

语音识别基本原理及应用



中国科学院语言声学与内容理解重点实验室
中国科学院声学研究所

2021年12月15日



- 基本概念
- 主要内容
 - 预处理
 - 特征提取
 - 声学模型
 - 语言模型
 - 解码搜索
 - 语音识别技术的应用

- 语音是人类日常使用的最有趣的信号之一，语音也是一个多学科交叉的研究主题
 - 语音是人类交流的最自然的方式
 - 语音跟语言有关，而语言学是社会科学的一个分支
 - 语音跟人类的生理能力有关，而生理学是医学科学的一个分支
 - 语音跟声音有关，而声学是物理科学的一个分支
 - 语音跟信号和信息有关，信号和信息处理则是电子、通讯和计算机等领域的重要主题

语音研究的范围

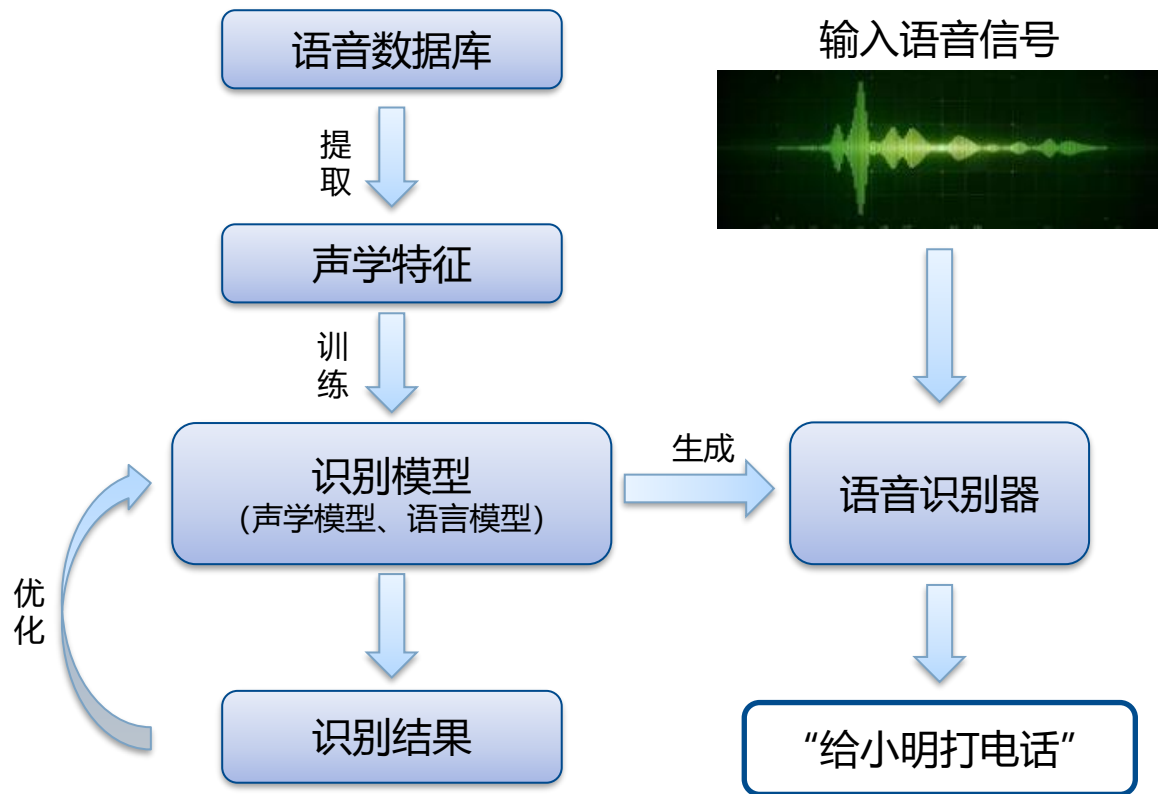


声 学						
声波频率	0Hz	20	200	8k	20k	1G
波段划分	次声波	语音 可听波			超声波	特超声或微波超声
应用领域	气象预测 军事用途	语音通信、信息检索、识别、 理解、交互等			工业探 测、医 疗保健	声学研 究 微电子学 医学
特点		物理学、数学、信息等学科高度 交叉融合，人类自然交流的方式！				

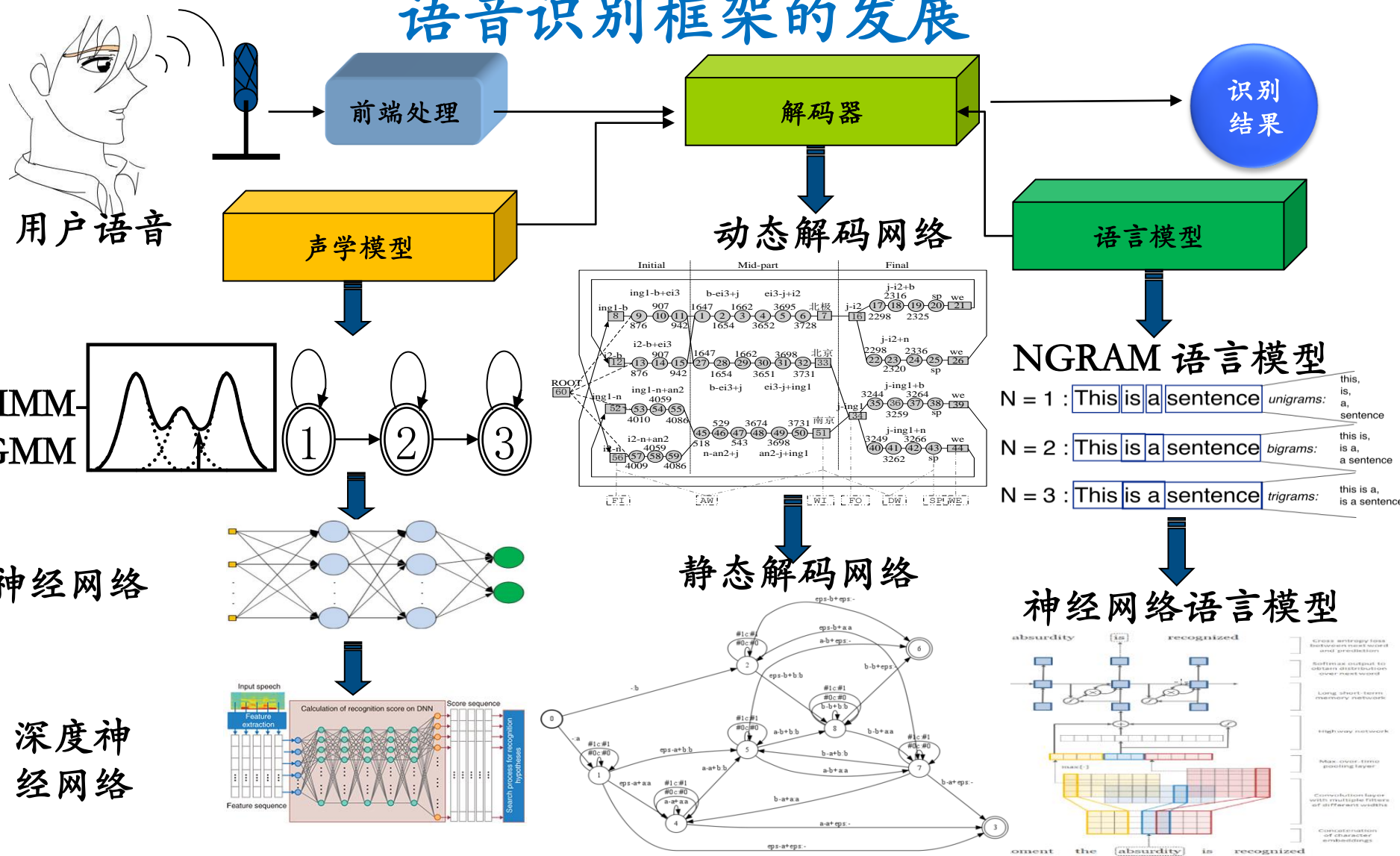
语音识别是以语音为研究对象，通过语音信号处理和模式识别让机器自动识别和理解人类口述的语言，就是**让机器听懂你说话**。



