(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110767244 B
(45)授权公告日 2024.03.29

(21)申请号 201810827229.0

(22)申请日 2018.07.25

(65)同一申请的已公布的文献号
申请公布号 CN 110767244 A

(43)申请公布日 2020.02.07

(73)专利权人 中国科学技术大学
地址 230026 安徽省合肥市包河区金寨路96号
专利权人 北京三星通信技术研究有限公司

(72)发明人 杜俊 高天 屠彦辉 王立众 杨磊 徐学淼

(74)专利代理机构 北京凯特来知识产权代理有限公司 11260
专利代理师 郑立明 郑哲

(51)Int.Cl.
*G10L 21/02*(2013.01)
*G10L 21/0208*(2013.01)
*G10L 25/30*(2013.01)

(56)对比文件
CN 101512573 A,2009.08.19
CN 103531204 A,2014.01.22
CN 107077860 A,2017.08.18
CN 107452389 A,2017.12.08
CN 107845389 A,2018.03.27
US 2016111107 A1,2016.04.21
US 2017061978 A1,2017.03.02
文仕学;孙磊;杜俊.渐进学习语音增强方法在语音识别中的应用.小型微型计算机系统.2018,(01),全文.

审查员 贾佳

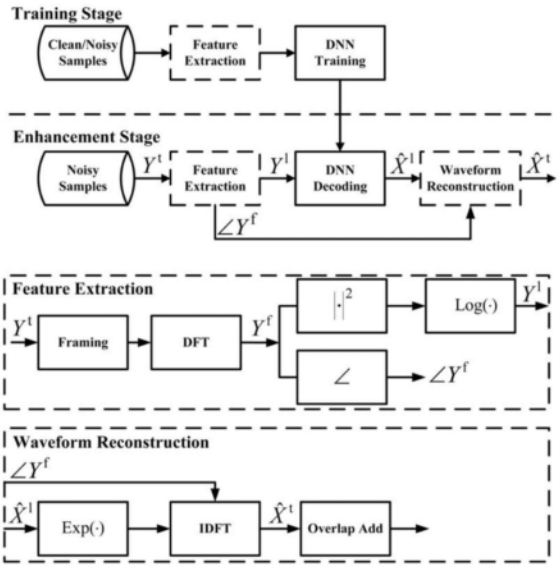
权利要求书2页 说明书7页 附图1页

(54)发明名称

# 语音增强方法

(57)摘要

本发明公开了一种语音增强方法,包括:提取各语音帧的声学特征;利用干净语音的与噪声语音的样本对渐进式双输出神经网络模型进行训练,利用训练后的渐进式双输出神经网络模型估计各语音帧的理想软掩蔽,并进行声学特征的增强处理;如果应用到人耳,则利用增强后的声学特征对波形进行重构,得到可主观测听的波形;如果应用到语音识别系统,则将估计到的理想软掩蔽应用到输入语音的声学特征上上,得到掩蔽后的声学特征,然后对波形进行重构得到增强后的语音。本发明上述方案可以满足人耳降噪需求和提升带噪语音的识别准确率。




CN 110767244 B

1.一种语音增强方法,其特征在于,包括:

提取各语音帧的声学特征;

利用干净语音的与噪声语音的样本对渐进式双输出神经网络模型进行训练,利用训练后的渐进式双输出神经网络模型估计各语音帧的理想软掩蔽,并进行声学特征的增强处理;

应用到人耳,则利用增强后的声学特征对波形进行重构,得到可主观测听的波形,包括:

首先,计算$\hat{X}'(d)$:

$$\hat{X}'(d)=\exp\{\hat{X}(d)/2\}\exp\{j\angle Y(d)\};$$

上式中,$\hat{X}(d)$为实数域上的定义,表示增强后的对数功率谱特征,$\hat{X}'(d)$也是增强后的对数功率谱特征,为复数域上的定义;∠Y(d)是指从输入语音中得到的相位信息;

然后,反向离散傅里叶变换重构得到增强后的时域语音$\hat{x}(l)$:

$$\hat{x}(l)=\frac{1}{L}\sum_{k=0}^{L-1}\hat{X}'(k)e^{j2\pi kl/L};$$

其中,L为提取各语音帧的声学特征时做离散傅里叶变换的点数;

最后,通过重叠相加算法合成整个句子的波形。

2.根据权利要求1所述的一种语音增强方法,其特征在于,所述提取各语音帧的声学特征包括:

对输入的语音信号进行分帧处理,获得语音帧序列;

声学特征采用对数功率谱特征,在提取各语音帧对数功率谱特征时,通过傅立叶变换和取模得到频域信号:

$$Y(d)'=\sum_{l}^{L-1}y(l)h(l)e^{-j2\pi dl/L}\quad d=0,1,\ldots,L-1$$

上式中,d为频率维度,h(l)为窗函数,L为做离散傅里叶变换的点数;

对数功率谱特征定义为:

$Y(d)=\log|Y(d)'|^2$ d=0,1,...,D-1;

上式中,D=L/2+1。

3.根据权利要求2所述的一种语音增强方法,其特征在于,该方法还包括:将提取的声学特征作为渐进式双输出神经网络模型的输入之前,还进行连续帧的拼接,拼接时以一定数量的帧拼接后的数据作为一个样本,样本的中心帧的标注作为其所在样本的标注。

4.根据权利要求1所述的一种语音增强方法,其特征在于,所述渐进式双输出神经网络模型按照信噪比逐渐增加的方式去学习最终目标,最终训练好的渐进式双输出神经网络模型能够预测各个时频点的理想软掩蔽,还能够对声学特征进行增强处理,即预测干净语音的对数功率谱特征。

5.根据权利要求1或4所述的一种语音增强方法,其特征在于,预测干净的对数功率谱特征的公式为:

$$\log(\hat{X}^2(t,d))\approx\log(I\hat{R}M(t,d))+\log((Y^2(t,d))$$

其中，$\hat{X}^2(t,d)$表示预测到的干净语音的对数功率谱特征，$\hat{IRM}(t,d)$表示理想软掩蔽，log((Y$^2$(t,d))＝Y(d)，Y(d)为提取的对数功率谱特征，d为频率维度，t为时间。

6.根据权利要求1所述的一种语音增强方法，其特征在于，基于最小批模式的随机梯度下降算法来提升渐进式双输出神经网络模型学习的收敛速度，表示为：

$$E=\frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}[(X_{n,d}^{l\hat{}ps1}\left(\mathbf{W}^l,\mathbf{b}^l\right)-X_{n,d}^{lps1})^2+X_{n,d}^{l\hat{}ps2}\left(\mathbf{W}^l,\mathbf{b}^l\right)-X_{n,d}^{lps2})^2+\ldots$$
$$+X_{n,d}^{l\hat{}psK}\left(\mathbf{W}^l,\mathbf{b}^l\right)-X_{n,d}^{lpsK})^2+(X_{n,d}^{i\hat{}rm}\left(\mathbf{W}^l,\mathbf{b}^l\right)-X_{n,d}^{irm})^2]$$

上式中，E是渐进式双输出神经网络模型学习的平均平方误差，$X_{n,d}^{l\hat{}ps(1\ldots K)}\left(\mathbf{W}^l,\mathbf{b}^l\right)$、$X_{n,d}^{lps(1\ldots K)}$对应的表示第1…K个渐进式学习目标在第n帧，第d个频率维的增强对数功率谱特征、目标的对数功率谱特征；$X_{n,d}^{i\hat{}rm}\left(\mathbf{W}^l,\mathbf{b}^l\right)$、$X_{n,d}^{irm}$对应的表示估计的理想软掩蔽、目标理想软掩蔽；N表示最小批的大小，即样本的数量；D对数功率谱特征向量的总维度；(W$^l$,b$^l$)表示在第l层有待学习的权重和偏置的参数。

# 语音增强方法

## 技术领域

[0001] 本发明涉及语音处理技术领域,尤其涉及一种语音增强方法。

## 背景技术

[0002] 语音识别是让机器听懂人说的话,也就是要将人类语音中的词汇内容转化为计算机可以识别的输入。近20年来,尤其是近几年深度学习的引入,使得语音识别技术取得了显著成效,开始从实验室走向市场。目前基于语音识别技术的语音输入,语音检索,语音翻译等得到了广泛的运用。众所周知,在噪声环境下如果我们不采取任何措施,那么自动语音识别的性能就会大幅下降,主要原因就是带噪语音分布于声学模型分布之间的差异。为了提高噪声环境下的识别准确率,语音增强算法通常用语音识别的前处理,其通过变换将带噪语音尽可能恢复到干净状态来匹配声学模型分布。

