语音识别基本原理介绍

杜俊

提纲

- 语音识别简单回顾
- 基于贝叶斯统计建模的理论框架
  - 声学特征
  - 声学模型
  - 语言模型
  - 解码搜索
- 语音识别的难点及其他相关技术介绍

语音识别领域发展现状

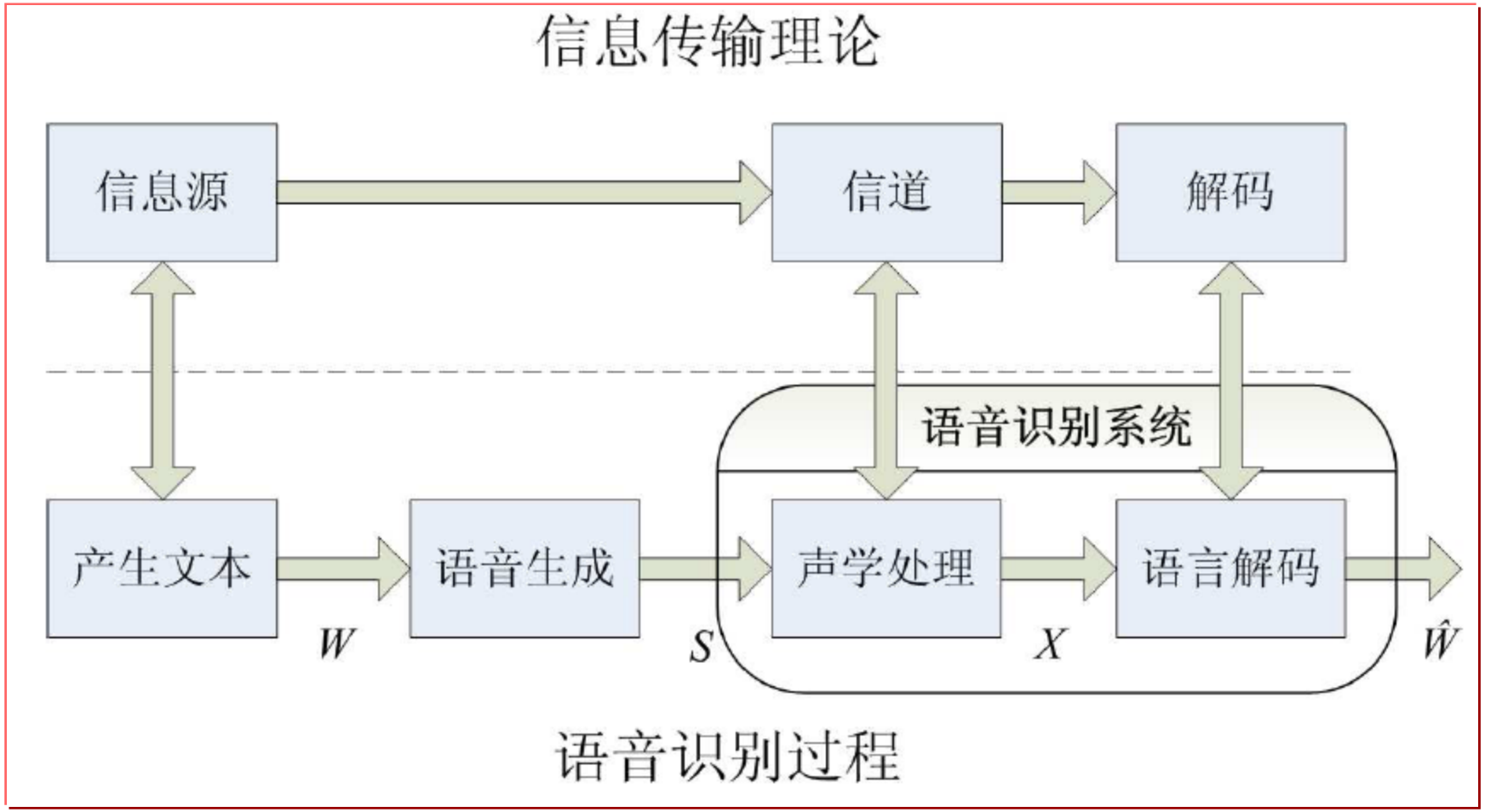
- 产业界
  - Nuance：全球最大的语音识别技术提供商
  - IBM：具有强大数学底蕴的老牌语音识别研究机构
  - Microsoft：Vista操作系统中首次加入语音识别功能
  - Google：凭借互联网方面的优势GOOG-411，音乐搜索
- 学术界
  - Cambridge：HTK工具对学术界研究推动巨大
  - CMU：SPHINX-李开复
  - SRI，MIT，RWTH，ATR

