(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 107845389 B
(45)授权公告日 2020.07.17

(21)申请号 201711397819.6

(22)申请日 2017.12.21

(65)同一申请的已公布的文献号
申请公布号 CN 107845389 A

(43)申请公布日 2018.03.27

(73)专利权人 北京工业大学
地址 100124 北京市朝阳区平乐园100号

(72)发明人 李如玮 刘亚楠 李涛 孙晓月

(74)专利代理机构 北京思海天达知识产权代理有限公司 11203
代理人 张慧

(51)Int.Cl.
G10L 21/0216(2013.01)
G10L 15/16(2006.01)
G10L 25/24(2013.01)

(56)对比文件
CN 102982801 A,2013.03.20
CN 103778920 A,2014.05.07
CN 107077860 A,2017.08.18
Arun Narayanan, DeLiang Wang.IDEAL RATIO MASK ESTIMATION USING DEEP NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION.<ICASSP>.2013,7092-7096.
CHEN J,WANG Y,WANG D L.A feature study for classification-based speech separation at low signal-to noise.<ACM Transactions on Audio>.2014,第22卷(第12期),1993-2002.
曹龙涛,李如玮,鲍长春,吴水才.基于噪声估计的二值掩蔽语音增强算法.《计算机工程与应用》.2015,第51卷(第17期),222-227.
李如玮,鲍长春,窦慧晶.基于双正交小波包分解的自适应阈值语音增强.《仪器仪表学报》.2008,第29卷(第10期),2135-2140.

审查员 谭雪艳

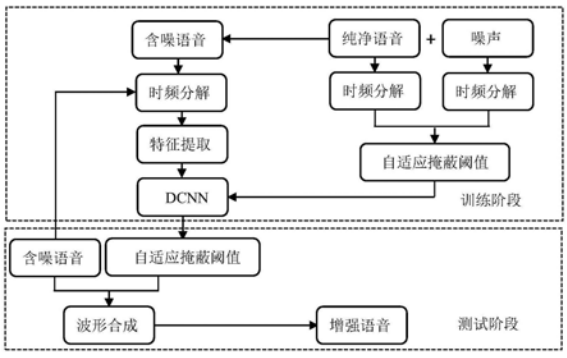
权利要求书2页 说明书7页 附图3页

(54)发明名称

一种基于多分辨率听觉倒谱系数和深度卷积神经网络的语音增强方法

(57)摘要

本发明提出了一种基于多分辨率倒谱系数和卷积神经网络的语音增强方法,首先构建了新的能够区分语音和噪声的特征参数—多分辨率听觉倒谱系数(MR-GFCC);其次,跟踪噪声变化构建了基于理想软掩蔽(IRM)和理想二值掩蔽(IBM)的自适应掩蔽阈值;然后将提取的新特征参数及其一二阶导数和自适应掩蔽阈值作为深度卷积神经网络(DCNN)的输入和输出,对构建的7层神经网络进行训练;最后利用DCNN估计的自适应掩蔽阈值对含噪语音进行增强。本发明充分利用了人耳的工作机理,提出了模拟人耳听觉生理模型的语音特征参数,不仅可以保留更多的语音信息,而且提取过程简单可行。



CN 107845389 B

1.一种基于多分辨率和深度卷积神经网络的语音增强方法,其特征在于,包括以下步骤:

步骤一,将含噪语音通过64通道的gammatone滤波器进行滤波,对每一个频率通道的输出做加窗分帧处理,得到其时频域表示形式;

步骤二,提取每个时频单元的特征参数

(1)帧长为20ms,帧移为10ms,求每个时频单元的能量,记作CG1;然后对每个时频单元的能量进行立方根非线性压缩变化来模拟人耳对语音的强度-响度感知特性;最后经过DCT到倒谱域,同时取前36维系数,得到CG1的倒谱系数,记作CG1-GFCC;

(2)将帧长改为200ms,重复步(1)特征提取过程,得到CG2-GFCC;

(3)使用一个长为11帧,宽为11子带的方形窗对CG1进行平滑,记作CG3,重复(1)中特征提取过程,得到CG3-GFCC;

(4)使用一个长为23帧,宽为23子带的方形窗对CG1进行平滑,记作CG4,重复(1)中特征提取过程,得到CG4-GFCC;

(5)将CG1-GFCC、CG2-GFCC、CG3-GFCC和CG4-GFCC合并得到36*4维的特征向量,得到多分辨率倒谱系数(MR-GFCC);

(6)MR-GFCC参数做一阶和二阶差分参数得到△MR-GFCC和△△MR-GFCC,MR-GFCC、△MR-GFCC和△△MR-GFCC相结合作为最后的语音特征参数;

步骤三,计算DCNN训练的目标

首先分别计算出IBM和IRM,然后通过跟踪噪声变化自适应的估计出IBM和IRM前面的系数,最后将二者结合起来计算出自适应的掩蔽阈值;具体为:

步骤三中自适应掩蔽阈值作为DCNN的训练目标,其公式为:

$$z(i,f_c)=\alpha*IBM(i,f_c)+(1-\alpha)*IRM(i,f_c) \quad (15)$$

(1)首先计算理想二值掩蔽(IBM),公式如下:

$$IBM(i,f_c)=\begin{cases}1 & E_s(i,f_c)\geq E_n(i,f_c)\cdot 10^{\frac{lc}{10}}\\ 0 & else\end{cases} \quad (16)$$

