



浙江大学 计算机学院
数字媒体与网络技术



群名称：数字视音频处理-数2019媒
群 号：398550295

数字语音处理II

浙江大学计算机学院

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

浙江大学外经贸楼520

2019年9月18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
- (12月30日) 冬8: 复习及项目成果展示 (周一实验课)

- 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



语音技术基础

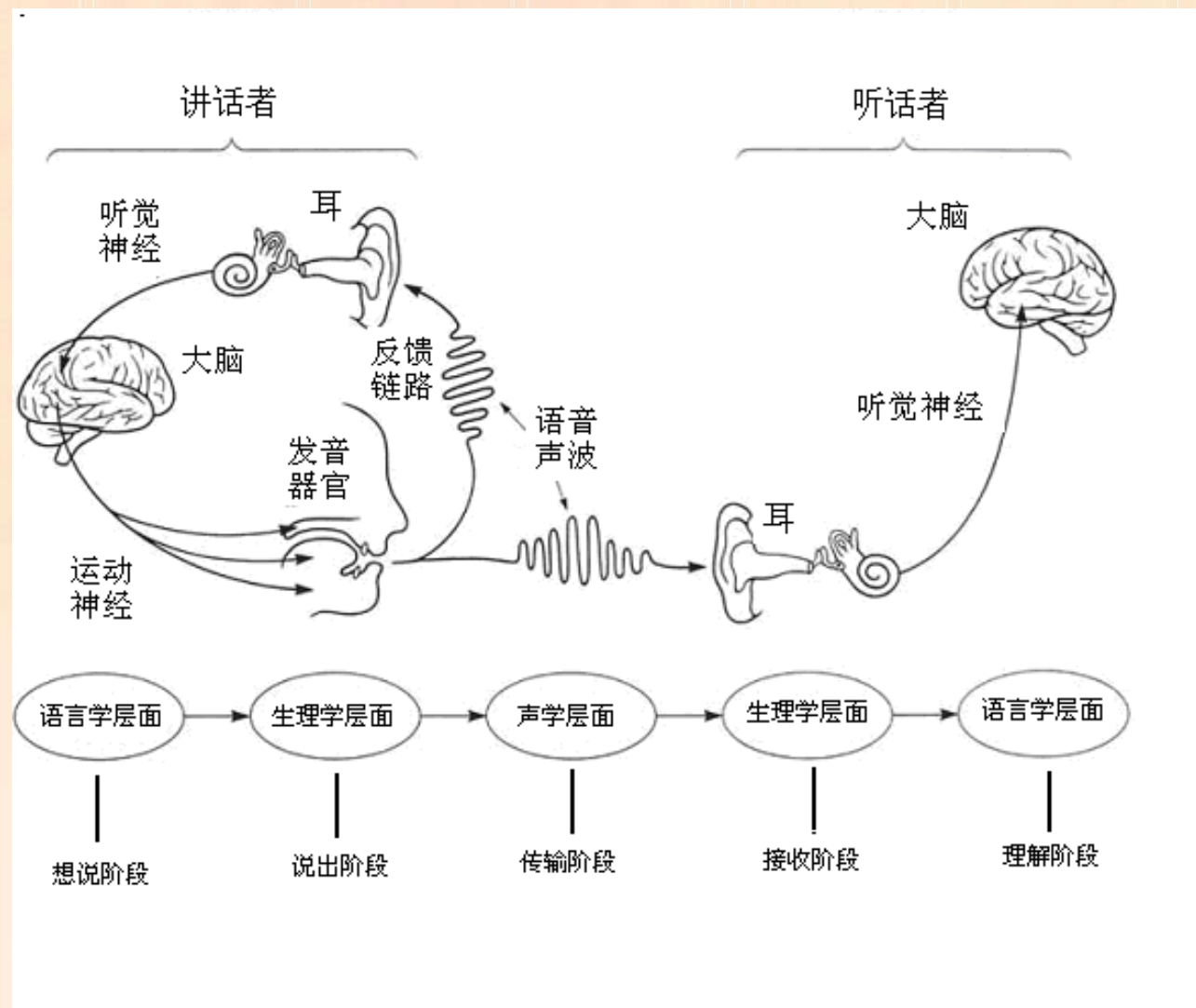
浙江大学计算机学院
数字媒体与网络技术



- 语言交际过程
- 语音产生过程



语音链





- “发音-传递-感知”三阶段
- 语音学 (Phonetics)
 - 发音语音学(Articulatory Phonetics)
 - 声学语音学(Acoustic Phonetics)
 - 感知语音学(Auditory Phonetics)



语音技术基础

浙江大学计算机学院
数字媒体与网络技术

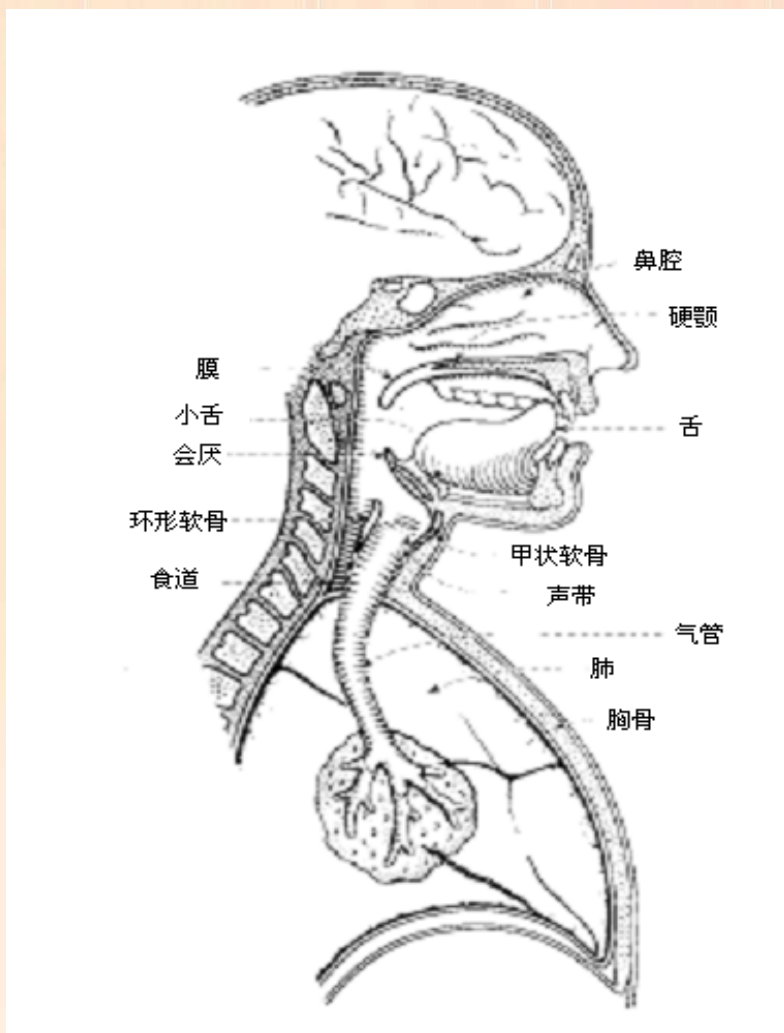


- 语言交际过程
- 语音产生过程



语音产生原理

浙江大学计算机学院
数字媒体与网络技术



- 1、动力源
- 2、发音体
- 3、共鸣器





人声的发音与接收流程，可以列出如下：

- 1、声门的快速打开与关闭
- 2、声道、口腔、鼻腔的共振
- 3、空气的波动
- 4、接收者耳膜的振动
- 5、内耳神经的接收
- 6、大脑的辨识





1、动力源

发音的原动力是呼吸时所产生的呼出的气流，因此可以说人类的呼吸器官就是发音的动力源。

呼吸器官主要是由肺、气管和支气管组成。胸廓、横膈膜是发起呼吸动作的外部辅助部分。

人们在说话或唱歌时的呼吸方式与平静呼吸时的方式有所不同。



语音产生原理



不同种类呼吸的参量变化情况（石锋）

呼吸种类	安静呼吸	言语呼吸	歌唱呼吸
标志			
呼吸目的	吸氧排二氧化碳进行气体交换	谈话	歌唱
呼吸控制	非意识地随便完成	受人意识控制	受人意识控制
呼吸比值	1: 1.2	1: 5~1: 8	1: 8~1: 12
呼吸次数	每分钟16~20次	每分钟8~10次	视歌曲而定
呼吸量	500~500ml	1000~1500ml	1500~2400ml
呼吸路径	主要经鼻	主要经口	主要经口
肌肉动作	吸气时胸部吸气肌群用力，膈略微下移 呼气时胸部吸气肌群放松，腹部肌微用力帮助膈复位	吸气肌群之中提高肋骨、胸骨、固定锁骨以及板直胸椎的各肌收缩，膈下降较明显。 呼气时，胸部吸气肌群放松，呼气肌群收缩，腹部各肌一齐用力，膈上升	吸气时，比言语呼吸又多几条肌肉参加作用 呼气时，收缩中的吸气肌群继续收缩用力，胸腹呼气肌联合作有控制性的收缩



