

定制音色歌声的合成方法、装置、电子设备和存储介质

本公开要求于 2021 年 1 月 20 日提交的申请号为 202110076168.0、名称为“定制音色
歌声的合成方法、装置、电子设备和存储介质”的中国专利申请的优先权，该中国专利申
5 请的全部内容通过引用全部并入本文。

技术领域

本公开涉及声音合成技术领域，尤其涉及一种定制音色歌声的合成方法、装置、电子
10 设备和存储介质。

背景技术

随着人工智能行业的高速发展，智能语音合成技术已经渗透多个领域，被应用于：智
能家居、语音导航、智能客服等业务，且人工智能合成的语音拟人度高，可达到替代人工
的标准。为满足用户音色多样性的需求，现有音色定制功能也日渐成熟，可以通过用户少
15 量语料音频训练出该用户专属音色。同时，随着虚拟偶像的知名度日渐提升，歌声合成也
成为了语音合成技术的主要发展方向之一。目前现有的歌声合成技术，例如虚拟偶像所依
托的“VOCALOID”合成器，主要依靠于通过真人语音数据建立语料库，再对用户提供的
歌词进行字词切割，进而从语料库中调取对应的语料，最后按照用户提供的乐谱调节合
成语音的节拍和音高来合成歌声。

20 相关技术中，由于歌声合成技术建立语料库需要用户长时间的语音数据，且调用语料
库生成语音的过程需要耗费大量时间，导致歌声合成的效率低。另外，由于语料库具备体
量大的特性，用户对于音色定制的需求只能通过替换整个语料库才可完成，过程繁琐且耗
时长。

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，
25 因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

本公开的目的在于提供一种定制音色歌声的合成方法、装置、电子设备和存储介质，
至少在一定程度上克服由于相关技术中定制音色歌声的合成效率低的问题。

30 本公开的其他特性和优点将通过后续的详细描述变得显然，或部分地通过本公开的实
践而习得。

根据本公开的一个方面，提供一种定制音色歌声的合成方法，包括：通过说话人录音
样本对第一神经网络进行训练，以得到说话人识别模型，第一神经网络输出训练结果为说
话人向量样本；通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清
35 唱合成模型；将待合成的说话人录音输入至说话人识别模型，获取说话人识别模型的中间

隐藏层输出的说话人信息；将待合成的清唱音乐信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声。

在本公开的一个实施例中，通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型包括：将说话人录音样本划分为测试录音样本和注册录音样本，并输入至第一神经网络；注册录音样本经第一神经网络输出注册录音特征，将注册录音特征进行平均池化处理，以得到注册录音向量；测试录音样本经第一神经网络输出测试录音向量；对注册录音向量和测试录音向量进行余弦相似度计算，以获得余弦相似度结果；通过余弦相似度结果和回归函数对第一神经网络进行参数优化，至回归函数的损失值最小为止；将参数优化后的第一神经网络确定为说话人识别模型。

在本公开的一个实施例中，清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：解析歌声清唱样本中的乐谱样本、歌词样本和音素时长样本；通过说话人向量样本、乐谱样本、歌词样本和音素时长样本对持续时间模型进行训练，持续时间模型的输出结果为时长预测样本。

在本公开的一个实施例中，清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：解析歌声清唱样本中的乐谱样本、歌词样本和音素时长样本；根据歌声清唱样本中的歌曲提取梅尔图谱样本；通过说话人向量样本、音素时长样本、乐谱样本、歌词样本和梅尔图谱样本对声学模型进行训练，声学模型输出结果为梅尔谱图预测样本。

在本公开的一个实施例中，清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：根据歌声清唱样本中的歌曲提取梅尔图谱样本；通过梅尔谱图样本对声码器模型进行训练，声码器模型的输出结果为音频预测样本。

在本公开的一个实施例中，将待合成的清唱音乐信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声包括：解析清唱音乐信息中的待合成乐谱和待合成歌词；将说话人信息、待合成乐谱和待合成歌词输入至持续时间模型，持续时间模型输出结果为待合成的时长预测结果；将时长预测结果、说话人信息、待合成乐谱和待合成歌词输入至声学模型，声学模型输出结果为待合成的梅尔谱图预测结果；将梅尔谱图预测结果输入至声码器模型，声码器模型输出结果为合成的定制音色歌声。

