(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117157708 A
(43)申请公布日 2023.12.01

(21)申请号 202280027518.7

(22)申请日 2022.04.27

(30)优先权数据
63/182,338 2021.04.30 US

(85)PCT国际申请进入国家阶段日
2023.10.09

(86)PCT国际申请的申请数据
PCT/US2022/026552 2022.04.27

(87)PCT国际申请的公布数据
WO2022/232284 EN 2022.11.03

(71)申请人 杜比实验室特许公司
地址 美国加利福尼亚州

(72)发明人 A·法内里 郧铭清 S·S·潘基 N·L·恩格尔 P·A·C·克鲁姆

(74)专利代理机构 北京市汉坤律师事务所 11602
专利代理师 魏小薇 吴丽丽

(51)Int.Cl.
G10L 25/84(2006.01)

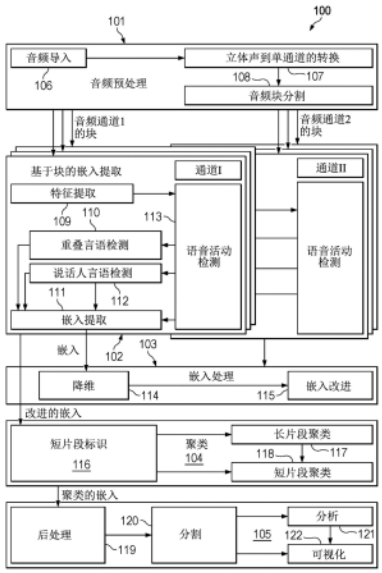
权利要求书3页 说明书15页 附图14页

(54)发明名称

支持插话式内容的说话人日志

(57)摘要

公开了用于支持插话式内容的说话人日志的实施例。在实施例中,一种方法包括:接收包括一个或多个话语的媒体数据;将媒体数据分成多个块;对多个块中与单个说话人相关联的每个块的片段进行标识;根据机器学习模型提取所标识片段的嵌入,其中,提取所标识片段的嵌入进一步包括统计地组合所提取的对应于与单个说话人相关联的相应连续话语的所标识片段的嵌入;将所标识片段的嵌入聚类成簇;以及根据聚类的结果为所标识片段的嵌入中的每个嵌入分配说话人标签。在一些实施例中,使用声纹来标识说话人标签对应的说话人和说话人身份。



CN 117157708 A

1.一种方法,包括:

利用至少一个处理器接收包括一个或多个话语的媒体数据;

利用所述至少一个处理器将所述媒体数据分成多个块;

利用所述至少一个处理器对所述多个块中与单个说话人相关联的每个块的片段进行标识;

利用所述至少一个处理器根据机器学习模型提取所标识片段的嵌入,其中,提取所标识片段的嵌入进一步包括统计地组合所提取的对应于与单个说话人相关联的相应连续话语的所标识片段的嵌入;

利用所述至少一个处理器将所标识片段的所述嵌入聚类成簇;

利用所述至少一个处理器根据所述聚类的结果为所标识片段的所述嵌入中的至少一个嵌入分配说话人标签;以及

利用所述至少一个处理器部分地基于所述说话人标签输出与所述媒体数据相关联的说话人日志信息。

2.如权利要求1所述的方法,进一步包括:

在将所述媒体数据分成多个块之前,利用所述至少一个处理器对所述媒体数据执行空间转换。

3.如权利要求2所述的方法,其中,对所述媒体数据执行所述空间转换包括:

将所述媒体数据的第一多个通道转换成与所述第一多个通道不同的第二多个通道;并且

将所述媒体数据分成多个块包括独立地将所述第二多个通道中的每个通道分成块。

4.如权利要求1至3中任一项所述的方法,进一步包括:

根据确定所述媒体数据对应于第一媒体类型,由第一训练数据集生成所述机器学习模型;以及

根据确定所述第一媒体数据对应于与所述第一媒体类型不同的第二媒体类型,由与所述第一训练数据集不同的第二训练数据集生成所述机器学习模型。

5.如权利要求1至4中任一项所述的方法,进一步包括:

在聚类之前并且根据确定满足优化标准,进一步优化所提取的所标识片段的嵌入。

6.如权利要求5所述的方法,进一步包括:

在聚类之前并且根据确定不满足优化标准,放弃进一步优化所提取的所标识片段的嵌入。

7.如权利要求5至6中任一项所述的方法,其中,优化所提取的所标识片段的嵌入包括:执行对所提取的嵌入进行降维或对所提取的嵌入进行嵌入优化中的至少一者。

8.如权利要求7所述的方法,其中,嵌入优化包括:

训练所述机器学习模型以使所提取的所标识片段的嵌入之间的可分离性最大化;以及

通过将用于使所提取的所标识片段的嵌入之间的所述可分离性最大化的所述机器学习模型应用于所提取的所标识片段的嵌入来更新所提取的嵌入。

9.如权利要求1至8中任一项所述的方法,其中,所述聚类包括:

对于每个标识的片段:

确定所述片段的相应长度;

根据确定所述片段的相应长度大于阈值长度,根据第一聚类过程分配与相应标识的片段相关联的嵌入;以及

根据确定所述片段的相应长度不大于阈值长度,根据与所述第一聚类过程不同的第二聚类过程分配与所述相应标识的片段相关联的嵌入。

10.如权利要求1至9中任一项所述的方法,进一步包括:

部分地基于确定与所述媒体数据相关联的不同说话人的数量从多个聚类过程中选择第一聚类过程。

11.如权利要求1至10中任一项所述的方法,其中,所述第一聚类过程包括谱聚类。

12.如权利要求1至10中任一项所述的方法,其中,所述媒体数据包括多个相关文件。

13.如权利要求12所述的方法,进一步包括:

将多个所述相关文件选择作为所述媒体数据,其中,选择所述多个相关文件是部分地基于以下各项中的至少一项:

与所述多个相关文件相关联的内容相似度;

与所述多个相关文件相关联的元数据相似度;或者

与处理特定一组文件的请求相对应的接收数据。

14.如权利要求12至13中任一项所述的方法,其中,所述机器学习模型是根据由多个相关音频文件中的每个文件共享的一个或多个属性而从多个机器学习模型中选择的。

15.如权利要求1所述的方法,进一步包括:

计算声纹嵌入与每个簇的参考点之间的声纹距离度量;

计算从每个参考点到属于所述簇的每个嵌入的距离;

针对每个簇计算所述嵌入与所述簇的所述参考点的距离的概率分布;

针对每个概率分布,计算所述声纹距离属于所述概率分布的概率;.

对所述概率进行排序;

基于所述排序为所述簇中的一个簇分配声纹;以及

将与所述声纹相关联的说话人身份与所述说话人日志信息进行组合。

16.如权利要求15所述的方法,其中,所述概率分布被建模为折叠高斯分布。

17.如权利要求15所述的方法,进一步包括:

将每个概率与置信度阈值进行比较;以及

基于所述比较确定与概率相关联的说话人是否说话。

18.如权利要求1至17中任一项所述的方法,进一步包括:

部分地基于所分配的说话人标签生成与所述媒体数据相关联的一个或多个分析文件或可视化。

19.一种方法,包括:

利用至少一个处理器接收包括一个或多个话语的媒体数据;

利用所述至少一个处理器将所述媒体数据分成多个块;

利用所述至少一个处理器对所述多个块中与单个说话人相关联的每个块的片段进行标识;

利用所述至少一个处理器根据机器学习模型提取所标识片段的嵌入;

利用所述至少一个处理器将所标识片段的所述嵌入谱聚类成簇;

利用所述至少一个处理器根据所述聚类的结果为所标识片段的所述嵌入中的至少一个嵌入分配说话人标签;以及

利用所述至少一个处理器部分地基于所述说话人标签输出与所述媒体数据相关联的说话人日志信息。

20.如权利要求19所述的方法,进一步包括:

在将所述媒体数据分成多个块之前,利用所述至少一个处理器对所述媒体数据执行空间转换。

21.如权利要求20所述的方法,其中,对所述媒体数据执行所述空间转换包括:

将所述媒体数据的第一多个通道转换成与所述第一多个通道不同的第二多个通道;以及

将所述媒体数据分成多个块包括独立地将所述第二多个通道中的每个通道分成块。

22.一种非暂态计算机可读存储介质,存储用于由电子设备的至少一个处理器执行的至少一个程序,所述至少一个程序包括用于执行如权利要求1至21中任一项所述的方法的指令。

23.一种系统,包括:

至少一个处理器;以及

存储器,所述存储器耦接到所述至少一个处理器,所述存储器存储用于由所述至少一个处理器执行的至少一个程序,所述至少一个程序包括用于执行如权利要求1至21中任一项所述的方法的指令。

支持插话式内容的说话人日志

[0001] 相关申请的交叉引用

[0002] 本申请要求于2021年4月30日提交的美国临时申请63/182,338的优先权。

技术领域

[0003] 本公开总体上涉及音频信号处理,并且更具体地涉及说话人日志(speaker diarization)。

背景技术

[0004] 说话人日志是将包含多个个体言语的输入音频流划分成与每个说话人相关联的同质片段的过程。说话人日志用于许多应用中,如了解录音对话、为视频加字幕等。说话人日志不同于说话人标识或说话人分离,因为说话人日志不需要说话人语音的“指纹”或对输入音频流中存在的说话人数量有先验了解。另外,说话人日志不同于源分离,因为说话人日志通常不应用于重叠言语。

发明内容

[0005] 公开了用于支持插话式内容的说话人日志的实施例。

[0006] 在一些实施例中,一种方法包括:利用至少一个处理器接收包括一个或多个话语的媒体数据;利用所述至少一个处理器将所述媒体数据分成多个块;利用所述至少一个处理器对所述多个块中与单个说话人相关联的每个块的片段进行标识;利用所述至少一个处理器根据机器学习模型提取所标识片段的嵌入,其中,提取所标识片段的嵌入进一步包括统计地组合所提取的对应于与单个说话人相关联的相应连续话语的所标识片段的嵌入;利用所述至少一个处理器将所标识片段的所述嵌入聚类成簇;利用至少一个处理器根据所述聚类的结果为所标识片段的所述嵌入中的每个嵌入分配说话人标签;以及利用所述至少一个处理器部分地基于所述说话人标签输出与所述媒体数据相关联的说话人日志信息。

[0007] 在一些实施例中,在将所述媒体数据分成多个块之前,对所述媒体数据执行空间转换。

[0008] 在一些实施例中,对所述媒体数据执行所述空间转换包括:将所述媒体数据的第一多个通道转换成与所述第一多个通道不同的第二多个通道;以及将所述媒体数据分成多个块包括独立地将所述第二多个通道中的每个通道分成块。

[0009] 在一些实施例中,根据确定所述媒体数据对应于第一媒体类型,由第一训练数据集生成所述机器学习模型;并且根据确定所述第一媒体数据对应于与所述第一媒体类型不同的第二媒体类型,由与所述第一训练数据集不同的第二训练数据集生成所述机器学习模型。

[0010] 在一些实施例中,在聚类之前并且根据确定满足优化标准,进一步优化所提取的所标识片段的嵌入。

[0011] 在一些实施例中,在聚类之前并且根据确定不满足优化标准,放弃进一步优化所

提取的所标识片段的嵌入。

[0012] 在一些实施例中,优化所提取的所标识片段的嵌入包括:执行对所提取的嵌入进行降维或对所提取的嵌入进行嵌入优化中的至少一者。

[0013] 在一些实施例中,嵌入优化包括:训练所述机器学习模型以使所提取的所标识片段的嵌入之间的可分离性最大化;以及通过将用于使所提取的所标识片段的嵌入之间的所述可分离性最大化的所述机器学习模型应用于所提取的所标识片段的嵌入来更新所述所提取的嵌入。

[0014] 在一些实施例中,所述聚类包括:对于每个标识的片段:确定所述片段的相应长度;根据确定所述片段的相应长度大于阈值长度,根据第一聚类过程分配与所述相应标识的片段相关联的嵌入;以及根据确定所述片段的相应长度不大于阈值长度,根据与所述第一聚类过程不同的第二聚类过程分配与所述相应标识的片段相关联的嵌入。

[0015] 在一些实施例中,前述方法中的任一方法进一步包括:部分地基于确定与所述媒体数据相关联的不同说话人的数量从多个聚类过程中选择第一聚类过程。

[0016] 在一些实施例中,所述第一聚类过程包括谱聚类。

[0017] 在一些实施例中,所述媒体数据包括多个相关文件。

[0018] 在一些实施例中,所述方法进一步包括:将多个所述相关文件选择作为所述媒体数据,其中,选择所述多个相关文件是部分地基于以下各项中的至少一项:与所述多个相关文件相关联的内容相似度;与所述多个相关文件相关联的元数据相似度;或者与处理特定一组文件的请求相对应的接收数据。

[0019] 在一些实施例中,所述机器学习模型是根据由所述多个相关音频文件中的每个文件共享的一个或多个属性而从多个机器学习模型中选择的。

[0020] 在一些实施例中,所述方法进一步包括:计算声纹嵌入与每个簇的质心之间的声纹距离度量;计算从每个质心到属于所述簇的每个嵌入的距离;针对每个簇计算所述嵌入与所述簇的所述质心的距离的概率分布;针对每个概率分布,计算所述声纹距离属于所述概率分布的概率;对所述概率进行排序;基于所述排序为所述簇中的一个簇分配声纹;以及将与所述声纹相关联的说话人身份与所述说话人日志信息进行组合。