2、发音体

喉是人类专职的发音器官。

(1) 喉软骨：甲状软骨、环状软骨、杓状软骨

(2) 喉关节：环甲关节两个，环杓关节两个。

(3) 喉肌：喉肌分内肌与外肌。

外肌控制喉位的上升、下降和固定。

内肌包括环甲肌、杓间肌、环杓侧肌、甲杓肌。

(4) 声带：甲杓肌分内肌和外肌，内肌就是声带肌。

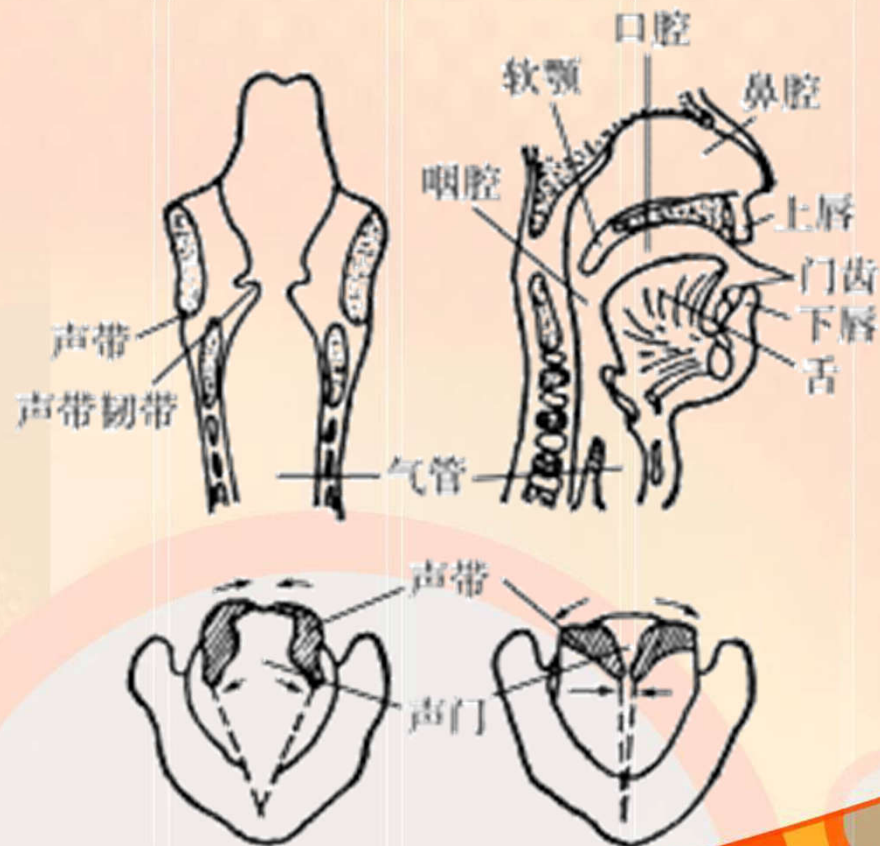


语音产生原理

浙江大学计算机学院
数字媒体与网络技术



- 声带构造
- 声门：音声门、气声门
- 声门状态——发声类型
- 声带振动原理：
伯努利效应
- 声带运动摄像
正常 异常



石锋，语音学原理，南开大学，2007





3、共鸣器

口腔：舌的重要作用

鼻腔：

咽腔：

石锋，语音学原理，南开大学，2007





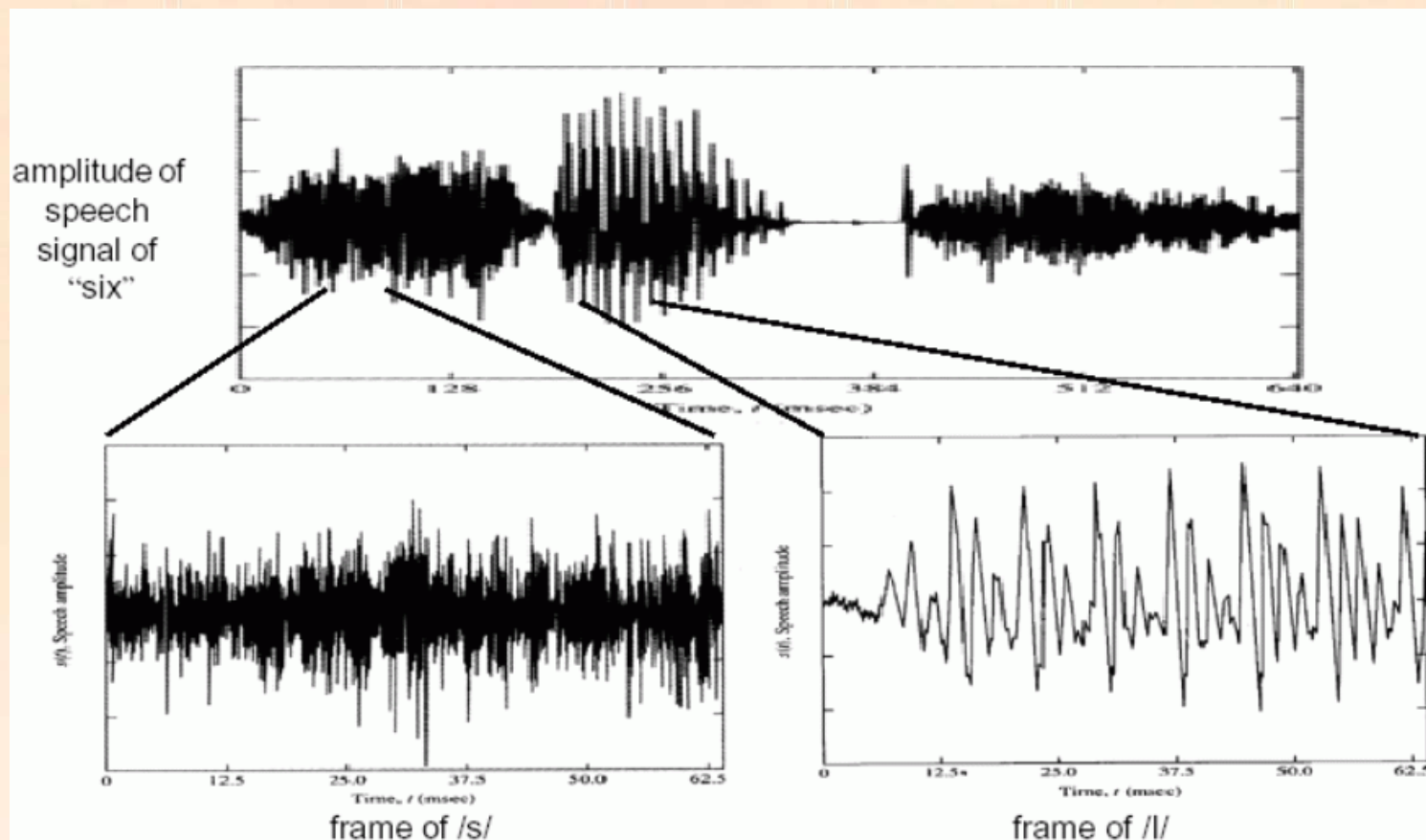
声门振动的快慢，决定声音的基本频率（即音高）。

口腔、鼻腔、舌头的位置、嘴型等，决定声音的内容（即音色）。

肺部压缩空气的力量大小，决定音量大小。



语音产生原理



语音产生原理



浊音 (Voiced): 如 /U/、 /d/、 /i/ 等音

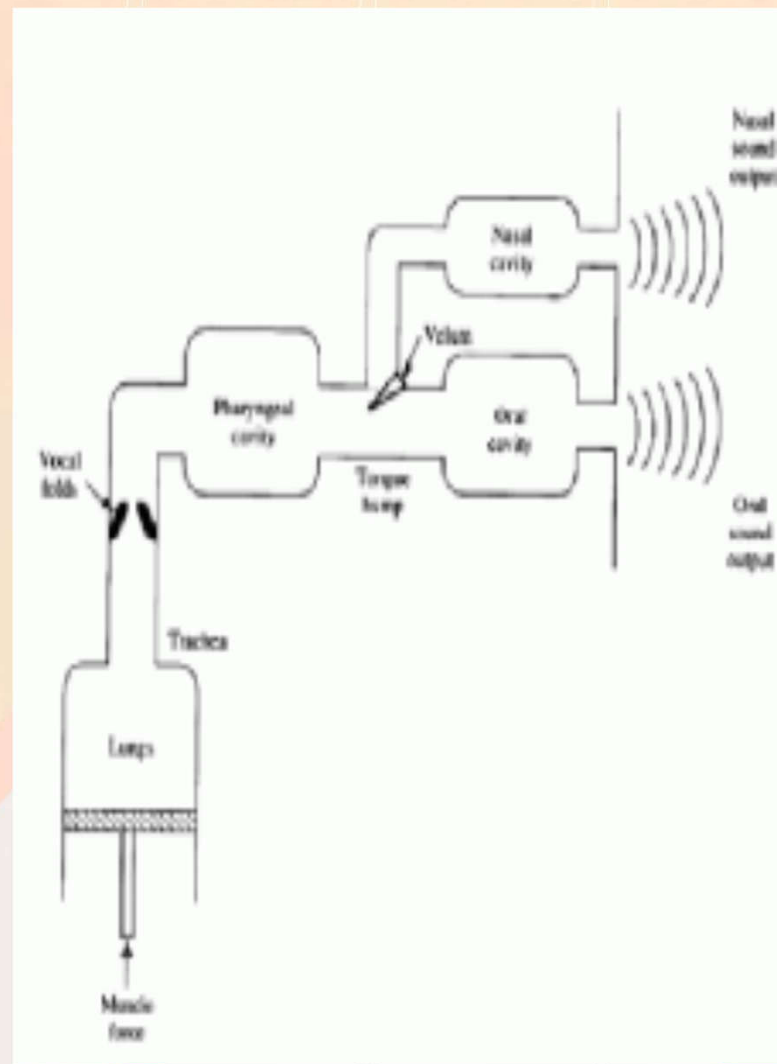
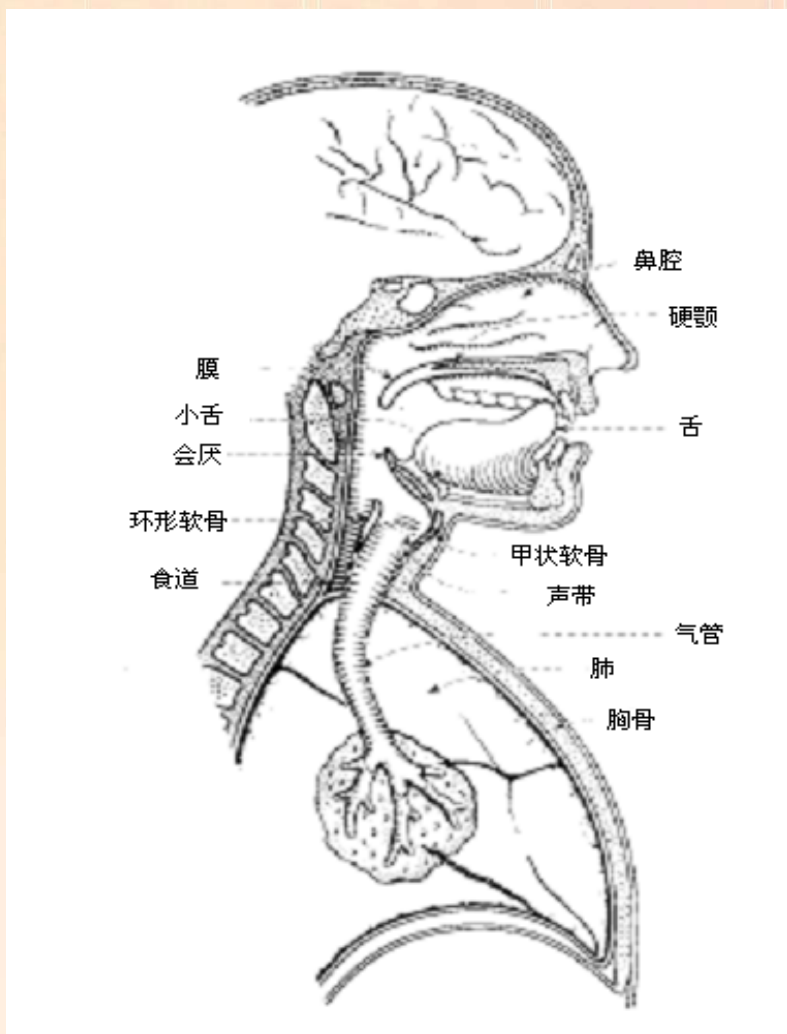


清音 (Unvoiced): 如 /s/ 音

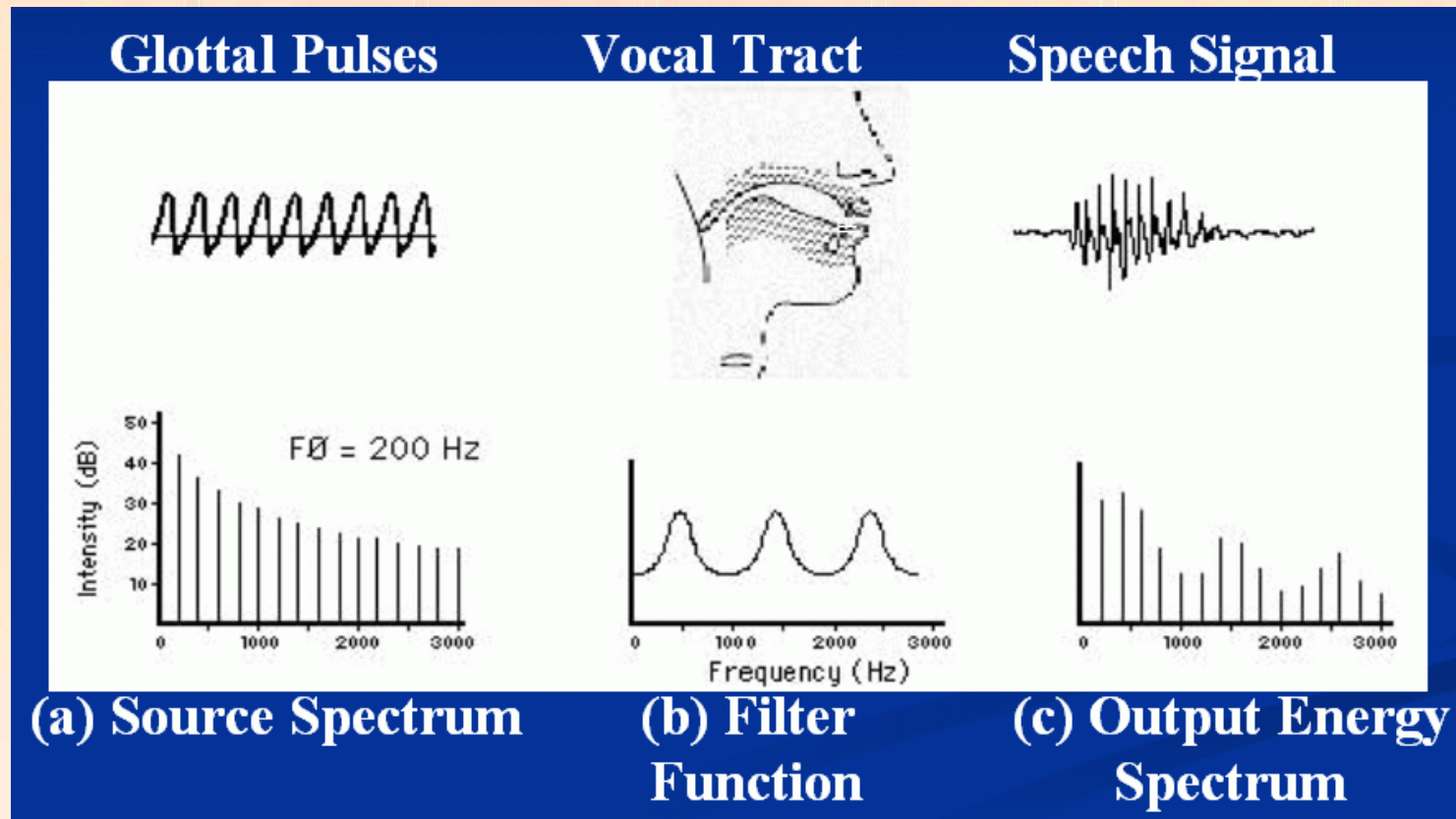


语音产生模型

浙江大学计算机学院
数字媒体与网络技术



语音产生模型



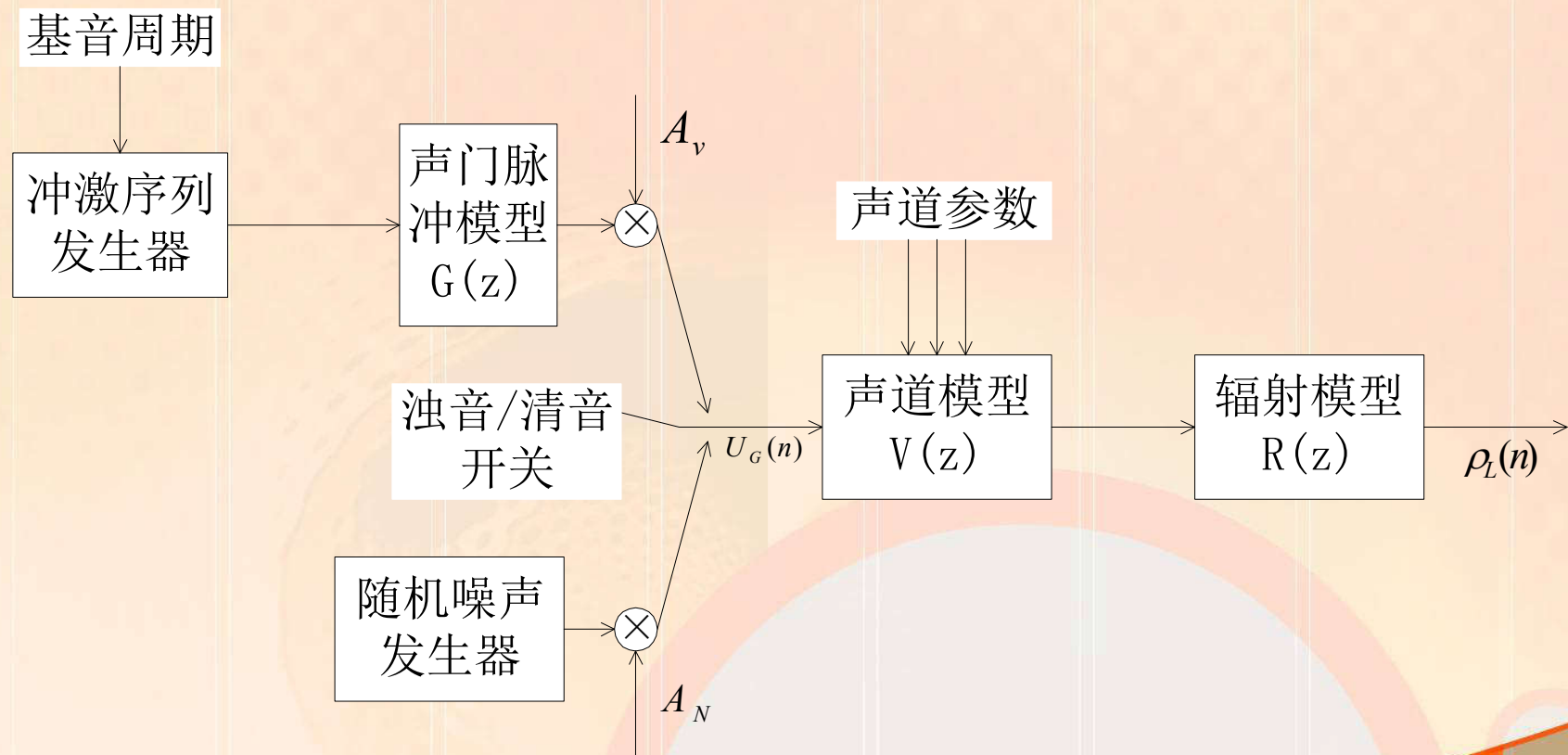
Source-filter model and the corresponding spectrum



语音产生模型



语音信号产生的数字模型

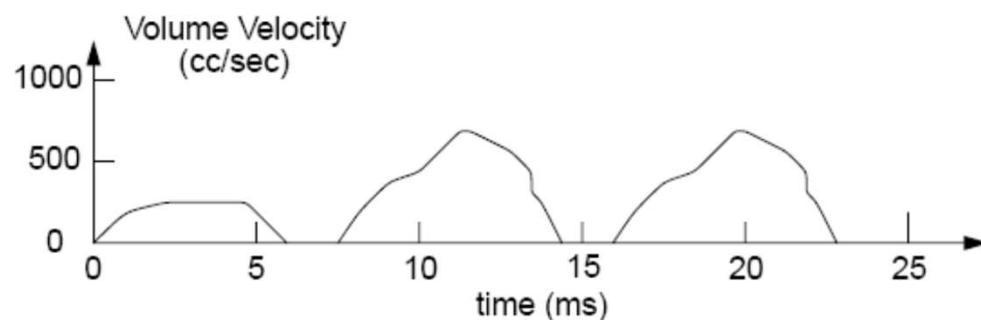


语音产生模型——激励模型



浊音信号产生的激励模型

For voiced sounds, the glottal volume velocity looks something like this:



浊音激励源是以基音周期为周期的斜三角形脉冲串

单个斜三角形脉冲时域表示

$$g(n) = \begin{cases} \frac{1}{2}[1 - \cos(n\pi / N_1)], & 0 \leq n \leq N_1 \\ \cos[(\pi(n - N_1) / N_2)], & N_1 \leq n \leq N_1 + N_2 \\ 0, & \text{其他} \end{cases}$$