[0003] 语音增强是语音信号处理领域的一个重要分支,早在对语音信号研究的起初,噪声就已经是一个被关注的问题,因为在现实生活环境中,语音都伴随着受噪声的干扰而产生。而语音信号处理一般只对语音的内容、说话人、语种等感兴趣,噪声作为干扰项,一般都需要预先被去除掉,但是考虑到语音和噪声的产生过程是非线性的和复杂的,因而去噪的过程很困难。在过去的几十年中,有很多无监督的语音增强方法被提出,它们都是通过先估计噪声的谱信息,然后从带噪语谱中将估计的噪声谱减掉以得到对干净语音谱的预测。但是由于噪声的随机性和突变性,让对噪声的跟踪和估计变得困难。同时在传统的语音增强方法中,考虑到噪声和语音间的相互作用关系很复杂,就需要一些对信号间的独立性假设以及对特征分布的高斯性假设。由于这些假设的存在,首先就导致了传统的语音增强方法残留很多噪声,甚至是音乐噪声。其次,语音的细节也在较大程度上受到破坏,这主要体现在对低信噪比语音的增强中。再者,极端非平稳噪声一直是传统语音增强方法中比较棘手的地方,因为非平稳噪声的突发性,使得它始终处于被欠估计状态,难以从带噪语音中去掉,可实际声学环境中,各种非平稳噪声又是大概率发生事件。最后,传统的语音增强方法易引入一些非线性失真,使得其对后端的语音识别产生破坏作用。

[0004] 谱减法是最经典的语音增强算法之一,最开始用于语音增强,后来逐步用到语音识别中。此方法分为噪声更新和噪声消除两部分。在噪声更新时,首先需要确定检测出非语音段,然后利用当前帧和历史长时信息结合的方法来估计出噪声谱,由于长时平均的存在,需要假设噪声是慢变的,因此在快变的非稳态噪声的情况下,谱减法就不起作用。在噪声消除时,通过从带噪语音谱中减去估计的噪声谱来得到干净语音谱的估计。由于谱减法存在过减的问题,需要在每个频带设置一个和信噪比相关的过减因子。谱减法的非线性操作会在增强的信号中产生残留的音乐噪声,在实际应用中要采用一些特定的解决方案。此外,一个可靠的语音/非语音检测模块也是至关重要的。

[0005] 与谱减法相关的一个方法是维纳滤波。该方法中,需要设计一个线性滤波器来最小化干净语音和对带噪语音滤波后的语音信号之间的均方误差。在此准则下得到的传输函数是干净信号和带噪信号的相关谱与待噪信号功率谱之间的比值,其中相关谱采用带噪谱

与估计的噪声谱的差值来近似。由此可见维纳滤波和谱减法关系密切,但前者运算复杂度更高,同样对非稳态噪声二者都会失效。随后,一种基于语音存在概率的软判决增强方法被提出(McAulay and Malpass,1980),它能尽最大可能地降低语音的失真。而革命性的语音增强方法是1984年Ephraim和Malah提出的基于最小均方误差的语音幅度谱估计,随后考虑到人耳对声强的感知是非线性的,因而对数谱域的最小均方误差的估计被提出。而目前该类方法中,用的最普遍的是Israel Cohen提出的最小控制的迭代平均的(Minima Controlled Recursive Averaging,MCRA)噪声估计方法。最小控制的迭代平均的噪声估计方法相比较之前的噪声估计方法而言,具有估计误差更小,对非平稳噪声跟踪的比较快的特点。可以认为该方法是传统单声道语音增强算法中目前为止最优的方法;然而,该方法在平稳噪声且信噪比较高的情况下,可以提高语音识别系统的识别准确率。但是在背景噪声较大的情况下,提升有限,甚至会降低识别率。主要原因是,目前的识别系统本身的鲁棒性已经很强,如果增强算法对目标语音有破坏,反而会带来负面影响。

[0006] 上面讨论的是传统单声道无监督语音增强方法,针对无监督语音增强算法中存在的一些问题,自从上世纪八十年代末,就有基于有监督训练的模型的方法用于语音增强任务中,例如基于非负矩阵分解的语音增强。非负矩阵分解(Non-negative matrix factorization,NMF)是指矩阵V可以被分解成矩阵W和矩阵H的乘积。但是特殊之处是,矩阵V,矩阵W和矩阵H都没有负的元素。考虑到音频的语谱图一般都是具有潜在的非负性,非负矩阵分解通常应用在音乐信号分离中,这是因为音乐信号相对于语音信号或者噪声信号而言,更具有固定的模式,因此比较方便对音乐信号进行训练并得到一些矩阵基。非负矩阵分解应用到信号分离领域中,主要的目标就是利用不同的信号基,从混合信号中分离出两路信号。利用以下步骤可以通过训练得到信号基。首先,定义目标函数,可以利用欧氏距离,也可以利用Kullback-Leibler散度,分别如下所示:

[0007] $$E=\left\|\mathbf{V}-\mathbf{WH}\right\|=\sum_{ij}(V_{ij}-(WH)_{ij})^2$$

[0008] $$D=\sum(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}}-V_{ij}+(WH)_{ij})$$

[0009] 基于非负矩阵分解的的语音降噪主要分为两个步骤:训练和降噪。在训练阶段,首先分别获得干净语音和噪声的特征基$W_{speech}$和$W_{noise}$。它们的大小都是$n_f\times n_b$,这里$n_f$指的是特征基的维度,而$n_b$指的是特征基的个数。目标公式如下所示:

[0010] $\min KLD(V_{speech}||W_{speech}H_{speech})$

[0011] $\min KLD(V_{noise}||W_{noise}H_{noise})$

[0012] 上式中,$H_{speech}$和$H_{noise}$指的是每帧信号产生对应基的实际信号的系数矩阵。它们的维度是特征的维度乘以总帧数。而在增强阶段,我们认为训练阶段得到的语音特征基和噪声的特征基,即$W_{speech}$和$W_{noise}$是可以在解码中沿用的,并且把二者拼接起来得到的总的特征基$W_{all}$,如下所示:

[0013] $W_{all}=[W_{speech}\ W_{noise}]$

[0014] 然后在给定的带噪语音信号的前提下,通过标准的梯度下降算法来求解$H_{all}$,最终可以将噪声信号和语音信号分离开来。基于非负矩阵分解的语音增强方法主要存在的问题是,噪声和干净语音的模型是单独训练的,并且假设了噪声和干净语音之间的独立性。这就

在无形中限制了该方法的性能上限。与此同时,基于深度神经网络的方法已经在语音分离相关任务中全面超过了非负矩阵分解方法的性能。

[0015] 对于有监督的语音增强方法,尤其是基于深度学习的方法,近些年来也蓬勃发展。随着深度神经网络在语音识别领域的成功应用,神经网络也可以被设计成一个精细的降噪滤波器。同时基于大数据训练,神经网络可以充分学习带噪语音和干净语音之间的复杂的非线性关系。另外神经网络的训练是离线学习的,如同人一样,它能记住一些噪声的模式,因而可以很好地抑制一些非平稳噪声;但是,如果训练数据跟测试数据不匹配的话,例如噪声类型不同,说话人差异性较大,系统性能则会大幅下降。

## 发明内容

[0016] 本发明的目的是提供一种语音增强方法,可以满足人耳降噪需求和提升带噪语音的识别准确率。

[0017] 本发明的目的是通过以下技术方案实现的:

[0018] 一种语音增强方法,包括:

[0019] 提取各语音帧的声学特征;

[0020] 利用干净语音的与噪声语音的样本对渐进式双输出神经网络模型进行训练,利用训练后的渐进式双输出神经网络模型估计各语音帧的理想软掩蔽,并进行声学特征的增强处理;

[0021] 如果应用到人耳,则利用增强后的声学特征对波形进行重构,得到可主观测听的波形;如果应用到语音识别系统,则将估计到的理想软掩蔽应用到输入语音的声学特征上上,得到掩蔽后的声学特征,然后对波形进行重构得到增强后的语音。

[0022] 由上述本发明提供的技术方案可以看出,基于渐进式双输出神经网络模型来做语音增强,既可以输出深度降噪后的语音来满足人耳的降噪需求,又可以同时输出部分降噪、提升了一定信噪比的语音来匹配后端数据驱动的识别模型;通过人工测听和客观指标度量,深度降噪后的语音在主观听感和各项指标上都取得了显著提升;通过结合语音识别模型,相比于不做降噪处理的识别结果,神经网络部分降噪后的语音有效提升了识别准确率。

## 附图说明

[0023] 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简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域的普通技术人员来讲,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附图。

[0024] 图1为本发明实施例提供的一种语音增强方法的流程图;