语音识别基本原理

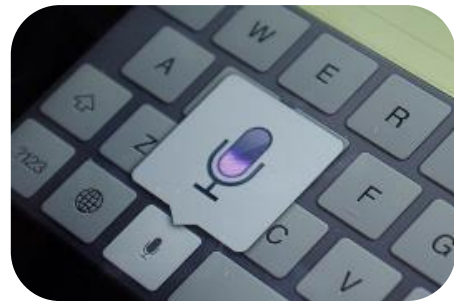


语音识别框架的发展



语音识别技术的现状和发展

- 语音搜索（朗读风格）准确率已达到90%以上，可以实用，未来搜索引擎中，通过语音进行搜索的比例会逐步提高。
- 安静环境下，自然口语对话识别准确率在80%左右，有待进一步提高。
- 复杂声学环境下（如互联网音视频），自然口语对话识别准确率下降比较厉害，这时比较实用的技术是采用关键词检索，通过特定敏感关键词的检测，进行语音内容分析。
- 目前的语音识别以国内语言为主，对多语言的支持与国外差距较大。



- 语音识别基础

- 语音的基本概念
- 语音识别基本原理

- 预处理

- 端点检测 (VAD)
- 归一化技术

- 语音特征提取

- MFCC特征
- PLP特征
- FBANK特征

- 声学模型

- 动态时间规整 (DTW)
- 隐马尔科夫模型 (HMM)
- 分类模型 (SVM)
- 高斯混合模型 (GMM)
- 神经网络声学模型

- 语言模型

- 统计语言模型 (NGRAM)
- 神经网络语言模型

- **解码器**
 - 搜索策略
 - WFST解码
- **自适应技术**
 - MLLR、MAP算法
 - 鲁棒特征、模型
- **深度神经网络**
 - 框架、算法和拓扑结构
 - 在语音识别中的作用
- **语音识别后处理技术**
 - 关键词识别
 - 发音评估
- **语音识别的应用**
 - 语音搜索
 - 客服质检
 - 智能家居
 - 国家安全



预处理



- 噪声消除
 - 两遍的维纳滤波消除平稳的背景噪声
- 垃圾语音过滤
 - 采用混合高斯建模的方法去除垃圾语音（震铃，彩铃，拨号音，笑声、咳嗽声等非自然人语音）
- 语音分段
 - 采用活动窗渐进寻找声学变更点的方法
 - 采用贝叶斯信息准则(BIC)



- 语音聚类
 - 采用层次聚类方法
- 语音端点检测 (Voice Activity Detection, VAD)
 - 判断语音信号中的语音段落的起始点和终结点的位置
 - 可以用于去掉多余的非有声信号，提高系统处理语音的速度，同时减少因非有声信号进入后端分析系统而产生的干扰。



