(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105023570 A
(43)申请公布日 2015.11.04

(21)申请号 201410182517.7

(22)申请日 2014.04.30

(71)申请人 安徽科大讯飞信息科技股份有限公司
地址 230088 安徽省合肥市高新开发区望江西路666号

(72)发明人 陈凌辉 江源 凌震华 胡国平 胡郁 刘庆峰

(74)专利代理机构 北京维澳专利代理有限公司 11252
代理人 王立民 吉海莲

(51)Int.Cl.
G10L 13/02(2013.01)
G10L 15/02(2006.01)

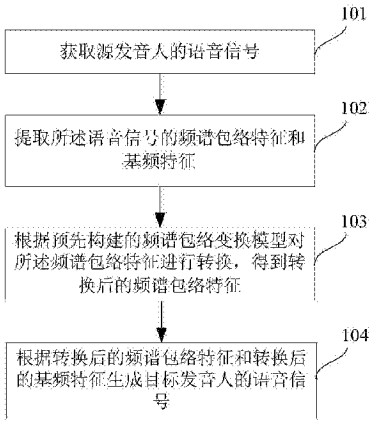
权利要求书3页 说明书11页 附图4页

(54)发明名称

一种实现声音转换的方法及系统

(57)摘要

本发明涉及语音合成技术领域，公开了一种实现声音转换的方法及系统，该方法包括：获取源发音人的语音信号；提取所述语音信号的频谱包络特征和基频特征；根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征；根据转换后的频谱包络特征和基频特征生成目标发音人的语音信号。利用本发明，可以有效提高转换语音的音质。



CN 105023570 A

1. 一种实现声音转换的方法，其特征在于，包括：

获取源发音人的语音信号；

提取所述语音信号的频谱包络特征和基频特征；

根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征；

根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号。

2. 根据权利要求 1 所述的方法，其特征在于，所述提取所述语音信号的频谱包络特征包括：

对于每一帧语音帧，提取其上下多帧的频谱包络特征作为所述语音帧的频谱包络特征。

3. 根据权利要求 1 所述的方法，其特征在于，按以下方式构建频谱包络变换模型：

获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据；

提取所述训练语音数据的频谱包络特征；

确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系；

确定源发音人和目标发音人的频谱包络变换模型拓扑结构；

根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数。

4. 根据权利要求 3 所述的方法，其特征在于，所述确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系包括：

以语音帧为单位，提取所述训练语音信号的美尔倒谱特征序列；

将源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列对齐；

根据所述源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列的对应关系，确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系。

5. 根据权利要求 3 所述的方法，其特征在于，所述确定源发音人和目标发音人的频谱包络变换模型拓扑结构包括：

利用第一 RBM 模型模拟源发音人频谱包络分布特点，并将其作为源发音人模型；

利用第二 RBM 模型模拟目标发音人频谱包络分布特点，并将其作为目标发音人模型；

利用 BBAM 模型模拟源发音人和目标发音人之间的参数传递关系，并将其作为转换模型；

拼接所述源发音人模型、转换模型、以及目标发音人模型，得到源发音人和目标发音人的频谱包络变换模型拓扑结构。

6. 根据权利要求 5 所述的方法，其特征在于，所述根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数包括：

获取训练数据，所述训练数据包括源发音人频谱包络数据和目标发音人频谱包络数据；

根据所述源发音人频谱包络数据训练源发音人模型参数，并根据所述目标发音人频谱包络数据训练目标发音人模型参数；

根据所述对应关系训练转换模型参数；

将所述源发音人模型参数、转换模型参数、以及目标发音人模型参数进行合并，得到所

述源发音人和目标发音人的频谱包络变换模型参数。

7. 根据权利要求 6 所述的方法，其特征在于，所述根据所述对应关系训练转换模型参数包括：

从所述训练数据中采样得到转换模型参数训练数据；

基于所述转换模型参数训练数据训练转换模型参数。

8. 根据权利要求 1 至 7 任一项所述的方法，其特征在于，所述根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征包括：

根据所述频谱包络变换模型计算所述频谱包络特征的条件概率分布；

根据所述条件概率分布确定转换后的频谱包络特征。

9. 一种实现声音转换的系统，其特征在于，包括：

语音信号获取模块，用于获取源发音人的语音信号；

特征提取模块，用于提取所述语音信号的频谱包络特征和基频特征；

频谱包络特征转换模块，用于根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征；

语音生成模块，用于根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号。

10. 根据权利要求 9 所述的系统，其特征在于，所述系统还包括：频谱包络变换模型构建模块，所述频谱包络变换模型构建模块包括：

训练语音数据获取单元，用于获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据；

特征提取单元，用于提取所述训练语音数据的频谱包络特征；

对应关系确定单元，用于确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系；

拓扑结构确定单元，用于确定源发音人和目标发音人的频谱包络变换模型拓扑结构；

参数训练单元，用于根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数。

11. 根据权利要求 10 所述的系统，其特征在于，所述对应关系确定单元包括：

第一提取单元，用于以语音帧为单位，提取所述训练语音信号的美尔倒谱特征序列；

对齐单元，用于将源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列对齐；

第一确定单元，用于根据所述源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列的对应关系，确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系。

