多媒体技术基础

第三章 多媒体数据压缩

§3.2 音频数据的压缩标准

2024年10月27日

授课内容

- 第一部分 多媒体的计算
  - 第一章 多媒体计算机系统
  - 第二章 媒体处理技术
  - 第三章 多媒体数据压缩
- 第二部分 多媒体的存储
  - 第四章 多媒体数据的数字存储
- 第三部分 多媒体信息的分析与处理
  - 第五章 多媒体信息分析与处理
- 第四部分 多媒体网络应用
  - 第六章 实时多媒体通信

{ynh,cxh}@ustc.edu.cn

第三章 多媒体数据压缩

- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MPEG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

音频信号处理

- 应用范围
  - 无线电广播、电话、电视信号中的声音
  - 移动通信、卫星通信、音频文件
- 趋势
  - Analog signals → Digital signals
  - 如家用音响的Hi-Fi功放 → AV功放

{ynh,cxh}@ustc.edu.cn

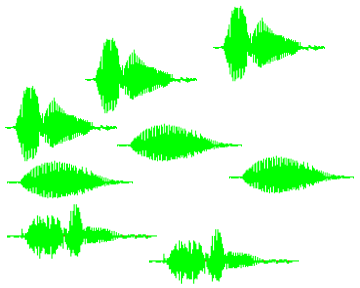
音频信号的冗余

- 时域信息的冗余度
  - 幅度的非均匀分布
  - 样本间的相关
  - 静音系数
- 频域信息的冗余度
  - 非均匀的长时功率谱密度
  - 语音特有的短时功率谱密度
- 人的听觉感知机理

{ynh,cxh}@ustc.edu.cn

幅度的非均匀分布

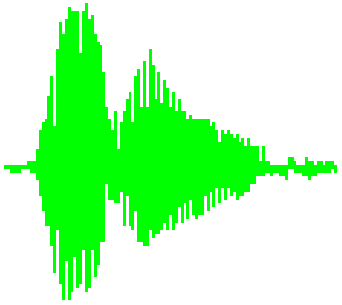
◆统计表明，语音中的小幅度样本比大幅度样本出现的概率要高。又由于通话中必然会有间隙，更出现了大量的低电平样本。此外，实际讲话信号功率电平也趋向于出现在编码范围的较低电平端。因此，语音信号取样值的幅度分布是非均匀的。

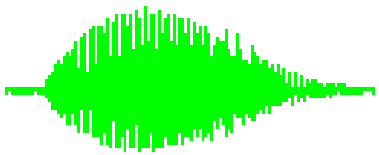


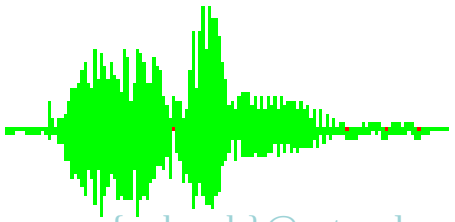
{ynh,cxh}@ustc.edu.cn

样本间的相关

◆对语音波形的分析表明，取样数据的最大相关性存在于邻近样本之间。当取样频率为8kHz时，相邻取样值间的相关系数大于0.85；甚至在相距10个样本之间，还可有0.3左右的数量级。如果取样速率提高，样本间的相关性将更强。因而根据这种较强的一维相关性，利用N阶差分编码技术，可以进行有效的数据压缩。



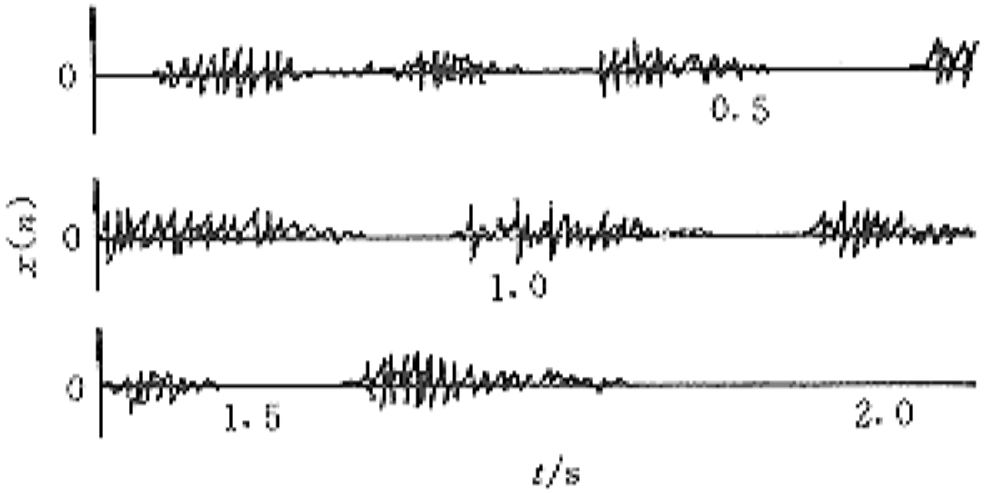




{ynh,cxh}@ustc.edu.cn

静止系数

◆两个人之间打电话，平均每人的讲话时间为通话总时间的一半，另一半时间听对方讲。听的时候一般不讲话，而即使是在讲话的时候，也会出现字、词、句之间的停顿。通过分析表明，话音间隙使得全双工话路的典型效率约为通话时间的40%（或静止系数为0.6）。显然，话音间隔本身就是一种冗余，若能正确检测出该静止段，便可“插空”传输更多的信息。



频域信息的冗余度

- 非均匀的长时功率谱密度
  - 在相当长的时间间隔内进行统计平均，可得到长时功率谱密度函数，其功率谱呈现强的非平坦性。从统计的观点看，这意味着没有充分利用给定的频段，或者说有着固有的冗余度。特别地，功率谱的高频能量较低，这恰好对应于时域上相邻样本间的相关性。
- 语音特有的短时功率谱密度
  - 语音信号的短时功率谱，在某些频率上出现峰值，而在另一些频率上出现谷值。这些峰值频率，也就是能量较大的频率，通常称为共振峰频率。此频率不止一个，最主要的是第一和第二个，由它们决定了不同的语音特征。另外，整个谱也是随频率的增加而递减。更重要的是，整个功率谱的细节以基音频率为基础，形成了高次谐波结构。这都与电视信号类似，仅有的差异在于直流分量较小。

{ynh,cxh}@ustc.edu.cn

两种思路

- 用比特串表示的声音的两种思路
  - 用一个“阶梯化”的波形尽可能精确地去模拟一个真实的声音波形（采样频率、样本精度）
  - 用基本信号尽可能逼真地合成一个模拟世界的声音

- WAV文件和MIDI文件对声音信号的记录是人们对“媒体信息”进行存储和再现的两种不同思路

- 人们在描述声音信号的时候采用了这样两种思想，同样在音频编码方面也存在类似的两种思想

{ynh,cxh}@ustc.edu.cn

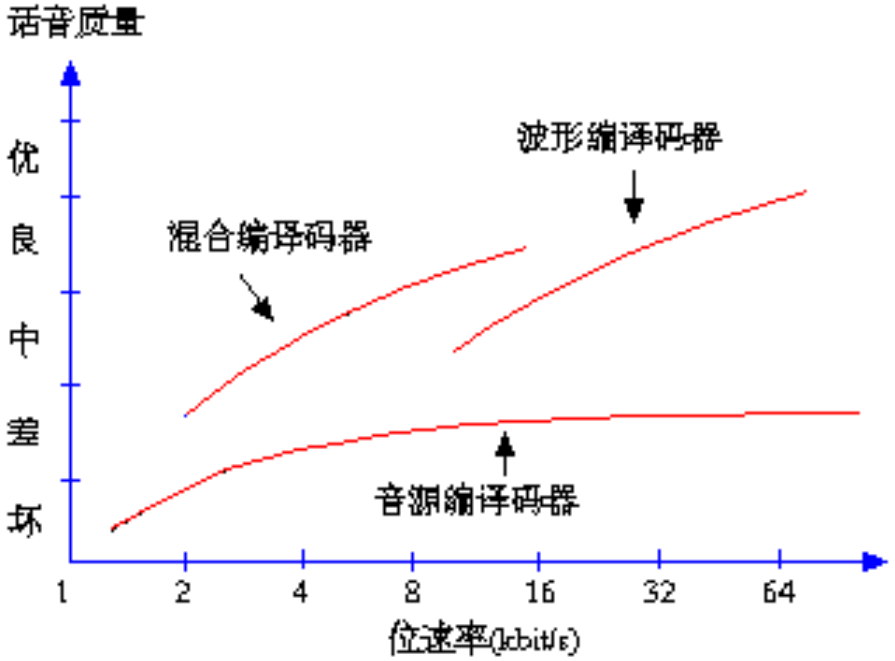
第三章 多媒体数据压缩

- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
    - 波形编译码器
    - 音源编译码器
    - 混合编译码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MPEG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

话音编译码器的分类

- 波形编码(waveform coding)
  - 不利用声音的任何知识，数据率较高，实现简单
- 参数编码(parametric coding)
  - 从声音的波形中提取生成话音的参数，数据率很低，实现复杂
- 混合编码(hybrid coding)
  - 混合编码：以上两种思想的结合



{ynh,cxh}@ustc.edu.cn

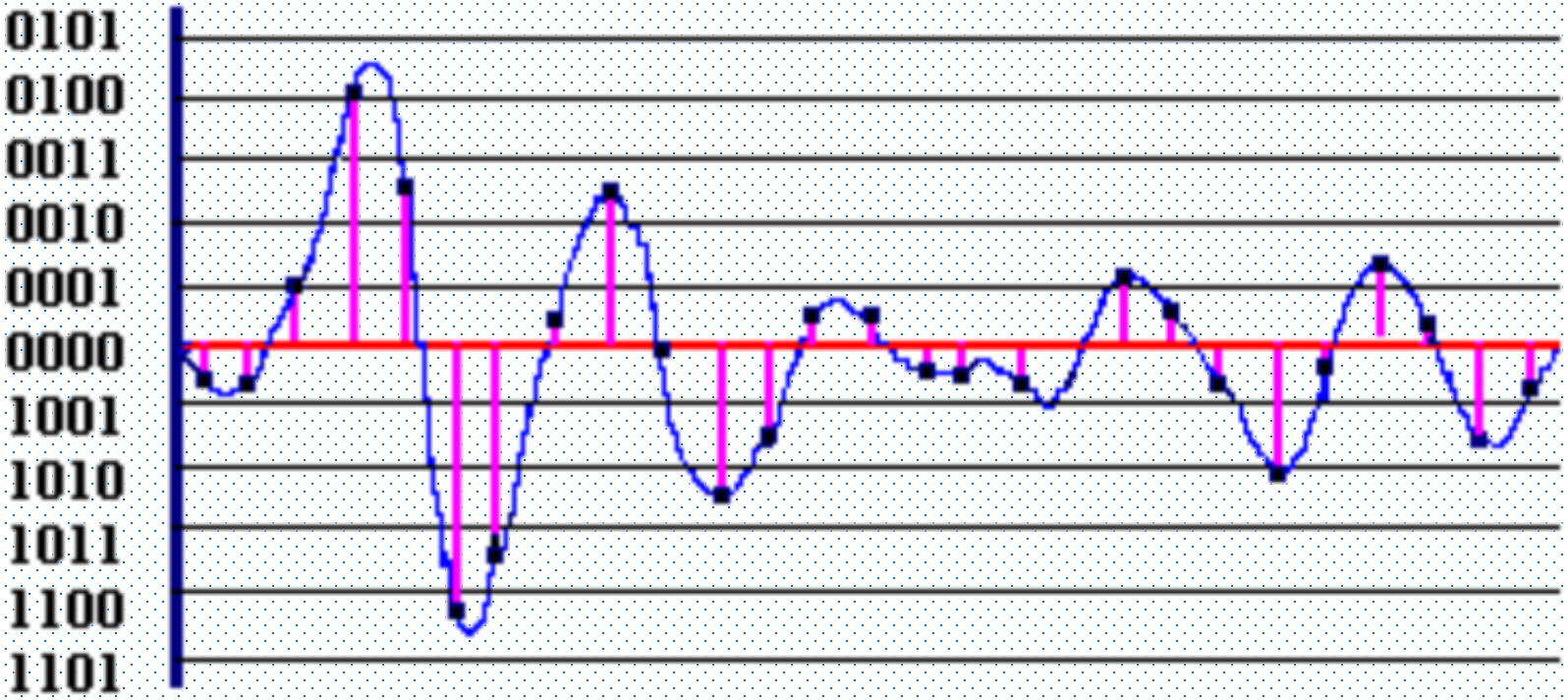
波形编译码器

- 波形编译码的想法
  - 不利用生成话音信号的知识产生而是产生一种重构信号，重构信号的波形和原始话音波形尽可能一致，这种编译码器的复杂程度低。
- 波形编码代表
  - PCM（脉冲编码调制）

{ynh,cxh}@ustc.edu.cn

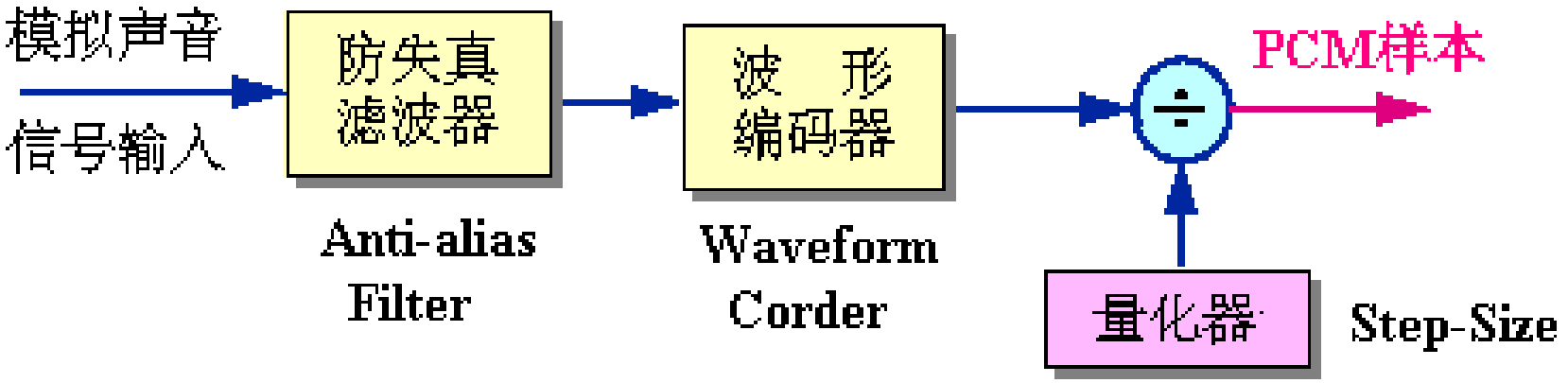
声卡的工作原理

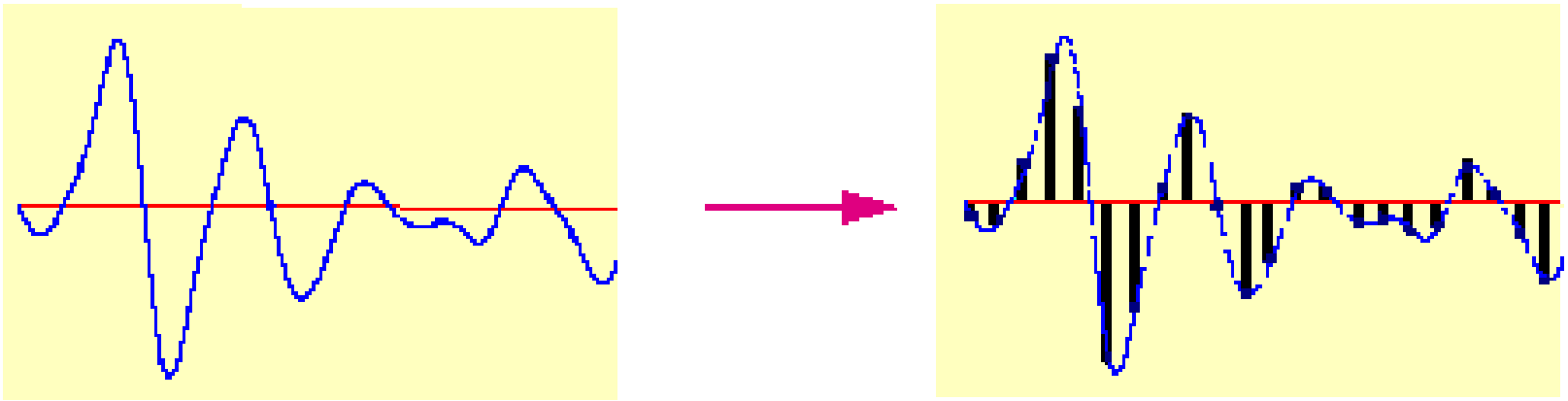
Review



采样 → 量化 → 编码

脉冲编码调制(PCM)

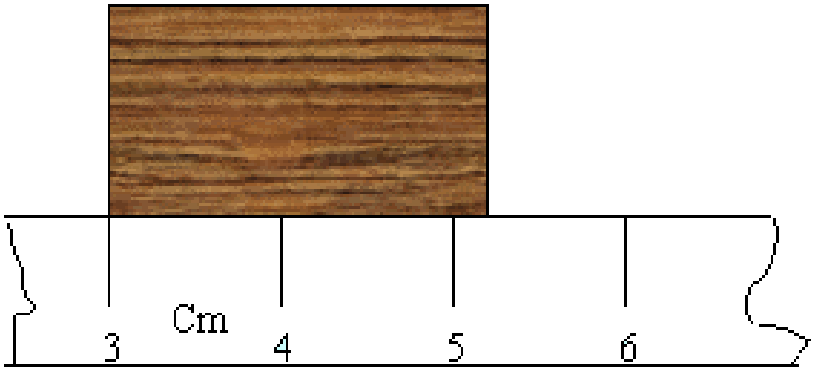


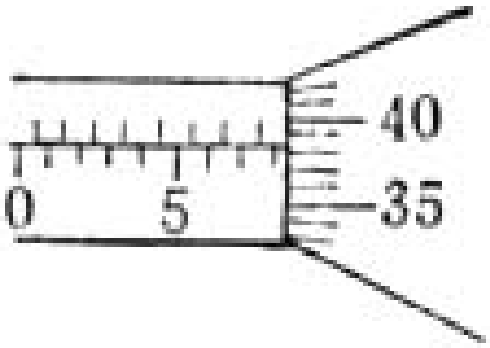


{ynh,cxh}@ustc.edu.cn

量化误差的概念

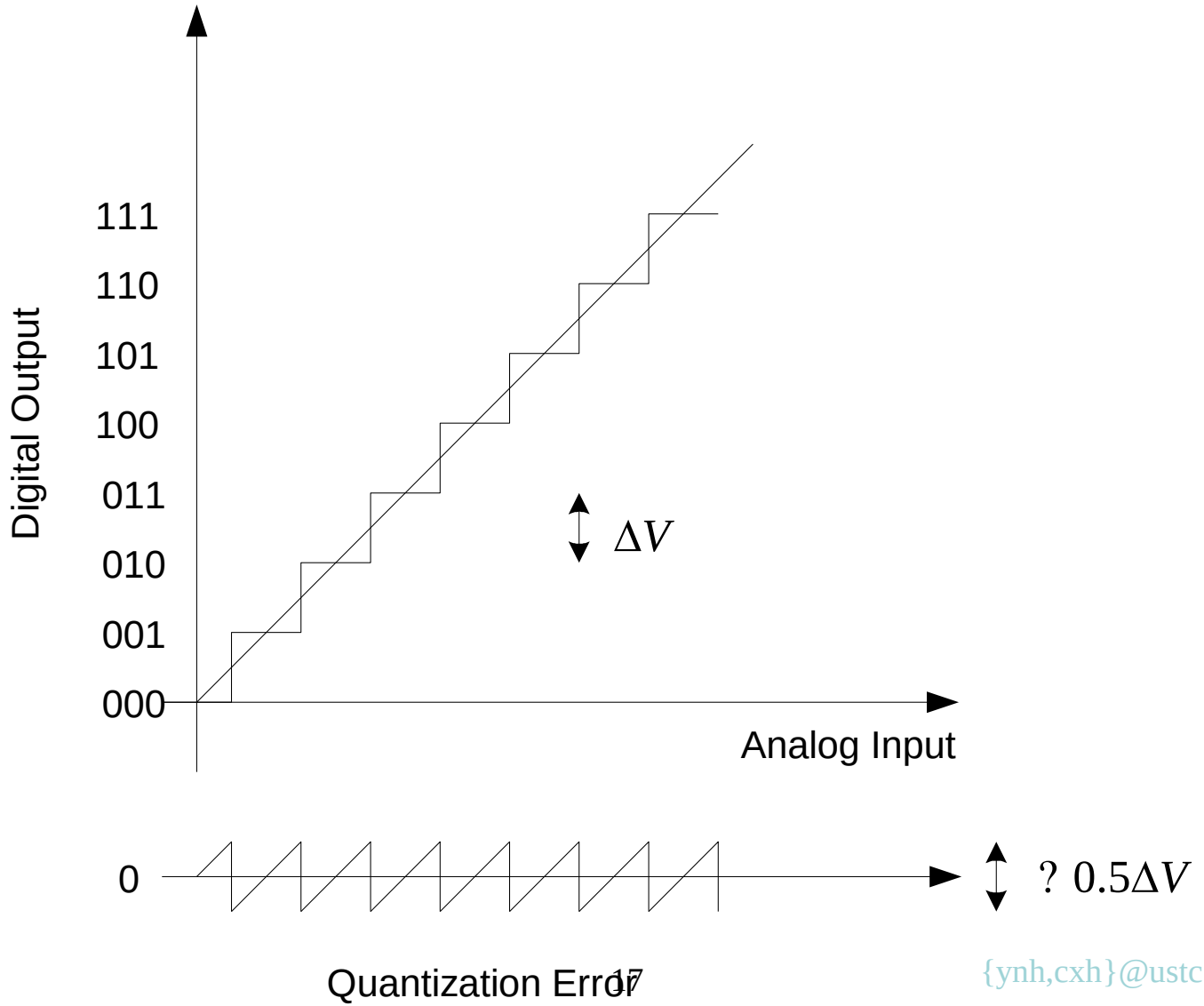
◆一道简单的概率计算题：某仪器表盘的刻度单位是0.2，读取刻度时选取偏差最小的刻度。请计算利用该仪器读取测量数值的误差小于0.04的概率是多大？误差大于0.05的概率是多大？





{ynh,cxh}@ustc.edu.cn

3bit量化过程中量化误差示意



{ynh,cxh}@ustc.edu.cn

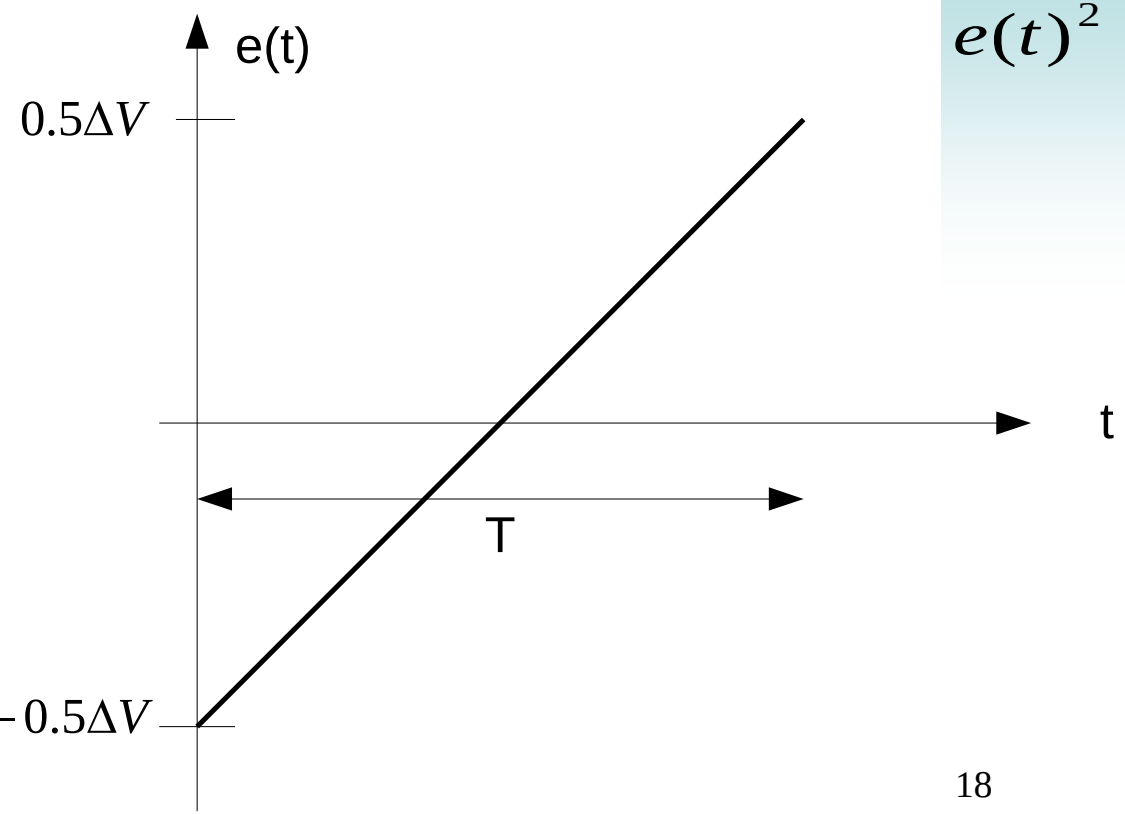
量化噪声

$$e(t) = mt + c$$

$$m = \frac{\Delta V}{T}$$

$$c = -\frac{\Delta V}{2}$$

$$e(t) = \frac{\Delta V}{T}t - \frac{\Delta V}{2}$$



$$e(t)^2 = \left(\frac{\Delta V}{T}t - \frac{\Delta V}{2}\right)^2$$

$$= \frac{\Delta V^2 t^2}{T^2} - \frac{\Delta V^2 t}{T} + \frac{\Delta V^2}{4}$$

$$RMS = \sqrt{\frac{1}{T}\int_0^T e(t)^2\,dt}$$

$$Noise_{RMS} = \sqrt{\frac{\Delta V^2}{12}}$$

{yhh,cxh}@ustc.edu.cn

SNR

假定信号为正弦波形 $2^{N-1}\Delta V \sin \omega t$

$$Signal_{RMS} = \frac{2^{N-1}\Delta V}{\sqrt{2}}$$

$$Noise_{RMS} = \sqrt{\frac{\Delta V^2}{12}}$$

$$SQNR = \frac{Signal_{RMS}}{Noise_{RMS}} = \frac{2^{N-1}\Delta V}{\sqrt{2}} \cdot \frac{\sqrt{12}}{\Delta V} = \sqrt{1.5} \cdot 2^N$$

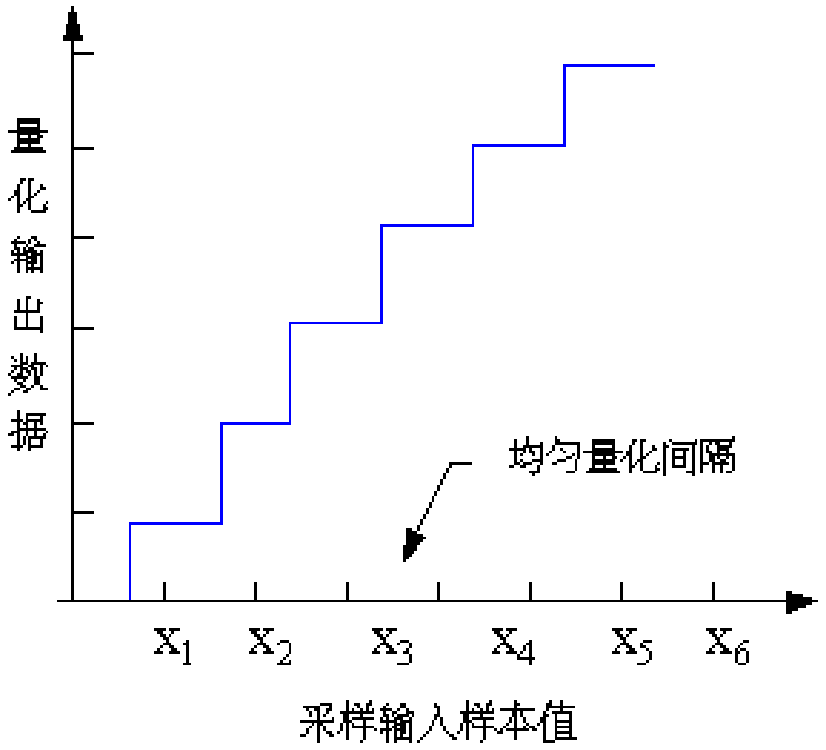
$$SQNR_{dB} = (6.02N + 1.76)dB$$

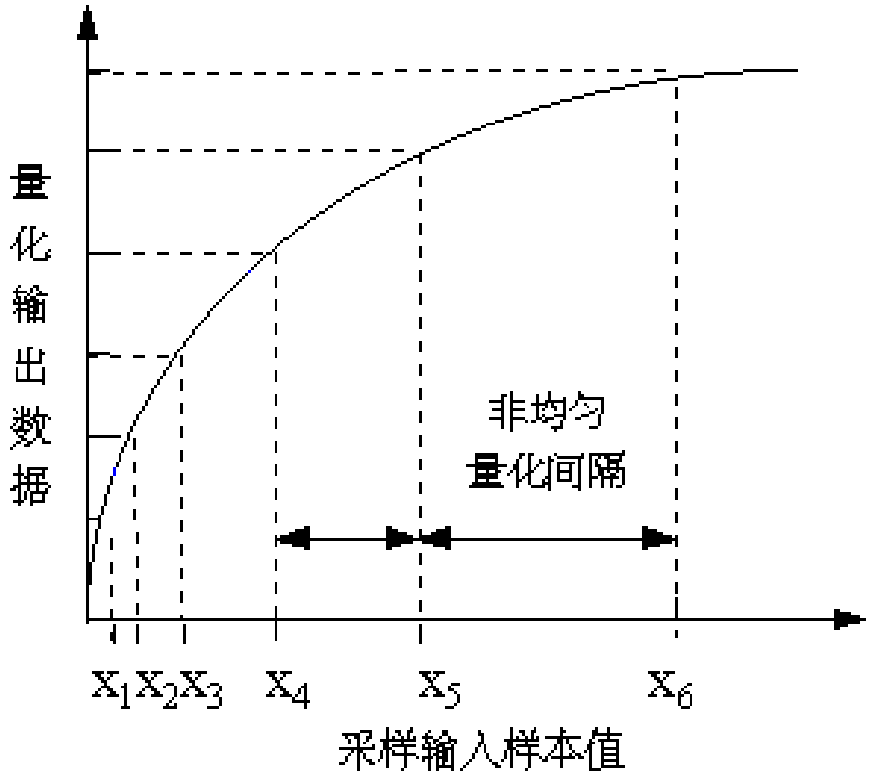
bits	SNR
4	25.8dB
8	49.9dB
12	74.0dB
16	98.1dB