在本公开的一个实施例中，解析清唱音乐信息中的待合成乐谱和待合成歌词包括：对清唱音乐信息中的乐谱和歌词进行文本分析和特征提取，以获取待合成乐谱和待合成歌词。

在本公开的一个实施例中，将时长预测结果、说话人信息、待合成乐谱和待合成歌词输入至声学模型，声学模型输出结果为待合成的梅尔谱图预测结果包括：对时长预测结果、待合成乐谱和待合成歌词进行帧级扩展；将帧级扩展的结果和说话人信息输入至声学模

型，声学模型输出结果为待合成的梅尔谱图预测结果。

根据本公开的另一个方面，提供一种定制音色歌声的合成装置，包括：第一训练模块，用于通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型，第一神经网络输出训练结果为说话人向量样本；第二训练模块，用于通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型；识别模块，用于将待合成的说话人录音输入至说话人识别模型，获取说话人识别模型的中间隐藏层输出的说话人信息；合成模块，用于将待合成的清唱音乐信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声

根据本公开的再一个方面，提供一种电子设备，包括：处理器；以及存储器，用于存储处理器的可执行指令；其中，处理器配置为经由执行可执行指令来执行上述任意一项的定制音色歌声的合成方法。

根据本公开的又一个方面，提供一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现上述任意一项的定制音色歌声的合成方法。

本公开的实施例所提供的定制音色歌声的合成方案，通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型，第一神经网络输出训练结果为说话人向量样本，并通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型，提高了模型合成的效率，不需要采集大量录音数据来生成语料库。

进一步地，通过将待合成的说话人录音输入至说话人识别模型，获取说话人识别模型的中间隐藏层输出的说话人信息，以及将待合成的清唱音乐信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声，只需少量语料即可训练出用户定制音色，并通过调整合成语音的节奏和音高达到歌声合成的效果，减少了定制音色歌声合成过程中需要的时间和训练语料，提升了定制音色歌声的合成效率。

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

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。

图 1 示出本公开实施例中一种定制音色歌声的合成方法的示意图；

图 2 示出本公开实施例中另一种定制音色歌声的合成方法的流程图；

图 3 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；

图 4 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；

图 5 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；

图 6 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 7 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 8 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 9 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
5 图 10 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 11 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 12 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 13 示出本公开实施例中又一种定制音色歌声的合成方法的流程图；
图 14 示出本公开实施例中一种定制音色歌声的合成装置的示意图；
10 图 15 示出本公开实施例中一种电子设备的示意图。

具体实施方式

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。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方式中。

此外，附图仅为本公开的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。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。

本公开提供的方案，通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型，第一神经网络输出训练结果为说话人向量样本，并通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型，提高了模型合成的效率，不需要采集大量录音数据来生成语料库。

进一步地，通过将待合成的说话人录音输入至说话人识别模型，获取说话人识别模型的中间隐藏层输出的说话人信息，以及将待合成的清唱音乐信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声，只需少量语料即可训练出用户定制音色，并通过调整合成语音的节奏和音高达到歌声合成的效果，减少了定制音色歌声合成过程中需要的时间和训练语料，提升了定制音色歌声的合成效率。

上述定制音色歌声的合成方案可以通过多个终端和服务器的交互实现。
终端可以是手机、游戏主机、平板电脑、电子书阅读器、智能眼镜、MP4(Moving Picture Experts Group Audio Layer IV, 动态影像专家压缩标准音频层面 4) 播放器、智能家居设备、AR (Augmented Reality, 增强现实) 设备、VR (Virtual Reality, 虚拟现实) 设备等移动终端，或者，终端也可以是个人计算机 (Personal Computer, PC)，比如膝上型便携

计算机和台式计算机等等。

其中，终端中可以安装有用于提供定制音色歌声的合成的应用程序。

终端与服务器集群之间通过通信网络相连。可选的，通信网络是有线网络或无线网络。

服务器集群是一台服务器，或者由若干台服务器组成，或者是一个虚拟化平台，或者是一个云计算服务中心。服务器集群用于为提供定制音色歌声的合成的应用程序提供后台服务。可选地，服务器集群承担主要计算工作，终端承担次要计算工作；或者，服务器集群承担次要计算工作，终端承担主要计算工作；或者，终端和服务器集群之间采用分布式计算架构进行协同计算。

