(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116312552 B
(45)授权公告日 2023.08.15

(21)申请号 202310569405.6

(22)申请日 2023.05.19

(65)同一申请的已公布的文献号
申请公布号 CN 116312552 A

(43)申请公布日 2023.06.23

(73)专利权人 湖北微模式科技发展有限公司
地址 430074 湖北省武汉市东湖开发区关东科技工业园七号地块

(72)发明人 申意萍 陈友斌 张志坚 徐一波

(74)专利代理机构 湖北高韬律师事务所 42240
专利代理师 鄢志波

(51)Int.Cl.
G10L 17/02(2013.01)
G10L 17/04(2013.01)
G10L 17/14(2013.01)
G10L 17/18(2013.01)
G10L 25/51(2013.01)
G10L 25/63(2013.01)

(56)对比文件
CN 114282621 A,2022.04.05
KR 20200129934 A,2020.11.18
US 2015088513 A1,2015.03.26
US 2019333522 A1,2019.10.31
US 2022321350 A1,2022.10.06
CN 114299953 A,2022.04.08
CN 114125365 A,2022.03.01
US 2020410265 A1,2020.12.31
CN 112906544 A,2021.06.04
CN 115937726 A,2023.04.07
CN 115050375 A,2022.09.13
Nishtha H. Tandel.Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey.2020 6th International Conference on Advanced Computing and Communication Systems.2020,第459-461页.

审查员 徐晶

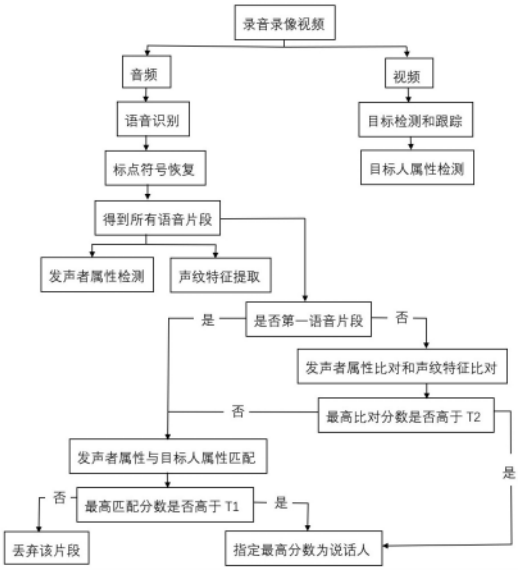
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种视频说话人日志方法及系统

(57)摘要

本发明提供了一种视频说话人日志方法及系统,所提出的方法将一段录音录像视频分离为音频部分和视频部分,一方面利用语音识别技术,对整个音频部分进行分割,得到仅包含单个说话人的语音片段,对每个语音片段进行发声者属性检测获取发声者属性信息并提取声纹特征;另一方面对视频部分中的人物目标进行目标检测和跟踪,得到目标人属性信息;最后结合发声者属性信息和目标人属性信息的匹配结构以及声纹特征比对来确定说话人。用于实现上述方法的视频说话人日志系统包括录音录像装备、存储器、处理器、显示设备、扬声器和输入设备。使用该方法及系统最终生成的说话人日志不仅包含说话人时间起止信息,还包含说话人图像信息和说话文字信息。



CN 116312552 B

1.一种视频说话人日志方法,该方法包括以下步骤:

S1、将录音录像视频分离为音频部分和视频部分;

S2、对音频部分,利用语音识别技术进行分割,得到仅包含单个说话人的语音片段,对每个语音片段进行发声者属性检测获取发声者属性信息并提取声纹特征,所述发声者属性信息包括发声者的年龄、性别、情绪;

S3、对视频部分,利用目标检测和跟踪技术得到每个人在画面中的时间,并记录目标ID,对每一个目标获取目标人属性信息,所述目标人属性信息包括目标人的年龄、性别、情绪;

S4、从第一语音片段开始,融合发声者属性信息与目标人属性信息的匹配结果以及声纹特征比对和发声者属性信息比对的结果,确定相应语音片段的说话人,得到最终的视频说话人日志;

S41、对第一个语音片段,对该片段持续期间所有在视频中出现的目标,取目标人属性信息与该片段的发声者属性信息进行匹配,根据匹配结果确定该语音片段说话人,同时保存该语音片段的声纹特征、语音长度、发声者属性信息和目标ID;

