语音信号数字处理基础
Digital Speech Signal Processing

清华大学深圳研究生院
吴志勇
zywu@sz.tsinghua.edu.cn

MEDIA JOINT Research Center

学习要点

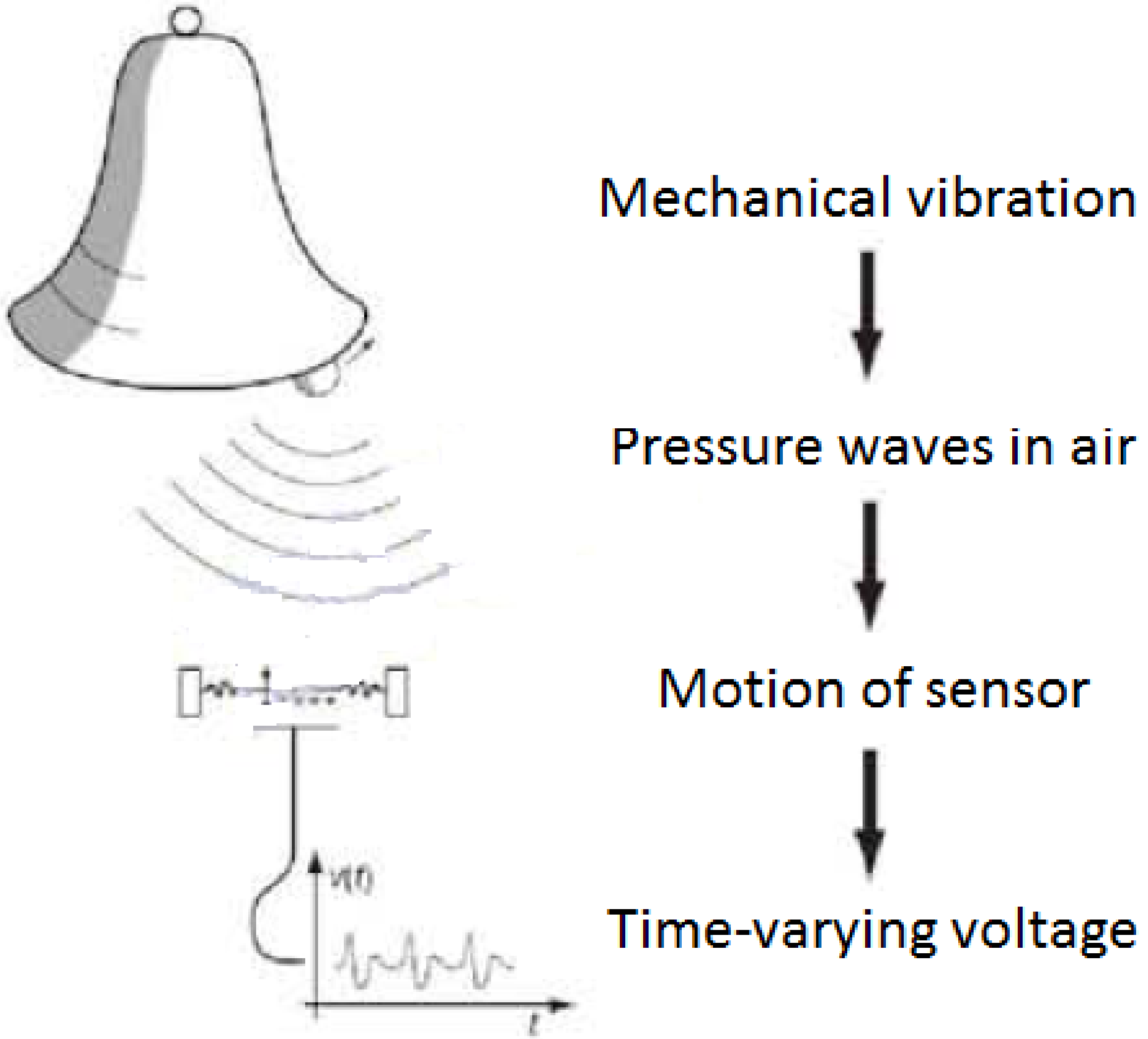
- **语音信号的数字化**
  - 信号的频谱特性
  - 抽样
  - 量化
- **语音信号的时域处理**
  - 语音信号的短时分析与预处理
  - 短时能量、短时平均幅度、短时平均过零率
  - 语音的端点检测
  - 短时自相关函数
  - 语音的短时基音估计
- **语音信号的频域分析**
  - 短时傅立叶变换
  - 语谱图

语音信号的数字化

ANALOG-TO-DIGITAL CONVERSION OF SPEECH SIGNAL

数字音频

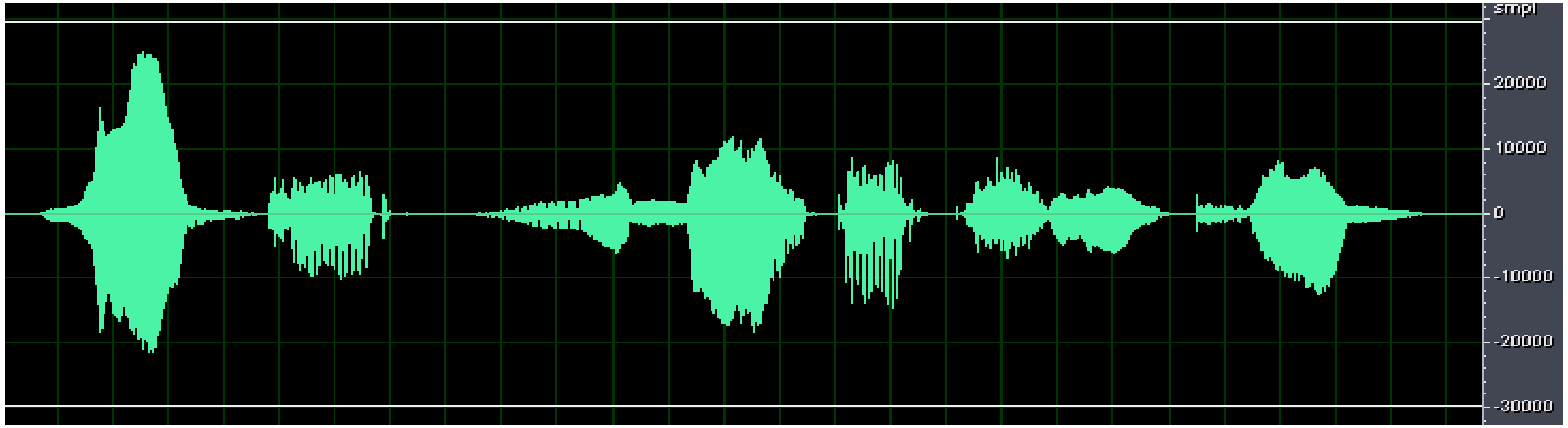
- **什么是数字音频?**
  - 声音是机械振动。振动越强，声音越大。
  - 话筒把机械振动转换成电信号。
  - 模拟音频中以模拟电压的幅度表示声音强弱。
  - 在数字音频中，数字声音是一个数据序列。通过离散的数值大小来表示声音强弱。
  - **数字音频**是由模拟声音经抽样、量化和编码后得到的。



音频数字化

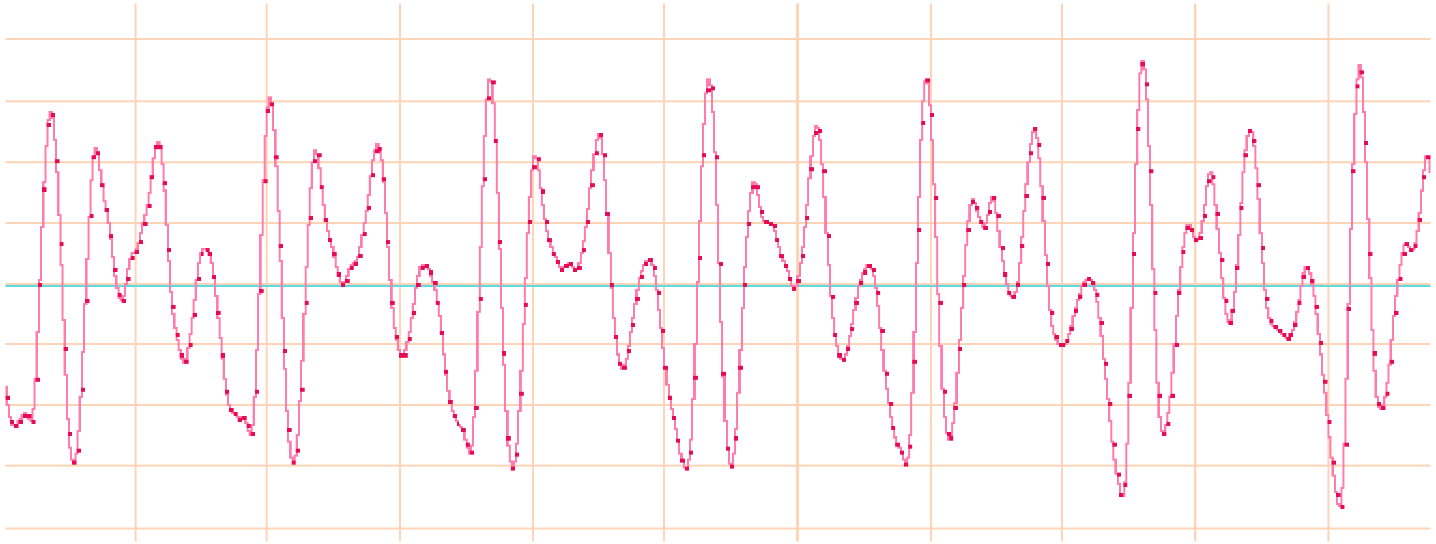
- **音频数字化**
  - 把模拟音频信号转换成有限个数字表示的离散序列，即实现音频数字化。它涉及到音频的抽样、量化和编码。
- **抽样**
  - 当把模拟声音变成数字声音时，每隔一个时间间隔在摸拟声音波形上取一个幅度值，这称之为**抽样**。该时间间隔称为**抽样周期**(其倒数称为**采样频率**)。
- **量化**
  - 在数字音频中，用数字来表示音频幅度时，只能把无穷多个电压幅度用有限个数字表示。即把某一幅度范围内的电压用一个数字表示，这称之为**量化**。量化时所采用的数字的上限称之为**量化精度**。
- **编码**
  - 对原始的音频数据进行压缩，便于存储和传输。

语音信号的数字化



望着无奈的秋天

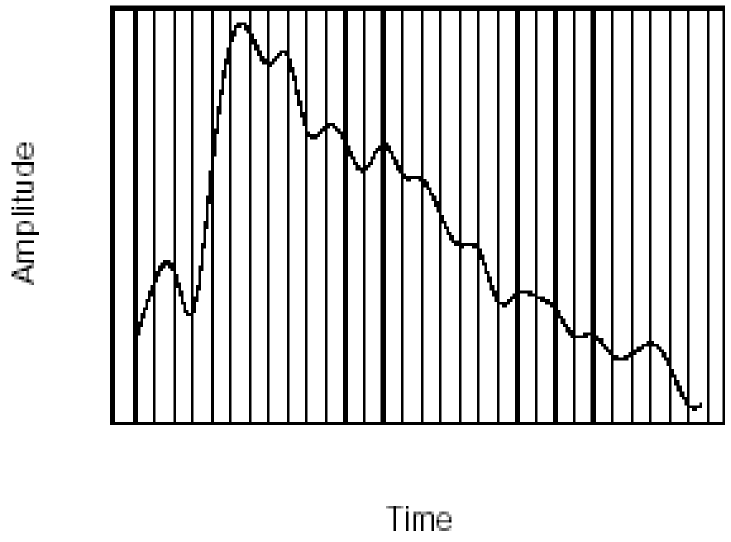
Wav文件格式：16KHz, 16Bit, 单声道



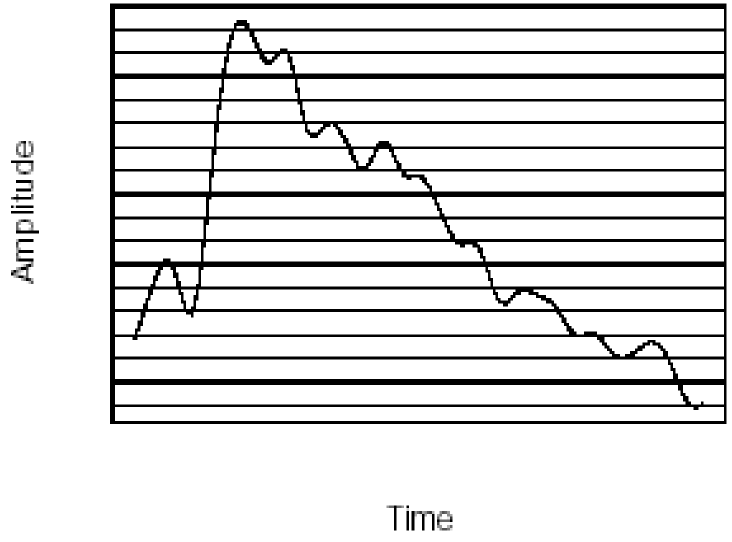
CoolEdit /
Adobe Audition

语音信号的抽样和量化

由模拟信号变成数字信号：

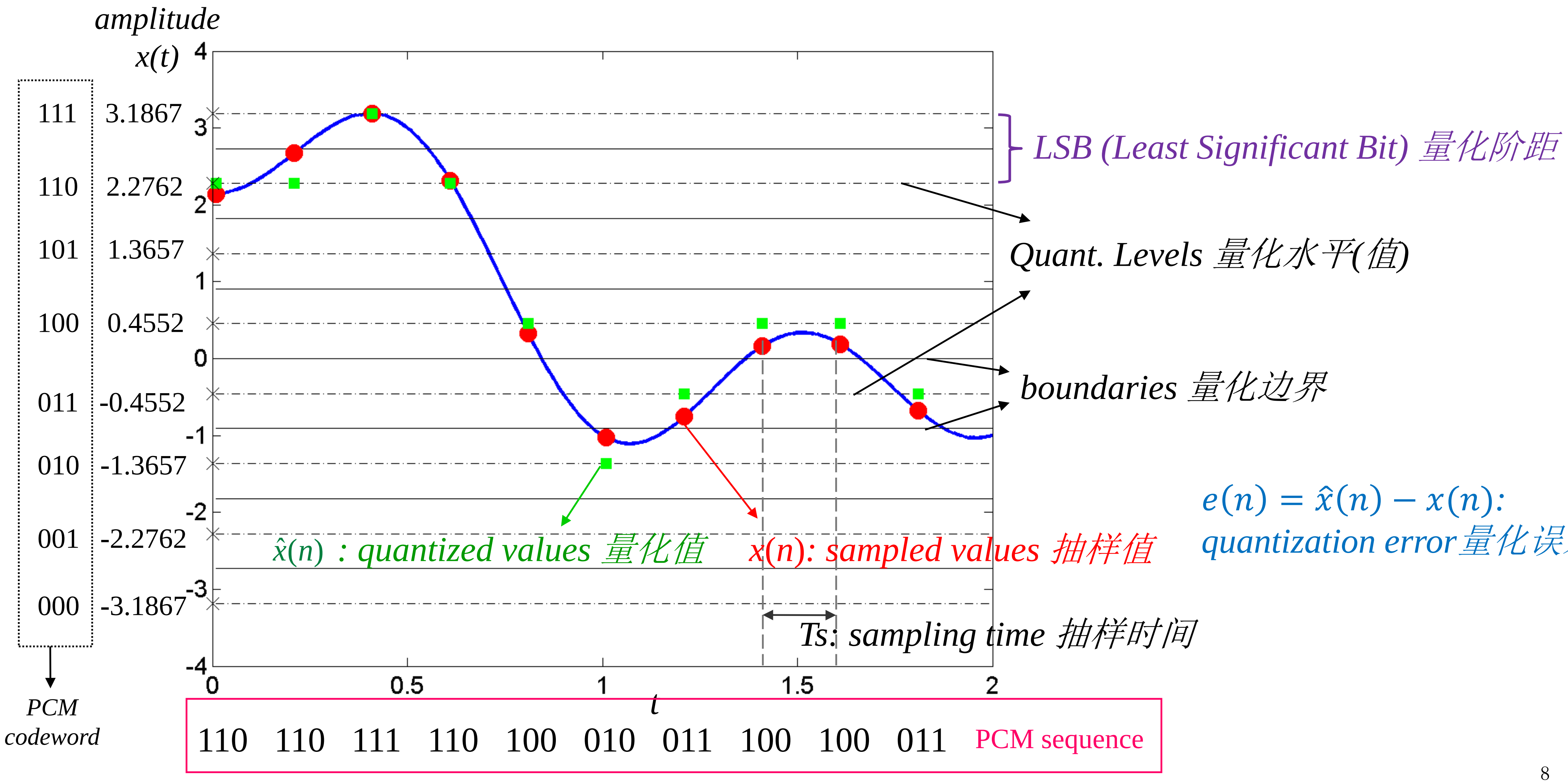


在时间上：抽样



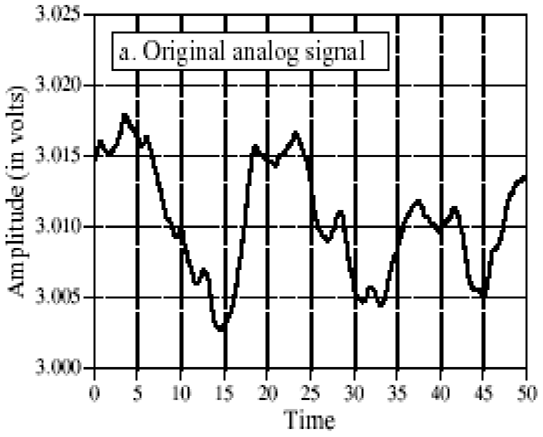
在幅度上：量化

语音信号的抽样和量化

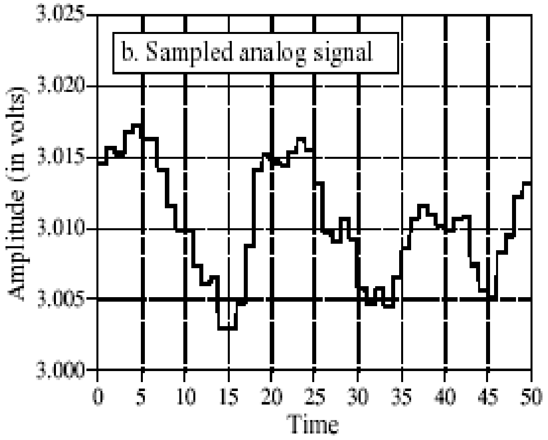


$e(n) = \hat{x}(n) - x(n)$:
quantization error 量化误差

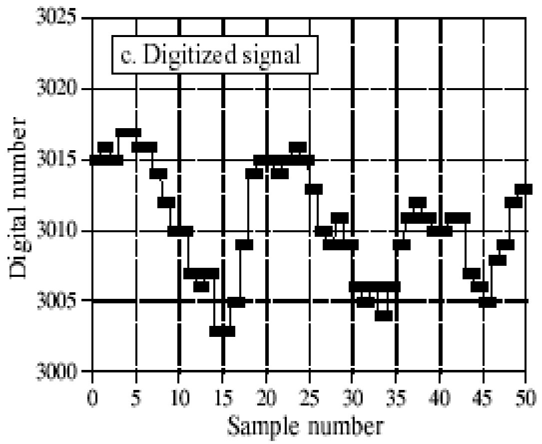
语音信号的抽样和量化



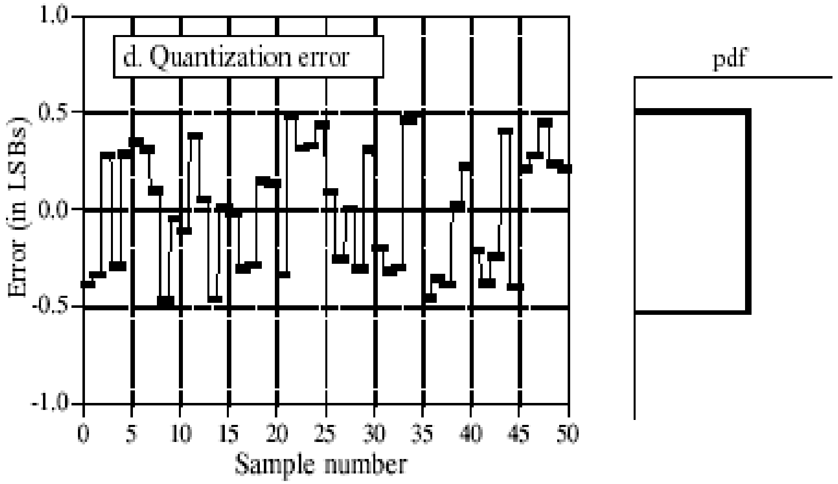
原始模拟信号



抽样后的信号



量化后的信号



量化误差(噪声)

