第5章 语音编码

语音信号的数字化传输一直是通信发展的主要方向之一，语音的数字通信与模拟通信相比，无疑具有更好的效率和性能，这主要体现在：①具有更好的话音质量；②具有更强的抗干扰性，并易于进行加密；③可节省带宽，能够更有效地利用网络资源；④更加易于存储和处理。最简单的数字化的方法是直接对语音信号进行模/数转换，只要满足一定的采样率和量化要求，就能够得到高质量的数字语音。但这时语音的数据量仍旧非常大，因此在进行传输和存储之前，往往要对其进行压缩处理，以减少其传输码率或存储量，即进行压缩编码。传输码率也称为数码率或编码速率，表示为传输每秒钟语音信号所需要的比特数。语音编码的目的就是要在保证语音音质和可懂度的条件下，采用尽可能少的比特数来表示语音。

早在20世纪30年代末期，语音编码技术的研究已经开始。而近十几年来，在数字通信领域实际需求的强力推动下，随着计算机技术的高速发展，语音编码技术的研究获得了突飞猛进的发展，并得到了广泛的应用，由此形成了比较完善的理论和技术体系。具体表现为，当今世界上存在着数量众多的语音编码的国际标准和地区性标准，并且该领域也成为国际标准化工作中最为活跃的研究领域。

最早提出的语音编码标准是数码率为64Kbps的PCM波形编码器，而在20世纪90年代中期出现了很多被广泛使用的语音编码国际标准，例如：数码率为5.3/6.4Kbps的G.723.1、数码率为8Kbps的G.729等。此外，也存在着各种未形成国际标准，但数码率更低的成熟的编码算法，有的算法数码率甚至可以达到1.2Kbps以下，但仍能提供可懂的语音。

语音编码方式有很多种划分方法。从数码率的角度可以将语音编码划分成五大类：高速率(32Kbps以上)、中高速率(16～32Kbps)、中速率(4.8～16Kbps)、低速率(1.2～4.8Kbps)和极低速率(1.2Kbps以下)。

从采用的编码方法的角度还可以分为三类：波形编码、参数编码和混合编码。波形编码是根据语音信号的波形导出相应的数字编码形式，其目的是尽量保持波形不变，使接收端能够忠实地再现原始语音。波形编码具有抗噪性能强、语音质量好等优点，但需要有较高的数码率，一般为16～64Kbps。参数编码又称为声码器技术，它通过对语音信号进行分析，提取参数来对参数进行编码。在接收端能够用解码后的参数重构语音信号，参数编码主要是从听觉感知的角度注重语音的重现，即让解码语音听起来与输入语音是相同的，而不是保证其波形相同。参数编码一般对数码率的要求要比波形编码低得多。混合编码是上述两种方法的有机结合，同时从两个方面构造语音编码，一方面增加语音的自然度，提高了语音质量，另一方面相对于波形编码实现较低的数码率指标。

在对语音信号压缩很多倍后仍可以得到可懂的语音，是因为语音信号中存在大量的冗余信息，而语音编码就是利用各种编码技术减少语音信号的冗余度。此外语音编码中也充分地利用了人耳的听觉掩蔽效应，一方面去除将会被掩蔽的语音信号，实现数据的压缩；另

一方面控制量化噪声,使其低于掩蔽阈值,即使在较低数码率的情况下,也能获得高质量的语音。

在本章中,5.1节主要介绍几种常用的波形编码算法;5.2节介绍参数编码器和混合编码器;5.3节介绍极低速率语音编码技术;在5.4节中对语音编码器的性能指标和质量评测方法进行讨论;最后在5.5节中对语音编码国际标准的情况进行介绍。

5.1 波形编码

5.1.1 均匀量化 PCM

最直接的语音数字化的方法是对其进行A/D转换,包括采样和量化两个过程。采样时,采样频率要高于信号中最高频率的两倍,以避免发生混叠失真。因此一般情况下在采样前应该进行抗混叠滤波,即进行低通滤波,以控制信号的最高频率。量化时将采样得到的样本的幅度用均匀量化的方法表示成二进制数字信号,相当于用一组二进制脉冲序列表示各量化后采样值,于是语音波形信号就被表示成一组用数字编码的脉冲序列。这种编码方法被称为脉冲编码调制(pulse coding modulation,PCM),其编码原理如图5-1所示。

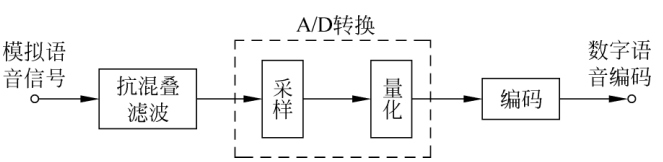


图5-1 PCM编码原理图

量化过程不可避免地会产生误差,量化误差$e(n)$可以定义为

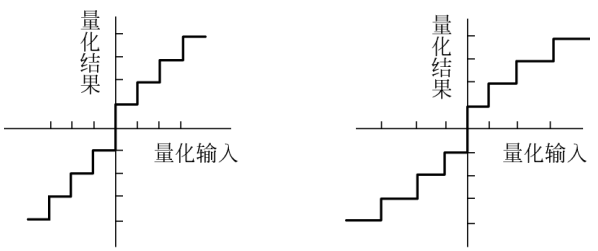
$$e(n)=\bar{x}(n)-x(n) \tag{5-1}$$

式中,$\bar{x}(n)$为量化后的信号,$x(n)$为量化前的采样信号。

量化误差也被称为量化噪声。对于均匀量化器来说,量化噪声的功率仅取决于量化间隔Δ,而与输入信号的功率及概率分布无关。如公式(3-3)所示,可以计算出当$B=12$时,采样频率8kHz的均匀量化器所产生的数字语音的信噪比可达60dB,基本上可以满足高质量的电话通信要求。此时PCM的编码速率为8kHz×12=96Kbps。

5.1.2 非均匀量化 PCM

均匀量化PCM编码器的主要问题是编码速率高。由于要满足一定信噪比的要求,所以量化间隔就不能太大,而当语音信号动态变化范围较大时,为了防止幅度较大的信号因超出量化范围而出现过载,必须使用较高的量化比特数。解决的方法是,依据语音信号的幅度统计分布特性,进行非均匀量化。在语音信号中,样本的幅度值不是均匀分布的,信号大量地集中在小幅度值上。如果对小幅度样本使用小的量化间隔,则可以进行精确量化;若对大幅度样本使用大的量化间隔,则既可成功地提高信噪比,又可避免大信号的过载。均匀量化和非均匀量化的特性如图5-2所示。



(a) 均匀量化特性　　(b) 非均匀量化特性

图 5-2　均匀与非均匀量化特性

最常用的非均匀量化方法是对数压扩方法。编码时，利用语音信号的幅度统计特性，对幅度按对数变换进行压缩，然后再进行均匀量化。解码时，则进行逆向的扩张变换。在实际使用中有各种不同的变换方法，如 μ 律变换、A 律变换等。

设 $x(n)$ 为语音波形的采样值，则 μ 律压缩定义为

$$y(n)=F_{\mu}[x(n)]=X_{\max}\frac{\ln\left[1+\mu\dfrac{|x(n)|}{X_{\max}}\right]}{\ln(1+\mu)}\mathrm{sgn}[x(n)] \quad (5\text{-}2)$$

即将输入语音 $x(n)$ 压缩变换为 $y(n)$，然后再进行均匀量化编码。式中，$X_{\max}$ 是 $x(n)$ 的最大幅值，μ 是常数，用于调节压缩的程度，μ 越大其压缩程度越高。当 $\mu=0$ 时表示不进行压缩，通常 μ 值在 100～500 之间取值。图 5-3 给出了不同 μ 值时 μ 律的压扩特性曲线。

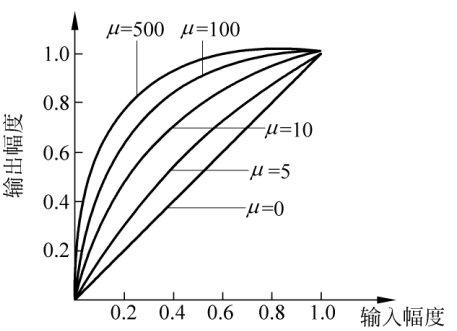


图 5-3　μ 律特性的输入输出关系

A 律的压缩方法与 μ 律相似，按如下公式进行：

$$y(n)=F_A[x(n)]=\begin{cases}\dfrac{A|x(n)|}{1+\ln A}\mathrm{sgn}[x(n)] & \left(0\leqslant\dfrac{|x(n)|}{X_{\max}}<\dfrac{1}{A}\right)\\ X_{\max}\dfrac{1+\ln[A|x(n)|/X_{\max}]}{1+\ln A}\mathrm{sgn}[x(n)] & \left(\dfrac{1}{A}\leqslant\dfrac{|x(n)|}{X_{\max}}\leqslant 1\right)\end{cases} \quad (5\text{-}3)$$

目前，非均匀量化的 PCM 编码广泛地应用在数字电话网中，北美和日本主要使用 μ 律压缩，我国则采用 A 律压缩。

5.1.3　自适应量化 PCM

除了上文介绍的非均匀量化的方法外，还可以通过自适应量化的方法来提高信噪比。由于语音信号的特性是随时间变化的，能量时大时小，因此可以采用自适应的方法，对短时能量比较大的信号，采用比较大的量化间隔进行量化，相反的，对短时能量比较小的信号，可以采用比较小的量化间隔进行量化，这样有助于减少量化噪声，提高量化后信号的信噪比。这种方法称为自适应量化 PCM(adaptive PCM，APCM)。它的量化器特性随着输入信号短时能量的变化而自适应地变化。在自适应量化器中，除了可以采用量化间隔作为量化器的特性外，还可以采用放大增益来作为量化器特性，实现时在固定量化器前加一个自适应的增

益控制,对能量较大的信号采用较小的放大增益,对能量较小的信号,采用较大的放大增益。可以看出,这种自适应改变放大增益的方法,与自适应的改变量化间隔的方法是等效的。显而易见,APCM 编码器除了要发送量化结果外,还需要发送自适应调整参数作为边信息,使解码端能获知当前采样点的量化器特性。

可以根据下式计算自适应参数:

$$\begin{cases}\Delta(n) = \Delta_0 \cdot \sigma(n) \\ G(n) = G_0/\sigma(n)\end{cases} \tag{5-4}$$

$\Delta(n)$和$G(n)$分别对应第n个采样点的量化间隔和放大增益。其中$\sigma^2(n)$为输入语音信号的方差。式(5-4)表明,$\Delta(n)$正比于输入信号方差$\sigma^2(n)$,通常认为,时变的方差$\sigma^2(n)$正比于信号的短时能量,因此$\Delta(n)$也就正比于信号的短时能量。而$G(n)$反比于信号的方差和短时能量。

APCM 的自适应方案又可分为前馈自适应和反馈自适应两种。采用前馈自适应方案,$\Delta(n)$和$G(n)$是由输入信号本身估算出来的。而采用反馈自适应方案,则是用量化器的输出来估算$\Delta(n)$和$G(n)$,即用前面信号的情况来估算后面信号的短时能量和方差。因此,前馈自适应能得到更好的信噪比指标,但需要一定的编码延迟,而反馈自适应方案不需要传输边信息。

采用自适应量化后可以提供更高的信噪比,一般可以得到约 4～6dB 的编码增益。

5.1.4 差分脉冲编码

语音编码就是通过减少语音信号中的信息冗余度来实现数据压缩,这种冗余度的最直接的证据,就是语音采样信号之间具有很强的相关性。分析表明,当采样频率为 8kHz 时,相邻采样值之间的自相关系数一般在 0.85 以上。可以利用这种相关性减小量化字长,从而降低编码速率。由于相邻采样值之间的差值远小于采样值本身,因此可以设计一种编码方法,对差值进行编码,而不是对采样值本身进行编码,这种编码方法称为差分脉冲编码(difference PCM,DPCM)。

产生差分信号的最简单的方法是直接存储前一次的采样值,然后用本次采样值去计算差值,经量化得到数字语音编码。解码端则做相反的处理,恢复原信号。其原理如图 5-4 所示。图中$x(n)$为输入语音,$d(n)$为差值信号,Q[·]为量化器,$c(n)$为语音编码,$\bar{x}(n)$为解码后的语音。

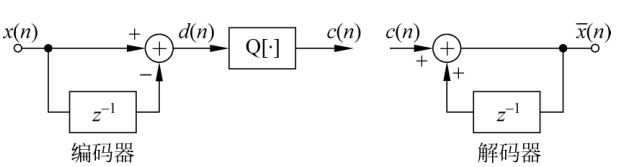


图 5-4 DPCM 原理图

用Z变换考察各点信号的时域关系,有

$$C(z) = X(z)(1 - z^{-1}) + E(z) \tag{5-5}$$

和

$$\overline{X}(z)=\frac{C(z)}{1-z^{-1}}=X(z)+\frac{E(z)}{1-z^{-1}} \tag{5-6}$$

式中，$E(z)$为量化器量化噪声 $e(n)$的 Z 变换。

由式(5-6)可以看出，量化器所产生的量化噪声被累积叠加到了输出信号中，即每次的量化噪声信号都被记忆下来，然后叠加到下一次输出中。如果量化噪声始终是同一方向，则输出信号会越来越偏离正常信号。为了解决这一问题，编码器应该用前一次解码后的采样值替代前一次的输入采样值，以生成差分信号。如图 5-5 所示，编码器通过反馈的方式由差分编码重构生成前一次的采样值。

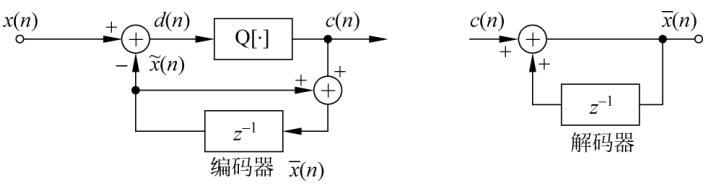


图 5-5　实际 DPCM 结构图

采用如图 5-5 所示的结构后，若一个采样点的量化噪声信号为正，则重构的采样值 $\bar{x}(n)$必将大于 $x(n)$，在下一个时刻，由于使用重构的采样值来计算差分，使差分信号变小而抵消上一次量化噪声的影响。从 Z 变换的角度进行分析会得到同样的结论，从图可知

$$\widetilde{X}(z)=\frac{C(z)z^{-1}}{1-z^{-1}} \tag{5-7}$$

编码结果为

$$C(z)=X(z)-\widetilde{X}(z)+E(z) \tag{5-8}$$

将式(5-7) 带入到式(5-8)中，得

$$C(z)=(X(z)+E(z))(1-z^{-1}) \tag{5-9}$$

因此有

$$\overline{X}(z)=\frac{C(z)}{1-z^{-1}}=X(z)+E(z) \tag{5-10}$$

可见，已经消除了量化噪声的积累。

上面所叙述的是差分脉冲编码的一种简单形式，它仅利用两个相邻采样值之间的相关性。实际上，当前输入的采样值不仅与上一时刻的采样值相关，而且也与前面若干个采样值相关，充分利用这些相关性无疑能够得到更多的编码增益。可以应用第 4 章曾详细讨论过的线性预测分析的方法来实现一般形式的差分脉冲编码。根据线性预测分析的原理，可以用过去的一些采样值的线性组合来预测和推断当前的采样值，得到一组线性预测系数，且预测所带来的误差 $e(n)$的动态范围和平均能量均比信号 $x(n)$要小得多，预测阶数越高，预测误差就越小，相应的编码速率就可以越低。图 5-6 为采用线性预测的 DPCM 的一般结构图。

图 5-6 中 $P(z)$为线性预测多项式，a_i为线性预测系数，p 为预测阶数。有

$$P(z)=\sum_{i=1}^{p}a_iz^{-i} \tag{5-11}$$

可以看出，当预测阶数为 1，且 $a_1=1$ 时，就得到前文所述简单形式的差分脉冲编码器。

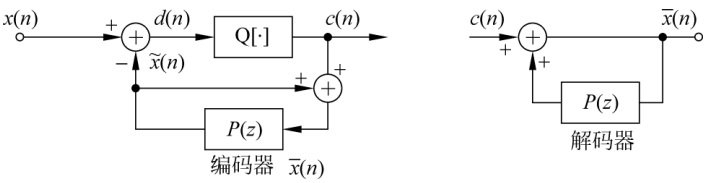


图 5-6 DPCM 的一般结构图

差分脉冲编码采用差分(预测误差)信号进行编码,由于差分信号能量比原输入信号能量要小得多,因此量化限幅电平也可以小得多。这样在量化电平数不变的条件下,差分量化器的量化间隔就可以比原输入信号的量化间隔小,从而减少量化噪声。因此差分编码的信噪比将比直接对原信号编码的 PCM 高,由此得到差分增益或称预测增益,其值等于原信号能量和差分信号能量之比。

从另一角度来讲,在保持信噪比不变的条件下,差分编码器可以通过减少量化字长,即减少量化电平数的方法来降低编码速率。分析表明,1 阶预测 DPCM 的差分增益可为 5dB,可比 PCM 减少 1 比特编码长度,即编码速率可降低到 56Kbps。3 阶预测 DPCM 能减少 1.5～2 比特编码长度,编码速率可降低到 48Kbps。

5.1.5 自适应差分脉冲编码

1. 自适应差分脉冲编码的原理

差分编码器的编码速率能降低到什么程度,主要取决于其预测精度,即其预测误差的大小。上节所述的 DPCM 采用的是固定系数的线性预测器,从第 4 章的内容可知,由于语音信号的不平稳性,显然不能保证其总是最佳预测器,从而使预测误差最小。比较好的方法是在编码的过程中,采用自适应技术动态地调整预测器系数。此外,用自适应量化技术对差分信号进行量化,也能进一步降低编码速率。一般将采用自适应量化及高阶自适应预测的 DPCM 称作自适应差分脉冲编码(adaptive DPCM,ADPCM)。

前馈型 ADPCM 的编码原理如图 5-7 所示,与图 5-6 相比较可知,系统的核心部分与 DPCM 相同,但 $P(z)$的系数受自适应逻辑控制,另外增加了自适应量化的功能。

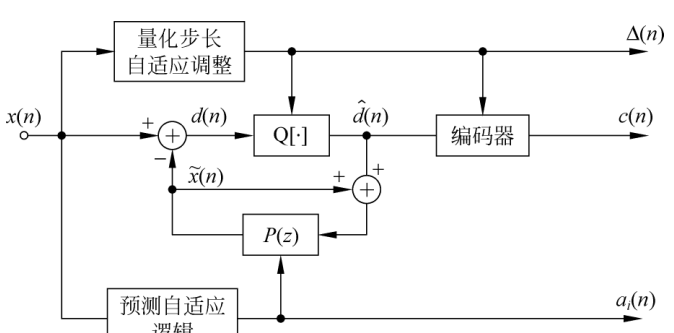


图 5-7 ADPCM 系统编码器原理图

从图 5-7 可知，当自适应量化采用前馈自适应时，编码器输出包括 3 类信息：

(1) 预测误差信号编码码字 $c(n)$；

(2) 预测器系数 $a_i(n)$；

(3) 量化间隔 $\Delta(n)$或者增益因子 $G(n)$。

如果自适应量化采用反馈自适应方法，编码器就不必传送 $\Delta(n)$和 $G(n)$，而由解码端根据前面的信号估算得到。

自适应线性预测以帧为单位进行，根据本帧语音波形的时间相关性确定预测系数，使预测误差信号的方差最小。可以采用第 4 章所述的自相关函数法等方法求取线性预测系数。自适应线性预测又可以分为前向预测和反向预测两种，前向预测采用当前帧的采样值计算出预测器系数，然后计算当前帧的预测信号，得出预测误差信号进行编码。其预测精度较高，并可获得较低的编码速率，代价是引入一帧时间的算法时延。反向预测采用上一帧的样本值算出预测器系数，以此预测器计算当前帧的预测信号，它虽然没有算法时延，但预测精度较低。

2. G.726 语音编码

ADPCM 已形成国际标准，ITU-T(原 CCITT)在 1988 年制定了 G.726 标准，将 1984 年和 1986 年分别制定的 ADPCM 标准 G.721 和 G.723 进行了合并，同时也删除了上述两个标准。G.726 能提供 4 种数码率：40Kbps、32Kbps、24Kbps、16Kbps。其语音质量相当于 64Kbps 的 PCM 编码，并具有很好的抗误码性能。图 5-8 为 G.726 的编码器方框图。编码器的输入为 8 位的 A 律或 μ 律 PCM 信号，首先通过转换器将其转换为 14 位的均匀量化的 PCM 编码。然后减去线性预测器输出的预测信号 $x_e(n)$，得到预测误差信号 $d(n)$，再经非均匀自适应量化器得到编码信号 $c(n)$。一方面将 $c(n)$传送给解码器；另一方面将其输入反向自适应量化器进行 D/A 转换，还原得到模拟量化差分信号 $d_q(n)$，供反馈回路生成重构信号和预测信号。自适应量化器和反向自适应量化器均受尺度因子 $y(n)$控制，其量化特性的变化与信号的动态范围相匹配。自适应量化速度控制器采用双模式自适应：对幅度变化较大的语音信号进行快速处理，其标尺因子为 $y_u(n)$；对幅度变化较小的带内数据和信令进行慢速自适应处理，其尺度因子为 $y_l(n)$。总的标尺因子 $y(n)$为 $y_u(n)$和 $y_l(n)$的线性组合，即

$$y(n) = k_1(n)y_u(n-1) + [1-k_1(n)]y_l(n-1) \tag{5-12}$$

式中，$k_1(n)$为自适应控制参数，有 $0 \leqslant k_1(n) \leqslant 1$。$k_1(n)$由自适应速率控制器模块根据差分信号变化速率确定。对于语音数据，$k_1(n)$趋于 1，对于带内数据或信令，$k_1(n)$趋于 0。$t_r(n)$和 $t_d(n)$为信号音检测信号，由信号音和转换检测器生成，供自适应控制模块转换适应模式。

自适应预测器根据量化差分信号 $d_q(n)$计算预测信号 $x_e(n)$，用一个两阶的全极点滤波器和一个六阶的全零点滤波器实现。G.726 采用反馈型自适应和反向预测的方法，编码中仅包括预测误差信号编码，不包含预测系数和自适应量化器的量化间隔或增益因子等参数。

