

语音转换模型的训练方法、装置、设备及介质

本申请要求于 2022 年 11 月 21 日提交的申请号为 202211455842.7、发明名称为“语音转换模型的训练方法、装置、设备及介质”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

本申请实施例涉及音频处理技术领域，特别涉及一种语音转换模型的训练方法、装置、设备及介质。

背景技术

随着网络技术的不断发展，越来越多用户开始使用虚拟形象在网络中进行直播、游戏、社交或者在线会议。

为了保护个人隐私安全，用户在使用虚拟形象过程中，可以设置虚拟形象的口音，使原始口音的用户语音被转换为所设置口音后播放，并保证用户语音内容保持不变。相关技术中，口音转换通常使用语音转换模型实现，而在训练语音转换模型过程中需要大量的平行语料。其中，该平行语料为相同语音内容的不同口音音频。

然而，平行语料通常需要人工录制，导致平行语料的获取难度较高，在平行语料不足的情况下，训练得到语音转换模型的质量较差，进而影响口音转换效果。

发明内容

本申请实施例提供了一种语音转换模型的训练方法、装置、设备及介质，能够在降低对人工录制的平行语料的需求的前提下，保证语音转换模型的训练质量。所述技术方案如下：

一方面，本申请实施例提供了一种语音转换模型的训练方法，该方法由计算机设备执行，包括：

基于第一样本音频训练第一 ASR（Automatic Speech Recognition，自动语音识别）模型，以及基于第二样本音频训练第二 ASR 模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一 ASR 模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二 ASR 模型对所述第二样本音频进行提取得到，所述平行样本数据包括不同内容特征，不同内容特征对应不同口音，且不同内容特征对应相同文本；

基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

基于训练得到的所述第一 ASR 模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

一方面，本申请实施例提供了一种语音转换方法，所述方法由计算机设备执行，所述计算机设备中设置有语音转换模型，所述语音转换模型包括第一 ASR 模型、第二转换模型和第三转换模型，所述方法包括：

获取第一口音音频，所述第一口音音频对应第一口音；

通过所述第一 ASR 模型对所述第一口音音频提取得到第一内容特征，所述第一内容特征对应所述第一口音；

通过所述第二转换模型将所述第一内容特征转换为第二内容特征，所述第二内容特征对应第二口音；

通过所述第三转换模型对所述第二内容特征进行音频转换，得到第二口音音频。

另一方面，本申请实施例提供了一种语音转换模型的训练装置，所述装置包括：

训练模块，用于基于第一样本音频训练第一 ASR 模型，以及基于第二样本音频训练第二 ASR 模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

所述训练模块，还用于基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一 ASR 模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

所述训练模块，还用于基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二 ASR 模型对所述第二样本音频进行提取得到，所述平行样本数据包括不同内容特征，不同内容特征对应不同口音，且不同内容特征对应相同文本；基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

所述训练模块，还用于基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

生成模块，用于基于训练得到的所述第一 ASR 模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

另一方面，本申请实施例提供了一种语音转换装置，其中，所述装置包括：

获取模块，用于获取第一口音音频，所述第一口音音频对应第一口音；

提取模块，用于通过所述第一 ASR 模型对所述第一口音音频提取得到第一内容特征，所述第一内容特征对应所述第一口音；

内容特征转换模块，用于通过所述第二转换模型将所述第一内容特征转换为第二内容特征，所述第二内容特征对应第二口音；

音频转换模块，用于通过所述第三转换模型对所述第二内容特征进行音频转换，得到第二口音音频。

另一方面，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，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行以实现如上述方面所述的语音转换模型的训练方法，或，如上述方面所述的语音转换方法。

另一方面，本申请实施例提供了一种计算机可读存储介质，所述可读存储介质中存储有至少一条指令，所述至少一条指令由处理器加载并执行以实现如上述方面所述的语音转换模型的训练方法，或，如上述方面所述的语音转换方法。

另一方面，本申请实施例提供了一种计算机程序产品，所述计算机程序产品包括计算机指令，所述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所述计算机可读存储介质读取所述计算机指令，所述处理器执行所述计算机指令，使得所述计算机设备执行如上述方面所述的语音转换模型的训练方法，或，如上述方面所述的语音转换方法。

本申请实施例中，在缺少第二口音的第二样本音频对应平行语料的情况下，首先基于第一口音的第一样本音频，训练用于将文本转换为内容特征的第一转换模型，从而利用该第一转换模型以及第二样本音频对应的第二样本文本，构建得到包含对应相同文本内容但对应不同口音的平行样本数据，进而利用该平行样本数据训练在不同口音间进行内容特征转换的第二转换模型，以及用于将内容特征转换为音频的第三转换模型，完成语音转换模型训练；模型训练过程中，利用训练得到的中间模型构建平行语料，无需在模型训练前录制不同口音的

平行语料，能够在保证模型训练质量的情况下，降低模型训练对人工录制的平行语料的需求，有助于提高模型训练效率，并提高样本不足情况下模型的训练质量。

附图说明

图1示出了本申请一个示例性实施例提供的语音转换系统的示意图；
图2示出了本申请一个示例性实施例提供的语音转换模型的训练方法的流程图；
图3示出了本申请一个示例性实施例提供的口音转换方法的流程图；
图4是本申请一个示例性实施例示出的语音设置界面的示意图；
图5是本申请一个示例性实施例提供的口音转换过程的实施示意图；
图6是本申请一个示例性实施例示出的文本转内容特征过程的流程图；
图7是本申请一个示例性实施例提供的FFT结构图；
图8是本申请一个示例性实施例示出的第一转换模型的结构示意图；
图9是本申请一个示例性实施例示出的第二转换模型训练过程的流程图；
图10是本申请一个示例性实施例示出的第二转换模型的结构示意图；
图11是本申请一个示例性实施例示出的第三转换模型的结构示意图；
图12是本申请一个示例性实施例示出的第三转换模型训练过程的流程图；
图13是本申请另一个示例性实施例提供的口音转换过程的实施示意图；
图14是本申请一个示例性实施例提供的语音转换模型的训练装置的结构框图；
图15是本申请一个示例性实施例提供的语音转换装置的结构框图；
图16示出了本申请一个示例性实施例提供的计算机设备的结构示意图。

具体实施方式

为了降低模型训练过程对预先录制的平行语料的依赖，本申请实施例中，语音转换模型由第一ASR模型（用于将音频转换为文本）、第二转换模型（用于在不同口音间进行内容特征转换）以及第三转换模型（用于将内容特征转换为音频）构成。并且，在训练过程中，完成第一ASR模型训练后，训练用于将文本转换为内容特征的第一转换模型，从而借助第一转换模型构建平行样本数据，以用于后续第二转换模型以及第三转换模型训练。训练过程中，借助训练得到的转换模型构建平行语料，无需预先人工录制大量平行语料，从而降低训练过程对平行语料的依赖，保证模型训练质量。

需要说明的是，本申请所涉及的信息（包括但不限于用户设备信息、用户个人信息等）、数据（包括但不限于用于分析的数据、存储的数据、展示的数据等）以及信号，均为经用户授权或者经过各方充分授权的，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。例如，本申请中涉及到的音频、口音、文本都是在充分授权的情况下获取的。

采用本申请实施例提供的训练方法所训练得到的语音转换模型，能够适用于各种需要进行口音转换的场景。如图1所示，其示出了本申请一个示例性实施例示出的语音转换系统的示意图。该语音转换系统中包括：音频采集设备110、终端120以及服务器130。

音频采集设备110是用于采集用户语音的设备，该音频采集设备110可以是耳麦、麦克风或者具有收音功能的AR/VR设备等等，本申请实施例对此不作限定。

音频采集设备110与终端120之间通过有线或无线方式相连，用于将采集到的用户语音传输至终端120，由终端120进一步对用户语音进行口音转换处理。其中，终端120可以是具有智能手机、平板电脑、个人计算机、车载终端等电子设备。

