

[19] 中华人民共和国国家知识产权局

[51] Int. Cl.

H04W 88/02 (2009.01)

H04M 1/27 (2006.01)



[12] 发明专利申请公布说明书

[21] 申请号 200710182354.2

[43] 公开日 2009年4月22日

[11] 公开号 CN 101415259A

[22] 申请日 2007.10.18

[21] 申请号 200710182354.2

[71] 申请人 三星电子株式会社

地址 韩国京畿道水原市灵通区梅滩洞416

共同申请人 北京三星通信技术研究有限公司

[72] 发明人 黄盈椿 金南勋 赵正美 金志渊

[74] 专利代理机构 北京铭硕知识产权代理有限公司

代理人 郭鸿禧 韩素云

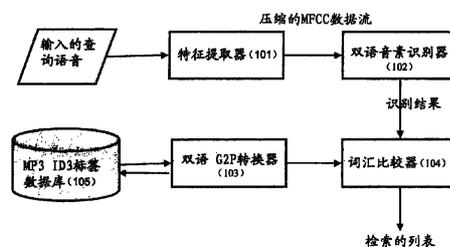
权利要求书4页 说明书19页 附图9页

[54] 发明名称

嵌入式设备上基于双语语音查询的信息检索系统及方法

[57] 摘要

提供了一种基于双语语音查询的信息检索的系统和方法。所述系统包括：特征提取器，将语音信号从 PCM 波形转换到 MFCC 特征参数，并在噪声抑制和帧压缩处理之后输出压缩的 MFCC 数据流；双语音素识别器，接收压缩的 MFCC 数据流，并通过将英语或者汉语语音自动转换到文本音素串来执行语音识别；双语文字音素转换器，将用于 MP3 ID3 标签数据库中的 MP3 文档的语音检索的可用内容的文字级的串转换成作为参考音素串的音素级的串；词汇比较器，将从双语音素识别器产生的识别的音素串和从双语文字音素转换器产生的参考音素串相比较，并输出最相关的前 N 个参考音素串。



1、一种基于双语语音查询的信息检索系统，所述系统包括：

特征提取器，将语音信号从PCM波形转换到MFCC特征参数，并在噪声抑制和帧压缩处理之后输出压缩的MFCC数据流；

双语音素识别器，接收压缩的MFCC数据流，并通过将英语或者汉语语音自动转换到文本音素串来执行语音识别；

双语文字音素转换器，将用于MP3 ID3标签数据库中的MP3文档的语音检索的可用内容的文字级的串转换成作为参考音素串的音素级的串；

词汇比较器，将从双语音素识别器产生的识别的音素串和从双语文字音素转换器产生的参考音素串相比较，并输出最相关的前N个参考音素串。

2、如权利要求1所述的系统，其中，双语音素识别器通过采用作为识别单元的汉语的声母/韵母以及英语的基础音素为总音素集来执行语音识别。

3、如权利要求2所述的系统，其中，双语音素识别器通过根据不同音素的平均持续时间来动态调整声学模型中Markov状态的数量来执行语音识别。

4、如权利要求3所述的系统，其中，双语音素识别器通过在汉语韵母具有2个子音素时增加2个状态并且在汉语韵母具有3个子音素时增加3个状态来调整Markov状态的数量。

5、如权利要求1所述的系统，其中，双语音素识别器通过采用一种新型语法网络来执行语音识别，所述新型网络通过这种方式被实现：添加特殊模型而在双语音素识别器中将两种语言分开以更好地区分两种完全不同的语言。

6、如权利要求5所述的系统，其中，所述新型网络这样被实现：通过将语法划分成两部分，其中，一部分是用于识别汉语的音节子环，另一部分是用于识别英语的基础音素子环；通过若干具有单一Markov状态的静音模型连接这两个子环。

7、如权利要求1所述的系统，其中，双语文字音素转换器包括：

语种识别器，通过操作系统所使用的不同的字符编码集来检测语种边界；

汉语文字音素转换器，通过采用最大匹配算法来执行汉语文字音素转换，并将句子中的所有汉字标记为汉语拼音串；

英语文字音素转换器，根据单词音节结构的结合来执行英语文字音素转

换;

特殊字符文字音素转换器, 将数字、英语缩写或其他汉语文字音素转换器和英语文字音素转换器丢弃的其他字符转换为音素级的串;

结合器, 将英语文字音素转换器、汉语文字音素转换器和特殊字符文字音素转换器的输出结合成它们的原始音素串。

8、如权利要求 7 所述的系统, 其中, 双语文字音素转换器还包括:

无意义字符确定器, 确定可用内容是否包含无效或无意义的字符;

文本过滤器, 滤除无效或无意义的字符;

格式规整器, 将可用内容规整为“标题_名字_专辑”的统一格式;

词间分割器, 将完整的音乐标题分离为很多条分词以执行分词查询。

9、如权利要求 7 所述的系统, 其中, 汉语文字音素转换器包括一个固定词典, 所述词典存储了最常用的汉字和汉语短语的发音。

10、如权利要求 7 所述的系统, 其中, 英语文字音素转换器存储了多于 20000 个英语音节的结构和其发音。

11、如权利要求 1 所述的系统, 其中, 词汇比较器包括:

混淆度矩阵模块, 创建混淆度矩阵 C, 其中, 混淆度矩阵 C 如下被定义:

$$C = \begin{pmatrix} P_{1,1} & \dots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \dots & P_{m,m} \end{pmatrix}$$

其中, 矩阵 C 中的 P_{ij} 表示在语音识别过程中第 j 音素被错误地识别为第 i 音素的概率;

相似度计算器, 通过使用修改的动态时间伸缩算法来计算识别的音素串和参考音素串之间的相似度, 其中, 相似度用一个数字格式的得分来表示;

分值确定器, 检测相似度得分是否大于预定阈值;

反转词典模块, 存储音素串和其原始文本串名称之间的链接以从缺乏理解力的音素串找到原始文字级的名称;

文字恢复器, 当得分大于阈值时将参考音素串恢复为文字级的串, 然后在 UI 显示装置中显示这些文字串。

12、如权利要求 11 所述的系统, 其中, 所述词汇恢复器还包括:

结果验证器, 防止当输入查询语音为非法或者 MP3 ID3 标签数据库不包括相关 MP3 文档时的错误输出。

13、一种基于双语语音查询的信息检索方法，所述方法包括：

将语音信号从 PCM 波形转换到 MFCC 特征参数，并在噪声抑制和帧压缩处理之后输出压缩的 MFCC 数据流；

通过将英语或者汉语语音自动转换到作为识别的音素串的文本音素串来对压缩的 MFCC 数据流执行语音识别；

将用于 MP3 ID3 标签数据库中的 MP3 文档的语音检索的可用内容的文字级的串转换成参考音素串；

将识别的音素串和参考音素串相比较，并输出前 N 个最相关参考音素串。

14、如权利要求 13 所述的方法，其中，通过将作为识别单元的汉语的声母/韵母以及英语的基础因素结合为总音素集来执行语音识别。

15、如权利要求 14 所述的方法，其中，通过根据音素的平均持续时间来动态调整声学模型中 Markov 状态的数量来执行语音识别。

16、如权利要求 15 所述的方法，其中，通过在汉语韵母具有 2 个子音素时增加 2 个状态而在汉语韵母具有 3 个子音素时增加 3 个状态执行调整 Markov 状态的状态数量的操作。

17、如权利要求 16 所述的方法，其中，通过采用一种新型语法网络来执行语音识别，所述新型语法网络通过这种方式被实现：添加特殊模型而将两种语言分开以更好地区分两种完全不同的语言。

18、如权利要求 17 所述的方法，其中，所述新型网络这样被实现：通过将语法划分成两部分，其中，一部分是用于识别汉语的音节子环，另一部分是用于识别英语的基础音素子环；通过若干具有单一 Markov 状态的静音模型连接这两个子环。

19、如权利要求 13 所述的方法，其中，将文字级的串转换为音素级的串的步骤包括：

通过 ASCII 字符集和汉语 GB2312 来检测语种边界；

当汉字被输入时，通过采用最大匹配算法来执行汉语文字音素转换，并将句子中的所有汉字标记为汉语拼音串；

当英语单词被输入时，根据单词音节结构的结合来执行英语文字音素转换；

当输入的文字是数字、英语缩写他不能进行汉语文字音素转换和英语文字音素转换的其他字符时，将数字、英语缩写或所述其他字符转换为音素级

的串;

将英语文字音素转换、汉语文字音素转换和特殊字符子文字音素转换的结果结合成它们的原始音素串。

20、如权利要求 19 所述的方法,其中,将文字级的串转换为音素级的串的步骤还包括:

确定可用内容是否包含无效或无意义的字符;

如果包含无效或无意义的字符,则滤除无效或无意义的字符;

将可用内容规整为“标题_名字_专辑”的统一格式;

将完整的音乐标题分离为很多条分词以执行分词查询。

21、如权利要求 19 所述的方法,其中,在执行汉语文字音素转换的步骤中,在固定词典中存储最常用的汉字和汉语短语的发音。

22、如权利要求 19 所述的方法,其中,在执行英语文字音素转换的步骤中,存储多于 20000 的英语音节结构和其发音。

23、如权利要求 13 所述的方法,其中,比较识别的音素串和参考音素串的步骤包括:

