第五章 语音识别 复习

目录

- 语音识别概念
  （音节、音素、音位）
- 特征提取
  流程
  MFCC
- 孤立词识别技术
  DTW
  GMM
- 语音识别技术
  HMM
  大词汇量的情况
- 语音识别系统
  基于 GMM-HMM的语音识别系统
  基于神经网络的语音识别系统

语音基本概念

- 音节

  听觉能够自然察觉到的最小语音单位，音节有声母、 韵母、声调三部分组成。一个汉字的读音就是一个音节，一个英 文单词可能有一个或多个音节构成。

- 音素

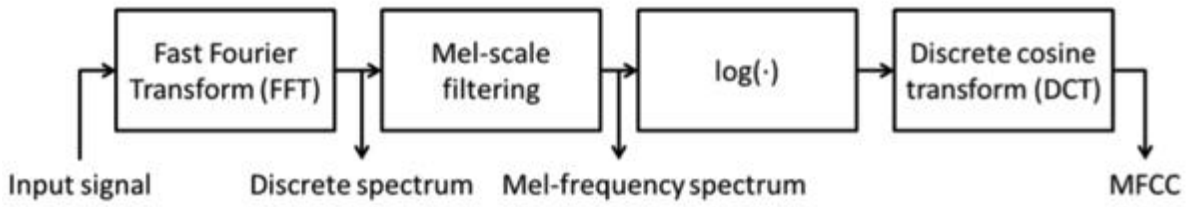
  从音节中分析出来的最小语音单位，语音分析到音素 就不能再分了。比如，“她穿红衣服”是5个音节，而“红”又可进一 步分为3个音素--h,o,ng。

- 音位:

  能够区分意义的音素，比如bian,pian,bu,pu就是靠b，p 两个音素来区分的，所以b，p就是两个音位。

特征提取

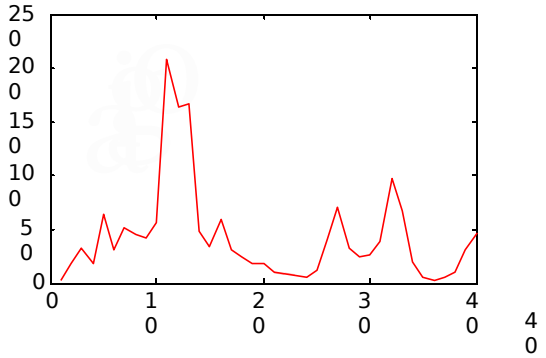
- 流程



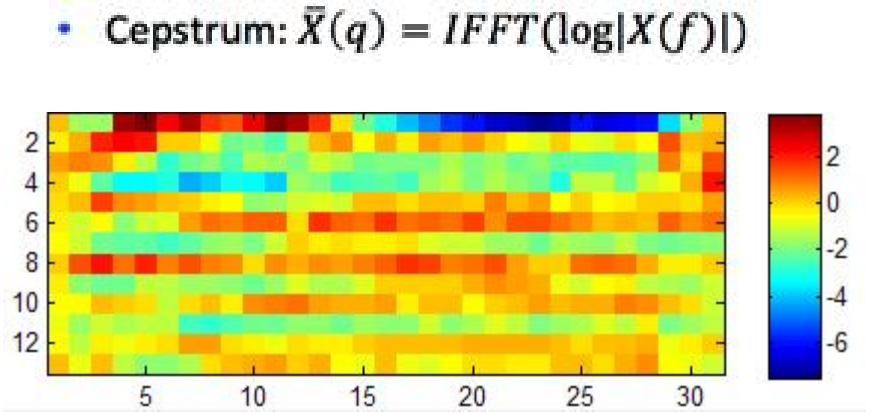
  - 截取出信号帧，通常 20 ~ 50 ms
  - 傅里叶变换计算频谱
    - –精细结构反映音高，用处较小
    - –包络反映音色，是主要信息
  - 三角滤波得到近似频谱包络（Mel spectrum）
    - 目的：模拟人耳响应随频率的特性
  - 近似频谱包络得到MFCC

