(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111524527 B
(45)授权公告日 2023.08.22

(21)申请号 202010365591.8

(22)申请日 2020.04.30

(65)同一申请的已公布的文献号
申请公布号 CN 111524527 A

(43)申请公布日 2020.08.11

(73)专利权人 合肥讯飞数码科技有限公司
地址 230088 安徽省合肥市高新区望江西路666号讯飞大厦1805、1807室

(72)发明人 方磊 蒋俊 方四安 柳林 方堃 丁奇

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 程琛

(51)Int.Cl.
G10L 17/06(2013.01)
G10L 17/02(2013.01)
G10L 21/028(2013.01)

(56)对比文件
CN 110299150 A,2019.10.01
EP 2808866 A1,2014.12.03
CN 107393527 A,2017.11.24
US 2016217793 A1,2016.07.28
US 2014074467 A1,2014.03.13
JP H07287592 A,1995.10.31
CN 110491392 A,2019.11.22
CN 110491411 A,2019.11.22
CN 110444223 A,2019.11.12
CN 111063341 A,2020.04.24
US 2018082689 A1,2018.03.22
CN 108766440 A,2018.11.06
CN 106782507 A,2017.05.31
US 5598507 A,1997.01.28
CN 110853666 A,2020.02.28
CN 108074576 A,2018.05.25
CN 106782563 A,2017.05.31
李锐;卓著;李辉.基于BIC和G_PLDA的说话人分离技术研究.中国科学技术大学学报.2015,(04),全文.

审查员 许李铭

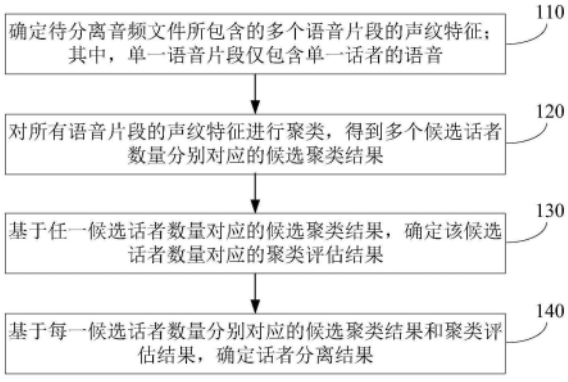
权利要求书2页 说明书11页 附图3页

(54)发明名称

话者分离方法、装置、电子设备和存储介质

(57)摘要

本发明实施例提供一种话者分离方法、装置、电子设备和存储介质，其中方法包括：确定待分离音频文件包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音；对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果；基于任一候选话者数量对应的候选聚类结果，确定该候选话者数量对应的聚类评估结果；基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果。本发明实施例提供的方法、装置、电子设备和存储介质，实现了不确定话者数量情况下的无源话者分割，避免了固定话者数量或通过固定阈值来确定话者数量导致话者数量不符合实际情况，影响无源话者分离准确性的问题。



CN 111524527 B

1.一种话者分离方法,其特征在于,包括:

确定待分离音频文件包含的多个语音片段的声纹特征;其中,单一语音片段仅包含单一话者的语音;

对所有语音片段的声纹特征进行聚类,得到多个候选话者数量分别对应的候选聚类结果,所述候选话者数量为所述待分离音频文件包含的话者数量的候选值,所述候选话者数量与所述候选聚类结果一一对应,所述候选聚类结果中的类别数为对应的候选话者数量;

基于任一候选话者数量对应的候选聚类结果,确定所述任一候选话者数量对应的聚类评估结果,所述聚类评估结果用于表征对应候选聚类结果的质量;

基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果,确定话者分离结果。

2.根据权利要求1所述的话者分离方法,其特征在于,所述确定待分离音频文件包含的多个语音片段的声纹特征,具体包括:

将任一语音片段输入至声纹提取模型,得到所述声纹提取模型输出的所述任一语音片段的声纹特征;其中,所述声纹提取模型用于提取所述任一语音片段的隐层特征,并基于所述隐层特征确定所述任一语音片段的声纹特征。

3.根据权利要求2所述的话者分离方法,其特征在于,所述声纹提取模型是联合话者分类模型和文本识别模型,基于样本语音片段及其对应的话者标签和文本标签训练得到的;

所述话者分类模型用于基于所述声纹提取模型提取的所述样本语音片段的样本声纹特征,对所述样本语音片段进行话者分类,所述文本识别模型用于基于所述声纹提取模型提取的所述样本语音片段的样本隐层特征,对所述样本语音片段进行文本识别。

4.根据权利要求2所述的话者分离方法,其特征在于,所述声纹提取模型是联合语音解码模型和语音增强判别模型,基于干净语音片段和带噪语音片段进行对抗训练得到的;

所述语音解码模型用于将所述声纹提取模型提取的所述带噪语音片段的隐层特征解码为增强语音片段,所述语音增强判别模型用于区分所述干净语音片段和所述增强语音片段。

5.根据权利要求1所述的话者分离方法,其特征在于,所述基于任一候选话者数量对应的候选聚类结果,确定所述任一候选话者数量对应的聚类评估结果,具体包括:

基于任一候选话者数量对应的候选聚类结果,确定每个语音片段的声纹特征分别属于所述候选聚类结果中每个类别的概率;

基于每个语音片段的声纹特征分别属于所述候选聚类结果中每个类别的概率,确定所述候选聚类结果的信息熵值,作为所述任一候选话者数量对应的聚类评估结果。

6.根据权利要求1至5中任一项所述的话者分离方法,其特征在于,所述对所有语音片段的声纹特征进行聚类,得到多个候选话者数量分别对应的候选聚类结果,具体包括:

基于任一语音片段的声纹特征与声纹库中每一库内声纹特征之间的相似度,确定所述任一语音片段的声纹在库状态;

对声纹在库状态为不在库的所有语音片段的声纹特征进行聚类,得到多个候选话者数量分别对应的候选聚类结果。

7.根据权利要求6所述的话者分离方法,其特征在于,所述基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果,确定话者分离结果,之后还包括:

基于所述话者分离结果，更新所述声纹库。

8.一种话者分离装置，其特征在于，包括：

片段声纹提取单元，用于确定待分离音频文件包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音；

片段声纹聚类单元，用于对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果，所述候选话者数量为所述待分离音频文件包含的话者数量的候选值，所述候选话者数量与所述候选聚类结果一一对应，所述候选聚类结果中的类别数为对应的候选话者数量；

聚类参数评估单元，用于基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果，所述聚类评估结果用于表征对应候选聚类结果的质量；