特征提取



特征提取：即对不同的语音寻找其内在特征，由此来判别出未知语音，所以每个语音识别系统都必须进行特征提取。

特征的选择对识别效果至关重要。同时，还要考虑特征参数的计算量。



语音信号的特征主要有时域和频域两种：

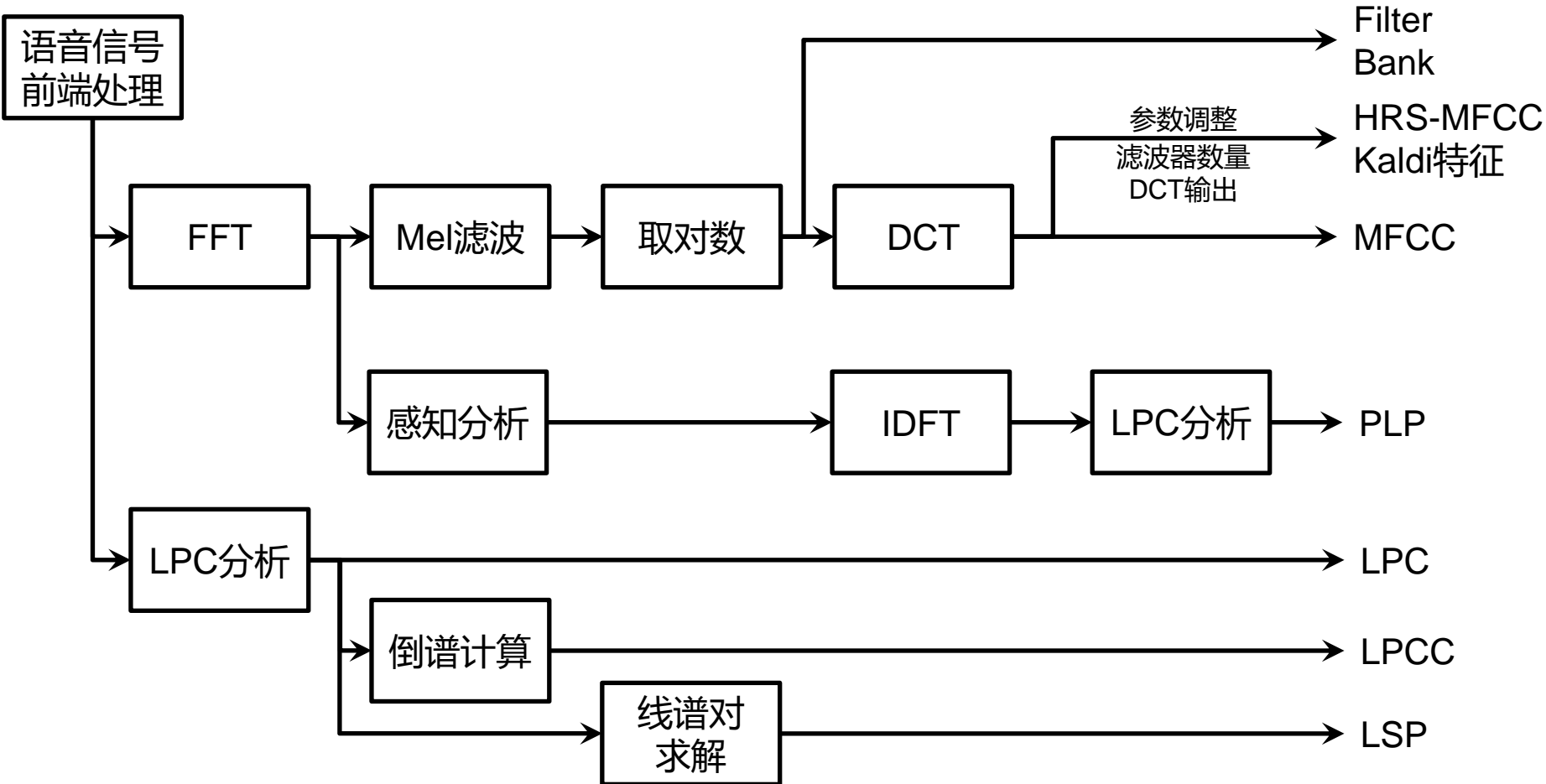
时域特征：短时平均能量、短时平均过零率、共振峰、基音周期等；

频域特征：线性预测系数(LPC)、LP倒谱系数(LPCC)、线谱对参数(LSP)、短时频谱、Mel频率倒谱系数(MFCC)、感知线性预测(PLP)等。

目前用的比较多的特征是Mel频率倒谱系数(MFCC)和感知线性预测(PLP)。

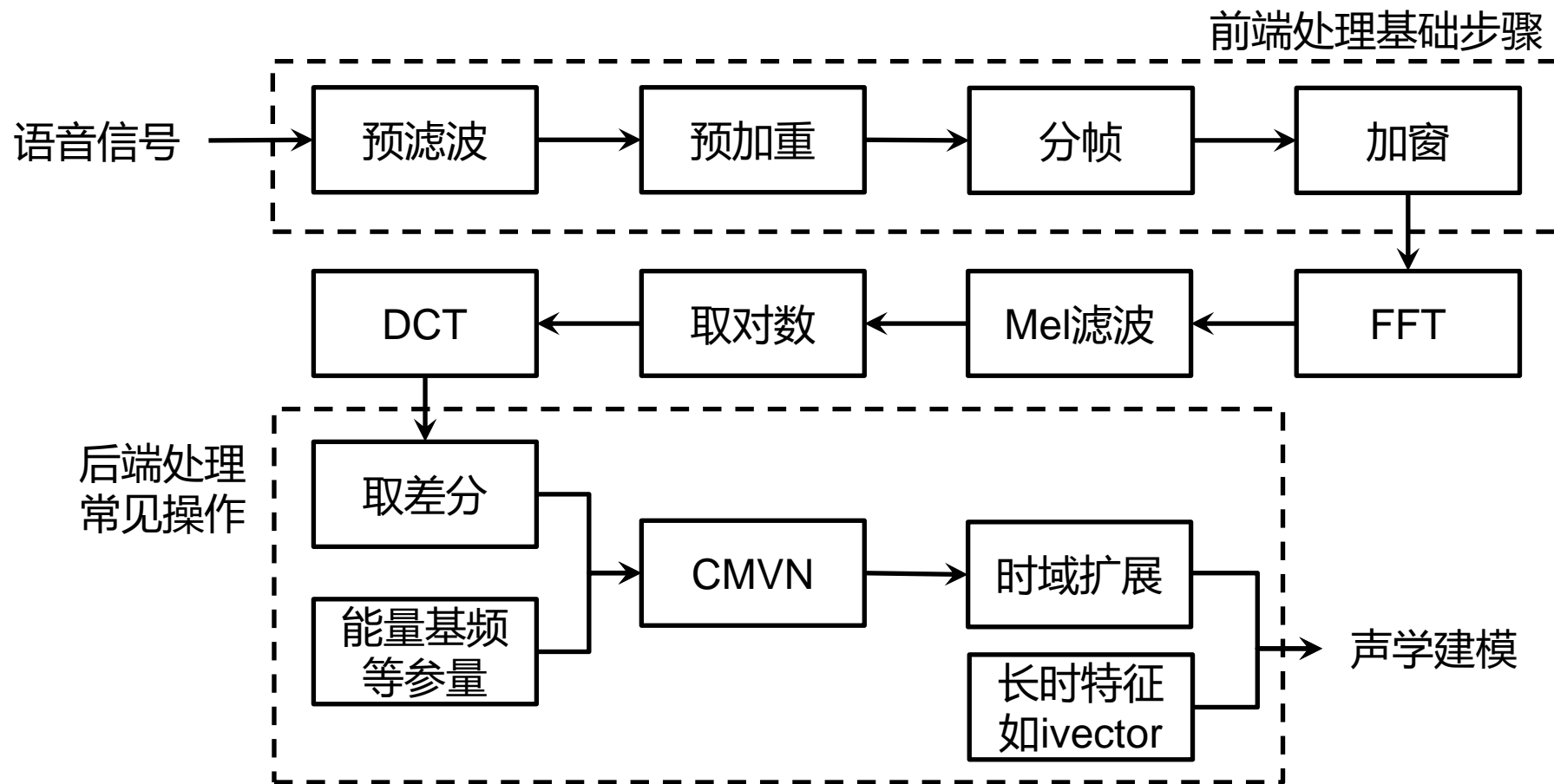


常见短时特征间的关系





实例：MFCC特征提取



■ 预滤波

- 抑制低频分量，防止50Hz交流电工频干扰
- 抑制1/2采样率以上的高频分量，防止混叠

■ 应用实例

- 8k采样率的电话信道数据通常的滤波范围为60Hz-3400Hz
- 16k麦克风录音数据可以不使用预滤波

■ 预加重

- 由于高频信号在传播过程中衰减更为严重，因此在所接收语音的有效频带内，频率越高的部分幅值越小
- 预加重旨在提升高频部分，使频谱整体更为平坦，以便于进行频谱分析
- 按时间序列依次计算

$$y(n) = x(n) - \alpha x(n - 1)$$

其中 $x(n)$ 为原始信号序列， $y(n)$ 为加重后序列，系数 α 在0.9到1之间，常取0.97。

■ 分帧

- 将所分析语音约束在一定范围之内，使其基本满足短时平稳性假设
- 语音识别领域的常规分帧方法为25ms帧长、10ms帧移

■ 加窗

- 分帧后的截断信号在进行傅里叶变换时会产生吉布斯效应，在时域加窗可以对其进行有效抑制
- 常用窗类型如汉明窗等

MFCC特征提取

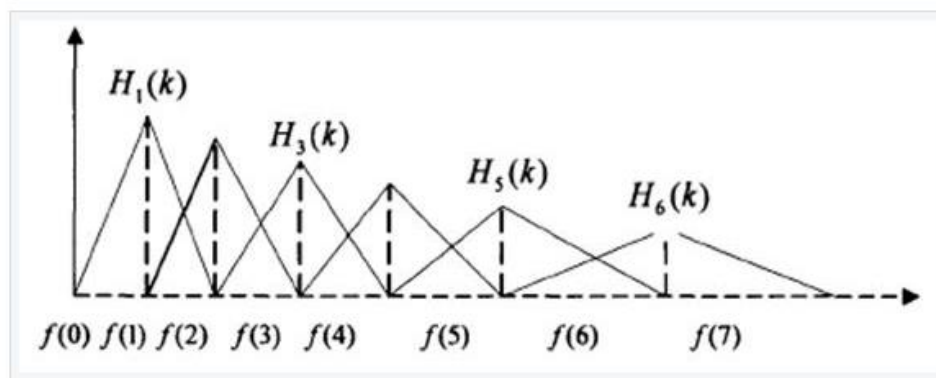
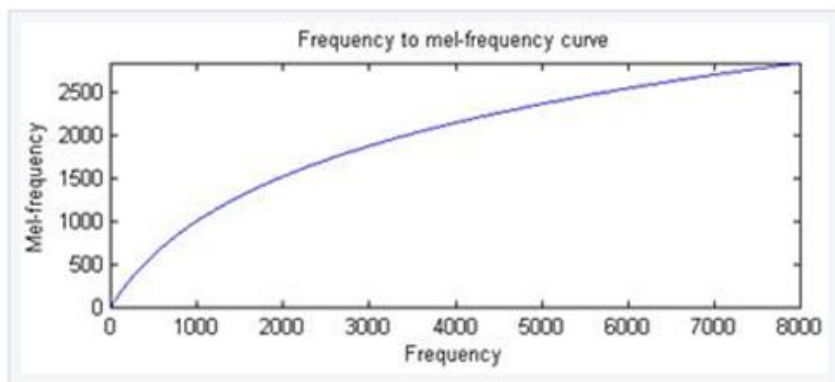
■ 快速傅里叶变换 (FFT)

■ 时频转换基本操作，语音在频域上的特性更易于观察

■ Mel滤波

■ 基于Mel频率曲线的等间隔三角窗滤波

■ 传统做法8k语音一般为15组，16k语音一般为23组





MFCC特征提取

■ 取对数

■ 压缩动态范围

■ 离散余弦变换 (DCT)

■ 从频域转换到倒谱域，抽取频谱包络

■ 细节可以参考“同态信号处理”理论

■ 取差分

- 求取基本特征变化率，称为动态特征

- 实验室计算方法：

$$\Delta x(n) = [x(n+1) - x(n-1) + 2 * (x(n+2) - x(n-2))]/10$$

■ 倒谱均值方差规整 (CMVN)