可选地，不同的终端中安装的应用程序的客户端是相同的，或两个终端上安装的应用程序的客户端是不同控制系统平台的同一类型应用程序的客户端。基于终端平台的不同，该应用程序的客户端的具体形态也可以不同，比如，该应用程序客户端可以是手机客户端、PC 客户端或者全球广域网客户端等。

本领域技术人员可以知晓，上述终端的数量可以更多或更少。比如上述终端可以仅为一个，或者上述终端为几十个或几百个，或者更多数量。本公开实施例对终端的数量和设备类型不加以限定。

可选的，该系统还可以包括管理设备，该管理设备与服务器集群之间通过通信网络相连。可选的，通信网络是有线网络或无线网络。

可选的，上述的无线网络或有线网络使用标准通信技术和/或协议。网络通常为因特网、但也可以是任何网络，包括但不限于局域网（Local Area Network，LAN）、城域网（Metropolitan Area Network，MAN）、广域网（Wide Area Network，WAN）、移动、有线或者无线网络、专用网络或者虚拟专用网络的任何组合）。在一些实施例中，使用包括超文本标记语言（Hyper Text Mark-up Language，HTML）、可扩展标记语言（Extensible MarkupLanguage，XML）等的技术和/或格式来代表通过网络交换的数据。此外还可以使用诸如安全套接字层（Secure Socket Layer，SSL）、传输层安全（Transport Layer Security，TLS）、虚拟专用网络（Virtual Private Network，VPN）、网际协议安全（Internet ProtocolSecurity，IPsec）等常规加密技术来加密所有或者一些链路。在另一些实施例中，还可以使用定制和/或专用数据通信技术取代或者补充上述数据通信技术。

下面，将结合附图及实施例对本示例实施方式中的定制音色歌声的合成方法的各个步骤进行更详细的说明。

图 1 示出本公开实施例中一种定制音色歌声的合成方法流程图。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。

如图 1 所示，电子设备执行定制音色歌声的合成方法，包括以下步骤：

步骤 S102，通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型，第一神经网络输出训练结果为说话人向量样本。

步骤 S104，通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到

清唱合成模型。

步骤 S106, 将待合成的说话人录音输入至说话人识别模型, 获取说话人识别模型的中间隐藏层输出的说话人信息。

5 步骤 S108, 将待合成的清唱音乐信息和说话人信息输入至清唱合成模型, 以获得合成的定制音色歌声。

在本公开的一个实施例中, 通过说话人录音样本对第一神经网络进行训练, 以得到说话人识别模型, 第一神经网络输出训练结果为说话人向量样本, 并通过歌声清唱样本和说话人向量样本对第二神经网络进行训练, 以得到清唱合成模型, 提高了模型合成的效率, 不需要采集大量录音数据来生成语料库。

10 进一步地, 通过将待合成的说话人录音输入至说话人识别模型, 获取说话人识别模型的中间隐藏层输出的说话人信息, 以及将待合成的清唱音乐信息和说话人信息输入至清唱合成模型, 以获得合成的定制音色歌声, 只需少量语料即可训练出用户定制音色, 并通过调整合成语音的节奏和音高达到歌声合成的效果, 减少了定制音色歌声合成过程中需要的时间和训练语料, 提升了定制音色歌声的合成效率。

15 基于图 1 所示的步骤, 如图 2 所示, 通过说话人录音样本对第一神经网络进行训练, 以得到说话人识别模型包括:

步骤 S2022, 将说话人录音样本划分为测试录音样本和注册录音样本, 并输入至第一神经网络。

20 步骤 S2024, 注册录音样本经第一神经网络输出注册录音特征, 将注册录音特征进行平均池化处理, 以得到注册录音向量。

在本公开的一个实施例中, 平均池化处理的前向传播就是把一个块中的值求取平均来做池化, 那么反向传播的过程也就是把一个元素的梯度等分为 n 份分配给前一层, 这样就保证池化前后的梯度(残差)之和保持不变, 平均池化处理能够减小邻域大小受限造成的估计值方差的误差, 更多的保留特征信息。

25 步骤 S2026, 测试录音样本经第一神经网络输出测试录音向量。

步骤 S2028, 对注册录音向量和测试录音向量进行余弦相似度计算, 以获得余弦相似度结果。

步骤 S2030, 通过余弦相似度结果和回归函数对第一神经网络进行参数优化, 至回归函数的损失值最小为止。

30 步骤 S2032, 将参数优化后的第一神经网络确定为说话人识别模型。

在本公开的一个实施例中, 通过余弦相似度结果和回归函数对第一神经网络进行参数优化, 至回归函数的损失值最小为止, 以获得能够对说话人进行识别的说话人识别模型, 仅需要几秒的说话人录音即可完成识别。