创建混淆度矩阵 C,其中,混淆度矩阵 C 如下被定义:

$$C = \begin{pmatrix} P_{1,1} & \cdots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \cdots & P_{m,m} \end{pmatrix}$$

其中,矩阵 C 中的 $P_{i,j}$ 表示在语音识别过程中第 j 音素被错误地识别为第 i 音素的概率;

通过使用修改的动态时间伸缩算法来计算识别的音素串和参考音素串之间的相似度,其中,相似度被表示为数字得分;

得分是否大于预定阈值;

如果得分大于预定阈值,借助音素串和它原始文字级的名称之间的链接将音素串恢复为文字级的串,然后输出这些文字串。

24、如权利要求 23 所述的方法,其中,比较识别的音素串和参考音素串的步骤还包括:

当输入查询语音为非法或者 MP3 ID3 标签数据库不包括相关 MP3 文档时防止错误输出。

嵌入式设备上基于双语语音查询的信息检索系统及方法

技术领域

根据本发明的系统和方法涉及一种用于在嵌入式设备上基于双语语音查询的信息检索系统和方法。

背景技术

现在, 移动电子产品(比如 PDA、蜂窝电话、MP3 播放器和 GPS 导航器)成为获取信息或多媒体内容的必要工具。在中国, 使用无线服务的用户有了很大增长。每天有超过 5 亿的移动电话用户在享受通信、彩铃下载和短消息服务。随着大存储容量的存储器的价格逐渐降低和高速 3G 无线网络的即将来临, 人们可从无线信道自由下载大量信息并将其存储在他们的移动装置中。在这种条件下, 如传统方式那样逐一搜索需要的数据是不实际的。对任何大存储装置, 快速而精确的信息检索(IR)是至关重要的。

移动装置通常具有小巧的外观, 这就不可避免地导致手动操作的不便。现在, 制造商渴望找到替代传统键盘和手写输入的替代品。如我们所知, 语音是人类交流的最自然的媒介。语音输入具有快速输入语音的能力, 并且与人机交互的其他媒介相比, 语音输入具有更友好的用户界面。因此, 在移动平台上加入语音识别(SR)应用(比如语音控制, 名称/数字拨号以及语音邮件应用)的兴趣在不断增长。在另一方面, 英语作为第二语言被全世界的人们广泛接受, 双语频繁出现在语音内容中。这种趋势在中国尤其显著。所以, 嵌入式信息检索的需要与支持多语种语音查询的结合成为新的研究课题。

内部音素识别器的性能在很大程度上影响整个语音查询信息检索系统的性能。在过去的几年中, 在探索建立多语种语音识别器的可能性上作了很多努力。直到现在, 还有多种方法来实现它。典型的一个方法是在多个单语种 SR 引擎前面采用外部语种识别器(LID)。所述语种识别器判断输入的话语的语种并将其发送到相应的语音识别器。显然整个语音识别器的性能在很大程度上取决于 LID 的准确性。理论上, 多语种 SR 的性能的上限分别等于每一单语种系统。通常, LID 需要至少 2 秒的语音波形来确保精度。然而, 对于我

们的应用来说,大多数的查询词汇集中在短词上。故通常的方法不适合我们的系统。第二种方法是将所有的语言有关的音素映射到一个总音素集中。迄今为止,语音学专家产生了一些总音素表示法,比如国际音标(IPA)和 SAMPA (Speech Assessment Methods Phonetic Alphabet, 具体可参见 <http://www.phon.ucl.ac.uk/home/sampa/home.htm>)和 Worldbet (具体可参见在 ICASSP 2007 的会议中由 Chien-lin Huang 和 Chung-Hsien Wu 发表的“Phone Set Generation Based on Acoustic and Contextual Analysis for Multilingual Speech Recognition”)。另外,可应用跨语言声学模型参数绑定(比如自底向上的聚类方法(bottom up clustering)和自顶向下的决策树分裂方法(top down decision tree splitting)以减少冗余并进行强健的参数估计(具体请参见由 Shengmin Yu, Sheng Hu, Shuwu Zhang 和 Bo Xu 在 Natural Language Processing and Knowledge Engineering, 2003 中发表的“Chinese-English Bilingual Speech Recognition”)。一般来说,由于不需要额外 LID 装置,所以总音素集更适合于移动装置中的多语种 SR 应用。可通过调节参数绑定的程度来容易地控制总音素集的大小。实际上,在大多数语言中研发识别器的结构和算法相似。这增加了建立总音素集的可行性。

结合语音识别和信息检索技术的语音查询 IR 系统继承了这二者的优点,从而用户可从存储在移动装置中的大数据库中快速检索他们请求的数据。最近,已经公开了一些关于这种基于语音查询的 IR 系统的发明。例如,Microsoft Research 公开了在 PDA 上实现汉语普通话的系统(具体请参见 IEEE trans on Speech and Audio processing 中 2002 年第 10 卷第 8 期中由 Eric Chang, Frank Seide, Helen M. Meng, Zhuoran Chen, Yu Shi 和 Yuk-Chi Li 发表的“A system for Spoken Query Information Retrieval on Mobile Devices”)。然而,它们中的大多数没有考虑全球化的背景和中国正在增长的双语用户。并且,它们也只面向小词库(small lexicon)检索任务(比如名称拨号、地点名称导航),而不支持中等或大词库检索。

发明内容

提供本发明的一方面在于解决上述和/或其他问题和缺点。

根据本发明,提供了基于双语语音查询检索系统允许用户在移动装置上执行双语语音查询。根据不同的功能,所述系统可由四部分组成:特征提取

器、文字音素转换器(G2P)、音素识别器和词汇比较器。为了解决与双语识别有关的问题,本发明基于两种语言的声学特性提出两种方法。一种方法是将汉语声母/韵母(I/F)和英语音素结合为总音素库;另一种方法是在双语音素识别器中改变不同音素的声学模型中的 Markov 状态的数量和使用新型语法网络结构来更好地区分两种完全不同的语言。

根据本发明的一方面,提供了一种基于双语语音查询的信息检索系统,所述系统包括:特征提取器,将语音信号从 PCM 波形转换到 MFCC 特征参数,并在噪声减小和帧压缩处理之后输出压缩的 MFCC 数据流;双语音素识别器,接收压缩的 MFCC 数据流,并通过将英语或者汉语语音自动转换到文本音素串来执行语音识别;双语文字音素转换器,将用于 MP3 ID3 标签数据库中的 MP3 文档的语音检索的可用内容的文字级的串转换成作为参考音素串的音素级的串;词汇比较器,将从双语音素识别器产生的识别的音素串和从双语文字音素转换器产生的参考音素串相比较,并输出前 N 个最相关的参考音素串。

根据本发明的一方面,双语音素识别器通过将作为识别单元的汉语的声母/韵母以及英语的基础音素结合为总音素集来执行语音识别。

根据本发明的一方面,双语音素识别器通过根据不同音素的平均持续时间来动态调整 Markov 状态的数量来执行语音识别。

根据本发明的一方面,双语音素识别器通过在汉语韵母具有 2 个子音素时增加 2 个状态而在汉语韵母具有 3 个子音素时增加 3 个状态来调整 Markov 状态的状态数量。

根据本发明的一方面,双语音素识别器通过采用一种新型语法网络来执行语音识别,所述新型网络通过这种方式被实现:添加特殊模型而在双语音素识别器中将两种语言分开以更好地区分两种完全不同的语言。

根据本发明的一方面,所述新型网络这样被实现:通过将语法划分成两部分,其中,一部分是用于识别汉语的音节子环,另一部分是用于识别英语的基础音素子环;通过添加若干具有单一 Markov 状态的静音模型连接这两个子环。

根据本发明的一方面,双语文字音素转换器包括:语种识别器,通过不同的字符集检测语种边界;汉语文字音素转换器,通过采用最大匹配算法来执行汉语文字音素转换,并将句子中的所有汉字标记为汉语拼音串;英语文

字音素转换器，根据单词音节结构的结合来执行英语文字音素转换；特殊字符文字音素转换器，将数字、英语缩写或其他汉语文字音素转换器和英语文字音素转换器丢弃的其他字符转换为音素级的串；结合器，将英语文字音素转换器、汉语文字音素转换器和特殊字符文字音素转换器的输出结合成它们的原始音素串。

根据本发明的一方面，双语文字音素转换器还包括：无意义字符确定器，确定可用内容是否包含无效或无意义的字符；文本过滤器，滤除无效或无意义的字符；格式规整器，将可用内容规整为“标题_名字_专辑”的统一格式；词间分割器，将完整的音乐标题分离为很多条分词以执行分词查询。

根据本发明的一方面，汉语文字音素转换器包括固定的词典，所述词典存储最常用的汉字和汉语短语的发音。

根据本发明的一方面，英语文字音素转换器存储多于 20000 的英语音节结构和其发音。

根据本发明的一方面，词汇比较器包括：混淆度矩阵模块，创建混淆度矩阵 C，其中，混淆度矩阵 C 如下被定义：

$$C = \begin{pmatrix} P_{1,1} & \dots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \dots & P_{m,m} \end{pmatrix}$$