解码器方框图如图 5-9 所示，其模块基本上与编码器中的反馈回路部分相同。其中同步编码调整模块的作用是防止同步级联情况下产生累计失真，调整 PCM 输出编码以消除后面一个 ADPCM 级的量化失真。

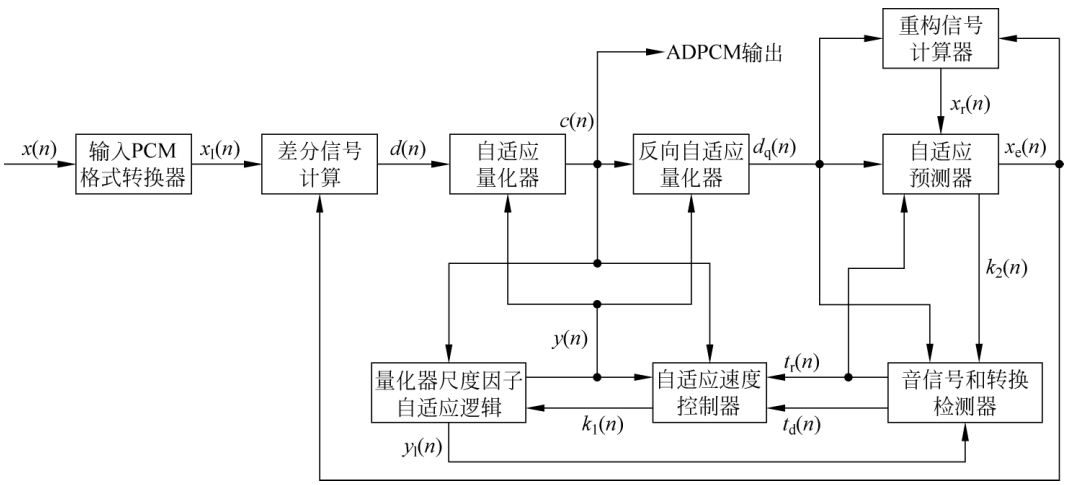


图 5-8 G.726 编码器方框图

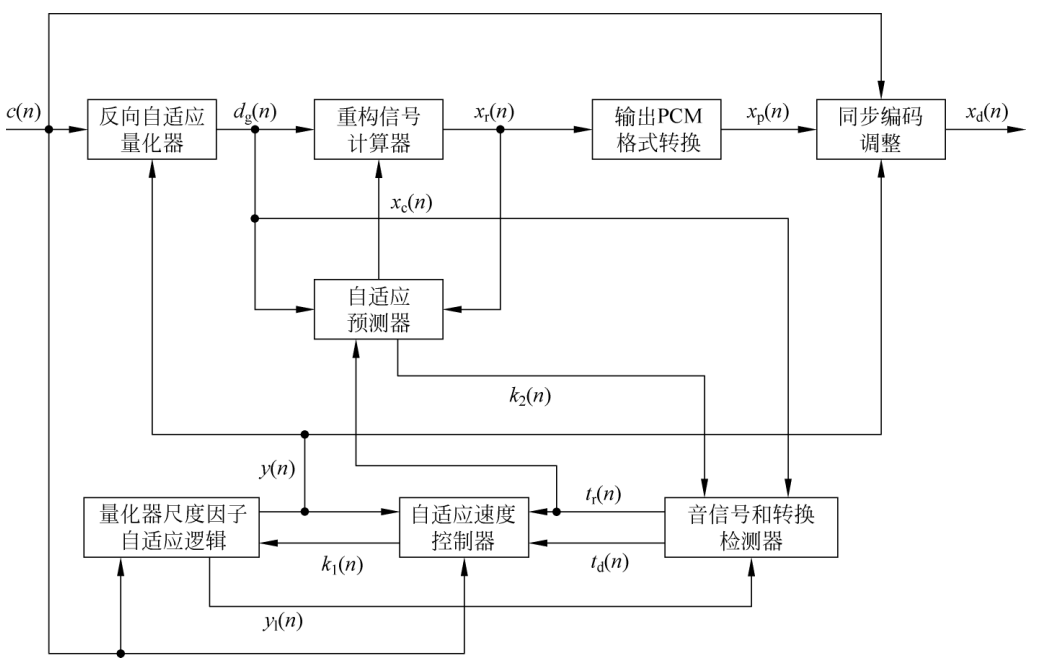


图 5-9 G.726 解码器方框图

3. 长时预测和噪声整形

在 ADPCM 系统中增加长时预测和噪声整形机制,可以进一步改善编码质量。ADPCM 中的线性预测器是利用相邻若干样本的采样值来预测当前样本的采样值,这种预测经常被称为短时预测。实际上,对短时预测所得到的预测误差信号还可以再次进行长时预测,从而得到功率更小的差分信号,获得更高的编码增益。浊音信号是准周期信号,其周期相当于基音周期,因此相邻周期的样本之间具有很大的相关性。经过短时预测之后,预测

误差序列仍然保持着这种相关性，从而显示出明显的周期性。利用这种周期性再次进行预测，预测器函数为

$$P(z) = \beta z^{-D} \tag{5-13}$$

式中，β 为预测系数，D 为基音周期。即用上一个基音周期的采样值来预测当前周期的采样值。这样，用预测信号计算获得的差分信号必然因去除了周期性而功率更小，从而可以进一步压缩量化字长。为了与短时预测的概念相区别，经常将这种基于基音周期的预测称为长时预测。

在语音编码中，量化器不可避免地会产生量化噪声。这种量化噪声可以近似地看做是高斯白噪声，即噪声谱是平坦的。但是由于人耳的听觉灵敏度在整个谱上并不是均匀分布的，因此方差最小的量化噪声信号对人耳的感觉来说不一定是最小的。如果能整形噪声谱，使其在人耳感觉灵敏的频段内噪声能量小，而相对地在人耳不灵敏的频段内噪声能量大，无疑会使噪声更不易被察觉，从而提高语音质量。噪声整形的工作原理如图 5-10 所示。

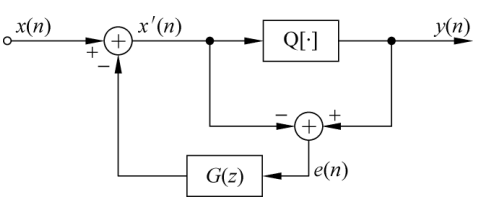


图 5-10　噪声谱整形工作原理图

量化噪声通过噪声整形滤波器 $G(z)$ 进行负反馈，$E(z)$ 为整形前的量化误差 $e(n)$ 的 Z 变换，$E'(z)$ 为整形后的量化误差，量化器输出为

$$Y(z) = X'(z) + E(z) = X(z) - E(z)G(z) + E(z) \tag{5-14}$$

$$E'(z) = [1 - G(z)]E(z) \tag{5-15}$$

对 $E(z)$ 的频谱按 $1-G(z)$ 进行整形，就得到整形后的量化误差的频谱。噪声整形技术的关键是如何选取合适的噪声整形滤波器 $G(z)$，以得到满意的噪声谱。选取的方法很多，这里介绍较常用的三种方法：

(1) 利用人耳的听觉掩蔽效应，使噪声谱的包络形状跟随语音频谱的包络变化，从而使量化噪声的能量集中在信号的高能量区域，如共振峰处。通过语音信号来掩盖噪声，获得更好的主观听觉效果。

(2) 整形噪声谱使其符合人耳的听觉灵敏度曲线，使噪声能量集中在听觉不敏感的区域内。国际标准组织认可的人耳听觉灵敏度曲线如：E-计权曲线、F-计权曲线等。

(3) 对量化噪声进行低频衰减、高频提升，从而把大部分量化噪声转移到信号频带以外，提高量化信号的信噪比。

5.1.6　增量调制和自适应增量调制

增量调制(delta modulation，DM)是 DPCM 的一种特殊形式。根据采样定理，采样频率必须大于奈奎斯特频率。当系统的采样频率大于奈奎斯特频率很多倍时，则相邻采样值之间的相关性会变得非常强，差分信号的幅值会在一个很小的动态范围内变化，这样就可以用正负两个固定的电平来表示差分信号。因此在 DM 中，仅用 1 比特就能量化差分信号，即只需指示极性。所采用的固定电平值被称为量化阶梯，在接收端，用上升下降的阶梯波形来逼近语音信号。

基本的 DM 使用固定的量化阶梯 Δ，当差分信号的幅值大于 Δ 时，量化为 0；小于 $-\Delta$ 时，量化为 1；若差分信号的绝对值小于 Δ，既可取 0 也可取 1，一般应让 0 和 1 交替出现。

如何选取适当的Δ值,要考虑两方面的因素:一方面若Δ值选取的太小,则当语音急剧变化时,重构信号会因不能反映信号的变化而产生斜率过载失真;另一方面,若Δ选取的太大,则当输入信号变化比较平稳时,量化输出将呈现0、1交替的序列,使重构信号围绕着某一固定电平重复增减,产生颗粒噪声。实际上,由于这两方面的因素相互矛盾,很难确定一个适当的Δ值。解决办法是采用自适应技术,实现自适应增量调制(adaptive DM,ADM)。

ADM的基本原理是使Δ值随信号的平均斜率而变化,斜率大时,Δ值自动增大;反之Δ值减小。这样Δ值跟随输入波形自适应的变化,使得斜率过载失真和颗粒噪声都减至最小。ADM一般采用反馈自适应方式,避免发送边信息。

5.1.7 子带编码

以上所介绍的都是基于时域的波形编码技术。下面介绍两种频域编码:子带编码和自适应变换域编码。本节主要介绍子带编码,而自适应变换域编码将在下节中详细介绍。

所谓子带编码(sub-band coding,SBC),就是首先将输入信号分割成几个不同的频带分量,然后再分别进行编码。这种编码方式主要有以下四个优点:

(1) 语音信号的频谱是非平坦的,且对人耳的听觉的贡献也是不均匀的。多数人的语音信号能量主要集中在500Hz～1kHz左右,并随着频率的升高衰减得很快。因此子带编码可以根据不同频段给各子带合理地分配量化字长,使编码速率更精确地与各子带的信源统计特性相匹配。例如可以用较高的比特数使低频带的基音和共振峰保存较高的精度,而对发生在高频带的摩擦音及噪声样值只分配较少的编码比特。

(2) 高频段的子带信号可以通过频谱平移变换成基带信号,然后用相对较低的采样频率进行欠采样后再进行编码。这样编码中各子带信号的采样率显然都远低于原信号的采样率,从而得到较低的编码速率。

(3) 调整不同子带的量化字长,就控制了总的量化噪声的频谱形状,进一步与语音心理-生理模型相结合,可将噪声谱按人耳主观噪声感知特性来成形。

(4) 各子带内的量化噪声都被束缚在本子带内,这样就避免能量较小频带内的输入信号被其他频段的量化噪声所掩盖。

子带编码的工作原理如图5-11所示,首先用一组带通滤波器(BPF)将输入信号频带分割成若干个子频带,然后用调制的办法将这些带通信号经过频谱平移变成基带信号,以利于降低采样率进行抽取(进行欠采样),抽取后的信号按波形编码技术(PCM、ADPCM等)进行编码。最后将各子带的编码数据复接成一个总编码数据发送给接收端。接收端首先通过内插恢复原始的采样率,然后经过频率平移恢复到原来的频段,最后各个频带的分量相加得到重构语音信号。

子带编码中各带通滤波器的宽度可以相同,也可以不同。等带宽子带编码虽然易于用硬件实现,但因为没有考虑人耳的听觉效果,难以获得很好的语音质量。一般情况下都采用不等带宽子带编码,而且按照对主观听觉贡献相等的原则来分配各子带的带宽。同时为了易于实现频谱平移,实际使用时往往采用“整数带”采样方法。所谓整数带,是指子带最低频率为子带带宽的整数倍,这样平移频谱成分时,可以不用调制器而直接实现。如图5-12所示。

子带编码中,重构信号的质量受带通滤波器组的性能影响很大。理想情况下,各子带之

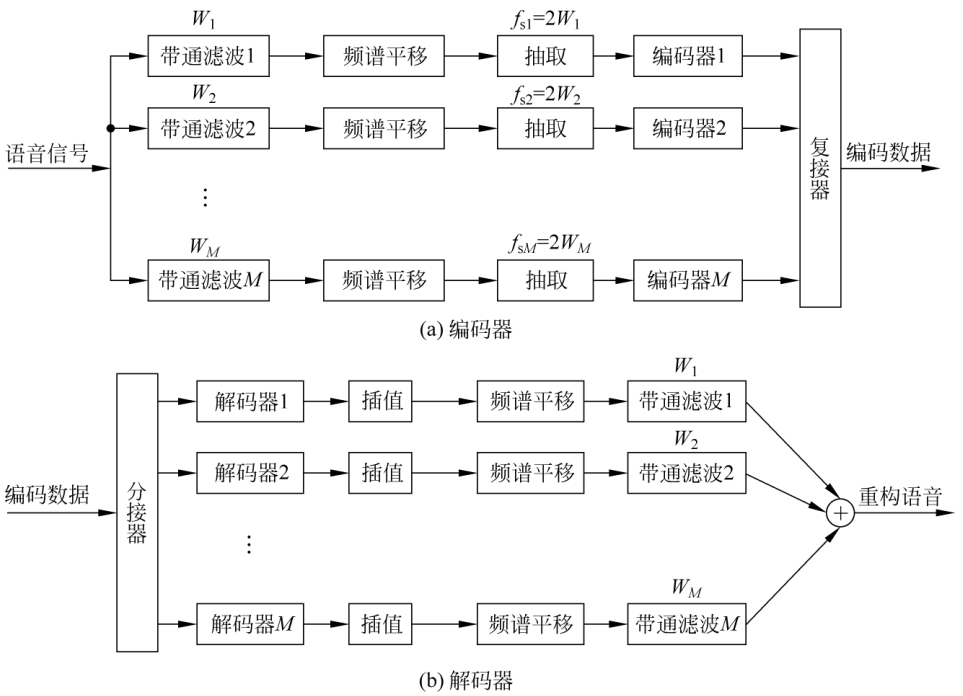


图 5-11　子带编码原理框图

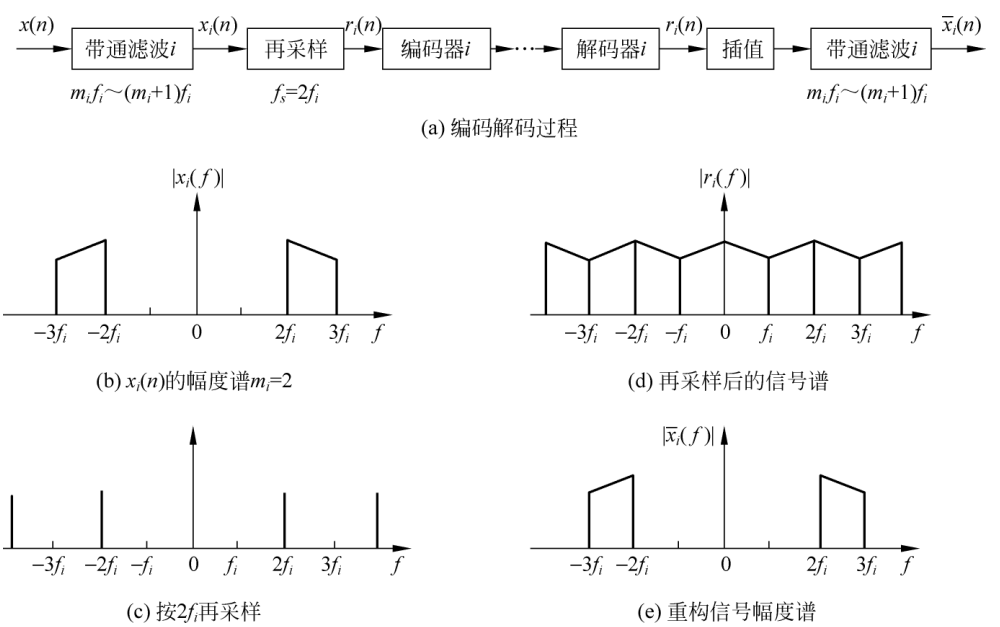


图 5-12　子带编码中的整数带采样方法及频谱的变化

和可以覆盖全部信号带宽，而不重叠。但实际上，数字滤波器的阻带和通带总存在波动，难以得到这种理想情况。如果子带滤波后的各频带重叠太多，将会需要更大的数码率；原来各独立子带的误差也会影响相邻的子带，造成混叠现象。早期的解决方法是让相邻子带间

留有间隙,尽管如此,这些间隙仍会引起输出结果的回声现象。现在多采用正交镜像滤波器(quadrature mirror filter,QMF)技术来解决这一问题。QMF允许编码器分解滤波中的混叠现象,而在解码端通过重构滤波器可以准确无误地消除混叠。

ITU-T制定的G.722标准就是基于SBC的编码器算法,它采用ADPCM技术对抽取后的信号进行编码,该算法将采样频率提高到16kHz,以适应高质量语音应用的场合,例如电话会议或视频会议等。它利用正交镜像滤波器将语音频带分成两个子带,高端子带采用16Kbps的ADPCM进行编码,低端采用48/40/32Kbps的ADPCM编码。因此,G.722可提供3种不同的数码率:64Kbps、56Kbps和48Kbps。

5.1.8 自适应变换域编码

自适应变换域编码(adaptive transform coding,ATC)与SBC一样,在频域上寻找语音的压缩途径。ATC与SBC都是在频域上分割信号的编码方式。

ATC对语音信号进行正交变换,以去除样本间的相关性,变换后的系数将集中在一个较小的范围内,所以对变换系数进行量化编码后,可以实现数码率的压缩。在接收端解码后,可用相应的逆变换重构语音信号。由于进行了正交变换,实际上等同于把时域的语音信号变换到另一个域中去,因此被称为变换域编码。它通过去除语音样本间的相关性,达到了减少语音中冗余信息的目的。

编码时,先将语音信号序列分帧,每帧表示为一个矢量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_N)^{\mathrm{T}}$,然后用正交变换矩阵 $\boldsymbol{A}$ 进行线性变换

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x} \tag{5-16}$$

式中,$\boldsymbol{A}$ 满足 $\boldsymbol{A}^{-1}=\boldsymbol{A}^{\mathrm{T}}$,$\boldsymbol{y}$ 中的元素就是变换域的系数,各元素可以看做是互不相关的,或基本上是互不相关。对其进行量化后得到矢量 $\bar{\boldsymbol{y}}$。在解码端通过逆变换重构出信号矢量 $\bar{\boldsymbol{x}}$ 为

$$\bar{\boldsymbol{x}}=\boldsymbol{A}^{-1}\bar{\boldsymbol{y}}=\boldsymbol{A}^{\mathrm{T}}\bar{\boldsymbol{y}} \tag{5-17}$$

变换域编码的关键是提供一种合适的正交变换。从去除相关性的意义来讲,KL变换(Karhunen-Loeve Transform)是最佳的,但是它需要计算变换矩阵及逆矩阵,不仅计算量大,而且需要传送边信息,很难实际应用。在变换域编码中,最常采用的正交变换是离散余弦变换(discrete cosine transform,DCT),它与KL变换相比,频域的概念比较直观,且与人的听觉频率分析机理相对应,因此容易控制量化噪声的频率范围。从信噪比的角度看,DCT变换比KL变换只相差1~2dB,计算复杂性却小得多。此外,其他正交变换,如快速傅里叶变换FFT、沃尔什—哈达马变换WHT等,因其计算上的优势,也有一定的实用价值。

变换域编码通常是按照各变换分量对语音质量贡献的程度来分配量化字长。在非自适应的情况下,码位分配和量化间隔均根据语音信号长时间统计特性来确定,是固定不变的。而自适应情况下,需要估计每帧变换谱的包络,使用估计的谱值代替方差,再计算出码位的分配。将表征估计谱的参数作为边信息传送到解码端,由解码端使用与编码端相同的步骤计算比特分配,解码变换域参数。

ATC的优劣取决于自适应的效果,即估计谱对语音信号短时DCT谱的逼近程度,因此码位的分配应使估计谱能正确反映变换域系数的能量分布,但是由于估计谱要作为边信息传送,所以它所占的比特数自然要受到一定的限制。在ATC中,谱估计常使用线性预测分析的方法或线性滤波器组的方法。ATC的原理如图5-13所示。

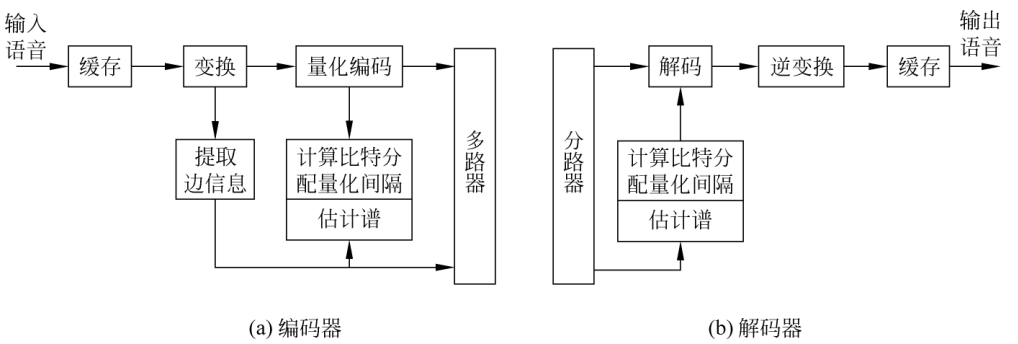


图 5-13　ATC 编码解码工作原理框图

5.2　参数编码和混合编码

参数编码器又称声码器(vocoder),其原理和设计思想与波形编码完全不同。波形编码的基本思路是忠实地再现语音的时域波形,它在 32Kbps 的编码速率下能够得到非常好的话音质量。在话务过载的情况下,还可降质使用 24Kbps 或 16Kbps 编码速率,但要进一步降低比特率就比较困难。因此,使用波形编码方式实现的语音编码器大多属于中高速率的编码器。参数编码根据声音形成机理的分析,着眼于构造语音生成模型,该模型以一定的精度模拟说话人的发音声道,接收端根据该模型还原生成合成语音。编码器发送的主要信息是该模型的参数,相当于语音的主要特征,而不是具体的语音波形的幅值。参数编码器是最早成功应用的语音编码器,它将分析与合成结合起来,实际上是一种语音分析合成系统。因为仅传输模型参数所需要的数据量要小得多,所以参数编码可以实现很低的编码速率,例如,可以达到 2.4Kbps 甚至 2.4Kbps 以下。但是参数编码器也有语音质量差,自然度较低,对环境噪声敏感等缺点。典型的参数编码器有通道声码器、共振峰声码器及线性预测声码器等,其中线性预测声码器目前得到了广泛的应用。

