CS 330 MIP – Lecture 13

音频信息处理 4 + 图像信息处理 1

Audio Information Processing 4 + Image Information Processing 1

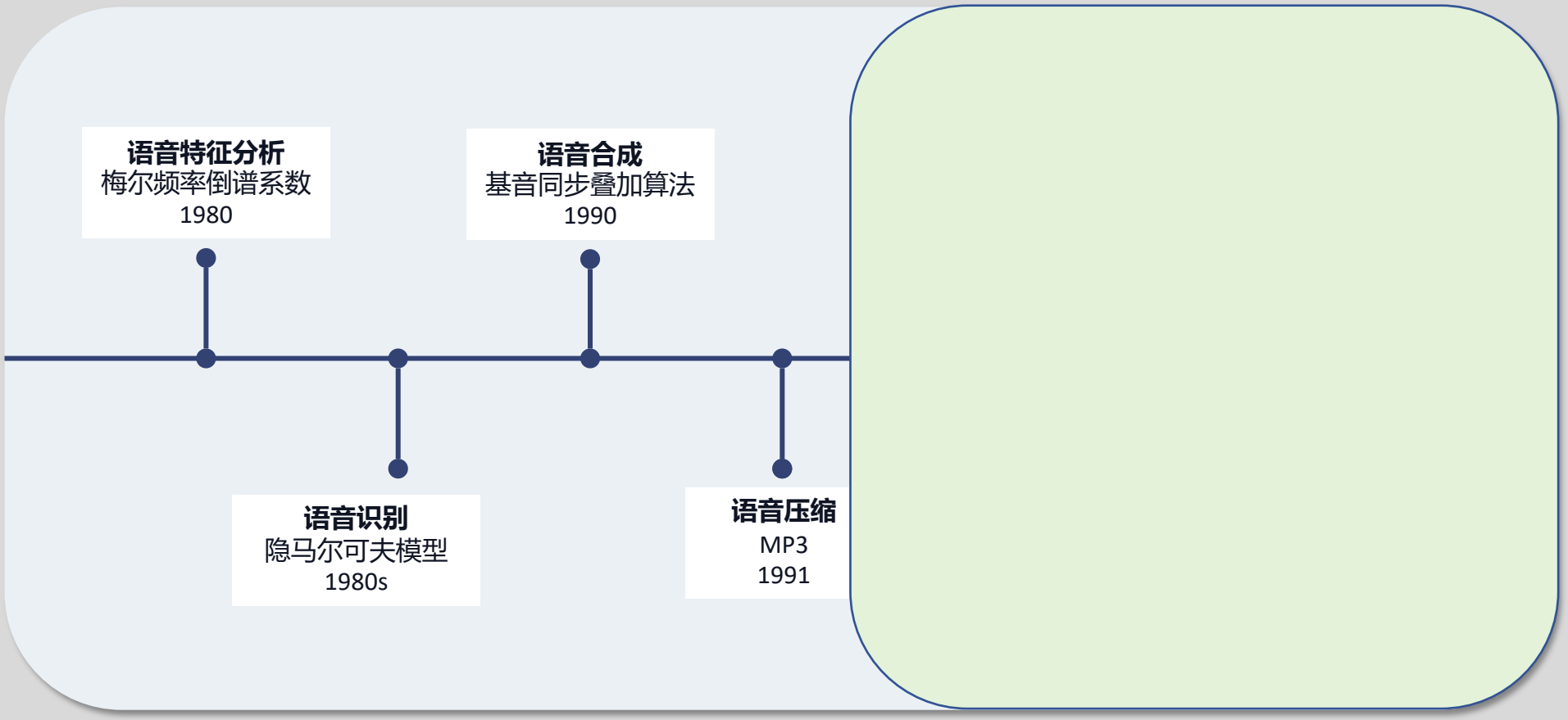
Jimmy Liu 刘江

2025-05-14

Lecture 13 Contents

1. Review of Lecture 12
2. 语音处理的7个里程碑之4-7
3. 图像信息处理7个里程碑及基本概念
4. 图像信息处理7个里程碑之1-2边缘检测和图像压缩

音频语音信息处理7个里程碑之4



里程碑4: 语音压缩之MP3

- 语音压缩是对编码后的数字语音进行压缩的方法。
  - MP3算法（MPEG Audio Layer-3）是一种广泛使用的数字音频压缩格式，其全称是动态影像专家压缩标准音频层面3。
  - 它是在1991年由德国的的一组工程师发明和标准化的。它被设计用来大幅度地降低音频数据量。
  - 利用 MPEG Audio Layer 3 的技术，将音乐以1:10 甚至 1:12 的压缩率，压缩成容量较小的文件，而对于大多数用户来说重放的音质与最初的不压缩音频相比没有明显的下降。
  - 用MP3形式存储的音乐就叫作MP3音乐，能播放MP3音乐的机器就叫作MP3播放器。
  - MP3算法的核心原理是基于人耳听觉特性的心理声学模型，通过去除人耳不易察觉的声音信息来实现数据压缩。

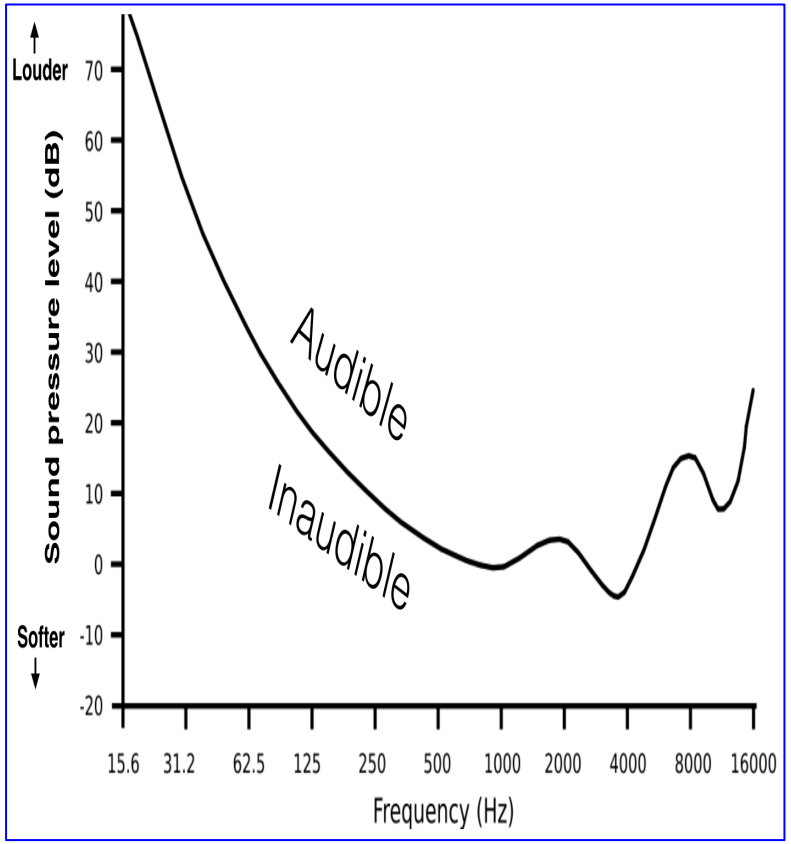
心理声学 Psycho-Acoustics

- **心理声学**：是声学的一个分支，涉及人类听觉感知，从耳朵的生物结构到大脑对声音信息的解释。MP3的压缩基于心理声学理论。

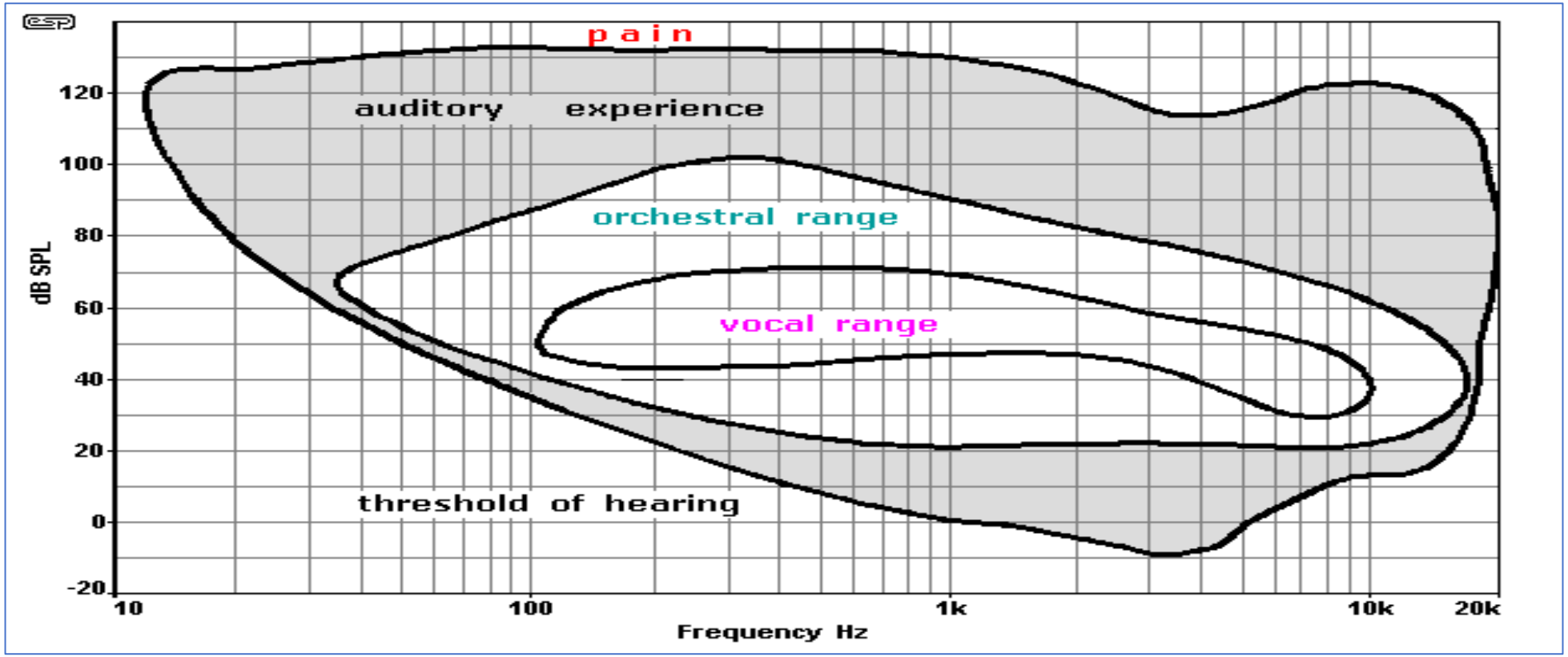
- 用于测量声音响度或人类耳朵感知的声音信号与噪声比（SNR）的单位是分贝（dB），1分贝是贝尔的十分之一。

心理声学特征 PAC 1:最小听觉阈值

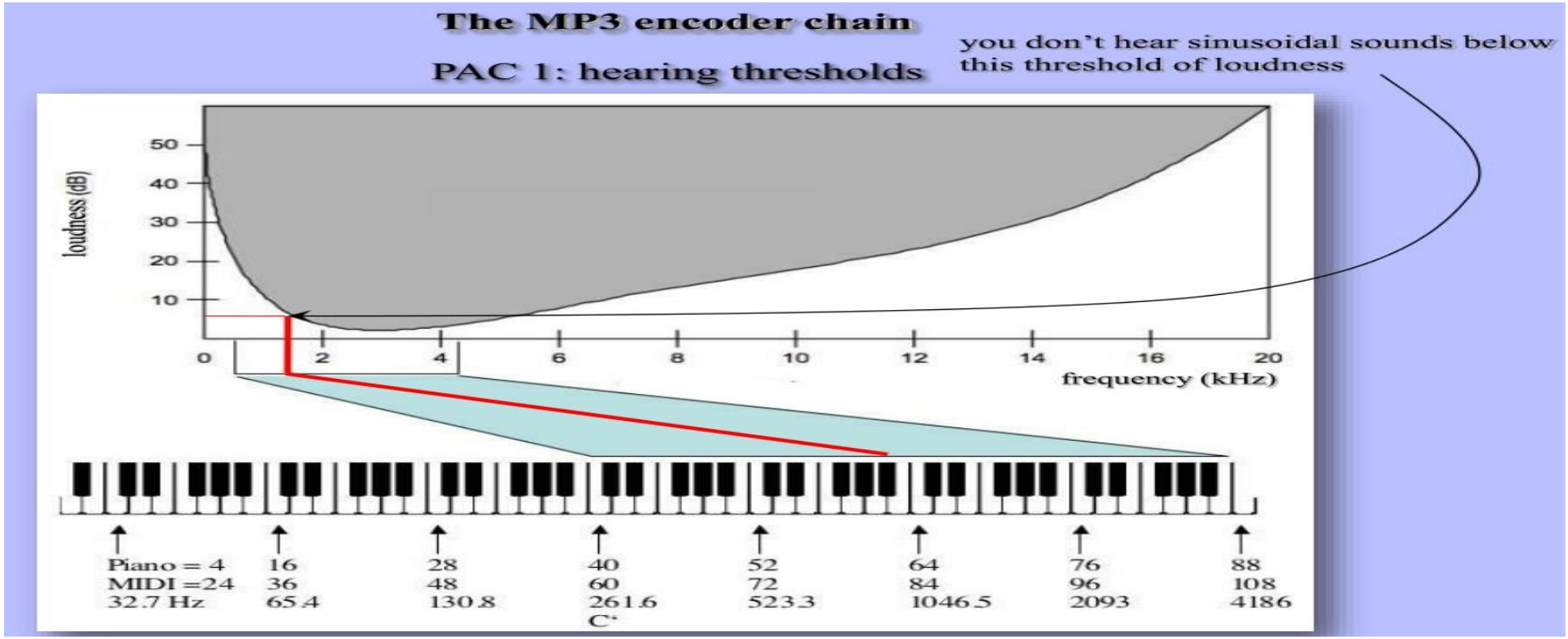
- **最小听觉阈值**是指正常人耳能够检测和听到的最微弱的声音。
- 然而，人耳的灵敏度是频率依赖的。最大灵敏度出现在1到5千赫兹（kHz）之间，而在低频和高频区域相对较不敏感（人声频率大约在85-1100赫兹（Hz）之间）。
- 当将最小阈值，即最微弱可听声音的振幅，与频率值相对应地绘制出来时，就会得到最小听觉阈值曲线。