12. 根据权利要求 10 所述的系统，其特征在于，所述拓扑结构确定单元包括：

源发音人模型单元，用于利用第一 RBM 模型模拟源发音人频谱包络分布特点，并将其作为源发音人模型；

目标发音人模型单元，用于利用第二 RBM 模型模拟目标发音人频谱包络分布特点，并将其作为目标发音人模型；

转换模型单元，用于利用 BBAM 模型模拟源发音人和目标发音人之间的参数传递关系，

并将其作为转换模型；

拼接单元，用于拼接所述源发音人模型、转换模型、以及目标发音人模型，得到源发音人和目标发音人的频谱包络变换模型拓扑结构。

13. 根据权利要求 12 所述的系统，其特征在于，所述参数训练单元包括：

获取频谱包络训练数据单元，用于获取训练数据，所述训练数据包括源发音人频谱包络数据和目标发音人频谱包络数据；

第一训练单元，用于根据所述源发音人频谱包络数据训练源发音人模型参数；

第二训练单元，用于根据所述目标发音人频谱包络数据训练目标发音人模型参数；

第三训练单元，用于根据所述对应关系训练转换模型参数；

合并单元，用于将所述源发音人模型参数、转换模型参数、以及目标发音人模型参数进行合并，得到所述源发音人和目标发音人的频谱包络变换模型参数。

14. 根据权利要求 9 至 13 任一项所述的系统，其特征在于，所述频谱包络特征转换模块包括：

条件概率分布计算单元，用于根据所述频谱包络变换模型计算所述频谱包络特征的条件概率分布；

转换特征确定单元，用于根据所述条件概率分布确定转换后的频谱包络特征。

一种实现声音转换的方法及系统

技术领域

[0001] 本发明涉及语音信号处理技术领域，具体涉及及一种实现声音转换的方法及系统。

背景技术

[0002] 声音转换即将一个发音人（源发音人）的语音转换为另一个发音人（目标发音人）的语音，使其具有目标发音人的发音特点。声音转换技术在实际生活中有广泛应用，可以帮助因发音器官受损而植入电子喉的病人发出高质量的语音，还可以丰富娱乐生活，通过模拟明星发音人的发音特点提高娱乐性等，具有广泛的应用前景。

[0003] 现有声音转换系统主要采用频谱变换和基频变换的方法，对源发音人的语音特征进行转换，使其具有目标发音人的发音特点，实现声音转换。相比于基频变换，由于频谱对于发音人身份信息的确认作用更加关键，因而基于基频和频谱变换的声音转换更为实用。

[0004] 现有的频谱变换技术主要采用数学统计模型训练源发音人和目标发音人的频谱特征的联合概率分布，确定源发音人和目标发音人的频谱变换关系。在接收到源发音人语音时，根据所述联合概率分布计算目标发音人特征的条件分布，并生成目标发音人特征样本。在数据统计方法中，训练数据越多，模型越准确，则模拟效果越好。然而由于应用场景的限制，能够获取的训练数据量往往较少，其应用模型往往较为简单，相应的转换得到的语音质量往往不高。

发明内容

[0005] 本发明实施例提供一种实现声音转换的方法及系统，以提高转换语音的音质。

[0006] 为此，本发明实施例提供如下技术方案：

[0007] 一种实现声音转换的方法，包括：

[0008] 获取源发音人的语音信号；

[0009] 提取所述语音信号的频谱包络特征和基频特征；

[0010] 根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征；

[0011] 根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号。

[0012] 优选地，所述提取所述语音信号的频谱包络特征包括：

[0013] 对于每一帧语音帧，提取其上下多帧的频谱包络特征作为所述语音帧的频谱包络特征。

[0014] 优选地，按以下方式构建频谱包络变换模型：

[0015] 获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据；

[0016] 提取所述训练语音数据的频谱包络特征；

[0017] 确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系；

[0018] 确定源发音人和目标发音人的频谱包络变换模型拓扑结构；

[0019] 根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数。

[0020] 优选地，所述确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系包括：

[0021] 以语音帧为单位，提取所述训练语音信号的美尔倒谱特征序列；

[0022] 将源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列对齐；

[0023] 根据所述源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列的对应关系，确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系。

[0024] 优选地，所述确定源发音人和目标发音人的频谱包络变换模型拓扑结构包括：

[0025] 利用第一 RBM 模型模拟源发音人频谱包络分布特点，并将其作为源发音人模型；

[0026] 利用第二 RBM 模型模拟目标发音人频谱包络分布特点，并将其作为目标发音人模型；

[0027] 利用 BBAM 模型模拟源发音人和目标发音人之间的参数传递关系，并将其作为转换模型；

[0028] 拼接所述源发音人模型、转换模型、以及目标发音人模型，得到源发音人和目标发音人的频谱包络变换模型拓扑结构。

[0029] 优选地，所述根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数包括：

[0030] 获取训练数据，所述训练数据包括源发音人频谱包络数据和目标发音人频谱包络数据；

[0031] 根据所述源发音人频谱包络数据训练源发音人模型参数，并根据所述目标发音人频谱包络数据训练目标发音人模型参数；

[0032] 根据所述对应关系训练转换模型参数；

[0033] 将所述源发音人模型参数、转换模型参数、以及目标发音人模型参数进行合并，得到所述源发音人和目标发音人的频谱包络变换模型参数。

[0034] 优选地，所述根据所述对应关系训练转换模型参数包括：

[0035] 从所述训练数据中采样得到转换模型参数训练数据；

[0036] 基于所述转换模型参数训练数据训练转换模型参数。

[0037] 优选地，所述根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征包括：

[0038] 根据所述频谱包络变换模型计算所述频谱包络特征的条件概率分布；

[0039] 根据所述条件概率分布确定转换后的频谱包络特征。

[0040] 一种实现声音转换的系统，包括：

[0041] 语音信号获取模块，用于获取源发音人的语音信号；

[0042] 特征提取模块，用于提取所述语音信号的频谱包络特征和基频特征；

[0043] 频谱包络特征转换模块，用于根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征；

[0044] 语音生成模块，用于根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号。

[0045] 优选地，所述系统还包括：频谱包络变换模型构建模块，所述频谱包络变换模型构

建模块包括：

[0046] 训练语音数据获取单元，用于获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据；

[0047] 特征提取单元，用于提取所述训练语音数据的频谱包络特征；

[0048] 对应关系确定单元，用于确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系；

[0049] 拓扑结构确定单元，用于确定源发音人和目标发音人的频谱包络变换模型拓扑结构；

[0050] 参数训练单元，用于根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数。

[0051] 优选地，所述对应关系确定单元包括：

[0052] 第一提取单元，用于以语音帧为单位，提取所述训练语音信号的美尔倒谱特征序列；

[0053] 对齐单元，用于将源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列对齐；

[0054] 第一确定单元，用于根据所述源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列的对应关系，确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系。

