

一种音频合成方法、电子设备及可读存储介质

本申请要求于 2022 年 4 月 28 日提交至中国专利局、申请号为 202210460941.8、发明名称为“一种音频合成方法、电子设备及可读存储 5 介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及音频处理技术领域，特别涉及一种音频合成方法、电子设备及计算机可读存储介质。

背景技术

10 随着人工智能的发展，当前已经可以通过语音合成技术将文字合成为非常逼真的语音音频。合成的音频通常用于新闻播报、书籍朗读、人机对话等场景。另外，用户也已经习惯收听各种类型的音乐作品。然而，如何将音乐和语音结合起来，获得一种新的声音格式，让人们在欣赏音乐的同时，也能够听到一些慰藉心灵或振奋人心的语音。当前的解决方案是电台 15 播放的形式，主播时而说话、时而播放既定的歌曲给广大听众。然而，这种方式需要人工参与，效果受限于主播的播音水平。

发明内容

有鉴于此，本申请的目的在于提供一种音频合成方法、电子设备及计算机可读存储介质，不受限于人工播音水平，播放效果较好。

20 为解决上述技术问题，第一方面，本申请提供了一种音频合成方法，包括：

获取待处理歌曲音频及相应的歌曲信息；

对所述待处理歌曲音频进行人声分离处理，得到人声音频；

基于所述人声音频，在多个候选音色信息中确定目标音色信息；

25 获取文本模板，并基于所述文本模板和所述歌曲信息生成待处理文本；

基于所述目标音色信息，对所述待处理文本进行语音合成处理，得到待合成音频；

对所述待合成音频和所述待处理歌曲音频进行合成处理，得到合成后

音频。

可选地，所述基于所述人声音频，在多个候选音色信息中确定目标音色信息，包括：

提取所述人声音频对应的声纹特征向量；

- 5 生成所述声纹特征向量分别与各个候选声纹特征向量之间的余弦相似度；各个所述候选声纹特征向量分别与各个所述候选音色信息相对应；

将最大余弦相似度对应的候选音色信息确定为所述目标音色信息。

可选地，所述基于所述人声音频，在多个候选音色信息中确定目标音色信息，包括：

- 10 将所述人声音频输入基于候选音频训练得到的音色匹配模型，得到所述人声音频与各个所述候选音频之间的相似概率；

将最大相似概率对应的候选音频的候选音色信息确定为所述目标音色信息。

- 15 可选地，所述歌曲信息包括多项歌曲子信息，所述基于所述文本模板和所述歌曲信息生成待处理文本，包括：

基于各个歌曲子信息的信息类型，将各个所述歌曲子信息插入所述文本模板中的对应位置，得到所述待处理文本。

可选地，所述基于所述目标音色信息，对所述待处理文本进行语音合成处理，得到待合成音频，包括：

- 20 将所述目标音色信息和所述待处理文本输入特征编码模型，得到文本编码；

将所述文本编码输入声码器模型，得到所述待合成音频。

可选地，所述特征编码模型和所述声码器模型的训练过程，包括：

获取各个候选音色信息分别对应的训练数据；其中，所述训练数据包

- 25 括训练文本和对应的候选音频；

将所述训练文本输入初始特征编码模型，得到训练编码；

将所述训练编码输入初始声码器模型，得到训练输出音频；

对所述候选音频提取与所述训练编码类型相同的标签编码，并利用所述标签编码和所述训练编码生成第一损失值；

- 30 利用所述候选音频与所述训练输出音频生成第二损失值；

利用所述第一损失值对所述初始特征编码模型进行参数调节，并利用所述第二损失值对所述初始声码器模型进行特征调节；

当检测到满足训练完成条件，将参数调节后的初始特征编码模型确定为所述特征编码模型，并将参数调节后的初始声码器模型确定为所述声码器模型。

可选地，所述对所述待合成音频和所述待处理歌曲音频进行合成处理，包括：

确定所述待处理歌曲音频对应的伴奏时间区间；

基于所述模板信息确定合成位置类型，并利用所述合成位置类型对所述伴奏时间区间进行筛选，得到目标伴奏时间区间；

若所述目标伴奏时间区间的长度不小于所述待合成音频的时间长度，则将所述待合成音频与所述待处理歌曲音频中所述目标伴奏时间区间对应的部分合成；

和/或，

若所述目标伴奏时间区间的长度小于所述待合成音频的时间长度，则基于所述目标伴奏时间区间的类型，对所述待合成音频和所述待处理歌曲音频进行拼接处理。

可选地，所述基于所述目标伴奏时间区间的类型，对所述待合成音频和所述待处理歌曲音频进行拼接处理，包括：

若所述目标伴奏时间区间的类型为前奏，则将所述待合成音频拼接至所述待处理歌曲音频之前；

