(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2023 年 6 月 29 日 (29.06.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/116660 A2

(51) 国际专利分类号:
G10L 13/02 (2013.01)

(21) 国际申请号: PCT/CN2022/140253

(22) 国际申请日: 2022 年 12 月 20 日 (20.12.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111577618.0 2021年12月22日 (22.12.2021) CN

(71) 申请人: 广州市百果园网络科技有限公司 (GUANGZHOU BAIGUOYUAN NETWORK TECHNOLOGY CO., LTD.) [CN/CN]; 中国广东省广州市番禺区市桥街兴泰路278号基盛商业中心4、5层, Guangdong 511402 (CN)。

(72) 发明人: 黄家鸿 (HUANG, Jiahong); 中国广东省广州市番禺区市桥街兴泰路278号基盛商业中心4、5层, Guangdong 511402 (CN)。 李玉乐 (LI, Yule); 中国广东省广州市番禺区市桥街兴泰路278号基盛商业中心4、5层, Guangdong 511402 (CN)。 项伟 (XIANG, Wei); 中国广东省广州市番禺区市桥街兴泰路278号基盛商业中心4、5层, Guangdong 511402 (CN)。

(74) 代理人: 北京泽方誉航专利代理事务所（普通合伙） (BEIJING ZFANG PATENT AGENCY); 中国北京市丰台区南四环西路186号四区2号楼7层23室徐濛, Beijing 100071 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(54) **Title:** MODEL TRAINING AND TONE CONVERSION METHOD AND APPARATUS, DEVICE, AND MEDIUM

(54) 发明名称: 一种模型训练以及音色转换方法、装置、设备及介质

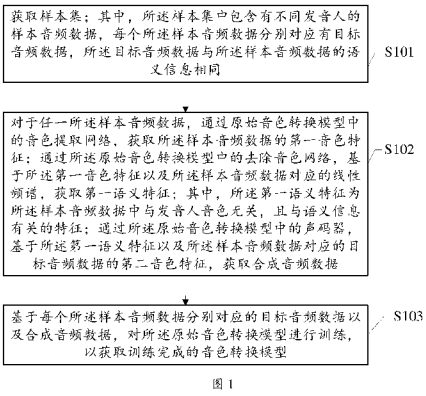

图 1

S101 Obtain a sample set; wherein the sample set comprises sample audio data of various speakers, each piece of the sample audio data corresponds to a piece of target audio data, and the target audio data and the sample audio data have the same semantic information

S102 For any one piece of the sample audio data, by means of a tone extraction network in an original tone conversion model, obtain a first tone feature of input sample audio data; by means of a tone-removing network in the original tone conversion model, and on the basis of the first tone feature and a linear spectrum corresponding to the sample audio data, obtain a first semantic feature; wherein the first semantic feature is a feature of the sample audio data that is not-related to the tone of the speaker but is related to the semantic information; by means of a vocoder in the original tone conversion model, and on the basis of the first semantic feature and a second tone feature of the target audio data corresponding to the sample audio data, obtain synthesized audio data

S103 On the basis of the target audio data and the synthesized audio data respectively corresponding to each piece of sample audio data, train the original tone conversion model, so as to obtain a trained tone conversion model

(57) **Abstract:** The present application provides a model training and tone conversion method and apparatus, a device and a medium. By means of a tone extraction network, a first tone feature of input sample audio data can be obtained, so as to obtain tone information of the input sample audio data, which facilitates subsequently obtaining synthesized audio data according to the tone feature, thereby improving the accuracy of the tone of the synthesized audio data. By means of a tone-removing network, and on the basis of the first tone feature, a first semantic feature of the sample audio data can be obtained, thereby accurately obtaining a feature of the sample audio data that is not-related to the tone of the speaker but is related to the spoken content, which facilitates subsequently obtaining synthesized audio data according to the first semantic feature, and ensures the accuracy of the spoken content of the synthesized audio data. After obtaining the trained tone conversion model, tone conversion is carried out by means of the tone conversion model, so that the conversion effect and reliability of tone conversion can be improved.

WO 2023/116660 A2

[见续页]

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 不包括国际检索报告，在收到该报告后将重新公布(细则48.2(g))。

(57) 摘要：本申请提供了一种模型训练以及音色转换方法、装置、设备及介质。由于通过该音色提取网络可以获取到输入的样本音频数据的第一音色特征，从而准确地获取到输入的样本音频数据的音色信息，有利于后续根据该音色特征获取合成音频数据，提高合成音频数据的音色的准确性，通过该去除音色网络，基于该第一音色特征，可以获取到该样本音频数据的第一语义特征，实现了准确地获取到样本音频数据中与发音人音色无关，且与发音内容有关的特征，从而有利于后续根据该第一语义特征获取合成音频数据，保证合成音频数据的发音内容的准确性。获取到训练完成的音色转换模型后，通过该音色转换模型进行音色转换，可以提高音色转换的转换效果以及可靠性。

一种模型训练以及音色转换方法、装置、设备及介质

本申请要求在2021年12月22日提交中国专利局，申请号为202111577618.0的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请涉及自然语言理解技术领域，尤其涉及一种模型训练以及音色转换方法、装置、设备及介质。

背景技术

音色转换技术是音频领域中一项重要的技术，广泛运用于音频内容生成、娱乐音频制作和保密通话等领域。音色转换技术是指将原始音频中的声音的音色转换为另外一个的说话人声音的音色。在音色转换过程中，需要保证转换音色之后的音频的音色与另外一个说话人声音的音色相似，而音频的内容保持不变。音色转换技术的难点在于如何保持原始音频的内容信息的同时进行音色变换。

有鉴于此，如何进行音色转换，获得稳定的音色转换效果，是亟待解决的技术问题。

发明内容

本申请实施例提供了一种模型训练以及音色转换方法、装置、设备及介质，用以解决现有音色转换的转换效果差，降低了音色转换的可靠性的问题。

本申请实施例提供了一种音色转换模型训练方法，所述方法包括：

获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

本申请实施例提供了一种音色转换方法，所述方法包括：

获取源音频数据以及目标发音人的音色特征；

通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

本申请实施例提供了一种音色转换模型训练装置，所述装置包括：

获取单元，配置为获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

处理单元，配置为对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以

及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

训练单元，配置为基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

本申请实施例提供了一种音色转换装置，所述装置包括：

获取模块，配置为获取源音频数据以及目标发音人的音色特征；

合成模块，配置为通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

本申请实施例提供了一种电子设备，所述电子设备至少包括处理器和存储器，所述处理器配置为执行存储器中存储的计算机程序时实现如上述所述音色转换模型训练方法的步骤，或者，实现如上述所述音色转换方法的步骤。

本申请实施例提供了一种计算机可读存储介质，其存储有计算机程序，所述计算机程序被处理器执行时实现如上述所述音色转换模型训练方法的步骤，或者，实现如上述所述音色转换方法的步骤。

本申请实施例还提供一种计算机程序产品，包括计算机程序，该计算机程序被执行时，可以实现如上述所述音色转换模型训练方法的步骤，或者，实现如上述所述音色转换方法的步骤。

由于在对音色转换模型进行训练的过程中，无需预先对样本集中的样本音频数据进行标注，减少了对样本音频数据进行标注所耗费的成本，方便后续基于样本集中的样本音频数据对音色转换模型的训练。由于原始音色转换模型中包括音色提取网络、去除音色网络以及声码器，通过该音色提取网络可以获取到输入的样本音频数据的第一音色特征，从而准确地获取到输入的样本音频数据的音色信息，有利于后续根据该音色特征获取合成音频数据，提高合成音频数据的音色的准确性，通过该去除音色网络，基于该第一音色特征，可以获取到输入的样本音频数据的第一语义特征，实现了准确地获取到样本音频数据中与发音人音色无关，且与发音内容有关的特征，从而有利于后续根据该第一语义特征获取合成音频数据，保证合成音频数据的发音内容的准确性。通过该声码器，基于该第一语义特征以及该样本音频数据对应的目标音频数据的第二音色特征，即可获取到合成音频数据。基于每个样本音频数据分别对应的目标音频数据以及合成音频数据，即可对原始音色转换模型进行训练，以获取训练完成的音色转换模型，实现了无监督训练音色转换模型，极大降低了获取音色转换模型的难度。后续基于训练完成的音色转换模型进行音色转换，可以提高音色转换的转换效果以及可靠性。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简要介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域的普通技术人员来讲，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 为本申请实施例提供的一种音色转换模型训练过程示意图；

图 2 为本申请实施例提供的一种音色转换模型的结构示意图；

图 3 为本申请实施例提供了一种音色转换过程示意图；

图 4 为本申请实施例提供的一种音色转换模型的结构示意图；

图 5 为本申请实施例提供的一种音色转换模型训练装置的结构示意图；

图 6 为本申请实施例提供的一种音色转换装置的结构示意图；

图 7 为本申请实施例提供的一种电子设备结构示意图；

图 8 为本申请实施例提供的再一种电子设备结构示意图。

具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请作进一步地详细描述，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它实施例，都属于本申请保护的范围。

本领域技术人员知道，本申请的实施方式可以实现为一种系统、装置、设备、方法或计算机程序产品。因此，本申请可以具体实现为以下形式，即：完全的硬件、完全的软件（包括固件、驻留软件、微代码等），或者硬件和软件结合的形式。

在本文中，需要理解的是，附图中的任何元素数量均用于示例而非限制，以及任何命名都仅用于区分，而不具有任何限制含义。

目前，可以采用基于文本转语音（Test to Speech，TTS）的音色转换技术、基于语音识别（Automatic Speech Recognition，ASR）的音色转换技术、基于生成式对抗网络（Generative Adversarial Network，GAN）的音色转换技术、以及基于变分自动编码器（Variational Auto Encoder，VAE）的音色转换技术，从而实现音色转换。在采用这些方法进行音色转换时，如果要保证合成音频数据的发音内容的准确性，则会无法保证合成音频数据中的语气等与语义内容无关的音色特征，比如叹气、尖叫等，降低合成音频数据的自然度以及音色转换效果。

以基于 ASR 的音色转换技术为例，在该方法中，一般会采用一个预训练好的 ASR 模型来提取音频的语义信息，并通过音色提取模型提取目标发音人的音频数据中的音色信息，根据该语义信息以及音色信息，生成目标发音人的合成音频数据。由于该种方法严重依赖预训练的 ASR 模型，ASR 模型的精度直接影响合成音频数据的音色转换效果。同时 ASR 模型主要是做语义提取的，其会忽略音频数据中的语气等与语义内容无关的音色信息，导致合成音频数据会丢失语气等与语义内容无关的信息。

