第 3 章

短时时域处理技术 «

在不同场合下，人们对声音信号携带的信息和特性的需求不同。例如，为了判断一段声音是不是语音信号，只需要提取人类语音信号的特征就可以了；为了区分清音还是浊音，就需要提取其基音频率，并进行功率谱分析；在存储或传输时，需要保留较多的声音信息以保证质量。

因此，数字声音信号的分析算法在所有声音信号处理中都占据重要的地位。不论是构建有效的语音编码器，还是进行高效语音识别、合成，或者对声音进行分析、比较、变换，都需要深入了解声音信号的特性。

声音信号的处理方式有很多，常用的有短时时域处理技术、短时频域处理技术、线性预测技术、同态滤波技术以及其他方法。本章主要讨论短时时域处理技术。

3.1 语音信号的短时处理方法

传统的语音信号处理中，在对语音信号进行特征提取之前，必须消除因为自身发声器官或者由于采集语音信号的设备可能会带来的高次谐波失真、混叠、高频等现象。预处理操作的目的就是降低这些因素对语音信号产生的负面影响，尽可能保证经过预处理后的信号更平滑、干净，方便有效地提取并表示语音信号所携带的信息。所以，需要对信号进行端点检测、预加重、分帧和加窗等预处理操作，如图 3.1 所示。

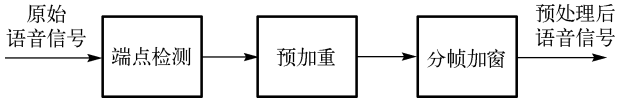


图 3.1　预处理流程图

3.1.1 语音端点检测

端点检测（Voice Activity Detection, VAD）也被称为语音活动检测，其主要目的是对一段音频区分语音部分与非语音部分，也可以理解为从携带噪声或者留白的语音中精确地定位语音的开始部分、结束部分，从而忽略噪声部分和静音部分，提取包含有效信息的语音端。

端点检测可分为三类：基于阈值的端点检测、基于分类器的端点检测、基于模型的端点检测。基于阈值的端点检测算法通过提取短时能力、短期过零率等时域特征或梅尔频率倒谱系数和谱熵等频域特征，再采用门限值作为判断的依据，对语音和非语音部分进行分离。基于分类器的端点检测算法将语音分为两类，利用机器学习的方法训练分类器，将训练完的模型对语音进行端点检测。基于声学模型的端点检测算法利用完整的声学模型，在解码的基础

上，通过全局信息对语音的开始端与结束端进行判别。

最常用的端点检测算法为基于短时能量的端点检测算法，该算法需要计算每一帧的短时能量，具体公式为

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \tag{3.1}$$

式中，$x(m)$ 为每一帧语音信号，$w(n)$ 为窗口函数，N 为帧长，E_n 为第 n 帧语音信号所有点的能量和。设定 E_0 为能量门限，将输入的语音信号的每一帧短时能量与 E_0 进行比较。当输入的能量均高于 E_0 时，认为该点为语音开始端点，并定义开始端点为连续帧的第一帧；相应地，当输入的能量低于 E_0 时，认为该点为语音结束端点，开始端点和结束端点之间的连续语音帧即认为是目标信号。

3.1.2 预加重

预加重是一种在发送端对输入信号高频分量进行补偿的信号处理方式。随着信号在人体传输过程中高频受到损失，为了在接收终端能得到比较好的信号波形，就需要对受损的语音信号进行补偿，预加重技术的思想就是增强信号的高频成分，以补偿高频分量在口腔、鼻腔传输过程中的衰减。而预加重对噪声并没有影响，因此能够有效地提高输出信噪比。

对语音信号进行预加重可提升语音的高频部分，提高语音的高频分辨率。预加重的传递函数为

$$H(z) = 1 - az^{-1} \tag{3.2}$$

其中，a 为预加重系数，设 n 时刻的语音时域信号采样值为 $x(n)$，经过预加重处理后，输出的语音时域信号 $y(n)$ 为

$$y(n) = x(n) - ax(n-1) \tag{3.3}$$

式中，根据经验设定，a 一般取 0.98。

3.1.3 分帧与加窗

语音信号通常是非平稳的，特别是低质量环境下的情感语音。因此，语音信号需要进行短时分析，即认为在短时间内该声音是平稳的，一般采取分帧与加窗处理。

分帧，即将语音片段进行分段处理，一般的语音信号以 10～30ms 为一帧进行划分，划分后假定认为每一帧短时平稳，考虑到帧与帧之间具有相关性，相邻帧之间会保留一部分重叠，从而上下帧之间平稳过渡，重叠部分称之为帧移，一般帧移为帧长的 1/4 至 2/3，图 3.2 所示为帧移与帧长比例为 1/4 的分帧示意图。

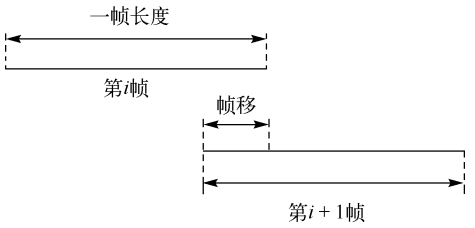


图 3.2　帧长与帧移示意图

分帧之后需要进行加窗处理，加窗的目的是让一帧信号的幅度在两端渐变到 0，而这种

渐变对傅里叶变换有好处，能够提高变换结果（即频谱）的分辨率。同时加窗能够使全局信息更加连续，避免出现吉布斯效应。

加窗的作用实际上是强调窗内的信号，削弱窗外信号。为了完全保留窗内信号的性质，理想的窗函数应尽可能相当于脉冲形式，用来提高其频率分辨率，并具有无旁瓣（即频率漏泄）的特性。实际上，这样的窗是不存在的，通常是对某种窗的折中选择。经常使用的窗函数有矩形、海宁、海明及布累克曼等，具体定义如下。

矩形（Rectangular）为

$$w(n)=\begin{cases}1, & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{3.4}$$

