

基于神经网络的端到端语音增强方法、装置

相关申请的交叉引用

本申请要求于 2021 年 04 月 06 日提交的申请号为 202110367186.4、名称为“基于神经网络的端到端语音增强方法、装置”的中国专利申请的优先权，该中国专利申请的全部内容通过引用全部并入本文。

技术领域

本公开涉及语音信号处理领域，具体而言，涉及一种基于神经网络的端到端语音增强方法、语音增强装置、计算机可读存储介质以及电子设备。

背景技术

近几年，随着深度学习技术的高速发展，语音识别技术的识别效果也得到很大提升，该技术在无噪音场景下语音的识别准确率，已达到可以替代人工的语音识别标准。

目前，语音识别技术主要可以应用于智能客服、会议录音转写、智能硬件等场景。但是，当背景环境有噪音时，如在智能客服通话时用户周围环境杂音或会议记录音频中的背景杂音等，受此类杂音影响，语音识别技术可能无法准确地识别说话人的语义，进而影响语音识别的整体准确率。

因此，如何提高有噪音情况下的语音识别准确率成为语音识别技术下一个需要攻关的难关。

需要说明的是，在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

根据本公开的第一方面，提供一种基于神经网络的端到端语音增强方法，包括：利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征；

对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。

在本公开的一种示例性实施例中，所述利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征，包括：

根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵；
对所述时域平滑参数矩阵作乘积运算得到所述时域卷积核的权重矩阵；
将所述时域卷积核的权重矩阵和所述原始语音信号作卷积运算，得到所述原始语音信号的时域平滑特征。

在本公开的一种示例性实施例中,所述根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵,包括:

初始化多个时域平滑因子;

基于预设的卷积滑窗和所述多个时域平滑因子得到时域平滑参数矩阵。

- 5 在本公开的一种示例性实施例中,所述对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取,得到增强语音信号,包括:

合并所述原始语音信号和所述原始语音信号的时域平滑特征,得到待增强语音信号;

以所述待增强语音信号为深度神经网络的输入,利用反向传播算法对所述时域卷积核的权重矩阵进行训练;

- 10 根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取,得到增强语音信号。

在本公开的一种示例性实施例中,所述以所述待增强语音信号为深度神经网络的输入,利用反向传播算法对所述时域卷积核的权重矩阵进行训练,包括:

将所述待增强语音信号输入深度神经网络中,并构建时域损失函数;

- 15 根据所述时域损失函数,利用误差反向传播算法对所述时域卷积核的权重矩阵进行训练。

在本公开的一种示例性实施例中,所述根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取,得到增强语音信号,包括:

将训练得到的权重矩阵与所述待增强语音信号中的原始语音信号作卷积运算,得到第

- 20 一时域特征图;

将训练得到的权重矩阵与所述待增强语音信号中的平滑特征作卷积运算,得到第二时域特征图;

组合所述第一时域特征图和所述第二时域特征图,得到所述增强语音信号。

根据本公开的第二方面,提供一种基于神经网络的端到端语音增强装置,包括:

- 25 时域平滑特征提取模块,用于利用时域卷积核对处理后的原始语音信号进行特征提取,得到所述原始语音信号的时域平滑特征;

组合特征提取模块,对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取,得到增强语音信号。

- 30 根据本公开的第三方面,提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现上述任意一项所述的方法。

根据本公开的第四方面,提供一种电子设备,包括:处理器;以及存储器,用于存储所述处理器的可执行指令;其中,所述处理器配置为经由执行所述可执行指令来执行上述任意一项所述的方法。

- 35 应当理解的是,以上的一般描述和下文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。显而易见地，下面描述中的附图仅仅是本公开的一些实施例，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。

图 1 示出了可以应用本公开实施例的一种端到端语音增强方法及装置的示例性系统架构的示意图；

图 2 示出了适于用来实现本公开实施例的电子设备的计算机系统的结构示意图；

图 3 示意性示出了根据本公开的一个实施例的端到端语音增强方法的流程图；

图 4 示意性示出了根据本公开的一个实施例的时域平滑特征提取的流程图；

图 5 示意性示出了根据本公开的一个实施例的增强语音信号获取的流程图；

图 6 示意性示出了根据本公开的一个实施例的组合特征提取的流程图；

图 7 示意性示出了根据本公开的一个实施例的端到端语音增强方法的流程图；

图 8 示意性示出了根据本公开的一个实施例的端到端语音增强装置的框图。

具体实施方式

现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本公开将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本公开的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本公开的技术方案而省略所述特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