单个斜三角形脉冲频域表示为低通滤波器

$$G(z) = \frac{1}{(1 - g_1 z^{-1})(1 - g_2 z^{-1})} \quad g_1, g_2 \text{ 接近 } 1$$



语音产生模型——激励模型

浙江大学计算机学院
数字媒体与网络技术



单个斜三角形脉冲频域表示为低通滤波器

$$G(z) = \frac{1}{(1 - g_1 z^{-1})(1 - g_2 z^{-1})} \quad g_1, g_2 \text{ 接近 } 1$$

斜三角形脉冲可以视作单位脉冲通过上述低通滤波器的输出

单位脉冲的Z变换

$$E(z) = \frac{A_v}{1 - z^{-1}}$$

整个激励源模型

$$U(z) = G(z)E(z) = \frac{A_v}{1 - z^{-1}} \times \frac{1}{(1 - g_1 z^{-1})(1 - g_2 z^{-1})}$$





发清音时，声带处于松弛状态，不发生振动，气流直接进入声道，声道被阻碍形成湍流，清音激励信号相当于一个随机白噪声。实际上用均值为0，方差为1，并在时间或幅值上用白色分布的序列来表示



语音产生模型——声道模型

浙江大学计算机学院
数字媒体与网络技术



声道的两种状态

(1) 发元音时，声道中的口腔为稳定的某种形状谐振腔。由声门带来的准周期脉冲激励声道而产生响应。

(2) 发辅音时，由声门带来的激励在声道某处形成湍流。

声道的两种模型

(1) 声管模型

将声道视作是由多个不同截面积的声管串联而成的系统。

(2) 共振峰模型

将声道视作谐振腔，腔体的谐振频率为共振峰。

语音的表示

(1) 元音 用前3个共振峰

(2) 辅音 用前5个以上共振峰



语音产生模型——声道模型

共振峰特性用如下全极点模型刻画

$$V(z) = \frac{1}{\sum_{i=0}^p a_i z^{-i}}$$

- 说明
- (1) p 为滤波器阶数，一般取8-12
 - (2) a_i 为声道模型参数，随声音的变化而不断变化
 - (3) 在10-30ms内，可以认为声道参数不变

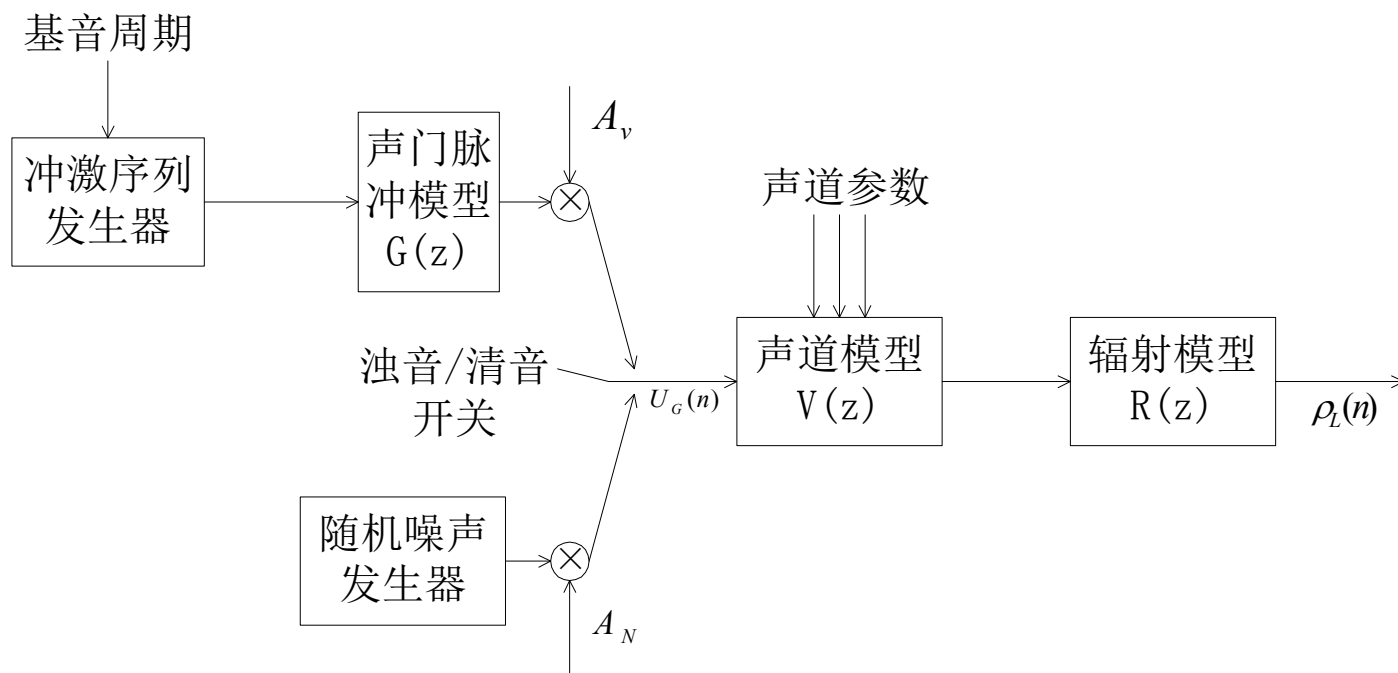
语音产生模型——声道模型

声道的终端是口唇，由口唇辐射引起能量损耗，此辐射效应在高频段较为明显，而在低频段影响较小，用以下高通滤波器表示：

$$R(z) = (1 - rz^{-1})$$

其中， r 接近1

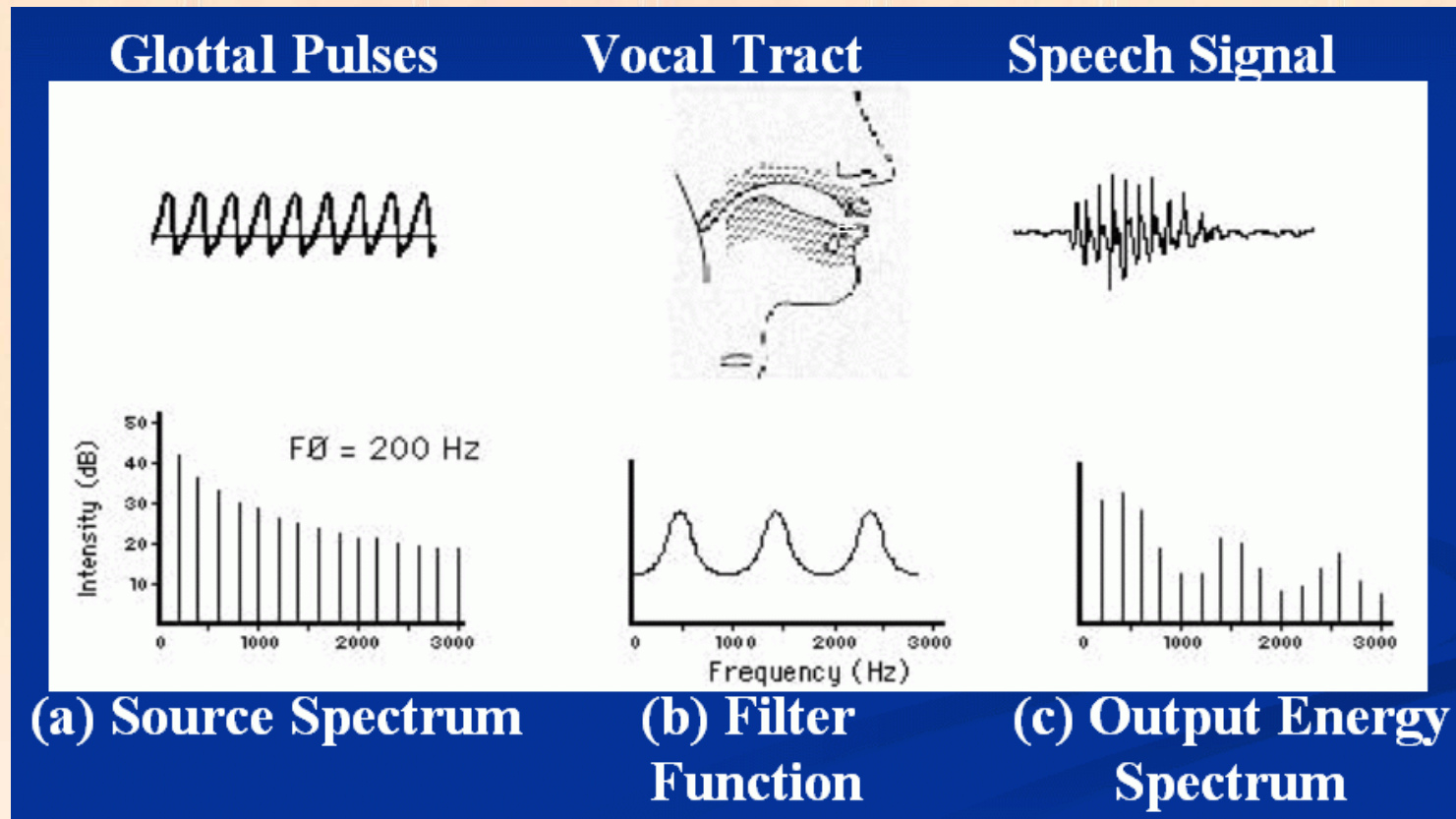
语音产生模型



语音信号产生的完整模型为

$$H(z) = U(z)V(z)R(z)$$

语音产生模型



Source-filter model and the corresponding spectrum

语音技术基础

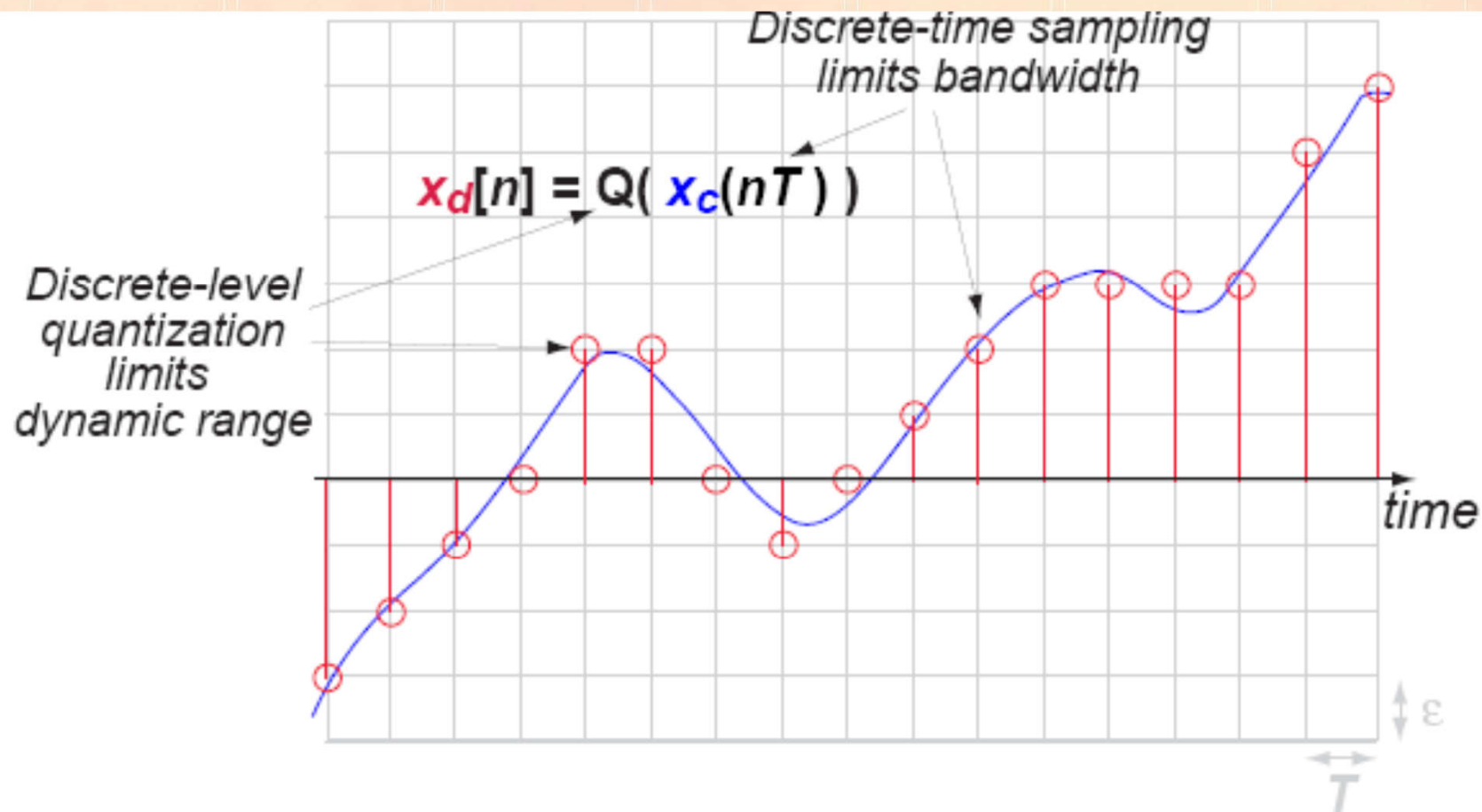
浙江大学计算机学院
数字媒体与网络技术



- 信号处理基础



信号处理——采样





Discrete-Time Signals

□ A sequence of numbers

□ Mathematical representation:

$$x = \{x[n]\}, \quad -\infty < n < \infty$$

□ Sampled from an analog signal, $x_a(t)$, at time $t = nT$,

$$x[n] = x_a(nT), \quad -\infty < n < \infty$$

□ T is called the **sampling period**, and its reciprocal,

$F_s = 1/T$, is called the **sampling frequency**

$$F_s = 8000 \text{ Hz} \leftrightarrow T = 1/8000 = 125 \mu\text{sec}$$

$$F_s = 10000 \text{ Hz} \leftrightarrow T = 1/10000 = 100 \mu\text{sec}$$

$$F_s = 16000 \text{ Hz} \leftrightarrow T = 1/16000 = 62.5 \mu\text{sec}$$

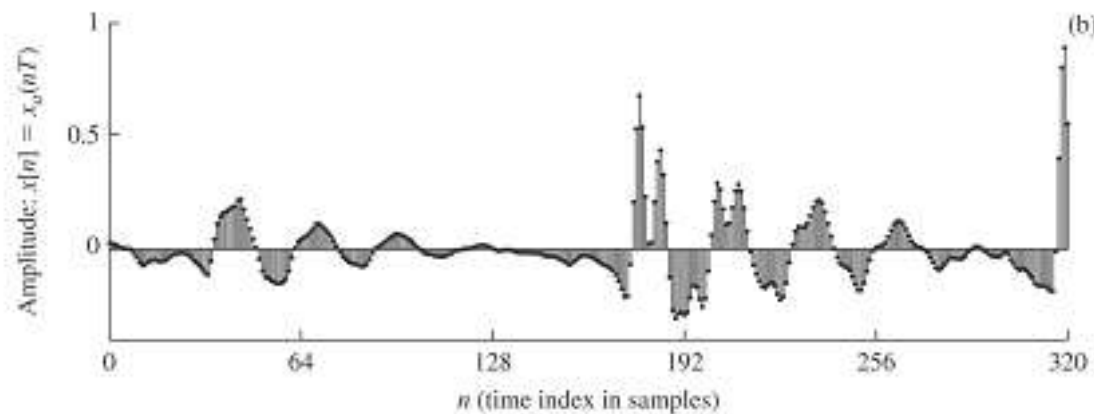
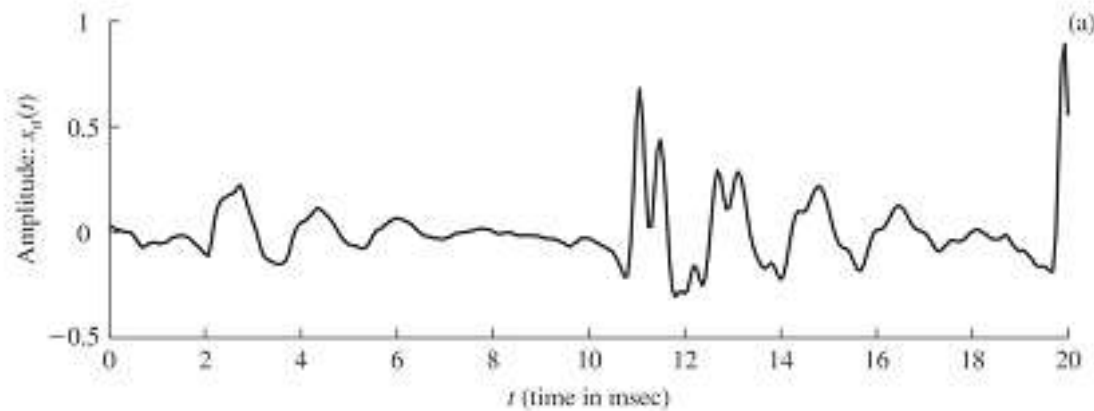
$$F_s = 20000 \text{ Hz} \leftrightarrow T = 1/20000 = 50 \mu\text{sec}$$