海宁（Hanning）为

$$w(n)=\begin{cases}0.5-0.5\cos\left(2\pi\dfrac{n}{N-1}\right), & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{3.5}$$

海明（Hamming）为

$$w(n)=\begin{cases}0.5-0.46\cos\left(2\pi\dfrac{n}{N-1}\right), & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{3.6}$$

布累克曼（Blackman）为

$$w(n)=\begin{cases}0.5-0.5\cos\left(2\pi\dfrac{n}{N-1}\right)+0.08\cos\left(2\pi\dfrac{n}{N-1}\right), & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{3.7}$$

这些窗在时域和频域中的形状不同。其中，矩形窗具有最窄主瓣、最高频率分辨率，同时也具有最大的频率漏泄。此外，在频域里，其基频倍数上的谐波幅度值较窄较尖，类似呈现更多的噪声。布累克曼窗有最低的频率分辨率和最小的频率泄漏，表现在频谱上比其他窗形更平滑。

表 3.1 描述了以上各种分析窗的特性。从以上分析及表 3.1 可以看出，海明的折中效果较好，在语音分析窗中的应用也最为广泛。

表 3.1　窗的特性

分析窗	矩形	海宁	海明	布累克曼
主瓣宽度（×π/*N*）	4	8	8	12
旁瓣漏泄（dB）	–13.3	–31.5	–42.7	–58.1

对于给定的窗函数，频率分辨率与窗长成反比。因为长分析窗可以得到信号在频域中的细节，如频谱中的谐波结构，声道谱包络中的幅度；而短分析窗可描述信号在时间上的精细结构，更接近假设的短时语音平稳性能，更好地表现谐波的变化和频谱包络，但却模糊了频谱的谐波，降低了谐波幅度。

具体来讲，窗的衰减实质上和窗长无关，增加长度只是减少了主瓣宽度。如果窗长很小，则短时能量会急剧变化；如果太长，则短时能量会在长时间里平均，不能恰当反映语音信号的性能变化。因此，窗长的选择是一个折中选择。为了能够精确表示信号的谐波结构，需要使窗里存在多于 1～7 个音调周期。但人的基音周期是变化的，而且男女老少的基音皆不同，一般选择窗的长度持续在 15～30ms。

3.2 短时能量和短时平均幅度

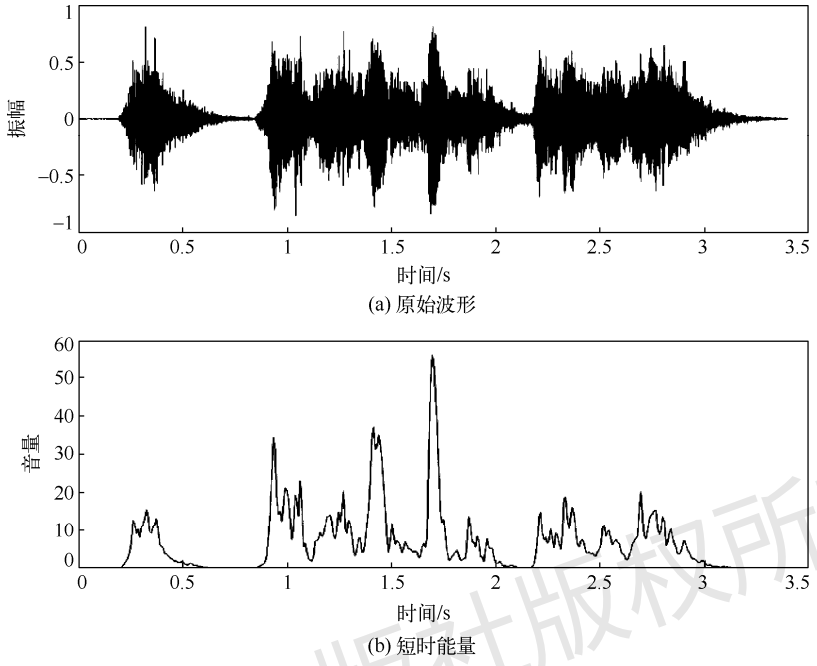
短时能量是声音信号中非常简单且常用的特征。每一帧的短时能量的具体公式见式（3.1），E_n 是第 n 帧语音信号所有点的能量和。

短时能量又称为音量，表示声音的强度、力度。一般而言，浊音的音量大于气音的音量，而气音的音量又大于噪声的音量。当然，它受到麦克风设定的影响，所以在计算前最好先减去声音信号的平均值，以避免信号的直流偏移（DC Bias）所导致的误差。

短时能量经常用在端点检测，估测有声之音母或韵母的开始位置及结束位置；也用于区分清浊音，当语音是浊音时，短时能量相对较高；当语音是清音时，短时能量较低；当语音是过渡音时，短时能量各不相同。因此，当语音段信噪比较高时，可以用短时能量进行语音分类。具体程序如下。

```
    waveFile = '/Users/multimodal/eNTERFACE wav/anger/an1.wav'; %声音文件
    frameSize = 256;                                          %帧长设置（点数）
    overlap = 128;                                            %帧间交叠（点数）
    % 读取语音文件，fs 为帧长
    [y, fs] = audioread（waveFile）;
    % 分帧
    frameMat = enframe（y, frameSize, overlap）;
    frameNum = size（frameMat, 2）;                           %帧数
    %计算音量
    volume = zeros（frameNum, 1）;
    for i = 1:frameNum
    frame = frameMat（:, i）;
    frame = frame - median（frame）;                          %减去均值
    volume（i） = sum（abs（frame .^ 2））;                    %式（3.1）的应用
    end
    % 画图
    sampleTime = （1:length（y））/ fs;                        %对应的时间
    frameTime = （（0:frameNum - 1） * （frameSize - overlap） + （frameSize/2））
/ fs; % 对齐时间轴
    subplot（2, 1, 1）;                                       %画子图
    plot（sampleTime, y）;                                    %x 轴与 y 轴的数据
    xlabel（{'时间/s';'（a）原始波形'}）;
    ylabel（'振幅'）;
    subplot（2, 1, 2）;
    plot（frameTime, volume）;
    xlabel（{'时间/s';'（b）短时能量'}）;
    ylabel（'音量'）;
```