以基于 TTS 的音色转换技术为例，在该方法中，需要预先采集目标发音人大量的音频数据后，基于采集到的每个音频数据以及每个音频数据分别对应的语义信息，训练得到该目标发音人的声学模型以及声码器。后续根据文本信息的文本特征以及训练完成的该目标发音人的声学模型以及声码器，获取目标发音人的合成音频数据。由于一般需要预先采集 3 万句以上、或者 30 小时以上的该目标发音人的音频数据，才能保证该目标发音人的声学模型以及声码器的精度，导致需要对每个音频数据的语义进行标注，增加了获取该目标发音人的声学模型以及声码器的难度以及所耗费的成本都非常的大，且获取到的合成音频数据中的语气等音色信息会比较固定，降低合成音频数据的自然度以及音色转换效果。

为了解决上述的问题，本申请实施例提供一种模型训练以及音色转换方法、装置、设备及介质。由于在对音色转换模型进行训练的过程中，无需预先对样本集中的样本音频数据进行标注，减少了对样本音频数据进行标注所耗费的成本，方便后续基于样本集中的样本音频数据对音色转换模型的训练。由于在对音色转换模型进行训练的过程中，无需预先对样本集中的样本音频数据进行标注，减少了对样本音频数据进行标注所耗费的成本，方便后续基于样本集中的样本音频数据对音色转换模型的训练。由于原始音色转换模型中包括音色提取网络、去除音色网络以及声码器，通过该音色提取网络可以获取到输入的样本音频数据的第一音色特征，从而准确地获取到输入的样本音频数据的音色信息，有利于后续根据该音色特征获取

合成音频数据，提高合成音频数据的音色的准确性，通过该去除音色网络，基于该第一音色特征，可以获取到输入的样本音频数据的第一语义特征，实现了准确地获取到样本音频数据中与发音人音色无关，且与发音内容有关的特征，从而有利于后续根据该第一语义特征获取合成音频数据，保证合成音频数据的发音内容的准确性。通过该声码器，基于该第一语义特征以及该样本音频数据对应的目标音频数据的第二音色特征，即可获取到合成音频数据。基于每个样本音频数据分别对应的目标音频数据以及合成音频数据，即可对原始音色转换模型进行训练，以获取训练完成的音色转换模型，实现了无监督训练音色转换模型，极大降低了获取音色转换模型的难度。后续基于训练完成的音色转换模型进行音色转换，可以提高音色转换的转换效果以及可靠性。

需要说明的是，上述实施例中所举出的应用场景仅是为了方便说明所提出的示例性的场景，并不是对本申请实施例所提供的一种模型训练以及音色转换方法、装置、设备及介质的应用场景的一种限定。本领域普通技术人员可知，随着新业务场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

实施例 1：

图 1 为本申请实施例提供的一种音色转换模型训练过程示意图，该过程包括：

S101：获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同。

本申请实施例提供的音色转换模型训练方法应用于电子设备，该电子设备可以为如机器人、移动终端等智能设备，也可以是服务器。

一般情况下，可以采集不同发音人的音频数据，并将这些音频数据确定为样本音频数据，以先通过这些样本音频数据，对原始音色合成模型进行训练，以获取训练完成的音色合成模型，以提高音色转换模型的鲁棒性以及音色转换模型可以合成的音色的种类。

本申请实施例中，S101 中获取样本音频数据包括以下至少一种方式：

方式 1、将录制得到的不同发音人的音频数据，确定为样本音频数据。

在采集样本音频数据的过程中，发音人可以在专业的录音环境中录制语音数据，将录制的语音数据确定为样本音频数据，也可以是通过智能终端（如手机、平板电脑等）录制语音数据，基于录制的语音数据确定样本音频数据。

例如，发音人可以向智能终端输入触发操作。其中，发音人向智能终端输入触发操作的方式有很多，该触发操作可以是发音人触发了智能终端的显示屏上显示的虚拟按钮，也可以是发音人向智能终端输入了语音信息，还可以是发音人在智能终端的显示屏上绘制了图形指令等，具体实施过程中，可以根据实际需求进行灵活设置，在此不做具体限定。智能终端接收到了发音人输入的触发操作后，可以将发音人选择的预先录制好的语音数据上传给电子设备，也可以进入语音录制功能，开始实时录制发音人的语音数据，并将录制的多条语音数据上传给电子设备，以使电子设备基于接收到的语音数据确定样本音频数据。

由于发音人无需到专业的录音环境录制语音数据，通过智能终端便可以录制语音数据，降低了获取发音人的语音数据的难度以及成本，极大地提高了用户体验。

当通过智能终端录制语音数据时，可以将通过智能终端录制得到的语音数据确定为发音人的原始语音数据。由于该原始语音数据中可能存在大量的工作环境中的噪声，因此可以先对录制得到的原始语音数据进行音频处理，比如，对该原始语音数据进行降噪处理，和/或去混响处理，以得到干净的语音数据。然后，将音频处理后的语音数据确定为样本音频数据。

需要说明的是，对该原始语音数据进行降噪处理，和/或去混响处理的具体过程，均属于现有技术，在

此不做具体赘述。

方式2、可以在上述方式1的基础上，将同一发音人的至少两个语音数据进行拼接处理，将拼接处理后获取到的拼接语音数据，确定为样本音频数据，这样可以实现对获取到的样本音频数据进行扩充，进一步降低获取样本音频数据的难度和所耗费的成本，也有利于根据获取到的大量的样本音频数据，对音色转换模型进行训练，提高获取到的音色转换模型的精度和鲁棒性。

在一种可能的实施方式中，在上述方式1的基础上，将同一发音人的至少两个语音数据进行拼接处理，将拼接处理后获取到的拼接语音数据，确定为样本音频数据可以包括如下几种方式：

方式一，可以将基于上述方式1录制得到的语音数据确定为基础语音数据。其中，该基础语音数据可以是上述实施例中的原始语音数据，也可以是上述实施例中的音频处理后的语音数据。针对不同的发音人，将该发音人的至少两个不同的基础语音数据进行拼接，确定出拼接语音数据（为了方便描述，记为第一拼接语音数据）。将每个基础语音样本以及每个第一拼接语音数据均确定为样本音频数据。

方式二，将基于上述方式1录制得到的语音数据确定为基础语音数据。针对不同的发音人，将该发音人的至少一个基础语音数据复制成设定倍数，将该至少一个复制的语音数据与对应的基础语音数据进行拼接，确定出拼接语音数据（为了方便描述，记为第二拼接语音数据）。可以理解的是，该第二拼接语音数据是由至少两个相同的语音数据拼接而成的。将每个基础语音样本以及每个第二拼接语音数据均确定为样本音频数据。

方式三，将基于上述方式1录制得到的语音数据确定为基础语音数据。针对不同的发音人，将该发音人的至少一个基础语音数据复制成设定倍数，将至少两个相同的语音数据（包括复制后的语音数据与基础语音数据），与该发音人的至少一个处该语音数据之外的其它语音数据进行拼接，确定出拼接语音数据（为了方便描述，记为第三拼接语音数据）。可以理解的是，该第三拼接语音数据包括至少两个相同的语音数据以及至少两个不同的语音数据。将每个基础语音样本以及每个第三拼接语音数据均确定为样本音频数据。

在一种可能的实施方式中，可以同时通过上述的方式一至方式三中的至少两种方式，对获取到的样本音频数据进行扩充。

由于在通过录制得到的发音人的语音数据确定样本音频数据的基础上，可以对发音人的至少两个语音数据进行拼接处理，并将拼接处理得到的语音数据确定为样本音频数据，从而实现对发音人的样本音频数据的扩充，进一步降低获取发音人的样本音频数据的难度和所耗费的成本，也有利于根据获取到的大量的样本音频数据，对音色转换模型进行训练，提高获取到的音色转换模型的精度和鲁棒性。

如果要获取到精度较高的音色转换模型，需要对该音色转换模型的音色转换效果进行监督，使得音色转换模型输出的合成音频数据，接近于目标发音人发出相同发音内容的音频数据（记为目标音频数据）。因此，在获取到样本集中的每个样本音频数据后，针对每个样本音频数据，可以确定该样本音频数据对应的目标音频数据，后续可以根据该目标音频数据，确定基于原始音色转换模型以及该样本音频数据所获取到的合成音频数据是否准确，从而确定原始音色转换模型的音色转换效果。

在一种可能的实施方式中，可以针对样本集中的每个样本音频数据，确定与该样本音频数据的语义信息相同的不同发音人的目标音频数据。其中，该目标音频数据包括样本音频数据以及非样本音频数据中的至少一种，即该目标音频数据包括样本集中与该样本音频数据的语义信息相同的不同发音人的样本音频数据，和/或，与该样本音频数据的语义信息相同的不同发音人的非样本音频数据。

在另一种可能的实施方式中，考虑到针对样本集中的每个样本音频数据，确定与该样本音频数据的语义信息相同的不同发音人的目标音频数据的过程，会耗费大量的成本。因此，在本申请实施例中，针对样

本集中的每个样本音频数据，可以将该样本音频数据确定为该样本音频数据对应的目标音频数据，从而实现无需再耗费资源确定与该样本音频数据音色不同的目标音频数据，降低了对音色转换模型进行训练的难度。

S102：对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据。

在本申请实施例中，预先配置有原始音色转换模型，该原始音色转换模型中的参数的参数值可以是随机设置的，也可以人工预先配置的。当基于上述的实施例获取到样本集后，可以基于样本集中包含的每个样本音频数据，对该原始音色转换模型进行训练，以获取训练完成的音色转换模型。

具体实施过程中，获取该本集中的任一样本音频数据，将该样本音频数据输入到原始音色转换模型。通过该原始音色转换模型，基于该样本音频数据以及该样本音频数据对应的目标音频数据的音色特征（记为第二音色特征），获取该样本音频数据对应的合成音频数据。

