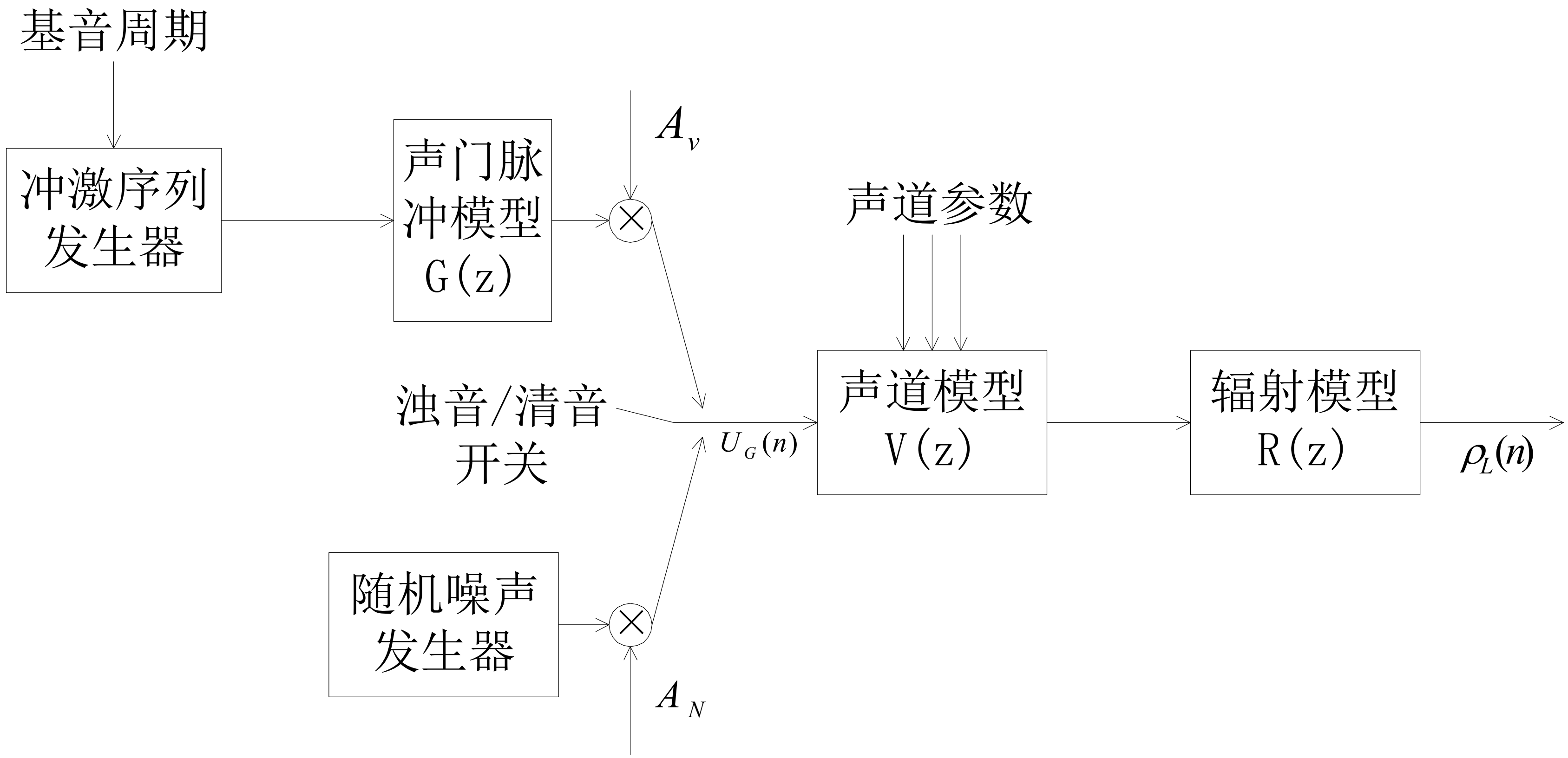
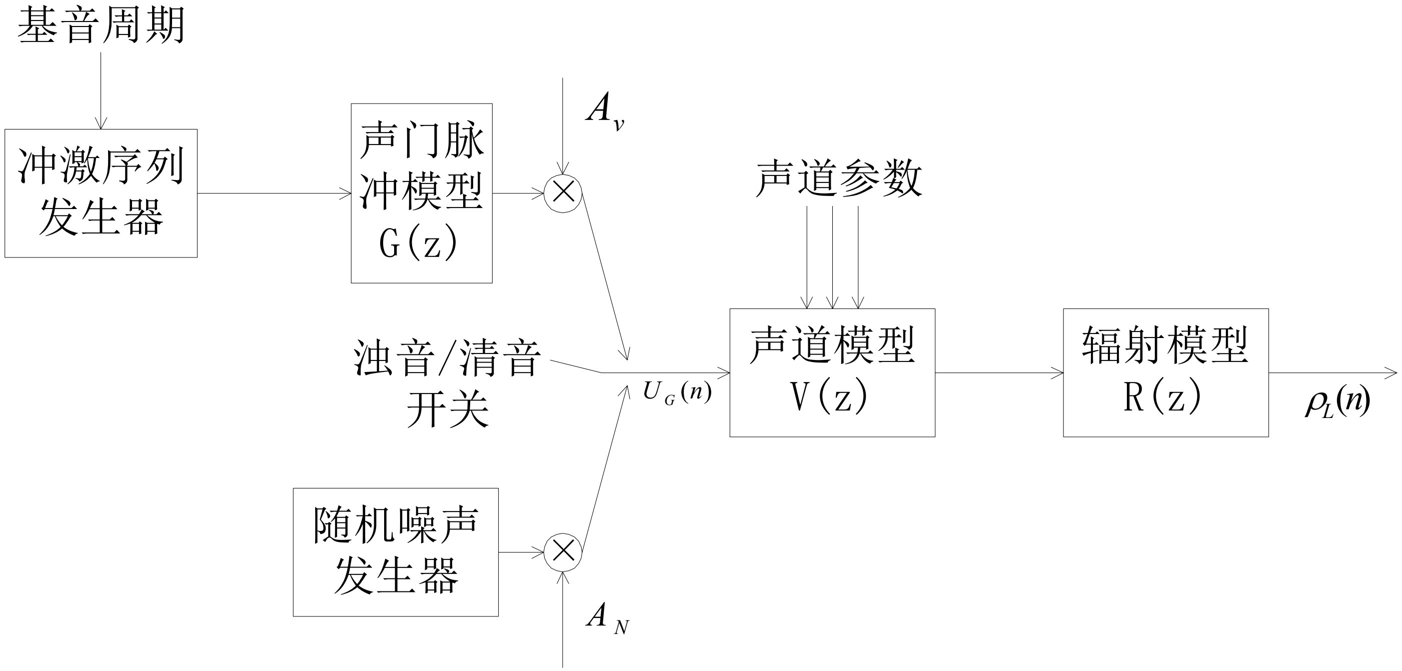
语音产生模型

语音信号产生的数字模型



语音产生模型



语音信号产生的完整模型为 $$H(z)=U(z)V(z)R(z)$$

数字语音处理III

Digital Speech Processing

https://courses.zju.edu.cn/course/join/8P84SXUIRG2

访问码: 8P84SXUIRG2

语音识别

Speech Recognition

杨莹春

yyc@zju.edu.cn

浙江大学教7-506

2021年10月15日

教学安排

讲授内容

（9月17日）秋1：课程简介＋语音技术引言
（9月24日）秋2：语音时域分析
（10月8日）秋4：语音频域分析&语音识别
（10月15日）秋5：语音识别
（12月17日）冬6：复习及项目成果展示（加实验课）

实验内容

1. PRAAT 语音分析 （9月24日）秋2
2. VOICEBOX说话人识别 （10月8日）
3、项目展示（12月17日）冬6

考试：2022年1月

语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

Applications

- Command and control
  - Manufacturing
  - Consumer products

http://www.speech.philips.com

Nuance, American Airlines:
1-800-433-7300, touch 1

- Database query
  - Resource management
  - Air travel information
  - Stock quote

- Dictation
  - http://www.lhsl.com/contacts/
  - http://www-4.ibm.com/software/speech
  - http://www.microsoft.com/speech/

Applications

Speech-recognition software

- IBM (Via Voice, Voice Server Applications,...)
  - Speaker independent, continuous command recognition
  - Large vocabulary recognition
  - Text-to-speech confirmation
  - Barge in (The ability to interrupt an audio prompt as it is playing)
- Dragon Systems, Lernout & Hauspie (L&H Voice Xpress™
- Philips
  - Dictation
  - Telephone
  - Voice Control (SpeechWave, VoCon SDK, chip-sets)
- Microsoft (Whisper, Dr Who)

Applications

The following is a transcript of a telephone conversation between Victor Zue and the Jupiter system for weather information. Jupiter's toll-free number is 1-888-573-8255.

- [Phone ringing]
- Jupiter: Hi. Welcome to Jupiter, the M.I.T. Lab for Computer Science Weather Information System. This call is being recorded for system development.
  You may hang up or ask for help at any time. How can I help you?
- V.Z.: What is the forecast for Boston today?
- Jupiter: In Boston this afternoon, partly sunny and windy. High around 50. Tonight, clear, lows around 35. What other information can I give you?
- V.Z.: Will it rain tomorrow?
- Jupiter: To my knowledge, the forecast calls for no rain in Boston tomorrow. Is there anything else?

- V.Z.: How about Seattle?
- Jupiter: To my knowledge, the forecast calls for no rain in Seattle tomorrow. Is there something else you would like to know?
- V.Z.: Is it sunny anywhere in the Caribbean?
- Jupiter: Sorry, I am not sure what you said.
- V.Z.: Is it sunny anywhere in the Caribbean?
- Jupiter: I am sorry, I may not know the city you are asking about. Try asking for the state or country. For example, what cities do you know about in Massachusetts?
- V.Z.: What is the temperature in Montego Bay?
- Jupiter: In Montego Bay today, high 86 and low 73. Is there something else?
- V.Z.: Good-bye.

语音识别的性能评价

原句：我 们 明 天 去 天 安 门

识别：我 × 明后天 去 天 坛 ×

删除错误 **Deletion** 插入错误 **Insertion** 替换错误 **Substitution**

正确率：

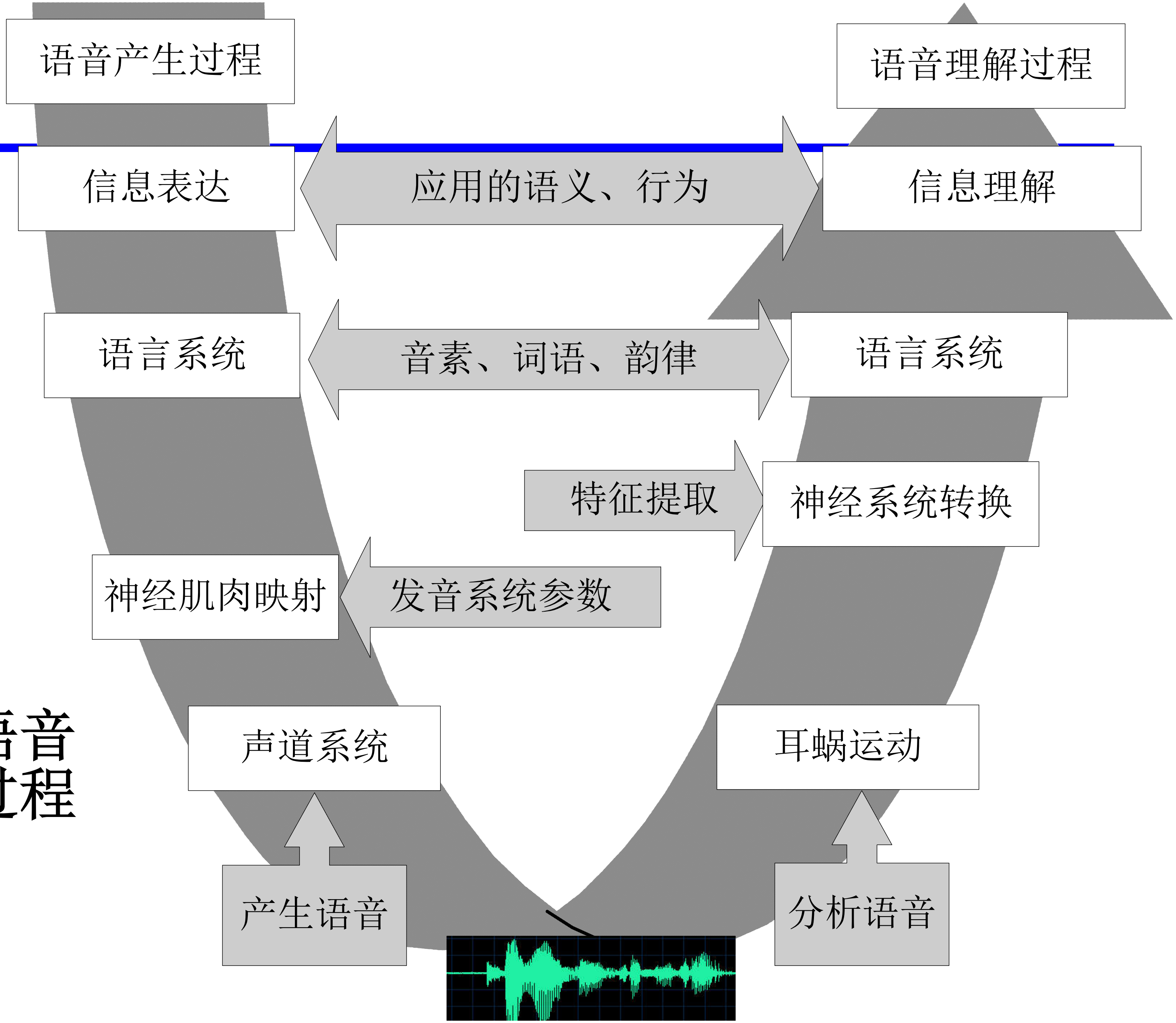
$$Correct = \frac{N - D - S}{N} \times 100\%$$

