

基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品

相关申请的交叉引用

5 本申请实施例基于申请号为 202011410814.4、申请日为 2020 年 12 月 03 日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请实施例作为参考。

技术领域

本申请涉及云技术和人工智能技术，尤其涉及一种基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品。

10 背景技术

人工智能（AI，Artificial Intelligence）是计算机科学的一个综合技术，通过研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科，涉及领域广泛，例如自然语言处理技术以及机器学习/深度学习等几大方向，随着技术的发展，人工智能技术将在更多的领域得到应用，并发挥越来越重要的价值。例如，在基于云技术的网络会议场景中，引入人工智能技术以提升音频质量。

15 然而，相关技术中对于音频的处理方式比较单一，这虽然会对音频中的噪声形成抑制效果，但不可避免地也降低音频中的有用信号（例如语音信号）的质量。

发明内容

20 本申请实施例提供一种基于人工智能的音频处理方法、装置、电子设备、计算机可读存储介质及计算机程序产品，能够基于音频场景的针对性的音频处理，提高音频质量。

本申请实施例的技术方案是这样实现的：

本申请实施例提供一种基于人工智能的音频处理方法，包括：

获取音频场景的音频片段，其中，所述音频片段中包括噪声；

25 基于所述音频片段执行音频场景分类处理，得到与所述音频片段中的噪声对应的音频场景类型；

确定与所述音频场景类型对应的目标音频处理模式，并根据所述音频片段中的噪声所带来的干扰度，对所述音频片段应用所述目标音频处理模式。

本申请实施例提供一种音频处理装置，包括：

30 获取模块，配置为获取音频场景的音频片段，其中，所述音频片段中包括噪声；

分类模块，配置为基于所述音频片段执行音频场景分类处理，得到与所述音频片段中的噪声对应的音频场景类型；

处理模块，配置为确定与所述音频场景类型对应的目标音频处理模式，并根据所述音频片段中的噪声所带来的干扰度，对所述音频片段应用所述目标音频处理模式。

本申请实施例提供一种用于音频处理的电子设备，所述电子设备包括：

35 存储器，用于存储可执行指令；

处理器，用于执行所述存储器中存储的可执行指令时，实现本申请实施例提供的基于人工智能的音频处理方法。

本申请实施例提供一种计算机可读存储介质，存储有可执行指令，用于引起处理器执行时，实现本申请实施例提供的基于人工智能的音频处理方法。

本申请实施例提供一种计算机程序产品，包括计算机程序或指令，所述计算机程序或指令使得计算机执行本申请实施例提供的基于人工智能的音频处理方法。

5 本申请实施例具有以下有益效果：

通过噪声与音频场景类型之间的关联，来识别音频对应的音频场景类型，从而基于音频场景类型进行针对性的音频处理，使得在音频场景中引入的音频处理模式能够与音频场景中所包括的噪声相适应，从而能够最大程度保留音频中的有用信息，提高音频处理的准确性。

10 附图说明

图 1 是本申请实施例提供的音频处理系统的应用场景示意图；

图 2 是本申请实施例提供的用于音频处理的电子设备的结构示意图；

图 3-图 5 是本申请实施例提供的基于人工智能的音频处理方法的流程示意图；

图 6 是本申请实施例提供的神经网络模型的结构示意图；

15 图 7 是本申请实施例提供的音频场景识别的整体流程示意图；

图 8 是本申请实施例提供的从时域的声音信号中提取频谱特征的流程示意图；

图 9 是本申请实施例提供的神经网络模型的结构示意图；

图 10 是本申请实施例提供的 ResNet 单元的结构示意图。

具体实施方式

20 对本申请实施例进行进一步详细说明之前，对本申请实施例中涉及的名词和术语进行说明，本申请实施例中涉及的名词和术语适用于如下的解释。

1) 卷积神经网络 (CNN, Convolutional Neural Networks): 一类包含卷积计算且具有深度结构的前馈神经网络 (FNN, Feedforward Neural Networks), 是深度学习 (deep learning) 的代表算法之一。卷积神经网络具有表征学习 (representation learning) 能力，
25 能够按其阶层结构对输入图像进行平移不变分类 (shift-invariant classification)。

2) 残差网络 (ResNet, Residual Network): 一种容易优化，并且能够通过增加相当的深度来提高准确率的卷积神经网络。其内部的残差块使用了跳跃连接，缓解了深度神经网络中增加深度带来的梯度消失问题。

3) 音频处理模式: 一种用于进行音频处理的模式，将音频处理模式应用于音频片段中，可以优化音频，以得到清晰、流畅的音频。本申请实施例中的音频处理模式包括
30 降噪处理模式以及码率切换处理模式。

相关技术中，视频码率自适应技术 (ABR, Adaptive Bitrate Streaming) 可以自适应调节视频码率，码率调节算法多应用于视频播放，根据网络状况或客户端播放缓冲 (buffer) 情况自动调整视频码率 (即清晰度)；基于普适噪声的降噪算法，以带噪语音的
35 频谱特征作为神经网络输入，干净语音 (clean speech) 作为神经网络的参照输出训练降噪模型，使用最小均方误差 (LMS, Least Mean Square) 作为优化的目标，开启降噪功能后，对各种场景环境都使用同一种降噪方法。

申请人在实施本申请的过程中发现，由于网络环境质量波动变化较为频繁，基于网
40 速的码率切换也会相应频繁变动，频繁切换清晰度极大影响用户体验。在实际环境下特定场景的特定噪声对基于普适噪声的降噪算法的鲁棒性提出更高要求与挑战。

为了解决上述问题,本申请实施例提供了一种基于人工智能的音频处理方法、装置、电子设备及计算机可读存储介质,能够基于音频场景的针对性的音频处理,提高音频质量。

本申请实施例所提供的基于人工智能的音频处理方法,可以由终端/服务器独自实现;也可以由终端和服务器协同实现,例如终端独自承担下文所述的基于人工智能的音频处理方法,或者,终端 A 向服务器发送针对音频的优化请求(包括音频片段),服务器根据接收的针对音频的优化请求执行基于人工智能的音频处理方法,响应于针对音频的优化请求,对音频场景的音频片段应用目标音频处理模式(包括降噪处理模式以及码率切换处理模式),并将处理后的音频片段发送至终端 B,从而使得终端 A 与终端 B 之间能够进行清晰的语音通话。

本申请实施例提供的用于音频处理的电子设备可以是各种类型的终端设备或服务器,其中,服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云计算服务的云服务器;终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等,但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。

以服务器为例,例如可以是部署在云端的服务器集群,向用户开放人工智能云服务(AIaaS, AI as a Service),AIaaS 平台会把几类常见的 AI 服务进行拆分,并在云端提供独立或者打包的服务,这种服务模式类似于一个 AI 主题商城,所有的用户都可以通过应用程序编程接口的方式来接入使用 AIaaS 平台提供的一种或者多种人工智能服务。

例如,其中的一种人工智能云服务可以为音频处理服务,即云端的服务器封装有本申请实施例提供的音频处理的程序。用户通过终端(运行有客户端,例如录音客户端、即时通信客户端)调用云服务中的音频处理服务,以使部署在云端的服务器调用封装的音频处理的程序,基于与音频场景类型匹配的目标音频处理模式,并对音频场景的音频片段应用目标音频处理模式。

作为一个应用示例,对于录音客户端,用户可以是某音频平台的签约主播,需要定期发布某有声书的音频,然而主播录音的场景可能是变化的,例如在家中、图书馆中甚至是在户外录音,这些场景各自存在不同的噪声。通过对当前音频场景进行录音,对记录的音频片段进行音频场景识别,以确定出音频场景类型,并基于音频场景类型对音频片段进行针对性的降噪处理,以存储去噪的音频片段,实现去噪录音功能。

