

视听觉信号处理 Visual and Auditory Signal Processing



1

为什么讲“视听觉”？
为什么视觉和听觉一起讲？
为什么有三门课程？

4

视听觉信息理解系列课程



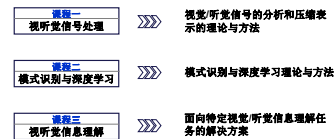
7

语音学概要



2

视听觉信息理解系列课程



5

课程设计的特点

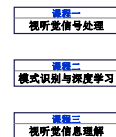
- 所涉及的信号处理、模式识别等理论在系列课程内讲述，使学生带着问题去学习，并很快学以致用；
- 同时包含视觉和听觉两方面内容，对比交融，使学生能够对人工智能技术形成更深层次的、更普适的理解和认识；
- 可以有更多的时间（一年半），逐层递进地以任务驱动的方式完成对知识的学习。

8

- 任课教师： 郑铁然
- 办公室地址： 综合楼603
- 办公室电话： 86417981-11
- 手机： 13313655979
- QQ： 2350562164
- Email: zhengtieran@hit.edu.cn

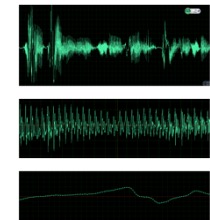
3

视听觉信息理解系列课程

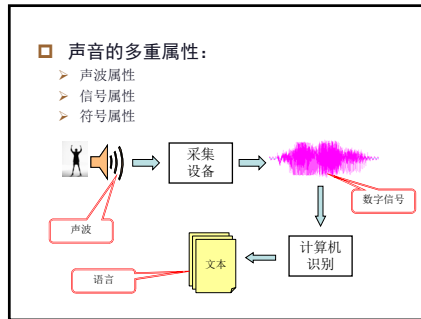


6

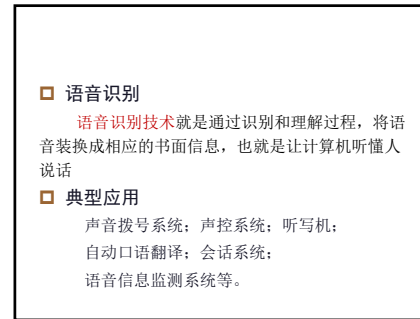
- 声音：
 - 语音
 - 音乐
 - 其它声音等
- 声音是一维信号
- 声音的感知和认知难度大
 - 承载语言
 - 差异性大
 - 环境声音始终存在
 - 声音的叠加性



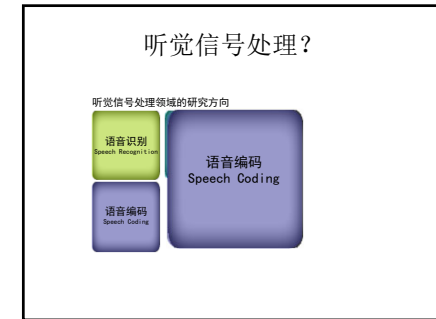
9



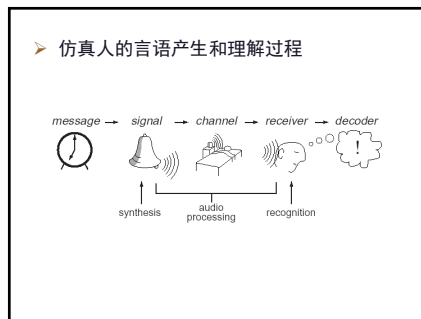
10



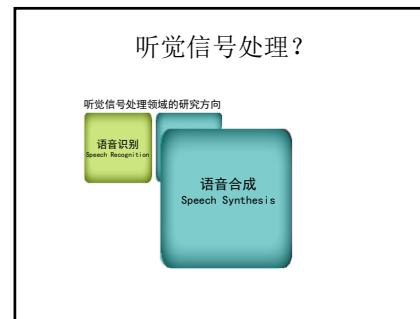
13



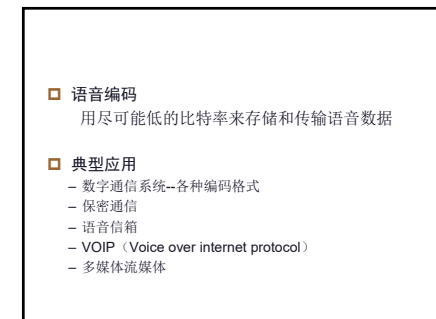
16



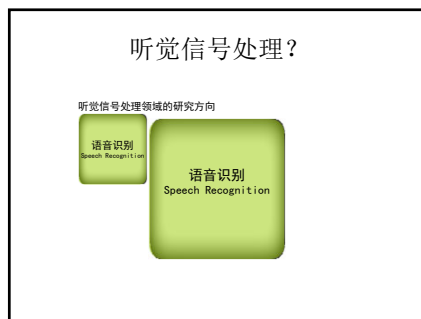
11



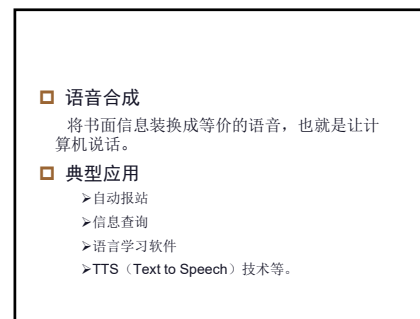
14



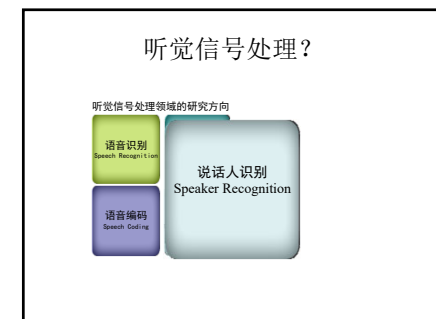
17



12



15



18

□ 说话人识别

根据语音辨认说话人的身份

□ 典型应用

- 声控门锁
- 电子商务
- 司法鉴定
- 情报搜集。

19

听觉信号处理？



22

语音识别技术

- 语音是语言的载体，是思维的依托，是人类有别于其它生物的重要标志，是智能的终极体现
- **语音识别技术**就是通过识别和理解过程，将语音转换成相应的书面信息，也就是让计算机听懂人说话
- 语音交互技术（语音识别+语音合成）将引领人类进入下一个交互时代



25

听觉信号处理？



20

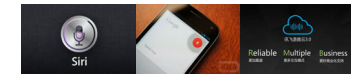
听觉信号处理？



23

历史和现状

- 语音识别研究起步很早，1952年Bell实验室研制成功能识别十个英语数字的识别器Audry系统。
- 几十年来，取得了许多重要的研究成果。
- 目前，正处于语音技术产品化的新浪潮之中。
- 然而，其性能还远未达到理想的水平



26

听觉信号处理？



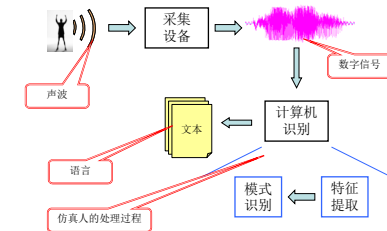
21

举 例：

语音识别技术的研究内容和发展轨迹

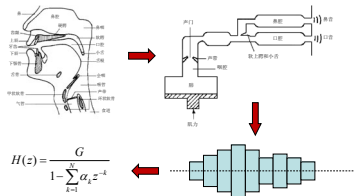
24

语音识别技术的框架



27

特征提取环节的仿真-发生机理



28

模式识别环节的仿真

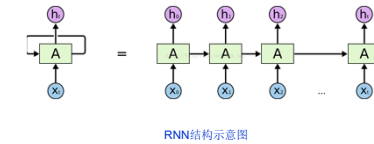
对大脑的仿真最难。一些里程碑式的研究工作：

- ◆ 基于**隐马尔可夫模型(HMM)**的语音识别技术
- ✓ 90年代初，**李开复**在CMU搭建了基于HMM的非特定人连续语音识别系统**SPHINX**；
- ✓ **三层结构**：声学语音层、词层、句法层；
- ✓ HMM不但对**声学内容**进行统计建模，也对其**时序变化**进行统计建模。

31

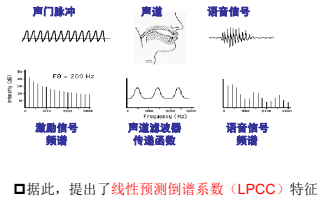
模式识别环节的仿真

◆ 基于RNN+DNN的语音识别技术



34

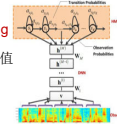
特征提取环节的仿真-发生机理



29

模式识别环节的仿真

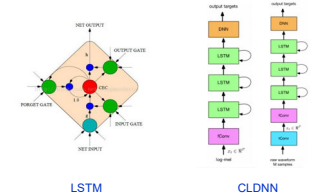
- ◆ 基于**深度学习(Deep Learning)**的语音识别技术
- ✓ 2011年，微软研究院**俞栋**等提出了基于DNN+HMM的语音识别方法
- ✓ 其训练分成**Pre-training**和**Fine-tuning**
- ✓ 引入RBM构建DBN，作为DNN的初值



32

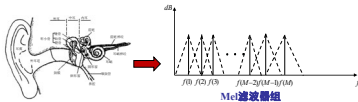
模式识别环节的仿真

◆ 基于RNN+DNN的语音识别技术



35

特征提取环节的仿真-感知机理



30

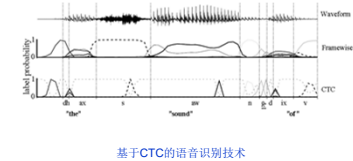
模式识别环节的仿真

- ◆ 基于**深度学习(Deep Learning)**的语音识别技术
- ✓ 近几年，许多深度学习技术被提出来，并被应用到语音识别技术中；
- ✓ 基于循环神经网络(RNN, Recurrent Neural Networks)的语音识别技术；
- ✓ 基于**LSTM** (Long-Short Term Memory) 网络的语音识别技术。

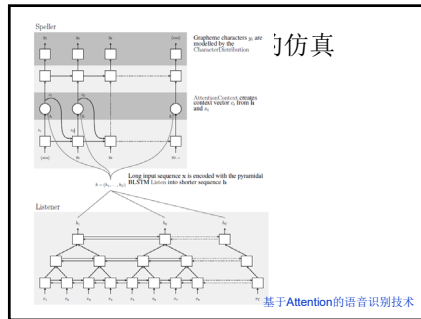
33

模式识别环节的仿真

◆ End to end语音识别技术



36



37

语音的声学特性

➤ 语音是以声波的方式在空气中传播。声波是一种纵波，它的振动方向和传播方向是一致的。

➤ 声波的基本物理量——频率：

- 单位时间内，声波的周期数。（波长 传播速度）
- 人耳对于声波频率高低的感受与实际频率近似成对数关系。
- 基频：60Hz~500Hz

40

语音的数字信号表示

43

- ### 参考书
- 韩纪庆、张磊、郑铁然《语音信号处理》清华大学出版社。
 - Huang X D, Acero A, Hon H, et al. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. New Jersey: Prentice Hall PTR, 2001
 - 余栋, 邓力. 解析深度学习——语音识别实践, 电子工业出版社, 2016
 - 易克初、田斌等《语音信号处理》国防工业出版社。
 - 蔡莲红、黄德智等《现代语音技术 基础与应用》清华大学出版社
 - Rabiner L, Juang B H. Fundamentals of Speech Recognition. New Jersey: Prentice Hall PTR, 北京: 清华大学出版社, 1999

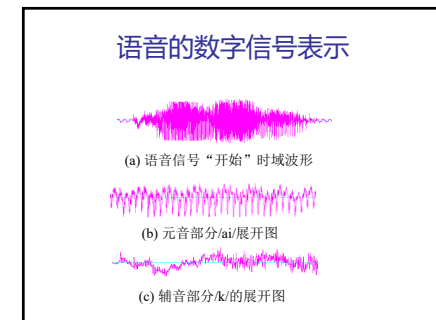
38

语音的声学特性

声波的基本物理量——振幅：

- 用声压或声强来表示声音的强度
- 声压 p 用来度量由于声波的传播而带来的气压的变化，单位为帕斯卡（Pa）。
- 声强 I 为单位时间内通过与声波传播方向垂直的某一单位面积上声能的平均值，单位 W/m^2 。
- 人耳对声音的强度非常敏感，且动态范围很大。能感受的最小声压称为闻阈，约为 $2 \times 10^{-5} Pa$ ；能承受的最大声压称为痛阈，约为 $200 Pa$ 。

41



44

语音的声学表示

39

语音的声学特性

- 习惯上采用相对强度，以闻阈 P_0 为基准，单位为dB
 - 声压级 $L = 20 \log_{10}(P/P_0)$ (dB)
 - 声强级 $L = 10 \log_{10}(I/I_0)$ (dB)
- 闻阈相当于0dB；痛阈相当于140dB；一个讲话时，离他一米远处的声强大约为60~80dB；
- 声波的基本物理量——共振和共振峰
 - 共振：当一个物体受迫震动时，所加驱动频率等于物体固有频率时，便以最大的振幅来震荡。
 - 语音中也有共振现象，元音的音色和区别特征主要取决于声道的共振峰特性。

42

数字信号处理的相关知识

➤ 离散时间信号与系统

- 离散时间信号
 - 在时间上是离散的，只在某些不连续的规定瞬间给出函数值。若幅值连续又称抽样信号。幅值离散又称数字信号
 - 通常函数值的离散时刻之间的间隔是均匀的。一般以序列 $x(n)(n=0, \pm 1, \pm 2, \dots)$ 来表示

45

数字信号处理的相关知识

- 离散信号序列的基本运算
序列中同序号的数值逐项运算而构成一个新序列。

加: $z(n) = x(n) + y(n) \quad (n = 0, \pm 1, \pm 2, \dots)$

乘: $z(n) = x(n) \cdot y(n) \quad (n = 0, \pm 1, \pm 2, \dots)$

时延: $z(n) = x(n-m) \quad (n = 0, \pm 1, \pm 2, \dots)$

求能量: $E = \sum_{n=-\infty}^{\infty} |x(n)|^2$

46

数字信号处理的相关知识

如果 $x(n)$ 在 $[0, N-1]$ 区间取值, 那么上式将改写为

$$y(n) = \sum_{m=0}^{N-1} x(m)h(n-m)$$

上两式被称作卷积运算, 记做 $y(n) = x(n) * h(n)$

卷积运算的性质:

- (1) 交换率 $a(n) * b(n) = b(n) * a(n)$
- (2) 结合率 $[a(n) * b(n)] * c(n) = a(n) * [b(n) * c(n)]$
- (3) 分配率 $[a(n) + b(n)] * c(n) = a(n) * c(n) + b(n) * c(n)$
- (4) 转移特性

在计算机中, 滤波主要是通过卷积运算来实现的。

49

数字信号处理的相关知识

➤ Z变换

它可以将离散系统的数学模型(差分方程)转化为简单的代数方程。

• Z变换的定义: $X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$

• Z变换存在收敛域问题, 在其收敛域 $R_1 < |z| < R_2$ 内满足:

$$\sum_{n=-\infty}^{\infty} |x(n)| \|z\|^{-n} < \infty$$

• DFT是特殊的Z变换, 取 $z = e^{j(2\pi/N)k}$ 即可得到DFT变换。在 z 平面的单位圆上, 取幅角为 $\omega = 2\pi k/N$ 计算其Z变换, 就得到DFT的第 k 个样值点 $X(k)$ 。有限长序列的DFT可以解释为它的Z变换在单位圆上的均匀抽样。

52

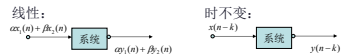
数字信号处理的相关知识

- 离散时间系统: 系统的激励和响应都是离散信号序列。离散时间系统的数学模型是差分方程。



最常见的系统是线性时不变系统。

设 $y_1(n)$ $y_2(n)$ 分别为激励 $x_1(n)$ $x_2(n)$ 的响应, 则:



47

数字信号处理的相关知识

➤ 离散傅立叶变换 (DFT)

连续非周期信号: 傅立叶变

连续周期信号: 傅立叶级

离散非周期信号: 离散时间

离散周期信号: 离散傅立

DFT:

对有限长序列 $x(n)(n=0, \dots, N-1)$ 进行延拓, 扩展成周期信号。

■ 变换:
$$\begin{cases} X(k) = DFT[x(n)] = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi nk}{N}} & (0 \leq k \leq N-1) \\ x(n) = IDFT[X(k)] = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{j\frac{2\pi nk}{N}} & (0 \leq n \leq N-1) \end{cases}$$

50

数字信号处理的相关知识

- Z变换的性质

1 线性: 若 $\begin{cases} Z[x(n)] = X(z) & (R_{x1} < |z| < R_{x2}) \\ Z[y(n)] = Y(z) & (R_{y1} < |z| < R_{y2}) \end{cases}$
则有 $Z[ax(n) + by(n)] = aX(z) + bY(z) \quad (R_1 < |z| < R_2)$
其中 $R_1 = \max(R_{x1}, R_{y1}) \quad R_2 = \min(R_{x2}, R_{y2})$

2 时移性: $Z[x(n-m)] = z^{-m}X(z)$

3 时域卷积定理: $Z[x(n) * y(n)] = X(z)Y(z) \quad R_1 < |z| < R_2$

53

数字信号处理的相关知识

- 卷积和滤波

单位冲激信号 $\delta(n) = \begin{cases} 1 & (n=0) \\ 0 & (n \neq 0) \end{cases}$ 作为激励就得到
单位冲激响应 $h(n)$



由于任意输入信号 $x(n)$ 可以表示为: $x(n) = \sum_{m=-\infty}^{\infty} x(m)\delta(n-m)$

根据线性时不变系统的特性, 其响应信号 $y(n)$ 可以写为:

$$y(n) = \sum_{m=-\infty}^{\infty} x(m)h(n-m)$$

48

数字信号处理的相关知识



“开始”中/ai/的频谱特性

51

数字信号处理的相关知识

➤ 离散余弦变换 (Discrete Cosine Transform)

$$C(k) = \sum_{n=0}^{N-1} x(n) \cos(\pi k(n+1/2)/N) \quad (0 \leq k \leq N-1)$$

$$x(n) = [C(0) + 2 \sum_{k=1}^{N-1} C(k) \cos(\pi k(n+1/2)/N)]/N \quad (0 \leq n \leq N-1)$$

➤ DCT变换可以从DFT变换推导得到

➤ DCT变换的优点在于能量的集中, 相比于DFT, 其系数主要集中在维数较低的部分, 这样就能用更少的系数来逼近原来的信号。

54

语音的语言表示

55

语音的语言表示

- 声母表
b p m f d t n l
g k h j q x
zh ch sh r z c s
- 韵母可以包括一个元音，也可以包括多个元音，也可以包括辅音。韵母共38个。
- 韵母表
l u ü A ia ua o uo e ie üe ai uai ei uei ao iao
ou iou an ian uan üan en in uen ün ang iang
uang eng ing ueng ong iong

58

语音的语言表示

- 句子 => 短语 => 词语 => 音节 => 音素
- 音素是语音的基本单位。可以分为元音（浊音）和辅音（清音）。
- 元音是指发音的过程中，对声腔气流无明显阻塞而发出的音段，如[a]、[i]等。
- 辅音是声腔气流明显受阻时所发出的音段，如[m][n]等。
- 此外还用半元音、双元音、半辅音等等。

56

语音的语言表示

- 汉语音素为64个，分为辅音、单元音、复元音和复鼻尾音。
- 汉语的每个字就是一个音节。音节由声母和韵母拼接而成，音节中也可以不包含声母。
- 无调音节415个。 [无调音节列表](#)
- 每个音节可以有四种声调，因此有调音节一千二百多个。

59

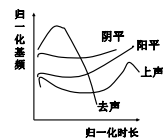
语音的语言表示

- 对一组语言来讲，可以用一组音素来描述。
- 美国英语包括42个音素，分为：元音12个；双元音6个；半元音4个；辅音20个。
- 汉语普通话是以北京语音为标准音，以北方话为基准，国际上常用的词为（mandarin）。
- 汉语采用声韵结构，每个字音分成两部分，前面的部分称为声母（initial），后一部分称为韵母（final）。
- 声母为辅音，但不是所有的辅音都可以做声母。声母共22个。

57

语音的语言表示

- 汉语音节的声调主要体现在信号的基音频率随时间而变的规律上。



60

视听觉信号处理

Visual and Auditory Signal Processing

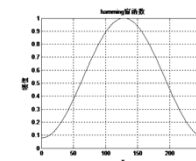


1

特征的分析时长

4

特征的时域分析方法



在进行频率分析（FFT）时，让信号具有周期性，消除吉布斯效应的影响

7

语音的时频域分析



2

特征的分析时长

短时分析

- 语音信号是非平稳信号，但是可以认为10~30ms的时间范围内，语音信号是平稳信号。
- 短时分析的最基本手段是对语音信号加窗。

$$x_w(n) = x(n)w(m)$$

常见窗函数：N为窗长

方窗： $w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$

5

特征的时域分析方法

8

如何理解语音信号？

如何用更少的属性数据来刻画声音的内容？

对识别类应用——特征提取

对表示类应用——表示机理

3

特征的时域分析方法

哈明（Hamming）窗

$$w(n) = \begin{cases} 0.54 - 0.46 \cos[2\pi n / (N-1)] & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$$

哈宁（Hanning）窗

$$w(n) = \begin{cases} 0.5[1 - \cos(2\pi n / (N-1))] & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$$

6

特征的时域分析方法

短时能量、短时平均幅度和短时过零率

- 短时能量

$$E_n = \sum_{m=-\infty}^{\infty} [x(n)w(m)]^2$$

短时能量可用于清浊判决、有声段和无声段进行判定、对声母和韵母分界，以及连字的分界等。经常是识别系统中特征的一维。

9

特征的时域分析方法

- 短时平均幅度

$$M_n = \sum_{m=-\infty}^{\infty} |x_n(m)|$$

- 短时过零率：单位时间内过零发生的次数。

$$Z_n = \sum_{m=-\infty}^{\infty} |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]| w(m)$$

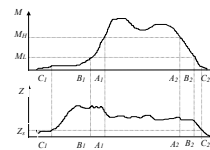
式中

$$\text{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \quad w(n) = \begin{cases} 1/2N & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$$

10

特征的时域分析方法

- 双门限法



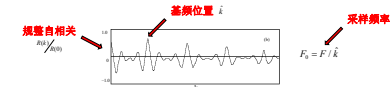
(1) 用较高的短时能量门限 M_0 确保A1-A2肯定是语音。

(2) 从A1 A2开始向两端搜索，短时能量>较低门限 M_0 的B1-B2还是语音段。

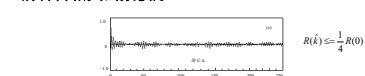
(3) 从B1开始向前搜索，短时过零率<门限 Z_0 的为清音部分。

13

浊音自相关函数波形



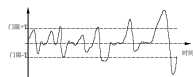
清音自相关函数波形



16

特征的时域分析方法

- 短时平均过零率容易受到噪声的干扰,因而提出了门限过零率的思想。



$$Z_n = \sum_{m=-\infty}^{\infty} \{ |\text{sgn}[x(m)-T] - \text{sgn}[x(m-1)-T]| + |\text{sgn}[x(m)+T] - \text{sgn}[x(m-1)+T]| \} w(m)$$

11

特征的时域分析方法

- 短时自相关函数

- 自相关函数

对于确定性离散信号 $x(n)$

$$R(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$$

$R(k)$ 表示一个信号和延迟 k 点后的该信号本身的相似性。

14

基音周期检测——短时自相关函数特征的应用示例

✓基音是指发浊音时声带振动所引起的周期性，它只是准周期性的。

✓在语音编解码器、语音识别、说话人确认和辨认，以及生理缺陷人的辅助系统等许多领域都是重要的一环。

✓浊音信号的自相关函数在基音周期的整数倍位置上出现峰值，而清音的自相关函数没有明显的峰值出现。

✓峰—峰值之间对应的就是基音周期。

✓在限定 K 值内的最大峰值出现的位置

17

特征的时域分析方法

- 端点检测——能量过零率特征的应用示例

- 对于语音进行“浊音/清音/无声”的判定。

- 在汉语中，若浊音处于音节的末尾，容易通过短时能量来区别，但在音节的前端，清音与环境噪声则很难区分。

- 浊音的能量高于清音，清音的过零率高于无声段。

12

特征的时域分析方法

自相关函数的性质：

1偶函数： $R(k) = R(-k)$

2 $k=0$ 时函数取最大值，对于确定性信号其值为能量。对于随机信号，其值为该信号的平均功率。

3 如果原序列是周期为 T 的周期信号，那么自相关函数也是周期为 T 的周期函数。

15

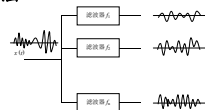
特征的频域分析方法

18

特征的频域分析方法

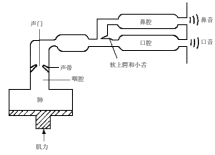
语音的感知过程与人类听觉系统具有频谱分析功能，是紧密相关的。因此，对语音信号进行频谱分析，是认识语音信号和处理语音信号的重要方法

• 滤波器组方法



19

特征的频域分析方法



语音产生的机理图

22

特征的频域分析方法

• 特征系统 D^* 

第一步是对信号进行Z变换，将卷积信号转变为乘积信号

$$Z[x(n)] = X(z) = E(z) \times H(z)$$

第二步是进行对数运算，将乘积信号变为加性信号

$$\log X(z) = \log E(z) + \log H(z) = \tilde{E}(z) + \tilde{H}(z) = \tilde{X}(z)$$

第三步进行反Z变换运算，变回时域信号

$$Z^{-1}[\tilde{X}(z)] = Z^{-1}[\tilde{E}(z) + \tilde{H}(z)] = \tilde{e}(n) + \tilde{h}(n) = \tilde{x}(n)$$

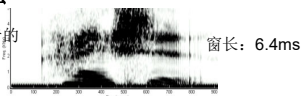
$\tilde{x}(n)$ 已知的条件下很容易通过线性运算得到 $\tilde{h}(n)$

25

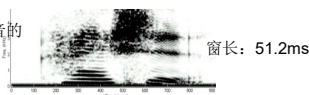
特征的频域分析方法

• 语谱图方法

“开始”语音的
宽带语谱图

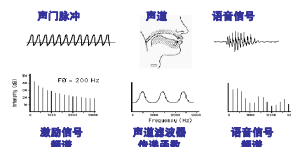


“开始”语音的
窄带语谱图



20

特征的频域分析方法



23

特征的频域分析方法

• 反特征系统 D^{*-1} : 它是特征系统的反运算。

复倒谱 (Complex Cepstrum): 将特征系统的输出称为复倒谱或对数复倒谱。

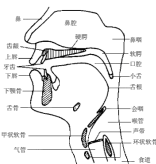
$$\tilde{x}(n) = Z^{-1}[\log[Z[x(n)]]]$$

其所在域称之为倒谱域。

26

特征的频域分析方法

• 语音的产生

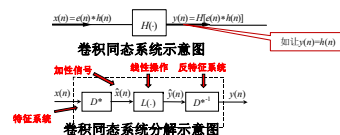


21

特征的频域分析方法

• 卷积同态信号处理方法

目的：乘积性组合信号或卷积性组合信号转化为加性信号。将非线性问题转化为线性问题来处理。



24

特征的频域分析方法

• 倒谱：仅对 $\tilde{x}(z)$ 的实部作逆Z变换

$$c(n) = Z^{-1}[\log |Z[x(n)]|]$$

倒谱不能通过逆特征系统还原成自身。

在绝大多数应用场合，特征系统和逆特征系统中的正反Z变换都可以用正反傅立叶变换 (DFT和IDFT) 来代替。

27

特征的频域分析方法

• 修正(2020.12.7):

短时自相关函数在假定窗外为0时是偶函数

$$R(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) = \sum_{m=0}^{N-1-k} x(m)x(m+k)$$

令 $m = n + k$, 则

$$R(k) = \sum_{m=0}^{N-1-k} x(m-k)x(m) = \sum_{m=0}^{N-1-k} x(m)x(m-k) = R(-k)$$

若不假定窗外为0, 做同样的变换

$$R(k) = \sum_{n=0}^{N-1-k} x(n)x(n+k) = \sum_{m=0}^{N-1-k} x(m-k)x(m) = \sum_{m=0}^{N-1-k} x(m-k)x(m) = R(-k)$$

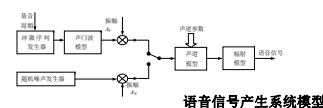
28

特征的频域分析方法

经过推导, 该模型系统的传递函数为如下形式

$$H(z) = \frac{G}{1 - \sum_{k=1}^N a_k z^{-k}}$$

N为级联声管的节数, 上式为全极点形式。



31

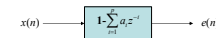
特征的频域分析方法

预测误差

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^N a_i x(n-i)$$

可以如下得到该预测器的预测误差

$$E(z) = [1 - \sum_{i=1}^N a_i z^{-i}] X(z)$$



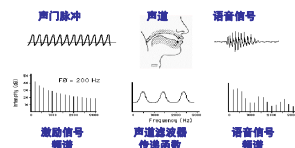
它的逆滤波器形式:

$$e(n) \rightarrow \frac{1}{1 - \sum_{i=1}^N a_i z^{-i}} \rightarrow x(n)$$

34

特征的频域分析方法

• 为人类的发声过程建立数学模型



29

特征的频域分析方法

▶ 线性预测 (Linear Prediction) 分析

• 根据语音信号的产生模型, 语音信号 $x(n)$ 可以看作以 $u(n)$ 为激励的一个全极点滤波器的响应。



问题: 如何在已知 $x(n)$ 的条件下, 求出系数 $\{a_i\} i=1, \dots, p$?