S42、对于后续每个语音片段,选取之前出现过的说话人,进行声纹特征比对和发声者属性信息比对,根据比对结果判断是否为之前出现过的说话人,如果是之前出现过的说话人,则可确定为该语音片段的说话人;

S43、如果不是之前出现过的说话人,先根据目标ID去掉之前出现过的说话人,对剩下的目标,进行发声者属性信息和目标人属性信息匹配,根据匹配结果确定该片段的说话人,同时保存该片段的声纹特征、语音长度、发声者属性信息和目标ID。

2.根据权利要求1所述的一种视频说话人日志方法,其特征在于:在步骤S2中,所述利用语音识别技术进行分割是先将语音信息转化为无标点的文本信息,然后恢复文本信息的标点符号,根据标点符号将音频部分分割为包含单个说话人的语音片段。

3.根据权利要求1所述的一种视频说话人日志方法,其特征在于:在步骤S3中,所述目标检测是人脸目标、人头目标、半身目标或全身目标,所述每个人在画面中的时间是指每个人在画面中的出现时间和持续时间。

4.一种基于上述权利要求1-3任一项方法的视频说话人日志系统,其特征在于:该系统包括录音录像设备、存储器、处理器、显示设备、扬声器和输入设备,所述录音录像设备分别与所述存储器、所述处理器、所述扬声器、所述显示设备连接,所述处理器还要与所述输入设备、所述存储器、所述扬声器和所述显示设备连接。

5.根据权利要求4所述的一种视频说话人日志系统,其特征在于:所述录音录像设备用于录制对话中的音频和视频。

6.根据权利要求4所述的一种视频说话人日志系统,其特征在于:所述存储器用于存储录制的音频和视频数据以及视频说话人日志可执行程序。

7.根据权利要求6所述的一种视频说话人日志系统,其特征在于:所述处理器用于执行所述视频说话人日志可执行程序,并把程序执行结果返回给所述存储器或者所述显示设备。

8.根据权利要求4所述的一种视频说话人日志系统,其特征在于:所述显示设备负责显示录制的视频和/或程序执行结果。

9.根据权利要求4所述的一种视频说话人日志系统,其特征在于:所述扬声器负责播放录制的音频。

10.根据权利要求4所述的一种视频说话人日志系统,其特征在于:所述输入设备用于输入一些指令,以控制程序的运行、停止和显示结果。

一种视频说话人日志方法及系统

技术领域

[0001] 本申请涉及音视频识别及分析技术领域，特别是涉及一种视频说话人日志方法及系统。

背景技术

[0002] 随着摄像头和麦克风设备的广泛普及，越来越多的领域使用这些设备进行录音和录像，获取现场音视频数据，作为存档证据或者自动稽核的数据。如医疗问诊、理财销售、保险销售、智能会议记录、智慧司法认罪认罚等。为了更好的理解这些音视频数据，需要生成说话人日志。一种通用的方法是对语音信号进行声纹聚类分割，具体流程为通过VAD技术提取语音信号，再利用信号分割(或者说话人转换检测)分成仅包含单个说话人的片段，然后对片段提取声纹特征，最后对声纹特征进行聚类得到说话人的日志。这种方法存在五个问题，第一，聚类分割的结果依赖于信号分割技术(或者说话人转换技术)，分割得过细会得到过短的语音片段，而在过短的语音片段上提取声纹特征会导致声纹信息不足，从而影响最终的聚类结果；如果分割的过粗有可能出现单个片段出现多个说话人。第二，过短的片段(如单字或双字回答)也会导致声纹信息不足。第三，当无法得知说话人数量时，聚类的结果往往不如人意。第四，在已知说话人数量的条件下，若说话长度极度不平衡，如某个人占据大量的长片段的发言，而剩下的人只有极少量的短句发言，也会导致聚类结果不理想甚至失败。第五，无法将声音的日志信息与视频的说话人对应起来。

发明内容

[0003] 针对现有技术存在的上述问题，本发明提出一种视频说话人日志方法及系统，所提出的方法一方面利用语音识别技术对整个语音进行分割，得到仅包含单个说话人的语音片段，提取语音片段的声纹特征和进行发声者属性检测；另一方面对视频中的目标进行检测和跟踪，对目标人进行属性检测；结合发声者属性和视频目标人属性匹配以及声纹特征比对来判定说话人。生成的说话人日志不仅包含说话人时间起止信息，还包含说话人图像信息和说话文字信息。所提出的系统可以实现上述功能，这个系统包括录音录像装备、存储器、处理器、显示设备、扬声器和输入设备。本发明的具体技术方案如下：

