(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 112802449 B
(45)授权公告日 2021.07.02

(21)申请号 202110298526.2
(22)申请日 2021.03.19
(65)同一申请的已公布的文献号
申请公布号 CN 112802449 A
(43)申请公布日 2021.05.14
(73)专利权人 广州酷狗计算机科技有限公司
地址 510660 广东省广州市天河区黄埔大道中315号自编1-17
(72)发明人 关迪聆 陈传艺 劳振锋 孙洪文
(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 李芳
(51)Int.Cl.
G10L 13/08(2013.01)
G10L 13/10(2013.01)
G10L 25/30(2013.01)
G10L 25/24(2013.01)
G10L 25/03(2013.01)
G10L 25/60(2013.01)

(56)对比文件
CN 102270449 A,2011.12.07
CN 111798832 A,2020.10.20
US 2004073428 A1,2004.04.15
US 5748838 A,1998.05.05
CN 1496554 A,2004.05.12

审查员 林登樟

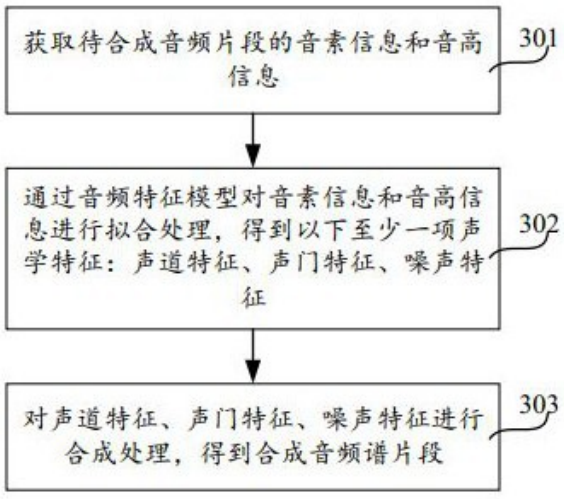
权利要求书3页 说明书9页 附图3页

(54)发明名称

音频合成方法、装置、计算机设备及存储介质

(57)摘要

本申请实施例提供一种音频合成方法、装置、计算机设备及存储介质，涉及深度学习技术领域。该方法包括：获取待合成音频片段的音素信息和音高信息；对音素信息和音高信息进行拟合处理，得到以下声学特征：声道特征、声门特征、噪声特征；对声道特征、声门特征、噪声特征进行合成处理，得到合成音频片段。本申请实施例提供的技术方案，拟合得到的声学特征是独立存在的，对某个声学特征进行修改时无需执行特征提取这一步骤，更为便捷，因此能提高对音频片段的修改效率。



CN 112802449 B

1.一种音频合成方法,其特征在于,所述方法包括:

获取待合成音频片段的音素信息和音高信息,所述音素信息包括所述待合成音频片段的最小语音单位,所述音高信息包括所述待合成音频片段的频率;

对所述音素信息和所述音高信息进行拟合处理,得到以下声学特征:声道特征、声门特征、噪声特征;

获取对应于对所述声学特征的修改指示,所述声学特征包括以下至少一项:所述噪声特征、所述声道特征、所述声门特征;

基于所述修改指示对所述声学特征进行修改,得到修改后声学特征;

基于所述修改后声学特征生成合成音频片段。

2.根据权利要求1所述的方法,其特征在于,所述基于所述修改后声学特征生成合成音频片段,包括:

基于修改后声道特征、所述声门特征、所述噪声特征进行合成处理,得到所述合成音频片段;或者,

基于所述声道特征、修改后声门特征、所述噪声特征进行合成处理,得到所述合成音频片段;或者,

基于所述声道特征、所述声门特征、修改后噪声特征进行合成处理,得到所述合成音频片段;或者,

基于修改后声道特征、修改后声门特征、所述噪声特征进行合成处理,得到所述合成音频片段;或者,

基于修改后声道特征、所述声门特征、修改后噪声特征进行合成处理,得到所述合成音频片段;或者,

基于所述声道特征、修改后声门特征、修改后噪声特征进行合成处理,得到所述合成音频片段;或者,

基于修改后声道特征、修改后声门特征、修改后噪声特征进行合成处理,得到所述合成音频片段。

3.根据权利要求2所述的方法,其特征在于,所述基于修改后声道特征、修改后声门特征、修改后噪声特征进行合成处理,得到所述合成音频片段,包括:

基于所述修改后声门特征和所述修改后声道特征,确定所述待合成音频片段的谐波分量;

基于所述修改后噪声特征,确定所述待合成音频片段的噪声分量;

基于所述待合成音频片段的谐波分量和所述待合成音频片段的噪声分量,生成所述合成音频片段。

4.根据权利要求3所述的方法,其特征在于,所述基于所述修改后声门特征和所述修改后声道特征,确定所述待合成音频片段的谐波分量,包括:

基于所述修改后声门特征,确定第一中间参数;

基于所述第一中间参数与所述修改后声道特征,确定第二中间参数;

对所述第二中间参数进行变换处理,得到第三中间参数;

获取所述合成音频片段中的每帧音频信号的时间信息;

