(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 109872714 A
(43)申请公布日 2019.06.11

(21)申请号 201910072525.9

(22)申请日 2019.01.25

(71)申请人 广州富港万嘉智能科技有限公司
地址 510000 广东省广州市黄埔区科学城南云五路11号光正科技产业园内501-1

(72)发明人 傅峰峰

(74)专利代理机构 广州市越秀区哲力专利商标事务所(普通合伙) 44288
代理人 罗晶 高淑怡

(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/183(2013.01)
G10L 15/25(2013.01)
G10L 15/26(2006.01)
G10L 21/0216(2013.01)

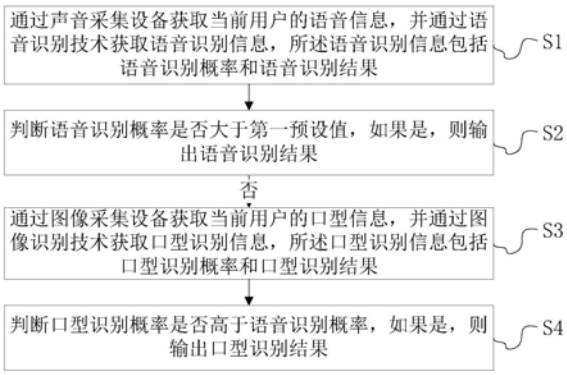
权利要求书1页 说明书6页 附图1页

(54)发明名称

一种提高语音识别准确性的方法、电子设备及存储介质

(57)摘要

本发明公开了一种提高语音识别准确性的方法，包括以下步骤：通过声音采集设备获取当前用户的语音信息，并通过语音识别技术获取语音识别信息，所述语音识别信息包括语音识别概率和语音识别结果；判断语音识别概率是否大于第一预设值，如果是，则输出语音识别结果；通过图像采集设备获取当前用户的口型信息，并通过图像识别技术获取口型识别信息；判断口型识别概率是否高于语音识别概率，如果是，则输出口型识别结果。本发明还提供了一种电子设备和计算机可读存储介质。本发明的提高语音识别准确性的方法通过综合比对语音识别结果和口型识别结果以得到准确率更高的识别结果，从而提高了识别的准确性。



CN 109872714 A

1.一种提高语音识别准确性的方法,其特征在于,包括以下步骤:

声音识别步骤:通过声音采集设备获取当前用户的语音信息,并通过语音识别技术获取语音识别信息,所述语音识别信息包括语音识别概率和语音识别结果;

第一判断步骤:判断语音识别概率是否大于第一预设值,如果是,则输出语音识别结果,如果否,则执行口型识别步骤;

口型识别步骤:通过图像采集设备获取当前用户的口型信息,并通过图像识别技术获取口型识别信息,所述口型识别信息包括口型识别概率和口型识别结果;

第二判断步骤:判断口型识别概率是否高于语音识别概率,如果是,则输出口型识别结果。

2.如权利要求1所述的一种提高语音识别准确性的方法,其特征在于,所述第二判断步骤包括以下子步骤:

计算步骤:计算得到口型识别概率与语音识别概率的差值;

结果步骤:判断该差值是否大于第二预设值,如果是,输出口型识别结果,且所述第二预设值为正值。

3.如权利要求2所述的一种提高语音识别准确性的方法,其特征在于,所述结果步骤中,判断该差值是否大于第二预设值,所述第二预设值为正值,如果是,则输出口型识别结果,如果否,则将语音识别结果和口型识别结果均输出并对其进行标记。

4.如权利要求2所述的一种提高语音识别准确性的方法,其特征在于,所述结果步骤中,判断该差值是否大于第二预设值,如果是,输出口型识别结果,如果否,通过自然语言处理技术分别计算语音识别结果和口型识别结果与上下文语句之间的语意相关性,采用语意相关性较高的作为预测结果进行输出。

