



(12)发明专利申请

(10)申请公布号 CN 112802446 A

(43)申请公布日 2021.05.14

(21)申请号 201911114561.3

(22)申请日 2019.11.14

(71)申请人 腾讯科技(深圳)有限公司

地址 518000 广东省深圳市南山区高新区
科技中一路腾讯大厦35层

(72)发明人 张黄斌 李辉

(74)专利代理机构 深圳市隆天联鼎知识产权代
理有限公司 44232

代理人 刘抗美

(51)Int.Cl.

G10L 13/04(2013.01)

G10L 13/08(2013.01)

G10L 13/10(2013.01)

G06N 3/08(2006.01)

G06N 3/04(2006.01)

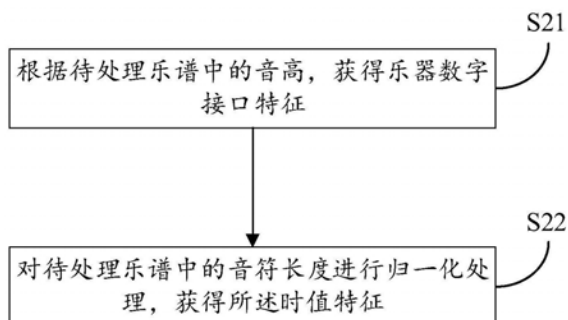
权利要求书2页 说明书19页 附图18页

(54)发明名称

音频合成方法及装置、电子设备和计算机可读存储介质

(57)摘要

本公开实施例提供了一种音频合成方法及装置、电子设备和计算机可读存储介质。该方法包括：获取待处理乐谱及其待处理歌词；从所述待处理乐谱中提取到音乐特征；从所述待处理歌词中提取到文本特征；通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理，获得频谱信息；根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。通过本公开实施例提供的方案，能够通过端到端神经网络模型实现歌唱音频的合成。



1. 一种音频合成方法,其特征在于,包括:
获取待处理乐谱及其待处理歌词;
从所述待处理乐谱中提取音乐特征;
从所述待处理歌词中提取文本特征;
通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息;
根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。
2. 根据权利要求1所述的方法,其特征在于,所述音乐特征包括乐器数字接口特征和时值特征;其中,从所述待处理乐谱中提取音乐特征,包括:
根据所述待处理乐谱中的音高,获得所述乐器数字接口特征;
对所述待处理乐谱中的音符长度进行归一化处理,获得所述时值特征。
3. 根据权利要求2所述的方法,其特征在于,所述端到端神经网络模型包括文本编码器、乐器数字接口编码器和时值编码器;其中,通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息,包括:
通过所述文本编码器对所述文本特征进行处理,获得文本嵌入向量;
通过所述乐器数字接口编码器对所述乐器数字接口特征进行处理,获得乐器数字接口嵌入向量;
通过所述时值编码器对所述时值特征进行处理,获得时值嵌入向量;
根据所述文本嵌入向量、所述乐器数字接口嵌入向量和所述时值嵌入向量,获得融合嵌入向量;
根据所述融合嵌入向量,获得所述频谱信息。
4. 根据权利要求3所述的方法,其特征在于,所述乐器数字接口编码器包括第一稠密神经网络;其中,通过所述乐器数字接口编码器对所述乐器数字接口特征进行处理,获得乐器数字接口嵌入向量,包括:
通过所述第一稠密神经网络对所述乐器数字接口特征进行处理,获得第一稠密向量;
根据所述第一稠密向量,获得所述乐器数字接口嵌入向量。
5. 根据权利要求4所述的方法,其特征在于,所述乐器数字接口编码器还包括第二稠密神经网络、正向门递归神经网络和反向门递归神经网络;其中,根据所述第一稠密向量,获得所述乐器数字接口嵌入向量,包括:
通过所述第二稠密神经网络对所述第一稠密向量进行处理,获得第二稠密向量;
通过所述正向门递归神经网络对所述第二稠密向量进行处理,获得第一特征图;
通过所述反向门递归神经网络对所述第一特征图进行处理,获得第二特征图;
级联所述第二特征图和所述第二稠密向量,获得所述乐器数字接口嵌入向量。
6. 根据权利要求3所述的方法,其特征在于,所述时值编码器包括第三稠密神经网络;其中,通过所述时值编码器对所述时值特征进行处理,获得时值嵌入向量,包括:
通过所述第三稠密神经网络对所述时值特征进行处理,获得所述时值嵌入向量。
7. 根据权利要求3所述的方法,其特征在于,所述端到端神经网络模型还包括注意力机制模块和频谱解码器;其中,根据所述融合嵌入向量,获得所述频谱信息,包括:
通过所述注意力机制模块对所述融合嵌入向量进行处理,获得注意力上下文向量;
通过所述频谱解码器对所述注意力上下文向量进行处理,获得所述频谱信息。

8. 根据权利要求1所述的方法, 其特征在于, 所述频谱信息包括梅尔谱参数和线性谱参数; 其中, 根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频, 包括:

通过神经网络声码器对所述梅尔谱参数和所述线性谱参数进行处理, 合成所述歌唱音频。

9. 根据权利要求1所述的方法, 其特征在于, 还包括:

获取样本乐谱、样本歌词及其样本歌唱音频;

从所述样本乐谱中提取到样本音乐特征;

从所述样本歌词中提取到样本文本特征;

根据所述样本歌唱音频获得样本频谱信息;

通过所述端到端神经网络模型对所述样本音乐特征和所述样本文本特征进行处理, 获得预测频谱信息;

根据所述样本频谱信息与所述预测频谱信息, 训练所述端到端神经网络模型。

10. 一种音频合成装置, 其特征在于, 包括:

乐谱歌词获取模块, 配置为获取待处理乐谱及其待处理歌词;

音乐特征提取模块, 配置为从所述待处理乐谱中提取到音乐特征;

文本特征提取模块, 配置为从所述待处理歌词中提取到文本特征;

频谱信息获得模块, 配置为通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理, 获得频谱信息;

歌唱音频合成模块, 配置为根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。

11. 一种电子设备, 其特征在于, 包括:

一个或多个处理器;

存储装置, 配置为存储一个或多个程序, 当所述一个或多个程序被所述一个或多个处理器执行时, 使得所述一个或多个处理器实现如权利要求1至9中任一项所述的音频合成方法。

12. 一种计算机可读存储介质, 所述计算机可读存储介质存储有计算机程序, 其特征在于, 所述计算机程序被处理器执行时实现如权利要求1至9中任一项所述的音频合成方法。

音频合成方法及装置、电子设备和计算机可读存储介质

技术领域

[0001] 本公开涉及计算机技术领域,具体而言,涉及一种音频合成方法及装置、电子设备和计算机可读存储介质。

背景技术

[0002] 现有歌唱合成技术方案中,采用DNN(Deep Neural Network,深度神经网络)进行建模,合成的歌唱干音自然度较低且音质不高,离真人歌唱水平较远。

[0003] 因此,需要一种新的音频合成方法及装置、电子设备和计算机可读存储介质。

[0004] 需要说明的是,在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解。

发明内容

[0005] 本公开实施例提供一种音频合成方法及装置、电子设备和计算机可读存储介质,能够通过端到端神经网络模型实现歌唱音频的合成,合成的歌唱音频自然度及音质较高,更接近真人歌唱水平。

[0006] 本公开的其他特性和优点将通过下面的详细描述变得显然,或部分地通过本公开的实践而习得。

[0007] 根据本公开实施例的一方面,提供一种音频合成方法,所述方法包括:获取待处理乐谱及其待处理歌词;从所述待处理乐谱中提取到音乐特征;从所述待处理歌词中提取到文本特征;通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息;根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。

[0008] 根据本公开实施例的一方面,提供一种音频合成装置,所述装置包括:乐谱歌词获取模块,配置为获取待处理乐谱及其待处理歌词;音乐特征提取模块,配置为从所述待处理乐谱中提取到音乐特征;文本特征提取模块,配置为从所述待处理歌词中提取到文本特征;频谱信息获得模块,配置为通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息;歌唱音频合成模块,配置为根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。

[0009] 在本公开的一些示例性实施例中,所述音乐特征包括乐器数字接口特征和时值特征。其中,所述音乐特征提取模块包括:乐器数字接口特征获得子模块,配置为根据所述待处理乐谱中的音高,获得所述乐器数字接口特征;时值特征获得子模块,配置为对所述待处理乐谱中的音符长度进行归一化处理,获得所述时值特征。

[0010] 在本公开的一些示例性实施例中,所述端到端神经网络模型包括文本编码器、乐器数字接口编码器和时值编码器。其中,所述频谱信息获得模块包括:文本嵌入向量获得子模块,配置为通过所述文本编码器对所述文本特征进行处理,获得文本嵌入向量;乐器数字接口嵌入向量获得子模块,配置为通过所述乐器数字接口编码器对所述乐器数字接口特征进行处理,获得乐器数字接口嵌入向量;时值嵌入向量获得子模块,配置为通过所述时值编

码器对所述时值特征进行处理,获得时值嵌入向量;融合嵌入向量获得子模块,配置为根据所述文本嵌入向量、所述乐器数字接口嵌入向量和所述时值嵌入向量,获得融合嵌入向量;频谱信息获得子模块,配置为根据所述融合嵌入向量,获得所述频谱信息。

[0011] 在本公开的一些示例性实施例中,所述乐器数字接口编码器包括第一稠密神经网络。其中,所述乐器数字接口嵌入向量获得子模块包括:第一稠密向量获得单元,配置为通过所述第一稠密神经网络对所述乐器数字接口特征进行处理,获得第一稠密向量;乐器数字接口嵌入向量获得单元,配置为根据所述第一稠密向量,获得所述乐器数字接口嵌入向量。

[0012] 在本公开的一些示例性实施例中,所述乐器数字接口编码器还包括第二稠密神经网络、正向门递归神经网络和反向门递归神经网络。其中,所述乐器数字接口嵌入向量获得单元包括:第二稠密向量获得子单元,配置为通过所述第二稠密神经网络对所述第一稠密向量进行处理,获得第二稠密向量;第一特征图获得子单元,配置为通过所述正向门递归神经网络对所述第二稠密向量进行处理,获得第一特征图;第二特征图获得子单元,配置为通过所述反向门递归神经网络对所述第一特征图进行处理,获得第二特征图;乐器数字接口嵌入向量获得子单元,配置为级联所述第二特征图和所述第二稠密向量,获得所述乐器数字接口嵌入向量。

[0013] 在本公开的一些示例性实施例中,所述时值编码器包括第三稠密神经网络。其中,所述时值嵌入向量获得子模块包括:时值嵌入向量获得单元,配置为通过所述第三稠密神经网络对所述时值特征进行处理,获得所述时值嵌入向量。

[0014] 在本公开的一些示例性实施例中,所述端到端神经网络模型还包括注意力机制模块和频谱解码器。其中,所述频谱信息获得子模块包括:注意力上下文向量获得单元,配置为通过所述注意力机制模块对所述融合嵌入向量进行处理,获得注意力上下文向量;频谱信息获得单元,配置为通过所述频谱解码器对所述注意力上下文向量进行处理,获得所述频谱信息。

[0015] 在本公开的一些示例性实施例中,所述频谱信息包括梅尔谱参数和线性谱参数。其中,所述歌唱音频合成模块包括:歌唱音频合成子模块,配置为通过神经网络声码器对所述梅尔谱参数和所述线性谱参数进行处理,合成所述歌唱音频。

[0016] 在本公开的一些示例性实施例中,所述装置还包括:样本信息获取模块,配置为获取样本乐谱、样本歌词及其样本歌唱音频;样本音乐特征提取模块,配置为从所述样本乐谱中提取到样本音乐特征;样本文本特征提取模块,配置为从所述样本歌词中提取到样本文本特征;样本频谱信息获得模块,配置为根据所述样本歌唱音频获得样本频谱信息;频谱预测模块,配置为通过所述端到端神经网络模型对所述样本音乐特征和所述样本文本特征进行处理,获得预测频谱信息;模型训练模块,配置为根据所述样本频谱信息与所述预测频谱信息,训练所述端到端神经网络模型。

[0017] 根据本公开实施例的一方面,提供了一种计算机可读存储介质,其上存储有计算机程序,所述程序被处理器执行时实现如上述实施例中所述的音频合成方法。

[0018] 根据本公开实施例的一方面,提供了一种电子设备,包括:一个或多个处理器;存储装置,配置为存储一个或多个程序,当所述一个或多个程序被所述一个或多个处理器执行时,使得所述一个或多个处理器实现如上述实施例中所述的音频合成方法。

[0019] 在本公开的一些实施例所提供的技术方案中,通过从待处理乐谱中提取到音乐特征,从待处理歌词中提取到文本特征,进而通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息,从而可以根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。采用端到端神经网络模型合成的歌唱音频,相较于DNN歌唱合成方案,合成的歌唱音频自然度更高、音质更好,更接近真人歌唱水平。

[0020] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

[0021] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理。显而易见地,下面描述中的附图仅仅是本公开的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。在附图中:

[0022] 图1示出了可以应用本公开实施例的音频合成方法或音频合成装置的示例性系统架构的示意图;

[0023] 图2示出了适于用来实现本公开实施例的电子设备的计算机系统的结构示意图;

[0024] 图3示出了相关技术中采用DNN实现歌唱干音合成的示意图;

[0025] 图4示意性示出了根据本公开的一实施例的音频合成方法的流程图;

[0026] 图5示出了图4中所示的步骤S2在一实施例中的处理过程示意图;

[0027] 图6示出了图4中所示的步骤S4在一实施例中的处理过程示意图;

[0028] 图7示出了图4中所示的步骤S4在另一实施例中的处理过程示意图;

