述对所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行合成处理，得到符合所述目标对象的音色且与所述目标文本对应的音频数据，包括：

对所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行对齐处理，得到所述目标文本的对齐向量；

对所述目标文本的对齐向量进行解码处理，得到所述目标文本的声学特征序列；

对所述声学特征序列进行声学特征转换，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

9、根据权利要求 8 所述的方法，其中，所述对所述声学特征序列进行声学特征转换，得到符合所述目标对象的音色且与所述目标文本对应的音频数据，包括：

对所述声学特征序列进行平滑处理，得到所述声学特征序列对应的频谱数据；

对所述声学特征序列对应的频谱数据进行傅里叶变换，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

10、根据权利要求 1-9 任一项所述的方法，其中，所述音频处理方法是通过调用神经网络模型实现的；所述神经网络模型的训练过程包括：

通过初始化的所述神经网络模型对对象样本的参考音频数据进行基于音色的注意力处理，得到所述对象样本的音色嵌入向量；

基于所述对象样本的音色嵌入向量进行对象预测处理，得到所述参考音频数据的预测对象；

基于所述参考音频数据的预测对象以及所述对象样本的对象标签，构建所述神经网络模型的第一损失函数；

基于所述第一损失函数更新所述神经网络模型的参数，将所述神经网络模型的更新的参数作为训练后的所述神经网络模型的参数。

11、根据权利要求 10 所述的方法，其中，所述基于所述第一损失函数更新所述神经网络模型的参数之前，所述方法还包括：

通过初始化的所述神经网络模型执行以下处理至少之一：

对文本样本进行文本编码处理，得到所述文本样本的内容嵌入向量，并基于所述文本样本的内容嵌入向量构建所述神经网络模型的第二损失函数；

基于所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量，构建所述神经网络模型的第三损失函数；

其中，所述第二损失函数以及所述第三损失函数中至少之一用于与所述第一损失函数结合，以更新所述神经网络模型的参数。

12、根据权利要求 11 所述的方法，其中，所述基于所述文本样本的内容嵌入向量构建所述神经网络模型的第二损失函数，包括：

基于所述文本样本的内容嵌入向量进行对象预测处理，得到所述文本样本的预测对象；

基于所述文本样本的预测对象以及所述对象样本的对象标签，构建第四损失函数；

对所述第四损失函数进行反转处理，得到所述神经网络模型的第二损失函数。

13、根据权利要求 11 所述的方法，其中，所述基于所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量，构建所述神经网络模型的第三损失函数，包括：

对所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量进行合成处理，得到符合所述对象样本的音色且与所述文本样本对应的音频数据样本；

对所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量进行对齐预测处理，得到所述文本样本的预测对齐信息；

对所述对象样本的音色嵌入向量以及所述文本样本的内容嵌入向量进行强制对齐处理，得到所述文本样本的强制对齐信息；

对所述文本样本的字符数以及所述音频数据样本的音频帧数进行线性映射处理，得到所述文本样本的线性对齐信息；

基于所述文本样本的预测对齐信息、所述文本样本的强制对齐信息以及所述文本样本的线性对齐信息，构建所述神经网络模型的第三损失函数。

14、根据权利要求 13 所述的方法，其中，所述基于所述文本样本的预测对齐信息、所述文本样本的强制对齐信息以及所述文本样本的线性对齐信息，构建所述神经网络模型的第三损失函数，包括：

基于所述文本样本的线性对齐信息与基于所述文本样本的预测对齐信息的差值，构建第一差异信息；

基于所述文本样本的强制对齐信息与基于所述文本样本的预测对齐信息的差值，构建第二差异信息；

对所述第一差异信息以及所述第二差异信息进行加权求和处理，得到所述神经网络模型的第三损失函数。

15、一种基于人工智能的音频处理装置，所述装置包括：

采样模块，配置为对目标对象的多段音频数据进行采样处理，得到所述目标对象的参考音频数据；

音频编码模块，配置为对所述目标对象的参考音频数据进行音频编码处理，得到所述参考音频数据的参考嵌入向量；

注意力模块，配置为对所述参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到所述目标对象的音色嵌入向量；