人类听觉

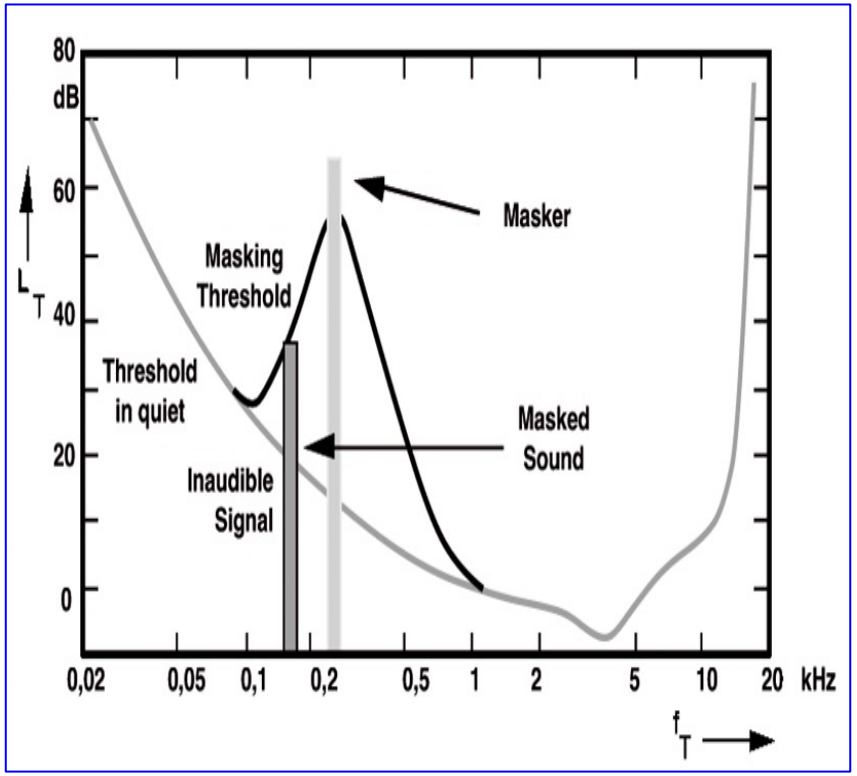

pain
auditory experience
orchestral range
vocal range
threshold of hearing
dB SPL
120
100
80
60
40
20
0
-20
10
100
1k
10k
20k
Frequency Hz

听觉阈值

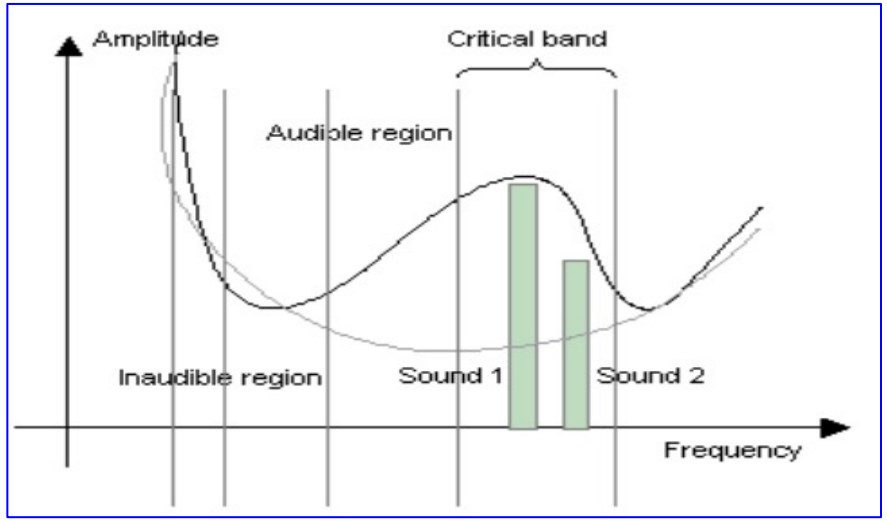

The MP3 encoder chain
PAC 1: hearing thresholds
you don't hear sinusoidal sounds below this threshold of loudness
loudness (dB)
50
40
30
20
10
0
2
4
6
8
10
12
14
16
20
frequency (kHz)
Piano = 4
MIDI =24
32.7 Hz
16
36
65.4
28
48
130.8
40
60
261.6
C'
52
72
523.3
64
84
1046.5
76
96
2093
88
108
4186

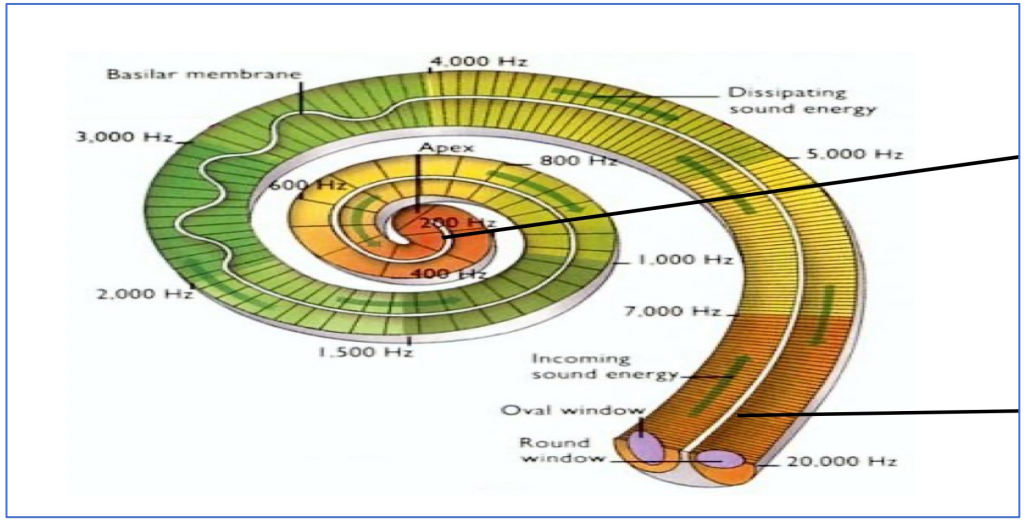
心理声学特征 PAC 2: 振幅掩蔽

- **振幅掩蔽**：在临界频带内，只有最强的声音会被听到，而其他声音则会被掩盖。
- 振幅掩蔽现象发生的原因是，可听声音倾向于扭曲阈值曲线并将其向上移动。曲线的扭曲程度被限制在最强声音周围的较小区域内。整个可听频率范围被划分为多个这样的区域，称为临界频带。



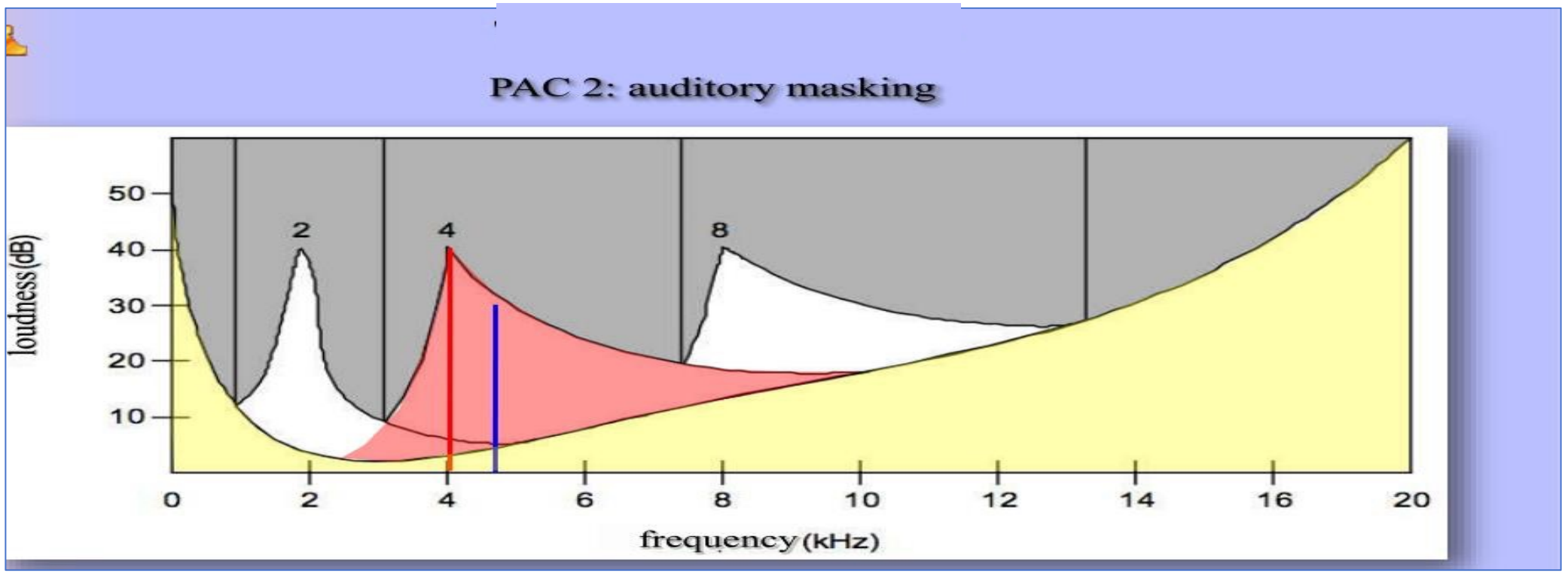
声音1的出现使声音2在同一临界频带内不可闻



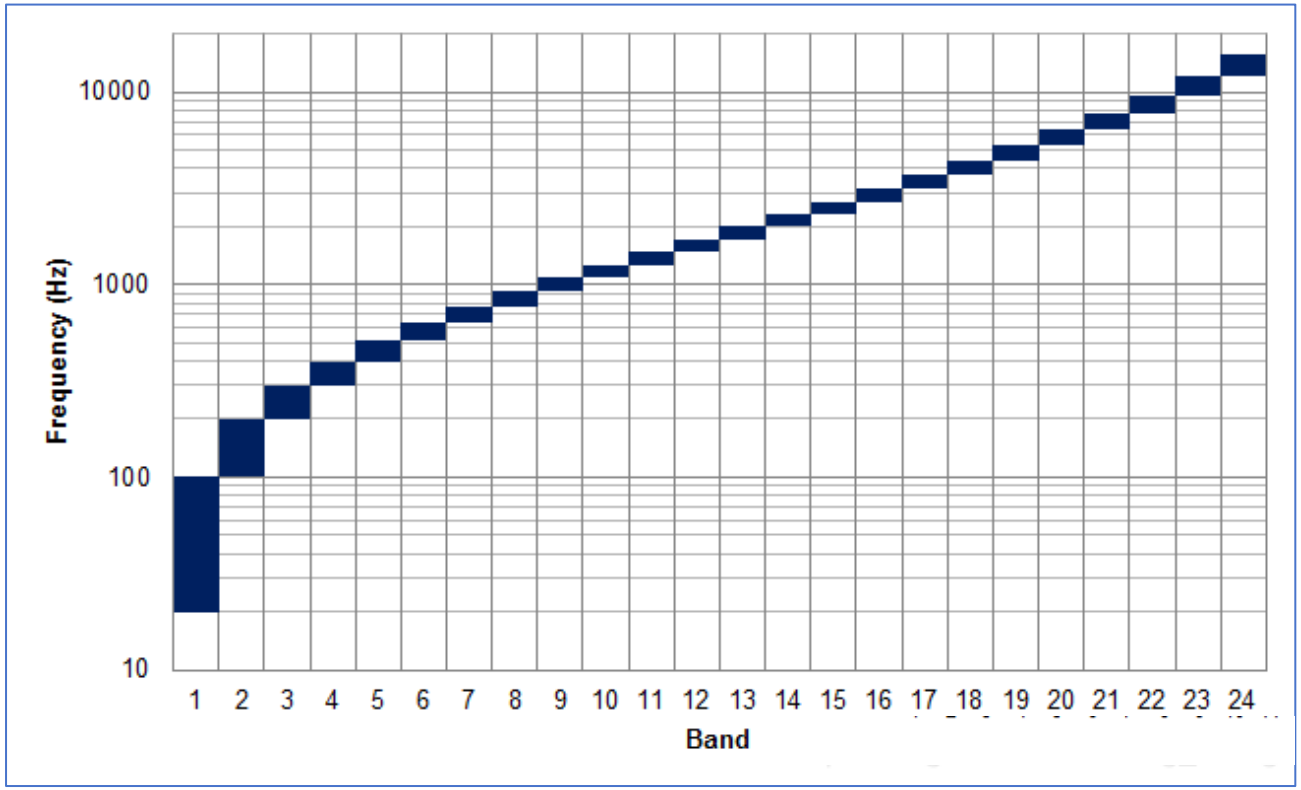


对于给定的**频率**，**临界频带**是指围绕该频率的最小频率范围，能够激活耳蜗中相同部分的**基底膜**。临界频带宽度代表了人耳对同时出现的音调或泛音的分辨能力。