20 世纪 70 年代中期,特别是 20 世纪 80 年代以来,语音编码技术有了突破性的进展,一些非常有效的处理方法被提出,产生了新一代的参数编码算法,也就是混合编码。混合编码克服了参数编码激励形式过于简单的缺点,成功地将波形编码和参数编码两者的优点结合起来,既利用了语音产生模型,通过对模型参数进行编码,减少被编码对象的动态范围和数据量;又使编码过程产生接近原始语音波形的合成语音,以保留说话人的各种自然特征,提高了语音质量。混合编码器在 4～16Kbps 的数码率上能够得到高质量的合成语音。目前比较成功的混合编码器有多脉冲激励线性预测编码(MPLPC)、规则脉冲激励线性预测编码(RPELPC)、码激励线性预测编码(CELP)以及多带激励(MBE)编码等。其中,MPLPC、RPELPC 和 CELP 是基于全极点语音产生模型的混合编码器,而 MBE 是基于正弦模型的混合编码器。

5.2.1　参数编码

参数编码的基础是语音产生的模型,如第 2 章的图 2-18 所示。根据该模型对语音信号进行分析可以得到谱包络、基音周期以及清浊音判别等信息,其中谱包络信息是一组定义声

道共振特性的滤波器系数。如果将上述参数编码后传输到接收端,那么就可以在同样的语音模型的基础上合成语音信号,合成器中所采用声道滤波器的形式与编码端的谱包络分析器的形式相对应,它们的不同形式决定了声码器的不同类型,如通道声码器、共振峰声码器和 LPC 声码器等。

1. 通道声码器

最古老的语音编码装置就是通道声码器,它是基于短时傅里叶变换的语音分析合成系统,发送端通过若干个并联的通道对语音信号进行粗略的频谱估计,而接收端产生一信号,使频谱与发送端规定的频谱相匹配。通道声码器的原理图如图 5-14 所示。

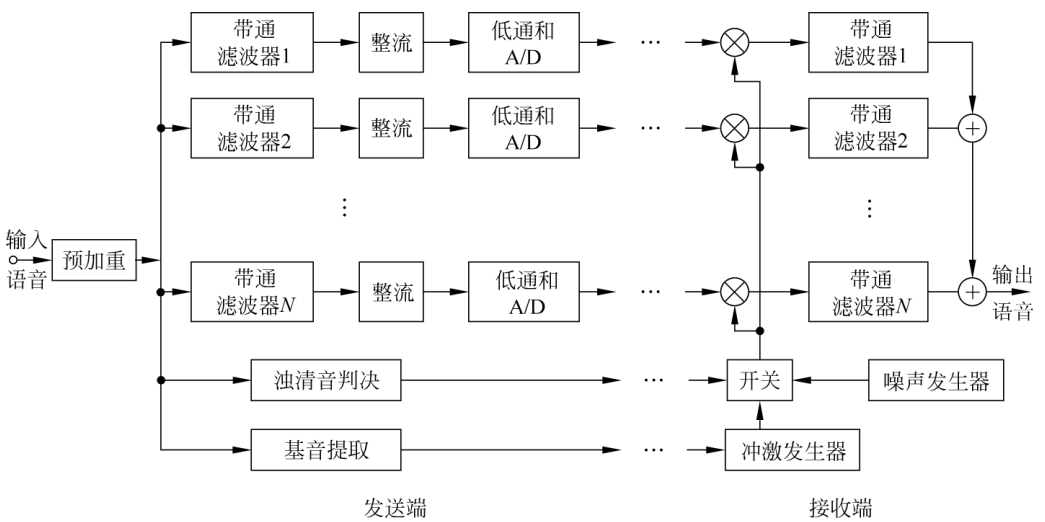


图 5-14 通道声码器原理图

在发送端,输入语音被加于滤波器组和基音提取器上。滤波器组将语音的频率范围分成许多相邻的频带或通道,滤波器的个数典型值为 10～20 个。这种频带的划分并不是均匀的,低频部分带宽较窄,以保证低频段有较高的频率分辨能力。整流电路取出各频段信号幅值,低通滤波器的目的是避免采样后产生混叠失真,同时完成信号的 A/D 转换。每一路通道输出对应频带的幅度谱的均值,这一组数据就反映了信号频谱的包络。将其与清浊音判决信号和基音周期一起编码后传送到接收端。

在接收端,通过清浊音判决信号和基音周期来提供声门激励信号,并用频谱包络信号对其进行调制,经带通滤波器输出后叠加在一起就合成为输出语音信号。

编码器中的预加重模块的作用是按 6dB/倍频程的比例补偿嘴唇辐射衰减,使得各通道输出信号的电平大致相同。相应地,在接收端应设置一个具有－6dB/倍频程衰减的逆滤波器进行去加重。

通道声码器的主要缺点是需要检测基音周期和进行清浊音判决,而精确地求出这两部分数据是相当困难的,其误差会对合成语音的质量造成很大的影响。此外,由于通道数量有限,可能几个谐波分量会落入同一个通道,在合成时它们将被赋予相同的幅度,结果导致频谱畸变。

2. 共振峰声码器

共振峰声码器不是将语音信号划分成多个频段，而是对整体进行分析，提取共振峰的位置、幅度和带宽等参数，构成两个声道滤波器。浊音滤波器采用全极点滤波器，由多个二阶滤波器级联而成；清音滤波器一般采用 1 个极点和 1 个零点的数字滤波器。这些滤波器的参数都是时变的。图 5-15 为共振峰声码器的合成器结构。其中共振峰 F_1、F_2、F_3为浊音滤波器的参数，极点 F_p和零点 F_z为清音滤波器的参数，F_0为基音频率，A_u、A_v为增益系数。

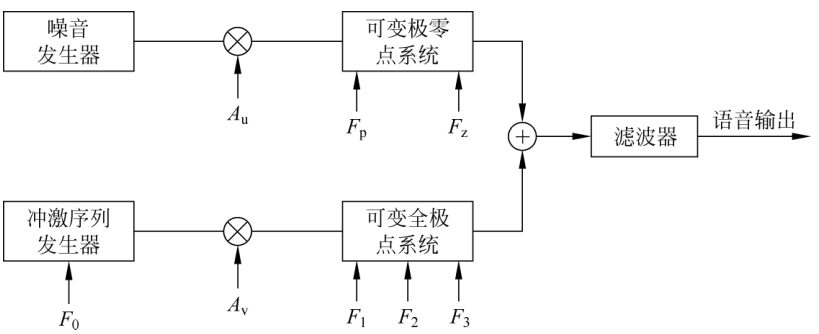


图 5-15　共振峰声码器的合成器结构

与通道声码器相比，共振峰声码器合成出语音的质量更好，比特率可压缩得更低。

3. 线性预测(LPC)声码器

LPC 声码器是应用最成功的低速率语音编码器。它基于全极点声道模型的假定，采用线性预测分析合成原理，对模型参数和激励参数进行编码传输。LPC 声码器遵循二元激励的假设，即浊音语音段采用间隔为基音周期的脉冲序列，清音语音段采用白噪声序列。因此，声码器只需对 LPC 参数、基音周期、增益和清浊信息进行编码。LPC 声码器可以得到很低的比特率(2.4Kbps 以下)。它的工作原理如图 5-16 所示。

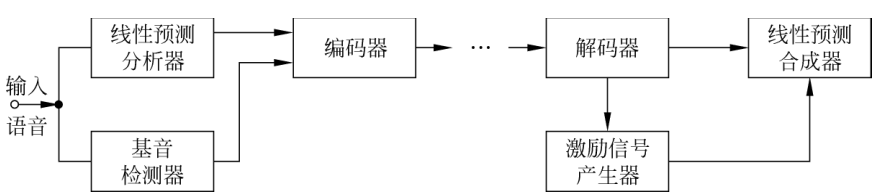


图 5-16　LPC 声码器原理图

虽然 LPC 声码器与 ADPCM 一样，都是基于线性预测分析来实现对语音信号的编码压缩，但是它们之间有本质的区别，LPC 声码器不考虑重建信号波形是否与原来信号的波形相同，而努力使重建信号在主观感觉上与输入语音一致，所以不必量化和传输预测残差，而只需传输 LPC 参数和重构激励信号的基音周期和清浊信息。

如第 4 章所述，LPC 分析存在多种推演参数，选用哪种参数进行编码，需要考虑如下两个因素。

(1) 参数的量化特性：参数的量化特性与参数的谱灵敏度是密切相关的，所谓谱灵敏

度是指参数的微小变化所引起的谱变化的程度。谱灵敏度比较均匀的参数,其量化特性就好,对于一定的谱失真允许范围,参数编码所需要的总比特数就比较小,合成滤波器的稳定性也会比较好。

(2) 参数的内插特性:在编码系统中,常需要将两组 LPC 参数进行线性内插,得到另一组 LPC 参数作为两者之间的过渡,以便使合成语音的频谱特性过渡更加自然平滑。如果参数的编码特性很好,但它内插所得到的参数不代表频谱的平滑过渡,甚至导致合成不稳定的滤波器,这样的参数显然也不适合用于编码传输。现在来比较几种 LPC 参数的编码性能。

1) 线性预测系数$\{a_i\}$

线性预测系数$\{a_i\}$显然不适合作为编码参数,它的谱灵敏度极不均匀,有些系数很小的变化,就可能会引起频谱发生很大的变化。而且线性预测系数的内插特性也很差,内插得到的新参数,不一定能够构成稳定的合成滤波器。

2) 反射系数$\{k_i\}$

用反射系数构成的格型滤波器是一种参数灵敏度较低的合成滤波器,它稳定的充分必要条件是$|k_i|<1$。这一点无论是在对参数进行量化编码时,还是在对参数进行线性内插时都容易保证。因此,反射系数被广泛地应用于语音的编码及合成。但是反射系数的谱灵敏度并不均匀,其绝对值越接近 1,谱灵敏度就越高。因此,采用反射系数进行编码时,一般都采用非线性量化,比特数分配也不是平均分配的。通常 k_1、k_2用 5～6 比特,其他各阶,随阶数增加量化比特数逐渐减少。

3) 对数面积比$\{g_i\}$

对数面积比参数可由下式计算:

$$g_i = \log(A_{i+1}/A_i) = \log[(1-k_i)/(1+k_i)], \quad 1 \leqslant i \leqslant p \tag{5-18}$$

式中,A_i就是多节无损声管中第 i 节的截面积。

由于式(5-18)将域$-1 \leqslant k_i \leqslant 1$映射到$-\infty \leqslant g_i \leqslant +\infty$,它使 g_i 呈现相当均匀的幅度分布,可以进行均匀量化。此外,对数面积比参数各维之间相关性很低,因此能够保证通过线性内插得到的滤波器的稳定性。

4) 预测多项式的根

对预测多项式 $A(z)$做简单的因式分解,有

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} = \prod_{i=1}^{p} (1 - z_i z^{-i}) \tag{5-19}$$

取 $A(z)=0$,即可求得一组根。其中每一对根与信号谱中的一个共振峰相对应。这种参数的优点是容易保证合成滤波器的稳定性。只要让$\{z_i\}$都在单位圆内就可以。其主要缺点是求解多项式的根需要相当大的计算量。

5) 线谱对参数 LSP

线谱对参数 LSP 是量化编码过程中最常用的 LPC 参数,实验证明,其量化特性和内插特性都明显优于其他参数。LSP 的 $P(z)$和 $Q(z)$的根均位于单位圆上,且相互交替间隔排列,利用这一性质,很容易保证合成滤波器的稳定性。LSP 的频谱灵敏度具有很好的频率选择性,单个 LSP 的误差只局限于该频率附近的频谱范围,这种误差相对独立的性质非常有利于 LSP 的量化和内插。

LPC 声码器在通信领域,尤其是军事通信领域得到了广泛的应用。1976 年美国确定用

LPC 声码器标准 LPC-10 作为 2.4Kbps 速率上的推荐编码方式。1981 年这个算法被官方接受，作为联邦政府标准 FS-1015 公布。利用这个算法可以合成清晰、可懂的语音，但是抗噪声能力和自然度尚有欠缺。自 1986 年以来，美国第三代保密电话装置(STU-Ⅲ)采用了速率为 2.4Kbps 的 LPC-10e(LPC-10 的增强型)作为语音处理手段。下面介绍 LPC-10 的工作原理和一些改进措施。

图 5-17 为 LPC-10 的编码器框图。原始语音经过一锐截止的低通滤波器之后，输入 A/D 转换器，以 8kHz 采样率 12 比特量化得到数字化语音，然后每 180 个采样点(22.5ms)为一帧，以帧为处理单元。编码器分两个支路同时进行，其中一个支路用于提取基音周期 D 和清浊音判决信息 V/UV，另一支路用于提取预测系数和增益因子 RMS。提取基音周期的支路把 A/D 变换后输出的数字化语音缓存，经过低通滤波、二阶逆滤波后，再用平均幅度差函数(AMDF)计算基音周期，经过平滑、校正得到该帧的基音周期。与此同时，利用模式匹配技术，基于低带能量、AMDF 函数最大值和最小值之比、过零率进行清/浊音判决，判决结果为以下 4 种状态中的一个：稳定的清音，清音向浊音转换，浊音向清音转换和稳定的浊音。在提取声道参数的支路，先进行预加重处理，然后增益因子 RMS 按如下形式计算：

$$\mathrm{RMS}=\left[\frac{1}{N}\sum_{i=1}^{N}x_i^2\right]^{\frac{1}{2}} \tag{5-20}$$

式中，N 为分析帧长，x_i 为经过预加重后的数字语音。

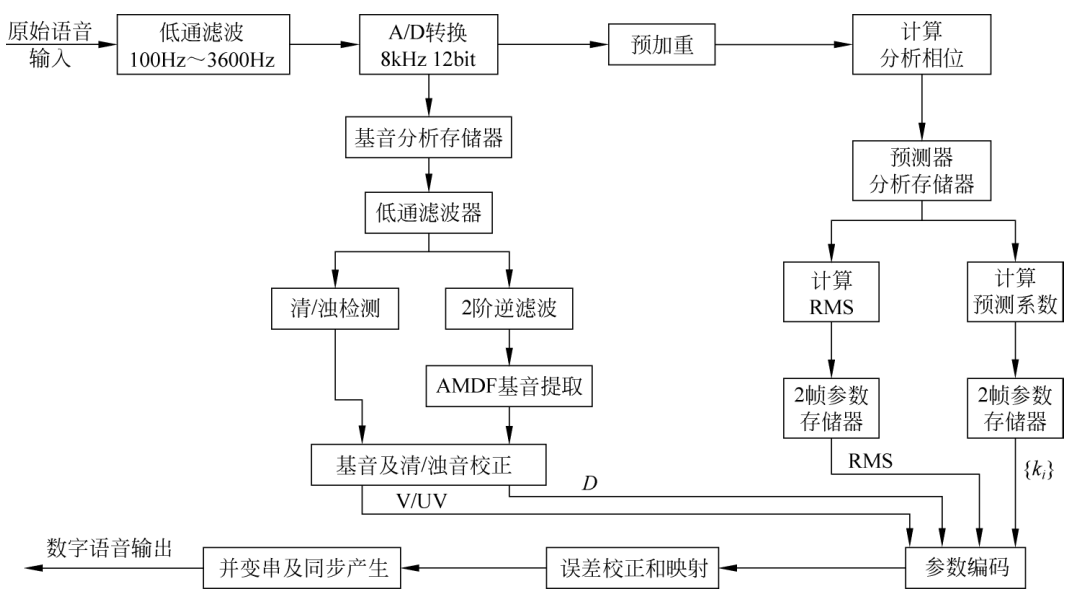


图 5-17　LPC-10 编码器框图

用协方差法求取 10 阶线性预测系数，将线性预测系数转换成反射系数$\{k_i\}$，$i=1,\cdots,10$。前两个反射系数被转化为对数面积比系数后进行量化编码，其余的直接按线性编码。$k_1\sim k_4$每个系数用 5 比特，$k_5\sim k_8$每个系数用 4 比特，k_9为 3 比特，k_{10}为 2 比特，基音周期和清浊判决用 7 比特，增益的对数用 5 比特，再加上同步信息用 1 比特，每帧共计 54 比特，因此总的编码速率为 2.4Kbps。

解码时，首先利用直接查表法对数码流进行检错和纠错。经过纠错解码后得到基音周

期、清浊音标志、增益及反射系数的数值。解码结果延时一帧输出。这样输出的数据可以在过去1帧、现在1帧、将来1帧共3帧内进行平滑,由于每帧语音只传输一组参数,考虑一帧之内可能有不止一个基音周期,因此要对接收数值进行由帧块到基音块的转换和插值,使基音周期、清浊音标志、增益及反射系数等参数值每个基音周期更新一次。在解码器中,根据莱文逊—杜宾递推算法将反射系数$\{k_i\}$变换为线性预测系数$\{a_i\}$,然后用直接型递归滤波器$H(z)=1\Big/\left(1-\sum_{i=1}^{p}a_iz^{-i}\right)$来合成语音。激励采用简单的二元激励,即用随机数来作为清音帧激励源,用周期性冲激序列通过一个全极点滤波器来生成浊音激励源。LPC-10的解码器框图如图5-18所示。

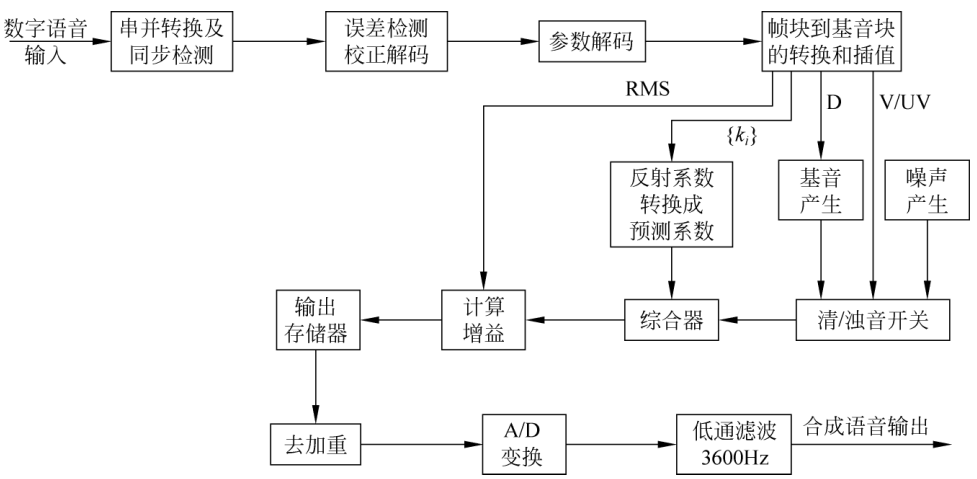


图5-18 LPC-10解码器框图

LPC-10虽然有编码速率低的优点,但是合成语音听起来很不自然,即使提高编码速率也无济于事。这主要是因为清浊音判决和浊音信号的基音检测很难做到十分可靠。有些摩擦音本身就清浊难分,在辅音与元音的过渡段或者有背景噪声的情况下,检测结果就更容易发生错误。这种错误对合成语音的清晰度影响特别严重。此外采用过分简化的二元激励形式,也不符合实际情况,因而造成自然度的下降。在增强型LPC—10e中采用了如下一些措施来改善语音的质量:

1) 激励源的改善

(1) 采用混合激励代替简单的二元激励。此时,浊音的激励源是由经过低通滤波的周期脉冲序列与经过高通滤波的白噪声相加而成的,周期脉冲与噪声的混合比例随输入语音的浊化程度变化。清音的激励源是白噪声加上位置随机的一个正脉冲跟随一个负脉冲的脉冲对形成的爆破脉冲。对于爆破音,脉冲对的幅度增大,与语音的突变成正比。采用混合激励可以使原来二元激励合成引起的金属声、重击声、音调噪声等得到改善。

(2) 采用激励脉冲加抖动(Jittler)的方式。将基音相关性不是很强或残差信号中有大的峰值的语音帧判定为抖动的浊音帧。除采用脉冲加噪声的混合激励外,激励信号中的周期脉冲的相位要做随机地抖动,即对每个基音周期的长度乘上一个0.75～1.25之间均匀分布的随机数,这样可以改善语音的自然度。

(3) 采用单脉冲与码本相结合的激励模式。可取多脉冲激励线性预测编码与码本激励线性预测编码各自的长处,对不同的语音段采用不同的激励模式。对于具有周期性的语音段用以基音周期重复的单脉冲作为激励源,非周期性语音段用从码本中选择的随机序列作为激励源。

2) 改进基音提取方法

计算线性预测残差信号或者语音信号的自相关函数,并利用动态规划的平滑算法来更准确地提取基音周期。将该帧的线性预测残差信号低通滤波后,求出所有可能的基音时延点上的归一化自相关系数,选出其中 L 个最大值,再用相邻 3 帧的每帧 L 个最大值,用动态规划算法求得最佳基音值。

3) 选择线谱对参数 LSP 作为声道滤波器的量化参数

5.2.2　基于全极点语音产生模型的混合编码

经过几十年的研究,人们已经认识到,导致 LPC 声码器性能差的原因不在于声道模型本身,而在于对激励信号的表示过于简化。多年来一直被广泛采用的,使用准周期性脉冲或白噪声作为激励源的方法,是进一步提高语音质量的障碍。基于这种认识,20 世纪 80 年代以来,人们提出了一系列高音质的混合编码算法,例如多脉冲激励线性预测声码器、规则脉冲激励线性预测声码器、码激励线性预测声码器等。这些混合编码算法在保留原有声道模型假定的基础上,引入高质量的波形编码准则来优化激励信号。以感觉加权均方误差最小为判决准则,采用闭环搜索的方法——合成分析法(analysis-by-synthesis,ABS)来选取最佳激励矢量,以得到最佳逼近原始语音的效果。

上文所列举的这三种编码都是基于全极点语音产生模型假定的,编码过程可以简述如下:首先通过线性预测分析方法提取声道滤波器参数;然后通过合成分析的方法确定最佳激励矢量;最后将滤波器参数和最佳激励矢量进行编码传输。有时也将它们统称为基于合成分析法的线性预测编码器(ABS-LPC)。本节首先将这类混合编码实现过程中所采用的主要分析方法做简要介绍,如:语音产生模型、合成分析法、感觉加权均方误差最小准则。然后分别介绍上文所列举的这三种编码算法。