信号时域分析



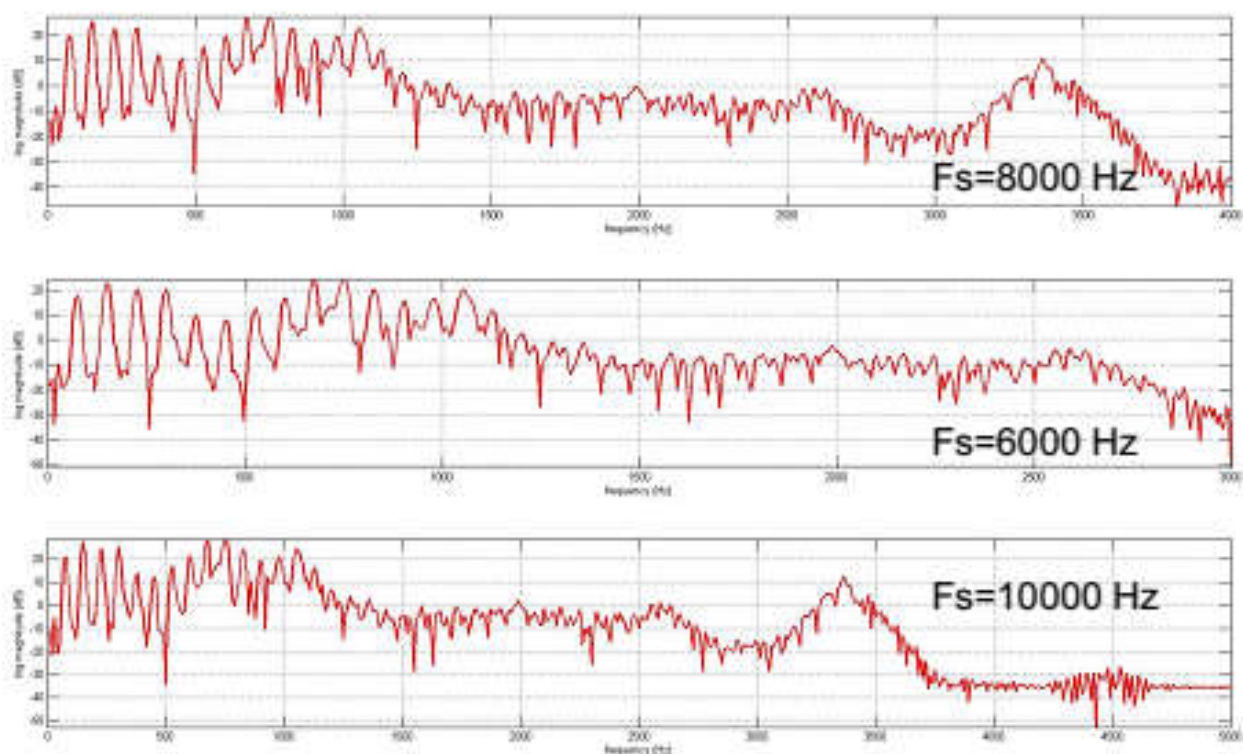
Speech Waveform Display



信号处理——采样



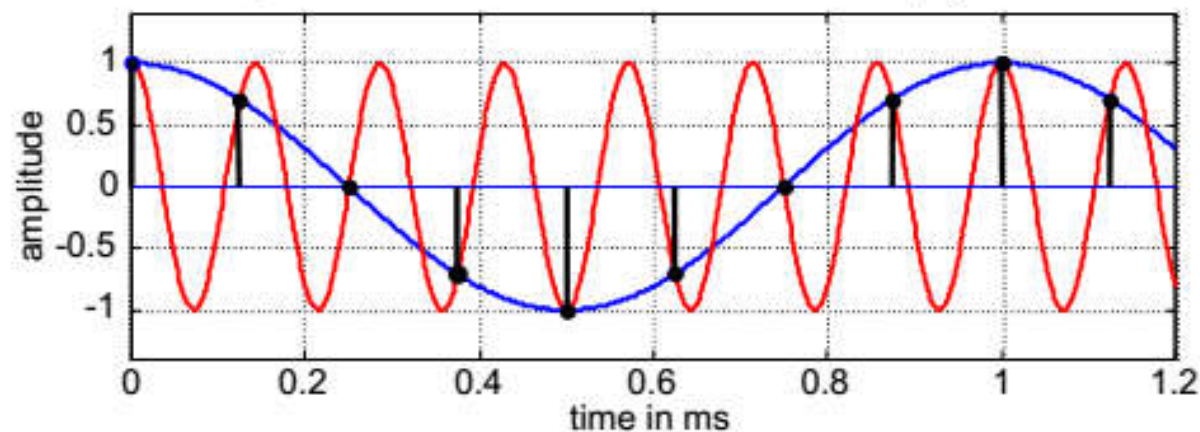
Varying Sampling Rates





The Sampling Theorem

Sampled 1000 Hz and 7000 Hz Cosine Waves; $F_s = 8000$ Hz



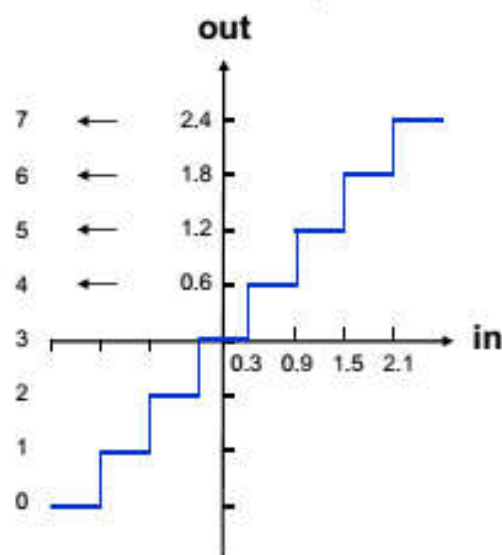
- A bandlimited signal can be reconstructed exactly from samples taken with sampling frequency

$$\frac{1}{T} = F_s \geq 2f_{\text{max}} \quad \text{or} \quad \frac{2\pi}{T} = \omega_s \geq 2\omega_{\text{max}}$$



Quantization

$x[n]$ can be quantized to one of a finite set of values which is then represented digitally in bits, hence a truly digital signal; the course material mostly deals with discrete-time signals (discrete-value only when noted).



A 3-bit uniform quantizer

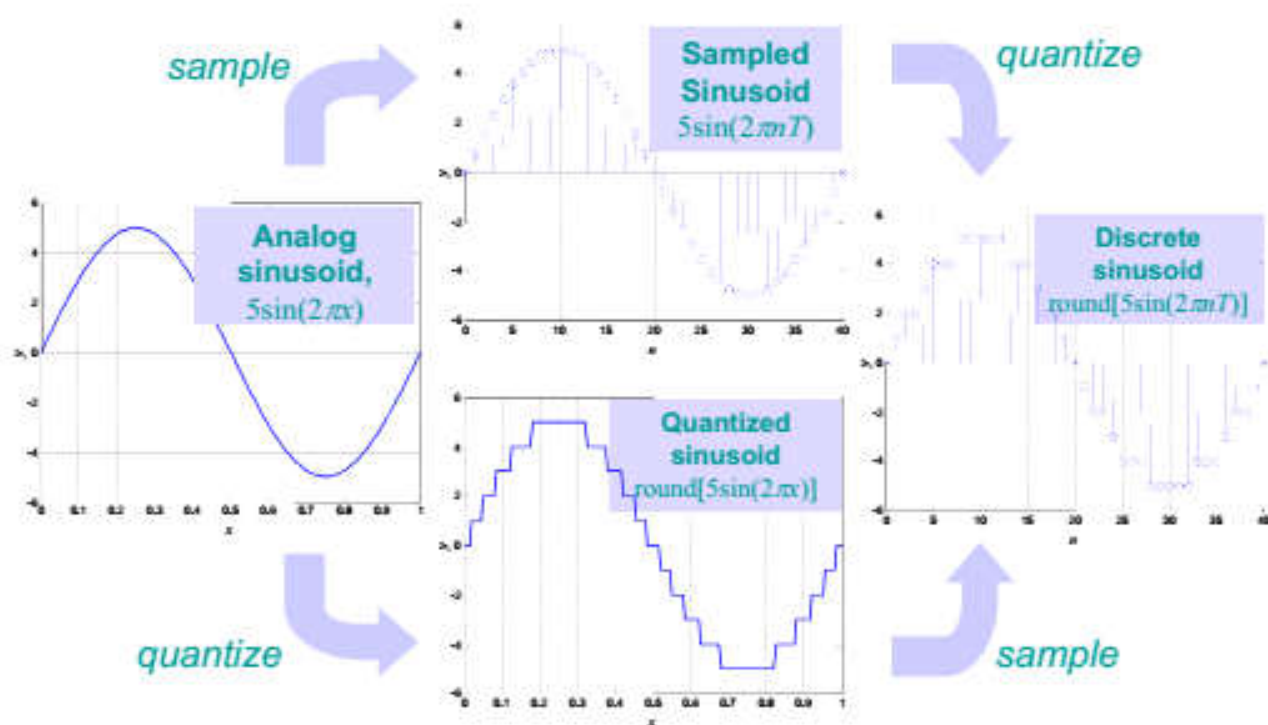
Quantization:

- Transforming a continuously-valued input into a representation that assumes one out of a finite set of values
- The finite set of output values is indexed; e.g., the value 1.8 has an index of 6, or $(110)_2$ in binary representation
- Storage or transmission uses binary representation; a quantization table is needed

信号处理——表示



Discrete Signals





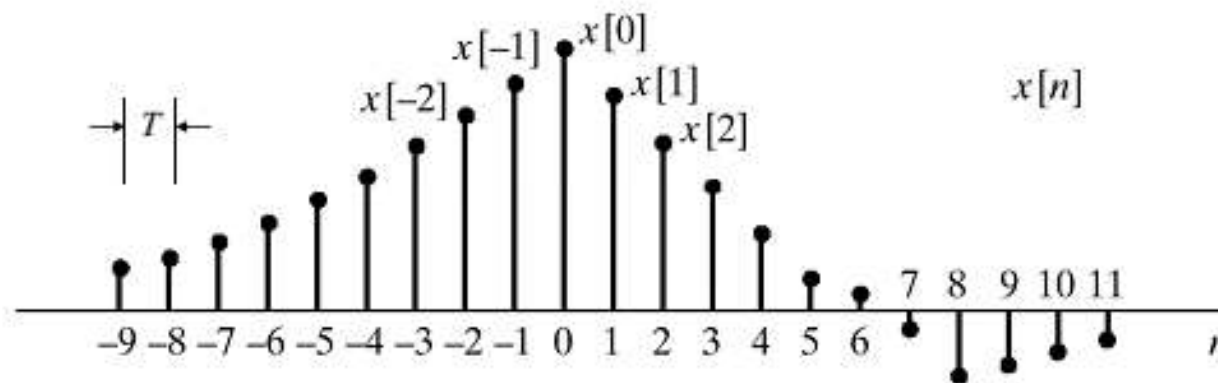
Issues with Discrete Signals

- what sampling rate is appropriate
 - 6.4 kHz (telephone bandwidth), 8 kHz (extended telephone BW), 10 kHz (extended bandwidth), 16 kHz (hi-fi speech)
- how many quantization levels are necessary at each bit rate (bits/sample)
 - 16, 12, 8, ... => ultimately determines the S/N ratio of the speech
 - speech coding is concerned with answering this question in an optimal manner





Discrete-Time (DT) Signals are Sequences



- $x[n]$ denotes the "sequence value at 'time' n "
- Sources of sequences:
 - Sampling a continuous-time signal
$$x[n] = x_c(nT) = x_c(t)|_{t=nT}$$
 - Mathematical formulas – generative system
e.g., $x[n] = 0.3 \cdot x[n-1] - 1$; $x[0] = 40$

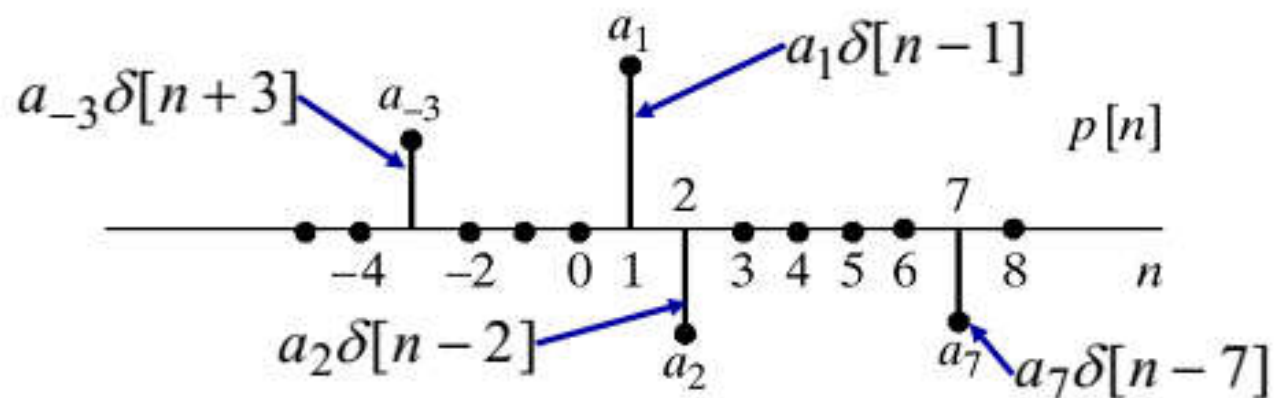


Impulse Representation of Sequences

A sequence, a function

$$x[n] = \sum_{k=-\infty}^{\infty} x[k] \delta[n-k]$$

Value of the function at k



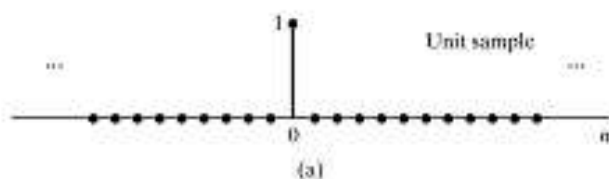
$$x[n] = a_{-3}\delta[n+3] + a_1\delta[n-1] + a_2\delta[n-2] + a_7\delta[n-7]$$