其中$E_s(i,f_c)$和$E_n(i,f_c)$分别代表第i帧中心频率为f_c的纯净语音和噪声的能量,lc是阈值,t表示帧数,f_c表示中心频率,$IBM(i,f_c)$表示第i帧中心频率为f_c的纯净语音和噪声的理想二值掩蔽值;

(2)理想软掩蔽值(IRM)是一个比值的时频掩蔽矩阵,通过纯净语音和噪声计算得到,其定义为:

$$IRM_{gamm}(i,f_c)=\left(\frac{E_s(i,f_c)}{E_s(i,f_c)+E_n(i,f_c)}\right)^{\beta} \quad (19)$$

其中,β是一个可调的尺度因子,

(3)α—自适应系数的估计

假设前6帧是噪声帧,由式计算出除去第1帧后5帧的噪声能量$n^2(i,f_c)$,由这5帧按式(20)(21)求出初始噪声能量$\overline{n}^2(i,f_c)$,作为第6帧的噪声能量,

$$n^2(i,f_c)=\frac{1}{L}\sum_{t=0}^{L-1}(n_t(i,f_c))^2 \quad (20)$$

$$\overline{n}^2(t,f_c)=\frac{1}{5}\sum_{m=0}^{4}n^2(t-m,f_c) \quad (21)$$

其中，m表示噪声前5帧的索引，之后各帧的带噪语音信号的能量按式(20)求出，而该帧的噪声能量按式(22)自适应估计：

$n_w^2(i,f_c)=\alpha(i,f_c)\times n^2(i-1,f_c)+(1-\alpha(i,f))\times n^2(i,f_c)$ (22)

每一帧每个通道的信噪比SNR(i,f_c)由式(23)计算：

$$SNR(i,f_c)=\frac{n^2(i,f_c)}{n_w^2(i-1,f_c)} \quad (23)$$

α(t,f_c)由s形函数产生，由式(24)定义：

$$\alpha(i,f_c)=\frac{1}{1+\exp(-SNR(i,f_c))} \quad (24)$$

(4)根据公式(15)计算出自适应的掩蔽阈值z(i,f_c)；

步骤四，构建深度卷积神经网络模型(DCNN)，构建了一个7层的深度卷积神经网络学习输入和输出之间的非线性关系；

步骤五，将提取的特征参数和计算的自适应混合掩蔽阈值分别作为DCNN的输入和输出，对构建的7层深度卷积神经网络进行训练，得到网络的权值和偏置；

步骤六，按步骤二的方法提取测试的含噪语音的特征参数，输入到步骤五已经训练好的深度卷积神经网络中，输出一个自适应的掩蔽阈值；

步骤七，利用含噪语音和估计出的自适应的掩蔽阈值去合成增强后的语音。

2.如权利要求1所述的基于多分辨率和深度卷积神经网络的语音增强方法，其特征在于，步骤四深度卷积神经网络模型(DCNN)由1个输入层，5个隐含层和1个输出层构成；输入层用来输入含噪语音的特征参数，隐含层由卷积层、池化层层和全连接层组成，输出层用来输出估计的目标；它们之间通过传递函数来传递数据。

3.如权利要求1所述的基于多分辨率和深度卷积神经网络的语音增强方法，其特征在于，步骤五，将步骤四中提取的特征参数v_a(i,f_c)和步骤三计算的自适应掩蔽阈值z(i,f_c)分别作为DCNN的输入和输出，采用自适应学习率的随机梯度下降算法对网络进行训练，训练完成以后，保存网络的权值和偏置，其训练采用的是线下训练。

4.如权利要求1所述的基于多分辨率和深度卷积神经网络的语音增强方法，其特征在于，步骤六，按步骤二的方法提取测试的含噪语音的特征参数，输入到步骤五已经训练好的深度卷积神经网络中，输出一个自适应的掩蔽阈值Y(i,f_c)，如公式(25)所示；

$Y(i,f_c)=\theta(X(i,f_c)$ (25)

其中，θ()表示训练好的DCNN网络模型参数，X(i,f_c)表示测试的含噪语音的特征参数，Y(i,f_c)表示DCNN估计出的自适应掩蔽阈值。

一种基于多分辨率听觉倒谱系数和深度卷积神经网络的语音增强方法

技术领域

[0001] 本发明属于语音信号处理技术领域,涉及到一种基于多分辨率听觉倒谱系数和深度卷积神经网络的语音增强方法。

背景技术

[0002] 语音增强技术是指当语音信号被各种各样的噪声(包括语音)干扰,甚至淹没后,从噪声背景中提取出尽可能纯净的语音信号,增强有用的语音信号,抑制、降低噪声干扰的技术。由于干扰的随机性,因而从带噪语音提取完全纯净语音信号几乎不可能。在这种情况下,语音增强的目的主要有两个:一是改进语音质量,消除背景噪声,使听者乐于接受,不感觉疲劳,这是一种主观度量;二是提高语音的可懂度,这是一种客观度量。这两个目的往往不能兼得。