- 全局模式、离线模式、在线模式

- 离线模式计算

- 对于单个语音片段，计算每一个维度的均值和方差并据此将其规整为标准正态分布 $N(0, 1)$



声学模型



□ 声学单元应该具有的特性

- 一致性：不同语音实例中相同的语音单元在声学上一致
- 可训练性：建模单元需要足够的训练数据进行参数估计
- 可共享性：不同的建模单元间共享某些具有共性的训练数据

□ 声学单元如何挑选

- 句子：中国科学院大学
- 单词：中国-科学院-大学
- 单字：中-国-科-学-院-大-学
- 音素：zh-ong1-g-uo2-k-e1-x-ve2-vv-van4-d-a4-x-ve2
- 考虑协同发音的三元音素:e1-x-ve2 和 a4-x-ve2

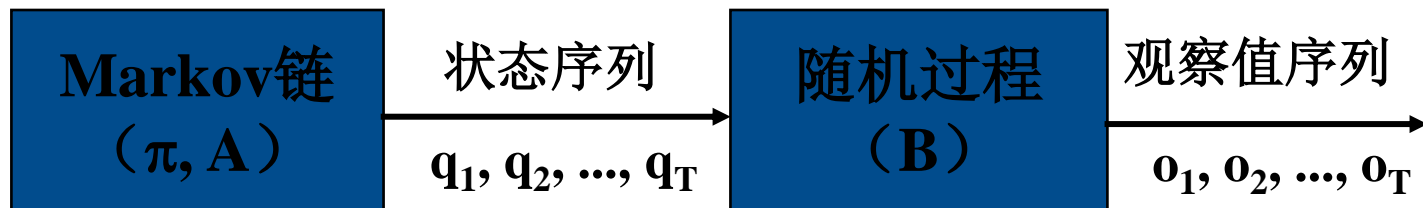
□ 声学单元对应的模型形式应该是什么

- HMM-GMM
- HMM-NN



HMM概念

- HMM的状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来
- 观察到的事件与状态并不是一一对应，而是通过一组概率分布相联系
- HMM是一个双重随机过程，两个组成部分：
 - 马尔可夫链：描述状态的转移，用转移概率描述。
 - 一般随机过程：描述状态与观察序列间的关系，用观察值概率描述。



HMM的组成示意图



HMM的基本要素

- 用模型五元组 $\lambda = (N, M, \pi, A, B)$ 用来描述HMM，或简写为 $\lambda = (\pi, A, B)$

参数	含义	实例
N	状态数目	缸的数目
M	每个状态可能的观察值数目	彩球颜色数目
A	与时间无关的状态转移概率矩阵	在选定某个缸的情况下，选择另一个缸的概率
B	给定状态下，观察值概率分布	每个缸中的颜色分布
π	初始状态空间的概率分布	初始时选择某口缸的概率

一个隐马尔可夫模型由下列参数来决定：

(1) N —模型的状态数目。

状态的集合表示为 $S = \{S_1, S_2, \dots, S_N\}$

(2) M —观测符号数。

即每个状态可能输出的观测符号的数目。

观测符号集合表示为 $O = \{O_1, O_2, \dots, O_M\}$

(3) A —状态转移概率分布。

状态转移概率构成的矩阵为

$$A = \{a_{ij}\}, a_{ij} = P[q_{t+1} = S_j | q_t = S_i], 1 \leq i, j \leq N$$



(4) B —状态的观测符号概率分布。

$$B = \{b_j(O_k)\}, b_j(O_k) = P[\text{在 } t \text{ 时刻的输出符号为 } O_k \mid q_t = S_j] \\ 1 \leq j \leq N, 1 \leq k \leq M$$

(5) π —初始状态分布。

$$\pi = \{\pi_i\}, \pi_i = P[q_1 = S_i], 1 \leq i \leq N$$

为了完整地描述一个隐马尔可夫模型，应当指定状态数 N ，观测符号数 M ，以及三个概率密度 A 、 B 和 π 。这些参数之间有一定的联系，因此为了方便，HMM 常用 $\lambda = (A, B, \pi)$ 来简记。



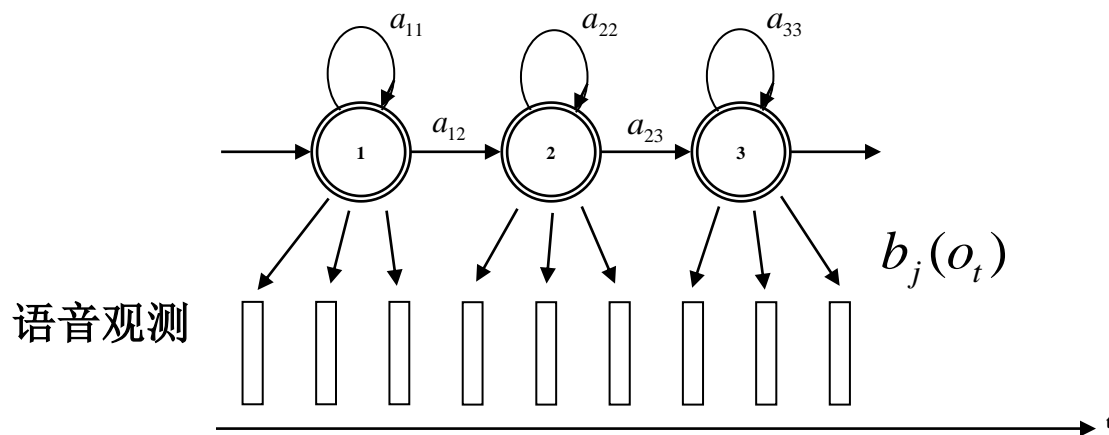
隐马尔可夫模型的三个基本问题

给定HMM的形式后，为了将其应用于实际，
必须解决以下三个基本关键问题：

- (1) 已知观测序列 $O = \{O_1, O_2, \dots, O_T\}$ 和模型 $\lambda = (A, B, \pi)$ ，如何有效的计算在给定模型条件下产生观测序列 O 的概率 $P(O|\lambda)$ 。
- (2) 已知观测序列 $O = \{O_1, O_2, \dots, O_T\}$ 和模型 $\lambda = (A, B, \pi)$ ，如何选择在某种意义上最佳的状态序列。
- (3) 给定观测序列，如何调整参数 (A, B, π) 使条件概率 $P(O|\lambda)$ 最大。

语音识别中常用的HMM-GMM模型结构

对应于一个音素，或者三音子



高斯混合

$$b_j(o_t) = \sum_k \omega_k N(\mu_{jk}, \Sigma_{jk})$$



语言模型



语言建模

- 从统计角度看，自然语言中的一个句子 s 可以由任何词串构成。

- $P(s)$ 有大有小。如：

s_1 = 我刚吃过晚饭

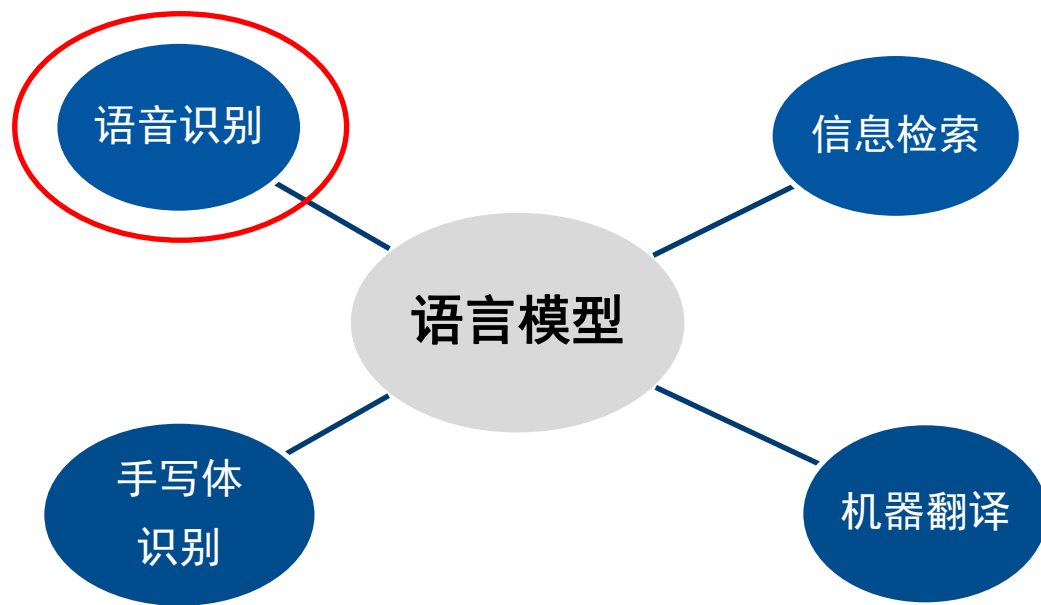
s_2 = 刚我过晚饭吃(并不要求语法是完备的,可对任意 s 给出概率)