此外，附图仅为本公开的示意性图解，并非一定是按比例绘制。图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

图 1 示出了可以应用本公开实施例的一种端到端语音增强方法及装置的示例性应用环境的系统架构的示意图。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103 中的一个或多个，网络 104 和服务 105。网络 104 用以在终端设备 101、102、103 和服务 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等

等。终端设备 101、102、103 可以是具有显示屏的各种电子设备，包括但不限于台式计算机、便携式计算机、智能手机和平板电脑等等。应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务。比如服务器 105 可以是多个服务器组成的服务器集群等。

5 本公开实施例所提供的端到端语音增强方法一般由服务器 105 执行，相应地，端到端语音增强装置一般设置于服务器 105 中。但本领域技术人员容易理解的是，本公开实施例所提供的端到端语音增强方法也可以由终端设备 101、102、103 执行，相应的，端到端语音增强装置也可以设置于终端设备 101、102、103 中，本示例性实施例中对此不做特殊限定。

10 图 2 示出了适于用来实现本公开实施例的电子设备的计算机系统的结构示意图。

需要说明的是，图 2 示出的电子设备的计算机系统 200 仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 2 所示，计算机系统 200 包括中央处理单元（CPU）201，其可以根据存储在只读存储器（ROM）202 中的程序或者从存储部分 208 加载到随机访问存储器（RAM）203 中的程序而执行各种适当的动作和处理。在 RAM 203 中，还存储有系统操作所需的各种程序和数
15 据。CPU 201、ROM 202 以及 RAM 203 通过总线 204 彼此相连。输入/输出（I/O）接口 205 也连接至总线 204。

以下部件连接至 I/O 接口 205：包括键盘、鼠标等的输入部分 206；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分 207；包括硬盘等的存储
20 部分 208；以及包括诸如 LAN 卡、调制解调器等的网络接口卡的通信部分 209。通信部分 209 经由诸如因特网的网络执行通信处理。驱动器 210 也根据需要连接至 I/O 接口 205。可拆卸介质 211，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 210 上，以便于从其上读出的计算机程序根据需要被安装入存储部分 208。

特别地，根据本公开的实施例，下文参考流程图描述的过程可以被实现为计算机软件
25 程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 209 从网络上被下载和安装，和/或从可拆卸介质 211 被安装。在该计算机程序被中央处理单元（CPU）201 执行时，执行本申请的方法和装置中限定的各种功能。

30 作为另一方面，本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该电子设备执行时，使得该电子设备实现如下述实施例中所述的方法。例如，所述的电子设备可以实现如图 3 至图 7 所示的各个步骤等。

35 需要说明的是，本公开所示的计算机可读介质可以是计算机可读信号介质或者计算机

可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF 等等，或者上述的任意合适的组合。

以下对本公开实施例的技术方案进行详细阐述：

在时域上，实际观测到的语音信号可以表示为纯净语音信号和噪声信号的加和，即：

$$y(n) = x(n) + w(n)$$

其中， $y(n)$ 表示时域带噪语音信号， $x(n)$ 表示时域纯净语音信号， $w(n)$ 表示时域噪声信号。

对语音信号做增强处理时，可以将带噪语音信号通过短时傅里叶变换（Short-Time Fourier Transform, STFT）从一维时域信号变为复数域二维变量 $Y(k, l)$ ，并取该变量的幅度信息，对应的有：

$$|Y(k, l)| = |X(k, l)| + |W(k, l)|$$

其中， $|Y(k, l)|$ 表示复数域语音信号的幅度信息， $|X(k, l)|$ 表示复数域纯净语音信号的幅度信息， $|W(k, l)|$ 表示复数域噪声信号的幅度信息， k 表示频率轴上第 k 个频率格， l 表示时间轴上第 l 个时间帧。