若所述目标伴奏时间区间的类型为尾奏，则将所述待合成音频拼接至所述待处理歌曲音频之后。

可选地，所述确定所述待处理歌曲音频对应的伴奏时间区间，包括：

利用人声检测模型对待处理歌曲音频进行人声检测，得到多个起始时间点和终止时间点；

将距离小于预设阈值的相邻的终止时间点和起始时间点去除，得到目标起始时间点和目标终止时间点；

利用所述目标起始时间点和目标终止时间点确定所述伴奏时间区间。

第二方面，本申请还提供了一种电子设备，包括存储器和处理器，其

中：

所述存储器，用于保存计算机程序；

所述处理器，用于执行所述计算机程序，以实现上述的音频合成方法。

第三方面，本申请还提供了一种计算机可读存储介质，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现上述的音频合成方法。

本申请提供的音频合成方法，获取待处理歌曲音频及相应的歌曲信息；对待处理歌曲音频进行人声分离处理，得到人声音频；基于人声音频，在多个候选音色信息中确定目标音色信息；获取文本模板，并基于文本模板和歌曲信息生成待处理文本；基于目标音色信息，对待处理文本进行语音合成处理，得到待合成音频；对待合成音频和待处理歌曲音频进行合成处理，得到合成后音频。

可见，该方法通过人声分离处理，并利用得到的人声音频在多个候选音色信息中确定目标音色信息，目标音色信息为与待处理歌曲音频中的人声部分音色相近的音色信息。在生成待处理文本后，基于目标音色信息生成的待合成音频中的人声语音的内容与待处理文本相同，且音色与待处理歌曲音频的演唱者的音色相近。将待合成音频和待处理歌曲音频合成得到的合成后音频中，人声部分音色前后保持相近，具有良好的听感。同时待合成音频的质量较高，不受限于人工播音水平，使得合成后音频的质量较高，播放效果较好，解决了需要人工参与，效果受限于主播的播音水平的问题。

此外，本申请还提供了一种电子设备及计算机可读存储介质，同样具有上述有益效果。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，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，还可以根据提供的附图获得其他的附图。

图1为本申请实施例提供的一种音频合成方法所适用的硬件组成框架示意图；

图2为本申请实施例提供的另一种音频合成方法所适用的硬件组成框

架示意图；

图3为本申请实施例提供的一种音频合成方法的流程示意图；

图4为本申请实施例提供的一种音频合成位置示意图；

图5为本申请实施例提供的一种音频合成全流程示意图。

5 具体实施方式

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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他实施例，都属于本申请保护的范围。

为了便于理解，先对本申请实施例提供的音频合成方法对应的方案所使用的硬件组成框架进行介绍。请参考图 1，图 1 为本申请实施例提供的一种音频合成方法所适用的硬件组成框架示意图。其中电子设备 100 可以包括处理器 101 和存储器 102，还可以进一步包括多媒体组件 103、信息输入/信息输出（I/O）接口 104 以及通信组件 105 中的一种或多种。

其中，处理器 101 用于控制电子设备 100 的整体操作，以完成音频合成方法中的全部或部分步骤；存储器 102 用于存储各种类型的数据以支持在电子设备 100 的操作，这些数据例如可以包括用于在该电子设备 100 上操作的任何应用程序或方法的指令，以及应用程序相关的数据。该存储器 102 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，例如静态随机存取存储器（Static Random Access Memory, SRAM）、电可擦除可编程只读存储器（Electrically Erasable Programmable Read-Only Memory, EEPROM）、可擦除可编程只读存储器（Erasable Programmable Read-Only Memory, EPROM）、可编程只读存储器（Programmable Read-Only Memory, PROM）、只读存储器（Read-Only Memory, ROM）、磁存储器、快闪存储器、磁盘或光盘中的一种或多种。在本实施例中，存储器 102 中至少存储有用于实现以下功能的程序和/或数据：