基于所述第三中间参数,以及所述每帧音频信号的时间信息,确定所述待合成音频片

段的谐波分量。

5.根据权利要求1所述的方法，其特征在于，所述获取对应于对所述声学特征的修改指示，包括：

获取对应于所述噪声特征的第一修改指示，所述第一修改指示用于指示修改所述待合成音频片段的信噪比；

和/或，

获取对应于所述声道特征的第二修改指示，所述第二修改指示用于指示修改所述声道特征对应的谐波的幅值；

和/或，

获取对应于所述声门特征的第三修改指示，所述第三修改指示用于修改所述声门特征对应的声门波形。

6.根据权利要求1至5任一项所述的方法，其特征在于，所述拟合处理由音频特征模型完成，所述音频特征模型的训练步骤包括：

获取训练音素信息和训练音高信息；

通过所述音频特征模型对所述训练音素信息和所述训练音高信息进行拟合处理，得到实际输出结果，所述实际输出结果包括：实际声门特征、实际声道特征、实际噪声特征；

通过所述实际输出结果与期望输出结果进行比对，得到损失函数，所述期望输出结果包括：期望声门特征、期望声道特征、期望噪声特征；

基于所述损失函数调整所述音频特征模型的参数。

7.一种音频合成装置，其特征在于，所述装置包括：

信息获取模块，用于获取待合成音频片段的音素信息和音高信息，所述音素信息包括所述待合成音频片段的最小语音单位，所述音高信息包括所述待合成音频片段的频率；

特征拟合模块，用于对所述音素信息和所述音高信息进行拟合处理，得到以下声学特征：声道特征、声门特征、噪声特征；

特征修改模块，用于获取对应于对所述声学特征的修改指示，所述声学特征包括以下至少一项：所述噪声特征、所述声道特征、所述声门特征，基于所述修改指示对所述声学特征进行修改，得到修改后声学特征；

音频合成模块，用于基于所述修改后声学特征生成合成音频片段。

8.根据权利要求7所述的装置，其特征在于，所述音频合成模块，用于：

基于修改后声道特征、所述声门特征、所述噪声特征进行合成处理，得到所述合成音频片段；或者，

基于所述声道特征、修改后声门特征、所述噪声特征进行合成处理，得到所述合成音频片段；或者，

基于所述声道特征、所述声门特征、修改后噪声特征进行合成处理，得到所述合成音频片段；或者，

基于修改后声道特征、修改后声门特征、所述噪声特征进行合成处理，得到所述合成音频片段；或者，

基于修改后声道特征、所述声门特征、修改后噪声特征进行合成处理，得到所述合成音频片段；或者，

基于所述声道特征、修改后声门特征、修改后噪声特征进行合成处理，得到所述合成音频片段；或者，

基于修改后声道特征、修改后声门特征、修改后噪声特征进行合成处理，得到所述合成音频片段。

9.一种计算机设备，其特征在于，所述计算机设备包括处理器和存储器，所述存储器存储有计算机程序，所述计算机程序由所述处理器加载并执行如权利要求1至6任一项所述的音频合成方法。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机程序，所述计算机程序由处理器加载并执行以实现如权利要求1至6任一项所述的音频合成方法。

音频合成方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请实施例涉及深度学习技术领域，特别涉及一种音频合成方法、装置、计算机设备及存储介质。

背景技术

[0002] 歌声合成技术是指使计算机设备合成出模拟人声的歌声的技术，被广泛应用于虚拟偶像领域。

[0003] 相关技术中，计算机设备通过音频特征模型拟合声学特征，之后通过声码器对上述声学特征进行转化处理，得到合成音频片段。其中，上述声学特征是指梅尔谱参数，梅尔谱参数是通过梅尔尺度滤波器组(Mel-scale filter banks)对声谱图进行转换得到的。

[0004] 相关技术中，若需要对合成音频片段进行修改，则需要从合成音频片段中提取出待修改的声学特征，修改难度较大。

发明内容

[0005] 本申请实施例提供一种音频合成方法、装置、计算机设备及存储介质，减小对声学特征进行修改的难度，提高修改效率。所述技术方案包括如下几方面。

[0006] 一方面，本申请实施例提供一种音频合成方法，所述方法包括如下步骤：

[0007] 获取待合成音频片段的音素信息和音高信息，所述音素信息包括所述待合成音频片段的最小语音单位，所述音高信息包括所述待合成音频片段的频率；

[0008] 对所述音素信息和所述音高信息进行拟合处理，得到以下声学特征：声道特征、声门特征、噪声特征；

[0009] 对所述声道特征、所述声门特征、所述噪声特征进行合成处理，得到合成音频片段。

[0010] 另一方面，本申请实施例提供一种音频合成装置，所述装置包括：

[0011] 信息获取模块，用于获取待合成音频片段的音素信息和音高信息，所述音素信息包括所述待合成音频片段的最小语音单位，所述音高信息包括所述待合成音频片段的频率；

[0012] 特征拟合模块，用于对所述音素信息和所述音高信息进行拟合处理，得到以下声学特征：声道特征、声门特征、噪声特征；

[0013] 音频合成模块，用于对所述声道特征、所述声门特征、所述噪声特征进行合成处理，得到合成音频片段。

[0014] 又一方面，本申请实施例提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器存储有计算机程序，所述计算机程序由所述处理器加载并执行以实现如一方面所述的音频合成方法。

[0015] 又一方面，本申请实施例提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，所述计算机程序由处理器加载并执行以实现如一方面所述的音

频合成方法。

[0016] 又一方面，本申请实施例提供了一种计算机程序产品，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上述音频合成方法。

