浙江大学 计算机学院

数字媒体与网络技术

数字语音处理IV

浙江大学计算机学院

群名称：数字视音频处理-数2019媒
群　号：398550295

杨莹春

yyc@zju.edu.cn, QQ:1169244241

QQ群：数字视音频处理-数2019媒（398550295）

验证信息/群名片：姓名学号口音

浙江大学曹光彪西楼304

2019年9月30日

数字语音处理课程安排

- 讲授内容
  - （9月11日）秋1：课程简介＋语音技术引言
  - （9月18日）秋2：语音分析（I）
  - （9月25、30日）秋3：语音分析（II）、语音识别（I）
  - （10月9日）秋5：语音识别（II）语音编码与合成
  - （1月7日）冬9：复习及项目成果展示
- 实验内容
  - 1. PRAAT 语音分析 （9月16日）秋2
  - 2. VOICEBOX说话人识别 （9月30日）秋4

考试：2020年1月16日08:00-10:00

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

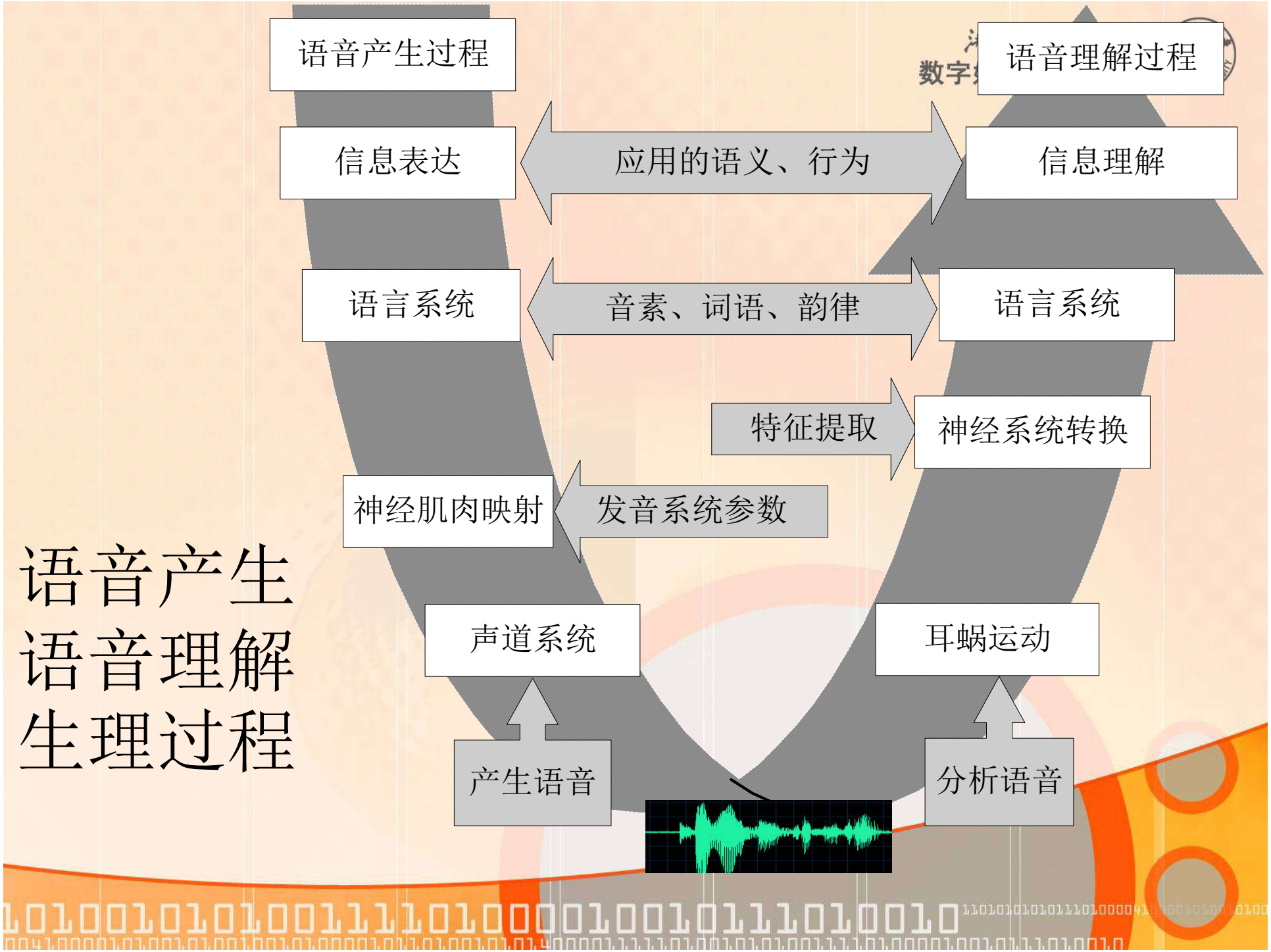
正确率：$$Correct = \frac{N - D - S}{N} \times 100\%$$

