第1章 语音识别概论

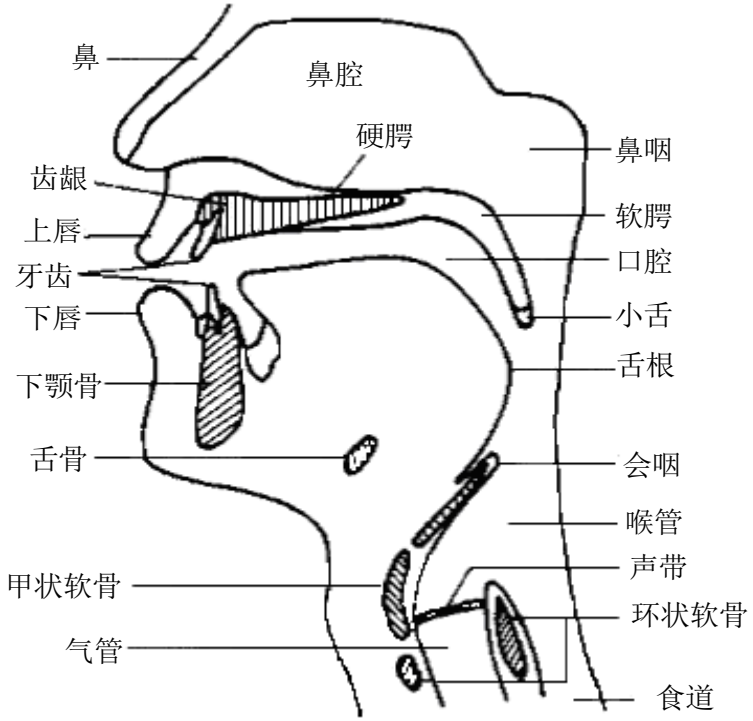
洪青阳
厦门大学信息学院
qyhong@xmu.edu.cn

纲要

1.1 语音的产生和感知
1.2 语音识别过程
1.3 语音识别发展历史
1.4 国内语音识别现状
1.5 语音识别建模方法
1.6 语音识别开源工具
1.7 语音识别常用数据库
1.8 语音识别评价指标

