

一种声音变换系统以及声音变换系统的训练方法

本公开要求在 2020 年 10 月 21 日提交中国专利局、申请号为 202011129857.5、发明名称为“一种声音变换系统、方法及应用”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及语音计算算法领域，尤其涉及一种声音变换系统以及声音变换系统的训练方法。

背景技术

随着计算机技术的不断发展，人工智能领域的不断深耕，以语音交互为目的的语音机器人逐渐进入大众视野当中。语音机器人的出现改变了现有电话业务的工作性质，目前语音机器人应用房产、教育、金融、旅游等行业中执行语音交互的功能，从而代替人工与用户进行语音交互。

为优化客户体验，利用语音转换技术变换语音机器人声音特征是其中一个重要改进方向。

语音转换技术是语音信号处理的一个研究分支，它涵盖了说话人识别、语音识别及语音合成等领域的内容，拟在保留原有的语义信息不变的情况下改变语音的个性化信息，使特定说话人(即源说话人)的语音听起来像另一个特定说话人(即目标说话人)的语音。语音转换技术的研究经过多年发展，语音转换领域已经涌现出多种不同的方法，其中以高斯混合模型为代表的统计转换方法已经成为该领域中的经典方法。但是这类算法还是存在某些缺陷，比如：使用高斯混合模型来进行语音转换的经典方法多是基于一对一的语音转换任务，要求使用平行文本，也即源说话人和目标说话人使用的训练语句内容相同，需将训练语句的频谱特征进行动态时间规整(Dynamic Time Warping, DTW)逐帧对齐，才能通过模型训练得到频谱特征间的映射关系，而非平行语料的文本是不平行的，也就不能使用上述高斯混合模型来进行语音转换。

发明内容

为解决上述问题，本公开实施例提出一种非平行语料训练的语音转换方案，摆脱对平行文本的依赖，可以在小样本下实现变声效果，解决在资源、设备有限的条件下难以实现语音转换的技术问题。

本公开实施例采用如下技术方案：

第一方面，提供一种声音变换系统，包括：

说话人独立的语音识别模型，至少包括瓶颈层，所述说话人独立的语音识别模型配置为，将输入的源语音的梅尔倒谱特征通过所述瓶颈层变换为源语音的瓶颈特征；

注意力变声网络，其配置为，将源语音的瓶颈特征变换为与目标语音相符的梅尔倒谱特征；

神经网络声码器，其配置为，将与目标语音相符的梅尔倒谱特征转化为语音输出。

第二方面，提供一种声音变换方法，包括：

把源语音的梅尔倒谱特征变换为源语音瓶颈特征；

把源语音的瓶颈特征变换为与目标语音相符的梅尔倒谱特征；

将与目标语音相符的梅尔倒谱特征转化为语音输出。

5 第三方面，提供一种说话人独立的语音识别模型的训练方法，包括：

把多人语音识别训练语料中的文字转换成的字符编码的编号和多人语音识别训练语料的梅尔倒谱特征一起输入说话人独立的语音识别模型，运行反向传播算法进行迭代优化，直到说话人独立的语音识别模型收敛。

第四方面，提供一种注意力变声网络的训练方法，包括：

10 将目标语音的梅尔倒谱特征变换为目标语音的瓶颈特征；

把目标语音的瓶颈特征输入基础注意力变声网络，以目标说话人对应的梅尔倒谱特征作为真实值，用深度迁移学习的方法训练注意力变声网络。

第五方面，提供一种神经网络声码器的训练方法，包括：

15 把目标语音的梅尔倒谱特征和目标语音的声音信号输入预训练神经网络声码器，用深度迁移学习的方法训练神经网络声码器。

第六方面，提供一种终端，包括前述的声音变换系统。

第七方面，提供一种计算机可读的存储介质，所述计算机可读的存储介质中存储有计算机程序，其中，所述计算机程序被设置为运行时执行前述方法之一。

第八方面，提供一种声音变换系统，其中，包括：

20 语音识别模型，至少包括瓶颈层，所述语音识别模型被配置为：将输入的源语音的梅尔倒谱特征，通过所述瓶颈层转换为瓶颈特征；并将所述瓶颈特征通过所述瓶颈层输入至注意力变声网络；