[0029] 图8示意性示出了根据本公开的一实施例的端到端神经网络模型的结构示意图;

[0030] 图9示出了图6中所示的步骤S42在一实施例中的处理过程示意图;

[0031] 图10示意性示出了根据本公开的一实施例的MIDI编码器的结构示意图;

[0032] 图11示出了图9中所示的步骤S422在一实施例中的处理过程示意图;

[0033] 图12示意性示出了根据本公开的另一实施例的MIDI编码器的结构示意图;

[0034] 图13示意性示出了根据本公开的一实施例的时值编码器的结构示意图;

[0035] 图14示意性示出了根据本公开的另一实施例的音频合成方法的流程图;

[0036] 图15示意性示出了根据本公开的一实施例的端到端神经网络模型的训练过程示意图;

[0037] 图16示意性示出了根据本公开的一实施例的端到端神经网络模型的预测过程示意图;

[0038] 图17示意性示出了应用不同的方式进行歌唱合成的效果评测示意图;

[0039] 图18-21示意性示出了应用本公开实施例提出的歌唱合成方法的应用场景示意图;

[0040] 图22示意性示出了根据本公开的一实施例的音频合成装置的框图。

具体实施方式

[0041] 现在将参考附图更全面地描述示例实施方式。然而,示例实施方式能够以多种形

式实施,且不应被理解为限于在此阐述的范例;相反,提供这些实施方式使得本公开将更加全面和完整,并将示例实施方式的构思全面地传达给本领域的技术人员。

[0042] 此外,所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中,提供许多具体细节从而给出对本公开的实施例的充分理解。然而,本领域技术人员将意识到,可以实践本公开的技术方案而没有特定细节中的一个或更多,或者可以采用其它的方法、组元、装置、步骤等。在其它情况下,不详细示出或描述公知方法、装置、实现或者操作以避免模糊本公开的各方面。

[0043] 附图中所示的方框图仅仅是功能实体,不一定必须与物理上独立的实体相对应。即,可以采用软件形式来实现这些功能实体,或在一个或多个硬件模块或集成电路中实现这些功能实体,或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0044] 附图中所示的流程图仅是示例性说明,不是必须包括所有的内容和操作/步骤,也不是必须按所描述的顺序执行。例如,有的操作/步骤还可以分解,而有的操作/步骤可以合并或部分合并,因此实际执行的顺序有可能根据实际情况改变。

[0045] 图1示出了可以应用本公开实施例的音频合成方法或音频合成装置的示例性系统架构100的示意图。

[0046] 如图1所示,系统架构100可以包括终端设备101、102、103中的一种或多种,网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0047] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的数目。比如服务器105可以是多个服务器组成的服务器集群等。

[0048] 用户可以使用终端设备101、102、103通过网络104与服务器105交互,以接收或发送消息等。终端设备101、102、103可以是具有显示屏的各种电子设备,包括但不限于智能手机、平板电脑、便携式计算机、可穿戴智能设备、智能家居设备和台式计算机、数字电影放映机等等。

[0049] 服务器105可以是提供各种服务的服务器。例如用户利用终端设备103(也可以是终端设备101或102)向服务器105发送各种请求。服务器105可以基于该请求中携带的相关信息,获取响应于所述请求的反馈信息给终端设备103,进而用户可以在终端设备103上查看显示的反馈信息。

[0050] 又如终端设备103(也可以是终端设备101或102)可以是智能电视、VR(Virtual Reality,虚拟现实)/AR(Augmented Reality,增强现实)头盔显示器、或者其上安装有即时通讯、视频应用程序(application,APP)等的移动终端例如智能手机、平板电脑等,用户可以通过该智能电视、VR/AR头盔显示器或者该即时通讯、视频APP向服务器105发送各种请求。服务器105可以基于该请求,获取响应于所述请求的反馈信息返回给该智能电视、VR/AR头盔显示器或者该即时通讯、视频APP,进而通过该智能电视、VR/AR头盔显示器或者该即时通讯、视频APP将返回的反馈信息显示。

[0051] 图2示出了适于用来实现本公开实施例的电子设备的计算机系统的结构示意图。

[0052] 需要说明的是,图2示出的电子设备的计算机系统200仅是一个示例,不应对本公开实施例的功能和使用范围带来任何限制。

[0053] 如图2所示,计算机系统200包括中央处理单元(CPU,Central Processing Unit) 201,其可以根据存储在只读存储器(ROM,Read-Only Memory) 202中的程序或者从储存部分208加载到随机访问存储器(RAM,Random Access Memory) 203中的程序而执行各种适当的动作和处理。在RAM 203中,还存储有系统操作所需的各种程序和数据。CPU 201、ROM 202以及RAM 203通过总线204彼此相连。输入/输出(I/O)接口205也连接至总线204。

[0054] 以下部件连接至I/O接口205:包括键盘、鼠标等的输入部分206;包括诸如阴极射线管(CRT,Cathode Ray Tube)、液晶显示器(LCD,Liquid Crystal Display)等以及扬声器等的输出部分207;包括硬盘等的储存部分208;以及包括诸如LAN(Local Area Network,局域网)卡、调制解调器等网络接口卡的通信部分209。通信部分209经由诸如因特网的网络执行通信处理。驱动器210也根据需要连接至I/O接口205。可拆卸介质211,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器210上,以便于从其上读出的计算机程序根据需要被安装入储存部分208。

[0055] 特别地,根据本公开的实施例,下文参考流程图描述的过程可以被实现为计算机软件程序。例如,本公开的实施例包括一种计算机程序产品,其包括承载在计算机可读存储介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分209从网络上被下载和安装,和/或从可拆卸介质211被安装。在该计算机程序被中央处理单元(CPU) 201执行时,执行本申请的方法和/或装置中限定的各种功能。

[0056] 需要说明的是,本公开所示的计算机可读存储介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM(Erasable Programmable Read Only Memory,可擦除可编程只读存储器)或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读存储介质,该计算机可读存储介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读存储介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF(Radio Frequency,射频)等等,或者上述的任意合适的组合。

[0057] 附图中的流程图和框图,图示了按照本公开各种实施例的方法、装置和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际

上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0058] 描述于本公开实施例中所涉及到的模块、子模块、单元和子单元可以通过软件的方式实现,也可以通过硬件的方式来实现,所描述的模块、子模块、单元和子单元也可以设置在处理器中。其中,这些模块、子模块、单元和子单元的名称在某种情况下并不构成对该模块、子模块、单元和子单元本身的限定。

[0059] 作为另一方面,本申请还提供了一种计算机可读存储介质,该计算机可读存储介质可以是上述实施例中描述的电子设备中所包含的;也可以是单独存在,而未装配入该电子设备中。上述计算机可读存储介质承载有一个或者多个程序,当上述一个或者多个程序被一个该电子设备执行时,使得该电子设备实现如下述实施例中所述的方法。例如,所述的电子设备可以实现如图4或图5或图6或图7或图9或图11或图14所示的各个步骤。

[0060] 人工智能(Artificial Intelligence, AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是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。

[0061] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0062] 语音技术(Speech Technology)的关键技术有自动语音识别技术(Automatic Speech Recognition, ASR)和语音合成技术(Text to Speech, TTS)以及声纹识别技术。让计算机能听、能看、能说、能感觉,是未来人机交互的发展方向,其中语音成为未来最被看好的人机交互方式之一。

[0063] 机器学习(Machine Learning, ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0064] 随着人工智能技术研究和进步,人工智能技术在多个领域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,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。

[0065] 本申请实施例提供的方案涉及人工智能的语音技术和机器学习等技术,具体通过如下实施例进行说明:

[0066] 近年来,歌唱合成技术一直备受社会各界的关注,歌唱合成技术最大的便利性在

于其可以让计算机唱出任何旋律的歌曲,这使得与歌唱紧密相关的音乐制作、娱乐等领域对歌唱合成技术的进步有着迫切的期待。

[0067] 相关技术中,歌唱合成的技术有基于隐马尔科夫模型 (Hidden Markov Model, HMM) 的参数合成系统以及基于波形拼接合成的系统。

[0068] 波形拼接的核心是预先录制某种语言中每个发音在不同音高的唱法,然后按照歌词和乐谱将这些录音连接起来。然而基于波形拼接合成的系统存在两大难点,一是在波形拼接的过程中极易产生波形扭曲,进而造成合成的声音不自然;二是波形拼接依赖非常庞大的录音数据,这就需要耗费大量的时间和人力来收集数据,从而导致歌唱合成的难度比较高。

[0069] 基于参数合成的歌唱合成系统首先分别确定各基本合成单元(如音节,音素等)的时长参数序列、基频参数序列以及频谱参数序列,随后再根据各参数序列利用参数合成器得到连续歌唱信号。

[0070] 图3示出了相关技术中采用DNN实现歌唱干音合成的示意图。

[0071] 如图3所示,相关技术中的歌唱合成技术方案采用DNN进行建模,该技术方案分为模型训练和预测两个部分。在模型训练部分,使用采集到的歌唱干音、乐谱信息和歌词信息,训练时长模型和声学模型。其中,时长模型用于预测每个音素的发音时长,声学模型用于预测音素的频谱参数。在模型预测部分,使用训练得到的声学模型和时长模型,根据用户输入的乐谱和歌词合成歌唱干音。这里的音素(phone)是从音色的角度划分出来的最小语音单位,例如汉语中的声母和韵母,英语中的辅音和元音等。

[0072] 具体的,在模型训练阶段,包括以下步骤:

[0073] 1.1、准备数据:从歌唱数据库中采集歌手的歌唱干音、乐谱信息和歌词信息。

[0074] 1.2、标注数据:将歌唱音频即上述采集的歌唱干音按句切开。人工标记每一句的字边界、音素边界、音符、音高、时值等信息。将标注数据作为训练数据。

[0075] 1.3、HTS (HMM-based speech synthesis system,基于隐马尔科夫模型的语音合成系统) 切分:使用上述步骤1.2中得到的训练数据训练HTS系统,然后将每个音素进一步切分到状态级别。

[0076] 具体的,预先构建一个HTS系统,然后将训练数据中的歌唱干音的每个句子分别输入至该HTS系统,利用该HTS系统预测每一句的字边界、音素边界、音符、音高、时值等信息,并与上述人工标记的每一句的字边界、音素边界、音符、音高、时值等信息进行比较,计算损失函数,训练该HTS系统。然后,使用这个HTS系统来进一步切分每个音素。

[0077] 例如“a”这个音,人工标记的是一整个完整的“a”音,但实际上,再根据频谱的特征还可以进一步切分成数份(即数个状态),在实际训练模型时,是预测状态级别的时长,不是直接预测音素的时长。HTS就是用来将每个音素切成5个状态,然后输出每个音素每个状态的长度。

[0078] 1.4、提取时长模型文本特征:从乐谱和歌词中提取时长模型文本特征。

[0079] 1.5、提取音素时长参数:从上述步骤1.3中得到的HTS切分结果中提取时长参数。这里利用音素状态级别的时长进行训练。

[0080] 1.6、训练时长模型:使用上述步骤1.4提取的时长模型文本特征和上述步骤1.5提取的音素时长参数,训练时长模型。

[0081] 1.7、提取歌唱音频声学参数:将上述步骤1.2中切分的每一句音频进一步切分成等长的N片,N为大于或等于1的正整数,每一片小音频都提取能表达音频频谱特征的声学参数和与音高相关的基频、颤音参数序列。

[0082] 1.8、提取声学模型文本特征:结合上述步骤1.4和1.5提取到的数据,通过序列扩充的手段,生成和音频长度一样的声学模型文本特征。

[0083] 1.9、训练声学模型:使用上述步骤1.7和1.8提取到的数据,训练声学模型。

[0084] 在模型预测阶段,包括以下步骤:

[0085] 2.1、提取用户提供的乐谱和歌词信息。

[0086] 2.2、从乐谱和歌词信息中提取时长模型文本特征序列。

[0087] 2.3、使用上述步骤2.2得到的时长模型文本特征序列,输入时长模型,预测每个音素的时长。

[0088] 2.4、结合上述步骤2.2得到的时长模型文本特征序列和上述步骤2.3预测的音素时长,通过序列扩充的手段,得到声学模型文本特征序列。

[0089] 2.5、使用上述步骤2.4得到的声学模型文本特征序列,输入声学模型,预测每一小片音频的声学参数、基频和颤音参数序列。

[0090] 2.6、使用上述步骤2.5得到的声学参数、基频和颤音参数序列,输入声码器中,获得合成的歌唱干音频。

[0091] 但是,上述相关技术中,采用DNN建模进行歌唱合成的歌唱干音的音素音频过渡不太协调,音色有机械感,不自然,整体听感上不“好听”,离真人歌唱水平相差较远。

[0092] 相关技术中,端到端语音合成技术是TTS技术,其是文本到语音的合成技术,但是合成的音频是人正常说话的语音,目前没有利用端到端技术成功的合成音高和时值都符合乐谱要求、高自然度的歌唱干音。

[0093] 图4示意性示出了根据本公开的一实施例的音频合成方法的流程图。本公开实施例提供的方法可以由任意具备计算处理能力的电子设备执行,例如如图1中的终端设备101、102、103中的任意一个或多个和/或服务器105。在下面的举例说明中,以服务器105为执行主体进行示例说明。

[0094] 如图4所示,本公开实施例提供的音频合成方法可以包括以下步骤。

[0095] 在步骤S1中,获取待处理乐谱及其待处理歌词。

[0096] 例如,用户通过客户端向服务器发送歌唱合成请求,该歌唱合成请求中携带有用用户所选择的当前待合成的歌唱音频对应的待处理乐谱及其待处理歌词信息,服务器接收到该歌唱合成请求后,从中获取到所述待处理乐谱及其所述待处理歌词。

