(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 12 月 3 日 (03.12.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/237855 A1

(51) 国际专利分类号:
G10L 21/0208 (2013.01) *G10L 25/57* (2013.01)
G10L 21/0272 (2013.01) *G06K 9/62* (2006.01)
G10L 25/30 (2013.01) *G06K 9/00* (2006.01)

(21) 国际申请号: PCT/CN2019/102199

(22) 国际申请日: 2019 年 8 月 23 日 (23.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910466401.9 2019年5月30日 (30.05.2019) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(72) 发明人: 王健宗 (WANG, Jianzong); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。
程宁 (CHENG, Ning); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所（普通合伙） (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路10号中浩大厦1528-1530室于志光, Guangdong 518000 (CN)。

(54) **Title:** SOUND SEPARATION METHOD AND APPARATUS, AND COMPUTER READABLE STORAGE MEDIUM

(54) **发明名称:** 声音分离方法、装置及计算机可读存储介质

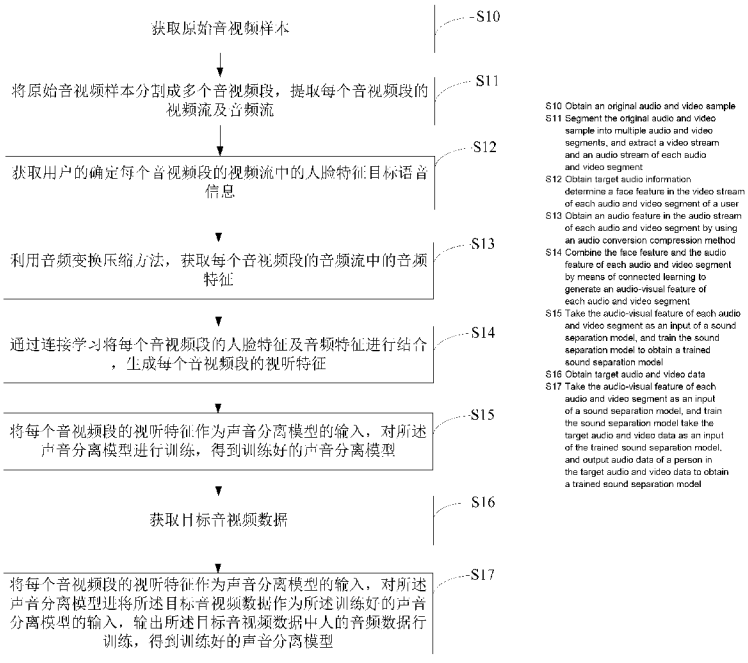

图 1

(57) **Abstract:** Disclosed in the present application is a sound separation method. The method comprises: segmenting an original audio and video sample into a plurality of audio and video segments, and extracting a video stream and an audio stream of each audio and video segment; determining a face feature in the video stream of each audio and video segment; obtaining an audio feature in the audio stream of each audio and video segment by using an audio conversion compression method; combining the face feature and the audio feature of each audio and video segment to generate an audio-visual feature of each audio and video segment; taking the

WO 2020/237855 A1

[见续页]

(81) 指定国(除另有指明，要求每一种可提供的国家保护)：AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

audio-visual feature of each audio and video segment as an input of a sound separation model, and training the sound separation model to obtain a trained sound separation model; and taking target audio and video data as an input of the trained sound separation model, and outputting audio data of a person in the target audio and video data. The present application also provides a sound separation device and a computer readable storage medium. The present application can achieve accurate mapping between a sound and a speaker and significantly improve the voice separation quality.

(57) 摘要：本申请公开了一种声音分离方法，该方法包括：将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；确定每个音视频段的视频流中的人脸特征；利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。本申请还提出一种声音分离装置以及一种计算机可读存储介质。本申请能实现声音与说话者的准确映射，显著提高语音分离的质量。

声音分离方法、装置及计算机可读存储介质

本申请要求于 2019 年 5 月 30 日提交中国专利局，申请号为201910466401.9、发明名称为“ 声音分离方法、装置及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，尤其涉及一种声音分离方法、装置及计算机可读存储介质。

