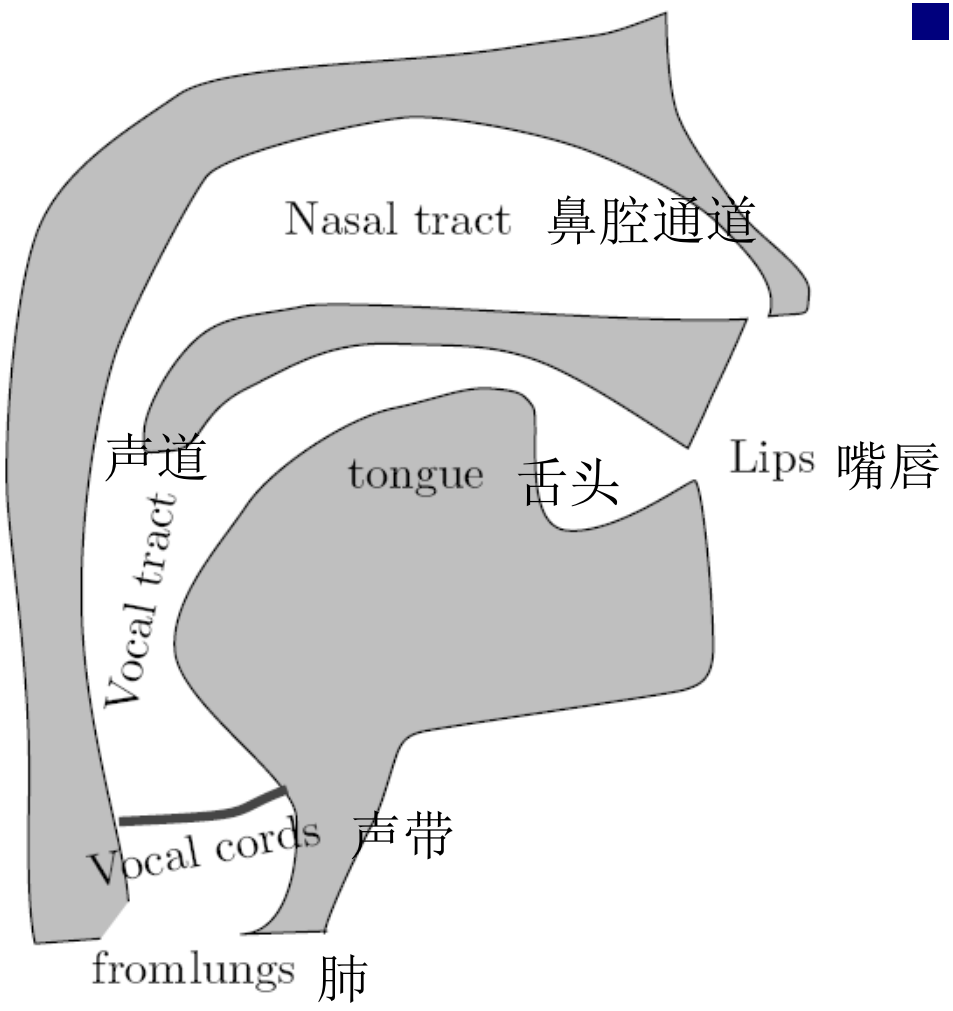
第八章　语音编码

主要内容

- 8.1 语音信号的特性
- 8.2 波形编码
- 8.3 声码编码/参数编码
- 8.4 混合编码

8.1 语音信号的特性



- 人的发声系统

当肺部中的受压空气沿着声道通过声门发出时就产生了话音。但声音是从声道（从声带延展到嘴，成人平均声道长度为17m）中产生的，声音的基音由**声道的形状变化**（主要通过移动舌头）和**移动嘴唇**控制。强度（响度）通过改变从肺部发出的气体的量改变。

人的声音变化很慢，肺的操作很慢，声道的形状变化很慢，所以语音的基音和强度变化也很慢。

表现在**相邻样本间很相似，即使是帧（20ms）也是强相关的**。这种相关性正是语音压缩的基础。

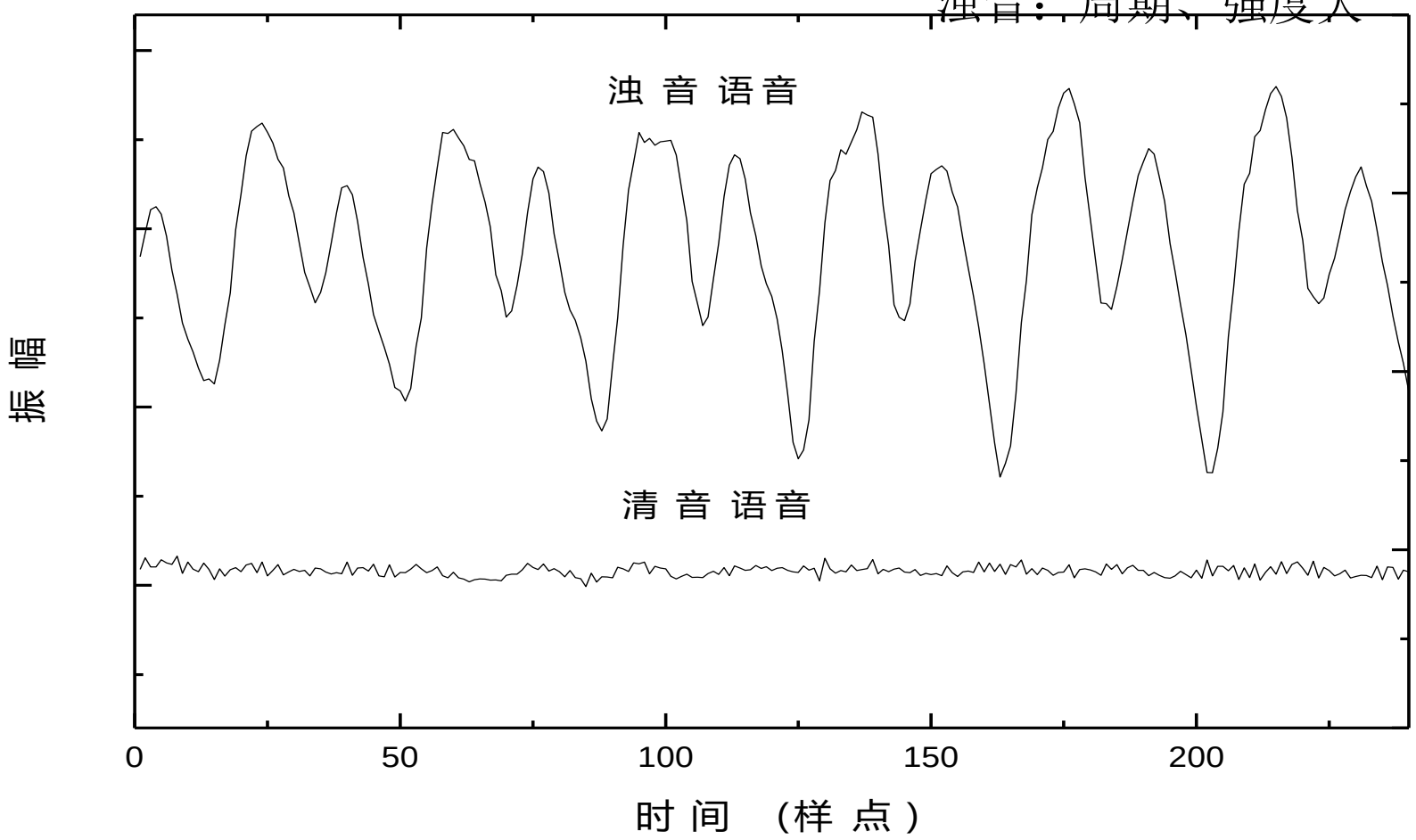
8.1.1 人的发声系统

声带可以打开和关闭。声带之间的打开称为声门（glottis）。声门和声道的移动产生不同类型的声音：

- 浊音（ Voiced sounds）：. 我们说话时发出的声音。声带振动，引起声门的打开和关闭，从而发送压力变化的脉冲到声道，在声道形成声波 。改变声带的形状和强度会改变声门的速率并控制声音的基音。人耳对频率从16Hz到大约20,000–22,000 Hz的声音敏感。而人类的声音范围更窄，通常在500 Hz到大约2 kHz的范围，等价于周期从2 ms到20 ms（对计算机而言，这是相当长的一短时间）。因此浊音信号有长时间周期性，这是语音压缩的关键
  - 浊音、m, n, l, r
- 清音（Unvoiced sounds）：清音是声门保持打开并将气体压进一个收缩声道的结果。清音样本表现出很少的相关性，是随机的或接近随机的
  - f, h, j, q, x, zh, ch, sh, z, c, s
- 爆破音（Plosive sounds）：声道关闭之后产生的压缩空气然后突然打开声道所发出的音
  - 爆破音： b, p, t, d, g, k

8.1.1 人的发声系统

浊音：周期、强度大



典型的浊音和清音的波形

8.1.1 人的发声系统

- 语音产生的物理过程
  - 当声音由三种激励方式产生后，便顺着声道进行传播
- 声道：具有某种谐振特性的腔体，腔体的一组谐振点称为共振峰
  - 类似滤波器，对输入信号进行调制
  - 这些共振峰的位置以及各个峰的宽度决定了声道的频谱特性，共振峰及带宽取决于声道的形状和尺寸

8.1.2 语音信号的时域冗余度

- 幅度非均匀分布
  - 小幅度样本出现的频率高
- 样本之间的相关性
  - 当取样频率为8KHz时，相邻样本间的相关系数大于0.85；
- 周期之间的相关性
  - 在特定瞬间，某段声音往往只是总频带300~3400Hz的少数几个频率分量在起作用→象某些振荡波一些，在周期与周期之间存在一定的相关性
- 基音之间的相关性
  - 男声基音周期为5~20ms，而典型的浊音持续100ms
- 静止系数（话音间隙）
  - 全双工话路的典型效率约为40%（静止系数为0.6）
- 长期相关性（long term correlation）
  - 如几十秒内的相关性