运行结果如图 3.3 所示。



(a) 原始波形

(b) 短时能量

图 3.3　音频的原始波形和短时能量

一般情况下，我们使用短时能量来表示声音的强弱，但是前述计算音量的方法，只是用数学的公式来接近人耳的感觉，和人耳的感觉有时会有相当大的差异。为了进行区分，主观音量被用来表示人耳听到的音量大小。例如，人耳对于同样振幅但不同频率的声音，所产生的主观音量就会非常不一样。若把以人耳为测试主体的等主观音量曲线（Curves of Equal Loudness）画出来，就可以得到图 3.4，该图也代表人耳对于不同频率的声音的灵敏程度，也就是人耳的频率响应（Frequency Response）。

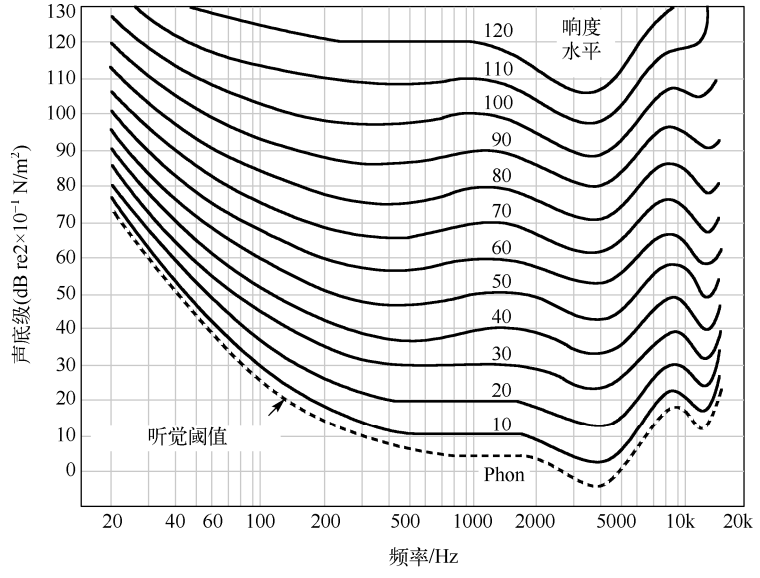


图 3.4　等主观音量曲线图

主观音量除了与频率有关外，也与声音的内容，如音色、基本周期的波形等有关，例如，可以尽量使用相同的主观音量来录下几个发音比较单纯的元音，再用音量公式来计算它们的音量，应该就可以看出音量公式和发音嘴型的关系了。具体程序如下。

```
waveFile='/Users/第 3 章/o-sound2.mp3';          %元音/o/音频文件的读取
au=myAudioRead（waveFile）;
opt=wave2volume（'defaultOpt'）;
opt.frameSize=512;                               %帧长
opt.overlap=0;                                   %帧间交叠为 0
time=（1:length（au.signal））/au.fs;             %每帧时间点

subplot（2,1,1）;                                %画子图
plot（time, au.signal）;                         %画原始波形
xlabel（{'时间 （s）';'（a）原始波形'}）;
ylabel（'幅度'）;
title（waveFile）;

subplot（2,1,2）;
opt1=opt;
opt1.frame2volumeOpt.method='absSum';            %运用子程序得到音量值
vollume1=wave2volume（au, opt1, 1）;
ylabel（opt1.frame2volumeOpt.method）;
xlabel（{'时间 （s）';'（b）音量'}）
ylabel（'音量'）
```

运行结果如图 3.5 所示，包括两个子图，分别对应“o”声音的原始波形和它的短时音量图。

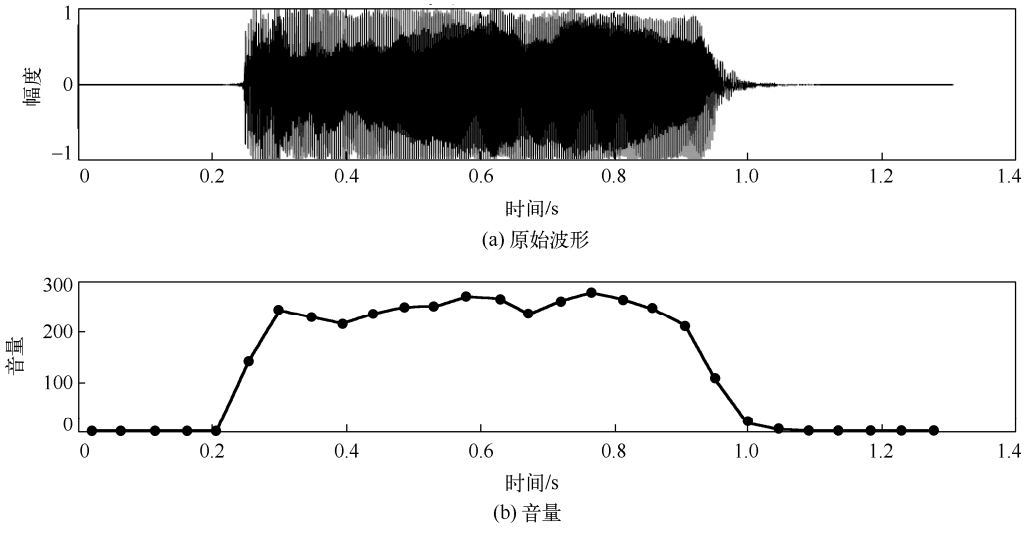


图 3.5　元音/o/的音量曲线图

主观音量容易受到频率和音色的影响，因此在进行语音或歌声合成时，常常根据声音的

频率和内容来对声音的振幅进行校正，以免造成主观音量忽大忽小的情况。

短时能量的一个主要问题是对信号电平值过于敏感。由于需要计算信号样值的平方和，在定点实现时很容易产生溢出。为了克服这个缺点，可以定义一个短时平均幅度函数来衡量语音幅度的变化，即

$$\mathrm{Amd}=\sum_{m=n-N+1}^{n}|x(n)|w(n-m) \tag{3.8}$$

与短时能量比较，短时平均幅度相当于用绝对值代替平方和，简化了运算，也能更好地表示清音的幅度变化。

3.3　短时过零率

过零分析是音频信号的一种时域分析方法，顾名思义，就是统计信号通过零值的次数。过零率特征粗略地描述了信号频谱特性，高的平均过零率意味着类似噪声的信号，低过零率意味着具有一定周期性的信号，这种高和低是相对的，没有准确的数值关系，所以经常和其他方法一起使用来进行判定。