35 基于图 1 所示的步骤, 如图 3 所示, 清唱合成模型包括持续时间模型、声学模型和声码器模型, 通过歌声清唱样本和说话人向量样本对第二神经网络进行训练, 以得到清唱合

成模型包括：

步骤 S3042，解析歌声清唱样本中的乐谱样本、歌词样本和音素时长样本。

步骤 S3044，通过说话人向量样本、乐谱样本、歌词样本和音素时长样本对持续时间模型进行训练，持续时间模型的输出结果为时长预测样本。

- 5 在本公开的一个实施例中，通过说话人向量样本、乐谱样本、歌词样本和音素时长样本对持续时间模型进行训练，持续时间模型的输出结果为时长预测样本，以实现合成后的清唱歌曲的时长预测结果，时长预测结果作为声学模型的一个输入量。

10 基于图 1 所示的步骤，如图 4 所示，清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：

步骤 S4042，解析歌声清唱样本中的乐谱样本、歌词样本和音素时长样本。

步骤 S4044，根据歌声清唱样本中的歌曲提取梅尔图谱样本。

步骤 S4046，通过说话人向量样本、音素时长样本、乐谱样本、歌词样本和梅尔图谱样本对声学模型进行训练，声学模型输出结果为梅尔谱图预测样本。

- 15 在本公开的一个实施例中，通过说话人向量样本、音素时长样本、乐谱样本、歌词样本和梅尔图谱样本对声学模型进行训练，声学模型输出结果为梅尔谱图预测样本，以得到合适大小的声音特征，通过梅尔谱图来模拟人耳对各种频率的声音的感知力，也即通过梅尔谱图强化低频部分，弱化高频部分，进而使清唱合成歌声更接近于自然人的歌声。

20 基于图 1 所示的步骤，如图 5 所示，清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：

步骤 S5042，根据歌声清唱样本中的歌曲提取梅尔图谱样本。

步骤 S5044，通过梅尔谱图样本对声码器模型进行训练，声码器模型的输出结果为音频预测样本。

- 25 在本公开的一个实施例中，通过梅尔谱图样本对声码器模型进行训练，声码器模型的输出结果为音频预测样本，从而得到符合说话人音色的音频预测样本。

基于图 1 和图 3 所示的步骤，如图 6 所示，将待合成的清唱音乐信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声包括：

步骤 S6082，解析清唱音乐信息中的待合成乐谱和待合成歌词。

- 30 步骤 S6084，将说话人信息、待合成乐谱和待合成歌词输入至持续时间模型，持续时间模型输出结果为待合成的时长预测结果。

步骤 S6086，将时长预测结果、说话人信息、待合成乐谱和待合成歌词输入至声学模型，声学模型输出结果为待合成的梅尔谱图预测结果。

- 35 步骤 S6088，将梅尔谱图预测结果输入至声码器模型，声码器模型输出结果为合成的定制音色歌声。

在本公开的一个实施例中，在定制音色歌声的合成过程中，通过说话人识别模型确定说话人，继而依次通过持续时间模型、声学模型和声码器模型根据说话人、待合成乐谱和待合成歌词，得到符合说话人音色的定制音色歌声。

5 基于图 1 和图 3 所示的步骤，如图 7 所示，解析清唱音乐信息中的待合成乐谱和待合成歌词包括：

步骤 S7082，对清唱音乐信息中的乐谱和歌词进行文本分析和特征提取，以获取待合成乐谱和待合成歌词。

10 在本公开的一个实施例中，通过对清唱音乐信息中的乐谱和歌词进行文本分析和特征提取，以获取待合成乐谱和待合成歌词，待合成歌词更符合说话人的吐字特点，待合成乐谱更符合说话人的清唱音色。

基于图 1 和图 3 所示的步骤，如图 8 所示，将时长预测结果、说话人信息、待合成乐谱和待合成歌词输入至声学模型，声学模型输出结果为待合成的梅尔谱图预测结果包括：