1. 主要分析方法

1) 计入长时相关性的语音产生模型

上一节讨论过语音中有两种类型的相关性,即在样本点之间的短时相关性和相邻基音周期之间的长时相关性。对语音信号用线性预测的方法分别进行这两种相关性的去相关处理后,可以得到更加平坦的预测残差信号,因而更加有利于进行量化编码。对应地,同时考虑这两种相关性的语音产生模型如图 5-19 所示。

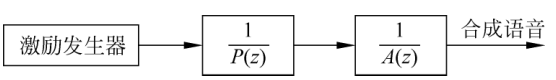


图 5-19　计入长时相关性的语音产生模型

在模型中,激励信号首先输入长时预测综合滤波器 $1/P(z)$,再将其输出作为短时预测综合滤波器 $1/A(z)$的输入,在输出端得到合成语音。

长时预测综合滤波器 $1/P(z)$是表示语音信号长时相关性的模型。它的一般形式为

$$1/P(z)=1\Big/\left[1-\sum_{i=-q}^{r}b_i z^{-(D+i)}\right] \tag{5-21}$$

式中,延时参数 D 等于基音周期,$\{b_i\}$是语音信号的长时预测系数。

通常长时预测系数的个数取在 $1(q=r=0)$到 $3(q=r=1)$之间。前文中的式(5-13)就是一阶预测器的情况。延时参数 D 和系数$\{b_i\}$可以从语音信号中提取,也可以从去除了短时相关性所得到的余量信号中提取。语音信号的长时相关性反映了谱的精细结构。

短时预测综合滤波器 $1/A(z)$与语音信号短时相关的模型相对应,可以用一个全极点模型来描述,它的传输函数 $H(z)$为

$$H(z)=1/A(z)=1\Big/\left[1-\sum_{i=1}^{p}a_i z^{-i}\right] \tag{5-22}$$

式中,$\{a_i\}$是语音信号的短时预测系数;p 是滤波器阶数。

一般称 $H(z)$为线性预测综合滤波器,$A(z)$为线性预测分析滤波器或逆滤波器,同时将 $Q(z)=\sum_{i=1}^{p}a_i z^{-i}$ 称为 p 阶预测器。短时相关性反映了语音信号谱包络信息。

编码时,对语音信号用线性预测分析的方法求取短时和长时预测系数后,构造短时和长时线性预测逆滤波器 $A(z)$和 $P(z)$,并将语音信号输入滤波器 $A(z)$和 $P(z)$,去除信号中的短时、长时相关性,在其输出端就可得到类似于噪声的波形,即线性预测残差信号。虽然在残差信号中浊音段可能还存在若干尖峰脉冲,但是与原语音信号相比要平坦得多,因此,编码时可以得到比较低的编码速率。如果用预测残差信号作为激励信号,则一定可以在语音产生模型上得到无失真的合成语音。但在事实上,从压缩数码率的角度来说,用残差作为激励信号进行语音编码是不现实的。必须采用某种技术,以较低的速率,有效地精确地对预测残差信号进行压缩编码,这也是 ABS-LPC 编码器中的核心问题。

根据具体编码方案的需要,也可以只进行短时预测,不进行长时预测,而在 LPC 激励模型中引入语音的长时相关性。

2) 合成分析法

近几年来,人们在 LPC 算法的基础上,对 16Kbps 以下的高质量语音编码技术进行了广泛深入的研究和实践。在此速率下,能用于残差信号编码的比特数比较少。若对残差信号进行直接的量化,并且使残差信号的量化误差达到最小,并不能保证原始语音与重建语音之间误差最小。必须采用合成分析的方法,以得到的重建语音能够最接近原始语音为目的,闭环搜索残差信号的编码量化值。

基于全极点语音产生模型的语音编解码算法,总是通过解码得到 LPC 系数,以构造综合滤波器,按一定的规则生成激励信号,并将激励信号输入到综合滤波器来合成重构语音。这一功能部件常被称为综合器。而合成分析法将综合滤波器引入到编码器中,使之与分析器相结合,将搜索到的每一残差信号的编码量化值作为激励,通过综合滤波器在编码器中产生与解码器端完全一致的合成语音,将此合成语音与原始语音相比较,按照一定的误差准则计算两者之间的误差,选择使误差最小的参数作为激励编码值。

3) 感觉加权滤波器(perceptually weighted filter)

感觉加权滤波器的依据是人耳的听觉掩蔽效应。在语音频谱中能量较高的频段,即共

振峰处的噪声相对于能量较低频段的噪声而言更不易被感知。因此，在度量原始语音与合成语音之间的误差时可以计入这一因素，在语音能量高的频段，允许两者的误差大一些，反之则小一些。为此可以引入一个频域感觉加权滤波器 $M(\omega)$来计算两者的误差，即

$$e=\int_0^{\omega_s}|x(\omega)-\bar{x}(\omega)|^2M(\omega)\mathrm{d}\omega \tag{5-23}$$

式中，f_s 是采样率，$\omega_s=2\pi f_s$；$x(\omega)$、$\bar{x}(\omega)$分别是原始语音与合成语音的傅里叶变换。

不难证明，为使 e 达到最小值，$|x(\omega)-\bar{x}(\omega)|^2M(\omega)$在整个积分域内应保持常数值。因此，在语音能量较大的语音频段内应使 $M(\omega)$较小，在能量较小的频段内使 $M(\omega)$较大，这就能抬高前者的误差能量，而降低后者的误差能量，为此可取的感觉加权滤波器 $M(\omega)$在 z 域的表达式 $M(z)$为

$$M(z)=\frac{A(z)}{A(z/\gamma)}=\frac{1-\sum_{i=1}^{p}a_iz^{-i}}{1-\sum_{i=1}^{p}a_i\gamma^iz^{-i}} \tag{5-24}$$

感觉加权滤波器的特性由预测系数$\{a_i\}$和加权因子 γ 来确定。γ 取值在 0～1 之间，由它控制共振峰区域误差的增加和减少。以两个极端情况为例，当 $\gamma=1$ 时，$M(z)=1$，此时没有进行感觉加权，当 $\gamma=0$ 时，$M(z)=1-\sum_{i=1}^{p}a_iz^{-i}$，它等于语音的 p 阶全极点模型谱的倒数。由此得到的噪声频谱能量分布与语音频谱的能量分布是一致的。显而易见，$M(z)$的作用就是使实际误差信号的谱不再平坦，而有着与语音信号谱具有相似的包络形状。这就使得误差度量的优化过程与感觉上的共振峰对误差的掩蔽效应相吻合，产生较好的主观听觉效果。实际上取 $\gamma=0$ 时听音效果并不很好，其原因是人耳对语音的共振峰更敏感，相应地对其信噪比要求也更高一些，实际听音的结果表明：在 8kHz 采样频率下，γ 取 0.8 左右较为适宜。将感觉加权滤波器 $M(z)$与滤波器 $H(z)$级联，即获得加权综合滤波器 $H(z/\gamma)$为

$$H(z/\gamma)=H(z)M(z)=\frac{1}{1-\sum_{i=1}^{p}a_iz^{-i}}\cdot\frac{1-\sum_{i=1}^{p}a_iz^{-i}}{1-\sum_{i=1}^{p}a_i\gamma^iz^{-i}}=\frac{1}{1-\sum_{i=1}^{p}a_i\gamma^iz^{-i}} \tag{5-25}$$

随着 γ 的减小，$H(z/\gamma)$的频谱中的各共振峰的带宽相应加大。因此，$H(z/\gamma)$有时又称为频带扩展滤波器或称为误差整形滤波器。若 $H(z)$的冲激响应为 $h(n)$，则 $H(z/\gamma)$的冲激响应为 $\gamma^nh(n)$。

2. 多脉冲激励线性预测声码器

人们对线性预测残差信号进行深入研究后发现，残差信号中的小信号对合成语音的质量影响不大。如果对残差信号进行削波处理，即将幅度低于某一阈值的所有信号皆置为零。这样只要适当调整阈值就可以使残差信号中 90%的样点值为零，用余下的幅度较大的信号作为语音产生模型的激励信号源，其合成语音并未产生明显的畸变。1982 年，Bishnu S. Atal 和 Joel R. Remde 提出了多脉冲线性预测编码（multi-pulse linear predictive coding，MPLPC）方案。在此方案中，首先规定激励脉冲序列在一定的时间间隔中只能出现数目有限的非零脉冲，然后对每个非零脉冲的位置和幅度用合成分析法和感觉加权误差最

小判决准则进行优化；最后用优化的脉冲序列表示残差信号，并作为合成滤波器的激励源。

图5-20为多脉冲激励线性预测声码器的原理框图。在MPLPC中，不再提取基音和进行清浊判决，原始语音信号$x(n)$以帧为单位进行处理，帧长通常取10～20ms。对每帧原始语音，首先采用线性预测分析方法计算出预测系数$\{a_i\}$；然后在当前帧范围内每5ms或10ms用合成分析法估计出一组激励脉冲的幅度和位置，将其输入合成器(虚线框内的部分)得到合成语音$\bar{x}(n)$，再将合成语音$\bar{x}(n)$与原始语音$x(n)$相减并输入感觉加权滤波器$M(z)$，得到加权误差信号$e_m(n)$；最后根据最小均方误差准则，分析估计出一组脉冲位置及幅度最佳的激励脉冲，与线性预测参数一起编码送入信道。

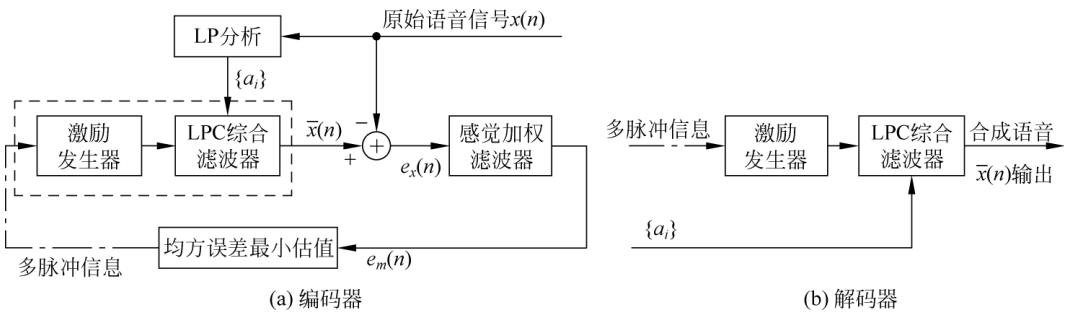


图5-20 多脉冲激励线性预测声码器的原理框图

MPLPC的关键问题是如何求出K个脉冲的位置和幅值，使合成语音与原始语音的感觉加权均方差误差最小。设帧长为N，K个脉冲的位置和幅值分别为$n_1,n_2,\cdots,n_K$和$g_1,g_2,\cdots,g_K$。将这K个脉冲形成的序列作为激励信号输入到LPC综合滤波器$H(z)=\dfrac{1}{1-\sum_{i=1}^{p}a_iz^{-i}}$，得到合成语音$\bar{x}(n)$。当前帧的$\bar{x}(n)$由两部分组成：一部分是LPC综合滤波器的零输入响应$\bar{x}_0(n)$，它是在当前帧不输入激励信号时，用以前各帧所有激励信号在合成器$H(z)$中存储的记忆值在当前帧产生的输出。在做逐帧分析时，当前帧的$\bar{x}_0(n)$为已知；另一部分是LPC综合滤波器$H(z)$的零状态响应，即在当前帧激励信号与$H(z)$的冲激响应$h(n)$的卷积。这样合成语音$\bar{x}(n)$可以表示为

$$\bar{x}(n)=\bar{x}_0(n)+\sum_{k=1}^{K}g_kh(n-n_k) \tag{5-26}$$

合成语音$\bar{x}(n)$与原始语音$x(n)$的误差$e_x(n)$为

$$e_x(n)=x(n)-\bar{x}(n)=x(n)-\bar{x}_0(n)-\sum_{k=1}^{K}g_kh(n-n_k)=\bar{e}(n)-\sum_{k=1}^{K}g_kh(n-n_k) \tag{5-27}$$

式中，$\bar{e}(n)=x(n)-\bar{x}_0(n)$表示输入的原始语音减去零输入响应，即当前帧内除去合成器中由历史记忆造成的输出后的等效语音。将$e_x(n)$输入到感觉加权滤波器$M(z)$，其输出$e_m(n)$为$e_x(n)$和感觉加权滤波器冲激响应$m(n)$的卷积，即

$$e_m(n)=\left[\bar{e}(n)-\sum_{k=1}^{K}g_kh(n-n_k)\right]*m(n)=\bar{e}_m(n)-\sum_{k=1}^{K}g_kh_m(n-n_k) \tag{5-28}$$

式中，$\bar{e}_m(n)$表示原始语音信号中除掉零输入响应的等效语音与 $m(n)$的卷积，$h_m(n)$是加权综合滤波器 $H(z/\gamma)$的冲激响应，感觉加权均方误差 E 为

$$E=\sum_{n=1}^{N}e_m^2(n)=\sum_{n=1}^{N}\left[\bar{e}_m(n)-\sum_{k=1}^{K}g_k h_m(n-n_k)\right]^2 \tag{5-29}$$

激励脉冲的位置与幅度的选择是使 E 最小。为了求取激励脉冲的最佳位置$\{n_k\}$和最佳幅度$\{g_k\}$，对 E 求偏导数，并使之等于 0，因此有

$$\frac{\partial E}{\partial n_j}=0,\quad j=1,\cdots,K \tag{5-30}$$

$$\frac{\partial E}{\partial g_j}=0,\quad j=1,\cdots,K \tag{5-31}$$

这样就能得到 $2K$ 个方程，由式(5-30)得到 K 个非线性方程，而由式(5-31)得到 K 个线性方程，它们是

$$\sum_{k=1}^{k}g_k R_{\mathrm{hh}}(n_k,n_j)=R_{\mathrm{eh}}(n_j),\quad j=1,\cdots,K \tag{5-32}$$

式中：

$$R_{\mathrm{eh}}(n_j)=\sum_{n=1}^{N}\bar{e}_m(n)\cdot h_m(n-n_j) \tag{5-33}$$

$$R_{\mathrm{hh}}(n_k,n_j)=\sum_{k=1}^{k}h_m(n-n_k)\cdot h_m(n-n_j) \tag{5-34}$$

当$\{n_k\}$，$\{g_k\}$($k=1,\cdots,M$)满足上述方程时，将式(5-33)和式(5-34)代入式(5-29)，得到当前帧最小加权均方误差为

$$E_{\min}=\sum_{n=1}^{N}[\bar{e}_m(n)]^2-\sum_{k=1}^{K}g_k R_{\mathrm{eh}}(n_k) \tag{5-35}$$

由于式(5-32)只包含 K 个方程，不可能求出 $2K$ 个未知数，要求出对应于 $E_{\min}$的$\{n_k\}$和$\{g_k\}\{k=1,\cdots,K\}$参数，需要同时解 K 个线性方程和 K 个非线性方程，这一过程是极其复杂的，考虑其实用性，可采用次优搜索算法，即用依次对每个激励脉冲的位置和幅度的顺序优化代替全面搜索的总体优化，这样可以大大简化计算复杂度。这种方法被称为准最优顺序优化激励参数估值法。

设 n_1、g_1 分别是第一个最优激励的位置和幅度，它们满足式(5-32)和式(5-35)，即

$$g_1 R_{\mathrm{hh}}(n_1,n_1)=R_{\mathrm{eh}}(n_1) \tag{5-36}$$

$$E_{\min}=\sum_{n=1}^{N}[\bar{e}_m(n)]^2-g_1 R_{\mathrm{eh}}(n_1) \tag{5-37}$$

将式(5-36)代入式(5-37)可得

$$E_{\min}=\sum_{n=1}^{N}[\bar{e}_m(n)]^2-\frac{R_{\mathrm{eh}}^2(n_1)}{R_{\mathrm{hh}}(n_1,n_1)} \tag{5-38}$$

由于$\bar{e}_m(n)$为固定的已知数，要在当前帧内搜索第一个激励脉冲的最佳位置 n_1，只要搜索到 $E_{\min}$，即只要搜索到使下式取得最大值的 n_1 即可。

$$\frac{R_{\mathrm{eh}}^2(n_1)}{R_{\mathrm{hh}}(n_1,n_1)} \tag{5-39}$$

然后再确定最佳幅度 g_1，即

$$g_1 = \frac{R_{eh}(n_1)}{R_{hh}(n_1, n_1)} \tag{5-40}$$

如果已逐个找到 $j-1$ 个激励脉冲的最优位置和幅度,现要找第 j 个激励脉冲的最优位置 n_j 和最佳幅值 g_j,它应满足式(5-36)和式(5-37),即

$$g_j R_{hh}(n_j, n_j) = R_{eh}(n_j) \tag{5-41}$$

$$E_{min} = \sum_{n=1}^{N} [\bar{e}_{m,j}(n)]^2 - g_j R_{eh}(n_j) \tag{5-42}$$

同样,将式(5-41)代入式(5-42)可得:

$$E_{min} = \sum_{n=1}^{N} [\bar{e}_{m,j}(n)]^2 - \frac{R_{eh}^2(n_j)}{R_{hh}(n_j, n_j)} \tag{5-43}$$

式中,$\bar{e}_{m,j}(n)$表示在输入的原始语音中,扣除了第 j 个以前的所有激励脉冲所产生合成语音的份额后的结果。起始条件为 $g_0=0$,$\bar{e}_{m,0}(n)=x_m(n)-\bar{x}_{m,0}(n)$,$\bar{x}_{m,0}(n)$是当前帧还未搜索出任何激励脉冲时,以前所有激励信号影响下所产生的 $H(z/\gamma)$的输出。在搜索第 j 个激励脉冲时,$\bar{e}_{m,j}(n)$是已知的。在顺序求各个激励脉冲时,它由下面的迭代公式更新:

$$\bar{e}_{m,j}(n) = \bar{e}_{m,j-1}(n) - g_{j-1} h_m(n - n_{j-1}), \quad j = 1, \cdots, K \tag{5-44}$$

式中,n_{j-1}和 g_{j-1}分别是在第 $j-1$ 次搜索中得到的第 $j-1$ 个激励脉冲的最优位置和最优幅值。相应地在每次搜索中 $R_{eh}(n_j)$的更新公式为

$$R_{eh}(n_j) = \sum_{n=1}^{N} \bar{e}_{m,j}(n) h_m(n - n_j) \tag{5-45}$$

由于$\bar{e}_{m,j}(n)$为固定的已知数,要在当前帧内搜索第 j 个激励脉冲的最优位置 n_j,只要搜索到 E_{min},即只要搜索到下式取最大值时的 n_j即可:

$$\frac{R_{eh}^2(n_j)}{R_{hh}(n_j, n_j)} \tag{5-46}$$

然后再确定最佳幅度 g_j:

$$g_j = \frac{R_{eh}(n_j)}{R_{hh}(n_j, n_j)} \tag{5-47}$$

在此搜索方案中,对于一帧内 K 个激励脉冲需要做 K 次搜索迭代,虽然可以方便地得到多脉冲激励中脉冲较优的位置和幅度,但它不是全局最优的,因此估值中会出现一些问题,应采取相应的措施来避免或克服。

MPLPC 合成的语音有较好的自然度,这种编码方法能保证一定的抗噪能力。但其最大的缺点是,即使采取了准最优顺序优化激励参数估值方法,分析时的运算量仍然很大,这使它难以实时实现,因此也很难推广应用。目前还没有见到采用这种算法的商用声码器或标准。

3. 规则脉冲激励线性预测声码器

规则脉冲激励线性预测声码器(regular pulse excitation linear predictive coding,RPELPC)是由 Ed. F. Deprettere 和 Peter Kroon 在 1985 年提出的,其编码思想与 MPLPC 很相似,但更实用。RPELPC 用一组间距一定的非零规则脉冲代替残差信号,该脉冲序列的相位(即第一个非零脉冲出现的位置)和每个非零脉冲的幅度可以按照 MPLPC 同样的方法进行优化。因为各个非零脉冲的相互位置是固定的,所以它的计算量和编码速率与

MPLPC 相比都要小得多。图 5-21 为规则脉冲激励线性预测声码器的原理框图。

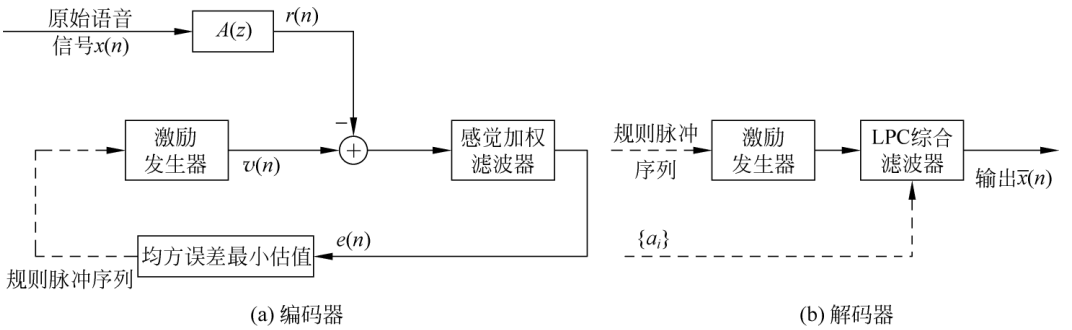


图 5-21　规则脉冲激励线性预测声码器的原理框图

语音信号首先经过 p 阶 LPC 逆滤波器 $A(z)$之后得到残差信号 $r(n)$，将 $r(n)$和激励信号 $v(n)$的差输入到感觉加权滤波器，可知滤波器的输出就应该是感觉加权误差 $e(n)$。通过调整激励信号 $v(n)$，可以使 $e(n)$在一定范围内取得平方和最小。

编码时将一帧语音激励信号分为若干个子帧，用 L 表示激励子帧的长度。8kHz 采样率时，L 的典型值是 40 个样点，相当于 5ms。在每个激励子帧内，都采用间隔相同的规则脉冲串作为激励信号。当脉冲间隔确定时，脉冲串所能采用的模式种类就应该是确定的，规则脉冲串的模式按照脉冲串的相位，即第一个非零脉冲出现的位置来区分。当脉冲间隔为 $R-1$ 个样点时，脉冲串的模式最多为 R 种。同理串中非零脉冲的数量 Q 也可以确定 $Q=L/R$。一种规则脉冲串的模式可以由位置脉冲矩阵$(Q\times L)$来表示，设 $\boldsymbol{B}_k$是相位为 k 的规则脉冲序列的位置脉冲矩阵，矩阵元素 b_{ij}^k可表示为

$$b_{ij}^k=\begin{cases}1, & j=i\times R+k;\ i=0,1,\cdots,(Q-1)\\ 0, & j\neq i\times R+k;\ j=0,1,\cdots,(L-1)\end{cases} \tag{5-48}$$