具体的，可以通过求解增益函数 $G(k, l)$ 实现语音信号的降噪。其中，可以将增益函数设为时变且频率依赖的函数，通过增益函数和带噪语音信号 $Y(k, l)$ ，可以得到预测的纯净语音信号 $\hat{x}(n)$ 的 STFT 参数 $\hat{X}(k, l)$ ，即：

$$\hat{X}(k, l) = G(k, l) \times |Y(k, l)|$$

也可以通过训练深度神经网络得到 $f_{\theta}(Y(k, l))$ 来估计纯净语音信号 $\hat{X}(k, l)$ ，即：

$$\hat{X}(k, l) = f_{\theta}(|Y(k, l)|)$$

上述语音增强方法中，在根据带噪语音信号 $Y(k, l)$ 中的幅度信息预测纯净语音信号 $\hat{x}(n)$ 时，并没有对 $Y(k, l)$ 的相位信息进行增强。如果不对相位信息进行增强，当 $Y(k, l)$ 的信噪比较高时，根据 $Y(k, l)$ 的相位信息和预测得到的 $\hat{X}(k, l)$ 恢复出的 $\hat{x}(n)$ 与实际的纯净语

音信号 $x(n)$ 差别不大。但是，当 $Y(k,l)$ 的信噪比较低，如信噪比为0db及以下时，如果只对幅度信息进行增强，而忽略相位信息，最终恢复出的 $\hat{x}(n)$ 和实际的纯净语音 $x(n)$ 差别就会变大，导致整体的语音增强效果较差。

基于上述一个或多个问题，本示例实施方式提供了一种基于神经网络的端到端语音增强方法，该方法可以应用于上述服务器 105，也可以应用于上述终端设备 101、102、103 中的一个或多个，本示例性实施例中对此不做特殊限定。参考图 3 所示，该端到端语音增强方法可以包括以下步骤 S310 和步骤 S320：

步骤 S310. 利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征；

步骤 S320. 对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。

在本公开示例实施方式所提供的语音增强方法中，通过利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征；对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。一方面，通过对原始语音信号中的幅度信息和相位信息均进行增强，可以提升语音增强的整体效果；另一方面，通过卷积神经网络对原始语音信号提取时域平滑特征，并结合深度神经网络可以实现时域降噪参数的自学习，进一步提升语音信号的质量。

下面，对于本示例实施方式的上述步骤进行更加详细的说明。

在步骤 S310 中，利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征。

端到端语音增强可以直接处理原始语音信号，避免通过中间变换提取声学特征。语音通信过程中环境噪声的干扰是不可避免的，实际观测到的原始语音信号一般为时域上的带噪语音信号。将原始语音信号进行特征提取之前，可以先获取该原始语音信号。

原始语音信号是一种连续变化的模拟信号，可以通过采样、量化及编码，将模拟的语音信号转化成离散的数字信号。示例性的，可以按一定的频率，每隔一段时间，测得模拟信号的模拟量的值，可以量化采样得到的点，并将量化的值用一组二进制来表示。因此，获取的原始语音信号可以用一个一维向量表示。

一种示例实施方式中，可以将原始语音信号输入深度神经网络中以进行时变的特征提取。例如，可以基于语音信号相邻帧之间的相关性，通过在时间维度进行平滑处理来计算该原始语音信号的局部特征，其中，可以对原始语音信号中的相位信息和幅度信息均进行语音增强。

可以对时域上的原始语音信号进行降噪处理，通过增强原始语音信号以提高语音识别的准确率。例如，可以利用深度神经网络模型进行语音增强，通过平滑算法对时域语音信号进行降噪处理时，可以将平滑算法并入深度神经网络的卷积模块当中，卷积模块中可以使用多层滤波器来实现不同特征的抽取，再由不同特征组合成新的不同特征。

示例性的,可以将时域平滑算法作为一维卷积模块并入深度神经网络中,该一维卷积模块可以是一个 TRAL (Time-Domain Recursive Averaging Layer, 时域递归平滑层) 模块,对应时间轴维度的噪声平滑。可以将原始语音信号作为 TRAL 模块的输入,通过 TRAL 模块对原始语音信号进行滤波处理,也就是进行时间轴维度的噪声平滑。例如,可以使用

5 加权移动平均法来预测待平滑时间轴上每个时间点的幅度谱信息,其中,加权移动平均法可以根据同一个移动段内不同时间的数据对预测值的影响程度(对应不同的权重)来预测未来值。

参考图 4 所示,可以根据步骤 S410 至步骤 S430 对时域语音信号进行噪声平滑:

步骤 S410. 根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵。

10 一种示例实施方式中,TRAL 模块可以采用多个时域平滑因子对原始输入信息进行处理,具体的,TRAL 模块对时域语音信号的平滑可以通过一个滑窗来实现,对应的平滑算法可以是:

$$R(n) = \sum_{i=1}^D \alpha^{D-i} (1 - \alpha) y(n), \quad \alpha = [\alpha_0 \quad \cdots \quad \alpha_N]$$

其中, n : 表示原始语音信号的采样点;

D : 表示滑窗宽度,其宽度可以根据实际情况进行设置,在本示例中,优选可以将滑窗宽度设置为 32 帧;

15

α : 时域平滑因子,表示对时域语音信号作平滑处理时,对滑窗宽度内每个采样点的语音信号 $y(n)$ 的利用程度, $[\alpha_0 \quad \cdots \quad \alpha_N]$ 为不同的平滑因子,每个平滑因子的取值范围为 $[0, 1]$, 对应于 α 的取值, TRAL 模块中的卷积核数量可以为 N ;

$y(n)$: 表示滑窗宽度内每个采样点的语音信号。本示例中,可以对每个采样点的语音信号加以利用,示例性的,第 32 帧采样点语音信号可以由滑窗宽度内的前面 31 帧采样点的语音信号组成;

20

另外,有 $i \in [1, D]$, 某一采样点离当前采样点越远时, α^{D-i} 的值越小,该采样点的语音信号的权重越小;离采样点的语音信号越近时, α^{D-i} 的值越大,该采样点的语音信号的权重越大;

25 $R(n)$: 表示由滑窗宽度内每个历史采样点的语音信号叠加得到新的语音信号,也是经过时域平滑得到的语音信号。

可以理解的是,在 TRAL 模块中,可以根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵,即可以根据滑窗宽度 D 和时域平滑因子 $\alpha = [\alpha_0 \quad \cdots \quad \alpha_N]$ 确定第一时域平滑参数矩阵 $[\alpha^0 \quad \cdots \quad \alpha^{D-1}]$ 和第二时域平滑参数矩阵 $[1 - \alpha]$ 。

30 步骤 S420. 对所述时域平滑参数矩阵作乘积运算得到所述时域卷积核的权重矩阵。

在对原始语音信号进行时域特征提取之前,可以先确定时域卷积核的权重矩阵。例如,可以初始化多个时域平滑因子 α , 如 $\alpha = [\alpha_0 \quad \cdots \quad \alpha_N]$, 并基于预设的卷积滑窗和多个时域平滑因子得到时域平滑参数矩阵。具体的,对时间轴进行平滑时,在 TRAL 模块中对

应可以有 N 个卷积核，每个卷积核对应不同的平滑因子，其中每个卷积核对应的第一时域平滑参数矩阵可以为 $[\alpha^0 \dots \alpha^{D-i}]$ ，结合第二时域平滑参数矩阵 $[1 - \alpha]$ ，如可以将第一时域平滑参数矩阵和第二时域平滑参数矩阵作乘积运算可以得到时域卷积核的最终权重矩阵 $N(\alpha)$ 。

- 5 步骤 S430. 将所述时域卷积核的权重矩阵和所述原始语音信号作卷积运算，得到所述原始语音信号的时域平滑特征。

可以将原始语音信号作为原始输入，该原始语音信号可以是一个 $1 * N$ 的一维向量，可以对该一维向量和时域卷积核的权重矩阵 $N(\alpha)$ 作卷积运算，得到原始语音信号的时域平滑特征。本示例中，利用卷积神经网络中卷积核的思想，将降噪算法做成卷积核，并通
10 过多卷积核的组合，在神经网络中实现了时变语音信号的降噪。而且，通过对时域上的带噪语音信号进行平滑，可以提高原始输入信息的信噪比，其中，输入信息可以包含带噪语音信号的幅度信息和相位信息。

在步骤 S320 中，对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。

- 15 参考图 5 所示，可以根据步骤 S510 至步骤 S530 得到增强语音信号：

步骤 S510. 合并所述原始语音信号和所述原始语音信号的时域平滑特征，得到待增强语音信号。

一种示例实施方式中，为了更好的保留原始输入的语音特征，可以将原始输入的特征和 TRAL 模块的输出进行拼接，这样既能保留原始语音信号的特征，又可以学习到深
20 层次特征。

