音频编码原理入门

Ginohu/胡子龙

- 常识
  - 声音的产生、传播和接收
  - 波形（waveform）
  - 采样率和位深度
  - 音频分帧
- 数学基础
  - Nyquist-Shannon采样定理
  - 傅里叶变换
  - 窗函数与瞬时噪声
- 声学基础
  - 人类发声原理
  - 外周听觉系统
  - 听力频率范围
  - 心理声学模型
- 编码器原理
  - MPEG AAC 框架
  - 心理声学模型
  - 滤波器组
  - 瞬时噪声整形
  - 线性预测编码
  - 立体声编码
  - 非均匀量化
  - 无噪声编码

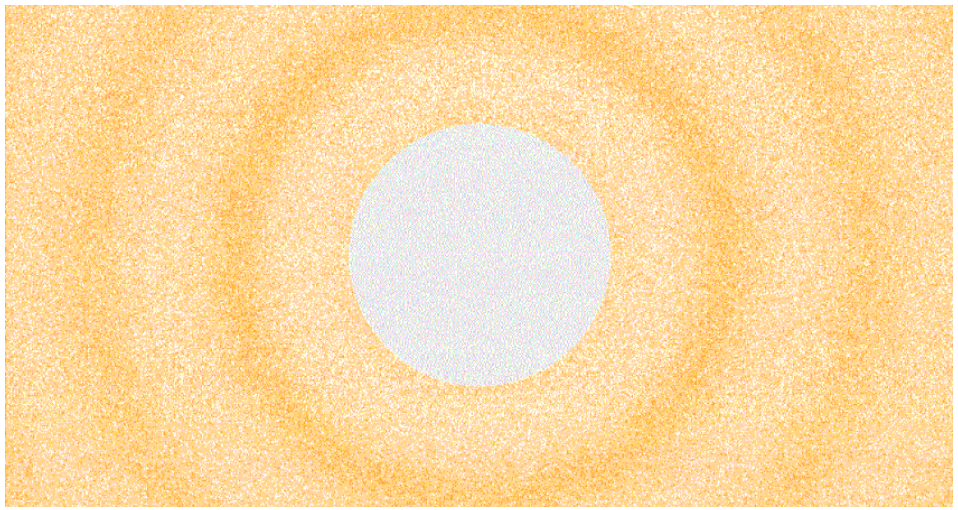
音频编码常识 – 声音的产生、传播与接收

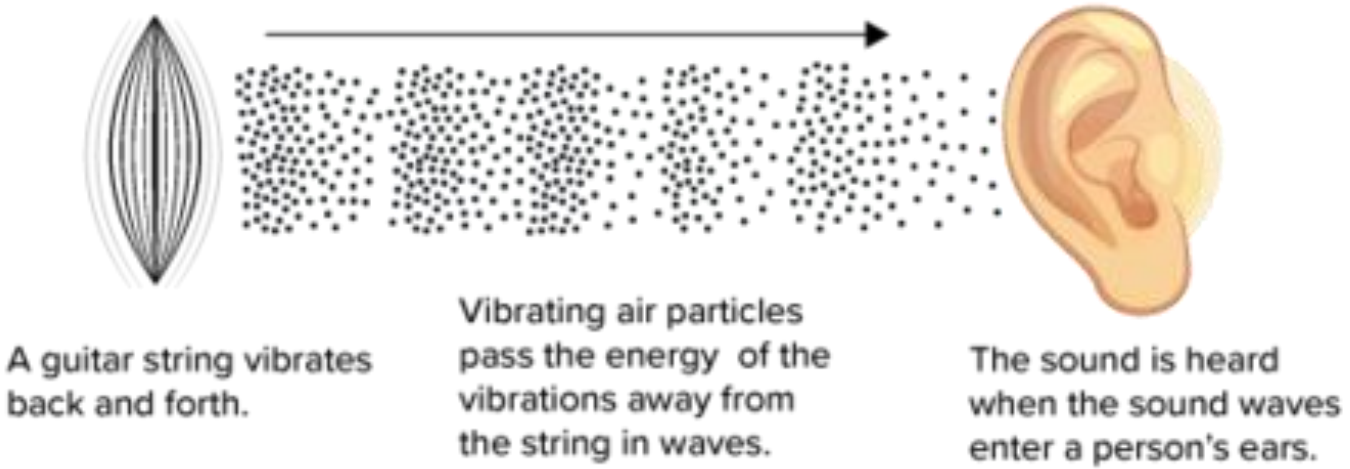
发声源

振动

传播

听觉

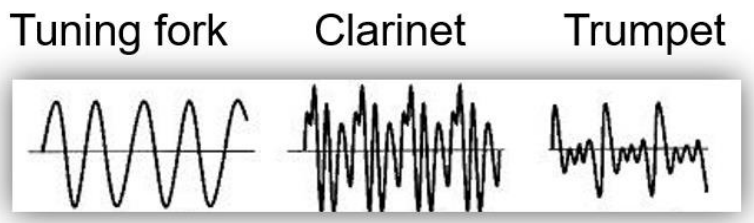




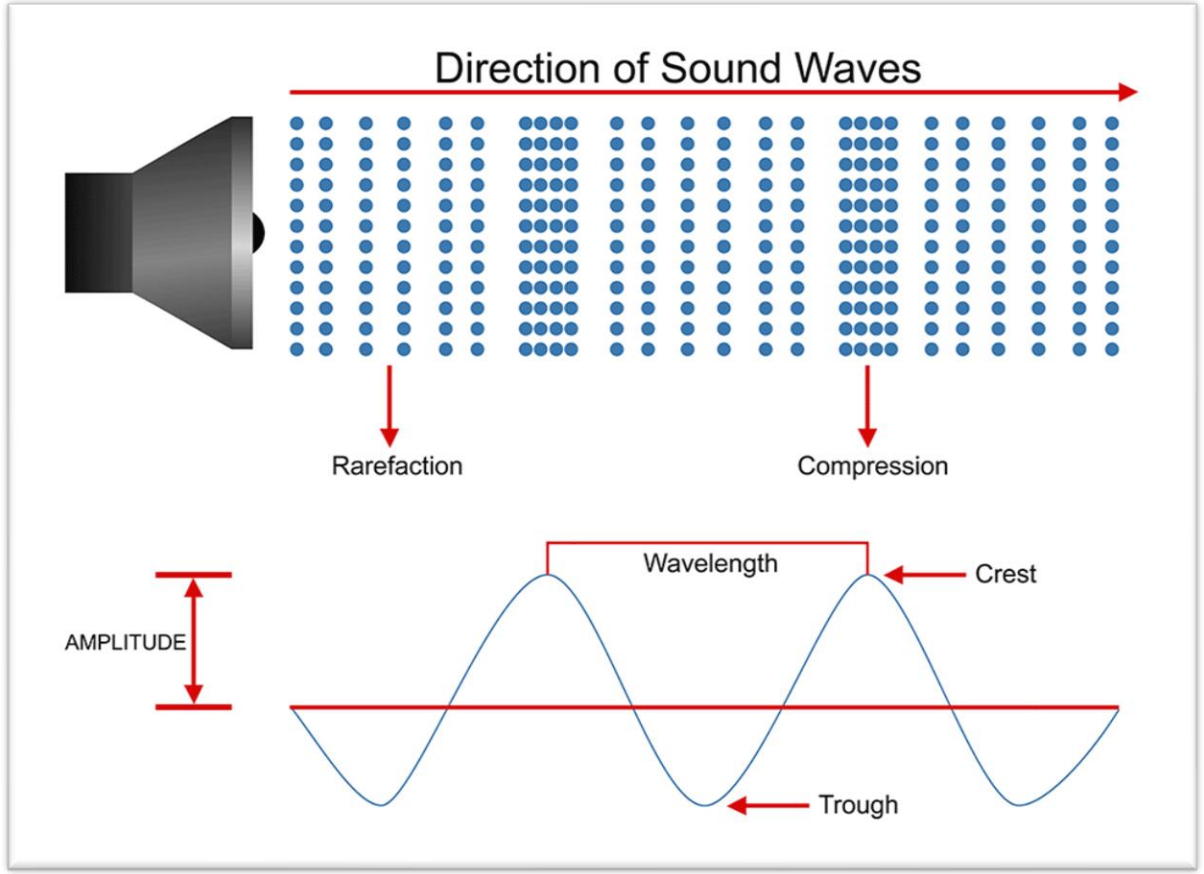
音频编码常识 – 波形（waveform）

特性

- **振幅**：表示声音的强度，又称音强
- **频率**：表示声音的音调
- **相位**：波形沿时间方向的位置，又称时域



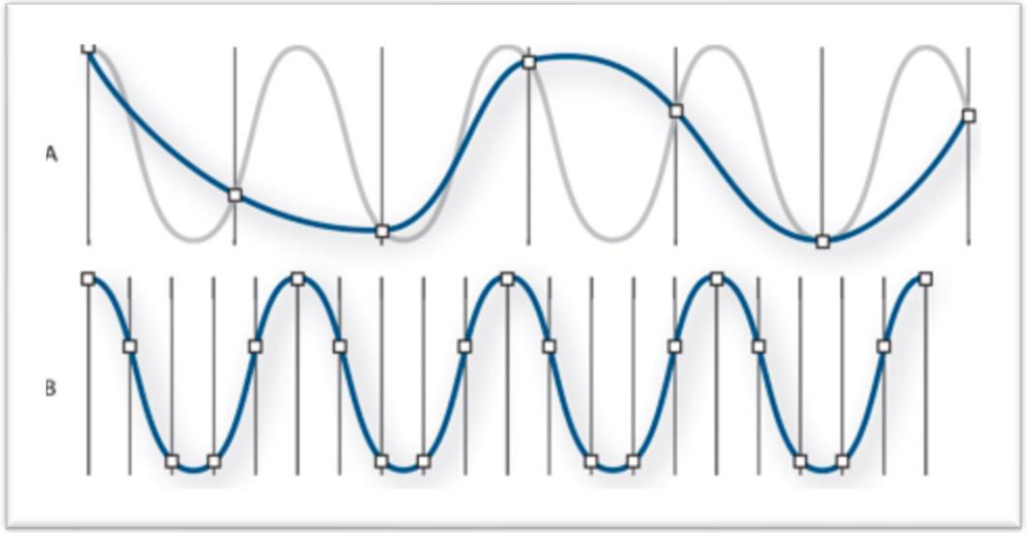
音叉、单管、小号的波形



音频编码常识 – 采样率和位深度

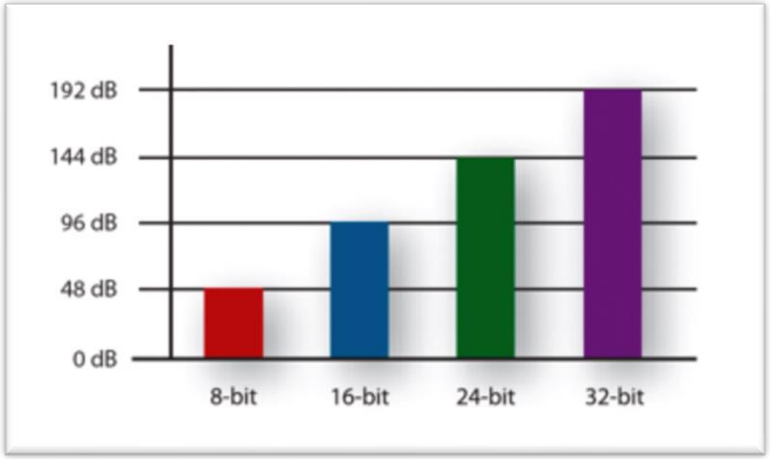
为了重现给定频率，采样率必须是该频率的两倍以上。例如，CD 的采样率为每秒 44,100 个采样，因此可重现最高为 22,050 Hz 的频率，此频率刚好超过人类的听力极限 20,000 Hz。

