(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120895028 A
(43)申请公布日 2025.11.04

(21)申请号 202511393740.0
(22)申请日 2025.09.28
(71)申请人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号
申请人 杭州市第七人民医院(杭州市心理危机研究与干预中心)
(72)发明人 江海腾 邓贵锋 饶姝颖 郭万军 李涛
(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
专利代理师 胡红娟
(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/26(2006.01)
G10L 25/78(2013.01)

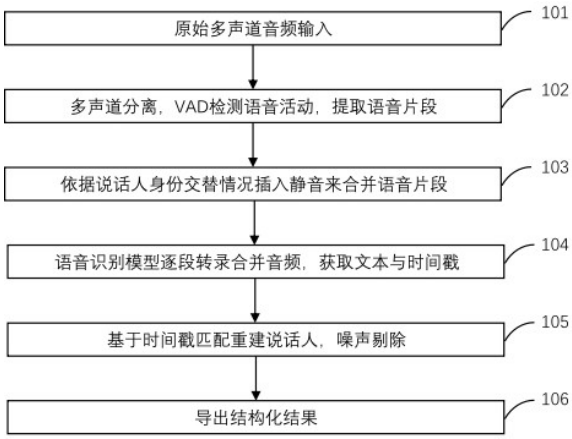
权利要求书1页 说明书7页 附图1页

(54)发明名称

基于单声道人工智能模型的多声道通话录音识别方法及装置

(57)摘要

本发明公开了一种基于单声道人工智能模型的多声道通话录音识别方法,包括:输入多声道通话音频数据;针对每个声道进行语音活动检测,以获取每个声道中的语音片段以及检测时对应的原始时间戳;基于原始时间戳的先后顺序对原始时间戳进行排序,构建一条单声道音频并记录合成时间戳;将构建获得的单声道音频输入至预训练的单声道语音识别模型,以生成识别文本序列并记录输出时间戳;基于合成时间戳和输出时间戳之间的重叠区间,以回溯匹配至原始时间戳;根据回溯匹配的结果构建包含说话人,时间戳以及识别文本的三元组。本发明还提供一种多声道通话录音识别装置。本发明提供的方法能实现在保持语义准确性的同时,识别说话人并重建通话逻辑顺序。



CN 120895028 A

1.一种基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,包括以下步骤:输入多声道通话音频数据,其包括多个声道以及每个声道内所包含至少一个说话人的语音数据;

针对每个声道进行语音活动检测,以获取每个声道中的语音片段以及检测时对应的原始时间戳;

基于原始时间戳的先后顺序对原始时间戳进行排序,并判断相邻两个语音片段是否来源同一声道,以选择不同时长的静音片段作为衔接从而构建一条单声道音频并记录合成时间戳;

将构建获得的单声道音频输入至预训练的单声道语音识别模型,以生成带有时间标注的识别文本序列并记录输出时间戳;

基于合成时间戳和输出时间戳之间的重叠区间,并采用最近邻对齐策略以回溯匹配至原始时间戳;

根据回溯匹配的结果构建包含说话人,时间戳以及识别文本的三元组。

2.根据权利要求1所述的基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,所述语音活动检测采用基于时间窗滑动的短时能量检测策略完成。

3.根据权利要求1所述的基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,所述不同时长的静音片段包括第一静音片段和第二静音片段;

当相邻两个语音片段来源于不同声道,则通过第一静音片段进行衔接;

当相邻两个语音片段来源于同一声道,则通过第二静音片段进行衔接;

所述第一静音片段的时长大于所述第二静音片段的时长。

4.根据权利要求3所述的基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,所述第一静音片段的时长选用1000毫秒,所述第二静音片段的时长选用50毫秒。

5.根据权利要求1所述的基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,所述预训练的单声道语音识别模型采用端到端语音识别模型。

6.根据权利要求1所述的基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,在构建包含说话人,时间戳以及识别文本的三元组时还需要进行数据过滤,所述数据过滤包括剔除由短时杂音、时间重叠识别冲突或无效空段导致的异常片段。