准确率：

$$Accuracy = \frac{N - D - S - I}{N} \times 100\%$$

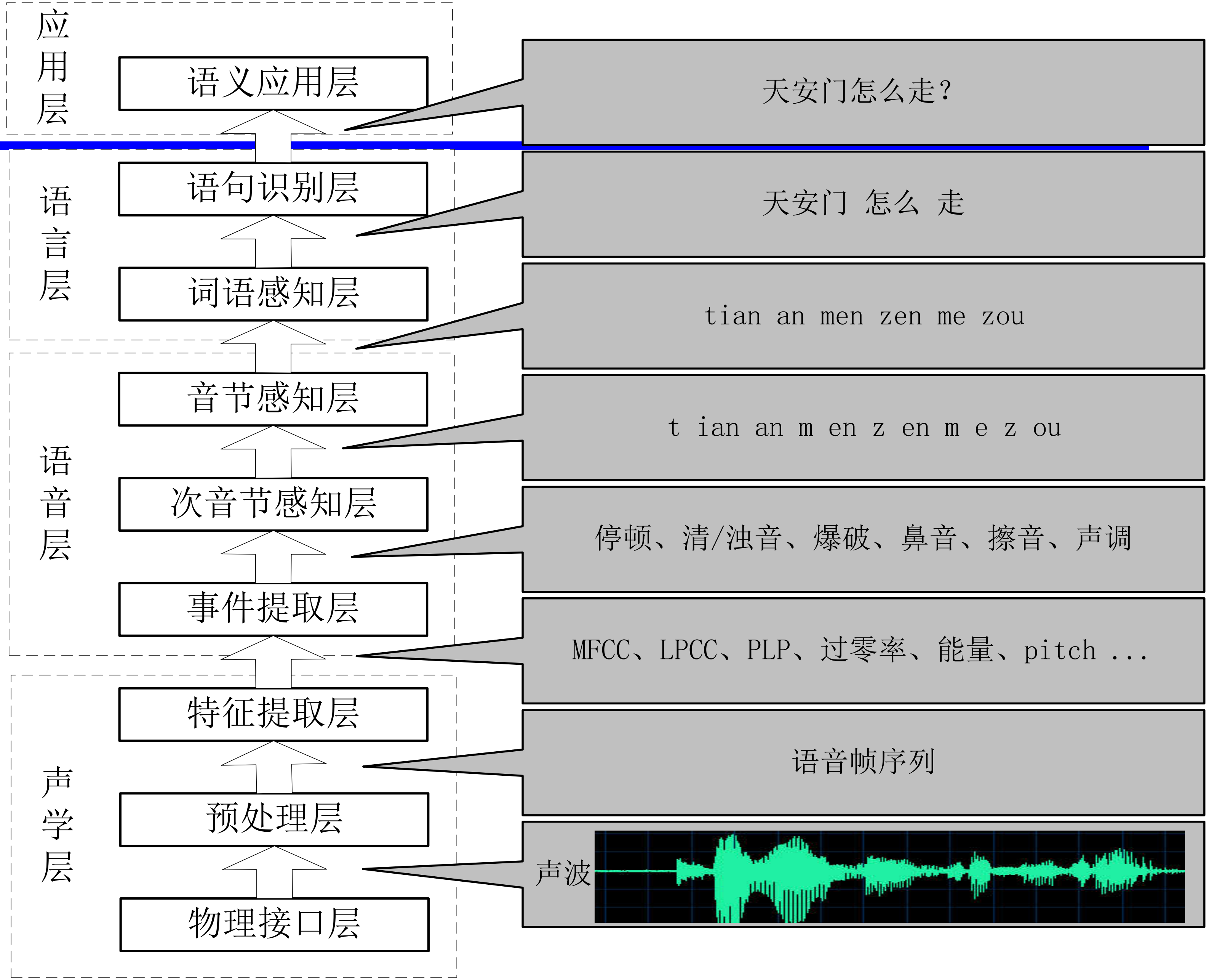
语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型



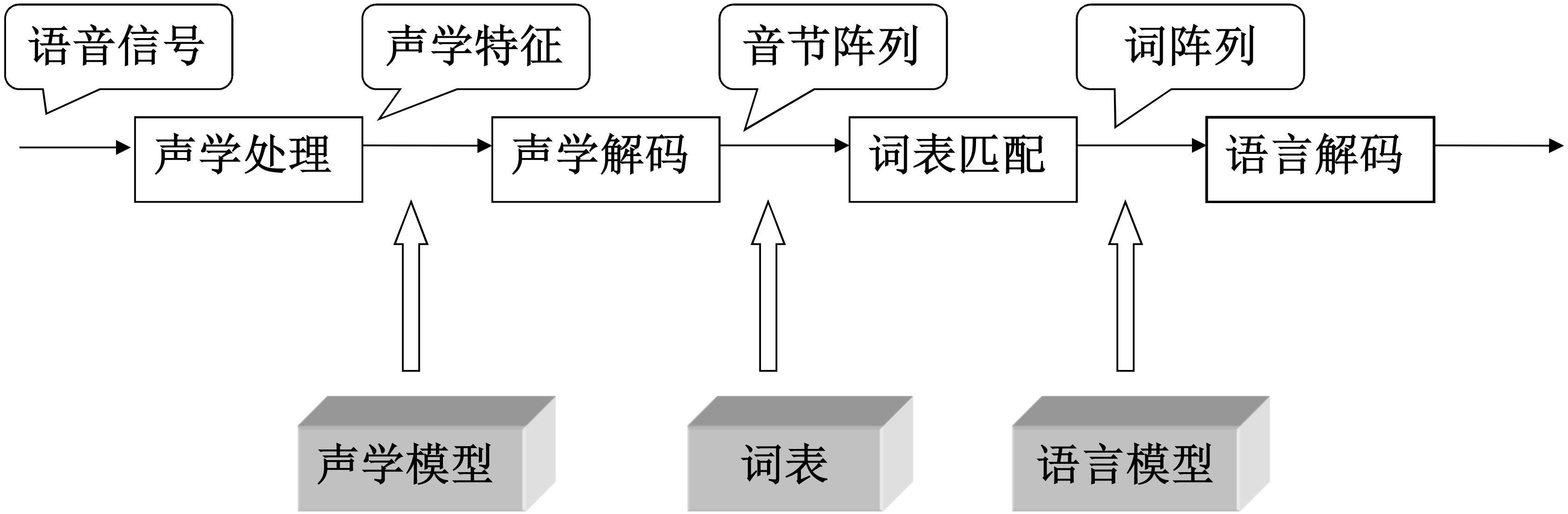
语音产生语音理解生理过程

语音识别层次模型


应用层
语义应用层
天安门怎么走?
语言层
语句识别层
天安门 怎么 走
词语感知层
tian an men zen me zou
语音层
音节感知层
t ian an m en z en m e z ou
次音节感知层
停顿、清/浊音、爆破、鼻音、擦音、声调
事件提取层
MFCC、LPCC、PLP、过零率、能量、pitch ...
声学层
特征提取层
语音帧序列
预处理层
声波
物理接口层

统一层次模型——系统设计


应用层
语义应用层
将查询语句转换成查询系统中语义解释模块可接受的标准格式
语言层
语句识别层
导入语法模型，识别查询语句
词语感知层
导入该任务中的有效词汇，识别词语
语音层
音节感知层
导入单音节拼合规则，识别音节
次音节感知层
导入音素模型，识别音素
事件提取层
声学层
特征提取层
提取13维MFCC（包括C_0）和一阶差分系数
预处理层
采样率16KHz、16bit线性量化、$H(z)=1-0.97z^{-1}$ 滤波、语音分帧（帧长20ms、帧移10ms）、Hamming窗加权、语音端点检测
物理接口层
采用强指向性抗噪麦克风、单声道录音


语音信号
声学特征
音节阵列
词阵列
声学处理
声学解码
词表匹配
语言解码
声学模型
词表
语言模型

Turning Sounds into Words - Current Norm

X = acoustic signal sequence; W = word sequence

$P_\Lambda(W|X) = P_{\lambda_X}(X|W)\, P_{\lambda_W}(W) / P(X)$

objective: maximize the *average* performance (accuracy rate)

$\max_\Lambda P_\Lambda(W|X)$ during training

$\max_W P_\Lambda(W|X)$ during decoding

⇨

$P_{\lambda_W}(W)$

- statistical language models (mostly for large vocabulary ASR)
- grammar expressions (finite-state, context-free, ..)

$P_{\lambda_X}(X|W)$

- hidden Markov model
- mixture density - close approx. to arbitrary distribution

<u>Data-driven methods led to major advances in speech recognition.</u>

语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

特征提取

- 预加重： $y[n]=x[n]-\alpha\cdot x[n-1]$ $\quad 0.9<\alpha<1.0$
- 分帧： 短时平稳**(10-30ms)**
- 加窗：**Hamming** $w[n]=0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right)$ $\quad 0\le n<N$
- 特征参数
- 倒谱均值归一化

特征参数

- 静态参数：Mel-Frequency Cepstrum Coefficients (MFCC)
- 帧能量
- 动态参数

Mel-频率

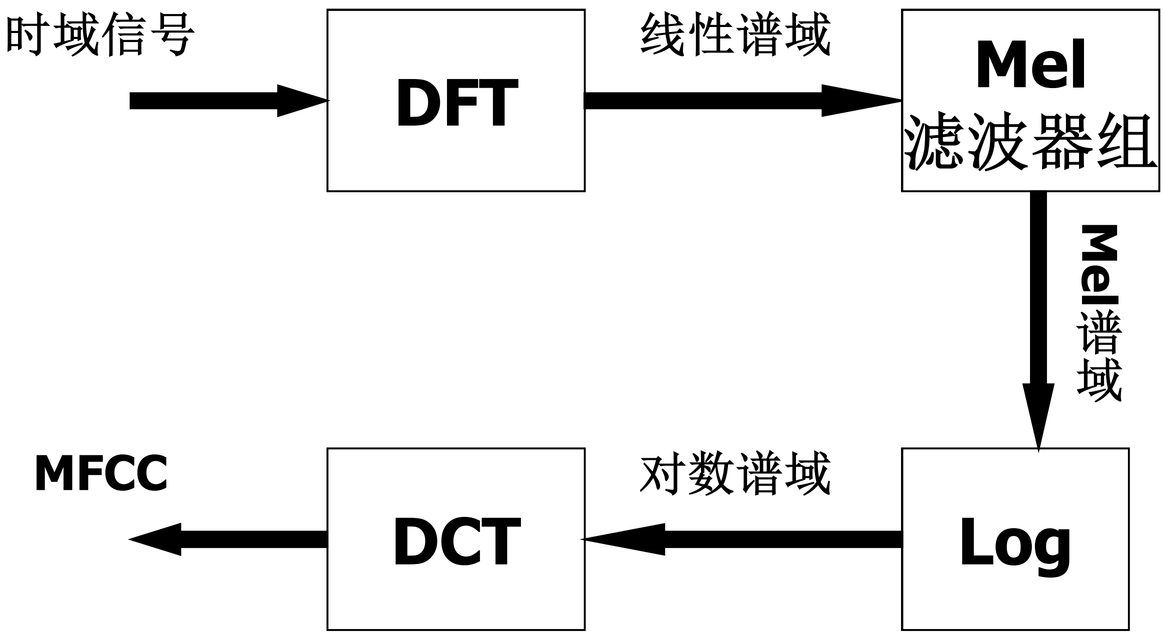
- 目的：模拟人耳对不同频率语音的感知

- 人类对不同频率语音有不同的感知能力
  - **1kHz**以下，与频率成线性关系
  - **1kHz**以上，与频率成对数关系

- **Mel**频率定义
  - **1Mel—1kHz**音调感知程度的**1/1000**

MFCC

- 计算流程：



Discrete Fourier Transform (DFT)

- 公式:

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j2\pi nk/N},0\le n<N$$

$x\left[n\right]$ -- 时域信号

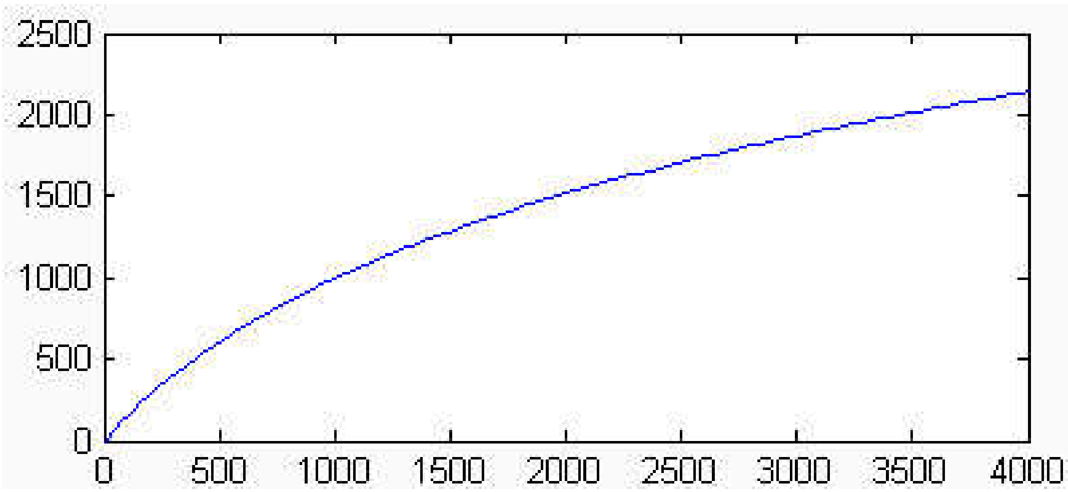
$X\left[k\right]$ -- 频域信号

Mel-频率

- 公式：

$$B(f) = 1125\ln(1 + f/700)$$

- 频率－Mel-频率：f -- 频率　　B -- Mel-频率



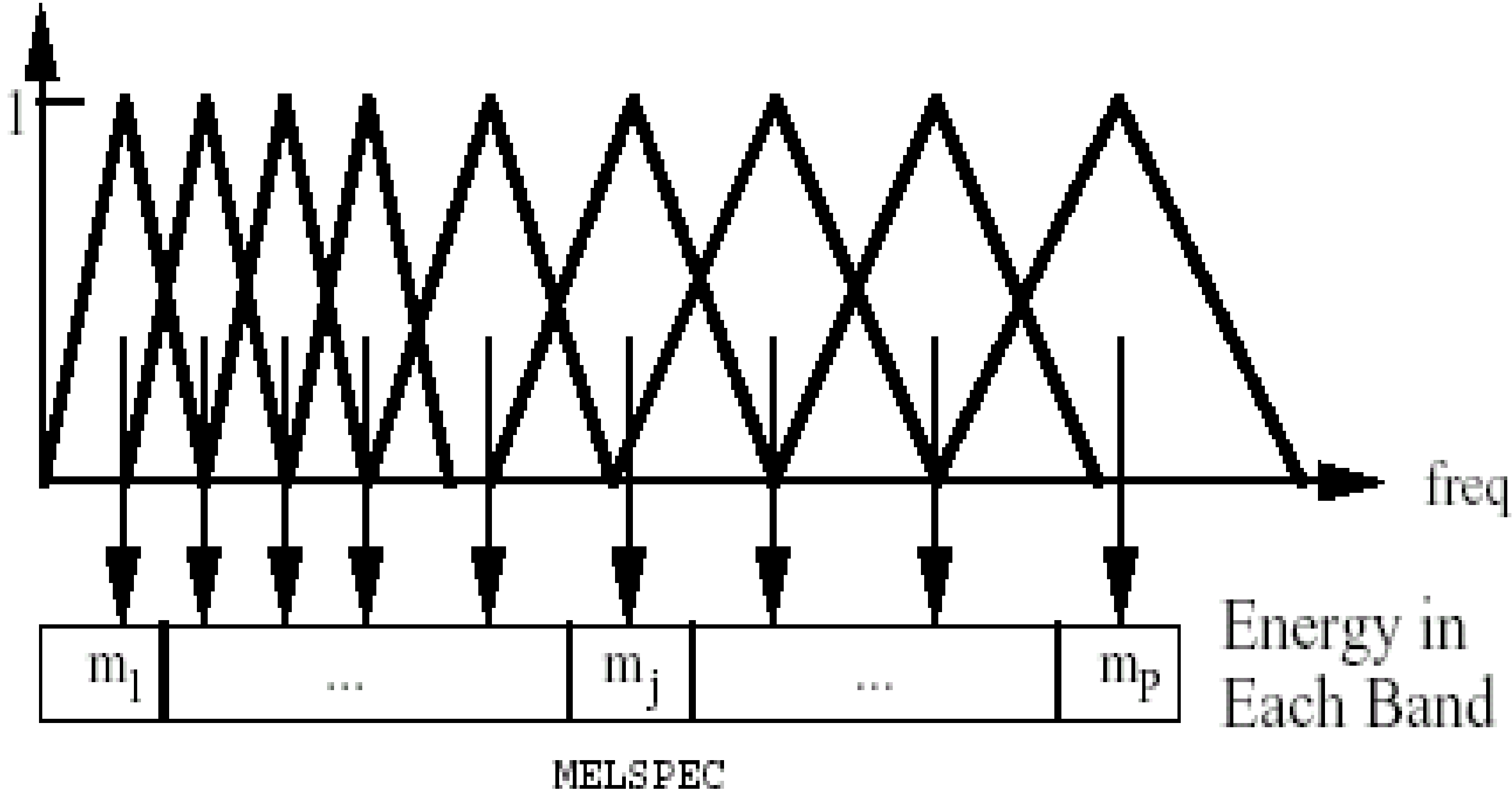
Mel-频率

频率(Hz)