Some Useful Sequences

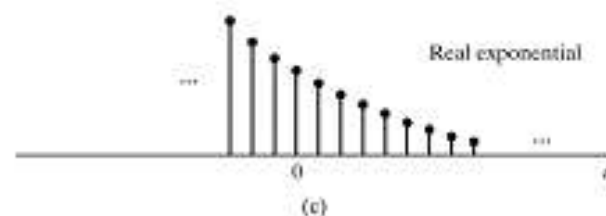
unit sample

$$\delta[n] = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases}$$



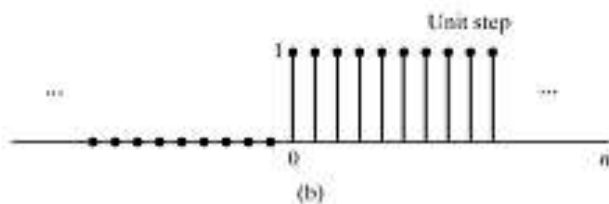
real exponential

$$x[n] = \alpha^n$$



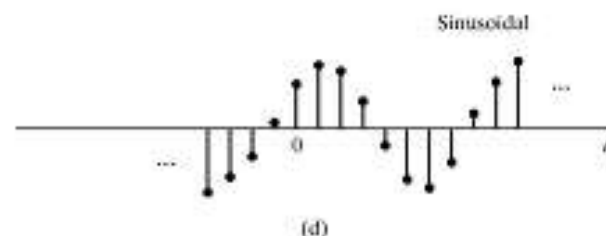
unit step

$$u[n] = \begin{cases} 1, & n \geq 0 \\ 0, & n < 0 \end{cases}$$



sine wave

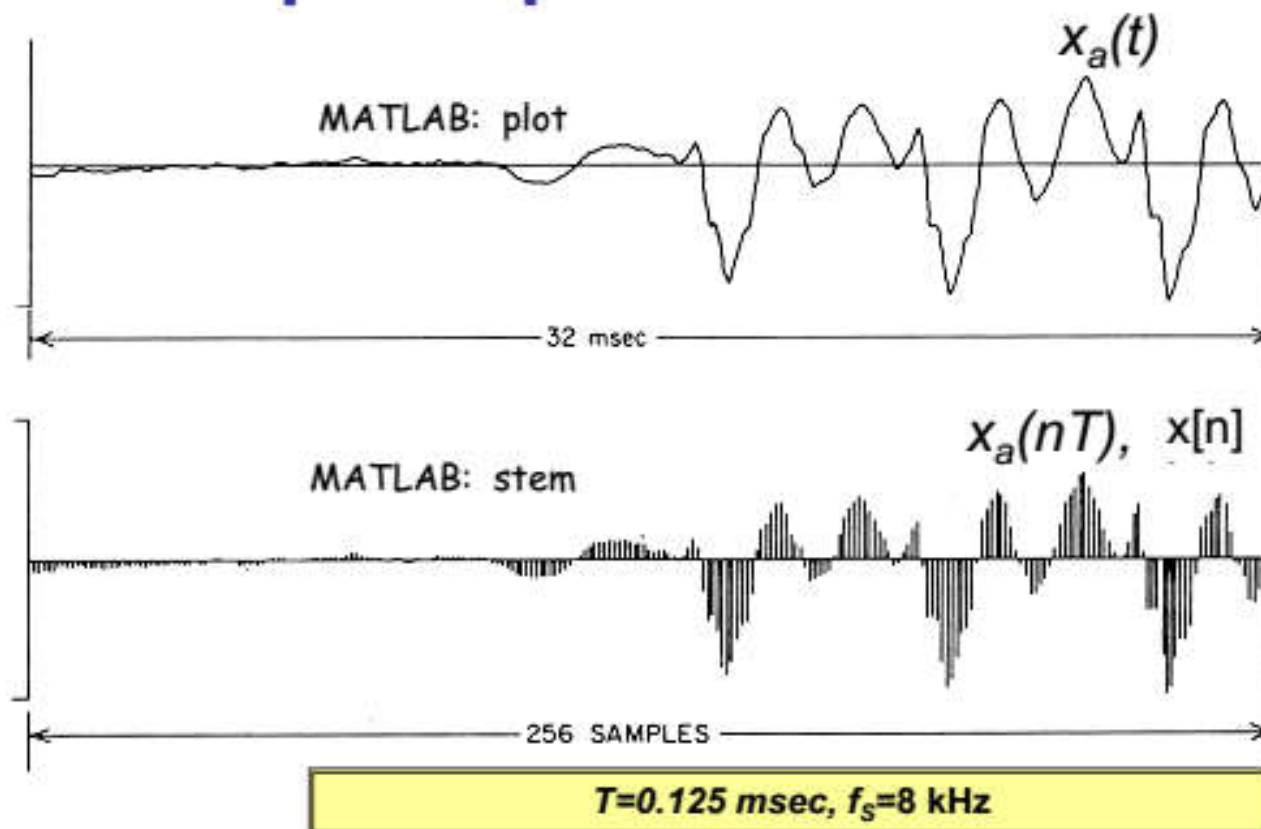
$$x[n] = A \cos(\omega_0 n + \phi)$$



信号处理——表示



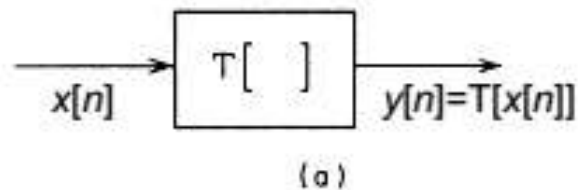
Sampled Speech Waveform



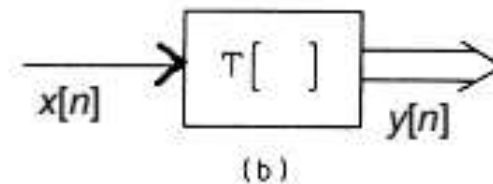


Signal Processing

- Transform digital signal into more desirable form



↑
single input—single output



↑
single input—multiple output,
e.g., filter bank analysis,
sinusoidal sum analysis, etc.





LTI Discrete-Time Systems



- Linearity (superposition):

$$T\{ax_1[n] + bx_2[n]\} = aT\{x_1[n]\} + bT\{x_2[n]\}$$

- Time-Invariance (shift-invariance):

$$x_1[n] = x[n - n_d] \Rightarrow y_1[n] = y[n - n_d]$$

- LTI implies discrete convolution:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = x[n] * h[n] = h[n] * x[n]$$





Convolution Example

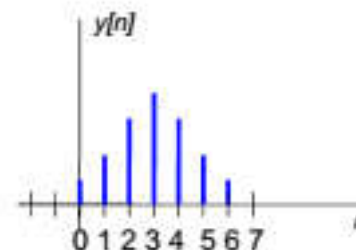
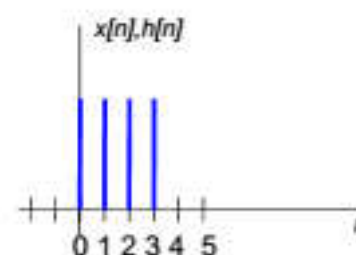
$$x[n] = \begin{cases} 1 & 0 \leq n \leq 3 \\ 0 & \text{otherwise} \end{cases} \quad h[n] = \begin{cases} 1 & 0 \leq n \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

What is $y[n]$ for this system?

Solution :

$$y[n] = x[n] * h[n] = \sum_{m=-\infty}^{\infty} h[m] x[n-m]$$

$$= \begin{cases} \sum_{m=0}^n 1 \cdot 1 = (n+1) & 0 \leq n \leq 3 \\ \sum_{m=n-3}^3 1 \cdot 1 = (7-n) & 4 \leq n \leq 6 \\ 0 & n \leq 0, n \geq 7 \end{cases}$$



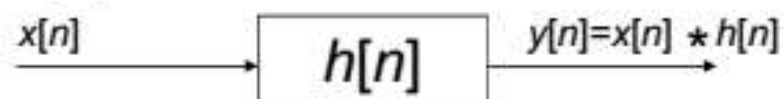


Linear Time-Invariant Systems

- easiest to understand
- easiest to manipulate
- powerful processing capabilities
- characterized completely by their response to unit sample, $h(n)$, via convolution relationship

$$y[n] = x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k] h[n-k] = \sum_{k=-\infty}^{\infty} h[k] x[n-k]$$

$y[n] = h[n] * x[n]$, where $*$ denotes discrete convolution

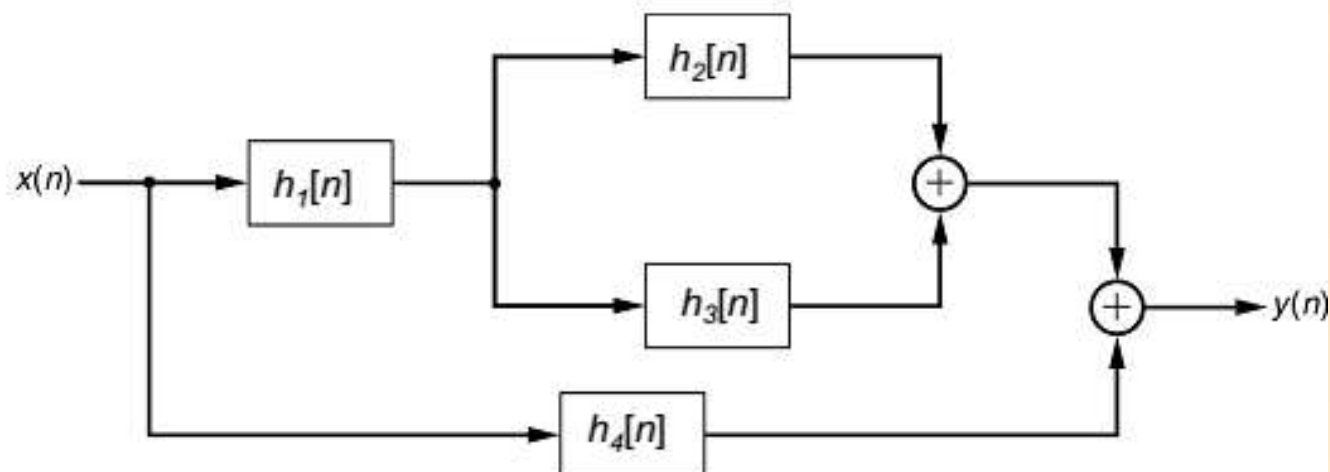


- basis for linear filtering
- used as models for speech production (source convolved with system)





More Complex Filter Interconnections



$$y[n] = x[n] * h_c[n]$$

$$h_c[n] = h_1[n] * (h_2[n] + h_3[n]) + h_4[n]$$



Transform Representations

- z-transform:

$$x[n] \longleftrightarrow X(z)$$

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$$

$$x[n] = \frac{1}{2\pi j} \oint_C X(z)z^{n-1}dz$$

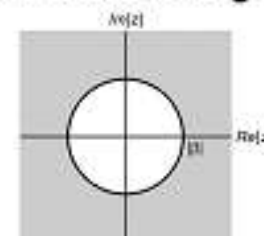
infinite power series in z^{-1} ,
with $x[n]$ as coefficients of
term in z^{-n}

- direct evaluation using residue theorem
- partial fraction expansion of $X(z)$
- long division
- power series expansion

- $X(z)$ converges (is finite) only for certain values of z :

$$\sum_{n=-\infty}^{\infty} |x[n]| |z^{-n}| < \infty \quad - \text{sufficient condition for convergence}$$

- region of convergence: $R_1 < |z| < R_2$





Transform Properties

Linearity	$ax_1[n] + bx_2[n]$	$aX_1(z) + bX_2(z)$
Shift	$x[n - n_0]$	$z^{-n_0} X(z)$
Exponential Weighting	$a^n x[n]$	$X(a^{-1}z)$
Linear Weighting	$n x[n]$	$-z \frac{dX(z)}{dz}$
Time Reversal	$x[-n]$ <small>non-causal, need $x[N_0 - n]$ to be causal for finite length sequence</small>	$X(z^{-1})$
Convolution	$x[n] * h[n]$	$X(z) H(z)$
Multiplication of Sequences	$x[n] w[n]$	$\frac{1}{2\pi j} \oint_C X(v) W(z/v) v^{-1} dv$ <small>circular convolution in the frequency domain</small>



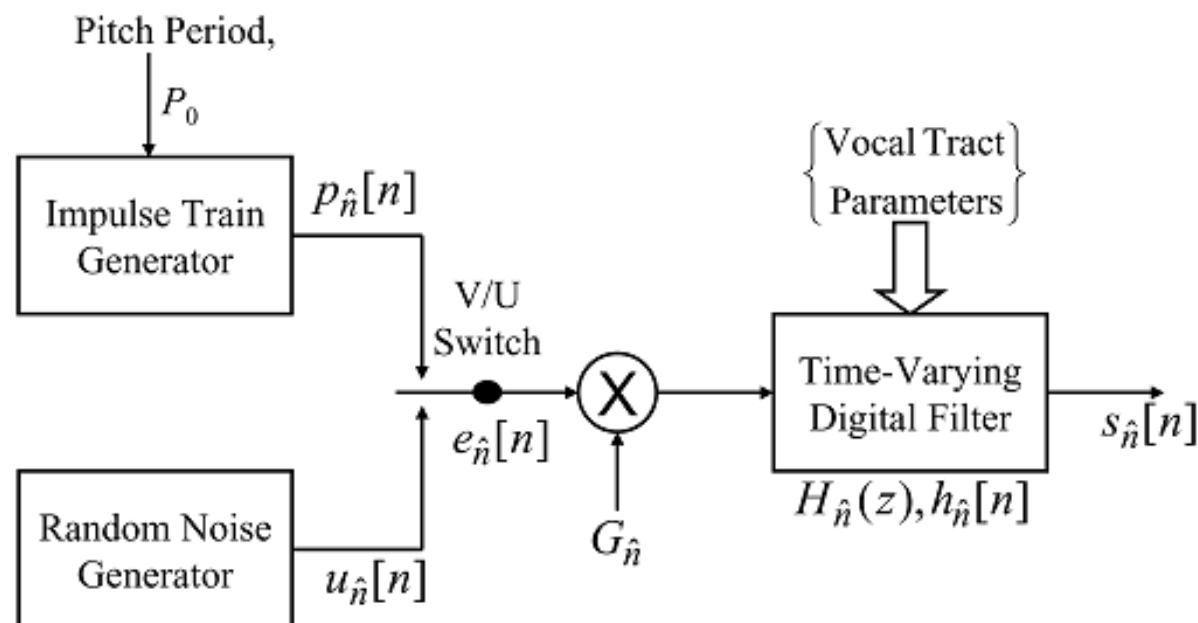


Fig. 4.1 Voiced/unvoiced/system model for a speech signal.

