(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108564952 A
(43)申请公布日 2018.09.21

(21)申请号 201810198543.7

(22)申请日 2018.03.12

(71)申请人 新华智云科技有限公司
地址 310012 浙江省杭州市西湖区文一西路460号文娱中心430室

(72)发明人 徐常亮 陈凌云 廖健 范梦真

(74)专利代理机构 上海百一领御专利代理事务所(普通合伙) 31243
代理人 陈贞健 王路丰

(51)Int.Cl.
G10L 15/26(2006.01)
G10L 21/02(2013.01)

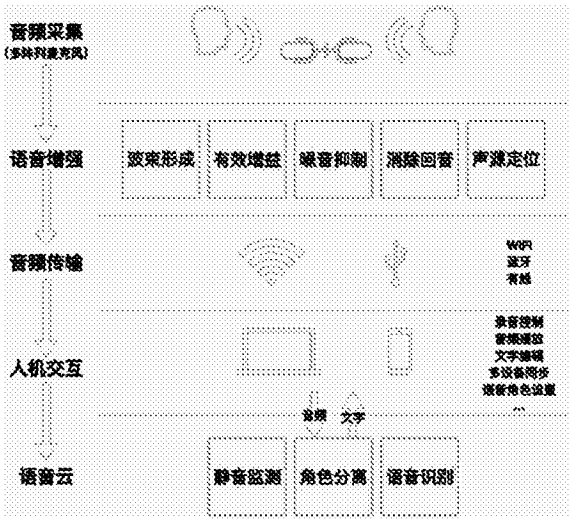
权利要求书3页 说明书9页 附图2页

(54)发明名称

语音角色分离的方法和设备

(57)摘要

本发明的目的是提供一种语音角色分离的方法和设备，通过采用多阵列指向性的麦克风，对不同人的声音，采用不同的硬件进行采集，结合算法+硬件的能力，比单纯凭借算法进行角色分离的准确率更高。记者在采访时无需了解技术细节，只需要针对不同的采访对象，摆放好相应录音设备，打开手机等人机交互设备上的App，既可将语音实时/非实时地转成文字，并拿到已经进行精准角色分离的文字结果，为记者的音频素材处理环节节约大量时间和精力。



CN 108564952 A

1.一种语音角色分离的方法,其中,该方法包括:

通过指向不同说话人的拾音头,采集指向不同说话角色对应的声道音频;

根据每个声道音频中对应所指向的说话角色,对每个声道音频进行增益处理;

根据每个声道音频中对应所指向的说话角色之外的侧面音频,对经过所述增益处理后的每个声道音频进行降噪处理;

对经过所述降噪处理处理后的每个声道音频进行消除回音的处理;

将经过消除回音处理的每个声道音频切分为音频片段,根据每个声道音频中对应所指向的说话角色,对每个音频片段标注对应的说话角色标签;

将每个音频片段转换为对应的文字,根据每个音频片段标注的说话角色标签,为对应的文字标注所述说话角色标签。

2.根据权利要求1所述的方法,其中,指向不同说话人的拾音头包括如下任一种:

单拾音头但是有多指向模式的麦克风;

手机上的两个以上麦克风;

录音笔上的两个以上麦克风;

两个以上独立设备的麦克风。

3.根据权利要求1所述的方法,其中,对经过所述降噪处理处理后的每个声道音频进行消除回音的处理,包括:

对经过所述降噪处理处理后的每个声道音频,采用基于ANC主动噪声对消的方法进行消除回音的处理。

4.根据权利要求1所述的方法,其中,根据每个声道音频中对应所指向的说话角色,对每个音频片段标注对应的说话角色标签,包括:

采用TDOA算法估计每个声道音频中的音频片段到达不同麦克风的时延差,根据所述时延差计算距离差,再通过计算得到的距离差和麦克风的空间几何来确定音频片段对应所指向的说话角色。

5.根据权利要求4所述的方法,其中,将经过消除回音处理的每个声道音频切分为音频片段,根据每个声道音频中对应所指向的说话角色,对每个音频片段标注对应的说话角色标签,包括:

人机交互单元接收经过消除回音处理的每个声道音频;