[0021] 在一些实施例中,所述概率分布被建模为折叠高斯分布。

[0022] 在一些实施例中,所述方法进一步包括:将每个概率与置信度阈值进行比较;以及基于所述比较确定与概率相关联的说话人是否说话。

[0023] 在一些实施例中,前述方法中的任一方法进一步包括部分地基于所分配的说话人标签生成与所述媒体数据相关联的一个或多个分析文件或可视化。

[0024] 在一些实施例中,一种非暂态计算机可读存储介质存储用于由电子设备的至少一个处理器执行的至少一个程序,所述至少一个程序包括用于执行上述方法中的任一方法的指令。

[0025] 在一些实施例中,一种系统包括:至少一个处理器;以及存储器,所述存储器耦接到所述至少一个处理器,所述存储器存储用于由所述至少一个处理器执行的至少一个程序,所述至少一个程序包括用于执行上述方法中的任一方法的指令。

[0026] 本文公开的特定实施例提供了以下优点中的至少一个或多个:1)优化了说话人日志的架构,与现有架构的标准日志结构相比有所改进;2)引入了预处理步骤,以在转换为单

声道之前利用立体声文件中存在的空间信息;3)引入了嵌入优化步骤,以最大化嵌入分离并基于多头注意力架构或VBx聚类改进聚类;4)引入了谱聚类作为流水线中的改进部件;5)引入了双重聚类步骤,以提高聚类的可靠性并改善说话人短片段的误分类;6)能够对任意长度的文件执行日志处理,从而降低内存占用和处理负荷;7)能够对不同的文件执行日志处理,从而允许对插话式内容执行日志处理;8)统计数据生成、误差量化和可视化,以便于评估日志成功率;以及9)使用输入声纹确定音频文件是否包含来自该人员的言语的日志流水线。

附图说明

[0027] 在附图中,为了便于描述,示出了示意性元素的特定布置或排序,如那些表示设备、单元、指令块和数据元素的示意性元素。然而,本领域技术人员应当理解,附图中示意性元素的特定排序或布置并不意味着暗示需要特定处理次序或顺序或者过程分离。进一步地,在附图中包括示意性元件并不意味着在所有实施例中都需要这种元件,或者在一些实施例中,由这种元件表示的特征可以不包括在其他元件中或与其他元件组合。

[0028] 进一步地,在附图中,在使用如实线或虚线或箭头等连接元素来说明两个或更多个其他示意性元素之间的连接、关系或关联的情况下,不存在任何这种连接元素并不意味着暗示不可能存在连接、关系或关联。换句话说,未在附图中示出元素之间的一些连接、关系或关联性,以免模糊本公开。另外,为了便于图示,使用单个连接元素来表示元素之间的多个连接、关系或关联性。例如,在连接元素表示信号、数据或指令的通信的情况下,本领域的技术人员应理解,这样的元素表示为了影响通信而可能需要的一条或多条信号路径。

[0029] 图1是根据实施例的说话人日志处理流水线的框图。

[0030] 图2是图示了根据实施例的、图1所示的流水线上的不同部件如何处理音频文件的流程图。

[0031] 图3是根据实施例的、在降维到两个维数后包含四个说话人的音频片段的嵌入表示。

[0032] 图4A是根据实施例的、图1所示的流水线中的过程的流程图,该过程使用交叉熵损失和二元分类模型来提供说话人变化检测和重叠言语检测。

[0033] 图4B是流水线100中的流程图400b,其可以使用三元组损失来解决嵌入提取过程。

[0034] 图5A是根据实施例的用于在日志流水线中处理立体声文件的流程图。

[0035] 图5B是根据实施例的用于在日志流水线中对立体声文件进行替代性处理的流程图。

[0036] 图6图示了根据实施例的块分割和插话式内容日志。

[0037] 图7图示了根据实施例的多头注意力加GE2E损失函数。

[0038] 图8是根据实施例的双重聚类过程的流程图。

[0039] 图9图示了根据实施例确定声纹嵌入与由日志流水线标识的簇的质心之间的距离。

[0040] 图10图示了根据实施例将质心与属于簇的嵌入的距离近似为折叠高斯分布。

[0041] 图11图示了根据实施例对概率进行排序以标识哪个簇包括声纹的概率最高。

[0042] 图12图示了根据实施例通过将说话人标识(ID)结果与日志流水线输出进行联合

来将说话人与所谈话的声纹进行关联。

[0043] 图13图示了根据实施例以地面真值作为基线的两个说话人的言语片段的日志性能。

[0044] 图14图示了根据实施例的、图13中所表示的每个说话人的言语片段。

[0045] 图15是根据实施例的支持插话式内容的说话人日志的过程的流程图，

[0046] 图16是根据实施例的用于实施参考图1至图15描述的特征和过程的示例设备架构的框图。

[0047] 各附图中使用的相同附图标记指示相似的元素。

具体实施方式

[0048] 在以下具体实施方式中，阐述了许多具体细节以提供对所描述的各种实施例的全面理解。对于本领域普通技术人员而言将明显的是，可以在没有这些具体细节的情况下实践所描述的各种实施例。在其他实例中，并未详细描述熟知方法、过程、部件以及电路以免不必要地模糊实施例的各方面。下文描述了几个特征，每个特征可以彼此独立使用或者与其他特征的任何组合一起使用。

[0049] 术语解释

[0050] 如本文所使用的，术语“包括”及其变体应被理解为意思是“包括但不限于”的开放式术语。除非上下文另外明确指出，否则术语“或”应被理解为“和/或”。术语“基于”应被理解为“至少部分地基于”。术语“一个示例实施例”和“示例实施例”应被理解为“至少一个示例实施例”。术语“另一个实施例”应理解为“至少一个其他实施例”。术语“确定”应被理解为获得、接收、计算、估算、估计、预测或得到。另外，在以下描述和权利要求中，除非另外定义，否则本文所使用的所有技术和科学术语具有与本公开所属领域的普通技术人员通常理解的含义相同的含义。

[0051] 示例系统

[0052] 图1是根据实施例的日志处理流水线100的框图。日志流水线100的特征在于5个主要部件：音频预处理部件101、基于块的嵌入提取部件102、嵌入处理部件103、聚类部件104和后处理部件105。

[0053] 日志流水线100将包含话语(例如，言语)的媒体数据(例如，单声道或立体声音频文件106)用作输入，并将带有在音频文件中检测到的每个说话人的相关联标记的对应片段(例如，言语片段)作为输出返回。在实施例中，日志流水线100还输出分析121和可视化120，以提供清晰的方式来向用户呈现日志结果。例如，日志流水线100可以标识音频文件中是否存在多个说话人(例如，音频文件中有3个不同的说话人)，并定义他们各自的言语的开始时间和结束时间。在一些实施例中，分析121包括但不限于：音频文件中说话人的数量、说话人标识的置信度、每个说话人参与对话的时间和百分比、对话中的主要说话人、重叠言语部分和常规轮次。

[0054] 在一些实施例中，日志流水线100的第一部件是音频预处理部件101，该音频预处理部件导入包括言语的媒体数据。在一些实施例中，媒体数据是音频文件106。在一些实施例中，音频文件106是单声道文件。在一些实施例中，音频文件106是立体声文件。在立体声文件的情况下，转换器107将音频文件106的左(L)和右(R)立体声通道转换成两个通道：通

道1和通道2。

[0055] 在一些实施例中,如果说话人被空间化并且在不同通道上平移,则音频文件106以最大化通道1和2中存在的空间信息的方式被转换。例如,如果期望保留说话人定位的空间信息,则可以使用独立成分分析(ICA)来生成其中保留空间信息的两个通道,并且将说话人分到这两个通道中。在一些实施例中,使用主成分分析(PCA)来执行该任务。在一些实施例中,使用深度学习或机器学习来执行该任务。在一些实施例中,音频文件106是通过对L通道和R通道进行子混音或通过选择这两个通道中的一个通道(例如,丢弃通道中的一个通道,以提高处理效率)而转换成单声道文件的立体声文件。

[0056] 在为了使平移说话人的空间信息的分离和保留最大化而将L通道和R通道转换成通道1和通道2后,每个通道由音频块分割部件108处理,该音频块分割部件将音频通道1和音频通道2分割成块,这些块被发送到基于块的嵌入提取部件102。如本文所使用的,“块”是处理单元。每个块可以包含一个或多个言语片段。“言语片段”是唯一说话人正在谈话的言语时间。

[0057] 注意的是,日志流水线100能够处理任何持续时间的音频文件106,并且还可以对其中可以存在相同说话人的多个文件进行处理和日志处理。例如,在一些实施例中,日志流水线100可以加载(例如,接收、取得)立体声音频文件,对音频文件106的两个通道应用ICA以最大化并利用与平移言语相关联的空间信息,并生成通道1和通道2。然后,日志流水线100将音频文件分成等长度的块(例如,1秒的块),其中,每个块由流水线100的下游部件处理。在一些实施例中,块包括多个长度(例如,0.5秒、0.1秒或不同持续时间)的块。

[0058] 每个音频通道的块由基于块的嵌入提取部件102处理,该基于块的嵌入提取部件对块执行特征提取109并对特征应用语音活动检测(VAD)113。在一些实施例中,VAD 113检测多个说话人的言语并执行重叠言语检测110。言语检测用于隔离标签言语片段,即,块中包含言语的部分。言语检测是基于来自VAD 113、说话人变化检测112和重叠言语检测110的结果的组合。

[0059] 将隔离的言语片段与用于重叠言语检测110和说话人变化检测112的数据一起输入到嵌入提取部件111中。嵌入提取部件111计算由说话人变化检测112标识的每个片段的嵌入,并丢弃重叠言语,这样就不会从重叠言语中提取嵌入。嵌入提取部件111和嵌入生成模型从隔离的言语片段中提取嵌入(例如,言语片段的多维向量表示)。在一些实施例中,嵌入提取部件111对嵌入执行降维114(参见图3),以生成要由聚类部件104进行聚类的改进的嵌入115。

[0060] 聚类部件104接收改进的嵌入115作为输入。短片段标识符116标识分别使用长片段聚类117和短片段聚类118进行单独聚类的短言语片段和长言语片段。嵌入簇被输入到后处理部件105中。后处理部件105对嵌入簇执行后处理119和分割120,使用这些嵌入簇生成分析121和可视化122。

[0061] 图2进一步详细地图示了根据实施例的、上文描述的日志流水线100的不同部件如何处理音频文件106。在言语分析中,嵌入是言语片段的多维向量表示。嵌入的生成是为了最大程度地区分多维特征空间中的不同说话人。例如,如果对话中有两个说话人在谈话,则这两个说话人生成的嵌入将是在多维空间的两个不同区域中进行聚类的多维向量。在一些实施例中,提取嵌入包括在多维子空间中表示单个说话人言语片段或单个说话人言语片段

的部分。在一些实施例中,提取嵌入包括将每个单个说话人言语片段转换成多维向量表示。

[0062] 生成嵌入的原因是促进对不同说话人进行聚类,从而成功地执行说话人日志。使用如嵌入生成模型等模型以通过将言语片段映射到或转换成多维空间中的表示来生成嵌入。这些模型可以使用不同的数据集并通过使用损失函数来训练以实现嵌入生成,这些损失函数使不同说话人的嵌入之间的距离最大化并且使从同一说话人生成的嵌入的距离最小化。在一些实施例中,嵌入模型是被训练为具有说话人判别嵌入层的深度神经网络(DNN)。

[0063] 参考图2,在所示示例中,原始音频包括来自三个不同说话人的言语以及音乐和噪音形式的环境声音。音频块分割产生第一个块(块1),其捕获说话人1(SPK1)的言语,然后是音乐,之后是来自说话人2(SPK2)的言语。将块1输入到检测SPK1言语的开始和结束的VAD 113中,从而产生言语片段1(SEGM1)。生成言语片段1的多个嵌入并从SEGM1的每个嵌入中提取特征,并且对嵌入进行统计组合(例如,通过根据多个嵌入计算平均嵌入)。音频块是音频文件的统一分割(例如,1s、2s、3s等)

[0064] 对块2和块3重复以上过程。例如,块2生成包含来自SPK3的言语的言语片段4(SEGM4)。块3生成包含来自SPK3的言语的言语片段7(SEGM7)。从SEGM4和SEGM7中提取特征的平均嵌入。然后,将块1、块2和块3的平均嵌入输入到聚类过程中(例如,短片段聚类和长片段聚类)。每个簇表示来自3个说话人(例如,多维特征空间的三个不同区域中的3个簇)中的一个说话人的言语。

[0065] 图3是根据实施例的、在使用t分布随机近邻嵌入(t-SNE)降维到两个维数后包含4个说话人的音频片段的嵌入表示。已使用例如以下文献中描述的t-SNE算法将512-D嵌入减少到2D以实现可视化目的:Van der Maaten、Laurens和Geoffrey Hinton.“Visualizing data using t-SNE.”Journal of machine learning research 9.11(2008)[使用t-SNE使数据可视化.机器学习研究杂志2008年9月11日]。图3中的每个点表示言语片段的嵌入。属于同一说话人的嵌入往往在空间中聚类,并且在图3中可以标识4个不同的簇,每个簇表示4个说话人中的一不同说话人。

[0066] 再次参考图1,流水线100包括在嵌入提取之前执行的3种言语检测:1)VAD 113;2)重叠言语检测110;以及3)说话人变化检测112。VAD 113将言语片段与音乐或噪音或其他音频源隔离。在一些实施例中,使用AC-4标准中描述的VAD执行言语活动检测。在一些实施例中,VAD 113是基于下文图4A、图4B所描述的架构。重叠言语检测110标识存在重叠言语的区域。通常会从日志任务中丢弃重叠言语。可以使用过程400a来检测重叠言语。也可以使用其他架构进行重叠言语检测,包括不同的CNN和/或RNN架构。说话人变化检测112检测对话轮次。对话轮次是从第一说话人到不同的第二说话人的过渡。对VAD 113的输出执行说话人变化检测112。也可以使用过程400a进行说话人变化检测。也可以使用其他架构进行说话人变化检测,如不同的CNN和/或RNN架构。