在一些实施例中，终端120中设置有具有口音转换功能的应用程序。通过该应用程序，用户可以设置口音转换目标，从而实现将用户语音由原始语音转换为目标语音。

在一种可能的实施方式中，口音转换可以由终端 120 在本地实现（语音转换模型设置在终端 120 中）；在另一种可能的实施方式中，口音转换可以由终端 120 借助服务器 130 实现（语音转换模型设置在服务器 130 中，终端 120 向服务器 130 传输口音转换需求）。

服务器 130 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，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络（Content Delivery Network, CDN）、以及大数据和人工智能平台等基础云计算服务的云服务器。本申请实施例中，服务器 130 可以为实现口音转换功能的后台服务器，用于提供不同口音间的转换服务。

在一些实施例中，服务器 130 中设置有多个语音转换模型，不同语音转换模型用于实现不同口音间的转换。比如，当支持将普通话转换为 n 种地方口音时，服务器 130 中设置有 n 个语音转换模型。

并且，在实现口音转换功能前，服务器 130 获取不同口音的口音语料，该口音语料由音频以及对应的文本构成，从而基于口音语料训练相应的语音转换模型。

如图 1 所示，在进行口音转换前，用户通过终端 120 设置将第一口音转换为第二口音，由终端 120 向服务器 130 发送口音转换请求，请求服务器 130 采用相应的语音转换模型（将第一口音转换为第二口音）进行口音转换。

音频采集设备 110 将采集到的第一口音的用户语音传输至终端 120，由终端 120 将第一口音的用户语音传输至服务器 130。服务器 130 通过语音转换模型将其转换为第二口音的用户语音，并反馈至终端 120，由终端 120 进行进一步处理。

其中，不同应用场景下，终端 120 对用户语音的处理方式不同。下面采用集中示例性的应用场景进行说明。

1、虚拟人内容制作场景

虚拟人内容制作场景下，终端获取转换得到的用户语音后，将该用户语音与制作的内容（比如虚拟人短视频、虚拟人长视频等等）进行融合，得到虚拟人内容。其中，在进行融合时，可以根据转换得到的用户语音对虚拟人的嘴部进行控制，提高虚拟人嘴部动作与语音的匹配度。

示例性地，在虚拟人内容制作场景下，以制作虚拟人的短视频为例，终端获取真实用户对应的第一口音音频，第一口音音频对应真实用户的第一口音，终端通过语音转换模型中的第一 ASR 模型对第一口音音频提取得到第一口音下的第一内容特征；终端通过语音转换模型中的第二转换模型将第一内容特征转换为第二内容特征，第二内容特征对应第二口音；在转换完口音后，终端通过语音转换模型中的第三转换模型对第二口音下的第二内容特征进行音频转换，得到虚拟人对应的第二口音音频。

2、虚拟主播直播场景

虚拟主播直播场景下，虚拟主播可以通过口音设置界面预先设置直播口音。直播过程中，终端将通过麦克风采集的用户语音发送至服务器，由服务器将原始口音的用户语音转换为直播口音的用户语音，并反馈至终端。终端将直播口音的用户语音与包含虚拟主播形象的视频流进行合并，从而通过推流服务器将合并得到的音视频流推送至直播间内的各个观众客户端。

示例性地，在虚拟主播直播场景下，可预先设置虚拟主播直播时的口音。终端通过麦克风采集真实用户对应的第一口音音频，第一口音音频对应真实用户的第一口音，终端通过语音转换模型中的第一 ASR 模型对第一口音音频提取得到第一口音下的第一内容特征；终端通过语音转换模型中的第二转换模型将第一内容特征转换为第二内容特征，第二内容特征对应第二口音；在转换完口音后，终端通过语音转换模型中的第三转换模型对第二口音下的第二内容特征进行音频转换，得到虚拟主播对应的第二口音音频，即，虚拟主播以第二口音音频进行直播。

3、元宇宙场景

元宇宙场景下，用户可以设置在元宇宙中进行交互时采用的口音。用户控制虚拟角色在元宇宙中与其他虚拟角色进行交互时，用户语音由耳麦、AR/VR 等设备采集并传输至终端，由终端进一步交由服务器进行口音转换，并在元宇宙中控制虚拟角色播放转换得到的口音音频，实现与其他虚拟角色之间的语音互动。

示例性地，在元宇宙场景下，可预先选择与其他虚拟角色进行交互时的第二口音。在交互时，终端通过麦克风采集真实用户对的第一口音音频，第一口音音频对应真实用户的第一口音，终端通过语音转换模型中的第一 ASR 模型对第一口音音频提取得到第一口音下的第一内容特征；终端通过语音转换模型中的第二转换模型将第一内容特征转换为第二内容特征，第二内容特征对应第二口音；在转换完口音后，终端通过语音转换模型中的第三转换模型对第二口音下的第二内容特征进行音频转换，得到元宇宙中的虚拟角色对应的第二口音音频，即，元宇宙中的虚拟角色以第二口音音频与其他虚拟角色进行交互。

上述应用场景仅作为示例性的说明，采用本申请实施例提供的方法训练得到的语音转换模型，还可以用于语音通话（方便不同口音的通话对象之间进行语音交流）、翻译等真实世界应用场景，本申请实施例并不对此构成限定。

并且，为了方便表述，下述各个实施例中，以语音转换模型的训练以及使用均用于计算设备（可以为终端或者服务器），且训练用于将第一口音转换为第二口音的语音转换模型为例进行说明（其他由源语音转换为目标语音的方案类似），但并不对此构成限定。

图 2 示出了本申请一个示例性实施例提供的语音转换模型的训练方法的流程图。该方法由计算机设备执行，该方法包括如下步骤。

步骤 201，基于第一样本音频训练第一 ASR 模型，以及基于第二样本音频训练第二 ASR 模型，第一样本音频对应第一口音，第二样本音频对应第二口音。

其中，第一口音为源口音，第二口音为目标语音，即训练得到的语音转换模型用于将第一口音的语音转换为第二口音的语音。

在一些实施例中，第一样本音频对应有第一样本文本，第二样本音频对应有第二样本文本。本申请实施例中，第一样本文本无需与第二样本文本相同，因此可以直接采用公开语音数据集用于模型训练。

在一个示意性的例子中，计算机设备采用 Wenet Speech 数据集作为第一样本音频，采用 KeSpeech 数据集作为第二样本音频，其中，Wenet Speech 数据集包括 1 万小时的 ASR 数据，可参见网址：<https://zhuanlan.zhihu.com/p/424118791> 中的介绍；KeSpeech 数据集包含不同地区方言的 ASR 数据，可参见网址：<https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/0336dcbab05b9d5ad24f4333c7658a0e-Abstract-round2.html> 中的介绍。

关于 ASR 模型的训练方式，在一种可能的实施方式中，计算机设备将样本音频输入 ASR 模型，得到 ASR 模型输出的预测文本，从而基于预测文本以及样本音频对应的样本文本对 ASR 模型进行训练。

可选的，ASR 模型的模型架构包括但不限于 Wenet、wav2vec2、Kaldi 等等，本申请实施例对此不作限定。Wenet 是出门问问语音团队联合西工大语音实验室开源的一款面向工业落地应用的语音识别工具包，该工具用一套简洁的方案提供了语音识别从训练到部署的一条龙服务，可参见网址：<https://zhuanlan.zhihu.com/p/349586567> 中的介绍。wav2vec 是 Interspeech 2019 收录的文章中提出的，作者用无监督预训练的卷积神经网络来改善语音识别任务，提出了一种噪声对比学习二分类任务，从而使得 wav2vec 可以在大量未标注的数据上进行训练，可参见网址：<https://zhuanlan.zhihu.com/p/302463174> 中的介绍。Kaldi 是一种开源语音识别工具，它使用 WFST 来实现解码算法，Kaldi 的主要代码是 C++ 编写，在此之上使用 bash 和 python 脚本做了一些工具，可参见网址：<https://zhuanlan.zhihu.com/p/84050431> 中的介

