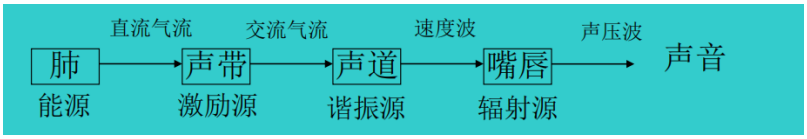
语音信号处理

基础知识

- 三个主要语言器官：

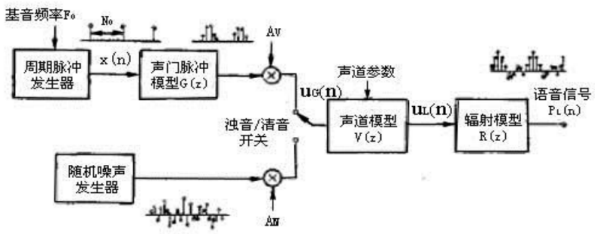


  - 肺：语音产生的能源所在
  - 声带：为产生语音提供主要的激励源
  - 声道：具有非均匀截面，且随时间变化，起谐振器的作用
- 三种语音类型：

  人耳能听到的频率在20~20K Hz

  - 浊音：声带振动产生准周期的声门脉冲激励声道产生浊音
    - 在时域是准周期的，在频域具有谐波结构
    - 周期脉冲的频率就是基频或基音
  - 清音：当气流在声道中受到阻碍时，产生湍流，此时生成清音
    - 在时域类似随机噪声，在频域具有宽带特征
  - 爆破音：
- 两个声学特性：
  - 基音频率：其值等于声带张开和闭合一次的时间的倒数
  - 共振峰：共振峰及其带宽取决于声道某一瞬间的形状和尺寸

    共振峰是声道的重要声学特性。声道对于一个激励信号的响应，可以用一个含有多对极点的线性系统来近似描述。每对极点都对应一个共振峰频率。这个线性系统的频率响应特性称为共振峰特性，它决定信号频谱的总轮廓，或称谱包络。
- 三个数字模型：



  - 激励模型：
    - 浊音激励

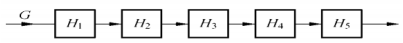
      准周期性脉冲波，其周期为基音周期，单个脉冲的波形类似于斜三角波
    - 清音激励

      随机白噪声

- 声道模型

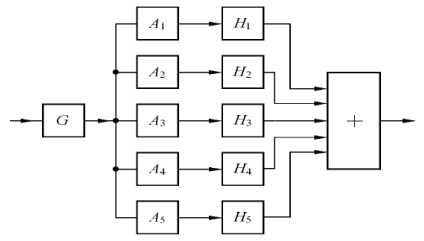
  谐振器

  - 级联型

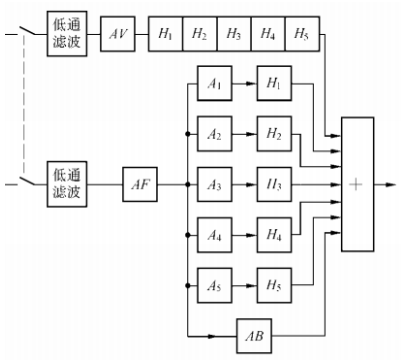
    适用于一般单元音

  - 并联型

    适用于鼻音、复合元音及大部分辅音

  - 混合型

- 辐射模型

- 语言学和语音学的区别
  - 语言学

    主要对控制语音中各个音的排列规则及其含义进行研究

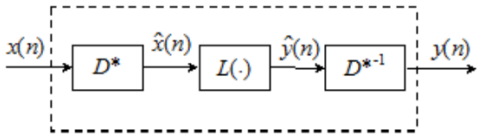
  - 语音学

    研究语音中各个音的物理特征和分类的学科。

- 语音的声学特性
  - 四个物理属性：
    - 音色：由共振峰决定
    - 音调：由基音频率决定
    - 音强
    - 音长
  - 两种结构：
    - 音素
      - 国际分类：浊音和清音

