视听觉信号处理

Visual and Auditory Signal Processing

哈爾濱工業大學
Harbin Institute of Technology

语 音 学 概 要

哈爾濱工業大學
Harbin Institute of Technology

- 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

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？

视听觉信息理解系列课程

课程一 视听觉信号处理	视觉/听觉信号的分析和压缩表示的理论与方法
课程二 模式识别与深度学习	模式识别与深度学习理论与方法
课程三 视听觉信息理解	面向特定视觉/听觉信息理解任务的解决方案

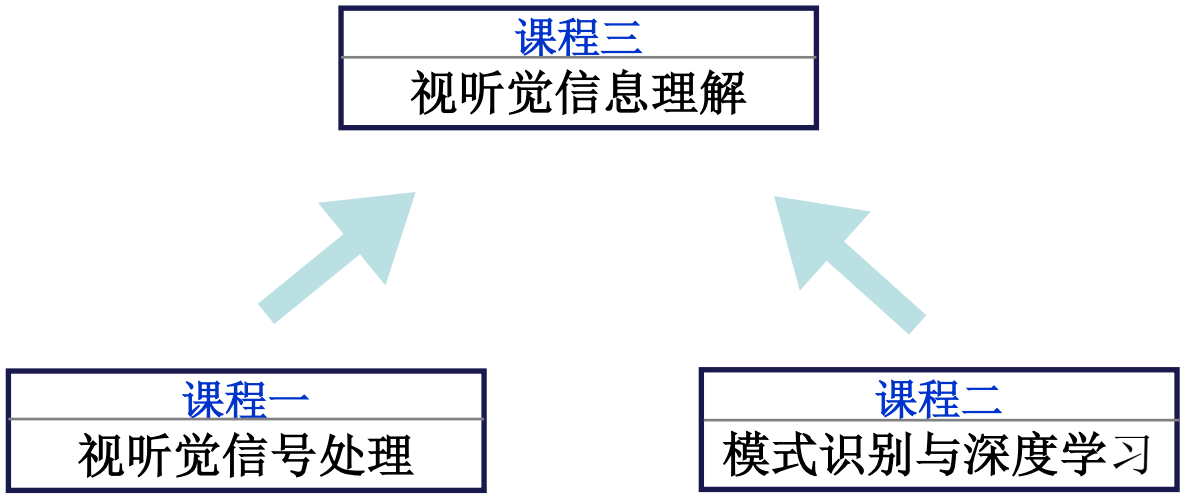
视听觉信息理解系列课程

课程一
视听觉信号处理

课程二
模式识别与深度学习

课程三
视听觉信息理解

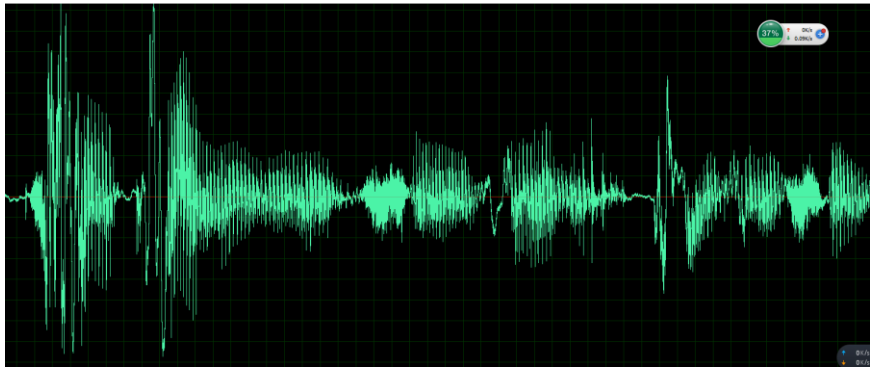
视听觉信息理解系列课程

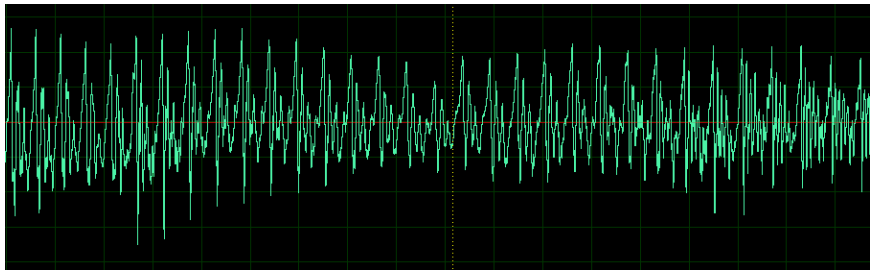


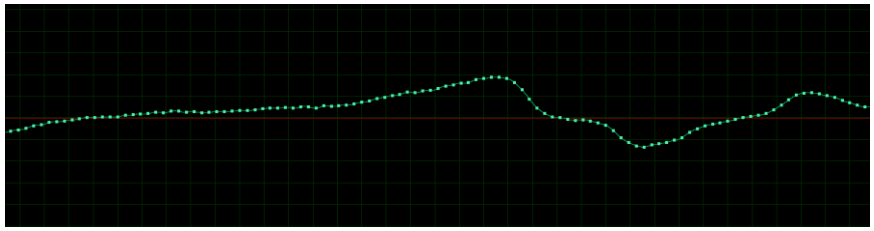
课程设计的特点

- 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，比对交融，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。