$$s_{\hat{n}}[n] = \sum_{m=0}^{\infty} h_{\hat{n}}[m] e_{\hat{n}}[n - m], \quad (4.1)$$





$$H(z) = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}}. \quad (4.2)$$

$$s[n] = \sum_{k=1}^p a_k s[n - k] + Ge[n], \quad (4.3)$$

$$X_{\hat{n}} = \sum_{m=-\infty}^{\infty} T\{x[m]w[\hat{n} - m]\}, \quad (4.4)$$



短时平稳假设

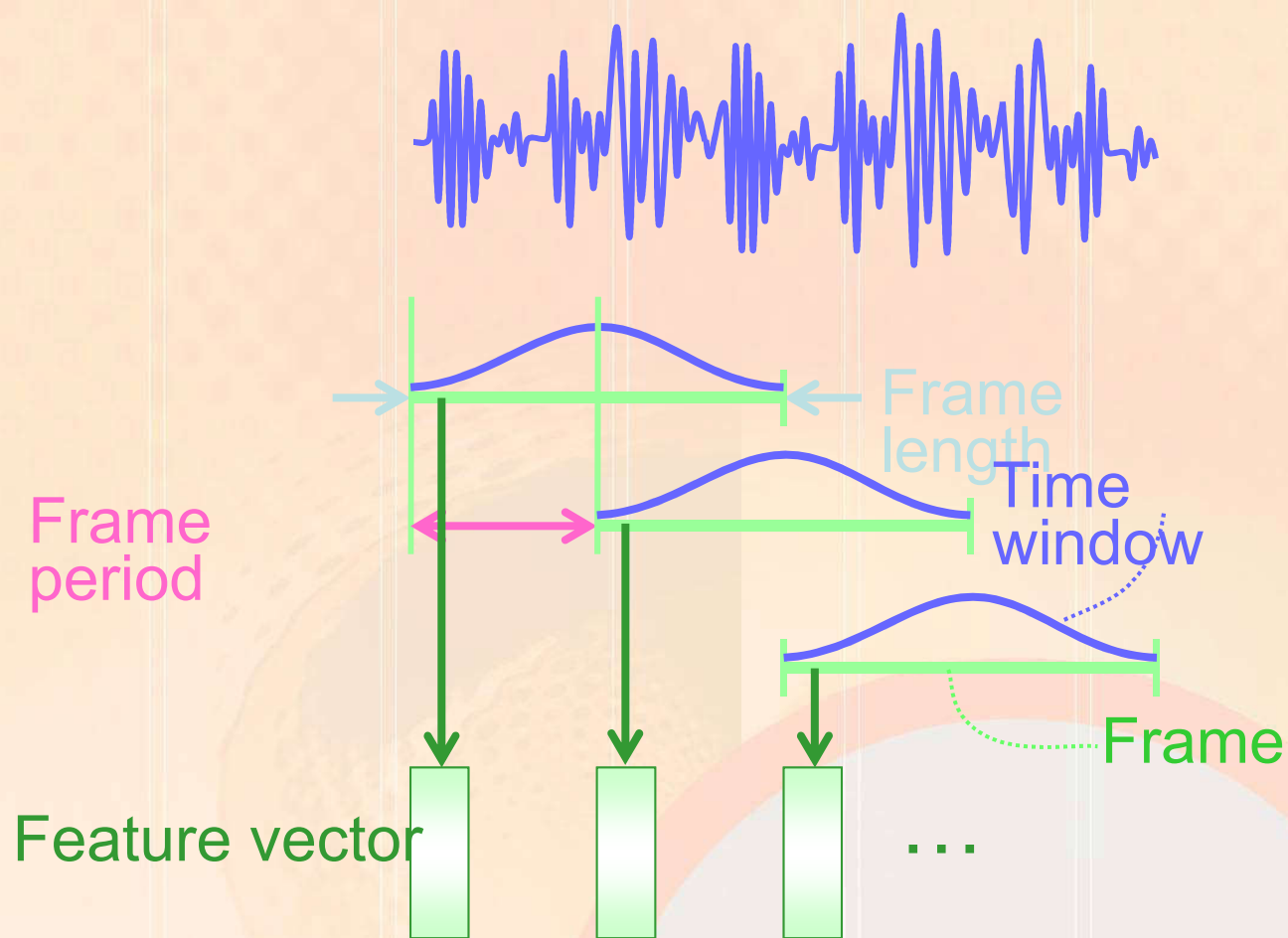


语音信号特性是随时间而变化的，本质上是一个非平稳过程。但不同的语音是由人的口腔肌肉运动构成声道的某种形状而产生的响应，而这种肌肉运动频率相对于语音频率来说是缓慢的，因而在一个短时间范围内，其特性基本保持不变，即相对稳定，可以视为一个准稳态过程。基于这样的考虑，对语音信号进行分段考虑，每一段称为一帧（**frame**）。一般假设为10-30ms的短时间隔。



语音特征提取

浙江大学计算机学院
数字媒体与网络技术



语音特征矢量(短时谱)提取



语音分析技术

浙江大学计算机学院
数字媒体与网络技术



- 语音时域分析
- 语音频域分析



语音时域分析



直接对语音的时域波形进行分析，简单直观、清晰易懂、运算量小‘物理意义明确。时域波形很难反映语音感知特性，且易受环境变化影响。主要的参数为：

音量（**Volume**）

过零率（**Zero Crossing Rate**）

音高/基音周期（**Pitch**）



语音频域分析



将语音的时域波形转换至频域进行分析。频谱特征具有具有实际的物理意义与明显的声学特性，且不易受环境变化影响。主要的参数为：

共振峰（**Formant**）

音高/基音周期（**Pitch**）

分析方法包括：

滤波器组法

傅里叶变化法

线性预测分析法



语音频域分析



将语音的对数功率谱进行反傅里叶变换后得到，可以进一步将声道特性和激励特性有效分开，可以更好地揭示语音的本质特征。主要的参数为：

MFCC（梅尔倒谱系数）

LPCC（线性预测倒谱系数）



语音分析技术

浙江大学计算机学院
数字媒体与网络技术



- 语音时域分析
- 语音频域分析



语音信号时域分析



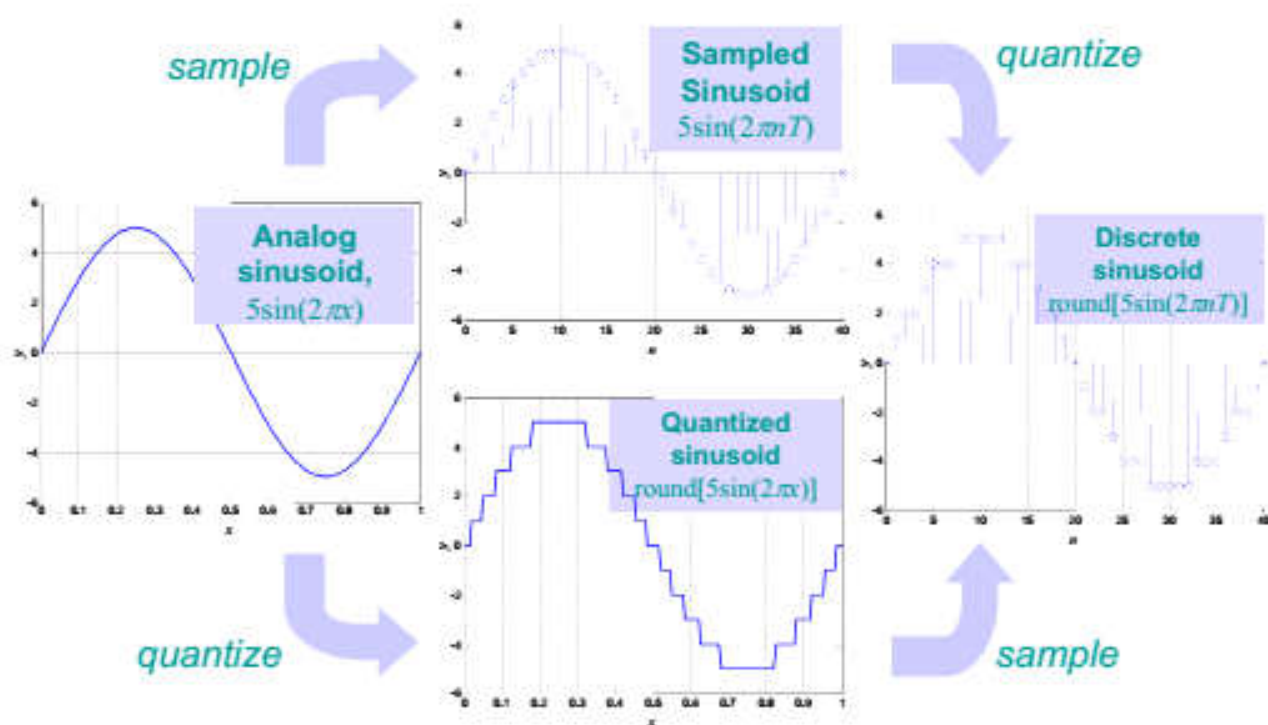
- 预处理
- 能量/音量 (Energy/Volume)
- 过零率 (Zero Crossing Rate)
- 端点检测 (End-Point Detection)
- 基频 (F0)



预处理——采样、量化



Discrete Signals



预处理——预加重

由于语音信号的平均功率谱受声门激励和口鼻辐射的影响，高频端大约在800Hz以上按-6dB/倍频程跌落，需要进行提升

$$H(z) = (1 - uz^{-1})$$

其中, u 取值为0.94-0.97

语音信号时域分析



- 预处理
- 能量/音量 (**Energy/Volume**)
- 过零率 (**Zero Crossing Rate**)
- 端点检测 (**End-Point Detection**)
- 基频 (**F0**)



能量/音量

音量：代表声音的强度，又称为“力度”、“强度”（**Intensity**）或“能量”（**Energy**），可由一帧内的语音采样点振幅大小来类比，基本上两种计算方式：

1、
$$E_n = \sum_{i=1}^n |S_i|$$

每帧内的语音采样点振幅的绝对值值的总和。此方法的计算较简单，只需要整数运算。

2、
$$E_n = 10 \times \log \sum_{i=1}^n S_i^2$$

每帧内的语音采样点振幅的平方值的总和，再取以 10 为底对数值，再乘以10：此方法得到的值是以分贝（**Decibels**）为单位，是一个相对强度的值，比较符合人耳对于大小声音的感觉。



短时平均能量指在一个短时音频帧内采样点信号所聚集的平均能量。假定一段连续音频信号流 x 到 K 个采样点，这 K 个采样点被分割成速加率为 50% 的 M 个短时帧。每个短时帧和窗口函数大小假定为 N ，对于第 m 个短时帧，其短时平均能量可以使用下面公式计算：

$$E_m = \frac{1}{N} \sum_n [x(n)w(n-m)]^2$$

其中， $x(n)$ 表示第 m 个短时帧信号中第 n 个采样信号值， $w(n)$ 是长度为 N 的窗口函数。

The short-time energy is defined as

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])^2 = \sum_{m=-\infty}^{\infty} x^2[m]w^2[\hat{n}-m]. \quad (4.6)$$





特点:

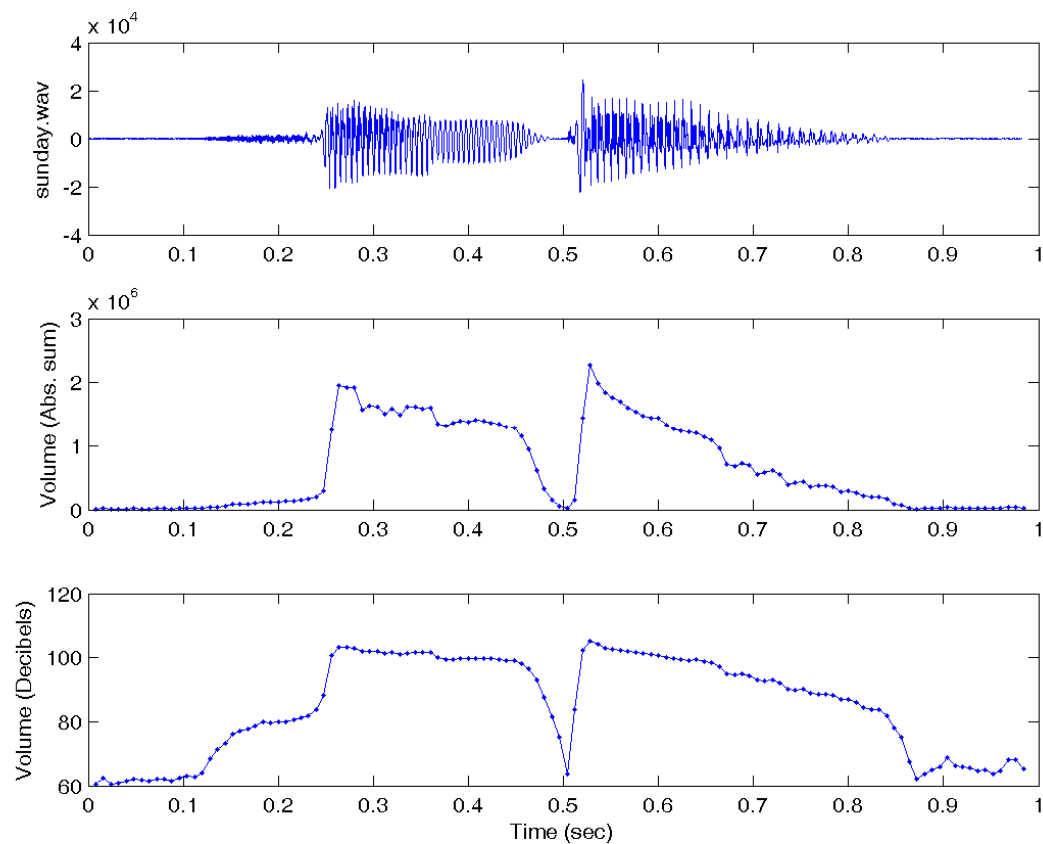
- 1、一般而言，有浊音的音量大语清音的音量，而清音的音量又大于噪音的音量。
- 2、是一个相对性的指标，受到麦克风设定的影响很大。
- 3、通常用在端点检测，检测浊音的声母或韵母的开始及结束位置。
- 4、在计算前最好先减去语音信号的平均值，以避免语音的直流偏移（DC Bias）所导致的误差。