[0097] 本公开实施例中,所述待处理乐谱可以是MIDI (Musical Instrument Digital Interface,乐器数字接口) 格式、ASCII (American Standard Code for Information Interchange,美国信息交换标准代码) 格式、MusicXML (Music Extensible Markup Language,音乐扩展标记语言) 格式等等中的任意一种。所述待处理乐谱可以为纸质乐谱和/或电子乐谱,所述待处理歌词与所述待处理乐谱对应,其也可以是纸质歌词和/或电子歌词。纸质乐谱和纸质歌词可以分别是印刷的乐谱和歌词,还可以是手写的乐谱和歌词,纸质乐谱和歌词可以通过扫描的方式分别转换为电子乐谱和歌词,再通过软件处理获得所需数字格式的乐谱和歌词;该电子乐谱和歌词可以为从网站上下载的电子乐谱和歌词,还可

以是用户自行编辑的电子乐谱和歌词,通过软件处理后获得所需数字格式的乐谱和歌词。

[0098] 在步骤S2中,从所述待处理乐谱中提取到音乐特征。

[0099] 本公开实施例中,所述音乐特征可以包括和音符音高相关的MIDI特征和与时值相关的时值(Interval)特征。具体的提取过程可以参照下图5实施例的描述。

[0100] 在步骤S3中,从所述待处理歌词中提取到文本特征。

[0101] 本公开实施例中,从所述待处理歌词中提取的文本特征可以包括:音素特征、韵律边界特征、是否乐句结束、当前拼音有多少音素、当前音素是否儿化音等中的任意一项或者多项的组合。

[0102] 例如,以待处理歌词为“拼音”为例,首先将文本转化为其对应的拼音“ping yin”,可以拆成“p”、“ing”、“y”和“in”四个序列,下面以当前拼音有多少个音子(特征2,相应的数字代表相应个数的音子)、当前拼音是不是儿化音(特征3,“0”代表不是儿化音,“1”代表是儿化音)、当前拼音是不是零声母音节(特征4,“0”代表不是零声母音节,“1”代表是零声母音节)这几个特征为例,对文本特征的生成过程进行说明:

[0103] 表1

[0104]	特征 1	特征 2	特征 3	特征 4
	p	2	0	0
[0105]	ing	2	0	0
	y	2	0	0
	in	2	0	0

[0106] 其中,零声母音节是指那些没有辅音声母,只有韵母构成的音节,实际上以元音(a、o、e、i、u、ü)开头的音节都是零声母音节。由此,合成的文本特征为:“p-2-0-0”、“ing-2-0-0”、“y-2-0-0”和“in-2-0-0”。

[0107] 应当理解的是,上述例举的特征仅用于举例说明,实际上可能包括数百维特征。

[0108] 本公开实施例中,音乐特征中的一个或者多个音符与文本特征中的一个歌词对应。音乐特征中的音符还具有时序性,即乐谱中的音符需要按照既定的时序演奏才能形成既定的旋律,文本特征中的歌词是与音符对应的,因此歌词对应也具有既定的时序。

[0109] 在步骤S4中,通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息。

[0110] 在步骤S5中,根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。

[0111] 具体的,可以将频谱信息输入至预先训练好的神经网络声码器中,自动输出与所述待处理乐谱及其待处理歌词相应的歌唱音频的采样信号,例如可以参照下图8有关神经网络声码器的内容。

[0112] 本公开实施例中,所述歌唱音频可以是歌唱干音,也可以最终合成的歌曲,以将其返回至发送歌唱合成请求的用户的客户端进行播放。其中,干音是指没有经过任何后期处理(混响、调制、压限、变速等操作)的原始歌唱音频。与之对应的是湿音,指的是经过后期处理的歌唱音频。当根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱干

音后,可以歌唱干音与相应的背景音乐合成得到最终合成的歌曲。

[0113] 本公开实施方式提供的音频合成方法,通过待处理乐谱中提取到音乐特征,从待处理歌词中提取到文本特征,进而通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息,从而可以根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。一方面,由于端到端神经网络模型是直接从文本到语音的建模,因此,不需要经过DNN建模中的时长模型,不再需要HTS系统来做数据预处理,本公开实施例提供的方案可以直接使用音素的长度,不需要进一步切分至音素的状态级别;另一方面,采用端到端神经网络模型合成的歌唱音频,相较于DNN歌唱合成方案,合成的歌唱音频自然度更高、音质更好,更接近真人歌唱水平。

[0114] 图5示出了图4中所示的步骤S2在一实施例中的处理过程示意图。本公开实施例中,所述音乐特征可以包括乐器数字接口特征和时值特征。

[0115] 如图5所示,本公开实施例中,上述步骤S2可以进一步包括以下步骤。

[0116] 在步骤S21中,根据所述待处理乐谱中的音高,获得所述乐器数字接口特征。

[0117] 本公开实施例中,所述MIDI特征是字级别的,歌词中的一个字对应一个MIDI。音高是指唱歌时声带的震动频率。可以将所述待处理乐谱中的音高 f 按照以下公式进行转换,获得MIDI特征 p :

$$[0118] \quad p = 69 + 12 \times \log_2 \left(\frac{f}{440} \right) \quad (1)$$

[0119] 例如,一首歌的简谱如下:

$$[0120] \quad 1 = {}^b B \quad \frac{6}{8}$$

$$[0121] \quad \text{♩} = 234$$

$$[0122] \quad \underline{6 \ 1 \ 3 \ 6 \ 1} \uparrow$$

你还记得吗

[0123] 上述谱子是以 ${}^b B$ 调为初始音高,即以233.08Hz为1音的频率,查下表2可以得到上述简谱对应的频率依次是196.00Hz、233.08Hz、293.66Hz、195.00Hz和456.16Hz,然后上述公式(1),转化为MIDI(取整)为55、58、62、55和70。

[0124] 表2音高与频率对照表

[0125]

频率, 单位为赫兹, 括号内为距离中央 C (261.63 赫兹) 的半音距离										
八度	0	1	2	3	4	5	6	7	8	9
→										
音名										
↓										

[0126]

C	16.352 (-48)	32.703 (-36)	65.406 (-24)	130.81 (-12)	261.63 (0)	523.25 (+12)	1046.5 (+24)	2093.0 (+36)	4186.0 (+48)	8372.0 (+60)
C [#] /D ^b	17.324 (-47)	34.648 (-35)	69.296 (-23)	138.59 (-11)	277.18 (+1)	554.37 (+13)	1108.7 (+25)	2217.5 (+37)	4434.9 (+49)	8869.8 (+61)
D	18.354 (-46)	36.708 (-34)	73.416 (-22)	146.83 (-10)	293.66 (+2)	587.33 (+14)	1174.7 (+26)	2349.3 (+38)	4698.6 (+50)	9397.3 (+62)
D [#] / ^b E	19.445 (-45)	38.891 (-33)	77.782 (-21)	155.56 (-9)	311.13 (+3)	622.25 (+15)	1244.5 (+27)	2489.0 (+39)	4978.0 (+51)	9956.1 (+63)
E	20.602 (-44)	41.203 (-32)	82.407 (-20)	164.81 (-8)	329.63 (+4)	659.26 (+16)	1318.5 (+28)	2637.0 (+40)	5274.0 (+52)	10548 (+64)
F	21.827 (-43)	43.654 (-31)	87.307 (-19)	174.61 (-7)	349.23 (+5)	698.46 (+17)	1396.9 (+29)	2793.8 (+41)	5587.7 (+53)	11175 (+65)
F [#] /G ^b	23.125 (-42)	46.249 (-30)	92.499 (-18)	185.00 (-6)	369.99 (+6)	739.99 (+18)	1480.0 (+30)	2960.0 (+42)	5919.9 (+54)	11840 (+66)
G	24.500 (-41)	48.999 (-29)	97.999 (-17)	196.00 (-5)	392.00 (+7)	783.99 (+19)	1568.0 (+31)	3136.0 (+43)	6271.9 (+55)	12544 (+67)
G [#] /A ^b	25.957 (-40)	51.913 (-28)	103.83 (-16)	207.65 (-4)	415.30 (+8)	830.61 (+20)	1661.2 (+32)	3322.4 (+44)	6644.9 (+56)	13290 (+68)
A	27.500 (-39)	55.000 (-27)	110.00 (-15)	220.00 (-3)	440.00 (+9)	880.00 (+21)	1760.0 (+33)	3520.0 (+45)	7040.0 (+57)	14080 (+69)
A [#] / ^b B	29.135 (-38)	58.270 (-26)	116.54 (-14)	233.08 (-2)	466.16 (+10)	932.33 (+22)	1864.7 (+34)	3729.3 (+46)	7458.6 (+58)	14917 (+70)
B	30.868 (-37)	61.735 (-25)	123.47 (-13)	246.94 (-1)	493.88 (+11)	987.77 (+23)	1975.5 (+35)	3951.1 (+47)	7902.1 (+59)	15804 (+71)

[0127] 在步骤S22中,对所述待处理乐谱中的音符长度进行归一化处理,获得所述时值特征。

[0128] 本公开实施例中,Interval是音符的时值。和时长Duration不是同一个概念,这里的时值是乐谱上标明的应该唱的长度。但是人的发音不可能和乐谱标明的完全一样。Duration是指人在实际发音时,每个音素(如a、ao等)的发音时间长度。

[0129] 可以根据以下公式将乐谱中的时值转化成所述时值特征interval_feature:

$$[0130] \quad \text{interval_feature} = \frac{\text{乐谱音符的理论长度 (毫秒)}}{1500} \quad (2)$$

[0131] 还是以上述简谱为例,谱子中指明一分钟有234拍,也就是一拍的长度为 $60\text{s}/234 = 0.256\text{s} = 256\text{ms}$,上述简谱序列对应的是一拍(256ms)、半拍(128ms)、半拍(128ms)、半拍(128ms)、半拍(128ms),所以interval_feature的值对应为: $256/1500 = 0.1707$, $128/1500 = 0.0853$, $128/1500 = 0.0853$, $128/1500 = 0.0853$, $128/1500 = 0.0853$ 。

[0132] 本公开实施例中,将上述提取到的乐器数字接口特征和时值特征作为所述音乐特征,由于所述音乐特征中的时值特征没有将乐谱中的时值进行离散化,例如转化成“0”和“1”代表的序列,而是将其通过上述公式(2)进行归一化处理,由于归一化处理的结果比离散化处理的效果要好,从而使得提取到的音乐特征较好。

[0133] 图6示出了图4中所示的步骤S4在一实施例中的处理过程示意图。本公开实施例中,所述端到端神经网络模型可以包括文本编码器(Text encoder)、乐器数字接口编码器(MIDI编码器)和时值编码器(Interval encoder)。

[0134] 如图6所示,本公开实施例中,上述步骤S4可以进一步包括以下步骤。

[0135] 在步骤S41中,通过所述文本编码器对所述文本特征进行处理,获得文本嵌入(Embedding)向量。

[0136] 其中,Text encoder用于将音素级别的文本转化成向量。具体实现过程例如可以参照下图8实施例中的文本编码器,用于将字符序列转换成隐特征表示,输入文本(input text)例如上述文本特征至文本编码器,这里的文本编码器例如可以包括依次连接的字符嵌入(Character Embedding)层、3卷积层(3 Conv Layers)和双向LSTM(Bidirectional LSTM(Long Short-Term Memory,长短期记忆神经网络)),字符嵌入层用于将文本特征转化为512维的字符嵌入向量表示,然后被输入至3个卷积层,每个卷积层包括512滤波器(形状5乘以1),例如每个滤波器跨越5个字符,后面接着批归一化(batch normalization,BN)和ReLU(Rectified Linear Unit,线性整流函数)函数,再将最后一个卷积层的输出结果输入至一个双向LSTM,其包括512单元(每个方向256个)以用于输出文本嵌入向量。

[0137] 在步骤S42中,通过所述乐器数字接口编码器对所述乐器数字接口特征进行处理,获得乐器数字接口嵌入(MIDI Embedding)向量。

[0138] 例如可以参照下图8实施例中的MIDI编码器,将上述提取的MIDI特征输入至MIDI编码器,输出MIDI Embedding向量。MIDI编码器的具体结构可以参照下图9至12所示实施例的描述。

[0139] 在步骤S43中,通过所述时值编码器对所述时值特征进行处理,获得时值嵌入(Interval Embedding)向量。

[0140] 例如可以参照下图8实施例中的时值编码器,将上述提取的时值特征输入至时值编码器,输出Interval Embedding向量。时值编码器的具体结构可以参照下图13所示实施例的描述。

[0141] 在步骤S44中,根据所述文本嵌入向量、所述乐器数字接口嵌入向量和所述时值嵌入向量,获得融合嵌入向量。

[0142] 例如可以参照下图8实施例,将文本编码器的双向LSTM、MIDI编码器和时值编码器三者的输出结果进行拼接,获得融合嵌入向量。

[0143] 在步骤S45中,根据所述融合嵌入向量,获得所述频谱信息。

[0144] 这里所述频谱信息可以包括梅尔谱和/或线性谱。梅尔谱可以经过转换变成线性谱。例如可以参照下图8实施例,将所述融合嵌入向量输入至基于位置敏感的注意力机制(Location Sensitive Attention)模块,生成注意力上下文向量,再输入至梅尔谱解码器(Mel Decoder)以生成梅尔谱(Mel Spectrogram)。