文本编码模块，配置为对目标文本进行文本编码处理，得到所述目标文本的内容嵌入向量；

合成模块，配置为对所述目标对象的音色嵌入向量以及所述目标文本的内容嵌入向量进行合成处理，得到符合所述目标对象的音色且与所述目标文本对应的音频数据。

16、一种电子设备，所述电子设备包括：

存储器，用于存储可执行指令；

处理器，用于执行所述存储器中存储的可执行指令时，实现权利要求 1 至 14 任一项所述的基于人工智能的音频处理方法。

17、一种计算机可读存储介质，存储有可执行指令，用于被处理器执行时实现权利要求 1 至 14 任一项所述的基于人工智能的音频处理方法。

18、一种计算机程序产品，包括计算机程序或指令，所述计算机程序或指令使得计算机执行如权利要求 1 至 14 中任一项所述的基于人工智能的音频处理方法。

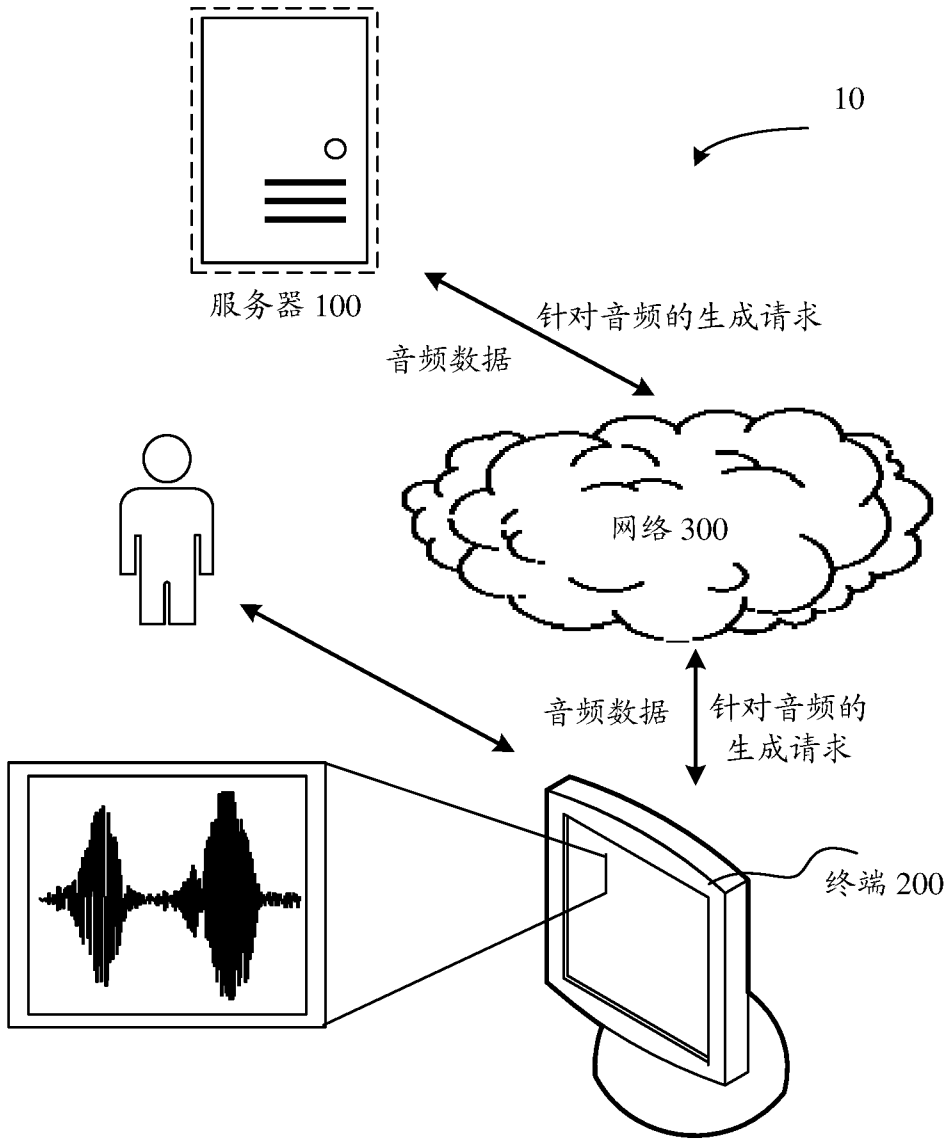

图 1

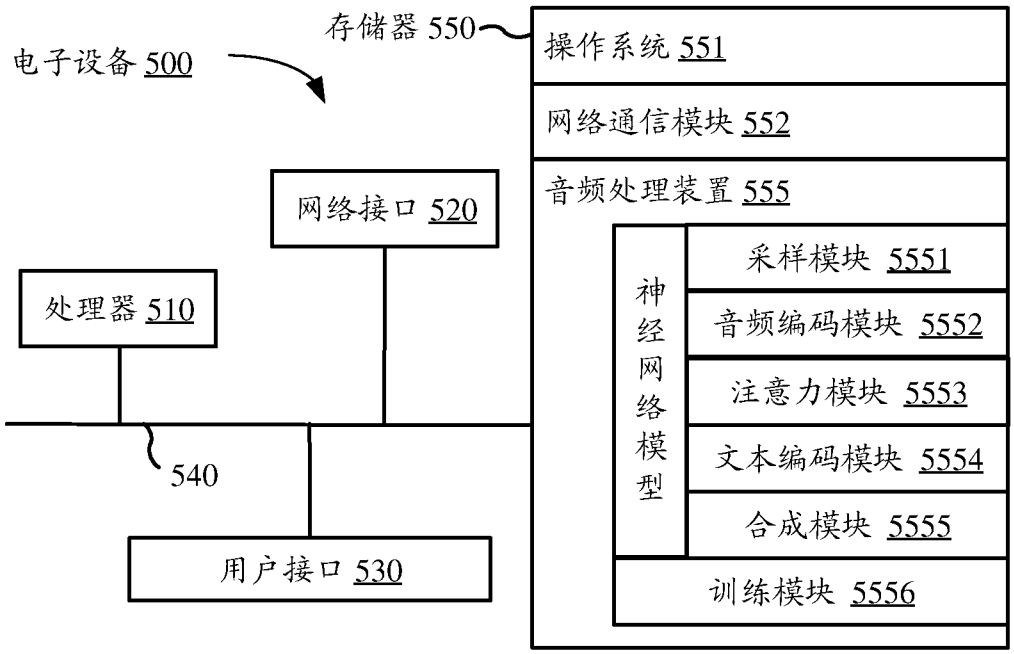
电子设备 500