[0003] 当前,语音增强已发展成为语音信号数字处理的一个重要分支。它在语音通信、语音编码、语音识别和数字助听器等诸多领域中得到了广泛应用。传统的语音增强方法有谱减法、维纳滤波法、最小均方误差法(MMSE)、基于统计模型和基于小波变换等方法,其在平稳噪声环境下有较好的性能,但对非平稳噪声处理效果不理想。随着计算听觉场景分析(CASA)出现,基于人耳听觉模型的方法被应用到语音增强当中。该方法根据估计的理想二值掩蔽值,利用人耳的听觉掩蔽效应实现语音增强。相对于其他语音增强算法,计算听觉场景分析对噪音没有任何假设,具有更好的泛化性能。但由于缺乏谐波结构很难处理语音中的清音成分。

[0004] 随着深度神经网络技术的发展,由于其具有良好的复杂特征提取表达能力,擅长对数据中的结构化信息进行建模,许多研究者把它引入到语音增强当中,该方法是利用深度神经网络学习一个从带噪特征到分离目标的特征函数。目前常用的基于深度学习的语音增强算法主要是基于目标语音的幅度谱和理想时频掩蔽这两方面展开的。

[0005] 基于深度神经网络的目标语音幅度谱的语音增强算法是直接估计目标语音的幅度谱,而幅度谱的变化范围较大,学习难度较大,对目标语音幅度谱的准确估计非常困难。

[0006] 基于深度神经网络的时频掩蔽的语音增强算法是估计目标语音的二值掩蔽或软掩蔽,二者的计算比较简单,但是前者对语音质量损害较大,后者残留的背景噪声较多。

[0007] 本发明提出了一种基于多分辨率听觉倒谱系数和卷积神经网络相结合的语音增强技术。该技术首先构建了新的能够区分语音和噪声的特征参数—多分辨率听觉倒谱系数(MR-GFCC);其次,跟踪噪声变化构建了基于理想软掩蔽(IRM)和理想二值掩蔽(IBM)的自适应掩蔽阈值;然后将提取的新特征参数及其一二阶导数和自适应掩蔽阈值作为深度卷积神经网络(DCNN)的输入和输出,对构建的7层神经网络进行训练;最后利用DCNN估计的自适应掩蔽阈值对含噪语音进行增强。

发明内容

[0008] 本发明的目的是针对目前的语音增强算法在非平稳噪声下算法性能不理想的问题以及语音特征参数提取过程中存在的问题，提出了一种基于多分辨率倒谱系数和深度卷积神经网络相结合的语音增强技术。首先，使用gammatone滤波器组和非线性压缩运算来更好地模拟人耳的听觉生理模型，得到一种新的语音特征参数。然后，跟踪噪声变化构建了基于理想软掩蔽(IRM)和理想二值掩蔽(IBM)的自适应掩蔽阈值；接着利用深度学习中的深度卷积神经网络(DCNN)模型具有提取复杂特征的能力，擅长对数据中的结构化信息进行建模对自适应的掩蔽阈值进行估计，可以解决传统的语音增强算法在非平稳噪声环境下性能不理想的问题。最后，利用DCNN估计的自适应掩蔽阈值对含噪语音进行增强。

[0009] 基于多分辨率和深度卷积神经网络的语音增强方法的实现步骤如下：

[0010] 步骤一，将含噪语音通过64通道的gammatone滤波器进行滤波，对每一个频率通道的输出做加窗分帧处理；得到其时频域表示形式(时频单元)；

[0011] 步骤二，提取每个时频单元的特征参数。

[0012] (1)帧长为20ms，帧移为10ms，求每个时频单元的能量，记作CG1；然后对每个时频单元的能量进行立方根非线性压缩变化来模拟人耳对语音的强度-响度感知特性，这不仅符合了人耳的听觉感知特性，而且计算过程简单；最后经过DCT(离散余弦变换)到倒谱域，同时取前36维系数，降低了算法复杂度，得到CG1的倒谱系数，记作CG1-GFCC；

[0013] (2)将帧长改为200ms，重复步(1)特征提取过程，得到CG2-GFCC；

[0014] (3)使用一个长为11帧，宽为11子带的方形窗对CG1进行平滑，记作CG3，重复(1)中特征提取过程，得到CG3-GFCC；

[0015] (4)使用一个长为23帧，宽为23子带的方形窗对CG1进行平滑，记作CG4，重复(1)中特征提取过程，得到CG4-GFCC；

[0016] (5)将CG1-GFCC、CG2-GFCC、CG3-GFCC和CG4-GFCC合并得到36*4维的特征向量，得到多分辨率倒谱系数(MR-GFCC)；

[0017] (6)MR-GFCC参数做一阶和二阶差分参数得到△MR-GFCC和△△MR-GFCC，MR-GFCC、△MR-GFCC和△△MR-GFCC相结合作为最后的语音特征参数；

[0018] 步骤三，计算DCNN训练的目标。首先分别计算出IBM和IRM，然后通过跟踪噪声变化自适应的估计出IBM和IRM前面的系数，最后将二者结合起来计算出自适应的掩蔽阈值；

[0019] 步骤四，构建深度卷积神经网络模型(DCNN)。构建了一个7层的深度卷积神经网络学习输入和输出之间的非线性关系；

[0020] 步骤五，将提取的特征参数和计算的自适应混合掩蔽阈值分别作为DCNN的输入和输出，对构建的7层深度卷积神经网络进行训练，得到网络的权值和偏置；

[0021] 步骤六，按步骤二的方法提取测试的含噪语音的特征参数，输入到步骤五已经训练好的深度卷积神经网络中，输出一个自适应的掩蔽阈值；

[0022] 步骤七，利用含噪语音和估计出的自适应的掩蔽阈值去合成增强后的语音。

[0023] 本发明提出了基于多分辨率倒谱系数和深度卷积神经络的语音增强技术。该技术首先提出了一个新的语音特征参数，它的提取过程中使用了可以模拟人耳听觉模型的gammatone滤波器组进行滤波处理，利用耳蜗对信号的感知机理，把信号分解成64个频带，得到信号的时频表示形式，然后求每个时频单元的能量。再用基于强度-响度感知变换的非