采样率	品质级别	频率范围
11,025 Hz	较差的 AM 电台（低端多媒体）	0–5,512 Hz
22,050 Hz	接近 FM 电台（高端多媒体）	0–11,025 Hz
32,000 Hz	好于 FM 电台（标准广播采样率）	0–16,000 Hz
44,100 Hz	CD	0–22,050 Hz
48,000 Hz	标准 DVD	0–24,000 Hz
96,000 Hz	蓝光 DVD	0–48,000 Hz



采样声波时，为每个采样指定最接近原始声波振幅的振幅值。较高的位深度可提供更多可能的振幅值，产生更大的动态范围、更低的噪声基准和更高的保真度。

位深度	品质级别	振幅值	动态范围
8 位	电话	256	48 dB
16 位	音频 CD	65,536	96 dB
24 位	音频 DVD	16,777,216	144 dB
32 位	最佳	4,294,967,296	192 dB

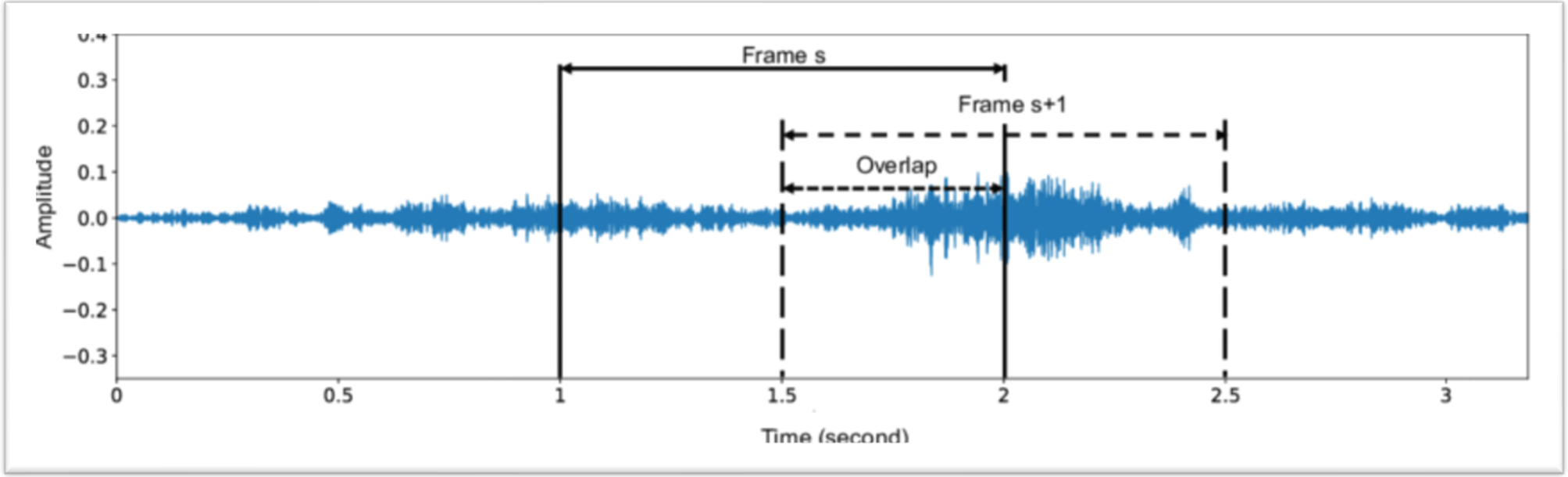


音频编码常识 – 音频分帧

音频分帧原则

- **宏观上**：分帧必须足够短来保证帧内信号是平稳的。口型变化是导致信号不平稳的原因，所以在一帧内口型不能有明显变化，即一帧的长度应当小于一个音素的长度。正常语速下，一个音素的持续时间约为50~200ms。因此，**帧长一般 < 50ms**。
- **微观上**：分帧必须包括足够多的振动周期，因为傅里叶变换是分析频率的，只有振动周期足够多才能分析。语音的基频，男声在100Hz左右，女声在200Hz左右，换算成周期就是10ms和5ms。既然一帧要包含多个周期，因此，**帧长一般 > 20ms**。

合理的帧长范围 = 20ms ~ 50ms



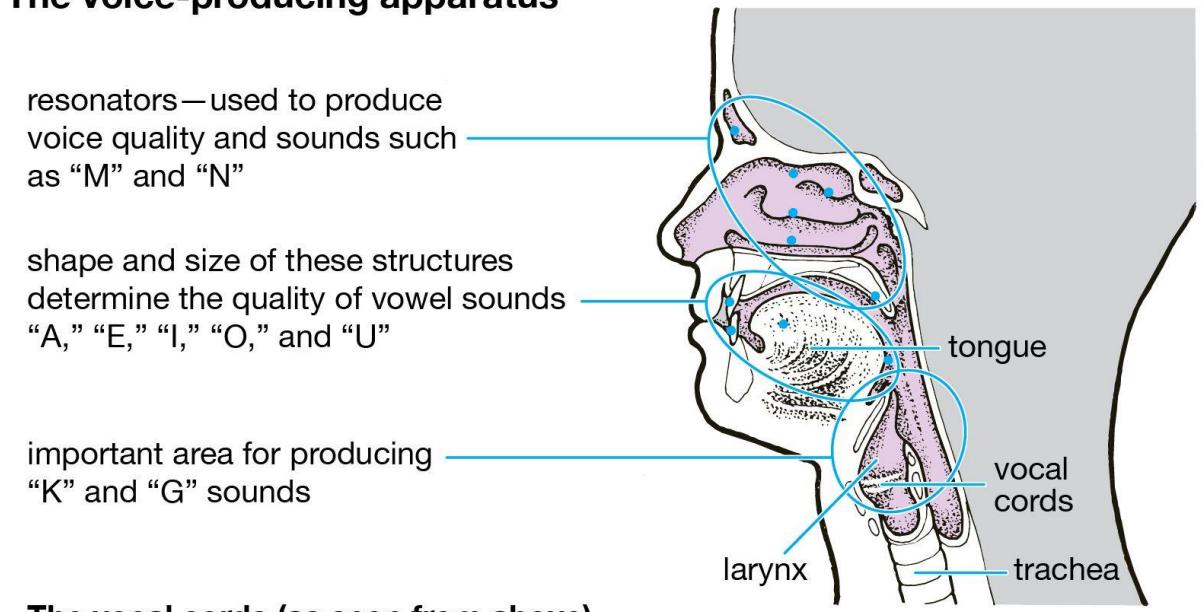
- 常识
  - 声音的产生、传播和接收
  - 波形（waveform）
  - 采样率和位深度
  - 音频分帧
- 数学基础
  - Nyquist-Shannon采样定理
  - 傅里叶变换
  - 窗函数与瞬时噪声
- 声学基础
  - 人类发声原理
  - 外周听觉系统
  - 听力频率范围
  - 心理声学模型
- 编码器原理
  - MPEG AAC 框架
  - 心理声学模型
  - 滤波器组
  - 瞬时噪声整形
  - 线性预测编码
  - 立体声编码
  - 非均匀量化
  - 无噪声编码

音频编码声学基础 – 人类发声原理

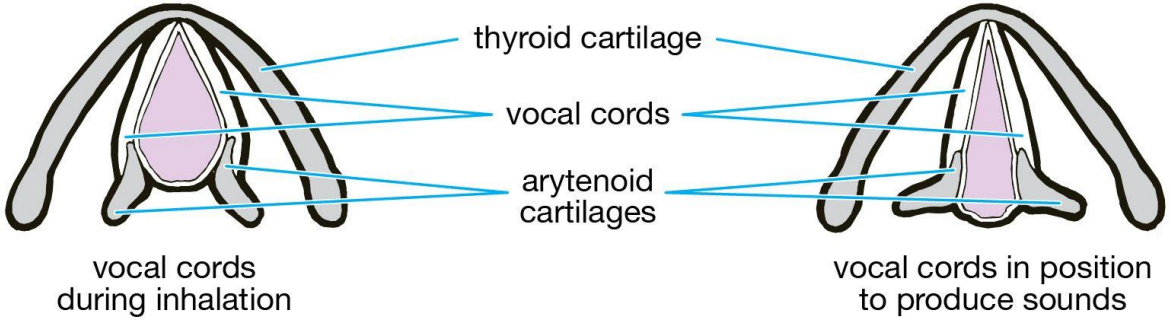
肺部呼气 -> 声带激励 -> 声道整形（口腔鼻腔共鸣）

浊音（voiced sound）：声带振动，产生周期性波形
清音（voiceless sound）：声带不振动，产生白噪声

The voice-producing apparatus



The vocal cords (as seen from above)

