(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 103578471 A
(43)申请公布日 2014.02.12

(21)申请号 201310489578.3

(22)申请日 2013.10.18

(71)申请人 威盛电子股份有限公司
地址 中国台湾新北市新店区中正路533号8楼

(72)发明人 张国峰 朱逸斐

(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
代理人 刘新宇

(51)Int.Cl.
G10L 15/183(2013.01)
G10L 15/28(2013.01)

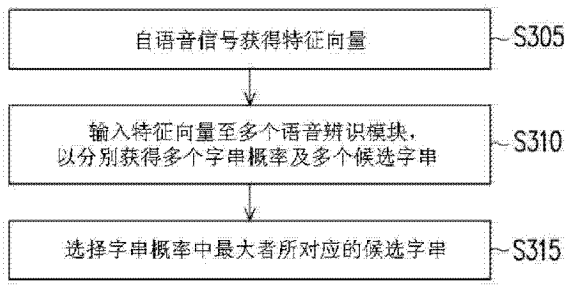
权利要求书2页 说明书6页 附图3页

(54)发明名称

语音辨识方法及其电子装置

(57)摘要

一种语音辨识方法及其电子装置。本语音辨识方法包括:将自语音信号获得的特征向量输入至多个语音辨识模块,而自上述语音辨识模块分别获得多个字串概率及多个候选字串,并且,选择上述字串概率中最大者所对应的候选字串,以作为语音信号的辨识结果。本发明可自动地辨识出语音信号所对应的语言。



CN 103578471 A

1. 一种语音辨识方法，用于一电子装置，其特征在于，该语音辨识方法包括：

自一语音信号获得一特征向量；

输入该特征向量至多个语音辨识模块，并自所述语音辨识模块分别获得多个字串概率及多个候选字串，其中所述语音辨识模块分别对应至多种语言；以及

选择所述字串概率中最大者所对应的候选字串，以作为该语音信号的辨识结果。

2. 根据权利要求 1 所述的语音辨识方法，其特征在于，输入该特征向量至所述语音辨识模块，并自所述语音辨识模块分别获得所述字串概率与所述候选字串的步骤包括：

输入该特征向量至每一个所述语音辨识模块的声学模型，并基于对应的声学词典，获得相对于每一种语言的候选词；以及

输入所述候选词至每一个所述语音辨识模块的语言模型，以获得所述语言对应的所述候选字串以及所述字串概率。

3. 根据权利要求 2 所述的语音辨识方法，其特征在于，还包括：

基于所述语言各自对应的语音数据库，经由训练而获得上述声学模型与上述声学词典；以及

基于所述语言各自对应的语料库，经由训练而获得上述语言模型。

4. 根据权利要求 1 所述的语音辨识方法，其特征在于，还包括：

通过一输入单元接收该语音信号。

5. 根据权利要求 1 所述的语音辨识方法，其特征在于，自该语音信号获得该特征向量的步骤包括：

将该语音信号切割为多个音框；以及

自每一所述音框取得多个特征参数，借以获得该特征向量。

6. 一种电子装置，其特征在于，包括：

一处理单元；

一储存单元，耦接至该处理单元，且储存多个程序码片段，以供该处理单元执行；以及

一输入单元，耦接至该处理单元，且接收一语音信号；

其中，该处理单元通过所述程序码片段来驱动多种语言所对应的多个语音辨识模块，并执行：自该语音信号获得一特征向量，并且输入该特征向量至所述语音辨识模块，而自所述语音辨识模块分别获得多个字串概率及多个候选字串；以及选出所述字串概率中最大者所对应的候选字串。

7. 根据权利要求 6 所述的电子装置，其特征在于，该处理单元输入该特征向量至每一所述语音辨识模块的声学模型，并基于对应的声学词典，获得相对于每一所述语言的候选词，并且输入所述候选词至每一所述语音辨识模块的语言模型，以获得所述语言对应的所述候选字串以及所述字串概率。

8. 根据权利要求 7 所述的电子装置，其特征在于，该处理单元基于所述语言各自对应的语音数据库，经由训练而获得上述声学模型与上述声学词典，并且基于所述语言各自对应的语料库，经由训练而获得上述语言模型。