步骤 S8082，对时长预测结果、待合成乐谱和待合成歌词进行帧级扩展。

15 步骤 S8084，将帧级扩展的结果和说话人信息输入至声学模型，声学模型输出结果为待合成的梅尔谱图预测结果。

下面参照图 9 来描述根据本公开的这种实施方式的定制音色歌声的合成方案的示意图。图 9 所示的定制音色歌声的合成方案的训练阶段 900 仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

20 如图 9 所示，定制音色歌声的合成方案的训练阶段 900 包括：将识别数据输入至说话人识别模型进行训练；说话人识别模型输出为说话人信息；将清唱数据和说话人信息输入至清唱合成模型进行训练。

下面参照图 10 来描述根据本公开的这种实施方式的定制音色歌声的合成方案的示意图。图 10 所示的定制音色歌声的合成方案的合成阶段 1000 仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

25 如图 10 所示，定制音色歌声的合成方案的合成阶段 1000 包括：将文本录制数据输入至说话人识别模型，以获得说话人信息；将说话人信息、乐谱和歌词输入至清唱合成模型，以获得清唱歌声。

30 下面参照图 11 来描述根据本公开的这种实施方式的定制音色歌声的合成方案的示意图。图 11 所示的定制音色歌声的合成方案的说话人识别模型 1100 仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 11 所示，定制音色歌声的合成方案的说话人识别模型 1100 执行阶段包括：

35 (1) 将测试录音、注册录音 1、.....注册录音 N 输入至 LSTM，N 为大于 1 的整数，其中，LSTM (Long Short-Term Memory) 是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。测试录音经 LSTM 处理后输出说话人向量 1，注册录音经 LSTM 处理后的向量进行平均池化处理，得到说话人向

量 2。

(2) 对说话人向量 1 和说话人向量 2 进行余弦相似度计算, 并进行打分函数的计算。

(3) 通过逻辑回归处理确定打分函数的结果为接收或拒绝。

下面参照图 12 来描述根据本公开的这种实施方式的定制音色歌声的合成方案的示意图。图 12 所示的定制音色歌声的合成方案的清唱合成模型 1200 仅仅是一个示例, 不应对本公开实施例的功能和使用范围带来任何限制。

如图 12 所示, 定制音色歌声的合成方案的清唱合成模型 1200 包括音素时长模型 1202、声学模型 1204 和声码器模型 1206, 各模块的训练过程可以独立执行如下:

(1) 说话人向量、乐谱和歌词、音素时长输入至音素时长模型 1202 进行训练。

(2) 说话人向量、乐谱和歌词、音素时长和梅尔图谱输入至声学模型 1204 进行训练。

(3) 梅尔图谱和歌曲输入至声码器模型 1206 进行训练。

具体地, 定制音色歌声的合成方案包括说话人识别模型和清唱合成模型, 清唱合成模型包含训练过程和推理过程, 清唱合成模型包括音素时长模型、声学模型和神经网络声码器模型。

可例如, 音素时长模型可以是三层全连接层组成的 DNN (Deep Neural Networks, 深度神经网络) 模型, 输入是乐谱和歌词, 输出是音素时长。在预测时, 我们只知道乐谱。

可例如, 说话人向量在训练的时候也被加入到音素时长模型当中, 用于根据不同的说话人, 来得到不同的音素时长模型。

可例如, 声学模型的输入是乐谱和音素时长, 输出是梅尔谱图, 说话人向量也被输入到声学模型中。

可例如, 声码器模型的输入是梅尔谱图, 输出是音频。声学模型的一种可能的实现方法是由三层 LSTM 组成的深度神经网络模型, 也可以是复杂的带有注意力机制的模型。

可例如, 声码器模型可以采用 LPCNet (Improving Neural Speech Synthesis Through Linear Prediction, 通过语音合成线性预测来改善神经系统) 声码器。

可例如, 在推理过程中, 已知乐谱和歌词, 以及歌唱者的说话人向量, 然后使用训练过程中得到的音素时长模型、声学模型和声码器模型, 可以输出合成的歌曲。

可例如, 在训练和推理的过程中, 都需要从乐谱和歌词出发, 进行文本分析、提取音素、分词等特征, 然后进行音素时长的预测或者音素时长模型的训练。

下面参照图 13 来描述根据本公开的这种实施方式的定制音色歌声的合成方案的示意图。图 13 所示的定制音色歌声的合成方案仅仅是一个示例, 不应对本公开实施例的功能和使用范围带来任何限制。

如图 13 所示, 定制音色歌声的合成方案的执行步骤包括:

步骤 S1302, 输入待合成的说话人向量。

步骤 S1304, 获取乐谱和歌词。

步骤 S1306, 文本分析。

步骤 S1308, 特征提取。

步骤 S1310, 音素时长预测。

步骤 S1312, 帧级扩展。

步骤 S1314, 梅尔图谱预测。

5 步骤 S1316, 合成歌曲。

其中, 说话人向量是根据说话人的声学特征, 使用深度神经网络提取的。通过说话人信息, 可以控制合成的清唱声音的音色。本公开训练一个清唱合成模型, 该清唱合成模型是在一个大量的说话人录制的朗读和清唱歌声数据集上训练得到的。当一个新的说话人需要合成清唱的歌声时, 只需要录制该说话人少量的朗读语料, 提取该说话人的说话人向量, 10 然后输入到该清唱合成模型, 结合乐谱和歌词, 通过该清唱合成模型的推理过程, 生成该说话人的清唱的声音, 即定制合成歌声。

可例如, 构建一个包含多音色、多歌声的清唱数据集, 用以训练一个给定乐谱和歌词可以合成清唱声音基础模型, 另外, 数据集还需要包含一部分指定文本的歌手录音数据。

可例如, 训练一个文本相关的说话人识别模型, 取模型中间隐藏层结果, 定义为说话 15 人向量。

可例如, 将歌手的指定文本录音, 送入说话人识别模型中, 得到说话人向量, 再将这个说话人向量作为清唱歌声模型的一部分, 通过大量多音色、多歌声清唱数据集, 训练一个清唱合成模型, 清唱合成模型中包含歌手身份信息。

下面参照图 14 来描述根据本公开的这种实施方式的定制音色歌声的合成装置 1400。 20 图 14 所示的定制音色歌声的合成装置 1400 仅仅是一个示例, 不应对本公开实施例的功能和使用范围带来任何限制。

如图 14 所示, 定制音色歌声的合成装置 1400 以硬件模块的形式表现。定制音色歌声的合成装置 1400 的组件可以包括但不限于: 第一训练模块 1402、第二训练模块 1404、识别模块 1406 和合成模块 1408。

25 第一训练模块 1402, 用于通过说话人录音样本对第一神经网络进行训练, 以得到说话人识别模型, 第一神经网络输出训练结果为说话人向量样本。

第二训练模块 1404, 用于通过歌声清唱样本和说话人向量样本对第二神经网络进行训练, 以得到清唱合成模型。

识别模块 1406, 用于将待合成的说话人录音输入至说话人识别模型, 获取说话人识别模型的中间隐藏层输出的说话人信息。 30

合成模块 1408, 用于将待合成的清唱音乐信息和说话人信息输入至清唱合成模型, 以获得合成的定制音色歌声。

下面参照图 15 来描述根据本公开的这种实施方式的电子设备 1500。图 15 显示的电子设备 1500 仅仅是一个示例, 不应对本公开实施例的功能和使用范围带来任何限制。

35 如图 15 所示, 电子设备 1500 以通用计算设备的形式表现。电子设备 1500 的组件可

以包括但不限于：上述至少一个处理单元 1510、上述至少一个存储单元 1520、连接不同系统组件（包括存储单元 1520 和处理单元 1510）的总线 1530。

其中，存储单元存储有程序代码，程序代码可以被处理单元 1510 执行，使得处理单元 1510 执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施方式的步骤。例如，处理单元 1510 可以执行本公开的定制音色歌声的合成方法中限定的步骤。

存储单元 1520 可以包括易失性存储单元形式的可读介质，例如随机存取存储单元（RAM）15201 和/或高速缓存存储单元 15202，还可以进一步包括只读存储单元（ROM）15203。

存储单元 1520 还可以包括具有一组（至少一个）程序模块 15205 的程序/实用工具 15204，这样的程序模块 15205 包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

总线 1530 可以为表示几类总线结构中的一种或多种，包括存储单元总线或者存储单元控制器、外围总线、图形加速端口、处理单元或者使用多种总线结构中的任意总线结构的局域总线。