注意：Signal幅度充满整个区间

{ynh,cxh}@nudt.edu.cn

脉冲编码调制(PCM)的量化方式





- μ 律(μ-Law)压扩(G.711)主要用在北美和日本等
- A律(A-Law)压扩(G.711)主要用在欧洲和中国大陆等

{ynh,cxh}@ustc.edu.cn

μ-Law & A-Law

- A-Law

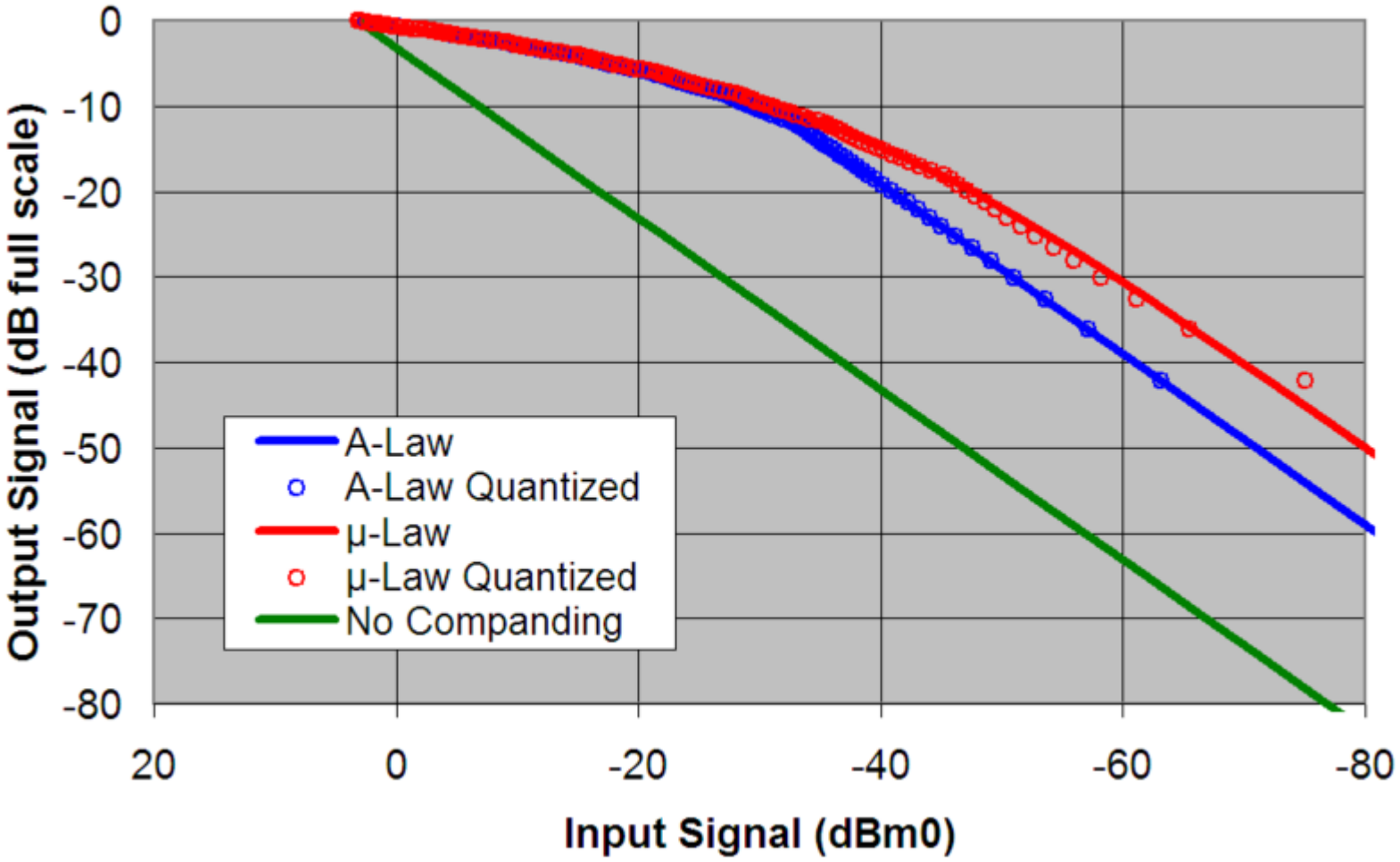
$$F(x) = \mathrm{sgn}(x) \begin{cases} \frac{A|x|}{1+\ln(A)}, & |x| < \frac{1}{A} \\ \frac{1+\ln(A|x|)}{1+\ln(A)}, & \frac{1}{A} \le |x| \le 1, \end{cases}$$

- μ-Law

$$F(x) = \mathrm{sgn}(x) \frac{\ln(1+\mu|x|)}{\ln(1+\mu)} \qquad -1 \le x \le 1$$

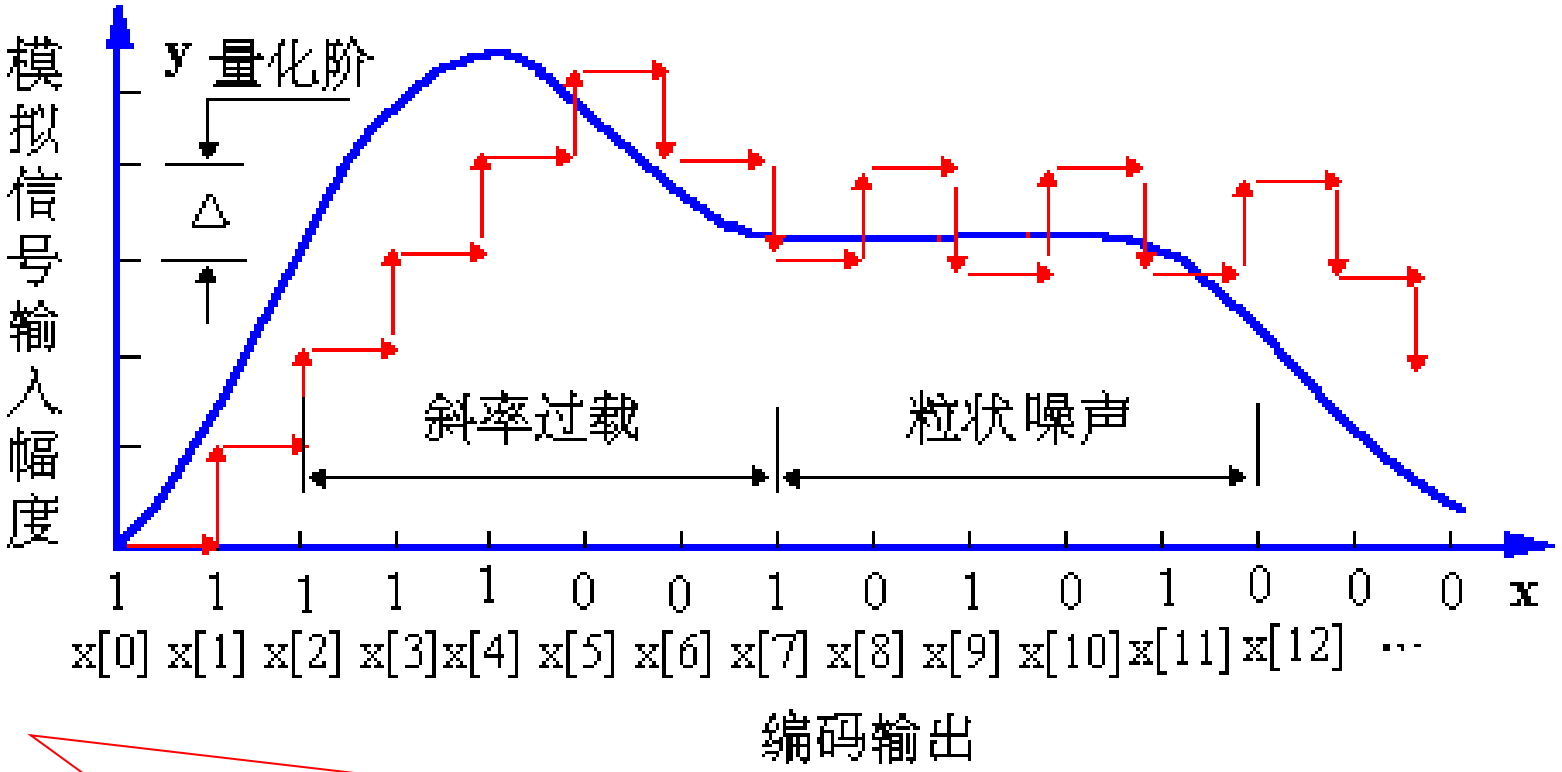
{ynh,cxh}@ustc.edu.cn

μ-Law vs. A-Law



{ynh,cxh}@ustc.edu.cn

增量调制(△M)



△调制(Delta Modulation,DM)是PCM编码的一种变形。PCM是对每个采样信号的整个幅度进行量化编码，因此它具有对任意波形进行编码的能力；DM是对实际的采样信号与预测的采样信号之差的极性进行编码，将极性变成“0”和“1”这两种可能的取值之一。由于DM编码只须用1位对话音信号进行编码，所以DM编码系统又称为“1位系统”。

{ynh,cxh}@ustc.edu.cn

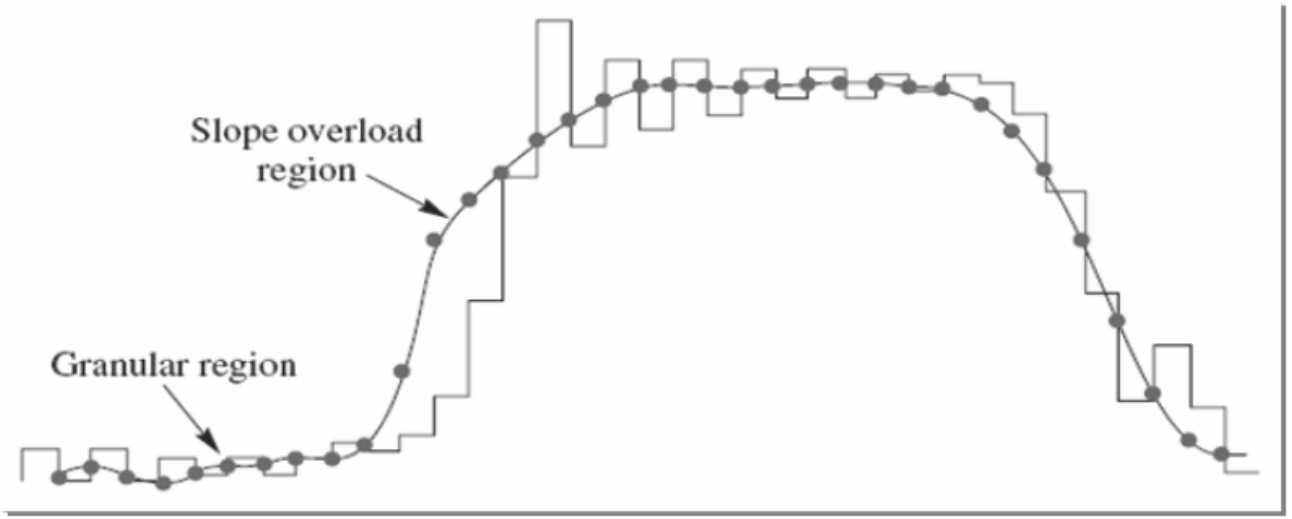
自适应增量调制

- 思路：自动调整量化阶△的大小；在检测到斜率过载的时候增大△，在输入信号斜率减小时降低△

- CFDM(Constant Factor Adaptive DM)
  - 根据量化器符号的判断当前区域是斜率过载还是颗粒噪声，进而改变△
- CVSD(Continuously Variable Slope DM)
  - 如果连续出现三个相同值△加大，反之减小

{ynh,cxh}@ustc.edu.cn

CFDM调整量化阶的过程



$$s_n = \begin{cases} 1 & \text{if } \hat{d}_n > 0 \\ -1 & \text{if } \hat{d}_n < 0 \end{cases} \qquad \Delta_n = \begin{cases} M_1 \Delta_{n-1} & \text{if } s_n = s_{n-1} \\ M_2 \Delta_{n-1} & \text{if } s_n \neq s_{n-1} \end{cases}$$

$$1 < M = M_1 = 1/M_2 < 2$$

{ynh,cxh}@ustc.edu.cn

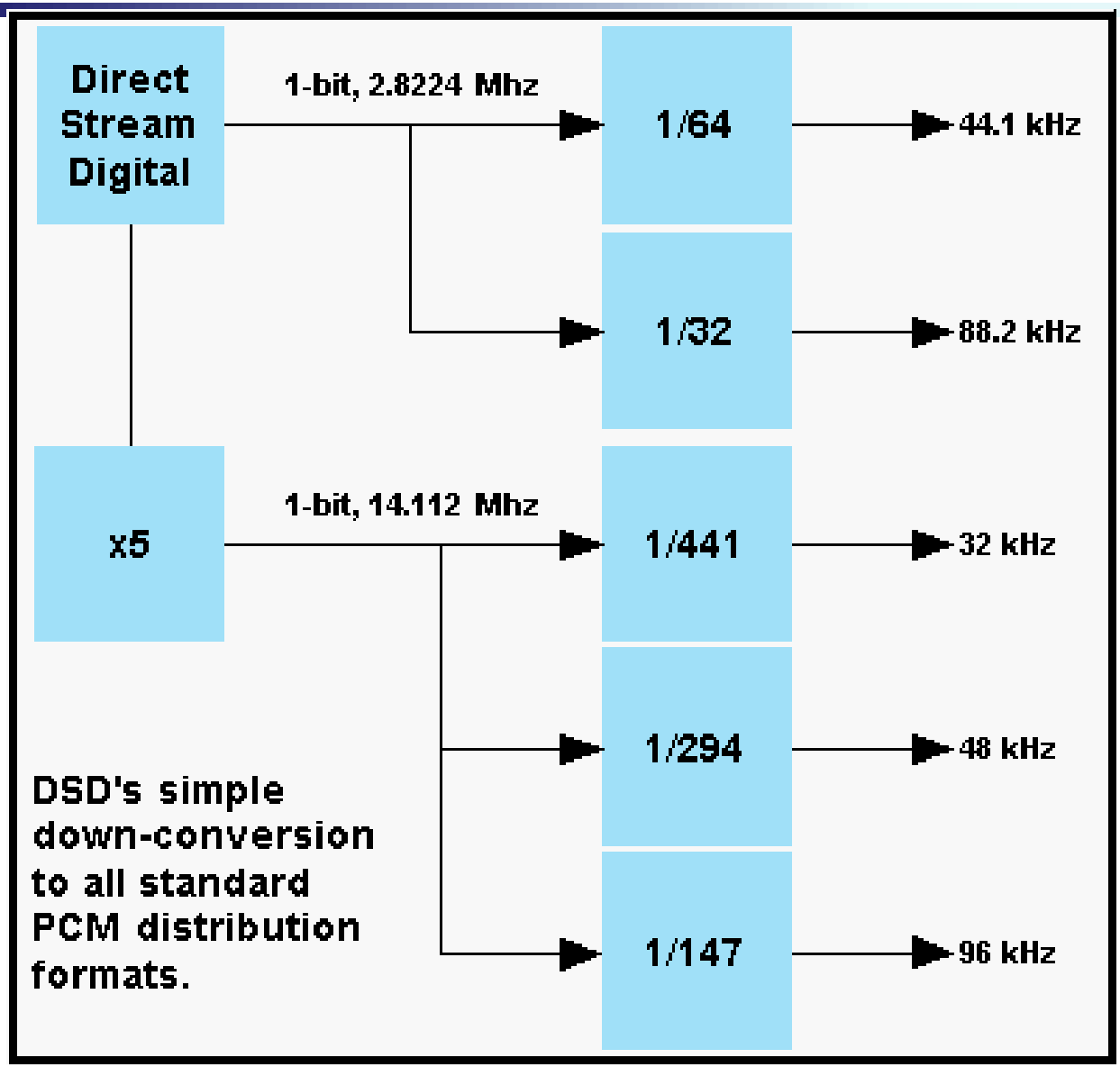
PCM vs. △M

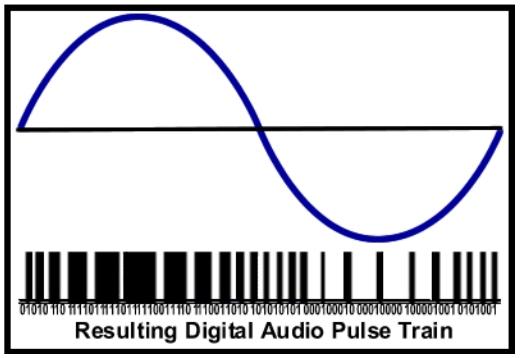
- 对于音频信号哪种更好？
- 各自怎样保证失真较小？

- CD
  - PCM（16bit/44.1kHz）
- SACD（Super Audio CD）
  - △M（1bit/2.8224MHz）

{ynh,cxh}@ustc.edu.cn

DSD(直接比特流数字编码)





SUPER AUDIO CD

{ynh,cxh}@ustc.edu.cn

“绝对” vs. “相对”

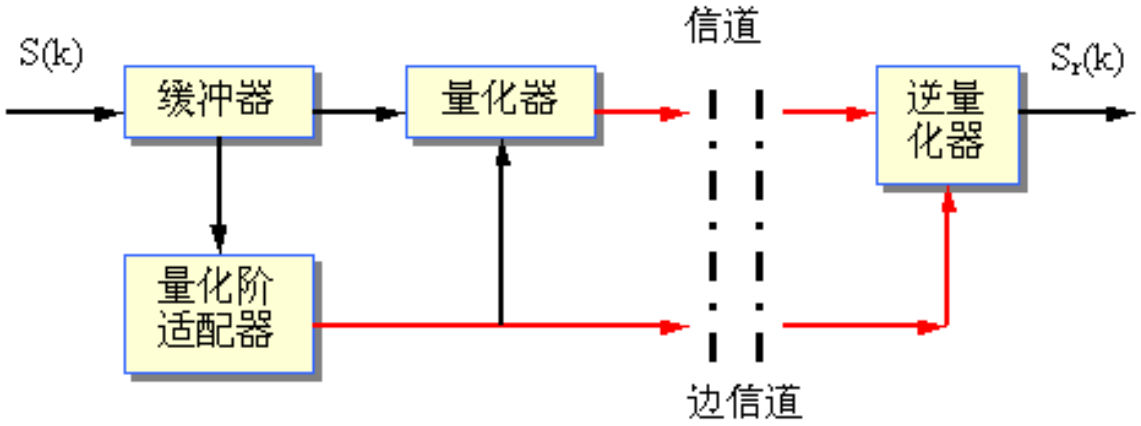
- 绝对数值和相对数值均可以表示信息，根据需要可以择其一
- 绝对和相对的互相转换

- 表示文件位置的绝对路径与相对路径
- 电脑屏幕上的绝对坐标与相对坐标

{ynh,cxh}@ustc.edu.cn

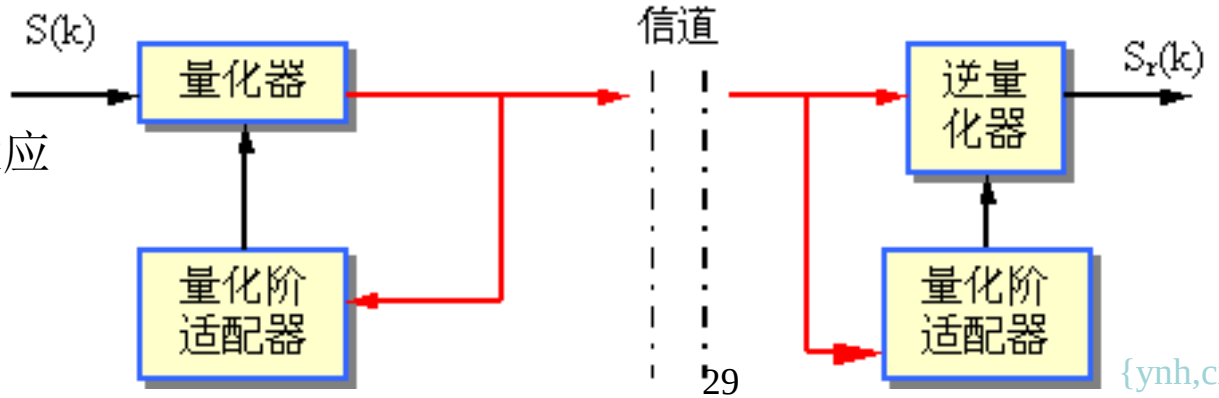
自适应脉冲编码调制
APCM, Adaptive Pulse Code Modulation

前向自适应



根据输入信号幅度大小来改变量化阶大小

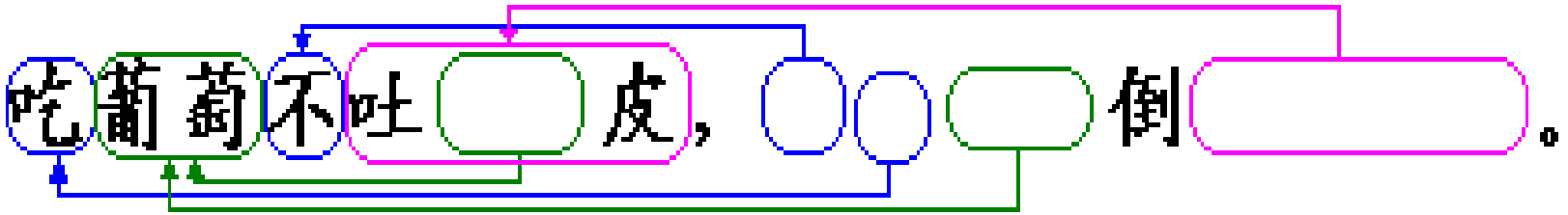
后向自适应



{ynh,cxh}@ustc.edu.cn

题外话：自适应的过程

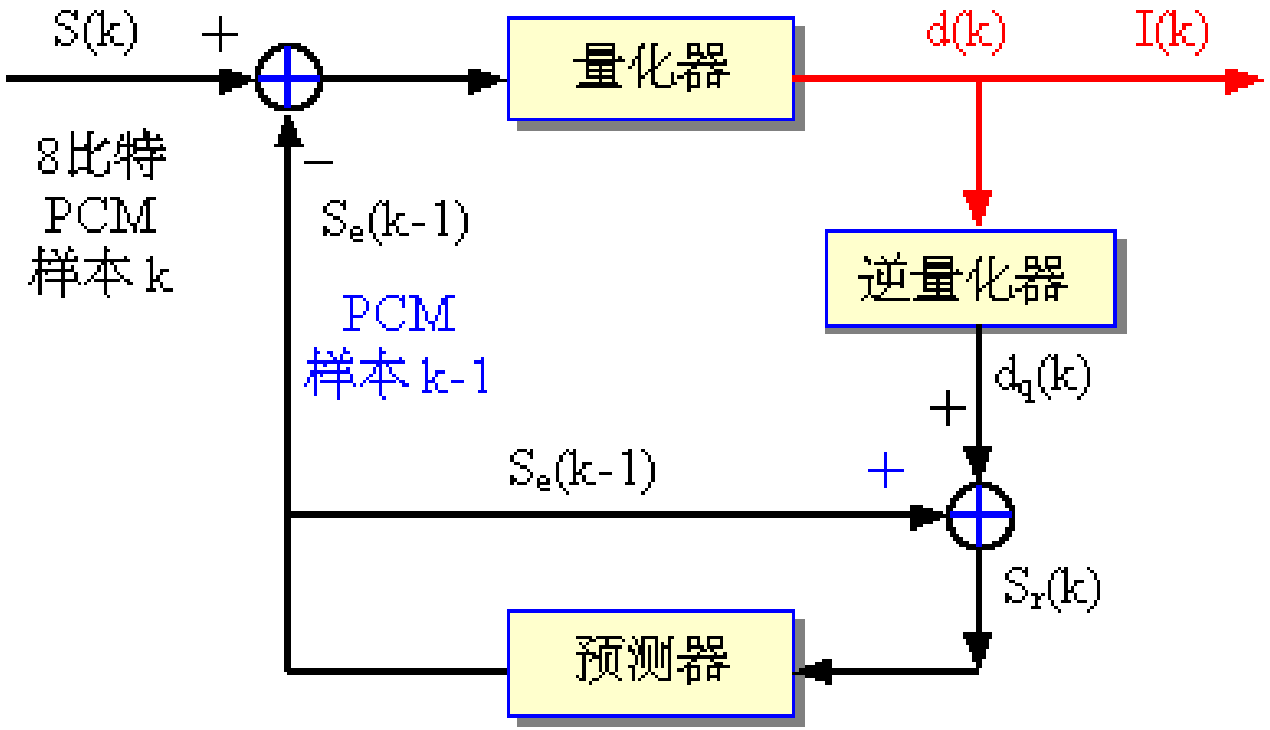
- 自适应：根据输入信息：
- 吃葡萄不吐葡萄皮，不吃葡萄倒吐葡萄皮
- 自动找出重复出现的词或短语



{ynh,cxh}@ustc.edu.cn

差分脉冲编码调制
DCPM, Differential Pulse Code Modulation

根据过去的样本估算下一个样本信号的幅度大小，这个值称为预测值，然后对实际信号值与预测值之**差进行量化编码**，从而就减少了表示每个样本信号的位数。



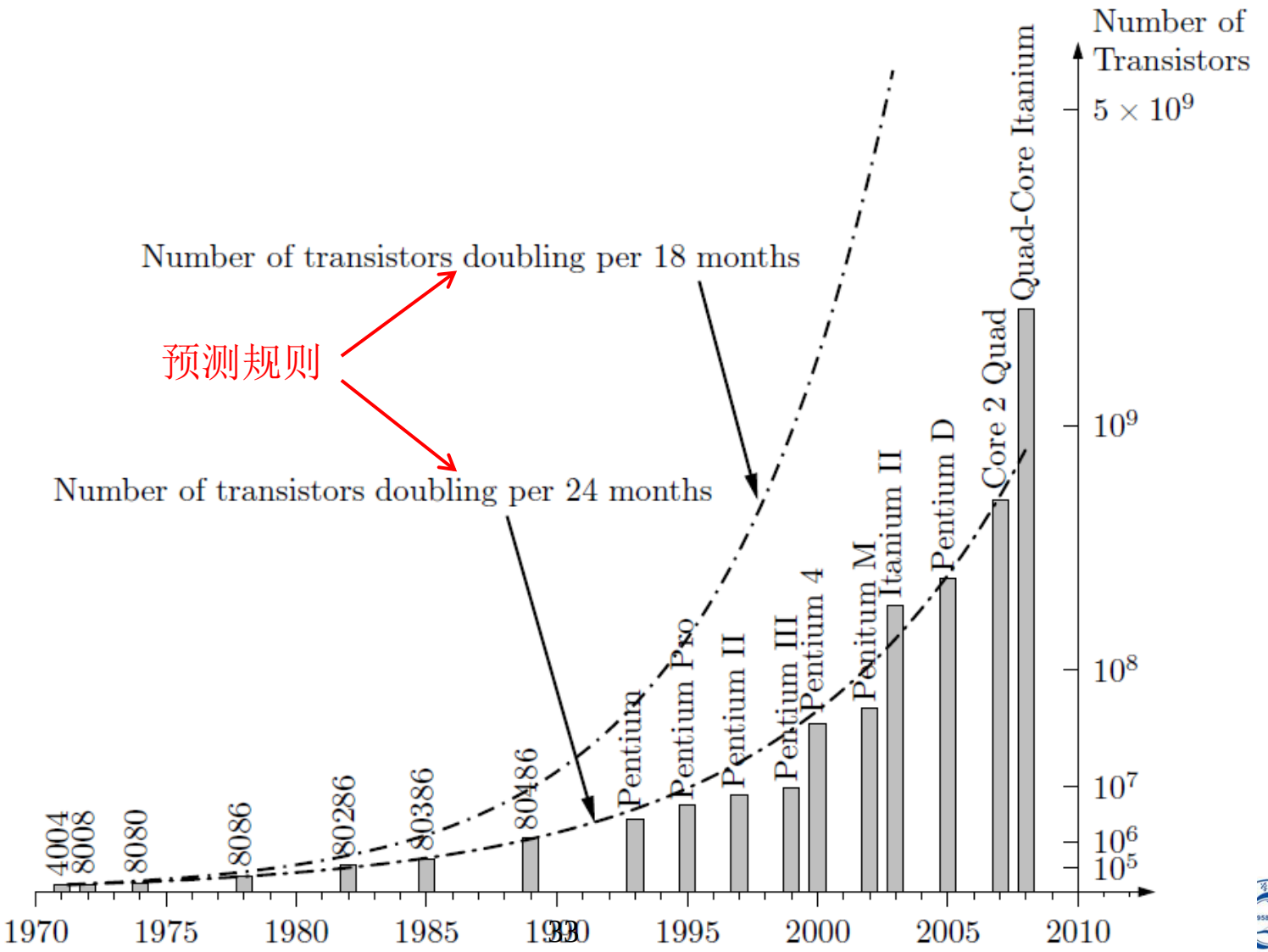
{ynh,cxh}@ustc.edu.cn

怎样做预测

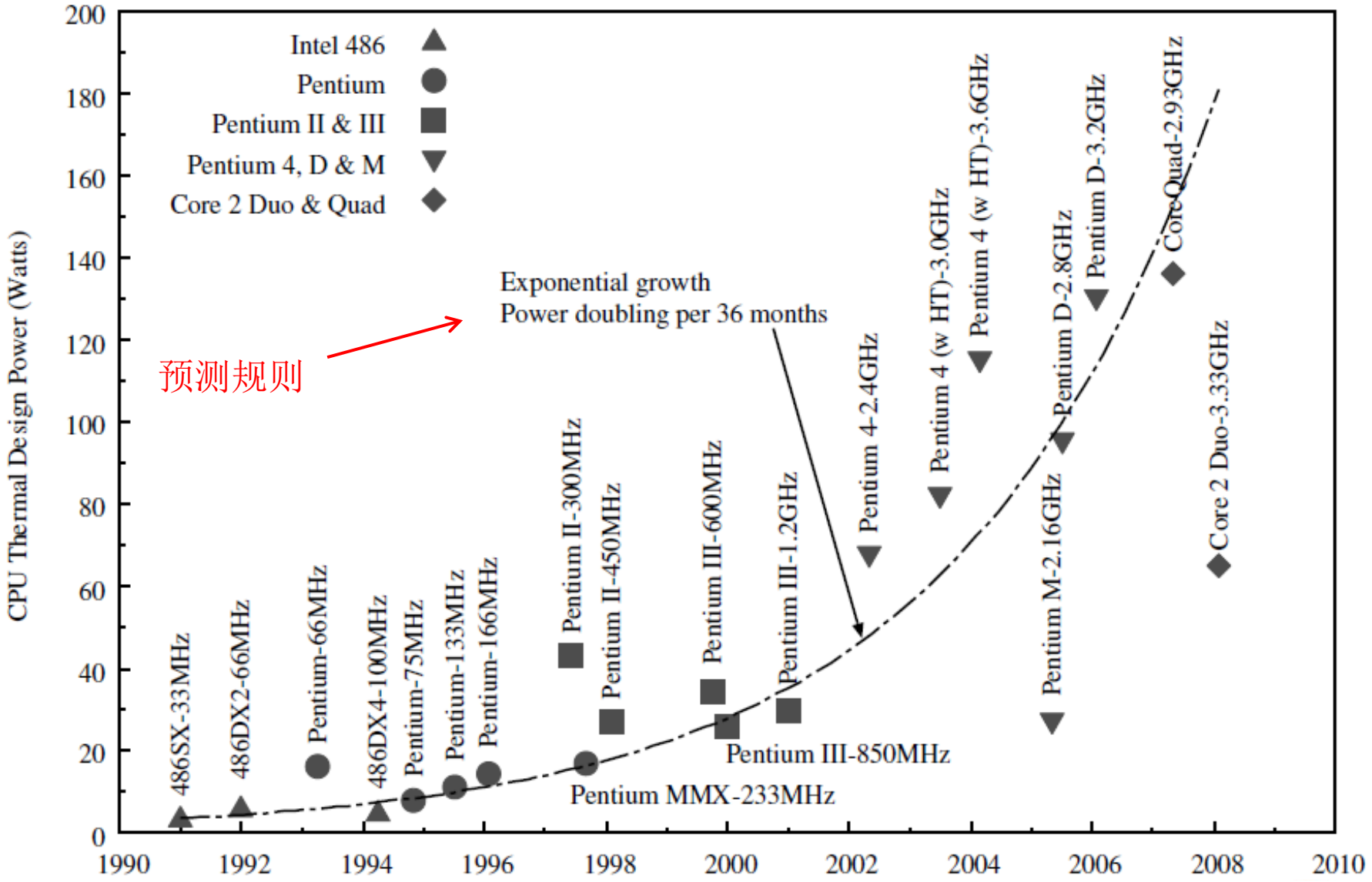
- 根据摩尔定律预测芯片发展
- 利用概率模型预测不及格学生人数
- 利用遥感技术预测农业产量
- 2025年中国老年人比例预测
- K线图的走势分析