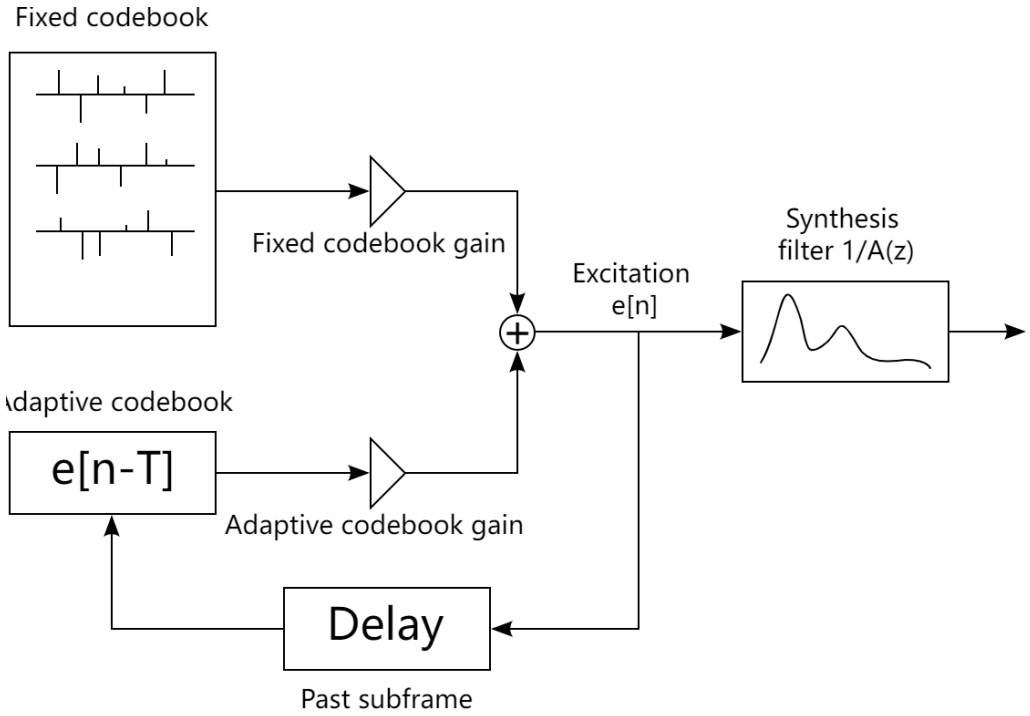


线性预测编码（LPC）

通过对发声原理进行建模，线性预测编码可以做到“**利用过去N个音素来预测当前音素**”的效果。

对于人声和大部分音调乐器，LPC能提供约70%的压缩率

数学建模

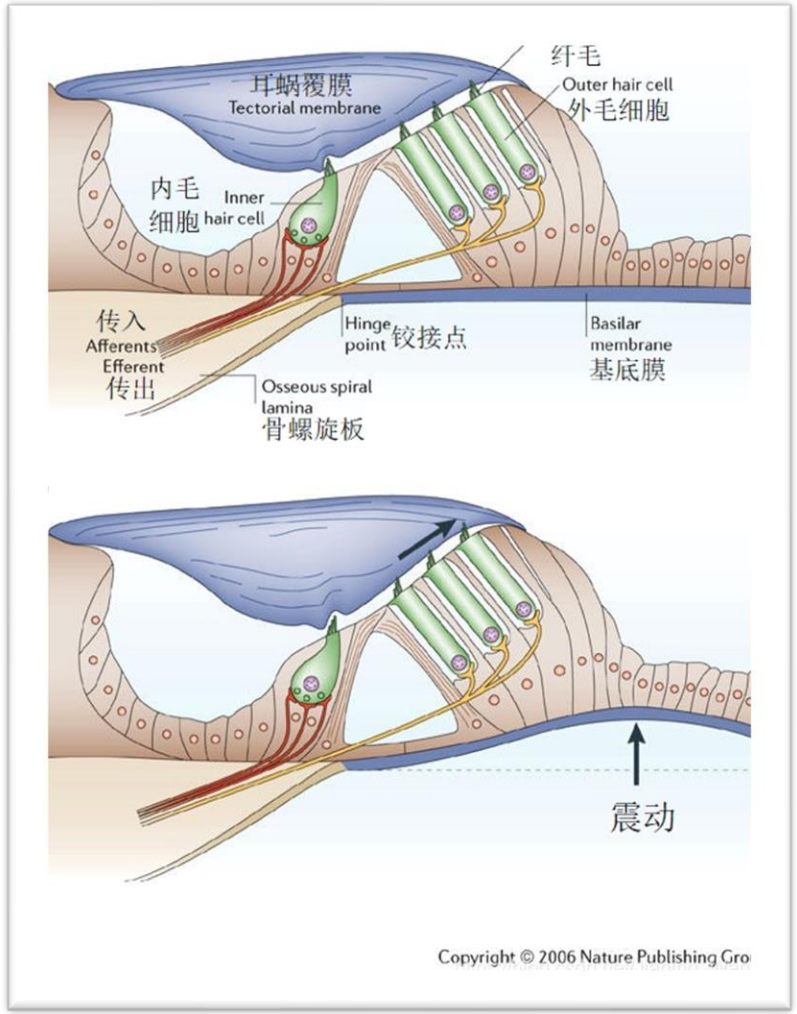


Linear Predictive Coding

音频编码声学基础 – 外周听觉系统

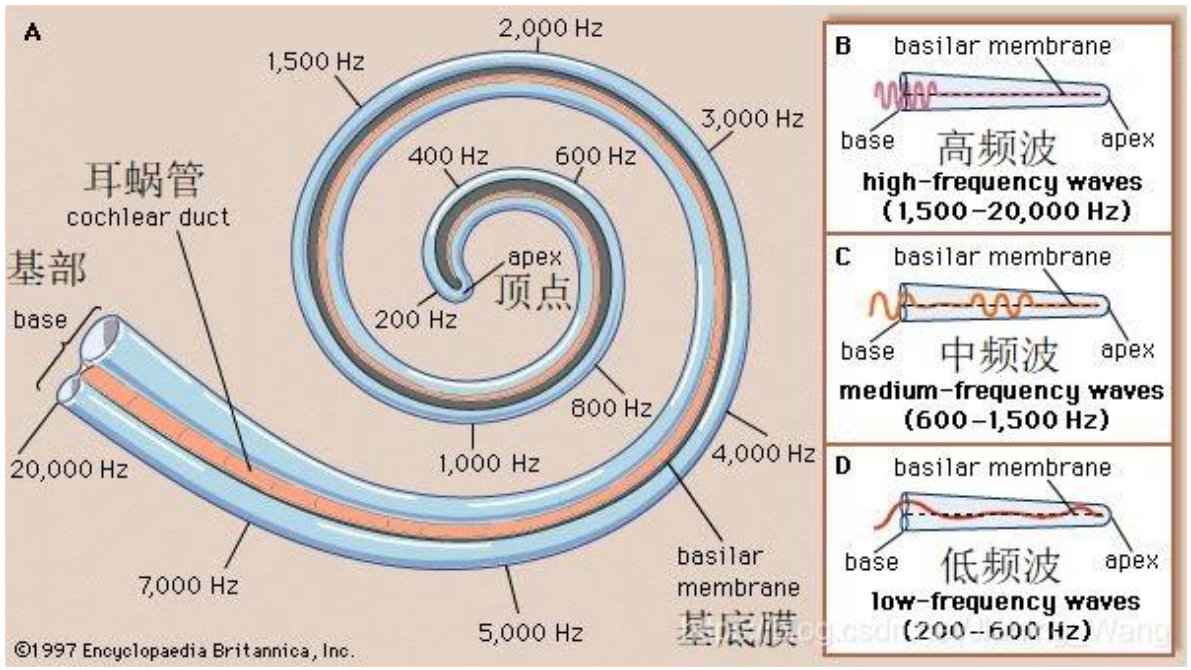
基底膜运动机理

耳蜗响应来自中耳的振动，从而给基底膜施加压力，使基底膜移动。基底膜移动导致纤毛运动，这种运动被转换为神经冲动，沿着听觉神经传递到大脑以进行处理。



基底膜感应频率的空间分布

- 基底膜在 **不同区域厚度不同**
- 基底膜在 **不同区域响应不同振动频率**
- 越靠近 **内耳对声音的敏感度越高**

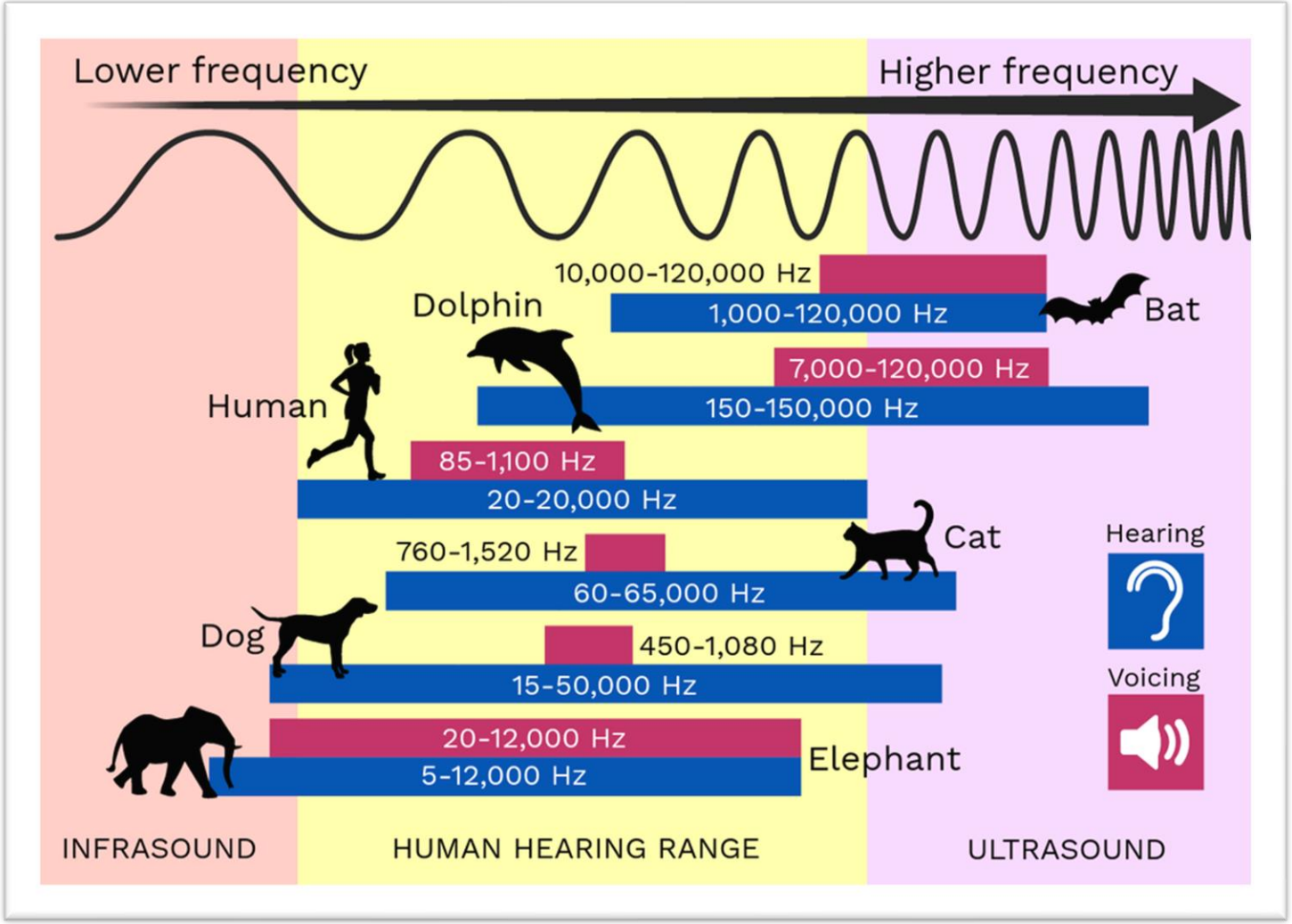


外周听觉系统

音频编码声学基础 – 听觉频率范围

- 人类听力频率极限范围：20~20000Hz
- 人类发声频率极限范围：85~1100Hz
- 人类和蝙蝠听力和发声频率范围交集很少，

双方几乎听不到对方发出的声音



音频编码声学基础 – 心理声学模型的听觉掩蔽效应

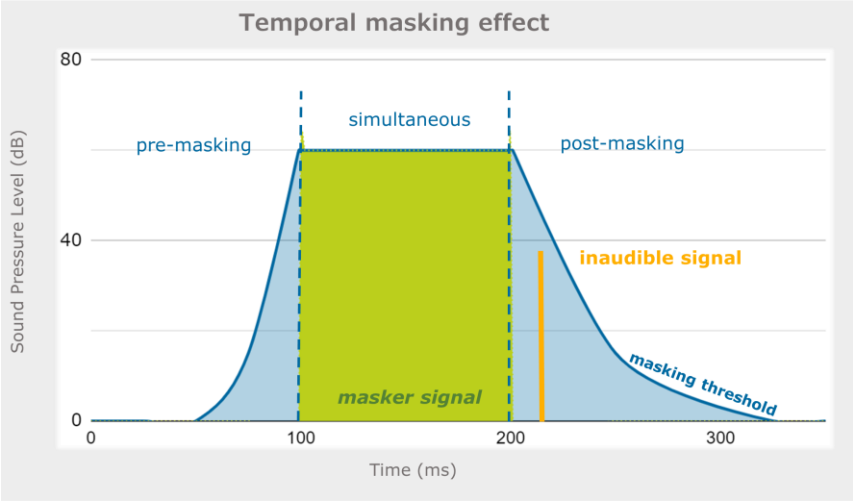
定义：当一种声音的感知受到另一种声音存在的影响时，就会出现**听觉掩蔽**。

时域掩蔽效应(Temporal masking)

Fastl和Zwicker（2007）提到，由于大脑处理信息需要花费一定的时间，所以感觉（sensation）并不是瞬间存在的，这就是存在时域掩蔽的原因。

- **超前掩蔽**（pre-masking）：在掩蔽声音发出前，会发生的掩蔽效应，在已知的研究中，超前掩蔽仅在非常短的时间内有效，即20毫秒
- **滞后掩蔽**（post-masking）：当掩蔽声音消失后，会产生滞后掩蔽效应，滞后掩蔽的强度随时间呈指数衰减，直到100~200ms后变为0。