[0145] 图7示出了图4中所示的步骤S4在另一实施例中的处理过程示意图。本公开实施例

中,所述端到端神经网络模型还可以包括注意力机制(Attention)模块和频谱解码器。

[0146] 如图7所示,本公开实施例中,上述步骤S4可以进一步包括以下步骤。

[0147] 在步骤S46中,通过所述注意力机制模块对所述融合嵌入向量进行处理,获得注意力上下文向量。

[0148] 例如可以参照下图8实施例,使用位置敏感的注意力机制,扩展增加的注意力机制至使用累计的注意力权重,将之前解码器时间步作为辅助特征。这样鼓励模型通过输入不断前进,减少解码器重复或忽略某些子序列的潜在故障模式。将输入和位置特征映射至128维隐特制表示后,计算注意力概率。位置特征使用32 1-D卷积滤波器计算,长度为31。

[0149] 在步骤S47中,通过所述频谱解码器对所述注意力上下文向量进行处理,获得所述频谱信息。

[0150] 例如可以参照下图8实施例,这里的频谱解码器为梅尔谱解码器,梅尔谱解码器包括5卷积层(5Conv Layer)Post-Net(将梅尔谱映射成线性谱的网络结构)、2层Pre-Net(2 Layer Pre-Net,2层预网络层)、2LSTM层和两个线性变换(Linear Projection)层,梅尔谱解码器是一个自回归递归神经网络,一次一帧从编码输入序列预测一个梅尔谱。从前一时间步的预测首先通过一个小的Pre-Net,其包括256个隐ReLU单元的2个全连接层。Pre-Net的输出和注意力上下文向量拼接并输入至2个单向LSTM层(有1024个单元)。LSTM输出和注意力上下文向量的级联通过线性变换映射到预测的目标频谱帧。最终,预测的梅尔谱被传输至一个5Conv Layer Post-Net,预测残差来相加以改善整体的重构。每个Post-Net包括512个滤波器(形状5乘以1)带有BN,后面接有tanh激活函数作为全部的最后一层。然后将梅尔谱输入至神经网络声码器(Neural Network Vocoder),输出音频采样(Audio Samples)即歌唱干音。

[0151] 图8示意性示出了根据本公开的一实施例的端到端神经网络模型的结构示意图。

[0152] 如图8所示,本公开实施例中的Mel Decoder,是一个自回归递归神经网络,结合Text encoder的输出,通过自回归的方式生成音频的梅尔谱。在输出Mel Spectrogram时,是递归来输出结果的。例如,一开始是使用1帧随机数结合上下文向量预测三帧的Mel Spectrogram,然后使用这三帧Mel Spectrogram的最后一帧作为输入结合上下文向量预测下一个三帧的Mel Spectrogram,如此循环,预测出所有的Mel Spectrogram。

[0153] 与频谱帧预测并行的,梅尔谱解码器LSTM输出与注意力上下文的级联被投射至一个标量,并传输至一个sigmoid激活函数以预测输出序列是否完成的概率。“停止令牌”(Stop Token)预测在推论期间被用来允许模型动态决定什么时间终止生成,而不是一直生成固定的时长。

[0154] 本公开实施例中,神经网络声码器可以采用WaveNet Vocoder,将Mel Spectrogram特征表示转化为时域波形样本。

[0155] 本公开实施例提供的方法,基于TTS的端到端合成技术,对模型结构做了修改,提出了适合于TTM(Text To Music,文本到歌唱的合成技术,合成的是唱歌的音频)的端到端合成技术MusicTactron。相较于DNN歌唱合成方案,本公开实施例合成的音频自然度更高、音质更好,非常接近于真人歌唱水平。

[0156] 需要说明的是,虽然图8中的端到端神经网络模型的模型结构,是以Tacotron2为例,添加了MIDI encoder和Interval encoder。但本公开并不限于此,可以对其他端到端

神经网络模型结构进行改进,例如,还可以对Tacotron1的结构进行改进,改造后的结构是把上述MIDI encoder和Interval encoder的输出结果和Tacotron1文本编码器的输出结果拼接在一起,输入到attention模块中。再例如,还可以对Deep Voice 3的结构进行改进,改造后的结构是把上述MIDI encoder、Interval encoder的输出结果和Tacotron1文本编码器的输出结果和Deep Voice 3文本编码器的输出结果拼接在一起,输入到AttentionBlock模块中。

[0157] 图9示出了图6中所示的步骤S42在一实施例中的处理过程示意图。本公开实施例中,所述乐器数字接口编码器可以包括第一稠密神经网络(Dense1,简写为D1)。

[0158] 如图9所示,本公开实施例中,上述步骤S42可以进一步包括以下步骤。

[0159] 在步骤S421中,通过所述第一稠密神经网络对所述乐器数字接口特征进行处理,获得第一稠密向量。

[0160] 在步骤S422中,根据所述第一稠密向量,获得所述乐器数字接口嵌入向量。

[0161] 图10示意性示出了根据本公开的一实施例的MIDI编码器的结构示意图。

[0162] 如图10所示,本公开实施例中,MIDI编码器包括一个第一稠密神经网络(D1),将上述提取的MIDI特征输入一层Dense1中,这层Dense1的节点数为1024,激活函数为ReLU,输出为MIDI嵌入向量。

[0163] 图11示出了图9中所示的步骤S422在一实施例中的处理过程示意图。本公开实施例中,所述乐器数字接口编码器还可以包括第二稠密神经网络、正向门递归神经网络(Recurrent Neural Network,RNN)和反向门递归神经网络。

[0164] 如图11所示,本公开实施例中,上述步骤S422可以进一步包括以下步骤。

[0165] 在步骤S4221中,通过所述第二稠密神经网络对所述第一稠密向量进行处理,获得第二稠密向量。

[0166] 本公开实施例中,可以将上述图10中的D1的计算结果输入下一层Dense2中,这层Dense2的节点数也可以为1024,激活函数为ReLU。

[0167] 在步骤S4222中,通过所述正向门递归神经网络对所述第二稠密向量进行处理,获得第一特征图。

[0168] 可以将上述步骤S4221的Dense2的计算结果输入至一层GRU1结构中,这层GRU1的节点数可以为128。

[0169] 在步骤S4223中,通过所述反向门递归神经网络对所述第一特征图进行处理,获得第二特征图。

[0170] 可以将上述步骤S4222的GRU1的计算结果输入下一层GRU2结构中,这层GRU2的节点数也可以为128。

[0171] 在步骤S4224中,级联所述第二特征图和所述第二稠密向量,获得所述乐器数字接口嵌入向量。

[0172] 可以将Dense2和GRU2的计算结果拼接在一起,作为MIDI Embedding向量,输出到注意力机制模块中。

[0173] 图12示意性示出了根据本公开的另一实施例的MIDI编码器的结构示意图。

[0174] 如图12所示,本公开实施例中,所述MIDI编码器的输入为MIDI特征,所述MIDI编码器可以包括一个第一稠密神经网络(Dense1,D1)、一个第二稠密神经网络(Dense2,D2)、一

个正向门递归神经网络 (GRU1) 和反向门递归神经网络 (GRU2) 以及一个级联层。D2的输出结果与GRU2的输出结果进行级联后,输出MIDI嵌入向量。

[0175] 图12实施例中,MIDI encoder中的两层Dense是MIDI embedding层,两个GRU组合成Bidirectional GRU(双向GRU)。通过级联Dense2和GRU2的计算结果来保证音高的稳定性,也就是同一个MIDI特征,Dense输出的部分是一样的。这里,级联为并联或结合的意思,第二特征图和所述第二稠密向量沿着通道维度进行拼接,重组为一个具有更大尺寸,包含更多特征信息的乐器数字接口嵌入向量。

[0176] 可以理解的是,所述MIDI encoder的网络结构并不限于上述图10和图12所列举的结构形式,例如还可以包括三层及以上的Dense,和/或包括三层及以上的GRU,GRU层和/或Dense层也可以用LSTM等其他递归神经网络或者任意的深度神经网络替代,本公开对此不做限定。

[0177] 在示例性实施例中,所述时值编码器可以包括第三稠密神经网络。其中,通过所述时值编码器对所述时值特征进行处理,获得时值嵌入向量,可以包括:通过所述第三稠密神经网络对所述时值特征进行处理,获得所述时值嵌入向量。

[0178] 图13示意性示出了根据本公开的一实施例的时值编码器的结构示意图。

[0179] 如图13所示,本公开实施例提供的时值编码器可以包括一个第三稠密神经网络(Dense3,D3),将上述提取的Interval特征输入一层Dense3中,这层Dense3的节点数可以为1024,激活函数可以为ReLU。将Dense3的计算结果作为Interval Embedding向量输入到注意力机制模块中。

[0180] 可以理解的是,所述时值编码器的网络结构并不限于图13的结构形式,例如其也可以是至少一层Dense叠加至少一层GRU的结构,本公开对此不做限定。

[0181] 图14示意性示出了根据本公开的另一实施例的音频合成方法的流程图。

[0182] 如图14所示,与上述实施例的不同之处在于,本公开实施例提供的方法还可以包括以下步骤。本公开实施例中,可以包括模型训练部分和模型预测部分。其中,模型训练部分包括以下步骤。

[0183] 在步骤S6中,获取样本乐谱、样本歌词及其样本歌唱音频。

[0184] 采集歌手的歌唱干音作为样本歌唱音频,以及该歌唱干音对应的乐谱和歌词分别作为样本乐谱和样本歌词。

[0185] 在步骤S7中,从所述样本乐谱中提取到样本音乐特征。

[0186] 将样本乐谱按句切开,标记每一句的音符、音高、时值等信息,根据音符音高转换获得MIDI特征,根据时值信息获取时值特征,所述MIDI特征和时值特征作为所述样本音乐特征。

[0187] 在步骤S8中,从所述样本歌词中提取到样本文本特征。

[0188] 将样本歌词按句切开,标记每一句的字边界和音素边界等信息。根据所述字边界、音素边界等信息,获得样本文本特征。

[0189] 在步骤S9中,根据所述样本歌唱音频获得样本频谱信息。

[0190] 使用频谱提取工具从所述样本歌唱音频中提取出样本频谱信息,例如所述样本频谱信息可以包括标注的真实的梅尔谱参数和线性谱参数。

[0191] 在步骤S10中,通过所述端到端神经网络模型对所述样本音乐特征和所述样本文

本特征进行处理,获得预测频谱信息。

[0192] 将所述样本音乐特征中的MIDI特征输入至MIDI编码器,将所述样本音乐特征中的时值特征输入至时值编码器,将所述样本文本特征输入至文本编码器,获得预测频谱信息。

[0193] 在步骤S11中,根据所述样本频谱信息与所述预测频谱信息,训练所述端到端神经网络模型。

[0194] 本公开实施例中,训练所述端到端神经网络模型时,可以采用以下损失函数:

[0195] $\text{loss} = \text{Mel_loss} + \text{Spec_loss}$ (3)

[0196] 上述公式中,loss代表总损失,Mel_loss代表梅尔谱损失,Spec_loss代表线性谱损失。

[0197] 其中,梅尔谱损失可以通过以下公式计算获得:

$$[0198] \quad \text{Mel_loss} = \sum_{i1=1}^{n1} (x_{i1} - y_{i1})^2 \quad (4)$$

[0199] 上述公式中,n1为梅尔谱帧数,n1为大于或等于1的正整数, y_{i1} 为标注的真实的第i1帧梅尔谱, x_{i1} 为预测的第i1帧梅尔谱,i1为大于或等于且小于或等于n1的正整数。

[0200] 线性谱损失可以通过以下公式计算获得:

$$[0201] \quad \text{Spec_loss} = \sum_{i2=1}^{n2} (x_{i2} - y_{i2})^2 \quad (5)$$

[0202] 上述公式中,n2为线性谱帧数,n2为大于或等于1的正整数, y_{i2} 为标注的真实的第i2帧线性谱, x_{i2} 为预测的第i2帧线性谱,i2为大于或等于且小于或等于n2的正整数。

[0203] 图15示意性示出了根据本公开的一实施例的端到端神经网络模型的训练过程示意图。

[0204] 如图15所示,可以从歌唱数据库中获取样本乐谱、样本歌词及其样本歌唱音频,对获取的数据进行数据切分与乐谱标注,提取样本文本特征和样本音乐特征,然后基于提取的样本文本特征和样本音乐特征预测梅尔谱参数和线性谱参数,并基于真实标注的梅尔谱参数和线性谱参数训练声学模型,这里的声学模型即上述的端到端神经网络模型。

[0205] 在示例性实施例中,所述频谱信息可以包括梅尔谱参数和线性谱参数。其中,根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频,可以包括:通过神经网络声码器对所述梅尔谱参数和所述线性谱参数进行处理,合成所述歌唱音频。

[0206] 图16示意性示出了根据本公开的一实施例的端到端神经网络模型的预测过程示意图。

[0207] 如图16所示,在模型预测部分,从用户提供的乐谱信息中提取音乐特征,从用户提供的歌词信息中提取文本特征,将提取的音乐特征和文本特征输入至上述MusicTactron模型,获得预测的梅尔谱和线性谱,将预测的梅尔谱和/或线性谱输入至神经网络声码器,生成歌词和曲调均符合用户要求的歌唱干音。

[0208] 本公开实施方式提供的语音合成方法,改进了相关技术中的端到端神经网络模型的结构,输入至注意力机制模块的信息增加了MIDI嵌入向量和Interval嵌入向量,使得原本用于合成正常说话语音的模型可以用于歌唱干音的合成,一方面,通过添加MIDI encoder,使得模型可以合成音高符合乐谱要求的歌唱干音;另一方面,通过添加符合时值编码器,使得模型可以合成音符时值和乐谱要求一致的歌唱干音,提出了一套可用于歌唱

合成的端到端歌唱合成技术,从而最终可以合成真人级别的歌唱干音,合成效果在自然度、音色拟真度上远强于DNN的技术方案。

[0209] 本公开实施例提出的新的端到端的技术模型结构,可以在歌唱合成领域获得非常好的效果,能按照乐谱合成高质量的歌唱干音。相对于相关技术中的DNN技术方案,在自然度、音色拟真度上有了较大的提高,50人的MOS (Mean Opinion Score,语音质量评测指标)评测结果如图17所述:

[0210] 本方案的MOS分为4.1分,远高于DNN技术方案的3.6分。真人歌唱的MOS分为4.3分,说明本方案合成的歌唱干音非常接近真人的歌唱水平。

[0211] 本公开实施例提供的技术方案是一个中台技术,可以用于支持音乐类、社交类产品。图18-21示意性示出了应用本公开实施例提出的歌唱合成方法的应用场景示意图。如图18所示,是该应用的首页,用户可以通过该应用实现自己作词甚至自己填曲,然后让应用自动唱出来。

[0212] 如图19所示,打开图18应用中的“音乐馆”,用户可以在这里面选择歌曲。如图20所示,自由创作这个大类目下可以包括自由创作、主题作词和热门推荐三个小类目。用户在主题栏中输入关键词或一句话,点击确认虚拟按钮,该应用可以根据用户输入的关键词或一句话,进行上下文的补全,自动生成一首歌的完整歌词,这样可以使得用户不需要自己手动输入全部的歌词。如图21所示,用户打开自由创作这个小类目,可以直接输入歌词,替换原来的歌词,这样用户可以直接使用原有歌曲的乐谱,重新编写自己想要的歌词。

[0213] 图22示意性示出了根据本公开的一实施例的音频合成装置的框图。

[0214] 如图22所示,本公开实施方式提供的音频合成装置2200可以包括:乐谱歌词获取模块2210、音乐特征提取模块2220、文本特征提取模块2230、频谱信息获得模块2240以及歌唱音频合成模块2250。

[0215] 其中,乐谱歌词获取模块2210可以配置为获取待处理乐谱及其待处理歌词。音乐特征提取模块2220可以配置为从所述待处理乐谱中提取到音乐特征。文本特征提取模块2230可以配置为从所述待处理歌词中提取到文本特征。频谱信息获得模块2240可以配置为通过端到端神经网络模型对所述音乐特征和所述文本特征进行处理,获得频谱信息。歌唱音频合成模块2250可以配置为根据所述频谱信息合成与所述待处理乐谱及其待处理歌词相应的歌唱音频。

[0216] 在示例性实施例中,所述音乐特征可以包括乐器数字接口特征和时值特征。其中,音乐特征提取模块2220可以包括:乐器数字接口特征获得子模块,可以配置为根据所述待处理乐谱中的音高,获得所述乐器数字接口特征;时值特征获得子模块,可以配置为对所述待处理乐谱中的音符长度进行归一化,获得所述时值特征。

[0217] 在示例性实施例中,所述端到端神经网络模型可以包括文本编码器、乐器数字接口编码器和时值编码器。其中,频谱信息获得模块2240可以包括:文本嵌入向量获得子模块,可以配置为通过所述文本编码器对所述文本特征进行处理,获得文本嵌入向量;乐器数字接口嵌入向量获得子模块,可以配置为通过所述乐器数字接口编码器对所述乐器数字接口特征进行处理,获得乐器数字接口嵌入向量;时值嵌入向量获得子模块,可以配置为通过所述时值编码器对所述时值特征进行处理,获得时值嵌入向量;融合嵌入向量获得子模块,可以配置为根据所述文本嵌入向量、所述乐器数字接口嵌入向量和所述时值嵌入向量,获

得融合嵌入向量；频谱信息获得子模块，可以配置为根据所述融合嵌入向量，获得所述频谱信息。

[0218] 在示例性实施例中，所述乐器数字接口编码器可以包括第一稠密神经网络。其中，所述乐器数字接口嵌入向量获得子模块可以包括：第一稠密向量获得单元，可以配置为通过所述第一稠密神经网络对所述乐器数字接口特征进行处理，获得第一稠密向量；乐器数字接口嵌入向量获得单元，可以配置为根据所述第一稠密向量，获得所述乐器数字接口嵌入向量。

[0219] 在示例性实施例中，所述乐器数字接口编码器还可以包括第二稠密神经网络、正向门递归神经网络和反向门递归神经网络。其中，所述乐器数字接口嵌入向量获得单元可以包括：第二稠密向量获得子单元，可以配置为通过所述第二稠密神经网络对所述第一稠密向量进行处理，获得第二稠密向量；第一特征图获得子单元，可以配置为通过所述正向门递归神经网络对所述第二稠密向量进行处理，获得第一特征图；第二特征图获得子单元，可以配置为通过所述反向门递归神经网络对所述第一特征图进行处理，获得第二特征图；乐器数字接口嵌入向量获得子单元，可以配置为级联所述第二特征图和所述第二稠密向量，获得所述乐器数字接口嵌入向量。

[0220] 在示例性实施例中，所述时值编码器可以包括第三稠密神经网络。其中，所述时值嵌入向量获得子模块可以包括：时值嵌入向量获得单元，可以配置为通过所述第三稠密神经网络对所述时值特征进行处理，获得所述时值嵌入向量。

[0221] 在示例性实施例中，所述端到端神经网络模型还可以包括注意力机制模块和频谱解码器。其中，所述频谱信息获得子模块可以包括：注意力上下文向量获得单元，可以配置为通过所述注意力机制模块对所述融合嵌入向量进行处理，获得注意力上下文向量；频谱信息获得单元，可以配置为通过所述频谱解码器对所述注意力上下文向量进行处理，获得所述频谱信息。

[0222] 在示例性实施例中，所述频谱信息可以包括梅尔谱参数和线性谱参数。其中，歌唱音频合成模块2250可以包括：歌唱音频合成子模块，可以配置为通过神经网络声码器对所述梅尔谱参数和所述线性谱参数进行处理，合成所述歌唱音频。

[0223] 在示例性实施例中，音频合成装置2200还可以包括：样本信息获取模块，可以配置为获取样本乐谱、样本歌词及其样本歌唱音频；样本音乐特征提取模块，可以配置为从所述样本乐谱中提取到样本音乐特征；样本文本特征提取模块，可以配置为从所述样本歌词中提取到样本文本特征；样本频谱信息获得模块，可以配置为根据所述样本歌唱音频获得样本频谱信息；频谱预测模块，可以配置为通过所述端到端神经网络模型对所述样本音乐特征和所述样本文本特征进行处理，获得预测频谱信息；模型训练模块，可以配置为根据所述样本频谱信息与所述预测频谱信息，训练所述端到端神经网络模型。

[0224] 本公开实施例提供的音频合成装置中的各个模块、子模块、单元和子单元的具体实现可以参照上述音频合成方法中的内容，在此不再赘述。

[0225] 应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块、子模块、单元和子单元，但是这种划分并非强制性的。实际上，根据本公开的实施方式，上文描述的两个或更多模块、子模块、单元和子单元的特征和功能可以在一个模块、子模块、单元和子单元中具体化。反之，上文描述的一个模块、子模块、单元和子单元的特征和功能可以进

一步划分为由多个模块、子模块、单元和子单元来具体化。

[0226] 通过以上的实施方式的描述,本领域的技术人员易于理解,这里描述的示例实施方式可以通过软件实现,也可以通过软件结合必要的硬件的方式来实现。因此,根据本公开实施方式的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中或网络上,包括若干指令以使得一台计算设备(可以是个人计算机、服务器、触控终端、或者网络设备等)执行根据本公开实施方式的方法。

[0227] 本领域技术人员在考虑说明书及实践这里公开的发明后,将容易想到本公开的其它实施方案。本申请旨在涵盖本公开的任何变型、用途或者适应性变化,这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的,本公开的真正范围和精神由下面的权利要求指出。