在一种示例中，该原始音色转换模型中至少包括音色提取网络、去除音色网络以及声码器，以通过该原始音色转换模型中包括的音色提取网络、去除音色网络以及声码器，对输入的样本音频数据进行准确地处理。其中，该音色提取网络分别与去除音色网络以及声码器连接，去除音色网络与声码器连接，通过该音色提取网络可以获取到输入的样本音频数据中的第一音色特征，从而准确地获取到输入的样本音频数据中与语义内容无关的音色信息，有利于后续根据该第一音色特征获取合成音频数据，提高合成音频数据的音色的准确性，通过该去除音色网络，可以获取到输入的样本音频数据的第一语义特征，该第一语义特征为与样本音频数据的发音人音色无关，只与样本音频数据的语义信息有关的特征，从而有利于后续根据该第一语义特征获取合成音频数据，保证合成音频数据的发音内容的准确性。通过该声码器，基于该第一语义特征以及该样本音频数据对应的目标音频数据的第二音色特征，即可获取到合成音频数据。

具体实施过程中，获取该样本集中的任一样本音频数据，将该样本音频数据输入到原始音色转换模型。通过该原始音色转换模型中的音色提取网络，对该样本音频数据进行相应的处理，获取该样本音频数据对应的第一音色特征。通过原始音色转换模型中的去除音色网络，基于该第一音色特征以及样本音频数据对应的线性频谱，获取语义特征（记为第一语义特征）。通过原始音色转换模型中的声码器，基于获取到的第一语义特征以及该样本音频数据对应的目标音频数据的第二音色特征，获取该样本音频数据对应的合成音频数据。

其中，第一语义特征为样本音频数据中与发音人音色无关，且与发音内容有关的特征，从而避免音色对语义信息的影响，提高获取到的语义信息的准确性。

需要说明的是，该样本音频数据对应的线性频谱可以是通过预设的线性频谱算法（比如，快速傅里叶变换算法等）获取到的。

在一种可能的实施方式中，在将样本集中的任一样本音频数据输入到原始音色转换模型之前，可以获取该样本音频数据的声学特征，将该样本音频数据的声学特征输入到该原始音色转换模型中，以通过该原始音色转换模型来对该声学特征进行处理，获取合成音频数据。

其中，该声学特征可以是梅尔频谱（Mel Spectrogram）梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient，MFCC）、树皮频率倒谱系数（Bark Frequency Cepstrum Coefficient，BFCC）、逆梅尔频率倒谱系数（Inverse Mel Frequency Cepstrum Coefficient，IMFCC）、伽马通频率倒谱系数（Gammatone Frequency

Cepstrum Coefficient，GFCC）、线性预测频率倒谱系数（Linear Prediction Cepstral Coefficients，LPCCs）等类型的声学特征中的任一种。

需要说明的是，该声学特征可以是通过声学特征提取算法获取的，也可以是通过声学特征提取模型获取的。

示例性的，以声学特征为梅尔频谱为例，获取该样本集中的任一样本音频数据的梅尔频谱，将该样本音频数据的梅尔频谱输入到原始音色转换模型中。通过该原始音色转换模型中的音色提取网络，对输入的梅尔频谱进行相应的处理，获取该样本音频数据的第一音色特征，比如，256 维的音色特征（tone_vector）。

需要说明的是，该音色提取网络可以根据声纹模型所包含的网络层确定的，比如端到端声纹网络（Deep Speaker RawNet, GE2E）等。

该原始音色转换模型中的去除音色网络中至少包括后验编码器，以实现准确地获取样本音频数据中与发音人音色无关，且与语义信息有关的特征，提高获取到的第一语义特征的准确性，避免样本音频数据的音色特征对样本音频数据的语义信息的影响。其中，该后验编码器（posterior encoder）与增强子网络连接，该后验编码器配置为从样本音频数据中获取与发音内容有关的隐向量，以根据该隐向量确定该样本音频数据的第一语义特征。

具体的，当基于上述的实施例获取到音色提取网络输出的第一音色特征后，通过该原始音色转换模型中去除音色网络所包含的后验编码器，可以基于该第一音色特征以及样本音频数据对应的线性频谱（linear spectrogrm），获取样本音频数据中语义信息的隐向量。然后根据该隐向量，确定该第一语义特征。

其中，可以直接将该隐向量确定为第一语义特征，也可以通过预设的数学函数（比如，对数函数等），对该隐向量进行处理，将处理后的隐向量确定为第一语义特征。

在一种可能的实施方式中，该原始音色转换模型中的去除音色网络中还包括增强子网络（例如，flow 网络），以增强样本音频数据中与发音人音色无关，且与语义信息有关的特征，提高语义信息的分布表示，进一步提高获取到的第一语义特征的准确性，避免样本音频数据的音色特征对样本音频数据的语义信息的影响。该增强子网络可以与去除音色网络中的后验编码器连接，该增强子网络配置为对后验编码器获取到的隐向量进行增强。可以理解的是，该增强子网络配置为从后验编码器获取到的隐向量中提取到更高维度，更加抽象的与发音人音色无关，且与语义信息有关的特征。

具体实施过程中，当基于上述的实施例获取到音色提取网络输出的第一音色特征后，通过该原始音色转换模型中去除音色网络所包含的后验编码器，可以基于该第一音色特征以及样本音频数据对应的线性频谱，获取样本音频数据中语义信息的隐向量。然后通过该原始音色转换模型中去除音色网络所包含的增强子网络，基于该隐向量，获取增强后的隐向量，即确定第一语义特征。

当基于上述的实施例获取到输入的样本音频数据的第一语义特征后，通过该音色转换模型中的声码器，基于该第一语义特征以及该样本音频数据对应的目标音频数据的第二音色特征，即可获取到发音内容满足第一语义特征，且音色满足第二音色特征的合成音频数据。

其中，该声码器可以是声码器，比如，高效（HiFiGAN）声码器、线性预测（Linear Predictive Coding ，LPC）声码器、World 声码器等，具体实施过程中，可以根据实际需求进行灵活设置，在此不做具体限定。

在一种可能的实施方式中，可以通过该原始音色转换网络中的音色提取网络，获取样本音频数据对应的目标音频数据的第二音色特征。

例如，若目标音频数据为样本音频数据，则可以直接将通过该原始音色转换网络中的音色提取网络所确定的样本音频数据的第一音色特征，确定为第二音色特征。

再例如，若目标音频数据不为样本音频数据，即该目标音频数据为样本集中与该样本音频数据的发音

人不同的样本音频数据，或，与该样本音频数据的发音人不同的非样本音频数据，则在将样本音频数据输入到原始音色转换模型中时，也将该样本音频数据对应的目标音频数据输入到该原始音色转换模型中，从而实现通过该原始音色转换网络中的音色提取网络，获取该目标音频数据的第二音色特征。

S103：基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

精度越高的音色转换模型所确定的样本音频数据对应的合成音频数据，与该样本音频数据对应的目标音频数据越相似。因此，当基于上述的实施例获取到每个样本音频数据分别对应的合成音频数据后，可以基于每个样本音频数据分别对应的目标音频数据以及每个样本音频数据分别对应的合成音频数据，对原始音色转换模型中的参数的参数值进行调整，从而获取训练完成的音色转换模型。示例性的，基于每个样本音频数据分别对应的目标音频数据分别与对应的合成音频数据，确定重构损失值，并根据该重构损失值，对原始音色转换模型中的参数的参数值进行调整，从而获取训练完成的音色转换模型。

在一种可能的实施方式中，可以针对每个样本音频数据，根据该样本音频数据对应的目标音频数据以及该样本音频数据对应的合成音频数据，确定子重构损失值。根据当前迭代获取到的所有子重构损失值的和，确定重构损失值，并根据该重构损失值，对该原始音色转换模型中的参数的参数值进行调整，从而获取训练完成的音色转换模型。

示例性的，可以通过如下公式根据该样本音频数据对应的目标音频数据以及该样本音频数据对应的合成音频数据，确定子重构损失值：

$$\text{recon\_loss}_k = ||\text{gen\_audio}_k - \text{target\_audio}_k||_1$$

其中，recon_loss_k表示第 k 个样本音频数据对应的子重构损失值，$\text{gen}_{\text{audio}_k}$表示第 k 个样本音频数据对应的合成音频数据，$\text{target}_{\text{audio}_k}$表示第 k 个样本音频数据对应的目标音频数据，$\|\cdot\|_1$表示 1 的范数运算符。

考虑到音色转换模型获取音频数据中的语义信息的精度也会影响到音色转换的效果。因此，在本申请实施例中，在对音色转换模型的训练过程中，对该音色转换模型提取到的语义特征的精确度进行监督。

在一种可能的实施方式中，可以预先对样本集中的每个样本音频数据的语义信息进行标注，以根据每个样本音频数据分别对应的标注语义以及每个样本音频数据分别对应的第一语义特征之间的差异，对该音色转换模型提取到的语义特征的精确度进行监督。

在另一种可能的实施方式中，如果要获取到精度较高的音色转换模型，则需要大量的样本音频数据，对每个样本音频数据进行标注，会带来大量的工作量，耗费大量的成本，进而增加获取音色转换模型的难度。因此，在本申请实施例中，可以通过在音色转换模型中添加一个语义提取网络，以通过该语义提取网络与音色转换模型中的去除音色网络对抗学习音频数据中的语义信息。在对原始音色转换模型进行训练的过程中，对于样本集中的任一样本音频数据，通过该原始音色转换模型中的语义提取网络，基于输入的样本音频数据，获取该样本音频数据中的语义特征（记为第二语义特征）。其中，该第二语义特征也为样本音频数据中与发音人音色无关，且与语义信息有关的特征。

在一种示例中，该语义提取网络中可以包括第一内容子网络、第二内容子网络以及第三内容子网络，以实现准确地获取样本音频数据中去除音色信息的语义信息。其中，该第一内容子网络与该第二内容子网络连接，该第一内容子网络配置为对样本音频数据进行处理，获取该样本音频数据中比较密集的内容特征，该第一内容子网络可以是 h-net 网络等。该第二内容子网络与第三内容自网络子网络连接，该第二内容子网络配置为对第一内容子网络输出的内容特征进行处理，获取离散化的内容特征，删除该内容特征中非必要的细节，使离散化的内容特征与样本音频数据的语义信息相关，比如，该第二内容子网络可以是向量化

层（Vector Quantization，VQ）等。该第三内容子网络配置为对第二内容子网络输出的离散化内容特征进行处理，激励该离散化内容特征学习样本音频数据中与语义信息有关的局部特征，从而获取样本音频数据的第二语义特征，比如，该第三内容子网络可以是 g-net 网络、基于循环神经网络（Recurrent Neural Network，RNN）的对比预测编码（Contrastive Predictive Coding，CPC）网络等。

