



第三章 语音信号的短时域分析

1

3.1 概述

2

3.2 语音信号的预处理

3

3.3 短时平均能量

4

3.4 短时平均幅度函数

5

3.5 短时平均过零率

6

3.6 短时自相关分析

7

3.7 基于能量和过零率的语音端点检测

8

3.8 基音周期估值



3.1 概述

语音信号是一种非平稳的时变信号，它携带着各种信息。在语音编码、语音合成、语音识别和语音增强等语音处理中都需要提取语音中包含的各种信息。

语音处理的目的是：对语音信号进行分析，提取特征参数，用于后续处理；加工语音信号。

总之，语音信号分析的目的就在于方便有效的提取并表示语音信号所携带的信息。





根据所分析的参数类型，语音信号分析可以分成时域分析和变换域（频域、倒谱域）分析。其中时域分析方法是最简单、最直观的方法，它直接对语音信号的时域波形进行分析，提取的特征参数主要有语音的**短时能量**和**平均幅度**、**短时平均过零率**、**短时自相关函数**和**短时平均幅度差函数**等。





3.2 语音信号的预处理

在对语音信号进行数字处理之前，首先要将模拟语音信号 $s(t)$ 离散化为 $s(n)$. 实际中获得数字语音的途径一般有两种，正式的和非正式的。

正式的是指大公司或语音研究机构发布的被大家认可的语音数据库，非正式的则是研究者个人用录音软件或硬件电路加麦克风随时随地录制的一些发音或语句。



语音信号的频率范围通常是300~3400Hz，一般情况下取采样率为8kHz即可。我们讨论的数字语音处理对象为语音数据文件，是已经数字化了的语音。

有了语音数据文件后，对语音的预处理包括：预加重、加窗分帧等。





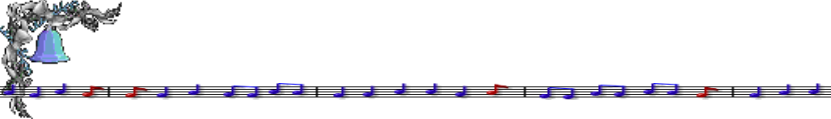
3.2.1 语音信号的预加重处理

预加重目的：为了对语音的高频部分进行加重，去除口唇辐射的影响，增加语音的高频分辨率。可通过一阶FIR高通数字滤波器来实现：

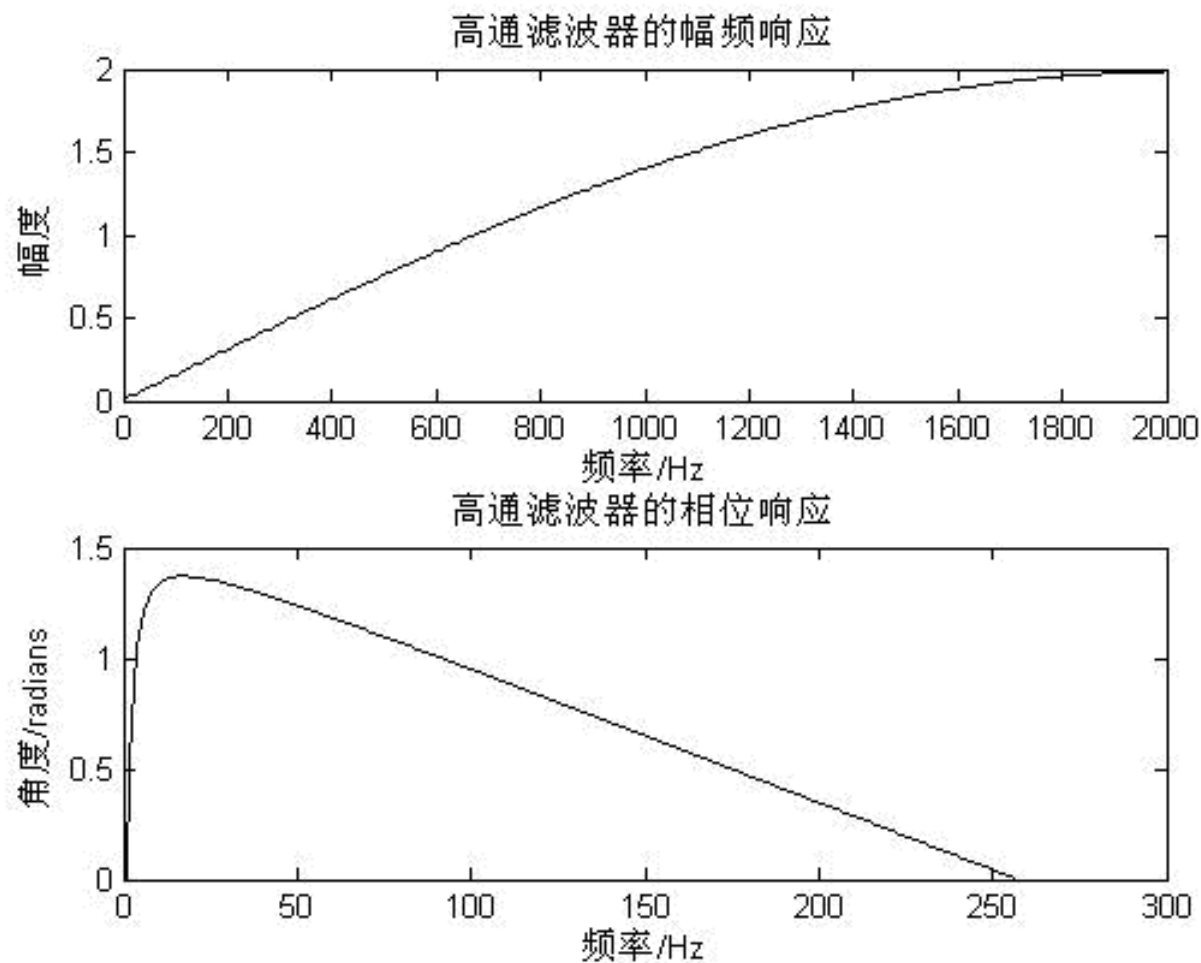
$$H(z) = 1 - \alpha z^{-1}$$

设 n 时刻的语音采样值为 $x(n)$ ，经过预加重处理后的结果为

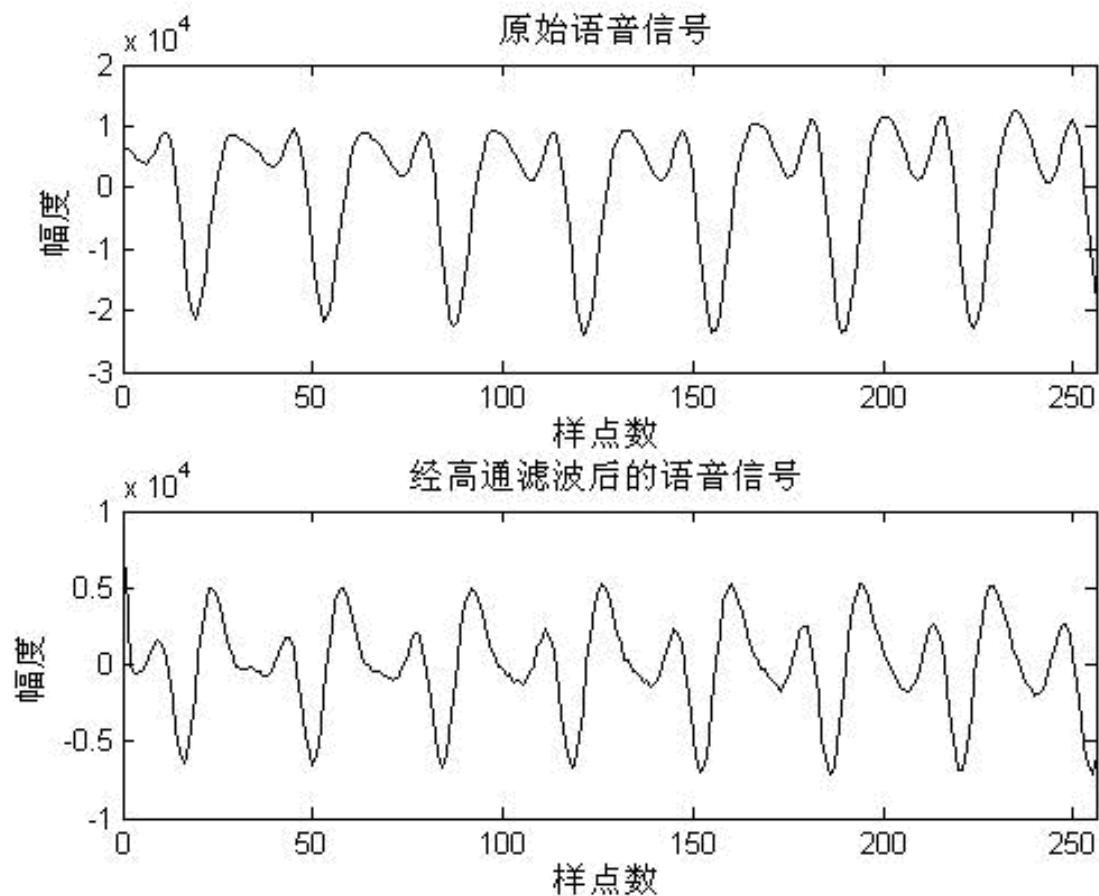
$$y(n) = x(n) - \alpha x(n-1)$$

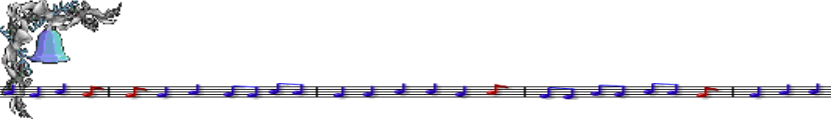


高通滤波器的幅频特性和相频特性如下

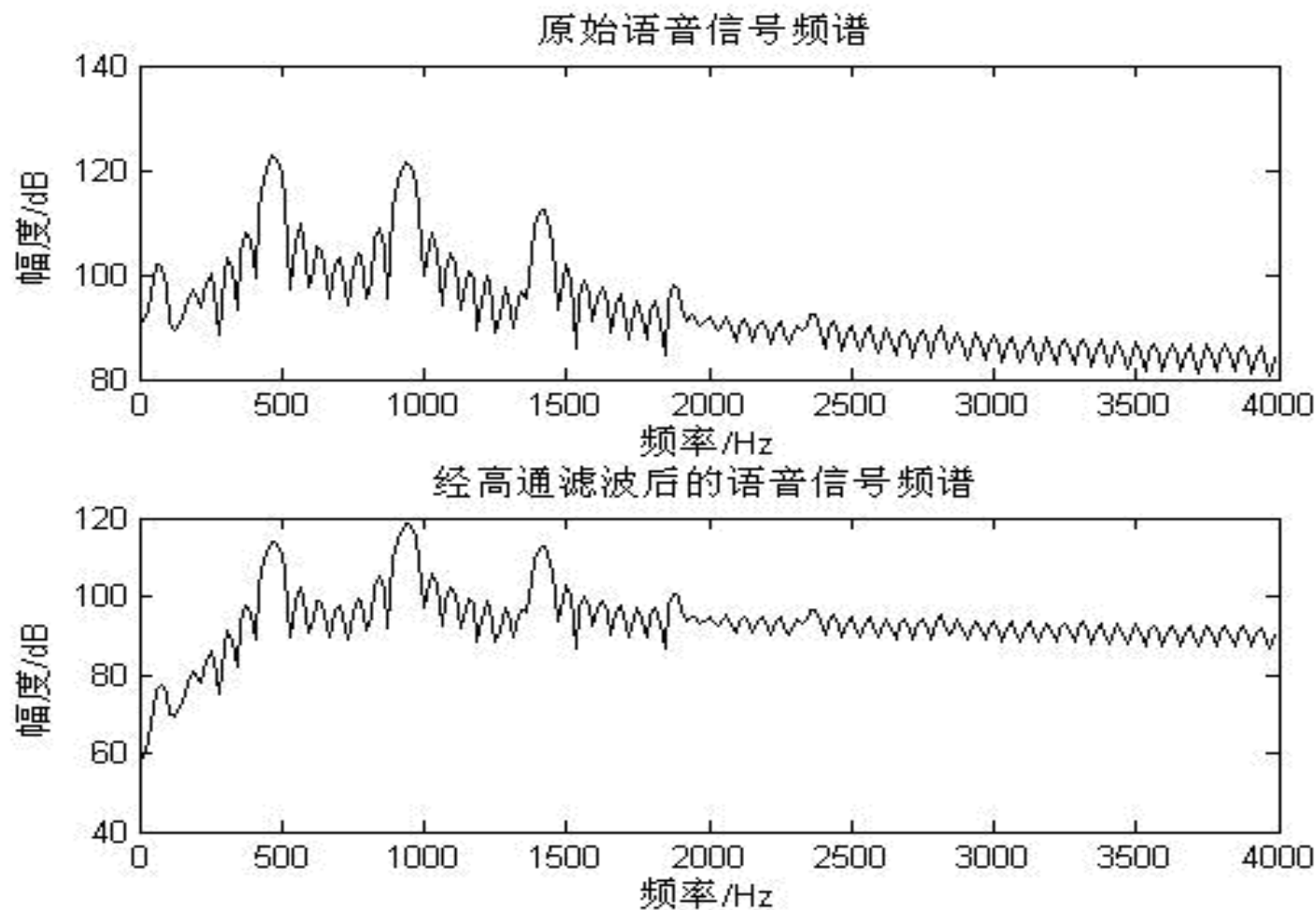


预加重前和预加重后的一段语音信号时域波形





预加重前和预加重后的一段语音信号频谱





3.2.2 语音信号的加窗处理

由于发音器官的惯性运动，可以认为在一小段时间里（一般为10ms~30ms）语音信号近似不变，即语音信号具有短时平稳性。这样，可以把语音信号分为一些短段（称为分析帧）来进行处理。