[0067] 图4A是根据一些实施例的流水线100中的过程400a的流程图,该过程使用交叉熵损失和二元分类模型来提供说话人变化检测112和重叠言语检测110。在一些实施例中,过程400a包括卷积神经网络(CNN)架构401a,然后是一个或多个卷积层402a,接着是一个或多个循环神经网络(RNN)层403a,之后是一个或多个前馈神经网络(FNN)层404a。在一些实施例中,这些层是针对二元分类问题(例如,言语的存在或不存在)使用如后向传播等合适的

训练技术来进行训练的。

[0068] 图4B是日志流水线100中的过程400b的流程图,该过程可以使用三元组损失来解决嵌入提取过程。在一些实施例中,过程400b包括卷积神经网络(CNN)层401b,然后是一个或多个卷积层402b,接着是一个或多个循环层403b,之后是时序池化404b,再之后是一个或多个前馈层404a。过程400b的输出是言语片段。

[0069] 在过程400a和400b中,CNN 401a、CNN 401b可以使用例如以下文献中描述的SincCONV架构来实施:Ravanelli,M.&Bengio,Y.(2019).Speaker Recognition from Raw Waveform with SincNet.2018IEEE Spoken Language Technology Workshop,SLT 2018-Proceedings[利用SincNet根据原始波形进行说话人识别.2018年IEEE口语技术研讨会],1021-1028.https://doi.org/10.1109/SLT.2018.8639585。在一些实施例中,SincCONV架构可以由手工设计特征(例如,梅尔频率倒谱系数(MFCC))代替。

[0070] 在将言语分割成仅存在一个说话人的片段(例如,每个片段不包括来自多于一个说话人的言语,如图2所示)并且隔离不同说话人的片段后,由基于块的嵌入提取部件102提取言语中的嵌入。在一些实施例中,可以使用x向量、d向量、i向量或其他策略来提取言语中的嵌入。出于相同目的,在对言语数据进行模型训练以使不同说话人之间的可分离性最大化后,可以使用图4B中的架构。在训练期间可以使用不同的成本函数来使损失最小化,如三元组损失、交叉熵损失、基于概率线性判别分析(PLDA)的损失等。

[0071] 在一些实施例中,在指定的窗口长度上提取嵌入,如图2所示。在一些实施例中,在恒定时间窗口长度(例如,20ms)上提取嵌入。在提取后,对属于来自同一说话人的连续言语片段的嵌入进行平均以减少噪音并提高准确度,如图2所示。被平均的嵌入的数量和定义用于进行平均的最大时间窗口的阈值可以根据输入音频文件106的长度进行修改。对于长输入音频文件,对属于来自同一说话人的整个言语片段的一部分的所有嵌入进行平均是可接受的。相比之下,在短输入音频文件的情况下,可以减少用于进行嵌入平均的最大可接受时间窗口,以生成总体上更大数量的嵌入,这些嵌入将被馈送到聚类部件104以提高聚类准确度。

[0072] 针对不同的内容类型,可以使用不同的训练数据集(例如,对应于相应内容或媒体类型的训练数据)来训练不同的嵌入模型。在一些实施例中,嵌入模型被优化为从与以下各项相关联的内容中提取(推断)嵌入:环境(例如,体育场地、音乐会场地、车辆座舱、录音棚)、用例(例如,播客、讲座、采访、体育广播、音乐)、学科或话题(例如,化学)、活动(例如,通勤、步行)等。在一些实施例中,嵌入模型被优化用于用户生成的内容(例如,业余记录、移动设备记录等)。在一些实施例中,需要进行日志处理的内容类型(例如,播客、电话呼叫、教育、用户生成的内容等)由应用程序或应用程序的用户指定,并且日志流水线100根据指定的内容类型为需要进行处理的内容类型选择最佳模型。在一些实施例中,根据各种方法(例如,查找表、内容分析)选择最佳模型。

[0073] 在提取嵌入111后,进一步处理嵌入,如图1所示。在实施例中,嵌入的处理由以下两个部件执行:1)嵌入降维部件114和2)嵌入优化部件115。在一些实施例中,例如,响应于确定提取的嵌入高于质量阈值,跳过部件114、115中的一个或两个部件。

[0074] 实施降维部件114是为了提高区分不同说话人并使聚类的成功率最大化的能力。可以使用PCA、t-SNE、统一流形逼近与投影(UMAP)、PLDA或其他降维策略来降低嵌入维数。

[0075] 嵌入优化部件115将在以下部分中进行更详细的描述。其旨在使用数据驱动的方法来训练模型以使嵌入之间的可分离性最大化并进一步促进聚类过程。在一些实施例中，这可以使用用于嵌入优化的多头注意力加广义端到端(GE2E)损失架构来实现。也可以使用其他架构来实现该目的。

[0076] 在嵌入提取和处理后，将嵌入聚类到不同的说话人。聚类部件105通过区分短言语片段117和长言语片段118来执行聚类，如图1所示，并且如下文进一步详细描述的。可以使用不同的聚类策略来实现短片段聚类或长片段聚类。在一些实施例中，在例如以下文献中描述了一种改进的谱聚类:Wang,Q.、Downey,C.、Wan,L.、Mansfield,P.A.&Moreno,I.L.(2018).c Speaker diarization with LSTM.ICASSP,IEEE International Conference on Acoustics,Speech and Signal Processing-Proceedings,2018-April[利用LSTM进行c说话人日志处理.ICASSP,IEEE国际声学、语音与信号处理会议,2018年4月],5239-5243。

[0077] 也可以使用其他聚类策略，如层次聚类、近邻传播、凝聚聚类或VBx聚类。在一些实施例中，音频文件106中的说话人数量是已知的或确定的。例如，可以获得指示接收到的(例如，经由应用程序编程接口(API)调用)或基于使用例如k均值聚类或其他合适的聚类算法对音频文件进行分析而得到的与音频文件相关联的说话人数量的数据。在实施例中，当事先不知道说话人数量时，可以将肘部法与k均值聚类算法一起使用。肘部法可以包括绘制解释变异作为簇的数量的函数的曲线，然后选取曲线的肘部作为要使用的簇的数量。

[0078] 在聚类后，对聚类的嵌入执行后处理119。执行后处理119是为了消除错误检测并提高聚类准确度。在一些实施例中，为了实现该目的，使用一种具有不同说话人之间的预定义过渡概率的方法(比如隐马尔可夫模型(HMM))。所得的经改进和后处理的嵌入由分割部件120使用以将音频文件分割成具有相关联的说话人标签/标识符的片段。

[0079] 分析121是根据聚类的嵌入生成以作为日志流水线100的输出。分析121可以包括信息，该信息包括但不限于:说话人的数量、说话人检测的置信度、说话人参与对话的百分比和时间、最相关说话人、以及生成的其他分析。日志流水线100还可以在说话人日志任务中生成检测准确度的概述。

[0080] 示例预处理块

[0081] 如先前所述，音频预处理部件101允许日志流水线100生成用于立体声输入文件和单声道输入文件的日志输出。在单声道文件的情况下，音频文件被视为单声道文件并由音频块分割块108处理，如下文所述。在立体声文件的情况下，可以应用过程500a、500b。

[0082] 图5A是根据实施例的用于在日志流水线100中处理立体声文件的流程图。过程500a(图5A所示)图示了在左通道501a和右通道502a不同并且说话人的语音经过平移并在左通道与右通道之间不对称时利用的实施例。平移言语具有可以用于日志任务的空间信息。对于每个通道501a、502a，空间转换块503a将空间信息分成两个不同的通道504a(通道1，通道2)，以提高日志任务的成功率。在一些实施例中，ICA或PCA由空间转换块503a实施。也可以使用深度学习策略。将日志流水线100的其余块单独应用于通道1和通道2。在块处理505a和嵌入的提取阶段506a后，对嵌入进行联合处理507a(例如，将来自每个通道的嵌入进行组合以用于后续处理)，以生成利用与说话人空间定位相关联的信息的嵌入的经处理版本。在一些实施例中，过程500a被应用于多通道音频(例如，Dolby® Atmos®)，而不是双耳音频。

[0083] 图5B是根据实施例的用于在日志流水线100中对立体声文件进行替代性处理的流程图。在一些实施例中,当左通道501a和右通道501b相同(图4B)时,将立体声文件转换成单声道503b,并且将日志流水线100的其余块应用于单个单声道文件,如参考图5A所述。单声道转换通过选择这两个通道中的一个通道或通过降混音来实现。

[0084] 示例音频块分割

[0085] 音频块分割108是流水线100中的另一个步骤(参见图1、图2和图6)。在多个音频块上从每个片段中提取到所有嵌入后,对由每个言语片段(片段是唯一说话人正在谈话的言语的时间窗口)提取的嵌入进行存储和聚类。该过程增加了处理任意长度的文件的可能性,并且在不影响性能的情况下限制了处理整个文件所需的处理量和内存使用量。该过程可以应用于输入文件列表,从而实现对插话式内容执行日志处理的可能性。

[0086] 参考图6,来自插话式内容(例如,播客季)的输入音频文件601的集合由音频块分割602进行分割并通过嵌入的提取部件603进行馈送,如先前参考图1所描述的。然后,提取的嵌入604在嵌入处理605中组合在一起、进行聚类606并输入到后处理/分割607中。这些步骤允许在多个音频文件106上执行说话人日志,从而实现对插话式内容计算分析的可能性。

[0087] 示例嵌入

[0088] 嵌入在说话人日志系统中起着重要的作用,因为嵌入是言语片段的多维向量表示。现有的解决方案利用基于高斯混合模型(GMM)的嵌入提取、i向量、x向量、d向量等。虽然这些现有方法在解决说话人日志问题上可能是稳健的并且展现出良好的性能,但这些方法并没有充分利用时间信息。因此,在日志流水线100中引入了新的模块,以通过利用时间信息进一步提高嵌入的有效性,从而提高下一个步骤中聚类的准确度。

[0089] 在文献中已经提出了利用言语的时间信息生成改进的嵌入的若干相关理念。例如,提出了基于长短期记忆(LSTM)的向量到序列评分模型,该评分模型利用相邻时间信息以根据嵌入生成相似度得分。然而,LSTM结构存在两个局限性:1)LSTM结构更侧重于局部信息并且可能无法完成长期依赖任务;以及2)LSTM结构既耗时又耗空间。

[0090] 为了解决LSTM结构的局限性,完全基于注意力机制的架构在序列到序列学习中展现出有希望的价值。除了提供显著更快的训练速度外,注意力网络还证明了长期依赖的高效建模。例如,可以使用位置多头注意力模型结构和三元组损失函数。位置编码是包含关于言语中特定位置的信息的多维向量。另外,三元组排序损失用于使用来自多头注意力模型的输出来学习相似度度量。

[0091] 在一些实施例中,可以使用新的损失函数——称为‘端对端损失函数’——来解决说话人验证问题。GE2E训练是基于以批的形式一次性处理大量话语,一个批包含N个说话人和来自每个说话人的平均M个话语。相似度矩阵$S_{ji,k}$被定义为每个嵌入向量e_{ji}与所有质心c_k之间的缩放余弦相似度。

[0092] $s_{ji,k}=w\cdot\cos(e_{ji},c_k)+b$

[0093] [1]

[0094] 此处,e_{ji}表示第j^{th}个说话人的第i^{th}个话语的嵌入向量($1<=i<=M,j<=N$),c_k表示嵌入向量的质心,其中,$k<=N$。

[0095] 由于说话人日志问题也是尝试使用提取的嵌入找到相似度矩阵,然后基于相似度矩阵执行聚类算法,因此,可以消除等式[1]中对M个话语的约束,并且可以使用类似的等式

来计算损失,因为在说话人日志问题中,说话人可以不限次数地谈话。与等式[1]类似,e_{ji}表示第j^{th}个说话人的第i^{th}个话语的嵌入向量(i>=1,j<=N),c_k表示嵌入向量的质心,其中,k<=N。然后,如果在等式[1]中使用softmax函数,则每个嵌入向量eji的损失可以定义为:

[0096] $\mathcal{L}(e_{ji}) = -S_{ji,k} + log\sum_{k=1}^{N}\exp(S_{ji,k})$。

[0097] [2]

[0098] 图7图示了根据实施例的多头注意力加GE2E损失函数。如图7所示,总体嵌入优化模块700将输入嵌入701发送到多头注意力模型中。在每个时间步长内应用位置编码702,这意味着训练过程使用时序池化706以在训练过程中考虑时间信息。训练过程使损失函数707(例如,等式[2]中所示的GE2E损失函数)最小化。损失函数707将每个嵌入向量推向靠近其质心的位置,并将其拉离所有其他质心。在测试期间,利用在多头注意力模型内训练的权重和偏差来预测每个输入的优化嵌入并形成相似度矩阵708,该相似度矩阵被输入到聚类709中。由于该训练过程考虑了时间信息,因此与原始嵌入相比,该训练过程应当能够构造出更好的相似度矩阵。

[0099] 示例聚类技术

[0100] 谱聚类是已经在包括日志的多个任务中使用的标准聚类算法。在日志流水线100中,使用修改版的谱聚类。图1图示了将修改版的谱聚类与参考图1描述的日志流水线100进行组合的实施例。经修改的谱聚类是由日志流水线100支持的一种示例聚类策略。在一些实施例中,可以使用其他聚类策略,包括但不限于:可以使用层次聚类、k均值、近邻传播、或凝聚聚类。

[0101] 如先前所述,为了消除噪音并提高嵌入的估计准确度,在一些实施例中,对在存在单个说话人的言语片段内生成的嵌入进行平均。出于该原因,短持续时间言语片段的信息可能没有详细到足够表征说话人。这可能是由于在长持续时间言语片段中被平均掉的嵌入的不准确性造成的。因此,提出了两步聚类方法,如参考图8所描述的。

