[19]中华人民共和国专利局

[51]Int.Cl[6]
G10L 5/06

[12] 发明专利申请公开说明书

[21] 申请号 95195955.7

[43]公开日 1997 年 10 月 15 日

[11] 公开号 CN 1162365A

[22]申请日 95.11.1

[30]优先权

[32]94.11.1 [33]EP[31]94308023.4

[86]国际申请 PCT/GB95/02563 95.11.1

[87]国际公布 WO96/13827 英 96.5.9

[85]进入国家阶段日期 97.4.29

[71]申请人 英国电讯公司

地址 英国英格兰伦敦

[72]发明人 S·P·A·林兰

[74]专利代理机构 中国专利代理(香港)有限公司

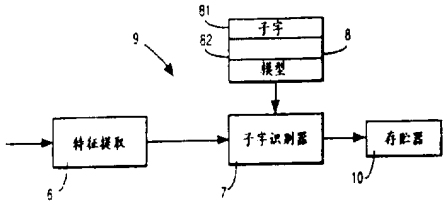
代理人 王 勇 张志醒

权利要求书 2 页 说明书 9 页 附图页数 3 页

[54]发明名称 语音识别

[57]摘要

一种语音识别器，其中识别字表由使用者自己的语音通过形成该用户发音的音素的记录产生，并且将这些记录用于或进一步的识别。该音素记录使用一个松散约束网络来产生，最好只受噪声约束。因此，所产生的记录同使用者的输入语音非常相似，但同已知的与说话人有关的字的表示法相比，它需要大大减少的存贮量。



权　利　要　求　书

1、为语音识别装置产生字表的一种方法，该方法包括:

接收表示一次发音的输入语音信号;

由每次发音产生一个代码，该代码从多个参考子字表示中识别一个与发音最相似的参考子字表示的序列;

为下一步识别存贮所产生的发音的代码。

2、如在权利要求 1 中声明的一种方法，其中子字表示序列是松散约束的。

3、如在权利要求 1 中声明的一种方法，其中子字表示序列是不受约束的。

4、如在权利要求 1，2 或 3 中声明的一种方法，其中表示相同发音的不只一个输入语言信号被输入用于产生该发音的代码。

5、如在权利要求 1，2，3 或 4 中声明的一种方法，其中子字是音素。

6、字表生成装置包括:

一个子字识别器，用于从每个输入语言信号产生一个代码，来从多个参考子字表示中识别一个与输入语言信号最相似的参考子字表示序列；和

一个用来存贮用于下一步识别的输入语言信号的代码的存贮器。

7、如权利要求 6 中声明的装置，其中子字识别器具有一个松散约束的语法。

8、如权利要求 6 中声明的装置，其中子字识别器被安排用于识别一个不受限制的参考子字模型序列。

9、语音识别装置包括一个语音识别器，用于比较输入语音信号和子字表示序列并输出一个表明识别（结果）的信号，其中子字表示序列由如权利要求 6，7 或 8 中声明的字表生成装置产生的代码（识别）确定。

10、由权利要求 9 中声明的装置，进一步包括存贮字的（一个）代码的第二存贮器，其中的代码以不同于那些存贮在第一存贮器中的代码的方式来产生。

11、如权利要求 10 中声明的装置，其中字的代码确定一个参考子

字表示的序列。

12、一种无线通信服务，使用如权利要求 9 到 11 中任一条所声明的装置。

13、根据权利要求 12 的一种无线通信服务，其中的服务是指一种指令拨号服务。

说　明　书

语音识别

该发明涉及语音处理特别是语音识别过程。

语音识别装置的开发者有一个生产出使人能够以完全自然的，没有限制的方式与之交互相应的机器的最终目标。人机之间的接口应该是理想的完全无缝的。

这是一个越来越接近实现的一个梦想，然而人机之间的完全流畅还仍然没有实现。为了实现流畅，一个自动识别装置将需要一个无限的字表而且必须不管每个使用者的口音及发音清晰程度如何，都能理解他们的语音。现有的技术以及我们对于人类怎样理解语音的有限的了解使得这点是不可行的。

当前的语音识别装置包括涉及到这个装置能够识别的有限词汇表的数据。这些数据通常涉及表示有限词汇表中字的统计模型或模板。在识别过程中，一个输入信号被与贮存的数据进行比较以判定该输入信号与存贮数据的相似性。如果找到一个足够相近的匹配，该输入信号通常被认为将被识别为提供最接近匹配的模型或模板（或模型或模板序列）。