[0228] 应当理解的是,本公开并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限制。

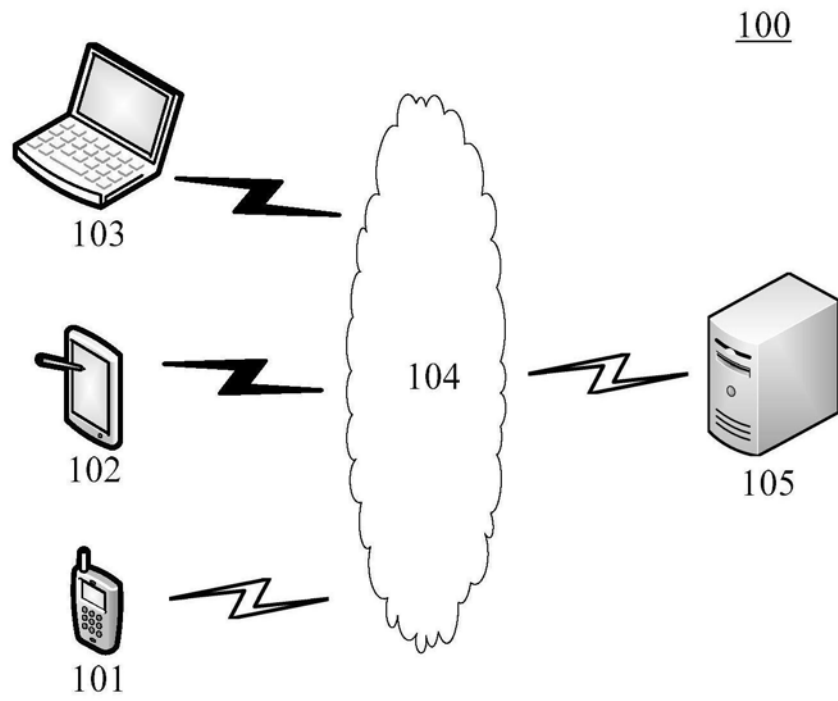


图1

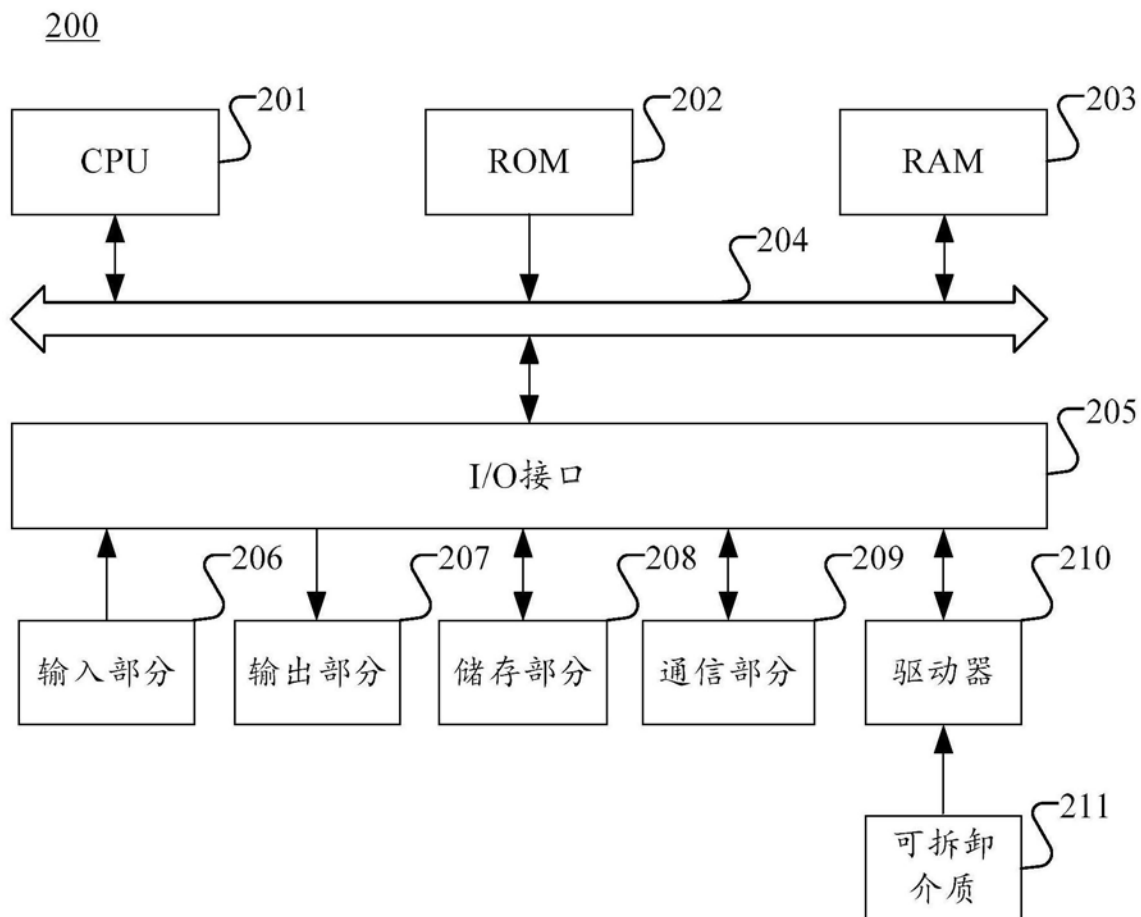


图2

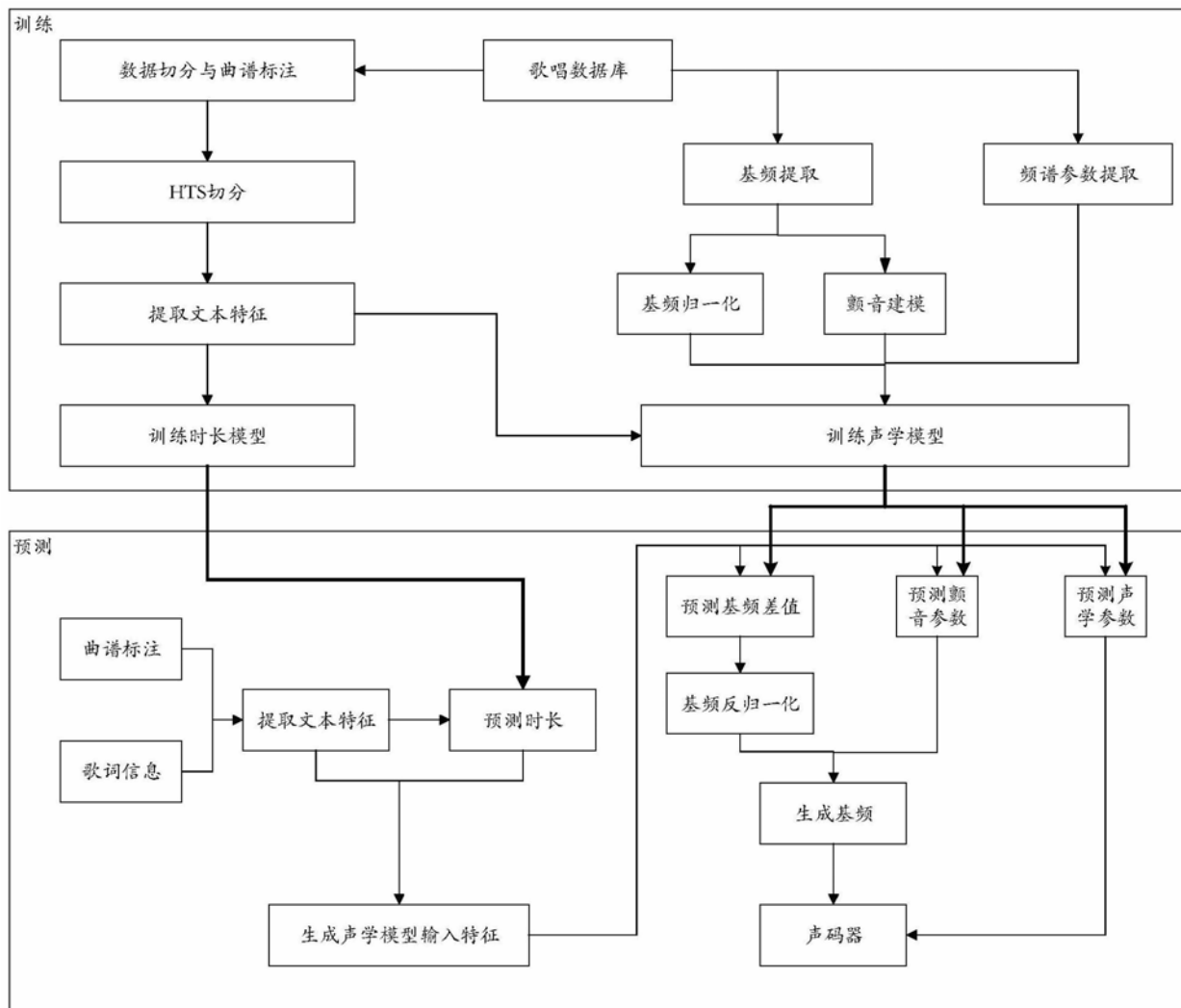


图3

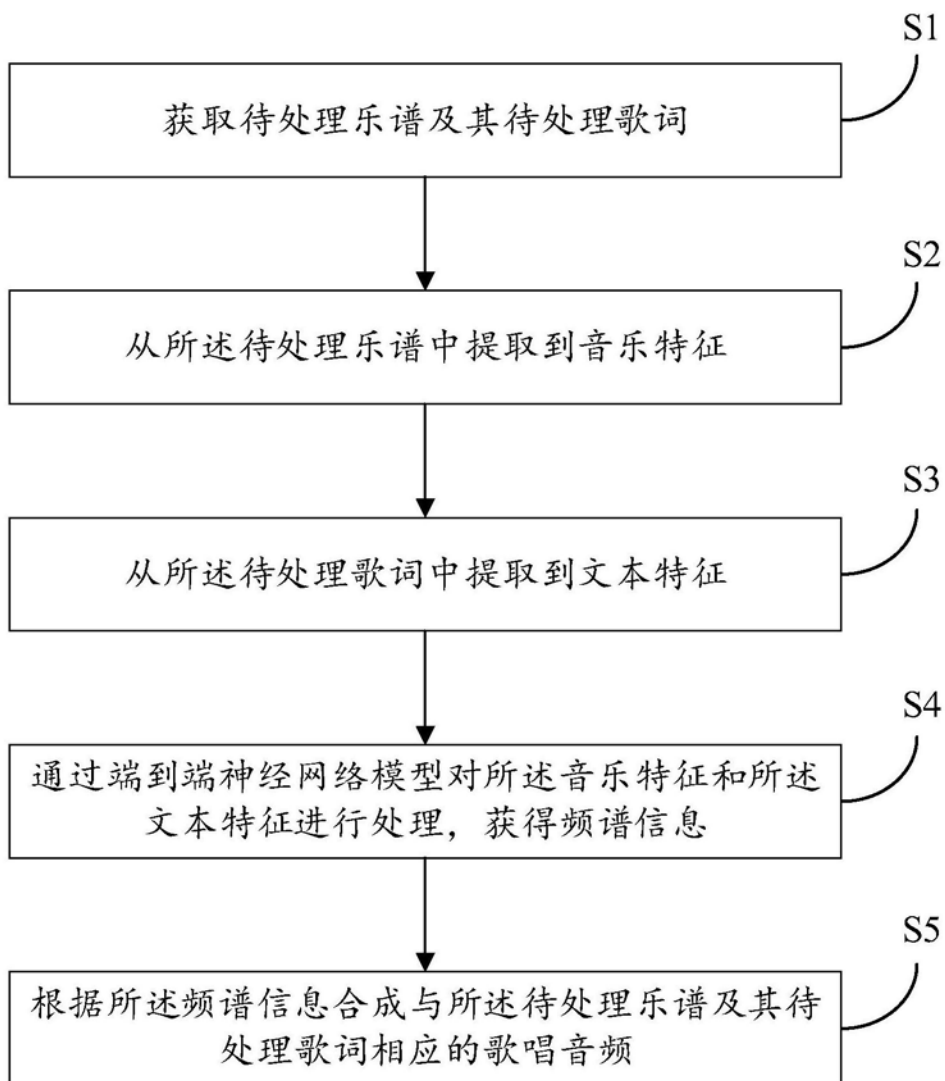


图4

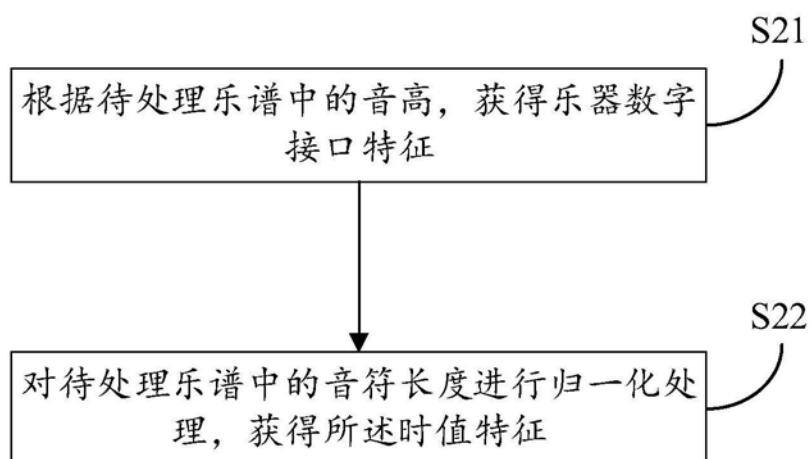


图5

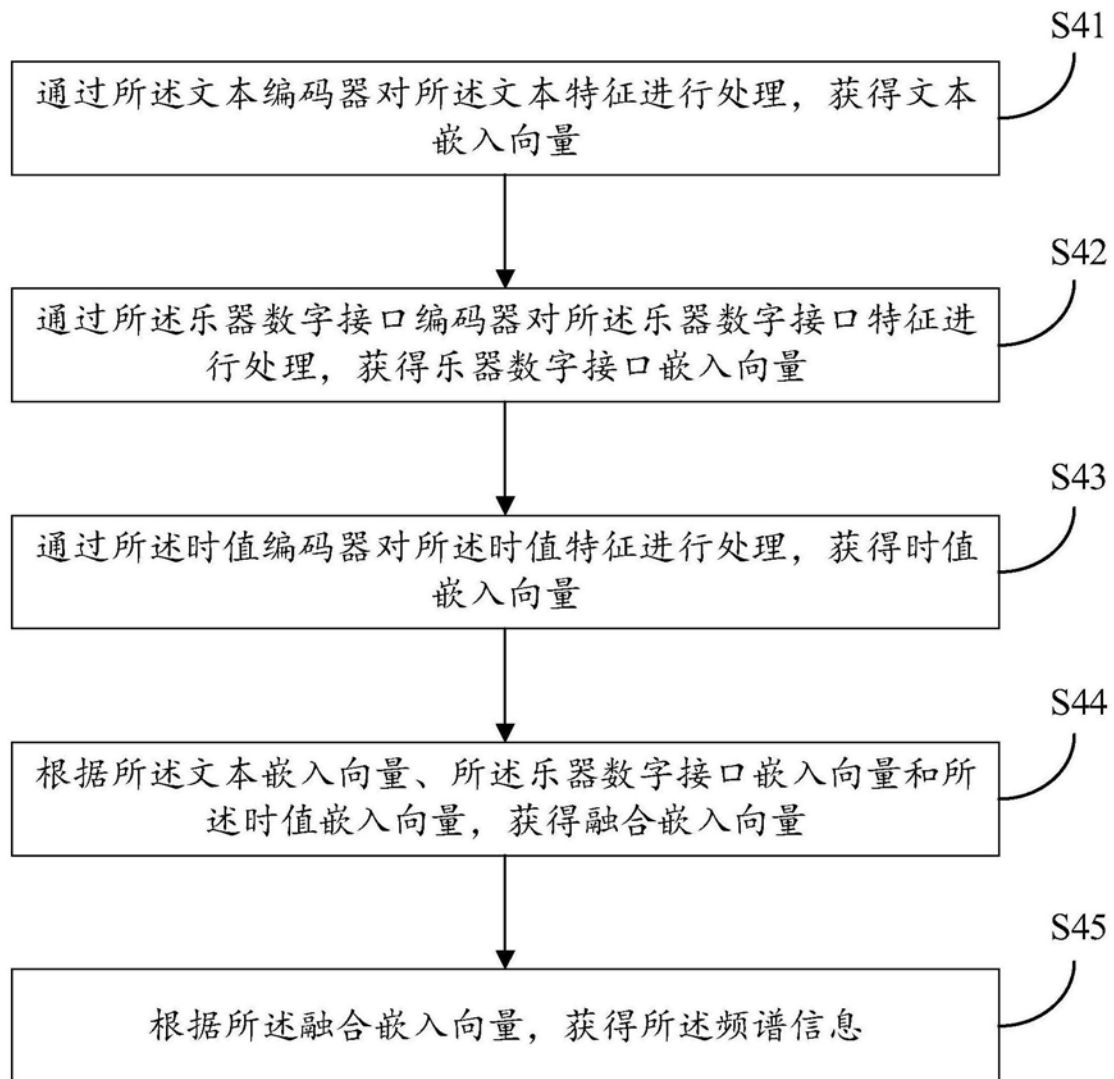