Mel 滤波器组—参数选择

- 以采样率8kHz，帧宽30ms为例:
  - FFT窗宽：512
  - 滤波器个数：26 (通常24-40)
  - 滤波器频率应用范围（电话频带）：
    - 最高：3400Hz
    - 最低：300Hz

Mel 滤波器组—图示



对数能量

- 公式：

$$S[m] = \ln\left(\sum_{k=0}^{N-1} \left|X[k]\right|^2 H_m[k]\right) \qquad 0 \le m < M$$

- 应用：对噪音和谱估计误差有更好的鲁棒性

$$S[m] = \sum_{k=0}^{N-1} \ln\left(\left|X[k]\right|^2 H_m[k]\right) \qquad 0 \le m < M$$

倒谱参数

- **Discrete Cosine Transform (DCT)**

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos\left(\pi n\left(m+1/2\right)/M\right) \quad 0 \le n < M$$

- 倒谱维数：前**12**维

帧能量

- 公式：

$$E=\sum_{n=0}^{N-1}\left(x[n]-\overline{x}\right)^2$$

其中：

$$\overline{x}=\frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

- 应用：

$$E=\sum_{n=0}^{N-1}\left|x[n]-\overline{x}\right| \qquad E=\ln\left(\sum_{n=0}^{N-1}\left(x[n]-\overline{x}\right)^2\right)$$

动态参数

- 反映帧间相关信息

- 一阶差分：

$$\Delta S_t = S_{t+1} - S_{t-1}$$

$$\Delta^2 S_t = \Delta S_{t+m} - \Delta S_{t-m} \qquad m = 1或2$$

- 二阶差分：

S_t **--** 静态参数，包括倒谱和帧能量

倒谱均值归一化

- **Cepstrum Mean Normalization (CMN)**
  – 目的：消除信道带来的影响
  – 应用：**T**通常为整个词的特征帧数

$$\widehat{O}_t = O_t - \overline{O} \qquad \text{其中} \qquad \overline{O} = \frac{1}{T}\sum_{t=1}^{T} O_t$$

- 一个变形：

$$\widehat{O}_t[i] = \frac{O_t[i] - \overline{O}[i]}{\sigma[i]} \quad \text{其中} \quad \sigma[i] = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(O_t[i] - \overline{O}[i]\right)^2}$$

语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

识别模型

- 动态时间规整(DTW)
- 矢量量化(VQ)
- 隐马尔科夫模型(HMM)
- 神经网络(TDNN)
- 模糊逻辑算法

识别模型

- DTW(Dynamic Time Warping)

- VQ(Vector Quantization)

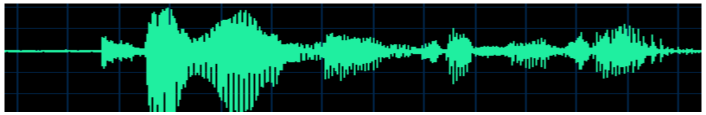
- HMM (Hidden Markov Models)

语音模型

- **DTW(Dynamic Time Warping)**
- VQ(Vector Quantization)
- HMM (Hidden Markov Models)

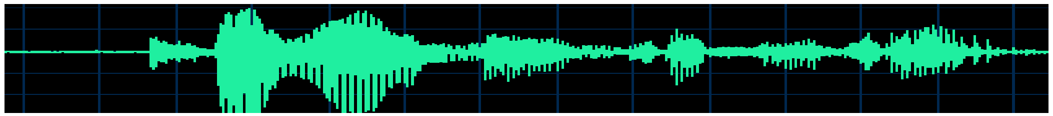
动态时间规整

- 语音识别模式匹配的问题—— 时间对准
  - 同一个人在不同时刻说同一句话、发同一个音，也不可能具有完全相同的时间长度
  - 语音的持续时间随机改变，相对时长也随机改变
- 方法**1**：线性时间规整



  - 均匀伸长或缩短
  - 依赖于端点检测

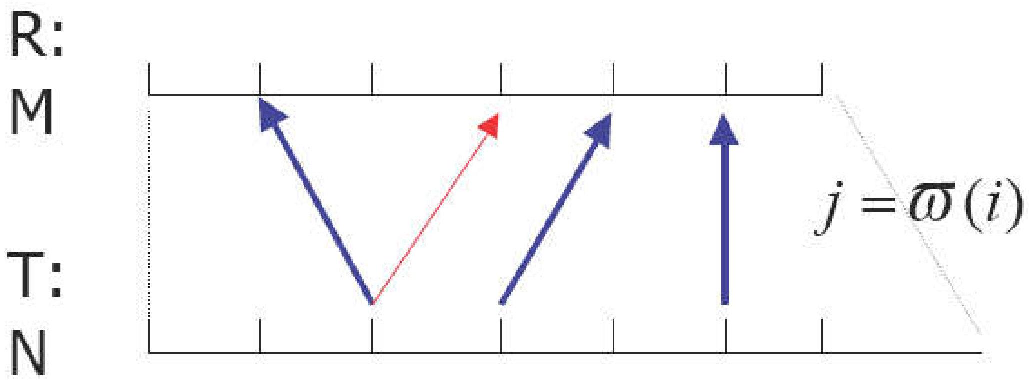
    通过时域分析进行，利用能量、振幅和过零率等特征
  - 缺点：仅扩展时间轴，无法精确对准



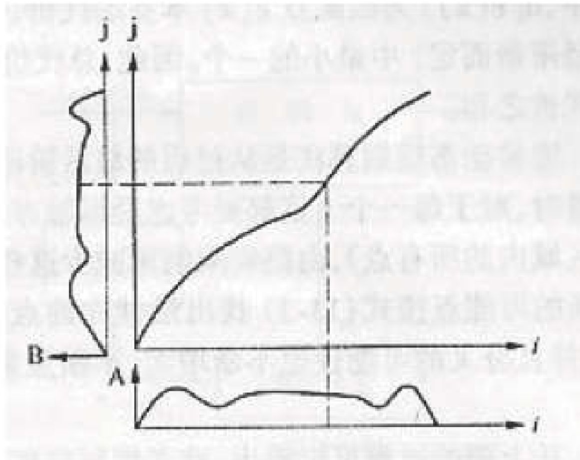
- 方法**2**：动态时间规整
  - **DTW－Dynamic Time Warping**

DTW的基本思想

- 一种非线性时间规整模式匹配算法
  - 将时间规整与距离测度结合起来，采用优化技术，以最优匹配为目标，寻找最优的时间规整函数**w(i)**，从而实现大小(长短)不同的模式的比较

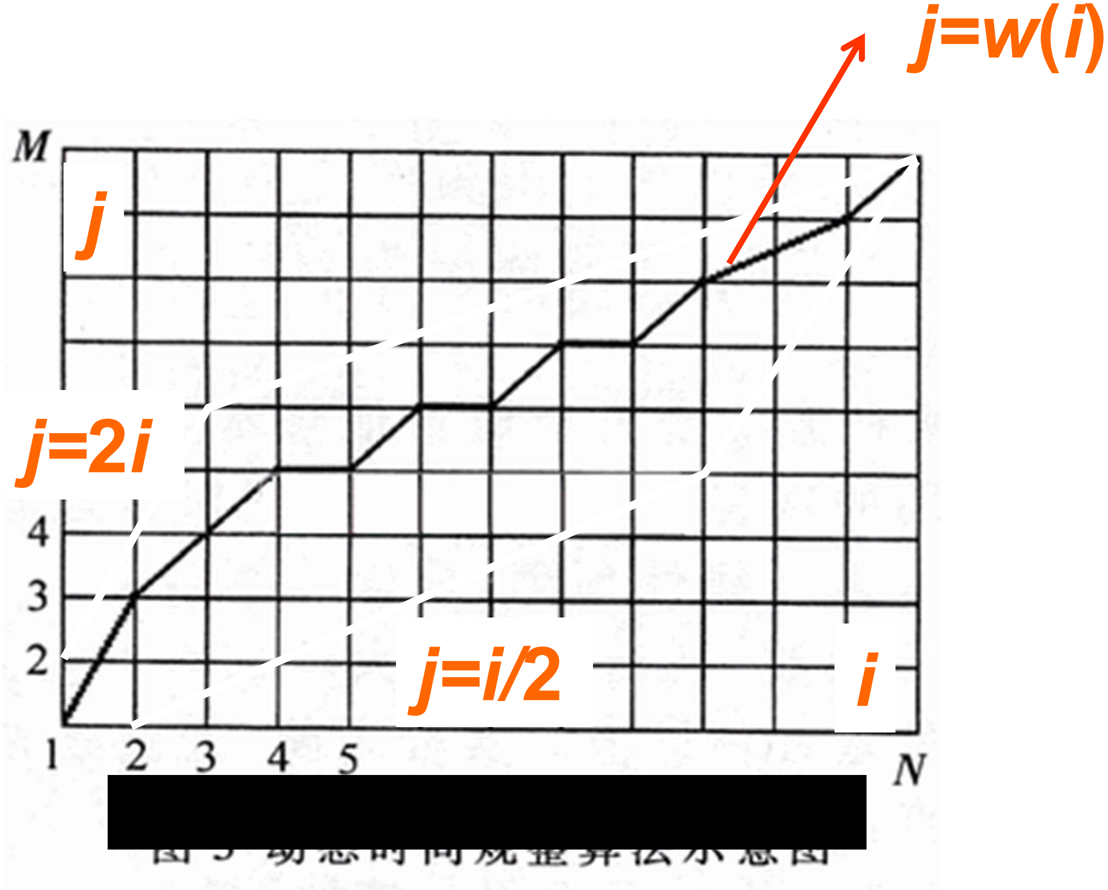


$$D=\min_{\varpi(i)}\sum_{i=1}^{M}d[T(i),R(\varpi(i))]$$



DTW的DP实现

- 动态规划 $D[c(k)]=d[c(k)] + \min D[c(k-1)]$
  - 搜索区域约束
    - 平行四边形
      - j=2i
      - j=i/2
  - 路径限制
    - W斜率
      - 0，1，2



[redacted]

DTW评价

- 适用场合
  - DTW适合于特定人、基元较少的场合
  - 多用于孤立词识别
- DTW的问题：
  - 运算量较大；
  - 识别性能过分依赖于端点检测；
  - 太依赖于说话人的原来发音；
  - 不能对样本作动态训练；
  - 没有充分利用语音信号的时序动态特性；

语音模型

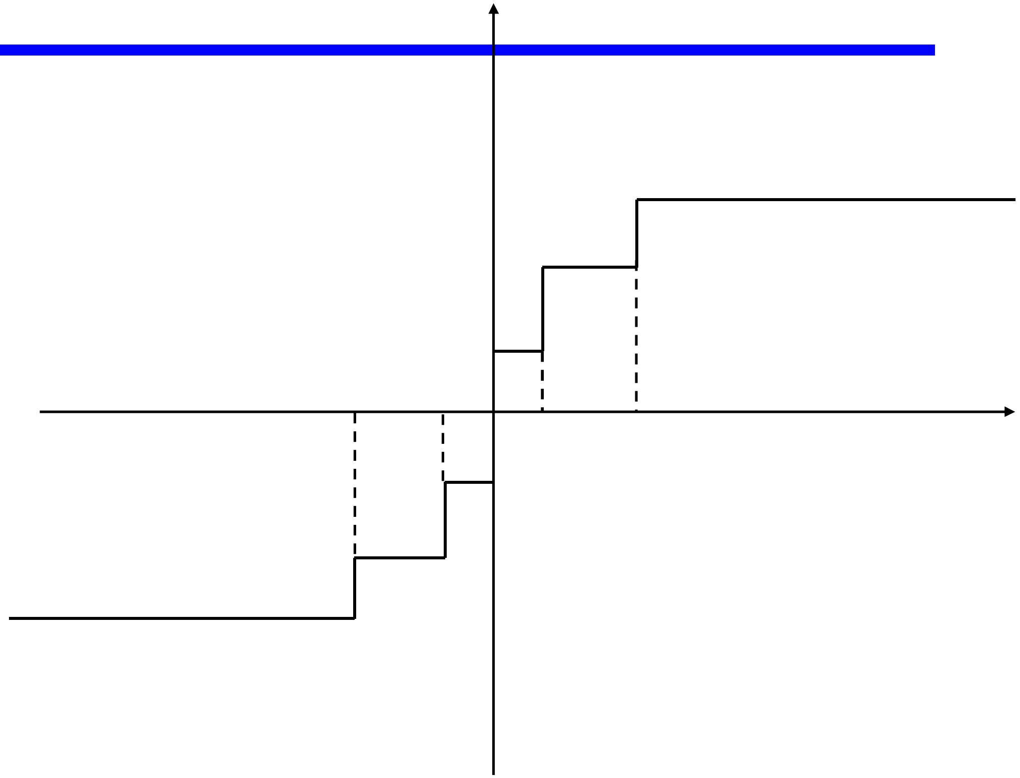
- DTW(Dynamic Time Warping)
- **VQ(Vector Quantization)**
- HMM (Hidden Markov Models)

VQ在语音分析中的应用

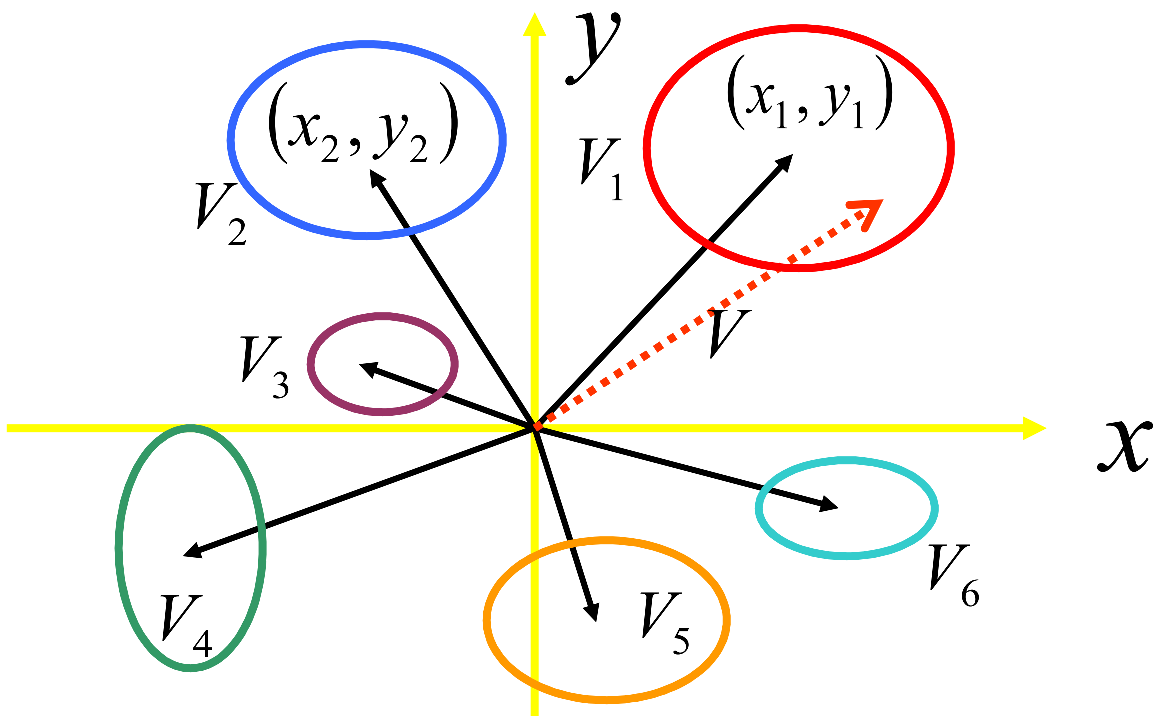
- 进入80年代以后，VQ技术引入语音处理领域，推动了语音技术发展，使之有了长足的进步
- 目前这项技术已经用于：
  - 语音识别；
  - 语音波形编码；
  - 线性预测编码；
  - ……