模板或模型通常通过测量输入语音的特定特征值而形成。这些特征值通常是某种形式的谱分析技术的结果，例如，一种滤波器组分析器、一种线性预测编码分析或一种离散变换分析。典型的，一个或多个对应于同一语音（例如，一个特殊的词，词组）的训练输入的特征值被用来创建一个或多个代表该声音特征的参考样本。该参考样本可以是一个从某种平均技术中得到的模板，或者是一个表征特定声音的训练输入特征值统计特性的模型。

未知输入被拿来与识别字表中的每一声音的参考样本比较，并计算未知输入与每一参考样本之间的相似性测量值。这个样本分类步骤可以包括一个全局时间校正程序（众所周知的动态时间弯折（ warp ）DTW ），用来补偿说话的不同速度。这些相似性测量值然后被用来判定哪个参考样本与未知输入最佳匹配，由此判定该未知输入被识别成什么。

语音识别装置原计划的用途也可决定系统的特性。例如一个设计为

与说话者有关的发音系统只需要从单一讲话者得到训练输入。那么这些模型或模板就代表一个特殊讲话者的输入语音而不是一些用户的平均语音了。尽管这样一个系统对于给出训练输入的讲话者有一个很好的识别率，该系统显然显然不适于被其它用户使用。

与讲话者无关的识别依赖于由大量讲话者的发音形成的字模型。为了连续识别的目的，形成了表示每一特殊语音输入的所有训练发音的统计模型或模板。尽管与说话人无关的系统对于大量的使用者来说运行相对良好。但是对于口音，音调和语调等与训练样本差别很大的使用者来说，与说话人无关的系统的表现就可能会差。

为了扩展可接受的词汇表，必须得到附加词汇表的充足的训练样本。这是一个耗费时间的操作，当词汇表反复改变时，这种操作可能会不合算。

众所周知的是提供一个语音识别系统，在这个系统中使得能被该系统识别的词汇表可以通过一个以文本方式输入附加词汇表的服务提供装置来扩展。这种系统的一个例子是 AT & T 的 Flexword 。在这样一个系统中，根据语言规则词汇被由文本方式转换成为它们的语音学记录。正是这些记录被应用于拥有每一个音素的声学模型的识别装置中。

一种语言中的音素数常常是一个判断的根据，它可能依赖于与之有关的特殊的语言学家。在英语中，大约有 40 个左右的音素，如表 1 所示。

这里参考的音素或子字涉及任何一种字的方便的构成结构块，例如音素、音素串、别音（ allophone ）等等。这里参考的任何音素或字节都是可互换的并且参考这种广义的解释。

为了达到识别的目的，根据代表单个音素的贮存模型可形成一个以音素方式记录的文本网络。在识别过程中，输入语音被与表示每个可用的字或词组的参考模型串比较。表示单个音素的模型可以以与说话者无关的方式，根据一定数目同说话者的发音产生。任何一种适合的模型都可被使用，例如隐马尔可夫模型。

一个这样的系统中，与字的标准音素记录的偏差是一点也不允许的，例如一个人有很重的口音。因此，即使一个使用者说了一个该系统字表中有的字，输入语音也有可能不被识别成那个字。

理想的是能够调整与说话者无关的系统使得它能被一个发音与模

型发音者发音不同的使用者使用。欧洲专利申请第 453649 号描述了这样一个装置，该装置字表中的许可字由表示字的子单元，例如音素的模型的链接来模拟。这种“字”模型，也就是说存贮的链接，于是通过从使用者语音估计字模型新参数的方式被训练为特定使用者的语音。这些已知的预定字模型（由音素模型的链接形成）被调整为适用于特定使用者。

同样的，欧洲专利申请第 508225 号描述了一个识别装置，其中要被识别的字被与表示该字的音素序列存贮在一起。在训练过程中，使用者说字表中的字，则音素模型的参数被调整以适应使用者的输入。

在这些已知系统中，要求一个以音素序列链接形式的预定字表。然而，在很多情况下，对于一个使用者来说理想的是往字表中加入字，这些字是特别针对该使用者的。唯一知道的提供给实际使用者这种灵活性的装置涉及使用与说话者有关的技术来形成新的字模型，这些字模型于是被存贮在一个独立的词典中。使用者必须一次或多次地讲每一个字以训练系统。这些与说话人有关的模型通常通过使用 DTW 或其它相似的要求相对大量内存以存贮每个使用者模板的技术来形成。典型地，每个使用者的每个字将至少占用 125 个字节（可能超过 2K 字节）。这意味着对于一个 20 字的字表，在识别开始之前，在 2.5 到 40K 之间的字节必须下载到识别器中。进一步来说，一个仅为 1000 个使用者服务提供的电话网将需要25 兆到20 兆字节的磁盘存贮空间仅仅用于使用者的模板存贮，这种服务的一个例子是指令拨号器，在这种服务中使用者定义他想要呼叫的人，使得之后通过说出想要接收者的名字，就能拨打电话。