7.根据权利要求1或6所述的基于单声道人工智能模型的多声道通话录音识别方法,其特征在于,所述三元组的格式包括CSV、Excel或JSON格式。

8.一种多声道通话录音识别装置,其特征在于,用于执行如权利要求1~7任一项所述的基于单声道人工智能模型的多声道通话录音识别方法的步骤。

基于单声道人工智能模型的多声道通话录音识别方法及装置

技术领域

[0001] 本发明属于语音识别技术领域,尤其涉及一种基于单声道人工智能模型的多声道通话录音识别方法及装置。

背景技术

[0002] 近年来,语音识别技术在深度学习推动下取得了显著进展;其代表性模型包括Facebook提出的自监督预训练模型wav2vec2.0和HuBERT,以及OpenAI推出的Whisper模型和阿里巴巴开源的SenseVoice模型等端到端语音识别系统。这些模型具备较强的鲁棒性,能够适应多语种、多场景的语言转写需求,已被广泛应用于字幕生成、语音助手、会议记录等场景。

[0003] 其中,现有的端到端语音识别模型(如Whisper、SenseVoice)仅支持单声道输入,对于多声道音频的处理常常存在以下不足:(1)无法准确还原对话内容中的说话人顺序;(2)多声道合并后直接识别易导致信息错位或时间重叠;(3)缺乏精细化的通话轮次结构划分与说话人标注机制。为实现识别“谁说了什么”的目标,需要额外引入说话人分离(Separation)与说话人分段(Diarization)模块,现有部分系统尝试使用端到端的说话人分离模型或说话人聚类算法来解决这一问题,但这类方法计算复杂度高、资源消耗大、可解释性差,难以部署于资源受限的真实场景中。

[0004] 与此同时,现实中大量语音通信数据以多声道方式采集,例如电话通话录音、线上会议等,每个声道对应一个说话人通道。这类多声道音频往往具有发言交叉、起止不整齐、说话时间不等长等特点,传统的单声道语音识别流程难以直接适配,容易引发文本错位、说话人识别错误和语义割裂等问题,严重影响识别结果的可用性。

[0005] 专利文献CN119296522A公开了一种多声道语音识别方法、装置、设备及介质,通过判断任意两个声道的有效语音段是否满足预先配置的信号混合条件,能够精确识别出哪些有效语音段之间可能发生了信号混合问题。信号混合条件结合了内容交集和时间交集两个方面的判断。内容交集确保了语音内容的相似性,而时间交集则确保了这种相似性是在同一时间段内发生的,通过该种信号混合条件,有利于精准识别出存在信号混合问题的两个有效语音段。当确定两个声道的有效语音段满足信号混合条件时,可以进一步比较内容交集在两个声道中的能量大小,并选择保留能量较大的语音段的内容交集。

[0006] 专利文献CN111883132A公开了一种语音识别方法、设备、系统及存储介质,包括:获取待识别的语音信号,其中,所述待识别的语音信号包括多声道语音信号,所述多声道语音信号中包括至少两个声道的语音信号;将所述多声道语音信号进行声道分离,得到至少两路第一单声道语音信号;获得所述第一单声道语音信号对应的至少一个第一语音片段;获得所述第一语音片段的第一语音识别结果;根据所述第一语音识别结果,得到所述第一单声道语音信号的语音识别结果。

发明内容

[0007] 本发明的目的在于提供一种基于单声道人工智能模型的多声道通话录音识别方法及装置,该方法能实现在保持语义准确性的同时,识别说话人并重建通话逻辑顺序。

[0008] 为了实现本发明的第一个目的,提供如下技术方案:一种基于单声道人工智能模型的多声道通话录音识别方法,包括以下步骤:

输入多声道通话音频数据,其包括多个声道以及每个声道内所包含至少一个说话人的语音数据;

针对每个声道进行语音活动检测,以获取每个声道中的语音片段以及检测时对应的原始时间戳;

基于原始时间戳的先后顺序对原始时间戳进行排序,并判断相邻两个语音片段是否来源同一声道,以选择不同时长的静音片段作为衔接从而构建一条单声道音频并记录合成时间戳;

将构建获得的单声道音频输入至预训练的单声道语音识别模型,以生成带有时间标注的识别文本序列并记录输出时间戳;

基于合成时间戳和输出时间戳之间的重叠区间,并采用最近邻对齐策略以回溯匹配至原始时间戳;

根据回溯匹配的结果构建包含说话人,时间戳以及识别文本的三元组。

[0009] 本发明通过通过语音活动检测(VAD)、静音插入合成策略、单声道语音识别模型识别与时间戳匹配算法,将原语音数据与识别文本进行准确配对,从而输出具有逻辑顺序的语音识别结果。

