



(12) 发明专利申请

(10) 申请公布号 CN 103971675 A

(43) 申请公布日 2014. 08. 06

(21) 申请号 201310033201. 7

(22) 申请日 2013. 01. 29

(71) 申请人 腾讯科技(深圳)有限公司

地址 518044 广东省深圳市福田区振兴路赛
格科技园 2 栋东 403 室

(72) 发明人 饶丰 卢鲤 陈波 岳帅 张翔
王尔玉 谢达东 李露 陆读羚

(74) 专利代理机构 北京德琦知识产权代理有限
公司 11018

代理人 张晓峰 宋志强

(51) Int. Cl.

G10L 15/02(2006. 01)

G10L 15/06(2013. 01)

G10L 21/06(2013. 01)

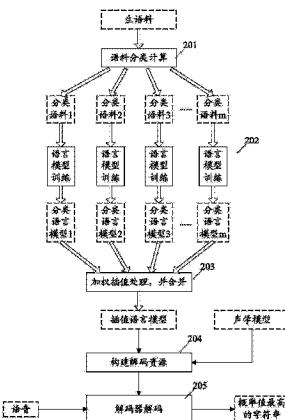
权利要求书3页 说明书7页 附图7页

(54) 发明名称

自动语音识别方法和系统

(57) 摘要

本申请公开了一种自动语音识别方法和系统,包括:对生语料进行语料分类计算,得到一个以上不同类别的分类语料;针对所述每个分类语料进行训练得到一个以上对应的分类语言模型;依据分类的生僻程度为所述各个分类语言模型进行加权插值处理,其中各分类的生僻程度与该分类对应的加权值成正相关关系,将加权插值处理后的分类语言模型合并,得到插值语言模型;依据声学模型和所述插值语言模型构建解码资源;依据所述解码资源,对输入的语音进行解码,输出概率值最高的字符串作为所述输入语音的识别结果。利用本发明,可以提高对生僻词语的语音的识别准确率。



1. 一种自动语音识别方法,其特征在于,包括:

对生语料进行语料分类计算,得到一个以上不同类别的分类语料;

针对所述每个分类语料进行语言模型训练计算,得到一个以上对应的分类语言模型;

依据分类的生僻程度为所述各个分类语言模型进行加权插值处理,其中各分类的生僻程度与该分类对应的加权值成正相关关系,将加权插值处理后的分类语言模型合并,得到插值语言模型;

依据声学模型和所述插值语言模型构建解码资源;

依据所述解码资源,对输入的语音进行解码,输出概率值最高的字符串作为所述输入语音的识别结果。

2. 根据权利要求 1 所述的方法,其特征在于,所述对生语料进行语料分类计算,得到一个以上不同类别的分类语料,具体包括:

根据生语料,计算词与词之间的亲和度矩阵;

利用词频 - 逆向文件频率 TF-IDF 方法从生语料中提取词特征;

根据所述亲和度矩阵,利用降维方法对所提取出的词特征进行降维处理;

将降维处理后的词特征输入分类器进行训练,输出一个以上不同类别的分类预料。

3. 根据权利要求 2 所述的方法,其特征在于,所述根据生语料,计算词与词之间的亲和度矩阵,具体包括:

对生语料进行分析,根据公式 $CO_{ij} = \frac{f_{ij} * f_{ji}}{d_{ij} * d_{ji} * f_i * f_j}$ 计算每个词与另一个词的词共现度,

并据此构建词与词的词共现矩阵;其中,所述 f_{ij} 为词 i 在词 j 前出现的次数, d_{ij} 为词 i 和词 j 的平均距离, f_i 为词 i 的词频, f_j 为词 j 的词频;

根据所述词共现矩阵,以及公式 $A_{ij} = \text{sqrt}(\sum OR(waf_{ik}, waf_{jk}) \sum OR(waf_{ki}, waf_{kj}))$ 计算词与词之间的亲和度,并据此构建词与词之间的亲和度矩阵。

4. 根据权利要求 2 所述的方法,其特征在于,所述降维方法为主成分分析 PCA 降维方法。

5. 根据权利要求 2 所述的方法,其特征在于,所述分类器为支持向量机 SVM 分类器。

6. 一种自动语音识别方法,其特征在于,包括:

根据生语料进行语言模型训练计算,得到主语言模型;

对生语料进行语料分类计算,得到一个以上不同类别的分类语料;

针对所述每个分类语料进行语言模型训练计算,得到一个以上对应的分类语言模型;

依据声学模型和所述主语言模型构建主解码资源,依据所述各分类语言模型构建对应的分类解码资源;

依据所述主解码资源对输入的语音进行解码,输出概率值 1 (w) 排在前 n 名的 n 个字符串;

依次根据所述各个分类语言模型对应的各分类解码资源,分别对所述 n 个字符串进行解码,得到每个字符串在每个分类语言模型中的概率值 n (w);将每个字符串在每个分类语言模型中的概率值 n (w) 乘以该字符串在主语言模型中的概率值 1 (w) 得到复合概率 p (w),输出复合概率 p (w) 最高的字符串作为所述输入语音的识别结果。

7. 根据权利要求 6 所述的方法,其特征在于,所述对生语料进行语料分类计算,得到一

个以上不同类别的分类语料,具体包括 :

根据生语料,计算词与词之间的亲和度矩阵;

利用 TF-IDF 方法从生语料中提取词特征;

根据所述亲和度矩阵,利用降维方法对所提取出的词特征进行降维处理;

将降维处理后的词特征输入分类器进行训练,输出一个以上不同类别的分类预料。

8. 根据权利要求 7 所述的方法,其特征在于,所述根据生语料,计算词与词之间的亲和度矩阵,具体包括 :

对生语料进行分析,根据公式 $CO_{ij} = \frac{f_{ij} * f_{ji}}{d_{ij} * d_{ji} * f_i * f_j}$ 计算每个词与另一个词的词共现度,

并据此构建词与词的词共现矩阵;其中,所述 f_{ij} 为词 i 在词 j 前出现的次数, d_{ij} 为词 i 和词 j 的平均距离, f_i 为词 i 的词频, f_j 为词 j 的词频;

根据所述词共现矩阵,以及公式 $A_{ij} = \text{sqrt}(\sum OR(waf_{ik} waf_{jk}) \sum OR(waf_{ki} waf_{kj}))$,计算词与词之间的亲和度,并据此计算词与词之间的亲和度矩阵。

9. 根据权利要求 7 所述的方法,其特征在于,所述降维方法为 PCA 降维方法。

10. 根据权利要求 7 所述的方法,其特征在于,所述分类器为 SVM 分类器。

11. 一种自动语音识别系统,其特征在于,包括 :