其中，矩阵 C 中的 $P_{i,j}$ 表示在语音识别过程中第 j 音素被错误地识别为第 i 音素的概率；相似度计算器，通过使用修改的动态时间伸缩算法来计算识别的音素串和参考音素串之间的相似度，其中，相似度被表示为某一数字得分；分值确定器，检测得分是否大于预定阈值；反转词典模块，存储音素串和其原始文本串名称之间的链接以从缺乏理解力的音素串找到原始文字级的名称；文字恢复器，当得分大于阈值时将参考音素串恢复为文字级的串，然后在 UI 显示装置中显示这些文字串。

根据本发明的一方面，所述词汇恢复器还包括：结果验证器，防止当查询语音非法或者 MP3 ID3 标签数据库不包括相关 MP3 文档时的错误输出。

根据本发明的另一方面，提供了一种基于双语语音查询的信息检索方法，所述方法包括：将语音信号从 PCM 波形转换到 MFCC 特征参数，并在噪声抑制和帧压缩处理之后输出压缩的 MFCC 数据流；通过将英语或者汉语语音自动转换到作为识别的音素串的文本音素串来对压缩的 MFCC 数据流执行语

音识别;将用于 MP3 ID3 标签数据库中的 MP3 文档的语音检索的可用内容的文字级的串转换成作为参考音素串的音素级的串;将识别的音素串和参考音素串相比较,并输出前 N 个最相关的参考音素串。

根据本发明的另一方面,通过将作为识别单元的汉语的声母/韵母以及英语的基础因素结合为总音素集来执行语音识别。

根据本发明的另一方面,通过根据音素的平均持续时间来动态调整各个音素声学模型中 Markov 状态的数量来执行语音识别。

根据本发明的另一方面,通过在汉语韵母具有 2 个子音素时增加 2 个状态而在汉语韵母具有 3 个子音素时增加 3 个状态执行调整 Markov 状态的状态数量的操作。

根据本发明的另一方面,通过采用一种新型语法网络来执行语音识别,所述新型语法网络通过这种方式被实现:添加特殊模型而将两种语言分开以更好地区分两种完全不同的语言。

根据本发明的另一方面,所述新型网络这样被实现:通过将语法划分成两部分,其中,一部分是用于识别汉语的音节子环,另一部分是用于识别英语的基础音素子环;通过添加若干具有单一状态的静音模型连接这两个子环。

根据本发明的另一方面,将文字级的串转换为音素级的串的步骤包括:通过 ASCII 字符集和汉语 GB2312 来检测语种边界;当汉字被输入时,通过采用最大匹配算法来执行汉语文字音素转换,并将句子中的所有汉字标记为汉语拼音串;当英语单词被输入时,根据单词音节结构的结合来执行英语文字音素转换;当输入的文字是数字、英语缩写等等不能进行汉语文字音素转换和英语文字音素转换的其他字符时,将数字、英语缩写或所述其他字符转换为音素级的串;将英语文字音素转换、汉语文字音素转换和特殊字符子文字音素转换的结果结合成它们的原始音素串。

根据本发明的另一方面,将文字级的串转换为音素级的串的步骤还包括:确定可用内容是否包含无效或无意义的字符;如果包含无效或无意义的字符,则滤除无效或无意义的字符;将可用内容规整为“标题_名字_专辑”的统一格式;将完整的音乐标题分离为很多条分词以执行分词查询。

根据本发明的另一方面,在执行汉语文字音素转换的步骤中,在固定词典中存储最常用的汉字和汉语短语的发音。

根据本发明的另一方面,在执行英语文字音素转换的步骤中,存储多于

20000 的英语音节结构和其发音。

根据本发明的另一方面，比较识别的音素串和参考音素串的步骤包括：创建混淆度矩阵 C，其中，混淆度矩阵 C 如下被定义：

$$C = \begin{pmatrix} P_{1,1} & \cdots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \cdots & P_{m,m} \end{pmatrix}$$

其中，矩阵 C 中的 $P_{i,j}$ 表示在语音识别过程中第 j 音素被错误地识别为第 i 音素的概率；通过使用修改的动态时间伸缩算法来计算识别的音素串和参考音素串之间的相似度，其中，相似度被表示为某一得分；得分是否大于预定阈值；如果得分大于预定阈值，借助音素串和它原始文字级的名称之间的链接将音素串恢复为文字级的串，然后输出这些文字串。

根据本发明的另一方面，比较识别的音素串和参考音素串的步骤还包括：当输入查询语音为非法或者 MP3 ID3 标签数据库不包括相关 MP3 文档时防止错误输出。

附图说明

从下面结合附图对示例性实施例的描述中，本发明的上述和其他方面将会更清楚并更容易理解，其中：

图 1 是示出根据本发明示例性实施例的基于双语语音查询的信息检索系统的框图；

图 2 是示出图 1 所示的系统的双语音素识别器 102 的具体结构的框图；

图 3A 示出传统的语法网络；

图 3B 示出建议的语法网络；

图 4 是示出双语 G2P 转换器 103 的结构的框图；

图 5 是示出词汇比较器 104 的结构的框图；

图 6 是分词分割的示例；

图 7 示出 DTW 网络；

图 8 是示出根据本发明示例性实施例的基于双语语音查询的信息检索系统的操作的简要流程图；

图 9 是 MP3 装置的外观。

具体实施方式

图 1 是示出根据本发明示例性实施例的基于双语语音查询信息检索系统的框图。如图 1 所示,所述系统包括特征提取器 101,双语音素识别器 102、双语字符-音素(G2P)转换器 103、词汇比较器 104。图 1 中所示的 MP3 ID3 标签数据库是双语 G2P 转换器 103 的输入。特征提取器 101 被用于从 PCM 波形文件中提取一种为语音识别所特有的 MFCC 特征参数。考虑到各种麦克风的通道失真,本发明采用减小噪声的算法来消除失真。另外,本发明增加了帧压缩模块以增加移动平台上的处理速度。帧压缩的基本想法在于通过线性插值将若干相邻的 MFCC 帧压缩为一帧。双语音素识别器 102 相似于音素环(phone-loop)识别器并将特征提取器 101 的结果作为输入。在特征提取器 101 的帮助下,双语音素识别器 102 可自动将英语或汉语语音转换为文本音素串。这里,MP3 ID3 标签数据库是系统的另一输入并包含大量 MP3 文档。所述每个文档具有语音检索的可用内容,比如音乐标题、专辑、流派、艺术家名字和歌词。然后,这些可用内容被用作双语 G2P 转换器 103 的输入。双语 G2P 转换器 103 对无意义的信息进行过滤,并对无序的格式进行规格化,将文字级的串转换为音素级的串,并为词汇比较器 104 生成可能的查询候选。在双语 G2P 转换器 103 中,每个英语或汉语文字串将被转换为音素级(或对于汉语也可称作拼音级)的串。双语 G2P 转换器 103 包括两种语言相关的子 G2P 转换器。因为处理汉语和处理英语的原理完全不同,所以一个子转换器用于汉语,而另一个子转换器用于英语。词汇比较器 104 主要用于进行信息检索程序。词汇比较器 104 将从双语音素识别器 102 产生的识别的音素串(用作测试模式)和从双语 G2P 转换器 103 产生的参考音素串(用作参考模式)相比较。最后,在词汇比较器 104 中执行相似性测量以选择最相关的参考音素串输出。

下面将分别对特征提取器 101、双语音素识别器 102、双语 G2P 转换器 103、词汇比较器 104 和 MP3 ID3 标签数据库 105 进行详细描述。

1.1 特征提取器 101

现在,很多移动装置提供麦克风以便于用户输入并与其他人通信。因此,在其上进行语音应用很方便。在图 1 所示的系统中,所述系统直接采用了接收用户的语音查询的麦克风并将语音查询以 PCM 格式存储。特征提取器 101 将 PCM 波形的语音信号转换为作为语音识别的典型矢量特征的 Mel 频率倒谱系数(Mel-frequency Cepstral Coefficient, MFCC)。考虑所述系统可被应用到

具有不同种类的麦克风的各种平台，并且这种通道失真有产生识别误差的趋势，这样进一步还会影响检索精度。在本发明的系统的特征提取器 101 采用噪声减小工具以去除这些不利影响。另外，由于在嵌入式平台中 CPU 和存储器的限制，帧压缩技术被合并到特征提取器 101 中以减轻系统的负担。帧压缩技术的主要想法是将若干相邻 MFCC 矢量插入到一帧。然而，特征提取器 101 的其他部分与现有方案相似，因此将不在这里描述。

1.2 双语音素识别器 102

图 2 是示出图 1 中所示的系统的双语音素识别器 102 的详细结构的框图。更具体地讲，双语音素识别器 102 将接收压缩的 MFCC 参数流，并通过使用 HMM 方法来识别音素串，所述 HMM 方法是一种统计模式识别方法。根据本发明示例性实施例，双语音素识别器 102 包括：双语声学模型 202、参数量化模块 205、语法网络模块 203 和使用 Viterbi 算法的双语音素识别单元 201。其中，双语声学模型表示在声学空间中每个音素的时变特性，从外部语音库中训练和估计声学模型的参数。在本发明示例性实施例中，汉语音素是 68 个声母/韵母；英语音素是 39 个基础音素(base phone)，在长时间跨度的汉语韵母中的 Markov 状态的数量增加时，提供了一种简单的方法来确定某些韵母的 Markov 状态的数量。另外，限制双语音素识别器 102 的可能输出的语法网络 203 被设置在双语音素识别器 102 中。下面将给出详细的描述。