获取待处理歌曲音频及相应的歌曲信息；

对待处理歌曲音频进行人声分离处理，得到人声音频；

基于人声音频，在多个候选音色信息中确定目标音色信息；

获取文本模板，并基于文本模板和歌曲信息生成待处理文本；

基于目标音色信息，对待处理文本进行语音合成处理，得到待合成音频；

对待合成音频和待处理歌曲音频进行合成处理，得到合成后音频。

5 多媒体组件 103 可以包括屏幕和音频组件。其中屏幕例如可以是触摸屏，音频组件用于输出和/或输入音频信号。例如，音频组件可以包括一个麦克风，麦克风用于接收外部音频信号。所接收的音频信号可以被进一步存储在存储器 102 或通过通信组件 105 发送。音频组件还包括至少一个扬声器，用于输出音频信号。I/O 接口 104 为处理器 101 和其他接口模块之间提供接口，上述其他接口模块可以是键盘，鼠标，按钮等。这些按钮可以是虚拟按钮或者实体按钮。通信组件 105 用于电子设备 100 与其他设备之间进行有线或无线通信。无线通信，例如 Wi-Fi，蓝牙，近场通信 (Near Field Communication, 简称 NFC)，2G、3G 或 4G，或它们中的一种或几种的组合，因此相应的该通信组件 105 可以包括：Wi-Fi 部件，蓝牙部件，
10
15 NFC 部件。

电子设备 100 可以被一个或多个应用专用集成电路 (Application Specific Integrated Circuit, 简称 ASIC)、数字信号处理器 (Digital Signal Processor, 简称 DSP)、数字信号处理设备 (Digital Signal Processing Device, 简称 DSPD)、可编程逻辑器件 (Programmable Logic Device, 简称 PLD)、
20 现场可编程门阵列 (Field Programmable Gate Array, 简称 FPGA)、控制器、微控制器、微处理器或其他电子元件实现，用于执行音频合成方法。

当然，图 1 所示的电子设备 100 的结构并不构成对本申请实施例中电子设备的限定，在实际应用中电子设备 100 可以包括比图 1 所示的更多或更少的部件，或者组合某些部件。

25 可以理解的是，本申请实施例中并不对电子设备的数量进行限定，其可以是多个电子设备共同协作完成音频合成方法。在一种可能的实施方式中，请参考图 2，图 2 为本申请实施例提供的另一种音频合成方法所适用的硬件组成框架示意图。由图 2 可知，该硬件组成框架可以包括：第一电子设备 11 和第二电子设备 12，二者之间通过网络 13 连接。

30 在本申请实施例中，第一电子设备 11 与第二电子设备 12 的硬件结构

可以参考图 1 中电子设备 100。即可以理解为本实施例中具有两个电子设备 100，两者进行数据交互。进一步，本申请实施例中并不对网络 13 的形式进行限定，即，网络 13 可以是无线网络（如 WIFI、蓝牙等），也可以是有线网络。

5 其中，第一电子设备 11 和第二电子设备 12 可以是同一种电子设备，如第一电子设备 11 和第二电子设备 12 均为服务器；也可以是不同类型的电子设备，例如，第一电子设备 11 可以是智能手机或其它智能终端，第二电子设备 12 可以是服务器。在一种可能的实施方式中，可以利用计算能力强的服务器作为第二电子设备 12 来提高数据处理效率及可靠性，进而提高
10 音频合成的处理效率。同时利用成本低，应用范围广的智能手机作为第一电子设备 11，用于实现第二电子设备 12 与用户之间的交互。可以理解的是，该交互过程可以为：智能手机获取歌曲信息，并将歌曲信息发送至服务器，由服务器进行合成后音频的生成。服务器将合成后音频发送至智能手机，由智能手机对合成后音频进行播放。

