语音信号的短时分析与处理

李军锋

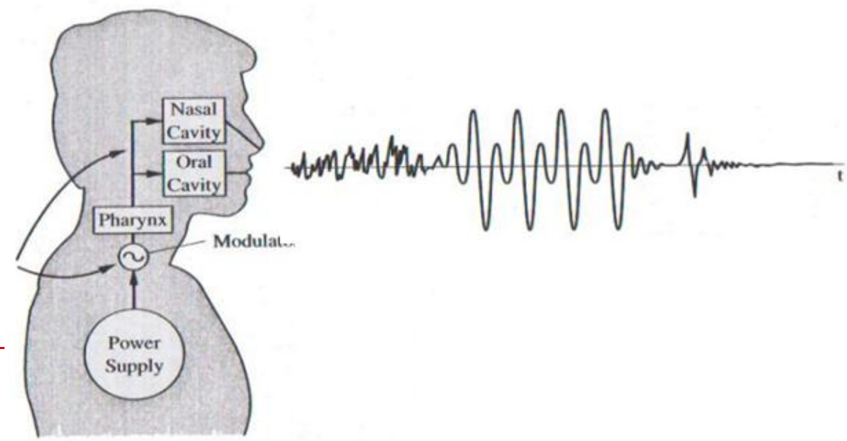
中国科学院声学研究所

提纲

- 短时分析的必要性
- 短时分析
  - 时域短时分析
  - 频域短时分析
- 常用短时分析技术
  - 短时能量
  - 短时平均幅度
  - 短时平均过零率
  - 短时自相关函数
  - 短时平均幅度差
  - 短时频谱
  - 短时功率谱

短时分析的必要性

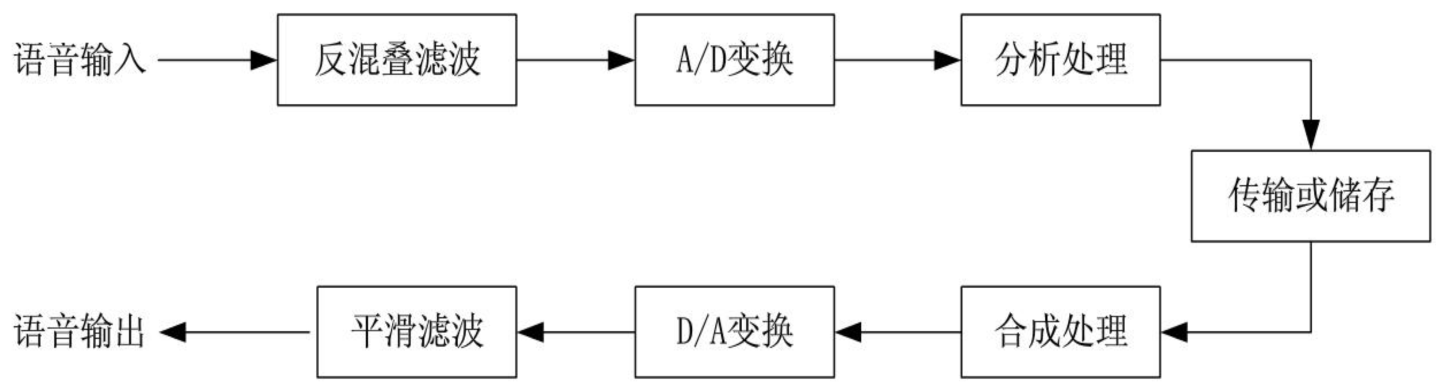
- 分析是处理的前提和基础
- 分析目的：提取需要的信息、获取特征表示参数
- 分类：时域分析/频域分析、模型分析/非模型分析等
- 分析技术：短时分析（10-30ms相对平稳）
- 分析帧长：20~30ms



语音信号的数字化与预处理

- 语音信号处理的根本方法 --- 短时分析技术

语音信号具有时变特性，但在一个短时间范围内其特性基本保持不变，因而可以将其看作是一个准稳态过程。语音的重要特性是它具有“短时性”，所以对语音的分析和处理必须建立在“短时”的基础上，即“短时分析”。



预滤波

- 预滤波的目的
  - 抑制输入信号各频率分量中频率超出fs/2的所有成分，防止混叠干扰
  - 抑制50Hz的工频干扰

实现:

预滤波实际上是一个带通滤波，其上下截止频率分别为f_H和f_L（如: f_H=3400Hz, f_L=60~100Hz)

采样

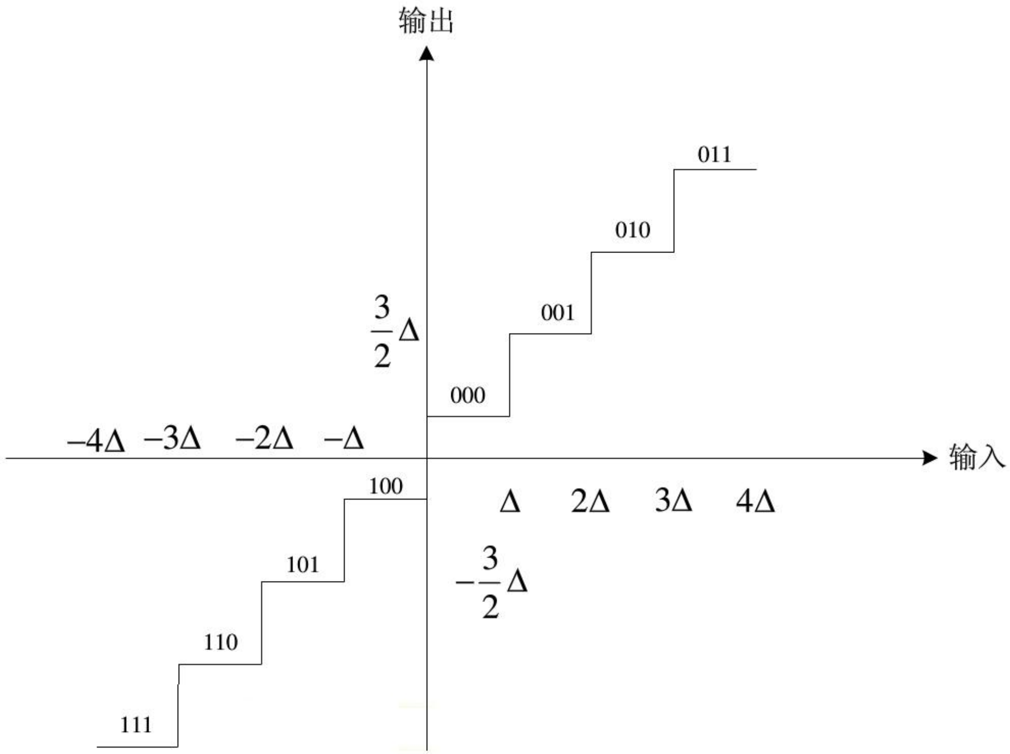
- 根据采样定理，当采样频率大于信号的两倍带宽时，采样过程中不会丢失信号，且从采样信号中可以精确地重构原始信号波形。在信号的带宽不明确时，在采样前应接入反混叠滤波器，使其带宽限制在某个范围内。

电话语音信号频率范围：300~3400Hz

采样率：8kHz

量化

- 将输入的整个幅值分成有限区间，把落入同一区间的波形样本都量化成同一幅度值。

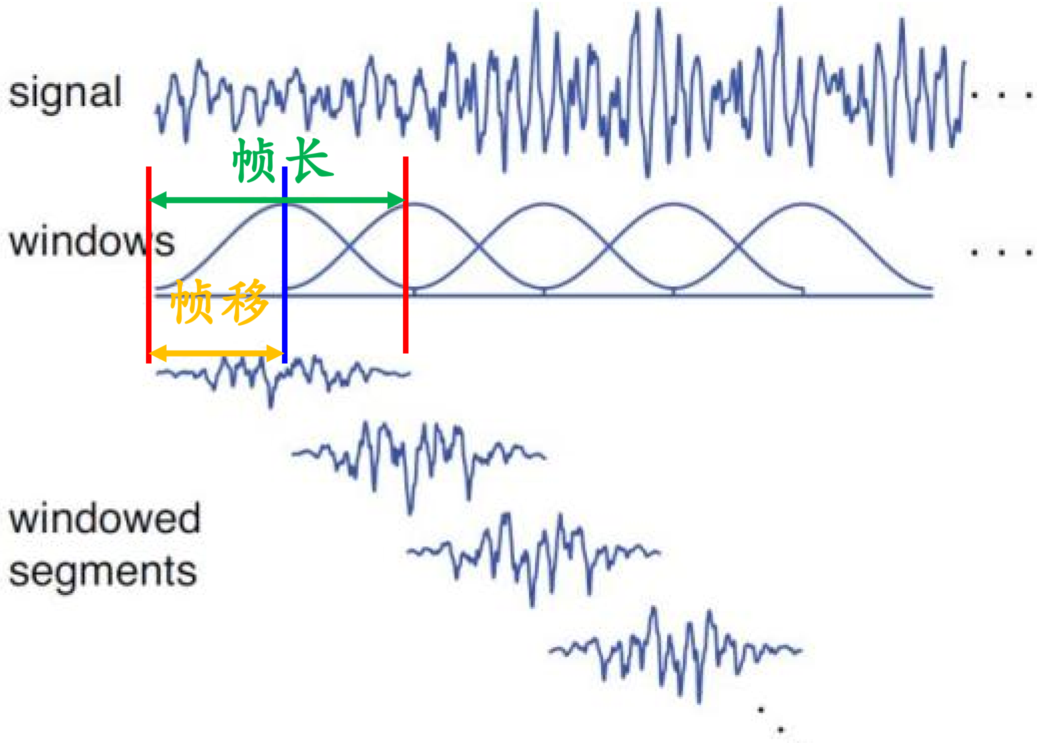


预处理：预加重

- 预加重实现：
  一阶高通滤波器 $H(Z)=1-\mu Z^{-1}$
  $$\mu=0.93\sim0.98$$
- 作用：对原输入信号 $x(n)$，得到新信号
  $$y(n)=x(n)-\mu x(n-1)$$
- 原因：语音信号平均功率谱受声门激励和口鼻辐射的影响，高频（大约在800Hz以上）按6dB/oct衰减。预加重可以提升高频部分，使得信号频谱变得平坦，以便进行频谱分析或声道参数分析。
- 位置：可在反混叠滤波之前进行，这样不仅能够进行预加重，而且可以压缩信号的动态范围，有效提高信噪比；也可以在A/D变换之后进行，用具有6dB/oct的提升高频特性的预加重滤波器实现。

预处理：加窗分帧

- 10~30ms内，语音信号可看作平稳信号
- 帧移图：



典型窗函数

矩形窗

$$w(n) = \begin{cases} 1 & 0 \le n \le M-1 \\ 0 & 其它 \end{cases}$$

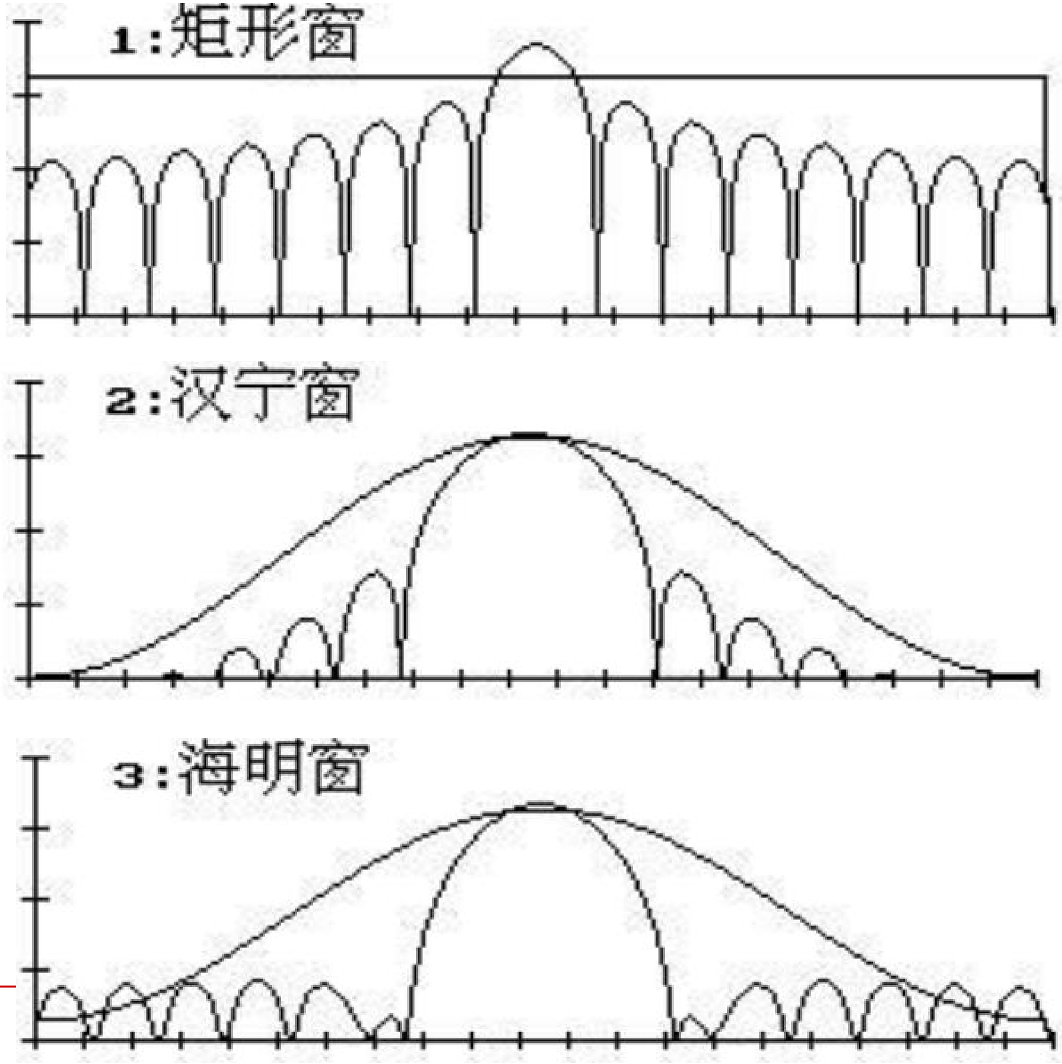
汉宁窗

$$w(n) = \begin{cases} 0.5 - 0.5\cos(2\pi n/(M-1)) \\ 0 \quad 其它 \end{cases}$$

海明窗

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(M-1)) \\ 0 \quad 其它 \end{cases}$$

