



(12) 发明专利申请

(10) 申请公布号 CN 105206257 A

(43) 申请公布日 2015. 12. 30

(21) 申请号 201510673278. X

(22) 申请日 2015. 10. 14

(71) 申请人 科大讯飞股份有限公司

地址 230088 安徽省合肥市高新区望江
西路 666 号

(72) 发明人 陈凌辉 江源 李栋梁 李啸
张卫庆 胡国平

(74) 专利代理机构 北京维澳专利代理有限公司
11252

代理人 刘路尧 逢京喜

(51) Int. Cl.

G10L 13/02(2013. 01)

G10L 13/08(2013. 01)

G10L 17/06(2013. 01)

G10L 17/14(2013. 01)

权利要求书2页 说明书9页 附图2页

(54) 发明名称

一种声音转换方法及装置

(57) 摘要

本发明公开了一种声音转换方法及装置，该方法包括：接收待转换语音数据；对所述待转换语音数据进行语音识别，得到识别结果及所述识别结果的时长信息；获取目标发音人的语音合成模型；利用所述语音合成模型及所述时长信息生成语音合成参数；利用所述语音合成参数对所述识别结果进行语音合成，得到目标发音人音色合成语音数据。利用本发明，可以使转换后的语音数据的时长与待转换语音数据的时长一致，提高合成语音的自然度。



1. 一种声音转换方法, 其特征在于, 包括 :

接收待转换语音数据 ;

对所述待转换语音数据进行语音识别, 得到识别结果及所述识别结果的时长信息 ;

获取目标发音人的语音合成模型 ;

利用所述语音合成模型及所述时长信息生成语音合成参数 ;

利用所述语音合成参数对所述识别结果进行语音合成, 得到目标发音人音色合成语音数据。

2. 根据权利要求 1 所述的方法, 其特征在于, 所述对所述待转换语音数据进行语音识别, 得到识别结果及所述识别结果的时长信息包括 :

利用预先训练的声学模型及语言模型构建解码网络 ;

提取所述待转换语音数据的特征参数 ;

基于所述解码网络及所述特征参数对所述待转换语音数据进行解码, 得到最优解码路径对应的文本序列及所述文本序列中各字和 / 或词的时长信息。

3. 根据权利要求 1 所述的方法, 其特征在于, 所述对所述待转换语音数据进行语音识别, 得到识别结果及所述识别结果的时长信息包括 :

利用预先训练的声学模型及语言模型构建解码网络 ;

提取所述待转换语音数据的特征参数 ;

基于所述解码网络及所述特征参数对所述待转换语音数据进行解码, 得到最优解码路径对应的语法单元序列及所述语法单元序列中各语法单元的时长信息。

4. 根据权利要求 1 所述的方法, 其特征在于, 所述获取目标发音人的语音合成模型包括 :

向用户展现可选的目标发音人信息, 并根据用户的选择确定目标发音人, 然后获取所述目标发音人的语音合成模型 ; 或者

接收用户提供的目标发音人语音数据, 并利用所述目标发音人语音数据训练得到目标发音人的语音合成模型。

5. 根据权利要求 1 至 4 任一项所述的方法, 其特征在于, 所述目标发音人合成模型包括 : 时长合成模型、基频合成模型、频谱合成模型 ;

所述利用所述语音合成模型及所述时长信息生成语音合成参数包括 :

利用所述时长信息及所述时长合成模型生成每个语法单元每个状态的时长合成参数 ;

利用目标发音人基频合成模型生成基频合成参数 ;

利用目标发音人频谱合成模型生成频谱合成参数。

6. 一种声音转换装置, 其特征在于, 包括 :

接收模块, 用于接收待转换语音数据 ;

语音识别模块, 用于对所述待转换语音数据进行语音识别, 得到识别结果及所述识别结果的时长信息 ;

模型获取模块, 用于获取目标发音人的语音合成模型 ;

合成参数生成模块, 用于利用所述语音合成模型及所述时长信息生成语音合成参数 ;

