(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108847249 B
(45)授权公告日 2020.06.05

(21)申请号 201810537499.8

(22)申请日 2018.05.30

(65)同一申请的已公布的文献号
申请公布号 CN 108847249 A

(43)申请公布日 2018.11.20

(73)专利权人 苏州思必驰信息科技有限公司
地址 215123 江苏省苏州市苏州工业园区新平街388号腾飞创新园14栋
专利权人 上海交通大学

(72)发明人 俞凯 陈宽 陈博

(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
代理人 方挺 黄谦

(51)Int.Cl.
G10L 21/003(2013.01)
G10L 21/007(2013.01)
G10L 21/013(2013.01)
G10L 25/24(2013.01)

(56)对比文件
CN 103345923 A,2013.10.09,
CN 104575495 A,2015.04.29,
CN 106504742 A,2017.03.15,
CN 106683666 A,2017.05.17,
CN 106486131 A,2017.03.08,
CN 107293302 A,2017.10.24,
CN 106504741 A,2017.03.15,
CN 101399044 A,2009.04.01,
CN 102982809 A,2013.03.20,
CN 107705802 A,2018.02.16,
Jiahao Lai.Phone-Aware LSTM-RNN for Voice Conversion.《IEEE》.2016,177-182.

审查员 韩金鑫

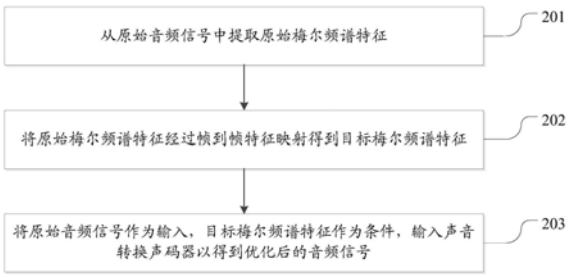
权利要求书2页 说明书10页 附图8页

(54)发明名称

声音转换优化方法和系统

(57)摘要

本发明公开声音转换优化方法和系统,方法包括:从原始音频信号中提取原始梅尔频谱特征;将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征;将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号。本发明提出了一个高品质的音频转换结构,摒弃了声学特征中常用的梅尔倒谱系数和基频F0,转而使用了非常低水平的梅尔频谱图作为声学特征,从而在简化结构和计算的同时还能比现有技术转换出的声音更加自然。



CN 108847249 B

1.一种声音转换优化方法,包括:

从原始音频信号中提取原始梅尔频谱特征;

将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征;

将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号。

2.根据权利要求1所述的方法,其中,所述将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号包括:

将所述目标梅尔频谱特征经过上采样层与所述原始音频信号对齐;

使用所述原始音频信号作为输入,对齐后的所述目标梅尔频谱特征及与所述原始音频信号对应的说话人身份信息作为条件;

将所述输入和所述条件输入多个残差网络,并将得到的所有输出相加,经过卷积层得到优化后的音频信号。

3.根据权利要求1所述的方法,其中,所述将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征包括:

对于具有相同文本的所述原始音频信号和目标音频信号,将来自所述原始音频信号的原始梅尔频谱特征$x=x_1,...,x_m$和来自所述目标音频信号的目标梅尔频谱特征$y=y_1,...,y_n$对齐到相同长度T;

将对齐后的特征序列$x=x_1,...,x_T$和$y=y_1,...,y_T$逐帧转换以得到所述目标梅尔频谱特征。

4.根据权利要求3所述的方法,其中,所述原始梅尔频谱特征和所述目标梅尔频谱特征之间的差距L通过下式获得:

$$\mathcal{L}=\sum_{i=1}^{T}|M_{xy}(x_i)-y_i|_2$$

其中,M_{xy}是从原始音频信号到目标音频信号梅尔频谱转换模型。

5.根据权利要求1-4中任一项所述的方法,其中,所述声音转换声码器为Wavenet声码器,所述Wavenet声码器中使用门控激活函数的机制调节梅尔频谱特征,

$z=\tanh(W_f*i+V_f*c)\odot\sigma(W_g*i+V_g*c)$,

其中,z表示非线性函数,*表示卷积运算符,而⊙表示单元乘法运算符,σ()表示一个S形函数,i表示输入,c表示梅尔频谱特征,f和g分别代表滤波器和门,W和V是可学习的权重。

6.一种声音转换优化系统,包括:

提取单元,配置为从原始音频信号中提取原始梅尔频谱特征;

映射单元,配置为将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征;

优化单元,配置为将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号。

7.一种电子设备,其包括:至少一个处理器,以及与所述至少一个处理器通信连接的存储器,其中,所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1至5任一项所述方法的步骤。

8.一种存储介质,其上存储有计算机程序,其特征在于,所述程序被处理器执行时实现权利要求1至5任一项所述方法的步骤。

声音转换优化方法和系统

技术领域

[0001] 本发明属于声音转换技术领域,尤其涉及声音转换优化方法和系统。

背景技术

[0002] 声音转换(Voice Conversion,VC)是一种技术,用于在保留语言内容的同时,修改源说话人的语音以使其听起来像目标说话人。传统的声音转换技术专注于使用源说话人和目标说话人说相同句子的一些并行数据来开发转换功能。一些转换模型,如高斯混合模型(Gaussian mixture model,GMM),深度神经网络已经被应用于将源说话人的声学特征转换为相应的目标说话人。

[0003] 转换后的语音的音质对研究人员来说总是很有吸引力。相关技术中转换的语音总会出现失真,例如,过度平滑,缺乏相似性等。在参数化声音转换中,已经提出了几种技术来增强音质,例如,建模附加特征(全局变差,频谱包络)和后滤波。然而,转换后的语音质量仍然不如目标说话人那么自然。

[0004] 现有技术中,转换模型的实现方法如图1所示,大多是基于梅尔倒谱的语音转换。使用了Mel-cepstrum(Mcep,梅尔倒谱),F0(基频),带非周期性(BAP,Band Aperiodicity)作为声码器的条件训练一个声码器,然后再使用GMM(高斯混合模型,Gaussian Mixture Model)转换原始说话的人的特征,生成目标说话人的特征,最后用训练好的声码器合成声音。

[0005] 发明人在实现本发明的过程中,发现之所以最后合成的声音不自然,一个重要的因素是用于参数化语音转换的声学特征通常是声码参数(例如,梅尔倒谱,基频F0),当用转换的声码参数产生波形时,其转换可能导致质量失真。

发明内容

[0006] 本发明实施例提供一种声音转换优化方法、系统及电子设备,用于至少解决上述技术问题之一。

[0007] 第一方面,本发明实施例提供一种声音转换优化方法,包括:从原始音频信号中提取原始梅尔频谱特征;将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征;将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号。

[0008] 第二方面,本发明实施例提供一种声音转换优化系统,包括:提取单元,配置为从原始音频信号中提取原始梅尔频谱特征;映射单元,配置为将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征;优化单元,配置为将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号。

[0009] 第三方面,提供一种电子设备,其包括:至少一个处理器,以及与所述至少一个处理器通信连接的存储器,其中,所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行本发明任一实施例

的声音转换优化方法的步骤。

[0010] 第四方面，本发明实施例还提供一种计算机程序产品，所述计算机程序产品包括存储在非易失性计算机可读存储介质上的计算机程序，所述计算机程序包括程序指令，当所述程序指令被计算机执行时，使所述计算机执行本发明任一实施例的声音转换优化方法的步骤。

[0011] 在本申请的方法和系统中，我们提出了一个高品质的音频转换结构，其中，我们摒弃了声学特征中常用的梅尔倒谱系数和基频F0，转而使用了非常低水平的梅尔频谱图作为声学特征，从而在简化结构和计算的同时还能比现有技术转换出的声音更加自然。

附图说明

[0012] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，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。