[0025] 图2为本发明实施例提供的渐进式双输出神经网络模型的示意图。

## 具体实施方式

[0026] 下面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,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明的实施例,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,都属于本发明的保护范围。

[0027] 语音增强方法通过变换将带噪语音尽可能恢复到干净状态,一般来说,语音增强主要是考虑人对增强信号的听觉感受是否变好,而对于识别系统来说更关心系统的错误率是否下降,这两者的目标有着一定的联系,但不能完全保持一致,原因很简单,人可能对语音信号中的某些畸变具有免疫功能,但识别系统也许会对它们很敏感。所以经常出现增强后的语音有很好的主观听感改善但是并不能带来语音识别准确率的提升。基于深度学习的语音增强面临的另一大挑战则是推广性的问题,这是基于深度学习的模型避免不了的难题。具体到语音增强,则难题出现在未知噪声、未知说话风格和极低信噪比等方面。

[0028] 对此,本发明实施例提供一种语音增强方法,通过多目标联合训练的方法,使得渐进式双输出神经网络模型能拥有更好地推广性并同时输出深度降噪后的语音和提升了一定信噪比的语音,既能满足人耳对带噪语音的降噪需求又能提升识别系统的准确率。如图1所示,本该方法的流程图,其主要包括如下几个部分:

[0029] 1、提取各语音帧的声学特征。

[0030] 1)对输入的语音信号进行分帧处理,获得语音帧序列。

[0031] 本步骤,可以通过加汉明窗进行对输入的语音(带噪语音)进行处理,得到每帧数据。示例性的,一般的汉明窗窗长可选择为32毫秒,窗移可为16毫秒,叠加部分可为16毫秒。

[0032] 2)声学特征采用对数功率谱特征,在提取各语音帧对数功率谱特征时,通过傅立叶变换和取模得到频域信号:

[0033] $$Y(d)'=\sum_{l}^{L-1}y(l)h(l)e^{-j2\pi dl/L}\quad d=0,1,\ldots,L-1;$$

[0034] 上式中,d为频率维度,h(l)为窗函数,L为做离散傅里叶变换的点数;做离散傅里叶变换的点数L如果能增加,即采样的信息点数更多,那么输入的特征将包含更多的信息,这也会有利于后续神经网络的学习。

[0035] 对数功率谱特征定义为:

[0036] $Y(d)=\log|Y(d)'|^2 d=0,1,\ldots,D-1$;

[0037] 本领域技术人员可以理解,由于STFT变换在频域上是对称的,所以只取前一半的点,即D=L/2+1,而对于后一半的点d=D,...,L-1,Y(d)通过对称准则获得,Y(d)=Y(L-d)。也即,在后续计算过程所涉及的Y(d)也只考虑前一半的点,在最后语音恢复时才考虑全部的点,具体如后文$\hat{x}(l)$公式所示,在计算$\hat{x}(l)$时,由于存在对称关系,因此,后一半点相应的数值也是已知量。

[0038] 示例性的,如果采样率是16kHz的波形文件,那么它的对数功率谱特征的维数就是257维(32ms*16Khz=每帧有512个采样点,D=512/2+1=257)。

[0039] 2)进行连续帧的拼接,拼接时以一定数量的帧拼接后的数据作为一个样本,样本的中心帧的标注作为其所在样本的标注。

[0040] 通常可以采用7帧拼接、11帧拼接或者15帧拼接;拼接得到的样本作为一个训练样本时中心帧的标注作为样本的标注。通常对于隐层节点数为1024或者2048的网络采用7帧进行拼接,对于3072或者4096甚至更大的网络可以采用11帧或者15帧拼接作为输入。在本案研究中对于隐层为1024的网络采用7帧拼接作为输入。

[0041] 2、利用干净语音与噪声语音样本对渐进式双输出神经网络模型进行训练,利用训

练后的渐进式双输出神经网络模型估计各语音帧的理想软掩蔽,并进行增强处理。

[0042] 所述渐进式双输出神经网络模型如图2所示,在每一层的学习目标都是随着信噪比不断增加,例如target1的信噪比是5db,target2的信噪比是10db。按照这样的信噪比逐渐增加的方式去学习最终目标,最终训练好的渐进式双输出神经网络模型能够预测各个时频点的理想软掩蔽$\hat{IRM}(t,d)$,还能够结合预测到的$\hat{IRM}(t,d)$对声学特征进行增强处理,即预测干净语音的对数功率谱特征(LPS)。