而在相位为 k 的规则脉冲序列中，Q 个非零脉冲的幅度可用行矢量 $\boldsymbol{g}^{(k)}$表示为

$$\boldsymbol{g}^{(k)}=[g^{(k)}(0),g^{(k)}(1),\cdots,g^{(k)}(Q-1)] \tag{5-49}$$

将一个子帧的激励信号表示为一个矢量，每一个采样点为矢量中的一维。则 L 维激励矢量 $\boldsymbol{v}^{(k)}$可表示为

$$\boldsymbol{v}^{(k)}=\boldsymbol{g}^{(k)}\cdot\boldsymbol{B}_k \tag{5-50}$$

设 $\boldsymbol{M}$ 是感觉加权滤波器 $M(z)$的冲激响应矩阵，这是一个 $L\times L$ 的上三角矩阵。它的第 j 行由 $M(z)$对单位冲激 $\delta(n-j)$的响应取前 $L-j$ 项组成，$j=0,1,\cdots,(L-1)$。$\boldsymbol{M}$ 矩阵为

$$\boldsymbol{M}=\begin{bmatrix}m(0) & m(1) & \cdots & m(L-1)\\ 0 & m(0) & \cdots & m(L-2)\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & m(0)\end{bmatrix} \tag{5-51}$$

如果用 $\boldsymbol{e}_0$ 表示 $M(z)$的零输入响应矢量，$\boldsymbol{r}$ 表示当前激励子帧的线性预测残差信号 $r(n)$形成的矢量，将 $\boldsymbol{r}$ 与第 k 个相位激励矢量 $\boldsymbol{v}^{(k)}$的差输入到感觉加权滤波器 $M(z)$，得到相应的输出感觉加权误差 $\boldsymbol{e}^{(k)}$为

$$\boldsymbol{e}^{(k)} = \boldsymbol{e}^{(0)} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k, \quad k = 0,1,\cdots,R-1 \tag{5-52}$$

式中:

$$\boldsymbol{e}^{(0)} = \boldsymbol{r} \cdot \boldsymbol{M} + \boldsymbol{e}_0 \tag{5-53}$$

$$\boldsymbol{M}_k = \boldsymbol{B}_k \cdot \boldsymbol{M} \tag{5-54}$$

优化过程的第一步就是求 $\boldsymbol{g}^{(k)}$,使 $\boldsymbol{e}^{(k)}$ 中各分量的平方和 $E^{(k)}$ 最小。$E^{(k)}$ 可表示为

$$E^{(k)} = \boldsymbol{e}^{(k)}\boldsymbol{e}^{(k)\mathrm{T}} \tag{5-55}$$

下面首先要解决的问题是:当 L、Q 和 k 都固定时,优化激励脉冲非零值的幅度使 $E^{(k)}$ 最小,将式(5-52)代入式(5-55)并展开,有

$$\begin{aligned} E^{(k)} &= [\boldsymbol{e}^{(0)} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k][\boldsymbol{e}^{(0)} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k]^{\mathrm{T}} \\ &= \boldsymbol{e}^{(0)}\boldsymbol{e}^{(0)T} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k\boldsymbol{e}^{(0)\mathrm{T}} - \boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}\boldsymbol{g}^{(k)\mathrm{T}} + \boldsymbol{g}^{(k)}\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}}\boldsymbol{g}^{(k)\mathrm{T}} \end{aligned} \tag{5-56}$$

为求幅度矢量 $\boldsymbol{g}^{(k)}$ 中的第 i 个分量的最佳幅度 $g^{(k)}(i)$,$(i=0,\cdots,Q-1)$,将式(5-56)两边对 $g^{(k)}(i)$ 求导,得

$$\begin{aligned} \frac{\partial E^{(k)}}{\partial g^{(k)}(i)} &= -\frac{\partial \boldsymbol{g}^{(k)}}{\partial g^{(k)}(i)}\boldsymbol{M}_k\boldsymbol{e}^{(0)\mathrm{T}} - \boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} + \frac{\partial \boldsymbol{g}^{(k)}}{\partial g^{(k)}(i)}\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}}\boldsymbol{g}^{(k)\mathrm{T}} + \boldsymbol{g}^{(k)}\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} \\ &= -\frac{\partial \boldsymbol{g}^{(k)}}{\partial g^{(k)}(i)}\boldsymbol{M}_k(\boldsymbol{e}^{(0)} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k)^{\mathrm{T}} - [\boldsymbol{e}^{(0)} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k]\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} \end{aligned} \tag{5-57}$$

将式(5-52)代入上式得到

$$\begin{aligned} \frac{\partial E^{(k)}}{\partial g^{(k)}(i)} &= -\frac{\partial \boldsymbol{g}^{(k)}}{\partial g^{(k)}(i)}\boldsymbol{M}_k\boldsymbol{e}^{(k)\mathrm{T}} - \boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} \\ &= -\left[\boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)}\right]^{\mathrm{T}} - \boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} \end{aligned} \tag{5-58}$$

令 $\dfrac{\partial E^{(k)}}{\partial g^{(k)}(i)} = 0$,$(i=0,\cdots,Q-1)$,则有

$$\left[\boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)}\right]^{\mathrm{T}} + \boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} = 0 \tag{5-59}$$

由于 $\boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\dfrac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)}$ 是一个标量,式(5-59)可写为

$$2\boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}}\frac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} = 0 \tag{5-60}$$

考虑到 $\dfrac{\partial \boldsymbol{g}^{(k)\mathrm{T}}}{\partial g^{(k)}(i)} = [0,\cdots,0,\underset{\text{第}i\text{位}}{1},0,\cdots,0]^{\mathrm{T}}$,因此有

$$\boldsymbol{e}^{(k)}\boldsymbol{M}_k^{\mathrm{T}} = 0 \tag{5-61}$$

将式(5-52)代入式(5-61),得到

$$[\boldsymbol{e}^{(0)} - \boldsymbol{g}^{(k)}\boldsymbol{M}_k]\boldsymbol{M}_k^{\mathrm{T}} = 0 \tag{5-62}$$

当 $\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}}$ 可逆时,得到相位为 k 的激励脉冲序列的最佳激励幅度矢量 $\boldsymbol{g}^{(k)}$

$$\boldsymbol{g}^{(k)} = \boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}(\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}})^{-1} \tag{5-63}$$

将式(5-63)代入式(5-56),求出相位为 k 的序列的最佳激励矢量 $\boldsymbol{v}^{(k)}$ 引起的误差 $E^{(k)}$

$$\begin{aligned} E^{(k)} &= \boldsymbol{e}^{(0)}\boldsymbol{e}^{(0)\mathrm{T}} - \boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}(\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}})^{-1}\boldsymbol{M}_k\boldsymbol{e}^{(0)\mathrm{T}} - \boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}[\boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}(\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}})^{-1}]^{\mathrm{T}} \\ &\quad + \boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}(\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}})^{-1}\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}}[\boldsymbol{e}^{(0)}\boldsymbol{M}_k^{\mathrm{T}}(\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}})^{-1}]^{\mathrm{T}} \\ &= \boldsymbol{e}^{(0)}[\boldsymbol{I} - \boldsymbol{M}_k^{\mathrm{T}}(\boldsymbol{M}_k\boldsymbol{M}_k^{\mathrm{T}})^{-1}\boldsymbol{M}_k]\boldsymbol{e}^{(0)\mathrm{T}} \end{aligned} \tag{5-64}$$

使 $E^{(k)}$ 最小的 k 就是最佳激励的模式号,它所对应的激励信号 $\boldsymbol{v}^{(k)}$ 就是最佳激励信号。$\boldsymbol{v}^{(k)}$

是由式(5-50)计算出来的。

从上述过程可以看出，最佳激励信号 $\boldsymbol{v}^{(k)}$ 是由相位信息 k 和幅度矢量 $\boldsymbol{g}^{(k)}$ 决定的，如式(5-63)所示，整个过程包含了 R 个线性方程组的求解，这种线性方程组有多种快速的解法，因此，RPELPC 的计算复杂度要比 MPLPC 小得多。

RPELPC 算法也可以增加长时预测机制来改善算法性能。一种比特率为 13Kbps 的长时预测 RPELPC 算法，已被欧洲电信标准协会(ETSI)的全球移动通信(GSM)分会定为其第一个 TDMA 数字蜂窝电话标准。

4. 码激励线性预测声码器

MPLPC 算法和 REPLPC 算法虽然克服了基音检测和清浊判决不精确导致的编码质量下降的问题，但是这两种算法表示激励脉冲所需要的比特数很难进一步压缩，当总的数码率低于 8Kbps 时，语音质量急剧下降。这就使其应用范围受到很大限制。1985 年，Manfred R. Schroeder 和 Bishnus S. Atal 提出了用矢量量化(VQ)技术对激励信号进行编码，VQ 码本中每一个存储的码字矢量都可以代替残差信号作为可能的激励信号源。在编码时对码本中码矢量逐个搜索，找到能产生与输入语音误差最小的合成语音的激励码矢量。只要将该码矢量的标号传送给接收端，在接收端用储存的同样的码本，就能根据收到的标号找到相应的码矢量作为激励。将这样的编码系统，称为码激励线性预测编码(code excited linear predictive coding，CELP)。CELP 在 4.8～16Kbps 的范围内可以获得质量相当高的合成语音，并且抗噪性能和多次转接的性能也很好。

CELP 采用分帧技术进行编码，帧长一般为 20～30ms，将每一语音帧分成 2～5 个子帧，在每个子帧内搜索最佳的码矢量作为激励信号。图 5-22 为 CELP 编码示意图。图中虚线框内是 CELP 综合器，它也是 CELP 解码器中的最主要功能部件。

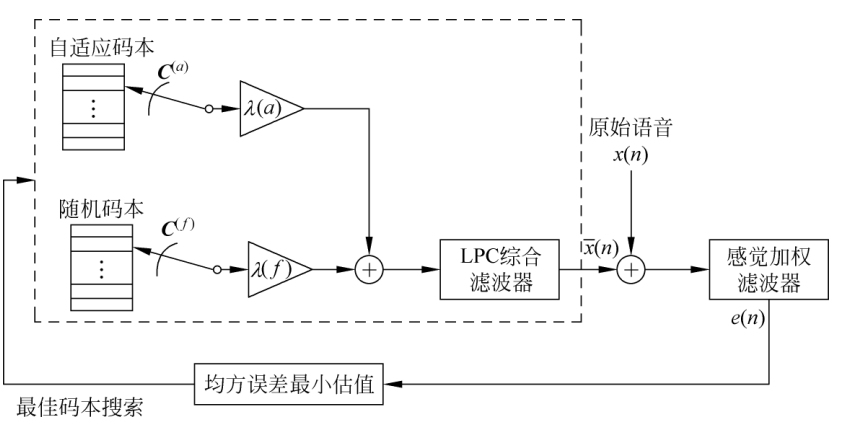


图 5-22　CELP 编码器示意图

CELP 一般都采用分阶段量化的方法将码本划分成两个，一个称为自适应码本，其码矢量逼近语音的长时周期性(基音)结构。另一个称为固定码本，其矢量为随机激励，对应语音经过短时预测和长时预测后的残差信号。当生成激励信号时，首先搜索确定自适应码本矢量，然后再搜索确定固定码本矢量。在搜索固定码本时，必须考虑自适应码本矢量的响应分量。两个码本矢量乘以各自的最佳增益后相加，其和就是 CELP 激励信号源。由于两个码

本的尺寸远小于未采用基音预测(自适应码本)的单码本尺寸,因此搜索效率将大大提高。将激励信号输入 p 阶线性预测综合滤波器 $1/A(z)$,得到合成语音信号 $\bar{x}(n)$,再将 $\bar{x}(n)$ 与原始语音 $x(n)$ 的误差经过感觉加权滤波器 $M(z)$,得到感觉加权误差 $e(n)$。CELP 用感觉加权的最小均方预测误差作为搜索最佳码矢量及其幅度的度量准则,使感觉加权误差的平方和最小的码矢量即是最佳码矢量。

设一个子帧内的信号为一个矢量,则输入语音矢量可表示为 $\boldsymbol{x}=[x(0),x(1),\cdots,x(L-1)]^{\mathrm{T}}$,激励矢量表示为 $\boldsymbol{e}=[e(0),e(1),\cdots,e(L-1)]^{\mathrm{T}}$,$L$ 为子帧的长度。

令 $\boldsymbol{C}_r^{(a)}$ 为标号为 r 的自适应码矢量,相应的增益因子为 $\lambda^{(a)}$;$\boldsymbol{C}_q^{(f)}$ 为标号为 q 的固定码矢量,相应的增益因子为 $\lambda^{(f)}$。则激励信号可表示为

$$\boldsymbol{e}_{rq}=\lambda^{(a)}\boldsymbol{C}_r^{(a)}+\lambda^{(f)}\boldsymbol{C}_q^{(f)} \tag{5-65}$$

搜索自适应码本,对所有的矢量计算其重构信号,每个矢量必须在同样的初始状态,即同样的零输入响应下输入线性预测合成滤波器。记 $\bar{\boldsymbol{x}}_r$ 为当激励输入是 $\boldsymbol{C}_r^{(a)}$ 时滤波器的合成信号,$\bar{\boldsymbol{x}}_0$ 为滤波器的零输入响应,则有

$$\bar{\boldsymbol{x}}_r=\lambda^{(a)}\boldsymbol{M}\boldsymbol{C}_r^{(a)}+\bar{\boldsymbol{x}}_0 \tag{5-66}$$

式中,$\boldsymbol{M}$ 是感觉加权滤波器 $M(z)$ 的冲激响应矩阵,如式(5-51)所示。则原信号和合成重构信号之均方差 $E_r^{(a)}$ 为

$$E_r^{(a)}=|\boldsymbol{x}-\bar{\boldsymbol{x}}_r|^2=\lambda^{(a)2}\boldsymbol{C}_r^{(a)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{C}_r^{(a)}-2\lambda^{(a)}\boldsymbol{C}_r^{(a)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}}_0)+|\boldsymbol{x}-\bar{\boldsymbol{x}}_0|^2 \tag{5-67}$$

对于给定的 $\boldsymbol{C}_r^{(a)}$,求最优增益 $\bar{\lambda}^{(a)}$,使 $E_r^{(a)}$ 为最小,应有 $\dfrac{\partial E_r^{(a)}}{\partial\lambda^{(a)}}=0$。由此得

$$\bar{\lambda}^{(a)}=\frac{\boldsymbol{C}_r^{(a)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}}_0)}{\boldsymbol{C}_r^{(a)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{C}_r^{(a)}} \tag{5-68}$$

将式(5-68)代入到式(5-67)中,并忽略常数项,得到误差判据:

$$E_r^{(a)}=\frac{[\boldsymbol{C}_r^{(a)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}}_0)]^2}{\boldsymbol{C}_r^{(a)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{C}_r^{(a)}} \tag{5-69}$$

对每一个自适应码矢量 $\boldsymbol{C}_r^{(a)}$ 按式(5-71)计算 $E_r^{(a)}$,选择使 $E_r^{(a)}$ 最小的 $\bar{\boldsymbol{C}}_r^{(a)}$ 作为激励信号中的自适应分量。显而易见,$\boldsymbol{x}-\bar{\boldsymbol{x}}_0$ 是自适应码本搜索过程中的目标矢量。

按同样的方法搜索固定码本,求得激励信号的固定码本分量。这时需要考察 $\bar{\boldsymbol{C}}_r^{(a)}$ 的响应分量 $\bar{\boldsymbol{x}}_r$,在固定码本搜索时依据下式进行计算,即

$$E_q^{(f)}=\frac{[\boldsymbol{C}_q^{(f)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}}_r)]^2}{\boldsymbol{C}_q^{(f)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{C}_q^{(f)}} \tag{5-70}$$

选择使 $E_q^{(f)}$ 最小的 $\bar{\boldsymbol{C}}_q^{(f)}$ 作为激励信号中的固定分量,可以看出,此时目标矢量变为 $\boldsymbol{x}-\bar{\boldsymbol{x}}_r$。

最佳码本矢量选定后,将 $\bar{\boldsymbol{C}}_r^{(a)}$ 代入式(5-68)计算最佳增益因子 $\bar{\lambda}^{(a)}$,同理可根据下式来计算 $\bar{\lambda}^{(f)}$。

$$\bar{\lambda}^{(f)}=\frac{\bar{\boldsymbol{C}}_q^{(f)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}}_r)}{\bar{\boldsymbol{C}}_q^{(f)\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\bar{\boldsymbol{C}}_q^{(f)}} \tag{5-71}$$

然后对两个增益因子 $\bar{\lambda}^{(a)}$ 和 $\bar{\lambda}^{(f)}$ 进行量化,自适应码本增益约需 3～4 比特,固定码本增益约需 4～5 比特。

CELP 解码器一般由两部分组成:综合器和后置滤波器滤波。综合器生成的合成语音一般还要经过后置滤波器滤波,以达到去除噪声和提高音质的目的。CELP 解码器的示意

图这里就不再给出。

在 CELP 的解码器中，解码操作也是按子帧进行的。首先对编码中的索引值执行查表操作，从激励码本中抽取对应的码矢量，通过相应的增益控制单元和合成滤波器生成合成语音，而合成滤波器系数和增益按照与编码器同样的方式定期更新。但是这样得到的重构信号往往仍旧包含可闻噪声，在低数码率编码的情况下尤其如此。为了降低噪声，同时又不降低语音质量，一般在解码器中要加入后置滤波器，它能够在听觉不敏感的频域对噪声进行选择性抑制。后置滤波既包括短时后置滤波，也包括长时后置滤波。其传输函数表示为

$$H(z) = GH_{\mathrm{L}}(z)H_{\mathrm{S}}(z) \tag{5-72}$$

式中，$H_{\mathrm{S}}(z)$和 $H_{\mathrm{L}}(z)$分别为短时和长时后置滤波器，G 为后置滤波增益控制因子。当然，后置滤波中也可以不包括长时部分，但加入长时后置滤波确实能够明显改善浊音段合成语音质量。

短时相关后置滤波器传递函数一般表示为

$$H_{\mathrm{S}}(z) = \frac{A(z/\alpha_1)}{A(z/\alpha_2)}(1-\mu z^{-1}) \tag{5-73}$$

参数 α_1 和 α_2 控制滤波器的频率响应，μ 为频谱斜率补偿因子，其作用是补偿由于后置滤波器扩展峰谷距离引起的频谱变化。μ 值可作为输入信号频谱的函数自适应调整，即

$$\mu = C\frac{r(1)}{r(0)} \tag{5-74}$$

式中，$r(1)/r(0)$为语音信号时延为 1 的归一化自相关函数，常数 C 用于限制 μ 的取值范围，典型值为 0.5。

长时相关后置滤波器的作用是增加浊音信号的周期性，其传递函数的一般表示式为

$$H_{\mathrm{L}}(z) = \frac{1+\lambda_1 z^{-D}}{1-\lambda_2 z^{-D}} \tag{5-75}$$

式中，λ_1、λ_2 为系统参数，D 为基音周期。常用的 $H_{\mathrm{L}}(z)$只含分子，即 $\lambda_2=0$，λ_1 为时延为 D 的归一化自相关系数，即

$$\lambda_1 = 0.5\frac{r(D)}{r(0)} \tag{5-76}$$

此时长时相关后置滤波器呈现为全零点滤波器。之所以采用全零点而不是全极点滤波的原因是，全零点滤波器能够反映波形快速变化的特性，能再生具有高度周期性的重构信号。

式(5-72)中增益因子 G 的作用是保证经后置滤波处理后的信号的能量和输入信号相同。由于滤波器本身是时变的，因此增益因子也需自适应调整。最常用的方法是取

$$G = \sqrt{\frac{\sum_n y_1^2(n)}{\sum_n y_2^2(n)}} \tag{5-77}$$

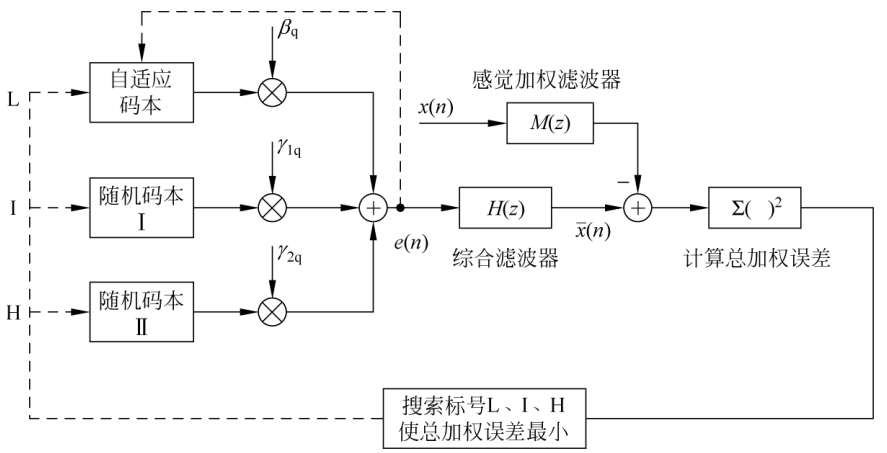
式中，$y_1(n)$和 $y_2(n)$分别为后置滤波前和后置滤波后的语音信号。

后置滤波器可以根据接收到的短时和长时预测系数导出，也可以通过线性预测分析的方法从解码后的语音信号中导出。

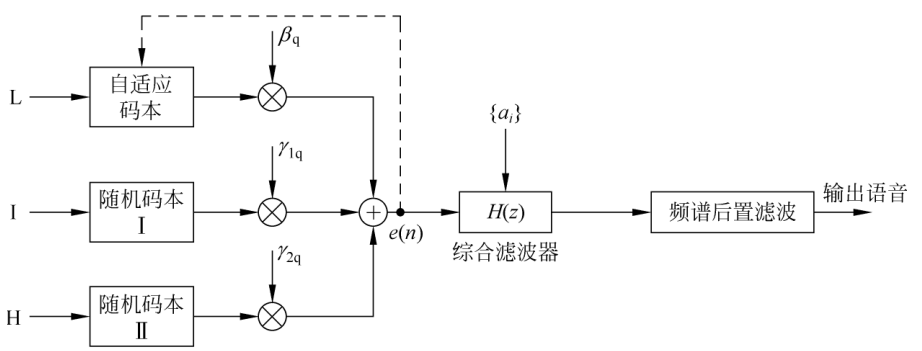
CELP 是 ABS-LPC 中最重要的形式，至今仍然是声码器研究中的热点之一。十几年来，减少 CELP 复杂度、增强 CELP 性能的新技术不断出现。下面简要介绍其中几种重要的方法。