- 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

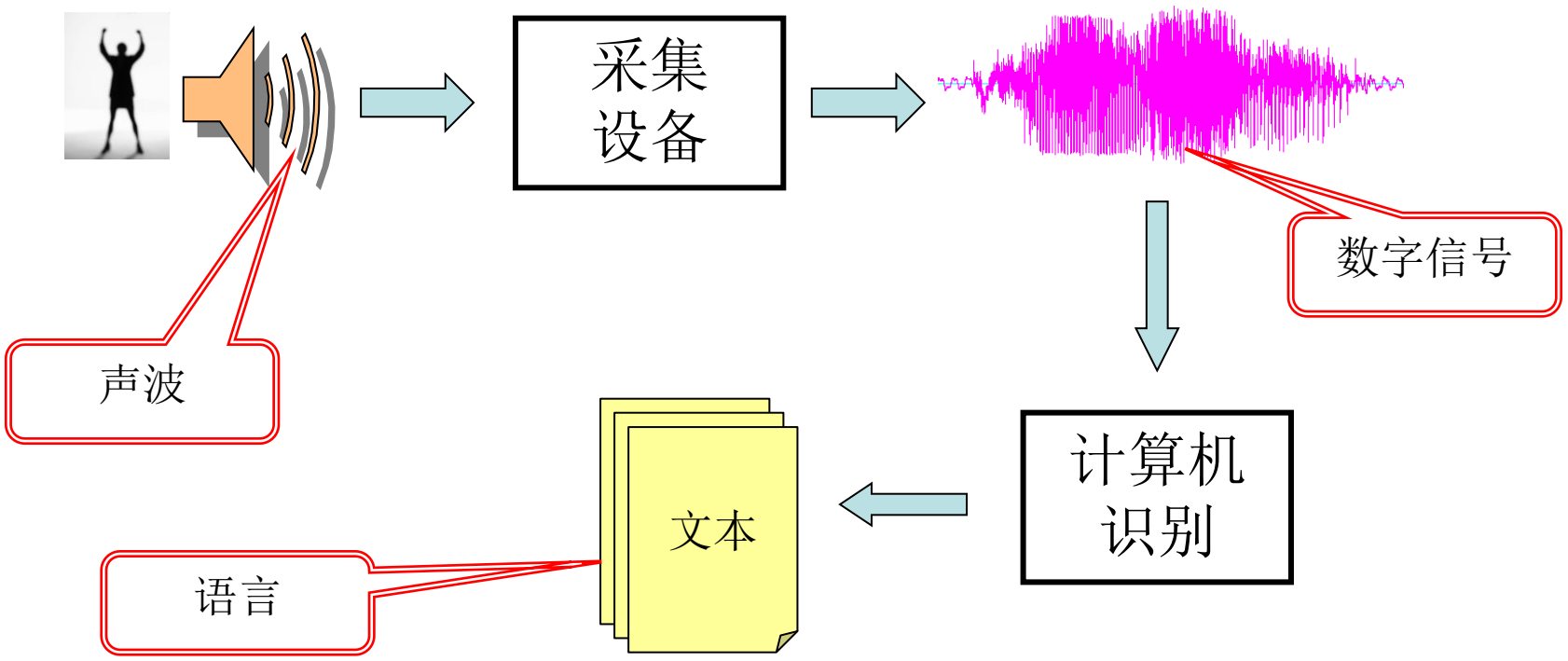




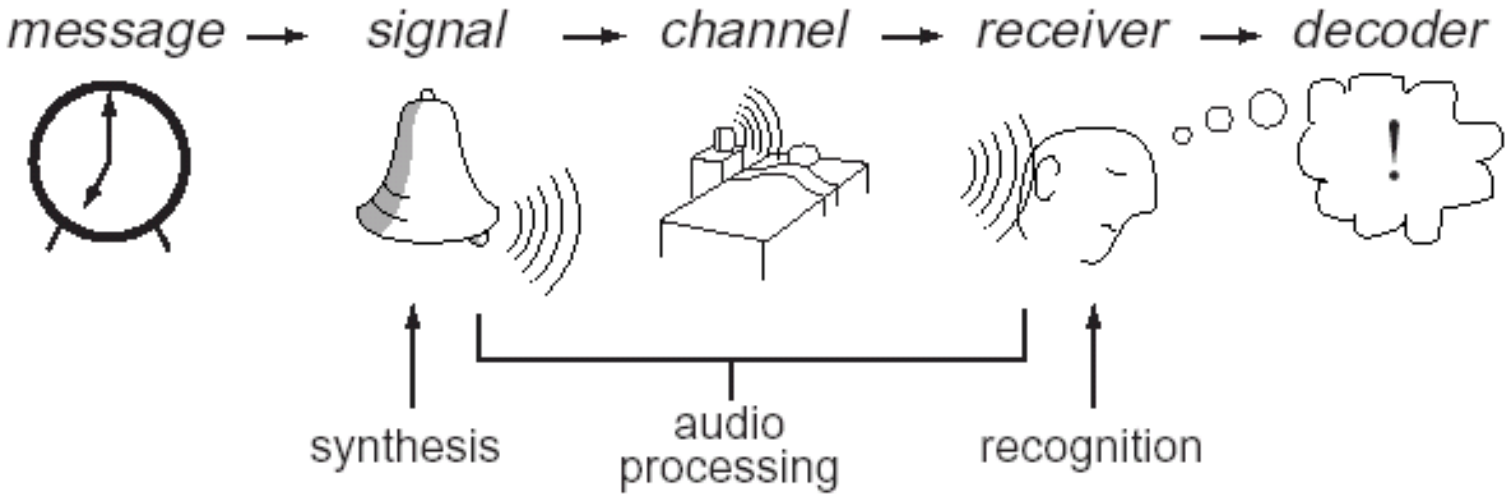


声音的多重属性：

- 声波属性
- 信号属性
- 符号属性

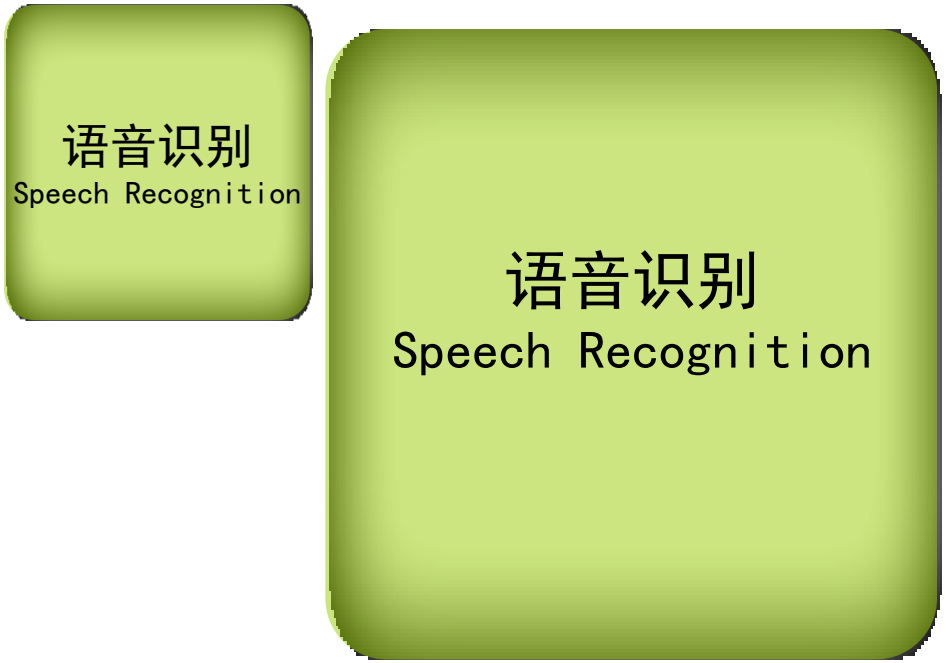


仿真人的言语产生和理解过程



听觉信号处理？

听觉信号处理领域的研究方向



语音识别

语音识别技术就是通过识别和理解过程，将语音装换成相应的书面信息，也就是让计算机听懂人说话

典型应用

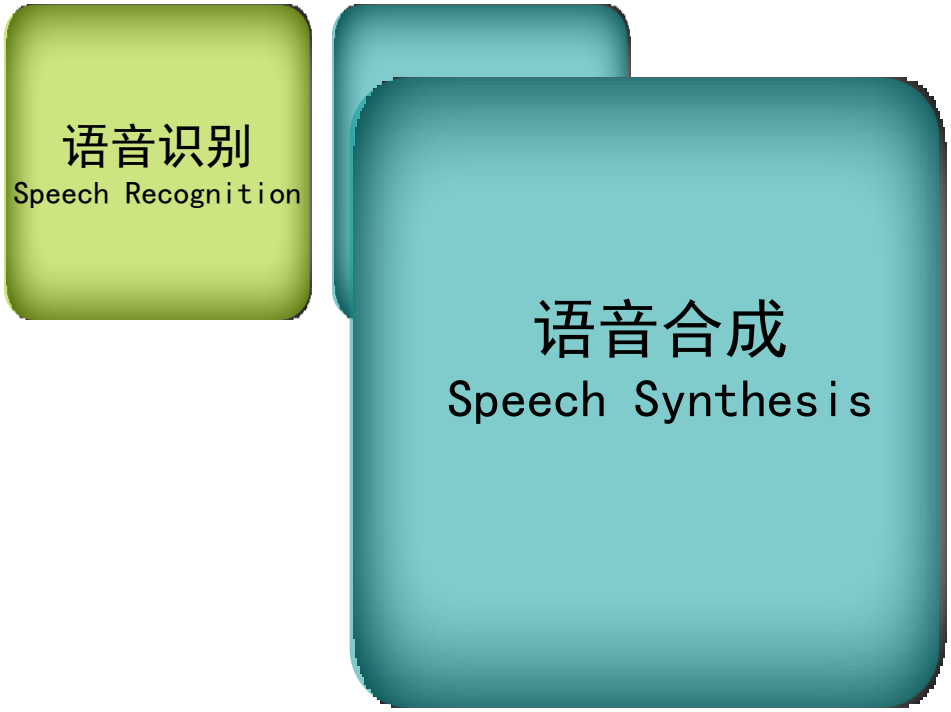
声音拨号系统；声控系统；听写机；

自动口语翻译；会话系统；

语音信息监测系统等。

听觉信号处理？

听觉信号处理领域的研究方向



语音合成