9. 根据权利要求 6 所述的电子装置，其特征在于，该处理单元通过所述程序码片段来驱动一特征撷取模块，以执行：切割该语音信号为多个音框，并自每一所述音框取得多个特征参数，借以获得该特征向量。

10. 根据权利要求 6 所述的电子装置，其特征在于，还包括：

一输出单元，输出所述字串概率中最大者所对应的候选字串。

语音辨识方法及其电子装置

技术领域

[0001] 本发明有关于一种语音辨识技术，且特别是有关于一种可用于识别不同语言的语音辨识方法及其电子装置。

背景技术

[0002] 语音辨识 (speech recognition) 毫无疑问的是一种热门的研究与商业课题。语音辨识通常是对输入的语音取出特征参数，再与数据库的样本相比对，找出与输入相异度低的样本取出。

[0003] 目前常见做法大都是先采集语音语料(如录下来的人的语音)，然后由人工进行标注(即，对每一句语音标注上对应的文字)，然后使用这些语料来训练声学模型和声学词典。声学模型是一种统计分类器。目前做法常使用混合高斯模型(Gaussian Mixture Model)，它将输入的语音分类到基本的音素(phone)。而音素是组成需要识别的语言的基本音标及音间过渡(transition between phones，每个音素可以有多个状态，比如 3 个，叫做三音素(tri-phone)，一个音标通常用一个音素表示，这个音素的前面的状态包含前面一个音素过渡到这个音素的状态，后面的状态包含这个音素过渡到下面一个音素的状态。)，此外，加上一些非语音的音素，如咳嗽声。而声学词典一般是由被识别语言的单词组成，通过隐藏式马可夫模型(Hidden Markov Model，HMM)将声学模型输出的音组成单词。

[0004] 然而，目前的做法存在如下问题。问题 1 ：倘若用户的非标准发音(如翘舌音不分、前后鼻音不分等)进入声学模型，将会造成声学模型的模糊性变大。如拼音“in”在声学模型中会给出比较大的概率为“ing”，而这个为了不标准发音的妥协，会导致整体错误率的升高。问题 2 ：由于不同地区的发音习惯不同，非标准发音有多种变形，导致声学模型的模糊性变得更大，因而使得识别准确率的进一步降低。问题 3 ：无法识别方言，如标准普通话、沪语、粤语、闽南语等。

发明内容

[0005] 本发明提供一种语音辨识方法及其电子装置，可自动地辨识出语音信号所对应的语言。

[0006] 本发明的语音辨识方法，用于电子装置。该语音辨识方法包括 ：自语音信号获得特征向量 ；输入特征向量至多个语音辨识模块，并自上述语音辨识模块分别获得多个字串概率及多个候选字串，其中上述语音辨识模块分别对应至多种语言 ；以及选择上述字串概率中最大者所对应的候选字串，以作为语音信号的辨识结果。

[0007] 在本发明的一实施例中，上述输入特征向量至上述语音辨识模块，并自上述语音辨识模块分别获得上述字串概率与上述字串的步骤包括 ：输入特征向量至上述各个语音辨识模块的声学模型，并基于对应的声学词典，获得相对于各种语言的候选词 ；以及输入上述候选词至上述各语音辨识模块的语言模型，以获得各种语言对应的候选字串以及字串概率。

[0008] 在本发明的一实施例中，上述语音辨识方法还包括：基于各种语言对应的语音数据库，经由训练而获得上述声学模型与上述声学词典；以及基于各种语言对应的语料库，经由训练而获得上述语言模型。

[0009] 在本发明的一实施例中，上述语音辨识方法还包括：通过输入单元接收语音信号。

[0010] 在本发明的一实施例中，上述自语音信号获得特征向量的步骤包括：将语音信号切割为多个音框，并自各音框取得多个特征参数，借以获得特征向量。

[0011] 本发明另提出一种电子装置，包括输入单元、储存单元以及处理单元。输入单元用以接收语音信号。储存单元中储存有多个程序码片段。处理单元耦接至输入单元以及储存单元。处理单元通过上述程序码片段来驱动多种语言所对应的多个语音辨识模块，并执行：自语音信号获得特征向量，并且输入特征向量至上述语音辨识模块，而自上述语音辨识模块分别获得多个字串概率及多个候选字串；以及选出上述字串概率中最大者所对应的候选字串。