滞后掩蔽在感知音频编码器的心理声学模型中有重要的影响和作用

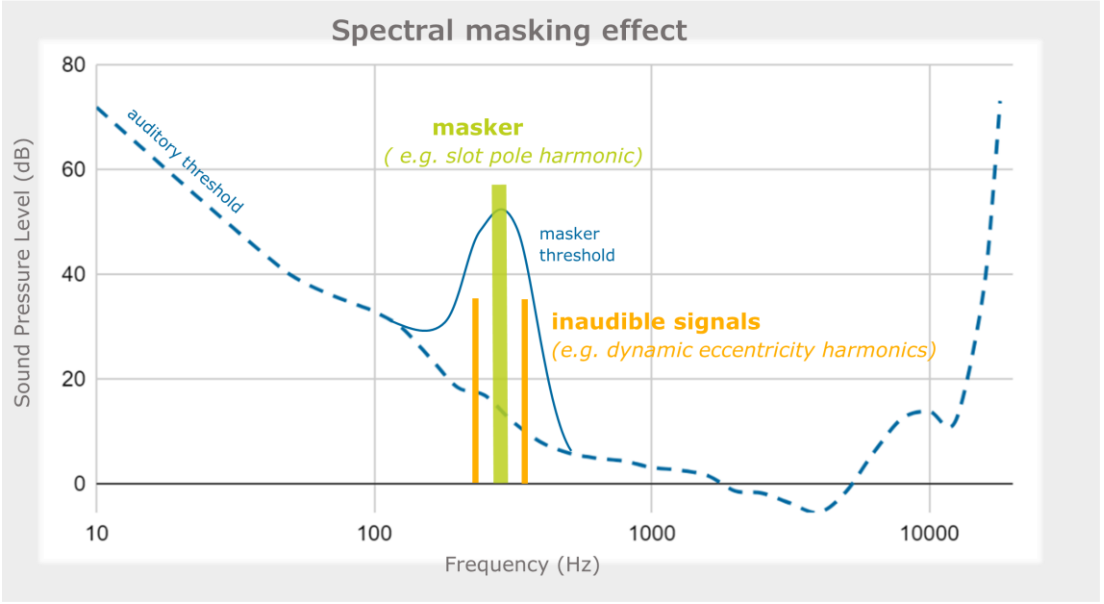


时域掩蔽效应(Temporal masking)

频域掩蔽效应(Frequency masking)

当掩蔽声音和被掩蔽声音同时存在时，会发生同时掩蔽，即频域掩蔽，一个强纯音会掩蔽在其频率附近同时发声的弱纯音。

带宽越大，掩蔽效果越强，但存在临界频带(critical band)。由于声音频率与掩蔽曲线不是线性关系，为了从感知上来统一度量声音频率，1961年德国声学家Eberhard Zwicker引入了巴克尺度（Bark scale），这个尺度的范围是从1到24，并且它们与听觉的临界频带相对应。

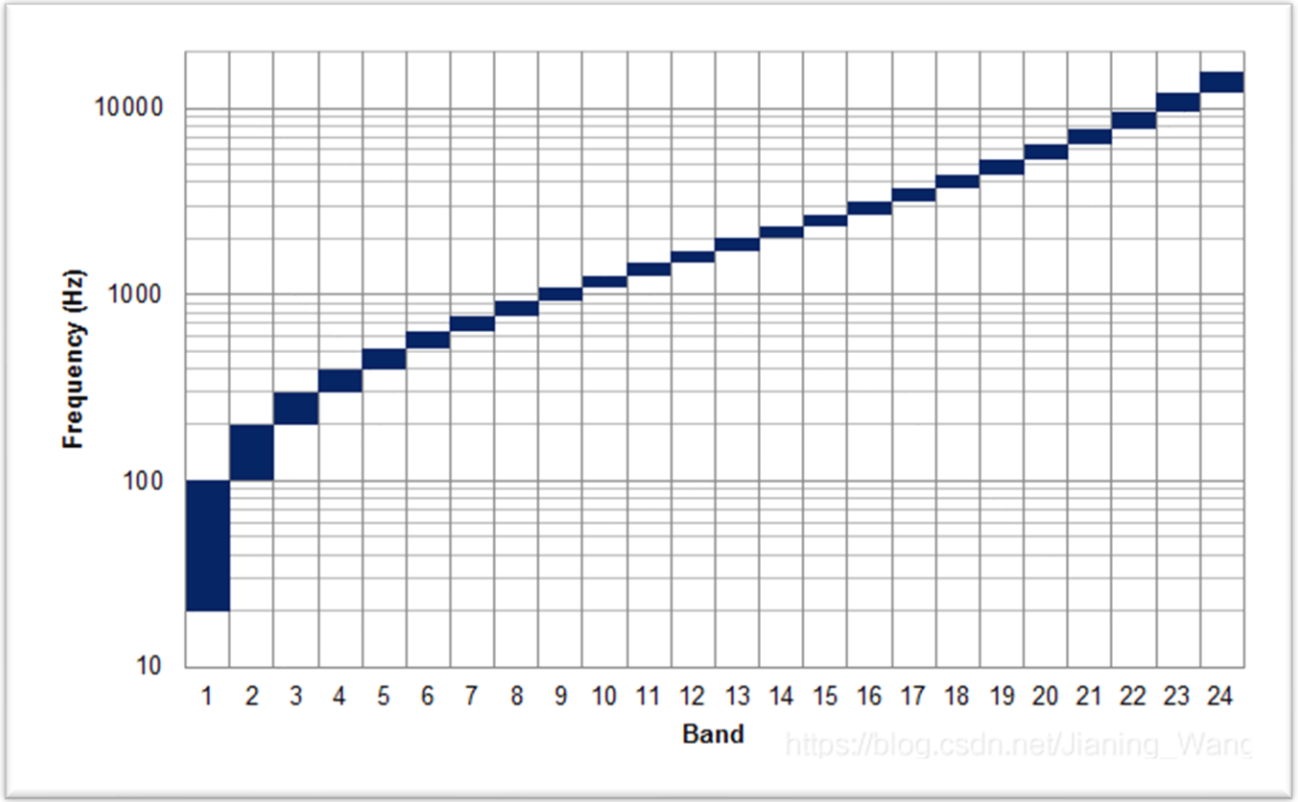


频域掩蔽效应(Temporal masking)

音频编码声学基础 – 心理声学模型的巴克尺度（Bark scale）

定义

- 巴克尺度是于1961年由德国声学家Eberhard Zwicker提出的一种心理声学的尺度。这个尺度下，实际的相同距离与感知上的相同距离一致。
- 由于人类沟通的原理，人耳对2500~3000Hz频率的声音更为敏感，所以巴克尺度在这个频率范围将频率分得更精细。



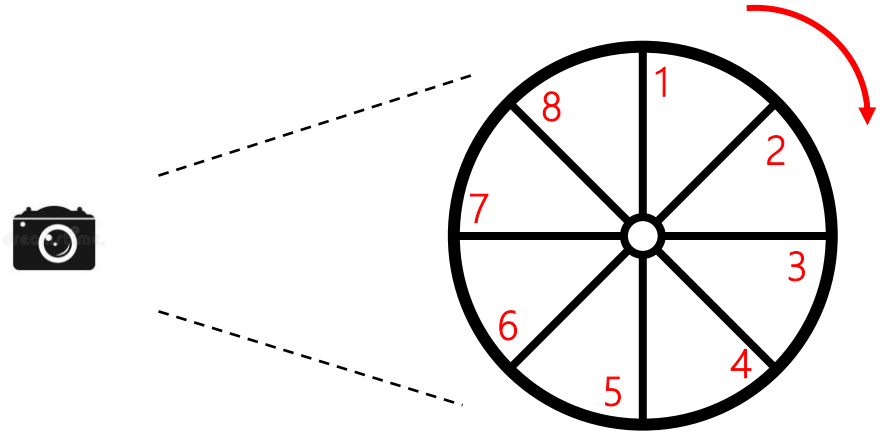
Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	450	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

- 常识
  - 声音的产生、传播和接收
  - 波形（waveform）
  - 采样率和位深度
  - 音频分帧
- **数学基础**
  - Nyquist-Shannon采样定理
  - 傅里叶变换
  - 窗函数与瞬时噪声
- 声学基础
  - 人类发声原理
  - 外周听觉系统
  - 听力频率范围
  - 心理声学模型
- 编码器原理
  - MPEG AAC 框架
  - 心理声学模型
  - 滤波器组
  - 瞬时噪声整形
  - 线性预测编码
  - 立体声编码
  - 非均匀量化
  - 无噪声编码

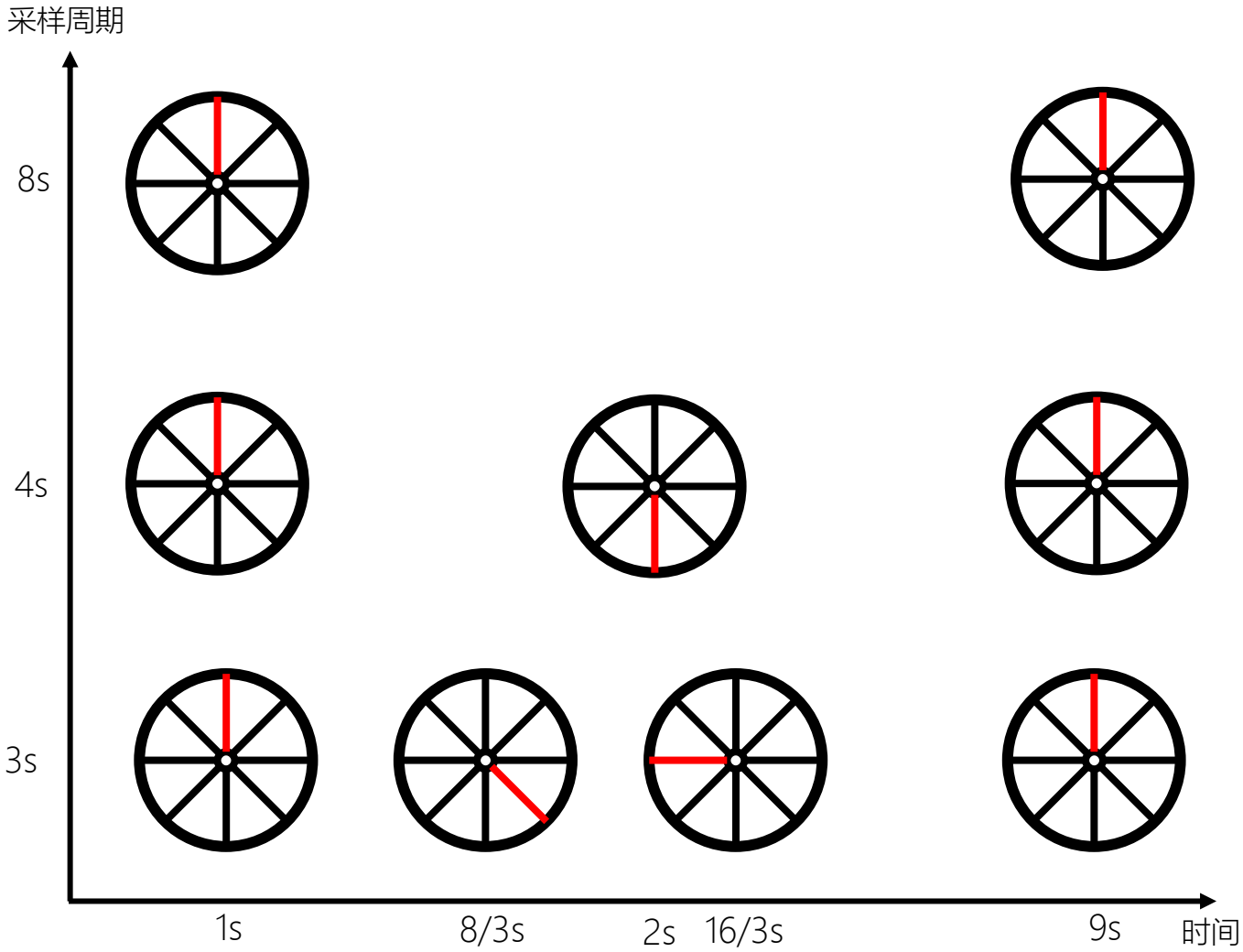
音频编码数学基础 – Nyquist-Shannon采样定理

用相机给车轮拍照，通过相片发现车轮的运动轨迹。

其中，相机相当于是采样设备，车轮是个正弦波形（相当于是音频信号）



采样周期：8s



[Nyquist–Shannon sampling theorem]

