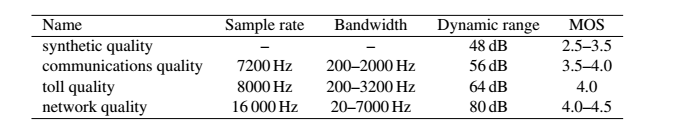
**Applied Speech and Audio Processing 第五章读书报告**

1. **语音通信**

本章主要是编码和压缩。

首先是量化，然后是通过参数化获得滤波器模型，然后用码激励线性预测(Code Excited Linear Prediction)做压缩，过程中会加入源滤波器模型、特征参数。

**语音音质评价的分级**

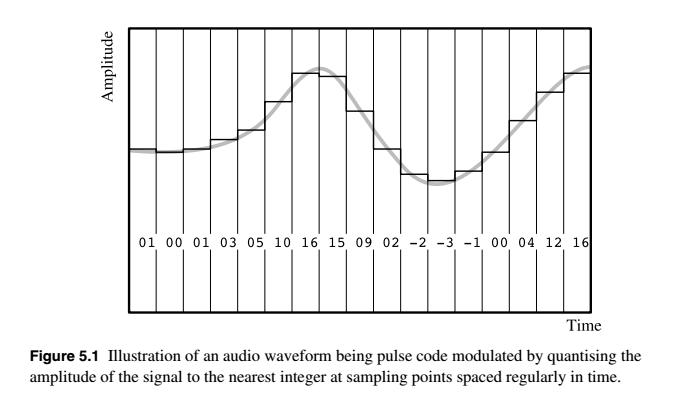
（顺序为合成-通信-电话-广播）

第三个为电话音质，是基于电话模拟通信发展起来的，但现在通常使用数字系统。

5.1 量化

本质上降低了需要传输的信息量。真实元件在带宽、存储量上受限。而语音通信系统的任务就是（其他通信系统也是）降低数据量、提高保真度。

5.1.1脉冲编码调制（Pusle Coded Modulation）



在计算机上对声音编码的一种方式：声音经过特定采样频率、比特数的模数转换（量化）后，变成了一组采样值。

采样方式有均匀非均匀两种，非均匀采样有一些理论发展，但实际应用以均匀采样为主。

5.1.2 △调制

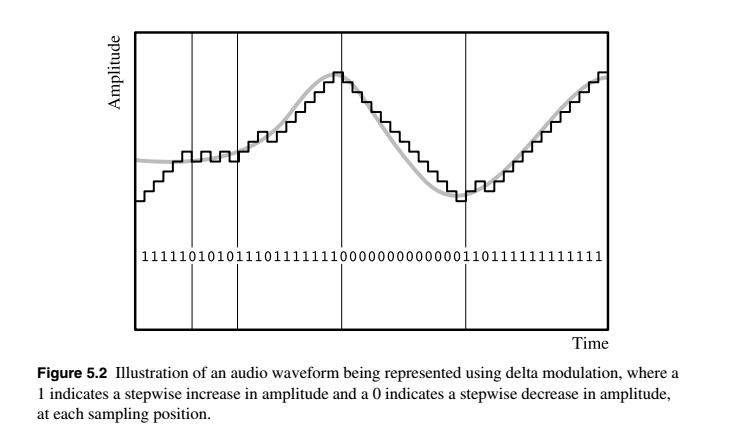
Delta指相邻时刻间的变化量，一般规定为+-1。该系统从0开始累加。每个采样点累加器将当前值与采样值比较，然后按比较出来的大小加一或减一，一次一次进行编码。

这种方式要求的采样率比PCM高很多。

斜率超载（slope overload） 为：在斜率较小波形较为平坦的地方，量化电平只能在每个采样周期间上下跳变，造成特有的失真。

粒状噪声：只能上下跳变的电平对较平缓的信号量化是产生的噪声。

只有提高采样率来减少粒状噪声和斜率超载。



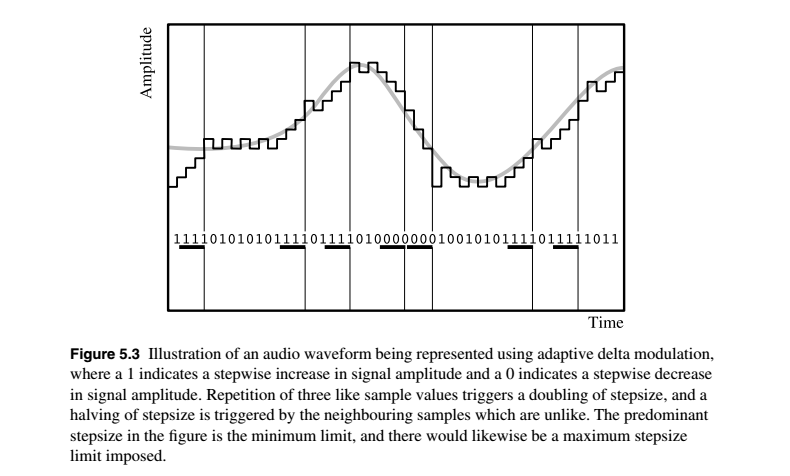
5.1.3 自适应增量调制

自适应增量调制为了解决斜率超载而提出的，通过改变跳变的幅度来工作，又名CVSDM（continuously variable slope delta modulation连续可变斜率增量调制），是一些较老的系统常用的语音压缩方式。