绍。

在一些实施例中，ASR模型可以基于样本音频重新训练得到（适用于样本音频数量较多的情况），也可以基于样本音频对预训练的ASR模型进行微调得到（适用于样本音频数量较少的情况）。

比如，当第一口音为普通话，第二口音为方言时，第一ASR模型基于第一样本音频重新训练得到，第二ASR模型在第一ASR模型的基础上，基于第二样本音频进行微调得到。

本申请实施例中，训练得到的ASR模型用于提取语音中的内容特征。在一些实施例中，该内容特征被称为BN（BottleNeck，瓶颈）特征，通常为ASR模型的最后一层特征，其保留了语音的内容特征，并剔除了诸如音色、音调等其他特征。

在一些实施例中，第一ASR模型的训练过程包括：计算机设备将第一样本音频输入第一ASR模型进行文本提取，得到第一预测文本；计算机设备计算得到第一预测文本与第一样本音频对应的第一样本文本之间的损失函数值；计算机设备基于第一预测文本与第一样本文本之间的损失函数值对第一ASR模型的模型参数进行更新，从而实现第一ASR模型的训练。

在一些实施例中，第二ASR模型的训练过程包括：计算机设备将第二样本音频输入第二ASR模型进行文本提取，得到第二预测文本；计算机设备计算得到第二预测文本与第二样本音频对应的第二样本文本之间的损失函数值；计算机设备基于第二预测文本与第二样本文本之间的损失函数值对第二ASR模型的模型参数进行更新，从而实现第二ASR模型的训练。

步骤202，基于第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，第一样本内容特征由第一ASR模型对第一样本音频进行提取得到，第一转换模型用于将文本转换为第一口音的内容特征。

由于采用不同口音表述相同文本内容时的发音并不相同，因此对相同文本对应的不同口音语音进行内容特征提取所得到的内容特征也不同，相应的，在不同口音间实现内容特征转换成为实现口音转换的关键。

本申请实施例中采用了数据增强方案实现非平行语料（即对应不同口音且对应不同文本的语料）之间的内容特征转换。

在一些实施例中，计算机设备通过训练得到的第一ASR模型对第一样本音频进行特征提取，得到第一样本音频的第一样本内容特征，从而基于第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型。其中，第一转换模型可以被称为文本内容特征转换模型（Text2BN模型），用于实现文本至源口音内容特征之间的转换。

在一些实施例中，第一转换模型的训练过程包括：计算机设备将第一样本文本输入第一转换模型，得到第一预测内容特征；计算机设备通过第一ASR模型对第一样本音频进行提取得到第一样本内容特征；计算机设备计算得到第一预测内容特征与第一样本内容特征之间的损失函数值；计算机设备基于第一预测内容特征与第一样本内容特征之间的损失函数值对第一转换模型的模型参数进行更新，从而实现第一转换模型的训练。

步骤203，基于第一转换模型、第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，第二样本内容特征由第二ASR模型对第二样本音频进行提取得到，平行样本数据包括不同内容特征，不同内容特征对应不同口音，且不同内容特征对应相同文本。

在一些实施例中，计算机设备基于第一转换模型对第二样本音频对应的第二样本文本进行文本转化，得到第二样本文本对应的第一口音的内容特征；计算机设备将第二样本文本对应的第一口音的内容特征和第二样本文本对应的第二口音的内容特征进行汇总，得到平行样本数据。

可选地，第二样本文本对应的第二口音的内容特征的由第二ASR模型提取得到的。

训练得到第一转换模型后，计算机设备基于第二样本音频对应的第二样本文本以及第一转换模型进行数据增强，从而根据第二样本内容特征以及数据增强得到的第一口音的内容特征，构建得到平行样本数据。其中，平行样本数据包括相同文本对应的第一口音的内容特征（由第一转换模型生成）以及第二口音的内容特征（由第二 ASR 模型提取得到）。

比如，在包含对应文本 A 的方言样本音频，而不包含对应文本 A 的普通话样本音频的情况下，计算机设备可以基于第一转换模型、文本 A 以及对应文本 A 的方言样本音频的方言样本内容特征，构建得到文本 A 对应的平行样本数据，该平行样本数据包含文本 A 对应的普通话以及方言的内容特征。

步骤 204，基于平行样本数据训练第二转换模型，第二转换模型用于对第一口音和第二口音间进行内容特征转换。

进一步的，计算机设备基于对应相同文本的平行样本数据，训练第二转换模型。其中，该第二转换模型可以被称为内容特征转换模型（BN2BN 模型），用于将源口音的内容特征转换为目标口音的内容特征。BN2BN 模型用于实现口音迁移任务，可参见网址：<https://zhuanlan.zhihu.com/p/586037409> 中的介绍。

比如，当第一口音为普通话，第二口音为方言时，计算机设备训练得到第二转换模型用于将普通话的内容特征转换为方言的内容特征。

需要说明的是，当第一样本音频与第二样本音频对应相同样本文本时，第一样本音频以及第二样本音频对应的样本内容特征可以直接被用于训练第二转换模型。

在一些实施例中，第二转换模型的训练过程包括：计算机设备通过第二 ASR 模型对第二样本音频进行提取得到第二样本内容特征；计算机设备通过第一转换模型对第二样本音频对应的第二样本文本进行转换得到第三样本内容特征，第三样本内容特征指采用第一口音表述第二样本文本所产生音频的内容特征；计算机设备将第三样本内容特征输入第二转换模型，得到第二预测内容特征；计算机设备计算得到第二预测内容特征与第二样本内容特征之间的损失函数值；计算机设备基于第二预测内容特征与第二样本内容特征之间的损失函数值对第二转换模型的模型参数进行更新，从而实现第二转换模型的训练。

步骤 205，基于不同样本音频的样本内容特征训练第三转换模型，第三转换模型用于将内容特征转换为音频。

其中，第三转换模型可以被称为内容音频转换模型，用于基于目标语音的内容特征转换得到目标语音的音频。

在一些实施例中，该第三转换模型可以包括声学模型以及声码器，其中，声学模型用于基于内容特征生成音频频谱，而声码器则用于基于音频频谱生成音频。

在一些实施例中，训练第三转换模型的样本可以为各种口音的样本音频。

需要说明的是，第三转换模型可以在 ASR 模型训练完成后执行，即第三转换模型可以与第一以及第二转换模型同步训练。本申请实施例并不对模型的训练时序构成限定。

在一些实施例中，第三转换模型的训练过程包括：计算机设备将样本内容特征以及样本音频对应的说话者标识输入第三转换模型进行音频的生成，得到预测音频；计算机设备计算得到预测音频与样本音频之间的损失函数值；计算机设备基于预测音频与样本音频之间的损失函数值对第三转换模型的模型参数进行更新，从而实现第三转换模型的训练。

步骤 206，基于训练得到的第一 ASR 模型、第二转换模型和第三转换模型生成语音转换模型，语音转换模型用于将第一口音的音频转换为第二口音的音频。

通过上述步骤训练得到第一 ASR 模型、第二转换模型和第三转换模型后，计算机设备将上述模型组合得到最终的语音转换模型。其中，模型之间的拼接顺序为第一 ASR 模型→第二转换模型→第三转换模型，即第一 ASR 模型的输出被输入第二转换模型，第二转换模型的输出被输入第三转换模型。