根据该发明，一种为识别装置产生字表的方法包括接收代表发音的输入语音信号；由每一个发音产生一个代码，该代码从大量参考子字表示中识别一个参考子字表示的序列，这个序列与该发音最类似；为下一步识别存贮产生的发音的代码。

这样一种方法允许使用者选用新字而不需为每一个字。产生一个字的声学模型，允许每一个字或词组被模拟为对使用者来说唯一的一序列参考子字表达式。这并不需要关于要被加入字表的字的以前的知识，因而允许使用者加入任何需要的字或词。

被使用者选择的字的代码可能比由文本形成的模型更相似于使用

者的口呼语音。此外，代码需要的内存容量至少比象 DTW 方法中存贮字表达式要低一个数量级（尽管在准确性上会有轻微的损失）。

更好的是代码的产生不受语法规则的约束，也就是说，任何子字表达式之后可跟随任何其它一个子字表达式。另一方法，可以使用二元语法，它在每一对子字，例如音素之间强加上转移概率。这样，一对在给定语言中通常不会出现的音素（例如英语中的 pH ）有一个低的转移概率。

多于一个的代表同样发音的语音信号的代码可能会被产生。代码中的任何异常情况将会被考虑。例如，如果发音是从有噪音的电话线上得到的，该发音的代码与从清晰电话线上得到的同一发音的代码可能几乎没有相似性。比较适合的是接收一个发音的三个训练输入并且抛弃那个与其它代码相差很大的那个代码。另一种方法是保留所有的代码。是否所有的代码都被存贮取决于该装置的开发者。

根据该发明的第二方面，字表产生装置包括从输入语音信号中提取特征样本的提取装置；一个子字识别器，用来从输入语音信号的每一个样本中产生一个代码，该代码从多个参考子字表达式中确定一个参考子字表达式序列，这个序列的子字表达式与输入语音信号最为接近；一个为了下一步识别而存贮输入语音信号的代码存贮器。

该装置旨在与一个被形成来识别由代码表征的发音的语音识别器有关。在识别中，语音识别器比较未知输入语音信号和由存贮在存贮器中的代码表征的子字表达式序列，并且输出一个表征识别与否的信号。

更好的是子字识别器的语法被松散的约束。例如，子字识别器可以，必然说，被约束以识别被线路噪声约束的任何子字单元序列。另一种方法，可以使用二元语法，它在每一对音素之间强加上转移概率。

语音识别装置可被形成以识别一些预先规定的字。最好的是，预先规定字也被存贮为预先规定字的子字描述的代码。预先规定字和使用者选择的字于是用相同的参考子字来模拟，语音识别器可以被形成来识别与使用者选择的字一起说出的预先规定的字。

最好的是参考子字表达式表征音素。每一个子字表达式可能是一个包含该特殊子字的多个说话者输入语音的统计模型。尽管其它模型可能会被使用，这些模型最好是隐马尔可夫模型。

该发明现在将仅通过例子进一步被描述，并参考附图。其中，

图 1 示意性地表示出根据该发明在无线电通信环境中，语音识别装置的使用。

图 2 是一个方框图，根据该发明，示意性地表示出字表生成器的功能块。

图 3 表示了一个象用于图 2 中字表生成器中一样的松散约束网络的例子。

图 4 表示了一个使用图 2 中的字表生成器的语音识别器。

图 5 表示一个如同使用了图 4 中语音识别器的识别网络的例子。

图 6 表示一个图 5 中示出的替换识别网络。

图 7 表示根据该发明语音识别装置的第二个具体实施方案。

参考图 1，一个包括语音识别的电信系统通常包括一个麦克风 1（一般地形成一个电话听筒的一部分）；一个电信网络 2（一般地为一个公共交换电信网（PSTN））；一个语音识别器 3，被连接来接收从网络 2 来的语音信号；以及一个应用装置 4 与语音识别器 3 相连，并被安排从中接收一个表示对一个特定字或词组识别或其它的声音识别信号，而且做出响应动作。例如，应用装置 4 可能是一个远程操作的指令拨号系统，在该系统中，使用者并不拨打想要的号码，而是简单的说出想要与之通话的人的名字。

很多情况下，应用装置 4 将产生一个使用者可听到的响应，通过网络 2 传到一个扬声器 5，扬声器 5 一般形成使用者电话听筒的一部分。

操作时，使用者向麦克风 1 喊话，信号从麦克风 1 传进网络 2，传到语音识别器 3。语音识别器分析该语音信号，并且一个表征对一特殊的字或词组识别与否的信号被产生并传给应用装置 4，然后应用装置 4 在识别该语音的情况下采取适当的行动。