结论

- 整数倍采样周期无法分析出相位
- 相位分析：采样周期要小于整数周期的1/2
- 上面轮子转动采样其实就是正弦波，对应到音频采样，原理是相同的
- 人类听觉极限频率在20kHz左右，要正确还原出音频，需要采样周期超过20kHz的2倍，通常44100Hz是一个合理的采样频率。

音频编码数学基础 – 傅里叶变换

定义

傅里叶变换（法语：Transformation de Fourier; 英语：Fourier transform; 简称：FT）是一种线性积分变换，用于函数（应用上称作"信号"）在时域和频域之间的变换。因其基本思想首先由法国学者约瑟夫·傅里叶系统地提出，所以以其名字来命名以示纪念。

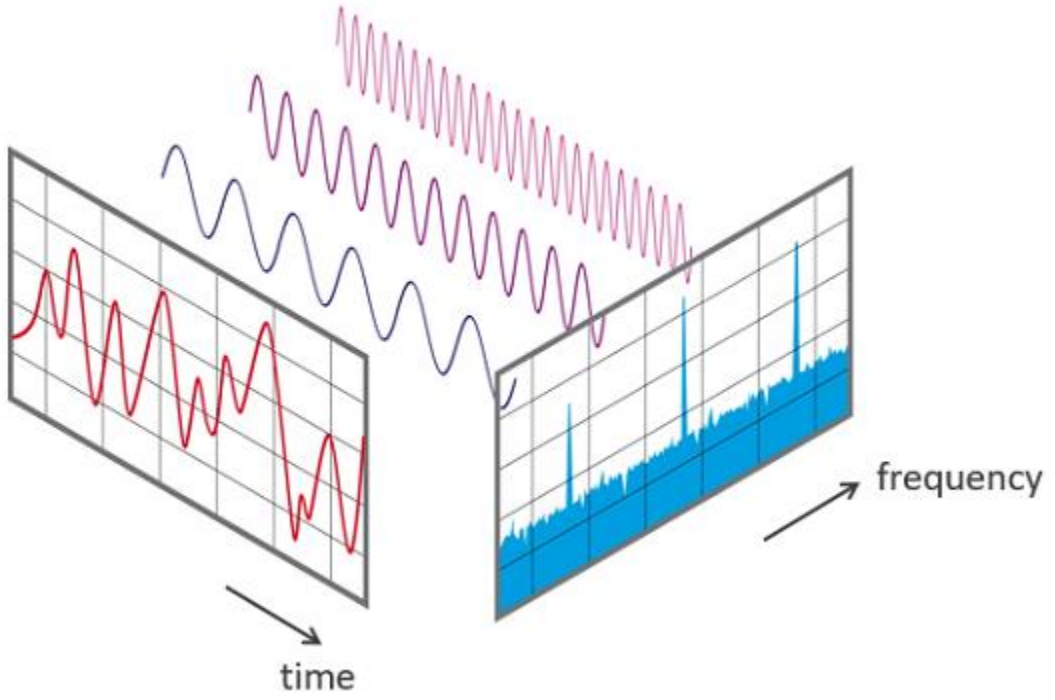
傅里叶变换 的作用是将函数分解为不同特征的正弦函数的和，如同化学分析来分析一个化合物的元素成分。对于一个函数，也可对其进行分析，来确定组成它的基本（正弦函数）成分。

（连续）傅里叶变换将可积函数 $f : \mathbb{R} \to \mathbb{C}$ 表示成复指数函数的积分形式或级数形式。

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx$$，ξ 为任意实数。ξ 的定义域为频域。

若约定自变量x表示时间（以秒为单位），变换变量ξ表示频率（以赫兹为单位）。在适当条件下，$\hat{f}$ 可由逆傅里叶变换（inverse Fourier transform）由下式得到 f。

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{2\pi i \xi x}\, d\xi$$，x 为任意实数。x 的定义域为时域。

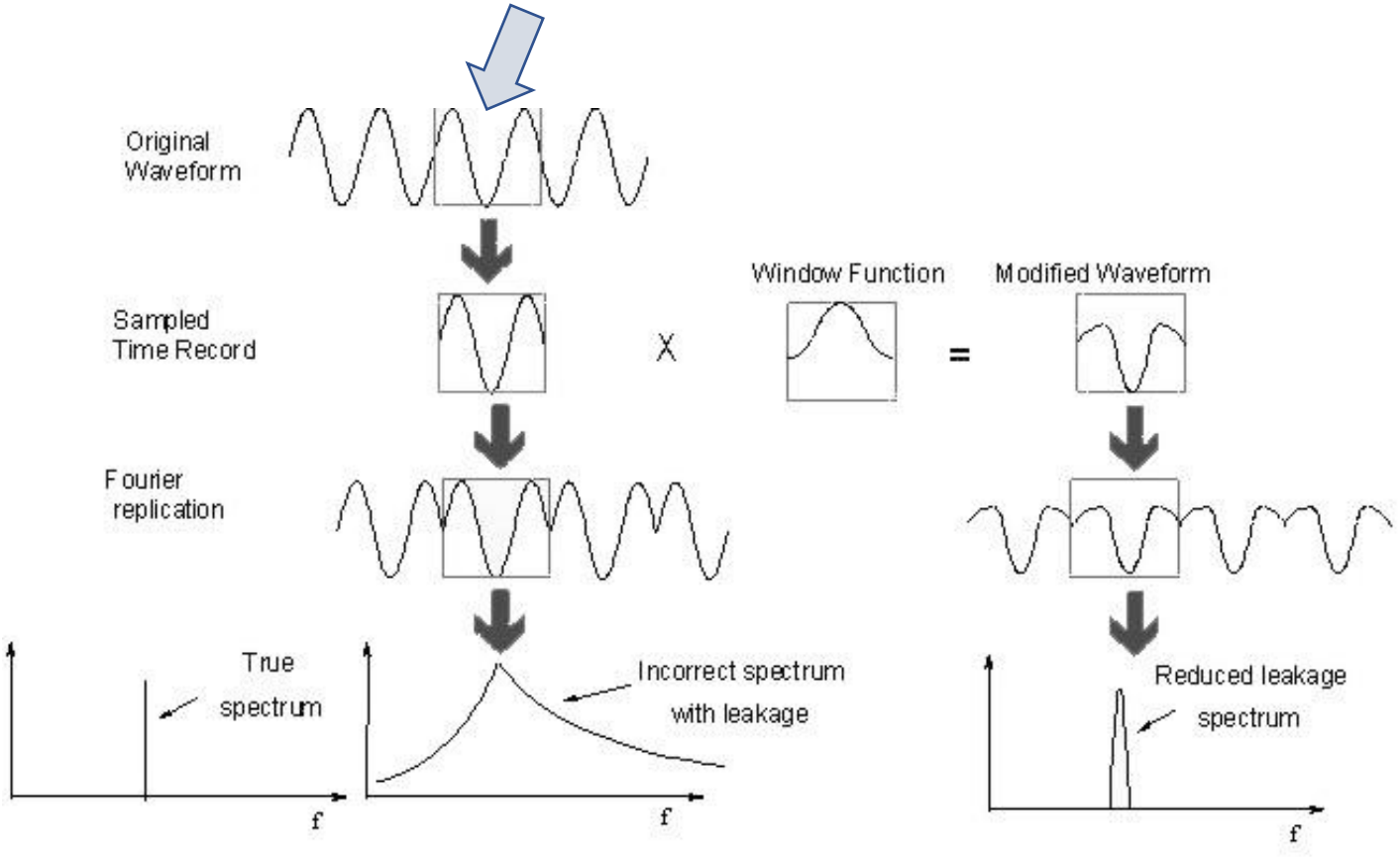


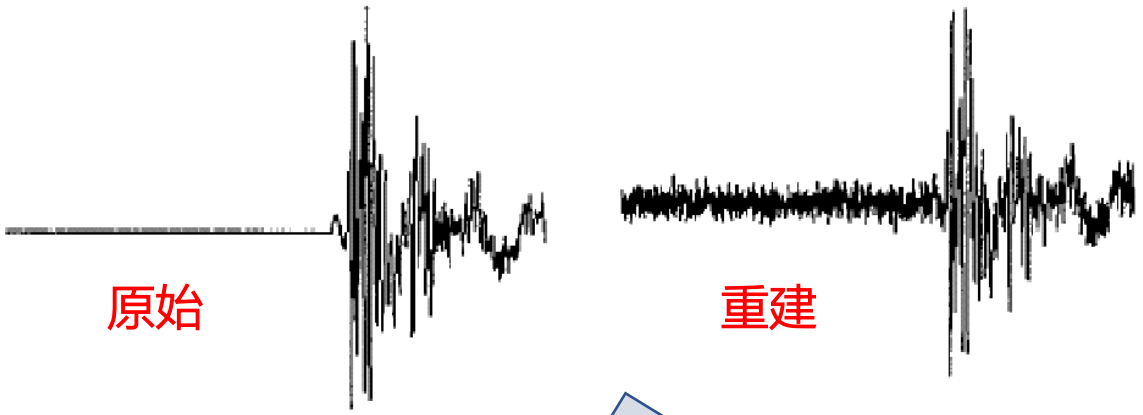
音频编码数学基础 – 窗函数与瞬时噪声

窗函数

当对时域信号进行正弦(DFT)/余弦(DCT)变换时，会用计算长度对波形画框，也就是下面的变换框。通常我们画框后，是不可能刚好取到采样的整数倍，这样会产生频域泄漏。

当对原始信号加窗函数之后，可以尽可能的减少信号在频域部分的泄漏。（**注意不是消除泄漏**）





瞬时噪声

当信号在瞬时产生短暂的声强之后，在进行变换编码再解码后，得到的信号会将这里产生的量化噪声扩散到整个时域，这个现象被称为瞬时噪声。

- 常识
  - 声音的产生、传播和接收
  - 波形（waveform）
  - 采样率和位深度
  - 音频分帧
- 数学基础
  - Nyquist-Shannon采样定理
  - 傅里叶变换
  - 窗函数与瞬时噪声
- 声学基础
  - 人类发声原理
  - 外周听觉系统
  - 听力频率范围
  - 心理声学模型
- **编码器原理**
  - MPEG AAC 框架
  - 心理声学模型
  - 滤波器组
  - 瞬时噪声整形
  - 线性预测编码
  - 立体声编码
  - 非均匀量化
  - 无噪声编码

音频编码器 – MPEG2 AAC框架

编码档位

(1) 主框架（Main）-- 本次主要讲这个档次

主框架对任何给定的码率都能提供最好音频编码质量，它不包括增益控制模块。主框架对内存和CPU处理能力要求比低复杂度LC框架要高。

(2) 低复杂度（LC）框架

在这层框架 不包括预测和预处理模块，并且TNS的阶数也受到限制。LC框架在质量很高时，对存储器和处理能力的需求都要比主框架少。

(3) 分级采样频率（SSR）框架

在这层框架中，增益控制模块是必需的。增益控制模块由一个多相正交滤波器（PQF）、几个增益检测器和几个增益调节器组成。预处理能够由控制模块完成。这层框架不需要预测模块，并且TNS的阶数和带宽都受到限制。采样率可分级框架的复杂度比主框架和低复杂度框架都低，并且它能产生一个频率可分级信号。