对应的，深度神经网络的输入可以由原始输入 $y(n)$ 变为组合输入，该组合输入可以是：

$$I_i(n) = \begin{cases} y(n), & i = 1 \\ R(n), & otherwise \end{cases}$$

其中， $I_i(n)$ 是组合得到的待增强语音信号， $y(n)$ 是原始输入的带噪语音信号， $R(n)$ 是 TRAL 模块的输出，即沿时间轴平滑后的语音信号。

本示例中，TRAL 模块中的一个滤波器的平滑因子为 0，即对原始信息不做平滑处理，
25 保持原始输入。其他滤波器通过不同的平滑因子可以实现对原始信息的不同平滑处理，从而既保持了原始信息的输入，又增加了深度神经网络的输入信息。而且，TRAL 模块兼具由专家知识开发出的降噪算法的可解释性和并入神经网络以后形成的强大拟合能力，是具有可解释性的神经网络模块，可以有效地将语音降噪领域的高级信号处理算法与深度神经网络进行结合。

- 30 步骤 S520. 以所述待增强语音信号为深度神经网络的输入，利用反向传播算法对所述时域卷积核的权重矩阵进行训练。

可以将待增强语音信号输入深度神经网络中，并构建时域损失函数，如均方误差损失函数。基于深度神经网络，在时域上的语音增强任务可以表示为：

$$\hat{x}(n) = f_{\theta}(I_i(n))$$

一种示例实施方式中，可以构建具有编码器-解码器结构的 U-Net 卷积神经网络模型作为端到端语音增强模型，并将 TRAL 模块并入该神经网络模型中。U-Net 卷积神经网络模型可以包括全卷积部分（Encoder 层）和反卷积部分（Decoder 层）。其中，全卷积部分可以用于提取特征，得到低分辨率的特征图，相当于时域中的滤波器，可以对输入信息进行编码，也可以对上一层 Encoder 层的输出信息再次进行编码，实现高层特征的抽取；反卷积部分可以将小尺寸的特征图通过上采样得到与原始尺寸相同的特征图，即可以对 Encoder 层编码后的信息进行解码。另外，Encoder 层和 Decoder 层之间可以进行跳跃连接，以增强解码效果。

具体的，可以根据：

$$f_{\theta}(I_i(n)) = g^L(w^L g^{L-1}(\dots g^1(w^1 * I_i(n))))$$

计算得到增强语音信号。其中， $I_i(n)$ 为 U-Net 卷积神经网络中的最终输入信息，即组合得到的待增强语音信号； w^L 可以表示 U-Net 卷积神经网络中第 L 层的权重矩阵； g^L 可以表示第 L 层的非线性激活函数。可以看出，Encoder 层和 Decoder 层的权重矩阵 w^L 可以通过参数自学习的方式实现，即滤波器可以通过梯度回传的方式，在训练过程中通过学习自动生成，先生成低层级特征，再从低层级特征组合出高层级特征。

根据时域损失函数，利用误差反向传播算法对时域卷积核的权重矩阵 $N(\alpha)$ 、神经网络的权重矩阵 w^L 进行训练。示例性的，神经网络模型的训练过程可以采用 BP（error Back Propagation，误差方向传播）算法，通过随机初始化参数，随着训练的加深，不断更新参数。例如，可以根据原始输入从前向后依次计算，得到输出层的输出；可以计算当前输出与目标输出的差距，即计算时域损失函数；可以利用梯度下降算法、Adam 优化算法等最小化时域损失函数，从后向前依次更新参数，也就是依次更新时域卷积核的权重矩阵 $N(\alpha)$ 、神经网络的权重矩阵 w^L 。

其中，误差回传过程可以是第 j 次的权重值就是第 $j-1$ 次的权重减去学习率与误差梯度，即：

$$\alpha_j = \alpha_{j-1} - \lambda \frac{\partial E(\alpha_{j-1})}{\partial \alpha_{j-1}} = \alpha_{j-1} - \lambda \frac{\partial E}{\partial I} \frac{\partial I}{\partial \alpha_{j-1}}$$