解答: 线性预测分析的方法。

32

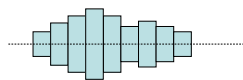
特征的频域分析方法

- ▶ 线性预测器与全极点模型一一对应
- ▶ 预测器也不容易确定, 系数不同就是不同的预测器, 有无数预测器
- ▶ 有一个特殊的预测器: 最佳线性预测器
- ▶ 最佳线性预测器的预测误差能量最小
- ▶ 求最佳线性预测器的过程可以被成为线性预测分析, 或者自回归 (Autoregressive, AR) 分析

35

特征的频域分析方法

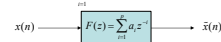
- 在声学上对均匀的无损耗的管道的声学特性有非常简单的数学描述。均匀: 截面积不变; 无损耗: 不考虑声波在管内的热损耗、粘滞摩擦损耗和管壁的热传导作用。
- 在此基础上, 可以将声道简化成一些截面积不等的均匀无损声管的级联。用该模型来逼近真实的声道, 称之为声道的时间离散模型。



30

特征的频域分析方法

线性预测器: $F(z) = \sum_{i=1}^N a_i z^{-i}$



在Z域, 是如下乘积关系

$$\hat{X}(z) = \sum_{i=1}^N a_i z^{-i} X(z)$$

反Z变换, 可得到如下时域差分方程:

$$\hat{x}(n) = \sum_{i=1}^N a_i x(n-i)$$

从时域角度可以理解, 用信号的前p个样本来预测当前的样本得到预测值

33

特征的频域分析方法

- ▶ 思路: 在数字信号处理中, 一个AR模型与一个最佳的线性预测器是等价的, 也就是说, 用AR模型的系数 a_i 构造的预测器必然是最佳预测器, 即在最小均方意义上, 预测误差能量最小。因此从 $x(n)$ 出发, 寻找其最佳预测器, 从而得到系数 a_i 。
- ▶ 系数被称为线性预测系数或LPC系数。

36

特征的频域分析方法

预测误差：

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^p a_i x(n-i)$$

短时预测均方差：

$$E_e = \sum_n e^2(n) = \sum_n [x(n) - \hat{x}(n)]^2 = \sum_n [x(n) - \sum_{i=1}^p a_i x(n-i)]^2$$

• 求解过程

使 $\partial E_e / \partial a_k = 0$, ($k=1, 2, \dots, p$) 则有：

$$\frac{\partial E_e}{\partial a_k} = -(2 \sum_n x(n)x(n-k) - 2 \sum_{i=1}^p a_i \sum_n x(n-k)x(n-i))$$

37

特征的频域分析方法

• 求LPC系数需考虑两个因素

(1) 模型阶数的选择 $p=2D+1$, D 是共振峰的个数(2) 考虑口唇的高频衰减特性, 在线性预测分析之前, 需要通过预加重进行高频提升 $1-\alpha z^{-1}$

40

特征的频域分析方法

将其转换成矩阵形式

$$\begin{bmatrix} R_x(0) & R_x(1) & R_x(2) & \dots & R_x(p-1) \\ R_x(1) & R_x(0) & R_x(1) & \dots & R_x(p-2) \\ R_x(2) & R_x(1) & R_x(0) & \dots & R_x(p-3) \\ \dots & \dots & \dots & \dots & \dots \\ R_x(p-1) & R_x(p-2) & R_x(p-3) & \dots & R_x(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \dots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} R_x(1) \\ R_x(2) \\ R_x(3) \\ \dots \\ R_x(p) \end{bmatrix}$$

这种方程为Yule-Walker方程, 其系数矩阵被称为托布利兹 (Toeplitz) 矩阵。具有如下性质:

- (1) $p \times p$ 阶的对称阵。
- (2) 沿着主对角线及任何一条与主对角线平行的斜线上的所有元素都相等。

43

特征的频域分析方法

得到线性方程组

$$\sum_n x(n)x(n-k) = \sum_{i=1}^p a_i \sum_n x(n-k)x(n-i) \quad k=1, 2, \dots, p$$

若定义 $\Phi(k, i) = \sum_n x(n-k)x(n-i) \quad k=1, 2, \dots, p \quad i=0, 1, 2, \dots, p$ 则方程组可简写为 $\sum_{i=1}^p a_i \Phi(k, i) = \Phi(k, 0)$

求解方程, 可得到LPC系数

一个由p个方程组成的有p个未知数的线性方程组

38

特征的频域分析方法

• 自相关法

我们定义 $\Phi(k, i) = \sum_n x(n-k)x(n-i)$ 时, 未将求和范围具体化。一种较直接的方法是, 认为语音段外的数据全为零, 只计算范围 n 以内 ($0 \leq n \leq N$) 的语音数据。

或

$$\Phi(k, i) = \sum_{n=0}^{N-k-i} x_n(n)x_n(n+k-i) \quad k=1, 2, \dots, p \quad i=0, 1, 2, \dots, p$$

 $x_n(n)$ 为加窗后的语音数据。

41

特征的频域分析方法

Yule-Walker方程可以用递推的方式来求解。典型的方法有:

- 莱文逊—杜宾 (Levinson—Durbin) 递推算法
- 舒尔 (Schur) 递推算法

44

特征的频域分析方法

要构造信号的AR模型, 还应估算增益因子

AR模型的差分方程形式 $x(n) = \sum_{i=1}^p a_i x(n-i) + Gu(n)$ 因此可计算预测误差 $e(n) = \sum_{i=1}^p a_i x(n-i) - \hat{x}(n) = Gu(n)$

且

$$E_e = G^2 \sum_n u^2(n)$$

激励信号 $u(n)$ 总能量可以认为近似为1, 因此有 $G = E_e^{1/2}$

39

特征的频域分析方法

由于短时自相关函数可以表示为:

$$R_x(k) = \sum_{n=0}^{N-k} x_n(n)x_n(n+k)$$

且有

则 $\Phi(k, i)$ 可以表示为

$$\Phi(k, i) = R_x(k-i) = R_x(i) \quad k=1, 2, \dots, p \quad i=0, 1, 2, \dots, p$$

求解LPC系数的方程组就可以写为:

$$\sum_{i=1}^p R_x(i) \hat{a}_i = R_x(k) \quad k=1, 2, \dots, p$$

42

特征的频域分析方法

莱文逊—杜宾 (Levinson-Durbin) 递推算法

- 不直接计算 p 阶预测器
- 从一阶预测器开始, 逐一递推各阶预测器
- 第 i 阶预测器的系数可以用第 $i-1$ 阶预测器的系数递推得到
- 直到递推出 p 阶预测器的系数
- 用到了 i 阶预测器的预测误差能量 $E^{(i)}$ 和一个中间系数 k_i

张贤达等,《现代信号处理》,清华大学出版社

45

特征的频域分析方法

- (1) 计算自相关系数 $R_k(j)$, $j = 0, 1, \dots, p$
 (2) 初值 $E^{(0)} = R_k(0)$ $i = 1$
 (3) 开始按如下公式进行递推运算

$$k_i = \frac{R_k(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R_k(i-j)}{E^{(i-1)}}$$

$$a_i^{(i)} = k_i$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_j^{(i-1)} \quad j = 1, \dots, i-1$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

- (4) $i=i+1$ 。若 $i > p$ 则算法结束退出, 否则返回第 (3) 步,

46

特征的频域分析方法

此时的系数矩阵不再是一个托布利兹矩阵, 它一般用乔里斯基 (Choleskey) 分解法来求解。

- 自相关法和协方差法的比较

➢ 自相关法必须对语音信号进行加窗处理, 规定了信号的长度范围, 假定窗外的语音样本值为零, 所以自相关法误差较大, 计算结果精度差, 但自相关法能够保证系统的稳定性。

➢ 协方差法因不需要加窗, 所给出的参数估值要比自相关法精确的多, 但不如自相关法稳定, 另外乔里斯基分解法因没有快速算法, 也需要较大的计算量。

49

特征的频域分析方法

得到 $\hat{h}(n)$ 和 a_i 间的递推关系为

$$\begin{cases} \hat{h}(1) = a_1 \\ \hat{h}(n) = a_n + \sum_{i=1}^{n-1} (1 - \frac{i}{n}) a_i \hat{h}(n-i), & 1 \leq n \leq p \\ \hat{h}(n) = \sum_{i=n}^p (1 - \frac{i-n}{p-n+1}) a_i \hat{h}(n-i), & n > p \end{cases}$$

52

特征的频域分析方法

经过递推计算后, 最终解为:

$$\hat{a}_j = a_j^{(p)}, \quad j = 1, 2, \dots, p \quad E^{(p)} = R_k(0) \prod_{i=1}^p (1 - k_i^2)$$

可以推知

$$|k_i| \leq 1, \quad i = 1, 2, \dots, p$$

k_i 称为反射系数, 也称PARCOR系数。

47

特征的频域分析方法

LPC倒谱系数 (LPPC)

➢ 倒谱是通过对信号进行Z变换, 取对数, 再反Z变换来得到的。

➢ 求单位冲激响应 $h(n)$ 的倒谱 $\hat{h}(n) = Z^{-1}[\log H(z)]$

➢ 它也反映了信号的谱包络信息。

有:

$$H(z) = \frac{1}{1 - \sum_{i=1}^p a_i z^{-i}}$$

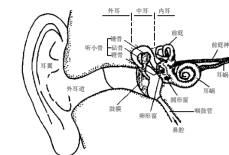
$$\hat{H}(z) = \log H(z) \text{ 可以展开成级数形式 } \hat{H}(z) = \sum_{n=1}^{\infty} \hat{h}(n) z^{-n}$$

$$\log \left[\frac{1}{1 - \sum_{i=1}^p a_i z^{-i}} \right] = \sum_{n=1}^{\infty} \hat{h}(n) z^{-n}$$

50

特征的频域分析方法

感知机理的仿真



53

特征的频域分析方法

- 协方差法:

重新定义求和范围

$$\Phi(k, i) = \sum_{m=i}^{N-1} x(m-k)x(m-i) \quad k = 1, 2, \dots, p \quad i = 0, 1, 2, \dots, p$$

设 $(n-i) = m$

$$\Phi(k, i) = \sum_{m=i}^{N-1} x(m+(i-k))x(m) \quad k = 1, 2, \dots, p \quad i = 0, 1, 2, \dots, p$$

此时不再满足 $\Phi(i+1, k+1) = \Phi(i, k)$, 因而系数矩阵变成如下形式

$$\begin{bmatrix} \Phi(1,1) & \Phi(1,2) & \Phi(1,3) & \dots & \Phi(1,p) \\ \Phi(2,1) & \Phi(2,2) & \Phi(2,3) & \dots & \Phi(2,p) \\ \Phi(3,1) & \Phi(3,2) & \Phi(3,3) & \dots & \Phi(3,p) \\ \dots & \dots & \dots & \dots & \dots \\ \Phi(p,1) & \Phi(p,2) & \Phi(p,3) & \dots & \Phi(p,p) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \dots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \Phi(1,0) \\ \Phi(2,0) \\ \Phi(3,0) \\ \dots \\ \Phi(p,0) \end{bmatrix}$$

48

特征的频域分析方法

将上式两边同时对 z^{-1} 求导

$$\frac{\partial}{\partial z^{-1}} \log \left[\frac{1}{1 - \sum_{i=1}^p a_i z^{-i}} \right] = \frac{\partial}{\partial z^{-1}} \sum_{n=1}^{\infty} \hat{h}(n) z^{-n}$$

有

$$\sum_{n=1}^{\infty} n \hat{h}(n) z^{-n+1} = \frac{\sum_{n=1}^{\infty} n a_n z^{-n+1}}{1 - \sum_{i=1}^p a_i z^{-i}}$$

$$(1 - \sum_{i=1}^p a_i z^{-i}) \sum_{n=1}^{\infty} n \hat{h}(n) z^{-n+1} = \sum_{n=1}^{\infty} n a_n z^{-n+1}$$

51

特征的频域分析方法

- 正常人耳能感知的频率范围为16.4Hz~16KHz; 强度范围为0dB~120dB。

- 音调是人耳对不同频率声音的一种主观感觉。单位为Mel, 与频率近似的满足方程:

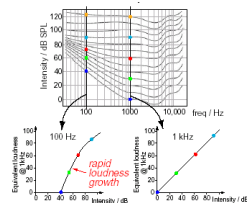
$$p_{Mel} \approx (1000/\lg 2) \times \lg(1 + 0.001 f_{Hz})$$

- 响度用来描述人耳对不同频率的纯音的辨别灵敏度。单位为Phon。1Phon等于1kHz纯音的1dB声强级。为了确定一个音的响度, 需要调节1kHz纯音的声强, 使其与目标音一样响, 此时的声强就是待求响度。

54

特征的频域分析方法

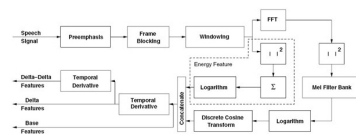
• 等响度曲线



• 掩蔽效应

55

特征的频域分析方法



58

特征的频域分析方法

• Mel频率倒谱系数 (MFCC)

人的耳蜗实质上的作用相当于一个滤波器组，耳蜗的滤波作用是在对数频率尺度上进行的，在1000Hz以下为线性尺度，而1000Hz以上为对数尺度，这就使得人耳对低频信号比对高频信号更敏感。

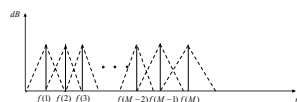
根据这一原则，研究者根据心理学实验得到了类似于耳蜗作用的一组滤波器组，这就是Mel频率滤波器组。

(1) 将时域信号 $x(n)$ 后补若干以形成长为 N （一般取 $N=512$ ）的序列，然后经过FFT变换的线性频谱 $X(k)$ 。

56

特征的频域分析方法

(2) 将线性频谱 $X(k)$ 通过Mel频率滤波器组得到Mel频谱。



(3) 对每个滤波器的输出信号取对数能量。

(4) 对这组对数能量值做DCT变换。

57

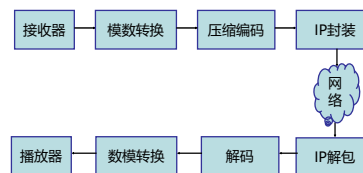
视听觉信号处理

Visual and Auditory Signal Processing



1

语音编码应用实例 (VoIP)



4

语音编码的类型:

- 波形编码
- 参数编码
- 混合编码

波形编码: 将时域模拟语音的波形信号经过采样、量化和编码形成数字语音信号, **解码后的语音信号基本上与输入语音信号波形相同。**

- 编码速率较高: 9.6k~64kbit/s
- 包括: PCM、压扩PCM、ADPCM、DM、ADM、SBC等
- 适应能力强、语音质量好; 编码速率高

7

语音编码



2

为什么语音是可以压缩的?