5.如权利要求4所述的一种提高语音识别准确性的方法,其特征在于,所述语音识别结果、口型识别结果和预测结果为会议记录或者点菜指令。

6.如权利要求1-5中任意一项所述的一种提高语音识别准确性的方法,其特征在于,所述第一预设值为85%。

7.如权利要求1-5中任意一项所述的一种提高语音识别准确性的方法,其特征在于,所述第二预设值为5%。

8.如权利要求1-5中任意一项所述的一种提高语音识别准确性的方法,其特征在于,所述声音采集设备采用环形麦克风,所述图像采集设备采用环形摄像头阵列。

9.一种电子设备,包括存储器、处理器以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计算机程序时实现权利要求1-8中任意一项所述的一种提高语音识别准确性的方法。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于:所述计算机程序被处理器执行时实现如权利要求1-8中任意一项所述的一种提高语音识别准确性的方法。

一种提高语音识别准确性的方法、电子设备及存储介质

技术领域

[0001] 本发明涉及一种识别技术领域，尤其涉及一种提高语音识别准确性的方法、电子设备及存储介质。

背景技术

[0002] 目前，语音识别是一种将数字语音转换为计算机可以理解的文字的技术。最近几年，语音识别技术取得显著进展，语音识别技术逐渐走入人们的生活，给我们的生活、工作带来便利。目前语音识别技术已经在工业、家电、通信、汽车电子、医疗、家庭服务、消费电子产品等各个领域开始应用。本发明主要聚焦语音识别(即对录音文件的识别)，比如会议记录、电话客服语音的识别分析和餐厅点菜等。虽然语音识别技术已经得到了空前的发展，准确率已经处于相对较高的水平，但是还是无法做到完全准确；因此，进一步提高语音识别准确性成为本领域技术人员亟待解决的技术问题。

发明内容

[0003] 为了克服现有技术的不足，本发明的目的之一在于提供一种提高语音识别准确性的方法，其能进一步提高识别准确性。

[0004] 本发明的目的之二在于提供一种电子设备，其能进一步提高识别准确性。

[0005] 本发明的目的之三在于提供一种计算机可读存储介质，其能进一步提高识别准确性。

[0006] 本发明的目的之一采用如下技术方案实现：

[0007] 一种提高语音识别准确性的方法，包括以下步骤：

[0008] 声音识别步骤：通过声音采集设备获取当前用户的语音信息，并通过语音识别技术获取语音识别信息，所述语音识别信息包括语音识别概率和语音识别结果；

[0009] 第一判断步骤：判断语音识别概率是否大于第一预设值，如果是，则输出语音识别结果，如果否，则执行口型识别步骤；

[0010] 口型识别步骤：通过图像采集设备获取当前用户的口型信息，并通过图像识别技术获取口型识别信息，所述口型识别信息包括口型识别概率和口型识别结果；

[0011] 第二判断步骤：判断口型识别概率是否高于语音识别概率，如果是，则输出口型识别结果。

[0012] 进一步地，所述第二判断步骤包括以下子步骤：

[0013] 计算步骤：计算得到口型识别概率与语音识别概率的差值；

[0014] 结果步骤：判断该差值是否大于第二预设值，如果是，输出口型识别结果，且所述第二预设值为正值。

[0015] 进一步地，所述结果步骤中，判断该差值是否大于第二预设值，所述第二预设值为正值，如果是，则输出口型识别结果，如果否，则将语音识别结果和口型识别结果均输出并对其进行标记。

[0016] 进一步地，所述结果步骤中，判断该差值是否大于第二预设值，如果是，输出口型识别结果，如果否，通过自然语言处理技术分别计算语音识别结果和口型识别结果与上下文语句之间的语意相关性，采用语意相关性较高的作为预测结果进行输出。

[0017] 进一步地，所述语音识别结果、口型识别结果和预测结果为会议记录或者点菜指令。

[0018] 进一步地，所述第一预设值为85％。

[0019] 进一步地，所述第二预设值为5％。

[0020] 进一步地，所述声音采集设备采用环形麦克风，所述图像采集设备采用环形摄像头阵列。

[0021] 本发明的目的之二采用如下技术方案实现：

[0022] 一种电子设备，包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现本发明目的之一中任意一项所述的一种提高语音识别准确性的方法。

[0023] 本发明的目的之三采用如下技术方案实现：

[0024] 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如本发明目的之一中任意一项所述的一种提高语音识别准确性的方法。

[0025] 相比现有技术，本发明的有益效果在于：

[0026] 本发明的提高语音识别准确性的方法通过综合比对语音识别结果和口型识别结果以得到准确率更高的识别结果，从而提高了识别的准确性。

附图说明

[0027] 图1为实施例一的提高语音识别准确性的方法的流程图。

具体实施方式

[0028] 下面，结合附图以及具体实施方式，对本发明做进一步描述，需要说明的是，在不相冲突的前提下，以下描述的各实施例之间或各技术特征之间可以任意组合形成新的实施例。