工作方式是在斜率小时量化更精细，斜率大时量化更粗糙（跳动幅度更大）。

制定规则如：过去n个值相同（即连续进行了n次增加/减少）时，加长一倍量化单位，反之减少一半。实际系统的量化规则也是变化的，比如慢慢地加长或减少步长（量化单位）。

一个量化过程如下（n=3）：有三次都是增加或减少是量化步长就增大一倍。

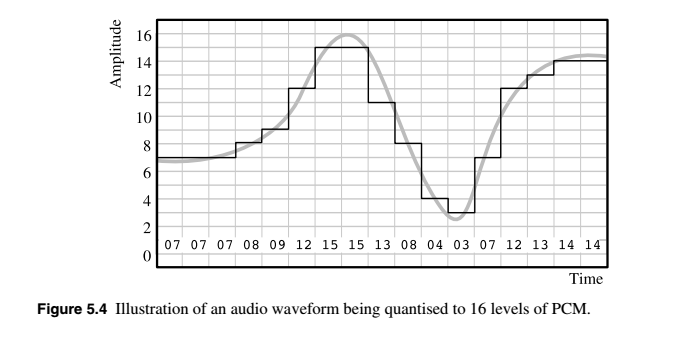


5.1.4 ADPCM

自适应音频脉冲编码(Adaptive differential pulse coded modulation)是自适应增量调制技术在PCM信号上的应用（上节CVSDM一般应用于增量调制码流）。'differential'表明该系统计算相邻采样时刻的差值（差分），'adaptive'表明步长随前若干时刻的采样值变化（适应性）。

从0开始累加，计算本时刻、下一时刻的差值，量化后送入累加器，然后重复上述步骤，量化的差值在每一次累加上传递。

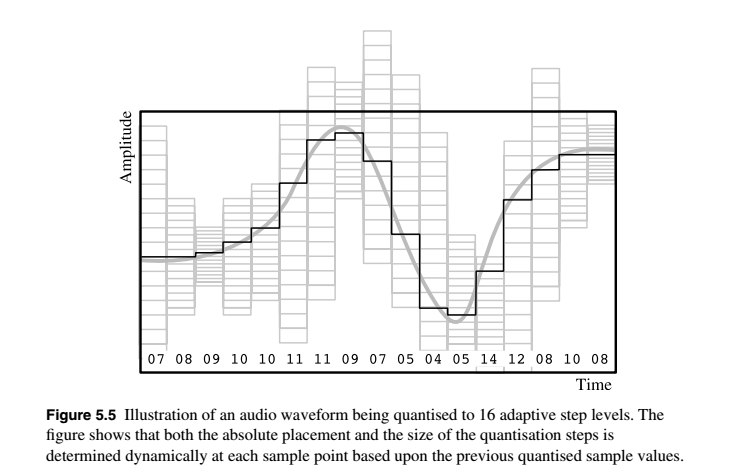
适应性体现在量化精度上，如最高16bit量化的原始音频的ADPCM精度可在3bit、4bit、5bit间变化。



上图为一例：每次量化值的跳变都不同，但由于只计算每次的变化量，信息量一定远小于原始比特率。ADPCM 根据以往每个采样值来决定下一个量化步长。

本例的规则是：如果量化值是处于中间的四个之一（6-10），下一步长减半，反之加长，以此放大小范围的波动，减小大范围的波动。

作用：减小量化噪声，较好地量化斜率大的波形。



5.1.5 SB-ADPCM

ITU标准G. 722提出的、包含两个独立的ADPCM编码单元的编码方式。每个操作一半的频率范围(分别是0-4kHz和4-8kHz)，低频的权重和量化精度高四倍，有更好的保真度（更适用于语音）。

