

音频处理的方法、装置、设备和存储介质

5 本申请要求 2024 年 1 月 23 日递交的、标题为“音频处理的方法、装置、设备和存储介质”、申请号为 202410096064.X 的中国发明专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

10 本公开的示例实施例总体涉及计算机领域，特别地涉及音频处理的方法、装置、设备和计算机可读存储介质。

背景技术

15 随着计算机技术的发展，互联网已经成为人们信息交互的重要平台。在人们通过互联网进行信息交互的过程中，各种类型的音频已经成为人们社交表达和交换信息的重要媒介。因此，期待通过对音频进行处理以达到在保留音色的基础上能够改变说话风格的变声技术。

发明内容

20 在本公开的第一方面，提供了一种音频处理的方法。该方法包括：获取用户输入的第一媒体内容，第一媒体内容包括第一音频内容；以及基于用户对目标风格的选择，提供第二媒体内容，第二媒体内容包括基于第一音频内容生成的第二音频内容，第二音频内容具有与第一音频内容相同的音色，并且第二音频内容具有与目标风格相对应的至少一项音频属性。

25 在本公开的第二方面，提供了一种用于音频处理的装置。该装置包括：获取模块，被配置为获取用户输入的第一媒体内容，第一媒体内容包括第一音频内容；以及提供模块，被配置为基于用户对目标风格的选择，提供第二媒体内容，第二媒体内容包括基于第一音频内容生成的第二音频内容，第二音频内容具有与第一音频内容相同的音色，并且第二音频内容具有与目标风格相对应的至少一项音频属性。

30 在本公开的第三方面，提供了一种电子设备。该设备包括至少一个处理单元；以及至少一个存储器，至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令。指令在由至少一个处理单元执行时使设备执行第一方面至第四方面中任一方面的方法。

在本公开的第四方面，提供了一种计算机可读存储介质。该计算机可读存储介质

上存储有计算机程序，计算机程序可由处理器执行以实现第一方面至第四方面中任一方面的方法。

应当理解，本内容部分中所描述的内容并非旨在限定本公开的实施例的关键特征或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的描述而变得容易理解。

附图说明

结合附图并参考以下详细说明，本公开各实施例的上述和其他特征、优点及方面将变得更加明显。在附图中，相同或相似的附图标记表示相同或相似的元素，其中：

- 图 1 示出了其中可以实施根据本公开的实施例的示例环境的示意图；
- 图 2 示出了根据本公开的一些实施例的用于音频处理的示例过程的流程图；
- 图 3A 至图 3C 示出了根据本公开的一些实施例的示例界面的示意图；
- 图 4 示出了根据本公开的一些实施例的用于音频处理的示例装置的示意性结构框图；以及
- 图 5 示出了能够实施本公开的多个实施例的电子设备的框图。

具体实施方式

下面将参照附图更详细地描述本公开的实施例。虽然附图中示出了本公开的某些实施例，然而应当理解的是，本公开可以通过各种形式来实现，而且不应该被解释为限于这里阐述的实施例，相反，提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是，本公开的附图及实施例仅用于示例性作用，并非用于限制本公开的保护范围。

需要注意的是，本文中所提供的任何节/子节的标题并不是限制性的。本文通篇描述了各种实施例，并且任何类型的实施例都可以包括在任何节/子节下。此外，在任一节/子节中描述的实施例可以以任何方式与同一节/子节和/或不同节/子节中描述的任何其他实施例相结合。

在本公开的实施例的描述中，术语“包括”及其类似用语应当理解为开放性包含，即“包括但不限于”。术语“基于”应当理解为“至少部分地基于”。术语“一个实施例”或“该实施例”应当理解为“至少一个实施例”。术语“一些实施例”应当理解为“至少一些实施例”。下文还可能包括其他明确的和隐含的定义。术语“第一”、“第二”等可以指代不同的或相同的对象。下文还可能包括其他明确的和隐含的定义。

