数字语音处理
Digital Speech Processing

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

访问码: 8P84SXUIRG2

课程引论
Course Introduction

杨莹春

yyc@zju.edu.cn

浙江大学教7-506

2021年9月17日

讲述提纲

- 语音研究简介
- 课程内容安排
- 语音技术基础

语言是衡量人类进化与智力水平高低的标志

语言是人类特有的一种交流工具,不仅承载了人类对物质世界的反映,而且促进了人类对物质世界的再创造.人类的发展离不开语言,语言的进化与人类的进化发展历程相伴随。

语言研究理由II

人类关于语言的千年梦想

IBM近日宣布，有八所大学与 **IBM** 研究人员正在合作推动沃森计算机系统背后“问答”技术的开发，该系统在**2011**年**2**月**14-16**日播出的智力竞赛节目**Jeopardy!** 《危险边缘》中与人类一较高下。麻省理工学院（**MIT**）、德州大学奥斯汀分校、南加州大学（**USC**）、伦斯勒理工学院（**RPI**）、纽约州立大学阿尔巴尼分校（**UAlbany**）、特兰托大学（意大利）、马萨诸塞大学安姆斯特分校以及卡内基梅隆大学。

语言研究理由II

人类关于语言的千年梦想

The implications of progress in A.I. are being brought into sharp relief now by the broadcasting of a recorded competition pitting the I.B.M. computing system named Watson against the two best human Jeopardy players, Ken Jennings and Brad Rutter.

The real value of Watson may ultimately be in forcing society to consider where the line between human and machine should be drawn.

语言是衡量人类进化与智力水平高低的标志

人类关于语言的千年梦想

现实社会的发展需求

DEMO

Universal Speech Interface CMU
C-Star Travel agent CMU
MonkeyBiz_2006 MIT MEDIA LAB

经典书籍

1. Xuedong Huang, Alex Acero , Hsiao-Wuen Hon，Spoken Language Processing : A Guide to Theory, Algorithm and System Development，1008 pages Prentice Hall; 1 edition (May 5, 2001)
2. Daniel Jurafsky , James H. Martin, Speech and Language Processing (2nd Edition) , 1024 pages， Pearson Prentice Hall; 2 edition (May 26, 2008)
3. Lawrence R. Rabiner, Ronald W. Schafer，Theory and Applications of Digital Speech Processing， 1056 pages， Pearson, 2010

重要会议

- ICASSP（International Conference on Acoustic, Speech and Signal Processing）

  每年一届，10月截稿，次年5月开会。

- Interspeech

  —— ICSLP（International Conference on Spoken Language Processing）

  偶数年举办，4月截稿，9月开会

  ——EuroSpeech（ European Conference on SpeechCommunication and Technology ）

  奇数年举办，4月截稿，9月开会

主要杂志

- Speech Communication
- Computer Speech and Language (CSL)
- IEEE Transactions on Audio, Speech and Language Processing (IEEE ASLP)

技术评测

- NIST Spoken Language Technology Evaluations Benchmark Tests http://www.nist.gov/speech/tests/index.htm

- 中国863语音技术评测
  http://www.863data.org.cn/2005eval.php

讲述提纲

- 语音研究简介
- 课程内容安排
- 语音技术基础

教学内容

数字语音处理

（1）语音分析技术：

主要包括语音产生基础知识、语音产生模型、时域分析技术、频域分析等；

（2）语音识别、编码与合成技术：

研究探索主题之一

语音识别与合成技术

教学安排

讲授内容

(9月17日）秋1：课程简介＋语音技术引言
(9月24日）秋2：语音分析（加实验课）
(10月8日）秋4：语音识别（加实验课）
(10月15日）秋5：语音编码及合成
(12月17日）冬6：复习及项目成果展示（加实验课）

实验内容

1. PRAAT 语音分析 （9月24日）秋2
2. VOICEBOX说话人识别 （10月8日）秋4
3、项目展示（ 12月17日）冬6

考试：2022年1月

考核方法

三重考核

（1）期末考试（40分）

期末统一闭卷考试，内容涉及语音、视频和音乐处理三块内容，每块内容比重相当。

（2）平时表现（30分）

根据同学在课程学习过程中的主动提问、讨论、测验、反馈等互动环节中的不同表现进行考核。

语音、音乐、视频3部分独立考核，每部分均为10分

（3）实验考核（30 分）

语音、音乐、视频3部分独立考核，每部分均为10分

所有未及时提交（迟交），则在得分基础上降2分；提交的报告不符合要求或报告为抄袭，相应的报告得分为“不及格”。未提交报告者给0分。

讲述提纲

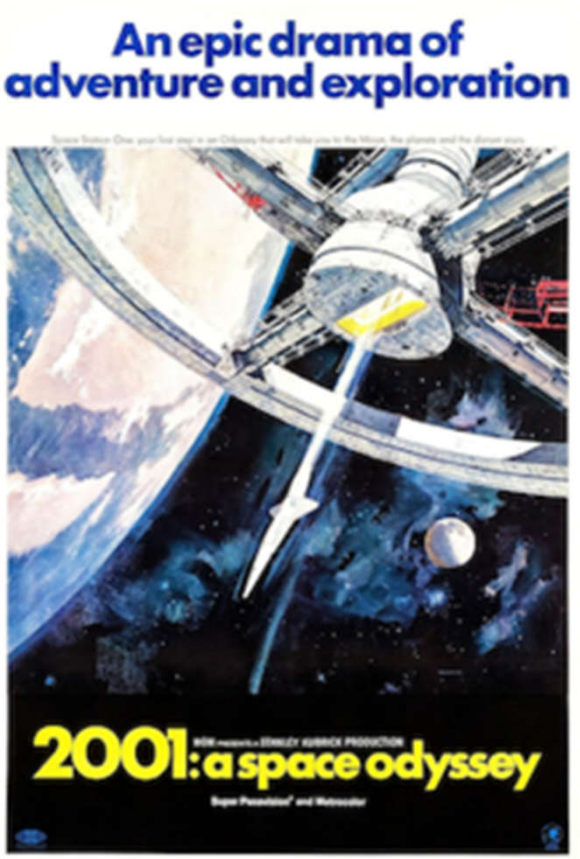
- 语音研究简介
- 课程内容安排
- 语音技术基础

语音技术基础

- 语言交际过程
- 语音产生过程

语音技术基础

https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_(film)



2001: A Space Odyssey is widely regarded as one of the greatest and most influential films ever made.[10] In 1991, it was deemed "culturally, historically, or aesthetically significant" by the United States Library of Congress and selected for preservation in the National Film Registry.[11] *Sight & Sound* magazine ranked *2001: A Space Odyssey* sixth in the top ten films of all time in its 2002[12] and 2012 critics' polls editions; it also tied for second place in the magazine's 2012 directors' poll. In 2010, it was named the greatest film of all time by *The Moving Arts Film Journal*.[13]

语言处理中的知识

语音学（**phonetics**）和音系学（**phonology**）的知识：

帮助我们建立词如何在话语中发音的模型

语言处理中的知识

形态学（**morphologic**）方面的知识：

能够产生并识别单词的这样或那样的变体，需要形态学方面的知识，这些知识能够反映关于上下文中词的形态和行为的有关信息。

语言处理中的知识

句法（**syntax**）：关于组词成句的知识。

语言处理中的知识

词汇语义学（**lexical semantics**）：
为了理解**Dave**的请求事实上是关于要求关闭分离舱门的一个命令，而不是讲关于当天中饭的菜单的事情，就要有复合词的语义知识、词汇语义学的知识。

语言处理中的知识

这种礼貌和委婉语言的用法属于语用学（**pragmatics**）的研究领域。

语言处理中的知识

正确地把这样的会话组织成结构，需要话语规约（**discourse convention**）的知识。

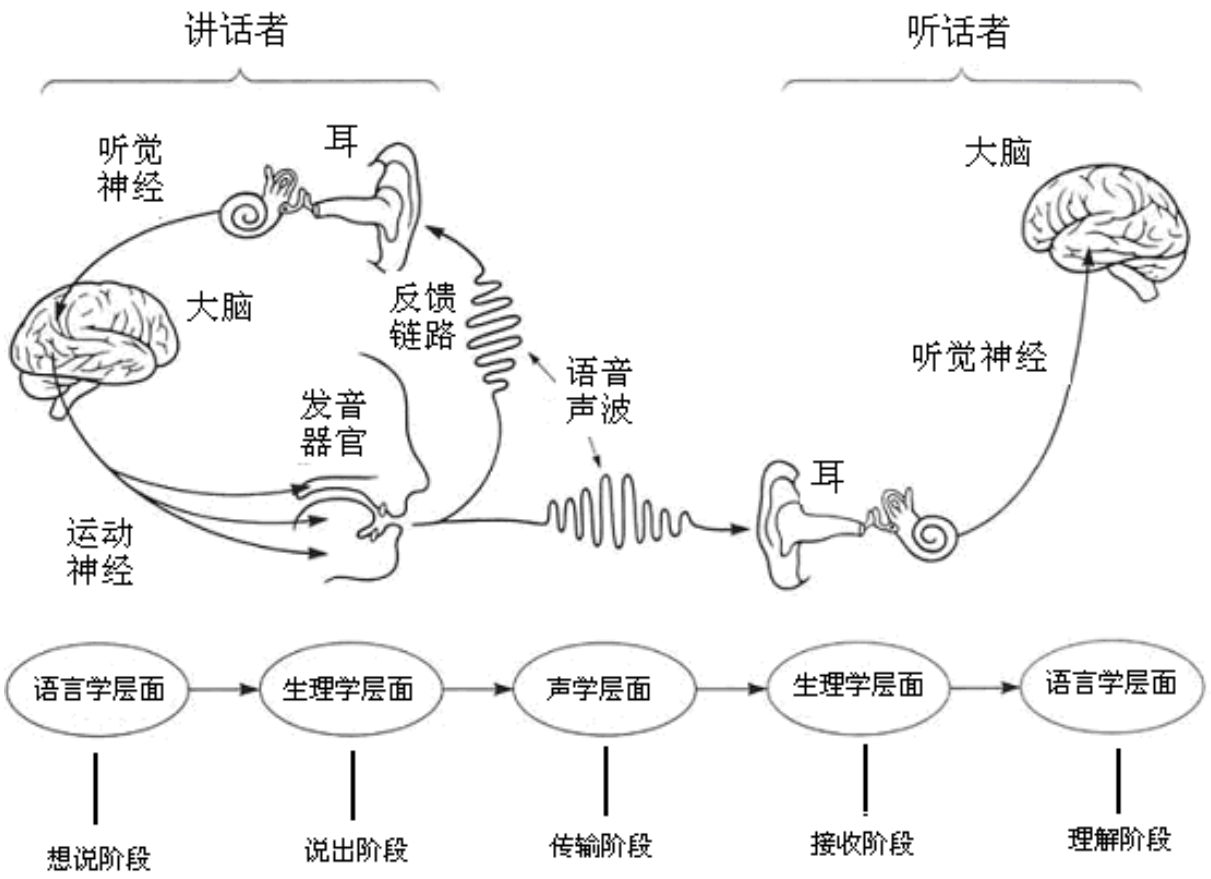
语言处理中的知识

- 语音学与音系学—研究语言的语音
- 形态学—研究词的有意义的组合
- 句法学—研究词与词之间的结构关系
- 语义学—研究意义
- 语用学—研究如何用语言来达成一定的目的
- 话语学—研究大于段的语言单位

模型和算法

- 状态机（**state machine**）
  - 包括状态、状态之间的转移、输入表示等
- 形式规则系统（**formal rule system**）
  - 正则语法、正则关系、上下文无关语法
- 逻辑（**logic**）
  - 逻辑表达方法是处理语义学、语用学和话语分析等方面知识的选择工具
- 概率论（**probability theory**）
  - 其他的各种模型都可以使用概率得到进一步提高
  - 也是一种机器学习（**machine learning**）的模型

语音链


讲话者
听话者
听觉神经
耳
大脑
反馈链路
语音声波
发音器官
运动神经
大脑
听觉神经
耳
语言学层面
生理学层面
声学层面
生理学层面
语言学层面
想说阶段
说出阶段
传输阶段
接收阶段
理解阶段

语音链

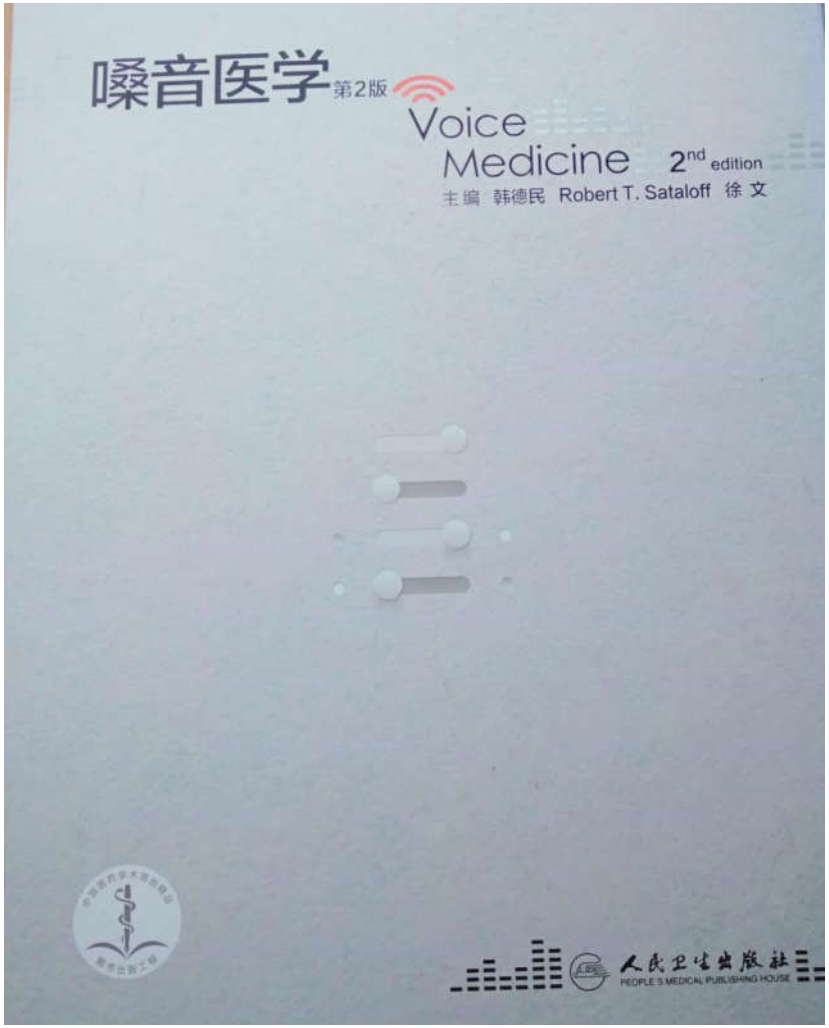
- “发音-传递-感知 ”三阶段

- **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

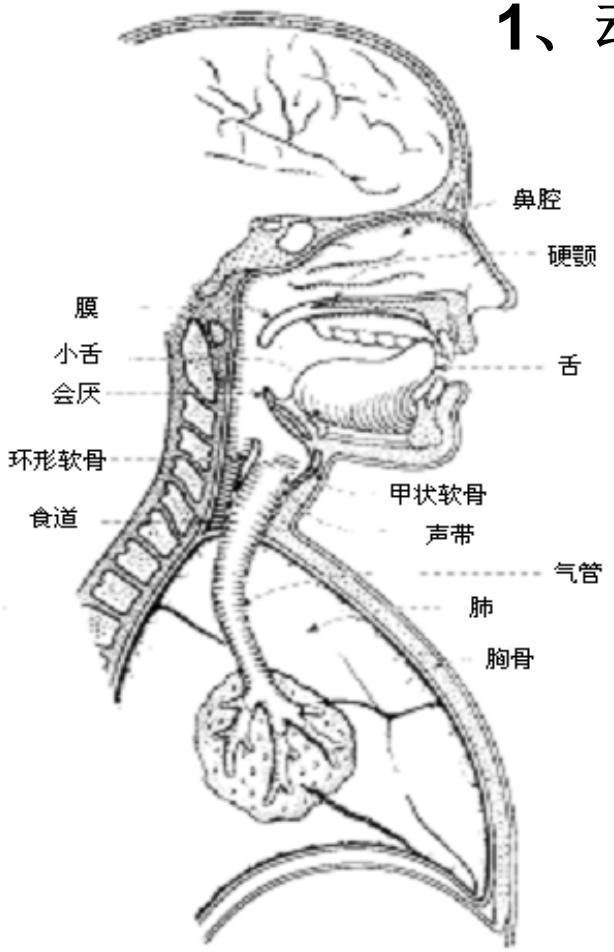
- 语言交际过程
- 语音产生过程

语音技术基础


嗓音医学 第2版
Voice Medicine 2nd edition
主编 韩德民 Robert T. Sataloff 徐 文
人民卫生出版社
PEOPLE'S MEDICAL PUBLISHING HOUSE