不同频率带中的振幅掩蔽

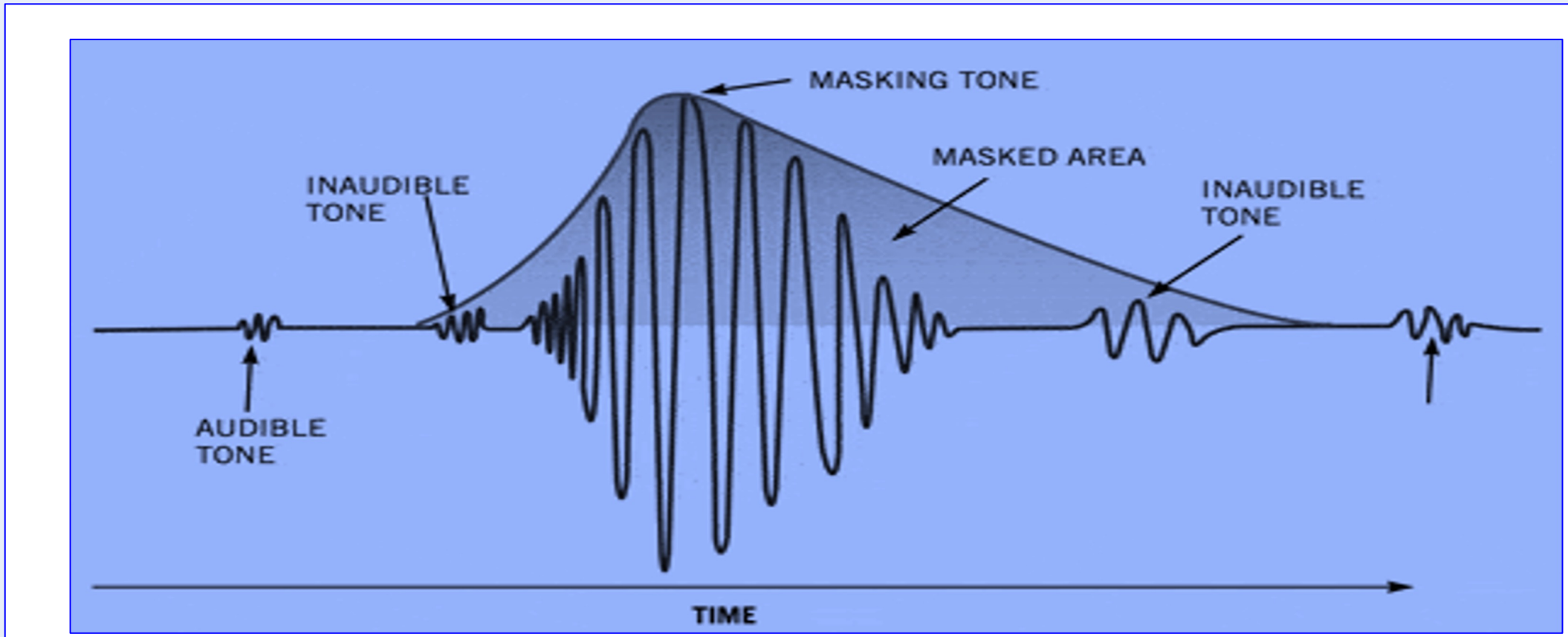


频率带——24巴克尺度

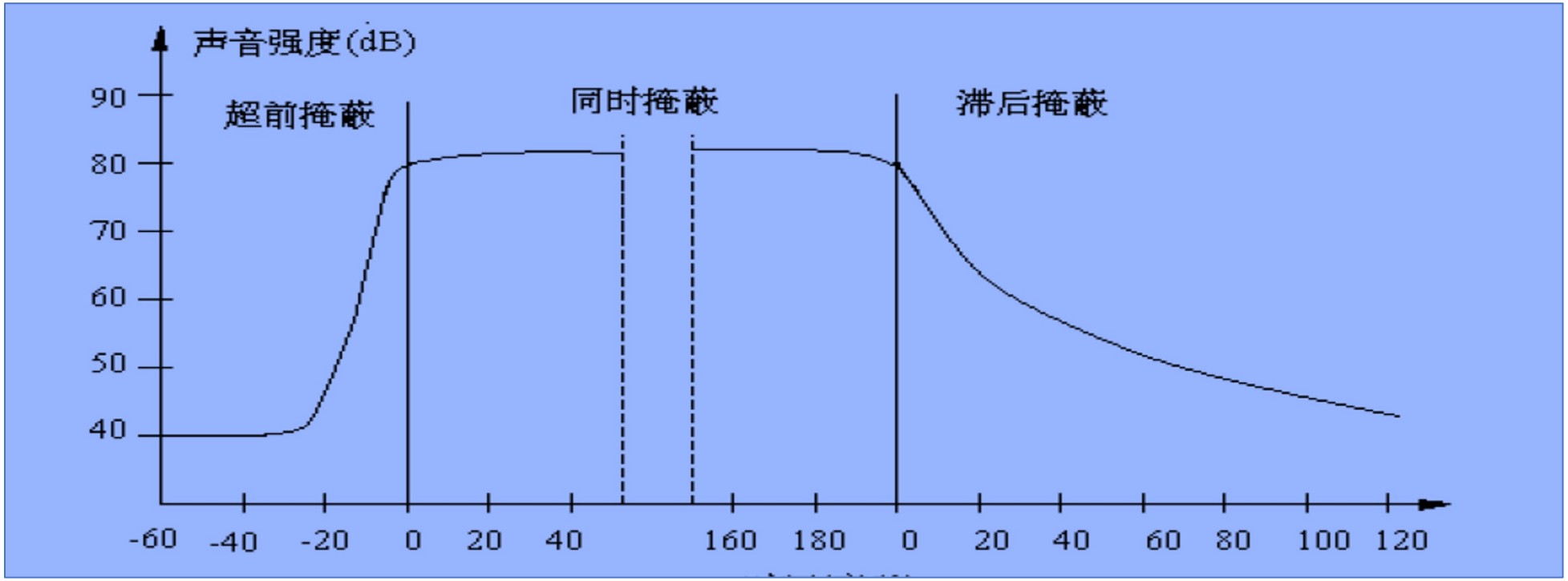


Index	Bark Scale		Mel Scale	
	Center Freq. (Hz)	BW (Hz)	Center Freq. (Hz)	BW (Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	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