作为另一个应用示例,对于即时通信客户端,用户可以将某指定好友发送语音,也可以向某指定群组发送语音,然而用户当前所处的场景可能是变化的,例如办公室、商场等,不同场景存在不同的噪声。通过对当前场景的语音进行语音场景识别,以确定出语音场景类型,并基于语音场景类型对语音进行针对性的降噪处理,以发送去噪后的音频片段,实现去噪语音发送功能。

作为另一个应用示例,对于会议客户端,参与会议的用户可以处于不同的环境进行语音通话,例如参会的用户 A 在办公室,参会的用户 B 在高铁上,不同场景存在不同的噪声,且通信信号不同,例如高铁存在车声且通信信号比较差。通过对各参会用户的语音通话进行语音场景识别,以确定出各参会用户的语音场景类型,并基于语音场景类型对各参会用户的语音进行针对性的码率切换处理,以实现自适应码率切换,提高会议通话的通话质量。

参见图 1,图 1 是本申请实施例提供的音频处理系统 10 的应用场景示意图,终端 200 通过网络 300 连接服务器 100,网络 300 可以是广域网或者局域网,又或者是二者的组合。

终端 200(运行有客户端,例如录音客户端、即时通信客户端、通话客户端等)可

以被用来获取针对音频的优化请求，例如，用户通过终端 200 输入或录入音频场景的音频片段，则终端 200 自动获取音频场景的音频片段，并自动生成针对音频的优化请求。

5 在一些实施例中，终端中运行的客户端中可以植入有音频处理插件，用以在客户端本地实现基于人工智能的音频处理方法。例如，终端 200 获取针对音频的优化请求（包括音频场景的音频片段）后，调用音频处理插件，以实现基于人工智能的音频处理方法，识别与音频片段中的噪声对应的音频场景类型，并基于与音频场景类型匹配的目标音频处理模式，并对音频场景的音频片段应用目标音频处理模式，例如，对于录音应用，用户

10 在当前音频场景进行录音，对记录的音频片段进行音频场景识别，以确定出音频场景类型，并基于音频场景类型对音频片段进行针对性的降噪处理，以存储去噪后的音频片段，实现去噪录音功能。

在一些实施例中，终端 200 获取针对音频的优化请求后，调用服务器 100 的音频处理接口（可以提供为云服务的形式，即音频处理服务），服务器 100 识别与音频片段中的噪声对应的音频场景类型，并基于与音频场景类型匹配的目标音频处理模式，对音频场景的音频片段应用目标音频处理模式，并将通过目标音频处理模式的音频片段（优化

15 后的音频片段）发送至终端 200 或者其他终端，例如，对于录音应用，用户在当前音频场景进行录音，终端 200 获取对应的音频片段，并自动生成针对音频的优化请求，并将针对音频的优化请求发送至服务器 100，服务器 100 基于针对音频的优化请求，对记录的音频片段进行音频场景识别，以确定出音频场景类型，并基于音频场景类型对音频片段进行针对性的降噪处理，以存储去噪后的音频片段，实现去噪录音功能；对于即时通

20 信应用，用户在当前语音场景进行语音发送，终端 200 获取对应的音频片段，并自动生成针对音频的优化请求，并将针对音频的优化请求发送至服务器 100，服务器 100 基于针对音频的优化请求，对音频片段进行语音场景识别，以确定出语音场景类型，并基于语音场景类型对音频片段进行针对性的降噪处理，以发送去噪后的音频片段，实现去噪语音发送功能，并基于语音场景类型对音频片段进行针对性的码率切换处理，以实现自

25 适应码率切换，提高语音通话质量；对于通话应用，用户 A 与用户 B 进行语音通话，用户 A 在当前语音场景进行语音通话，终端 200 获取用户 A 对应的音频片段，并自动生成针对音频的优化请求，并将针对音频的优化请求发送至服务器 100，服务器 100 基于针对音频的优化请求，对用户 A 的音频片段进行语音场景识别，以确定出语音场景类型，并基于语音场景类型对音频片段进行针对性的降噪处理，并向用户 B 发送去噪的用户

30 用户 A 的音频片段，实现去噪语音通话功能。

下面说明本申请实施例提供的用于音频处理的电子设备的结构，参见图 2，图 2 是本申请实施例提供的用于音频处理的电子设备 500 的结构示意图，以电子设备 500 是服务器为例说明，图 2 所示的用于音频处理的电子设备 500 包括：至少一个处理器 510、

35 存储器 550、至少一个网络接口 520 和用户接口 530。电子设备 500 中的各个组件通过总线系统 540 耦合在一起。可理解，总线系统 540 用于实现这些组件之间的连接通信。总线系统 540 除包括数据总线之外，还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见，在图 2 中将各种总线都标为总线系统 540。

处理器 510 可以是一种集成电路芯片，具有信号的处理能力，例如通用处理器、数字信号处理器（DSP，Digital Signal Processor），或者其他可编程逻辑器件、分立门或者

40 晶体管逻辑器件、分立硬件组件等，其中，通用处理器可以是微处理器或者任何常规的处理器等。

存储器 550 包括易失性存储器或非易失性存储器，也可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（ROM，Read Only Memory），易失性存储器可以是随机存取存储器（RAM，Random Access Memory）。本申请实施例描述

45

的存储器 550 旨在包括任意适合类型的存储器。存储器 550 可选地包括在物理位置上远离处理器 510 的一个或多个存储设备。

在一些实施例中，存储器 550 能够存储数据以支持各种操作，这些数据的示例包括程序、模块和数据结构或者其子集或超集，下面示例性说明。

- 5 操作系统 551，包括用于处理各种基本系统服务和执行硬件相关任务的系统程序，例如框架层、核心库层、驱动层等，用于实现各种基础业务以及处理基于硬件的任务；
- 网络通信模块 552，用于经由一个或多个（有线或无线）网络接口 520 到达其他计算设备，示例性的网络接口 520 包括：蓝牙、无线相容性认证（WiFi）、和通用串行总线（USB，Universal Serial Bus）等；

- 10 在一些实施例中，本申请实施例提供的音频处理装置可以采用软件方式实现，例如，可以是上文所述的终端中的音频处理插件，可以是上文所述的服务器中音频处理服务。当然，不局限于此，本申请实施例提供的音频处理装置可以提供为各种软件实施例，包括应用程序、软件、软件模块、脚本或代码在内的各种形式。

- 15 图 2 示出了存储在存储器 550 中的音频处理装置 555，其可以是程序和插件等形式的软件，例如音频处理插件，并包括一系列的模块，包括获取模块 5551、分类模块 5552、处理模块 5553 以及训练模块 5554；其中，获取模块 5551、分类模块 5552、处理模块 5553 用于实现本申请实施例提供的音频处理功能，训练模块 5554 用于训练神经网络模型，其中，音频场景分类处理是通过神经网络模型实现的。

- 20 如前所述，本申请实施例提供的基于人工智能的音频处理方法可以由各种类型的电子设备实施。参见图 3，图 3 是本申请实施例提供的基于人工智能的音频处理方法的流程图示意图，结合图 3 示出的步骤进行说明。

在下面步骤中，音频场景是指产生有音频的环境，例如居家环境、办公环境、乘坐高铁等各类交通工具时的外在环境等。

- 25 在步骤 101 中，获取音频场景的音频片段，其中，音频片段中包括噪声。

作为获取音频片段的示例，用户在当前音频场景通过终端（运行有客户端）输入音频，终端 200 获取对应的音频片段，并自动生成针对音频的优化请求，并将针对音频的优化请求发送至服务器，服务器解析针对音频的优化请求，以获取音频场景的音频片段，以便后续基于音频片段进行音频场景识别。

