(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101154379 B
(45)授权公告日 2011.11.23

(21)申请号 200610152758.2

(22)申请日 2006.09.27

(73)专利权人 夏普株式会社
地址 日本大阪府

(72)发明人 李丰芹 吴亚栋 杨庆涛 陈晨

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 王波波

(51)Int.Cl.
G10L 15/08(2006.01)
G10L 15/00(2006.01)
G10L 15/02(2006.01)
G10L 15/28(2006.01)

(56)对比文件

CN 1455389 A,2003.11.12, 权利要求5、说明书第5-6页.

EP 0838803 B1,2002.09.11, 全文.

吴旭辉等. 一种基于特征空间轨迹匹配方式的语音关键词检测法.《第六届全国人机语音通讯学术会议论文集》.2001,215-218.

吴旭辉等. 基于特征空间轨迹匹配方式的语音关键词检测法.《计算机工程与应用》.2003,(第36期),83-86.

审查员 万济萍

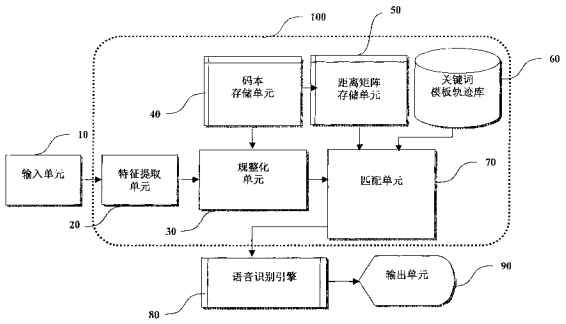
权利要求书 2 页 说明书 11 页 附图 7 页

(54)发明名称

定位语音中的关键词的方法和设备以及语音识别系统

(57)摘要

公开了一种定位语音中的关键词的方法和设备以及语音识别系统。所述方法包括步骤:提取构成待识别语音的各个帧的特征参数矢量,以形成用于描述待识别语音的特征参数矢量序列;利用包含多个码本矢量的码本对特征参数矢量序列进行规整化处理,以得到待识别语音在矢量空间中的特征轨迹;以及匹配预先存储的关键词模板轨迹和所述特征轨迹,以确定关键词的位置。利用本发明,由于基于同一码本来描述关键词模板轨迹和待识别语音的特征空间轨迹,所以在具有相同音韵特征结构的音频帧进行线性移动匹配时不需要重新刻度,这提高了定位和识别的速度,并同时保证了识别的精度。



CN 101154379 B

1. 一种定位待识别语音中的关键词的方法，包括步骤：

提取构成所述待识别语音的各个帧的特征参数矢量，以形成用于描述待识别语音的特征参数矢量序列；

利用包含多个码本矢量的码本对特征参数矢量序列进行规整化处理，以得到待识别语音在矢量空间中的特征轨迹；以及

匹配预先存储的关键词模板轨迹和所述特征轨迹，以确定关键词的位置，

其中，所述码本是通过下面的步骤获得的：

从基于隐马尔可夫模型的声学模型中提取描述各个音素模型的状态的统计参数，形成各个状态的特征矢量；以及

通过用编号表示各个状态的特征矢量，来生成所述码本中的各个码本矢量。

2. 如权利要求 1 所述的方法，其中，所述规整化步骤包括：

从所述特征参数矢量序列中提取代表性特征参数矢量，来表征所述特征轨迹。

3. 如权利要求 2 所述的方法，其中，所述提取代表性特征参数矢量的步骤包括步骤：

在所述码本中搜索与所述特征参数矢量最接近的码本矢量；

用搜索的码本矢量的编号表示所述特征参数矢量；以及

对连续用相同的码本矢量表示的特征参数矢量进行合并，以表示所述特征轨迹。

4. 如权利要求 1 到 3 之一所述的方法，其中，所述匹配步骤包括；

利用各个关键词模板轨迹，针对每一轨迹帧，依次计算表示所述关键词模板轨迹的各个模板特征矢量与表示所述特征轨迹的各个代表性特征参数矢量之间的距离，所述轨迹帧是所述特征轨迹的时域表示；

确定所述距离中的最小值；以及

将与所述最小值所对应的关键词模板轨迹在矢量空间中的位置确定为关键词的位置。

5. 如权利要求 4 所述的方法，其中，所述特征参数矢量包括下面参数中的至少一个：

倒谱参数、倒谱参数的能量值、倒谱参数的一阶导数、倒谱参数的二阶导数、能量值的一阶导数以及能量值的二阶导数。

6. 如权利要求 4 所述的方法，其中，所述距离是用码本矢量距离表示的，所述码本矢量距离是所述码本中任意两个码本矢量之间的距离，所述关键词模板轨迹是基于所述码本而生成的。

7. 如权利要求 6 所述的方法，其中，所述码本矢量距离是以矩阵的形式预先存储的。

8. 如权利要求 6 所述的方法，其中，通过下面的步骤获得关键词模板轨迹：

通过音节和音素的隐马尔可夫模型之间的对应关系获得关键词的各个音节的音素名；

通过音素的隐马尔可夫模型和该音素的隐马尔可夫模型的状态之间的对应关系获得关键词的各个状态，形成状态矢量；

用所述码本中与状态矢量对应的编号表示音素的隐马尔可夫模型的码本矢量序列编号；

对连续用相同码本矢量编号表示的音素的隐马尔可夫模型的码本矢量序列编号进行合并；以及

顺序连接合并后的各个码本矢量序列编号，来得到关键词模板轨迹。

9. 如权利要求 6 所述的方法，其中，通过下面的步骤获得关键词模板轨迹 ：

对输入的关键词的音频帧进行切分，提取各个音频帧的特征参数矢量，以形成该关键词的特征参数矢量序列 ；以及

通过用所述码本对特征参数矢量序列进行规整化处理，来产生表示所述关键词模板轨迹的模板矢量序列。

10. 如权利要求 1 所述的方法，其中，所述音素模型是通过多个应用领域的语料训练而生成的。

11. 一种定位待识别语音中的关键词的设备，包括 ：

特征提取装置，用于提取构成所述待识别语音的各个帧的特征参数矢量，以形成用于描述待识别语音的特征参数矢量序列 ；

规整化装置，用于利用包含多个码本矢量的码本对特征参数矢量序列进行规整化处理，以得到待识别语音在矢量空间中的特征轨迹 ；以及

匹配装置，用于匹配预先存储的关键词模板轨迹和所述特征轨迹，以确定关键词的位置，

其中，所述码本是通过下面的方式获得的 ：

从基于隐马尔可夫模型的声学模型中提取描述各个音素模型的状态的统计参数，形成各个状态的特征矢量 ；以及

通过用编号表示各个状态的特征矢量，来生成所述码本中的各个码本矢量。

12. 如权利要求 11 所述的设备，其中，所述规整化装置从所述特征参数矢量序列中提取代表性特征参数矢量，来表征所述特征轨迹。

13. 如权利要求 12 所述的设备，其中，所述规整化装置在所述码本中搜索与所述特征参数矢量最接近的码本矢量，用搜索到的码本矢量的编号表示所述特征参数矢量，以及对连续用相同的码本矢量表示的特征参数矢量进行合并，以表示所述特征轨迹。

14. 如权利要求 11 到 13 之一所述的设备，其中，所述匹配装置利用各个关键词模板轨迹，针对每一轨迹帧，依次计算表示所述关键词模板轨迹的各个模板特征矢量与表示所述特征轨迹的各个代表性特征参数矢量之间的距离，以及确定所述距离中的最小值，并将最小值所对应的关键词模板轨迹在矢量空间中的位置确定为关键词的位置，所述轨迹帧是所述特征轨迹的时域表示。

15. 一种语音识别系统，包括 ：

如权利要求 11 所述的设备 ；以及

语音识别装置，用于基于所述设备所确定的关键词的位置识别关键词的内容。

16. 如权利要求 15 所述的语音识别系统，其中，所述语音识别装置是基于隐马尔可夫模型而进行识别的。

定位语音中的关键词的方法和设备以及语音识别系统

技术领域

[0001] 本发明涉及一种用于定位语音中的关键词的方法和设备，以及语音识别系统，具体地讲，涉及一种基于语音特征空间轨迹匹配来对语音中的关键词进行定位的方法和设备，以及利用该设备的语音识别系统，能够提高定位和识别的速度，并同时保证识别的精度。

背景技术

[0002] 近年来，越来越多的信息终端装置，例如 PC（个人计算机）、PDA（个人数字助理）、移动电话和遥控器等具备了语音输入功能，从而允许使用者通过发出语音来输入命令或者输入关键词。也就是说，需要这样的信息终端装置能够从用户输入的未知语音串中识别出用户想要输入的关键词。因此，如何准确和快速地确定关键词的位置是语音识别技术中一个重要的问题。