[0055] 优选地，所述拓扑结构确定单元包括：

[0056] 源发音人模型单元，用于利用第一 RBM 模型模拟源发音人频谱包络分布特点，并将其作为源发音人模型；

[0057] 目标发音人模型单元，用于利用第二 RBM 模型模拟目标发音人频谱包络分布特点，并将其作为目标发音人模型；

[0058] 转换模型单元，用于利用 BBAM 模型模拟源发音人和目标发音人之间的参数传递关系，并将其作为转换模型；

[0059] 拼接单元，用于拼接所述源发音人模型、转换模型、以及目标发音人模型，得到源发音人和目标发音人的频谱包络变换模型拓扑结构。

[0060] 优选地，所述参数训练单元包括：

[0061] 获取频谱包络训练数据单元，用于获取训练数据，所述训练数据包括源发音人频谱包络数据和目标发音人频谱包络数据；

[0062] 第一训练单元，用于根据所述源发音人频谱包络数据训练源发音人模型参数；

[0063] 第二训练单元，用于根据所述目标发音人频谱包络数据训练目标发音人模型参数；

[0064] 第三训练单元，用于根据所述对应关系训练转换模型参数；

[0065] 合并单元，用于将所述源发音人模型参数、转换模型参数、以及目标发音人模型参数进行合并，得到所述源发音人和目标发音人的频谱包络变换模型参数。

[0066] 优选地，所述频谱包络特征转换模块包括：

[0067] 条件概率分布计算单元，用于根据所述频谱包络变换模型计算所述频谱包络特征的条件概率分布；

[0068] 转换特征确定单元，用于根据所述条件概率分布确定转换后的频谱包络特征。

[0069] 本发明实施例提供的实现声音转换的方法及系统，基于频谱包络变换模型将源发音人语音信号的频谱包络特征变换为目标发音人的频谱包络特征，然后，基于变换后的频谱包络特征及基频特征生成目标发音人的语音信号。由于频谱包络特征是从高维频谱中提取出来的，是语音信号最直接、准确的表示，因此可以大大提高频谱变换的有效性和准确性，进而提高声音转换的效果。

附图说明

[0070] 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。

[0071] 图 1 是本发明实施例实现声音转换的方法的流程图；

[0072] 图 2 是本发明实施例中构建频谱包络变换模型的流程图；

[0073] 图 3 是本发明实施例中 RBM 模型结构示意图；

[0074] 图 4 是本发明实施例中源发音人和目标发音人的频谱包络变换模型拓扑结构及参数训练过程示意图；

[0075] 图 5 是本发明实施例中用于模拟源发音人和目标发音人之间的参数传递关系的转换模型参数训练流程图；

[0076] 图 6 是本发明实施例中基于频谱包络变换模型获得转换后的频谱包络特征的流程图；

[0077] 图 7 是本发明实施例实现声音转换的系统的结构示意图；

[0078] 图 8 是本发明实施例中频谱包络变换模型构建模块的结构示意图。

具体实施方式

[0079] 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。

[0080] 由于传统的基于频谱变换的声音转换系统主要采用 GMM 模型模拟源发音人和目标发音人的联合频谱特征空间的概率分布，采取的是低维频谱特征，在从频谱中提取低维的特征过程中丢失了很多的频谱细节信息，直接影响了转换语音的音质。而且，GMM 模型存在过平滑效应，导致了合成语音中的过平滑效应。为此，本发明实施例提供一种实现声音转换的方法及系统，基于频谱包络变换模型将源发音人语音信号的频谱包络特征变换为目标发音人的频谱包络特征，然后，基于变换后的频谱包络特征及基频特征生成目标发音人的语音信号。由于频谱包络特征是从高维频谱包络中提取出来的，是语音信号最直接、准确的表示，因此可以大大提高频谱包络变换的有效性和准确性，进而提高声音转换的效果。

[0081] 如图 1 所示，是本发明实施例实现声音转换的方法的流程图，包括以下步骤：

[0082] 步骤 101，获取源发音人的语音信号。

[0083] 步骤 102，提取所述语音信号的频谱包络特征和基频特征。

[0084] 在具体应用中，可以采用现有的频谱包络提取方法，比如，对语音信号加平滑窗做 FFT 变换等。特别地，在本发明实施例中，对于每帧语音帧，可以提取其上下多帧频谱包络特

征作为当前语音帧的频谱包络特征，比如，以连续的三帧频谱包络为例，当前语音帧的频谱包络特征为：

[0085] $x_t = [x_{t-1}^{(s)^T}, x_t^{(s)^T}, x_{t+1}^{(s)^T}]$。

[0086] 步骤 103，根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征。

[0087] 步骤 104，根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号。

[0088] 需要说明的是，在进行语音合成时，所述基频特征也需要进行一定的转换，具体转换方式可以采用现有的一些转换方式，比如，均值方差规整法等，对此本发明实施例不做限定。

[0089] 不同于传统的声音转换系统中的基于 GMM 模型的联合概率分布，在本发明实施例中，基于频谱包络变换模型实现对源发音人语音信号频谱包络的变换，以提高变换后的频谱包络的准确性。

[0090] 下面对本发明实施例中构建频谱包络变换模型的具体过程进行详细说明。

[0091] 如图 2 所示，是本发明实施例中构建频谱包络变换模型的流程图，包括以下步骤：

[0092] 步骤 201，获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据。

[0093] 步骤 202，提取所述训练语音数据的频谱包络特征。

[0094] 具体地，需要分别提取对应相同文本的源发音人语音和目标发音人语音的频谱包络特征。

[0095] 步骤 203，确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系。

[0096] 由于相同语料不同发音人的语音时长可能并不一致，因此在得到源发音人语音和目标发音人语音的频谱包络特征后，需要对不同时长的特征对进行对齐，得到一一对应的频谱包络特征对。