[0010] 具体的,所述语音活动检测采用基于时间窗滑动的短时能量检测策略完成,以过滤非语音段落,有效分离清晰语句区间。

[0011] 具体的,所述不同时长的静音片段包括第一静音片段和第二静音片段;

当相邻两个语音片段来源于不同声道,则通过第一静音片段进行衔接;

当相邻两个语音片段来源于同一声道,则通过第二静音片段进行衔接;

所述第一静音片段的时长大于所述第二静音片段的时长。

[0012] 具体的,所述第一静音片段的时长选用1000毫秒,所述第二静音片段的时长选用50毫秒。

[0013] 具体的,所述预训练的单声道语音识别模型采用端到端语音识别模型进行构建。

[0014] 具体的,在构建包含说话人,时间戳以及识别文本的三元组时还需要进行数据过滤,所述数据过滤包括剔除由短时杂音、时间重叠识别冲突或无效空段导致的异常片段。

[0015] 具体的,所述三元组的格式包括CSV、Excel或JSON格式。

[0016] 为了实现本发明的第二个目的,提供了如下技术方案:一种多声道通话录音识别装置,用于执行如上述的基于单声道人工智能模型的多声道通话录音识别方法的步骤。

[0017] 与现有技术相比,本发明的有益效果:

避免对多声道数据进行复杂建模,仅使用单声道语音识别模型(如Whisper和SenseVoice等)即可完成准确识别,无需额外训练或微调,兼容性与部署效率高;

利用插入不同长度静音片段的策略有效增强说话人区分度,提升语音识别准确率;

通过时间戳比对策略实现说话人重建,提升对话结构还原能力,适用多种场景。

附图说明

[0018] 图1为本实施例提供的基于单声道人工智能模型的多声道通话录音识别方法的流程图;

图2为本实施例提供的多声道音频向单声道音频合成及时间戳匹配过程的示意图。

具体实施方式

[0019] 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,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。通常在此处附图中描述和示出的本发明实施例的组件可以以各种不同的配置来布置和设计。因此,以下对在附图中提供的本发明的实施例的详细描述并非旨在限制要求保护的本发明的范围,而是仅仅表示本发明的选定实施例。基于本发明中的实施例,本领域普通技术人员在没有做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。

[0020] 如图1所示,为本实施例提供的一种基于单声道人工智能模型的多声道通话录音识别方法,包括以下具体步骤:

步骤S101、输入多种声道通话语音数据,每个声道记录一位说话人完整的语音内容,即将输入的多声道通话录音(如电话录音、远程会议录音等)加载至系统,每一声道通常对应一个独立的说话人发言。

[0021] 步骤S102、如图2所示,针对每个声道进行语音活动检测,以获取每个声道中的语音片段以及检测时对应的原始时间戳,即多声道分离后VAD算法检测语音活动,提取语音片段,将多个声道分别提取为独立音轨,基于语音活动检测算法检测语音起止位置,去除无声段与背景噪声,保留有效语音段,同时标记原始时间戳,该原始时间戳在图2中通过T1表示。

[0022] 步骤S103、基于原始时间戳的先后顺序对原始时间戳进行排序,并判断相邻两个语音片段是否来源同一声道,以选择不同时长的静音片段作为衔接从而构建一条单声道音频并记录合成时间戳,即将所有提取出的语音片段根据时间顺序重新组合为一条单声道音轨,若前后片段来自不同声道(即不同说话人),则在片段之间插入预设长度的第一静音片段;若连续片段来自同一声道(即相同说话人),则插入较短的第二静音片段,以模拟自然停顿,并记录新合成音频的合成时间戳,该合成时间戳在图2中通过T2表示。

[0023] 步骤S104、基于合成时间戳和输出时间戳之间的重叠区间,并采用最近邻对齐策略以回溯匹配至原始时间戳,根据回溯匹配的结果构建包含说话人,时间戳以及识别文本的三元组,即将合成后的单声道音频输入至语音识别模型(如Whisper)进行语音转写,逐段获取转录文本,记录转录文本在合成音频中的输出时间戳,该输出时间戳在图2中通过T3表示。

[0024] 步骤S105、基于合成合成时间戳(T2)与单声道语音识别模型输出时间戳(T3)之间的重叠区间和最近邻对齐策略,回溯匹配至原始时间戳(T1)对比,将转录文本归类至对应声道及说话人,并对转录内容中不符合上下文的短噪声或无效识别段进行剔除,形成“说话