其中， λ 为学习率， ∂E 为由 U-Net 卷积神经网络回传到 TRAL 的误差， $\partial E(\alpha_{j-1})/\partial \alpha_{j-1}$ 为由 U-Net 卷积神经网络回传到 TRAL 的误差梯度，并且可以根据：

$$\frac{\partial E}{\partial I} = (w^1)^T * (g^1)' * (w^2)^T * (g^2)' \dots (w^L)^T * (g^L)'$$

$$\frac{\partial I}{\partial \alpha_{j-1}} = -\alpha_{j-1}^{D-1} I_i(n) - \sum_{i=1}^{D-1} \alpha_{j-1}^{D-i} I_i(n)$$

对平滑因子矩阵 $\alpha = [\alpha_0 \dots \alpha_N]$ 进行更新。具体的，可以先设置深度神经网络的初始权重 w_0^L ；将第 i 个样本语音信号作为参考信号，添加噪声信号构建对应的第 i 个原始语音

信号; 根据第 i 个原始语音信号, 通过深度神经网络前向计算, 获取对应的第 i 个第一特征; 根据第 i 个第一特征以及第 i 个样本语音信号, 计算均方误差, 获取第 i 个均方误差; 将第 i 个样本语音信号求平方、取平均, 并与获取的第 i 个均方误差作比值, 获取训练过后每一层的最优权重系数 w^L ; 可以根据该最优权重系数, 计算深度神经网络的输出值。

- 5 步骤 S530. 根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取, 得到增强语音信号。

可以将原始语音信号输入 TRAL 模块, 并将原始语音信号以及 TRAL 模块的输出合并输入到 U-NET 卷积神经网络模型中, 对各个权重因子进行训练后, 可以对原始输入、TRAL 模块输出进行组合特征的抽取。

- 10 参考图 6 所示, 可以根据步骤 S610 至步骤 S630 实现组合特征提取:

步骤 S610. 将训练得到的权重矩阵与所述待增强语音信号中的原始语音信号作卷积运算, 得到第一时域特征图;

- 15 可以将原始语音信号作为深度神经网络的输入, 该原始语音信号可以是一个 $1 * N$ 的一维向量, 可以对该一维向量和训练得到的权重矩阵 $\bar{w}_0$ 作卷积运算, 得到第一时域特征图。

步骤 S620. 将训练得到的权重矩阵与所述待增强语音信号中的平滑特征作卷积运算, 得到第二时域特征图;

可以将平滑特征作为深度神经网络的输入, 以对该平滑特征和训练得到的权重矩阵 $\bar{w}_a$ 作卷积运算, 得到第二时域特征图。

- 20 步骤 S630. 组合所述第一时域特征图和所述第二时域特征图, 得到所述增强语音信号。

- 本示例中, 通过将时域信号平滑算法做成一维 TRAL 模块, 并且可以成功并入深度神经网络模型, 与卷积神经网络、递归神经网络、全连接神经网络均能理想结合, 实现梯度传导, 使得 TRAL 模块内的卷积核参数, 也即降噪算法参数可以由数据驱动, 无需专家知识作为先验信息, 就可以得到统计意义上的最优权重系数。另外, 通过直接对带噪的时域语音信号做语音增强来预测出纯净语音信号时, 可以利用该时域语音信号中的幅度信息和相位信息, 该语音增强方法更实际、语音增强效果更好。

图 7 示意性的给出了 TRAL 模块与深度神经网络结合的语音增强的流程图, 该过程可以包括步骤 S701 至步骤 S703:

- 30 步骤 S701. 输入语音信号 $y(n)$, 该信号为带噪语音信号, 包括纯净语音信号和噪声信号;

步骤 S702. 将该带噪语音信号输入 TRAL 模块, 对该带噪语音信号的相位信息和幅度信息提取时域平滑特征, 得到沿时间轴降噪后的语音信号 $R(n)$;

- 35 步骤 S703. 输入深度神经网络: 将该带噪语音信号 $y(n)$ 和沿时间轴降噪后的语音信号 $R(n)$ 合并输入深度神经网络中, 以进行组合特征的提取, 得到增强后的语音信号。