将书面信息装换成等价的语音，也就是让计算机说话。

典型应用

- 自动报站
- 信息查询
- 语言学习软件
- TTS（Text to Speech）技术等。

听觉信号处理？

听觉信号处理领域的研究方向

语音编码

用尽可能低的比特率来存储和传输语音数据

典型应用

- 数字通信系统--各种编码格式
- 保密通信
- 语音信箱
- VOIP（Voice over internet protocol）
- 多媒体流媒体

听觉信号处理？

听觉信号处理领域的研究方向

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。

听觉信号处理？

听觉信号处理领域的研究方向

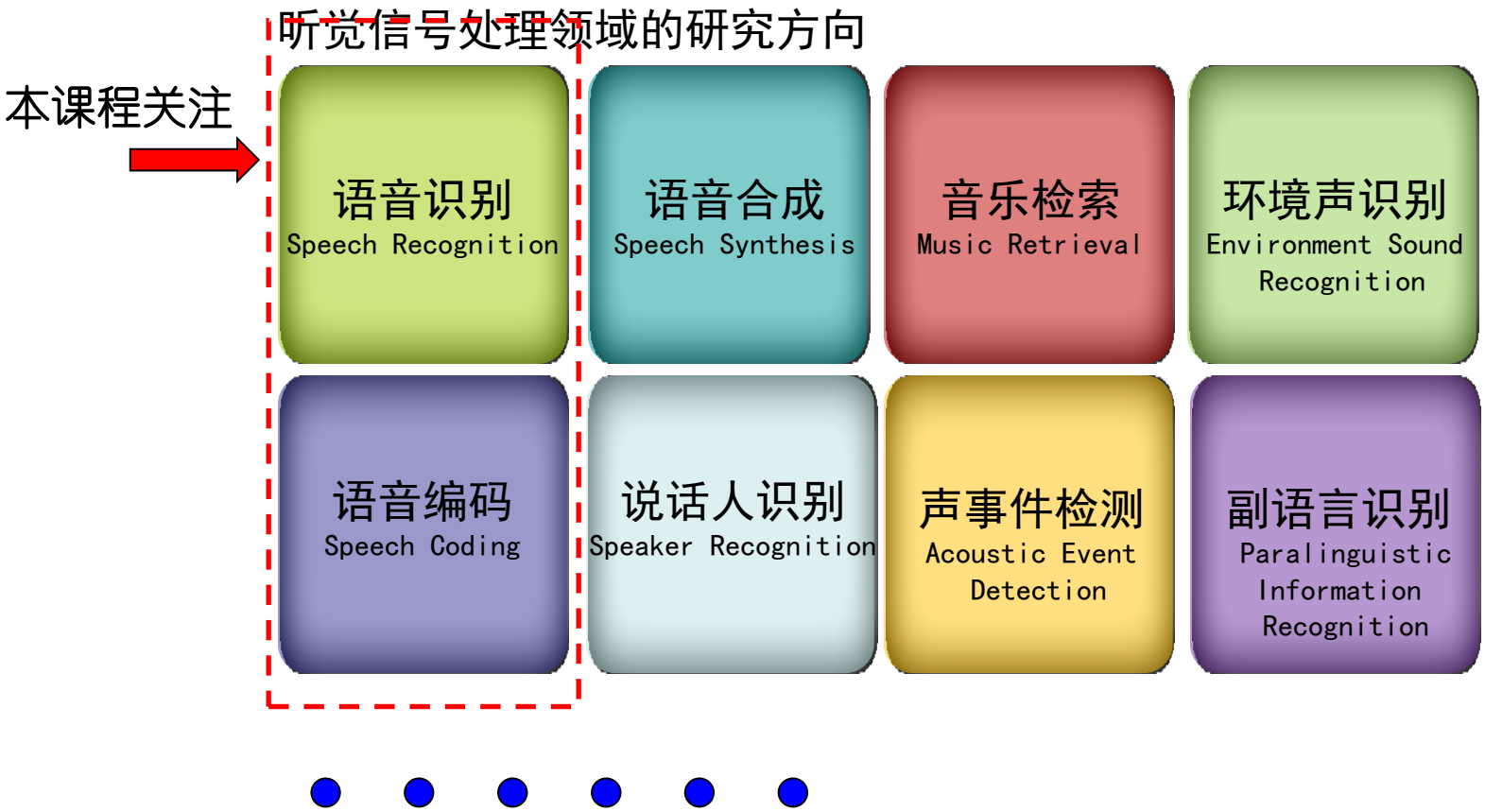
听觉信号处理？

听觉信号处理领域的研究方向

听觉信号处理？

听觉信号处理领域的研究方向

听觉信号处理？


听觉信号处理领域的研究方向
本课程关注
语音识别
Speech Recognition