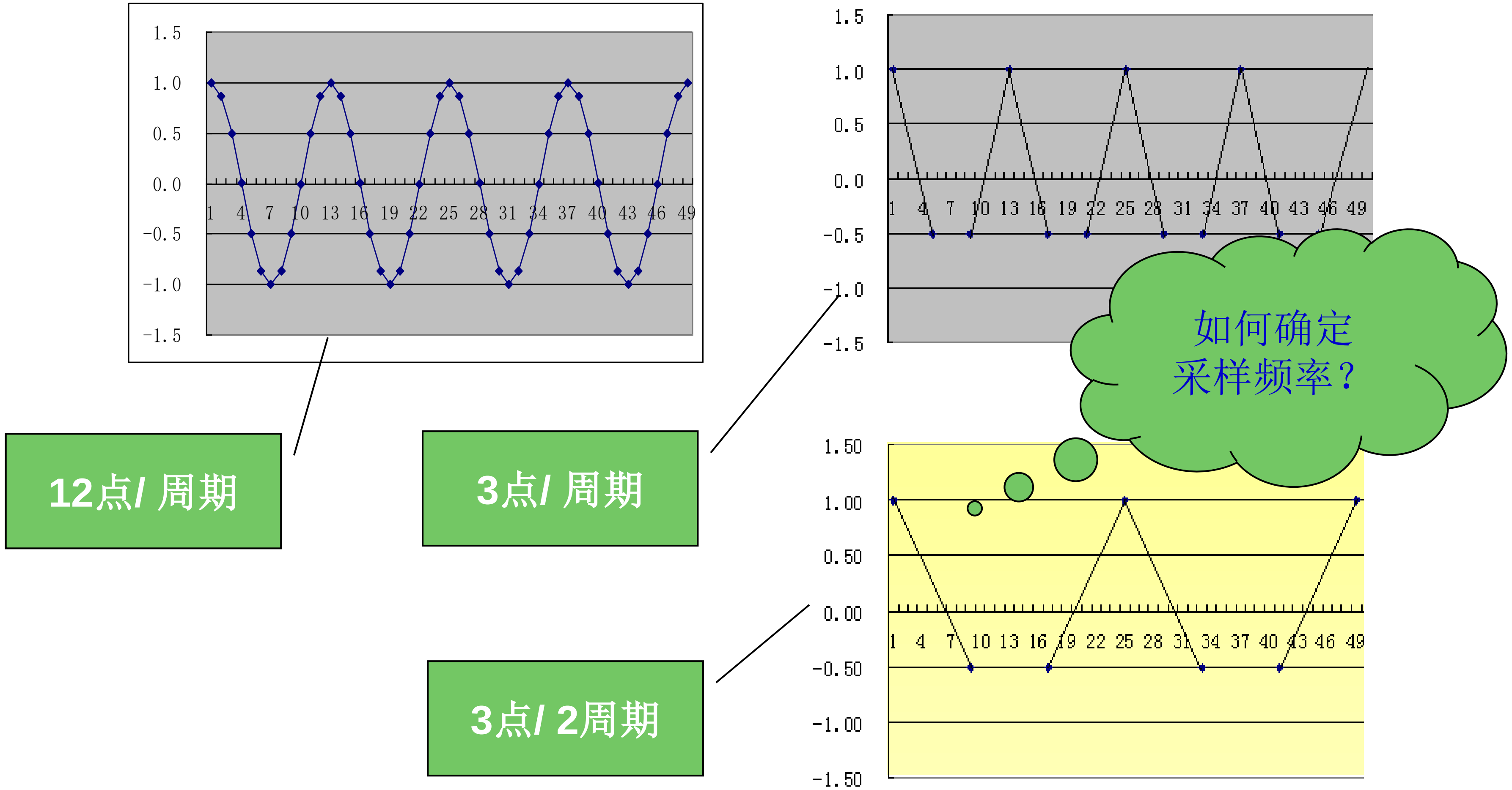
* LSB: Least Significant Bit，量化阶距

抽样与混叠：

采样频率如何确定？混叠时发生了什么？

SAMPLING AND ALIASING OF SPEECH SIGNAL

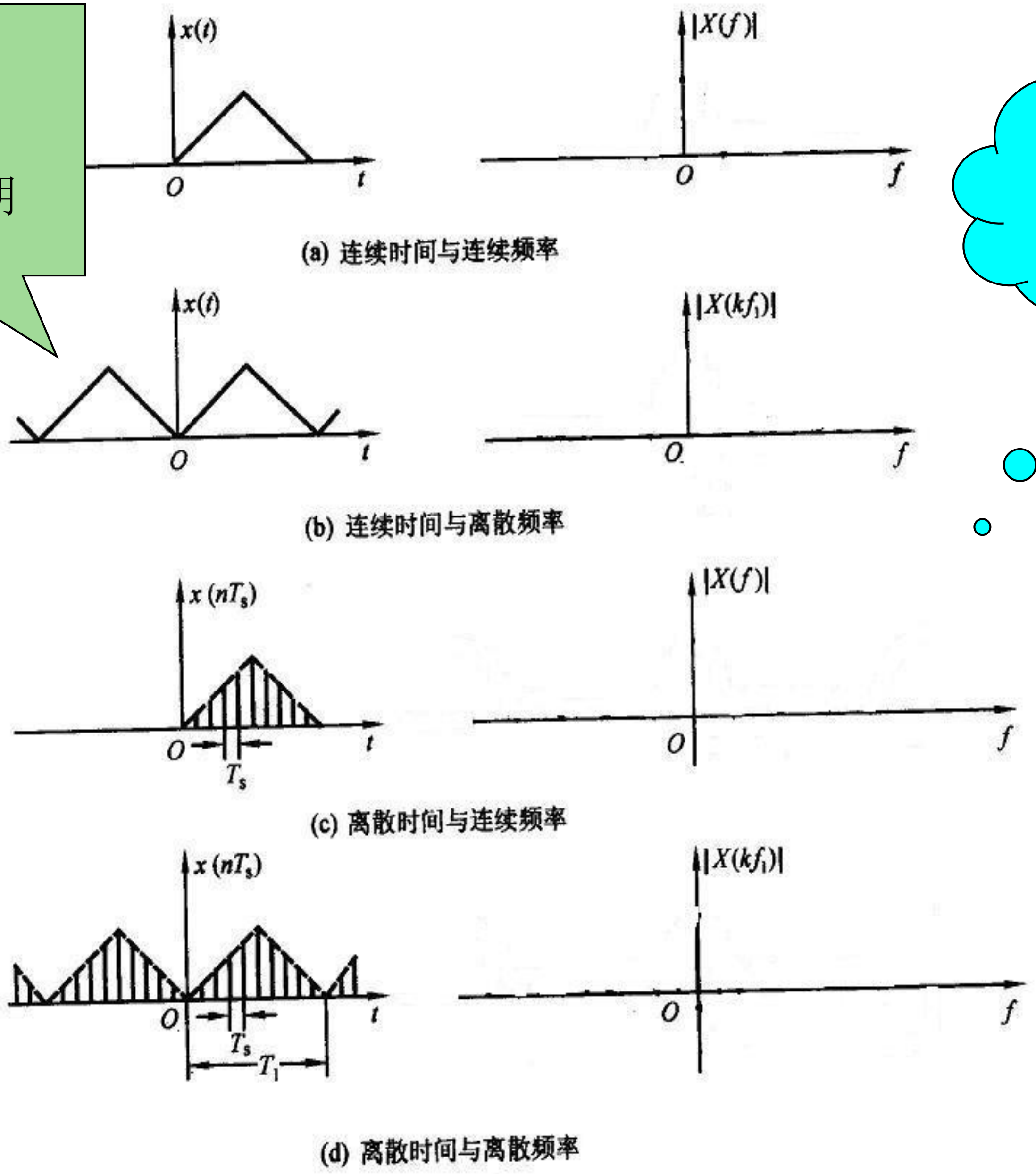
语音信号的抽样



信号的频谱特性

连续/离散

周期/非周期



频谱特性？

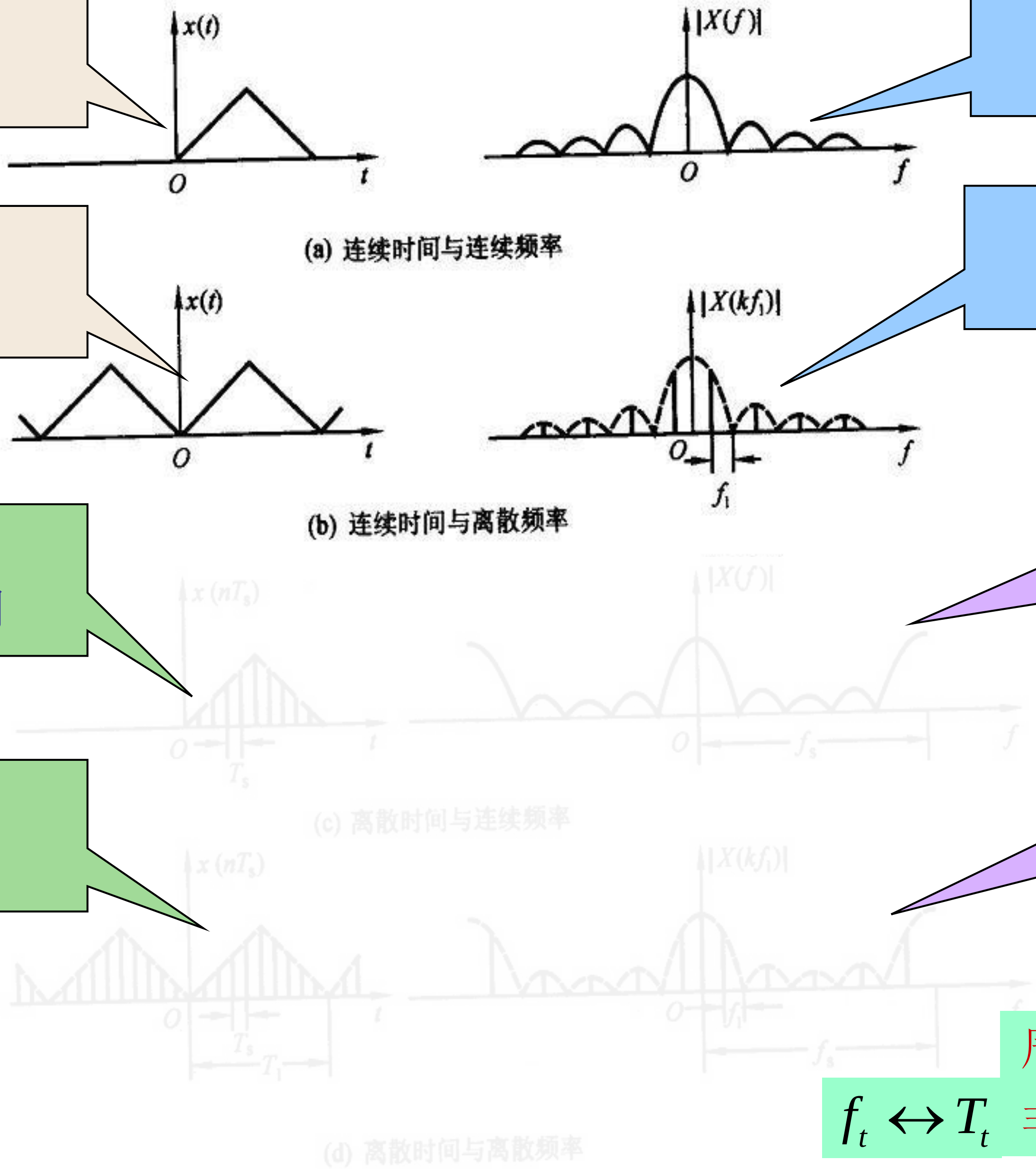
信号的频谱特性

非周期

周期

非周期

周期



连续

离散

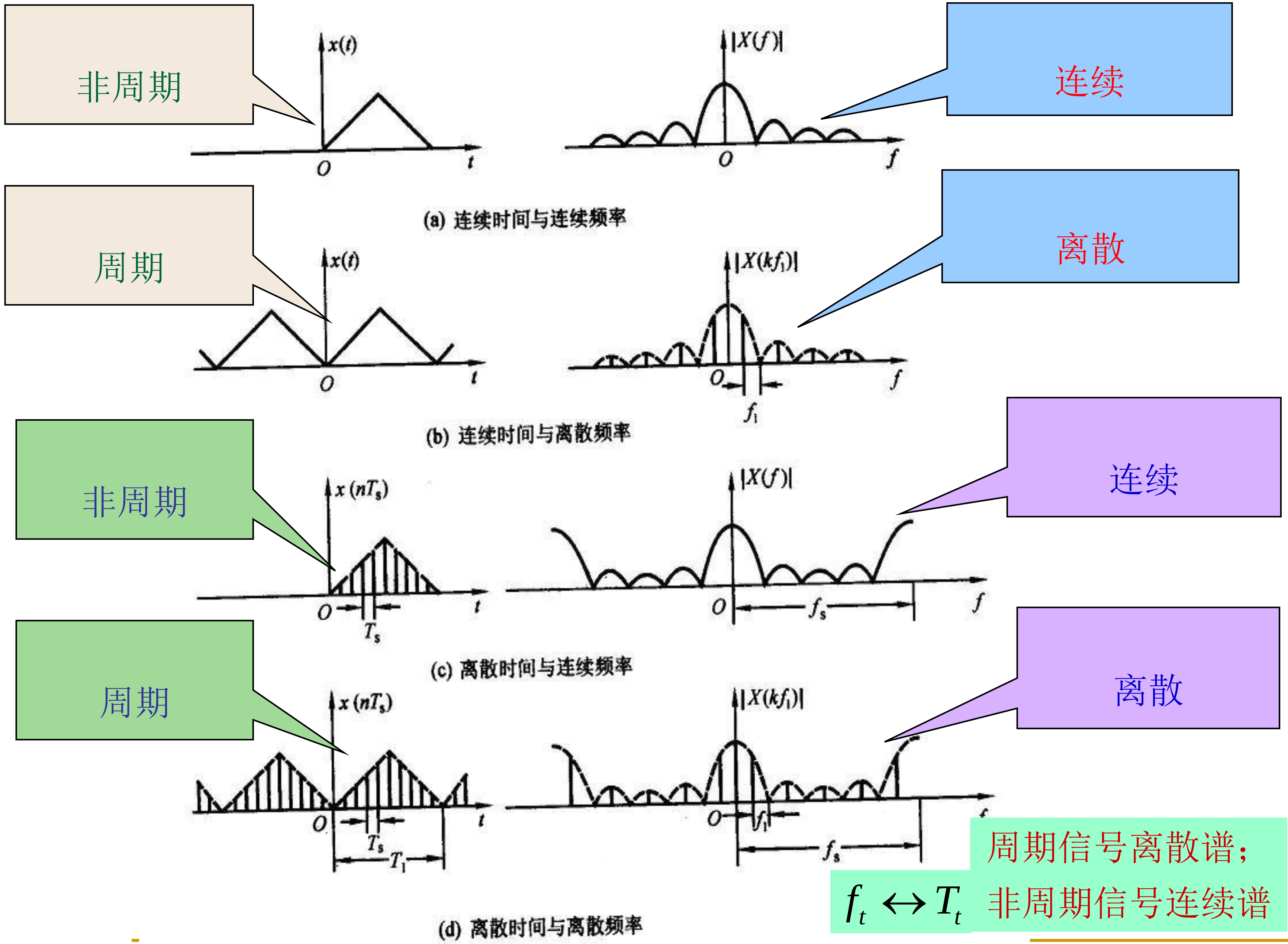
连续

离散

周期信号离散谱；

$f_t \leftrightarrow T_t$ 非周期信号连续谱

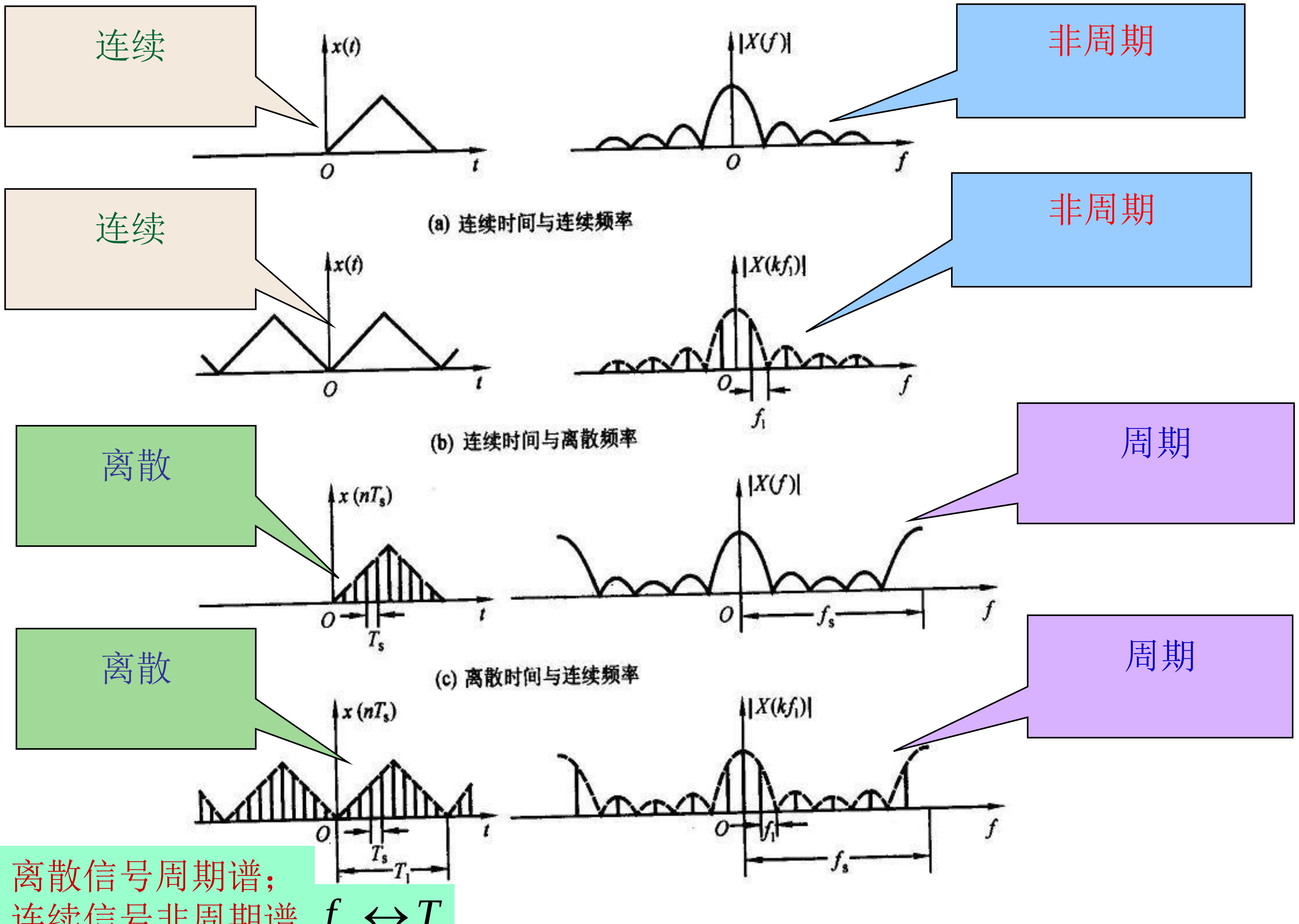
信号的频谱特性



周期信号离散谱；

$f_t \leftrightarrow T_t$ 非周期信号连续谱

信号的频谱特性



(a) 连续时间与连续频率

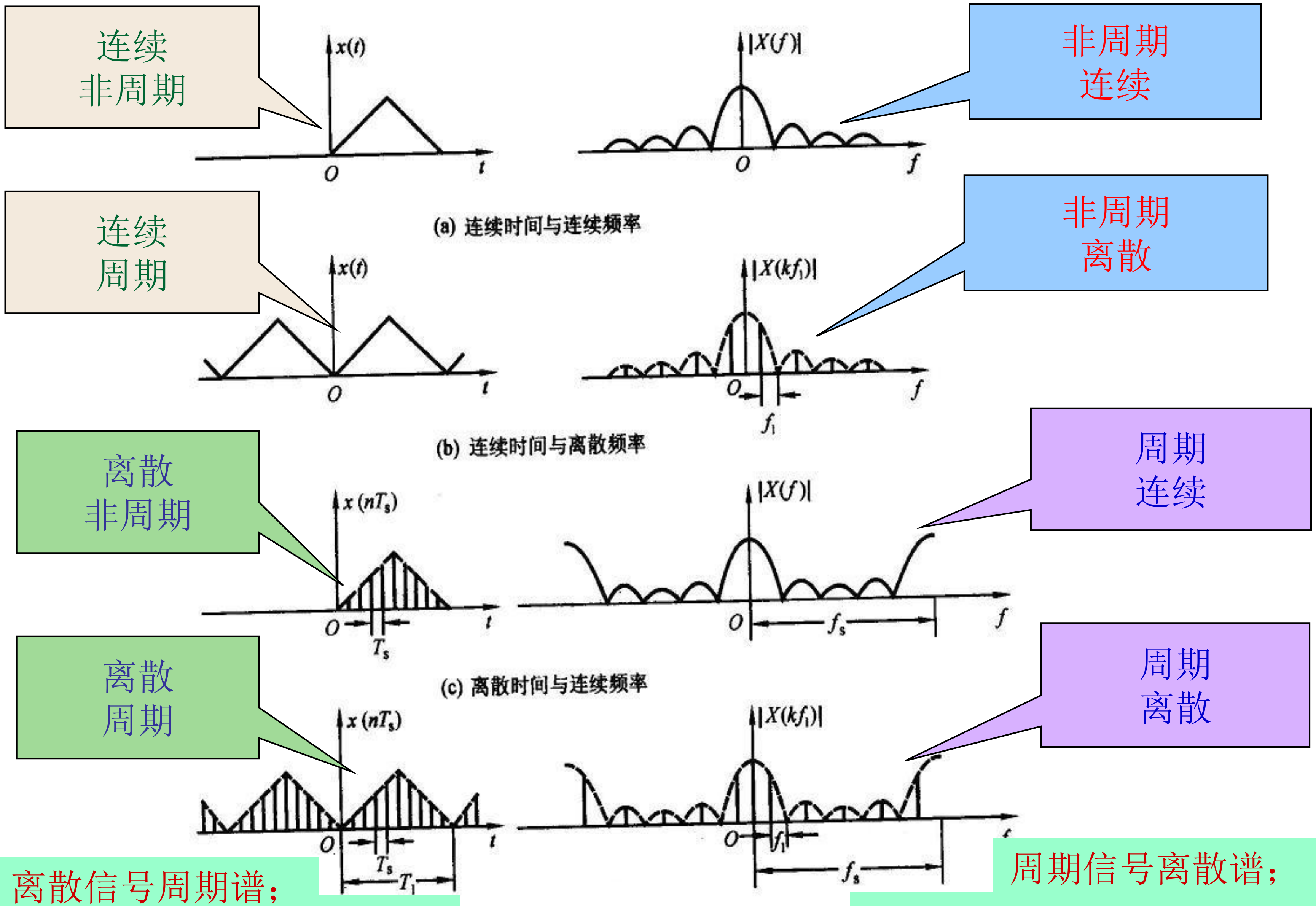
(b) 连续时间与离散频率

(c) 离散时间与连续频率

(d) 离散时间与离散频率

离散信号周期谱；
连续信号非周期谱 $f_s \leftrightarrow T_s$

信号的频谱特性



(d) 离散时间与离散频率

离散信号周期谱；
连续信号非周期谱 $f_s \leftrightarrow T_s$

周期信号离散谱；
$f_t \leftrightarrow T_t$ 非周期信号连续谱

语音信号的抽样

语音信号的理想抽样输出为：

$$f_s(t)=\sum_{n=-\infty}^{\infty} f(t)\delta(t-nT_s)=\sum_{n=-\infty}^{\infty} f(nT_s)\delta(t-nT_s)$$

根据时域相乘等于频域卷积，可求抽样信号的频谱

$$F_s(j\omega)=\frac{1}{2\pi}[F(j\omega)*\Delta_{T_s}(j\omega)]$$

其中

$$F(j\omega)=DTFT[f(t)]=\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt$$

$$\Delta_{T_s}(j\omega)=DTFT[\delta_{T_s}(t)]=\omega_s\sum_{k=-\infty}^{\infty}\delta(\omega-k\omega_s)$$

语音信号的抽样

$$F_s(j\omega)=\frac{1}{2\pi}[F(j\omega)*\frac{2\pi}{T_s}\sum_{k=-\infty}^{\infty}\delta(\omega-k\omega_s)]$$

$$=\frac{1}{T_s}\int_{-\infty}^{\infty}F(j\theta)\sum_{k=-\infty}^{\infty}\delta(\omega-k\omega_s-\theta)d\theta$$

$$=\frac{1}{T_s}\sum_{k=-\infty}^{\infty}\int_{-\infty}^{\infty}F(j\theta)\delta(\omega-k\omega_s-\theta)d\theta$$

$$=\frac{1}{T_s}\sum_{k=-\infty}^{\infty}F(j\omega-jk\omega_s)$$

一个连续时间信号经过理想抽样后，其频谱将以抽样频率：

$\omega_s=\frac{2\pi}{T_s}$ 为间隔而重复，也即频谱产生周期延拓。

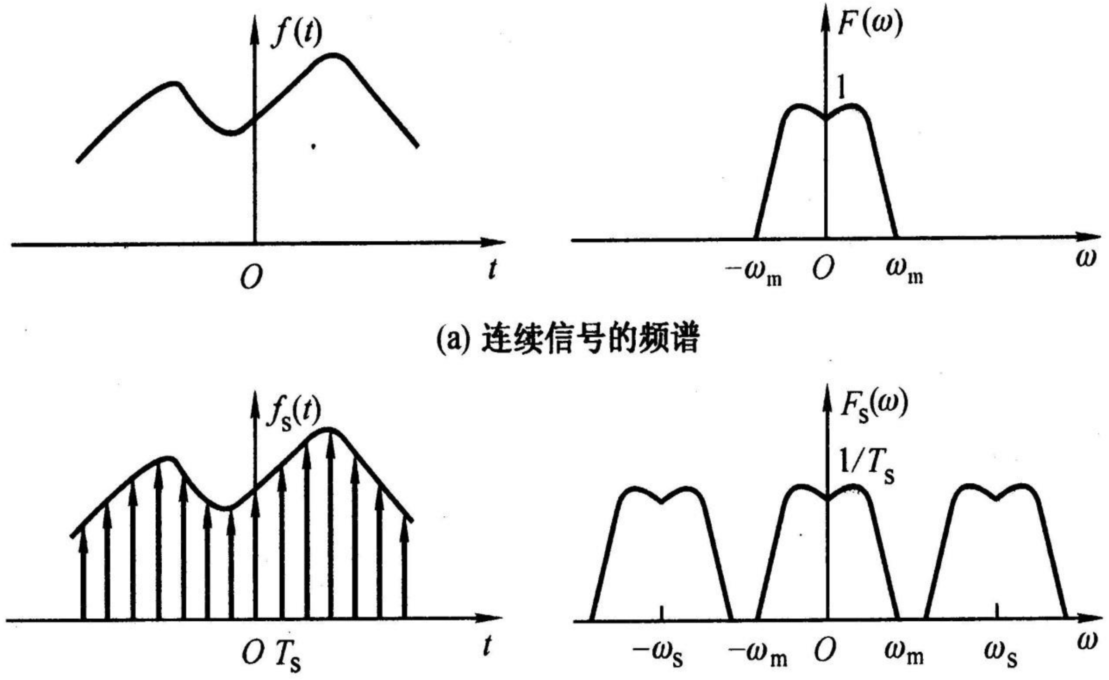
奈奎斯特(Nyquist)抽样定理

- **Nyquist Sampling Theorem**
  - 要从抽样信号中无失真地恢复（重建、还原）原信号，采样频率必须大于等于**两倍**信号谱的最高频率（截止频率）：

$$\omega_s \geq 2\omega_m$$

或

$$f_s \geq 2 f_m$$



(a) 连续信号的频谱

(b) 高抽样率时的抽样信号及频谱(不混叠)