8.1.3 语音信号的频域冗余度

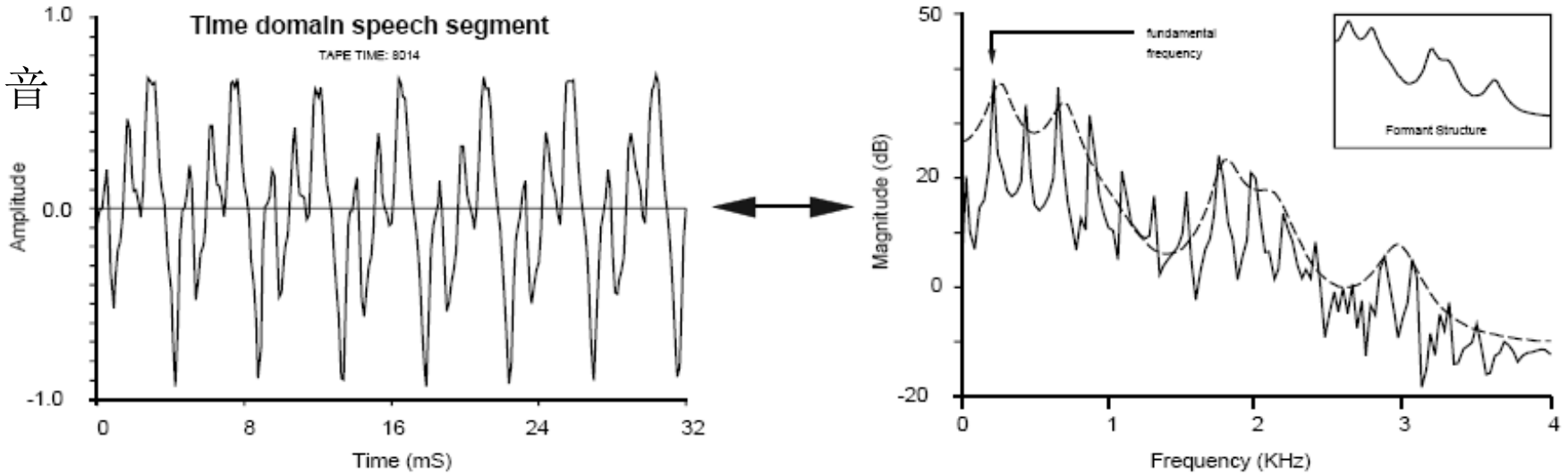
从频域考察语音信号的功率谱密度：

- 非均匀的长时间功率谱密度
  - 长时间功率谱呈现强烈的非平坦性，高频能量较低→时域上相邻样本相关
- 语音特有的短时功率谱密度
  - 语音的短时功率谱，在某些频率出现峰值（该频率称为共振峰频率），在另外一些频率上出现谷值。
  - 出现共振峰的频率不止一个，最主要的是前两个，决定了不同的语音特征
  - 整个谱也随频率增加而递减
  - 功率谱的细节以基音频率为基础，形成高次谐波结构

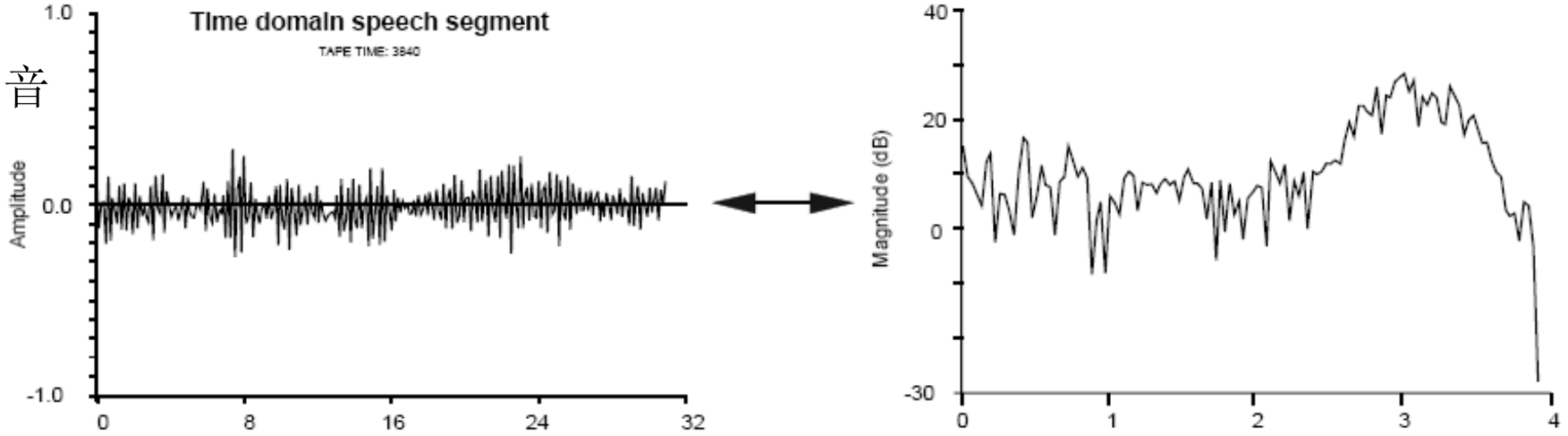
另外，人的声道形状及其变化规律是有限的→按一定的时间段（帧）来计算声道滤波器的参数或语音谱包络

语音信号的短时功率谱

浊音



清音



语音信号的预测编码

- 波形编码：使重构语音信号和原始语音信号在波形上尽可能相同
  - PCM、Delta M、DPCM/ADPCM
  - 子带-ADPCM
  - 基于变换的编码：STC、DCT
- 音源编码/参数编码：根据语声模型提取语音信号的特征参数，接收端根据特征参数重建语音信号
  - LPC
- 混合编码：既保留参数编码的声道模型，又像波形编码那样传递预测误差
  - CELP 、MP-LPC

语音编码技术的发展

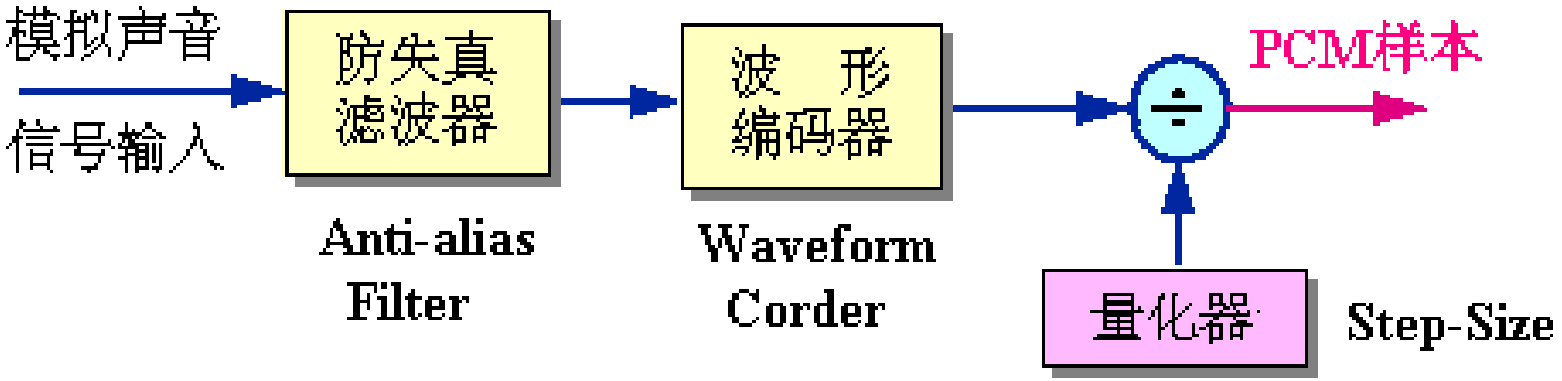
- 1972年的G.711建议：64kb/s PCM
  - 语音质量高
  - 用于数字通信、数字交换机等领域
- 1984的G.721建议：32kb/s ADPCM
  - 语音质量与PCM相同，抗误码性能优良
  - 用于卫星，海缆及数字语音插空设备以及可变速率编码器中
- 1992年的G.728建议：16kb/s低延迟码激励线性预测（LD-CELP）
  - 延迟较小、速率较低、性能较高
  - 在实际中得到广泛的应用，如可视电话伴音、无绳电话机、单路单载波卫星和海事卫星通信、数字插空设备、存储和转发系统、语音信息录音、数字移动无线系统、分组化语音等
- 1996年的G.729建议： 8kb/s共轭代数码激励线性预测（CS-ACELP）
  - 延迟小，节省87.5%%的带宽，语音质量与32kb/s的ADPCM相同，且在噪声较大的环境中也会有较好的语音质量
  - 广泛应用于个人移动通信、数字卫星通信、高质量移动无线通信、存储/检索、分组语音和数字租用信道等领域

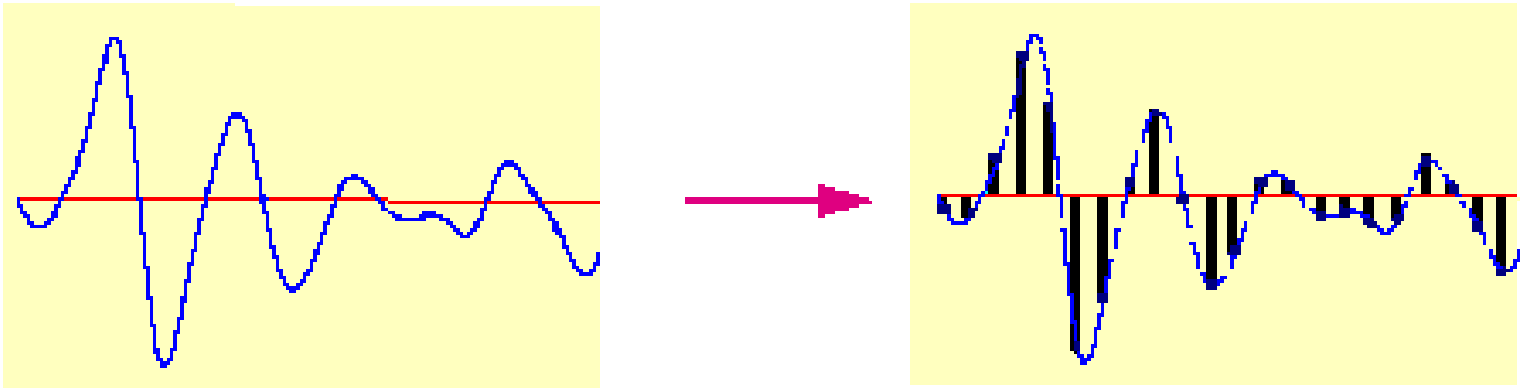
	算法	名称	数据率	标准	应用	质量
波形编码	PCM	均匀量化			ISDN	4.0-4.5
	u/A律	折线量化	64kbps	G.711		
	APCM	自适应量化				
	DPCM	差分预测				
	ADPCM	自适应差分预测	32kbps	G.721		
	SB-ADPCM	子带- 自适应差分预测	64kbps	G.722		
			5.3kbps/ 6.3kbps	G.723		
参数编码	LPC	线性预测编码	2.4kbps		保密通信	2.5-3.5
混合编码	CELPC	码激励LPC	4.8kbps		移动通信	3.7-4.0
	LD-CELP	低延时码激励LPC	16kbps	G.728 G.729		