$$P(s_1) > P(s_2)$$

- 对于给定的句子 s 而言，通常 $P(s)$ 是未知的。
- 对于一个服从某个未知概率分布 P 的语言 L ,

根据给定的语言文字样本估计 P 的过程被称作统计语言建模。

■ 用数学的方式描述语言学中词与词之间的约束现象

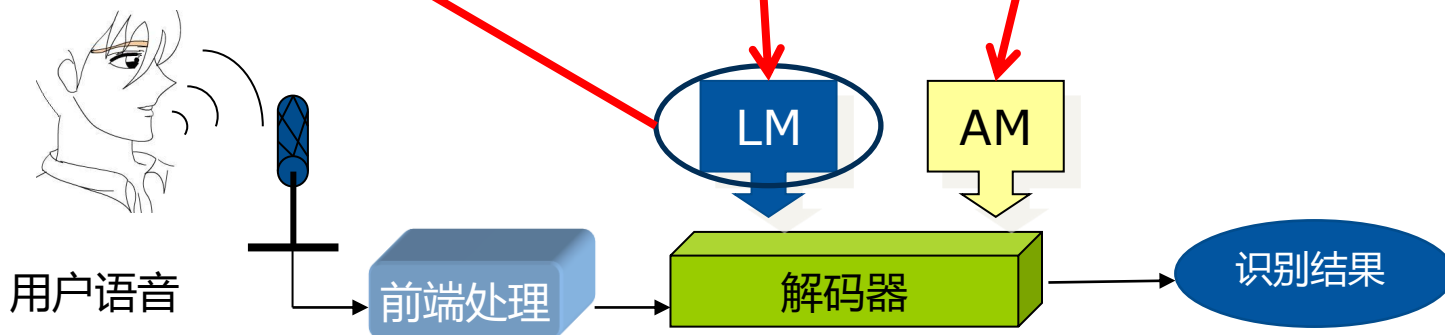


语言模型广泛应用场景

语言模型在语音识别中的作用

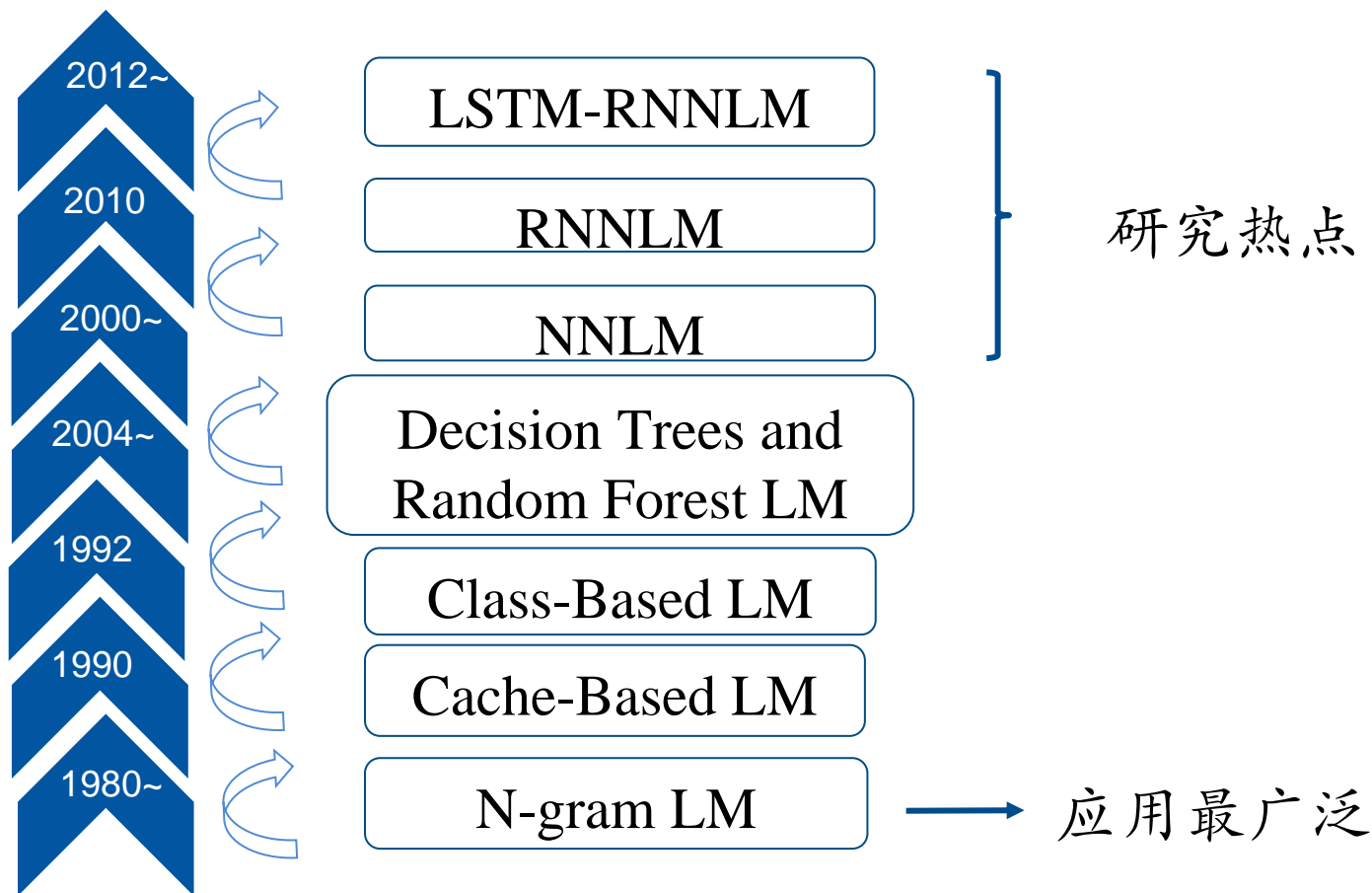
$$\begin{aligned} W^* &= \arg \max P(W | O) \\ &= \arg \max \frac{P(W) * P(O | W)}{P(O)} \\ &\approx \arg \max P(W) * P(O | W) \end{aligned}$$

描述信源的先验知识，用于估计每个候选词序列的概率





- 有的词序列听起来很像，但并不都是正确的句子
 - Ni xian zai zai gan shen me.
 - 词序列：
 - 你现在在干什么？ ✓
 - 你先在载感什么？