[0003] 文献 1（‘变帧速率技术在语音识别中的应用’，孙放，胡光锐，虞晓，上海交通大学学报，第 32 卷，第 8 期，1998 年 8 月）提出了将变帧速率技术应用于语音识别，用来丢弃那些特征非常相似的音频帧，从而达到快速识别输入语音的目的。在文献 1 中通过使用数学求导方法并定义合适的阈值，来进行语音特征的归并，进而获得具有音韵特征结构的语音特征矢量序列。但是上述阈值的设定非常困难，并且会直接影响到识别的精度。此外，文献 1 所提出的方法使用非线性匹配，因此需要在关键词识别过程中进行大量的计算。

[0004] 文献 2（‘KEYWORD SPOTTING METHOD BASED ON SPEECHFEATURE SPACE TRACE MATCHING’，Ya-dong Wu，Bao-long Liu，Proceedings of the Second Conference on Machine Learning andCybernetics，2003 年 11 月 2-5 日）提出通过计算特征矢量空间中矢量之间的距离并定义合适的阈值来进行语音特征归并，进而获得具有音韵特征结构的语音特征矢量序列。但是，由于这种归并是针对特定说话人语音的，因此表征同一音韵特征的代表特征点本身含有很多说话人的特征信息，变动较大。这样，在后续匹配过程中，不得不对语音轨迹进行重新刻度才能进行识别，由此增加了识别的复杂度。并且，文献 2 的技术没有很好地解决重新刻度的问题，因此识别的精度也很难保证。此外，计算矢量间距离所需的计算量非常大，为进行特征归并而设定合适的阈值也非常困难，且这个阈值的设定直接影响到具有音韵结构特征的语音轨迹估计是否准确。这些原因导致后续的基于此语音特征空间轨迹的匹配准确度不高。

[0005] 另外，在关键词模板建立方面，文献 2 的技术也是采用这种归并方法进行关键词语音特征空间轨迹估计，其中关键词内容是针对特定识别任务域设计的。具体来说，该关键词语音轨迹不是通过多种应用领域的语料训练而产生的，很难直接应用到非特定人领域。当任务域改变时，必须重新制作关键词语音模板。因此，在文献 2 的技术中，关键词语音轨迹模板不具有通用性，在实际应用中有一定困难。

[0006] 由于上述的问题，文献 1 和文献 2 所提出的方法无法实际应用到信息终端装置中。因此，需要一种能够快速定位输入的语音中的关键词并且能够减小计算量的技术。

发明内容

[0007] 鉴于现有技术的上述问题，完成了本发明。本发明的目的是提供一种基于语音特征空间轨迹匹配的、定位输入语音中的关键词的方法和设备，以及采用该设备的语音识别系统，能够提高定位和识别的速度，并同时保证识别的精度。

[0008] 在本发明的一个方面，提出了一种定位待识别语音中的关键词的方法，包括步骤：提取构成所述待识别语音的各个帧的特征参数矢量，以形成用于描述待识别语音的特征参数矢量序列；利用包含多个码本矢量的码本对特征参数矢量序列进行规整化处理，以得到待识别语音在矢量空间中的特征轨迹；以及匹配预先存储的关键词模板轨迹和所述特征轨迹，以确定关键词的位置。

[0009] 此外，根据本发明的实施例，所述规整化步骤包括：从所述特征参数矢量序列中提取代表性特征参数矢量，来表征所述特征轨迹。

[0010] 此外，根据本发明的实施例，所述提取代表性特征参数矢量的步骤包括步骤：在所述码本中搜索与所述特征参数矢量最接近的码本矢量；用搜索的码本矢量的编号表示所述特征参数矢量；对连续用相同的码本矢量表示的特征参数矢量进行合并，以表示所述特征轨迹。

[0011] 此外，根据本发明的实施例，所述匹配步骤包括：利用各个关键词模板轨迹，针对每一轨迹帧，依次计算表示所述关键词模板轨迹的各个模极特征矢量与表示所述特征轨迹的各个代表性特征参数矢量之间的距离，所述轨迹帧是所述特征轨迹的时域表示；确定所述距离中的最小值；以及将与所述最小值所对应的关键词模板轨迹在矢量空间中的位置确定为关键词的位置。

[0012] 此外，根据本发明的实施例，所述特征参数矢量包括下面参数中的至少一个：倒谱参数、倒谱参数的能量值、倒谱参数的一阶导数、倒谱参数的二阶导数、能量值的一阶导数以及能量值的二阶导数。

[0013] 此外，根据本发明的实施例，所述距离是用码本矢量距离表示的，所述码本矢量距离是码本中任意两个码本矢量之间的距离。

[0014] 此外，根据本发明的实施例，所述码本矢量距离是以矩阵的形式预先存储的。

[0015] 此外，所述关键词模板轨迹是基于所述码本而生成的。

[0016] 此外，根据本发明的实施例，通过下面的步骤获得关键词模板轨迹：通过音节和音素模型之间的对应关系获得关键词的各个音节的音素名；通过音素模型和状态之间的对应关系获得关键词的各个状态，形成状态矢量；用所述码本中与状态矢量对应的编号表示音素模型的码本矢量序列编号；对连续用相同码本矢量编号表示的音素模型的码本矢量序列编号进行合并；以及顺序连接合并后的各个码本矢量序列编号，来得到关键词模板轨迹。

[0017] 此外，根据本发明的实施例，通过下面的步骤获得关键词模板轨迹：对输入的关键词的音频帧进行切分，提取各个音频帧的特征参数矢量，以形成该关键词的特征参数矢量序列；以及通过用所述码本对特征参数矢量序列进行规整化处理，来产生表示所述关键词模板轨迹的模板矢量序列。

[0018] 此外，根据本发明的实施例，所述码本是通过下面的步骤获得的：从基于隐马尔可夫模型的声学模型中提取描述各个音素模型的状态的统计参数，形成各个状态的特征矢

量；以及通过用编号表示各个状态的特征矢量，来生成所述码本中的各个码本矢量。

[0019] 此外，根据本发明的实施例，所述音素模型是通过多个应用领域的语料训练而生成的。

[0020] 在本发明的另一个方面，提出了一种定位待识别语音中的关键词的设备，包括：特征提取装置，用于提取构成所述待识别语音的各个帧的特征参数矢量，以形成用于描述待识别语音的特征参数矢量序列；规整化装置，用于利用包含多个码本矢量的码本对特征参数矢量序列进行规整化处理，以得到待识别语音在矢量空间中的特征轨迹；以及匹配装置，用于匹配预先存储的关键词模板轨迹和所述特征轨迹，以确定关键词的位置。

[0021] 此外，根据本发明的实施例，所述规整化装置从所述特征参数矢量序列中提取代表性特征参数矢量，来表征所述特征轨迹。

[0022] 此外，根据本发明的实施例，所述规整化装置在所述码本中搜索与所述特征参数矢量最接近的码本矢量，用搜索到的码本矢量的编号表示所述特征参数矢量，对连续用相同的码本矢量表示的特征参数矢量进行合并，以表示所述特征轨迹。

[0023] 此外，根据本发明的实施例，所述匹配装置利用各个关键词模板轨迹，针对每一轨迹帧，依次计算表示所述关键词模板轨迹的各个模板特征矢量与表示所述特征轨迹的各个代表性特征参数矢量之间的距离，以及确定所述距离中的最小值，并将最小值所对应的关键词模板轨迹在矢量空间中的位置确定为关键词的位置，所述轨迹帧是所述特征轨迹的时域表示。

[0024] 在本发明的又一方面，提出了一种语音识别系统，它包括如上所述的设备；以及语音识别装置，用于基于所述设备所确定的关键词的位置识别关键词的内容。

[0025] 此外，根据本发明的实施例，所述语音识别装置是基于隐马尔可夫模型而进行识别的。

[0026] 利用本发明的方法和设备，由于基于同一码本来描述关键词模板轨迹和待识别的语音的特征空间轨迹，在具有相同音韵特征结构的音频帧进行线性移动匹配时，不需要重新刻度，从而降低了计算量，且提高了定位和识别的精度。

[0027] 另外，由于通过预先存储的码本矢量距离来描述待识别语音的特征空间轨迹和关键词模板轨迹之间的距离，使得在匹配过程中计算待识别语音和模板语音之间的距离时，可以通过查找的方式来获得匹配结果，进一步减小了匹配所需的计算量，提高了定位和识别的精度。

[0028] 另外，由于形成码本所需的音素模型是通过多个应用领域的语料训练而生成的，使得本发明的方案具有通用性。也就是可以应用在不同的领域。

[0029] 另外，将本发明的关键词定位方法和设备与现有的基于 HMM（隐马尔可夫模型）的语音识别系统相结合，可以避免 HMM 识别方法中难以建立废料模型的缺点，从而进一步提高了识别精度。