当一个使用者头一次使用应用装置 4 提供的服务时，语音识别器 3 需要得到涉及字表的数据，并对照该字表来检验后序的未知的语音信号。数据获得由在训练模式下运行的字表产生器 9 来完成，在训练模式下的运行中，使用者提供训练输入语音样本，为了后续识别目的从中产生训练输入语音的子字内容的代码。

根据该发明，图 2 表示了字表产生器 9 的功能块。字表生成器 9 包括一个特征提取器 6 来从输入语音信号中提取特征数据，该输入语音信号已经被划分成邻接样本的一系列的帧。通常，帧表示一个 16 毫秒的

输入语音样本，每个样本被加窗（例如用汉明窗）。适合的特征提取器的例子在技术上是众所周知的，并且可能包括某种形式的谱分析技术，例如，滤波器组分析器，线性预测编码分析和离散变换分析。

该特征可能，比如包括倒谱系数（例如，象发表于 1982proc IEEEp 2026 中， Chollet 和 Gagnoulet 著的“使用参考系统的语音识别器及数据库的评价”中描述的 LPC 倒谱系数或 mel 频率倒谱系数）。或者象发表于 1988 IEEE Trans 声学、语音与信号处理卷 36 第 6 号 871 页，Soong 和 Rosenberg 著的“在说话者识别中瞬时和转移谱信息的使用”中描述的这些系数的微分值，对每个系数来说，该值包括系数之间的差值以及相应前面矢量中系数值。同样的，有可能使用几种特征系数的混和。特征提取器由一个适当编程的数字信号处理器（ DSP ）设备提供。特征提取器 6 的输出数据组形成子字识别器 7 的输入。

子字识别器 7 是涉及具有表示表 1 中给出的 40 个音素的 HMM 模型的子字模型存贮器 8 。对大量子字中的每一个模型，存贮器 8 包括域 81 、 82 …… 。例如，子字识别器被设计为识别音素、并相应地为每一个音素在模型存贮器中提供一个域。

子字识别器 7 被安排为顺序读取存贮器 8 中的每个域，并用当前输入特征系数组对每个域计算其输入特征组与相应域符合的概率。一个表征最可能的子字模型的信号被输出并存贮在字存贮器 10 中。因此，对于一个单个发音来说，字存贮器 10 存贮一个表征子字识别器认为最接近地代表输入语音的参考子字模型序列的代码。

上述的计算使用了如同在 1988 年 4 月的英国电通技术期刊第 6 卷第 2 号中发表的 S J Cox 著的“自动语音识别中的隐马尔可夫模型：理论与应用”中讨论的著名的 HMM，为方便起见，子字识别器 7 中执行的 HMM 处理使用著名的 Viterbi 算法。子字识别器 7 可能，比如说，是一个诸如 Intel™ i-486™微处理器或 Motorola™68000 微处理器这样的一个微处理器，或者另一种选择为一个 DSP 设备（例如，象特征提取器 6 一样的 DSP 设备）。

象前面描述的与子字识别器有关的子字模型以一种与说话人无关的方式获得。因此，子字识别器 7 产生的代码只在它们表示一个给定使用者如何发音一个字的音素记录这种程度上是与说话人有关的。

子字识别器 7 有一个识别网络，该识别网络对也许会产生的可能子

字单元序列施加很少或不施加约束。图 3 中示出一个松散约束网络的例子。该网络允许对一个单一连接的受噪声限制的音素序列识别。该音素序列是完全不受限制的，并且因而在操作的语言中（在描述的例子中为英语）不发生的音素序列可能会被产生。

图3中表示的识别网络目前提供给电话语音的记录结果要比一个完全不受约束的网络，即在音素模型前后没有噪声模型的网络要好。它不允许被噪声跟随的音素跟随有音素。对于一个实际应用系统这点的重要性是对于分离的字或连接的词组来说，它将增强系统的正确性，但是如果使用者输入的词组中有的字之间有空隙，就将会有问题。例如，在一个指令拨号器中，如果使用者说“ John Smith ”时在名和姓之间没有空隙，这种语法将不会造成任何问题。另一方面，如果他们确实在它们之间留下一个空隙，性能就会受损失。然而，子字识别器的识别网络将被设计以符合系统的要求，比如说，分离的字，连接的字等等。