[0012] 在本发明的一实施例中，该处理单元输入该特征向量至每一所述语音辨识模块的声学模型，并基于对应的声学词典，获得相对于每一所述语言的候选词，并且，该处理单元输入所述候选词至每一所述语音辨识模块的语言模型，以获得所述语言对应的所述候选字串以及所述字串概率。

[0013] 在本发明的一实施例中，该处理单元基于所述语言各自对应的语音数据库，经由训练而获得上述声学模型与上述声学词典，并且基于所述语言各自对应的语料库，经由训练而获得上述语言模型。

[0014] 在本发明的一实施例中，该处理单元通过所述程序码片段来驱动一特征撷取模块，以执行：切割该语音信号为多个音框，并自每一所述音框取得多个特征参数，借以获得该特征向量。

[0015] 在本发明的一实施例中，上述电子装置还包括有一输出单元。此输出单元用以输出上述字串概率中最大者所对应的候选字串。

[0016] 基于上述，本发明将语音信号分别在不同的语音辨识模块中来进行解码，借以获得每个语音辨识模块所对应的候选字串的输出以及候选字串的字串概率。并且，以字串概率最大者作为语音信号的辨识结果。据此，可自动地辨识出语音信号所对应的语言，而不用使用者事先手动选择所欲使用的语音辨识模块的语言。

[0017] 为让本发明的上述特征和优点能更明显易懂，下文特举实施例，并配合所附图式作详细说明如下。

附图说明

[0018] 图 1A 是依照本发明一实施例的电子装置的方块图。

[0019] 图 1B 是依照本发明另一实施例的电子装置的方块图。

[0020] 图 2 是依照本发明一实施例的语音辨识模块的示意图。

[0021] 图 3 是依照本发明一实施例的语音辨识方法的流程图。

[0022] 图 4 是依照本发明一实施例的多语言模型的架构示意图。

[0023] 附图中符号的简单说明如下：

[0024] 110：处理单元

[0025] 120：储存单元

[0026] 130：输入单元

[0027] 140：输出单元

[0028] 21：语音数据库

[0029] 22：语料库

[0030] 200、A、B、C：语音辨识模块

[0031] 210：声学模型

[0032] 220：声学词典

[0033] 230：语言模型

[0034] 240：解码器

[0035] 410：特征撷取模块

[0036] 411A：第一声学模型

[0037] 411B：第二声学模型

[0038] 411C：第三声学模型

[0039] 412A：第一声学词典

[0040] 412B：第二声学词典

[0041] 412C：第三声学词典

[0042] 413A：第一语言模块

[0043] 413B：第二语言模块

[0044] 413C：第三语言模块

[0045] 414A：第一解码器

[0046] 414B：第二解码器

[0047] 414C：第三解码器

[0048] S：语音信号

[0049] S305 ～ S315：步骤。

具体实施方式

[0050] 在传统语音辨识方法中，普遍存在如下问题，即，由于在不同地区的方言中的模糊音、使用者发音习惯的不同、或是不同的语言，会导致辨识率的精准度受到影响。为此，本发明提出一种语音辨识方法及其电子装置，可在原有语音识别的基础上，改进辨识率的精准度。为了使本发明的内容更为明了，以下特举实施例作为本发明确实能够据以实施的范例。

[0051] 图 1A 是依照本发明一实施例的电子装置的方块图。请参照图 1A，电子装置 100 包括处理单元 110、储存单元 120 以及输入单元 130。电子装置 100 例如为手机、智能手机、个人数字助理（Personal Digital Assistant，PDA）、平板计算机、笔记型计算机、桌上型计算机、车用计算机等具有运算功能的装置。

[0052] 在此，处理单元 110 耦接至储存单元 120 以及输入单元 130。处理单元 110 例如为中央处理单元（Central Processing Unit，CPU）或微处理器（microprocessor）等，其用以执行电子装置 100 中的硬件、固件以及处理软件中的数据。储存单元 120 例如为非易失性存储器（Non-volatile memory，NVM）、动态随机存取存储器（Dynamic Random Access

Memory，DRAM）或静态随机存取存储器（Static Random Access Memory，SRAM）等。

[0053] 在此，以程序码来实现电子装置100的语音辨识方法而言，储存单元120中储存有多个程序码片段。上述程序码片段在被安装后，会由处理单元110来执行。这些程序码片段包括多个指令，处理单元110通过这些指令来执行语音辨识方法的多个步骤。在本实施例中，电子装置100仅包括一个处理单元110，而在其他实施例中，电子装置100亦可包括多个处理单元，而由这些处理单元来执行被安装的程序码片段。