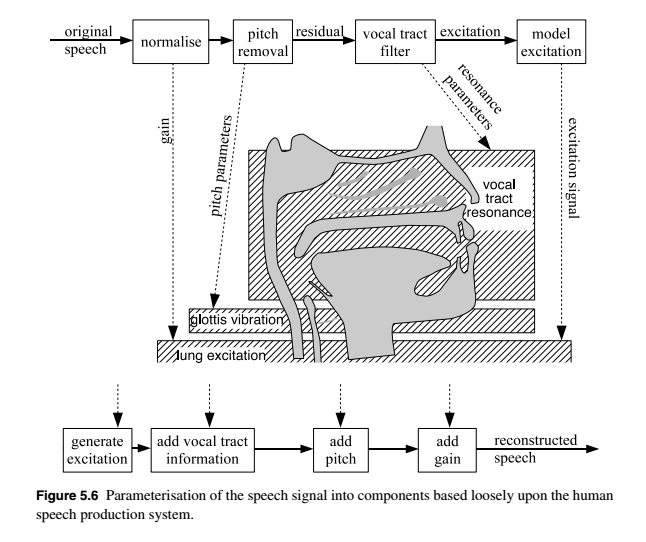
5.2 参数化

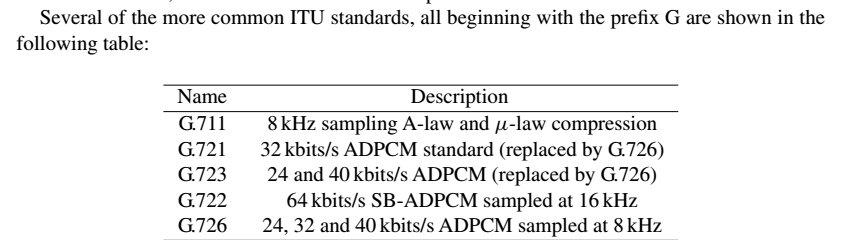
编码技术只能预测跟踪简单有规律的信号，对于信号本身的特征描述有限，能描述的仅包括最大范围和斜率。这种在比特率、编码质量间的协调的策略在高保真理想低比特率编码方案上很有成效，而为了进一步降低数据量，工程师们又提出了参数化用来编码、解码恢复原信号。

合适的参数能是重建波形有更高的保真度、合理的参数量化方式可以提高音质——比如在重要参数上采用更高的精度。

更高级的算法对声音成分、听音系统相关的取参数用以语音合成或识别。

取参方式：整体增益、被编码的语音向量幅度、音调信息、声道共振、肺激励等。



**ITU（International Telecommunications Union 国际电信联合会）的一些编码标准**

5.2.1 线性预测

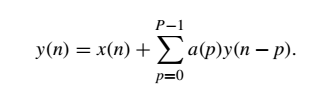
1971年开始应用的古老技术。基于人体发音器官（肌肉）提取一些参数——这些肌肉的运动形成了人的语音，因为肌肉运动速度有限，说话时产生的信号30ms内是近似稳定的。声门的开合对音调的限制一般短于30ms。

伪平稳意味着，240个采样在8Hz/s采样率(对应的持续时间30毫秒)可用更少的参数表示，如8或10个预测系数，这些系数由能够得出原始音频特征的数字滤波器的生成多项式构成。所有多项式的时域响应可能不同，但频域跟原始音频一定是匹配的。线性预测编码(LPC)在应用上获得很大成功，1975年美国联邦标准的1015 2.4kbit/s算法就是基于LPC的。它的缺点是音质不好，不满足军事通信的要求。

5.2.1.1 LPC滤波器

假设有一些参数来表示线性预测，则这些参数有两种可能的用法：用合成滤波器加入到一组采样中去，或者用分析滤波器从一组采样中去除掉。

以下有两种p阶线性滤波器：

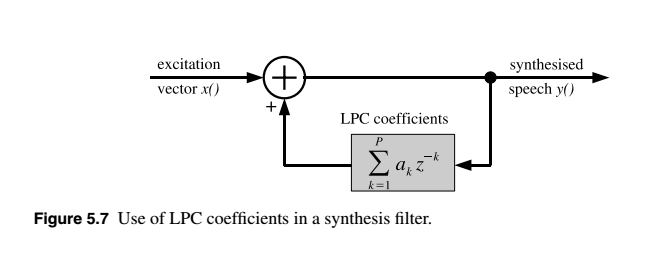


对于上述公式（阶数为P，a(p)是系数），这是一个无限脉冲响应(IIR)滤波器。Matlab中用Durbin–Levinson–Itakura方法来产生所需系数的递归方程。

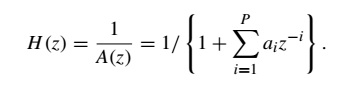
代码：

seg=speech(1:160);%加窗为8k采样率的音频的160个点，也就是20ms，本案例是手动加窗。且采用交叠形式。  
wseg=seg.\*hamming(160);%汉明窗。  
a=lpc(wseg,10); %这里的十阶算上了前向通道的增益1。

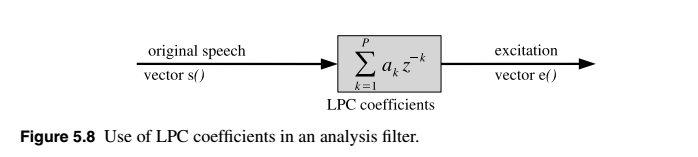
如a=[1; -1.6187;2.3179; -2.9555;2.8862; -2.5331; 2.2299;-1.3271; 0.9886; -0.6126;0.2354]; %a是10阶的，第一个系数为1，表示此滤波器是标准FIR。