在第一次使用该服务时，应用装置提示使用者提供他希望加入识别器字表中的字。使用者对麦克风说一个选择的字作为对应用装置的可听到的提示的响应。在一个指令拨号系统中，这个字可能是使用者想要呼叫的人的名字，例如“ Janc ”。字表生成器从输入提取特征，该特征被传给子字识别器 7。当输入语音被接收时，它被依照存贮器 8 中的模型来匹配。具有一个图 3 中表示的识别网络的子字识别器 7 产生一个口呼输入的代码。该代码确定与输入语音最相似的模型序列。于是输入语音的一个音素的记录被产生。产生的训练发音的代码被存贮在存贮器 10 中。然后使用者被提示重复输入以形成一个输入语音的更健全的表示。

从实验发现，当只提供一个训练发音时达到的正确率为 87.8 %，然而当提供 3 个训练发音时，正确率显著提高到 93.7 %。明显地，低质量的电话线对于产生的结果将有很重要的影响。提供 3 个训练发音时所达到的正确率也高于替代子字表示的从文本输入中得到理想接收发音记录时的正确率。接收到的发音是标准南部英国英语口音。

一个进一步的提示被给予使用者，询问是否有更多的字打算加入。如果使用者给出肯定回答（例如使用预先制定的 DTMF 键），就为下一个字重复识别过程。如果使用者给出否定回答，系统就切换到识别模式，也就是说，语音识别器 3 开始工作。存贮器 10 中存贮着为每一个附加字表条目识别一个序列的参考子字表示的代码。

一旦字表中需要的每一个字的表示都被产生，该字表就可被语音识别器 3 使用。图 4 表示了识别器 3 的组成部分。语音识别器 3 包括一个特征提取器 6，一个子字模型存贮器 8 及一个字表生成器 9 产生的代码的存贮器 10。一个网络生成器 12 涉及存贮器 10 并形成一个由被代码表示的参考子字表示序列形成的识别网络。这样一个网络可以通过，例如，将存贮器 10 中的单个的代码组合成象图 5 中示出的网络并行可替换物来产生。或者将代码组成树型结构，如图 6 中表示的，两者都表示了在字“Six”和“Seven”的一个发音中被识别的一个音素序列的例子。

在识别过程中，输入语音信号被传给特征提取装置 6，特征值被传递给涉及由网络生成器 12 形成的网络的识别器 16。未知输入语音当网络形成后与网络比较，如果在未知语音输入与网络的一个分枝之间发现一个接近的匹配，就从识别器 16 输出一个信号，然后是用代码表示的一个字或词组。一旦识别已经发生，应用装置 4 就根据该服务执行下一适当的步骤例如，该服务为一个指令拨号服务，并且识别器 16 认为字“Jane”已被识别，应用装置就拨叫涉及名字“Jane”的号码。

图 7 举例说明该发明的第二个实施方案。图 2 和 3 显示字表生成器 9 及语音识别器 3 为分离的部件，而图 7 显示它们组合在语音识别装置 20 中。字表生成器 9 与识别器 16 共享公共的部件，也就是说，特征提取器 6，子字模型存贮器 8 及使用者选择字的存贮器 10。语音识别装置 20 另外包括一个预定字存贮器 14，它存贮适合所欲应用装置的预先制定字的音素记录的预先制定代码。例如，对于一个指令拨号系统，这些预先制定的字可能是数字 0 到 9“dial”，“no”，“yes”，“add”等等。

语音识别装置 20 通常处于识别模式，也就是说，输入语音信号被传给识别器 16。当使用者想向系统字表中加字时，使用者说字“add”。该信号被传给特征提取器 6，特征值被传给识别器 16。网络生成器 12 产生一个包括所有表示在存贮器 14 和 10（一开始，可能存贮器 10 中设有任何字）中的字的网络。识别器 16 匹配将输入与网络匹配并识别出输入信号为字“add”，并以将输入接续（输送）到字表生成器 9 的方式进入训练模式做为响应。

使用者于是象前一个实施方案中那样通过说出要加入系统字表中

的名字进行下面操作。字表生成器 9 的子字识别器 7 产生代码以存贮在使用者选择存贮器 10。然而，使用者可以通过说“Yes”或“No”以口呼的方式对应用装置的提示做出响应。如果一个特定的响应是预期的，输入语音信号被输送给识别器 16。

一旦使用者已经选择了想要的字，由网络生成器 12 产生的后序网络组合存贮器 14 中的预定的字及存贮器 10 中的用户选择字。结果识别器有一种语法，其中一些字由从使用者语音提取的音素序列定义，一些字由从另一个来源提取的序列预先制定。这两个存贮器中的字可以被组合以使得，例如，如果字“dial”被预先制定，识别网络可以被形成来组合“dial”和每一个选择的字，使得系统语法允许连接语音“dial Jane”，“dial”是预先制定的，“Jane”是使用者选择的。