- 30 在步骤 102 中，基于音频片段执行音频场景分类处理，得到与音频片段中的噪声对应的音频场景类型。

- 例如，获取到音频场景的音频片段后，可以通过神经网络模型基于音频片段进行音频场景分类处理，以得到与音频片段中的噪声对应的音频场景类型。其中，可以将音频片段输入至神经网络模型，还可以将音频片段的时域特征或者频域特征输入至神经网络模型，神经网络模型基于音频片段的时域特征或者频域特征进行音频场景分类处理，以得到与音频片段中的噪声对应的音频场景类型。以音频片段的频域特征为例，获取到音频片段后，先对音频片段的时域信号进行分帧处理，得到多帧音频信号，然后对多帧音频信号进行加窗处理，并对加窗处理后的音频信号进行傅里叶变换，得到音频片段的频域信号，对频域信号的梅尔频带进行对数处理，得到音频片段的频域特征，即用于进行音频场景分类的音频片段。
- 40

为了使得神经网络模型能够对多通道输入进行处理，可以对对数处理得到的音频片段的频域特征进行求导处理，得到音频片段的一阶导数，然后对一阶导数进行求导处理，得到音频片段的二阶导数，最后将音频片段的频域特征、一阶导数以及二阶导数组合成三通道输入信号，并将三通道输入信号作为用于进行音频场景分类的音频片段。

- 45 在一些实施例中，音频场景分类处理是通过神经网络模型实现的，神经网络模型学

习到音频片段中所包括的噪声与音频场景类型的关联关系；基于音频片段执行音频场景分类处理，得到与音频片段中的噪声对应的音频场景类型，包括：基于音频片段调用神经网络模型以执行音频场景分类处理，得到与音频片段中所包括的噪声存在关联关系的音频场景类型。

5 例如，如图6所示，神经网络模型包括映射网络、残差网络以及池化网络；通过映射网络对音频片段进行特征提取处理，得到音频片段中噪声的第一特征向量；通过残差网络对第一特征向量进行映射处理，得到音频片段的映射向量；通过映射网络对音频片段的映射向量进行特征提取处理，得到音频片段中噪声的第二特征向量；通过池化网络对第二特征向量进行池化处理，得到音频片段的池化向量；对音频片段的池化向量进行非线性映射处理，得到与音频片段中所包括的噪声存在关联关系的音频场景类型。

10 承接上述示例，映射网络包括多个级联的映射层；通过映射网络对音频片段进行特征提取处理，得到音频片段中噪声的第一特征向量，包括：通过多个级联的映射层中的第一个映射层，对音频片段进行特征映射处理；将第一个映射层的映射结果输出到后续级联的映射层，通过后续级联的映射层继续进行特征映射和映射结果输出，直至输出到最后一个映射层，并将最后一个映射层输出的映射结果作为音频片段中噪声的第一特征向量。

其中，映射网络可以有效提取音频片段中的场景噪声特征，映射层可以是卷积神经网络，但本申请实施例并不局限于卷积神经网络，还可以是其他神经网络。

20 在一些实施例中，残差网络包括第一映射网络以及第二映射网络；通过残差网络对第一特征向量进行映射处理，得到音频片段的映射向量，包括：通过第一映射网络对第一特征向量进行映射处理，得到音频片段的第一映射向量；对第一映射向量进行非线性映射处理，得到音频片段的非映射向量；通过第一映射网络对音频片段的非映射向量进行映射处理，得到音频片段的第二映射向量；将音频片段的第一特征向量与音频片段的第二映射向量的加和结果作为音频片段的映射向量。

25 其中，残差网络可以有效预防神经网络训练误差传递中的梯度消失问题，以加快神经网络模型的训练。

30 在一些实施例中，需要训练神经网络模型，以使训练后的神经网络模型能够进行音频场景分类，其训练方法如下：基于无噪声的音频信号以及多个不同音频场景分别对应的背景噪声，构建多个不同音频场景分别对应的音频样本；基于多个不同音频场景分别对应的音频样本对神经网络模型进行训练，得到用于音频场景分类的神经网络模型。

为了增强样本数据的多样性，其构建音频样本的方法如下：针对多个不同音频场景中的任一音频场景执行以下处理：基于音频场景的背景噪声与无噪声的音频信号的融合比例，对音频场景的背景噪声以及无噪声的音频信号进行融合，得到音频场景的第一融合音频信号；在第一融合音频信号中融合对应第一随机系数的音频场景的背景噪声，得到音频场景的第二融合音频信号；在第二融合音频信号中融合对应第二随机系数的无噪声的音频信号，得到音频场景的音频样本。

例如，在保留人声（无噪声的音频信号）与背景噪声按融合比例1:1进行融合后，生成部分随机比例以叠加数据，例如噪声叠加系数（第一随机系数）为0.3~0.5中的随机数，人声叠加系数（第二随机系数）为0.5~0.7中的随机数。

40 在一些实施例中，基于多个音频场景分别对应的音频样本对神经网络模型进行训练，得到用于音频场景分类的神经网络模型，包括：通过神经网络模型对多个不同音频场景分别对应的音频样本进行音频场景分类处理，得到音频样本的预测音频场景类型；基于音频样本的预测音频场景类型、音频样本的音频场景标注以及音频样本的权重，构建神经网络模型的损失函数；更新神经网络模型的参数直至损失函数收敛，将损失函数收敛

时神经网络模型的更新的参数，作为用于音频场景分类的神经网络模型的参数。

例如，基于音频样本的预测音频场景类型、音频样本的音频场景标注以及音频样本的权重，确定神经网络模型的损失函数的值后，可以判断神经网络模型的损失函数的值是否超出预设阈值，当神经网络模型的损失函数的值超出预设阈值时，基于神经网络模型的损失函数确定神经网络模型的误差信号，将误差信息在神经网络模型中反向传播，并在传播的过程中更新各个层的模型参数。

这里，对反向传播进行说明，将训练样本数据输入到神经网络模型的输入层，经过隐藏层，最后达到输出层并输出结果，这是神经网络模型的前向传播过程，由于神经网络模型的输出结果与实际结果有误差，则计算输出结果与实际值之间的误差，并将该误差从输出层向隐藏层反向传播，直至传播到输入层，在反向传播的过程中，根据误差调整模型参数的值；不断迭代上述过程，直至收敛。

在步骤 103 中，确定与音频场景类型对应的目标音频处理模式，并根据音频片段中的噪声所带来的干扰度，对音频片段应用目标音频处理模式。

例如，在获得音频场景类型后，首先确定与音频场景类型匹配的目标音频处理模式，然后将目标音频处理模式应用于音频场景的音频片段，以进行针对性的音频优化，提高音频处理的准确性。

参见图 4，图 4 是本申请实施例提供的基于人工智能的音频处理方法的一个可选的流程示意图，图 4 示出图 3 中的步骤 103 可以通过图 4 示出的步骤 1031A 至步骤 1032A 实现：目标音频处理模式包括降噪处理模式；在步骤 1031A 中，基于音频场景对应的音频场景类型，查询不同候选音频场景类型与候选降噪处理模式的对应关系，得到与音频场景类型匹配的降噪处理模式；在步骤 1032A 中，将音频场景类型匹配的噪声类型，与音频片段中的噪声进行匹配处理；对与噪声类型匹配成功的噪声进行抑制处理，得到抑制后的音频片段；其中，抑制后的音频片段中的语音信号强度与噪声信号强度的比值低于信噪比阈值。