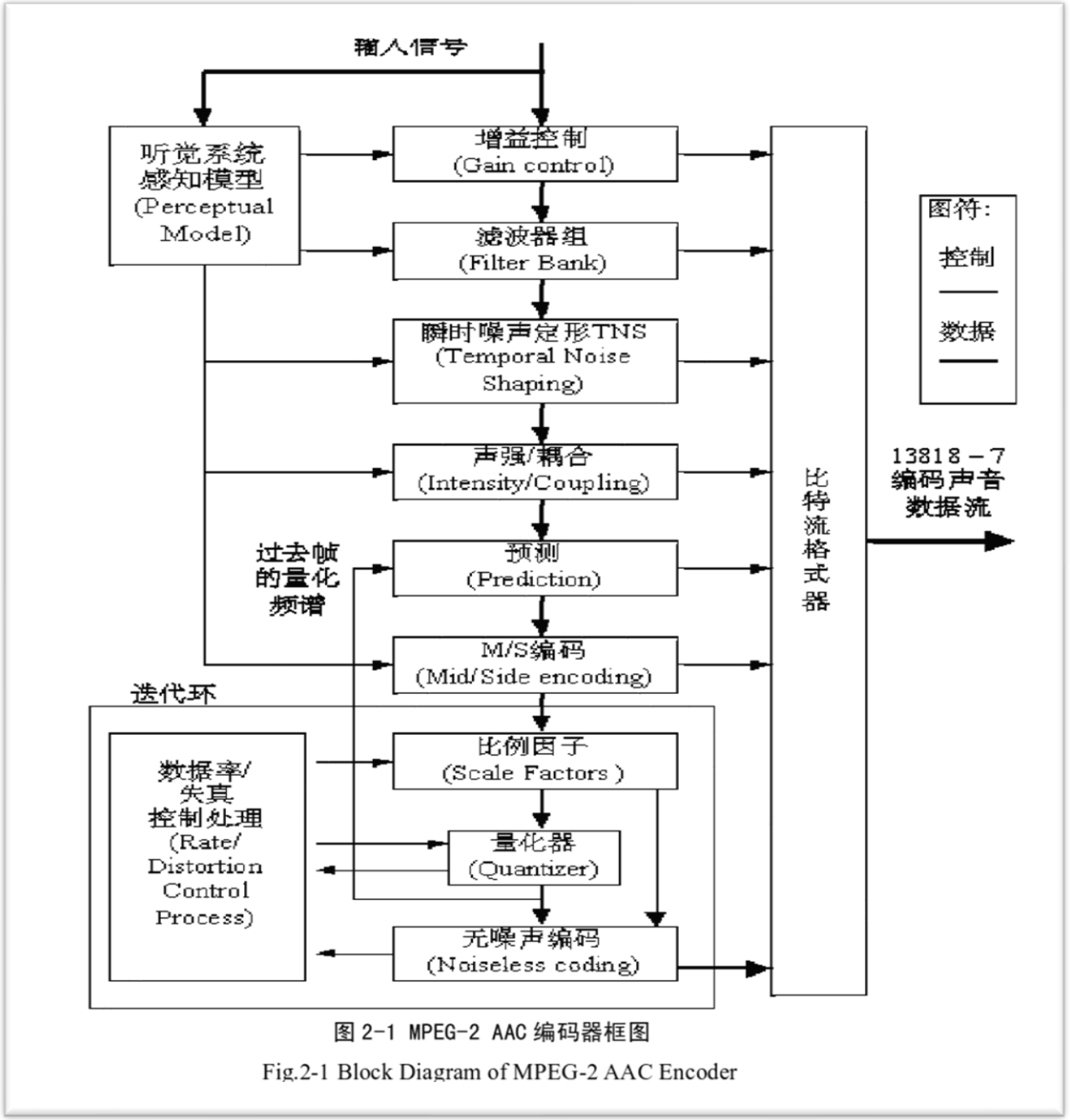


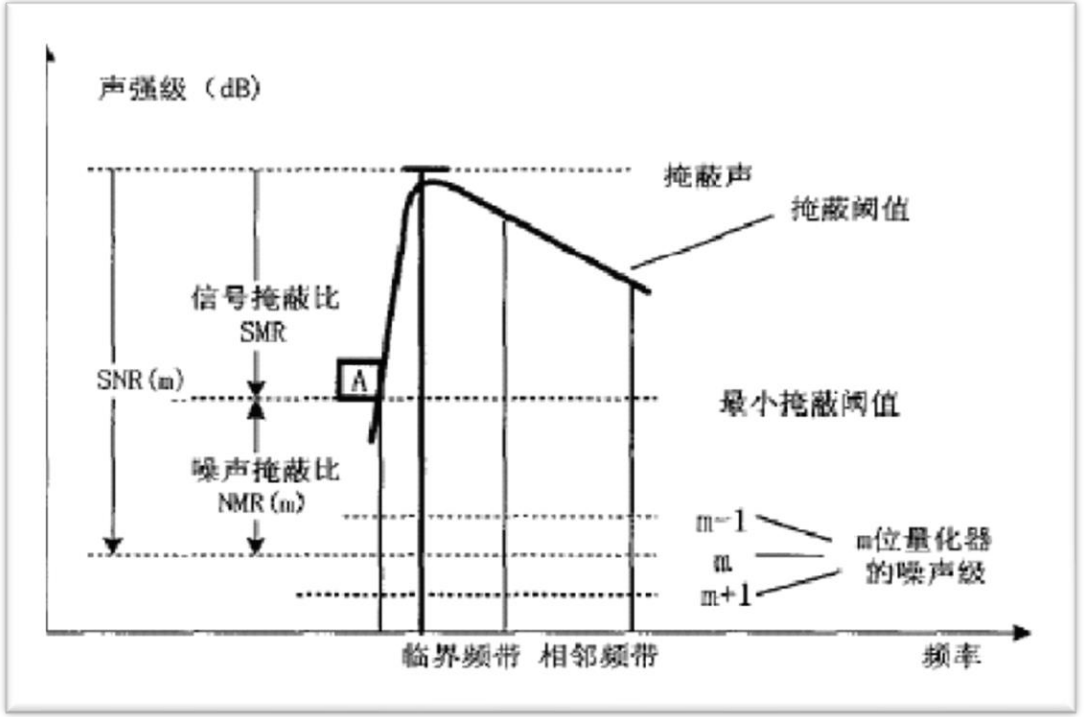
图 2-1 MPEG-2 AAC 编码器框图

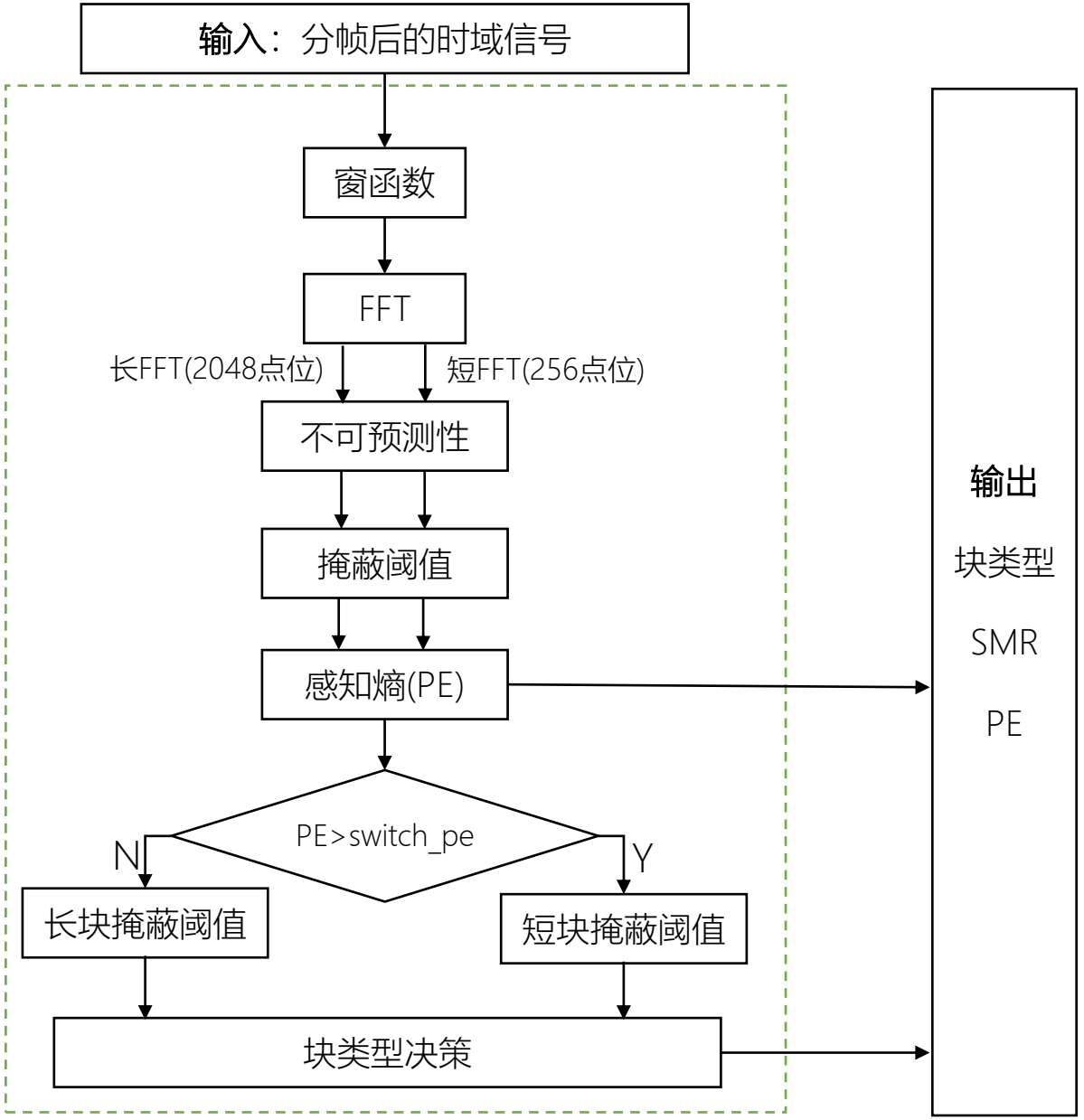
Fig.2-1 Block Diagram of MPEG-2 AAC Encoder

音频编码器 – 心理声学模型

步骤

1. 对分帧后的时域信号s(i)加 **窗函数**，得到加窗信号sw(i)
2. 对加窗信号sw(i)进行 **FFT变换**，转换到频域，得到频域信号f(w)
3. 对频域信号f(w)进行 **巴克尺度分区**
4. 用前两块静态线性预测能量和相位，得到r_pred(w)和f_pred(w)
5. 计算预测块与真实块的 **不可预测性**（欧几里得距离）
6. 用扩展函数计算预测块的 **掩蔽阈值**
7. 用预测的能量掩蔽阈值和不可预测性来计算 **感知熵PE**，得到预测块的 **信号变化剧烈程度**
8. 用PE来决定滤波器组是用2048点位还是256点位的MDCT变换