分类处理模块,用于对生语料进行语料分类计算,得到一个以上不同类别的分类语料;

分类语言模型训练模块,用于针对所述每个分类语料进行语言模型训练计算,得到一个以上对应的分类语言模型;

加权合并模块,用于依据分类的生僻程度为所述各个分类语言模型进行加权插值处理,其中各分类的生僻程度与该分类对应的加权值成正相关关系,将加权插值处理后的分类语言模型合并,得到插值语言模型;

资源构建模块,用于依据声学模型和所述插值语言模型构建解码资源;

解码器,用于依据所述解码资源,对输入的语音进行解码,输出概率值最高的字符串作为所述输入语音的识别结果。

12. 根据权利要求 11 所述的系统,其特征在于,所述分类处理模块具体包括 :

亲和度矩阵模块,用于根据生语料,计算词与词之间的亲和度矩阵;

特征提取模块,用于利用 TF-IDF 方法从生语料中提取词特征;

降维模块,用于根据所述亲和度矩阵,利用降维方法对所提取出的词特征进行降维处理;

分类器,用于对降维处理后的词特征进行训练,输出一个以上不同类别的分类预料。

13. 根据权利要求 12 所述的系统,其特征在于,所述降维模块为 PCA 降维模块。

14. 根据权利要求 12 所述的系统,其特征在于,所述分类器为 SVM 分类器。

15. 一种自动语音识别系统,其特征在于,包括 :

主语言模型训练模块,用于根据生语料进行语言模型训练计算,得到主语言模型;

分类处理模块,用于对生语料进行语料分类计算,得到一个以上不同类别的分类语料;

分类语言模型训练模块,用于针对所述每个分类语料进行语言模型训练计算,得到一

个以上对应的分类语言模型；

主资源构建模块，用于依据声学模型和所述主语言模型构建主解码资源；

分类资源构建模块，用于依据所述各分类语言模型构建对应的分类解码资源；

第一解码器，用于依据所述主解码资源对输入的语音进行解码，输出概率值 $l(w)$ 排在前 n 名的 n 个字符串；

第二解码器，用于依次根据所述各个分类语言模型对应的各分类解码资源，分别对所述 n 个字符串进行解码，得到每个字符串在每个分类语言模型中的概率值 $n(w)$ ；将每个字符串在每个分类语言模型中的概率值 $n(w)$ 乘以该字符串在主语言模型中的概率值 $l(w)$ 得到复合概率 $p(w)$ ，输出复合概率 $p(w)$ 最高的字符串作为所述输入语音的识别结果。

16. 根据权利要求 15 所述的系统，其特征在于，所述分类处理模块具体包括：

亲和度矩阵模块，用于根据生语料，计算词与词之间的亲和度矩阵；

特征提取模块，用于利用 TF-IDF 方法从生语料中提取词特征；

降维模块，用于根据所述亲和度矩阵，利用降维方法对所提取出的词特征进行降维处理；

分类器，用于对降维处理后的词特征进行训练，输出一个以上不同类别的分类预料。

17. 根据权利要求 16 所述的系统，其特征在于，所述降维模块为 PCA 降维模块。

18. 根据权利要求 16 所述的系统，其特征在于，所述分类器为 SVM 分类器。

自动语音识别方法和系统

技术领域

[0001] 本申请涉及自动语音识别(ASR, Automatic Speech Recognition)技术领域,尤其涉及一种自动语音识别方法和系统。

背景技术

[0002] 自动语音识别技术是将人类的语音中的词汇内容转换为计算机可读的输入字符的一项技术。语音识别具有复杂的处理流程,主要包括声学模型训练、语言模型训练、解码资源构建、以及解码四个过程。图 1 为现有自动语音识别系统的一种主要处理流程的示意图。参见图 1,主要处理过程包括:

[0003] 步骤 101 和 102,需要根据声学原料进行声学模型训练得到声学模型,以及根据生语料进行语言模型训练得到语言模型。

[0004] 所述声学模型是语音识别系统中最为重要的部分之一,目前的主流语音识别系统多采用隐马尔科夫模型(HMM, Hidden Markov Model)进行建模,隐马尔可夫模型是统计模型,它用来描述一个含有隐含未知参数的马尔可夫过程。在隐马尔可夫模型中,状态并不是直接可见的,但受状态影响的某些变量则是可见的。在声学模型中描述了语音与音素的对应概率。所述音素是根据语音的自然属性划分出来的最小语音单位。从声学性质来看,音素是从音质角度划分出来的最小语音单位;从生理性质来看,一个发音动作形成一个音素。

[0005] 所述语言模型主要构建为字符串 s 的概率分布 $p(s)$,反映了字符串 s 作为一个句子出现的概率。假设 w 为字符串 s 中的每个词,则:

$$p(s) = p(w_1 w_2 w_3 \dots w_n) = p(w_1) p(w_2 | w_1) p(w_3 | w_1 w_2) \dots p(w_k | w_1 w_2 \dots w_{k-1})$$

[0007] 步骤 103,依据所述声学模型和语言模型,以及预设的词典,构建相应的解码资源。所述解码资源为加权优先转换机(WFST, weighted finite state transducer) 网络。

[0008] 步骤 104、将语音输入到解码器,解码器依据所构建的解码资源对所述语音进行解码,输出概率值最高的字符串作为所述输入语音的识别结果。

[0009] 但是,现有的语音识别技术多基于普适性的语音识别应用,即针对常用说话识别来进行模型搭建,这种情况下,语言模型的训练语料主要根据数据采集以及实际用户的输入,虽然从某种程度上较好地反映了用户的说话习惯,针对日常用语往往有较好的识别效果;但是,由于语言模型的训练语料中关于生僻词语较少出现,例如医药名和地名等,不能形成有效的概率统计模型,语言模型中生僻词语对应字符串的概率值非常低,因此当需要识别用户说出的较为生僻的词语的时候,往往会发生数据偏移问题,即识别出的字符串不是用户说出的词语,也就是说对于生僻词语的语音的识别准确率较低,难以取得较好的识别结果。