[0029] 实施例一

[0030] 在本实施例的描述中，主要是针对于会议记录和用户点餐两个场景来进行描述的，但是在进行具体实施的时候其不仅仅可以应用于这两种场景中，还可以根据实际的需求应用于其他的场景中。

[0031] 如图1所示，本实施例提供了一种提高语音识别准确性的方法，包括以下步骤：

[0032] S1：通过声音采集设备获取当前用户的语音信息，并通过语音识别技术获取语音识别信息，所述语音识别信息包括语音识别概率和语音识别结果；所述声音采集设备采用环形麦克风；通过环形麦克风可以更为高效准确的获取圆桌四周的声音信息，获取到的声音源信息越清晰，那么后期进行语音翻译也就会使得其越准确。

[0033] 语音识别技术主要包括特征提取技术、模式匹配准则及模型训练技术三个方面。语音识别系统主要组成包括语音信号采样模块、语音信号前期处理模块、语音信号特征参数提取模块、语音信号识别核心模块、语音信号识别后期处理模块。模式识别匹配是语音识

别的主要过程。首先对人的语音进行分析，提取特点建立针对性的语音模型，通过语音模型建立语音识别所需的模式。利用语音识别的整体模型，在语音识别过程中将得到的语音信号的特征与前期建立的语音模式进行匹配比较，通过预设的搜索策略和匹配策略，可以得出最好的且与输入的语音信号相匹配的模式。最后，就可以计算机输出的识别结果。

[0034] 一般来说，语音识别的方法有三种：基于声道模型和语音知识的方法、模板匹配的方法以及利用人工神经网络的方法。模板匹配的方法发展比较成熟，目前已达到了实用阶段。在模板匹配方法中，要经过四个步骤：特征提取、模板训练、模板分类、判决。常用的技术有三种：动态时间规整(DTW)、隐马尔可夫(HMM)理论、矢量量化(VQ)技术。

[0035] 上述仅仅是描述了在语音识别领域中我们可能会采用到的技术，接下来具体针对与一段语音详细描述一下识别过程：当我们要对一段语音进行识别时，首先需要进行的是对语音特征的提取。这一步所做的工作其实就是从输入的语音信号(时域信号)中提取出可以进行建模的声学观测特征向量序列O。通俗地解释就是把需要识别的一段语音进行特征提取，之后得到了一组可以表征这一段语音的向量，后续对语音进行的一系列操作都是基于这组向量的。

[0036] 在观测特征向量O的条件下，找到一组词向量W使得P(W|O)的概率最大。这个也正是人听到一段语音的时候做的事情——找所有已知文字中和这段语音最匹配的。但是，仅仅依靠这个公式，我们是无法解决语音识别问题的。还需要利用贝叶斯定理对其进行转换，将其转换成我们能够分别进行建模求解的形式。转换如下：W＝argmaxP(W|O)＝argmax P(O|W)P(W)/P(O)；

[0037] 其中，P(O)是声学观测的先验概率，在自动语音识别过程中，由于输入的声学观测特征序列是固定的，可以认为上述公式中的P(O)是常量，因此P(O)在上述公式的最大化的过程中不起作用，可以忽略。那么我们现在只剩下P(O|W)和P(W)需要考虑。而在声学模型和语言模型分别提供了对P(O|W)和P(W)进行计算的方法，通过声学模型和语言模型计算得到P(O|W)和P(W)。

[0038] 我们利用上述的声学模型、语言模型和发音词典就可以构建起一个解码空间，之后利用解码器，结合每一组输入的语音特征向量在空间中进行搜索，找到一条最优的词序列，就是找到一条路径使得P(O|W)P(W)概率最大。那么，最终得到的这个词序列就是我们想要的识别结果。也即是在我们获取对应的识别结果的时候，也可以获取到对应的识别概率，从一定程度上来说，这个概率显示的是该识别结果的准确率。