[0043] 对对数功率谱直接估计的方法我们称之为直接映射,直接映射的方法可以去除大量的噪声,在语音客观指标和主观听感上都有很大的改善。但是,对于对接后端识别模型的应用场景来说,这种方法在不重训练后端识别模型的情况下,增强后的语音与识别模型会有较大的不匹配存在,难以提升识别准确率。所以,在对接识别后端时,我们采用时频掩蔽(masking)的方法。该方法只去除部分噪声来提高一定程度的信噪比,这样就能保持与后端识别模型尽可能的匹配。

[0044] 下面介绍时频掩蔽的语音增强方法。理想二元时频掩蔽(Ideal binary mask,IBM)是一个从预先混合的语音和噪声的信号中构建的时频掩蔽。对于每一个IBM单元,如果它的信噪比大于预先设定的本地阈值(local SNR criterion,LC),那么它就是语音主导的单元,否则就是噪声主导的单元。IBM可以被定量地定义成:

[0045] $$IBM(t,d)=\begin{cases}1 & if\ \ SNR(t,d)>LC\\ 0 & otherwise\end{cases}$$

[0046] 虽然基于分类的IBM估计的语音增强能得到比较好的可懂度,但是语音质量损失严重,说话人信息等也丢失。而且十分依赖于IBM分类的正确与否,一旦IBM被判错,信息损失就非常严重。针对IBM的缺陷,本发明实施例采用理想软掩蔽的目标来训练深层神经网络,它可以既保证语音的可懂度,又保证语音的质量的特性。IRM现在也成为基于掩蔽方法中的主流。IRM的定义为:

[0047] $$IRM(t,d)=(\frac{X^2(t,d)}{X^2(t,d)+N^2(t,d)})^{\beta}=(\frac{SNR(t,d)}{SNR(t,d)+1})^{\beta};$$

[0048] 上式中,$X^2(t,d)$和$N^2(t,d)$分别是干净语音和噪声语音的对数功率谱特征,SNR(t,d)表示信噪比。β是一个可调的参数,一般设置成0.5,它就变成平方根的维纳增益。这里实际上是假设了噪声和语音之间的关系是独立的。

[0049] 利用$X^2(t,d)$和$N^2(t,d)$对渐进式双输出神经网络模型进行训练,可以通过学习他们之间的映射关系,来准确估计出IRM。因为在测试的时候,假设通过深层神经网络对某个T-F单元上的IRM已经有一个准确的估计值,即理想软掩蔽$\hat{IRM}(t,d)$。

[0050] 则通过$\hat{IRM}(t,d)$可以对干净语音的对数功率谱特征(也即增强后的对数功率谱特征)进行预测,公式如下:

[0051] $$\hat{X}^2(t,d)=\hat{IRM}(t,d)(X^2(t,d)+N^2(t,d))\approx\hat{IRM}(t,d)(Y^2(t,d));$$

[0052] 上式中,$Y^2(t,d)$是带噪语音的功率谱。具体到对数功率谱特征,干净语音的对数功率谱特征预测可写为:

[0053] $$\log(\hat{X}^2(t,d)) \approx \log(\hat{IRM}(t,d)) + \log((Y^2(t,d))$$ ;

[0054] 其中,$\log((Y^2(t,d))=Y(d)$。

[0055] 本领域技术人员可以理解,上述各个参数中所出现的t表示的是某个时间,例如,$X^2(t,d)$、$N^2(t,d)$;当考虑整个时间轴时,则省略其中的t,例如Y(d)其考虑的是整个时间轴。

[0056] 在本发明实施例中,反向传播算法是基于干净语音对数功率谱和增强后语音对数功率谱之间的最小均方误差,而采取对数功率谱的这种方式是和人的听觉系统更为一致。因为人耳对声强的感知是非线性关系,且声强越大压抑的程度越高。基于最小批模式的随机梯度下降算法可用来提升渐进式双输出神经网络模型学习的收敛速度,表示为:

[0057] $$E=\frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}[(X_{n,d}^{l\hat{pps}1}(\mathbf{W}^l,\mathbf{b}^l)-X_{n,d}^{lps1})^2+X_{n,d}^{l\hat{ps}2}(\mathbf{W}^l,\mathbf{b}^l)-X_{n,d}^{lps2})^2+\ldots$$ $$+X_{n,d}^{l\hat{psK}}(\mathbf{W}^l,\mathbf{b}^l)-X_{n,d}^{lpsK})^2+(X_{n,d}^{i\hat{rm}}(\mathbf{W}^l,\mathbf{b}^l)-X_{n,d}^{irm})^2]$$ ;