15 具体的，请参考图 3，图 3 为本申请实施例提供的一种音频合成方法的流程示意图。该实施例中的方法包括：

S101：获取待处理歌曲音频及相应的歌曲信息。

20 其中，歌曲信息，是指用于对歌曲进行标识的信息，其可以包括歌曲名称、歌手名称等。此外，歌曲信息能够参与生成待合成音频，为了丰富待合成音频的内容，歌曲信息还可以包括附加信息，例如歌曲创作背景、创作目的等信息。

待处理歌曲音频，是指需要被进行合成处理以得到具有播音效果的原始歌曲音频，播音效果即在歌曲中插入与歌曲相关的语音后取得的效果，其能够使得歌曲播放时具有更好的听感。

25 根据歌曲信息的内容的不同，获取待处理歌曲音频的方式不同。例如当歌曲信息中本身即包括待处理歌曲音频，则可以直接从歌曲信息中筛选得到待处理歌曲音频。若歌曲信息中不包括待处理歌曲音频，则可以根据歌曲名、歌手名等信息唯一确定待处理歌曲音频的身份，并从预设的歌曲音频文件库中获取待处理歌曲音频。

30 S102：对待处理歌曲音频进行人声分离处理，得到人声音频。

为了能够使得待合成音频中的人声音色与待处理歌曲音频的歌手人声音色相近，进而使得整个合成后音频中的人生音色保持相对一致，提高播放效果，本申请以待处理歌曲音频的歌手人声作为基础，在候选的多种音色中选择最为相近的一个，并生成待合成音频。为实现该目的，首先需要

5 确定待处理歌曲音频的歌手人声音色。具体的，对待处理歌曲音频进行人声分离处理，得到人声音频，人声音频在听感上是歌手在没有伴奏器乐时的清唱效果。本实施例并不限定人声分离处理的具体方式，例如可以利用训练好的人声分离模型进行处理，人声分离模型可以利用清澈人声音频和干净伴奏音频训练得到。在进行人声分离处理时，将待处理歌曲音频输入

10 人声分离模型，人声分离模型将人声音频和伴奏音频进行拆分并输出。人声分离模型的具体类型不做限定，例如可以选择卷积神经网络（Convolutional Neural Network, CNN）、深度神经网络（Deep Neural Network, DNN）、循环神经网络（Recurrent Neural Network, RNN）或复数域的 UNET 模型等。

15 S103：基于人声音频，在多个候选音色信息中确定目标音色信息。

候选音色信息，是指多个候选人声音色分别对应的标识信息，目标音色信息是其中与人声音频中的歌手音色最相近的一个。目标音色信息用于在语音合成处理时选择采用的人声音色，使得待合成音频中的人声音色与歌手人生音色近似。音色信息的具体形式不做限定，例如可以为候选人声

20 编号或名称。

本实施例并不限定目标音色信息的具体确定方式，具体的，可以提取人声音频中能够表征音色的特征，并将其与各个候选音色信息分别对应的表征音色的特征进行匹配，将其中最为接近的确定为目标音色信息。在一种实施方式中，可以提取人声音频对应的声纹特征向量，声纹特征向量的

25 具体提取方式不做限定，其能够表征人声音频的音色即可。此外，预先生成各个候选声纹特征向量，各个候选声纹特征向量分别与各个候选音色信息相对应，其通过对各个候选音色信息对应的候选人说话音频进行声纹特征提取得到。可以理解的是，所有声纹特征向量的提取方式相同，例如可以采用 I-Vector 模型或 X-Vector 模型进行特征提取。生成声纹特征向量分

30 别与各个候选声纹特征向量之间的余弦相似度，并将最大余弦相似度对应

的候选音色信息确定为目标音色信息。可以理解的是，最大余弦相似度可以表明二者最为接近。

在另一种实施方式中，可以将人声音频输入基于候选音频训练得到的音色匹配模型，得到人声音频与各个候选音频之间的相似概率。具体的，
5 利用各个候选人说话音频训练得到音色匹配模型，将人声音频输入其中后，音色匹配模型能够输出人声音频与各个候选音频（即候选人说话音频）之间的相似概率，相似概率是指音色匹配模型认为人声音频与候选音频的音色相同的概率。将最大相似概率对应的候选音频的候选音色信息确定为目标音色信息，最大相似概率表明人声音频与候选音频音色相同的可能性最大，即二者最为相似。
10