VQ基本概念

- 标量量化
  - 均匀
  - 非均匀
- 矢量/向量量化VQ
  - **Vector Quantization**
  - **VQ**就是将某一区域（范围）内的矢量归为某一类
- 矢量量化的基本要素
  - 聚类（**Cluster**）
  - 量化（**Quantization**）



VQ基本原理



上图的两维矢量空间里，存在6类矢量，每一类都有一个中心，称为室心(x_i, y_i)，每一室心对应一个码字矢量V_i=(x_i, y_i)，表征第i类矢量。集合{V_i}称为码本(codebook)。

VQ基本原理

- 任意一个矢量V应该归为哪一类，要看它是“靠近”哪一类矢量，或者说它离哪一个室心最“近”
  - 例如上图中虚线画出的矢量V最靠近V1，则将其规定为V1类，并用V1表示V，或者说V被量化为V1

- 把本来无限多的矢量只用有限个码字矢量来表示
  - 上例中为6个（只需要不到3个bits表示）
  - 假如码本中的码字矢量是有序的，则被量化的矢量可用码字序号来表示。因此，可以大大压缩信息量。

VQ基本原理

- 可见**VQ**技术包含两个步骤
  - 先要生成码本，这是将语音的特征矢量空间首先进行划分的过程——也称为聚类；
  - 将语音参数序列作为矢量，参照码本进行归类的过程——也称为量化。
- 在语音处理中
  - 通常把一帧**(**短时窗**)**语音对应的特征参数（**LPCC**，**MFCC…**）用矢量表示，并称为特征矢量或特征向量；

将训练矢量集TVS中的T个矢量用聚类算法，在总体失真最小的情况下划分为N个子类，在每类的中心设置一个码字，共得N个码字，组成一个码本

训练矢量集 → 聚类算法 → 码本

输入矢量 → 最小失真映射 → 编码

在已有码本的情况下，将矢量$V(t)$与码本$\{V_i\}$对照，按照最小失真原则去寻找与之最近邻关系的码字矢量V_k，并用其代表$V(t)$

VQ的数学描述

- 假定x是一个K维向量，其各维分量都是实值随机变量。在VQ中，向量x要映射成另一个K维向量y，这称作把x量化成y，写作y=VQ(x)。
- y在一个有限集中取值，这个有限集就是一个码本，我们记作CB={CW_i: 1≤i≤NC}，NC为码本大小。显然，VQ的过程就是样本空间x到有限空间CB的映射：

$$x \in X \subset E^K \to y = VQ(x) \in CB \subset E^K$$

VQ的数学描述

- 当把x量化为y后，它们之间存在一个量化失真或称距离度量**d(x, y)**
- 一个量化器**VQ(·)**称为最优的是说它是所有量化器中平均/期望量化失真最小的，其中**|X|**表示集合**X**中元素的个数。

$$D = \frac{1}{|X|} \sum_{x \in X} d\left(x, VQ(x)\right)$$

VQ应用

- 在实际的实现中，某一向量**x**对某一码本**CB**量化成CW_i后，为运算方便，只用该码字在**CB**中的编号**i**来表示量化结果。这样，**VQ**可以表示为：

$$c = VQ(x) = i \ \text{ iff } \ d(x, CW_i) \le d(x, CW_j), \text{对所有} j \ne i$$

或

$$c = VQ(x) = \arg\min_i d(x, CW_i)$$

参考文献

1.吴朝晖，杨莹春，说话人识别模型与方法，清华大学出版社，2009, 2

2. Roger Jang (張智星)

Audio Signal Processing and Recognition (音訊處理與辨識)

http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/index.asp

课后任务

- 阅读文献
  - L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing

    Ch4：4.2，4.3，4.4，4.5

    Ch5：5.1，5.6.3，5.7

    Ch9：9.1，9.2

语音模型

- DTW(Dynamic Time Warping)
- VQ(Vector Quantization)
- **HMM (Hidden Markov Models)**

HMM在语音识别中的应用

- 隐式马尔可夫模型(HMM)最开始出现在Baum等人的文章[Baum 72]中，紧接其后，它分别被CMU的Baker等人[Baker 75]和IBM的Bakis、Jelink等人[Bakis 76, Jelink 76]引入语音识别领域。在八十年代初美国Bell Lab的Rabiner等人提出了这一方法用于非特定人的语音识别[Rabiner 83]。
- HMM成为语音识别中一种很有效的技术，它不仅能用来作为(以音素、音节或词为单位的)语音产生的声学模型，而且能作为词法、语法、语义等高层次的语言模型，在很多领域都取得很大的应用。

Markov模型

- Andrei A. Markov
- Russian statistician
- 1856 – 1922

1.Markov propose Markov framework from 俄国文学家普希金名著<叶夫盖尼.奥涅金>

2.Baum and his collegue introduced and studied Hidden Markov Model in 1960s and 1970s

3.Became popular in 1980s. work very well for several important applications such as speech recognirion.

L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

4. David Haussler etc. described preliminary results on modeling protein sequence multiple alignments in 1992. HMM has been applied in Bioinformatics since then.

Markov Model

- 有N个可观测状态
  - $S_1, S_2 \ldots S_N$
- 存在一个离散的时间序列
  - $t=0,1, \ldots, T$
- 观测序列
  - $q_1, q_2, \ldots, q_T$
- 一阶马尔可夫假设
  - **当前状态qt只与前面相邻的一个状态qt-1有关，与其他状态无关**

$$q_t \in \{S_1, S_2, \ldots, S_N\}$$

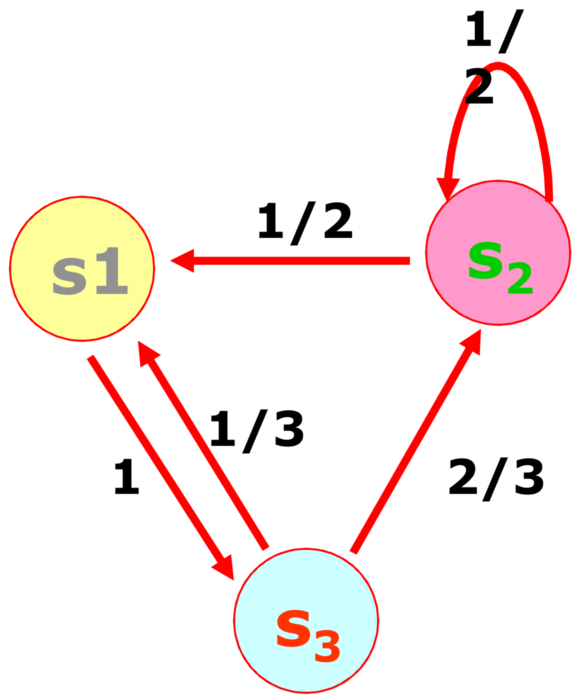
$$P(q_t = S_j \mid q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j \mid q_{t-1} = S_i)$$

一阶MM示例

$P(q_{t+1}=s_1|q_t=s_1) = 0$

$P(q_{t+1}=s_2|q_t=s_1) = 0$

$P(q_{t+1}=s_3|q_t=s_1) = 1$



$P(q_{t+1}=s_1|q_t=s_2) = 1/2$

$P(q_{t+1}=s_2|q_t=s_2) = 1/2$

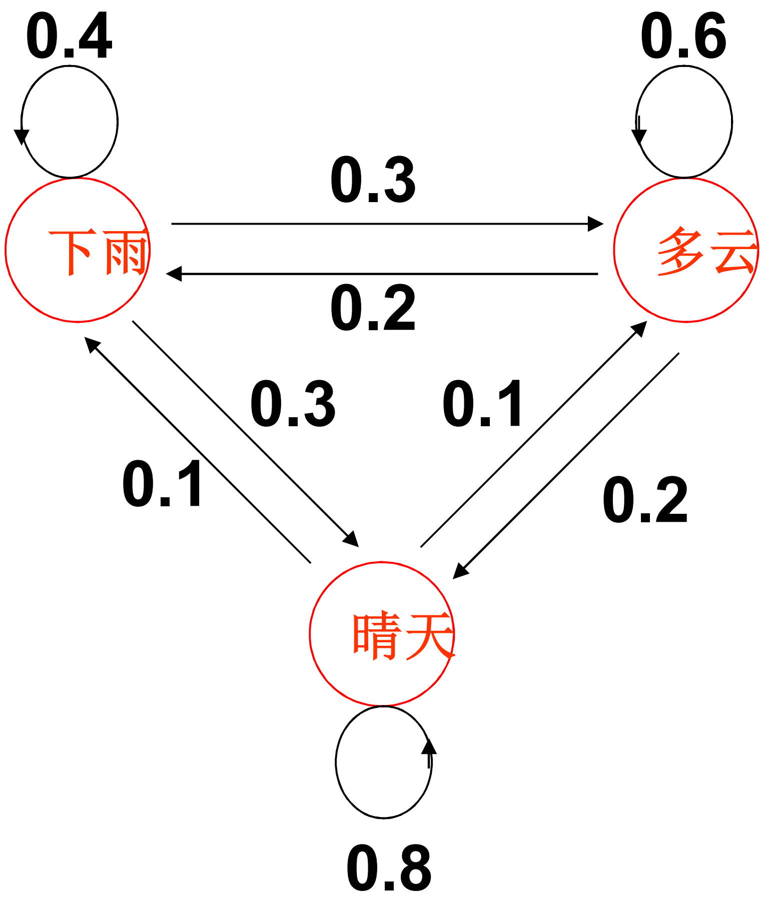
$P(q_{t+1}=s_3|q_t=s_2) = 0$

$P(q_{t+1}=s_1|q_t=s_3) = 1/3$

$P(q_{t+1}=s_2|q_t=s_3) = 2/3$

$P(q_{t+1}=s_3|q_t=s_3) = 0$

一阶MM实例



- 下雨---状态1--R
- 多云---状态2--C
- 晴天---状态3--S

$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

Markov Model

- **Markov Model**
- **状态转移矩阵A**

$$\lambda = \{\Pi, A\}$$

- **满足**

$$a_{ij} = P(q_t = S_j \mid q_{t-1} = S_i), 1 \le i, j \le N$$

$$a_{ij} \ge 0 \qquad \forall i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad \forall i$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{Nj} & \cdots & a_{NN} \end{bmatrix}$$

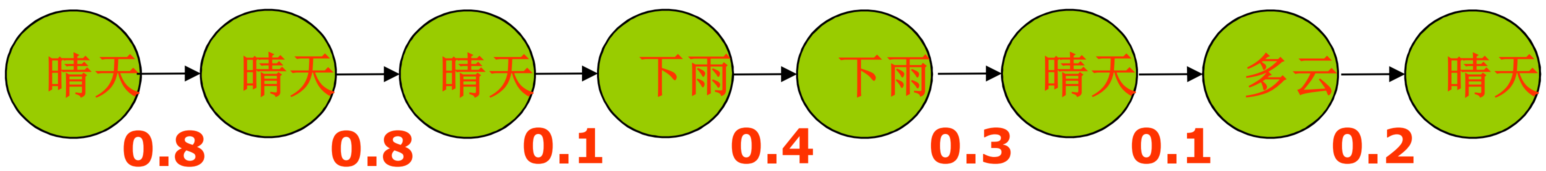
- **初始概率**

$$\Pi = \{\pi_i \mid i = 1,2,\ldots,N\}, \pi_i = P(q_1 = S_i)$$

例子

- **问题**
  - **今天是晴天，从今天开始连续8天的天气状况为“晴天-晴天-晴天-下雨-下雨-晴天-多云-晴天”的概率是多少?**
  - **计算P(SSSRRSCS| λ)**

$$\lambda = \{\Pi, A\}$$

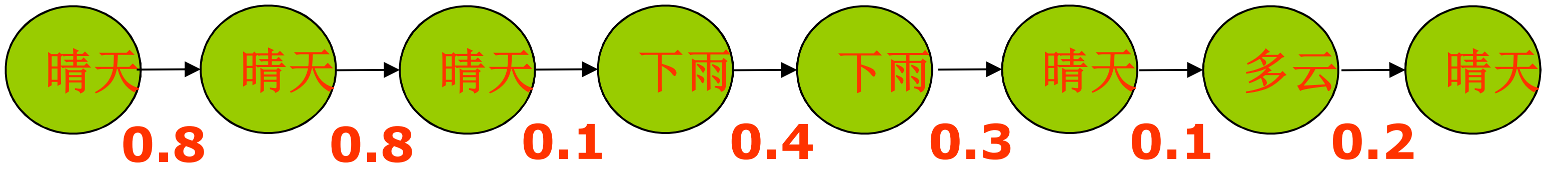


马尔可夫链规则

- **基本条件概率公式**
  - **P(A,B)=P(A|B)P(B)**
- **马尔可夫链规则**

$$
\begin{aligned}
&P(q_1, q_2, \ldots q_{\mathrm{T}}) \\
&= P(q_T \mid q_1, q_2, \ldots q_{\mathrm{T-1}}) P(q_1, q_2, \ldots q_{\mathrm{T-1}}) \\
&= P(q_T \mid q_{\mathrm{T-1}}) P(q_1, q_2, \ldots q_{\mathrm{T-1}}) \\
&= P(q_T \mid q_{\mathrm{T-1}}) P(q_{T-1} \mid q_{\mathrm{T-2}}) P(q_1, q_2, \ldots q_{\mathrm{T-2}}) \\
&= P(q_T \mid q_{\mathrm{T-1}}) P(q_{T-1} \mid q_{\mathrm{T-2}}) \ldots P(q_2 \mid q_1) P(q_1)
\end{aligned}
$$

例子



$$P(O \mid Model) = P([\mathrm{S,S,S,R,R,S,C,S}] \mid Model)$$

$$= P(\mathrm{S})P(S \mid S)^2 P(R \mid S)P(R \mid R)P(S \mid R)P(C \mid S)P(S \mid C)$$

$$= \pi_3 (a_{33})^2 a_{31} a_{11} a_{13} a_{32} a_{23}$$

$$= (1.0)(0.8)^2 (0.1)(0.4)(0.3)(0.1)(0.2)$$

$$= 1.536 * 10^{-4}$$

例：连续保持某状态的概率

- **例子**
  - 连续5天晴第6天阴/雨的概率是多少?
- **抽象**
  - 连续d个时间单位内保持某状态S_i，而到d+1时刻状态改变的概率

$$
\begin{aligned}
p_i(d) &= P(q_1 = i, q_2 = i, \ldots, q_d = i, q_{d+1} \neq i, \ldots) \\
&= \pi_i (a_{ii})^{d-1} (1 - a_{ii})
\end{aligned}
$$

例：连续保持某状态的概率

- 问题
  - 平均的连续晴天时间是多少天?
  - 平均的连续雨天时间是多少天?
  - 平均的连续阴天时间是多少天?
- 抽象
  - 求连读d天保持某状态i的期望