能量/音量



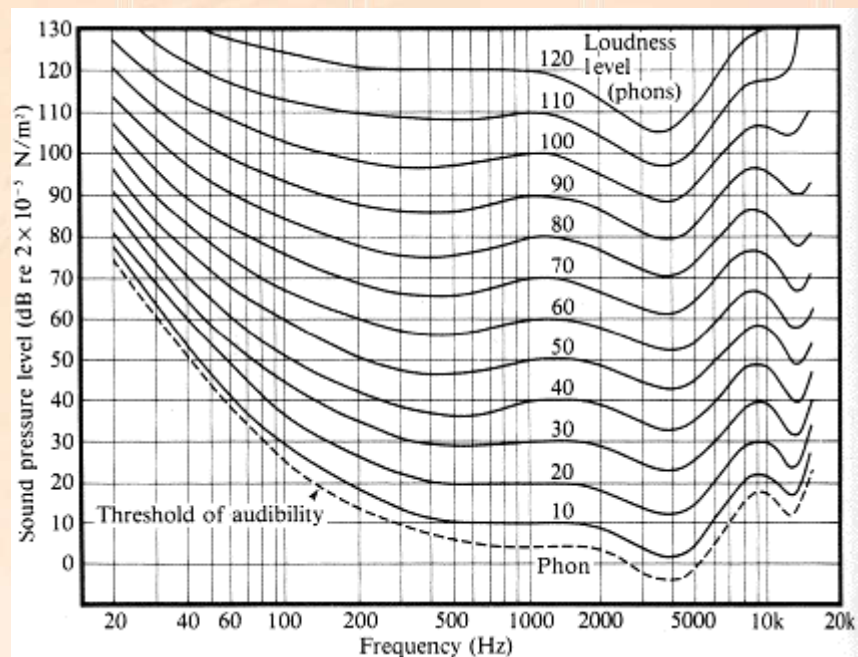
举例1: Sunday



能量/音量



基本上我们使用音量来表示声音的强弱，但是前述两种计算音量的方法，只是用数学的公式来逼近人耳的感觉，和人耳的感觉有时候会有相当大的差别，为了区分，我们使用“主观音量”来表示人耳所听到的音量大小。例如，人耳对于同样振幅但不同频率的声音，所产生的主观音量就会非常不一样。若把以人耳为测试主体的「等主观音量曲线」（Curves of Equal Loudness）画出来，就可以得到下图：

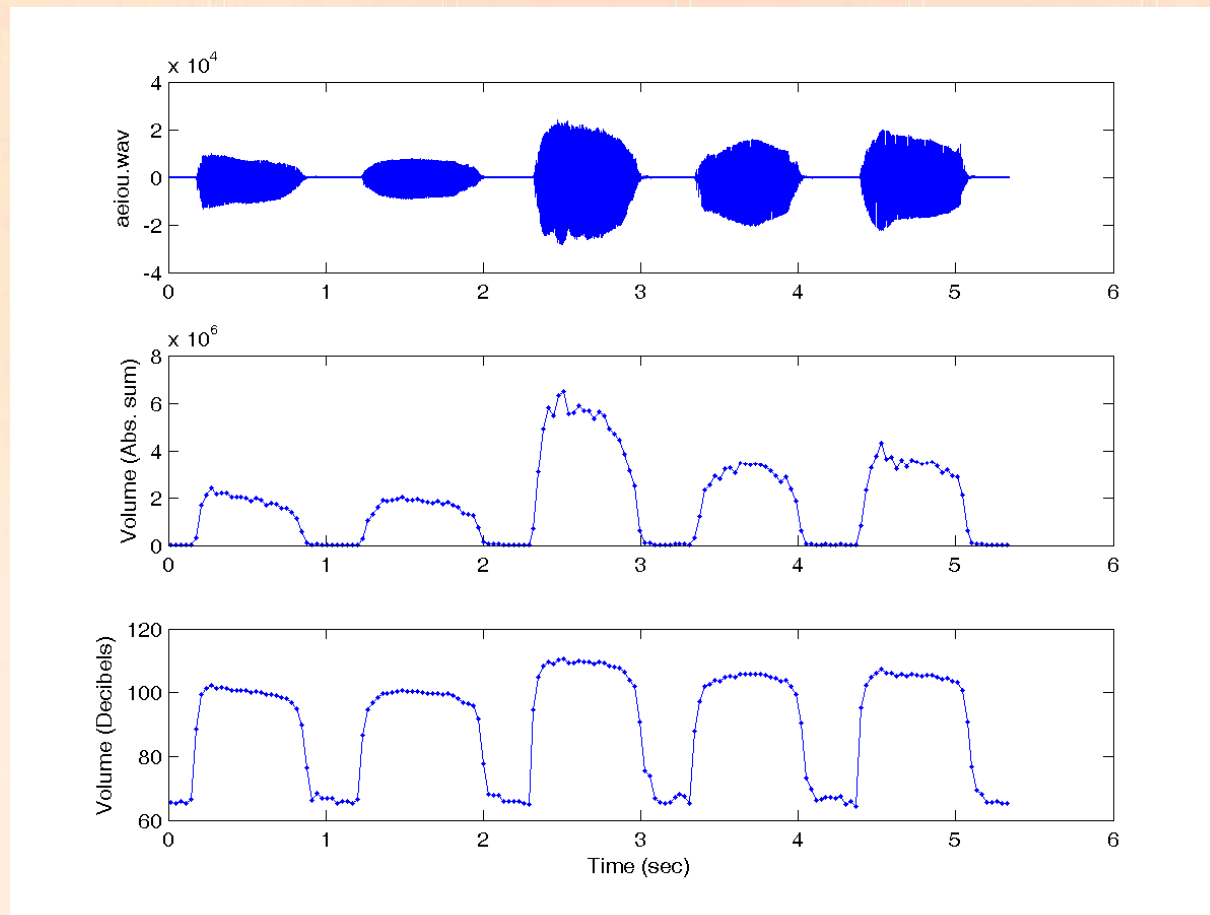


能量/音量



主观音量除了和频率有关外，也和语音的内容（音色）有关

举例2: a,i,u,e,o



语音信号时域分析



- 预处理
- 能量/音量 (Energy/Volume)
- 过零率 (Zero Crossing Rate)
- 端点检测 (End-Point Detection)
- 基频 (F0)





“过零率(Zero-crossing Rate)”指在一个短时帧内，离散采样信号值由正到负和由负到正变化的次数，这个量大概能够反映信号在短时帧内里的平均频率^[14]。对于音频信号流 x 中第 m 帧，其过零率计算如下：

$$Z_{\hat{n}} = \sum_{m=-\infty}^{\infty} 0.5 |\text{sgn}\{x[m]\} - \text{sgn}\{x[m-1]\}| w[\hat{n} - m], \quad (4.7)$$

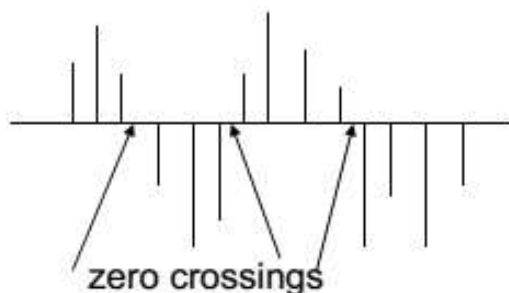
where

$$\text{sgn}\{x\} = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases} \quad (4.8)$$





Short-Time Average ZC Rate



zero crossing => successive samples
have different algebraic signs

- zero crossing rate is a simple measure of the 'frequency content' of a signal—especially true for narrowband signals (e.g., sinusoids)
- sinusoid at frequency F_0 with sampling rate F_s has F_s/F_0 samples per cycle with two zero crossings per cycle, giving an average zero crossing rate of

$$z_1 = (2) \text{ crossings/cycle} \times (F_0 / F_s) \text{ cycles/sample}$$

$$z_1 = 2F_0 / F_s \text{ crossings/sample (i.e., } z_1 \text{ proportional to } F_0)$$

$$z_M = M (2F_0 / F_s) \text{ crossings/(M samples)}$$





Sinusoid Zero Crossing Rates

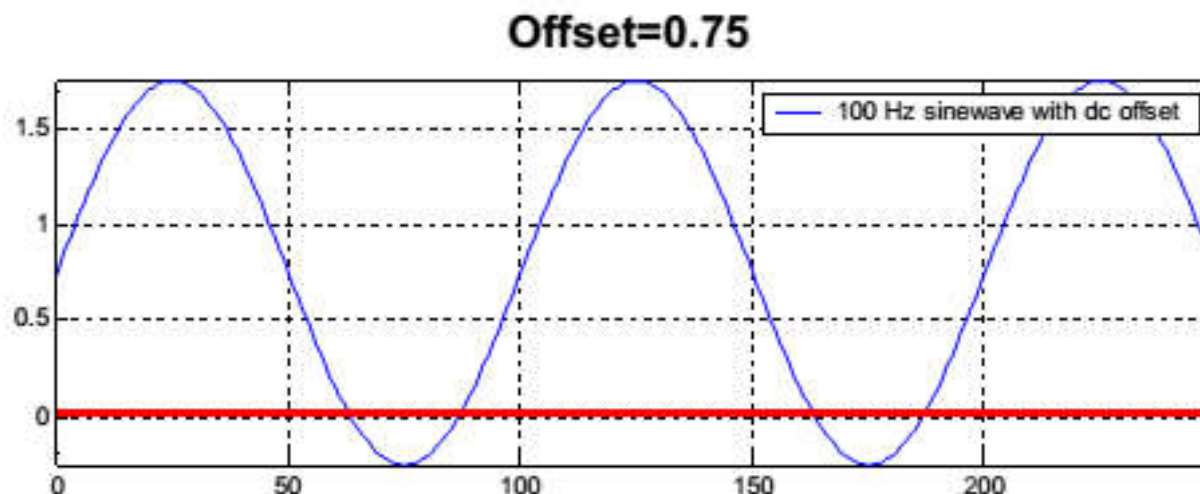
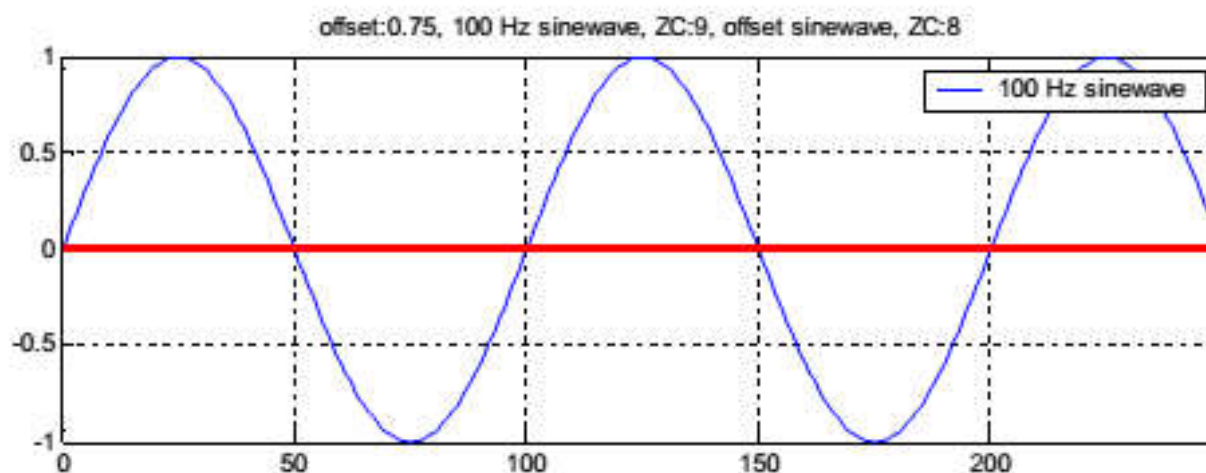
Assume the sampling rate is $F_s = 10,000$ Hz

1. $F_0 = 100$ Hz sinusoid has $F_s / F_0 = 10,000 / 100 = 100$ samples/cycle;
or $z_1 = 2 / 100$ crossings/sample, or $z_{100} = 2 / 100 * 100 =$
2 crossings/10 msec interval
2. $F_0 = 1000$ Hz sinusoid has $F_s / F_0 = 10,000 / 1000 = 10$ samples/cycle;
or $z_1 = 2 / 10$ crossings/sample, or $z_{100} = 2 / 10 * 100 =$
20 crossings/10 msec interval
3. $F_0 = 5000$ Hz sinusoid has $F_s / F_0 = 10,000 / 5000 = 2$ samples/cycle;
or $z_1 = 2 / 2$ crossings/sample, or $z_{100} = 2 / 2 * 100 =$
100 crossings/10 msec interval



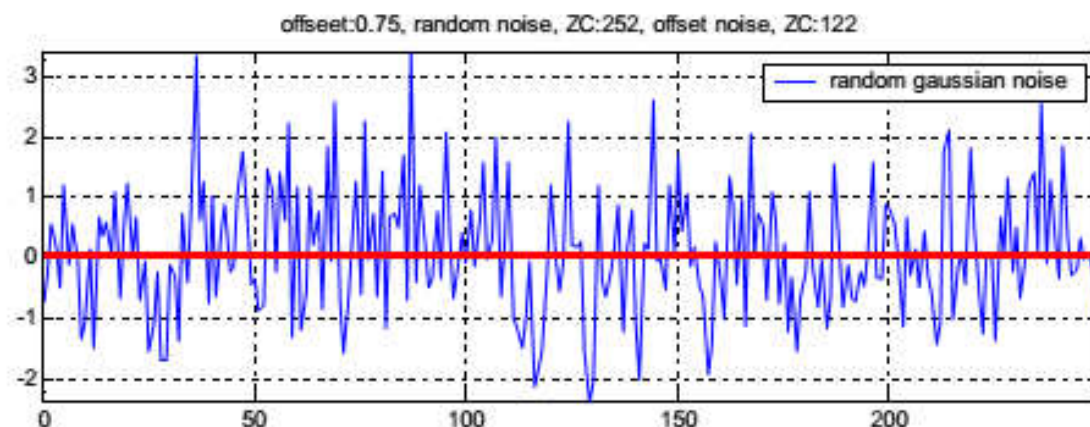


Zero Crossing for Sinusoids

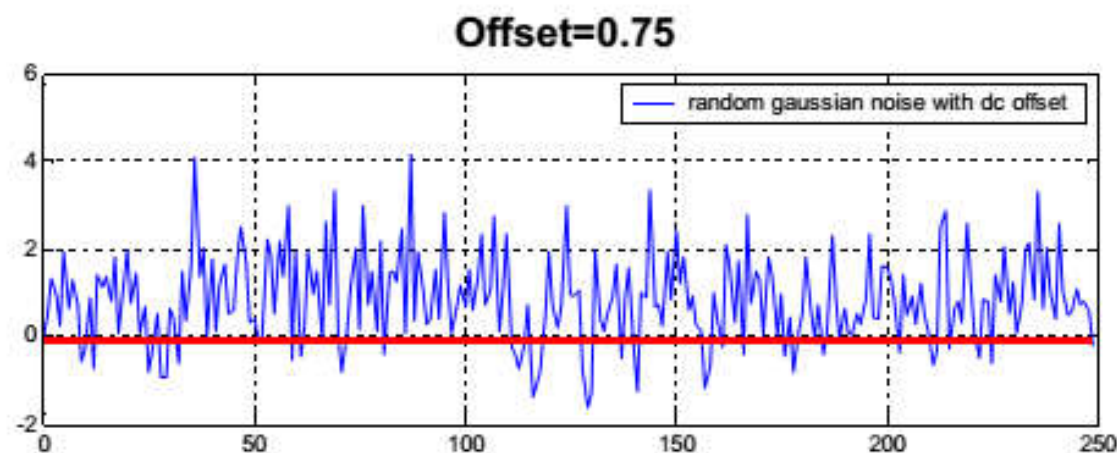




Zero Crossings for Noise



ZC=252



ZC=122



ZC Rate Definitions

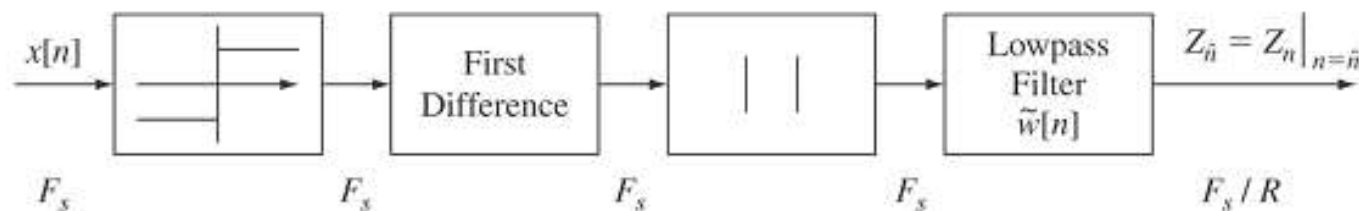
$$Z_{\hat{n}} = \frac{1}{2L_{\text{eff}}} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}(x[m]) - \text{sgn}(x[m-1])| \tilde{w}[\hat{n}-m]$$

$$\begin{aligned} \text{sgn}(x[n]) &= 1 & x[n] \geq 0 \\ &= -1 & x[n] < 0 \end{aligned}$$

□ simple rectangular window:

$$\begin{aligned} \tilde{w}[n] &= 1 & 0 \leq n \leq L-1 \\ &= 0 & \text{otherwise} \end{aligned}$$

$$L_{\text{eff}} = L$$



Same form for $Z_{\hat{n}}$ as for $E_{\hat{n}}$ or $M_{\hat{n}}$





ZC Normalization

□ The formal definition of $Z_{\hat{n}}$ is:

$$Z_{\hat{n}} = z_1 = \frac{1}{2L} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}(x[m]) - \text{sgn}(x[m-1])|$$

is interpreted as the number of zero crossings per sample.