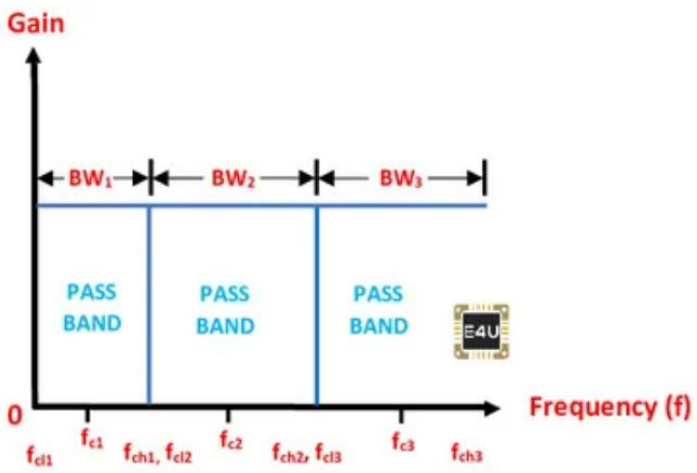
音频编码器 – 滤波器组（filter bank）

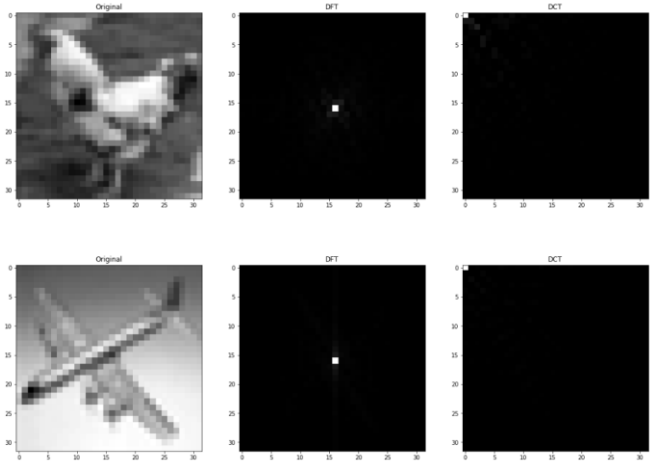
定义

在信号处理中，滤波器组是一组带通滤波器，它将输入信号分离为多个分量，每个分量携带原始信号的一个子频带。

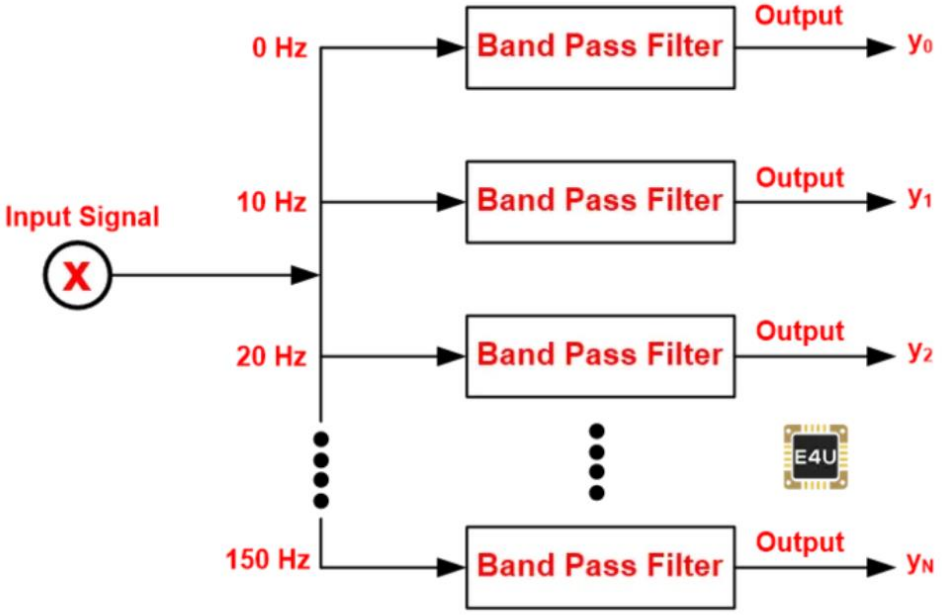
由于人类对信号的低频和高频部分感知不同，滤波器组变换函数的设置通常是以巴克尺度来决定。

MPEG-AAC的滤波器组函数是MDCT/IMDCT，同时在应用DCT变换前，需要对信号加窗函数。





FFT vs DCT

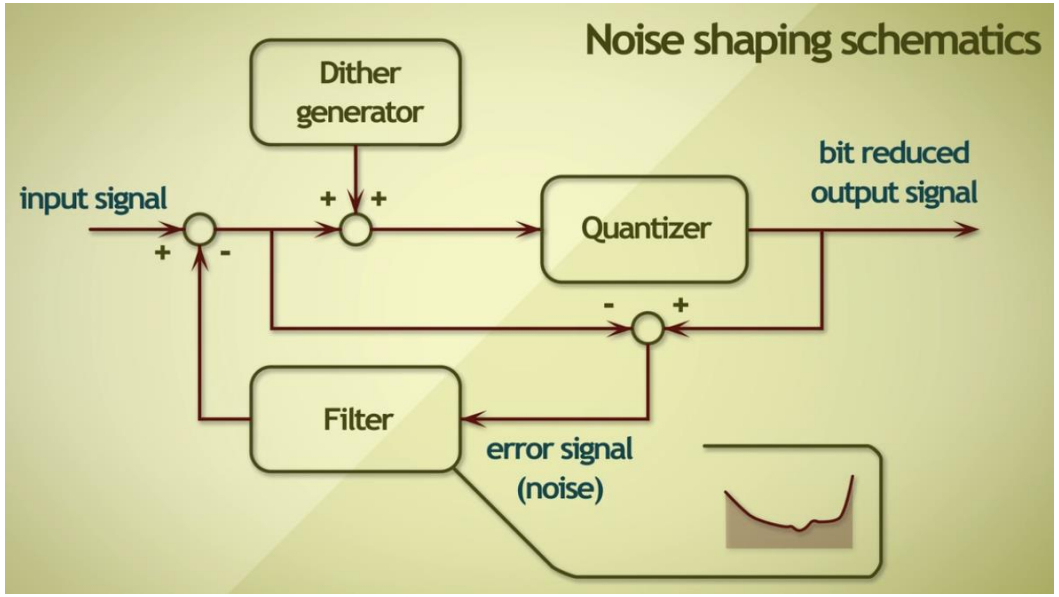


Structure of a Filter Bank

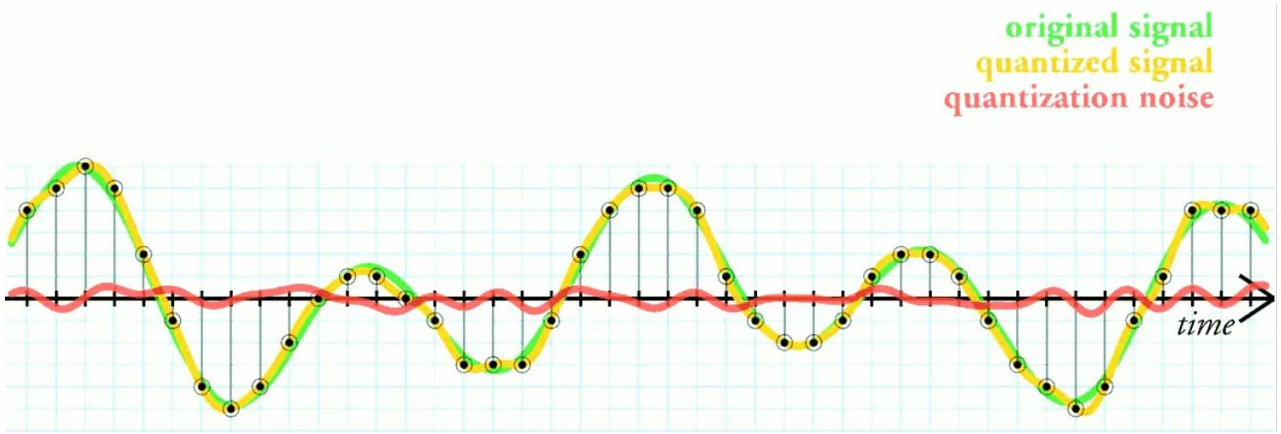
音频编码器 – 瞬时噪声整形（Temporal Noise Shaping）

原理

噪声是信号量化时被引入的，TNS技术为信号增加抖动（dither），并结合人类听觉特性，在500Hz~5kHz之间为信号做抖动减法，在6kHz以上为信号做抖动加法。从而实现在总噪声相同的情况下，更好的为噪声整形。



简单模型



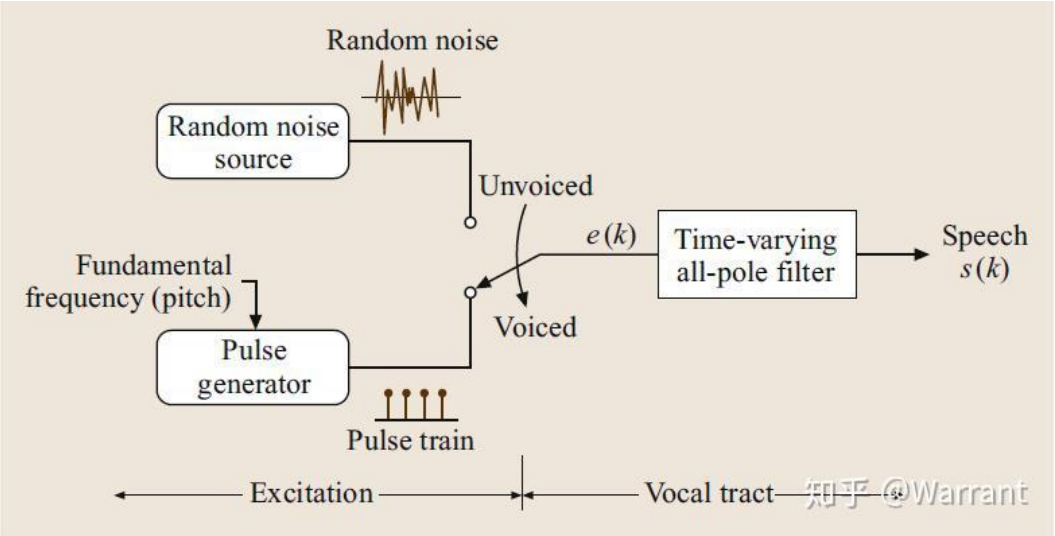
噪声产生过程

[Noise Shaping]

