人工智能基础与进阶

感知信息处理

上海交通大学

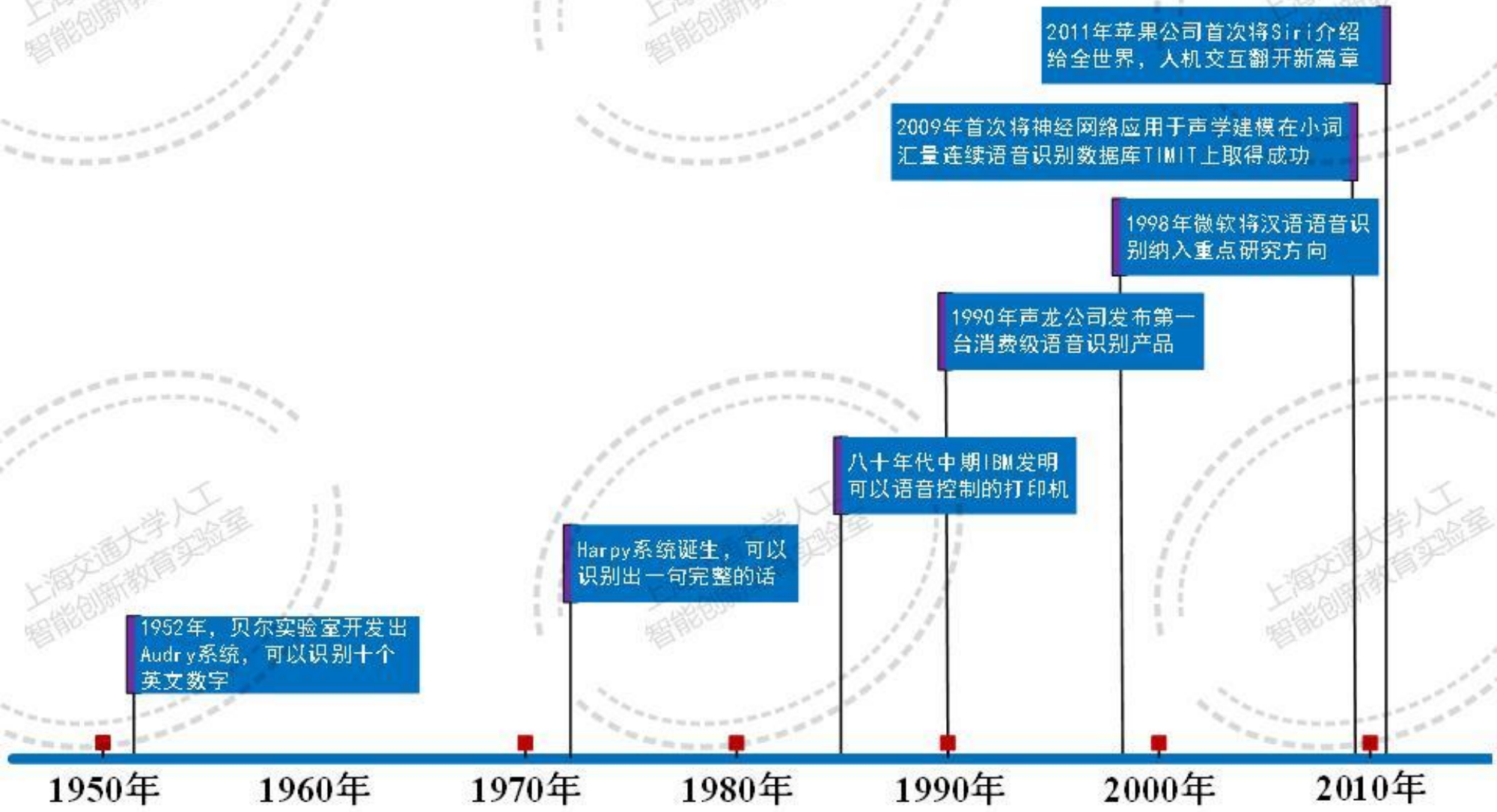
目录 content

第一节　语音信息处理基础认识

语音识别发展历史

- 1952年，贝尔实验室实现了第一个可以识别十个英文数字的语音识别系统——Audry系统
- Harpy系统诞生后，可以识别出一句完整的话
- 20世纪80年代中期，IBM发明了一个可以用语音控制的打字机。

语音识别发展历史



语音识别发展历史

- 1990年，声龙公司发布了第一款消费级别的语音识别产品。
- 1998年，微软在北京成立亚洲研究院，将汉语语音识别纳入重点研究方向。
- 2009年，深度神经网络首次应用于语音的声学建模并且获得一定成功
- 2011年10月，苹果公司揭开了个人手机助理Siri的神秘面纱，人机交互领域翻开了新的篇章。

几种常见语音助手

- 苹果公司Siri
- 微软公司Cortana
- 亚马逊公司Alexa
- 谷歌公司Google Assistant
- 科大讯飞公司灵犀

灵犀

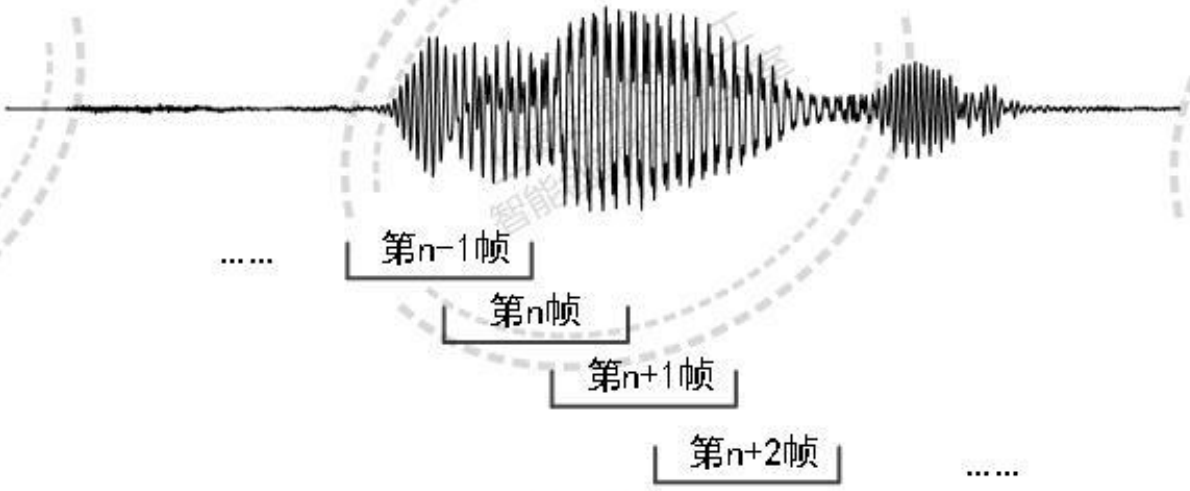
听话的语音小秘书

语音信息时域特征

- 时域波形图横坐标是时间，纵坐标是振幅
- 时域上的语音信息特征直观，有明确的物理意义

数字1-10语音波形

语音信息时域特征



语音信号的分帧

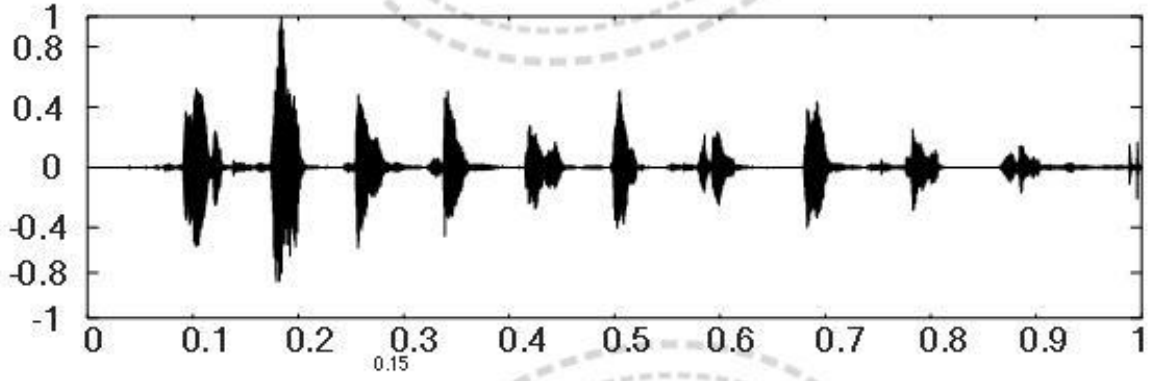
- 默认语音信号在10ms到30ms近似内不变
- 常将10ms到30ms内语音切成一帧
- 相邻两帧有重叠部分，重叠部分称为帧移
- 分帧过程如左图所示