心理声学特征 PAC 3


MASKING TONE
MASKED AREA
INAUDIBLE TONE
INAUDIBLE TONE
AUDIBLE TONE
TIME

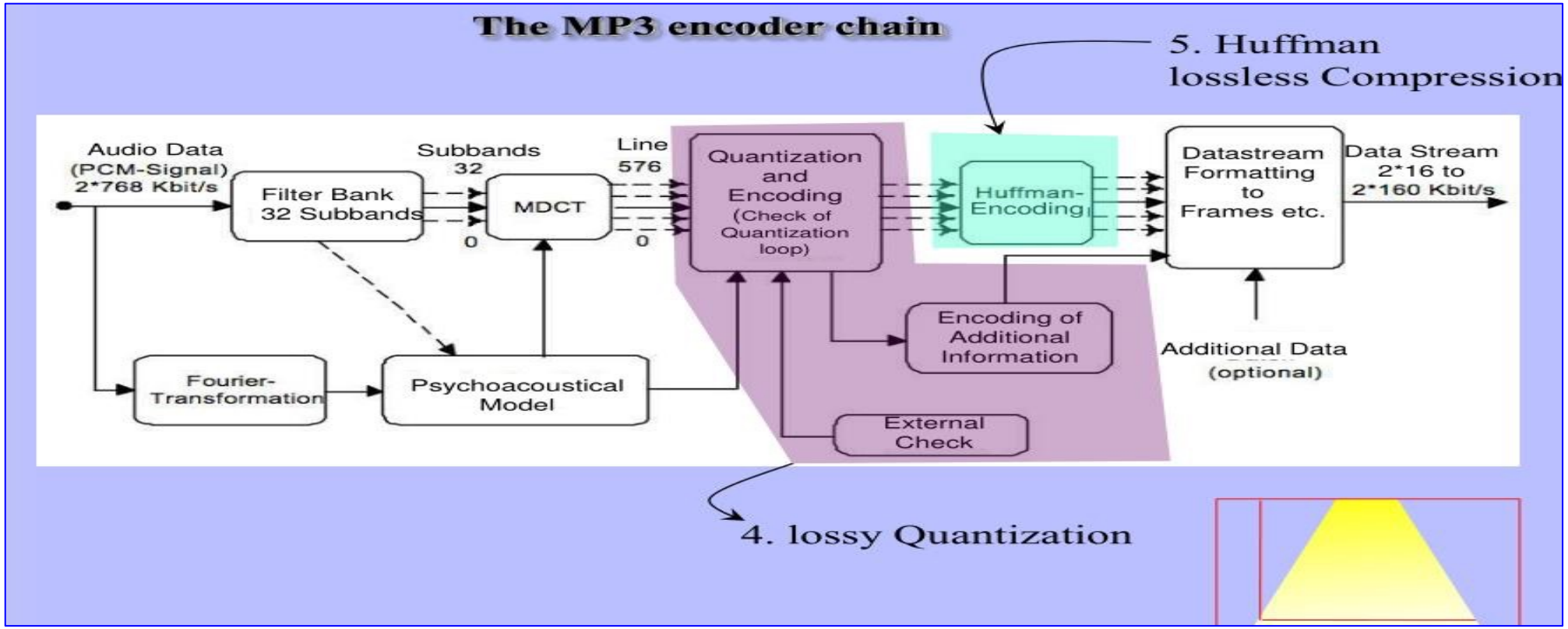
前掩蔽可能持续5到20毫秒，而后掩蔽可能持续50到200毫秒



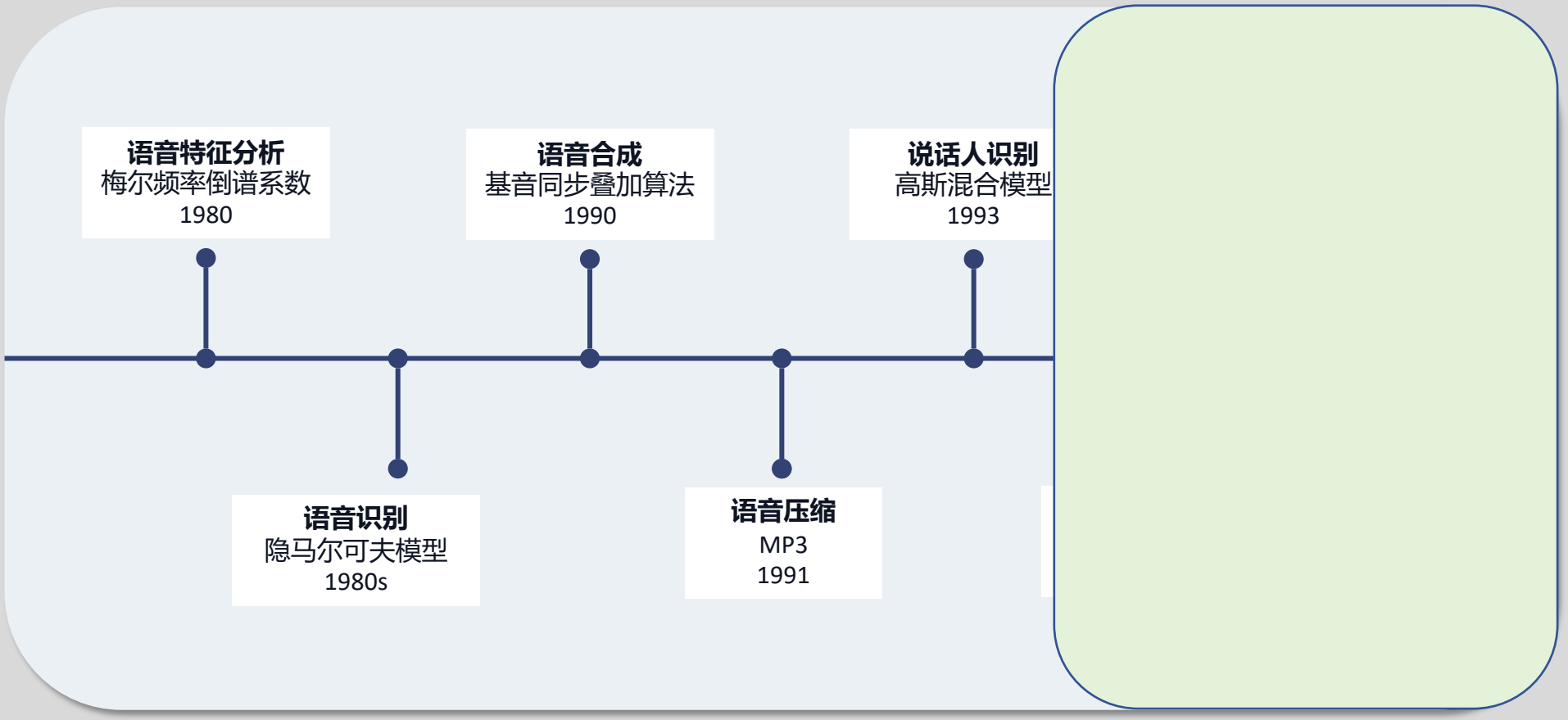
MP3算法步骤

步骤	描述
采样率转换	将输入音频信号的采样率转换为固定的值，通常为44.1 kHz。这是为了匹配人耳对于音频的感知范围，削弱或删除高于人耳感知范围的频率。
分帧	将音频信号分成一系列短时窗口，通常为23.2 ms至46.4 ms的长度。每个窗口内的音频数据被视为一个帧。使用重叠窗口技术减少帧之间的不连续性。
快速傅里叶变换（FFT）	对每个帧应用FFT变换，将时域中的音频信号转换为频域中的频谱表示。
声学模型	基于人耳的听觉特性，使用心理声学模型来确定哪些频率成分对人耳更重要，并舍弃部分人耳感知不灵敏的部分以进行更多的压缩。
量化和编码	使用掩蔽模型为每个频率成分确定对应的量化器步长。然后，将量化后的频谱系数进行熵编码，通常使用霍夫曼编码。霍夫曼编码是一种无损数据压缩算法，它基于频率较高的字符用较短的编码表示，而频率较低的字符用较长的编码表示的原理，实现对数据的高效压缩。

MP3压缩全流程


The MP3 encoder chain
5. Huffman lossless Compression
Audio Data (PCM-Signal) 2*768 Kbit/s
Filter Bank 32 Subbands
Subbands 32
MDCT
Line 576
0
0
Quantization and Encoding (Check of Quantization loop)
Huffman-Encoding
Datastream Formatting to Frames etc.
Data Stream 2*16 to 2*160 Kbit/s
Encoding of Additional Information
Additional Data (optional)
Fourier-Transformation
Psychoacoustical Model
External Check
4. lossy Quantization

音频语音信息处理7个里程碑之5



说话人识别 Automatic Speaker Recognition

- 说话人识别，也称声纹(Voiceprint)识别，是一种利用说话人的语音特征进行身份辨认或确认的技术，属于生物识别技术的一种。它的原理是通过分析处理说话人的语音信号，提取出包含在其中的个性因素，如发音器官和发音习惯的差异，从而将不同人的声音进行有效区分。在实际应用中，说话人识别可以分为说话人确认和说话人辨认两个应用范畴。
- 说话人识别的理论基础是每一个声音都具有独特的特征，通过该特征能将不同人的声音进行有效的区分。
  - 这种独特的特征主要由两个因素决定，第一个是声腔的尺寸，具体包括咽喉、鼻腔和口腔等，这些器官的形状、尺寸和位置决定了声带张力的大小和声音频率的范围。因此不同的人虽然说同样的话，但是声音的频率分布是不同的，听起来有的低沉有的洪亮。每个人的发声腔都是不同的，就像指纹一样，每个人的声音也就有独特的特征。
  - 第二个决定声音特征的因素是发声器官被操纵的方式，发声器官包括唇、齿、舌、软腭及腭肌肉等，他们之间相互作用就会产生清晰的语音。而他们之间的协作方式是人通过后天与周围人的交流中随机学习到的。人在学习说话的过程中，通过模拟周围不同人的说话方式，就会逐渐形成自己的声纹特征。

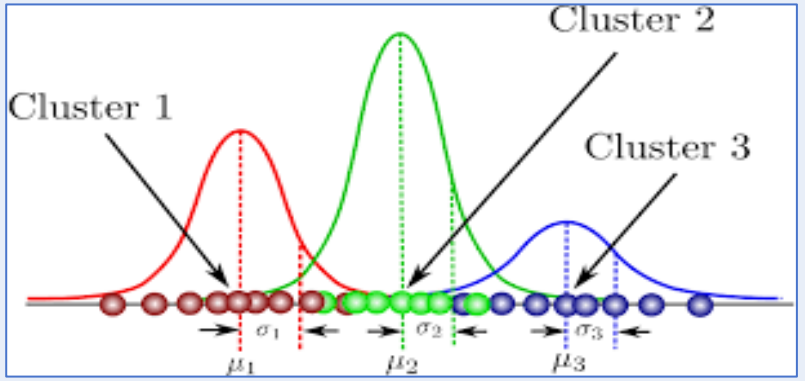
说话人识别与语音识别

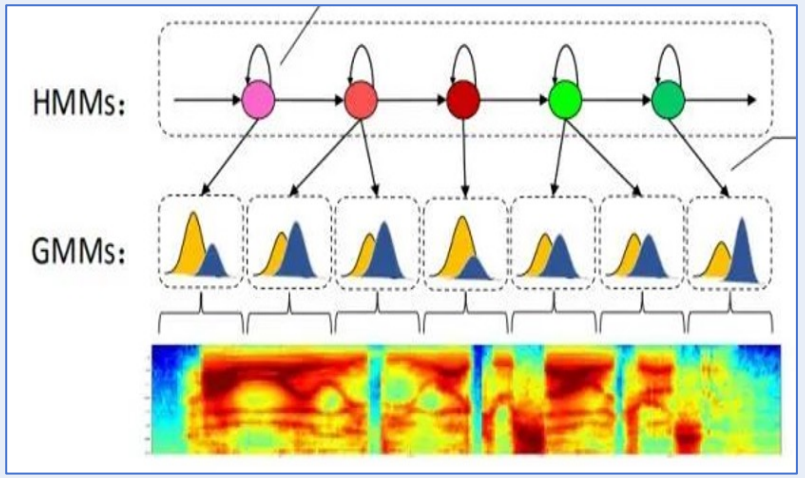
- 从目的上看，说话人识别是通过分析处理说话人的语音信号，提取出包含在其中的个性因素，如发音器官和发音习惯的差异，从而进行身份鉴别与认证。而语音识别技术的目标则是将人类语音中的词汇内容转换为计算机可读的输入，即让机器“听懂”人类口述的语言，包括理解口述语言中的要求或询问并做出正确响应。
- 从原理上看，说话人识别是基于声纹识别的一种生物识别技术，通过分析语音信号中的个性因素来识别说话人的身份。而语音识别则是将语音信号转变为文本，然后将理解转变为指令的技术。

里程碑5: 说话人识别之GMM

- 高斯混合模型（Gaussian Mixture Model, GMM）是一种概率模型，它假设数据集是由多个高斯分布混合而成的。每个高斯分布被称为一个“成分”或“簇”，由其均值（mean）和协方差（covariance）定义。

- 在说话人识别任务中，GMM通过训练为每个目标说话人语音建立一个特征模型，再通过匹配处理来获得最终的识别结果。在语音识别任务中，GMM也可以与HMM结合来提高语音识别的准确性。

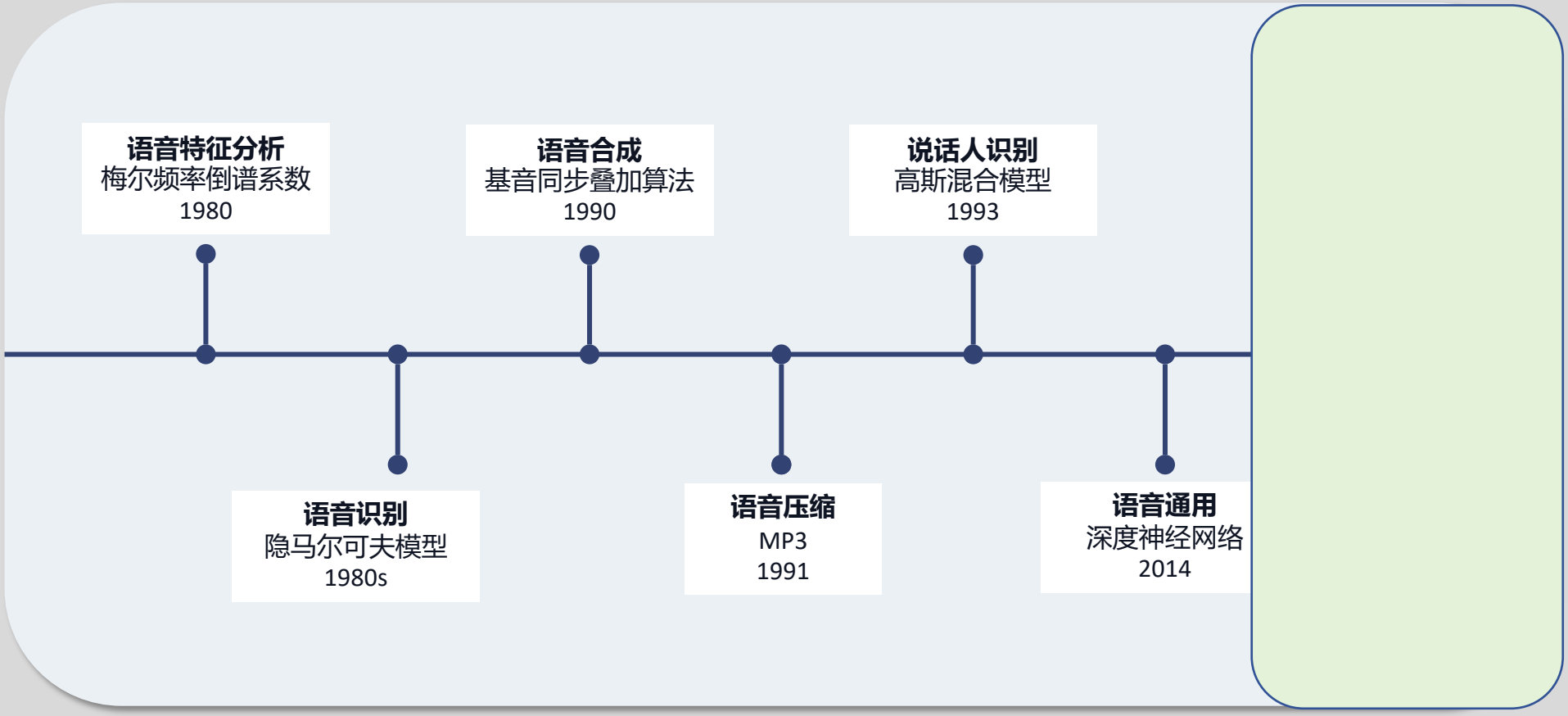




GMM说话人识别

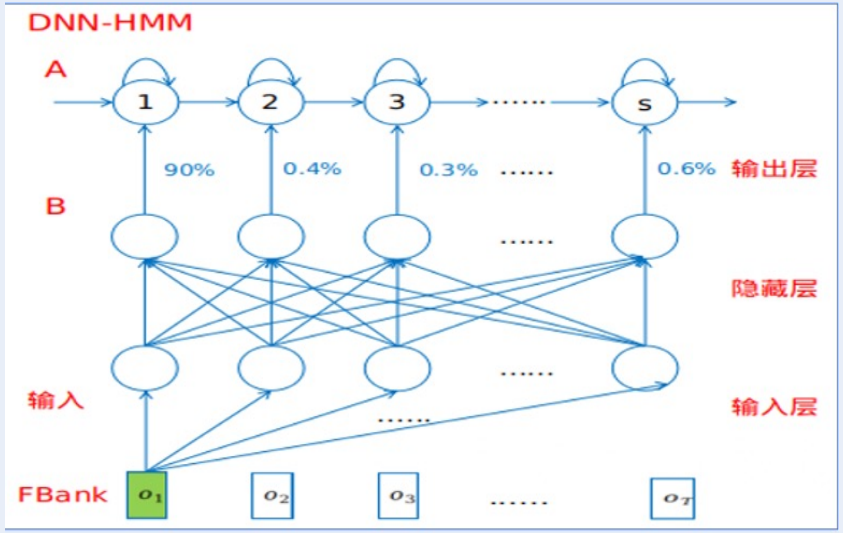
- 在说话人识别中，GMM的核心设定是将每个说话人的音频特征用一个高斯混合模型来表示。这种设定基于一个直观的理解：每个说话人的声纹特征可以分解为一系列简单的发音子概率分布，这些子概率分布可以近似的认为是正态分布（高斯分布）。
- GMM如需要训练出每个说话人的声音模型，因此需要大量的训练数据和时间成本。此外，如果针对开集新人员进行识别，需要重新训练模型，限制了GMM实用性。
- GMM-UBM（高斯混合模型-通用背景模型）主要用于开集的说话者辨认。UBM是从大量不同的说话人的背景数据中训练而来的，用于建模整个数据集中存在的变异性。然后，通过使用UBM的信息，对特定说话者的GMMs进行调整以更好地匹配其特征。UBM不会受到训练数据不足以及隐性数据（unseen data）的影响。