语音合成模块, 用于利用所述语音合成参数对所述识别结果进行语音合成, 得到目标

发音人音色合成功音数据。

7. 根据权利要求 6 所述的装置，其特征在于，所述语音识别模块包括：

第一解码网络构建单元，用于利用预先训练的声学模型及语言模型构建解码网络；

特征提取单元，用于提取所述待转换语音数据的特征参数；

第一解码单元，用于基于所述解码网络及所述特征参数对所述待转换语音数据进行解码，得到最优解码路径对应的文本序列及所述文本序列中各字和 / 或词的时长信息。

8. 根据权利要求 6 所述的装置，其特征在于，所述语音识别模块包括：

第二解码网络构建单元，用于利用预先训练的声学模型及语言模型构建解码网络；

特征提取单元，用于提取所述待转换语音数据的特征参数；

第二解码单元，用于基于所述解码网络及所述特征参数对所述待转换语音数据进行解码，得到最优解码路径对应的语法单元序列及所述语法单元序列中各语法单元的时长信息。

9. 根据权利要求 6 所述的装置，其特征在于，

所述模型获取模块包括：

展现单元，用于向用户展现可选的目标发音人信息；

目标发音人确定单元，用于根据用户的选择确定目标发音人；

模型获取单元，用于获取所述目标发音人的语音合成模型；

或者，所述目标发音人确定模块包括：

接收单元，用于接收用户提供的目标发音人语音数据；

模型训练单元，用于利用所述目标发音人语音数据训练得到目标发音人的语音合成模型。

10. 根据权利要求 6 至 9 任一项所述的装置，其特征在于，所述目标发音人合成模型包括：时长合成模型、基频合成模型、频谱合成模型；

所述合成参数生成模块包括：

时长合成参数生成单元，用于利用所述时长信息及所述时长合成模型生成每个语法单元每个状态的时长合成参数；

基频合成参数生成单元，用于利用目标发音人基频合成模型生成基频合成参数；

频谱合成参数生成单元，用于利用目标发音人频谱合成模型生成频谱合成参数。

一种声音转换方法及装置

技术领域

[0001] 本发明涉及语音信号处理技术领域，具体涉及一种声音转换方法及装置。

背景技术

[0002] 在日常的生活交流中，一个人的声音往往就是他的身份名片，听到自己熟悉人的声音后，就可辨认出这个人。声音转换技术由于可以将一个发音人的声音转换为另一个发音人的声音，使人听起来像是另一个人的发音，有着广泛的应用前景，如用户可以将自己的声音转换成自己喜欢的明星的声音，或转换成用户自己熟悉人的声音。

[0003] 现有的声音转换方法一般是将待转换语音数据进行语音识别，得到识别文本后，利用目标发音人合成模型对所述识别文本进行语音合成，从而得到目标发音人音色的合成语音数据。这种方法对识别文本进行语音合成时，容易出现合成的语音数据与待转换语音数据的时长不一致的情况，从而使合成语音听起来较机械，韵律感差，大大降低了合成语音的自然度。