发明内容

[0010] 有鉴于此,本发明的主要目的在于提供一种自动语音识别方法和系统,以提高对生僻词语的语音的识别准确率。

- [0011] 本发明的一种技术方案是这样实现的：
- [0012] 一种自动语音识别方法，包括：
- [0013] 对生语料进行语料分类计算，得到一个以上不同类别的分类语料；
- [0014] 针对所述每个分类语料进行语言模型训练计算，得到一个以上对应的分类语言模型；
- [0015] 依据分类的生僻程度为所述各个分类语言模型进行加权插值处理，其中各分类的生僻程度与该分类对应的加权值成正相关关系，将加权插值处理后的分类语言模型合并，得到插值语言模型；
- [0016] 依据声学模型和所述插值语言模型构建解码资源；
- [0017] 依据所述解码资源，对输入的语音进行解码，输出概率值最高的字符串作为所述输入语音的识别结果。
- [0018] 一种自动语音识别系统，包括：
- [0019] 分类处理模块，用于对生语料进行语料分类计算，得到一个以上不同类别的分类语料；
- [0020] 分类语言模型训练模块，用于针对所述每个分类语料进行语言模型训练计算，得到一个以上对应的分类语言模型；
- [0021] 加权合并模块，用于依据分类的生僻程度为所述各个分类语言模型进行加权插值处理，其中各分类的生僻程度与该分类对应的加权值成正相关关系，将加权插值处理后的分类语言模型合并，得到插值语言模型；
- [0022] 资源构建模块，用于依据声学模型和所述插值语言模型构建解码资源；
- [0023] 解码器，用于依据所述解码资源，对输入的语音进行解码，输出概率值最高的字符串作为所述输入语音的识别结果。
- [0024] 与现有技术相比，本发明的上述技术方案对生语料进行语料分类计算和训练，得到一个以上对应的分类语言模型，从而使得生僻词语可以被分类到某一个或某几个分类语言模板中，然后依据分类的生僻程度为所述各个分类语言模型进行加权插值处理，其中各分类的生僻程度与该分类对应的加权值成正相关关系，即生僻程度越高，则对应的加权值越高，将加权插值处理后的分类语言模型合并，得到插值语言模型。这样在插值语言模板中，生僻词语所对应的字符串的概率值就会相应提高，从而减少与常用词语对应字符串的概率值的差距，后续解码过程中，当需要识别用户说出的较为生僻的词语的时候，由于生僻词语对应的字符串的概率值显著提高，因此会降低发生数据偏移的几率，提高了对于生僻词语的语音的识别准确率。
- [0025] 本发明的再一种技术方案是这样实现的：
- [0026] 一种自动语音识别方法，包括：
- [0027] 根据生语料进行语言模型训练计算，得到主语言模型；
- [0028] 对生语料进行语料分类计算，得到一个以上不同类别的分类语料；
- [0029] 针对所述每个分类语料进行语言模型训练计算，得到一个以上对应的分类语言模型；
- [0030] 依据声学模型和所述主语言模型构建主解码资源，依据所述各分类语言模型构建对应的分类解码资源；

[0031] 依据所述主解码资源对输入的语音进行解码,输出概率值 $1(w)$ 排在前 n 名的 n 个字符串;

[0032] 依次根据所述各个分类语言模型对应的各分类解码资源,分别对所述 n 个字符串进行解码,得到每个字符串在每个分类语言模型中的概率值 $n(w)$;将每个字符串在每个分类语言模型中的概率值 $n(w)$ 乘以该字符串在主语言模型中的概率值 $1(w)$ 得到复合概率 $p(w)$,输出复合概率 $p(w)$ 最高的字符串作为所述输入语音的识别结果。

[0033] 一种自动语音识别系统,包括:

[0034] 主语言模型训练模块,用于根据生语料进行语言模型训练计算,得到主语言模型;

[0035] 分类处理模块,用于对生语料进行语料分类计算,得到一个以上不同类别的分类语料;

[0036] 分类语言模型训练模块,用于针对所述每个分类语料进行语言模型训练计算,得到一个以上对应的分类语言模型;

[0037] 主资源构建模块,用于依据声学模型和所述主语言模型构建主解码资源;

[0038] 分类资源构建模块,用于依据所述各分类语言模型构建对应的分类解码资源;

[0039] 第一解码器,用于依据所述主解码资源对输入的语音进行解码,输出概率值 $1(w)$ 排在前 n 名的 n 个字符串;

[0040] 第二解码器,用于依次根据所述各个分类语言模型对应的各分类解码资源,分别对所述 n 个字符串进行解码,得到每个字符串在每个分类语言模型中的概率值 $n(w)$;将每个字符串在每个分类语言模型中的概率值 $n(w)$ 乘以该字符串在主语言模型中的概率值 $1(w)$ 得到复合概率 $p(w)$,输出复合概率 $p(w)$ 最高的字符串作为所述输入语音的识别结果。

[0041] 与现有技术相比,本发明的上述方案对生语料进行语料分类计算和训练,得到一个以上对应的分类语言模型,从而使得生僻词语可以被分类到某一个或某几个分类语言模型中,而生僻词语在其所属的最相关的分类语言模型中的概率值 $n(w)$ 是较高的;在对输入语音进行解码时,先利用主语言模型所构建的主解码资源进行一次解码,输出的概率值 $1(w)$ 排在前 n 的 n 个字符串,生僻词语对应的字符串虽然在主语言模型中的概率值 $1(w)$ 往往不是最高的,但是通常能够排在前 n 名;接下来,再对该 n 个字符串分别根据每个分类语言模型对应的分类解码资源进行二次解码,得到每个字符串在每个分类语言模型中的概率值 $n(w)$;将每个字符串在每个分类语言模型中的概率值 $n(w)$ 乘以该字符串在主语言模型中的概率值 $1(w)$ 得到复合概率 $p(w)$,该复合概率 $p(w)$ 可以修正生僻词语的过低概率值 $1(w)$,因此按照该复合概率 $p(w)$ 的高低输出的字符串作为所述输入语音的识别结果,可以降低生僻词语的语音发生数据偏移的几率,提高了对于生僻词语的语音的识别准确率。

[0042] 由于本发明的技术方案没有对原始的生语料提出特殊要求,以生僻词出现频率较少的生语料为基础进行训练即可达到本发明的发明目的,因此能够在不影响普通用户日常使用的情况下,满足了某些特殊用户对生僻词语的语音识别需求。

附图说明

[0043] 图 1 为现有自动语音识别系统的一种主要处理流程的示意图;

[0044] 图 2 为本发明所述自动语音识别方法的一种处理流程图;

- [0045] 图 3 为本发明所述自动语音识别方法的又一种处理流程图；
- [0046] 图 4 为本发明所述对生语料进行语料分类计算，得到一个以上不同类别的分类语料的具体处理流程图；
- [0047] 图 5 为本发明所述一种语音识别系统的一种组成示意图；
- [0048] 图 6 为本发明所述又一种语音识别系统的一种组成示意图；
- [0049] 图 7 为所述图 5 和图 6 中所述的分类处理模块的一种组成示意图。