语音信号的分帧实现方法：

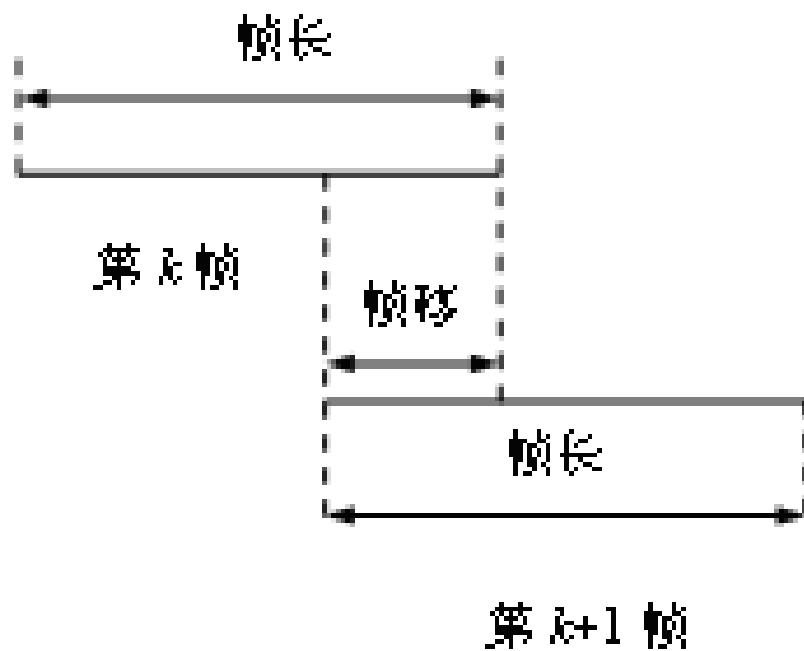
采用可移动的有限长度窗口进行加权的方法来实现的。一般每秒的帧数约为33~100帧。

分帧一般采用交叠分段的方法，这是为了使帧与帧之间平滑过渡，保持其连续性。前一帧和后一帧的交叠部分称为帧移，帧移与帧长的比值一般取为0~1/2。





图3.3给出了帧移与帧长示意图。





加窗常用的两种方法：

矩形窗，窗函数如下：

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$$

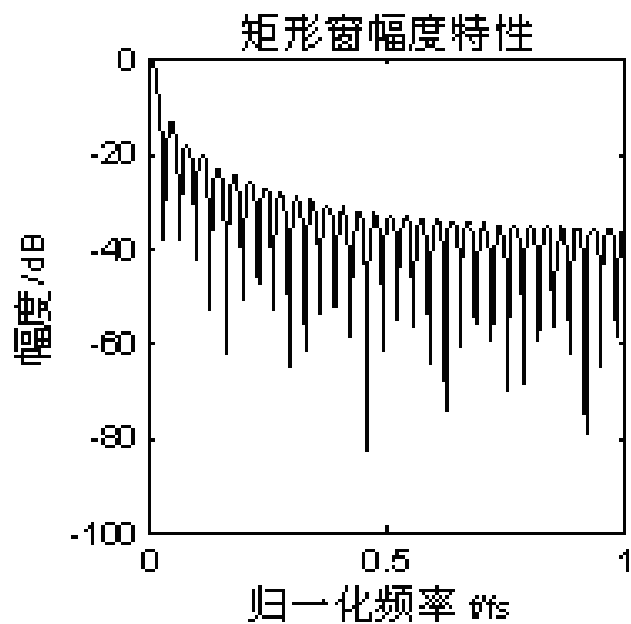
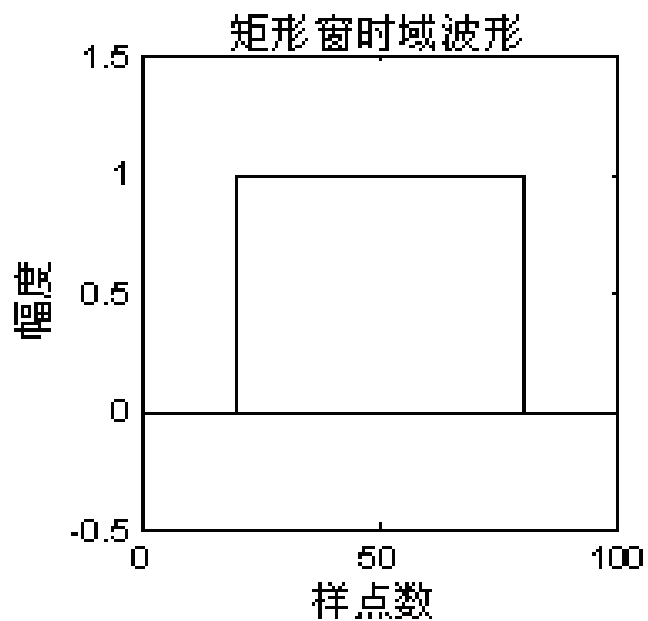
汉明(Hamming)窗，窗函数如下

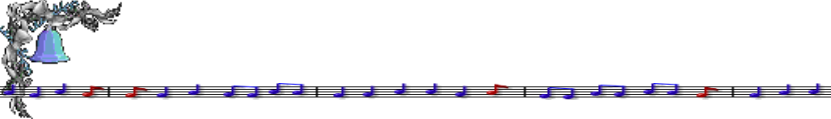
$$w(n) = \begin{cases} 0.54 - 0.46 \cos \left[2\pi n / (N-1) \right] , & 0 \leq n \leq N \\ 0 & \text{其它} \end{cases}$$



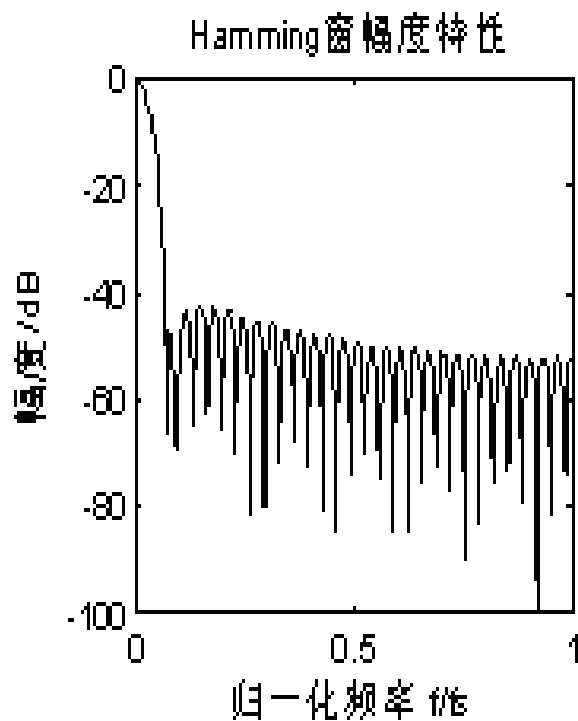
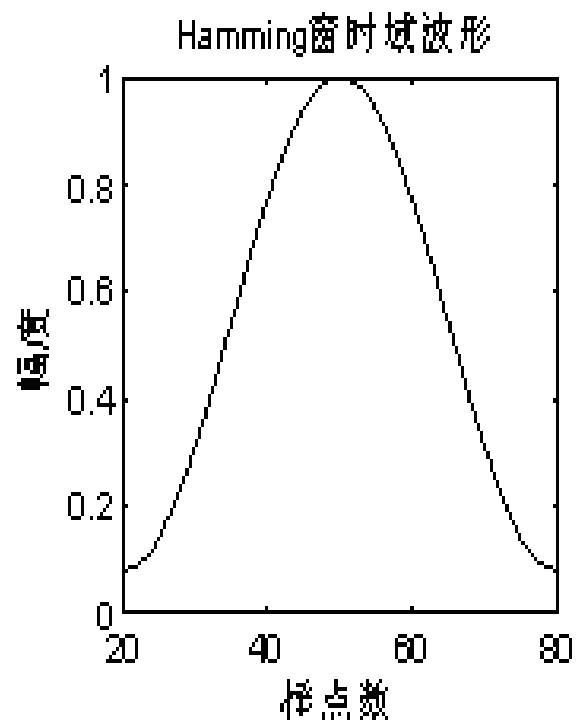


矩形窗及其频谱如下



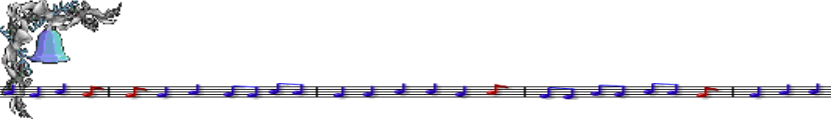


汉明窗及其频谱如下

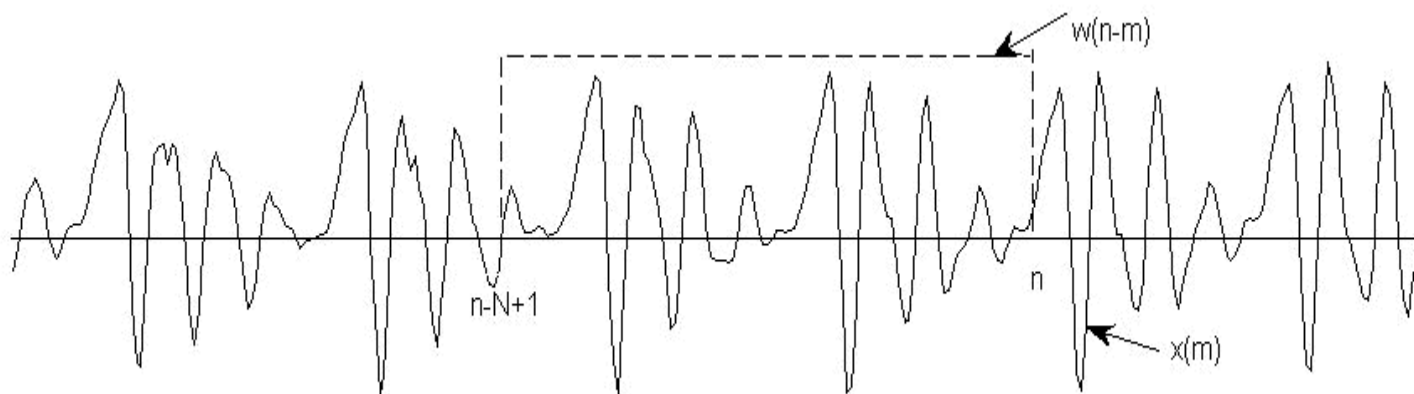


思考：两种窗效果有何异同？





加窗方法示意图：





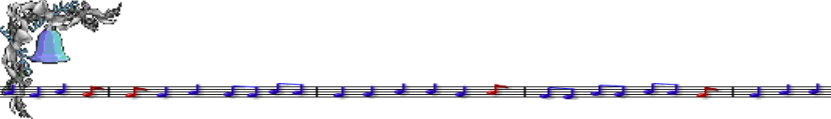
窗长的选择

一般选取100~200。原因如下：

当窗较宽时，平滑作用大，能量变化不大，故反映不出能量的变化。

当窗较窄时，没有平滑作用，反映了能量的快变细节，而看不出包络的变化。





语音信号的分帧处理，实际上就是对各帧进行某种变换或运算。设这种变换或运算用 $T[]$ 表示， $x(n)$ 为输入语音信号， $w(n)$ 为窗序列， $h(n)$ 是与 $w(n)$ 有关的滤波器，则各帧经处理后的输出可以表示为：