    雨天：$1/(1-a_{11})=1/(1-0.4)=1.67$天

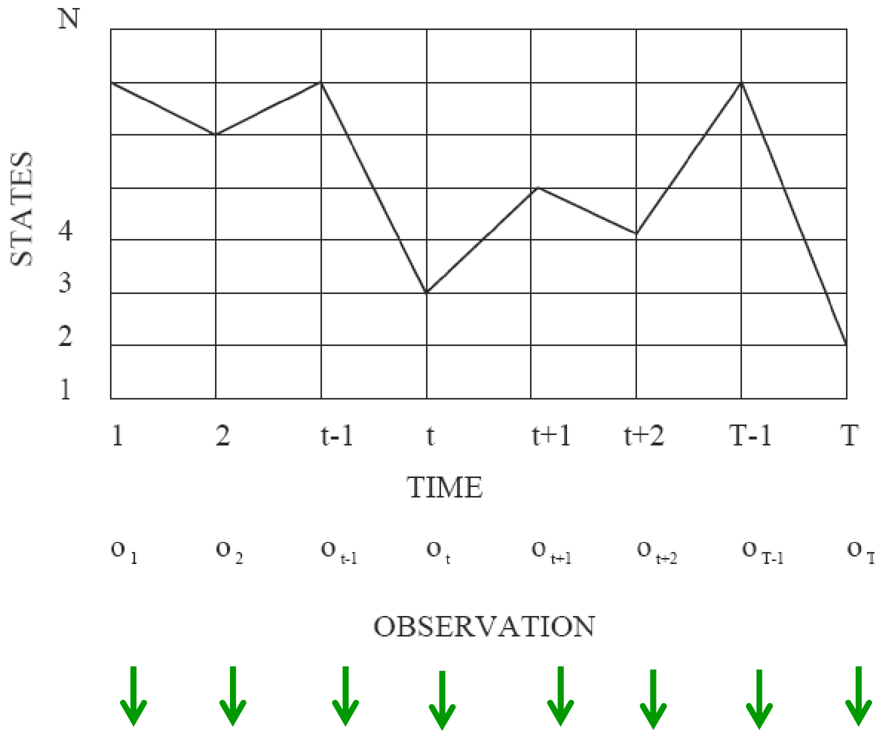
    阴天：$1/(1-a_{22})=1/(1-0.6)=2.5$天

    晴天：$1/(1-a_{33})=1/(1-0.8)=5$天

$$
\begin{aligned}
\bar{d}_i &= \sum_{d=1}^{\infty} d p_i(d) \\
&= \sum_{d=1}^{\infty} d(a_{ii})^{d-1}(1-a_{ii}) \\
&= (1-a_{ii}) \sum_{d=1}^{\infty} d(a_{ii})^{d-1} \\
&= (1-a_{ii}) \frac{\partial}{\partial a_{ii}} \sum_{d=1}^{\infty} (a_{ii})^d \\
&= (1-a_{ii}) \frac{\partial}{\partial a_{ii}} \left(\frac{a_{ii}}{1-a_{ii}} \right) \\
&= \frac{1}{1-a_{ii}}
\end{aligned}
$$

MM→HMM

- MM
  - 状态可见，状态即观测结果
- HMM
  - 状态不可见，但状态之间的转移仍然是概率的
  - 观测/输出结果是状态的概率函数



举例：从罐子里取颜色球

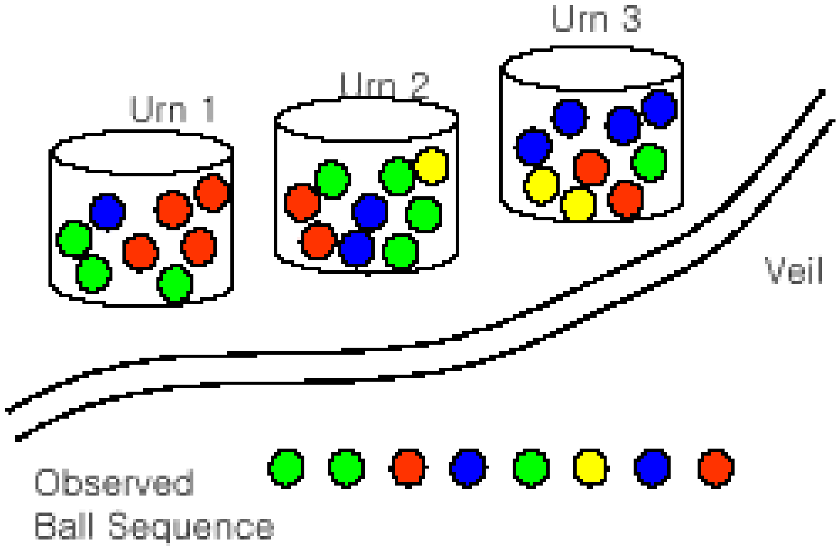
- N个罐子，内装各种颜色的球
- 共有M个不同颜色的球
- 每个罐子装的球的颜色分布可能不同
- 序列产生过程
  - 1. 随机选择一个初始罐子
  - 2. 从选中的罐子中随机取一个球，然后放回
  - 3. 根据一个与当前罐子有关的随机过程再选择一个罐子
  - 4. 重复2和3

实验内容：

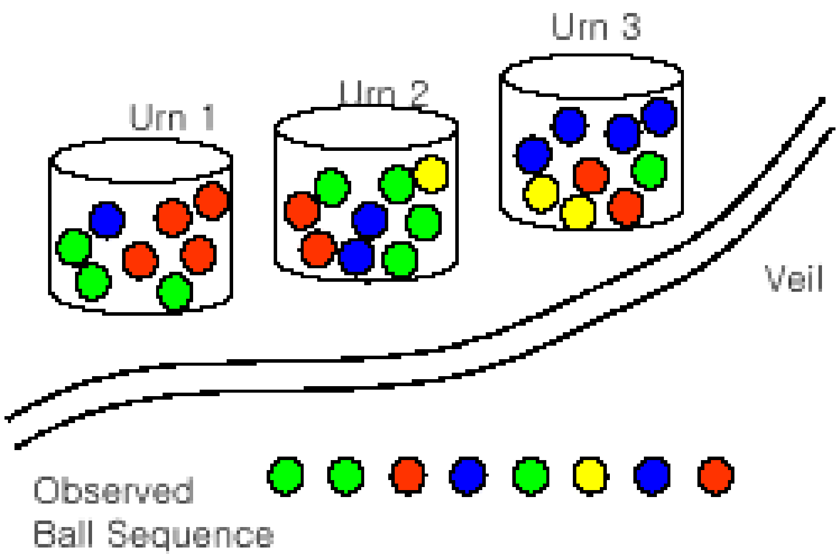
1. 根据某个初始概率分布，随机选择**N**个缸中的一个，例如第**i**个缸，再根据这个缸中彩色球颜色的概率分布，随机的选择一个球。记下球的颜色 ，再把球放回缸中。
2. 又根据缸的转移概率随机选出下一O_1个缸，比如第**j**个缸，再从缸中随机取出一个球，记下球的颜色 ，再把球放回缸中。
3. 一直进行下去。可以得到一个描述球的颜色的序列

$$O_2$$

$$O = O_1, O_2, ..., O_T$$



- $O = O_1, O_2, \ldots, O_T$ 是观察到的事件，称之为观察值序列。
- 缸之间的转移以及每次选取的缸被隐藏
- 从每个缸中选取的球的颜色并不是与缸一一对应，而是由该缸
中彩球颜色概率分布随机决定
- 每次选取哪个缸则是由一组转移概率决定



HMM分类

- 根据观察输出函数是基于VQ、连续密度还是二者的综合，HMM又分为：
  - 离散HMM (DHMM, Discrete HMM)；
  - 连续密度HMM (CDHMM, Continuous Density HMM, 简称CHMM)
  - 半连续HMM (SCHMM, Semi-Continuous HMM)
- 下面以DHMM为例介绍HMM

DHMM模型基本要素

- **1. 状态S_l (l=1,2,…, L)**
  - 所有状态构成了状态空间
  - x_n表示**n(=1,2,…, N)**时刻系统所处的状态

    $x_n \in \{S_1, S_2, \dots, S_L\}$
- **2. 初始状态概率**
  - $\pi=(\pi_1, \pi_2,\dots, \pi_L)$

表示**1(初始)时刻**系统处于状态S_l的概率

$$\pi_l = \Pr\left(x_1 = S_l\right),\quad l = 1, 2, \cdots, L$$

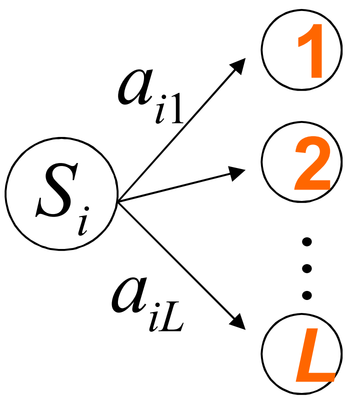
DHMM模型基本要素

- **3. 状态转移矩阵A ={a_{ij}}$_{LxL}$**
  - a_{ij}表示 **n**时刻系统处在S_i状态下，n**+1**时刻系统转移到S_j的概率（一步转移概率）

$$a_{ij} = \Pr\left(x_{n+1} = S_j \middle| x_n = S_i\right), \quad n \ge 1 \quad i, j = 1, 2, \cdots, L$$

$$\sum_{j=1}^{L} a_{ij} = 1, \qquad \forall i$$

  - 有了**A**，对长度为**N**的输出，系统可能产生L^N种互异的有限的状态序列，任何一种状态序列**X=(x$_1$, x$_2$, …, x$_N$)**的出现概率可写成：



$$\Pr\left(X \middle| \pi, A\right) = \pi_{x_1} a_{x_1 x_2} a_{x_2 x_3}, \cdots, a_{x_{N-1} x_N} = \pi_{x_1} \cdot \prod_{n=1}^{N-1} a_{x_n x_{n+1}}$$

DHMM模型基本要素

- **4.** 观察矢量序列**Y=(*y*$_1$, *y*$_2$, …, *y*$_N$)**
  - 任意时刻n，系统的状态x_n隐藏在系统内部，外界能得到一个观察矢量y_n
  - 如y_n具有离散分布：**n**时刻系统处于S_l状态下，观察矢量y_n的概率分布函数为

$$P_{x_n=S_l}(y_n)=\Pr(y_n|x_n=S_l),\quad n\geq 1\quad l=1,2,\cdots,L$$

  - 如y_n具有连续分布：**n**时刻系统处于S_l状态下，观察矢量y_n的概率密度函数为

$$P_{x_n=S_l}(y_n)=p(y_n|x_n=S_l),\quad n\geq 1\quad l=1,2,\cdots,L$$

DHMM模型基本要素

Pr和***p***只取决于$\boldsymbol{S_l}$，可直接用$\Pr_{S_l}(y)$或$p_{S_l}(y)$表示

有***L***个状态：$S_1, S_2, \cdots, S_L$

对应***L***个概率密度函数

$$B = \left(p_{S_1}(y), p_{S_2}(y), \cdots, p_{S_L}(y) \right)$$

或 ***L***个概率分布函数

$$B = \left(\Pr_{S_1}(y), \Pr_{S_2}(y), \cdots, \Pr_{S_L}(y) \right)$$

以后用***P***表示**Pr**或***p***

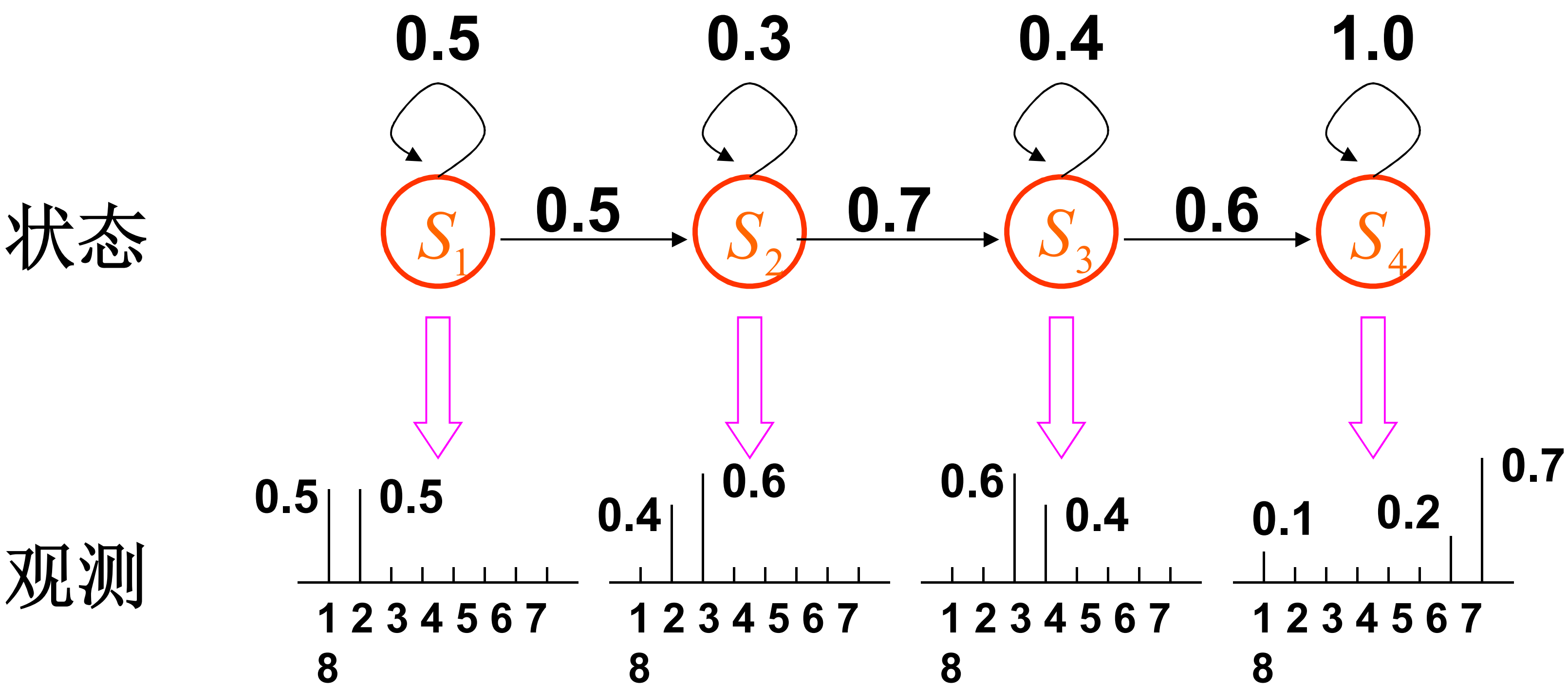
DHMM模型基本要素

- **HMM**模型常用**λ=(π, A, B)**来简记
- **HMM**系统从n**=1**时刻运行到N时刻，给出有N个随机矢量的矢量序列**Y=**($y_1, y_2, \ldots, y_N$)，称为观测矢量序列
- 该**HMM**产生**Y**的概率由**π, A, B**三者决定**(**由全概率公式**)**：对所有可能状态序列**X**的积分（求期望）

$$P\left(Y \middle| \pi, A, B\right) = \sum_{X}\left[\Pr\left(X\right)\cdot\left\{\prod_{n=1}^{N} P_{x_n = S_l}\left(y_n\right)\right\}\right]$$

举例

- **4个状态， 8个VQ码字，单链的拓扑结构**



初始状态概率

$$\pi = (1,\ 0,\ 0,\ 0)$$

A

$$\begin{array}{cc|cccc} & \text{到达} & S_1 & S_2 & S_3 & S_4 \\ \hline \text{出发} & S_1 & \boxed{0.5} & \boxed{0.5} & 0 & 0 \\ & S_2 & 0 & \boxed{0.3} & \boxed{0.7} & 0 \\ & S_3 & 0 & 0 & \boxed{0.4} & \boxed{0.6} \\ & S_4 & 0 & 0 & 0 & \boxed{1.0} \end{array}$$