本公开的实施例中可能涉及用户的数据、数据的获取和/或使用等。这些方面均遵循相应的法律法规及相关规定。在本公开的实施例中，所有数据的采集、获取、处理、加工、转发、使用等，都是在用户知晓并且确认的前提下进行的。相应地，在实现本

公开的各实施例时，均应根据相关法律法规通过适当的方式，将可能所涉及的数据或信息的类型、使用范围、使用场景等告知用户并获得用户的授权。具体的告知和/或授权方式可以根据实际情况和应用场景而变化，本公开的范围在此方面不受限制。

本说明书及实施例中方案，如涉及个人信息处理，则均会在具备合法性基础（例如征得个人信息主体同意，或者为履行合同所必需等）的前提下进行处理，且仅会在规定或者约定的范围内进行处理。用户拒绝处理基本功能所需必要信息以外的个人信息，不会影响用户使用基本功能。

在人们通过互联网进行信息交互的过程中，人们期望使用优质的音频处理方法来便捷地达到需要的变声效果。传统的音频处理方法，是通过变声器，进行变声。然而，变声器只可以改变说话人的音色。

鉴于此，本公开的实施例提出了一种音频处理方案。根据该方案，可以获取用户输入的第一媒体内容，第一媒体内容包括第一音频内容。进一步地，可以基于用户对目标风格的选择，提供第二媒体内容，第二媒体内容包括基于第一音频内容生成的第二音频内容，第二音频内容具有与第一音频内容相同的音色，并且第二音频内容具有与目标风格相对应的至少一项音频属性。通过此方式，本公开的实施例能够在保留音色的基础上，提高变声效果。

以下进一步结合附图来详细描述该方案的各种示例实现。

示例环境

图 1 示出了本公开的实施例能够在其中实现的示例环境 100 的示意图。如图 1 所示，示例环境 100 可以包括终端设备 110。

在该示例环境 100 中，终端设备 110 可以运行有支持对音频进行处理的平台。例如，对音频进行变声处理。用户 140 可以经由终端设备 110 和/或其附接设备来与该平台进行交互。

在图 1 的环境 100 中，如果平台处于活动状态，终端设备 110 可以通过平台呈现用于支持界面交互的界面 150。

在一些实施例中，终端设备 110 与服务器 130 通信，以实现对平台的服务的供应。终端设备 110 可以是任意类型的移动终端、固定终端或便携式终端，包括移动手机、台式计算机、膝上型计算机、笔记本计算机、上网本计算机、平板计算机、媒体计算机、多媒体平板、掌上电脑、便携式游戏终端、VR/AR 设备、个人通信系统（Personal Communication System，PCS）设备、个人导航设备、个人数字助理（Personal Digital Assistant，PDA）、音频/视频播放器、数码相机/摄像机、定位设备、电视接收器、无线电广播接收器、电子书设备、游戏设备或者前述各项的任意组合，包括这些设备的配件和外设或者其任意组合。在一些实施例中，终端设备 110 也能够支持任意类型的

针对用户的接口（诸如“可佩戴”电路等）。

服务器 130 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，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、内容分发网络、以及大数据和人工智能平台等基础云计算服务的云服务器。服务器 130 例如可以包括计算系统/服务器，诸如大型机、边缘计算节点、云环境中的计算设备，等等。服务器 130 可以为终端设备 110 中支持虚拟场景的应用 120 提供后台服务。

服务器 130 与终端设备 110 之间可以建立有通信连接。通信连接可以通过有线方式或无线方式建立。通信连接可以包括但不限于蓝牙连接、移动网络连接、通用串行总线（Universal Serial Bus，USB）连接、无线保真（Wireless Fidelity，WiFi）连接等，本公开的实施例在此方面不受限制。在本公开的实施例中，服务器 130 与终端设备 110 可以通过二者之间的通信连接实现信令交互。