抽样与混叠

- 常用的音频采样率
  - 8kHz、11.025kHz、22.05kHz、16kHz、37.8kHz、44.1kHz、48kHz

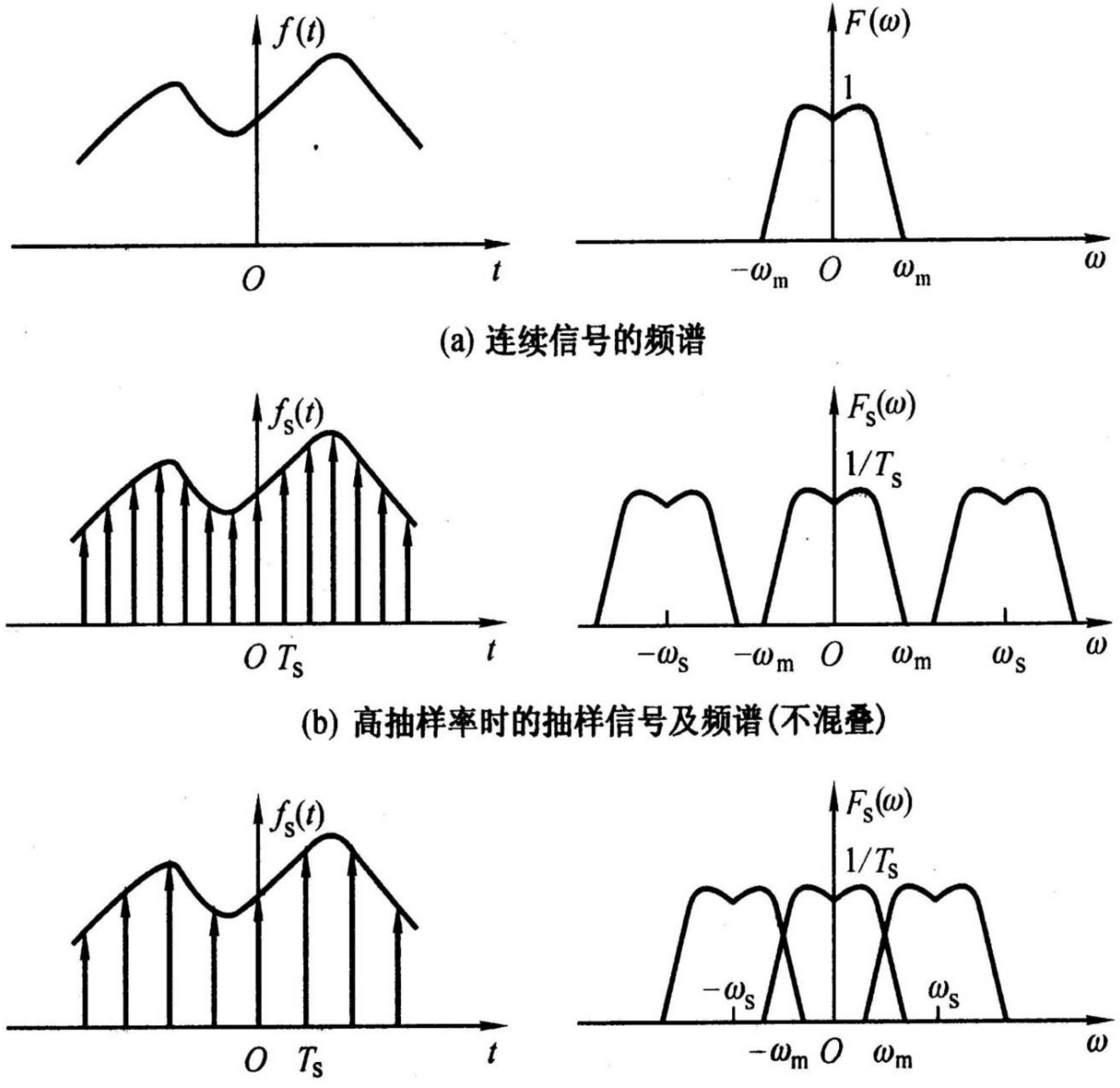
- 重建原信号的必要条件
  - Nyquist 抽样定理：

$$f_s \geq 2 f_m$$

  - 否则，就要发生频谱**混叠**现象

抽样与混叠

$$\omega_s \ge 2\omega_m$$



(a) 连续信号的频谱

(b) 高抽样率时的抽样信号及频谱(不混叠)

(c) 低抽样率时的抽样信号及频谱(混叠)

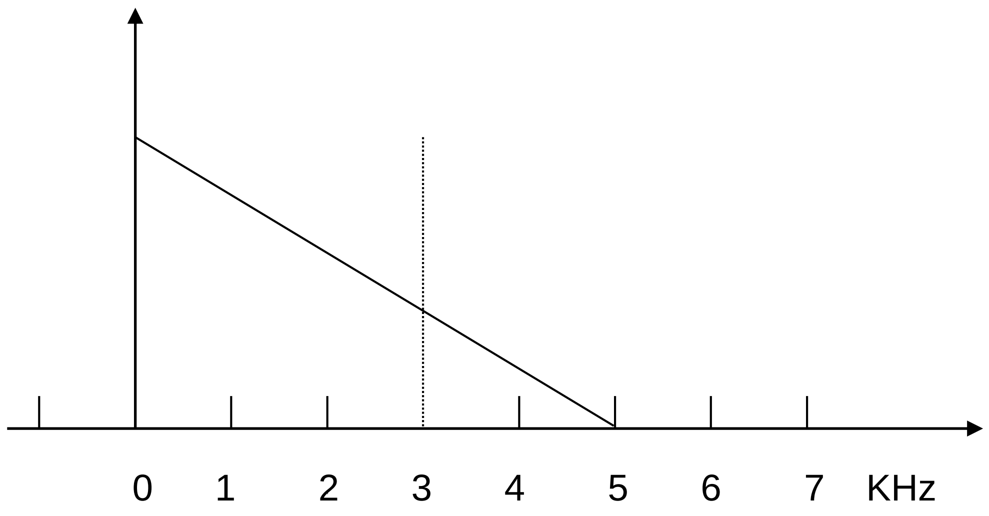
抽样与混叠

- **信号的恢复与频谱混叠（alias）**
  - 要从抽样信号中无失真地恢复（重建、还原）原信号，采样频率必须要大于等于两倍信号谱的最高（截止）频率，否则就会发生频谱混叠（alias）现象
- **折叠频率**
  - 采样频率的一半，称之为折叠频率
  - 当语音信号频谱分布超过折叠频率时，就会被折叠回来，造成频谱的混叠

抽样与混叠

设音频信号的高频截止频率为5kHz，抽样频率为6kHz，

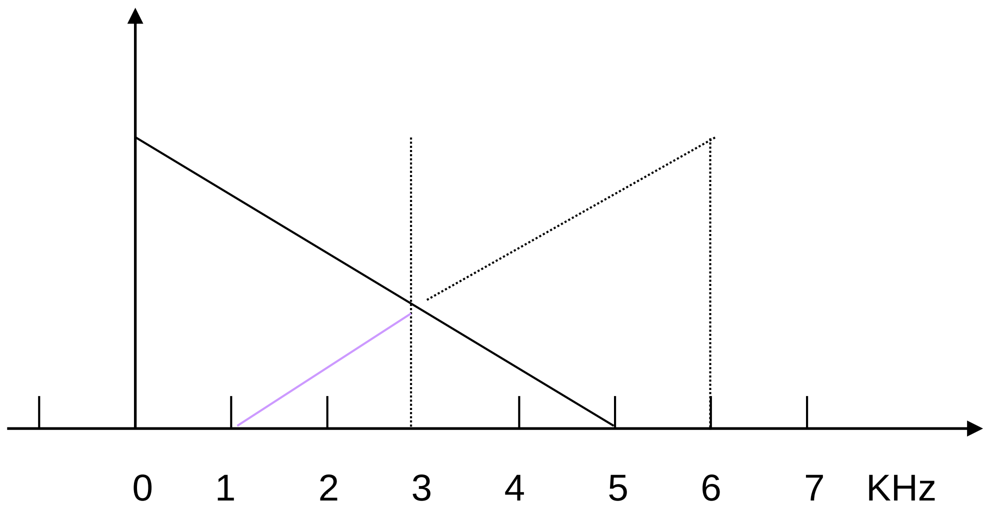
问：2kHz信号中混有哪些频率的信号？



抽样与混叠

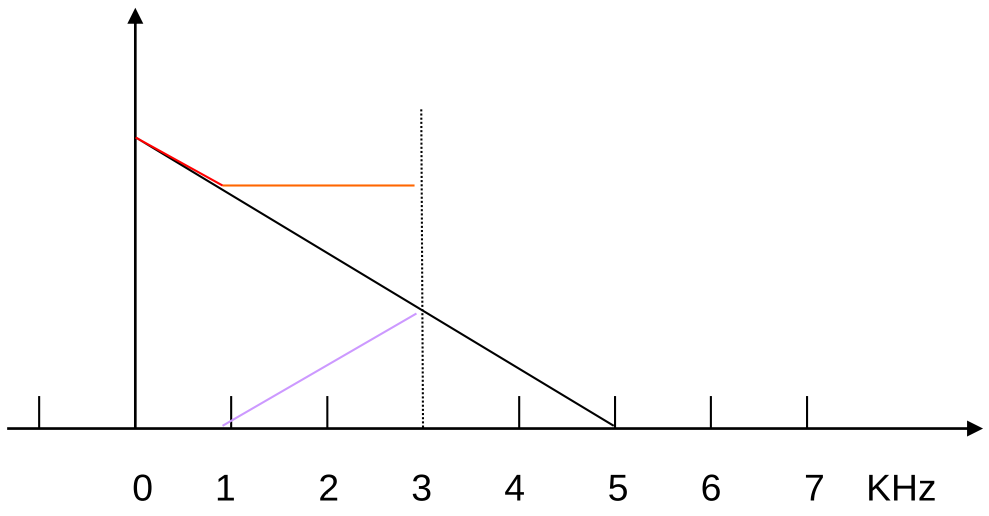
设音频信号的高频截止频率为5kHz，抽样频率为6kHz，

问：2kHz信号中混有哪些频率的信号？

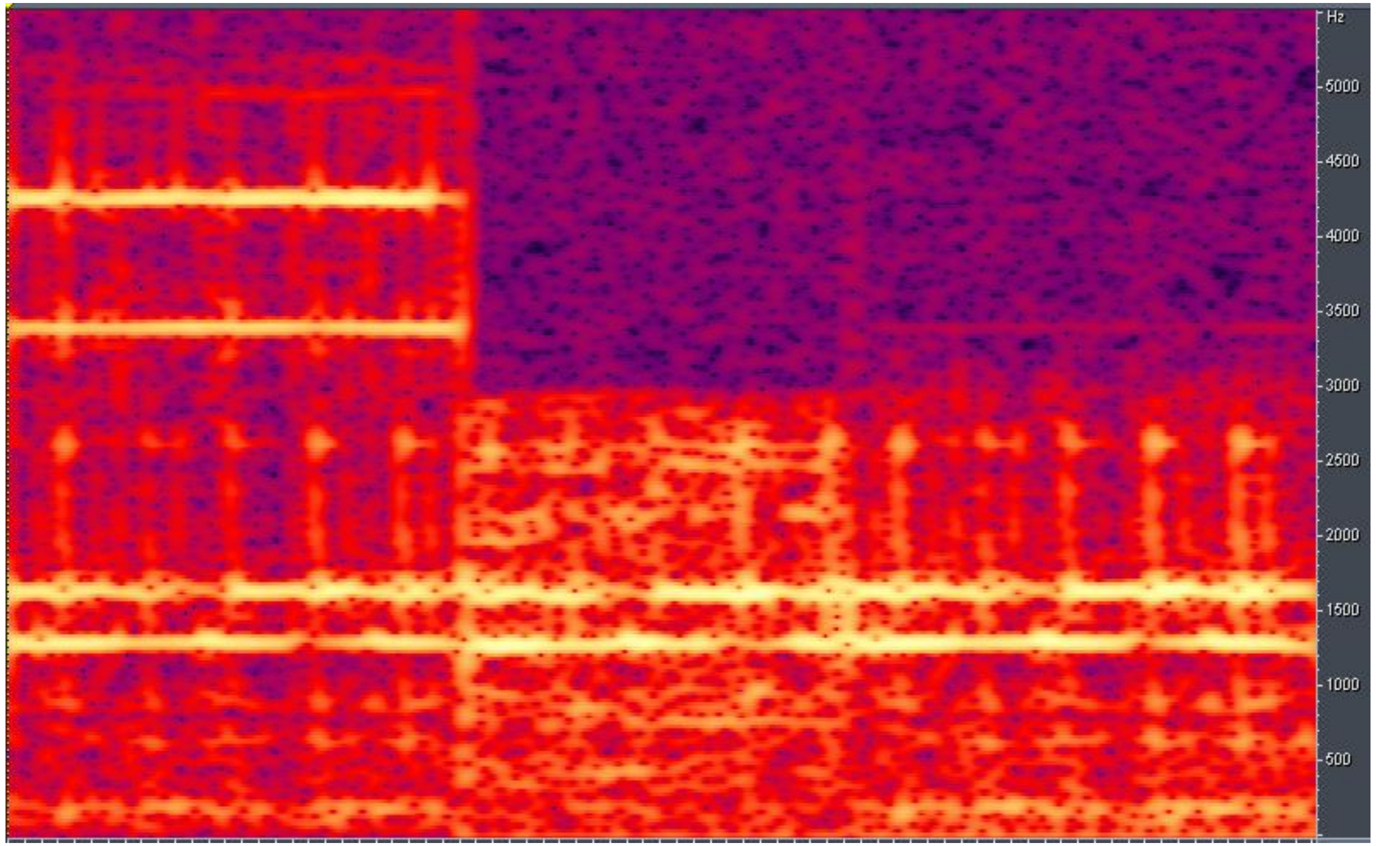


设音频信号的高频截止频率为5kHz，抽样频率为6kHz，

问：2kHz信号中混有哪些频率的信号？



抽样与混叠



11KHz
16 bits

6KHz
16 bit

6KHz
16 bit
3KHz Cutoff

CoolEdit

语音信号抽样

- **理想的数字音频信号的采样率**
  - 达到模拟音频的质量
  - 采样率
    - F_s = 44.1kHz
    - 人耳听觉特性：20Hz ~ 20kHz
    - 奈奎斯特 (Nyquist) 抽样定理：采样频率大于等于2倍信号最大频率（截止频率）
  - 高保真音响
- **桌面计算机语音的采样率**
    - F_s=16kHz
    - 语音的频率范围：60Hz ~ 8kHz
- **电话语音的采样率**
    - F_s=8kHz

量化与噪声：

数字音频在什么情况下质量达到或优于模拟音频？

QUANTIZATION OF SPEECH SIGNAL

量化

量化：为了把抽样序列 $x(n)$ 存入计算机，必须将样值量化成一个有限个幅度值的集合 $\hat{x}(n)$ 。

用二进制数字表示量化后的样值。

用 B 位二进制码字可以表示 2^B 个不同的量化电平。

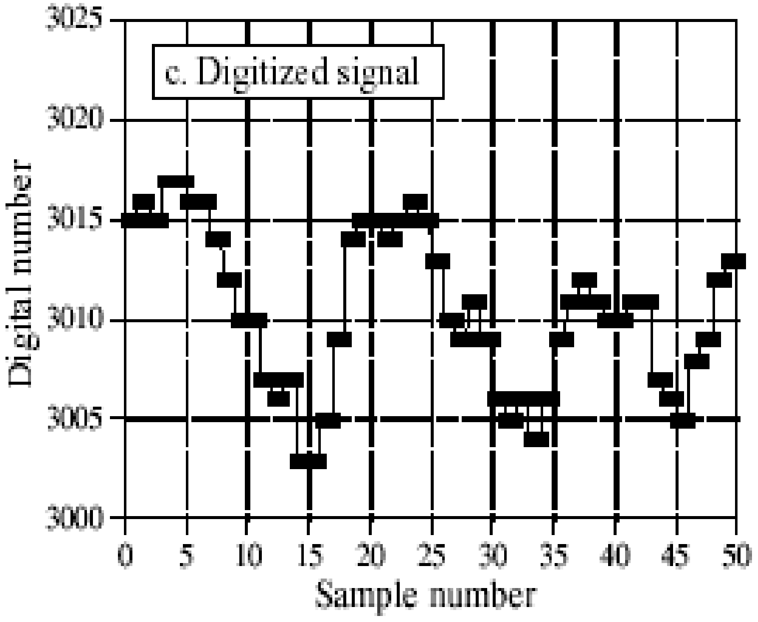
存储数字音频信号的比特率为：

$$I = B \cdot fs \quad \text{(比特/秒)}$$

fs 是抽样率 (抽样/秒)

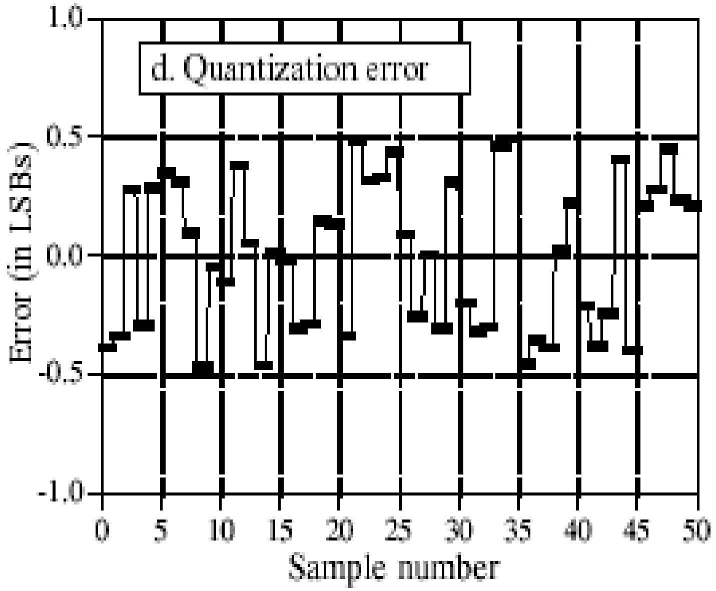
B 是每个样值的比特数 (比特/抽样)

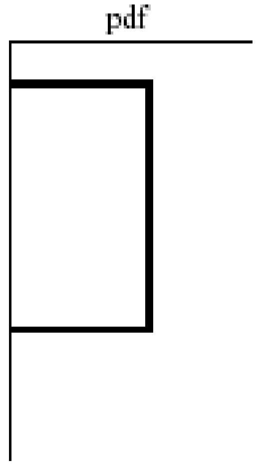
语音信号的量化



量化后的信号

量化误差（量化噪声）





* LSB: Least Significant Bit，量化阶距

量化噪声

量化抽样的过程：先将整个幅度划分成为有限个小幅度（量化阶距）的集合，把落入某个阶距内的样值归为一类，并赋予相同的量化值。

如果量化值是均匀分布的，我们称之为**均匀量化**。设△为量化阶距，量化器的最大范围是 X_{max}，则：

$$\Delta = \frac{2X_{\max}}{2^B}$$

对于小于 $(i+\frac{1}{2})\Delta$，而大于 $(i-\frac{1}{2})\Delta$ 的样值，均规定为相同的量化值 $i\Delta$。

量化样值 $\hat{x}(n)$ 与未量化样值 $x(n)$ 的关系是：

$$\hat{x}(n) = x(n) + e(n)$$

$e(n)$ 是量化误差（量化噪声），$-\frac{\Delta}{2} \le e(n) \le \frac{\Delta}{2}$

量化噪声的特点

1. 语音信号是一个复杂信号，若量化阶距足够小，那么量化噪声与输入信号不相关，即

$$E[x(n)e(n+m)]=0 \qquad m\ \text{为任意值}$$

2. 量化噪声是平稳白噪声过程，其均值为 0，且量化噪声之间不相关，即

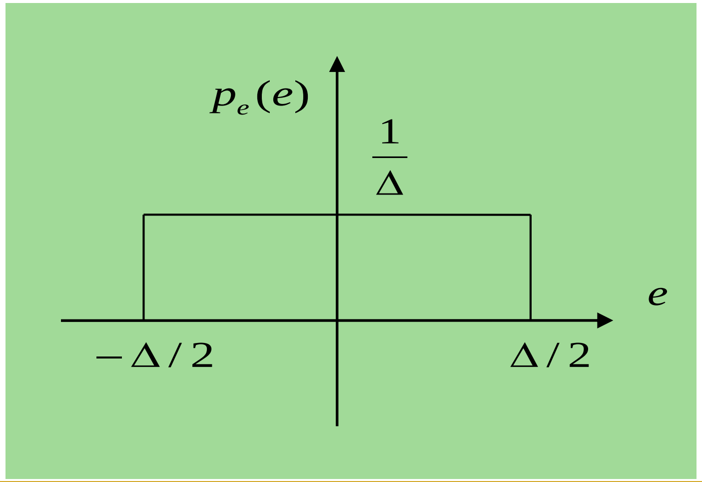
$$E[e(n)e(n+m)]=\sigma_e^2 \quad m=0 \qquad \sigma_e\ \text{量化误差}\ e(n)\ \text{的均方差}$$

$$=0 \qquad \text{其它}$$

3. 对于阶距为△的均匀量化器，量化噪声的幅度分布是均匀的，量化误差的概率密度函数与阶距的关系是:

$$p_e(e)=\frac{1}{\Delta} \qquad -\frac{\Delta}{2}\le e(n)\le\frac{\Delta}{2}$$

$$=0 \qquad \text{其它}$$



量化性能评价

- **SNR: Signal-to-Noise Ratio, 信噪比**
  - 信号与量化噪声的功率比

$$SNR = \frac{E[x^2(n)]}{E[e^2(n)]} = \frac{E\{[x(n) - E(x(n))]^2\}}{E\{[e(n) - E(e(n))]^2\}} = \frac{\sigma_x^2}{\sigma_e^2}$$

  - 均匀量化器
    - 假设量化器量化范围是$2X_{max}$ (X_{max} 为峰值)。量化器位数是B，则均匀量化器的阶距△为:

$$\Delta = \frac{2X_{\max}}{2^B}$$

    - 量化噪声具有均匀幅度分布，则：

$$\sigma_e^2 = \int_{-\Delta/2}^{\Delta/2} \frac{1}{\Delta} e^2(n) de = \frac{1}{3\Delta} e^3(n)\Big|_{-\Delta/2}^{\Delta/2} = \frac{\Delta^2}{12} = \frac{X_{max}^2}{(3)\cdot 2^{2B}}$$

$$\mathrm{SNR} = \frac{\sigma_x^2}{\sigma_e^2} = \frac{3\cdot 2^{2B}}{(\frac{X_{max}}{\sigma_x})^2}$$

量化性能评价

信噪比用分贝表示：

$$SNR(dB) = 10\log[\frac{\sigma_x^2}{\sigma_e^2}] = 4.77 + 6.02B - 20\log[\frac{X_{\max}}{\sigma_x}]$$

假设输入信号均方差 σ_x 的四倍刚好是 $X_{\max}$，

即 $X_{\max} = 4\sigma_x$，则上式变为：

$$SNR(dB) = 6.02B - 7.27$$

我们常用此公式近似计算量化器的信噪比，如：

B=6　SNR(dB)=28.85

B=8　SNR(dB)=40.89

量化器每增加一位编码，信噪比增大 6dB。

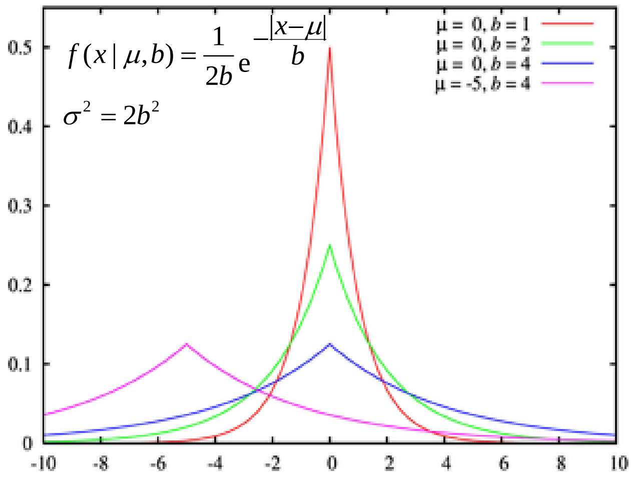
在高保真的音响系统中，信噪比大于 90dB。

为达致90dB左右的信噪比，量化精度B=16

量化性能评价

- 语音信号的幅度分布
  - Laplace distribution
  - The *pdf* function

$$p(x)=\frac{1}{\sqrt{2}\sigma_x}e^{-\frac{\sqrt{2}|x|}{\sigma_x}}$$



  - The probability for amplitude gets over $4\sigma_x$ is only 0.35%

$$p(x\,|\,x>4\sigma_x)=0.35\% \quad \Rightarrow \quad X_{max}\cong 4\sigma_x$$

语音信号数字化：总结

- 为了达到模拟音频信号的质量，理想的数字音频信号的采样率和量化精度是多少？
  - 采样率
    - Fs = 44.1kHz
    - 人耳听觉特性：20Hz ~ 20kHz
    - 奈奎斯特 (Nyquist) 抽样定理：采样频率大于等于2倍信号最大频率 (截止频率)
  - 量化精度
    - B = 16bit
    - 量化误差与量化性能：SNR(dB)=6.02B-7.27
    - 高保真音响系统 (模拟音频信号)，其信噪比SNR>=90dB