话者分离单元，用于基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1至7中任一项所述的话者分离方法的步骤。

10.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现如权利要求1至7中任一项所述的话者分离方法的步骤。

话者分离方法、装置、电子设备和存储介质

技术领域

[0001] 本发明涉及智能语音技术领域，尤其涉及一种话者分离方法、装置、电子设备和存储介质。

背景技术

[0002] 话者分离是指将一段音频文件中分属于每一话者的音频数据进行分割，将同一话者的音频数据合并成一类，不同话者的音频数据分开，并获得每个话者音频数据的时间位置信息，即解决什么话者在什么时候说的问题。根据事先是否掌握话者信息的情况，话者分离可细分为无源话者分离与有源话者分离。其中，无源话者分离是在事先不知道音频文件所涉及的话者及人数的情况下执行的。

[0003] 目前，对于电话信道获取的音频文件，无源话者分离默认话者人数为两人，并在此基础上将分割的语音片段聚成两类。但对于人数不确定的多人对话场景，无法提前确定聚类的类别数。同时，由于不同话者的风格差异大、聚类片段的时长不固定等因素，很难通过一个统一的门限阈值来自动确定类别数，导致上述无源话者分离技术难以在话者人数不确定的场景下推广应用。

发明内容

[0004] 本发明实施例提供一种话者分离方法、装置、电子设备和存储介质，用以解决话者人数不确定的场景下，无源话者分离技术难以应用的问题。

[0005] 第一方面，本发明实施例提供一种话者分离方法，包括：

[0006] 确定待分离音频文件包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音；

[0007] 对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果；

[0008] 基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果；

[0009] 基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果。

[0010] 优选地，所述确定待分离音频文件包含的多个语音片段的声纹特征，具体包括：

[0011] 将任一语音片段输入至声纹提取模型，得到所述声纹提取模型输出的所述任一语音片段的声纹特征；其中，所述声纹提取模型用于提取所述任一语音片段的隐层特征，并基于所述隐层特征确定所述任一语音片段的声纹特征。

[0012] 优选地，所述声纹提取模型是联合话者分类模型和文本识别模型，基于样本语音片段及其对应的话者标签和文本标签训练得到的；

[0013] 所述话者分类模型用于基于所述声纹提取模型提取的所述样本语音片段的样本声纹特征，对所述样本语音片段进行话者分类，所述文本识别模型用于基于所述声纹提取

模型提取的所述样本语音片段的样本隐层特征，对所述样本语音片段进行文本识别。

[0014] 优选地，所述声纹提取模型是联合语音解码模型和语音增强判别模型，基于干净语音片段和带噪语音片段进行对抗训练得到的；

[0015] 所述语音解码模型用于将所述声纹提取模型提取的所述带噪语音片段的隐层特征解码为增强语音片段，所述语音增强判别模型用于区分所述干净语音片段和所述增强语音片段。

[0016] 优选地，所述基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果，具体包括：

[0017] 基于任一候选话者数量对应的候选聚类结果，确定每个语音片段的声纹特征分别属于所述候选聚类结果中每个类别的概率；

[0018] 基于每个语音片段的声纹特征分别属于所述候选聚类结果中每个类别的概率，确定所述候选聚类结果的信息熵值，作为所述任一候选话者数量对应的聚类评估结果。

[0019] 优选地，所述对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果，具体包括：

[0020] 基于任一语音片段的声纹特征与声纹库中每一库内声纹特征之间的相似度，确定所述任一语音片段的声纹在库状态；

[0021] 对声纹在库状态为不在库的所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果。

[0022] 优选地，所述基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果，之后还包括：

[0023] 基于所述话者分离结果，更新所述声纹库。

[0024] 第二方面，本发明实施例提供一种话者分离装置，包括：

[0025] 片段声纹提取单元，用于确定待分离音频文件包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音；

[0026] 片段声纹聚类单元，用于对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果；

[0027] 聚类参数评估单元，用于基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果；

[0028] 话者分离单元，用于基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果。

[0029] 第三方面，本发明实施例提供一种电子设备，包括处理器、通信接口、存储器和总线，其中，处理器，通信接口，存储器通过总线完成相互间的通信，处理器可以调用存储器中的逻辑命令，以执行如第一方面所提供的方法的步骤。

[0030] 第四方面，本发明实施例提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如第一方面所提供的方法的步骤。

[0031] 本发明实施例提供的一种话者分离方法、装置、电子设备和存储介质，通过多个候选话者数量分别对应的候选聚类结果，分别得到多个候选话者数量的聚类评估结果，并基于此确定话者分离结果，实现了不确定话者数量情况下的无源话者分割，避免了固定话者数量或通过固定阈值来确定话者数量导致话者数量不符合实际情况，影响无源话者分离准

确性的问题,有利于无源话者分离在话者人数不确定的场景下推广应用。

附图说明

[0032] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施例,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。

[0033] 图1为本发明实施例提供的话者分离方法的流程示意图;

[0034] 图2为本发明实施例提供的多任务联合训练示意图;

[0035] 图3为本发明实施例提供的对抗训练示意图;

[0036] 图4为本发明实施例提供的聚类评估结果确定方法的流程示意图;

[0037] 图5为本发明实施例提供的聚类方法的流程示意图;

[0038] 图6为本发明实施例提供的声纹提取模型的训练示意图;

[0039] 图7为本发明实施例提供的话者分离装置的结构示意图;

[0040] 图8为本发明实施例提供的电子设备的结构示意图。

具体实施方式

[0041] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0042] 目前无源话者分离技术主要应用于电话信道声纹识别的前端处理,具体通过分割与聚类两个阶段实现,详细的实现步骤包括:将待分离的音频文件分割为多个语音片段,目标是使得每个语音片段中均只包含一个话者的语音;随即,对多个语音片段进行聚类,直至聚成两类为止。在此过程中,待分割的音频文件中是否只包含两个话者的语音,会直接影响话者分离的效果,如果待分割的音频文件中只包含一个话者的语音,上述方法也会强制将一个话者的语音切割两个部分,如果待分割的音频文件中包含多于两个话者的语音,那么聚类纯度会严重受损。由此可见,在不确定话者数量的情况下,如何实现准确的无源话者分割,仍然是话者分割领域亟待解决的问题。

[0043] 对此,本发明实施例提供了一种话者分离方法。图1为本发明实施例提供的话者分离方法的流程示意图,如图1所示,该方法包括:

[0044] 步骤110,确定待分离音频文件所包含的多个语音片段的声纹特征;其中,单一语音片段仅包含单一话者的语音。

[0045] 此处,待分离音频文件即需要进行话者分离的音频文件,待分离音频文件可以包含多个语音片段。在同一时刻仅一个话者发言,不存在多个话者同时发言的场景下,后一话者在前一话者发言结束后发言,两段发言之间存在间隔,待分离音频文件中包含的多个语音片段可以通过语音端点检测(Voice Activity Detection,VAD)进行分割得到。又或者,可以基于BIC(Bayesian Information Criterion,贝叶斯信息准则)对待分离音频文件进行话者变化点检测,依据检测的结果进行音频分割,得到多个语音片段。

[0046] 在得到多个语音片段后,可以分别获取每个语音片段的声纹特征。任一语音片段的声纹特征具体是指该语音片段中话者所体现的声音特征。语音片段的声纹特征可以通过将语音片段输入到预先训练好的声纹特征提取模型得到。

[0047] 步骤120,对所有语音片段的声纹特征进行聚类,得到多个候选话者数量分别对应的候选聚类结果。

[0048] 具体地,候选话者数量有多个,可以是预先设定的待分离音频文件中可能包含的话者数量,候选话者数量的设定可以与待分离音频文件的获取场景相关联,例如飞航过程中飞行员和相关人员进行通话的场景下,话者数量可能在3至6人之间,对应的候选话者数量分别为3、4、5、6;又例如待分离音频文件是在小型会客室录制的,小型会客室设置有4个座位,则话者数量可能在2至4人之间,对应的候选话者数量分别为2、3、4。

[0049] 经过步骤110得到所有语音片段的声纹特征后,可以对所有声纹特征进行聚类,此处应用的聚类算法可以是EM算法(Expectation-maximization algorithm,最大期望值算法),也可以是K-Means(K均值)聚类算法或层次聚类算法等,本发明实施例对此不作具体限定。需要说明的是,通过聚类所得到的并不是传统聚类算法最终输出的唯一的聚类结果,而是分别对应于多个候选话者数量的多个候选聚类结果。此处,每个候选话者数量均对应一个候选聚类结果,候选聚类结果中的类别数即对应的候选话者数量。例如,候选话者数量为3时,对应的候选聚类结果中包含3个类别,候选话者数量为4时,对应的候选聚类结果中包含4个类别。

[0050] 步骤130,基于任一候选话者数量对应的候选聚类结果,确定该候选话者数量对应的聚类评估结果。

[0051] 具体地,聚类评估结果即对候选话者数量的候选聚类结果进行评估的得到的评估结果,聚类评估结果用于表征对应候选聚类结果的质量,具体可以表示为候选聚类结果中各个类别的类内聚集程度、类间离散程度等,还可以表示为候选聚类结果可能发生的概率,本发明实施例对此不做具体限定。

[0052] 步骤140,基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果,确定话者分离结果。

[0053] 具体地,在得到每一候选话者数量的聚类评估结果后,可以基于每一候选话者数量的聚类评估结果,对每一候选话者数量分别对应的候选聚类结果的质量进行比对,进而从中挑选出聚类评估结果最优的候选聚类结果,将聚类评估结果最优的候选聚类结果作为待分离音频文件的话者分离结果,将其对应的候选话者数量作为待分离音频文件中实际包含的话者数量。

[0054] 进一步地,在基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果确定话者分离结果时,针对任一聚类评估结果,候选聚类结果中各个类别的类内聚集程度和类间离散程度越高,则候选聚类结果的质量越高,越可能被选为话者分离结果;候选聚类结果可能发生的概率越高,则候选聚类结果的质量越高,越可能被选为话者分离结果。

[0055] 本发明实施例提供的方法,通过多个候选话者数量分别对应的候选聚类结果,分别得到多个候选话者数量的聚类评估结果,并基于此确定话者分离结果,实现了不确定话者数量情况下的无源话者分割,避免了固定话者数量或通过固定阈值来确定话者数量导致话者数量不符合实际情况,影响无源话者分离准确性的问题,有利于无源话者分离在话者

人数不确定的场景下推广应用。

[0056] 基于上述实施例，步骤110具体包括：将任一语音片段输入至声纹提取模型，得到声纹提取模型输出的该语音片段的声纹特征；其中，声纹提取模型用于提取该语音片段的隐层特征，并基于隐层特征确定该语音片段的声纹特征。

[0057] 具体地，可以将待分离音频文件中的任一语音片段输入到预先训练好的声纹提取模型中，由声纹提取模型对该语音片段进行编码并提取该语音片段编码后的隐层特征，并在此基础上，对该语音片段的隐层特征进行声纹特征提取，输出该语音片段的声纹特征。

[0058] 进一步地，声纹提取模型可以包括隐层特征提取层和声纹特征提取层；其中，隐层特征提取层用于对输入的语音片段进行编码并提取该语音片段编码后的隐层特征，声纹特征提取层用于对隐层特征提取层输出的隐层特征进行声纹特征提取，并输出声纹特征。

[0059] 在执行步骤110之前，还可以预先训练得到声纹提取模型，例如可以通过如下方法训练得到声纹提取模型：首先，收集大量样本语音片段及其对应的样本声纹特征，应用样本语音片段和样本声纹特征对初始模型进行训练，从而得到声纹提取模型。

[0060] 考虑到在一些特定的场景，例如飞航过程中飞行员和相关人员进行通话的场景，主题明确的会议讨论场景等，待分离音频文件中包含语音所对应的文本内容实际上十分有限，多为行业术语，其中存在相同文本的概率较高，且文本内容可以形成相对稳定的闭集。

[0061] 基于上述任一实施例，图2为本发明实施例提供的多任务联合训练示意图，如图2所示，声纹提取模型是联合话者分类模型和文本识别模型，基于样本语音片段及其对应的话者标签和文本标签训练得到的；话者分类模型用于基于声纹提取模型提取的样本语音片段的样本声纹特征，对样本语音片段进行话者分类，文本识别模型用于基于声纹提取模型提取的样本语音片段的样本隐层特征，对样本语音片段进行文本识别。

[0062] 具体地，训练过程中，将样本语音片段输入至声纹提取模型，由声纹提取模型对样本语音片段进行编码，提取样本语音片段编码后的样本隐层特征，并对样本隐层特征进行声纹特征提取，输出样本语音片段的样本声纹特征。

[0063] 将声纹提取模型输出的样本声纹特征输入至话者分类模型，由话者分类模型预测样本声纹特征所对应的话者身份并输出。此外，将声纹提取模型中间产生的样本隐层特征输入至文本识别模型，由文本识别模型基于样本隐层特征对样本语音片段进行文本识别，并输出识别文本。