8.2 波形编码

- 脉冲编码调制

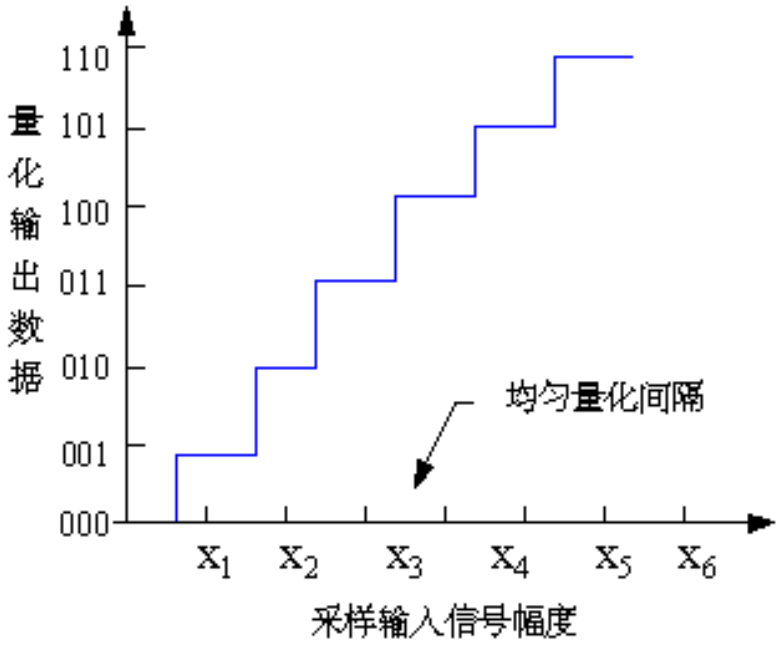
PCM(Pluse Code Modulation)





8.2.1 PCM

- 线性量化
  - 用相等的量化间隔对采样得到的信号作量化
  - 为适应幅度大的输入信号，同时又满足精度要求，需要增加样本的位数
  - 对语音信号来说，大信号出现的机会不多，增加的样本位数没有充分利用，因此出现了非均匀量化的方法

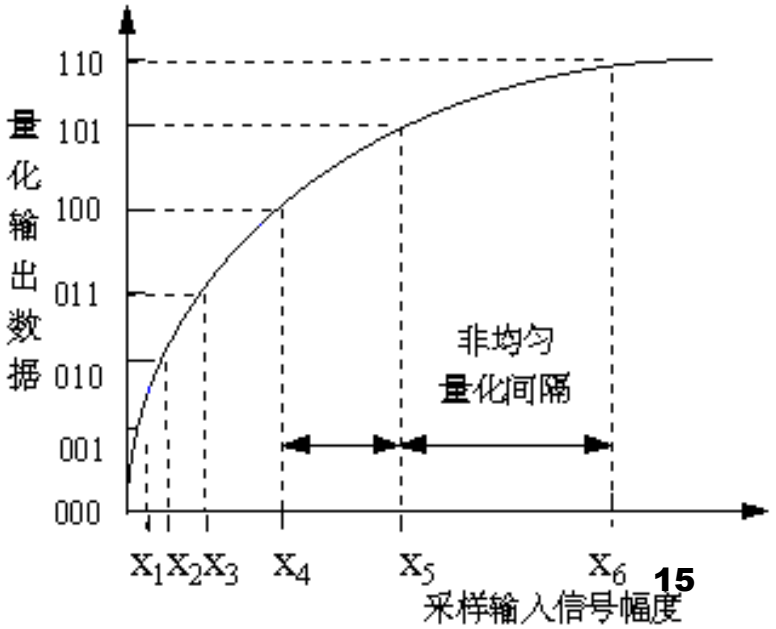


8.2.2 u/A律

- 非线性量化：
  - 对输入信号进行量化时，大输入信号采用大量化间隔;小输入信号采用小量化间隔→满足精度要求下用较少位数表示
  - 先用一非线性变换 $F(x)$ 将信号“压缩”，然后再均匀量化，通信网中已经使用两种对数变换形式（压扩量化）
    - μ律（北美和日本等地的数字电话通信）
    - ***A***律（欧洲和中国大陆等地的数字电话通信）

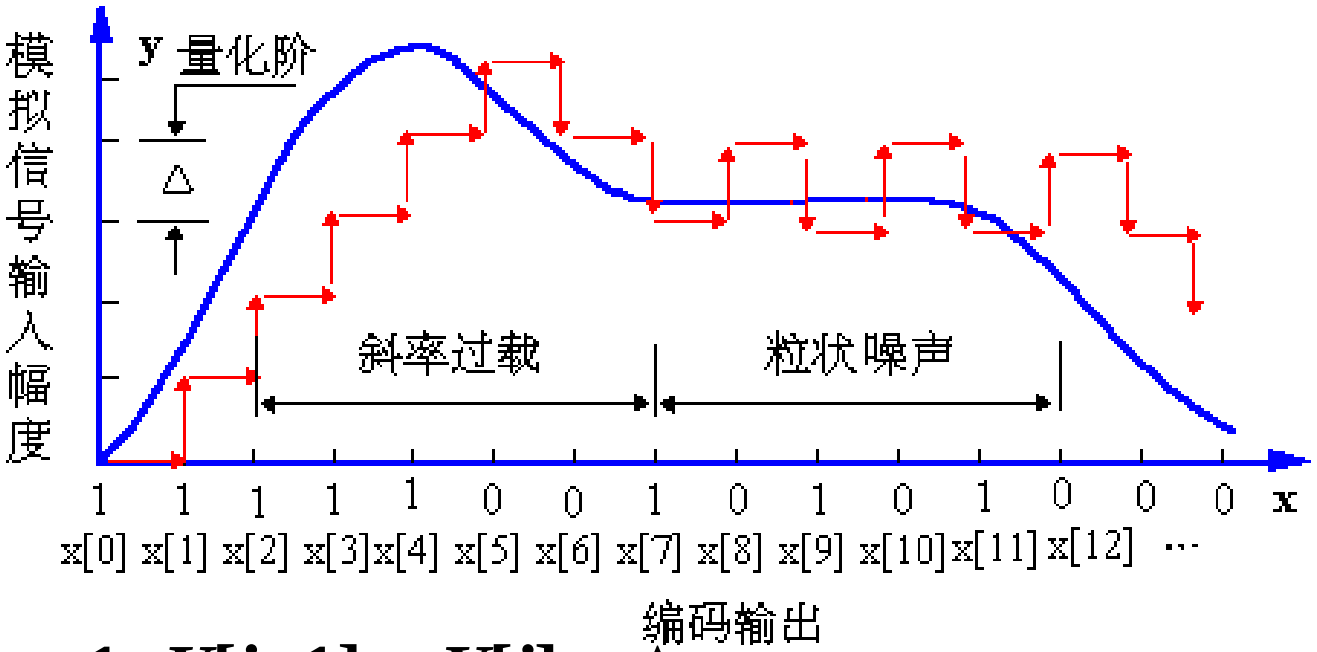
μ律：$F(x)=\dfrac{\ln(1+\mu x)}{\ln(\mu x)},\ \mu=255$

A律：$F(x)=\begin{cases}\dfrac{Ax}{1-\ln x} & 0\le x\le 1/A\\ \dfrac{1+Ax}{1+\ln A} & 1/A<x\le 1\end{cases},\ A=87.6$



8.2.3 增量调制(Delta Modulation)

- 增量调制也称△调制DM(**d**elta **m**odulation)，它是一种预测编码技术。如果实际的采样信号与预测的采样信号之差的极性为“正”，则用“1”表示；相反则用“0”表示，或者相反。由于DM编码只须用1比特对话音信号进行编码，所以DM编码系统又称为“1比特系统”。



- $X[i] = 1: Y[i+1] = Y[i] + \Delta$
- $X[i] = 0: Y[i+1] = Y[i] - \Delta$

8.2.3 增量调制(Delta Modulation)

- 斜率过载现象：主要是反馈回路输出信号的最大变化速率受到量化阶大小的限制，因为量化阶的大小是固定的
- 粒状噪声：在输入信号缓慢变化部分，即输入信号与预测信号的差值接近零的区域，增量调制器的输出出现随机交变的“0”和“1”。

- 在输入信号变化快的区域，斜率过载是关心的焦点，而在输入信号变化慢的区域，关心的焦点是粒状噪声。为了尽可能避免出现斜率过载，就要加大量化阶Δ，但这样做又会加大粒状噪声；相反，如果要减小粒状噪声，就要减小量化阶Δ，这又会使斜率过载更加严重。这就促进了对自适应增量调制ADM(**a**daptive **d**elta **m**odulation)的研究。