预测模型

电脑芯片中晶体管数量每18个月翻一番



CPU热设计功耗


CPU Thermal Design Power (Watts)
200
180
160
140
120
100
80
60
40
20
0
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
Intel 486
Pentium
Pentium II & III
Pentium 4, D & M
Core 2 Duo & Quad
Exponential growth
Power doubling per 36 months
预测规则
486SX-33MHz
486DX2-66MHz
Pentium-66MHz
486DX4-100MHz
Pentium-75MHz
Pentium-133MHz
Pentium-166MHz
Pentium II-300MHz
Pentium II-450MHz
Pentium III-600MHz
Pentium III-1.2GHz
Pentium 4-2.4GHz
Pentium 4 (w HT)-3.0GHz
Pentium 4 (w HT)-3.6GHz
Pentium D-2.8GHz
Pentium D-3.2GHz
Pentium M-2.16GHz
Core Quad-2.93GHz
Core 2 Duo-3.33GHz
Pentium III-850MHz
Pentium MMX-233MHz

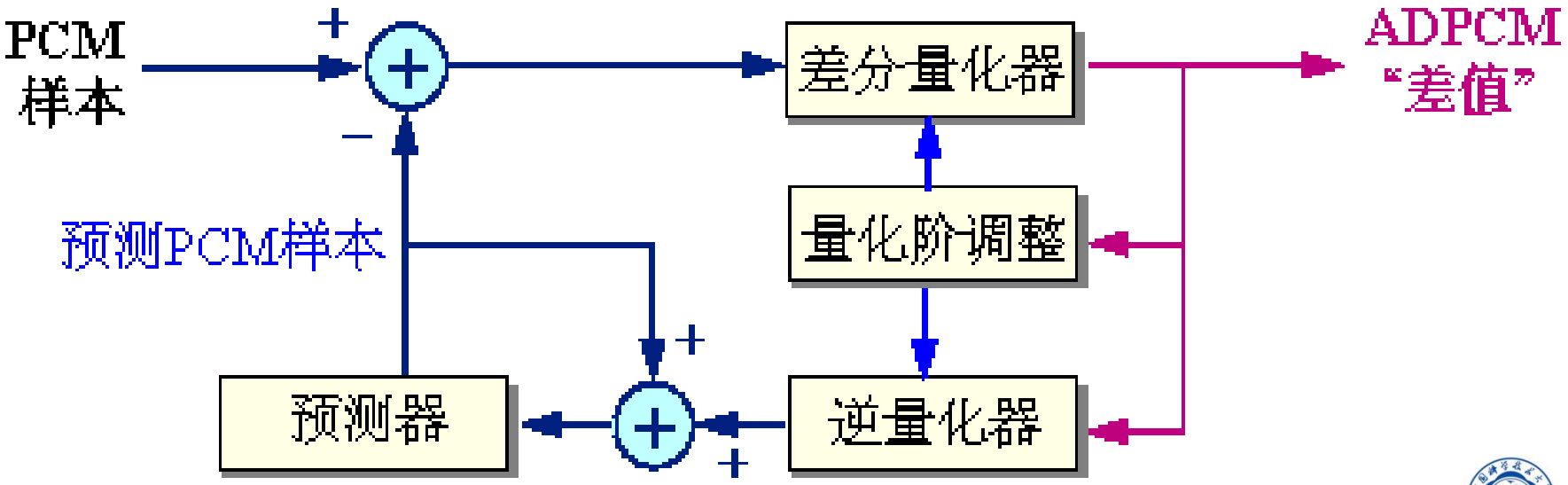
预测结果评判

- 利用误差最小的原则来确定预测模型的各项参数
  - 误差 – 实际值与预测值之间的偏差
  - 绝对平均偏差(MAD:mean absolute deviation)
    - 误差的绝对值的平均值
  - 均方差(MSE:mean square error)
    - 误差的平方的平均值
  - 累积误差(RSFE: running sum of forecast error)
- Such as：在企业效益不变的前提下，老板可以按照总体工薪成本最低的原则确定每年的人力资源计划

{ynh,cxh}@ustc.edu.cn

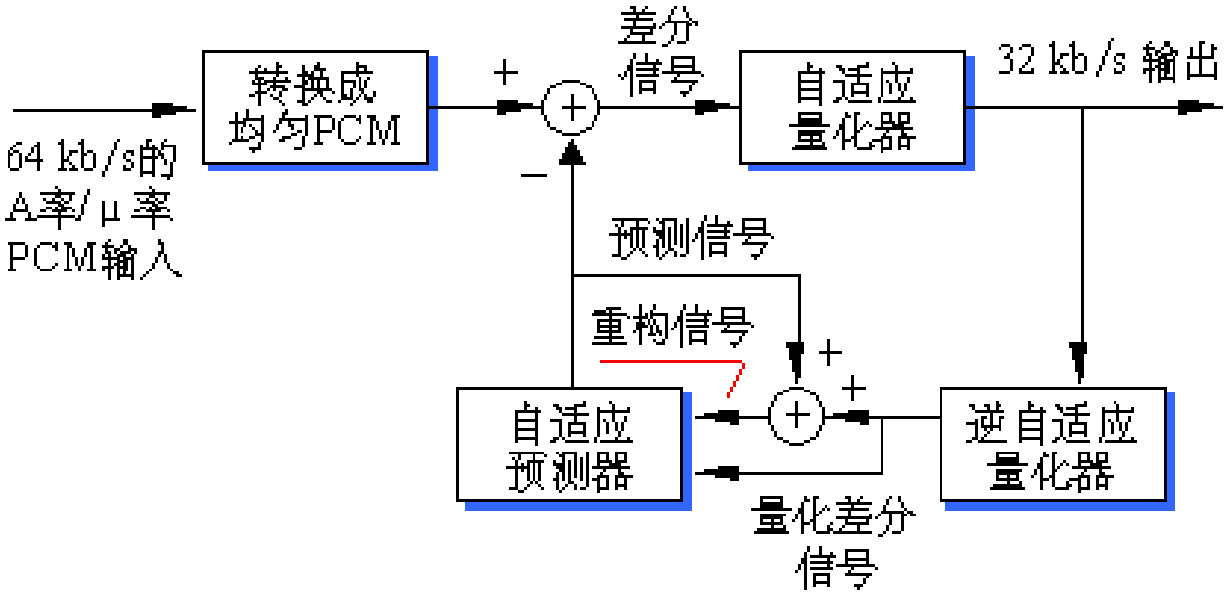
自适应差分脉冲编码调制(ADPCM)
Adaptive Difference Pulse Code Modulation

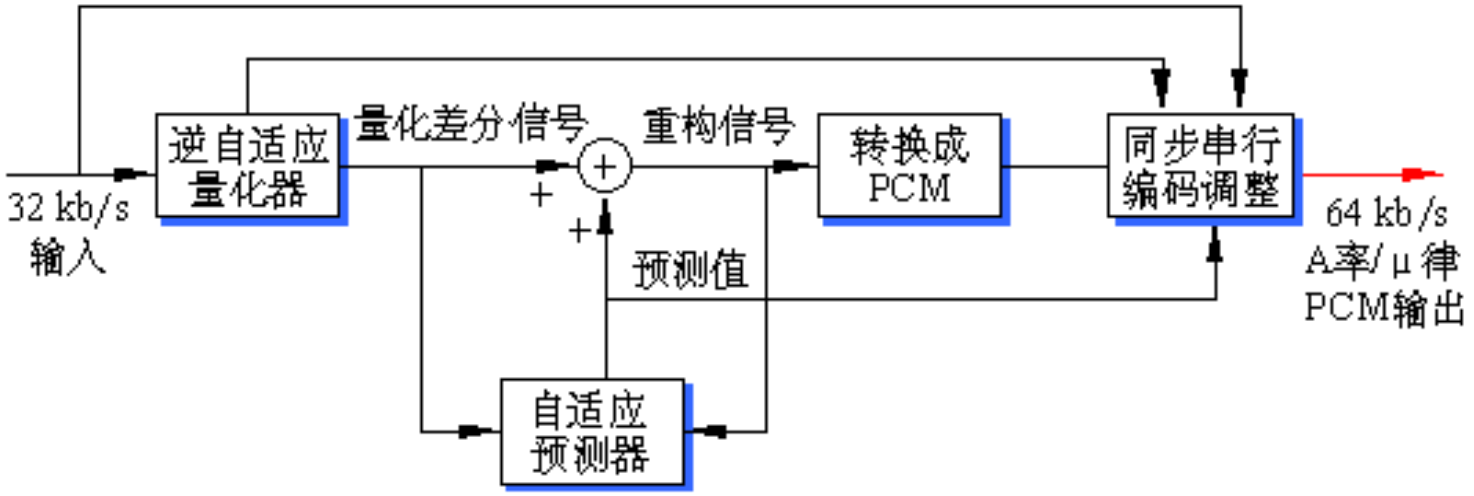
- 结合APCM自适应特性和DPCM差分特性，思路：
  - 利用自适应改变量化阶大小，用小的量化阶(step-size)编码小的差值，用大的量化阶编码大的差值
  - 使用过去的样本值估算当前输入样本的预测值，使实际样本值和预测值之间的差值总是最小



{ynh,cxh}@ustc.edu.cn

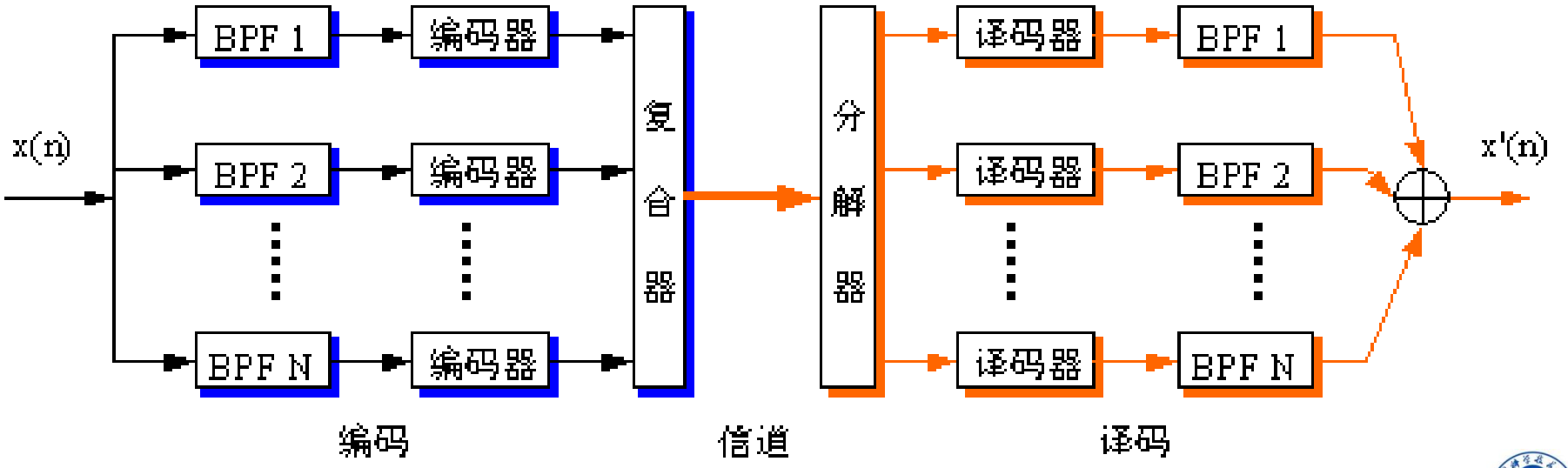
G.721





子带编码
SBC(subband coding)

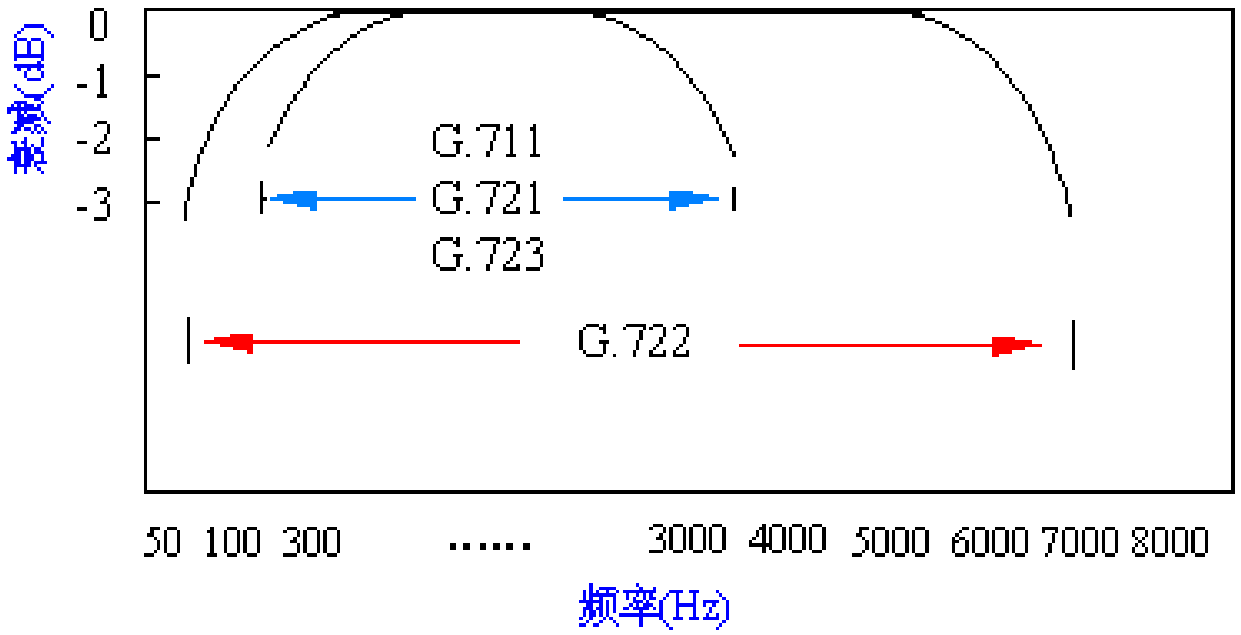
◆用一组带通滤波器BPF(band-pass filter)把输入音频信号的频带分成若干个子带。每个子带中的音频信号采用单独的编码方案编码。传送时，将每个子带的代码复合起来。在接收端译码时，将每个子带的代码单独译码，然后把它们组合起来。



{ynh,cxh}@ustc.edu.cn

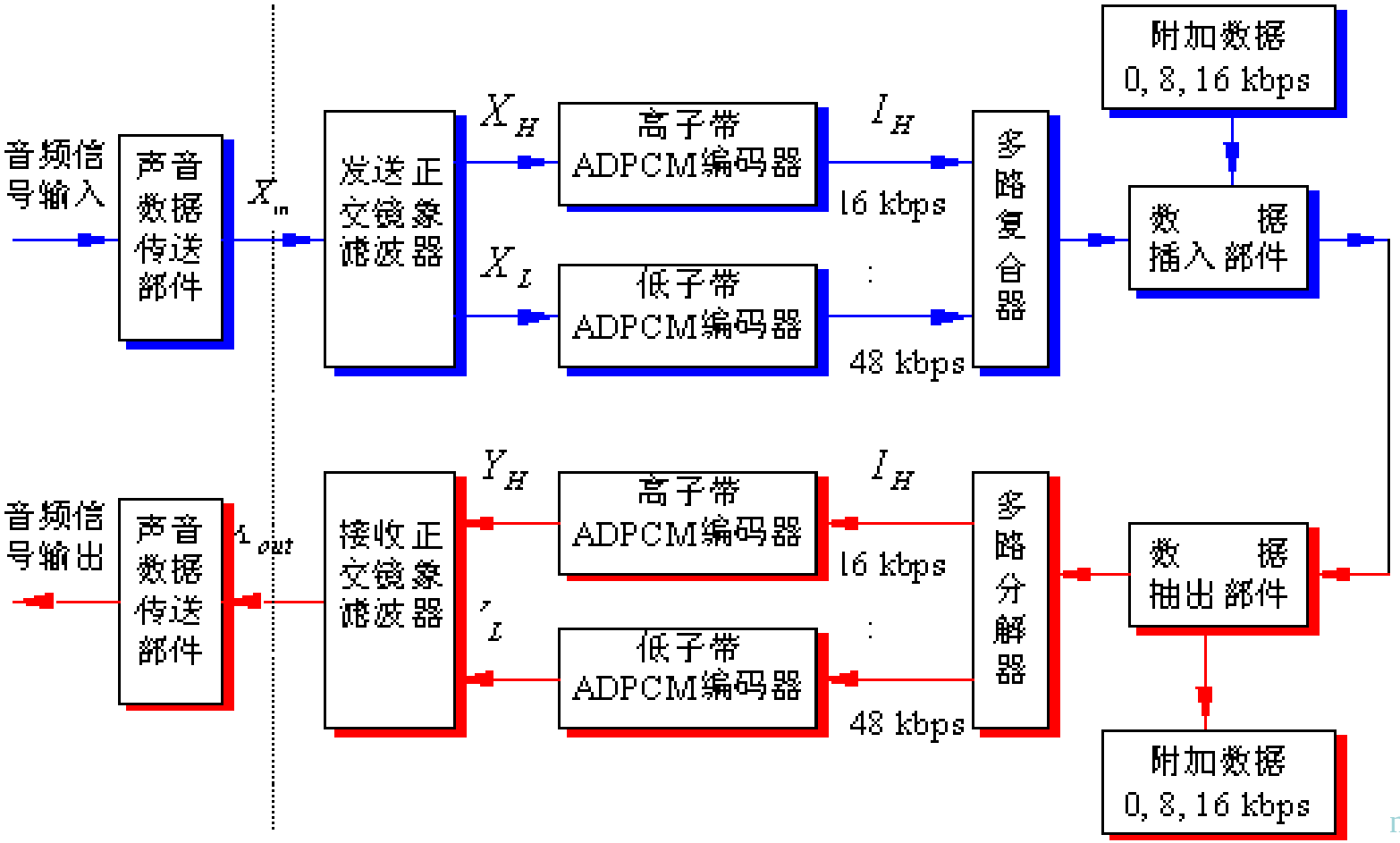
子带-自适应差分脉冲编码调制(SB-ADPCM)

- G.722标准把采样频率由8kHz提高到16kHz
- 信号频率由原来的3.4 kHz扩展到7 kHz
- 低频端把截止频率扩展到50 Hz



G.722 SB-ADPCM

- 低频段6bit/sample；高频段2bit/sample



小结：波形编译码

- PCM(G.711 64kb/s):幅度的非均匀分布→非均匀量化
- DPCM:利用相邻样本之间冗余信息
- APCM:信号的动态范围
- △M:斜率过载、颗粒噪声
- ADPCM(G.721 32kb/s):4bit
- SB-ADPCM(G.722 48kb/s + 16kb/s):低频/高频子带

- 目标：减小量化误差

{ynh,cxh}@ustc.edu.cn

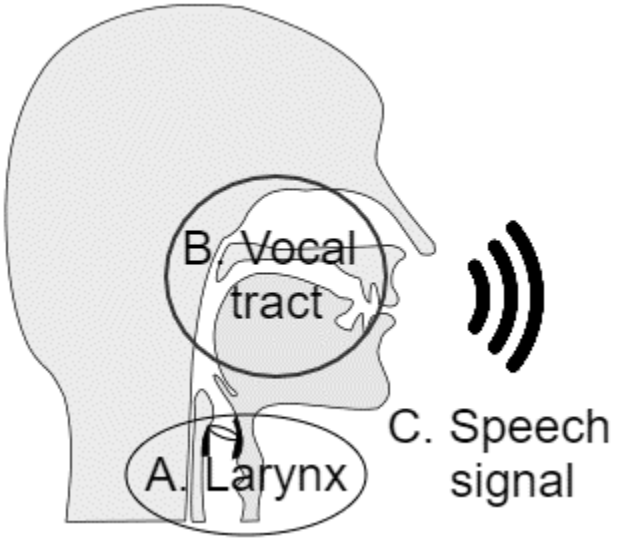
第三章 多媒体数据压缩

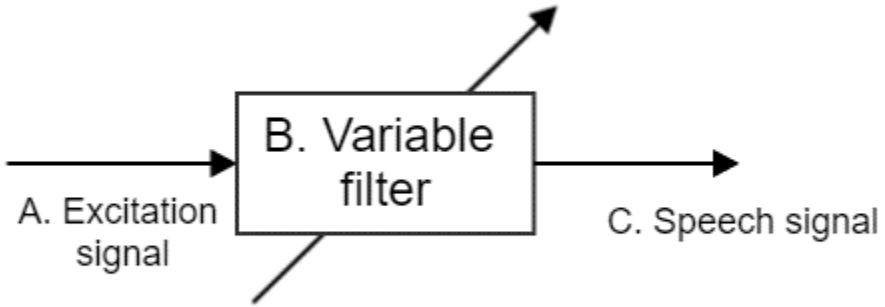
- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
    - 波形编译码器
    - 音源编译码器
    - 混合编译码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MPEG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

源——滤波器模型 source-filter model

◆源——滤波器模型：人发音系统由独立的两部分组成，其中声带作为源振动发声，发出的信号经过一个由声道、喉咙、口腔、鼻腔、牙齿与嘴唇构成的滤波器系统拥有了特定的频谱。





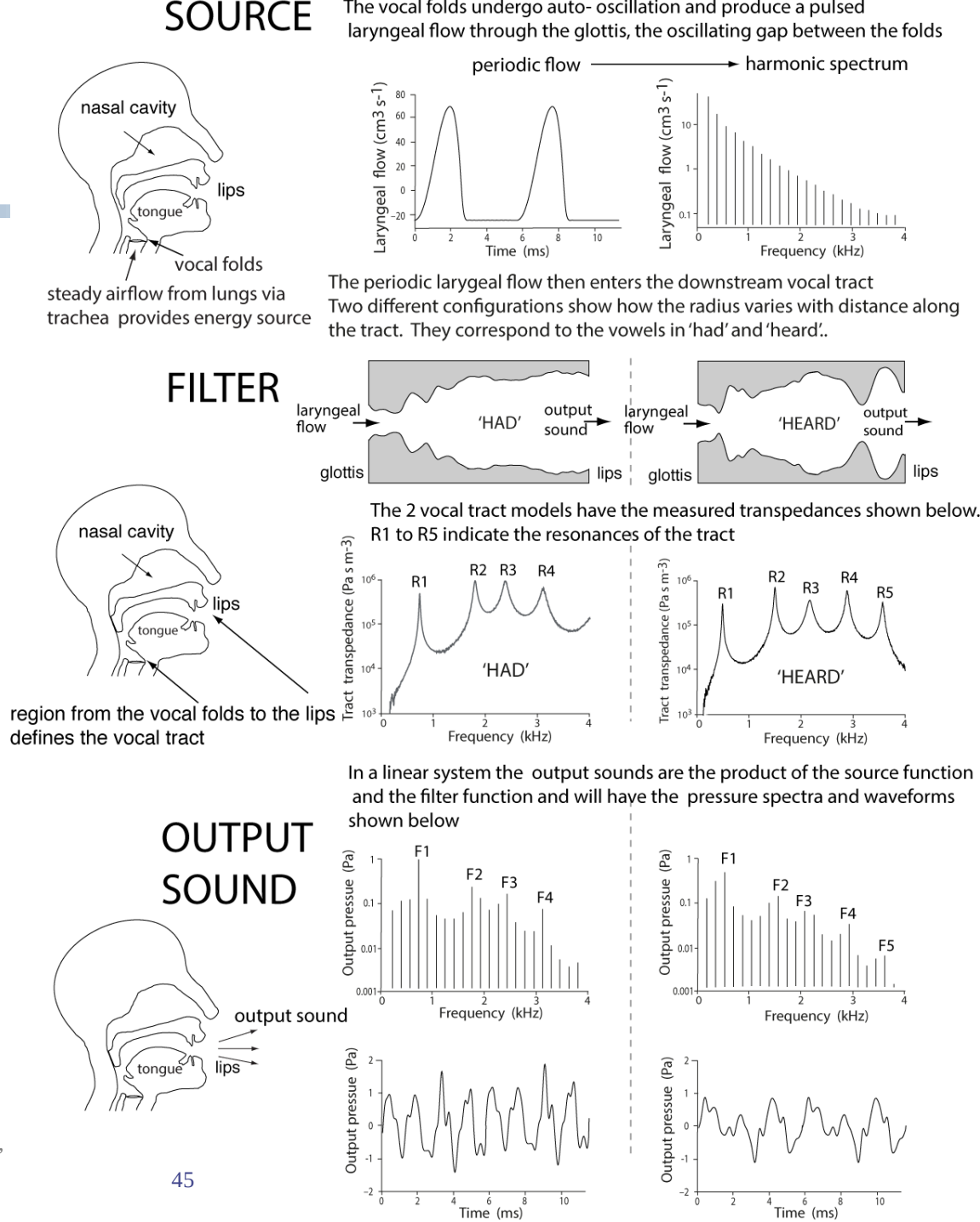
{ynh,cxh}@ustc.edu.cn

Chiba, Tsutomu, and Masato Kajiyama. The Vowel, Its Nature and Structure. Tokyo: Phonetic Society of Japan, 1958.