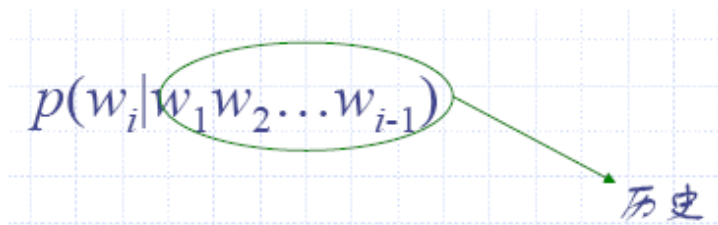
n-gram语言模型

- 一般来说，如果用变量 s 代表文本中一个任意的词序列，它由顺序排列的 L 个词组成，即 $s=w_1w_2\dots w_L$ ，则统计语言模型就是该词序列 s 在文本中出现的概率 $P(s)$
- 利用概率的乘积公式（链式法则）， $P(s)$ 可展开为：

$$\begin{aligned} p(s) &= p(w_1)p(w_2 | w_1)p(w_3 | w_1w_2)\dots p(w_L | w_1w_2\dots w_{L-1}) \\ &= \prod_{i=1}^L p(w_i | w_1w_2\dots w_{i-1}) \end{aligned}$$

不难看出，为了预测词 w_n 的出现概率，必须知道它前面所有词的出现概率。从计算上来看，这种方法太复杂了。

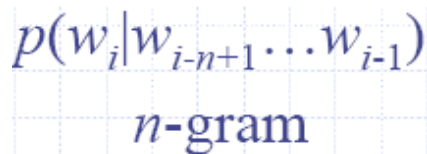
“马尔科夫假设” 下一个词的出现仅仅依赖于它前面的一个词或者几个词.


$$p(w_i | w_1 w_2 \dots w_{i-1})$$

历史

为了便于计算，通常考虑的历史不能太长，一般只考虑前面**n-1**个词构成的历史。

即：


$$p(w_i | w_{i-n+1} \dots w_{i-1})$$

n-gram

■ N元文法语言模型

N元
文法
语言
模型

优点：简单、高效、易用

缺点：数据稀疏、参数空间过大
无法捕捉词与词之间的相似度
无法有效利用长距离的上下文信息

研究内容：基于N-gram语言模型平滑算法的研究

仍然是现今语音识别系统第一遍解码的首选模型

■ 神经网络语言模型

神经网络语言模型



优点: NNLM能够自动学习词语的连续空间表达,使得词义和用法相近的词在连续空间聚集在一起



缺点: 训练复杂度过高
训练时间成本高
不适合大量语料的训练



研究内容: 把NNLM应用在语音识别第二遍多候选解码中

在困惑度和语音识别性能上
优于N-gram语言模型

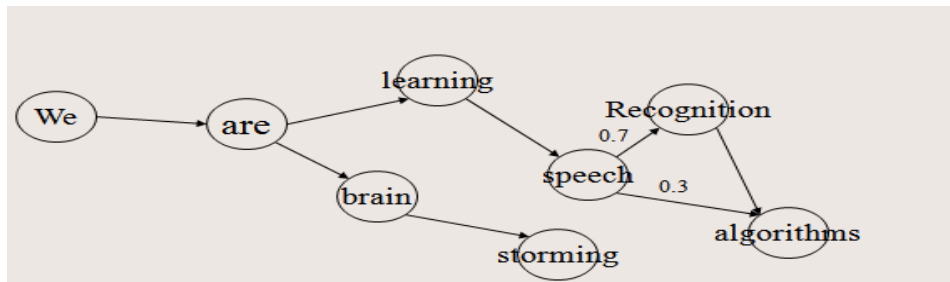


解码器



- 问题:
未知语音信号对应于一个句子文本.
- 解码器的任务: 找出解码网络中具有最高积累概率的路径
 - 搜索网络构建
 - 令牌传递算法
- 最佳路径:
对应于一个词序列, 其在所有可能的词序列组合中具有最高匹配概率。
- 解码器关键技术:
 - 优化WFST网络的编译方法和构建步骤
 - 优化基于WFST网络的搜索算法

- 语言模型
 - 词图/有限状态图



- N元统计语言模型

$$P(w_N | w_{N-1}, w_{N-2}, \dots, w_1) \approx P(w_N | w_{N-1}, w_{N-2})$$

- 发音字典

speech	s p <u>iy</u> <u>ch</u>
recognition	r eh k ax n <u>ih</u> <u>sh</u> ax n

- 声学上下文

speech	sil-s+p s-p+iy p-iy+ch iy-ch+sil
recognition	<u>sil-r+eh</u> <u>r-eh+k</u> <u>eh-k+ax</u> <u>k-ax+n</u> <u>ax-n+ih</u> <u>n-ih+sh</u> <u>ih-sh+ax</u> <u>sh-ax+n</u> <u>ax-n+sil</u>

- 声学模型

- 隐含马尔科夫模型 (HMM)

- 特征

- MFCC, PLP, LPCC....



搜索空间的优化

- LVCSR搜索空间

- 动态

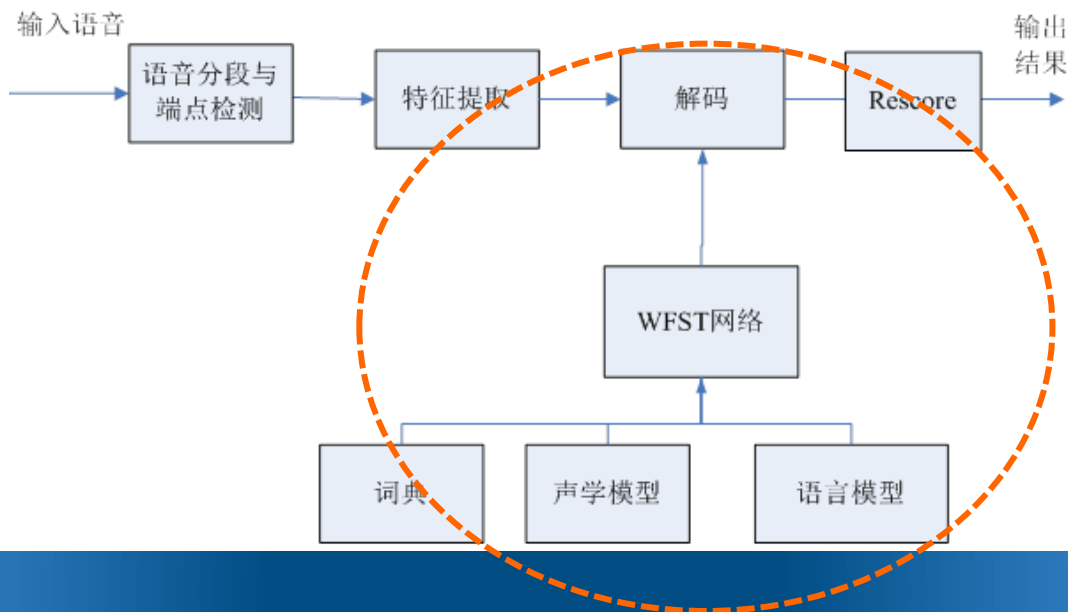
- 动态加载知识源
 - 解码过程需要查询知识源信息
 - 内存占用小，空间构建速度快
 - 解码速度不够快

- 静态

- 预先编译好知识源
 - 解码过程就是一个FST的搜索问题
 - 内存占用大，空间构建速度慢
 - 解码速度快

解码算法:

- ✓ 采用基于WFST的静态搜索空间构建方法，有效地单遍集成各种知识源，将声学模型、声学上下文、发音词典、语言模型等静态编译成状态网络





基于深度神经网络的语音识别

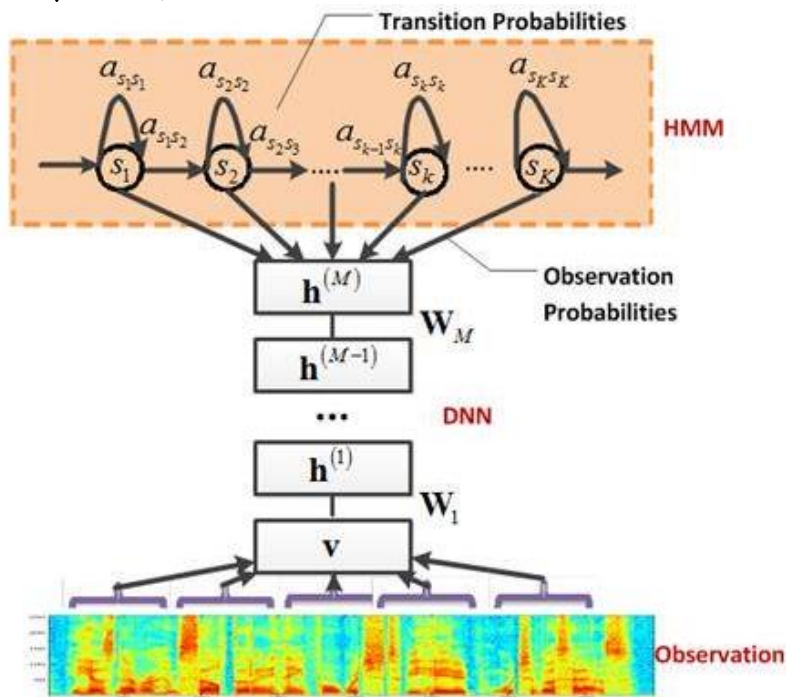


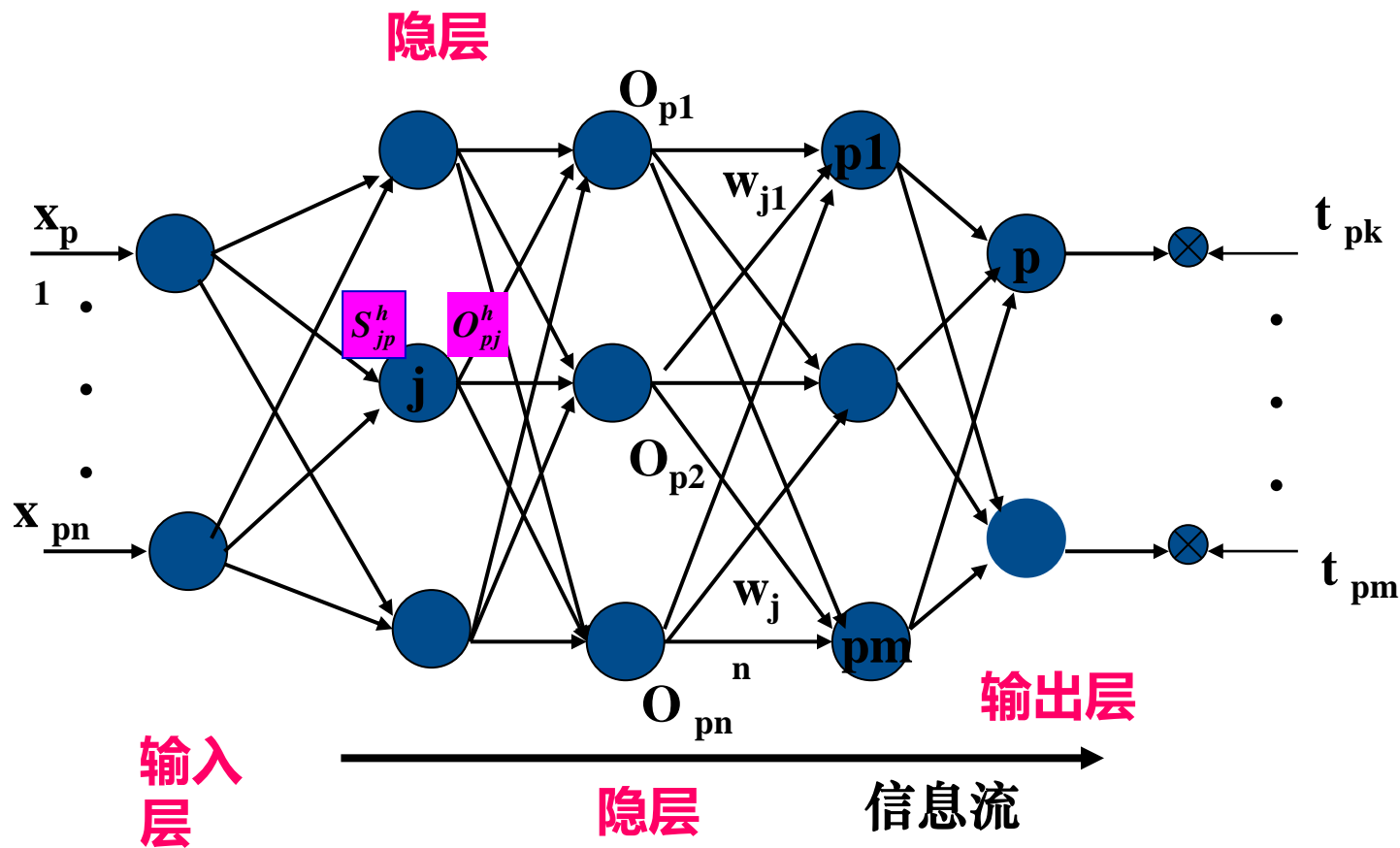
神经网络 (Neural Network, NN)

- **生物神经网络**主要是指人脑的神经网络，它是人工神经网络的技术原型。

人脑是人类思维的物质基础，思维的功能定位在大脑皮层，后者含有大约10¹¹个神经元，每个神经元又通过神经突触与大约10³个其它神经元相连，形成一个高度复杂高度灵活的动态网络。作为一门学科，生物神经网络主要研究人脑神经网络的结构、功能及其工作机制，意在探索人脑思维 and 智能活动的规律。
- **人工神经网络**是生物神经网络在某种简化意义下的技术复现，作为一门学科，它的主要任务是根据生物神经网络的原理和实际应用的需要建造实用的人工神经网络模型，设计相应的学习算法，模拟人脑的某种智能活动，然后在技术上实现出来用以解决实际问题。因此，生物神经网络主要研究智能的机理；人工神经网络主要研究智能机理的实现，两者相辅相成。

深度神经网络架构







- 传统神经网络的学习算法
 - 前馈型神经网络学习算法
 - 反馈型神经网络学习算法
 - 自组织神经网络学习算法



- 几种典型的神经网络
 - NN (前向神经网络)
 - RNN (递归神经网络)
 - BRNN (双向递归神经网络)
 - LSTM/LSTM-DBRNN (长短时记忆-双向递归神经网络)
 - CNN (卷积神经网络)