[0013] 图1为现有技术的转换模型的实现方法的示意图；

[0014] 图2为本发明一实施例提供的一种声音转换优化方法的流程图；

[0015] 图3为本发明一实施例提供的另一种声音转换优化方法的流程图；

[0016] 图4a和图4b为本发明一实施例提供的一种对偶方法的原理图；

[0017] 图5为本发明一实施例提供的一种声音转换优化方法的转换架构图；

[0018] 图6为本发明一实施例提供的一种声音转换优化系统的转换原理图；

[0019] 图7为本发明一实施例提供的一种声音转换优化系统的条件WaveNet的体系结构图；

[0020] 图8为本发明一实施例提供的F0转换语音分布图；

[0021] 图9为本发明一实施例提供的F0轮廓的一个示例；

[0022] 图10为本发明一实施例提供的不同说话人转换语音的语义比较结果；

[0023] 图11为本发明一实施例提供的不同说话人转换语音的自然度比较结果；

[0024] 图12a和图12b为本发明一实施例提供的不同系统与目标说话人相比较的结果；

[0025] 图13为本发明一实施例提供的声音转换优化系统的框图；

[0026] 图14为本发明一实施例提供的电子设备的结构示意图。

具体实施方式

[0027] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0028] 下面，先介绍本申请的实施方式，之后将用实验数据证实本申请的方案与现有技术相比有什么不同，能实现什么有益效果。

[0029] 请参考图1，其示出了本发明的声音转换优化方法一实施例的流程图，本实施例的声音转换优化方法可以适用于各种变声器等声音转换装置。

[0030] 最近，一种用WaveNet语音产生模型的高质量的声码器已经被提出。WaveNet是最先进的自然波形生成技术，可以生成高质量的语音波形。其优点之一是WaveNet语音生成模型能够在特定条件下生成波形，如语言信息或声学特征。它已被应用于许多应用，如文本到语音，声音转换和语音声码器。类似于WaveNet声码器，声学特征主要是梅尔频率倒谱系数(Mel-cepstral，Mcep)和基频(F0)，它们广泛用于语音声码。WaveNet声码转换语音的音质与STRAIGHT-vocoded语音相当。

[0031] 如图2所示，步骤201，从原始音频信号中提取原始梅尔频谱特征；

[0032] 在步骤202中，将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征；

[0033] 在步骤203中，将原始音频信号作为输入，目标梅尔频谱特征作为条件，输入声音转换声码器以得到优化后的音频信号。

[0034] 在本实施例中，对于步骤201，本申请的声音转换优化装置只需要提取原始音频信号的原始梅尔频谱特征(Msp，Mel-spectrogram)。之后，对于步骤202，将该原始梅尔频谱特征进行帧对帧映射转换成目标梅尔频谱特征。最后，对于步骤203，将原始音频信号作为一个输入，目标梅尔倒谱特征作为一个条件，将上述输入和条件输入至声码器中以得到转换优化后的音频信号。

[0035] 本实施例的方法，放弃了现有技术中在进行声音转换时普遍使用的梅尔倒谱系数和基频F0，转而使用了声学特征中非常低水平的梅尔频谱图作为条件，没有了之前F0，BAP转换不够准确的缺陷，从而在简化结构和计算的同时还能比现有技术转换出的声音更加自然。具体改进多少，请参见后续的实验及相关数据，在此不再赘述。

[0036] 在一些可选的实施例中，上述步骤201包括：对于具有相同文本的原始音频信号和目标音频信号，将来自原始音频信号的原始梅尔频谱特征$x=x_1,\ldots,x_m$和来自目标音频信号的的目标梅尔频谱特征$y=y_1,\ldots,y_n$对齐到相同长度T；将对齐后的特征序列$x=x_1,\ldots,x_T$逐帧转换至$y=y_1,\ldots,y_T$以得到目标梅尔频谱特征。从而可以通过上述方式将原始梅尔频谱特征转换成目标梅尔频谱特征。在一些可选的实现中，可以采用GMM或者LSTM NN(Long Short-Term Memory Neural Network，长短时记忆神经网络)。

[0037] 进一步可选的，原始梅尔频谱特征到目标梅尔频谱特征的训练成本L通过下式获得：

[0038] $$\mathcal{L}=\sum_{i=1}^{T}|M_{xy}(\boldsymbol{x}_i)-\boldsymbol{y}_i|_2$$ ，

[0039] 其中，Mxy是从原始音频信号到目标音频信号梅尔频谱转换模型。这个L是表示原始梅尔频谱和目标梅尔频谱之间的差距，L越小，表示目标频谱和原始频谱更加相似，模型的预测能力更准确。

[0040] 在另一些可选的实施例中，声音转换声码器可以采用Wavenet声码器，Wavenet声码器中使用门控激活函数的机制调节梅尔频谱特征，

[0041] $z=\tanh(W_f*i+V_f*c)\odot\sigma(W_g*i+V_g*c)$，

[0042] 其中，*表示卷积运算符，而⊙表示单元乘法运算符。σ()表示一个S形函数，i表示输入，c表示梅尔频谱特征，f和g分别代表滤波器和门，W和V是可学习的权重，z代表一种非线性函数，能够调节模型中的变量。

[0043] 进一步参考图3,其示出了本发明一实施例提供的另一种声音转换优化方法的流程图,本流程图主要是针对图2中步骤203的细化步骤的流程。

[0044] 如图3所示,在步骤301中,将目标梅尔频谱特征经过上采样层与原始音频信号对齐;

[0045] 在步骤302中,使用原始音频信号作为输入,对齐后的目标梅尔频谱特征及与原始音频信号对应的说话人身份信息作为条件;

[0046] 在步骤303中,将输入和条件输入多个残差网络,并将得到的所有输出相加,经过卷积层得到优化后的音频信号。

[0047] 在本实施例中,对于步骤301,声音转换优化装置首先将目标梅尔频谱特征经过声码器的上采样层与原始音频信号对齐。之后,对于步骤302,将原始音频信号作为输入,与该原始音频信号对齐后的目标梅尔频谱特征以及与原始音频信号对应的说话人身份信息作为条件,在步骤303中输入至声码器中的多个残差网络(Residual Block)中,然后将经由残差网络后得到的所有输出想加,再经过卷积层得到优化后的音频信号。

[0048] 本实施例的方法由于使用了声学特征中非常低水平的梅尔频谱图作为声码器的条件,没有了之前F0,BAP转换不够准确的缺陷,可以最终输出的声音更加真实。

[0049] 需要说明的是,对于前述的各方法实施例,为了简单描述,故将其都表述为一系列的动作合并,但是本领域技术人员应该知悉,本发明并不受所描述的动作顺序的限制,因为依据本发明,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本发明所必须的。在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述的部分,可以参见其他实施例的相关描述。

[0050] 接下来,以一个具体的示例以及实验和实验结果对比分析来论述本发明的实现过程和效果。

[0051] 本申请背景技术部分提到的基于Mcep的声音转换系统将在后续的实验中将用于与本申请的方案进行比较。不同系统中的Msp和Mcep使用相似的LSTM-RNN神经网络进行帧到帧特征映射的训练。使用STRAIGHT声码器和基于Mcep的WaveNet声码器将转换后的Mcep和F0编码为波形。针对不同系统详细分析了转换波形的F0轮廓,这是语音质量的一个重要因素。人类听众主观地评估自然性,相似性和可理解性。结果表明,使用梅尔频谱特征进行声音转换可以产生高质量的转换语音,尤其是在相似度方面。

[0052] 发明人发现,在现有技术中,使用了F0作为条件,但是F0的转换并不是特别精确,只是通过线性的拉伸或者通过全局均值方差变换,所以和目标说话人的F0会有较大的差别,同时使用GMM作为转换模型在特征变换上也没有充分利用历史信息,因此转换的特征不够准确。合成的声音和目标说话人的音色不太像,并且在音调上也有比较大的差别。

[0053] 以往,为了解决这些缺陷,通常都会使用比较大的数据集或者修改转换模型去提高特征的转换精度,但是F0的转换一直都没有较大的变化,而本申请的方案不需要用到F0,直接修改传统方法的特征,使得只需要一种特征就可以完成所有的操作。

[0054] 在实现本申请的过程中,发明人还曾经尝试过另外的一些方案。对比本申请的最终方案,之前的方案中存在通过修改生成语音的时长来使得合成的语音更加自然。之前的方案在整体设计上与现有的设计方案一致,但是需要首先通过对原始说话人和目标说话人