应当理解，仅出于示例性的目的描述环境 100 中各个元素的结构和功能，而不暗示对于本公开的范围的任何限制。

以下将继续参考附图描述本公开的一些示例实施例。

示例过程

图 2 示出了根据本公开的一些实施例的用于音频处理的示例过程 200 的流程图。过程 200 可以被实现在终端设备 110 处。下面参考图 1 来描述过程 200。

如图 2 所示，在框 210，终端设备 110 获取用户 140 输入的第一媒体内容。第一媒体内容包括第一音频内容。

在一些实施例中，终端设备 110 所获取的用户 140 输入的第一媒体内容可以为用户 140 录制的第一媒体内容。例如，用户拍摄的一段视频、或录制一段语音以作为第一媒体内容。第一媒体内容可以包括一段音频内容，这样的音频内容例如可以包括用户的说话内容、或者歌唱内容等。

在一些实施例中，终端设备 110 所获取的用户 140 输入的第一媒体内容可以为用户 140 上传的第一媒体内容。例如，终端设备 110 上存储的先前的一段视频、或存储的先前录制的一段语音以作为第一媒体内容。

以下将参考图 3A 至图 3C 来描述过程 200。图 3A 至图 3C 示出了根据本公开的一些实施例的示例界面 301 至 303 的示意图。界面 301 至界面 303 例如可以由图 1 所示的终端设备 110 所提供。

如图 3A 所示，终端设备 110 获取用户 140 在拍摄页面录制一段视频，该视频包括第一音频。例如，用户 140（拍摄者）的说的话语音信息。随后，用户 140 点击界面 301 中的声音控件 311。

终端设备 110 基于用户 140 的点击，将呈现选择面板 320。如图 3B 所示，终端设备 110 可以在界面 302 中显示选择面板 320。作为示例，该选择面板 320 可以显示不同风格的音频效果，例如，风格一 321、风格二 322、风格三，等等。

5 在一些实施例中，终端设备 110 显示的选择面板 320，可以提供用于提供一组候选音频效果。在一些示例中，终端设备 110 获取用户 140 输入的说话内容，则选择面板 320 可以提供针对说话内容可以转换的不同的风格。

10 在一些示例中，终端设备 110 获取用户 140 以普通话说的一段话，选择面板 320 可以提供“英文”风格的音频效果、“方言”风格的音频效果，等等。可以理解的，如果用户 140 选择选择面板 320 上呈现的“英文”风格的音频效果，终端设备 110 可将用户 140 以普通说的普通话，转换为以用户 140 的音色形成的英文版本的一段话。如果用户 140 选择选择面板 320 上呈现的“方言”风格的音频效果，终端设备 110 可将用户 140 以方言说的普通话，转换为以用户 140 的音色形成的方言版本的一段话。这仅仅是示例性的本公开对此不做限定。

15 在一些实施例中，终端设备 110 接收用户针对一组候选音频效果中的目标音频效果的选择。该目标音频效果对应于目标风格。例如，终端设备 110 可以接收用户 140 所选择的该组候选音频效果中“英文”风格、或者“某某方言”风格，等等。

20 继续参考图 2，在框 220，终端设备 110 基于用户 140 对目标风格的选择，提供第二媒体内容。在一些实施例中，第二媒体内容包括第二音频内容。第二音频内容包括基于第一音频内容生成的第二音频内容。第二音频内容具有与第一音频内容相同的音色。第二音频内容具有与目标风格相对应的至少一项音频属性。在一些示例中，终端设备 110 在保留第一音频对应音色的基础上，基于用户 140 选择的目标风格，将第一音频内容转换为第二音频内容。该第二音频内容至少具有与目标风格对应的一项音频属性。

25 在一些实施例中，第二音频内容具有与目标风格相对应的至少一项音频属性。目标音频属性至少包括音调、节奏等等。例如，如果第一音频内容的音调包括 A 调，则第二音频内容至少保留 A 调。