如图 2 所示，具有 MFCC 特征的压缩的数据流从特征提取器 101 被发送到双语音素识别器 102。与现有技术中的大多数音素识别器一样，双语音素识别器 102 基于隐马尔可夫模型(HMM)以更好的表现声学特性，所述 HMM 是一种需要从大量外部训练库中估计声学模型的每个参数的训练过程的一种统计学模式识别技术的方法。然而，先前的大多数对语音识别的研究都集中在单语系统。因为全球化的显著的发展趋势，英语作为一种通用语言被全世界的人们所接受。同样，英语符号频繁出现在中国的报纸、杂志和网站中。这种不断增长的趋势迫使我们在我们的系统中增加双语要素。直到现在，有两种主要的方法来建立双语音素识别器。一种典型的方法是在很多单语 SR 引擎前采用一种外部语言识别器(LID)。另一种方法是将所有语言相关的音素映射到总音素集，然后建立总声学模型。经过若干实验，由于后面的这种方法在嵌入式平台上的强健性和负荷小的特点，本发明选择后面的这种方法。

在建立声学模型的过程中，与建立总音素集有关的第一个问题是如何选

择表现汉语普通话和英语的声学特性的最好的识别单位。有若干个候选对象，比如单词、汉字、音节、声母/韵母和基础音素。对于英语 HMM，以前的研究人员证实基础音素可最好地描述英语和拉丁语系的其他语言的发音体系。由于基础音素是发音体系中最小的发音单位，所以可通过在特定规则内结合若干基础音素而容易地标记每个英语单词的发音。另一方面，只有 39 个基础音素(由 CMU 美国英语音素集得出)，这小于英语音节(多于 2 万个)或英语单词的数量。因此，选择基础音素作为识别单位将在很大程度上减小模型的大小并增加英语识别器的强健性。同时，这不同于汉语。我们知道，汉语是音调语言，它的书写体系基于汉字，发音体系基于音节(即汉语拼音)。每个汉字对应于一个带音调的音节。通常 3000 多最常用的汉字(根据 GB2321 定义)可满足日常使用。然而，如果汉字被选择为建模单位，则将会产生很多替换误差，这是因为通常若干个汉字共享相同的音节并且总共有 400 多个无音调的音节。通过“声母+韵母”的限制每个音节可分成两个半音节。在现有技术中，证实基于 I/F(声母/韵母)的普通话音素识别器与基于基础音素的系统相比，可以获得较高的性能，这是因为对音节结构的限制在总体上能够防止半音节的非法结合。而且，声母和韵母的总数是 68，远远小于汉语中音节或汉语单词的数目。因此，本发明选择声母/韵母作为汉语识别单位。

第二个问题是如何建立双语音素集。最简单的方法是将两种系统结合起来而形成一种大的音素集。从现在开始，采取上述最简单的方法的系统被称作基线系统，因此本发明在收集的库的基础上继续构造所述基线系统。在总音素库中总共有 107 个元素，其中包含 68 个汉语声母/韵母和 39 个英语基础音素。具体地讲，它不能在嵌入式平台上执行大词汇量的语音识别(LVSR)。因此，将通过 HTK 训练工具来产生总音素集中的每个音素的 HMM 以表现英语和汉语特性。

传统的语法网络如图 3A 所示。在图 3A 中带阴影的部分为英语基础音素，而不添加阴影部分表示汉语音节。为基线系统提供传统的语法网络，并可以发现在双语语音识别中有很多可混淆的跨语言转换。在这种语法中，所有汉语声母/韵母间的组合遵循特定的音节结构，而英语基础音素具有聚合在一起的更多的自由度。显然转变跨语言相当随意。为了方便，根据本发明示例性实施例的系统选择单高斯概率密度函数(PDF)来基于 HTK 工具建立 HMM。这种实验结果显示基本系统的音素准确度与各个单音系统相比减小了很多。

可以得出在发音体系中，半音节可被进一步分为一个或若干个基础音素串。在 Viterbi 束搜索中，长时间跨度的汉语韵母（比如“iong”、“iang”和“van”）通常被若干个短跨度英语基础音素代替。在另一方面，在一些短跨度 I/F 和英语音素之间在声学域存在严重的重叠。

这里提出了两种方法，它们被结合到一起来解决上述问题。一种方法是扩展每个长时间跨度汉语韵母的 HMM 中的 Markov 状态的数量。在理论上，长时间跨度半音节应该比基础音素占用更多的特征帧，并且长时间跨度韵母可进一步被分离为（头）+主体+（尾）。例如，“iong”可被分为“i”+“o”+“ng”；“van”可被分为“v”+“a”+“n”。因此，对于“iong”和“van”，传统的 3 状态 HMM 不能满足声学域中的时变。最好的方法是将每个长时间跨度韵母的状态数量进行扩展以与其平均持续时间成比例。在本示例性实施例中，采用了一种非常简单的方法，该方法根据特定韵母所具有的基础音素的多少来调整该韵母的状态数量。在基线系统中，为每个 HMM 保持 3 状态，而在建议的方法中，考虑到平衡，当韵母包含两个子音素时，又添加了 2 个状态（总共 5 个状态），而当韵母包含 3 个子音素时，又添加了 3 个状态（总共 6 个状态）。在表 1 中列出了更多细节内容。实验结果显示：与基线系统相比，增加了英语和汉语部分的音素准确度。总之，这种方法在某种程度上防止了替代误差。

汉语韵母的种类	状态数量
ai an ang ao en eng ia in ing ong ou ua ui un uo ve vn	5 (3+2)
ian iang iao iong uai uan uang ueng van	6 (3+3)
其他	3

表 1 总音素集的状态数量

从表 1 中可以看出，在基线系统中，每个韵母的状态数量保持为 3，而在本发明中，当汉语的韵母包含两个子音素时，又添加了 2 个状态，而当韵母包含 3 个子音素时，又添加了 3 个状态。

双语声学模型以及语法网络，是双语音素识别器的必要的输入。语法网络是很重要的部分，它限制了一个识别器的可能的输出。上面提到的两种方法的另一种方法是新的语法网络。语法网络模块 203 与声学模型模块 202 一样，也是双语音素识别单元 201 的必要的输入。

由于在双语语音识别期间有很多可混淆的跨语言转换，因此提出了如图

3B 所示的语法网络，如 3B 所示的语法网络基于如图 3A 所示的传统语法网络。在图 3B 中带阴影的部分为英语基础音素，而不添加阴影部分表示汉语音节。图 3B 的可能输出串与图 3A 的相同。然而，它们之间的主要不同在于添加了具有单一状态的静音（在图中对应于 sil）模型以阻止跨语言转换。换句话说，如果发生跨语言转换，则必须通过无声模型，并且无声模型占用不少于两个 MFCC 帧。其原理相似于一种惩罚以有效防止可混淆的跨语言转换的发生。双语语音识别的准确度增加很多。在传统语法网络中，对跨语言转换没有限制，从而长时间跨度的汉语韵母可自由地被若干应用基础音素所替代。而且，两种语言的相似音素还可被混淆地相互交换。这种现象主要发生在汉语声母和其相似英语辅音之间。我们发现大多数的目标查询词是或者单以汉语或者单以英语的小短语。很少发生混合语言查询词。即使有，跨语言转换的频率小于 1 次。例如，“活着 viva”，“爱的 complain”。因此，传统语法的识别结果在很大程度上与事实相背。通常可以添加额外的语言模型，通过使用语言模型中大量的构词规则来防止它。然而，考虑到 CPU 和存储器的成本，在嵌入式平台上增加语言模型是一个障碍。有希望的一个途径是对识别器的语法增加特殊的限制。在本发明的示例性实施例中，传统语法被分成若干部分，其中，一部分是用于识别汉语的音节子环，另一部分是用于识别英语的基础音素子环，这两个子环通过若干 1 状态的无声模型连接。忽略无声模型的效果，我们建议的语法可输出通常的语法也支持的所有可能的音素串。主要的区别在于当发生跨语言转换时，无声模型必须占用若干帧。这些无声模型有效地防止了不必要的跨语言转换并保持了一种语言的连续性。

基于第一种方法，本发明将这种新语法应用到语法网络。实验结果显示与所述基线系统相比，根据本发明示例性实施例的双语系统的准确率有了很大进步。其性能更接近于每种单语系统。证实了该建议的语法具有防止不必要的跨语言转换的能力。

为了在嵌入式平台上提高处理速度，利用参数量化模块 205 通过矢量量化方法来将双语声学模型的连续参数改变成离散参数。矢量量化的优点如下所述。通常，对于输入的特征矢量和连续的声学模型，应该计算特征矢量属于特定模型的概率。所述概率被定义为高斯分布，所述高斯分布的参数（比如均值、方差）被存储在声学模型中。通常计算高斯分布的输出概率是件很耗时的的工作。因此，本发明利用参数量化模块 205。在参数量化模块 205 中，