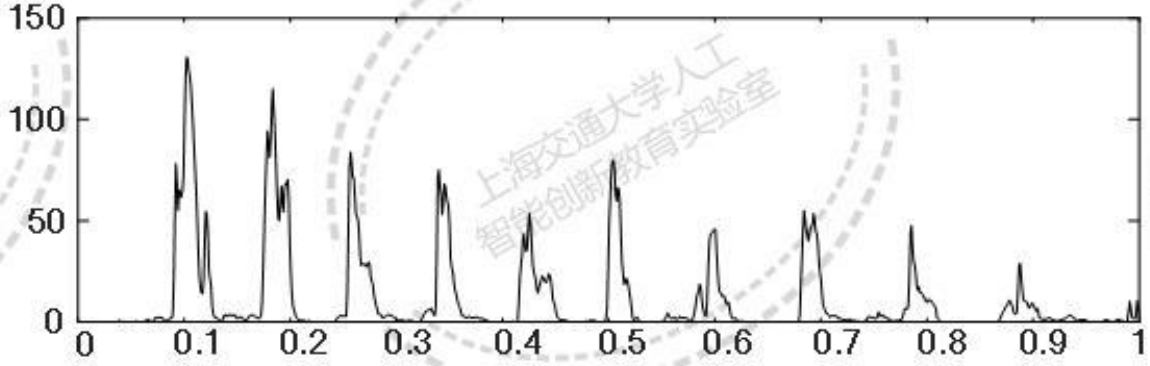
语音信息时域特征

音量：音量代表声音的强度，可以由一帧内信号振幅的大小来衡量

度量方法一：用每一帧内信号幅值的绝对值总和来度量音量大小。公式为：

$$volume=\sum_{i=1}^{n}|s_i|$$



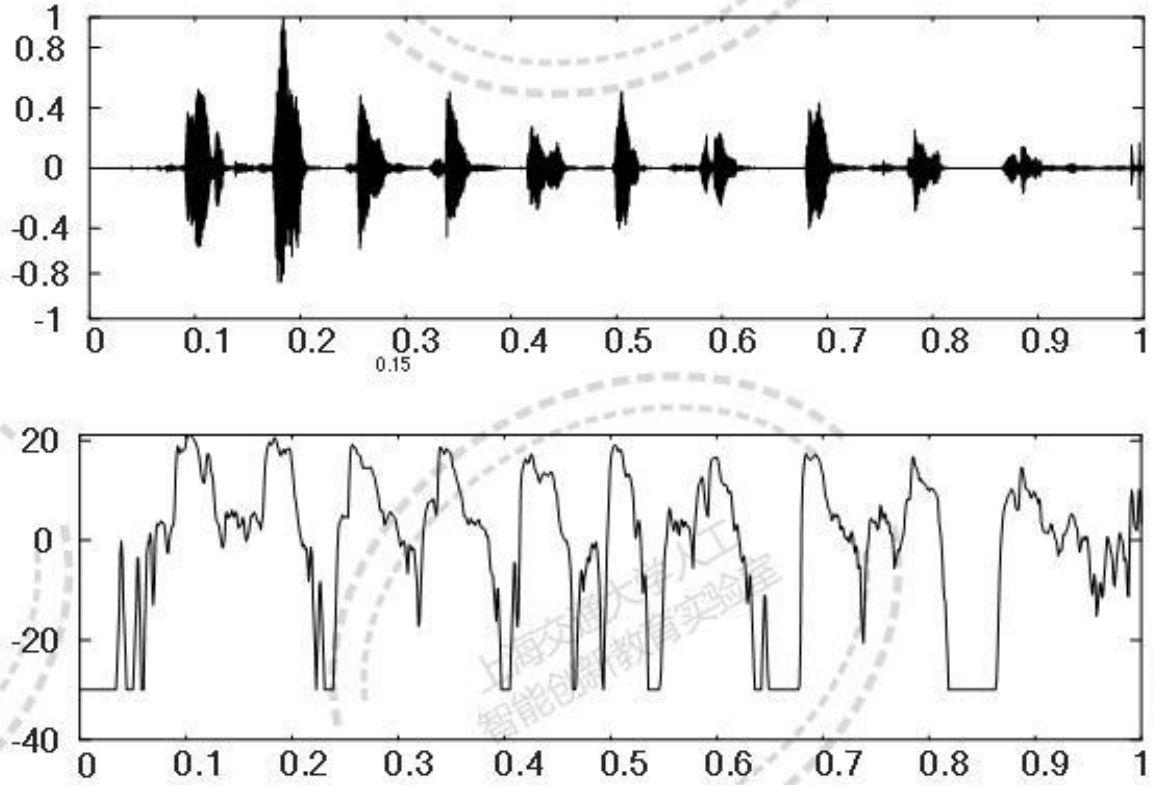


语音信息时域特征

音量：音量代表声音的强度，可以由一帧内信号振幅的大小来衡量

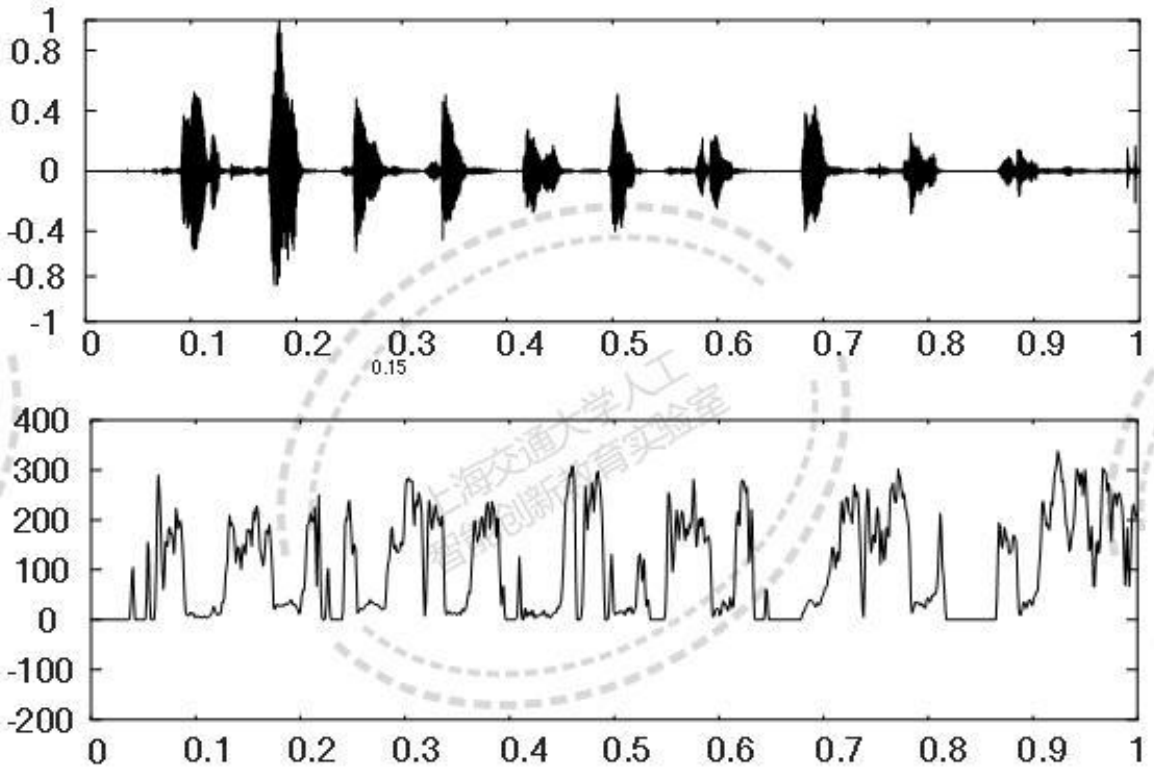
度量方法二：对每一帧信号的幅值平方和的常数对数取10倍。公式为：

$$volume=10\times\log_{10}\sum_{i=1}^{n}s_i^2$$



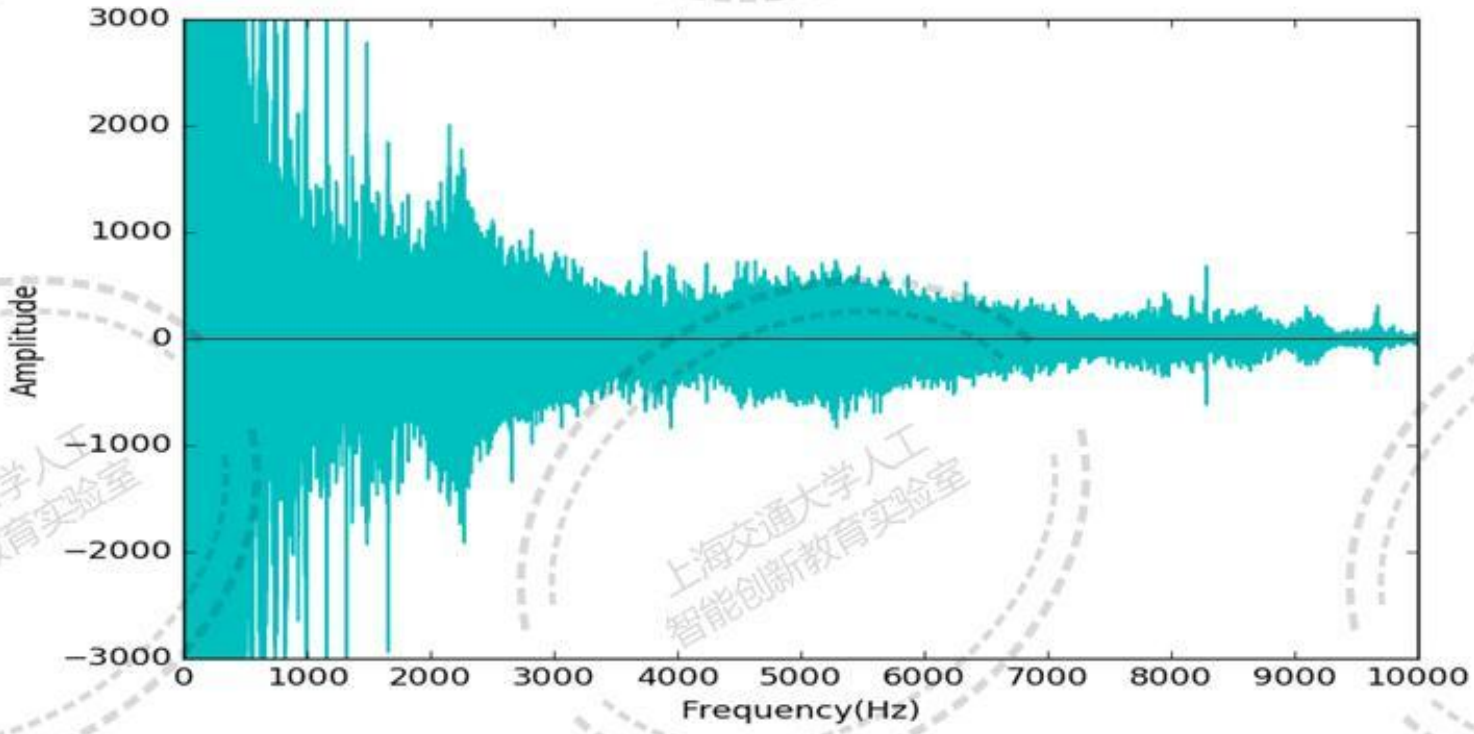
语音信息时域特征

- **清音**：发音时声带不振动发出的音
- **浊音**：发音时声带要振动才能发出的音
- **过零率（ZCR）**：指每一帧语音信号通过零点的次数。
  - 清音和环境噪音的过零率都大于浊音
  - 无法通过过零率区分清音和环境噪音
  - 过零率常用来进行端点检测



语音信息频域特征

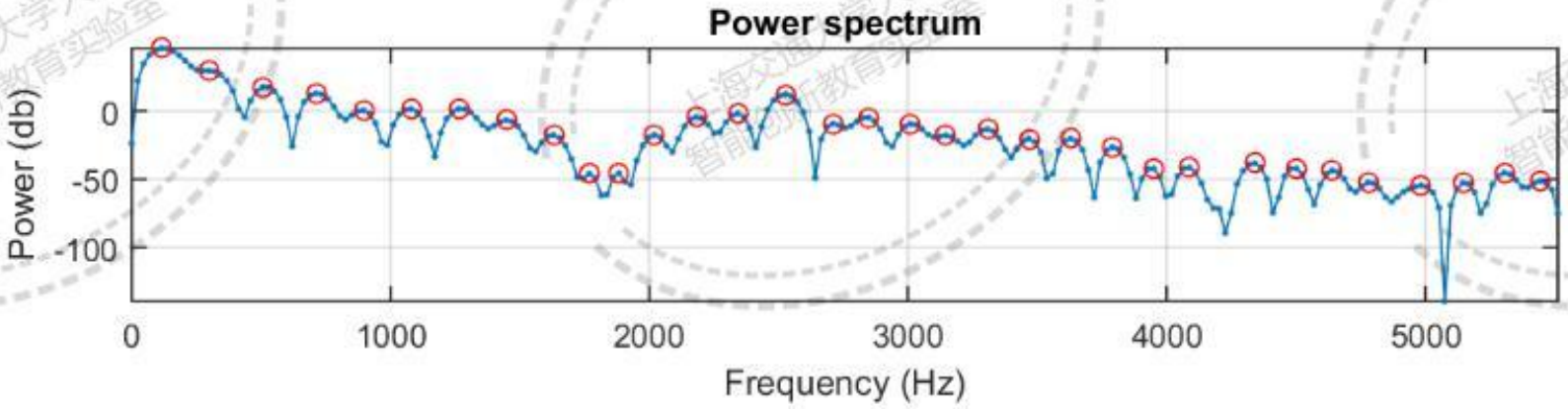
- 语音信息频域波形图的横坐标表示频率，纵坐标表示频谱幅度
- 频域波形图相对时域波形图来说更加简练，便于进一步剖析问题