具体实施方式

- [0050] 下面结合附图及具体实施例对本发明再作进一步详细的说明
- [0051] 图 2 为本发明所述自动语音识别方法的一种处理流程图。参见图 2，该流程包括：
- [0052] 步骤 201、对生语料进行语料分类计算，得到一个以上不同类别的分类语料。例如，所述分类语料可以分为人名类、地名类、计算机术语类、医药术语类等等。例如“板蓝根”属于医药术语类的词。一个词也有可能属于多个分类。
- [0053] 步骤 202、针对所述每个分类语料进行语言模型训练计算，得到一个以上对应的分类语言模型。
- [0054] 步骤 203、依据分类的生僻程度为所述各个分类语言模型进行加权插值处理，其中各分类的生僻程度与该分类对应的加权值成正相关关系，即生僻程度越高，则对应的加权值越高，并将加权插值处理后的分类语言模型合并，得到插值语言模型。这样在插值语言模板中，生僻词语所对应的字符串的概率值就会相应提高，从而减少与常用词语对应字符串的概率值的差距，提高生僻词的语音被识别的几率。
- [0055] 步骤 204、依据声学模型和所述插值语言模型构建解码资源。此处假设声学模型已经训练好，本发明可以直接利用现有的声学模型。另外，本领域技术人员知道，在构建解码资源的过程中，还需要词典的参与，来构建解码资源。
- [0056] 步骤 205、依据所述解码资源，对输入的语音进行解码，输出概率值最高的字符串作为所述输入语音的识别结果。
- [0057] 图 3 为本发明所述自动语音识别方法的又一种处理流程图。参见图 3，该流程包括：
- [0058] 步骤 301、根据生语料进行语言模型训练计算，得到主语言模型。此处的语言模型训练为现有的常规语言模型训练。
- [0059] 步骤 302、对生语料进行语料分类计算，得到一个以上不同类别的分类语料。
- [0060] 步骤 303、针对所述每个分类语料进行语言模型训练计算，得到一个以上对应的分类语言模型。
- [0061] 步骤 304~305、依据声学模型和所述主语言模型构建主解码资源，依据所述各分类语言模型构建对应的分类解码资源。所述主解码资源用于在第一次解码时使用，所述分类解码资源用于在第二次解码时使用。
- [0062] 步骤 306、依据所述主解码资源对输入的语音进行解码，即第一次解码，输出概率值 $l(w)$ 排在前 n 名的 n 个字符串。所述概率值 $l(w)$ 为语音对应的字符串在主语言模型中的概率值。
- [0063] 步骤 307、依次根据所述各个分类语言模型对应的各分类解码资源，分别对所述 n

个字符串进行解码，得到每个字符串在每个分类语言模型中的概率值 $n(w)$ 。假设此处有 m 个分类语言模型，则会得到 $n \times m$ 个概率值 $n(w)$ 。然后，将每个字符串在每个分类语言模型中的概率值 $n(w)$ 乘以该字符串在主语言模型中的概率值 $l(w)$ 得到 $n \times m$ 个复合概率 $p(w)$ ，输出复合概率 $p(w)$ 最高的字符串作为所述输入语音的识别结果。

[0064] 在所述步骤 201 和步骤 302 中，所述对生语料进行语料分类计算，得到一个以上不同类别的分类语料的具体方法如图 4 所示，具体包括：

[0065] 步骤 401、根据生语料，计算词与词之间的亲和度矩阵。

[0066] 所述生语料是一种训练文本。本发明通过建立词的亲和度矩阵(也称为词共现矩阵)来描述词之间的语义关系。在人的认知层面上，一个词总是与其它词有关联，而不是孤立存在的。这种关联用一种激活效应可以表示，例如，听到“医生”这个词，马上会联想到“患者”或者“护士”；听到“猫”这个词，立刻会联想到“狗”；听到“男孩”，反应出“女孩”；“喝”联想到“水”。

[0067] 因此在该步骤 401 中，首先要计算每个词与另一个词的词共现度。具体包括：

[0068] 对生语料进行分析，根据公式 $CO_{ij} = \frac{f_{ij} * f_{ji}}{d_{ij} * d_{ji} * f_i * f_j}$ 计算每个词与另一个词的词共现度，并据此构建词与词的词共现矩阵；其中，所述 f_{ij} 为词 i 在词 j 前出现的次数， d_{ij} 为词 i 和词 j 的平均距离， f_i 为词 i 的词频， f_j 为词 j 的词频。

[0069] 根据所述词共现矩阵，以及公式 $A_{ij} = \sqrt{(\sum OR(waf_{ik}, waf_{jk}) - \sum OR(waf_{ki}, waf_{kj}))}$ 计算词与词之间的亲和度，并据此构建词与词之间的亲和度矩阵。

[0070] 所述的亲和度，被定义为两个词入链与入链的重叠部分、出链与出链的重叠部分的几何平均值。显然词亲和度矩阵是一个对称矩阵，即无向的网络。按亲和度大小排序，排在前面的词基本都是同义、近义或非常相关的词。在亲和度网络中，两结点间的边的亲和度越强，说明他们越相关；如果强度很弱甚至两结点不存在边，则表明它们几乎不相关。通过计算 A_{ij} ，可以构建一个词与词之间的协方差矩阵，该协方差矩阵就是亲和度矩阵，该亲和度矩阵中，由于是按亲和度排序，对于亲和度很小的部分可以忽略，因此该亲和度矩阵的维度相比原始的生语料的词特征向量的维度会小很多。

[0071] 步骤 402、利用词频-逆向文件频率(TF-IDF, term frequency - inverse document frequency) 方法从生语料中提取词特征。

[0072] 本发明文本分类中主要应用的模型是文本的向量空间模型(VSM, Vector Space Model)。向量空间模型的基本思想是以文本的特征向量 $\langle W_1, W_2, W_3, \dots, W_n \rangle$ 来表示文本，其中 W_i 为第 i 个特征项的权重。因此基于向量空间模型的分类中关键一步就是如何从文本中提取反映类别的有效特征。在本步骤 402 中，本发明采用 TF-IDF 方法从生语料中提取词特征，用 TF-IDF 特征来表示 w 的权重。

[0073] 在一份给定的文件里，词频(TF, term frequency) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化，以防止它偏向长的文件。同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否。逆向文件频率(IDF, inverse document frequency) 是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，