[0102] 在一些实施例中,可以在日志流水线100中使用x向量序列(VBx)的贝叶斯HMM聚类,其在以下文献中进行了描述:Landini Federico、Ján Profant、Mireia Diez和Lukáš Burget.“Bayesian HMM clustering of x-vector sequences(VBx)in speaker diarization:theory,implementation and analysis on standard tasks.”Computer Speech&Language[说话人日志中的x向量序列(VBx)的贝叶斯HMM聚类:关于标准任务的理论、实施和分析.计算机语音和语言]71(2022):101254。在日志流水线100中,VBx被修改为采用x向量和其他嵌入。

[0103] 图8是根据实施例的双重聚类过程的流程图。如图8所示,提取每个言语片段801的嵌入。这些言语片段基于持续时间阈值而被分成长持续时间言语片段802和短持续时间言语片段803。长持续时间言语片段802先于803进行聚类,并且计算每个簇的中心804。计算每个短持续时间言语片段与每个簇之间的余弦距离(例如,艾森(Eisen)余弦相关距离)805,并且将短持续时间言语片段803分配给最接近的簇。可以使用其他距离度量(例如,欧氏(Euclidean)、曼哈顿(Manhattan)、皮尔逊(Pearson)、斯皮尔曼(Spearman)、肯德尔(Kendall)相关距离),或者可以采用其他策略以将每个短言语片段803分配给最佳簇。总体聚类结果是短持续时间言语片段803和长持续时间言语片段802的聚类结果的组合。

[0104] 示例后处理

[0105] 处理流水线100中的另一个部件是后处理部件105。在流水线100中引入后处理部件105是为了减少分割和聚类错误。后处理105分析由聚类生成的注释,并标识和纠正可能的错误。可以在流水线100中使用若干后处理策略。例如,可以使用具有相关联的说话人变化概率的HMM。出于相同目的,可以使用利用注释之间的时间关系的其他算法。

[0106] 示例说话人标识

[0107] 说话人标识(说话人ID)是基于输入声纹了解说话人身份的能力。在一些实施例中,日志流水线100使用输入声纹(例如,人员谈话的10秒至20秒)来确定音频文件是否包含来自该人员的言语。另外,基于对多个说话人声纹的了解,可以将说话人ID标签分配给日志流水线100的输出。如上所述,日志流水线100计算每个言语片段的嵌入,并且随后将这些嵌入聚类成有限数量的簇。

[0108] 图9和图10图示了根据实施例如何将参考点与属于簇的嵌入的距离近似为折叠高斯分布。参考点可以是质心或任何其他合适的参考点,如在围绕中心点划分(PAM)聚类中的中心点。如图9所示,说话人ID过程的第一步骤是计算声纹距离度量,在该示例中,该声纹距离度量是声纹vp(905)与每个簇的每个参考点(例如,质心c_i(对于簇901至904))之间的余弦距离度量d_vp_i,其中,i为对簇进行标识的索引,其范围为1到N,其中,N为簇的数量。在所示示例中,N=4个簇表示4个主要说话人。

[0109] 对于每个簇,计算从簇的参考点到属于该簇的每个嵌入的余弦距离。对于每个簇901至904,还计算嵌入与属于该簇的参考点的距离的分布。在一些实施例中,分布被建模为折叠高斯分布(p_i),如图10所示。随后,计算声纹距离属于对应分布的概率。如图11所图示的,可以对概率进行排序并将声纹分配给对应的簇,其中,在该示例中Spk1的概率最高。通过限定置信度阈值,可以更好地了解声纹人员在任何情况下在音频文件中说话的概率。如图12所图示的,通过将说话人ID结果与日志流水线100的输出联合起来可以确定了解与声纹相关联的说话人何时说话。在图12所示的示例中,使用说话人ID结果来标识言语片段1201-1到1201-4,其中,Spk1正在说话。

[0110] 示例分析/可视化

[0111] 日志流水线100生成分析、度量和可视化作为输出。输出使得用户能够对日志流水线100的输出进行客观且清晰的量化。在一些实施例中,为单个文件、为多个插话式文件和/或为整个数据集生成分析和可视化。下表1和表2汇总了可以由日志流水线100生成的示例分析。在一些实施例中,可视化包括根据从聚类的嵌入得出的一个或多个统计数据促使显示表示第一说话人的第一对象、表示相应说话人的第二对象以及在外观(长度、大小、颜色、形状、样式等)上变化的连接第一对象和第二对象的连接器对象。

[0112] 表1—示例单个文件分析汇总

[0113]

名称	目的
谈话者的数量	得到对话中谈话者的总数量
每个谈话者检测的置信度	了解对对话中每个被检测到的谈话者的置信度如何
每个谈话者参与对话的百分比	了解谁是主导谈话者/每个谈话者在对话中谈话的时间是多少
每个谈话者的言语片段	知道每个谈话者的每个言语片段的开始时间和结束时间
最相关的谈话者	了解在对话中谁是主要说话人、播客的主持人、主要发言人等
每个谈话者的最长独白	了解每个谈话者的最长独白
对话变化	标识对话中说话人何时改变
说话人重叠	标识何时有两个说话人同时说话
礼貌等级	基于其他说话人中断的次数来量化说话人的礼貌程度
沉默次数	量化对话中的沉默。

[0114] 表2—示例插话式分析汇总

[0115]

名称	目的
谈话者的数量	得到插话式内容中谈话者的总数量
每个谈话者的对话时间	在多个插话中，每个谈话者在对话中谈话的时间是多少。
谈话者之间的交互	使用交互图（图8）定义在该季插话中谈话者间的交互。
每个插话的主要谈话者	定义每个插话的主要谈话者
将单个文件分析扩展到整季	针对单个文件描述的所有分析也可以用于描述季播内容。
……	

[0116] 另外，可以评估性能日志流水线100。利用日志地面真值，表3中的评估度量可以用于生成允许用户更好地评估日志流水线100的性能的报告。

[0117] 表3—示例评估度量汇总

[0118]

模块	评估度量
言语活动检测	检测成本函数
	检测错误率
	检测准确度
	检测精确度
	检测召回率
分割	分割纯度
	分割覆盖率
日志	日志错误率
	雅卡尔（Jaccard）错误率
	日志纯度
	日志覆盖率
	日志完整度
	日志均匀度
以说话人为中心的评估	说话人的数量
	说话人持续时间
	每个说话人的精确度

[0119]

	每个说话人的召回率
	每个说话人的准确度

[0120] 在一些实施例中,日志流水线100包括可视化模块,以帮助用户以更高效的方式了解日志性能。例如,图13示出了每个说话人谈话时间的百分比,这有助于了解每个说话人的相对持续时间以及有多少言语被正确检测到。图14示出了每个说话人的言语片段,其中地面真值作为基线示出,这有助于更直观地了解日志流水线100的性能。

[0121] 示例过程

[0122] 图15是根据实施例的情境感知音频处理的过程1500的流程图。过程1500可以使用例如参考图16描述的设备架构1500来实施。

[0123] 过程1500包括以下步骤:接收包括一个或多个话语的媒体数据(1501);将媒体数据分成多个块(1502);标识多个块中与单个说话人相关联的每个块的片段(1503);根据机器学习模型提取所标识片段的嵌入(1504);将所标识片段的嵌入聚类成簇(1505);以及根据聚类的结果为所标识片段的嵌入中的每个嵌入分配说话人标签(1506)。先前上文参考图1至图14详细描述了这些步骤中的每个步骤。

[0124] 示例系统架构

[0125] 图16示出了适合于实施参考图1至图15描述的示例实施例的示例系统1600的框图。系统1600包括中央处理单元(CPU)1601,所述中央处理单元能够根据存储在例如只读存

储器(ROM)1602中的程序或者从例如存储单元1608加载到随机存取存储器(RAM)1603的程序来执行各种进程。在RAM 1603中,根据需要,还存储CPU 1601执行各种进程时所需的数据。CPU 1601、ROM 1602和RAM 1603经由总线1604相互连接。输入/输出(I/O)接口1605也连接到总线1604。

[0126] 以下部件连接到I/O接口1605:输入单元1606,其可以包括键盘、鼠标等;输出单元1607,其可以包括如液晶显示器(LCD)等显示器以及一个或多个扬声器;存储单元1608,其包括硬盘或另一种合适的存储设备;以及通信单元1609,其包括如网卡(例如,有线或无线)等网络接口卡。

[0127] 在一些实施例中,输入单元1606包括位于不同位置(取决于主机设备)的一个或多个麦克风,所述一个或多个麦克风使得能够捕获各种格式(例如,单声道、立体声、空间、沉浸式和其他合适的格式)的音频信号。

[0128] 在一些实施例中,输出单元1607包括具有各种数量的扬声器的系统。输出单元1607可以呈现各种格式(例如,单声道、立体声、沉浸式、双耳和其他合适的格式)的音频信号。

[0129] 通信单元1609被配置为(例如,经由网络)与其他设备通信。根据需要,驱动器1610也连接到I/O接口1605。根据需要,如磁盘、光盘、磁光盘、闪存驱动器或其他合适的可移动介质等可移动介质1611被安装在驱动器1610上,使得从中读取的计算机程序被安装到存储单元1608中。本领域技术人员将理解,尽管系统1600被描述为包括上文所描述的部件,但是在实际应用中,可以添加、移除和/或替换这些部件中的一些部件,并且所有这些修改或变更都落入本公开的范围内。

[0130] 根据本公开的示例实施例,上文所描述的过程可以实施为计算机软件程序或者在计算机可读存储介质上实施。例如,本公开的实施例包括计算机程序产品,所述计算机程序产品包括有形地体现在机器可读介质上的计算机程序,所述计算机程序包括用于执行方法的程序代码。在这种实施例中,计算机程序可以经由通信单元709从网络下载和安装,和/或从可移动介质1611安装,如图16所示。

[0131] 通常,本公开的各种示例实施例可以以硬件或专用电路(例如,控制电路)、软件、逻辑或其任何组合来实施。例如,上文所讨论的单元可以由控制电路(例如,与图16的其他部件组合的CPU)执行,因此,控制电路可以执行本公开中描述的动作。一些方面可以以硬件来实施,而其他方面可以以可以由控制器、微处理器或其他计算设备(例如,控制电路)执行的固件或软件来实施。尽管本公开的示例实施例的各个方面被图示和描述为框图、流程图或使用一些其他图形表示,但应当理解,本文所描述的框、装置、系统、技术或方法可以以硬件、软件、固件、专用电路或逻辑、通用硬件或控制器、或其他计算设备、或其某种组合来实施,作为非限制性示例。

[0132] 另外,流程图中所示的各个框可以被视为方法步骤、和/或由计算机程序代码的操作产生的操作、和/或被构造为执行相关联的(多个)功能的多个耦接逻辑电路元件。例如,本公开的实施例包括计算机程序产品,所述计算机程序产品包括有形地体现在机器可读介质上的计算机程序,所述计算机程序包含被配置为执行上文所描述的方法的程序代码。

[0133] 在本公开的上下文中,机器可读介质可以是可以包含或存储用于由指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合使用的程序的任何有形介质。机器

可读介质可以是机器可读信号介质或机器可读存储介质。机器可读介质可以是非暂态的并且可以包括但不限于电子、磁性、光学、电磁、红外或半导体系统、装置或设备、或前述各项的任何合适的组合。机器可读存储介质的更具体的示例将包括具有一条或多条导线的电连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或闪速存储器)、光纤、便携式致密盘只读存储器(CD-ROM)、光存储设备、磁存储设备、或者前述各项的任何合适的组合。

[0134] 用于执行本公开的方法的计算机程序代码可以用一种或多种编程语言的任何组合来编写。这些计算机程序代码可以被提供给通用计算机、专用计算机或具有控制电路的其他可编程数据处理装置的处理器,使得程序代码在由计算机的处理器或其他可编程数据处理装置的处理器执行时,实施流程图和/或框图中指定的功能/操作。程序代码可以完全在计算机上执行,部分在计算机上执行,作为独立的软件包,部分在计算机上执行、并且部分在远程计算机上执行,或者完全在远程计算机或服务器上执行,或者分布在一个或多个远程计算机和/或服务器上。

[0135] 虽然本文档包含许多具体实施例细节,但是这些细节不应被解释为对可能要求保护的事物的范围的限制,而是被解释为对可能特定于特定实施例的特征的描述。本说明书中在单独实施例的上下文中描述的特定特征还可以在单个实施例中以组合形式实施。相反,在单一实施例的上下文中描述的各种特征也可以被单独地或以任何适合的子组合的方式实施在多个实施例中。此外,尽管特征在上文可以被描述为以某些组合起作用并且甚至最初也是如此要求保护的,但是在一些情况下可以从组合中去除要求保护的组合的一个或多个特征,并且所要求保护的组合可以涉及子组合或子组合的变体。在附图中描绘的逻辑流程不需要所示出的特定顺序或者有序顺序来实现期望的结果。另外,可以从所描述的流程中提供其他步骤,或者可以从中删除步骤,并且可以向所描述的系统添加其他部件,或者从中去除其他部件。因此,其他实施例也在所附权利要求的范围内。