语音信息频域特征

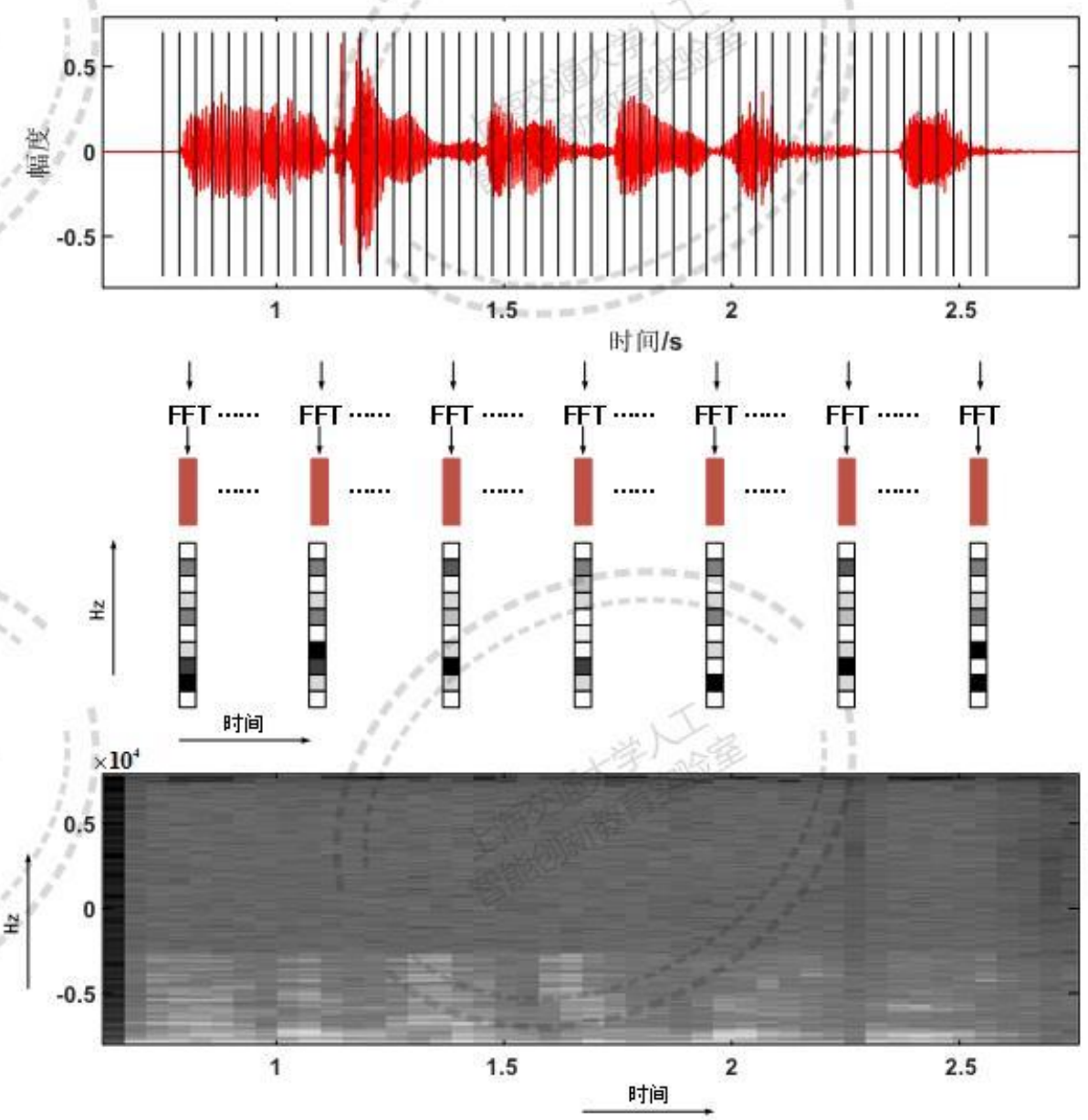
- **音调**：音调表示人听到的调子的高低
- 声音信号频率越高，音调就越高；声音信号的频率越低，音调就越低
- 音调的高低由频率大小决定
- 频谱图可以直接描述声音的音调特征

语音信息频域特征



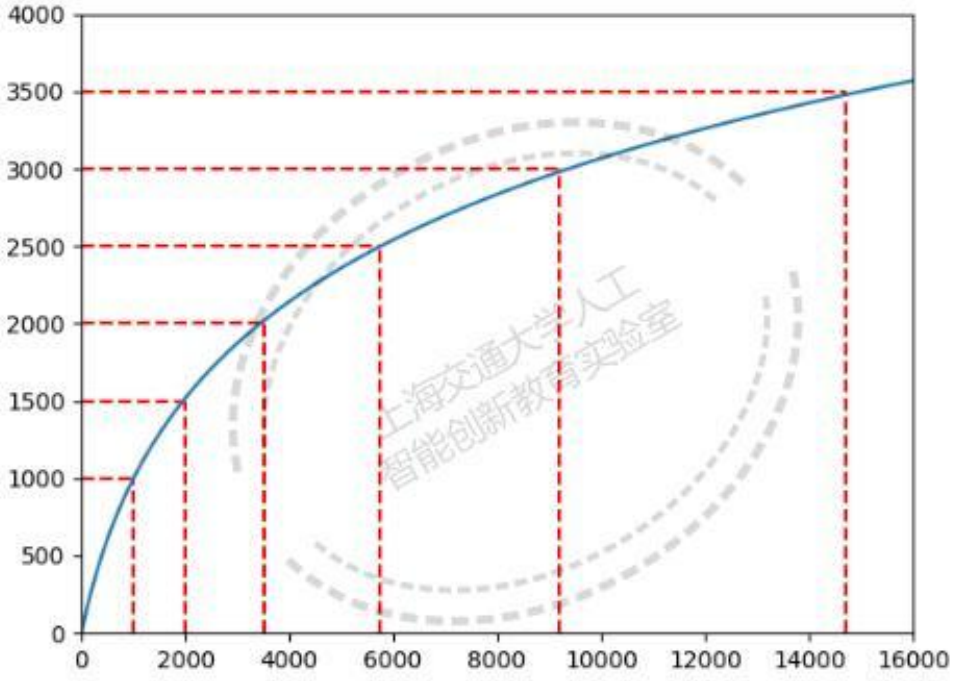
- **共振峰**：声音频谱上能量相对集中的一些区域
- 共振峰是音质的决定因素之一
- 反映了声道的物理特征，代表发音信息的直接来源
- 是语音信号处理中非常重要的特征参数
- 上图红色区域是语音信息频谱中的共振峰

语音信息频域特征



语音信息倒谱特征

- 梅尔频率倒谱系数——MFCC
- MFCC是经典的声学特征之一
- MFCC的优点在于：大大降低了特征维数；粗略刻画频谱形状，描述不同频率下声音能量的高低；表达出共振峰的特性

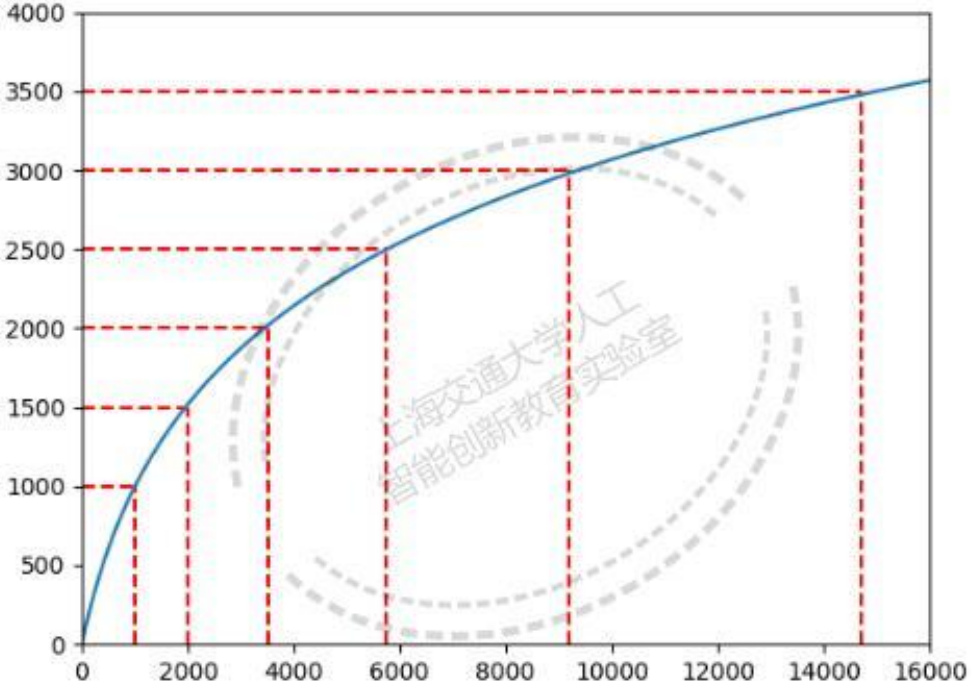


语音信息倒谱特征

- 梅尔倒谱系数(MFCC)提取的第一步是将普通频率转换成梅尔频率
- 梅尔频率与普通频率转换公式：

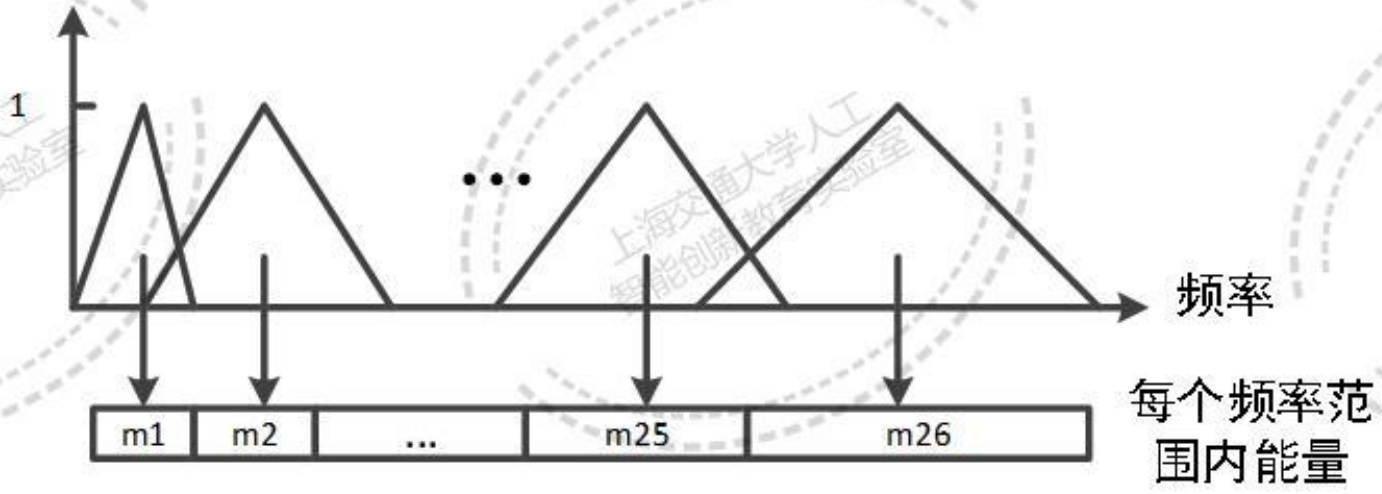
$$mel(f) = 1125 \times \ln(1 + f / 700)$$