线性压缩-立方根压缩对每个时频单元的能量进行压缩，这样在特征参数提取的过程中能更好地符合人耳的听觉感知特性，而且计算过程简单，使得计算复杂度低且运行时间较短。最后经过DCT变换到倒谱域，取前36维系数以及其一、二阶导数作为最后提取的特征参数，进一步降低算法的复杂度。其次，利用IBM和IRM各自的优点构造出一个跟踪噪声变化的自适应的掩蔽阈值。接着，构建了一个7层的深度卷积神经网络，利用它强大的非线性映射能力估计出自适应的掩蔽阈值。最后，利用含噪语音和估计出的自适应掩蔽阈值合成增强后的语音。该技术充分利用了人耳的工作机理，提出了模拟人耳听觉生理模型的语音特征参数，提取过程简单可行，算法复杂度低，同时利用深度卷积神经网络训练得到自适应掩蔽阈值对含噪语音进行增强，使得该算法在非平稳噪声环境中也有较好的性能。

附图说明

[0024] 图1本发明的实现流程图

[0025] 图2语音特征参数的提取流程图

[0026] 图3gammatone滤波器组中每个滤波器的频率相应

[0027] 图4gammatone滤波器组中每个滤波器合成后的频率相应

[0028] 图5DCNN的网络结构图

[0029] 图6自适应掩蔽阈值计算的流程图

具体实施方式

[0030] 为了更好地理解本发明，下面将详细描述本发明的具体实施方式：

[0031] 如图1所示，本发明提供一种基于多分辨率听觉倒谱系数和深度卷积神经网络的语音增强方法，包括以下步骤：

[0032] 步骤一，对输入的信号进行时频分解，然后进行加窗分帧处理，得到输入信号的时频表示形式；

[0033] (1)首先对输入的信号进行时频分解；

[0034] 语音信号是典型的时变信号，而时频分解正是着眼于真实语音信号组成成分的这种时变谱特征，将一维的语音信号分解成时间-频率表示的二维信号，旨在揭示语音信号中包含多少频率分量级及每个分量随时间是如何变化的。Gammatone滤波器即是时频分解的一种良好的工具。它能模拟人耳基底膜的时频分解机制，为此本文采用Gammatone滤波器组对含噪语音进行时频分解。gammatone滤波器组中每个滤波器的频率相应，如图3所示，gammatone滤波器组中每个滤波器合成后的频率相应，如图4所示。Gammatone滤波器的冲击响应为：

[0035] $$g(t,f_c)=\begin{cases}t^{l-1}e^{-2\pi B(f_c)t}\cos(2\pi f_c+\phi) & if \quad t\geq 0\\ 0 & esle\end{cases} \qquad (1)$$

[0036] 其中t表示采样点，f_c表示第c个gammatone滤波器通道的中心频率，心理声学的研究表明，人耳对声音信号的听觉感知依赖于临界频带。因此将人耳的临界频带的中心频率作为Gammatone滤波器中心频率。本文所用的实验数据采样率为16KHz，所以中心频率的范围设置为[50Hz,8000Hz]，将其划分为64个通道的Gammatone滤波器组可以更好地反映此频

带内语音的基频和谐波特性。φ为滤波器的初始相位，为简化模型，将φ设置为0。l为滤波器的阶数，实验表明当l＝4时，Gammatone滤波器可以很好地模拟耳蜗的听觉滤波特性，因此本文中设置l＝4。$B(f_c)$为滤波器带宽，它被定义为：

[0037] $B(f_c)=b*ERB(f_c)$ (2)

[0038] 其中b表示衰减因子，由实验数据分析当b＝1.019时，可以得到最好的滤波效果，所以本文设置b＝1.019。$ERB(f_c)$代表等效矩形带宽(equivalent rectangle bandwidth, ERB)，与中心频率f_c关系可定义为：

[0039] $ERB(f_c)=24.7+0.108f_c$ (3)

[0040] 其中24.7和0.108为实验中得到的经验值。

[0041] 输入信号的表达式如公式(4)所示：

[0042] $x(t)=s(t)+n(t)$ (4)

[0043] 式中x(t)代表含噪语音信号，s(t)代表纯净语音信号，n(t)表示噪声信号，它们的采样率均设置为16kHz。

[0044] 将x(t)通过64通道的gammatone滤波器进行滤波，将x(t)分解成64个子带信号$G_c(t,f_c)$，如公式(5)所示：

[0045] $G_c(t,f_c)=g(t,f_c)\cdot U(t)\cdot x(t)$ (5)

[0046] 其中U(t)单位阶跃函数，c表示子带编号。

[0047] (2)对每一个子带信号用汉明窗进行分帧处理，得到其时频域表示形式$y_i(t,f_c)$(时频单元)，如公式(6)所示：

[0048] $y_i(t,f_c)=w(t)*G_c(((i-1)*inc+t),f_c)$ (6)