短时过零率具体计算公式为式（3.9）和式（3.10），其中 $x(n)$为音频采样信号，$w(n)$为窗函数，sgn[ ]为符号函数，即

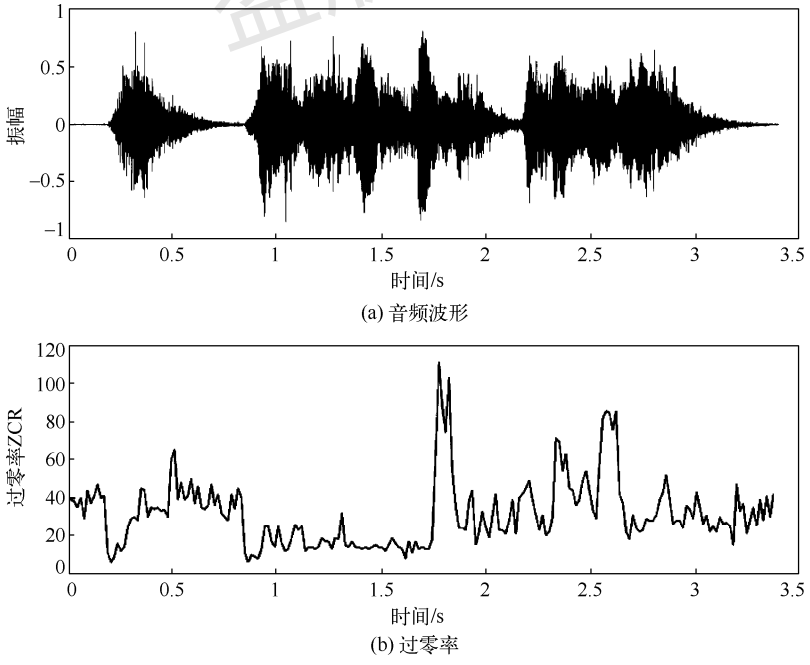
$$\begin{aligned}\mathrm{z}_n&=\sum_{m=-\infty}^{+\infty}|\mathrm{sgn}[x(n)]-\mathrm{sgn}[x(n-1)]|w(n)\\&=|\mathrm{sgn}[x(n)]-\mathrm{sgn}[x(n-1)]|\,w(n)\end{aligned} \tag{3.9}$$

$$\mathrm{sgn}[x(n)]=\begin{cases}1, & x(n)\geqslant 0\\-1, & x(n)<0\end{cases} \tag{3.10}$$

以说话声音频信号为例，具体程序如下，提取其短时过零率，运行结果如图 3.6 所示。

```
waveFile = '/Users/第 3 章/multimodal/eNTERFACE wav/anger/an1.wav';
frameSize = 256;
overlap = 0;
audio = myAudioRead（waveFile）;
y = audio.signal;
fs = audio.fs;

% 分帧
mat = enframe（y, frameSize, overlap）;
num = size（mat, 2）;

% 预分配过零率数组
zcr = zeros（1, num）;
```

```
for i = 1:num
    % 计算每一帧的过零率
    frame = mat(:, i);
    % 过零点是信号通过零点的地方
    zcr(i) = sum(abs(diff(sign(frame)))) / 2;    %过零率，对应式(3.9)
end

% 计算时间向量
sampleTime = (1:length(y)) / fs;
frameTime = (0:num-1) * (frameSize - overlap) / fs;

% 绘制波形
subplot(2, 1, 1);
plot(sampleTime, y);
xlabel({'时间 (s)';'(a) 音频波形'});
ylabel('振幅');

% 绘制过零率
subplot(2, 1, 2);
plot(frameTime, zcr);
xlabel({'时间 (s)';'(b) 过零率'});
ylabel('过零率 ZCR');
```

(a) 音频波形

(b) 过零率

图 3.6　音频的短时过零率

一般而言，在计算过零率时存在以下问题。

（1）由于有的信号恰好位于零点，因此过零率的计算就有两种，出现的效果也不同。因此，必须多加观察，才能选用最好的做法。

（2）大部分使用声音的原始整数值来进行计算，这样才不会因为在使用浮点数信号减去直流偏移（DC Bias）时，造成过零率的增加。

3.4　短时自相关函数

短时自相关函数用于衡量信号自身时间波形的相似性。在人类语音中，清音和浊音的发声机理不同，因而在波形上也存在着较大的差异。浊音的时间波形呈现出一定的周期性，波形之间相似性较好；清音的时间波形呈现出随机噪声的特性，杂乱无章，样点间的相似性较差。这样，可以用短时自相关函数来测定语音的相似特性。