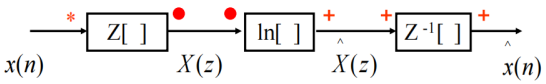
        - 我国分类：
            - 元音：属于浊音
            - 辅音：分浊辅音和清辅音
        - 音节：一个或多个音素构成
            - 汉语语音是单音节结构，语音音调包含语义
            - 英语语音是多音节结构，语音音调不包含语义
    - 两类语谱图
        - 窄带语谱图：频带宽度约为45Hz，具有良好的频率分辨率，但时间分辨率较差
        - 宽带语谱图：频带宽度约为300Hz，具有良好的时间分辨率，但频率分辨率较差
    - 听觉器官
        - 外耳：相当于共振腔
        - 中耳：相当于低通滤波器
        - 内耳
    - 掩蔽效应：当某一频率的声音，有一特定音强存在时，另一个不同频率的声音要将音强提高才会被听到，这就是听觉掩蔽效应。
    -
- # 短时时域分析
    - 短时分析技术
        - 因为语音信号具有很强的时变特性，但在较短时间内可以认为语音特征保持不变
        - 预处理
            - 分帧
            - 加窗：减小语音帧的截断效应
                > 哈明窗的主瓣最宽，旁瓣高度最低，可以有效的克服泄漏现象，具有更平滑的低通特性
    - 常用的时域短时分析技术
        - 短时能量：平方使对高电平非常敏感
        - 短时平均幅度
        - 短时过零率：信号频率的简单度量，在一帧信号中，信号波形穿过横轴的次数
        - 短时自相关函数
        - 短时平均幅度差函数：解决了短时自相关函数计算量大的问题
    - 浊音、清音、无声的短时特性
        - 浊音的短时平均幅度最大，无声的短时平均幅度最小
        - 清音的短时过零率最大，无声居中，浊音的短时过零率最小。
        - 浊音是周期信号，浊音的短时自相关函数也呈现明显的周期性，自相关函数的周期就是浊音信号的周期；清音接近于随机噪声，请音的短时自相关函数不具有周期性
- # 短时频域分析

- 同态分析(倒谱分析)：设法将非线性问题转化为线性问题来处理的一种方法。
  - 卷积同态分析：

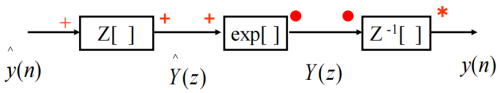
    通过特征系统将两个信号的卷积运算转换为加性运算(非线性变为线性)，并通过逆特征系统恢复为卷积信号

    - 特征系统

      把卷积转换为和，把非线性变为线性
    - 线性系统

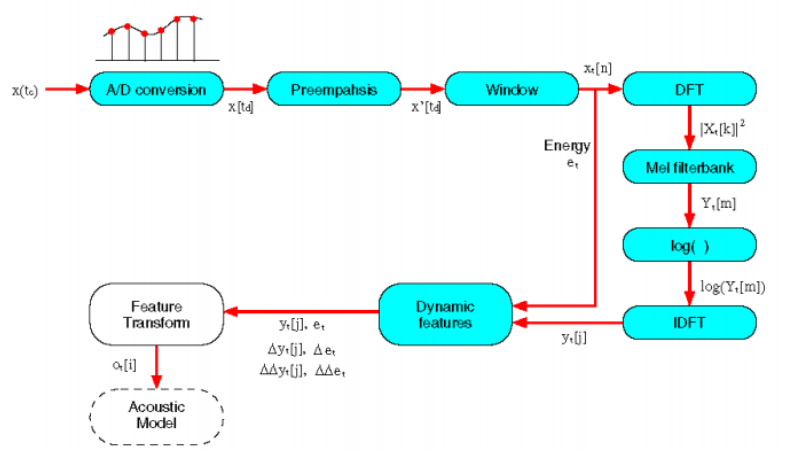
      真正需要的处理算法，可利用信号与系统中所学过的各种处理手段，满足叠加原理
    - 逆特征系统

      把和转换为卷积，把线性变为非线性
  - 复倒谱和倒谱
    - 复倒谱：经过特征系统后的值
      - 有幅频特性和相频特性
    - 倒谱：复倒谱去掉相位后的值
      - 由于人的听觉对相位不敏感，为了减小计算，丢掉相位
      - 由于相位丢失，不可还原为复倒谱
  - 声门激励和声道相应
    - 声门激励

      复倒谱是无限冲激序列，幅度变、周期不变
    - 声道响应

      集中在原点附近 、双边衰减
    - 综上：
      - 倒谱通过低倒谱窗后经过逆特征系统恢复可得到声道响应
      - 倒谱通过高倒谱窗后经过逆特征系统恢复可得到声门激励
- 特征参数提取