图6

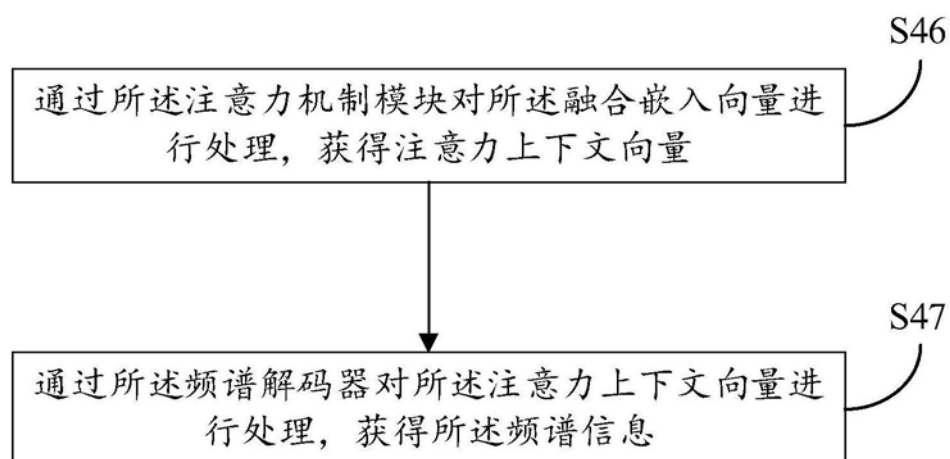


图7

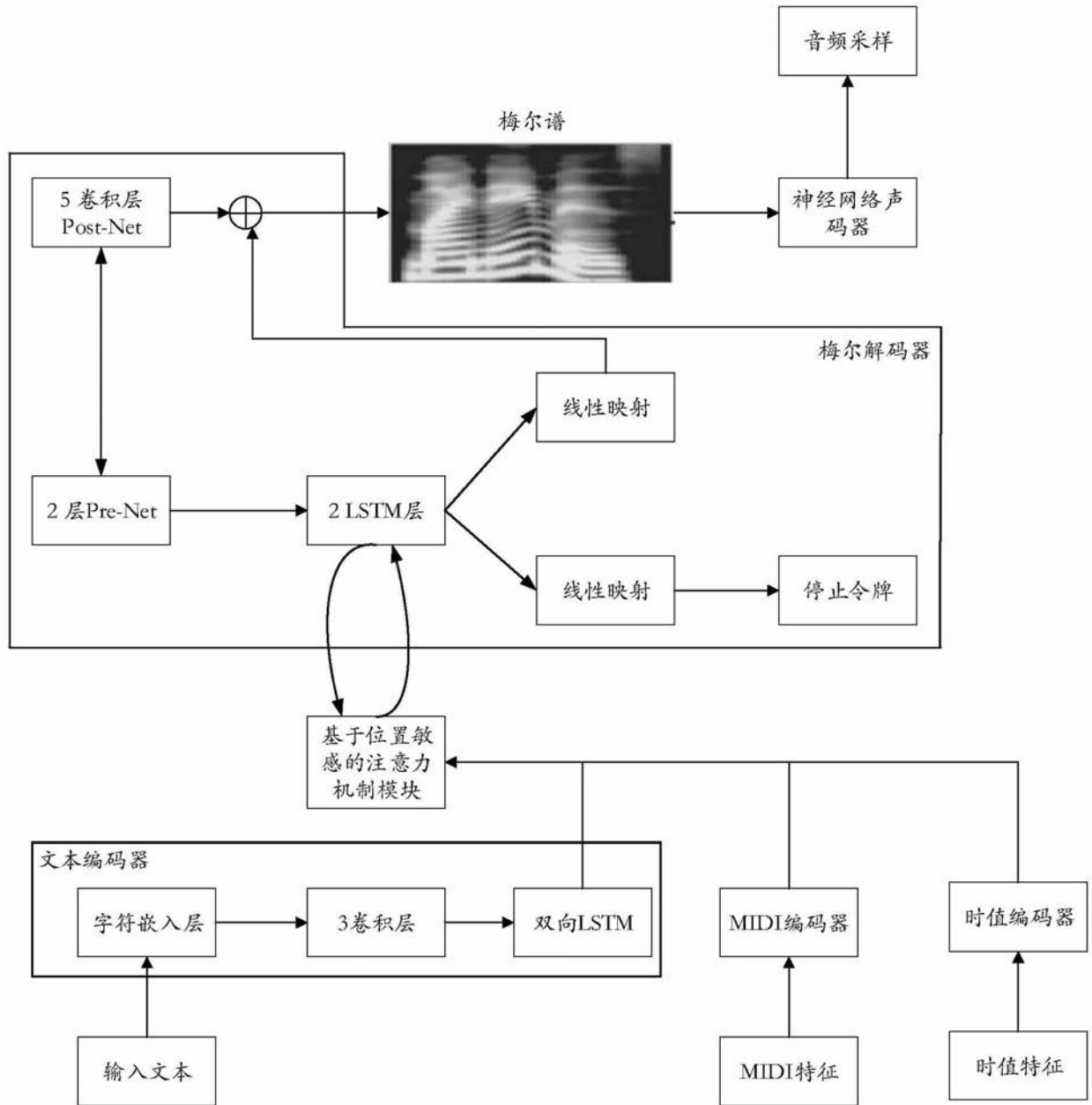


图8

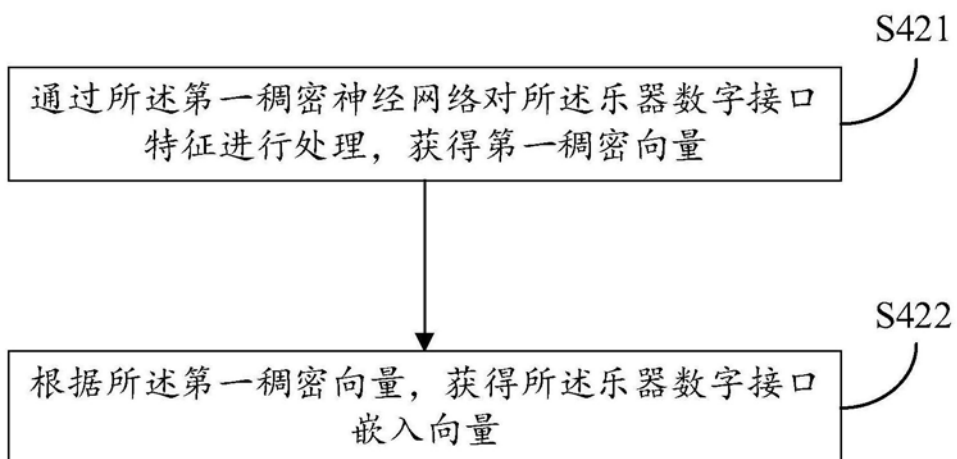


图9

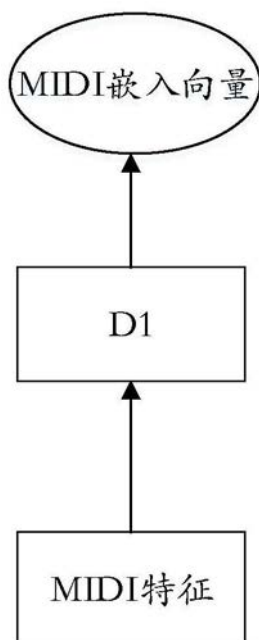


图10

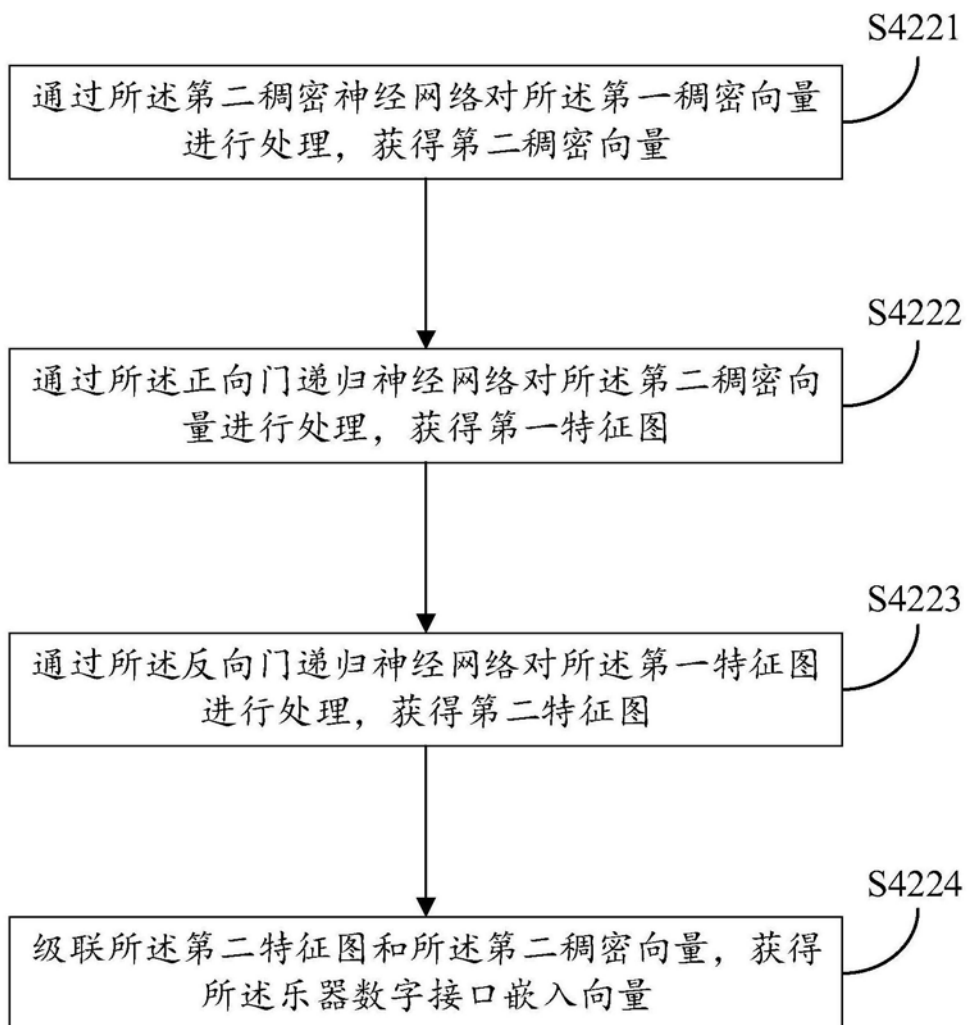


图11

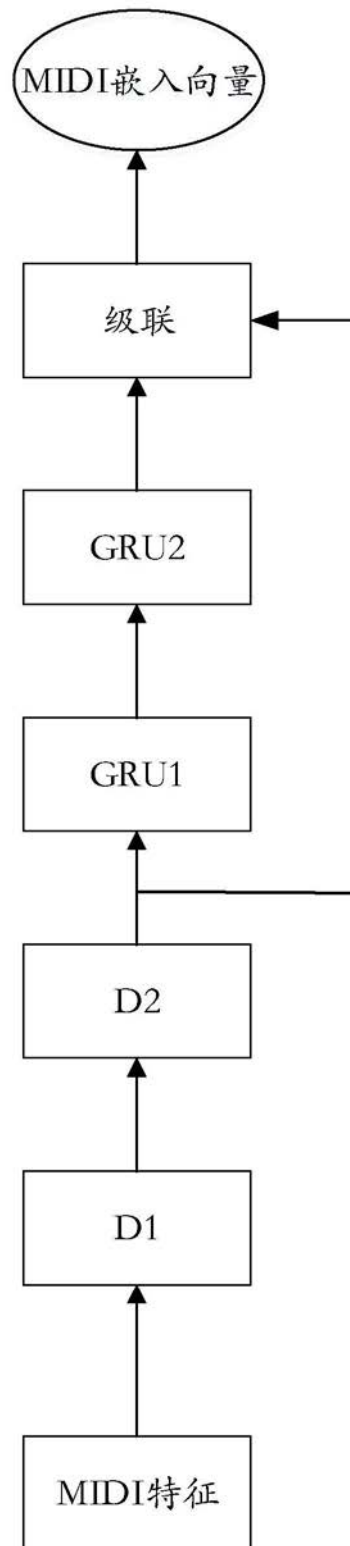


图12

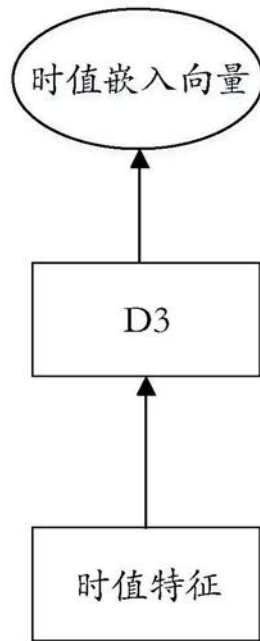