在多维空间中建立了成百的质心，并且每个质心给出了一个索引号。如果特征矢量或者模型参数最接近于一个特定的质心，那么我们使用该特定质心来近似地代替原始连续参数。然后，通过查找为每对质心存储预先计算的概率的二维表来快速地查找输出概率。例如，如果发现特征矢量最接近于第 I 质心，并且发现一个模型的高斯分布的参数最接近于第 J 质心，则系统将在预先计算的表中查找第 I 行第 J 列的元素，并选择该值作为输入。显然，与使用连续参数相比，嵌入式系统的负荷很大大地减小。因此，参数量化模块 205 的主要作用是用于两个声学模型的参数，以方便于下面在双语音素识别单元 201 中的计算。

双语音素识别单元 201 是双语音素识别器 102 的核心部件，并且需要声学模型和语法网络作为它的输入。所述双语音素识别单元 201 自动识别不同语种的语音特征，然后将执行两步识别过程以将语音特征转换为识别的文本音素串。

与基于 HTK 的识别器不同，所述双语音素识别器 102 利用两步搜索，它使用单音素 HMM 和 Viterbi 算法来创建音素网格，然后使用三音素 (tri-phone) HMM 和 A-star 算法来从网格中跟踪 N 个最好的路径。实验证实这种两步识别方法比传统的基于 HTK 的识别器具有更快的解码速度，适合于移动装置。

1.3 双语 G2P 转换器 103

如上所讨论的，总音素集作为双语音素识别器 102 的识别单位而被建立。所述总音素集包括 68 个汉语 I/F 和 39 个英语基础音素。为了与双语音素识别器 102 保持一致，将该总音素集选择为检索系统的检索单位。双语 G2P 转换器 103 被用于过滤冗余信息，将参考文本转换为音素串，并输出可能的查询词汇。将在图 4 中详细显示双语 G2P 转换器 103 的详细结构。在示例性实施例中，双语 G2P 转换器 103 的可能输入是 MP3 ID3 标签数据库中的内容，包括标题、专辑、艺术家名字、流派、歌词。在 MP3 ID3 标签数据库中，包含大量 MP3 文档，所述文档的每个文档都具有用于语音检索的可用内容，比如音乐标题、专辑、流派、艺术家名字和类型。通常，这些内容不仅是无序的并且还充满无意义的字符。

图 4 示出双语 G2P 转换器 103 的结构。根据本发明示例性实施例，双语 G2P 转换器 103 包括语种识别器 405、汉语 G2P 转换器 406、英语子 G2P 转换器 407 和特殊字符 G2P 转换器 408。另外，双语 G2P 转换器 103 还可包括

无意义字符确定器 401、文本过滤器 402、格式规整器 403 和词间分割器 404 以用于进行分词查询。无意义字符确定器 401 确定 MP3 ID3 标签中的内容是否包含无效或者无意义的字符，例如，各种标点符号和认为定义的符号（比如 ^、_、@ 和 © 等）。如果包含无意义的字符，则文本过滤器 402 将它们滤除。然后格式规整器 403 将无序的内容规整为“标题_名字_专辑”的统一格式。词间分割器 404 可将完整的音乐标题分离成很多条分词。考虑到双语背景，采用两种语种相关的子 G2P 转换器 406 和 407 以及一种另外的特殊字符 G2P 转换器。我们知道，英语习惯于使用 ASCII 字符编码，而汉语使用 GB2321（简化汉字）。由于在两种字符编码集之间没有重叠，通过语种识别器 405 容易找到语种边界，然后发送与语种相关的子 G2P 转换器 406 和 407 的每一片段。在英语 G2P 转换器 407 中，因为有太多英语单词要被包括在一个词库中，最好存储 20000 多的英语音节而非单词。因此，当输入了英语单词时，双语 G2P 转换器 103 可根据单词的音节结构结合来拼出它的发音。拼写方法可有效解决英语中严重的超词汇量(OOV)的问题，尽管拼写发音与其实际具有一些误匹配。对比地，汉语声母/韵母被选择为检索单位，从而在汉语 G2P 转换器 406 中有效地避免了 OOV 问题。通常，多于 400 的无音调字节可覆盖了所有的汉字的发音。然而，对于很多汉字，比如“和”和“绿”，当在各种不同词汇上下文中使用时存在多个发音（即多音）。因此，如何检测汉字边界以及如何正确地标记汉语多音字的发音，依然是研究课题。考虑到在嵌入式平台上的 CPU 限制，采用了最大匹配技术，所述最大匹配技术是一种为汉语句子找到正确发音的贪婪算法。最大匹配算法是：首先创建存储拼音级的 20000 多常用词和 6000 多汉字发音的库。库中各项的等级按词的长度分类。当输入了汉语句子时，使用库中最长的词来检查它是否与句子的前面部分匹配。如果匹配，则将词的边界标记为音频，如果不匹配，则继续检查第二最长的词，第三最长的词……直到找到匹配的词。然后为剩余的部分重复这种处理直到句子中的所有汉字被标记为拼音。

此外，将一些在歌词中频繁出现的特殊词和单个的汉字添加到汉语 G2P 转换器 406 中用户定义的库中以便于进行音乐检索。无疑汉语字符到音素的转换的标记准确度与预先存储的库中有多少多音字成正比。换句话说，如果 G2P 可包括很多多音现象，则可获得较高的准确度。然而，较大的库则意味着它将在单词匹配上花费非常多的时间。因此，应该根据不同任务而在速

度和性能之间进行折衷。特殊字符 G2P 转换器 408 被用于处理特殊情况(比如数字、英语缩写或者一些被语言相关的 G2P 转换器 406 和 407 丢弃的其他字符)。例如,当输入字符是数字时,我们首先考虑其上下文背景。如果它属于汉语环境,则将其标记为汉语发音。最后,各个语言相关的 G2P 输出的转换结果将在结合器 409 中按照原先输入文本字串的排列方式进行组合。

1.4 词汇比较器 104

词汇比较器 104 的结构如图 5 所示,词汇比较器 104 包括相似度计算器 501、混淆度矩阵模块 502、分值确定器 506 和文字恢复器 503。在词汇比较器 104 中,将执行信息检索处理。词汇比较器 104 具有两个输入:来自双语音素识别器 102 的识别的音素串(现称作“测试模式”)和来自双语 G2P 转换器 103 的很多参考音素串(现称作“参考模式”)。为了与音素识别器和双语 G2P 转换器保持一致,汉语 I/F 和英语基础音素的结合被选择为词汇比较器 104 的检索单位。由于参考音素串的数量相当大,修改的快速 DTW(动态时间伸缩)算法被应用于相似度计算器 501 以快速计算测试模式和参考模式之间的接近程度。修改的 DTW 算法的主要目的被显示在稍后将要详细描述图 7 中。

这里,来自双语音素识别器 102 的识别的音素串将被用作测试模式以产生一系列相关的 MP3 文档。如上所述,为了与双语音素识别器 102 保持一致,汉语 I/F 和英语基础音素的结合被选择为词汇比较器 104 的检索单位。MP3 ID3 标签数据库中的所有可用内容被双语 G2P 转换器 103 提取并被用作参考模式的参考音素串。对于一些音乐标题,特别是在英文歌曲中,标题通常包含多于 5 个的单词。它们中的大多数对于非母语的人们是无意义的或者是模糊的。为了方便,引入了“分词查询”的新功能。如果用户仅记得一个音乐标题的一些分词,他可讲这些分词以代替整个标题名称,并且系统将会自动地对其进行分析。最后,其标题包含这些分词的相关文档将被返回。为了实现这种新功能,词间分割器 404 被添加到双语 G2P 转换器 103 中。在词间分割器 404 中,完整的音乐标题将被分离为很多个分词。例如,对于汉语分词的最小单位是汉字,而对于英语,该最小单位是单词。词间分割器 404 将不管语法正确与否,试着列出分词的所有结合,以覆盖所有可能的实例。然而,存在这样的规则:只有包含多于两个音节的结合才能成为可能的查询词。以音乐文件名为“David's love is the wave”作为例子,标题可被分离为六个分词:

“David”、“s”、“love”、“is”、“the”和“wave”。考虑到不同的结合，词间分割器 404 可如图 6 所示输出若干可能的候选，比如“David”、“David’s”、“David’s love”、“David the”、“David wave”和其他很多情况。分词的所有结合都是用户检索其标题包含这些分词的特定文档所使用的可能的查询词。这样做的一个重要的优点是用户可讲一些关键词而不是整个名称，特别是当标题很长而难于复述的时候。

显然，当引入分词查询时库大小会立即扩大。通常，1000 首歌曲的数据库可产生具有多于 5000 的分词的结合的库。为了加速检索处理，我们必须利用相似度计算器 501 来计算测试模式和参考模式之间是的接近程度。

可以以各种方式来进行相似度测量。一种方式是基于规则的方法，它需要声学知识。另一种方式是数据驱动的方法。在示例性实施例中，采用修改的 DTW 算法来计算测试模式和参考模式之间的相似度。混淆度矩阵模块 502 被用于在应用 DTW 之前创建混淆度矩阵 C。这里，混淆度矩阵 C 被定义如下：

$$C = \begin{pmatrix} P_{1,1} & \cdots & P_{1,m} \\ \vdots & \ddots & \vdots \\ P_{m,1} & \cdots & P_{m,m} \end{pmatrix}$$