1) 矢量和激励线性预测(VSELP)编码

VSELP 与 CELP 的基本区别在于激励序列形成的方法。如图 5-23 所示,VSELP 有 3 个激励源。一个激励源来自于基音(长时)预测器的状态,即自适应码本。另外两个分别来自于具有 128 个码字的结构化随机码本。3 个激励源的输出分别乘以各自的增益,然后相加得到最终的激励序列。其中 LPC 合成滤波器由具有 10 个极点的滤波器构成,分析帧长 20ms。在合成端,通过内插,激励参数和 LPC 预测系数每 5ms 更新一次。



(a) 编码器框图



(b) 解码器框图

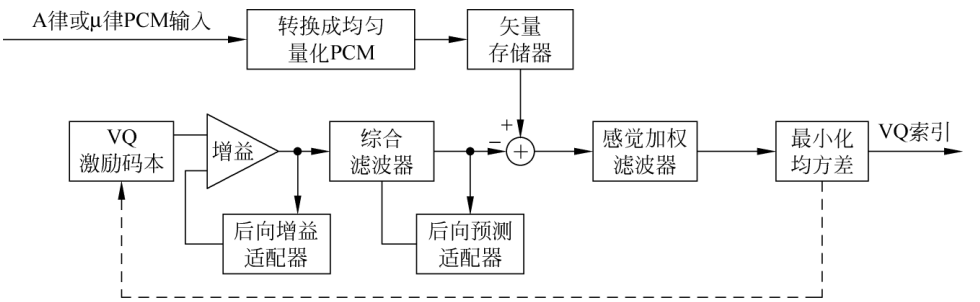
图 5-23 VSELP 编码器/解码器原理框图

VSELP 是一个比较理想的 CELP 改进形式,它保留了 CELP 高效编码的优点,同时又使运算量大大降低。两个随机码本可在保持一定的复杂度下提高语音质量。而结构化码本不仅减少了运算量,也增强了抗信道误码的能力。1989 年 8Kbps 的 VSELP 已被美国电子工业协会(EIA)下属的电信工业协会(TIA)选为北美 TDMA 数字蜂窝电话系统语音编码标准(IS-54),其语音质量与 32Kbps 的 CVSD 和 13Kbps 的 RPELPC 语音质量相当。一种 6.7Kbps 的 VSELP 也被日本采纳为 TDMA 数字蜂窝(JDC)系统全速率语音编码器标准。

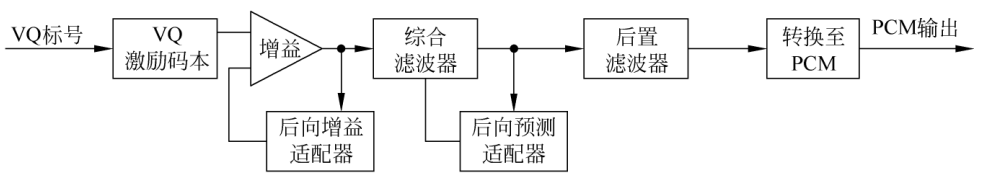
2) 短时延 CELP(LD-CELP)编码

16Kbps 的 LD-CELP 编码算法已标准化为 ITU-T 建议的 G.728 标准。前面所述几种声码器都是利用前馈自适应预测去除语音信号的相关性,它们都需要足够的编码时延和存

储空间，典型的编码时延在 40～60ms 之间。而 LD-CELP 在 CELP 算法基础上，采用带有增益参数的后馈自适应预测和 5 维激励矢量来达到高音质和低时延的效果。它的算法时延是 0.625ms，一路编码时延小于 2ms。LD-CELP 编码器/解码器原理如图 5-24 所示。



(a) 编码器框图



(b) 解码器框图

图 5-24　LD-CELP 编码器/解码器原理框图

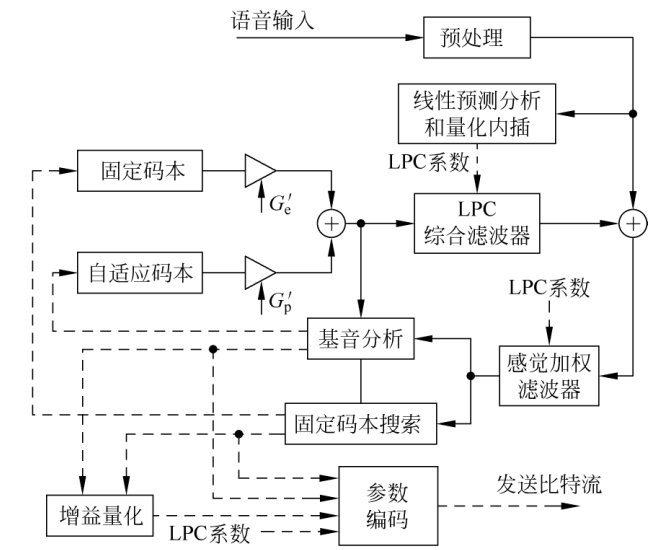
在编码端，5 个连续的语音样点形成一个 5 维语音矢量。激励码本中共有 1024 个 5 维矢量。对于每个输入语音矢量，编码器利用合成分析法从码本中搜索出最佳矢量，然后将 10 比特的 VQ 标号送出去。激励的增益和线性预测系数都是用先前量化过的语音信号来提取和更新的。每 4 个相邻的输入矢量(共 20 个采样点)构成一个子帧，每个子帧更新一次线性预测系数。

3) 共轭结构代数码激励线性预测(CS-ACELP)编码

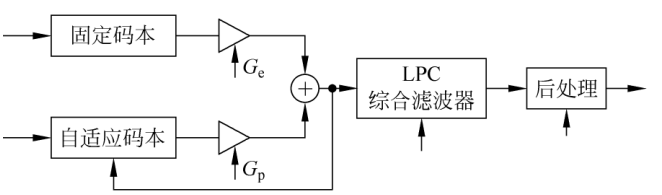
ITU-T 的编码建议的 G.729 标准就是采用这种语音编码方案。其编码原理如图 5-25 所示。

CS-ACELP 的思想是基于 CELP 的编码模式，编码器对增益的矢量量化过程中，采用了共轭结构(conjugate structure)。CS-ACELP 的码本搜索过程也可分为固定码本的搜索过程和自适应码本的搜索过程两部分，其中固定码本采用了代数(algebraic)结构。代数码本的特点是：算法简单，码本不需要存储，其码矢量为 40 维，其中有 4 个非零脉冲，它们的幅度为+1 或−1，位置也在限定的范围内。在解码端，只要从编码中获得非零脉冲的幅度和位置信息，就可直接得到对应的输出矢量。

在发送端要进行线谱对 LSP 参数的量化、基音分析、固定码本的搜索和增益的量化 4 个步骤。编码器首先对输入的信号(8kHz 采样 16 比特 PCM 信号)进行预处理，然后对每帧(10ms)语音进行线性预测分析，得到 LPC 系数，并将其转换为 LSP 参数，接着对 LSP 参数进行二级矢量量化。基音分析采用开环基音分析和自适应码本搜索相结合，每一帧搜索到最佳基音时延 T 的一个候选 T_{op}，然后依据 T_{op} 在每一个子帧内搜索出各自的最佳基音时



(a) 编码器



(b) 解码器

图 5-25 CS-ACELP 编码器/解码器原理框图

延。固定码本的搜索主要是找到 4 个非零脉冲的位置和幅度。最后还需对自适应码本增益和固定码本增益进行量化。除 LSP 参数每帧更新外,其他编码参数每子帧更新一次。

在解码器端,通过对接收到的各种参数标志进行解释得到编码器参数,依次进行激励生成、语音合成和后处理工作。在参数中,对 LSP 参数进行内插,以使其每子帧更新一次,再将其转换成线性预测滤波器系数。

实际上,前文的 LPC 声码器可以看成是只有两类激励矢量的开环 CELP 语音编码器。基于 CELP 编码的变化形式还有很多,例如基音同步刷新码激励线性预测(PSI-CELP)编码、变速率码激励线性预测(QCELP)编码等。1996 年 ITU-T 制定的 G.723.1 编码算法,在网络多媒体通信领域获得了广泛的应用,它提供两种编码速率 6.3Kbps 和 5.3bps。在 5.3bps 编码速率下,编码器采用的是 ACELP 编码算法,而在 6.3Kbps 的速率下,采用的是多脉冲激励线性预测编码算法。

5.2.3 基于正弦模型的混合编码

前文所介绍的 MPLPC、REPLPC 及 CELP 都是基于全极点声道模型,采用线性预测分析方法实现的语音编码算法。这些编码算法通过采用矢量量化技术、合成分析的方法以及感觉加权误差最小判决准则等,在 4.8～16Kbps 速率范围内获得了巨大的成功。然而当速

率进一步降低时，合成语音质量迅速下降。由于全极点声道模型完全是基于人的发音物理机制而提炼出来的，因此上述线性预测编码器在分析和合成非语音声音和数据时（例如语音段中包含很强的噪声），语音的质量就难以满足要求。这里介绍的正弦模型编码所采用的是从语音信号的频谱分解角度出发而建立的正弦分析合成模型。这种模型的主要优点就是，对于一般声音的表示和重建也能给出很好的效果，例如海洋动物的声音、乐音、有音乐背景的语音、多人同时讲话的语音等。基于正弦模型的编码算法同样容易与人耳的听觉模型相结合，改善合成语音的主观音质。

正弦模型的思想是 R. J. Mcaulay 等人在 20 世纪 80 年代提出的，它是相位声码器的进一步发展。语音信号 $x(t)$可以表示为线性时变声道滤波器受声门激励信号 $e(t)$激励而产生的输出，即

$$x(t) = \int_0^t h(t-\tau,t)e(\tau)\mathrm{d}\tau \tag{5-78}$$

式中，$h(\tau,t)$是线性时变声道滤波器的单位冲激响应，设其频率响应为 $H(\omega,t)$。并有

$$H(\omega,t) = M(\omega,t)\exp[\mathrm{j}\Phi(\omega,t)] \tag{5-79}$$

式中，$M(\omega,t)$和 $\Phi(\omega,t)$分别为 $H(\omega,t)$的幅值分量和相位分量。同时可以用一组时变的正弦波来描述激励信号：

$$e(t) = \sum_{k=1}^{N(t)} a_k(t)\sin(V_k(t)+\phi_k) \tag{5-80}$$

其中

$$V_k(t) = \int_{t_1}^t \omega_l(\sigma)\mathrm{d}\sigma \tag{5-81}$$

t_1 是第 k 个正弦波的开始时间。适当地选择幅值 $a_k(t)$、频率 $\omega_l(t)$、相位 ϕ_k 可以形成浊音、清音或过渡音所需要的声门激励信号 $e(t)$。将式(5-80)代入式(5-78)并经推导 $x(t)$可以化简为如下形式（详细推导过程请参见文献[6]）：

$$x(t) = \sum_{k=1}^{N(t)} A_k(t)\sin[V_k(t)+\phi_k+\Phi(\omega_k(t),t)] \tag{5-82}$$

式中，$A_k(t)$为

$$A_k(t) = a_k(t)\cdot M[\omega_k(t),t] \tag{5-83}$$

式(5-82)就是语音信号的正弦模型，即可以将语音信号表示成基音信号及其各次谐波的叠加，这样短时语音信号就可以用基音频率、谐波振幅及其相位参数来表示。其中的振幅和频率是缓慢时变的，可以用帧间峰值匹配算法来估计，而相位常用一种具有去卷绕能力的内插方法来实现其平滑变化。$N(t)$的变化说明语音信号的正弦分量的生灭现象，语音的过渡段主要靠正弦分量的生灭来实现语音特征的急剧过渡，而对于较平稳的浊音段，因可视为准周期性信号，所以也可以用正弦模型很好地描述。数学上已证明，正弦模型可以描述各种准周期性信号。

采用正弦模型对语音信号进行分析与合成具有诸多优点，许多基于这种思想的编码方法，在低速率范围内表现出良好的性能。典型的基于正弦模型的语音编码有正弦变换编码和多带激励编码等。这类编码器都是在分析端通过提取和量化某些参数来表示语音的短时谱，特别注重在浊音语音中的基音谐波；在合成端用一组正弦波相加来合成浊音语音，并通过仔细修正每帧正弦波的频率和相位来跟踪浊音语音的短时谱特性。从这一点来说，基于

正弦模型的语音编码与波形编码有相似之处。

1. 正弦变换编码

正弦变换编码(sine transform coding,STC)是通过对语音进行傅里叶分析,提取最能表示语音信号的几个频率成分,并用这几个频率的正弦波合成语音。

正弦变换编码的原理如图 5-26 所示。在编码端分析语音帧的基音及谐波成分(谱峰),并对这些谱峰和相位的信息进行编码和传输。这样,在接收端通过这些参数控制一组正弦波的幅度和相位来重构语音信号,使合成语音具有与原始语音相似的时变谱结构。

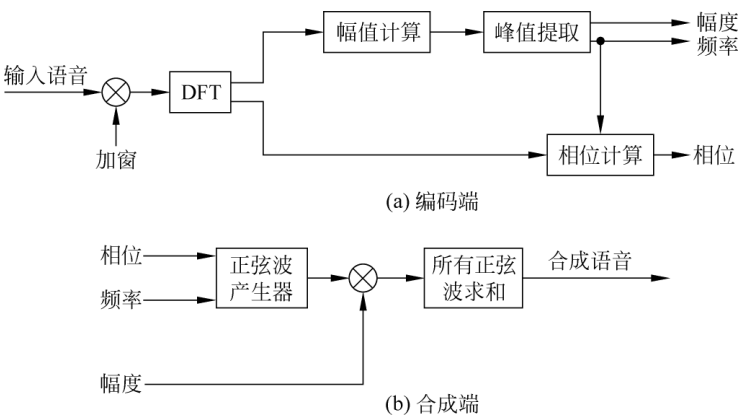


图 5-26 正弦变换编码原理图

STC 编码与波形编码相结合可以产生另一类称之为波形内插(WI)的编码方法。

2. 多带激励(MBE)编码

语音信号短时段中往往既含有周期性分量,又含有非周期性分量,这种特性在频谱上的表现就是在某些频段上语音谱呈现周期谱的特征,而在某些频段上呈现噪声谱的特征。

美国 MIT 林肯实验室于 1984 年提出了多带激励语音编码方案(multi-band excitation,MBE)。它将语音谱按各基音谐波频率分成若干个带,对各带信号分别判断是浊音(V),还是清音(U)。然后根据各带是清音还是浊音,采用不同的激励信号产生其合成信号;最后将各带信号相加,形成全带合成语音。分析过程采用类似于 ABS 的方法,提高了语音参数提取的准确度。MBE 在 2.4～4.8Kbps 速率上能够合成出比传统声码器好得多的语音,并且具有较好的自然度和抗噪性能。

这种算法提出了一种由正弦模型引出的频域模型—多带激励模型,其模型结构如图 5-27 所示。在 MBE 模型中,加窗后的短时语音信号可以表示为

$$X_w(\omega) = H_w(\omega)E_w(\omega) \tag{5-84}$$

即将语音信号的频谱看作系统函数的频谱 $H_w(\omega)$ 与激励信号的频谱 $E_w(\omega)$ 的乘积。而重构语音信号可以表示为

$$\overline{X}_w(\omega) = \overline{H}_w(\omega)\overline{E}_w(\omega) \tag{5-85}$$

式中,$\overline{H}_w(\omega)$ 和 $\overline{E}_w(\omega)$ 分别是 $H_w(\omega)$ 和 $E_w(\omega)$ 的估计,根据原始信号计算得到。

在 LPC 声码器中,$\overline{H}_w(\omega)$ 用全极点函数来逼近。而激励信号 $\overline{E}_w(\omega)$ 采用二元激励形

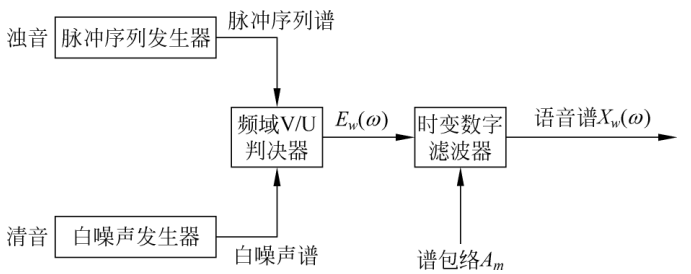


图 5-27　MBE 语音信号产生模型

式。而在 MBE 模型中，首先按基音的各谐波频率，将一帧语音的频谱分成若干个谐波带，然后以若干个谐波带为一组进行分带，例如以 3 个相邻的谐波带为一组进行分带。分别对各带进行清浊(V/U)判决，对于浊音带，用以基音周期为周期的脉冲序列谱作为激励信号谱；对于清音带，则使用白噪声谱作为激励信号谱。总的激励信号由各带激励信号相加构成。系统函数 $\overline{H}_w(\omega)$的作用是确定各频带的相对幅度和相位，起到将这种混合的激励信号谱映射成语音谱的作用。这种模型使得合成语音谱同原始语音谱在细致结构上能够拟合得很好，更符合实际语音的特性。同时在每一谐波带内可以认为 $\overline{H}_w(\omega)$保持不变，用一个常数 A_m 来表示，它描述了各谐波带内的谱包络情况。

MBE 编码器就是通过调整 A_m 和 $\overline{E}_w(\omega)$，使得原始语音谱模值$|X_w(\omega)|$与合成语音谱模值$|\overline{X}_w(\omega)|$之差的加权积分达到最小，即令下式为最小

$$\varepsilon=\frac{1}{2\pi}\int_{-\pi}^{\pi}M(\omega)(|X_w(\omega)|-|\overline{X}_w(\omega)|)^2\mathrm{d}\omega \tag{5-86}$$

式中，$M(\omega)$为感觉加权频率函数。

由图 5-27 可知，对于每一帧语音，必须已知如下参数才能完成对 MBE 模型的分析：基音频率 ω_0、清浊音判决和谱包络参数 A_m(实际是谐波处的谱抽样)。基音频率和谱包络参数的估计是同时进行的。估计时采用搜索算法和最小均方误差准则，依次假设基音频率 ω_0 为各种可能出现的值。对每一个 ω_0，按谐波带宽将 $\omega=-\pi\sim\pi$ 分成 M 个谐波带。各频带频率的上、下限分别为 $b_m=(m+1/2)\omega_0$ 和 $a_m=(m-1/2)\omega_0$，$m=-M\sim M$，则式(5-86)可以写成如下形式：

$$\varepsilon=\sum_{m=-M}^{M}\left[\frac{1}{2\pi}\int_{a_m}^{b_m}M(\omega)(|X_w(\omega)|-|A_m\|\overline{E}_w(\omega)|)^2\mathrm{d}\omega\right] \tag{5-87}$$

可以证明，当

$$|A_m|=\frac{\int_{a_m}^{b_m}M(\omega)|X_w(\omega)\|\overline{E}_w(\omega)|\,\mathrm{d}\omega}{\int_{a_m}^{b_m}M(\omega)|\overline{E}_w(\omega)|^2\mathrm{d}\omega} \tag{5-88}$$

时，式(5-87)取最小值。在未做清浊音判定之前，所有频带均假设为浊音。

基音频率搜索和估计由以下方法实现。

为减少运算的复杂性，先在时域内进行粗估。将式(5-87)转化为时域形式，并加入修正项，得到无偏估计式：

$$\varepsilon_{\mu b} \approx \frac{\sum_{n=-N}^{N} w^2(n)x^2(n) - D\sum_{k=-L}^{L}\phi(kD)}{\left[1 - D\sum_{n=-N}^{N} w^4(n)\right]\left[\sum_{n=-N}^{N} w^2(n)x^2(n)\right]} \tag{5-89}$$

式中,$x(n)$和$w(n)$分别是原始语音信号和窗函数,且有$\sum_{n=-\infty}^{\infty} |w(n)|^2 = 1$。$D$为假定的基音周期,$\phi(m) = \sum_{n=-\infty}^{\infty} w^2(n)x(n)w^2(n-m)x(n-m)$,它实际上是$w^2(n)x(n)$的自相关函数。做估计时,设窗长为$(2N+1)$,并绕原点对称,同时假设在窗长范围内有$L$个假设的基音周期,即

$$L = \left\lfloor \frac{2N+1}{D} \right\rfloor \tag{5-90}$$

符号$\lfloor x \rfloor$表示取小于或等于x的最大整数。通过搜索,可以得到一个基音周期的初次估计值D_{I}。为保证估计的精确度,还要在频域内根据式(5-87)进一步搜索初次估计D_{I}附近的值。当最终确定了ω_0后,可由式(5-88)直接计算对应的$|A_m|$。

对每个频带都要进行V/U判决,首先计算下式

$$\xi_m = \frac{\varepsilon_m}{\frac{1}{2\pi}\int_{a_m}^{b_m} |X_w(\omega)|^2 \mathrm{d}\omega} \tag{5-91}$$

由于在估计谱时假设语音为浊音,因此浊音带误差ξ_m较小,而清音带误差较大。所以可以将ξ_m与一预先设定的门限值η_m比较,从而做出V/U判决。确定V/U后,可以对各谐波的幅度做最后的确定。对于浊音带有$a_m = |A_m|$;对于清音带,其幅度值就是原始语音该谐波带的平均幅度值。

MBE合成算法是以MBE模型为依据,利用分析算法得到的参数来合成语音。清浊音分别进行合成操作,然后将两者相加得到最终的合成语音。

1) 清音语音合成

清音合成是在频域进行的。设U_w是一单位方差白噪声信号的加窗谱。用V/U判决结果来修正U_w,使白噪声信号在频率分布和能量上与原始语音的清音相吻合。用于在谐波带的浊音区,令$U_w(\omega) = 0$,所以修正的效果相当于用一组带通滤波器滤除了浊音带的信号。修正后的U_w再做傅里叶反变换就得到了合成的清音语音序列。为保证前后帧语音的连续性,此序列还要经过前后帧的线性插值,最后得到当前帧语音的清音部分$\bar{x}_{wu}(t)$。