当只提供一个训练发音时，字表 50 %预先制定的语音识别装置与字表完全由使用者选择识别装置具有相同的正确性，然而当提供三个训练发音时，这种装置的正确性就远比字表完全由用户选择的装置差。

因而，在该发明的进一步实施方案中，语音识别器 20 有一定数目的预先制定的字存在预先制定存贮器 14，并有一定数目的在训练模式下定义的使用者选择字存在存贮器 10 中。在该装置的使用中，通过将特征值从特征提取装置 6 传给子字识别器 7 及识别器 16，从使用者的输入语音中产生预先制定字的子字表示。子字识别器为该发音产生的代码被加入到存贮器 10。后序发音与存贮器 10 中的存贮的表示的匹配应该比与存贮器 14 中的存贮表示更接近，这导致预先定义字识别的正确性的提高。

说 明 书 附 图

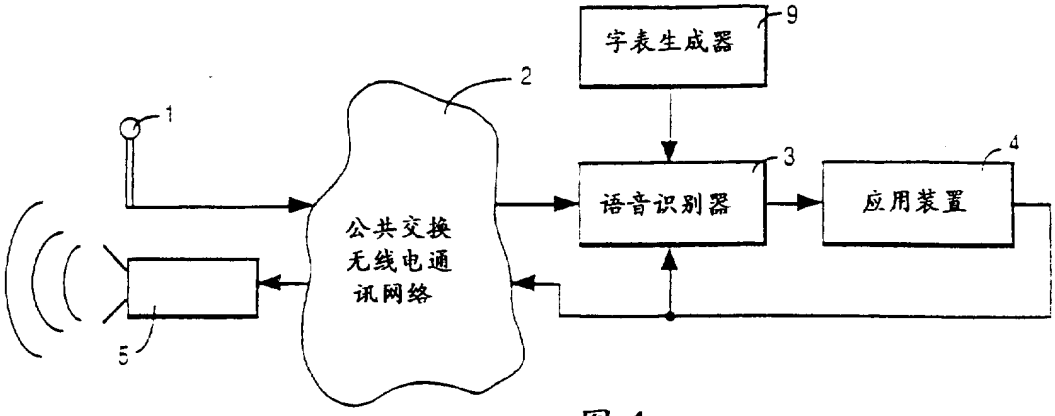


图 1

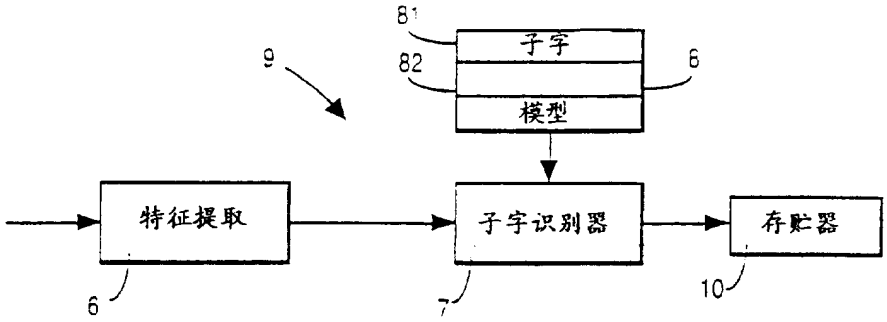


图 2

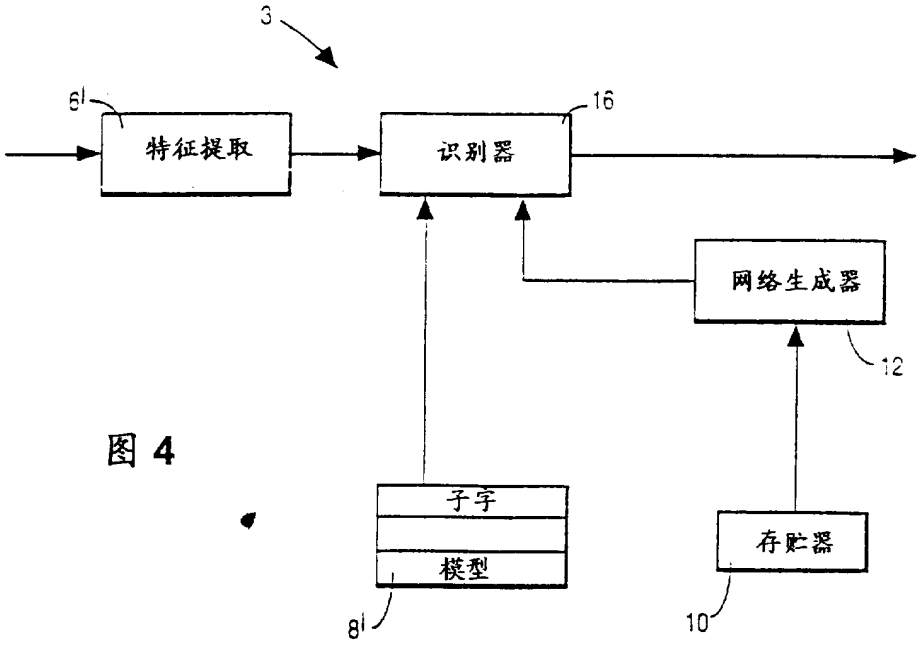


图 4

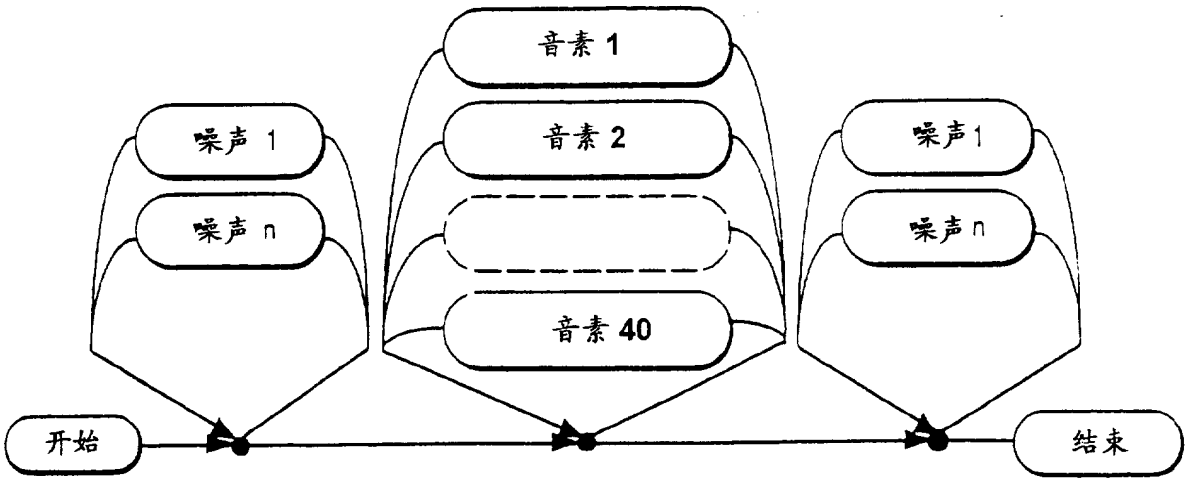

音素 1
音素 2
音素 40
噪声 1
噪声 n
噪声1
噪声n
开始
结束

图 3

图 5

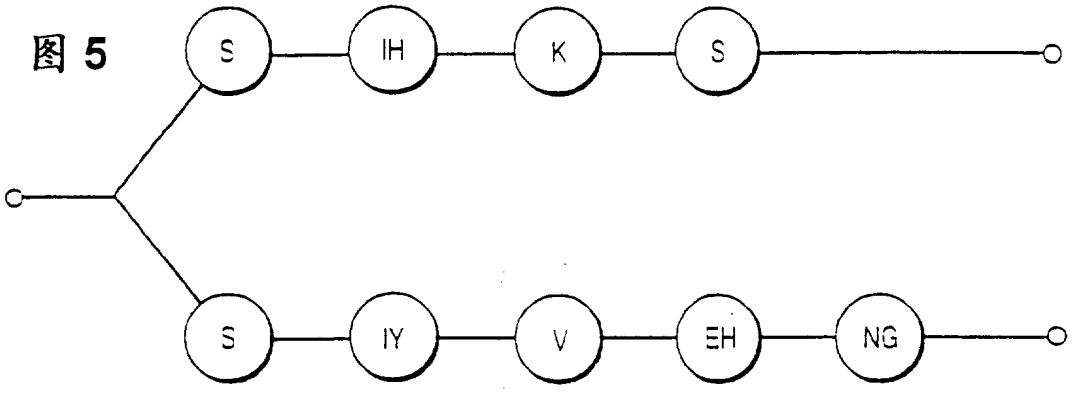
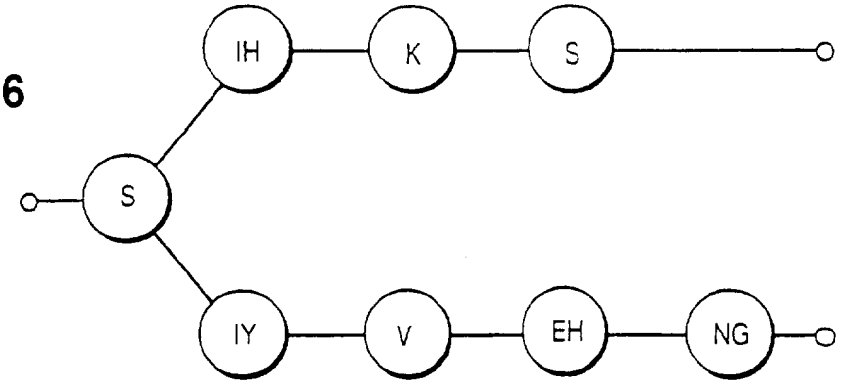