音频编码器 – 线性预测编码（Linear predictive coding）

线性预测编码

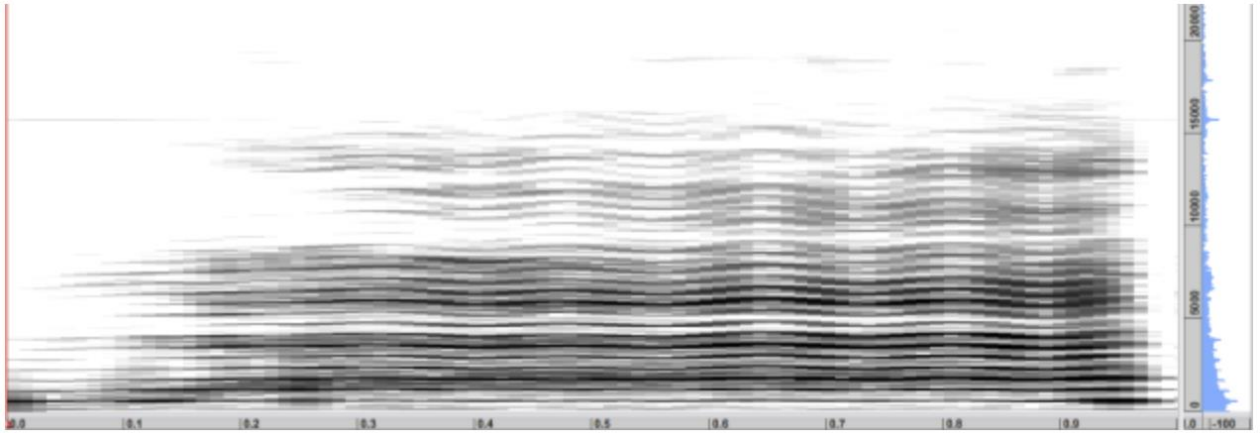
预测编码基于频域共振峰形成的原理，对发声模型进行数学建模，可以有效的预测波形。



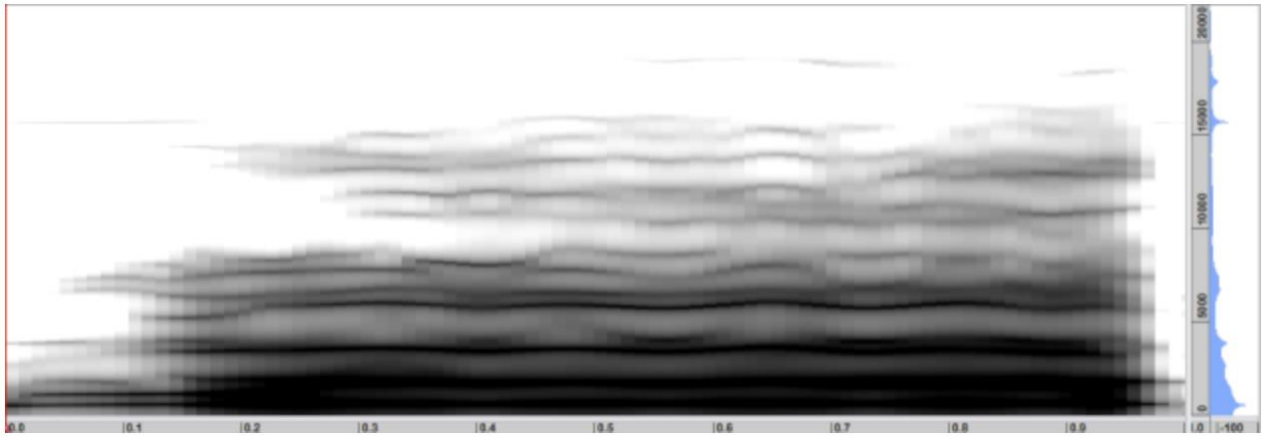
线性预测编码算法的原理是"一个语音取样的现在值可以用若干个语音取样过去值的加权线性组合来逼近"。以下方程式中，s(k) 代表真实值，e(k) 表示误差，线性加权部分表示预测值。

$$s(k) = \sum_{p=1}^{P} a_p s(k-p) + e(k),$$

其中， P是滤波器的阶数，a_p是滤波器系数。LPC就是在已知s(k)的情况下获取 a_p ， 常用方式是求解MSE。



一个元音的FFT语谱图



20阶系数LPC预测语谱图

音频编码器 – 立体声编码

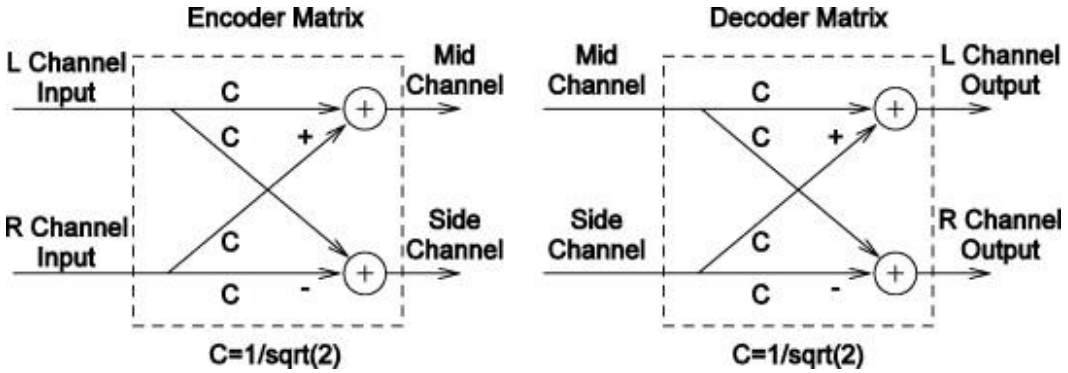
Mid/Side 立体声编码

在音频编码研究中，人们发现在左右声道之间往往存在一定的相关性和冗余度。例如，在欣赏立体声歌曲时，左右声道发送的是近似相同的音乐。如果能够消除这些相关性和冗余，则可以进一步提高压缩比。

M/S立体声编码可以消除左右声道的相关性，它用M、S声道来代替L、R声道。

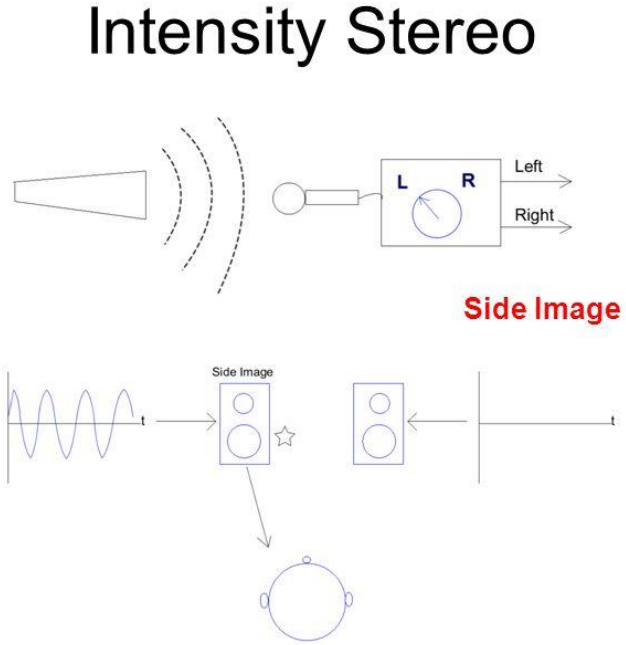
$$M = \frac{L+R}{2}, S = \frac{L-R}{2}$$

用M代替L，用S代替R。当L、R正相似时，M与L能量相近，S能量则远小于R的能量；当L、R反相似时，S的能量与R能量相近，M的能量远小于L的能量，由此可以实现减少编码比特数的目的。在解码端，将M、S声道恢复为L、R声道。



强度立体声（IS）编码

在高频段，人耳对声音的感知主要是基于声音的强度，对相位并不敏感。因此对于高频频谱，IS编码将左声道用于表示实际左右声道的联合强度，将右声道替换为零，同时记录左右声道各子带的能量比。实践证明，在音频编码的频带范围（20Hz ~ 20kHz）内，平均高于6kHz以上的频带就可以使用IS编码，IS编码至少能使右声道的编码比特数减少1/3。



音频编码器 – 非均匀量化

量化

用小集合表示更大的集合（可能是无限大）的过程，我们称为量化。

可以理解成是量化器拿着一把尺，把值从小刻度映射到大刻度。

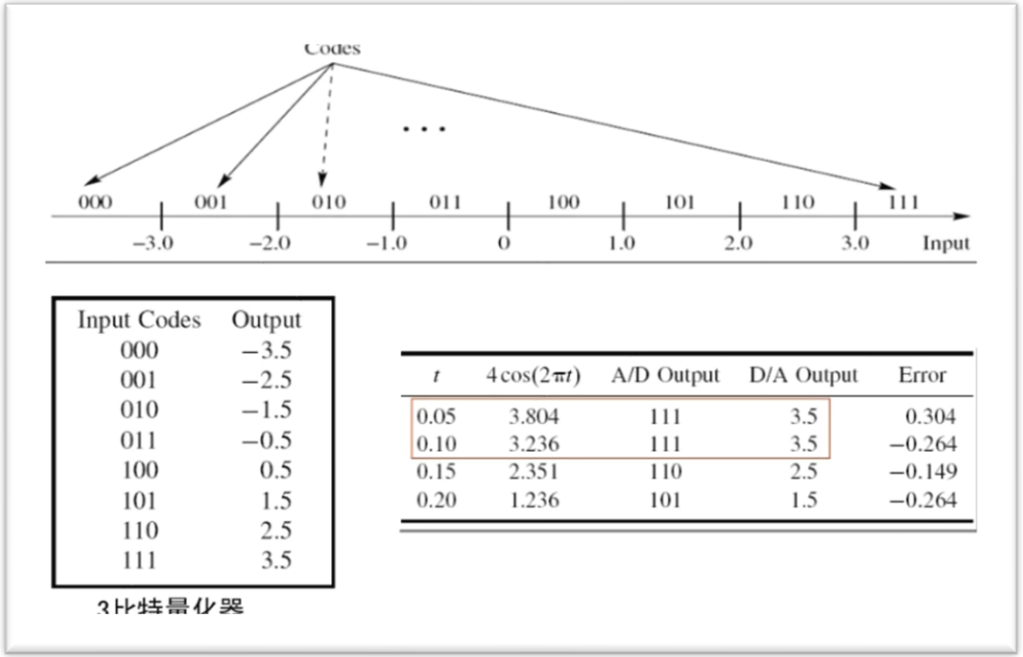
音频在通过MDCT滤波器组后，能量会被聚集在矩阵的左上角，更方便量化。

均匀量化

这把尺的刻度是均匀分布的。

非均匀量化

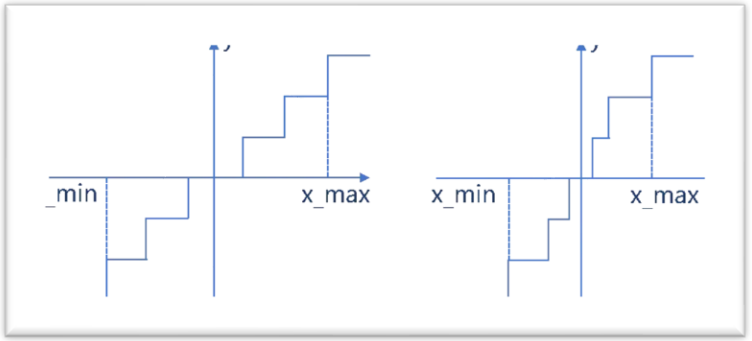
由于人耳对声音频率的感知度不同，因此，音频数据的量化在低频部分可以保留很多的信息，而在高频部分则保留相对少的信息。



Input Codes	Output
000	−3.5
001	−2.5
010	−1.5
011	−0.5
100	0.5
101	1.5
110	2.5
111	3.5

3比特量化器

t	$4\cos(2\pi t)$	A/D Output	D/A Output	Error
0.05	3.804	111	3.5	0.304
0.10	3.236	111	3.5	−0.264
0.15	2.351	110	2.5	−0.149
0.20	1.236	101	1.5	−0.264



均匀量化　　非均匀量化

原始：1 2 3 4 5 6 7 8 → 20%冗余 → 计算冗余：0.8 1.6 2.4 3.2 4.0 4.8 5.6 6.6 → 量化 → 量化后2：1 2 2 3 4 5 6 6

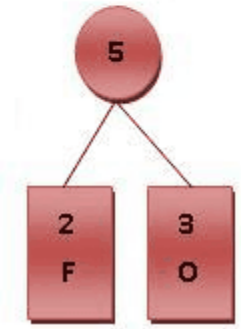
音频编码器 – 无噪声编码（霍夫曼编码）

定义

霍夫曼是一种无损（无噪声）字码编码，它使用变长编码表对信号进行编码，其中变长编码表是通过一种评估来源字码出现频率的方法得到的，对出现频率高的字母使用较短的编码，反之出现频率低的则使用较长的编码，这使编码后的字码平均长度、尽可能的少，从而达到无损压缩数据的目的。

Symbol	F	O	R	G	E	T
Frequency	2	3	4	4	5	7

生成树



生成码表

Symbol	F	O	R	G	E	T
Frequency	2	3	4	4	5	7
CODE	000	001	100	101	01	11

引用

- Nyquist采样定理
- 心理声学模型
- 线性预测编码
- 噪声整形
- 傅里叶变换
- 滤波器组
- 语谱图
- 音频分帧