具体实施过程中，当基于上述的实施例将任一样本音频数据输入到原始音色转换模型后，通过该原始音色转换模型中的语义提取网络所包含的第一内容子网络，基于该样本音频数据，获取内容特征；通过该语义提取网络中的第二内容子网络，基于该内容特征，获取离散化内容特征；通过语义提取网络中的第三内容子网络，基于该离散化内容特征，获取该样本音频数据的第二语义特征。

图 2 为本申请实施例提供的一种音色转换模型的结构示意图。下面结合图 2 对本申请实施例提供的一种音色转换模型训练方法进行说明。

获取该样本集中的任一样本音频数据的梅尔频谱，将该样本音频数据的梅尔频谱输入到原始音色转换模型。通过该原始音色转换模型中的音色提取网络（speaker encoder），对该样本音频数据的梅尔频谱进行相应的处理，获取该样本音频数据对应的第一音色特征（tone_vector）。通过该原始音色转换模型中去除音色网络所包含的后验编码器（posterior encoder），可以基于该第一音色特征以及样本音频数据对应的线性频谱（linear spectrogrm），获取样本音频数据中语义信息的隐向量（z_{sq}）。通过该原始音色转换模型中去除音色网络所包含的增强子网络（flow），基于该隐向量，获取增强后的隐向量，即确定第一语义特征。同时通过该原始音色转换模型中的语义提取网络（VQCPC net），基于输入的样本音频数据，获取该样本音频数据中的第二语义特征。然后通过图 2 所示的声码器（decoder），基于该第一语义特征以及样本音频数据对应的目标音频数据的第二音色特征（speaker inner embedding），获取合成音频数据（raw waveform）。

具体的，通过该原始音色转换模型中的语义提取网络，基于输入的样本音频数据，获取该样本音频数据中的第二语义特征的过程包括：通过该原始音色转换模型中的语义提取网络所包含的第一内容子网络（h-net），基于该样本音频数据的梅尔频谱，获取内容特征 Z，如图 2 所示的第 k 个样本音频数据的内容特征 Z 包括第 k 个样本音频数据所包含的每个音频帧的内容数据，例如，第 k 个样本音频数据所包含的第 n 个音频帧的内容数据为$z_{k,n}$。通过该语义提取网络中的第二内容子网络（Vector Quantization（VQ）），基于该内容特征 Z，获取离散化内容特征$\hat{Z}$，如图 2 所示的第 k 个样本音频数据的离散化内容特征$\hat{Z}$包括第 k 个样本音频数据所包含的每个音频帧的离散化内容数据，例如，第 k 个样本音频数据所包含的第 n 个音频帧的离散化内容数据为$\hat{z}_{k,n-1}$。通过语义提取网络中的第三内容子网络（g-net），基于该离散化内容特征$\hat{Z}$，获取该样本音频数据的第二语义特征 R，如图 2 所示的第 k 个样本音频数据的第二语义特征 R 包括第 k 个样本音频数据所包含的每个音频帧的第二语义特征，例如，第 k 个样本音频数据所包含的第 n 个音频帧的离散化内容数据为$r_{k,n}$。

需要说明的是，该 h-net 网络中可以包括卷积层，规范层、线性变换层、逻辑函数层。图 2 所示的 h-net 网络中包括一个卷积层（conv layer）与 4 个相同连接结构的子网络顺序连接，对于任一子网络，该子网络中包括 1 层规范层（layer normalization）、2 层线性变换层（linear layer）以及 1 层逻辑函数（ReLU）层。

当基于上述实施例获取到每个样本音频数据的第二语义特征后，可以基于每个样本音频数据分别对应的目标音频数据及每个样本音频数据分别对应的合成音频数据，以及每个样本音频数据分别对应的第一语义特征及每个样本音频数据分别对应的第二语义特征，对原始音色转换模型进行训练，以获取训练完成的音色转换模型。也就是说，基于每个样本音频数据分别对应的目标音频数据分别与对应的合成音频数据之间的差异，以及每个样本音频数据分别对应的第一语义特征分别与对应的第二语义特征之间的差异，对该

原始音色转换模型中的参数的参数值进行调整，以获取训练完成的音色转换模型，从而实现无监督地对音色转换模型提取语义信息的能力进行训练。

具体实施过程中，基于每个样本音频数据分别对应的目标音频数据及每个样本音频数据分别对应的合成音频数据，确定重构损失值；并基于每个样本音频数据分别对应的第一语义特征及每个样本音频数据分别对应的第二语义特征，确定语义损失值。根据该重构损失值以及语义损失值，确定综合损失值。根据该综合损失值，对该原始音色转换模型中的参数的参数值进行调整，以获取训练完成的音色转换模型。

示例性的，针对样本集中的每个样本音频数据，根据该样本音频数据对应的目标音频数据与对应的合成音频数据之间的差异，确定子重构损失值，并根据该样本音频数据的第一语义特征与第二语义特征之间的差异，确定子语义损失值。根据当前迭代确定的所有重构损失值的和，确定重构损失值；并根据当前迭代确定的所有子语义损失值的和，确定语义损失值。根据该重构损失值以及语义损失值，确定综合损失值。根据该综合损失值，对该原始音色转换模型中的参数的参数值进行调整，以获取训练完成的音色转换模型。

在一种示例中，在根据该重构损失值以及语义损失值，确定综合损失值时，可以根据重构损失值及其对应的第一权重值，以及语义损失值及其对应的第二权重值，确定综合损失值。例如，获取重构损失值与对应的第一权重值的乘积（记为第一乘积），以及语义损失值与对应的第二权重值的乘积（记为第二乘积）。根据第一乘积与第二乘积的和，确定综合损失值。

在一种可能的实施方式中，在通过原始音色转换模型中去除音色网络所包含的后验编码器，基于样本音频数据对应的线性频谱以及原始音色转换模型中音色提取网络所确定的第一音色特征，获取到隐向量的同时，还可以获取到该隐向量的均值向量以及方差向量。也就是说，通过原始音色转换模型中去除音色网络所包含的后验编码器，基于样本音频数据对应的线性频谱以及原始音色转换模型中音色提取网络所确定的第一音色特征，还可以获取到隐向量的均值向量以及方差向量。后续基于每个样本音频数据分别对应的第一语义特征以及每个样本音频数据分别对应的第二语义特征，确定语义损失值时，可以基于每个样本音频数据分别对应的第一语义特征、所述第二语义特征、均值向量以及方差向量，确定语义损失值。也就是说，可以针对每个样本音频数据，根据该样本音频数据的第一语义特征、第二语义特征、均值向量以及方差向量，确定子语义损失值。根据当前迭代确定的所有子语义损失值的和，确定语义损失值。

示例性的，可以通过如下公式根据该样本音频数据的第一语义特征、第二语义特征、均值向量以及方差向量，确定子语义损失值：

$$KL\_loss_k = log_{sq_k} - log_{sp_k} - 0.5 + 0.5 * (z_{p_k} - m_{p_k}) ** 2 + \exp(-2 * logs_{p_k})$$

其中，KL_loss 表示第 k 个样本音频数据对应的子语义损失值，log_{sq_k}表示第 k 个样本音频数据对应的隐向量的方差向量的 log 值，log_{sp_k}是第 k 个样本音频数据的第二语义特征的方差向量的 log 值，z_{p_k}表示第 k 个样本音频数据的第一语义特征，m_{p_k}表示第 k 个样本音频数据对应的隐向量的均值向量。

在一种可能的实施方式中，若原始音色转换模型中的语义提取网络包括第一内容子网络、第二内容子网络以及第三内容子网络，则在对原始音色转换模型进行训练的过程中，也需要考虑该语义提取网络中存在损失值。因此，在本申请实施例中，可以基于每个样本音频数据分别对应的内容特征及每个样本音频数据分别对应的离散化内容特征，确定量化损失值；并基于每个样本音频数据分别对应的离散化内容特征及每个样本音频数据分别对应的第二语义特征，确定对比学习损失值，以方便后续基于每个样本音频数据分别对应的目标音频数据及每个样本音频数据分别对应的合成音频数据，每个样本音频数据分别对应的第一语义特征及每个样本音频数据分别对应的第二语义特征，该量化损失值以及对比学习损失值，对该原始音色转换模型中的参数的参数值进行调整，以获取训练完成的音色转换模型。

在一种示例中，在确定量化损失值时，可以针对每个样本音频数据，根据该样本音频数据包含的每个

音频帧所对应的内容特征及离散化内容特征，确定该样本音频数据对应的子量化损失值。根据当前迭代确定的所有子量化损失值的和，确定量化损失值。

例如，可以通过如下公式基于每个样本音频数据分别对应的内容特征及每个样本音频数据分别对应的离散化内容特征，确定量化损失值：

$$VQ\_loss = \frac{1}{KT}\sum_{k=1}^{K}\sum_{n=1}^{N/2}\left\|z_{k,n} - sg(\hat{z}_{k,n})\right\|_2^2$$

其中，VQ_loss表示量化损失值，K 表示样本集包含的样本音频数据的总数量，N 表示从每个样本音频数据中分别包含的音频帧的总数量，n 表示当前第 k 个样本音频数据中包含的第 n 个音频帧， $z_{k,n}$表示第 k 个样本音频数据中包含的第 n 个音频帧所对应的内容特征，$\hat{z}_{k,n}$表示第 k 个样本音频数据中包含的第 n 个音频帧所对应的离散化内容特征，sg(.)表示停止梯度运算符，$\|\cdot\|_2$表示 2 的范数运算符。

在一种示例中，在确定对抗学习损失值时，可以针对每个样本音频数据，可以根据该样本音频数据包含的每个音频帧所对应的离散化内容特征以及第二语义特征，确定该样本音频数据对应的子对比学习损失值。根据当前迭代确定的所有子对比学习损失值的和，确定对比学习损失值。

例如，可以通过如下公式基于每个样本音频数据分别对应的离散化内容特征及每个样本音频数据分别对应的第二语义特征，确定对比学习损失值：

$$CPC\_loss = -\frac{1}{KN'M}\sum_{k=1}^{K}\sum_{n=1}^{N'}\sum_{m=1}^{M}\log\left[\frac{\exp\left(\hat{z}_{k,n+m}^{T} w_m r_{k,n}\right)}{\sum_{\tilde{z}\in\Omega_{k,n,m}}\exp\left(\tilde{z}^{T} w_m r_{k,n}\right)}\right]$$