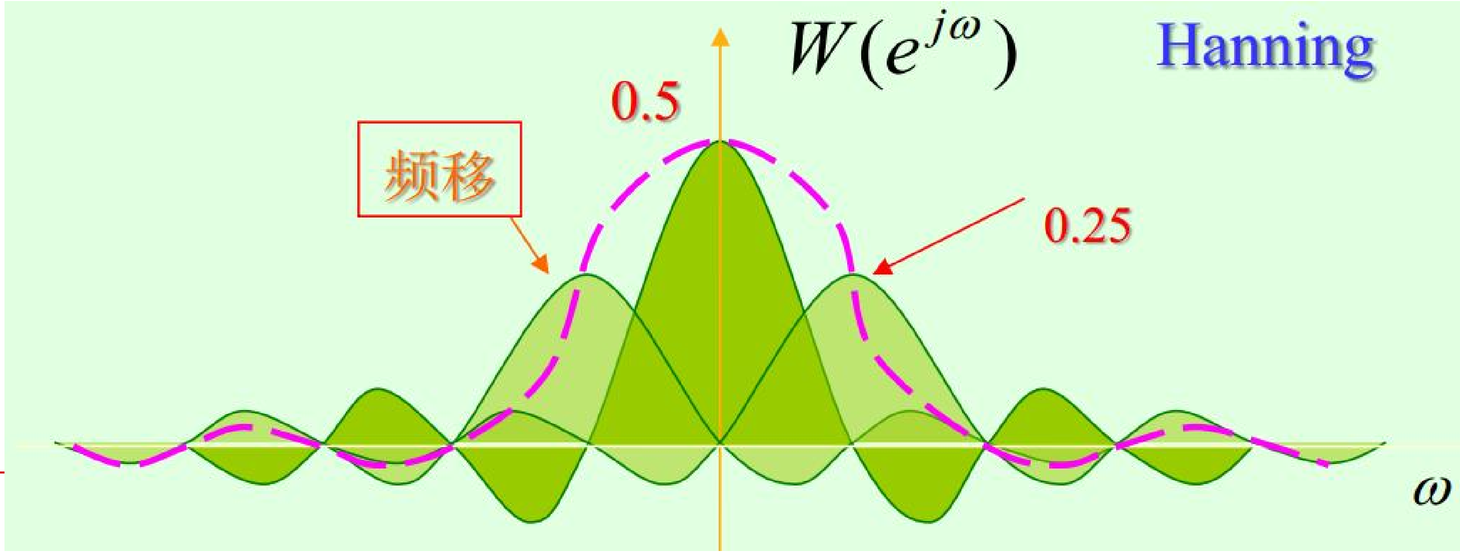
典型窗函数



窗函数的Fourier变换

$$W[n]=0.5+0.5\frac{e^{-j2\pi n/\mathrm{N}}+e^{j2\pi n/\mathrm{N}}}{2}=\frac{1}{2}\left[1+\cos(\frac{2n\pi}{\mathrm{N}})\right]$$

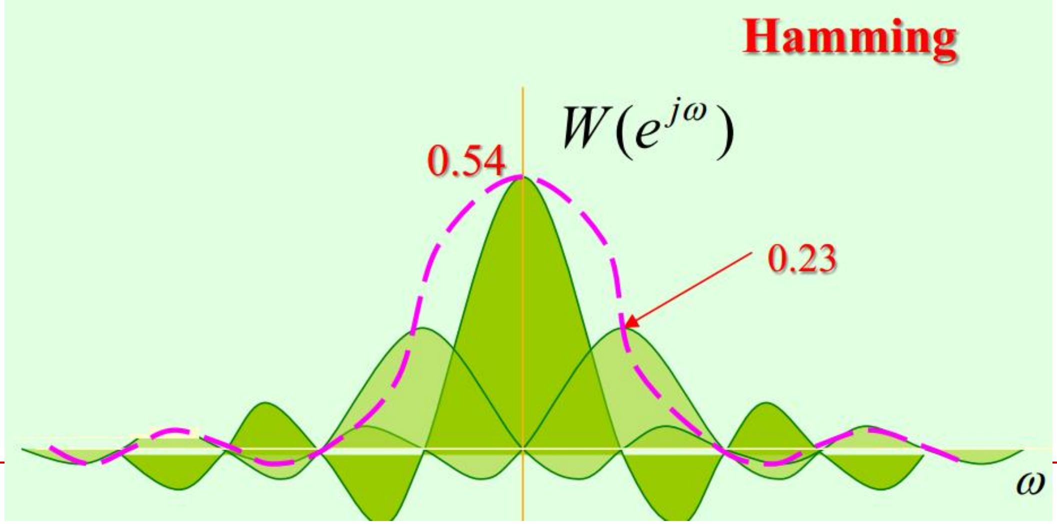
$$\left|W\left(e^{j\omega}\right)\right|=\frac{1}{2}\left|W_R\left(e^{j\omega}\right)\right|+\frac{1}{4}\left[\left|W_R\left(e^{j(\omega-\frac{2\pi}{N})}\right)\right|+\left|W_R\left(e^{j(\omega+\frac{2\pi}{N})}\right)\right|\right]$$



窗函数的Fourier变换

$$W[n]=0.54+0.46\frac{e^{-j2\pi n/\mathrm{N}}+e^{j2\pi n/\mathrm{N}}}{2}=0.54-0.46\cos\left(\frac{2n\pi}{\mathrm{N}}\right)$$

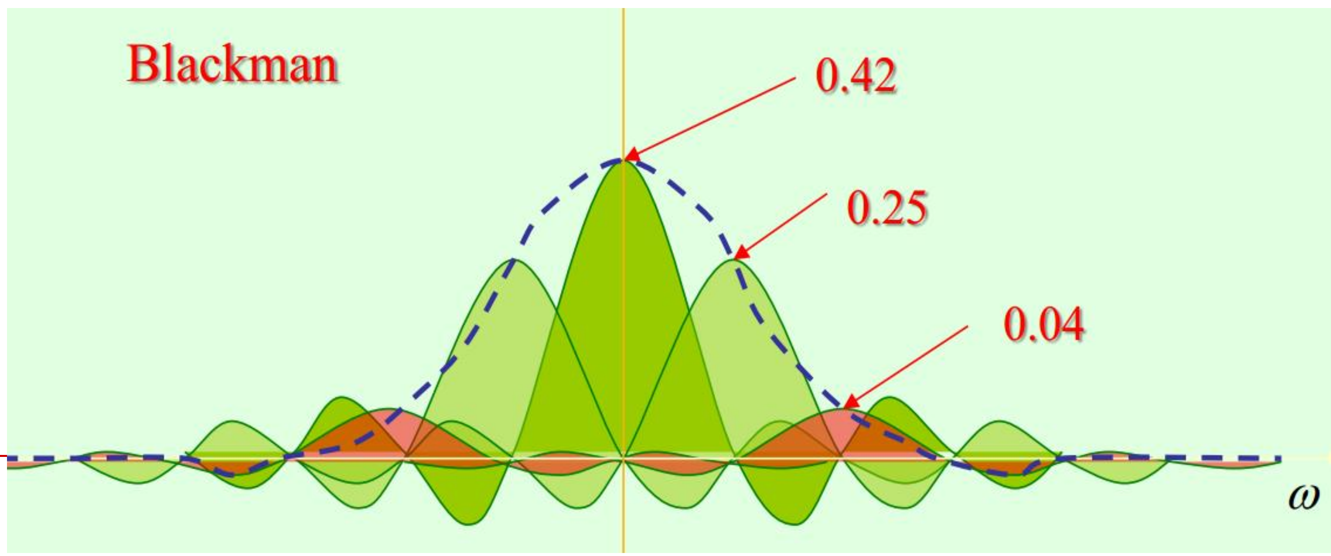
$$\left|\mathrm{W}\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|=0.54\left|\mathrm{W}_{\mathrm{R}}\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|+0.23\left[\left|\mathrm{W}_{\mathrm{R}}\left(\mathrm{e}^{\mathrm{j}(\omega-\frac{2\pi}{N})}\right)\right|+\left|\mathrm{W}_{\mathrm{R}}\left(\mathrm{e}^{\mathrm{j}(\omega+\frac{2\pi}{N})}\right)\right|\right]$$



窗函数的Fourier变换

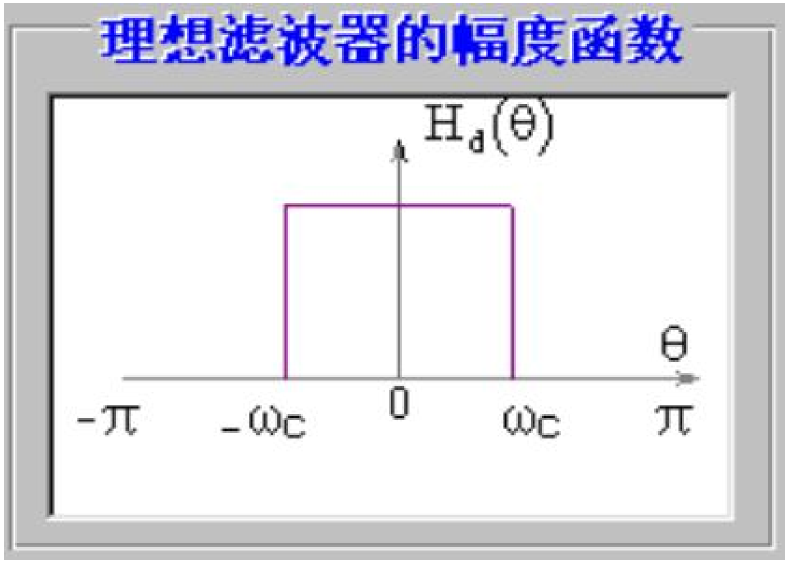
$$W[n]=0.42+0.5\frac{e^{-j2\pi n/N-1}+e^{j2\pi n/N-1}}{2}+0.08\frac{e^{-j4\pi n/N-1}+e^{j4\pi n/N-1}}{2}$$

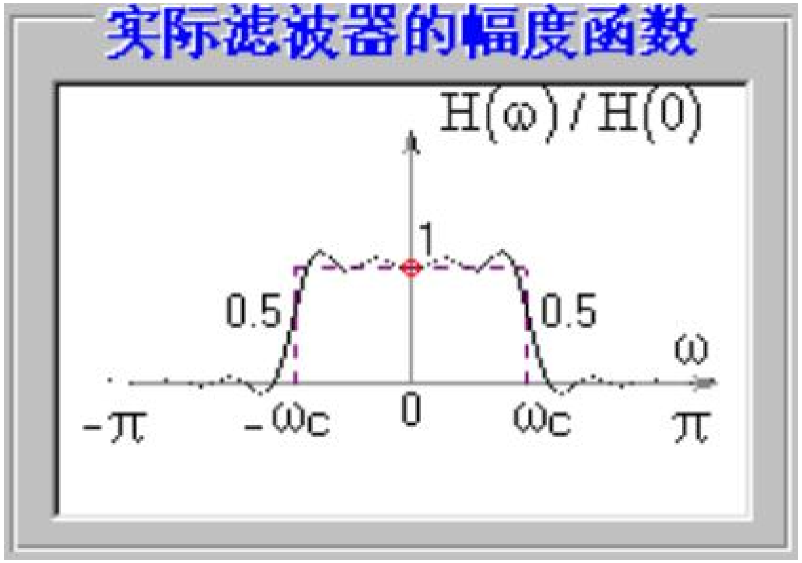
$$\left|W(e^{j\omega})\right|=0.42\left|W_R(e^{j\omega})\right|+0.25\left[\left|W_R(e^{j(\omega-\frac{2\pi}{N-1})})\right|+\left|W_R(e^{j(\omega+\frac{2\pi}{N-1})})\right|\right]$$
$$+0.04\left[\left|W_R(e^{j(\omega-\frac{4\pi}{N-1})})\right|+\left|W_R(e^{j(\omega+\frac{4\pi}{N-1})})\right|\right]$$