1.1 语音的产生机制—发音器官

- 人的发音器官包括：肺、气管、声带、喉、咽、鼻腔、口腔、唇。肺部产生的气流冲击声带，产生震动。



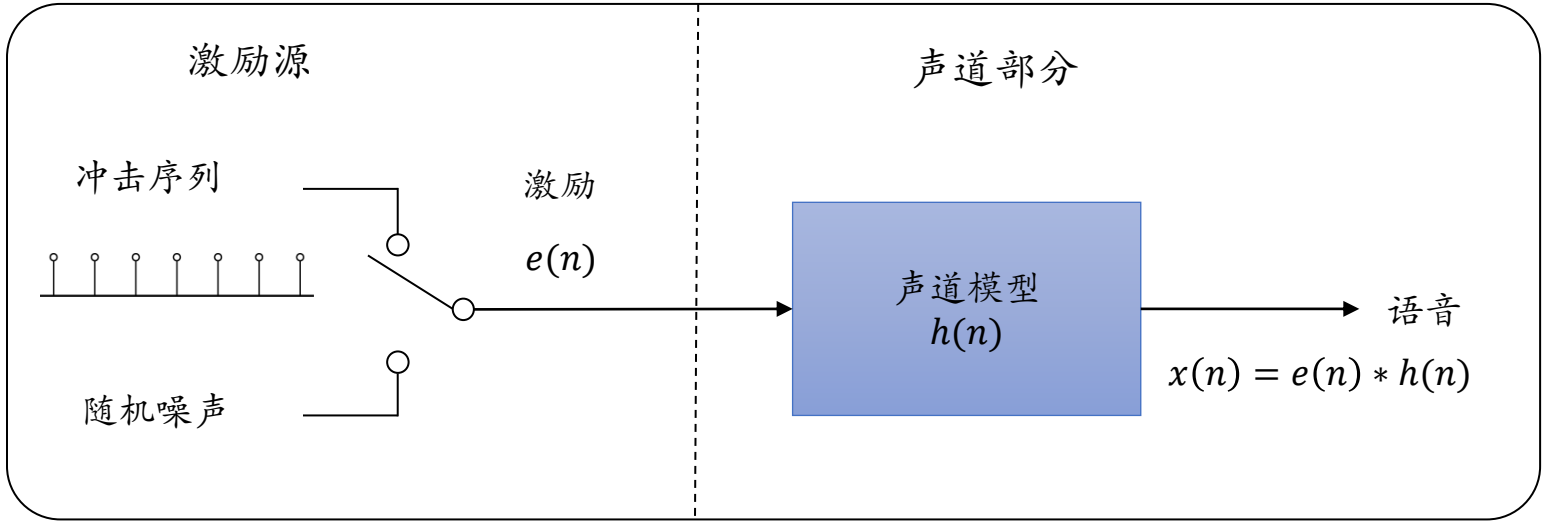
（图片来源：《语音信号处理》，韩纪庆等编著）

1.1 语音的产生机制

- 声带每开启和闭合一次的时间是基音周期（Pitch period, T），其倒数为基音频率（F_0=1/T，基频），范围在70-450Hz。基频越高，声音越尖细，如小孩的声音比大人尖，就是因为其基频更高。基频随时间的变化，也反映汉语的声调变化。

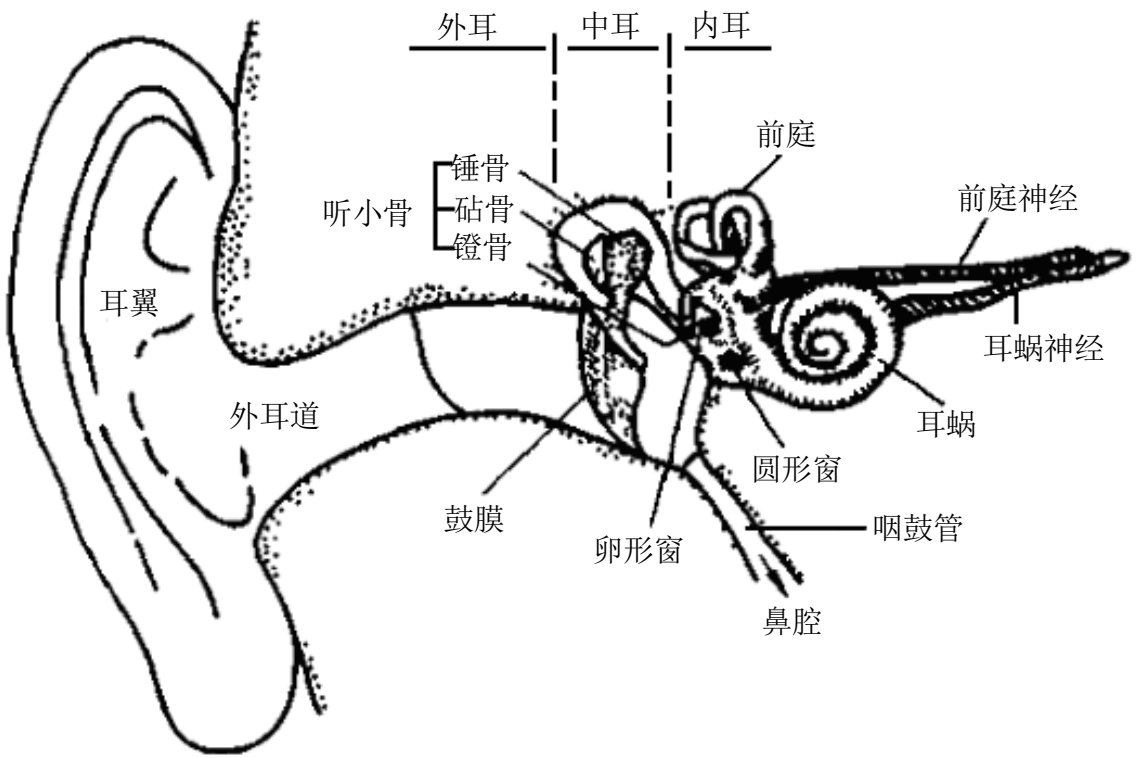
- 声道主要由口腔和鼻腔组成，是对发音起重要作用的器官，气流在声道会产生共振。前五个共振峰频率（F1, F2, F3, F4和F5），反映了声道的主要特征。共振峰的位置、带宽和幅度决定元音音色，改变声道形状可改变共振峰，改变音色。

1.1 语音的产生机制—激励模型



- **激励源：**
  - 发浊音：声带振动形成准周期的冲击序列；
  - 发清音：声带松弛，相当于一个随机噪声。
- **声道部分：**模拟口腔、鼻腔等声道器官，最后产生语音信号。