发明内容

[0004] 本发明提供一种声音转换方法及装置，以使转换后的语音数据的时长与待转换语音数据的时长一致，提高合成语音的自然度。

[0005] 为此，本发明提供如下技术方案：

[0006] 一种声音转换方法，包括：

[0007] 接收待转换语音数据；

[0008] 对所述待转换语音数据进行语音识别，得到识别结果及所述识别结果的时长信息；

[0009] 获取目标发音人的语音合成模型；

[0010] 利用所述语音合成模型及所述时长信息生成语音合成参数；

[0011] 利用所述语音合成参数对所述识别结果进行语音合成，得到目标发音人音色合成语音数据。

[0012] 优选地，所述对所述待转换语音数据进行语音识别，得到识别结果及所述识别结果的时长信息包括：

[0013] 利用预先训练的声学模型及语言模型构建解码网络；

[0014] 提取所述待转换语音数据的特征参数；

[0015] 基于所述解码网络及所述特征参数对所述待转换语音数据进行解码，得到最优解码路径对应的文本序列及所述文本序列中各字和 / 或词的时长信息。

[0016] 优选地，所述对所述待转换语音数据进行语音识别，得到识别结果及所述识别结果的时长信息包括：

[0017] 利用预先训练的声学模型及语言模型构建解码网络；

[0018] 提取所述待转换语音数据的特征参数；

- [0019] 基于所述解码网络及所述特征参数对所述待转换语音数据进行解码,得到最优解码路径对应的语法单元序列及所述语法单元序列中各语法单元的时长信息。
- [0020] 优选地,所述获取目标发音人的语音合成模型包括:
- [0021] 向用户展现可选的目标发音人信息,并根据用户的选择确定目标发音人,然后获取所述目标发音人的语音合成模型;或者
- [0022] 接收用户提供的目标发音人语音数据,并利用所述目标发音人语音数据训练得到目标发音人的语音合成模型。
- [0023] 优选地,所述目标发音人合成模型包括:时长合成模型、基频合成模型、频谱合成模型;
- [0024] 所述利用所述语音合成模型及所述时长信息生成语音合成参数包括:
- [0025] 利用所述时长信息及所述时长合成模型生成每个语法单元每个状态的时长合成参数;
- [0026] 利用目标发音人基频合成模型生成基频合成参数;
- [0027] 利用目标发音人频谱合成模型生成频谱合成参数。
- [0028] 一种声音转换装置,包括:
- [0029] 接收模块,用于接收待转换语音数据;
- [0030] 语音识别模块,用于对所述待转换语音数据进行语音识别,得到识别结果及所述识别结果的时长信息;
- [0031] 模型获取模块,用于获取目标发音人的语音合成模型;
- [0032] 合成参数生成模块,用于利用所述语音合成模型及所述时长信息生成语音合成参数;
- [0033] 语音合成模块,用于利用所述语音合成参数对所述识别结果进行语音合成,得到目标发音人音色合成语音数据。
- [0034] 优选地,所述语音识别模块包括:
- [0035] 第一解码网络构建单元,用于利用预先训练的声学模型及语言模型构建解码网络;
- [0036] 特征提取单元,用于提取所述待转换语音数据的特征参数;
- [0037] 第一解码单元,用于基于所述解码网络及所述特征参数对所述待转换语音数据进行解码,得到最优解码路径对应的文本序列及所述文本序列中各字和/或词的时长信息。
- [0038] 优选地,所述语音识别模块包括:
- [0039] 第二解码网络构建单元,用于利用预先训练的声学模型及语言模型构建解码网络;
- [0040] 特征提取单元,用于提取所述待转换语音数据的特征参数;
- [0041] 第二解码单元,用于基于所述解码网络及所述特征参数对所述待转换语音数据进行解码,得到最优解码路径对应的语法单元序列及所述语法单元序列中各语法单元的时长信息。
- [0042] 优选地,所述模型获取模块包括:
- [0043] 展现单元,用于向用户展现可选的目标发音人信息;
- [0044] 目标发音人确定单元,用于根据用户的选择确定目标发音人;

[0045] 模型获取单元,用于获取所述目标发音人的语音合成模型 ;
[0046] 或者,所述目标发音人确定模块包括 :
[0047] 接收单元,用于接收用户提供的目标发音人语音数据 ;
[0048] 模型训练单元,用于利用所述目标发音人语音数据训练得到目标发音人的语音合成模型。
[0049] 优选地,所述目标发音人合成模型包括 :时长合成模型、基频合成模型、频谱合成模型 ;
[0050] 所述合成参数生成模块包括 :
[0051] 时长合成参数生成单元,用于利用所述时长信息及所述时长合成模型生成每个语法单元每个状态的时长合成参数 ;
[0052] 基频合成参数生成单元,用于利用目标发音人基频合成模型生成基频合成参数 ;
[0053] 频谱合成参数生成单元,用于利用目标发音人频谱合成模型生成频谱合成参数。
[0054] 本发明实施例提供的声音转换方法及装置,首先接收待转换语音数据,然后对待转换语音数据进行语音识别,得到识别结果及其时长信息,最后利用目标发音人的语音合成模型及所述时长信息生成语音合成参数,利用该语音合成参数对所述识别结果进行语音合成,得到目标发音人音色合成功能语音数据。该方法及装置对待转换语音数据进行语音识别时,不仅获取识别结果,而且还要获取该识别结果的时长信息,利用该时长信息生成目标发音人的语音合成参数,有效保证了合成功能语音数据的时长与待转换语音数据的时长一致,提高了转换后语音的自然度。

附图说明

[0055] 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发明中记载的一些实施例,对于本领域普通技术人员来讲,还可以根据这些附图获得其他的附图。

[0056] 图 1 是本发明实施例声音转换方法的一种流程图 ;
[0057] 图 2 是本发明实施例声音转换方法的一种具体应用流程图 ;
[0058] 图 3 是本发明实施例声音转换装置的一种结构示意图。

具体实施方式

[0059] 为了使本技术领域的人员更好地理解本发明实施例的方案,下面结合附图和实施方式对本发明实施例作进一步的详细说明。

[0060] 针对现有技术进行声音转换时容易出现合成的语音数据与待转换语音数据的时长不一致的情况,使转换后的声音韵律感差、自然度低的问题,本发明实施例提供一种声音转换方法及装置,在对待转换语音数据进行语音识别时,获取识别结果对应的时长信息,利用该时长信息生成目标发音人的语音合成参数,从而使最终得到的目标发音人音色的合成功能语音数据与待转换语音数据的时长保持一致,提高转换后语音的自然度。

[0061] 如图 1 所示,是本发明实施例声音转换方法的一种流程图,包括以下步骤 :
[0062] 步骤 101,接收待转换语音数据。
[0063] 步骤 102,对所述待转换语音数据进行语音识别,得到识别结果及所述识别结果的