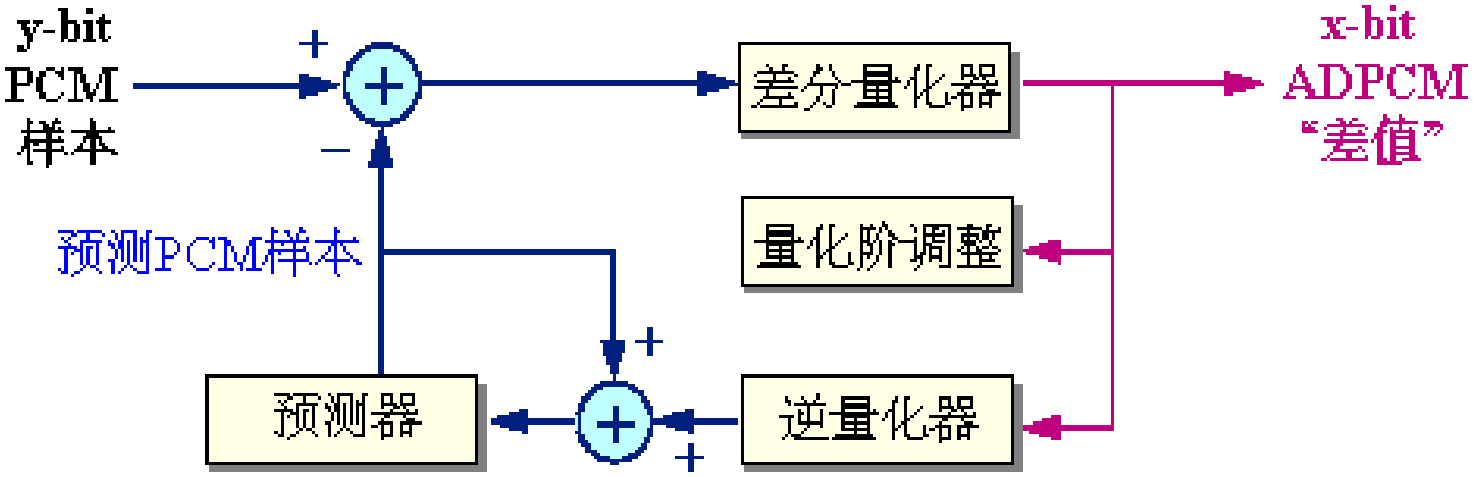
8.2.4 自适应增量调制(ADM)

- 目的：为了使增量调制器的量化阶Δ能自适应，也就是根据输入信号斜率的变化自动调整量化阶Δ的大小，以使斜率过载和粒状噪声都减到最小

- CVSD（连续可变斜率增量调制）：检测到斜率过载时开始增大量化阶Δ，而在输入信号的斜率减小时降低量化阶Δ
- **Sigma-DM:** 在DM编码器中的量化器之前包含一个数字积分器，然后对信号的积分而不是信号本身进行编码，使得在感兴趣带宽内的噪声减小

8.2. ADPCM

- ①利用自适应的思想改变量化阶的大小，即使用小的量化阶(step-size)去编码小的差值，使用大的量化阶去编码大的差值
- ②使用过去的样本值估算下一个输入样本的预测值，使实际样本值和预测值之间的差值总是最小。



例：IMA-ADPCM

- IMA：Interactive Multimedia Association，一个计算机硬件制造商和软件制造商的联盟，其目标是开发制定多媒体应用的标准

- IMA-ADPCM：
  - 目标：简单公用、快速
  - 将16bit的样本经过预测和量化后变成4bit，压缩因子为4
- 简单快速：
  - （非自适应）预测：只用前一个样本来预测，且系数为1
  - 自适应量化：量化阶的调整通过查两个表实现

8.3 声码编码(参数编码)

- 用一个数学模型表示信源，该模型取决于一些参数
- 编码器根据输入信号计算模型参数，然后对模型参数进行编码
- 解码器接收到模型参数，再利用数学模型重建原始数据

对其他信源，如果能找到一个合适数学模型及其参数计算方式，也可用类似思想进行编码。这一大类方法称为信源模型编码技术。

分析—综合编码

- 本质是随机信号的参数建模问题——谱估计问题
- 回顾：维纳滤波器是用来从噪声中提取信号问题的一种过滤（滤波）方法
  - X(n)=s(n)+v(n)　(s(n)是信号，v(n)是噪声）
  - 希望x(n)经过线性时不变系统后得到的y(n)尽可能接近于s(n)
- 维纳滤波的三种类型
  - 滤波（过滤）：利用直到当前时刻的随机过程的观察值，得到当前信号值的估计
  - 平滑（内插）：利用直到当前时刻的随机过程的观察值，得到过去某个时刻信号的估值
  - 预测（外推）：利用直到当前时刻的随机过程的观察值，得到将来某个时刻信号的估值。
    - 在最小均方误差准则下，线性预测器的误差序列是白噪声
    - 一个平稳的随机过程序列可以分解成一个可预测序列和一个不可预测序列

平稳随机信号的参数建模法

平稳随机信号的参数模型

- **时间序列**可模型化为**白噪声**序列作用于**数字滤波器**$H(z)$的输出。
- $H(z)$通常为**有理分式**的形式：

$$H(z)=G\frac{1+\sum_{l=1}^{q}b_l z^{-l}}{1-\sum_{i=1}^{p}a_i z^{-i}}$$

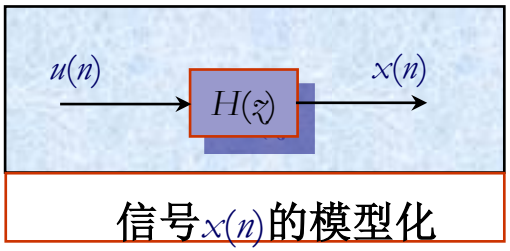
，**模型参数**：a_i, b_i ——**系数**，

，G ——**增益因子**。

- 信号$x(n)$的**模型化**。

$x(n)$、$X(z)$ —— 模型化的**信号**和其 z 变换，

$u(n)$、$U(z)$ —— 模型的**激励**和其 z 变换。



信号$x(n)$的模型化

◆ z **域关系式**：$X(z)=H(z)U(z)$

◆**时域关系式**：

- **物理意义**：$x(n)$由其**过去值**及**模型输入**的**线性组合**来**预测**得到。

$$x(n)=\sum_{i=1}^{p}a_i x(n-i)+G\sum_{l=0}^{q}b_l u(n-l);\quad b_0=1$$

- $x(n)$ 为**零均值**的随机信号时，

  系统的**输出、输入**关系可用**相关函数**或**功率谱**来表征：

$$R_{xx}(z)=H(z)H(z^{-1})R_{uu}(z)$$

式中，$R_{xx}(z)$ —— **信号** $x(n)$ 的**自相关函数**的 z 变换；

$R_{uu}(z)$ —— **输入** $u(n)$ 的**自相关函数**的 z 变换。

◆通常，$u(n)$ 是**零均值**、σ_u^2 **方差白噪声序列**，因此有：

$$R_{uu}(n)=\sigma_u^2\delta(n) \Longrightarrow R_{uu}(z)=\sigma_u^2$$

◆**假设** $\sigma_u^2=1$，则本页第一式的变换写成**功率谱**形式，有：

$$\left|X(\mathrm{e}^{\mathrm{j}\omega})\right|^2=\left|H(\mathrm{e}^{\mathrm{j}\omega})\right|^2$$

◆上式表明，信号 $x(n)$ 的**功率谱**完全由**滤波器的幅频响应**决定。

即系统 $H(z)$ 确实可以用来**模型化**信号 $x(n)$。

- 上式是用**模型参数**分析法**估计随机信号**的**理论依据**。

➢ **信号模型**分三种（按滤波器的有理分式）：

◆ **ARMA模型**：传递函数含有**极点**和**零点**（**零极点**模型）。
（**自回归–滑动平均模型**）
ARMA模型产生的序列称为**ARMA**过程序列。

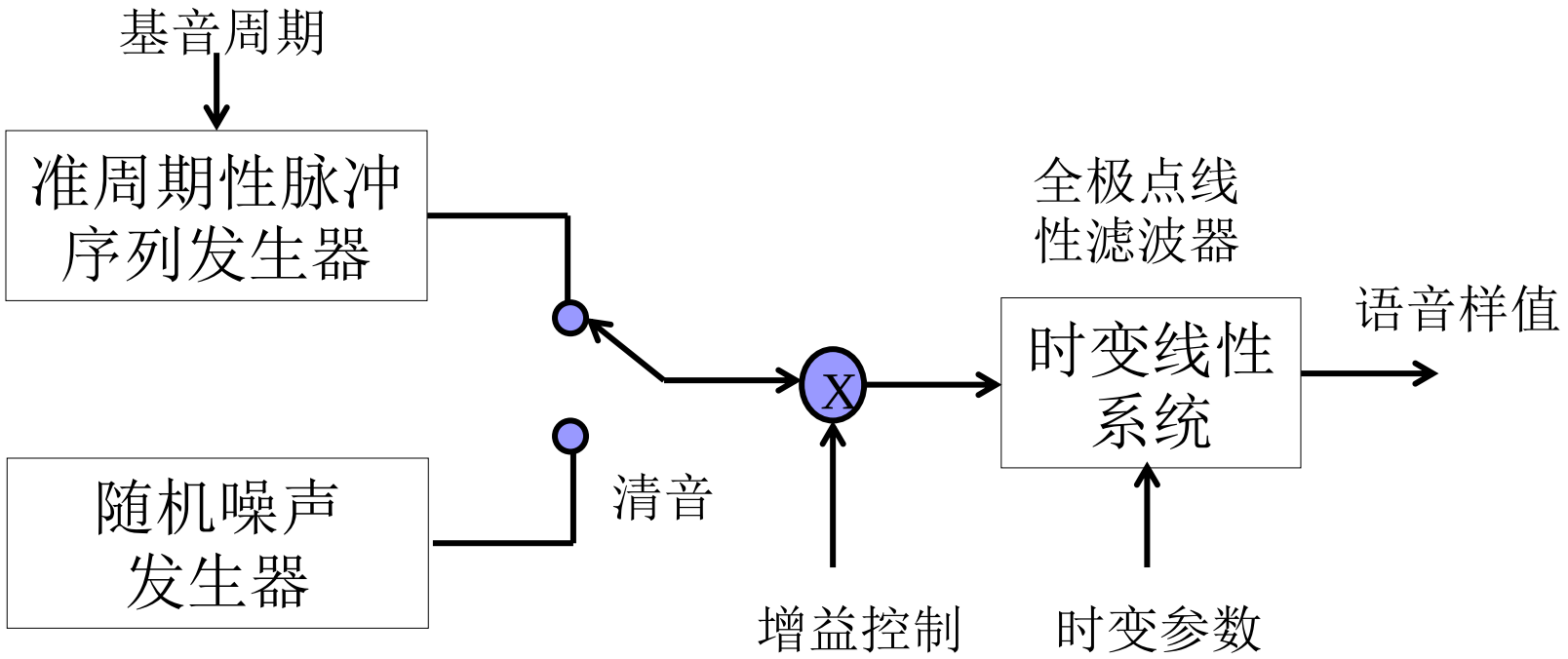
◆ **AR模型**：传递函数的**分子多项式**为常数（**全极点**模型）。
（**自回归模型**）
输出只取决于过去的信号值。
AR模型产生的序列称为**AR**过程序列。