所述人机交互单元将每个声道音频切分为音频片段,根据每个声道音频中对应所指向的说话角色,对每个音频片段标注对应的说话角色标签;

所述人机交互单元将标注对应的说话角色标签的音频片段上传至云端。

6.根据权利要求5所述的方法,其中,将每个音频片段转换为对应的文字,根据每个音频片段标注的说话角色标签,为对应的文字标注所述说话角色标之后,还包括:

人机交互单元获取标注的说话角色标签后的音频片段和对应的文字;

所述人机交互单元获取用户选择的某一说话角色的对应音频和文字的请求;

所述人机交互单元基于所述请求,获取标注对应说话角色标签的音频片段和对应的文字进行播放。

7.根据权利要求1所述的方法,其中,将每个音频片段转换为对应的文字,包括:

通过通过VAD算法,识别并剔除每个音频片段中不包含语音信号的音频帧;

采用ASR计算，将识别并剔除不包含语音信号的音频帧后的音频片段转换为对应的文字。

8.根据权利要求1所述的方法，其中，指向不同说话人的拾音头的数量为2～4个，拾音头与说话角色之间的距离小于1米。

9.一种语音角色分离的设备，其中，该设备包括：

语音信号采集单元，用于通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

增强处理单元，用于根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

降噪处理单元，用于根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

自适应波束形成单元，用于对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

声源定位单元，用于将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

角色分离单元，用于将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

10.一种基于计算的设备，其中，包括：

处理器；以及

被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器：

通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

11.一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，该计算机可执行指令被处理器执行时使得该处理器：

通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

语音角色分离的方法和设备

技术领域

[0001] 本发明涉及计算机领域，尤其涉及一种语音角色分离的方法和设备。

背景技术

[0002] 随着社会各行业信息化及自动化程度的不断提升，人们对更加精准的数据的需求越来越高。以采访场景为例，录音是记者采访不可或缺的一个环节，记者们需要对音频内容的进行记录、对音频素材中的内容进行分析，摘取有效的信息，并最后写成一篇稿件，工作繁重。语音识别技术的发展，为该音频素材的处理场景提供了解决方案。

[0003] 说话人角色分离是采访音频素材处理里面的一个重要步骤。目前，大多数实现角色分离的方案主要是基于说话人的声纹特征，即接收到语音信号后，先基于BIC(英文：Bayesian Information Criterion，中文：贝叶斯信息准则)对语音信号进行说话人转折点检测，将语音信号分割成多个语音片段；然后通过采用GMM(Gaussian Mixture Model-高斯混合模型)和HMM(Hidden Markov Model-隐马尔科夫模型)对每个角色的声音进行建模。从而对说话人的声音片段进行剥离，达到角色分离的目的。

[0004] 其中，BIC(Bayesian Information Criterion-贝叶斯信息准则)是对模型的拟合效果进行评价的一个指标，BIC值越小，则模型对数据的拟合越好，BIC＝-2ln(L)+ln(n)*k。GMM(Gaussian Mixture Model-高斯混合模型)是用高斯概率密度函数精确地量化事物，将一个事物分解为若干的基于高斯概率密度函数形成的模型。HMM(Hidden Markov Model-隐马尔科夫模型)是一种统计模型，用来描述一个含有隐含未知参数的马尔科夫过程

[0005] 上述解决方案，在理想录音环境下的分离效果较好。但是在采访场景下，由于采访空间的不确定，声音传播受空间影响较大，由于空间反射，衍射，麦克风收到的信号除了直达信号以外，还有多径信号叠加，使得信号被干扰，即为混响。在室内环境中，受房间边界或者障碍物衍射，反射导致声音延续，极大程度的影响语音的可懂度，再加之说话人数的不确定，角色分离的准确率可能会大打折扣。

发明内容

[0006] 本发明的一个目的是提供一种语音角色分离的方法和设备，能够解决现有的语音角色分离的方案准确率不高的问题。

[0007] 根据本发明的一个方面，提供了一种语音角色分离的方法，该方法包括：

[0008] 通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0009] 根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0010] 根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0011] 对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0012] 将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0013] 将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0014] 进一步的，上述方法中，指向不同说话人的拾音头包括如下任一种：