[0097] 考虑到频谱包络特征的维数太高，计算复杂度过高，此外频谱包络太过精细，其距离并不能反映真实频谱的差异性。为此，在本发明实施例中，可以选取频谱包络特征中的任意一种特征进行动态规划对齐，下面以 MCEP(Mel Cepstrum，美尔倒谱）特征为例说明获取频谱包络对应关系的详细过程。

[0098] 首先，提取语音信号的美尔倒谱特征，具体可由美尔域对数功率谱经过逆 FFT 变换得到；然后，对于每帧语音帧，按照其 MCEP 特征对齐源发音人和目标发音人的 MCEP 特征序列，由于频谱包络与 MCEP 是一一对应的，根据 MCEP 特征序列的对应关系，即可得到频谱包络序列的对应关系。

[0099] 步骤 204，确定源发音人和目标发音人的频谱包络变换模型拓扑结构。

[0100] 在本发明实施例中，可以分别采用 RBM(Restricted Boltzmann Machine，受限波尔兹曼机）模型模拟源发明人和目标发音人频谱包络分布特点，为了描述方便，将其分别称为第一 RBM 模型（也可称为源发明人模型）和第二 RBM 模型（也可称为目标发音人模型）。RBM 也可以被视为一个无向图模型，如图 3 所示，其中，v 为可视层，用于表示观测数

据，h 为隐含层，W 为两层之间的连接权重。

[0101] 建立 BBAM(Bernoulli Bidirectional Associative Memory，伯努利双向联想记忆器）模型，所述 BBAM 模型用于模拟源发音人和目标发音人之间的参数传递关系。

[0102] 然后，将上述三个模型，即第一 RBM 模型、BBAM 模型、第二 RBM 模型进行拼接，得到源发音人和目标发音人的频谱包络变换模型拓扑结构，如图 4 所示。

[0103] 其中，第一 RBM 模型为源发音人的模型拓扑，包含频谱包络变量 x 和隐变量 h_x，第二 RBM 模型为目标发音人的模型拓扑，包含频谱包络变量 y 和隐变量 h_y，W_x 为 x 和 h_x 之间的连接权重，W_y 为 y 和 h_y 之间的连接权重，W_h 为 h_x 和 h_y 之间的连接权重。

[0104] 在该拼接模型中，通过源发音人的 RBM 模型可以得到源发音人频谱包络的隐变量表示，通过目标发音人的 RBM 模型可以得到目标发音人频谱包络的隐变量表示，然后使用 BBAM 建立起两个发音人隐变量的联合分布，从而建立起两个发音人频谱包络之间的转换关系。

[0105] 需要说明的是，在实际应用中，也可以用更深层次的网络替代上述 RBM 模型及 BBAM 模型，如图 3 中两个 RBM 可以换成两个更深层的随机神经网络，如 DBN(Deep Belief Network，深度置信网络）或 DBM(Deep Boltzmann Machie，深层波尔兹曼机），DBN 和 DBM 可以由多个 RBM 级联得到，以形成更深层次的网络）。

[0106] 步骤 205，根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数。

[0107] 在参数训练过程中，首先需要分别独立训练源发音人及目标发音人的模型参数。下面以源发音人模型参数训练为例进行详细说明。

[0108] 如图 4 所示，对于源发音人模型拓扑，包含频谱包络变量 x 和一个隐变量 h_x。在本发明实施例中，可以采用一个全局模型模拟源发音人声学空间中的频谱包络参数概率分布，其描述的概率分布为：

[0109] $$P(x)=\frac{1}{Z}\sum_{h_x}e^{-E(x,h_x)}$$

[0110] 其中：$Z=\sum_{h_x}\int_x e^{-E(x,h_x)}dx$ 为配分函数，

[0111] $E(x,h_x)=(x-b_x)^{\mathrm{T}}\Sigma_x^{-1}(x-b_x)-h_x{}^{\mathrm{T}}b_{h_x}-(\Sigma_x^{-\frac{1}{2}}x)^{\mathrm{T}}W_xh_x$ 为该模型的一个能量函数，

[0112] Σ_x 为训练数据的对角协方差矩阵。

[0113] 源发音人的模型参数为 $\theta_x=\{W_x,b_x,b_{h_x}\}$。其中 W_x 为 x 与 h_x 之间的连接权重，b_x、b_{h_x} 分别为 x 层和 h_x 层的偏置。

[0114] 模型的训练准则是使模型达到一个稳态，也就是能量达到最低，对应到概率模型上就是似然值最大化。RBM 的模型参数可以通过 CD(Contrastive Divergence，最小对比散度）算法来高效地训练得到。此外，DBN 和 DBM 的模型参数则可以使用多个 RBM 级联得到，训练过程以无监督的形式进行。

[0115] 目标发音人的模型参数训练过程与上述类似，训练一个描述目标发音人的频谱包络参数概率分布。训练得到的目标发音人的模型参数为 $\theta_y=\{W_y,b_y,b_{h_y}\}$。

[0116] 在得到源发音人的模型参数及目标发音人的模型参数后，需要训练源发音人和目标发音人之间的参数传递关系模型参数。在本发明实施例中，可以采用有监督训练方法来获取源发音人和目标发音人频谱包络之间的映射关系。如图 4 中所示，利用 BBAM 来对两个发音人相关模型的隐变量的联合分布进行建模，从而得到两个发音人频谱包络之间的映射关系。

[0117] 图 4 中 BBAM 描述的概率分布为：

[0118] $$P(h_x,h_y)=\frac{1}{Z}e^{-E(h_x,h_y)}$$

[0119] 其中：$Z=\sum_{h_x,h_y}e^{-E(h_x,h_y)}$ 为配分函数；

[0120] $E(h_x,h_y)=-h_x^{\mathrm{T}}b_{h_x}-h_y^{\mathrm{T}}b_{h_y}-h_y^{\mathrm{T}}W_h h_x$ 为该模型的能量函数。