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



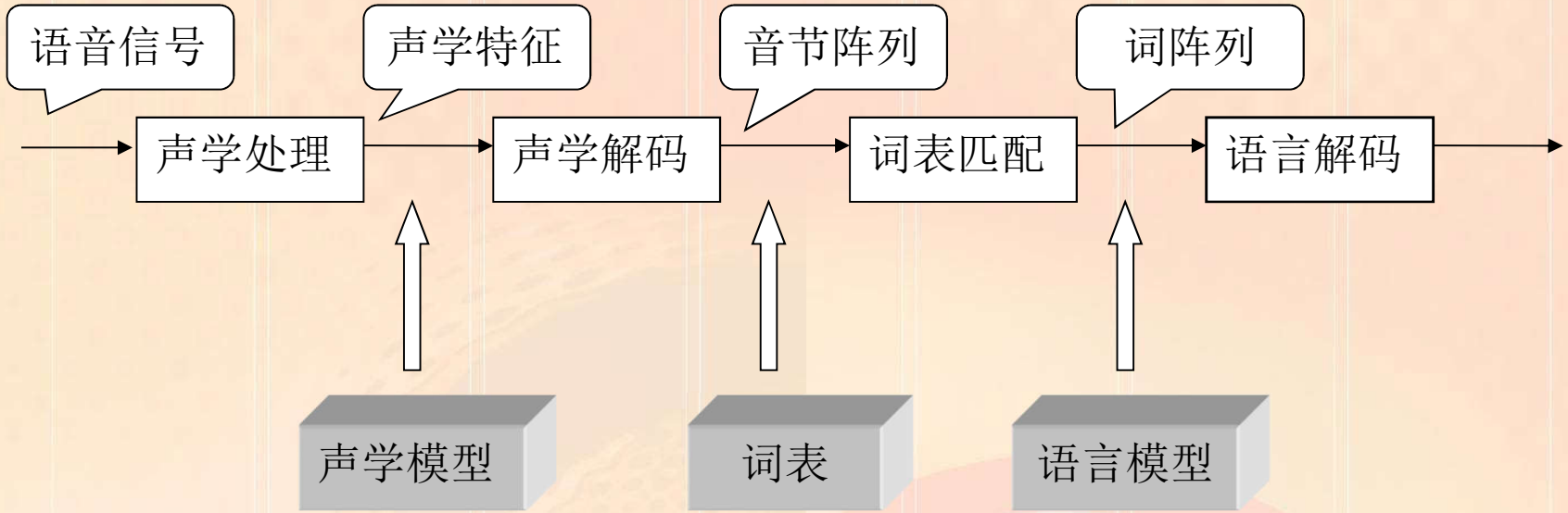
语音识别层次模型


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

Statistical Speech Recognition Architectures



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

Data-driven methods led to major advances in speech recognition.