S104：获取文本模板，并基于文本模板和歌曲信息生成待处理文本。

待处理文本，是指待合成音频对应的内容文本，其基于歌曲信息和文本模板生成，基于歌曲信息生成的待处理文本与待处理歌曲音频具有较强的相关关系，进而使得待合成音频与待处理歌曲音频相关。通常情况下，
15 歌曲信息的内容较多，例如可以包括多项歌曲子信息，在这种情况下，可以基于各个歌曲子信息的信息类型，将各个歌曲子信息插入文本模板中的对应位置，得到待处理文本。例如，歌曲信息包括歌曲名、歌手名称和附加信息三项歌曲子信息，文本模板为：“欢迎收听 A 演唱的歌曲 B，这首歌曲想要 C”。其中 A 为歌手名，B 为歌曲名，C 为歌曲目的信息，例如可以
20 以为“鼓励大家在困境中保持积极心态”，其属于附加信息。文本模板可以具有不同类型，通过这种文字组织方式，将歌曲信息和想要表达的文字内容进行整合，获得多种不同类型的待处理文本，模板类型包括但不限于鼓励类型、平铺直叙类型、开心分享类型、搞怪类型等。

本实施例并不限定获取文本模板的具体方式，在一种实施方式中，可
25 以根据模板信息获取文本模板。模板信息，是指用于描述生成待合成音频对应的待处理文本的文本模板的信息，其具体可以为标识文本模板的标识符，例如序号或名称。本实施例并不限定模板信息和歌曲信息的具体获取方式，例如在一种实施方式中，可以通过数据交互部件，例如键盘、鼠标、触摸屏等与用户进行交互，根据用户通过数据交互部件传输的数据得到上
30 述两种信息中的任意一种。在另一种实施方式中，可以与其他电子设备进

行数据交互，或从指定的存储路径下获取歌曲信息和/或模板信息。

需要说明的是，待处理文本的数量可以为一个或多个，即对于一个待处理歌曲音频来说，其可以对应多个待合成音频，每个待合成音频对应的待处理文本的内容可以不同。

- 5 S105：基于目标音色信息，对待处理文本进行语音合成处理，得到待合成音频。

利用目标音色信息，可以指定待合成音频的人声音色，利用待处理文本可以确定待合成音频的人声内容。因此基于目标音色信息，对待处理文本进行语音合成处理，即可得到待合成音频，本实施例并不限定语音合成处理的10 具体方式，在一种可行的实施方式中，将目标音色信息和待处理文本输入特征编码模型，得到文本编码，特征编码模型用于基于目标音色信息确定编码方式，编码方式与人声音色相关，并采用该编码方式对待处理文本进行编码。声码器模型用于根据文本编码输出对应的音频，将文本编码输入声码器模型，得到待合成音频。

15 可以理解的是，声码器模型对文本编码的识别方式应当与特征编码模型相匹配。为了保证这一点，可以将声码器模型和特征编码模型共同训练。在训练过程中，首先获取各个候选音色信息分别对应的训练数据；其中，训练数据包括训练文本和对应的候选音频，候选音频对应的各个候选人音色与 S102 步骤中的候选音频的候选人音色一致，具体可以由几个候选人朗20 诵相同的不同的内容得到。训练文本数据的具体内容不做限定，例如可以包括文字、汉语拼音数据、停顿标识数据等，停顿标识数据用于表示文字在何处断句或间隔停顿。将训练文本输入初始特征编码模型，初始特征编码模型基于文本提取对应的编码，例如可以为梅尔频谱编码提取，得到训练编码。

25 将训练编码输入初始声码器模型，初始声码器模型对训练编码进行识别，即可训练输出音频。对候选音频提取与训练编码类型相同的标签编码，例如当初始特征编码模型进行梅尔频谱编码的提取，则对候选音频提取梅尔频谱编码，得到标签编码，并利用标签编码和训练编码生成第一损失值。相应的，将候选音频作为另一个标签，利用候选音频与训练输出音频生成30 第二损失值。利用第一损失值对初始特征编码模型进行参数调节，并利用