音频语音信息处理7个里程碑之6

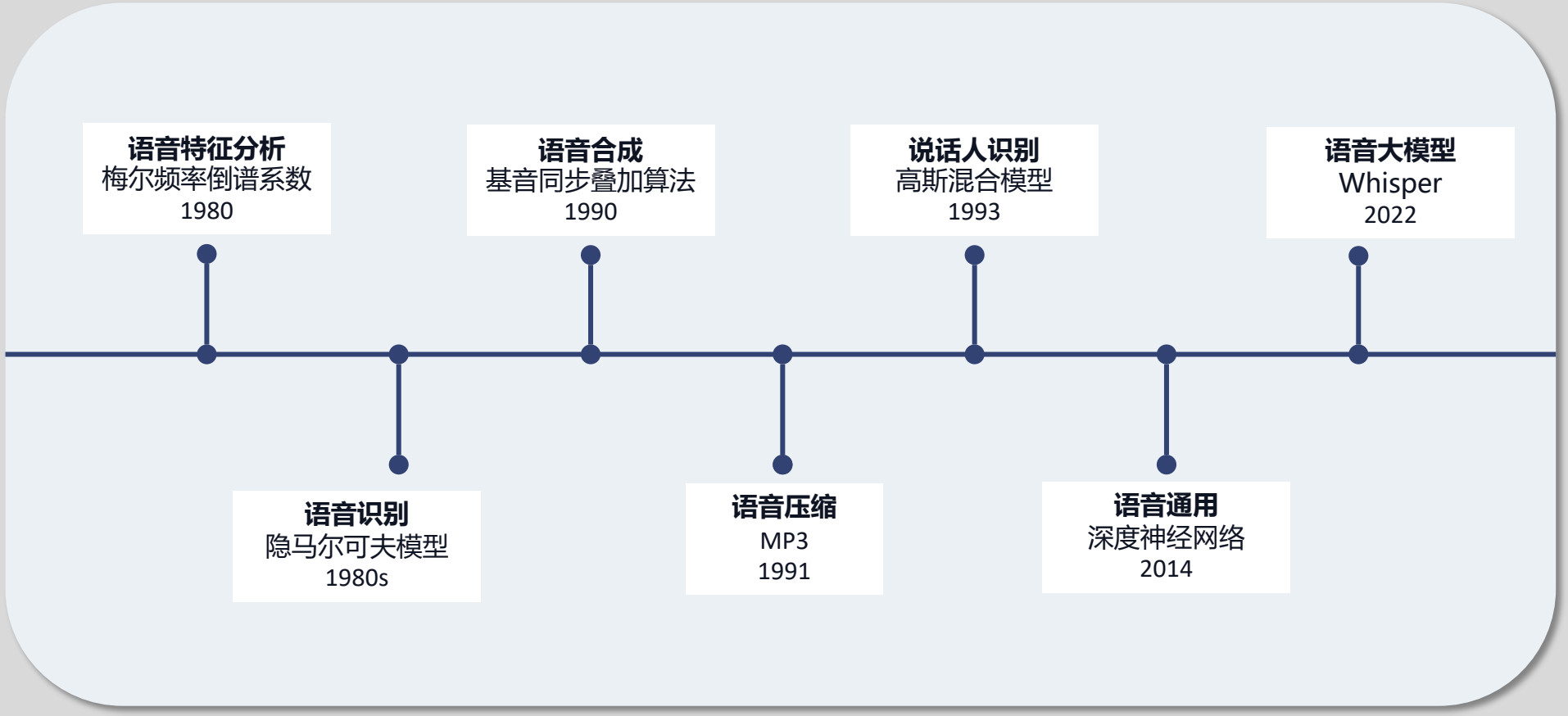


里程碑6:语音通用深度神经网络

- 语音通用深度神经网络（通常简称为语音深度神经网络或语音DNN）是深度学习领域中的一种模型，专门用于处理和分析语音数据。这些网络利用大量的参数和复杂的结构来捕获语音信号中的复杂模式，并在各种语音处理任务中取得显著的效果。
- 语音DNN可以与HMM结合，语音DNN学习特征表示和预测HMM中的状态转移概率，而HMM则描述语音信号的时序结构和状态转移关系。语音DNN-HMM模型使用深度神经网络对语音信号进行特征提取和转换，并输入到HMM中进行建模和解码，提高语音识别的性能。



音频语音信息处理7个里程碑之7



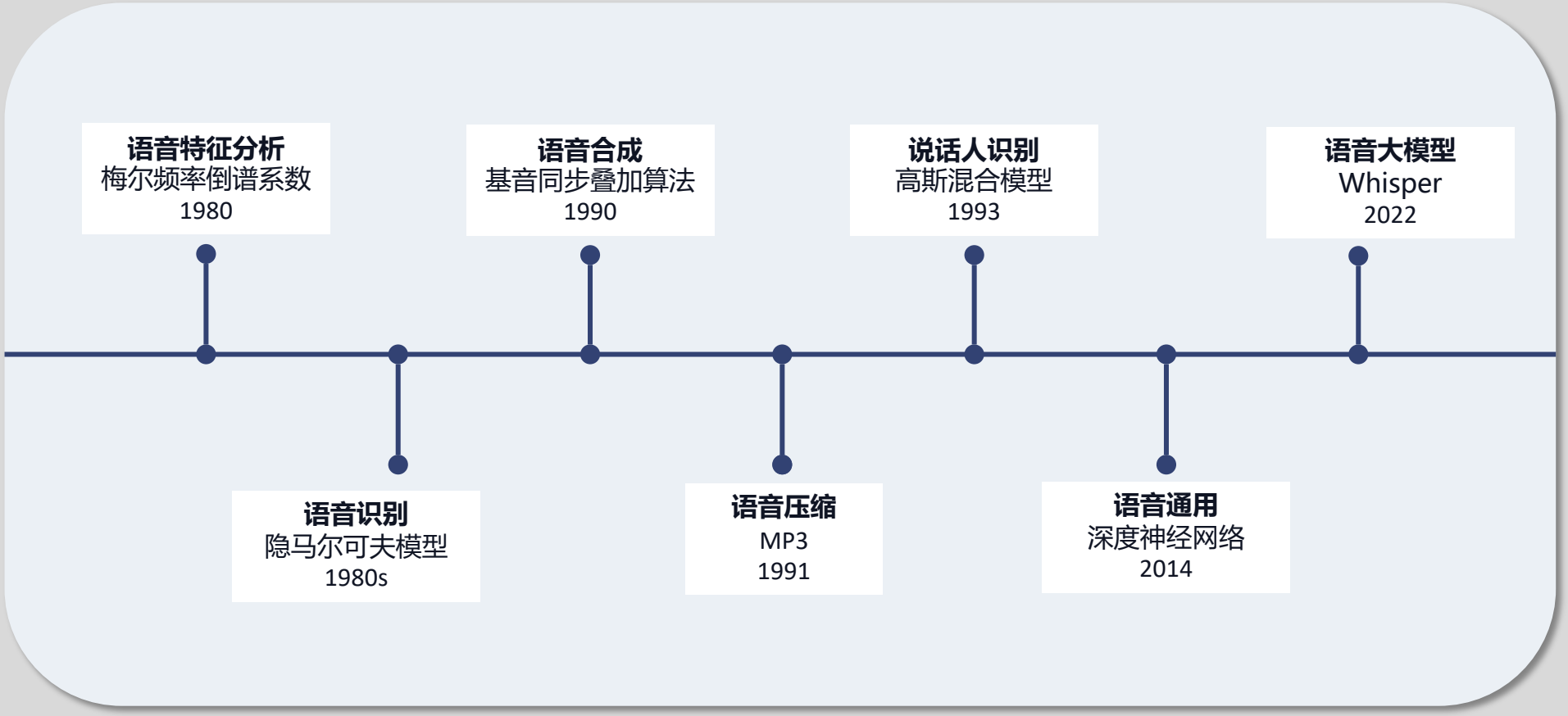
里程碑7: 语音大模型之Whisper

- Whisper模型是2022年 OpenAI公司开发的一种的基于Transformer 模型的预训练端到端模型，它集成了多语种ASR、语音翻译、语种识别的功能。
  - 在预处理阶段，它使用25毫秒的窗口和10毫秒的步幅计算80通道的log Mel谱图表示。
  - 在训练阶段，超过68万小时的标记多语言和多任务监督训练数据使其能够适应不同的口音、背景噪音和技术术语。
  - 在识别阶段，Whisper模型使用的CTC（Connectionist Temporal Classification）解码算法将神经网络输出的概率分布映射到最可能的文本序列。
  - 在后处理阶段，它通过语言模型纠正拼写纠错，进一步提高识别准确率，其转译效果已经接近人类专家。
- 2024年OpenAI推出的GPT-4o 进一步提升了系统的语音和多媒体功能

ASR与TTS的端到端模型

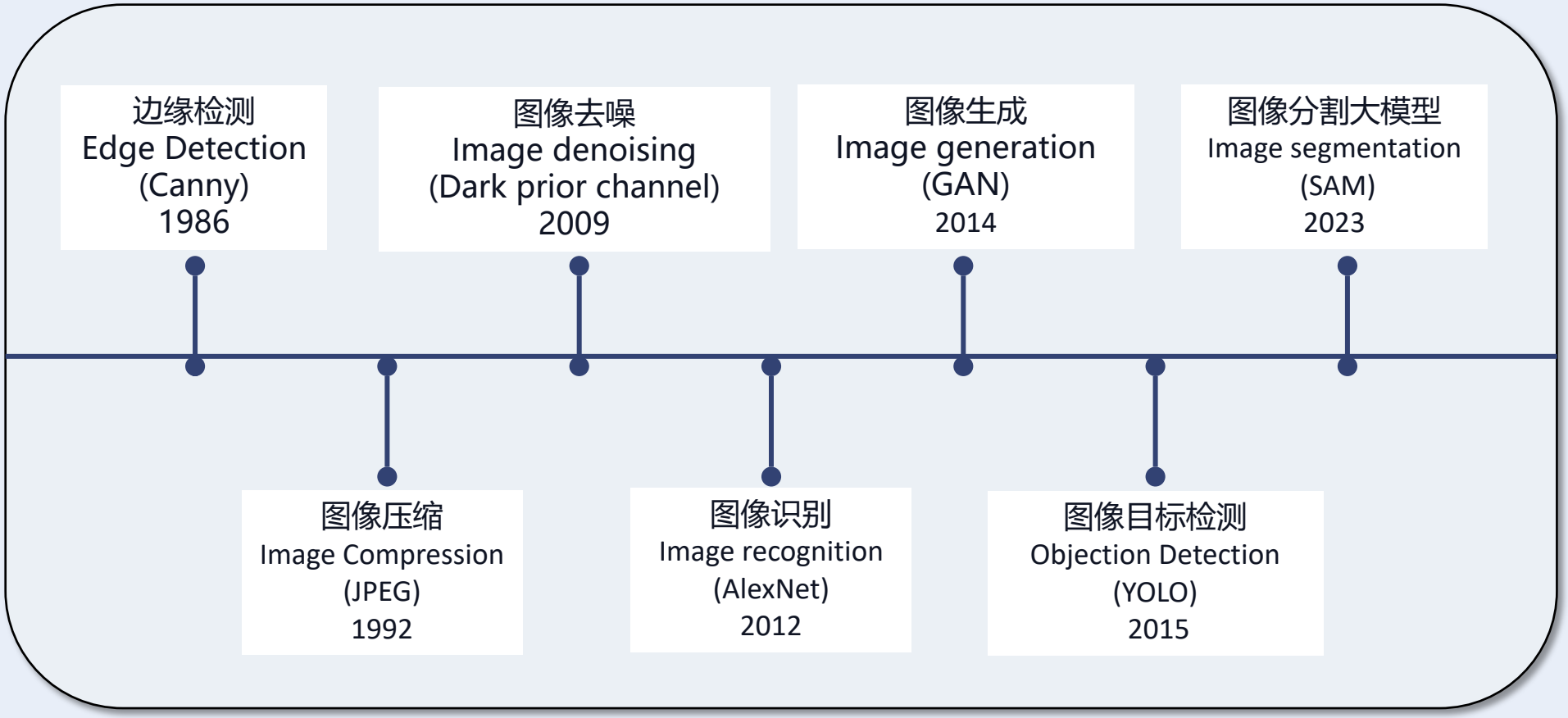
- 传统的语音识别ASR系统需要首先提取声学特征，语音或者音素特征，语音统计特征构建声学模型/语言模型/语音模型3个步骤。传统的语音生成TTS系统需要构建文本分析，声学模型，语音合成几个步骤。这些步骤需要大量的行业知识。
- 端到端的技术将以上的每个步骤直接用深度学习网络取代，简化了系统设计。ASR使用的CTC， RNN-T(Transducer)， LAS（Listen Attend Spell)和TTS使用的Tacotron, WaveNet, DeepVoice等的效果已经超越传统模型