存储器 550
操作系统 551
网络通信模块 552
音频处理装置 555
网络接口 520
处理器 510
神经网络模型
采样模块 5551
音频编码模块 5552
注意力模块 5553
文本编码模块 5554
合成模块 5555
训练模块 5556
540
用户接口 530

图 2

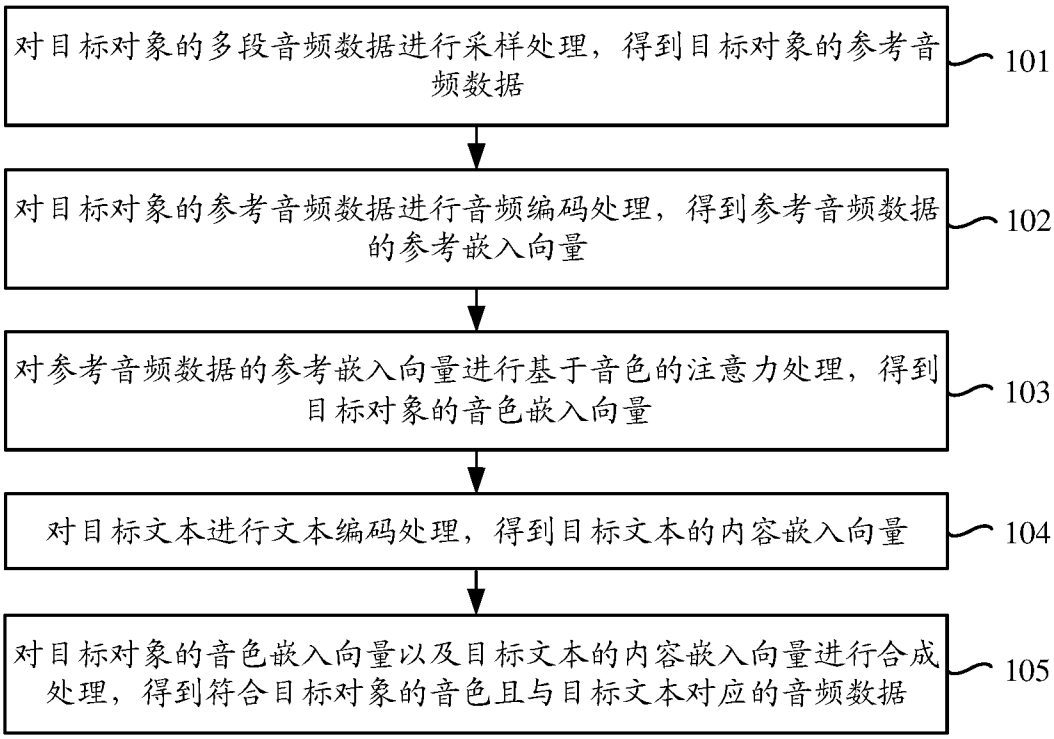
对目标对象的多段音频数据进行采样处理，得到目标对象的参考音频数据
101
对目标对象的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量
102
对参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到目标对象的音色嵌入向量
103
对目标文本进行文本编码处理，得到目标文本的内容嵌入向量
104
对目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据
105