例如，根据实际的应用场景，预先构建包括不同候选音频场景类型与候选降噪处理模式的对应关系的映射表，并将该映射表存储至存储空间，通过读取存储空间中映射表所包括的不同候选音频场景类型与候选降噪处理模式的对应关系，可以基于音频场景对应的音频场景类型快速查询到与音频场景类型匹配的降噪处理模式，从而将降噪处理模式应用于音频场景的音频片段，以去除音频场景的音频片段的噪声，从而实现针对性地去噪，提高音频片段的音频质量（即音频的清晰度）。

在一些实施例中，目标音频处理模式包括降噪处理模式；确定与音频场景类型对应的目标音频处理模式，包括：基于音频场景对应的音频场景类型，确定与音频场景类型匹配的噪声类型；基于与音频场景类型匹配的噪声类型，查询不同候选噪声类型与候选降噪处理模式的对应关系，得到与音频场景类型对应的降噪处理模式；其中，与不同的音频场景类型匹配的噪声类型不完全相同。

例如，先通过音频场景对应的音频场景类型，确定与音频场景类型匹配的噪声类型，然后通过与音频场景类型匹配的噪声类型，获取与音频场景类型匹配的降噪处理模式，即实现音频场景类型与候选降噪处理模式的解耦，以便后续可以灵活地调整音频场景类型与候选降噪处理模式的对应关系。

例如，由于客户端的开发者针对不同噪声分配降噪处理模式的策略可能是变化的，或者，不同用户针对不同噪声的降噪处理模式的需求是变化的，因此，如果通过神经网络模型实现音频片段的音频场景类型与降噪处理模式的映射关系，需要针对性训练大量模型，且一旦针对不同噪声分配降噪处理模式发生变化，则需要重新训练神经网络模型，将消耗大量的计算资源。

然而，如果仅通过神经网络模型实现音频片段的音频场景类型与噪声的映射关系，则训练一个神经网络模型即可满足实际应用中针对降噪处理模式的各种需求，只需要在客户端中实现噪声类型与降噪处理模式的策略设定即可，即使针对不同噪声分配降噪处理模式发生变化，仅在客户端中调整噪声类型与降噪处理模式的策略设定即可，从而避免消耗大量的计算资源训练神经网络模型。

在一些实施例中，根据音频片段中的噪声所带来的干扰度，对音频片段应用目标音频处理模式，包括：确定音频片段中的噪声所带来的干扰度；当干扰度大于干扰度阈值时，对音频场景的音频片段应用与音频场景类型对应的降噪处理模式。

例如，当音频片段中的噪声对音频片段的影响不大时，可以不进行降噪处理，只有当音频片段中的噪声影响到音频片段时，才对音频片段进行降噪处理，例如，在用户进行录音时，虽然录音时会收录音频场景的一些噪声，但是这些噪声并不会影响录音的效果，则可以不对录音进行降噪处理；当这些噪声影响到录音的效果（例如听不清录音的内容）时，则可以对录音进行降噪处理。

参见图 5，图 5 是本申请实施例提供的基于人工智能的音频处理方法的一个可选的流程示意图，图 5 示出图 3 中的步骤 103 可以通过图 4 示出的步骤 1031B-1032B 实现：目标音频处理模式包括码率切换处理模式；在步骤 1031B 中，基于与音频场景对应的音频场景类型，查询不同候选音频场景类型与候选码率切换处理模式的对应关系，得到与音频场景类型对应的码率切换处理模式；在步骤 1032B 中，对音频片段应用与音频场景类型对应的码率切换处理模式。

例如，根据实际的应用场景，预先构建包括不同音频场景类型与候选码率切换处理模式的对应关系的映射表，并将该映射表存储至存储空间，通过读取存储空间中映射表所包括的不同音频场景类型与候选码率切换处理模式的对应关系，可以基于音频场景对应的音频场景类型快速查询到与音频场景类型匹配的码率切换处理模式，从而将码率切换处理模式应用于音频场景的音频片段，以切换音频片段的码率，从而实现针对性地码率切换，提高音频片段的流畅性。

在一些实施例中，目标音频处理模式包括码率切换处理模式；确定与音频场景类型对应的目标音频处理模式，包括：将音频场景类型与预设音频场景类型比对；当比对确定音频场景类型是预设音频场景类型时，将与预设音频场景类型关联的码率切换处理模式，作为与音频场景类型对应的码率切换处理模式。

例如，并不是所有的音频场景都需要进行码率切换，例如办公环境的通信信号比较稳定，不需要进行码率切换，而高铁环境信号较弱且不稳定，需要进行码率切换。因此，在确定码率切换处理模式之前，需要将音频场景类型与需要进行码率切换的预设音频场景类型比对，当比对确定音频场景类型属于需要进行码率切换的预设音频场景类型时，才将与预设音频场景类型关联的码率切换处理模式，确定为音频场景类型匹配的码率切换处理模式，以避免所有场景都进行码率切换所造成的资源浪费。

在一些实施例中，对音频场景的音频片段应用目标音频处理模式，包括：当音频场景的通信信号强度小于通信信号强度阈值时，按照第一设定比例或第一设定值降低音频片段的音频码率；当音频场景的通信信号强度大于或者等于通信信号强度阈值时，按照第二设定比例或第二设定值提升音频片段的音频码率。

以语音通话场景为例，多人处于不同的环境进行语音通话，并通过客户端向服务器发送音频片段，服务器接收到各客户端发送的音频片段，基于音频片段执行音频场景分类处理，以得到与音频片段中的噪声对应的音频场景类型，并确定与音频场景类型匹配的码率切换处理模式后，确定音频场景的通信信号强度，当音频场景的通信信号强度小于通信信号强度阈值时，则说明当前音频场景的信号弱，需要降低码率，因此按照与音

频场景类型匹配的码率切换处理模式中的第一设定比例或第一设定值,降低音频片段的音频码率,以便后续进行流畅音频交互,避免语音通话中断;当音频场景的通信信号强度大于或者等于通信信号强度阈值时,则说明当前音频场景的通信信号强,即使不调低码率也不会使得通话中断,不需要降低码率,因此按照与音频场景类型匹配的码率切换处理模式中的第二设定比例或第二设定值,提升音频片段的音频码率,提高音频交互的流畅性。需要说明的是,第一设定比例与第二设定比例可以相同,也可以不同;第一设定值比例与第二设定值可以相同,也可以不同。第一设定比例、第二设定比例、第一设定值、第二设定值可以是根据实际需求设置的。

其中,获取音频场景的通信信号强度的方式示例如下:对在音频场景中多次采样得到的通信信号强度进行平均,将平均结果作为音频场景的通信信号强度。例如,将某用户的语音通话开始时至今的多次采样结果平均,将平均结果作为该用户在所处音频场景的通信信号强度。

在一些实施例,对音频场景的音频片段应用目标音频处理模式,包括:基于在音频场景中多次采样得到的通信信号强度,确定音频场景中的通信信号强度的抖动信息;当抖动信息表征通信信号呈现不稳定的状态时,按照第三设定比例或第三设定值降低音频片段的音频码率。

例如,在音频场景中多次采样得到的通信信号强度,并通过正态分布的方式获取音频场景中的通信信号强度的抖动变化情况(即抖动信息),当表征抖动变化情况的正态分布中的方差大于方差阈值时,说明正态分布中的数据(即通信信号)比较分散,则通信信号强度抖动剧烈,则说明通信信号不稳定,为了在保证音频流畅性的基础上,避免后续来回切换音频码率,可以按照与音频场景类型匹配的码率切换处理模式中的预设比例或预设值,降低音频片段的音频码率。其中,第三设定比例、第三设定值可以是根据实际需求设置的。

通过判断音频场景中的通信信号强度的抖动变化情况,进一步确定是否需要切换码率,从而在保证音频流畅性的基础上,避免频繁切换音频码率,提高用户的体验感。

在一些实施例,对音频场景的音频片段应用目标音频处理模式,包括:当用于传输音频片段的通信网络的类型属于设定类型时,按照第四设定比例或第四设定值降低音频片段的音频码率。