语音产生原理

1、动力源　2、发音体　3、共鸣器



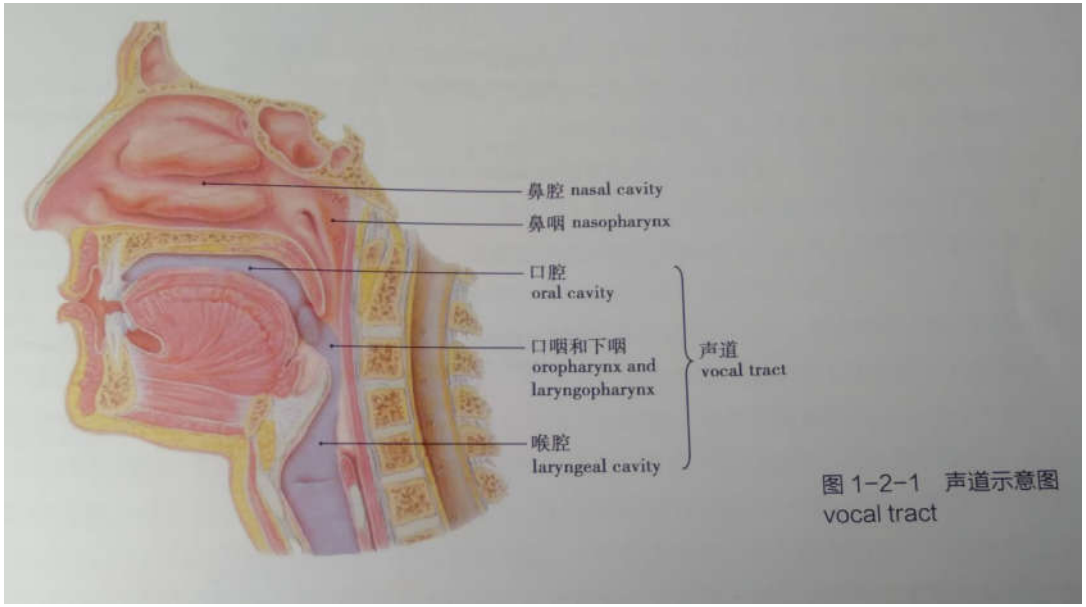


图 1-2-1　声道示意图
vocal tract

语音产生原理

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 标志	安静呼吸	言语呼吸	歌唱呼吸
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。

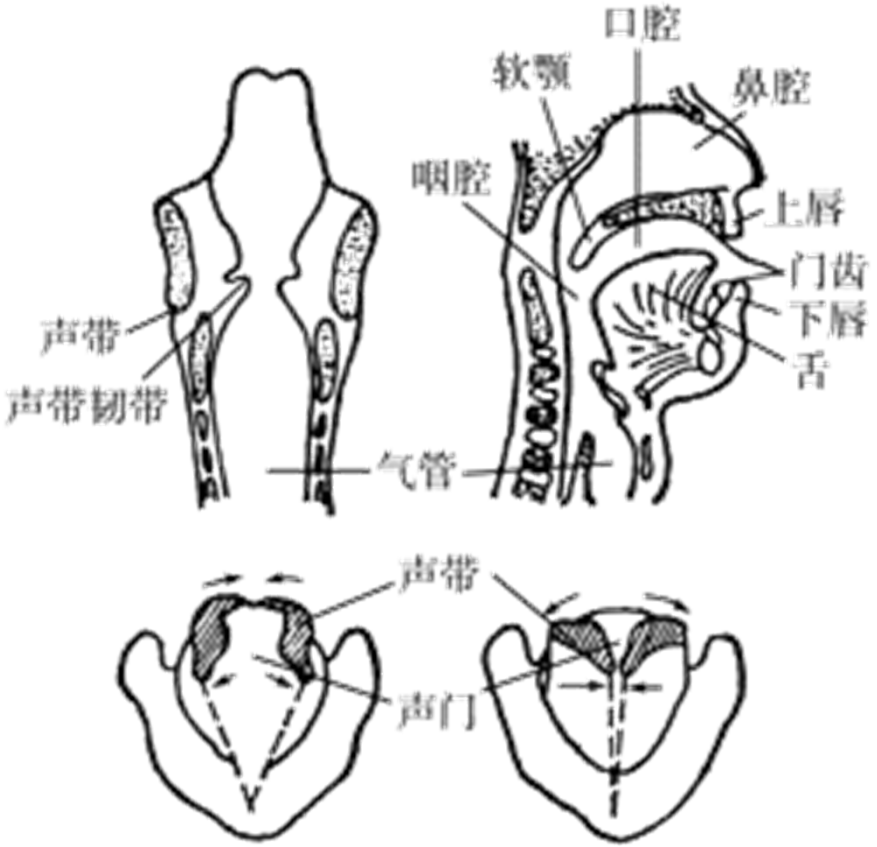
内肌包括环甲肌、杓间肌、环杓侧肌、甲杓肌。

（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

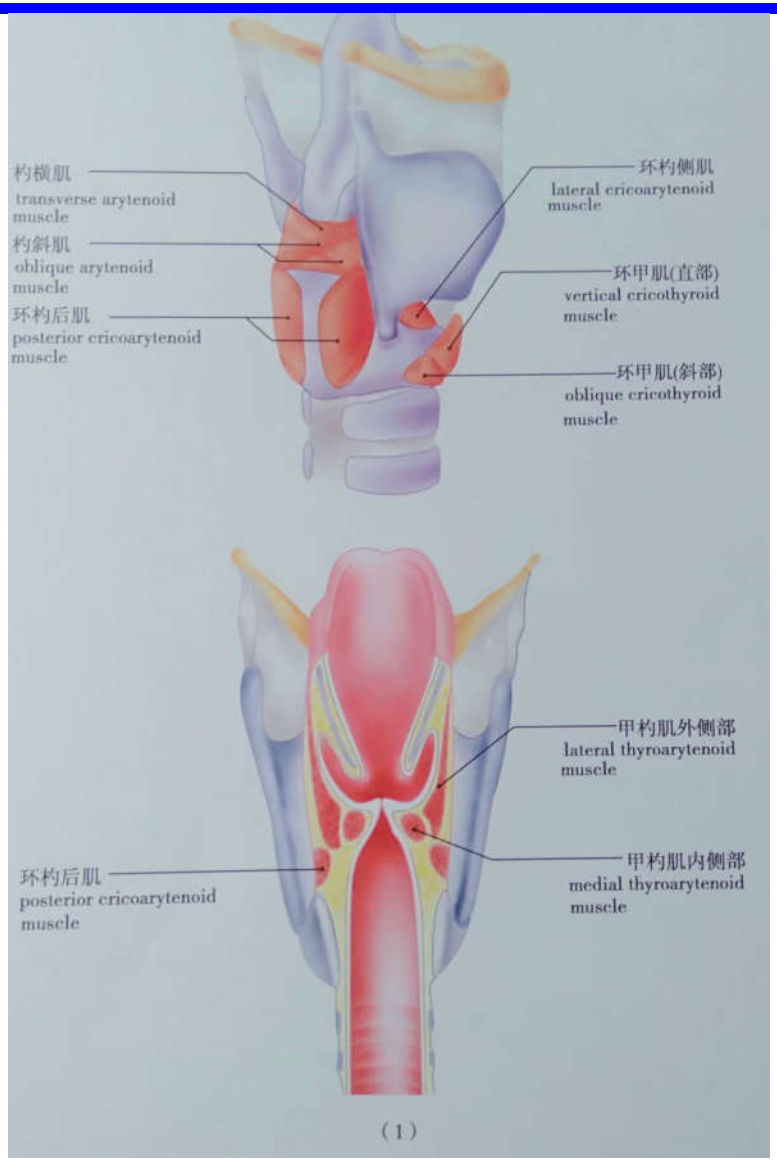
- 声带构造
- 声门：音声门、气声门
- 声门状态---发声类型
- **声带振动原理：伯努利效应**
- 声带运动摄像

  正常　异常



石锋，语音学原理，南开大学，2007

语音产生原理


杓横肌
transverse arytenoid muscle
杓斜肌
oblique arytenoid muscle
环杓后肌
posterior cricoarytenoid muscle
环杓侧肌
lateral cricoarytenoid muscle
环甲肌(直部)
vertical cricothyroid muscle
环甲肌(斜部)
oblique cricothyroid muscle
甲杓肌外侧部
lateral thyroarytenoid muscle
甲杓肌内侧部
medial thyroarytenoid muscle
环杓后肌
posterior cricoarytenoid muscle
(1)

语音产生原理

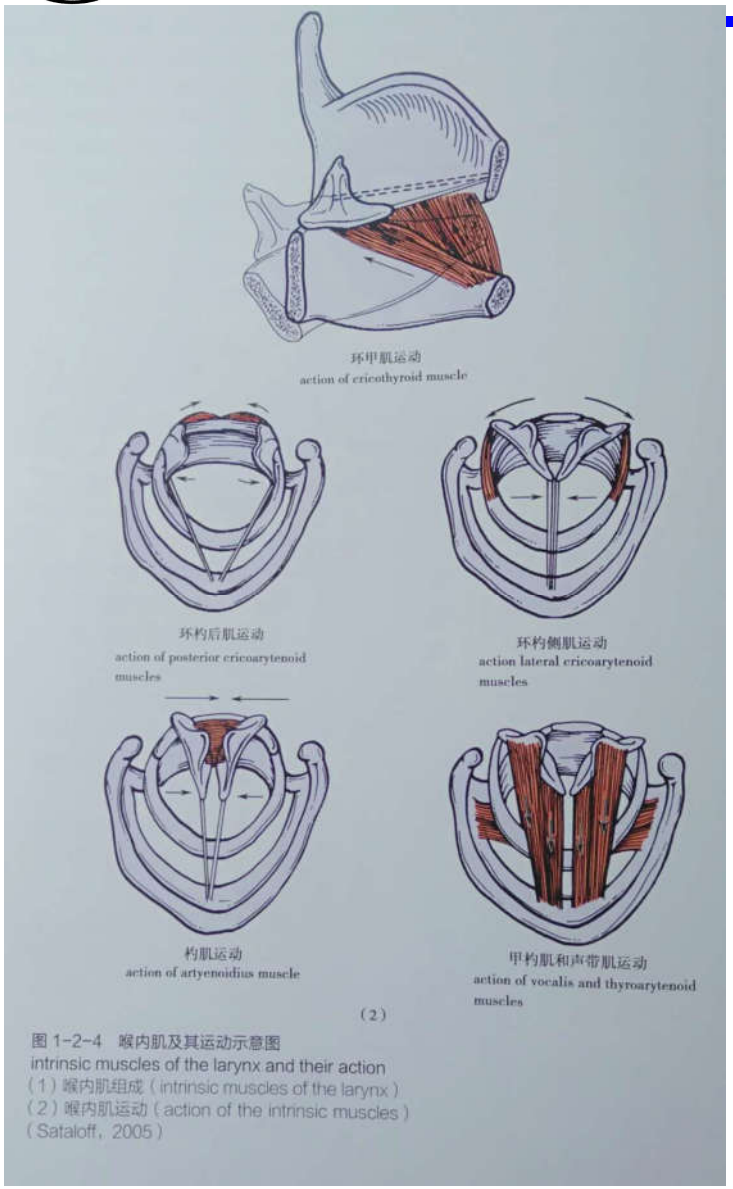


图 1-2-4 喉内肌及其运动示意图
intrinsic muscles of the larynx and their action
(1) 喉内肌组成 (intrinsic muscles of the larynx)
(2) 喉内肌运动 (action of the intrinsic muscles)
(Sataloff, 2005)

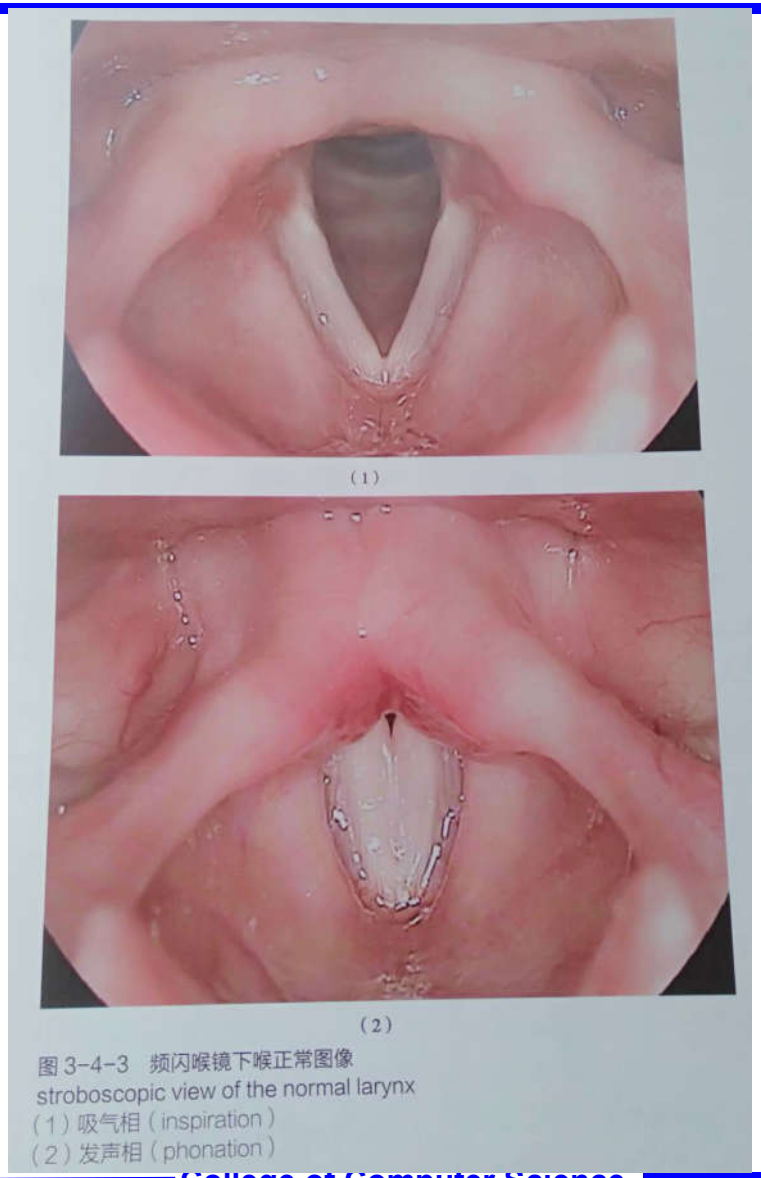
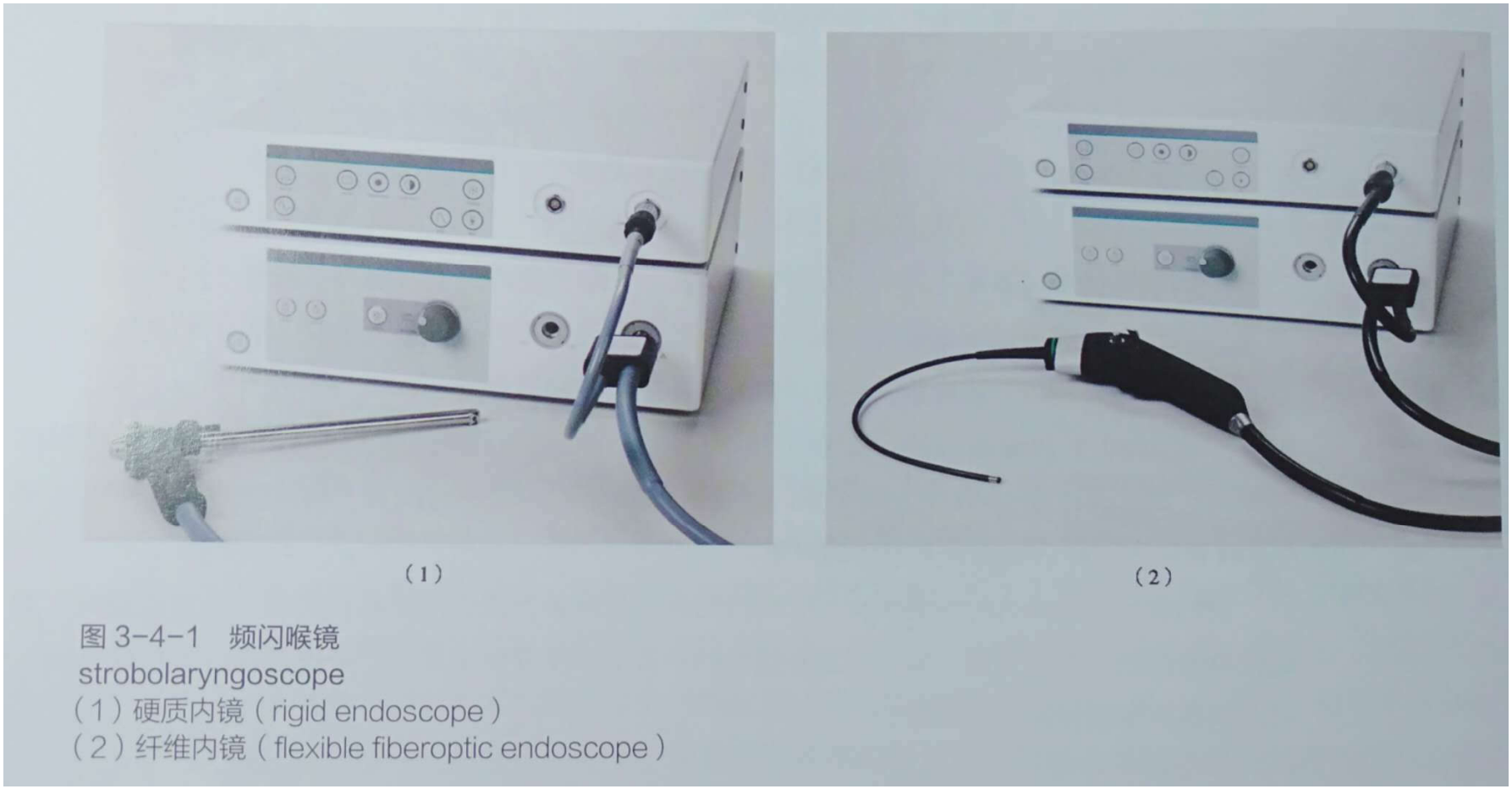