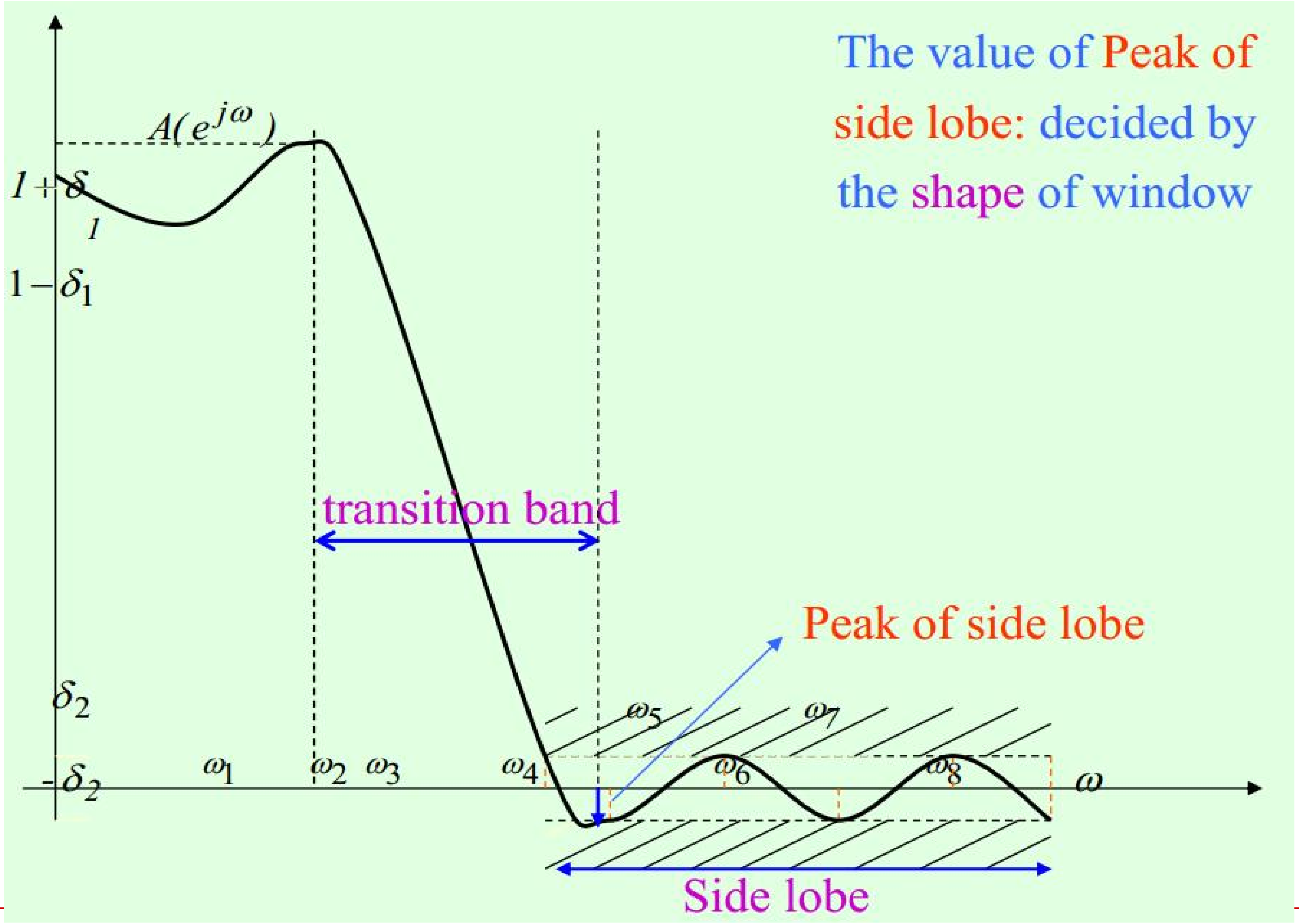
窗函数的影响

$$h[n]=w[n]h_d[n]\Leftrightarrow H(e^{j\omega})=\frac{1}{2\pi}\int_{-\pi}^{\pi}H_d(e^{j\theta})W(e^{j(\omega-\theta)})d\theta$$





窗函数的影响



The value of Peak of side lobe: decided by the shape of window
$A(e^{j\omega})$
$1+\delta_1$
$1-\delta_1$
δ_2
$-\delta_2$
ω_1 ω_2 ω_3 ω_4 ω_5 ω_6 ω_7 ω_8 ω
transition band
Peak of side lobe
Side lobe

窗函数的选择

- 时域：decrease the slope at the ends to get smooth transition, and reduce the effect of being cut off
- 频域：narrow transition band, and low peak of side lobe

各种窗函数特征比较

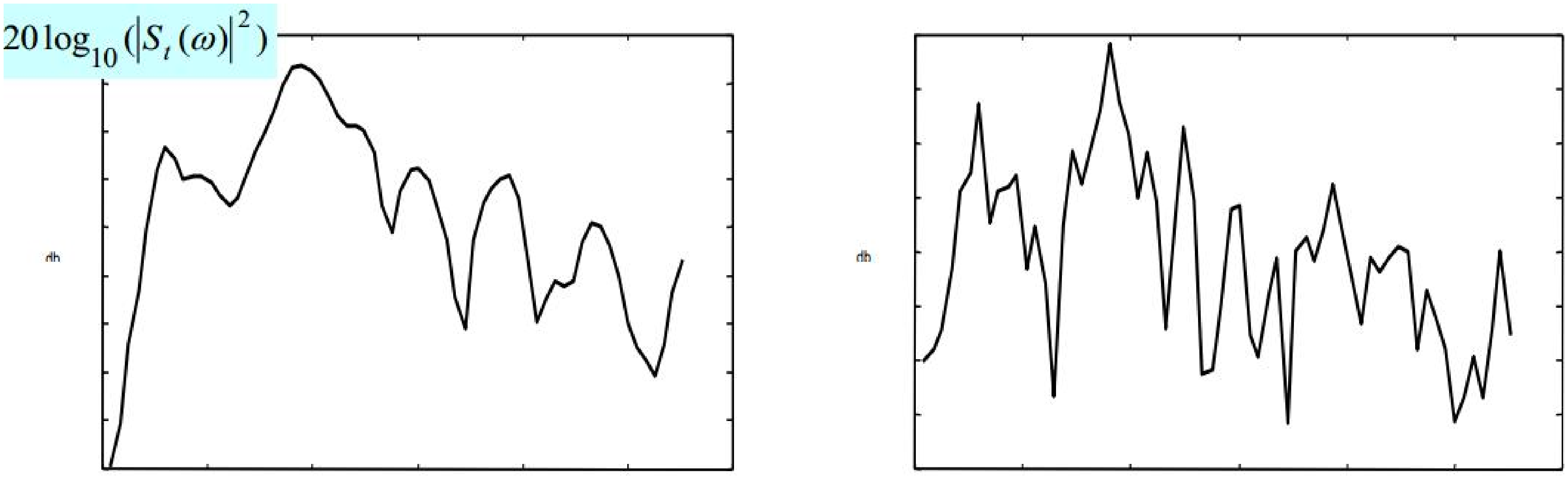
窗的类型	旁瓣峰值	主瓣宽度
Rect.	-13	$4\pi/N$
Bartlett	-25	$8\pi/N$
Hanning	-31	$8\pi/N$
Hamming	-41	$8\pi/N$
Blackman	-57	$12\pi/N$

矩形窗的过渡带最窄，但是波动最明显；其他几种窗沿两边平滑过渡，频响旁辫显著降低，但代价是过渡带变宽

窗长的选择

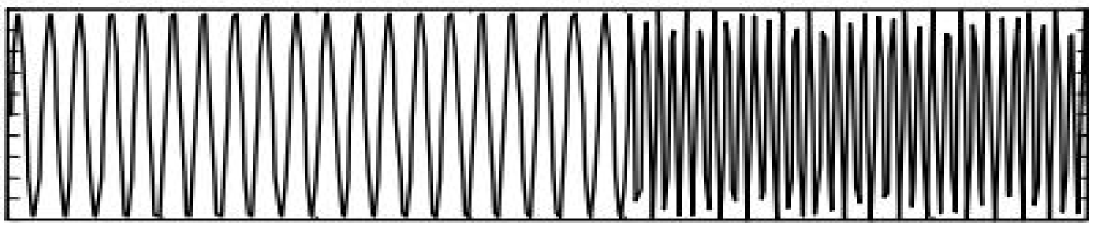
窗长的选择： 窗长的选择对反映语音信号的幅度变化起着重要的作用。如果窗长太长，它就可以看作一个很窄的低通滤波器，此时随时间的变化很小，不能反映语音信号的幅度变化，信号的变化细节就看不出来了；反之，如果窗长太短，滤波器的通带太宽，随时间有急剧的变化，不能得到平滑的能量函数等。

标准： 一帧内含有1~7个基音周期。10kHz采样下，窗长可以去100~200个点。

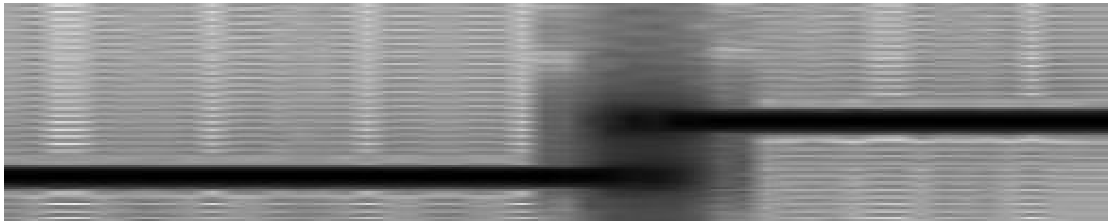


N = 64　　N = 128

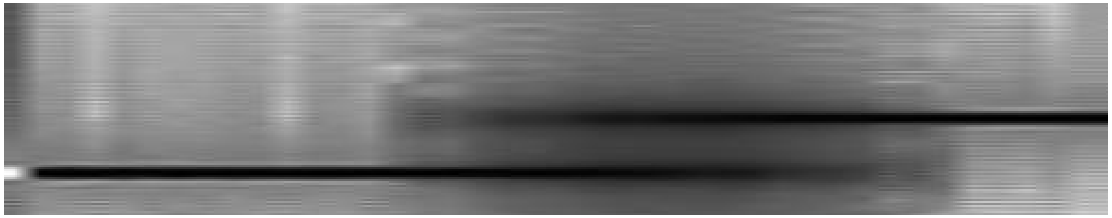
音节[fei]某一帧的短时频谱(横轴表示频率(Hz),
纵轴表示对数能量密度谱



波形



N = 64



N = 128

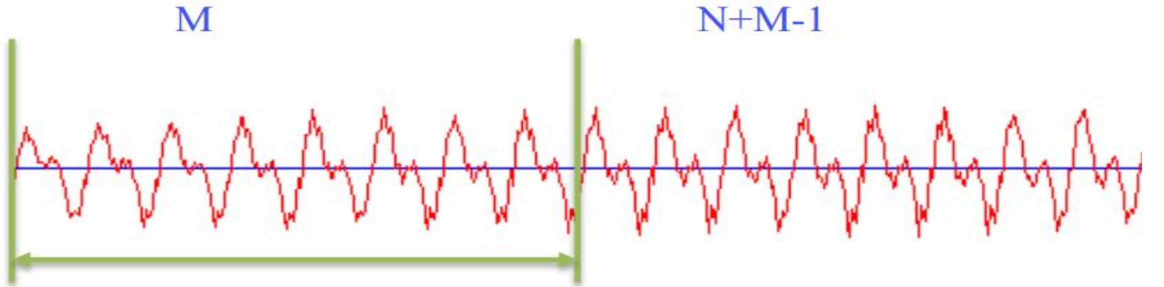
常用的短时分析技术

常用的短时分析技术

- 短时能量
- 短时平均幅度
- 短时平均过零率
- 短时自相关函数
- 短时平均幅度差
- 短时频谱
- 短时功率谱

短时能量分析

- 语音信号s(n)



- 短时能量

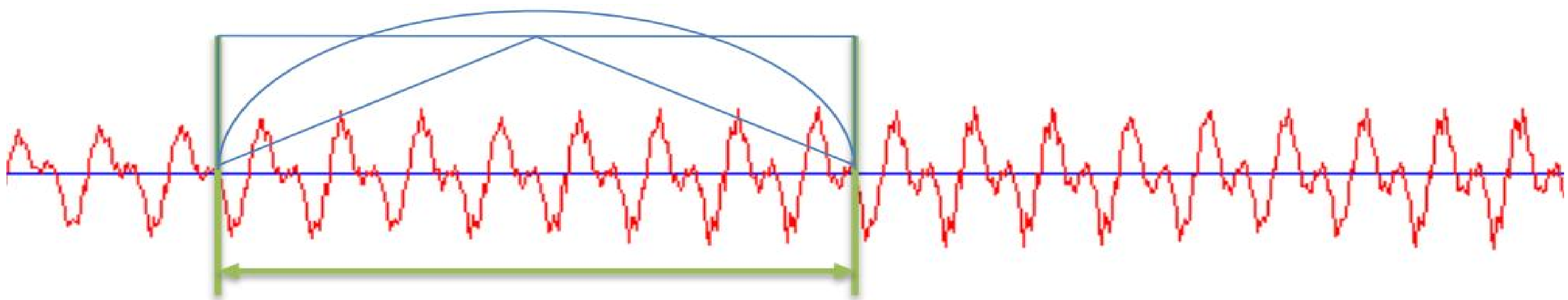
$$E_M = \sum_{m=M}^{N+M-1} [s(m)]^2 \qquad E_{n\times M} = \sum_{m=n\times M}^{N+n\times M-1} [s(m)]^2$$