所述注意力变声网络被配置为：将输入的瓶颈特征转换为与目标语音相符的梅尔倒谱特征；

25 神经网络声码器，所述神经网络声码器被配置为：将与目标语音相符的梅尔倒谱特征转换为仿目标语音并输出；所述仿目标语音为源语音通过转换生成的音频。

第九方面，提供一种声音变换系统的训练方法，应用于前述第五方面所述的声音变换系统，其中，包括：

30 根据多人语音识别训练语料中的文字与字符编码之间的映射关系，将所述多人语音识别训练语料中的文字转换为编号；

将转换而得的编号和所述多人语音识别训练语料的梅尔倒谱特征输入至所述语音识别模型；

运行反向传播算法；

进行迭代优化至所述语音识别模型收敛，以训练所述语音识别模型。

35 本公开实施例通过以上技术方案，摆脱对平行文本的依赖，实现任意说话人对多说话人的转换，提高了灵活性，解决在资源、设备有限的条件下难以实现语音转换的技术问题。

具体来说：

40 1、训练好的说话人独立的语音识别模型可用于任意的源说话人，即说话人独立（Speaker-Independent, SI）；说话人独立的语音识别模型训练方法只需要训练一次，后

续小样本只需要用训练好的模型提取对应的特征，即可对声音实现实时变换，实现实时变换声音的需求。

2、音频的瓶颈（Bottleneck）特征相比语音后验概率(Phonetic PosteriorGram, PPG)特征，更加抽象，既能反映说话内容和说话人音色解耦，同时又和音素类别的绑定没有那么紧密，不是明确的一一对应关系，所以一定程度上能缓解 ASR（Automatic Speech Recognition 语音自动识别技术）识别错误导致发音不准的问题。在实际测试时，使用 Bottleneck 特征作声音变换得到的音频，发音的准确率明显高于 PPG 方法，并且音色没有显著区别。

3、本公开实施例实现了声音变换的快速训练，与一般变声网络相比，数据量需求明显降低，本公开实施例系统训练时间可缩短至 5 到 20 分钟，大大减少训练语料的依赖，系统实用性明显增强。

附图说明

为了更清楚地说明本公开的技术方案，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，对于本领域普通技术人员而言，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 为本公开实施例说话人独立的语音识别模型（SI-ASR 模型）训练流程图；

图 2 为本公开实施例注意力变声网络（Attention 变声网络）训练流程图；

图 3 为本公开实施例神经网络声码器训练流程图；

图 4 为本公开实施例声音变换流程图；

图 5 为本公开实施例注意力变声网络（Attention 变声网络）架构图；

图 6 为本公开实施例神经网络声码器网络架构图；

图 7 为本公开实施例说话人独立的语音识别模型（SI-ASR 模型）网络架构图。

具体实施方式

以下将结合附图和具体实施例对本公开进行详细说明，显然，所描述的实施例仅仅只是本公开一部分实施例，而不是全部的实施例，基于本公开中的实施例，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他实施例，都属于本公开保护的范围。

本公开实施例提供一种声音变换系统，包括：

（1）说话人独立的语音识别（SI-ASR）模型，采用五层 DNN（Deep Neural Network 深度神经网络）结构，其中第四层使用瓶颈（Bottleneck）层，把源语音的梅尔倒谱特征（MFCC）变换为源语音瓶颈特征（Bottleneck 特征）；

ASR 模型将语音转化为文字，模型会输出音频对应每个词的概率，而 PPG 就是这种概率的载体。基于 PPG 的方法，使用 PPG 作为 SI-ASR 模型的输出。

PPG 即 Phonetic PosteriorGram，它是一个把每个音频时间帧映射为某个音素类别的后验概率的矩阵。PPG 在一定程度上可以表示一段语音说话内容的节奏和韵律信息，同时又去掉了和说话人音色相关的特征，因此它是说话人独立的。PPG 定义如下：

$$P_t = (p(s|X_t), s=1, 2, 3, \dots, C);$$

其中，C 是音素个数；s 是一个音素（数字表示）； X_t 是音频第 t 帧的 MFCC 特征；

$P(s|X_t)$ 是音素 s 的后验概率。

实践中发现,虽然 PPG 特征可以去掉说话人的音色特征,但由于 ASR 模型对文字的识别存在一定的错误率,模型输出的后验概率有可能是不准确的。这会导致最终变换输出的音频出现个别字发音不准,或者发音错误的情况,甚至会出现噪声等情况。