[0054] 输入单元130接收一语音信号。例如，输入单元130为麦克风，其接收使用者所发出的模拟语音信号，并将模拟语音信号转换为数字语音信号后，传送至处理单元110。

[0055] 具体而言，处理单元110通过上述程序码片段来驱动多种语音所对应的多个语音辨识模块，并执行如下步骤：自语音信号获得特征向量，并且输入特征向量至上述语音辨识模块，而自上述语音辨识模块分别获得多个字串概率及多个候选字串；以及选出字串概率中最大者所对应的候选字串。

[0056] 另外，在其他实施例中，电子装置100还可包括一输出单元。举例来说，图1B是依照本发明另一实施例的电子装置的方块图。请参照图1B，电子装置100包括处理单元110、储存单元120、输入单元130以及输出单元140。处理单元110耦接至储存单元120、输入单元130及输出单元140。关于处理单元110、储存单元120及输入单元130相关描述已阐明于上述，故在此不再赘述。

[0057] 输出单元140例如为阴极射线管（Cathode Ray Tube，CRT）显示器、液晶显示器（Liquid Crystal Display，LCD）、等离子显示器（Plasma Display）、触控显示器（Touch Display）等显示单元，以显示所获得的字串概率中最大者所对应的候选字串。或者，输出单元140亦可以是扬声器，以播放所获得的字串概率中最大者所对应的候选字串。

[0058] 在本实施例中，针对不同的语言或方言，建立不同的语音辨识模块，即，针对不同的语言或方言，分别建立一套声学模型（acoustic model）与语言模型（language model）。

[0059] 声学模型是语音辨识模块中最为重要的部分之一，一般可采用隐藏式马可夫模型（Hidden Markov Model，HMM）进行建模。语言模型（language model）主要是利用机率统计的方法来揭示语言单位内在的统计规律，其中N元语法（N-Gram）简单有效而被广泛使用。

[0060] 下面举一实施例来说明。

[0061] 图2是依照本发明一实施例的语音辨识模块的示意图。请参照图2，语音辨识模块200主要包括声学模型210、声学词典220、语言模型230以及解码器240。

[0062] 其中，声学模型210与声学词典220是由语音数据库21经训练而获得，语言模型230是由语料库（text corpus）22经训练而获得。

[0063] 具体而言，声学模型210多是采用基于一阶HMM进行建模。声学词典220包含语音辨识模块200所能处理的词汇及其发音。语言模型230对语音辨识模块200所针对的语言进行建模。例如，语言模型230是基于历史信息的模型（History-based Model）的设计理念，即，根据经验法则，统计先前已出现的一连串事件与下一个出现的事件之间的关系。解码器240是语音辨识模块200的核心之一，其任务是对输入的语音信号，根据声学模型210、声学词典220以及语言模型230，寻找能够以最大概率输出的候选字串。

[0064] 举例来说，利用声学模型210获得对应的音素（phone）或音节（syllable），再由声学词典220来获得对应的字或词，之后由语言模型230来判断一连串的字成为句子的概率。

[0065] 如下即搭配上述图 1A 的电子装置 100 来进一步说明语音辨识方法的各步骤。图 3 是依照本发明一实施例的语音辨识方法的流程图。请同时参照图 1A 及图 3，在步骤 S305 中，处理单元 110 自语音信号获得特征向量。

[0066] 举例来说，模拟的语音信号会转成数字的语音信号，并将语音信号切割为多个音框，而这些音框中的两相邻音框之间可以有一段重叠区域。之后，再从每个音框中取出特征参数而获得一特征向量。例如，可利用梅尔倒频谱系数（Mel-frequency Cepstral Coefficients，MFCC）自音框中取出 36 个特征参数，而获得一个 36 维的特征向量。

[0067] 接着，在步骤 S310 中，处理单元 110 将特征向量输入至多个语音辨识模块，而分别获得多个字串概率以及多个候选字串。具体而言，将特征向量输入至各语音辨识模块的声学模型，并基于对应的声学词典，而获得相对于各种语言的候选词。并且，将各种语言的候选词输入至各语音辨识模块的语言模型，以获得各种语言对应的候选字串以及字串概率。