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]h(n-m)$$





几种常见的短时处理方法是：

1. $T[x(m)] = x^2(m)$, $h(n) = w^2(n)$ Q_n 对应于能量；

2. $T[x(m)] = |\text{sgn}[x(m)] - \text{sgn}[x(m-1)]|$, $h(n) = w(n)$

Q_n 对应于平均过零率；

3. $T[x(m)] = x(m)x(m+k)$, $h(n) = w(n)w(n+k)$

Q_n 对应于自相关函数；





3.3 短时平均能量

1. 短时平均能量定义

定义 n 时刻某语音信号的短时平均能量 E_n 为：

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

当窗函数为矩形窗时，有

$$E_n = \sum_{m=n-(N-1)}^n x^2(m)$$



若令

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

则短时平均能量可以写成：

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

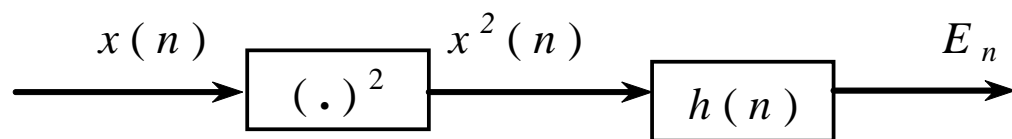


图 3.7 语音信号的短时平均能量实现方框图





2. E_n 特点： E_n 反映语音信号的幅度或能量随时间缓慢变化的规律。

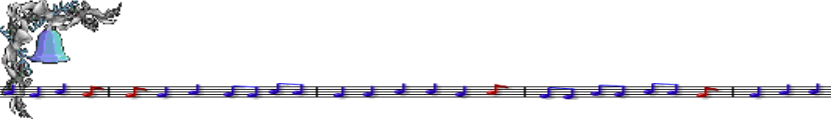
3. 窗的长短对于能否由短时能量反映语音信号的幅度变化，起着决定性影响。

如果窗选得很长， E_n 不能反映语音信号幅度变化。

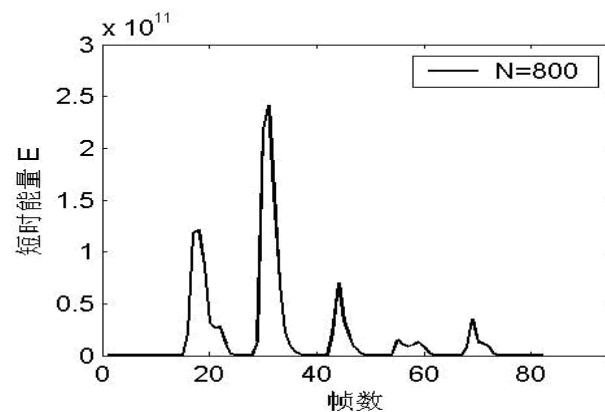
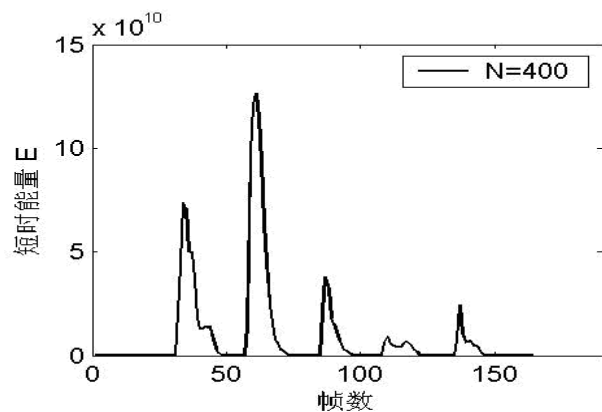
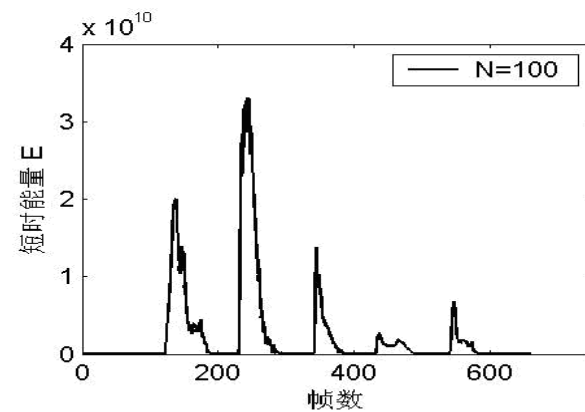
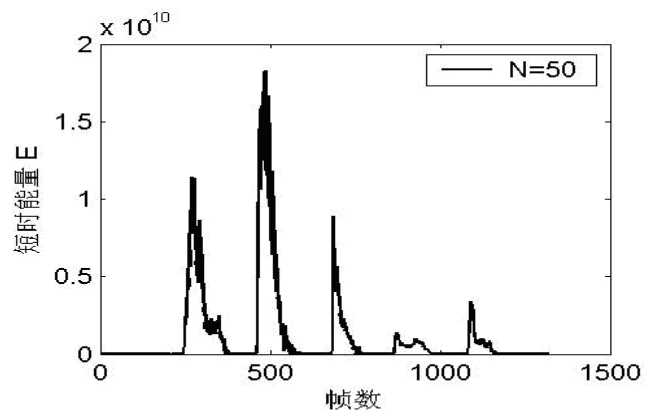
窗选得太窄， E_n 将不够平滑。

通常，当取样频率为 10kHz 时，选择窗宽度 $N=100\sim 200$ 是比较合适的。





不同矩形窗长 N 时的短时能量函数

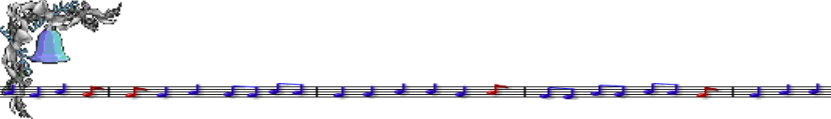




短时平均能量的主要用途如下：

- 1) 可以作为区分清音和浊音的特征参数。**
- 2) 在信噪比较高的情况下，短时能量还可以作为区分有声和无声的依据。**
- 3) 可以作为辅助的特征参数用于语音识别中。**





MATLAB的具体实现如下：

- 1、用Cooledit读入语音“我到北京去”。**
- 2、将读入的语音文件wav保存为txt文件，设置采样率为8kHz，16位，单声道。**

- 3、把保存的文件zqq.txt读入Matlab。**

```
fid=fopen('zqq.txt','rt'); x=fscanf(fid,'%f');
```

```
fclose(fid);
```

- 4、对采集到的语音样点值进行分帧。**





3.4 短时平均幅度函数

为了克服短时能量函数计算 $x^2(m)$ 的缺点，
定义了短时平均幅度函数：

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

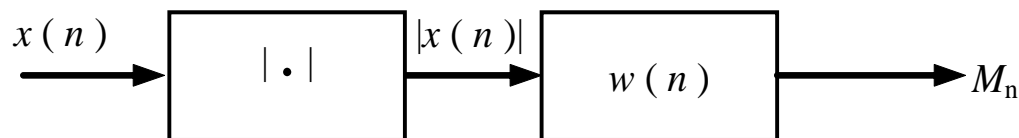
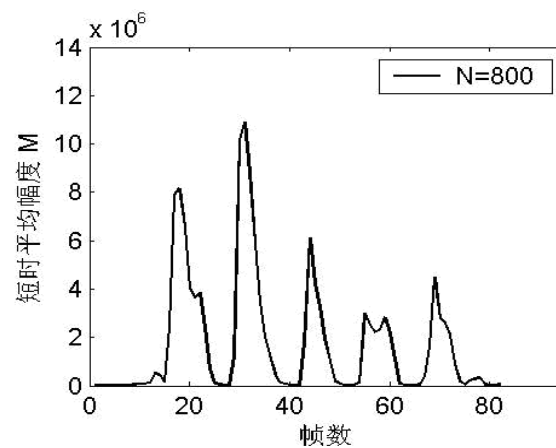
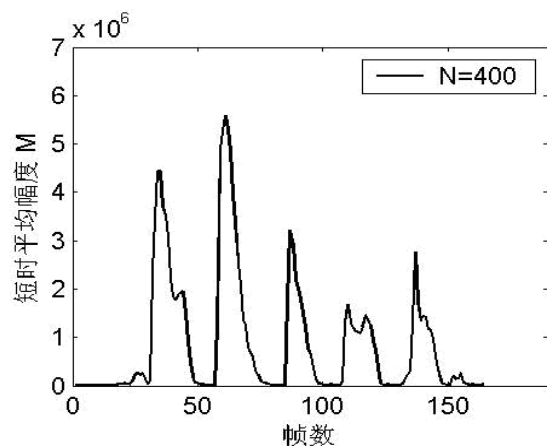
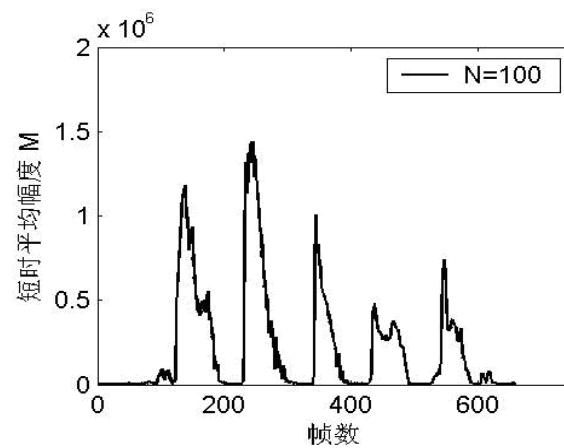
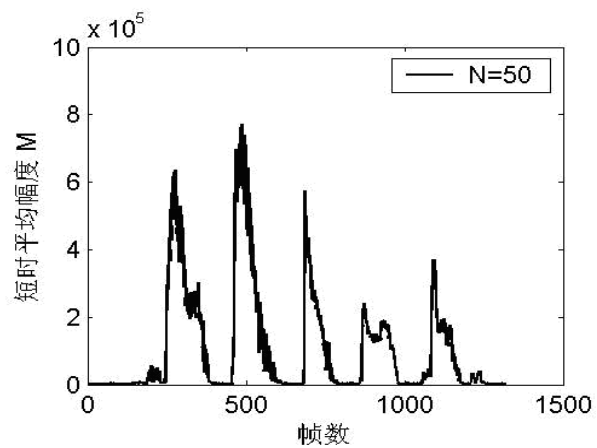


图 3.9 短时平均幅度

M_n 与 E_n 的比较:

1. M_n 能较好地反映清音范围内的幅度变化;
2. M_n 所能反映幅度变化的动态范围比 E_n 好;
3. M_n 反映清音和浊音之间的电平差次于 E_n 。





短时平均幅度函数随矩形窗窗长 N 变化的情况

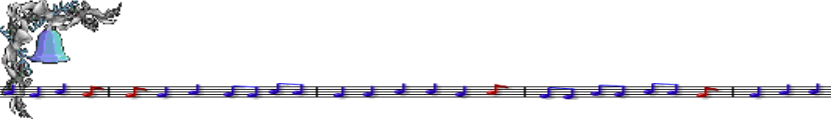


3.5 短时平均过零率

1. 定义

在离散时间语音信号情况下，如果相邻的采样具有不同的代数符号就称为发生了过零。单位时间内过零的次数就称为过零率。短时平均过零率的定义为

$$\begin{aligned} Z_n &= \sum_{m=-\infty}^{+\infty} \left| \text{sgn}[x(m)] - \text{sgn}[x(m-1)] \right| w(n-m) \\ &= \left| \text{sgn}[x(n)] - \text{sgn}[x(n-1)] \right| * w(n) \end{aligned}$$



$$w(n) = \begin{cases} \frac{1}{2N} & 0 \leq n \leq N-1 \\ 0 & \text{其它} \end{cases}$$

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

在上式中，用 $1/2N$ 作为幅值，是考虑了对该窗口范围内的过零数取平均的意思。





考虑到 $w(n-m)$ 的非零值范围为 $n-m \geq 0$ ，即 $m \leq n$ ，以及 $n-m \leq N-1$ ，故 $m \geq n-N+1$ ，因此短时平均过零率可以改写为：

$$Z_n = \frac{1}{2N} \sum_{m=n-(N-1)}^n |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]|$$

$$\begin{aligned} Z_n &= \sum_{m=-\infty}^{+\infty} |\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]| w(n-m) \\ &= |\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]| * w(n) \end{aligned} \quad (\text{定义式})$$



