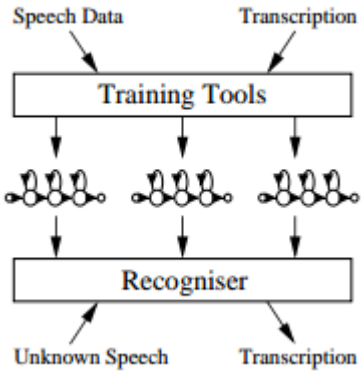
HTK基础



HTK是一个用于搭建HMM模型的工具箱。HMM可以用来对任何时间序列进行建模，所以HTK的核心功能也和这个类似。不过HTK本质上还是作为一个基于HMM的语音处理工具来设计的，特别是语音识别，因此，HTK的许多基础支持（infrastructure support）都是针对识别这个任务而设计的。由上图可知，一个识别器涉及到两个主要处理阶段，首先HTK训练工具使用训练的语音（utterances）和对应的文本（transcriptions）来估计**多个HMMs**的参数，然后一段未知的语音将会被HTK识别工具来转换成文本（transcribed）。

这本书的主要内容是关于这两个处理方法的。但是，在深入讲解之前，有必要对HMM的基本原理有一定程度的了解。

这本书第一部分的第一章将会介绍HMM的基本原理以及在语音识别里面的应用。第二章会对HTK工具有一个简单的概览，对于老版本的使用者，在第二章里面会标明主要的不同点。之后的第三章描述了如何使用HTK工具来搭建一个基于HMM的，简单的小规模词汇的连续语音的识别器。

这本书的第二部分将会重新回到并深入第一部分略过的主题上。第二部分可以和第三部分还有第四部分结合起来阅读。第四部分提供了一个关于HTK的参考手册，包含了对于每个工具的描述，对于用来配置HTK的各种参数的总结以及一个关于错误信息的列表。

1.1 HMMs的基本原理

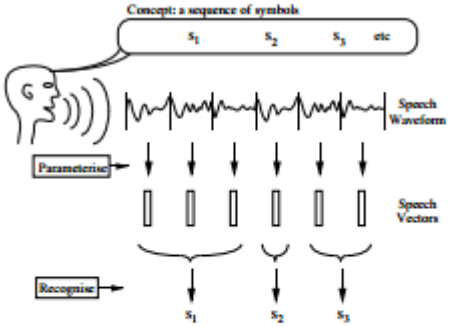


Fig. 1.1 Message Encoding/Decoding

语音识别系统通常假设语音信号是一个由一个或者多个符号组成的序列经过编码之后实例化得到的（如图1.1）。为了实现上述操作的逆操作，也就是在给定语音的情况下识别出隐藏的符号序列，连续的语音波形首先会被转换称一个等间隔的离散参数的向量序列（例如MFCC）。这个参数向量的序列被假设为一段语音波形的精确表达，因为每一个单个的向量只对应了语音信号里面的很短一个片段（一般是10ms），在这么短的时间里面，语音可以被认为是稳定的。虽然这不是事实，但是是一个很合理的近似。常用的一些典型的参数表示包括平滑化的频谱或者是线性预测系数加上各种其他表示。

识别器的扮演的角色是实现一个语音向量序列到隐藏符号序列的一个映射。这有两个很难的问题。首先，从符号到语音的映射并不是一一对应的，这是因为不同过得隐藏符号可以给出相似的发音。另外，相同的符号的发音也会因为说话人的多样性，感情和环境等存在很多变化。第二，符号和符号之间的边界并不能从语音波形显示的辨认出来。因此，把语音波形当做一个拼接起来的静态模式是不行的。

第二个关于边界的问题，可以通过限制任务为孤立词识别来避免。如图1.2所示，这里表示这个语音波形对应这单个从固定词汇表中选出来的隐藏符号（例如词(word)）。先不管这个简单的问题在某种程度上是人造的这一事实，它还是有着很广泛的应用。另外，在处理更为复杂的连续语音识别的情况之前，这个问题可以用来介绍基于HMM的识别原理。因此，接下来首先讲使用HMMs的孤立词识别。

