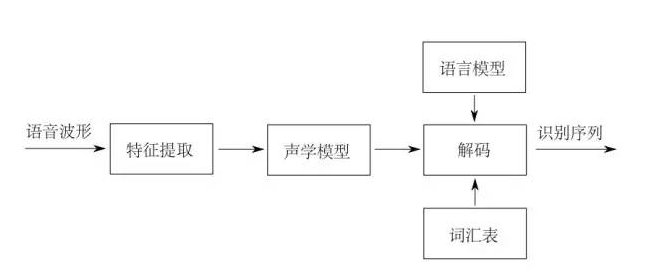
语音识别

<https://www.sohu.com/a/160550466_197042>



所谓语音识别，就是将一段语音信号转换成相对应的文本信息，系统主要包含特征提取、声学模型，语言模型以及字典与解码四大部分，其中为了更有效地提取特征往往还需要对所采集到的声音信号进行滤波、分帧等预处理工作，把要分析的信号从原始信号中提取出来；之后，特征提取工作将声音信号从时域转换到频域，为声学模型提供合适的特征向量；声学模型中再根据声学特性计算每一个特征向量在声学特征上的得分；而语言模型则根据语言学相关的理论，计算该声音信号对应可能词组序列的概率；最后根据已有的字典，对词组序列进行解码，得到最后可能的文本表示。

1. ****声学信号预处理****

在最终进行模板匹配的时候，是将输入语音信号的特征参数同模板库中的特征参数进行对比。

声音信号进行滤波与采样：为了排除非人体发声以外频率的信号与50Hz电流频率的干扰，该过程一般是用一个带通滤波器、设定上下戒指频率进行滤波，再将原有离散信号进行量化处理实现的，之后需要平滑信号的高频与低频部分的衔接段，从而可以在同一信噪比条件下对频谱进行求解，使得分析更为方便快捷；

分帧加窗操作：为了将原有频域随时间变化的信号具有短时平稳特性，即将连续的信号用不同长度的采集窗口分成一个个独立的频域稳定的部分以便于分析，此过程主要是采用预加重技术；

端点检测工作：也就是对输入语音信号的起止点进行正确判断，这主要是通过短时能量（同一帧内信号变化的幅度）与短时平均过零率（同一帧内采样信号经过零的次数）来进行大致的判定。

1. ****声学特征提取****

能用于语音识别的特征参数必须满足以下几点:特征参数能够尽量描述语音的根本特征；尽量降低参数分量之间的耦合，对数据进行压缩；

目前主流研究机构最常用到的特征参数有：线性预测倒谱系数（LPCC）和 Mel 倒谱系数（MFCC）。两种特征参数在倒谱域上对语音信号进行操作，前者以发声模型作为出发点，利用 LPC 技术求倒谱系数。后者则模拟听觉模型，把语音经过滤波器组模型的输出做为声学特征，然后利用离散傅里叶变换（DFT）进行变换。

所谓基音周期，是指声带振动频率（基频）的振动周期，因其能够有效表征语音信号特征。

所谓共振峰，是指语音信号中能量集中的区域，因其表征了声道的物理特征。

****3. 声学模型****

声学模型是语音识别系统中非常重要的一个组件，对不同基本单元的区分能力直接关系到识别结果的好坏。语音识别本质上一个模式识别的过程，而模式识别的核心是分类器和分类决策的问题。

孤立词、中小词汇量识别中使用动态时间规整（DTW）分类器会有良好的识别效果，并且识别速度快，系统开销小，是语音识别中很成功的匹配算法；在大词汇量、非特定人语音识别的时候，DTW 识别效果就会急剧下降，这时候使用隐马尔科夫模型（HMM）；深度神经网络来完成声学建模，形成所谓的DNN-HMM构架来取代传统的GMM-HMM构架。

使用高斯混合模型对短时特征向量建模有以下几个好处：首先，高斯混合模型的具有很强的建模能力，只要分量总数足够多，高斯混合模型就可以以任意精度来逼近一个概率分布函数；另外，使用 EM 算法可以很容易地使模型在训练数据上收敛。对于计算速度和过拟合等问题，除了使用 EM 算法作最大似然估计以外，我们还可以使用和词或音素错误率直接相关的区分性的误差函数来训练高斯混合模型，能够极大地提高系统性能。因此，直到在声学模型中使用深度神经网络的技术出现之前，高斯混合模型一直是短时特征向量建模的不二选择。

但是，高斯混合模型同样具有一个严重的缺点：高斯混合模型对于靠近向量空间上一个非线性流形 (manifold) 上的数据建模能力非常差。

对应到语音识别系统中，我们使用隐马尔可夫模型来刻画一个音素内部子状态变化，来解决特征序列到多个语音基本单元之间对应关系的问题。在语音识别任务中使用隐马尔可夫模型需要计算模型在一段语音片段上的可能性。而在训练的时候，我们需要使用 Baum-Welch 算法[23] 学习隐马尔可夫模型参数，进行最大似然估计 (Maximum Likelihood Estimation, MLE)。Baum-Welch 算法是EM (Expectation-Maximization) 算法的一种特例，利用前后项概率信息迭代地依次进行计算条件期望的 E 步骤和最大化条件期望的 M 步骤。

****4. 语言模型****

语言模型主要是刻画人类语言表达的方式习惯，着重描述了词与词在排列结构上的内在联系。语言模型分为规则模型和统计模型两类，统计语言模型用概率统计的方法来刻画语言单位内在的统计规律。

最简单又却又最常用的语言模型是 N 元语言模型 (N-gram Language Model,N-gram LM) 。

****5. 解码与字典****

解码器是识别阶段的核心组件，通过训练好的模型对语音进行解码，获得最可能的词序列，或者根据识别中间结果生成识别网格 (lattice) 以供后续组件处理。解码器部分的核心算法是动态规划算法 Viterbi。由于解码空间非常巨大，通常我们在实际应用中会使用限定搜索宽度的令牌传递方法 (token passing)。