2. 实现短时平均过零率

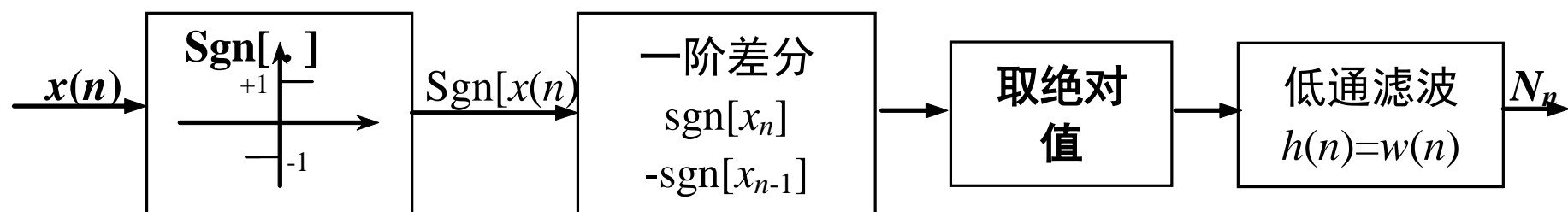
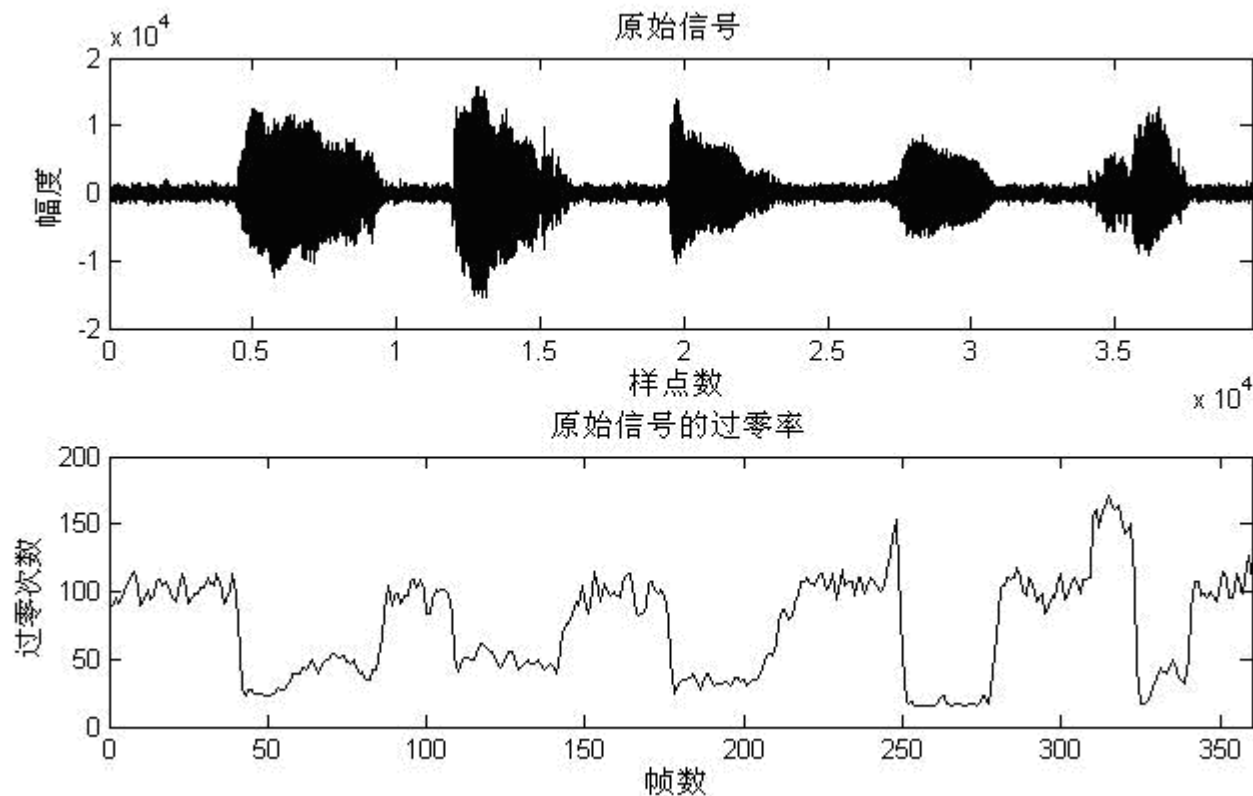


图 3.11 语音信号的短时平均跨零数



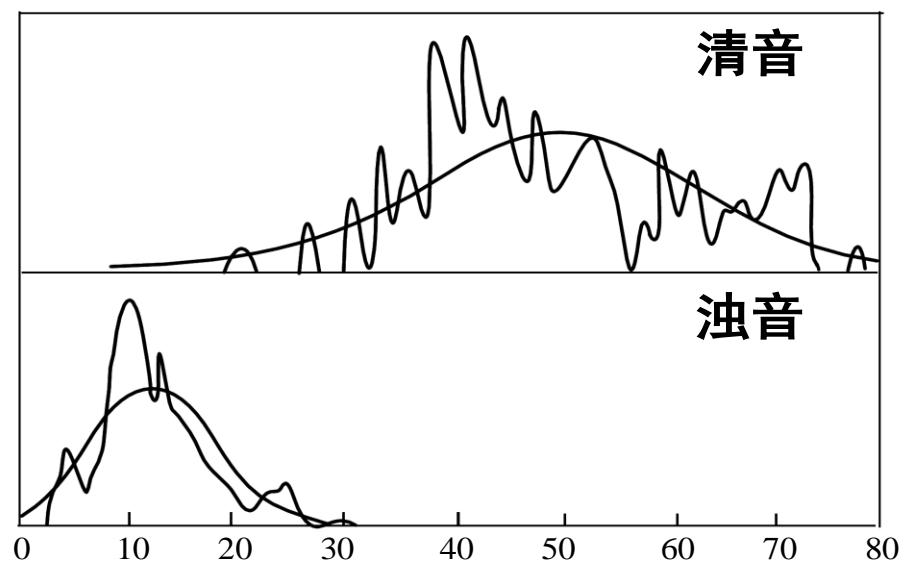
女声“我到北京去”的短时平均过零次数的变化曲线：



3. 应用

清音过零率高，浊音过零率低。

局限性：浊音和清音重叠区域只根据短时平均过零率不可能明确地判别清、浊音。



每 10ms 内的过零数

过零率概率分布



端点检测

端点检测目的：从包含语音的一段信号中确定出语音的起点及结束点。

有效的端点检测不仅能使处理时间减到最少，而且能抑制无声段的噪声干扰，提高语音处理的质量。





3.6 短时自相关分析

3.6.1 短时自相关函数

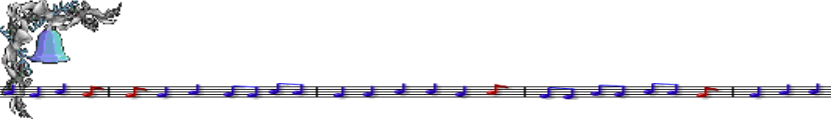
时域离散确定信号的自相关函数定义为：

$$R(k) = \sum_{m=-\infty}^{+\infty} x(m)x(m+k)$$

时域离散随机信号的自相关函数定义为：

$$R(k) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{m=-N}^N x(m)x(m+k)$$

周期为 P 的周期信号满足： $R(k) = R(k+P)$



自相关函数具有下述性质：

- (1) 对称性 $R(k) = R(-k)$**
- (2) 在 $k = 0$ 处为最大值，即对于所有 k 来说，
 $|R(k)| \leq R(0)$**
- (3) 对于确定信号， $R(0)$ 对应于能量**
对于随机信号， $R(0)$ 对应于平均功率





3.6.2 语音信号的短时自相关函数

采用短时分析方法，定义语音信号短时自相关函数为

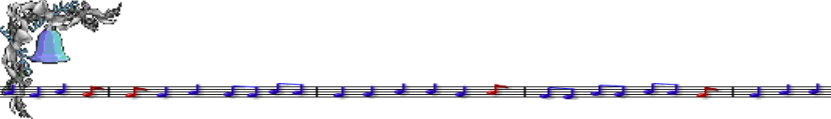
$$R_n(k) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)x(m+k)w(n-k-m)$$

因为

$$R_n(-k) = R_n(k)$$

所以

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



定义
$$h_k(n) = w(n)w(n+k) \quad (3-18)$$

那么短时自相关函数可以写成：

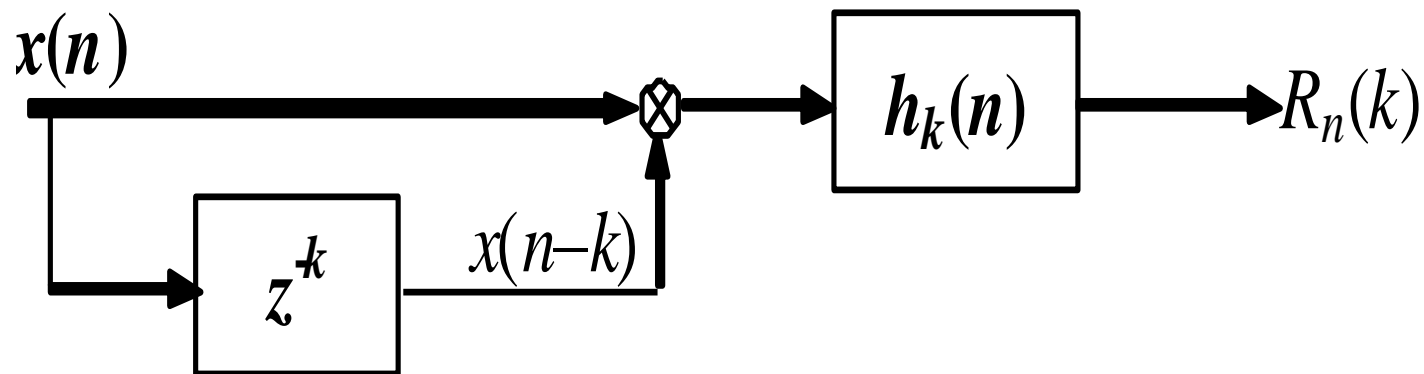
$$R_n(k) = \sum_{m=-\infty}^{+\infty} x(m)x(m-k)h_k(n-m)$$

上式表明，序列 $x(n)x(n-k)$ 经过一个冲激响应为 $h_k(n)$ 的数字滤波器滤波即得到短时自相关函数 $R_n(k)$





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





也可采用直接运算的方法，令

$$m = n + m'$$

$$w(-m) = w'(m)$$

则可得：

$$\begin{aligned} R_n(k) &= \sum_{m'=-\infty}^{+\infty} [x(n+m')w(-m')] [x(n+m'+k)w'(k+m')] \\ &= \sum_{m=-\infty}^{+\infty} [x(n+m)w'(m)][x(n+m+k)w'(k+m)] \end{aligned}$$

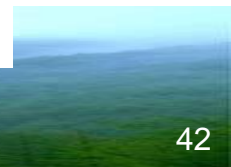
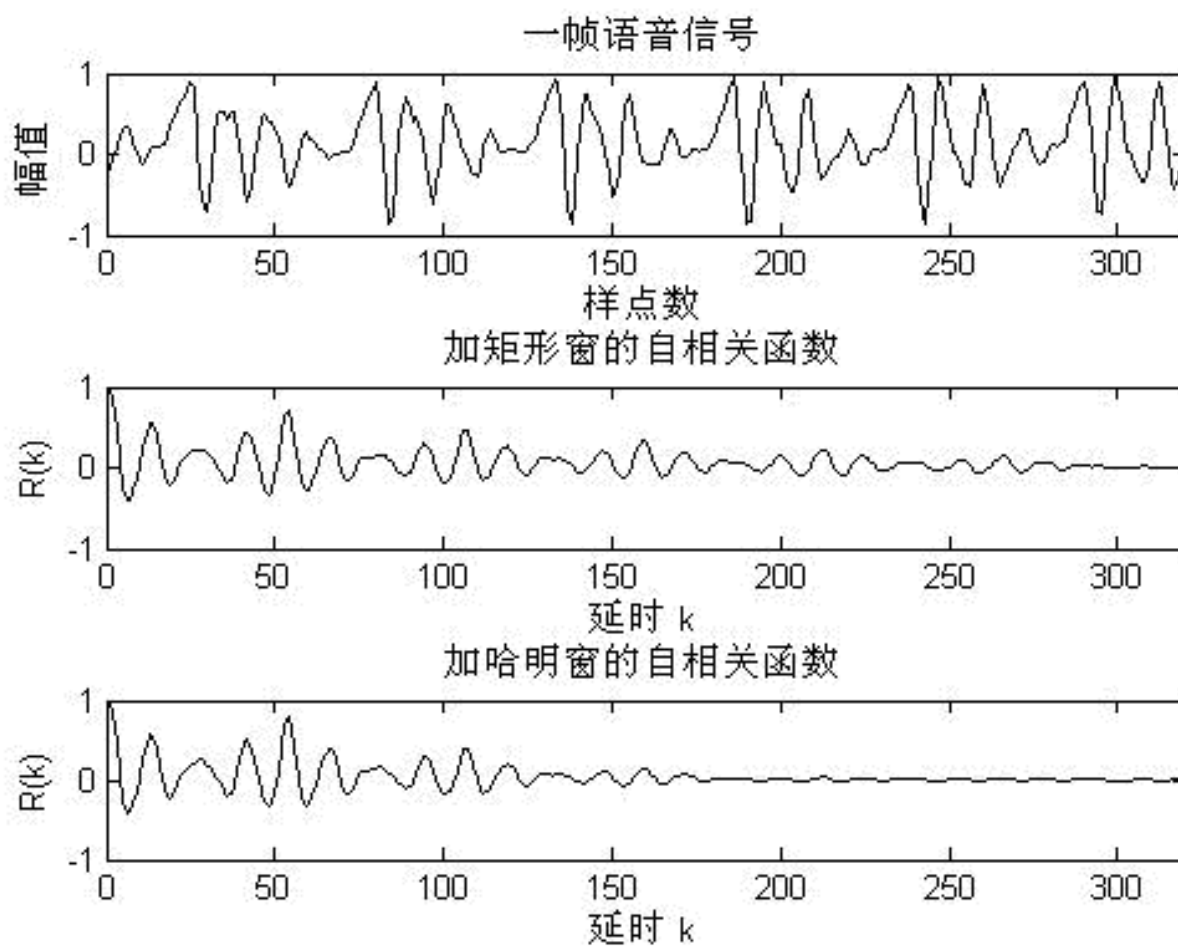
上式可以写成