[0064] 在得到话者分类模型输出的话者身份和文本识别模型输出的识别文本后，可以将话者身份和识别文本分别与样本语音片段对应的话者标签和文本标签进行比对，从而更新声纹提取模型、话者分类模型与文本识别模型的模型参数，实现针对声纹提取模型的多目标训练。

[0065] 参考图2示出的模型结构，话者分类模型和文本识别模型在分别进行话者分类和文本识别时，共用了声纹提取模型中用于提取隐层特征的部分，即图2中的隐层特征提取层，隐层特征提取层的共用使得多目标训练过程中话者分类模型和文本识别模型能够实现信息共享，从而充分利用特定场景下待分离音频文件中包含语音所对应的文本内容相对固定的优势，使得声纹提取模型能够更好地区分相同文本内容下不同话者的音频片段所表征的声纹特征，提高声纹提取模型输出声纹特征的准确性。

[0066] 本发明实施例提供的方法，联合话者分类模型和文本识别模型实现声纹提取模型

的多目标训练，优化声纹提取模型针对相同文本内容的不同话者声纹特征的区分性，从而提高输出声纹特征的可靠性，进而实现准确可靠的话者分离。

[0067] 待分离音频文件中可能包含大量的环境噪声，如果不进行降噪处理，基于语音片段提取的声纹特征中必然也会包含噪声带来的影响，严重影响话者分离的聚类纯度。针对这一问题，基于上述任一实施例，图3为本发明实施例提供的对抗训练示意图，如图3所示，声纹提取模型是联合语音解码模型和语音增强判别模型，基于干净语音片段和带噪语音片段进行对抗训练得到的；语音解码模型用于将声纹提取模型提取的带噪语音片段的隐层特征解码为增强语音片段，语音增强判别模型用于区分干净语音片段和增强语音片段。

[0068] 具体地，可以预先收集干净语音片段和带噪语音片段。此处，干净语音片段是指不包含环境噪声的语音片段，带噪语音片段即包含环境噪声的语音片段，带噪语音片段可以通过对干净语音片段进行加噪处理得到。

[0069] 在训练过程中，将带噪语音片段输入至声纹提取模型，由声纹提取模型对带噪语音片段进行编码，提取带噪语音片段编码后的样本隐层特征。随即将带噪语音片段对应的样本隐层特征输入至语音解码模型，由语音解码模型对样本隐层特征进行解码还原，得到并输出带噪语音片段对应的增强语音片段。然后将增强语音片段输入至语音增强判别模型，由语音增强判别模型判别输入的语音片段是干净语音片段还是增强语音片段。

[0070] 将声纹提取模型联合语音解码模型和语音增强判别模型进行对抗训练的目的，在于通过声纹提取模型和语音解码模型得到的增强语音片段能够无限接近真实的干净语音片段，使得语音增强判别模型无法区分输入的语音片段是真实的干净语音片段，还是经过声纹提取模型和语音解码模型得到的增强语音片段。经过对抗训练后的声纹提取模型中用于提取隐层特征的部分，即图3示出的隐层特征提取层，具备在提取隐层特征的同时尽量滤除语音片段中包含的环境噪声的能力。

[0071] 本发明实施例提供的方法，通过对抗训练在实现声纹提取功能的同时，实现了语音增强功能，从而保证声纹提取模型在进行语音片段的声纹提取时，能够有效抑制语音片段中裹挟的环境噪声干扰，提高输出声纹特征的准确性，从而实现准确可靠的话者分离。

[0072] 基于上述任一实施例，图4为本发明实施例提供的聚类评估结果确定方法的流程示意图，如图4所示，步骤130具体包括：

[0073] 步骤131，基于任一候选话者数量对应的候选聚类结果，确定每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率。

[0074] 具体地，任一候选话者数量对应的候选聚类结果包括该候选话者数量个类别。在得到候选聚类结果后，可以计算每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率。

[0075] 例如，候选话者数量为3，对应的候选聚类结果包括3个类别，分别表示为c1、c2和c3，假设待分离音频文件共包含n个语音片段，则其中第i个语音片段的声纹特征属于3个类别的概率可以表示为$pi_{cm3}=(pi_{c1},pi_{c2},pi_{c3})'$，式中$pi_{c1}$、$pi_{c2}$和$pi_{c3}$分别为第i个语音片段的声纹特征属于类别c1、c2和c3的概率。

[0076] 在此基础上，可以得到每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率，具体表示为$P3=\{p1_{cm3},p2_{cm3},...,pi_{cm3},...,pn_{cm3}\}_{n*3}$。

[0077] 步骤132，基于每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概

率，确定候选聚类结果的信息熵值，作为该候选话者数量对应的聚类评估结果。

[0078] 具体地，在得到每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率后，即可计算该候选聚类结果的信息熵值。此处，候选聚类结果的信息熵值可以反映该候选聚类结果的发生概率，信息熵值越小，则该候选聚类结果发生的概率越大，该候选聚类结果越稳定。

[0079] 本发明实施例提供的方法，将候选聚类结果的信息熵值作为候选话者数量对应的聚类评估结果用于确定话者数量和话者分离，从而解决了待分离音频文件所包含的话者数量不确定的问题，有利于无源话者分离在话者人数不确定的场景下推广应用。

[0080] 基于上述任一实施例，步骤140具体包括：将最小信息熵值对应的候选话者数量作为所述最终话者数量。

[0081] 具体地，将候选聚类结果的信息熵值作为对应候选话者数量的聚类评估结果后，在确定话者分离结果时，仅需要比较各个候选话者数量对应的信息熵值大小，可以将信息熵值最小的候选话者数量作为最终的话者数量。此处，信息熵值最小的候选话者数量，其对应的候选聚类结果即多个候选聚类结果中最稳定、发生概率最高的聚类结果，由此可以确定话者分离结果。

[0082] 此外，还可以在得到多个候选话者数量的信息熵值的最小值之后，还可以将最小值与预先设定的信息熵值阈值进行比较，如果最小值小于信息熵值阈值，则将最小值对应的候选话者数量作为最终的话者数量，将最小值对应的候选聚类结果作为话者分离结果；如果最小值大于信息熵值阈值，则确认每个候选话者数量均不是最终的话者数量，可以重新设置候选话者数量可以聚类。