总结


语音特征分析
梅尔频率倒谱系数
1980
语音识别
隐马尔可夫模型
1980s
语音合成
基音同步叠加算法
1990
语音压缩
MP3
1991
说话人识别
高斯混合模型
1993
语音通用
深度神经网络
2014
语音大模型
Whisper
2022

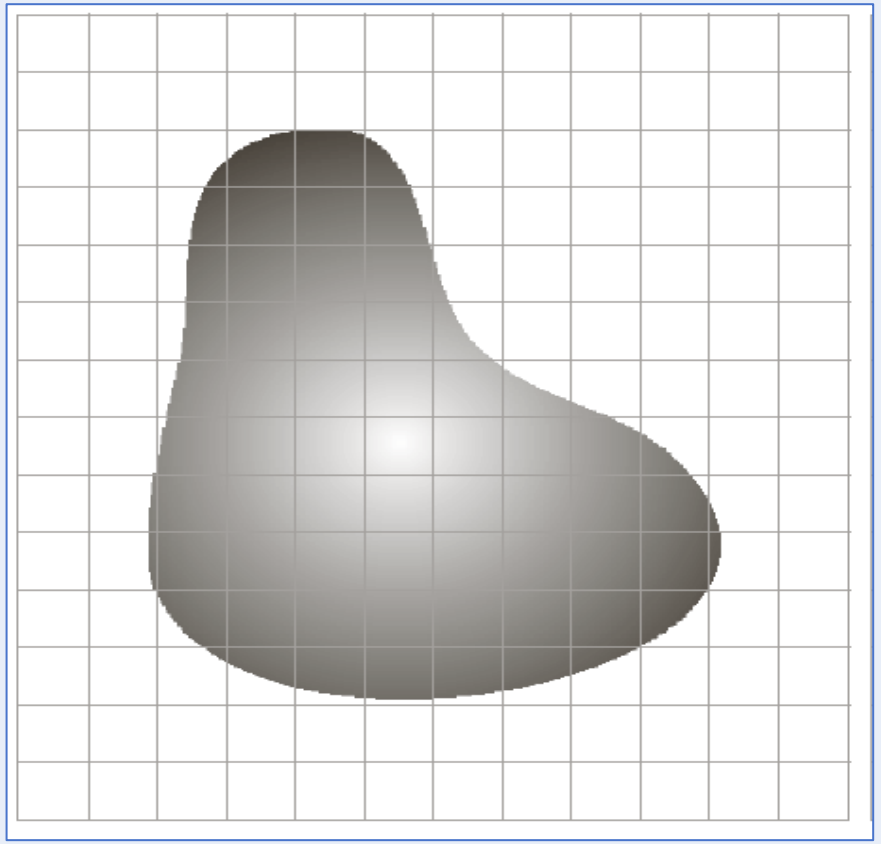
Lecture 13 Contents

图像信息处理7个里程碑



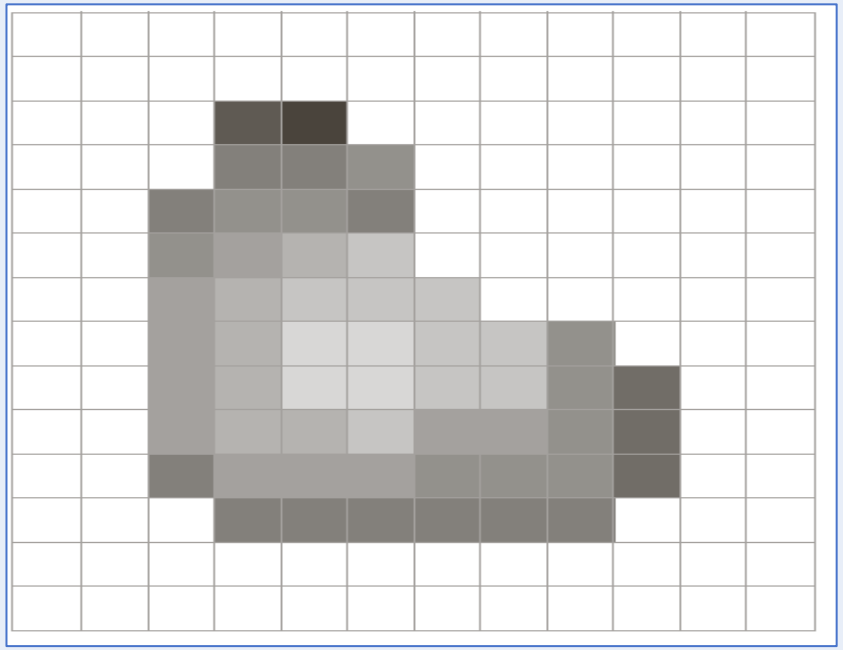
图像数字化步骤1-图像采样

图像采样是对二维空间上连续的图像进行在水平、垂直方向上等间距的分割，分割结果为矩形网状结构，连续图像被划分为一个个小的区域（称为像素或像素点），每个像素的位置由其坐标（x, y）表示。设连续图像f(x,y)经过数字化后，可以用一个离散量组成的矩阵g(i,j)（即二维数组）来表示。g(i,j)代表的点（i,j）即为采样点（sampling point），也称灰度值。



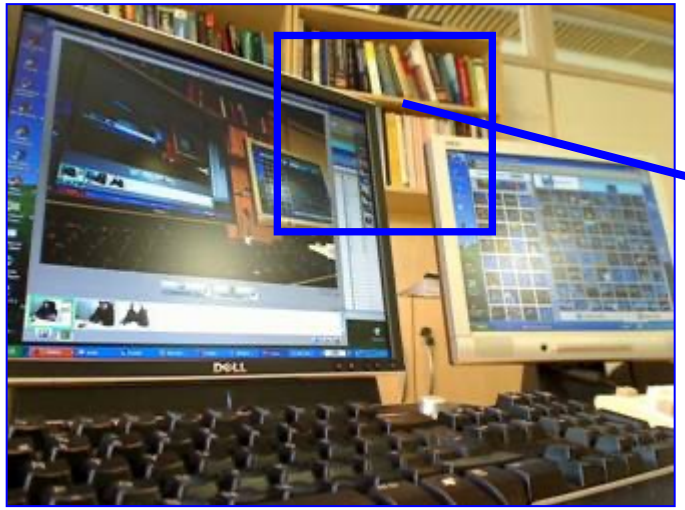
图像数字化步骤2-图像量化

图像量化是将采样后得到的像素值（即灰度值）进行离散化处理的过程。原来在一个图像块中，幅值是连续变化的，量化操作将这些幅值量化成有限个离散值。量化决定了图像能够容纳的颜色或者灰度总数，反映了采样的质量。



图像数字化步骤3-图像编码成像素

像素是对所见世界的普通的、低级的表示。



123 33 234 45 67
90 12 134 34 56
89 54 67 98 111
56 67 90 65 34 ….

像素值表示所看物体的亮度和颜色，但没有指示这些数字指的是什么物体，例如书籍、显示器——因此是低级别和直接的。

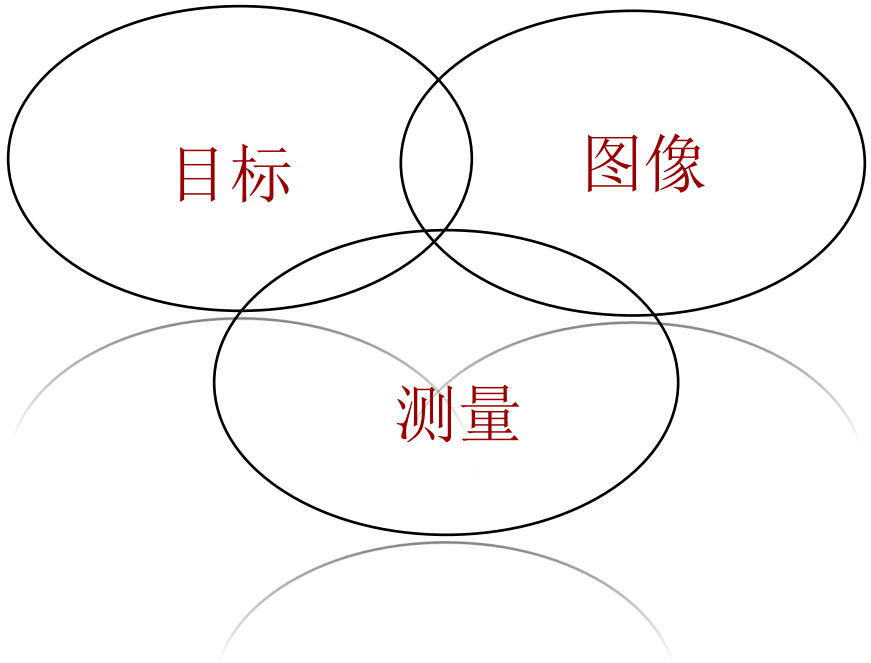
数字图像处理

- 图像处理（Image Processing）是对图像进行分析、加工和处理，以改善图像的视觉效果或从中获取有用信息的过程。根据抽象程度和处理方法的不同，主要可分为以下三个层次：

图像处理级别	处理的抽象程度描述	示例技术/方法
低级处理	基本图像处理，改善视觉效果	噪声降低、对比度增强、锐化
中级处理	提取图像特征	边缘检测、图像分割、特征提取
高级处理	图像分析和理解	图像识别、图像解释、场景理解