本示例中，在端到端（即序列到序列）的语音增强任务中加入了时域信号平滑算法，并将该算法做成一维卷积模块，即 TRAL 模块，相当于增加包含专家知识的滤波器，可以提高原始输入信息的信噪比，以及增加深度神经网络的输入信息，进而可以提升 PESQ（Perceptual Evaluation of Speech Quality，语音质量感知评价指标）、STOI（Short-Time Objective Intelligibility，短时客观可懂度指标）、fw SNR（frequency-weighted SNR，频率加权信噪比）等语音增强评测指标。另外，TRAL 模块与深度神经网络可以通过梯度回传的方式连接，能够实现降噪参数的自学习，进而可以得到统计意义的最优参数，该过程无需人工设计算子或专家知识作为先验。即此 TRAL 模块既纳入了具有信号处理领域的专家知识，又结合了深度神经网络的梯度回传算法进行参数寻优。将两者的优势进行了融合，提升了最终的语音增强效果。

在本公开示例实施方式所提供的语音增强方法中，通过利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征；对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。一方面，通过对原始语音信号中的幅度信息和相位信息均进行增强，可以提升语音增强的整体效果；另一方面，通过卷积神经网络对原始语音信号提取时域平滑特征，并结合深度神经网络可以实现时域降噪参数的自学习，进一步提升语音信号的质量。

应当注意，尽管在附图中以特定顺序描述了本公开中方法的各个步骤，但是，这并非要求或者暗示必须按照该特定顺序来执行这些步骤，或是必须执行全部所示的步骤才能实现期望的结果。附加的或备选的，可以省略某些步骤，将多个步骤合并为一个步骤执行，以及/或者将一个步骤分解为多个步骤执行等。

进一步的，本示例实施方式中，还提供了一种基于神经网络的端到端语音增强装置，该装置可以应用于一服务器或终端设备。参考图 8 所示，该端到端语音增强装置 800 可以包括时域平滑特征提取模块 810 和组合特征提取模块 820，其中：

时域平滑特征提取模块 810，用于利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征；

组合特征提取模块 820，对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。

在一种可选的实施方式中，时域平滑特征提取模块 810 包括：

参数矩阵确定单元，根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵；

权重矩阵确定单元，用于对所述时域平滑参数矩阵作乘积运算得到所述时域卷积核的权重矩阵；

时域运算单元，用于将所述时域卷积核的权重矩阵和所述原始语音信号作卷积运算，得到所述原始语音信号的时域平滑特征。

在一种可选的实施方式中，参数矩阵确定单元包括：

数据初始化子单元，用于初始化多个时域平滑因子；

矩阵确定子单元,用于基于预设的卷积滑窗和所述多个时域平滑因子得到时域平滑参数矩阵;

在一种可选的实施方式中,组合特征提取模块 820 包括:

5 输入信号获取单元,用于合并所述原始语音信号和所述原始语音信号的时域平滑特征,得到待增强语音信号;

权重矩阵训练单元,用于以所述待增强语音信号为深度神经网络的输入,利用反向传播算法对所述时域卷积核的权重矩阵进行训练;

增强语音信号获取单元,用于根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取,得到增强语音信号。

10 在一种可选的实施方式中,权重矩阵训练单元包括:

数据输入子单元,用于将所述待增强语音信号输入深度神经网络中,并构建时域损失函数;

数据训练子单元,用于根据所述时域损失函数,利用误差反向传播算法对所述时域卷积核的权重矩阵进行训练。

15 在一种可选的实施方式中,增强语音信号获取单元包括:

第一特征图获取子单元,用于将训练得到的权重矩阵与所述待增强语音信号中的原始语音信号作卷积运算,得到第一时域特征图;

第二特征图获取子单元,用于将训练得到的权重矩阵与所述待增强语音信号中的平滑特征作卷积运算,得到第二时域特征图;

20 特征组合子单元,用于组合所述第一时域特征图和所述第二时域特征图,得到所述增强语音信号。

上述端到端语音增强装置中各模块的具体细节已经在对应的语音增强方法中进行了详细的描述,因此此处不再赘述。

25 应当注意,尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元,但是这种划分并非强制性的。实际上,根据本公开的实施方式,上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之,上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

应当理解的是,本公开并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本公开的范围仅由所附的权利要求来限制。