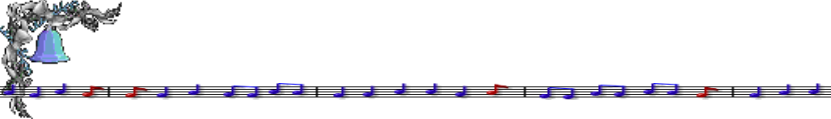
$$R_n(k) = \sum_{m=0}^{N-1-k} [x(n+m)w'(m)][x(n+m+k)w'(k+m)]$$



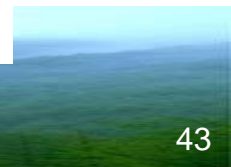
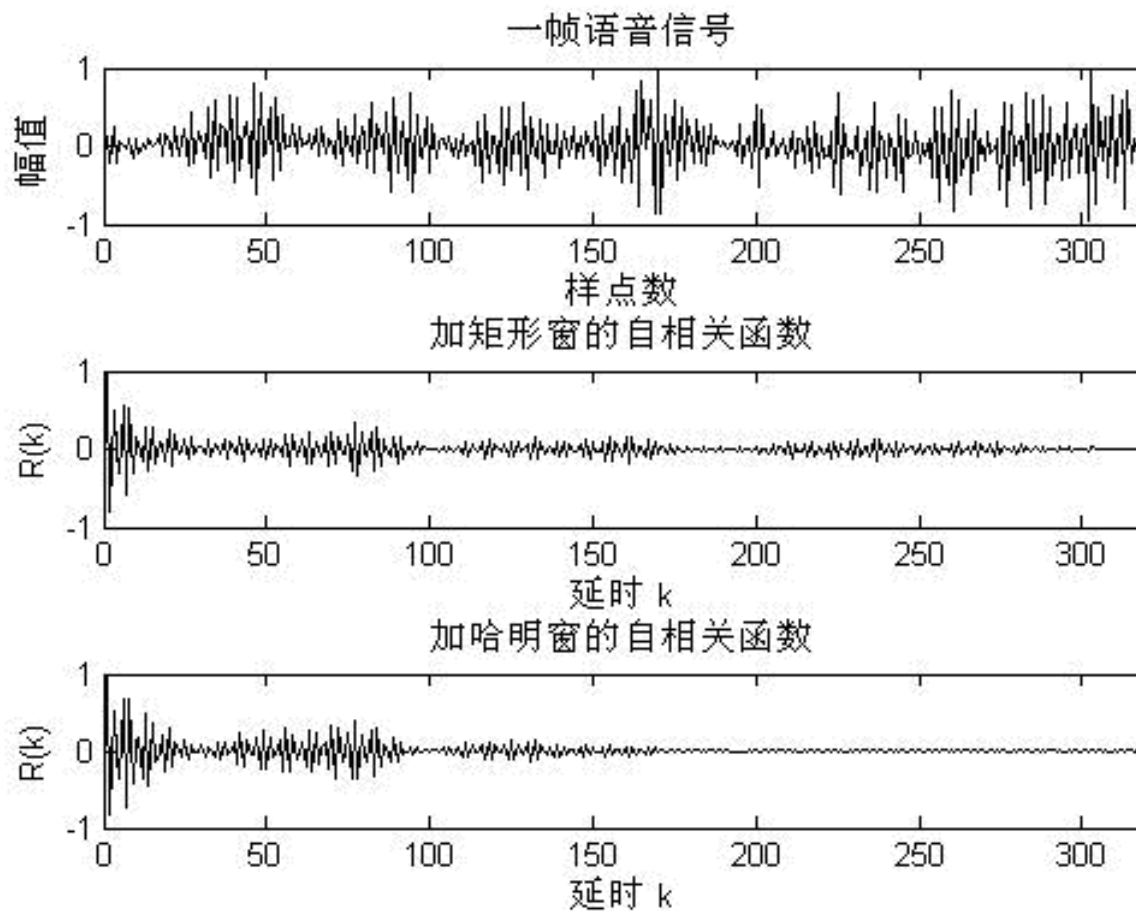


浊音的短时自相关函数





清音的短时自相关函数





浊音和清音的短时自相关函数有如下几个特点：

- 1) 短时自相关函数可以很明显的反映出浊音信号的周期性。
- 2) 清音的短时自相关函数没有周期性，也不具有明显突出的峰值，其性质类似于噪声。
- 3) 不同的窗对短时自相关函数结果有一定的影响。



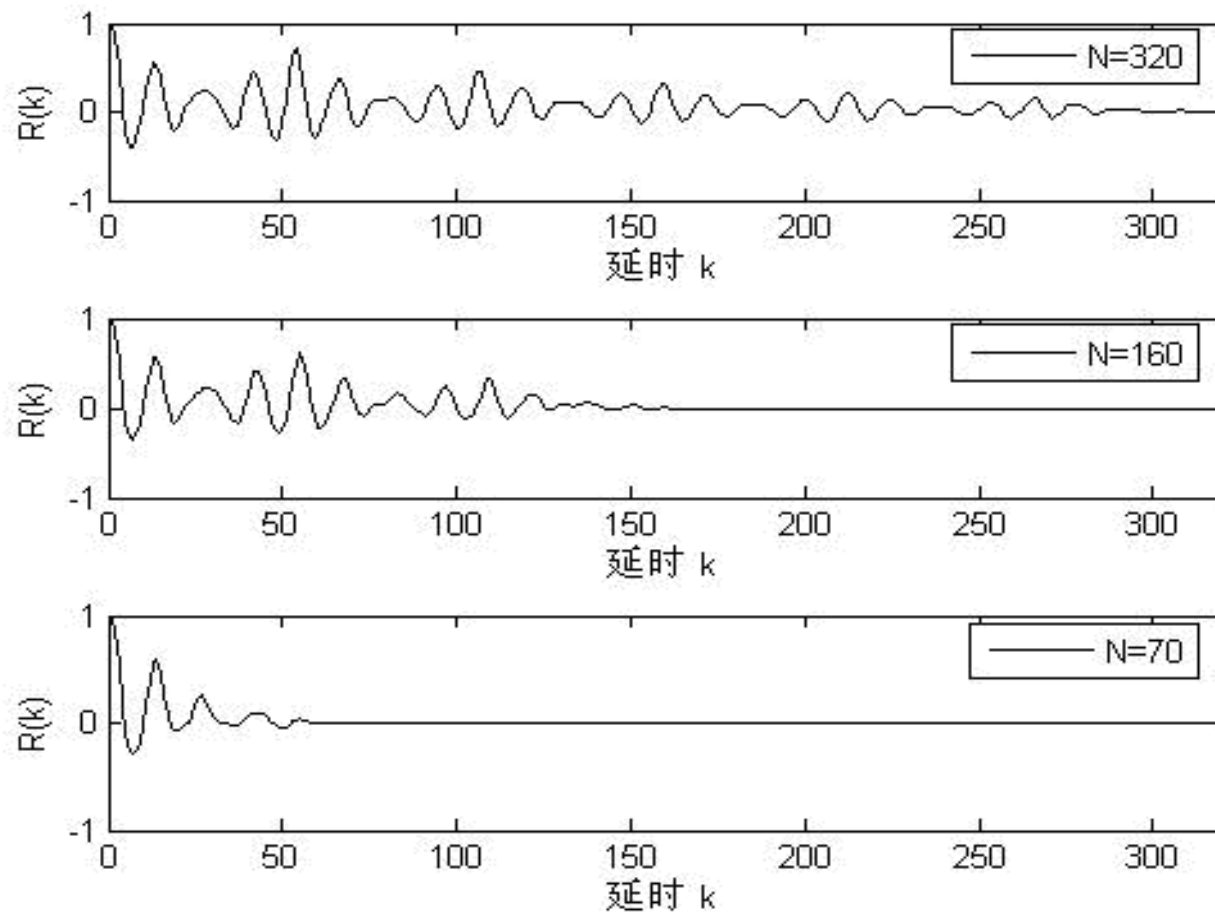


图3.16 不同矩形窗长时的短时自相关函数





3.6.3 修正的短时自相关函数

修正的短时自相关函数，其定义如下



$$\hat{R}_n(k) = \sum_{m=-\infty}^{+\infty} x(m)w_1(n-m)x(m+k)w_2(n-m-k)$$

若令 $m = n + m'$ ，代入上式中，可得

$$\hat{R}_n(k) = \sum_{m'=-\infty}^{+\infty} x(n+m')w_1(-m')x(n+m'+k)w_2(-m'-k)$$

定义 $\begin{cases} \hat{w}_1(m) = w_1(-m) \\ \hat{w}_2(m) = w_2(-m) \end{cases}$

$$\hat{R}_n(k) = \sum_{m=-\infty}^{+\infty} x(n+m)\hat{w}_1(m)x(n+m+k)\hat{w}_2(m+k)$$


$$\hat{w}_1(m) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{其它} \end{cases}$$

$$\hat{w}_2(m) = \begin{cases} 1, & 0 \leq n \leq N-1+K \\ 0, & \text{其它} \end{cases}$$

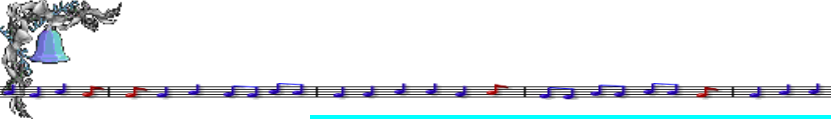
式中， K 为 k 的最大值，即 $0 \leq k \leq K$ 。

要使 $\hat{w}_2(m+k)$ 为非零值，必须使 $m+k \leq N-1+K$

考虑到 $k \leq K$ ，可得 $m \leq N-1$

修正的短时自相关函数可以写成：

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k)$$



$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k)$$

因为求和上限是 $N-1$ ，与 k 无关，故当 k 增加时， $\hat{R}_n(k)$ 值不下降。





3.6.4 短时平均幅度差函数

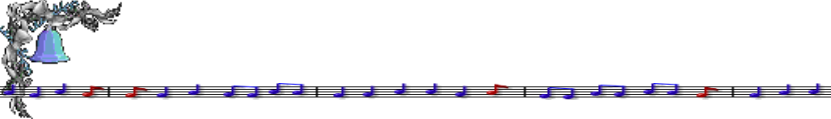
一个周期为 P 的周期信号，在 $k=0, \pm P, 2P, \dots$ 时，