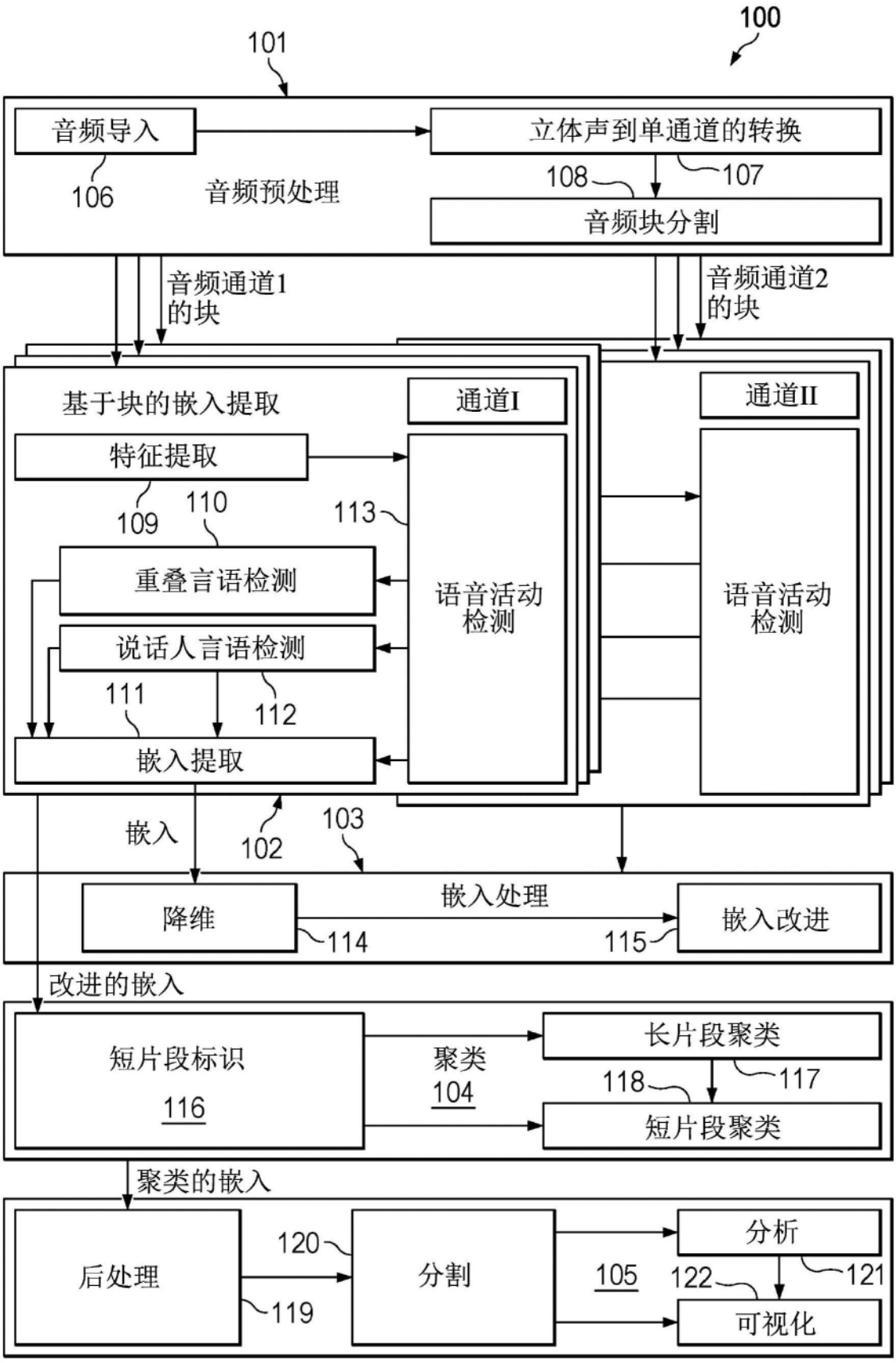

100
101
音频导入
106
音频预处理
立体声到单通道的转换
108
107
音频块分割
音频通道1的块
音频通道2的块
基于块的嵌入提取
通道I
通道II
特征提取
109
110
113
重叠言语检测
说话人言语检测
语音活动检测
语音活动检测
111
112
嵌入提取
嵌入
102
103
嵌入处理
降维
114
115
嵌入改进
改进的嵌入
短片段标识
116
聚类
104
长片段聚类
118
117
短片段聚类
聚类的嵌入
后处理
120
119
分割
105
分析
122
121
可视化

图1

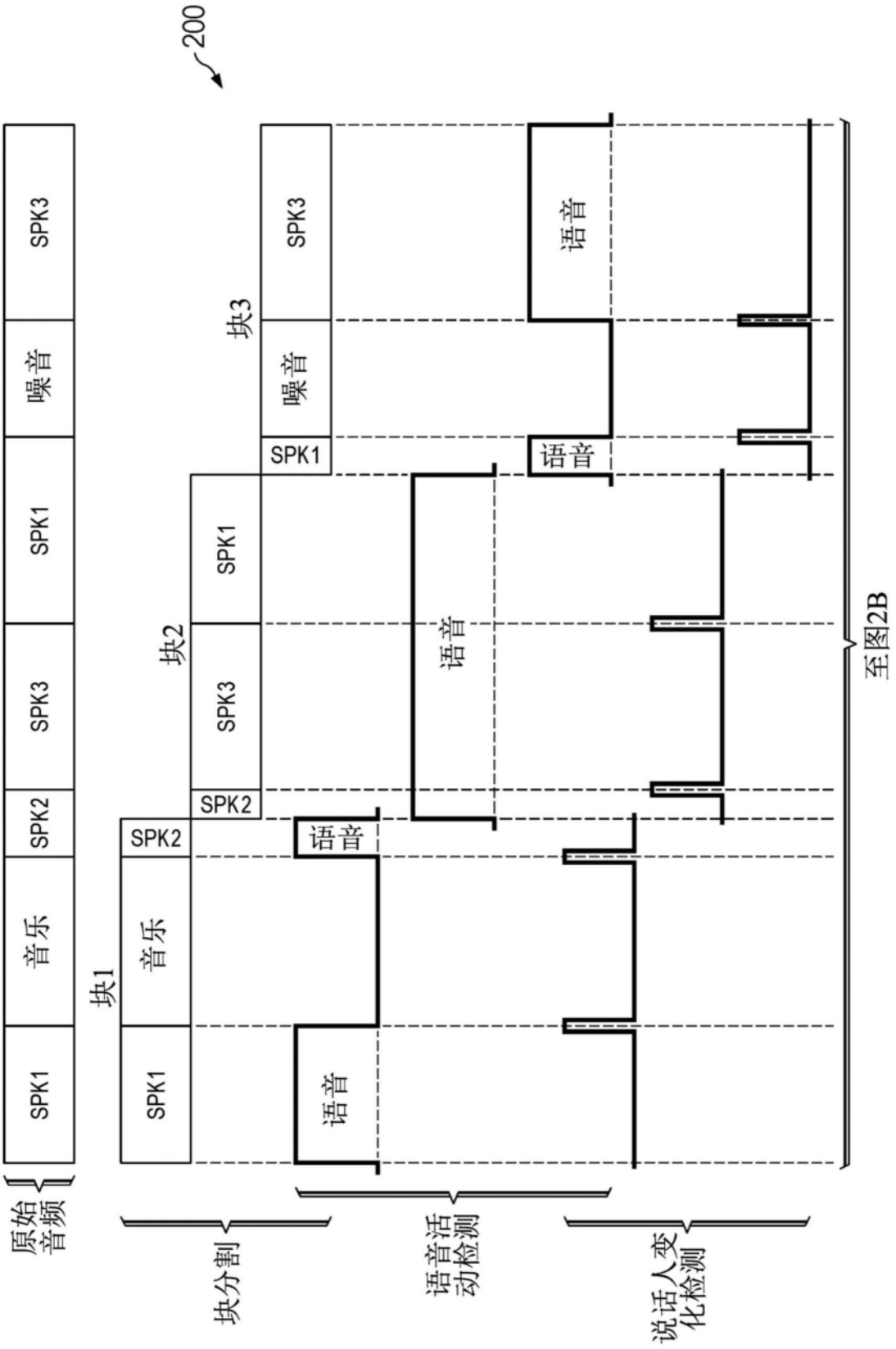

200
SPK1
音乐
SPK2
SPK3
SPK1
噪音
SPK3
原始音频
块1
SPK1
音乐
SPK2
语音
块2
SPK2
SPK3
SPK1
语音
块3
SPK1
噪音
SPK3
语音
至图2B
块分割
语音活动检测
说话人变化检测

图2A

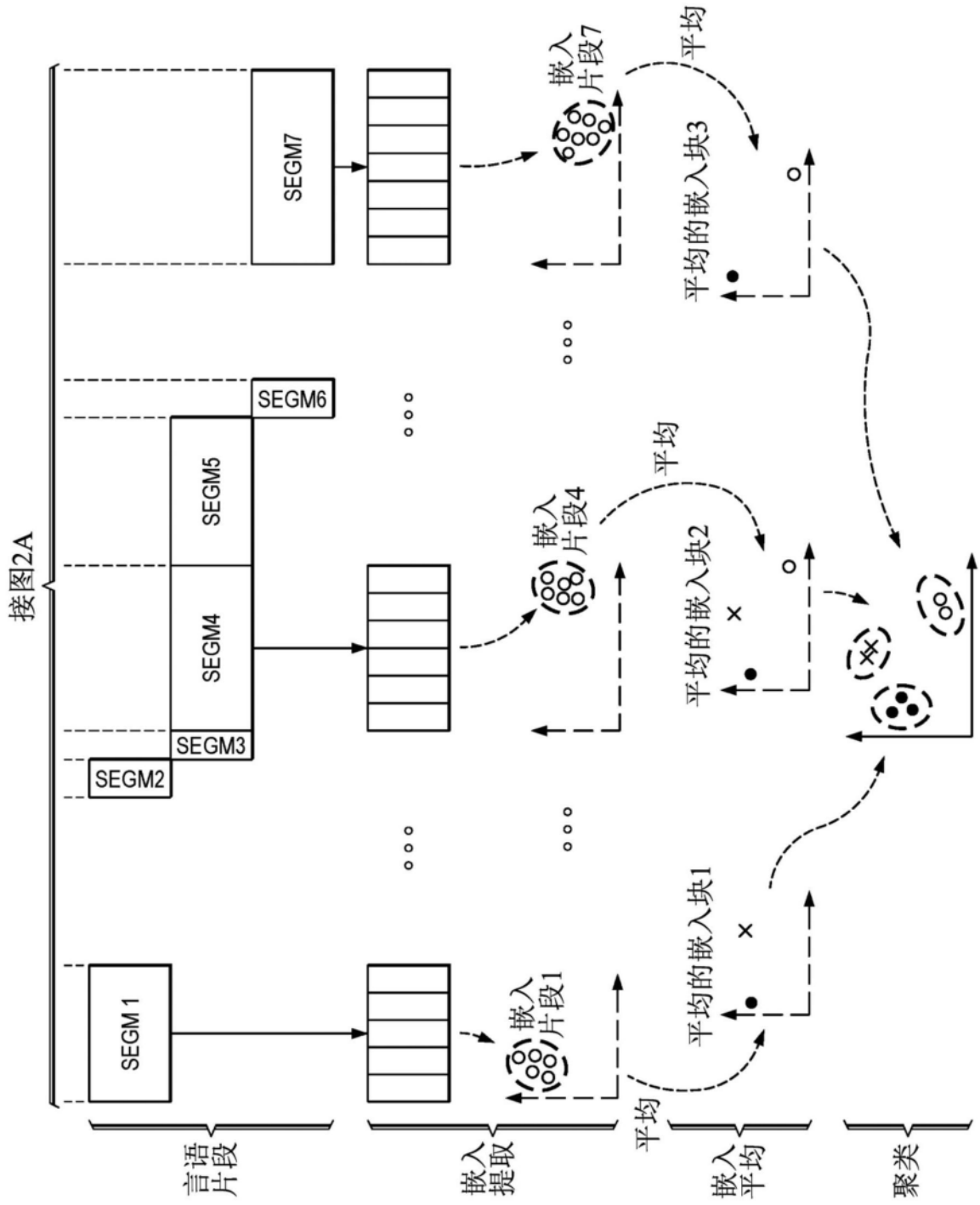

接图2A
SEGM 1
SEGM2
SEGM3
SEGM4
SEGM5
SEGM6
SEGM7
嵌入片段1
嵌入片段4
嵌入片段7
平均的嵌入块1
平均的嵌入块2
平均的嵌入块3
平均
平均
言语片段
嵌入提取
嵌入平均
聚类