在一个示意性的例子中，训练得到的用于将普通话转换为方言的语音转换模型由普通话

ASR 模型, 普通话-方言内容转换模型以及内容音频转换模型构成。

综上所述, 本申请实施例中, 在缺少第二口音的第二样本音频对应平行语料的情况下, 首先基于第一口音的第一样本音频, 训练用于将文本转换为内容特征的第一转换模型, 从而利用该第一转换模型以及第二样本音频对应的第二样本文本, 构建得到包含对应相同文本内容但对应不同口音的平行样本数据, 进而利用该平行样本数据训练在不同口音间进行内容特征转换的第二转换模型, 以及用于将内容特征转换为音频的第三转换模型, 完成语音转换模型训练; 模型训练过程中, 利用训练得到的中间模型构建平行语料, 无需在模型训练前录制不同口音的平行语料, 能够在保证模型训练质量的情况下, 降低模型训练对人工录制的平行语料的需求, 有助于提高模型训练效率, 并提高样本不足情况下模型的训练质量。

下面对采用上述方案训练得到的语音转换模型的应用过程进行说明。通过利用该语音转换模型可实现语音转换方法, 该语音转换方法由计算机设备执行, 该语音转换模型包括第一 ASR 模型、第二转换模型和第三转换模型。在进行语音转换时, 计算机设备获取第一口音音频, 第一口音音频对应第一口音; 计算机设备通过第一 ASR 模型对第一口音音频提取得到第一内容特征, 第一内容特征对应第一口音; 计算机设备通过第二转换模型将第一内容特征转换为第二内容特征, 第二内容特征对应第二口音; 计算机设备通过第三转换模型对第二内容特征进行音频转换, 得到第二口音音频, 从而完成语音转换。

示例性地, 计算机设备接收到第一口音的第一口音音频后, 通过语音转换模型中的第一 ASR 模型进行内容特征提取, 得到第一内容特征。

在一些实施例中, 计算机设备将第一 ASR 模型提取到的第一内容特征输入第二转换模型, 由第二转换模型在第一口音和第二口音之间进行内容特征转换, 得到第二口音下的第二内容特征。

其中, 第一内容特征和第二内容特征对应相同文本 (均为第一口音音频对应的文本)。

可选地, 第二转换模型包括卷积层以及 N 层堆叠的 FFT; 计算机设备通过第二转换模型中的卷积层对第一内容特征进行卷积处理后, 将卷积结果输入至 N 层堆叠的 FFT 进行转换得到第二内容特征。

在一些实施例中, 计算机设备将第二内容特征以及目标音色对应说话者的说话者标识输入第三转换模型, 得到第二口音音频。

其中, 不同说话者对应不同说话者标识。

可选地, 第三转换模型包括第三转换子模型以及声码器, 其中, 第三转换子模型用于将内容特征转换为音频谱特征, 声码器则用于基于音频谱特征生成音频。

可选的, 该第三转换子模型包括卷积层以及 N 层堆叠的 FFT, 该音频谱特征可以为 Mel (梅尔) 谱特征、MFCC (Mel Frequency Cepstrum Coefficient, 梅尔频率倒谱系数) 特征等等, 本申请实施例对此不作限定。

可选的, 该声码器可以是采用自回归的 Wavenet 或 WaveRNN, 或者采用非自回归的 Hifigan 或 Melgan 等等, 本申请实施例对此不作限定。

为了方便表述, 下述实施例中以音频谱特征为 Mel 谱特征, 声码器为 hifigan 为例进行说明, 但并不对此构成限定。

在一些实施例中, 计算机设备将第二内容特征以及说话者标识输入第三转换子模型, 得到音频谱特征; 计算机设备将音频谱特征输入声码器, 得到第二口音音频。

下面对采用上述方案训练得到的语音转换模型的应用过程进行说明。图 3 示出了本申请一个示例性实施例提供的口音转换方法的流程图。该方法由计算机设备执行, 该方法包括如下步骤。

步骤 301, 响应于口音转换指令, 通过第一 ASR 模型提取第一口音音频的第一内容特征, 第一内容特征对应第一口音, 口音转换指令用于指示将音频由第一口音转换为第二口

音。

在一些实施例中，该口音转换指令在完成口音设置后触发。在一种可能的场景中，如图4所示，在元宇宙虚拟角色设置界面41中，除了包含虚拟角色形象设置选项外，还包括语音设置选项。用户可以通过语音设置选项设置虚拟角色的音色以及口音。当完成语音以及形象设置后，通过触发进入按键42即可进入元宇宙。其中，触发进入按键42后，计算机设备接收到口音转换指令，该口音转换指令中包含源口音以及目标口音的口音标识。本实施例中，以源口音为第一口音，目标口音为第二口音为例进行说明。

计算机设备接收到第一口音的第一口音音频后，通过语音转换模型中的第一ASR模型进行内容特征提取，得到第一内容特征，该第一内容特征提出了音色、音调等干扰，仅保留所表达内容层面的特征。

在一些实施例中，计算机设备将第一ASR模型最后一层BN特征作为第一内容特征。

示意性的，如图5所示，当需要将普通话转换为方言时，计算机设备通过普通话ASR模型52对普通话音频51进行特征提取，得到普通话内容特征53。

步骤302，通过第二转换模型将第一内容特征转换为第二内容特征，第二内容特征对应第二口音。

进一步的，计算机设备将第一ASR模型提取到的第一内容特征输入第二转换模型，由第二转换模型在第一口音和第二口音之间进行内容特征转换，得到第二口音下的第二内容特征。其中，第一内容特征和第二内容特征对应相同文本（均为第一口音音频对应的文本）。

示意性的，如图5所示，BN2BN模型54用于在普通话和方之间进行内容特征转换。得到普通话内容特征53后，计算机设备进一步通过BN2BN模型54对普通话内容特征53进行特征转换，得到方言内容特征55。

步骤303，通过第三转换模型对第二内容特征进行音频转换，得到第二口音音频。

进一步的，计算机设备将第二内容特征输入第三转换模型，由第三转换模型基于内容特征生成第二口音音频。

示意性的，如图5所示，计算机设备将方言内容特征55输入BN2Wav模型56，得到BN2Wav模型56输出的方言音频57。

第一转换模型作为构建平行样本数据的关键模型，在训练第一转换模型的过程中，计算机设备将第一样本文本输入第一转换模型，得到第一转换模型输出的第一预测内容特征，从而以第一样本内容特征为第一预测内容特征的监督，训练第一转换模型。

在一些实施例中，计算机设备以第一样本内容特征为第一预测内容特征的监督，基于第一预测内容特征与第一样本内容特征之间的特征差异确定第一转换模型损失，从而基于该第一转换模型损失训练第一转换模型。其中，该损失可以为MSE（Mean Square Error，均方误差）损失或其他类型损失，本实施例对此不作限定。

均方误差是指第一预测内容特征与第一样本内容特征之间的特征差异值的平方和的平均数，也就是误差平方和的平均数。

可选的，第一转换模型的损失 $loss_{Text2BN}$ 可以表示为：

$$loss_{Text2BN} = \|BN_{na} - \widehat{BN}_{na}\|$$

其中， BN_{na} 为第一ASR模型提取到的第一样本内容特征， $\widehat{BN}_{na}$ 为第一转换模型输出的第一预测内容特征。

为了提升文本到内容特征的转换质量，在一种可能的设计中，该第一转换模型包括第一转换子模型、时长预测子模型以及第二转换子模型，其中，第一转换子模型用于实现文本到文本编码特征之间的转换，时长预测子模型用于预测文本的表述时长，而第二转换子模型则用于将文本编码特征转换为内容特征。

相应的，第一转换模型将文本转换为内容特征的过程如图6所示。

步骤 601, 通过第一转换子模型对第一样本文本进行编码, 得到第一文本编码特征。

由于文本表述具有前后关联性, 因此本申请实施例中, 为了提高后续特征转换质量, 在一种可能的设计中, 采用 N 层堆叠的 FFT (Feed Forward Transformer, 前馈转换) 构成第一转换子模型。其中, 该 FFT 用于通过线性变换, 先将数据映射到高纬度的空间再映射到低纬度的空间, 以此提取更深层次的特征。