[0017] 本申请实施例提供的技术方案可以带来的有益效果至少包括：通过对音素信息以及音高信息进行拟合处理，得到声道特征、声门特征、噪声特征等声学特征，上述三个声学特征被作为音频合成的素材信息来合成音频片段，相比于相关技术中通过梅尔谱这类综合性声学特征来合成音频片段，上述三个声学特征是独立存在的，对某个声学特征进行修改时无需对合成音频片段执行特征提取这一步骤，更为便捷，因此能提高对音频片段的修改效率。

附图说明

[0018] 图1是本申请一个实施例提供的应用场景的示意图；

[0019] 图2是本申请另一个实施例提供的应用场景的示意图；

[0020] 图3是本申请一个实施例提供的音频合成方法的流程图；

[0021] 图4是本申请一个实施例提供的音频合成的示意图；

[0022] 图5是本申请一个实施例提供的训练音频特征模型的流程图；

[0023] 图6是本申请一个实施例提供的音频合成装置的框图；

[0024] 图7是本申请一个实施例示出的终端的结构框图。

具体实施方式

[0025] 为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

[0026] 首先对本申请实施例涉及的相关名词进行介绍。

[0027] 音素：根据语音的自然属性划分出来的最小语音单位，其基于音节里的发音动作分析得到，一个发音动作构成一个因素。如汉语音节ā有一个音素，ài有两个音素，dāi有三个音素。

[0028] 音高：音的高度，基于声波的频率决定。声速一定时，频率高，波长短，则音高较高，反之，频率低，波长长，则音高较低。

[0029] 音频特征模型：基于音素信息以及音高信息拟合声学特征的模型，其中，音频特征模型的输入数据为音素信息和音高信息，输出数据为声道特征、声门特征、噪声特征。

[0030] 声道特征包括合成音频片段中的频谱包络，频谱包络是指将不同频率的振幅最高点连接起来形成的曲线。声门特征包括声门波形，声门是指位于喉部的两侧声带之间的区域，声门波形用于描述声门开闭运动中，气流通过声门时的波动形式。声门波形通常为尖峰和平坦交替出现，平坦表示声门处于闭合状态，尖峰表示声门处于开启状态。声道特征与声门特征用于合成音频片段中的谐波分量，噪声特征用于合成音频片段中的噪声分量。

[0031] 音频特征模型是通过训练音素信息以及训练音高信息对神经网络模型进行训练得到。可选地，神经网络模型为以下任意一项：tacotron2模型、Deepvoice3模型、Wavenet模

型。tacotron2模型是一种端到端的生成式文本转语音模型。Deepvoice3模型是一个基于注意力机制的全卷积神经元文本转语音(Text To Speech)模型。Wavenet模型是一种序列生成模型,可以用于语音生成建模。

[0032] 本申请实施例提供的技术方案,通过对音素信息以及音高信息进行拟合处理,得到声道特征、声门特征、噪声特征等声学特征,上述三个声学特征被作为音频合成的素材信息来合成音频片段,相比于相关技术中通过梅尔谱这类综合性声学特征来合成音频片段,上述三个声学特征是独立存在的,对某个声学特征进行修改时无需对合成音频片段执行特征提取这一步骤,更为便捷,因此能提高对音频片段的修改效率。

[0033] 本申请实施例提供的技术方案,各步骤的执行主体可以是计算机设备。在一种可能的实现方式中,该计算机设备是智能手机、平板电脑、个人计算机之类的终端设备。在另一种可能的实现方式中,该计算机设备是智能音箱。

[0034] 本申请实施例提供的技术方案,可以应用在虚拟偶像场景以及智能音箱场景。下面对这两个场景进行介绍。

[0035] 虚拟偶像场景:参见图1,计算机设备设置有虚拟偶像11,用户可以设置期望虚拟偶像11演唱的歌曲,计算机设备基于歌曲的歌词文本确定音素信息,并获取虚拟偶像对应的音高信息,通过音频特征模型对上述音素信息和音高信息进行拟合处理,得到声道特征、声门特征、噪声特征等声学特征,最后基于上述声学特征合成歌声信号,从而实现控制虚拟偶像11歌唱。

[0036] 智能音箱场景:参见图2,用户对智能音箱21进行提问,智能音箱21查询该问题的答案后,将其划分得到音素信息,之后获取自身对应的音高信息,通过音频特征模型对上述音素信息和音高信息进行拟合处理,得到声道特征、声门特征、噪声特征等声学特征,最后基于上述声学特征合成语音片段,并播放语音片段以回答用户提出的问题。

[0037] 图3示出了本申请一个实施例提供的音频合成方法的流程图。该方法包括如下步骤。

[0038] 步骤301,获取待合成音频片段的音素信息和音高信息。

[0039] 待合成音频片段的音素信息包括待合成音频片段的最小语音单位,比如待合成音频片段中的各个音素,以及各个音素的排列顺序。可选地,待合成音频片段为一首歌曲时,计算机设备获取该歌曲对应的歌词文本,之后将歌词文本划分成音素,得到音素信息。

[0040] 音高信息包括待合成音频片段的频率。待合成音频片段中的不同片段对应的音高可以相同,也可以不相同。该音高信息由计算机设备默认设定,或者,由技术人员自定义设定。

[0041] 步骤302,对音素信息和音高信息进行拟合处理,得到以下声学特征:声道特征、声门特征、噪声特征。

[0042] 声道特征以及声门特征用于决定合成音频片段中的谐波分量。由于谐波分量对应的谐波特征的维度取决于音高,为避免谐波特征的维度不一致,则将谐波特征转化为维度固定的频谱包络(也即声道特征)以及声门特征。噪声特征用于决定合成音频片段中的噪声分量。