[0049] 式中，w(t)为汉明窗函数，汉明窗与矩形窗相比，其频率分辨率相对较低，但是它的低通特性更加平滑，能够更好地反映语音信号的频率特性，所以本文选用汉明窗。i表示帧数，inc为帧移，设置为10ms(160点)，t的范围为[1,L]，L表示帧长，设置为20ms(320点)。

[0050] 步骤二，对输入信号的时频单元进行特征参数的提取，如图2所示；

[0051] (1)计算输入信号的每个时频单元(帧长为20ms)的听觉滤波器输出能量(cochleagram)CG1(i,f_c)，如公式(7)表示：

[0052] $$CG1(i,f_c)=\sum_{t=0}^{L-1}y_i^2(t,f_c) \quad (7)$$

[0053] (2)然后对每个时频单元的能量进行立方根非线性压缩变化来模拟人耳对语音的强度-响度感知特性，立方根非线性压缩能量CG_1(i,f_c)的计算公式为：

[0054] $CG\_1(i,f_c)=[CG1(i,f_c)]^{1/3}$ (8)

[0055] (3)最后经过DCT(离散余弦变换)到倒谱域获得帧长为20ms的听觉倒谱系数F(i,f_c)，其数学表达式为：

[0056] $$F(i,f_c)=(\frac{2}{M})^{0.5}\sum_{c=1}^{M}CG\_1(i,f_c)\cos(\frac{\pi n(2c-1)}{2M}) \quad (9)$$

[0057] 式中M为总通道数，本发明取M＝64。当c>36时，F(i,f_c)的值较小，因此取前36维特征，记作CG1-GFCC(i,f_c)；

[0058] (4)只将上述帧长20ms改为200ms，和CG1-GFCC(i,f_c)的提取过程一样，得到的特征参数记为CG2-GFCC(i,f_c)；

[0059] (5)使用一个长为11帧，宽为11子带的方形窗对CG1(i,f_c)进行平滑，得到CG3(i,f_c)，如公式(10)所示：

[0060] $$CG3(i,f_c)=\sum_{j=c-5}^{c+5}\sum_{k=i-5}^{i+5}sum(sum(CG1(i,f_c))/(11*11) \quad (10)$$

[0061] 再对CG3(i,f_c)进行2中的(2)、(3)操作，得到CG3-GFCC(i,f_c)；

[0062] (6)使用一个长为23帧，宽为23子带的方形窗对CG1(i,f_c)进行平滑，得到CG4(i,f_c)，如公式(11)所示：

[0063] $$CG4(i,f_c)=\sum_{j=c-11}^{c+11}\sum_{k=i-11}^{i+11}sum(sum(CG1(i,f_c))/(23*23) \quad (11)$$

[0064] 再对CG4(i,f_c)进行2中的(2)、(3)操作，得到CG4-GFCC(i,f_c)；

[0065] (7)将CG1-GFCC(i,f_c)、CG2-GFCC(i,f_c)、CG3-GFCC(i,f_c)和CG4-GFCC(i,f_c)进行合并，得到多分辨率听觉倒谱系数-MR-GFCC(i,f_c)，如公式(12)所示：

[0066] MR-GFCC(i,f_c)＝[CG1-GFCC(i,f_c)；CG2-GFCC(i,f_c)；CG3-GFCC(i,f_c)；CG4-GFCC(i,f_c)] (12)

[0067] (8)动态特征的提取。动态特征可以保留语音时域信息，与原始MR-GFCC相互补充可以保留更多的语音信息，有利于提高DCNN对目标估计的准确性。动态特征可以通过对公式(12)中的MR-GFCC(i,f_c)参数做一阶和二阶差分参数得到△MR-GFCC(i,f_c)和△△MR-GFCC(i,f_c)。其定义分别由公式(13)和(14)所示：

[0068] $$\Delta MR-GFCC(i,f_c)=\frac{\sum_{k=1}^{K}k(MR-GFCC(i+k,f_c)-MR-GFCC(i-k,f_c))}{2\sum_{k=1}^{K}k^2} \quad (13)$$

[0069] $$\Delta\Delta MR-GFCC(i,f_c)=\frac{\sum_{k=1}^{K}k(\Delta MR-GFCC(i+k,f_c)-\Delta MR-GFCC(i-k,f_c))}{2\sum_{k=1}^{K}k^2} \quad (14)$$

[0070] 式中k表示帧数差，k通常取1。

[0071] 最后提取的特征参数为v_a(i,f_c)＝[MR-GFCC(i,f_c)；△MR-GFCC(i,f_c)；△△MR-GFCC(i,f_c]，a为特征维数，a＝432。

[0072] 步骤三，计算DCNN的目标，如图6所示；

[0073] 本发明提出的自适应掩蔽阈值作为DCNN的训练目标。其公式为：

[0074] z(i,f_c)＝α*IBM(i,f_c)+(1-α)*IRM(i,f_c) (15)

[0075] (1)首先计算理想二值掩蔽(IBM)，公式如下：

[0076] $$IBM(i,f_c)=\begin{cases}1 & E_s(i,f_c)\geq E_n(i,f_c)\cdot 10^{\frac{lc}{10}}\\ 0 & else\end{cases} \quad (16)$$

[0077] 其中E_s(i,f_c)和E_n(i,f_c)分别代表第i帧中心频率为f_c的纯净语音和噪声的能量，计算公式由公式(17)、(18)得到。lc是阈值，通常取比含噪语音信噪比低5dB。t表示帧数，f_c表示中心频率。IBM(i,f_c)表示第i帧中心频率为f_c的纯净语音和噪声的理想二值掩蔽值。

