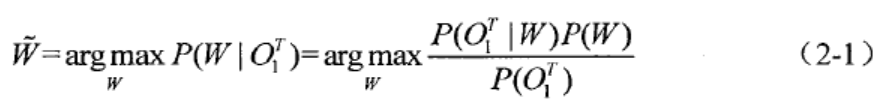
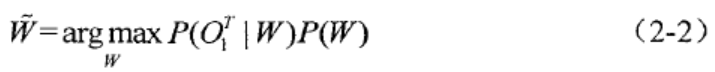
本周我阅读了**《automatic speech recognition a deep learning approach》，阅读并翻译之后有一点自己对算法和公式的理解**

文章开头介绍了语音识别发展概述，语音识别的发展与现状，和深度学习对语音识别的意义

第二部分主要讲了语音识别相关理论研究，语音识别主要作用就是把一段语音信号转换成相对应的文本信息，系统主要由声学特征提取、语言模型、声学模型和解码器等组成。训练识别的过程是从原始波形语音数据中提取的声学特征经过训练得到声学模型，与发声词典、语言模型组成网络，对新来的语音提取特征，经过声学模型表示，通过维特比解码得出

识别结果。大词汇量连续语音识别系统都是在隐马尔科夫模型（HMM）为框架进行统计训练，在给定语音特征序列后，结合声学模型和语言模型，根据最大后验概率算法输出词序列中，其数学表示如下式所示

上式中，是语言模型，代表特定词序列『出现的先验概率；为声学模型，表示给定词序列为W时，输出的声学特征时的概率；是声学特征的概率，与词序列W的无关，所以可以忽略掉。因而，式2-1可变为：

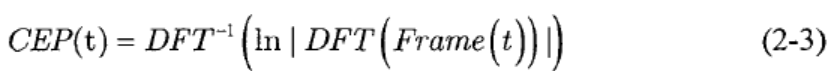
声学特征的提取可以认为是对原始波形信号压缩的过程，也可以认为是对信

号进行解卷积的过程，这样做是为了达到最好的分类效果。语音信号在短时内随

时间变化很缓慢，在范围内信号特征保持相对稳定，这样对语音信号

处理应可以通过短时分析。语音识别中的特征包括：线性预测参数（LPC），倒谱系数（CEP）、梅尔频率倒谱系数（MFCC）和感知线性预测系数（PLP）

1.线性预测即某一时刻语音信号可以用之前若干时刻的信号来线性表示。其基本问题就是由语音信号直接求出一组线性预测系数，使得在一短段语音波形中均方预测误差最小。

2.倒谱系数是重要的语音特征参数，是以同态处理方法为基础实现的。处理公式如下：

这里用表示第t帧语音信号，分别用表示离散傅里叶变换（DFT）和反傅里叶变换（IDFT）通过上式可以看出，由于进行了同态分析，只需要前几阶系数就可以包含语音信号的很大一部分信息，这样可以有效的压缩数据。

3. 与线性预测和倒谱特征不同，梅尔倒谱系数和感知线性预测主要基于人耳的听觉系统的研究成果。提取特征过程为，首先通过计算将时域信号转换到频域上，然后对它的对数能量谱用刻度分布的幅度相同的三角滤波器组进行卷积运算，之后再对滤波器输出进行离散余弦变换（DCT）取前几个作为最后的特征参数。

对语音波形信号提取特征之后，为了增强鲁棒性，或者降低特征的维度，需要对原始特征做归一化处理或者用一些线性变换的方法。常见的特征处理的方法有倒谱均值归一化（ＣＭＮ）、说话人声道长度规整（ＶＴＬＮ）、ＲＡＳＴＡ滤波、线性判别分析（ＬＤＡ）、最大似然线性转换（ＭＬＬＴ），异方差线性判别分析（HLDA）等。

接下来文章又对声学模型的搭建进行了介绍

声学模型在语音识别系统中起着至关重要的作用，描述了声学基元产生特征

序列的变换过程。给定一个声学特征矢量，依据声学模型来计算它属于每个基元

的概率值，通过最大似然准则得出与特征序列对应的状态序列。

声学基元选择

声学基元的选择是声学建模中很关键的问题，合适粒度的基元对系统性能的

提升有很大帮助。语音识别把词、音节、声韵母以及音素（等作为常用基元。基元的选择和设计通常考虑的是语音学知识，基元也可基于数据驱动的方法产生，虽然有些在语音学上意义不明确，但是有的效果也不错。在语音识别中为了应对协同发音现象从而能更准确的描述语音，通常选用上下文相关的方法来对语音建模，这其中有两种情况，当只考虑前一音子对当前音子的影响被称为，同时考虑前一音子和后一音子的影响被称为。对英语识别中通常采用上下文相关的三音素作为基元，对于汉语建模往往采用音节或者上下文相关的声韵母作为基元。

我在其他文献中也了解到，HMM隐马尔科夫模型是大部分语音识别系统声学建模时常采用到的模型，它用来描述语音内在隐含状态和时间序列的转换关系。是一种使用参数来描述随机过程统计特性状态的概率模型，它对动态时间序列的建模能力很突出。主要分为两部分，具体为有固定状态数的隐式马尔科夫链模型还有显式的随机函数集。这中间每个随机函数值都与链中一个状态有关联，这样就可以通过显式过程产生的观察序列来获取隐式过程的相关信息。