其中，混淆度矩阵 C 中 P_{ij} 表示在语音识别期间第 j 音素被错误地识别为第 i 音素的概率。通过训练样本的强制对齐方法而容易地获得混淆度矩阵 C。如图 7 所示，创建了包含 $M \times N$ 个节点的 2D 网络，其中，M 和 N 分别表示测试模式和参考模式中音素的数量。在图 7 中，每个节点具有从混淆度矩阵获得的数字格式的音素相似度的得分。测试模式来自于识别的结果。通常，在根据本发明实施例的音素识别器中得分最高的结果的准确度接近 60%。因此我们合并了得分最高的前 N 个结果以确保检索性能。由于所有可能的分词已经作为参考而被列出，所以与一般的 DTW 不同，为了提高速度而固定了开始和结束节点。然后将利用动态规划算法以找到到达最高分数的最佳路径。具体地讲，对选择后续路径上存在另一限制。将图 7 作为示例，因为忽略了删除错误，所以禁止两个节点在同一行上的路径。因此，只有虚线表示的菱形的路径在路径搜索期间被考虑。这种限制在很大程度上减少了网络的复杂性。根据这种方式，数据库中的所有参考项被重复地输入到相似度计算器 501 以用当前查询来计算相似度得分。然后系统对所有得分进行分类并返回一列

最相关的 N 个节点。反转词典模块 505 存储音素串和其原始文字级的名称之间的链接。翻转词典模块的主要功能是从缺乏理解力的音素串找到其原始文字级的名称。分值确定器 506 确定得分是否大于预定阈值。当得分大于所述阈值时，文字恢复器 503 在反转词典模块 505 的帮助下将 N 个最相关的项的音素串恢复为原始文字级的文件名称、艺术家的姓名、专辑……，然后将这些文字串输出到 UI 显示装置。当得分低于阈值时，根据本发明的系统将从 MP3 ID3 标签数据库 105 重新加载新纪录，并重复上述的相似度测量的处理直到所有的记录已经与测试模式相比较。另外，词汇比较器 104 还可包括结果验证器 504，结果验证器 504 用于在查询语音非法或者 MP3 ID3 标签数据库没有包含相关 MP3 文档时防止错误的输出。

1.5 MP3 ID3 标签数据库 105

在本发明示例性实施例中，采用基于双语语音查询的信息检索系统以将 MP3 文档安置于特定的嵌入式产品中。因此，需要提供用于查询的内容的大数据库。在中国，人们可经过互联网和无线信道从著名的音乐网站将 MP3 文档容易地下载到用户的 MP3 播放器或者蜂窝电话。在这种情况下，如果用户的 MP3 播放器包含很多文档，那么前端程序将自动地递归地扫描每个目录，找到后缀为“.mp3”、“.wma”等与 MP3 有关的所有文件，并从所有 MP3 文档中提取 ID3 标签信息。根据 MP3 ID3 的定义，项目的格式被很好地组织。在 ID3 标签中，包含所有重要的信息，比如“艺术家的名字”、“专辑标题”、“音乐标题”、“流派”和“歌词”等。

然后所述前端程序将对来自 ID3 标签的所有信息分裂为数据库。所述数据库与传统的关系数据库相同。在理论上，一个关系数据库包括一个或多个表。根据本发明，每条记录包含三个属性，它包括：“音乐标题”、“专辑标题”和“艺术家的名字”。通常，上述的三个属性可最好地表现 MP3 文档。在它们之中，我们可将“音乐标题”设置为关键字段以识别每条记录，这是因为对于其它属性，不同的记录共享相同的属性“专辑名称”或“艺术家名字”。

在这种条件下，数据库可以二进制格式被存储在固定的目录中。根据本发明的检索系统中的双语 G2P 转换器 103 通过结构化查询语言 (SQL) 或者预先定义的命令首先将查询命令发送到数据库。其次，将返回所有相关记录的集合。所述双语 G2P 转换器 103 将使用它们作为输入来实现检索任务。

图 8 是示出根据本发明示例性实施例的基于双语语音查询的信息检索系

统的操作的简要流程图。

在操作 810, 特征提取器 101 将语音信号从 PCM 波形转换到 MFCC 特征参数, 并在噪声减小和帧压缩处理后输出压缩的 MFCC 数据流。

在操作 820, 双语音素识别器 102 接收压缩的 MFCC 数据流, 通过自动地将英语和汉语语音改变成相应的文本音素串来执行语音识别。所述双语音素识别器将作为识别单位的汉语声母/韵母和英语的基础音素结合成总音素集, 并通过在汉语韵母具有 2 个子因素时多添加 2 个状态, 而在汉语韵母具有 3 个子音素时多添加 3 个状态来调整 Markov 状态的状态数量。另外, 所述双语音素识别器 102 采用新的语法网络以阻止跨语言的转换。

在步骤 830, 双语 G2P 转换器 103 将 MP3 ID3 标签数据库中 MP3 文档的用于语音检索的可用内容的文字级的串转换到作为参考音素串的音素级的串。具体的转换过程已经参照了双语 G2P 转换器 103 进行了描述, 为了清楚和简明起见, 这里将省略对它的详细描述。

在步骤 840, 词汇比较器 104 将从双语音素识别器 102 产生的识别的音素串与从双语 G2P 转换器产生的参考音素串相比较, 并输出最相关的音素串。已经参照词汇比较器 104 的结构详细描述了比较操作, 所以为了清楚和简明, 将省略对该操作的详细描述。

所述建议的系统的目标平台是大存储容量的 MP3 播放器或者音乐蜂窝电话。通常, 这些装置可拥有 1000 多个 MP3 文档。与传统的 MP3 播放装置不同, 通过手动操作来搜索他想要的文档是不实际的。因为语音输入具有输入速度快和容易理解的优点, 所以基于语音查询的检索系统变得很有必要。在中国, 因为英语符号频繁出现在作为查询文字的所有可能内容的音乐标题、专辑和歌词中, 我们添加了具体特别的技术以执行双语检索。

目标平台的外观如图 9 所示。通常, 当用户聆听某一 MP3 歌曲时, 他(或她)必须操作导航条来在显示 UI 上浏览菜单, 并选择期望的一首歌曲。此外, 如果用户想要聆听属于某一专辑、流派或某一艺术家的一系列歌曲, 他必须重复的逐个选择这些项目, 所以很不方便。随后他可点击按钮 II 来播放音乐列表。

在应用了根据本发明的双语检索系统之后, 操作变得完全不同。在移动装置的顶部, 提供了一个麦克风让用户输入他的语音。通常, 他存储容量的 MP3 播放器和音乐蜂窝电话可拥有 1000 多首 MP3 歌曲。当然, 用户不可能

记住装置中的所有音乐标题。大多数人只是记得若干个关键词或代表性的短语。以音乐标题“思念是一种病”作为例子，人们通常记得它的若干分词，比如“思念”或“病”。因此，本发明添加了新的方法，即，分词查询。每个分词和它自己的音乐文档之间有一个链接以加快检索。

在采用根据本方面示例性实施例基于双语查询的检索系统的装置中，提供了两种检索方法。一种方法是像传统方式那样使用导航器和显示屏幕来进行检索，另一种方法是语音查询系统。当用户选择并点击语音查询的图标时，根据本发明的系统将会启动。现在除了按钮 I 之外的所有按钮都是无效的，系统将弹出欢迎消息以等待用户的语音输入。然后用户一直按着按钮 I 直到他（或她）讲完话。在这期间，用户的嘴一直对这麦克风。讲话的内容可以是与特定内容音乐标题、专辑名称或艺术家的姓名相关的多个分词。例如，为了检索“思念是一种病”，用户可说“思念”、“病”、“是一种病”或者甚至是它的整个名字。对于查询也可能是所有可能的结合。这是根据本发明示例性实施例的基于双语语音查询的信息检索系统的第一个创新点所在。

在释放按钮 I 之后，根据本发明的基于双语语音查询信息检索系统将会迅速地将用户的语音转化为相应的文本、通过相似度测量来检索相关文档并在屏幕上显示它们的列表。现在所有的按钮恢复正常使用状态。如果所述列表包含多个需要的文档，则用户可使用导航器来对它们进行多项选择。当前，通过导航器进行选择的范围远小于传统的方法。然而，如果列表不包含需要的文档，则用户可再按下按钮 I 并重复上述的处理直到用户满意。最后，用户点击按钮 II 来播放音乐列表。

根据本发明的基于双语语音查询的检索系统的目的是检索与用户输入的语音最相关的文档列表。总之，根据本发明的基于双语语音查询的检索系统是语音识别和信息检索技术的有效结合。该系统直接使用装置上的麦克风来输入查询语音。更重要的是，该系统为双语用户设计。换句话说，该系统能够处理汉语和英语语音查询并对不同语言保持同样的高性能。

另一创新点是支持双语查询。换句话说，用户可将英语和汉语以进行查询，并且根据本发明的系统能够很好的处理两种语言。这个创新点很大程度上取决于内部语音识别模块和信息检索模块。在图 9 中显示的装置的外观中，不存在为不同语言切换的切换器。可不经任何切换器来进行用户的多语种语音顺利识别。