[0043] 可选地,计算机设备获取预先训练的音频特征模型,通过音频特征模型对音素信息和音高信息进行拟合处理,得到以下声学特征:声道特征、声门特征、噪声特征。音频特征

模型是通过训练音素信息以及训练音高信息对神经网络模型进行训练得到的，其具有拟合声学特征的功能。音频特征模型的训练过程在下文实施例进行介绍。

[0044] 在本申请实施例中，所拟合的并非是梅尔谱这类综合性声学特征，而是独立的声学特征，包括声道特征、声门特征与噪声特征等，对上述独立的声学特征进行修改时，无需执行特征提取这一步骤，使得对声学特征的修改更为便捷。

[0045] 步骤303，对声道特征、声门特征、噪声特征进行合成处理，得到合成音频谱片段。

[0046] 可选地，步骤303包括如下子步骤。

[0047] 步骤303a，基于声门特征和声道特征，确定待合成音频片段的谐波分量。

[0048] 可选地，终端基于声门特征，确定第一中间参数；基于第一中间参数与声道参数，确定第二中间参数；对第二中间参数进行变换处理，得到第三中间参数；获取合成音频片段中的每帧音频信号的时间信息；基于第三中间参数，以及每帧音频信号的时间信息，确定合成音频片段的谐波分量。

[0049] 第一中间参数是指频域表示的声门波模型。可选地，第二中间参数是指频域表示的声门模型与声道特征的乘积。第三中间参数是时域表示的谐波模型，其通过对第二中间参数进行傅里叶反变换得到。每帧音频信号的时间信息基于帧周期信息以及生成声门源的相位信息来确定，之后将第三中间参数，按照时间信息进行叠加，得到待合成音频片段的谐波分量。

[0050] 在一个示例中，声道特征记为vt，声门特征记为rd，噪声特征记为psd，计算机设备通过合成函数将第i帧的声门特征rd(i)转换为频域表示的声门波模型G(i)(也即第一中间参数)，将G(i)与第i帧的声道特征vt(i)相乘得到第二中间参数H(i)，再对第二中间参数H(i)进行傅里叶反变换，得到通过时域表示的谐波模型h(i)(也即第三中间参数)，之后将每帧音频信号对应的第三中间参数按照时间信息进行叠加，得到待合成音频片段的谐波分量。其中，i为正整数，i的最大取值为音频信号的数量。

[0051] 步骤303b，基于噪声特征，确定待合成音频片段的噪声分量。

[0052] 可选地，计算机设备通过噪声特征与高斯白噪声来获取待合成音频片段在时域上的噪声分量。

[0053] 步骤303c，基于待合成音频片段的谐波分量和待合成音频片段的噪声分量，生成合成音频片段。

[0054] 可选地，计算机设备将合成音频片段的谐波分量和合成音频片段的噪声分量相加，得到合成音频片段。

[0055] 参考图4，其示出了本申请一个实施例提供的音频合成的示意图。音素信息以及音高信息被输入音频特征模型，音频特征模型对输入的信息进行特征拟合，得到声道特征、声门特征以及噪声特征等声学特征，最后对上述声学特征进行合成处理，得到合成音频片段。

[0056] 综上所述，本申请实施例提供的技术方案，通过对音素信息以及音高信息进行拟合处理，得到声道特征、声门特征、噪声特征等声学特征，上述三个声学特征被作为音频合成的素材信息来合成音频片段，相比于相关技术中通过梅尔谱这类综合性声学特征来合成音频片段，上述三个声学特征是独立存在的，对某个声学特征进行修改时无需对合成音频片段执行特征提取这一步骤，更为便捷，因此能提高对合成音频片段的修改效率。

[0057] 在本申请实施例中，拟合处理得到的声学特征并非综合性的声学特征，而是独立

的声学特征，因此在进行音频合成之前，可以便捷地对一项或多项声学特征进行修改，以使得合成出的音频片段相应修改。下面对修改声学特征进行讲解。在基于图3所示实施例提供的可选实施例中，在步骤303之前，该音频合成方法还包括如下步骤。

[0058] 步骤401，获取对应于对声学特征的修改指示。

[0059] 声学特征包括以下至少一项：噪声特征、声道特征、声门特征。

[0060] 在一种可能的实现方式中，步骤401实现为：获取对应于噪声特征的第一修改指示。

[0061] 第一修改指示用于指示修改待合成音频片段的信噪比。计算机设备基于第一修改指示对噪声特征执行以下至少一项修改操作：增强操作、削弱操作、放大操作、缩小操作。

[0062] 增强操作用于指示对噪声特征进行增强。计算机设备可以对整个音频片段的噪声特征进行增强，也可以对部分音频片段的噪声特征进行增强。增强量由技术人员自定义设定，或者，由计算机设备默认设定。

[0063] 削弱操作用于指示对噪声特征进行削弱。计算机设备可以对整个音频片段的噪声特征进行削弱，也可以对部分音频片段的噪声特征进行削弱。削弱量由技术人员自定义设定，或者，由计算机设备默认设定。

[0064] 放大操作用于指示对噪声特征进行放大处理。计算机设备可以对整个音频片段的噪声特征进行放大，也可以对部分音频片段的噪声特征进行放大。放大倍数由技术人员自定义设定，或者，由计算机设备默认设定。可选地，放大倍数大于1。