[0039] S2：判断语音识别概率是否大于第一预设值，如果是，则输出语音识别结果，如果否，则执行口型识别步骤；所述第一预设值为85％；在常规的会议记录中，不管这个语音识别概率的高或者低，其都会对其进行记录，这样就会产生很大的正确性的隐患；为了避免这样的情况出现，可以有以下的方式去进行实施，当会议记录中，当语音识别概率低于第一预设值的时候，可以将其标记出来，留给后续整理会议记录者去进行复核，标记方式可以是加粗或者斜体或者变换颜色等等。但是如果是在点餐过程中，如果语音识别得到的准确率较低的话，则不会下单，会发出语音提示来进一步与用户进行沟通，以判断是否继续下单。

[0040] 但是在本申请中并非采用这样的方式，在本实施例中，为了实现该语音识别系统更高的自动化，设置了另外一种方式对语音信息进行复核；也即是通过图像识别口型的方式来进行，由于两者采用的是不同的模型与识别逻辑，所以避免了在一定程度上的错误的

交叠，使得该语句信息是经过两次的计算校验，能够从一定程度上避免出错，使得其语音识别的准确率进一步的提高。在进行设置的时候，可以设置两者同时进行识别然后比对，也可以设置先进行口型识别然后再进行语音识别这样的方式。

[0041] 最为优选地实施方式如下：就是先对语音信息进行检测，然后再通过口型识别去进行语句检测。由于相对于图像来说，处理语音的数据量还是相对较小，所以在这里设置一个判断的步骤，只有当语音识别的得到的概率较低的时候，才使用口型识别，这样不仅能够保证一定的识别准确性，还可以使得整体的处理速度较快。使得其可以消耗的计算资源相对较少，提升了整体自动化识别的效率。

[0042] S3：通过图像采集设备获取当前用户的口型信息，并通过图像识别技术获取口型识别信息，所述口型识别信息包括口型识别概率和口型识别结果；所述图像采集设备采用环形摄像头阵列；最为优选地方式是声音采集设备中麦克风的数量与环形摄像头中的摄像头的数量相同，并且使得两者有一个一一对应的关系，这样获取到的信息就可以直接进行对应，而不必再去寻找各自的对应关系。

[0043] 为了实现上述两者的结合，可以在麦克风上增设摄像头，这样使得麦克风不仅可以采集用户在执行说话过程中生成的语音信号，还可以采集生成的口型图像信号，该图像信号中至少包含人脸的唇部部位的图像，当然为了更好的识别出口型的变化，图像信号中也可以包括人脸的其它部位的图像，这是由于有时口型变化和人脸表情变化相关。

[0044] 在进行口型识别的过程中，也是依靠模板匹配来实现的，需要先构建口型识别库，这个口型识别库通过获取大量的图形或者视频信息来进行训练，使得该模型比较健壮。

[0045] 具体的实现过程中如下：图像采集设备通过摄像头获取只包含用户口型变化的视频序列并输入视频解码单元；视频解码单元将输入的唇动视频利用关键帧采集技术获取视频流中具有代表性的关键帧，并将提取的关键帧序列(归一化的唇部色彩静态图片)送入图像预处理单元；图像预处理单元对上一单元获得的关键帧图像，利用OpenCV库函数进行灰度化和中值滤波处理，而后对图片进行二值化处理，最后对图片进行扫描去噪获得规格化的口型二值化图片。

[0046] 特征提取单元针对经过图像处理后的规格化二值化图片，利用模板法进行口型特征提取，获得表示口型特征的特征向量；口型模板库是预先建立的用于存储标准口型特征向量的模块，储存了先期试验中采集的标准口型模板，包括所有汉语拼音字母发音时的唇动图像(单张或多张)样本及针对口型图像利用模板法提取的特征向量；口型识别单元对处理后的规格化二值化图像进行识别，从特征提取单元中获得序列中每张图片的特征向量，最后得到口型识别信息，同样的，口型识别结果也有对应的口型识别概率；这个概率用于后续进一步判断是否进行输出操作。

[0047] 对于会议记录来说，肯定是需要获取所有的口型特征，然后存储有足够大量的信息才能够对会议过程中出现的所有的情况进行比较完整的预测和记录。同样的在点餐过程中，也可以通过这样的方式去进行实现，但是除此之外，在点餐过程中还有另外一种操作方式去进行口型识别，由于每个商家的菜的样式是固定的，然后对于用户通常点菜的时候所采用的语句进行收集，得到所有在点餐过程中出现下单这样的祈使句与疑问句的差别，然后对收集到的所有的字音与口型进行匹配，然后去提取训练这样对应字音的口型特征；大大降低了数据库构建的繁琐，由于数据库量的减小，也可以使得在匹配的过程中速度大大