◆ **MA模型**：传递函数的**分母多项式**为常数（**全零点**模型）。
（**滑动平均模型**）
输出只由模型的输入来决定。
MA模型产生的序列称为**MA**过程序列。

➢ **ARMA**模型是**AR**模型和**MA**模型的**混合结构**。

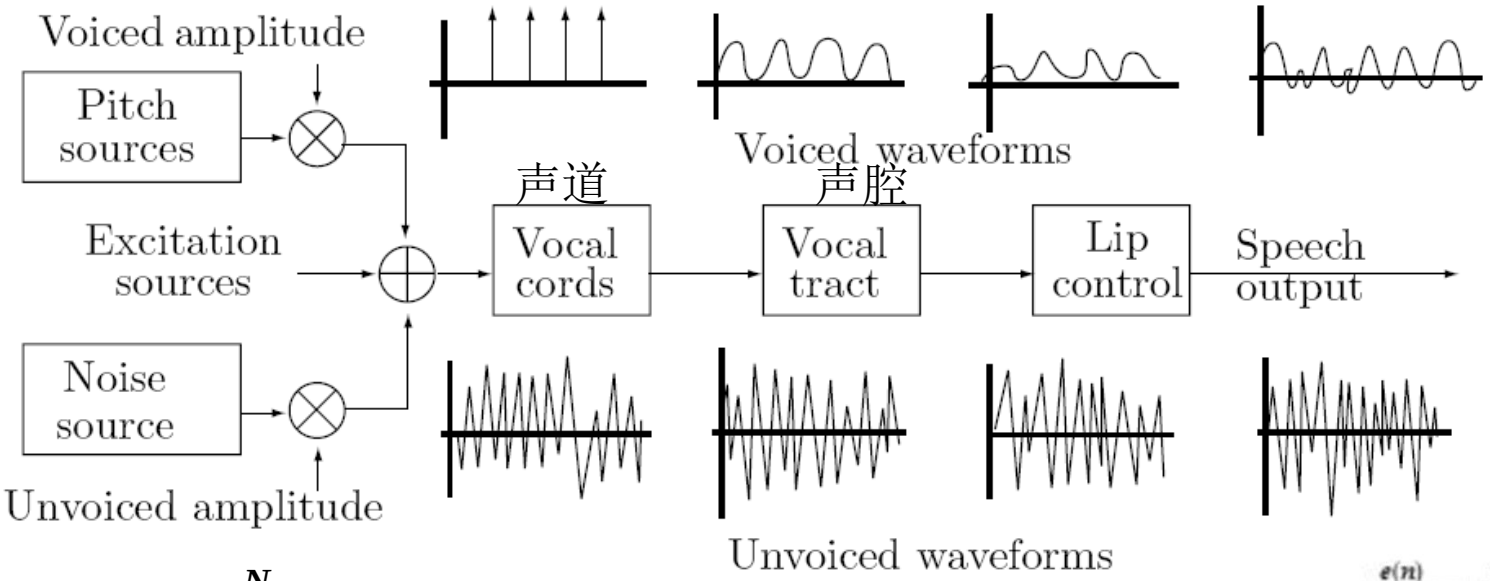
8.3 声码编码(参数编码)

- 描述语音产生过程的离散时间信号模型



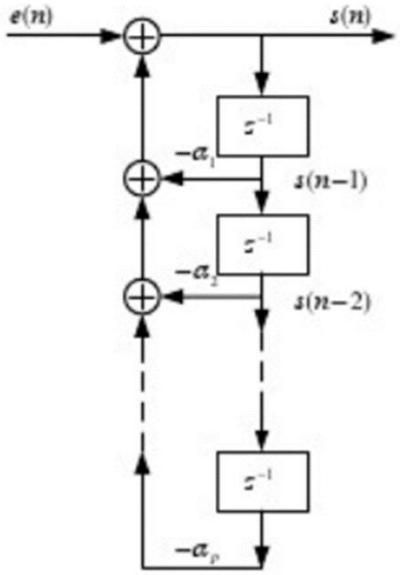
- 激励模型(显然，获得准确的清浊音分析结果需要耗费计算资源)
  浊音：周期脉冲信号
  清音：随机噪声
- 声道模型：M阶全极点滤波器/**AR模型** → 线性预测

8.3 声码编码(参数编码)