B

$$\begin{array}{c|cccccccc} \text{状态} \backslash \text{码字} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline S_1 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_2 & 0 & 0.4 & 0.6 & 0 & 0 & 0 & 0 & 0 \\ S_3 & 0 & 0 & 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ S_4 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 \end{array}$$

HMM的三个基本问题

- 问题1：Training Problem (训练/建模问题)
  - 输入：给定若干个矢量序列$Y^{(m)}$——训练集
  - 目标：调整模型参数λ=(π, A, B)，使得该HMM产生训练集中所有矢量序列概率的(算术或几何或某种)平均值最大
  - 第1个问题的解决用于获得HMM模型的参数，以便建立模型

$$\prod_m P\left(Y^{(m)} \middle| \lambda\right) \to \max$$

HMM的三个基本问题

- **问题2: Evaluation Problem (估计问题)**
  - 给定一个观察矢量序列Y——待识别语音，和一个HMM模型λ=(π, A, B)，如何计算该模型λ产生该序列Y的概率P(Y|λ)?

$$P(Y|\pi, A, B) = \sum_{X}\left[\Pr(X)\cdot\left\{\prod_{n=1}^{N} P_{x_n=S_l}(y_n)\right\}\right]$$

  - 该问题的解决可以用于根据观察序列，计算每个模型的得分，从而实现对未知语音的识别，适用于孤立词识别系统

HMM的三个基本问题

- 问题3：Hidden State Sequence Uncovering Problem (状态序列选择问题)
  - 给定一个观察矢量序列Y和一个HMM模型λ=(π, A, B)，如何选择一个在某种意义下最优的状态序列(S_1, S_2, …, S_N)？比如

$$X^* = \arg\max_{X} \left[\Pr(X) \cdot \left\{ \prod_{n=1}^{N} P_{x_n = S_l}(y_n) \right\} \right]$$

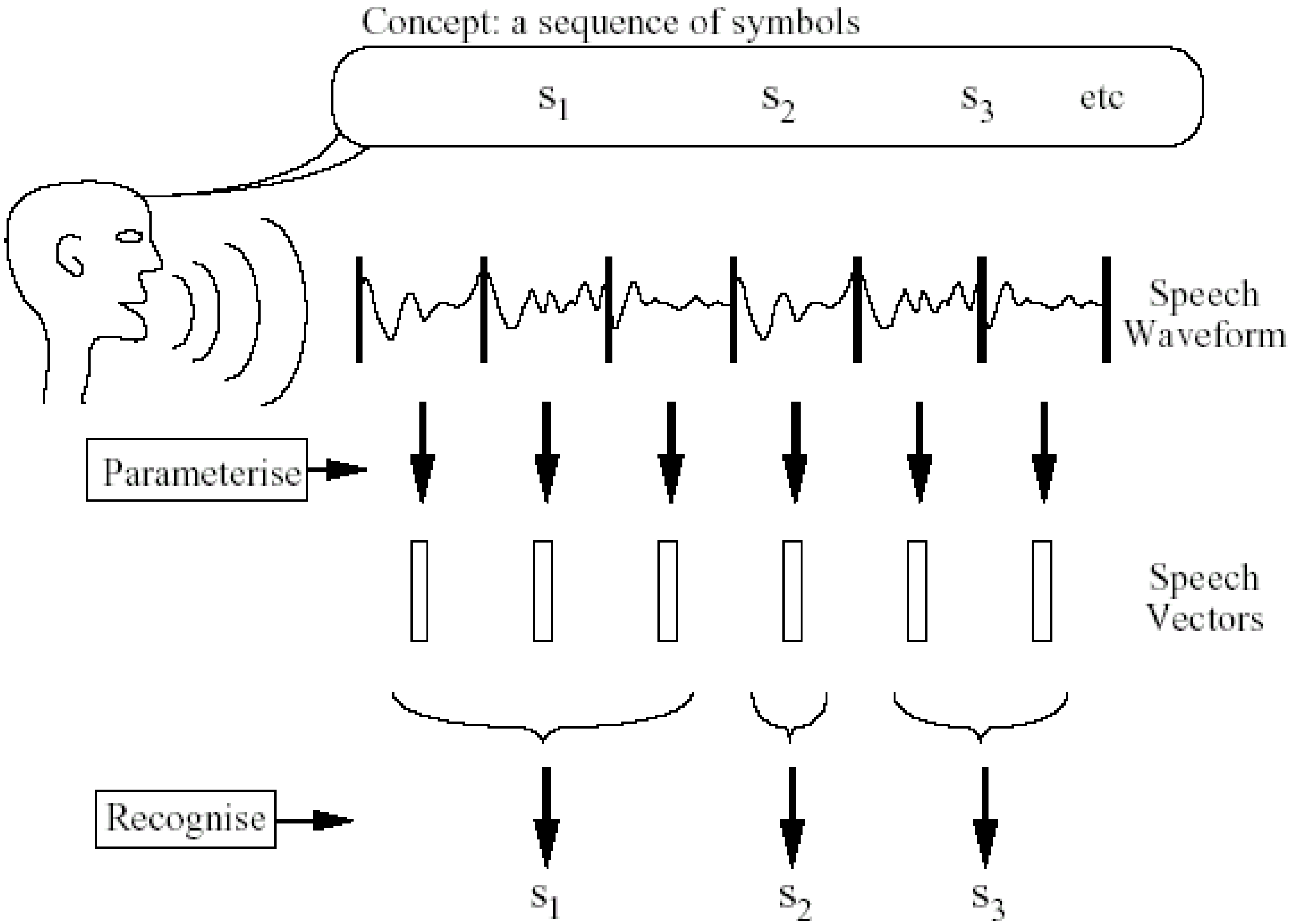
  - 也称为解码/识别问题，其解决使HMM在连续语音识别中发挥作用

HMM三个基本问题的求解

- 问题1：训练问题
  - 根据已知观测确定模型参数
  - Baum-Welch算法
- 问题2：估计问题
  - 根据已知模型求未知观测似然度
  - Forward-Backward算法
- 问题3：最优路径搜索、状态序列分割问题
  - Viterbi算法

语音信号与HMM

- 语音信号的短时平稳假设
  - 特征序列可以分成若干段(状态)
  - 在每个状态内观察特征是服从相同的分布的
- 可以用两个过程去刻画：
  - 状态之间的转移（隐藏的）
  - 在特定状态下的特征输出（可见的）


Concept: a sequence of symbols
s_1 s_2 s_3 etc
Speech Waveform
Parameterise
Speech Vectors
Recognise
s_1 s_2 s_3

两个基本假定

- **问题简化的数学模型**
  - 当前状态只与前一状态有关，而与更早的状态无关（无后效性或马尔可夫性）
    - 一阶马尔可夫链(过程)

$$\Pr\left(x_{n+1}=S_{n+1}\middle|x_1=S_1,x_2=S_2,\cdots,x_n=S_n\right)=\Pr\left(x_{n+1}=S_{n+1}\middle|x_n=S_n\right)$$

  - 当前状态下的输出只与当前状态有关，而与其他任何状态均无关
    - 状态间输出的独立性

HMM的三个基本问题求解的应用

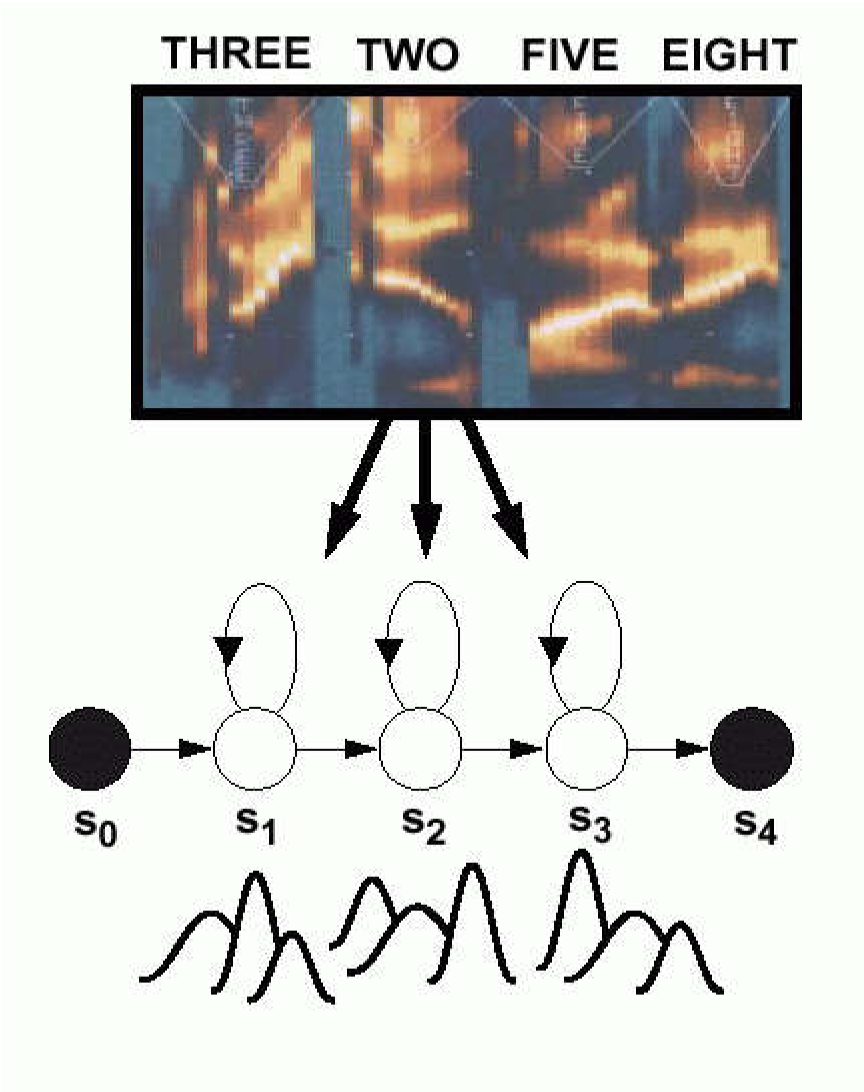
- 问题1：Training Problem
  - 给定每个基元(词/音素)的m个训练样本(表示为m个特征矢量)，学习得到基元的HMM模型
- 问题2: Evaluation Problem (估计问题)
  - 给定某测试样本Y，可以给出HMM模型所有可能状态序列产生Y的似然概率
- 问题3：解码问题/状态序列选择问题
  - 给定某测试样本Y，可以给出HMM模型产生Y的似然概率最大的状态序列/路径

HMM(Hidden Markov Model)

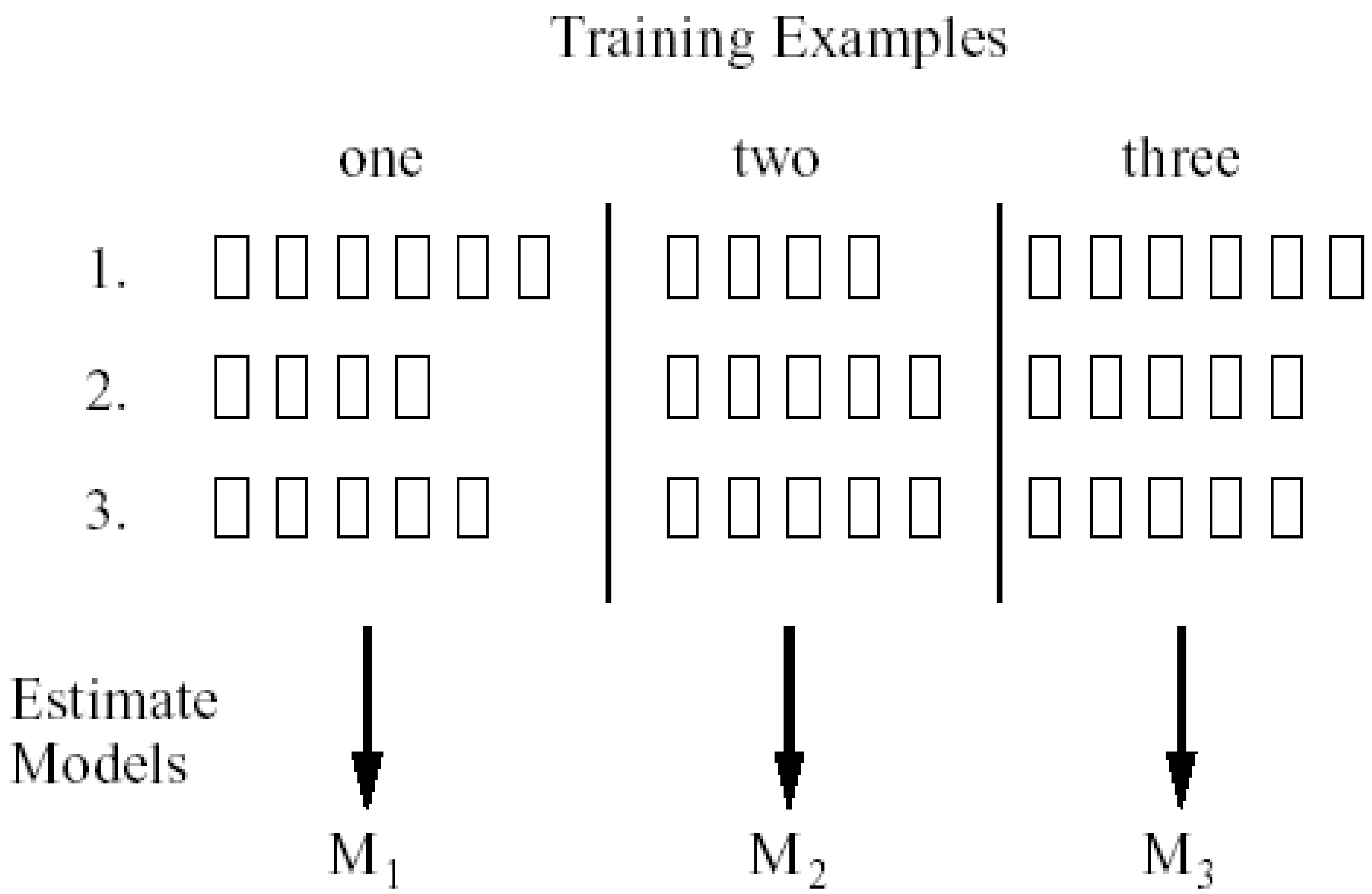
- HMM是描述说话人发音的统计模型

- 高斯混合密度分布刻划了语音状态（如音素）以及语音状态之间的时序变迁的统计规律

- 基本算法：
  -评估： 给定观测向量Y和模型，利用前向后向（Forward-Backward）算法计算得分；
  -匹配：给定观测向量Y，用Viterbi算法确定一个优化的状态序列；
  -训练：用Baum-Welch 算法（类似于EM）重新估计参数，使得分最大。