附图说明

[0030] 通过下面结合附图对发明进行的详细描述，将使本发明的上述特征和优点更加明显，其中：

[0031] 图 1 示出了根据本发明实施例的语音识别系统的构成框图；

[0032] 图 2 是说明根据本发明实施例的状态特征码本生成过程和码本矢量距离矩阵生成过程的示意图；其中图 2(A) 示出了说明码本生成过程和码本矢量距离矩阵生成过程的流程图；图 2(B) 示出了说明状态特征码本的一个实例；

[0033] 图 3 是说明本发明实施例的语音特征空间轨迹规整化过程的示意图，其中图 3(A) 示出了如何获得输入语音的特征矢量的示意图；图 3(B) 是获得的特征矢量在矢量空间中的示意图；图 3(C) 示出了如何对获得的特征矢量进行规整化的示意图；图 3(D) 是规整化的特征矢量在矢量空间中的示意图；

[0034] 图 4 是用来说明根据本发明实施例的语音特征空间轨迹的生成过程的示意图；其中图 4(A) 示出了输入语音的特征矢量序列在矢量空间中的示意图；图 4(B) 示出了规整化的特征矢量序列在矢量空间中的示意图；图 4(C) 示出了用来说明语音特征空间轨迹生成过程的流程图；以及

[0035] 图 5 是说明根据本发明实施例的基于语音特征空间轨迹匹配来定位输入语音中的关键词的过程的原理示意图；其中图 5(A) 示出了包括非关键词语音和关键词语音的一段语音波形；图 5(B) 示出了关键词语音和非关键词语音的特征矢量在矢量空间中的轨迹；图 5(C) 是用来说明关键词模板的轨迹与输入语音的轨迹进行线性移动匹配的过程的示意图；

[0036] 图 6 是说明根据本发明实施例的关键词模板的生成过程的示意图；其中图 6(A) 示出了说明关键词模板生成过程的流程图；图 6(B) 示出了关键词模板生成过程的一个实例；以及

[0037] 图 7 是说明根据本发明实施例的线性移动匹配过程的示意图。

具体实施方式

[0038] 下面，参考附图详细说明本发明的优选实施方式。在附图中，相同的参考标记在不同的附图中表示相同的或相似的组件。为了清楚和简明，包含在这里的已知功能和结构的详细描述将被省略，以避免它们使本发明的主题不清楚。

[0039] 图 1 示出了根据本发明实施例的语音识别系统的构成框图。如图 1 所示，本发明的语音识别系统包括：诸如麦克风之类的输入单元 10，用于输入待识别的语音；与输入单元连接的关键词定位设备 100，用于确定待识别语音中的关键词的位置；语音识别引擎 80，与关键词定位设备 100 连接、用于基于关键词定位设备 100 所确定的关键词的位置对关键词进行识别；以及输出单元 90，用于输出语音识别引擎 80 的识别结果。

[0040] 如图 1 所示，根据本发明实施例的关键词定位设备 100 包括：特征提取单元 20，与输入单元 10 连接，用于提取切分的语音帧的特征参数；诸如磁存储器或者半导体存储器之类的码本存储单元 40，用于存储矢量量化码本；诸如磁存储器或者半导体存储器之类的距离矩阵存储单元 50，用于以矩阵的形式存储码本中的码本矢量之间的距离；规整化单元 30，根据码本存储单元 40 中存储的码本对特征提取单元 20 提取的特征参数所形成的各个特征参数矢量进行规整化，以得到待识别的语音在矢量空间中的特征轨迹；关键词模板轨迹库 60，用于存储用户感兴趣的关键词在矢量空间中的轨迹；以及匹配单元 70，根据距离矩阵存储单元 50 中存储的码本矢量之间的距离和关键词模板轨迹库 60 中存储的关键词模板轨迹，通过匹配关键词模板轨迹和待识别语音的特征轨迹，来确定关键词的位置。

[0041] 在本发明实施例的语音识别系统中，利用诸如麦克风之类的语音输入单元 10 输入待识别的语音或者模板语音。当然，也可以从存储设备中直接获得预先记录的语音数据或者直接调用语音文件来输入语音。

[0042] 特征提取单元 20 按照预定的参数配置，例如格式、采样频率、编码位数、声道类型、帧长、帧移以及特征参数类型等等，对输入的语音数据的各帧进行特征提取，以得到输入语音的特征参数矢量序列 $\{V_1, V_2, \cdots\cdots V_N\}$，其中每个矢量 V_i 都是预定维数 K 的特征矢量，$i = 1, \cdots\cdots, N$。在本实施例中，特征提取单元 20 将输入的语音切分成音频帧，然后针对各个音频帧提取相应的特征参数，形成特征参数矢量。所提取的特征参数包括：倒谱参数、倒谱参数的能量值、倒谱参数的一阶导数、倒谱参数的二阶导数、能量值的一阶导数和能量值的二阶导数。这里所述的倒谱参数，是例如 12 维的 FFT（快速傅立叶变换）倒谱参数。在这种情况下，特征参数矢量的维数 K 是 39，具体如下：

[0043] 倒谱参数：$C_1, C_2, \cdots\cdots, C_{12}$；

[0044] 能量值：E；

[0045] 倒谱参数的一阶导数：$dC_1, dC_2, \cdots\cdots, dC_{12}$；

[0046] 倒谱参数的二阶导数：$DC_1, DC_2, \cdots\cdots, DC_{12}$；

[0047] 能量值的一阶导数：dE；

[0048] 能量值的二阶导数：DE。

[0049] 在本实施例中，特征提取单元 20 所提取的特征参数矢量序列 $\{V_1, V_2, \cdots\cdots V_N\}$ 被输入到关键词定位设备 100 中的规整化单元 30 中，以估计该输入语音在矢量空间中的轨迹，并且对该轨迹进行规整化处理，输出该输入语音的特征轨迹，用于后续的匹配处理。规整化单元 30 利用码本存储单元 40 中预先存储的“码本”对特征参数矢量序列中的各个特征参数矢量进行矢量量化后，输出以特征矢量量化序列表示的、该输入语音在矢量空间的特征轨迹。这里，码本存储单元 40 中存储的是用 HMM 方法生成的特征参数的标准矢量，用于对输入的待量化矢量进行量化。

[0050] 另外，距离矩阵存储单元 50 中存储了码本存储单元 40 中所存储的多个标准矢量（即码本矢量）中的任意两个码本矢量之间的距离，该码本矢量距离将被用来描述，在矢量空间中，模板语音的特征参数矢量与待识别语音的特征矢量之间的相似程度。

[0051] 关键词模板轨迹库 60 中预先存储了以特征矢量量化序列表示的、用户感兴趣的关键词（即模板关键词）在矢量空间中的特征轨迹，该特征轨迹用于与输入的待识别语音的特征轨迹进行匹配。

[0052] 规整化单元 30 把输入语音的规整化的特征矢量序列提供给匹配单元 70。匹配单元 70 从关键词模板轨迹库 60 中依次取出各个关键词的模板轨迹，并沿着由规整化的特征参数矢量序列所表示的点在矢量空间中形成的轨迹（它表示待识别的语音在矢量空间中的特征轨迹），移动所取出的模板轨迹，逐个轨迹帧进行匹配操作。这里轨迹帧是规整化的语音帧，也就是与构成规整化的轨迹的各个特征矢量相对应的音频帧。在移动过程中，每移动一个轨迹帧，匹配单元 70 基于距离矩阵存储单元 50 中存储的码本矢量距离，通过求和运算来计算该关键词模板轨迹与待识别的语音在矢量空间中的特征轨迹之间的距离。在整个轨迹匹配结束之后，获得利用该关键词模板轨迹匹配得到的最小距离。然后匹配单元 70 针对存储的各个关键词执行上述的过程，得到了各个关键词模板轨迹与待识别的语音在矢量

空间中的特征轨迹之间的相应最小距离。

[0053] 接下来，匹配单元 70 通过比较确定这些针对各个关键词模板轨迹的最小距离中的最小值，并将与该最小值相对应的那个模板轨迹的关键词识别为候选关键词。应该指出，在不需要较高识别精度的情况下，匹配单元 70 也可以直接将该候选关键词识别为最终的关键词。

[0054] 然后，匹配单元 70 将该候选关键词在矢量空间轨迹上的位置映射回时域中的相应音频帧中，从而能够确定该关键词在待识别语音中的位置。

[0055] 这样，语音识别引擎 80 可以直接利用特征提取单元 20 提取的、已经定位的候选关键词位置处的待识别语音的特征参数，进行进一步识别，以得到最终的识别结果，即关键词的内容。在最终确定关键词内容时可以参考候选关键词结果。