图 3

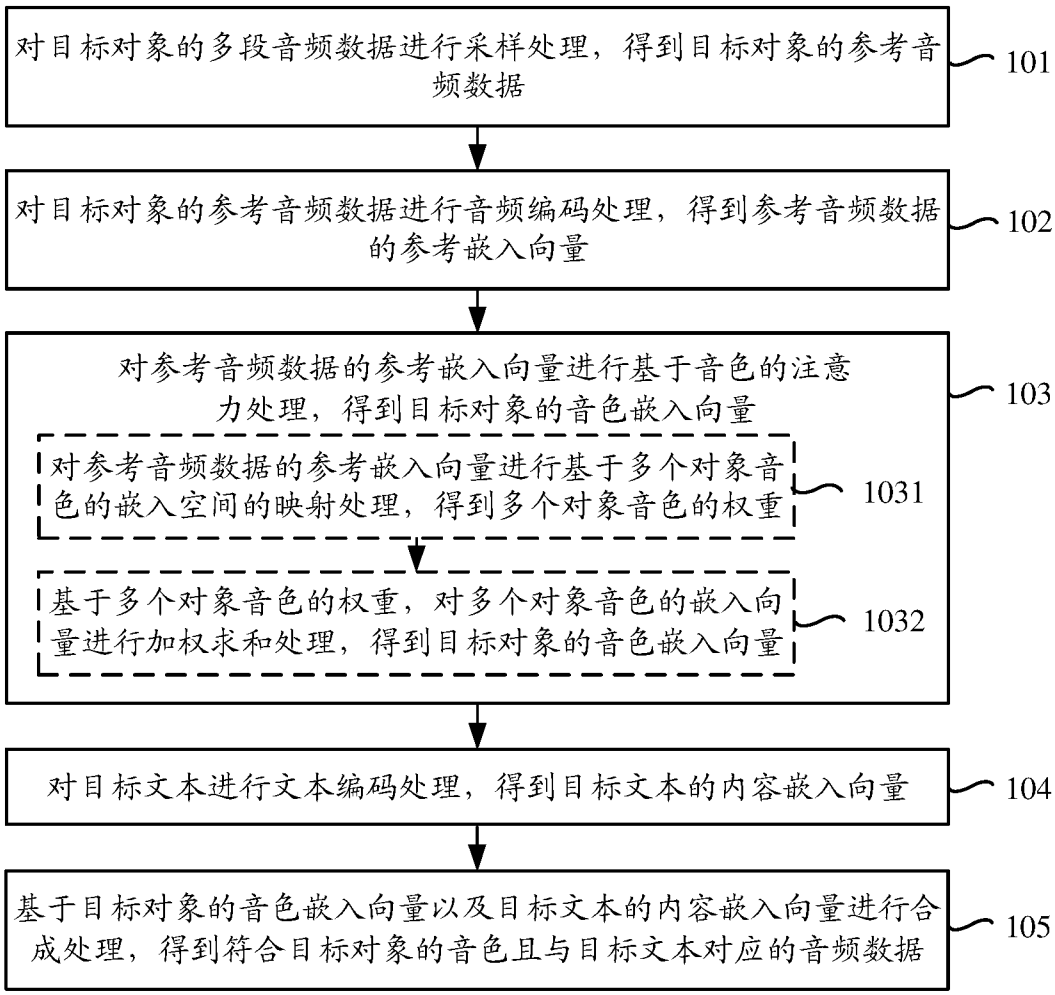
对目标对象的多段音频数据进行采样处理，得到目标对象的参考音频数据
101
对目标对象的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量
102
对参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到目标对象的音色嵌入向量
103
对参考音频数据的参考嵌入向量进行基于多个对象音色的嵌入空间的映射处理，得到多个对象音色的权重
1031
基于多个对象音色的权重，对多个对象音色的嵌入向量进行加权求和处理，得到目标对象的音色嵌入向量
1032
对目标文本进行文本编码处理，得到目标文本的内容嵌入向量
104
基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据
105