图 3-4-3 频闪喉镜下喉正常图像
stroboscopic view of the normal larynx
(1) 吸气相 (inspiration)
(2) 发声相 (phonation)

语音产生原理



（1）　　（2）

图 3-4-1　频闪喉镜
strobolaryngoscope
（1）硬质内镜（rigid endoscope）
（2）纤维内镜（flexible fiberoptic endoscope）

语音产生原理

图 3-4-7　一个声门周期的蒙太奇图像
photo-montage of one glottal cycle
一个声门周期中声带的开放和闭合相，双侧声带对称（one glottal cycle has been assembled as a photo-montage of one glottal cycle. note the open and closed phase, the symmetry of vocal fold between the right and the left and the opening and closing phase are all equal between the two sides）

图 3-4-8　男性以频率 127Hz、声强 71dB 轻声发音时的喉高速摄影图像
the high-speed videoendoscopy images of male soft phonation in soft modal voicewith the frequency of 127Hz at 71dB
两侧声带对称开放和闭合，开放相约占声门周期的 60%（note the symmetric opening and closing of each vocal fold. The opening phase is about 60% of the glottal cycle）

图 3-4-9　男性以频率 127Hz、声强 82dB 大声发音时的喉高速摄影图像
the high-speed videoendoscopy images of the production at loud voice register at the frequency 127Hz and the amplitude 82dB by a male
图中声带比图 3-4-8 中声带短而厚，声带越短开放相越短（note the shorter vocal folds with short open phase. The vocal folds are also shorter and thicker than on figure 3-4-8）

图 3-4-10　男性以 400Hz、74dB 发假声时的喉高速摄影图像
the high-speed videoendoscopy images of the male production of falsetto voice of at 400Hz and at 74dB by a male
声带长而薄，膜部几乎没有接触（note the long thin vocal folds with very little membranous vocal fold approximation）

语音产生原理

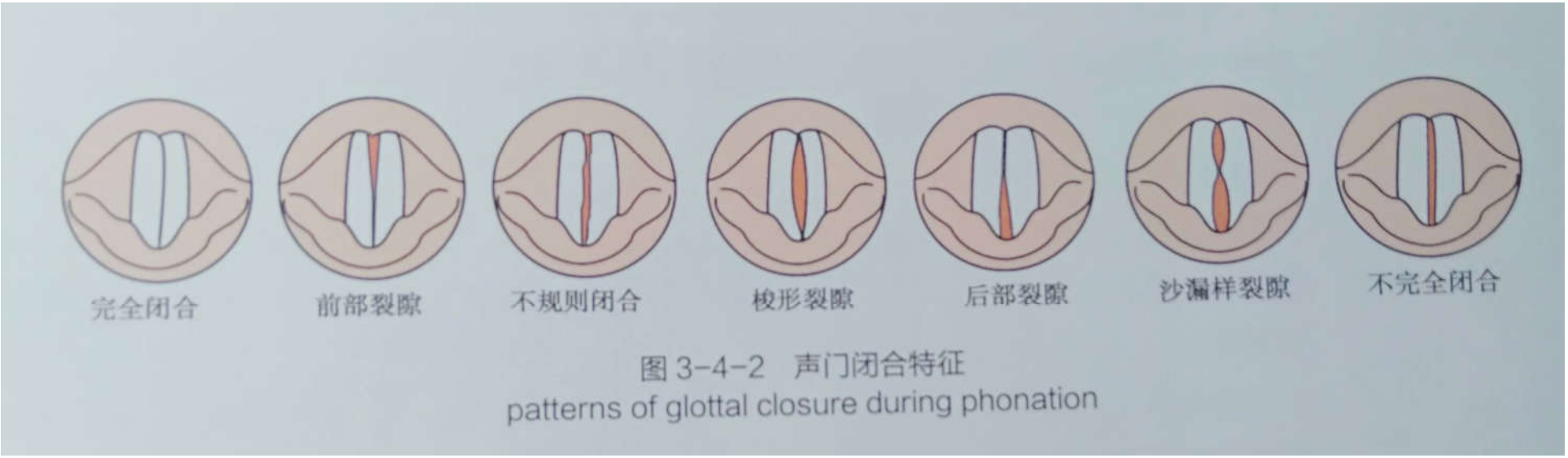


图 3-4-2　声门闭合特征

patterns of glottal closure during phonation

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

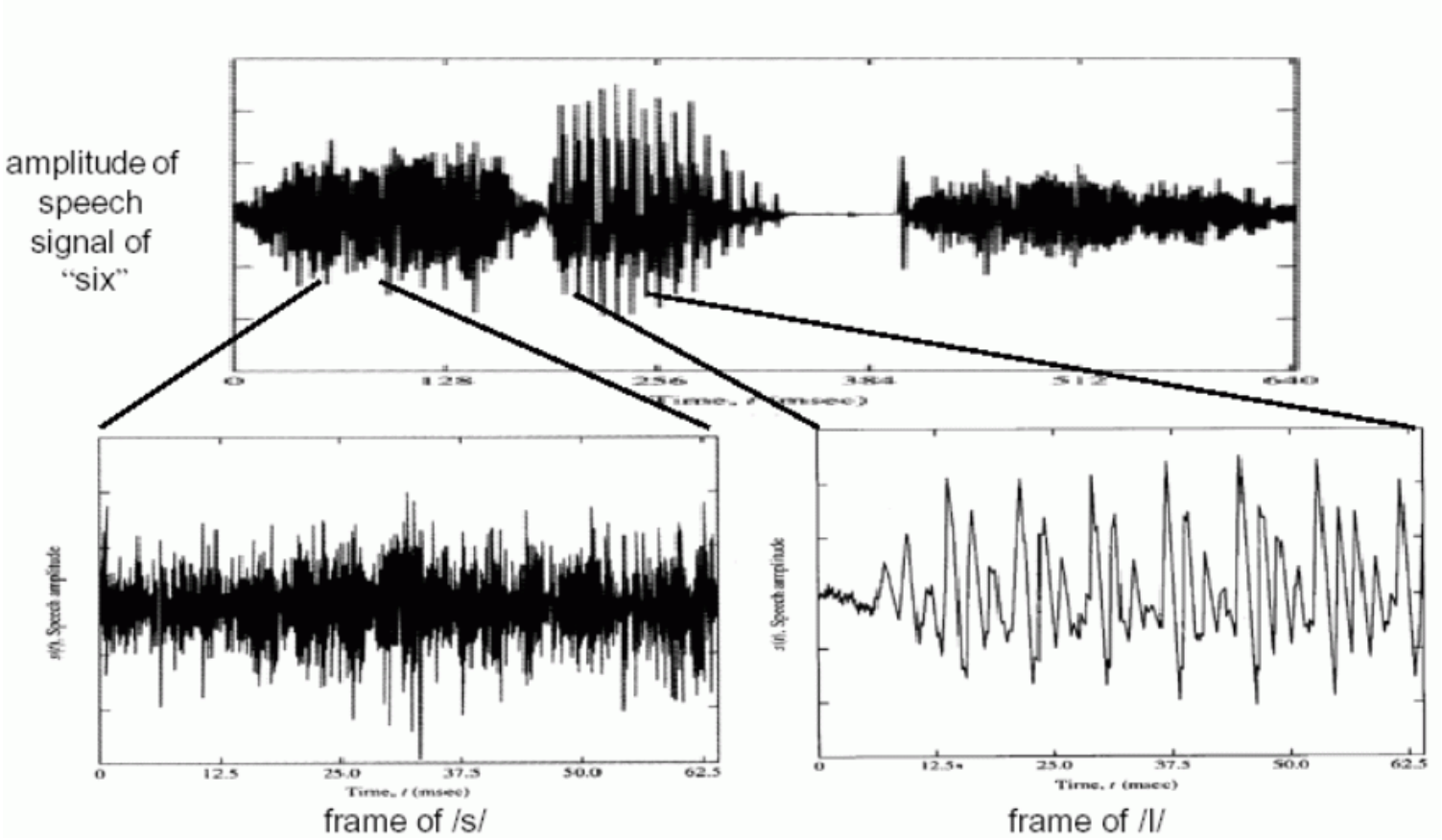
语音产生原理

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）。

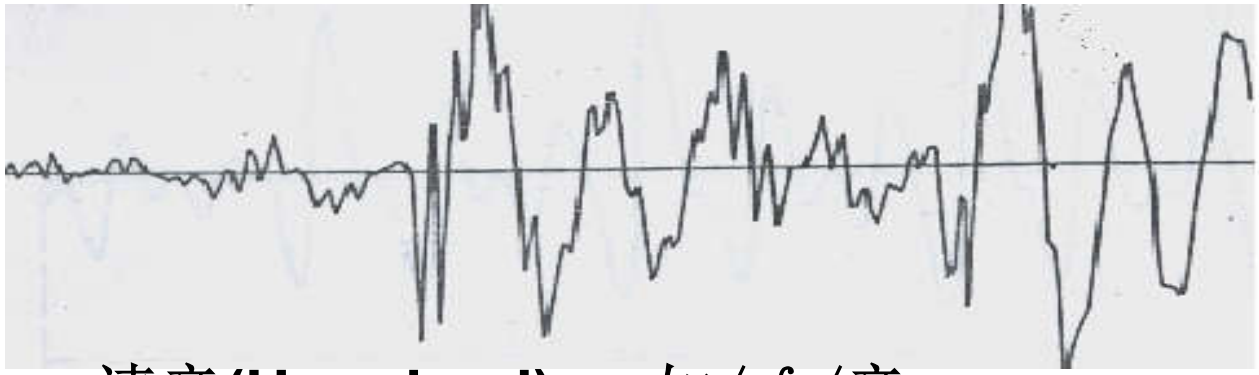
肺部压缩空气的力量大小，决定音量（响度）大小。

语音产生原理


amplitude of speech signal of "six"
0
128
256
384
512
640
Time, t (msec)
frame of /s/
frame of /l/
0
12.5
25.0
37.5
50.0
62.5
Time, t (msec)