图2B

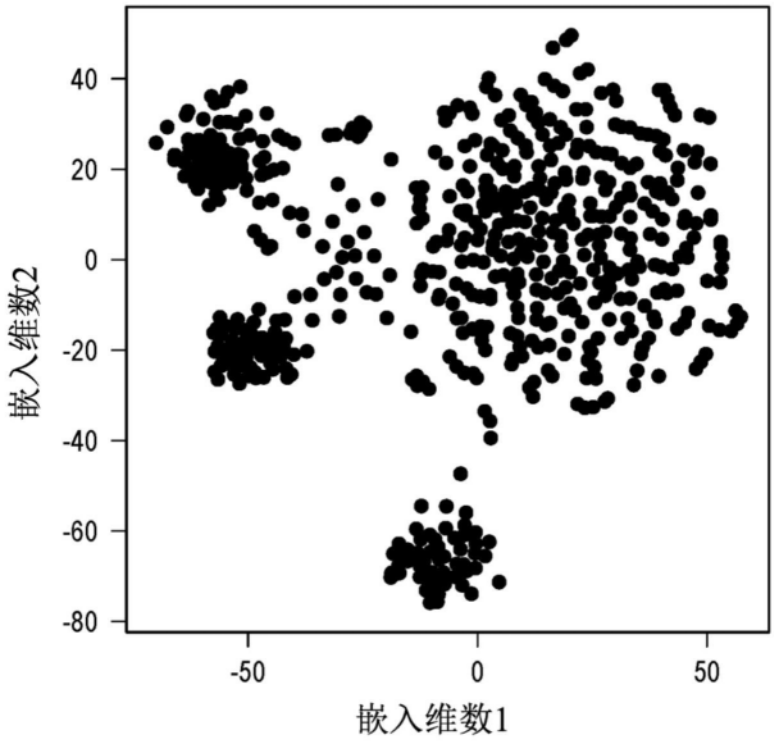

40
20
0
-20
-40
-60
-80
嵌入维数2
-50
0
50
嵌入维数1

图3

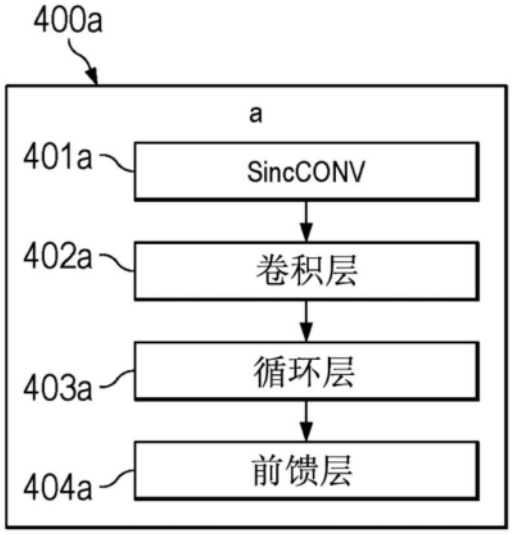

400a
a
401a
SincCONV
402a
卷积层
403a
循环层
404a
前馈层

图4A

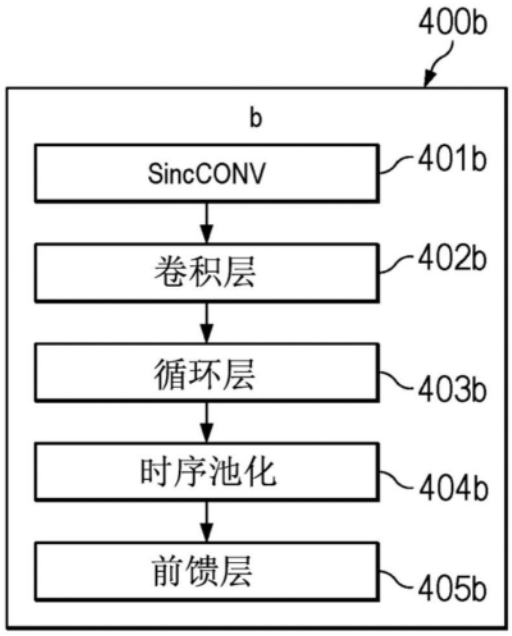

400b
b
SincCONV
401b
卷积层
402b
循环层
403b
时序池化
404b
前馈层
405b

图4B

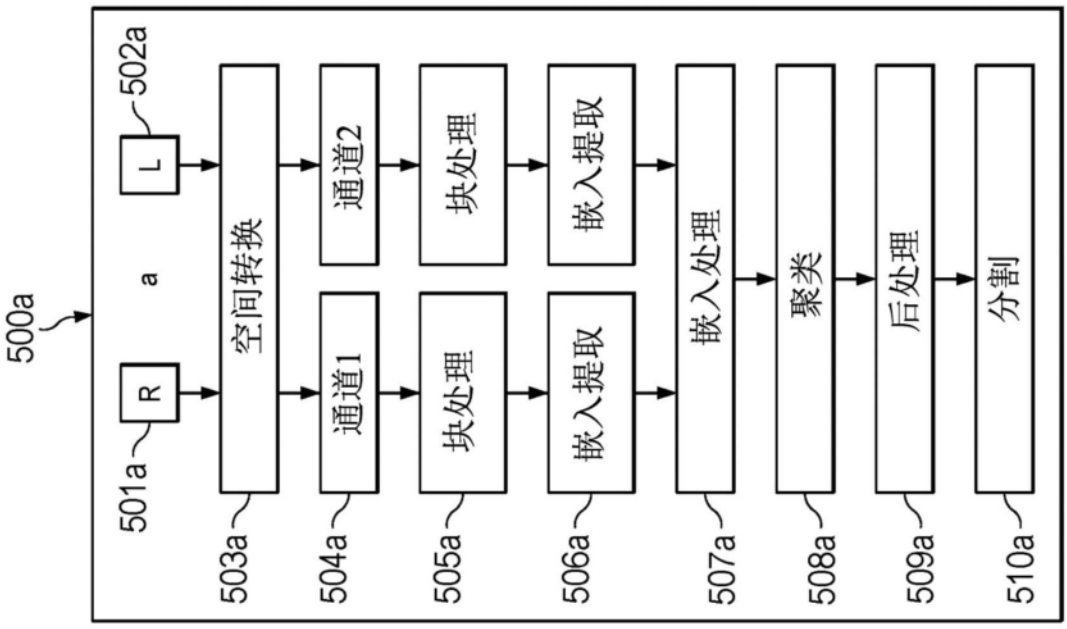

500a
a
L
502a
R
501a
空间转换
503a
通道2
504a
通道1
块处理
505a
块处理
嵌入提取
506a
嵌入提取
嵌入处理
507a
聚类
508a
后处理
509a
分割
510a

图5A

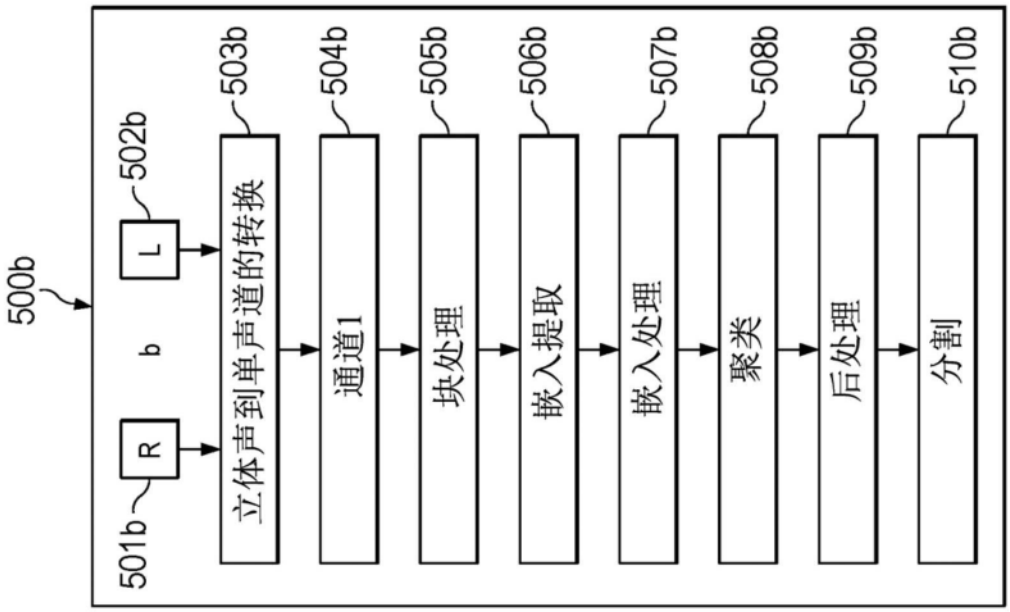

500b
501b
R
b
502b
L
503b
立体声到单声道的转换
504b
通道1
505b
块处理
506b
嵌入提取
507b
嵌入处理
508b
聚类
509b
后处理
510b
分割

图5B

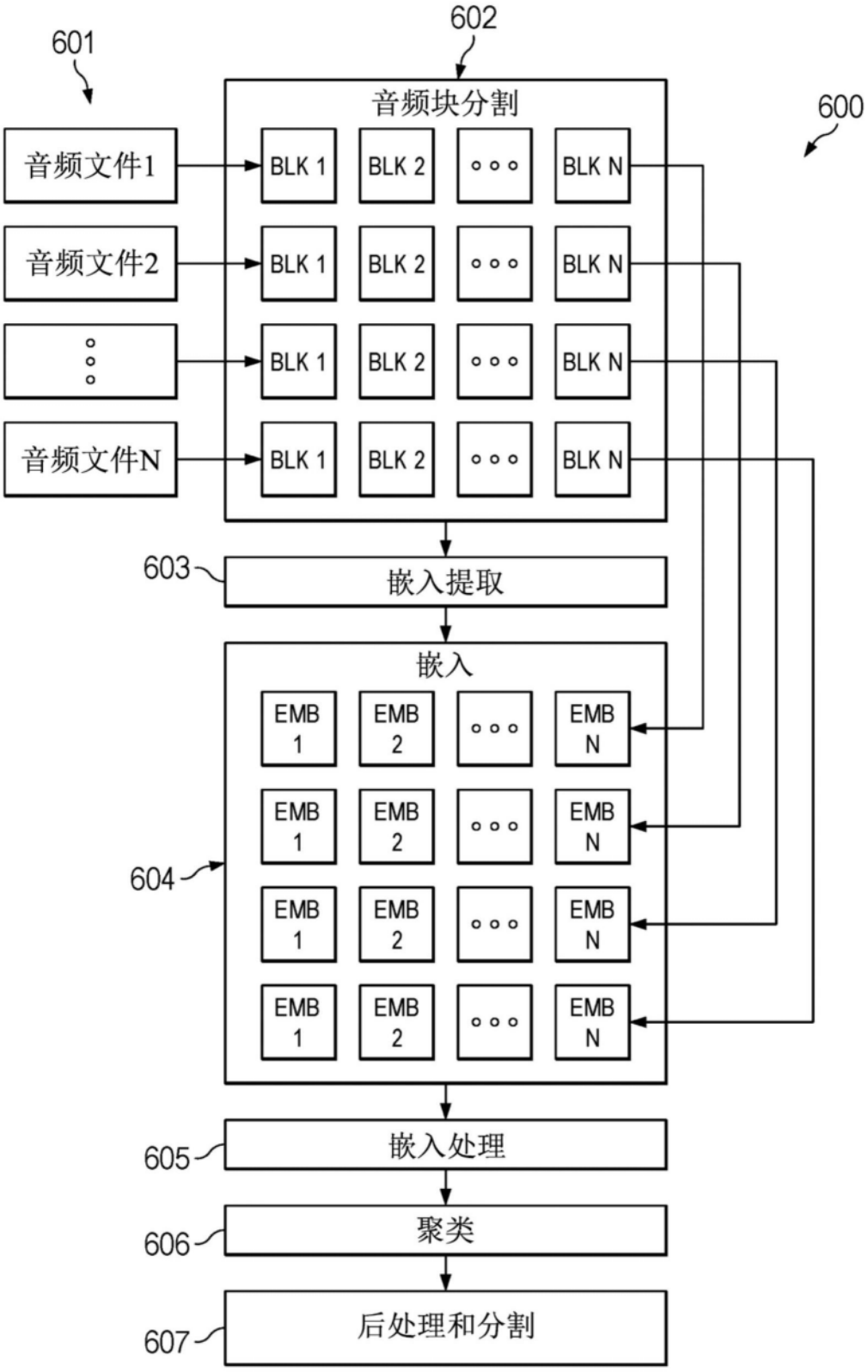

601
602
600
音频块分割
音频文件1
音频文件2
音频文件N
BLK 1
BLK 2
BLK N
603
嵌入提取
嵌入
EMB 1
EMB 2
EMB N
604
605
嵌入处理
606
聚类
607
后处理和分割

图6

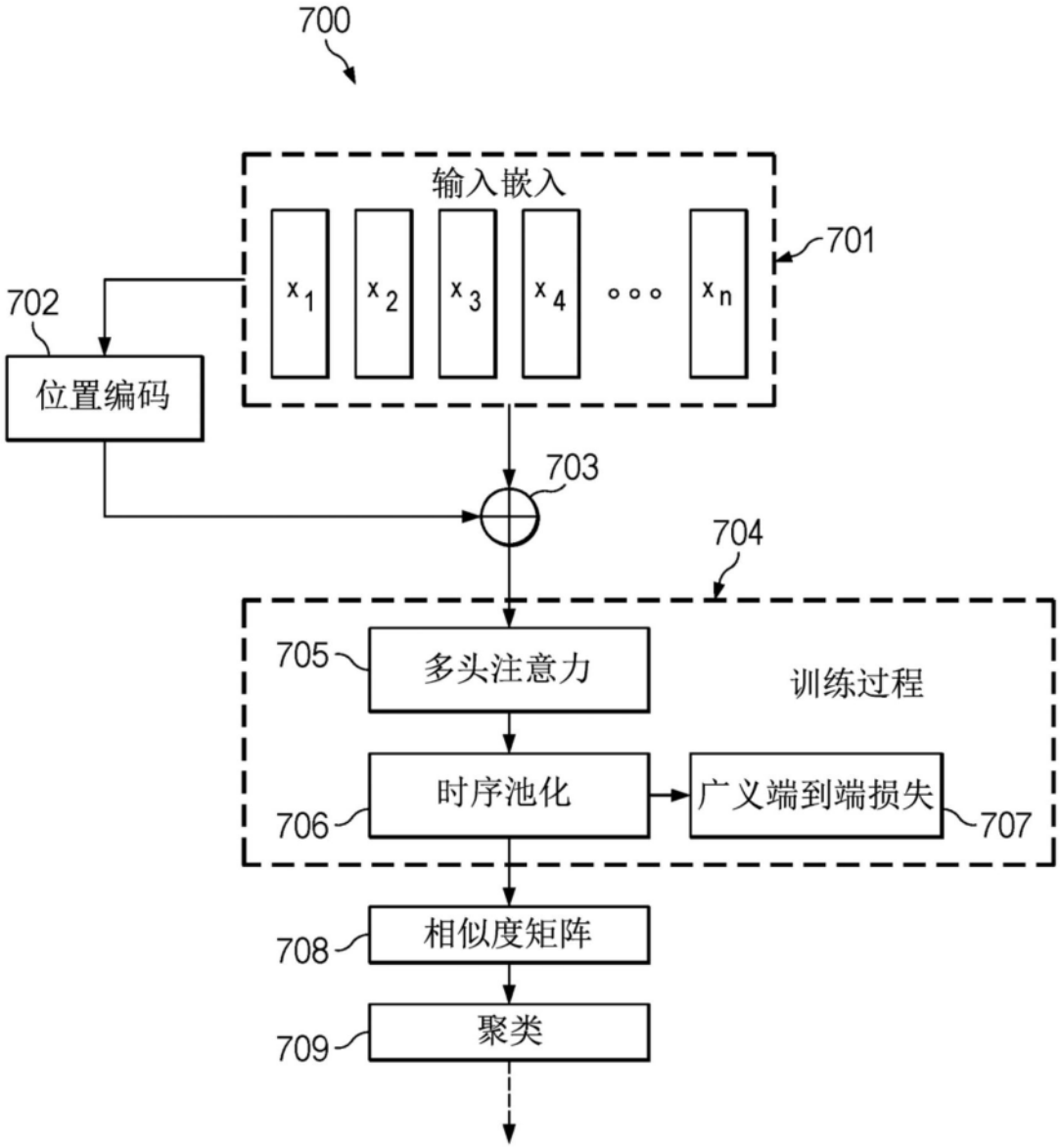

700
输入嵌入
701
x1
x2
x3
x4
ooo
xn
702
位置编码
703
704
705
多头注意力
训练过程
时序池化
广义端到端损失
706
707
708
相似度矩阵
709
聚类