的语音数据做语音识别得到各自的phone(音素)的时长,通过建立每一个phone之间的对应关系,得到每一个phone的时长比例,因此最后我们合成的时候可以根据这个比例对声音的长度进行调整,从而使声音的语速更加接近目标说话人。

[0055] 在该方案中,我们发现合成的声音会出现中间个别音节没有发出来的问题,因此我们采用了一种对偶学习的方法,具体请参照图4a和图4b。其中,MSE表示最小均方误差。

[0056] 步骤一:从目标说话人和原始说话人的数据中分别提取特征:声学特征A,声学特征B;

[0057] 步骤二:设计一种对偶神经网络A→B'→A',这种网络要求输出的特征A能够得到预测的特征B',同时预测的特征B'能够通过网络得到预测的特征A',这样我们可以计算的损失就不仅仅是B和B',同时还要加上A和A'

[0058] 步骤三:我们将步骤二种的网络切开成B→A'→B',这样我们可以保持两个网络参数共享并且在这个网络中以预测的真实特征B作为输入,输出我们的原始真实特征A的预测值,最后还要预测B'。

[0059] 通过这种对偶网络能够增加训练数据的数量,同时加强网络对特征转换之间的学习能力使得转换的效果更为准确。该方案确实能提高声音的自然度。

[0060] 但是通过我们的实验结果发现,由于合成的语音中本来就有个别phone发不出来,并且改变时长同样会对发音的准确性产生干扰,而这种方法并不能有效的减少未发出音的数量,虽然有一定的效果但是需要仔细的听,对听音测试的帮助不大,而且过程比较繁琐,因此在最终的版本中我们抛弃了这种方案。

[0061] 后来,发明人采用了Msp(梅尔频谱,Mel-spectrogram)作为转换特征,没有了之前F0,BAP转换不够准确的缺陷,并且使用了最新的WaveNet模型并采用Msp作为条件,将两者结合在一起形成一个完整的系统。以下,介绍一些与本申请最终方案相关的内容。

[0062] 首先,介绍并行数据声音转换框架。图1示出了基于Mcep的并行数据声音转换系统的体系结构。原始说话人的声学特征在不同的特征流中被转换为目标说话人。然后将转换后的功能编码为音频信号。这种架构也是一种通用的参数化声音转换框架,其中一些常规处理被特定方法(例如BLSTM NN,WaveNet Vocoder(声码器))所取代。

[0063] 对于具有相同文本的语音对,来自源扬声器的声学特征$x=x_1,\ldots,x_m$和来自目标扬声器的相应声学特征$y=y_1,\ldots,y_n$首先被对齐到相同长度T。对齐通常直接由动态时间包裹(Dynamic Time Wrapping,DTW)提供。另外,还有一些技术可以在自动语音识别技术的帮助下获得更准确的特征对齐。对齐的特征序列$x=x_1,\ldots,x_T$和$y=y_1,\ldots,y_T$然后在不同的方法(例如GMM,LSTM)中被逐帧地转换。如公式(1)所示,训练成本简单地通过均方误差来测量,其中M_{xy}是从原始音频信号到目标音频信号的Mcep转换模型。F0线性转换,非周期性不转换。

[0064] $$\mathcal{L}=\sum_{i=1}^{T}|M_{xy}(x_i)-y_i|_2 \quad (1)$$

[0065] 我们观察到,转换后语音的清晰度可能会因WaveNet Vocoder而降低。

[0066] 梅尔频谱转换

[0067] 梅尔频谱是语音波形的非常低水平的声学呈现。它还没有作为声音转换任务中的

声学特征导入,因为之前没有好的梅尔频谱声码器。

[0068] 如图5所示,我们提出了一个架构来将语音波形与梅尔谱图进行转换。语音波形仅被分析成梅尔谱图。然后按照图6中的体系结构逐帧转换梅尔频谱。与传统的基于Mcep的声音转换相比,F0不需要明确地作为单独的特征流转换。相关技术中已经对F0和持续时间模式进行了参数化,以适当地处理它们的超节段特性,在帧逐步转换过程中没有很好地转换。然而,在所提出的系统中,在转换梅尔频谱的同时,F0被非相干地转换。F0转换的性能将在后续实验中详细分析。

[0069] WaveNet声码器

[0070] 传统的声音转换声码器做出了各种各样的假设,这些假设通常会导致转换语音的声音质量下降。因此,相关技术中提出了一种主要基于Mel倒谱和F0的Wavenet Vocoder来克服这个问题。现有的实验结果也表明,该方案能产生更好的波形。

[0071] 本申请的基于Msp的WaveNet可以在端到端的文本到语音任务中产生高质量的语音波形。条件WaveNet的体系结构如图7所示。它由一堆扩张的因果卷积层组成,每个层可以并行处理输入向量。两个转置的卷积层被添加用于上采样。此外,在WaveNet中使用门控激活函数的机制可调节声音或语言特征等额外信息:

[0072] $z=\tanh(W_f*i+V_f*c)\odot\sigma(W_g*i+V_g*c)$ (2)

[0073] 其中*表示卷积运算符,而⊙表示单元乘法运算符。σ()表示一个S形函数。i是输入向量,c是像梅尔谱图那样的额外条件特征,并且是说话者身份的一个热点。f和g分别代表滤波器和门。W和V是可学习的权重,,z代表一种非线性函数,能够调节模型中的变量。

[0074] 实验设置

[0075] 这些实验使用PyTorch在CMU ARCTIC数据集上进行。数据集中的句子随机分为训练集,开发集和测试集,每个句子有957,107,55个句子。波形以16kHz采样率采样。通过短时傅里叶变换(STFT),使用50ms帧大小,12.5ms帧跳跃和Hann窗口函数提取梅尔频谱。基线系统使用LSTM-RNN声音转换系统。使用基于MLSA(Mel Log Spectrum Approximation,近似对数梅尔频谱)和Mcep的WaveNet声码器将转换后的声学特征编码为语音波形。Mceps在5ms帧移位时被提取。我们使用8位μ律训练了一个与扬声器相关的WaveNet声码器。

[0076] 对于本申请提出的系统,除了测试集中的话语外,我们在CMU ARCTIC数据集中的所有波形上初步训练了一个与说话人无关的WaveNet声码器。WaveNet网络训练了1000k步,Adam optitimizer(adaptive moment estimation optimizer,自适应矩优化优化器)在4个GTX1080TI上配备了16个小批量(minibatch),它有24层,分为4组。剩余连接和门控层的隐藏单元是512,输出层的跳转连接是256。我们还使用10个混合组件作为物流输出分配的混合物。然后,我们训练了一个基于LSTM网络的转换模型,该网络有两层,隐藏单元为256。在LSTM层之前,我们使用了两个PReLU(Rectified Linear Units,整流线性单位)激活的密集层。我们对源语言和焦点语者应用全局均值-方差变换。确保两个WaveNet声码器都训练有素。在WaveNet声码器可以在训练集上产生令人信服的语音之后,训练过程停止。

[0077] F0是影响语音质量的重要声学特征。在基于梅尔谱图的声音转换中,所有声学信息都保持在低级谱图表示中。因此,在梅尔谱图转换过程中,固有地转换F0。我们首先给出转换语音的F0轮廓的评估。

[0078] F0轮廓使用WORLD从自然和转换的语音中提取。图9显示了F0轮廓的一个例子,音

频从bdl转换为slt。由于bdl和slt具有相似的语速，因此我们可以直接查看它们的F0轮廓。我们可以看到，从Msp转换后的语音的F0轮廓更接近目标语音，即使F0没有明确转换。我们在图8中绘制了F0的分布，我们提出的系统和基于Msp的系统都与目标语音具有接近的均值和标准差。确切地说，基于Msp的系统中的F0是通过源语言和目标语句之间的全局均值-方差变换来传递的。所以可以肯定的是，本申请提出的系统在没有任何先验条件下可以获得更好的F0。

[0079] 其中，以下两个表格中bdl、rms、clb、slt分别代表四个不同说话人的数据，bdl和rms是男性数据集，clb和slt是女性数据集。

[0080]

系统	bdl-rms	clb-rms	bdl-slt	clb-slt
MSP-WaveNet	10.18	10.28	9.15	9.1
Mcep-WaveNet	11.22	10.85	11.76	11.06