- 语音识别实用化方面的两种论调
  - 悲观：缺少杀手级应用，与人类的语音识别水平还有很大差距
  - 乐观：Nuance能如此成功，计算机存储和运算能力的不断提高

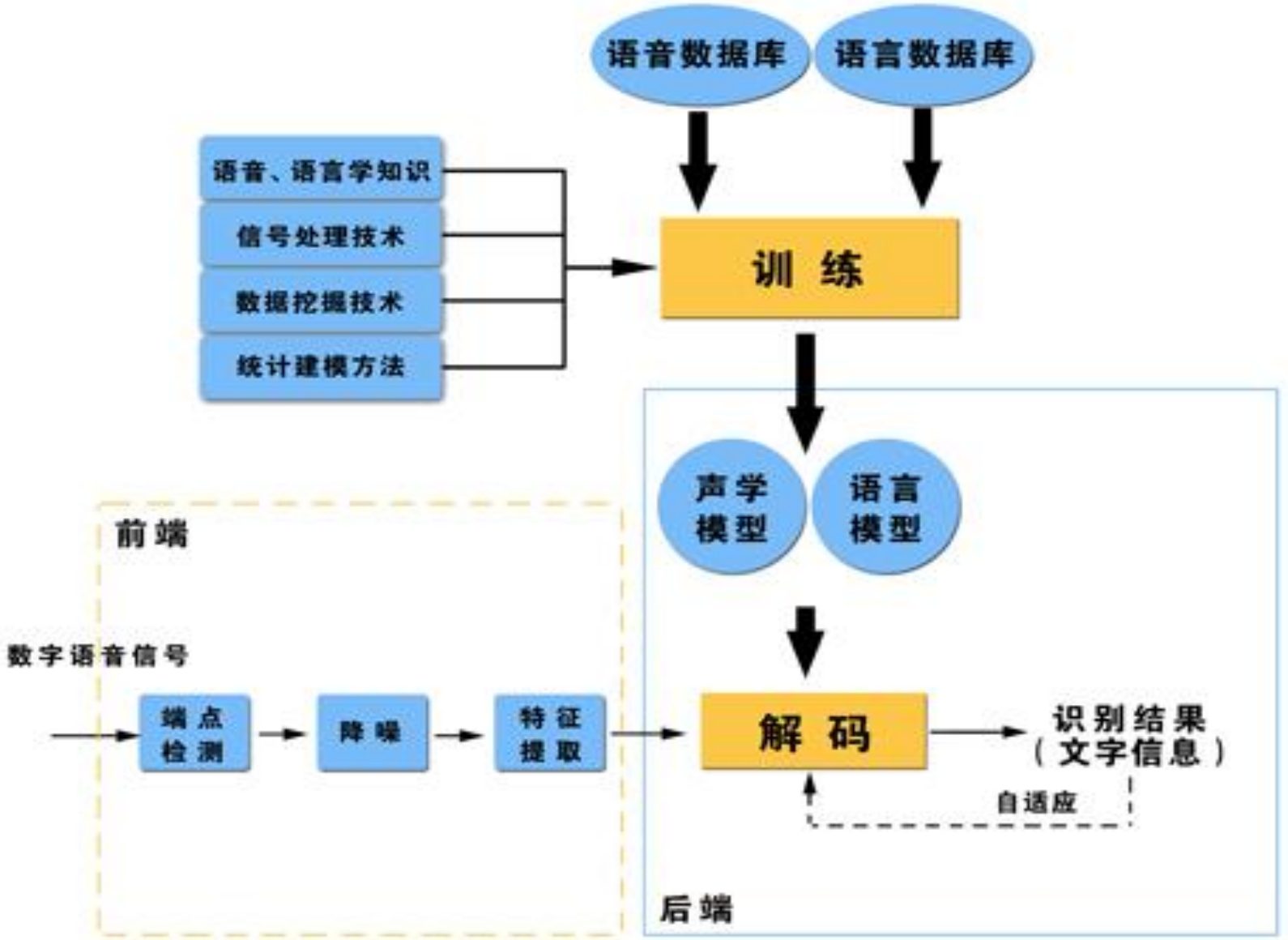
语音识别各种具体应用

- 命令词系统
  - 识别语法网络相对受限，对用户要求较严格
  - 菜单导航，语音拨号，车载导航，数字字母识别等等
- 智能交互系统
  - 对用户要求较为宽松，需要识别和其他领域技术的结合
  - 呼叫路由，POI语音模糊查询，关键词检出
- 大词汇量连续语音识别系统
  - 海量词条，覆盖面广，保证正确率的同时实时性较差
  - 音频转写
- 结合互联网的语音搜索
  - 实现语音到文本，语音到语音的搜索

从信道传输理论来看语音识别



语音识别基本框图


语音数据库