以图 3C 作为示例，用户 140 在选择面板 320 上选择风格三 331 作为目标风格。终端设备 110 接收到用户 140 的选择，将把第一媒体内容中的第一音频内容转换为第二媒体内容中的第二音频内容。

30 在一些示例中，终端设备 110 获取用户 140 输入的说话内容，则可以更改不同风格的说话内容。例如，将用户 140 输入的以普通话说说的“一段话”（第一音频内容）转换为与用户 140 相同音色的、且以“方言”风格述说的“一段话”（第二音频内容）。终端设备 110 基于用户对目标风格的选择，调用服务器 130 转换第一音频内容，并提供第二音频内容。

在本公开实施例中，通过此种方式，用户可以低成本听到以自己的音色用其他风格说话是什么效果。

以下对生成第二媒体内容进行说明。终端设备 110 根据第一媒体内容包括的第一音频内容，生成第二音频内容。在一些实施例中，终端设备 110 根据第二音频内容，
5 调整第一媒体内容包括的第一视频内容的视频内容的播放速度。在一些实施例中，终端设备 110 通过确定第二音频内容中与目标内容对应的音频部分，以及视觉内容中与目标内容对应的视频部分。从而，调整视频部分的播放速度，使得视频部分与音频部分同步。在一些示例中，目标内容可以为“一段话”、“一句话”、“一个字”、或者“一个词”，等等。

10 在一些示例中，终端设备 110 将用户 140 输入的以普通话述说的“一段话”转换为与用户 140 相同音色的、且以“方言”风格述说的“一段话”。然后，终端设备 110 确定第二音频内容中与“一段话”对应的音频部分，以及确定第一视频内容的视觉内容中与“一段话”对应的视频部分。随后，终端设备 110 根据其确定的音频部分和视频部分，调整视频部分的播放速度，进而，使得视频部分与音频部分同步。

15 在一些实施例中，终端设备 110 根据第二音频内容和经调整的视觉内容，生成第二视频内容，以作为第二媒体内容。本公开实施例，通过调整第一视频内容的视觉内容的播放速度，已适配变声后的第二音频内容，能够在保留音色的基础上，缩小音画不一致性。

以下对生成第二音频内容进行说明。终端设备 110 从第一媒体内容中提取第一音频内容。随后，终端设备 110 将第一音频内容输入目标模型，以获取第二音频内容。
20 在一些实施例中，目标模型可以是服务器根据目标风格对应的样本数据训练的模型。

在一些实施例中，目标模型包括语音识别模块、风格转换模块、语音生成模块。语言识别模块被配置为：终端设备 110 确定与第一音频内容对应的文本内容。在一些示例中，终端设备 110 将用户 140 输入的第一音频内容发送至服务器 130。服务器 130
25 通过调用目标模型中的语音识别模块，将第一音频内容转换为文本内容。服务器 130 将第一音频内容对应的文本把内容发送给终端设备 110。终端设备 110 确定与第一音频内容对应的文本内容。

在一些实施例中，风格转换模块被配置为：将文本内容转换为与目标风格对应的第一特征。在一些实施例中，第一特征至少指示与目标风格对应的重音转换风格。在
30 一些示例中，终端设备 110 将第一音频内容对应的文本内容以及用户 140 选择的目标风格发送至服务器 130。服务器 130 利用目标模型中的风格转换模块，将文本内容转换为与目标风格对应的第一特征。例如，将源口音或语音风格中话语的时变内容特征映射到目标口音或语音风格中的内容特征。

在一些实施例中，语音生成模块被配置为：基于第一特征和第二特征，生成中间

音频内容。第二特征是用于表征第一音频内容的音色的。在一些实施例中，语音生成模块还包括扩散模型。在一些示例中，服务器 130 根据其生成的第一特征和用于表征第一音频内容的音色的第二特征，生成中间音频内容。服务器 130 将生成的中间音频发送至终端设备 110。