N为语音短时分析的帧长

M为语音短时分析的帧移

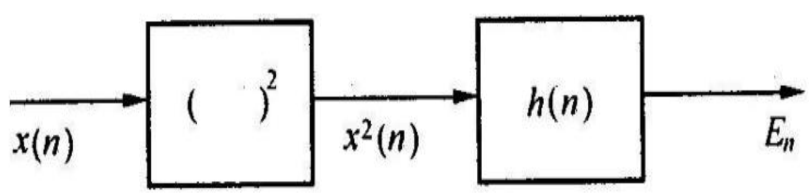
短时能量分析

决定短时能量特性有两个条件：不同的窗口形状和长度。窗长越长，频率分辨率越高；而时间分辨率越低。



$$E_{n\times M} = \sum_{m=n\times M}^{N+n\times M-1}[s(m)w(m-n\times M)]^2$$

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



短时过零率

- 一帧语音中语音信号波形穿过横轴(零电平)的次数
- 离散信号：样本改变符号的次数
- 定义

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

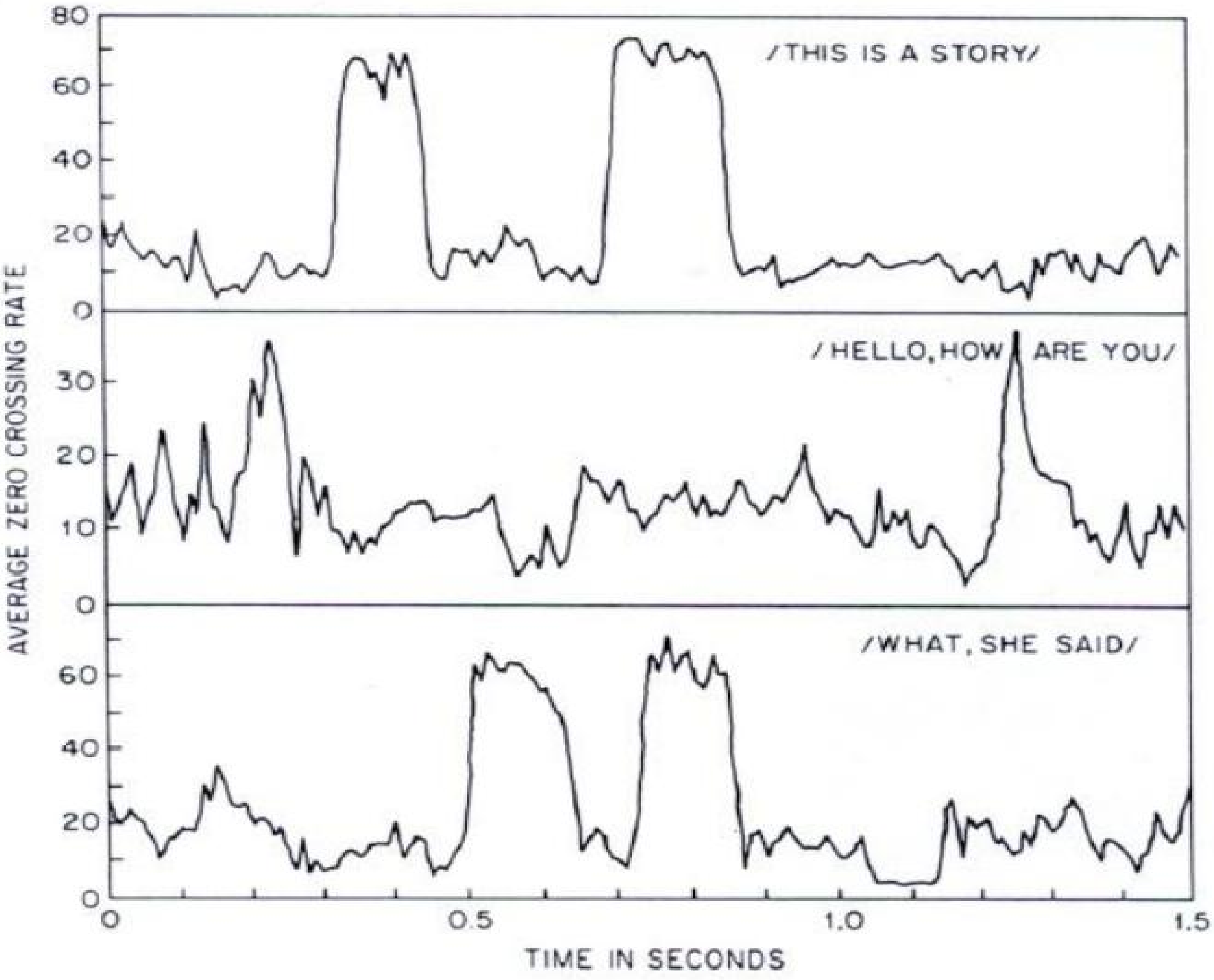
$$= \left| \mathrm{sgn}[x_w(m)] - \mathrm{sgn}[x_w(m-1)] \right| * w(n)$$

其中

计算之前，需去除直流分量

$$\mathrm{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \qquad w(n) = \begin{cases} 1/2N & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$$

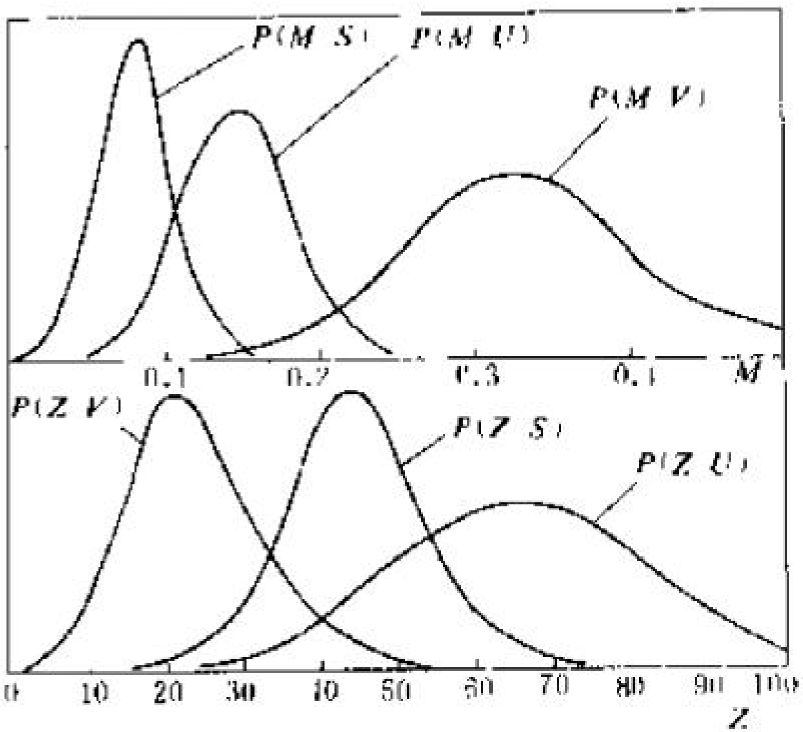
短时过零率



/THIS IS A STORY/
/HELLO, HOW ARE YOU/
/WHAT, SHE SAID/
AVERAGE ZERO CROSSING RATE
TIME IN SECONDS
0
20
40
60
80
10
30
0.5
1.0
1.5

短时过零率分析的意义

- 可以区分清音与浊音：浊音时具有较低的平均过零数，而清音时具有较高的平均过零数。
- 利用它可以从背景噪声中找出语音信号，可用于判断寂静无语音和有语音的起点和终点位置。
- 在背景噪声较小时用平均能量识别较为有效，而在背景噪声较大时用平均过零数识别较为有效。



在 S、U、V 三种情况下，短时平均幅度 M 和短时过零率 Z 的条件概率密度函数示意图

无声： S
清音： U
浊音： V

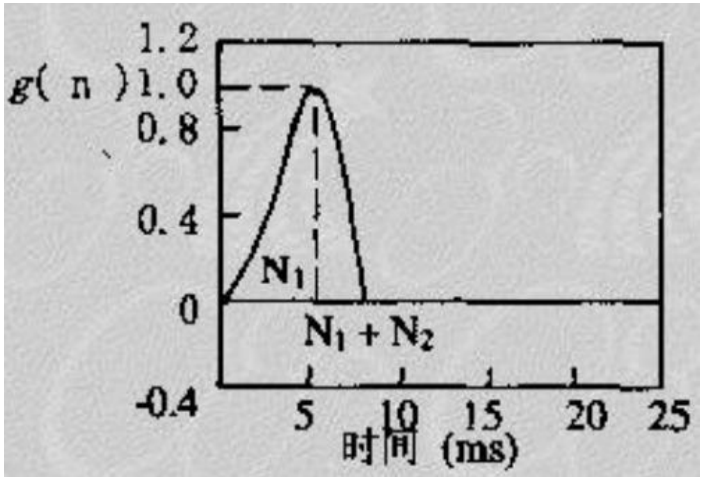
浊音的短时平均幅度最大，过零率最低

清音的短时平均幅度居中，过零率最高

无声的短时平均幅度最低，过零率居中

短时过零率

- 浊音：声带振动发出的语音，激励是以基音周期为周期的斜三角脉冲



- 清音：激励是随机白噪声

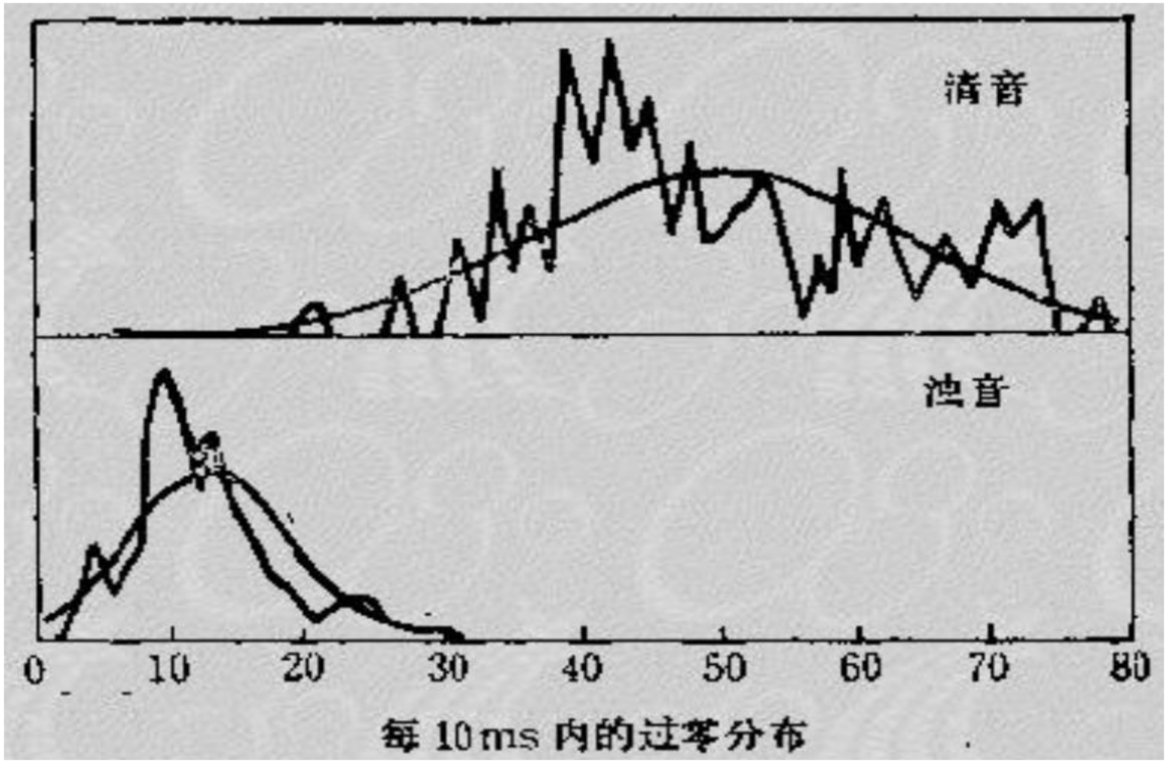
短时过零率

- 浊音频谱高频跌落，能量在3kHZ以下，过零率低
- 清音能量在高频上，过零率高

- 经验值：每10毫秒内
  - 清音： Zn >= 49
  - 浊音： Zn <＝14

- 短时过零率可用来判断清浊音

短时过零率

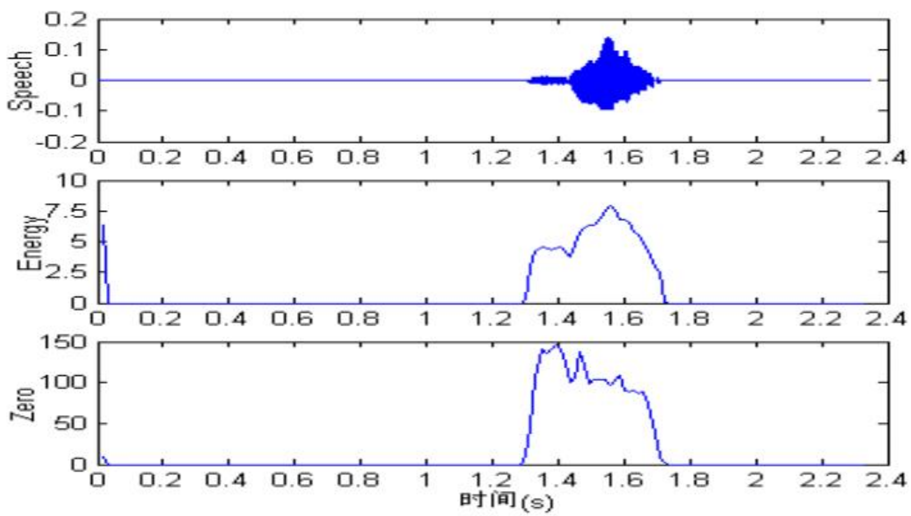


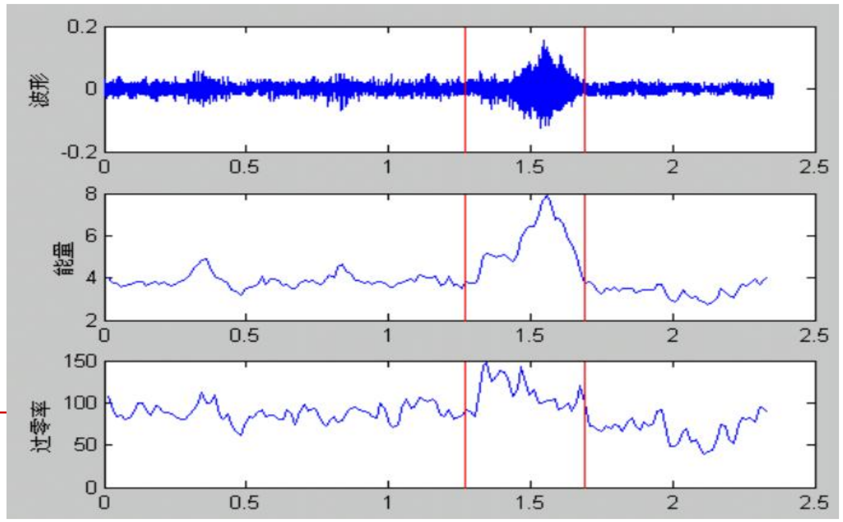
横轴为平均过零数，纵轴为概率

光滑曲线为模拟高斯分布

短时过零率的应用—语音端点检测

- 目的：从一段信号中确定语音的起始和结束点
- 困难： 1. 无声段噪声 2. 发音前后人为呼吸的杂音 3. 大多起点为清声母、塞擦音，与无声段噪音差别不大
- 末点检测影响不大，起点检测困难，且对语音识别影响大