时长信息。

[0064] 语音识别的具体过程与现有技术相同,即利用预先训练的声学模型及语言模型构建解码网络;提取语音数据的特征参数,比如,线性预测参数(LPCC)、和/或 Mel 频率倒谱系数(MFCC)参数,然后基于所述解码网络及所述特征参数对所述语音数据进行解码,得到最优解码路径对应的识别文本,即由字和/词组成的文本序列。不同的是,在本发明实施例中,不仅要获取识别结果,还要获取与该识别结果对应的时长信息,也就是说,所述文本序列中各字和/或词的时长信息。所述时长信息可以根据所述字、词对应的语音段的时长信息来得到,在此不再详述。

[0065] 步骤 103,获取目标发音人的语音合成模型。

[0066] 所述目标发音人语音合成模型主要包括目标发音人的时长合成模型、基频合成模型和频谱合成模型。

[0067] 在实际应用中,目标发音人的语音合成模型的获取可以有多种方式。

[0068] 比如,向用户展现可选的目标发音人信息,根据用户的选择确定目标发音人,然后即可从模型库中获取所述目标发音人的语音合成模型。所述发音人信息可以是目标发音人编号、目标发音人名称等,对此本发明实施例不做限定。当然,给出目标发音人同时,还可以给出对每个目标发音人发音特点的简单描述,如发音人:小明,发音特点:浑厚有力、语速较慢。所述目标发音人的语音合成模型可以通过预先收集大量目标发音人语音数据训练得到。当然,目标发音人的确定还可以有其它方式,比如由系统随机给出目标发音人等,在此不再一一列举。

[0069] 再比如,也可以利用用户提供的目标发音人语音数据得到目标发音人的语音合成模型,具体地,接收用户提供的目标发音人语音数据,然后利用所述目标发音人语音数据训练得到目标发音人的语音合成模型;或者根据用户提供的目标发音人语音数据进行模型自适应得到,具体训练过程或自适应过程与现有技术相同,在此不再详述。

[0070] 步骤 104,利用所述语音合成模型及所述时长信息生成语音合成参数。

[0071] 所述语音合成参数包括时长参数、基频参数、频谱参数,各种参数生成方法具体如下:

[0072] 对于识别文本,使用语音合成本文分析器将其解析成对应的语法单元序列,所述语法单元为语音合成时,使用的最小语法单元,如音素;每个语法单元包含多个状态,如 5 个,每个状态的时长分布假设服从单高斯分布:

$$[0073] P(d_n^i | p_n, i) = N(d; \mu_n^i, \sigma_n^{i^2}) \quad (1)$$

[0074] 其中, p_n 为第 n 个语法单元, d_n^i 为第 n 个语法单元第 i 个状态的时长, μ_n^i 和 $\sigma_n^{i^2}$ 为第 n 个语法单元第 i 个状态的时长合成模型均值和方差。

[0075] 为了保证合成语音数据与待转换语音数据时长一致,本发明实施例对生成的时长参数进行约束,即在待转换语音时长范围内生成时长合成参数,如对各字或词的时长进行约束,具体约束方法如式(2)所示:

$$[0076] \sum_{n \in C_j} \sum_{i=1}^{i=S} d_n^i = D_j \quad (2)$$

[0077] 其中, C_j 为第 j 个字或词所包含的语法单元集合, D_j 为第 j 个字或词的时长, S 为每个语法单元的状态数。

[0078] 使用最大似然准则估计得到每个语法单元每个状态的时长参数集合 $\{d_n^{i*}\}$, 如式(3) 所示:

$$[0079] \quad \{d_n^{i*}\} = \arg \max_{\{d_n^i\}} \prod_{i=1}^S P(d_n^i | p_n, i) \quad (3)$$

[0080] 其中, d_n^{i*} 为第 n 个语法单元第 i 个状态估计得到的时长参数。

[0081] 将式(1) 和式(2) 代入式(3) 求解即可得每个语法单元每个状态的时长参数, 如式(4) 所示。

$$[0082] \quad d_n^{i*} = \mu_n^i + \frac{\sigma_n^{i2} (d_{pn} - \sum_{n \in C_j} \sum_{i=1}^{i=S} \mu_i)}{\sum_{n \in C_j} \sum_{i=1}^{i=S} \sigma_n^{i2}} \quad (4)$$

[0083] 频谱、基频参数的生成与传统方法一致。

[0084] 步骤 105, 利用所述语音合成参数对所述识别结果进行语音合成, 得到目标发音人音色合成语音数据。

[0085] 本发明实施例提供的声音转换方法, 首先接收待转换语音数据, 然后对待转换语音数据进行语音识别, 得到识别结果及其时长信息, 最后利用目标发音人的语音合成模型及所述时长信息生成语音合成参数, 利用该语音合成参数对所述识别结果进行语音合成, 得到目标发音人音色合成语音数据。该方法对待转换语音数据进行语音识别时, 不仅获取识别结果, 而且还要获取该识别结果的时长信息, 利用该时长信息生成目标发音人的语音合成参数, 有效保证了合成语音数据的时长与待转换语音数据的时长一致, 提高了转换后语音的自然度。