语音合成
Speech Synthesis
音乐检索
Music Retrieval
环境声识别
Environment Sound Recognition
语音编码
Speech Coding
说话人识别
Speaker Recognition
声事件检测
Acoustic Event Detection
副语言识别
Paralinguistic Information Recognition

举 例：

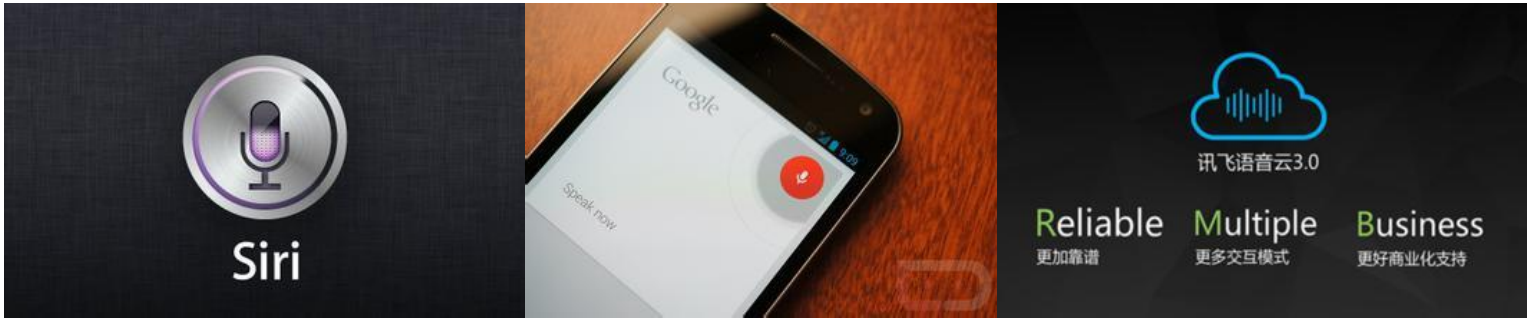
语音识别技术的研究内容和发展轨迹

语音识别技术

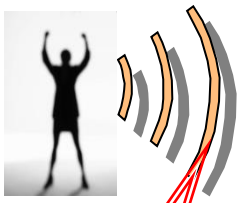
- 语音是语言的载体，是思维的依托，是人类有别于其它生物的重要标志，是智能的终极体现
- 语音识别技术就是通过识别和理解过程，将语音装换成相应的书面信息，也就是让计算机听懂人说话
- 语音交互技术（语音识别+语音合成）将引领人类进入下一个交互时代