语言数据库
语音、语言学知识
信号处理技术
数据挖掘技术
统计建模方法
训 练
声学模型
语言模型
前端
数字语音信号
端点检测
降噪
特征提取
解 码
识别结果（文字信息）
自适应
后端

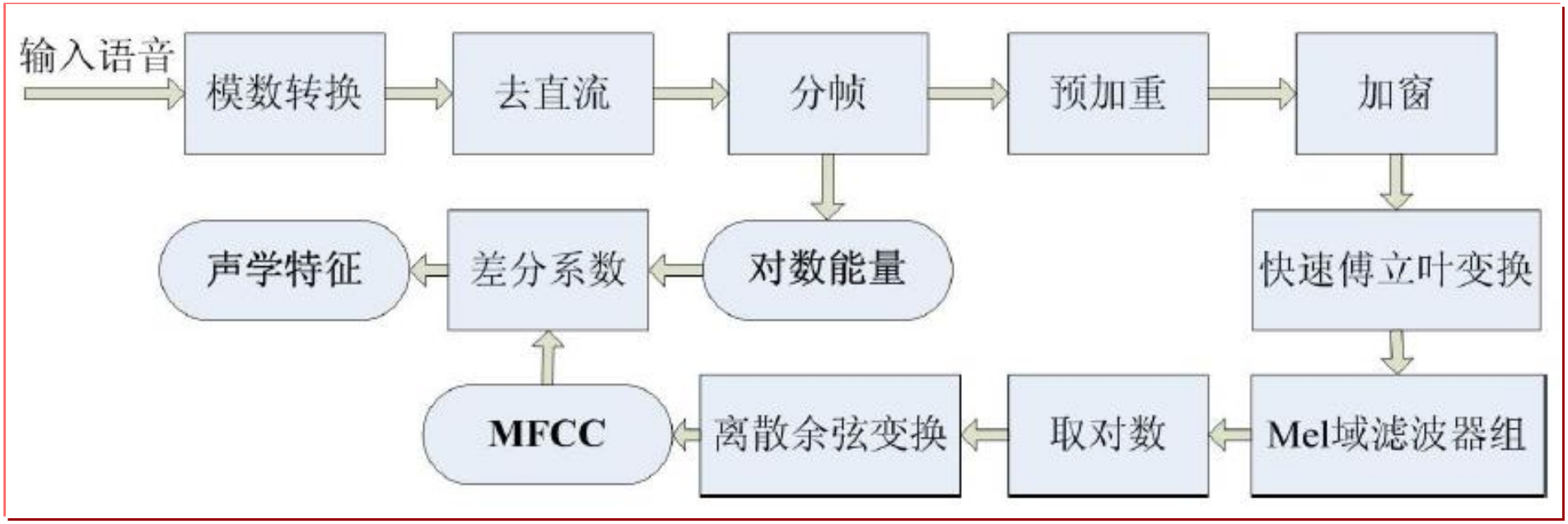
语音识别基本原理

- 贝叶斯统计建模框架（MAP/最大后验概率决策准则）
  - Plug-In MAP

$$\hat{W} = \arg\max_{W \in \Gamma} p(W \mid X) = \arg\max_{W \in \Gamma} P(W) \cdot p(X \mid W)$$
$$= \arg\max_{W \in \Gamma} \overline{P}_{\Gamma}(W) \cdot \overline{p}_{\Lambda}(X \mid W)$$