5 针对这个问题,本公开实施例说话人独立的语音识别模型设置包括瓶颈层,将输入的源语音的梅尔倒谱特征通过所述瓶颈层变换为源语音的瓶颈特征,并将所述源语音瓶颈特征由所述瓶颈层输出至注意力变声网络。

Bottleneck 特征的提取和 ASR 模型构建有关系,不同的 ASR 模型对应的深度也不同,以五层 DNN 神经网络结构为例,可以设置其中某一层使用 Bottleneck 层(瓶颈层),即
10 一层 Bottleneck 层(DNN)放在四层 DNN 结构中组成 ASR 模型,为取得较佳效果,优选 Bottleneck 层置于 ASR 模型的第三层或第四层的输出较佳,即将第三层或第四层的输出作为 Bottleneck 特征。需要说明的是,本公开实施例中优选 Bottleneck 层置于 ASR 模型的第四层,以避免其余位置可能的音色信息的干扰。

以下作为本公开实施例的一个优选实施方案予以详述,不作为对本公开实施例保护
15 范围的限定,即以本公开实施例权利要求确定的保护范围内,其他仍可解决本公开实施例技术问题。

本实施例采用五层 DNN 结构,其中第四层使用 Bottleneck 层(瓶颈层),即三层 DNN 结构、一层 Bottleneck 层(DNN)和一层 DNN 结构组成 ASR 模型,模型训练好后,输入音频特征,输出 Bottleneck 层结果作为变声特征。Bottleneck 特征与音频的声学特征
20 不同,是音频的一种语言特征,不包含说话人的音色等信息。用包含大量不同说话人的训练语料训练 ASR 模型,就可以捕获所有说话人的共有信息,去掉说话人的个性信息。Bottleneck 层一般是模型的中间层,但是为了让输出的特征尽量脱离音色信息,本公开实施例改进网络设计,使 Bottleneck 层尽量靠近 ASR 网络的输出。只有这样,提取的特征才不包含音色信息。因此本公开实施例提取 ASR 网络的倒数第二层的输出作为 Bottleneck
25 特征。实践证明这样可以很好地去除音色信息,同时保留语言信息。

Bottleneck 特征相比 PPG 特征更加抽象,既能反映说话内容和说话人音色解耦,同时又和音素类别的绑定没有那么紧密,不是明确的一一对应关系,所以一定程度上能缓解 ASR 识别错误导致发音不准的问题。在实际测试时,使用 Bottleneck 特征作声音变换得到的音频,发音的准确率明显高于 PPG 方法,并且音色没有显著区别。

30 (2) 注意力(Attention)变声网络,把源语音的瓶颈特征变换为与目标语音相符的梅尔倒谱特征;

Attention 变声网络基于 seq2seq(端对端)架构。主要改进点在于:第一、利用一层双向 RNN(Bidirectional Recurrent Neural Network,循环神经网络)把 SI-ASR 模型输出的 BN 特征编码为高维特征。第二、结合 Attention 机制,将 encode(编码)和 decode(解
35 码)联系起来,避免人工对齐产生的不稳定性。第三、简化 decoder 网络,利用两层 DNN 网络结构,后面接一层 RNN 结构,再利用带残差连接的多层 Self Attention(自注意)作为 Post-Net,将 BN 特征转化为声学特征(详见附图 5)。GRU(Gated Recurrent Unit 门控循环单元)编码是指门控循环单元编码。

相比 PPG 方法,本公开实施例的变声网络使用更简单直接的声学特征。PPG 方法将
40 PPG 转化为声码器所需要的特征,还需要结合 F0(fundamental frequency,基频)和 AP

(aperiodic component, 非周期分量) 特征, 再利用声码器还原音频。因为 F0 特征中包含说话人信息, 虽然可以使变换后的声音更加饱和, 但也牺牲了音色作为代价。而本公开实施例的 Attention 变声网络, 可以直接预测输出所有需要的声码器参数, 而不需要再手动提取 F0 和 AP 等滤波器特征。这样, 一方面大大简化了网络的输入输出和流程设计, 使模型更加简洁高效, 另一方面也使得变声以后的音频更像目标说话人。另外, 本公开实施例的网络规模比较小, 运行速度快, 可以实现实时变声, 目前 10s 音频, 只需要 1s 变换时间, 通过工程化流式封装可以达到实时变声的效果。