[0065] 缩小操作用于指示对噪声特征进行缩小处理。计算机设备可以对整个音频片段的噪声特征进行缩小，也可以对部分音频片段的噪声特征进行缩小。缩小倍数由技术人员自定义设定，或者，由计算机设备默认设定。可选地，缩小倍数小于1。

[0066] 在另一种可能的实现方式中，步骤401实现为：获取对应于声道特征的第二修改指示。

[0067] 第二修改指示用于指示修改声道特征对应的谐波的幅值。计算机设备基于第二修改指示获取声道特征对应的谐波，之后对上述谐波的幅值执行增强操作、削弱操作、放大操作、缩小操作中的至少一种。

[0068] 在又一种可能的实现方式中，步骤401实现为：获取对应于声门特征的第三修改指示。第三修改指示用于修改声门特征对应的声门波形。可选地，计算机设备基于第三修改指示对声门特征对应的声门波形进行替换操作。

[0069] 在其他可能的实现方式中，计算机设备接收到对应于噪声特征的第一修改指示，且接收到对应于声道特征的第二修改指示；或者，计算机设备接收到对应于噪声特征的第一修改指示，且接收到对应于声门特征的第三修改指示；计算机设备接收到对应于声道特征的第二修改指示，且接收到对应于声门特征的第三修改指示；或者，计算机设备接收到对应于噪声特征的第一修改指示、对应于声道特征的第一修改指示以及对应于声门特征的第三修改指示。

[0070] 步骤402，基于修改指示对声学特征进行修改，得到修改后声学特征，修改后声学特征用于合成合成音频片段。

[0071] 当修改指示包括第一修改指示时，若第一修改指示用于指示对噪声特征进行增强操作，则对噪声特征执行增强操作；若第一修改指示用于指示对噪声特征进行削弱操作，则

对噪声特征执行削弱操作;若第一修改指示用于指示对噪声特征进行放大操作,则对噪声特征执行放大操作;若第一修改指示用于指示对噪声特征进行缩小操作,则对噪声特征执行缩小操作。噪声特征被修改后,待合成音频片段的噪声功率谱密度相应改变,此时合成音频片段的呈现效果相应改变。

[0072] 当修改指示包括第二修改指示时,计算机设备获取声道特征后,通过预设算法获取声道特征对应的幅值,之后对上述幅值执行增强操作、削弱操作、放大操作、缩小操作中的至少一种,在完成幅值修改后,再将修改后幅值还原为声道特征。其中,预设方法可以是插值法。声道特征被修改后,待合成音频片段的谐波分量相应改变,此时合成音频片段的呈现效果相应改变。

[0073] 当修改指示包括第三修改指示时,计算机设备获取用户期望的声门效果,基于该期望的声门效果确定相应的声门波形,之后将确定出的声门波形替换声门特征对应的声门波形,得到修改后声门特征。声门特征被修改后,合成音频片段的声门效果相应改变,此时合成音频片段的呈现效果相应改变。

[0074] 可选地,计算机设备基于修改后声学特征生成合成音频片段。

[0075] 在一种可能的实现方式中,计算机设备基于修改后声道特征、声门特征、噪声特征进行合成处理,得到合成音频片段。在另一种可能的实现方式中,计算机设备基于声道特征、修改后声门特征、噪声特征进行合成处理,得到合成音频片段。在一种可能的实现方式中,计算机设备基于声道特征、声门特征、修改后噪声特征进行合成处理,得到合成音频片段。

[0076] 在一种可能的实现方式中,计算机设备基于修改后声道特征、修改后声门特征、噪声特征进行合成处理,得到合成音频片段。在一种可能的实现方式中,计算机设备基于修改后声道特征、声门特征、修改后噪声特征进行合成处理,得到合成音频片段。在一种可能的实现方式中,计算机设备基于声道特征、修改后声门特征、修改后噪声特征进行合成处理,得到合成音频片段。

[0077] 在一种可能的实现方式中,计算机设备基于修改后声道特征、修改后声门特征、修改后噪声特征进行合成处理,得到合成音频片段。在该种实现方式中,合成处理包括如下步骤:基于修改后声门特征和修改后声道特征,确定待合成音频片段的谐波分量;基于修改后噪声特征,确定待合成音频片段的噪声分量;基于待合成音频片段的谐波分量和待合成音频片段的噪声分量,生成合成音频片段。以上步骤的实现细节参考步骤303的介绍,此处不作赘述。

[0078] 其中,基于修改后声门特征和修改后声道特征,确定待合成音频片段的谐波分量,包括:基于修改后声门特征,确定第一中间参数;基于第一中间参数与修改后声道特征,确定第二中间参数;对第二中间参数进行变换处理,得到第三中间参数;获取合成音频片段中的每帧音频信号的时间信息;基于第三中间参数,以及每帧音频信号的时间信息,确定待合成音频片段的谐波分量。以上步骤的实现细节参考步骤303的介绍,此处不作赘述。

[0079] 综上所述,本申请实施例提供的技术方案,通过对噪声特征、声道特征以及声门特征中的一项或多项进行修改,以修改合成音频片段。

[0080] 在上文实施例中介绍到,对噪声特征、声门特征、声道特征进行拟合处理由音频特征模型来完成,下面对音频特征模型的训练过程进行介绍。参考图5,该训练过程包括如下