[0121] 模型的参数为 $\theta_h=\{W_h\}$。

[0122] 如图 5 所示，是本发明实施例中用于模拟源发音人和目标发音人之间的参数传递关系的转换模型参数训练流程图，包括以下步骤：

[0123] 步骤 501，获取转换模型参数训练数据。

[0124] 不同于源发音人和目标发音人的模型参数训练时训练数据的获取情况，在训练源发音人和目标发音人之间的参数传递关系模型参数时，训练数据可以从频谱包络中使用相应的 RBM 模型提取得到。

[0125] 由图 4 所示可知，在本发明实施例中，将源发音人和目标发音人频谱包络之间的转换关系转化为其对应模型的隐变量之间的转换关系来间接地建模，该模型用于模拟隐性的映射关系，其隐变量是假想的，并没有原始的训练数据。所述隐变量可以根据原始的频谱包络（即源发音人和目标发音人的频谱包络）及训练得到的源发音人和目标发音人相关模型中采样得到。

[0126] 比如，给定源发音人的一帧频谱包络 x，可以从下式描述的概率分布中以随机采样的方式得到对应的隐变量 h_x：

[0127] $$h_x\sim p(h_x=1\mid x,\theta_x)=g(W_x^{\mathrm{T}}\Sigma_x^{-\frac{1}{2}}x+b_{h_x})$$

[0128] 其中，$g(x)=1/(1+e^{x})$，所述采样可以以均值采样的形式进行，即 $h_x=g(W_x^{\mathrm{T}}\Sigma_x^{-\frac{1}{2}}x+b_{h_x})$。

[0129] 采用同样的方式，可以得到目标发音人对应的隐变量 h_y。

[0130] 这样，得到源发音人对应的隐变量 h_x 和目标发音人对应的隐变量 h_y，并将其作为所述转换模型的训练数据。

[0131] 步骤 502，基于所述转换模型参数训练数据训练转换模型参数。

[0132] 针对图 4 所示的 BBAM 模型，模型参数为 $\theta_h=\{W_h\}$。在本发明实施例中，可以采用梯度下降算法来训练更新该模型参数，具体训练过程如下：

[0133] (1) 采用高斯随机数初始化模型参数 W_h^0。

[0134] (2) 计算参数 W_h 的梯度 ∂W_h：

[0135] $\partial W_h = E_d[h_x h_y^{\mathrm{T}}] - E_m[h_x h_y^{\mathrm{T}}]$

[0136] 其中，$E_d[\cdot]$ 表示在数据分布上计算的期望，可以通过训练样本即转换模型参数训练数据计算得到，即 $E_d[h_x h_y^{\mathrm{T}}] = h_x h_y^{\mathrm{T}}$。

[0137] $E_m[\cdot]$ 表示在真实分布上计算的期望，可以通过 Gibbs 采样算法从训练样本即转换模型参数训练数据中采样得到，具体采样过程如下：

[0138] 首先，根据转换模型参数训练数据，得到初始样本 h_x^0、h_y^0；然后，给定 h_y^0，从条件概率 $P(h_x^1 = 1 | h_y^0) = g(W_h^{\mathrm{T}} h_y^0 + b_{h_y})$ 中进行采样，得到样本 h_x^1；给定 h_x^1，从条件概率 $P(h_y^1 = 1 | h_x^1) = g(W_h h_x^1 + b_{h_x})$ 中进行采样，得到样本 h_y^1；使用 h_x^1 和 h_y^1 近似计算 $E_m[h_x h_y^{\mathrm{T}}] = h_x^1 h_y^{1\mathrm{T}}$。

[0139] (3) 利用计算得到的梯度 ∂W_h 更新模型参数，即：

[0140] $W_h + \alpha \partial W_h \rightarrow W_h$，其中，α 为更新的步长。

[0141] (4) 判断是否结束训练；如果是，则执行步骤 (5)；否则转入步骤 (2)。

[0142] 其中训练结束的条件可以根据应用需要预先设置，比如可以是迭代次数超过设定的次数阈值，或者是模型参数更新似然值增长幅度超过设定的幅度阈值等。

[0143] (5) 结束。

[0144] 基于上述构建的频谱包络变换模型对从源发音人的语音信号中提取的频谱包络特征进行转换，得到转换后的频谱包络特征，具体流程如图 6 所示，包括以下步骤：

[0145] 步骤 601，对从源发音人的语音信号中提取的频谱包络特征，计算其输出的条件概率分布。

[0146] 为了简化计算，提高运算效率，在实际应用中，可以将所述条件概率分布近似为一个单高斯分布，即：

[0147]
$$P(y_t | x_t) \simeq P(y_t | h_y^*, \theta_y) = N(y_t; \mu_t, \Sigma_y)$$

[0148] 该分布完全由目标发音人模型确定，其中：

[0149] x_t 为输入的频谱包络，y_t 为输出的频谱包络；

[0150] $\mu_t = \Sigma_y^{\frac{1}{2}} (W_y^{\mathrm{T}} h_y^* + b_y)$；

[0151] $h_y^* = \arg\max_{h_y} P(h_y | h_x^*, \theta_h)$；

[0152] $h_x^* = \arg\max_{h_x} P(h_x | x_t, \theta_x)$；

[0153] 在本发明实施例中，对任意的输入，输出的条件单高斯分布共享相同的方差，即所有目标频谱包络训练数据的对角方差。

[0154] 步骤 602，根据所述条件概率分布确定转换后的频谱包络特征。

[0155] 具体地，可以使用最大似然准则从步骤 601 中得到的条件概率分布中得到转换的单帧序列，即：

[0156] $$y^{(s)*} = \arg\max_{y^{(s)}} \prod_{t} P(y_t \mid x_t)$$

[0157] 然后可以求解得到转换的频谱包络。

[0158] 然后再根据上下相关的频谱包络特征$y_t = [y_{t-1}^{(s)T}, y_t^{(s)T}, y_{t+1}^{(s)T}]^T$，获取静态频谱包络$y_t^{(s)T}$，作为转换后的频谱包络特征。