Source filter

“HAD” 和 “HEARD” 所对应的共鸣腔形状不同，其对不同频率的透射率也不同，但都具有数个峰值（R1、R2、R3 等）。

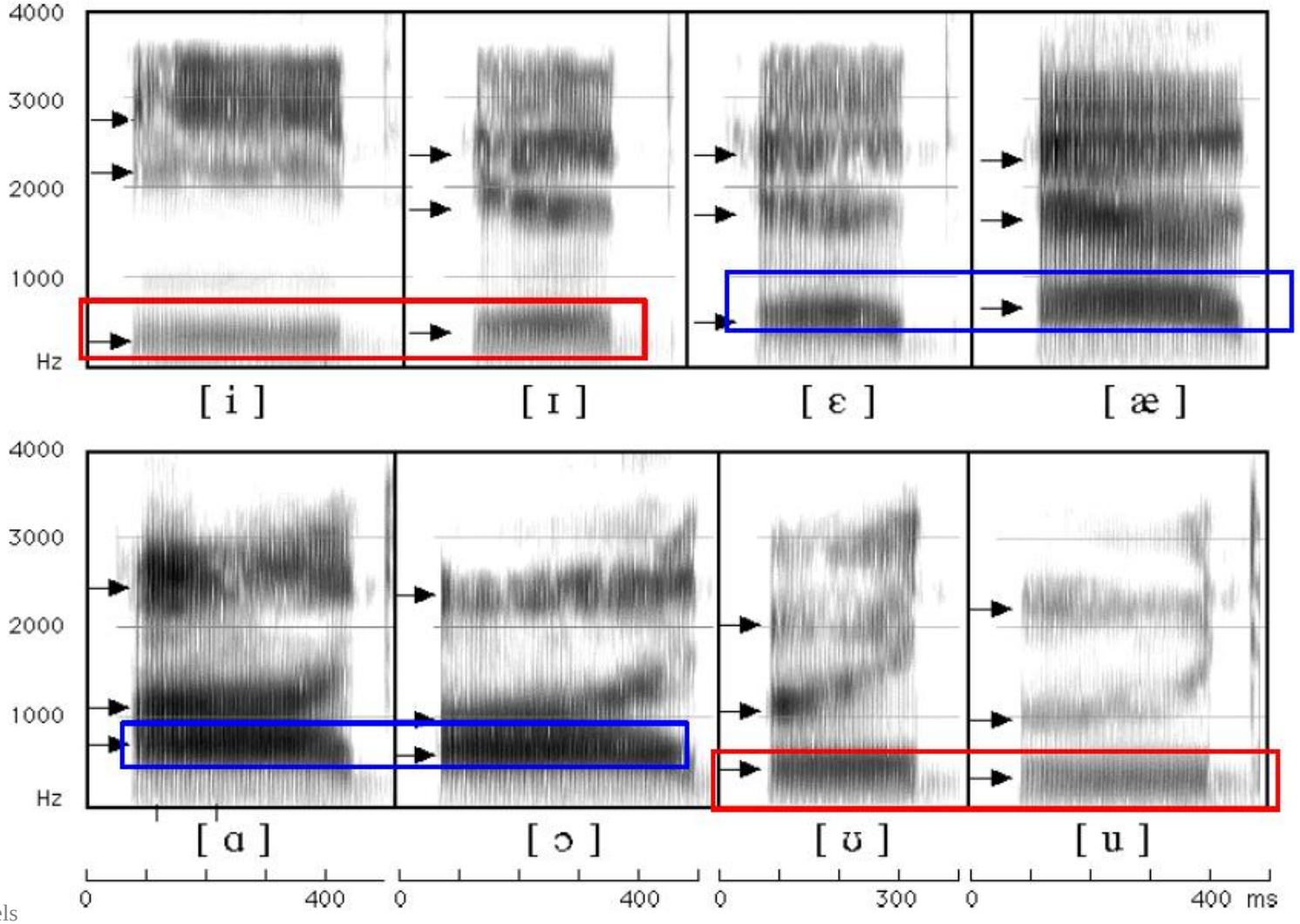
源发出的脉冲在穿过共鸣腔滤波器后出现了数个频率峰值（F1、F2、F3 等）。



"An experimentally measured Source-Filter model: glottal flow, vocal tract gain and radiated sound from a physical model," Wolfe, J., Chu, D., Chen, J.-M. and Smith J. (2016) Acoust. Australia https://www.phys.unsw.edu.au/jw/reprints/Source-Filter.pdf

英语中元音的频谱结构

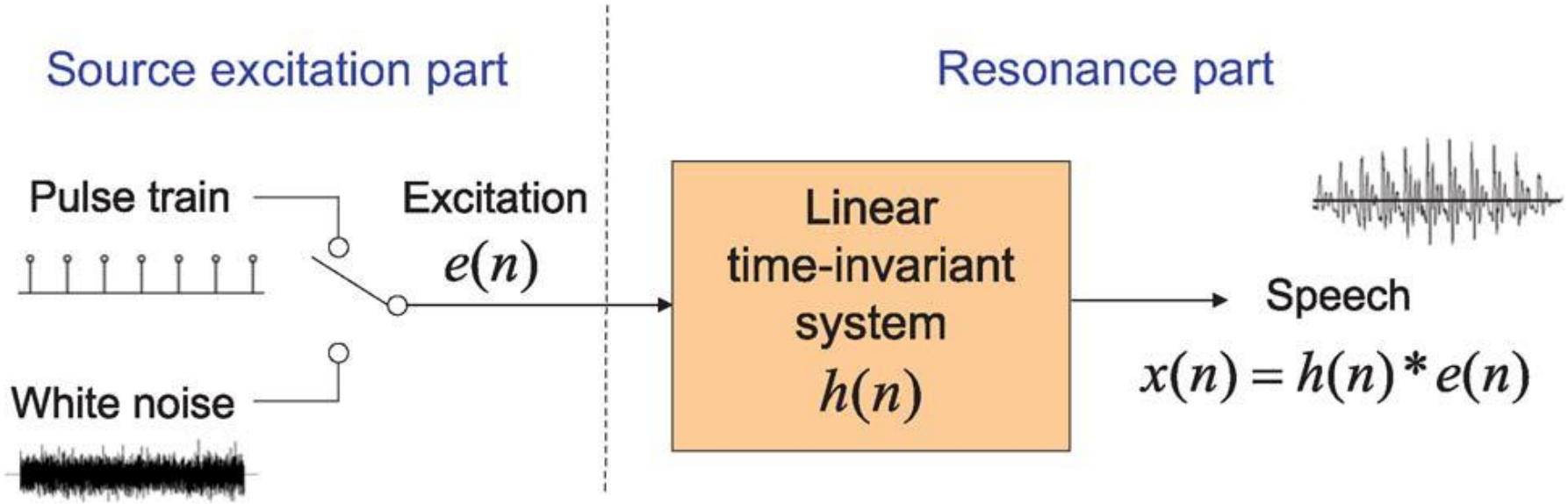
Red = high vowels, low F1　　Blue = mid/low vowels, higher F1



Formants of Vowels
https://corpus.eduhk.hk/english_pronunciation/index.php/2-2-formants-of-vowels/

话音产生的数字模型
Source-filter model

◆源-滤波器语音合成方法模拟人的语音产生过程，认为语音由人的声带振动产生的激励经过声道传播产生，声道中的传播过程可以用一个滤波器近似表示。



K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, "Speech Synthesis Based on Hidden Markov Models," in Proceedings of the IEEE, vol. 101, no. 5, pp. 1234-1252, May 2013, doi: 10.1109/JPROC.2013.2251852.

{ynh,cxh}@ustc.edu.cn

音源编译码器

- 通过话音波形的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。
- 在模型中声道被等效成一个随时间变化的滤波器，叫时变滤波器，激励函数是由白噪声、无声话音段激励或者由有声话音段激励。
- 传送的是解码器的信息就是滤波器的规格、发声或不发声的标志和有声话音的音节周期，每10~20ms更换一次。
- 数据率2.4bps，产生的语音质量很低，可以听懂而已。增加数据率对于话音质量没有用，因为这是由模型限制的，但保密性好。

{ynh,cxh}@ustc.edu.cn

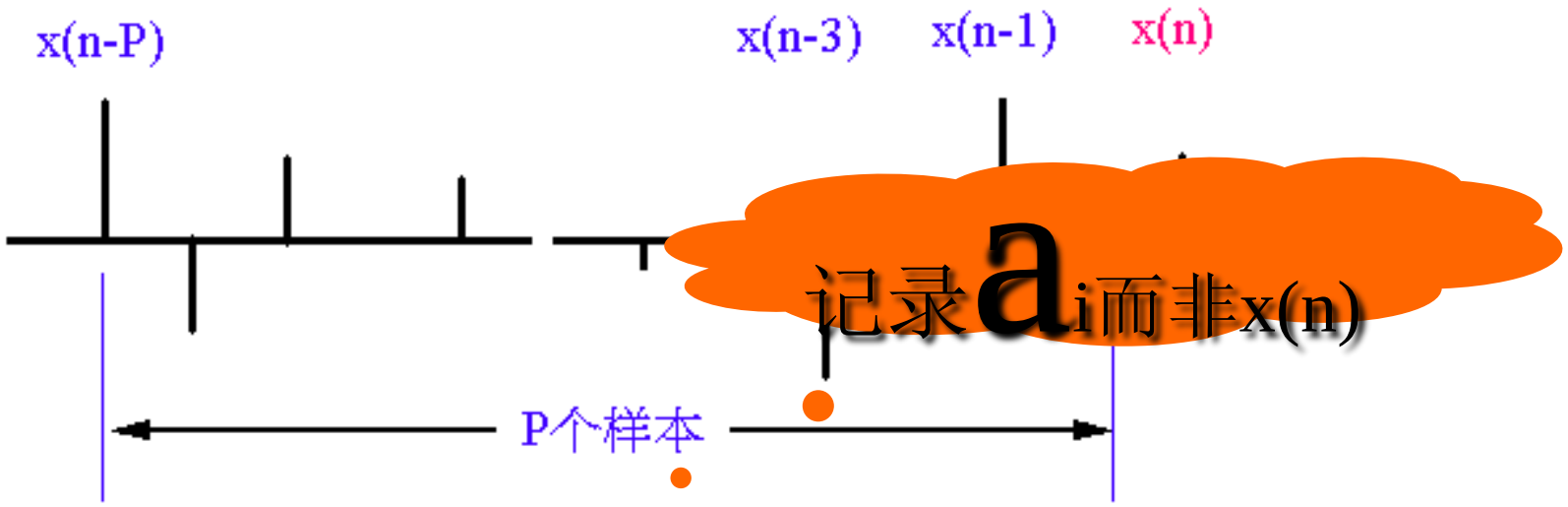
音源编译码数据速率低

- 一般的语音传输每隔20ms传输一次
  - 话音在短时间周期(20 ms的数量级)里可以被认为是准定态(quasi-stationary)的，也就是说基本不变的。
- 波形编码的数据量大
  - 20ms的CD音乐的存储量
    - 20ms/1000ms*44.1k*2Bytes*2=3.528kB
  - 20ms的G.721的存储量
    - 20ms/1000ms*32kbits=0.64kb
- 用声道参数表示声音
  - LPC速率2.4kbps(平均20ms传输48bit)

{ynh,cxh}@ustc.edu.cn

线性预测编码(LPC)

- 发送端产生声道激励和转移函数的参数
- 接收端通过话音合成器重构话音
- 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。



$$x_{pre}(n) = -[a_1 x(n-1) + a_2 x(n-2) + \cdots\cdots + a_p (n-p)]$$

第三章 多媒体数据压缩

- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
    - 波形编译码器
    - 音源编译码器
    - 混合编译码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MEPG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

混合编译码

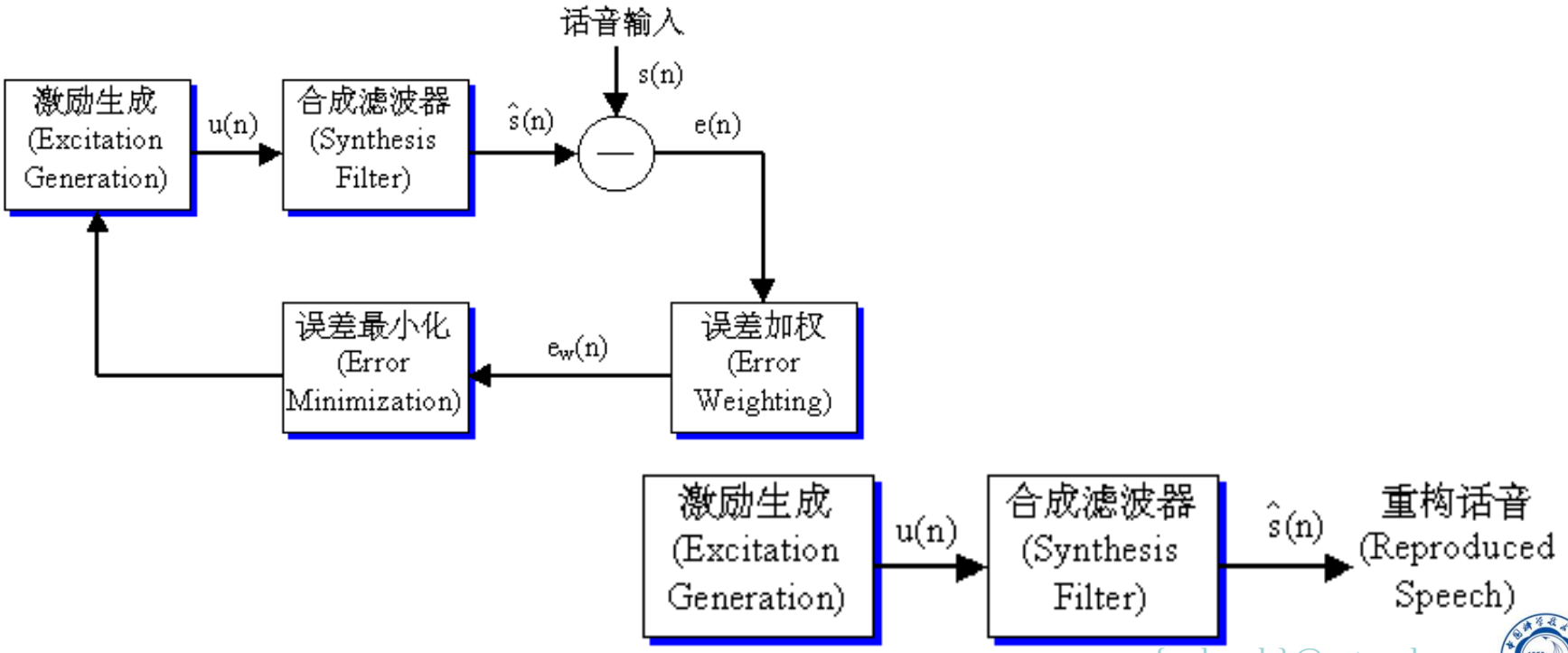
◆混合编译码的想法是企图填补波形编译码和音源编译码之间的间隔。波形编译码器虽然可提供高话音的质量，但数据率低于16 kb/s的情况下，在技术上还没有解决音质的问题；声码器的数据率虽然可降到2.4 kb/s甚至更低，但它的音质根本不能与自然话音相提并论。

◆为了得到音质高而数据率又低的编译码器，历史上出现过很多形式的混合编译码器，但最成功并且普遍使用的编译码器是时域合成-分析(analysis-by-synthesis，AbS)编译码器。

{ynh,cxh}@ustc.edu.cn

时域合成-分析编码 (AbS)

◆使用的声道线性预测滤波器模型与LPC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，而是企图寻找这样一种激励信号，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。



{ynh,cxh}@ustc.edu.cn

AbS的激励

- 多脉冲激励MPE (multi-pulse excited)
  - 每5ms使用4个脉冲，每个脉冲的位置和幅度由编码器确定
- 等间隔脉冲激励RPE (regular-pulse excited)
  - 每5ms用10个脉冲，记录第一个脉冲的位置和所有脉冲的幅度
- 码激励线性预测CELP(code excited linear predictive)
  - 激励信号由一个矢量量化大码簿的表项给出

{ynh,cxh}@ustc.edu.cn

小结：音频编码的分类

- 基于音频数据的统计特性进行**波形编码**
  - 基于音频数据的统计特性进行编码，其典型技术是波形编码。其目标是使重建语音波形保持原波形的形状，如PCM、DPCM、ADPCM、SB-ADPCM等。
- 基于音频的声学参数进行**参数编码**
  - 其目标是使重建音频保持原音频特性。常用的音频参数有共振峰、线性预测系数、滤波器组等。这种编码技术的优点是数据率低，但还原信号的质量较差，自然度低。
- 将上述两种编码算法结合起来的**混合编码**
  - 能在较低码率上得到较高的音质。如MPE、RPE、CELP等。
- 基于人的听觉特性进行编码
  - 利用人听觉系统的特性，设计心理声学模型，从而实现更高效率的数字音频的压缩。以MPEG的音频编码和Dolby AC-3最有影响。

{ynh,cxh}@ustc.edu.cn

第三章 多媒体数据压缩

- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MPEG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

GSM网络中的话音编码

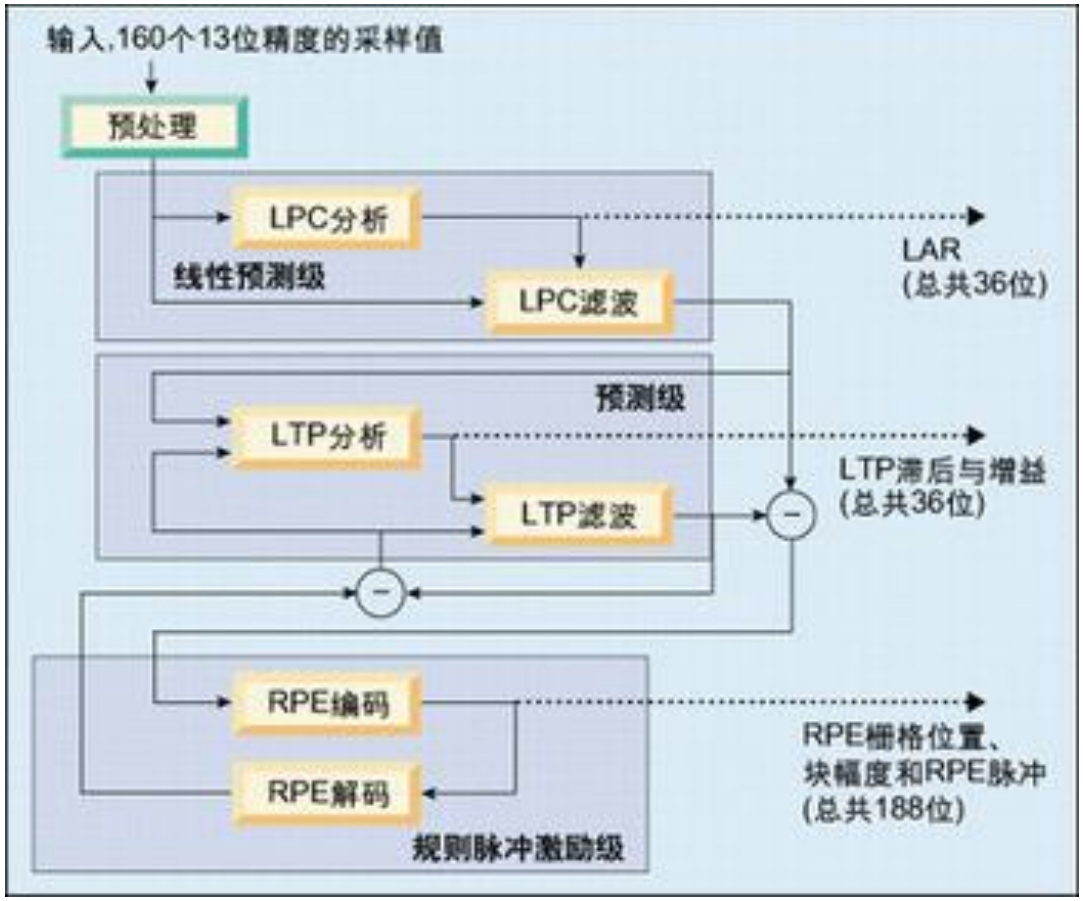
◆如果以8 kHz采样率及13位精度来对来自GSM蜂窝手机麦克风的音频数据进行采样，得到104kbps的源数据速率。GSM系统中有四种编解码器：全速率(Full Rate, FR)、增强型全速率(Enhanced Full Rate, EFR)、自适应多速率(Adaptive Multi-rate, AMR)及半速率(Half Rate, HR)语音压缩。

编解码器	位速率	压缩比	编码器类型
全速率	13Kbps	8	RPE-LTP LPC
EFR	12.2Kbps	8.5	ACELP
半速率	5.6Kbps	18.4	VSELP
AMR	12.2~4.75Kbps	8.5~21.9	ACELP

{ynh,cxh}@ustc.edu.cn

全速率 (Full Rate, FR)

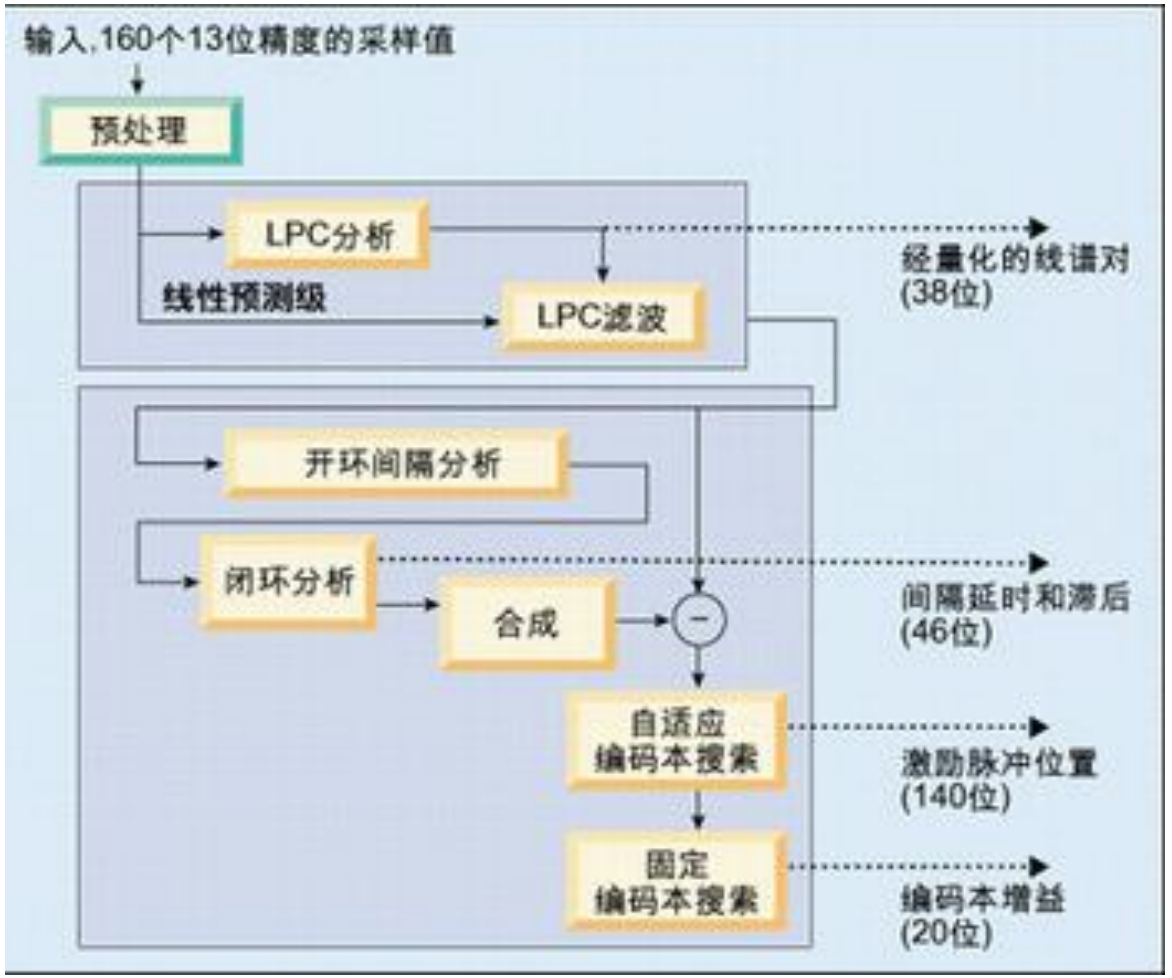
- 全速率编解码器就被称为RPE-LTP线性预测编码器
  - 长期预测(LTP)
  - 规则脉冲激励(RPE)
- 8个系数被变换成可以更少的位数来进行更佳量化的LAR(log-area ratio)。



{ynh,cxh}@ustc.edu.cn

增强型全速率 (Enhanced Full Rate, EFR)

◆EFR声码器是一种代数码激励线性预测(ACELP)编码器，EFR声码器的12.2kbps输出等于每帧244位。但编码语音是通过拥有260位容量的常规GSM全速率空中信道来传输，其余16位被填以CRC以及重复一些用于冗余的最重要编解码器参数。



{ynh,cxh}@ustc.edu.cn

自适应多速率 (Adaptive Multi-rate, AMR)

- 当全部参数均能解码时，全速率及EFR编解码器可实现良好的语音再现。但当参数丢失或错误时，所接收信号的质量将迅速下降。
- AMR声码器采用ACELP (Algebraic Code Excited Linear Prediction)编码方式，提供了8种编码速率(4.75 ~ 12.20 kbps)，故可提供87%至480%的冗余。在很糟的情况下，4.75kbps编解码器数据仍能恢复。
  - 1999年3GPP采纳了由爱立信、诺基亚、西门子提出的AMR作为第三代移动通信中语音编解码器的标准。提出了AMR－NB，AMR-WB和AMR-WB+三种不同的协议。AMR-NB应用于窄带，而AMR-WB和AMR-WB+则应用于宽带通信中。

{ynh,cxh}@ustc.edu.cn

半速率(Half Rate, HR)

- GSM所采用的空中接口允许使用两个完全独立的半速率子信道，故能使蜂窝单元的语音容量加倍。
- 半速率声码器采用矢量和激励线性预测VSELP(Vector Sum Exited Linear Prediction)编码器，它以一种类似EFR及AMR编解码器的分析加合成方式工作，速率为5.7kbps。
- 人们对半速率语音的感觉普遍不佳，在空口资源足够时一般不采用半速率编码。

{ynh,cxh}@ustc.edu.cn

4G之后的话音编码EVS
Enhanced Voice Services (EVS) Codec

涵盖语音和音乐的音频编解码器EVS

Band-width	Bitrates [kbps]											
FB 20 kHz						16.4	24.4	32.0	48.0	64.0	96.0	128.0
SWB ≥ 14 kHz				9.6	13.2	16.4	24.4	32.0	48.0	64.0	96.0	128.0
WB 8 kHz	5.9 VBR	7.2	8.0	9.6	13.2	16.4	24.4	32.0	48.0	64.0	96.0	128.0
NB 4 kHz	5.9 VBR	7.2	8.0	9.6	13.2	16.4	24.4					

ACELP/MDCT ⟷ MDCT

{ynh,cxh}@ustc.edu.cn

速率小结

Codec	Rate (kHz)	Bitrate (kbps)	Delay (ms)
EVS-Narrowband (NB)	8	5.9 kbps (VBR) 7.2-24.4 kbps	
EVS-Wideband (WB)	16		
EVS-Super-Wideband (SWB)	32	7.2-128 kbps	
EVS-Fullband (FB)	48	16.4-128 kbps	
AMR-NB	8	4.75-12.2	20
AMR-WB (G.722.2)	16	6.6-23.85	20
G.729	8	8	15
GSM-FR	8	13	20
GSM-EFR	8	12.2	20
G.723.1	8	5.3 6.3	37.5
G.728	8	16	0.625
G.711 (μ/A-law)	8	64	
G.722	16	48 56 64	

{ynh,cxh}@ustc.edu.cn

第三章 多媒体数据压缩

- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MPEG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

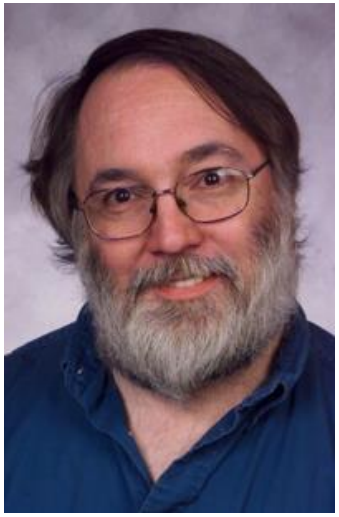
人的听觉感知机理

- 人的听觉具有掩蔽效应
  - 当几个强弱不同的声音同时存在时，强声使弱声难以听见的现象称为同时掩蔽，它受掩蔽声音和被掩蔽声音之间的相对频率关系影响很大；声音在不同时间先后发生时，强声使其周围的弱声难以听见的现象称为异时掩蔽。
- 人耳对不同频段的声音的敏感程度不同
  - 通常对低频端较之对高频端更敏感。对同样声压级的声音，人耳的实际感觉到的音量也是随频率而变化的。
- 人耳对语音信号的相位变化不敏感
  - 人耳听不到或感知极不灵敏的声音分量都视为冗余

{ynh,cxh}@ustc.edu.cn

Perceptual Entropy(感知熵)

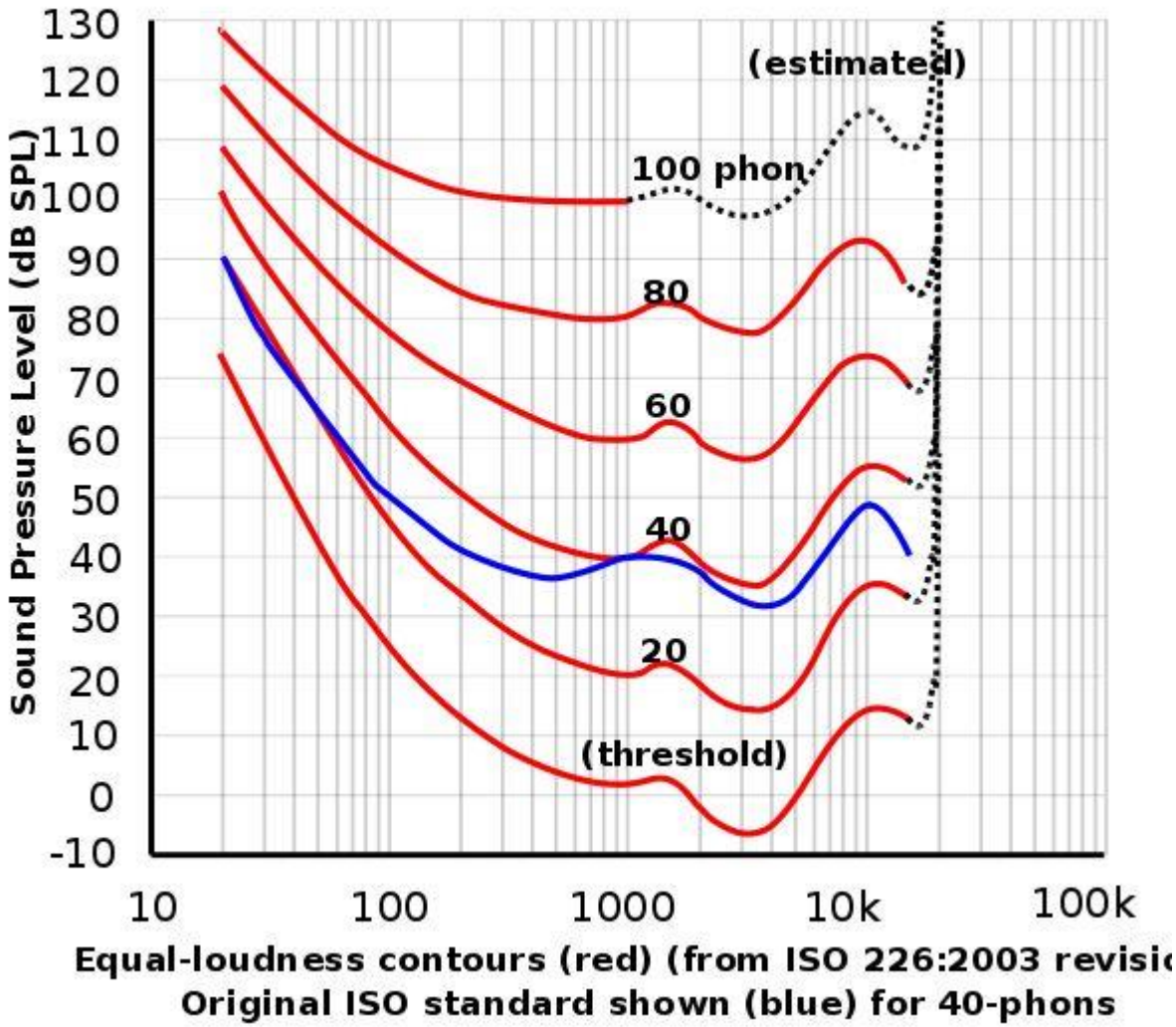
- Perceptual Entropy, objective metric of perceptually relevant introduced by J. Johnston
- The perceived information from an audio signal is only a fraction of the total information emanated by the source.



Ted Painter and Andreas Spanias. *Perceptual coding of digital audio*. Proceedings of the IEEE, 88(4):449-513. April 2000.