[0015] 单拾音头但是有多指向模式的麦克风；

[0016] 手机上的两个以上麦克风；

[0017] 录音笔上的两个以上麦克风；

[0018] 两个以上独立设备的麦克风。

[0019] 进一步的，上述方法中，对经过所述降噪处理处理后的每个声道音频进行消除回音的处理，包括：

[0020] 对经过所述降噪处理处理后的每个声道音频，采用基于ANC主动噪声对消的方法进行消除回音的处理。

[0021] 进一步的，上述方法中，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签，包括：

[0022] 采用TDOA算法估计每个声道音频中的音频片段到达不同麦克风的时延差，根据所述时延差计算距离差，再通过计算得到的距离差和麦克风的空间几何来确定音频片段对应所指向的说话角色。

[0023] 进一步的，上述方法中，将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签，包括：

[0024] 人机交互单元接收经过消除回音处理的每个声道音频；

[0025] 所述人机交互单元将每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0026] 所述人机交互单元将标注对应的说话角色标签的音频片段上传至云端。

[0027] 进一步的，上述方法中，将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标之后，还包括：

[0028] 人机交互单元获取标注的说话角色标签后的音频片段和对应的文字；

[0029] 所述人机交互单元获取用户选择的某一说话角色的对应音频和文字的请求；

[0030] 所述人机交互单元基于所述请求，获取标注对应说话角色标签的音频片段和对应的文字进行播放。

[0031] 进一步的，上述方法中，将每个音频片段转换为对应的文字，包括：

[0032] 通过通过VAD算法，识别并剔除每个音频片段中不包含语音信号的音频帧；

[0033] 采用ASR计算，将识别并剔除不包含语音信号的音频帧后的音频片段转换为对应的文字。

[0034] 进一步的，上述方法中，指向不同说话人的拾音头的数量为2～4个，拾音头与说话角色之间的距离小于1米。

[0035] 根据本发明的另一方面，还提供了一种语音角色分离的设备，其中，该设备包括：

[0036] 语音信号采集单元，用于通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0037] 增强处理单元，用于根据每个声道音频中对应所指向的说话角色，对每个声道音

频进行增益处理；

[0038] 降噪处理单元，用于根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0039] 自适应波束形成单元，用于对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0040] 声源定位单元，用于将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0041] 角色分离单元，用于将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0042] 根据本发明的另一方面，还提供了一种基于计算的设备，其中，包括：

[0043] 处理器；以及

[0044] 被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器：

[0045] 通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0046] 根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0047] 根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0048] 对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0049] 将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0050] 将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0051] 根据本发明的另一方面，还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，该计算机可执行指令被处理器执行时使得该处理器：

[0052] 通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0053] 根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0054] 根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0055] 对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0056] 将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0057] 将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0058] 与现有技术相比，本发明由于采用多阵列指向性的麦克风，对不同人的声音，采用不同的硬件进行采集，结合算法+硬件的能力，比单纯凭借算法进行角色分离的准确率更高。记者在采访时无需了解技术细节，只需要针对不同的采访对象，摆放好相应录音设备，打开手机等人机交互设备上的App，既可将语音实时/非实时地转成文字，并拿到已经进行精准角色分离的文字结果，为记者的音频素材处理环节节约大量时间和精力。

附图说明

[0059] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本发明的其它特征、目的和优点将会变得更明显：

[0060] 图1示出本发明一实施例的语音角色分离的方法的流程图；

[0061] 图2示出本发明一实施例的语音角色分离的方法和设备的原理图；

[0062] 图3示出根据本发明一实施例的自适应噪声对消器的示意图。

[0063] 附图中相同或相似的附图标记代表相同或相似的部件。

具体实施方式

[0064] 下面结合附图对本发明作进一步详细描述。

[0065] 在本申请一个典型的配置中，终端、服务网络的设备和可信方均包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0066] 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

[0067] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0068] 如图1和2所示，本发明提供一种语音角色分离的方法，包括：

[0069] 步骤S1，通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0070] 在此，语音信号采集单元可以通过多向拾音麦克风阵列来获取不同说话人的声音，即通过多个枪型麦克风分别指向不同的说话人，来获取多路不同的音频信号，由于采用多阵列指向性的麦克风，对不同人的声音，采用不同的硬件进行采集，结合算法+硬件的能力，比单纯凭借算法进行角色分离的准确率更高，提升语音角色分离的准确率；