例如,在确定与音频场景类型匹配的码率切换处理模式后,还可以先确定用于传输音频片段的通信网络的类型是否属于设定类型(例如WiFi网络、蜂窝网络等),例如当确定用于传输音频片段的通信网络的类型属于wifi网络,则说明当前音频片段处于不稳定的环境,为了在保证音频流畅性,可以按照与音频场景类型匹配的码率切换处理模式中的预设比例或预设值,降低音频片段的音频码率。其中,第三设定比例、第三设定值可以是根据实际需求设置的。

下面,将说明本申请实施例在一个实际的应用场景中的示例性应用。

本申请实施例可以应用于各种语音的应用场景中,例如,对于录音应用,用户通过终端中所运行的录音客户端在当前语音场景进行录音,录音客户端对记录的音频片段进行语音场景识别,以确定出语音场景类型,并基于语音场景类型对音频片段进行针对性的降噪处理,以存储去噪的音频片段,实现去噪录音功能;对于即时通信应用,用户通过终端中所运行的即时通信客户端在当前语音场景进行发语音,该即时通信客户端获取对应的音频片段,对音频片段进行语音场景识别,以确定出语音场景类型,并基于语音场景类型对音频片段进行针对性的降噪处理,通过即时通信客户端发送去噪的音频片段,实现去噪语音发送功能;对于通话应用,用户A与用户B进行语音通话,用户A通过终端中所运行的通话客户端在当前语音场景进行语音通话,通话客户端获取用户A的音

频片段,并基于用户A的音频片段自动生成针对音频的优化请求,并将针对音频的优化请求发送至服务器,服务器基于接收到的针对音频的优化请求对用户A的音频片段进行语音场景识别,以确定出语音场景类型,并基于语音场景类型对音频片段进行针对性的降噪处理,并向用户B发送去噪的用户A的音频片段,实现去噪语音通话功能。

5 本申请实施例提供一种基于人工智能的音频处理方法,针对音频片段提取梅尔频率对数能量特征,将归一化后的特征输入到神经网络当中,得到音频片段对应的场景预测。由于通话场景相较而言更为稳定,基于场景的码率控制因而具有更好的稳定性。针对不同场景的噪声特性,可以使用适应学习、迁移学习等方式获得针对特定场景的个性化降噪方案,基于场景识别的结果相应切换特定场景的专用降噪模式将获得更好的降噪表现,提升通话质量与用户体验。

10 例如,在实时通信会议中,随着会议移动终端的不断完善,用户可能于各种不同环境(音频场景、语音场景)中入会,例如,办公室环境,居家环境,地铁、高铁等移动交通工具环境等。不同场景给实时音频信号处理带来了具有场景特性的挑战。例如,高铁等场景信号较弱且不稳定,音频通信时常出现卡顿,严重影响通信质量;不同场景具有的特定背景噪声(例如居家环境中的孩童嬉闹、电视背景声、餐厨噪声等)对降噪算法的鲁棒性提出更高要求。

为了满足用户在各个场景下进行会议的需求,提高用户在复杂环境下的会议音频体验,基于环境特性提供场景个性化的解决方案是音频处理算法优化的重要趋势。准确识别音频发生的场景是实现场景个性化方案的重要依据与基础。本申请实施例提出一种音频场景分类(Audio Scene Classification)的方案,针对音频场景的分类结果启用场景个性化的音频处理算法,例如,针对信号较弱且不稳定的高铁场景进行自动码率(Bit rate)切换,降低音频码率避免卡顿;根据识别出的场景应用针对特定场景的降噪方案,提高用户入会体验。

25 其中,针对性降噪方案主要针对场景特性的噪声,例如办公环境的键盘声,纸片摩擦声;居家环境的餐厨噪声,儿童嬉闹,电视背景声;移动交通工具的报站声等场景特性的噪声。基于通用降噪模型通过自适应等方式训练针对各场景的降噪模型(对场景特性噪声重点进行消除的模型),在识别出场景后启用场景对应的降噪模型进行降噪。针对性码率切换主要针对特定场景,比如在高铁等信号比较弱的移动交通工具环境,降低会议通信的码率(例如从16k降低到8k),减轻传输负担,减少卡顿,提升参会体验。

30 本申请实施例提出的音频场景分类方案,首先根据采集到的时域音频信号,提取出对应的频域频谱特征,即梅尔频率对数能量谱(Mel Frequency Log Filterbank Energy),再对这些频谱特征进行归一化(Normalization)处理。归一化处理之后,将这些归一化之后的频谱特征输入到神经网络模型,例如基于卷积神经网络(CNN, Convolutional Neural Network)的深度残差网络(ResNet, Residue Network),通过神经网络模型对归一化后的频谱特征进行建模。实际测试时,首先对输入音频信号的对数能量谱归一化,输入到已建立的神经网络模型,神经网络模型对输入的对数能量谱输出场景分类结果。依据神经网络模型识别的场景结果,会议系统可以自动切换适应的音频码率并启用适合该场景的针对性降噪方案等,整体提高语音通话质量和用户体验。

40 如图7所示,图7是本申请实施例提供的音频场景识别的整体流程示意图,音频场景识别包含训练和测试两个阶段,其中包含5个模块,分别是:1)场景噪声语料收集;2)训练数据构建;3)训练数据特征提取;4)神经网络模型训练;5)场景预测。

1) 场景噪声语料收集

收集不同场景下的背景噪声,例如办公环境的键盘声,纸片摩擦声;居家环境的餐厨噪声,儿童嬉闹,电视背景声;移动交通工具的报站声等场景特性的噪声。

2) 训练数据构建

将收集到的不同场景下的背景噪声与不同的干净音频（无噪声的语音）在时域上相叠加，生成场景噪声与干净音频叠加的混合信号，作为神经网络模型训练的输入语料。叠加时为了防止叠加后语音幅值超过系统阈值，同时增强数据多样性，更好的模拟真实环境下的音频，在保证人声与噪声按原比例 1:1 叠加的数据的同时，生成部分随机比例的叠加数据，例如人声叠加系数为 0.5~0.7 中的随机数，噪声叠加系数为 0.3~0.5 中的随机数。

3) 训练数据特征提取

对训练数据中的音频信号进行分帧、加窗、傅立叶变换等操作得到梅尔对数能量频谱特征。

如图 8 所示，图 8 是本申请实施例提供的从时域的声音信号中提取频谱特征的流程图，首先依据短时瞬态假设对时域信号进行分帧操作，以将连续信号转为离散向量；此后对每帧音频信号进行加窗平滑，以消除边缘不连续性；随后对每帧进行傅里叶变换（FT, Fourier Transform）得到频域信号；再对频域信号应用梅尔频带获取每个频带内的能量，基于人耳对于音频的响应非线性，此处使用梅尔频带取代线性频带以更好地模拟人耳响应，最后取对数操作得到梅尔对数频谱特征。

4) 神经网络模型训练

神经网络模型的输入是场景噪声叠加干净音频的三通道梅尔对数能量频谱特征，神经网络模型的输出是场景识别的分类结果。训练过程中，采用交叉熵误差（Cross Entropy Loss）作为损失函数，并以最小化损失函数作为训练目标： $\min_{\theta} L(\theta) = -\sum_i t_i \log o_i$ 。其中， t_i 表示输入音频的正确场景标注， o_i 为神经网络模型预测的场景类别。

如图 9 所示，图 9 是本申请实施例提供的神经网络模型的结构示意图，神经网络模型由 2 个 ResNet 单元（残差网络），多个 CNN 网络与平均池化层（Pooling Layer）组成，梅尔对数能量特征与其一阶导数与二阶导数组成三通道输入信号，最终输出场景分类结果。