[0081] 表1:基频(RMSE)

[0082]

系统	bdl-rms	clb-rms	bdl-slt	clb-slt
MSP-WaveNet	3.38	3.1	2.63	4.01
Mcep-WaveNet	3.46	3.21	2.71	3.63

[0083] 表2:有声/无声判定误差比较(%)

[0084] 表1显示了F0误差的客观测量。在我们评估之前，DTW被用来对准自然目标话语和转换后的话语。我们提出的系统比基于Mcep的系统具有更高的精度。表2列出了清音/浊音(U/V)判决误差。相信所提出的系统能够和基于Mcep的系统以相当精确的速度捕捉U/V信息。

[0085] 主观测试

[0086] 所有的主观测试都是在性别上和性别上进行的。在听力测试中，我们使用(clb→slt)作为内部性别对，(bdl→slt)作为跨性别对。测试集中的所有55个句子都用于听力测试。在每个测试中，每个句子都会呈现给至少6个听众。听众都是非母语的人。

[0087] 自然:我们对语音自然度进行平均评价评分(MOS)评估。经过评估的实验设置如下：

[0088] -自然语言(N)

[0089] -在自然Msp上的WaveNet-vocoded语音(WNS)

[0090] -在自然Mcep上的WaveNet-vocoded语音(WNC)

[0091] -转换后的Msp上的WaveNet语音编码语音(WCS)

[0092] -WaveNet-vocoded声音转换Mcep(WCC)

[0093] -转换后的Mcep上的MLSA语音编码语音(MCC)

[0094] 其中，以上缩写中第一个字符是指声码器类型(WaveNet/MLSA)；第二个字符是指声学特征(Natural/Converted)；第三个字符是指声学特征类型(Mel-Spectrogram/Mel-Cepstrum)

[0095] 清晰度:我们观察到，使用WaveNet声码器(Msp和Mcep)可能会扭曲上下文信息。因此我们还对转换后的语音的可理解性进行了MOS评估。

[0096] 相似:我们运行偏好测试来评估相似性。将来自两个系统的转换后的语音以随机顺序与来自目标讲话者的相同句子的自然语音一起提供给收听者。听众被要求选择哪个句

子听起来更像目标说话人。

[0097] 实验结果

[0098] 图11显示了转换语音的自然结果。我们可以看到,WNS比WNC表现更好,这意味着Mel谱图转换在语音自然度上有更高的上限,这可以进一步研究。除此之外,与WCC和MCC相比,WCS实现了更好的性能,这表明基于梅尔谱图的声音转换可以实现良好的自然度。

[0099] 图10显示了转换语音的可理解性结果。MCC实现比WCS和WCC更好的性能。其中一个原因是MCC可以在所有帧中生成质量稳定的转换语音,另一个原因是WaveNet Vocoder有时会产生嗡嗡声,这可以被视为缺乏WaveNet Vocoder的训练数据。这也可以说明为什么基于Mcep的WaveNet声码器具有与MLSA相似的语音质量MOS的原因,即使具有更高的自然度。

[0100] 除此之外,我们还可以看到WNS的表现比WNC好得多,这意味着Msp比Mcep包含更多的信息。

[0101] 图12a和图12b显示了不同系统与目标说话人相比较的结果。其中,图12a示出了bdl到slt的实验结果,图12b示出了clb到slt的实验结果。它显示MspWavenet在性别和跨性别案例上的表现明显优于McepWaveNet和Mcep STAIGHT。

[0102] 结论和未来的工作

[0103] 本申请提出了一种声音转换技术,利用LSTM神经网络和基于Mel-spectrogram的WaveNet Vocoder从原始音频信号到目标目标音频信号生成高质量语音。我们不使用STAIGHT的传统特征,而是在所提出的系统的所有流水线中应用梅尔频谱。实验表明,基于Mel-spectrogram的WaveNet Vocoder在自然度,相似度和清晰度方面的性能优于基于Mcep的WaveNet Vocoder在声音转换任务中的性能。

[0104] 本申请的方案由于直接修改了传统方法的特征,减少了特征的数量,在特征提取的层次上更为方便,同时更加有利于对特征的建模,使得从模型中生成的特征更加准确。另外,由于简化了流程,使的整个建模过程中的损失更少,模型的准确度大大提高,生成的声音在自然度和相似度上都较传统的方案有很大的提升。

[0105] 请参考图13,其示出了本发明一实施例提供的声音转换优化系统的框图。如图13所示,本发明的声音转换优化系统1300,包括提取单元1310、映射单元1320和优化单元1330。

[0106] 其中,提取单元1310,配置为从原始音频信号中提取原始梅尔频谱特征;映射单元1320,配置为将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征;以及优化单元1330,配置为将所述原始音频信号作为输入,所述目标梅尔频谱特征作为条件,输入声音转换声码器以得到优化后的音频信号。

[0107] 在一些可选的实施例中,优化单元1330配置为:将所述目标梅尔频谱特征经过上采样层与所述原始音频信号对齐;使用所述原始音频信号作为输入,对齐后的所述目标梅尔频谱特征及与所述原始音频信号对应的说话人身份信息作为条件;将所述输入和所述条件输入多个残差网络,并将得到的所有输出相加,经过卷积层得到优化后的音频信号。

[0108] 应当理解,图13中记载的诸模块与参考图2和图3中描述的方法中的各个步骤相对应。由此,上文针对方法描述的操作和特征以及相应的技术效果同样适用于图13中的诸模块,在此不再赘述。

[0109] 值得注意的是,本公开的实施例中的模块并不用于限制本公开的方案,例如提取

单元可以描述为从原始音频信号中提取原始梅尔频谱特征的单元。另外，还可以通过硬件处理器来实现相关功能模块，例如提取单元也可以用处理器实现，在此不再赘述。

[0110] 在另一些实施例中，本发明实施例还提供了一种非易失性计算机存储介质，计算机存储介质存储有计算机可执行指令，该计算机可执行指令可执行上述任意方法实施例中的声音转换优化方法；

[0111] 作为一种实施方式，本发明的非易失性计算机存储介质存储有计算机可执行指令，计算机可执行指令设置为：

[0112] 从原始音频信号中提取原始梅尔频谱特征；

[0113] 将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征；

[0114] 将所述原始音频信号作为输入，所述目标梅尔频谱特征作为条件，输入声音转换声码器以得到优化后的音频信号。

[0115] 作为一种非易失性计算机可读存储介质，可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块，如本发明实施例中的声音转换优化方法对应的程序指令/模块。一个或者多个程序指令存储在非易失性计算机可读存储介质中，当被处理器执行时，执行上述任意方法实施例中的声音转换优化方法。

[0116] 非易失性计算机可读存储介质可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据声音转换优化装置的使用所创建的数据等。此外，非易失性计算机可读存储介质可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实施例中，非易失性计算机可读存储介质可选包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至声音转换优化装置。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0117] 本发明实施例还提供一种计算机程序产品，计算机程序产品包括存储在非易失性计算机可读存储介质上的计算机程序，计算机程序包括程序指令，当程序指令被计算机执行时，使计算机执行上述任一项声音转换优化方法。

[0118] 图14是本发明实施例提供的电子设备的结构示意图，如图14所示，该设备包括：一个或多个处理器1410以及存储器1420，图14中以一个处理器1410为例。声音转换优化方法的设备还可以包括：输入装置1430和输出装置1440。处理器1410、存储器1420、输入装置1430和输出装置1440可以通过总线或者其他方式连接，图14中以通过总线连接为例。存储器1420为上述的非易失性计算机可读存储介质。处理器1410通过运行存储在存储器1420中的非易失性软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例声音转换优化方法。输入装置1430可接收输入的数字或字符信息，以及产生与信息投放装置的用户设置以及功能控制有关的键信号输入。输出装置1440可包括显示屏等显示设备。

[0119] 上述产品可执行本发明实施例所提供的方法，具备执行方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节，可参见本发明实施例所提供的方法。

[0120] 作为一种实施方式，上述电子设备应用于神经网络语言模型中，包括：至少一个处理器；以及，与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够：