[0071] 步骤S2，根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0072] 在此，可通过以增强处理单元对如枪型麦克风指向的方向，获取的音频波束进行增益处理；

[0073] 步骤S3，根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0074] 在此，可以通过一降噪处理单元对各个指向的麦克风的侧面输入的音频信号进行抑制，从而进行降噪处理；

[0075] 步骤S4，对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0076] 步骤S5，将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0077] 步骤S6，将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0078] 在此，可以通过一角色分离单元根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0079] 通过本发明的方案，采用多个指向性的拾音头(如枪型麦克风)进行录音，这样可以最大程度的获取不同说话人的声音信号，避免噪音干扰。在双人说话场景下，最好是双拾音头进行拾音，效果更佳。记者在采访时无需了解技术细节，只需要针对不同的采访对象，摆放好相应录音设备，既可将语音实时/非实时地转成文字，并拿到已经进行精准角色分离的文字结果，为记者的音频素材处理环节节约大量时间和精力。

[0080] 本发明的语音角色分离的方法以实施例中，步骤S1中，指向不同说话人的拾音头包括如下任一种：

[0081] 单拾音头但是有多指向模式的麦克风；

[0082] 手机上的两个以上麦克风；

[0083] 录音笔上的两个以上麦克风；

[0084] 两个以上独立设备的麦克风。

[0085] 在此，为了区分不同的说话人，还可以支持以下方法来采集语音信号：

[0086] a)单拾音头但是有多指向模式的麦克风作为音频输入口，这样，可以将来源于不同指向的麦克风获取的音频通过不同的声道进行传输；

[0087] b)采用具有两个以上麦克风的手机作为音频输入源，如三星GALAXY S6；

[0088] c)带有两个以上麦克风的录音笔；

[0089] d)捕获直播推流：在视频采访场景中，可以通过获取来自不同设备的直播推流，来得到不同说话人的声音；

[0090] e)其他的多通道音频捕获和传输装置，如电脑自带的麦克风+独立麦克风，或手机自带的麦克风+独立麦克风。

[0091] 本发明的语音角色分离的方法以实施例中，步骤S4，对经过所述降噪处理处理后的每个声道音频进行消除回音的处理，包括：

[0092] 对经过所述降噪处理处理后的每个声道音频，采用基于ANC主动噪声对消的方法进行消除回音的处理。

[0093] 在此，可以通过一自适应波束形成单元采用ABF(Adaptive Beam Forming-自适应波束)的方案从而降低噪音和回音的干扰。ABF(Adaptive Beam Forming-自适应波束)它使用天线阵将信号能量聚集为一个很窄的波束，提高天线的传播效率和无线链路的可靠性和频率的重复使用率。如图3所示，ABF中的一种GSLC(generalized sidelobe canceller-子阵级广义旁瓣抵消器)是一种基于ANC(Auto-adapted noise cancellation自适应噪声对消器)主动噪声对消的方法，带噪信号同时通过主通道和辅助通道，而辅助通道的阻塞矩阵将语音信号滤除，得到仅包含多通道噪声的参考信号、各通道根据噪声信号得到一个最优信号估计，得到纯净语音信号估计。ANC(Auto-adapted noise cancellation自适应噪声对消器)通过将被噪音污染的声音信号与参考信号进行抵消运算，从而消除带噪信号中的噪

声。

[0094] 本发明的语音角色分离的方法以实施例中，步骤S5中，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签，包括：

[0095] 采用TDOA算法估计每个声道音频中的音频片段到达不同麦克风的时延差，根据所述时延差计算距离差，再通过计算得到的距离差和麦克风的空间几何来确定音频片段对应所指向的说话角色。

[0096] 在此，可以通过一声源定位单元采用TDOA(Time Difference of Arrival-到达时间差)算法估计每个声道音频中的音频片段(声源)到达不同麦克风的时延差，并计算距离差，再通过距离差和麦克风的空间几何来确定音频片段的位置(对应所指向的说话角色)。