[0083] 基于上述任一实施例，步骤120中的声纹特征聚类可以通过EM算法实现的，对n个音频片段的声纹特征x_i进行无监督聚类，对应得到聚类的结果为一个高斯混合模型$P(\lambda)=\sum_j^m w_j * N(\mu_j,\Sigma_j)$，其中m为任一候选话者数量，即任一候选聚类结果中的类别数，j为小于等于m的正整数，j表示候选聚类结果中的类别序号，w_j为第j个类别在高斯混合模型中的权重，$N(\mu_j,\Sigma_j)$为第j个类别的高斯模型。例如，候选话者数量可以是3、4、5、6，则对应地m的取值可以是3、4、5、6。

[0084] 候选话者数量m＝3时，可以通过如下公式计算第i个音频片段的声纹特征x_i属于3个类别中任一类别中心λ_c的高斯占用率，作为x_i属于3个类别中任一类别的概率$p(x_i|\lambda_c)$：

[0085] $$p(x_i|\lambda_c)=\frac{w_c N(x_i;\lambda_c)}{\sum_{j=1}^{m} w_j N_j(x_i;\lambda_j)}$$

[0086] 例如，可以通过如下公式计算x_i属于3个类别中第1个类别c1的概率pi_{c1}：

[0087] $$pi_{c1}=\frac{w_{c1} N(x_i;\lambda_{c1})}{\sum_{j=1}^{3} w_j N_j(x_i;\lambda_j)}$$

[0088] 式中，λ_{c1}为类别c1的中心，λ_j为第j个类别的类别中心。

[0089] 通过上述公式即可得到m＝3时n个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率$P3=\{p1_{cm3},p2_{cm3},\ldots,pi_{cm3},\ldots,pn_{cm3}\}_{n*3}$，其中$pi_{cm3}$为$x_i$分别属于候选结果中每个类别的概率，$pi_{cm3}=(pi_{c1},pi_{c2},pi_{c3})'$。

[0090] 在此基础上，可以将P3代入信息熵值公式，从而得到m＝3时的信息熵值作为候选

话者数量为3时的聚类评估结果，信息熵值公式如下：

[0091] $$E = -\sum_{i=1}^{n}\sum_{c=1}^{m} p(x_i|\lambda_c)\, log(\, p(x_i|\lambda_c)$$

[0092] 式中，$p(x_i|\lambda_c)$即x_i属于任一类别的概率。

[0093] 由此可得m＝3时的信息熵值公式具体为：

[0094] $$E_{cm3} = -\sum_{i=1}^{n} pi_{c1} * log(pi_{c1}) + pi_{c2} * log(pi_{c2}) + pi_{c3} * log(pi_{c3})$$

[0095] 式中，E_{cm3}即候选话者数量为3时的信息熵值，$pi_{c1}*\log(pi_{c1})$、$pi_{c2}*\log(pi_{c2})$和$pi_{c3}*\log(pi_{c3})$分别为x_i对应于三个类别的信息熵，E_{cm3}即每个语音片段的声纹特征分别对应三个类别的信息熵的总和。

[0096] 基于上述任一实施例，图5为本发明实施例提供的聚类方法的流程示意图，如图5所示，步骤120具体包括：

[0097] 步骤121，基于任一语音片段的声纹特征与声纹库中每一库内声纹特征之间的相似度，确定该语音片段的声纹在库状态。

[0098] 具体地，在得到任一语音片段的声纹特征之后，可以将该语音片段的声纹特征和声纹库中已有的声纹特征，即库内声纹特征进行匹配。

[0099] 在将该语音片段的声纹特征和库内声纹特征进行匹配时，可以计算得到该语音片段的声纹特征和每一库内声纹特征之间的相似度，若该语音片段的声纹特征和任一库内声纹特征之间相似度大于等于预先设定的相似度阈值，则确定该语音片段的声纹特征与该库内声纹特征属于同一话者；若语音片段的声纹特征和每一库内声纹特征之间相似度均小于相似度阈值，则确定该语音片段的声纹特征不同于任何库内声纹特征。此处，语音片段的声纹在库状态可以是在库或不在库。

[0100] 步骤122，对声纹在库状态为不在库的所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果。

[0101] 具体地，根据步骤121的判断，若已经确定任一语音片段属于库内的已有的声纹特征，则无需再对该语音片段进行聚类。步骤122中，仅对于声纹在库状态为不在库，即不属于库内的已存储的声纹特征所属话者的声纹特征进行聚类，从而减小聚类的语音片段数，提高聚类精度，避免聚类后形成的新的话者对应的声纹特征与声纹库中的已知话者重合造成数据混乱。

[0102] 基于上述任一实施例，步骤140之后还包括：基于所述话者分离结果，更新所述声纹库。

[0103] 具体地，在得到话者分离结果后，将话者分离结果中不同类别的声纹特征分别存储到声纹库中，从而不断充实声纹库，以减小话者分离中的不确定性，逐渐将无源话者分离问题转化为有源话者分离问题，以降低话者分离的解决难度，从而实现更加高效、准确的话者分离。

[0104] 基于上述任一实施例，一种话者分离方法，包括如下步骤：

[0105] 确定待分离音频文件，此处的待分离音频文件包括1个语音片段，其中每个语音片

段的持续时长为0.5-3秒，只含一个话者的语音，信噪比低，语音对应的文本内容相对固定，且片段之间存在长度不一的时间间隔。

[0106] 首先，采用VAD算法剔除上述l个语音片段中的噪声数据，得到l′个纯人声的语音片段。

[0107] 其次，利用预先训练好的声纹提取模型，将l′个语音片段映射成l′个512维的声纹特征向量集，此处记为j-vector向量集$X=\{x_1,x_2,...x_i,...,x_{l'}\}$。图6为本发明实施例提供的声纹提取模型的训练示意图，如图6所示，针对场景特点，此处的声纹提取模型是多目标学习优化后的模型，同时考虑了语音增强、文本识别及声纹识别三个目标，充分抑制了噪声干扰，并利用文本信息，使得声纹识别任务中的声纹特征提取层的输出向量能够更好的表征场景所述的声纹信息，有利于后续的说话人无监督聚类。

[0108] 图6中，声纹提取模型的隐层特征提取层与语音解码模型相结合以实现带噪音频片段的自动编码和语音增强，隐层特征提取层与语音解码模型作为生成器，与作为判别器的语音增强判别模型构成生成对抗网络，其目的在于通过声纹提取模型和语音解码模型得到的增强语音片段能够无限接近真实的干净语音片段，使得语音增强判别模型无法区分输入的语音片段是真实的干净语音片段，还是经过声纹提取模型和语音解码模型得到的增强语音片段，从而抑制噪声干扰。同时，话者分类模型和文本识别模型在分别进行话者分类和文本识别时，共用声纹提取模型中的隐层特征提取层，使得话者分类模型和文本识别模型能够实现信息共享，从而充分利用特定场景下待分离音频文件中包含语音所对应的文本内容相对固定的优势，使得声纹提取模型能够更好地区分相同文本内容下不同话者的音频片段所表征的声纹特征。