[0056] 最后，输出单元 90 根据语音识别引擎 80 的识别结果，输出识别的关键词的内容，例如将识别的关键词显示在屏幕上。

[0057] 下面结合附图 2 ～ 7 详细说明上述各个单元中的具体操作过程。

[0058] 图 2 是用来说明根据本发明实施例的状态特征码本生成过程和码本矢量距离矩阵生成过程的示意图；其中图 2(A) 示出了用来说明码本生成过程和码本矢量距离矩阵生成过程的流程图；图 2(B) 是用来说明状态特征码本的示意图。

[0059] 码本是由矢量量化所用的标准矢量构成的集合。在本实施例中，码本的物理意义是用来描述 HMM 声学模型的状态特征。

[0060] 码本矢量矩阵是保存了码本中任意两个码本矢量之间距离的二维数组，该码本和码本矢量矩阵被预先存储在诸如 ROM（只读存储器）或者 HD（硬盘）之类的存储器中。可以将码本和码本矢量矩阵分别存储在一个单独的存储器中，例如码本存储单元 40 和距离矩阵存储单元 50，或者将它们存储在单个存储器的不同存储区域中。

[0061] 在本实施例中，码本是在 HMM 声学模型的基础上生成的，具体的产生过程如下所述：

[0062] 1)HMM 声学模型是用 HMM 模型定义文件 (hmmdefs) 来描述的，各个音素模型的 hmmdefs 是通过多种应用领域的语料训练而得到的，hmmdefs 的结构如下：

[0063] ～h″ iz2″ //声学模型名

[0064] <BEGINHMM>

[0065] <NUMSTATES>5 //状态数，5 个，但只有 2，3，4 三个有效状态

[0066] <STATE>2 //状态编号

[0067] <NUMMIXES>6 //混合高斯分布数

[0068] <MIXTURE>1 1.250000e-001 //高斯分布编号及权重

[0069] <MEAN>39 //高斯分布的均值参数，39 维

[0070] 7.702041e+0006.226375e+000……2.910257e-001-8.276044e-002

[0071] <VARIANCE>39 //高斯分布的协方差参数，39 维

[0072] 7.258195e+001 5.090110e+001……3.907018e-001 2.388687e-002

[0073] ……

[0074] <MIXTURE>6 1.250000e-001 //高斯分布编号及权重

[0075] <MEAN>39 //高斯分布的均值参数，39 维

[0076] 8.864381e-001 5.187749e-001……-2.090234e-001-2.064035e-001

[0077] <VARIANCE>39 //高斯分布的协方差参数,39维

[0078] 7.258195e+001 5.090110e+001……3.907018e-001 2.388687e-002

[0079] <STATE>3 //状态编号

[0080] <NUMMIXES>6 //混合高斯分布数,各高斯分布也用均值和协方差两个

[0081] //参数来描述

[0082] ……

[0083] <STATE>4 //状态编号

[0084] <NUMMIXES>6 //混合高斯分布数,各高斯分布也用均值和协方差两个

[0085] //参数来描述

[0086] ……

[0087] <TRANSP>5 //状态转移概率矩阵

[0088] 0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000

[0089] 0.000000e+000 6.800905e-001 3.199094e-001 0.000000e+000 0.000000e+000

[0090] 0.000000e+000 0.000000e+000 6.435547e-001 3.564453e-001 0.000000e+000

[0091] 0.000000e+000 0.000000e+000 0.000000e+000 5.890240e-001 4.109760e-001

[0092] 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000

[0093] <ENDHMM>。

[0094] 2)特征状态提取(S110),也就是,按照具体应用来提取形成码本所需的特征参数。HMM模型定义文件中存储了描述音素模型各状态的统计参数,包括高斯分布均值(39维)、协方差(39维)、权重和状态转移矩阵(描述音素模型中各状态间转移的概率,每个音素用5个状态来描述,故为5×5的矩阵)。在本实施例中,抽取了描述每个状态的6个高斯分布的均值参数部分(12维CEP),并根据各高斯分布的权重求取它们的算术平均,利用计算得到的均值平均值(12维CEP)作为表征该状态的码本特征参数。

[0095] 然后,对所有状态进行编号,每个状态都有惟一确定的ID编号,例如图2(B)所示的1,2,……,M-1,M,它代表该状态的特征矢量,并用于生成状态特征码本。所生成的状态特征码本被存储在码本存储单元40中。码本中所含码本矢量的个数(M)定义为码本的大小。

[0096] 此外,在生成上述HMM声学模型的过程中,还可以得到音节-音素模型对应表和音素模型名-状态名对应表。这里,状态名由构成码本矢量的各个状态的编号来表示,如上所述。

[0097] 3)计算码本中任意两个状态特征矢量间的距离,得到M*M的码本矢量的距离矩阵MATRIX(S120),然后将其存储在距离矩阵存储单元50中。

[0098] 图3是用来说明本发明实施例的语音特征空间轨迹的规整化过程的示意图,其中图3(A)示出了如何获得输入语音波形的特征矢量的示意图;图3(B)是获得的特征矢量在矢量空间中的示意图;图3(C)示出了如何对获得的特征矢量进行规整化的示意图;图3(D)是规整化的特征矢量在矢量空间中的示意图。

[0099] 根据文献2,对同一字(词)音的两个特征矢量的时间序列X(tx)和R(tr),按相同的轨迹长s沿其各自的轨迹所提取出的新的特征矢量的时间序列X'(s)和R'(s)具有

对时间轴伸缩的不变性。这就是基于特征空间轨迹对语音进行时间轴规整的基本原理。由于经时间规整化后的特征矢量序列可采用线性匹配方式，故可大幅度地减少识别时的计算量。

[0100] 在本实施例中，语音特征空间轨迹是基于 HMM 声学模型生成的。首先提取输入语音文件的基本特征参数，如图 3(A) 所示。

[0101] 假设表示符号序列 S_1，S_2 和 S_3 的连续音频信号（波形）经过分帧处理成为 7 个音频帧。针对这 7 个音频帧提取相应的特征参数，得到 7 个特征矢量 V_i($1 \leqslant i \leqslant 7$)，以构成特征矢量序列，其中 V_i 是一个指定维数 (K 维）的特征矢量。

[0102] 需要指出的是，本领域的普通技术人员应该理解，虽然在各附图中以三维矢量空间来表示上述的 K 维矢量空间，但是这仅仅是出于清楚演示本发明的目的，而不意味着上述的 K 维就是 3 维。

[0103] 如图 3(B) 所示，这些特征序列 V_i 可以视作在 K 维空间上分布的坐标点，下面将其称为特征点。如果把这些特征点按照时间顺序连接起来 ($V_1 \rightarrow V_2 \rightarrow V_3 \cdots\cdots \rightarrow V_6 \rightarrow V_7$)，就可以得到在 K 维空间上的一条轨迹。从图 3(B) 可以看出，特征点 V_1 与其他的特征点分开比较远，特征点 V_2 和 V_3 相距比较近，而特征点 $V_4 \sim V_7$ 大致散落在一个比较集中的范围内。

[0104] 对语音特征空间轨迹进行规整化的关键在于如何准确地估计得到语音的特征空间轨迹。因为在实际应用中，语音中的各特征矢量在时间上是离散的，而且它除了要受到音速变化的影响之外，还将受到其它多种变动因素的影响，从而导致即使是同一音韵特征空间区域，该区域内的各帧的谱特性也将发生某些变动，其反映在语音特征空间中即为一簇相邻的特征点，对发音长的音韵，其簇中的特征点较多 (V_4, V_5, V_6, V_7)；对发音短的音韵，其簇中特征点较少 (V_2, V_3)。如图 3(B) 所示，称同一音韵的特征点散布区域为准平稳区 (Semi-stabilityArea)，而称不同音韵的特征点散布区域为非平稳区 (Non-stabilityArea)。为此，可以提取该簇特征点（矢量）中具有代表性的特征点来表征该音韵的特征，并以这些代表性的特征矢量 (F_j, $j = 1,2,3$) 来估计语音轨迹。这里，特征矢量 F_1 表示第一簇特征点中的代表性特征点，特征矢量 F_2 表示第二簇特征点中的代表性特征点，而特征矢量 F_3 表示第三簇特征点中的代表性特征点，如图 3(D) 所示。

[0105] 另外，如图 3(C) 所示，音频帧 1 的特征参数由经过规整化的特征点（矢量）F_1 来表示，音频帧 2 和 3 的特征参数由同一个特征点（矢量）F_2 来表示，而音频帧 4 ～ 7 的特征参数由另一个特征点（矢量）F_3 来表示。