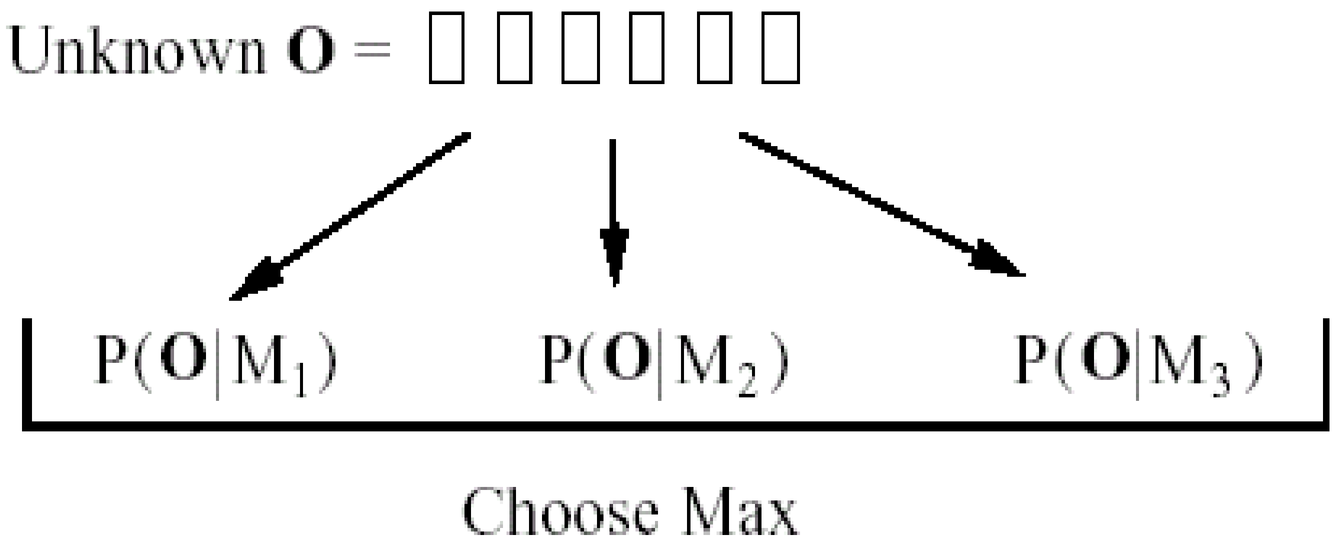


训练


Training Examples
one
two
three
1.
2.
3.
Estimate Models
M_1
M_2
M_3

HMM用于孤立词识别

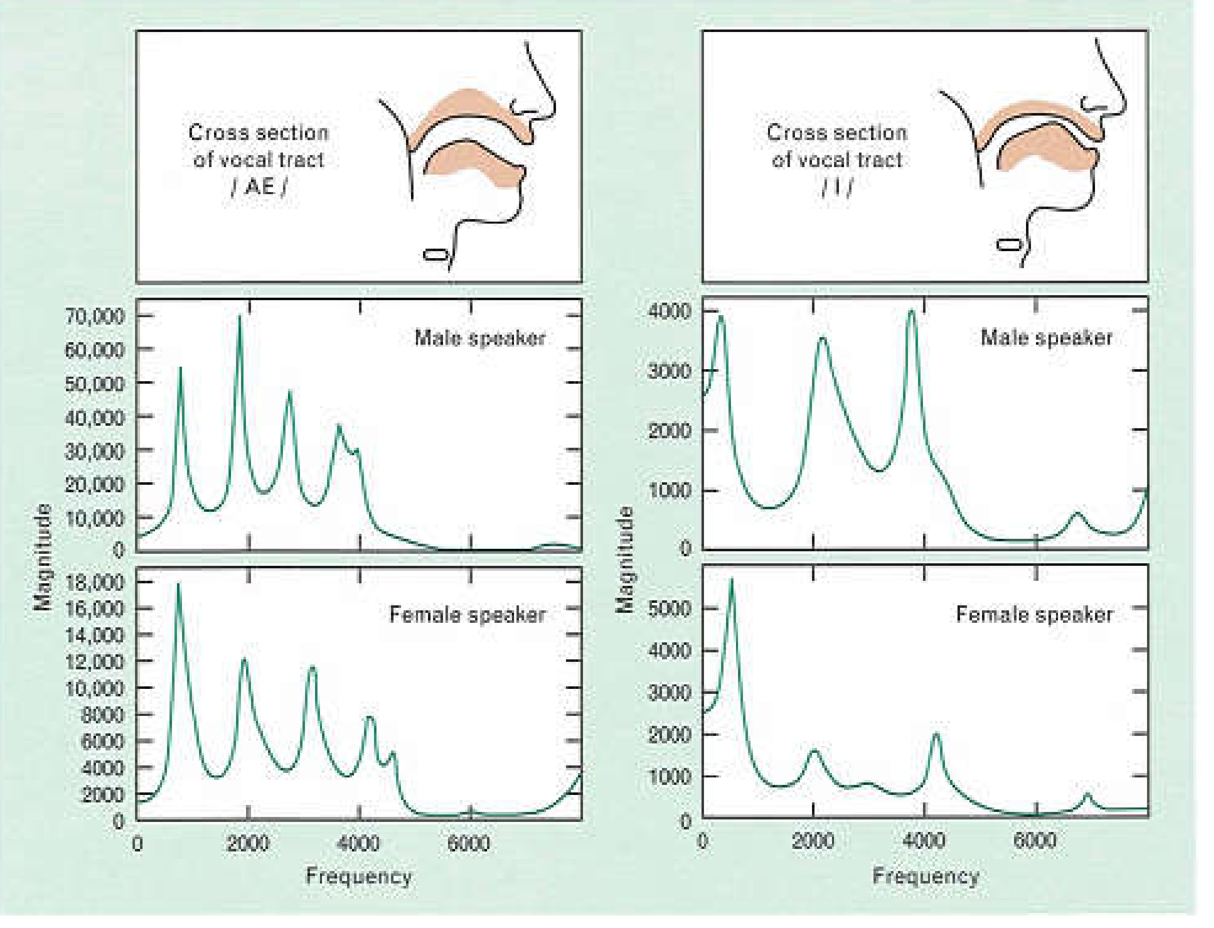
- 计算观察特征矢量序列Y与任意一个模型λh∈{λ1, λ2, …, λH}之间的匹配得分，并认为argmax P(Y| λh)对应的就是识别结果。



讲述提纲

- 说话人识别

Problem Statement



Problem Statement

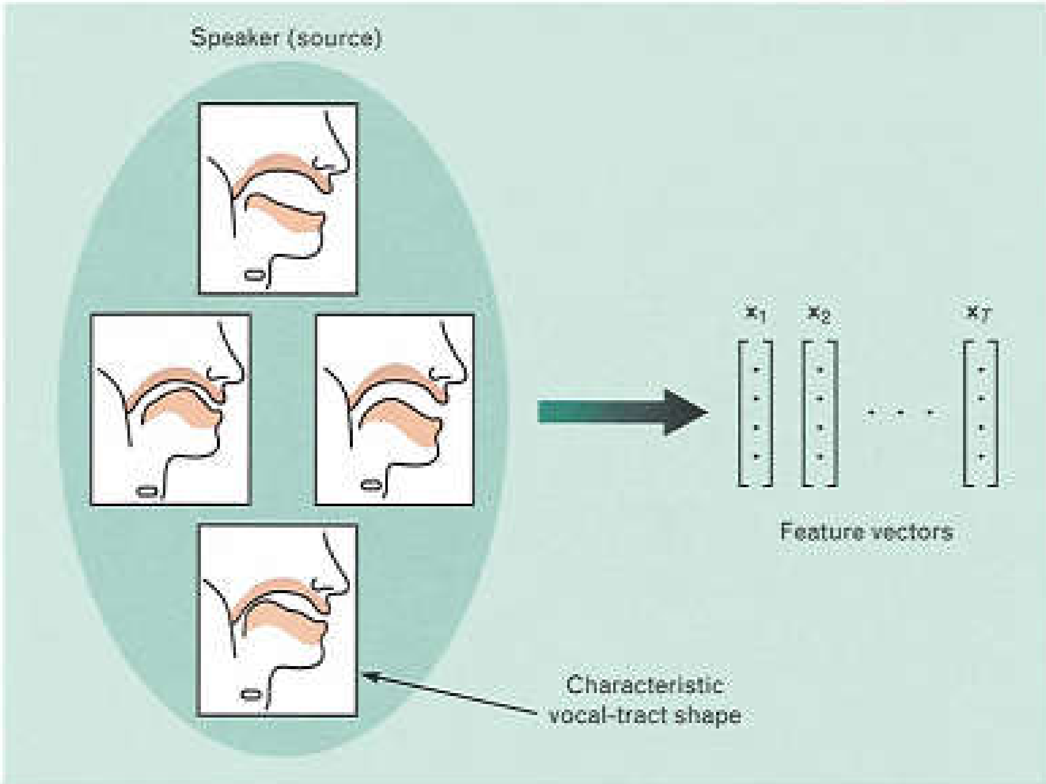
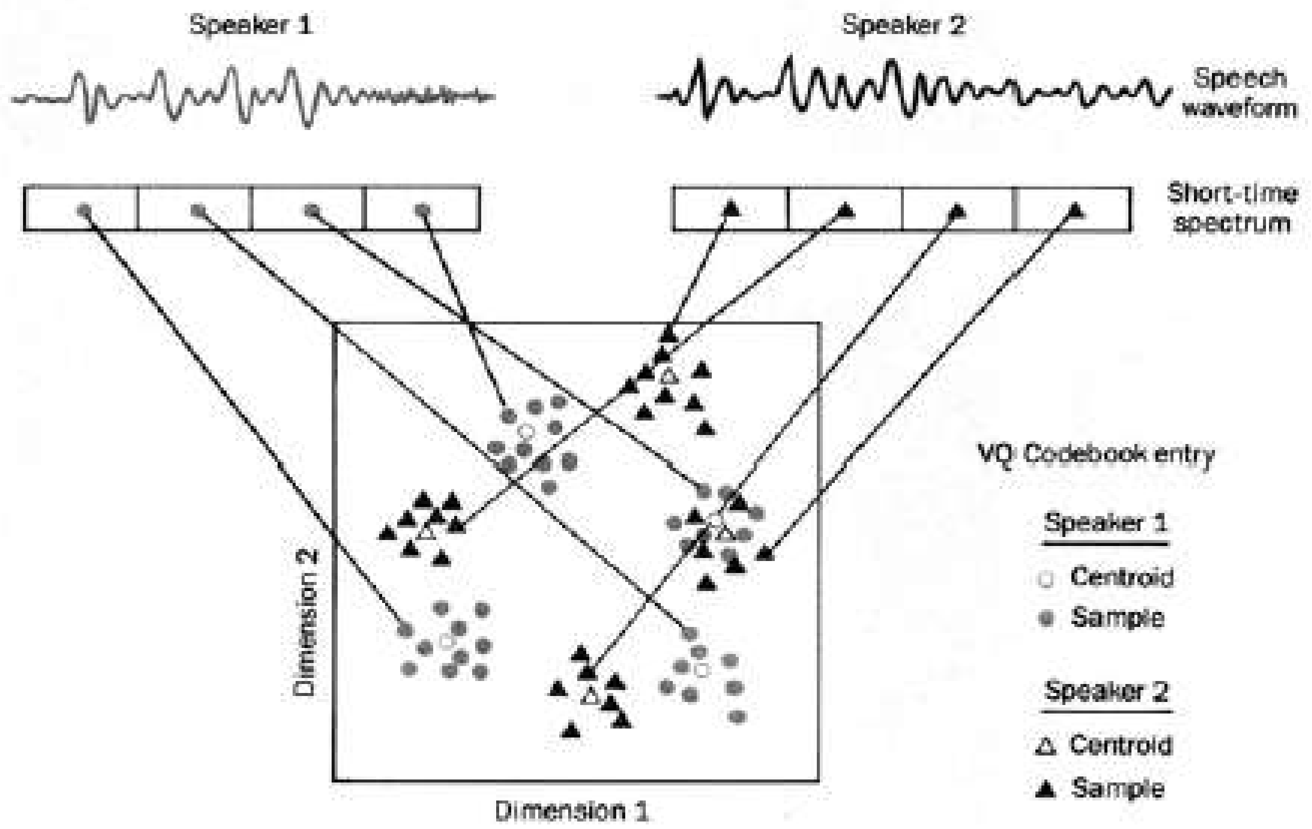


FIGURE 5. Statistical speaker model. The speaker is modeled as a random source producing the observed feature vectors. Within the random source are states corresponding to characteristic vocal-tract shapes.

VQ (Vector Quantization)

An Example of Speaker Modeling

[F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition," *AT&T Technical Journal*, Vol. 66, pp. 14-26, Mar/Apr 1987]



(Text-Independent) Speaker Modeling Revisited

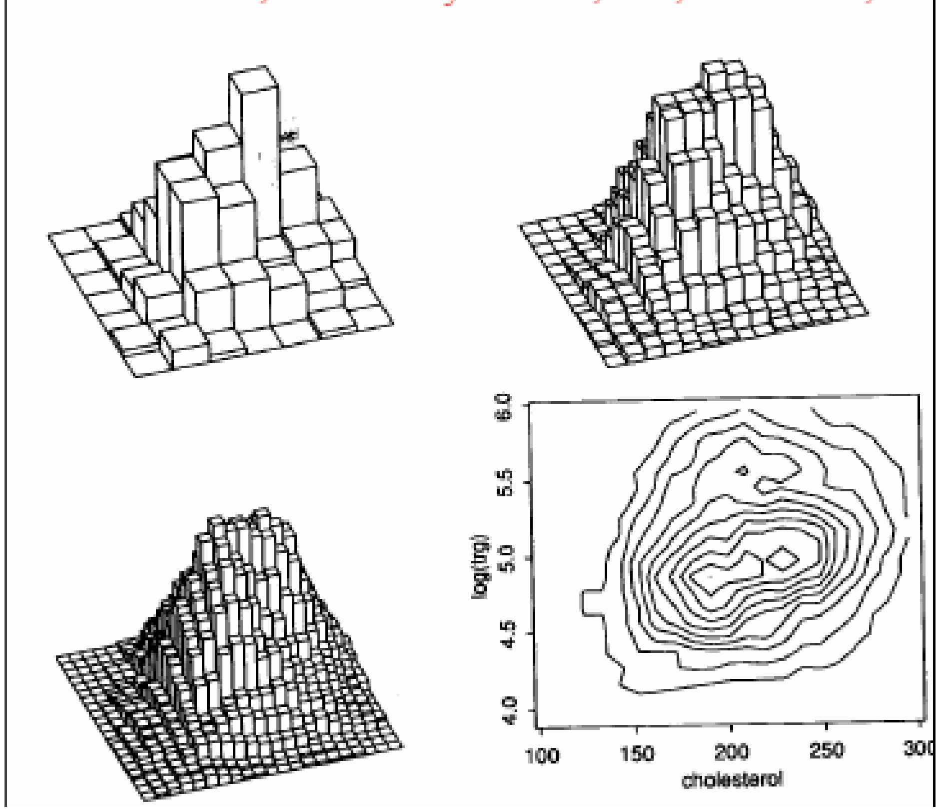
- The purpose of speaker modeling is to characterize the source that generated the feature vectors
- Since the same source (in this case, the speaker) produces the vectors, these should follow some probability distribution that is characteristic for this source.
- In statistics and pattern recognition, the problem of estimating the probability distribution of observation vectors is known as ***density estimation***
- The more training data (vectors) we have, the better estimate we have
- A simple example of a density estimator : ***histogram***

1. Select the number of histogram bins M
2. Divide the data range $[x_{min}, x_{max}]$ into M bins of width $(x_{max} - x_{min}) / M$
3. For each bin i, the density estimate is given by $p_i = N_i / N$

Examples of Histogram Density Estimates

(2-dimensional data)

[D.W.Scott, *Multivariate Density Estimation – Theory, Practise, and Visualization*, John Wiley & Sons, Inc., New York, 1992.]