TF-IDF 倾向于保留文档中较为特别的词语,过滤常用词。因此通过这种 TF-IDF 的方式,可以从生语料中提取出较生僻的词语的词特征。

[0074] 步骤 403、根据所述亲和度矩阵,利用降维方法对所提取出的词特征进行降维处理。

[0075] 在本步骤 403 中,所述降维方法可以有多种。但是在一种优选实施方式中,可以采用主成分分析(PCA, Principal Components Analysis) 降维方法来实现。由于在步骤 402 中所提取出的词特征向量的维度较高,例如此处假设为 N 维,而步骤 401 所述的亲和度矩阵的维度较少,例如此处假设为 M 维,N 远大于 M。那么经过降维处理后,所述 N 维的词特征向量的维度则被降为 M 维。即通过降维处理,可以降低噪声数据的影响,降低时间复杂度和空间复杂度等,可以将那些亲和度小的词与词的组合过滤掉。

[0076] 步骤 404、将降维处理后的词特征输入分类器进行训练,输出一个以上不同类别的分类语料。

[0077] 分类器是一种计算机程序,可以自动将输入的数据分到已知的类别。本步骤 404 中,所述分类器可以采用现有的某种分类器。例如在一种优选实施方式中,所述分类器为支持向量机(SVM, Support Vector Machine) 分类器。经过测试,本发明在 20 个类的分类效果能够达到 92% 的准确率。

[0078] 当然,除了图 4 所述的对生语料进行语料分类计算的方法,本发明还可以采用其它现有的语料分类计算方法对生语料进行分类。但是,图 4 所述的方法的准确率更高,速度更快。

[0079] 与上述方法相对应,本发明还公开了语音识别系统,用于执行上述的方法。

[0080] 图 5 为本发明所述一种语音识别系统的一种组成示意图。参见图 5,该系统包括:

[0081] 分类处理模块 501,用于对生语料进行语料分类计算,得到一个以上不同类别的分类语料。

[0082] 分类语言模型训练模块 502,用于针对所述每个分类语料进行语言模型训练计算,得到一个以上对应的分类语言模型;

[0083] 加权合并模块 503,用于依据分类的生僻程度为所述各个分类语言模型进行加权插值处理,其中各分类的生僻程度与该分类对应的加权值成正相关关系,即生僻程度越高,则对应的加权值越高,将加权插值处理后的分类语言模型合并,得到插值语言模型。

[0084] 资源构建模块 504,用于依据声学模型和所述插值语言模型构建解码资源。

[0085] 解码器 505,用于依据所述解码资源,对输入的语音进行解码,输出概率值最高的字符串作为所述输入语音的识别结果。

[0086] 图 6 为本发明所述又一种语音识别系统的一种组成示意图。参见图 6,该系统包括:

[0087] 主语言模型训练模块 601,用于根据生语料进行语言模型训练计算,得到主语言模型。此处的语言模型训练为现有的常规语言模型训练。

[0088] 分类处理模块 602,用于对生语料进行语料分类计算,得到一个以上不同类别的分类语料。

[0089] 分类语言模型训练模块 603,用于针对所述每个分类语料进行语言模型训练计算,得到一个以上对应的分类语言模型。

- [0090] 主资源构建模块 604,用于依据声学模型和所述主语言模型构建主解码资源。
- [0091] 分类资源构建模块 605,用于依据所述各分类语言模型构建对应的分类解码资源。
- [0092] 第一解码器 606,用于依据所述主解码资源对输入的语音进行解码,输出概率值 $l(w)$ 排在前 n 名的 n 个字符串;
- [0093] 第二解码器 607,用于依次根据所述各个分类语言模型对应的各分类解码资源,分别对所述 n 个字符串进行解码,得到每个字符串在每个分类语言模型中的概率值 $n(w)$;将每个字符串在每个分类语言模型中的概率值 $n(w)$ 乘以该字符串在主语言模型中的概率值 $l(w)$ 得到复合概率 $p(w)$,输出复合概率 $p(w)$ 最高的字符串作为所述输入语音的识别结果。
- [0094] 图 7 为所述图 5 和图 6 中所述的分类处理模块的一种组成示意图。参见图 7,所述分类处理模块具体包括:
- [0095] 亲和度矩阵模块 701,用于根据生语料,计算词与词之间的亲和度矩阵。具体的计算方法请参考上述步骤 401 至步骤 404。
- [0096] 特征提取模块 702,用于利用 TF-IDF 方法从生语料中提取词特征。
- [0097] 降维模块 703,用于根据所述亲和度矩阵,利用降维方法对所提取出的词特征进行降维处理。在一种优选实施方式中,所述降维模块为 PCA 降维模块。
- [0098] 分类器 704,用于对降维处理后的词特征进行训练,输出一个以上不同类别的分类预料。在一种优选实施方式中,所述分类器为 SVM 分类器。
- [0099] 本发明所述的语音识别方法和系统可以应用在垂直领域的语音识别、语音关键字的识别,以及语音问答系统等技术领域中。而且可以支持多平台,包括嵌入式平台和 PC 平台。
- [0100] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本发明保护的范围之内。