特征提取

- MFCC



log,
DCT



  - Log
  - DCT，信号回到cepstral（正比于频率倒数）而非时域

    -优点：排除基频，符合听觉，维度低
    -缺点：视野小，受噪声、回声、滤波影响严重

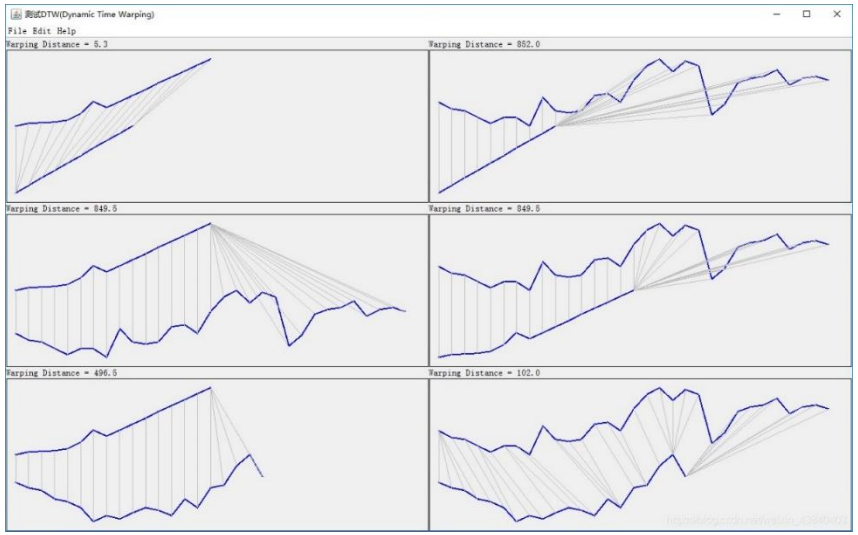
孤立词识别技术

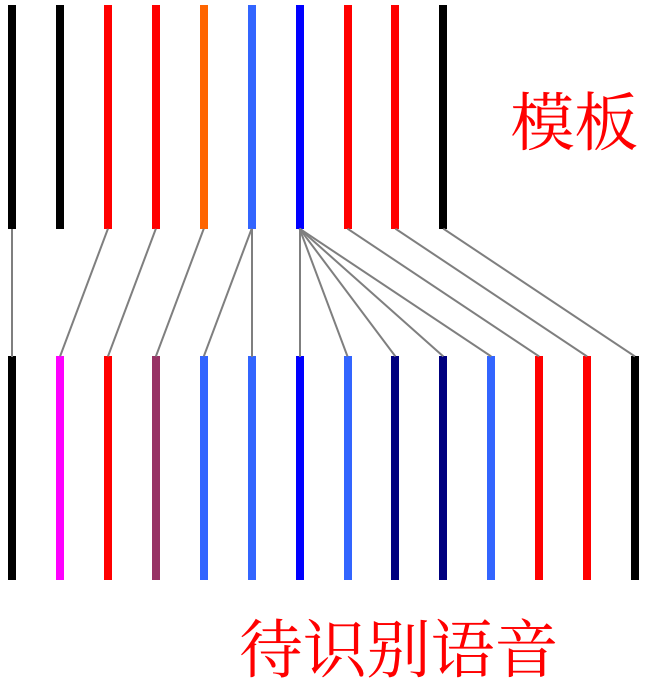
- 模板比较法
  - 1个模板：动态时间调整算法（DTW）
  - 多个模板：高斯混合模型（GMM）

孤立词识别技术

- DTW

  计算两个不同长度、不同节奏的时间序列的相似度
  采用动态规划DP方法，扭曲时间序列的形态
  按照顺序，让每一帧与模板中最相似的一帧匹配
  总距离为各帧的 欧氏距离之和





孤立词识别技术

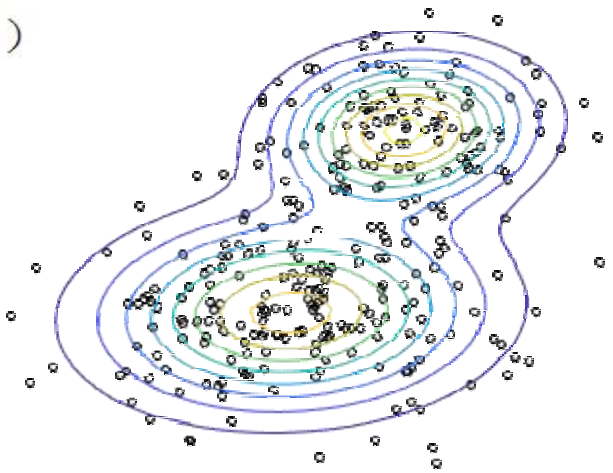
- GMM

  多个高斯分布函数的叠加，理论上能够拟合任意分布

  用GMM概率密度代替特征向量间的欧氏距离

高斯分布: $f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

GMM: $f(x) = \sum_{i}^{k} \phi_i \frac{1}{\sqrt{2\sigma_i^2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$ （$\sum_{i}^{k} \phi_i = 1$）