短时自相关函数分析

- 对一帧内的语音 $x_n(m)$，短时自相关函数定义为

$$R_n(k)=\sum_{m=0}^{N-1-k}x_n(m)x_n(m+k)$$

$$=\sum_{m=-\infty}^{\infty}x(m)w(n-m)\bullet x(m+k)w(n-(m+k))$$

令$m+k=m'$,则

$$R_n(k)=\sum_{m'=-\infty}^{\infty}x(m'-k)x(m')w(n-m'+k)w(n-m')$$

$$=R_n(-k)$$

- 自相关用于研究信号本身，如信号波形的同步性、周期性等。

短时自相关函数分析

$$R_n(k)=\sum_{m=0}^{N-1-k}x_n(m)x_n(m+k)$$

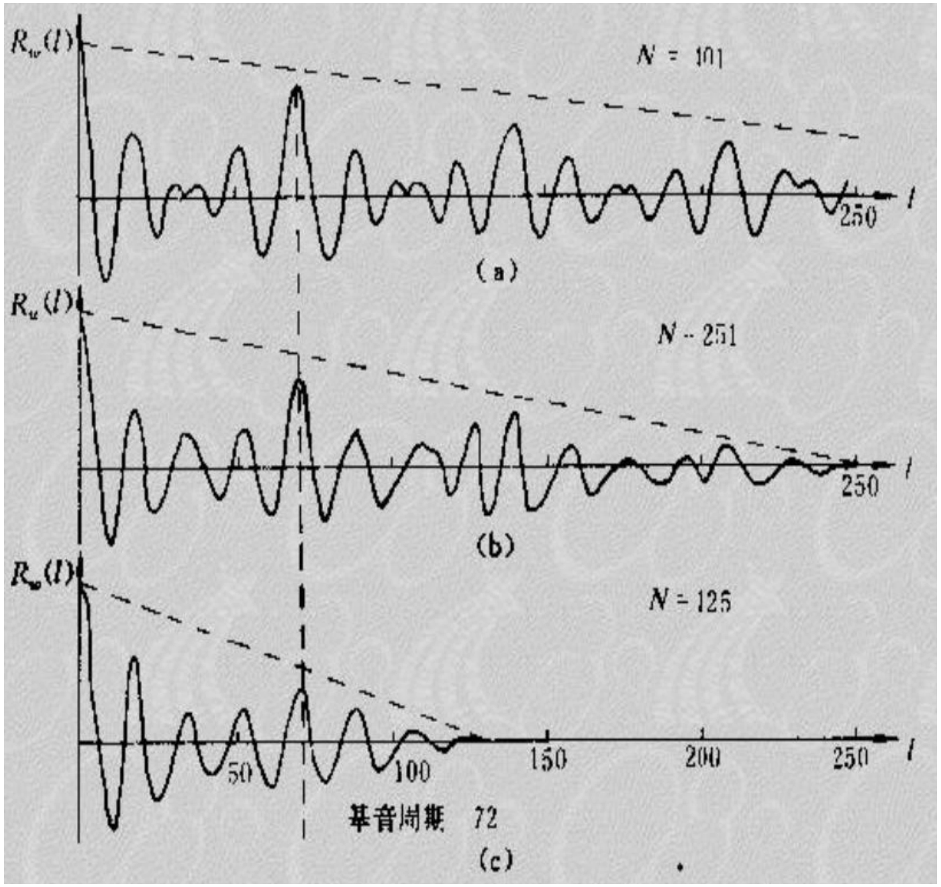
$$=\sum_{m=-\infty}^{\infty}x(m)w(n-m)\bullet x(m+k)w(n-(m+k))$$

讨论m的取值范围

$$0\le m\le N-1-k$$

缺点：

随着k的增加，进行乘积和的项数减少，总体上自相关函数的幅度值随着k增加而减少



不同窗长的短时自相关函数

短时自相关函数

- 修正的短时自相关函数
  - 用两个不同长度的窗口，长度相差最大的延迟点数K，保持乘积和项数不变—始终为短窗的长度。

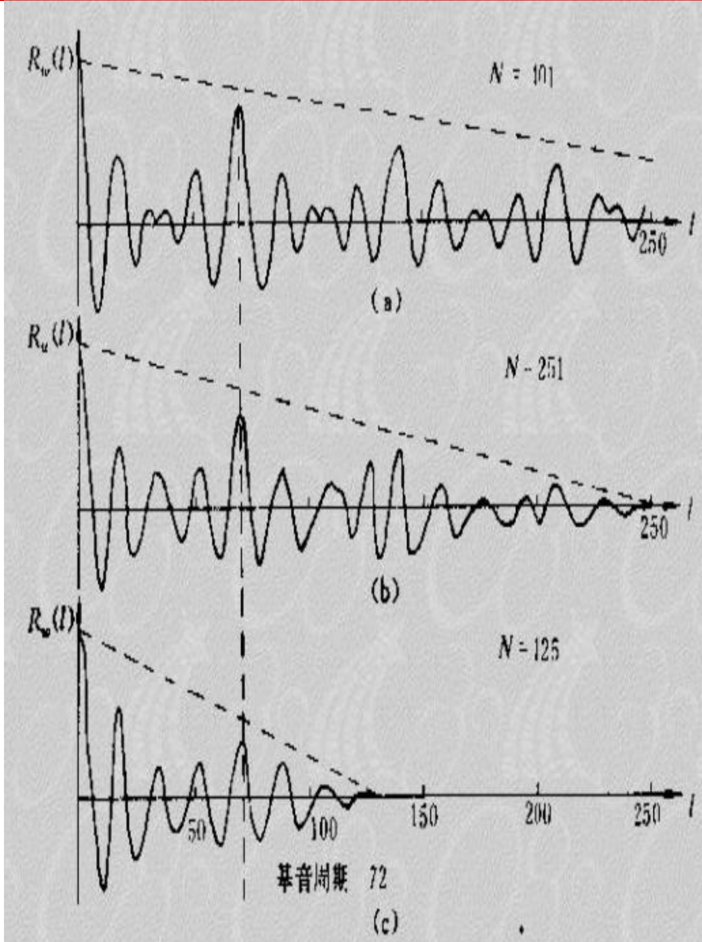
$$\hat{R}_n(k)=\sum_{m=0}^{N-1}x_n(m)x'_n(m+k) \qquad 0\le k\le K$$

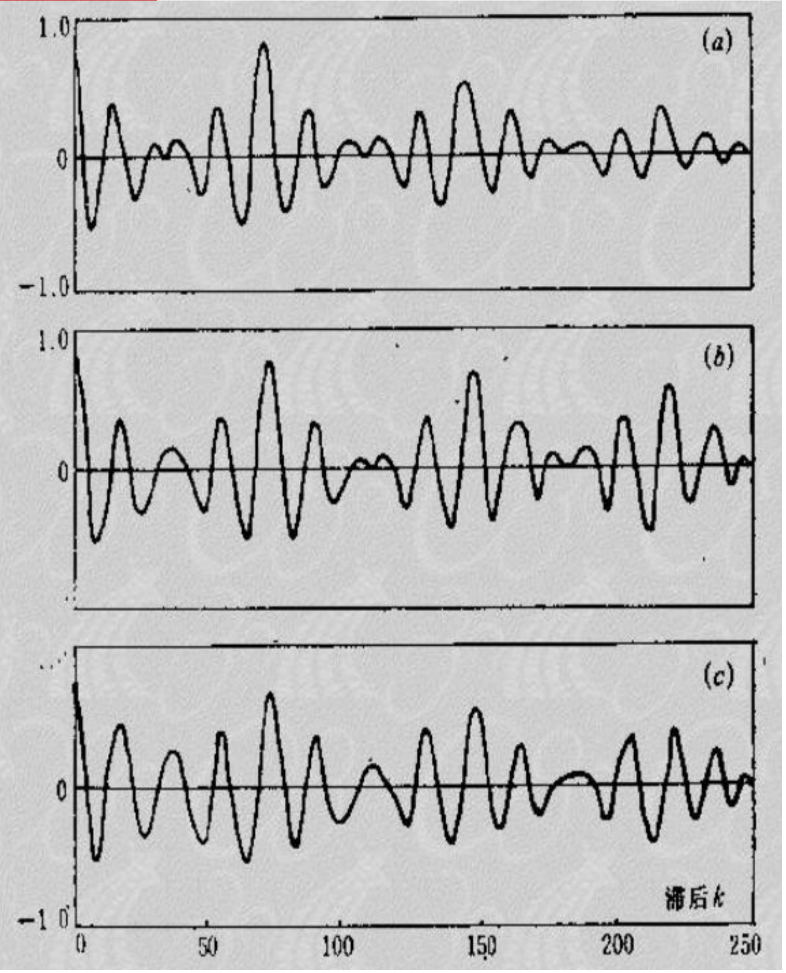
$$x_n(m)=w(m)x(n+m) \qquad 0\le m\le N-1$$

窗长相差恒为K

$$x'_n(m)=w'(m)x(n+m) \qquad 0\le m\le N-1+K$$

短时自相关函数





短时自相关函数　　　　修正短时自相关函数

- 短时自相关分析在语音识别中可有下面两个方面的应用:
  - 用来区分清音和浊音，因为浊音信号是准周期性的，对浊音语音可以用自相关函数求出语音波形序列的基音周期;
  - 另外在进行语音信号的线性预测分析时，也要用到短时自相关函数。

短时平均幅度差函数

- 如果信号是周期的，周期为N，则相距为周期的整数倍的样点上的幅值是相等的。

$$d(n)=x(n)-x(n+k), k=0,\pm N,\pm 2N\cdots$$

- 实际语音信号不为零，但值很小，这些极小值出现在整数倍周期位置上。

  定义如下：

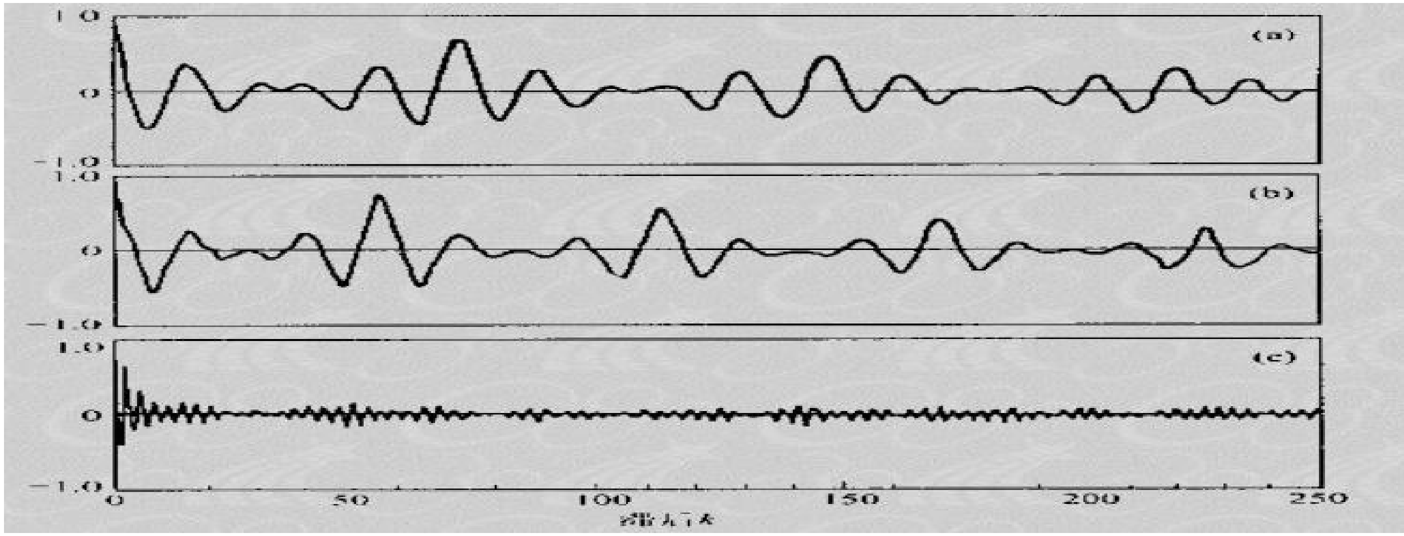
$$F_n(k)=\sum_{m=0}^{N-1-k}\left|s(n+m)w_1(m)-s(n+m-k)w_2(m-k)\right|$$

- 特性：
  - 若x(n)在窗口取值范围内有周期性（Np)，则 $F_n(k)$在 k=Np， 2Np…. 上出现极小值， 也$F_n(k)$准周期性
  - 与自相关函数的关系

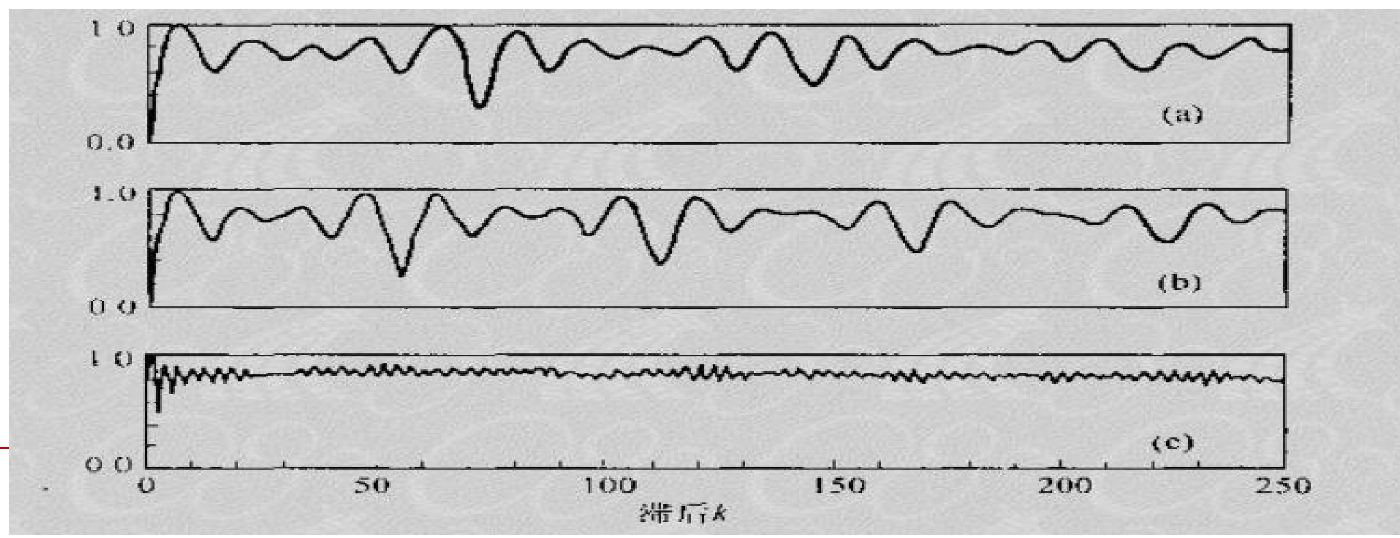
$$F_n(k) \approx \frac{\sqrt{2}}{R}\beta(k)[\hat{R}_n(0) - \hat{R}_n(k)]^{1/2}$$