背景技术

现有业内或产品的缺陷或不足或问题：在嘈杂的室内环境中，比如在鸡尾酒会中，同时存在着许多不同的声源，人类听觉非常容易从嘈杂的环境中专注于某一声音，自动“屏蔽”掉其他声音，而目前已有的计算机言语识别的智能系统无法准确在嘈杂环境中识别目标语句，无法加强选中人的语音，同时减弱同一时间其他人的音量，现有的系统未能解决“鸡尾酒会效应”。

发明内容

本申请提供一种声音分离方法、装置及计算机可读存储介质，其主要目的在于实现将目标声音从混杂声音中分离出来，实现了特定目标人的语音增强，同时削弱了其他杂音。

为实现上述目的，本申请提供一种声音分离方法，所述方法包括：

获取原始音视频样本；

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定每个音视频段的视频流中的人脸特征；

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

获取目标音视频数据；

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

为了实现上述目的，本申请还提供一种声音分离装置，所述装置包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的声音分离程序，所述声音分离程序被所述处理器执行时实现如下步骤：

获取原始音视频样本；

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定每个音视频段的视频流中的人脸特征；

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

获取目标音视频数据；

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有声音分离程序，所述声音分离程序可被一个或者多个处理器执行，以实现如上所述的声音分离方法的步骤。

本申请获取原始音视频样本；将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；确定每个音视频段的视频流中的人脸特征；利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；获取目标音视频数据；将所述目标音视频数据作为所述训练好的声音分离模型的输入，输

出所述目标音视频数据中人的音频数据。本申请结合听视觉信号来分离输入视频中的语音，实现了声音与说话者的准确映射，显著提高语音分离的质量，净化了可视化图像分离出的音轨；并通过深度学习，将目标声音从混杂声音中分离出来，生成纯净的视频，实现了特定目标人的语音增强，同时削弱了其他杂音。

附图说明

图1为本申请一实施例提供的声音分离方法的流程示意图；

图2为本申请一实施例提供的声音分离装置的内部结构示意图；

图3为本申请一实施例提供的声音分离装置中声音分离程序的模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种声音分离方法。参照图 1 所示，为本申请一实施例提供的声音分离方法的流程示意图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，声音分离方法包括：

S10、获取原始音视频样本。

在本实施例中，所述原始音视频样本包括多个应用场景的音视频。例如获取会议室的历史音视频文件，从中选择大约 10000 个小时的音视频数据。

S11、将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流。

由于音视频文件往往很大，这对数据的导入、处理、分析等操作造成效率低下、无法处理、分析等问题，为了能高效正常处理数据，本申请将原始音视频文件分割成预设大小的音视频段。将所述音视频数据切成预设大小，如大约 3 秒到 10 秒时长，的多个片段，其中，每秒包括大概 25 帧静态图像 。

S12、确定每个音视频段的视频流中的人脸特征。

在本实施例中，所述确定每个音视频段的视频流中的人脸特征包括：

将每个音视频段的视频流输入到人脸检测器中；

利用所述人脸检测器，在每个音视频段的视频流中的每帧中查找人脸图像；

使用人脸特征提取网络模型，从检测到的人脸图像中提取人脸特征作为每个音视频段的视频流中的人脸特征。

具体地，所述人脸检测器模型的训练数据可以来自一个考勤系统中的人脸数据。在一具体实现中本申请利用大量的人脸数据训练所述人脸识别模型，该人脸识别模型随数据的不断更新训练，可以得到较好的人脸识别效果。本案中基于该人脸识别模型从所述视频片段中提取检测到的人脸图像，并丢弃人脸图像间无关的变化，如由光照原因造成的变化等。

具体地，所述人脸特征提取网络模型为扩张卷积神经网络结构，所述人脸特征提取网络模型包括：卷积层、降采样层、全链接层，每一层有多个特征图。其中卷积层通过卷积运算，使原信号特征增强，同时降低噪音。降采样层根据人脸图像局部相关性的原理，对人脸图像进行子采样可以减少计算量，同时保持人脸图像旋转不变形。全链接层：采用损失函数全连接，得到的激活值即扩张卷积神经网络提取的人脸特征。