- 梅尔频率在低频部分的分辨率较高，高频部分分辨率较低，这一点与人耳类似

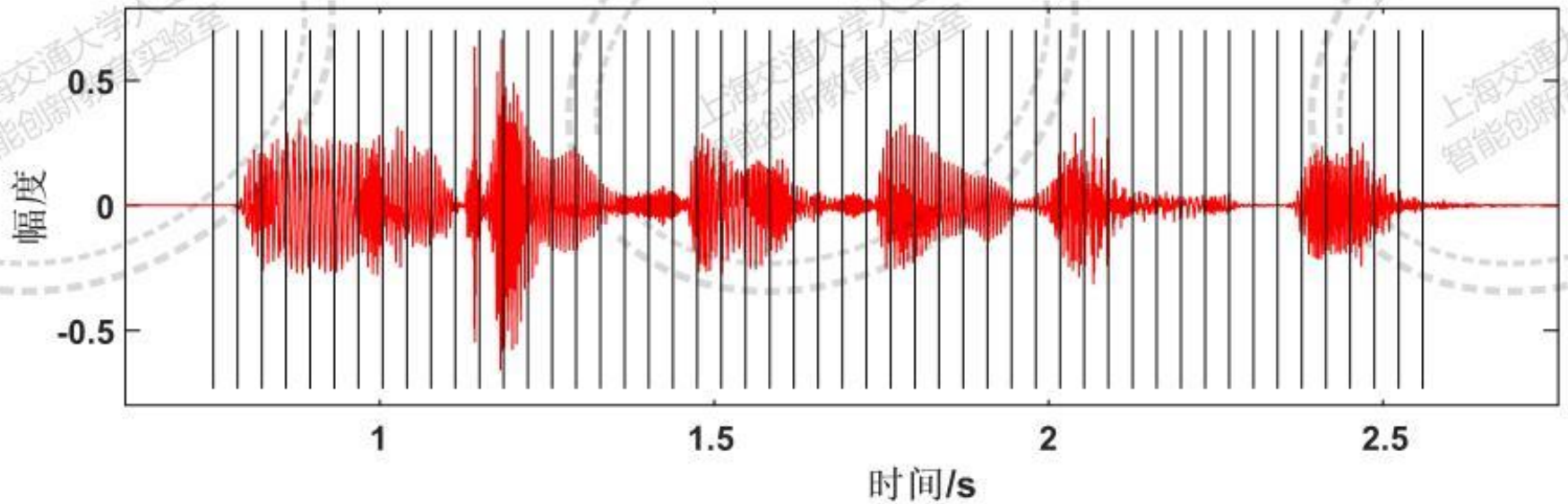


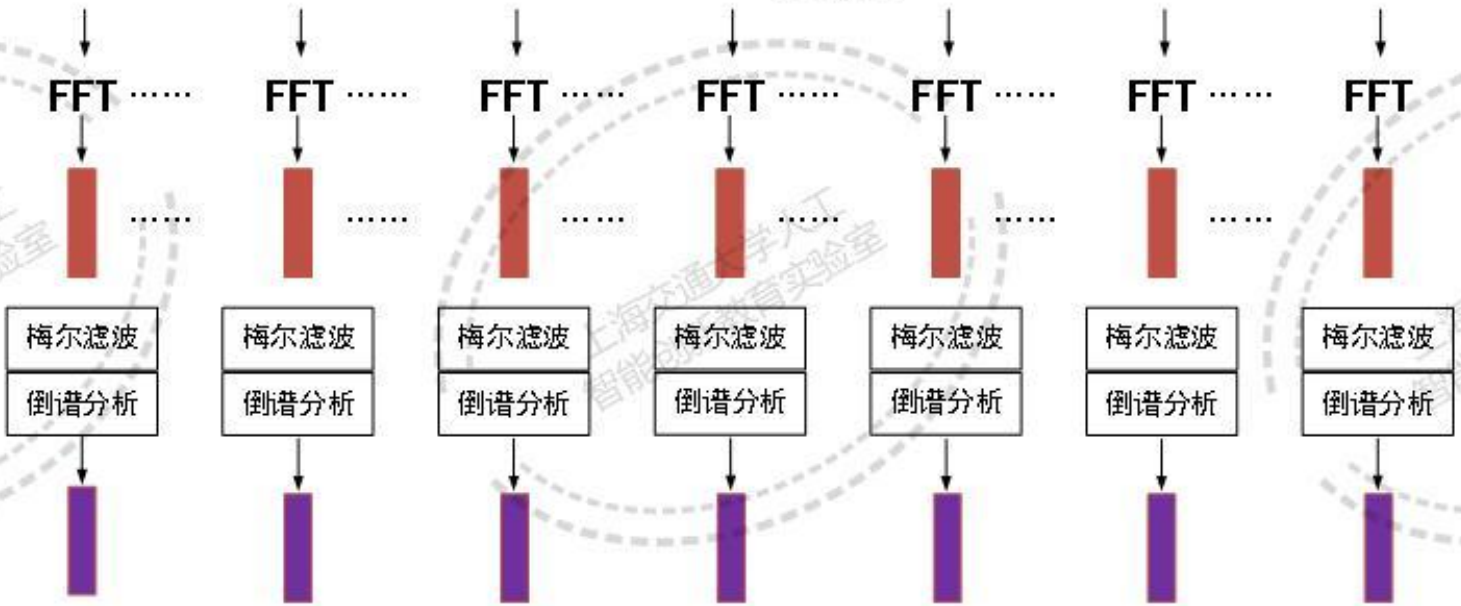
语音信息倒谱特征

- 将频率范围分为26个区间，对每个区间内的频谱求平均值，该平均值代表每个频率范围内声音能量的大小
- 这26个均值表示一个26维的特征
- 将这个26维的特征经过一系列数学运算后得到一个13维的特征
- 真正的梅尔倒谱系数特征是这个13维的特征
- MFCC在保留重要音频信号特点的同时，尽量降低了特征的维数



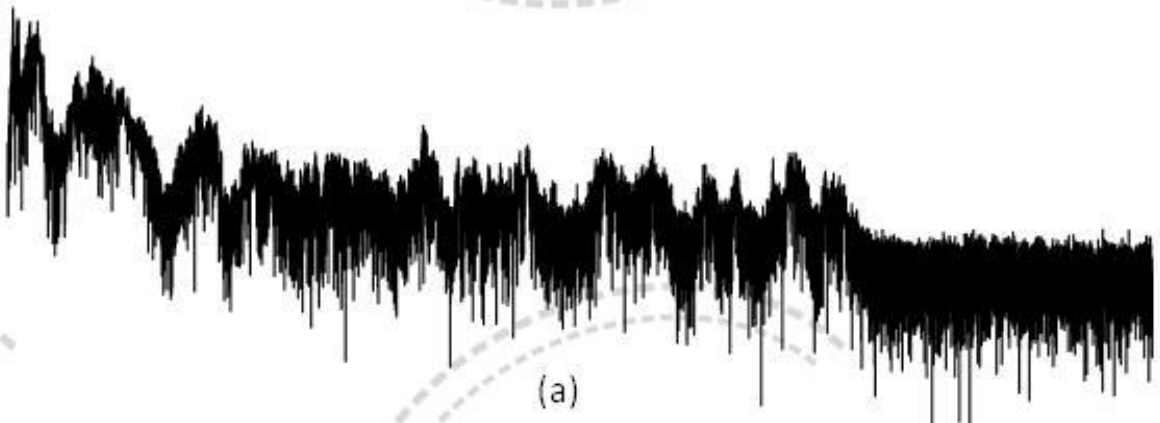
语音信息倒谱特征





深度学习方法提取语音特征

- 假设上图(a)图中所提取出的MFCC特征为(b)图中长向量
- (b)图中短向量表示用来提取音频特征的卷积核
- 卷积运算的结果可以理解为提取出来的语音特征



(a)

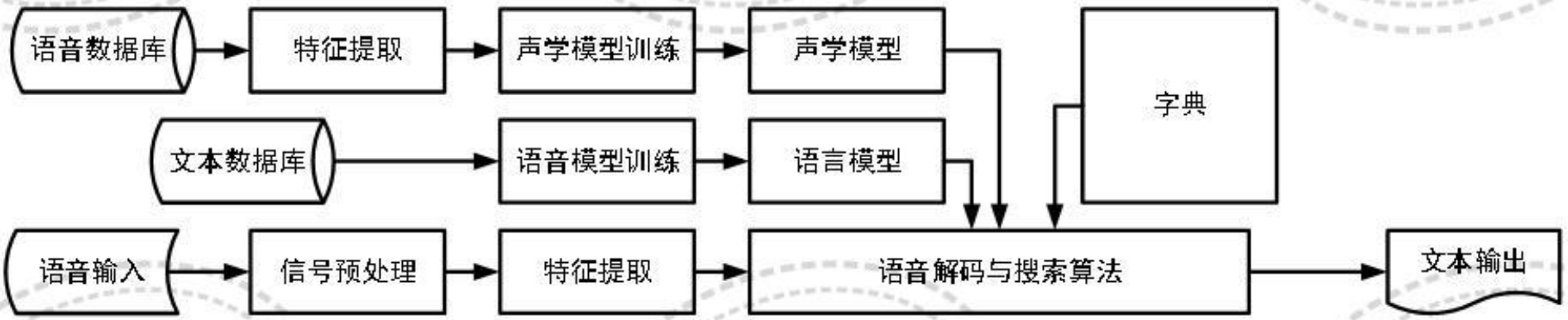
(b)