图7

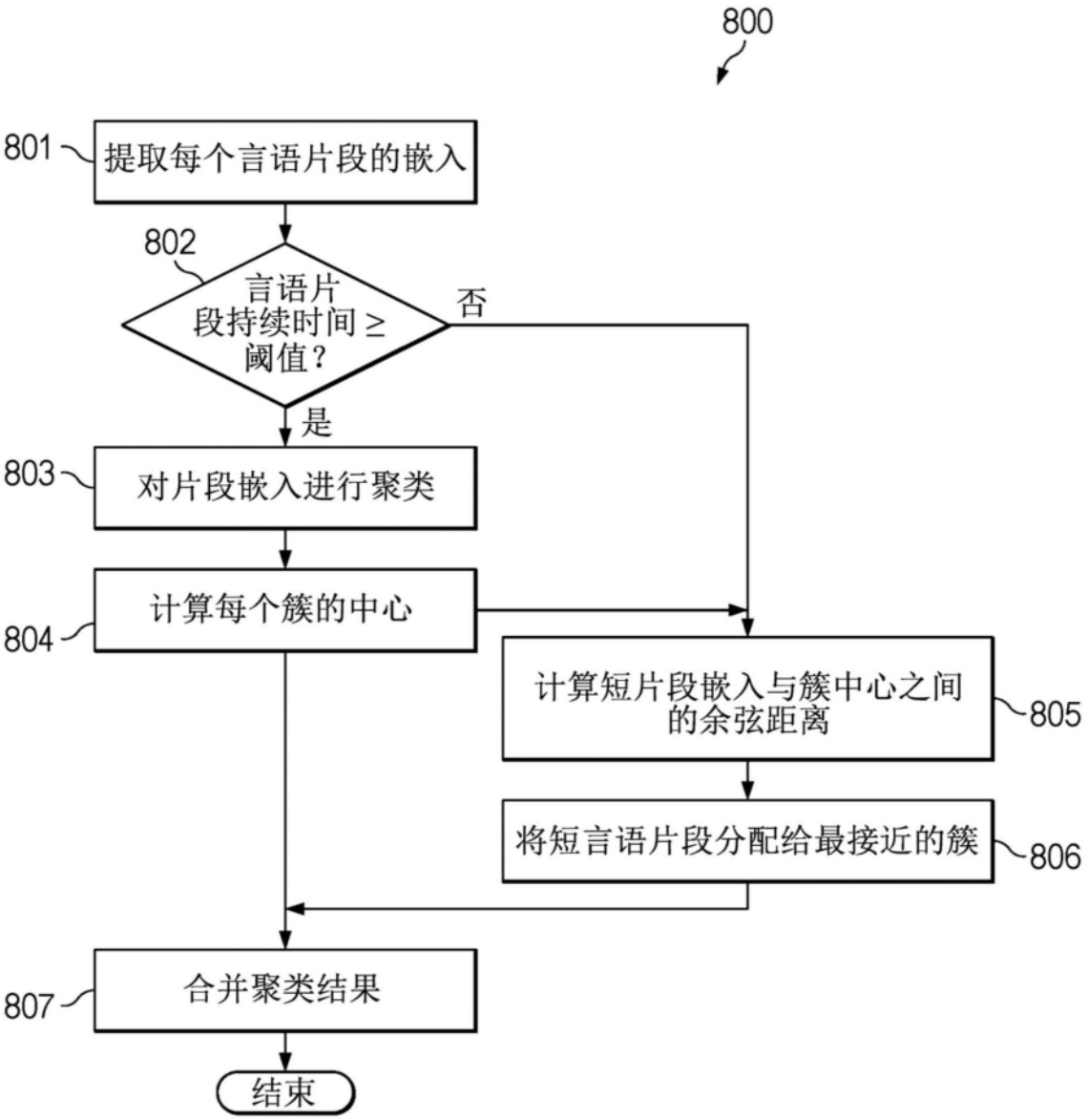

800
801
提取每个言语片段的嵌入
802
言语片段持续时间≥阈值?
否
是
803
对片段嵌入进行聚类
804
计算每个簇的中心
计算短片段嵌入与簇中心之间的余弦距离
805
将短言语片段分配给最接近的簇
806
807
合并聚类结果
结束

图8

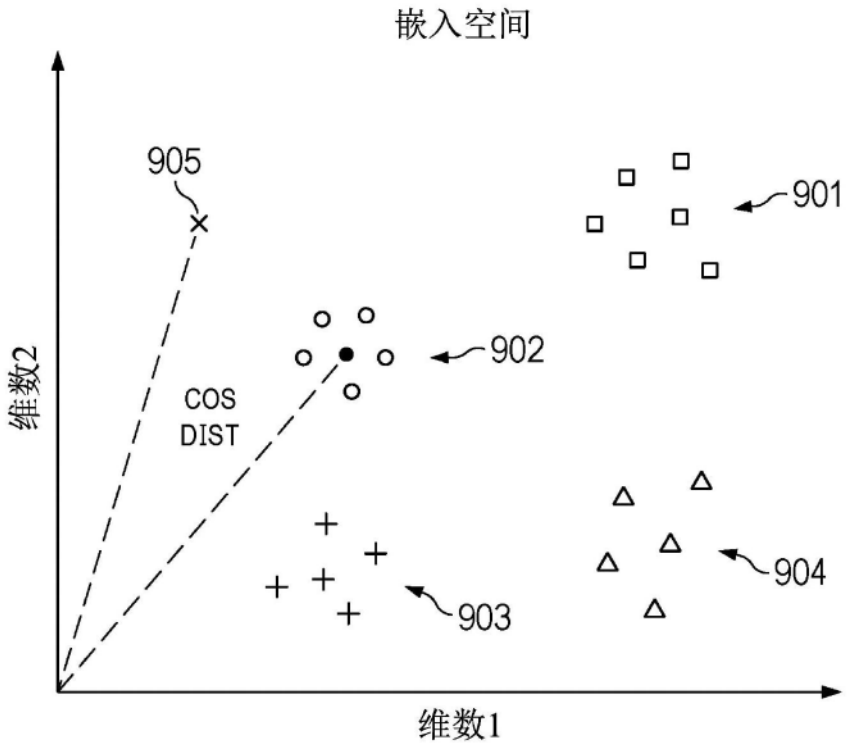

嵌入空间
905
901
902
COS
DIST
维数2
904
903
维数1

图9

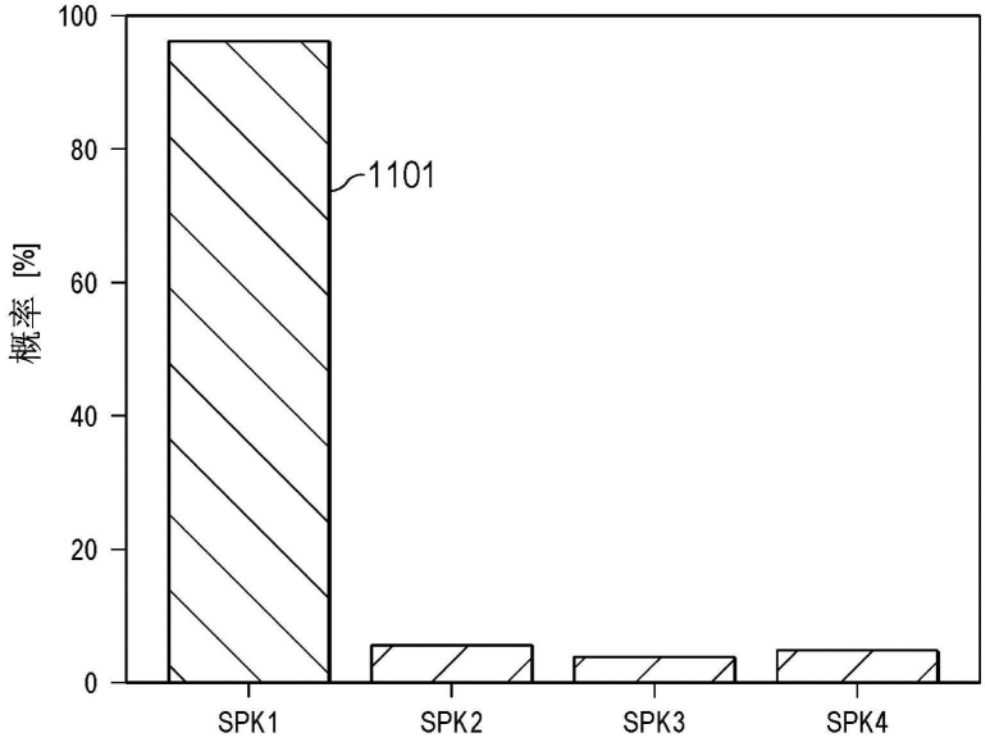

100
80
60
40
20
0
概率 [%]
1101
SPK1
SPK2
SPK3
SPK4

图11

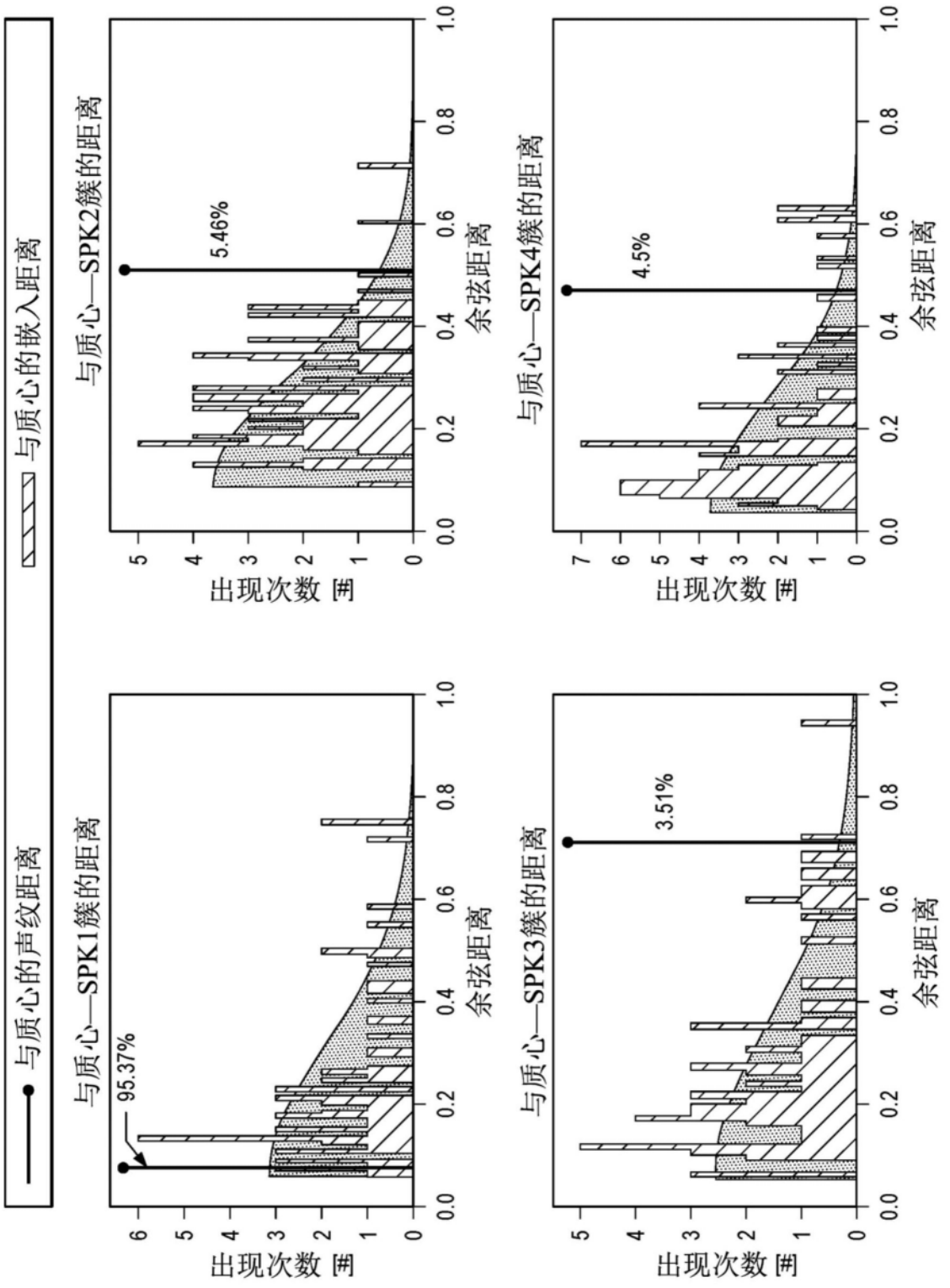

与质心的声纹距离
与质心的嵌入距离
与质心—SPK1簇的距离
95.37%
与质心—SPK2簇的距离
5.46%
与质心—SPK3簇的距离
3.51%
与质心—SPK4簇的距离
4.5%
出现次数 [#]
余弦距离