[0121] 从原始音频信号中提取原始梅尔频谱特征；

[0122] 将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征；

[0123] 将所述原始音频信号作为输入，所述目标梅尔频谱特征作为条件，输入声音转换声码器以得到优化后的音频信号。

[0124] 本申请实施例的电子设备以多种形式存在，包括但不限于：

[0125] (1)移动通信设备：这类设备的特点是具备移动通信功能，并且以提供话音、数据通信为主要目标。这类终端包括：智能手机(例如iPhone)、多媒体手机、功能性手机，以及低端手机等。

[0126] (2)超移动个人计算机设备：这类设备属于个人计算机的范畴，有计算和处理功能，一般也具备移动上网特性。这类终端包括：PDA、MID和UMPC设备等，例如iPad。

[0127] (3)服务器：提供计算服务的设备，服务器的构成包括处理器、硬盘、内存、系统总线等，服务器和通用的计算机架构类似，但是由于需要提供高可靠的服务，因此在处理能力、稳定性、可靠性、安全性、可扩展性、可管理性等方面要求较高。

[0128] (4)其他具有数据交互功能的电子装置。

[0129] 以上所描述的装置实施例仅仅是示意性的，其中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的劳动的情况下，即可以理解并实施。

[0130] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分的方法。

[0131] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

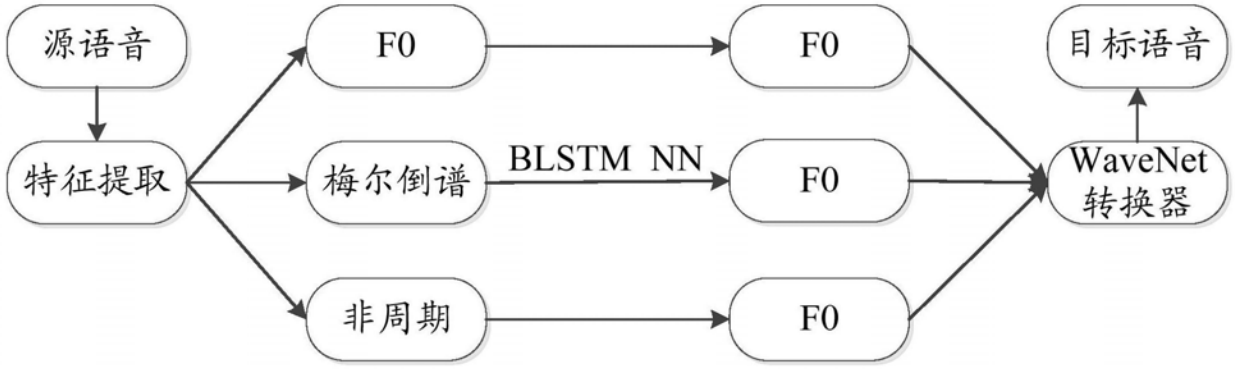

源语音
特征提取
F0
梅尔倒谱
非周期
BLSTM NN
F0
F0
F0
WaveNet
转换器
目标语音

图1

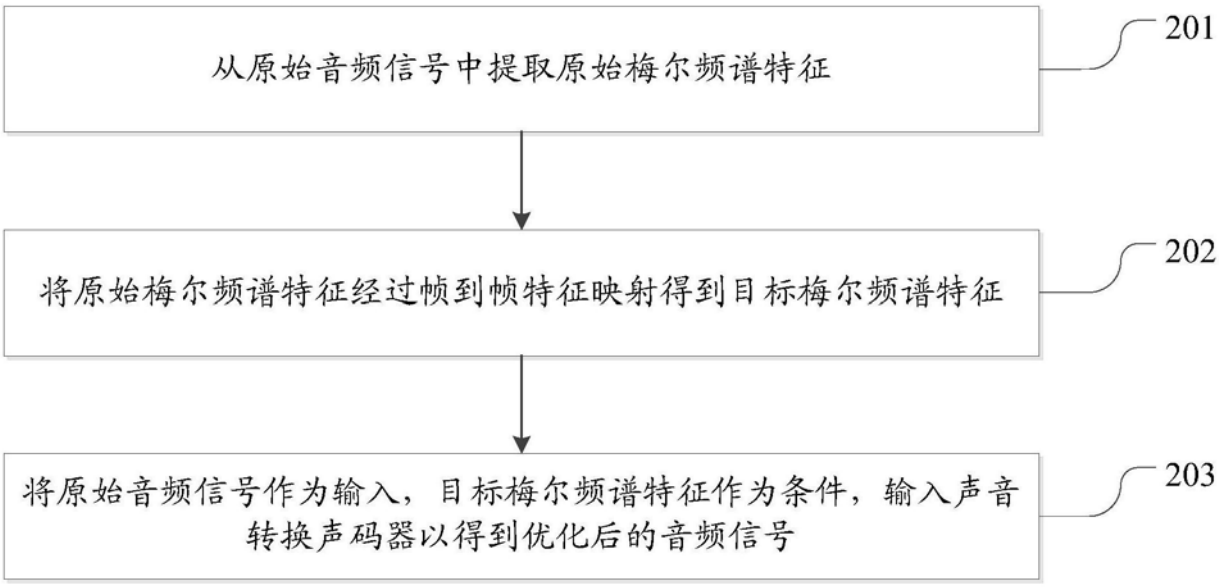

从原始音频信号中提取原始梅尔频谱特征
201
将原始梅尔频谱特征经过帧到帧特征映射得到目标梅尔频谱特征
202
将原始音频信号作为输入，目标梅尔频谱特征作为条件，输入声音转换声码器以得到优化后的音频信号
203

图2

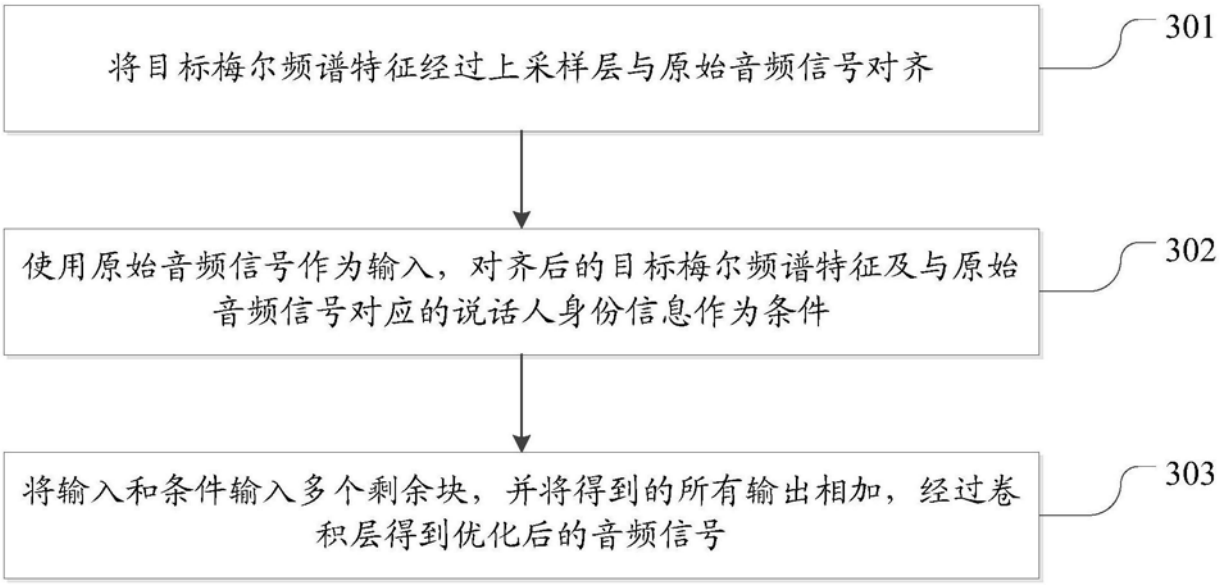

将目标梅尔频谱特征经过上采样层与原始音频信号对齐
301
使用原始音频信号作为输入，对齐后的目标梅尔频谱特征及与原始音频信号对应的说话人身份信息作为条件
302
将输入和条件输入多个剩余块，并将得到的所有输出相加，经过卷积层得到优化后的音频信号
303