[0109] 随后，将声纹特征向量集中的l′个声纹特征分别与声纹库中的已有声纹进行比对，剔除相似度超过相似度阈值的语音片段对应的声纹特征，减小聚类的语音片段数，从而提高聚类精度。剔除超过相似度阈值后，得到n个声纹特征，记为$X'=\{x_1,x_2,...x_i,...,x_n\}$。

[0110] 接着，利用EM算法，对X′进行无监督聚类，分别计算多个候选话者数量下n个语音片段的声纹特征在不同类别中的概率，通过信息熵公式进一步计算得到不同候选话者数量分别对应的信息熵值，以熵值最小者为最终话者数量，将最终话者数量下的聚类结果作为话者分离结果。

[0111] 基于上述任一实施例，图7为本发明实施例提供的话者分离装置的结构示意图，如图7所示，话者分离装置包括片段声纹提取单元710、片段声纹聚类单元720、聚类参数评估单元730和话者分离单元740；

[0112] 片段声纹提取单元710用于确定待分离音频文件包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音；

[0113] 片段声纹聚类单元720用于对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果；

[0114] 聚类参数评估单元730用于基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果；

[0115] 话者分离单元740用于基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果。

[0116] 本发明实施例提供的装置，通过多个候选话者数量分别对应的候选聚类结果，分别得到多个候选话者数量的聚类评估结果，并基于此确定话者分离结果，实现了不确定话者数量情况下的无源话者分割，避免了固定话者数量或通过固定阈值来确定话者数量导致话者数量不符合实际情况，影响无源话者分离准确性的问题，有利于无源话者分离在话者人数不确定的场景下推广应用。

[0117] 基于上述任一实施例，片段声纹提取单元710具体用于：

[0118] 将任一语音片段输入至声纹提取模型，得到所述声纹提取模型输出的所述任一语音片段的声纹特征；其中，所述声纹提取模型用于提取所述任一语音片段的隐层特征，并基于所述隐层特征确定所述任一语音片段的声纹特征。

[0119] 基于上述任一实施例，所述声纹提取模型是联合话者分类模型和文本识别模型，基于样本语音片段及其对应的话者标签和文本标签训练得到的；

[0120] 所述话者分类模型用于基于所述声纹提取模型提取的所述样本语音片段的样本声纹特征，对所述样本语音片段进行话者分类，所述文本识别模型用于基于所述声纹提取模型提取的所述样本语音片段的样本隐层特征，对所述样本语音片段进行文本识别。

[0121] 基于上述任一实施例，所述声纹提取模型是联合语音解码模型和语音增强判别模型，基于干净语音片段和带噪语音片段进行对抗训练得到的；

[0122] 所述语音解码模型用于将所述声纹提取模型提取的所述带噪语音片段的隐层特征解码为增强语音片段，所述语音增强判别模型用于区分所述干净语音片段和所述增强语音片段。

[0123] 基于上述任一实施例，聚类参数评估单元730具体用于：

[0124] 基于任一候选话者数量对应的候选聚类结果，确定每个语音片段的声纹特征分别属于所述候选聚类结果中每个类别的概率；

[0125] 基于每个语音片段的声纹特征分别属于所述候选聚类结果中每个类别的概率，确定所述候选聚类结果的信息熵值，作为所述任一候选话者数量对应的聚类评估结果。

[0126] 基于上述任一实施例，片段声纹聚类单元720具体用于：

[0127] 基于任一语音片段的声纹特征与声纹库中每一库内声纹特征之间的相似度，确定所述任一语音片段的声纹在库状态；

[0128] 对声纹在库状态为不在库的所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果。

[0129] 基于上述任一实施例，该装置还包括声纹库更新单元，所述声纹库更新单元用于基于所述话者分离结果，更新所述声纹库。

[0130] 图8为本发明实施例提供的电子设备的结构示意图，如图8所示，该电子设备可以包括：处理器(processor)810、通信接口(Communications Interface)820、存储器(memory)830和通信总线840，其中，处理器810，通信接口820，存储器830通过通信总线840完成相互间的通信。处理器810可以调用存储器830中的逻辑命令，以执行如下方法：确定待分离音频文件包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音；对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果；基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果；基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分

离结果。

[0131] 此外，上述的存储器830中的逻辑命令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案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一个存储介质中，包括若干命令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0132] 本发明实施例还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现以执行上述各实施例提供的方法，例如包括:确定待分离音频文件包含的多个语音片段的声纹特征;其中，单一语音片段仅包含单一话者的语音;对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果;基于任一候选话者数量对应的候选聚类结果，确定所述任一候选话者数量对应的聚类评估结果;基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果。

[0133] 以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

[0134] 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各实施方式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。基于这样的理解，上述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，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干命令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0135] 最后应说明的是:以上实施例仅用以说明本发明的技术方案，而非对其限制;尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换;而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