长向量	3	2	5	1	4
短向量	1	1	2		
结果向量	15	9	14		

深度学习方法提取语音特征

- 用深度学习方法提取到的语音特征更加强大
- 神经网络的应用大大提高了语音识别的准确率

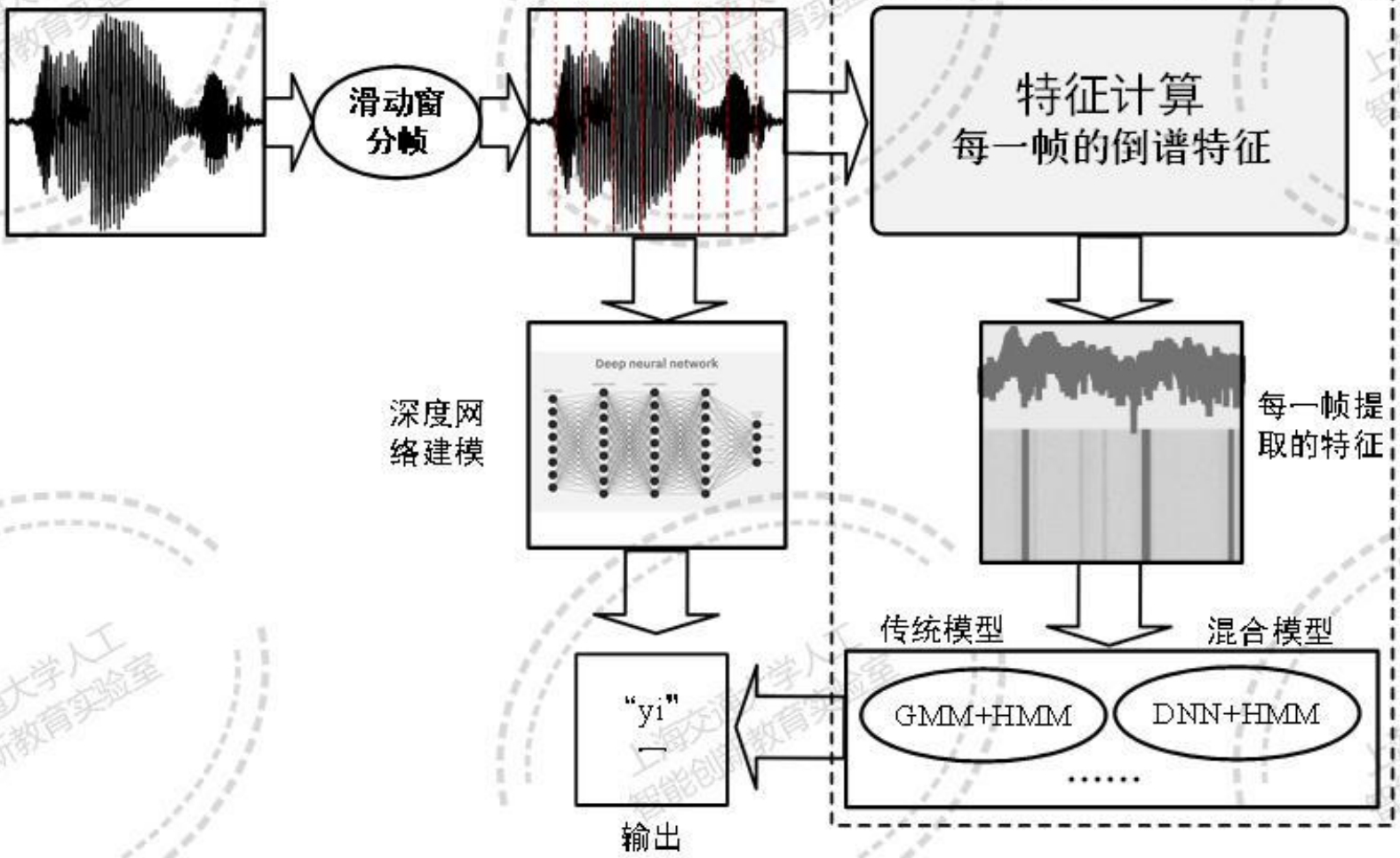
语音信识别系统框架



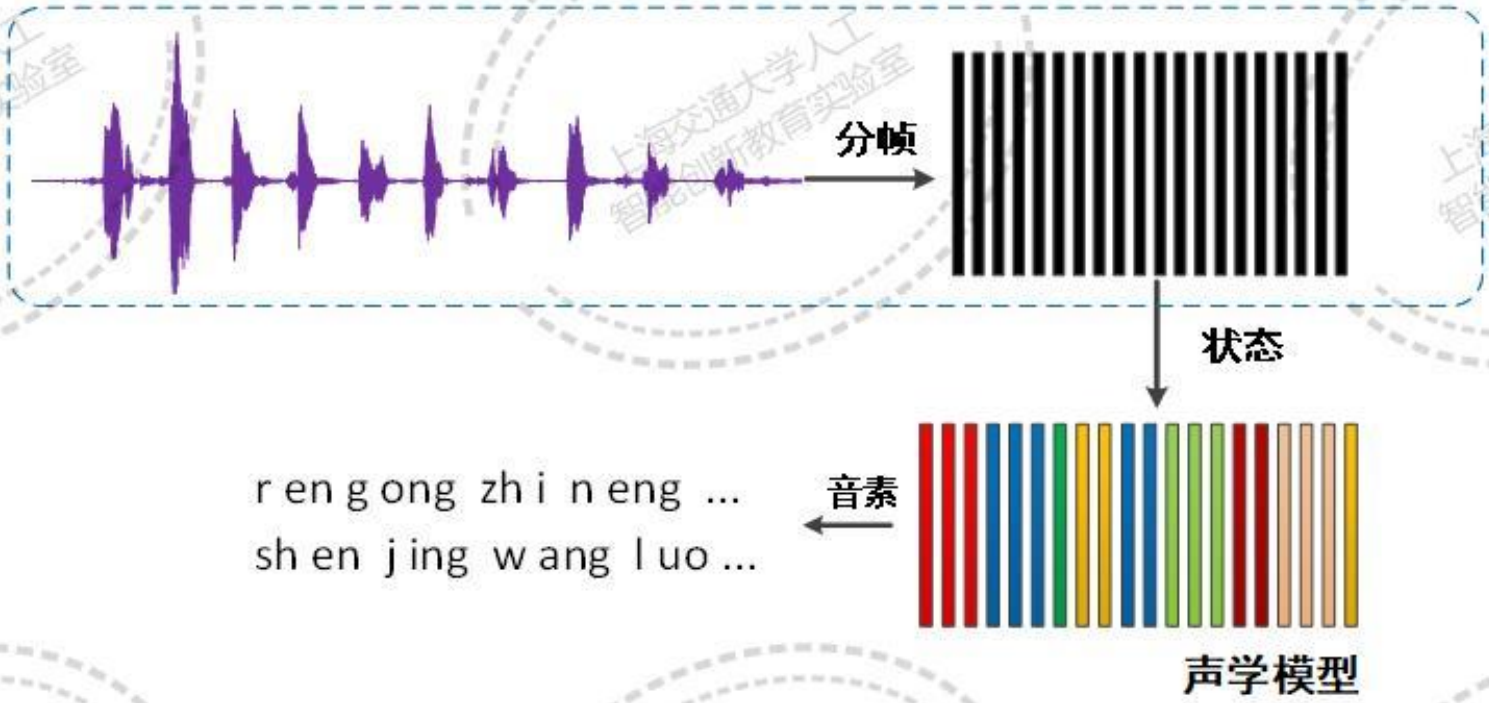
语音信识别系统框架

- **信号预处理**：分帧、静音切除，尽可能降低环境噪音对后续步骤的干扰
- **特征提取**：从预处理后的声音信号中提取出特征，就是把每一帧的声音波形都变换为一个包含声音信息的多维向量
- **声学模型**：输入的是语音的特征向量，输出的是该特征所对应的音素信息，根据语音数据库中的各种语音进行建模
- **语言模型**：对语音进行建模，通过文本数据库中的文本信息进行训练，得到的是单个字或者词语相互关联的概率
- **字典**：声学模型和语音模型单元之间的映射，类似拼音和汉字的对应，音标和单词的对应
- **解码器**：对于输入的信号，根据以上各个模块提取到的特征进行处理解码，找出能够以最大概率输出该信号的词序列，将音频数据变成文本输出

语音信识别系统框架-声学模型



语音信识别系统框架-声学模型

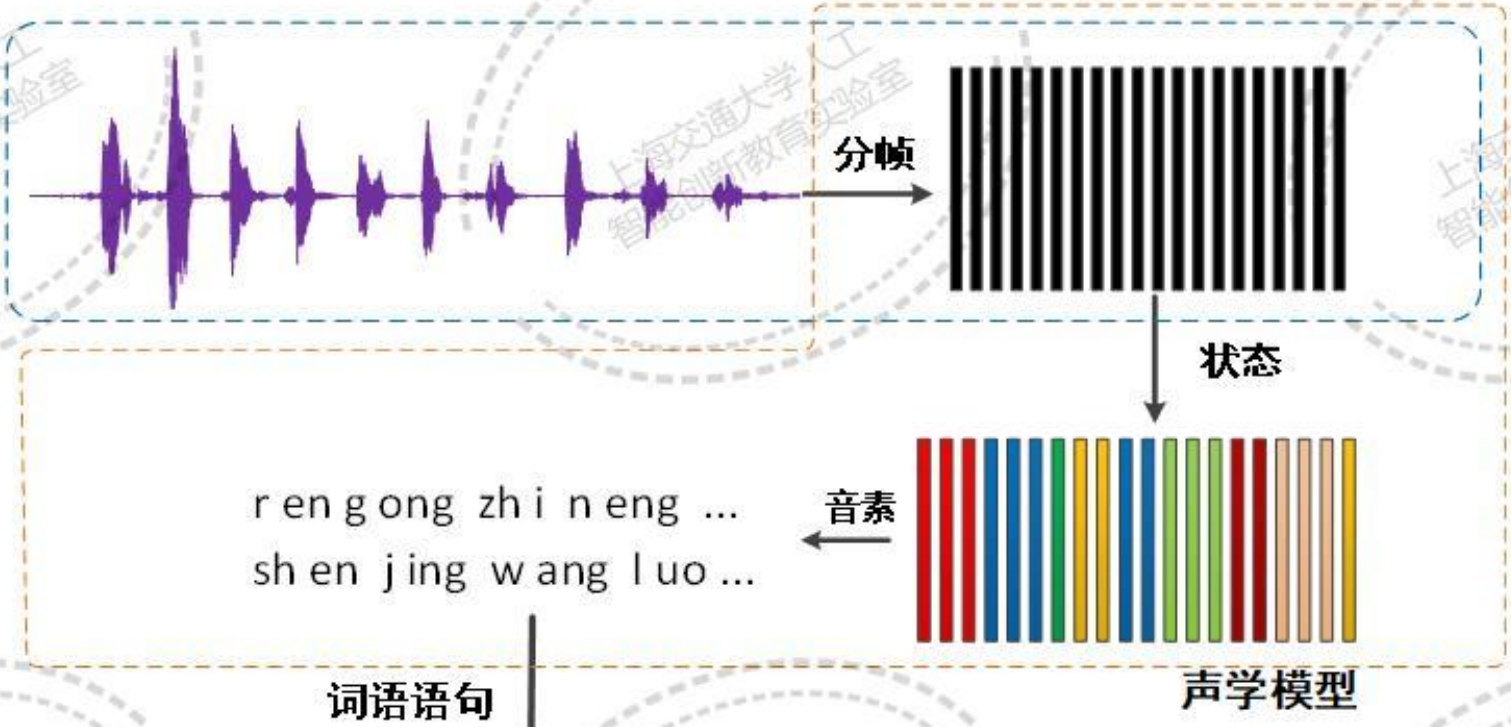


- 语音分帧后，每一帧语音被识别成一个状态
- 音素可以理解成拼音中的声母或韵母
- 三个状态可以组合成一个音素，状态是比音素更加细节的语音单位
- 这部分就是所谓的声学模型

语音信识别系统框架-语言模型

- 语音识别系统中的声学模型可以将一系列语音帧转换成对应的音素。
- 如何将音素转换成文字并且使语句通顺且具有实际意义，是语言模型的任务

语音信识别系统框架


分帧
状态
r en g ong zh i n eng ...
sh en j ing w ang l uo ...
音素
声学模型
词语语句
人 工 智 能 … 神 经 网 络 …
人 工 只 能 … 深 井 网 罗 …
人 公 职 能 … 神 经 网 罗 …
人 工 智 能 … 神 经 网 络 …
语言模型

语音识别系统小结

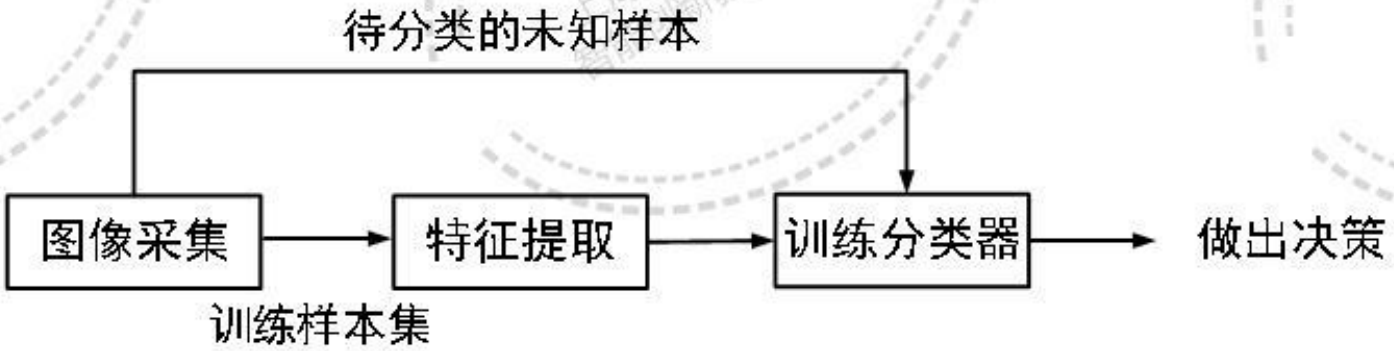
- 对语音信号进行分析和处理，除去冗余信息
- 提取影响语音识别的关键信息和表达语言含义的特征信息
- 紧扣特征信息，用最小单元识别字词
- 按照不同语言的各自语法，依照先后次序识别字词
- 把前后意思当作辅助识别条件
- 按照语义分析，给关键信息划分区间，取出所识别出的字词并连接起来，同时根据语句意思调整句子构成
- 结合语义，分析上下文之间的相互联系，对当前语句进行适当修正，最后输出语音识别结果