电子设备 1500 也可以与一个或多个外部设备 1540（例如键盘、指向设备、蓝牙设备等）通信，还可与一个或者多个使得用户能与该电子设备交互的设备通信，和/或与使得该电子设备 1500 能与一个或多个其它计算设备进行通信的任何设备（例如路由器、调制解调器等等）通信。这种通信可以通过输入/输出（I/O）接口 1550 进行。并且，电子设备 1500 还可以通过网络适配器 1560 与一个或者多个网络（例如局域网（LAN），广域网（WAN）和/或公共网络，例如因特网）通信，网络适配器 1560 通过总线 1530 与电子设备 1500 的其它模块通信。应当明白，尽管图中未示出，可以结合电子设备使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RA 标识系统、磁带驱动器以及数据备份存储系统等。

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、终端装置、或者网络设备等）执行根据本公开实施方式的方法。

在本公开的示例性实施例中，还提供了一种计算机可读存储介质，其上存储有能够实现本说明书上述方法的程序产品。在一些可能的实施方式中，本公开的各个方面还可以实现为一种程序产品的形式，其包括程序代码，当程序产品在终端设备上运行时，程序代码用于使终端设备执行本说明书上述“示例性方法”部分中描述的根据本公开各种示例性实施方式的步骤。

根据本公开的实施方式的用于实现上述方法的程序产品，其可以采用便携式紧凑盘只

读存储器 (CD-ROM) 并包括程序代码, 并可以在终端设备, 例如个人电脑上运行。然而, 本公开的程序产品不限于此, 在本文件中, 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。

计算机可读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, 其中承载了可读程序代码。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。可读信号介质还可以是可读存储介质以外的任何可读介质, 该可读介质可以发送、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。

可读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、有线、光缆、RF 等等, 或者上述的任意合适的组合。

可以以一种或多种程序设计语言的任意组合来编写用于执行本公开操作的程序代码, 程序设计语言包括面向对象的程序设计语言—诸如 Java、C++ 等, 还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中, 远程计算设备可以通过任意种类的网络, 包括局域网 (LAN) 或广域网 (WAN), 连接到用户计算设备, 或者, 可以连接到外部计算设备 (例如利用因特网服务提供商来通过因特网连接)。

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或者单元, 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或者单元的特征和功能可以在一个模块或者单元中具体化。反之, 上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

此外, 尽管在附图中以特定顺序描述了本公开中方法的各个步骤, 但是, 这并非要求或者暗示必须按照该特定顺序来执行这些步骤, 或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的, 可以省略某些步骤, 将多个步骤合并为一个步骤执行, 以及/或者将一个步骤分解为多个步骤执行等。

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、移动终端、或者网络设备等) 执行根据本公开实施方式的方法。

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。本公开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所附的权利要求指出。

工业实用性

- 5 本公开提供的方案，通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型，第一神经网络输出训练结果为说话人向量样本，并通过歌声清唱样本和说话人向量样本对第二神经网络进行训练，以得到清唱合成模型，提高了模型合成的效率，不需要采集大量录音数据来生成语料库。进一步地，通过将待合成的说话人录音输入至说话人识别模型，获取说话人识别模型的中间隐藏层输出的说话人信息，以及将待合成的清唱音乐
- 10 信息和说话人信息输入至清唱合成模型，以获得合成的定制音色歌声，只需少量语料即可训练出用户定制音色，并通过调整合成语音的节奏和音高达到歌声合成的效果，减少了定制音色歌声合成过程中需要的时间和训练语料，提升了定制音色歌声的合成效率。

权利要求

1.一种定制音色歌声的合成方法，其特征在于，包括：

通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型，所述第一神经网络输出训练结果为说话人向量样本；

5 通过歌声清唱样本和所述说话人向量样本对第二神经网络进行训练，以得到清唱合成模型；

将待合成的说话人录音输入至所述说话人识别模型，获取所述说话人识别模型的中间隐藏层输出的说话人信息；

10 将待合成的清唱音乐信息和所述说话人信息输入至所述清唱合成模型，以获得合成的定制音色歌声。

2.根据权利要求1所述的定制音色歌声的合成方法，其特征在于，通过说话人录音样本对第一神经网络进行训练，以得到说话人识别模型包括：

将所述说话人录音样本划分为测试录音样本和注册录音样本，并输入至所述第一神经网络；

15 所述注册录音样本经所述第一神经网络输出注册录音特征，将所述注册录音特征进行平均池化处理，以得到注册录音向量；

所述测试录音样本经所述第一神经网络输出测试录音向量；

对所述注册录音向量和所述测试录音向量进行余弦相似度计算，以获得余弦相似度结果；

20 通过所述余弦相似度结果和回归函数对所述第一神经网络进行参数优化，至所述回归函数的损失值最小为止；

将参数优化后的所述第一神经网络确定为所述说话人识别模型。

3.根据权利要求1所述的定制音色歌声的合成方法，其特征在于，所述清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和所述说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：