图3

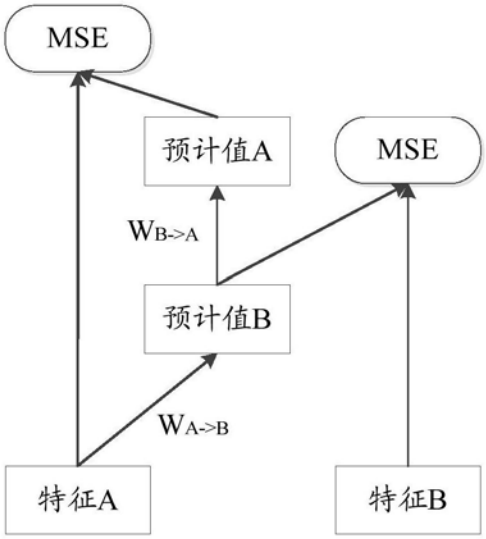

MSE
预计值A
MSE
WB->A
预计值B
WA->B
特征A
特征B

图4a

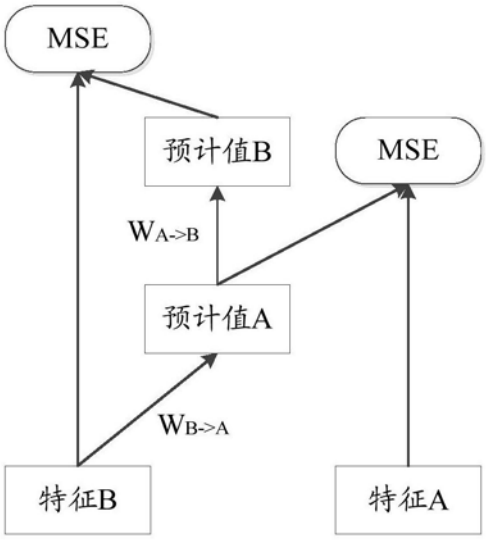

MSE
预计值B
MSE
WA->B
预计值A
WB->A
特征B
特征A

图4b

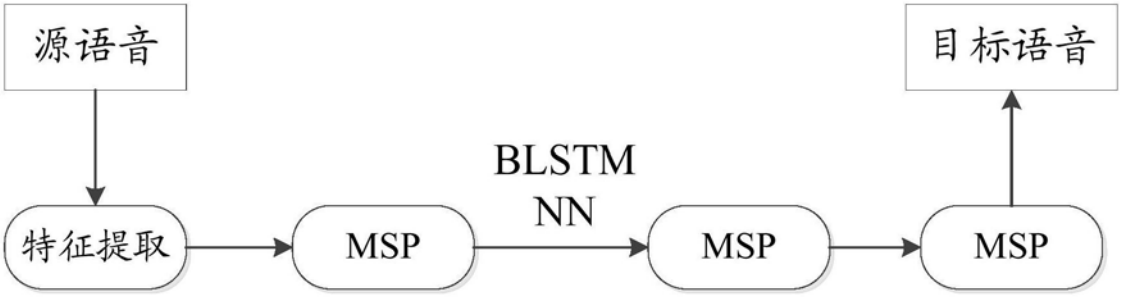

源语音
目标语音
BLSTM
NN
特征提取
MSP
MSP
MSP

图5

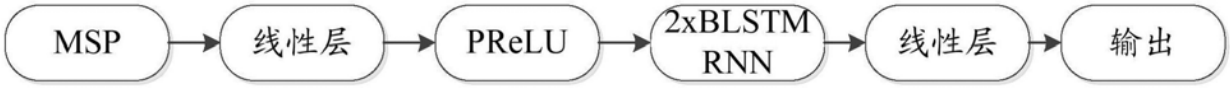

MSP
线性层
PReLU
2xBLSTM
RNN
线性层
输出

图6

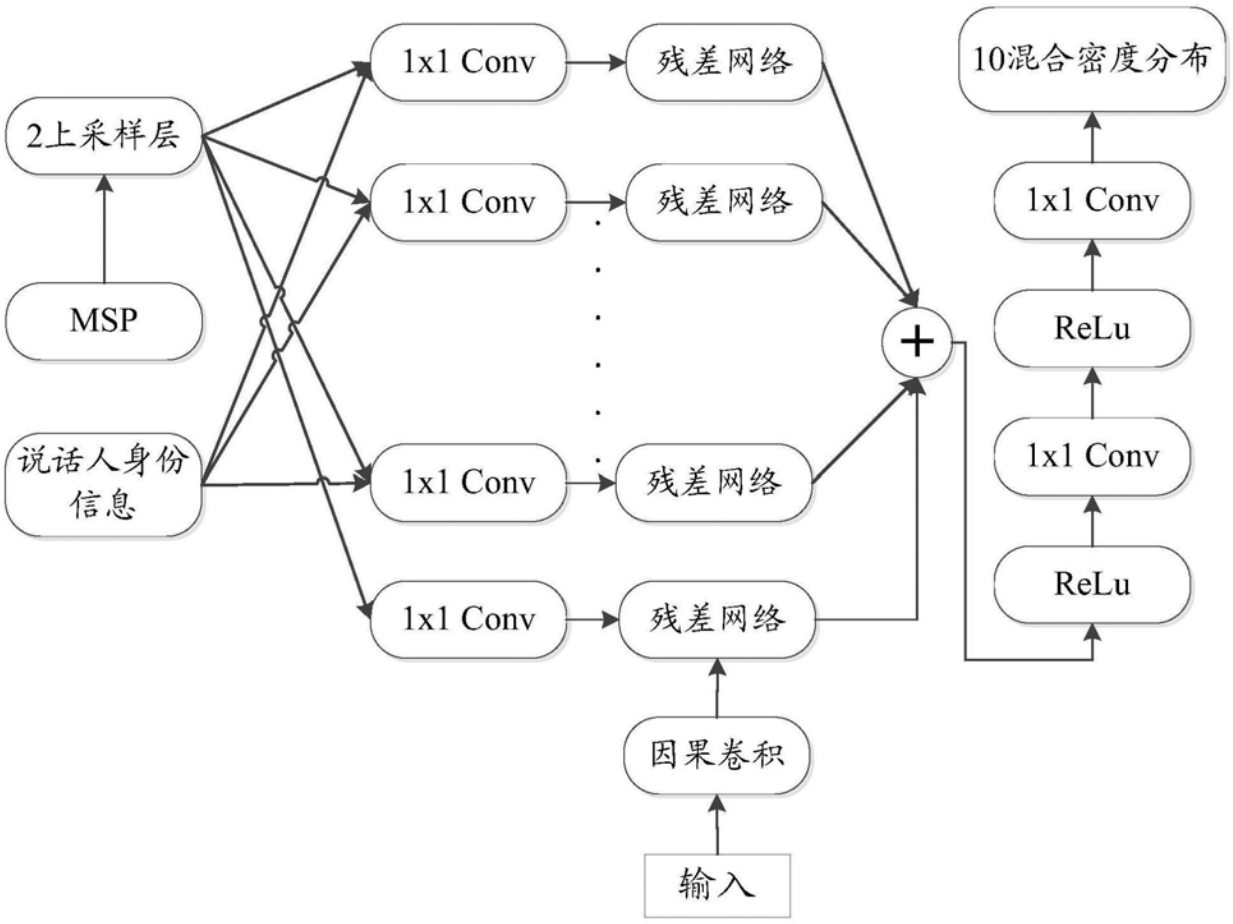

2上采样层
MSP
说话人身份信息
1x1 Conv
1x1 Conv
1x1 Conv
1x1 Conv
残差网络
残差网络
残差网络
残差网络
因果卷积
输入
+
10混合密度分布
1x1 Conv
ReLu
1x1 Conv
ReLu