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

特征参数

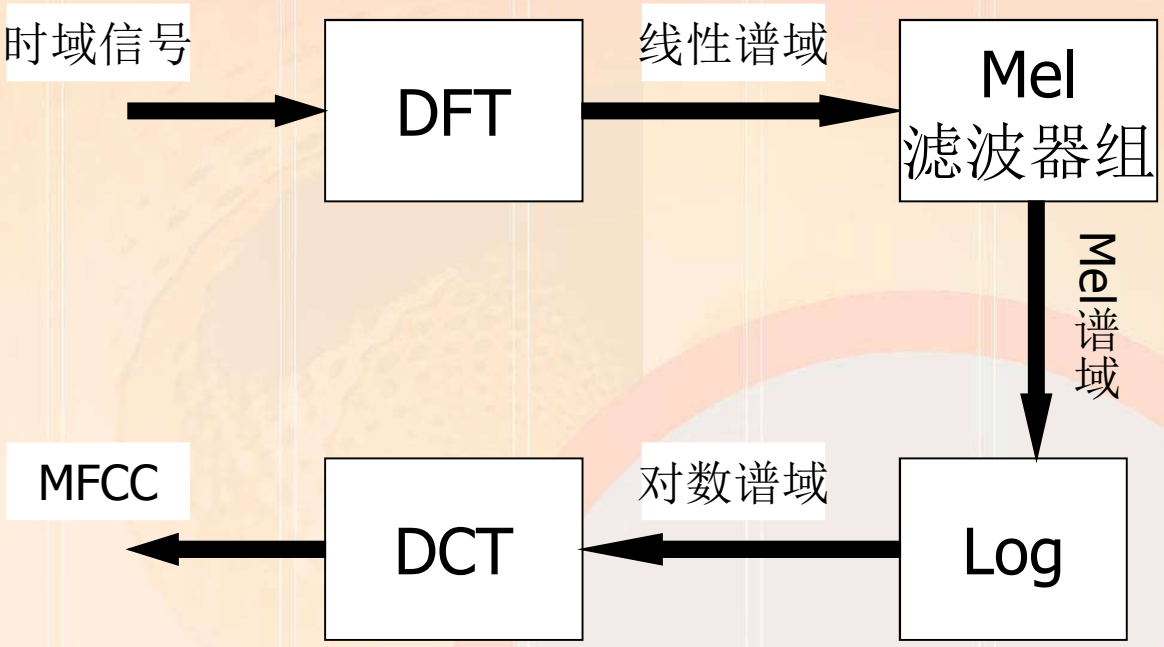
- 静态参数： Mel-Frequency Cepstrum Coefficients (MFCC)
- 帧能量
- 动态参数

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:

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

$x[n]$ -- 时域信号

$X[k]$ -- 频域信号

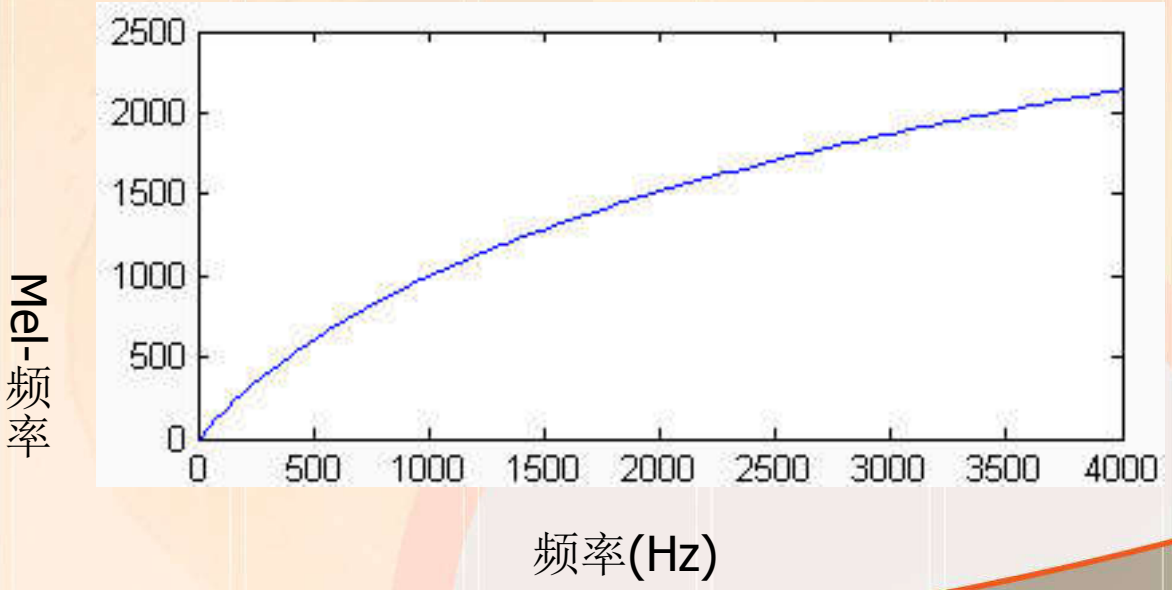
Mel-频率

- 公式：

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

f -- 频率　　B -- Mel-频率

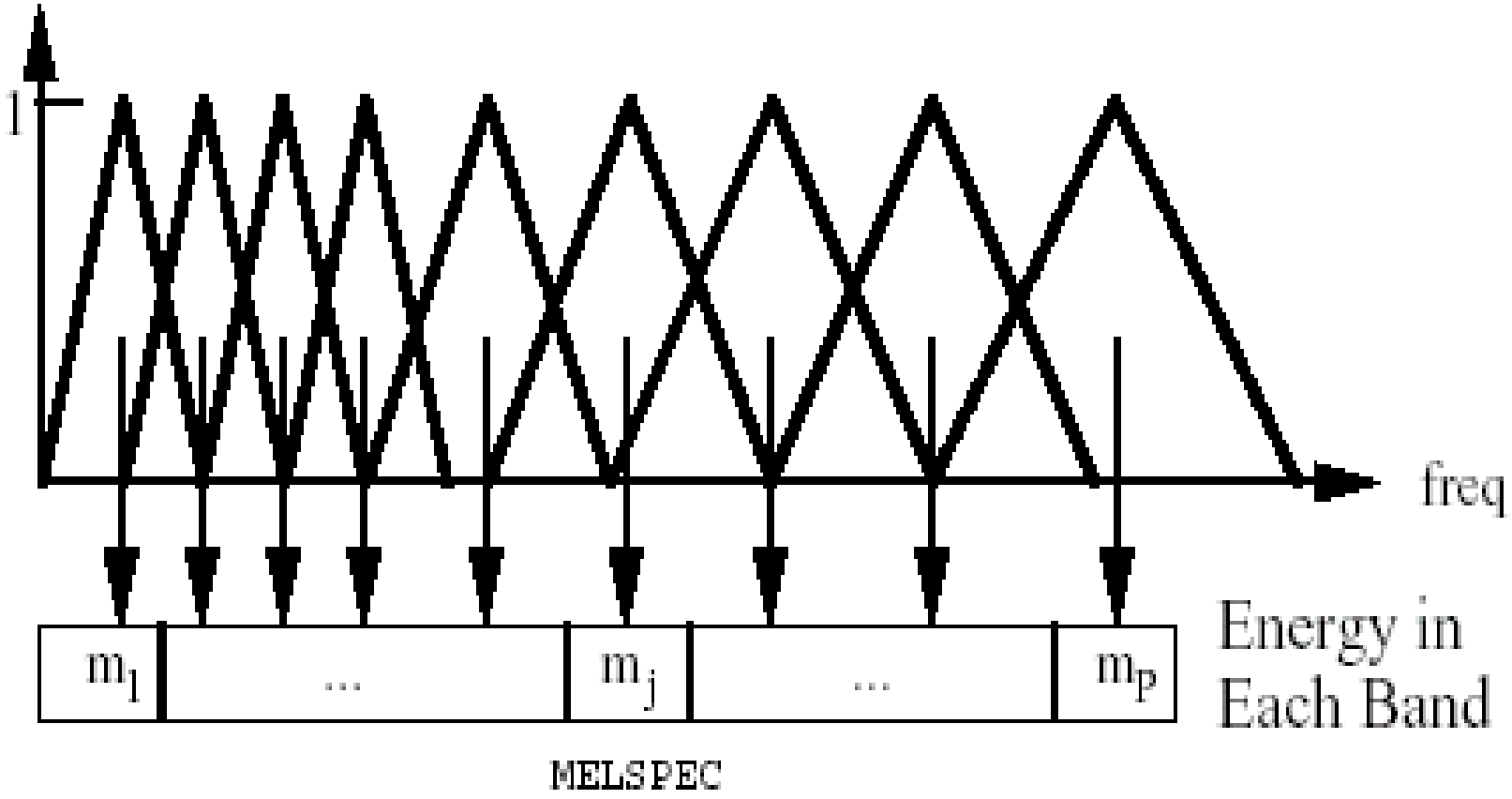
- 频率—Mel-频率：



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：

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

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

- 应用：

$$E=\sum_{n=0}^{N-1}\left|x[n]-\bar{x}\right|$$

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

动态参数

- 反映帧间相关信息
- 一阶差分：

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

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