$$d(n) = x(n) - x(n - k) = 0 \quad , \quad (k = 0, \pm P, \pm 2P, \dots)$$

对于浊音语音，在基音周期的整数倍上， $d(n)$ 总是很小，但不是零，因此，我们可以定义短时平均幅度差函数AMDF为

$$r_n(k) = \sum_{m=-\infty}^{+\infty} |x(n+m)w_1(m) - x(n+m-k)w_2(m-k)|$$





使用矩形窗时，短时平均幅度差函数可写成：

$$r_n(k) = \sum_{n=0}^{N-1} |x(n) - x(n+k)|, \quad k = 0, 1, \dots, N-1$$

$r_n(k)$ 与 $\hat{R}_n(k)$ 之间的关系为：

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





3.7 基于能量和过零率的语音端点检测

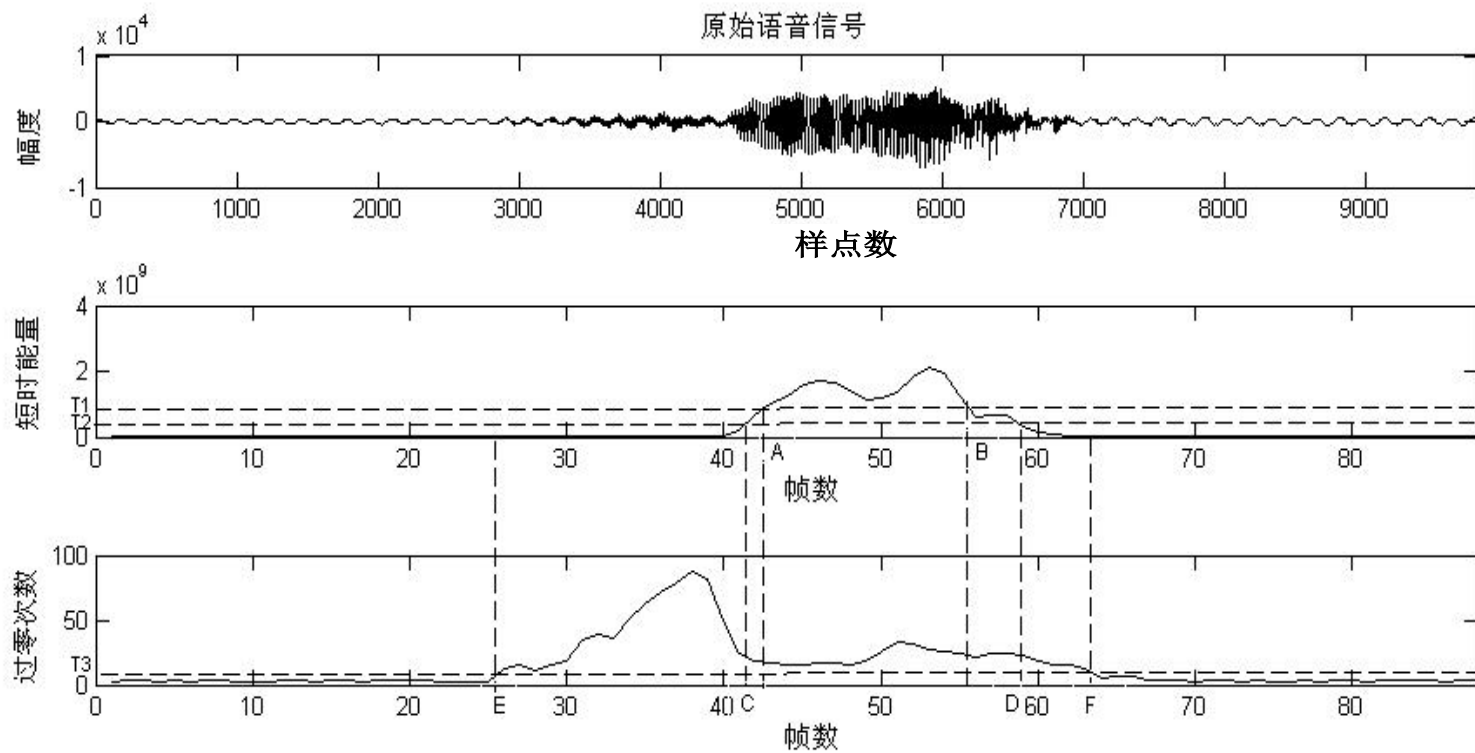
语音端点检测就是指从包含语音的一段信号中确定出语音的起始点和结束点。

正确的端点检测对于语音识别和语音编码系统都有重要的意义。

本节介绍基于能量和过零率的语音端点检测方法——两级判决法及程序实现。



两级判决法示意图





采用双门限比较法的两级判决法，具体如下

第一级判决：

1. 先根据语音短时能量的轮廓选取一个较高的门限 T_1 ，进行一次粗判：语音起止点位于该门限与短时能量包络交点所对应的时间间隔之外(即AB段之外)。

2. 根据背景噪声的平均能量确定一个较低的门限 T_2 ，并从A点往左、从B点往右搜索，分别找到短时能量包络与门限 T_2 相交的两个点C和D，于是CD段就是用双门限方法根据短时能量所判定的语音段。



第二级判决：

以短时平均过零率为标准，从C点往左和从D点往右搜索，找到短时平均过零率低于某个门限T3的两点E和F，这便是语音段的起止点。门限T3是由背景噪声的平均过零率所确定的。

注意：门限T2，T3都是由背景噪声特性确定的，因此，在进行起止点判决前，T1，T2，T3，三个门限值的确定还应当通过多次实验。





基于MATLAB程序实现能量与过零率的端点检测算法步骤如下：

- (1) 语音信号 $x(n)$ 进行分帧处理。
- (2) 得到语音的短时帧能量。
- (3) 计算每一帧语音的过零率，得到短时帧过零率。
- (4) 考察语音的平均能量设置一个较高的门限 T_1 ，用以确定语音开始，然后根据背景噪声的平均能量确定一个稍低的门限 T_2 ，用以确定第一级语音结束点。第二级判决同样根据背景噪声平均过零率 Z_N ，设置一个门限 T_3 ，判断语音前端清音和后端尾音。





3.8 基音周期估值

基音周期估值在语音信号处理应用中具有十分重要的作用。本节介绍语音信号基音周期估值最基本的两种方法：

基于短时自相关法的基音周期估值

基于短时平均幅度差函数法的基音周期估值





3.8.1 基于短时自相关法的基音周期估值

语音的浊音信号具有准周期性，其自相关函数在基音周期的整数倍处取最大值。计算两相邻最大峰值间的距离，就可以估计出基音周期。

为了突出反映基音周期的信息，同时压缩其他无关信息，减小运算量，自相关计算之前需要对语音信号进行适当预处理。





预处理的两种方法

第一种方法：先对语音信号进行低通滤波，再进行自相关计算。

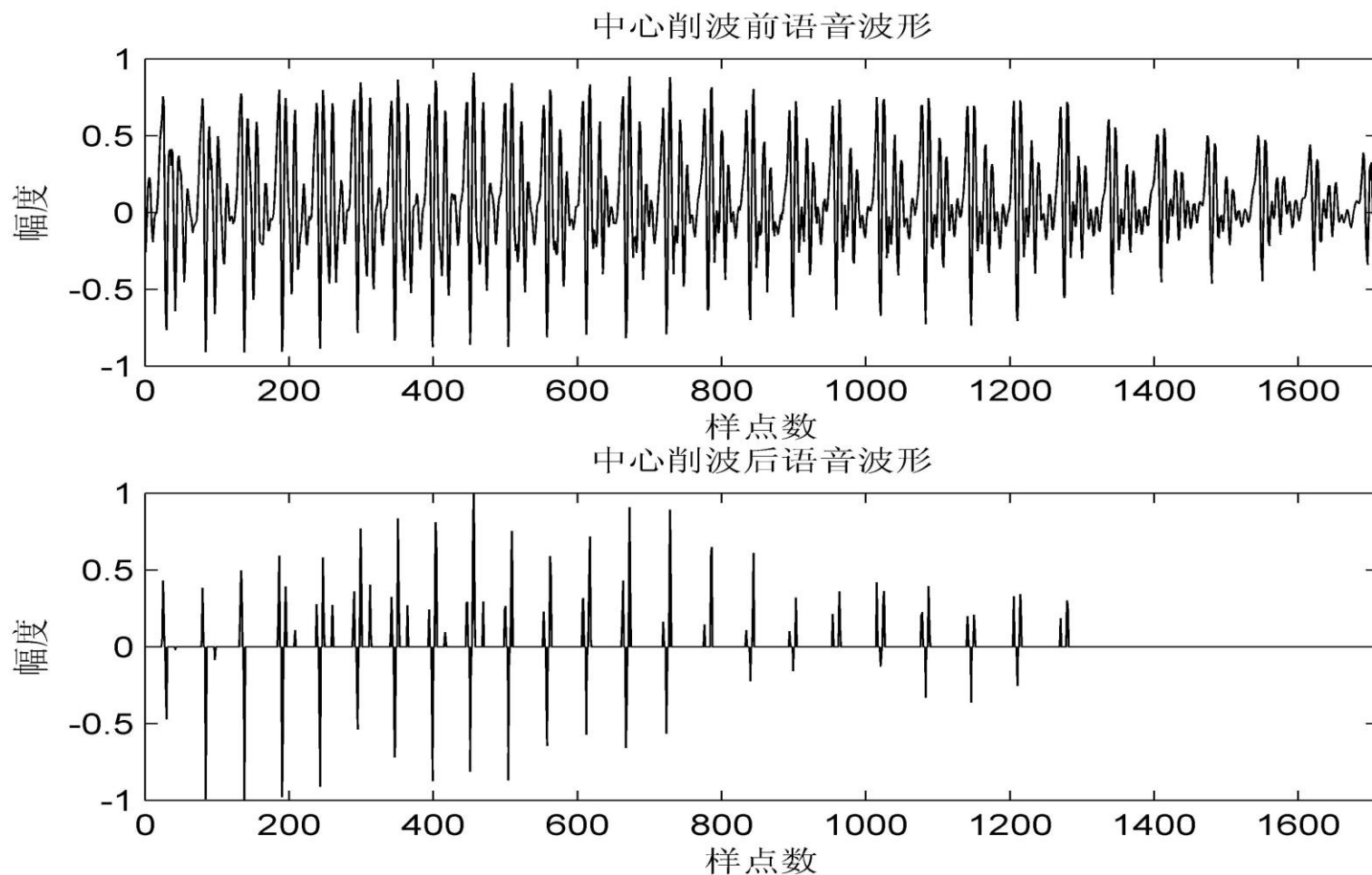
第二种方法：先对语音信号进行中心削波处理，再进行自相关计算。常用的有两种削波函数，下面分别介绍。

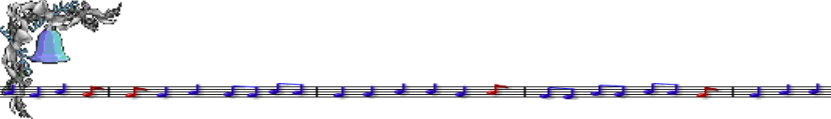
1.中心削波

中心削波函数为

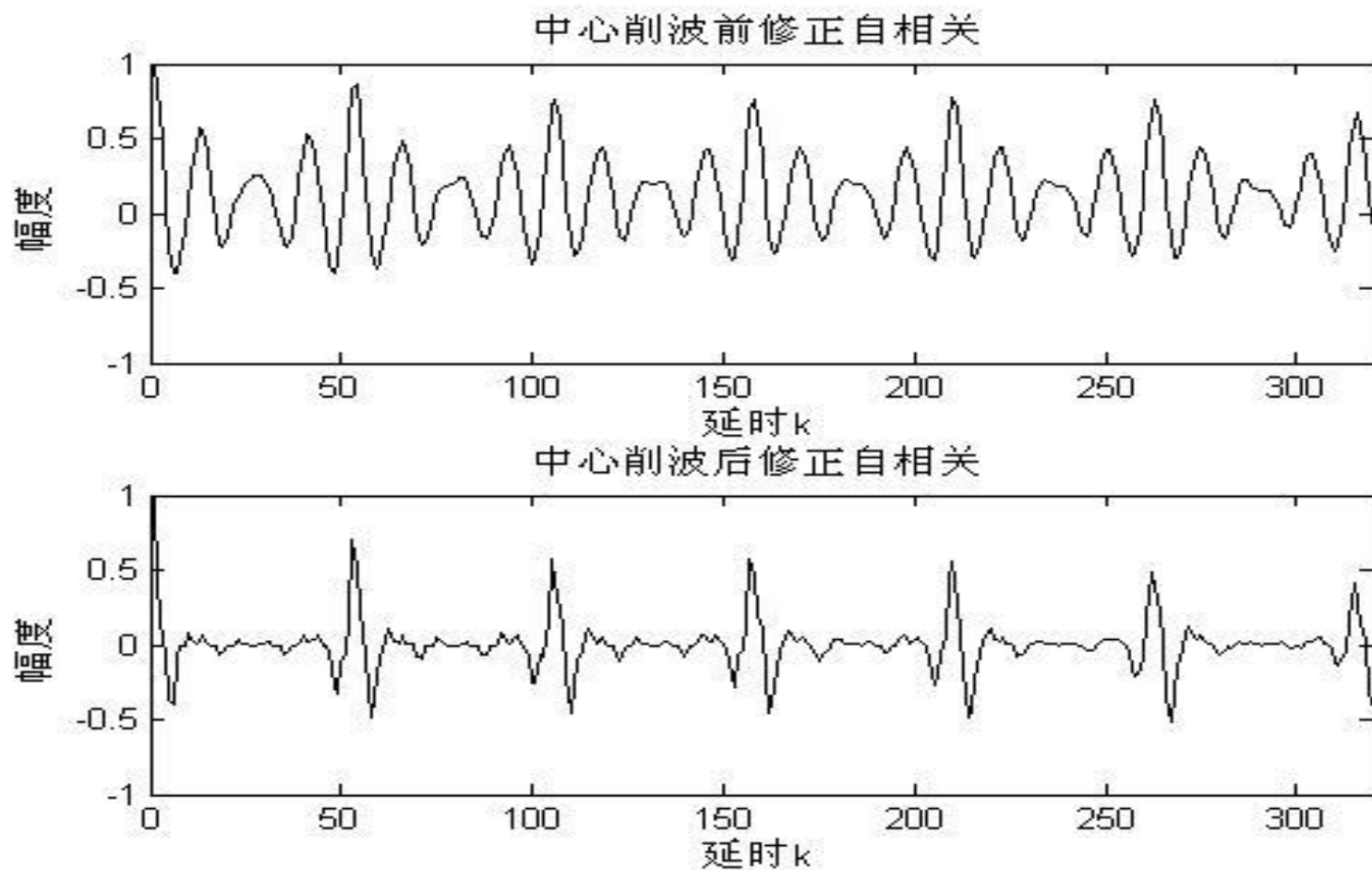
$$f(x) = \begin{cases} x - x_L & (x > x_L) \\ 0 & (-x_L \leq x \leq x_L) \\ x + x_L & (x < -x_L) \end{cases}$$