[0058] 上式中,E是渐进式双输出神经网络模型学习的平均平方误差,$X_{n,d}^{l\hat{ps}(1\ldots K)}(\mathbf{W}^l,\mathbf{b}^l)$、$X_{n,d}^{lps(1\ldots K)}$对应的表示第1…K个渐进式学习目标在第n帧,第d个频率维的增强对数功率谱特征、目标的对数功率谱特征,也即渐进式双输出神经网络模型的输出;$X_{n,d}^{i\hat{rm}}(\mathbf{W}^l,\mathbf{b}^l)$、$X_{n,d}^{irm}$对应的表示估计的理想软掩蔽、目标理想软掩蔽;N表示最小批的大小,即样本的数量;D是对数功率谱特征向量的总维度;$(W^1,b^1)$表示在第1层有待学习的权重和偏置的参数。

[0059] 本领域技术人员可以理解,$X_{n,d}^{lps(1\ldots K)}$、$X_{n,d}^{irm}$的具体值可以由用户根据实际情况来设定。

[0060] 用L表示整个隐层的数目,那么L+1就表示输出层。另外,需要注意的是渐进式双输出神经网络模型的输入特征都是经过高斯归一化后的,即整个训练数据的均值被规整到0,方差被规整到1。无论是带噪语音还是干净语音都用带噪的训练数据的全局均值和全局方差来规整。这样处理的一个好处,就是可以使得渐进式双输出神经网络模型的输入数据和输出数据经过了相同的变换,使得神经网络的学习更为容易。在准备好输入数据和输出数据之后,就可以用一个学习速率λ来开始更新网络的权重和偏置参数。

[0061] 3、神经网络解码和语音恢复

[0062] 1)如果应用到人耳,则利用增强后的声学特征(对数功率谱特征)对波形进行重构,得到可主观测听的波形。

[0063] 首先,计算$\hat{X}'(d)$:

[0064] $$\hat{X}'(d)=\exp\{\hat{X}(d)/2\}\exp\{j\angle Y(d)\}$$ ;

[0065] 上式中,$\hat{X}(d)$为实数域上的定义,为增强后的对数功率谱特征,其考虑整个时间轴,故省略了t,$\hat{X}'(d)$也是增强后的对数功率谱特征,但是为复数域上的定义;∠Y(d)是指从输入语音中得到的相位信息,因为人耳对相位的微小变化并不敏感。

[0066] 然后,反向离散傅里叶变换重构得到增强后的时域语音$\hat{x}(l)$:

[0067] $$\hat{x}(l)=\frac{1}{L}\sum_{k=0}^{L-1}\hat{X}'(k)e^{j2\pi kl/L}$$；

[0068] 最后，通过经典的重叠相加算法合成整个句子的波形。

[0069] 2)如果应用到语音识别系统，则将估计到的理想软掩蔽应用到输入语音的声学特征上(即带噪语音的对数功率谱特征)，得到掩蔽后的声学特征(即增强后的声学特征)，然后对波形进行重构得到增强后的语音。

[0070] 其中掩蔽后的声学特征也即增强后的对数功率谱特征，之后的对波形进行重构得到增强后的语音的实现方式可以参见前述应用到人耳时的处理方式。

[0071] 由于在机器学习中，最大问题是训练数据和测试数据的不匹配，从而使网络的估计存在偏差。具体到实际语音增强的应用中，当测试语音和训练语音差异性较大时，网络直接输出的增强后的对数功率谱特征会对干净语音有破坏，从而影响识别准确性。而理想掩蔽信号计算增强后的对数功率谱特征会减少对干净语音的破坏，同时会残留较多的噪声。对于语音识别系统，其噪声的鲁棒性很强，但是对语音破坏很敏感，所以用理想掩蔽信号计算增强后的对数功率谱特征更加适合识别系统。

[0072] 本发明实施例上述方案，基于渐进式双输出神经网络模型来做语音增强，既可以输出深度降噪后的语音来满足人耳的降噪需求，又可以同时输出部分降噪、提升了一定信噪比的语音来匹配后端数据驱动的识别模型；通过人工测听和客观指标度量，深度降噪后的语音在主观听感和各项指标上都取得了显著提升；通过结合语音识别模型，相比于不做降噪处理的识别结果，神经网络部分降噪后的语音有效提升了识别准确率。