S13、利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征。

在本实施例中，所述利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征包括：

将每个音视频段的音频流中的时间和频率参数输入到短时傅里叶变换模型，得到变换后的信号

对变换后的信号执行幂律压缩，生成每个音视频段的音频流中噪声信号和纯净参考信号；

将每个音视频段的音频流中噪声信号和纯净参考信号输入至训练好的音频特征提取模型中，并输出每个音视频段的音频流中的音频特征。

通过上述实施例，将音频的时间和频率参数传入到短时傅里叶变换模型，同时执行幂律压缩，将声音进行分离，生成噪声信号和纯净参考信号数据集，并使用扩张卷积神经网络提取音频特征，从而保证训练数据的准确性，更好

的训练模型。

S14、通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征。

在本实施例中，每个音视频段的人脸特征对应每个音视频段的音频特征，从而实现人脸特征与音频特征的映射集，从而更好的训练模型。

S15、将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型。

在本实施例中，所述将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型包括：

基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型。

具体地，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

S16、获取目标音视频数据。

在本实施例中，获取目标原始音视频数据，按照类似于对训练样本的处理步骤得到所述目标原始音视频数据中每个音视频段的人脸特征及音频特征，作为所述目标音视频数据。

S17、将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

本申请获取原始音视频样本；将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；确定每个音视频段的视频流中的人脸特征；利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；获取目标音视频数据；将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。本申请结合听视觉信号来分离输入视频中的语音，实现了声音与说话者的准确映射，显著提高语音分离的质量，净化了可视化图像分离出的音轨；并通过深度学习，将目标声音从混杂声音中分离出来，生成纯净的视频，实现了特定目标人的语音增强，同时削弱了其他杂音。

本申请还提供一种声音分离装置。参照图 2 所示，为本申请一实施例提供的声音分离装置的内部结构示意图。

在本实施例中，声音分离装置 1 可以是个人电脑（Personal Computer，PC），也可以是智能手机、平板电脑、便携计算机等终端设备。该声音分离装置 1 至少包括存储器 11、处理器 12，通信总线 13，以及网络接口 14。

其中，存储器 11 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、磁性存储器、磁盘、光盘等。存储器 11 在一些实施例中可以是声音分离装置 1 的内部存储单元，例如该声音分离装置 1 的硬盘。存储器 11 在另一些实施例中也可以是声音分离装置 1 的外部存储设备，例如声音分离装置 1 上配备的插接式硬盘，智能存储卡（Smart Media Card, SMC），安全数字（Secure Digital, SD）卡，闪存卡（Flash Card）等。进一步地，存储器 11 还可以既包括声音分离装置 1 的内部存储单元也包括外部存储设备。存储器 11 不仅可以用于存储安装于声音分离装置 1 的应用软件及各类数据，例如声音分离程序 01 的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器或其他数据处理芯片，用于运行存储器 11 中存储的程序代码或处理数据，例如执行声音分离程序 01 等。

通信总线 13 用于实现这些组件之间的连接通信。

网络接口 14 可选的可以包括标准的有线接口、无线接口（如 WI-FI 接口），通常用于在该装置 1 与其他电子设备之间建立通信连接。

可选地，该装置 1 还可以包括用户接口，用户接口可以包括显示器（Display）、输入单元比如键盘（Keyboard），可选的用户接口还可以包括标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管（Organic Light-Emitting Diode，OLED）触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在声音分离装置 1 中处理的信息以及用于显示可视化的用户界面。

图 2 仅示出了具有组件 11-14 以及声音分离程序 01 的声音分离装置 1，本领域技术人员可以理解的是，图 1 示出的结构并不构成对声音分离装置 1 的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

在图 2 所示的装置 1 实施例中，存储器 11 中存储有声音分离程序 01；处理器 12 执行存储器 11 中存储的声音分离程序 01 时实现如下步骤：

获取原始音视频样本。

在本实施例中，所述原始音视频样本包括多个应用场景的音视频。例如获取会议室的历史音视频文件，从中选择大约 10000 个小时的音视频数据。

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流。

由于音视频文件往往很大，这对数据的导入、处理、分析等操作造成效率低下、无法处理、分析等问题，为了能高效正常处理数据，本申请将原始音视频文件分割成预设大小的音视频段。将所述音视频数据切成预设大小，如大约 3 秒到 10 秒时长，的多个片段，其中，每秒包括大概 25 帧静态图像。