[0159] 本发明实施例实现声音转换的方法，基于频谱包络变换模型将源发音人语音信号的频谱包络特征变换为目标发音人的频谱包络特征，然后，基于变换后的频谱包络特征及基频特征生成目标发音人的语音信号。由于频谱包络特征是从高维频谱包络中提取出来的，是语音信号最直接、准确的表示，因此可以大大提高频谱包络变换的有效性和准确性，进而提高声音转换的效果。

[0160] 相应地，本发明实施例还提供一种实现声音转换的系统，如图 7 所示，是该系统的一种结构示意图。

[0161] 在该实施例中，所述系统包括：

[0162] 语音信号获取模块 701，用于获取源发音人的语音信号；

[0163] 特征提取模块 702，用于提取所述语音信号的频谱包络特征和基频特征；

[0164] 频谱包络特征转换模块 703，用于根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征；

[0165] 语音生成模块 704，用于根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号。

[0166] 上述特征提取模块 702 可以采用现有的频谱包络提取方法，比如，对语音信号加平滑窗做 FFT 变换等。特别地，在本发明实施例中，对于每帧语音帧，可以提取其上下多帧频谱包络特征作为当前语音帧的频谱包络特征。

[0167] 上述频谱包络特征转换模块 703 需要利用预先构建的频谱包络变换模型对所述频谱包络特征进行转换，为此，在本发明实施例的系统中，还可进一步包括：频谱包络变换模型构建模块（未图示），图 8 示出了该频谱包络变换模型构建模块的结构示意图。

[0168] 所述频谱包络变换模型构建模块包括：

[0169] 训练语音数据获取单元 801，用于获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据；

[0170] 特征提取单元 802，用于提取所述训练语音数据的频谱包络特征；

[0171] 对应关系确定单元 803，用于确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系；

[0172] 拓扑结构确定单元 804，用于确定源发音人和目标发音人的频谱包络变换模型拓扑结构；

[0173] 参数训练单元 805，用于根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数。

[0174] 需要说明的是，在本发明实施例中，上述特征提取单元 802 需要分别提取对应相同文本的源发音人语音和目标发音人语音的频谱包络特征。

[0175] 由于相同语料不同发音人的语音时长可能并不一致，因此在上述特征提取单元

802 得到源发音人语音和目标发音人语音的频谱包络特征后，相应地，上述对应关系确定单元 803 需要对不同时长的特征对进行对齐，得到一一对应的频谱包络特征对。具体地，上述对应关系确定单元 803 可以选取频谱包络特征中的任意一种特征进行动态规划对齐，比如 MCEP 特征等。相应地，上述对应关系确定单元 803 的一种具体结构可以包括：第一提取单元、对齐单元和第一确定单元，其中：

[0176] 所述第一提取单元用于以语音帧为单位，提取所述训练语音信号的美尔倒谱特征序列；

[0177] 所述对齐单元用于将源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列对齐；

[0178] 所述第一确定单元用于根据所述源发音人的美尔倒谱特征序列与目标发音人的美尔倒谱特征序列的对应关系，确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系。

[0179] 前面提到，在本发明实施例中，可以分别采用 RBM 模型模拟源发明人和目标发音人频谱包络分布特点，利用 BBAM 模型模拟源发音人和目标发音人之间的参数传递关系。然后，将上述三个模型，即第一 RBM 模型、BBAM 模型、第二 RBM 模型进行拼接，得到源发音人和目标发音人的频谱包络变换模型拓扑结构。

[0180] 相应地，上述拓扑结构确定单元 804 可以包括以下各单元：

[0181] 源发音人模型单元，用于利用第一 RBM 模型模拟源发音人频谱包络分布特点，并将其作为源发音人模型；

[0182] 目标发音人模型单元，用于利用第二 RBM 模型模拟目标发音人频谱包络分布特点，并将其作为目标发音人模型；

[0183] 转换模型单元，用于利用 BBAM 模型模拟源发音人和目标发音人之间的参数传递关系，并将其作为转换模型；

[0184] 拼接单元，用于拼接所述源发音人模型、转换模型、以及目标发音人模型，得到源发音人和目标发音人的频谱包络变换模型拓扑结构。

[0185] 相应地，上述参数训练单元 805 可以包括以下各单元：

[0186] 获取频谱包络训练数据单元，用于获取训练数据，所述训练数据包括源发音人频谱包络数据和目标发音人频谱包络数据；

[0187] 第一训练单元，用于根据所述源发音人频谱包络数据训练源发音人模型参数；

[0188] 第二训练单元，用于根据所述目标发音人频谱包络数据训练目标发音人模型参数；

[0189] 第三训练单元，用于根据所述对应关系训练转换模型参数；

[0190] 合并单元，用于将所述源发音人模型参数、转换模型参数、以及目标发音人模型参数进行合并，得到所述源发音人和目标发音人的频谱包络变换模型参数。

[0191] 基于上述构建的频谱包络变换模型，上述频谱包络特征转换模块 703 对从源发音人的语音信号中提取的频谱包络特征进行转换，得到转换后的频谱包络特征。上述频谱包络特征转换模块 703 的一种具体结构包括：条件概率分布计算单元和转换特征确定单元。其中：所述条件概率分布计算单元于根据所述频谱包络变换模型计算所述频谱包络特征的条件概率分布；所述转换特征确定单元用于根据所述条件概率分布确定转换后的频谱包络