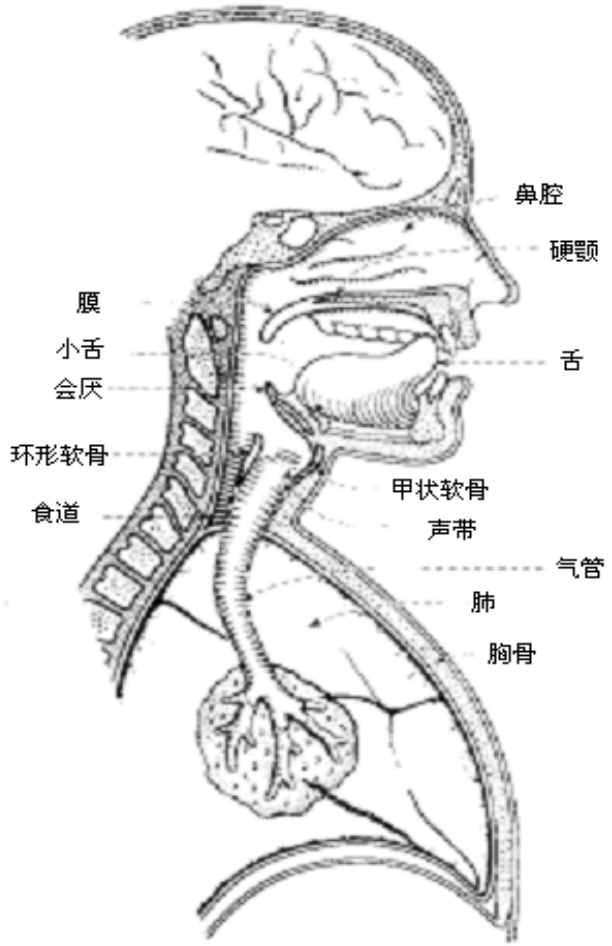
语音产生原理

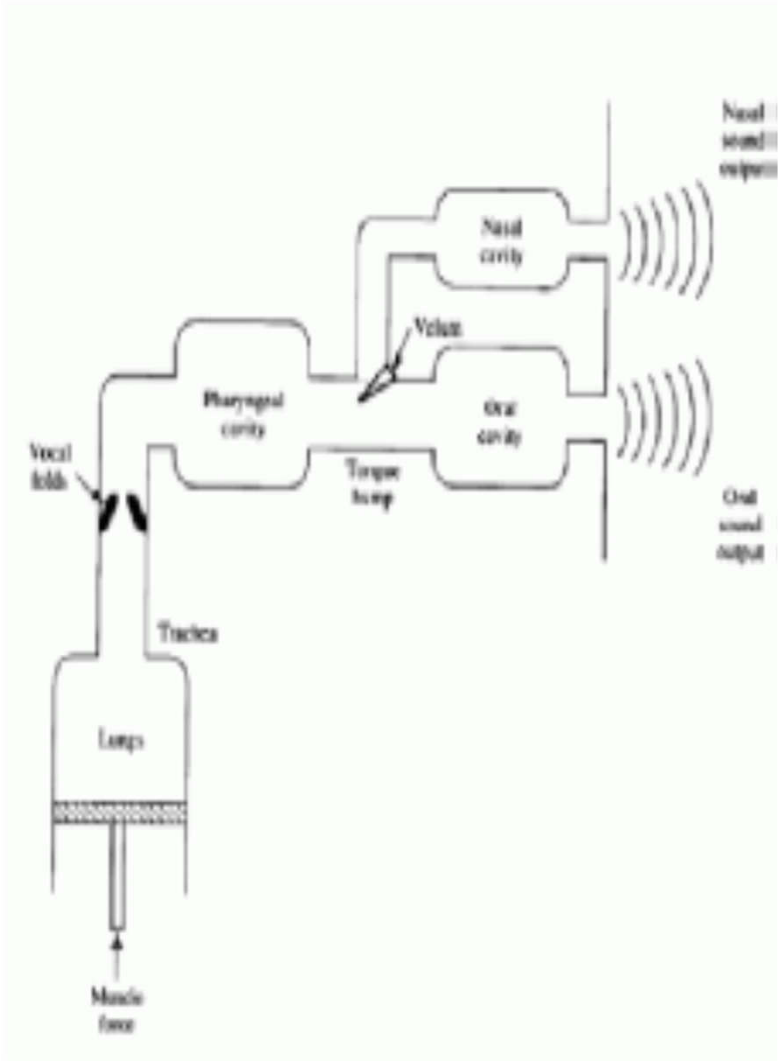
浊音 (Voiced)：如/U/、/d/、/i/等音



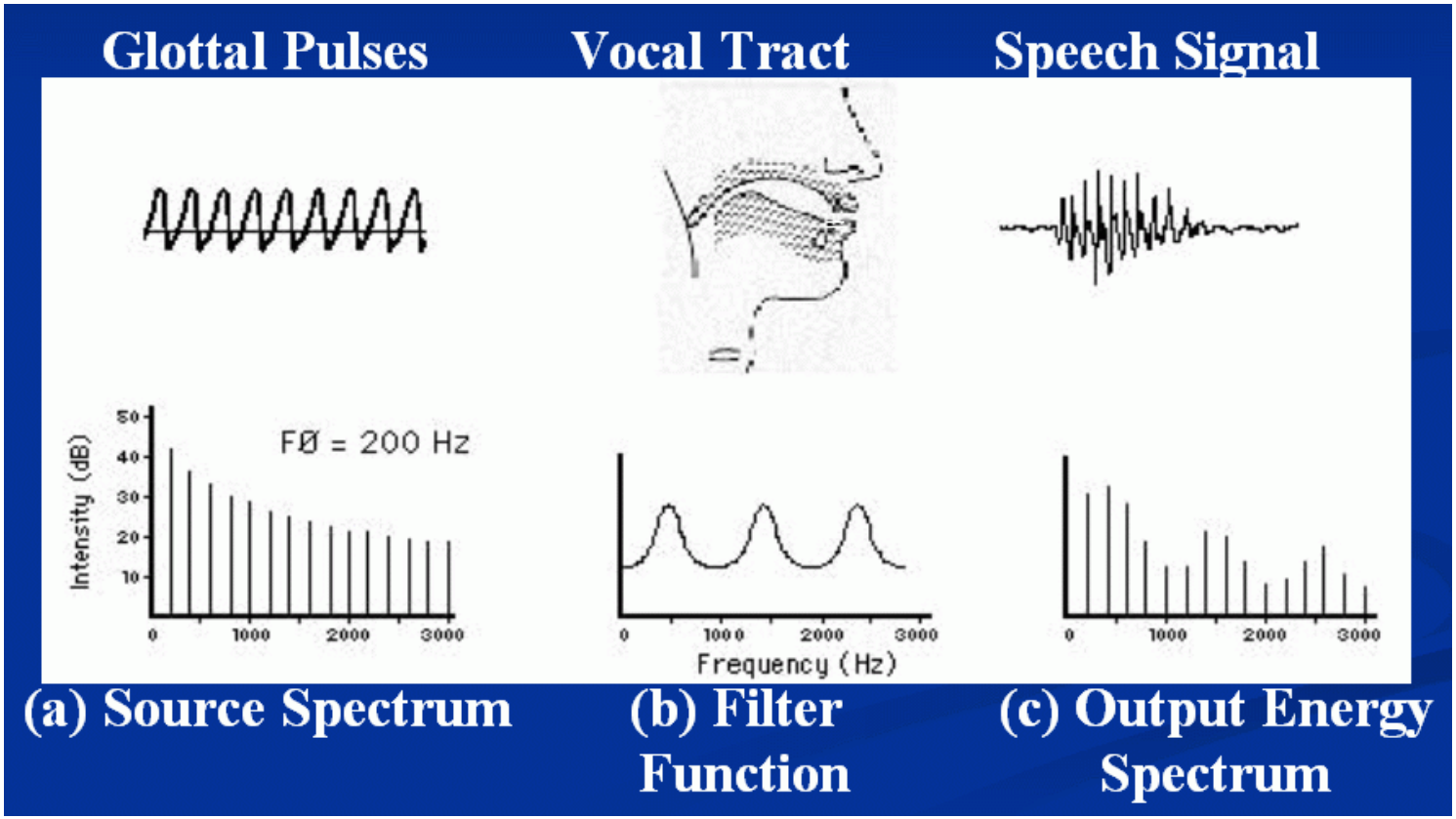
清音(Unvoiced) ：如/ʃ/音

语音产生模型


鼻腔
硬颚
膜
小舌
会厌
舌
环形软骨
甲状软骨
食道
声带
气管
肺
胸骨


Nasal cavity
Velum
Pharyngeal cavity
Oral cavity
Vocal folds
Tongue hump
Trachea
Lungs
Muscle force

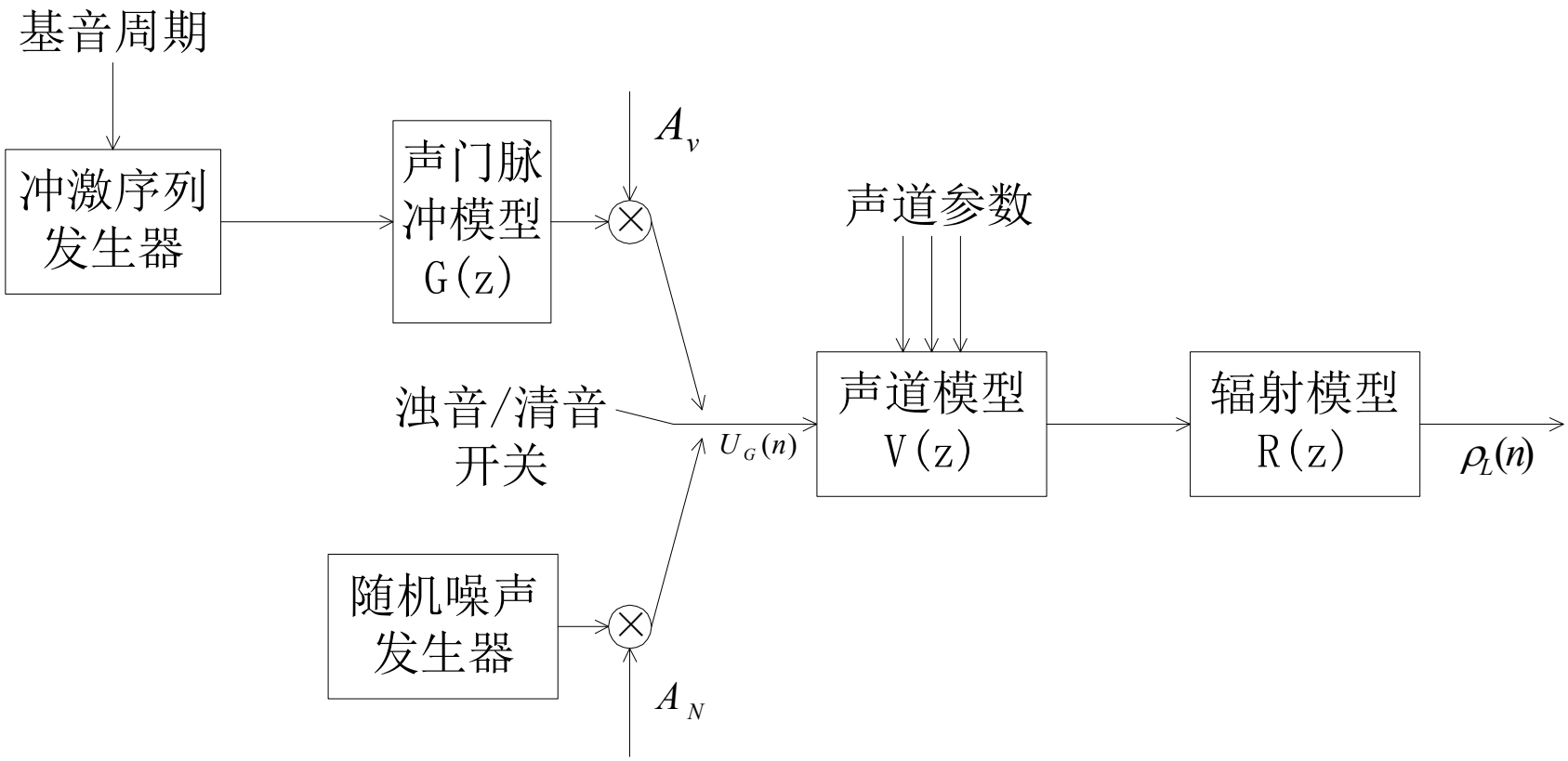
语音产生模型



Source-filter model and the corresponding spectrum

语音产生模型

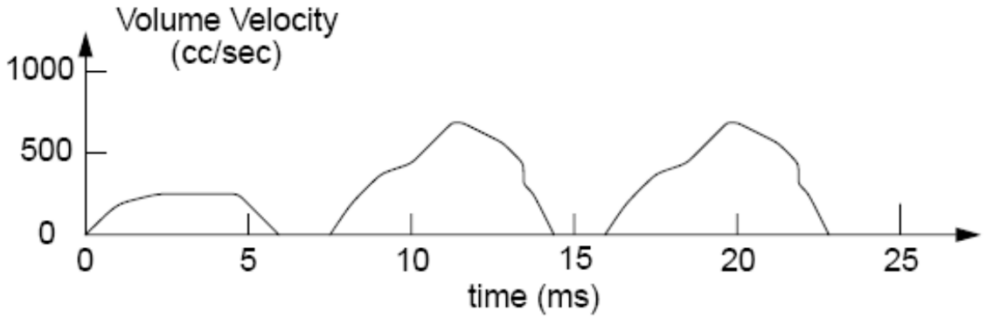
语音信号产生的数字模型



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

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

单个斜三角形脉冲频域表示为低通滤波器

$$G(z)=\frac{1}{(1-g_1z^{-1})(1-g_2z^{-1})}$$

g_1, g_2接近1

语音产生模型——激励模型

单个斜三角形脉冲频域表示为低通滤波器

$$G(z)=\frac{1}{(1-g_1z^{-1})(1-g_2z^{-1})}$$

g_1, g_2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

$$U(z)=G(z)E(z)=\frac{A_v}{1-z^{-1}}\times\frac{1}{(1-g_1z^{-1})(1-g_2z^{-1})}$$

语音产生模型——激励模型

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

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

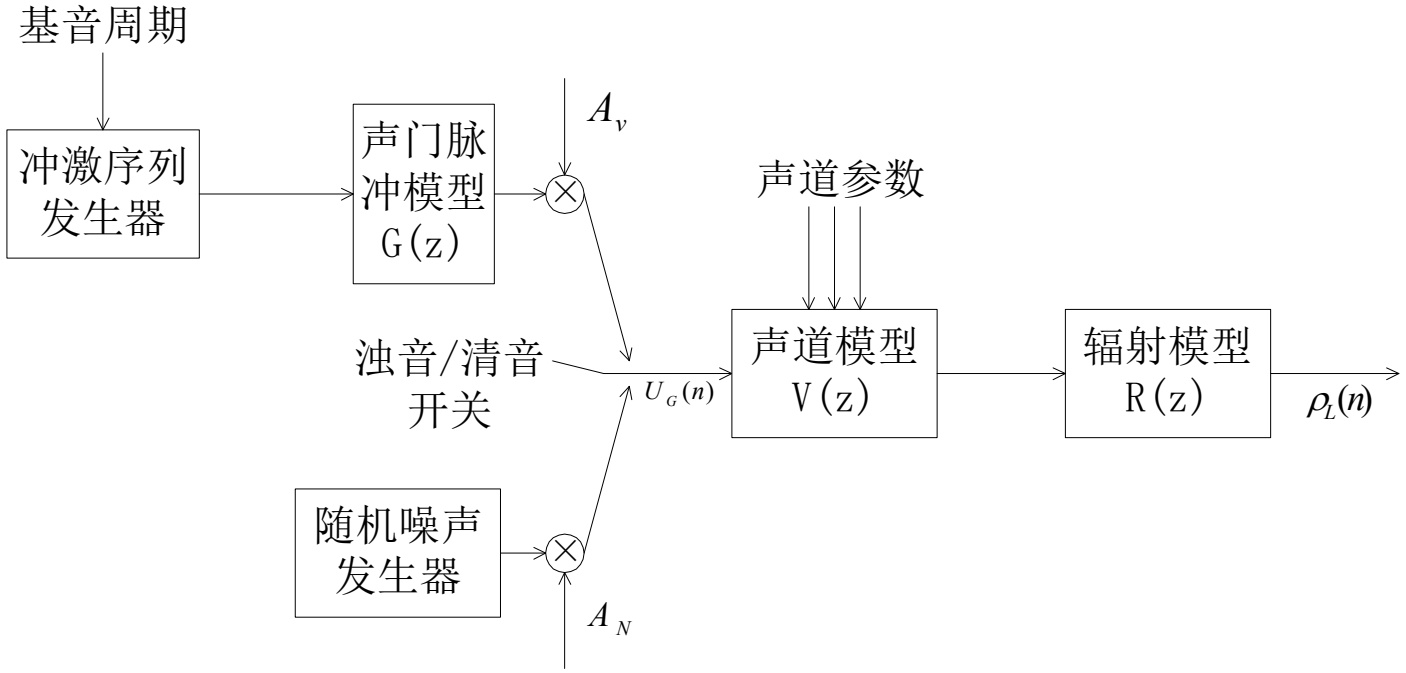
语音产生模型——声道模型

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：

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

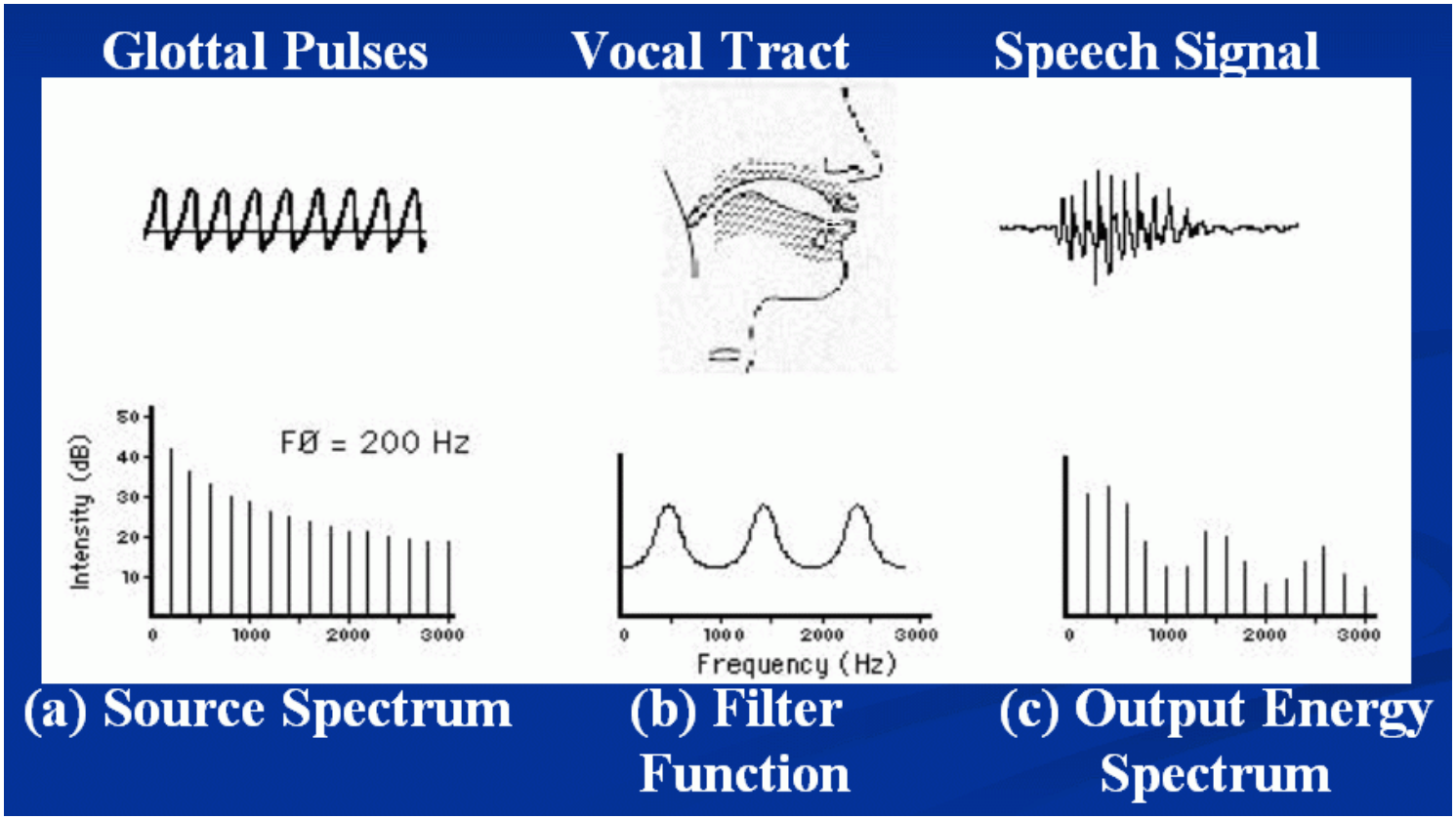
其中，　　r接近1

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

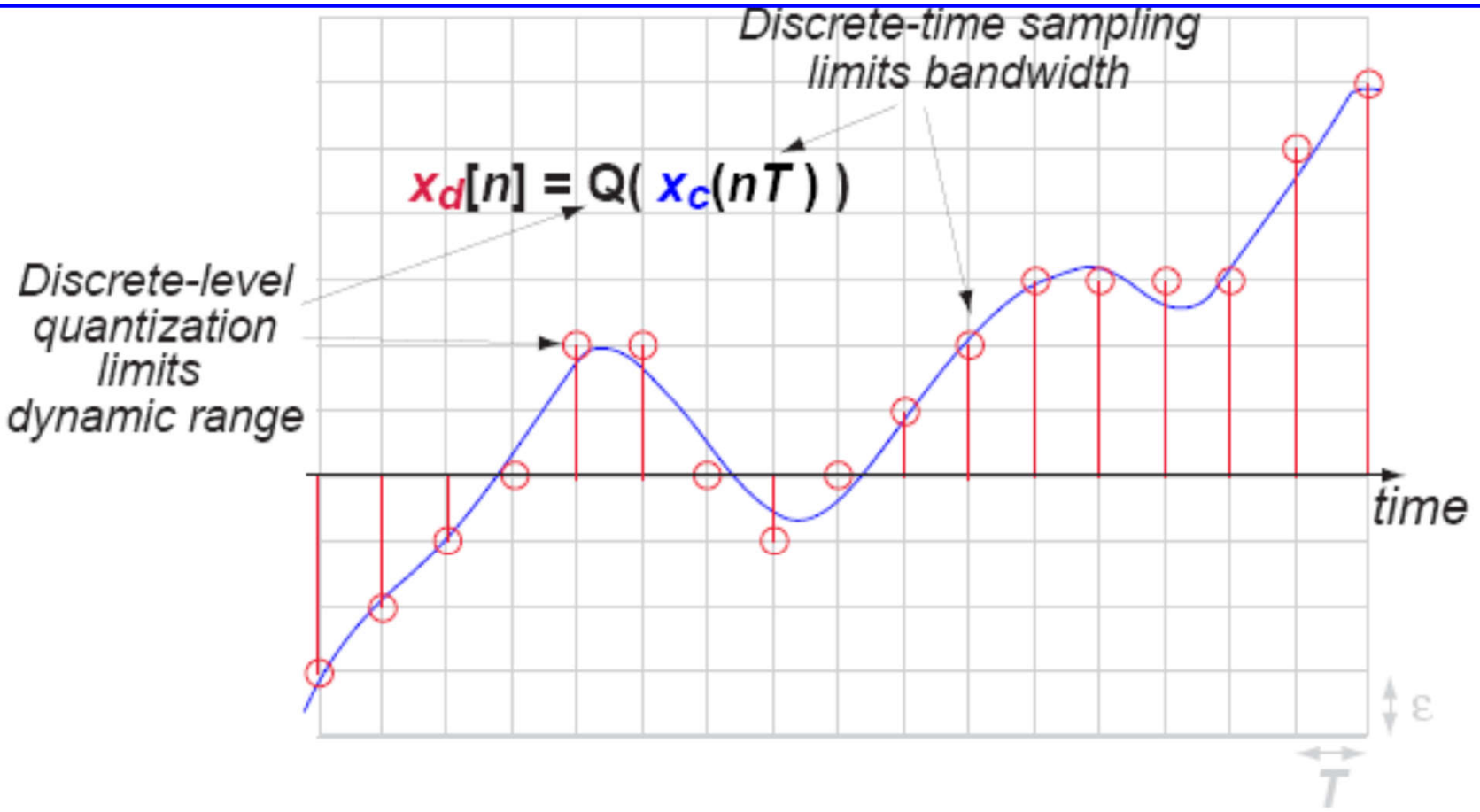


Source-filter model and the corresponding spectrum

语音技术基础

- 信号处理基础

信号处理——采样


Discrete-time sampling
limits bandwidth
$x_d[n] = Q(x_c(nT))$
Discrete-level
quantization
limits
dynamic range
time
ε
T