5 例如，目标模型 bottleneck-to-bottleneck（即 BN2BN）模型，该模型将源口音或语音风格中话语的时变内容特征映射到目标口音或语音风格中的内容特征。然后，通过去噪扩散概率模型（扩散模型）执行零镜头的语音转换的条件。

 对于面向产品的模型，使用预训练语音识别技术 ASR 模型中提取的中间瓶颈特征来处理时变内容特征（例如，BN L10）。该扩散模型采用 BN2BN 模型的重音转换内容特征以及作为调节信号的话语级说话者嵌入，能够进行快速的语音转换，即保留训练期间未见的任何任意源说话者的音色。

 本公开实施例通过利用目标模型处理第一音频内容，生成第二音频内容，能够实现零镜头的身份保护口音和语音风格转换。

 综上，本公开实施例通过获取用户输入的第一媒体内容包括的第一音频内容，并根据用户对目标风格的选择，提供包括具有与第一音频内容相同的音色的第二音频内容的第二媒体内容。与之相应的，第二音频内容具有与目标风格相对应的至少一项音频属性，由此，能够在保留音色的基础上，提高变声效果。进一步的，通过调整第一视频内容的视觉内容的播放速度，以适配变声后的第二音频内容，能够在保留音色的基础上，达到音画同步的效果。

20

示例装置和设备

 本公开的实施例还提供了用于实现上述方法或过程的相应装置。图 4 示出了根据本公开的某些实施例的用于音频处理的示例装置 400 的示意性结构框图。装置 400 可以被实现为或者被包括在终端设备 110 中。装置 400 中的各个模块/组件可以由硬件、软件、固件或者它们的任意组合来实现。

25 如图 4 所示，装置 400 包括获取模块 410，被配置为获取用户输入的第一媒体内容，第一媒体内容包括第一音频内容。

 装置 400 还包括提供模块 420，被配置为基于用户对目标风格的选择，提供第二媒体内容，第二媒体内容包括基于第一音频内容生成的第二音频内容，第二音频内容具有与第一音频内容相同的音色，并且第二音频内容具有与目标风格相对应的至少一项音频属性。

30 在一些实施例中，至少一项音频属性包括以下至少一项：音调、节奏。

 在一些实施例中，提供模块 420 还包括选择模块，选择模块被配置为显示选择面板，其中选择面板至少提供一组候选音频效果；以及接收用户针对一组音频效果中的

目标音频效果的选择，其中目标音频效果对应于目标风格。

在一些实施例中，提供模块 420 还包括生成模块，生成模块被配置为基于第一音频内容，生成第二音频内容；基于第二音频内容，调整第一视频内容的视觉内容的播放速度；以及基于第二音频内容和经调整的视觉内容，生成第二视频内容，以作为第二媒体内容。

在一些实施例中，提供模块 420 还包括调整模块，调整模块被配置为确定第二音频内容中与目标内容对应的音频部分以及视觉内容中与目标内容对应的视频部分；以及调整视频部分的播放速度，使得视频部分与音频部分同步。

在一些实施例中，第一媒体内容包括第一视频内容，获取模块 410，还被配置为获取用户录制的第一媒体内容；以及获取用户上传的第一媒体内容。

在一些实施例中，生成模块，还被配置为从第一媒体内容中提取第一音频内容；以及利用目标模型处理第一音频内容，以生成第二音频内容，其中目标模型是基于与目标风格对应的样本数据而被训练。

在一些实施例中，目标模型包括：语音识别模块，被配置为确定与第一音频内容对应的文本内容；风格转换模块，被配置为将文本内容转换为与目标风格对应的第一特征；以及语音生成模块，被配置为基于第一特征和第二特征生成中间音频内容，第二特征用于表征第一音频内容的音色。