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



语音识别技术的框架



声波

采集设备

数字信号

计算机识别

文本

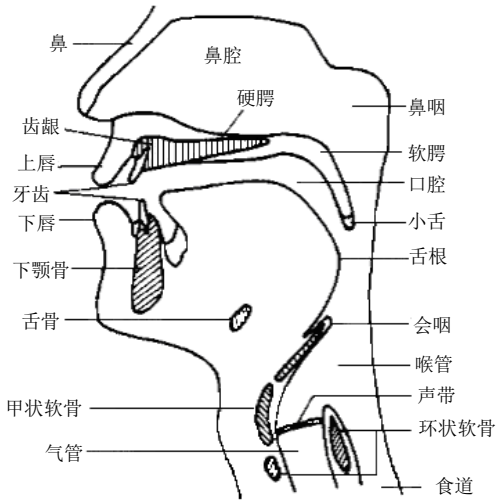
语言

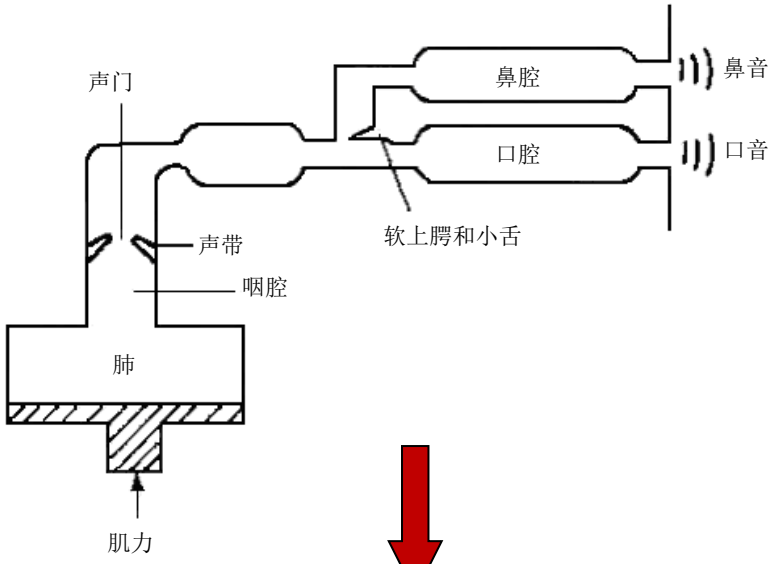
模式识别

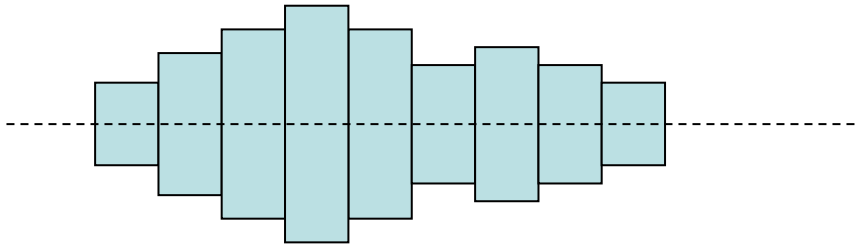
特征提取

仿真人的处理过程

特征提取环节的仿真-发生机理



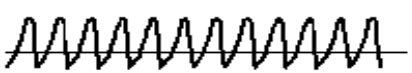




$$H(z)=\frac{G}{1-\sum_{k=1}^{N}\alpha_k z^{-k}}$$

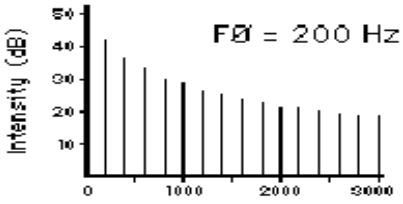
特征提取环节的仿真-发生机理

声门脉冲

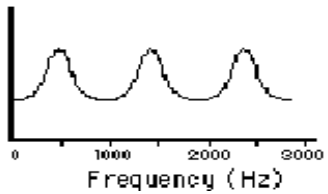


声道

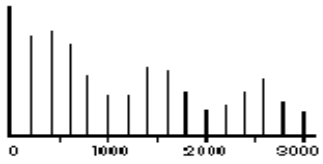
语音信号



激励信号
频谱



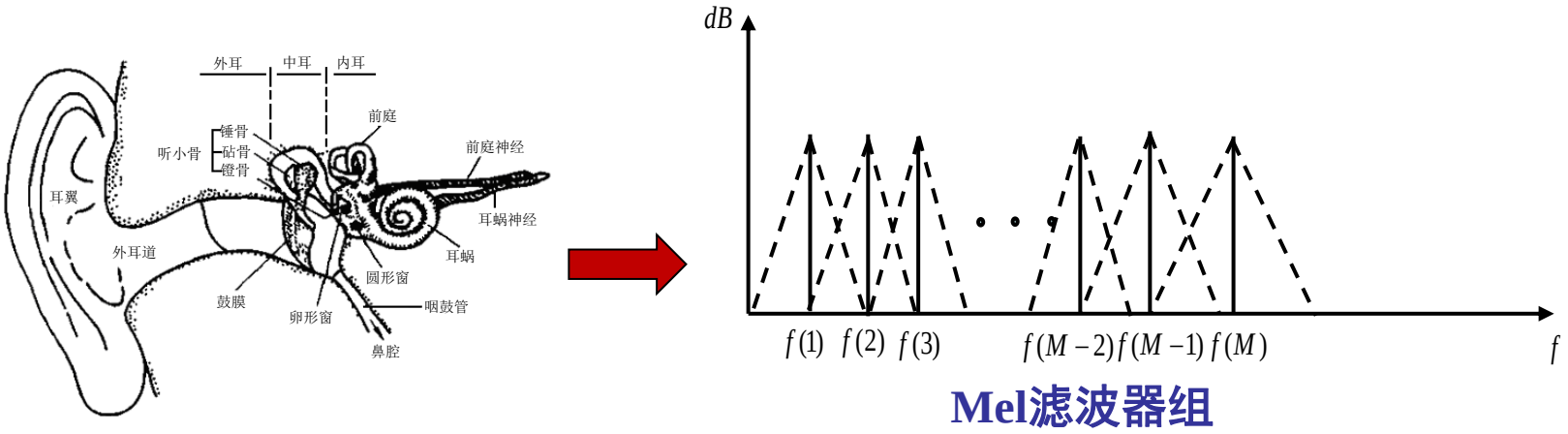
声道滤波器
传递函数



语音信号
频谱

- 据此，提出了线性预测倒谱系数（LPCC）特征

特征提取环节的仿真-感知机理



- 据此，提出了Mel频域倒谱系数（MFCC）特征

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

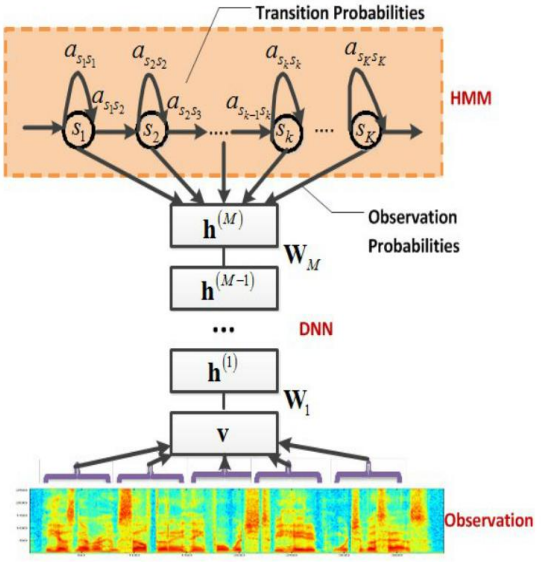
✓ 90年代初，李开复在CMU搭建了基于HMM的非特定人连续语音识别系统SPHINX；