步骤。

[0081] 步骤501,获取训练音素信息和训练音高信息。

[0082] 训练音素信息和训练音高信息也即是一组训练样本。训练样本的数量基于音频特征模型的训练精度实际设定。音频特征模型的训练精度越高,则训练样本的数量越多。

[0083] 步骤502,通过音频特征模型对训练音素信息和训练音高信息进行拟合处理,得到实际输出结果。

[0084] 可选地,在训练开始前的音频特征模型是以下任意一种:tacotron2模型、Deepvoice3模型、Wavenet模型。实际输出结果包括以下声学特征:实际声门特征、实际声道特征与实际噪声特征。

[0085] 实际输出结果包括以下声学特征:实际声门特征、实际声道特征与实际噪声特征。

[0086] 步骤503,通过实际输出结果与期望输出结果进行比对,得到损失函数。

[0087] 期望输出结果包括以下声学特征:期望声门特征、期望声道特征与期望噪声特征。计算机设备将实际输出结果与期望输出结果进行逐一比对,得到损失函数。

[0088] 步骤504,基于损失函数调整音频特征模型的参数。

[0089] 计算机设备通过预设算法基于损失函数,来调节音频特征模型的参数,之后重复上述步骤502-504,直至损失函数符合预设条件,此时得到完成训练的音频特征模型。

[0090] 以下为本申请装置实施例,对于装置实施例中未详细阐述的部分,可以参考上述方法实施例中公开的技术细节。

[0091] 请参考图6,其示出了本申请一个示例性实施例提供的音频合成装置的框图。该音频合成装置可以通过软件、硬件或者两者的组合实现成为终端的全部或一部分。该音频合成装置包括如下模块。

[0092] 信息获取模块601,用于获取待合成音频片段的音素信息和音高信息,所述音素信息包括所述待合成音频片段的最小语音单位,所述音高信息包括所述待合成音频片段的频率。

[0093] 特征拟合模块602,用于对所述音素信息和音高信息进行拟合处理,得到以下声学特征:声道特征、声门特征、噪声特征。

[0094] 音频合成模块603,用于对所述声道特征、所述声门特征、所述噪声特征进行合成处理,得到合成音频片段。

[0095] 综上所述,本申请实施例提供的技术方案,通过对音素信息以及音高信息进行拟合处理,得到声道特征、声门特征、噪声特征等声学特征,上述三个声学特征被作为音频合成的素材信息来合成音频片段,相比于相关技术中通过梅尔谱这类综合性声学特征来合成音频片段,上述三个声学特征是独立存在的,对某个声学特征进行修改时无需对合成音频片段执行特征提取这一步骤,更为便捷,因此能提高对合成音频片段的修改效率。

[0096] 在基于图6所示实施例提供的可选实施例中,所述音频合成模块603,用于:基于所述声门特征和所述声道特征,确定所述待合成音频片段的谐波分量;基于所述噪声特征,确定所述待合成音频片段的噪声分量;基于所述待合成音频片段的谐波分量和所述待合成音频片段的噪声分量,生成所述合成音频片段。

[0097] 可选地,所述音频合成模块603,用于:基于所述声门特征,确定第一中间参数;基于所述第一中间参数与所述声道特征,确定第二中间参数;对所述第二中间参数进行变换

处理，得到第三中间参数；获取所述合成音频片段中的每帧音频信号的时间信息；基于所述第三中间参数，以及所述每帧音频信号的时间信息，确定所述待合成音频片段的谐波分量。

[0098] 在基于图6所示实施例提供的可选实施例中，所述装置还包括特征修改模块(图6未示出)。

[0099] 特征修改模块，用于：获取对应于对所述声学特征的修改指示，所述声学特征包括以下至少一项：所述噪声特征、所述声道特征、所述声门特征；基于所述修改指示对所述声学特征进行修改，得到修改后声学特征，所述修改后声学特征用于合成所述合成音频片段。

[0100] 可选地，所述特征修改模块，用于：获取对应于所述噪声特征的第一修改指示，所述第一修改指示用于指示修改所述待合成音频片段的信噪比；和/或，获取对应于所述声道特征的第二修改指示，所述第二修改指示用于指示修改所述声道特征对应的谐波的幅值；和/或，获取对应于所述声门特征的第三修改指示，所述第三修改指示用于修改所述声门特征对应的声门波形。

[0101] 在基于图6所示实施例提供的可选实施例中，所述拟合处理由音频特征模型完成，所述音频特征模型的训练步骤包括：获取训练音素信息和训练音高信息；通过所述音频特征模型对所述训练音素信息和所述训练音高信息进行拟合处理，得到实际输出结果，所述实际输出结果包括：实际声门特征、实际声道特征、实际噪声特征；通过所述实际输出结果与期望输出结果进行比对，得到损失函数，所述期望输出结果包括：期望声门特征、期望声道特征、期望噪声特征；基于所述损失函数调整所述音频特征模型的参数。

[0102] 需要说明的是，上述实施例提供的装置在实现其功能时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的装置与方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0103] 图7示出了本申请一个示例性实施例提供的计算机设备700的结构框图。该计算机设备700可以是：智能手机、平板电脑、MP3播放器、MP4播放器、笔记本电脑或台式电脑。计算机设备700还可能被称为用户设备、便携式计算机设备、膝上型计算机设备、台式计算机设备等其他名称。