平稳过程的自相关函数 $R(k)$ 具有以下性质。

（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，$|R(k)|\leqslant R(0)$。

（3）对于确定信号，$R(0)$ 对应能量；而对于随机信号，$R(0)$ 对应平均功率。

上述的第（2）个性质中，如果是一个周期为 P 的信号，则在取样处，其自相关函数也是最大值，因此可以根据自相关函数的最大值的位置来估计周期信号的周期值。

若一段语音是浊音信号，则其短时自相关函数也呈现周期现象，且其自相关函数的周期等于原语音信号的周期；若一段语音是清音信号，则其自相关函数不存在周期性。自相关函数可表示为

$$R(k)=\sum_{i=0}^{N-1-k}s(k)s(i+k),\qquad k=0,\cdots,N \tag{3.11}$$

其中，$s(k)$ 为输入的语音帧信号，N 为语音帧长度点数。

3.5　短时时域处理技术案例：基音提取

基音是基于发声器官如声门、声道和鼻腔的生理结构而提取的参数，能够很好地刻画说话人的声带特征，在很大程度上反映了人的个性特征。因此，基音检测算法在语音信号领域中具有广泛应用。

目前已经存在的很多基音检测算法是根据所在语音帧的清浊音分类结果进行检测的。在许多有噪声环境下的语音信号处理中，精确的基音提取尤其重要。例如，基于基音的语音合成器，要求标注每个基音的准确位置；文本到语音的合成器依靠精确基音来提高语音合成质量；在语音增强系统中，准确的基音提取直接关系到增强处理后的语音质量；在语音参数编码和混合编码中，基音常常作为编解码的重要参数和要传输的信息。

但是到目前为止，并没有一个简单、可靠的方法精确地检测基音，也没有客观的评价标准衡量基音的准确率。这是因为基音容易受到以下几个方面的影响而有所变化。

（1）声道滤波的影响使声门激励呈现出非完美的周期性。例如，放松的说话和用力说

话能使声门波平滑或猛烈地关闭，基音也会随之变化。即使说话人努力地想保持说话方式或者声道的形状，基音也会随机地抖动，连续声门波的幅度也会放大或者削弱而无法令基音周期保持不变。

（2）在清浊音语音类型变化处，由于语音的平稳性遭到破坏，基音特性变化速度快。

（3）基音范围比较大，为 50～400Hz，难以非常精确地检测基音。

（4）当清音、浊音同时存在时，基音难以准确检测。

（5）当有丰富的谐波信息存在时，基音难以准确被检测。

（6）由于环境噪声的存在，如人声喧哗处、汽车内或有其他声音的干扰，难以准确检测基音。

因此，基音检测方法成为语音信号处理中一项具有重要意义的研究课题。在不同的语音分析方法中，出现了不同的基音检测方法。目前已经存在多种时域、频域等若干基音检测方法。前者常用的基音提取方法主要是自相关方法和短时平均幅度差函数法等方法，其优点是简单、计算量小。但当语音信号频率或者幅度快速变化时，基音提取准确率会明显下降。在同态分析中，利用倒频谱中存在的峰值位置提供基音估计。当对不同尺度的滤波器组利用最大值相关方法时，子波变换可以导出基音的估计。在此基础上，还存在一些改进型检测方法。频域的基音检测方法主要是频谱的相似度方法，如谐波峰值检测、频谱相似度检测及正弦语音模型的基音确定方法等。

3.5.1 基音检测估计方法 1：三电平削波法

自相关函数会在语音信号基音周期的整数倍处取得最大值，凭借这一特性可以获得基音周期，但在实际操作过程中，在基音轨迹图上看到的第一个最高峰值点往往会在共振峰处，虽然在算法过程中应用了削噪、中心滤波、求取 LPC 残差等一系列的方法去减少共振峰的干扰，将共振峰干扰降到最低，但在一些语音帧处仍可能会出现一些处在基音周期整数倍以外的峰。鉴于此，为了保证基音周期处的峰值最高，对自相关提取基音的方法进行改进，即采用削波法提升其准确率。中心削波提取基音法如图 3.7 所示。

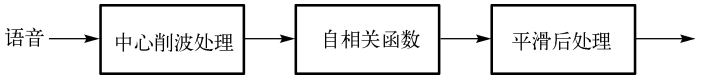


图 3.7　中心削波提取基音法

中心削波是降低短时自相关法所产生的倍频和半频错误过程中必不可少的一步，对于长度为 N 的一帧语音信号，对其进行加窗处理后记为 $s_w(n)$，其函数表达式为

$$y(n)=\begin{cases} s_w(n)-\mathrm{th0}, & s_w(n)>\mathrm{th0} \\ 0, & |s_w(n)|\leqslant \mathrm{th0} \\ s_w(n)+\mathrm{th0}, & s_w(n)<-\mathrm{th0} \end{cases} \tag{3.12}$$

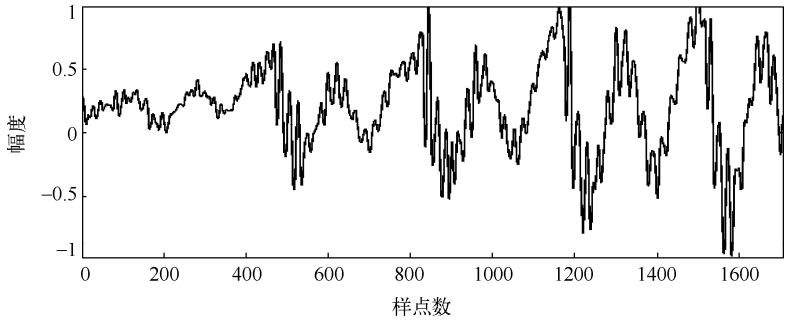
式中，$n=1,2,\cdots,N$，表示采样的点数，th0 是中心削波所用的电平阈值，一般取这帧语音幅度最大值的 50%～60%。语音信号的低幅值部分有着很多的共振峰信息，它的高幅值部分一般包含基音信息，因此，经过中心削波后，将更便于正确地提取语音的基音周期。