用来加入声带特征的滤波器的系统示意图。



此滤波器的系统函数（z域）。



上图是用来去除声带特征的系统示意图。如在z域用ejw代替z可获得频谱响应。

freqz(1, a); %观察a的共振峰

freqz(a); %

freq函数可以画出频响和相位响应，还可以返回一些后续分析所需的参数。比如在n点等间隔幅频分析（fft）中找到第一个共振峰：

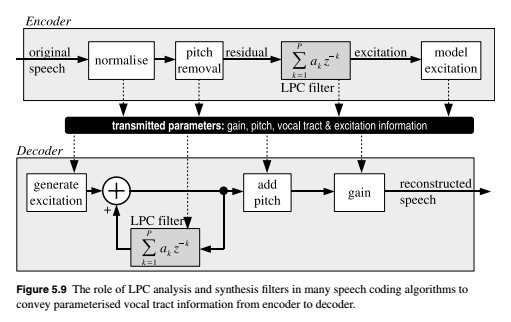
Fs=8000; %采样率

N=100; %频域分辨率  
[H, F] = freqz(1,a,N);

semilogy(0:Fs/(N-1):Fs,abs(H));%y轴对数坐标曲线  
[y,n]=max(abs(H)); %返回共振峰值的坐标  
PeakF=(n-1)\*Fs/(N-1); %返回共振峰值的频率值

上述做法采用LPC（线性预测编码），原理在下图：

编码过程是音量标准化-移除频率-计算残差-得到模型系统函数-模型；解码是逆过程。



5.2.1.2 LPC稳定性讨论

由于参数的保存和传输的格式（双精度的定点数）原因，解码时可能会造成声音不稳定，也许会出现冲激信号。传输误码是可预计和有手段消除的，但LPC量化参数导致的不稳定性问题仍没得到解决。

5.2.1.3 语音信号预增强

实践表明LPC对语音的模拟在低频优于高频，所以可在编码前先增强高频。在生理结构上，发声是从一个很高气压的地方传播到气压低的地方，频谱在高频呈滚降的特点。在发声腔体内部和外部声音是不同的。做预增强的一个原因就是抵消口腔对语音的效应，更好模拟声带发音特征。

对于给定信号s(n)， 施加单输出滤波器(1-az-1) ，a是一个接近1的系数，此处取0.9375。

易得表达式： s’(n)= s(n) - 0.9375 × s(n - 1)

这个函数的作用是滤除不属于声带的特征，以及解决高频不足的问题。

移除这个函数效应使用r(n) = s'(n) + 0.9375 × r(n - 1)

以下假定语音信号为s，程序增强后为es，去增强后为ds：

% Create the emphasis/de-emphasis filter coefficients  
h=[1, -0.9375];  
% Apply the emphasis filter  
es=filter(h, 1, s);  
% Apply the de-emphasis filter  
ds=filter(1, h, es);

%psd()函数为功率谱密度函数，可观察新波形的特性。

Es听起来会有鼻音。

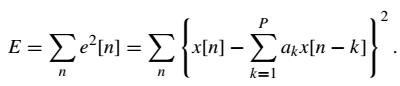
5.2.2 反射系数

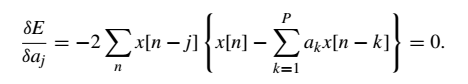
反射系数记录了系统中不同连接点反射声音的状况。有时叫PARCOR（部分相关系数；偏相关系数；高阶偏相关系数）。

递推方程：x'[n] = a1x[n - 1] + a2x[n - 2] + a3x[n - 3] + · · · + aPx[n - P]

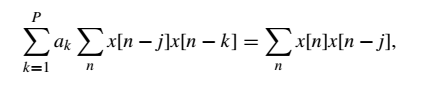
估计误差：e[n] = x[n] - x’[n]

总误差最小时滤波器为最优：



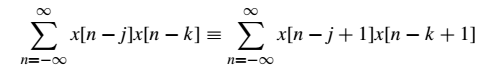
得到最优解是误差对滤波器系数的偏微分应为0：

变形后得：

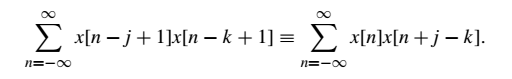


解方程组有两类方法：协方差法和自相关法。前者把声音分成加矩形窗的帧，以减少每一帧的误差；后者假定信号是稳定且能量有限的，求和范围无限。协方差对较小的帧更精确，但对矩形窗造成的截断不解决容易导致不稳定，所以大多使用的是自相关法，常用偏相关系数PARCOR。