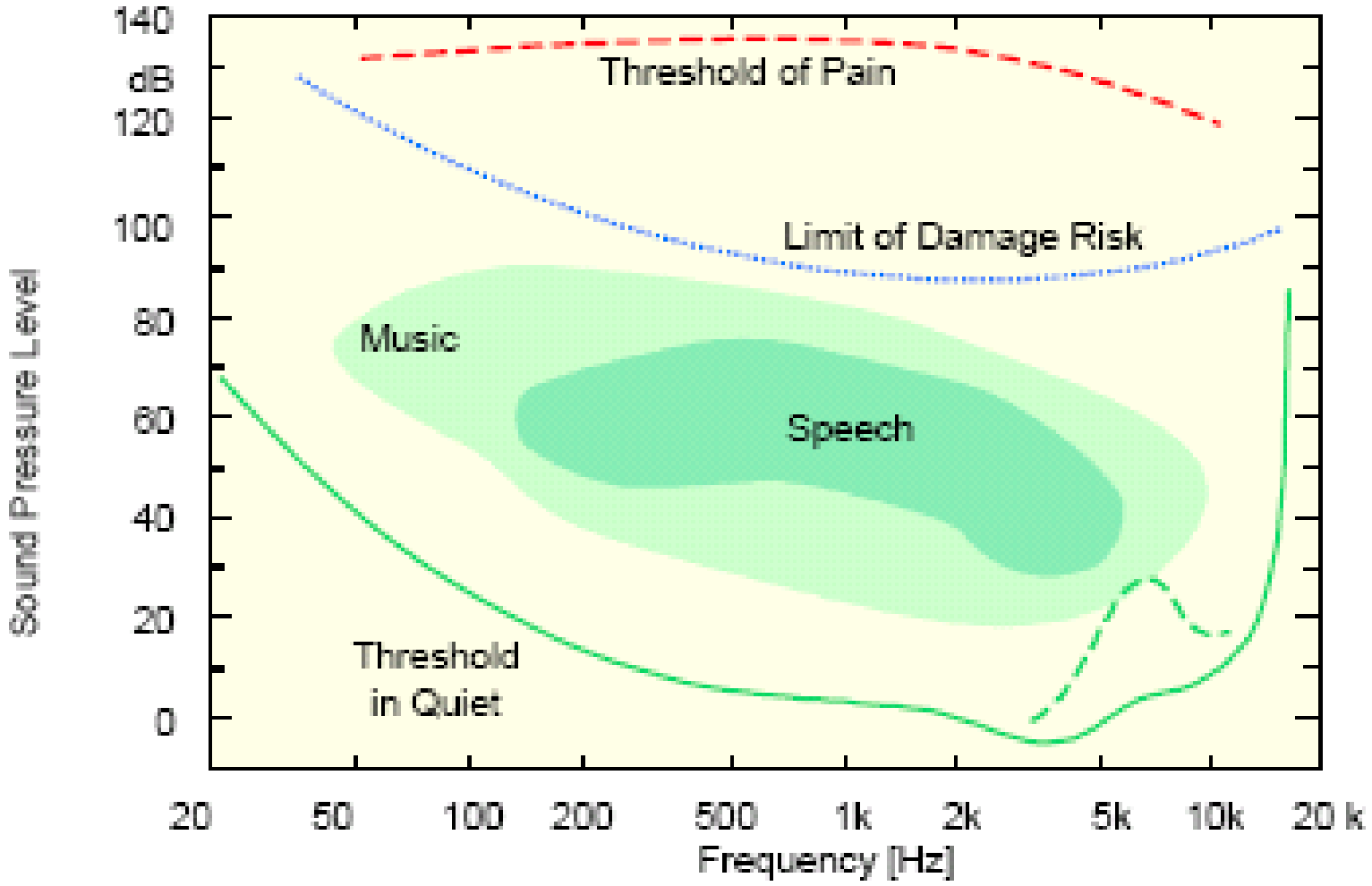
{ynh,cxh}@ustc.edu.cn

人耳对不同频率声音的敏感度不同



{ynh,cxh}@ustc.edu.cn

Sound Pressure Level

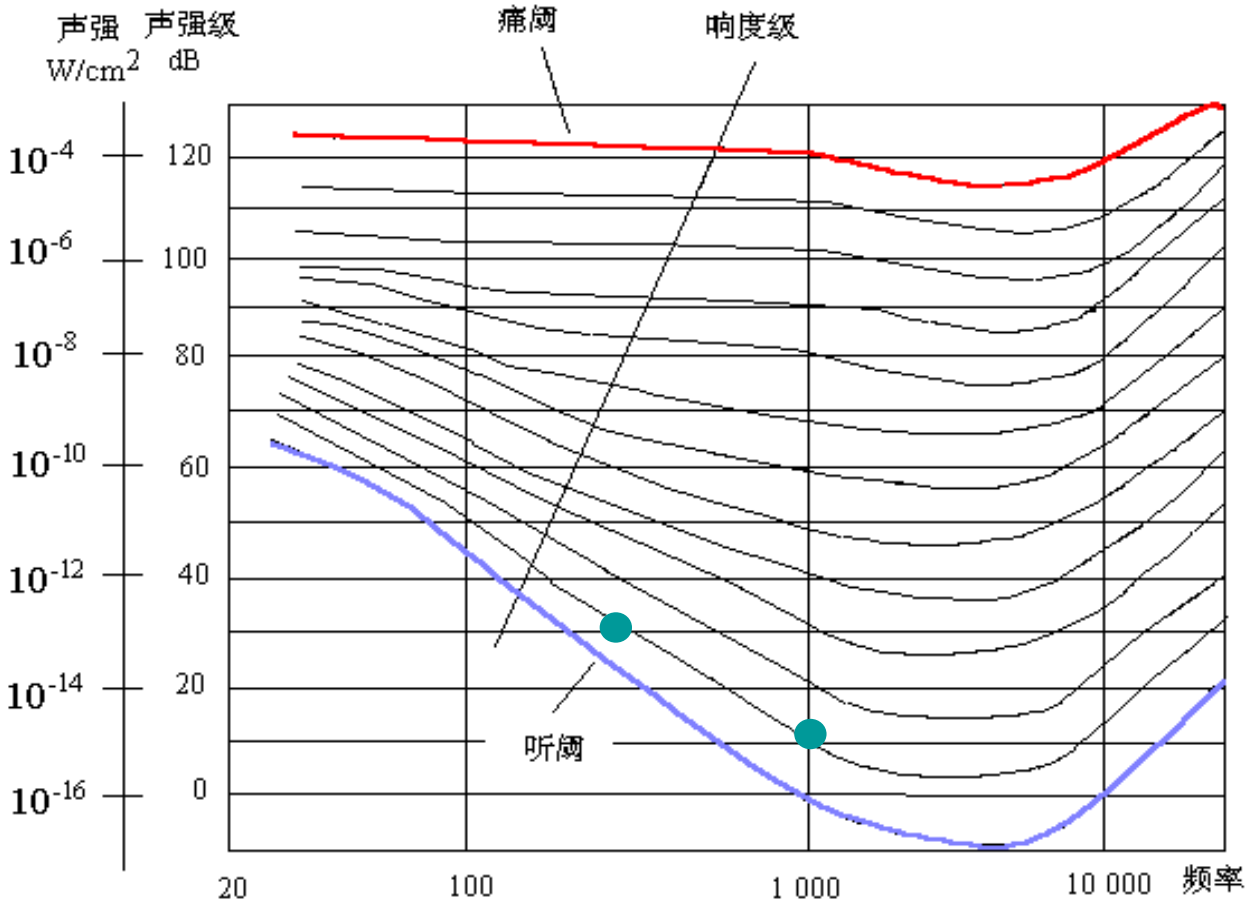


http://www.transport07.co.nz/uploads/File/ppt/C%20Roberts.ppt

{ynh,cxh}@ustc.edu.cn

“听阈—频率”和“痛阈—频率”曲线

◆1 kHz的10 dB的声音和200 Hz的30 dB的声音，在人耳听起来具有相同的响度



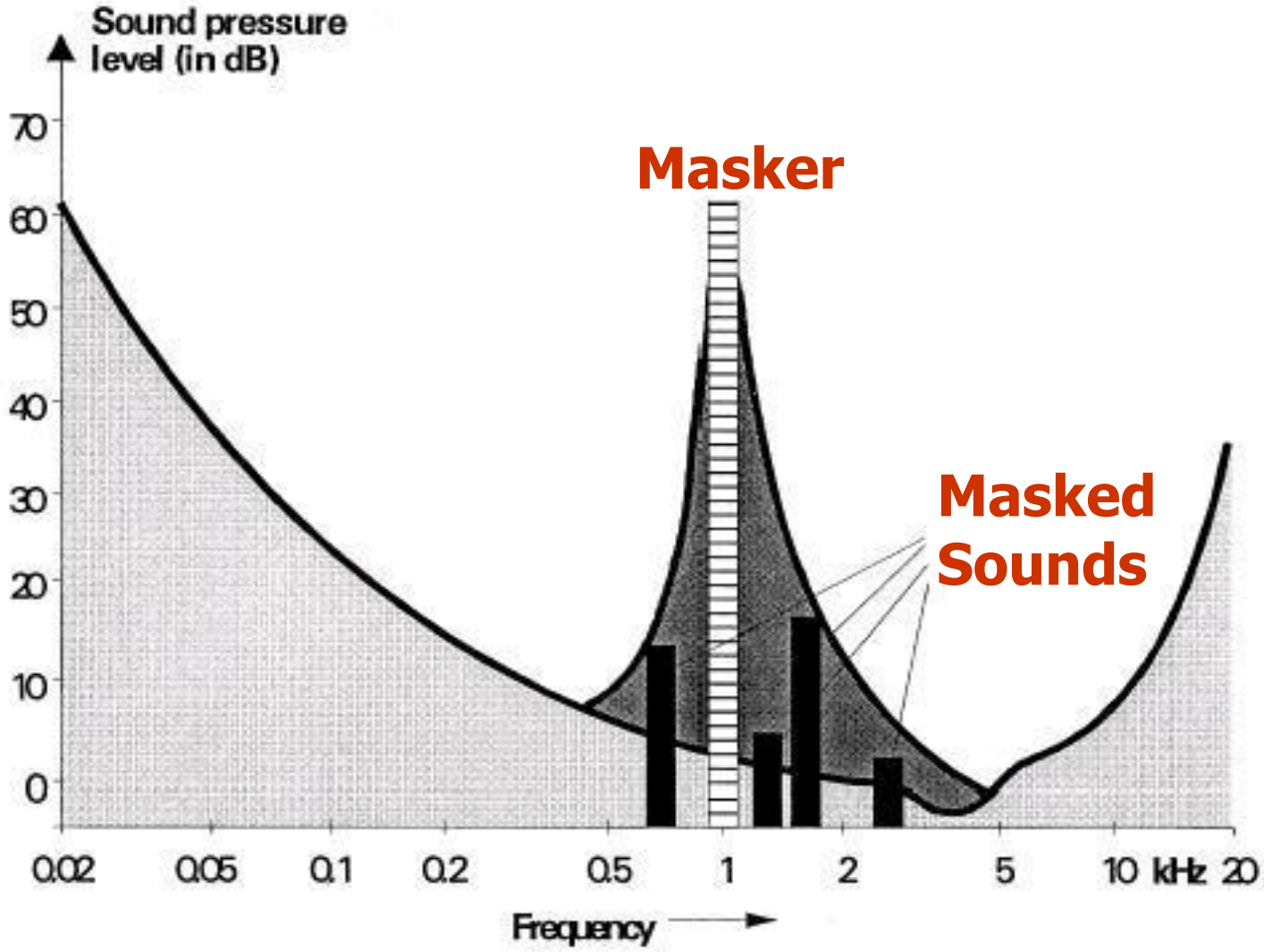
{ynh,cxh}@ustc.edu.cn

听觉掩蔽效应

- 一种频率的声音阻碍听觉系统感受另一种频率的声音的现象称为掩蔽效应。
- 频域掩蔽：同时发出的频率接近的两个纯音，声强低的纯音会被声强高的纯音淹没
- 时域掩蔽：在时间上相邻的声音之间也有掩蔽现象，称为时域掩蔽。产生的主要原因是人的大脑处理信息需要花费一定的时间。

{ynh,cxh}@ustc.edu.cn

频域掩蔽


Sound pressure level (in dB)
70
60
50
40
30
20
10
0
Masker
Masked Sounds
0.02
0.05
0.1
0.2
0.5
1
2
5
10 kHz 20
Frequency

时域掩蔽

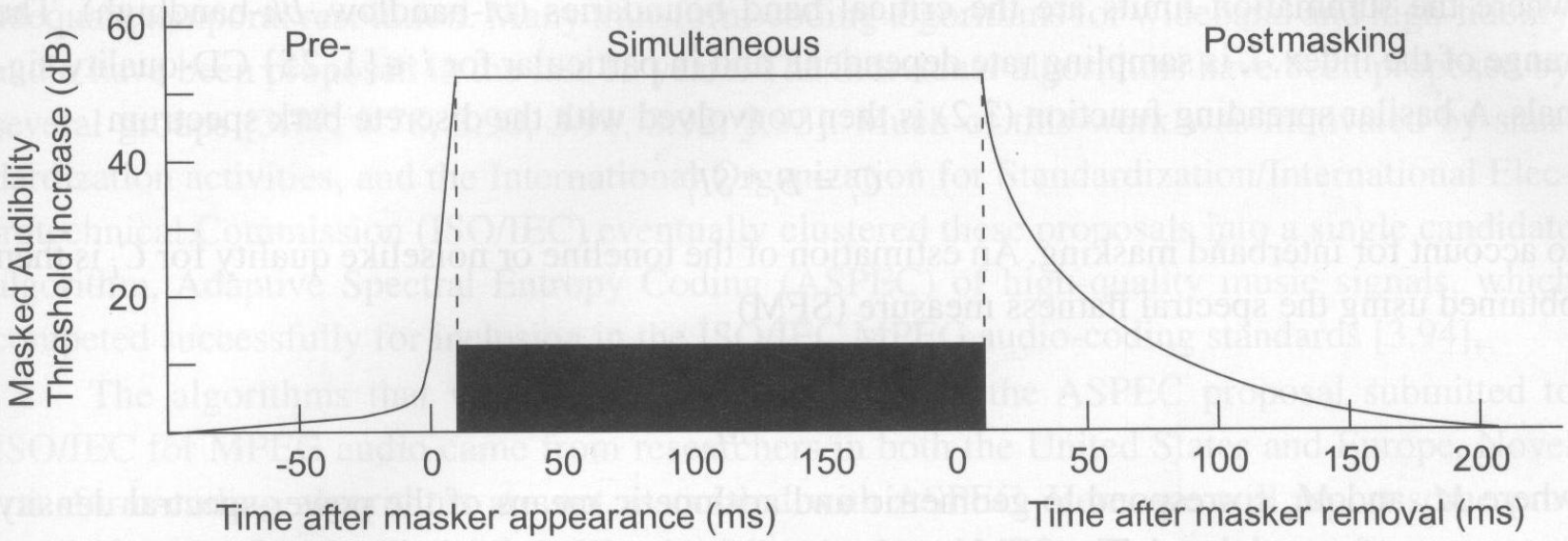
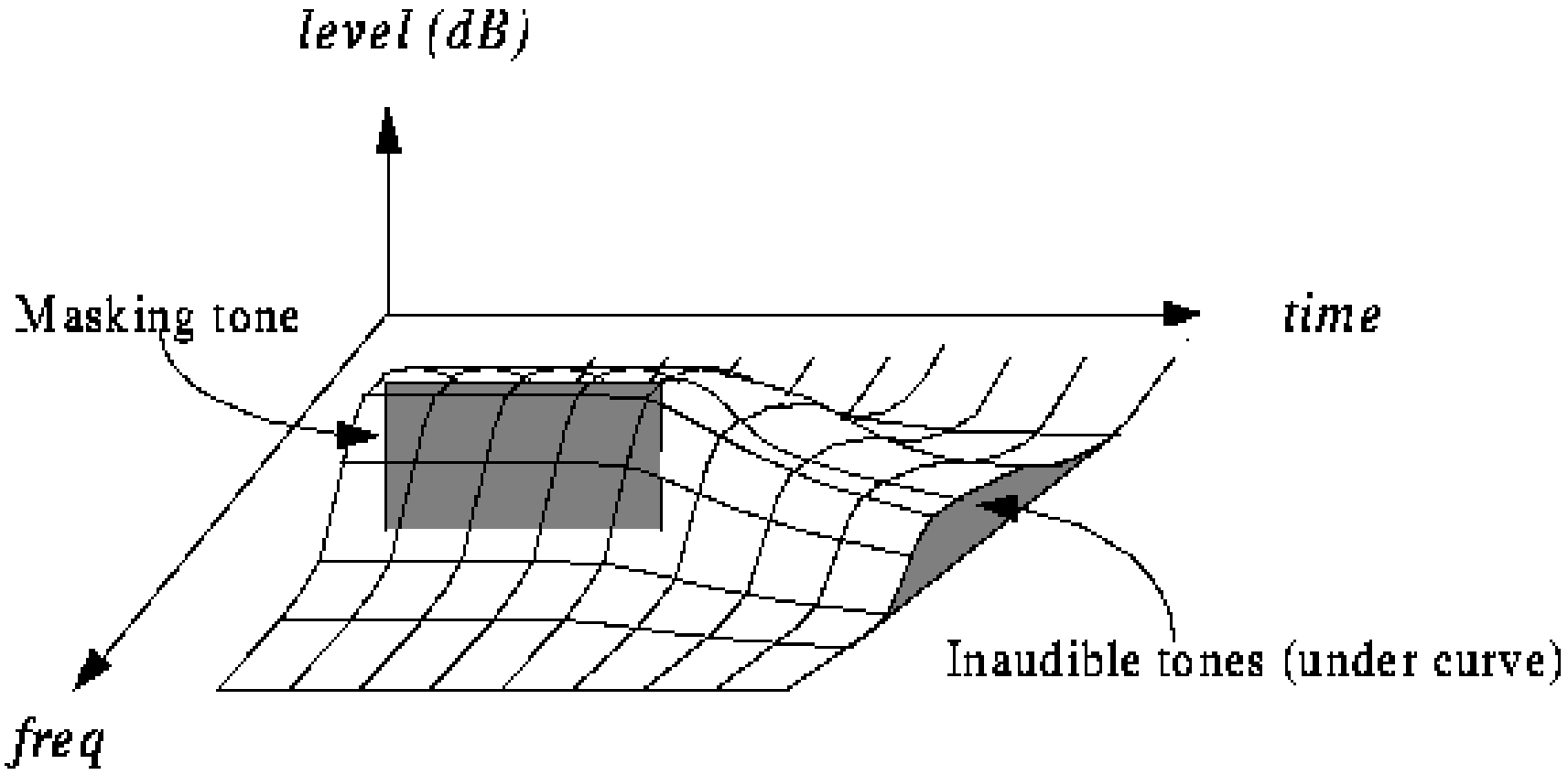


Figure 3.7 Schematic representation of temporal masking properties of the human ear [3.82]. ©1997 IEEE.

{ynh,cxh}@ustc.edu.cn

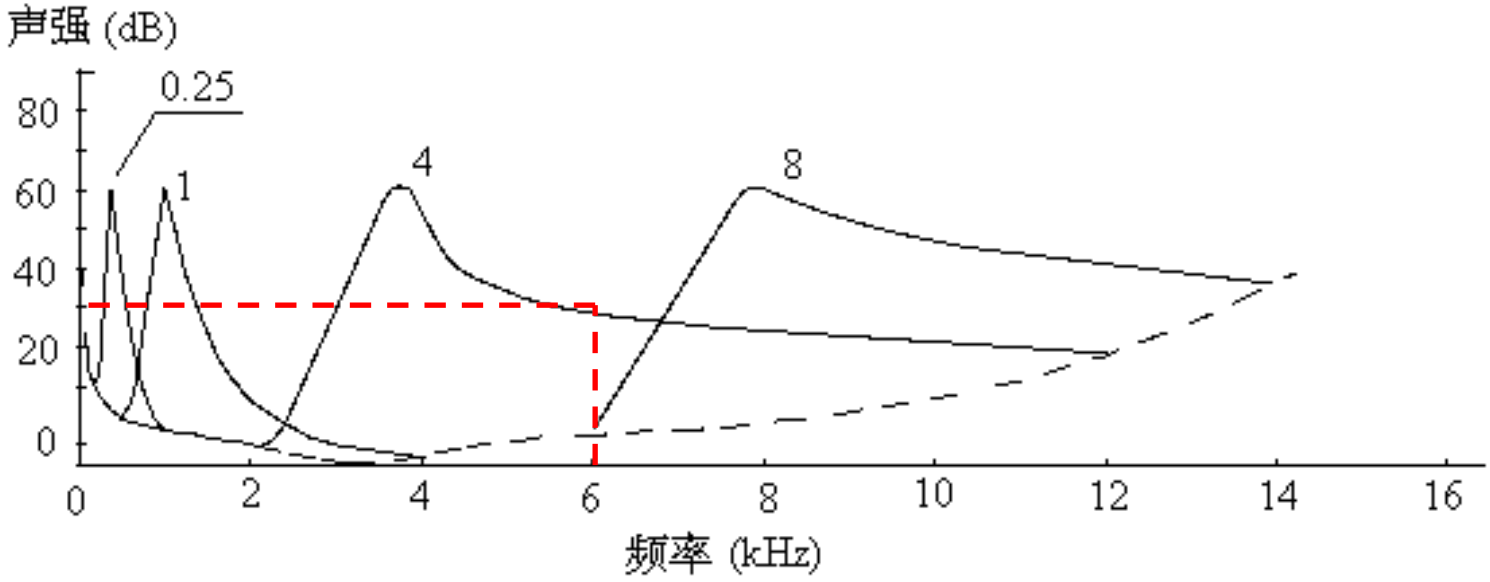
频域掩蔽+时域掩蔽



{ynh,cxh}@ustc.edu.cn

不同纯音的掩蔽效应曲线

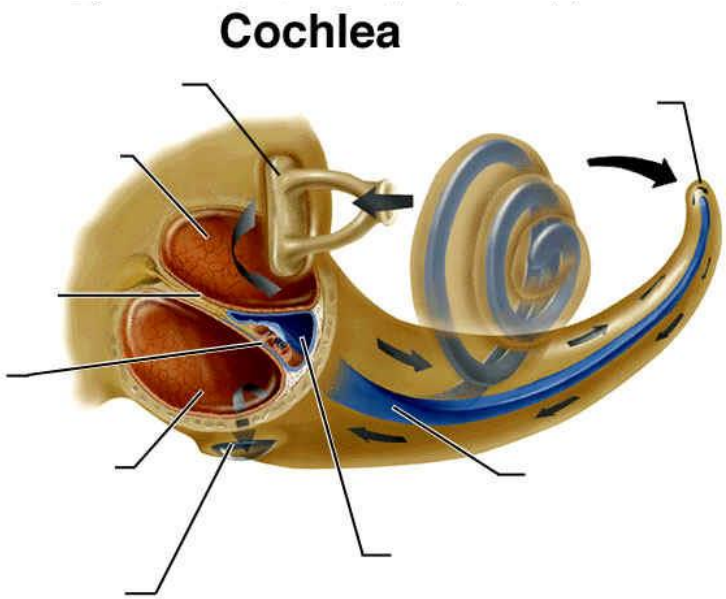
◆图中的一组曲线分别表示频率为250 Hz、1 kHz、4 kHz和8 kHz纯音的掩蔽效应。从图中可以看到：①在纯音附近，对其他纯音的掩蔽效果最明显，②低频纯音可以有效地掩蔽高频纯音，但高频纯音对低频纯音的掩蔽作用则不明显。



临界频带

Critical Bands

Concept introduced by Harvey Fletcher 1940.



耳蜗

Frequency to Place Transform.
Function of frequency that quantifies the cochlear filter passbands.

Example: The critical band for a 1kHz is about 160Hz in width.

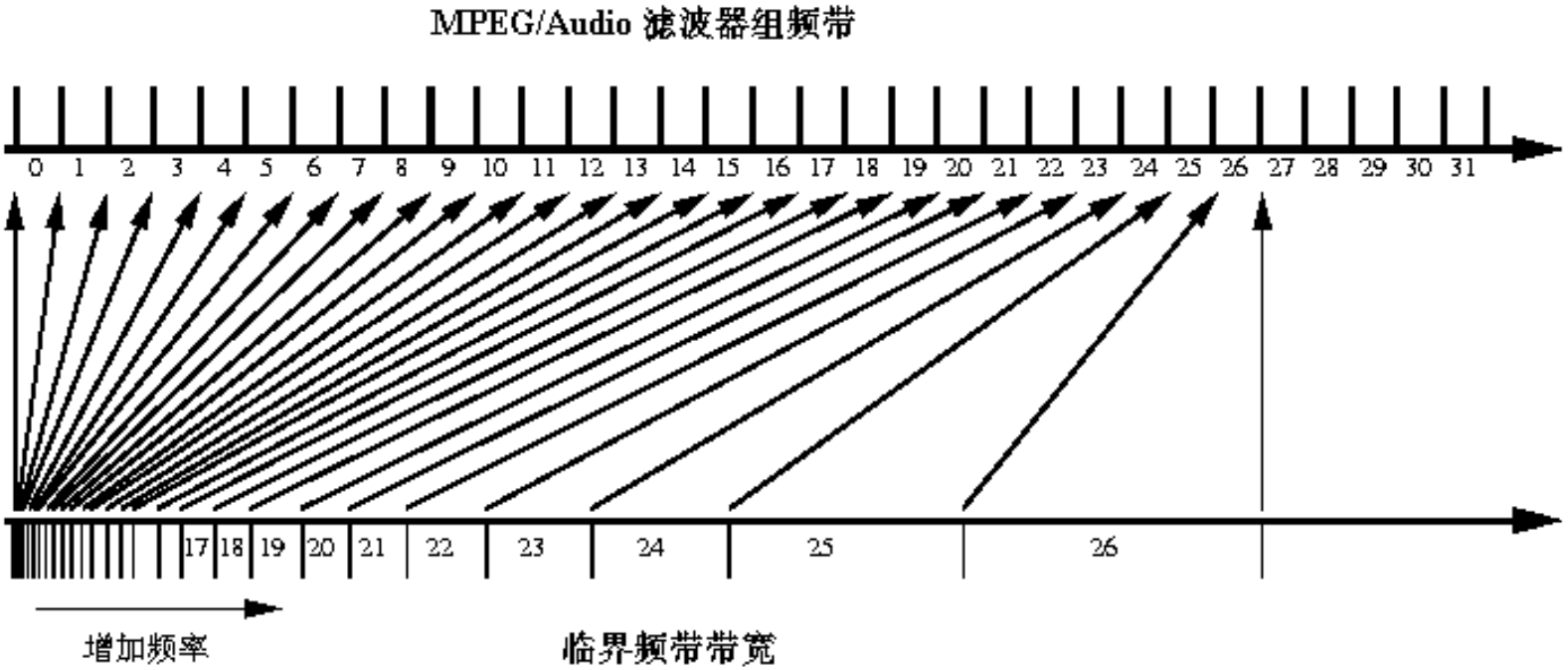
A narrow band noise centered at 1kHz is perceived with the same loudness as long as the width < 160Hz.

$$BW_c(f) = 25 + 75[1 + 1.4(f/1000)^2]^{0.69} \text{ (Hz)}$$

{ynh,cxh}@ustc.edu.cn

子带的划分（临界频带）

◆方法有两种，一种是线性划分，另一种是非线性划分。如果把声音频带划分成带宽相等的子带，这种划分就不能精确地反映人耳的听觉特性，因为人耳的听觉特性是以“临界频带”来划分的，在一个临界频带之内，很多心理声学特性都是一样的。

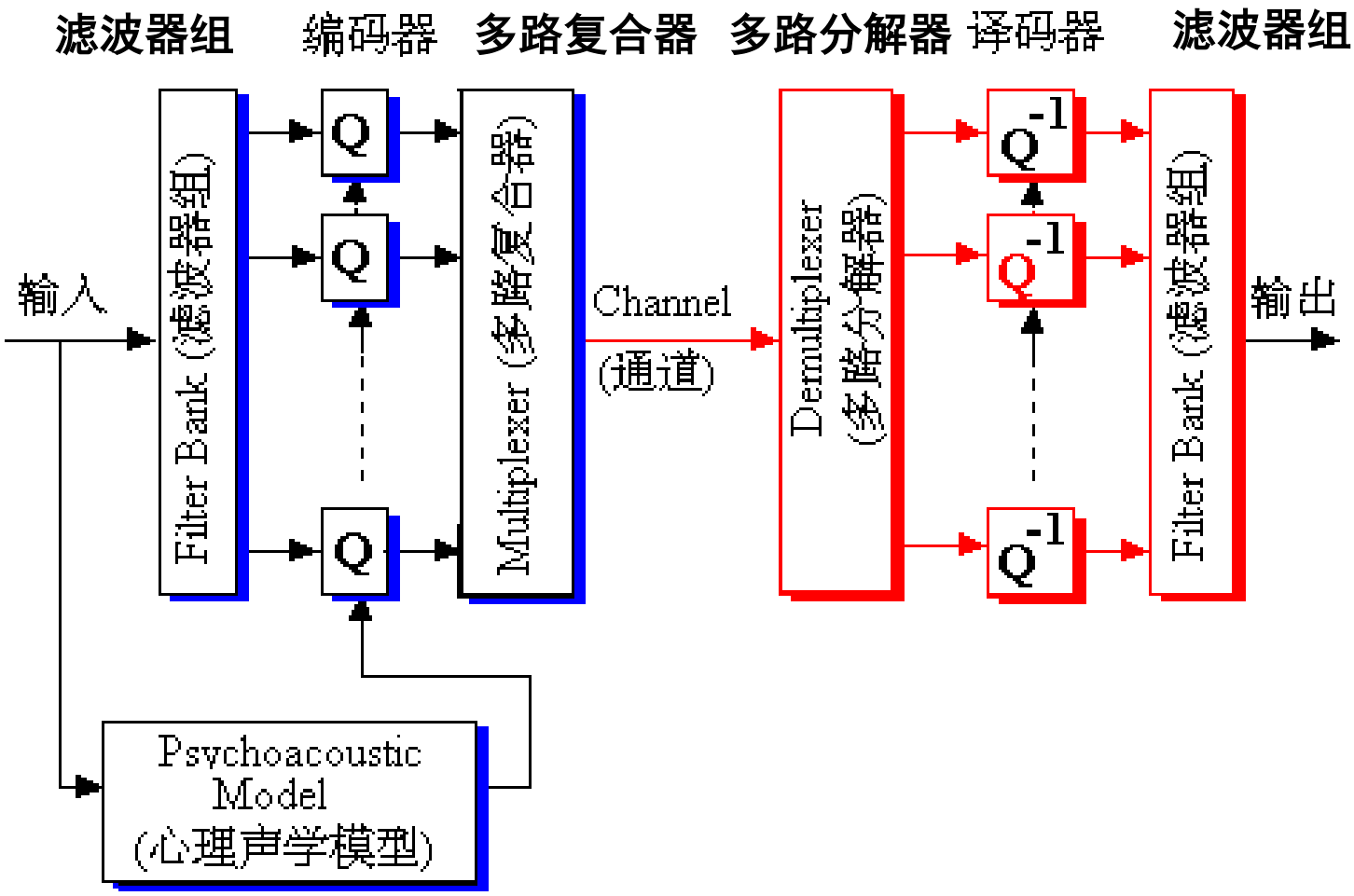


利用感知特性进行压缩编码

- 人耳对不同临界频带感知不同，可以将临界频带分成子带采用不同的量化阶
- 听觉系统中存在一个听阈电平，低于这个电平的声音信号就听不到，因此就可以把这部分信号去掉。
- 利用掩蔽效应将人耳无法感知的频率分量消除