第二节　图像信息处理基础认识

人眼识别标志牌

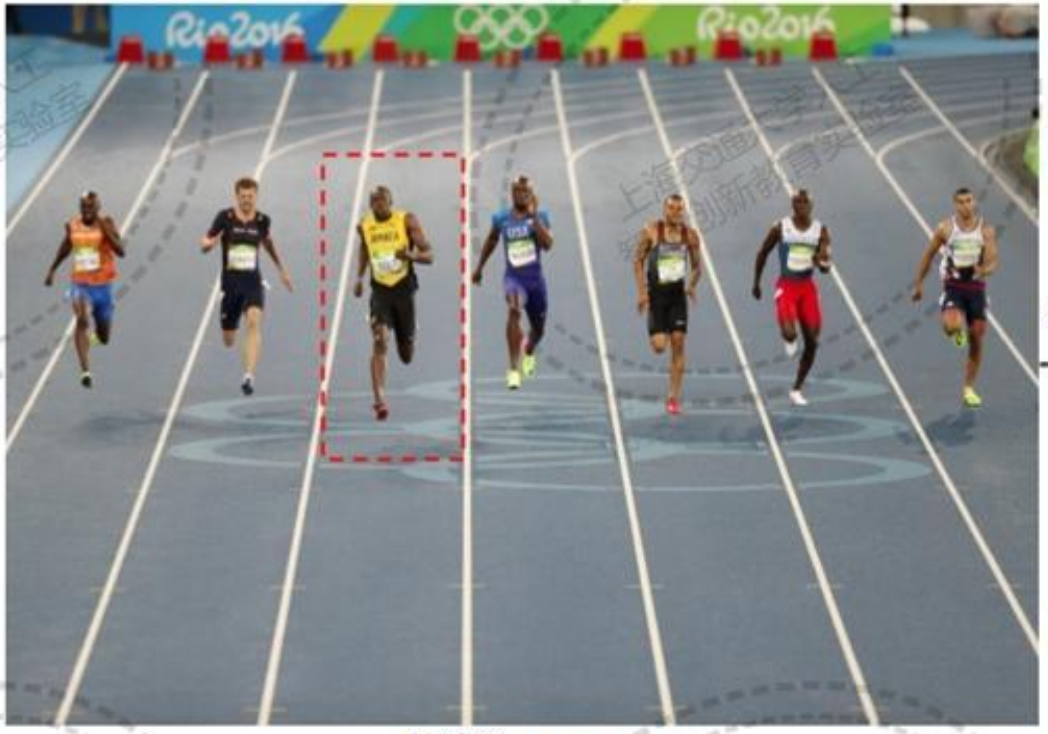
- 找到标志牌上最引人注目的区域
- 观察得到该区域特征
- 将该特征与大脑中之前见过的已知的标志牌特征进行匹配，判断出该标志牌是什么标志牌
- 依据判断结果做出相应决策

计算机识别标志牌



- 图像采集主要负责准备训练数据
- 训练数据由目标所在区域和标注数据组成
- 交通标志牌的标注数据就是，直行标志牌、限速标志牌等

图像获取与预处理



原图像 720 x 475

剪裁

100 x 200

缩放

64 x 128

- 在左图中找出目标人物所在位置
- 将目标人物裁剪出来，是一个大小为100x200的图片块
- 将该图片块缩放成64x128大小的图片块
- 给该图片块标注相应信息，如博尔特

图像采集

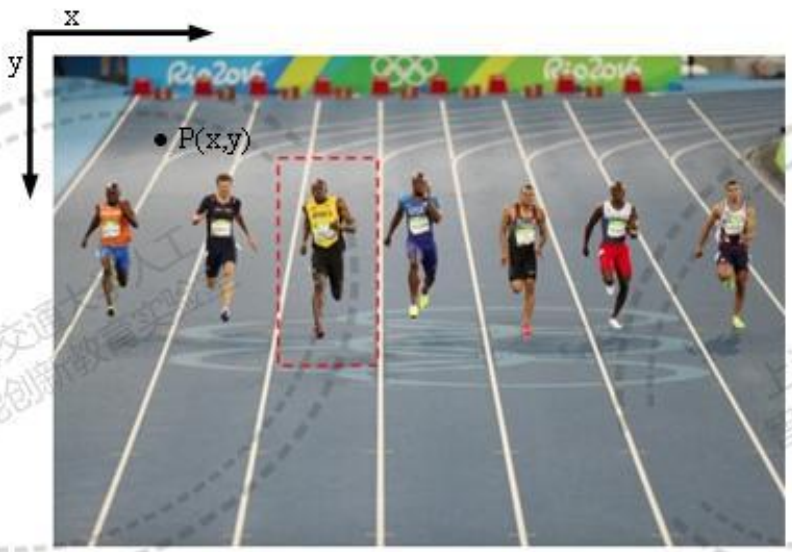
原图像 400 x 300

- 找出原图中标志物所在区域
- 经过裁剪得到168x168大小的标志物
- 将标志物缩放成64x64大小
- 对该64x64大小的标志物进行标注，限速标志

图像梯度直方图特征提取

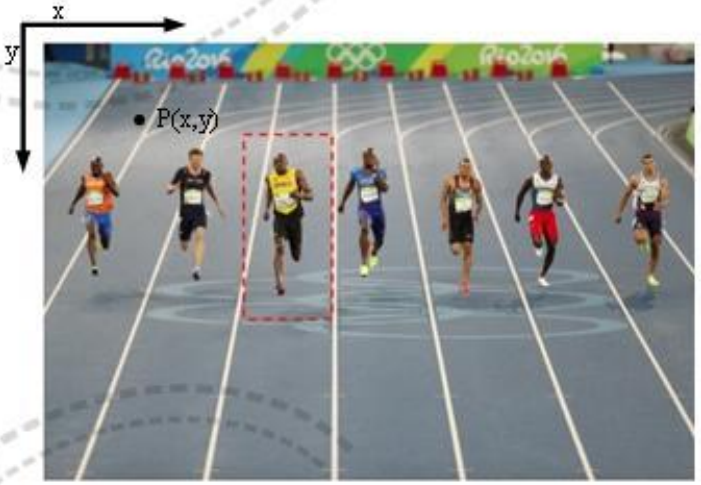
- 图像的坐标表示，P点坐标(x, y)，P点像素值为 I(x, y)

- P点的水平梯度和垂直梯度定义分别如下：



$$G_x(x,y) = I(x+1,y) - I(x-1,y)$$
$$G_y(x,y) = I(x,y+1) - I(x,y-1)$$

图像梯度直方图特征提取

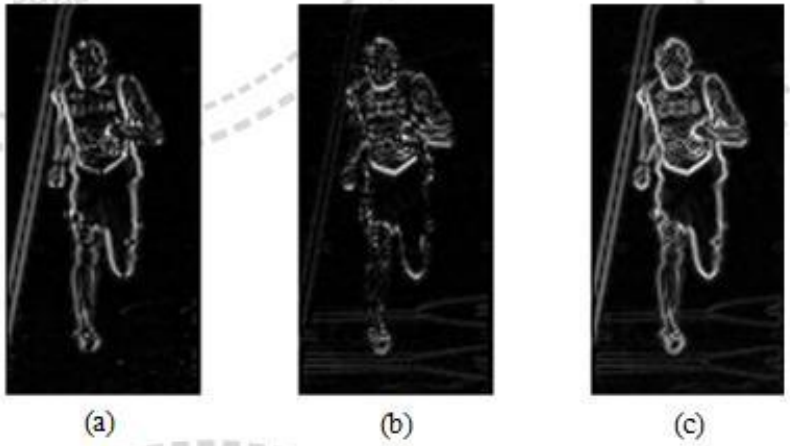


P点梯度幅值和梯度方向分别为：

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}$$

$$\theta(x,y) = \arctan\left(\frac{G_y(x,y)}{G_x(x,y)}\right)$$

图像梯度直方图特征提取



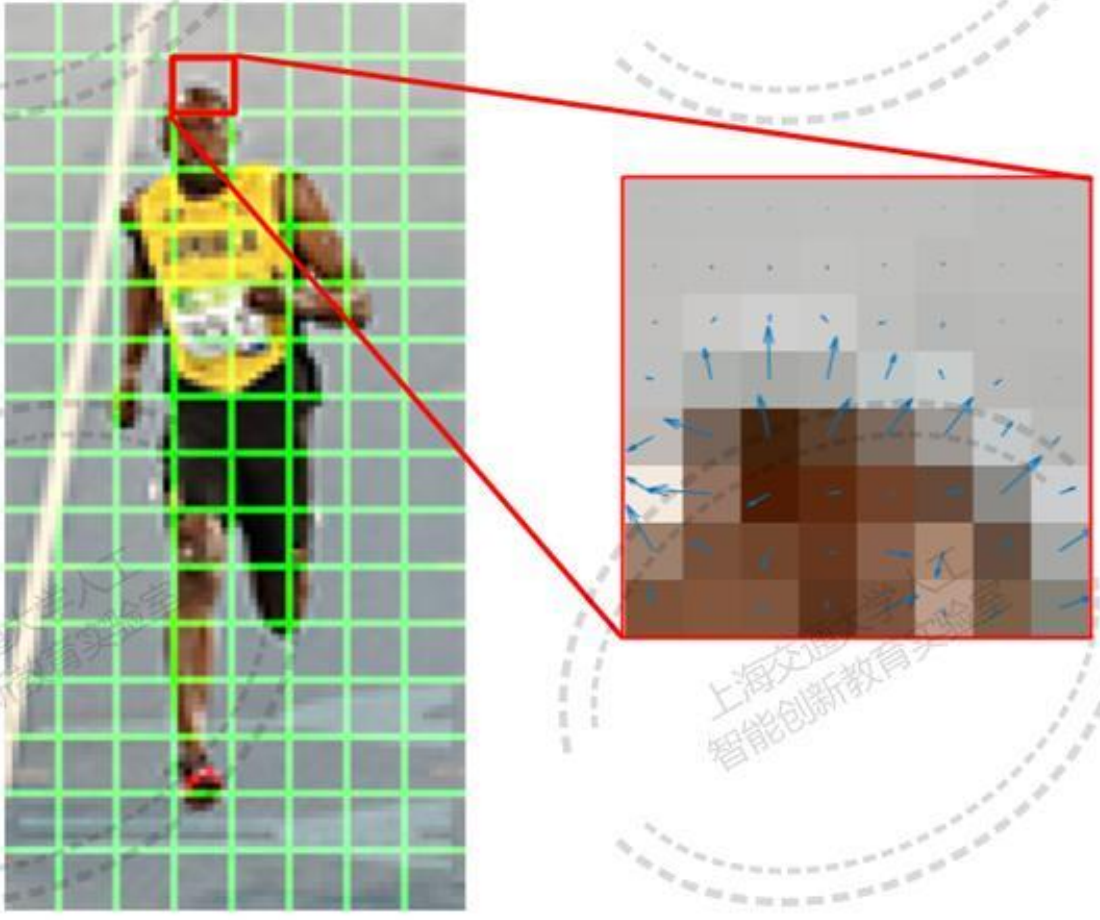
- 图像中每个像素点都有梯度幅值和梯度方向
- 在RGB彩色图像中，分别单独计算R、G、B每个通道上的梯度幅值和梯度方向
- 取像素点最大梯度幅值，以及该幅值对应的梯度方向
- 梯度图像是单通道的灰度图像
- 从左到右依次为图像的水平梯度图像、垂直梯度图像和梯度幅值图像