第二损失值对初始声码器模型进行特征调节，使得初始特征编码模型和初始声码器模型学习相关的内容。当检测到满足训练完成条件，将参数调节后的初始特征编码模型确定为特征编码模型，并将参数调节后的初始声码器模型确定为声码器模型。训练完成条件的数量和具体内容可以根据需要设置，例如可以为准确率条件、训练时间条件或训练轮次条件。

需要说明的是，声码器模型和特征编码模型的模型类型不做限制，例如声码器模型可以采用 Tacotron, FastSpeech 等模型结构。

S106：对待合成音频和待处理歌曲音频进行合成处理，得到合成后音频。

10 通过将二者进行合成，即可得到质量较高，全程音色相近且具有良好的听感的合成后音频。可以理解的是，为了保持待处理歌曲音频的完整性，可以在待处理歌曲音频中没有人声的位置合成待合成音频。具体的，可以确定待处理歌曲音频对应的伴奏时间区间，伴奏时间区间是指不具有人声，仅有伴奏的时间区间。模板信息可以表征其适配的位置，例如通常可以在歌曲开始前对即将播放的歌曲进行介绍，在歌曲播放结束后对歌曲的创作目的进行介绍，在歌曲播放过程中对歌曲的创作背景进行介绍等，具体可以进行设置。基于模板信息确定合成位置类型，并利用合成位置类型对伴奏时间区间进行筛选，得到目标伴奏时间区间。每个待合成音频可以对应于一个或多个目标伴奏时间区间。

20 若目标伴奏时间区间的长度不小于待合成音频的时间长度，则说明在目标伴奏时间区间内，待合成音频可以播放完毕，在这种情况下，可以将待合成音频与待处理歌曲音频中目标伴奏时间区间对应的部分合成，以便在播放伴奏的同时播放待合成音频。若目标伴奏时间区间的长度小于待合成音频的时间长度，为了保持待处理歌曲音频的完整性，可以基于目标伴奏时间区间的类型，对待合成音频和待处理歌曲音频进行拼接处理，即二者在时间上不会重叠，演唱待处理歌曲的时长。具体的，若目标伴奏时间区间的类型为前奏，则将待合成音频拼接至待处理歌曲音频之前，若目标伴奏时间区间的类型为尾奏，则将待合成音频拼接至待处理歌曲音频之后。请参考图 4，图 4 为本申请实施例提供的一种音频合成位置示意图，可以看出，若目标伴奏时间区间的类型为前奏，则可以将待合成音频（即语音）

拼接到歌曲音频的前方，若为尾奏则拼接到歌曲音频的后方。本实施例并不限定为间奏时的处理方式，例如可以随机选择拼接至前方或后方，或者可以指定拼接在一个位置，例如后方。

此外，在将待合成音频与待处理歌曲音频中目标伴奏时间区间对应的部分合成时，为了取得更好的听感，可以对待合成音频和目标伴奏时间区间对应的部分进行音量调节。当目标伴奏时间区间对应的部分为前奏时，可以将前奏的音量大小设置为从小到大，最终恢复到正常音量，而将待合成音频的音量设置为从大到小，起到逐渐引入歌曲的效果。若目标伴奏时间区间对应的部分为尾奏，则可以将尾奏的音量大小设置为从大到小，而将待合成音频的音量设置为从小到达，起到逐渐引出歌曲的效果，若目标伴奏时间区间对应的部分对应为间奏，则可以将间奏的音量设置为一直较小，小于待合成音频的音量，或采取和前奏相同的设置方式。