  - $\beta(k)$对不同语音段在0.6～1.0之间变化

短时自相关函数



短时平均幅度差函数



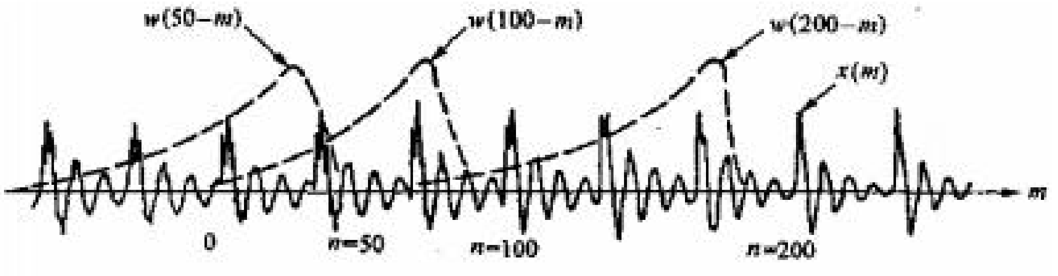
短时平均幅度差计算加、 减法和取绝对值的运算， 与自相关函数的相加与相乘的运算相比，其运算量大大减小，尤其在硬件实现语音信号分析时有很大好处。为此，AMDF已被用在许多实时语音处理系统中。

短时傅里叶变换

- 定义:

$$X_n(e^{jw}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-jwm}$$

短时傅里叶变换有两个自变量: n 和 w;
所以它既是关于时间 的离散函数, 又是关于角频率的连续函数。



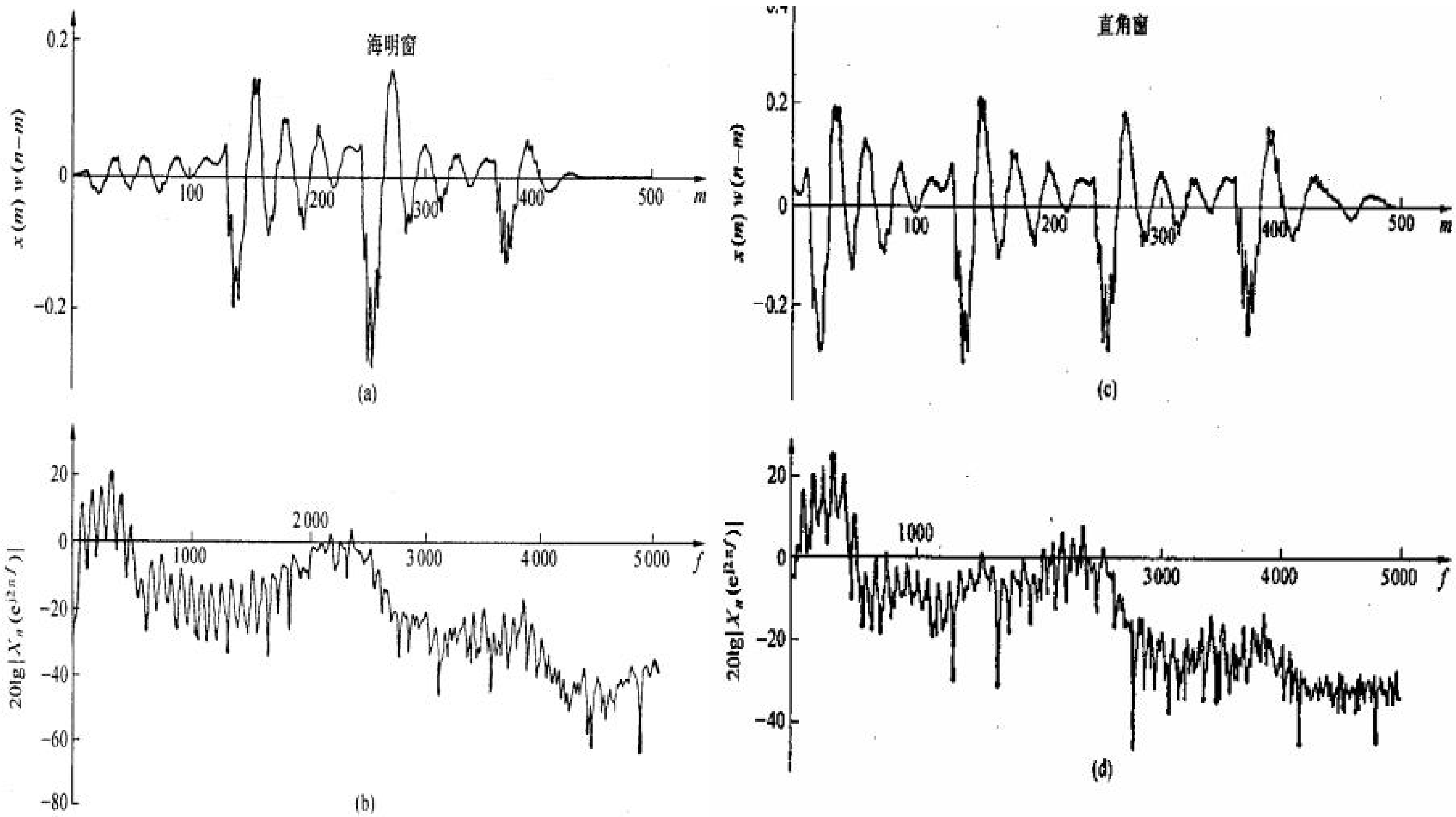
在几个 n 值上 x(m) 与 w(n − m) 的示意图

- 根据功率谱的定义，短时功率谱和短时傅里叶变换之间的关系为：

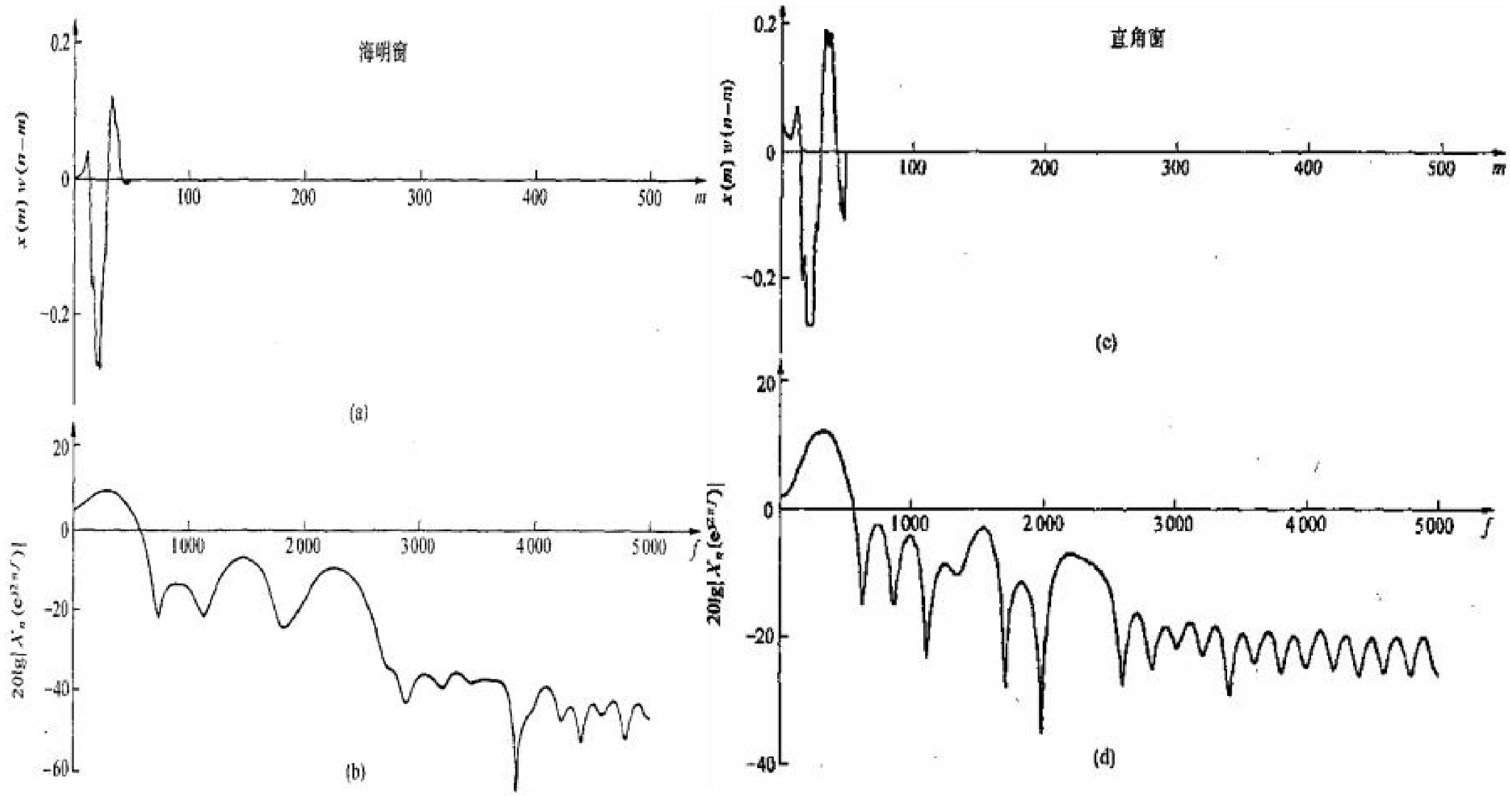
$$S_n(e^{j\omega}) = X_n(e^{j\omega})X^*{}_n(e^{j\omega}) = \left|X_n(e^{j\omega})^2\right|$$