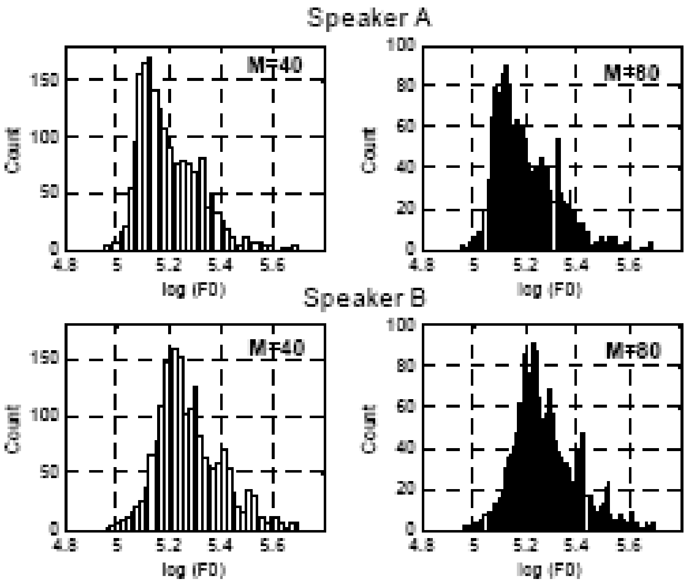


Beyond the Histogram Density Estimator

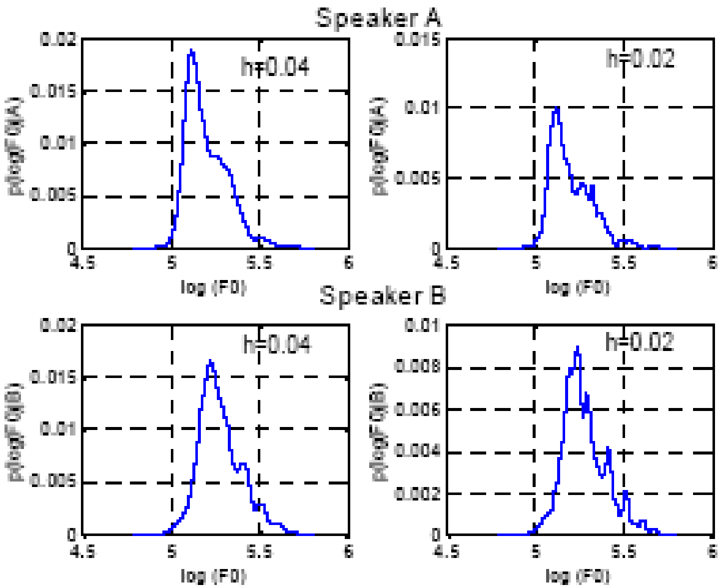
- The histogram method is simple and intuitive, but not the best one: for instance, the density estimates that it generates are "ragged" which violates the nature of our data (continuous in most cases)

- A more general approach that generates smoother density estimates is so-called ***kernel density estimator***

Histogram method



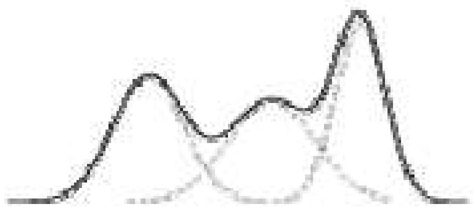
Kernel density method



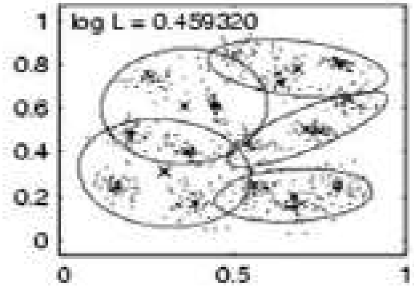
The Gaussian Mixture Model (cont.)

Examples:

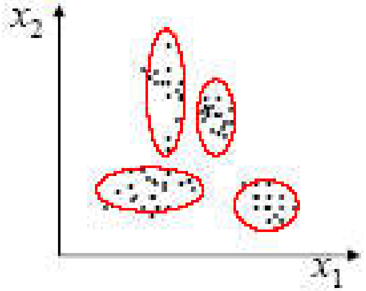
Dimensionality=1, K=3 :



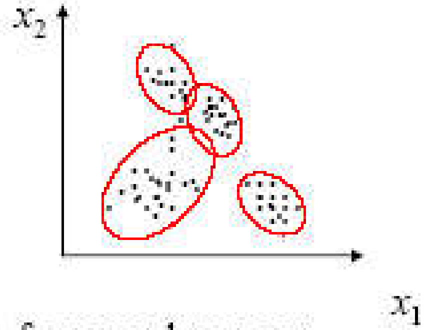
Dimensionality=2, K=5 :



Diagonal covariance GMM :



Full covariance GMM :



- Usually the diagonal covariance GMM is used, for several reasons:
  - Some typically used features have rather low inter-correlations (or, they should have at least! Remember p. 54, requirement (6).)
  - Computational complexity, memory usage
  - Numerical stability
  - Ease of implementation

GMM(Gaussian Mixture Model)

a set of acoustic feature vectors representing an utterance: $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, the likelihood of those feature vectors given a GMM model λ is the following:

$$p(\vec{x}|\lambda) = p(\vec{x}|w_i, \vec{\mu}_i, \Sigma_i) = \sum_{i=1}^{M} w_i p_i(\vec{x}) \tag{1.1}$$

where

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x}-\vec{\mu}_i)}$$

and

$$\sum_{i=1}^{M} w_i = 1$$

Here, there are M gaussians in the GMM and each mixture i is associated with a weight w_i, a mean $\vec{\mu}_i$, and a covariance Σ_i.

GMM(Gaussian Mixture Model)

With the GMM as the basic speaker representation, we can then apply this model to specific speaker-recognition tasks of identification and verification. The identification system is a straightforward maximum-likelihood classifier. For a reference group of S speaker models $\{\lambda_1, \lambda_2, \ldots, \lambda_S\}$, the objective is to find the speaker identity $\hat{s}$ whose model has the maximum posterior probability for the input feature-vector sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. The minimum-error Bayes' rule for this problem is

$$\hat{s} = \arg\max_{1 \le s \le S} \Pr(\lambda_s | X) = \arg\max_{1 \le s \le S} \frac{p(X|\lambda_s)}{p(X)} \Pr(\lambda_s).$$

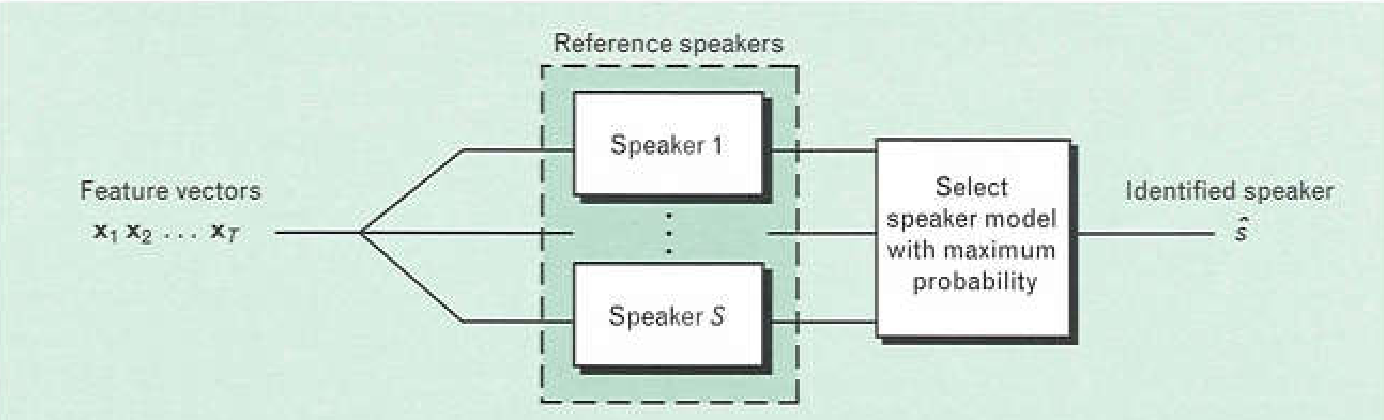
GMM(Gaussian Mixture Model)

Assuming equal prior probabilities of speakers, the terms $\Pr(\lambda_s)$ and $p(X)$ are constant for all speakers and can be ignored in the maximum. By using logarithms and assuming independence between observations, the decision rule for the speaker identity becomes

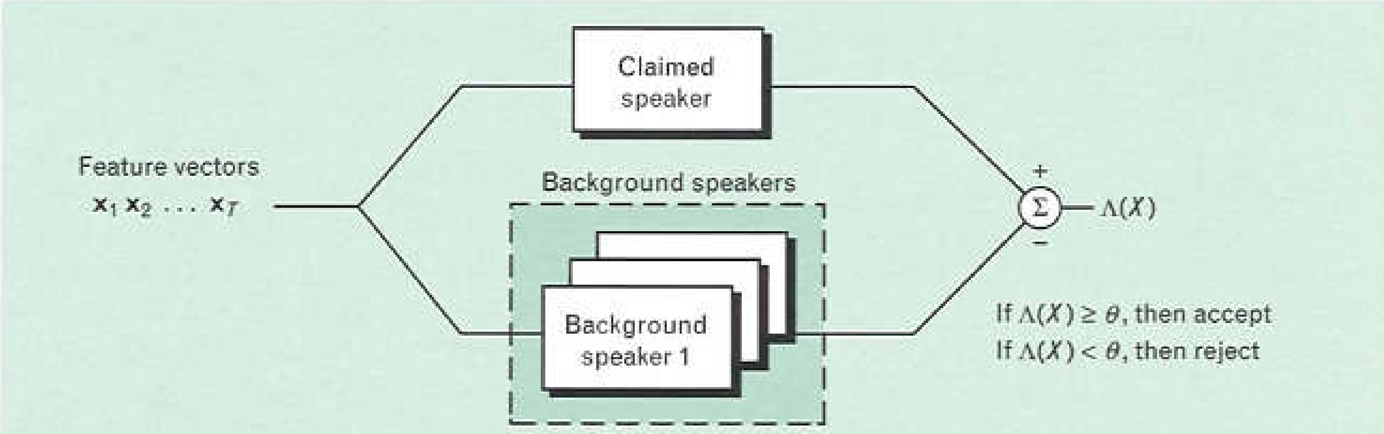
$$\hat{s} = \arg \max_{1 \le s \le S} \sum_{t=1}^{T} \log p(\mathrm{x}_t | \lambda_s),$$

in which T is the number of feature vectors and $p(\mathrm{x}_t | \lambda_s)$ is given in Equation 1. Figure 6(a) shows a diagram of the speaker-identification system.

GMM(Gaussian Mixture Model)



(a)



(b)

GMM(Gaussian Mixture Model)

E-step: Given the following statistic for mixture i of a GMM model:

$$P(i|\vec{x}_t) = \frac{w_i p_i(\vec{x}_t)}{\sum_{j=1}^{M} w_j p_j(\vec{x}_t)}$$

we have:

$$n_i = \sum_{t=1}^{T} P(i|\vec{x}_t)$$

$$E_i(\vec{x}) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|\vec{x}_t)\vec{x}_t$$

$$E_i(\vec{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|\vec{x}_t)\vec{x}_t^2$$

GMM(Gaussian Mixture Model)

M-step: New model parameters obtained using statistics computed during E-step as follows:

$$\hat{w}_i = [\alpha_i n_i/T + (1-\alpha_i)\hat{w}_i]\,\gamma$$

$$\hat{\vec{\mu}}_i = \alpha_i E_i(\vec{x}) + (1-\alpha_i)\vec{\mu}_i$$

$$\hat{\vec{\sigma}}_i^2 = \alpha_i E_i(\vec{x}^2) + (1-\alpha_i)(\vec{\sigma}_i^2 + \vec{\mu}_i^2) - \hat{\vec{\mu}}_i^2$$

where the scale factor γ ensures that the new weights $\hat{w}_i$ sum to unity. In addition, α is the relevance factor, controlling the balance between the UBM prior and new estimates obtained in the E-step.

GMM-UBM (Universal Background Model)

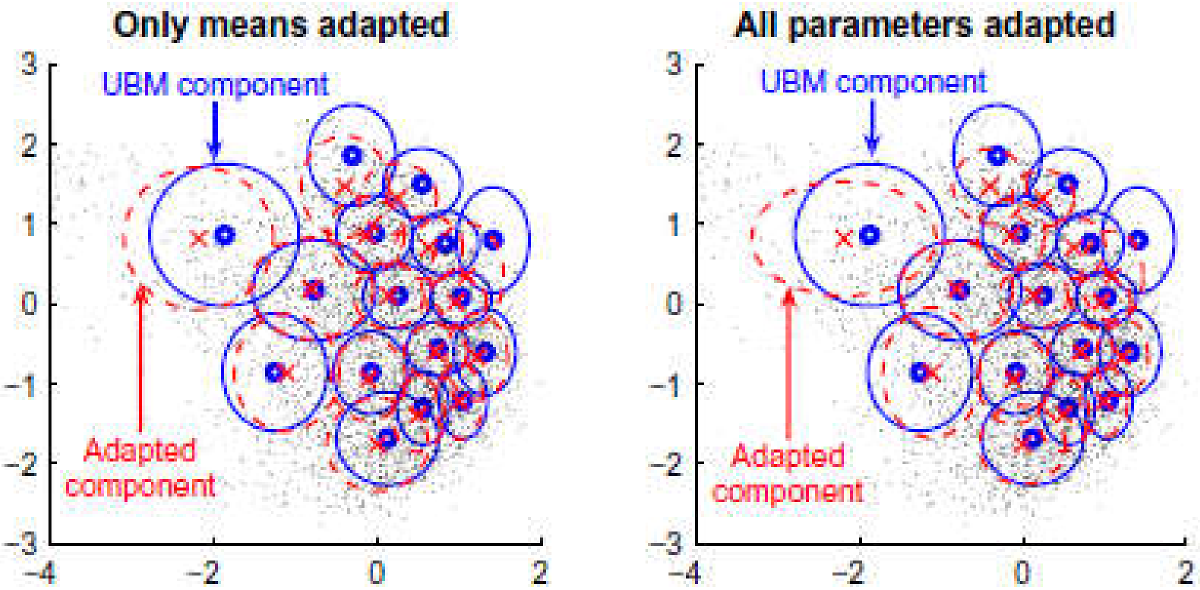


Fig. 8. Examples of GMM adaptation using *maximum a posteriori* (MAP) principle. The Gaussian components of a universal background model (solid ellipses) are adapted to the target speaker's training data (dots) to create speaker model (dashed ellipses).

GMM-UBM (Universal Background Model)

$$\mathrm{LLR}_{\mathrm{avg}}(\mathcal{X}, \lambda_{\mathrm{target}}, \lambda_{\mathrm{UBM}}) = \frac{1}{T}\sum_{t=1}^{T}\{\log p(\boldsymbol{x}_t|\lambda_{\mathrm{target}}) - \log p(\boldsymbol{x}_t|\lambda_{\mathrm{UBM}})\}, \qquad (13)$$

The use of a common background model for all speakers makes the match score ranges of different speakers comparable.

Score normalization

the "raw" match score is normalized relative to a set of other speaker models known as cohort.

$$s' = \frac{s - \mu_I}{\sigma_I} \qquad (15)$$

zero normalization ("Z-norm")
test normalization ("T-norm")

参考文献

1.吴朝晖，杨莹春，说话人识别模型与方法，清华大学出版社，2009, 2

2.杨莹春，陈华，吴飞，视音频信号处理，浙江大学出版社，待出版

3. Roger Jang (張智星)

Audio Signal Processing and Recognition (音訊處理與辨識)

http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/index.asp

课后任务

- 阅读文献
  - *Douglas A. Reynolds. Automatic Speaker Recognition Using Gaussian Mixture Speaker Models*
  - **L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition". *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.**（可选读）