其中，神经网络模型使用 ResNet 单元。如图 10 所示，图 10 是本申请实施例提供的 ResNet 单元的结构示意图，每个 ResNet 单元包含两层 CNN，其中 x 和 y 分别为残差单元的输入和输出， f_1 、 f_2 分别表示两个 CNN 层的函数映射， W_1 、 W_2 分别表示两个 CNN 层对应的权重参数（Weights）。CNN 层可以有效捕捉频谱信息中的场景噪声特征，残差网络可以有效预防神经网络训练误差传递中的梯度消失问题。

5) 场景预测

在训练完神经网络模型之后，选取最优的模型参数保存为训练好的模型。测试时，将带噪语音归一化后提取频谱特征输入训练好的模型中，训练好的模型输出预测的音频场景。后续针对音频场景的分类结果启用场景个性化的音频处理算法，例如，针对信号较弱且不稳定的高铁场景进行自动码率（Bit rate）切换，降低音频码率避免卡顿；根据识别出的场景应用针对特定场景的降噪方案，提高用户入会体验。

综上，本申请实施例构建轻量音频场景识别模型，对存储空间要求低，预测速度快。作为前端算法可以作为后续复杂算法优化的依据和基础。根据音频场景识别结果调整控制音频码率、启用场景特定降噪方案等场景个性化音频解决方案。

至此已经结合本申请实施例提供的服务器的示例性应用和实施，说明本申请实施例提供的基于人工智能的音频处理方法。本申请实施例还提供音频处理装置，实际应用中，音频处理装置中的各功能模块可以由电子设备（如终端设备、服务器或服务器集群）的硬件资源，如处理器等计算资源、通信资源（如用于支持实现光缆、蜂窝等各种方式通信）、存储器协同实现。图 2 示出了存储在存储器 550 中的音频处理装置 555，其可以是

程序和插件等形式的软件,例如,软件 C/C++、Java 等编程语言设计的软件模块、C/C++、Java 等编程语言设计的应用软件或大型软件系统中的专用软件模块、应用程序接口、插件、云服务等实现方式,下面对不同的实现方式举例说明。

示例一、音频处理装置是移动端应用程序及模块

5 本申请实施例中的音频处理装置 555 可提供为使用软件 C/C++、Java 等编程语言设计的软件模块,嵌入到基于 Android 或 iOS 等系统的各种移动端应用中(以可执行指令存储在移动端的存储介质中,由移动端的处理器执行),从而直接使用移动端自身的计算资源完成相关的信息推荐任务,并且定期或不定期地通过各种网络通信方式将处理结果传送给远程的服务器,或者在移动端本地保存。

10 示例二、音频处理装置是服务器应用程序及平台

本申请实施例中的音频处理装置 555 可提供为使用 C/C++、Java 等编程语言设计的应用软件或大型软件系统中的专用软件模块,运行于服务器端(以可执行指令的方式在服务器端的存储介质中存储,并由服务器端的处理器运行),服务器使用自身的计算资源完成相关的信息推荐任务。

15 本申请实施例还可以提供为在多台服务器构成的分布式、并行计算平台上,搭载定制的、易于交互的网络(Web)界面或其他各用户界面(UI, User Interface),形成供个人、群体或单位使用的信息推荐平台(用于推荐列表)等。

示例三、音频处理装置是服务器端应用程序接口(API, Application Program Interface)及插件

20 本申请实施例中的音频处理装置 555 可提供为服务器端的 API 或插件,以供用户调用,以执行本申请实施例的基于人工智能的音频处理方法,并嵌入到各类应用程序中。

示例四、音频处理装置是移动设备客户端 API 及插件

本申请实施例中的音频处理装置 555 可提供为移动设备端的 API 或插件,以供用户调用,以执行本申请实施例的基于人工智能的音频处理方法。

25 示例五、音频处理装置是云端开放服务

本申请实施例中的音频处理装置 555 可提供为向用户开发的信息推荐云服务,供个人、群体或单位获取推荐列表。

30 其中,音频处理装置 555 包括一系列的模块,包括获取模块 5551、分类模块 5552、处理模块 5553 以及训练模块 5554。下面继续说明本申请实施例提供的音频处理装置 555 中各个模块配合实现音频处理方案。

35 获取模块 5551,配置为获取音频场景的音频片段,其中,所述音频片段中包括噪声;分类模块 5552,配置为基于所述音频片段执行音频场景分类处理,得到与所述音频片段中的噪声对应的音频场景类型;处理模块 5553,配置为确定与所述音频场景类型对应的目标音频处理模式,并根据音频片段中的噪声所带来的干扰度,对所述音频片段应用所述目标音频处理模式。

在一些实施例中,所述目标音频处理模式包括降噪处理模式;所述处理模块 5553 还配置为基于所述音频场景对应的音频场景类型,查询不同候选音频场景类型与候选降噪处理模式的对应关系,得到与所述音频场景类型对应的降噪处理模式。

40 在一些实施例中,所述目标音频处理模式包括降噪处理模式;所述处理模块 5553 还配置为基于所述音频场景对应的音频场景类型,确定与所述音频场景类型匹配的噪声类型;基于与所述音频场景类型匹配的噪声类型,查询不同候选噪声类型与候选降噪处理模式的对应关系,得到与所述音频场景类型对应的降噪处理模式;其中,与不同的所述音频场景类型匹配的噪声类型不完全相同。

在一些实施例中,在所述对所述音频场景的音频片段应用所述目标音频处理模式之

前,所述处理模块 5553 还配置为确定所述音频片段中的噪声所带来的干扰度;当所述干扰度大于干扰度阈值时,对所述音频片段应用与所述音频场景类型对应的降噪处理模式。

5 在一些实施例中,所述处理模块 5553 还配置为将所述音频场景类型匹配的噪声类型,与所述音频片段中的噪声进行匹配处理;对与所述噪声类型匹配成功的噪声进行抑制处理,得到抑制后的所述音频片段;其中,抑制后的所述音频片段中的语音信号强度与噪声信号强度的比值低于信噪比阈值。

10 在一些实施例中,所述目标音频处理模式包括码率切换处理模式;所述处理模块 5553 还配置为基于与所述音频场景对应的音频场景类型,查询不同候选音频场景类型与候选码率切换处理模式的对应关系,得到与所述音频场景类型对应的码率切换处理模式。

在一些实施例中,所述目标音频处理模式包括码率切换处理模式;所述处理模块 5553 还配置为将所述音频场景对应的音频场景类型,与预设音频场景类型比对;当比对确定所述音频场景类型是所述预设音频场景类型时,将与所述预设音频场景类型关联的码率切换处理模式,作为与所述音频场景类型对应的码率切换处理模式。

15 在一些实施例中,所述处理模块 5553 还配置为获取所述音频场景的通信信号强度;当所述音频场景的通信信号强度小于通信信号强度阈值时,按照第一设定比例或第一设定值降低所述音频片段的音频码率;当所述音频场景的通信信号强度大于或者等于所述通信信号强度阈值时,按照第二设定比例或第二设定值提升所述音频片段的音频码率。

20 在一些实施例中,所述处理模块 5553 还配置为基于在所述音频场景中多次采样得到的通信信号强度,确定所述音频场景中的通信信号强度的抖动信息;当所述抖动信息表征所述通信信号呈现不稳定的状态时,按照第三设定比例或第三设定值降低所述音频片段的音频码率。