[0106] 另外，为了提高语音特征轨迹估计的准确性，应该考虑：(1) 对语音信号按较小的帧移 (frame shift) 周期分帧，以提高非平稳区域内特征点的密度。例如现有技术的帧移周期是 20ms，而本实施例采用 10ms 或者 8ms 的帧移周期；(2) 对散布在准平稳区域内的特征点进行一定的修剪，即保留其具有代表性的特征点，删除其余不必要的特征点。一种可选的修剪方法是依次计算特征点间的导数，将导数小于设定阀值的那些特征点作为同一个准平稳区的点，然后这些特征点的平均作为该准平稳区域的代表性特征点。另一种可选的修剪方法是计算各特征点间的矢量距离，将矢量距离小于设定阀值的那些特征点作为同一个准平稳区的点，然后将准平稳区域内的特征点的平均作为该准平稳区域的代表性特征点。又一种方法是对连续用相同码本矢量表示的特征矢量（点）帧进行压缩合并。将在下面描述这种方法。

[0107] 图 4 是用来说明根据本发明实施例的待识别语音特征空间轨迹（特征矢量序列）的生成过程的示意图；图 4(A) 示出了输入语音的特征矢量序列在矢量空间中的示意图；图 4(B) 示出了规整化的特征矢量序列在矢量空间中的示意图；图 4(C) 示出了待识别语音的特征空间轨迹生成过程的流程图；

[0108] 下面参照图 4 描述在对特征矢量进行压缩的情况下的特征矢量序列生成过程。考虑到前面提到的语音轨迹规整化过程中应该注意的两点，提出了基于 HMM 声学模型的、采用矢量量化来规整化输入语音的方法。

[0109] 如图 4(A) 所示，输入待识别的语音 (S210)。这里，假设输入的语音表示为：

[0110] $X_i(t) = (X_1(t), X_2(t), \cdots\cdots X_6(t), X_7(t))$ （i：音频帧号）

[0111] 然后，对输入的语音进行特征提取操作 (S220)，以得到相同数量的特征矢量：

[0112] $V_i(t) = (V_1(t), V_2(t), \cdots\cdots V_6(t), V_7(t))$ （i：音频帧号）

[0113] 对照之前生成的状态特征码本，规整化单元 30 搜索特征矢量在码本中最为匹配的码本矢量，并用该码本矢量的 ID 编号表示该特征矢量，并对连续用相同码失表示的特征矢量帧进行压缩合并 (S230)，规整化单元 30 输出得到的语音特征空间的 VQ 矢量 (S240)。图 4(B) 示出了特征矢量是 k = 3 个的情况：

[0114] $V_j(t) = (ID_1(t), ID_2(t), \cdots\cdots ID_{k-1}(t), ID_k(t))$

[0115] 其中，j = 1,2，……，k，ID_j 表示码本矢量编号，k 表示待识语音状态特征矢量的总数，通常情况下 k 小于音频帧的数目。

[0116] 图 5 是用来说明根据本发明实施例的基于语音特征空间轨迹匹配过程的示意图；其中图 5(A) 示出了包括非关键词语音和关键词语音的一段语音波形；图 5(B) 示出了关键词语音和非关键词语音在矢量空间中的轨迹；图 5(C) 是用来说明关键词模板的轨迹与输入的语音的轨迹进行线性移动匹配的过程的示意图。

[0117] 如图 5(A) 所示，通常情况下关键词的波形位于输入的待识别语音波形中的某个位置处。如图 5(B) 所示，输入的语音在矢量空间中的轨迹是一条连续的曲线，该曲线上大致位于中间部分的那一段是关键词语音在矢量空间中的轨迹。

[0118] 如上所述，在已经知道了输入语音的特征轨迹的情况下，通过将关键词模板轨迹沿着输入语音的特征轨迹移动，来对二者进行匹配。同时，每移动与特征轨迹对应的时域信号中的一个音频帧，即轨迹帧，就通过线性加和模板轨迹上的特征点和与其对应的待识别语音的特征点间的距离来计算两个轨迹之间的距离，该距离表示该模板在此位置与待识别语音的空间轨迹的相似度。在该关键词模板轨迹匹配结束之后，得到针对该关键词模板轨迹的最小距离。然后针对不同的关键词模板轨迹来匹配所得到的各个最小距离。最后，从这些最小距离中找到最小值，将与该最小值所对应的关键词识别为候选关键词，进而将该候选关键词与输入语音的轨迹之间距离最小的那个位置映射回时域，得到该候选关键词所在的音频帧的帧号。

[0119] 图 6 是用来说明根据本发明实施例的关键词模板的生成过程的示意图；其中图 6(A) 示出了用来说明关键词模板生成过程的流程图；图 6(B) 示出了关键词模板生成过程的一个实例。

[0120] 模板的输入可以分为语音输入和文本输入两种方式。然而，本发明不限于此，也可通过其它方式输入。下面以文本输入和语音输入为例具体描述如下：

[0121] 【文本输入】

[0122] 如图 6(A) 所示，输入输入关键词文本 (S310)，例如‘上海’。然后，进行音节切分和拼音转换操作，例如将‘上海’切分成‘上 / 海’，并且得到‘上’和‘海’的字符串表达式，即拼音‘shang4’和‘hai3’(S320)，如图 6(B) 所示。

[0123] 接下来，通过音节 - 音素模型名对应表可以得到 shang4 和 hai3 的音素表达方式，即音素模型名，分别为“sh a4 ng4”和“haa3”(S330)，如图 6(B) 所示。

[0124] 在得到关键词的音素名后，利用该音素名，在音素模型名 - 状态名对应表中搜索与该音素模型名相对应的状态名，得到组成音素的各模型的码本矢量编号 (S240)。如图 6(B) 中，关键词“上海”的码本矢量序列编号为 :3,6,9,9,8,1, ……。

[0125] 接下来，将这些编号按照它们在关键词中的原始顺序连接起来，并将连续用相同码本矢量表示的特征矢量帧进行压缩合并，就得到了关键词模板的轨迹。如图 6(B)，表示关键词“上海”的模板轨迹的特征矢量序列的编号包括 :3,6,9,8,1, ……，其中相同的两个编号‘9’被压缩为同一个编号。最后，将该关键词模板轨迹与该关键词相对应地存储在关键词模板轨迹库 60 中。

[0126] 【语音输入】

[0127] 对于语音形式输入的关键词，即语音波形，首先将其按音频帧进行切分，提取各个音频帧的特征参数矢量，以获得描述该语音波形的特征参数矢量序列。在矢量空间中，利用上述矢量量化码本的各个特征参数矢量进行规整化，输出由各个特征点 (矢量) 表示的特征矢量序列。同样，该特征矢量序列中的各个特征矢量的元素是状态编号。

[0128] 图 7 是用来说明根据本发明实施例的轨迹移动匹配过程的示意图。

[0129] 如图 7 所示，规整化单元 30 根据基于 HMM 声学模型矢量量化的规整化算法，对待识语音进行规整化并得到各自的特征矢量序列 $T' = \{T'_m\}$ ($m = 0,1, \cdots, L$)，其中 L 为待识语音的轨迹总长度 (S410)。

[0130] 如上所述，模板语音 w($w = 1,2, \cdots, W$) 事先被进行了规整化。并且，将得到的规整化的特征矢量序列 $X'_w = \{X'_{n,w}\}$ ($n = 0,1, \cdots, L_w$) 存储在关键词模板轨迹库 60 中，其中 W 为模板总个数，L_w 为规整化后模板 w 的轨迹的总长度。

[0131] 然后，将模板语音轨迹 X'_w 从待识语音特征矢量序列 $\{T'_m\}$ 的第 0 轨迹帧 ($m = 0$) 开始，逐帧和待识语音轨迹段 $S_{m,m+Lw} = \{T'_m, T'_{m+1}, \cdots, T'_{m+Lw}\} \in \{T'_m\}$ ($m = 0,1, \cdots, L-L_w$)) 做线性移动匹配，利用码失距离矩阵 MATRIX 中存储的码本矢量距离，通过加和，记录每移动一轨迹帧时的匹配距离 :$D_{m,w} = \sum_n MATRIX(X'_{n,w}, T'_{m+n})$ ($m = 0,1, \cdots, L-L_w, n = 0, 1, \cdots, L_w$)，直至待识别语音轨迹的匹配终点 ($m = L-L_w$)。然后，记录此模板语音 w 的轨迹与待识语音的轨迹之间的最小匹配距离 $D_{m^*,w} = min(D_{m,w})$ ($0 \leqslant m^* \leqslant (L-L_w)$) (S420)。