[0073] 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上述实施例可以通过软件实现，也可以借助软件加必要的通用硬件平台的方式来实现。基于这样的理解，上述实施例的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中，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)执行本发明各个实施例所述的方法。

[0074] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明披露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求书的保护范围为准。

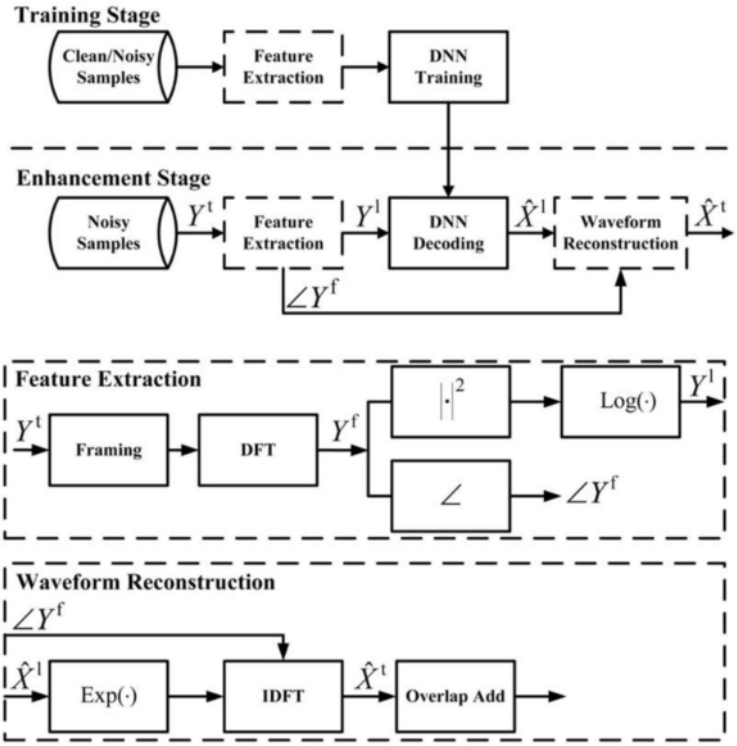

Training Stage
Clean/Noisy Samples
Feature Extraction
DNN Training
Enhancement Stage
Noisy Samples
$Y^t$
Feature Extraction
$Y^l$
DNN Decoding
$\hat{X}^l$
Waveform Reconstruction
$\hat{X}^t$
$\angle Y^f$
Feature Extraction
$Y^t$
Framing
DFT
$Y^f$
$|\cdot|^2$
Log(·)
$Y^l$
$\angle$
$\angle Y^f$
Waveform Reconstruction
$\angle Y^f$
$\hat{X}^l$
Exp(·)
IDFT
$\hat{X}^t$
Overlap Add

图1

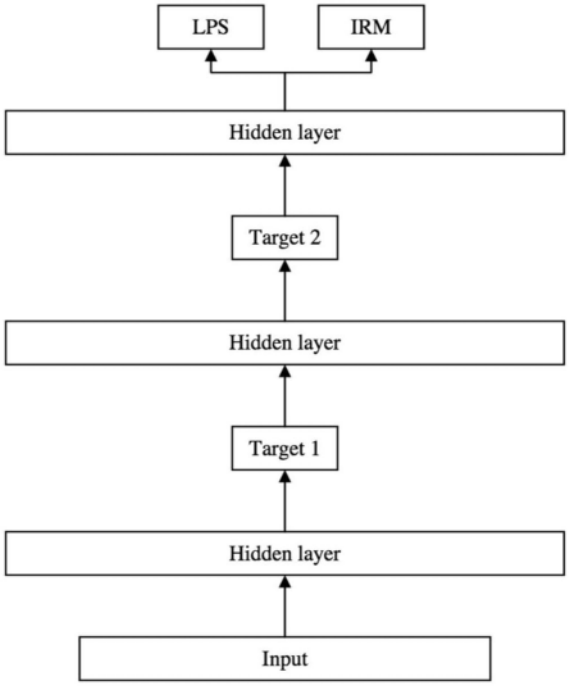

LPS
IRM
Hidden layer
Target 2
Hidden layer
Target 1
Hidden layer
Input

图2