并且, FFT 包括多头注意力机制层和卷积层。在一个示意性的例子中, 该 FFT 结构如图 7 所示。原始输入首先经过多头注意力层 701 处理, 多头注意力层 701 处理得到的多路结果和原始输入共同经过加权和标准化 702 处理后, 输入卷积层 703 进行卷积处理。卷积层 703 的输入与输出相加后继续进行加权和标准化 702 处理最终输出。

由于 FFT 通过多头注意力机制和卷积层实现, 且使用了残差网络思想, 因此利用多层 FFT 叠加得到的第一转换子模型进行文本编码, 能够提高文本编码质量。

当然, 第一转换子模型除了可以采用堆叠的 FFT 外, 还可以采用 LSTM (Long Short-Term Memory, 长短期记忆) 等其他类型的模块 (需要包含注意力机制, 且保持输入与输出的尺寸一致) 实现, 本申请实施例对此不作限定。

步骤 602, 通过时长预测子模型对第一文本编码特征进行时长预测, 得到预测时长, 预测时长用于表征第一样本文本的发音时长。

由于采用口语表述文本时, 具有一定的表述时长, 因此为了提高后续转换得到的音频的真实性 (使转换的得到的语音符合真人语速), 计算机设备通过时长预测子模型进行时长预测, 得到第一样本文本的发音时长。

在一些实施例中, 该预测时长包括第一样本文本中各个子文本对应的发音子时长。比如, 第一样本文本为“今天天气真好”, 该预测时长包括“今”、“天”、“天”、“气”、“真”、“好”各自对应的发音时长。

步骤 603, 基于预测时长对第一文本编码特征进行特征扩充, 得到第二文本编码特征。

进一步的, 计算机设备基于预测时长对第一文本编码特征进行特征扩充, 复制第一文本编码特征中的子特征, 使复制后子特征对应的持续时长与对应子文本的发音子时长保持一致。

在一个示意性的例子中, 第一文本编码特征为“abcd”, 经过特征扩充后的第二文本编码特征为“aabbcdcccc”。

步骤 604, 通过第二转换子模型将第二文本编码特征转换得到第一预测内容特征。

在一些实施例中, 第二转换子模型输出的第一预测内容特征与输出的第二文本编码特征的特征尺寸保持一致。

在一些实施例中, 第二转换子模型包括 N 层 FFT, 以此提升文本编码特征到内容特征的转换质量。

在一个示意性的例子中, 如图 8 所示, 第一转换子模型 81 首先对第一样本文本进行特征编码, 得到第一文本编码特征, 并将第一文本编码特征输入时长预测子模型 82, 得到预测时长, 并基于预测时长对第一文本编码特征进行特征扩充处理, 得到第二文本编码特征。最终通过第二转换子模型 83 对第二文本编码特征进行特征转换, 得到第一预测内容特征。

针对上述平行样本数据的构建, 以及基于平行样本数据训练第二转换模型的过程, 在一种可能的实施方式中, 如图 9 所示, 该过程可以包括如下步骤。

步骤 901, 通过第一转换模型将第二样本文本转换得到第三样本内容特征, 第三样本内容特征指采用第一口音表述第二样本文本所产生音频的内容特征。

在基于第二样本音频构建平行样本数据时, 计算机设备对第二样本音频对应的第二样本文本进行内容特征转换, 得到第三样本内容特征。由于第一转换模型用于将文本转换为第一口音的内容特征, 因此利用第一转换模型对第二样本文本进行内容特征转换, 得到第三样本

内容特征即为采用第一口音表述第二样本文本所产生音频的内容特征。

借助第一转换模型，即便缺少第二样本音频对应的平行语料，也能够生成平行语料的内容特征，免去了人工录制平行语料，以及对平行语料进行内容特征提取的流程。

步骤 902，基于第二样本内容特征和第三样本内容特征构建平行样本数据。

由于第三样本内容特征与第二样本内容特征对应不同口音，且对应相同文本，因此两者组合集合构建得到平行样本数据。

步骤 903，将第三样本内容特征输入第二转换模型，得到第二预测内容特征。

在一种可能的设计中，为了提高内容特征转换质量，第二转换模型包括卷积层以及 N 层堆叠的 FFT，其中，FFT 的具体结构可以参考图 7，本实施例在此不做赘述。进行内容特征转换时，内容特征首先经过卷积层卷积处理，然后经过 N 层 FFT 处理，得到转换后的内容特征。

示意性的，如图 10 所示，计算机设备通过卷积层 1001 对第三样本内容特征进行卷积处理后，将卷积结果输入 N 层 FFT 1002，得到第二预测内容特征。

步骤 904，以第二样本内容特征为第二预测内容特征的监督，训练第二转换模型。

为了使第二转换模型的内容转换结果接近第二 ASR 模型输出的第二样本音频的第二样本内容特征，在一种可能的实施方式中，计算机设备基于第二样本内容特征与第二预测内容特征的差异确定第二转换模型损失，从而基于该第二转换模型损失训练第二转换模型。

其中，该损失可以为 MSE 损失或其他类型损失，本实施例对此不作限定。

可选的，第二转换模型的损失 $loss_{BN2BN}$ 可以表示为：

$$loss_{BN2BN} = \|BN_{ac} - \widehat{BN}_{ac}\|$$

其中， BN_{ac} 为第二 ASR 模型提取到的第二样本内容特征， $\widehat{BN}_{ac}$ 为第二转换模型输出的第二预测内容特征。

由于内容特征剔除了音色等因素的影响，而样本音频具有音色特征，因此在训练第三转换模型过程中，需要将样本音频的说话者标识作为输入的一部分，使训练得到的第三转换模型能够输出具有特定音色的音频。

在一种可能的实施方式中，计算机设备将样本内容特征以及样本音频对应的说话者标识输入第三转换模型，得到预测音频，从而基于预测音频以及样本音频，训练第三转换模型。其中，预测音频与样本音频对应相同音频内容，且具有相同音色。

可选的，不同说话者对应不同说话者标识。在一些实施例中，预先将说话者划分为不同音色，从而为同一音色对应的不同说话者分配相同说话者标识。

在一种可能的设计中，第三转换模型包括第三转换子模型以及声码器，其中，第三转换子模型用于将内容特征转换为音频谱特征，声码器则用于基于音频谱特征生成音频。

可选的，该第三转换模型包括卷积层以及 N 层堆叠的 FFT，该音频谱特征可以为 Mel（梅尔）谱特征、MFCC（Mel Frequency Cepstrum Coefficient，梅尔频率倒谱系数）特征等等，本申请实施例对此不作限定。

可选的，该声码器可以是采用自回归的 Wavenet 或 WaveRNN，或者采用非自回归的 hifigan 或 melgan 等等，本申请实施例对此不作限定。

为了方便表述，下述实施例中以音频谱特征为 Mel 谱特征，声码器为 hifigan 为例进行说明，但并不对此构成限定。

相应的，训练过程中，计算机设备将样本内容特征以及说话者标识输入第三转换子模型，得到预测音频谱特征，从而将预测音频谱特征输入声码器，得到预测音频。

示意性的，如图 11 所示，BN2Wav 模型包括 BN2Mel 子模型 1101 以及 hifigan 子模型 1102，其中，BN2Mel 子模型 1101 包括卷积层 11011 以及 N 层堆叠的 FFT 11012。模型训练过程中，计算机设备将样本音频的样本内容特征 BN 以及说话者标识 spk_id 输入 BN2Mel 子

模型 1101。BN2Mel 子模型 1101 将转换得到的 Mel 频谱输入 hifigan 子模型 1102，由 hifigan 子模型 1102 转换得到预测音频。