[0086] 考虑到直接对识别文本进行语音合成, 容易将语音识别过程出现的错误带入到语音合成中, 如多音字问题, 造成合成后语音数据的语义相比待转换语音数据的语义发生了变化, 如待转换语音数据为“办张美国信用卡”, 识别文本为“办张没过信用卡”, 出现了识别错误, 利用目标发音人语音合成模型对识别文本进行合成后, 得到的合成语音为“办张没过信用卡”, 合成语音的语义发生了变化, 这是不希望出现的结果。因此, 在实际应用中, 还可以将根据声学模型得到的语法单元序列作为所述识别结果, 同时获取所述语法单元序列中各语法单元的时长信息。这样, 在进行语音合成时, 直接对待转换语音数据对应的语法单元序列进行语音合成, 从而避免了将语音识别过程出现的错误带入到语音合成中, 保证了合成后的语音数据的语义与待转换语音数据的语义的一致性。

[0087] 下面结合图 2 所示流程对上述声音转换方法做进一步详细说明。

[0088] 如图 2 所示, 是本发明实施例声音转换方法的一种具体应用流程图, 包括以下步骤:

[0089] 步骤 201, 接收待转换语音数据。

[0090] 步骤 202, 利用预先训练的声学模型及语言模型构建解码网络。

- [0091] 步骤 203, 提取所述待转换语音数据的特征参数。
- [0092] 所述特征参数可以是 LPCC、和 / 或 MFCC。
- [0093] 步骤 204, 基于所述解码网络及所述特征参数对所述语音数据进行解码, 得到最优解码路径对应的语法单元序列及所述语法单元序列中各语法单元的时长信息。
- [0094] 所述语法单元是指语音识别时使用的最小语法单元, 如音素。
- [0095] 步骤 205, 获取目标发音人的语音合成模型。
- [0096] 步骤 206, 利用所述语音合成模型及所述时长信息生成语音合成参数。
- [0097] 所述语音合成参数包括时长参数、基频参数、频谱参数, 各种参数生成方法具体如下:
- [0098] 1) 利用语法单元序列时长信息及目标发音人时长合成模型生成时长合成参数
- [0099] 每个语法单元序列采用多个状态表示, 如 5 个状态; 每个状态的时长模型假设服从单高斯分布, 如式 (5) 所示:

$$[0100] P(d_n^i | p_n, i) = N(d; \mu_n^i, \sigma_n^{i2}) \quad (5)$$

[0101] 其中, p_n 为第 n 个语法单元, d_n^i 为第 n 个语法单元第 i 个状态的时长, μ_n^i 和 σ_n^{i2} 为第 n 个语法单元第 i 个状态的时长合成模型均值和方差。

[0102] 为了保证合成功语音数据与待转换语音数据时长一致, 本发明实施例对生成的时长参数进行约束, 即在待转换语音时长范围内生成时长合成参数, 具体约束方法如式 (6) 所示:

$$[0103] \sum_{i=1}^{i=S} d_n^i = d_{pn} \quad (6)$$

[0104] 其中, d_{pn} 为待转换语音中第 n 个语法单元时长, S 为语法单元的状态总数。

[0105] 根据待转换语音数据语法单元对应的时长约束, 及目标发音人时长合成模型, 采用最大似然准则估计得到每个语法单元每个状态的时长合成参数 $\{d_n^{i*}\}$, 如式 (7) 所示:

$$[0106] \{d_n^{i*}\} = \arg \max_{\{d_n^i\}} \prod_{i=1}^S P(d_n^i | p_n, i) \quad (7)$$

[0107] 其中, d_n^{i*} 为第 n 个语法单元第 i 个状态估计得到的时长参数。

[0108] 将式 (5) 和式 (6) 代入式 (7) 进行计算, 可以得到语法单元每个状态的时长, 具体如式 (8) 所示:

$$[0109] d_n^{i*} = \mu_n^i + \frac{\sigma_n^{i2} (d_{pn} - \sum_{i=1}^{i=S} \mu_i)}{\sum_{i=1}^{i=S} \sigma_n^{i2}} \quad (8)$$

[0110] 2) 利用目标发音人基频合成模型生成基频合成参数

[0111] 基频合成参数的生成过程如下:

[0112] 首先, 对识别得到的语法单元序列进行扩展, 扩展成上下文相关的语法单元序

列,如语法单元序列为“xx-y-u-y-in-h-e-ch-eng-xx”,将所述语法单元序列扩展成上下文相关的语法单元序列为：“xx-y+u:/A, y-u+y:/A, u-y+in:/A, y-in+h:/A, in-h+e:/A, h-e+ch:/A, e-ch+eng:/A, ch-eng+xx:/A”,其中“-”和“+”之间语法单元为当前语法单元,“:/A”为当前语法单元的上下文相关信息,如声调信息,当然所述上下文相关的语法单元序列的表示方法不限于上述表示方法;

[0113] 然后,利用基频合成模型预测得到当前语法单元各状态的基频模型,具体预测方法与现有技术相同,在此不再详述;

[0114] 随后,根据语法单元序列的状态时长信息对各语法单元相应状态进行复制,根据每个语法单元预测得到的各状态的基频模型,得到复制后的语法单元序列的基频分布,即语法单元序列预测得到的基频模型;

[0115] 最后,根据语法单元序列的基频分布生成基频合成参数,如式(9)所示:

$$c = (W^T UW)^{-1} W^T UW \quad (9)$$

[0117] 其中,W为计算语法单元序列动态参数的窗函数矩阵,c为待生成的基频合成参数,M和U分别为预测得到的语法单元序列所有状态基频模型的均值及协方差矩阵。

[0118] 3) 利用目标发音人频谱合成模型生成频谱合成参数

[0119] 频谱合成参数的生成过程与上述基频合成参数的生成过程类似,在此不再赘述。

[0120] 步骤207,利用所述语音合成参数对所述识别结果进行语音合成,得到目标发音人音色合成语音数据。

[0121] 语音合成的具体实现过程与现有技术相同,在此不再赘述。

[0122] 本发明实施例提供的声音转换方法,不仅有效保证了合成语音数据的时长与待转换语音数据的时长一致,提高了转换后语音的自然度;而且还进一步将根据声学模型得到的语法单元序列作为所述识别结果,这样,在进行语音合成时,直接对待转换语音数据对应的语法单元序列进行语音合成,从而避免了将语音识别过程出现的错误带入到语音合成中,保证了合成后的语音数据的语义与待转换语音数据的语义的一致性。

[0123] 相应地,本发明实施例还提供一种声音转换装置,如图3所示,是本发明实施例声音转换装置的一种结构示意图。

[0124] 在该实施例中,所述装置包括:

[0125] 接收模块301,用于接收待转换语音数据;

[0126] 语音识别模块302,用于对所述待转换语音数据进行语音识别,得到识别结果及所述识别结果的时长信息;

[0127] 模型获取模块303,用于获取目标发音人的语音合成模型;

[0128] 合成参数生成模块304,用于利用所述语音合成模型及所述时长信息生成语音合成参数;

[0129] 语音合成模块305,用于利用所述语音合成参数对所述识别结果进行语音合成,得到目标发音人音色合成语音数据。

[0130] 在实际应用中,语音识别模块302可以对待识别语音数据进行语音识别,得到待识别语音数据对应的文本序列及所述文本序列中各字和/或词的时长信息。相应地,语音识别模块302的一种具体结构包括以下各单元:

[0131] 第一解码网络构建单元,用于利用预先训练的声学模型及语言模型构建解码网

络；

[0132] 特征提取单元，用于提取待转换语音数据的特征参数；

[0133] 第一解码单元，用于基于所述解码网络及所述特征参数对所述待转换语音数据进行解码，得到最优解码路径对应的文本序列及所述文本序列中各字和 / 或词的时长信息。

[0134] 考虑到直接对识别文本进行语音合成，容易将语音识别过程出现的错误带入到语音合成中，如多音字问题，造成合成后语音数据的语义相比待转换语音数据的语义发生了变化。因此，在实际应用中，语音识别模块 302 还可以将根据声学模型得到的语法单元序列作为所述识别结果，同时获取所述语法单元序列中各语法单元的时长信息。这样，在进行语音合成时，直接对待转换语音数据对应的语法单元序列进行语音合成，从而避免了将语音识别过程出现的错误带入到语音合成中，保证了合成后的语音数据的语义与待转换语音数据的语义的一致性。相应地，语音识别模块 302 的另一种具体结构包括以下各单元：

[0135] 第二解码网络构建单元，用于利用预先训练的声学模型及语言模型构建解码网络；

[0136] 特征提取单元，用于提取待转换语音数据的特征参数；

[0137] 第二解码单元，用于基于所述解码网络及所述特征参数对所述待转换语音数据进行解码，得到最优解码路径对应的语法单元序列及所述语法单元序列中各语法单元的时长信息。

[0138] 另外，上述模型获取模块 303 也可以有多种实现方式。

[0139] 比如，模型获取模块 303 的一种具体结构可以包括：展现单元、目标发音人确定单元、以及模型获取单元。其中，所述展现单元用于向用户展现可选的目标发音人信息；所述目标发音人确定单元用于根据用户的选择确定目标发音人；所述模型获取单元用于获取所述目标发音人的语音合成模型。