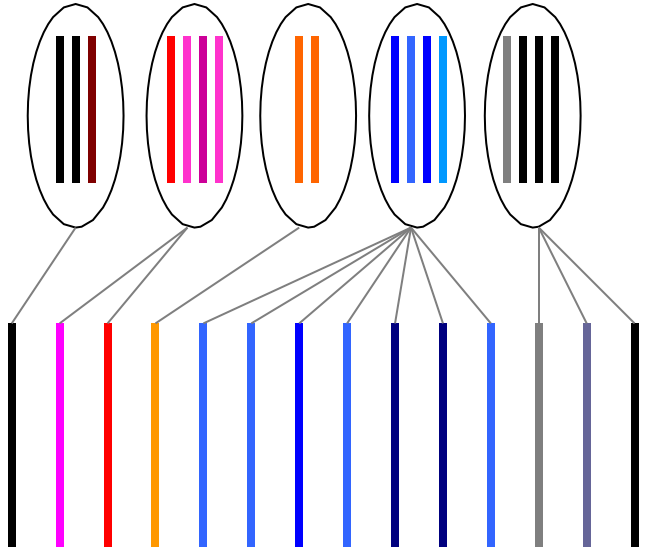


孤立词识别技术

- GMM
  - DTW对齐待识别语音与模型
  - 计算模型每个状态的GMM
  - 对于一帧的MFCC，计算每个状态的GMM模型的概率P(待识别语音|模型)
  - 取概率最大的模型为识别结果

模型

待识别语音



语音识别技术

- 隐马尔可夫模型（HMM)

包含：

$\Pi = (\pi_i)$ ：初始化概率向量；

$A = (a_{ij})$ ：状态转移矩阵；$Pr(x_{i_t} | x_{j_{t-1}})$

$B = (b_{ij})$ ：混淆矩阵；$Pr(y_i | x_j)$

状态转移矩阵：

π 向量：定义系统初始化时每一个状态的概率

		Today		
		sun	cloud	rain
Yesterday	sun	0.50	0.375	0.125
	cloud	0.25	0.125	0.625
	rain	0.25	0.375	0.375

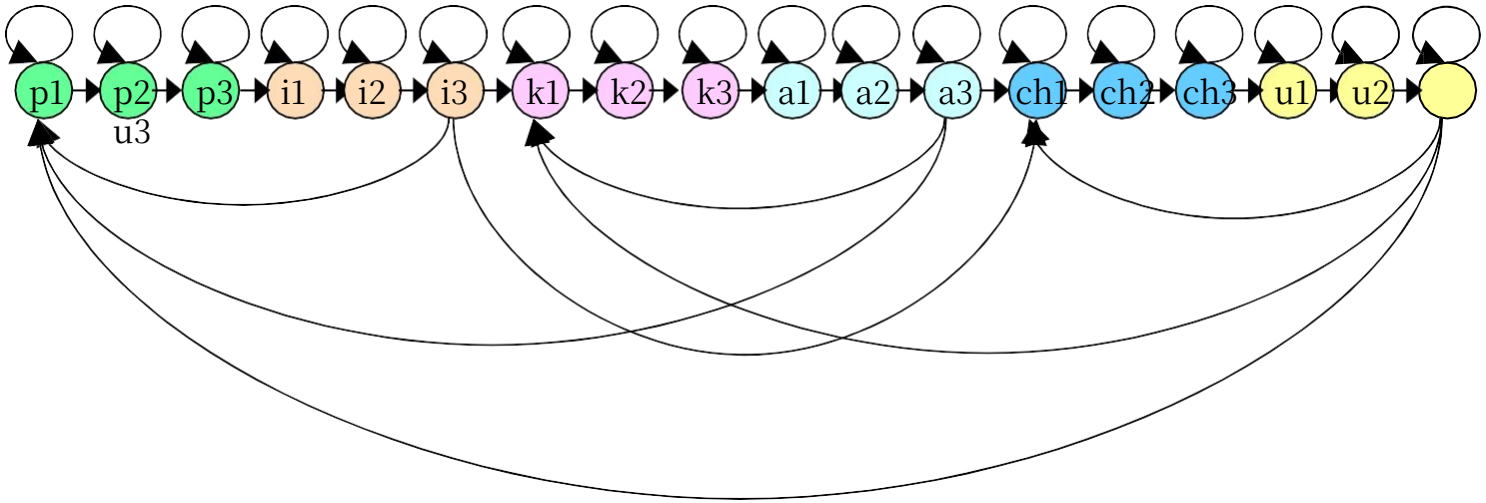
语音识别技术

- 隐马尔可夫模型（HMM)
  1.DTW对齐待识别语音与模型
  2.计算每个音素的HMM
  3.对于一帧的MFCC，计算每个音素的HMM模型的概率P(待识别语音|模型)
  4.取概率最大的模型为识别结果