{ynh,cxh}@ustc.edu.cn

MPEG Audio压缩算法框图



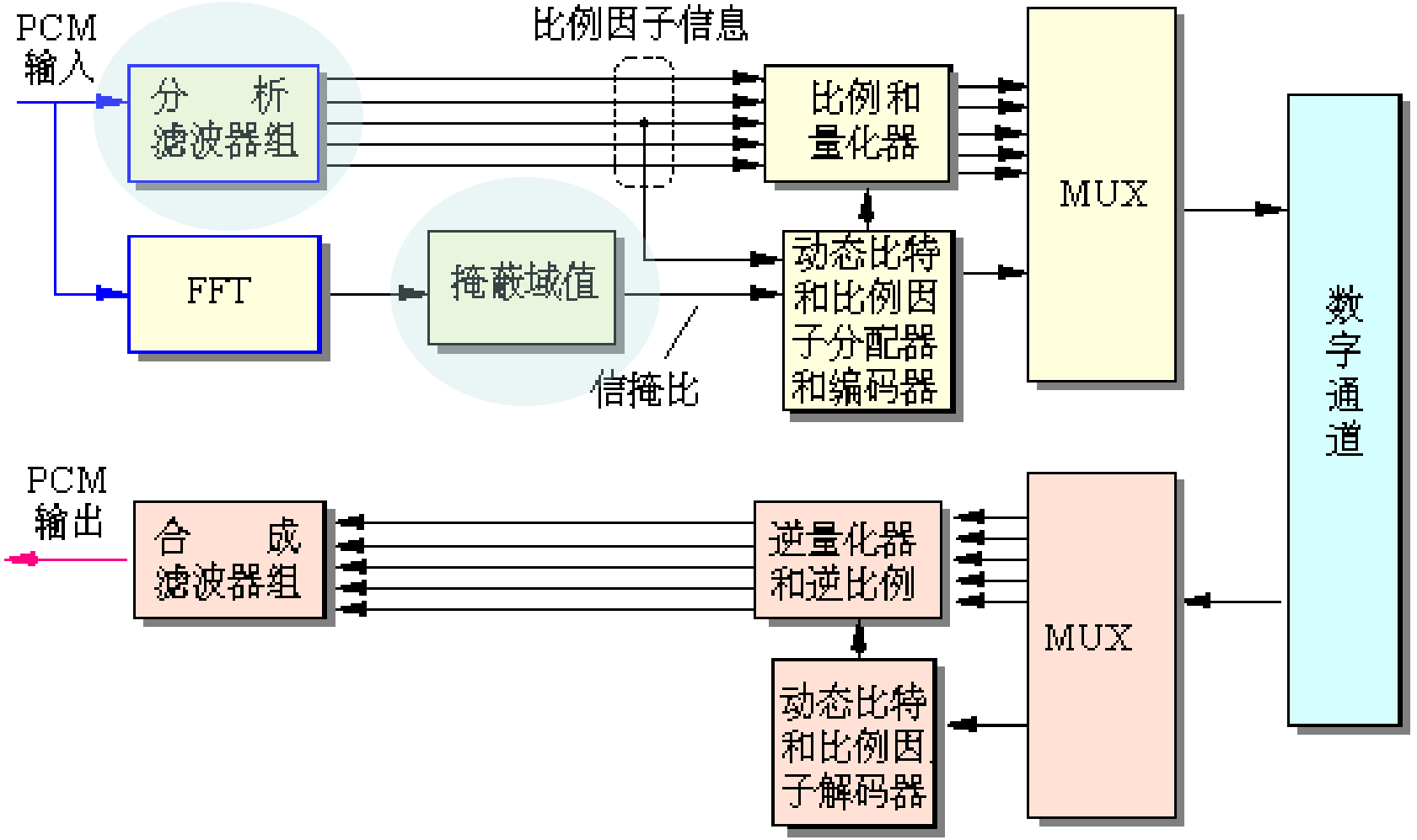
{ynh,cxh}@ustc.edu.cn

MPEG1 Audio

- Layer1
  - 频带相等的子带，使用频域掩蔽特性， 每个子带用6比特量化。
- Layer2
  - 频带相等的子带，除了使用频域掩蔽特性之外还利用了时间掩蔽特性，低频段的子带用4比特，中频段的子带用3比特，高频段的子带用2比特。
- Layer3
  - 层3使用了临界频带滤波器，把声音频带分成非等带宽的子带，除了使用频域掩蔽特性和时间掩蔽特性之外，还考虑了立体声数据的冗余，并且使用了霍夫曼(Huffman)编码器。还使用了MDCT (modified discrete cosine transform) 把子带的输出在频域里进一步细分以达到更高的频域分辨率。

{ynh,cxh}@ustc.edu.cn

MPEG1 audio层1和层2编解码



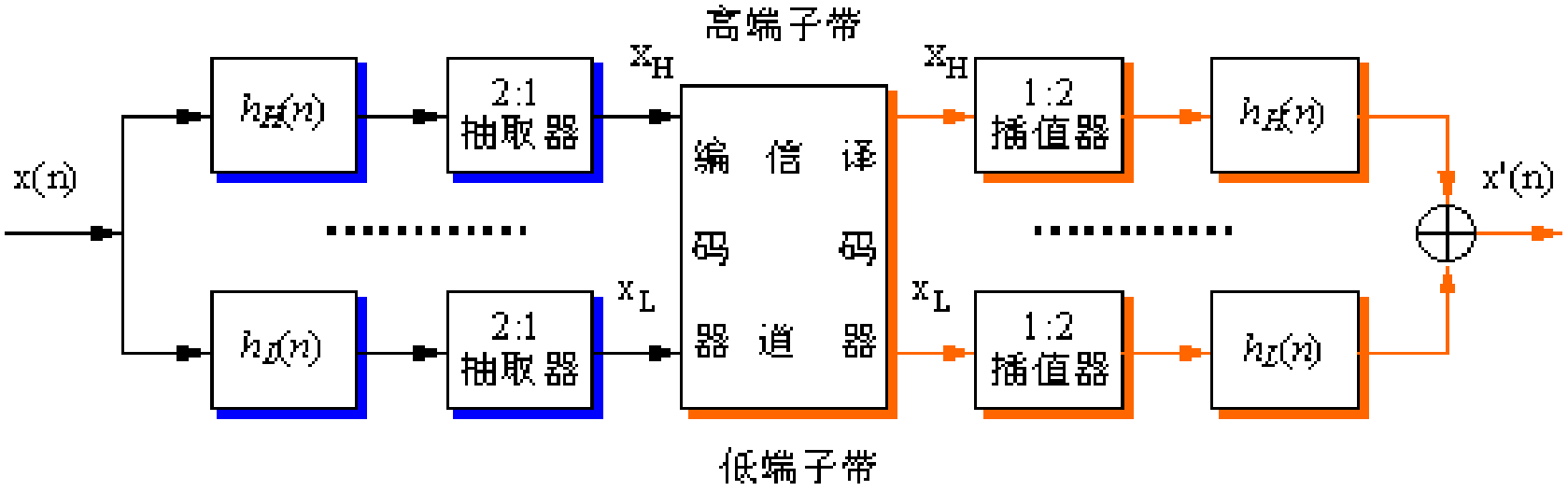
{ynh,cxh}@ustc.edu.cn

临界频带

临界	频率（Hz）			临界	频率（Hz）		
频带	低端	高端	宽度	频带	低端	高端	宽度
0	0	100	100	13	2000	2320	320
1	100	200	100	14	2320	2700	380
2	200	300	100	15	2700	3150	450
3	300	400	100	16	3150	3700	550
4	400	510	110	17	3700	4400	700
5	510	630	120	18	4400	5300	900
6	630	770	140	19	5300	6400	1100
7	770	920	150	20	6400	7700	1300
8	920	1080	160	21	7700	9500	1800
9	1080	1270	190	22	9500	12000	2500
10	1270	1480	210	23	12000	15500	3500
11	1480	1720	240	24	15500	22050	6550
12	1720	2000	280				

子带分割

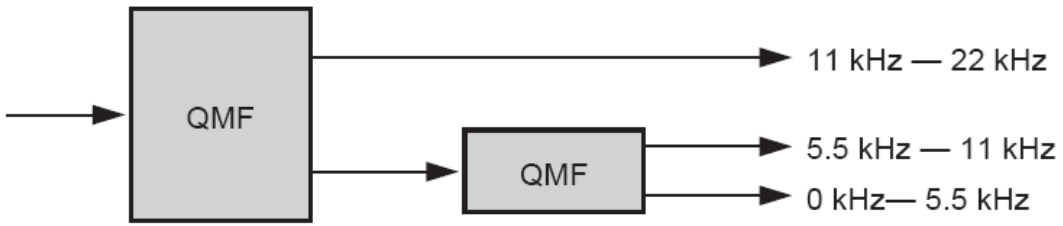
◆把音频信号分割成相邻的子带分量之后，用2倍于子带带宽的采样频率对子带信号进行采样，就可以用它的样本值重构出原来的子带信号。

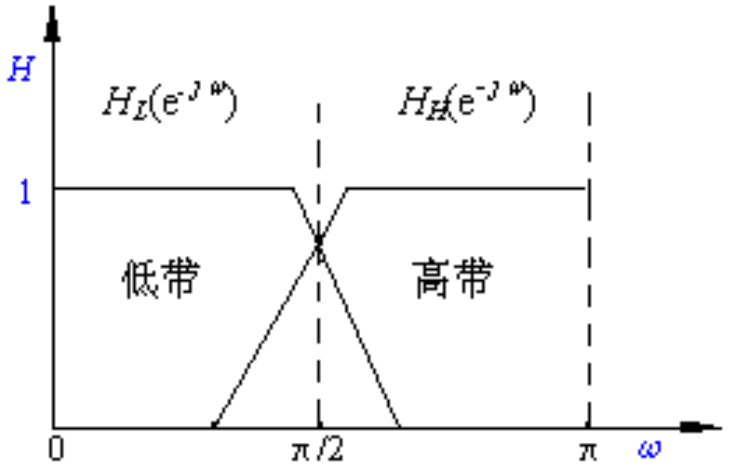


{ynh,cxh}@ustc.edu.cn

正交镜象滤波器
QMF, quandrature mirror filter

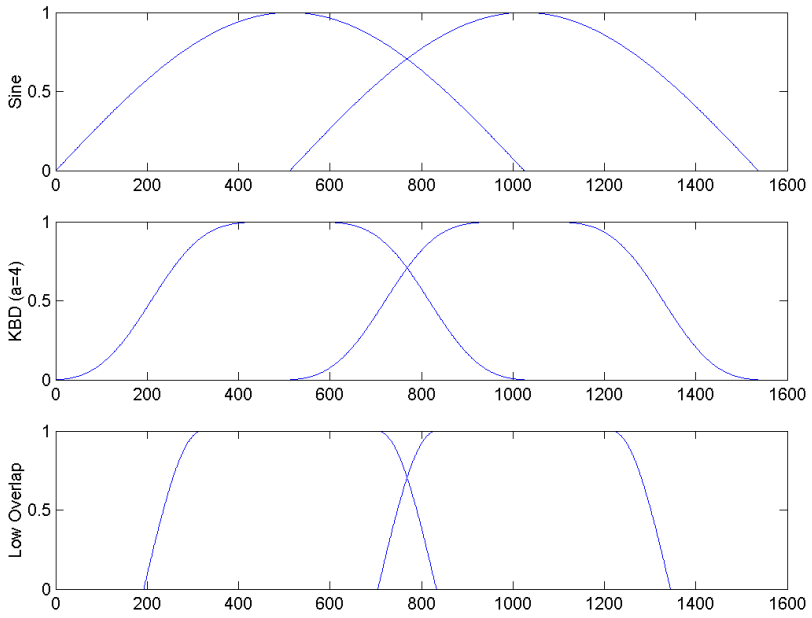
◆由于分割频带所用的滤波器不是理想的滤波器，经过分带、编码、译码后合成的输出音频信号会有混迭效应。采用正交镜象滤波器QMF来划分频带，混迭效应在最后合成时可以抵消。

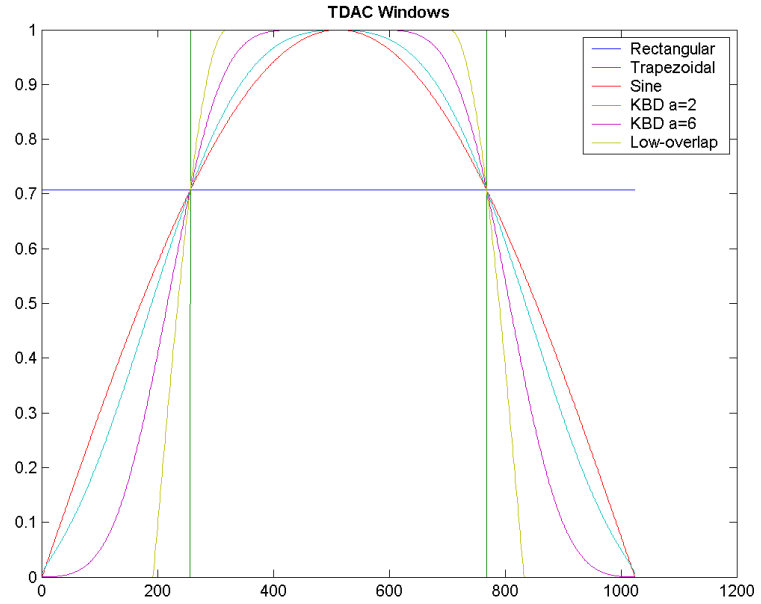




Kaiser-Bessel Derived (KBD) window

- A 50% overlap add (OLA) structure with certain pre and post, time domain aliasing cancellation (TDAC) windowing, the initial signal can be completely recovered.





MDCT

In particular, it is a linear function $F: \mathbf{R}^{2N} \to \mathbf{R}^{N}$ (where **R** denotes the set of real numbers). The $2N$ real numbers $x_0, \ldots, x_{2N-1}$ are transformed into the N real numbers $X_0, \ldots, X_{N-1}$ according to the formula:

$$X_k = \sum_{n=0}^{2N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right]$$

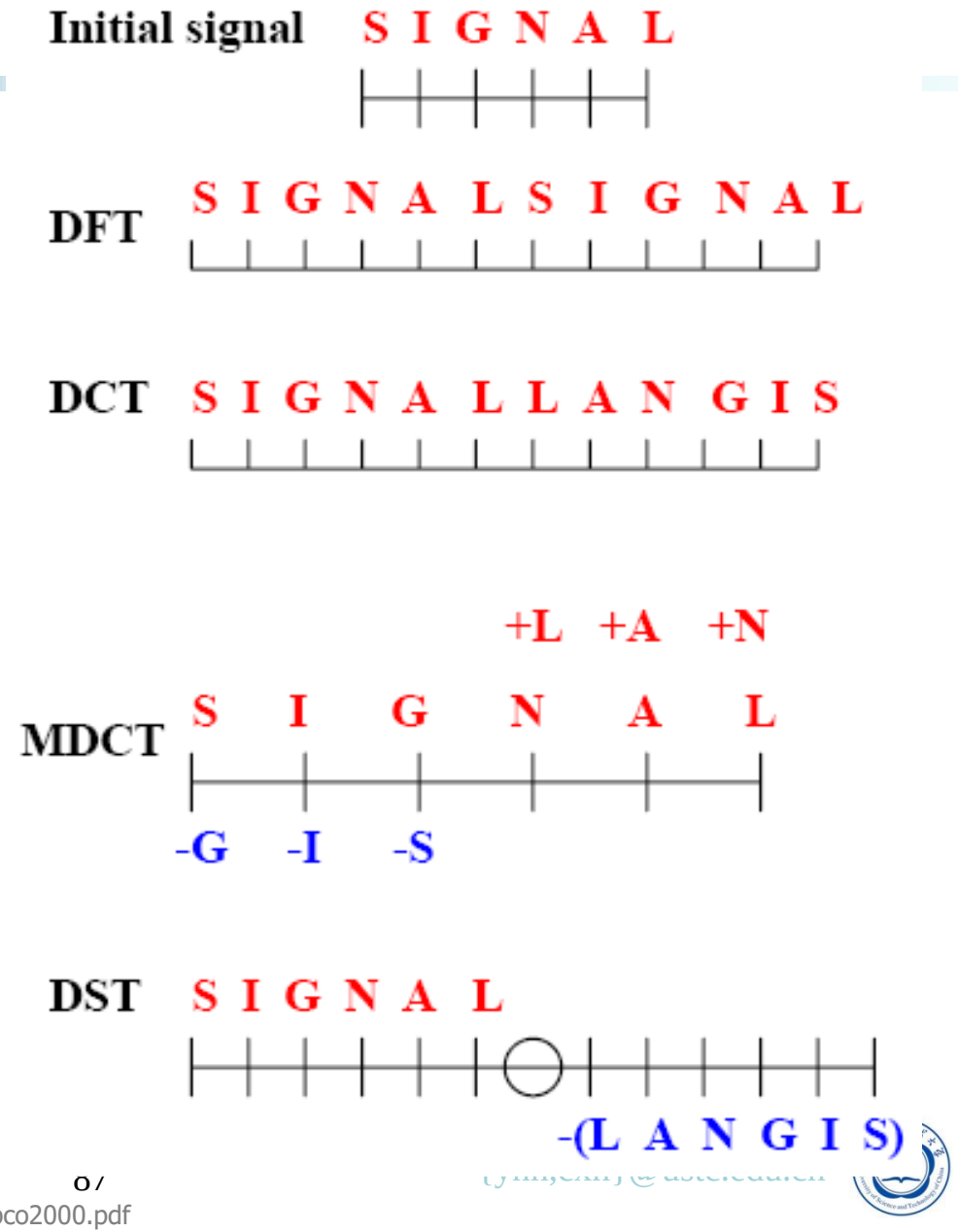
The inverse MDCT is known as the **IMDCT**. Because there are different numbers of inputs and outputs, at first glance it might seem that the MDCT should not be invertible. However, perfect invertibility is achieved by *adding* the overlapped IMDCTs of subsequent overlapping blocks, causing the errors to *cancel* and the original data to be retrieved; this technique is known as *time-domain aliasing cancellation* (**TDAC**).

The IMDCT transforms N real numbers $X_0, \ldots, X_{N-1}$ into $2N$ real numbers $y_0, \ldots, y_{2N-1}$ according to the formula:

$$y_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right]$$

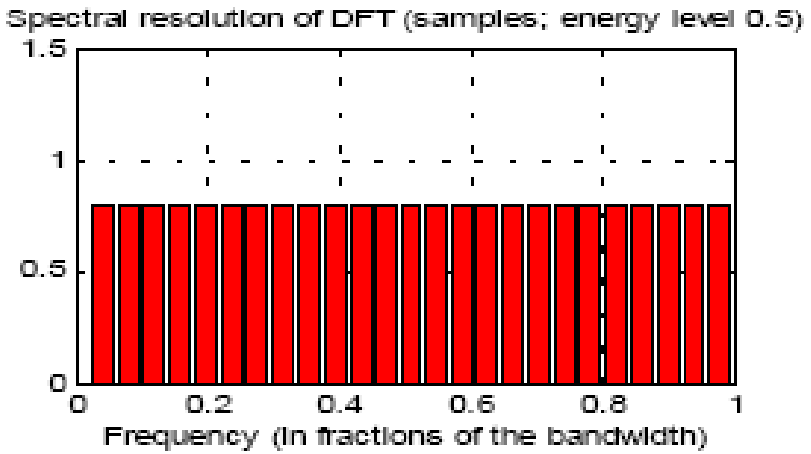
Sampling

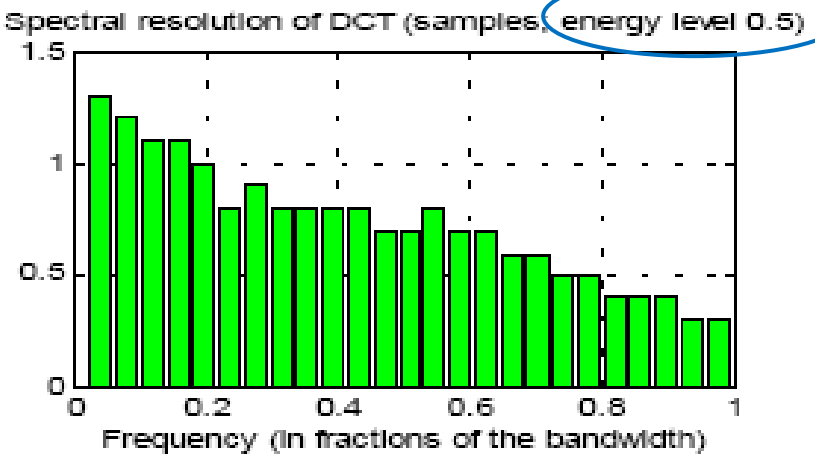
◆不同变换在时域采样的方式是不同的；对应的获得了不同的频域分辨率

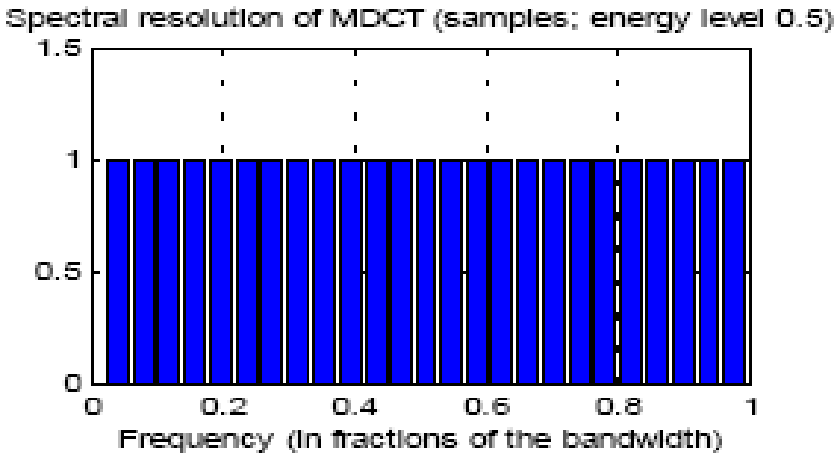


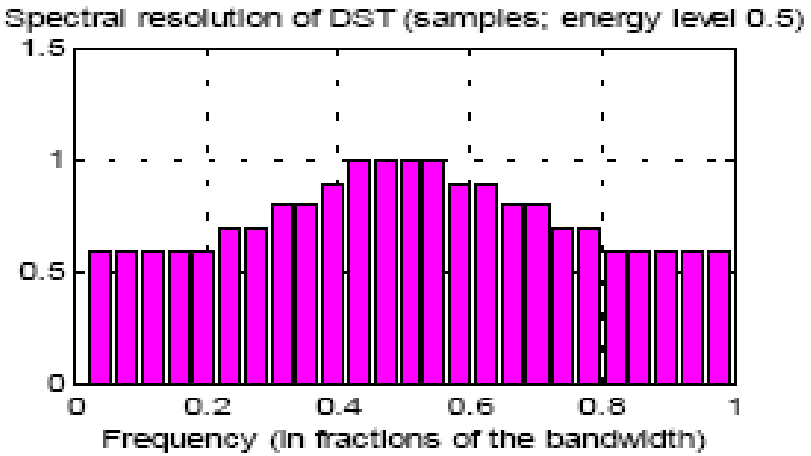
URL http://www.comp.nus.edu.sg/~wangye/papers/Others/cr1207_eusipco2000.pdf

Transform energy compaction capability



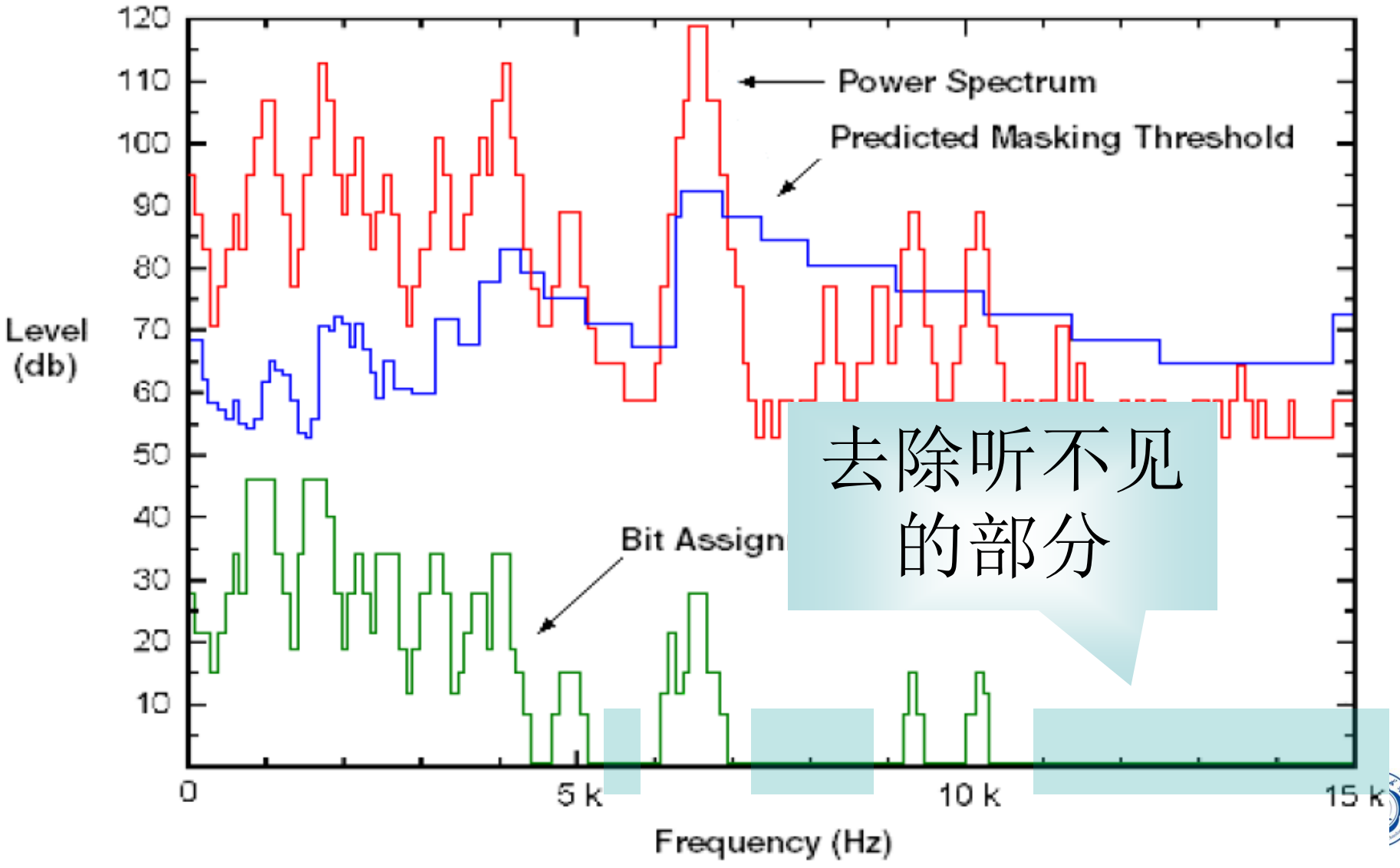






{ynh,cxh}@ustc.edu.cn

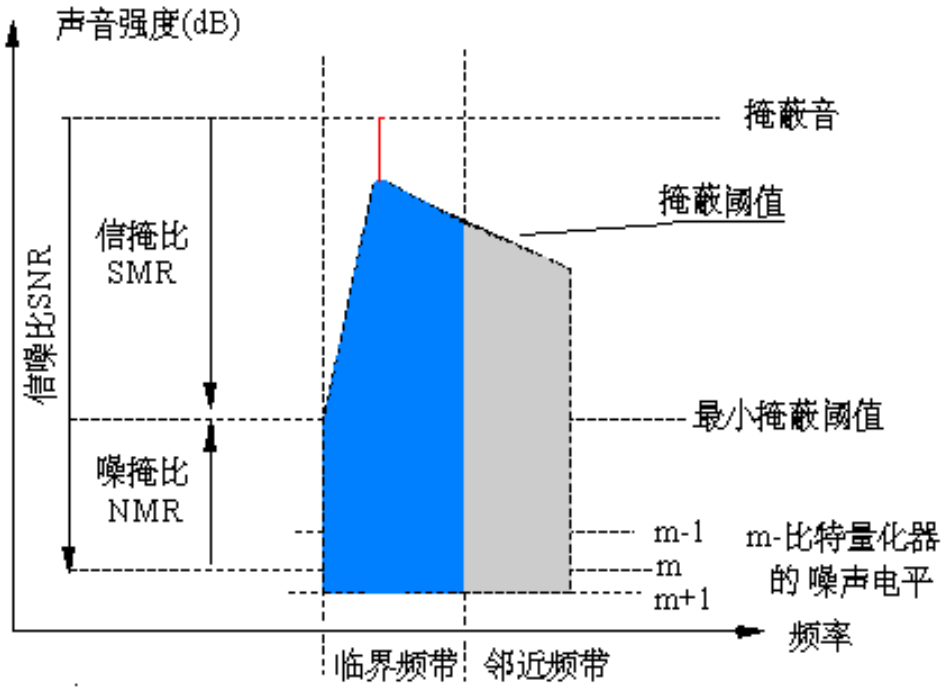
掩蔽特性应用举例

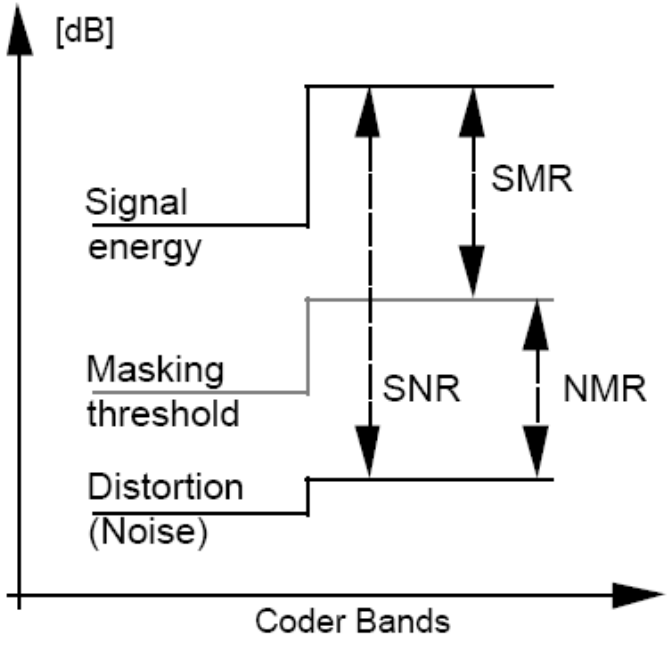

120
110
100
90
80
70
60
50
40
30
20
10
Level (db)
Power Spectrum
Predicted Masking Threshold
Bit Assign
去除听不见的部分
0
5 k
10 k
15 k
Frequency (Hz)