- 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
- 一个变形：

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

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

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

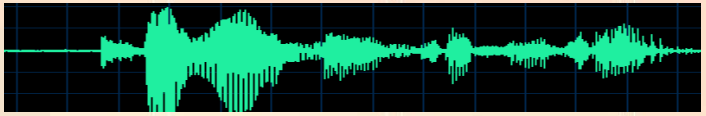
- DTW(Dynamic Time Warping)
- VQ(Vector Quantization)
- HMM (Hidden Markov Models)

语音模型

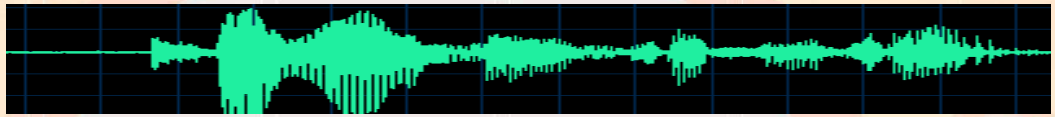
- DTW(Dynamic Time Warping)
- VQ(Vector Quantization)
- HMM (Hidden Markov Models)

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



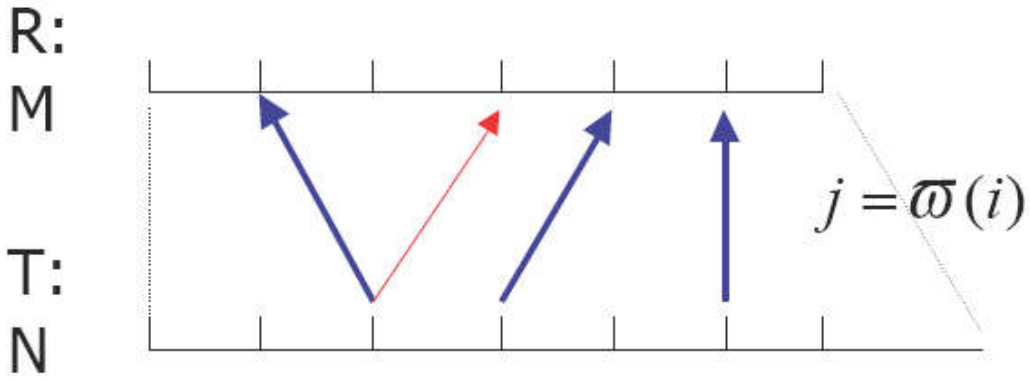
  - 均匀伸长或缩短
  - 依赖于端点检测
    - 通过时域分析进行，利用能量、振幅和过零率等特征
  - 缺点：仅扩展时间轴，无法精确对准