$$W^* = \arg\max_W P(W \mid X) = \arg\max_W \frac{P(X \mid W)P(W)}{P(X)} = \arg\max_W P(X \mid W)P(W)$$

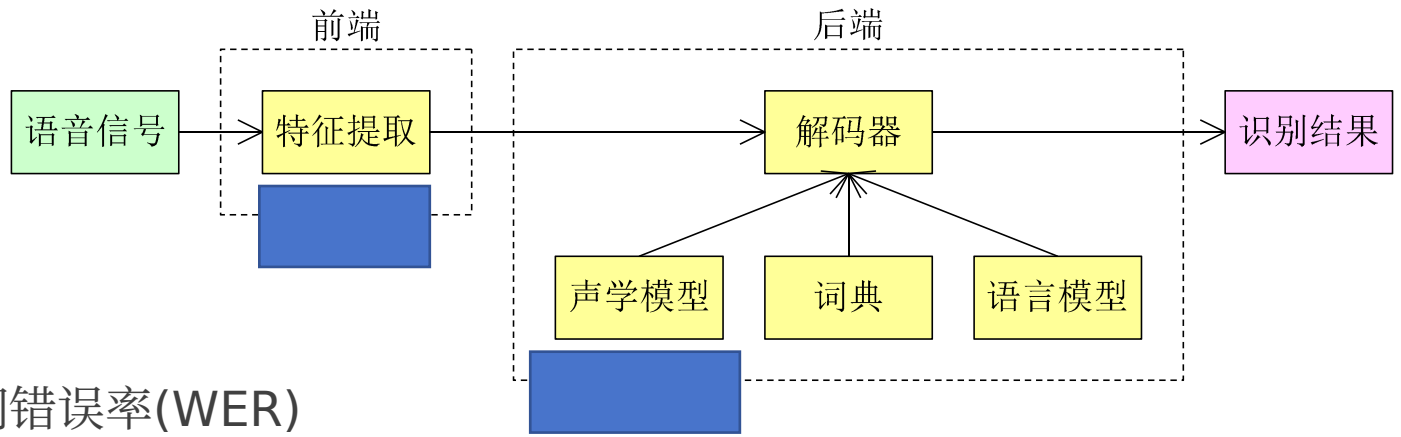
语音识别技术

- 大词汇量语音识别
  - 不能为每个单词训练单独的HMM
    - 改成为每个音素训练一个HMM
  - HMM的复合：
    - 音素HMM按词典拼接成单词HMM
    - 单词HMM与语言模型复合成语言HMM