✓ 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。

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

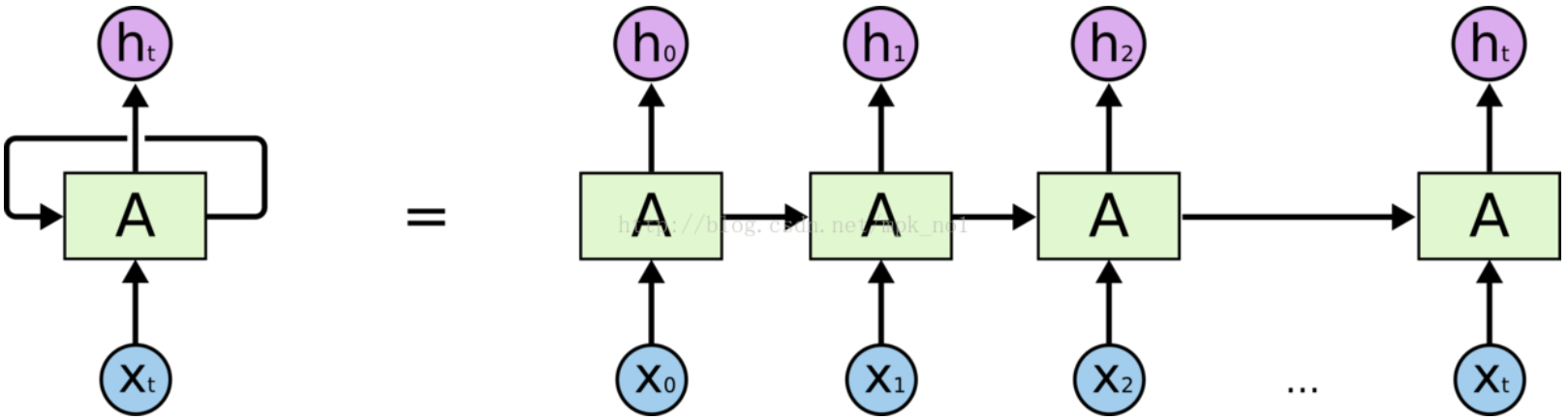


模式识别环节的仿真

- 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(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。