以下是进行削波处理后的参考程序代码。运行该代码，获得图 3.8 和图 3.9。

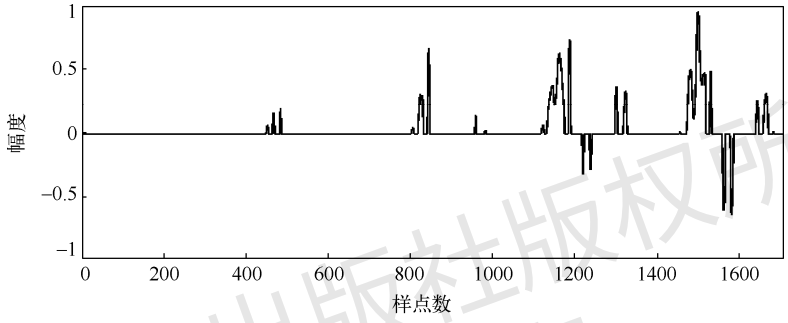
```
%读入数据，采样 fs=8kHz，采样位数为 16bit，长度为 320 样点
fid=fopen（'voice.txt','rt'）;                    %打开语音文件
[a,count]=fscanf（fid,'%f',[1,inf]）;             %读语音文件
L=length（a）;                                    %测定语音的长度
m=max（a）;
for i=1:L
a（i）=a（i）/m;                                  %数据归一化
end
%找到归一化以后数据的最大值和最小值
m=max（a）;                                       %找到最大的正值
n=min（a）;                                       %找到最小的负值
%为保证幅值与横坐标轴对称，采用计算公式 n+(m-n)/2，合并为(m+n)/2
ht=（m+n）/2;
for i=1:L;                                        %数据中心下移，保持和横坐标轴对称
a（i）=a（i）-ht;
end
figure（1）;                                      %画第一幅图
subplot（2,1,1）;                                 %第一个子图
plot（a,'r'）;                                    %r 表示画图采用红色
axis（[0,1711,-1,1]）;                            %确定横纵坐标的范围
title（'（a） 三电平削波前语音波形'）;             %图标题
xlabel（'样点数'）;                               %横坐标
ylabel（'幅度'）;                                 %纵坐标
coeff=0.5;                                        %中心削波函数系数取 0.5
th0=max（a）*coeff;                               %求三电平削波函数门限（threshold）
for k=1:L ;                                       %三电平削波
 if a（k）>=th0
 a（k）=a（k）-th0;
elseif a（k）<=（-th0）;
 a（k）=a（k）+th0;
 else
 a（k）=0;
 end
end
m=max（a）;
for i=1:L;                                        %三电平削波函数幅度的归一化
a（i）=a（i）/m;
end
subplot（2,1,2）;                                 %第二个子图
plot（a,'r'）;
```

```
axis（[0,1711,-1,1]）;                          %确定横纵坐标的范围
title（'（b） 三电平削波后语音波形'）;          %图标题
xlabel（'样点数'）;                             %横坐标
ylabel（'幅度'）;                               %纵坐标
fclose（fid）;                                  %关闭文件
%没有经过三电平削波的修正自相关计算
fid=fopen（'voice.txt','rt'）;
[b,count]=fscanf（fid,'%f',[1,inf]）;
fclose（fid）;
N=320;                                          %选择的窗长
ax=xcorr（a,320）;                              %削波前信号求自相关函数
a1x=xcorr（a1,320）;                            %削波后信号求自相关函数
figure（2）;                                    %画第二幅图
subplot（2,1,1）;                               %第一个子图
plot（ax,'k'）;
title（'（a） 三电平削波前修正自相关'）;        %图标题
xlabel（'延时 k'）;                             %横坐标
ylabel（'幅度'）;                               %纵坐标
%axis（[0,320,-1,1]）;
%三电平削波函数和修正的自相关方法结合

subplot（2,1,2）;                               %第二个子图
plot（a1x,'k'）;
title（'（b） 三电平削波后修正自相关'）;        %图标题
xlabel（'延时 k'）;                             %横坐标
ylabel（'幅度'）;                               %纵坐标
%axis（[0,320,-1,1]）;axis（[0,320,-1,1]）;
%三电平削波函数和修正的自相关方法结合
N=320;                                          %选择的窗长
A=[];
for k=1:320;                                    %选择延迟长度
sum=0;
for m=1:N;
sum=sum+a（m）*a（m+k-1）;                      %对削波后的函数计算自相关
end
```

对如图 3.8(a)所示的带噪语音信号进行中心削波处理，得到图 3.8(b)，此处削波电平为语音信号的 50%，可以看到只留下峰值较大的信号。再将原始信号与削波后信号进行自相关计算，分别得到图 3.9(a)和图 3.9(b)，经过比对可以很明显地看出中心削波后其自相关图中基音周期位置的峰值变得更加尖锐，也可以大大减少基音提取的半频错误。此外，也可以改变阈值参数比较其结果。