[0004] 一种视频说话人日志方法，该方法包括以下步骤：

[0005] S1、将录音录像视频分离为音频部分和视频部分；

[0006] S2、对音频部分，利用语音识别技术进行分割，得到仅包含单个说话人的语音片段，对每个语音片段进行发声者属性检测获取发声者属性信息并提取声纹特征；

[0007] S3、对视频部分，利用目标检测和跟踪技术得到每个人在画面中的时间，并记录目标ID，对每一个目标获取目标人属性信息；

[0008] S4、从第一语音片段开始，融合发声者属性信息与目标人属性信息的匹配结果以及声纹特征比对和发声者属性信息比对的结果，确定相应语音片段的说话人，得到最终的视频说话人日志；

[0009] S41、对第一个语音片段，对该片段持续期间所有在视频中出现的目标，取目标人属性信息与该片段的发声者属性信息进行匹配，根据匹配结果确定该语音片段说话人，同时保存该语音片段的声纹特征、语音长度、发声者属性信息和目标ID；

[0010] S42、对于后续每个语音片段，选取之前出现过的说话人，进行声纹特征比对和发声者属性信息比对，根据比对结果判断是否为之前出现过的说话人，如果是之前出现过的说话人，则可确定为该语音片段的说话人；

[0011] S43、如果不是之前出现过的说话人，先根据目标ID去掉之前出现过的说话人，对剩下的目标，进行发声者属性信息和目标人属性信息匹配，根据匹配结果确定该片段的说话人，同时保存该片段的声纹特征、语音长度、发声者属性信息和目标ID。

[0012] 具体地，在步骤S2中，所述利用语音识别技术进行分割是先将语音信息转化为无标点的文本信息，然后恢复文本信息的标点符号，根据标点符号将音频部分分割为包含单个说话人的语音片段；所述发声者属性信息包括发声者的年龄、性别、情绪等。

[0013] 具体地，在步骤S3中，所述目标检测可以是人脸目标、人头目标、半身目标或全身目标，所述每个人在画面中的时间是指每个人在画面中的出现时间和持续时间，所述目标人属性信息包括目标人的年龄、性别、情绪等。

[0014] 同时，本发明还提供了一种使用上述方法的视频说话人日志系统，该系统包括录音录像设备、存储器、处理器、显示设备、扬声器和输入设备，所述录音录像设备分别与所述存储器、所述处理器、所述扬声器、所述显示设备连接，所述处理器还要与所述输入设备、所述存储器、所述扬声器和所述显示设备连接。

[0015] 具体地，所述录音录像设备用于录制对话中的音频和视频。

[0016] 具体地，所述存储器用于存储录制的音视频数据以及视频说话人日志可执行程序。

[0017] 具体地，所述处理器用于执行所述视频说话人日志可执行程序，并把程序执行结果返回给所述存储器或者所述显示设备。

[0018] 具体地，所述显示设备负责显示录制的视频和(或)程序执行结果。

[0019] 具体地，所述扬声器负责播放录制的音频。

[0020] 具体地，所述输入设备用于输入一些指令，以控制程序的运行、停止和显示结果。

[0021] 基于本发明的以上技术方案，本发明的有益效果如下：

[0022] 1.依赖于自动语音识别技术和标点符号恢复技术得到的语音片段，可以更好的挖掘上下文信息，准确度更高。

[0023] 2．发声者属性信息的获取，对语音片段的长度要求较低，例如：人耳可以通过听单字或者双字就能判断说话人的性别。

[0024] 3．根据发声者属性信息和目标人属性信息的匹配结果确定说话人，既可以获得说话人的语音信息，也可以获得说话人的图像信息，生成的日志信息更加丰富完整。

[0025] 4．融合声纹特征比对和发声者属性信息比对来判定说话人，避免声纹不可靠时的误匹配问题，同时使用比对来代替聚类，既解决了无法预知说话人数量导致的聚类不准确问题，也解决了不同说话人说话长度极度不均衡导致的聚类失败问题。

附图说明

[0026] 图1本发明一种视频说话人日志方法流程图；

[0027] 图2本发明一种视频说话人日志系统结构图。

具体实施方式

[0028] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本发明，并不用于限定本发明。此外，下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