图 4

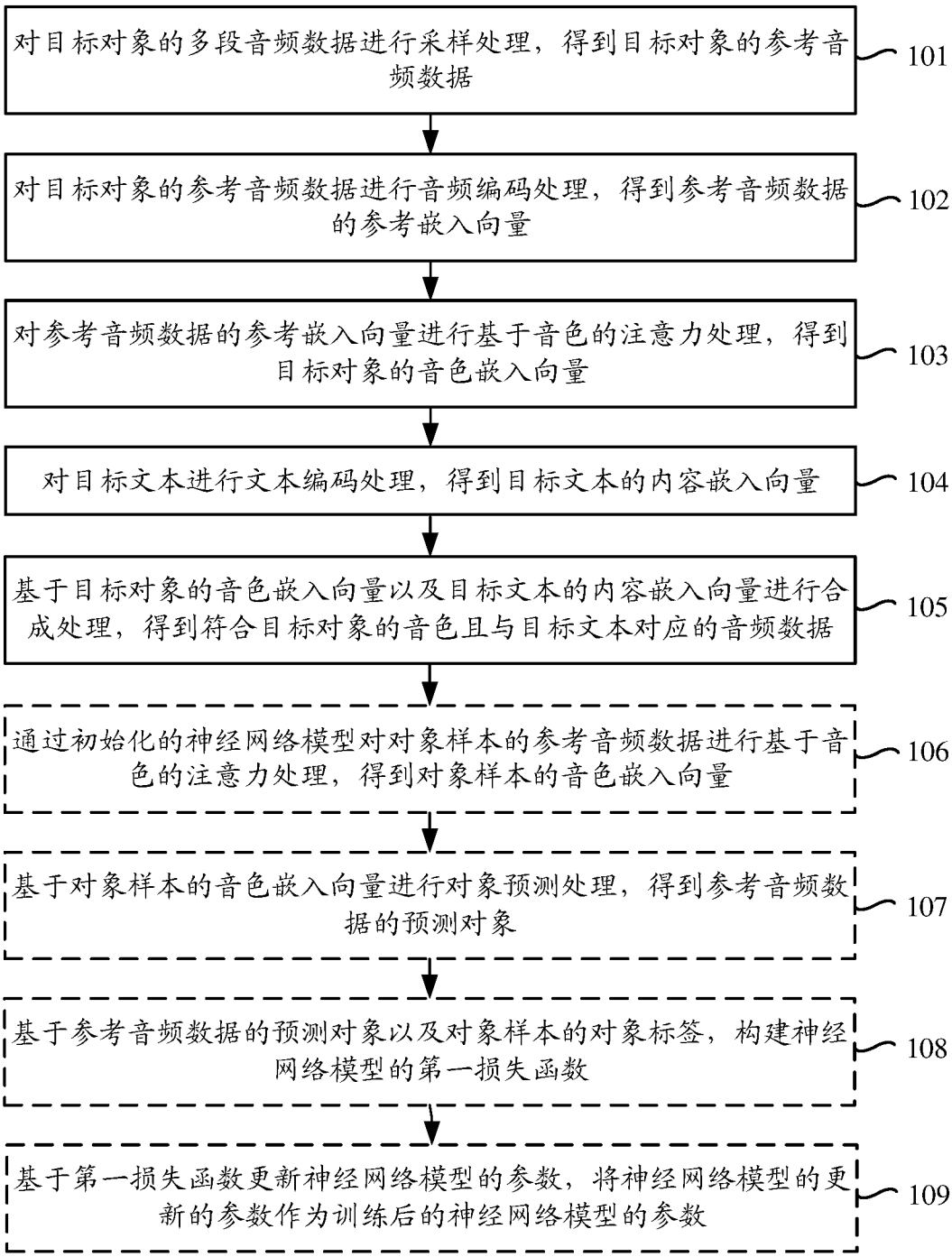
对目标对象的多段音频数据进行采样处理，得到目标对象的参考音频数据
101
对目标对象的参考音频数据进行音频编码处理，得到参考音频数据的参考嵌入向量
102
对参考音频数据的参考嵌入向量进行基于音色的注意力处理，得到目标对象的音色嵌入向量
103
对目标文本进行文本编码处理，得到目标文本的内容嵌入向量
104
基于目标对象的音色嵌入向量以及目标文本的内容嵌入向量进行合成处理，得到符合目标对象的音色且与目标文本对应的音频数据
105
通过初始化的神经网络模型对对象样本的参考音频数据进行基于音色的注意力处理，得到对象样本的音色嵌入向量
106
基于对象样本的音色嵌入向量进行对象预测处理，得到参考音频数据的预测对象
107
基于参考音频数据的预测对象以及对象样本的对象标签，构建神经网络模型的第一损失函数
108
基于第一损失函数更新神经网络模型的参数，将神经网络模型的更新的参数作为训练后的神经网络模型的参数
109