确定每个音视频段的视频流中的人脸特征。

在本实施例中，所述确定每个音视频段的视频流中的人脸特征包括：

将每个音视频段的视频流输入到人脸检测器中；

利用所述人脸检测器，在每个音视频段的视频流中的每帧中查找人脸图像；

使用人脸特征提取网络模型，从检测到的人脸图像中提取人脸特征作为每个音视频段的视频流中的人脸特征。

具体地，所述人脸检测器模型的训练数据可以来自一个考勤系统中的人

脸数据。在一具体实现中本申请利用大量的人脸数据训练所述人脸识别模型，该人脸识别模型随数据的不断更新训练，可以得到较好的人脸识别效果。本案中基于该人脸识别模型从所述视频片段中提取检测到的人脸图像，并丢弃人脸图像间无关的变化，如由光照原因造成的变化等。

具体地，所述人脸特征提取网络模型为扩张卷积神经网络结构，所述人脸特征提取网络模型包括：卷积层、降采样层、全链接层，每一层有多个特征图。其中卷积层通过卷积运算，使原信号特征增强，同时降低噪音。降采样层根据人脸图像局部相关性的原理，对人脸图像进行子采样可以减少计算量，同时保持人脸图像旋转不变形。全链接层：采用损失函数全连接，得到的激活值即扩张卷积神经网络提取的人脸特征。

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征。

在本实施例中，所述利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征包括：

将每个音视频段的音频流中的时间和频率参数输入到短时傅里叶变换模型，得到变换后的信号

对变换后的信号执行幂律压缩，生成每个音视频段的音频流中噪声信号和纯净参考信号；

将每个音视频段的音频流中噪声信号和纯净参考信号输入至训练好的音频特征提取模型中，并输出每个音视频段的音频流中的音频特征。

通过上述实施例，将音频的时间和频率参数传入到短时傅里叶变换模型，同时执行幂律压缩，将声音进行分离，生成噪声信号和纯净参考信号数据集，并使用扩张卷积神经网络提取音频特征，从而保证训练数据的准确性，更好的训练模型。

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征。

在本实施例中，每个音视频段的人脸特征对应每个音视频段的音频特征，从而实现人脸特征与音频特征的映射集，从而更好的训练模型。

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型。

在本实施例中，所述将每个音视频段的视听特征作为声音分离模型的输

入，对所述声音分离模型进行训练，得到训练好的声音分离模型包括：

基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型。

具体地，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

获取目标音视频数据。

在本实施例中，获取目标原始音视频数据，按照类似于对训练样本的处理步骤得到所述目标原始音视频数据中每个音视频段的人脸特征及音频特征，作为所述目标音视频数据。

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

本申请获取原始音视频样本；将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；确定每个音视频段的视频流中的人脸特征；利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；获取目标音视频数据；将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。本申请结合听视觉信号来分离输入视频中的语音，实现了声音与说话者的准确映射，显著提高语音分离的质量，净化了可视化图像分离出的音轨；并通过深度学习，将目标声音从混杂声音中分离出来，生成纯净的视频，实现了特定目标人的语音增强，同时削弱了

其他杂音。

可选地，在其他实施例中，声音分离程序还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由一个或多个处理器（本实施例为处理器 12）所执行以完成本申请，本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段，用于描述声音分离程序在声音分离装置中的执行过程。

例如，参照图 3 所示，为本申请声音分离装置一实施例中的声音分离程序的程序模块示意图，该实施例中，声音分离程序可以被分割为获取模块 10、提取模块 20、确定模块 30、生成模块 40、训练模块 50 及输出模块 60，示例性地：

获取模块 10 获取原始音视频样本；

提取模块 20 将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定模块 30 确定每个音视频段的视频流中的人脸特征；

所述获取模块 10 利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

生成模块 40 通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

训练模块 50 将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

所述获取模块 10 获取目标音视频数据；

输出模块 60 将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

上述获取模块 10、提取模块 20、确定模块 30、生成模块 40、训练模块 50 及输出模块 60 等程序模块被执行时所实现的功能或操作步骤与上述实施例大体相同，在此不再赘述。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有声音分离程序，所述声音分离程序可被一个或多个处理器执行，以实现如下操作：

获取原始音视频样本；

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定每个音视频段的视频流中的人脸特征；

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

获取目标音视频数据；

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

本申请计算机可读存储介质具体实施方式与上述声音分离装置和方法各实施例基本相同，在此不作累述。

需要说明的是，上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。并且本文中的术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