在一些实施例中，语音生成模块包括扩散模型。

在一些实施例中，第一特征至少指示与目标风格对应的重音转换风格。

图 5 示出了其中可以实施本公开的一个或多个实施例的电子设备 500 的框图。应当理解，图 5 所示出的电子设备 500 仅仅是示例性的，而不应当构成对本文所描述的实施例的功能和范围的任何限制。图 5 所示出的电子设备 500 可以用于实现图 1 的终端设备 110。

如图 5 所示，电子设备 500 是通用电子设备的形式。电子设备 500 的组件可以包括但不限于一个或多个处理器或处理单元 510、存储器 520、存储设备 530、一个或多个通信单元 540、一个或多个输入设备 550 以及一个或多个输出设备 560。处理单元 510 可以是实际或虚拟处理器并且能够根据存储器 520 中存储的程序来执行各种处理。在多处理器系统中，多个处理单元并行执行计算机可执行指令，以提高电子设备 500 的并行处理能力。

电子设备 500 通常包括多个计算机存储介质。这样的介质可以是电子设备 500 可访问的任何可以获取的介质，包括但不限于易失性和非易失性介质、可拆卸和不可拆卸介质。存储器 520 可以是易失性存储器（例如寄存器、高速缓存、随机访问存储器（RAM））、非易失性存储器（例如，只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、闪存）或它们的某种组合。存储设备 530 可以是可拆卸或不可拆

卸的介质，并且可以包括机器可读介质，诸如闪存驱动、磁盘或者任何其他介质，其可以能够用于存储信息和/或数据并且可以在电子设备 500 内被访问。

电子设备 500 可以进一步包括另外的可拆卸/不可拆卸、易失性/非易失性存储介质。尽管未在图 5 中示出，可以提供用于从可拆卸、非易失性磁盘（例如“软盘”）进行读取或写入的磁盘驱动和用于从可拆卸、非易失性光盘进行读取或写入的光盘驱动。在这些情况中，每个驱动可以由一个或多个数据介质接口被连接至总线（未示出）。存储器 520 可以包括计算机程序产品 525，其具有一个或多个程序模块，这些程序模块被配置为执行本公开的各种实施例的各种方法或动作。

通信单元 540 实现通过通信介质与其他电子设备进行通信。附加地，电子设备 500 的组件的功能可以以单个计算集群或多个计算机器来实现，这些计算机器能够通过通信连接进行通信。因此，电子设备 500 可以使用与一个或多个其他服务器、网络个人计算机（PC）或者另一个网络节点的逻辑连接来在联网环境中进行操作。

输入设备 550 可以是一个或多个输入设备，例如鼠标、键盘、追踪球等。输出设备 560 可以是一个或多个输出设备，例如显示器、扬声器、打印机等。电子设备 500 还可以根据需要通过通信单元 540 与一个或多个外部设备（未示出）进行通信，外部设备诸如存储设备、显示设备等，与一个或多个使得用户与电子设备 500 交互的设备进行通信，或者与使得电子设备 500 与一个或多个其他电子设备通信的任何设备（例如，网卡、调制解调器等）进行通信。这样的通信可以经由输入/输出（I/O）接口（未示出）来执行。

根据本公开的示例性实现方式，提供了一种计算机可读存储介质，其上存储有计算机可执行指令，其中计算机可执行指令被处理器执行以实现上文描述的方法。根据本公开的示例性实现方式，还提供了一种计算机程序产品，计算机程序产品被有形地存储在非瞬态计算机可读介质上并且包括计算机可执行指令，而计算机可执行指令被处理器执行以实现上文描述的方法。

这里参照根据本公开实现的方法、装置、设备和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理单元，从而生产出一种机器，使得这些指令在通过计算机或其他可编程数据处理装置的处理单元执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上，使得在计算机、其他可编程数据处理装置或其他设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其他可编程数据处理装置、或其他设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