(3) 神经网络声码器, 将与目标语音相符的梅尔倒谱特征转化为语音输出。

神经网络声码器采用 WaveRNN 的一种变体, 将声学特征还原为音频输出。本公开实施例将声学特征编码成高维空间中的特征, 然后利用循环神经网络高维特征还原为音频输出, 具体神经网络声码结构参见附图 6, 其中, 2GRU (Gated Recurrent Unit 门控循环单元) 网络为门控循环单元网络, 2 层 FC (Fully Connected 对全连接) 是指 2 个对全连接层, softmax 为归一化指数函数。

本公开实施例还介绍一种声音变换系统训练方法, 包括以下 A1-A3 三个部分:

A1、SI-ASR 模型 (说话人独立的语音识别模型) 训练阶段。该阶段训练得到 Attention 变声网络 (注意力变声网络) 训练阶段的特征和声音转换阶段提取 Bottleneck 特征 (直译为瓶颈特征, 也简称为 BN 特征) 用到的 SI-ASR 模型; 该模型的训练使用包含很多说话人的训练语料训练而成。训练好之后, 可以用于任意的源说话人, 也就是说, 它是说话人独立的 (Speaker-Independent, SI), 因此称为 SI-ASR 模型; 训练好以后, 后面直接使用, 不需要重复训练。

SI-ASR 模型 (说话人独立的语音识别模型) 训练阶段包括以下步骤 (参见附图 1):

B1、对多人 ASR 训练语料进行预处理;

该预处理包括对训练音频去空白和归一化。去空白即检测并裁剪掉音频中过长的停顿、静音 (不包括字词之间正常的停顿)。归一化就是把音频数据统一在一定范围内。

对训练语料中的文本进行清洗校对, 修正文字和音频对应不准确的情况; 对文本进行正则化处理, 将数字、年月日、小数、单位符号等转换为汉字。如果模型训练以词为单位, 就需要调用分词工具 (如 jieba, pkuseg 等) 对文本进行分词。然后把文字转换为拼音和音素。统计语料中出现的所有字、词、拼音或音素, 生成词汇表, 并统一编码为整数表示, 训练以音素为单位。

B2、参数提取, 得到训练语料音频的 MFCC 特征 (梅尔倒谱特征);

梅尔频率倒谱系数 (Mel-scale Frequency Cepstral Coefficients, 简称 MFCC) 是语音识别和语音合成领域最常用到的语音特征之一。这种特征不依赖于信号的性质, 对输入信号不做任何的假设和限制, 有很好的鲁棒性, 又利用了听觉模型的研究成果, 更符合人耳的听觉特性, 而且当信噪比降低时仍然具有较好的识别性能。MFCC 的提取过程包括以下步骤: 预加重、分帧、加窗、快速傅里叶变换、梅尔滤波、对数运算、离散余弦变换。

B3、SI-ASR 模型训练。

SI-ASR 模型的训练使用 Kaldi 框架。Kaldi 是一个研究用途的开源语音识别框架。模型采用基于深度神经网络的 (Deep Neural Network, DNN) 架构。

把训练语料中的文字转换成字符编码的编号, 和音频 MFCC 特征一起输入 SI-ASR

模型，运行反向传播算法进行迭代优化，直到模型收敛。

说话人独立的语音识别模型（SI-ASR 模型）网络架构见图 7。

5 A2、Attention 变声网络（注意力变声网络）训练阶段。该阶段训练得到基于 Sequence-to-sequence（端对端）架构和 Attention 机制的变声网络（后面简称 Attention 变声网络），用于声音转换阶段把 SI-ASR 模型提取的 BN 特征变换为声码器所需声学特征。该阶段的变声网络需要针对不同目标说话人分别进行训练。训练好以后，可以将任意源说话人的声音音色转换为目标说话人音色；

Attention 变声网络训练阶段包括以下步骤（参见附图 2）：

10 C1、对目标说话人的目标语音进行预处理；
训练语音预处理包括降噪，去空白，音量归一化等。音频去空白，归一化和 B1 步骤相同。降噪利用现有的降噪模型对训练语音进行处理，减小噪声的影响。

变换一个人的声音需要大约 5~20min，50~200 句录音，大大简化了繁琐的录音工作，而且变声不需要文本校验，所以这些训练音频既用于训练 Attention 变声网络，也用于训练神经网络声码器。