2) 浊音语音合成

浊音可以用一组以基频ω_0及其谐波为振荡频率的正弦波在时域直接合成。即

$$\bar{x}_{wv}(t) = \sum_m a_m(t)\sin(\theta_m(t)) \tag{5-92}$$

式中,$a_m(t)$为第m次谐波带的幅度;而

$$\theta_m(t) = \int_0^t \omega_m(\xi)\mathrm{d}\xi + \phi_0 \tag{5-93}$$

是相位函数,ϕ_0是初始相位,$\omega_m(t)$是经前后帧线性插值的频率轨迹。最后合成语音为

$$\bar{x}_w(t) = \bar{x}_{wv}(t) + \bar{x}_{wu}(t) \tag{5-94}$$

MBE编码在速率降至2.4Kbps时,仍能保持相当的可懂度和自然度。由于MBE不需

要码本，其复杂度也较低，所以基于 MBE 的编码器在多项语音编码标准评选中均显示了强有力的竞争力。一种改进的 MBE 编码器(IMBE)在 1990 年被 INMARSAT 和 AUSAT 采纳，作为其移动卫星通信的语音编码标准，编码速率为 6.4Kbps，EIA/TIA 也选择了 MBE 编码器作为北美陆地移动通信系统(Project25)的语音编码标准，编码速率为 7.2Kbps。

5.3 极低速率语音编码技术

前面介绍的各种编码算法，主要是针对中低速率语音编码应用的。通常将数码率低于 1.2Kbps 以下的语音编码器称为极低速率语音编码器，这类编码器在算法上有着不同的特点，本节专门进行讨论。

现代通信一方面扩展信道，实现“宽带通信”，另一方面仍然追求更加有效、经济实用的信道。其中最重要一项就是要压缩信源频带或编码速率。在语音的通信信道中，有的信道难以扩宽，且质量很差，例如短波信道；有的信道正在广泛使用，短期内难以更新，如市话和载波信道；有的信道通信环境比较复杂，例如在强的“人为干扰”或环境噪声下的军用通信、数字语音保密通信、因特网语音通信；还有的信道十分昂贵，例如卫星、宇宙通信等。在这些条件下，极低速率语音编码颇具吸引力。

5.3.1 400bps～1.2Kbps 的声码器

400bps～1.2Kbps 的语音编码算法一般是在 2.4Kbps 的 LPC 声码器的基础上，利用矢量量化技术和帧间相关性做进一步的数据压缩。

1. 帧填充技术

在 2.4Kbps 声码器的码序列中，相邻帧之间仍存在相关性，尤其在较平稳的语音段，如浊音段，帧与帧之间变化并不大。因此，编码时可以每隔一帧做一次编码传输，并通过边信息通知合成端如何填充空白帧，填充时可以使用前邻帧，也可以使用后邻帧。这样处理大概可以再压缩一半的编码速率。在这种构想的基础上，还可以再做一些更加细致的考虑，比如，使填充帧的基频、能量按既定的规则生成，而不是完全复制相邻帧。采用帧填充技术后，可以在数码率降低一半后，保证合成语音的音质基本保持不变。

2. 矢量量化技术

利用矢量量化技术可以进一步减少帧间编码参数的相关性。在码激励线性预测编码器中，利用矢量量化技术对激励信号进行编码，实现了对编码的压缩，实际上，还可以利用矢量量化技术对声道滤波器系数等参数进行编码，进一步降低编码速率。其基本思路是：把一帧或多帧需要传输的某些参数划分在一起，组成一个矢量。根据感觉误差最小准则，在一个已训练好的码本中搜索该矢量对应的最佳码字，在传输时只传送该码字在码本中的序号，这样就可以进一步降低编码速率，而不过多地影响音质。

在极低速率声码器中，利用矢量量化技术来压缩编码速率的一个典型的例子是 VQ-LPC 声码器。它在 LPC 声码器的基础上，结合 VQ 技术进一步降低了编码速率，而语音质量并没有明显下降。从 5.2.1 节可以看出，LPC 声码器 LPC—10 的参数量化比特分配的情况

为：基音6比特、清浊标志1比特、增益5比特，这些参数已没有进一步压缩的余地。然而 p 个LP参数仍然还具有较大的压缩余地，它们本身就是一种典型的矢量信号。每组LP参数代表一种与能量大小无关的谱形，它反映声道的一种形态。对于这样的矢量，已经找到了与主观感觉有较好对应关系的失真测度方法。既然它是声道形态的表征，那么它在 p 维空间中的分布必然是比较集中的，而人类听觉系统对于语音信号的谱形的分辨能力有限，允许一定程度的量化失真，因此用VQ技术进行量化编码时，码本不必很大。一般情况下，码本中码字的数量为256(2^8)，最多为1024(2^{10})。这样用VQ技术对LP参数进行编码，可以提高其数据压缩比，以 $p=10$ 为例，在量化编码前，若每个参数用4个字节的浮点数表示，则一帧数据总共需要40个字节。若用码本大小为256的矢量量化器编码，一帧数据仅用1个字节，压缩了40倍，就是与前述LPC—10中每个参数孤立地进行编码(即标量量化)时相比，其压缩比也要高四、五倍。

采用VQ技术对LPC参数编码，不必考虑每个参数的量化特性，只要考虑这种参数矢量在多维空间中的失真测度。例如：增益归一化似然比失真测度就是一种用于VQ的良好失真测度，然而计算这种测度所用的参数(被测信号的增益归一化自相关数和参考信号线性预测系数的自相关数)的量化特性都不大好。当然，合成滤波器参数的插值特性仍然是重要的。可以用两类不同的参数存储两个相对应的码本，一个用于VQ编码，一个用于合成和参数内插。

VQ用于数据压缩的所有优势在LP参数的编码中都能得到充分的体现。A. Buzo等人在首次提出VQ技术的应用时，就是用VQ-LPC声码器作为例证来证明VQ压缩数据的强大威力的。这一例证对于新型语音编码器和低速率声码器的发展更是起了重要的推动作用。

5.3.2 识别合成型声码器

从信息论的观点来看，语音所含信息量的信息率下界是50bps左右(对英语而言)。但是，已有的大量研究表明：要将数码率压缩至400bps以下，目前的各种基于分析合成的算法都不能满足要求，所提供的语音质量无法达到公众能接受的程度。其根本原因在于这种分析合成型声码器的编码单元是一帧或几帧语音信号，每帧约为10～30ms的一段，其特性变化无穷，用一个太小的有限符号集来编码，意味着恢复的语音信号难免产生不可容忍的失真。要接近这个下界，只有采用语音识别与合成技术，以语音基元为编码单位进行编码。这一思想早在20世纪50年代就已提出，20世纪80年代还曾有多个研究机构申请过发明专利，但由于面临语音识别和语音合成两大难题，一直没能进行实用化研究。近十几年来，非特定人、连续语音识别和按规则语音合成已取得突破性进展，因此，现在开发这种声码器应该说已经具备了较好的基础。

识别合成型声码器就是采用语音识别与合成技术，以语音单位(或称语音基元)为编码单元对语音信号进行编码。语音基元可以是音素、音节或词，任何一种语言的音素或音节都是一个有限数目的集合，用它们作为基元进行编码可以实现无限词汇的语音编码。这种声码器的结构如图5-28所示，在发送部分采用语音识别技术进行语音基元识别和编码，接收部分根据收到的语音基元代码串和某些附加的韵律信息重新合成语音。因此这种声码器需要在信道中传输的参数很少，可以以极低的编码速率传输或存储语音参数，而且能恢复出高质量的语音。

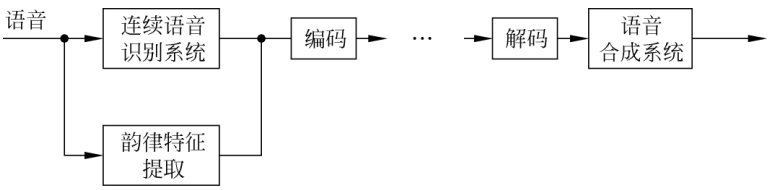


图 5-28　识别合成声码器示意图

这种独特的语音编码技术，至少对于汉语来讲应该是现实可行的，且很有发展前景。这主要是因为汉语语音有其独特的语言结构，其音节基本上是以声母、韵母和声调巧妙地结合而成的。汉语音节总数只有一千多种，它们在语音流中具有一定的独立性和稳定性，比较容易基于音节基元自动识别，也容易以音节为基元合成无限的词汇。

识别合成型语音编码的基础是语音识别和语音合成技术。目前，汉语的非特定人连续语音识别技术和高清晰度、高自然度的语音合成技术已取得重大的进展，因而发展这种识别合成型编码技术的时机已经到来。但是在基于语音识别与合成技术构成的识别合成声码器中，还存在一些在通常的识别合成研究中不曾遇到的问题：

1）如何从语音信号中提取韵律特征参数并对它们进行压缩编码

所谓韵律就是语句中各音节的声学特征，如音长、音强、基音轮廓线、共振峰轨迹等的变化规律；在接收端利用这些韵律参数可以获得较高质量的输出语音。汉语语音音节虽然在语句中有一定的稳定性和独立性，但音节之间的相互影响也是十分明显，特别是同一词内相邻的音节之间存在着明显的协同发音的情况，它们的基音轮廓线和共振峰走向等特征之间的相互影响有时十分显著。因此，合成语句时若不对所有的音节进行适当的韵律修改，合成语音不仅自然度差，可懂度也很低。

2）如何在语音识别中保证获得较高的音节识别正确率

例如使用特定人语音识别技术。虽然汉语非特定人连续语音识别技术已经取得了重大的进展，但是大量的文献表明，非特定人语音识别系统的性能仍然无法和特定人语音识别的性能相比拟。然而在特定人系统中，对于大词汇量语音识别系统而言，由于有大量的参数需要训练，需要使用者录入大量的训练数据，这是一件非常繁琐的工作，而且在很多情况下也是不可能的。一种可行的方法是采用说话人自适应技术，研究表明在语音识别系统中，应用各种快速说话人自适应算法是提高系统性能的一种有效途径。另外，大量研究也表明，适当的语言模型对提高系统的识别率也可发挥重要的作用。而在识别合成声码器中的语音模型又与一般语音识别系统的语言模型有所不同，它可以在保证音节发音正确的情况下，不必区分音节所对应的不同汉字的情况。而且在模型中，韵律信息也可以有效地加以利用，以得到更高的识别率性能。因而研究适用于识别合成型声码器的语言模型也是该编码算法的一项重要任务。

5.4　语音编码器的性能指标和质量评测方法

一般总是通过衡量比较各种语音编码器或语音编码算法的性能指标来评价它们的好坏，这些指标包括编码速率、语音质量、顽健性、时延、计算复杂性和算法的扩展性等。从前

面的分析知道,对同一种编码算法而言,这些性能指标之间往往存在矛盾,必须根据实际情况进行取舍和折中。

5.4.1 编码速率

降低编码速率往往是语音编码的首要目标,它直接关系到传输资源的有效利用和网络容量的提高。根据编码速率和输入语音的关系可将编码器分成两类:固定速率编码器和可变速率编码器。

现有大部分编码标准都是固定速率编码,其范围为0.8～64Kbps。其中,保密电话的编码速率最低,为0.8～4.8Kbps,其原因是它的通信信道带宽限定在4.8Kbps以下。数字蜂窝移动电话和卫星电话编码器的编码速率为3.3～13Kbps,它使数字蜂窝系统的容量可以达到模拟系统的3～5倍。需要注意的是,蜂窝系统中常伴有信道编码,使总的编码速率达到20～30Kbps。普通电话网的编码速率为16～64Kbps。其中有一类特别的编码器称为宽带(wideband)编码器,其编码速率为48/56/64Kbps用于传送50Hz～7kHz的高质量音频信号,主要应用于会议电视系统。在固定速率的编码器中,有些编码器采用了一些特殊的技术来提高信道利用率,例如语音插空技术,它利用语音信号之间的自然停顿传送另一路语音或数据。

可变速率编码是近年来出现的新技术。根据统计,两方通话大约只有40%的时间是真正有声音的,因此一个自然的想法是采用通、断二状态编码。通状态对应有声期,采用固定编码速率;断状态对应无声期,传送极低编码速率信息(如背景噪声特征等),甚至不传送任何信息。更复杂的多状态编码还可以根据网络负荷、剩余存储容量等外部因素调整其编码速率。可变速率编码主要包括两个算法。一是有声检测(voice activity detection,VAD),主要用于确定输入信号是语音还是背景噪声,其难点在于正确识别语音段的起始点,确保语音的可懂度。二是舒适噪声生成(comfortable noise generation,CNG),主要用于接收端重建背景噪声,其设计必须保证发送端和接收端的同步。可变速率编码的典型应用是数字电路倍增设备、非实时的语音存储和CDMA移动通信系统。

5.4.2 顽健性

编码器的顽健性(robustness)是通过取多种不同来源的语音信号进行编码解码,并对输出语音质量进行比较测试得到的一种指标。例如,取不同发音人的语音、各种背景噪声下的语音、用各种麦克风或不同频响的放大器录制的语音、非语音声音等。在应用于通信系统时,编码器要适应各种各样的情况。

多级编码解码(tandem encoding)情况下的输出语音质量,也是衡量编码器顽健性的一项重要指标。在逐步发展起来的数字通信网中,既有数字电话又有模拟电话,从端到端的路由中,语音信号会在模拟信号和数字化压缩编码之间多次进行转换,即出现一种异步级联多级编解码的情况。在这样的情况下,有些编码算法的语音质量就会明显下降,例如ADPCM编码器级联,其音质就大为降低。就是在全数字化网络中,也存在从“64KbpsPCM—数字化压缩编码”这样的多级级联编解码的情况。这种同步多级级联编码形式对于一些复杂的编码算法,例如ATC等的影响非常大。64Kbps的μ律PCM对以上两种类型的多级级联编码、解码的情况都具有很好的顽健性。

此外，在存在部分数据丢失的情况下，语音编码器顽健性的研究也有重要的意义。特别是在异步传输方式(asynchronous transfer mode，ATM)下，通信数据基元丢失是很难避免的。如果不采取一定的措施，即使是64Kbps的μ律PCM的语音质量也会因部分数据丢失而明显降低。解决这一问题的方法有3种，即替代法、插值法和嵌入式编码方案。采用此类方法，可以有效地提高数据丢失时编码器的顽健性。

5.4.3 时延

编码器时延由以下4部分组成：

1) 算法时延

编码和解码操作通常是以帧为单位进行的，有些算法中还需要知道下一帧的部分数据，称为“前视(lookhead)”。因此，算法时延就等于帧长和前视长度之和，其值完全取决于算法，与具体的实现无关。PCM编码的算法时延为125μs。对于低速率编码来说，其典型值为20～30ms。

2) 计算时延

即编码器分析时间和解码器的重建时间，其值取决于硬件速度。通常可认为计算时延等于或略小于帧长，以确保下一帧数据到齐后，当前帧已处理完毕。算法时延和计算时延之和称为单向编解码器时延。

3) 复用时延

即装配时延。编码器发送之前和解码器解码之前，必须将整个数据块的所有比特装配好。

4) 传输时延

其值离散性很大，取决于是采用专用线还是共享信道。对于共享信道而言，常假设传输时延和复用时延之和约为1个帧长。

上述4部分时延之和称为单向系统时延，粗略估计至少为3个帧长。语音通信对于时延有较高的要求。对于交互式通信来说，单向时延大于150ms就可感受到通话连续性受到影响，最大可容忍时延为400～500ms，超过此值只能进行半双工通信。对于具有回声的情况，单向时延不能超过25ms，否则就需要装备回声抑制功能。

需要指出的是，单向系统时延不单决定于语音编码，它还与网络环境等多种外部条件有关。对于不同的系统，即使采用相同的编码器，其系统时延也会有很大的差异。

5.4.4 计算复杂度和算法的可扩展性

计算复杂度主要影响硬件实现的成本。能否推广应用，设备成本当然是一个不容忽视的因素。对于一些复杂的编码算法，如混合编码算法，一般采用处理每一秒钟信号所需的DSP(数字信号处理器)指令条数来衡量其计算复杂度。

所谓算法的可扩展性是指一种编码算法不仅能解决当前的实际应用，而且可以兼顾将来的发展，随着运算器件性能的增强，算法稍加修改就可获得更高的语音质量。这就是要求算法具有可扩展性。

5.4.5 语音质量及其评价方法

编解码后的语音质量受到很多条件的制约,例如编码器速率的高低、环境噪声的情况、传输信道误码的影响、多重编解码的影响、不同发音者(如高音和低音)的影响、不同语言的影响等。在这些制约关系中,数码率等是非常定量的概念,而音质则易受主观因素的影响,然而在对编码器进行性能评价的时候,的确需要一种可重复的、意义明确的、可靠的方法对输出语音质量进行量化。实际上,不只是语音编码领域需要对语音质量定量分析,在语音合成和语音增强等领域同样需要进行音质的评价。

目前用于评价输出语音质量的方法可分为主观和客观两种。主观评价是基于一个或一组评听者对原始语音和失真语音(即经编解码后的重构语音)进行对比试听的基础上,根据某种预先约定的尺度对失真语音来划分质量等级,它反映了听者对语音质量好坏程度的一种主观印象。语音主观评价方法种类很多,其中又可分为音质(quality)评价和可懂度(intelligibility)评价两类。音质直接反映评听人对输出语音质量好坏的综合意见,包括自然度和可辨识说话人能力等方面;而可懂度则反映了评听人对输出语音内容的识别程度。音质高,一般意味着可懂度也高,但反过来却不一定。

1. 可懂度评价方法

可懂度评价方法有以下几种。

1) 判断韵字测试(diagnostic rhyme test,DRT)

DRT是衡量通信系统可懂度的ANSI标准之一。它主要用于低速率语音编码的质量测试。这种测试方法使用若干对(通常是96对)同韵母单字或单音节词进行测试,例如中文的"为"和"费",英文的"veal"和"feel"等。测试中让评听人每次听一对韵字中的某个音,然后让他判断所听到的音是哪一个字,全体评听人判断正确的百分比就是DRT得分。通常认为DRT为95%以上时其清晰度为优,85%～94%为良,75%～84%为中,65%～75%为差,而65%以下为不可接受。在实际通信中,清晰度为50%时,整句可懂度大约为80%,这是因为整句中具有较高的冗余度,即使个别字听不清楚,人们仍然能理解整句话的意思。当清晰度为90%时,整句话的可懂度已接近100%。

2) 改进的韵字测试(modefied rhyme test,MRT)

MRT也是评测通信系统语音可懂度的ANSI标准之一。测试材料由6组,每组50个同韵母的字或词组成,例如汉语中"干、汉、烂、但、半、乱",英语中的"pin、sin、tin、fin、din、win",主要用于区分起始辅音或末尾辅音。评听人针对所听内容选择出6个词中哪个与之相符。

其他还有拼写字母测试(spelling alphabet test,SpAT)以及语音平衡字表法(phonetically balance word list,PB)等。

2. 音质的评价方法

音质评价方法有以下几种。

1) 平均意见得分(mean opinion score,MOS)

MOS法从绝对等级评价法(abosolute category rating,ACR)发展而来,用于对语音整

体满意度或语音通信系统质量的评价。ACR是用于针对电话通信的总体质量评价。MOS与ACR一样采用5级评分标准，如表5-1所示，参加测试的评听人在听完受测语音后，从这5个等级中选择其中某一级作为他对所测语音质量的评价。全体试验者的平均分就是所测语音质量的MOS的得分。MOS是目前应用最为广泛的测试方法。由20～60个非专职测试者参加评听，当MOS≥4.0时认为测试语音是高质量的语音，达到长途电话网的质量要求，接近于透明信道编码，也常称之为网络质量或长途质量。MOS在3.5左右称作通信质量，这时感到重建语音质量下降，但不妨碍正常通话，可以满足话音系统的使用要求。MOS在3.0以下称为合成语音质量，系指一些声码器合成的语音所能达到的质量，它一般具有足够的可懂度，但在自然度及讲话人确认等方面不够好。

表5-1 MOS判分五级标准

MOS得分	质量级别	MOS得分	质量级别
5	Excellent(优)	2	Poor(差)
4	Good(良)	1	Bad(不可接受)
3	Fair(中)		

2) 判断满意度测量(diagnostic acceptability measure，DAM)

DAM的方法是由Dynasta公司推出的一种评价语音通信系统和通信连接的主观语音质量和满意度的评测方法。它具有一些独特的优点。首先，它将直接途径和间接途径结合在一起进行主观质量评价。这里所谓的直接途径是指要求评听人针对语音样本给出个人主观感觉，而不是依赖于人为评价等级的划分；间接途径则是指评听人根据已有的评测标准，脱离开评听人的主观喜好来评分。这样评听人既有机会表达个人主观喜好，又能依标准对每项指标进行评测。例如，在背景噪声下两名评听人或许对语音样本的整体满意度意见相左，但他们很有可能会对语音样本中掺入噪声的多少这一指标达成共识。其次，DMA方法要求评听人可将评价过程划分为总共21个等级，其中10级是考虑信号的感觉质量，8级考虑背景情况，另外3级是可懂度、清晰度和总体满意度。总之，DAM是对语音质量的综合评价，是在多种条件下对话音质量可接受程度的一种度量。它采用百分比评分。

语音主观评价当然是最准确，也是最容易理解的一类方法，但同时也是十分消耗时间、人力和费用的，并且经常受到人的反应的内在不可重复性的影响。针对这些不足，许多基于客观测度的语音质量的客观评价方法相继被提出来，它们都是建立在原始语音信号和失真语音信号的数学对比基础上的。大多数客观评价方法是用数值距离，或者描述听觉系统如何感知质量的模型来量化语音质量。可以说，无法找到一个绝对完善的测度和十分理想的测试方法。一般地，客观评价都要借鉴主观评价的那种高度智能和人性化的过程，其优劣也往往取决于与主观评价结果在统计意义上的相关程度。目前所用的客观测度方法可以分为时域测度、频域测度和其他测度3类方法。时域客观测度定义为被测系统的输入语音与输出语音在时域波形比较上的失真度。主要有信噪比(SNR)和分段信噪比(SNR_{seg})等几种方法。其信噪比取值越大，语音质量越好。频域客观测度采用的是谱失真测度的方法，并模仿人耳的一些听觉特性，使测度结果尽量与主观感觉相吻合。具体测度方法有：对数谱距离测度、LPC倒谱距离测度、Bark谱测度、Mel谱测度等。在频域测度中，一般计算结果取值越小，说明失真语音与原始语音越接近，即语音质量越好。除时域客观测度和频域客观测度