短时功率谱是短时自相关函数的傅里叶变换：

$$R_n(k) = \sum_{m=-\infty}^{\infty} w(n-m)x(m)w(n-k-m)x(m+k)$$

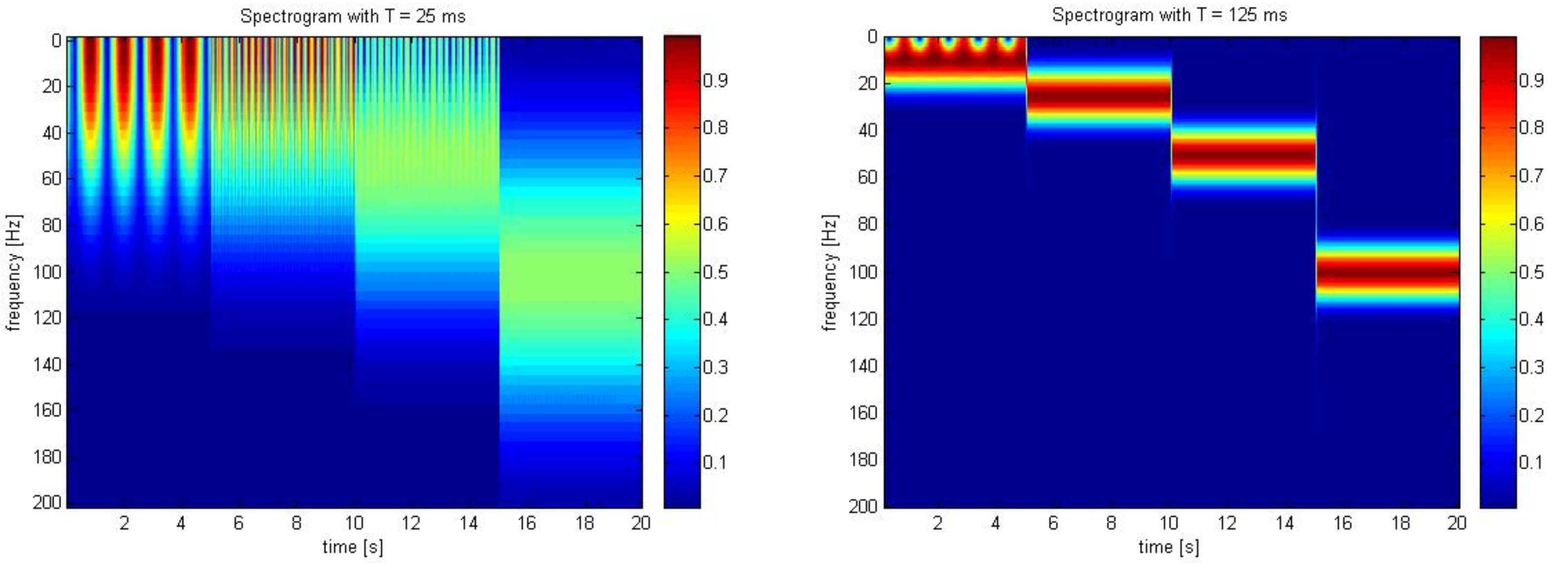


N＝500时海明窗与直角窗的浊音谱分析



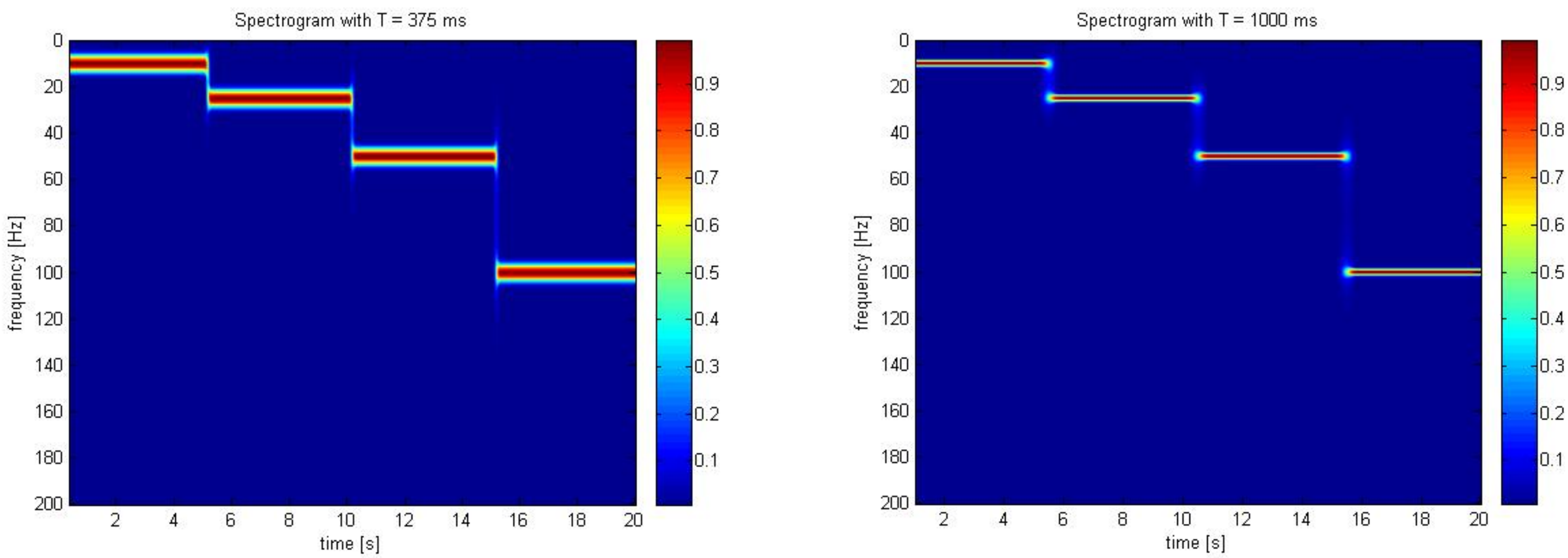
N＝50时海明窗与直角窗的浊音谱分析

不同窗口长度对STFT分辨率的影响



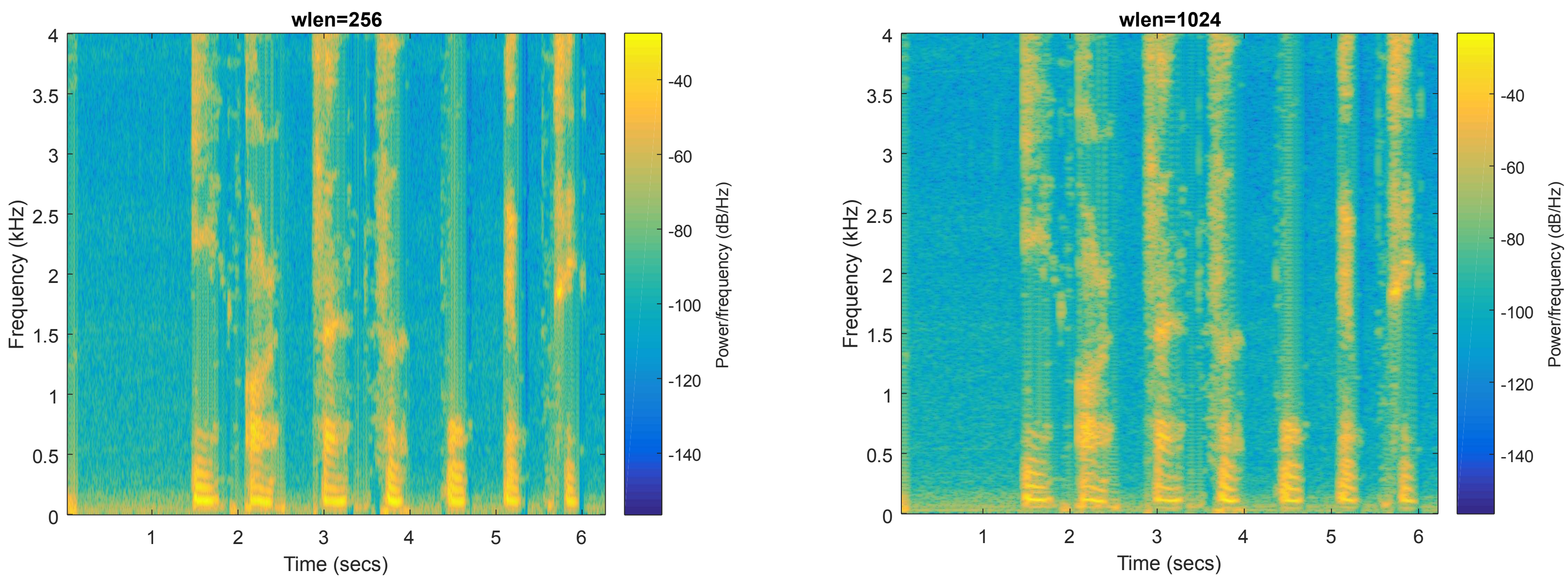
窄窗口提供良好的时间分辨率,可以准确地判断出信号变化的具体时刻，但信号频率分辨率低。

不同窗口长度对STFT分辨率的影响



宽窗口提供更好的频率分辨率，可以得到更准确的信号频率，但信号在时间上的变化时刻变得模糊。

窗长度对语谱图的影响



用于分类模型时，可以采用较大的窗长进行STFT变换，频域分辨率高，得到的时频图具有更明显的特征，易于分类。

STFT的局限性

- STFT的局限性在于它具有固定的分辨率。
- 加窗函数的宽度决定了是否有良好的频率分辨率（相近的频率分量可以分辨出来）或时间分辨率（精确反映出频率变化的时间），但二者不可兼得。
- 给出如下信号x(t)，4个单频信号每隔 5 秒依次出现，用不同长度的窗(winlen=25,125,375,1000 ms)进行短时傅里叶变换：

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

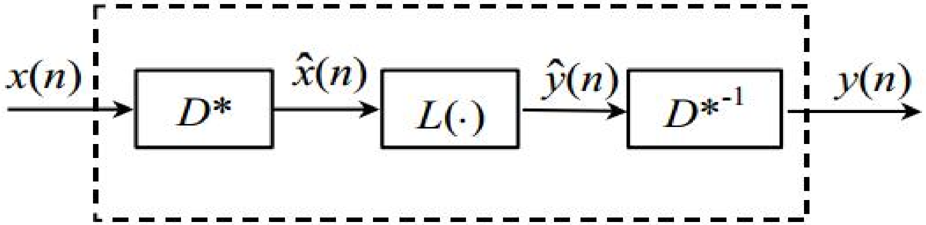
语音信号的倒谱分析

- 语音倒谱特征参数是由同态处理来实现的。
- 同态处理（同态滤波）：解卷将卷积关系变为求和处理。将语音信号的声门激励和声道响应分离开。

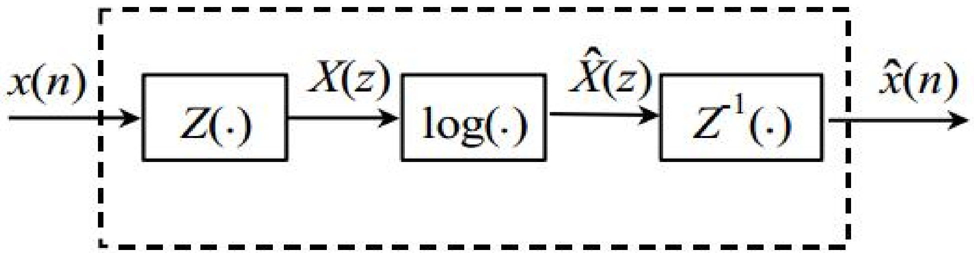
同态信号处理的基本原理

- 信号分类：加性信号、乘积性信号、卷积性信号等。
- 同态信号处理目的：将非线性问题转化为线性问题来处理。
- 同态信号处理分类：乘积同态处理和卷积同态处理两种。

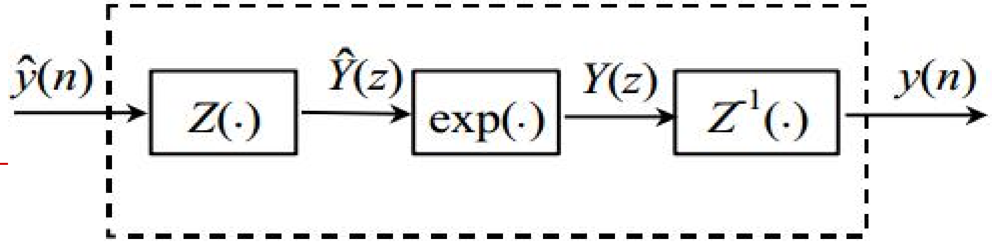
- 卷积同态系统:



- 特征系统D^*:



- 逆特征系统D^{*-1}:



- 特征系统D*

$$\begin{cases} F[x(n)] = X(e^{j\omega}) \\ \hat{X}(e^{j\omega}) = \ln[X(e^{j\omega})] \\ \hat{x}(n) = F^{-1}[\hat{X}(e^{j\omega})] \end{cases}$$

- 逆特征系统D*

$$\begin{cases} \hat{Y}(e^{j\omega}) = F[\hat{y}(n)] \\ Y(e^{j\omega}) = \exp[\hat{Y}(e^{j\omega})] \\ y(n) = F^{-1}[Y(e^{j\omega})] \end{cases}$$

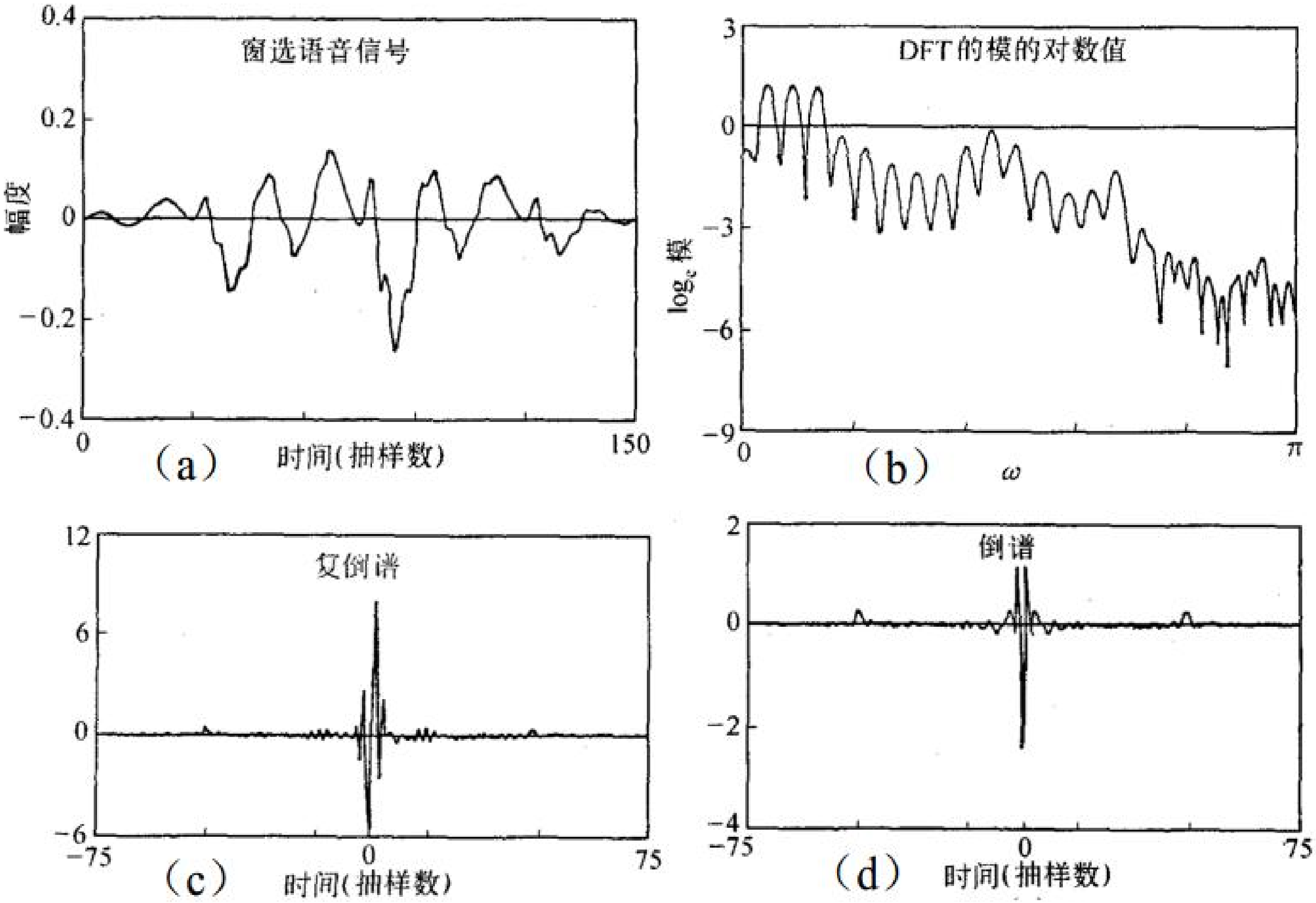
$$X(e^{j\omega}) = \left|X(e^{j\omega})\right| e^{j\arg[X(e^{j\omega})]}$$

$$\hat{X}(e^{j\omega}) = \ln[X(e^{j\omega})] = \ln\left|X(e^{j\omega})\right| + j\arg[X(e^{j\omega})]$$

只考虑$\hat{X}(e^{j\omega})$的实部：$c(n) = F^{-1}[\ln\left|X(e^{j\omega})\right|]$

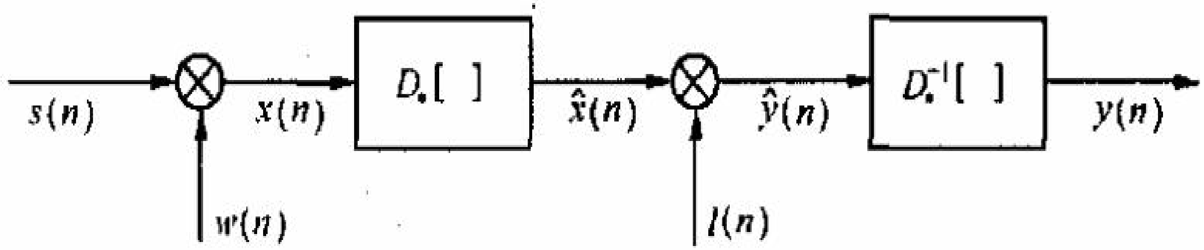
c(n)是序列x(n)对数幅度谱的傅里叶逆变换， c(n)称为“倒频谱”或简称为“倒谱”，有时也称“对数倒频谱”。