人ID-起止时间-文本”三元组结构。

[0025] 步骤S106、导出结构化结果。将最终识别出的说话人标签、起止时间、转录文本等信息导出为结构化格式(如CSV或JSON),用于后续的数据分析、质检或智能问答等应用场景。

[0026] 在本实施例中所提及的三元组结构可导出为标准数据格式(如CSV、Excel、JSON等),供后续系统使用。

[0027] 本实施例还提供了一种多声道通话录音识别装置,用于执行上述实施例提出的基于单声道人工智能模型的多声道通话录音识别方法的步骤,其包括输入单元,数据处理单元以及输出单元。

[0028] 该输入单元,用于输入多声道通话音频数据。

[0029] 该数据处理单元,用于将输入的多声道通话音频数据进行数据重构,以生成“说话人ID-起止时间-文本”三元组结构的语音数据。

[0030] 该输出单元,根据用户选择的文件形式将数据处理单元生成的语音数据进行结构化输出。

[0031] 通过上述步骤,本发明实现了从原始多声道录音中自动提取出有语义结构的文本及说话人序列,极大地简化了多声道转录与对话结构重建的流程,适用于电话通话录音、远程会诊、司法调查记录、线上教育等实际应用场景。

[0032] 为使本发明的技术方案、实施过程及技术效果更加清晰明确,以图1、图2以及某心理援助热线的实际应用场景,对本实施例所提供的基于单声道人工智能模型的多声道通话录音识别方法进行详细说明。

[0033] 本实施例以该热线的双声道通话录音处理为具体应用场景,所述双声道分别对应热线接线员与求助者,录音格式为MP3,采样率16kHz,位深16bit。

[0034] 本实施例依托某心理援助热线的真实通话数据开展,其通话录音采用双声道采集模式(声道1对应接线员,声道2对应求助者),具有发言交叉、起止不整齐、包含背景噪声等典型多声道通话特征。为验证本发明方法的有效性,在2023年1月1日至12月31日期间的21527个热线电话样本中随机选取60个(.mp3格式),截取每个样本前5分钟有效音频,总时长300分钟。由2名具备心理热线经验的专业人员对音频进行人工标注,标注内容包括说话人身份(接线员/求助者)、每段语音的起止时间戳(精确到毫秒)及对应文本内容,作为性能评估的基准数据。

[0035] 实验硬件环境为Intel Xeon Gold 6330 CPU、NVIDIA A800 GPU,软件环境基于Python 3.9,依赖Librosa库进行音频处理、Whisper-large-v3模型(端到端单声道语音识别模型)进行文本转写,时间戳匹配算法基于NumPy实现。

[0036] 具体执行过程如下:

步骤1:多声道通话音频数据输入,通过音频输入接口加载60个双声道热线录音样本,系统自动解析音频的声道数量、采样率等元数据,将声道1(接线员)与声道2(求助者)的语音数据分离为独立的音频流,每个音频流关联唯一的声道标识(ID1对应接线员,ID2对应求助者)。

[0037] 步骤2:语音活动检测与原始时间戳提取(对应图1中102、图2中T1),针对分离后的每个声道音频流,采用基于时间窗滑动的短时能量检测策略执行语音活动检测(VAD):设置

时间窗长度为20ms,滑动步长为10ms,计算每个时间窗内的短时能量值,将能量阈值设为该声道背景噪声能量均值的1.5倍。当连续3个时间窗的能量值超过阈值时,判定为语音起始;当连续5个时间窗的能量值低于阈值时,判定为语音结束。

[0038] 通过上述检测,提取每个声道中的有效语音片段,同步记录各片段的原始时间戳T1(包含起始时间T1_start与结束时间T1_end)。原始时间戳以音频起始时刻为时间原点(0ms),精确到毫秒级。检测过程中自动过滤纯噪声段(能量低于阈值的连续片段)。

[0039] 步骤3:合成声道构建与合成时间戳记录(对应图1中103、图2中合成声道与T2);

语音片段排序:将两个声道提取的所有有效语音片段按原始时间戳T1_start的升序进行全局排序,形成统一的片段序列;