[0068] 举例来说，图 4 是依照本发明一实施例的多语言模型的架构示意图。本实施例以 3 种语言为例，而在其他实施例中，也可以为 2 种语言或 3 种以上的语言。

[0069] 请参照图 4，本实施例提供有 3 种语言的语音辨识模块 A、B、C。例如，语音辨识模块 A 用以识别标准普通话，语音辨识模块 B 用以识别粤语，语音辨识模块 C 用以识别闽南话。在此，将所接收的语音信号 S 输入至特征撷取模块 410，借以获得多个音框的特征向量。

[0070] 语音辨识模块 A 包括第一声学模型 411A、第一声学词典 412A、第一语言模块 413A 以及第一解码器 414A。其中，第一声学模型 411A 与第一声学词典 412A 是由标准普通话的语音数据库经由训练而获得，而第一语言模块 413A 则是由标准普通话的语料库经由训练而获得。

[0071] 语音辨识模块 B 包括第二声学模型 411B、第二声学词典 412B、第二语言模块 413B 以及第二解码器 414B。其中，第二声学模型 411B 与第二声学词典 412B 是由粤语的语音数据库经由训练而获得，而第二语言模块 413B 则是由粤语的语料库经由训练而获得。

[0072] 语音辨识模块 C 包括第三声学模型 411C、第三声学词典 412C、第三语言模块 413C 以及第三解码器 414C。其中，第三声学模型 411C 与第三声学词典 412C 是由闽南话的语音数据库经由训练而获得，而第三语言模块 413C 则是由闽南话的语料库经由训练而获得。

[0073] 接着，将特征向量分别输入至语音辨识模块 A、B、C，而由语音辨识模块 A 获得第一候选字串 SA 及其第一字串概率 PA；由语音辨识模块 B 获得第二候选字串 SB 及其第二字串概率 PB；由语音辨识模块 C 获得第三候选字串 SC 及其第三字串概率 PC。

[0074] 即，语音信号 S 会经由各个语音辨识模块而识别出在各种语言下的声学模块与语言模块中具有最高概率的候选字串。

[0075] 之后，在步骤 S315 中，处理单元 110 选择字串概率最大者所对应的候选字串。以图 4 而言，假设第一字串概率 PA、第二字串概率 PB、第三字串概率 PC 分别为 90%、20%、15%，因此，处理单元 110 选择第一字串概率 PA（90%）对应的第一候选字串 SA，以作为语音信号的辨识结果。另外，还可进一步将所选出的候选字串，如第一候选字串 SA，输出至如图 1B 所示的输出单元 140。

[0076] 综上所述，对于不同的语言或方言，建立不同的声学模型和语音模型，并分别训练。而对于语音信号的输入，分别在不同的声学模型和语言模型中来进行解码，解码结果不仅可以得到每个语言模型所对应的候选字串的输出，同时也能得到这个候选字串的概率。

据此，在具备多种语言模型的状况下，选出概率最大的输出，作为语音信号的辨识结果。相比于传统方法，本发明中使用单独的语言模型都是准确的，不会存在语言混淆的问题。此外，不仅可以正确进行声音至文字的转换，同时还可知道语言或方言的类型。这对后续的机器语音对话会有帮助，例如对粤语发音的输入直接用粤语回答。另外，在新引入另一种语言或方言的情况下，亦不会对原有的模型产生混淆。

[0077] 以上所述仅为本发明较佳实施例，然其并非用以限定本发明的范围，任何熟悉本项技术的人员，在不脱离本发明的精神和范围内，可在此基础上做进一步的改进和变化，因此本发明的保护范围当以本申请的权利要求书所界定的范围为准。