削波后的序列用短时自相关函数估计基音周期，在基音周期处峰值更加尖锐，可减少倍频或半频错误。





削波前后语音信号对比图及修正自相关对比图

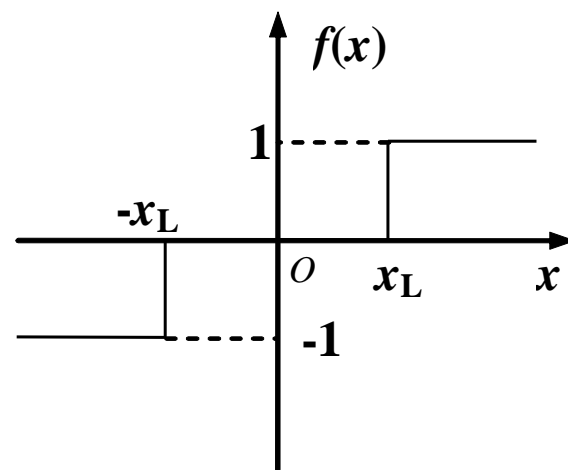




2. 三电平削波

为了克服短时自相关函数计算量大的问题，在中心削波法的基础上，还可以采用三电平削波法，削波函数如下式

$$f(x) = \begin{cases} 1 & x > x_L \\ 0 & -x_L \leq x \leq x_L \\ -1 & x < -x_L \end{cases}$$



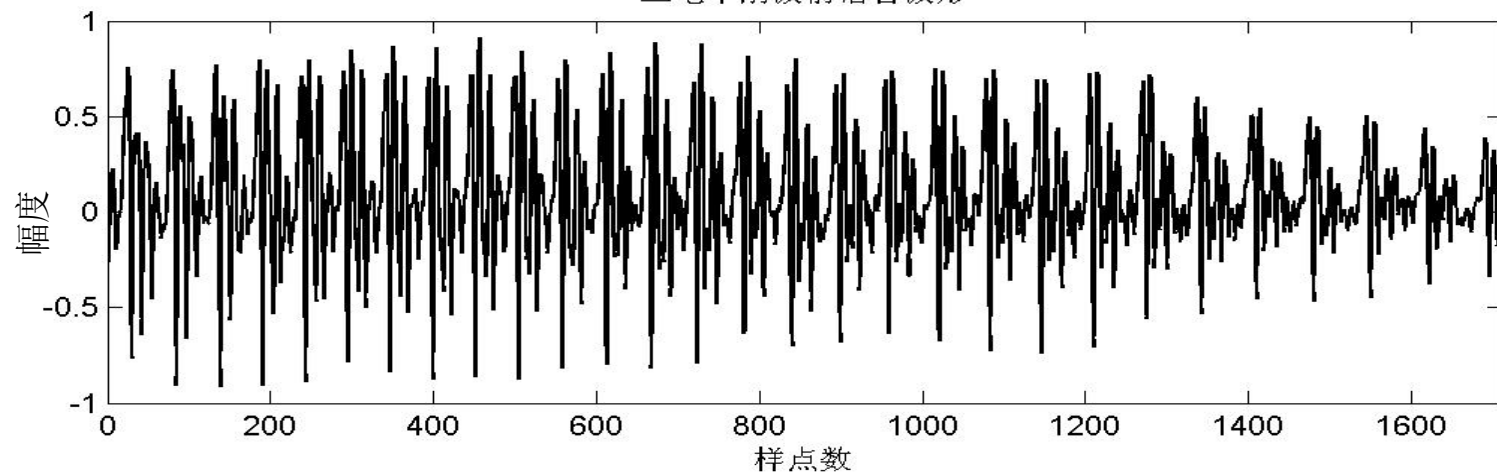


经削波后的取样值仅有三种可能情况，即+1，0，-1。显然，这种信号的短时自相关函数的计算实际上是不需要乘法运算的，这就大大节省了计算时间。

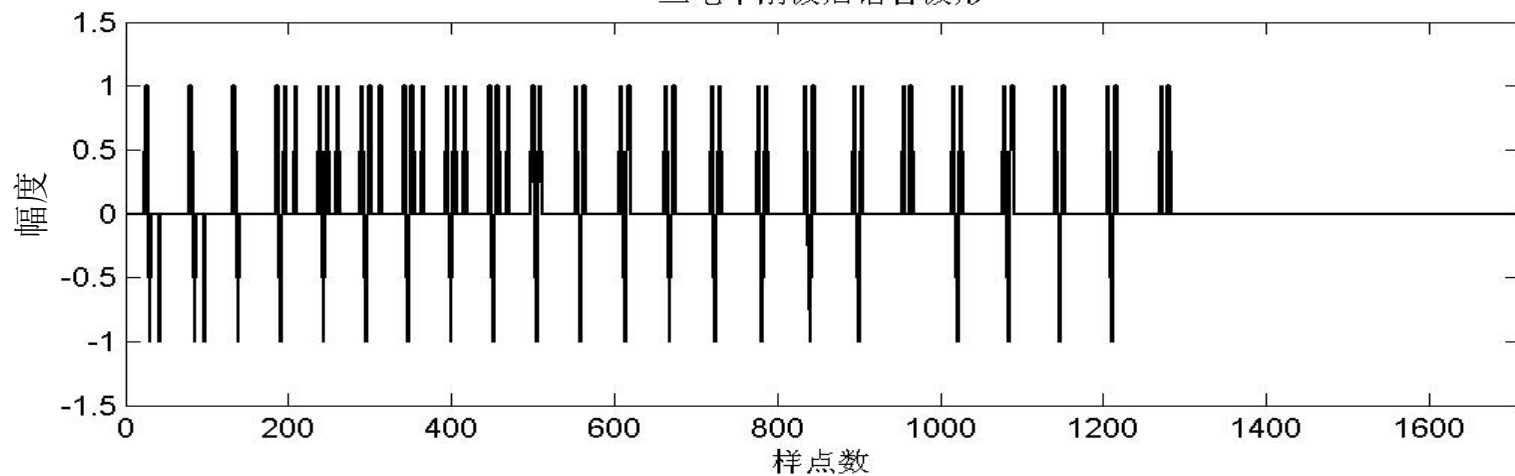


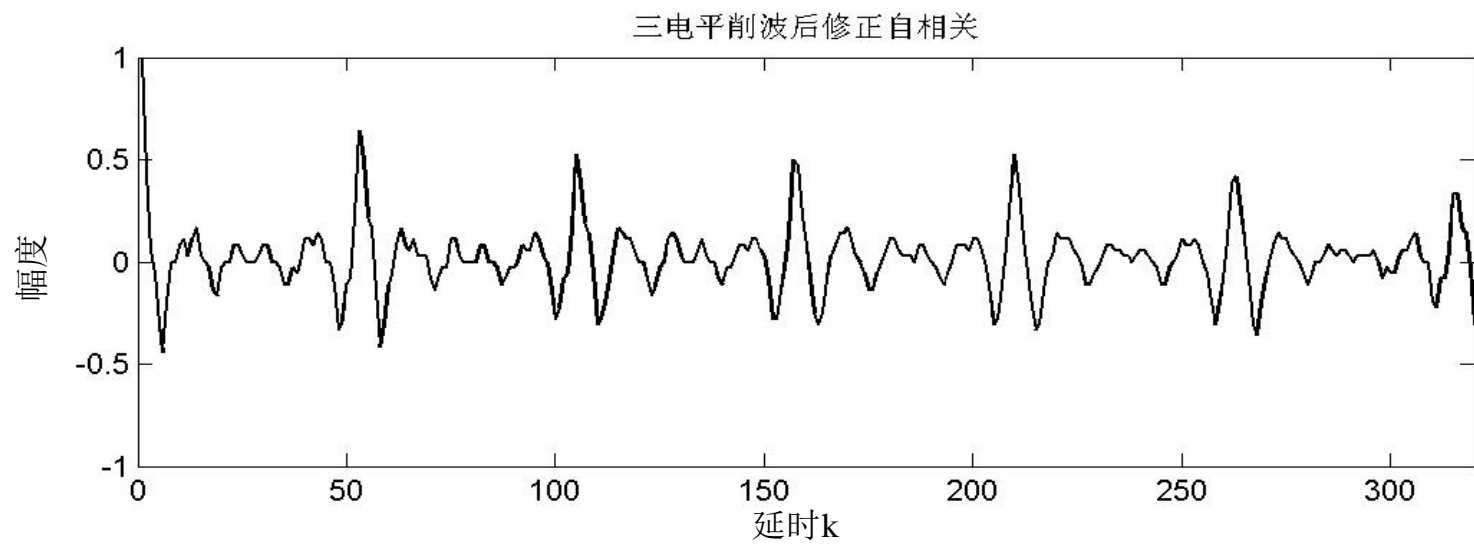
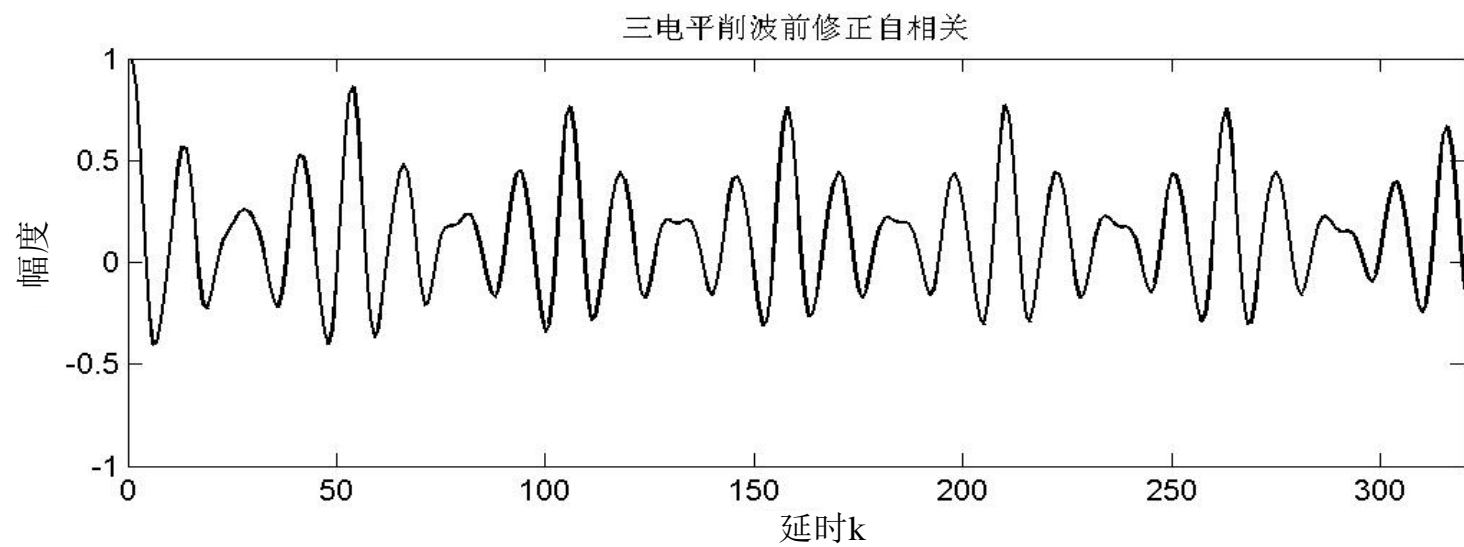


三电平削波前语音波形



三电平削波后语音波形







3.8.2 基于短时平均幅度差函数 AMDF法的基音周期估值

对于浊音语音，在基音周期的整数倍上的幅度差值不是零，但总是很小，因此，可以通过计算短时平均幅度差函数中两相邻谷值间的距离来进行基音周期估值。这里使用修正的短时平均幅度差函数并加矩形窗，得到：