其中，CPC_loss表示对比学习损失值，K 表示样本集包含的样本音频数据的总数量， $N' = \frac{N}{2} - M$，N 表示从每个样本音频数据中分别包含的音频帧的数量，M 为任一样本音频数据中包含的正样本音频帧的总数量，n 表示当前第 k 个样本音频数据中第 n 个音频帧， $\hat{z}_{k,n+m}$表示第 k 个样本音频数据中第 n+m 个音频帧所对应的离散化内容特征，w_m表示关联帧数为 m 所对应的权重矩阵，$r_{k,n}$表示第 k 个样本音频数据中包含的第 n 个音频帧所对应的第二语义特征，T 为转置符，$\tilde{z}$表示从第 k 个样本音频数据中包含的除正样本音频帧之外的任一负样本音频帧，$\Omega_{k,n,m}$表示包含有第 k 个样本音频数据中包含的负样本音频帧的集合。

具体实施过程中，在对该原始音色转换模型进行训练时，针对每个样本音频数据，根据该样本音频数据对应的目标音频数据与对应的合成音频数据之间的差异，确定子重构损失值；根据该样本音频数据的第一语义特征与第二语义特征之间的差异，确定子语义损失值；根据该样本音频数据包含的每个音频帧所对应的内容特征及离散化内容特征，确定该样本音频数据对应的子量化损失值；并根据该样本音频数据包含的每个音频帧所对应的离散化内容特征以及第二语义特征，确定该样本音频数据对应的子对比学习损失值。根据当前迭代确定的所有子重构损失值的和，确定重构损失值；根据当前迭代确定的所有子语义损失值的和，确定语义损失值；根据当前迭代确定的所有子对比学习损失值的和，确定对比学习损失值；根据当前迭代确定的所有子量化损失值的和，确定量化损失值。根据确定的重构损失值、语义损失值、对比学习损失值以及量化损失值，确定综合损失值。根据该综合损失值，对该原始音色转换模型中的参数的参数值进行调整，以获取训练完成的音色转换模型。

示例性的，可以根据确定的重构损失值及其对应的第一权重值、语义损失值及其对应的第二权重值、对比学习损失值及其对应的第三权重值以及量化损失值及其对应的第四权重值，确定综合损失值。

考虑到合成音频数据的音色转换效果也会受声码器的精度的影响。因此，在本申请实施例中，在对原始音色转换模型进行训练的过程中，还会考虑该原始音色转换模型中的声码器中可能存在的损失值，以根据声码器的损失值，对音色转换模型进行训练。也就是说，在对原始音色转换模型进行训练的过程中，基

于每个样本音频数据分别对应的目标音频数据及合成音频数据，以及该原始音色转换模型中的声码器的损失值，对原始音色转换模型进行训练，以获取训练完成的音色转换模型。

示例性的，以声码器为 HiFiGAN 声码器为例，HiFi-GAN 声码器是一种深度神经网络模型，采用端到端的前向网络结构，训练多尺度的判别器，能够实现高效的、高质量的语音合成。HiFiGAN 声码器包含有生成器以及判别器，该判别器有两种，分别是多尺度判别器和多周期判别器，以从两种不同角度分别鉴定 HiFiGAN 声码器中的生成器所生成的音频数据。HiFiGAN 声码器将特征匹配损失作为训练生成器的额外损失，通过提取判别器每个中间特征，计算每个特征空间中目标音频数据和合成音频数据之间的距离 L1，从而稳定 GAN。因此，该 HiFiGAN 声码器的损失值包括特征匹配损失值，以根据该特征匹配损失值，对该原始音色转换模型进行训练。

示例性的，该特征匹配损失值可通过如下公式确定：

$$L_{FM}(G;D) = E_{x,s}[\sum_{j=1}^{J}\frac{1}{Q_j}\left\|D^j(x) - D^i(G(s))\right\|_1$$

其中，J 表示声码器所包含的判别器中提取特征的层数，$D^j()$表示判别器中第 j 个提取特征层提取到的特征，Q_j表示判别器中第 j 个提取特征层提取到的特征的数量，x 为目标音频数据，s 为生成器生成的合成音频数据的梅尔频谱。

由于 HiFiGAN 声码器的本质仍然是一个生成对抗网络，HiFiGAN 声码器中的判别器计算合成音频数据是目标音频数据的概率，HiFiGAN 声码器中的生成器配置为合成音频数据，在对 HiFiGAN 声码器进行训练的过程中，希望该 HiFiGAN 声码器中的生成器可以合成接近目标音频数据的合成音频数据，以致于 HiFiGAN 声码器中的判别器无法区分是目标音频数据还是合成音频数据。基于此，该 HiFiGAN 声码器的损失值还包括生成对抗损失值。例如，可以根据目标音频数据以及 HiFiGAN 声码器中的生成器生成的合成音频数据的梅尔频谱，确定生成对抗损失值。

示例性的，可以通过如下公式生成对抗损失值：

$$L_{adv}(D;G) = E_{x,s}[(D(x)-1)^2 + D(G(s))^2]$$

$$L_{adv}(G;D) = E_s[\left(D\left(G(s)\right)-1\right)^2]$$

其中，$L_{adv}(D;G)$表示 HiFiGAN 声码器中的判别器的生成对抗损失值，$L_{adv}(G;D)$表示 HiFiGAN 声码器中的生成器的生成对抗损失值，x 为目标音频数据，s 为生成器生成的合成音频数据的梅尔频谱。

基于上述的实施例获取到声码器的损失值后，即可根据该声码器的损失值以及每个样本音频数据分别对应的目标音频数据及合成音频数据，对该原始音色转换模型中的各个参数的参数值进行调整，以获取训练完成的音色转换模型。

例如，基于每个样本音频数据分别对应的目标音频数据及合成音频数据，确定重构损失值，并确定原始音色转换模型中的声码器的损失值。根据该重构损失值以及声码器的损失值，确定综合损失值。根据该综合损失值，对该原始音色转换模型中的各个参数的参数值进行调整。

其中，在根据该重构损失值以及声码器的损失值，确定综合损失值时，可以根据该重构损失值及其对应的第一权重值、以及声码器的损失值及其对应的第五损失值，确定综合损失值。

再例如，基于每个样本音频数据分别对应的目标音频数据及合成音频数据，确定重构损失值；基于每个样本音频数据分别对应的第一语义特征以及每个样本音频数据分别对应的第二语义特征，确定语义损失值；基于每个样本音频数据分别对应的内容特征及每个样本音频数据分别对应的离散化内容特征，确定量化损失值；基于每个样本音频数据分别对应的离散化内容特征及每个样本音频数据分别对应的第二语义特征，确定对比学习损失值；确定原始音色转换模型中的声码器的损失值。根据该重构损失值、语义损失值、

量化损失值、对比学习损失值以及声码器的损失值，确定综合损失值。根据该综合损失值，对该原始音色转换模型中的各个参数的参数值进行调整。

示例性的，根据该重构损失值、语义损失值、量化损失值、对比学习损失值以及声码器的损失值，确定综合损失值可通过如下公式表示：

$$total\_loss = recon\_loss + KL\_loss + vq\_loss + cpc\_loss + decoder\_loss$$

其中，total_loss 表示综合损失值，recon_loss 表示重构损失值，KL_loss 表示语义损失值，vq_loss 表示量化损失值，cpc_loss 表示对比学习损失值，decoder_loss 表示声码器的损失值。

需要说明的是，若声码器为 HiFiGAN 声码器，则 decoder_loss 可以包括特征匹配损失值（fm_loss）和生成对抗损失值（adv_loss）。

由于包含有若干个配置为训练原始音色转换模型的样本音频数据，针对每个样本音频数据，均执行上述的步骤，直到满足预设的收敛条件。

其中，满足预设的收敛条件可以为当前迭代所确定的综合损失值小于预设的损失阈值，或对原始音色转换模型进行训练的迭代次数达到预先设置的最大迭代次数等。具体实施中可以灵活进行设置，在此不做具体限定。

在一种可能的实施方式中，在对原始音色转换模型训练时，把样本音频数据分为训练样本和测试样本，先基于训练样本对原始音色转换模型进行训练，再基于测试样本对上述已训练的音色转换模型的可靠程度进行验证。

在获取到训练完成的音色转换模型之后，可以基于该音色转换模型以及样本集中的每个样本音频数据，确定样本集中的每个发音人分别对应的音色特征。然后对应保存每个发音人分别对应的音色特征以及每个发音人分别对应的对象标识，这样后续在通过该训练完成的音色转换模型，合成该样本集中任一发音人的合成音频数据时，可以直接根据对象标识与音色特征的对应关系，确定该发音人的对象标识的音色特征，有利于根据该发音人的音色特征以及音色转换模型，获取该发音人的合成音频数据，提高音色转换的效率。可以理解的是，该样本集中任一样本音频数据的发音人的音色，均是该训练完成的音色转换模型所支持的音色。

实施例 2：

本申请实施例还提供了一种音色转换方法，图 3 为本申请实施例提供了一种音色转换过程示意图，该过程包括：

S301：获取源音频数据以及目标发音人的音色特征。

本申请实施例提供的音色转换方法应用于电子设备，该电子设备可以为如机器人等智能设备，也可以为服务器。其中，本申请实施例中进行音色转换的电子设备可以与上述进行音色转换模型训练的电子设备相同，也可以不同。

在一种可能的实施方式中，由于在进行音色转换模型训练的过程中，一般采用离线的方式，进行音色转换模型训练的，因此，当获取到训练完成的音色转换模型后，可以将该音色转换模型部署到进行音色转换的电子设备，以方便进行音色转换的电子设备可以通过该音色转换模型进行音色转换。

需要说明的是，具体训练音色转换模型的过程已在上述实施例中进行描述，重复之处不做赘述。

在一种可能的实施方式中，若训练完成的音色转换模型中包括语义提取网络，由于该语义提取网络主要用于对去除音色网络提取到的语义特征进行监督的，因此，在将训练完成的音色转换模型部署到进行音色转换的电子设备时，可以将该音色转换模型中除语义提取网络之外的其它网络部署到进行音色转换的电子设备，以减少该音色转换模型所包含的参数量，降低数据传输所耗费的成本以及进行音色转换的电子设

备的内存空间的压力。

当需要进行音色转换时，用户可以通过在智能设备上输入合成请求，以通过该合成请求可以控制智能设备合成目标发音人发出某一发音内容的音频数据。其中，具体输入合成请求的方式有很多，比如，输入合成请求的方式可以是通过输入语音信息的方式输入，也可以对智能设备的显示屏上显示的虚拟按钮进行操作的方式输入等，具体实施过程中可以根据需求进行灵活设置，在此不做具体限定。当智能设备获取到合成请求后，可以将该合成请求、源音频数据以及目标发音人的信息发送至进行音色转换的电子设备。