语音信号的短时分析：

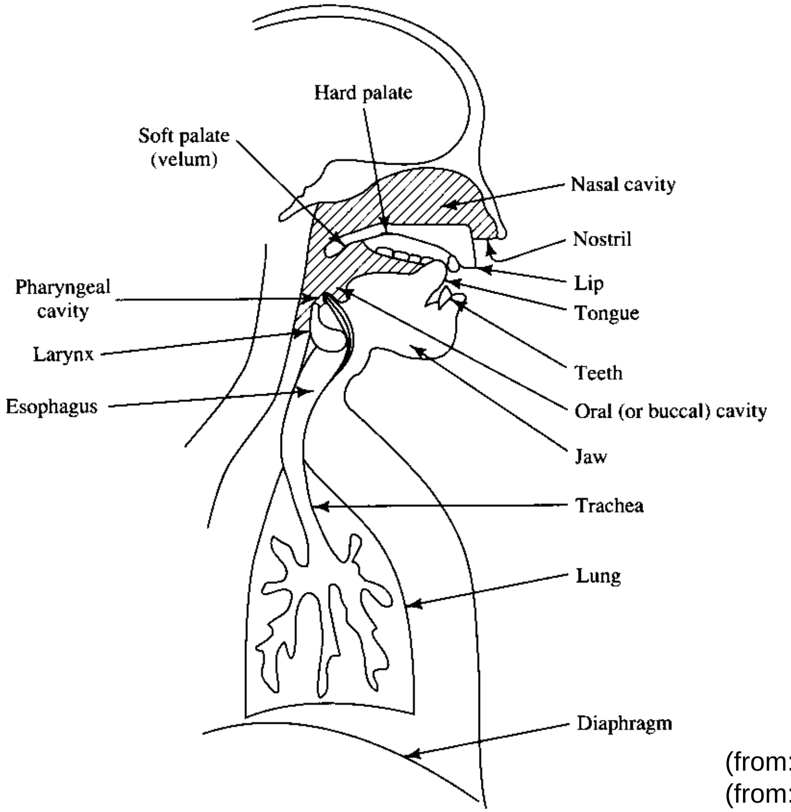
短时分析对什么产生影响？

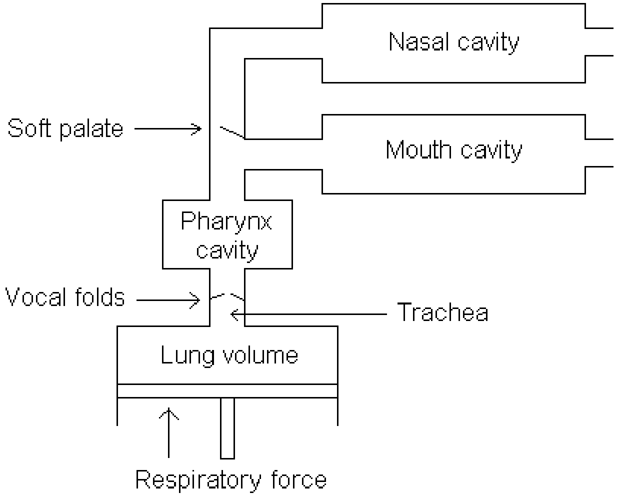
SHORT-TIME PROCESSING OF SPEECH SIGNAL

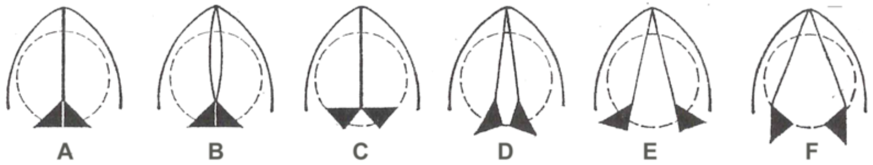
Speech Production: 语音产生

Speech Production

- Speech is produced by air-pressure waves emanating from the mouth and nostrils of a speaker
- 由于声门 (glottis) 的肌肉张力，加上由肺部压迫出来的空气，就会造成声门的快速打开与关闭，这一疏一密的空气压力，即为语音源头，再经过声道、口腔、鼻腔的共振，就会产生不同声音。

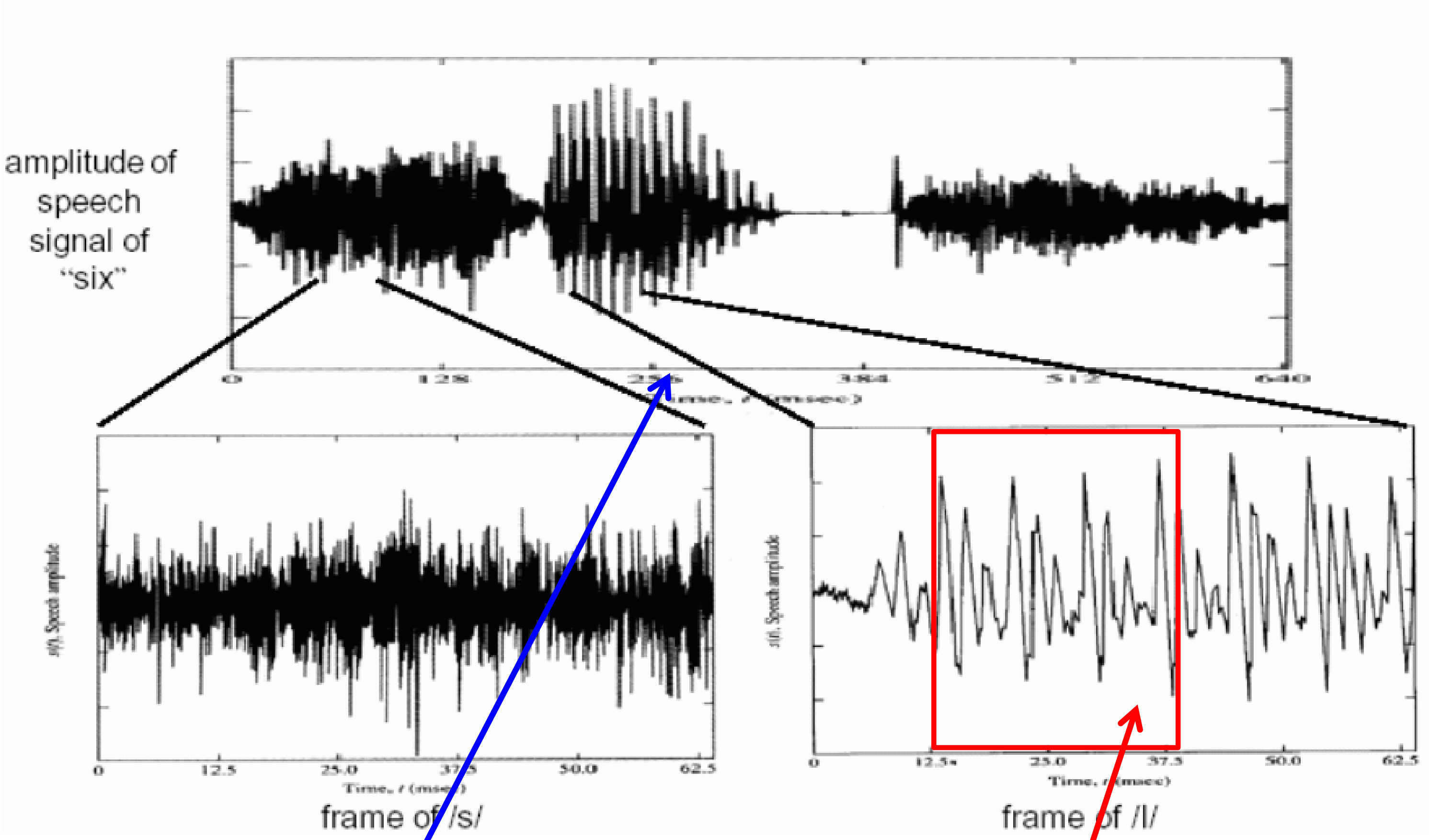






(from: http://imp.lss.wisc.edu/~jrvalent/AIS/Grammar/Phonology/Phonol002a.html)
(from: http://clas.mq.edu.au/phonetics/phonetics/introduction/index.html)

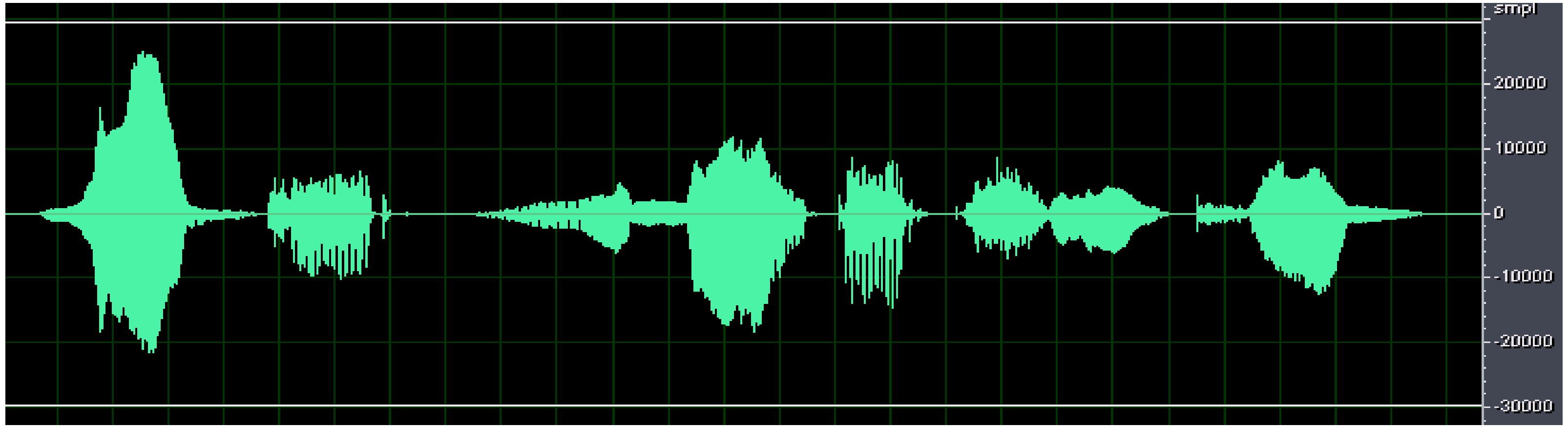
An Example Speech



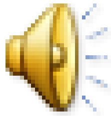
语音信号是长时非平稳的

但在10ms~30ms的时间段内，
语音信号又具有短时平稳的周期性

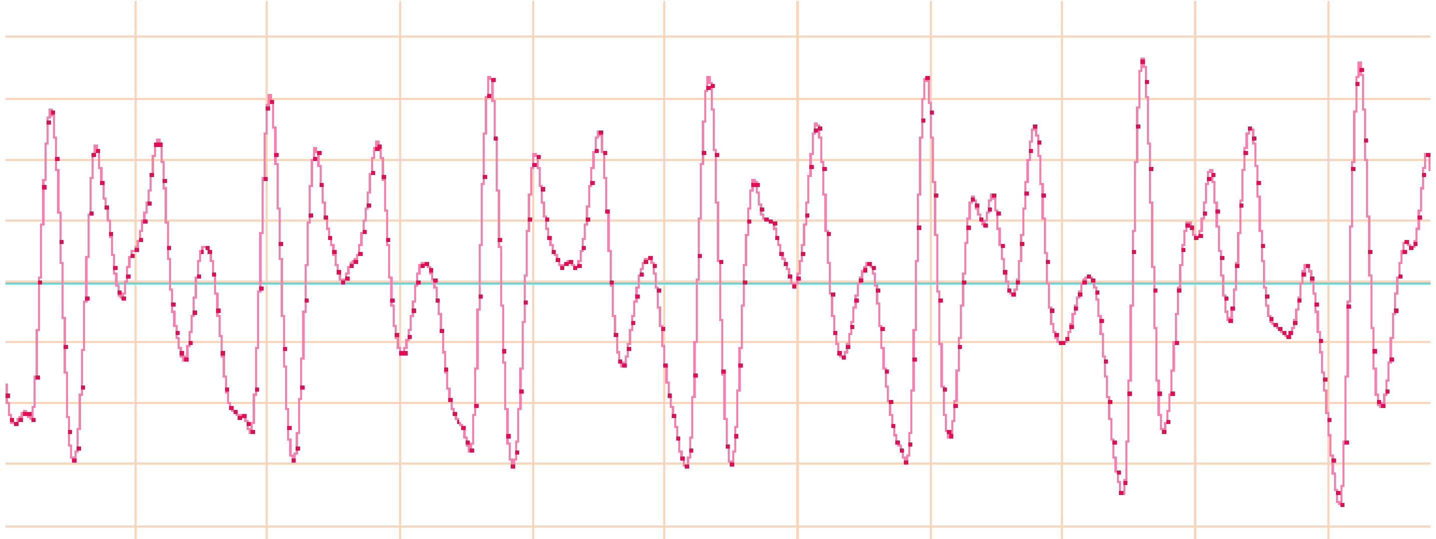
Another Example



望着无奈的秋天



Wav文件格式：16KHz, 16Bit, 单声道



Short-time Speech Processing

- **语音信号是一种典型的非平稳信号**
  - 由于人自身发音器官运动的过渡性特点
- **但是，语音信号具有短时平稳的周期性**
  - 10ms-30ms的时间段内
  - 短时平稳
  - 具有一定的周期性

几乎所有的语音信号处理方法都是基于语音信号短时平稳的假设！

Short-time Speech Processing

如何进行
短时分析？

几乎所有的语音信号处理方法都是基于语音信号短时平稳的假设！

Windowing: 加窗处理

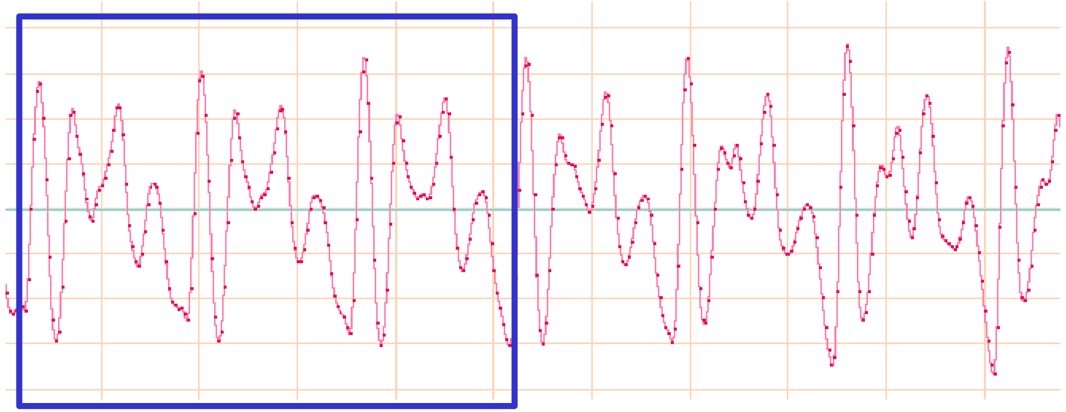
- 短时分析的最基本手段是对语音加窗
  - 用一个有限长度的窗序列截取一段语音信号进行分析

$$s_w(n)=\sum_{m=-\infty}^{\infty} s(m)w(n-m)=s(n)*w(n)$$

  - 窗函数可以按时间方向滑动，以便分析任一时刻附近的信号

$$s_w(n_0)=\sum_{m=n_0}^{n_0-(N-1)} s(m)w(n_0-m)$$

  - 加窗运算实际上是一种卷积运算



Window Function: 窗函数

- **Rectangular Window: 矩形窗/方窗**

$$w(n)=\begin{cases}1 & 0\le n\le N-1\\ 0 & n<0 \text{ or } n<N\end{cases}$$

- **Hamming Window: 哈明窗**

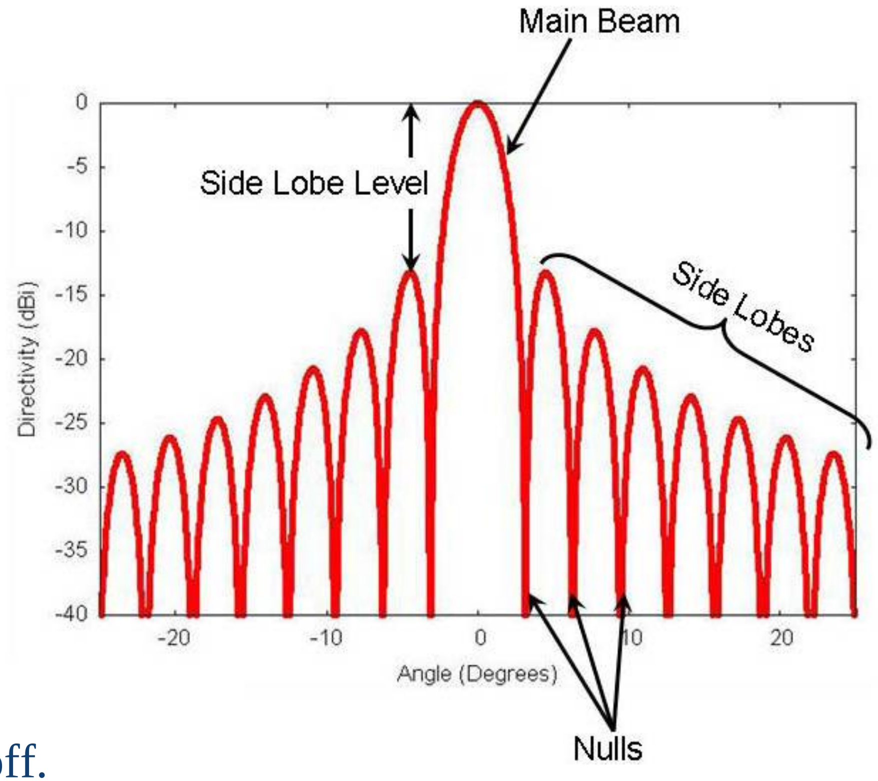
$$w(n)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{N-1}) & 0\le n\le N-1\\ 0 & n<0 \text{ or } n<N\end{cases}$$

- **Hann Window: 汉宁窗**

$$w(n)=\begin{cases}0.5(1-\cos(\frac{2\pi n}{N-1})) & 0\le n\le N-1\\ 0 & n<0 \text{ or } n<N\end{cases}$$

Frequency Response: 频率响应

- **Frequency Response of Window Function**
  - The Width of the Main Lobe / Main Beam (主瓣宽度)
    - ***Ideally:*** The main lobe will be **narrow** (corresponding to high frequency resolution).
    - 与窗长成反比
  - The Side Lobe Level (旁瓣高度)
    - The attenuation at the maximum height of a side lobe, generally the first side lobe.
    - ***Ideally***: The first side lobe will be **low** (corresponding to noise suppression).
  - The Side Lobe Fall-off (旁瓣衰减速度)
    - The rate at which the peaks of the side lobes fall-off.
    - ***Ideally***: The side lobes fall-off **rapidly**.

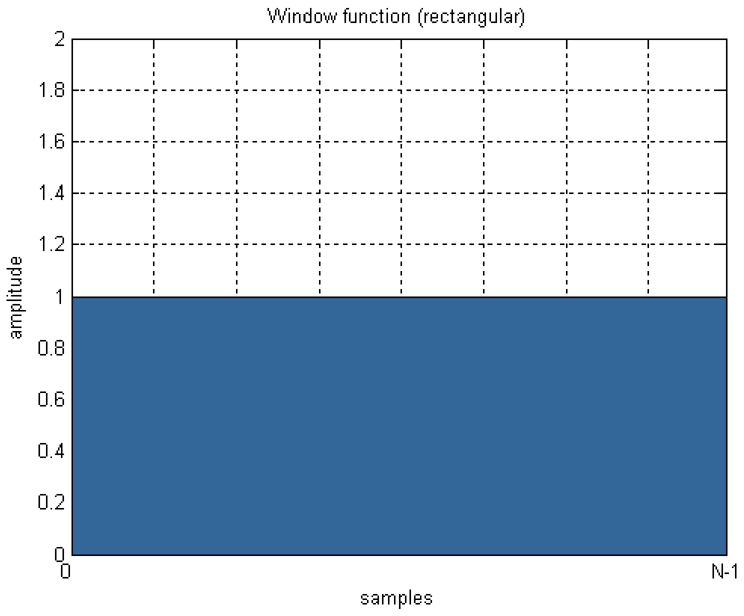


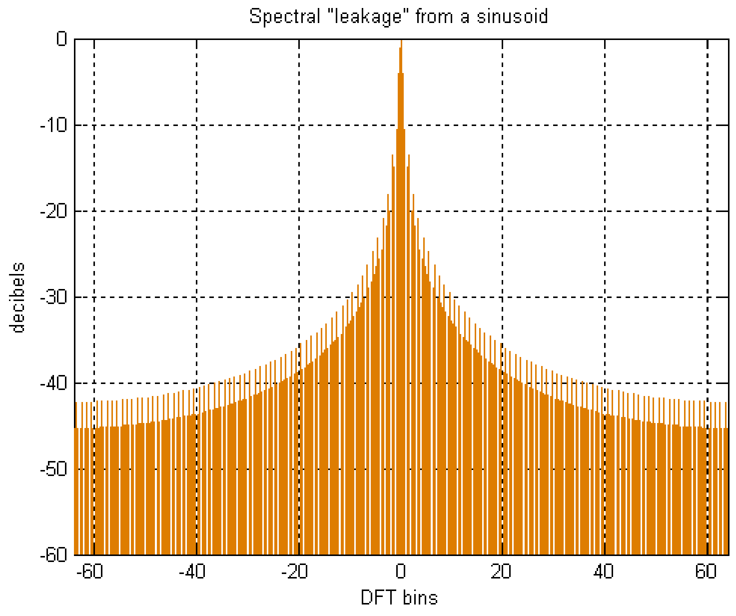
(from: http://en.wikipedia.org/wiki/Spectral_leakage)

Rectangular Window: 矩形窗/方窗

- 矩形窗的频率响应幅度特性

$$w(n)=\begin{cases}1 & 0\le n\le N-1\\ 0 & n<0\ \text{or}\ n<N\end{cases}$$

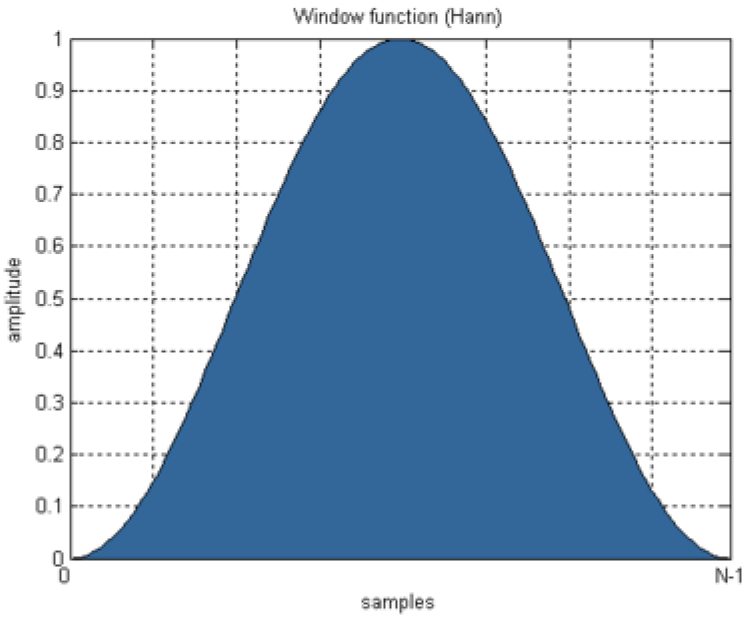


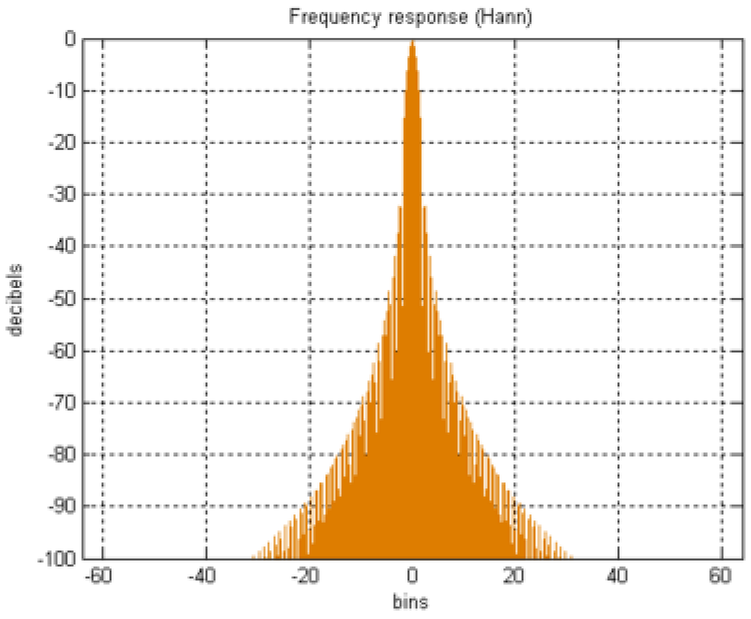


Hann Window: 汉宁窗

- 汉宁窗的频率响应幅度特性

$$w(n)=\begin{cases}0.5(1-\cos(\frac{2\pi n}{N-1})) & 0\le n\le N-1\\ 0 & n<0 \text{ or } n<N\end{cases}$$