[0078] $$E_s(i,f_c)=\sum_{t=0}^{L-1}s_t^2(i,f_c) \quad (17)$$

[0079] $$E_n(i,f_c)=\sum_{t=0}^{L-1}n_t^{\ 2}(i,f_c) \quad (18)$$

[0080] (2)理想软掩蔽值(IRM)是一个比值的时频掩蔽矩阵,通过纯净语音和噪声计算得到,其定义为:

[0081] $$\mathrm{IRM}_{\mathrm{gamm}}(i,f_c)=\left(\frac{\mathrm{E}_s(i,f_c)}{\mathrm{E}_s(i,f_c)+\mathrm{E}_n(i,f_c)}\right)^{\beta} \quad (19)$$

[0082] 式中β是一个可调的尺度因子,大量的实验表明β=0.5是最好的选择。

[0083] (3)α—自适应系数的估计。

[0084] 假设前6帧是噪声帧,由式计算出除去第1帧后5帧的噪声能量$n^2(i,f_c)$,由这5帧按式(20)(21)求出初始噪声能量$\bar{n}^2(i,f_c)$,作为第6帧的噪声能量。

[0085] $$n^2(i,f_c)=\frac{1}{L}\sum_{t=0}^{L-1}(n_t(i,f_c))^2 \quad (20)$$

[0086] $$\bar{n}^2(t,f_c)=\frac{1}{5}\sum_{m=0}^{4}n^{\ 2}(t-m,f_c) \quad (21)$$

[0087] 式中m表示噪声前5帧的索引,之后各帧的带噪语音信号的能量按式(20)求出,而该帧的噪声能量按式(22)自适应估计:

[0088] $n_w^2(i,f_c)=\alpha(i,f_c)\times n^2(i-1,f_c)+(1-\alpha(i,f))\times n^2(i,f_c)$ (22)

[0089] 每一帧每个通道的信噪比SNR(i,f_c)由式(23)计算:

[0090] $$SNR(i,f_c)=\frac{n^2(i,f_c)}{n_w^{\ 2}(i-1,f_c)} \quad (23)$$

[0091] α(t,f_c)由s形函数产生,由式(24)定义:

[0092] $$\alpha(i,f_c)=\frac{1}{1+\exp(-SNR(i,f_c))} \quad (24)$$

[0093] (4)根据公式(15)计算出自适应的掩蔽阈值z(i,f_c)。

[0094] 步骤四,构建深度卷积神经网络模型(DCNN);

[0095] 由于深度学习中的深度卷积神经网络(DCNN)有对复杂特征优秀的抽象和建模能力,所以本文通过DCNN对含噪语音提取的特征进行建模去估计。然后再用估计的IRM和IBM与含噪语音去合成增强后的语音。DCNN模型的结构一般由3部分组成:输入层、隐含层和输出层。输入层用来输入含噪语音的特征参数,隐含层由卷积层、池化层层和全连接层组成,输出层用来输出估计的目标。它们之间通过传递函数来传递数据。

[0096] 本文构建的DCNN模型的网络结构如图5所示。由1个输入层,5个隐含层和1个输出层构成。因为随着隐含层数目太少,不能很好得学习输入和输出之间的映射关系,但随着隐层数目的增多,网络结构变得复杂,它的建模能力下降。实验中发现隐含层数目为5时,它的性能较好。其中输入层各节点代表MR-GFCC的特征参数(432维);隐含层中的卷积层1有64个卷积滤波器,大小为7x7;池化层2采用的Max-Poling,滤波器个数为64,大小为3x3;卷积层2有128个卷积滤波器,大小为3x3,池化层4也采用Max-Poling,滤波器个数为128,大小3x3;全连接层5的神经元个数为1024;输出层的各节点代表一帧的gammatone滤波器组64个频率通道的自适应掩蔽值。输入层和隐含层之间的传递函数采用sigmoid函数,sigmoid函数是一种非线性函数,输出范围在(0,1)之间,使数据在DCNN模型传递的过程中不容易发散,输

出层的传递函数是softmax函数。

[0097] 步骤五，将步骤四中提取的特征参数$v_a(i,f_c)$和步骤三计算的自适应掩蔽阈值$z(i,f_c)$分别作为DCNN的输入和输出，采用自适应学习率的随机梯度下降算法对网络进行训练，训练完成以后，保存网络的权值和偏置，其训练采用的是线下训练。

[0098] 步骤六，按步骤二的方法提取测试的含噪语音的特征参数，输入到步骤五已经训练好的深度卷积神经网络中，输出一个自适应的掩蔽阈值$Y(i,f_c)$，如公式(25)所示；

[0099] $$Y(i,f_c)=\theta(X(i,f_c)) \quad (25)$$

[0100] 式中$\theta()$表示训练好的DCNN网络模型参数，$X(i,f_c)$表示测试的含噪语音的特征参数，$Y(i,f_c)$表示DCNN估计出的自适应掩蔽阈值。