特征。具体的计算过程可参照前面本发明方法实施例中的描述，在此不再赘述。

[0192] 本发明实施例实现声音转换的系统，基于频谱包络变换模型将源发音人语音信号的频谱包络特征变换为目标发音人的频谱包络特征，然后，基于变换后的频谱包络特征及基频特征生成目标发音人的语音信号。由于频谱包络特征是从高维频谱包络中提取出来的，是语音信号最直接、准确的表示，因此可以大大提高频谱包络变换的有效性和准确性，进而提高声音转换的效果。

[0193] 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，对于系统实施例而言，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。以上所描述的系统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。

[0194] 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设备；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。

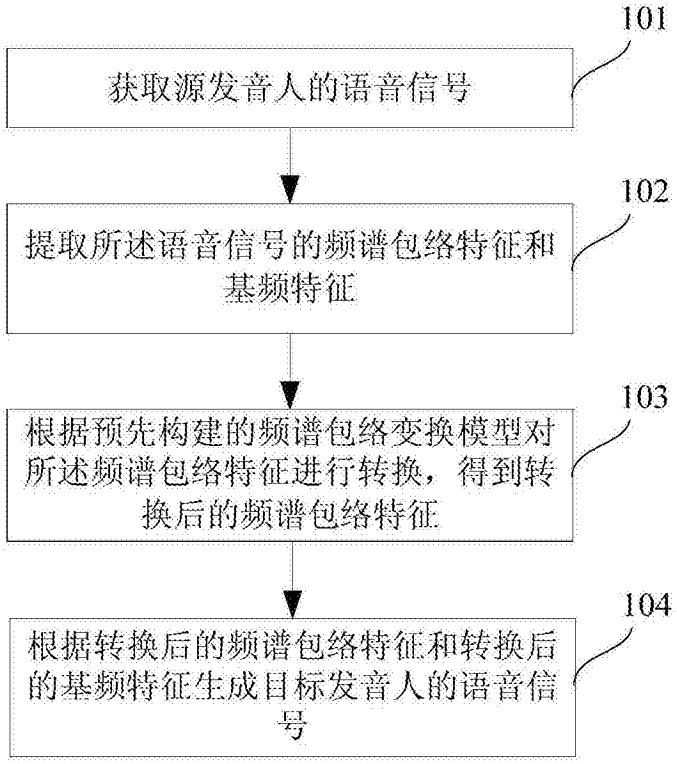

101
获取源发音人的语音信号
102
提取所述语音信号的频谱包络特征和基频特征
103
根据预先构建的频谱包络变换模型对所述频谱包络特征进行转换，得到转换后的频谱包络特征
104
根据转换后的频谱包络特征和转换后的基频特征生成目标发音人的语音信号

图 1

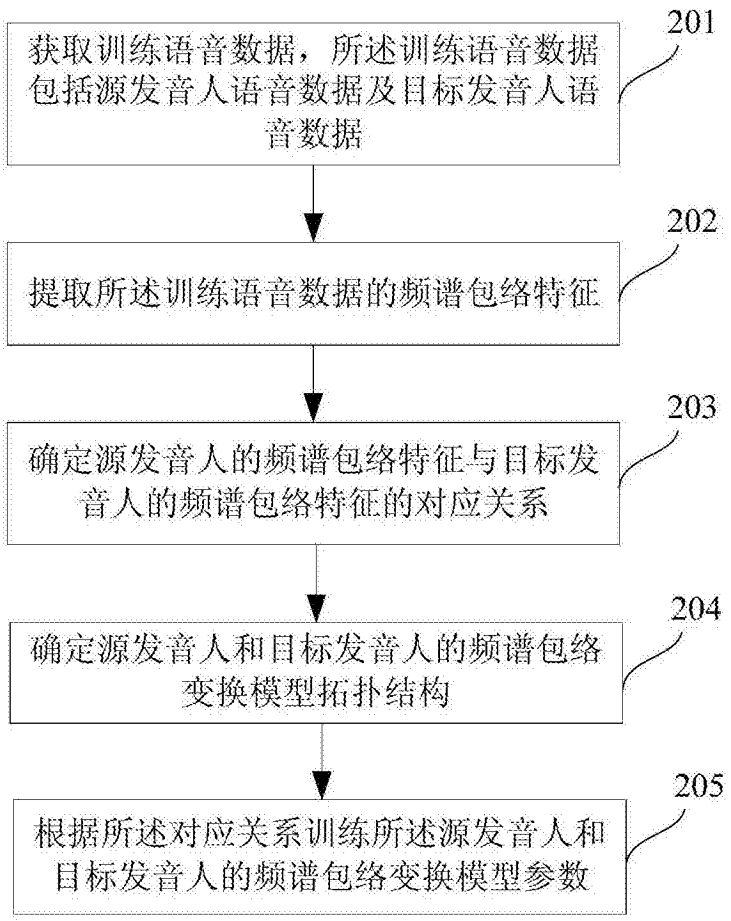

获取训练语音数据，所述训练语音数据包括源发音人语音数据及目标发音人语音数据
201
提取所述训练语音数据的频谱包络特征
202
确定源发音人的频谱包络特征与目标发音人的频谱包络特征的对应关系
203
确定源发音人和目标发音人的频谱包络变换模型拓扑结构
204
根据所述对应关系训练所述源发音人和目标发音人的频谱包络变换模型参数
205