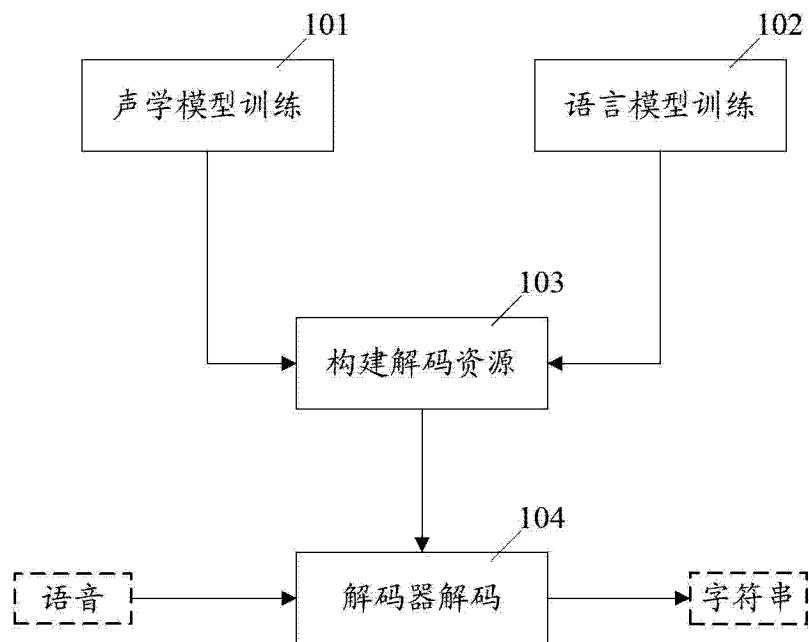


图 1

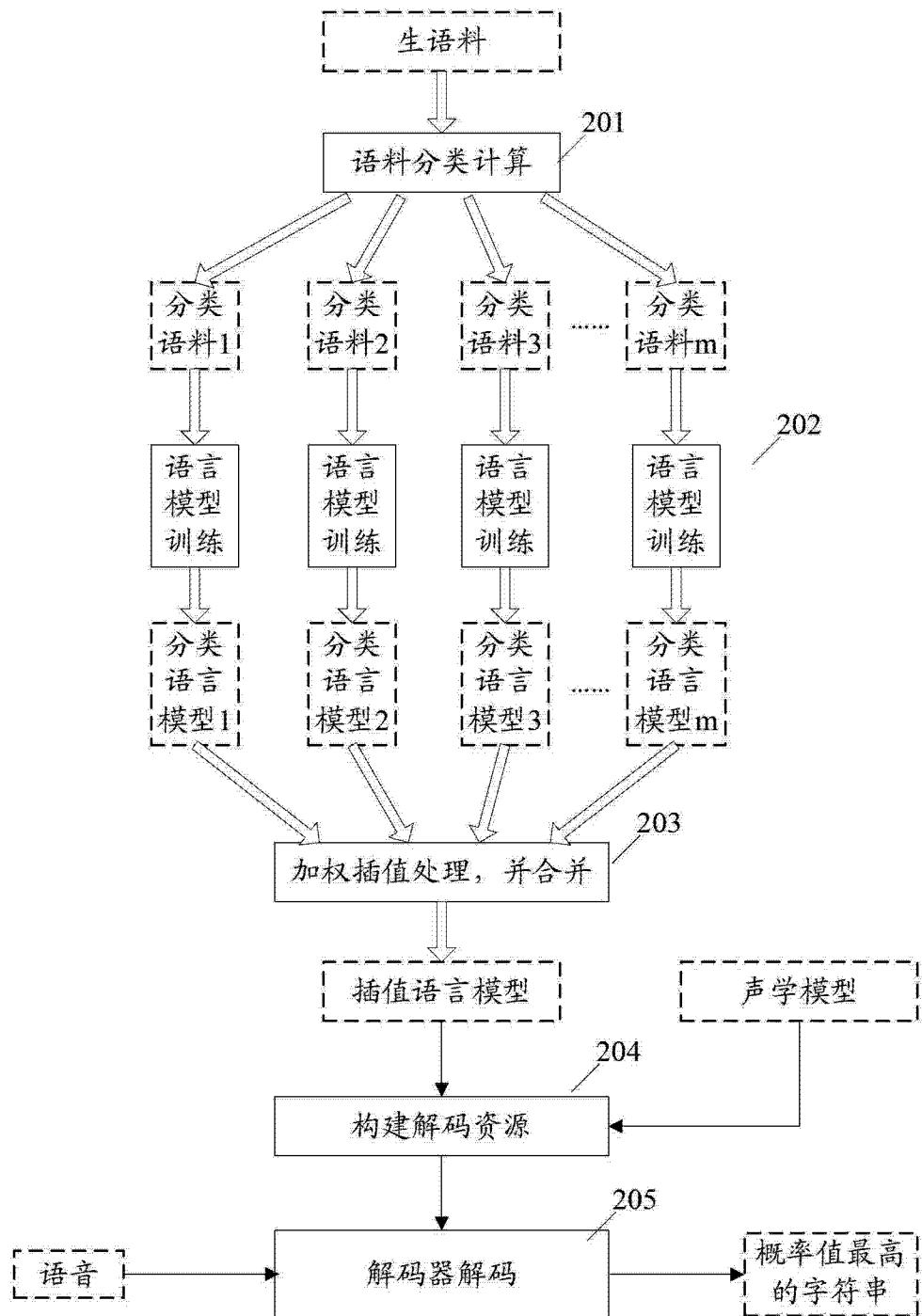


图 2