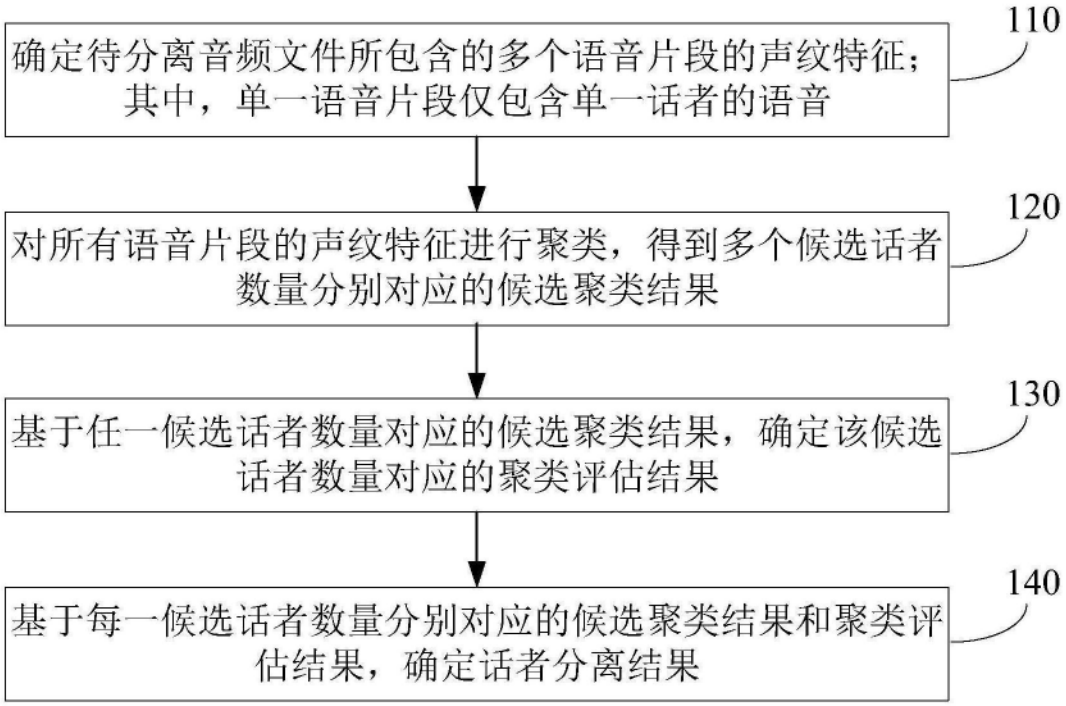

确定待分离音频文件所包含的多个语音片段的声纹特征；其中，单一语音片段仅包含单一话者的语音
110
对所有语音片段的声纹特征进行聚类，得到多个候选话者数量分别对应的候选聚类结果
120
基于任一候选话者数量对应的候选聚类结果，确定该候选话者数量对应的聚类评估结果
130
基于每一候选话者数量分别对应的候选聚类结果和聚类评估结果，确定话者分离结果
140

图1

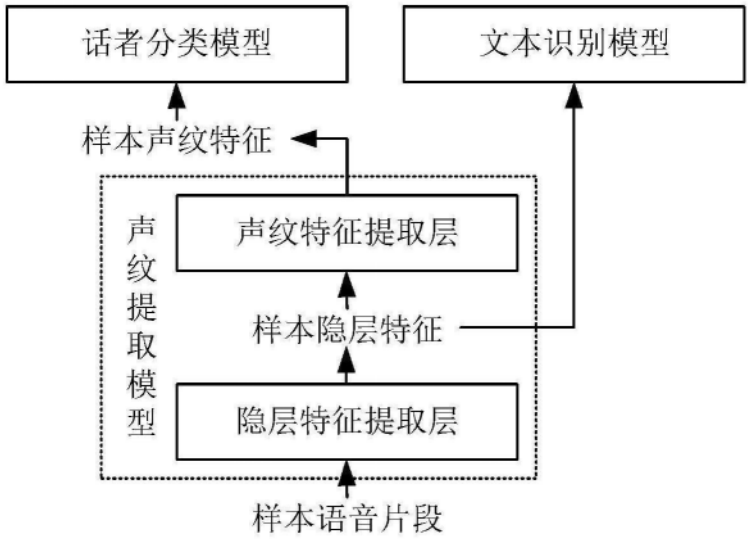

话者分类模型
文本识别模型
样本声纹特征
声纹提取模型
声纹特征提取层
样本隐层特征
隐层特征提取层
样本语音片段

图2

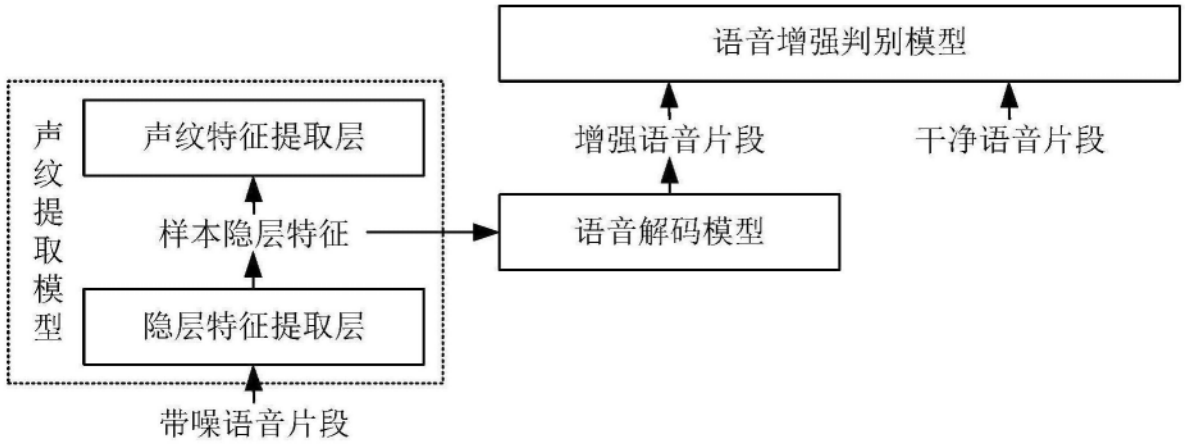

语音增强判别模型
增强语音片段
干净语音片段
语音解码模型
声纹提取模型
声纹特征提取层
样本隐层特征
隐层特征提取层
带噪语音片段

图3

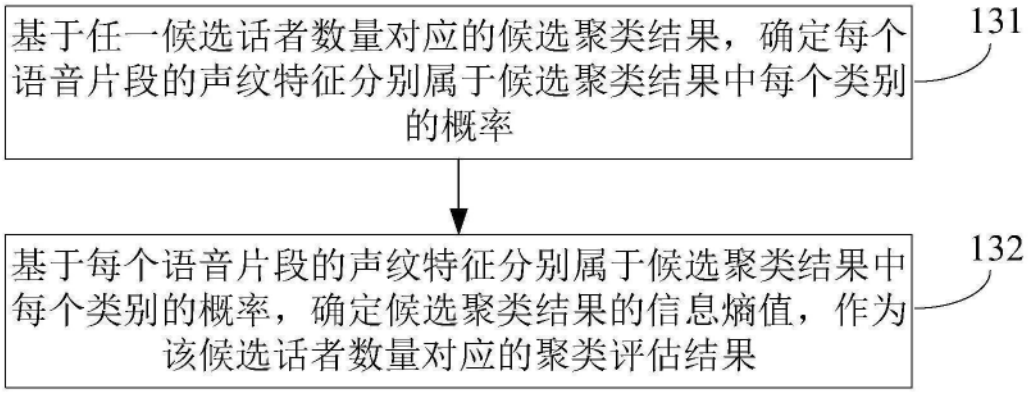

基于任一候选话者数量对应的候选聚类结果，确定每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率
131
基于每个语音片段的声纹特征分别属于候选聚类结果中每个类别的概率，确定候选聚类结果的信息熵值，作为该候选话者数量对应的聚类评估结果
132