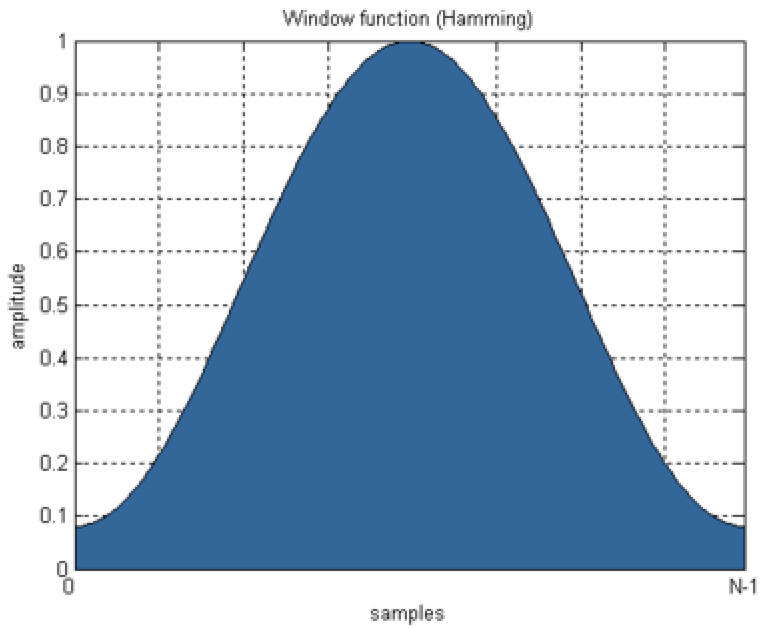


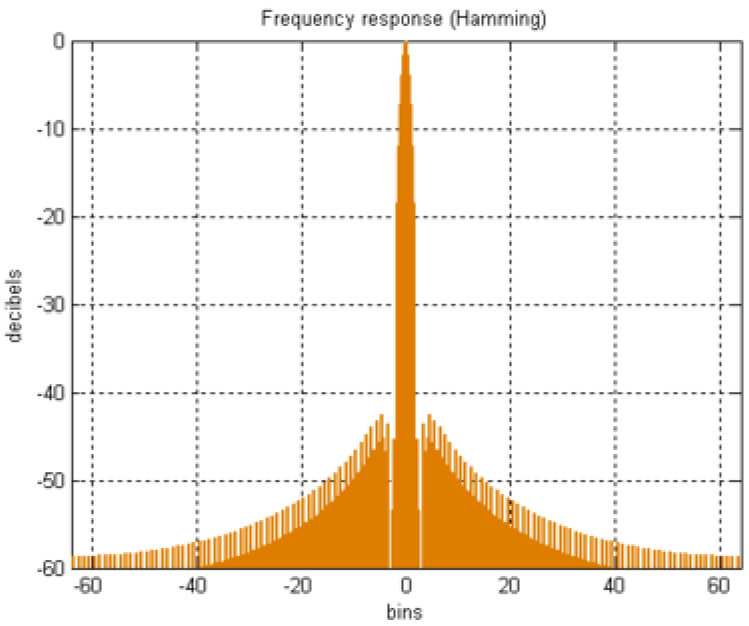


Hamming Window: 哈明窗

- 哈明窗的频率响应幅度特性

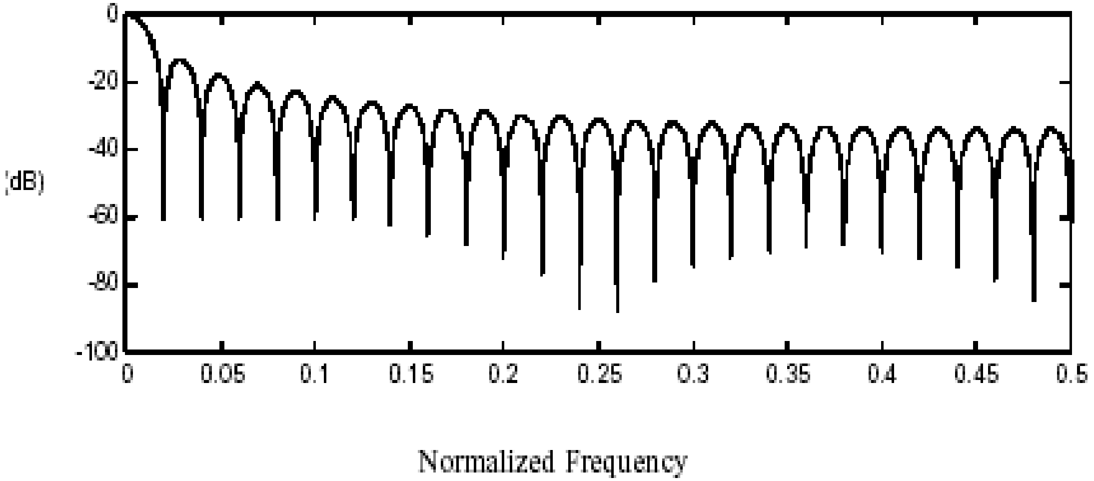
$$w(n) = \begin{cases} 0.54 - 0.46\cos(\dfrac{2\pi n}{N-1}) & 0 \le n \le N-1 \\ 0 & n < 0 \text{ or } n < N \end{cases}$$





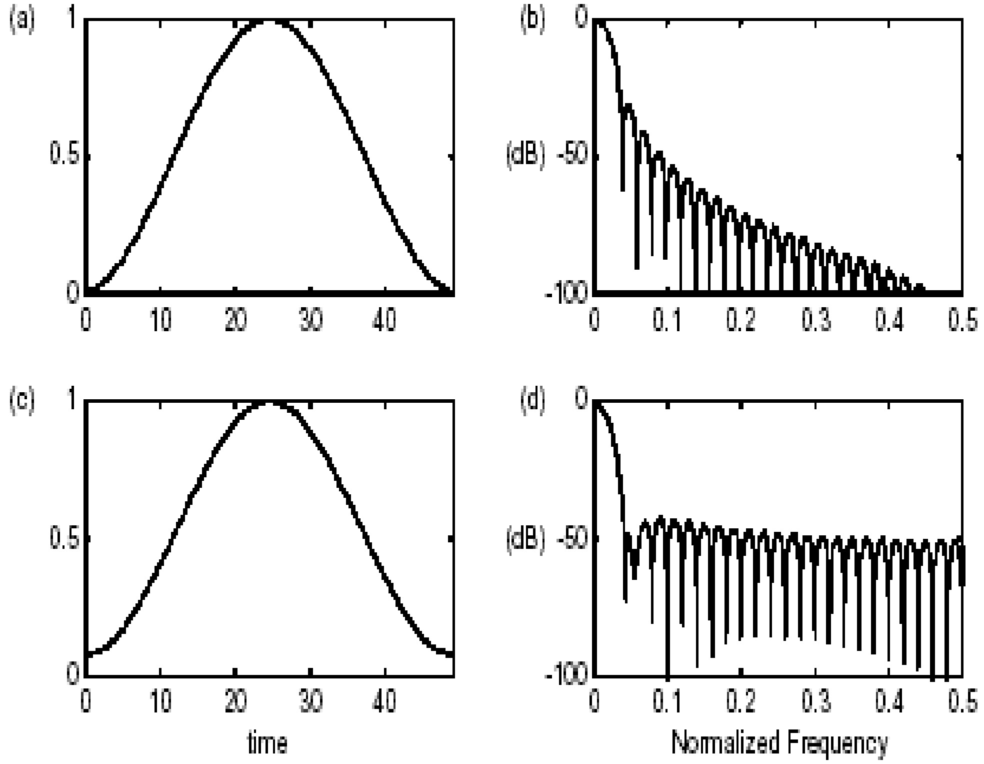
Frequency Response: 频率响应

矩形窗频率响应幅度特性:



汉宁窗频率响应幅度特性:

哈明窗频率响应幅度特性:



语音信号的时域处理：

如何进行特征参数的计算？

特征采样的采样频率如何确定？

TIME-DOMAIN PROCESSING OF SPEECH SIGNAL

特征计算与短时处理

窗函数对短时处理的影响

- 加窗处理等于对语音特性进行了低通滤波：
  - 矩形窗的截止频率：$f_c=f_s/N$
  - 哈明窗的截止频率：$f_c=2f_s/N$
    - 窗特性的影响
    - 窗长的影响

语音特征的采样频率如何确定？

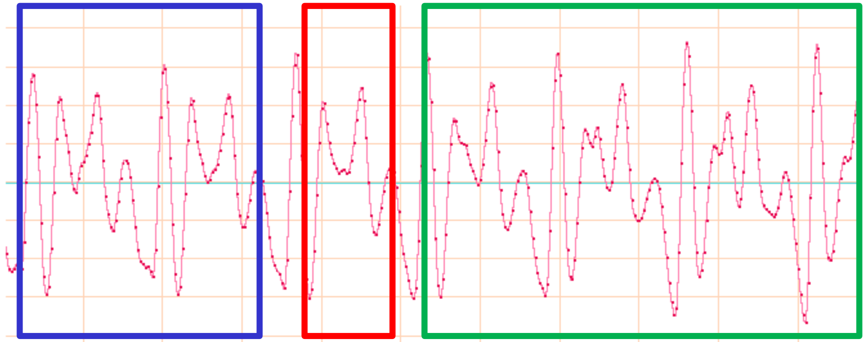
- 窗移如何确定？
  - 窗移的长度不等于窗长
  - 窗移的长度应小于等于1/2窗长

Window Length: 窗长的选择

- 窗长对短时能量计算的影响

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

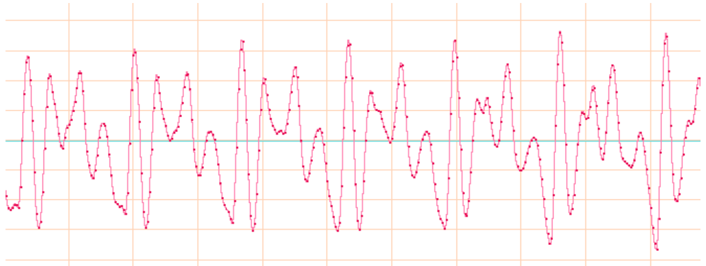
$$h(n) = w^2(n)$$



- 窗长的选择
  - 短时能量可以看作语音信号的平方通过一个冲激响应为 $h(n)$ 的线性滤波器后的输出
  - 窗太长
    - 平滑作用明显，短时能量曲线随时间变化缓慢，不能体现语音变化
  - 窗太短
    - 短时能量随时间变化剧烈，无法得到平滑的能量函数
  - 窗长的选择应该包含1~7个周期
    - 因男女老少基音周期差异大，折衷选择：10ms~30ms作为窗长

语音信号的短时分析

- 语音信号是一种典型的非平稳信号
  - 由于人自身发音器官运动的过渡性特点
- 语音信号具有短时平稳的周期性
  - 10ms-30ms的时间段内
  - 短时平稳
  - 具有一定的周期性



- 语音信号的短时分析
  - 基本手段是对语音加窗：矩形窗、哈明窗
  - 确定窗长：10ms-30ms，一般取30ms
  - 确定窗移：小于等于1/2窗长，一般取1/3窗长

短时能量/短时平均幅度

- 短时能量

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=n}^{n+N-1} [x(m)w(n-m)]^2$$

- 短时平均幅度

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|\, w(n-m) = |x(n)| * w(n)$$

- 应用：
  - 能量是语音的一个重要特性
  - 区分清音和浊音
    - 清音的能量较小
    - 浊音的能量较大

短时平均过零率

- 过零
  - 时域波形穿过坐标轴，表现在离散信号序列上是相邻采样值异号
- 短时过零率
  - 单位时间内过零发生的次数称作短时过零率

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

$$= |\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]| * w(n)$$

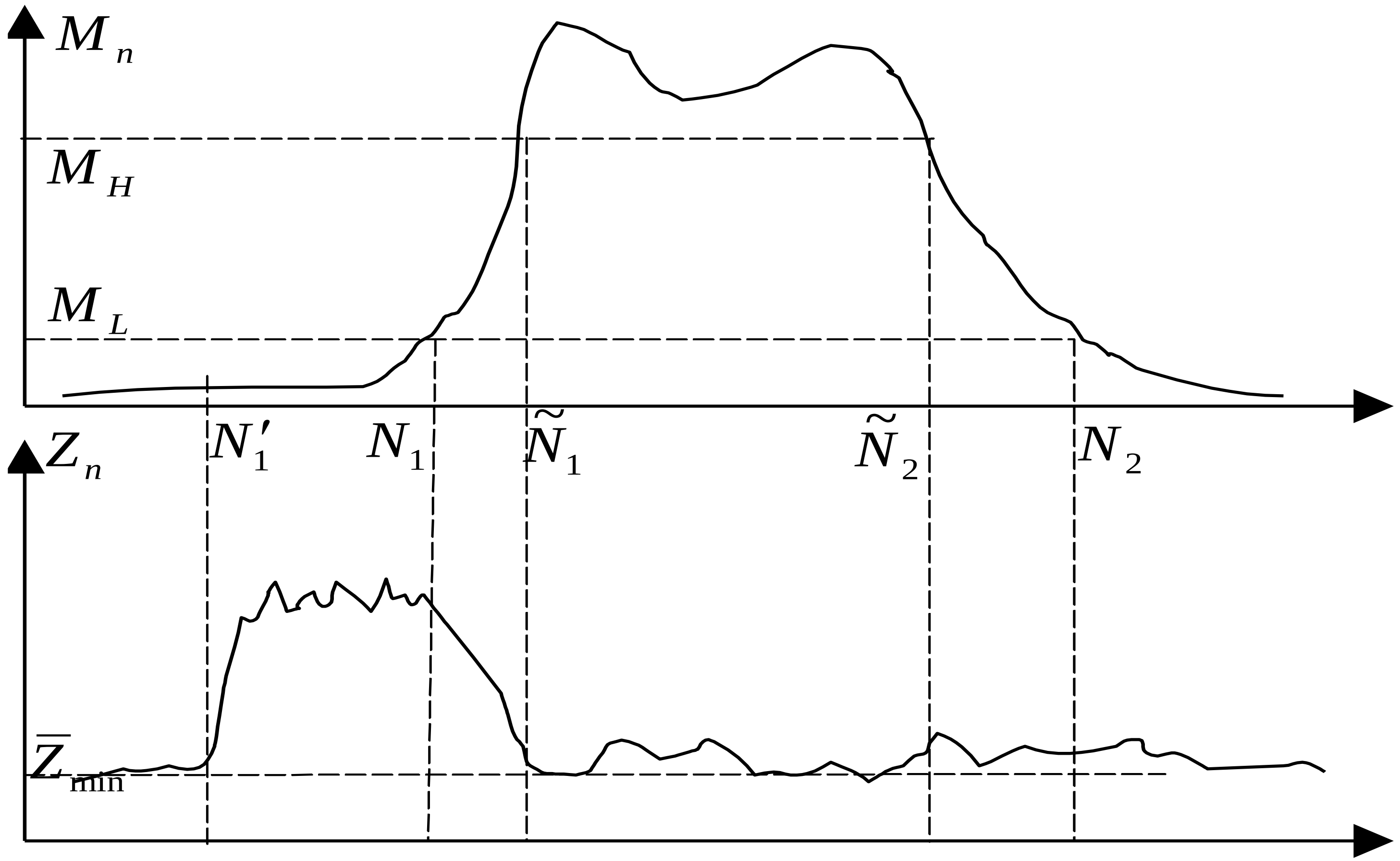
  - 短时过零率对噪声的存在非常敏感
    - 为避免“虚假”的过零，提高过零率计算的鲁棒性，引入门限：|T|

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

- 应用
  - 区分有声和无声（噪声）
    - 有声的短时平均过零率大
    - 无声（噪声）的短时平均过零率小

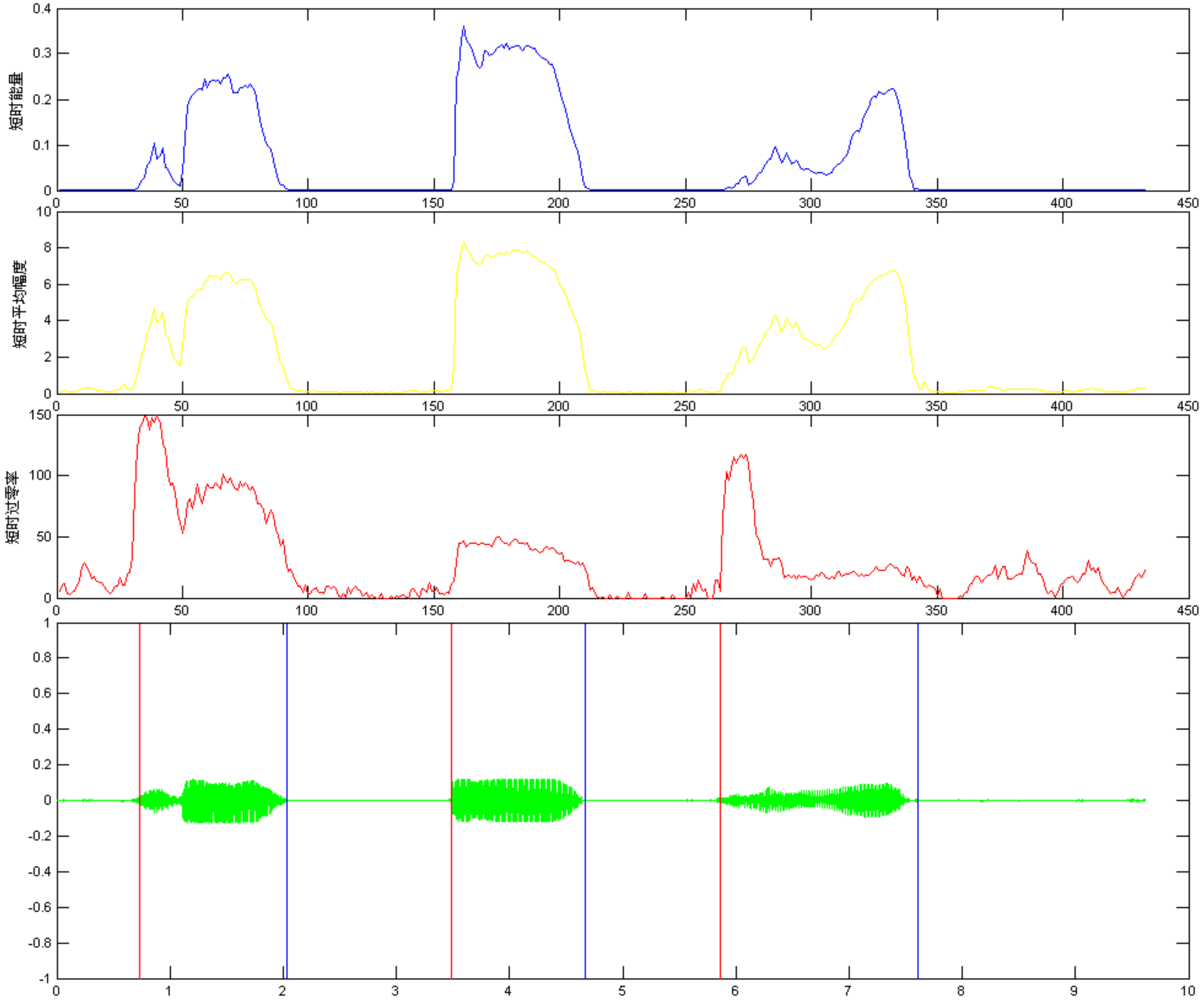
语音信号的端点检测

- **语音的端点检测：双门限法**
  - 语音信号短时特征的一种应用

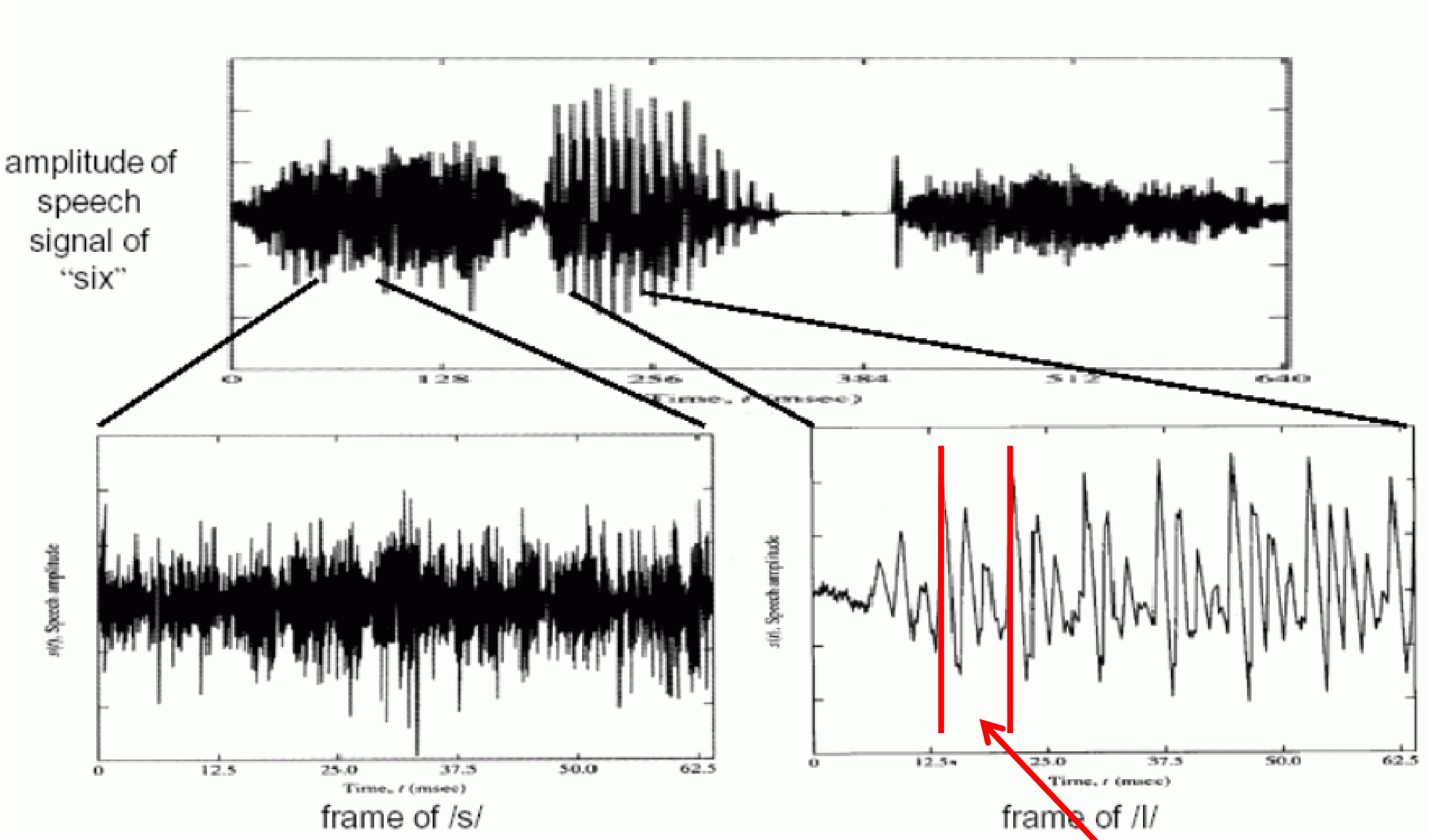


语音信号的端点检测

- 语音的端点检测：双门限法
  - 语音“7，8，9”的端点检测



Fundamental Frequency: 基频



- **Unvoiced Speech: /s/, /k/**
- **Voiced Speech: /i/**

浊音的波形呈现周期性
声带振动的频率：基频 F_0

自相关函数

- 相关分析
  - 常用的时域波形分析方法

- 自相关函数

$$R(k)=\sum_{m=-\infty}^{\infty}\left[x(m)\cdot x(m+k)\right]$$

- 特性
  - 1) 自相关函数是偶函数：$R(k) = R(-k)$
  - 2) $k = 0$时函数取得最大值，取值为信号的能量
  - 3) 如果原序列是周期为 T 的周期信号，那么自相关函数也是周期为 T 的周期函数：$R(k) = R(T+k)$