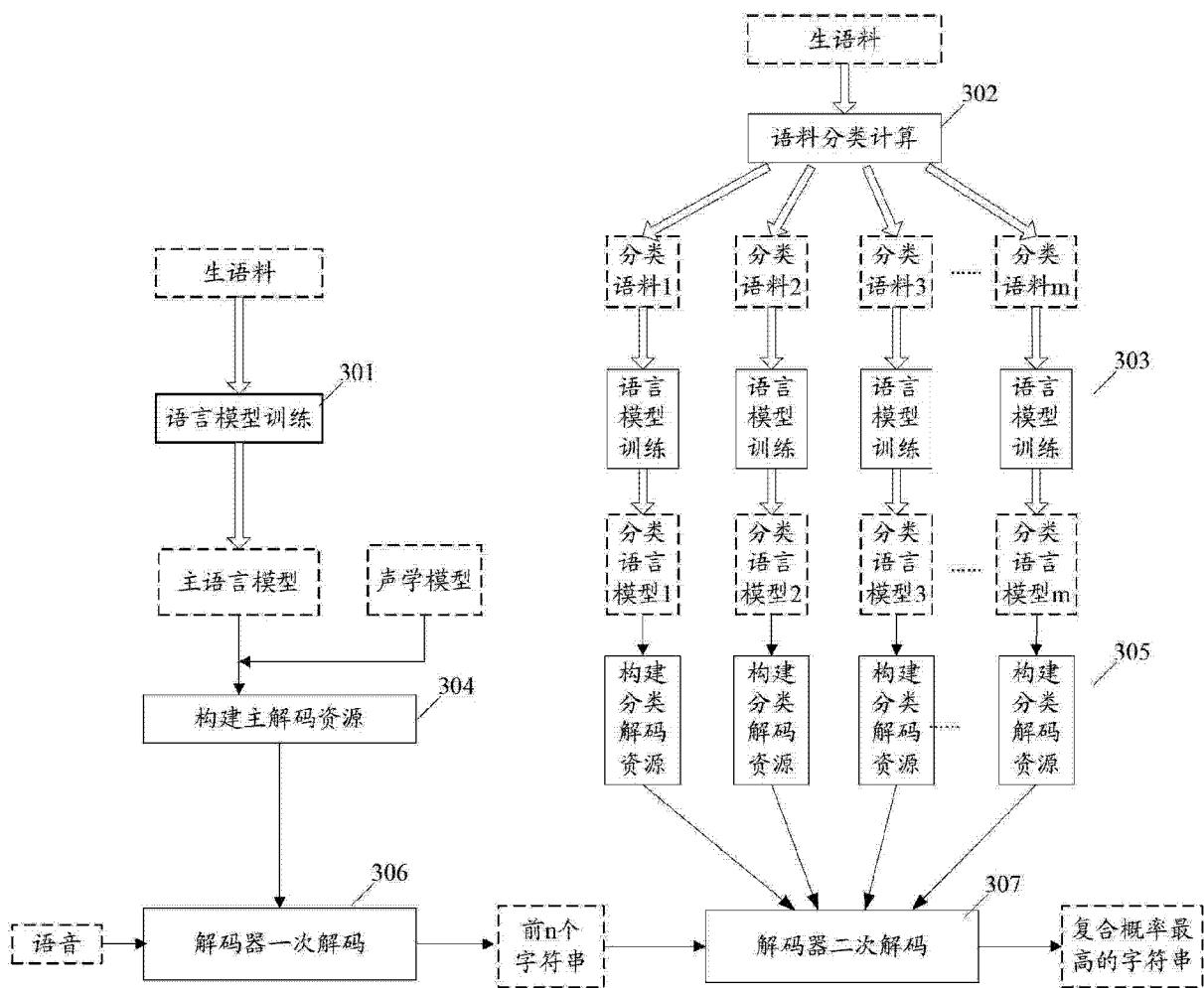


图 3

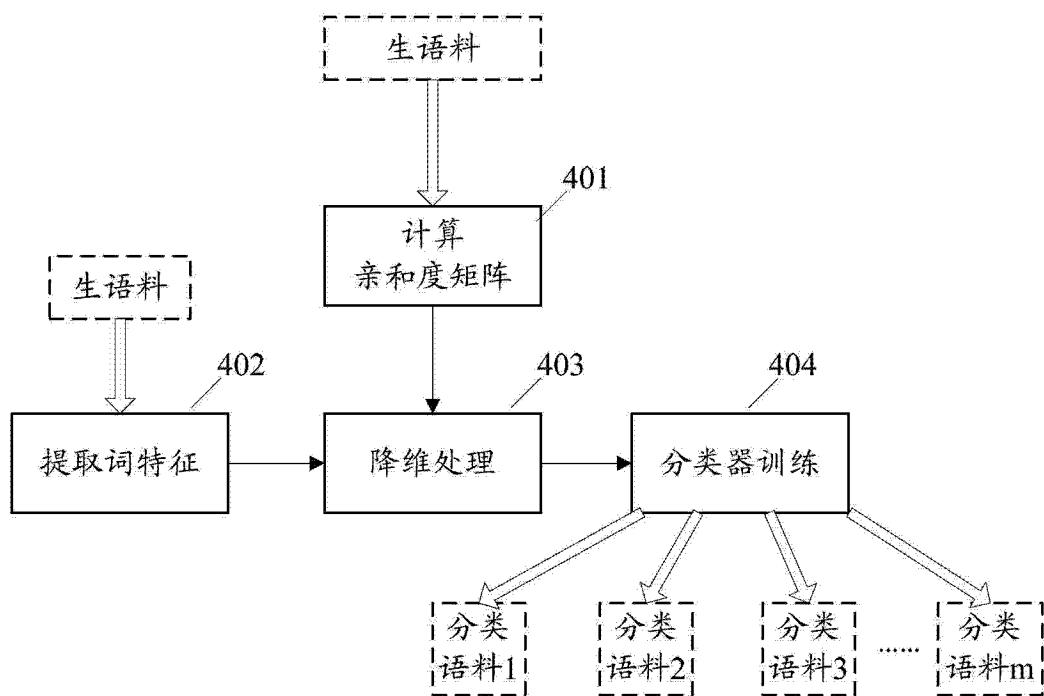


图 4

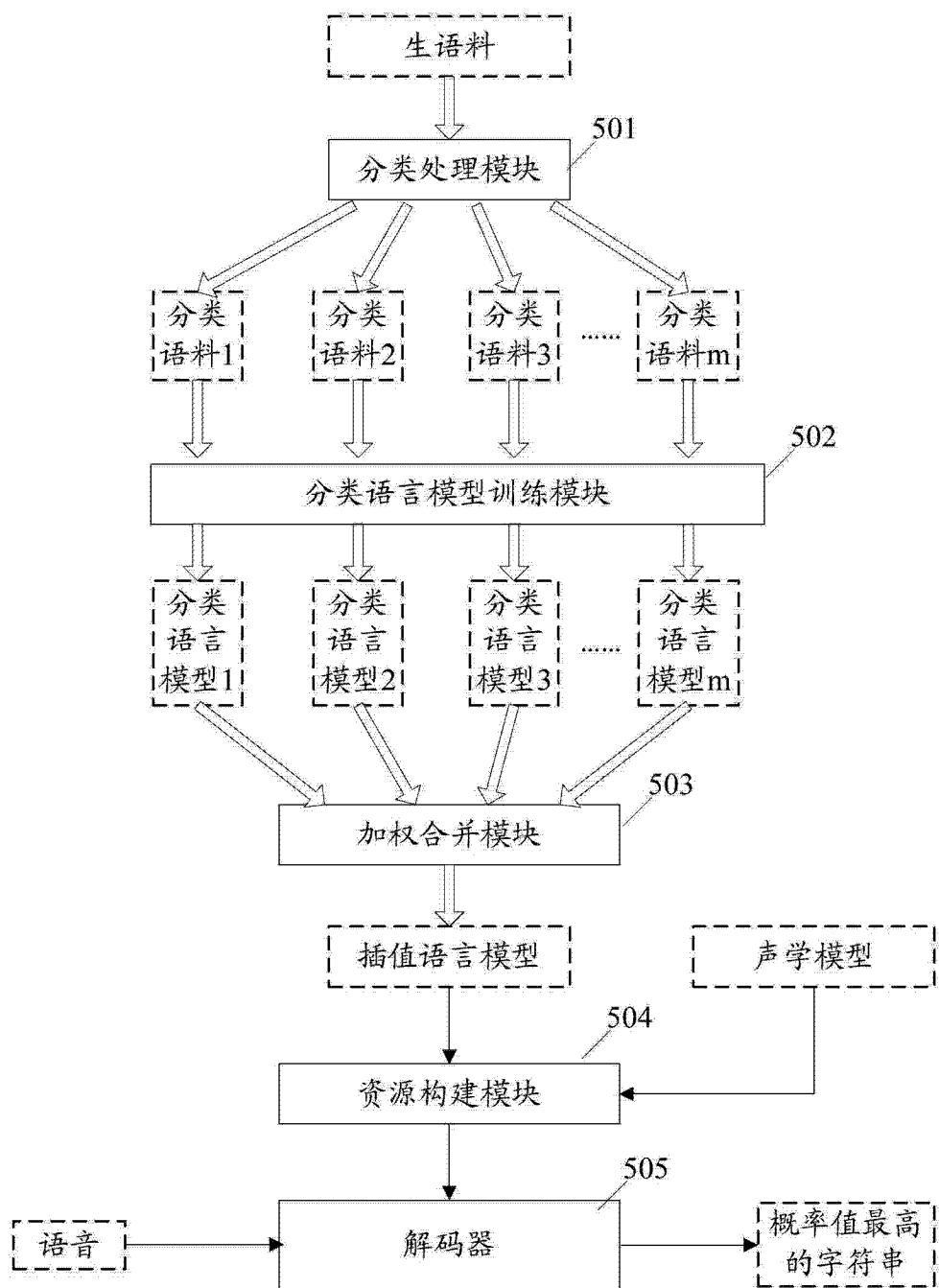