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方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本申请各个实施例所述的方法。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是

利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种声音分离方法，其特征在于，所述方法包括：

获取原始音视频样本；

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定每个音视频段的视频流中的人脸特征；

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

获取目标音视频数据；

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

2、如权利要求 1 所述的声音分离方法，其特征在于，所述确定每个音视频段的视频流中的人脸特征包括：

将每个音视频段的视频流输入到人脸检测器中；

利用所述人脸检测器，在每个音视频段的视频流中的每帧中查找人脸图像；

使用人脸特征提取网络模型，从检测到的人脸图像中提取人脸特征作为每个音视频段的视频流中的人脸特征。

3、如权利要求 2 所述的声音分离方法，其特征在于，所述人脸特征提取网络模型为扩张卷积神经网络结构，所述人脸特征提取网络模型包括：卷积层、降采样层、全链接层，每一层有多个特征图。

4、如权利要求 1 所述的声音分离方法，其特征在于，所述利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征包括：

将每个音视频段的音频流中的时间和频率参数输入到短时傅里叶变换模型，得到变换后的信号

对变换后的信号执行幂律压缩，生成每个音视频段的音频流中噪声信号

和纯净参考信号；

将每个音视频段的音频流中噪声信号和纯净参考信号输入至训练好的音频特征提取模型中，并输出每个音视频段的音频流中的音频特征。

5、如权利要求 1 所述的声音分离方法，其特征在于，所述将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型包括：

基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型。

6、如权利要求 1 所述的声音分离方法，其特征在于，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

7、如权利要求 2-5 任一项所述的声音分离方法，其特征在于，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

8、一种声音分离装置，其特征在于，所述装置包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的声音分离程序，所述声音分离程序被所述处理器执行时实现如下步骤：

获取原始音视频样本；

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定每个音视频段的视频流中的人脸特征；

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

获取目标音视频数据；

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

9、如权利要求 8 所述的声音分离装置，其特征在于，所述确定每个音视频段的视频流中的人脸特征包括：

将每个音视频段的视频流输入到人脸检测器中；

利用所述人脸检测器，在每个音视频段的视频流中的每帧中查找人脸图像；

使用人脸特征提取网络模型从检测到的人脸图像中提取人脸特征作为每个音视频段的视频流中的人脸特征。

10、如权利要求 9 所述的声音分离装置，其特征在于，所述人脸特征提取网络模型为扩张卷积神经网络结构，所述人脸特征提取网络模型包括：卷积层、降采样层、全链接层，每一层有多个特征图。

11、如权利要求 8 所述的声音分离装置，其特征在于，所述利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征包括：

将每个音视频段的音频流中的时间和频率参数输入到短时傅里叶变换模型，得到变换后的信号

对变换后的信号执行幂律压缩，生成每个音视频段的音频流中噪声信号

和纯净参考信号；

将每个音视频段的音频流中噪声信号和纯净参考信号输入至训练好的音频特征提取模型中，并输出每个音视频段的音频流中的音频特征。

12、如权利要求 8 所述的声音分离装置，其特征在于，所述将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型包括：

基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型。

13、如权利要求 8 所述的声音分离装置，其特征在于，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

14、如权利要求 9-12 任一项所述的声音分离装置，其特征在于，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

15、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有声音分离程序，所述声音分离程序可被一个或者多个处理器执行，以实现如下步骤：

获取原始音视频样本；

将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流；

确定每个音视频段的视频流中的人脸特征；

利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征；

通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征；

将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型；

获取目标音视频数据；

将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据。

16、如权利要求 15 所述的计算机可读存储介质，其特征在于，所述确定每个音视频段的视频流中的人脸特征包括：

将每个音视频段的视频流输入到人脸检测器中；

利用所述人脸检测器，在每个音视频段的视频流中的每帧中查找人脸图像；

使用人脸特征提取网络模型从检测到的人脸图像中提取人脸特征作为每个音视频段的视频流中的人脸特征。

17、如权利要求 16 所述的计算机可读存储介质，其特征在于，所述人脸特征提取网络模型为扩张卷积神经网络结构，所述人脸特征提取网络模型包括：卷积层、降采样层、全链接层，每一层有多个特征图。