图 5

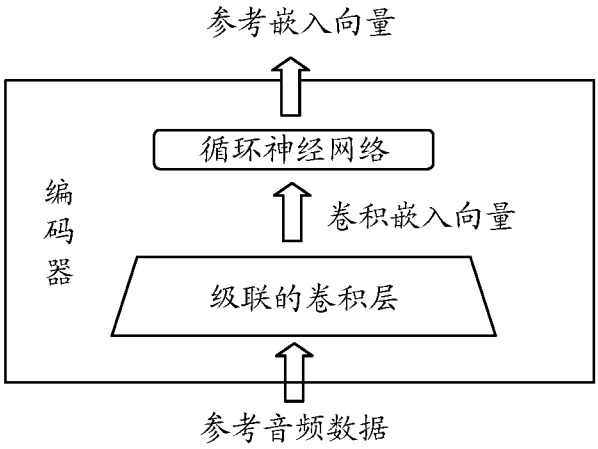
参考嵌入向量
循环神经网络
编码器
卷积嵌入向量
级联的卷积层
参考音频数据

图 6

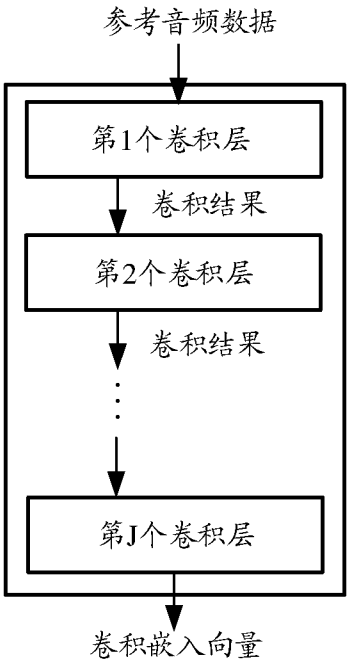
参考音频数据
第1个卷积层
卷积结果
第2个卷积层
卷积结果
第J个卷积层
卷积嵌入向量

图 7

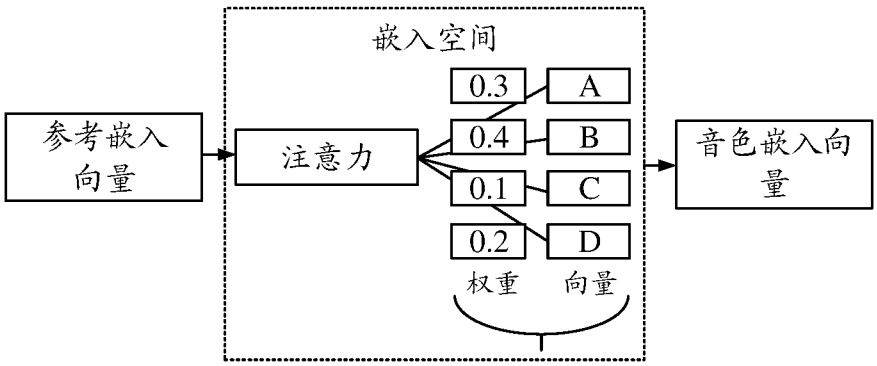
嵌入空间
参考嵌入向量
注意力
0.3
A
0.4
B
0.1
C
0.2
D
权重
向量
音色嵌入向量