1. 存在冗余度:

- (1) 幅度非均匀分布
- (2) 语音信号样本间的相关性很强
- (3) 浊音具有准周期
- (4) 声道的形状及其变化缓慢
- (5) 语音间隙

5

参数编码: 基于人类语音的产生机理建立数学模型, 根据输入语音得出模型参数并传输, 在收端恢复, **重建的语言信号与原始信号样本之间没有一一对应关系, 但内容相同。**

- 编码速率较低: 2.4k~4.8kbit/s
- 包括各线性预测编码 (LPC) 方法和余弦声码器等
- 编码速率低; 语音质量差、自然度低、对环境噪声敏感

混合编码: 波形编码+参数编码

- 编码速率较低: 16k~2.4kbit/s
- 包括多脉冲激励线性预测编码(MPLPC)、规则脉冲激励线性编码(RPE-LPC)、码本激励线性预测编码(CELP)

8

为什么要进行编码?

语音数据有多大?

--fs: 每秒钟样本数 (8k~44.1k)

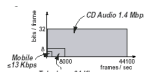
"c": 每个样本的通道数 (1 or 2)

"b": 每个样本点的位数 (8 or 16)

比特率: $f_s * c * b$, bit/s 或 bps

压缩语音信号的传输带宽, 降低信道的传输速率

语音编码就是使比特率尽可能小



3

为什么语音是可以压缩的?

2. 人的听觉感知机理

- (1) 人类的听觉特性具有掩蔽效应
- (2) 人耳对不同频段声音的敏感程度不同
- (3) 人耳对语音相位不敏感

6

在语音通信中, 语音质量分为以下四等:

① **广播质量:** 宽带, 语音质量高, 感觉不出噪声存在。

② **长途电话质量:** 指通过电话网传输后得到的语音质量, **信噪比大于30dB**, 谐波失真小于**2%-3%**。

③ **通信质量:** 可以听懂, 但和长途电话质量相比, 略有较大失真。

④ **合成质量:** **80%-90%**可懂度, 听起来像机器说话, 失去了讲话者的个人特征。

$$\text{波形失真度} \quad \text{SNR} = 10 \cdot \log \left(\frac{\sum_{n=1}^N (x(n))^2}{\sum_{n=1}^N (\hat{x}(n) - \bar{x})^2} \right)$$

9

已经标准化的语音编码

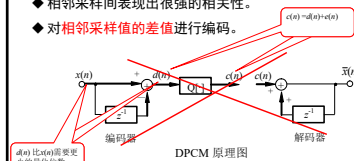
指定组织：国际电信联盟ITU-T, (<http://www.itu.int>)

| 标准 | 编码类型 | 比特率 (kbps) | MOS | 复杂度 | 时延 (ms) |
|-----------|----------|---------------|-----|-----|------------|
| G.711 | PCM | 64 | 4.3 | 1 | 0.125 |
| G.726 | ADPCM | 32 | 4.0 | 10 | 0.125 |
| G.723 | LD-CELP | 16 | 4.0 | 50 | 0.625 |
| GSM | RPE-LTP | 13 | 3.7 | 5 | 20 |
| G.729 | CSA-CELP | 8 | 4.0 | 30 | 15 |
| G.729A | | | | | 15 |
| G.723.1 | ACELP | 6.3 | 3.8 | 25 | 37.5 |
| | MP-MELQ | 5.3 | | | |
| ITU G.728 | LPC-10 | 2.4 | | 10 | 22.5 |
| FS1015 | | | | | |

10

DPCM

- ◆ 在PCM中, 各采样值都独立编码, 需要较多位数, 比特率较高。
- ◆ 相邻采样间表现出很强的相关性。
- ◆ 对相邻采样值的差值进行编码。



13

DPCM

$$\bar{X}(z) = \frac{C(z)z^{-1}}{1-z^{-1}}$$

$$C(z) = X(z) - \bar{X}(z) + E(z)$$

$$C(z) = (X(z) + E(z))(1-z^{-1})$$

$$\bar{X}(z) = \frac{C(z)}{1-z^{-1}} = X(z) + E(z)$$

$$\text{有: } \bar{x}(n) = x(n) + e(n)$$

可以看出, 已经消除了量化噪声的累积。

16

DPCM

分析存在的问题。用z变换考察各点信号的时域关系, 有:

$$C(z) = X(z)(1-z^{-1}) + E(z)$$

$$\bar{X}(z) = \frac{C(z)}{1-z^{-1}} = X(z) + \frac{E(z)}{1-z^{-1}}$$

其中E(z)为量化器量化噪声e(n)的z变换。有:

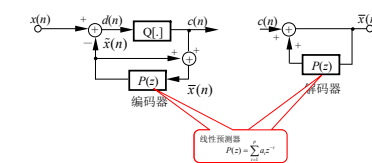
$$\bar{x}(n) = x(n) + \sum_{m=1}^{\infty} e(m)$$

可以看出, 量化器所产生的量化噪声被累积叠加到了输出信号中。

14

ADPCM

- ◆ 仅仅利用到了相邻的两个采样点之间的相关性。
- ◆ 用线性预测来刻画更多采样间的相关性。
- ◆ 对采样值与预测值间的差值进行编码。



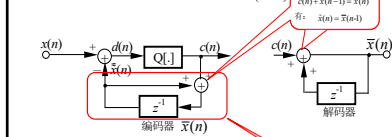
17

波形编码之ADPCM

12

DPCM

- ◆ 原因: 在编码端与x(n-1)做差, 而在解码端则与x̂(n-1)计算差值。合理的方式是都采用x̂(n-1)

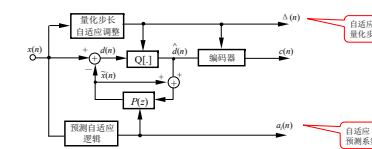


- ◆ 在编码器中包含解码器

15

ADPCM

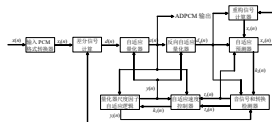
- ◆ 最佳线性预测器系数每帧(段)不同, 需要每帧(段)传送一次。被称之为边信息。
- ◆ 每帧(段)可以采用不同的量化步长(位数)。根据预测残差自适应确定。



18

ADPCM

- ◆ ADPCM已形成国际标准，ITU-T（原CCITT）在1988年制定了G.726标准，将1984年和1986年分别制定的ADPCM标准G.721和G.723进行了合并，同时也删除了上述两个标准。
- ◆ G.726能提供四种数码率：40kbit/s、32kbit/s、24kbit/s、16kbit/s。其语音质量相当于64kbit/s的PCM编码，并具有很好的抗误码性能。



G.726 编码器方框图

19

LPC编码

- ◆ 美国确定LPC-10作为2.4kb/s速率上的推荐编码形式，用于第三代保密电话中。
- ◆ 在其发送端，原始语音信号采用8kHz采样，然后每180个采样值分为一帧（22.5ms），提取语音特征参数并加以编码传送。每帧总共编码为54bits，每秒传输44.4帧，因此总传输速率为2.4kb/s。
- ◆ 其增强版本为LPC-10c

22

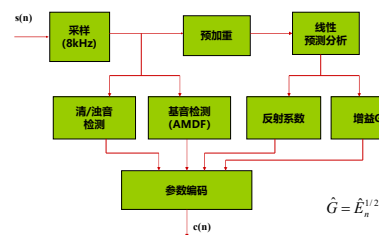
混合编码之CELP

25

参数编码之LPC-10

20

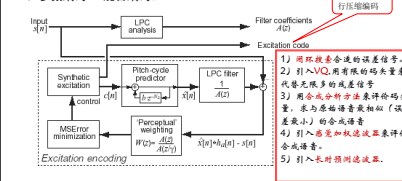
LPC-10编码器发送端



23

混合编码

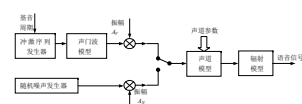
- ◆ LPC-10语音质量是“合成语音”级别。
- ◆ 对声门波的假定并不符合实际。（声门波=预测误差？）
- ◆ 参数编码 → 混合编码。



26

LPC编码

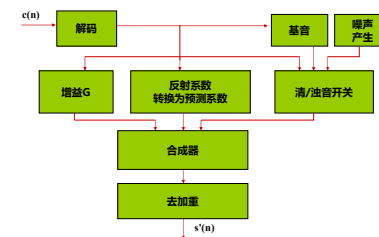
- ◆ 完全基于语音信号的产生模型。
- ◆ 在编码端计算模型参数，作为编码传输，在解码端基于该模型参数合成语音。
- ◆ 解码后语音波形一般都会发生改变。



语音信号产生系统模型

21

LPC-10编码器接收端



24

混合编码

- 预测误差信号仍有较强的相关性，可以用长时预测去相关，使其更平坦。
- 对应声门波要经过一个长时预测综合滤波器，表示语音信号长时相关性的模型。它的一般形式为：

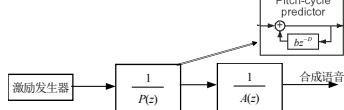
$$1/P(z) = 1/[1 - \sum_{i=0}^{r-1} b_i z^{-(D+i)}]$$

- 其中延时参数D等于基音周期， $\{b_i\}$ 是语音信号的长时预测系数
- 预测系数的个数取1（ $q=r=0$ ）或3（ $q=r=1$ ）

27

混合编码

◆声道系统由两个级联的滤波器组成

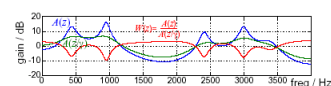


- 两种相关性：样本点之间的短时相关性和相邻基音周期之间的长时相关性。
- 对这两种相关性进行去相关后，可以得到更加平坦的预测残差信号，因而更加有利于进行量化编码。

28

混合编码

感觉加权滤波器的频谱



- 由于掩蔽效应，在语音频谱中，能量较高的频段（共振峰处）的噪声相对于能量较低的频段的噪声不易被感觉。在度量原始语音和合成语音之间的误差时，在高能量段应允许误差大。

31

语音编码的评测方法

编码速率

降低编码速率往往是语音编码的首要目标。

分成两类：**固定速率编码器**和**可变速率编码器**。

□ 固定编码速率

现有大部分编码标准都是固定速率编码，其范围为0.8 kbit/s ~ 64 kbit/s。

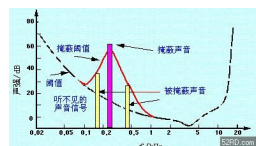
34

混合编码

感觉加权滤波器

- 语音质量与信噪比等价么？

- 掩蔽效应（频域）



'Perceptual' weighting
 $W(z) = \frac{A(z)}{A(z/\gamma)}$

29

CELP

- CELP是近10年来最成功的语音编码算法。

- CELP语音编码算法，用一个包含许多典型的激励矢量的码本作为激励参数，每次编码时都在这个码本中搜索一个最佳的激励矢量。

- 这个激励矢量的编码值就是这个序列的码本中的序号。

- 码本的获得：LBG算法，双重矢量量化

32

语音编码的评测方法

□ 可变编码速率

可变速率编码是近年来出现的新技术。两方通话大约只有40%的时间是真正有声音的，因此可采用通/断二状态编码。可变速率编码主要包括两个算法。一是**有声检测**，主要用于确定输入信号是语音还是背景噪声。二是**舒适噪声生成**，主要用于接收端重建背景噪声，其设计必需保证发送端和接收端的同步。

35

混合编码

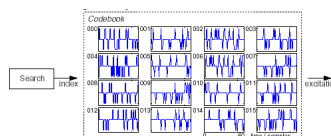
感觉加权滤波器的传递函数

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^p a_i z^{-i}}{1 - \sum_{i=1}^p a_i \gamma^i z^{-i}}$$

加权因子 γ 取值在0~1之间

30

CELP



- 以码本激励线性预测（CELP）原理为基础的G.729、G.729可将经过采样的64kb/s语音以几乎不失真的质量压缩至8kb/s。

- G.723.1有两种编码速率：6.3kb/s和5.4kb/s

33

语音编码的评测方法

□ 强健性

通过取多种不同来源的语音信号进行编解码，并对输出语音质量进行比较测试得到的一种指标。

对存在**部分数据丢失**的情况下，语音编码器强健性的研究也有重要的意义。

36

语音编码的评测方法

□ 时延

- (1) **算法时延**。等于帧长和前视长度之和，其值完全取决于算法，与具体的实现无关。
- (2) **计算时延**。即编码器分析时间和解码器的重建时间，其值取决于硬件速度。
- (3) **复用时延**。编码器发送之前和解码器解码之前，必需将整个数据块的所有比特装配好。
- (4) **传输时延**。取决于是采用专用线还是共享信道。

37

语音编码的评测方法

□ 语音质量及其评价方法

用于评价输出语音质量的方法可分为主观和客观两种。语音主观评价方法种类很多，其中又可分为音质评价和可懂度评价两类。

可懂度评价方法有：

- (1) 判断韵字测试
- (2) 改进的韵字测试

音质的评价方法有：

- (1) 平均意见得分
- (2) 判断满意度测量

40

语音编码的评测方法

□ 时延

单向时延大于150ms就可感受到通话连续性受到影响，最大可容忍时延为400ms~500ms，超过此值只能进行半双工通信。

对于具有回声的情况，单向时延不能超过25ms，否则就需要装备回声抑制功能。

38

语音编码的评测方法

□ 语音质量及其评价方法

目前所用的客观测度方法可以分为时域测度、频域测度和其它测度三类方法。

- (1) 时域测度：信噪比和分段信噪比等；
- (2) 频域测度：对数谱距离测度、LPC例谱距离测度

还有在此二者的基础上发展起来的其它测度方法。

41

语音编码的评测方法

□ 计算复杂度和算法的可扩展性

计算复杂度主要影响硬件实现的成本。算法的可扩展性是指一种编码算法不仅能解决当前的实际应用，而且可以兼顾将来的发展。

39

视听觉信号处理

Visual and Auditory Signal Processing



1

语音识别算法

► 语音识别任务的分类

- 按词汇表 (Vocabulary) 的大小分
 - 小词汇表系统: 包括10~100个词条
 - 中词汇表系统: 包括100~1000个词条
 - 大词汇表系统: 至少包含1000个以上的词条
- 按照发音方式分
 - 孤立词 (Isolated Word) 识别
 - 连接词 (Connected Word) 识别
 - 连续语音 (Continuous Speech) 识别

4

语音识别算法

► 如何计算两个矢量序列 X_1 和 X_2 之间的相似度???