恒定质量量化方案

- 能量在掩蔽线以下的噪声人耳就不能感知了
- 故可以根据SMR来决定分配给子带信号的量化位数，使量化噪声低于掩蔽阈值

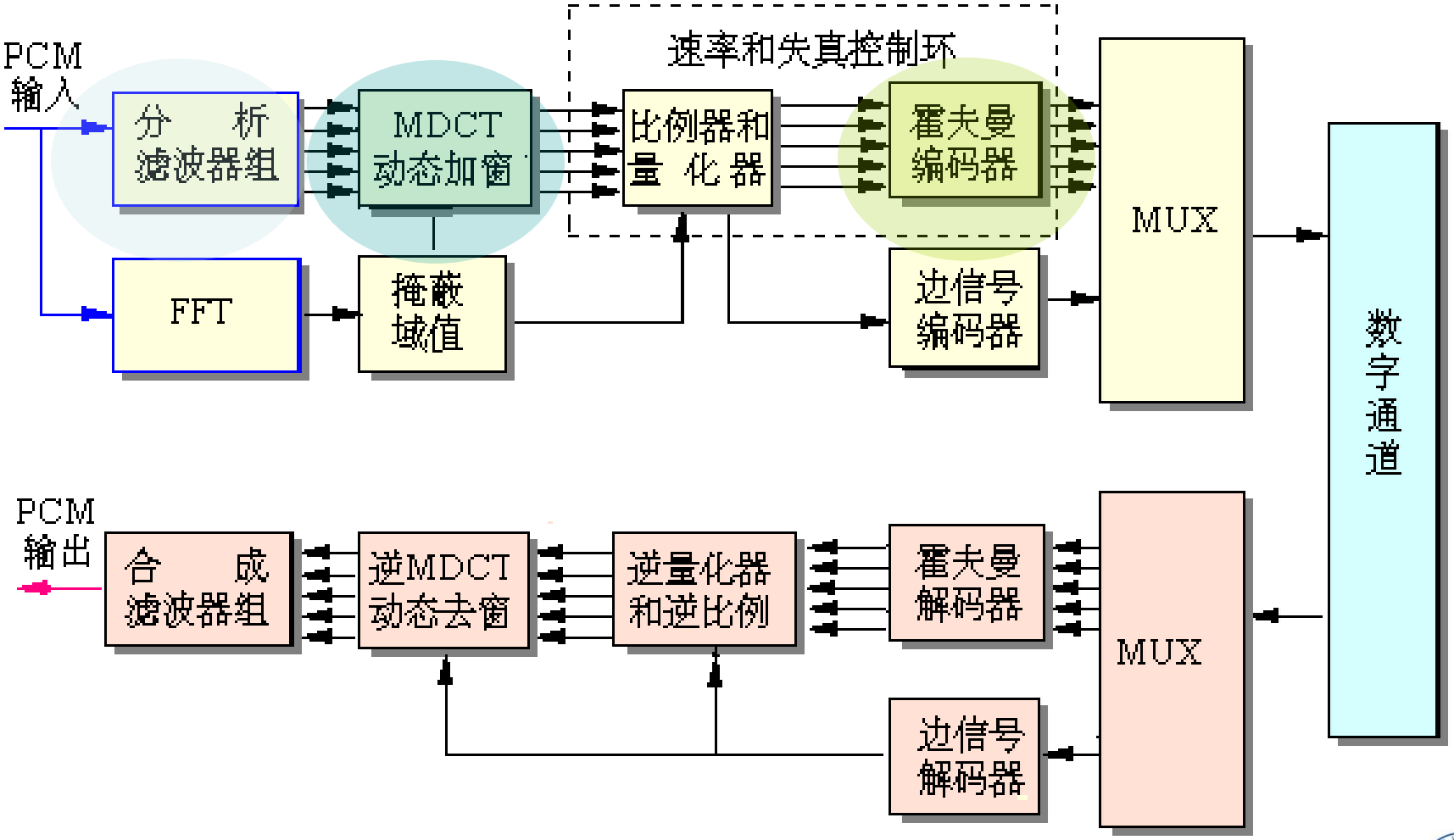
http://www.ee.columbia.edu/~marios/courses/e6820y02/





{ynh,cxh}@ustc.edu.cn

MP3编解码



{ynh,cxh}@ustc.edu.cn

讨论

- SB-ADPCM和MPEG Audio都是利用子带分割的思想，它们之间存在哪些不同？

- Most important
- SB-ADPCM（时域）
- MPEG Audio（频域）

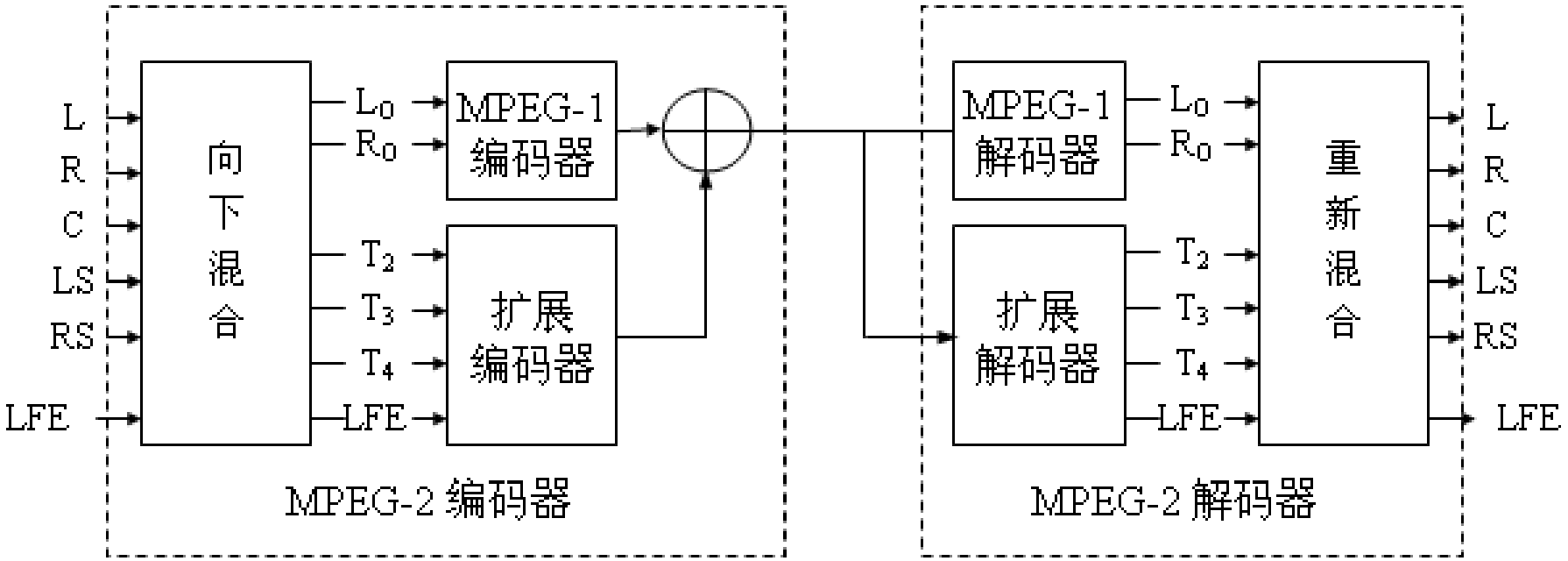
{ynh,cxh}@ustc.edu.cn

MPEG2 Audio

- MPEG-2 BC (Backward Compatible)
  - 增加了16 kHz, 22.05 kHz和24 kHz采样频率
  - 输出速率由32～384 kb/s扩展到8～640 kb/s
  - 支持5.1声道和7.1声道的环绕声
  - 支持Linear PCM(线性PCM)和Dolby AC-3(Audio Code Number 3)编码
- MPEG-2 AAC (Advanced Audio Coding)
  - 利用掩蔽特性减少数据量，并把量化噪声分散到各个子带中，用全局信号把噪声掩蔽掉。
  - 采用频率可从8 kHz到96 kHz，可支持声道数目极多

{ynh,cxh}@ustc.edu.cn

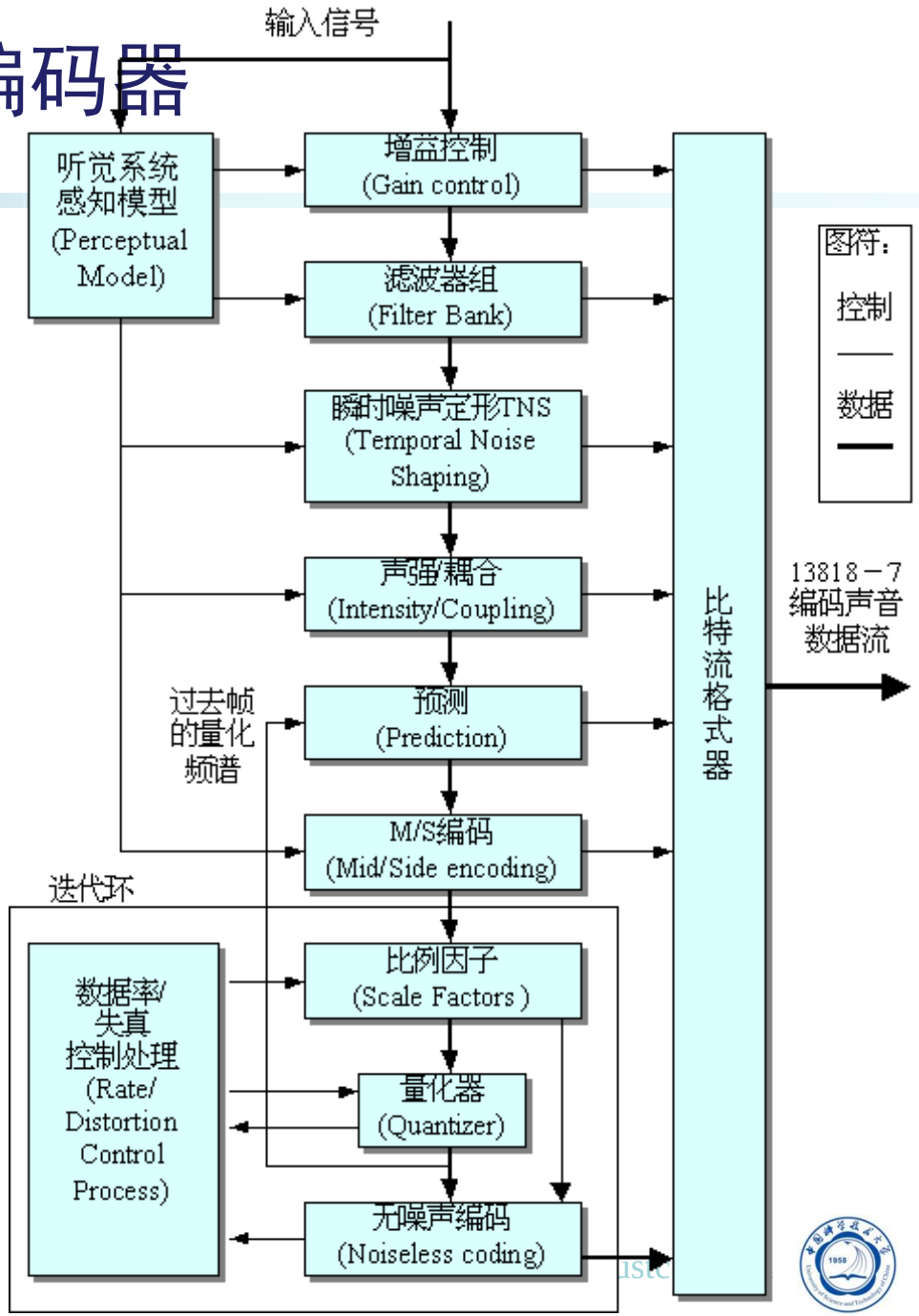
MPEG-2 BC编码器/解码器



{ynh,cxh}@ustc.edu.cn

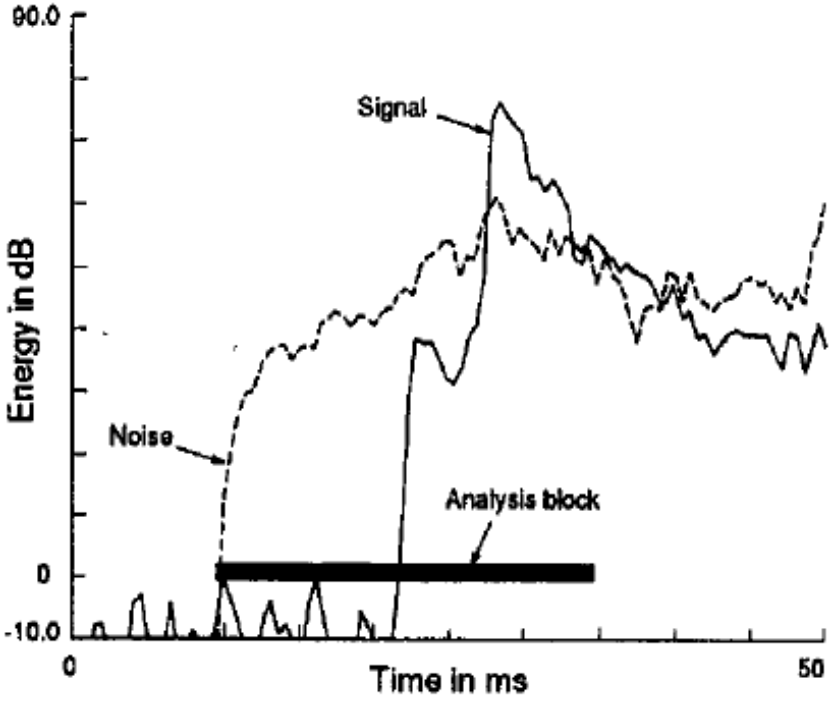
MPEG-2 AAC编码器

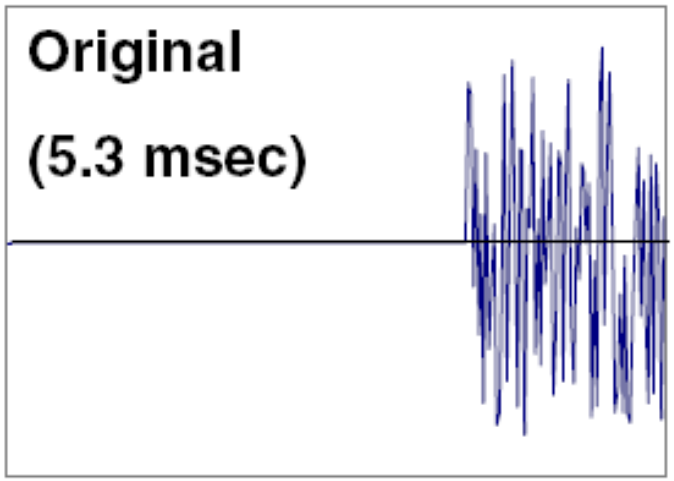
- 增益控制
- 滤波器组
  - MDCT/TDAC
- 瞬时噪声定形
- 声强/耦合
  - 单声道信号加上位置信息
- 预测
- Mid/Side
  - 左右声道转换为中央M(middle)和边S(side)声道
- 比例因子
- 量化器
- 无噪声编码
  - 霍夫曼编码

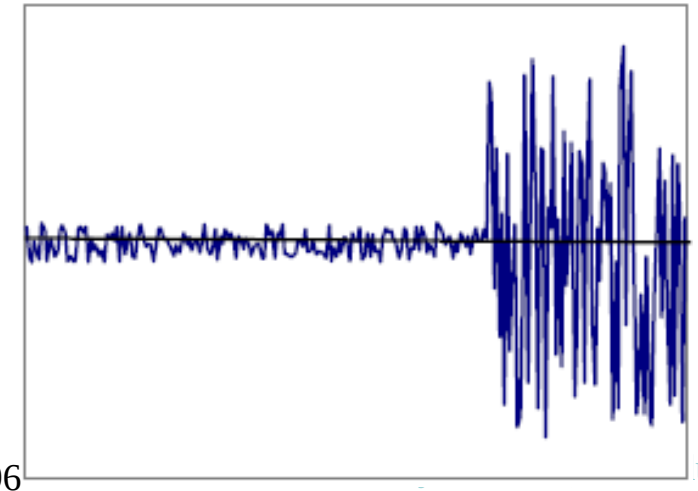


Pre-Echo

- 频域系数在编码过程中的量化产生的量化误差在时域被扩展了
- 当时域上出现能量突然变化的信号时（频域系数变化也比较大），量化噪声明显变大







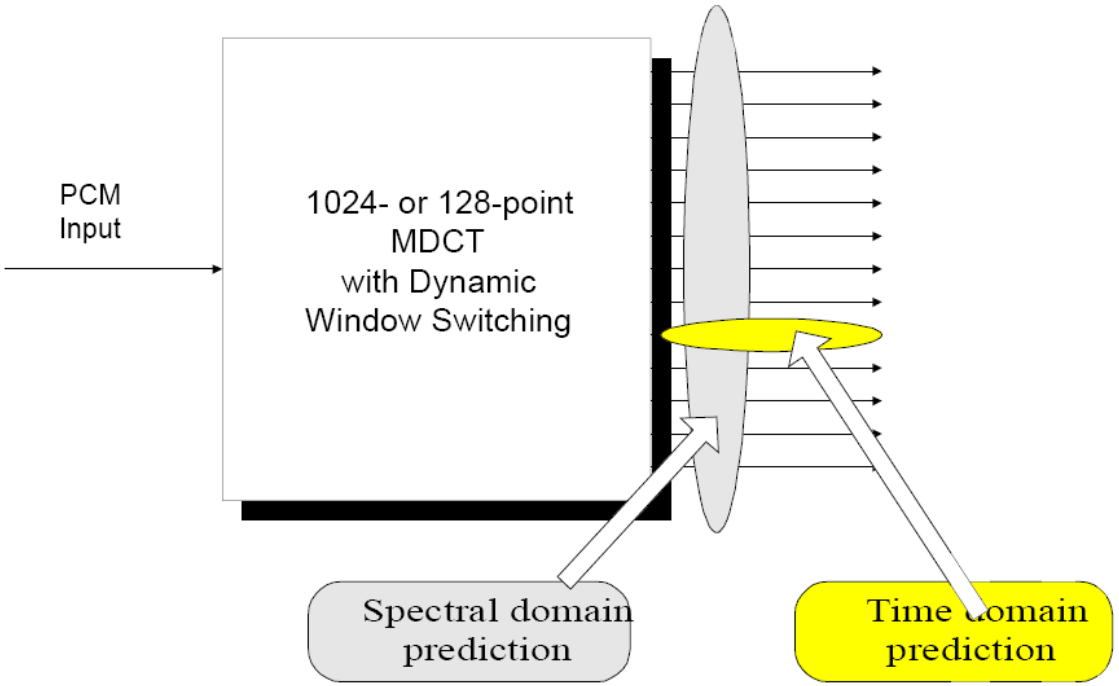
消除pre echo思路

- 频域系数采用尽量小的量化阶（需要更多的比特）
- 自适应的窗口大小，对于stationary信号采用长的窗口，transient信号采用短的窗口（增加了复杂度）
- 采用增益控制，通过控制频域系数的动态范围使量化误差减小

{ynh,cxh}@ustc.edu.cn

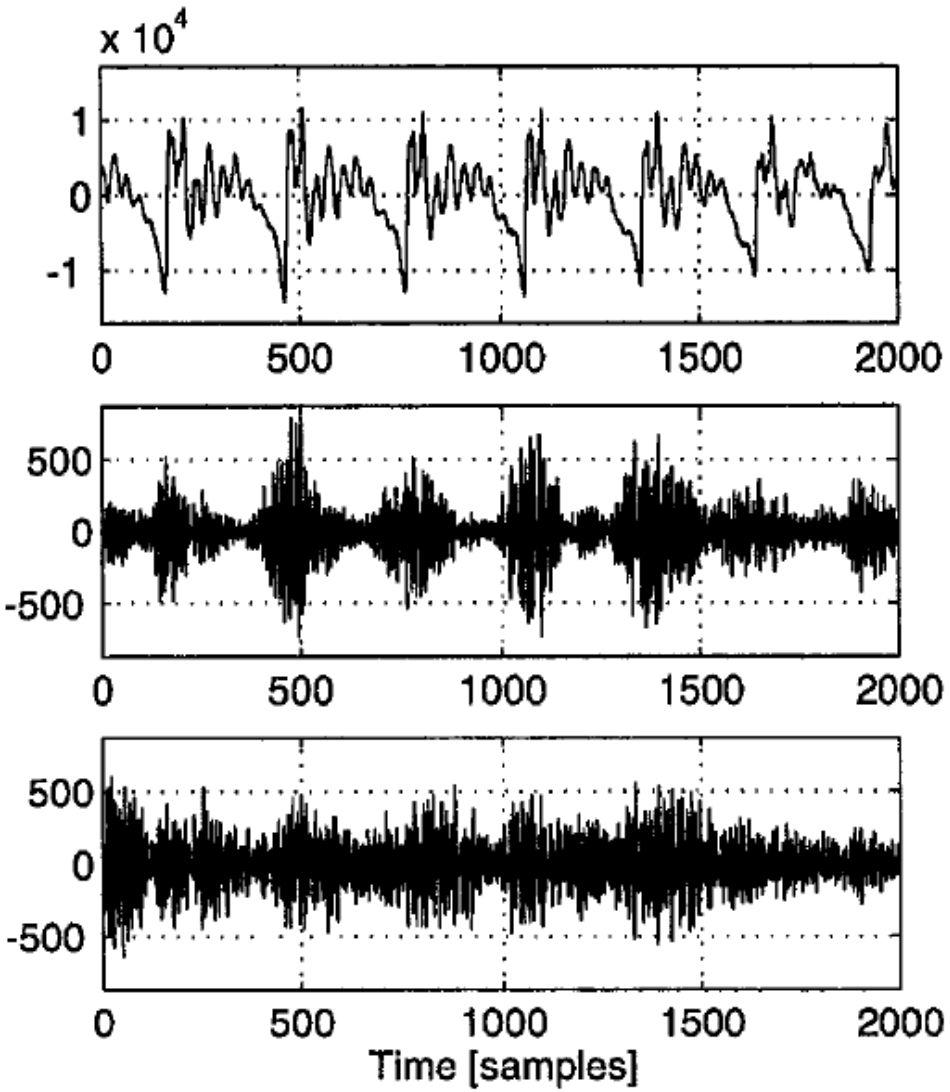
TNS(Temporal Noise Shaping)

- 正常情况下，频域上的系数通过PCM进行编码；并随时对频率系数进行预测。当预测器发现频域系数变化超过一定阈值的时候，对频域系数采用DPCM编码
- 即通过对频域系数编码的调整降低频域上量化给时域带来的噪声

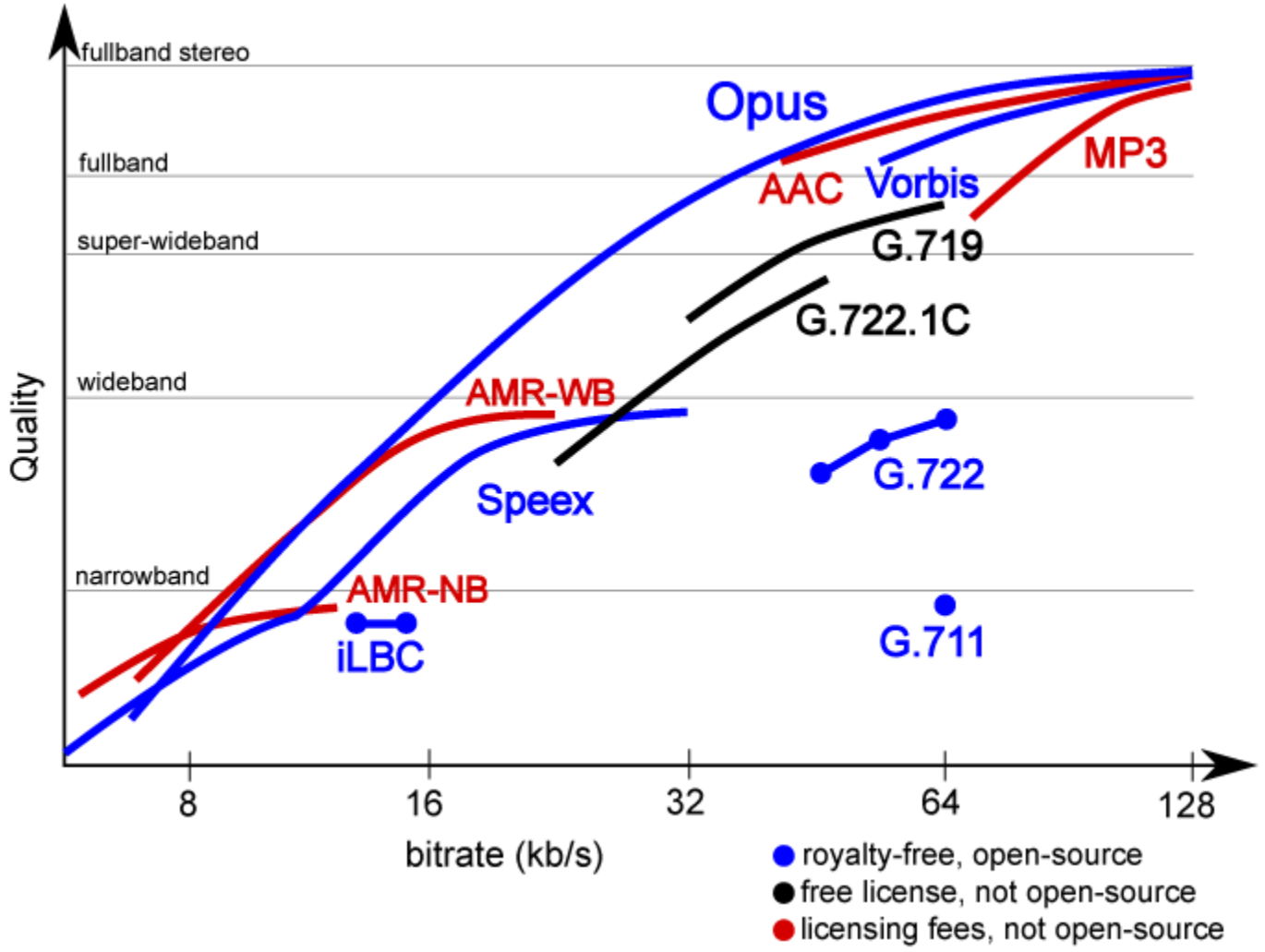


TNS效果

- 原始信号
- Coding with TNS
- Coding without TNS



AAC和MP3的性能对比



{ynh,cxh}@ustc.edu.cn

https://www.maketecheasier.com/difference-between-mp3-aac-flac-other-audio-formats/

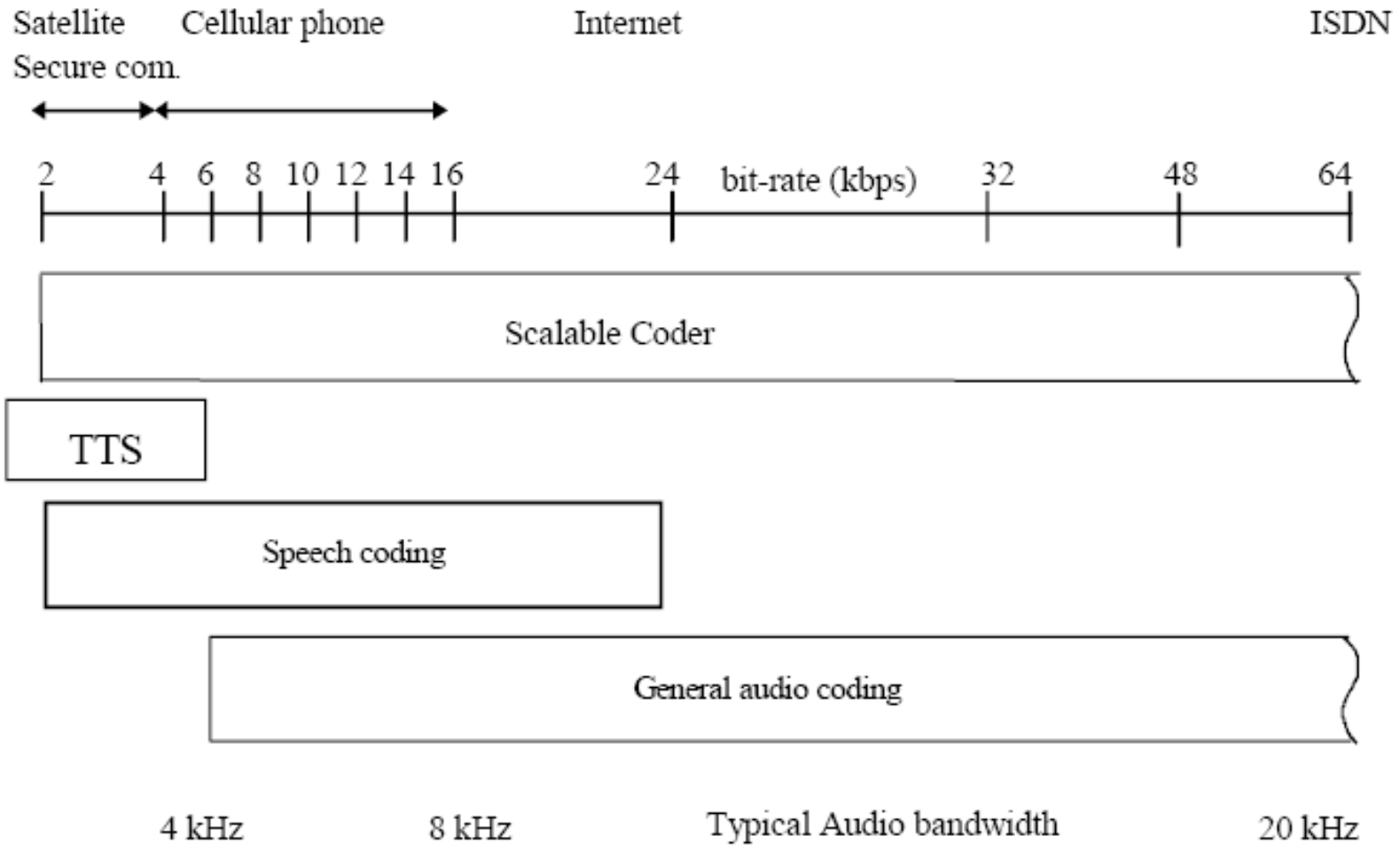
MPEG4 Audio

- MPEG-4 Audio标准可集成从话音到高质量的多通道声音，从自然声音到合成声音，编码方法包括

- 参数编码(parametric coding)
- 码激励线性预测(code excited linear predictive，CELP)编码
- 时间/频率T/F(time/frequency)编码
- 结构化声音SA(structured audio)编码
- 文本-语音TTS(text-to-speech)系统的合成声音