在一种可能的实施方式中，计算机设备联合训练第三转换子模型和声码器。

在另一种可能的实施方式中，计算机设备首先训练第三转换子模型，然后基于训练完成的第三转换子模型训练声码器，以此提高训练效率。

如图 12 所示，第三转换模型的训练过程可以包括如下步骤。

步骤 1201，将样本内容特征以及说话者标识输入第三转换子模型，得到预测音频谱特征。

在一种可能的实施方式中，计算机设备将样本内容特征以及说话者标识输入第三转换子模型，得到样本音频对应的预测 Mel 频谱。

步骤 1202，以样本音频的样本音频谱特征为预测音频谱特征的监督，训练第三转换子模型。

在一些实施例中，计算机设备对样本音频进行音频谱特征提取，得到样本音频谱特征，从而基于预测音频谱特征与样本音频谱特征的差异确定第三转换子模型损失，从而基于该第三转换子模型损失训练第三转换子模型。

其中，该损失可以为 MSE 损失或其他类型损失，本实施例对此不作限定。

可选的，第三转换子模型的损失 $loss_{BN2Mel}$ 可以表示为：

$$loss_{BN2Mel} = \|Mel - \widehat{Mel}\|$$

其中， Mel 为直接对样本音频进行音频谱特征提取到的样本音频谱特征， $\widehat{Mel}$ 为第三转换子模型输出的预测音频谱特征。

步骤 1203，在第三转换子模型训练完成的情况下，将训练完成后第三转换子模型输出的预测音频谱特征输入声码器，得到预测音频。

完成第三转换子模型训练后，计算机设备将样本内容特征以及说话者标识输入训练得到的第三转换子模型，得到预测音频谱特征，然后将该预测音频谱特征输入声码器，得到声码器输出的预测音频。

在一个示意性的例子中，计算机设备将训练完成的 BN2Mel 子模型输出的预测 Mel 频谱特征输入 hifigan，得到 hifigan 输出的预测音频。

步骤 1204，基于预测音频以及样本音频，训练第三转换模型中的声码器。

可选的，计算机设备以样本音频为预测音频的监督，确定声码器的转换损失，从而基于该损失训练声码器。

在一些实施例中，当声码器采用对抗网络时，以 hifigan 为例，计算机设备采用对抗训练思想，通过生成器（Generator）和判别器（Discriminator）对抗训练。对抗训练过程中生成器的损失可以表示为：

$$L_G = L_G(G; D) + L_{FM}(G; D) + L_{mel}(G)$$

其中， $L_{mel}(G) = \|\varphi(x) - \varphi(G(s))\|$ ， $\varphi(G(s))$ 为生成器生成的音频 $G(s)$ 重新转换得到的 Mel 谱特征， $\varphi(x)$ 为从样本音频中提取到的 Mel 谱特征； $L_{FM}(G; D)$ 为生成音频与样本音频的特征匹配损失； $L_G(G; D)$ 为生成音频的判别损失。

对抗训练过程中判别器的损失可以表示为：

$$L_D(G; D) = (D(x) - 1)^2 + (D(G(s)))^2$$

其中， $D(x)$ 为判别器对样本音频的判别结果， $D(G(s))$ 为判别器对预测音频的判别结果。

显然，通过上述方式训练得到的第三转换模型，处理能够将内容特征转换为音频外，还能够在转换得到的音频中添加特定的音色。相应的，在应用过程中，用户除了选择目标口音

外，还可以选择目标音色。

在一种可能的实施方式中，在口音转换指令中包含目标音色的情况下，计算机设备将第二内容特征以及目标音色对应说话者的说话者标识输入第三转换模型，得到第二口音音频，其中，第二口音音频具有第二口音以及目标音色。

示意性的，如图 13 所示，当需要将普通话转换为方言，且具有目标音色时，计算机设备通过普通话 ASR 模型 1302 对普通话音频 1301 进行特征提取，得到普通话内容特征 1303。计算机设备进一步通过 BN2BN 模型 1304 对普通话内容特征 1303 进行特征转换，得到方言内容特征 1305。计算机设备将方言内容特征 1305 以及目标音色 1306 对应的说话者标识输入 BN2Wav 模型 1307，得到 BN2Wav 模型 1307 输出的带有目标音色的方言音频 1308。

需要说明的是，当需要保持口音转换前后音色一致时，需要预先获取当前用户的语料数据（比如累计 30 分钟时长的语音数据），并为当前用户分配说话者标识，从而基于当前用户的语料数据以及说话者标识训练第三转换模型，本实施例在此不作赘述。

本实施例中，在训练第三转换模型的过程中，除了将样本音频的内容特征作为输入外，还将样本音频对应说话者标识作为输入，使第三转换模型在训练中能够基于内容特征以及说话者的音色特征进行音频转换。后续使用过程中，对于相同文本内容，通过输入不同的说话者标识，第三转换模型能够输出不同音色的音频，实现口音以及音色的双重转换。

图 14 是本申请一个示例性实施例提供的语音转换模型的训练装置的结构框图，该装置包括：

训练模块 1401，用于基于第一样本音频训练第一 ASR 模型，以及基于第二样本音频训练第二 ASR 模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

所述训练模块 1401，还用于基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一 ASR 模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

所述训练模块 1401，还用于基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二 ASR 模型对所述第二样本音频进行提取得到，所述平行样本数据包括不同内容特征，不同内容特征对应不同口音，且不同内容特征对应相同文本；基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

所述训练模块 1401，还用于基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

生成模块 1402，用于基于训练得到的所述第一 ASR 模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

可选的，所述训练模块 1401，用于：

通过所述第一转换模型将所述第二样本文本转换得到第三样本内容特征，所述第三样本内容特征指采用所述第一口音表述第二样本文本所产生音频的内容特征；

基于所述第二样本内容特征和所述第三样本内容特征构建所述平行样本数据。

可选的，所述训练模块 1401，用于：将所述第三样本内容特征输入所述第二转换模型，得到第二预测内容特征；

以所述第二样本内容特征为所述第二预测内容特征的监督，训练所述第二转换模型。

可选的，所述训练模块 1401，用于：将所述第一样本文本输入所述第一转换模型，得到所述第一转换模型输出的第一预测内容特征；

以所述第一样本内容特征为所述第一预测内容特征的监督，训练所述第一转换模型。

可选的，所述第一转换模型中包括第一转换子模型、时长预测子模型以及第二转换子模型；

所述训练模块 1401，用于：

通过所述第一转换子模型对所述第一样本文本进行编码，得到第一文本编码特征；

通过所述时长预测子模型对所述第一文本编码特征进行时长预测，得到预测时长，所述预测时长用于表征所述第一样本文本的发音时长；

基于所述预测时长对所述第一文本编码特征进行特征扩充，得到第二文本编码特征；

通过所述第二转换子模型将所述第二文本编码特征转换得到所述第一预测内容特征。

可选的，所述第一转换子模型和所述第二转换子模型包括 FFT，所述 FFT 包括多头注意力机制层和卷积层。

可选的，所述训练模块 1401，用于：

将所述样本内容特征以及所述样本音频对应的说话者标识输入所述第三转换模型，得到预测音频，所述预测音频与所述样本音频对应相同音频内容，且具有相同音色，其中，不同说话者对应不同说话者标识；

基于所述预测音频以及所述样本音频，训练所述第三转换模型。

可选的，所述第三转换模型包括第三转换子模型以及声码器；

所述训练模块 1401，用于：

将所述样本内容特征以及所述说话者标识输入所述第三转换子模型，得到预测音频谱特征；

将所述预测音频谱特征输入所述声码器，得到所述预测音频。

可选的，所述训练模块 1401，用于：

以所述样本音频的样本音频谱特征为所述预测音频谱特征的监督，训练所述第三转换子模型；

在所述第三转换子模型训练完成的情况下，将训练完成后所述第三转换子模型输出的所述预测音频谱特征输入所述声码器，得到所述预测音频；

基于所述预测音频以及所述样本音频，训练所述第三转换模型中的所述声码器。

可选的，所述装置还包括：

转换模块，用于响应于口音转换指令，通过所述第一 ASR 模型提取第一口音音频的第一内容特征，第一内容特征对应所述第一口音，所述口音转换指令用于指示将音频由所述第一口音转换为所述第二口音；