图13

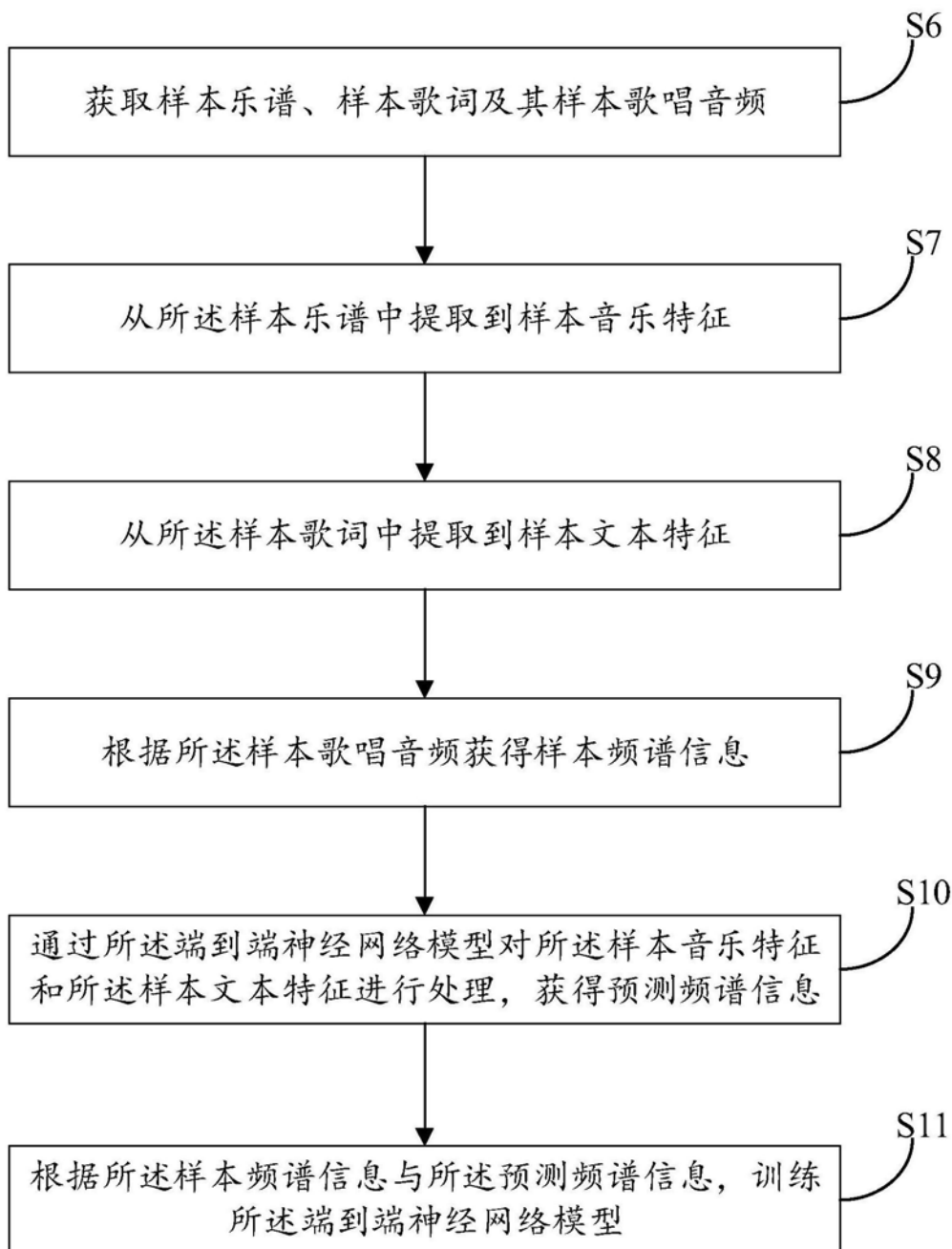


图14

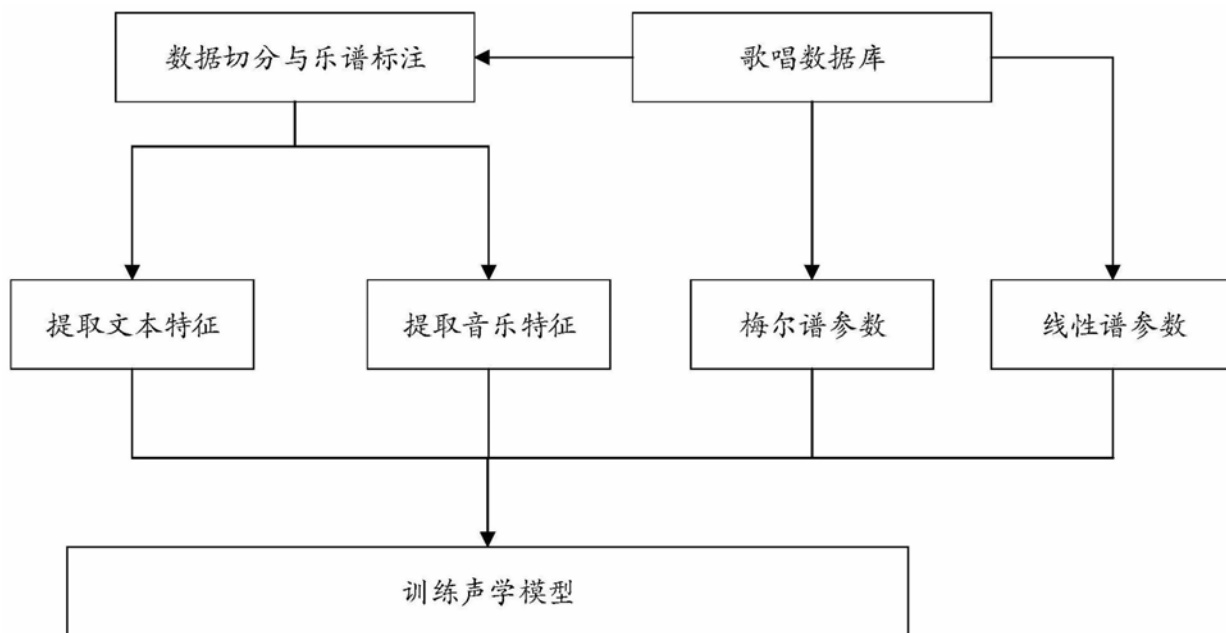


图15

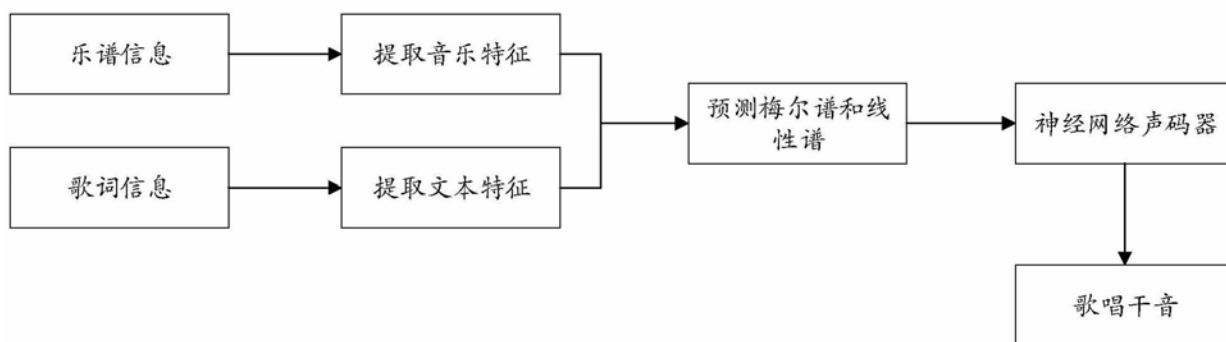


图16

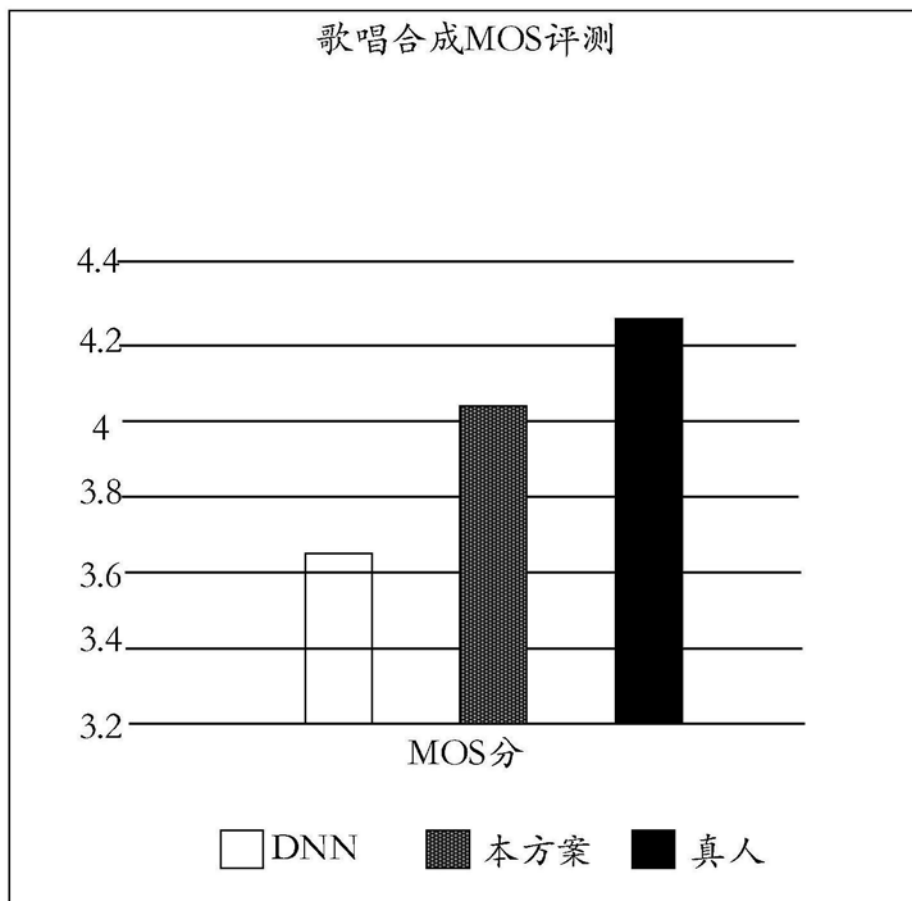


图17



图18



图19



图20

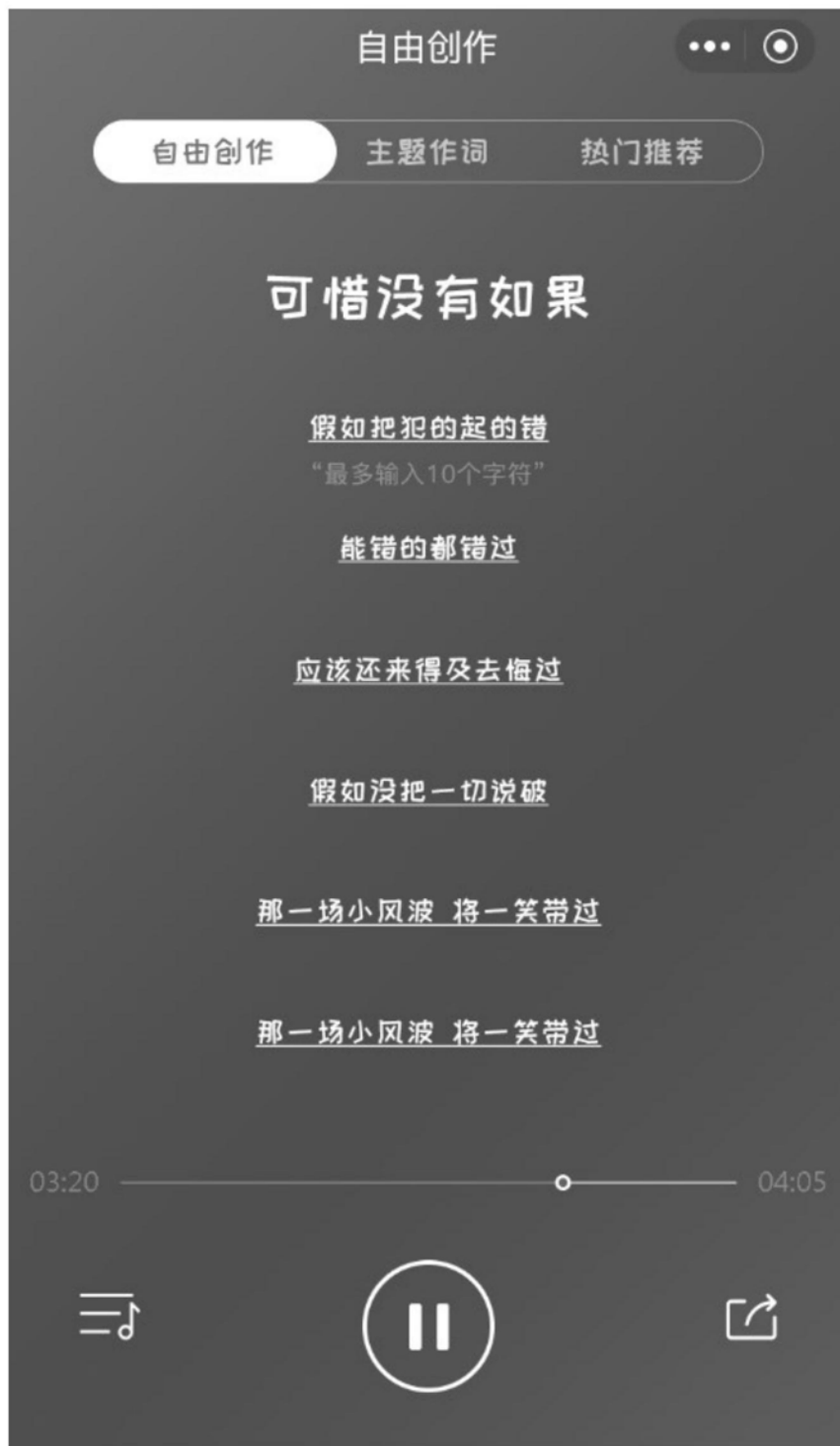


图21

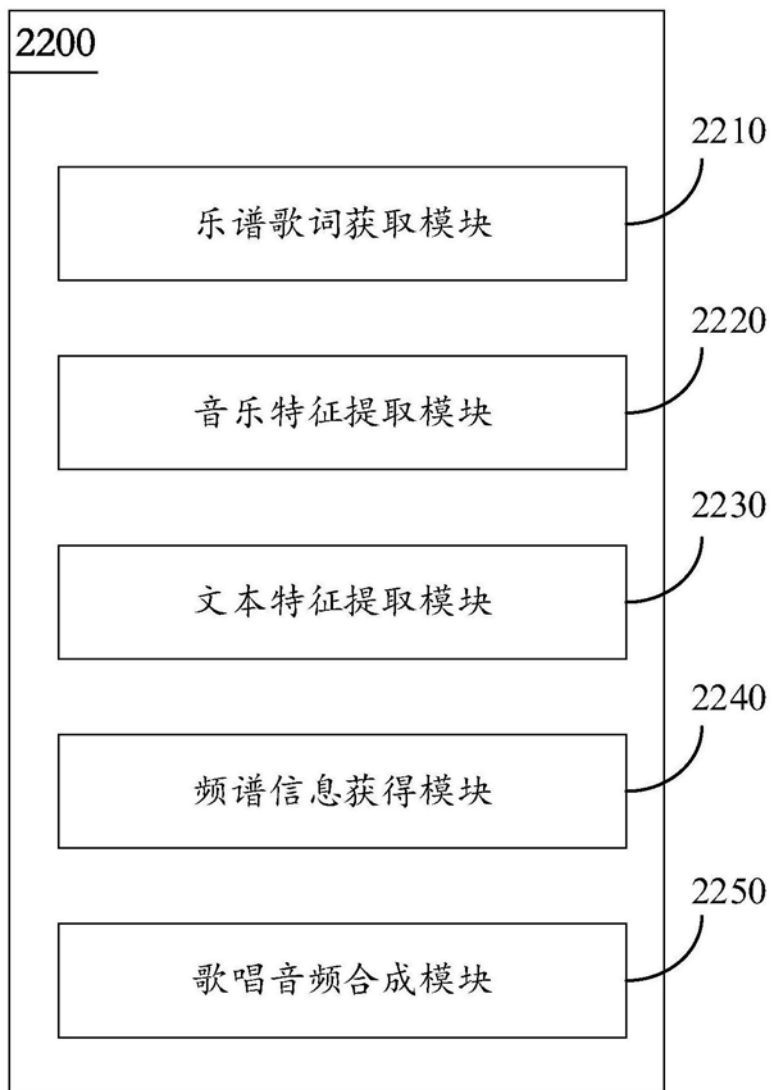


图22