18、如权利要求 15 所述的计算机可读存储介质，其特征在于，所述利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征包括：

将每个音视频段的音频流中的时间和频率参数输入到短时傅里叶变换模型，得到变换后的信号

对变换后的信号执行幂律压缩，生成每个音视频段的音频流中噪声信号

和纯净参考信号；

将每个音视频段的音频流中噪声信号和纯净参考信号输入至训练好的音频特征提取模型中，并输出每个音视频段的音频流中的音频特征。

19、如权利要求 15 所述的计算机可读存储介质，其特征在于，所述将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型包括：

基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型。

20、如权利要求 15-19 任一项所述的计算机可读存储介质，其特征在于，所述基于每个音视频段的视听特征，利用 LSTM 网络及三层全链接网络训练所述声音分离模型包括：

获取第一预设数量的训练数据；

将所述第一预设数量的训练数据依次输入所述声音分离模型，利用 LSTM 网络及三层全链接网络训练所述声音分离模型，并利用第二预设数量的训练数据校验训练后的声音分离模型；

利用第二预设数量的训练数据校验训练后的声音分离模型，若训练后的声音分离模型的识别准确率大于或等于预设阈值，则训练完成；

若训练后的声音分离模型的识别准确率小于预设阈值，则发出提醒信息，提醒用户增加样本数量重新训练所述声音分离模型。

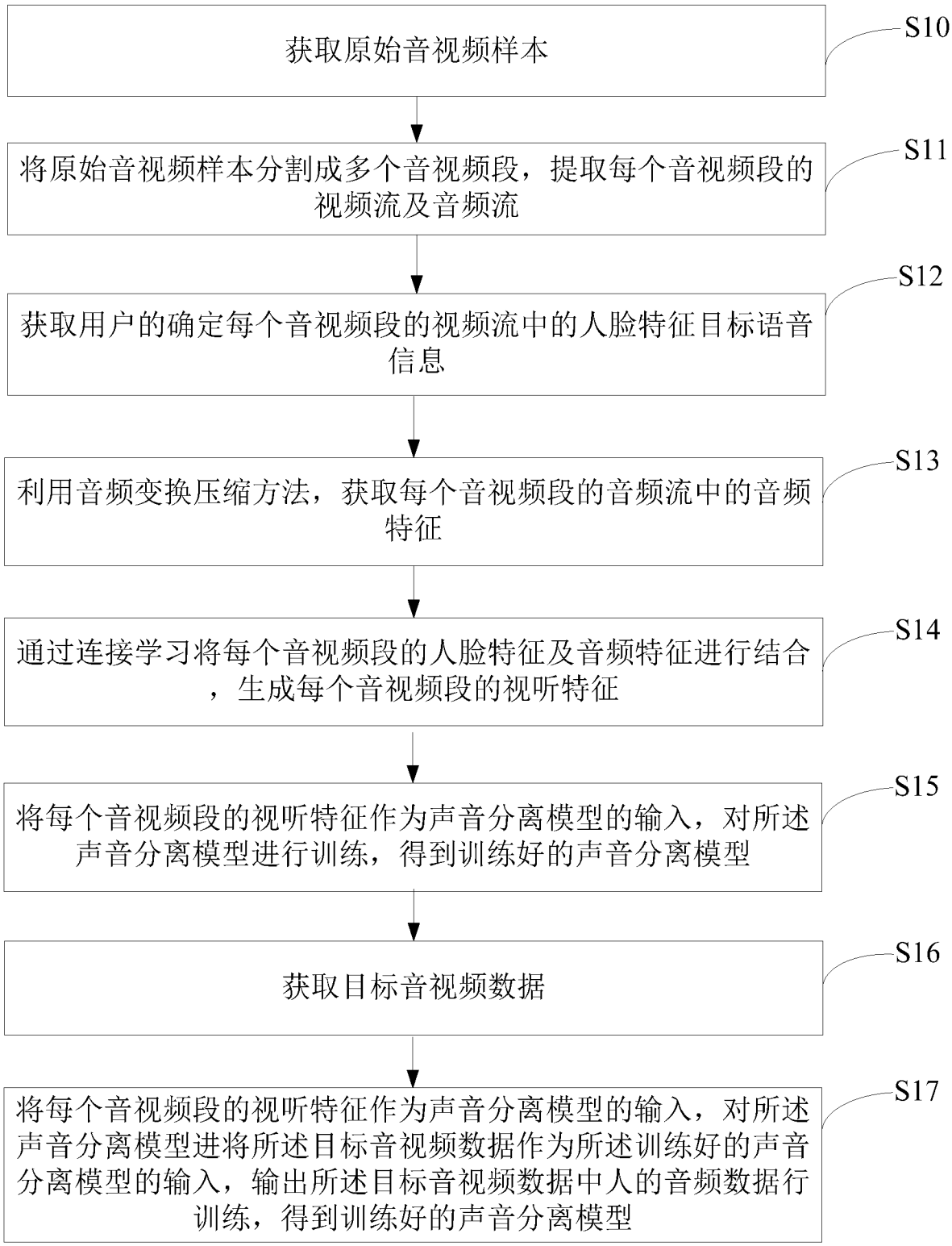
获取原始音视频样本
S10
将原始音视频样本分割成多个音视频段，提取每个音视频段的视频流及音频流
S11
获取用户的确定每个音视频段的视频流中的人脸特征目标语音信息
S12
利用音频变换压缩方法，获取每个音视频段的音频流中的音频特征
S13
通过连接学习将每个音视频段的人脸特征及音频特征进行结合，生成每个音视频段的视听特征
S14
将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进行训练，得到训练好的声音分离模型
S15
获取目标音视频数据
S16
将每个音视频段的视听特征作为声音分离模型的输入，对所述声音分离模型进将所述目标音视频数据作为所述训练好的声音分离模型的输入，输出所述目标音视频数据中人的音频数据行训练，得到训练好的声音分离模型
S17