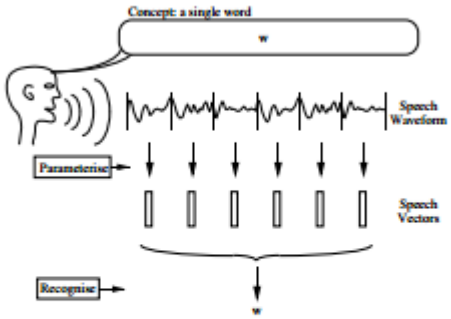


Fig. 1.2 Isolated Word Problem

1.2 孤立词识别

让每一个口语词都被表示成一个语音向量的序列或者是观测值O, O的定义如下：

$$O = o_1, o_2, \ldots, o_T \tag{1}$$

这里o_t表示在时间t观测到的语音向量。孤立词识别的问题可以被看成计算下面的表达式：

$$arg\max_i P(\omega_i|O) \tag{2}$$

这里ω_i表示第i个单词。这个概率不能直接计算，但是通过贝叶斯共识可以得到：

$$P(\omega_i|O) = \frac{P(O|\omega_i)P(\omega_i)}{P(O)} \tag{3}$$

因此，对于先验概率$P(\omega_i)$的一个给定集合，最可能的口语词只取决于似然度$P(O|\omega_i)$。给定观测序列的长度，由一些口语词的样本来直接计算联合概率密度$P(o_1, o_2, \ldots|\omega_i)$是不可行的。然而，如果先假设一个词语发音的参数模型例如马尔科夫模型，然后再用数据进行估计是可行的，因为估计条件概率密度$P(O|\omega_i)$这个问题被估计马尔科夫模型参数这个简单的多的问题取代了。

在基于HMM的语音识别里面，会假设可观测到的每个词语的语音向量序列是由对应的马尔科夫模型产生的，如图1.3所示。一个马尔科夫模型是一个有限状态机，会在每个时间点上改变一次状态， 并且在每个时间点t进入状态j之后，一个语音观测向量o_t将会根据概率密度$b_j(o_t)$产生。还要，从状态i转义到状态j同样是概率化的，这个离散概率我们记为a_ij。图1.3显示了一个关于这个过程的例子，6状态的HMM模型，为了生成观测序列o_1到o_6，其状态转移序列为$X = 1, 2, 2, 3, 4, 4, 5, 6$。注意，在HTK里面，HMM的进入和退出状态是没有发射概率的。这样对于复合模型的搭建来说是很有帮助的。关于复合模型将会在后面提到。

观测序列O由模型M经过状态序列X产生这样的一个联合概率分布可以被简单的计算为状态转移概率和输出概率的乘积。因此对于图1.3中的状态序列X，有

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)\ldots \tag{4}$$

然而，在实际中，只有观测序列是已知的，状态序列是隐藏的。这也是为啥会被成为隐马尔科夫模型的原因。

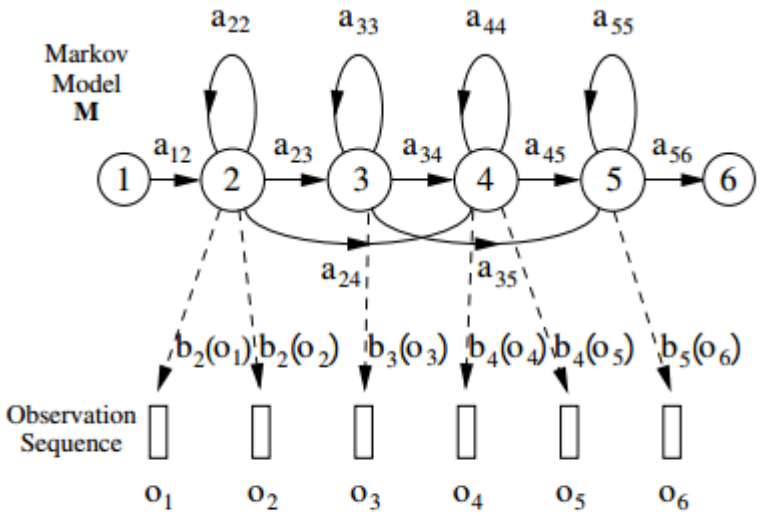


Fig. 1.3 The Markov Generation Model

在X未知的条件下，我们想要的似然度可以通过对所有可能的状态序列$X = x(1), x(2), x(3)\ldots x(T)$求和得到，表达式如下：

$$P(O|M) = \sum_X a_x(0)x(1)\prod_{t=1}^{T} b_{x(t)}(o_t)a_{x(t)x(t+1)} \tag{5}$$

这里$x(0)$被限制为模型的进入状态，$x(T+1)$被限制为模型的退出状态。

作为公式（5）的一种近似，这个似然度可以只考虑最有可能的状态序列，表达式如下：

$$\hat{P}(O|M) = \max_{X} a_x(0)x(1)\prod_{t=1}^{T} b_{x(t)}(o_t)a_{x(t)x(t+1)} \tag{6}$$

1.3 输出概率详述

1.4 Baum-Welch估计

1.5 识别和Viterbi解码

1.6 连续语音识别

1.7 说话人自适应