语音识别系统

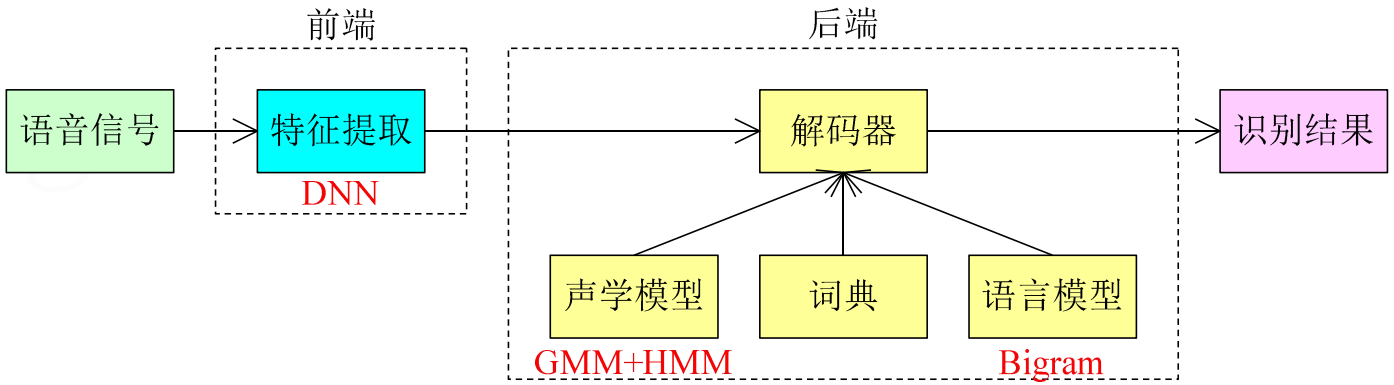
- 基于 GMM-HMM的语音识别系统
  - 结构



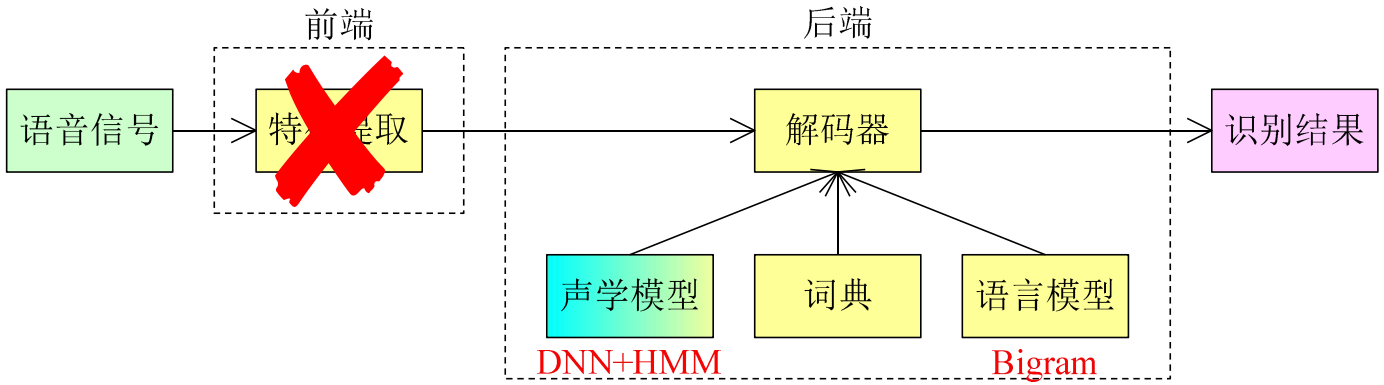
  - 评价指标：词错误率(WER)
    - 用插入、删除、替换错误的总数除以标准 答案的长度

语音识别

- 基于神经网络的语音识别系统
  - Tandem结构



  - Hybrid结构



语音识别

- 基于神经网络的语音识别系统

  3.直接使用

  循环神经网络（recurrent neural network，RNN）

  理想：直接替代

  实际：代替DNN用于特征提取或声学模型，

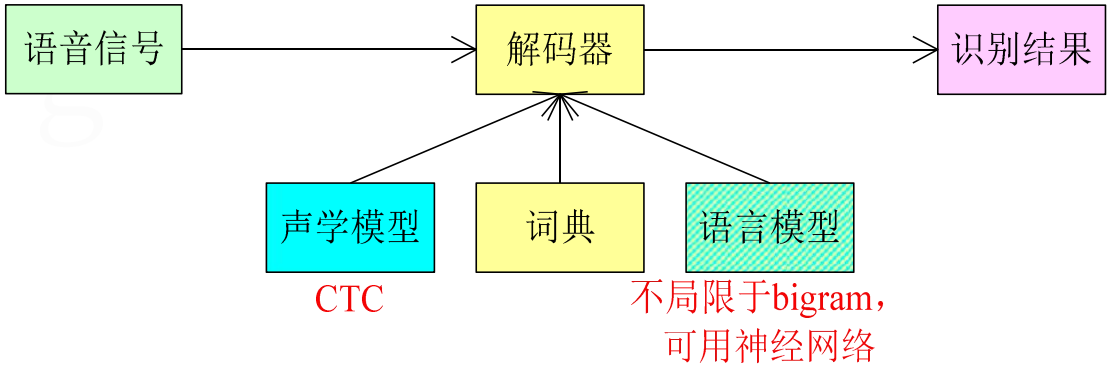
  HMM保留

  - 训练时提供各音素起止时间
  - 解码时提共状态转移概率

语音识别

- 基于神经网络的语音识别系统

  4. Grapheme系统



CTC

端到端，只关心预测输出的序列是否和真实的序列接近

不需要预先做对齐或转移概率