的提高;使得其达到更好的效果。这一步主要是通过图像识别口型信息从而得到对应的识别结果。

[0048] S4:判断口型识别概率是否高于语音识别概率,如果是,则输出口型识别结果。在具体实施的过程中有如下方式,一种是当得到的口型识别概率高于语音识别概率的时候,则可以直接将识别概率较高的那个作为结果来进行输出,这个时候,操作相对简单,但是会存在这样一个问题,就是虽然口型识别概率相对较高,但是两者结果对于正常语音识别来说,这样较低的准确率都是不能够接受的,比如两者一个80%,另外一个是81%;虽然说,口型识别概率比语音识别概率高1%,但是在会议记录过程中,这样低的识别率也是不能够接受的;并且由于语音识别结果和口型识别结果都是通过概率得到的,所以其在一定范围内存在有误差,而这种1%的差距有时候也是可以忽略的。

[0049] 故而更为优选的实施方式如下:

[0050] 所述步骤S4具体包括以下子步骤:

[0051] 计算步骤:计算得到口型识别概率与语音识别概率的差值;这个差值是有正负的,而不是绝对值。

[0052] 结果步骤:判断该差值是否大于第二预设值,如果是,输出口型识别结果,且所述第二预设值为正值,如果否,则将语音识别结果和口型识别结果均输出并对其进行标记。所述第二预设值为5%。在这里也就是两者识别得到的准确率都处于较低的水平,比如两者的识别率都是80%,如果是在会议记录这类型的识别过程中,则对这两者分别进行标记,然后并注明各自的对应关系,以便于后期复查者能够更明确的定位并进行复查。而如果是在点餐这样类型的识别过程中,则直接挑选其中一条结果发送至音频输出端进行音频阅读,使得点餐人员进行进一步确认。

[0053] 虽然上述方式可以从一定程度上解决提升精确度的问题,但是为了更好的能够进行自动化,使得会议记录复查者减少工作量,使得点餐人员少一点确认即可点到正确的菜品,在本实施例中还采用了另一种方式去进行识别。所述结果步骤中,判断该差值是否大于第二预设值,如果是,输出口型识别结果,如果否,通过自然语言处理技术分别计算语音识别结果和口型识别结果与上下文语句之间的语意相关性,采用语意相关性较高的作为预测结果进行输出。

[0054] 通过自然语言处理技术对其进行分析的时候,首选需要做的是对两种识别结果进行分词,只有进行完分词之后才有可能对两者进行进一步的比对分析。在具体比对分析中有如下方式去进行;可以先通过对句法来识别两种识别结果中不符合句话规范的内容,然后对分词结果中,各个词语之间的相关性进行语义分析。

[0055] 对于不同的语言单位,语义分析的任务各不相同。在词的层次上,语义分析的基本任务是进行词义消歧(WSD),在句子层面上是语义角色标注(SRL);有监督词义消歧根据上下文和标注结果完成分类任务。而无监督词义消歧通常被称为聚类任务,使用聚类算法对同一个多义词的所有上下文进行等价类划分,在词义识别的时候,将该词的上下文与各个词义对应上下文的等价类进行比较,通过上下文对应的等价类来确定词的词义。此外,除了有监督和无监督的词义消歧,还有一种基于词典的消歧方法。通过自然语言识别技术来计算得到语意相关性更高的作为结果进行输出。所述语音识别结果、口型识别结果和预测结果为会议记录或者点菜指令。

[0056] 本实施例的方案通过综合语音识别和口型识别进一步提高了语音识别的准确率，并且在两者都识别得到的结果准确性都较低的时候，进一步通过自然语言处理技术来进一步进行修饰，选取语意相关性较高的结果，通过多方面综合性进行验证和判断，使得语音识别的准确性大大提高；也就使得该系统的实用性更好。