短时自相关函数

$$R_n(k)=\sum_{m=-\infty}^{\infty}\left[x(m)\cdot w(n-m)\cdot x(m+k)\cdot w(n-(m+k))\right]$$

根据偶函数的特性：

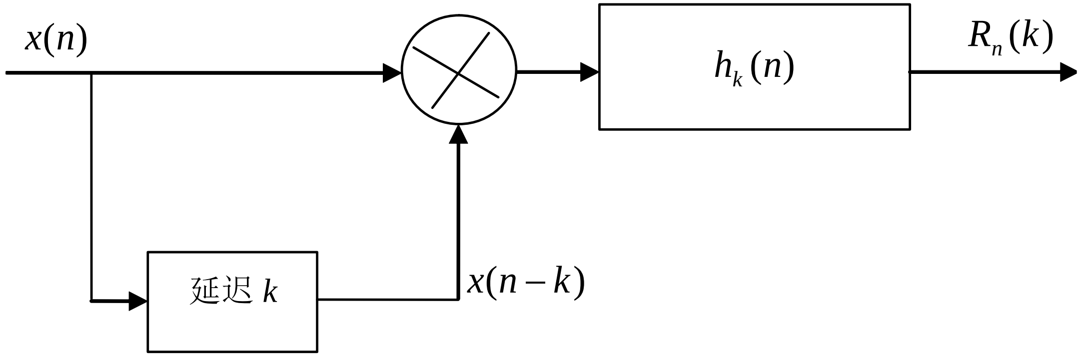
$$R_n(k)=R_n(-k)=\sum_{m=-\infty}^{\infty}\left[x(m)w(n-m)x(m-k)w(n-(m-k))\right]$$

定义：

$$h_k(n)=w(n)w(n+k)$$

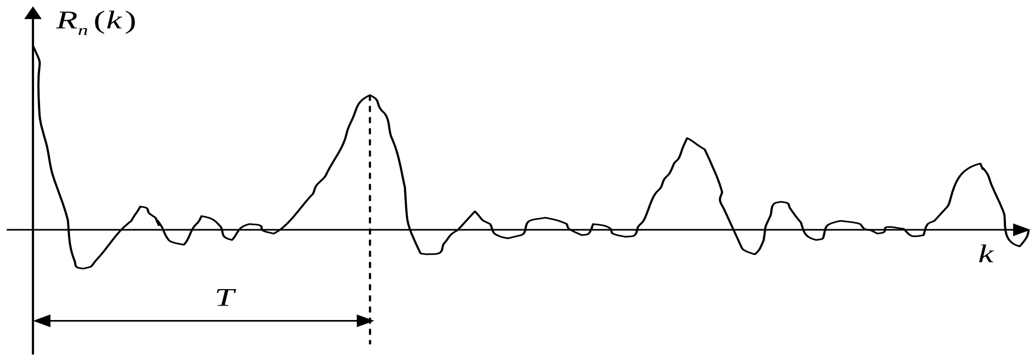
有：

$$R_n(k)=\sum_{m=-\infty}^{\infty}[x(m)x(m-k)]h_k(n-m)=[x(n)x(n-k)]*h_k(n)$$



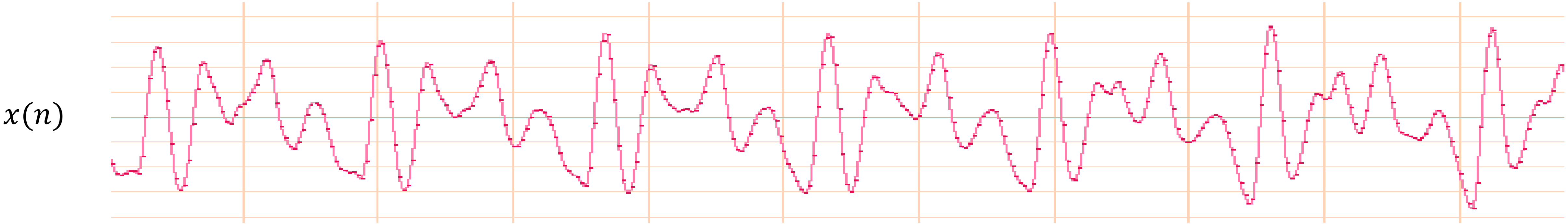
Pitch Detection：短时基音周期估计

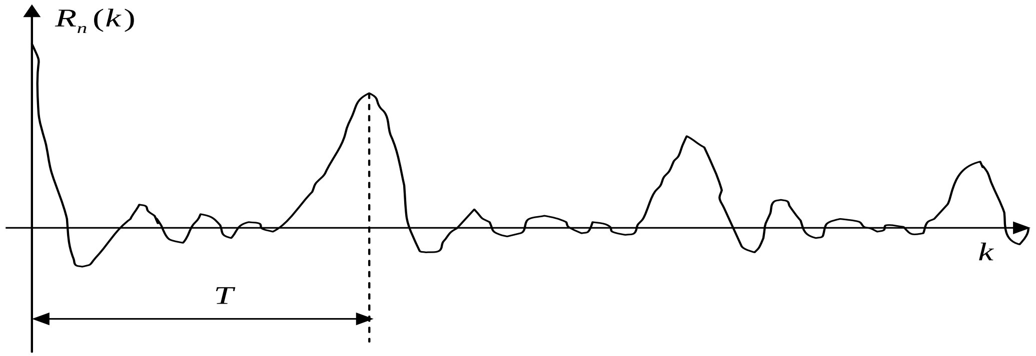
- 基于自相关函数的基音周期估计
  - 预处理
    - 除去声道共振峰对基音周期的干扰：共振峰频率一般大于1000Hz
    - 带通滤波器滤波：60Hz～900Hz
  - 基于短时自相关函数的估计算法



    - 短时自相关函数在基音周期的各个整数倍点上有很大的峰值
    - 第一最大峰值点与零点的距离就是**基音周期**
    - 自相关函数窗长的选择
      - 至少应大于两个基音周期才能有较好的效果
      - 语音频率下限约为50Hz，基音周期最长约为20ms，因此窗长应大于40ms

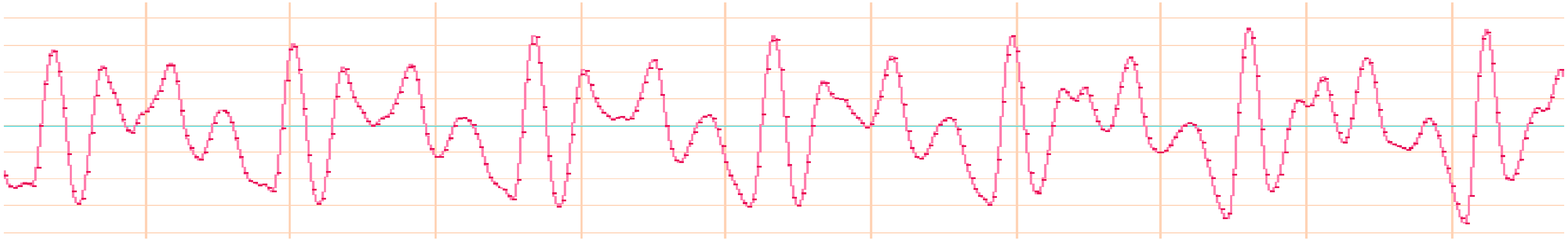
Pitch Detection：短时基音周期估计



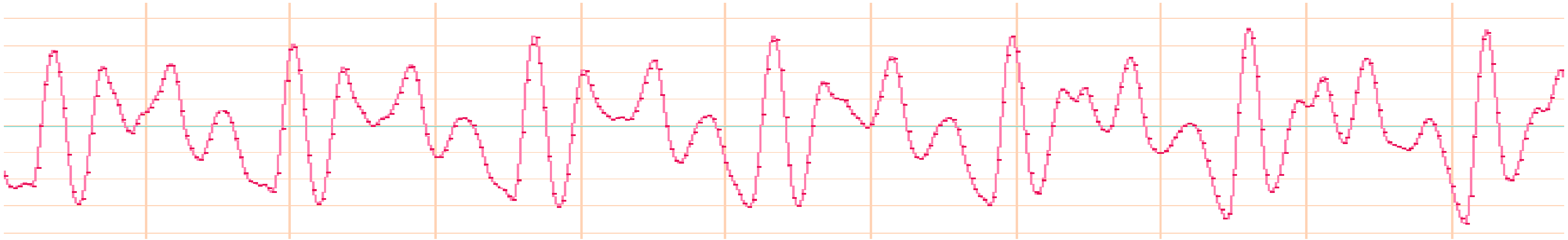


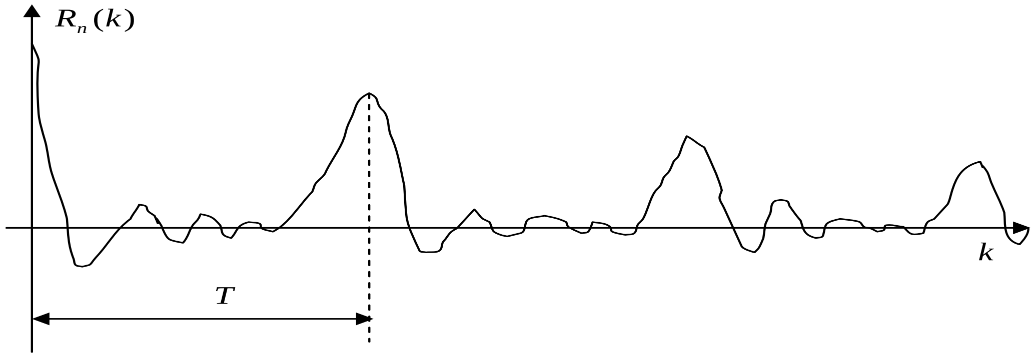
Pitch Detection：短时基音周期估计

$x(n)$



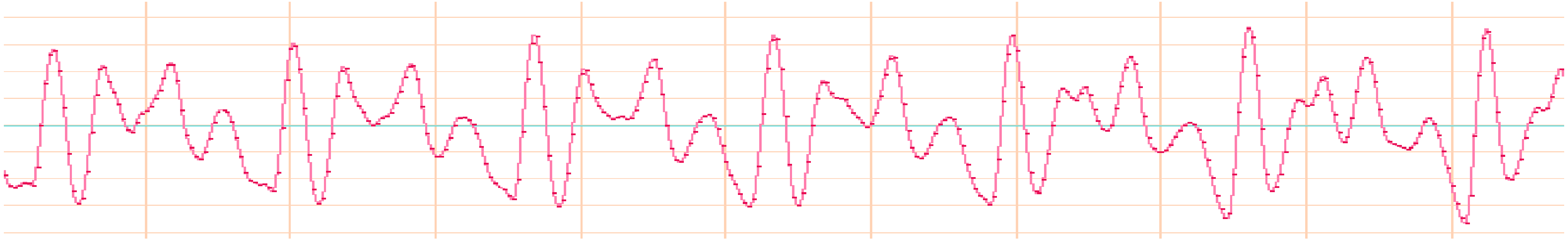
$x(n-k)$



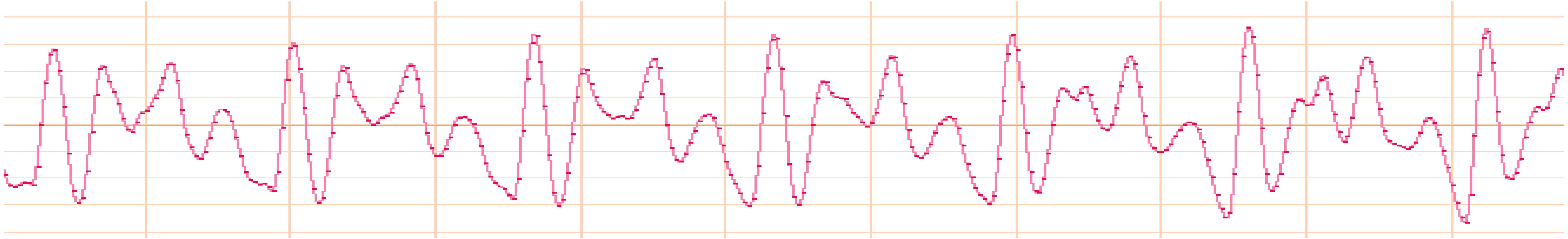


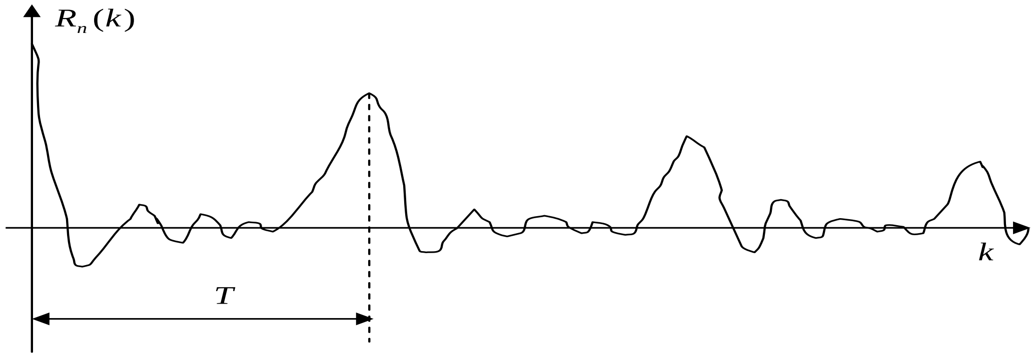
Pitch Detection：短时基音周期估计

$x(n)$



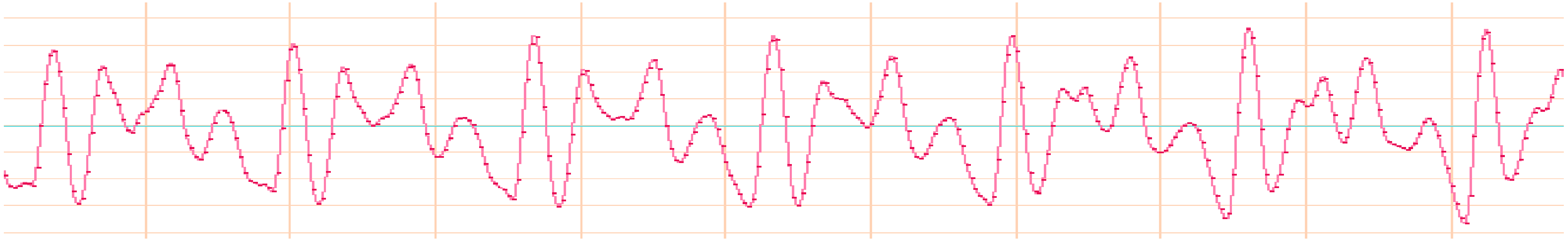
$x(n-0)$
$k=0$



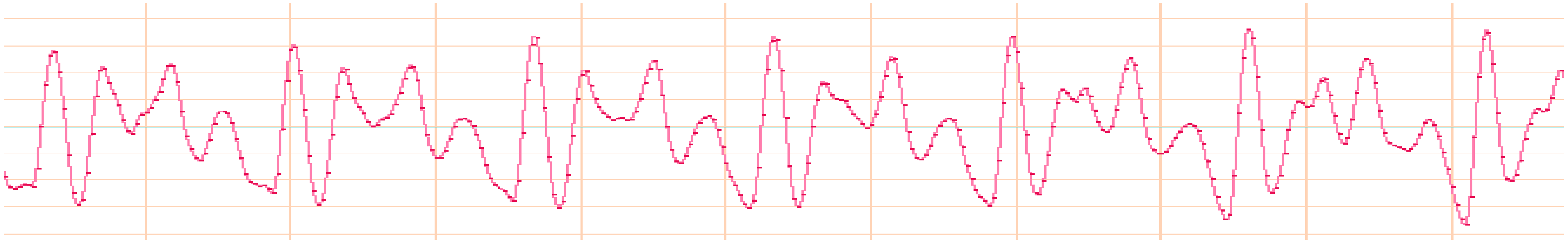


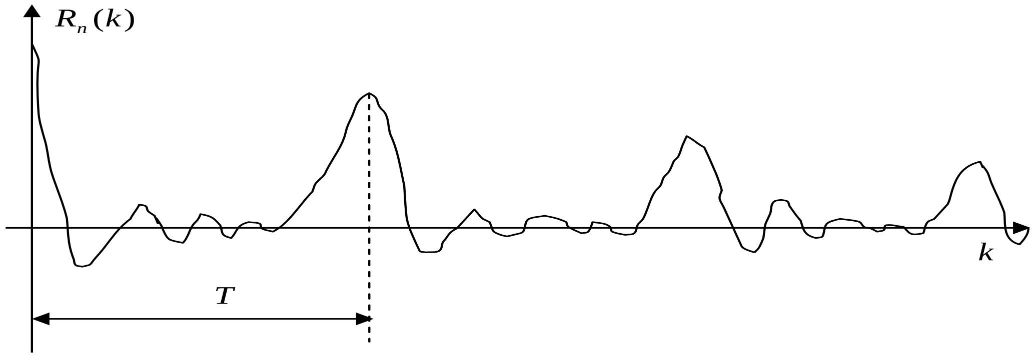
Pitch Detection：短时基音周期估计

$x(n)$



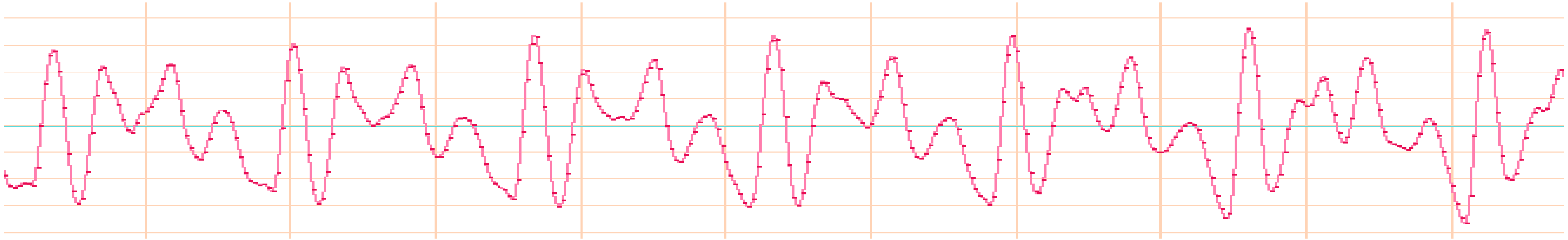
$x(n-T/3)$
$k=T/3$



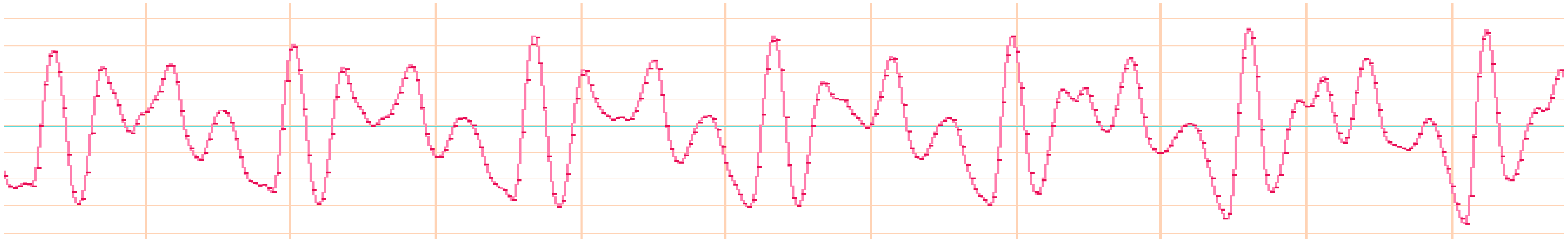


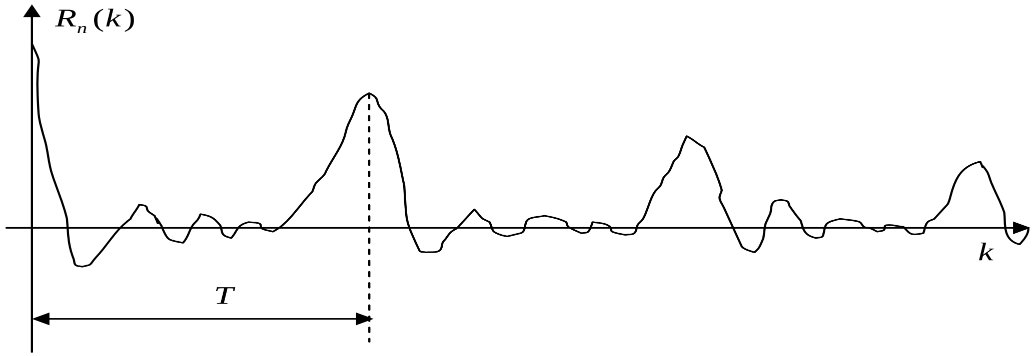
Pitch Detection：短时基音周期估计

$x(n)$



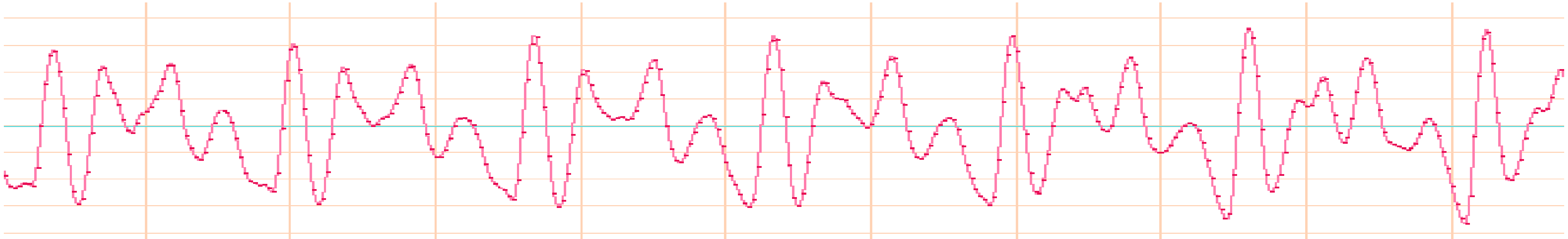
$x(n-2T/3)$
$k=2T/3$



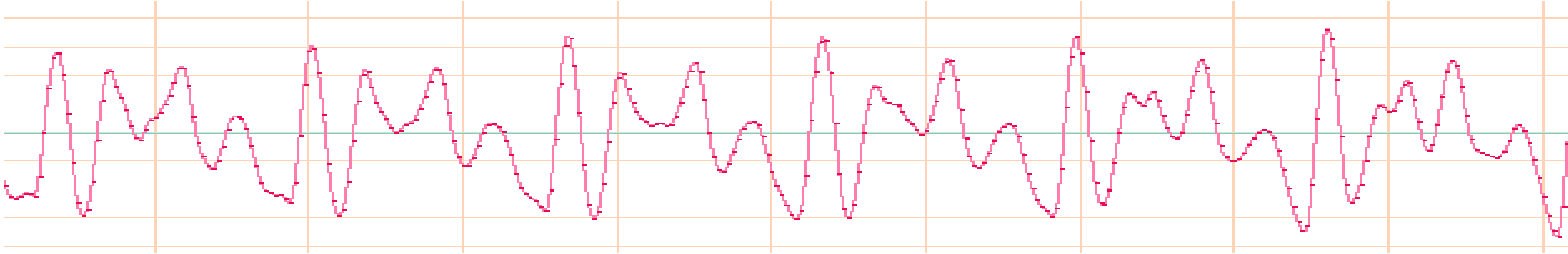


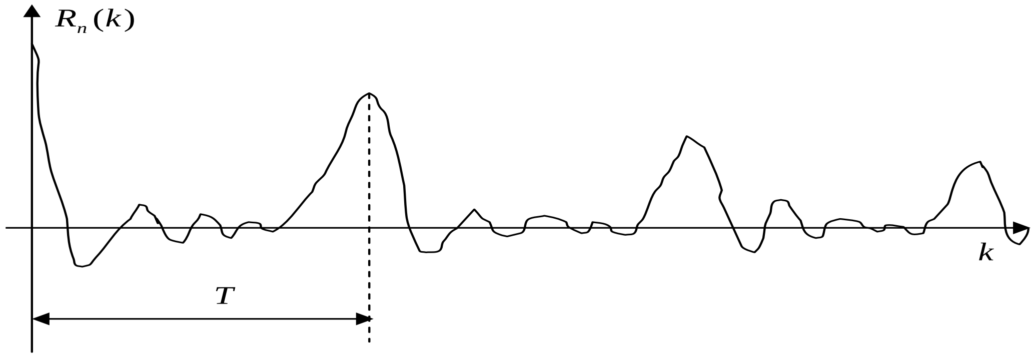
Pitch Detection：短时基音周期估计

$x(n)$



$x(n-T)$
$k=T$





语音信号的频域处理

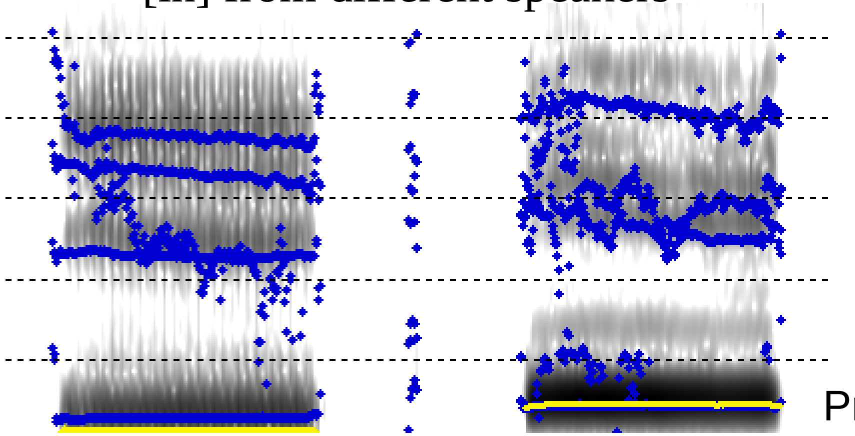
FREQUENCY-DOMAIN PROCESSING OF SPEECH SIGNAL