图7

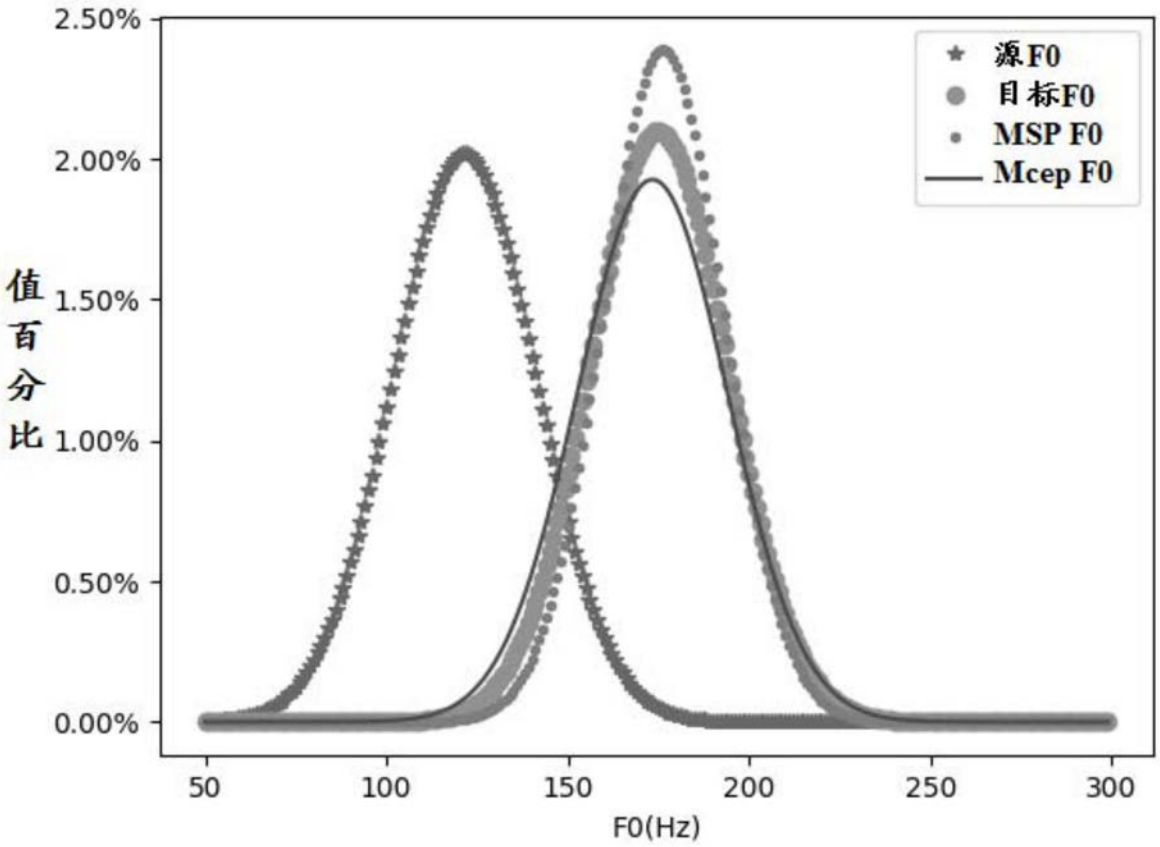

2.50%
2.00%
1.50%
1.00%
0.50%
0.00%
值百分比
50
100
150
200
250
300
F0(Hz)
源F0
目标F0
MSP F0
Mcep F0

图8

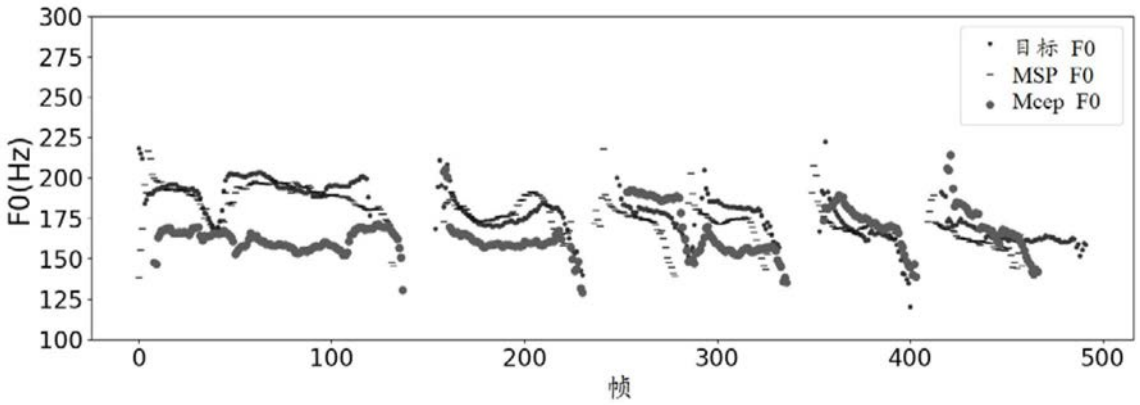

300
275
250
225
200
175
150
125
100
F0(Hz)
0
100
200
300
400
500
帧
目标 F0
MSP F0
Mcep F0

图9

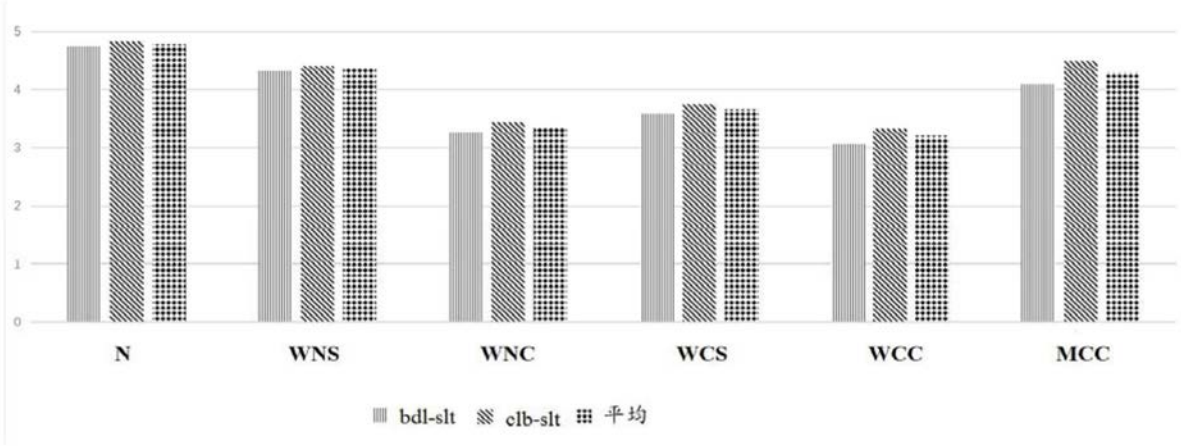

5
4
3
2
1
0
N
WNS
WNC
WCS
WCC
MCC
bdl-slt
clb-slt
平均

图10

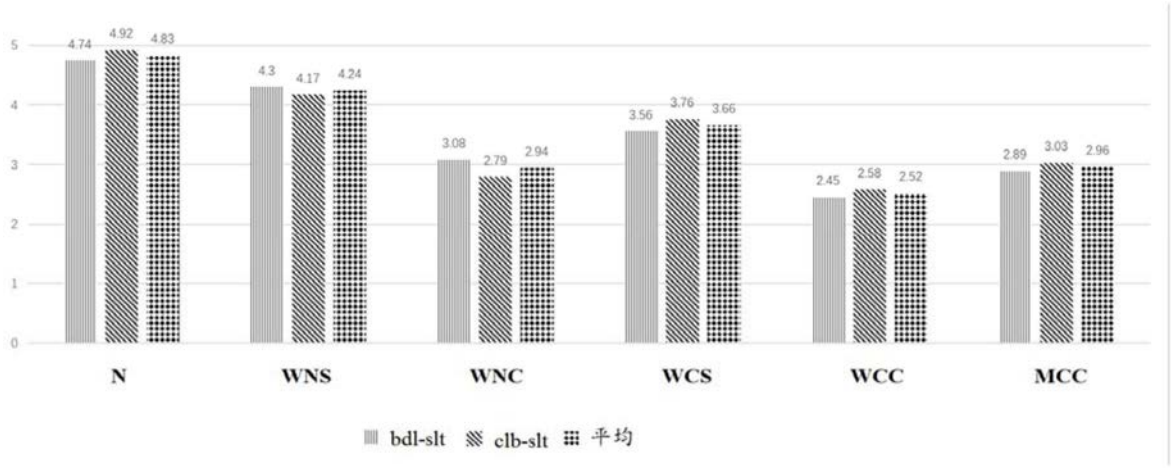

4.74
4.92
4.83
4.3
4.17
4.24
3.08
2.79
2.94
3.56
3.76
3.66
2.45
2.58
2.52
2.89
3.03
2.96
N
WNS
WNC
WCS
WCC
MCC
bdl-slt
clb-slt
平均