一个直接的想法 $D(X_1, X_2) = \sum_{i=1}^M d(\vec{x}_{1i}, \vec{x}_{2i})$

存在问题:

- 长度不同, $M_1 \neq M_2$
- 对不准

DTW: 将表示两个语音段的矢量序列对准后再计算相似度。
或者说在时间上归正后再计算相似度。

7

语音识别技术概述



2

语音识别算法

- 按说话人的限定范围分
 - 特定人 (Speaker Dependent, SD) 识别
 - 非特定人 (Speaker-Independent, SI) 识别

特定人小词表孤立词系统

动态时间归正方法 (DTW)

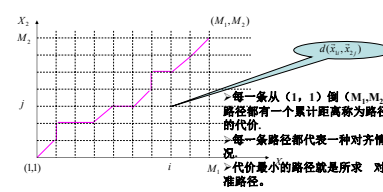
非特定人大词表连续语音识别任务

隐马尔科夫模型方法 (HMM)

5

语音识别算法

► 如何对准

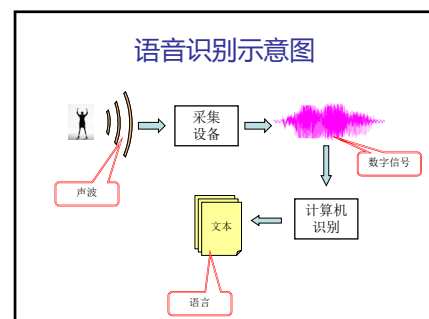


每一条从 $(1, 1)$ 到 (M_1, M_2) 的路径都有一个累计距离称为路径的代价

每一条路径都代表一种对齐情况

M_1 代价最小的路径就是所求 对准路径。

8



3

语音识别算法

► 动态时间归正

DTW (Dynamic Time Warping) 是一种模板匹配技术, 是基于相似度计算与匹配实现的识别方法。

- 计算两个标量 x_1 和 x_2 的相似度

$$d = |x_1 - x_2|$$
- 计算两个矢量 $\vec{x}_1 = \{x_{11}, \dots, x_{1n}\}$ 和 $\vec{x}_2 = \{x_{21}, \dots, x_{2n}\}$ 的相似度

$$d(\vec{x}_1, \vec{x}_2) = \sum_{i=1}^n (x_{1i} - x_{2i})^2$$

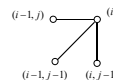
欧氏距
- 经过预处理和特征提取后的语音可以看作矢量的序列

$$X = (\vec{x}_1, \vec{x}_2, \dots, \vec{x}_M)$$

6

语音识别算法

- 将对准问题, 或者说求两个语音段的相似度问题, 转化成了搜索代价最小的最优路径问题。
- 事实上, 在搜索过程中, 往往要进行路径的限制
 - 起点/终点的限制
 - 连续性限制



再此限制条件下, 可以将全局最优化问题转化为许多局部最优化问题一步一步地来求解, 这就动态规划 (Dynamic Programming, 简称DP) 的思想。

9

语音识别算法

定义一个代价函数 $\Phi(i, j)$ 表示从起始点(1,1)出发, 到达 (i, j) 点最小代价路径的累计距离。

有: $\Phi(i, j) = \min_{(i', j') \rightarrow (i, j)} \{\Phi(i', j') + d(\tilde{x}_i, \tilde{x}_{j'})\} w_a$

则: $\Phi(M_1, M_2) = \min\{\Phi(M_1 - 1, M_2) + d(\tilde{x}_{M_1}, \tilde{x}_{M_2})w_a,$

$\Phi(M_1, M_2 - 1) + d(\tilde{x}_{M_1}, \tilde{x}_{M_2})w_a,$

$\Phi(M_1 - 1, M_2 - 1) + d(\tilde{x}_{M_1}, \tilde{x}_{M_2})w_a\}$

依次类推, $\Phi(M_1 - 1, M_2)$ 、 $\Phi(M_1, M_2 - 1)$ 、 $\Phi(M_1 - 1, M_2 - 2)$ 可由更低一层的代价函数计算得到。

10

语音识别算法

(2) 递推求累计距离 并记录回溯信息

$\Phi(i, j) = \min\{\Phi(i-1, j) + d(\tilde{x}_i, \tilde{x}_j) \cdot W_a(1); \Phi(i-1, j-1) + d(\tilde{x}_i, \tilde{x}_j) \cdot W_a(2);$

$\Phi(i, j-1) + d(\tilde{x}_i, \tilde{x}_j) \cdot W_a(3)\}$

$i = 2, 3, \dots, M_1; j = 2, 3, \dots, M_2; (i, j) \in \text{Reg}$

一般取距离加权值为, $W_a(1) = W_a(3) = 1$ $W_a(2) = 2$

并将 (i, j) 点的回溯信息记录在 $p(i, j)$ 中

13

语音识别算法

若 $D(X_1, X_2) < \sigma$ 且最优路径为

$(i(1), j(1)), (i(2), j(2)), \dots, (i_{T_y}(T_y), j_{T_y}(T_y))$

则可得到新的模板 Y , 长度为 T_y

$$Y_k = \frac{1}{2}(x_{i(k)} + x_{j(k)}), \quad k = 1, 2, \dots, T_y$$

比偶然训练法可靠, 但不充分。当识别任务是针对非特定人时, 这种问题更为突出。

16

语音识别算法

• 这样就可从

$\Phi(1,1)$ 逐步向上搜索。

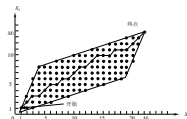
• 加权系数的取值与局部路径有关

$$w_a = \begin{cases} 2 & (i-1, j-1) \rightarrow (i, j) \\ 1 & \text{其它} \end{cases}$$

• 定义回溯函数

$P(i, j)$

• 平行四边形区域约束



11

语音识别算法

(3) 回溯求出所有的匹配点对: 根据每步的上一步最佳局部路径 $p(i, j)$, 由匹配点 (M_1, M_2) 对向前回溯一直到 (1,1)。这个回溯过程对于求平均模板或聚类中心来讲是必不可少的, 但在识别过程往往不必进行。

• 对所求得的 $\Phi(M_1, M_2)$ 还需用 $\sum W_a$ 来归正

14

语音识别算法

► 非特定人识别的模板训练算法—聚类方法

令 Ω 为 L 个训练序列的集合, $\Omega = \{X_1, X_2, \dots, X_L\}$, 其中, 每个元素为某特定语音的一次实现, 即一次发音。对每两次发音的特征矢量序列进行匹配计算, 得到的匹配距离 $d(X_i, X_j)$, 则可构成一个 $L \times L$ 的距离矩阵。聚类的目的是将训练集 Ω 聚成 V 个不同的类 $\{c_k; k=1, 2, \dots, V\}$, 使 $\Omega = \bigcup_{k=1}^V c_k$, 在同一类中的语音模式比较相近。

[MKM聚类算法.doc](#)

17

语音识别算法

DTW 路径搜索算法

(1) 初始化: $i = j = 1, \Phi(1, 1) = d(\tilde{x}_{11}, \tilde{x}_{21})$

$$\Phi(i, j) = \begin{cases} 0 & \text{当 } (i, j) \in \text{Reg} \\ \text{huge} & \text{当 } (i, j) \notin \text{Reg} \end{cases}$$

其中约束区域 Reg 可以假定是这样一平行四边形, 它有两个顶点位于 (1,1) 和 (M_1, M_2) , 相邻两条边的斜率分别为 2 和 1/2。

12

语音识别算法

• 模板的训练

1 偶然训练法

将每个词的每一遍语音形成一个模板。在识别时, 待识别矢量序列用 DTW 算法分别求得与每个模板的累计失真, 综合在一起形成总失真。这种方法具有很大的偶然性。

2 稳健模板训练方法

这种方法将每个词重复说多遍, 直到得到一对一致性较好的特征矢量序列。最终得到的模板是在一致性较好的特征矢量序列对在沿 DTW 的路径上求平均。

15

语音识别算法

课堂练习:

要求: 编制 DTW 匹配程序

输入: 语音矢量序列 X_1, X_2

输出: X_1, X_2 的相似度得分

18

马尔可夫链

19

马尔可夫性质示例

◆ 抛硬币输赢模型

假设甲乙两人以抛硬币的方式进行赌博，每次抛同一枚硬币；若出现正面，则甲付给乙一元钱，若出现反面，则乙付给甲一元钱。记 X_n 为第 n 局之后甲赢的总钱数。则 $\{X_n, n \geq 0\}$ 是马尔可夫链。

22

马尔科夫链的定义

定义 设随机过程 $\{X_n, n \geq 0\}$ 的状态空间为： $S = \{0, 1, 2, 3, \dots\}$

若对任意的 $n \geq 0$ ，及 $i_0, i_1, \dots, i_{n-1}, i, j \in S$ 有

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\}$$

则称 $\{X_n, n \geq 0\}$ 为离散时间、离散状态的马尔可夫过程，或简称为马尔可夫链。

马氏性

25

马尔可夫链

- ◆ 因俄国数学家安德烈·马尔可夫而得名，是他在1906年提出来的。
- ◆ 状态空间是有限的或可列；
- ◆ 是一种离散时间随机过程。即指标集 $T = \{0, 1, 2, \dots\}$ ；
- ◆ 具有马尔可夫性质（无后效性）。即在给定当前知识或信息的情况下，过去（即当期以前的历史状态）对于预测将来（即当期以后的未来状态）是无关的。

20

马尔可夫性质示例

- ◆ 有时是计算处理的需要；
- ◆ 计算符号串的概率

如拼音输入法
拼音串：wo zai deng ni

对应字串：我在等你

我在瞪你

我载邓妮

窝仔灯拟

.....

决策时需要计算每个可能的字串的概率

23

马尔科夫链的定义

- ◆ 字符串的概率可计算为：

$$\begin{aligned} P(X_0 = i_0, X_1 = i_1, \dots, X_{k-1} = i_{k-1}, X_k = i_k) \\ = P(X_k = i_k | X_{k-1} = i_{k-1}) \cdot P(X_{k-1} = i_{k-1} | X_{k-2} = i_{k-2}) \cdot \\ \dots \cdot \\ P(X_2 = i_2 | X_1 = i_1) \cdot P(X_1 = i_1 | X_0 = i_0) \cdot P(X_0 = i_0) \end{aligned}$$

- ◆ 即马尔可夫链 $\{X_n, n \geq 0\}$ 的有限维分布完全由初始分布

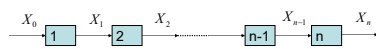
$P\{X_0 = i\}$ 和条件概率 $P\{X_n = j | X_{n-1} = i\}$ 确定。

26

马尔可夫性质示例

◆ 简单信号模型

在某数字通讯系统中，只传输 0、1 两种信号，且传输要经过很多级，且每级中由于噪声的存在会引起误差。假设每级输入 0、1 信号后，其输出不产生误差的概率为 X_n 。记 $\{X_n, n \geq 0\}$ 为第 n 级的输出信号。则它是状态有限的马尔可夫链。



21

定义----马尔可夫性质示例

- ◆ 计算字符串 $\{X_n, n \geq 0\}$ 的发生概率

$$\begin{aligned} P(X_0 = i_0) \\ = P(X_k = i_k) \end{aligned}$$

各条件概率理论上是可以估计的，因而可以事先得到并存储。

计算特定字符串的概率时，通过查表来得到。

然而，存储各条件概率的数据表需要多少存储空间呢？

$$\begin{aligned} P(X_2 = i_2 | X_0 = i_0, X_1 = i_1) \\ P(X_1 = i_1 | X_0 = i_0) P(X_0 = i_0) \end{aligned}$$

24

马尔科夫链的定义

定义1 设 $\{X_n, n \geq 0\}$ 是马尔可夫链，记

$$a_{ij}(n) = P\{X_{n+1} = j | X_n = i\}$$

称 $a_{ij}(n)$ 为马尔可夫链 $\{X_n, n \geq 0\}$ 在时刻 n 时的一步转移概率。有：

$$\begin{aligned} a_{ij}(n) \geq 0, \quad \forall i, j \in S, n > 0; \\ \sum_{j \in S} a_{ij}(n) = 1, \quad \forall i \in S, n > 0. \end{aligned}$$

若其一步转移概率 $a_{ij}(n)$ 与时间 n 无关，即：

$$a_{ij} = P\{X_{n+1} = j | X_n = i\} = P\{X_1 = j | X_0 = i\}$$

则称 $\{X_n, n \geq 0\}$ 为齐次马尔可夫链

27

马尔科夫链的定义

齐次马尔科夫链的一步转移概率矩阵,

$$A = (a_{ij}) = \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0j} & \cdots \\ a_{10} & a_{11} & a_{12} & \cdots & a_{1j} & \cdots \\ a_{20} & a_{21} & a_{22} & \cdots & a_{2j} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\ a_{i0} & a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix}$$

矩阵的每一行都是一条件分布律

记 $\pi = (\pi_0, \pi_1, \cdots)$, $(\pi_i = P\{X_0 = i\}, i \in S)$, 称 π 为齐次马尔科夫链的初始分布.