- 在语音识别中的应用
 - 基于DNN的声学模型建模
 - 基于DNN的语言模型建模
 - 基于DNN的语音端点检测
 -



语音识别的应用



移动
互联网

所需技术

- 语音识别
- 自然语言问答
- 噪声处理技术
- 大数据



智能
家电

所需技术

- 语音识别
- 自然语言问答
- 远讲及噪声处理
- 自动内容识别



车载
电子

所需技术

- 语音识别
- 自然语言问答
- 远讲及噪声处理



智能
客服

所需技术

- 语音识别
- 自然语言问答
- 情绪识别



可穿戴
设备

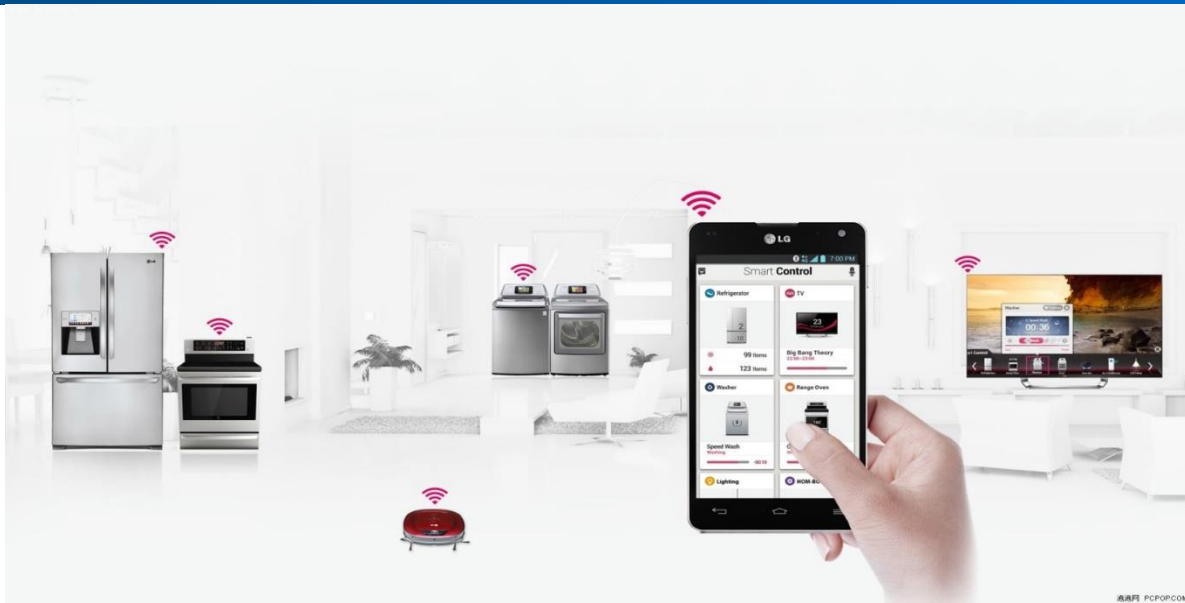
所需技术

- 语音识别
- 语音激活
- 远讲及噪声处理
- 自然语言问答



语音助手





智能家电是将微处理器、传感器技术、网络通信技术引入家电设备，能够自动感知和控制住宅空间状态和家电自身状态、家电服务状态。用户通过手机，语音等在住宅内或远程的执行控制指令。

智能家电构建了高效的住宅设施与家庭日程事务的管理系统，提升家居安全性、便利性、舒适性、艺术性，并实现环保节能的居住环境。

通过语音指令控制
智能电器开关：
控制灯，电视，音
乐播放等。



回答用户提出问题：识
别语音，识别语义内容，
检索查询用户所需信息，
并作出相应的应答；
查询天气，路线，交通
状况等。



监控调节室内温度，
湿度等环境：
向用户播报室内环境
状态，通过语音指令，
进行调控，开关空调，
窗帘，加湿器等。



日程和通讯管理：
根据用户的日程表，对用
户进行语音提醒；根据用
户指令，连接手机等通讯
设备，进行发信息，邮件，
uber叫车等操作。



常用形式:

- 车用导航
- 汽车信息系统 (智能仪表盘)
- 智能后视镜



中关村在线
ZOL.COM.CN



语音问答

- 阿里小蜜



- 微软小冰

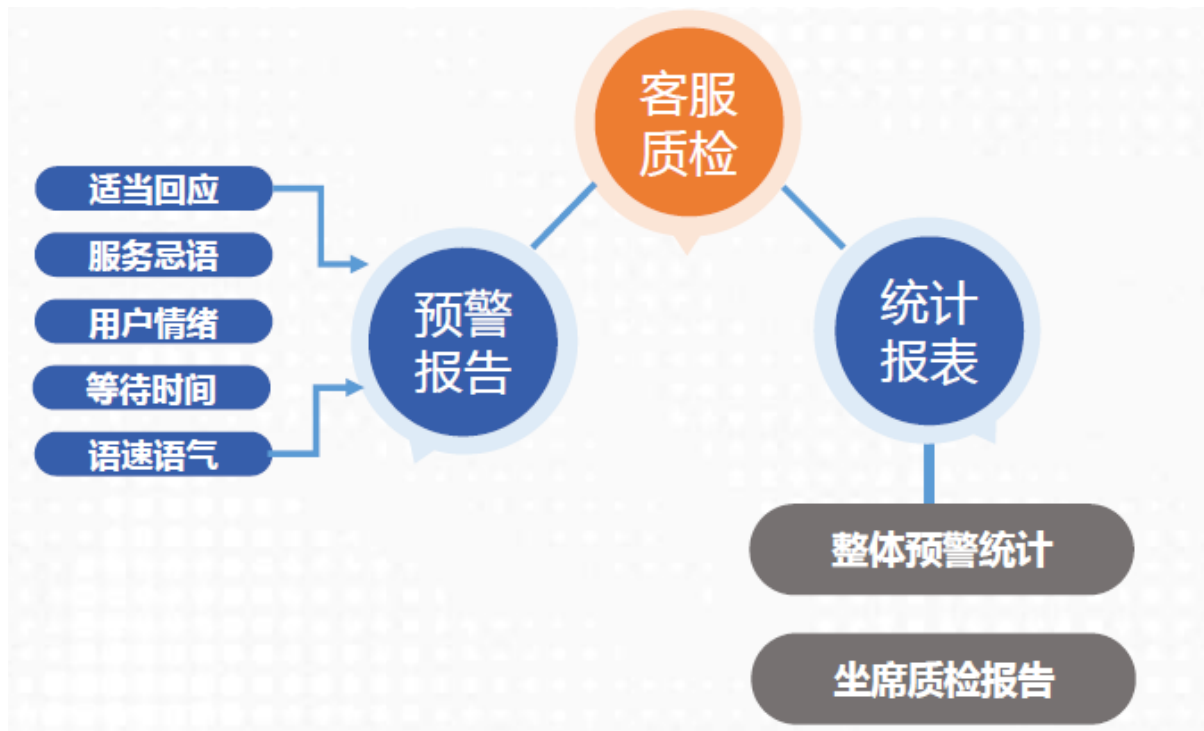


- 娇娇



- “阿里小蜜”、“微软小冰”等交互软件凭借在大数据、自然语义分析、机器学习方面技术积累，精炼为几千万条真实而有趣，并且实用的语料库，通过理解对话的语境与语义，实现了超越简单人机问答的自然交互。
- 娇娇是一款交通银行智能客服机器人，也是国内首个大规模投入到银行业中的实体机器人，被网友们称为“史上最萌大堂经理”。在研发过程中不但解决了营业厅嘈杂环境下的降噪技术，大大提高了语音识别的识别率，同时对远讲语音识别进行了优化，提升用户体验。

语音质检



智能语音质检分析系统包含语音质检、语音分析两大部分，可以实现对呼叫中心的话务内容进行服务质量监控、客户需求挖掘等语音分析功能。

可穿戴设备

受硬件形态的约束，市场上有不少可穿戴设备都引入了智能语音的技术。主要的表现形式分两种，一种是通过先进的语音识别，听懂人类的指令，并作出一定的互动；另一种则是通过某种感应设备，监测当前用户是不是处于合理的状态下，否则会给予语音提示，起到纠错的作用。



- Rabiner L R, Juang B H. Fundamentals of speech recognition[J]. 1993.
- Huang X D, Ariki Y, Jack M A. Hidden Markov models for speech recognition[M]. Edinburgh: Edinburgh university press, 1990.
- Bourlard H A, Morgan N. Connectionist speech recognition: a hybrid approach[M]. Springer Science & Business Media, 2012.
- Jelinek F. Statistical methods for speech recognition[M]. MIT press, 1997.
- 王炳锡等著，实用语音识别基础，国防工业出版社，2005年1月出版
- 俞栋，邓力著，解析深度学习：语音识别实践，电子工业出版社，2016年7月出版
- 韩纪庆等编著，语音信号处理，清华大学出版社，2013年4月出版
- 吴岸城著，神经网络与深度学习，电子工业出版社，2016年6月出版
- 边肇祺等编著，模式识别，清华大学出版社，2000年1月出版
- 陈果果，都家宇，那兴宇，张俊博著，Kaldi语音识别实战，中国工信出版集团，2020年4月出版
- 洪青阳等著，语音识别原理与应用，中国工信出版集团，2020年6月出版



谢谢