[0132] 如果 $w < W$，则对其余的关键词模板重复 S420 步骤，否则，从各模板的最小匹配距离 $D_{m^*,w}$ ($w = 1,2, \cdots, W$) 中取出最小值所对应的关键词 $w^* = \arg\min_{1 \leq w \leq W}(D_{m^*,w})$，即检测出的待识别语音中含有的候选关键词 w^*(S430)，并将 w^* 对应特征点位置 m^* 转换成时域中的原始音频帧编号，结束检测 (S440)。

[0133] 此外，在上述的实施例中，以码本矢量距离的和来表示匹配距离，但是，这不应该被看作是对本发明范围的限定，因为本领域的普通技术人员也可以采用诸如码本矢量距离

的平方和、方根和或者范数和来表示。

[0134] 如上所述，在匹配单元 70 确定了关键词的位置之后，语音识别引擎 80 利用特征提取单元 20 提取的、已经定位的候选关键词位置处的待识别语音的特征参数进行进一步识别，以获得最终的识别结果，即识别内容。在最终确定关键词内容时可以参考候选关键词结果。

[0135] 另外，输出单元 90 根据语音识别引擎 80 的识别结果，输出识别的关键词的内容，例如将识别的关键词显示在屏幕上。

[0136] 上面的描述仅用于实现本发明的实施方式，本领域的技术人员应该理解，在不脱离本发明的范围的任何修改或局部替换，均应该属于本发明的权利要求来限定的范围，因此，本发明的保护范围应该以权利要求书的保护范围为准。

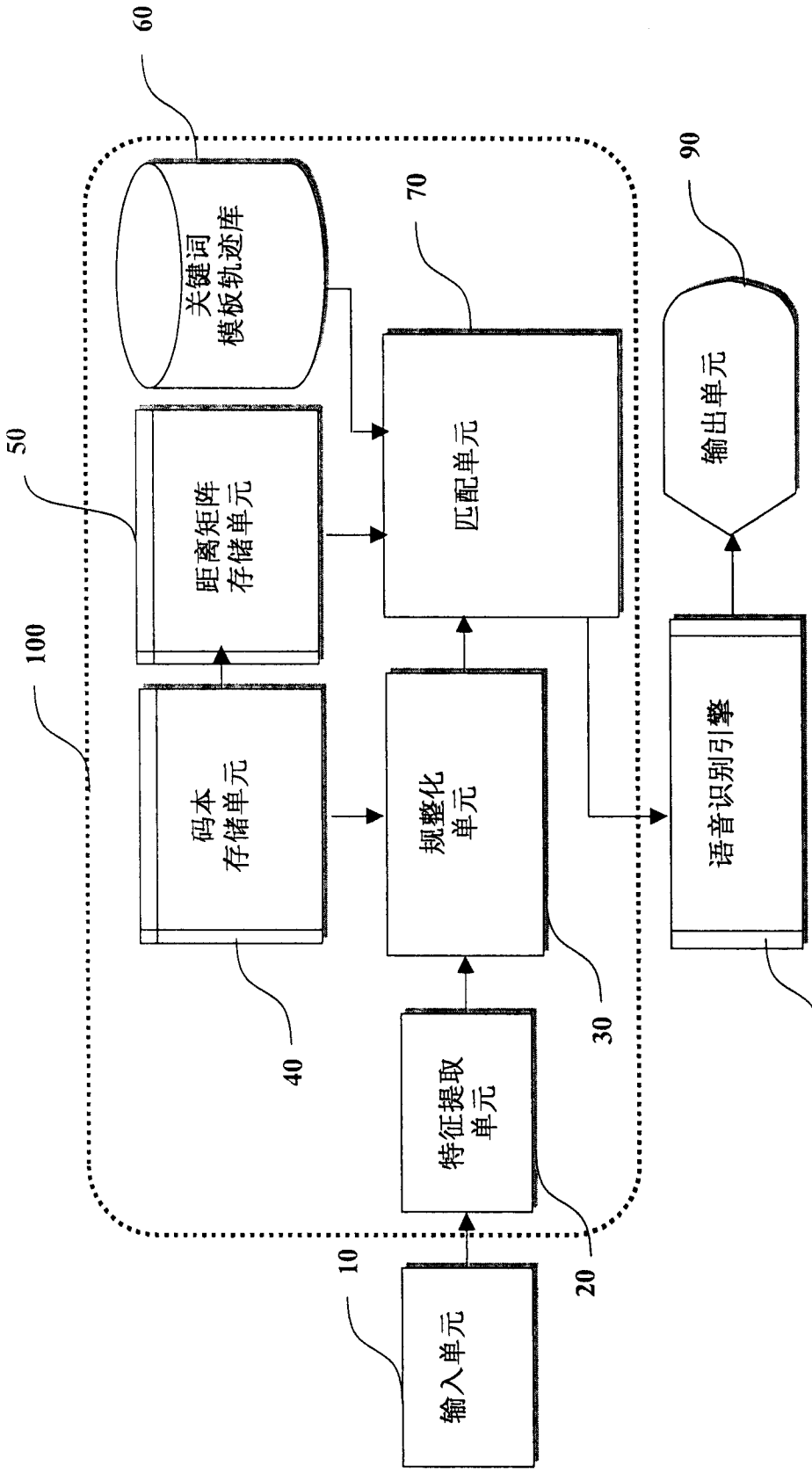
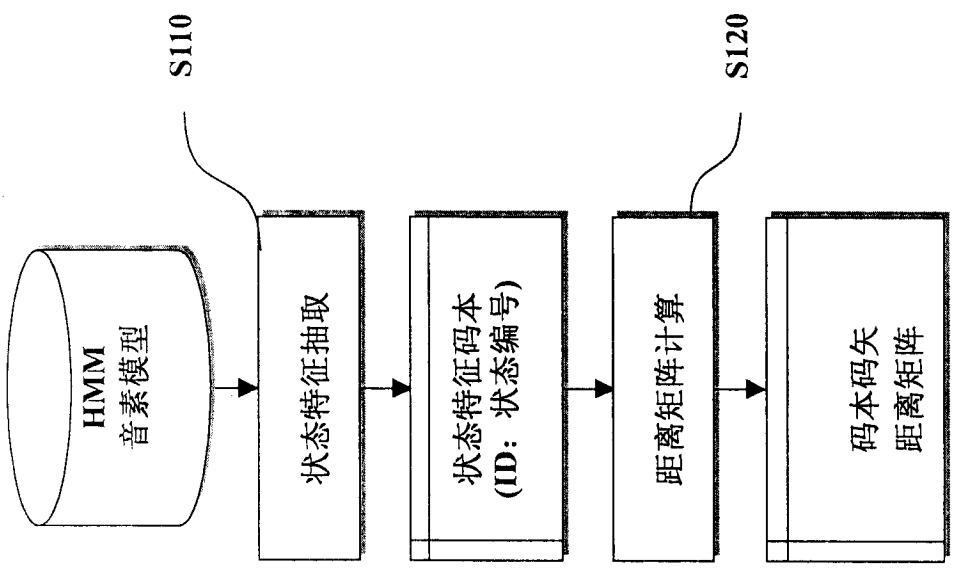
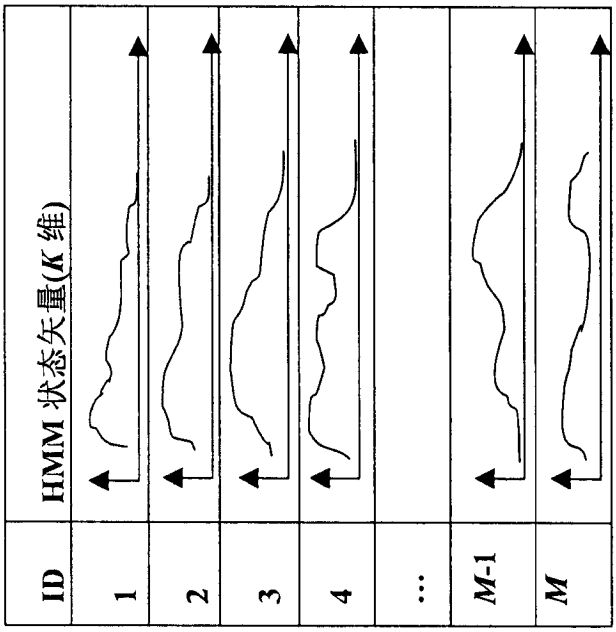

60
关键词
模板轨迹库
70
90
输出单元
50
距离矩阵
存储单元
匹配单元
100
码本
存储单元
规整化
单元
语音识别引擎
40
30
80
特征提取
单元
10
20
输入单元

图 1


HMM
音素模型
S110
状态特征抽取
状态特征码本
(ID：状态编号)
S120
距离矩阵计算
码本码矢
距离矩阵

(A)


ID
HMM 状态矢量(K 维)
1
2
3
4
...
M-1
M

(B)