图像梯度直方图特征提取

- 选取8x8像素大小的图像小块对图像进行分块
- 64x128的博尔特图像可以分为128个8x8的小区域
- 64x64的标志牌可以分成64个小区域

图像梯度直方图特征提取

每个8x8的小图像块中，每个像素点都有对应的梯度幅值和梯度方向



2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

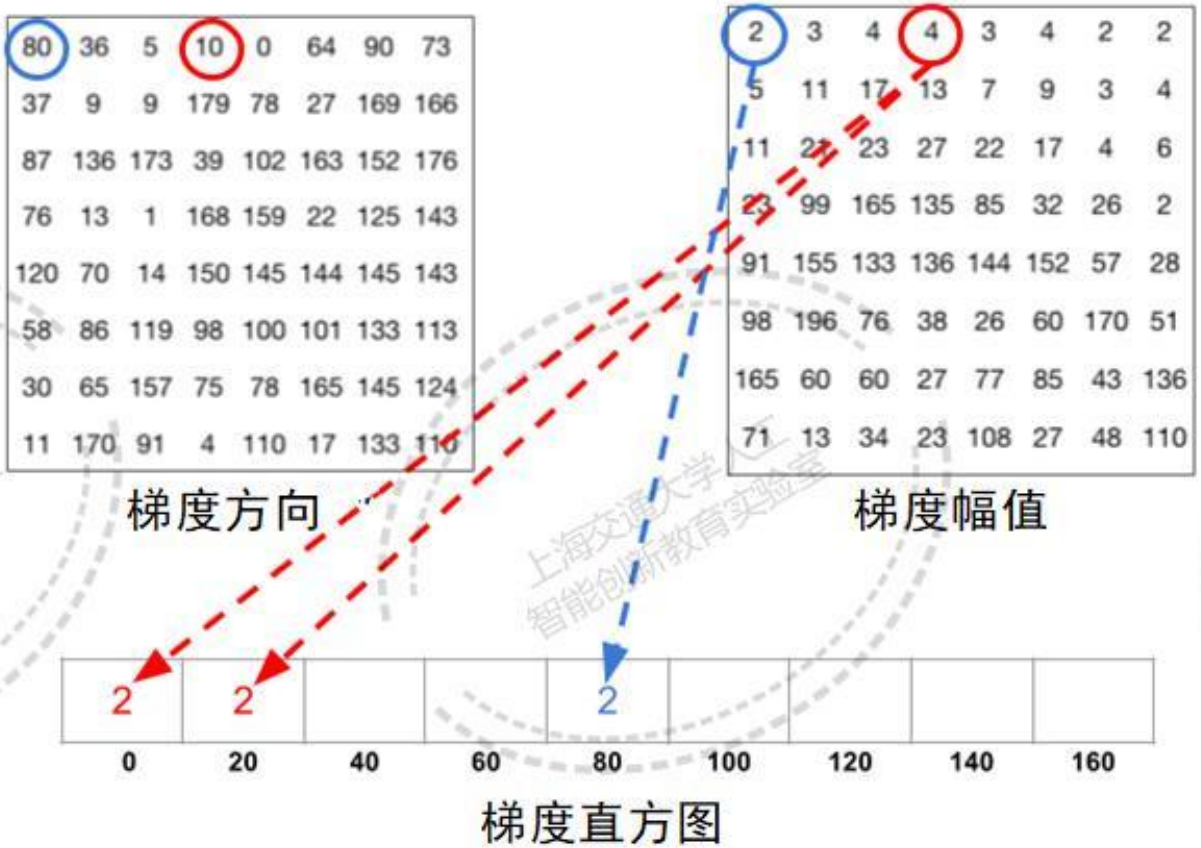
梯度幅值

80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

梯度方向

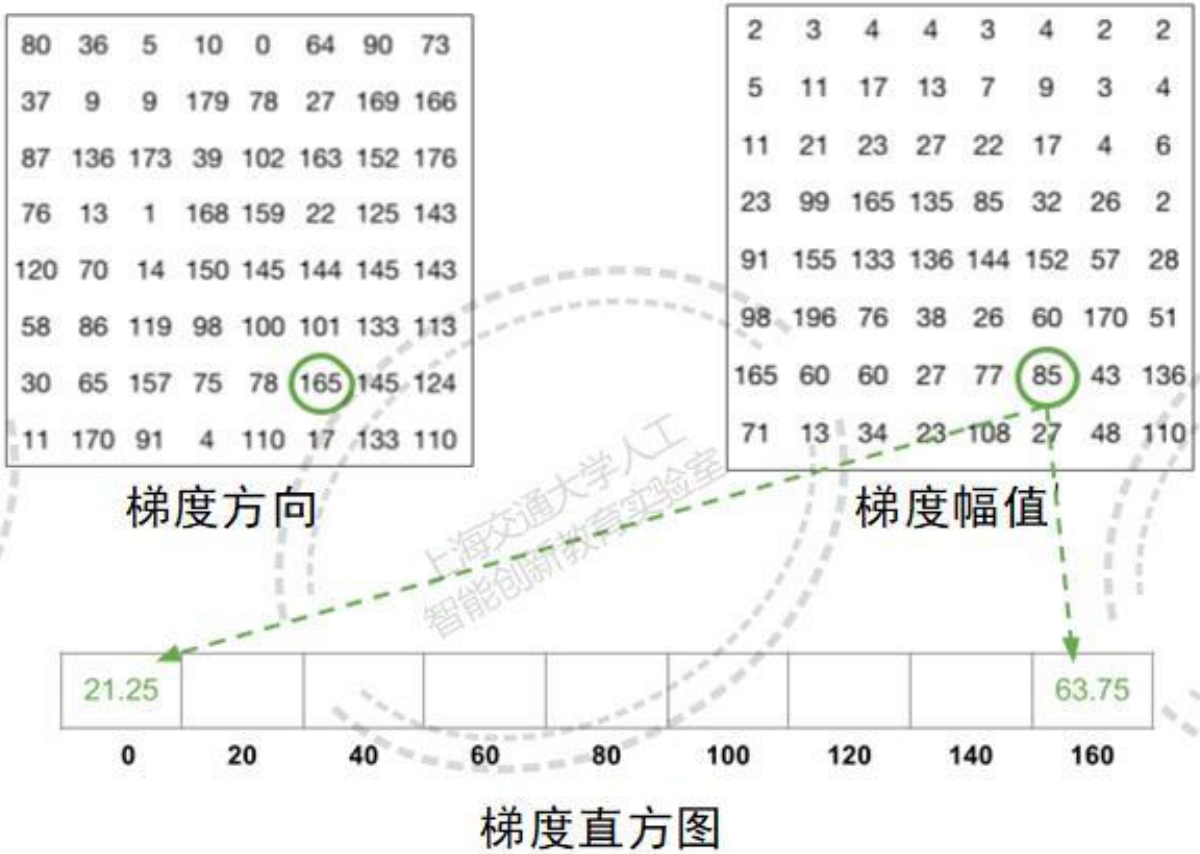
图像梯度直方图特征提取

- 计算这个8x8大小的图像块的梯度直方图
- 梯度方向0到180度平均分成9个区域，分别以0，20，40，60，80，100，120，140，160为中心点
- 根据每个像素点的梯度方向，将梯度幅值按照规则计入相应梯度直方图中



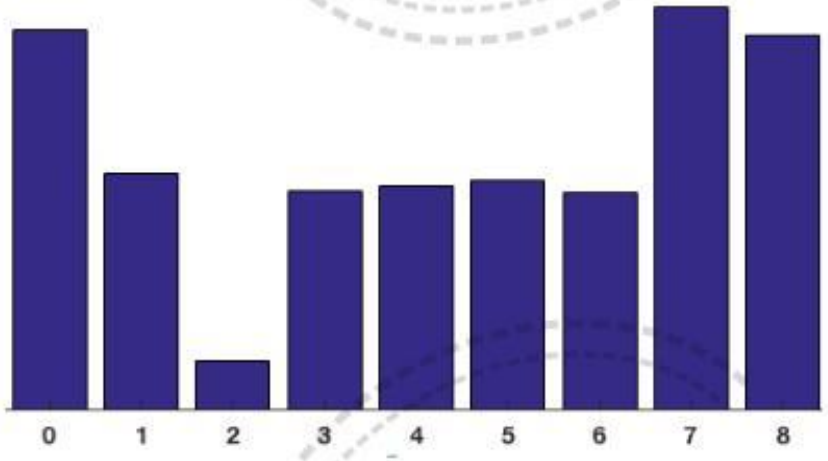
图像梯度直方图特征提取

- 对于梯度方向角度刚好在区间中点位置时候，该点的梯度幅值直接累加计入即可，如上页中蓝色的像素点
- 当像素点梯度方向角度落在两个区间中点之间时，该点的梯度幅值按比例分给这两个区间，如左图中绿色像素点所示

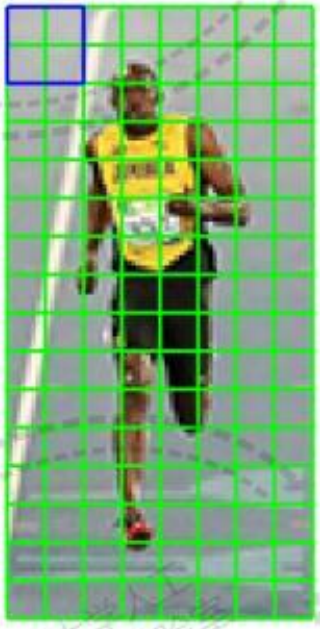


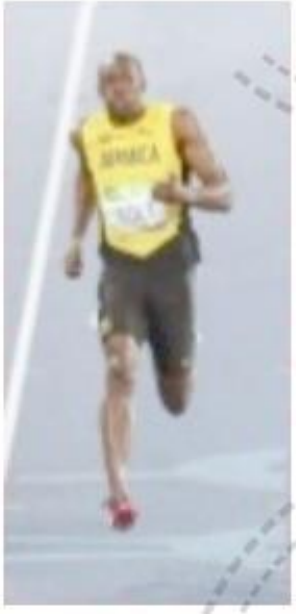
方向梯度直方图

某一小块图像的方向梯度直方图



归一化操作

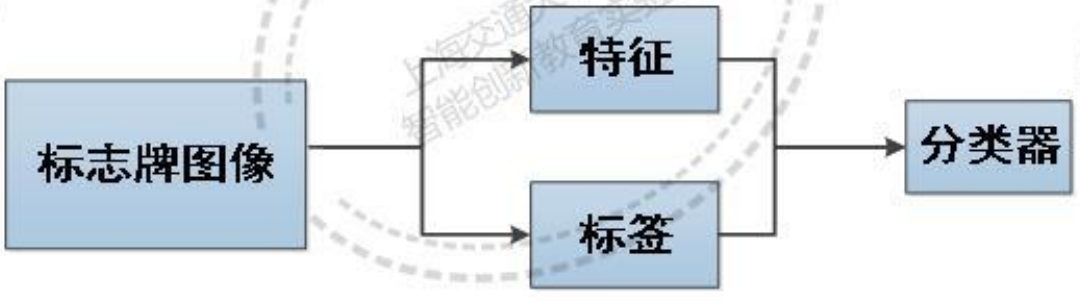




- 为了实现梯度方向直方图对光照的不变性，需要对原图进行归一化操作
- 归一化操作指，将四个8x8的图像块看成一个向量，计算出该向量幅值，再对每个像素值除以幅值，得到归一化后的像素值
- 原图是三通道的彩色图像，归一化操作要分别对每个通道上的像素都计算一遍