图 5

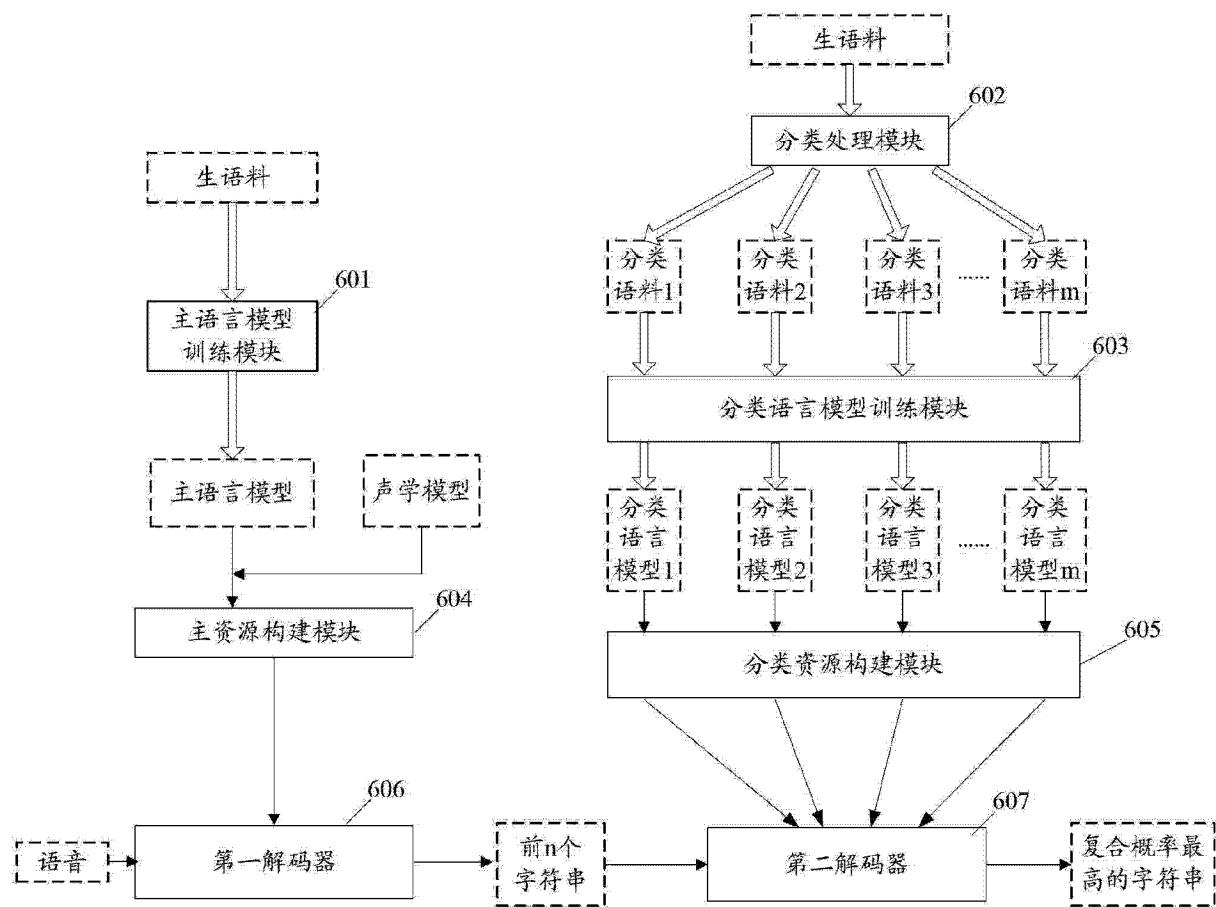


图 6

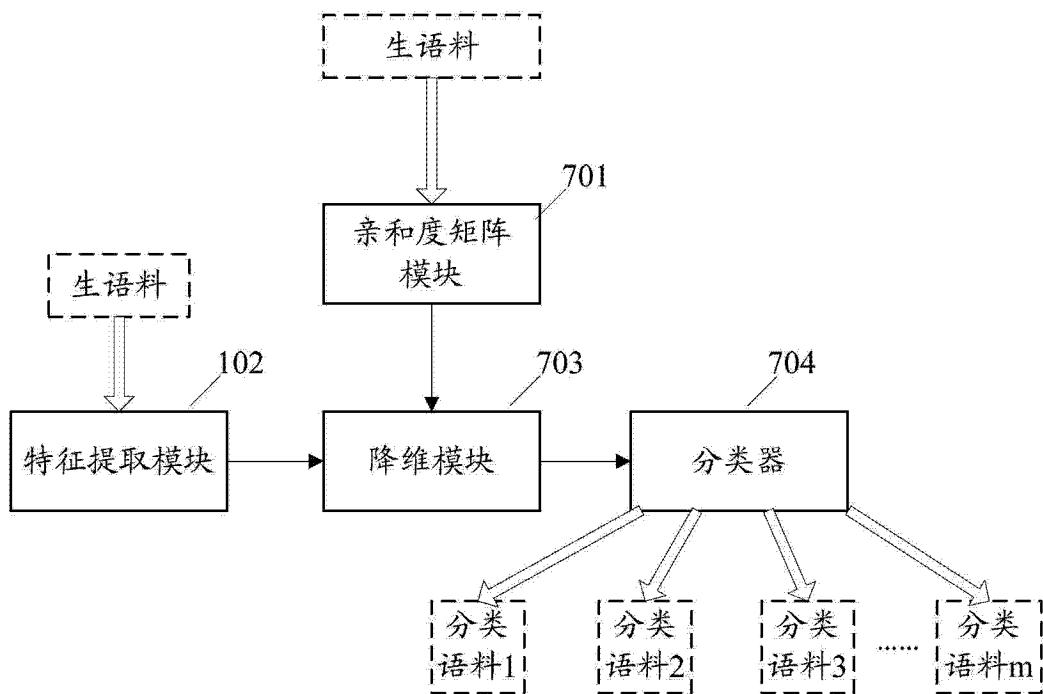


图 7