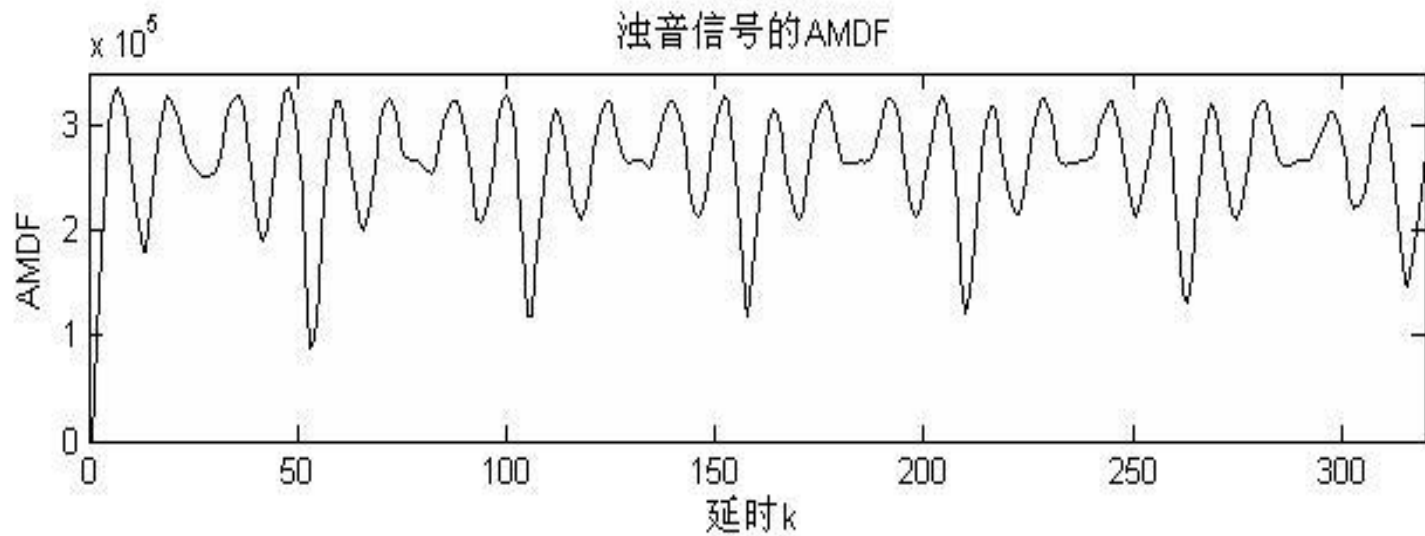
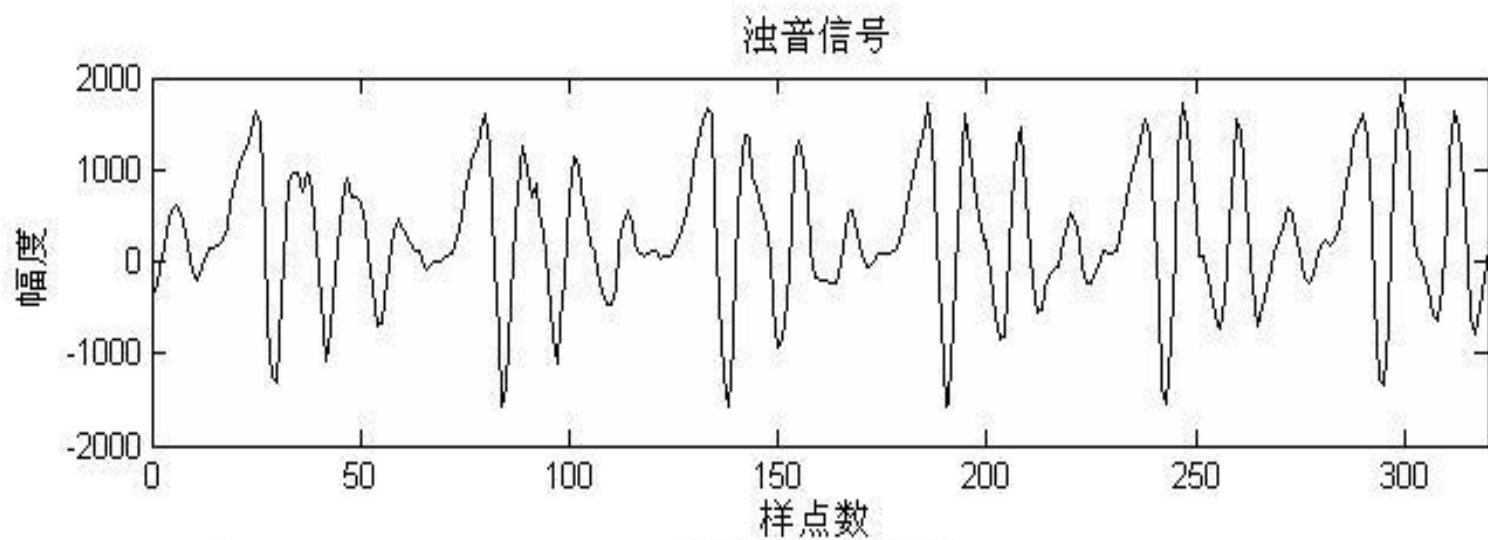
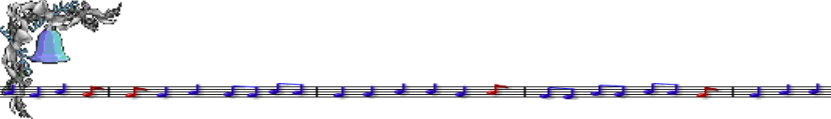
$$r_n(k) = \sum_{n=0}^{N-1} |x(n) - x(n+k)|, \quad k = 0, 1, \dots, N-1$$



AMDF函数与短时自相关函数的不同是：

自相关函数进行基音周期估计时寻找的是最大峰值点的位置，而AMDF寻找的是它的最小谷值点的位置。由于清音没有周期性，所以它的自相关函数和平均幅度差函数均不具有准周期性的峰值或谷值。







3.8.3 基音周期估值的后处理

在提取基音时，无论采用哪种方法提取的基音频率轨迹与真实的基音频率轨迹都不可能完全吻合。实际情况是大部分段落吻合，而在一些局部段落和区域中有一个或几个基音频率估计值偏离，甚至远离正常轨迹，通常是偏离到正常值的2倍或1/2处，即实际基音频率的倍频或分频处，称这种偏离点为基音轨迹的“野点”。

为了去除“野点”，常用的平滑技术主要有：中值滤波平滑处理、线性平滑、动态规划平滑处理。



1. 中值平滑处理

基本原理：设 $x(n)$ 为输入信号， $y(n)$ 为中值滤波器的输出，采用一滑动窗，则 n_0 处的输出值 $y(n_0)$ 就是将窗的中心移到 n_0 处时窗内输入样点的中值。即在 n_0 点的左右各取 L 个样点。连同被平滑点共同构成一组信号采样值（共 $(2L+1)$ 个样值），然后将这 $(2L+1)$ 个样值按大小次序排成一队，取此队列中的中间者作为平滑器的输出。 L 值一般取为1或2，即中值平滑的“窗口”一般包括3至5个样值，称为3点或5点中值平滑。



2. 线性平滑处理

线性平滑是用滑动窗进行线性滤波处理

$$y(n) = \sum_{m=-L}^L x(n-m)w(m)$$

$\{w(m), m = -L, -L+1, \dots, 0, 1, 2, \dots, L\}$ 为 $2L+1$ 点平滑窗，满足

$$\sum_{m=-L}^L w(m) = 1$$





3. 组合平滑处理

为了改善平滑的效果可以将两个中值平滑串接，图3.26(a)所示是将一个5点中值平滑和一个3点中值平滑串接。另一种方法是将中值平滑和线性平滑组合，如图3.26(b)所示。为了使平滑的基音轨迹更为贴近，还可以采用二次平滑的算法。

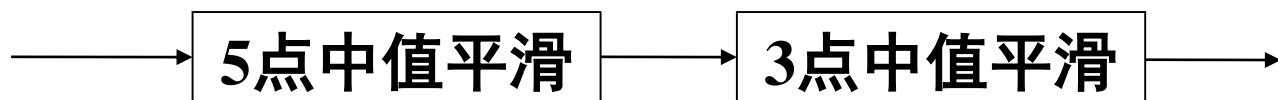
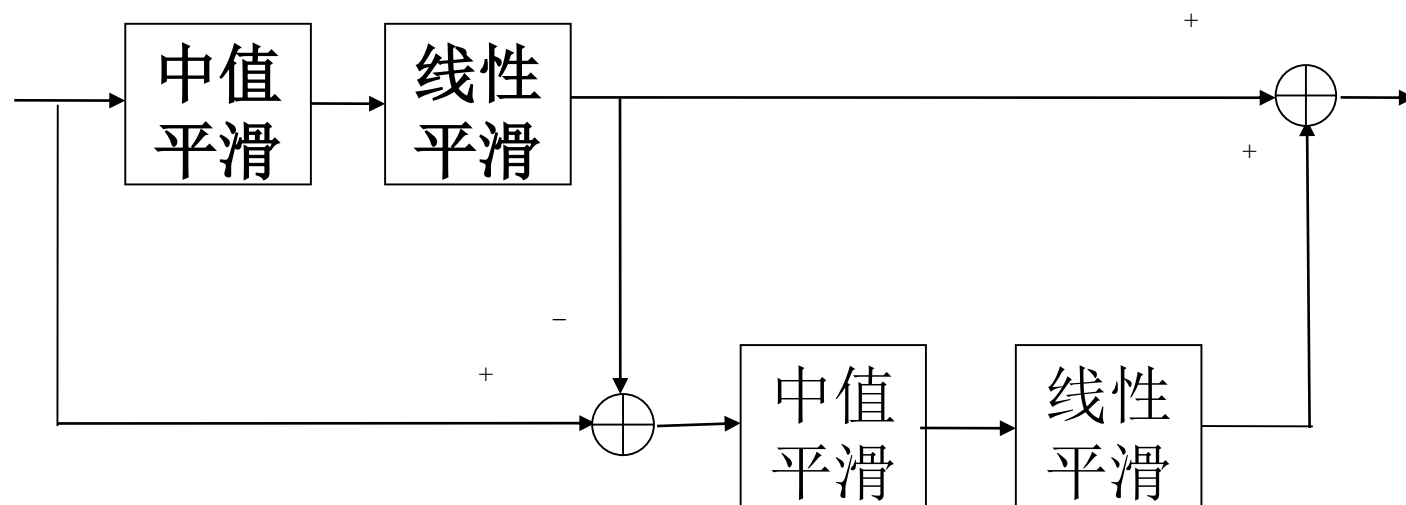
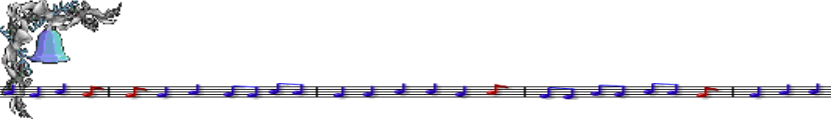


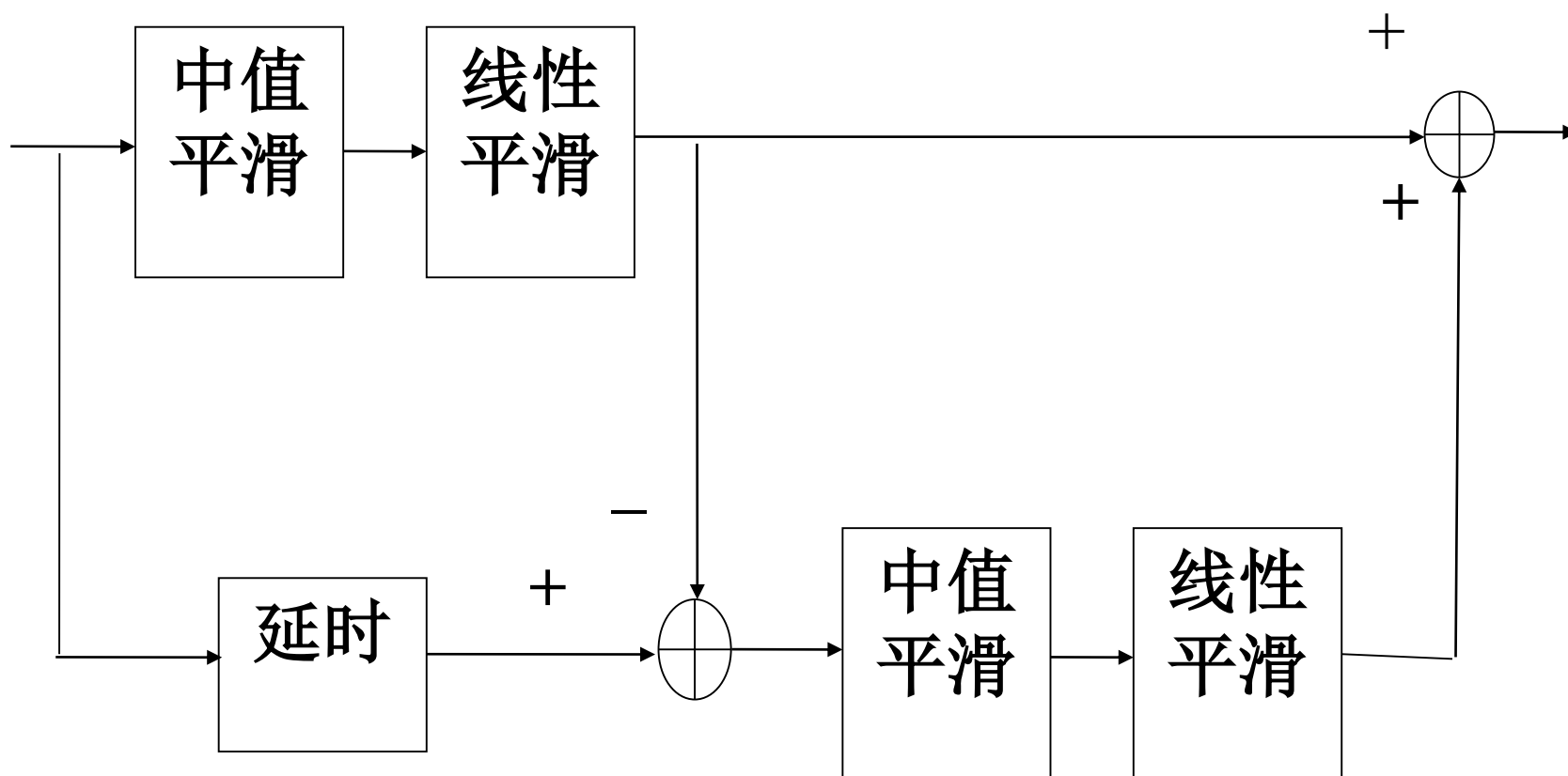
图3.26(a)



图3.26(b)







各种组合平滑算法