训练标志牌分类器

- 标志牌分类器的训练过程
- 输入标志牌图像
- 提取出该图像特征，与标志牌图像标签相对应
- 训练完成后得到标志牌分类器



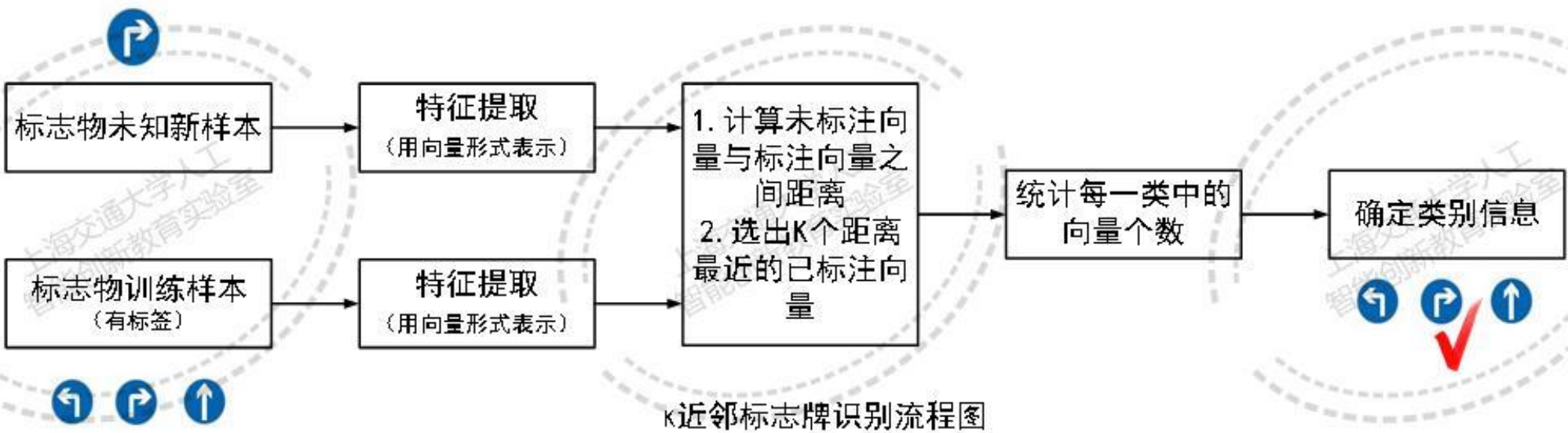
基于kNN算法识别标志牌—K近邻算法

对于传统k近邻算法，对于给定的数据集，有n个数据样本是已标记的，另一部分数据样本是未标记的，对于未标记的数据样本，通过如下方式进行分类：

① 度量每个未标记数据样本与所有已标记数据样本的距离；

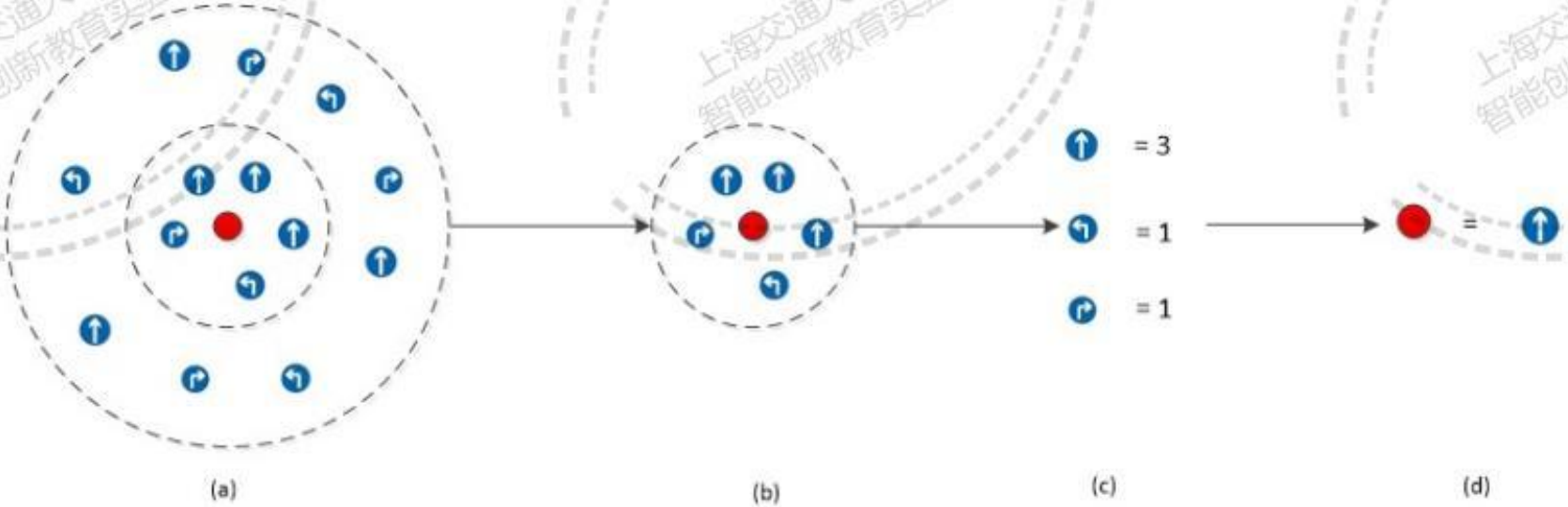
② 对所有求出的距离选择与未标记数据样本距离最近的k（k≤n）个已标记数据样本；

③ 统计这k个已标记的数据样本，那一类的数据样本个数最多，则未标记的拘束样本标记为该类样本



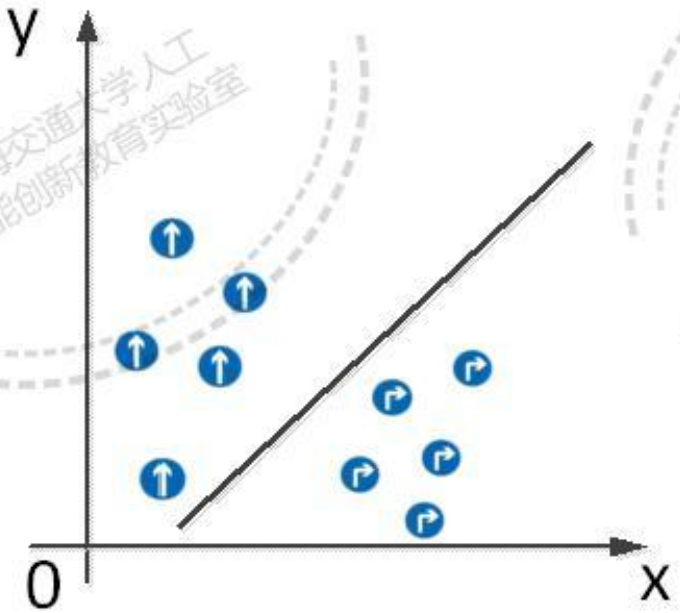
K近邻标志牌识别流程图

基于kNN算法识别标志牌

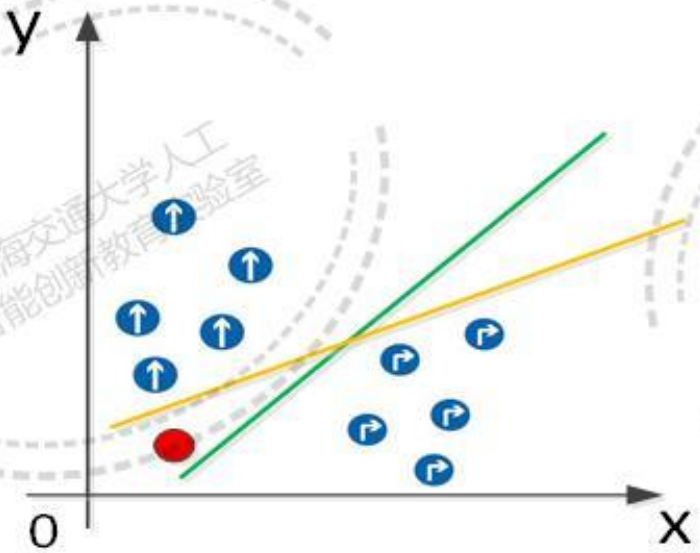


- 提取输入图像的HOG特征
- 计算样本训练集中每张图片的HOG特征与输入图像HOG特征之间的距离
- 在与输入图像特征相距最近的k个邻居特征中投票，输入图像属于投票最多的样本类别

基于SVM算法识别标志牌

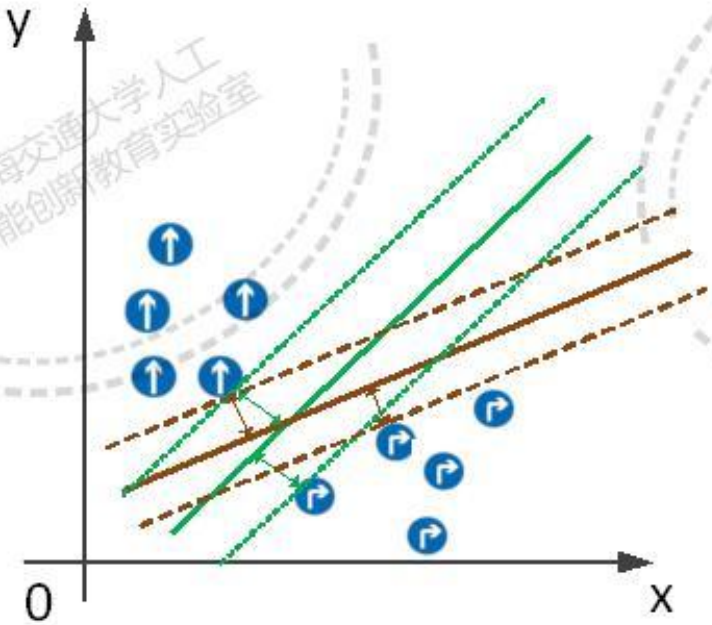


- 支持向量机(SVM)算法
- 仅靠一条直线区分分类结果
- 当输入图像特征位于直线上方时，该图像属于直行图像
- 当输入图像特征位于直线下方时，该图像属于右转图像

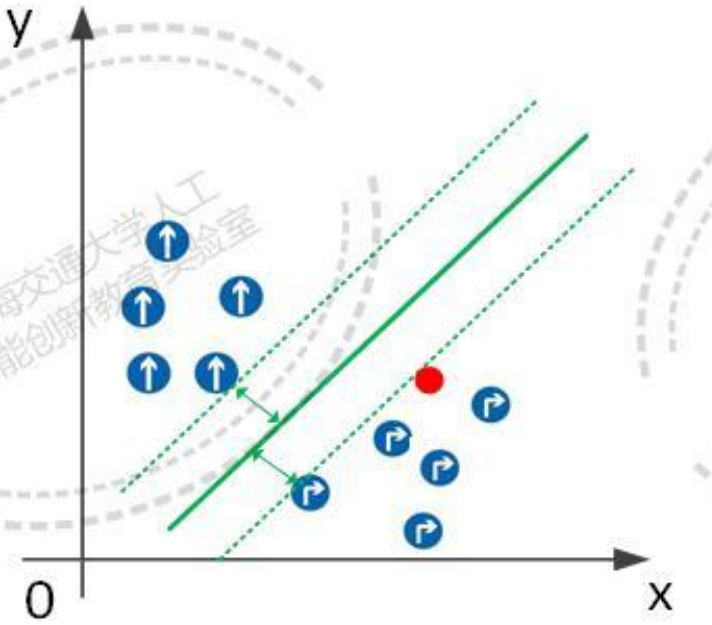


- 可以区分两类标志牌的直线不止一条
- 选择不同直线，黄色直线和绿色直线会导致输入同一张图像的分类结果不同
- 如何选择合适的分类直线

基于SVM算法识别标志牌



- 选择合适的分类线
- 选择具有最大的最小边界距离的分类线
- 最小边界距离由支持向量决定
- 最小边界距离越大，分类器缓冲区域越大，分类器正确性更高



支持向量实际上就是边界附近的若干个数据点

支持向量决定了最小边界距离，影响分类线的选择

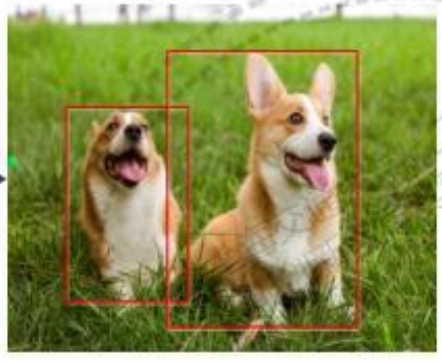
用SVM算法进行其他检测

SVM算法还可以进行更为复杂的行人检测

用SVM算法进行其他检测

车辆检测

检测小狗



识别性能评价

指标	说明
识别准确率	准确率是指正确分类的测试实例个数占测试实例总数的比例，用于衡量模型正确预测新的或者先前从未见过的数据的能力。
识别误分率	错误分类的测试实例个数占测试实例总数的比例，表示分类器做出错误分类的可能性有多大。
识别精确率	正确分类的正例的个数与分类为正例的实例的个数之比，又称为查准率
识别查全率	正确分类的正例个数占实际正例个数的比例，又称为召回率。
识别速度	识别速度可以理解为识别出一张标识物所需的时间，识别所需时间越短，说明该系统性能越好，实时性越强。