[0057] 实施例二

[0058] 实施例二公开了一种电子设备，该电子设备包括处理器、存储器以及程序，其中处理器和存储器均可采用一个或多个，程序被存储在存储器中，并且被配置成由处理器执行，处理器执行该程序时，实现实施例一的一种提高语音识别准确性的方法。该电子设备可以是手机、电脑、平板电脑等等一系列的电子设备。

[0059] 实施例三

[0060] 实施例三公开了一种计算机可读存储介质，该存储介质用于存储程序，并且该程序被处理器执行时，实现实施例一的一种提高语音识别准确性的方法。

[0061] 当然，本发明实施例所提供的一种包含计算机可执行指令的存储介质，其计算机可执行指令不限于如上所述的方法操作，还可以执行本发明任意实施例所提供的方法中的相关操作。

[0062] 通过以上关于实施方式的描述，所属领域的技术人员可以清楚地了解到，本发明可借助软件及必需的通用硬件来实现，当然也可以通过硬件实现，但很多情况下前者是更佳的实施方式。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，如计算机的软盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、闪存(FLASH)、硬盘或光盘等，包括若干指令用以使得一台电子设备(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的方法。

[0063] 值得注意的是，上述基于内容更新通知装置的实施例中，所包括的各个单元和模块只是按照功能逻辑进行划分的，但并不局限于上述的划分，只要能够实现相应的功能即可；另外，各功能单元的具体名称也只是为了便于相互区分，并不用于限制本发明的保护范围。

[0064] 上述实施方式仅为本发明的优选实施方式，不能以此来限定本发明保护的范围，本领域的技术人员在本发明的基础上所做的任何非实质性的变化及替换均属于本发明所要求保护的范围。

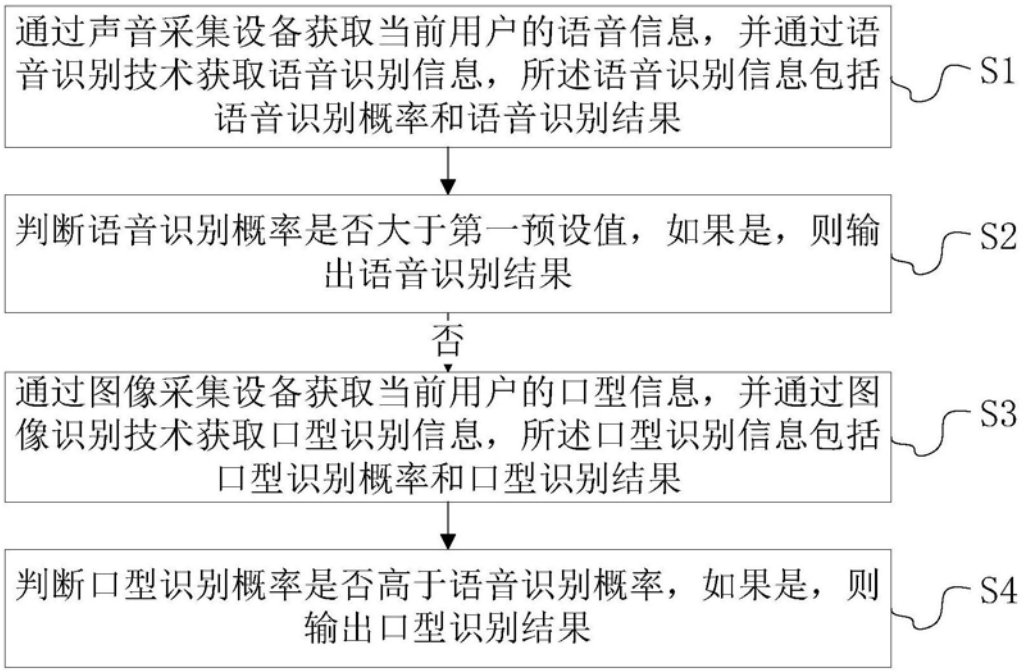

通过声音采集设备获取当前用户的语音信息，并通过语音识别技术获取语音识别信息，所述语音识别信息包括语音识别概率和语音识别结果
S1
判断语音识别概率是否大于第一预设值，如果是，则输出语音识别结果
S2
否
通过图像采集设备获取当前用户的口型信息，并通过图像识别技术获取口型识别信息，所述口型识别信息包括口型识别概率和口型识别结果
S3
判断口型识别概率是否高于语音识别概率，如果是，则输出口型识别结果
S4

图1