15 C2、参数提取，得到目标语音的 MFCC 特征和声学特征；

MFCC 特征的提取和 B2 步骤相同。声学特征即神经网络声码器需要的特征。目前本公开实施例使用梅尔谱作为声学特征，为了更贴近人耳感知，本公开实施例利用梅尔谱作为声学特征。

20 C3、使用 B3 中训练好的 SI-ASR 模型把 MFCC 特征变换为 BN 特征；

C4、加载用大规模语料训练好的基础 Attention 变声网络；

C5、把 BN 特征输入基础 Attention 变声网络，以声学特征作为真实值 ground truth，用深度迁移学习的方法训练模型。

25 理论上各种语言的发音差距不是很大，因此可以使用深度迁移学习方法，复用以前训练好的网络参数和学习到的特征，大大降低了模型训练的难度、训练数据需求量和采集成本。

A3、神经网络声码器训练阶段。该阶段训练得到基于深度神经网络的声码器（Deep Neural Network Vocoder，即神经网络声码器），用于把声学特征变换为目标语音信号。

声码器训练阶段包括以下步骤（参见附图 3）：

30 D1、对目标说话人的目标语音进行预处理；

这一步和 C1 中的预处理操作相同。

D2、参数提取，得到目标语音的声学特征；

不同的声码器使用不同的声学特征，目前本公开实施例使用梅尔谱作为声学特征。

D3、加载预训练声码器模型；

35 D4、把声学特征和目标语音的声音信号输入预训练神经网络声码器模型，用深度迁移学习的方法训练模型。

本公开实施例还包括一种声音变换方法。对输入的源语音进行声音变换，变换为目标语音信号输出，也就是声音符合目标说话人声音特点，但说话内容和源语音相同的语音。声音转换阶段包括以下步骤（参见附图 4）：

40 E1、对待转换的源语音进行参数提取，得到 MFCC 特征；

E2、使用 B3 中训练好的 SI-ASR 模型把 MFCC 特征变换为 BN 特征；

E3、使用 C5 中训练好的 Attention 变声网络把 BN 特征变换为声学特征（梅尔谱）；

E4、使用 D4 中训练好的神经网络声码器将声学特征（梅尔谱）转换为语音输出。

通过这一方式，训练好的说话人独立的语音识别模型可用于任意的源说话人，即说话人独立（Speaker-Independent, SI）；说话人独立的语音识别模型训练方法只需要训练一次，后续小样本只需要用训练好的模型提取对应的特征即可。

本公开实施例还包括一种终端，该终端使用实施例一所述的声音变换系统。

所述终端可以为搭载有自动语音应答或提示服务系统的移动终端、PC 设备、可穿戴设备等，或者，可以为具有自动语音应答或提示服务功能的语音机器人等，本公开实施例对此不作限定。

本公开实施例还包括一种计算机可读的存储介质，所述计算机可读的存储介质中存储有计算机程序，其中，所述计算机程序被设置为运行时执行上述实施例 A4 所述方法进行声音变换，模块训练使用上述实施例 A1-A3 所述方法。

本公开还提供一种声音变换系统，其中，包括：语音识别模型，至少包括瓶颈层，所述语音识别模型可以被配置为：将输入的源语音的梅尔倒谱特征，通过所述瓶颈层转换为瓶颈特征；并将所述瓶颈特征通过所述瓶颈层输入至注意力变声网络。

本公开实施例提供的技术方案，应用在需要将源说话人说出的语音（源语音），转换成与目标说话人相匹配的语音的场景，也就是说，所述源语音为转换开始前，源说话人说出的源语音。在实际应用中，所述语音识别模型是由大量训练语料训练而成的，训练后的语音识别模型，可以应用于任一源说话人，也就是说，所述语音识别模型是独立的。

具体的，所述语音识别模型可以包括五层 DNN 结构，其中第三层或者第四层可以为所述瓶颈层。可选地，Bottleneck 层置于 ASR 模型的第四层，以避免其余位置可能的音色信息的干扰。

所述注意力变声网络可以被配置为：将输入的瓶颈特征转换为与目标语音相符的梅尔倒谱特征；

在本实施例中，所述目标语音是目标说话人说出的语音。因此，为了将源语音转换为与目标语音相符的语音，使用注意力变声网络将源语音的瓶颈特征转换为与目标语音相符的梅尔倒谱特征。