[0140] 再比如，模型获取模块 303 的另一种具体结构可以包括：接收单元和模型训练单元。其中，所述接收单元用于接收用户提供的目标发音人语音数据；所述模型训练单元用于利用所述目标发音人语音数据训练得到目标发音人的语音合成模型。

[0141] 所述目标发音人合成模型包括：时长合成模型、基频合成模型、频谱合成模型。

[0142] 相应地，所述合成参数生成模块 304 包括：

[0143] 时长合成参数生成单元，用于利用所述时长信息及所述时长合成模型生成每个语法单元每个状态的时长合成参数；

[0144] 基频合成参数生成单元，用于利用目标发音人基频合成模型生成基频合成参数；

[0145] 频谱合成参数生成单元，用于利用目标发音人频谱合成模型生成频谱合成参数。

[0146] 本发明实施例提供的声音转换装置，首先接收待转换语音数据，然后对待转换语音数据进行语音识别，得到识别结果及其时长信息，最后利用目标发音人的语音合成模型及所述时长信息生成语音合成参数，利用该语音合成参数对所述识别结果进行语音合成，得到目标发音人音色合成语音数据。该方法及系统对待转换语音数据进行语音识别时，不仅获取识别结果，而且还要获取该识别结果的时长信息，利用该时长信息生成目标发音人的语音合成参数，有效保证了合成语音数据的时长与待转换语音数据的时长一致，提高了转换后语音的自然度。进一步地，可以将根据声学模型得到的语法单元序列作为所述识别结果，这样，在进行语音合成时，直接对待转换语音数据对应的语法单元序列进行语音合

成,从而避免了将语音识别过程出现的错误带入到语音合成中,保证了合成后的语音数据的语义与待转换语音数据的语义的一致性。

[0147] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于装置实施例而言,由于其基本相似于方法实施例,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0148] 以上对本发明实施例进行了详细介绍,本文中应用了具体实施方式对本发明进行了阐述,以上实施例的说明只是用于帮助理解本发明的方法及装置;同时,对于本领域的一般技术人员,依据本发明的思想,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发明的限制。