图 8

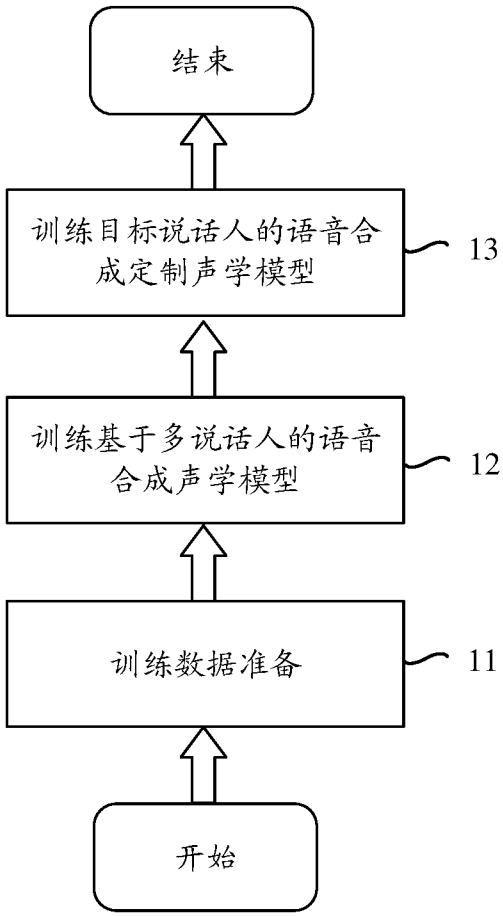
结束
训练目标说话人的语音合成定制声学模型
13
训练基于多说话人的语音合成声学模型
12
训练数据准备
11
开始

图 9

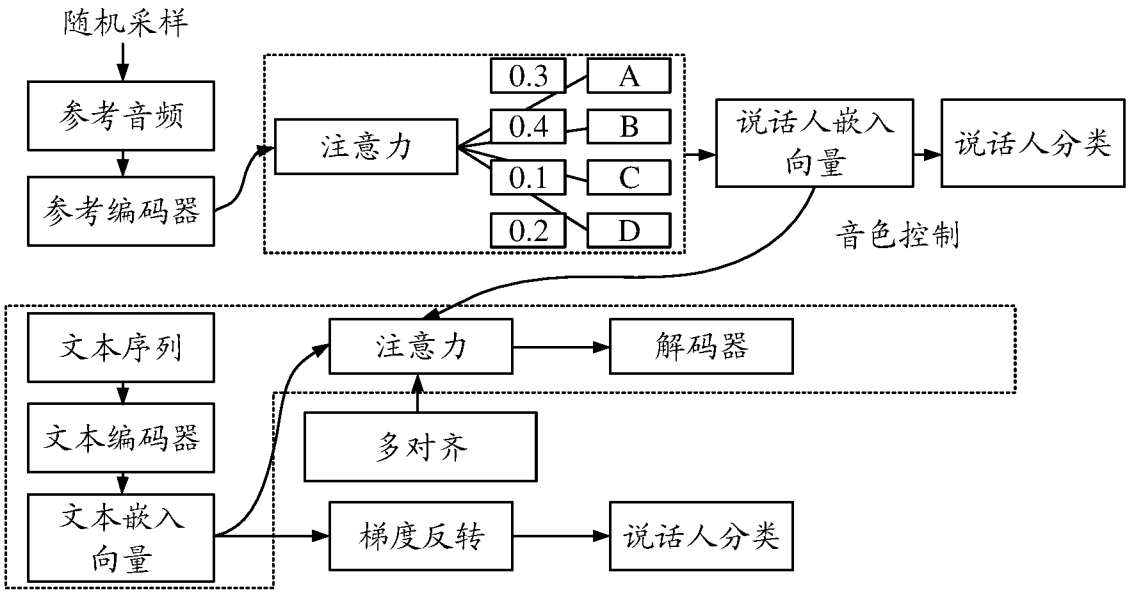
随机采样
参考音频
参考编码器
注意力
0.3
A
0.4
B
0.1
C
0.2
D
说话人嵌入向量
说话人分类
音色控制
文本序列
文本编码器
文本嵌入向量
注意力
解码器
多对齐
梯度反转
说话人分类