[0101] 步骤七，利用测试的含噪语音和步骤六估计出的自适应的掩蔽阈值去合成增强后的语音。

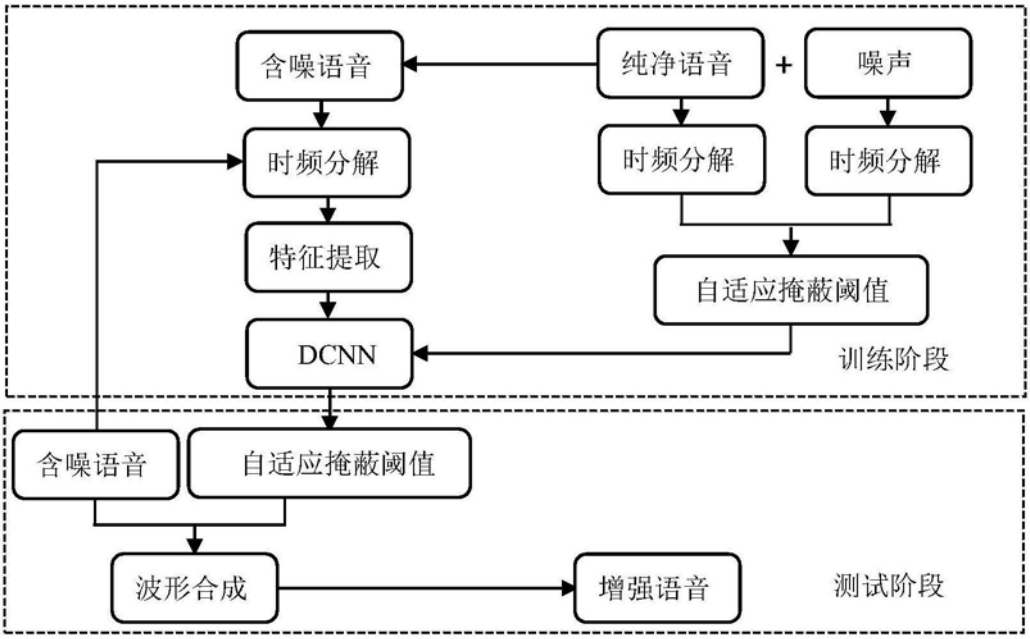

含噪语音
纯净语音
+
噪声
时频分解
时频分解
时频分解
特征提取
自适应掩蔽阈值
DCNN
训练阶段
含噪语音
自适应掩蔽阈值
波形合成
增强语音
测试阶段

图1

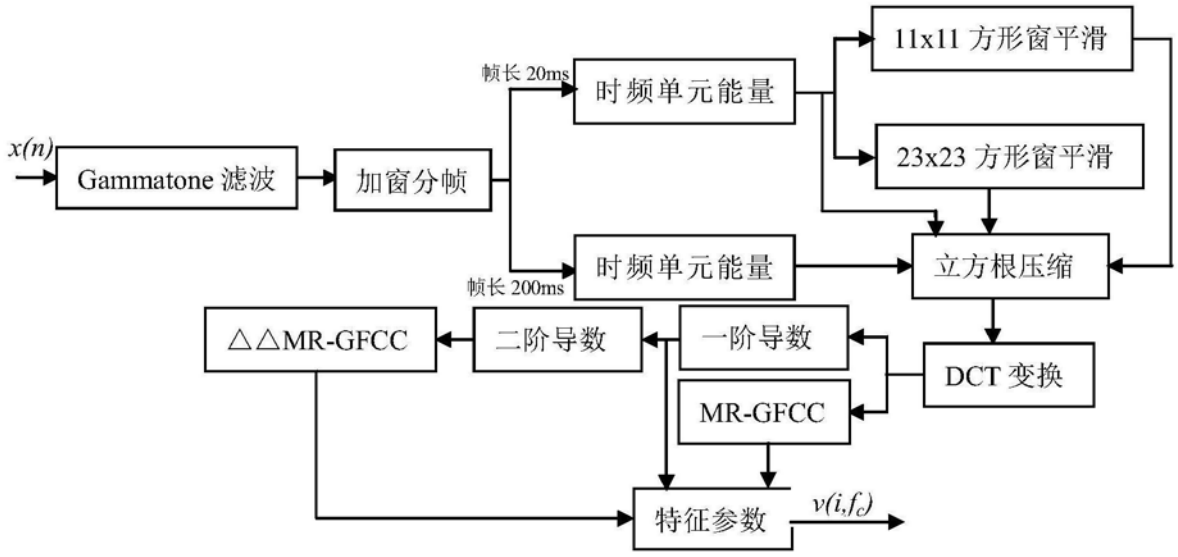

x(n)
Gammatone 滤波
加窗分帧
帧长 20ms
帧长 200ms
时频单元能量
时频单元能量
11x11 方形窗平滑
23x23 方形窗平滑
立方根压缩
DCT 变换
一阶导数
二阶导数
△△MR-GFCC
MR-GFCC
特征参数
v(i,f_c)