图10

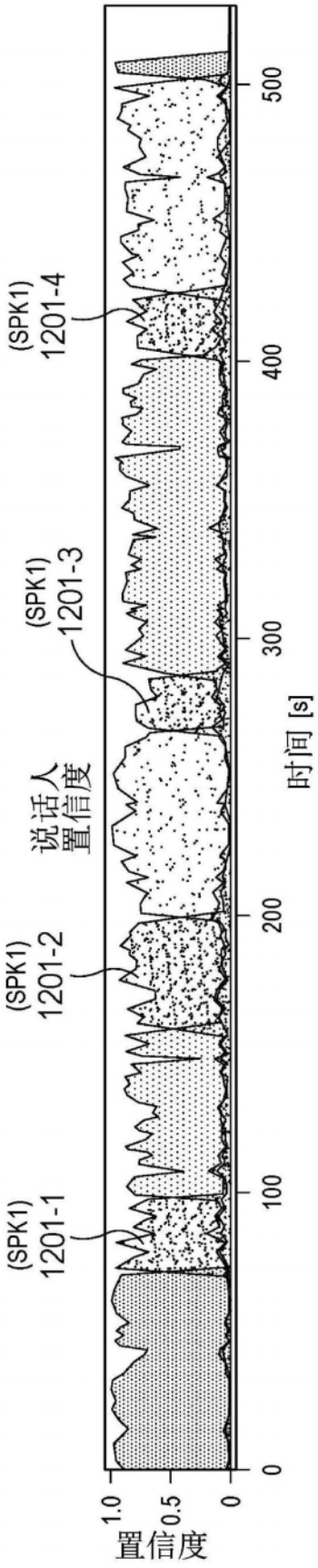

(SPK1)
1201-4
(SPK1)
1201-3
(SPK1)
1201-2
(SPK1)
1201-1
说话人
置信度
0
100
200
300
400
500
时间 [s]
1.0
0.5
0
置信度

图12

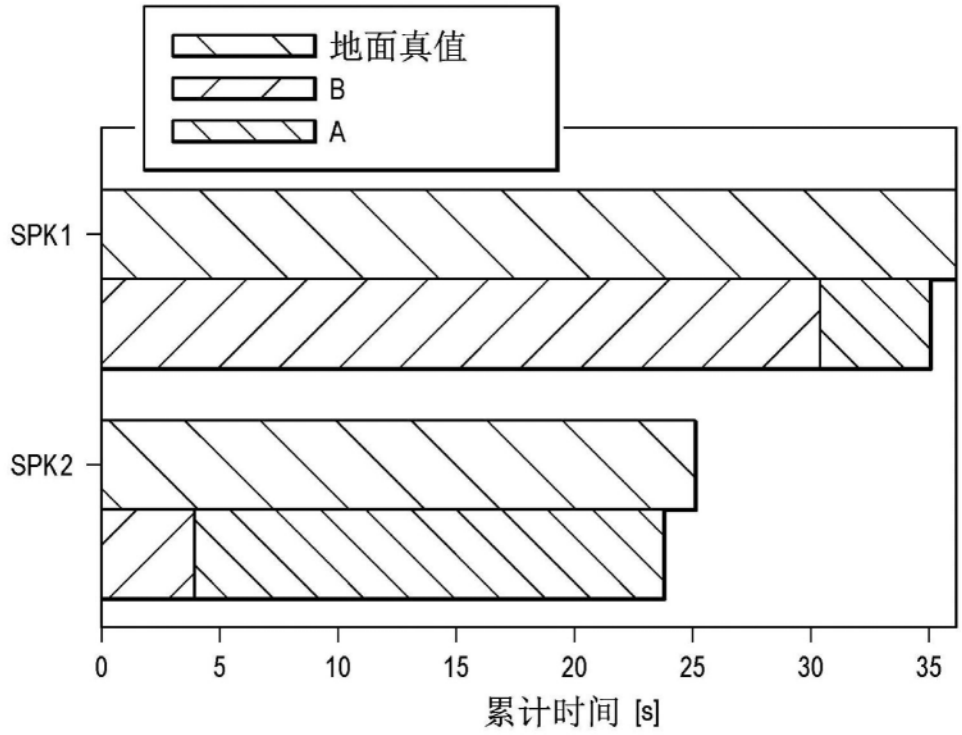

地面真值
B
A
SPK1
SPK2
0
5
10
15
20
25
30
35
累计时间 [s]

图13

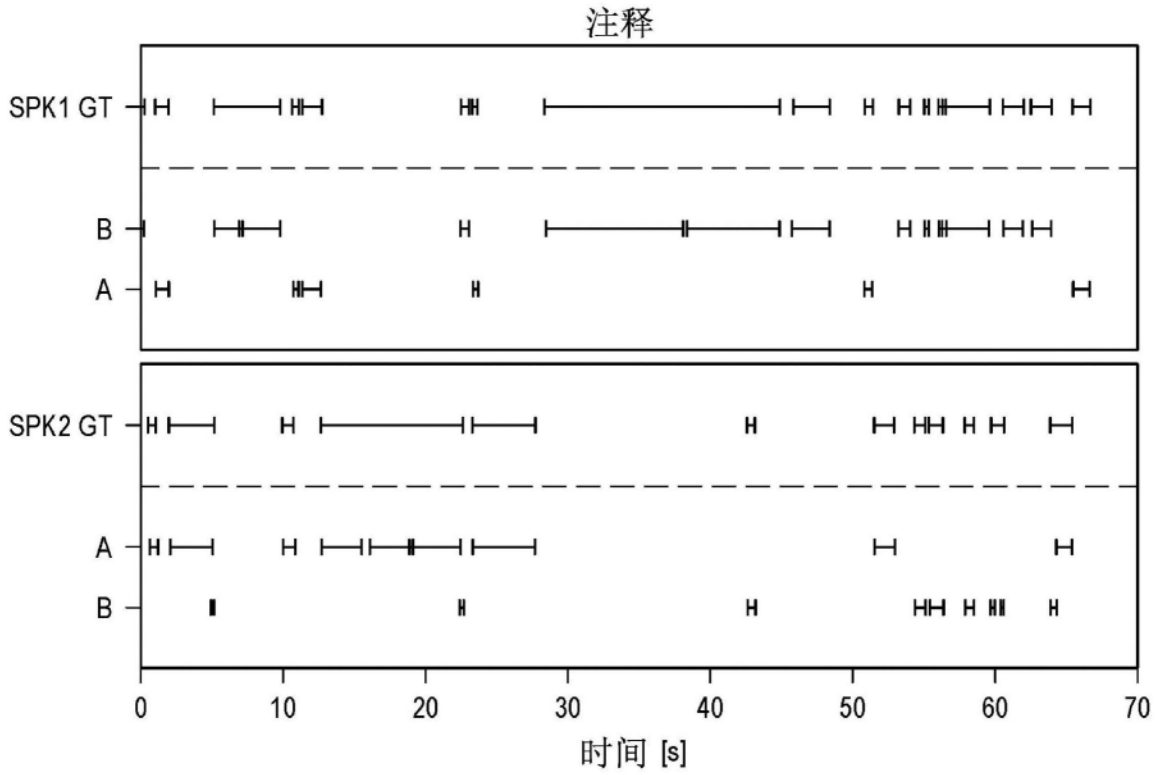

注释
SPK1 GT
B
A
SPK2 GT
A
B
0
10
20
30
40
50
60
70
时间 [s]

图14

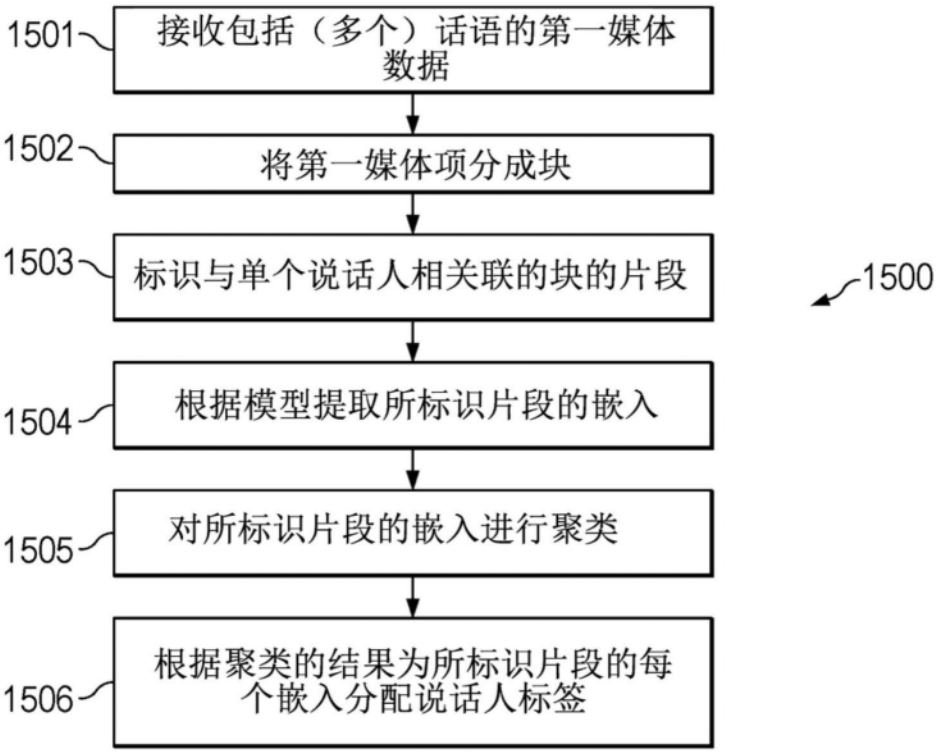

1501
接收包括（多个）话语的第一媒体数据
1502
将第一媒体项分成块
1503
标识与单个说话人相关联的块的片段
1500
1504
根据模型提取所标识片段的嵌入
1505
对所标识片段的嵌入进行聚类
1506
根据聚类的结果为所标识片段的每个嵌入分配说话人标签

图15

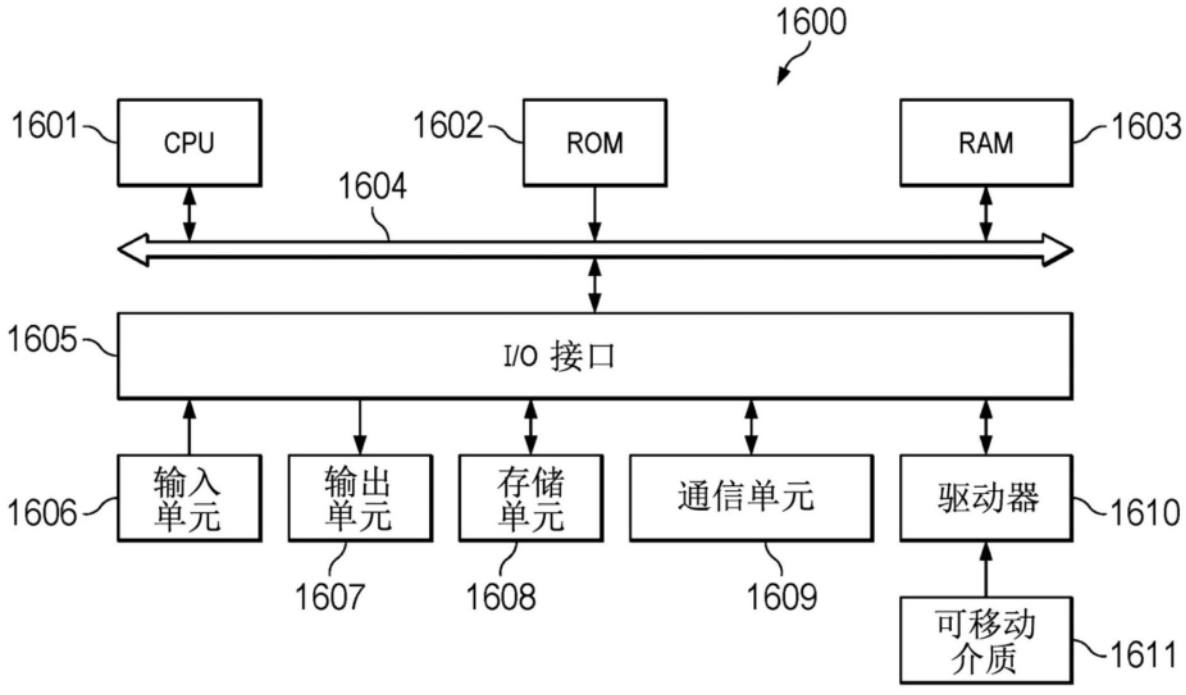

1600
1601
CPU
1602
ROM
RAM
1603
1604
1605
I/O 接口
1606
输入单元
输出单元
存储单元
通信单元
驱动器
1610
1607
1608
1609
可移动介质
1611

图16