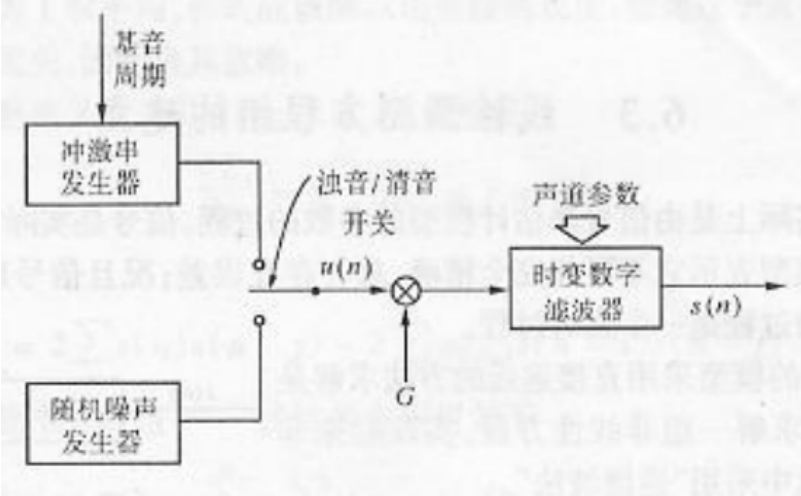


- A/D转换：将模拟信号转换为数字信号
- 预加重：增大高频部分的幅度，平衡频谱(频谱倾斜现象)
- 加窗：分帧作用，保持信号的短时不变性；合适的窗函数还能减小语音帧的截断效应
- DFT变化：特征系统的第一步，然后用频谱得到能量谱
- Mel滤波器组：Mel刻度在在低频分辨率高，高频分辨率低，符合人耳特性
- 计算对数(log)幅度平方：这一步去除了相位信息，减小计算量。并且在这一步输出FBANK特征
- IDFT变换：转换为倒谱域，提取倒谱特征。取前12个倒谱特征系数作为为MFCC特征
- 动态特征：在12维MFCC特征中加入1维能量及它们的一阶和二阶动态特征构成39维GMM/HMM特征

- MFCC特征和FBANK特征对比
  - MFCC特征是经过逆变换的倒谱特征，提取12维特征。特征之间相关性小，更适合GMM/HMM系统
  - FBANK特征没有经过逆变换。特征之间相关性大，更适合DNN/HMM系统。
- MFCC和GMM的适应性
  - GMM中每个高斯成分的协方差都取对角协方差矩阵，各维度相关性小；而MFCC经过逆特征变换，特征之间相关性小。因此两者相适应
- FBANK和DNN的适应性
  - DNN不需要特征向量之间不相关，可以使用具有相关性的特征向量；而实验表明FBANK比MFCC效果更好。因此FBANK适合于DNN。

线性预测分析

- 语音样本之间存在相关性，一个语音信号的样本可以用过去若干个样本的线性组合来逼近；
- 常用来合成语音



- # 高斯混合模型
  - 三个问题
    - Likelihood(概率评估)：前向算法
    - Decoding and alignment(最优状态序列)：Viterbi算法
    - Training(参数估计)：前后向算法和EM算法
  - 前向算法
  - Viterbi算法
  - 例子

    假设有3个盒子，编号为1,2,3，每个盒子都装有红白两种颜色的小球，数目如下：

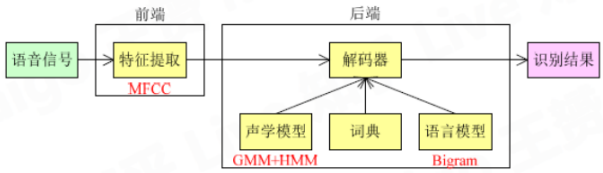
盒子号	1	2	3
红球数	5	4	7
白球数	5	6	3

    $$\pi=\begin{pmatrix}0.2\\0.4\\0.4\end{pmatrix}\quad A=\begin{bmatrix}0.5&0.2&0.3\\0.3&0.5&0.2\\0.2&0.3&0.5\end{bmatrix}\quad B=\begin{bmatrix}0.5&0.5\\0.4&0.6\\0.7&0.3\end{bmatrix}$$

- # 语音识别
  - 概念：把语音转换成文字
  - 语音识别系统的分类
    - 识别单位：孤立词、连接词、连续语音、语义理解、会话语音识别
    - 词汇量：小(<50个)、中(2000以下)、大(2000以上)
    - 讲话人范围：特定人、非特定人
    - 使用环境：公共场合、录音室
  - 性能的评价
    - 三种错误：删除错误、插入错误、替换错误
    - 评价指标：词错误率(word error rate——WER)