图4

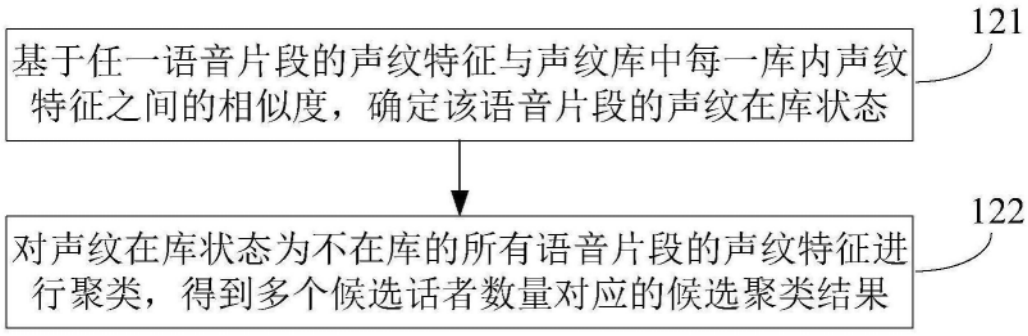

基于任一语音片段的声纹特征与声纹库中每一库内声纹特征之间的相似度，确定该语音片段的声纹在库状态
121
对声纹在库状态为不在库的所有语音片段的声纹特征进行聚类，得到多个候选话者数量对应的候选聚类结果
122

图5

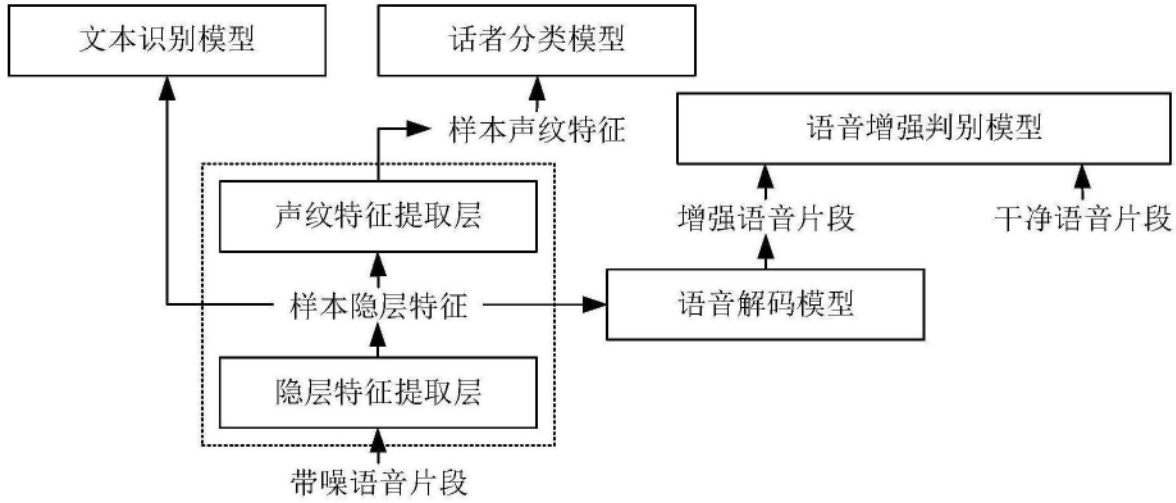

文本识别模型
话者分类模型
样本声纹特征
语音增强判别模型
声纹特征提取层
增强语音片段
干净语音片段
样本隐层特征
语音解码模型
隐层特征提取层
带噪语音片段

图6

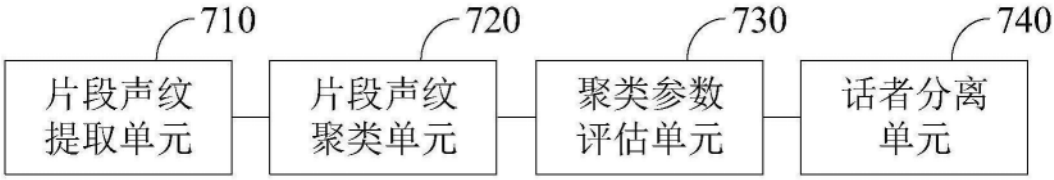

710
720
730
740
片段声纹提取单元
片段声纹聚类单元
聚类参数评估单元
话者分离单元

图7

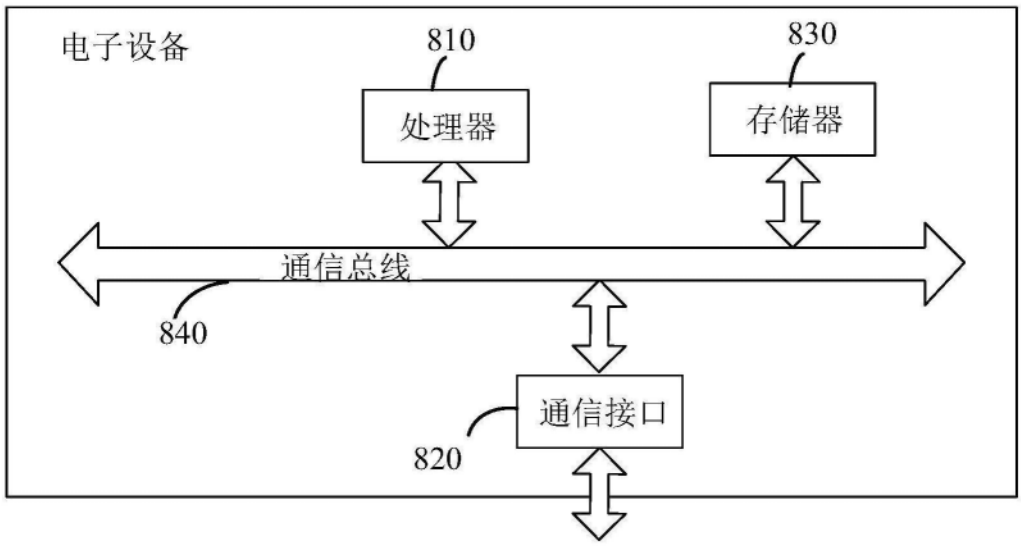

电子设备
810
处理器
830
存储器
通信总线
840
通信接口
820

图8