静音片段选择与插入:遍历排序后的片段序列,判断相邻两个片段的声道来源。若相邻片段来源于不同声道(如前一片段为ID1,后一片段为ID2),插入1000ms的第一静音片段;若相邻片段来源于同一声道(如连续两个片段均为ID1),插入50ms的第二静音片段;

合成时间戳生成:将排序后的语音片段与插入的静音片段依次拼接,形成单声道音频流,同时记录合成时间戳T2——即每个原始语音片段在合成音频中的起始时间T2_start与结束时间T2_end,静音片段不分配说话人标识,仅作为片段分隔标记。合成过程的时间轴映射关系如图2所示,声道1与声道2的语音片段通过不同长度静音衔接后整合为单一合成声道。

[0040] 步骤4:单声道语音识别与输出时间戳获取(对应图1中104、图2中T3),将构建的单声道音频输入预训练的Whisper-large-v3端到端语音识别模型,设置识别语言为中文,模型输出带有时间标注的识别文本序列,同步记录每个文本片段在合成音频中的输出时间戳T3(包含T3_start与T3_end)。该模型无需针对多声道场景额外微调,直接复用预训练权重,符合兼容性设计要求。

[0041] 步骤5:时间戳回溯匹配(对应图1中105、图2中T1-T2-T3映射)

采用最近邻对齐策略实现时间戳匹配:计算合成时间戳T2与输出时间戳T3的重叠区间,重叠率计算公式为:Overlap = (min(T2_end, T3_end) - max(T2_start, T3_start)) / (T2_end - T2_start),保留重叠率≥50%的匹配对;

对每个匹配对,通过合成时间戳T2反向追溯至对应的原始语音片段,获取其原始时间戳T1及所属声道标识(说话人ID),完成“识别文本-合成时间戳-原始时间戳-说话人”的关联映射。

[0042] 步骤6:三元组构建与数据过滤(对应图1中106),其包括:

数据过滤:用于剔除除异常片段。

[0043] 短时杂音:剔除时长<100ms的识别文本片段。

[0044] 时间重叠冲突:当多个识别文本对应同一原始语音片段时,保留ROUGE-L(F1)最高的文本。

[0045] 无效空段:剔除纯标点、语气词或语义空白的文本片段。

[0046] 三元组生成:将过滤后的说话人ID、原始时间戳(T1_start、T1_end)及识别文本构建为结构化三元组,支持导出为JSON格式,示例如下:

```json
[
```

```
  {
    "说话人ID": "ID1",
    "起始时间": "00:00:12.350",
    "结束时间": "00:02:36.120",
    "识别文本": "您好,这里是心理援助热线,请问有什么可以帮您?"
  },
  {
    "说话人ID": "ID2",
    "起始时间": "00:01:05.780",
    "结束时间": "00:04:44.210",
    "识别文本": "我最近总是失眠,感觉生活没有意义……"
  }
]
```

。

[0047] 实施效果评估:以人工标注结果为基准,对本发明方法的性能进行定量评估,同时与“直接将原始多声道音频输入Whisper模型”的传统方式进行对比,评估指标及结果如表1所示。

表1

[0048]

评估指标	定义说明	本发明方法	传统直接识别方式
说话人识别准确率	正确匹配说话人身份的文本片段数 / 总有效文本片段数（总有效文本片段数为人工标注中非空文本片段总数）	0.9643	N/A（无法区分说话人，无身份输出）
CER（字符错误率）	（插入错误字符数 + 删除错误字符数 + 替换错误字符数）/ 人工标注文本总字符数（仅统计非空文本）	0.0857	0.3786
ROUGE-L(F1, 字符级)	识别文本与人工标注文本的最长公共子序列相似度，计算方式为 2×（精确率 × 召回率）/（精确率 + 召回率），综合衡量文本匹配度	0.9315	0.7723

注:评估数据基于60个热线样本的300分钟音频计算,所有指标均保留4位小数。

[0049] 由表1可知,本发明方法通过静音片段插入与时间戳匹配策略,实现了96.43%的说话人识别准确率,解决了传统单声道模型无法区分说话人的核心痛点;同时,CER较传统方式降低29.3%,ROUGE-L(F1)提升15.9%,证明其在文本转录准确性上的显著优势。