实施例1

[0029] 如图1所示，本实施例公开了一种视频说话人日志方法，该方法包括以下步骤：

[0030] S1、将录音录像视频分离为音频部分和视频部分；

[0031] S2、对音频部分，利用语音识别技术进行分割，得到仅包含单个说话人的语音片段，对每个语音片段进行发生者属性检测获取发声者属性信息并提取声纹特征；

[0032] 通过自动语音识别技术，将语音信息转化为无标点的文本信息，将文本信息输入到利用词汇特征和韵律特征并结合监督学习技术和深度学习技术对大规模数据进行训练得到的模型中，为文本信息恢复其标点符号，根据标点符号对音频部分进行分割，得到包含单个说话人的语音片段；

[0033] 对每个语音片段进行发声者属性检测，根据语音信号可以预测说话人的年龄(老年、中年、小孩)及相应的置信度、性别及置信度、情绪(开心、难过、生气、恶心、害怕、惊讶)及置信度，获取发声者属性信息；同时提取每个语音片段的声纹特征，声纹特征可以是传统的i-vector，也可以是基于深度学习的d-vector，x-vector或者其他方法。

[0034] S3、对视频部分，利用目标检测和跟踪技术得到每个人在画面中的时间，并记录目标ID，对每一个目标获取目标人属性信息；

[0035] 利用目标(人脸目标、人头目标、半身或全身人)检测和跟踪技术，获取每个人在画面中的出现时间和持续时间，并标记每一个目标ID；对每一个目标，通过人脸图像、人头图像、半身图像或者全身图像预测，也可以是融合多种类型图像预测的获取目标人属性信息，获取的目标人属性信息包括年龄(老年、中年、小孩)及置信度、性别及置信度、情绪(开心，难过，生气，恶心，害怕，惊讶)及置信度。

[0036] S4、从第一语音片段开始，融合发声者属性信息与目标人属性信息的匹配结果以及声纹特征比对和发声者属性信息比对的结果，确定相应语音片段的说话人，得到最终的视频说话人日志；

[0037] S41、对第一个语音片段，对该片段持续期间所有在视频中出现的目标，取目标人属性信息与该片段的发声者属性信息进行匹配，根据匹配结果确定该语音片段说话人，同时保存该语音片段的声纹特征、语音长度、发声者属性信息和目标ID；属性匹配分数的计算方式为：

[0038] $$\mathrm{Attri\_score}(j)=\sum_{i\in\{所有使用的属性\}} w_i sim(A_i,V_{i,j})$$

[0039] 上式中，w_i为第i发声者属性信息和第i目标人属性信息匹配的权重，A_i表示当前语音片段的第i发声者属性信息，$V_{i,j}$表示ID为j的说话人的第i目标人属性信息，$sim(A_i,V_{i,j})$为第i发声者属性信息和第i目标人属性信息的匹配相似度，当第i发声者属性信息类别和第i目标人属性信息类别相同时(如性别属性结果都是男性)，匹配相似度为两者置信度之积；当第i发声者属性信息类别和第i目标人属性信息类别不相同时(如语音的性别属性为男性，而图像的性别属性为女性)，则匹配相似度为0。假设目标J为最高相似度的匹配，且匹配分数高于阈值T1，判定该语音片段的说话人为目标J，记录其ID为J，语音长度为L，声纹特征为f_J和发声者属性信息特征为$A_{i,J}$，$i\epsilon${所有属性}；若匹配分数均低于阈值T1，则丢弃该语音片段。该步骤根据发声者属性信息和目标人属性信息的匹配来指定语音片段的说话人，得到第一个说话人；

[0040] S42、对于后续语音片段k，先根据声纹特征比对和发声者属性信息比对的结果判断是否为之前出现过的说话人。假设之前出现过的说话人为N人，任取一个说话人，其ID、声纹特征和发声者属性特征分别为j、f_j和$A_{i,j}$，则该说话人与当前语音片段k的比对分数的计算方式为：

[0041] $$\text{score}=\alpha \sum_{i\in\{所有使用的属性\}} w_i^A sim(A_{i,k},A_{i,j}) + \beta sim(f_j,f_k)$$