(a) 真实模型

$$x_k = e_k + \sum_{i=1}^{N} a_i x_{k-i}$$



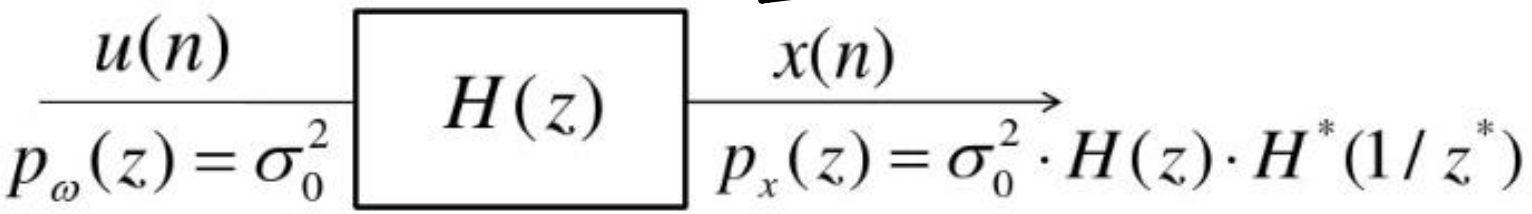
可解释为用信号e_k激励全极点滤波器

$$H(Z) = \frac{1}{1 - \sum_{i=1}^{M} a_i Z^{-i}}$$

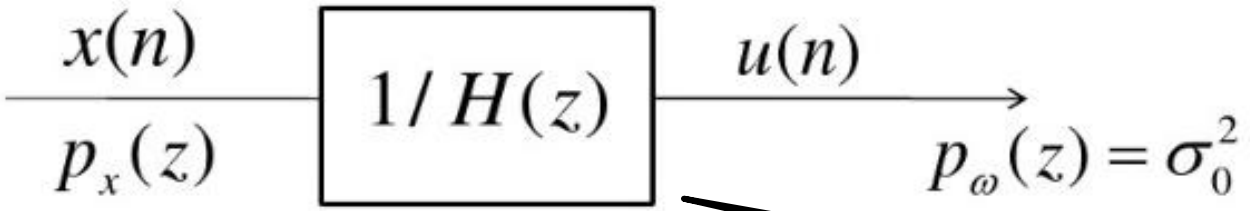
假定当前语音样本能由过去语音样本与驱动项加权求和表示

清音的白噪声和浊音的准周期性脉冲都是驱动项

参数编码系统

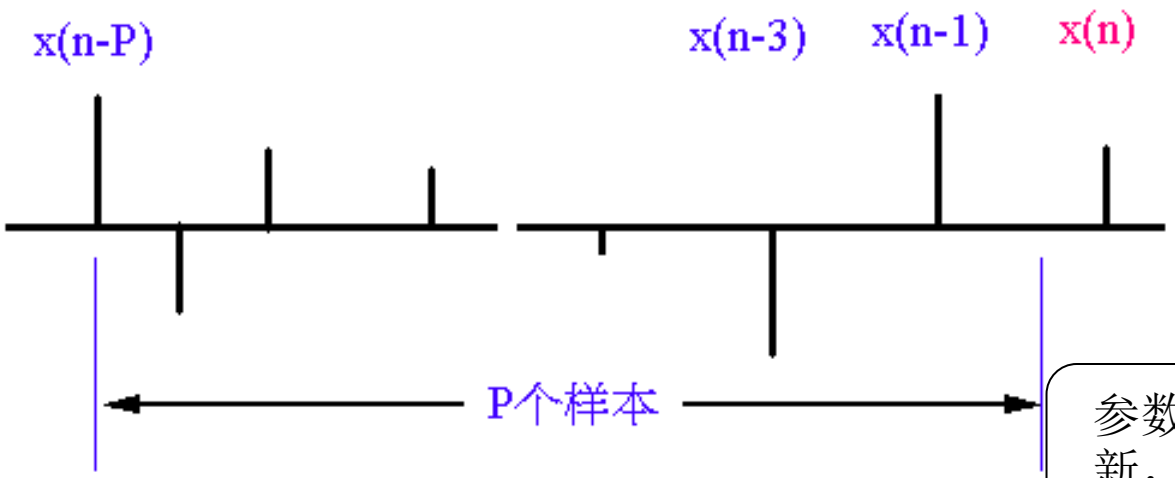


随机信号的新息表示

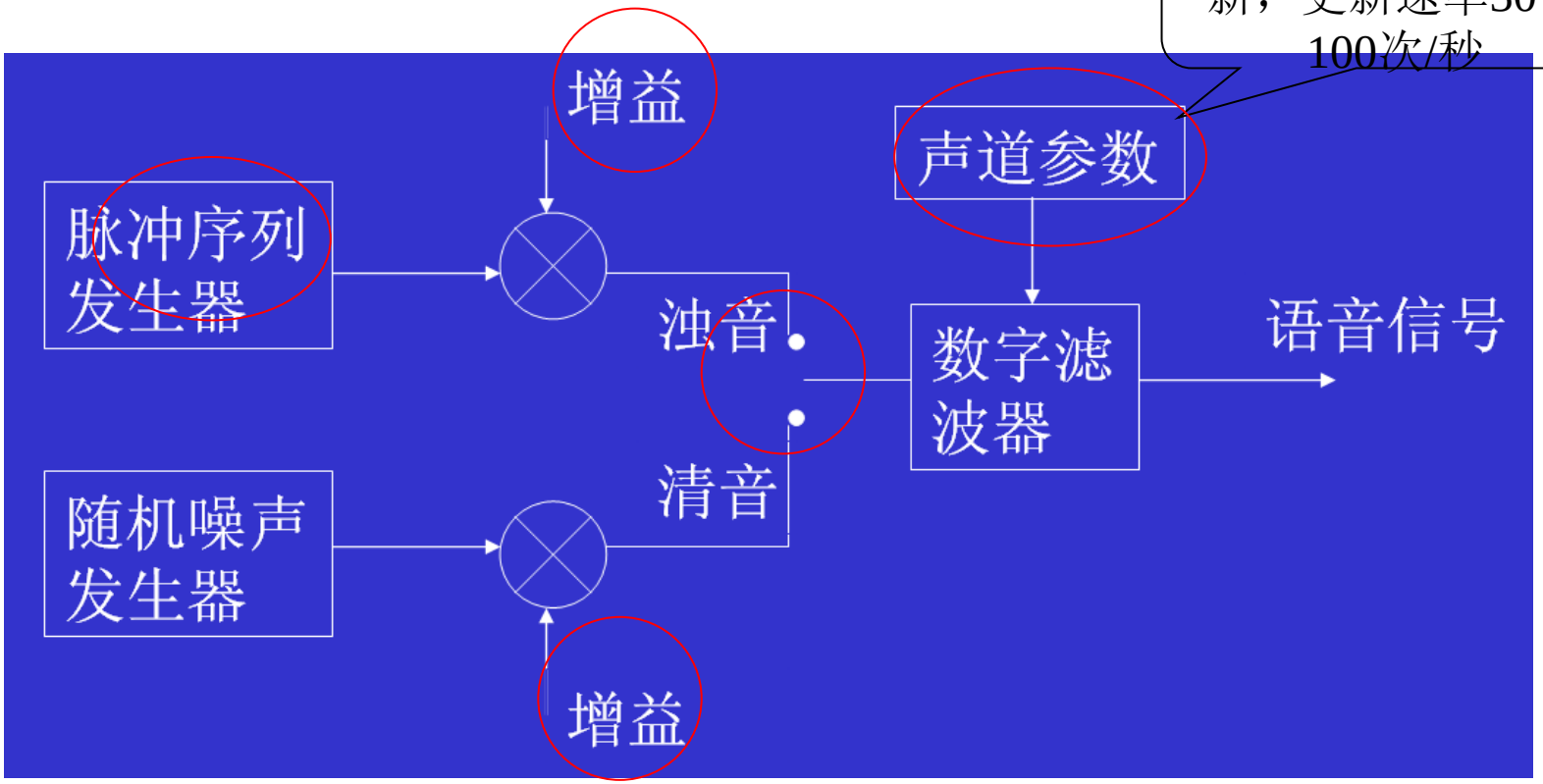


随机信号的白化

差分预测编码系统



参数随时间逐帧更新，更新速率30～100次/秒



8.3.1 LPC语音合成模型

- 从话音波形信号中提取生成话音的参数，使用这些参数通过话音生成模型重构出话音。
- 在话音生成模型中，声道被等效成一个随时间变化的滤波器，叫做时变滤波器(time-varying filter)，它由白噪声—清音话音段激励，或者由脉冲串——浊音段激励。
- 因此需要传送给解码器的信息就是滤波器的规格、发声或者不发声的标志和有声话音的音节周期，并且每隔10～20 ms更新一次。
- 声码器的模型参数（基音周期、清/浊音判别）既可使用时域的方法也可以使用频域的方法确定，这项任务由编码器完成。

8.3.1 LPC语音合成模型

- 通过分析话音波形来产生声道激励和转移函数的参数，对声音波形的编码实际就转化为对这些参数的编码，这就使声音的数据量大大减少。
- 在接收端使用LPC分析得到的参数，通过话音合成器重构话音。合成器实际上是一个离散的随时间变化的时变线性滤波器，它代表人的话音生成系统模型。
- 时变线性滤波器既当作预测器使用，又当作合成器使用。分析话音波形时，主要是当作预测器使用，合成话音时当作话音生成模型使用。随着话音波形的变化，周期性地使模型的参数和激励条件适合新的要求。

LPC分析的频域解释

- 由于语音产生模型中全极点滤波器的频率特性主要反映了声道的共振特性，而语音信号的LPC系数就是语音信号产生模型中全极点合成滤波器$H(z)$的分母多项式的系数，因此当根据一帧语音的取样值计算出语音信号的LPC系数后，只要将 $z = e^{j\omega}$ 代入H（z）进行计算，就意味着求得了这帧语音信号产生模型的频率特性。
- 本质：参数模型功率谱估计问题！

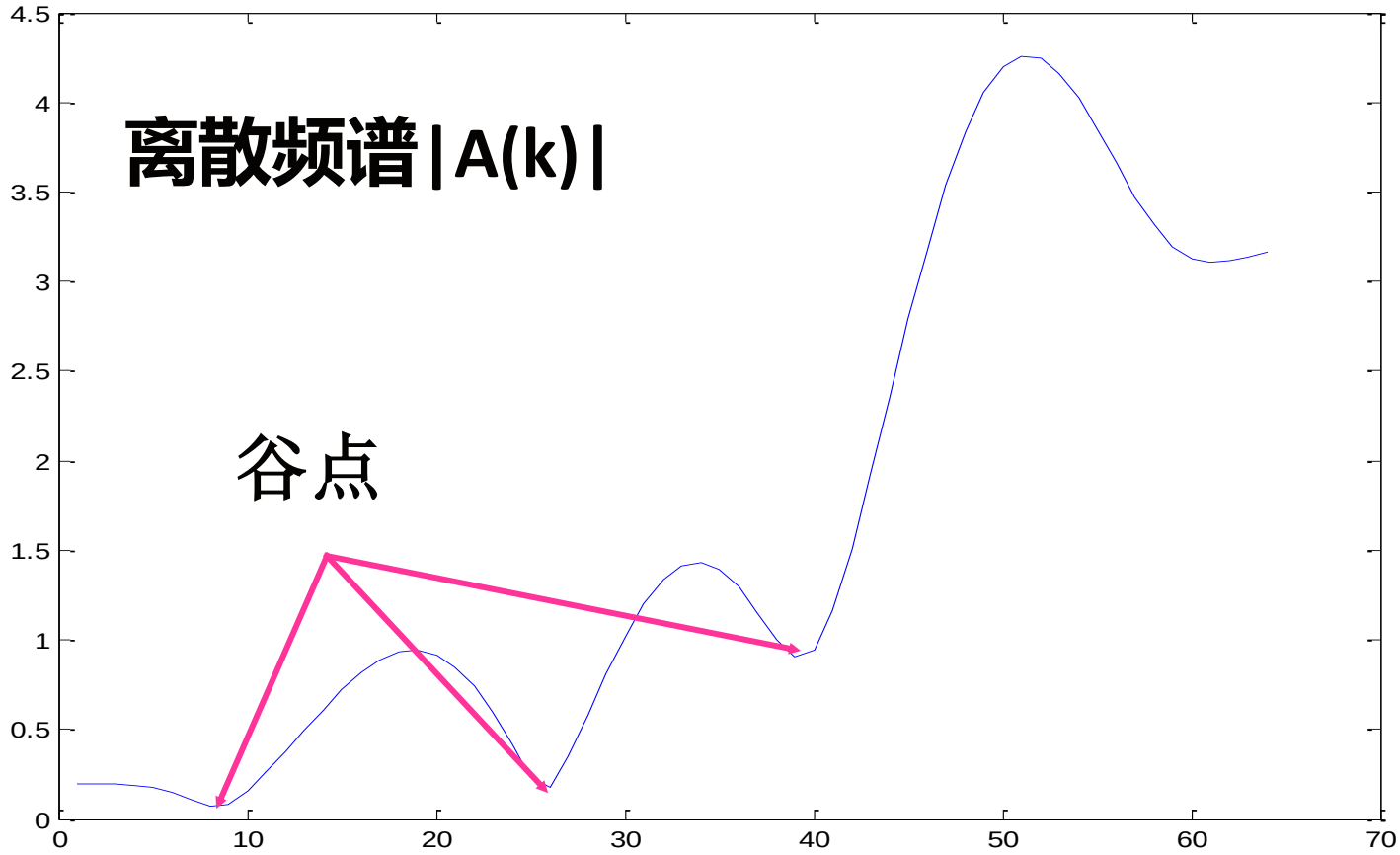
利用线性预测系数求共振峰，离散频谱|A(k)|的谷点就是共振峰的位置。通过求A(z)多项式的系数序列{1,a_1,a_2,…a_p}的DFT，就可以得到|A(k)|。

$$H(z)=\frac{G}{1+\sum_{k=1}^{p}a_k z^{-k}}$$

$|H(e^{jw})|$ 的峰值对应共振峰

$$A(z)=1+\sum_{k=1}^{p}a_k z^{-k}$$

$|A(e^{jw})|$ 的谷点对应共振峰



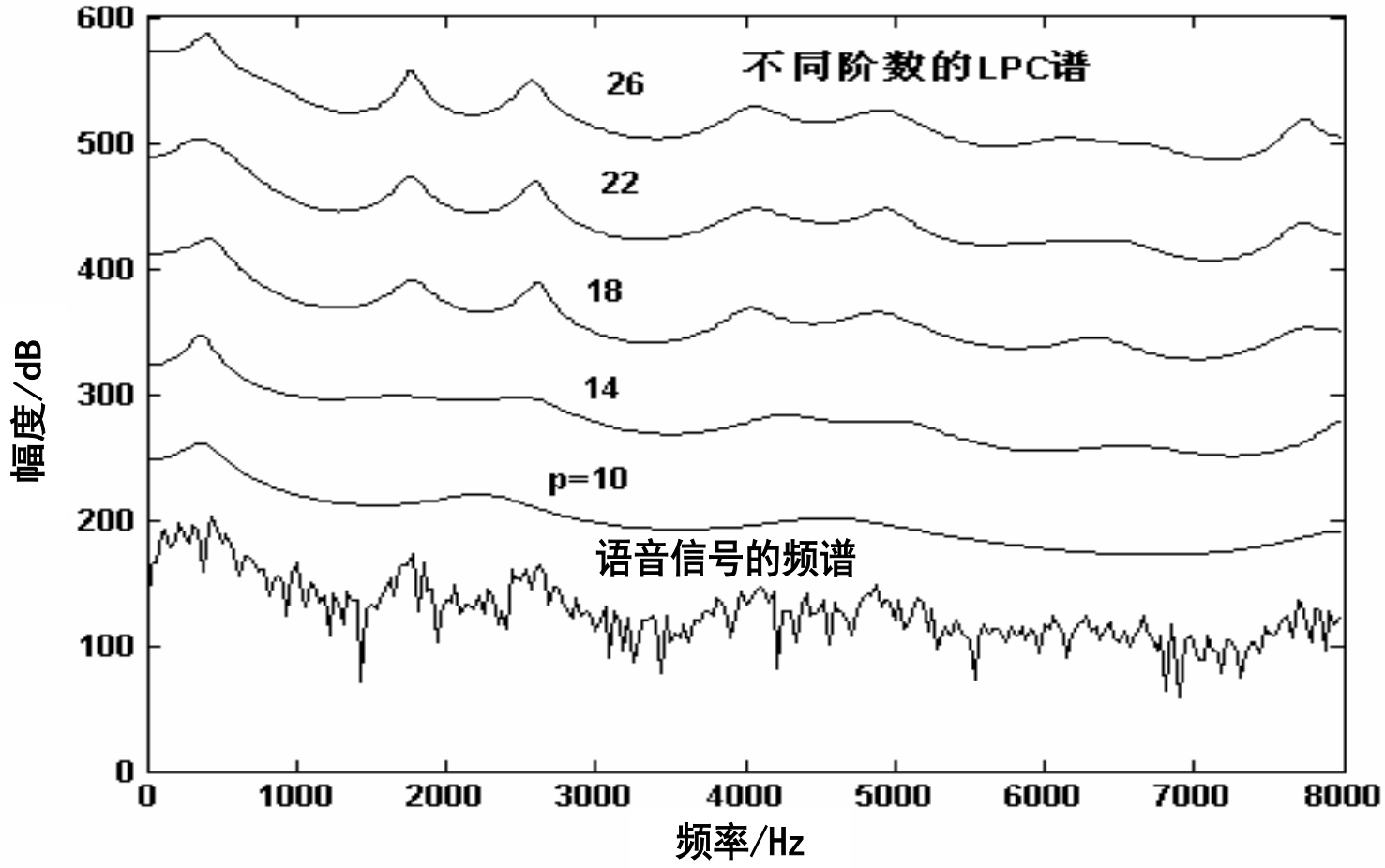
{1,-1.45,0.9,-0.45,-0.12,0.36,-0.30,0.39,0.12,-0.34,0.06}