其中，目标发音人指的是通过音色转换技术获取到的合成音频数据的音色所归属的发音人，该目标发音人的信息包括目标发音人的对象标识，或，目标发音人的音频数据。该源音频数据指的是配置为提供语义信息，且不提供音色信息的音频数据。

需要说明的是，该源音频数据以及目标发音人的音频数据均可以是用户通过智能设备录制的，也可以是预先配置在智能设备中的音频数据。

在一种可能的应用场景中，若用户将样本集中任一样本音频数据的发音人确定为目标发音人，则该目标发音人的信息包括目标发音人的对象标识。例如，智能设备输出有多个发音人的对象标识，用户可以对智能设备输出的多个对象标识进行选择，若智能设备检测到用户选择的对象标识，则可以将选择的对象标识确定为目标发音人的信息，后续智能设备可以将合成请求、该目标发音人的对象标识以及该源音频数据发送至进行音色转换的电子设备。

其中，该对象标识可以通过数字、字符串等形式表示，也可以通过其他形式表示，只要可以唯一标识该发音人的形式均可应配置为本申请实施例。

需要说明的是，该样本集中的样本音频数据配置为训练音色转换模型。可以理解的是，该样本集中任一样本音频数据的发音人的音色，均是该训练完成的音色转换模型所支持的音色。

在另一种可能的应用场景中，若用户将样本集中之外的其他发音人确定为目标发音人，即不将样本集中任一样本音频数据的发音人确定为目标发音人，则该目标发音人的信息包括目标发音人的音频数据。例如，智能设备输出有多个发音人的对象标识，若智能设备未检测到用户选择的对象标识，或检测到用户输入的添加目标发音人的触发操作，则可以提示用户输入目标发音人的音频数据，并将用户输入的音频数据确定为目标发音人的信息。后续智能设备可以将合成请求、该目标发音人的音频数据以及该源音频数据发送至进行音色转换的电子设备。

进行音色转换的电子设备接收到该合成请求、目标发音人的信息以及源音频数据后，可以基于该目标发音人的信息，确定目标发音人的音色特征。后续基于该目标发音人的音色特征以及源音频数据，获取目标发音人发出源音频数据的发音内容的音频数据，即获取目标发音人的合成音频数据。

在一种可能的实施方式中，获取所述目标发音人的音色特征，包括：

获取所述目标发音人的信息；

若确定所述目标发音人的信息为对象标识，则根据保存的对象标识与音色特征的对应关系，确定所述目标发音人的对象标识所对应的音色特征；

若确定所述目标发音人的信息为音频数据，则通过所述音色转换模型中的音色提取网络，获取所述音频数据的音色特征。

在本申请实施例中，基于该目标发音人的信息，确定目标发音人的音色特征包括如下两种情况：

情况一、若该目标发音人的信息为目标发音人的对象标识，说明该目标发音人为样本集中任一样本音频数据的发音人，则根据保存的对象标识与音色特征的对应关系，确定目标发音人的对象标识所对应的音色特征，将确定的音色特征确定为该目标发音人的音色特征。

其中，音色转换模型是基于所述样本集中的样本音频数据训练得到的。

情况二、若该目标发音人的信息为音频数据，说明无法确定该目标发音人是否为样本集中任一样本音频数据的发音人，则通过音色转换模型中的音色提取网络，获取该音频数据的音色特征，将获取到的音色特征确定为该目标发音人的音色特征。

S302：通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

当基于上述的实施例获取到源音频数据以及目标发音人的音色特征后，可以将该源音频数据以及目标发音人的音色特征输入到预先训练的音色转换模型中，以通过该音色转换模型，基于源音频数据以及目标发音人的音色特征，确定合成音频数据。

具体实施过程中，通过预先训练的音色转换模型中的音色提取网络，获取源音频数据的音色特征。通过音色转换模型中的去除音色网络，基于音色特征以及源音频数据对应的线性频谱，获取语义特征；其中，该语义特征为源音频数据中与发音人音色无关，且与语义信息有关的特征；通过音色转换模型中的声码器，基于语义特征以及目标发音人的音色特征，获取合成音频数据。

其中，获取源音频数据对应的线性频谱的过程已在上述实施例进行描述，在此不做具体限定。

在一种可能的实施方式中，在将源音频数据输入到预先训练的音色转换模型之前，可以获取该源音频数据的梅尔倒谱，将该梅尔倒谱代替该源音频数据输入到音色转换模型中，以降低音色转换模型所需的计算量，提高音色转换效率，方便音色转换模型进行音色转换。

图 4 为本申请实施例提供的一种音色转换模型的结构示意图。下面结合图 4 对本申请实施例提供的一种音色转换方法进行说明。

首先，获取源音频数据以及目标发音人的信息。

若该目标发音人的信息为目标发音人的对象标识，说明该目标发音人为样本集中任一样本音频数据的发音人，当前希望获取到目标发音人发出任一音频数据中的发音内容的音频数据，则根据保存的对象标识与音色特征的对应关系，确定目标发音人的对象标识所对应的音色特征，将确定的音色特征确定为该目标发音人的音色特征。通过该种目标发音人的信息，后续可以实现 any-to-many 的音色转换方式，即选择样本集中任一样本音频数据的发音人的音色特征，通过预先训练的音色转换模型，基于该音色特征以及任意发音内容的源音频数据，即可获取到该发音人发出该发音内容的音频数据。

若该目标发音人的信息为音频数据，说明无法确定该目标发音人是否为样本集中任一样本音频数据的发音人，当前希望获取到任一目标发音人发出任一音频数据中的发音内容的音频数据，则通过音色转换模型中的音色提取网络，获取该音频数据的音色特征，将获取到的音色特征确定为该目标发音人的音色特征。通过该种目标发音人的信息，后续可以实现 many-to-many 的音色转换方式，即通过预先训练的音色转换模型，基于任意音频数据的音色特征以及任意发音内容的源音频数据，即可获取到满足该音色特征以及发音内容的合成音频数据。

将该源音频数据（Reference-wav）以及目标发音人的音色特征输入到预先训练的音色转换模型。

通过该音色转换模型中的音色提取网络（如图 4 所示的 speaker encoder），对该源音频数据的梅尔频谱（Mel spectrogrm）进行相应的处理，获取该源音频数据对应的音色特征，如图 4 所示的 tone_vector。

通过该音色转换模型中去除音色网络所包含的后验编码器，如图 4 所示的 posterior encoder，可以基于该源音频数据对应的音色特征以及源音频数据对应的线性频谱，如图 4 所示的 linear spectrogrm，获取源音

频数据中语义信息的隐向量，如图 4 所示的z_{sq}。

通过该音色转换模型中去除音色网络所包含的增强子网络（flow），基于该隐向量，获取增强后的隐向量，即确定语义特征。

通过图 4 所示的声码器(decoder)，基于该语义特征以及目标发言人的音色特征，如图 4 所示的 speaker inner embedding，获取合成音频数据，即获取图 4 所示的 raw waveform。

实施例 3：

本申请实施例提供了一种音色转换模型训练装置，图 5 为本申请实施例提供的一种音色转换模型训练装置的结构示意图，该结构包括：

获取单元 51，配置为获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

处理单元 52，配置为对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

训练单元 53，配置为基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

由于上述音色转换模型训练装置解决问题的原理与音色转换模型训练装置方法相似，因此上述音色转换模型训练装置的实施以及对应的有益效果可以参见方法的实施和有益效果，重复之处不再赘述。

实施例 4：

本申请实施例提供了一种音色转换装置，图 6 为本申请实施例提供的一种音色转换装置的结构示意图，该结构包括：

获取模块 61，配置为获取源音频数据以及目标发音人的音色特征；

合成模块 62，配置为通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

由于上述音色转换装置解决问题的原理与音色转换装置方法相似，因此上述音色转换装置的实施可以参见方法的实施，重复之处不再赘述。

实施例 5：

图 7 为本申请实施例提供的一种电子设备结构示意图，该电子设备，包括：处理器 71、通信接口 72、存储器 73 和通信总线 74，其中，处理器 71，通信接口 72，存储器 73 通过通信总线 74 完成相互间的通信；

所述存储器 73 中存储有计算机程序，当所述程序被所述处理器 71 执行时，使得所述处理器 71 执行如下步骤：

获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数

据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

由于上述电子设备解决问题的原理与音色转换模型训练方法相似，因此上述电子设备的实施可以参见方法的实施例 1，重复之处不再赘述，相应的有益效果部分相同，此处不再赘述。

实施例 6：

图 8 为本申请实施例提供的再一种电子设备结构示意图，该电子设备，包括：处理器 81、通信接口 82、存储器 83 和通信总线 84，其中，处理器 81，通信接口 82，存储器 83 通过通信总线 84 完成相互间的通信；

所述存储器 83 中存储有计算机程序，当所述程序被所述处理器 81 执行时，使得所述处理器 81 执行如下步骤：

获取源音频数据以及目标发音人的音色特征；

通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

由于上述电子设备解决问题的原理与音色转换方法相似，因此上述电子设备的实施可以参见方法的实施例 2，重复之处不再赘述。

上述电子设备提到的通信总线可以是外设部件互连标准（Peripheral Component Interconnect，PCI）总线或扩展工业标准结构（Extended Industry Standard Architecture，EISA）总线等。该通信总线可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。通信接口 82 配置为上述电子设备与其他设备之间的通信。存储器可以包括随机存取存储器（Random Access Memory，RAM），也可以包括非易失性存储器（Non-Volatile Memory，NVM），例如至少一个磁盘存储器。可选地，存储器还可以是至少一个位于远离前述处理器的存储装置。

上述处理器可以是通用处理器，包括中央处理器、网络处理器（Network Processor，NP）等；还可以是数字指令处理器（Digital Signal Processing，DSP）、专用集成电路、现场可编程门陈列或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。

实施例 7：

在上述各实施例的基础上，本申请实施例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内存储有可由处理器执行的计算机程序，当所述程序在所述处理器上运行时，使得所述处理器执行时实现如下步骤：

获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述

样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

由于计算机可读存储介质解决问题的原理以及相应得到的有益效果与上述实施例中的音色转换模型训练方法相似，因此具体实施可以参见音色转换模型训练方法的实施。

实施例 8：

在上述各实施例的基础上，本申请实施例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内存储有可由处理器执行的计算机程序，当所述程序在所述处理器上运行时，使得所述处理器执行时实现如下步骤：