- 计算方法：
  - 将标准答案和识别结果对齐
  - 用插入、删除、替换错误的总数除以标准答案的长度
  - 对齐应使得错误数最少
- 孤立词语音识别：
  - 概念：发音认真、单词之间有停顿、端点检测较易。前后单词之间是孤立的，识别基础建立在数学方法之上，不含“语言”知识。
  - 识别技术：
    - DTW
    - HMM
    - 混合技术(VQ/HMM)
- 连续语音识别：
  - 语言模型：
    - 单词序列的可能性的定量排序(统计角度)
    - 如何创建单词序列或句子的一组规则(语法角度)
  - 连接数字串的语音识别
    - 连接词与孤立词语音识别的差异：
      - 连续语流中的识别基元受发音时的上下文等影响
      - 连续语流中的识别基元之间的边界预先未知
  - 大词汇量连续语音识别：
    - 不能为每个单词训练单独的HMM，改成为每一个音素训练一个HMM
    - HMM的复合：
      - 音素HMM按照词典拼接成单词HMM
      - 单词HMM与语言模型复合成语音HMM
    - 识别基元的选择与切分：

      对单词进行识别显然是不可能的，因此，必须选择恰当的识别基元。这种选择应考虑用尽量少且又易于从连续语流中切分出来的基元。
    - 语音识别层次模型：
      - 特征层
      - 语音层
      - 语言层
      - 应用层
  - 语音识别系统结构：
    - 框图



- 改进：
  - 上下文有关模型
  - 区分式训练
  - 说话人适应
  - 二次打分

说话人识别

- 分类：
  - 说话人辨认(Identification)：
    - 从一组已知的声音中确定谁在说话
    - 不需要用户的进行声明(一对多映射)
  - 说话人确认(Verification)：
    - 确定这个人是否是他所声称的人
    - 用户提出身份声明（一对一映射）
- Speech Modalities：
  - **Text-dependent(T-D)**：训练和测试的文本相同
  - **Text-independent(T-I)**：训练和测试的文本任意
- 表征说话人特点的基本特征，这些特征应该具有如下特点：
  - 能够有效地区分不同的说话人，但又能在同 一说话人的语音发生变化时相对保持稳定。
  - 易于从语音信号中提取。
  - 不易被模仿。
- 说话人确认：
  - 本质：是一个两类假设检验
    - H0：语音S来自冒名顶替者
    - H1：语音S来自声明者
    - 对数似然比(LLR)：

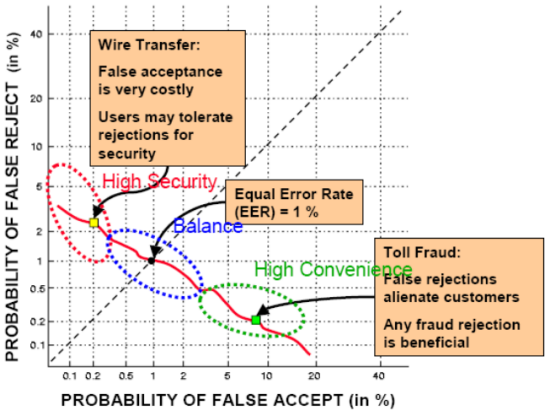
$$LLR = \log p(S|H1) - \log p(S|H0) \quad \begin{matrix} LLR > \theta & \text{Accept} \\ LLR < \theta & \text{Reject} \end{matrix}$$

    - UBM(Universal Background Model)
      - 通过大量的演讲训练来代表一般的语音模型

$$\log p(S|H0) = \log p(S|UBM)$$

      - GMM-UBM

- 两类错误
  - **错误拒绝率False Reject**： 拒绝真实的说话人而造成的错误
  - **错误接收率False Accept**： 接受假冒者而造成的错误
- Equal Error Rate：



- Decision Cost Function

$$DCF = C_{fa} \cdot FA \cdot P_{imp} + C_{fr} \cdot FR \cdot P_{tar}$$

C_{fa} = **Cost of a false alarm**

P_{imp} = **Prior probability of impostor attempt**

C_{fr} = **Cost of a miss/ false reject**

$P_{tar} = 1 - P_{imp}$ = **Prior probability of true speaker attempt**

- 读语谱图
- 实验