p=10

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{i=1}^{p} a_i e^{-j\omega i}} = \frac{G}{A(e^{j\omega})}$$

- **LPC分析可以看成是对语音信号短时谱进行估计的一种有效方法。**
- **在语音产生模型中，语音的功率谱等于激励源功率谱与全极点合成滤波器频率特性模的平方的乘积，而激励源是准周期冲击序列或白噪声，其功率谱是平坦的。所以语音的功率谱主要由全极点滤波器的特性来决定。**

线性预测分析的阶数*p*可有效控制所得谱的平滑度：



8.3.1 LPC语音合成模型

LPC参数

- 1. 参数 *V* (voiced)：对应声带振动； 表示为无声声音（unvoiced）；
  - 可以通过检测帧内是否有特定的主频率实现
- 2. *T*：声带振动周期
- 3. *G* (gain)：对应声音强度
- 4. 线性预测器的系数 $\hat{x}_k = \sum_{i=1}^{10} w_i x_{k-i}$

- LPC模型可用13元组表示：**A** = (a_1, a_2, . . . , a_{10},*G*, *V/UV*,*T*)

8.3.1 LPC语音合成模型

- 模型假设参数集合**A**在20 ms内保持平稳，因此每20 ms更新一次
  - 若取样频率为8 kHz，则20 ms 内共有160个样本。模型根据160个样本计算集合**A**中的13个参数的值，并写入压缩流；然后进行下一个20 ms
- 线性预测器的系数 w_i 的确定：
  - 根据一帧的160样本

$$\hat{x}_k = \sum_{i=1}^{10} w_i x_{k-i}$$

$$\begin{bmatrix} R(0) \\ R(1) \\ \cdots \\ R(10) \end{bmatrix} = \begin{bmatrix} R(0) & R(1) & \cdots & R(9) \\ R(1) & R(0) & \cdots & R(8) \\ \cdots & \cdots & \cdots & \cdots \\ R(10) & R(8) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_{10} \end{bmatrix}, \quad R(i) = \sum_{j=1}^{160-j} x_j x_{j+i}$$

8.3.1 LPC语音合成模型

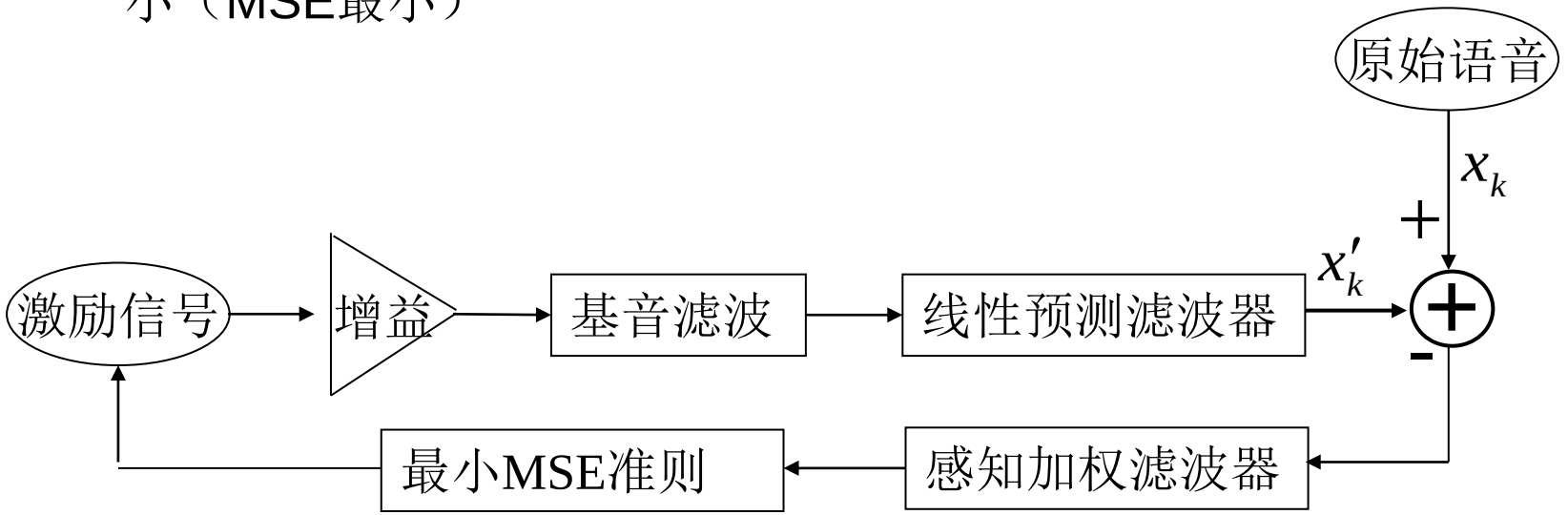
- 其他参数的确定：
  - 若160个样本呈现周期性，则*T* 为基音周期，且1-bit的参数 *V/UV* 设置为 *V* 。
  - 若160个样本没有呈现良好的周期性，则则*T* 保持不变，且参数 *V/UV* 设置为 *UV*。
  - *G* 对应声音强度，由最大的样本决定。
- 例：2.4-kbps的LPC声码器将10个参数w_i进一步转化成10个线性频谱对（ LSP ）参数ω_i，且$0 < \omega_1 < \omega_2 ... < \omega_{10} < \pi$，对该10个参数的编码位数为

ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
3	4	4	4	4	3	3	3	3	3

  - *G* 编码为7比特，周期*T*量化为6比特，*V/UV*参数为1比特
  - 13个参数为48比特，每帧20ms，每秒共50帧，50 ×48 = 2.4 kbps.
  - 原始数据为8000 ×8bps，压缩比为：(8000 ×8) /2400 = 26.67

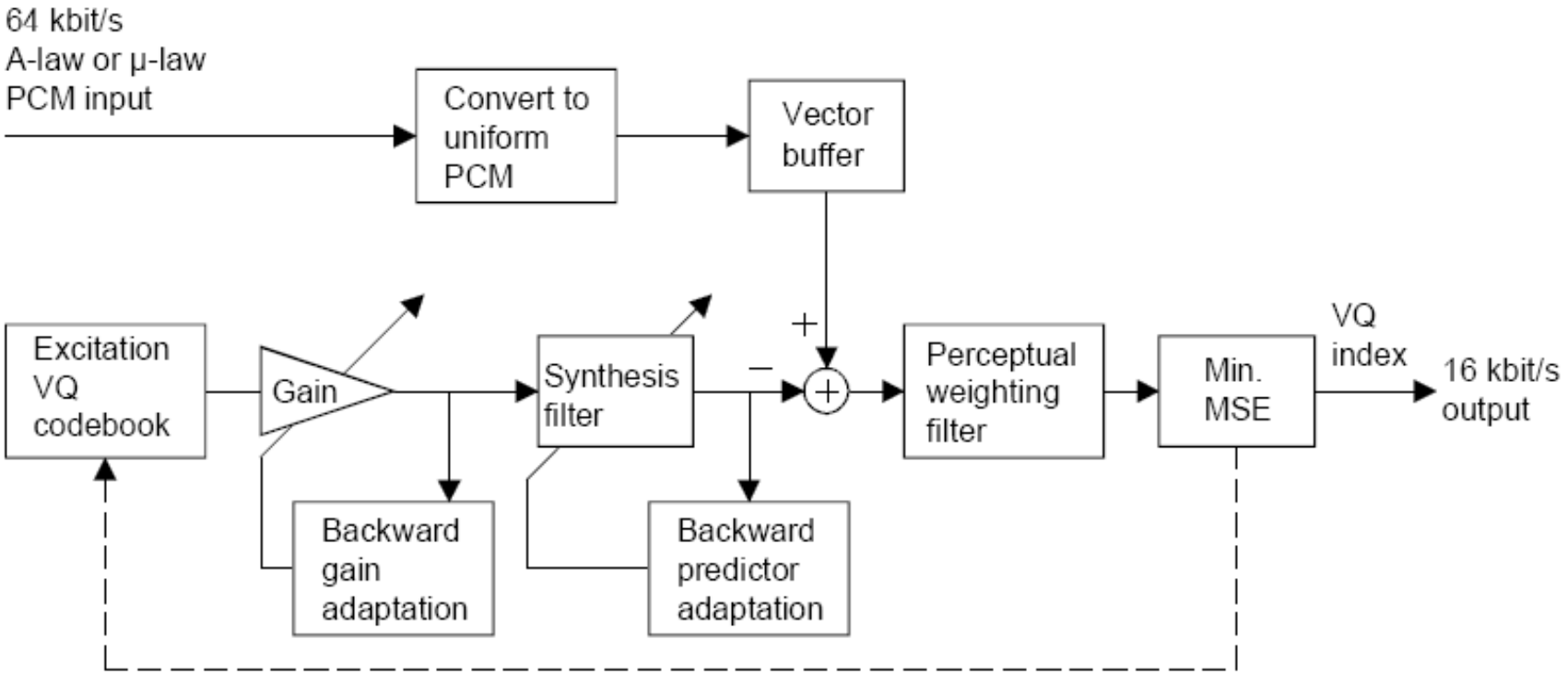
8.4 混合编码（ hybrid codecs ）

- 混合编码器结合了波形编码器和声源模型编码器的特点。最流行的混合编码为时域合成-分析AbS(**A**nalysis-**b**y-**S**ynthesis)算法
- 编码使用的声道线性预测滤波器模型与线性预测编码使用的模型相同，不使用两个状态(有声/无声)的模型来寻找滤波器的输入激励信号，而是寻找一种激励信号，使用这种信号激励产生的波形尽可能接近于原始话音的波形
  - 类似DPCM，引入负反馈，使得重构信号与原始信号之间的差值最小（MSE最小）