在一些实施例中,所述处理模块 5553 还配置为当用于传输所述音频片段的通信网络的类型属于设定类型时,按照第四设定比例或设定值降低所述音频片段的音频码率。

25 在一些实施例中,所述音频场景分类处理是通过神经网络模型实现的,所述神经网络模型学习到所述音频片段中所包括的噪声与所述音频场景类型的关联关系;所述分类模块 5552 还配置为基于所述音频片段调用所述神经网络模型执行音频场景分类处理,得到与所述音频片段中所包括的噪声存在关联关系的音频场景类型。

30 在一些实施例中,所述神经网络模型包括映射网络、残差网络以及池化网络;所述分类模块 5552 还配置为通过所述映射网络对所述音频片段进行特征提取处理,得到所述音频片段中噪声的第一特征向量;通过所述残差网络对所述第一特征向量进行映射处理,得到所述音频片段的映射向量;通过所述映射网络对所述音频片段的映射向量进行特征提取处理,得到所述音频片段中噪声的第二特征向量;通过所述池化网络对所述第二特征向量进行池化处理,得到所述音频片段的池化向量;对所述音频片段的池化向量进行非线性映射处理,得到与所述音频片段中所包括的噪声存在关联关系的音频场景类型。

40 在一些实施例中,所述映射网络包括多个级联的映射层;所述分类模块 5552 还配置为通过所述多个级联的映射层中的第一个映射层,对所述音频片段进行特征映射处理;将所述第一个映射层的映射结果输出到后续级联的映射层,通过所述后续级联的映射层继续进行特征映射和映射结果输出,直至输出到最后一个映射层,并将所述最后一个映射层输出的映射结果作为所述音频片段中噪声的第一特征向量。

在一些实施例中,所述残差网络包括第一映射网络以及第二映射网络;所述分类模块 5552 还配置为通过所述第一映射网络对所述第一特征向量进行映射处理,得到所述音频片段的第一映射向量;对所述第一映射向量进行非线性映射处理,得到所述音频片

段的非映射向量；通过所述第一映射网络对所述音频片段的非映射向量进行映射处理，得到所述音频片段的第二映射向量；将所述音频片段的第一特征向量与所述音频片段的第二映射向量的加和结果作为所述音频片段的映射向量。

5 在一些实施例中，所述装置还包括：训练模块 5554，配置为基于无噪声的音频信号以及多个不同音频场景分别对应的背景噪声，构建所述多个不同音频场景分别对应的音频样本；基于所述多个不同音频场景分别对应的音频样本对神经网络模型进行训练，得到用于音频场景分类的神经网络模型。

10 在一些实施例中，所述训练模块 5554 还配置为针对所述多个不同音频场景中的任意一音频场景执行以下处理：基于所述音频场景的背景噪声与所述无噪声的音频信号的融合比例，对所述音频场景的背景噪声以及所述无噪声的音频信号进行融合，得到所述音频场景的第一融合音频信号；在所述第一融合音频信号中融合对应第一随机系数的所述音频场景的背景噪声，得到所述音频场景的第二融合音频信号；在所述第二融合音频信号中融合对应第二随机系数的所述无噪声的音频信号，得到所述音频场景的音频样本。

15 在一些实施例中，所述训练模块 5554 还配置为通过所述神经网络模型对所述多个不同音频场景分别对应的音频样本进行音频场景分类处理，得到所述音频样本的预测音频场景类型；基于所述音频样本的预测音频场景类型、所述音频样本的音频场景标注以及所述音频样本的权重，构建所述神经网络模型的损失函数；更新所述神经网络模型的参数直至所述损失函数收敛，将所述损失函数收敛时所述神经网络模型的更新的参数，作为用于音频场景分类的神经网络模型的参数。

20 在一些实施例中，在所述基于所述音频片段执行音频场景分类处理之前所述获取模块 5551 还配置为对所述音频片段的时域信号进行分帧处理，得到多帧音频信号；对所述多帧音频信号进行加窗处理，并对加窗处理后的音频信号进行傅里叶变换，得到所述音频片段的频域信号；对所述频域信号的梅尔频带进行对数处理，得到用于进行所述音频场景分类的所述音频片段。

25 本申请实施例提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行本申请实施例上述的基于人工智能的音频处理方法。

30 本申请实施例提供一种存储有可执行指令的计算机可读存储介质，其中存储有可执行指令，当可执行指令被处理器执行时，将引起处理器执行本申请实施例提供的基于人工智能的音频处理方法，例如，如图 3-5 示出的基于人工智能的音频处理方法。

35 在一些实施例中，计算机可读存储介质可以是 FRAM、ROM、PROM、EPROM、EEPROM、闪存、磁表面存储器、光盘、或 CD-ROM 等存储器；也可以是包括上述存储器之一或任意组合的各种设备。

40 在一些实施例中，可执行指令可以采用程序、软件、软件模块、脚本或代码的形式，按任意形式的编程语言（包括编译或解释语言，或者声明性或过程性语言）来编写，并且其可按任意形式部署，包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

45 作为示例，可执行指令可以但不一定对应于文件系统中的文件，可以可被存储在保存其它程序或数据的文件的一部分，例如，存储在超文本标记语言（HTML，Hyper Text Markup Language）文档中的一个或多个脚本中，存储在专用于所讨论的程序的单个文件中，或者，存储在多个协同文件（例如，存储一个或多个模块、子程序或代码部分的文件）中。

作为示例，可执行指令可被部署为在一个计算设备上执行，或者在位于一个地点的

多个计算设备上执行，又或者，在分布在多个地点且通过通信网络互连的多个计算设备上执行。

- 5 以上所述，仅为本申请的实施例而已，并非用于限定本申请的保护范围。凡在本申请的精神和范围之内所作的任何修改、等同替换和改进等，均包含在本申请的保护范围之内。

权利要求书

- 1、一种基于人工智能的音频处理方法，所述方法包括：
获取音频场景的音频片段，其中，所述音频片段中包括噪声；
基于所述音频片段执行音频场景分类处理，得到与所述音频片段中的噪声对应的音
5 频场景类型；
确定与所述音频场景类型对应的目标音频处理模式，并根据所述音频片段中的噪声
所带来的干扰度，对所述音频片段应用所述目标音频处理模式。
- 2、根据权利要求1所述的方法，其中，
所述目标音频处理模式包括降噪处理模式；
10 所述确定与所述音频场景类型对应的目标音频处理模式，包括：
基于所述音频场景对应的音频场景类型，查询不同候选音频场景类型与候选降噪处
理模式的对应关系，得到与所述音频场景类型对应的降噪处理模式。
- 3、根据权利要求1所述的方法，其中，
所述目标音频处理模式包括降噪处理模式；
15 所述确定与所述音频场景类型对应的目标音频处理模式，包括：
基于所述音频场景对应的音频场景类型，确定与所述音频场景类型匹配的噪声类型；
基于与所述音频场景类型匹配的噪声类型，查询不同候选噪声类型与候选降噪处
理模式的对应关系，得到与所述音频场景类型对应的降噪处理模式；
其中，与不同的所述音频场景类型匹配的噪声类型不完全相同。
- 20 4、根据权利要求2或3所述的方法，其中，所述根据所述音频片段中的噪声所带
来的干扰度，对所述音频片段应用所述目标音频处理模式，包括：
确定所述音频片段中的噪声所带来的干扰度；
当所述干扰度大于干扰度阈值时，对所述音频片段应用与所述音频场景类型对应的
降噪处理模式。
- 25 5、根据权利要求2或3所述的方法，其中，所述对所述音频片段应用所述目标音
频处理模式，包括：
将所述音频场景类型匹配的噪声类型，与所述音频片段中的噪声进行匹配处理；
对与所述噪声类型匹配成功的噪声进行抑制处理，得到抑制后的所述音频片段；
其中，抑制后的所述音频片段中的语音信号强度与噪声信号强度的比值低于信噪比
30 阈值。
- 6、根据权利要求1所述的方法，其中，
所述目标音频处理模式包括码率切换处理模式；
所述确定与所述音频场景类型对应的目标音频处理模式，包括：
基于与所述音频场景对应的音频场景类型，查询不同候选音频场景类型与候选码率
35 切换处理模式的对应关系，得到与所述音频场景类型对应的码率切换处理模式。
- 7、根据权利要求1所述的方法，其中，
所述目标音频处理模式包括码率切换处理模式；
所述确定与所述音频场景类型对应的目标音频处理模式，包括：
将所述音频场景对应的音频场景类型，与预设音频场景类型比对；
40 当比对确定所述音频场景类型是所述预设音频场景类型时，将与所述预设音频场景
类型关联的码率切换处理模式，作为与所述音频场景类型对应的码率切换处理模式。
- 8、根据权利要求6或7所述的方法，其中，所述对所述音频片段应用所述目标音
频处理模式，包括：
获取所述音频场景的通信信号强度；