Frequency Domain Analysis: 频域分析

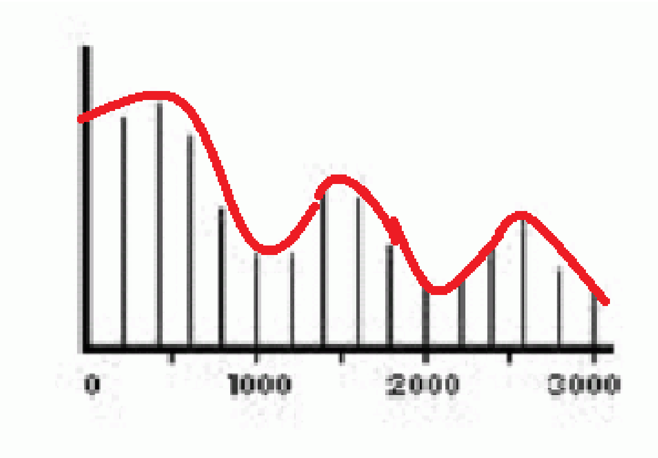
Importance of Frequency Domain Information

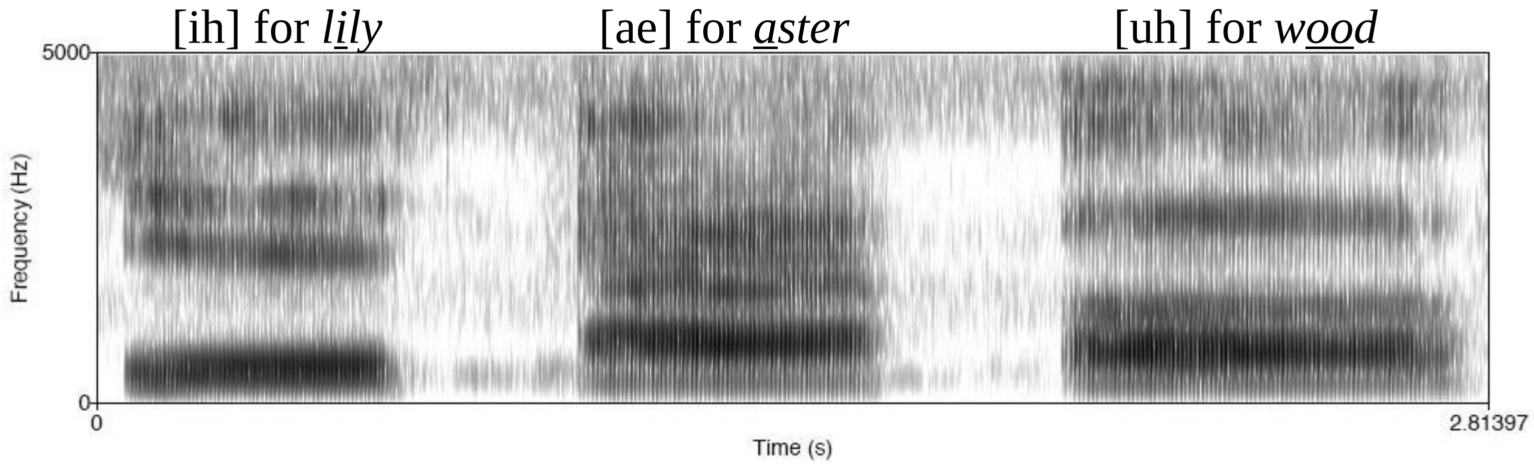
- The frequency domain information of the waveform are very important for some applications such as speech recognition, speaker identification, etc.
- Spectrum: 语谱
- Spectrogram: 语谱图
- Formant: 共振峰

[ih] from different speakers



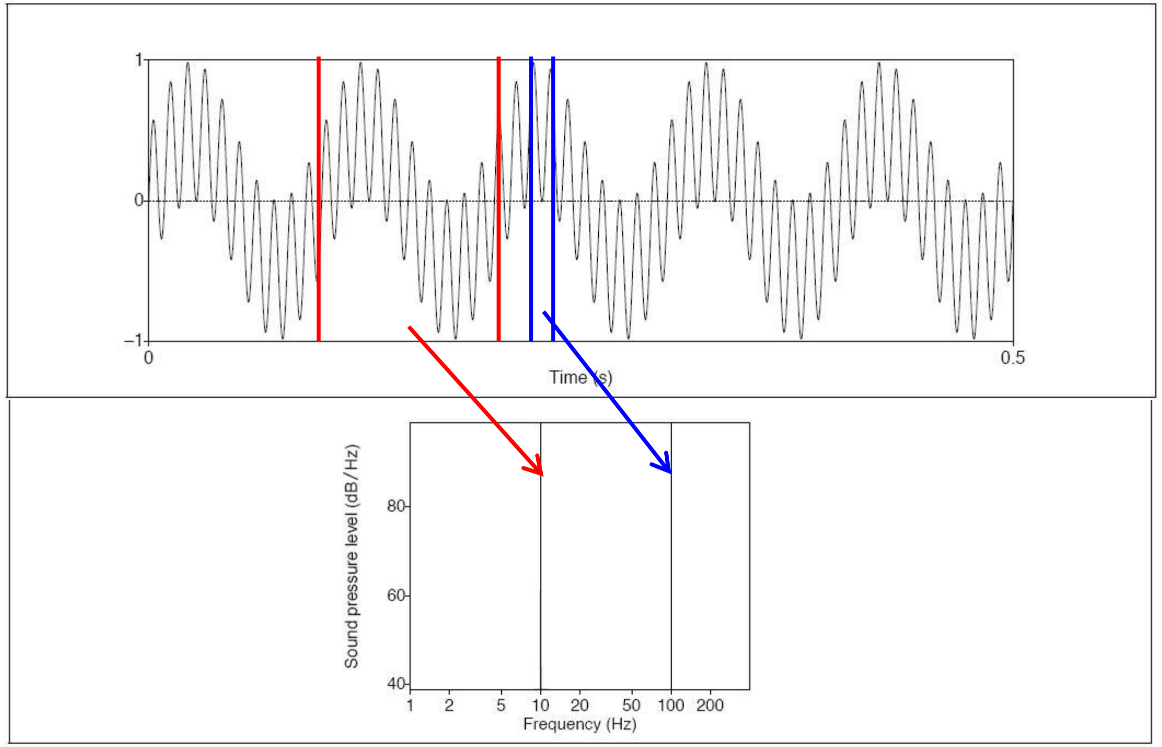
Praat





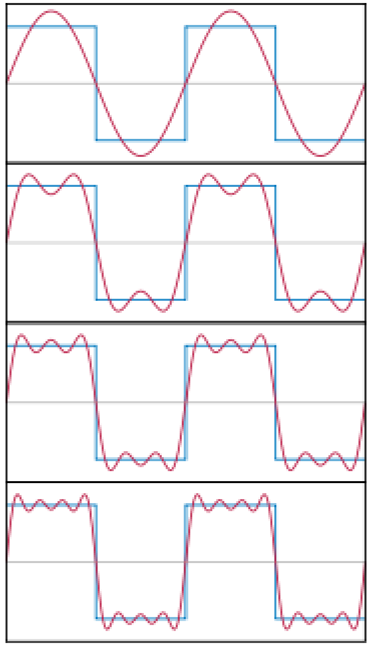
Fourier Transform: 傅立叶变换

- **Fourier Analysis: 傅立叶分析**
- **语谱分析工具**
  - 每个复杂的波形都是由不同频率的正弦波组合而成
  - 将原始信号由时域特征转换为频域特征进行分析，为信号的频域分析奠定了基础

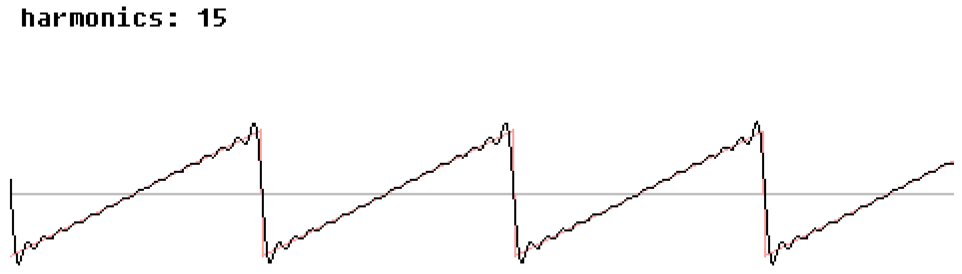


Fourier Transform

- A function or a signal can be decomposed into a sum of simple oscillating functions, namely sines and cosines.



$$x_{\text{square}}(t) = \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{\sin\left((2k-1)2\pi f t\right)}{(2k-1)}$$
$$= \frac{4}{\pi}\left(\sin(2\pi f t) + \frac{1}{3}\sin(6\pi f t) + \frac{1}{5}\sin(10\pi f t) + \cdots\right).$$



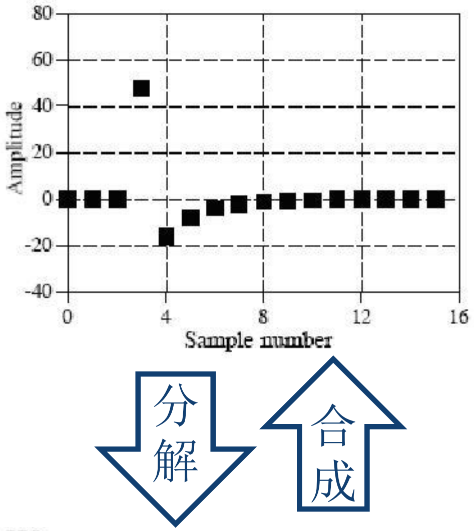
$$x_{\text{sawtooth}}(t) = -\frac{2}{\pi}\sum_{k=1}^{\infty}\frac{\sin(2\pi k f t)}{k}$$

(from: wikipedia)

Illustration of Fourier Series (傅立叶级数)

Fourier Transform: 傅立叶变换

- **Fourier Transform**
  - A function or a signal can be decomposed into a sum of simple oscillating functions, namely sines and cosines.



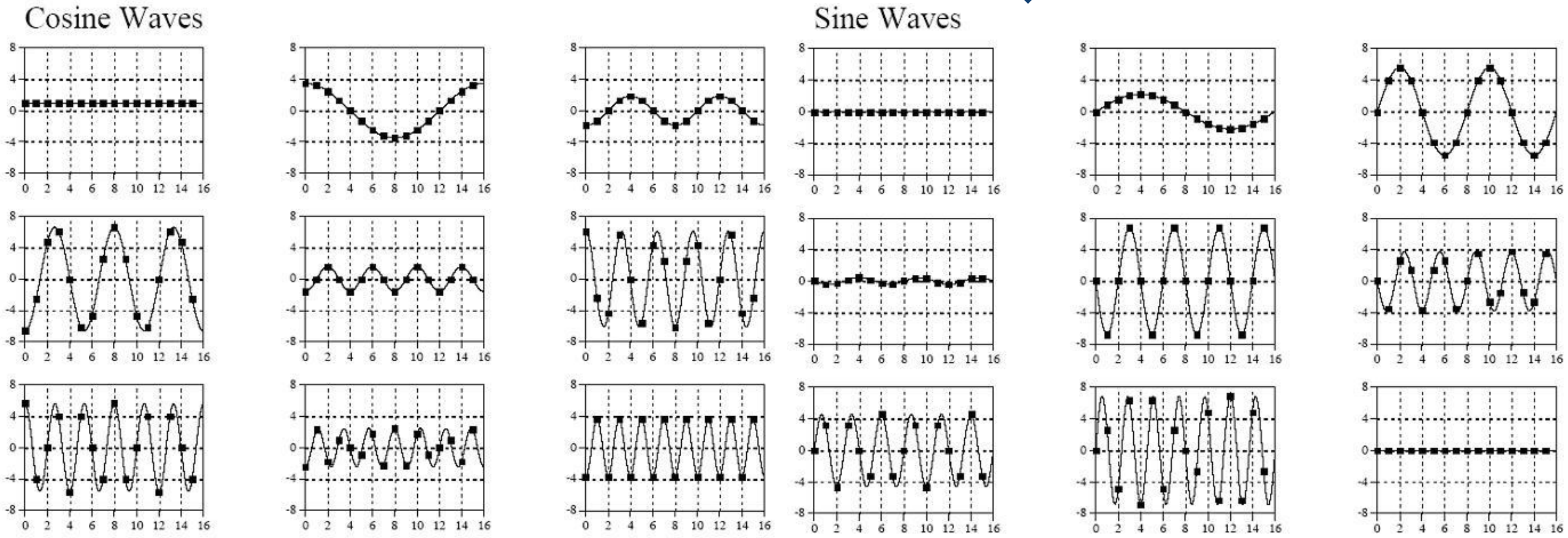


Illustration of Fourier Decomposition/Synthesis (傅立叶分解/合成)

Fourier Transform: 傅立叶变换

The Family of Fourier Transform

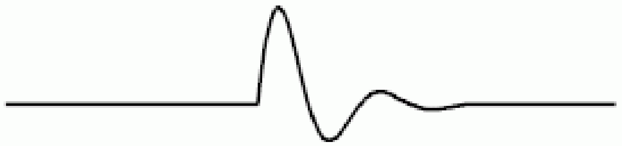
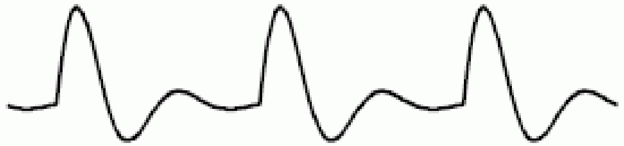
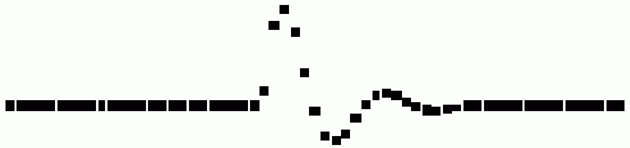
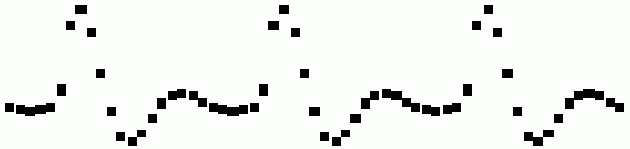
	Type of Transform	Example Signal
傅立叶变换 连续非周期信号	Fourier Transform *signals that are continious and aperiodic*	
傅立叶级数 连续周期信号	Fourier Series *signals that are continious and periodic*	
DTFT 离散时间傅立叶变换 离散非周期信号 时间离散、频域连续	Discrete Time Fourier Transform *signals that are discrete and aperiodic*	
DFT 离散傅立叶变换 离散周期信号 时域和频域均离散	Discrete Fourier Transform *signals that are discrete and periodic*	

FIGURE 8-2
Illustration of the four Fourier transforms. A signal may be continuous or discrete, and it may be periodic or aperiodic. Together these define four possible combinations, each having its own version of the Fourier transform. The names are not well organized; simply memorize them.

(from: http://www.dspguide.com/ch8/1.htm)

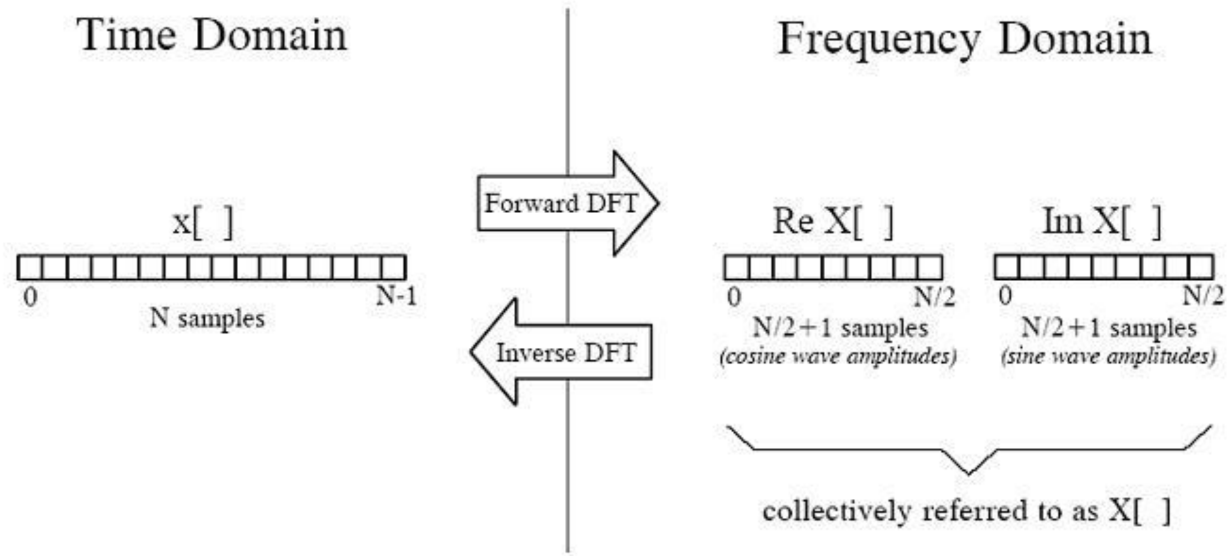
DFT: Discrete Fourier Transform

- **Discrete Fourier Transform: DFT**
  - 离散傅立叶变换

$$X(k) = \mathrm{DFT}[x(n)] = \sum_{n=0}^{N-1} x(n) \mathrm{e}^{-\mathrm{j}\frac{2\pi nk}{N}} \qquad (0 \le k \le N-1)$$

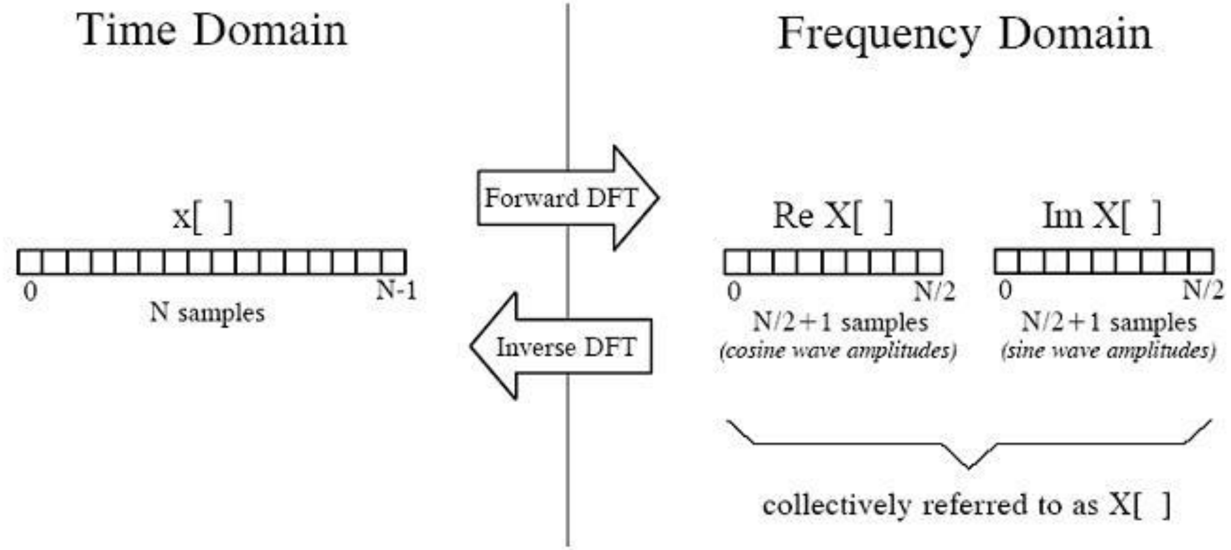
- **Inverse Discrete Fourier Transform: IDFT**
  - 离散傅立叶变换的反变换

$$x(n) = \mathrm{IDFT}[X(k)] = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \mathrm{e}^{\mathrm{j}\frac{2\pi nk}{N}} \qquad (0 \le n \le N-1)$$



DFT: Discrete Fourier Transform

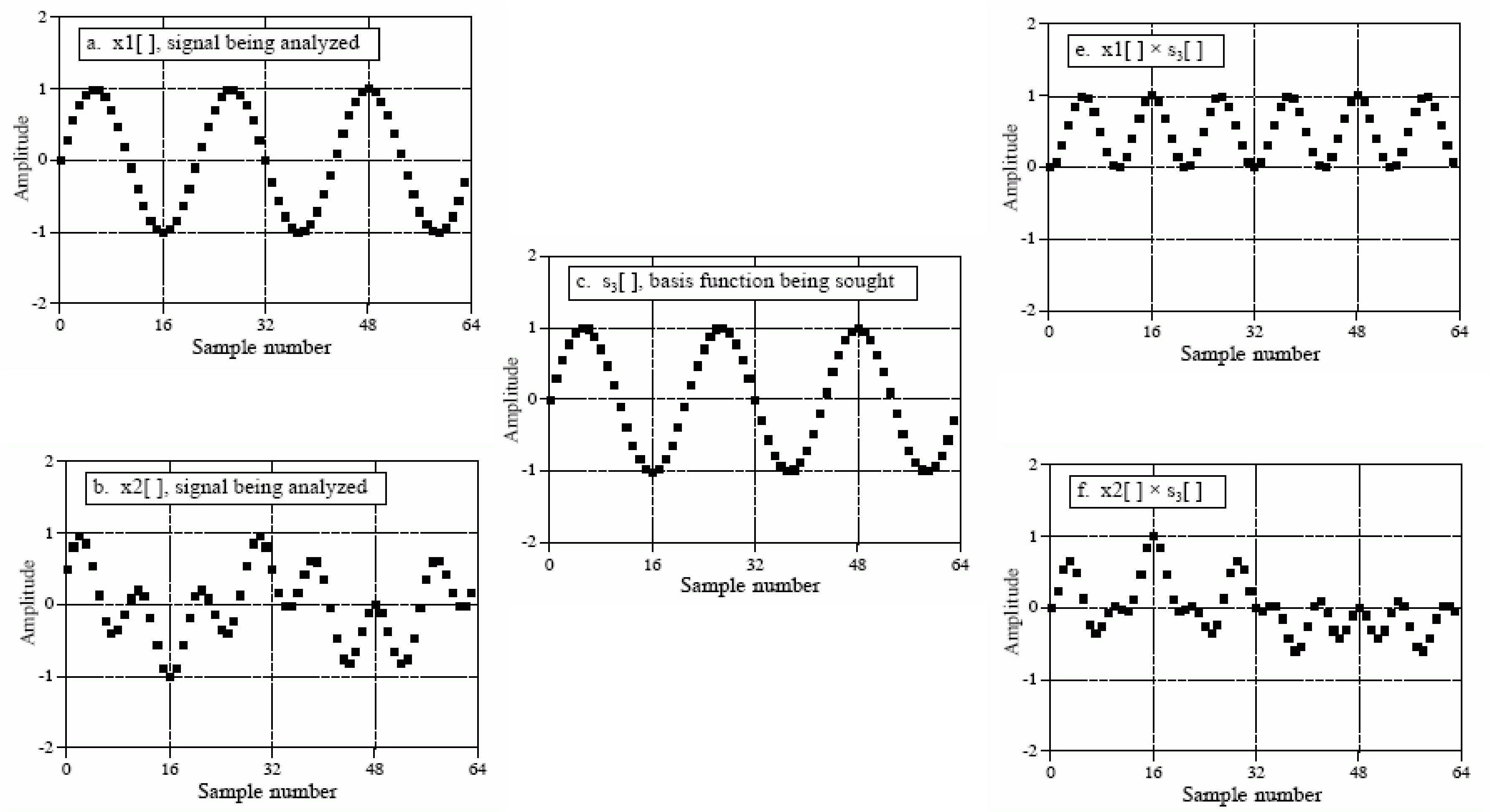
- **DFT and IDFT**
  - DFT transforms any signal from *time domain* to *frequency domain*.
  - The frequency domain contains exactly the same information as the time domain. If you know one domain, you can calculate the other.
  - DFT: *decomposition*, *analysis*, or *forward* DFT
  - IDFT: *synthesis*, or *inverse* DFT
  - N: the number of samples in the time domain
    - Can be any positive integer
    - A power of two is usually chosen: 128, 256, 512, 1024, etc.