图 10

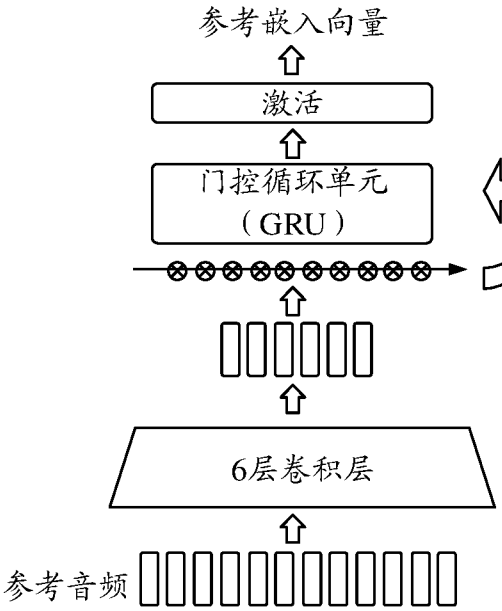

图 11

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/090951

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/126(2020.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G10L

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

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

CNPAT, CNKI, WPI, EPODOC, IEEE: 音频, 音色, 文本, 内容, 注意力, 编码, 嵌入, 向量, audio, speech, text, attention, cod +, embed+ w vector

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113822017 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 21 December 2021 (2021-12-21) claims 1-17, and description, paragraph 0221	1-18
X	CN 112786009 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 11 May 2021 (2021-05-11) abstract, description, paragraphs 0031-0059, and claims 9, 10	1-18
A	CN 112802448 A (YIWISE TECHNOLOGY LTD.) 14 May 2021 (2021-05-14) entire document	1-18
A	CN 111326136 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 23 June 2020 (2020-06-23) entire document	1-18
A	CN 112687258 A (BEIJING CENTURY TAL EDUCATION TECHNOLOGY CO., LTD.) 20 April 2021 (2021-04-20) entire document	1-18
A	US 2021142783 A1 (NEOSAPIENCE INC.) 13 May 2021 (2021-05-13) entire document	1-18

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
21 July 2022	**03 August 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/090951

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 113822017 A	21 December 2021	None	
CN 112786009 A	11 May 2021	None	
CN 112802448 A	14 May 2021	None	
CN 111326136 A	23 June 2020	None	
CN 112687258 A	20 April 2021	None	
US 2021142783 A1	13 May 2021	KR 20210082153 A	02 July 2021
		KR 20200119217 A	19 October 2020
		WO 2020209647 A1	15 October 2020

国际检索报告

国际申请号
PCT/CN2022/090951

A. 主题的分类

G06F 40/126(2020.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; G10L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT,CNKI,WPI,EPODOC,IEEE：音频，音色，文本，内容，注意力，编码，嵌入，向量，audio，speech，text，attention，cod+，embed+ w vector

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 113822017 A (腾讯科技深圳有限公司) 2021年12月21日 (2021-12-21) 权利要求1-17、说明书第0221段	1-18
X	CN 112786009 A (平安科技深圳有限公司) 2021年5月11日 (2021-05-11) 摘要、说明书第0031-0059段、权利要求9、10	1-18
A	CN 112802448 A (杭州一知智能科技有限公司) 2021年5月14日 (2021-05-14) 全文	1-18
A	CN 111326136 A (腾讯科技深圳有限公司) 2020年6月23日 (2020-06-23) 全文	1-18
A	CN 112687258 A (北京世纪好未来教育科技有限公司) 2021年4月20日 (2021-04-20) 全文	1-18
A	US 2021142783 A1 (NEOSAPIENCE., INC.) 2021年5月13日 (2021-05-13) 全文	1-18

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2022年7月21日	2022年8月3日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 霍玉明 电话号码 010-53961327

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/090951

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	113822017	A	2021年12月21日	无			
CN	112786009	A	2021年5月11日	无			
CN	112802448	A	2021年5月14日	无			
CN	111326136	A	2020年6月23日	无			
CN	112687258	A	2021年4月20日	无			
US	2021142783	A1	2021年5月13日	KR	20210082153	A	2021年7月2日
				KR	20200119217	A	2020年10月19日
				WO	2020209647	A1	2020年10月15日