之后我有百度了一下关于HMM的具体讲解和介绍。

马尔科夫链（Markov Chains），表示一个状态到另一个状态转换的随机过程。该过程要求具备“无记忆”的性质：下一状态的概率分布只能由当前状态决定，在时间序列中它前面的事件均与之无关。这种特定类型的“无记忆性”称作**马尔可夫性质**。   
假设状态空间*qt*∈*s*(*j*),*j*=1,2,...,*N*，马尔科夫性质可以表示如下：

*P*(*qt*+1=*s*(*t*+1)|*q*0=*s*0,...,*qt*=*st*)=*P*(*qt*+1=*s*(*t*+1)|*qt*=*st*)

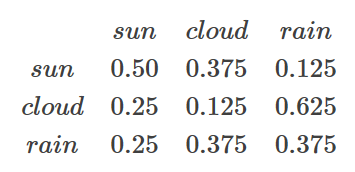
转移概率可以定义如下：

*P*(*qt*=*sj*|*qt*−1=*si*)=*aij*(*t*),*i*,*j*=1,2,...,*N*

如果转移概率跟t无关，只由i和j决定，那么说马尔科夫链是齐次（homogeneous）的，转移概率矩阵可以写成一下形式：

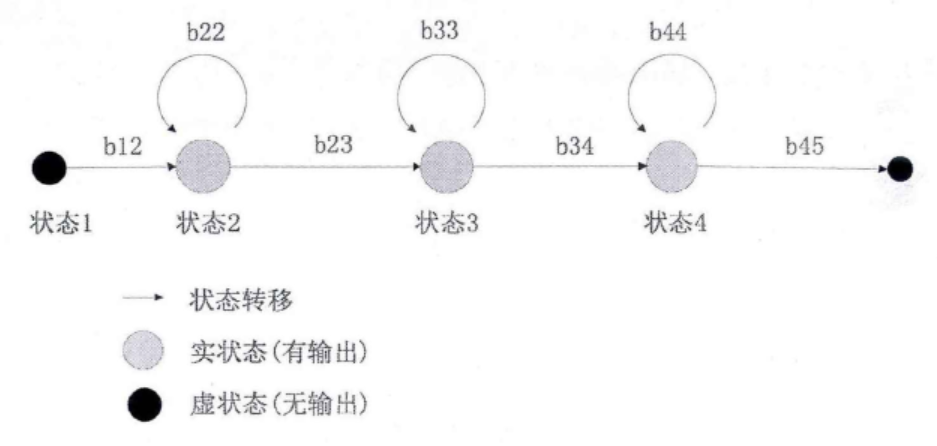
*A*=[*aij*],*whereaij*≥0,∀*i*,*j*;*and*∑*j*=1*Naij*=1,∀*i*

马尔科夫链关注的状态转移概率（Hidden Markov Sequences）。比如天气的三个状态分别是{sun, cloud, rain}，状态转移概率矩阵如下



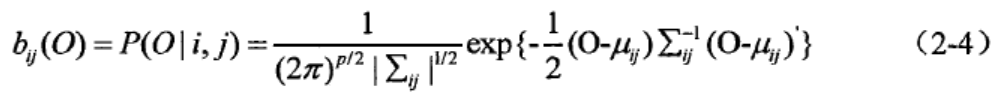
马尔科夫链求解的问题是：如果今天晴天(sun)，后天下雨(rain)的概率。   
马尔科夫链的观察序列和状态序列是一一对应的，具有确定性（比如今天是sun，那么sun的概率将会是1，不可能为cloud或者rain），从而导致马尔科夫链很难去拟合有些现实问题。   
为了增加马尔科夫链的随机性，对每一个状态，增加一个观察值的生成概率函数，即为隐马尔科夫序列。使用隐马尔科夫序列去建模解决实际问题，称为Hidden Markov model（HMM）。

HMM可以用以下五个参数来进行描述，即在上式中，S表示模型所包含的有限隐含状态的集合；O表示输出的可观测序列的集合；A表示状态之间的转移概率的集合；B表示给定状态的下输出相应输出观察值的概率；π表示系统初始状态概率的集合。用作声学模型时，其结构如下图所示，其中表示状态到状态的转移概率。HMM用作声学模型时，其结构如下图所示，其中表示状态i到状态j的转移概率。



语音信号特征分布情况用简单高斯概率密度函数难以描述，实际应用中常采

用混合高斯模型来拟合语音信号，主要通过混合高斯函数来表示输出概率B。即：



上式中，O表示输出状态；i和j分别对应相邻两个要转换的状态；p表示

混合数；分别表示从状态i转换到输出状态j的均值和方差。

接下来我又了解了声学模型的训练准则

最大似然准则（MLE）仍是目前常用的声学模型训练准则，它只釆用与模型有关的语料训练，需要有一定的假设条件，即训练数据足够多且服从一定分布。而且ＭＬＥ是通过单一类的类内训练数据去学习，没有考虑类间的影响。相比于ＭＬＥ这种生成性准则，区分性训练的目标不是最大相似度，而是最小分类错误。由于区分性准则在声学模型训练中突显出的种种优势，近年来得到了更多的关注。语音识别中的区分性训练准则主要包括：最大互信息估计（ＭＭＩＥ），最小分类错误（ＭＣＥ）、最小音素错误（ＭＰＥ）和ｆＭＰＥ等。上述这些准则在一定程度上提高了系统的识别率，深度学习的方法在参数调整的过程其实也是一种区分性学习，其中深度网络训练中也可以用准则取代交叉熵来调整网络。