[0097] TDOA(Time Difference of Arrival-到达时间差)是一种利用时间差进行定位的方法。通过测量信号到达监测站的时间，可以确定信号源的距离，利用信号源到各个监测站的距离(以监测站为中心，距离为半径作圆)，就能确定信号的位置。

[0098] 本发明的语音角色分离的方法以实施例中，步骤S5，将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签，包括：

[0099] 人机交互单元接收经过消除回音处理的每个声道音频；

[0100] 所述人机交互单元将每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0101] 所述人机交互单元将标注对应的说话角色标签的音频片段上传至云端。

[0102] 在此，可以通过一音频传输单元将音频信号，通过有线(数据线)或者无线(WiFi、蓝牙、其他无线传输信道等)的方式传输给手机/电脑端/智能硬件等人机交互单元(如手机App和网页应用、智能音箱等)，并通过该人机交互单元，将音频信号传输至云端，以通过云端进行后续的文本转换处理。

[0103] 在音频传输方面，采用USB等有线的方案进行信号传输，避免音频信号的数据丢失(无线传输比较容易受到信道传输的限制，并在传输的过程中容易丢包)。

[0104] 以录音文件(而非音频流数据)的形式上传音频数据，并转成文字，可以获取较佳的准确率。

[0105] 本发明的语音角色分离的方法以实施例中，步骤S6，将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标之后，还包括：

[0106] 人机交互单元获取标注的说话角色标签后的音频片段和对应的文字；

[0107] 所述人机交互单元获取用户选择的某一说话角色的对应音频和文字的请求；

[0108] 所述人机交互单元基于所述请求，获取标注对应说话角色标签的音频片段和对应的文字进行播放。

[0109] 在此，所述人机交互单元可提供相应的应用程序(如手机APP和网页应用)给用户使用，可以具体包括如下功能：

[0110] α)录音控制：开启录音、暂停、结束、实时保存录音；

[0111] b)用户可以在录音过程中对重要的段落进行标记，并在后续的使用过程中进行查看；

[0112] c)对不同的麦克风进行命名:设置采访对象的名称;

[0113] d)选择性播放不同说话人的音频:选择任意说话人,即可单独播放对应说话人的音频;

[0114] e)选择性展示不同说话人的文字:选择任意说话人,即可展示对应说话人音频转出来的文字;

[0115] f)点哪播哪:以句子为维度,用户可以选择不同的字词/段落进行播放;

[0116] g)对文本内容进行编辑:对录音之后转出来的文字内容进行编辑、删除、重命名;

[0117] h)搜索特定关键词:结合搜索引擎技术,用户可以输入关键词,对自有的录音、文字及说话人进行搜索;

[0118] i)下载并导出音频文件:把录音文件进行导出;

[0119] j)云端同步:可以在多设备同时使用,并对音频内容进行云端同步,避免数据的丢失。

[0120] 本发明的语音角色分离的方法以实施例中,步骤S6中,将每个音频片段转换为对应的文字,包括:

[0121] 通过通过VAD算法,识别并剔除每个音频片段中不包含语音信号的音频帧;

[0122] 采用ASR计算,将识别并剔除不包含语音信号的音频帧后的音频片段转换为对应的文字。

[0123] 在此,可以通过一静音消除单元,采用VAD算法(Voice Activity Detection一语音活动监测)识别并剔除每个音频片段中不包含语音信号的音频帧,以减少后续语音转文字的不必要的计算量。VAD(Voice Activity Detection-语音活动监测)目的是从声音信号流里识别和消除长时间的静音期,从而节约语音转写成本的方案。

[0124] 另外,ASR(Automatic Speech Recognition-语音识别)将人的语音转化成文本的技术,可通过一语音转文字单元,采用ASR(Automatic Speech Recognition-语音识别)技术,将上述每个音频片段转成文字并返回给音频转写程序。本实施例可以支持多种使用场景,包括:实时音频流转文字,和离线录音文件转文字。