{ynh,cxh}@ustc.edu.cn

MPEG4的3种编码



{ynh,cxh}@ustc.edu.cn

Scalable Coding(可伸缩编码)

http://www.ntt.co.jp/tr/0403/files/ntr200403053.pdf

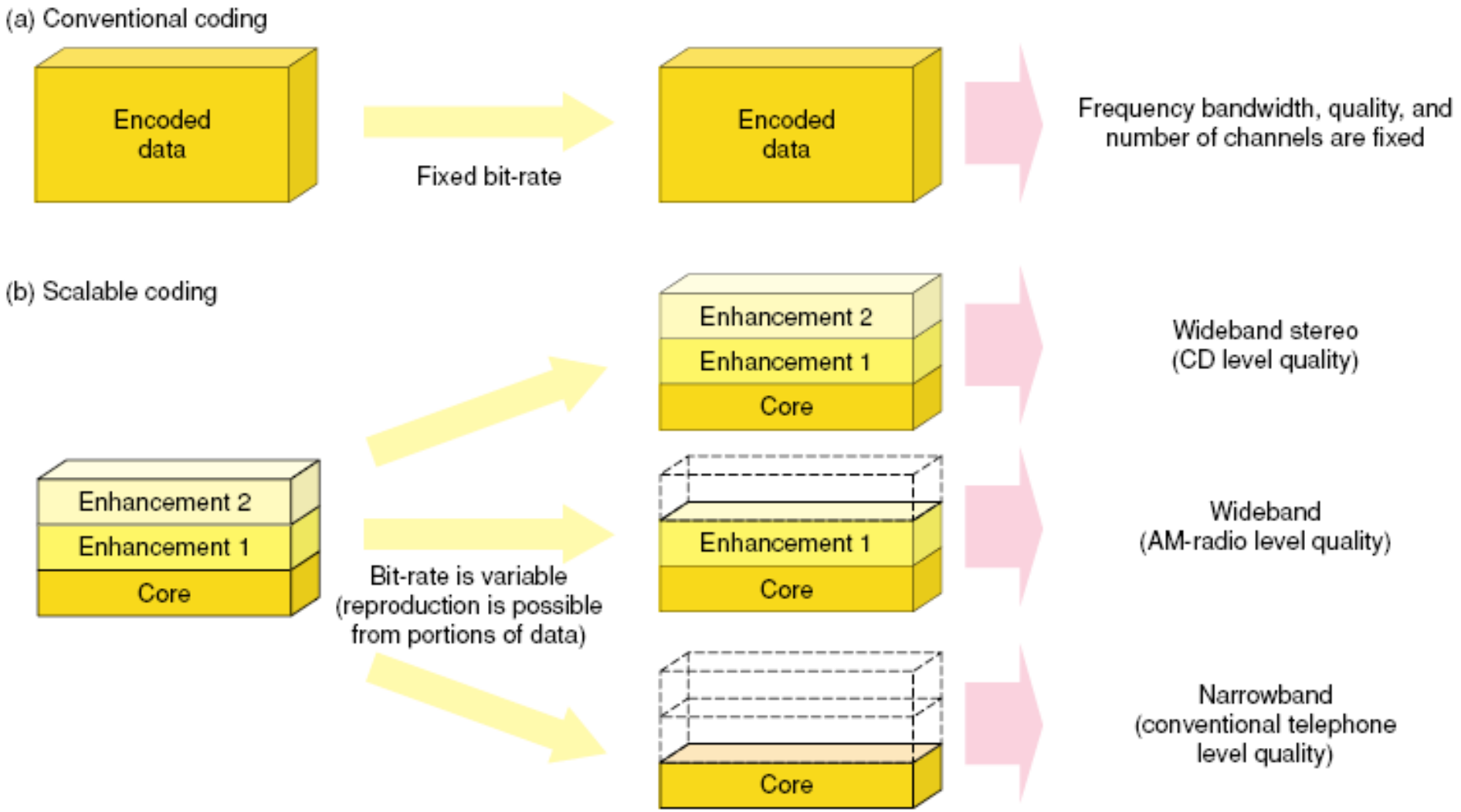
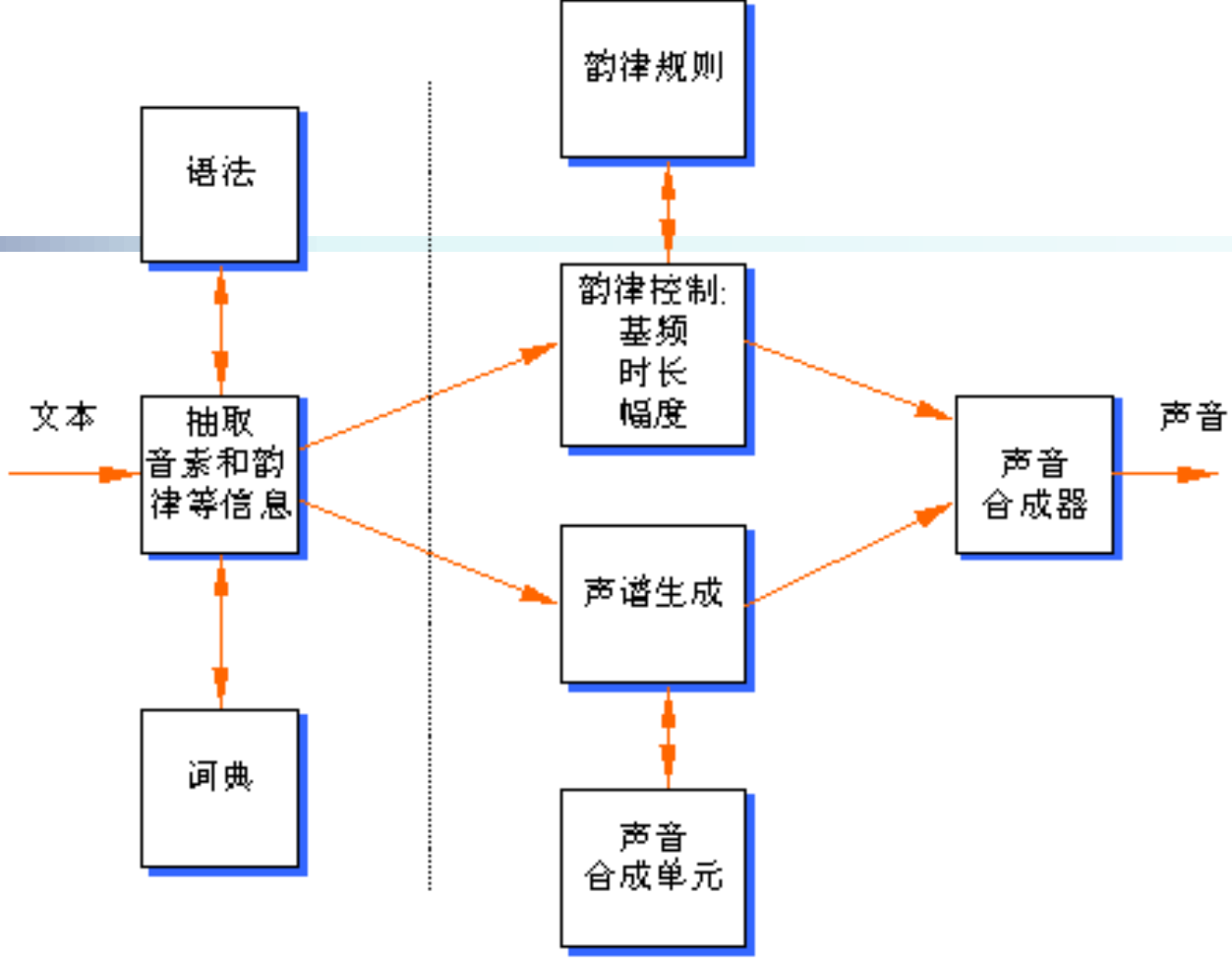


Fig. 2. Comparison of conventional coding and scalable coding.

TTS



◆图中，虚线左边的部分是文本分析部分，通过对输入文本进行词法分析、语法分析，甚至语义分析，从文本中抽取音素和韵律等发音信息。虚线右边的部分是语音合成部分，它使用从文本分析得到的发音信息去控制合成单元的谱特征(音色)和韵律特征(基频、时长和幅度)，送入声音合成器(软件或硬件)产生相应的语音输出。

{ynh,cxh}@ustc.edu.cn

音频压缩思路小结

- 基于音频数据的统计特性进行编码
  - μ 律(μ-Law)或A律(A-Law) 非均匀量化实现压缩
  - △M通过记录差值实现压缩
  - DPCM通过记录预测值与实际信号的差实现压缩
  - APCM通过调整量化阶实现压缩
  - ADPCM是DPCM和APCM思想的集合
  - SB-ADPCM通过改变不同子带样本的比特分配实现压缩（听觉特性）
- 基于音频的声学参数进行参数编码
  - LPC记录的是信道模型的参数
- 混合编码
  - MPE、RPE改变激励获取不同的效果，CELP通过建立码本进一步压缩
- 基于人的听觉特性进行编码
  - MPEG1 Layer1/2/3，基于听觉特性的变换域编码
  - MPEG2 BC & AAC ，基于听觉特性的变换域编码
  - MPEG4 Audio使用了参数编码和混合编码

{ynh,cxh}@ustc.edu.cn

音频编码算法和标准一览

	算法	名称	数据率	标准	应用	质量
波形编码	PCM	均匀量化			公用网 ISDN 配音	4.0–4.5
	μ(A)	μ(A)	64kb/S	G.711		
	APCM	自适应量化				
	DPCM	差值量化				
	ADPCM	自适应差值量化	32kb/S	G.721		
	SB-ADPCM	子带－自适应差值量化	64kb/S	G.722		
			5.3kb/S 6.3kb/S	G.723		
参数编码	LPC	线性预测编码	2.4kb/S		保密话声	2.5–3.5
混合编码	CELPC	码激励 LPC	4.8kb/S		移动通信	4.0–3.7
	VSELP	矢量和激励 LPC	8kb/S		语音邮件	
	RPE-LTP	长时预测规则码激励	13.2kb/S		ISDN	
	LD-CELP	低延时码激励 LPC	16kb/S	G.728 G.729		
	MPEG	多子带 感知编码	128kb/S		CD	5.0
	AC-3	感知编码			音响	5.0

{ynh,cxh}@ustc.edu.cn

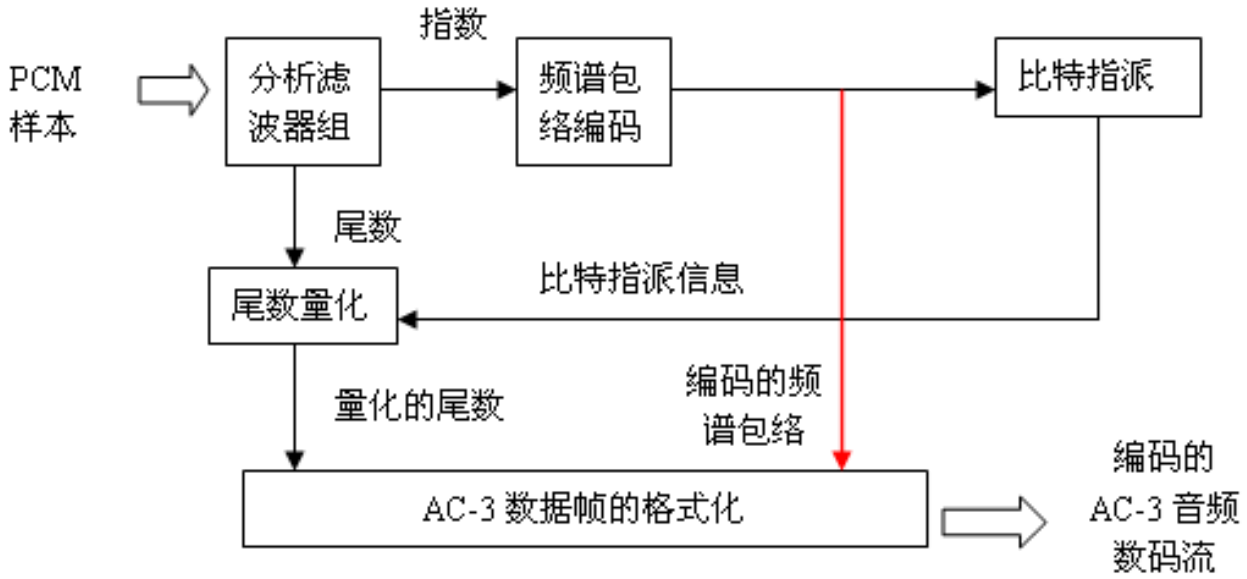
第三章 多媒体数据压缩

- §3.1 无损数据压缩
- §3.2 音频数据的压缩标准
  - §3.2.1 话音编码基础
  - §3.2.2 三种话音编码器
  - §3.2.3 移动通信网中的话音编码
  - §3.2.4 MPEG Audio
  - §3.2.5 其他音频标准
- §3.3 图像数据的压缩标准
- §3.4 视频数据的压缩标准

{ynh,cxh}@ustc.edu.cn

Dobly AC-3

◆512个时间样本的相互重叠样本块被乘以时间窗而变换到频域。由于样本块相互重叠，每个PCM输入样本将表达在两个相继的变换样本块中。频域表达式则可以二取一，使每个样本块包含256个频率系数。这些单独的**频率系数用二进制指数记数法表达为一个二进制指数和一个尾数**。这个指数的集合被编码为信号频谱的粗略表达式，称作频谱包络。核心的比特指派例行程序用这个频谱包络，确定每个单独尾数需要用多少比特进行编码。将频谱包络和6个音频样本块粗略量化的尾数，格式化成一个AC-3数据帧。



{ynh,cxh}@ustc.edu.cn

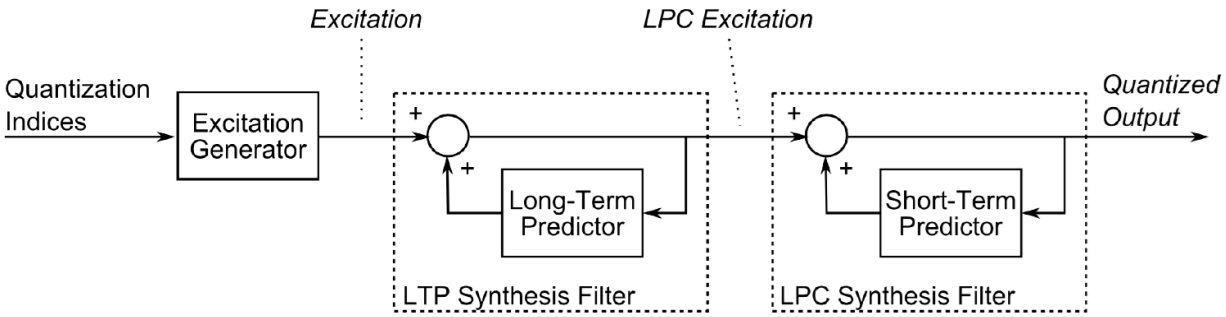
Opus音频编码

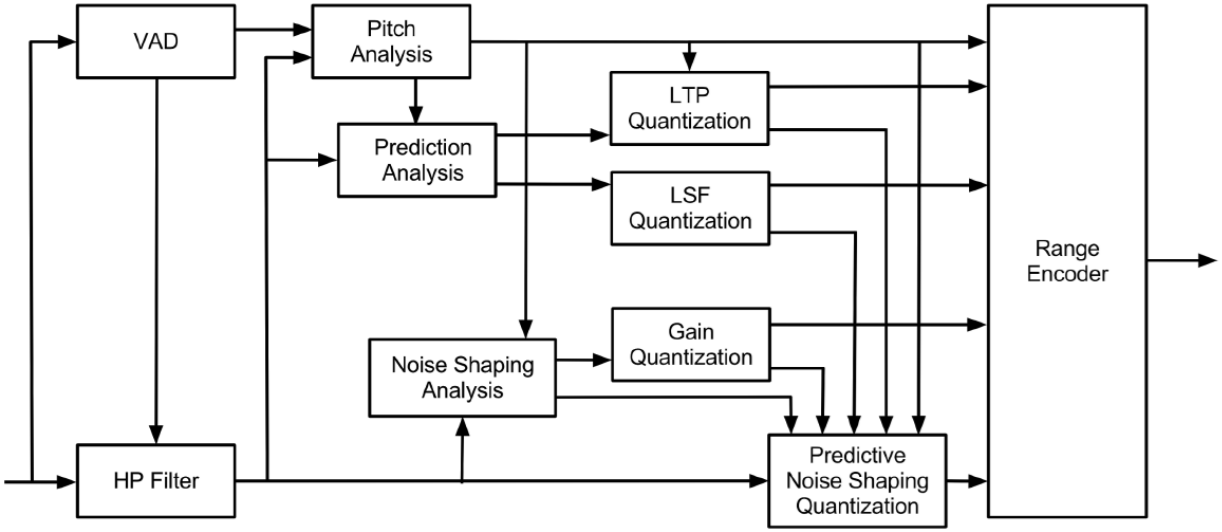
- Opus是一个有损声音编码的格式，由Xiph.Org基金会开发，之后由IETF（互联网工程任务组）标准化，单一格式包含声音和语音，适用于网络上低延迟的即时声音传输，标准格式定义于RFC 6716文件。
  - Opus格式是一个开放格式，使用上没有专利限制。
  - 支持多种帧长度，包括2.5ms、5ms、10ms、20ms、40ms、60ms，可以将多个帧组成一个120ms的包进行处理。
  - 支持从8 kHz到48 kHz的采样频率。
  - 编解码技术：Opus结合了线性预测编码（LPC）和修正散余弦变换（MDCT）技术，能够在低比特率下高效压缩语音信号，同时在高保真压缩方面表现出色。

Opus编码器

Opus合并了 Skype's SILK 和 Xiph.Org's CELT（MDCT）

SILK使用LP
对**话音**进行编码



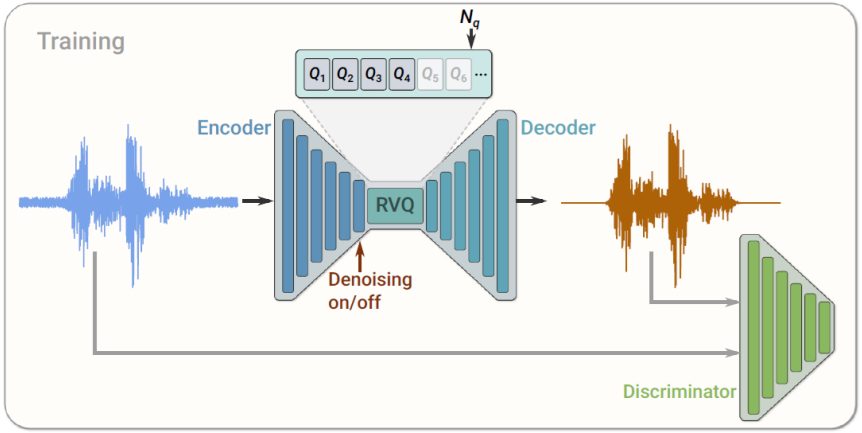


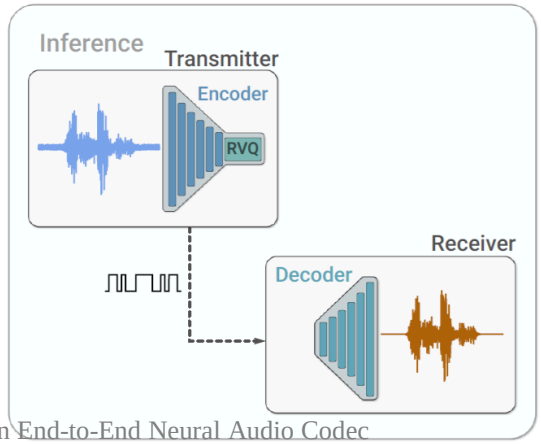
CELT使用MDCT
对**音频**进行编码

{ynh,cxh}@ustc.edu.cn

基于深度学习的音频压缩
Google: Lyra

Lyra 是谷歌2021年开源的低比特率（ 3.2kbps到9.2kbps ）音频编解码器，2022年Lyra v2效果更好。





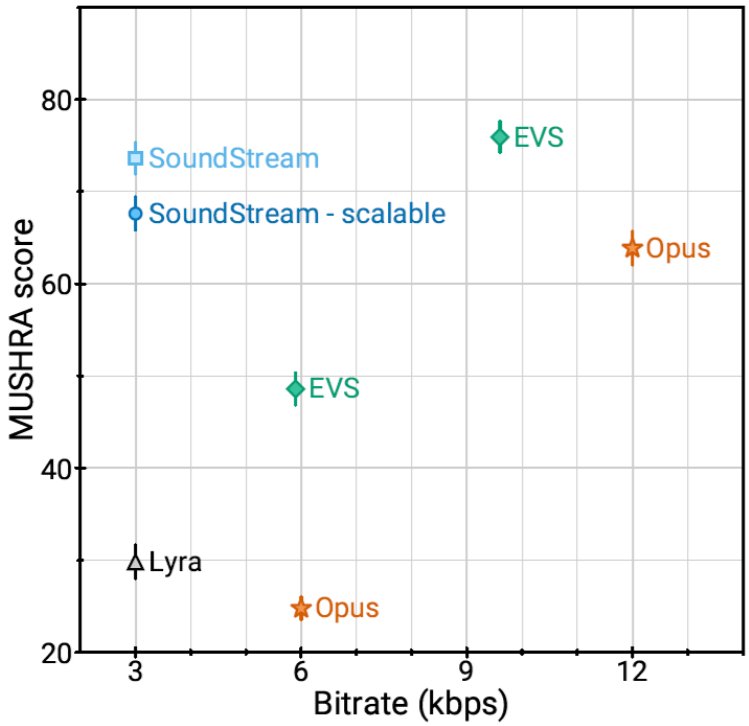


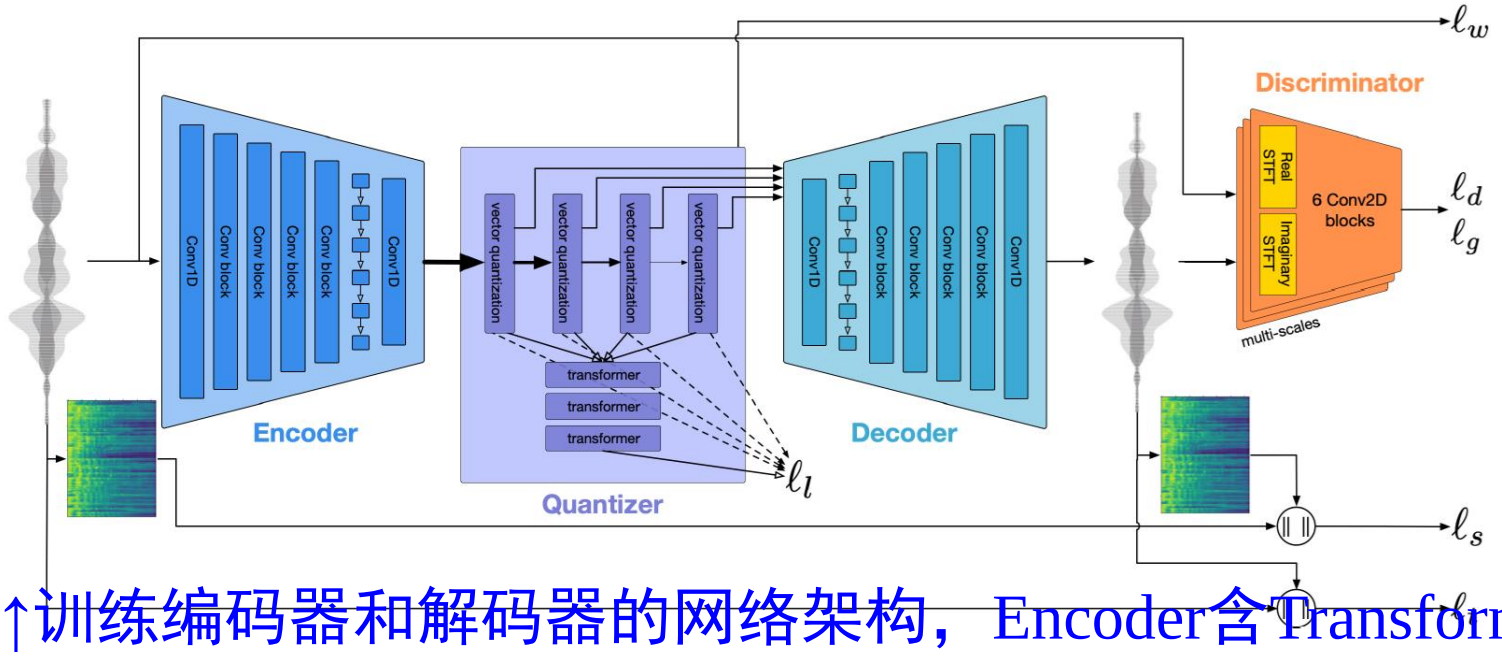
Fig. 1: *SoundStream* @3 kbps vs. state-of-the-art codecs.

Opus、EVS、Lyra、Lyra v2
主观评分对比

SoundStream: An End-to-End Neural Audio Codec
https://arxiv.org/pdf/2107.03312

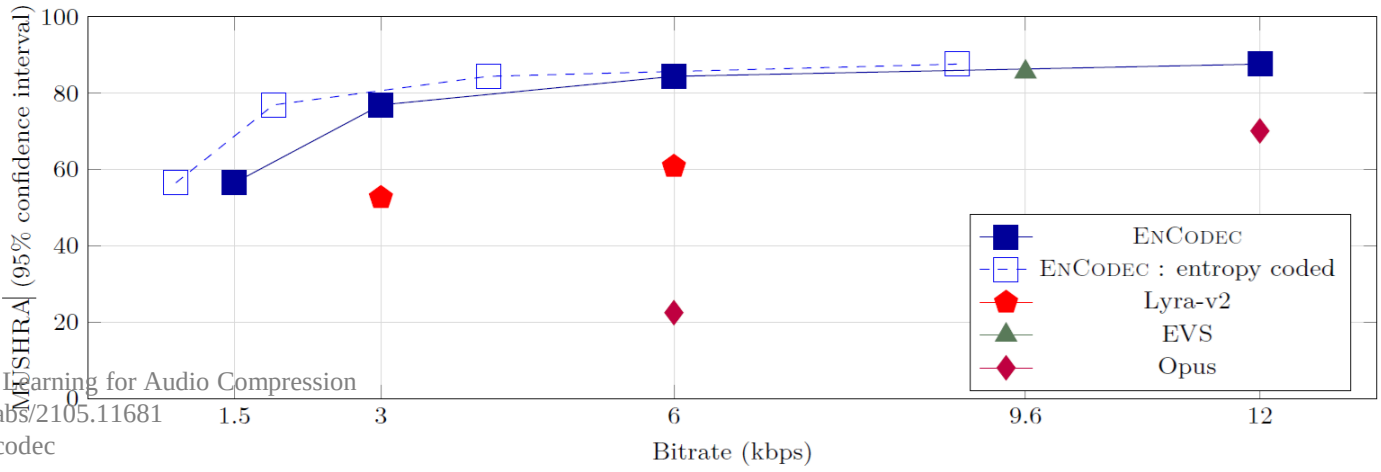
{ynh,cxh}@ustc.edu.cn

基于深度学习的音频压缩
facebook: EnCodec



↑训练编码器和解码器的网络架构，Encoder含Transformer结构

编码结果的主观评分优于Opus→



Deep Neural Networks and End-to-End Learning for Audio Compression
[v2] Tue, 13 Jul 2021. https://arxiv.org/abs/2105.11681
https://github.com/facebookresearch/encodec

历史的步伐

- 降低比特率
  - 波形编码 → 音源编码
- 降低比特率的同时提高质量
  - 混合编码（预测、自适应）
- 从语音到音频不断提高音频信号带宽
  - G.721 8kHz → G.722 16k → CD 44.1k
  - Dolby AC-3和MPEG Audio针对全频段
- 从单声道到立体声到多声道
  - G.711 → CD → Dolby AC-3和MPEG Audio
- 充分利用听觉特性
  - 均匀量化到非均匀量化(G.711)
  - 各种感知编码的产生(SBC)
- 可伸缩编码（Scalable Coding）

{ynh,cxh}@ustc.edu.cn

小结

- 冗余
  - 幅度非均匀/样本、周期、基音相关静止系数长时自相关
  - 非均匀的长时功率谱密度/语音特有的短时功率谱密度
- 编译码器
  - 波形编译码器：PCM、DPCM、ADPCM、SB-ADPCM
  - 音源编译码器：LPC
  - 混合编译码器：MPE、RPE、CELP、MPEG4 Audio
    - 移动通信中话音编码：GSM-HR/GSM-FR/GSM-EFR/AMR-NB/AMR-WB、EVS
  - 感知编码：mpeg1 Layer1/2/3、mpeg2 BC & AAC、Dolby AC-3
  - 基于深度学习的编码

{ynh,cxh}@ustc.edu.cn