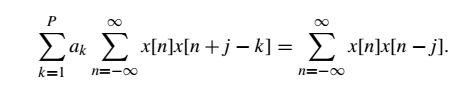
无限求和公式：



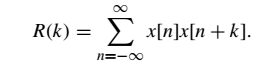
和



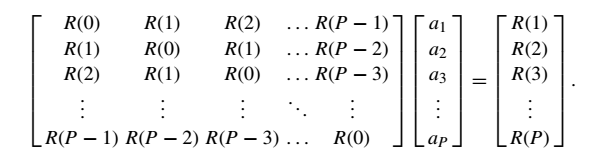
与之前的式子比较得到



自相关的标准公式为：

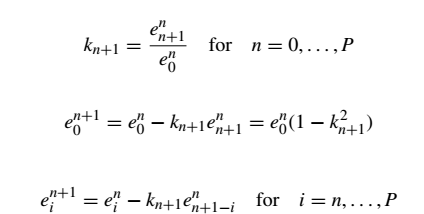


那么得到一系列方程：



实际中常使用汉明窗进行处理，加窗操作在运行自相关计算之前，整个相关结果通常除以r(0)来标准化。标准化后参数记为r(i)。

这个方程组有标准解法，即Durbin–Levinson–Itakura方法，或Le Roux 方法，后一种方法是一系列略微低效但紧凑简单的公式：



递推的结果就是e0i = R(i) ，k即反射系数。

5.2.3 反射系数和LPC间的转换

转换方程：

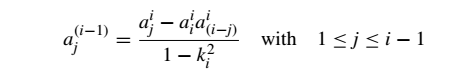


上式i表示时刻，j表示系数的阶数。

从LPC转换到反射系数的公式，初始值为：



后面的为：

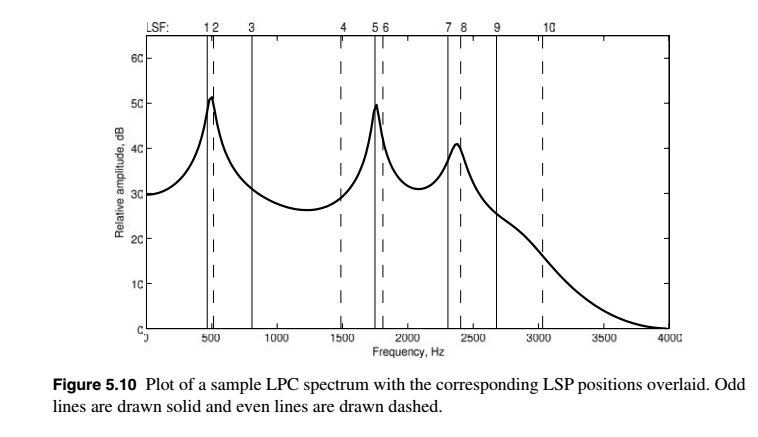


aij是i从P递减到1，1到P之间初始值。

5.2.4线谱对

也叫线谱频率。它描述的是人类声带发声时管道连接处的两种振动状况，包括嘴唇和鼻腔，是由物理意义的计量方式。有两组多项式，分别估计的是声门全开或者全闭的时候的状态。

采用的物理模型是等长度半径不同的管相连接，前述两种情况即连接管的端口状况为打开或关闭。谐振的情况取决于管子的数量和构造。每个谐波频率在奇数和偶数的倍频率上交错分布。



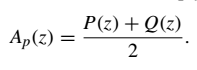
上图的频谱上有十根线，共振峰总是落于实线和虚线之间，谷值也落在虚线实线之间。

5.2.4.1 线谱对的推导

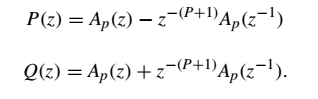
当LPC多项式为：



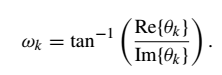
定义两个P+1阶多项式P(z) 、 Q(z) ，一个称为反对称多项式，对应声门全闭，后者称为对称多项式，对应声门全开。如前文所述，声带不会全开或全闭，所以有：



P(z) 、 Q(z)的表达式：



两个多项式有一系列复根θk，每个根在复平面上对应一个角度wk。系统稳定时，θk都在单位圆上。



根θK的求解可以采用试错法（通过不断试验和消除误差求解的方法）、Newton–Raphson 逼近（[牛顿](http://baike.baidu.com/subview/1511/5237091.htm" \t "http://baike.baidu.com/view/_blank)[迭代法](http://baike.baidu.com/view/649495.htm" \t "http://baike.baidu.com/view/_blank)、牛顿-拉夫逊（拉弗森）方法，使用函数f(x)的[泰勒级数](http://baike.baidu.com/view/400903.htm" \t "http://baike.baidu.com/view/_blank)的前两项来寻找方程f(x) = 0的根）、在单位元附近估计符号变化等。不同算法复杂度、精度、数据处理量都不同。

Matlab中可直接用roots()求解。

{wk}这个集合既包含P的根也包含Q的根。

LPC到LSP的转换代码（函数）如下：

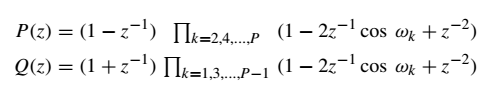
function lsp=lpc\_lsp(a)  
p=length(a); A(1)=1; B(1)=1;  
for k=2:p  
A(k)=(a(k) - a(p+2-k)) + A(k-1);  
B(k)=(a(k) + a(p+2-k)) - B(k-1);  
end  
r1=roots(A);  
r2=roots(B);  
for k=1:p-1  
if (real(r1(k)) < 0)  
theta1(k)=pi-abs(atan(imag(r1(k))/real(r1(k))));else  
theta1(k)=abs(atan(imag(r1(k))/real(r1(k))));  
end  
if (real(r2(k)) < 0)  
theta2(k)=pi-abs(atan(imag(r2(k))/real(r2(k))));  
else  
theta2(k)=abs(atan(imag(r2(k))/real(r2(k))));  
end  
end%求相位  
p=p-1;  
for k=1:p/2  
theta(k)=theta1(k\*2);  
theta(k+(p/2))=theta2(k\*2);  
end %给所有θk赋值  
lsp=sort(theta); %所有线谱频率排序