5 附图中的流程图和框图显示了根据本公开的多个实现的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也
10 要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本公开的各实现，上述说明是示例性的，并非穷尽性的，并且也
15 不限于所公开的各实现。在不偏离所说明的各实现的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实现的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其他普通技术人员能理解本文公开的各个实现方式。

权利要求书

1. 一种音频处理方法，包括：
获取用户输入的第一媒体内容，所述第一媒体内容包括第一音频内容；以及
5 基于所述用户对目标风格的选择，提供第二媒体内容，所述第二媒体内容包括基于所述第一音频内容生成的第二音频内容，所述第二音频内容具有与所述第一音频内容相同的音色，并且所述第二音频内容具有与所述目标风格相对应的至少一项音频属性。
2. 根据权利要求 1 所述的方法，其中所述至少一项音频属性包括以下至少一项：
10 音调、节奏。
3. 根据权利要求 1 或 2 所述的方法，还包括：
显示选择面板，其中所述选择面板提供一组候选音频效果；以及
接收所述用户针对所述一组音频效果中的目标音频效果的选择，其中所述目标音频效果对应于所述目标风格。
- 15 4. 根据权利要求 1 或 2 所述的方法，其中所述第一媒体内容包括第一视频内容，所述第二媒体内容是基于以下过程而被生成：
基于所述第一音频内容，生成所述第二音频内容；以及
基于所述第二音频内容，调整所述第一视频内容的视觉内容的播放速度；以及
基于所述第二音频内容和经调整的所述视觉内容，生成第二视频内容，以作为所
20 述第二媒体内容。
5. 根据权利要求 4 所述的方法，其中基于所述第二音频内容调整所述第一视频内容的视觉部分的播放速度包括：
确定所述第二音频内容中与目标内容对应的音频部分以及所述视觉内容中与所述目标内容对应的视频部分；以及
25 调整所述视频部分的播放速度，使得所述视频部分与所述音频部分同步。
6. 根据权利要求 1 或 4 所述方法，其中获取用户输入的第一媒体内容包括：
获取所述用户录制的所述第一媒体内容；以及
获取所述用户上传的所述第一媒体内容。
7. 根据权利要求 1 或 5 所述的方法，其中所述第二音频内容是通过以下过程所
30 生成：
从所述第一媒体内容中提取所述第一音频内容；以及
利用目标模型处理所述第一音频内容，以生成所述第二音频内容，其中所述目标模型是基于与所述目标风格对应的样本数据而被训练。

8. 根据权利要求 7 所述的方法，其中所述目标模型包括：
语音识别模块，被配置为确定与所述第一音频内容对应的文本内容；
风格转换模块，被配置为将所述文本内容转换为与所述目标风格对应的第一特征；
以及
- 5 语音生成模块，被配置为基于所述第一特征和第二特征生成中间音频内容，所述第二特征用于表征所述第一音频内容的音色。
9. 根据权利要求 8 所述的方法，其中所述语音生成模块包括扩散模型。
10. 根据权利要求 8 所述的方法，其中所述第一特征至少指示与所述目标风格对应的重音转换风格。
- 10 11. 一种用于音频处理的装置，包括：
获取模块，被配置为获取用户输入的第一媒体内容，所述第一媒体内容包括第一音频内容；以及
提供模块，被配置为基于所述用户对目标风格的选择，提供第二媒体内容，所述第二媒体内容包括基于所述第一音频内容生成的第二音频内容，所述第二音频内容具有与
15 所述第一音频内容相同的音色，并且所述第二音频内容具有与所述目标风格相对应的至少一项音频属性。
12. 一种电子设备，包括：
至少一个处理单元；以及
至少一个存储器，所述至少一个存储器被耦合到所述至少一个处理单元并且存储
20 用于由所述至少一个处理单元执行的指令，所述指令在由所述至少一个处理单元执行时使所述电子设备执行根据权利要求 1 至 10 中任一项所述的方法。
13. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序可由处理器执行以实现根据权利要求 1 至 10 中任一项所述的方法。