[0042] 上式中，α是属性分数的权重，β是声纹相似度的权重，w_i^A表示第i发声者属性的权重，$sim(A_{i,k},A_{i,j})$表示语音片段k与说话人j在表示第i发声者属性上的相似度，其计算方式可类似于第i发声者属性信息和第i目标人属性信息的相似度计算方式，$sim(f_j,f_k)$表示根据声纹特征f_j和f_k的声纹相似度。β的取值可以是固定值，也可以是由语音片段k长度和语音片段j长度共同决定的值，当长度较长，说明声纹特征较可靠，那么β就越大，反之则越小。对N个目标，计算得到N个比对分数，取最高比对分数，其ID为M，若最高比对分数高于阈值T2，则认为该语音片段来自目标M，更新目标M的发声者属性值、声纹特征为当前该说话人最长语音片段的发声者属性值和声纹特征。若比对分数均低于阈值T2，则认为不是之前出现过的说话人；

[0043] S43、如果不是之前出现过的说话人，对于当前语音片段持续时间内出现在视频画面中的人，先根据目标ID去掉之前出现过的说话人，对剩下的目标，计算其目标人属性信息与该语音片段的发声者属性信息的匹配分数Attri_score，取匹配分数最高者，且其匹配分数高于阈值T1，确定为该语音片段的说话人，同样记录目标ID、语音长度、声纹特征和发声者属性信息；若匹配分数均低于阈值T1，，则丢弃该语音片段。

实施例2

[0044] 如图2所示，本实施例公开了一种视频说话人日志系统，该系统包括录音录像设备、存储器、处理器、扬声器、输入设备、显示设备，录音录像设备分别与存储器、处理器、扬声器、显示设备连接，另外处理器还要与输入设备、存储器、扬声器和显示设备连接。

[0045] 其中，录音录像设备负责录制对话中的音频和视频，录制结束后，录制的音频和视频数据存储于存储器，同时也可通过显示设备和扬声器播放；

[0046] 存储器除了存储录制的音频和视频数据,同时也存储了视频说话人日志可执行程序;

[0047] 处理器负责执行视频说话人日志可执行程序,并把程序执行结果返回给存储器,也可以返回给显示设备和扬声器;

[0048] 输入设备用于输入一些指令,以控制程序的运行、停止和显示结果;

[0049] 显示设备除了负责播放录制的视频,也可以播放处理器返回的程序执行结果,也可以同时播放相对应的录制视频和处理器返回的程序执行结果。

[0050] 本文中所描述的具体实施仅仅是对本发明精神作具体说明,本发明所属技术领域的技术人员可以对所描述的具体实施例进行各种微调修改或补充或采用类似的方法替代,均包含在本发明的保护范围之内。