[0125] 本发明的语音角色分离的方法以实施例中,步骤S1中,指向不同说话人的拾音头的数量为2～4个,拾音头与说话角色之间的距离小于1米。

[0126] 在此,麦克风数量＞＝2,对于远场的录音场景,双麦克风的结构难度更低、功耗低、使用成本更低、方案更成熟。

[0127] 或者是多个拾音头进行拾音,如果是单个拾音头,可以通过左右声道分离的方式,分别抽取不同说话人的声音。

[0128] 麦克风的距离和角度可以由用户进行自定义,用户可以根据采访场景和采访距离,对麦克风的位置进行个性化的处理。

[0129] 麦克风与说话人之间的距离最好不超过1米,此时可以取得90%以上的准确率。

[0130] 由于蓝牙传输带宽的限制,若采用蓝牙对音频信号进行传输,需要将原始的pcm格式的信号进行压缩之后再传输,麦克风数量应该小于4。

[0131] 根据本发明的另一方面,还提供了一种语音角色分离的设备,其中,该设备包括:

[0132] 语音信号采集单元,用于通过指向不同说话人的拾音头,采集指向不同说话角色对应的声道音频;

[0133] 增强处理单元，用于根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0134] 降噪处理单元，用于根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0135] 自适应波束形成单元，用于对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0136] 声源定位单元，用于将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0137] 角色分离单元，用于将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0138] 根据本发明的另一方面，还提供了一种基于计算的设备，其中，包括：

[0139] 处理器；以及

[0140] 被安排成存储计算机可执行指令的存储器，所述可执行指令在被执行时使所述处理器：

[0141] 通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0142] 根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0143] 根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0144] 对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0145] 将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0146] 将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0147] 根据本发明的另一方面，还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，该计算机可执行指令被处理器执行时使得该处理器：

[0148] 通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频；

[0149] 根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理；

[0150] 根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理；

[0151] 对经过所述降噪处理处理后的每个声道音频进行消除回音的处理；

[0152] 将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签；

[0153] 将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签。

[0154] 显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

[0155] 需要注意的是，本发明可在软件和/或软件与硬件的组合体中被实施，例如，可采用专用集成电路(ASIC)、通用目的计算机或任何其他类似硬件设备来实现。在一个实施例

中，本发明的软件程序可以通过处理器执行以实现上文所述步骤或功能。同样地，本发明的软件程序(包括相关的数据结构)可以被存储到计算机可读记录介质中，例如，RAM存储器，磁或光驱动器或软磁盘及类似设备。另外，本发明的一些步骤或功能可采用硬件来实现，例如，作为与处理器配合从而执行各个步骤或功能的电路。

[0156] 另外，本发明的一部分可被应用为计算机程序产品，例如计算机程序指令，当其被计算机执行时，通过该计算机的操作，可以调用或提供根据本发明的方法和/或技术方案。而调用本发明的方法的程序指令，可能被存储在固定的或可移动的记录介质中，和/或通过广播或其他信号承载媒体中的数据流而被传输，和/或被存储在根据所述程序指令运行的计算机设备的工作存储器中。在此，根据本发明的一个实施例包括一个装置，该装置包括用于存储计算机程序指令的存储器和用于执行程序指令的处理器，其中，当该计算机程序指令被该处理器执行时，触发该装置运行基于前述根据本发明的多个实施例的方法和/或技术方案。

[0157] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附图标记视为限制所涉及的权利要求。此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。装置权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第一，第二等词语用来表示名称，而并不表示任何特定的顺序。