  - 声学特征--- X 通过前端特征提取获得
  - 声学模型--- $\overline{p}_{\Lambda}(X \mid W)$ 对声学特征进行统计建模
  - 语言模型--- $\overline{P}_{\Gamma}(W)$ 对词串进行统计建模
  - 解码搜索---通过设计算法得到最优词串 $\hat{W}$

声学特征提取示例



- 简单来说，X 是一个帧序列，而每帧就是一个多维向量

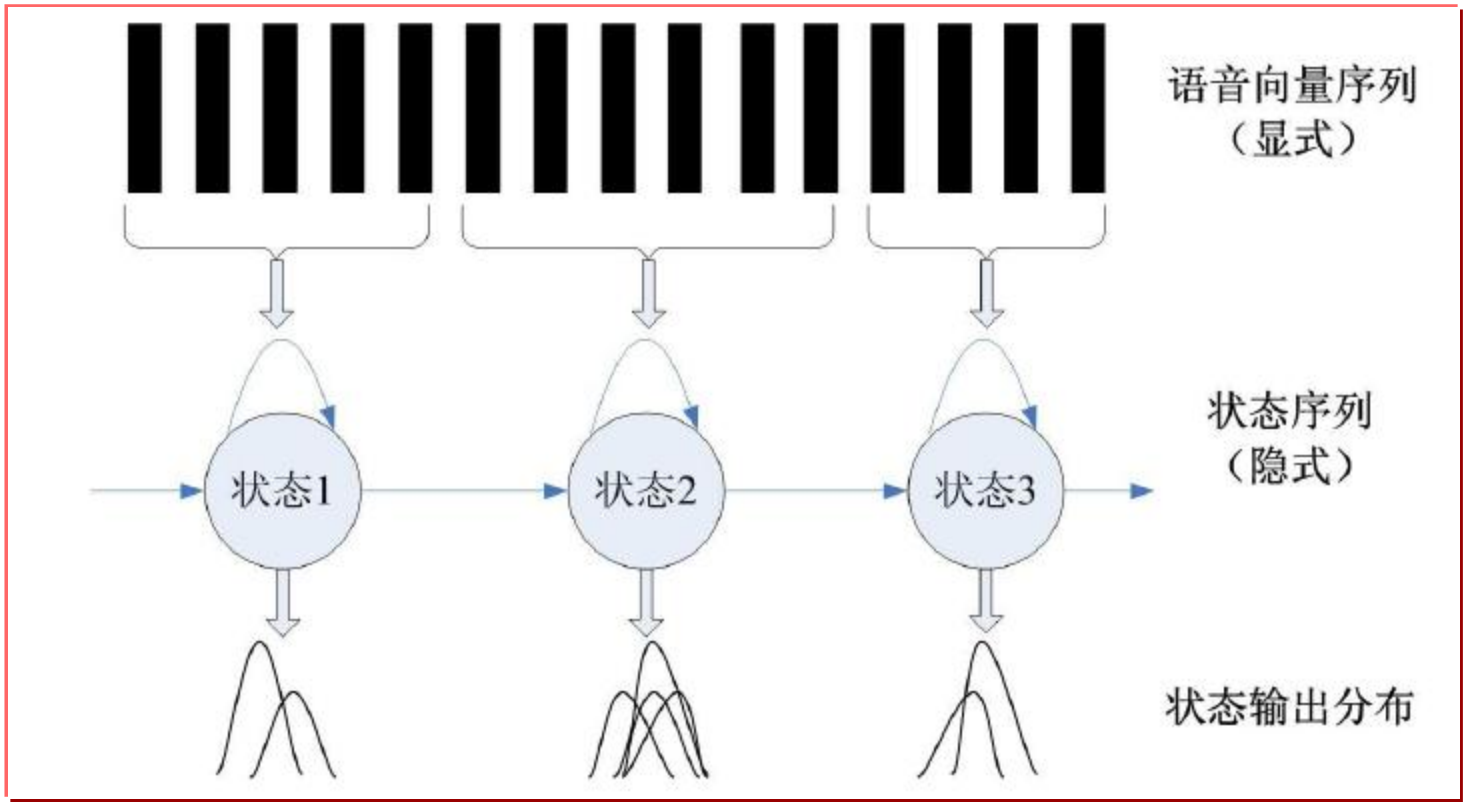
声学模型 $\overline{p}_{\Lambda}(X|W)$

- 声学单元应该具有的特性
  - 一致性：不同语音实例中相同的语音单元在声学上一致
  - 可训练性：建模单元需要足够的训练数据来进行参数估计
  - 可共享性：不同的建模单元之间共享某些具有共性的训练数据
- 声学单元如何挑选？
  - 句子(sentence)：　科大讯飞实验室
  - 单词(word)：　科大-讯飞-实验室
  - 单字(syllable)：　科-大-讯-飞-实-验-室
  - 音素(phone)：　k-e-d-a-x-un-f-ei-sh-i-y-an-sh-i
  - 考虑协同发音的三元音素(tri-phone)：ei-sh+i和an-sh+I
  - 精细建模和训练数据量之间的矛盾如何解决？参数绑定
- 声学单元对应的模型形式应该是什么？
  - 隐Markov模型（HMM），神经网络（NN）

声学模型 $\overline{p}_{\Lambda}(X|W)$

- Markov过程和Markov链
  - 描述了一个最小记忆系统的随机行为
- 隐Markov模型（HMM）：双重随机过程

安德雷·安德耶维齐·马尔可夫



声学模型 $\overline{p}_{\Lambda}(X|W)$

- HMM的几要素
  - 观测向量 $X = \{x\}$
  - 状态集合 $\Omega = \{s_i\}$
  - 初始状态概率 $\pi = \{\pi_i\}$
  - 状态转移矩阵 $\mathbf{A} = \{a_{ij}\}$
  - 状态输出概率分布 $\mathbf{B} = \{b_i(x)\}$

$$a_{ij} \geq 0, \quad b_i(x) \geq 0, \quad \pi_i \geq 0$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \int b_i(x)\,\mathrm{d}x = 1 \qquad \sum_{i=1}^{N} \pi_i = 1$$