通过所述第二转换模型将所述第一内容特征转换为第二内容特征，所述第二内容特征对应所述第二口音；

通过所述第三转换模型对所述第二内容特征进行音频转换，得到第二口音音频。

可选的，所述口音转换指令中包含目标音色；

所述转换模块，用于：

将所述第二内容特征以及所述目标音色对应说话者的说话者标识输入所述第三转换模型，得到所述第二口音音频，其中，不同说话者对应不同说话者标识。

综上所述，本申请实施例中，在缺少第二口音的第二样本音频对应平行语料的情况下，首先基于第一口音的第一样本音频，训练用于将文本转换为内容特征的第一转换模型，从而利用该第一转换模型以及第二样本音频对应的第二样本文本，构建得到包含对应相同文本内容但对应不同口音的平行样本数据，进而利用该平行样本数据训练在不同口音间进行内容特征转换的第二转换模型，以及用于将内容特征转换为音频的第三转换模型，完成语音转换模型训练；模型训练过程中，利用训练得到的中间模型构建平行语料，无需在模型训练前录制不同口音的平行语料，能够在保证模型训练质量的情况下，降低模型训练对人工录制的平行语料的需求，有助于提高模型训练效率，并提高样本不足情况下模型的训练质量。

需要说明的是：上述实施例提供的装置，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，即将装置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，其实现过程详见方法实施例，这里不再赘述。

图 15 是本申请一个示例性实施例提供的语音转换装置的结构框图，该装置包括：

获取模块 1501，用于获取第一口音音频，所述第一口音音频对应第一口音；

提取模块 1502，用于通过所述第一 ASR 模型对所述第一口音音频提取得到第一内容特征，所述第一内容特征对应所述第一口音；

内容特征转换模块 1503，用于通过所述第二转换模型将所述第一内容特征转换为第二内容特征，所述第二内容特征对应第二口音；

音频转换模块 1504，用于通过所述第三转换模型对所述第二内容特征进行音频转换，得到第二口音音频。

在一些实施例中，内容特征转换模块 1503，还用于将第一 ASR 模型提取到的第一内容特征输入第二转换模型，由第二转换模型在第一口音和第二口音之间进行内容特征转换，得到第二口音下的第二内容特征。

在一些实施例中，第二转换模型包括卷积层以及 N 层堆叠的 FFT。

在一些实施例中，内容特征转换模块 1503，还用于通过第二转换模型中的卷积层对第一内容特征进行卷积处理，将卷积结果输入至 N 层堆叠的 FFT 转换得到第二内容特征。

在一些实施例中，第三转换模型包括第三转换子模型以及声码器，其中，第三转换子模型用于将内容特征转换为音频谱特征，声码器则用于基于音频谱特征生成音频。

在一些实施例中，音频转换模块 1504，还用于将第二内容特征以及说话者标识输入第三转换子模型，得到音频谱特征。

在一些实施例中，音频转换模块 1504，还用于将音频谱特征输入声码器，得到第二口音音频。

在一些实施例中，第三转换子模型包括卷积层以及 N 层堆叠的 FFT。

请参考图 16，其示出了本申请一个示例性实施例提供的计算机设备的结构示意图，该计算机设备可以为上述实施例中的投屏设备或终端。具体来讲：所述计算机设备 1600 包括中央处理单元（Central Processing Unit, CPU）1601、包括随机存取存储器 1602 和只读存储器 1603 的系统存储器 1604，以及连接系统存储器 1604 和中央处理单元 1601 的系统总线 1605。所述计算机设备 1600 还包括帮助计算机内的各个器件之间传输信息的基本输入/输出系统（Input/Output, I/O 系统）1606，和用于存储操作系统 1613、应用程序 1614 和其他程序模块 1615 的大容量存储设备 1607。

所述基本输入/输出系统 1606 包括有用于显示信息的显示器 1608 和用于用户输入信息的诸如鼠标、键盘之类的输入设备 1609。其中所述显示器 1608 和输入设备 1609 都通过连接到系统总线 1605 的输入输出控制器 1610 连接到中央处理单元 1601。所述基本输入/输出系统 1606 还可以包括输入输出控制器 1610 以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地，输入输出控制器 1610 还提供输出到显示屏、打印机或其他类型的输出设备。

所述大容量存储设备 1607 通过连接到系统总线 1605 的大容量存储控制器（未示出）连接到中央处理单元 1601。所述大容量存储设备 1607 及其相关联的计算机可读介质为计算机设备 1600 提供非易失性存储。也就是说，所述大容量存储设备 1607 可以包括诸如硬盘或者驱动器之类的计算机可读介质（未示出）。

不失一般性，所述计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括随机存取记

忆体 (RAM, Random Access Memory)、只读存储器 (ROM, Read Only Memory)、闪存或其他固态存储其技术, 只读光盘 (Compact Disc Read-Only Memory, CD-ROM)、数字通用光盘 (Digital Versatile Disc, DVD) 或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然, 本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器 1604 和大容量存储设备 1607 可以统称为存储器。

存储器存储有一个或多个程序, 一个或多个程序被配置成由一个或多个中央处理单元 1601 执行, 一个或多个程序包含用于实现上述方法的指令, 中央处理单元 1601 执行该一个或多个程序实现上述各个方法实施例提供的方法。

根据本申请的各种实施例, 所述计算机设备 1600 还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即计算机设备 1600 可以通过连接在所述系统总线 1605 上的网络接口单元 1611 连接到网络 1612, 或者说, 也可以使用网络接口单元 1611 来连接到其他类型的网络或远程计算机系统 (未示出)。

本申请实施例还提供一种计算机可读存储介质, 该可读存储介质中存储有至少一条指令, 至少一条指令由处理器加载并执行以实现上述实施例所述的语音转换模型的训练方法, 或, 如上述方面所述的语音转换方法。

可选地, 该计算机可读存储介质可以包括: ROM、RAM、固态硬盘 (SSD, Solid State Drives) 或光盘等。其中, RAM 可以包括电阻式随机存取记忆体 (ReRAM, Resistance Random Access Memory) 和动态随机存取存储器 (DRAM, Dynamic Random Access Memory)。

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上述实施例所述的语音转换模型的训练方法, 或, 如上述方面所述的语音转换方法。

权 利 要 求 书

1.一种语音转换模型的训练方法，所述方法由计算机设备执行，所述方法包括：

基于第一样本音频训练第一 ASR 模型，以及基于第二样本音频训练第二 ASR 模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一 ASR 模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二 ASR 模型对所述第二样本音频进行提取得到，所述平行样本数据包括不同内容特征，不同内容特征对应不同口音，且不同内容特征对应相同文本；

基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

基于训练得到的所述第一 ASR 模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

2.根据权利要求 1 所述的方法，其中，所述基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，包括：

通过所述第一转换模型将所述第二样本文本转换得到第三样本内容特征，所述第三样本内容特征指采用所述第一口音表述第二样本文本所产生音频的内容特征；

基于所述第二样本内容特征和所述第三样本内容特征构建所述平行样本数据。

3.根据权利要求 2 所述的方法，其中，所述基于所述平行样本数据训练第二转换模型，包括：

将所述第三样本内容特征输入所述第二转换模型，得到第二预测内容特征；

以所述第二样本内容特征为所述第二预测内容特征的监督，训练所述第二转换模型。

4.根据权利要求 3 所述的方法，其中，所述以所述第二样本内容特征为所述第二预测内容特征的监督，训练所述第二转换模型，包括：

根据所述第二样本内容特征与所述第二预测内容特征之间的差异确定第二转换模型损失，基于所述第二转换模型损失训练所述第二转换模型。

5.根据权利要求 1 至 4 任一所述的方法，其中，所述基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，包括：