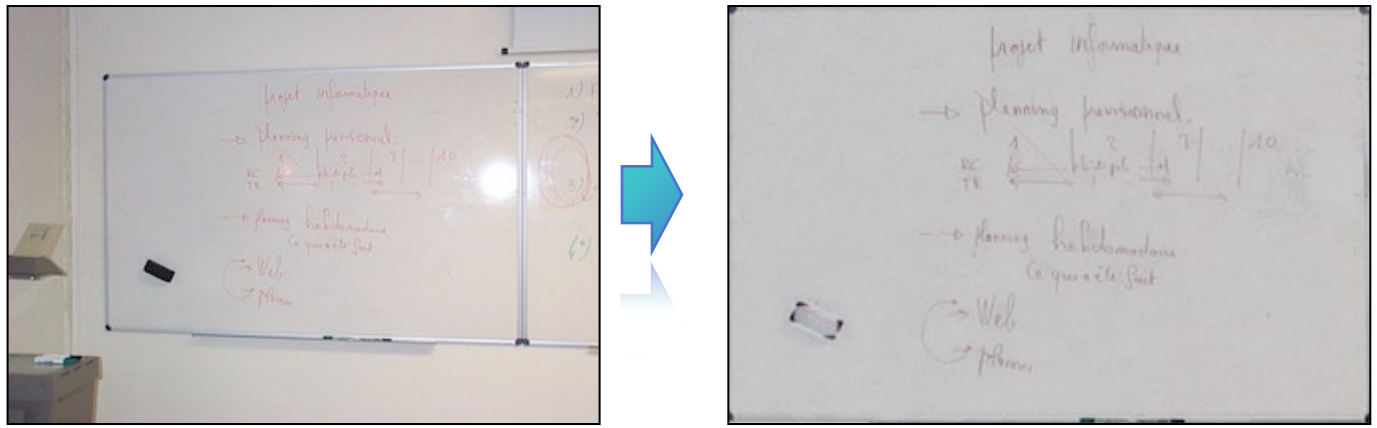
图像处理

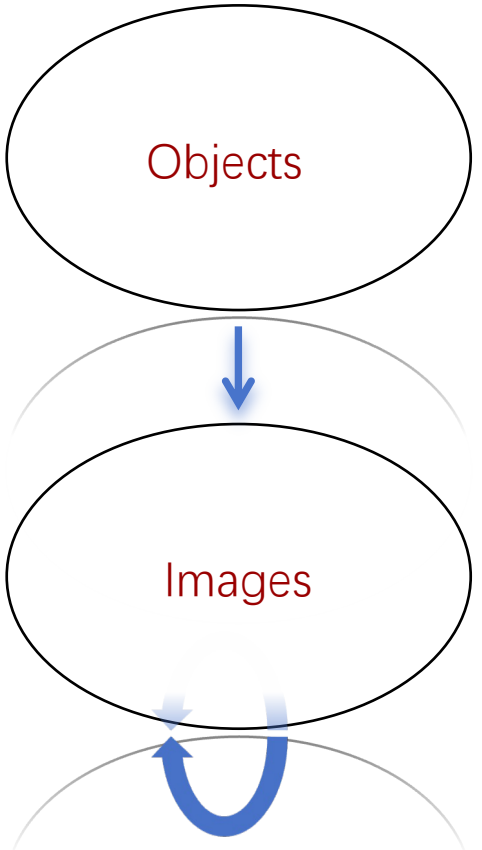
- 这四个术语经常一起用，有些有时会混淆
  - 图像处理
  - 图像分析
  - 计算机视觉
  - 计算机图形学
- 它们都共享表示法、基础数学和一些算法
- 它们的目标非常不一致



传统图像处理

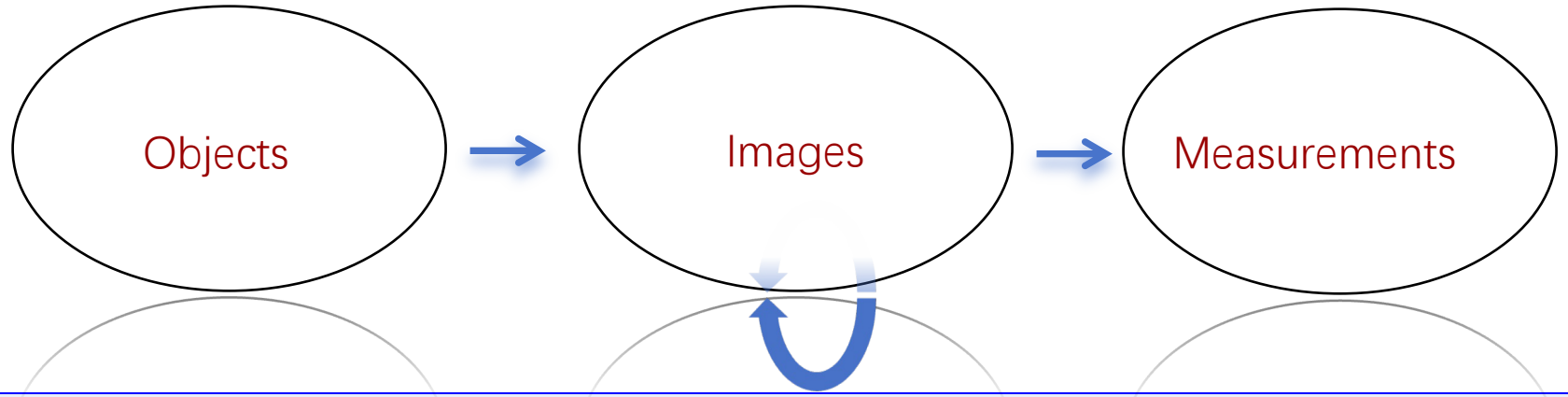
- 输入图像, 输出图像
- 关键信息更容易看到/提取出来
- 更美观



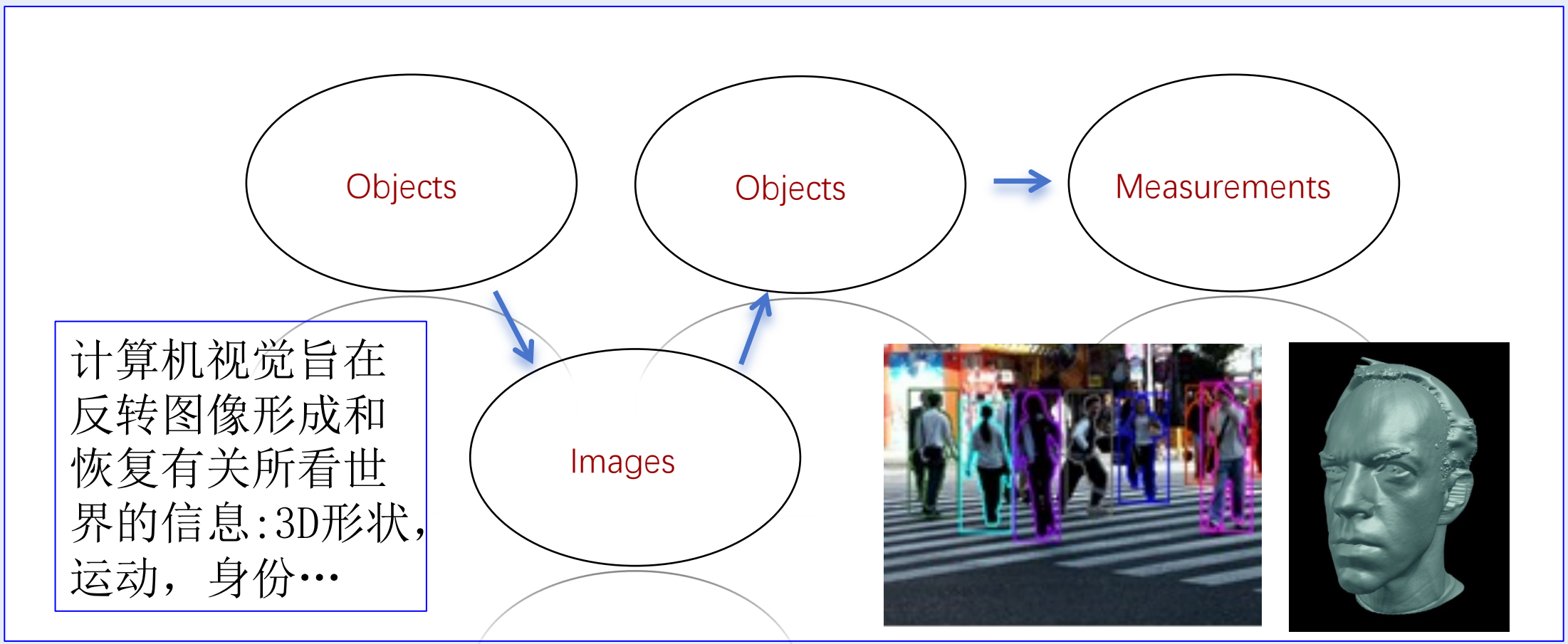


图像分析vs 图像处理

- 图像分析对图像进行定量分析：
  - 图像采集受到限制，因此图像测量是一些真实世界值的代理
  - 介于图像处理和计算机视觉之间
  - 深度图像处理覆盖了图像分析中广泛使用的方法



图像处理不是计算机视觉

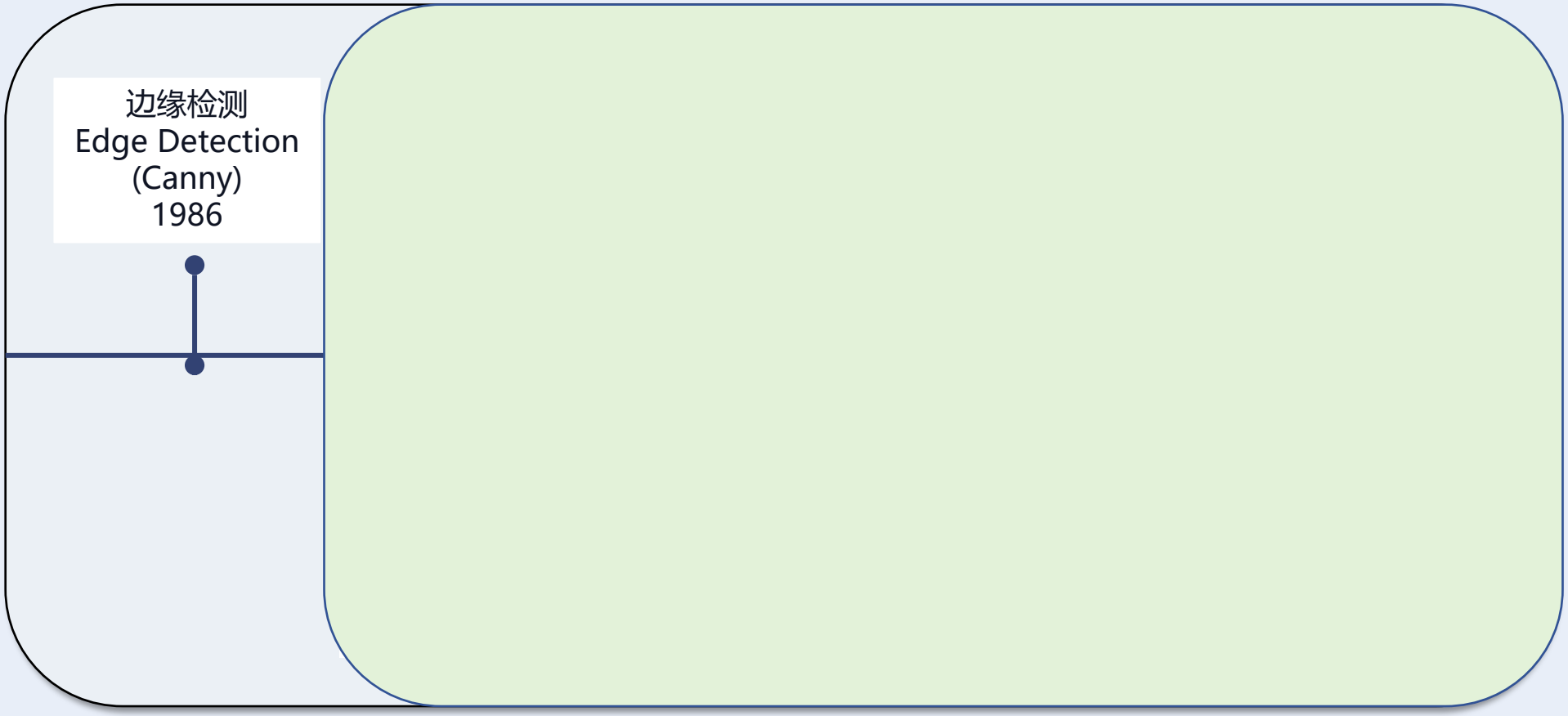


概念比较及交叉融合

概念	描述	主要内容	应用领域
图像处理 (Image Processing)	对图像进行各种操作以改善其质量或提取所需信息的过程	– 图像增强和复原， 图像压缩，图像分割， 图像识别和匹配等	– 医学影像分析，卫星遥感，指纹识别， 安全监控
图像分析 (Image Analysis)	利用数学模型和图像处理技术分析图像，以提取图像中的有用信息	– 特征提取， 目标检测，场景识别， 语义理解等	– 模式识别，遥感图像处理
计算机视觉 (Computer Vision)	赋予计算机从图像或视频中获取、处理、理解和分析信息的能力	– 目标检测，目标跟踪，三维重建，运动分析，场景理解等	– 自动驾驶，机器人导航，安防监控， 虚拟现实/增强现实
计算机图形学 (Computer Graphics)	利用计算机技术生成、处理和展示图形和图像的科学	– 建模，渲染，动画， 人机交互，虚拟现实等	– 视频游戏，影视特效，虚拟人