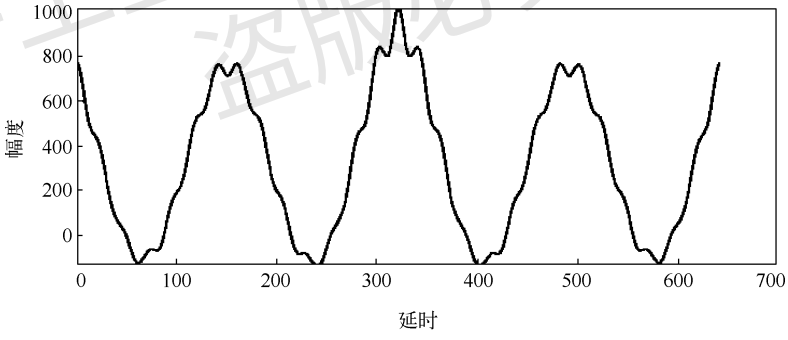


(a) 三电平削波前语音波形

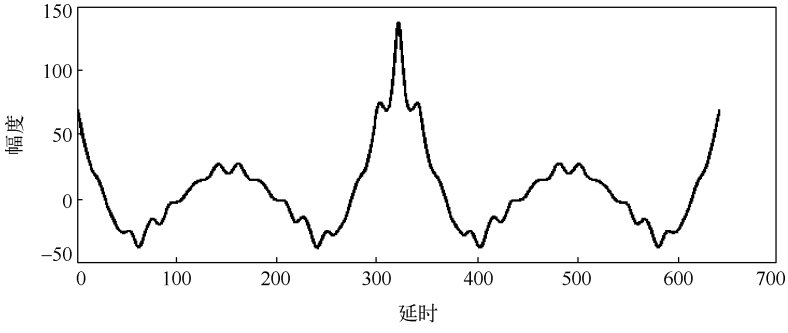


(b) 三电平削波后语音波形

图 3.8　三电平削波前后的语音波形



(a) 三电平削波前修正自相关



(b) 三电平削波后修正自相关

图 3.9　中心削波前后的自相关波形

3.5.2 基音检测估计方法 2：SHR 谐波检测法

由于发声时产生的抖动或声门处脉冲幅度不同等因素的影响，在语音中表现为倍频或者半频处也存在周期性信号。尤其当声壁在几何或者机械上不对称时，容易产生子谐波，表现在语音有连续声门激励时，在语音波形准周期的半频处出现与周期幅度成一定比率的周期成分，称为幅度交变；或出现与频率交错出现的成分，称为频率交变。因此，提取基音时经常取到基音的倍频或者半频，引起基音检测的误差和基音提取的误判。故研究高精度基音检测方法仍然是语音分析中有实际意义的难题之一。

子谐波和谐波与声音组成成分非常相似。目前这种方法已经广泛应用在语音及音频分析与合成算法中。基于子谐波和谐波的基音提取方法自提出以来有一定进展。在此基础上得到的子谐波-谐波比率（Subharmonic-harmonic Ratio，SHR）基音提取方法已经被验证具有一定的鲁棒性。但在基音周期变换较快时由于周期个数增多会引起判断误差。本节进一步分析了语音特性，结合 SHR 算法及语音频率变换的调制率，用频率自相关方法提取相应峰值的周期频率共同判定基音，再用正弦语音模型的最小均方误差对判别结果进行精选，从而得到准确的基音。

根据 SHR 基音提取理论，在浊音周期信号中除了存在基音脉冲周期，还存在幅度交变和频率交变的周期信号。通常将它们设为子谐波周期。这两种与基音周期交替发生的子谐波周期信号相当于对基音周期语音进行了幅度调制和频率调制，其调制率可分别表示为

$$M_{AM}=\frac{A_i-A_{i+1}}{A_i+A_{i+1}},\quad F_{FM}=\frac{F_i-F_{i+1}}{F_i+F_{i+1}} \tag{3.13}$$

其中，i 和 $i+1$ 表示波形上的点，A 表示幅度，F 表示频率，M_{AM} 和 F_{FM} 分别表示 i 点与 $i+1$ 点的幅度调制率与频率调制率。根据调制率可以计算语音子谐波和谐波之间的比率。设基音频率为 f_0，对数频谱的幅值为 $\mathrm{LOGA}(\log f)$，谐波出现于 $\log f_0$、$\log(2f_0)$、$\log(3f_0)$、…，则所有谐波幅值之和，即基音的 n 倍频率处的幅值之和为

$$\mathrm{SH}=\sum_{n=1}^{N}\mathrm{LOGA}(\log n+\log f_0) \tag{3.14}$$

其中，N 为谐波数目总数。如果认为子谐波频率在谐波一半频率处，则子谐波的幅值之和为

$$\mathrm{SS}=\sum_{n=1}^{N}\mathrm{LOGA}[\log(n-0.5)+\log f_0] \tag{3.15}$$

在对数频率轴上移动频谱，分别得到偶数点及奇数点幅值，即

$$\mathrm{SUMA}(\log f)_{\mathrm{even}}=\sum_{n=1}^{N}\mathrm{LOGA}\left[\log(2n)+\log f\right] \tag{3.16}$$

$$\mathrm{SUMA}(\log f)_{\mathrm{odd}}=\sum_{n=1}^{N}\mathrm{LOGA}\ [\log(2n-1)+\log f\,] \tag{3.17}$$

将式（3.16）和（3.17）分别代入式（3.14）和式（3.15），则 SH 与 SS 又相当于

$$\mathrm{SH}=\mathrm{SUMA}\left[\log(0.5f)\right]_{\mathrm{even}} \tag{3.18}$$

$$\mathrm{SS}=\mathrm{SUMA}\left[\log(0.5f)\right]_{\mathrm{odd}} \tag{3.19}$$

则设式（3.16）和（3.17）之差为差分函数，定义 DA 函数为

$$\mathrm{DA}(\log f)=\mathrm{SUMA}(\log f)_{\mathrm{even}}-\mathrm{SUMA}(\log f)_{\mathrm{odd}} \tag{3.20}$$

当频率取 50～450Hz 时，得到 DA 频谱，并可以得到

$$\mathrm{DA}\left[\log(0.5f_0)\right]=\mathrm{SH}-\mathrm{SS},\quad \mathrm{DA}\left[\log(0.25f_0)\right]=\mathrm{SH}+\mathrm{SS} \tag{3.21}$$

当不存在谐波时，认为最大点出现在 DA 频谱的 $0.5f_0$ 处；当存在子谐波时，认为最大值出现在 DA 频谱的 $0.5f_0$ 或者 $0.25f_0$ 处。根据式（3.13），定义子谐波−谐波比率（SHR）为

$$\mathrm{SHR}\approx 0.5\frac{\mathrm{DA}\left[\log(0.25f_0)\right]-\mathrm{DA}\left[\log(0.5f_0)\right]}{\mathrm{DA}\left[\log(0.25f_0)\right]+\mathrm{DA}\left[\log(0.5f_0)\right]}<\frac{\mathrm{SS}}{\mathrm{SH}} \tag{3.22}$$

故取得 DA 频谱中的最大值和第二最大值，由式（3.22）可以得到基音值。但当语音基音频率较大时，在 DA 频谱值会出现另一个周期的最大值，可能误选为此周期最大值，造成基音计算错误。而且，当子谐波的频率出现在基音频率一半处并且其幅值等于谐波幅值时，则有