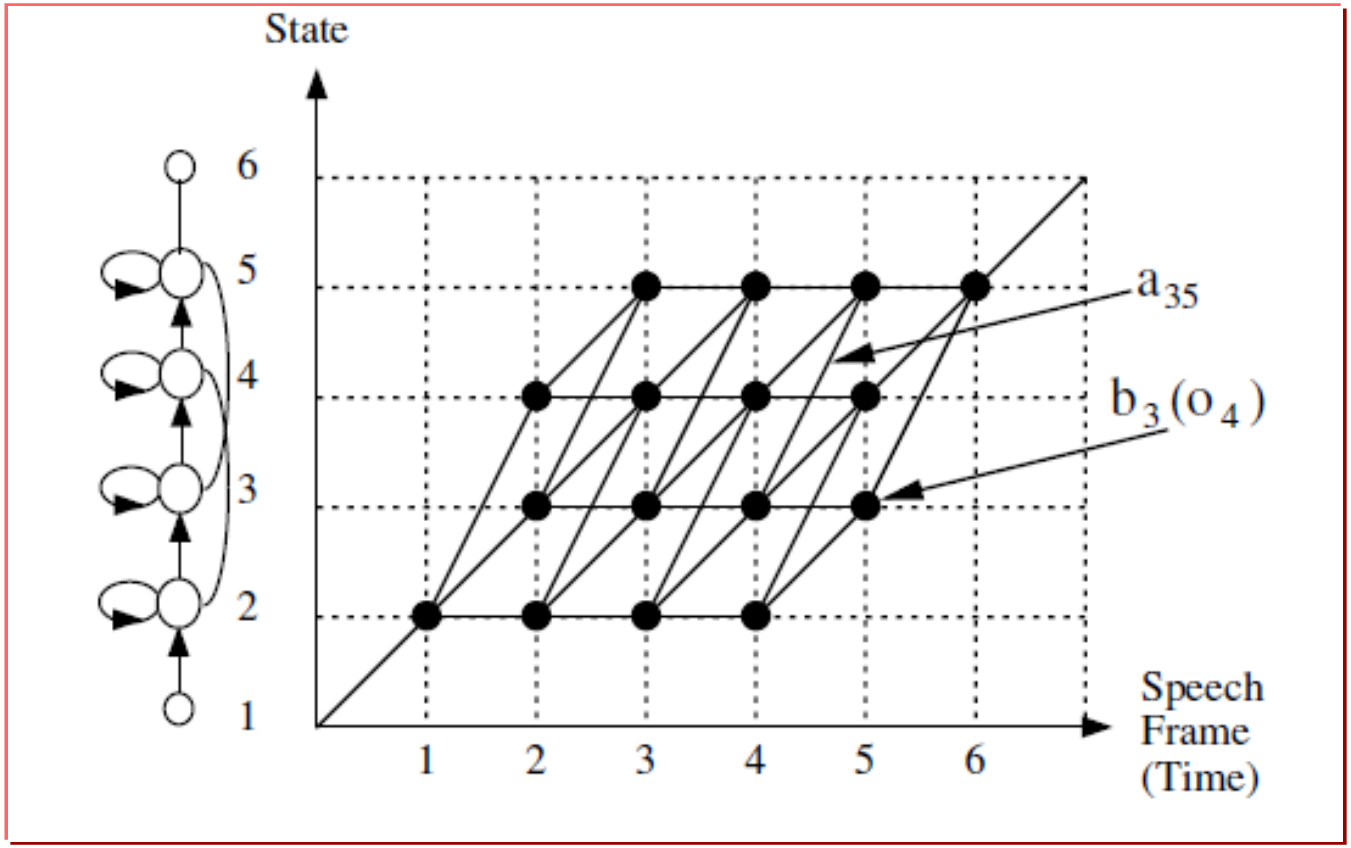
声学模型 $\overline{p}_{\Lambda}(X|W)$

- HMM的两假设
  - 一阶Markov假设 $p(s_t \mid s_1^{t-1}) = p(s_t \mid s_{t-1})$
  - 输出无关假设 $p(x_t \mid x_1^{t-1}, s_1^t) = p(x_t \mid s_t)$

- HMM的三个问题
  - 评估问题
    - 给定HMM模型参数以及一串观测序列，如何求得观测序列的似然度
  - 解码问题
    - 给定HMM模型参数以及一串观测序列，如何搜索出最优的状态序列
  - 训练问题
    - 给定观测序列，如何得到模型参数

声学模型 $\overline{p}_{\Lambda}(X \mid W)$

- 由观测和状态组成的网格



声学模型 $\overline{p}_{\Lambda}(X \mid W)$

- 评估问题

$$p(X \mid \Phi) = \sum_{S} p(X, S \mid \Phi) = \sum_{S} p(S \mid \Phi) p(X \mid S, \Phi)$$

$$p(S \mid \Phi) = p(s_1 \mid \Phi) \prod_{t=2}^{T} p(s_t \mid s_{t-1}, \Phi) = \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1} s_t}$$

$$p(X \mid S, \Phi) = \prod_{t=1}^{T} b_{s_t}(x_t)$$

$$p(X \mid \Phi) = \sum_{S} \pi_{s_1} b_{s_1}(x_1) a_{s_1 s_2} b_{s_2}(x_2) \ldots a_{s_{T-1} s_T} b_{s_T}(x_T)$$

声学模型 $\overline{p}_{\Lambda}(X \mid W)$

- 评估问题
  - 运算量太大，需要寻找快速算法—前向算法（Forward Algorithm）

$$\alpha_t(i) = p(x_1^t, s_t = i \mid \Phi)$$

前向算法描述

第一步：初始化

$$\alpha_1(i) = \pi_i b_i(x_1) \qquad 1 \le i \le N$$

第二步：递归计算

$$\alpha_t(i) = \left[\sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji} \right] b_i(x_t) \qquad 1 \le i \le N; \quad 2 \le t \le T$$