25 解析所述歌声清唱样本中的乐谱样本、歌词样本和音素时长样本；

通过所述说话人向量样本、所述乐谱样本、所述歌词样本和所述音素时长样本对所述持续时间模型进行训练，所述持续时间模型的输出结果为时长预测样本。

4.根据权利要求1所述的定制音色歌声的合成方法，其特征在于，所述清唱合成模型包括持续时间模型、声学模型和声码器模型，通过歌声清唱样本和所述说话人向量样本对第二神经网络进行训练，以得到清唱合成模型包括：

30 解析所述歌声清唱样本中的乐谱样本、歌词样本和音素时长样本；

根据所述歌声清唱样本中的歌曲提取梅尔图谱样本；

35 通过所述说话人向量样本、所述音素时长样本、所述乐谱样本、所述歌词样本和所述梅尔图谱样本对所述声学模型进行训练，所述声学模型输出结果为梅尔谱图预测样本。

5.根据权利要求1所述的定制音色歌声的合成方法,其特征在于,所述清唱合成模型包括持续时间模型、声学模型和声码器模型,通过歌声清唱样本和所述说话人向量样本对第二神经网络进行训练,以得到清唱合成模型包括:

根据所述歌声清唱样本中的歌曲提取梅尔图谱样本;

5 通过所述梅尔谱图样本对所述声码器模型进行训练,所述声码器模型的输出结果为音频预测样本。

6.根据权利要求1-5中任一项所述的定制音色歌声的合成方法,其特征在于,所述清唱合成模型包括持续时间模型、声学模型和声码器模型,将待合成的清唱音乐信息和所述说话人信息输入至所述清唱合成模型,以获得合成的定制音色歌声包括:

10 解析所述清唱音乐信息中的待合成乐谱和待合成歌词;

将所述说话人信息、所述待合成乐谱和所述待合成歌词输入至所述持续时间模型,所述持续时间模型输出结果为待合成的时长预测结果;

将所述时长预测结果、所述说话人信息、所述待合成乐谱和所述待合成歌词输入至所述声学模型,所述声学模型输出结果为待合成的梅尔谱图预测结果;

15 将所述梅尔谱图预测结果输入至所述声码器模型,所述声码器模型输出结果为所述合成的定制音色歌声。

7.根据权利要求6所述的定制音色歌声的合成方法,其特征在于,解析所述清唱音乐信息中的待合成乐谱和待合成歌词包括:

20 对所述清唱音乐信息中的乐谱和歌词进行文本分析和特征提取,以获取所述待合成乐谱和所述待合成歌词。

8.根据权利要求6所述的定制音色歌声的合成方法,其特征在于,将所述时长预测结果、所述说话人信息、所述待合成乐谱和所述待合成歌词输入至所述声学模型,所述声学模型输出结果为待合成的梅尔谱图预测结果包括:

对所述时长预测结果、所述待合成乐谱和所述待合成歌词进行帧级扩展;

25 将所述帧级扩展的结果和所述说话人信息输入至所述声学模型,所述声学模型输出结果为待合成的梅尔谱图预测结果。

9.一种定制音色歌声的合成装置,其特征在于,包括:

第一训练模块,用于通过说话人录音样本对第一神经网络进行训练,以得到说话人识别模型,所述第一神经网络输出训练结果为说话人向量样本;

30 第二训练模块,用于通过歌声清唱样本和所述说话人向量样本对第二神经网络进行训练,以得到清唱合成模型;

识别模块,用于将待合成的说话人录音输入至所述说话人识别模型,获取所述说话人识别模型的中间隐藏层输出的说话人信息;

35 合成模块,用于将待合成的清唱音乐信息和所述说话人信息输入至所述清唱合成模型,以获得合成的定制音色歌声。

10.一种电子设备，其特征在于，包括：

处理器；以及

存储器，用于存储所述处理器的可执行指令；

其中，所述处理器配置为经由执行所述可执行指令来执行权利要求 1-8 中任一项所述的定制音色歌声的合成方法。

11.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，

所述计算机程序被处理器执行时实现权利要求 1-8 中任一项所述的定制音色歌声的合成方法。