DFT: Discrete Fourier Transform

- **Understanding the DFT**

$$X(k) = \mathrm{DFT}[x(n)] = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi nk}{N}} \qquad (0 \le k \le N-1)$$



STFT: Short-time Fourier Transform

- **Short-time Fourier Transform: STFT**
  - DFT: 普通离散傅立叶变换：适用于平稳、周期信号
    - 语音信号是典型的非平稳信号
    - 语音信号是短时平稳的：10ms~30ms的窗函数进行加窗处理
  - STFT: 短时傅立叶变换
    - 对语音信号首先进行短时加窗处理，然后对所得的加窗后的短时语音信号进行傅立叶变换

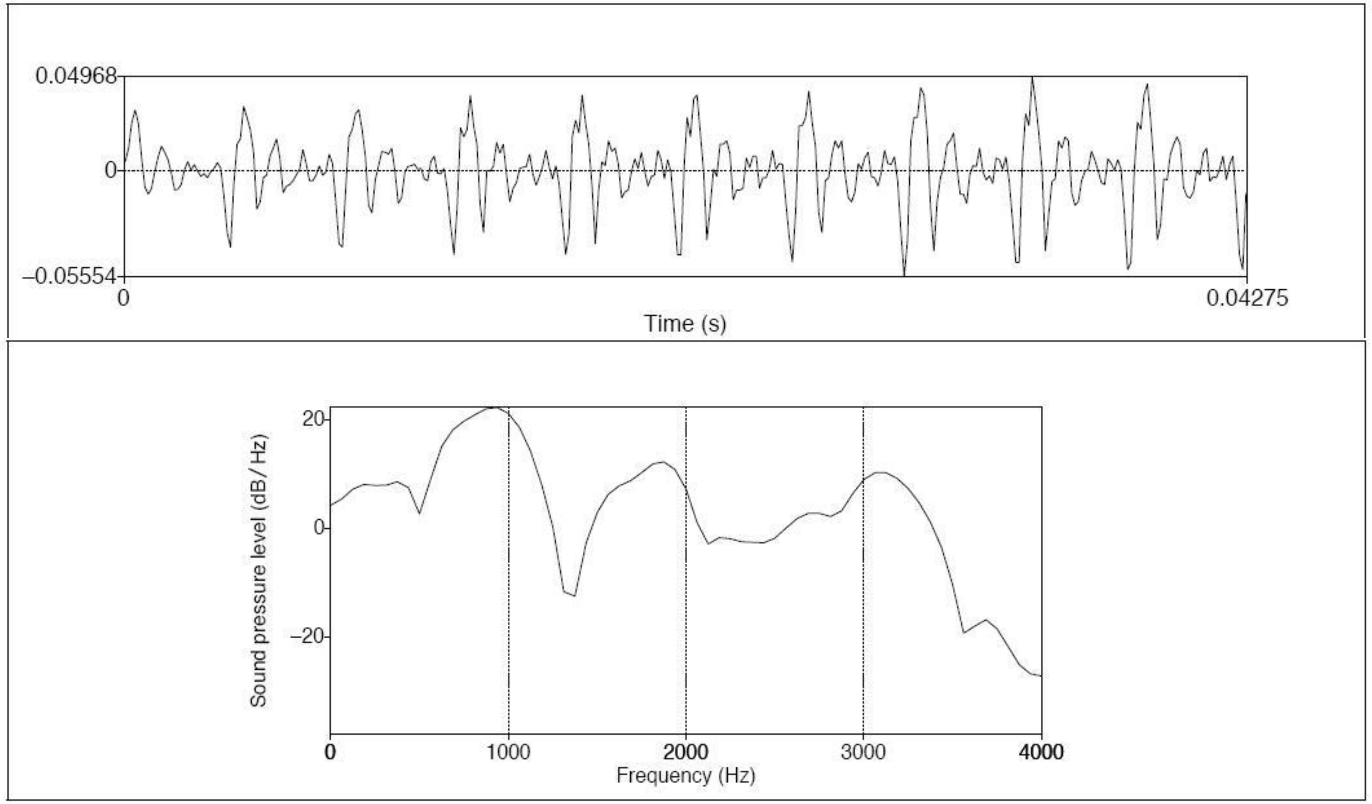
$$\mathrm{STFT}\{x(n)\} \equiv X(n,\omega) = X_n(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)\mathrm{e}^{-\mathrm{j}\omega m}$$

- **Spectrogram: 语谱图**
  - The **magnitude squared** of the STFT yields the spectrogram of the function.

$$\mathrm{spectrogram}\{x(n)\} \equiv S_n(\mathrm{e}^{j\omega}) = |X_n(\mathrm{e}^{\mathrm{j}\omega})|^2$$

Spectrum and Spectrogram: 语谱和语谱图

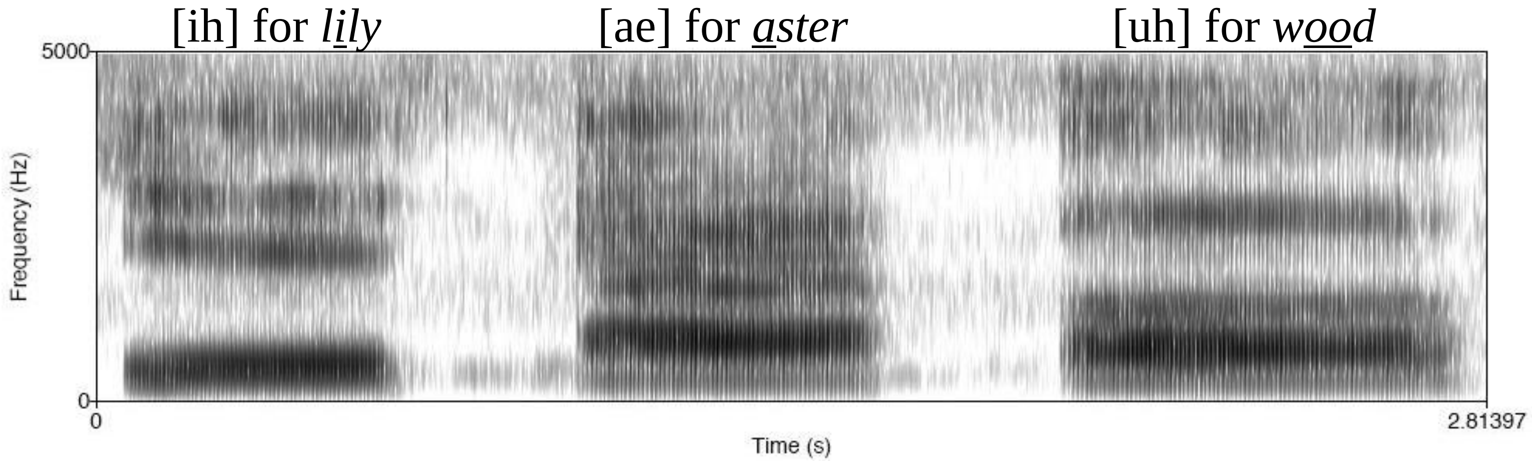
- **Spectrum: 语谱**
  - The spectrum of a signal is a representation of each of its frequency components and their amplitudes.
  - 语音信号的频域波形，描述信号包含的频率成分和它们的幅度
  - 描述的是某个特定时刻的加窗后短时语音信号的频率分布，又称为spectral slice



The waveform for the vowel /ae/ from *h<u>a</u>d*, and its spectrum computed from DFT

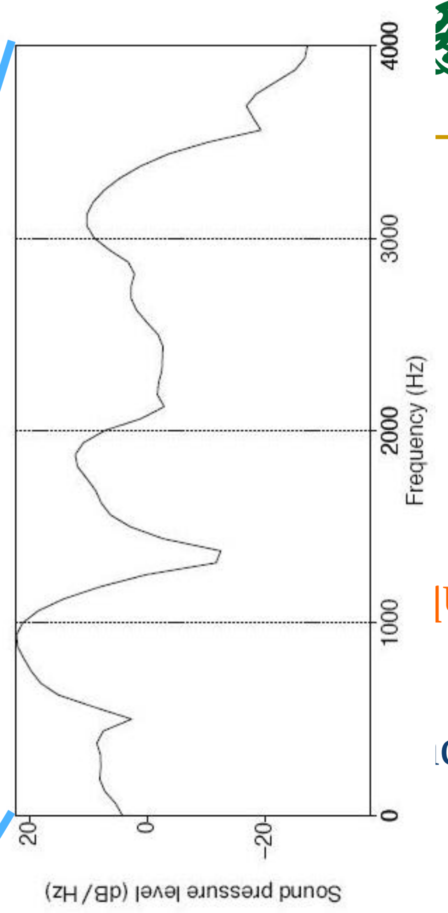
Spectrum and Spectrogram: 语谱和语谱图

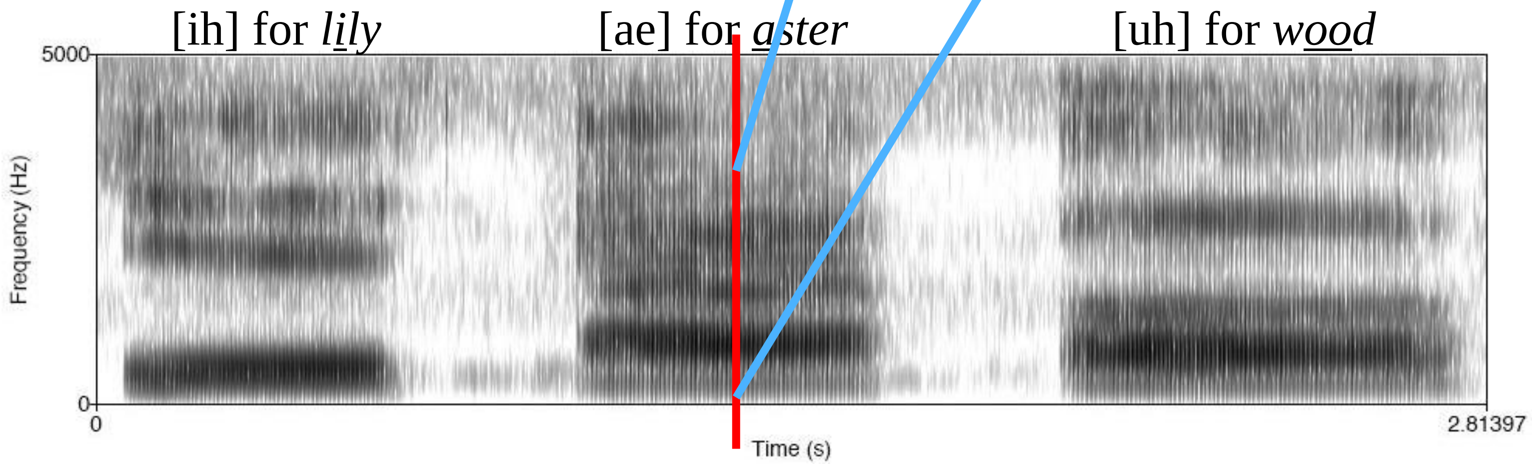
- **Spectrogram: 语谱图**
  - A spectrogram is a way of envisioning how the different frequencies that make up a waveform change over time.
  - The *x*-axis shows time, as it did for the waveform.
  - The *y*-axis now shows frequencies in Hertz.
  - The darkness of a point on a spectrogram corresponding to the amplitude of the frequency component.
    - Very dark points have high amplitude, light points have low amplitude.
  - The spectrogram is a useful way of visualizing the three dimensions (time x frequency x amplitude).



Spectrum and Spectrogram: 语 … 图

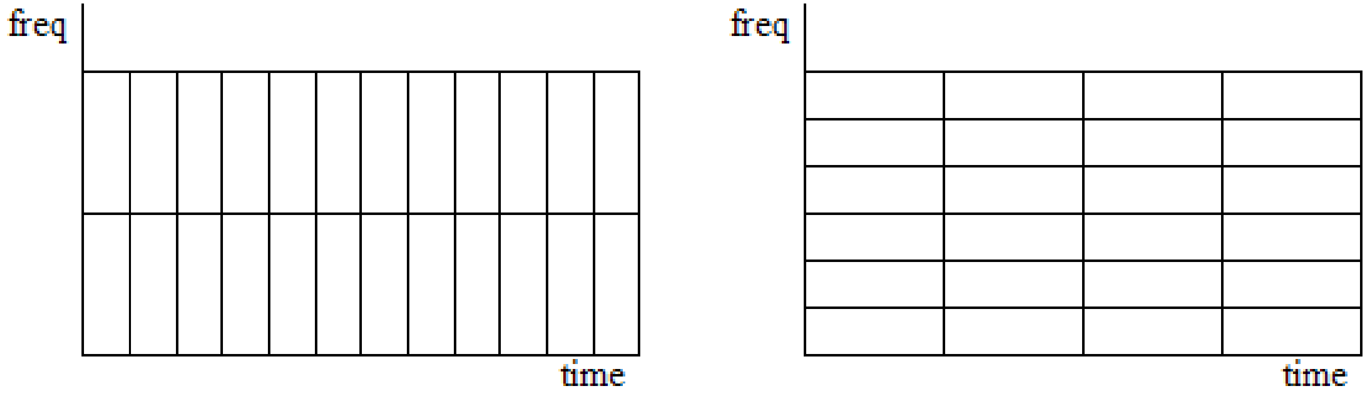
- **Spectrogram: 语谱图**
  - A spectrogram is a way of envisioning how the different frequ … waveform change over time.
  - The *x*-axis shows time, as it did for the waveform.
  - The *y*-axis now shows frequencies in Hertz.
  - The darkness of a point on a spectrogram corresponding to the … uency component.
    - Very dark points have high amplitude, light points have lo
  - The spectrogram is a useful way of visualizing the three dime … cy x amplitude).





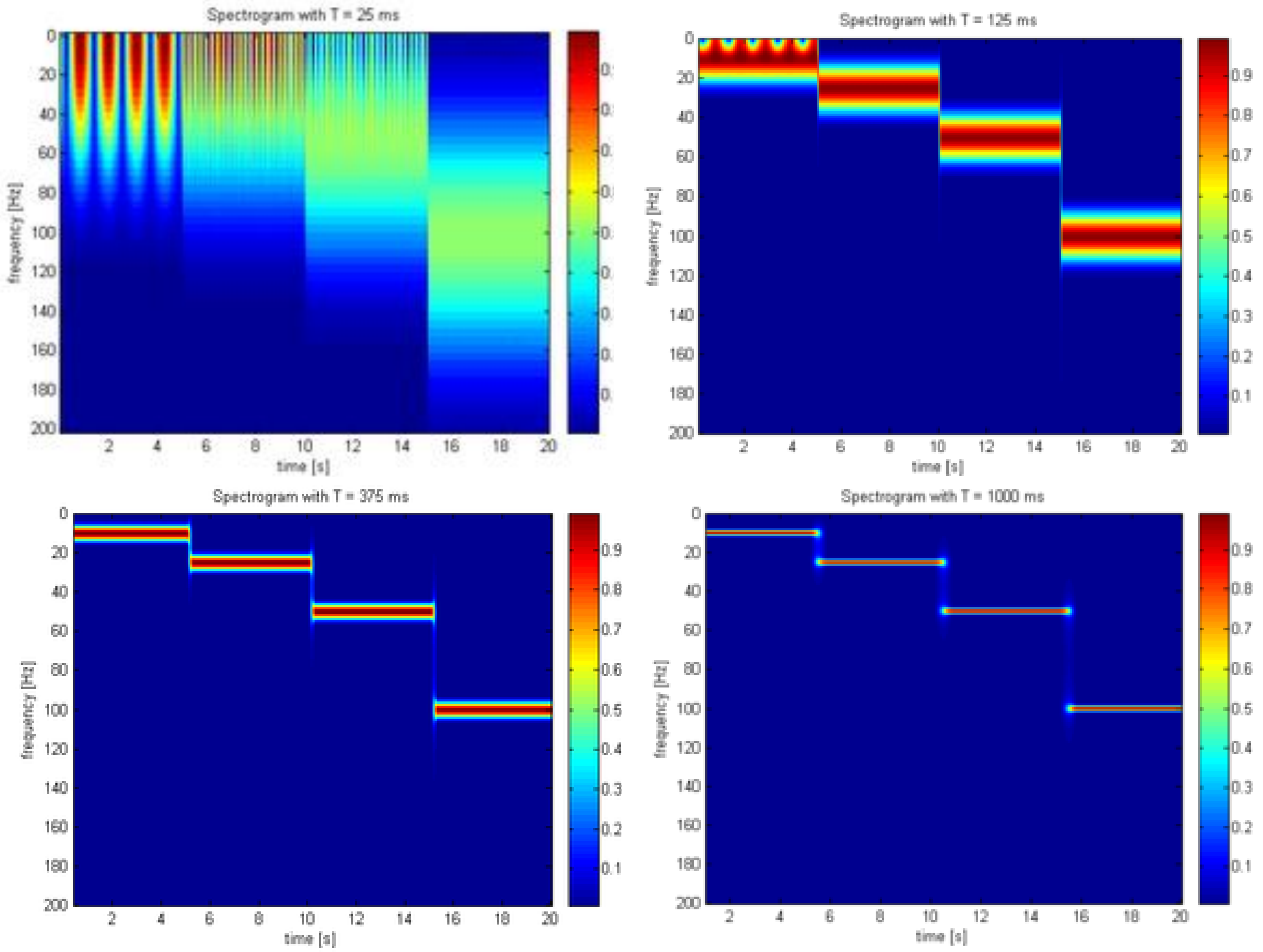
Spectrogram: 语谱图

- **Resolution Issues**
  - Frequency Resolution: 频率分辨率
    - 语谱图纵坐标(频率)的分辨率
    - 频率分辨率高，靠在一起的频率分量能较容易分开
  - Time Resolution: 时间分辨率
    - 语谱图横坐标(时间)的分辨率
    - 时间分辨率高，更能反映频率随时间变化的情况



Spectrogram: 语谱图

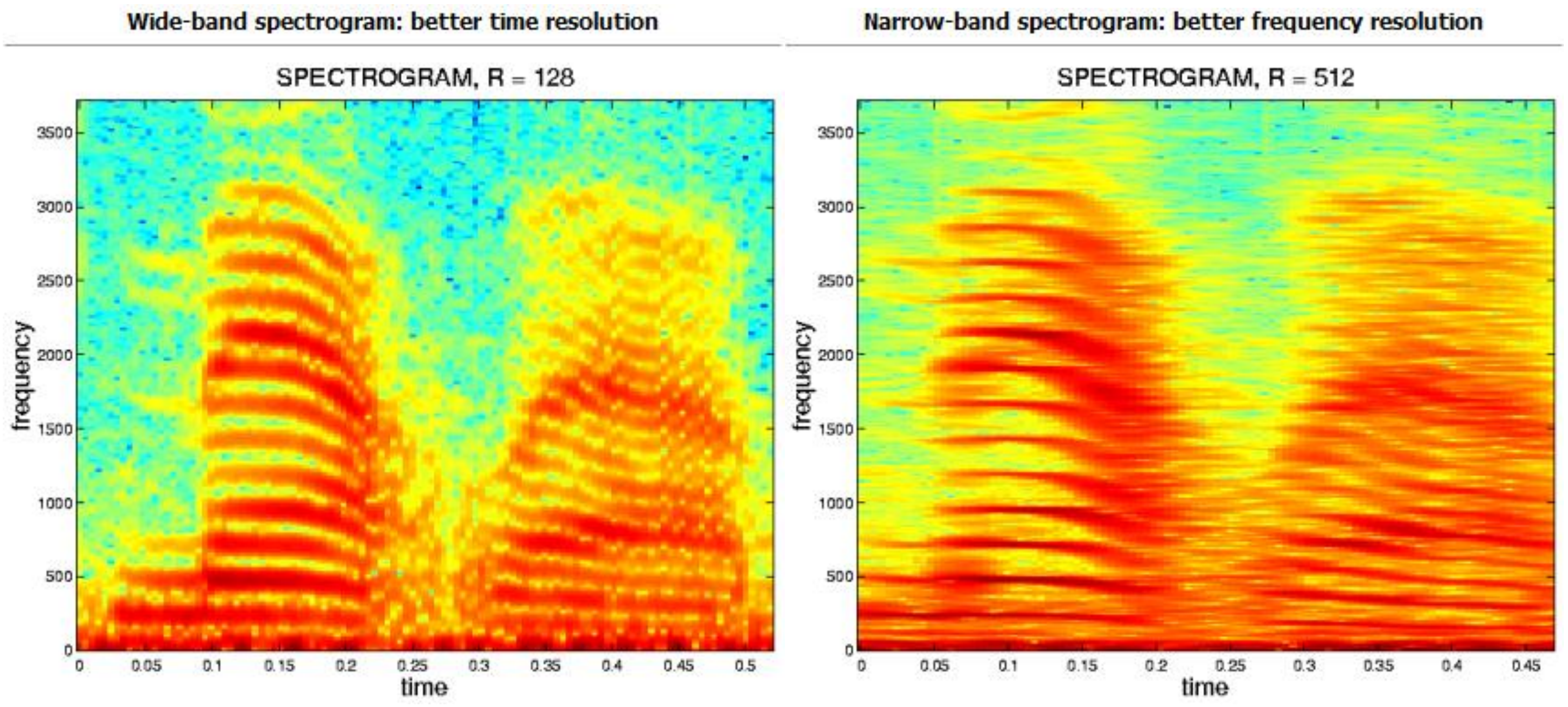
- **Resolution Issues**
  - 两种分辨率均受窗函数的影响
    - Wide window gives better frequency but poor time resolution.
    - Narrow window gives good time but poor frequency resolution.
  - Explanation
    - Frequency Resolution:
      Frequency space between
      2 consecutive coefficients: fs/N



$$x(t) = \begin{cases} \cos(2\pi 10 t); & 0 \le t < 5s \\ \cos(2\pi 25 t); & 5 \le t < 10s \\ \cos(2\pi 50 t); & 10 \le t < 15s \\ \cos(2\pi 100 t); & 15 \le t < 20s \end{cases}$$

Spectrogram：语谱图

- **Wide-band Spectrogram: 宽带语谱图**
  - 频率分辨率取300-400Hz，时间分辨率2-5ms，良好的时间分辨率，频率分辨率较差
- **Narrow-band Spectrogram: 窄带语谱图**
  - 频率分辨率取50-100Hz，时间分辨率5-10ms，良好的频率分辨率，时间分辨率较差



(from: http://cnx.org/content/m10570/latest/)

praat

Q&A

MEDIA JOINT Research Center