获取源音频数据以及目标发音人的音色特征；

通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

由于计算机可读存储介质解决问题的原理与上述实施例中的音色转换方法相似，因此具体实施可以参见音色转换方法的实施。

本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本申请实施例还提供一种计算机程序产品，包括计算机程序，该计算机程序被执行时，可以实现如上述所述音色转换模型训练方法的步骤，或者，实现如上述所述音色转换方法的步骤。

本申请是参照根据本申请的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

权 利 要 求 书

1、一种音色转换模型训练方法，其中，所述方法包括：

获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

2、根据权利要求 1 所述的方法，其中，所述样本音频数据对应的目标音频数据包括以下中的至少一种：所述样本音频数据、与所述样本音频数据的发音人不同的样本音频数据、以及与所述样本音频数据的发音人不同的非样本音频数据。

3、根据权利要求 2 所述的方法，其中，获取所述样本音频数据对应的目标音频数据的第二音色特征，包括：

若所述目标音频数据为所述样本音频数据，则将所述样本音频数据的第一音色特征，确定为所述第二音色特征；

若所述目标音频数据不为所述样本音频数据，则通过所述原始音色转换模型中的音色提取网络，获取所述目标音频数据的第二音色特征。

4、根据权利要求 1 所述的方法，其中，所述通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征，包括：

通过所述去除音色网络中的后验编码器，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取所述样本音频数据中语义信息的隐向量；

通过所述去除音色网络中的增强子网络，基于所述隐向量，获取所述第一语义特征。

5、根据权利要求 4 所述的方法，其中，所述方法还包括：

对于任一所述样本音频数据，通过所述原始音色转换模型中的语义提取网络，基于所述样本音频数据，获取第二语义特征；

所述基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，包括：

基于每个所述样本音频数据分别对应的目标音频数据及每个所述样本音频数据分别对应的合成音频数据，以及每个所述样本音频数据分别对应的第一语义特征及每个所述样本音频数据分别对应的第二语义特征，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

6、根据权利要求 5 所述的方法，其中，所述通过所述原始音色转换模型中的语义提取网络，基于所述样本音频数据，获取第二语义特征，包括：

通过所述语义提取网络中的第一内容子网络，基于所述样本音频数据，获取内容特征；

通过所述语义提取网络中的第二内容子网络，基于所述内容特征，获取离散化内容特征；

通过所述语义提取网络中的第三内容子网络，基于所述离散化内容特征，获取所述第二语义特征。

7、根据权利要求 5 或 6 所述的方法，其中，所述基于每个所述样本音频数据分别对应的目标音频数据及每个所述样本音频数据分别对应的合成音频数据，以及每个所述样本音频数据分别对应的第一语义特

征及每个所述样本音频数据分别对应的第二语义特征，对所述原始音色转换模型进行训练，包括：

基于每个所述样本音频数据分别对应的目标音频数据及每个所述样本音频数据分别对应的合成音频数据，确定重构损失值；

基于每个所述样本音频数据分别对应的第一语义特征以及每个所述样本音频数据分别对应的第二语义特征，确定语义损失值；

根据所述重构损失值以及所述语义损失值，确定综合损失值；

根据所述综合损失值，对所述原始音色转换模型中的参数的参数值进行调整，以获取训练完成的音色转换模型。

8、根据权利要求 1-7 中任一项所述的方法，其中，所述方法还包括：

通过所述去除音色网络中的后验编码器，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取所述隐向量的均值向量以及方差向量；

所述基于每个所述样本音频数据分别对应的第一语义特征以及每个所述样本音频数据分别对应的第二语义特征，确定语义损失值，包括：

基于每个所述样本音频数据分别对应的第一语义特征、第二语义特征、均值向量以及方差向量，确定语义损失值。

9、根据权利要求 6 所述的方法，其中，所述基于每个所述样本音频数据分别对应的目标音频数据及每个所述样本音频数据分别对应的合成音频数据，以及每个所述样本音频数据分别对应的第一语义特征及每个所述样本音频数据分别对应的第二语义特征，对所述原始音色转换模型进行训练，包括：

基于每个所述样本音频数据分别对应的内容特征及每个所述样本音频数据分别对应的离散化内容特征，确定量化损失值；并

基于每个所述样本音频数据分别对应的离散化内容特征及每个所述样本音频数据分别对应的第二语义特征，确定对比学习损失值；

基于每个所述样本音频数据分别对应的目标音频数据及每个所述样本音频数据分别对应的合成音频数据，每个所述样本音频数据分别对应的第一语义特征及每个所述样本音频数据分别对应的第二语义特征，所述量化损失值以及所述对比学习损失值，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

10、根据权利要求 1-9 中任一项所述的方法，其中，所述获取训练完成的音色转换模型之后，所述方法还包括：

基于所述音色转换模型以及样本集中每个所述样本音频数据，确定所述不同发音人分别对应的音色特征；

并对应保存所述不同发音人分别对应的对象标识以及音色特征。

11、一种音色转换方法，其中，所述方法包括：

获取源音频数据以及目标发音人的音色特征；

通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

12、根据权利要求 11 所述的方法，其中，获取所述目标发音人的音色特征，包括：

获取所述目标发音人的信息；

若确定所述目标发音人的信息为对象标识，则根据保存的对象标识与音色特征的对应关系，确定所述目标发音人的对象标识所对应的音色特征；

若确定所述目标发音人的信息为音频数据，则通过所述音色转换模型中的音色提取网络，获取所述音频数据的音色特征。

13、一种音色转换模型训练装置，其中，所述装置包括：

获取单元，配置为获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同；

处理单元，配置为对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据；

训练单元，配置为基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型。

14、一种音色转换装置，其中，所述装置包括：

获取模块，配置为获取源音频数据以及目标发音人的音色特征；

合成模块，配置为通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据。

15、一种电子设备，其中，所述电子设备至少包括处理器和存储器，所述处理器配置为执行存储器中存储的计算机程序时实现如权利要求 1-10 中任一所述音色转换模型训练方法的步骤，或者，实现如权利要求 11-12 中任一所述音色转换方法的步骤。

16、一种计算机可读存储介质，其中，其存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求 1-10 中任一所述音色转换模型训练方法的步骤，或者，实现如权利要求 11-12 中任一所述音色转换方法的步骤。

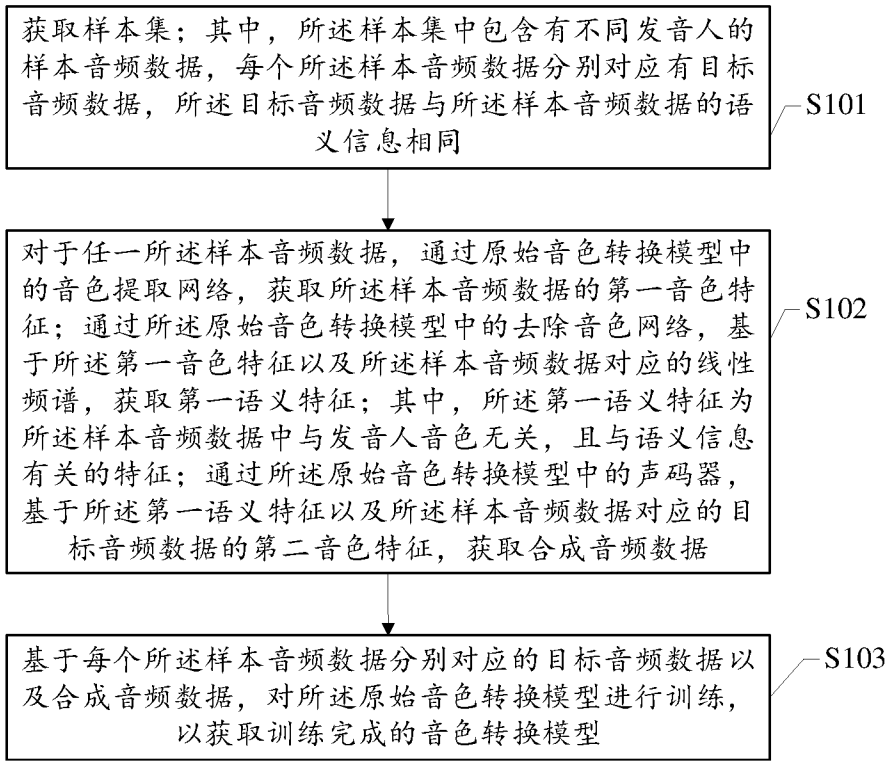
获取样本集；其中，所述样本集中包含有不同发音人的样本音频数据，每个所述样本音频数据分别对应有目标音频数据，所述目标音频数据与所述样本音频数据的语义信息相同
S101
对于任一所述样本音频数据，通过原始音色转换模型中的音色提取网络，获取所述样本音频数据的第一音色特征；通过所述原始音色转换模型中的去除音色网络，基于所述第一音色特征以及所述样本音频数据对应的线性频谱，获取第一语义特征；其中，所述第一语义特征为所述样本音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述原始音色转换模型中的声码器，基于所述第一语义特征以及所述样本音频数据对应的目标音频数据的第二音色特征，获取合成音频数据
S102
基于每个所述样本音频数据分别对应的目标音频数据以及合成音频数据，对所述原始音色转换模型进行训练，以获取训练完成的音色转换模型
S103