外,还有在此两者的基础上发展起来的其他测度方法,例如相关函数法、转移概率距离测度以及组合距离测度等。

5.5 语音编码国际标准

由于各种运算、存储器件的迅速发展,以及语音通信和存储领域对高质量语音编码需求的日益增加,语音编码技术在近十几年得到了突破性的发展,出现了许多实用的高质量的语音编码算法。针对不同的应用,国际电联ITU和一些地区标准协会已制订了一系列的语音编码标准。这些标准的制订为应用在通信网络中的各种语音编码器的兼容性提供了有力的保证。

关于波形编码的国际标准主要由ITU-T制订,为G系列标准,如表5-2所示。其中G.726为G.721与G.723的合成,G.726推出后,G.723和G.721就删除了。

表5-2 波形编码国际标准

标 准	制订年份	编码速率(kbit/s)	编码算法	话音质量
G.711	1972	64	μ/A律PCM	长途
G.726 (G.721,G.723)	1988 (1984,1986)	40/32/24/16	ADPCM	长途
G.727	1990	40/32/24/16	ADPCM	长途
G.722	1988	64/56/48	SB+ADPCM	长途

有影响的混合编码国际标准和地区性标准主要由ITU-T与数字蜂窝标准组织制订,如表5-3所示。

表5-3 混合编码国际和地区性标准

标 准	制订机构	制订年份	编码速率(kbit/s)	编码算法	话音质量
G.728	ITU-T	1994	16	LD-CELP	长途
G.729	ITU-T	1996	8	CS-ACELP	长途
G.729A	ITU-T	1996	8	CS-ACELP	长途
G.723.1	ITU-T	1995	6.3/5.3	多脉冲CELP	长途
GSM全速率	ETSI(欧)	1987	13	RPE-LTP	长途
GSM半速率	ETSI(欧)	1994	5.6	VSELP	长途
IS54	TLA(美)	1989	7.95	VSELP	=RPE-LTP
IS96	TLA(美)	1993	8.5/4/2/0.8	QCELP	<IS54
JDC全速率	RCR(日)	1990	6.7	VSELP	<IS54
JDC半速率	RCR(日)	1993	3.45	PSI-CELP	同全速率

注:ETSI——欧洲电信标准学会;TLA——电信工业协会;RCR——无线电系统研发中心。

5.6 感知音频编码

前面介绍的是针对语音信号的编码原理和编码方法。然而现实世界中存在大量非语音的其他音频信号,如音乐、音效等,这些音频信号的带宽比语音信号要宽,其产生机理也与语

音有很大的差异，所以语音编码算法并不能很好地适用于这些音频信号。近十几年来，出现了不少针对一般音频信号的压缩编码技术，例如 MPEG-1 Layer3、MPEG-2 AAC、Dolby 实验室的 AC-3、微软的 WMA、Xiph 公司的 Ogg Vorbis、Lucent 科技的 EPAC 和索尼的 ATRAC-3 等。这些编码在时频域分析环节所采用的技术各不相同，如 MPEG-1 Layer3 采用了 5.1.7 节中所述的子带编码方法，而 Dolby 的 AC-3 则采用了 5.1.8 节所述的变换域编码方法。

虽然采用的是不同的时频域分析方法，但这些音频信号的编码技术也有一些共性的特点，它们都在编码的量化环节充分利用了人耳的感知机理，保留人耳能听到的音频信号，而对感知灵敏度小或接近不可感知的音频信号进行大幅度的压缩，从而在保证主观听觉效果的前提下，达到最好的压缩效果，即用最少的比特数来代表原始信号。由于这类编码技术充分利用了人耳的感知机理，因而常被统称为感知编码。本节将对感知编码技术进行概要介绍。

5.6.1　感知编码的一般框架

对一个典型的感知音频编码器，它先将时域的声音信号转换成频域的信号，再借由听觉感官模型在频域上计算出人耳听觉可允许的量化误差，然后利用此量化误差值对音频进行编码，使编码后的误差人耳感觉不出来或者在可以忍受的范围内。在音频感知编码中使用的听觉感官模型又常被称为心理声学模型。

一般的感知音频编码器的主要架构如图 5-29 所示，包含了心理声学模型的分析、信号的时频域转换分析、量化及比特分配和无损熵编码等基本部分。

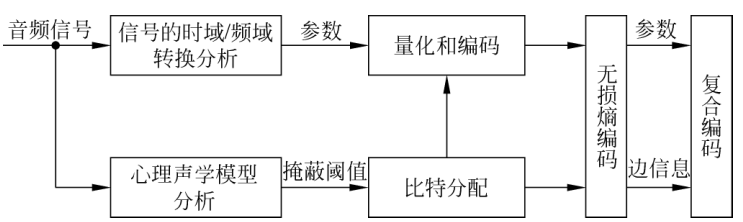


图 5-29　感知编码的一般框架

对音频信号首先进行时频域分析，提取时频域参数，然后对时频域参数进行量化编码。对大多数音频感知编码方法而言，一般在频域上计算编码参数，如 MPEG-1 Layer3 将子带编码和 MDCT 变换编码相结合来得到频域编码参数。

在量化编码过程中，一个重要的问题就是如何在比特分配过程中，将有限的比特数合理地分配给各个子带或变换系数。感知编码的一个重要特征，就是基于心理声学模型的分析结果来分配比特数。音频信号的接受方是人耳，虽然声音是客观存在的，但是人的主观感觉和客观实际并不完全一致，人类的听觉系统对声音的音高、音强和动态频谱等具有分析感知能力。这些听觉特性在心理声学模型分析时需要加以考虑。目前，音频感知编码的心理声学模型主要利用的是听觉掩蔽效应，通过采用一种近似的数学模型，对掩蔽效应进行定量分析，计算出掩蔽阈值曲线，从而在比特分配过程中确保所引入的量化噪声尽可能处于掩蔽阈值曲线下方，这样就可保证在量化时即使引入了量化噪声也无法被人耳听见。

上述量化和编码过程显然是一种有损压缩,在感知编码中通常会在有损编码的基础上引入一个无损熵编码环节,对有损压缩的结果进一步的压缩。霍夫曼(huffman)编码是最常采用的技术,它合理利用信源的统计特性,采用非等长编码,对概率大的信源符号赋予长度较小的码字,对概率小的信源符号赋予长度较大的码字,使平均码长尽可能小。霍夫曼码的译码具有唯一性。

5.6.2 心理声学模型

心理声学模型是感知编码算法的核心,它是否能真实地反映人耳的主观感知特性决定了整个编码器编码质量的优劣。心理声学模型的基本思想就是不依据音频波形本身的相关性和人的发音机理,而利用人的听觉系统的特性来达到压缩音频数据的目的,同时使失真尽可能不被觉察出来。在 MPEG-1 Layer3 和 AAC 标准及 AC-3 标准中都采用了心理声学模型。这些模型将听阈、临界频带、时域掩蔽和频率掩蔽等概念紧密相连,用客观的参数指标反映主观的听觉效果,以使量化、编码过程中产生的量化噪声不易被感知,达到高效率、高保真编码的目的。

在 MPEG 音频标准中给出了两种心理声学模型,心理声学模型Ⅰ和心理声学模型Ⅱ。前者结构较为简单,计算复杂度较小,适用于对压缩比要求不高的场合,主要应用在 MPEG-1 Layer1 和 Layer2 中。后者计算复杂度大,但能够提供更为精确的声学参数,已被 MPEG-1 Layer3、MPEG-2 AAC 以及 MPEG-4 AAC 所采用。两个心理声学模型都通过计算信号的信掩比(signal-to-mask ratios,SMR)来为编码器服务,基于 SMR 值对每个频带进行比特分配,SMR 值越大给予的比特数越多,反之则越少。在比特率一定的条件下,编码质量的优劣取决于对每个频带中比特分配是否得当。本节以心理声学模型Ⅱ的计算过程为例来介绍其算法思想。

心理声学模型Ⅱ采用 FFT 滤波器组对输入信号进行频域分析,这一变换过程与编码器的频域分析是相互独立的,如在 MPEG-1 Layer3 和 MPEG-2 AAC 中,编码器采用改进的离散余弦变换(MDCT)分析滤波器组来获得频域参数,而其心理声学模型则基于 FFT 进行频谱分析。

首先对音频信号 $x(i)$进行加窗处理,然后对其进行 FFT 变换,使用极坐标表示,得到其频谱幅值 $r(\omega)$和相位 $f(\omega)$。由于编码器需要有效平衡音频编码的频域分辨率和时域分辨率,所以其 MDCT 变换可以采取两种不同的块变换类型(长块和短块)。对应地,在心理声学模型计算中,也需要对同一帧音频信号分别计算出两套频域表示,如计算一组 2048 点的 FFT 和八组 256 点的 FFT。

然后根据频谱系数得到各临界频带内的信号能量和不可预测性测度(unpredictability measurement)。先根据前两帧的 $r(\omega)$和 $f(\omega)$来得到当前帧的预测频谱 $r_pred(\omega)$和 $f\_pred(\omega)$,有

$$r\_pred(\omega) = 2.0 \times r_{t-1}(\omega) - r_{t-2}(\omega) \tag{5-95}$$

$$f\_pred(\omega) = 2.0 \times f_{t-1}(\omega) - f_{t-2}(\omega) \tag{5-96}$$

$r_{t-1}(\omega)$和 $f_{t-1}(\omega)$为当前帧前面第一帧的频谱幅值和相位,$r_{t-2}(\omega)$和 $f_{t-2}(\omega)$为当前帧前面第二帧的频谱幅值和相位。接着,根据频谱幅值和相位的预测值,以及频谱幅值和相位的实际值进行信号不可预测性 $c(\omega)$的计算:

$$c(\omega)=[(r(\omega)\cos(r(\omega))-r\_pred(\omega)\cos(r\_pred(\omega)))^2+(f(\omega)\cos(f(\omega))-f\_pred(\omega)\cos(f\_pred(\omega)))^2]^{\frac{1}{2}}/(r(\omega)+abs(r\_pred(\omega))) \tag{5-97}$$

预测值与实际值间的差距越大，则不可预测性也越大。在心理声学模型Ⅱ中，不可预测性 $c(\omega)$ 表现为频率的函数。在每个临界频带上计算该频带的不可预测性 $c(b)$ 和能量 $e(b)$，计算方法如下。

$$c(b)=\sum_{\omega=b\_low}^{b\_high}c(\omega)r(\omega)^2 \tag{5-98}$$

$$e(b)=\sum_{\omega=b\_low}^{b\_high}r(\omega)^2 \tag{5-99}$$

式中，b 是特定临界频带的序号，b_low 和 b_high 分别为该临界频带的频率下界和上界。在实际计算中，还需要将这两项分别与扩展函数进行卷积运算，得到新的不可预测性和能量，从而考虑了其他临界频带对本临界频带的掩蔽影响。

音频信号中的音调（纯音）成分和非音调（噪声）成分具有不同的掩蔽性，这会影响到附近的掩蔽阈值，因此为了计算一个临界频带的总掩蔽阈值，必须对音调成分和非音调成分加以区分。可以根据频带的不可预测性做出该频带是否是音调成分的判断。MPEG 的心理声学模型Ⅱ没有直接区分音调成分和非音调成分，而是将音调指标表达成一个音调索引函数。该函数反映该频段是音调成分的概率大小，避免了直接区分判决而引入的判决误差。临界频带 b 的音调索引函数 $tb(b)$ 计算如下：

$$tb(b)=-0.299-0.43\log_e(c(b)) \tag{5-100}$$

$c(b)$ 为临界频带的不可预测性。$tb(b)$ 的取值在 0～1 之间，越趋向 1 表明信号更接近音调，反之则接近非音调。

根据音调索引函数，可以进一步计算每个临界频带中的信噪比 $SNR(b)$。

$$SNR(b)=tb(b)\times TMN(b)+(1-tb(b))\times NMT(b) \tag{5-101}$$

式中，$TMN(b)$ 为临界频带 b 的音调对噪声的掩蔽（tone masking noise），$NMT(b)$ 为临界频带 b 的噪声对音调的掩蔽（noise masking tone）。一般所有临界频带上的 $NMT(b)$ 设为 6dB，$TMN(b)$ 设为 18dB。

根据信噪比 $SNR(b)$ 和能量 $e(b)$，可以如下计算临界频带的掩蔽阈值 $nb(b)$：

$$nb(b)=e(b)\times 10^{-SNR(b)/10} \tag{5-102}$$

式中，$10^{-SNR(b)/10}$ 的部分为功率比，所以 $nb(b)$ 给出了此临界频带的噪声阈值。在实际计算中还要引入听阈对 $nb(b)$ 进行修正。听阈又被称为绝对听觉门限，是指一个人在没有噪声的环境下，就声音的某一个频率点（纯音），信号能产生听觉感知的最低能量幅度。即若纯音信号幅度小于该频率的听阈，人就无法感知了。显见我们计算得到的临界频带的掩蔽阈值若小于其听阈是没有意义的，此时应将掩蔽阈值设为听阈。听阈是根据大量心理声学实验得出的，对心理声学模型而言是预制的。MPEG 标准根据输入 PCM 信号的采样率的不同制定了“频率、临界频带比率和听阈”表，从表中可以查出频谱的听阈的值。

通过上述计算，我们得到了各临界频带的掩蔽阈值，然而编码器频域分析所采用的是 MDCT 滤波器组，其对频带的划分与临界频带的划分方法并不相同，因而还需要将在临界频带上得到的参数转换到 MDCT 所得到的各子带上去，这些子带被称为缩放因子频带

(scalefactor band)。基于缩放因子频带上的掩蔽阈值进而可以得到信掩比SMR,它表示为FFT频谱能量和噪声的比值。

在心理声学模型Ⅱ中还需要计算感知熵(perceptual entropy)。感知熵是1988年Johnson等利用心理声学模型的掩蔽现象和信号的量化原理定义的,用来测量音频信号中感知相关的信息。感知熵一般以位(bit)作为单位,实际上表示音频信号压缩的理论极限。感知熵 PE 可以由各临界频带的能量 $e(b)$ 和掩蔽阈值 $nb(b)$ 来求得

$$PE = -\sum_{b}(b\_high - b\_low)\log_{10}(nb(b)/(e(b)+1)) \tag{5-103}$$

式(5-103)对所有的临界频带求和,b_low 和 b_high 分别为临界频带 b 的频率下界和上界。

首先通过感知熵信息可以为编码器MDCT变换选择块变换类型,判断使用长块还是短块。将感知熵与一个切换阈值相比较,并参考前一帧的块类型情况决定当前的块类型。此外,感知熵信息也可以在无损熵编码环节用于确定所需要的比特数。

5.6.3 常用的感知编码标准

1. MPEG-1 Layer3

通常被简称为MP3,是MPEG-1的衍生编码方案(ISO/IEC11172—3,1992)。MP3是1993年由德国Fraunhofer IIS研究院和汤姆生公司合作研制的,是目前最为普及的音频压缩格式。它采用了子带分解、分析滤波器组、变换域编码、熵编码、动态比特分配、非同一量化编码和心理声学分析等技术,支持32kHz、44.1kHz和48kHz采样频率下对16比特PCM信号进行编码,同时,提供单声道、立体声道、两个独立双声道和联合立体声等四种音频声道模式。

随着网络的普及,这种开放式的音频编码格式,受到了数以亿计的用户的欢迎,各种与MP3相关的软件产品层出不穷,而且更多的硬件产品也开始支持MP3,我们能够买到的VCD/DVD播放机有很多都能够支持MP3,还出现了许多便携的MP3播放器等。

MP3编码流程见图5-30所示。PCM信号分两路进入编码器,一路进入多相滤波器组中分解为32个等带宽的关键采样的子带,然后再经过MDCT变换得到频域内的频谱系数;另一路PCM输入数据进行FFT变换,进行心理声学分析,得到每个子带的信掩比SMR等参数送入其他模块。把心理声学模型分析模块输出的心理声学参数送到量化编码模块,计算出编码所需的比特数,然后在信掩比和所需比特数的指导下,对经滤波器组输出的频谱系数进行非线性量化和霍夫曼无损编码。最后由比特率、采样率和量化编码后的频谱等共同形成最终的比特流。

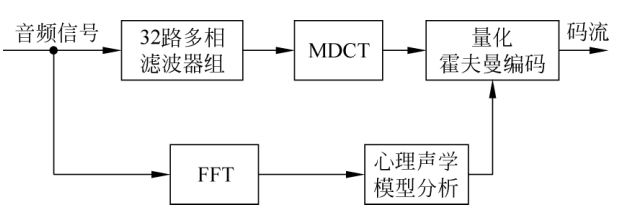


图5-30 MP3编码的简略框图

2. AC-3

Dolby AC-3 是美国 Dolby 实验室于 1990 年提出的，到了 1997 年初，Dolby 实验室正式将其改为“Dolby 数码环绕声”(dolby surround digital)，常称为 Dolby Digital。它是适用于宽频带数字音频信号的变换编码算法，也是数字音频信号压缩的典型应用。该算法可以满足单声道到 5.1 声道数字音频的编码要求，采用时域混叠抵消技术，并运用人耳掩蔽效应，从而对 PCM 信源进行高效压缩，恢复质量与原音相差无几。

AC-3 编码采用的 5.1 声道环绕立体声系统，所有的 5 个全带宽声道和低频效果声道实行统一编码，使之成为复合数据流，其比特流所允许的采样频率可以为 48kHz、44.1kHz 或 32kHz 中的任何一种，声音样本精度为 20 比特，并且所支持的码率从 32Kbps 到 640Kbps 不等。目前，数字音频压缩 AC-3 算法已在很多领域得到广泛应用，如 DVD、激光视盘、HDTV、多媒体等，它是发展家庭影院的关键技术之一。

图 5-31 显示的是 AC-3 的编码流程。PCM 音频信号在进入 MDCT 滤波器组进行时频域变换之前，需要先经过暂态检测器判断音频信号的突变性，若信号变化比较平缓，则在进行 MDCT 变换时使用长窗，即对每个音频块进行 512 点 MDCT 变换；若信号变化剧烈，则将音频块划分成 2 个 256 点 MDCT 变换。得到的频域系数按照指数形式分解为指数和尾数两个部分，其中尾数为规整化后的大于 0 小于 1 的数，指数为 0～24 之间的整数。然后，这些指数和尾数分别送到指数编码器和尾数量化器中进行编码，而在进行尾数的量化时，必须将 MDCT 变换后的频谱包络送到感知模型中，通过频谱包络计算出掩蔽阈值，再通过比特分配模块计算出量化比特数。最后，经过编码后的尾数和指数信息，感知模型参数及某些比特信息参数组合成 AC-3 码流，即完成 AC-3 编码过程。

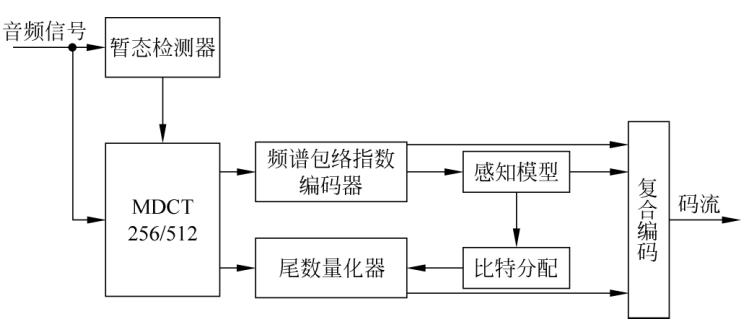


图 5-31　AC-3 编码的简略框图

3. AAC

AAC 是 1997 年制定的 MPEG-2 advanced audio coding 的缩写，它是由 MP3 专利的拥有者 Fraunhofer IIS 联合 Dolby、AT&T、索尼、苹果等产业巨头共同开发出的一种数字音频压缩方式。它增加了诸如对立体声的完美再现、比特流效果音扫描、多媒体控制、版权保护、降噪等 MP3 没有的特性，在音频压缩后仍能完美地再现 CD 的音质。它对大部分立体声信号在 128Kbps 码率下具有感知透明的特性，在 96Kbps 码率的表现超过了 128Kbps 的 MP3 格式，但是对早期的标准不具有后向兼容性。

相对MP3等以往的音乐格式,AAC具备了不少优点,如:压缩率高,可以有更小的文件尺寸(音频压缩比可达到15∶1～20∶1)获得更高的音质;支持多声道,最多可达48个全音域声道;更高的解析度,可支持8～96kHz的采样频率;提升的解码效率,解码播放所占的资源更少;允许对多媒体信息进行编解码等。

AAC的算法复杂度比MP3高很多,也具有多声道、高采样率和低码率下的高音质等特点,非常适合未来的DVD应用。AAC也得到了诺基亚、苹果、松下等多家移动娱乐产品巨头的鼎力支持,另外,出现了一些编码软件,如FAAC,Nero AAC,苹果公司的QuickTime/iTunes等。AAC在移动通信、网络电话、在线广播等领域,被认为是立体声与多声道音频信号编码的下一代通用标准。

后续发展的MPEG-4音频标准,MPEG-4 AAC,是在MPEG-2 AAC的基础上,增加了一些新的编码特性,从而进一步降低音频码率、提高编码效率。

参考文献

[1] 杨行峻,迟惠生等.语音信号数字处理[M].北京:电子工业出版社,1995.
[2] 易克初,田斌,付强.语音信号处理[M].合肥:国防工业出版社,2000.
[3] 王炳锡.语音编码[M].西安:西安电子科技大学出版社,2002.
[4] 糜正琨.IP网络电话技术[M].北京:人民邮电出版社,2000.
[5] 胡航.语音信号处理[M].哈尔滨:哈尔滨工业大学出版社,2000.
[6] 蔡莲红,黄得智,蔡锐.现代语音技术基础与应用[M].北京:清华大学出版社,2003.
[7] 李琳.音频感知编码及关键技术研究[D].合肥:中国科学技术大学,2008.
[8] McAulay R J,Quatieri T F. Speech analysis-synthesis based on asinusoidal representation[J]. IEEE Trans Acoustic Speech Signal Process,1986,744-754.