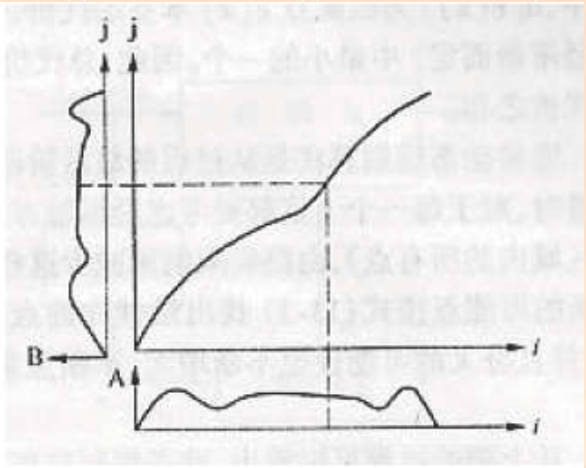
- 方法2：动态时间规整
  - DTW－Dynamic Time Warping

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



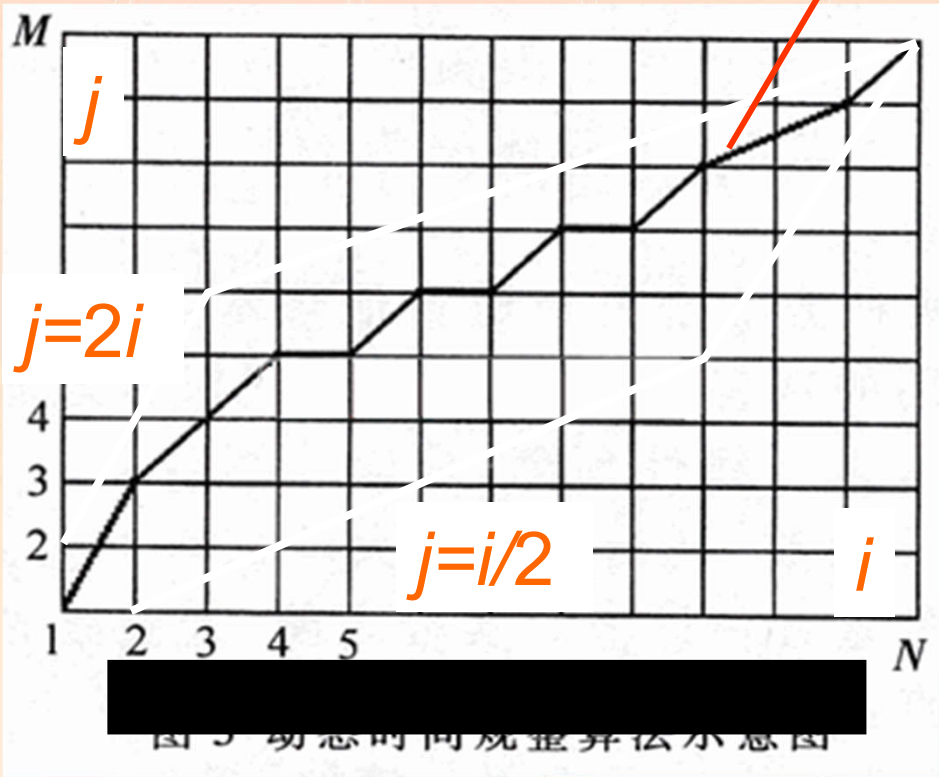
$$D=\min_{\varpi(i)}\sum_{i=1}^{M}d[T(i),R(\varpi(i)]$$