图 2

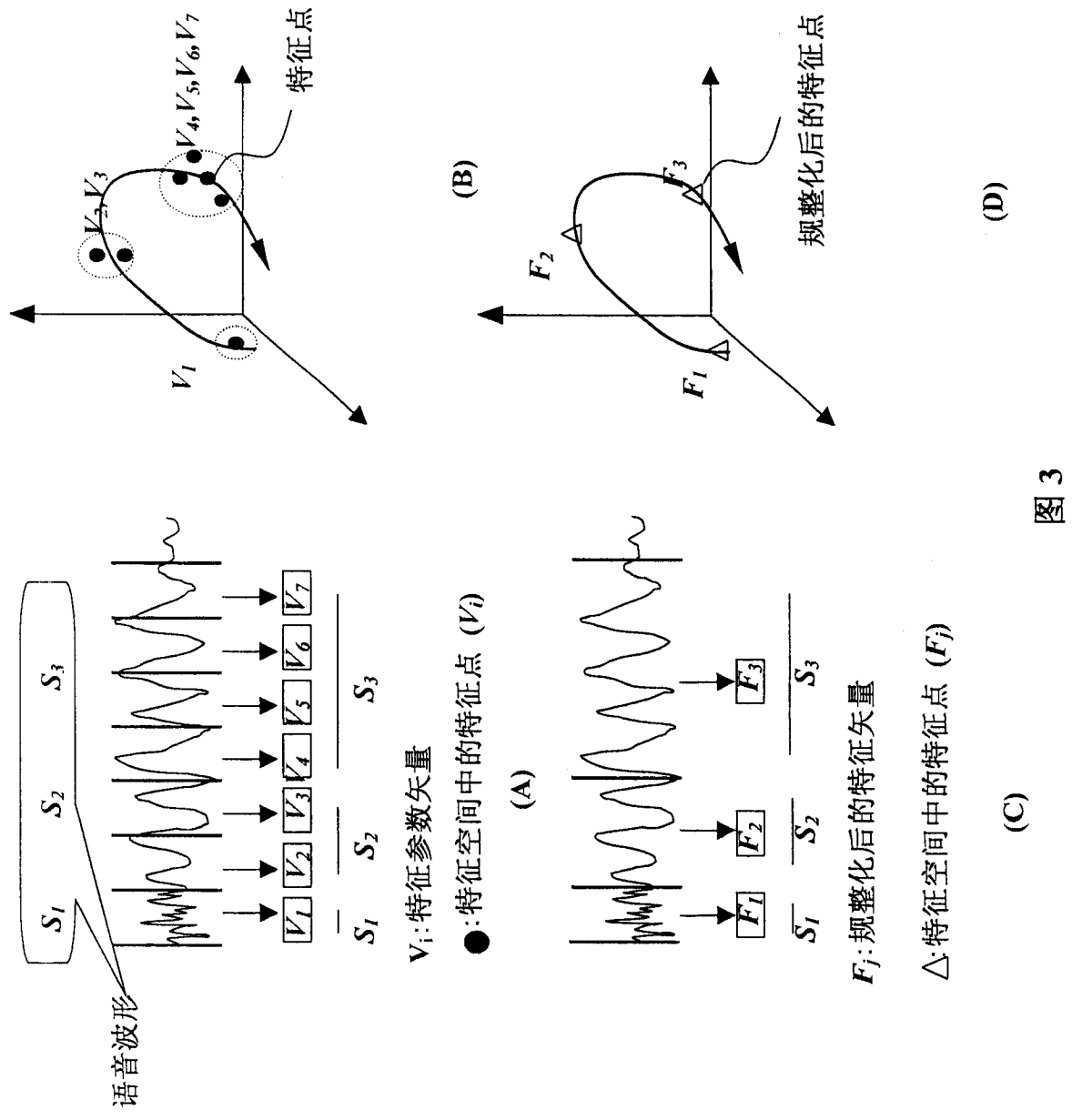
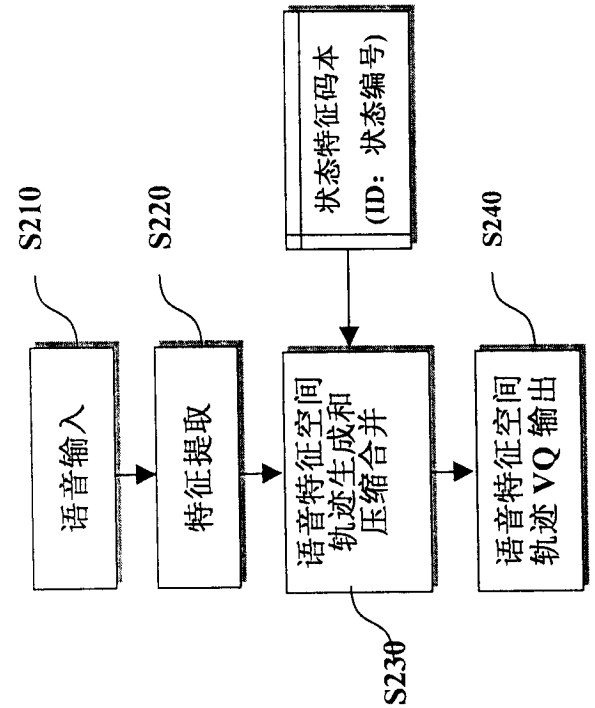
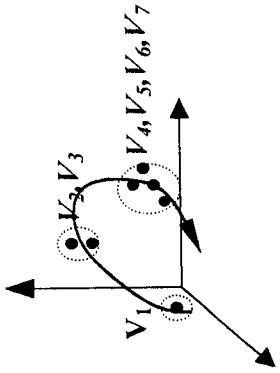

语音波形
S_1 S_2 S_3
V_1 V_2 V_3 V_4 V_5 V_6 V_7
S_1 S_2 S_3
V_i:特征参数矢量
●:特征空间中的特征点 (V_i)
(A)
V_1
V_2, V_3
V_4, V_5, V_6, V_7
特征点
(B)
F_1 F_2 F_3
S_1 S_2 S_3
F_j:规整化后的特征矢量
△:特征空间中的特征点 (F_j)
(C)
规整化后的特征点
(D)

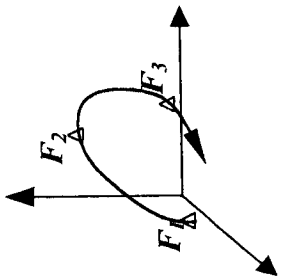
图 3


S210
S220
S230
S240
语音输入
特征提取
语音特征空间轨迹生成和压缩合并
状态特征码本 (ID：状态编号)
语音特征空间轨迹 VQ 输出

(C)


V1
V2,V3
V4,V5,V6,V7

(A)


F1
F2
F3

(B)