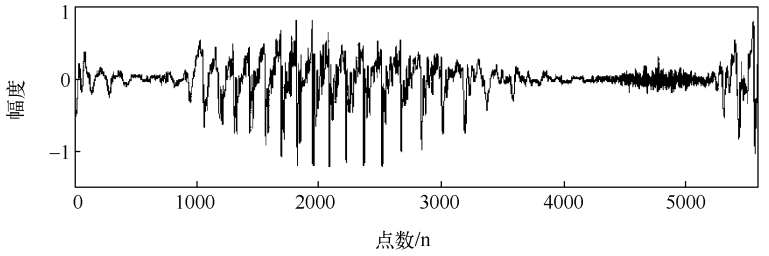
$$\mathrm{SS}=\sum_{n=1}^{N}\mathrm{LOGA}\left[\log\left(\frac{1}{2}n\right)+\log f_0\right],\quad \mathrm{SH}=\mathrm{SS} \tag{3.23}$$

此时有

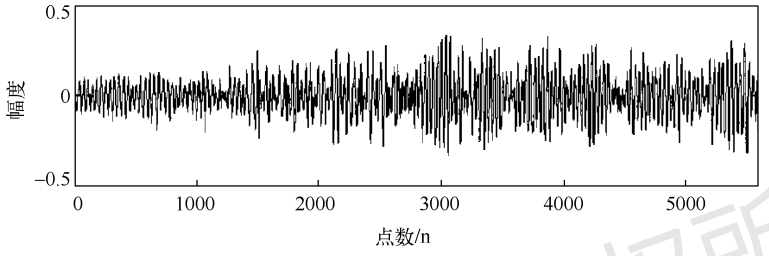
$$\mathrm{DA}\left[\log\left(\frac{1}{8}f_0\right)\right]=\mathrm{DA}\left[\log\left(\frac{1}{8}f_0\right)\right] \tag{3.24}$$

即 $\frac{1}{8}f_0$ 处也会出现最大值。因此，需要对频率再进行谐波比率检测，以得到准确的周期。

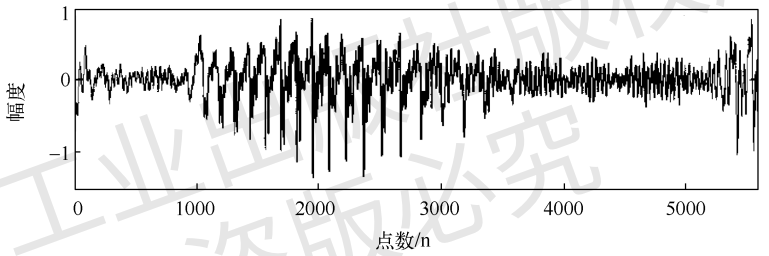
扫描右侧二维码查看程序。从实验结果可以得到、实际带噪声语音基音检测的结果。由于语音的真实基音难以确定，因此用一个客观的误差测度准则来衡量某个基音检测算法的优劣受到了限制。本节先截取一段语音作为被检测的原始语音，其本身带有一定的随机噪声，经过 50～7000Hz 带通滤波，采样频率是 16000Hz，帧速率是 40Hz，基音检测范围是 50～400Hz。先对原始语音提取基音，然后对原始语音加上餐厅录制的背景噪声，再进行基音检测。原始语音波形如图 3.10(a)所示。在无噪声语音基音提取中，SHR 算法和 ISHR 算法提取的基音基本相同，而由自相关方法提取的基音与另两种算法提取的基音有一定误差。当信噪比在 8dB 时，由图 3.10(d)可以看出，利用 ISHR 算法提取的基音基本保持不变；而用自相关方法提取的基音由于噪声的影响，偏离的误差比较大，不能准确取得基音。当信噪比为 6.5dB 时，利用 SHR 算法提取的基音与无噪声时提取的基音都有一定跳变，取得的基音有所偏离。因此，利用自相关方法提取基音的抗干扰性和准确率都低于 SHR 算法。



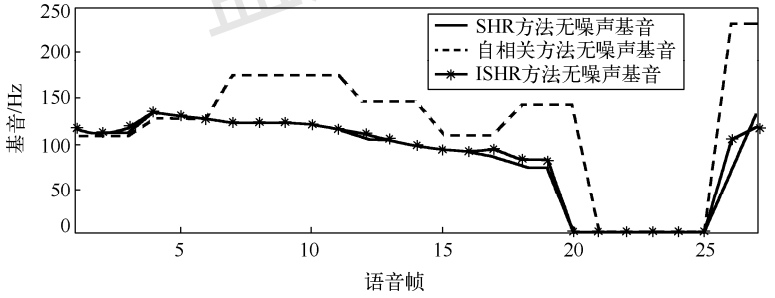
(a) 原始语音波形



(b) 噪声波形



(c) 原始语音加噪声波形



(d) 无噪声时，利用SHR、自相关方法提取语音信号的基音

图 3.10　真实语音波形与带噪声语音时域波形

练　习　题

1．为什么要对语音信号进行短时处理？是怎么实现的？

2．假设一段语音的采样频率为 8000Hz，请实现语音信号的分帧和加窗，要求帧长为 20ms，帧移为 10ms，窗型为海明窗；并画出其中一帧语音加窗前后的波形。

3．基音表示声音信号的什么特征？它由什么因素决定？对于男声、女声、小孩的声音，基音有什么特性？

4．可以用什么特征来区分声音和噪声？在一段语音信号中，可以用什么方法判断语音的起点和终点？

5．怎么提取基音？请用一种方法实现基音的提取，并比较利用该算法在纯净语音、10dB 信噪比、5dB 信噪比、3dB 信噪比和 0dB 信噪比的情况下的结果。

6．为什么中心削波处理的基音提取方法比一般的自相关方法提取的基音更准确？