1.1 语音的感知机制—听觉器官



（图片来源：《语音信号处理》，韩纪庆等编著）

- 人耳三部分：
  - 外耳：外耳包含耳翼和外耳道，耳翼具有定向作用，外耳道同其它管道一样也有共振频率，大约是3400Hz。鼓膜位于外耳道内端，声音的振动通过鼓膜传到内耳。
  - 中耳：由三块听小骨组成，作用包括放大声压和保护内耳。中耳通过咽鼓管与鼻腔相通，其作用是调节中耳压力。
  - 内耳：耳蜗实现声震动到神经冲动的转换，并传递到大脑。
- 正常人耳感知的频率范围：20Hz~20kHz。

1.2 语音识别过程

- 人的大脑从出生婴儿开始，就不断在学习外界的声音，经过长时间的潜移默化，最终才听懂人类的语言。**机器跟人一样，需要学习语言的共性和发音的规律，才能进行语音识别。**
- 音素（phone）是构成语音的最小单位。英语中有48个音素（20个元音和28个辅音）。采用元音和辅音来分类，汉语普通话有32个音素，包括元音10个，辅音22个。
- 但普通话的韵母很多是复韵母，不是简单的元音，因此拼音一般分为声母（initial）和韵母（final），经过扩充和调整后，包含27个声母和38个韵母（不带声调）。

声母（27个）	b p m f d t n l g k h j q x zh ch sh r z c s _a _o _e _y _w _v
韵母（38个）	a o e i -i(前) -i(后) u v er ai ei ao ou ia ie ua uo ve(üe) iao iou uai uei an ian uan van(üan) en in uen vn(ün) ang iang uang eng ing ueng ong iong

1.2 语音识别过程

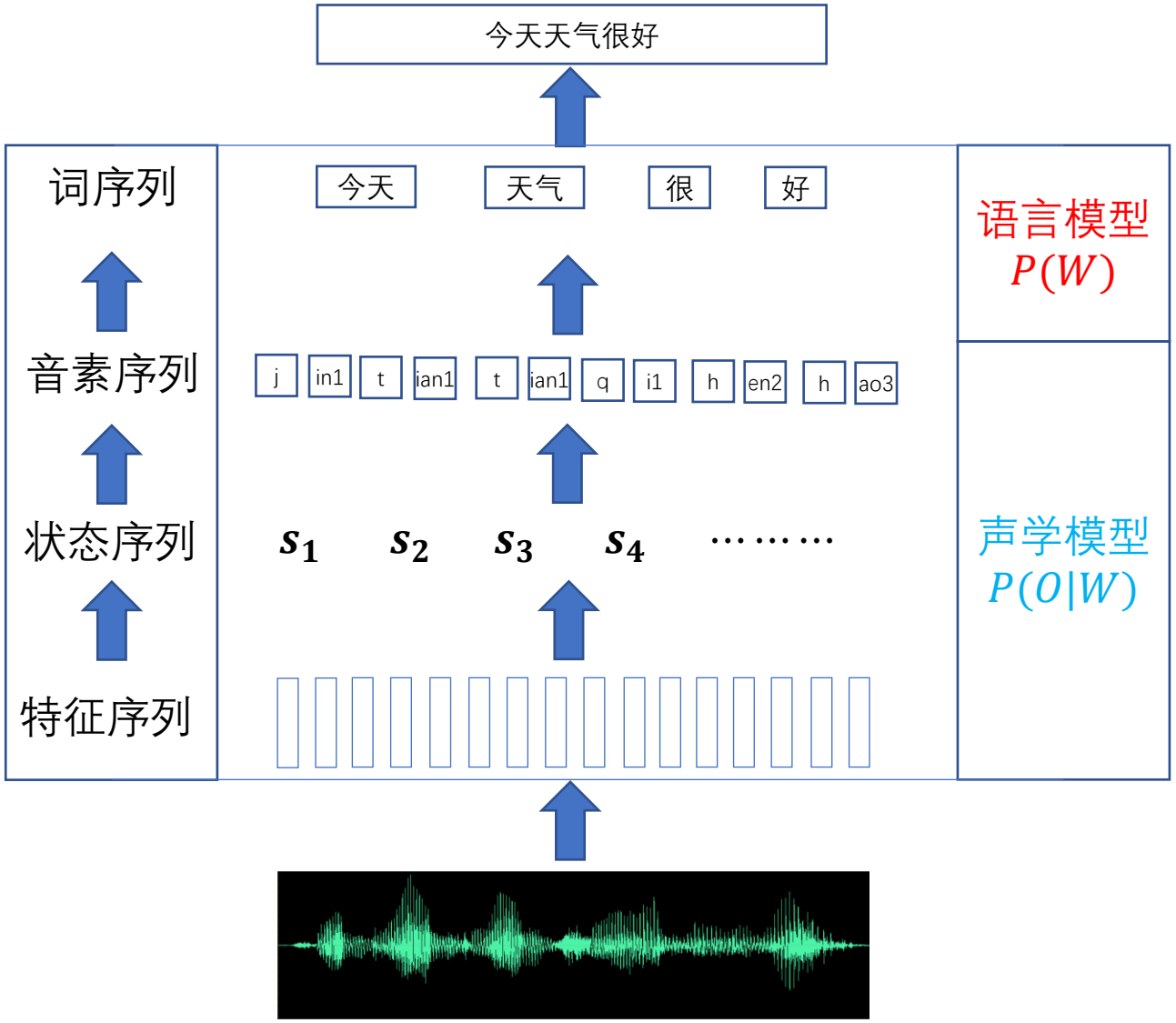
- 语音识别的任务为，找到对应观察序列 O 的最可能的词序列 $\widehat{W}$ 。按贝叶斯准则：