图 4

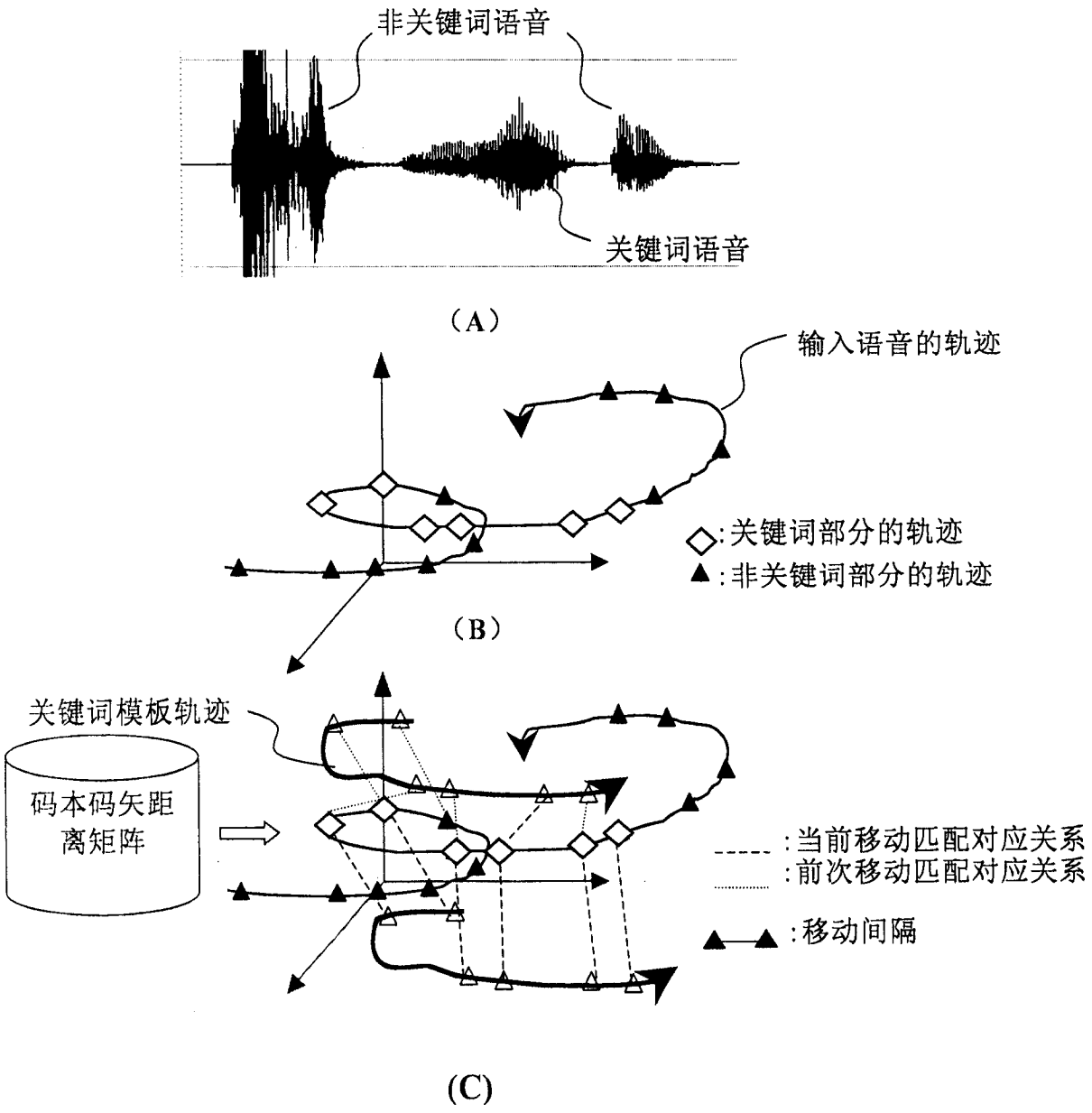

非关键词语音
关键词语音
(A)
输入语音的轨迹
:关键词部分的轨迹
:非关键词部分的轨迹
(B)
关键词模板轨迹
码本码矢距离矩阵
:当前移动匹配对应关系
:前次移动匹配对应关系
:移动间隔
(C)

图 5

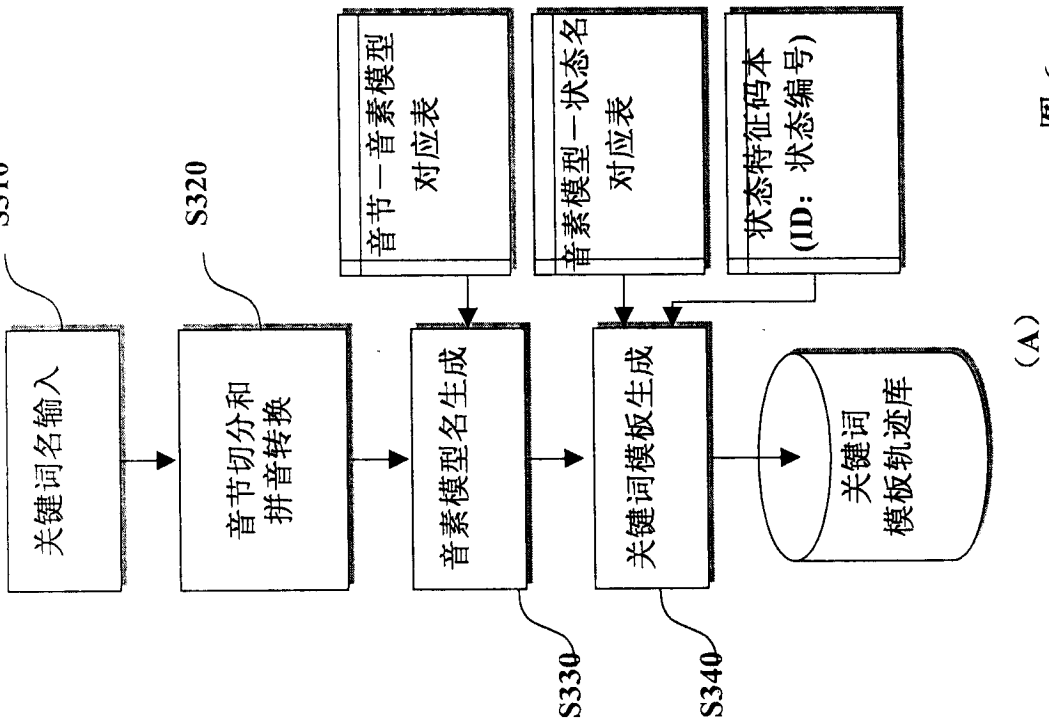

S310
S320
关键词名输入
音节切分和拼音转换
音节—音素模型对应表
音素模型—状态名对应表
状态特征码本(ID：状态编号)
音素模型名生成
关键词模板生成
关键词模板轨迹库
S330
S340
（A）
上海
上 / 海
Shang4 / Hai3
拼音 音素模型名
Shang4 : sh a4 ng4
Hai3 : h aa3 i3
模型名 状态名 ID
sh 3，6，9
a4 9，8，1
...
（B）

图 6

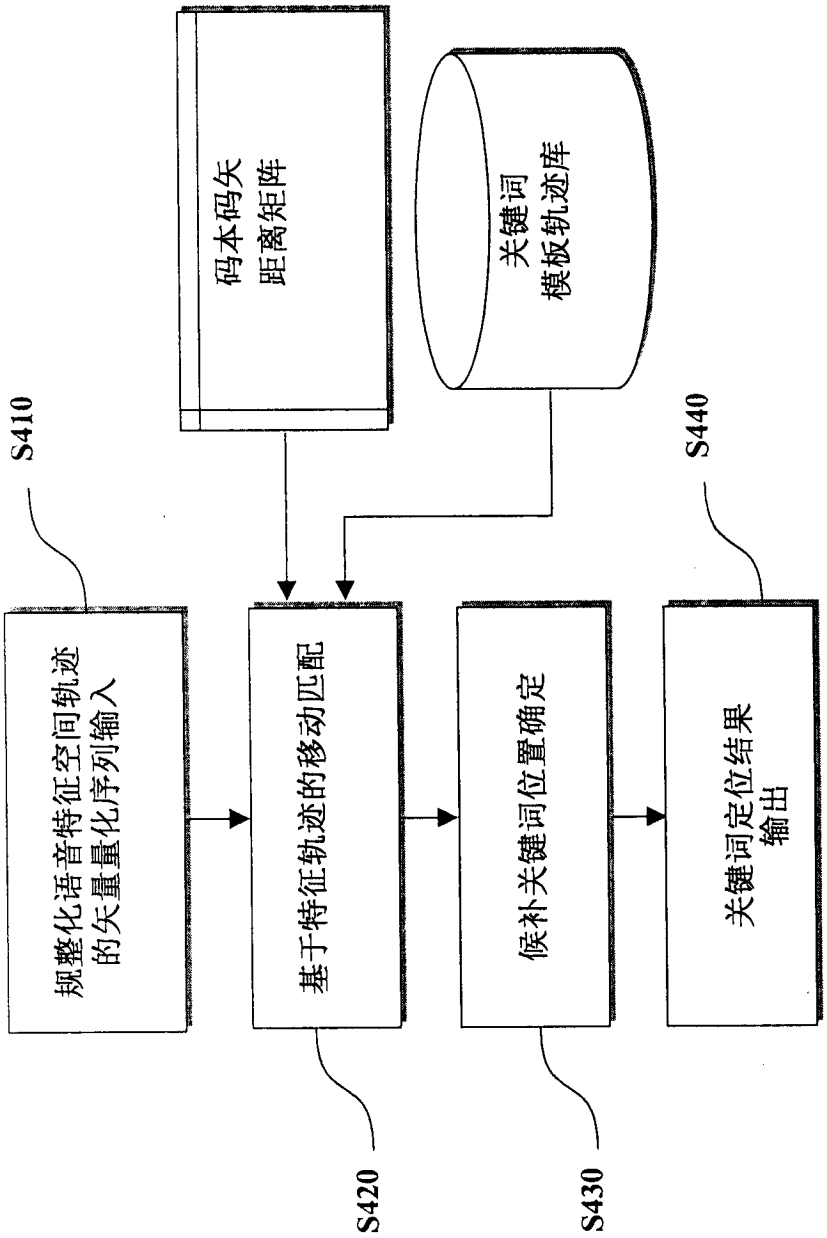

码本码矢
距离矩阵
关键词
模板轨迹库
S410
S440
规整化语音特征空间轨迹
的矢量量化序列输入
基于特征轨迹的移动匹配
候补关键词位置确定
关键词定位结果
输出
S420
S430

图 7