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 $t = nT$,

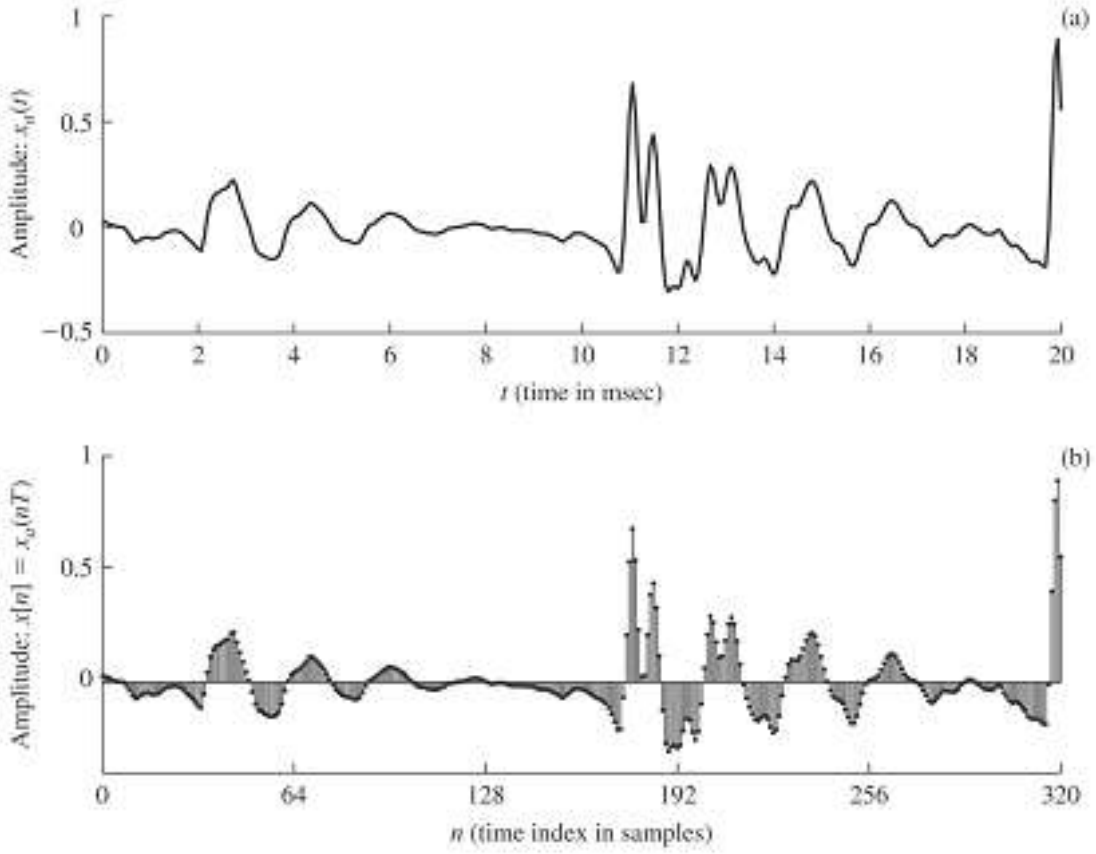
$$x[n] = x_a(nT), \quad -\infty < n < \infty$$

- 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

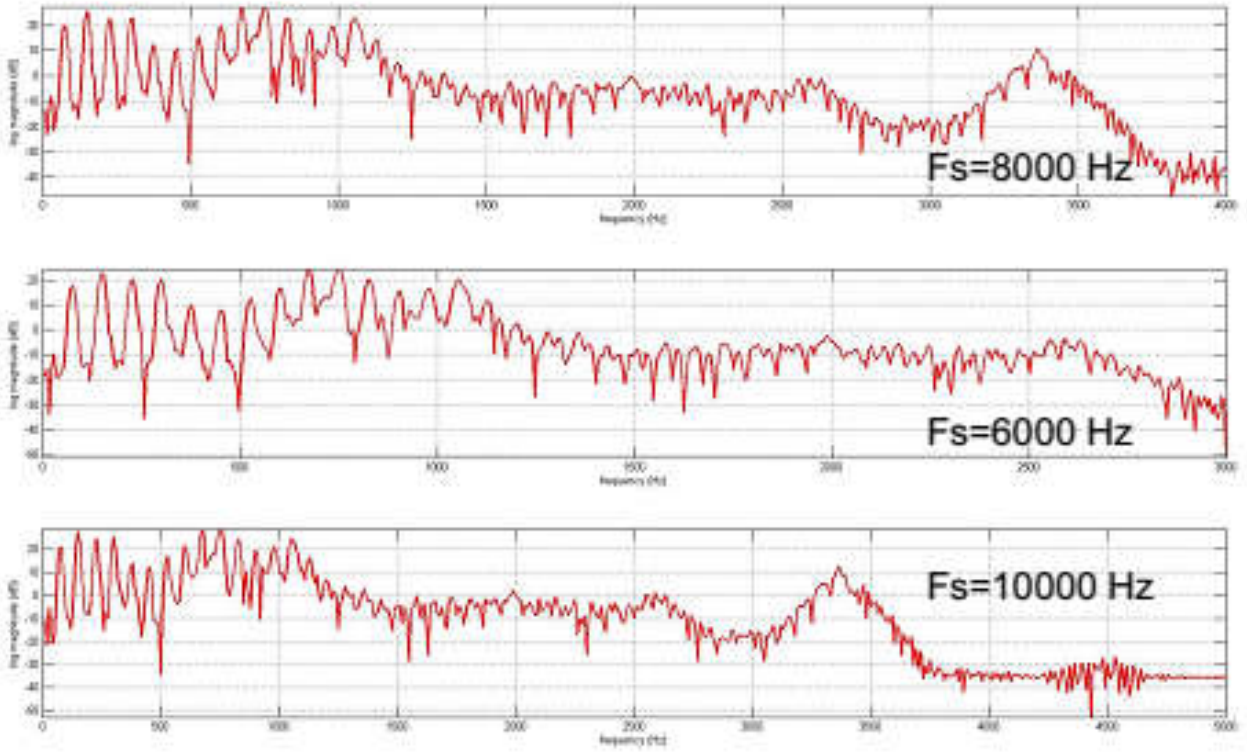
Speech Waveform Display



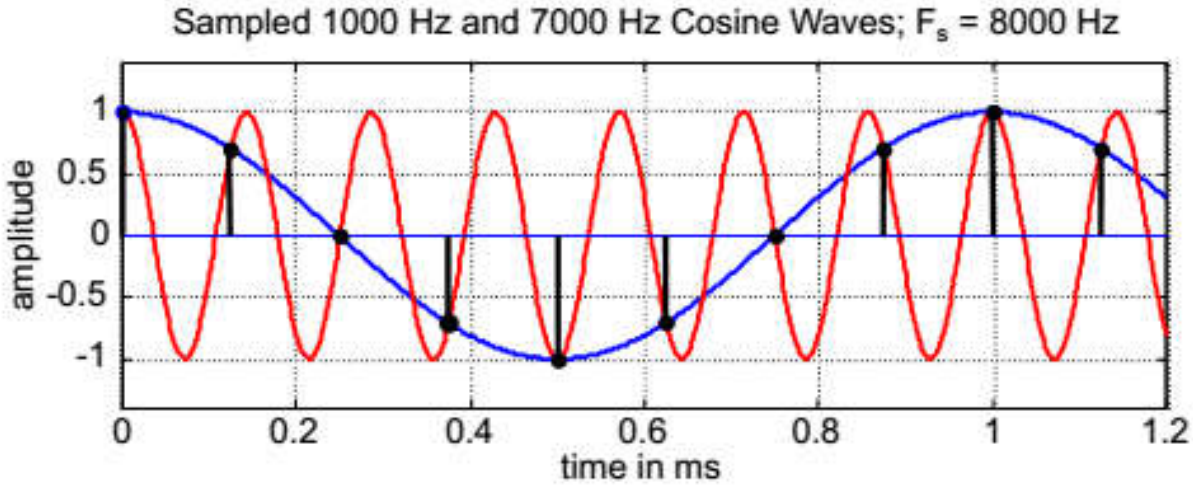
plot( );

stem( );

Varying Sampling Rates



The Sampling Theorem

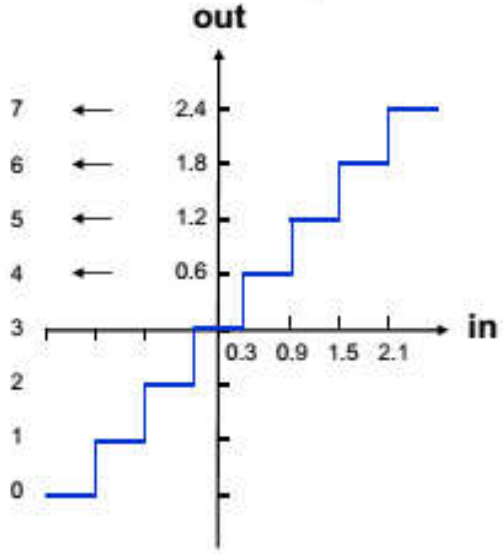


- A bandlimited signal can be reconstructed exactly from samples taken with sampling frequency

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

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).

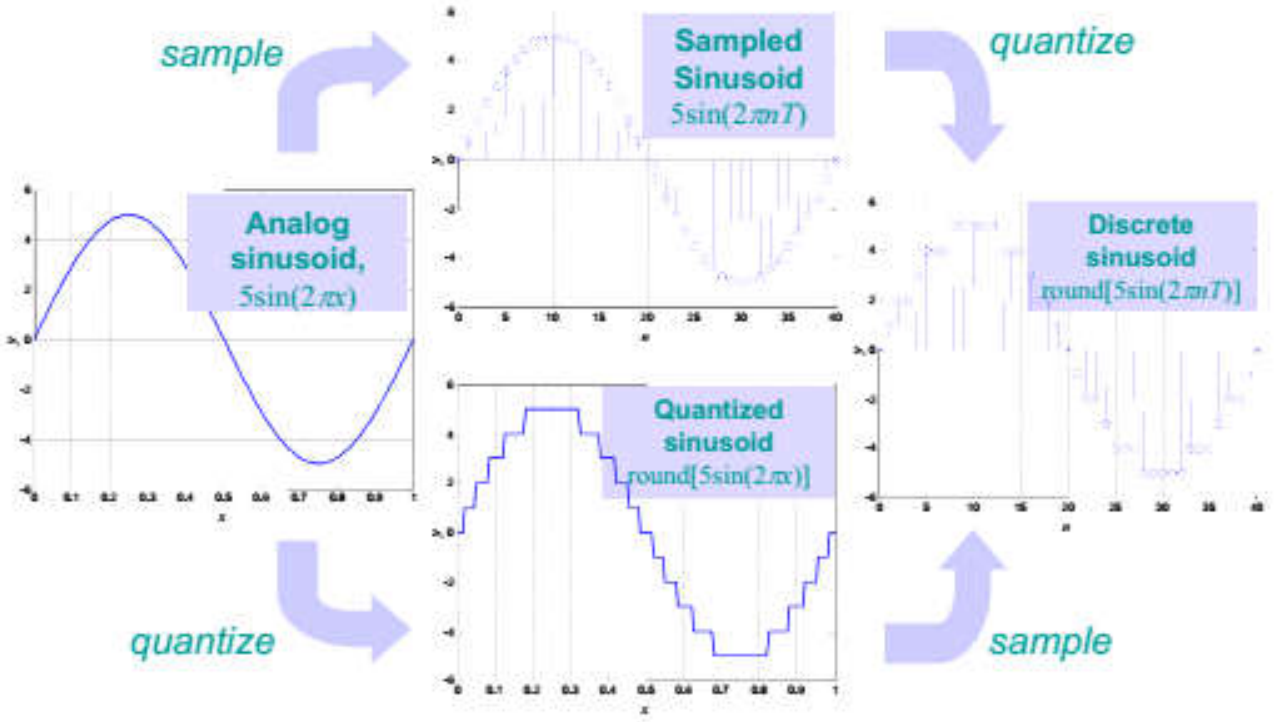


A 3-bit uniform quantizer

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

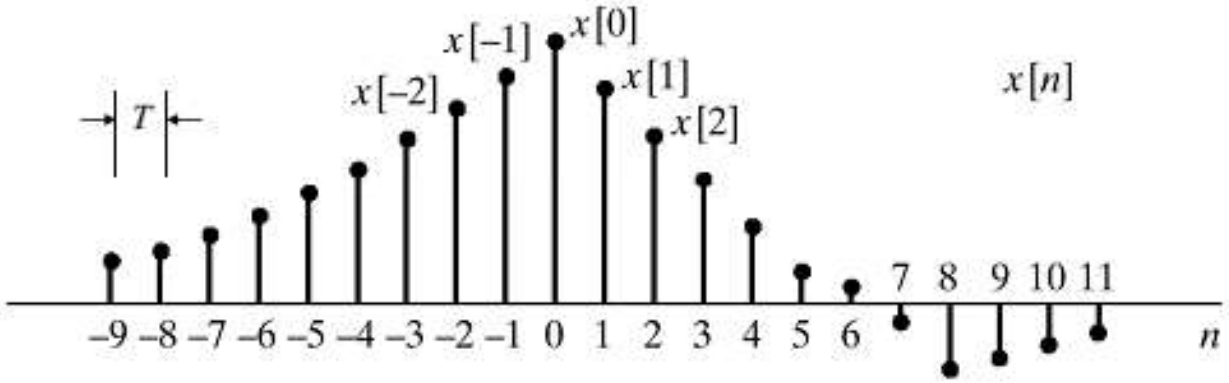
Discrete Signals



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



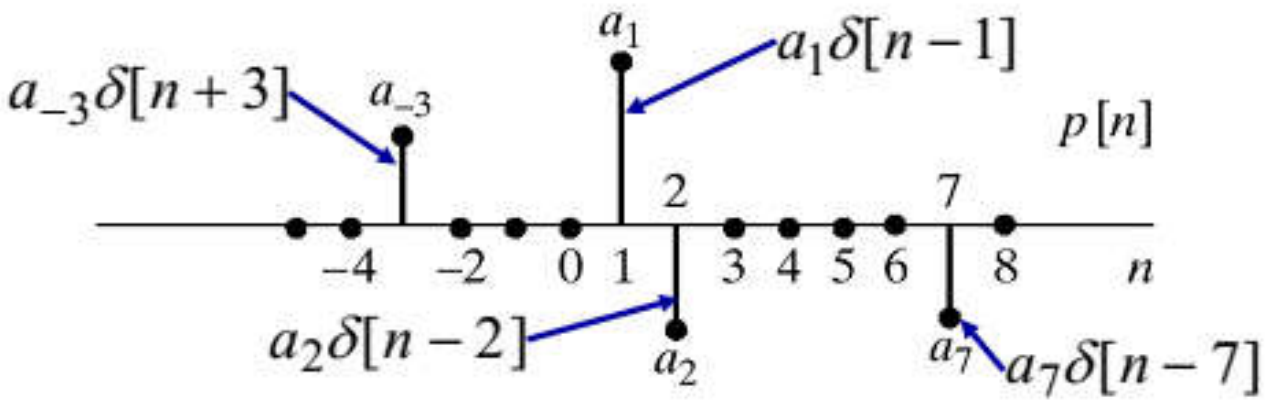
- $x[n]$ denotes the "sequence value at 'time' n"
- Sources of sequences:
  - Sampling a continuous-time signal
    $x[n] = x_c(nT) = x_c(t)|_{t=nT}$
  - Mathematical formulas – generative system
    e.g., $x[n] = 0.3 \cdot x[n\text{-}1] - 1$; $x[0] = 40$