图 1

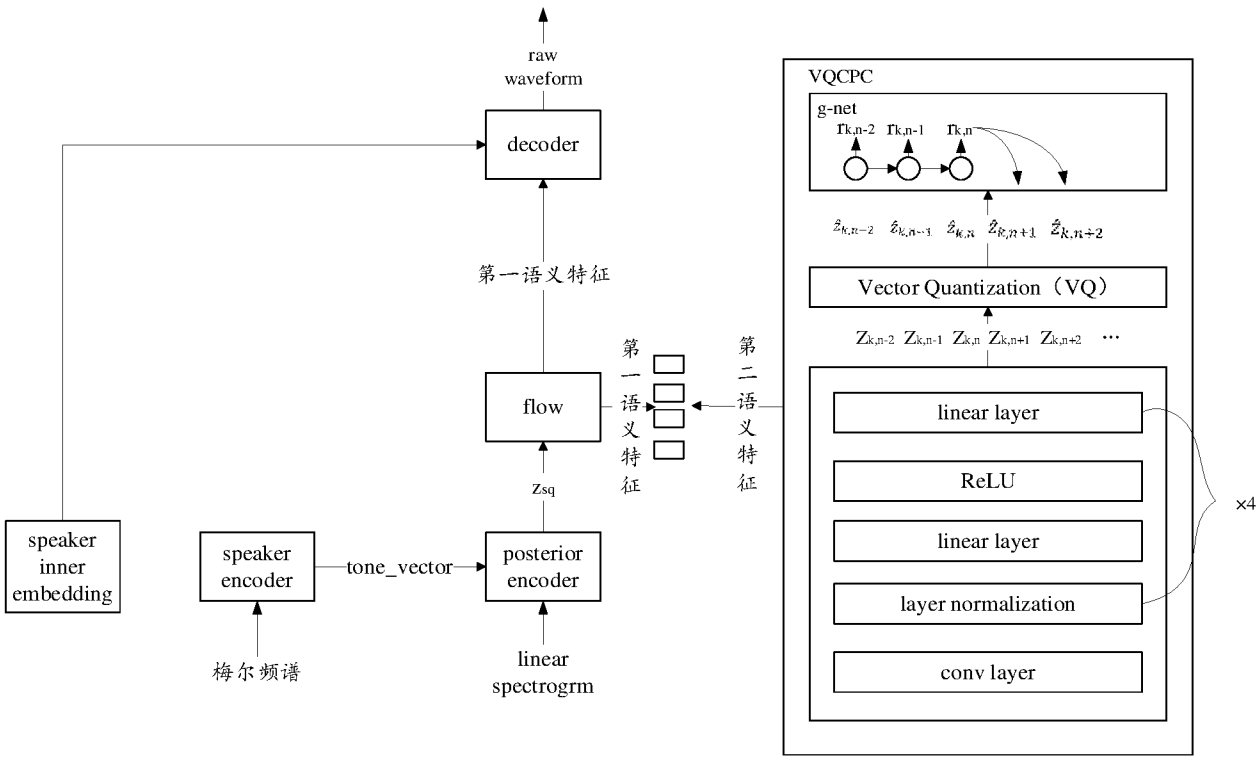
raw waveform
decoder
第一语义特征
flow
第一语义特征
第二语义特征
Zsq
speaker inner embedding
speaker encoder
tone_vector
posterior encoder
梅尔频谱
linear spectrogrm
VQCPC
g-net
Vector Quantization（VQ）
linear layer
ReLU
linear layer
layer normalization
conv layer
×4

图 2

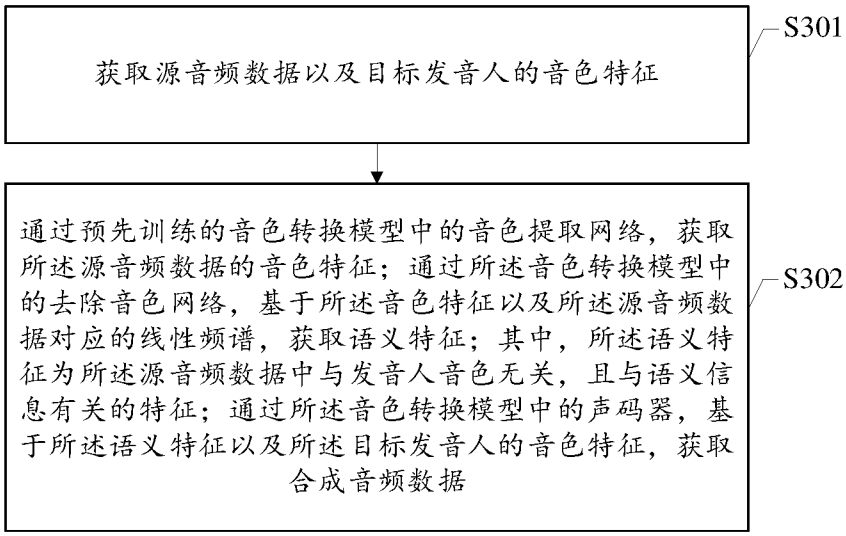
获取源音频数据以及目标发音人的音色特征
S301
通过预先训练的音色转换模型中的音色提取网络，获取所述源音频数据的音色特征；通过所述音色转换模型中的去除音色网络，基于所述音色特征以及所述源音频数据对应的线性频谱，获取语义特征；其中，所述语义特征为所述源音频数据中与发音人音色无关，且与语义信息有关的特征；通过所述音色转换模型中的声码器，基于所述语义特征以及所述目标发音人的音色特征，获取合成音频数据
S302

图 3

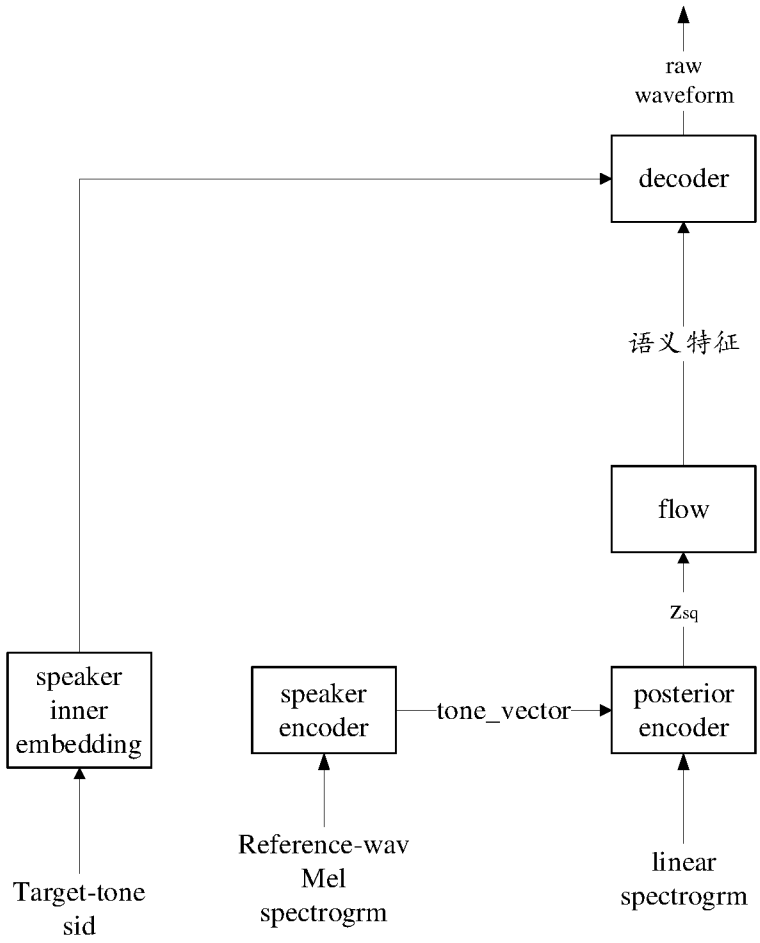
raw waveform
decoder
语义特征
flow
Zsq
speaker inner embedding
speaker encoder
tone_vector
posterior encoder
Target-tone sid
Reference-wav Mel spectrogrm
linear spectrogrm

图 4

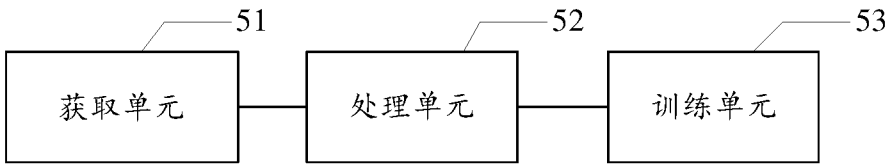
51
52
53
获取单元
处理单元
训练单元

图 5

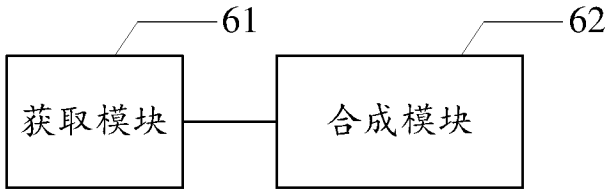

图 6

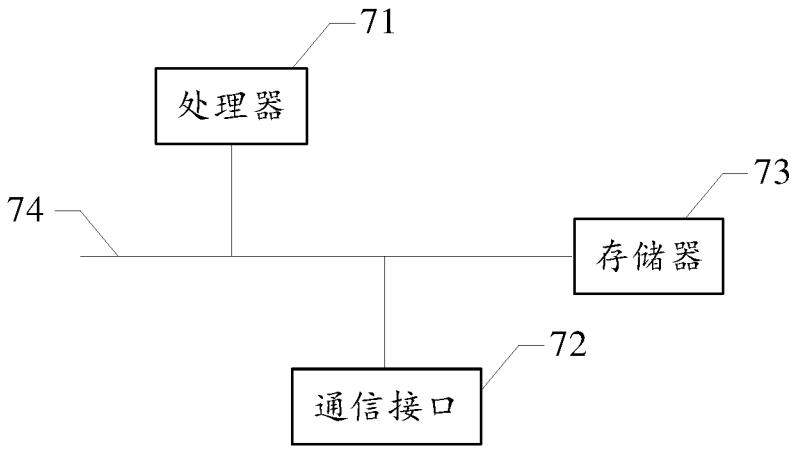

图 7

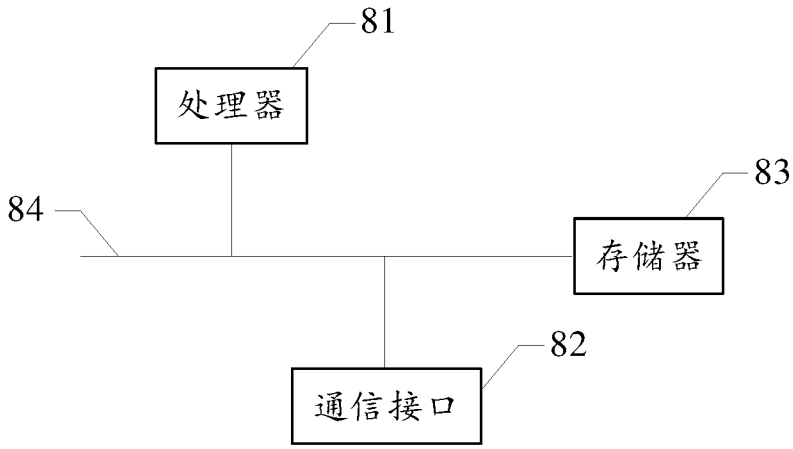

图 8