图 1

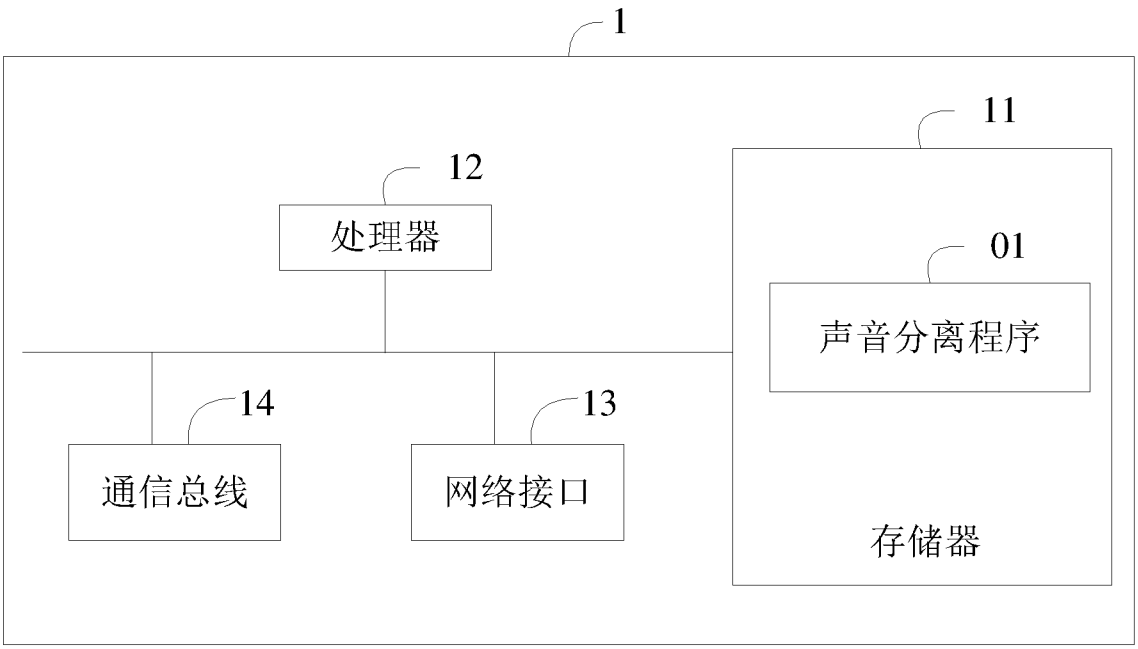
1
11
12
处理器
01
声音分离程序
14
13
通信总线
网络接口
存储器