Impulse Representation of Sequences

A sequence, a function

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

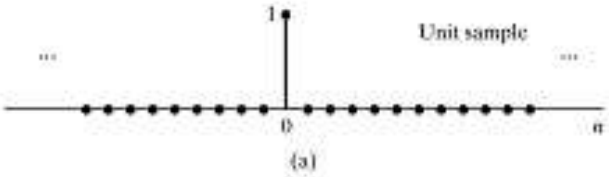
Value of the function at k



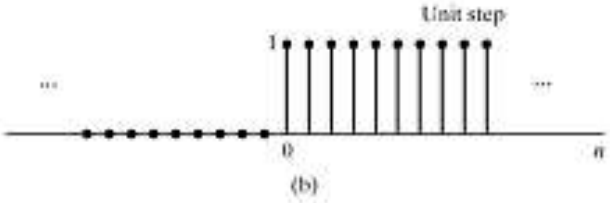
$$x[n]=a_{-3}\delta[n+3]+a_1\delta[n-1]+a_2\delta[n-2]+a_7\delta[n-7]$$

Some Useful Sequences

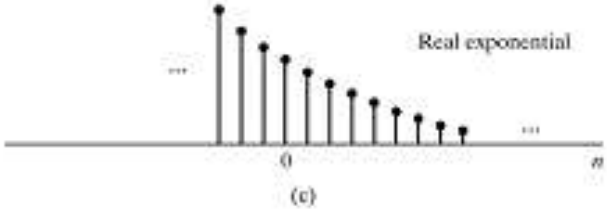
unit sample $\delta[n]=\begin{cases}1, & n=0\\ 0, & n\neq 0\end{cases}$



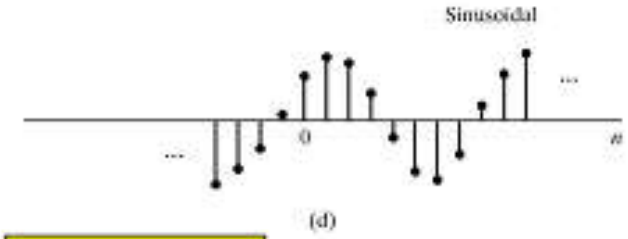
unit step $u[n]=\begin{cases}1, & n\geq 0\\ 0, & n<0\end{cases}$

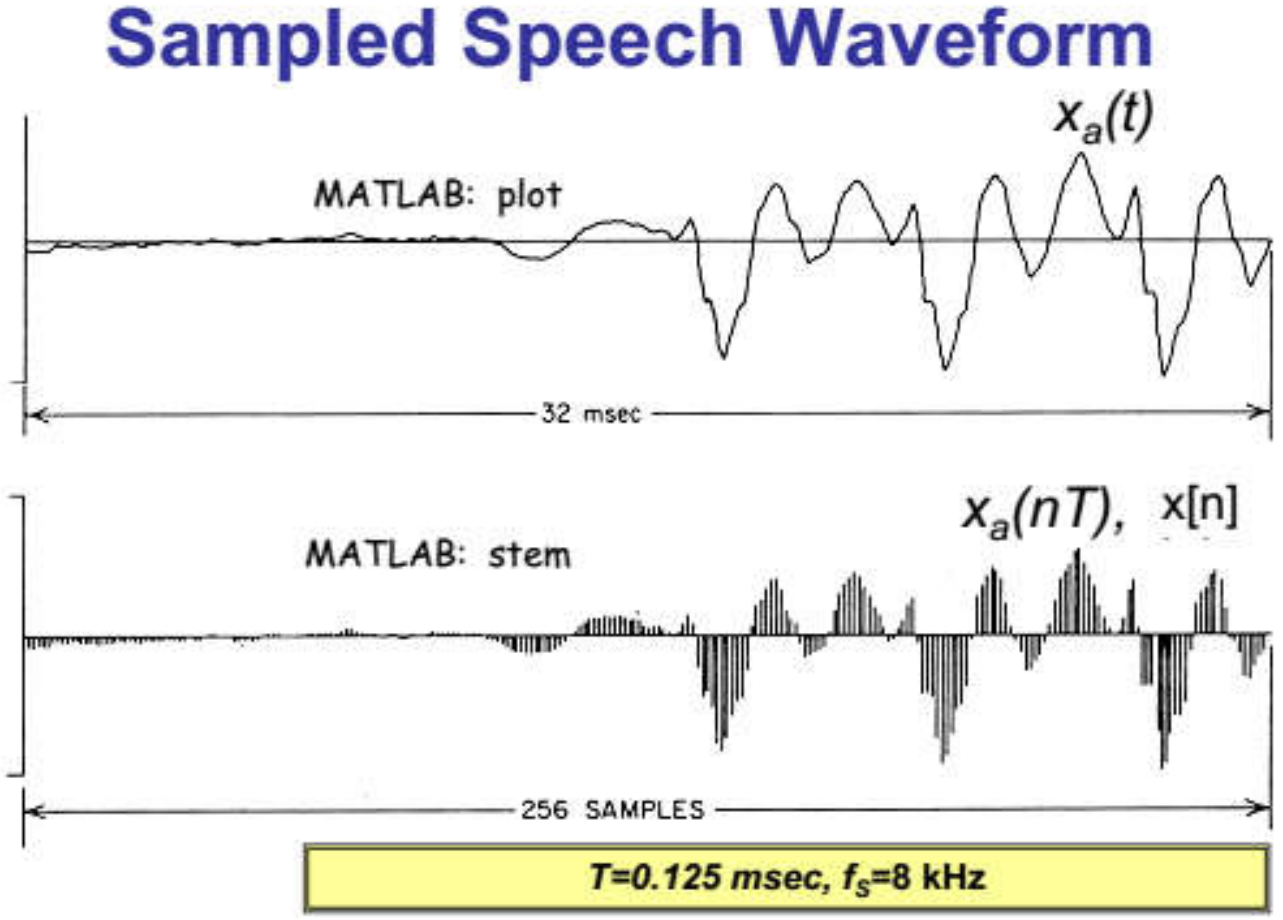


real exponential $x[n]=\alpha^n$



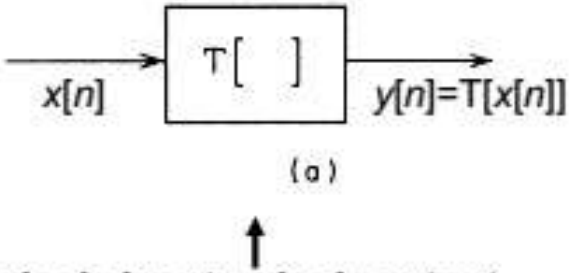
sine wave $x[n]=A\cos(\omega_0 n+\phi)$



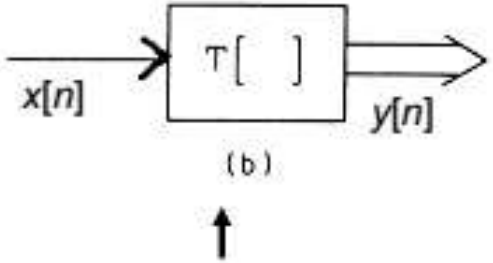

Sampled Speech Waveform
$x_a(t)$
MATLAB: plot
32 msec
$x_a(nT)$, x[n]
MATLAB: stem
256 SAMPLES
T=0.125 msec, f_S=8 kHz

Signal Processing

- Transform digital signal into more desirable form

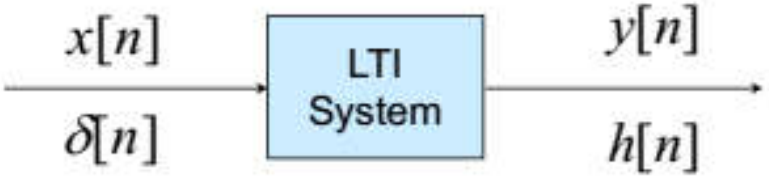


single input—single output



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):

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

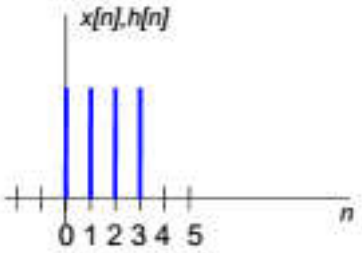
- LTI implies discrete convolution:

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

Convolution Example

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

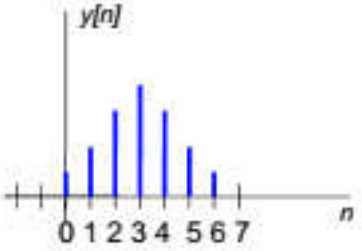
What is $y[n]$ for this system?



Solution :

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

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

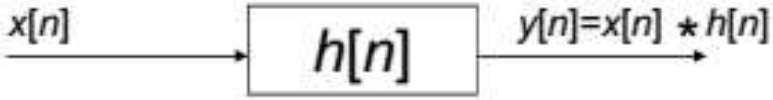


Linear Time-Invariant Systems

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

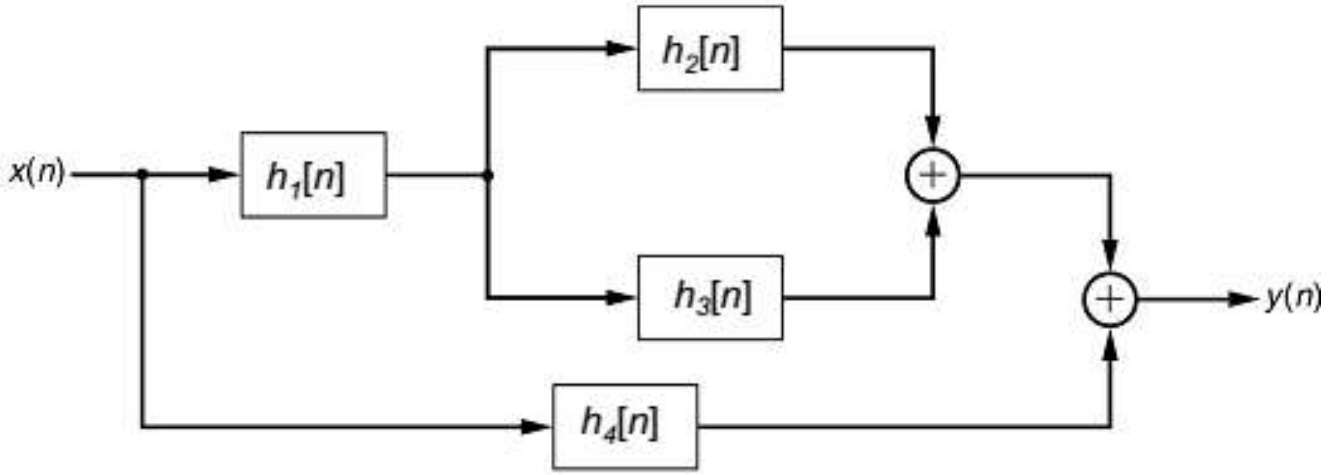
$$y[n] = x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = \sum_{k=-\infty}^{\infty} h[k]x[n-k]$$

$y[n] = h[n] * x[n]$, 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:

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

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

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

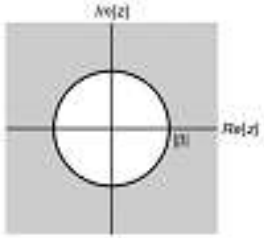
infinite power series in z^{-1}, with *x[n]* as coefficients of term in z^{-n}

- 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*:**

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

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



39

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	$n\,x[n]$	$-z\,dX(z)/dz$
Time Reversal	$x[-n]$ non-causal, need $x[N_0-n]$ to be causal for finite length sequence	$X(z^{-1})$
Convolution	$x[n] * h[n]$	$X(z)\,H(z)$
Multiplication of Sequences	$x[n]\,w[n]$	$\frac{1}{2\pi j}\oint_C X(v)W(z/v)v^{-1}dv$ circular convolution in the frequency domain

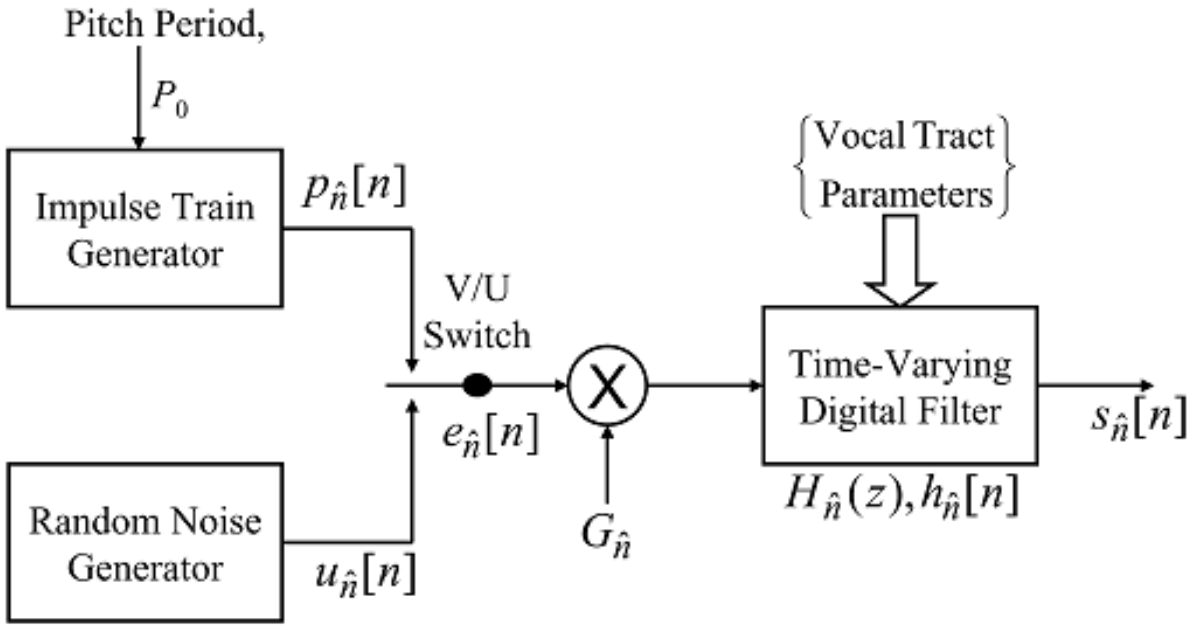


Fig. 4.1 Voiced/unvoiced/system model for a speech signal.