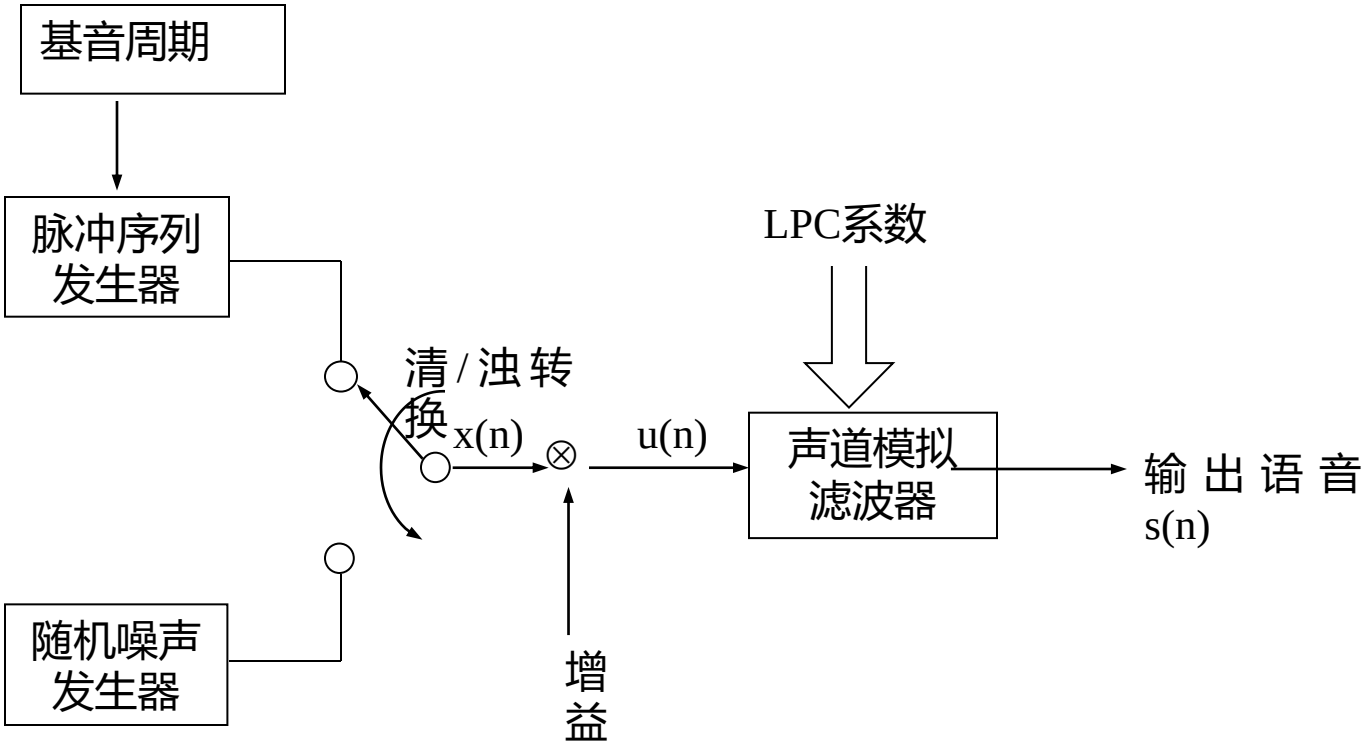
8.4 混合编码

- 码激励线性预测编码（CELP）是AbS编码的一种
  - 将线性预测和矢量量化相结合
  - 从典型激励矢量的码书中选择最佳的矢量作为语音信号表示，只输出码本的索引　　-使得MSE最小的矢量

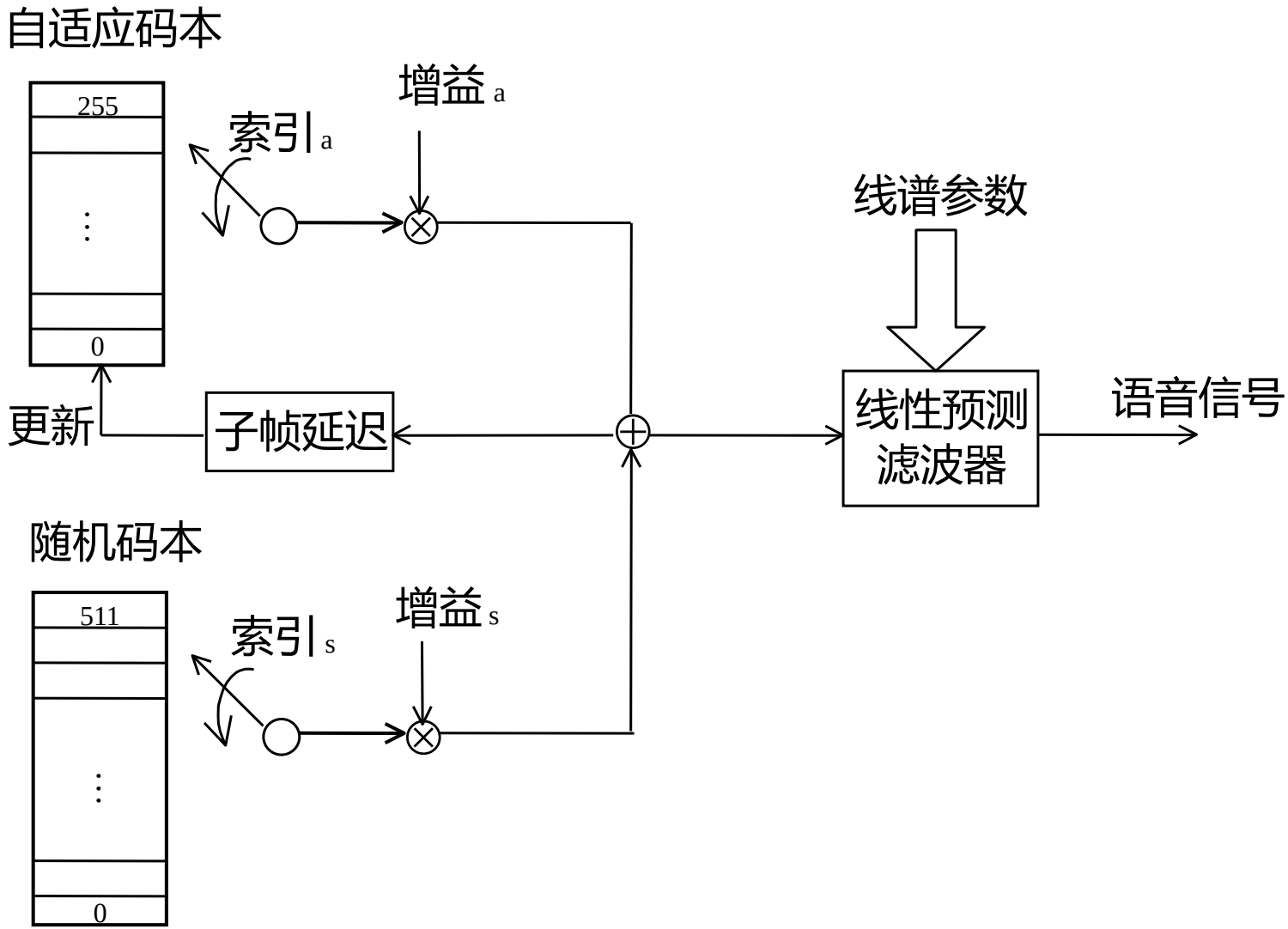


a) LD-CELP encoder

回顾：LPC 语音合成图



CELP 语音合成示意图



8.4 混合编码

改善语音的质量

- 对误差信号进行感觉加权，利用人类听觉的掩蔽特性来提高语音的主观质量；
- 用分数延迟改进基音预测，使浊音的表达更为准确，尤其改善了女性语音的质量；
- 使用修正的MSPE（最小平方预测误差）准则来寻找“最佳”的延迟，使得基音周期延迟的外形更为平滑；
- 根据长时预测的效率，调整随机激励矢量的大小，提高语音的主观质量；
- 使用基于信道错误率估计的自适应平滑器，在信道误码率较高的情况下也能合成自然度较高的语音。

8.4 混合编码

结论

- CELP算法在低速率编码环境下可以得到令人满意的压缩效果。
- 使用快速算法，可以有效地降低CELP算法的复杂度，使它完全可以实时地实现。
- CELP可以成功地对各种不同类型的语音信号进行编码，这种适应性对于真实环境，尤其是背景噪声存在时更为重要。

语音编码技术的发展方向

- CELP的码率较低，但复杂度高，可以在4.8kpbs左右的码率上获得较高质量的语音，是当今中低码率语音编码技术的主流技术之一

- 中低码率的语音编码的实用化，以及实用化过程中进一步降低码率和提高抗干扰能力、抗噪声能力
- 如何进一步降低码率
  - 目前已能在5kbps-6kbps的码率上获得高质量的重建语音，下一个目标是在4kbps的码率上获得短延时、高质量的重建语音
  - 更低码率（400-1200bps）的编码算法（CELP达不到此要求）
  - 引入新的分析技术，如高阶统计分析技术、非线性预测、多精度时频分析等技术。这些技术能进一步挖掘人耳的听觉屏蔽感知机制，从而在低码率语音编码研究上取得突破