实例分析



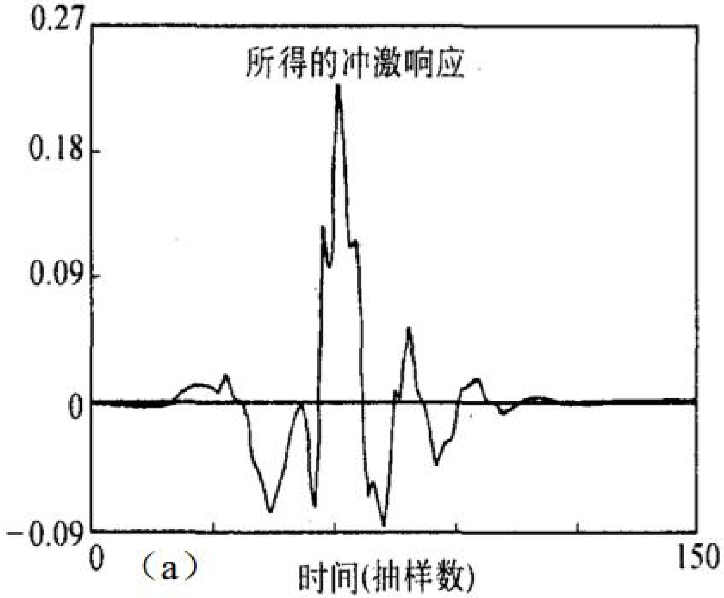
浊语音的倒谱和复倒谱实例

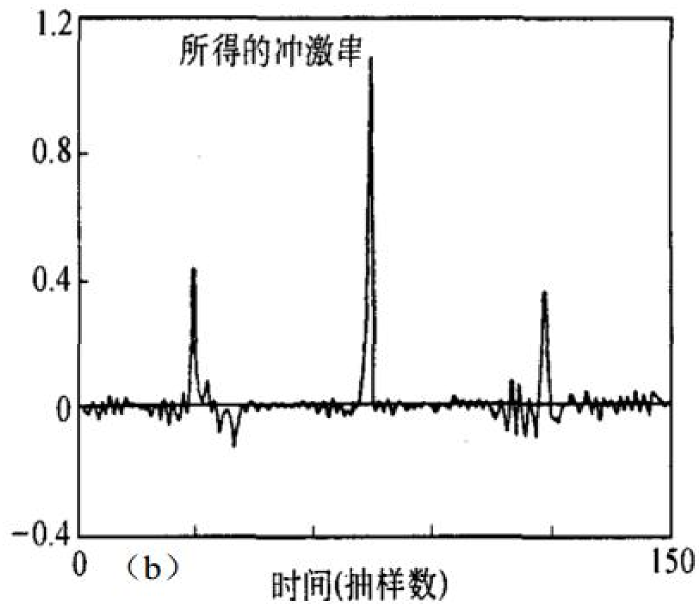
窗长为15ms，fs=10kHz，因此共包括150个语音样点。这段语音用海明窗加权，基音周期为Np=45。



语音同态滤波系统的构成

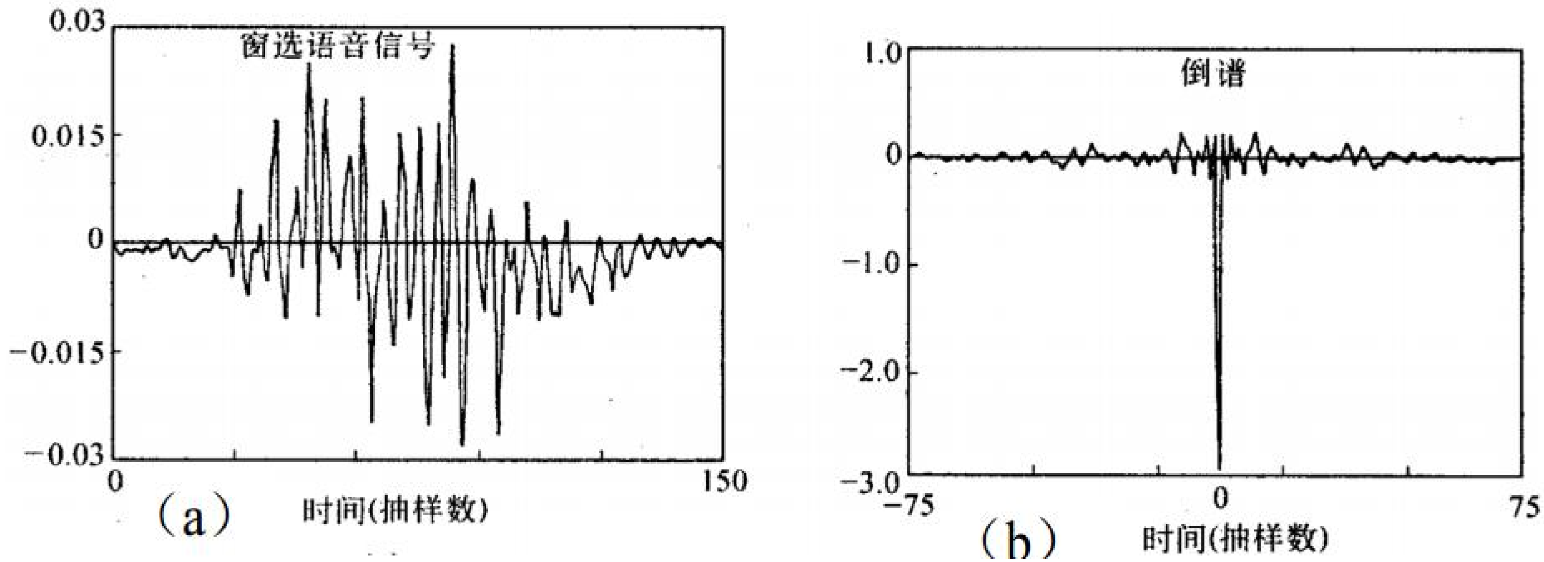
先用窗w(n)选择一个语音段，再计算复倒谱，然后将欲得到的复倒谱分量用一个“复倒谱窗”分离出来。所得到的窗选复倒谱用逆特征系统进行处理以恢复所需的卷积分量





浊音语音用同态滤波分离出**声门激励**和**声道响应**的示例

上图给出了经过滤波和逆特征系统处理后的结果。图(a)为经过低复倒谱窗l(n)和 之后的输出波形即声道冲击响应，图(b)给出了声门激励信号。可以看出声门激励波形近似于一个冲击串，其幅度随时间变化保持了用来加权输入信号所用的海明窗形状。



清语音的同态分析

上图给出了相同条件下一段加窗语音的时域波形及其倒谱。图(a)是一个海明窗乘过的清音语音段，图(b)为相应的倒谱。可见倒谱中没有出现在浊音情况下的那种尖峰，然而倒谱的低时域部分包含了关于声道冲击响应的信息。

Mel滤波器组倒谱系数

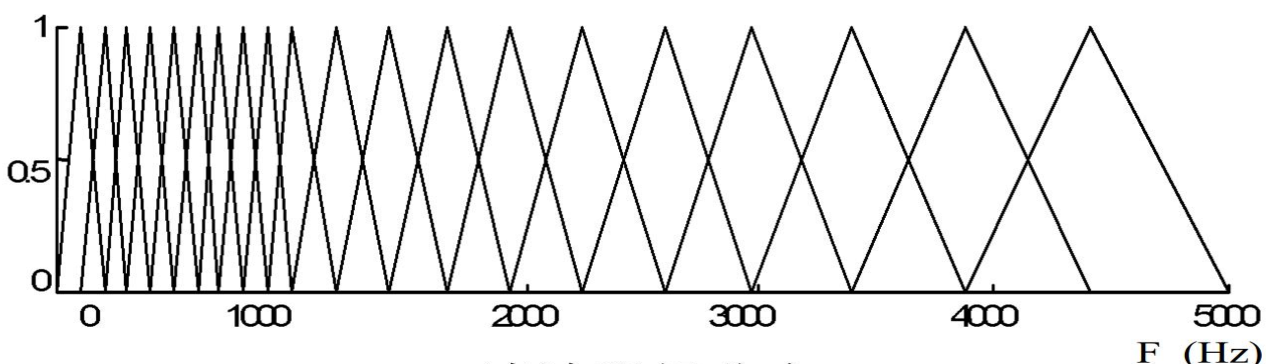
- Mel频率尺度—基于人耳的听觉特性

$$Mel(f) = 700\log_2(1 + f/700) = 2595\lg(1 + f/700)$$

Mel尺度均匀分布， f在1000Hz以下， 大致呈线性分布，带宽为100Hz左右， 在1000Hz以上， 呈对数增大。

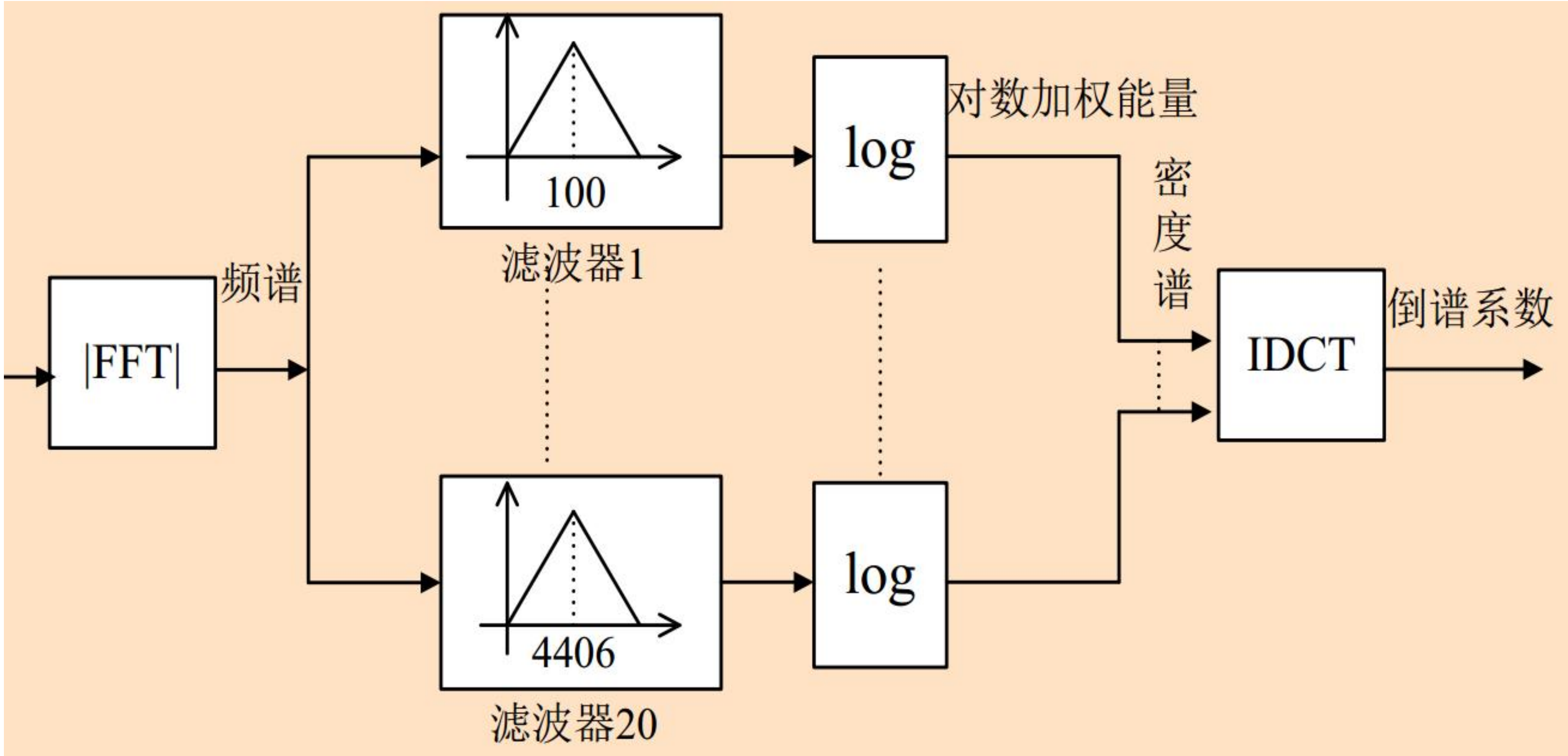
- Mel滤波器组 filter bank
  在Mel频率轴上配置L个三角形滤波器（triangle filter），其中心频率在Mel频率轴上等间隔分配（uniformly distributed）

$$c(l) \qquad \begin{cases} h(l) & \text{——第1个滤波器的上限频率} \\ c(l) & \text{——第1个滤波器的中心频率} \\ o(l) & \text{——第1个滤波器的下限频率} \end{cases}$$



Mel 滤波器组分布

MFCC计算



MFCC计算

- 计算每一个三角形滤波器的输出

$Xn(k)$：一帧语音的能量密度谱

$$m(l)=\sum_{k=c(l)}^{h(l)} w_l(k)X_n(k)$$

- 计算MFCC

$$W_l(k)=\begin{cases}\dfrac{k-o(l)}{c(l)-o(l)} & o(l)\le k\le c(l)\\[2ex] \dfrac{h(l)-k}{h(l)-c(l)} & c(l)\le k\le h(l)\end{cases}$$

$$MFCC(i)=\sqrt{\frac{2}{N}}\sum_{l=1}^{L}\log m(l)\cdot\cos\left\{(l-2)\frac{i\pi}{L}\right\}$$

i=1···M，

M—Mel倒谱系数维数

小作业(需随课布置):特征提取

- **基础要求**：从 PHONE_001.wav 提取基础 filter bank 特征，观察特征分布特点
  - 原始数据形式： 8k16bit pcm
  - 截止频率： 60Hz 3400Hz
  - 三角窗数量： 15 组
- **进阶自选 1** ： 在 filter bank 特征基础上进一步提取 MFCC 特征
- **进阶自选 2** ： 求取特征 3 阶差分并进行离线 cepstral mean and variance normalization (CMVN) CMVN
- **提交要求**
  - 特征
  - 源代码
  - 演示 PPT, 每周会挑选出优秀作业做报告, 额外加分
- **提交时间**
  - 10月15号

谢谢！