28

转移概率和初始分布

例3 某计算机机房的一台计算机经常出故障, 研究者每隔15分钟观察一次计算机的运行状态, 收集了24个小时的数(共作97次观察), 用1表示正常状态, 用0表示不正常状态, 所得的数据序列如下:

11100100111111100111101111110011111111000110110
11110110110101111011110111110011011111100111

设 X_n 为第 $n(n=1, 2, \dots, 97)$ 个时段的计算机状态, 可以认为它是一个齐次马氏链. 求

(1) 一步转移概率矩阵;

(2) 已知计算机在某一时段(15分钟)的状态为0, 问在此条件下, 从此时段起, 该计算机能连续正常工作45分钟(3个时段)的条件概率.

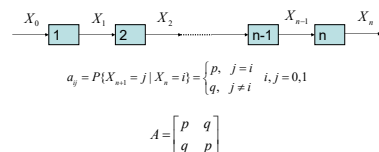
31

隐马尔可夫模型

34

转移概率和初始分布

例1 简单信号模型的转移概率矩阵



29

转移概率和初始分布

解: (1) 设 X_n 为第 $n(n=1, 2, \dots, 97)$ 个时段的计算机状态, 可以认为它是一个齐次马氏链, 状态空 $S = \{0, 1\}$,

96次状态转移情况是: $0 \rightarrow 0$: 8次; $0 \rightarrow 1$: 18次;

$1 \rightarrow 0$: 18次; $1 \rightarrow 1$: 52次;

因此一步转移概率可用频率近似地表示为:

$$a_{00} = P(X_{n+1} = 0 | X_n = 0) \approx \frac{8}{8+18} = \frac{8}{26}$$

$$a_{01} = P(X_{n+1} = 1 | X_n = 0) \approx \frac{18}{8+18} = \frac{18}{26} \quad \text{即: } A = \begin{bmatrix} \frac{8}{26} & \frac{18}{26} \\ \frac{18}{70} & \frac{52}{70} \end{bmatrix}$$

$$a_{10} = P(X_{n+1} = 0 | X_n = 1) \approx \frac{18}{18+52} = \frac{18}{70}$$

$$a_{11} = P(X_{n+1} = 1 | X_n = 1) \approx \frac{52}{18+52} = \frac{52}{70}$$

32

语音识别算法

- 实际问题比Markov链模型所描述的更为复杂. 观察到的事件并不是与状态一一对应, 而是通过一组概率分布相联系.
- 使用双重随机过程来描述模型, 一个是Markov链, 描述状态的转移; 另一个随机过程描述状态和观察值之间的统计对应关系.
- 由于状态是不可见的, 因此称之为“隐” Markov 模型.

35

转移概率和初始分布

例2 (一个简单的疾病死亡模型)

考虑一个包含两个健康状态 S_1 和 S_2 以及两个死亡状态 S_3 和 S_4 (即由不同原因引起的死亡) 的模型. 若个体病愈, 则认为它处于状态 S_1 ; 若患病, 则认为它处于 S_2 . 个体可以从 S_1 , S_2 进入 S_3 和 S_4 . 易见这是一个马氏链, 转移矩阵为

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

30

转移概率和初始分布

(2) 某一时段的状态为0, 定义其为初始状态, 即 $X_0 = 0$, 所求概率为:

$$P(X_1 = 1, X_2 = 1 | X_0 = 0)$$

$$= P(X_1 = 1 | X_0 = 0) P(X_2 = 1 | X_0 = 0, X_1 = 1)$$

$$\times P(X_3 = 1 | X_0 = 0, X_1 = 1, X_2 = 1)$$

$$= a_{01} a_{11} a_{11}$$

$$= \frac{18}{26} \frac{52}{70} \frac{52}{70} = 0.382$$

33

语音识别算法

• 一个HMM的例子: Ball and Urn



$$P(\text{红}) = b_1(1) \quad P(\text{红}) = b_2(1) \quad P(\text{红}) = b_N(1)$$

$$P(\text{绿}) = b_1(2) \quad P(\text{绿}) = b_2(2) \quad P(\text{绿}) = b_N(2)$$

$$P(\text{蓝}) = b_1(3) \quad P(\text{蓝}) = b_2(3) \quad P(\text{蓝}) = b_N(3)$$

36

语音识别算法

◆一个HMM可以由下列参数描述

初始状态概率 $\pi = (\pi_1, \dots, \pi_N)$

$$\pi_i = P(q_1 = \theta_i) \quad 1 \leq i \leq N$$

状态转移概率矩阵 $A = (a_{ij})_{N \times N}$

$$a_{ij} = P(q_{t+1} = \theta_j | q_t = \theta_i) \quad \text{单高斯概率密度函数 或 混合高斯概率密度函数}$$

观察概率序列 $B = (b_1(o), b_2(o), \dots, b_N(o))$

$$b_i(o) = \sum_{j=1}^K c_j N(o; \mu_j, \Sigma_j)$$

◆一个HMM的参数组为:

$$\lambda = (\pi, A, B)$$

37

语音识别算法

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) \\ = \sum_{Q} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T)$$

计算量: $2TN^T$ 当 $N=5, T=100$ 时, 计算量达 10^{12}

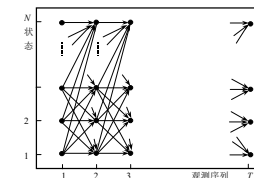
■ 前向—后向算法

定义前向变量为: $\alpha_i(i) = P(o_1 o_2 \cdots o_i, q_i = i | \lambda)$

40

语音识别算法

格型结构



43

语音识别算法

◆ HMM的三个基本问题

- 1 已知一个HMM参数组 $\lambda = (\pi, A, B)$, 和给定一个观察序列 $O = o_1 o_2 \cdots o_T$ 的条件下, 如何计算在给定模型 λ 条件下观察序列 O 的概率 $P(O|\lambda)$ 。
- 2 如何确定最佳状态序列 $Q = q_1 q_2 \cdots q_T$, 以最好的解释观察序列 O 。
- 3 给定一组观察序列的集合 $\{O_n\}$, 如何调整参数 λ , 以使 $P(\{O_n\}|\lambda)$ 达到最大值。

38

语音识别算法

前向变量有如下性质:

- (1) 初值易求 $\alpha_i(i) = P(o_1, q_1 = i) = \pi_i b_i(o_1)$
- (2) 可以计算 $P(O|\lambda) = \sum_{i=1}^N \alpha_T(i)$
- (3) 有递推关系 $\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1})$

因此可以利用递推关系, 逐层递推, 计算出全部 $\alpha_t(j) \quad 1 \leq t \leq T-1 \quad 1 \leq j \leq N$ 。最后再由 $\alpha_T(i)$ 计算得到 $P(O|\lambda)$

41

语音识别算法

定义后向变量为

$$\beta_i(i) = P(o_{i+1} o_{i+2} \cdots o_T | q_i = i, \lambda)$$

有初值

$$\beta_T(i) = b_i(o_T)$$

有递推关系

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

且易知

$$P(O|\lambda) = \sum_{j=1}^N \alpha_1(j) \beta_1(j)$$

44

语音识别算法

■ 计算概率 $P(O|\lambda)$ 先计算 $P(O, Q|\lambda)$, 其中 Q 为一给定的状态序列

$$Q = q_1 q_2 \cdots q_T$$

有: $P(O, Q|\lambda) = P(O|Q, \lambda) P(Q|\lambda)$

而

$$P(O|Q, \lambda) = \prod_{t=1}^T P(o_t | q_t, \lambda) = b_{q_1}(o_1) b_{q_2}(o_2) \cdots b_{q_T}(o_T)$$

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} \cdots a_{q_{T-1} q_T}$$

所以 $P(O, Q|\lambda) = \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T)$

39

语音识别算法

(a) 初始化: 对 $1 \leq i \leq N$

$$\alpha_i(i) = \pi_i b_i(o_1)$$

(b) 递推: 对 $1 \leq t \leq T-1, 1 \leq j \leq N$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1})$$

(c) 终止:

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i)$$

计算量为 $N^2 T$, $N=5, T=100$ 时, 只需2500次乘法运算

42

语音识别算法

• 最佳状态链的确定

确定一个最佳状态序列 $Q^* = q_1^*, q_2^*, \dots, q_T^*$, 使 $P(O, Q^*|\lambda)$ 为最大。

$$Q^* = \arg \max_Q P(O, Q|\lambda)$$

Viterbi算法

定义 $\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P(q_1 q_2 \cdots q_{t-1}, q_t = i, o_1 o_2 \cdots o_t | \lambda)$

为在时刻 t , 沿一条路径 q_1, q_2, \dots, q_t , 且 $q_t = i$, 产生出 $o_1 o_2 \cdots o_t$ 的最大概率

45

语音识别算法

是否满足DP算法的三个条件

1 初值易求:

$$\delta_1(i) = P(q_1 = i, a_1 | \lambda) = \pi_i b_i(a_1)$$

2 能够解决问题:

$$P(O, Q^* | \lambda) = \max_{q_1, \dots, q_T} P(o_1, o_2, \dots, o_T, q_1, \dots, q_{T-1}, q_T = i | \lambda) = \max_i \delta_T(i)$$

3 有递推关系:

46

语音识别算法

c) 终止:

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)]$$

$$q_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)]$$

路径回溯, 确定最佳状态序列:

$$q_t = \varphi_{t+1}(q_{t+1}), \quad t = T-1, T-2, \dots, 1$$

49

语音识别算法

$$\begin{aligned} \hat{\mu}, \hat{\sigma} &= \arg \max_{\mu, \sigma} \sum_{i=1}^N \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \ln \sigma - \ln \sqrt{2\pi} \right) \\ &= \arg \max_{\mu, \sigma} \sum_{i=1}^N \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \ln \sigma \right) \quad \text{目标函数 } J(\mu, \sigma) \\ &= \arg \max_{\mu, \sigma} \ln \left(\sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} + \ln \sigma \right) \\ \text{求极值} \quad \frac{\partial J}{\partial \mu} &= \sum_{i=1}^N \frac{-2(x_i - \mu)}{2\sigma^2} = -\frac{1}{\sigma^2} \sum_{i=1}^N (x_i - \mu) \\ &= -\frac{1}{\sigma^2} \left(N\mu - \sum_{i=1}^N x_i \right) = 0 \quad \downarrow \\ \hat{\mu} &= \frac{1}{N} \sum_{i=1}^N x_i \\ \frac{\partial J}{\partial \sigma} &= \sum_{i=1}^N \frac{-1}{\sigma^3} = -\frac{N}{\sigma^3} \\ &= -\frac{1}{\sigma^3} \left(N - \sum_{i=1}^N \frac{(x_i - \mu)^2}{\sigma^2} \right) = 0 \quad \downarrow \\ \hat{\sigma}^2 &= \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \end{aligned}$$

52

语音识别算法

若已知 $\delta_t(i)$ 对应的最佳状态链 $q_1, q_2, \dots, q_{t-1}, q_t = j$

可以证明 $\delta_{t+1}(j)$ 对应的最佳状态链为 q_1, q_2, \dots, q_{t-1}

即若已知 $q_{t-1} = j$ 则 $\delta_t(i) = \delta_{t-1}(j) a_{ji} b_j(a_t)$

实际上不知道 $q_{t-1} = j$, 可以遍历所有的 q_{t-1} 求最大值:

$$\delta_t(i) = \max_{1 \leq j \leq N} [\delta_{t-1}(j) a_{ji} b_j(a_t)]$$

可以用回溯的方式求出 Q^*

47

语音识别算法

$\max_Q P(O, Q | \lambda)$ 实际上是 $\sum_Q P(O, Q | \lambda)$ 中举足轻重的唯一成分, 因此, 常常等价地使用 $\max_Q P(O, Q | \lambda)$ 来近似 $\sum_Q P(O, Q | \lambda)$ 。即Viterbi算法也就能用来计算 $P(O | \lambda)$ 。

在连接词和连续语音识别中, 更多地采用Viterbi算法来进行识别操作。因为它不仅能计算得分, 还能通过最佳状态链获得**词的边界信息**。

50

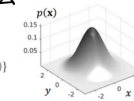
语音识别算法

MLE—多元高斯分布

$$p(x) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

有

$$\begin{aligned} \hat{\mu}, \hat{\Sigma} &= \arg \max_{\mu, \Sigma} \ln \prod_{i=1}^N p(x_i | \mu, \Sigma) \\ &= \arg \max_{\mu, \Sigma} \sum_{i=1}^N \ln p(x_i | \mu, \Sigma) \\ &= \arg \max_{\mu, \Sigma} \sum_{i=1}^N \left(-\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) - \frac{1}{2} \ln |\Sigma| - \frac{D}{2} \ln (2\pi) \right) \\ &= \arg \max_{\mu, \Sigma} \sum_{i=1}^N \left(\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \frac{1}{2} \ln |\Sigma| \right) \end{aligned}$$



53

语音识别算法

那么, 求取最佳状态序列 Q 的过程为

(a) 初始化: 对 $1 \leq i \leq N$

$$\hat{q}_1(i) = \pi_i b_i(a_1)$$

$$q_1(i) = 0$$

(b) 递推: 对 $2 \leq t \leq T, 1 \leq j \leq N$

$$\hat{q}_t(j) = \max_{1 \leq i \leq N} [\hat{q}_{t-1}(i) a_{ij} b_j(a_t)]$$

$$q_t(j) = \arg \max_{1 \leq i \leq N} [\hat{q}_{t-1}(i) a_{ij}]$$

48

语音识别算法

• MLE和EM

— MLE—一元高斯分布

$$\hat{\mu}, \hat{\sigma} = \arg \max_{\mu, \sigma} p(\{x_i\} | \mu, \sigma) = \arg \max_{\mu, \sigma} \prod_{i=1}^N p(x_i | \mu, \sigma)$$

改变目标函数

$$\begin{aligned} \hat{\mu}, \hat{\sigma} &= \arg \max_{\mu, \sigma} \ln \left\{ \prod_{i=1}^N p(x_i | \mu, \sigma) \right\} \\ &= \arg \max_{\mu, \sigma} \sum_{i=1}^N \ln p(x_i | \mu, \sigma) \end{aligned}$$