$$\widehat{W} = arg\max_{W} P(W|O) = arg\max \frac{P(O|W)P(W)}{P(O)}$$

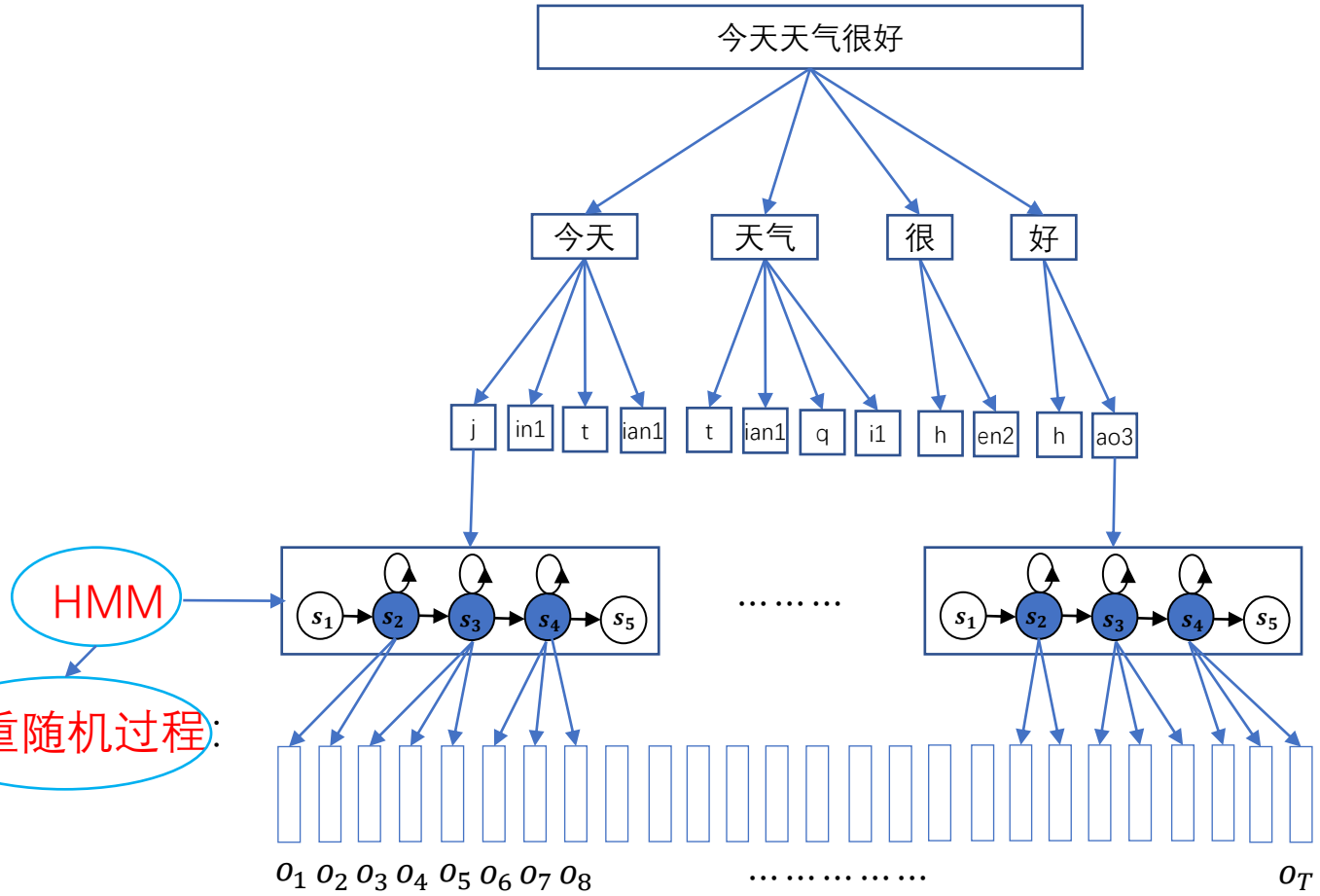
$$\widehat{W} = arg\max_{W} P(O|W)P(W)$$