当所述音频场景的通信信号强度小于通信信号强度阈值时,按照第一设定比例或第一设定值降低所述音频片段的音频码率;

当所述音频场景的通信信号强度大于或者等于所述通信信号强度阈值时,按照第二设定比例或第二设定值提升所述音频片段的音频码率。

- 5 9、根据权利要求 6 或 7 所述的方法,其中,所述对所述音频片段应用所述目标音频处理模式,包括:

基于在所述音频场景中多次采样得到的通信信号强度,确定所述音频场景中的通信信号强度的抖动信息;

- 10 当所述抖动信息表征所述通信信号呈现不稳定的状态时,按照第三设定比例或第三设定值降低所述音频片段的音频码率。

10、根据权利要求 6 或 7 所述的方法,其中,所述对所述音频片段应用所述目标音频处理模式,包括:

当用于传输所述音频片段的通信网络的类型属于设定类型时,按照第四设定比例或设定值降低所述音频片段的音频码率。

- 15 11、根据权利要求 1 所述的方法,其中,

所述音频场景分类处理是通过神经网络模型实现的,所述神经网络模型学习到所述音频片段中所包括的噪声与所述音频场景类型的关联关系;

所述基于所述音频片段执行音频场景分类处理,得到与所述音频片段中的噪声对应的音频场景类型,包括:

- 20 基于所述音频片段调用所述神经网络模型执行音频场景分类处理,得到与所述音频片段中所包括的噪声存在关联关系的音频场景类型。

12、根据权利要求 11 所述的方法,其中,

所述神经网络模型包括映射网络、残差网络以及池化网络;

所述基于所述音频片段调用所述神经网络模型以执行音频场景分类处理,包括:

- 25 通过所述映射网络对所述音频片段进行特征提取处理,得到所述音频片段中噪声的第一特征向量;

通过所述残差网络对所述第一特征向量进行映射处理,得到所述音频片段的映射向量;

- 30 通过所述映射网络对所述音频片段的映射向量进行特征提取处理,得到所述音频片段中噪声的第二特征向量;

通过所述池化网络对所述第二特征向量进行池化处理,得到所述音频片段的池化向量;

对所述音频片段的池化向量进行非线性映射处理,得到与所述音频片段中所包括的噪声存在关联关系的音频场景类型。

- 35 13、根据权利要求 12 所述的方法,其中,

所述映射网络包括多个级联的映射层;

所述通过所述映射网络对所述音频片段进行特征提取处理,得到所述音频片段中噪声的第一特征向量,包括:

- 40 通过所述多个级联的映射层中的第一个映射层,对所述音频片段进行特征映射处理;

将所述第一个映射层的映射结果输出到后续级联的映射层,通过所述后续级联的映射层继续进行特征映射和映射结果输出,直至输出到最后一个映射层,并

将所述最后一个映射层输出的映射结果作为所述音频片段中噪声的第一特征向量。

14、根据权利要求 12 所述的方法,其中,

所述残差网络包括第一映射网络以及第二映射网络;

所述通过所述残差网络对所述第一特征向量进行映射处理,得到所述音频片段的映射向量,包括:

通过所述第一映射网络对所述第一特征向量进行映射处理,得到所述音频片段的第一映射向量;

5 对所述第一映射向量进行非线性映射处理,得到所述音频片段的非映射向量;

通过所述第一映射网络对所述音频片段的非映射向量进行映射处理,得到所述音频片段的第二映射向量;

将所述音频片段的第一特征向量与所述音频片段的第二映射向量的加和结果作为所述音频片段的映射向量。

10 15、根据权利要求1所述的方法,其中,所述方法还包括:

基于无噪声的音频信号以及多个不同音频场景分别对应的背景噪声,构建所述多个不同音频场景分别对应的音频样本;

基于所述多个不同音频场景分别对应的音频样本对神经网络模型进行训练,得到用于音频场景分类的神经网络模型。

15 16、根据权利要求15所述的方法,其中,所述基于无噪声的音频信号以及多个不同音频场景分别对应的背景噪声,构建所述多个不同音频场景分别对应的音频样本,包括:

针对所述多个不同音频场景中的任一音频场景执行以下处理:

20 基于所述音频场景的背景噪声与所述无噪声的音频信号的融合比例,对所述音频场景的背景噪声以及所述无噪声的音频信号进行融合,得到所述音频场景的第一融合音频信号;

在所述第一融合音频信号中融合对应第一随机系数的所述音频场景的背景噪声,得到所述音频场景的第二融合音频信号;

25 在所述第二融合音频信号中融合对应第二随机系数的所述无噪声的音频信号,得到所述音频场景的音频样本。

17、根据权利要求15所述的方法,其中,所述基于所述多个不同音频场景分别对应的音频样本对神经网络模型进行训练,得到用于音频场景分类的神经网络模型,包括:

通过所述神经网络模型对所述多个不同音频场景分别对应的音频样本进行音频场景分类处理,得到所述音频样本的预测音频场景类型;

30 基于所述音频样本的预测音频场景类型、所述音频样本的音频场景标注以及所述音频样本的权重,构建所述神经网络模型的损失函数;

更新所述神经网络模型的参数直至所述损失函数收敛,将所述损失函数收敛时所述神经网络模型的更新的参数,作为用于音频场景分类的神经网络模型的参数。

35 18、根据权利要求1所述的方法,其中,在所述基于所述音频片段执行音频场景分类处理之前,所述方法还包括:

对所述音频片段的时域信号进行分帧处理,得到多帧音频信号;

对所述多帧音频信号进行加窗处理,并对加窗处理后的音频信号进行傅里叶变换,得到所述音频片段的频域信号;

40 对所述频域信号的梅尔频带进行对数处理,得到用于进行所述音频场景分类的所述音频片段。

19、一种音频处理装置,所述装置包括:

获取模块,配置为获取音频场景的音频片段,其中,所述音频片段中包括噪声;

分类模块,配置为基于所述音频片段执行音频场景分类处理,得到与所述音频片段中的噪声对应的音频场景类型;

处理模块，配置为确定与所述音频场景类型对应的目标音频处理模式，并根据所述音频片段中的噪声所带来的干扰度，对所述音频片段应用所述目标音频处理模式。

20、一种电子设备，所述电子设备包括：

存储器，用于存储可执行指令；

5 处理器，用于执行所述存储器中存储的可执行指令时，实现权利要求 1 至 18 任一项所述的基于人工智能的音频处理方法。

21、一种计算机可读存储介质，存储有可执行指令，用于被处理器执行时实现权利要求 1 至 18 任一项所述的基于人工智能的音频处理方法。

10 22、一种计算机程序产品，包括计算机程序或指令，所述计算机程序或指令使得计算机执行如权利要求 1 至 18 任一项所述的基于人工智能的音频处理方法。