51

语音识别算法

$$\begin{aligned} \frac{\partial J}{\partial \mu} &= \frac{\partial}{\partial \mu} \sum_{i=1}^N \left[\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \frac{1}{2} \ln |\Sigma| \right] \\ &= \frac{\partial}{\partial \mu} \sum_{i=1}^N \left[\frac{1}{2} (x_i^T \Sigma^{-1} x_i - x_i^T \Sigma^{-1} \mu - \mu^T \Sigma^{-1} x_i + \mu^T \Sigma^{-1} \mu) \right] \\ &= \frac{\partial}{\partial \mu} \sum_{i=1}^N \left[\frac{1}{2} \mu^T \Sigma^{-1} \mu - \mu^T \Sigma^{-1} x_i \right] \\ &= \Sigma^{-1} \sum_{i=1}^N (x_i - \mu) \\ \frac{\partial J}{\partial \Sigma} &= \frac{\partial}{\partial \Sigma} \sum_{i=1}^N \left[\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) + \frac{1}{2} \ln |\Sigma| \right] \\ &= \frac{\partial}{\partial \Sigma} \sum_{i=1}^N \left[\frac{1}{2} \text{tr} \left(\Sigma^{-1} (x_i - \mu)(x_i - \mu)^T \right) + \frac{1}{2} \ln |\Sigma| \right] \\ &= \frac{\partial}{\partial \Sigma} \sum_{i=1}^N \left[\frac{1}{2} \text{tr} \left(\Sigma^{-1} x_i x_i^T - \Sigma^{-1} x_i \mu^T - \mu x_i^T \Sigma^{-1} + \mu \mu^T \Sigma^{-1} \right) + \frac{1}{2} \ln |\Sigma| \right] \end{aligned}$$

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

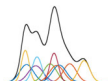
$$\begin{aligned} \frac{dx^T A}{dx} &= A \\ \frac{dAx}{dx} &= A^T \\ \frac{\partial u^T v}{\partial x} &= \frac{\partial u^T}{\partial x} v + \frac{\partial v^T}{\partial x} u^T \\ &= \mu^T \end{aligned}$$

54

语音识别算法

EM—混合高斯分布

$$p(x) = \sum_{k=1}^K w_k g_k(x | \mu_k, \Sigma_k)$$



- 由于求和式出现对数函数中，因而求极值而得到的方程组不是线性方程组，无法求解
- 用迭代的方法求解，先给个初值，然后认为均值、协方差矩阵已知，去估计权重，再根据权重，去估计均值和方差

55

语音识别算法

可以证明：若 $Q(\bar{\lambda}, \lambda) \geq Q(\lambda, \lambda)$ 则有 $P(O|\bar{\lambda}) \geq P(O|\lambda)$

$$\begin{aligned} Q(\bar{\lambda}, \lambda) - Q(\lambda, \lambda) &= \sum_Q P(Q, O|\lambda) \log \frac{P(Q, O|\bar{\lambda})}{P(Q, O|\lambda)} \\ &\leq \sum_Q P(Q, O|\lambda) \left(\frac{P(Q, O|\bar{\lambda})}{P(Q, O|\lambda)} - 1 \right) \\ &= P(O|\bar{\lambda}) - P(O|\lambda). \end{aligned}$$

58

语音识别算法

 a_y 的重估

$$\varphi(\lambda, \bar{\lambda}) = \sum_{t=1}^T \sum_{i=1}^N \log \bar{a}_y p(O, q_{t-1} = s, q_t = s_j | \lambda) + \varphi_2$$

以及约束条件 $\sum_{i=1}^N a_y = 1$

以此类推，去重估观察概率。

61

语音识别算法

• HMM模型的训练

给定一个观察值序列 $O = o_1, o_2, \dots, o_T$ 确定一个 $\lambda = (\pi, A, B)$ ，使 $P(O|\lambda)$ 最大。

实际上，不存在一种方法直接估计最佳的 λ 。

替代的方法是：

根据观察值序列选取初始模型 $\lambda = (\pi, A, B)$ ，然后依据某种方法求得一组新参数 $\bar{\lambda} = (\bar{\pi}, \bar{A}, \bar{B})$ ，保证有 $P(O|\bar{\lambda}) > P(O|\lambda)$ 。重复这个过程，逐步改进模型参数，直到 $P(O|\bar{\lambda})$ 收敛。

56

语音识别算法

计算 $\frac{\partial Q(\lambda, \bar{\lambda})}{\partial \lambda} = 0$ ，得到一组求取 $\bar{\lambda}$ 的公式，这一组公式就称为重估 (Re-Estimation) 公式，它们是 Baum-Welch 算法的核心内容。

$$\begin{aligned} Q(\bar{\lambda}, \lambda) &= \sum_Q P(Q, O|\lambda) \log P(Q, O|\bar{\lambda}) \\ &= \sum_Q \log(\bar{\pi}_n \bar{b}_{n1}(o_1)) \prod_{t=1}^T \bar{a}_{n_t, n_{t+1}} \bar{b}_{n_{t+1}}(o_{t+1}) P(O, Q|\lambda) \\ &= \sum_Q \log \bar{\pi}_n P(O, Q|\lambda) + \sum_Q \left(\sum_{t=1}^T \log \bar{a}_{n_t, n_{t+1}} \right) P(O, Q|\lambda) + \sum_Q \left(\sum_{t=1}^T \log \bar{b}_{n_{t+1}}(o_{t+1}) \right) P(O, Q|\lambda) \end{aligned}$$

59

语音识别算法

Baum-Welch算法

定义 $\xi_t(i, j)$ 为给定训练序列 O 和模型 λ 时，HMM模型在 t 时刻处于状态 i ， $t+1$ 时刻处于状态 j 的概率。

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j | O, \lambda)$$

易证

$$\xi_t(i, j) = [\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)] / P(O|\lambda)$$

定义HMM模型在 t 时刻处于状态 i 的概率为 $\gamma_t(i)$ 。

$$\gamma_t(i) = P(q_t = i | O, \lambda) = \sum_{j=1}^N \xi_t(i, j) = \alpha_t(i) \beta_t(i) / P(O|\lambda)$$

62

语音识别算法

- 这一方法，未必能求得全局最大值、而有可能得到一局部极值点
- 经典的方法：Baum-Welch算法。
- Baum-Welch算法的理论基础是EM算法。

定义辅助函数

$$Q(\bar{\lambda}, \lambda) = \sum_Q P(Q, O|\lambda) \log P(Q, O|\bar{\lambda}).$$

57

语音识别算法

 π_i 的重估

$$Q(\lambda, \bar{\lambda}) = \sum_Q \log \bar{\pi}_n P(O, Q|\lambda) + \varphi_1 = \sum_{n=1}^N \log \bar{\pi}_n P(O, q_1 = n | \lambda) + \varphi_1$$

以及约束条件 $\sum_{i=1}^N \pi_i = 1$

根据拉格朗日乘子法

$$\begin{aligned} \frac{\partial}{\partial \bar{\pi}_n} \left(\sum_{i=1}^N \log \bar{\pi}_i P(O, q_1 = i | \lambda) + \varphi_1 + r \left(1 - \sum_{i=1}^N \bar{\pi}_i \right) \right) &= 0 \\ \bar{\pi}_i &= \frac{P(O, q_1 = i | \lambda)}{P(O|\lambda)} = \frac{\alpha_i(1) \beta_i(1)}{\sum_{i=1}^N \alpha_i(1) \beta_i(1)} \end{aligned}$$

60

语音识别算法

重估公式可写成如下形式

$$\bar{\pi}_i = \gamma_1(i) \\ \bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

若观察概率采用离散值

$$\bar{b}_{ik} = \frac{\sum_{t=1}^T \gamma_t(i)}{\sum_{i_1=i_2}^T \gamma_t(i)}$$

63

语音识别算法

若观察概率为多维连续高斯概率密度函数形式，即

$$b_i(o) = N(o; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{(o - \mu_i)^T \Sigma_i^{-1} (o - \mu_i)}{2} \right\}$$

则

$$\bar{\mu}_i = \frac{\sum_{t=1}^T \gamma_t(i) o_t}{\sum_{t=1}^T \gamma_t(i)} \quad \bar{\Sigma}_i = \frac{\sum_{t=1}^T \gamma_t(i) (o_t - \bar{\mu}_i)(o_t - \bar{\mu}_i)^T}{\sum_{t=1}^T \gamma_t(i)}$$

64

语音识别算法

HMM算法实现中的问题

- 初始模型选取

初始模型的选取对Baum-Welch算法的结果有巨大影响。只有选取好的初始模型，才能使最后求出的局部极大与全局最大相接近。

最常采用的是一种基于Viterbi算法的初始模型选取方法。

67

语音识别算法

将 $\bar{\mu}_i$ 估计为所有状态标号为 i 的特征矢量的样本均值
将 $\bar{\Sigma}_i$ 估计为所有状态标号为 i 的特征矢量第 d 维的方差

若观察概率为混合高斯概率密度函数形式

$$b_i(o) = \sum_{k=1}^K c_{ik} N(o; \mu_{ik}, \Sigma_{ik})$$

需要将状态标号为 i 的特征矢量进行聚类，聚成 K 类，在每一类的样本中估计 $\bar{\mu}_{ik}, \bar{\Sigma}_{ik}$

$$\bar{c}_{ik} = \frac{\text{状态标号为 } i \text{ 的特征矢量中属于第 } k \text{ 类的数量}}{\text{状态标号为 } i \text{ 的特征矢量的数量}}$$

70

语音识别算法

若观察概率为混合高斯分布形式，即

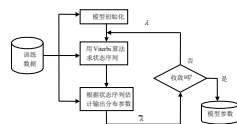
$$b_j(o_t) = \sum_{k=1}^K c_{jk} N(o_t; \mu_{jk}, \Sigma_{jk})$$

则重估公式写为

$$\bar{c}_{jk} = \frac{\sum_{t=1}^T v_t(j, k)}{\sum_{t=1}^T v_t(j)} \quad \bar{\mu}_{jk} = \frac{\sum_{t=1}^T v_t(j, k) o_t}{\sum_{t=1}^T v_t(j, k)}$$

65

语音识别算法



68

语音识别算法

- 多个观察值序列训练

用 L 个观察序列训练HMM时，要对Baum-Welch算法的重估公式加以修正。

$$\bar{\pi}_i = \frac{\sum_{l=1}^L \alpha_i^{(l)}(i) \beta_i^{(l)}(i) / P(O^{(l)} | \lambda)}{\sum_{l=1}^L \sum_{t=1}^{T-1} \sum_{j=1}^N \alpha_i^{(l)}(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}^{(l)}(j) / P(O^{(l)} | \lambda)}$$

$$\bar{a}_{ij} = \frac{\sum_{l=1}^L \sum_{t=1}^{T-1} \alpha_i^{(l)}(i) \beta_i^{(l)}(j) / P(O^{(l)} | \lambda)}{\sum_{l=1}^L \sum_{t=1}^{T-1} \sum_{j=1}^N \alpha_i^{(l)}(i) \beta_i^{(l)}(j) / P(O^{(l)} | \lambda)}$$

71

语音识别算法

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^T v_t(j, k) (o_t - \bar{\mu}_{jk})(o_t - \bar{\mu}_{jk})^T}{\sum_{t=1}^T v_t(j, k)}$$

式中

$$v_t(j, k) = \frac{\sum_{i=1}^N \alpha_{t-1}(i) a_{ij} c_{jk} b_{jk}(o_t) \beta_t(j)}{P(O | \lambda)}$$

66

语音识别算法

- 根据状态序列重估

$\bar{\pi}_i$ = 以状态 i 起始的语音段数量 / 语音段总数量

\bar{a}_{ij} = 状态 i 后出现状态 j 的数量 / 状态 i 的数量

若观察概率为单高斯概率密度函数形式

$$b_i(o) = N(o; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{(o - \mu_i)^T \Sigma_i^{-1} (o - \mu_i)}{2} \right\}$$

一般 Σ_i 采用对角阵形式

$$\Sigma_i = \begin{bmatrix} \sigma_{i1}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{iD}^2 \end{bmatrix}$$

69

语音识别算法

$$\bar{b}_{jk} = \frac{\sum_{l=1}^L \sum_{t=1}^{T-1} \alpha_i^{(l)}(i) \beta_i^{(l)}(j) / P(O^{(l)} | \lambda)}{\sum_{l=1}^L \sum_{t=1}^{T-1} \sum_{j=1}^N \alpha_i^{(l)}(i) \beta_i^{(l)}(j) / P(O^{(l)} | \lambda)}$$

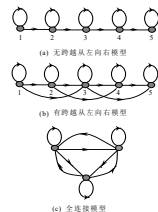
- 数据下溢问题

用对数似然度，取代概率值

72

语音识别算法

• Markov链的形状和HMM的类型



73

语音识别算法

➤ 连接词语音识别技术

• 连接词

(1) 连续发音，不知道语音中词的个数和词的边界信息。

(2) 词表有限，可以象孤立词识别一样，以词为单位建模。

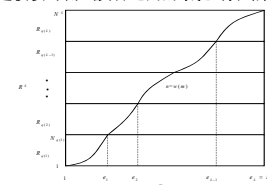
• 连接词识别

连接词识别，就是指系统存储的模板或模型是针对孤立词的，但是识别的语音却是由这些词构成的词串。

76

语音识别算法

超模板与测试发音之间的最优对齐路径



79

语音识别算法

HMM的类型:

- 离散HMM
- 连续HMM

74

语音识别算法

➤ 连接词识别问题的一般描述(从DTW的角度)

设给定测试发音的特征矢量序列为 $O = \{o(1), o(2), \dots, o(M)\}$ ，词表中 V 个词的模板分别为 R_1, R_2, \dots, R_V 。某一个参考模板 R_i 具有如下的形式：

$$R_i = \{r_i(1), r_i(2), \dots, r_i(N_i)\}$$

其中 N_i 是第 i 个词参考模板的帧数。

连接词识别的问题变为，寻找与 O 序列最优匹配的参考模板序列 R^* ， R^* 是 L 个参考模板的连接：

$$R^* = \{R_{q(1)} \oplus R_{q(2)} \oplus R_{q(3)} \oplus \dots \oplus R_{q(L)}\}$$