第三步：得出评估结果

$$p(X \mid \Phi) = \sum_{i=1}^{N} \alpha_T(i)$$

$O(N^T)$ → $O(N^2T)$

声学模型 $\overline{p}_{\Lambda}(X \mid W)$

- 解码问题—Viterbi算法 $S^* = \arg\max_S p(X, S \mid \Phi)$

$V_t(i) = p(x_1^t, s_1^{t-1}, s_t = i \mid \Phi)$

Viterbi算法描述

第一步：初始化

$$V_1(i) = \pi_i b_i(x_1), \quad B_1(i) = \emptyset \qquad 1 \le i \le N$$

第二步：递归计算

$$V_t(i) = \max_{1 \le j \le N} \left[V_{t-1}(j) a_{ji} \right] b_i(x_t) \qquad 1 \le i \le N; \quad 2 \le t \le T$$

$$B_t(i) = \arg\max_{1 \le j \le N} \left[V_{t-1}(j) a_{ji} \right]$$

第三步：得出解码结果

$$p(X, S^* \mid \Phi) = \max_{1 \le i \le N} V_T(i)$$

$$s_T^* = \arg\max_{1 \le i \le N} V_T(i)$$

第四步：回朔

$$s_t^* = B_{t+1}(s_{t+1}^*) \qquad t = T-1, T-2, \ldots, 1$$

$$S^* = (s_1^*, s_2^*, \ldots, s_T^*)$$

声学模型 $\overline{p}_{\Lambda}(X|W)$

- 训练问题—最大似然估计
  - EM算法（ Expectation-Maximization Algorithm）
  - Baum-Welch算法/前后向算法（Forward-Backward ）

前后向算法描述

第一步：选择初始模型Φ

第二步：E步基于Φ计算辅助函数$Q(\Phi,\hat{\Phi})$

第三步：M步根据优化的更新公式计算$\hat{\Phi}$，所需统计量可由前后向概率得到

第四步：$\Phi=\hat{\Phi}$，返回第二步进行迭代直到收敛

EM算法概述

- 解决什么样的问题
  - 存在隐藏或者丢失数据时的参数估计，无法直接获得
  - 1977年由哈佛的A. P. Dempster等人正式提出EM算法
- 具体都有哪些应用
  - 语音识别：训练(MLE,MAP)和自适应(MLLR,MAPLR)
  - 信号处理：降噪
- 算法流程是怎样的(迭代过程)
  - E step：根据不完整数据构造完整数据的似然度
  - M step: 最大化此似然度，得到参数更新

EM算法性质及其扩展

- 递增性和收敛性
  - 局部最优

- 扩展算法
  - ECM: Expectation Conditional Maximization
  - GEM: Generalized Expectation Maximization

EM算法在语音识别中的应用

- 单高斯模型的最大似然估计
- 混合高斯模型的最大似然估计
  - 为何要用EM
  - 隐藏数据或者隐变量是什么
  - E步构造辅助函数
  - M步最大化辅助函数
- 基于混合高斯HMM的最大似然估计
  - Baum-Welch算法和Forward-Backward算法

语言模型 $\overline{P}_\Gamma(W)$

- N-Gram语言模型

$$\Pr(W) = \Pr(w_1, w_2, \cdots, w_M) = \prod_{i=1}^{M} p(w_i \mid h_i)$$

  - Uni-gram
  - Bi-gram
  - Tri-gram

- Context-Free Grammar (CFG)

解码搜索

- Viterbi算法
  - 时间同步和时间异步
  - 搜索空间裁减
  - N-best和Word-Graph

- 对于命令词/孤立词识别网络，情况要简化很多
  - 对于每条命令词先扩展成HMM序列，然后计算得分
  - 选择得分最大的作为识别输出结果

语音识别的难点及其他相关技术介绍

- 说话人的差异
  - 不同说话人：发音器官，口音，说话风格
  - 同一说话人：不同时间，不同状态
- 噪声影响
  - 背景噪声
  - 传输信道，麦克风频响

- 鲁棒性技术
  - 区分性训练
  - 特征补偿和模型补偿

语音识别的难点及其他相关技术介绍

- 说话人的差异
  - 不同说话人：发音器官，口音，说话风格
  - 同一说话人：不同时间，不同状态
- 噪声影响
  - 背景噪声
  - 传输信道，麦克风频响

- 鲁棒性技术
  - 区分性训练
  - 特征补偿和模型补偿

谢谢大家！

用正确的方法，做有用的研究！