□ For most practical applications, we need the rate of zero crossings per fixed interval of M samples, which is

$$z_M = z_1 \cdot M = \text{rate of zero crossings per } M \text{ sample interval}$$

Thus, for an interval of τ sec., corresponding to M samples we get

$$z_M = z_1 \cdot M; \quad M = \tau F_s = \tau / T$$

□ $F_s = 10,000$ Hz; $T = 100$ μ sec; $\tau = 10$ msec; $M = 100$ samples

□ $F_s = 8,000$ Hz; $T = 125$ μ sec; $\tau = 10$ msec; $M = 80$ samples

□ $F_s = 16,000$ Hz; $T = 62.5$ μ sec; $\tau = 10$ msec; $M = 160$ samples

Zero crossings/10 msec interval as a function of sampling rate





ZC Normalization

- For a 1000 Hz sinewave as input, using a 40 msec window length (L), with various values of sampling rate (F_s), we get the following:

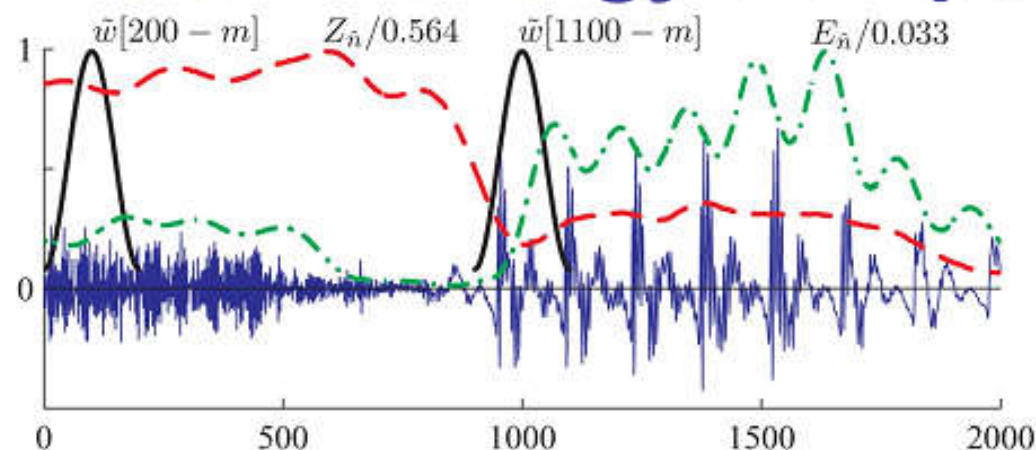
F_s	L	z_1	M	z_M
8000	320	1/4	80	20
10000	400	1/5	100	20
16000	640	1/8	160	20

- Thus we see that the normalized (per interval) zero crossing rate, z_M , is independent of the sampling rate and can be used as a measure of the dominant energy in a band.

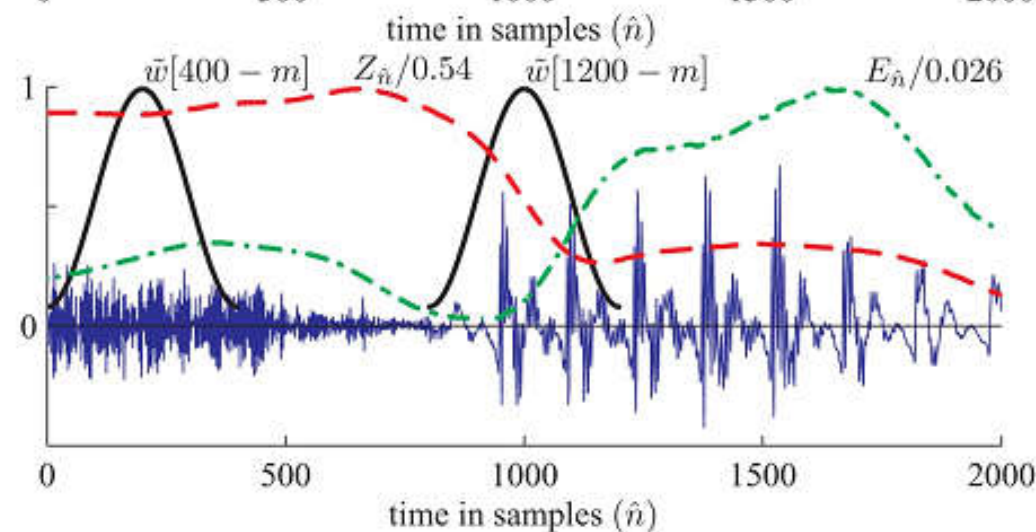




ZC and Energy Computation



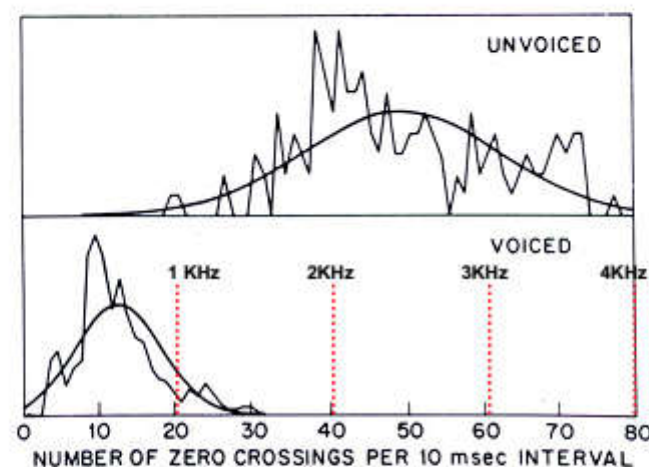
Hamming window
with duration
 $L=201$ samples
(12.5 msec at
 $F_s=16$ kHz)



Hamming window
with duration
 $L=401$ samples
(25 msec at
 $F_s=16$ kHz)



ZC Rate Distributions



Unvoiced Speech:
the dominant energy
component is at
about 2.5 kHz

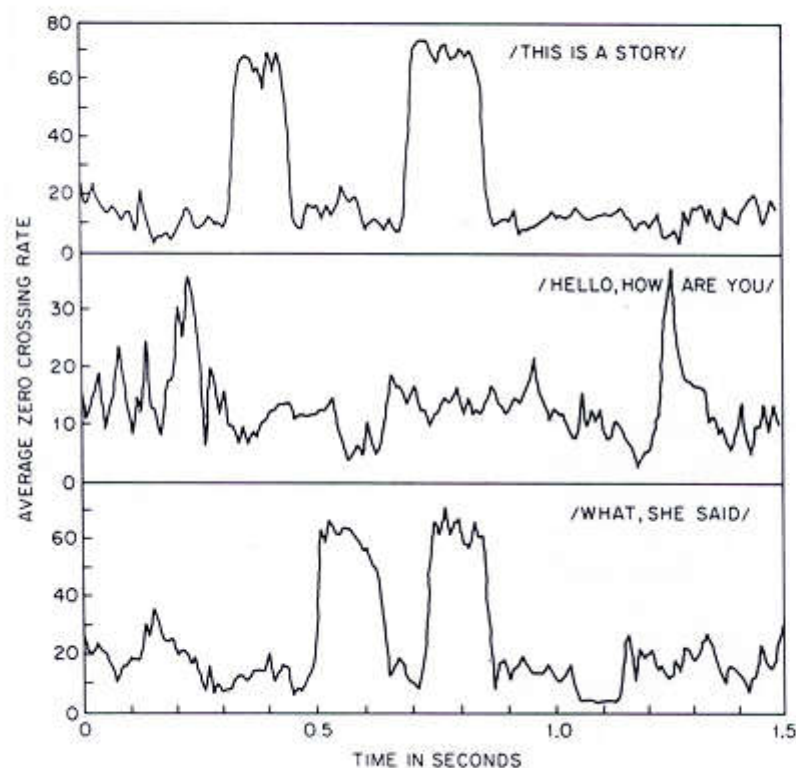
Voiced Speech: the
dominant energy
component is at
about 700 Hz

Fig. 4.11 Distribution of zero-crossings for unvoiced and voiced speech.

- for voiced speech, energy is mainly below 1.5 kHz
- for unvoiced speech, energy is mainly above 1.5 kHz
- mean ZC rate for unvoiced speech is 49 per 10 msec interval
- mean ZC rate for voiced speech is 14 per 10 msec interval



ZC Rates for Speech



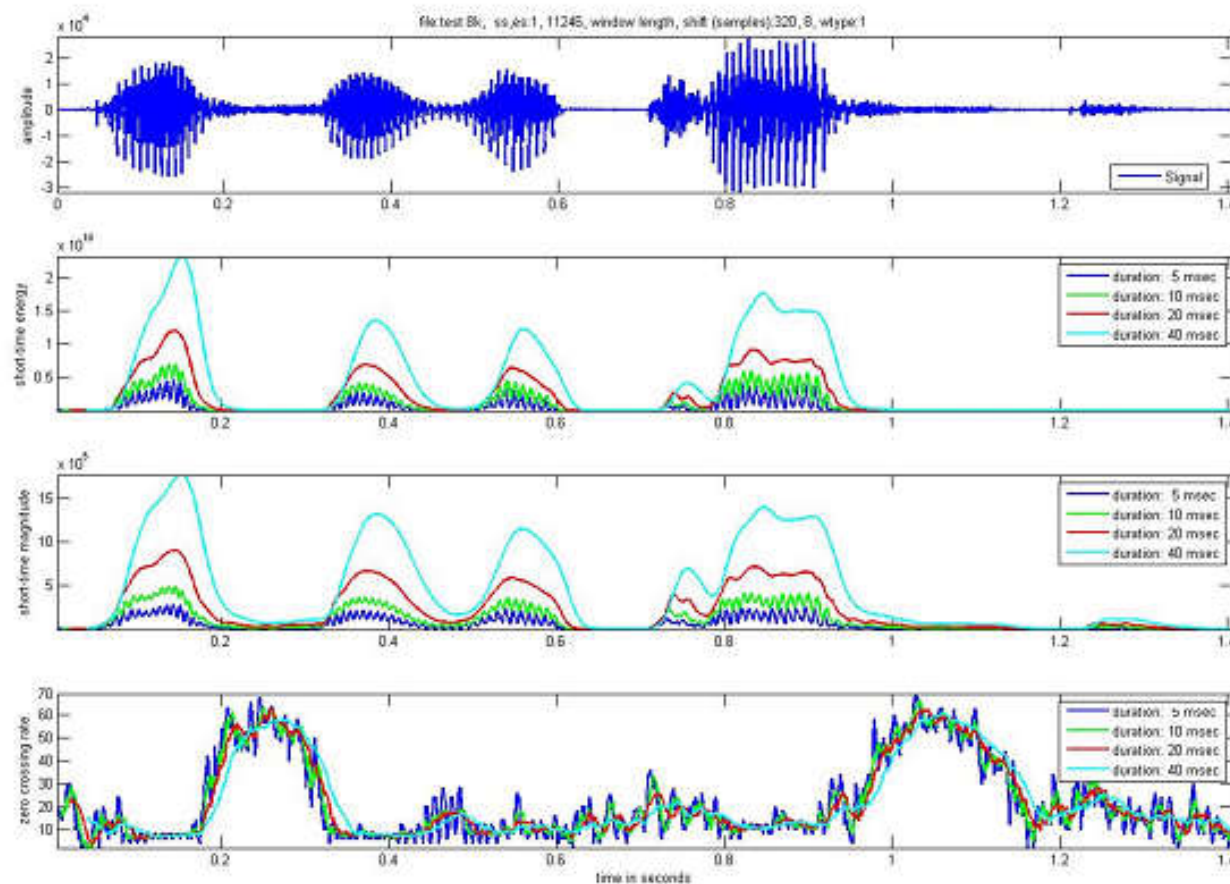
- 15 msec windows
- 100/sec sampling rate on ZC computation

Fig. 4.12 Average zero-crossing rate for three different utterances.

过零率

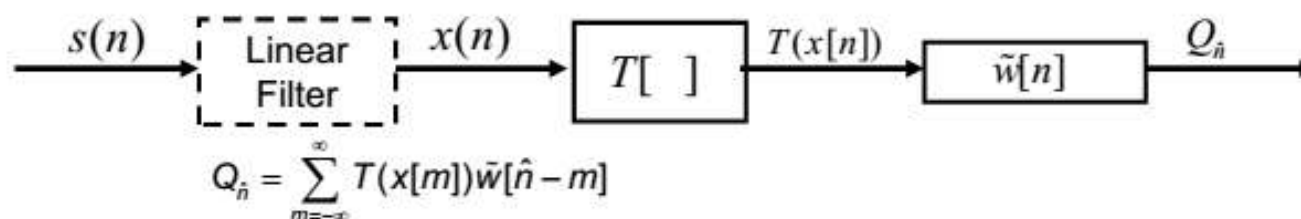


Short-Time Energy, Magnitude, ZC





Summary of Simple Time Domain Measures



1. Energy:

$$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} x^2[m]\tilde{w}[\hat{n}-m]$$

□ can downsample $E_{\hat{n}}$ at rate commensurate with window bandwidth

2. Magnitude:

$$M_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} |x[m]|\tilde{w}[\hat{n}-m]$$

3. Zero Crossing Rate:

$$Z_{\hat{n}} = z_1 = \frac{1}{2L} \sum_{m=\hat{n}-L+1}^{\hat{n}} |\text{sgn}(x[m]) - \text{sgn}(x[m-1])|\tilde{w}[\hat{n}-m]$$

where $\text{sgn}(x[m]) = 1 \quad x[m] \geq 0$
 $\quad \quad \quad = -1 \quad x[m] < 0$

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
 - Ch1_Introduction***
 - Ch2_The Speech Signal***
 - Ch4_Short-Time Analysis of Speech (Pre)***