本实施例并不限定伴奏时间区间的具体获取方式，在一种实施方式中，可以获取待处理歌曲音频对应的歌词，根据歌词的时间戳确定伴奏的时间区间。例如，对于周杰伦的《七里香》这首歌曲，通过歌词解析，得到首句歌词“窗外的麻雀在电线杆上多嘴”是从约 28 秒开始，那么 0 至 28 秒即为前奏；而对于歌词“雨下整夜我的爱溢出就像雨水”这句是在约 162 秒开始，而其上一句是在约 135 秒处结束，其间约有 27 秒的无人声部分，即为间奏；对于最后一句歌词在约 270 秒结束，而歌曲是在 299 秒处结束，这 29 秒的无人声部分即为尾奏。从 0 至 28 秒、从 135 秒至 162 秒、从 270 秒至 299 秒为三个伴奏时间区间，且三个伴奏时间区间对应于不同的合成位置类型，分别为前、中和后。

在另一种实施方式中，若没有歌词，则可以利用人声检测模型对待处理歌曲音频进行人声检测，得到多个起始时间点和终止时间点。由于歌手演唱时，相邻的两句歌词中间同样具有间隔，但是该间隔中不应当插入待合成音频。为了避免存在较短的待合成音频被插入相邻的两句歌词中间，可以将距离小于预设阈值的相邻的终止时间点和起始时间点去除，得到目标起始时间点和目标终止时间点，其中距离是指时间间隔距离，预设阈值的具体大小不做限定。利用目标起始时间点和目标终止时间点可以确定演唱部分的时间区间，同样可以确定伴奏时间区间。

请参考图 5，图 5 为本申请实施例提供的一种音频合成全流程示意图。其中，文字即表示上述的文本模板，利用歌曲信息，可以确定待处理歌曲音频，进而利用待处理歌曲音频进行人声分离和音色识别处理，得到目标音色信息。利用文字和歌曲信息生成待处理文本后，利用语音合成技术，

5 基于目标音色信息对待处理文本进行语音合成，得到合成语音，即上述的待合成音频，并利用自动混叠技术对合成语音和歌曲信息对应的待处理歌曲音频进行音频合成，得到长音频，即合成后音频。

应用本申请实施例提供的音频合成方法，通过人声分离处理，并利用得到的人声音频在多个候选音色信息中确定目标音色信息，目标音色信息

10 为与待处理歌曲音频中的人声部分音色相近的音色信息。在生成待处理文本后，基于目标音色信息生成的待合成音频中的人声语音的内容与待处理文本相同，且音色与待处理歌曲音频的演唱者的音色相近。将待合成音频和待处理歌曲音频合成得到的合成后音频中，人声部分音色前后保持相近，具有良好的听感。同时待合成音频的质量较高，不受限于人工播音水平，

15 使得合成后音频的质量较高，播放效果较好，解决了需要人工参与，效果受限于主播的播音水平的问题。

下面对本申请实施例提供的计算机可读存储介质进行介绍，下文描述的计算机可读存储介质与上文描述的音频合成方法可相互对应参照。

本申请还提供一种计算机可读存储介质，计算机可读存储介质上存储

20 有计算机程序，计算机程序被处理器执行时实现上述的音频合成方法的步骤。

该计算机可读存储介质可以包括：U 盘、移动硬盘、只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

25 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，各个实施例之间相同或相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

本领域技术人员还可以进一步意识到，结合本文中所公开的实施例描

30 述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的

结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件的方式来执行，取决于技术方案的特定应用和设计约束条件。本领域技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应该认为超出本申请的范围。

结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系属于仅仅用来将一个实体或者操作与另一个实体或者操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语包括、包含或者其他任何变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。