- 要找到最可能的词序列，必须使上式右侧两项的乘积最大。第一项由声学模型决定，第二项由语言模型决定。

1.2 语音识别过程


今天天气很好
词序列
今天
天气
很
好
语言模型
$P(W)$
音素序列
j
in1
t
ian1
t
ian1
q
i1
h
en2
h
ao3
状态序列
s_1
s_2
s_3
s_4
……
声学模型
$P(O|W)$
特征序列

1.2 语音识别过程



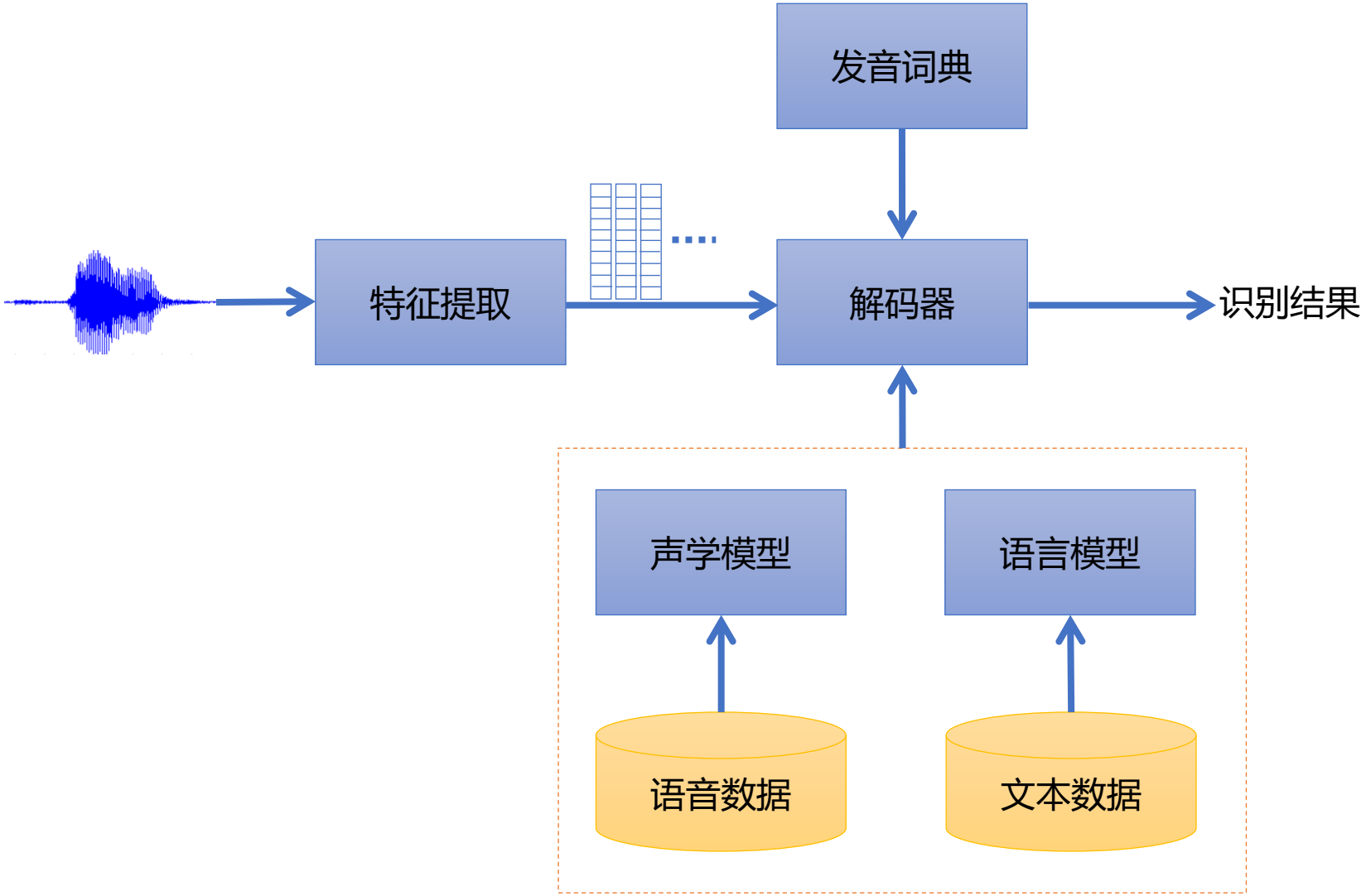
词典：

今天 j in1
天气 t ian1 q i1
很 h en2
好 h ao3

人的发音包含双重随机过程：

- **说什么不确定**
- **怎么说不确定**

1.2 语音识别过程—系统框架



1.3 语音识别发展历史

模型	模板匹配 Audry-识别 10个英文数字	DTW VQ	GMM-HMM	DNN	DNN-HMM	End-to-End: CTC RNN-T Attention Transformer
时间	**1950s**	**1970s**	**1980~1990s**	**2006**	**2012**	**2015以来**

孤立词识别 → 大规模连续语音识别 → 复杂场景识别

第一阶段：模板匹配 | **第二阶段：统计模型** | **第三阶段：深度学习**

1.3 语音识别发展历史

第一阶段：模板匹配（DTW）

- 1978年，日本人Sakoe and Chiba在前苏联科学家Vintsyuk的工作基础上，成功地使用动态规划算法将两段不同长度的语音在时间轴上进行了对齐，这就是我们现在经常提到的**动态时间规整**（Dynamic Time Warping，DTW）。

第二阶段：统计模型（GMM-HMM）

- 在80年代中期，Bell实验室的**L.R. Rabiner**等人对**隐马尔可夫模型**（Hidden Markov Modl，HMM）做了深入浅出的介绍，并出版了语音识别专著《Fundamentals of Speech Recognition》。