对应的参数为：

a=[1;-1.6187;2.3179;-2.9555;2.8862;-2.5331;2.2299;-1.3271;  
0.9886;-0.6126;0.2354];  
lsp=[0.3644;0.4023;0.6334;1.1674;1.3725;1.4205;1.8111;  
1.8876;2.1032;2.3801];

5.2.4.2 用LSP推导LPC系数

转换公式：



书上有一系列求解中间过程。

代码（函数）：

function a=lsp-lpc(theta)  
p=length(theta);  
q=cos(theta(1:p));  
f1(10)=1; f1(9)=0;  
for i=1:p/2  
f1(10+i)=-2\*q(2\*i-1)\*f1(10+i-1) + 2\*f1(10+i-2);  
for k=i-1:-1:1  
f1(10+k)=f1(10+k) - 2\*q(2\*i-1)\*f1(10+k-1) + f1(10+k-2);  
end  
end  
f2(10)=1; f2(9)=0;  
for i=1:p/2  
f2(10+i)=-2\*q(2\*i)\*f2(10+i-1) + 2\*f2(10+i-2);  
for k=i-1:-1:1  
f2(10+k)=f2(10+k) - 2\*q(2\*i)\*f2(10+k-1) + f2(10+k-2);  
end  
end  
f1b(1)=f1(11)+1;  
f2b(1)=f2(11)-1;  
for i=2:p/2  
f1b(i) = f1(10+i) + f1(10+i-1);  
f2b(i) = f2(10+i) - f2(10+i-1);  
end  
for i=1:p/2  
a2(i) = 0.5\*(f1b(i) + f2b(i));  
a2(i+p/2) = 0.5\*(f1b((p/2)-i+1) - f2b((p/2)-i+1));  
end  
a=[1,a2];

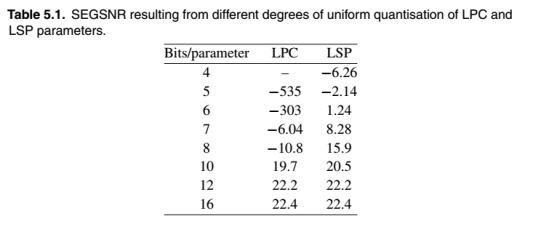
5.2.4.3 线谱对可视化

函数

function lpcsp(a, lsp)  
[HH, FF]=freqz(1, a, 100);  
semilogy(abs(HH),’m-’);  
hold on V=axis;  
axis([1,length(FF),V(3),V(4)]);  
hold on;  
lsc=100/pi;  
for lp=1:length(lsp)  
line([1+lsp(lp)\*lsc,1+lsp(lp)\*lsc], [V(3),V(4)]);  
end  
hold off;

作用是将频谱图和线谱整合到一起方便观察。

5.2.5 量化问题

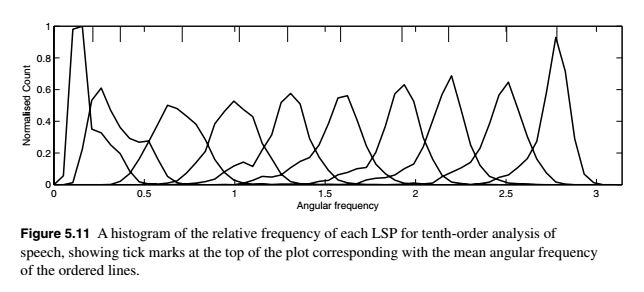


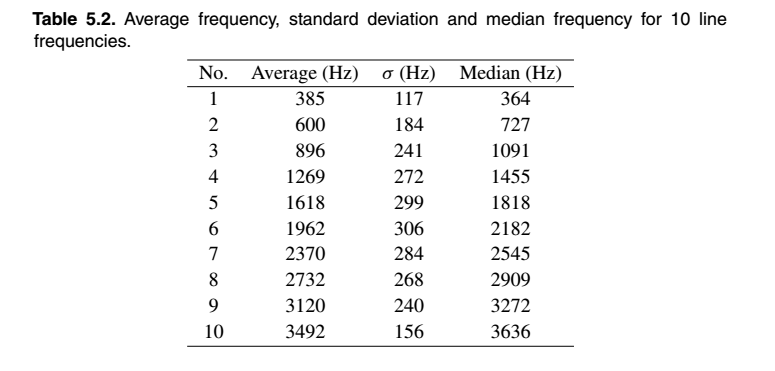
如表，lsp的信噪比受精度影响的程度比lpc小。除此外lsp还有其他优越性，所以传输时更倾向使用lsp编码。