将所述第一样本文本输入所述第一转换模型，得到所述第一转换模型输出的第一预测内容特征；

以所述第一样本内容特征为所述第一预测内容特征的监督，训练所述第一转换模型。

6.根据权利要求 5 所述的方法，其中，所述第一转换模型中包括第一转换子模型、时长预测子模型以及第二转换子模型；

所述将所述第一样本文本输入所述第一转换模型，得到所述第一转换模型输出的第一预测内容特征，包括：

通过所述第一转换子模型对所述第一样本文本进行编码，得到第一文本编码特征；

通过所述时长预测子模型对所述第一文本编码特征进行时长预测，得到预测时长，所述预测时长用于表征所述第一样本文本的发音时长；

基于所述预测时长对所述第一文本编码特征进行特征扩充，得到第二文本编码特征；

通过所述第二转换子模型将所述第二文本编码特征转换得到所述第一预测内容特征。

7.根据权利要求 6 所述的方法,其中,所述第一转换子模型和所述第二转换子模型包括 FFT,所述 FFT 包括多头注意力机制层和卷积层。

8.根据权利要求 5 所述的方法,其中,所述以所述第一样本内容特征为所述第一预测内容特征的监督,训练所述第一转换模型,包括:

根据所述第一样本内容特征与所述第一预测内容特征之间的差异确定第一转换模型损失,基于所述第一转换模型损失训练所述第一转换模型。

9.根据权利要求 1 至 8 任一所述的方法,其中,所述基于不同样本音频的样本内容特征训练第三转换模型,包括:

将所述样本内容特征以及所述样本音频对应的说话者标识输入所述第三转换模型,得到预测音频,所述预测音频与所述样本音频对应相同音频内容,且具有相同音色,其中,不同说话者对应不同说话者标识;

基于所述预测音频以及所述样本音频,训练所述第三转换模型。

10.根据权利要求 9 所述的方法,其中,所述第三转换模型包括第三转换子模型以及声码器;

所述将所述样本内容特征以及所述样本音频对应的说话者标识输入所述第三转换模型,得到预测音频,包括:

将所述样本内容特征以及所述说话者标识输入所述第三转换子模型,得到预测音频谱特征;

将所述预测音频谱特征输入所述声码器,得到所述预测音频。

11.根据权利要求 10 所述的方法,其中,所述将所述预测音频谱特征输入所述声码器,得到所述预测音频之前,所述方法还包括:

以所述样本音频的样本音频谱特征为所述预测音频谱特征的监督,训练所述第三转换子模型;

所述将所述预测音频谱特征输入所述声码器,得到所述预测音频,包括:

在所述第三转换子模型训练完成的情况下,将训练完成后所述第三转换子模型输出的所述预测音频谱特征输入所述声码器,得到所述预测音频;

所述基于所述预测音频以及所述样本音频,训练所述第三转换模型,包括:

基于所述预测音频以及所述样本音频,训练所述第三转换模型中的所述声码器。

12.根据权利要求 11 所述的方法,其中,所述以所述样本音频的样本音频谱特征为所述预测音频谱特征的监督,训练所述第三转换子模型,包括:

对所述样本音频进行音频谱特征提取,得到样本音频谱特征;

根据所述预测音频谱特征与所述样本音频谱特征的差异确定第三转换子模型损失,基于所述第三转换子模型损失训练所述第三转换子模型。

13.根据权利要求 1 至 12 任一所述的方法,其中,所述方法包括:

响应于口音转换指令,通过所述第一 ASR 模型提取第一口音音频的第一内容特征,第一内容特征对应所述第一口音,所述口音转换指令用于指示将音频由所述第一口音转换为所述第二口音;

通过所述第二转换模型将所述第一内容特征转换为第二内容特征,所述第二内容特征对应所述第二口音;

通过所述第三转换模型对所述第二内容特征进行音频转换,得到第二口音音频。

14.根据权利要求 13 所述的方法,其中,所述口音转换指令中包含目标音色;

所述通过所述第三转换模型对所述第二内容特征进行音频转换,得到第二口音音频,包括:

将所述第二内容特征以及所述目标音色对应说话者的说话者标识输入所述第三转换模型,得到所述第二口音音频,其中,不同说话者对应不同说话者标识。

15.一种语音转换方法，所述方法由计算机设备执行，所述计算机设备中设置有语音转换模型，所述语音转换模型包括第一 ASR 模型、第二转换模型和第三转换模型，所述方法包括：

获取第一口音音频，所述第一口音音频对应第一口音；

通过所述第一 ASR 模型对所述第一口音音频提取得到第一内容特征，所述第一内容特征对应所述第一口音；

通过所述第二转换模型将所述第一内容特征转换为第二内容特征，所述第二内容特征对应第二口音；

通过所述第三转换模型对所述第二内容特征进行音频转换，得到第二口音音频。

16.一种语音转换模型的训练装置，其中，所述装置包括：

训练模块，用于基于第一样本音频训练第一 ASR 模型，以及基于第二样本音频训练第二 ASR 模型，所述第一样本音频对应第一口音，所述第二样本音频对应第二口音；

所述训练模块，还用于基于所述第一样本音频对应的第一样本文本以及第一样本内容特征，训练第一转换模型，所述第一样本内容特征由所述第一 ASR 模型对所述第一样本音频进行提取得到，所述第一转换模型用于将文本转换为所述第一口音的内容特征；

所述训练模块，还用于基于所述第一转换模型、所述第二样本音频对应的第二样本文本以及第二样本内容特征，构建平行样本数据，所述第二样本内容特征由所述第二 ASR 模型对所述第二样本音频进行提取得到，所述平行样本数据包括不同内容特征，不同内容特征对应不同口音，且不同内容特征对应相同文本；基于所述平行样本数据训练第二转换模型，所述第二转换模型用于对所述第一口音和所述第二口音间进行内容特征转换；

所述训练模块，还用于基于不同样本音频的样本内容特征训练第三转换模型，所述第三转换模型用于将内容特征转换为音频；

生成模块，用于基于训练得到的所述第一 ASR 模型、所述第二转换模型和所述第三转换模型生成语音转换模型，所述语音转换模型用于将第一口音的音频转换为第二口音的音频。

17.一种语音转换装置，其中，所述装置包括：

获取模块，用于获取第一口音音频，所述第一口音音频对应第一口音；

提取模块，用于通过所述第一 ASR 模型对所述第一口音音频提取得到第一内容特征，所述第一内容特征对应所述第一口音；

内容特征转换模块，用于通过所述第二转换模型将所述第一内容特征转换为第二内容特征，所述第二内容特征对应第二口音；

音频转换模块，用于通过所述第三转换模型对所述第二内容特征进行音频转换，得到第二口音音频。

18.一种计算机设备，其中，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行以实现如权利要求 1 至 14 任一所述的语音转换模型的训练方法，或，如权利要求 15 所述的语音转换方法。

19.一种计算机可读存储介质，其中，所述可读存储介质中存储有至少一条指令，所述至少一条指令由处理器加载并执行以实现如权利要求 1 至 14 任一所述的语音转换模型的训练方法，或，如权利要求 15 所述的语音转换方法。

20.一种计算机程序产品，其中，所述计算机程序产品包括计算机指令，所述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所述计算机可读存储介质读取所述计算机指令，所述处理器执行所述计算机指令，使得所述计算机设备执行如权利要求 1 至 14 任一所述的语音转换模型的训练方法，或，如权利要求 15 所述的语音转换方法。