权利要求

1. 一种基于神经网络的端到端语音增强方法，包括：
利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特
5 征；
对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。
2. 根据权利要求 1 所述的端到端语音增强方法，所述利用时域卷积核对原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征，包括：
10 根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵；
对所述时域平滑参数矩阵作乘积运算得到所述时域卷积核的权重矩阵；
将所述时域卷积核的权重矩阵和所述原始语音信号作卷积运算，得到所述原始语音信号的时域平滑特征。
3. 根据权利要求 2 所述的端到端语音增强方法，所述根据卷积滑窗和时域平滑因子
15 确定时域平滑参数矩阵，包括：
初始化多个时域平滑因子；
基于预设的卷积滑窗和所述多个时域平滑因子得到时域平滑参数矩阵。
4. 根据权利要求 1 所述的端到端语音增强方法，所述对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号，包括：
20 合并所述原始语音信号和所述原始语音信号的时域平滑特征，得到待增强语音信号；
以所述待增强语音信号为深度神经网络的输入，利用反向传播算法对所述时域卷积核的权重矩阵进行训练；
根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取，得到增强语音信号。
5. 根据权利要求 4 所述的端到端语音增强方法，所述以所述待增强语音信号为深度
25 神经网络的输入，利用反向传播算法对所述时域卷积核的权重矩阵进行训练，包括：
将所述待增强语音信号输入深度神经网络中，并构建时域损失函数；
根据所述时域损失函数，利用误差反向传播算法对所述时域卷积核的权重矩阵进行训练。
6. 根据权利要求 4 所述的端到端语音增强方法，所述根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取，得到增强语音信号，包括：
30 将训练得到的权重矩阵与所述待增强语音信号中的原始语音信号作卷积运算，得到第一时域特征图；
将训练得到的权重矩阵与所述待增强语音信号中的平滑特征作卷积运算，得到第二时域特征图；
35 域特征图；

组合所述第一时域特征图和所述第二时域特征图，得到所述增强语音信号。

7. 一种基于神经网络的端到端语音增强装置，包括：

时域平滑特征提取模块，用于利用时域卷积核处理后的原始语音信号进行特征提取，得到所述原始语音信号的时域平滑特征；

5 组合特征提取模块，对所述原始语音信号和所述原始语音信号的时域平滑特征进行组合特征提取，得到增强语音信号。

8. 根据权利要求7所述的端到端语音增强装置，所述时域平滑特征提取模块，包括：

参数矩阵确定单元，用于根据卷积滑窗和时域平滑因子确定时域平滑参数矩阵；

10 权重矩阵确定单元，用于对所述时域平滑参数矩阵作乘积运算得到所述时域卷积核的权重矩阵；

时域运算单元，用于将所述时域卷积核的权重矩阵和所述原始语音信号作卷积运算，得到所述原始语音信号的时域平滑特征。

9. 根据权利要求8所述的端到端语音增强装置，所述参数矩阵确定单元，包括：

数据初始化子单元，用于初始化多个时域平滑因子；

15 矩阵确定子单元，用于基于预设的卷积滑窗和所述多个时域平滑因子得到时域平滑参数矩阵。

10. 根据权利要求7所述的端到端语音增强装置，所述组合特征提取模块，包括：

输入信号获取单元，用于合并所述原始语音信号和所述原始语音信号的时域平滑特征，得到待增强语音信号；

20 权重矩阵训练单元，用于以所述待增强语音信号为深度神经网络的输入，利用反向传播算法对所述时域卷积核的权重矩阵进行训练；

增强语音信号获取单元，用于根据训练得到的权重矩阵对所述待增强语音信号进行组合特征提取，得到增强语音信号。

11. 根据权利要求10所述的端到端语音增强装置，所述权重矩阵训练单元，包括：

25 数据输入子单元，用于将所述待增强语音信号输入深度神经网络中，并构建时域损失函数；

数据训练子单元，用于根据所述时域损失函数，利用误差反向传播算法对所述时域卷积核的权重矩阵进行训练。

30 12. 根据权利要求10所述的端到端语音增强装置，所述增强语音信号获取单元，包括：

第一特征图获取子单元，用于将训练得到的权重矩阵与所述待增强语音信号中的原始语音信号作卷积运算，得到第一时域特征图；

第二特征图获取子单元，用于将训练得到的权重矩阵与所述待增强语音信号中的平滑特征作卷积运算，得到第二时域特征图；

35 特征组合子单元，用于组合所述第一时域特征图和所述第二时域特征图，得到所述增

强语音信号。

13. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现权利要求 1-6 任一项所述方法。

14. 一种电子设备，包括：

5 处理器；以及

存储器，用于存储所述处理器的可执行指令；

其中，所述处理器配置为经由执行所述可执行指令来执行权利要求 1-6 任一项所述的方法。