Lsp编码有两种差分方式，一种是前后帧之间的，还有一种是每一帧和所有帧的平均值差分的。有时也分别用“短时”和“长时”来区分。

长时平均差分（long-term average differential quantisation ）计算每一帧lsp与平均值的差值，也叫插值lsf（lsfi）。其外另有一种方式是每四个子帧（一个标准帧）做一次插值，一定程度上避免传输中突发错误的影响。

线谱坐落在不重要的、对清晰度影响低的频段，所以受精度影响小。





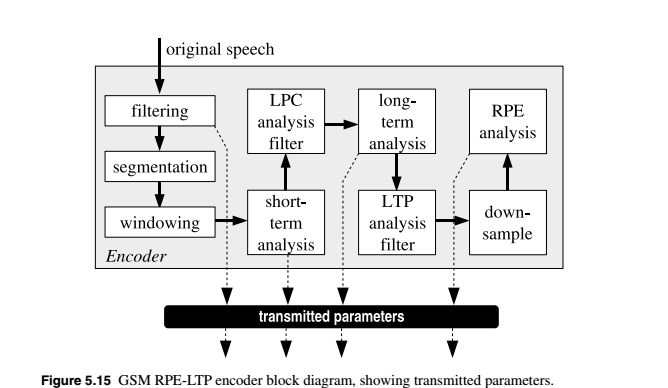
实验证明使用差分比不使用效果好，矢量量化比标量量化效果好。

5.3音高模型

l pc和lsp是声带的建模，音高模型是声门的建模。对它的参数影响因素有：脉冲的形状、正负向尖峰的高度、最大尖峰振幅、脉冲之间的间距等。

5.3.1 Regular pulse excitation规则脉冲激励

Gsm标准的一种模型。对于一帧（160个采样）先是用lpc滤除声带的影响，保留残差。残差分为四个子帧，每个子帧分别做成音高模型后相互连接，用ADPCM参与传输，无音高是频率模型为白噪音。



5.3.2 LTP音高提取（LTP pitch extraction ）

定义一系列节拍和延迟为长期预测的措施。

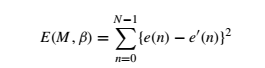
一阶如x(n) = c(n) + βx(n − M )

三阶如x(n) = c(n) + β1x(n − M − 1) + β2x(n − M ) + β3x(n − M + 1).

β是延迟，M是对应的基音的响应幅度。

这个措施用来提高音调的质量。通常M和beita已知，预测时gsm标准中常常声道模型不变但音高会变，也就是不同的帧M经常不同。

5.3.2.1 音高提取

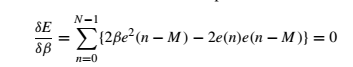
即估计M的方式，基于最小均方误差准则的LPC方法。

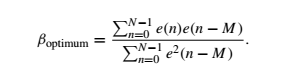
e(n)表示残差，e’(n)表示预测的音高模型。假设是一阶预测那么：



N为分析窗长度。

同样地误差对系数求导应为0：



得到最优系数：

最优延迟：

由公式，最佳系数和延迟时间是用整个帧的采样来估计的。

function [B,M]=ltp(sp)

n=length(sp);

%Establish upper and lower pitch search limits

pmin=50; pmax=200;

sp2=sp.ˆ2; %pre-calculate to save time

for M=pmin:pmax

e\_del=sp(1:n-M);

e=sp(M+1:n);

e2=sp2(M+1:n);

E(1+M-pmin)=sum((e\_del.\*e).ˆ2)/sum(e2);

end

%Find M, the optimum pitch period

[null, M]=max(E);

M=M+pmin;

%Find B, the pitch gain

e\_del=sp(1:n-M);

e=sp(M+1:n);

e2=sp2(M+1:n);

B=sum(e\_del.\*e)/sum(e2);