图 2

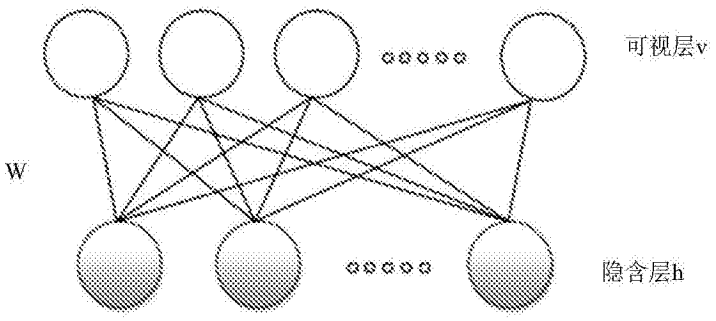

可视层v
W
隐含层h

图 3

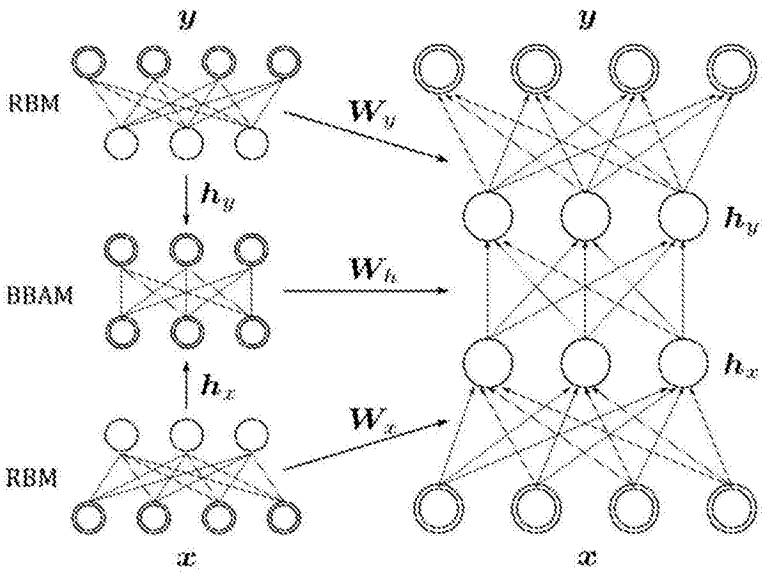

y
RBM
W_y
h_y
BBAM
W_h
h_x
RBM
W_x
x
y
h_y
h_x
x

图 4

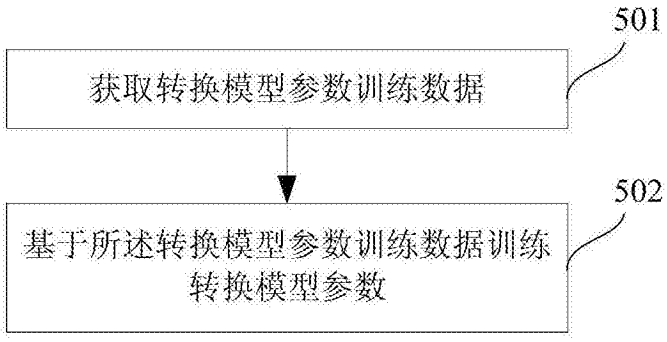

501
获取转换模型参数训练数据
502
基于所述转换模型参数训练数据训练转换模型参数

图 5

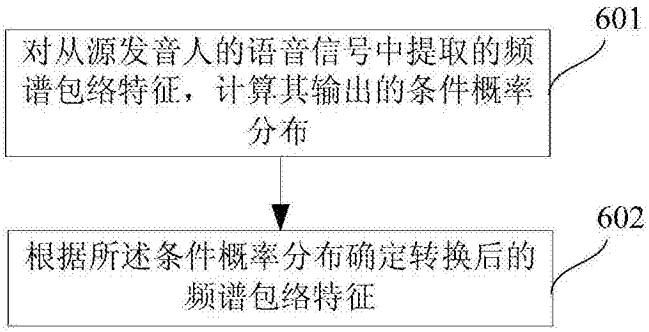

601
对从源发音人的语音信号中提取的频谱包络特征，计算其输出的条件概率分布
602
根据所述条件概率分布确定转换后的频谱包络特征

图 6

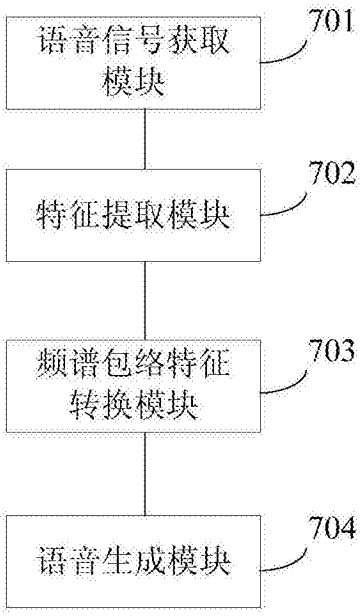

语音信号获取模块
701
特征提取模块
702
频谱包络特征转换模块
703
语音生成模块
704

图 7

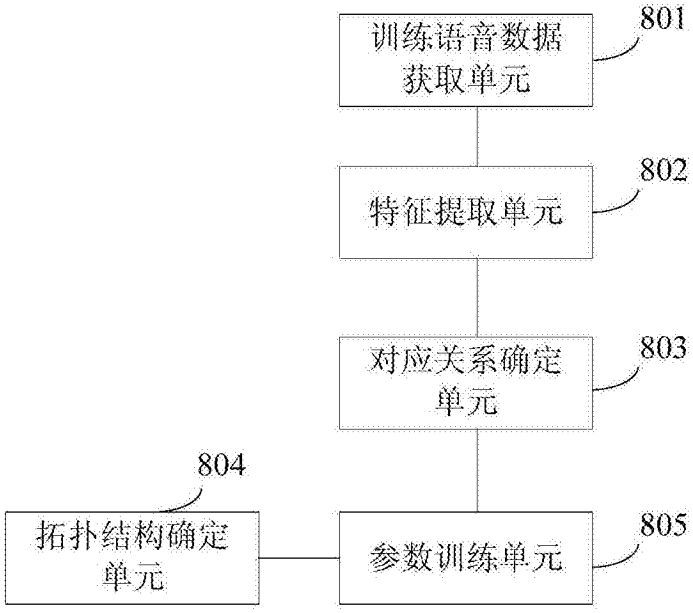

训练语音数据获取单元
801
特征提取单元
802
对应关系确定单元
803
804
拓扑结构确定单元
参数训练单元
805

图 8