图 2

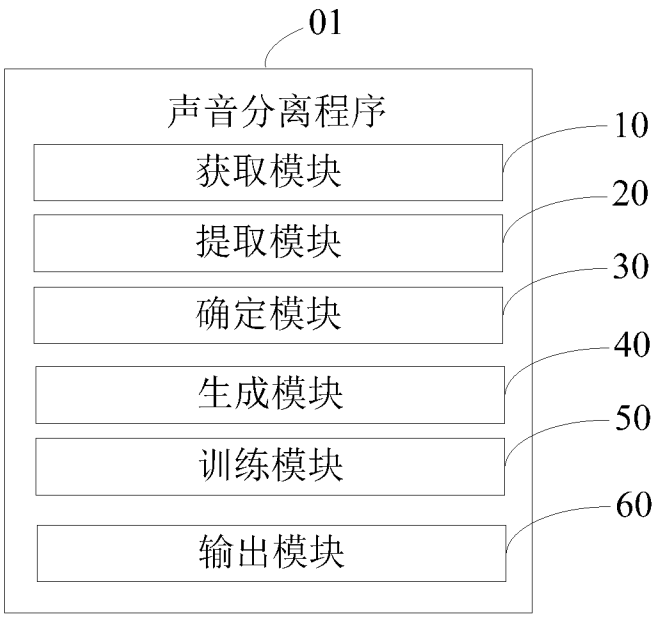
01
声音分离程序
10
获取模块
20
提取模块
30
确定模块
40
生成模块
50
训练模块
60
输出模块

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/102199

A. CLASSIFICATION OF SUBJECT MATTER

G10L 21/0208(2013.01)i; G10L 21/0272(2013.01)i; G10L 25/30(2013.01)i; G10L 25/57(2013.01)i; G06K 9/62(2006.01)i; G06K 9/00(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; VEN; WOTXT; EPTXT; USTXT; CNKI; IEEE: 声音, 噪声, 分离, 音频, 视频, 人脸, 识别, voice, noise, audio, video, face, recogn+

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 108847238 A (NORTHEASTERN UNIVERSITY) 20 November 2018 (2018-11-20) description, paragraphs [0005]-[0031], and figures 1-4	1-20
A	CN 107483445 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 15 December 2017 (2017-12-15) entire document	1-20
A	US 2016284346 A1 (QUALCOMM INC.) 29 September 2016 (2016-09-29) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 January 2020	**27 February 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/102199

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108847238 A	20 November 2018	None	
CN 107483445 A	15 December 2017	US 2019066695 A1	28 February 2019
US 2016284346 A1	29 September 2016	US 9666183 B2	30 May 2017

国际检索报告

国际申请号
PCT/CN2019/102199

A. 主题的分类

G10L 21/0208(2013.01)i; G10L 21/0272(2013.01)i; G10L 25/30(2013.01)i; G10L 25/57(2013.01)i; G06K 9/62(2006.01)i; G06K 9/00(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G10L G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNABS;CNTXT;VEN;WOTXT;EPTXT;USTXT;CNKI;IEEE:声音，噪声，分离，音频，视频，人脸，识别，voice，noise，audio，video，face，recogn+

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 108847238 A (东北大学) 2018年 11月 20日 (2018-11-20) 说明书第[0005]-[0031]段，附图1-4	1-20
A	CN 107483445 A (百度在线网络技术北京有限公司) 2017年 12月 15日 (2017-12-15) 全文	1-20
A	US 2016284346 A1 (QUALCOMM INC) 2016年 9月 29日 (2016-09-29) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：
“A” 认为不特别相关的表示了现有技术一般状态的文件
“E” 在国际申请日的当天或之后公布的在先申请或专利
“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）
“O” 涉及口头公开、使用、展览或其他方式公开的文件
“P” 公布日先于国际申请日但迟于所要求的优先权日的文件
“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件
“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性
“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性
“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 1月 17日	2020年 2月 27日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 颜彦 电话号码 86-（0512）-88997267

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/102199

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 108847238 A	2018年 11月 20日	无	
CN 107483445 A	2017年 12月 15日	US 2019066695 A1	2019年 2月 28日
US 2016284346 A1	2016年 9月 29日	US 9666183 B2	2017年 5月 30日