DTW的DP实现

$$D[c(k)]=d[c(k)] + \min D[c(k-1)]$$

- 动态规划
  - 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



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

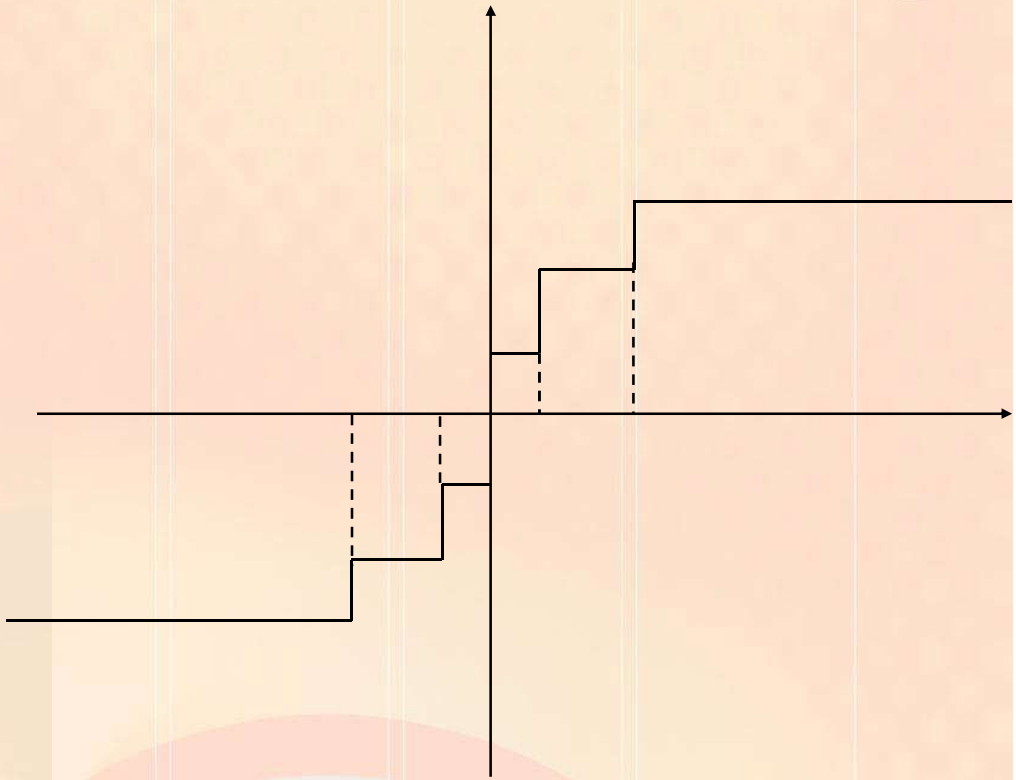
- DTW(Dynamic Time Warping)
- VQ(Vector Quantization)
- HMM (Hidden Markov Models)

VQ在语音分析中的应用

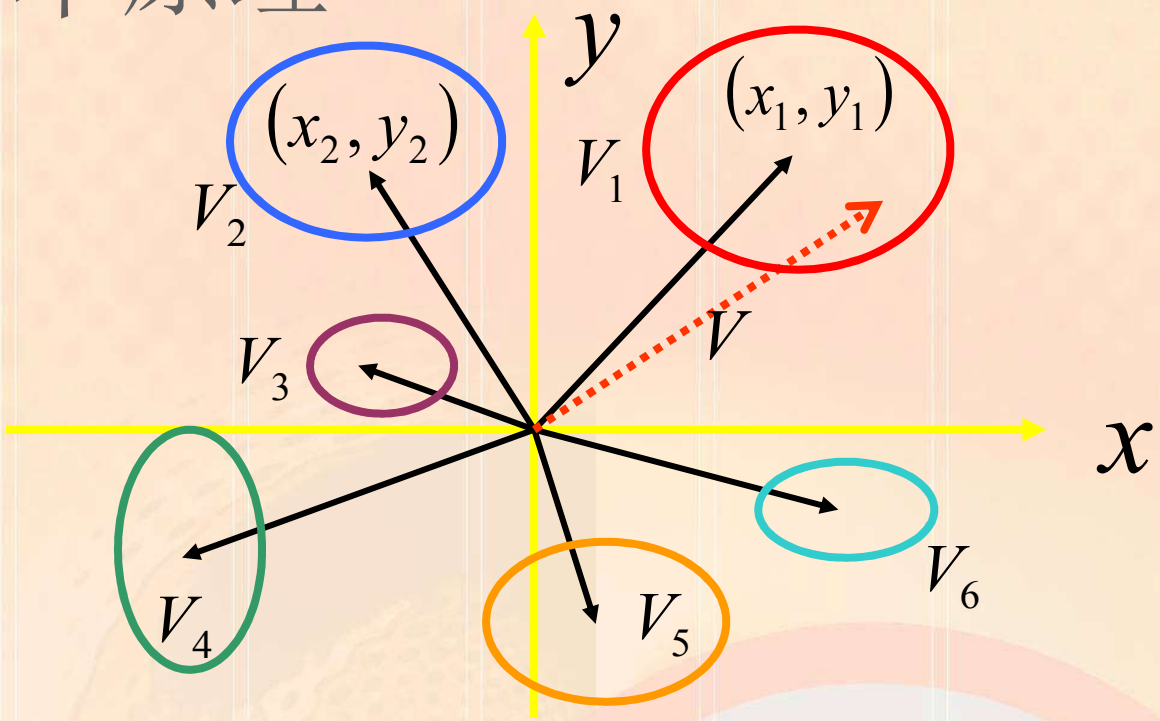
- 进入80年代以后，VQ技术引入语音处理领域，推动了语音技术发展，使之有了长足的进步
- 目前这项技术已经用于：
  - 语音识别；
  - 语音波形编码；
  - 线性预测编码；
  - ……