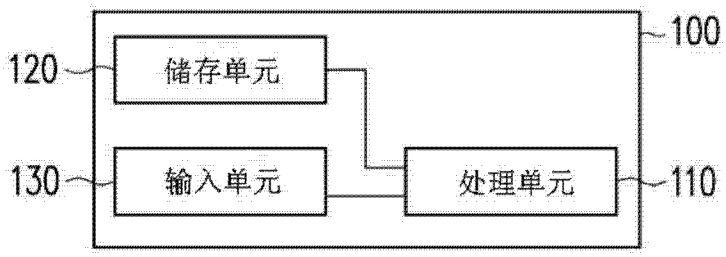

100
120
储存单元
130
输入单元
处理单元
110

图 1A

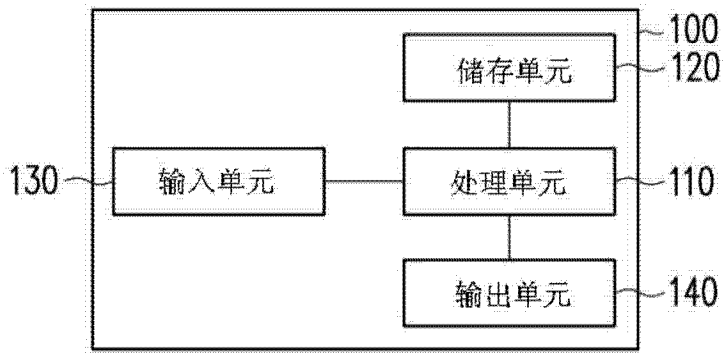

100
储存单元
120
130
输入单元
处理单元
110
输出单元
140

图 1B

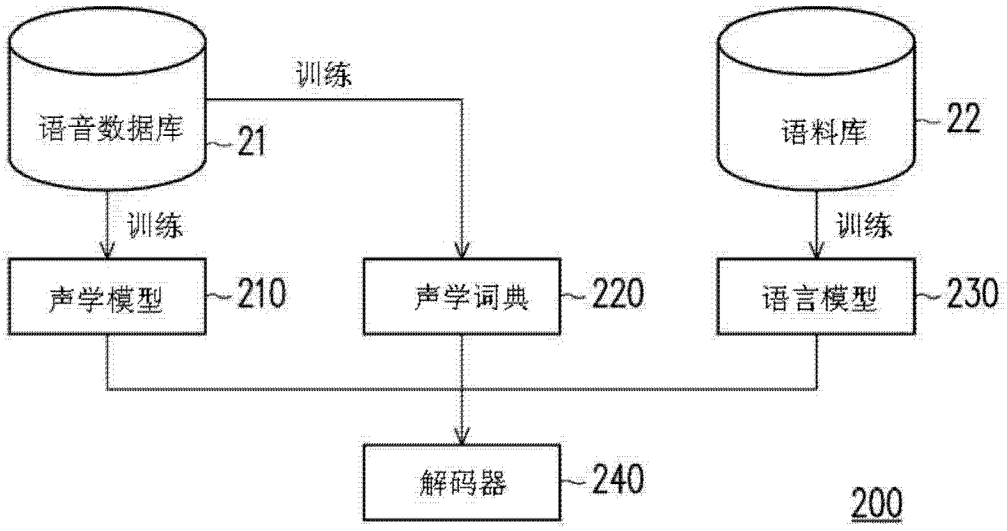

训练
语音数据库
21
语料库
22
训练
训练
声学模型
210
声学词典
220
语言模型
230
解码器
240
200

图 2

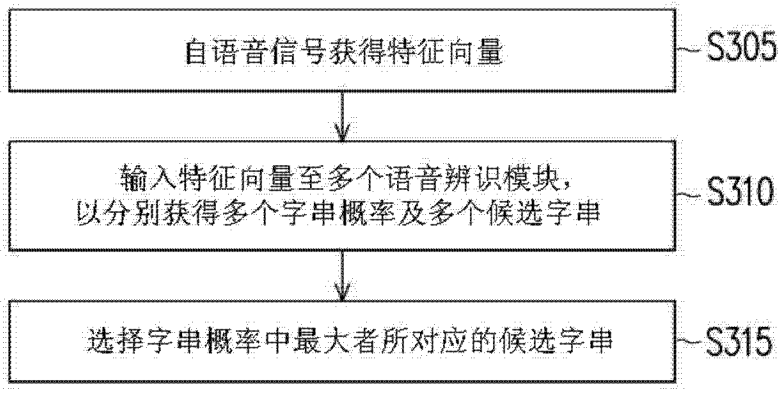

自语音信号获得特征向量
S305
输入特征向量至多个语音辨识模块，
以分别获得多个字串概率及多个候选字串
S310
选择字串概率中最大者所对应的候选字串
S315