S
IH
K
S
S
IY
V
EH
NG

图 6


IH
K
S
S
IY
V
EH
NG

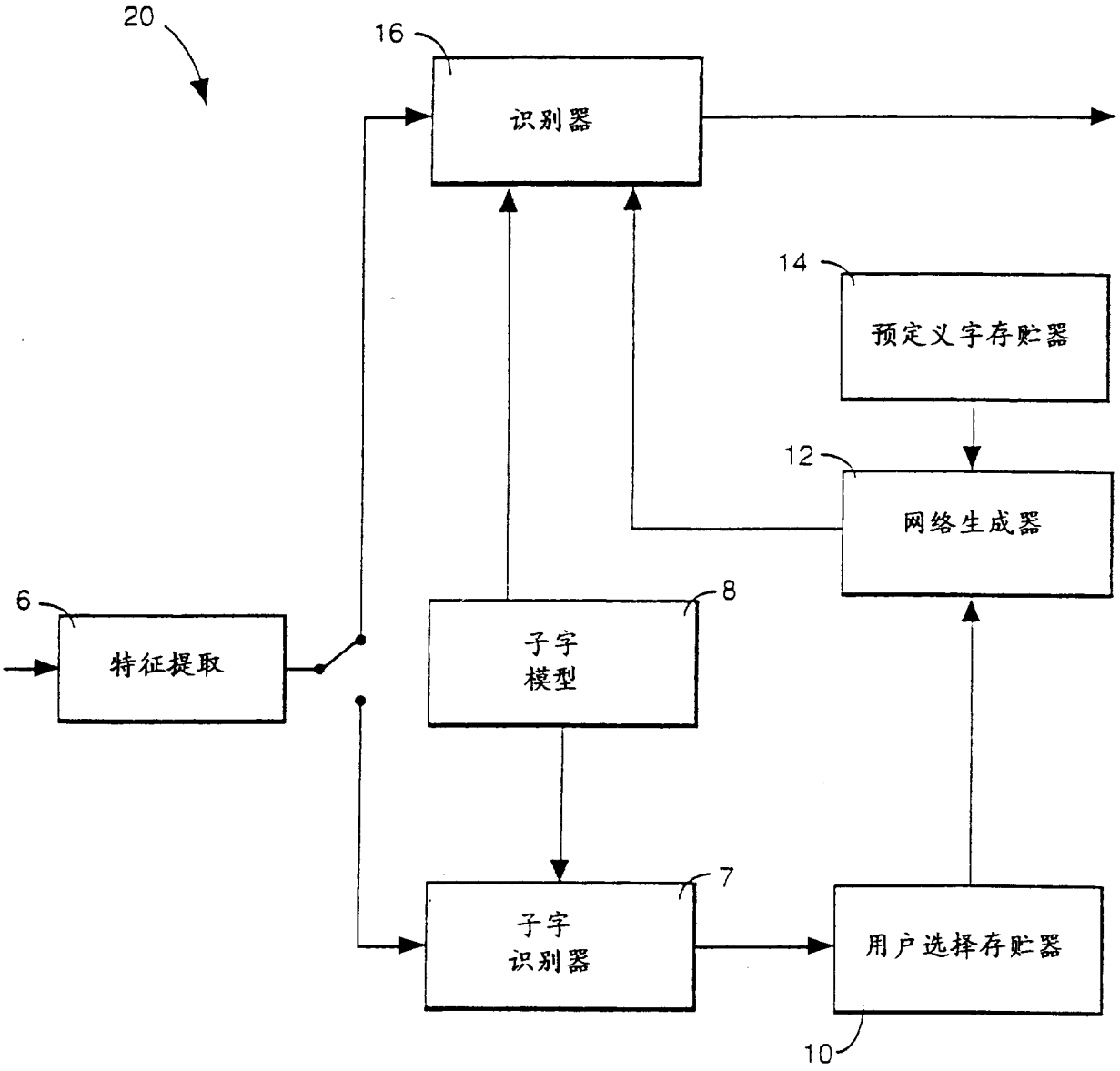


图 7