所述注意力变声网络可以包括一层双向 RNN 结构。

本公开提供的技术方案还包括神经网络声码器，所述神经网络声码器可以被配置为：将与目标语音相符的梅尔倒谱特征转换为仿目标语音并输出；所述仿目标语音为源语音通过转换生成的音频。

在实际应用中，所述神经网络声码器能够将梅尔倒谱特征转换成音频输出，产生的音频即为仿目标语音。也就是说，所述仿目标语音是与目标语音相符的语音，经过上述过程的处理，实现了声音变换。

本公开还提供一种声音变换系统的训练方法，应用于上述实施例所述的声音变换系统，其中，包括：

S1：根据多人语音识别训练语料中的文字与字符编码之间的映射关系，将所述多人语音识别训练语料中的文字转换为编号。

在本公开技术方案中，所述字符编码是指把输入任意字符转换为固定形式的一种转

换形式，注意：汉字需要转化为对应的音素，以便可以进行神经网络计算，所述字符编码可以采用 ASCII 码，也可以采取其他形式，对此本公开不做具体限定。

S2：将转换而得的编号和所述多人语音识别训练语料的梅尔倒谱特征输入至所述语音识别模型；

5 S3：运行反向传播算法；

S4：进行迭代优化至所述语音识别模型收敛，以训练所述语音识别模型。

对语音识别模型进行训练，是建立模型与训练语料之间联系的过程。在本实施例中，所述训练语料的具体文字内容本公开不做限定。

所述方法还包括对所述多人语音识别训练语料进行预处理；预处理包括去空白和归
10 一化。在本实施例中，去空白的过程能够去除训练语料中过长的停顿以及静音等，提高训练语料的质量。归一化的过程即为对训练语料进行音量归一化，训练语料中如果存在音量时而大时而小的情况，会影响训练的效果，因此通过归一化将训练语料的音量控制在一定范围内，具体范围可以根据实际情况进行设计，本公开不做具体限定。

所述方法还包括：

15 S5：将目标语音的梅尔倒谱特征转换为瓶颈特征；

S6：将所述目标语音的瓶颈特征输入至注意力变声网络；将所述目标语音的梅尔倒谱特征作为真实值输入至所述注意力变声网络；

S7：使用深度迁移学习的方法训练所述注意力变声网络。

在本实施例中，所述目标语音是指在训练过程中，目标说话人说出的语音。使用目
20 标语音训练所述注意力变声网络，能够建立目标语音与注意力变声网络之间的联系，以使所述注意力变声网络能够实现将源语音的瓶颈特征转换为与目标语音相符的梅尔倒谱特征。

在本公开部分实施例中，将目标语音的梅尔倒谱特征转换为瓶颈特征的步骤，由预先训练的所述语音识别模型进行。

25 所述方法还包括：

S8：将所述目标语音的梅尔倒谱特征以及所述目标语音输入至神经网络声码器；

S9：使用深度迁移学习的方法训练所述神经网络声码器。

在本实施例中，通过目标语音的梅尔倒谱特征与目标语音的声音信号之间的关系，训练神经网络声码器，以使所述神经网络声码器能够将与目标语音相符的梅尔倒谱特征
30 转换为音频并输出。

所述方法还包括对所述目标语音进行预处理；预处理包括去空白和归一化。

在本实施例中，去空白及归一化，能够避免音频中过长的停顿、静音以及过大或多小的音量等对后续训练注意力变声网络和神经网络声码器产生影响。

所述方法还包括：参数提取，以得到多人语音识别训练语料的梅尔倒谱特征、目标
35 语音的梅尔倒谱特征以及源语音的梅尔倒谱特征。

由以上技术方案可知，本公开提供的声音变换系统及声音变换系统的训练方法，能够以任意源说话人说出的源语音以及目标说话人说出的目标语音为基础，将源语音变换为与目标语音相匹配的音频输出，实用性强。

以上实施例仅用以说明本公开的技术方案而非限制，尽管参照较佳实施例对本公开进行了详细说明，本领域的普通技术人员应当理解，可以对本公开的技术方案进行修改
40

或者等同替换，而不脱离本公开技术方案的宗旨和范围，其均应涵盖在本公开的权利要求范围当中。本公开未详细描述的技术、形状、构造部分均为公知技术。