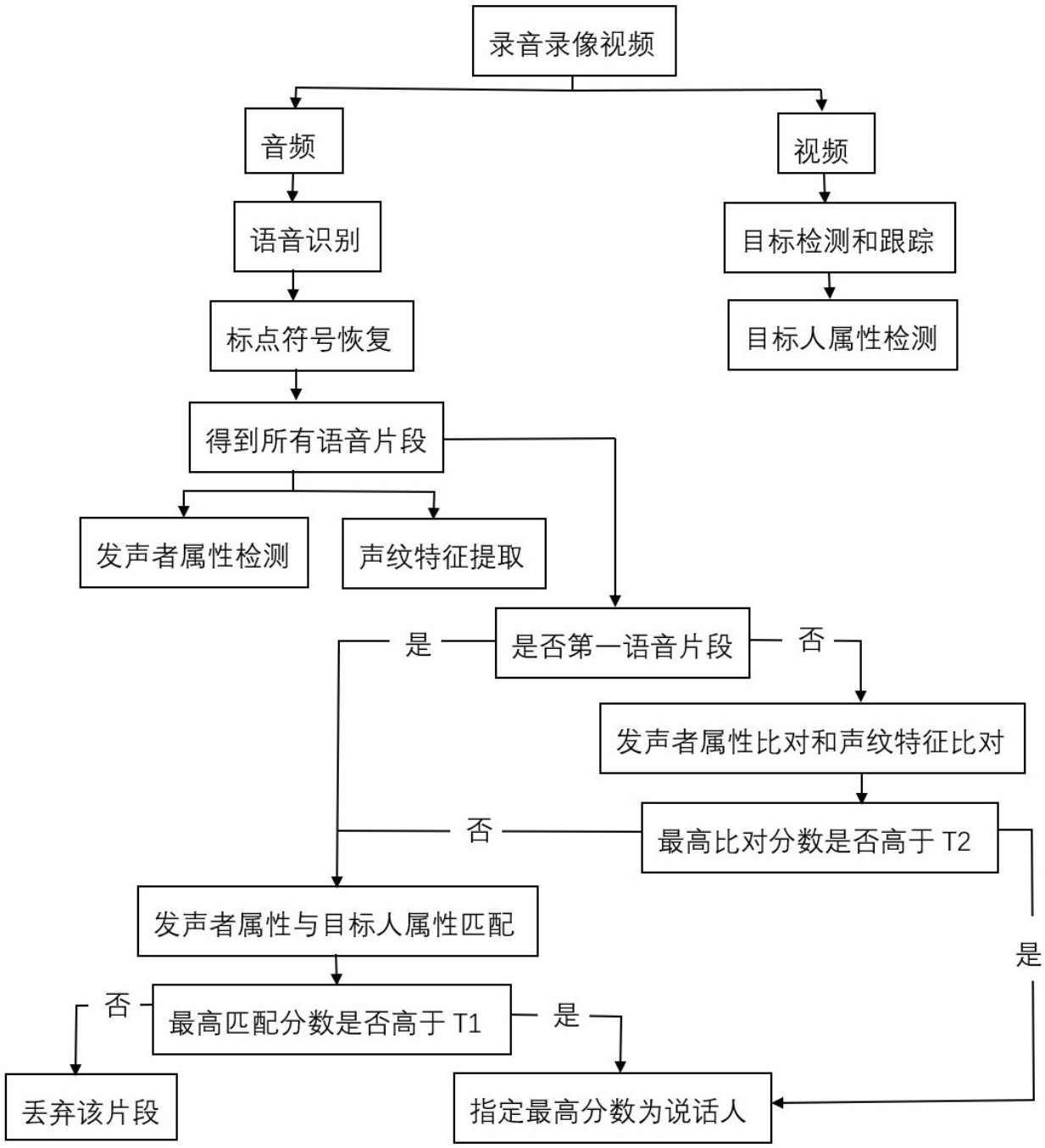

录音录像视频
音频
视频
语音识别
目标检测和跟踪
标点符号恢复
目标人属性检测
得到所有语音片段
发声者属性检测
声纹特征提取
是否第一语音片段
是
否
发声者属性比对和声纹特征比对
最高比对分数是否高于 T2
否
是
发声者属性与目标人属性匹配
否
最高匹配分数是否高于 T1
是
丢弃该片段
指定最高分数为说话人

图 1

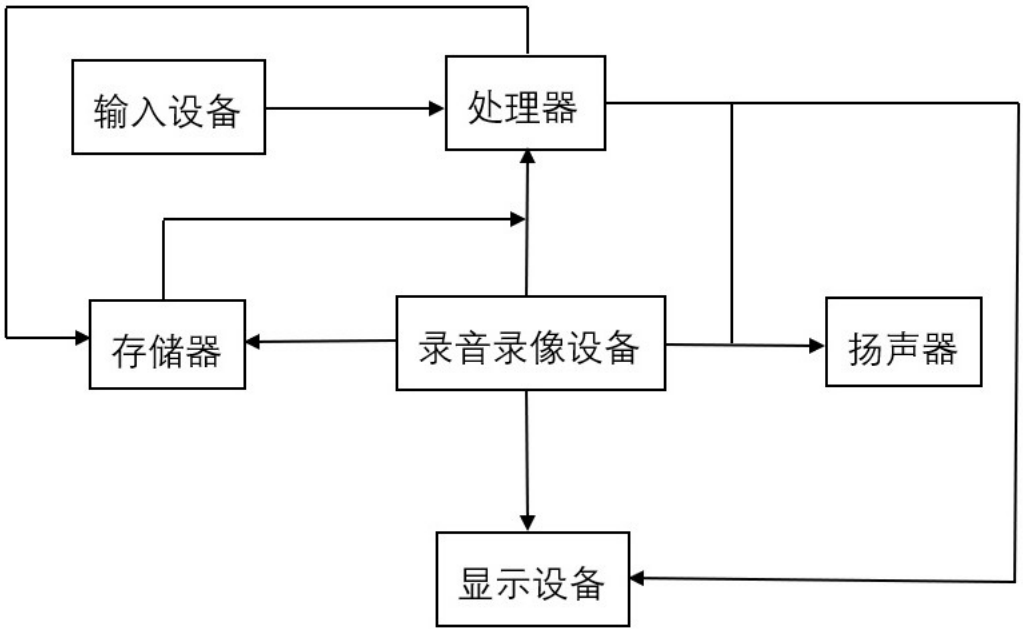


图 2