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

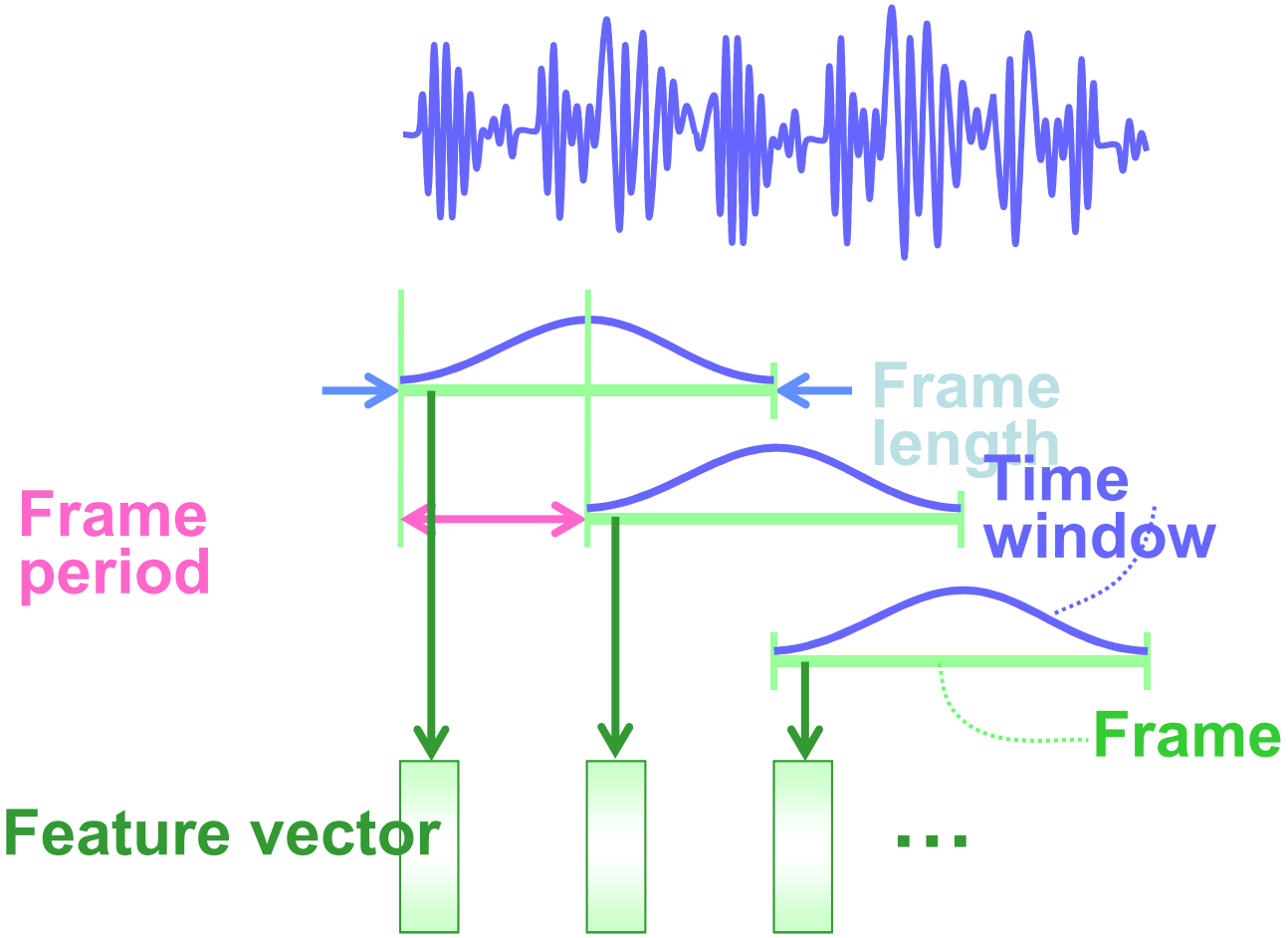
$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \qquad (4.2)$$

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

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

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，可以视作一个准稳态过程。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。



语音特征矢量(短时谱)提取

语音分析技术

- 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）

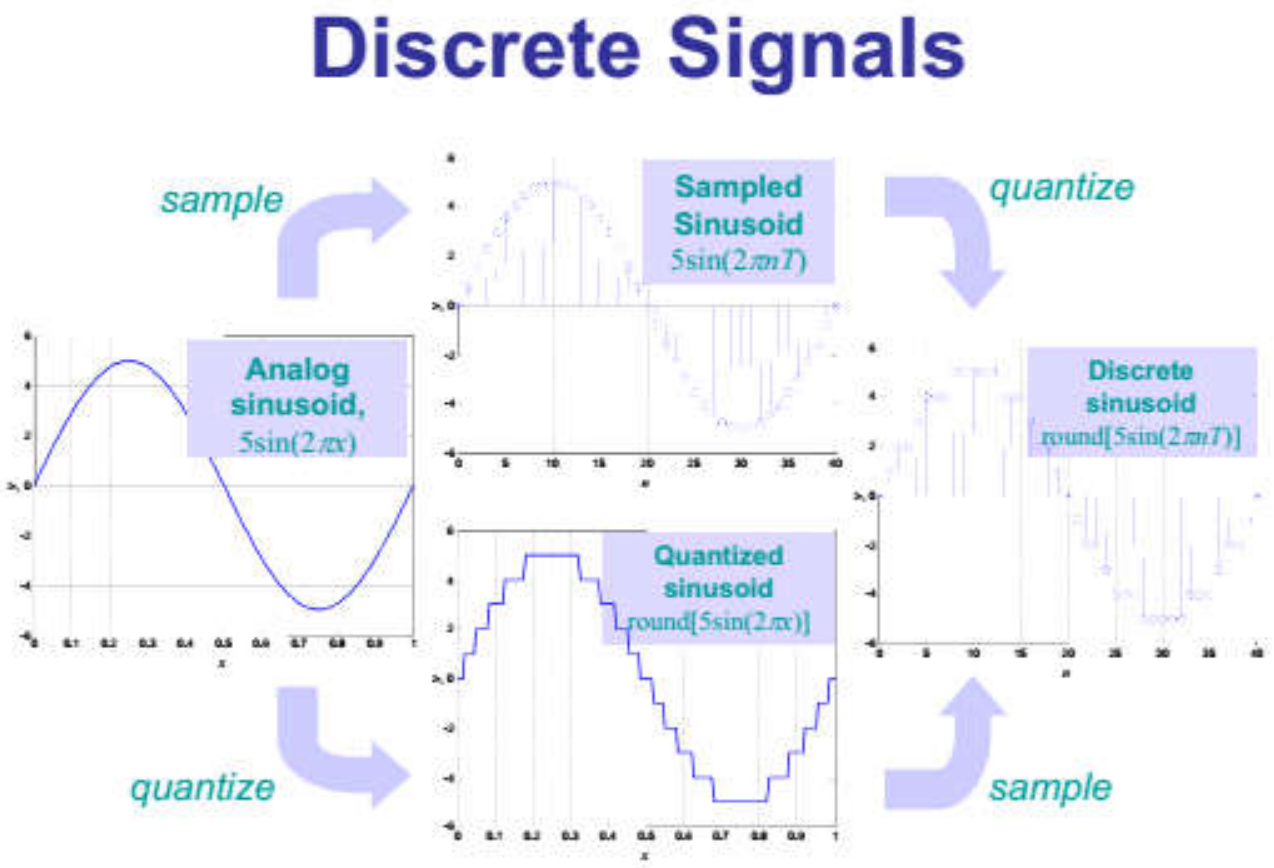
语音分析技术

- 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
sample
Sampled
Sinusoid
5sin(2πnT)
quantize
Analog
sinusoid,
5sin(2πx)
Discrete
sinusoid
round[5sin(2πnT)]
Quantized
sinusoid
round[5sin(2πx)]
quantize
sample

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”、”強度”（**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**）为單位，是一个相对強度的值，比较符合人耳对于大小声音的感觉。

能量/音量

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 个采样点被分割成迭加率为 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：

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

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

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

能量/音量

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。

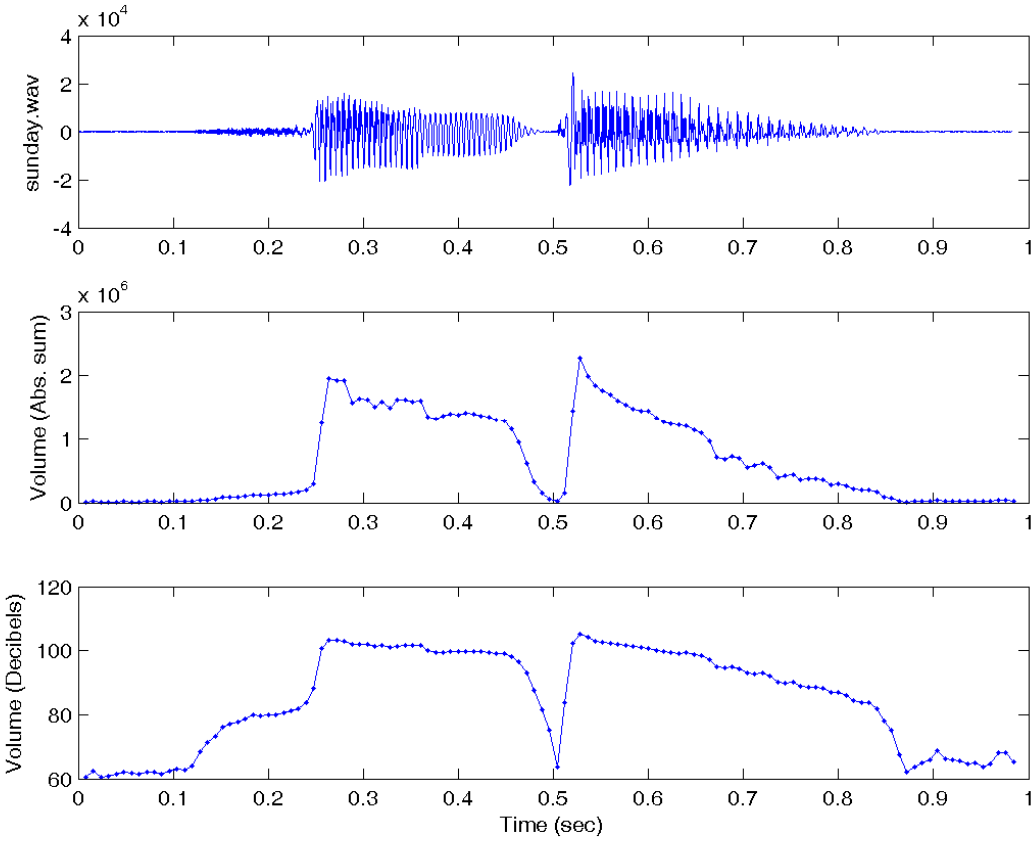
2、是一个相对性的指标，受到麦克风设定的影响很大。

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。

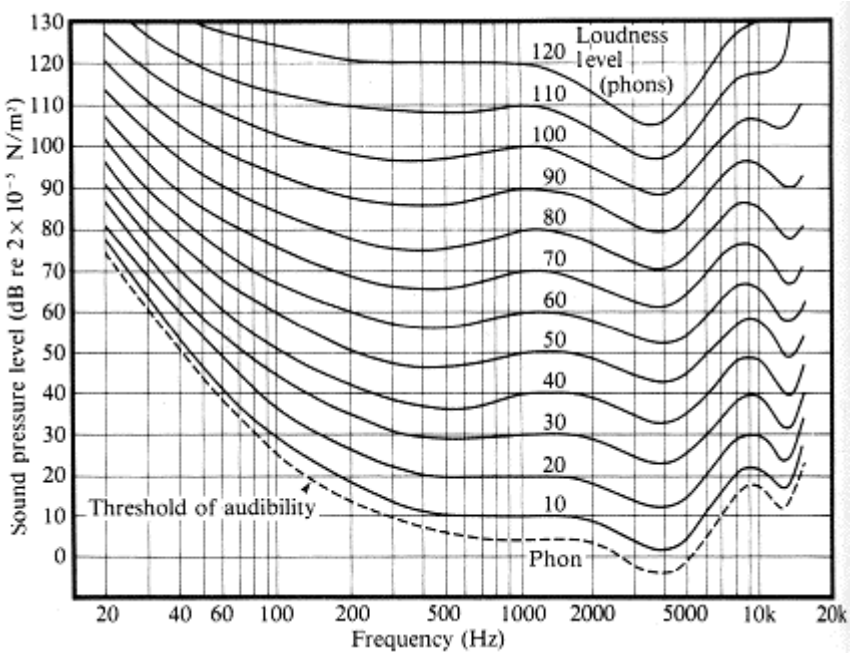
能量/音量

举例1: Sunday



能量/音量

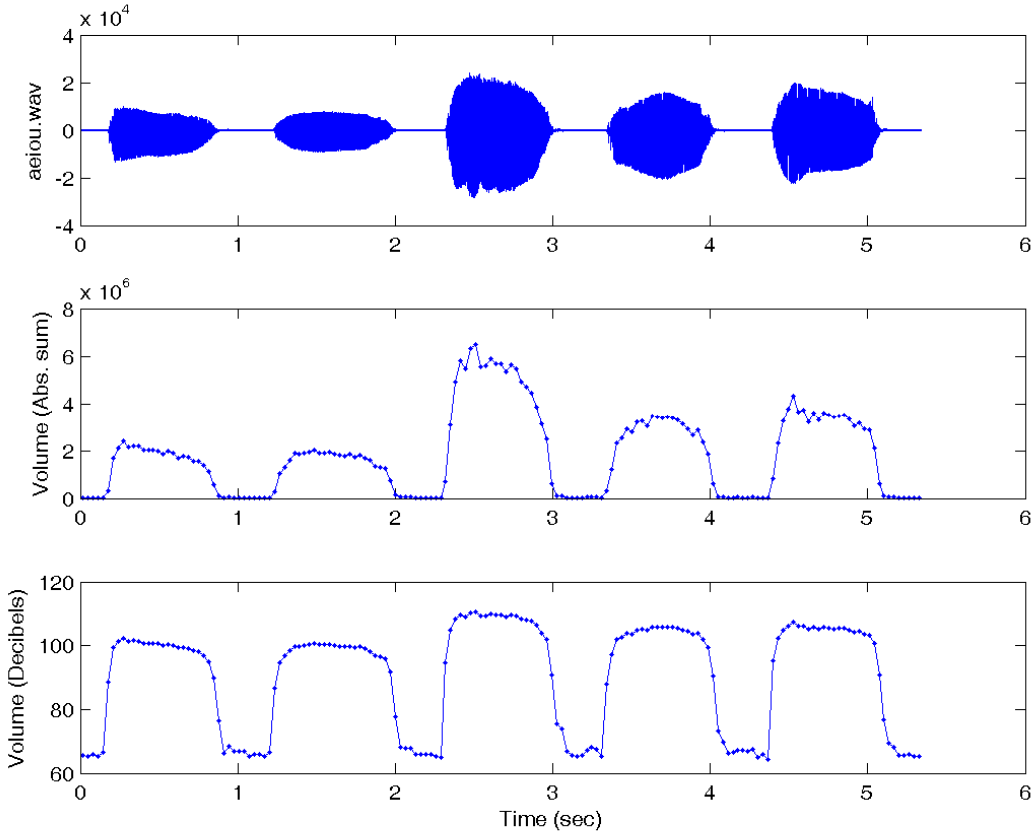
基本上我们使用音量来表示声音的強弱，但是前述两种计算音量的方法，只是用数学的公式来逼近人耳的感觉，和人耳的感觉有時候会有相当大的差别，为了区分，我们使用”主观音量”来表示人耳所听到的音量大小。例如，人耳对于同样振幅但不同频率的声音，所产生的主观音量就会非常不一样。若把以人耳为测试主体的「等主观音量曲线」（**Curves of Equal Loudness**）画出來，就可以得到下图：



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

“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：

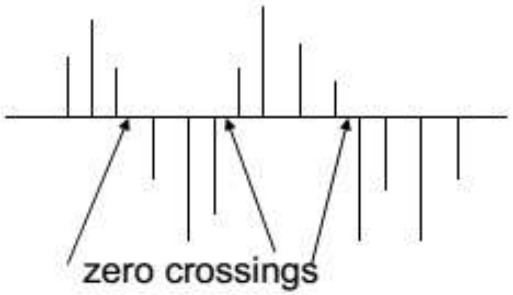
$$Z_{\hat{n}} = \sum_{m=-\infty}^{\infty} 0.5\,|\mathrm{sgn}\{x[m]\} - \mathrm{sgn}\{x[m-1]\}|\,w[\hat{n}-m], \qquad (4.7)$$

where

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

过零率

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 pe cycle with two zero crossings per cycle, giving an average zero crossing rate of

$z_1=(2)$ crossings/cycle x (F_0/F_S) cycles/sample

$z_1=2F_0/F_S$ crossings/sample (i.e., **z_1 proportional to F_0**)

$z_M=M\,(2F_0/F_S)$ crossings/(M samples)

Sinusoid Zero Crossing Rates

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

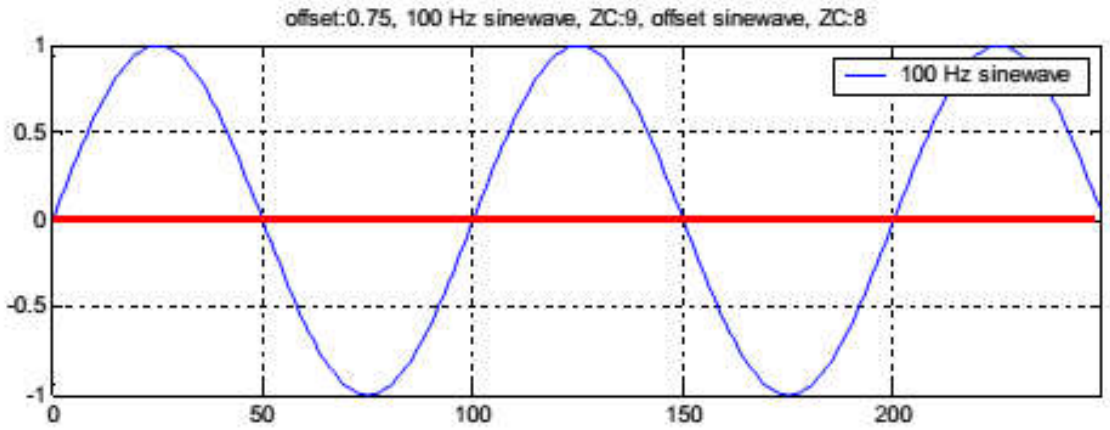
1. $F_0 = 100$ Hz sinusoid has $F_S / F_0 = 10,000 / 100 = 100$ samples/cycle;
   or $z_1 = 2 / 100$ crossings/sample, or $z_{100} = 2 / 100 * 100 =$
   2 crossings/10 msec interval