模式识别环节的仿真

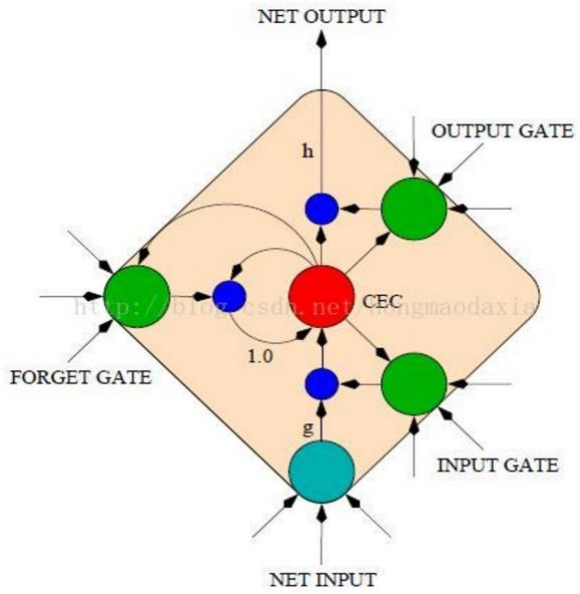
◆ **基于RNN+DNN的语音识别技术**

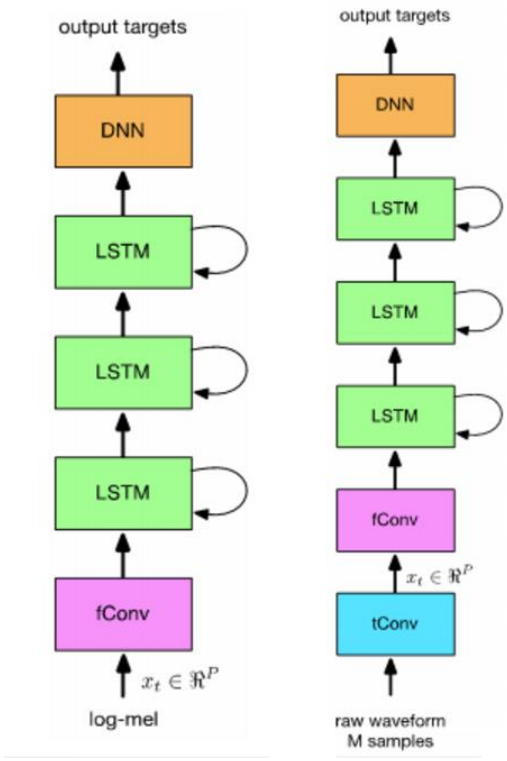


RNN结构示意图

模式识别环节的仿真

◆ **基于RNN+DNN的语音识别技术**



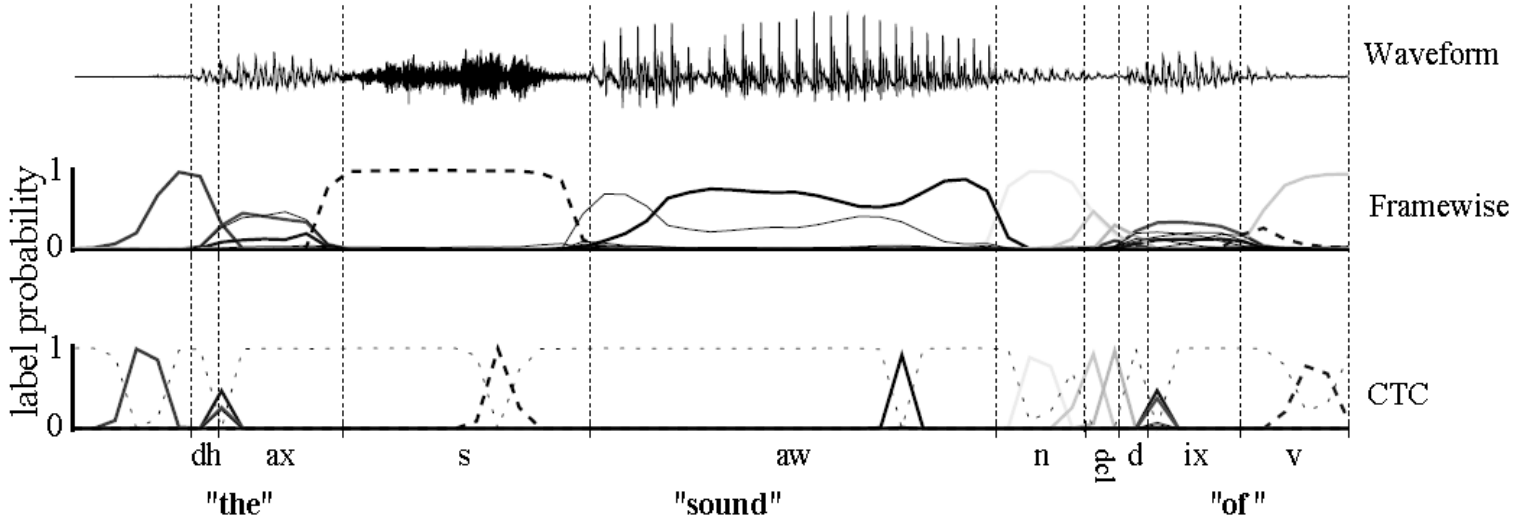


LSTM

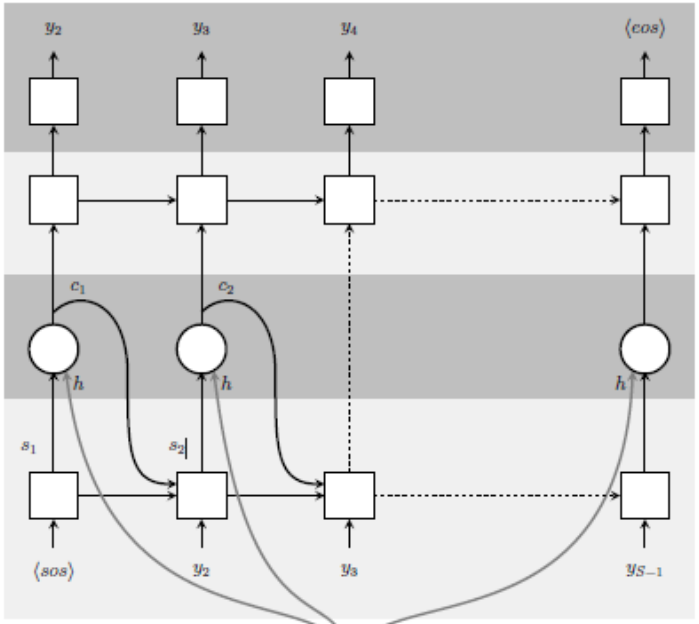
CLDNN

模式识别环节的仿真

◆ **End to end语音识别技术**



Speller



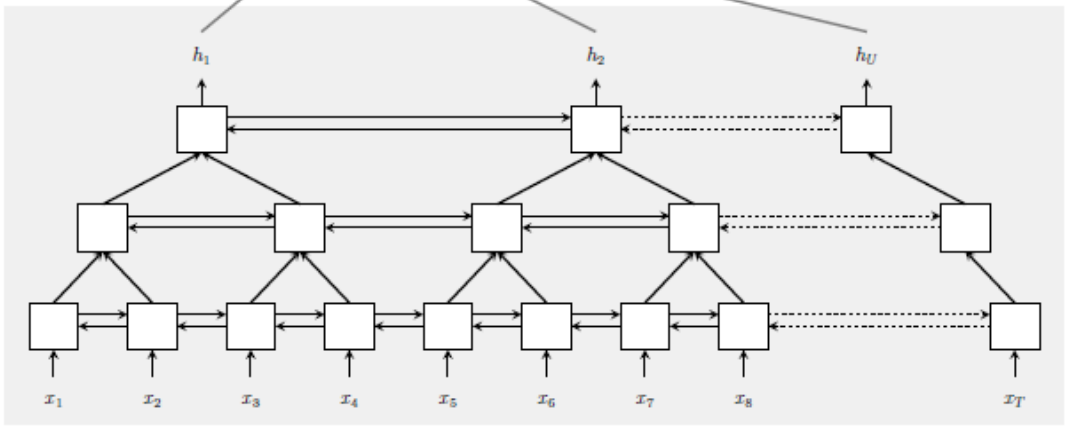
Grapheme characters y_i are modelled by the CharacterDistribution

AttentionContext creates context vector c_i from $\mathbf{h}$ and s_i

$h = (h_1, \ldots, h_U)$

Long input sequence $\mathbf{x}$ is encoded with the pyramidal BLSTM Listen into shorter sequence $\mathbf{h}$

Listener



的仿真

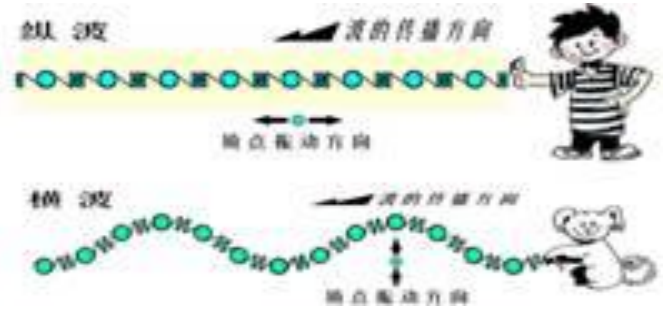
参考书

- 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, etal. 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**