其中每个 $q^*(l)$ 可能是 $[1, V]$ 中任意一个模板。

77

语音识别算法

R^* 可以如下计算

$$D^* = \min_{R^*} D(R^*, O)$$

$$= \min_{1 \leq i_1 \leq L, 1 \leq i_2 \leq L, \dots, 1 \leq i_L \leq L} \min_{q(1) \in V, q(2) \in V, \dots, q(L) \in V} \sum_{m=1}^M d(o(m), r^{q(i)}(W(m)))$$

$$R^* = \arg \min_{R^*} D(R^*, O)$$

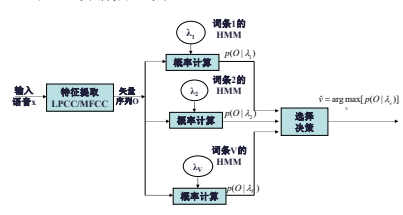
计算量太大，因此必须寻找快速算法：

- 二阶动态规划算法
- 分层构筑算法

80

语音识别算法

• 孤立词识别原理图



75

语音识别算法

一个超模板 R^*

$$R^* = R_{q(1)} \oplus R_{q(2)} \oplus R_{q(3)} \oplus \dots \oplus R_{q(L)} = \{r^*(n)\}_{n=1}^{N^*}$$

超模板与测试发音 O 之间的相似度可以用DTW来计算

$$D(R^*, O) = \min_{w(n)} \sum_{n=1}^M d(o(n), r^*(w(n)))$$

$d(\cdot, \cdot)$ 为局部特征匹配距离， $w(\cdot)$ 是时间弯折函数

78

语音识别算法

➤ 二阶动态规划算法

先定义两个函数

$$\hat{D}(b, e) = \min_{1 \leq v \leq V} [\hat{D}(v, b, e)]$$

$$\hat{N}(b, e) = \arg \min_{1 \leq v \leq V} [\hat{D}(v, b, e)]$$

$\hat{D}(v, b, e)$ 表示起始点为 b ，结束点为 e 的语音段与模板 v 之间的DTW距离

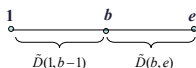
81

语音识别算法

设 $\bar{D}_l(e)$ 为有 l 个词时, 以 e 为终点的 D' , 有初值

$$\bar{D}_1(e) = \bar{D}(1, e) \quad 1 \leq e \leq M$$

看一看当词的数目确定为2时



$$D' = \min_{1 \leq b < e} [\bar{D}(1, b-1) + \bar{D}(b, e)]$$

$$\bar{D}_2(e) = \min_{1 \leq b < e} [\bar{D}_1(b-1) + \bar{D}(b, e)] \quad 1 \leq e \leq M$$

82

语音识别算法

(4) 最优解

$$D' = \min_{1 \leq b < e} [\bar{D}_l(M)]$$

(5) 回溯: 利用 D' 所对应的 $\bar{D}(b, e)$, 可以找到其对应标号 $\bar{N}(b, e)$, 以及最优路径上第 l 个模板的起始位置 b , $b-1$ 即为第 $l-1$ 个模板的结束位置 e , 通过 $\bar{D}_{l-1}(e)$ 可以找到第 $l-1$ 个模板的起始位置, 以及它的前一个模板的结束位置, 以此类推, 就可以逐步找出所有的最优模板。

85

语音识别算法

• 在上个世纪90年代初期, 取得了里程碑式的成果 (李开复和他的Sphinx)

• 基于HMM的LVCSR系统的统一框架, 将整个识别系统分为三层: 声学—语音层、词层和句法层。

❖ 声学—语音层是识别系统的底层, 它接受输入语音, 并以一种“子词 (Subword)”单位作为其识别输出, 每个子词单位对应一套HMM结构和参数。

❖ 词层规定词汇表中每个词是由什么音素—音子串接而成的

❖ 句法层中规定词按照什么规则组合成句子。

88

语音识别算法

词的数目为 l 个时的递推式

$$\bar{D}_l(e) = \min_{1 \leq b < e} [\bar{D}_{l-1}(b-1) + \bar{D}(b, e)] \quad 1 \leq e \leq M$$

而 D^*

$$D' = \min_{1 \leq b < e} [\bar{D}_l(M)]$$

83

语音识别算法

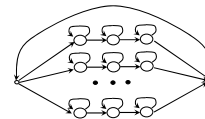
• 基于HMM的连接词识别

举例: 数字串识别

■ 0-9共10个数字, 采用3状态从左至右无跨越HMM

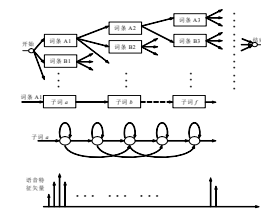
■ 形成一个新的HMM(识别网络)

■ Viterbi解码



86

语音识别算法



89

语音识别算法

算法描述:

(1) 初始化

$$\bar{D}_1(e) = \infty, \quad 1 \leq l \leq L_{\max}, \quad 0 \leq e \leq M$$

(2) 对 $l=1$

$$\bar{D}_1(e) = \bar{D}(1, e), \quad 2 \leq e \leq M$$

(3) 递推, 对 e 从 $l=2$ 到 M 进行循环

$$\bar{D}_l(e) = \min_{1 \leq b < e} [\bar{D}_{l-1}(b-1) + \bar{D}(b, e)], \quad 3 \leq e \leq M$$

$$\bar{D}_l(e) = \min_{1 \leq b < e} [\bar{D}_{l-1}(b-1) + \bar{D}(b, e)], \quad 4 \leq e \leq M$$

$$\bar{D}_l(e) = \min_{1 \leq b < e} [\bar{D}_{l-1}(b-1) + \bar{D}(b, e)], \quad l+1 \leq e \leq M$$

84

语音识别算法

► 大词汇量连续语音识别技术 (LVCSR)

• 语音识别研究中意义最重大、应用成果最丰富, 同时最具有挑战性的研究课题。

• 大词汇量非特定人的连续语音识别系统的词误识率大体为小词汇量、特定人的孤立词识别系统词误识率的50倍左右。

• 特有的问题:

❖ 词 (模式) 的数量太多, 语料不够。

❖ 发音相近的内容多, 误识严重。

87

语音识别算法

在句法层, 每个句子由若干词条组成, 需要通过语言模型评价所有可能的句子候选的合理性。

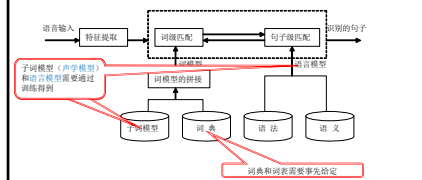
在词层, 每一个词条由若干子词串接而成, 为此需要一部词典来描述这种串接关系。

在语音层, 每一个子词用一个HMM模型及一套参数来表示。

90

语音识别算法

基于子词单元的连续语音识别系统总体框图



91

语音识别算法

声学模型

(1) 基本声学单元的选择

- 以词作为基本单元建立模型会造成大量不必要的冗余存储和计算。因此一般采用比词小的子词单元，如音节、半音节、音素等。
- 声学单元越小，其数量也就越少，训练模型的工作量也就越小；
- 但是，单元越小，对上下文的敏感性越大，越容易受到前后相邻的影响而产生变异，因此其类型设计和训练样本的采集更困难。

94

语音识别算法

- 对中文而言，音节、半音节、音素的数量都是**固定不变的**。
- 半音节 (Subsyllable) 是主要建模单元
- 汉字是单音节的，是声韵结构的，这种独特而规则的结构，使对音节、以及词条的表示变得比较规则和统一
- 使用半音节作为单元，其上下文有特殊的约束规则，Triphone的数目比较少，不是单元数的立方。

97

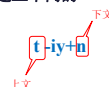
语音识别算法

$$\begin{aligned}
 W^* &= \arg \max_W P(W|O) \\
 &= \arg \max_W \frac{P(O|W)P(W)}{P(O)} \quad \text{语言学得分} \\
 &= \arg \max_W P(O|W)P(W) \quad \text{声学得分}
 \end{aligned}$$

92

语音识别算法

- 子词的数量应该是**固定不变的**。因而，**英语**只能选择**音素**作为建模单元。
- 单音素模型 (Monophone)**：每个音素建立一个HMM模型。
- 三音素模型 (Triphone)**：考虑协同发音效应，上下文不同则建立不同的HMM模型。例如：



95

语音识别算法

(2) 如何得到词模型

- 在词层用一部词典 (Dictionary) 来规定词表中每一个词是用哪些子词单元以何种方式构筑而成的。
- 最简单实用的方案是每个词用若干子词单元串接而成。
- 然而每个词的发音可能有多种变化方式
 - 替换：即词中的某个子词可能被其它相似而略有差异的子词单元所替换。
 - 插入和删除：词中有时增加了一个不是本词成分的子词单元，有时又将本词成分中的某个子词删除。

98

语音识别算法

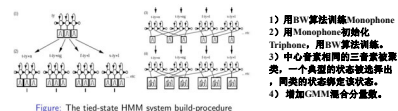
几个问题：

- 1) 基本声学单元 (子词) 的选择？
- 2) 如何得到词模型？
- 3) 如何训练子词模型？
- 4) 如何利用语言学知识？
- 5) 如何识别？

93

语音识别算法

- Triphone数量太多 (48×48×48)，建模时需要太多数据。
- 解决方法：**状态绑定**。
- Young, S., "Tree-based state tying for high accuracy acoustic modeling," Proc. ARPA Human Language Technology Workshop 1994.

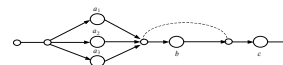


96

语音识别算法

解决方案

- 方案1：每一个词建立多套子词单元串接规则。
- 方案2：将子词单元构成词的规则用一个网络图来描述。



99

语音识别算法

(3) 基于子词单元的HMM训练

- 子词单元的HMM一般采用从左到右的结构，状态数固定为2到4个。
- 在语音段中，子词太短，无法精确标出语音的边界。
- 已知句子内容，因此可以将子词模型串接成句子。
- 用分段K均值算法进行多次迭代，对各子词模型进行重估。最终它会自动收敛于一个最佳模型估计，同时达到合理的子词分段

100

语音识别算法

➤ 语言模型

- 从一个词表中任意选择若干词所构成的序列不一定能构成自然语言中的句子，只有合乎句法者才能算是句子。
- 语言模型分为基于文法的语言模型和基于统计的语言模型。
- 在大词汇量的语音识别系统中，统计语言模型被广泛的应用。

103

语音识别算法

❖ 训练数据稀疏时的解决方法：为了避免出现 $F(W) = 0$ 或接近于零的情况，可以用三元、二元和一元相对频率做插值。

$$\hat{P}(w_3 | w_1 w_2) = p_1 \frac{F(w_1 w_2 w_3)}{F(w_1 w_2)} + p_2 \frac{F(w_3 w_2)}{F(w_2)} + p_3 \frac{F(w_3)}{\sum_i F(w_i)}$$

其中 $\sum_{i=1}^3 p_i = 1$ ， $\sum_i F(w_i)$ 是训练语料的总词数。

106

语音识别算法

分段K均值算法

- ❖ 初始化：将每个训练语句线性分割成子词单元，将每个子词单元线性分割成状态，即假定在一个语句中，子词单元及其内部的状态驻留时间是均匀的。
- ❖ 聚类：对每个给定子词单元的每一个状态，其在所有训练语句段中特征矢量用K均值算法聚类。
- ❖ 参数估计：根据聚类的结果计算均值、各维方差和混合权重系数。
- ❖ 分段：根据上一步得到的新的子词单元模型，通过Viterbi算法对所有训练语句再分成子词单元和状态，重新迭代聚类参数估计，直到收敛。

101

语音识别算法

- 统计语言模型的基本原理是，采用大量的文本资料，统计各个词的**出现概率**以及其**相互关联的条件概率**。

● 理想情况：对词串 $W = w_1, w_2, \dots, w_G$ ，

$$P(W) = P(w_1, w_2, \dots, w_G) \\ = P(w_1)P(w_2 | w_1)P(w_3 | w_1 w_2) \dots P(w_G | w_1 w_2 \dots w_{G-1})$$

● 一般采用简化模型

104

语音识别算法

(4) **N元词类文法模型**：每个词 w_i 只与其所在类 c_i 有关，而与前一时间的词所在类 c_{i-1} 中的成员无关。

$$P(W) = \sum_{c_1, c_2, \dots, c_N} \prod_{i=1}^N P(c_i | c_1 c_2 \dots c_{i-N+1}) P(w_i | c_i)$$

107

语音识别算法

- 这一过程也被称之为“**强制对齐**”(Force Alignment)
- 对齐过程合理分割了语音段，并得到**子词边界**
- 也初步估计出了每个子词的**HMM参数**。
- 以此参数为初值，采用BW算法迭代若干次即完成子词训练。

102

语音识别算法

(1) **N元文法模型**：条件概率计算时，只考虑与前 $N-1$ 个词相关。

$$P_N(W) = \prod_{i=1}^N P(w_i | w_{i-1} w_{i-2} \dots w_{i-N+1})$$

- ❖ 通常系统中采用的也只是**二元**和**三元**文法。
- ❖ N元文法统计语言模型的建立，一般是通过相对频率计数得到：

$$\hat{P}(w_i | w_{i-1} w_{i-2} \dots w_{i-N+1}) = \frac{F(w_i w_{i-1} w_{i-2} \dots w_{i-N+1})}{F(w_{i-1} w_{i-2} \dots w_{i-N+1})}$$

$F(W)$ 是指词串 W 在训练数据中出现的次数

105

语音识别算法

➤ 如何识别

❖ 用viterbi算法做最优路径搜索，找出概率最大的状态序列：

- ✓ HMM转移概率控制子词内状态间的转移；
- ✓ 词典控制词内子词间的状态转移；
- ✓ 语言模型控制词间的状态转移。

❖ 状态空间太大，例如：词表中有十万个词，平均每个词有10状态，则计算复杂度为：

$$O(100万^2 \times T)$$

108

语音识别算法

- 在此状态空间上，计算量非常大，无法保证识别算法的实时性
- 解决方案：Viterbi Beam 搜索算法
- 核心思想是剪枝，每个时刻仅保留少量状态可以向下个时刻扩展路径。例如，若仅保留100个状态，时间复杂度为： $O(100万 \times 100 \times T)$
- 剪枝的依据是当前局部路径的概率
- 贪心算法

109

语音识别算法

Viterbi Beam搜索算法

- 初始化
 - 初始化活动路径（最高层）
- 递推
 - For $m=1$ 到 M
 - For 每一层次（指各个层次的语言和声学模型）
 - For HMM的每个活动状态
 - 把每个活动路径向后扩展一帧至所有可以到达的状态
 - 执行Viterbi计算
 - 裁剪路径
 - End 活动状态
 - End 每一层次
 - End 观察矢量序列
- 终止：选择最可能的路径

110