图2

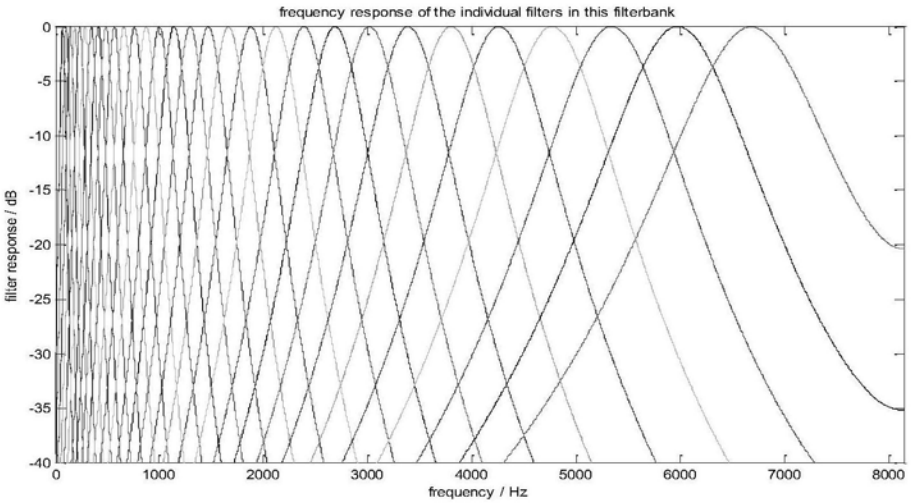

frequency response of the individual filters in this filterbank
filter response / dB
frequency / Hz

图3

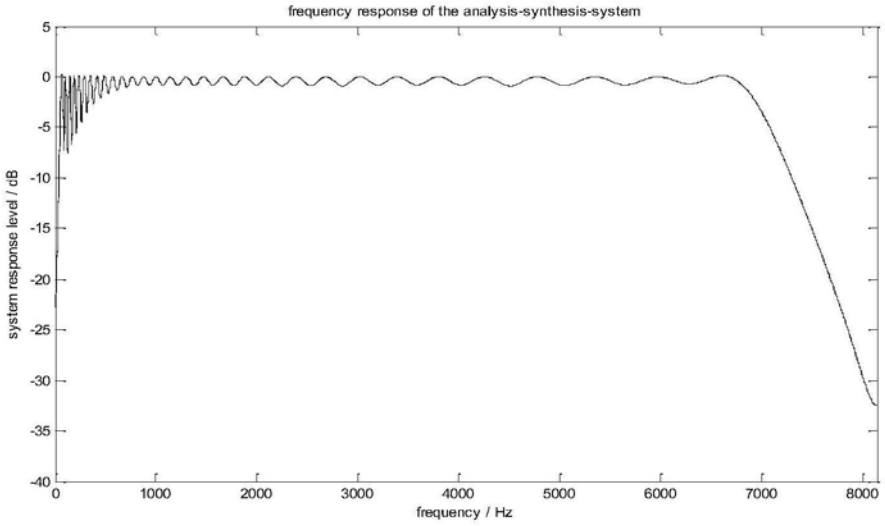

frequency response of the analysis-synthesis-system
system response level / dB
frequency / Hz

图4

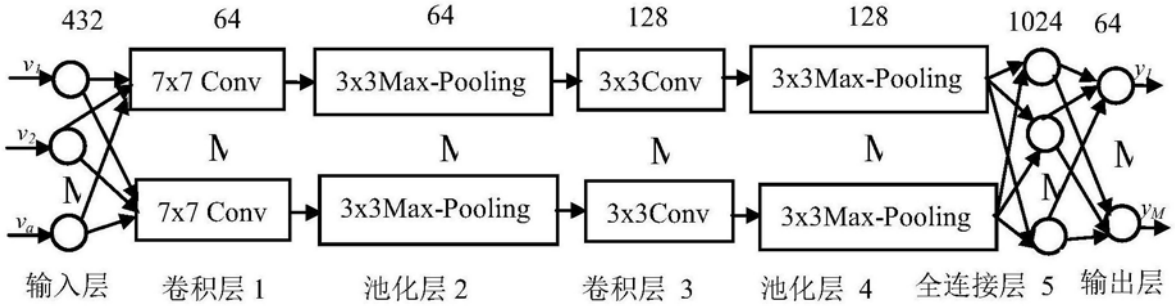

432
64
64
128
128
1024
64
7x7 Conv
3x3Max-Pooling
3x3Conv
3x3Max-Pooling
7x7 Conv
3x3Max-Pooling
3x3Conv
3x3Max-Pooling
输入层
卷积层 1
池化层 2
卷积层 3
池化层 4
全连接层 5
输出层

图5

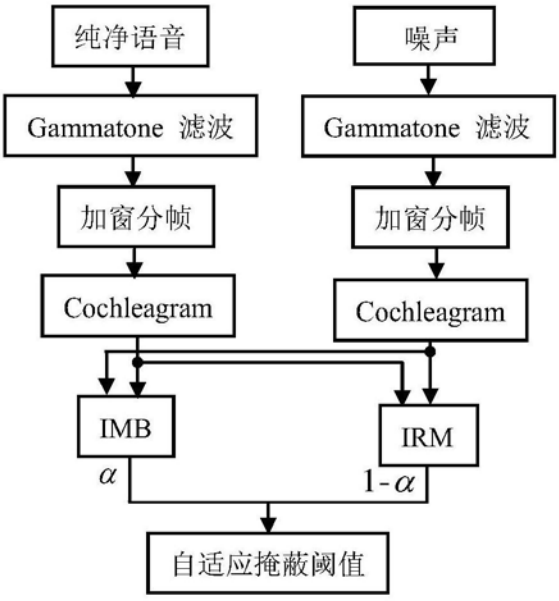

纯净语音
噪声
Gammatone 滤波
Gammatone 滤波
加窗分帧
加窗分帧
Cochleagram
Cochleagram
IMB
IRM
α
1-α
自适应掩蔽阈值

图6