图11

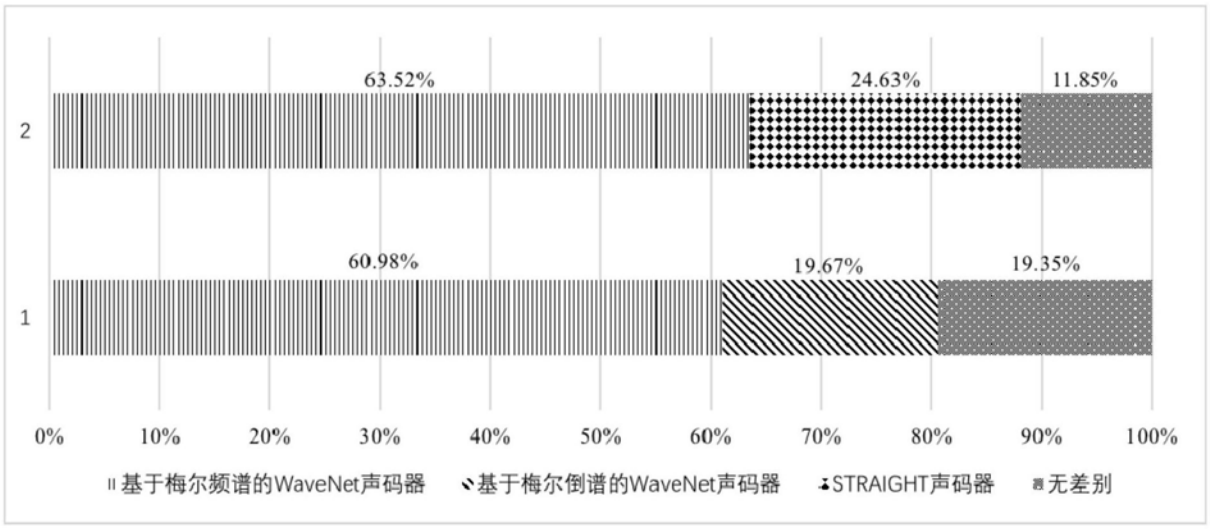

63.52%
24.63%
11.85%
60.98%
19.67%
19.35%
2
1
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
基于梅尔频谱的WaveNet声码器
基于梅尔倒谱的WaveNet声码器
STRAIGHT声码器
无差别

图12a

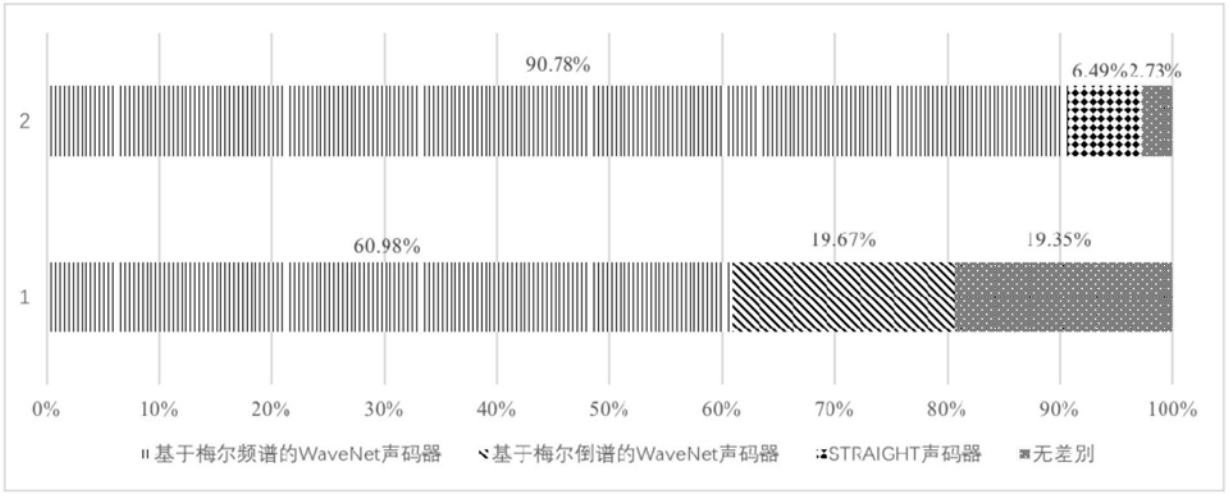

90.78%
6.49%2.73%
60.98%
19.67%
19.35%
2
1
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
基于梅尔频谱的WaveNet声码器
基于梅尔倒谱的WaveNet声码器
STRAIGHT声码器
无差别

图12b

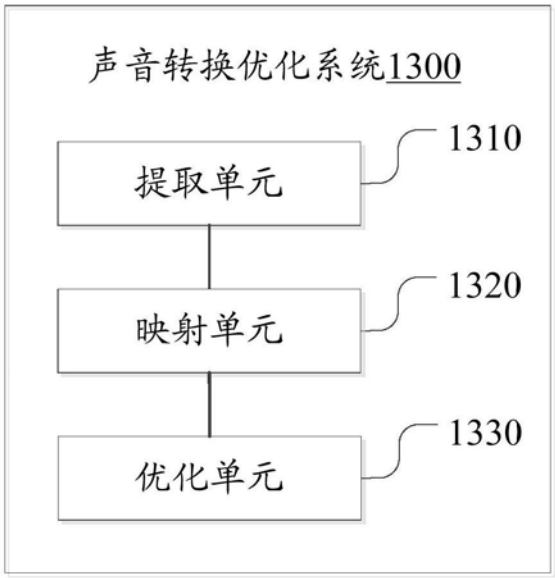

声音转换优化系统1300
提取单元
1310
映射单元
1320
优化单元
1330

图13

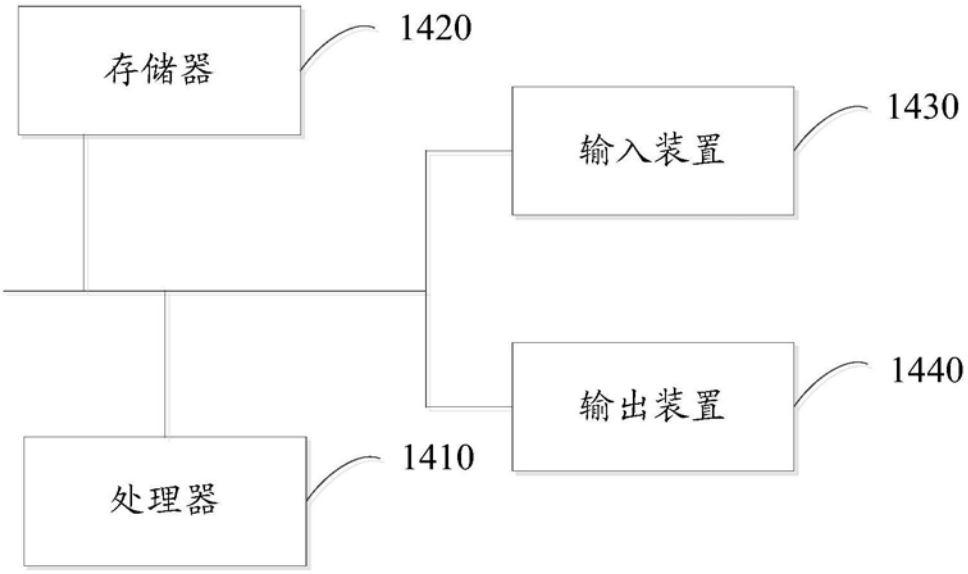

存储器
1420
输入装置
1430
输出装置
1440
处理器
1410

图14