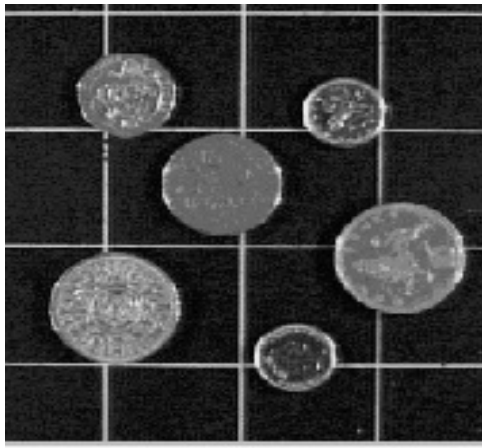
Lecture 13 Contents

图像处理发展里程碑1-边缘检测

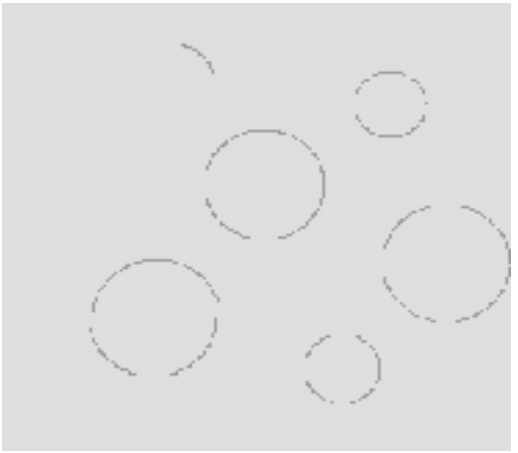


边缘检测

边缘检测算法是计算机视觉领域中一种常用的图像处理技术，用于检测图像中的边缘信息。边缘通常指的是图像中灰度级发生突变的区域，这些区域通常表示物体的轮廓或对象的边界。



6 coins



边缘检测

边缘检测算法	描述	特点
Sobel算子	结合高斯平滑和微分求导的边缘检测算法	简单、快速，可检测水平和垂直边缘，对斜向边缘检测效果较差
Prewitt算子	类似于Sobel算子的边缘检测算法	可检测水平、垂直和斜向边缘，但斜向边缘检测精度可能较低
Roberts算子	计算图像像素点与其对角线方向上的邻域像素点差异	对具有陡峭的低噪声图像效果较好，但定位准确性较差
Canny边缘检测算法	包含高斯滤波、梯度计算、非极大值抑制和双阈值处理	能检测到真正的弱边缘，同时抑制噪声产生的假边缘，经典算法
基于深度学习的边缘检测	利用卷积神经网络等深度学习模型提取边缘特征	具有更高的检测精度和更广泛的应用场景，需要训练数据

边缘检测（Canny）

- John F. Canny在1986年提出的一个多级边缘检测算法—Canny边缘检测算子是一个广泛使用的边缘检测算法，它遵循了Canny提出的三个边缘检测准则：
  - 低错误率：检测到的边缘点应尽可能接近实际边缘中心
  - 高定位精度：检测到的边缘点应准确地定位在边缘中心
  - 单一响应：对于单一的边缘，检测器仅应返回一个响应

Canny边缘检测的流程

原始图像

↓

高斯卷积平滑，以减少噪声和细节对边缘检测的影响

↓

应用微分算子来计算图像像素点在x和y方向上的梯度幅值和方向

↓

在图像梯度中，通过非极大值抑制来寻找局部最大值，即边缘的峰值

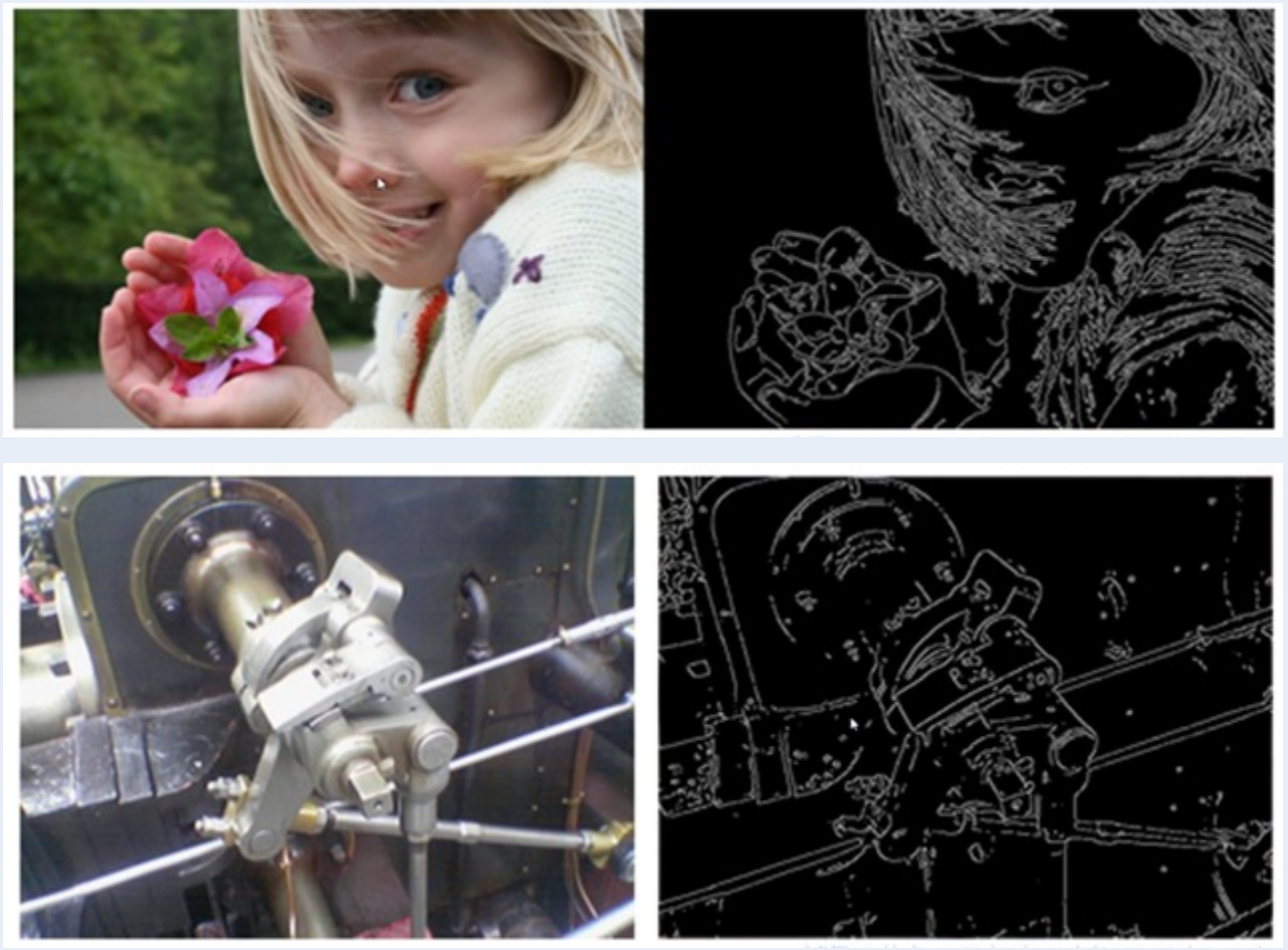
↓

使用双阈值（高阈值和低阈值）来区分强边缘和弱边缘
强边缘被直接接受为边缘线条，而弱边缘则需要与强边缘相连才能被接受
这个过程有助于连接断裂的边缘，并减少噪声引起的假边缘

↓

二值化的边缘映射图像

Canny边缘检测示例



图像处理发展里程碑2-图像压缩

图像压缩

- 图像压缩是指以较少的比特有损或无损地表示原来的像素矩阵的技术，也称图像编码。图像压缩主要是为了减少表示数字图像时所需的数据量。
  - 图像数据之所以能被压缩，是因为其中存在着冗余。这种冗余主要表现为：图像中相邻像素间的相关性引起的空间冗余，不同彩色平面或频谱带的相关性引起的频谱冗余等。数据压缩的目的就是通过去除这些数据冗余来减少表示数据所需的比特数。
  - 图像压缩可以分为有损压缩和无损压缩两种。无损压缩方法适用于需要保持图像完整性的情况。有损压缩方法则非常适合于自然的图像，如将色彩空间化减到图像中常用的颜色，色度抽样（利用人眼对于亮度变化的敏感性远大于颜色变化），以及变换编码）等。这些方法可能会带来一些微小的图像损失。

图像压缩与语音压缩

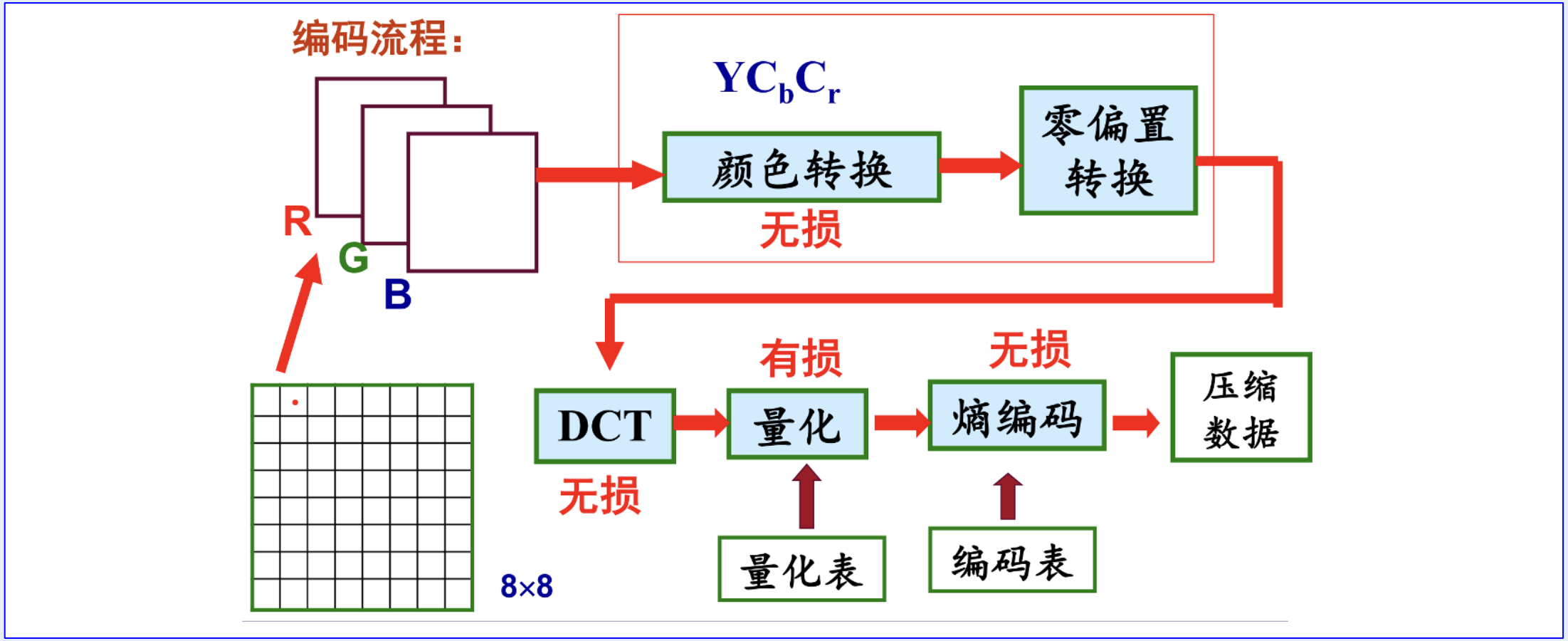
	图像压缩	语音压缩
定义	使用尽可能少的比特数代表图像或图像中所包含的信息	提高通信网中的信息传输效率及实现语音的高效存储，对编码后的数字语音进行压缩
目的	减少表示数字图像时所需的数据量，以便于存储、传输和处理	降低语音信号的编码比特率，以满足窄带信道低码率传输的要求及实现语音的高效存储
压缩类型	无损压缩（如PNG、GIF）、 有损压缩（如JPEG）	根据压缩率的不同，可以分为 高、中、低速率编码
应用场景	社交媒体分享、网页设计开发、 医疗图像存档等	移动通信、卫星通信、多媒体技术、 IP电话通信等
压缩依据	图像数据中的冗余，如空间冗余 和频谱冗余	语音信号中的冗余和人类的听觉感知机理

JPEG

什么是JPEG文件?

JPEG（Joint Photographic Experts Group）即联合图像专家组，是用于连续色调静态图像压缩的一种标准，文件后缀名为.jpg或.jpeg，是最常用的图像文件格式。其主要是采用预测编码（DPCM）、离散余弦变换（DCT）以及熵编码的联合编码方式，以去除冗余的图像和彩色数据，属于有损压缩格式，它能够将图像压缩在很小的储存空间，一定程度上会造成图像数据的损伤。尤其是使用过高的压缩比例，将使最终解压缩后恢复的图像质量降低，如果追求高品质图像，则不宜采用过高的压缩比例。

JPEG 图像压缩流程



JPEG压缩实例

245,760 bytes

69,632 bytes

5,951 bytes

CS 330 MIP – Lecture 13

音频信息处理 4 + 图像信息处理 1

Audio Information Processing 4 + Image Information Processing 1

Jimmy Liu 刘江

2025-05-14