[0050] 本实施例基于某心理援助热线的真实多声道通话场景,完整呈现了本发明方法的

执行流程,结合附图1、图2明确了各步骤的技术细节与数据流转关系,并通过定量实验验证了方法的有效性。实施过程中未引入额外的模型训练或复杂硬件依赖,仅通过音频结构化重构与时间戳对齐策略,即实现了“说话人-时间-文本”的精准关联,充分体现了本发明兼容性强、部署成本低、识别效果优的技术优势,可广泛应用于电话录音、远程咨询等多声道语音转写场景。

[0051] 相比于传统依赖人工标注或高成本分离模型的方案,本发明具备如下优势:一是避免了说话人分离模型的训练依赖与计算负担;二是基于静音片段插入和时间戳匹配策略,在不引入高复杂度算法的前提下,实现了多轮对话说话人重建;三是结合Whisper等主流单声道语音识别模型,实现了高效、稳定的文本生成能力。此外,该方法结构灵活,模块划分清晰,适合部署于多种通话录音场景,实际使用中还可扩展接入情绪识别模型、文本摘要模型、多语言转写模块等,从而构建更加丰富的语音理解系统。

[0052] 本发明不仅在工程实施上具有较高的可操作性,在语义保真、说话人追踪、处理效率等关键指标上也具备显著性能优势,具备良好的产业转化前景。

[0053] 此外,术语“上”、“下”、“内”、“外”、“前”、“后”仅用于描述目的,而不能理解为指示或暗示相对重要性。除非另外具体说明,否则在这些实施例中阐述的部件和步骤的相对步骤、数字表达式和数值并不限制本发明的范围。

[0054] 当然,以上所述仅为本发明的具体实施例而已,并非来限制本发明实施范围,凡依本发明申请专利范围所述构造、特征及原理所做的等效变化或修饰,均应包括于本发明申请专利范围内。

[0055] 最后应说明的是:以上所述实施例,仅为本发明的具体实施方式,用以说明本发明的技术方案,而非对其限制,本发明的保护范围并不局限于此,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:任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特征进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应所述以权利要求的保护范围为准。

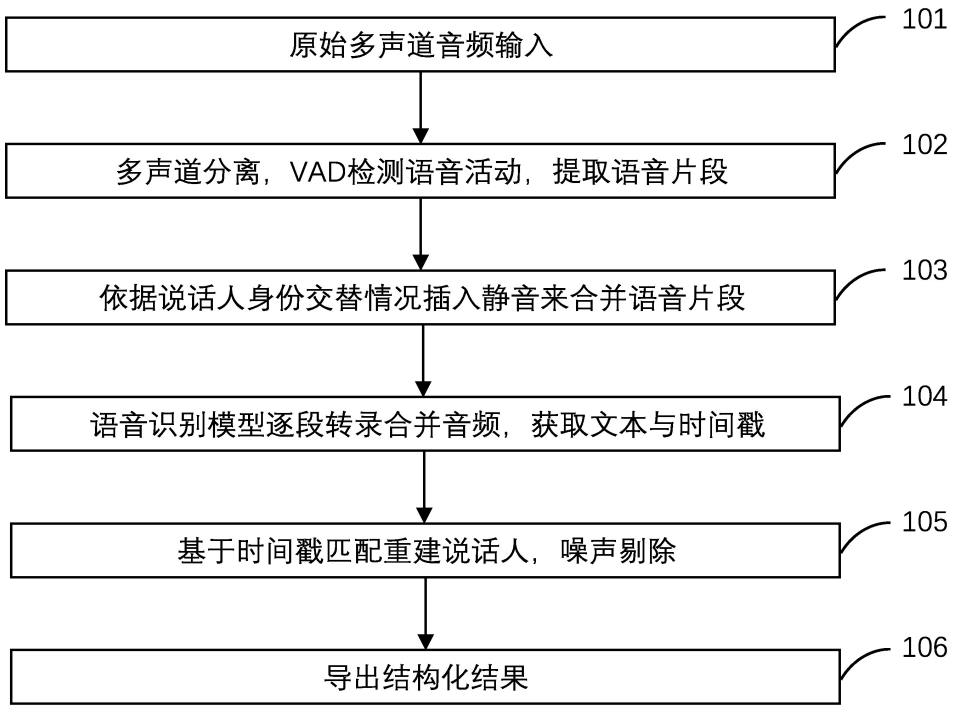

原始多声道音频输入
101
多声道分离，VAD检测语音活动，提取语音片段
102
依据说话人身份交替情况插入静音来合并语音片段
103
语音识别模型逐段转录合并音频，获取文本与时间戳
104
基于时间戳匹配重建说话人，噪声剔除
105
导出结构化结果
106