[0104] 通常，计算机设备700包括有：处理器701和存储器702。

[0105] 处理器701可以包括一个或多个处理核心，比如4核心处理器、7核心处理器等。处理器701可以采用数字信号处理(Digital Signal Processing，DSP)、现场可编程门阵列(Field－Programmable Gate Array，FPGA)、可编程逻辑阵列(Programmable Logic Array，PLA)中的至少一种硬件形式来实现。处理器701也可以包括主处理器和协处理器，主处理器是用于对在唤醒状态下的数据进行处理的处理器，也称中央处理器(Central Processing Unit，CPU)；协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器701可以在集成有图像处理器(Graphics Processing Unit，GPU)，GPU用于负责显示屏所需要显示的内容的渲染和绘制。

[0106] 存储器702可以包括一个或多个计算机可读存储介质，该计算机可读存储介质可以是非暂态的。存储器702还可包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。在一些实施例中，存储器702中的非暂态的计算机可

读存储介质用于存储计算机程序，该计算机程序用于被处理器701所执行以实现本申请中方法实施例提供的音频合成方法。

[0107] 在一些实施例中，计算机设备700还可选包括有：外围设备接口703和至少一个外围设备。处理器701、存储器702和外围设备接口703之间可以通过总线或信号线相连。各个外围设备可以通过总线、信号线或电路板与外围设备接口703相连。具体地，外围设备包括：射频电路704、触摸显示屏705、摄像头组件706、音频电路707、定位组件708和电源709中的至少一种。

[0108] 本领域技术人员可以理解，图7中示出的结构并不构成对计算机设备700的限定，可以包括比图示更多或更少的组件，或者组合某些组件，或者采用不同的组件布置。

[0109] 在示例性实施例中，还提供了一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，所述计算机程序由终端的处理器加载并执行以实现上述方法实施例中的音频合成方法。

[0110] 可选地，上述计算机可读存储介质可以是只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、磁带、软盘和光数据存储设备等。

[0111] 在示例性实施例中，还提供了一种计算机程序产品，该计算机程序产品包括计算机指令，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上述一方面或者一方面的各种可选实现方式中提供的音频合成方法。

[0112] 以上所述仅为本申请的示例性实施例，并不用以限制本申请，凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。


请演唱下雨这首歌
11

图1

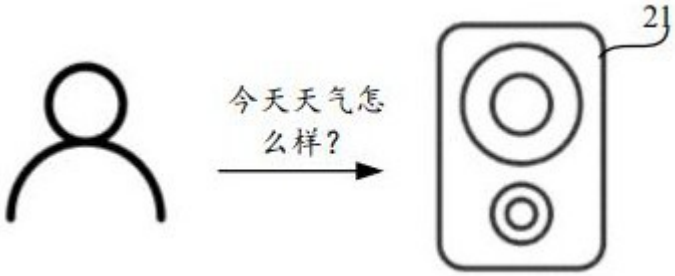

今天天气怎么样?
21

图2

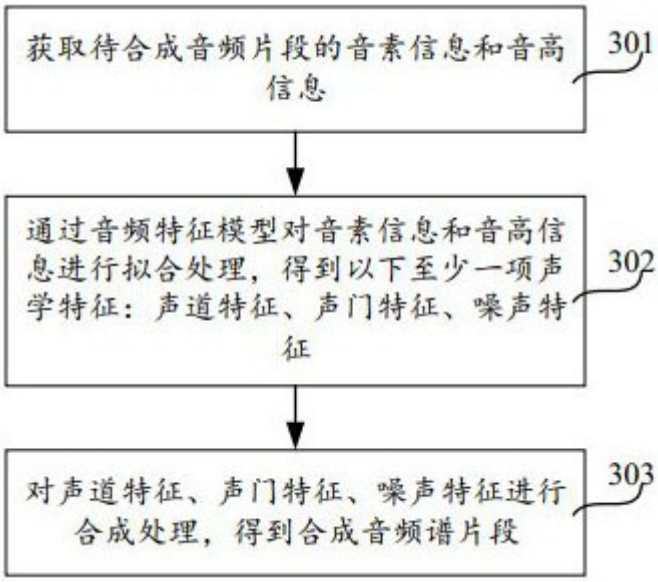

获取待合成音频片段的音素信息和音高信息
301
通过音频特征模型对音素信息和音高信息进行拟合处理，得到以下至少一项声学特征：声道特征、声门特征、噪声特征
302
对声道特征、声门特征、噪声特征进行合成处理，得到合成音频谱片段
303