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

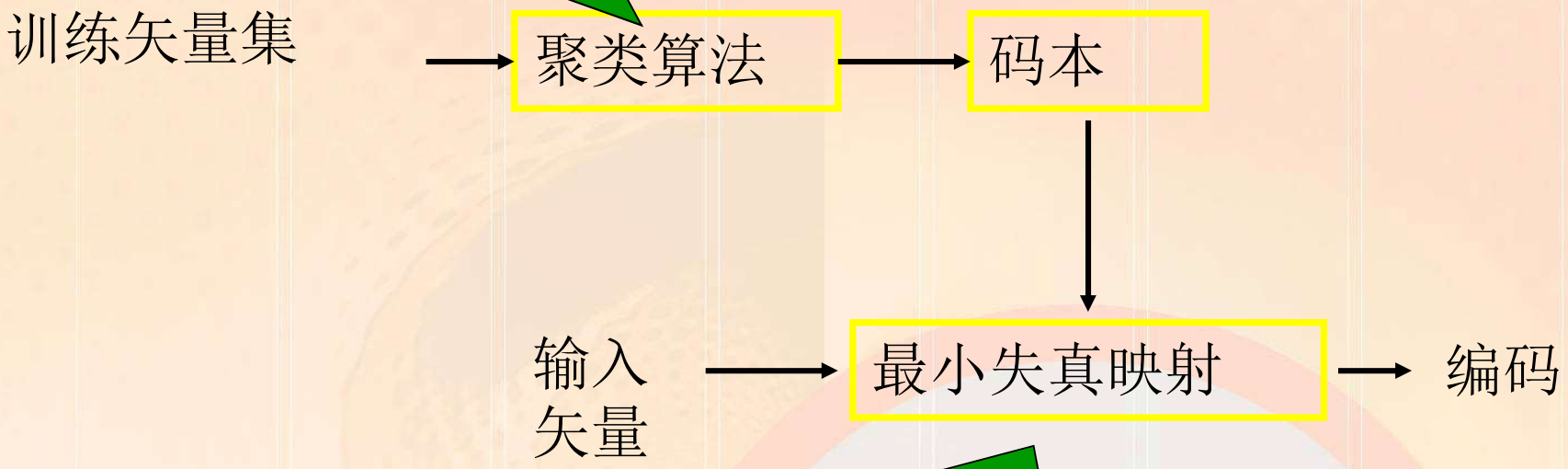
- 任意一个矢量V应该归为哪一类，要看它是“靠近”哪一类矢量，或者说它离哪一个室心最“近”
  - 例如上图中虚线画出的矢量V最靠近V1，则将其规定为V1类，并用V1表示V，或者说V被量化为V1

- 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



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$y=\mathrm{VQ}(x)$。

- 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$\mathrm{CB}=\{CW_i: 1\le i\le NC\}$，$NC$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：

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

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.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
  - 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

Brief History

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.

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）