语音的声学表示

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觉与实际频率近似成对数关系。
  - 基频：60Hz~500Hz

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2×10^{-5}Pa；能承受的最大声压称为痛阈，约为200Pa。

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)

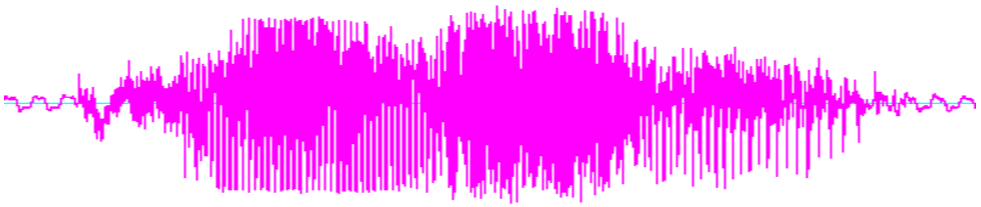
- 闻阈相当于0dB；痛阈相当于140dB；一个讲话时，离他一米远处的声强大约为60~80dB；

➢ **声波的基本物理量---共振和共振峰**

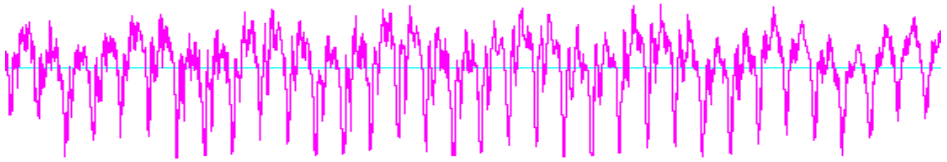
- 共振：当一个物体受迫震动时，所加驱动频率等于物体固有频率时，便以最大的振幅来震荡。
- 语音中也有共振现象，元音的音色和区别特征主要取决于声道的共振峰特性。

语音的数字信号表示

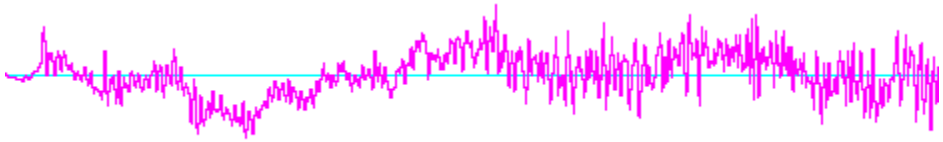
语音的数字信号表示



(a) 语音信号“开始”时域波形



(b) 元音部分/ai/展开图



(c) 辅音部分/k/的展开图

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 $x(n)(n=0,\pm1,\pm2,...)$ 来表示

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：$z(n)=x(n)+y(n) \quad (n=0,\pm1,\pm2,\ldots)$

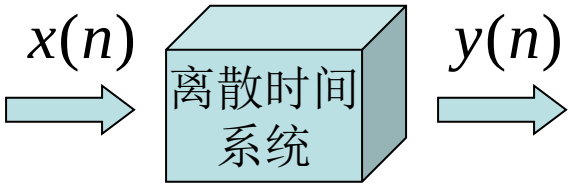
乘：$z(n)=x(n)\cdot y(n) \quad (n=0,\pm1,\pm2,\ldots)$

时延：$z(n)=x(n-m) \quad (n=0,\pm1,\pm2,\ldots)$

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

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：

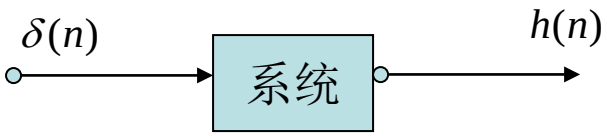
线性：

时不变：

数字信号处理的相关知识

- 卷积和滤波

单位冲激函数 $\delta(n)=\begin{cases}1 & (n=0)\\ 0 & (n\neq 0)\end{cases}$ 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：

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

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 $[a(n)+b(n)]*c(n)=a(n)*b(n)+a(n)*c(n)$

（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。

数字信号处理的相关知识

- 离散傅立叶变换（Discrete Fourier Transform）

连续非周期信号：傅立叶变换

连续周期信号：　傅立叶级数

离散非周期信号：离散时间傅立叶变换

离散周期信号：　离散傅立叶变换

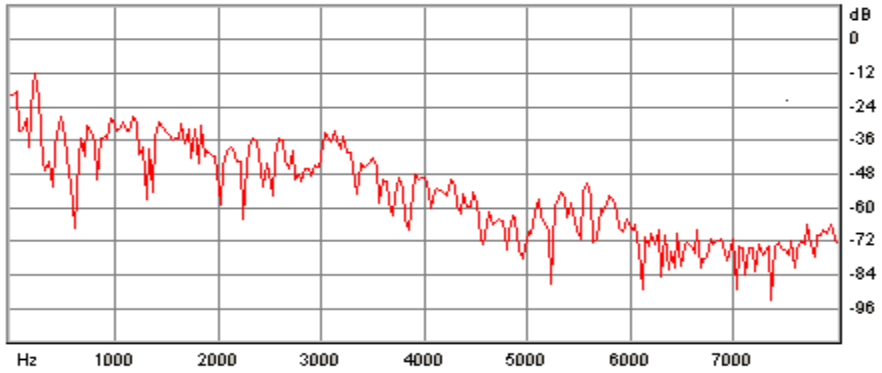
DFT:

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

■变换：

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

数字信号处理的相关知识



“开始”中/ai/的频谱特性

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平面的单位园上，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变换在单位园上的均匀抽样。

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 位移性： $Z[x(n-m)] = z^{-m}X(z)$

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

数字信号处理的相关知识

- 离散余弦变换（Discrete Cosine Transform）

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

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

  - 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。

语音的语言表示

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。

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。

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

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。

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。