图3

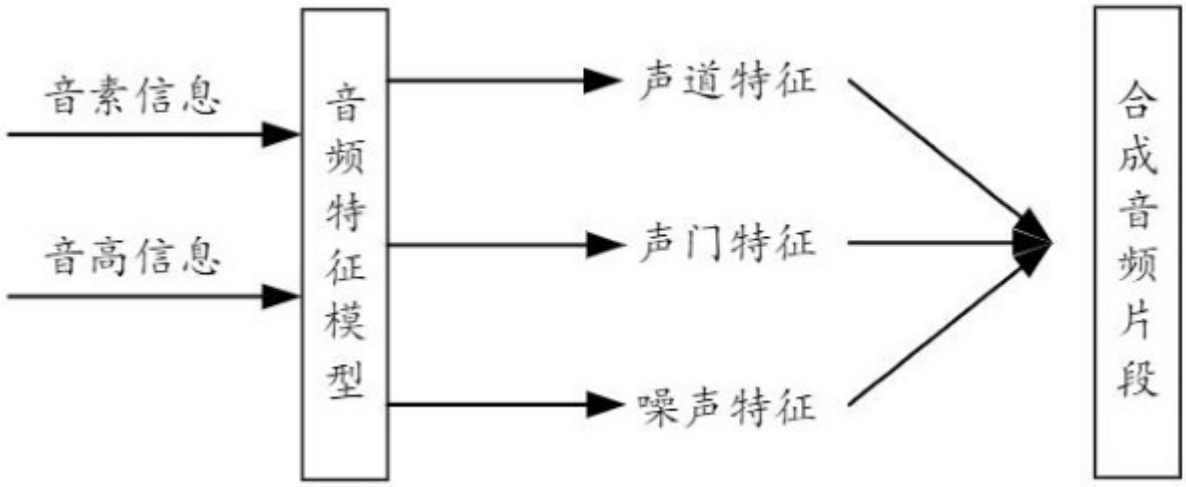

音素信息
音高信息
音频特征模型
声道特征
声门特征
噪声特征
合成音频片段

图4

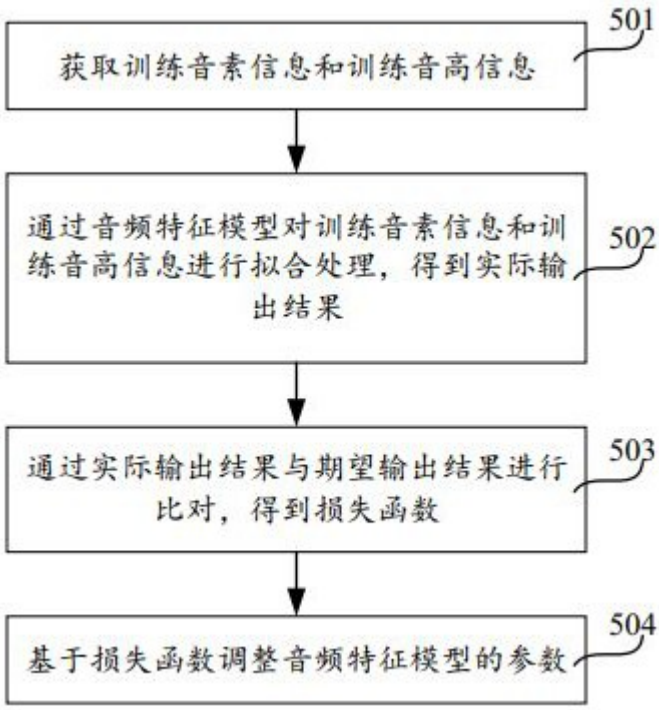

501
获取训练音素信息和训练音高信息
502
通过音频特征模型对训练音素信息和训练音高信息进行拟合处理，得到实际输出结果
503
通过实际输出结果与期望输出结果进行比对，得到损失函数
504
基于损失函数调整音频特征模型的参数

图5


信息获取模块
601
特征拟合模块
602
音频合成模块
603

图6

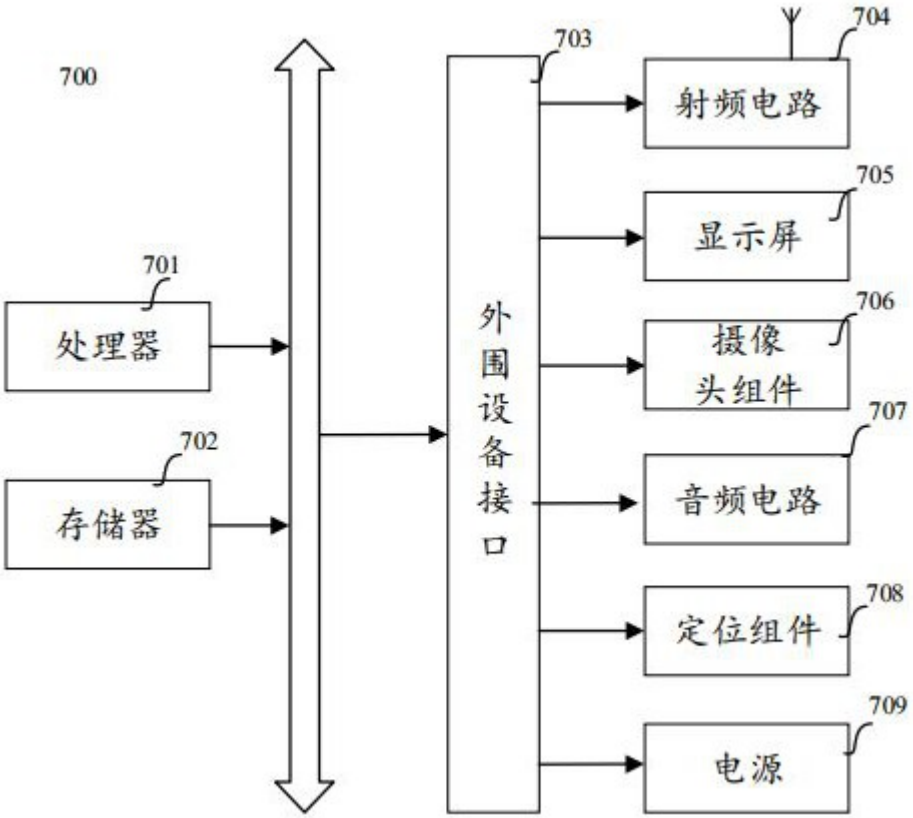

700
701
处理器
702
存储器
703
外围设备接口
704
射频电路
705
显示屏
706
摄像头组件
707
音频电路
708
定位组件
709
电源

图7