5.3.3 音高问题

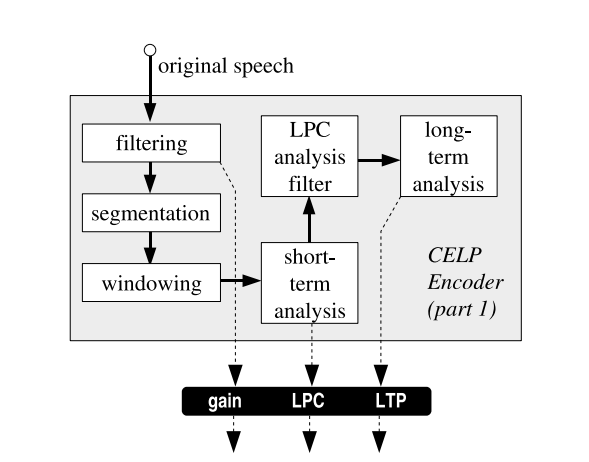
上述方法容易遇到音高减半增倍的问题，可限定频率范围解决，限定方式主观、凭经验。

5.4 分析合成法

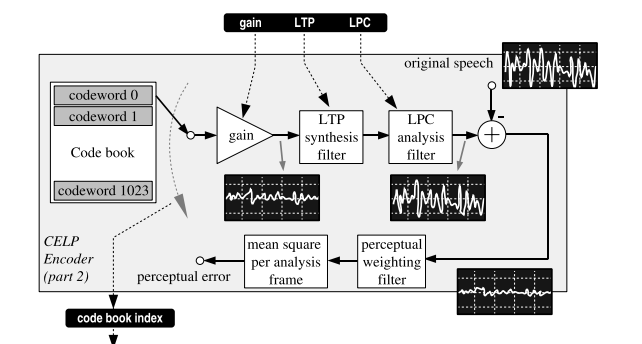
过程为提取特征-重建，评价方式不用最小均方误差，因为人耳听能的特征。常用

perceptual weighting filter和频谱失真来评价。

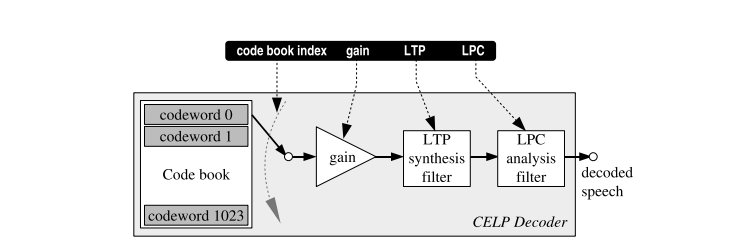
5.4.1 基本celp （码激励线性预测编码）



上图的分析前文都出现过。



这个部分将编码与码本比较，输出码的序号，同时计算输出编码主观误差，误差是由LPC编码造成的，编码后的声音与原音频有区别（即存在误差）。



上图为解码流程。得到码本中的码后进行LTP合成、LPC分析得到原码。

5.4.1.1 码本

码本由码字组成。最接近采样的码字被用来表示采样。20ms一帧的系统（8kHz采样率）中，每一帧由160个采样组成，每一个子帧的长度是五毫秒，LPC、LTP的等参数都是每帧计算四次。码本由1024个160位的随机数组成，另外有一些技术可以增强码本中码的实用性。

CELP编码器内部包含一个解码器，以预测解码后的情况，和对两边信号进行同步。

5.4.2 代数码激励线性预测

一种减少运算的循环次数的方式。基本celp的操作都是线性时不变的，实践证明放大这个步骤在最后处理更好。

一个lpc滤波器

y(n) = a(1) × x(n) + a(2) × x(n − 1) + a(3) × x(n − 2) + · · · + a(m + 1)× x(n − m)

由于ACELP的码字是稀疏的，上式的多项式会出现很多0，从而简化运算。

且由于代数码码字简单，也方便实时产生许用码字从而节约存储空间。

5.4.3 分离码本（Split codebook schemes）

不同与ACELP，分离码本的工作方式是减少搜索空间。

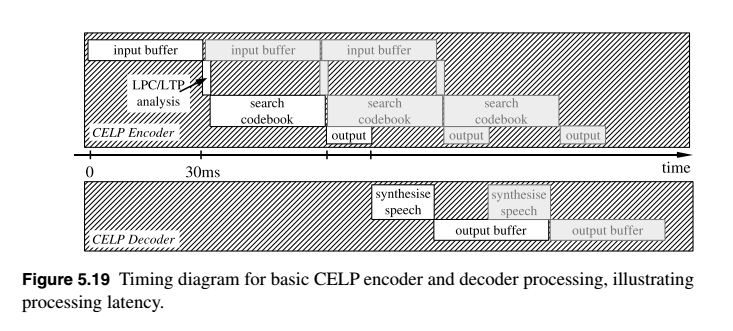
一种模式是将码字归类，在符合描述的类下搜索。

另一种采用两个相互正交的码本（此处正交定义为：两个码本中的码字是一一对应的）。

混合激励线性预测（MELP）不需要码本就能分离出激励信号（也不采取分析合成法）。

矢量和激励线性预测（VSELP）的码字对不同的码本分别搜索，然后求和来表示向量，每个码本的缩放比例也是不同的。可以替代LTP。

5.4.4 前向-后向码激励线性预测Forward–backward CELP



数字系统输入输出缓冲会造成系统延迟（硬件延迟），中间的运算步骤同上。一般系统的延迟约200-300ms，人耳是容易察觉的，且较长的暂停可能造成听者心理上的不适，如能将暂停缩短到100-150ms则能改善很多。

I TU G.728标准解决了此问题，方式即前向-后向码激励线性预测。该方式并行执行码本搜索和缓冲，编码使用的参数是前一帧的。

这种模式可以极大缩短系统延迟（硬件延迟）。

**总结：**

这一章主要内容是语音处理传输系统的介绍，主要涉及业界常用规格和基础方法。

开始比较有疑问的是残差这个概念。应该是指去除LPC后剩下的那一部分，它被用来构造音高模型。LPC是线性预测，对于实际上不稳定的语音信号预测会有误差，这个误差就是残差，这样理解对吗？经常在处理过程中认为声带模型（LPC）变化比音高模型慢很多。

分析合成法即是使用分析模型来对语音信号编码，用合成方式来解码。

码激励线性预测也是LPC、音高模型实际应用的一种形式，用有限的码本来表示数据。