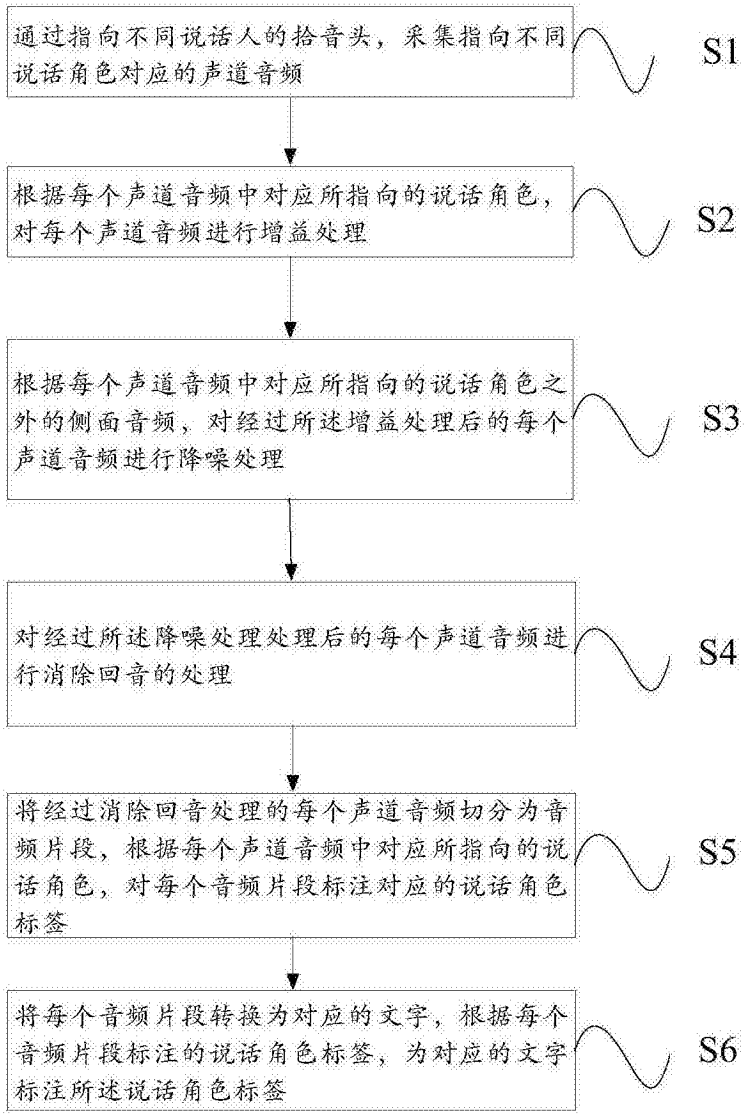

通过指向不同说话人的拾音头，采集指向不同说话角色对应的声道音频
S1
根据每个声道音频中对应所指向的说话角色，对每个声道音频进行增益处理
S2
根据每个声道音频中对应所指向的说话角色之外的侧面音频，对经过所述增益处理后的每个声道音频进行降噪处理
S3
对经过所述降噪处理处理后的每个声道音频进行消除回音的处理
S4
将经过消除回音处理的每个声道音频切分为音频片段，根据每个声道音频中对应所指向的说话角色，对每个音频片段标注对应的说话角色标签
S5
将每个音频片段转换为对应的文字，根据每个音频片段标注的说话角色标签，为对应的文字标注所述说话角色标签
S6