尽管已经显示并描述了本发明的示例性实施例，但是本领域的技术人员应该理解，在不脱离本发明的原理和精神的情况下，可对这些实施例进行各种改变。本发明的范围由所附的权利要求和等同物所限定。

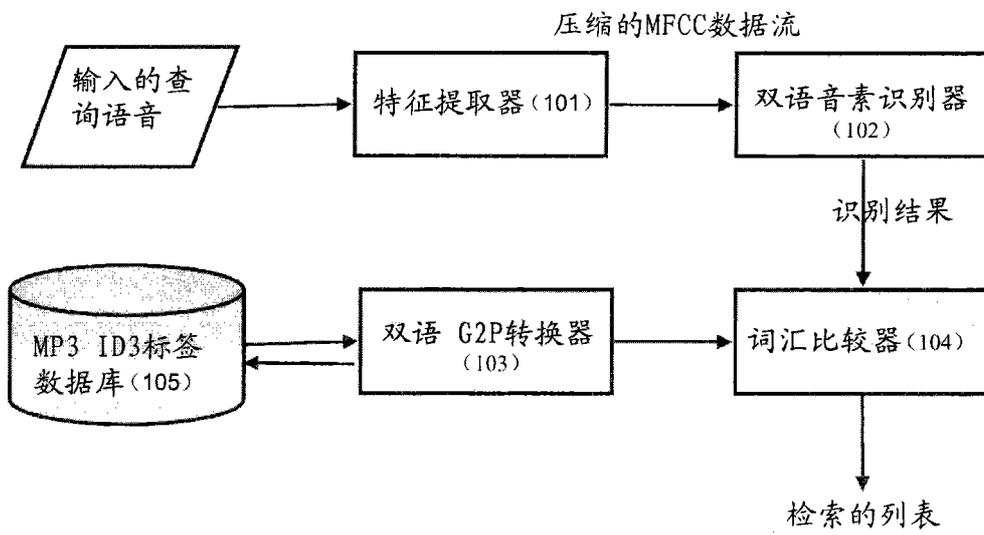


图1

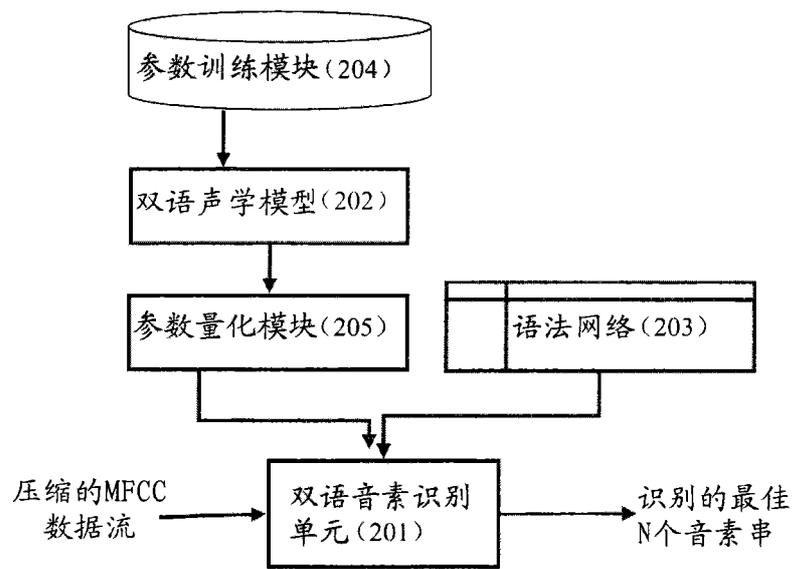


图2

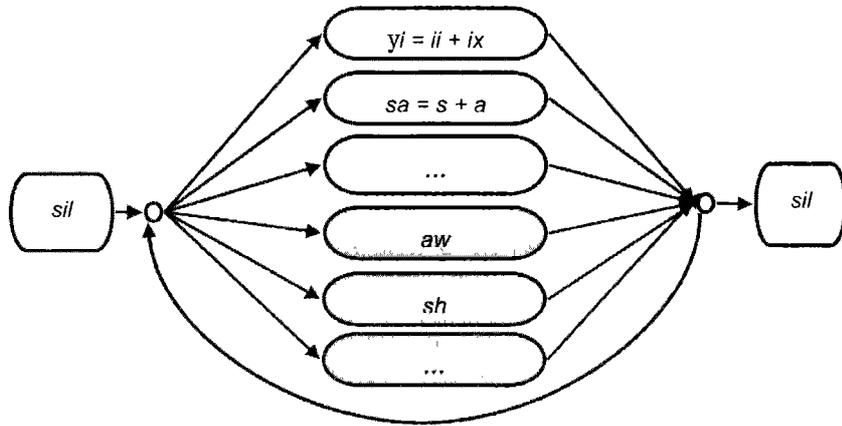


图 3A

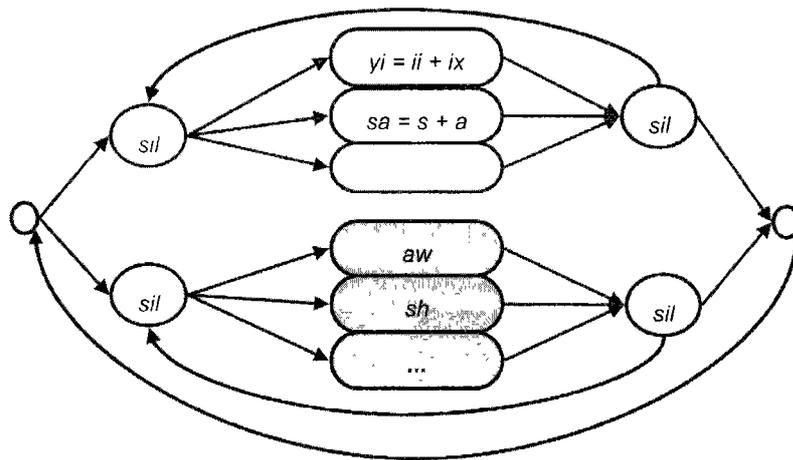


图 3B