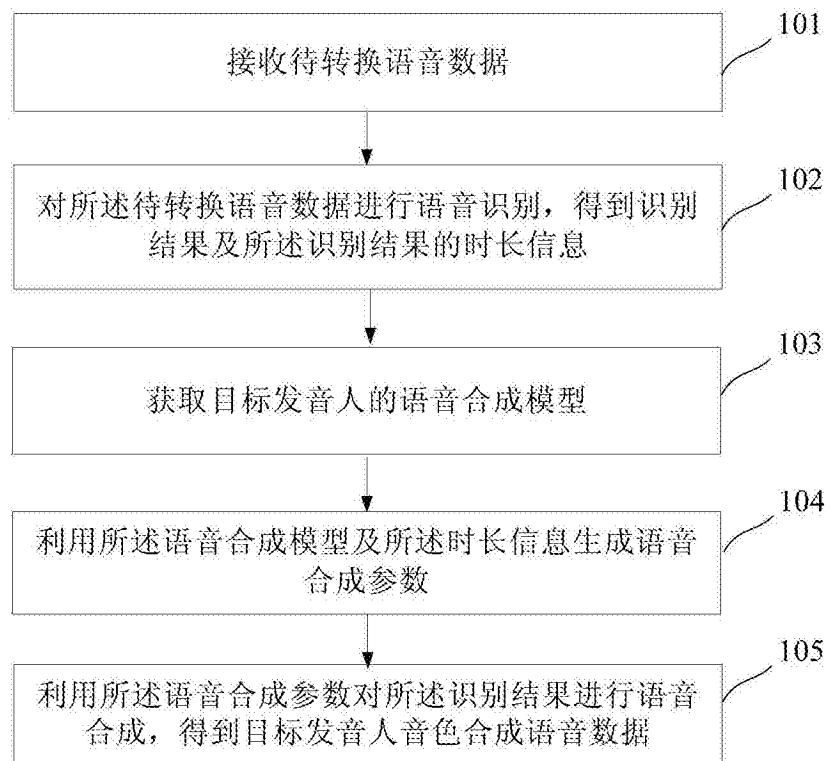


图 1

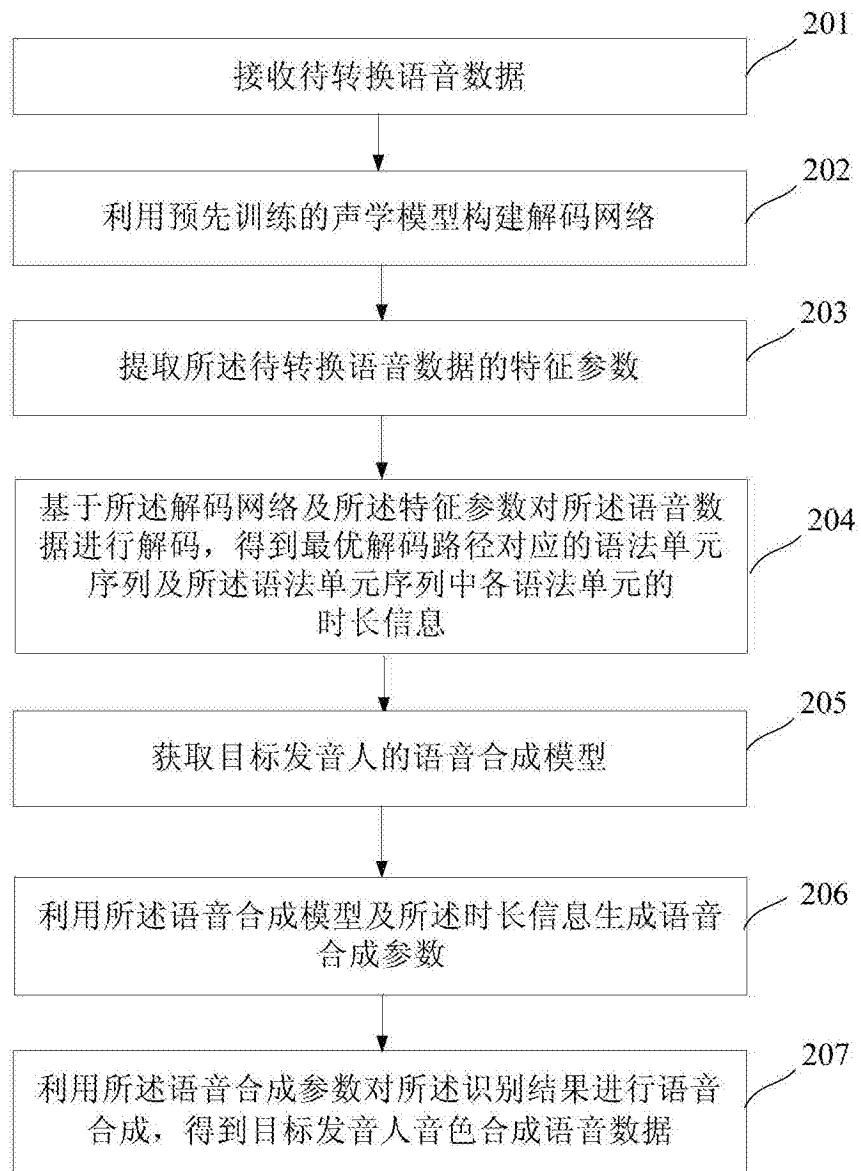


图 2

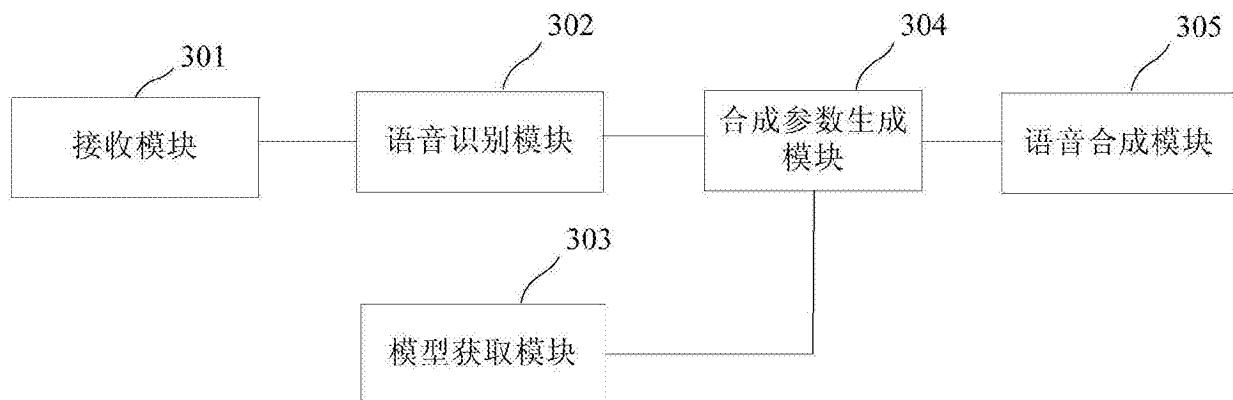


图 3