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

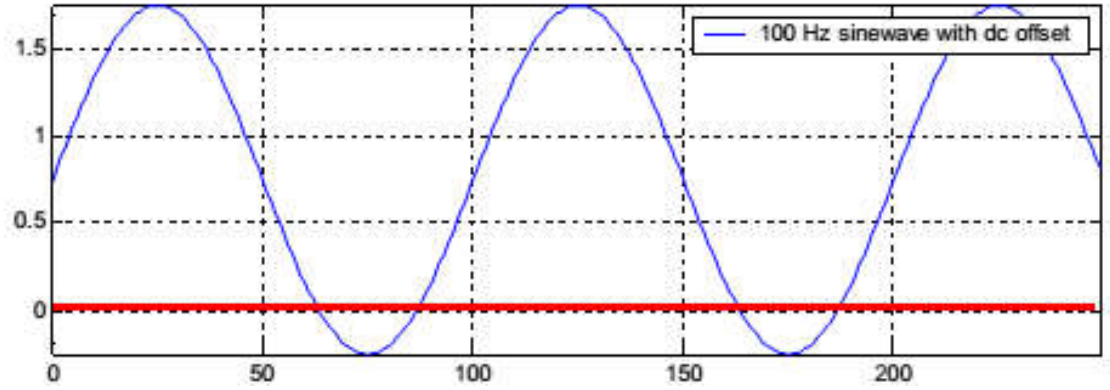
过零率

Zero Crossing for Sinusoids



ZC=9

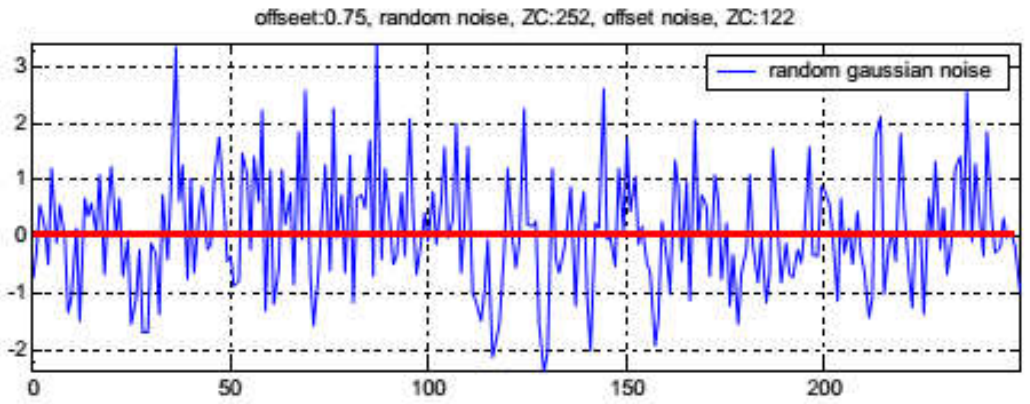
Offset=0.75



ZC=8

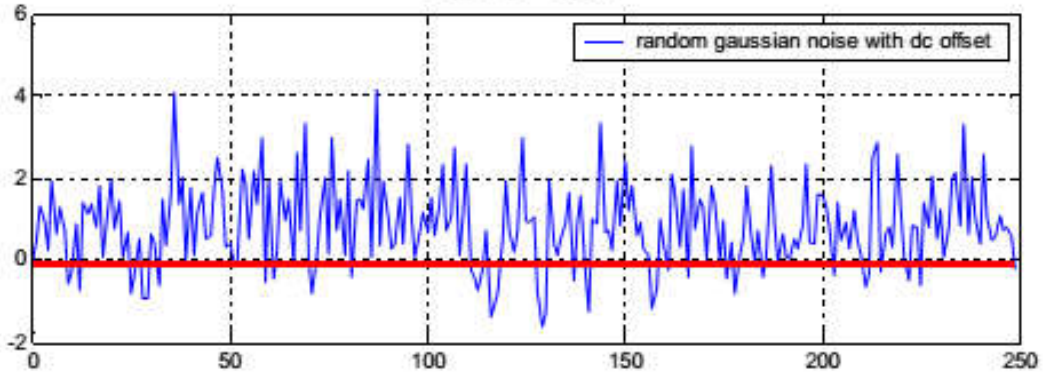
过零率

Zero Crossings for Noise



ZC=252

Offset=0.75



ZC=122

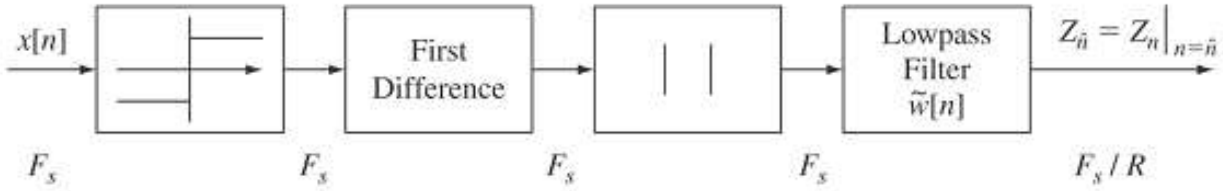
ZC Rate Definitions

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

$$\text{sgn}(x[n]) = 1 \qquad x[n] \geq 0$$
$$= -1 \qquad x[n] < 0$$

- simple rectangular window:

$$\tilde{w}[n] = 1 \qquad 0 \leq n \leq L-1$$
$$= 0 \qquad \text{otherwise}$$
$$L_{\text{eff}} = L$$



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:

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

  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

  $z_M = z_1 \cdot M$ = rate of zero crossings per M 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; $T = 100$ μsec; τ=10 msec; $M = 100$ samples
- $F_S = 8{,}000$ Hz; $T = 125$ μsec; τ=10 msec; $M = 80$ samples
- $F_S = 16{,}000$ Hz; $T = 62.5$ μsec; τ=10 msec; $M = 160$ samples

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

ZC Normalization

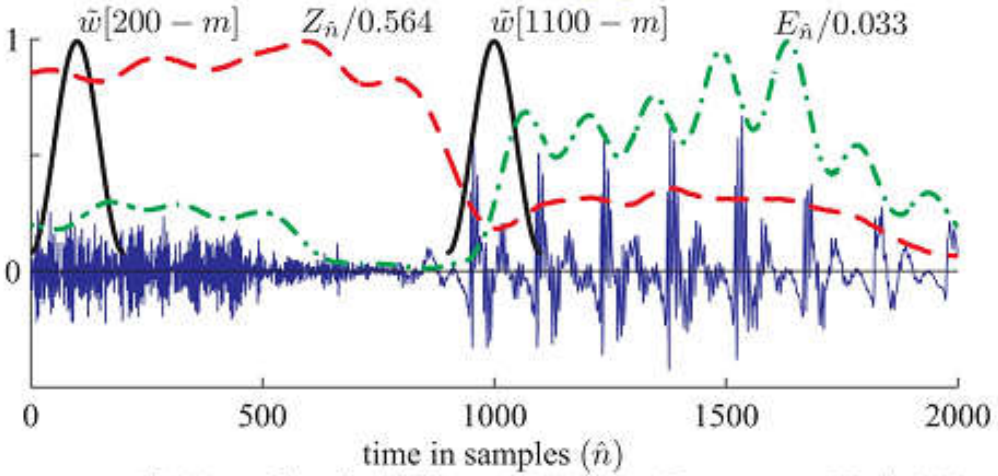
- For a 1000 Hz sinewave as input, using a 40 msec window length (L), with various values of sampling rate (F_S), we get the following:

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

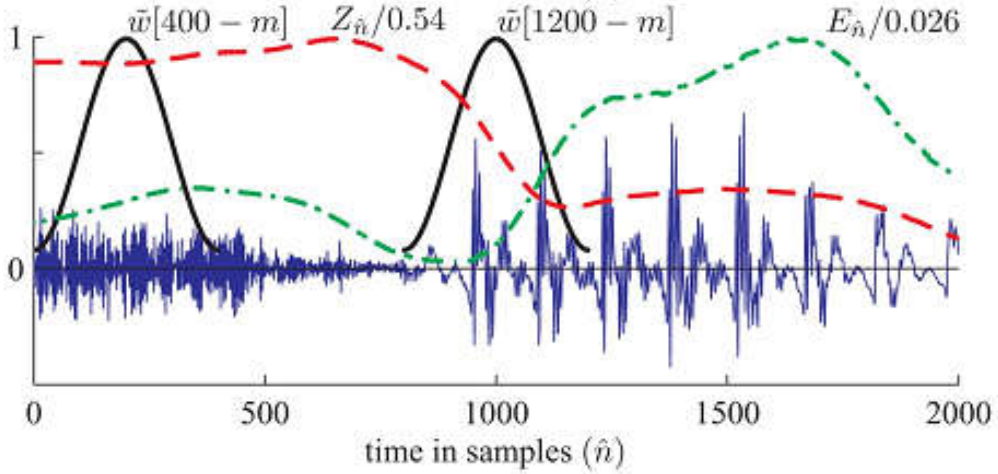
- Thus we see that the normalized (per interval) zero crossing rate, z_M, is independent of the sampling rate and can be used as a measure of the dominant energy in a band.

过零率

ZC and Energy Computation



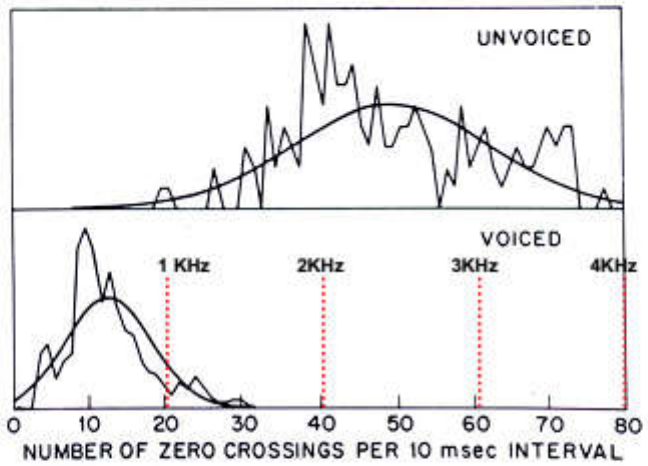
Hamming window with duration *L=201* samples (12.5 msec at *Fs*=16 kHz)



Hamming window with duration *L=401* samples (25 msec at *Fs*=16 kHz)

过零率

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

过零率

ZC Rates for Speech

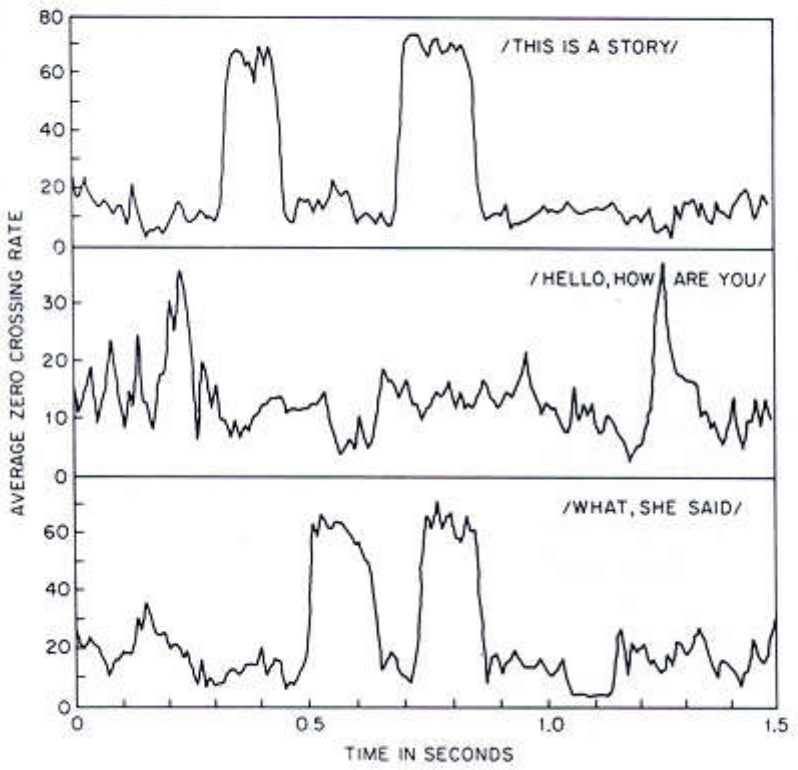


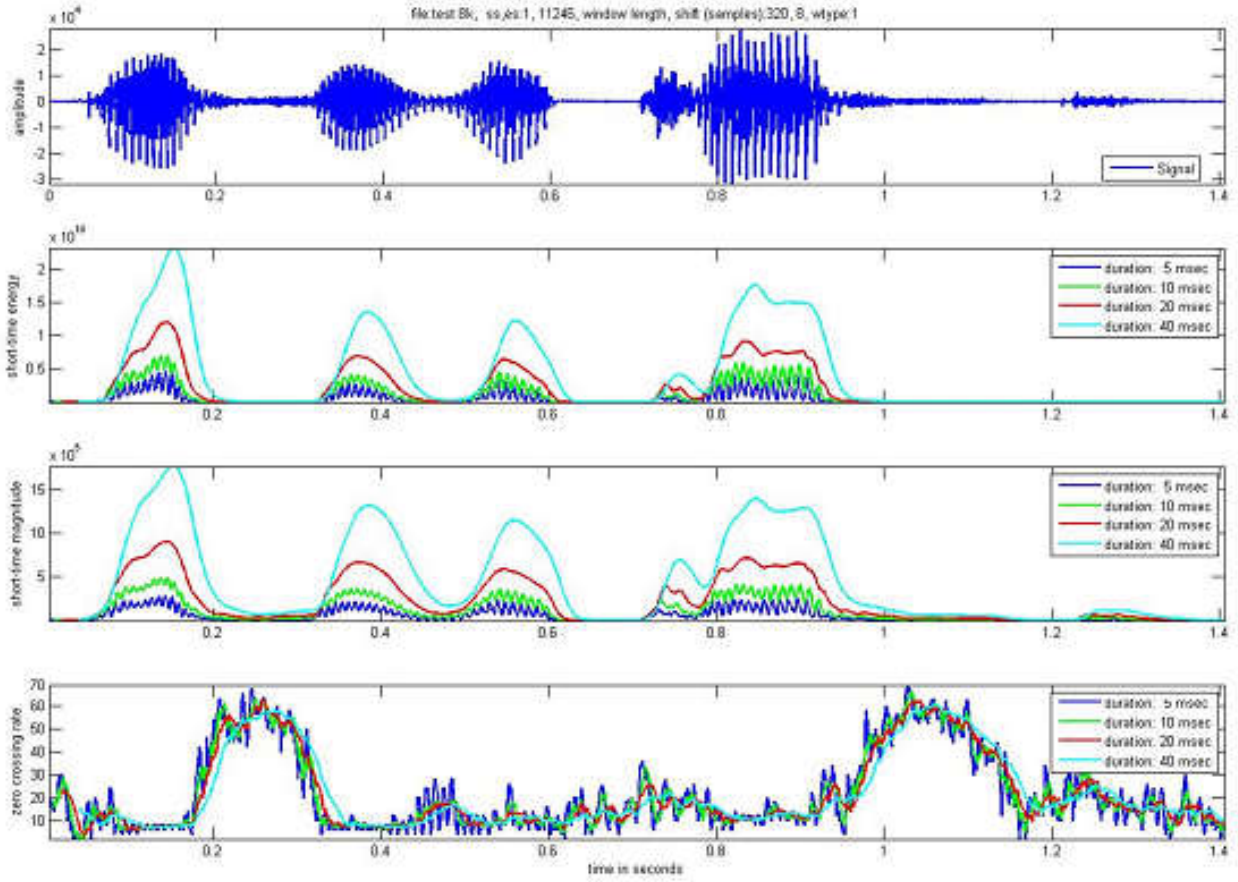
Fig. 4.12 Average zero-crossing rate for three different utterances.

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

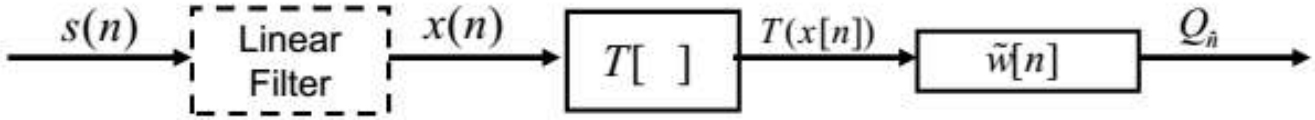
48

过零率

Short-Time Energy, Magnitude, ZC



Summary of Simple Time Domain Measures



$$Q_{\hat{n}} = \sum_{m=-\infty}^{\infty} T(x[m])\tilde{w}[\hat{n}-m]$$

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:

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

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

$= -1 \quad x[m] < 0$

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)

课后任务I

- 下载安装实验软件

  – **PRAAT** 语音分析

  doing Phonetics by Computer

  – **VOICEBOX**说话人识别

  Speech Processing Toolbox for MATLAB

  http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html