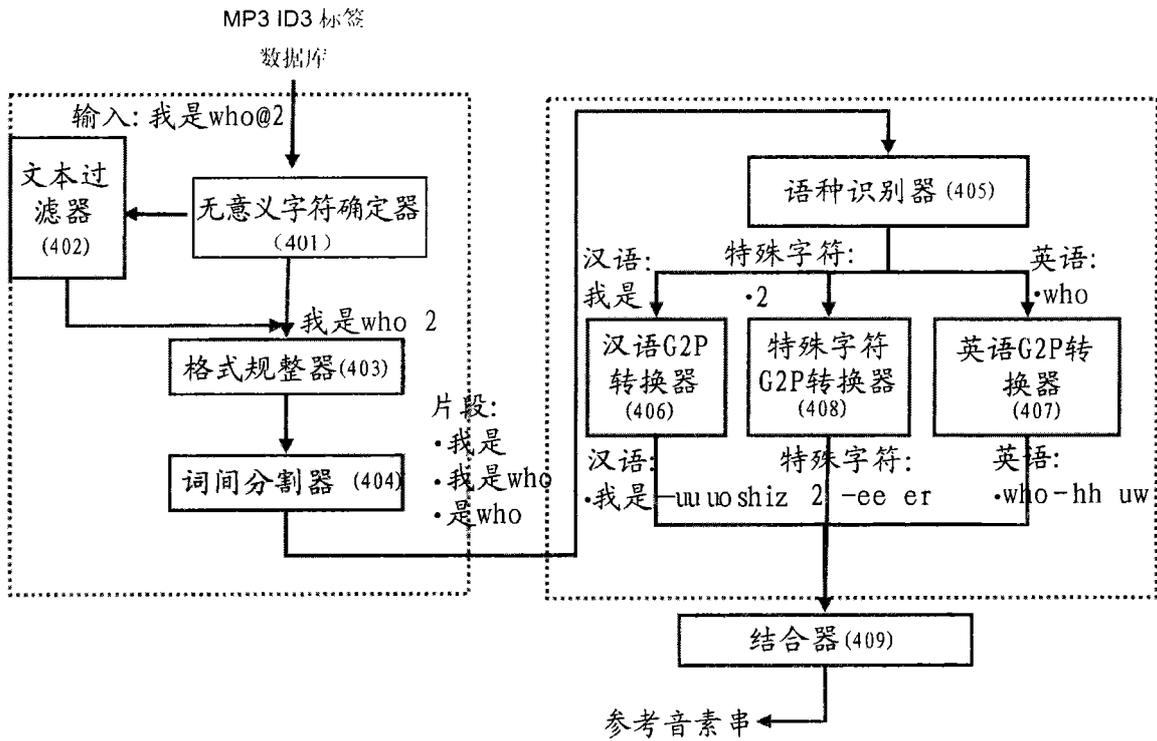


图4

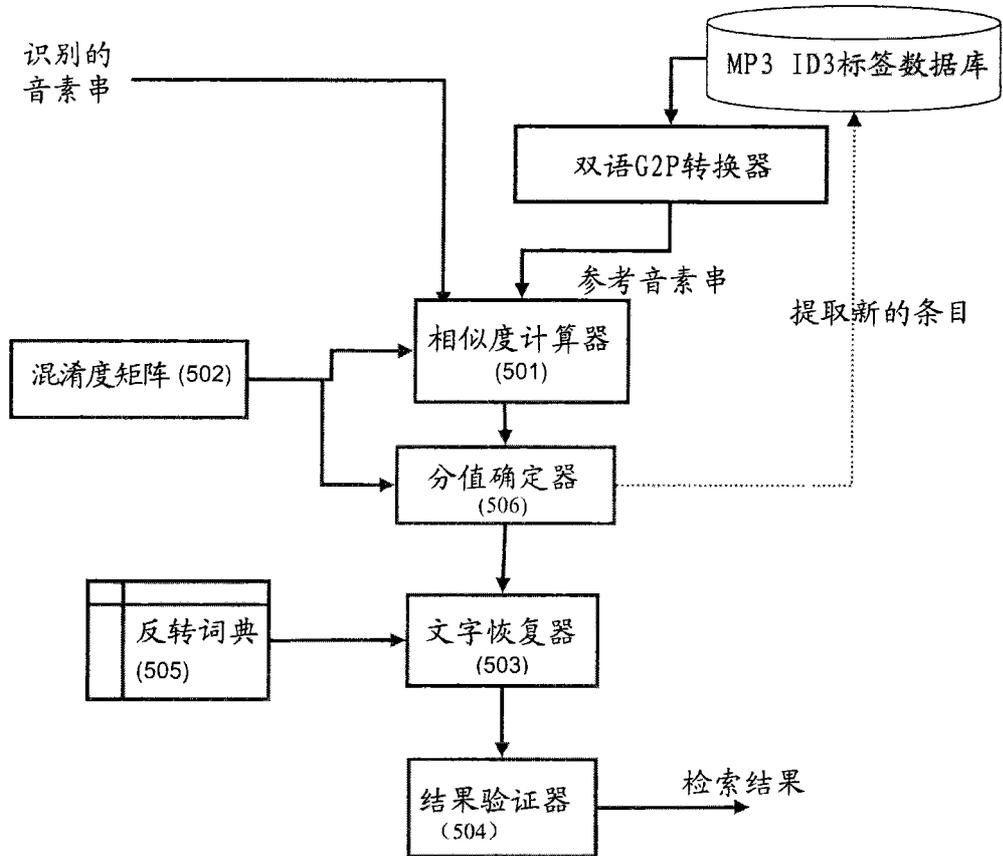


图5

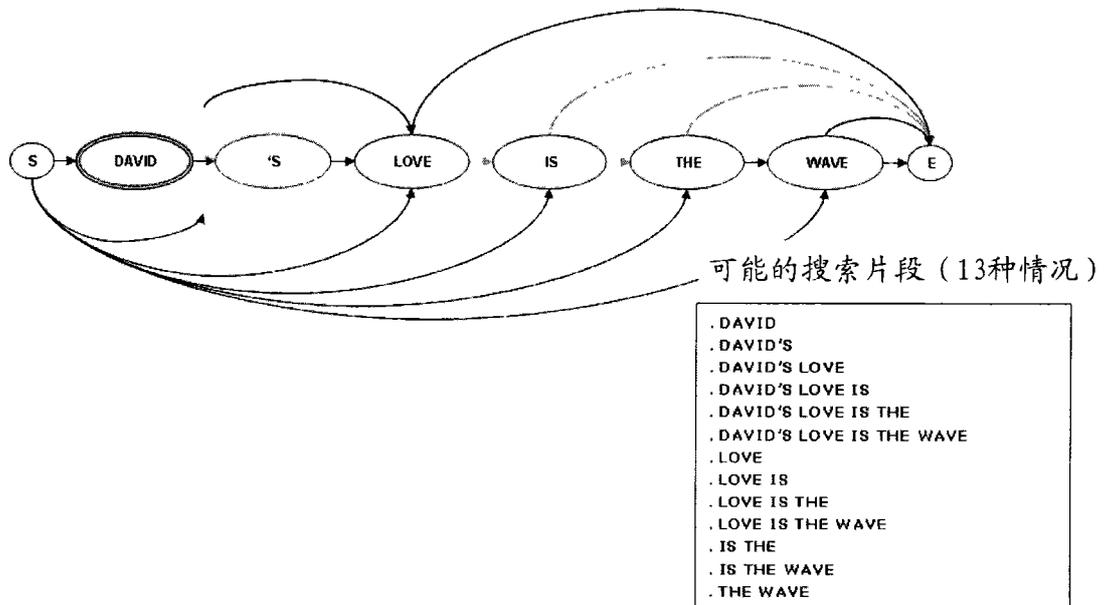


图6

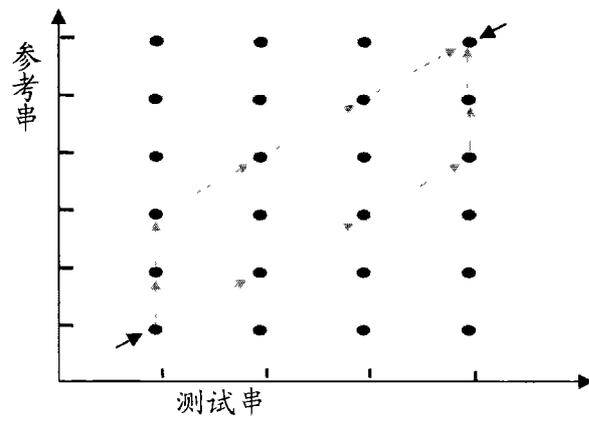


图7

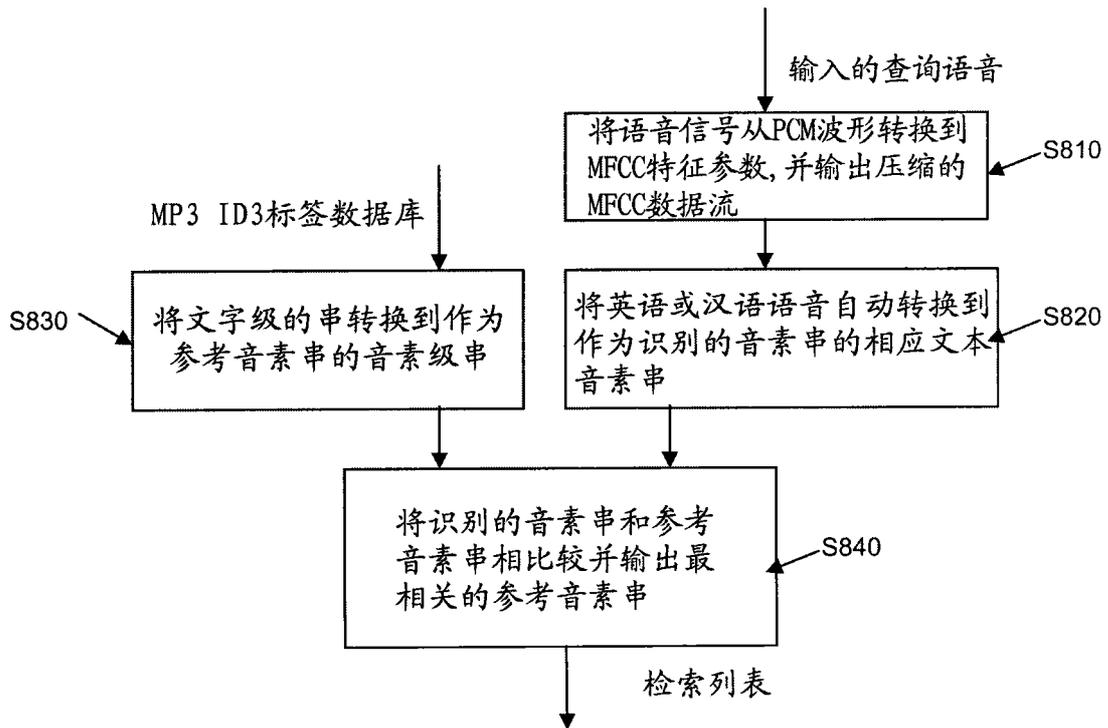


图8

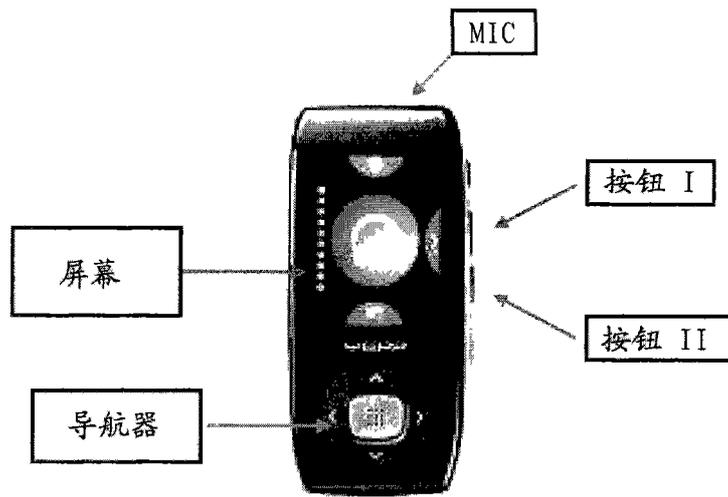


图9