图 1

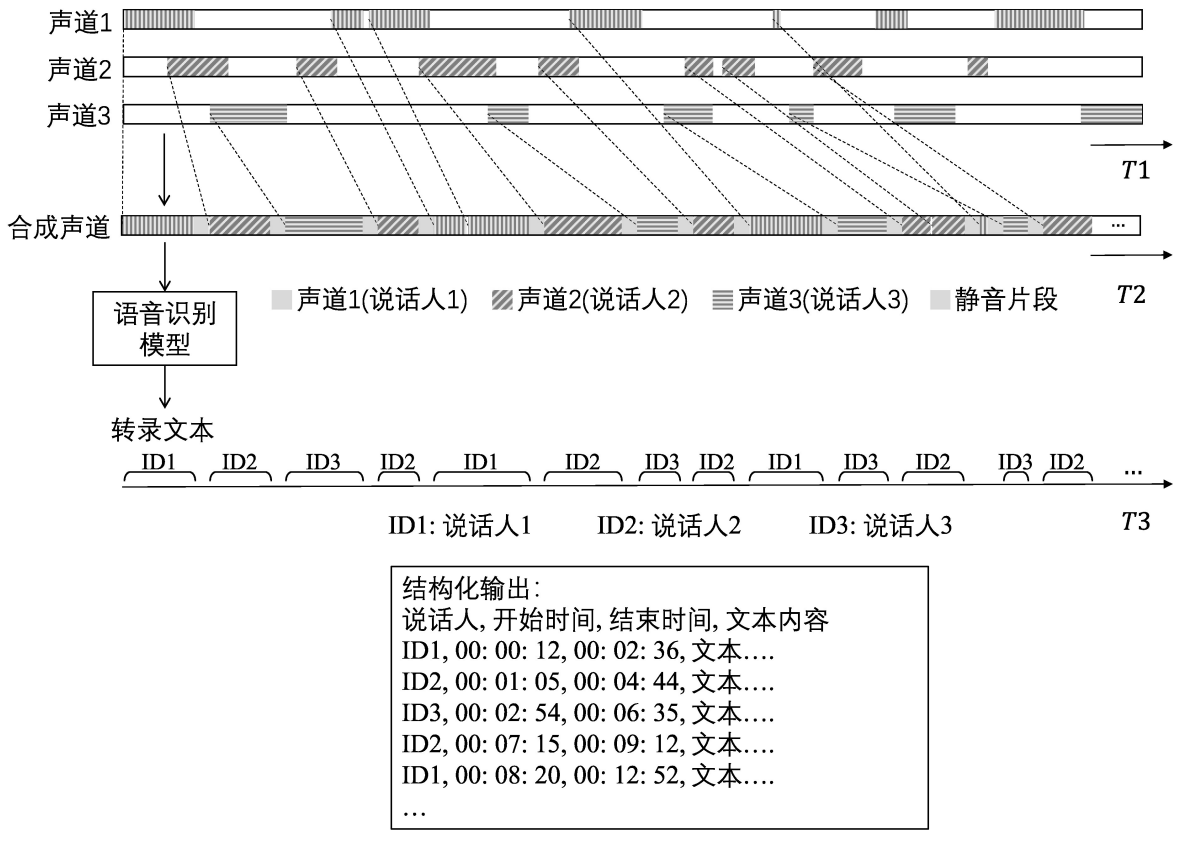

声道1
声道2
声道3
T1
合成声道
...
T2
语音识别模型
声道1(说话人1)
声道2(说话人2)
声道3(说话人3)
静音片段
转录文本
ID1 ID2 ID3 ID2 ID1 ID2 ID3 ID2 ID1 ID3 ID2 ID3 ID2 ...
ID1: 说话人1
ID2: 说话人2
ID3: 说话人3
T3
结构化输出:
说话人, 开始时间, 结束时间, 文本内容
ID1, 00: 00: 12, 00: 02: 36, 文本….
ID2, 00: 01: 05, 00: 04: 44, 文本….
ID3, 00: 02: 54, 00: 06: 35, 文本….
ID2, 00: 07: 15, 00: 09: 12, 文本….
ID1, 00: 08: 20, 00: 12: 52, 文本….
…

图 2