图 3

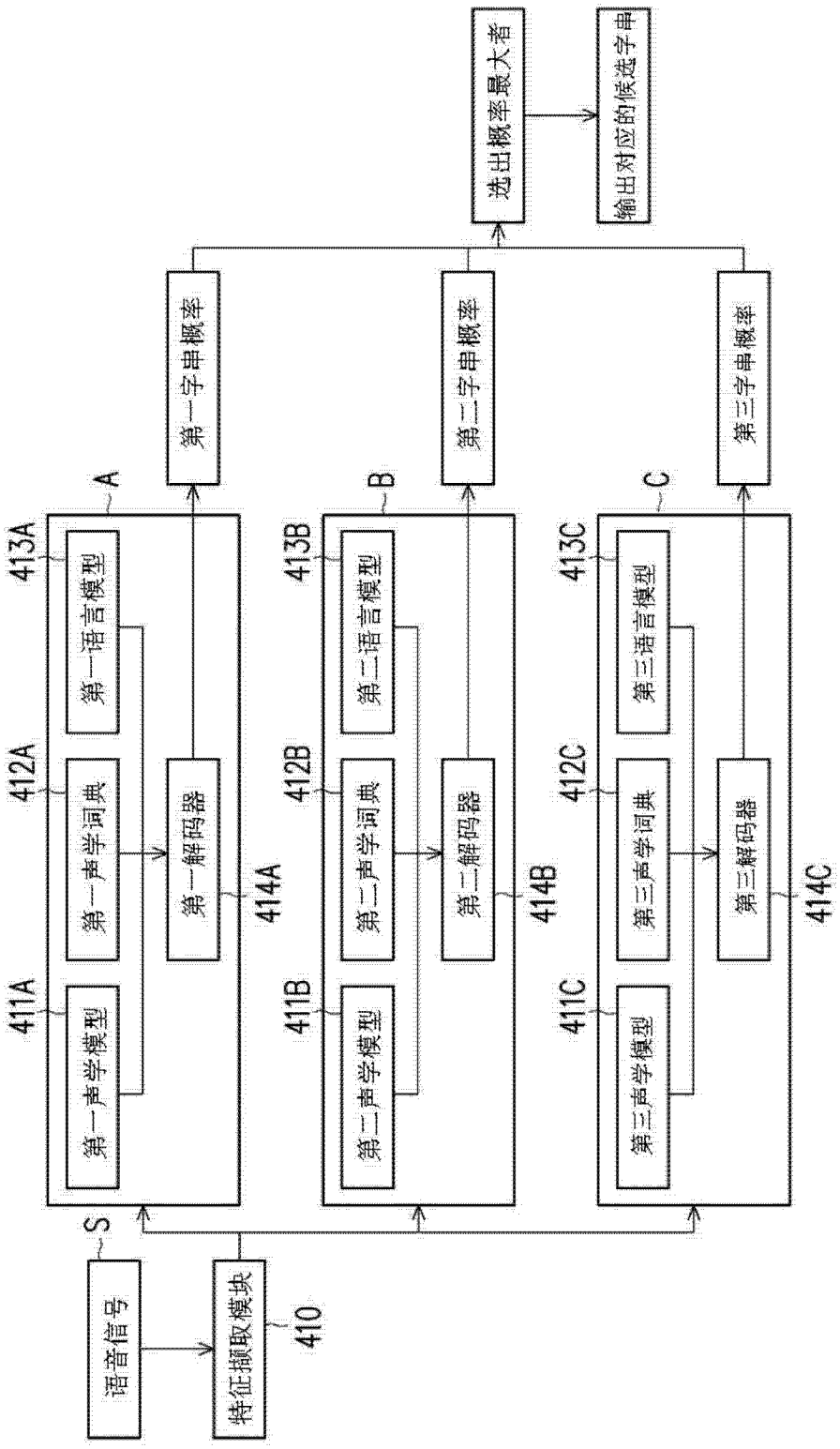

选出概率最大者
输出对应的候选字串
第一字串概率
第二字串概率
第三字串概率
A
B
C
413A
412A
411A
第一语言模型
第一声学词典
第一声学模型
第一解码器
414A
413B
412B
411B
第二语言模型
第二声学词典
第二声学模型
第二解码器
414B
413C
412C
411C
第三语言模型
第三声学词典
第三声学模型
第三解码器
414C
S
语音信号
特征撷取模块
410

图 4