本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权 利 要 求

- 1、一种音频合成方法，其特征在于，包括：
获取待处理歌曲音频及相应的歌曲信息；
对所述待处理歌曲音频进行人声分离处理，得到人声音频；
5 基于所述人声音频，在多个候选音色信息中确定目标音色信息；
获取文本模板，并基于所述文本模板和所述歌曲信息生成待处理文本；
基于所述目标音色信息，对所述待处理文本进行语音合成处理，得到待合成音频；
对所述待合成音频和所述待处理歌曲音频进行合成处理，得到合成后
10 音频。
- 2、根据权利要求1所述的音频合成方法，其特征在于，所述基于所述人声音频，在多个候选音色信息中确定目标音色信息，包括：
提取所述人声音频对应的声纹特征向量；
生成所述声纹特征向量分别与各个候选声纹特征向量之间的余弦相似
15 度；各个所述候选声纹特征向量分别与各个所述候选音色信息相对应；
将最大余弦相似度对应的候选音色信息确定为所述目标音色信息。
- 3、根据权利要求1所述的音频合成方法，其特征在于，所述基于所述人声音频，在多个候选音色信息中确定目标音色信息，包括：
将所述人声音频输入基于候选音频训练得到的音色匹配模型，得到所
20 述人声音频与各个所述候选音频之间的相似概率；
将最大相似概率对应的候选音频的候选音色信息确定为所述目标音色信息。
- 4、根据权利要求1所述的音频合成方法，其特征在于，所述歌曲信息包括多项歌曲子信息，所述基于所述文本模板和所述歌曲信息生成待处理
25 文本，包括：
基于各个歌曲子信息的信息类型，将各个所述歌曲子信息插入所述文本模板中的对应位置，得到所述待处理文本。
- 5、根据权利要求1所述的音频合成方法，其特征在于，所述基于所述目标音色信息，对所述待处理文本进行语音合成处理，得到待合成音频，
30 包括：

将所述目标音色信息和所述待处理文本输入特征编码模型，得到文本编码；

将所述文本编码输入声码器模型，得到所述待合成音频。

6、根据权利要求 5 所述的音频合成方法，其特征在于，所述特征编码模型和所述声码器模型的训练过程，包括：

获取各个候选音色信息分别对应的训练数据；其中，所述训练数据包括训练文本和对应的候选音频；

将所述训练文本输入初始特征编码模型，得到训练编码；

将所述训练编码输入初始声码器模型，得到训练输出音频；

10 对所述候选音频提取与所述训练编码类型相同的标签编码，并利用所述标签编码和所述训练编码生成第一损失值；

利用所述候选音频与所述训练输出音频生成第二损失值；

利用所述第一损失值对所述初始特征编码模型进行参数调节，并利用所述第二损失值对所述初始声码器模型进行特征调节；

15 当检测到满足训练完成条件，将参数调节后的初始特征编码模型确定为所述特征编码模型，并将参数调节后的初始声码器模型确定为所述声码器模型。

7、根据权利要求 1 所述的音频合成方法，其特征在于，所述对所述待合成音频和所述待处理歌曲音频进行合成处理，包括：

20 确定所述待处理歌曲音频对应的伴奏时间区间；

基于所述模板信息确定合成位置类型，并利用所述合成位置类型对所述伴奏时间区间进行筛选，得到目标伴奏时间区间；

25 若所述目标伴奏时间区间的长度不小于所述待合成音频的时间长度，则将所述待合成音频与所述待处理歌曲音频中所述目标伴奏时间区间对应的部分合成；

和/或，

若所述目标伴奏时间区间的长度小于所述待合成音频的时间长度，则基于所述目标伴奏时间区间的类型，对所述待合成音频和所述待处理歌曲音频进行拼接处理。

30 8、根据权利要求 7 所述的音频合成方法，其特征在于，所述基于所述

目标伴奏时间区间的类型，对所述待合成音频和所述待处理歌曲音频进行拼接处理，包括：

若所述目标伴奏时间区间的类型为前奏，则将所述待合成音频拼接至所述待处理歌曲音频之前；

- 5 若所述目标伴奏时间区间的类型为尾奏，则将所述待合成音频拼接至所述待处理歌曲音频之后。

9、根据权利要求7所述的音频合成方法，其特征在于，所述确定所述待处理歌曲音频对应的伴奏时间区间，包括：

- 10 利用人声检测模型对待处理歌曲音频进行人声检测，得到多个起始时间点和终止时间点；

将距离小于预设阈值的相邻的终止时间点和起始时间点去除，得到目标起始时间点和目标终止时间点；

利用所述目标起始时间点和目标终止时间点确定所述伴奏时间区间。

- 10、一种电子设备，其特征在于，包括存储器和处理器，其中：
15 所述存储器，用于保存计算机程序；
 所述处理器，用于执行所述计算机程序，以实现如权利要求1至9任一项所述的音频合成方法。

- 11、一种计算机可读存储介质，其特征在于，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求1至9任一项所述的
20 的音频合成方法。