图1

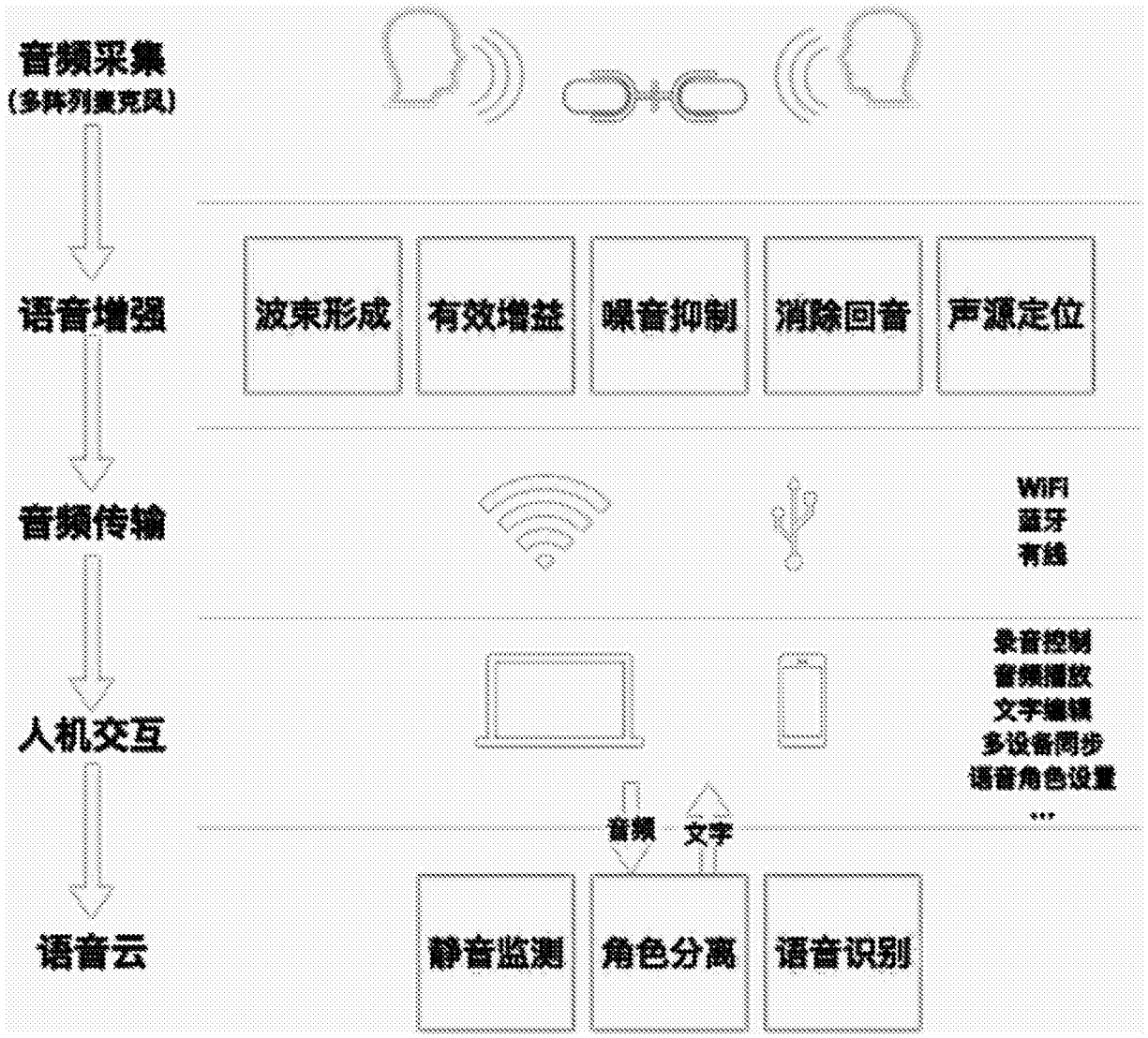

音频采集
(多阵列麦克风)
语音增强
波束形成
有效增益
噪音抑制
消除回音
声源定位
音频传输
WiFi
蓝牙
有线
人机交互
录音控制
音频播放
文字编辑
多设备同步
语音角色设置
...
音频
文字
语音云
静音监测
角色分离
语音识别

图2

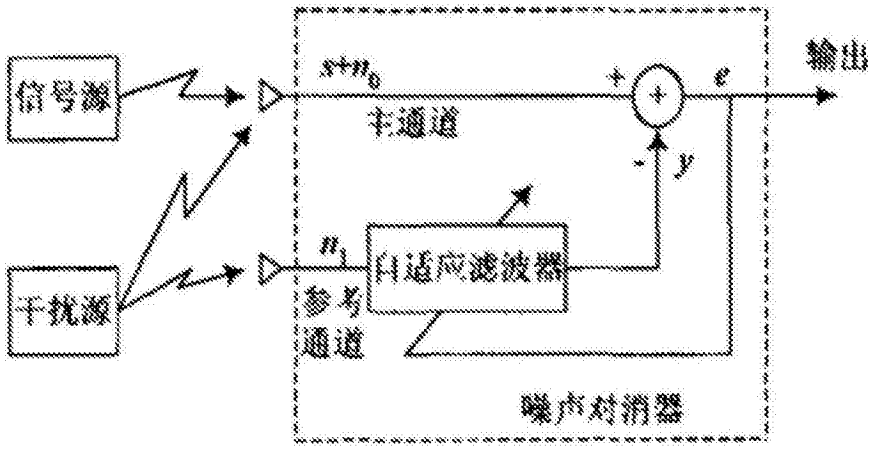

信号源
干扰源
s+n0
主通道
n1
参考
通道
自适应滤波器
+
+
-
e
y
输出
噪声对消器

图3