- 在DARPA的语音研究计划的资助下，又诞生出了一批著名的语音识别系统，其中包括**李开复（K.F. Lee）**在CMU攻读博士学位时开发的**SPHINX**系统。
- 英国剑桥大学**Steve Young**开创的的语音识别工具包**HTK**（Hidden Markov Tool Kit），自1995年发布以来，持续近20年被广泛采用。
- 但GMM-HMM系统对日常对话、新闻广播等场景的识别率达到80%左右，就无法再取得突破。2000年前后这个阶段语音识别的研究陷入了瓶颈期。

1.3 语音识别发展历史

第三阶段：深度学习（DNN-HMM，E2E）

- 2006年，**Hinton**揭开了属于**深度学习**的时代序幕。
- 2009年，**Hinton**和他的学生Mohamed将**深度神经网络（DNN）**应用于声学建模，他们的尝试在TIMIT音素识别任务上取得了成功。
- 2012年，微软研究院的**俞栋和邓力**等人提出了上下文相关的深度神经网络与隐马尔可夫模型融合的声学模型（**CD-DNN-HMM**），相比于传统的GMM-HMM系统获得超过20%的相对性能提升，这项工作是深度学习在语音识别上具有重大意义的成果。
- 由**Daniel Povey**领衔开发在2011年发布的**Kaldi**，是DNN-HMM系统的基石，在工业界得到广泛应用。
- 深度学习技术在近十几年中，一直保持着飞速发展的状态，从DNN到**RNN、CNN、LSTM、GRU、TDNN**等模型，推动着语音识别技术不断取得突破。
- 最近几年，基于**端对端（End-to-End，E2E）**的语音识别方案逐渐成为了行业中的关注重点，并开始落地应用。

1.4 国内语音识别现状

- 主要科研机构:
  - 中科院声学所、中科院自动化所
  - 清华大学、中国科技大学、哈尔滨工业大学
  - 上海交通大学、西北工业大学、厦门大学
- 工业界:
  - 老牌厂商：科大讯飞、捷通华声、中科信利
  - 互联网巨头：百度、腾讯、阿里巴巴
  - 互联网新贵：小米、京东、字节跳动、美团、贝壳找房等
  - 专业公司：云知声、思必驰、出门问问、天聪智能等
  - 工业巨头：华为等

1.4 国内语音识别现状—产业发展

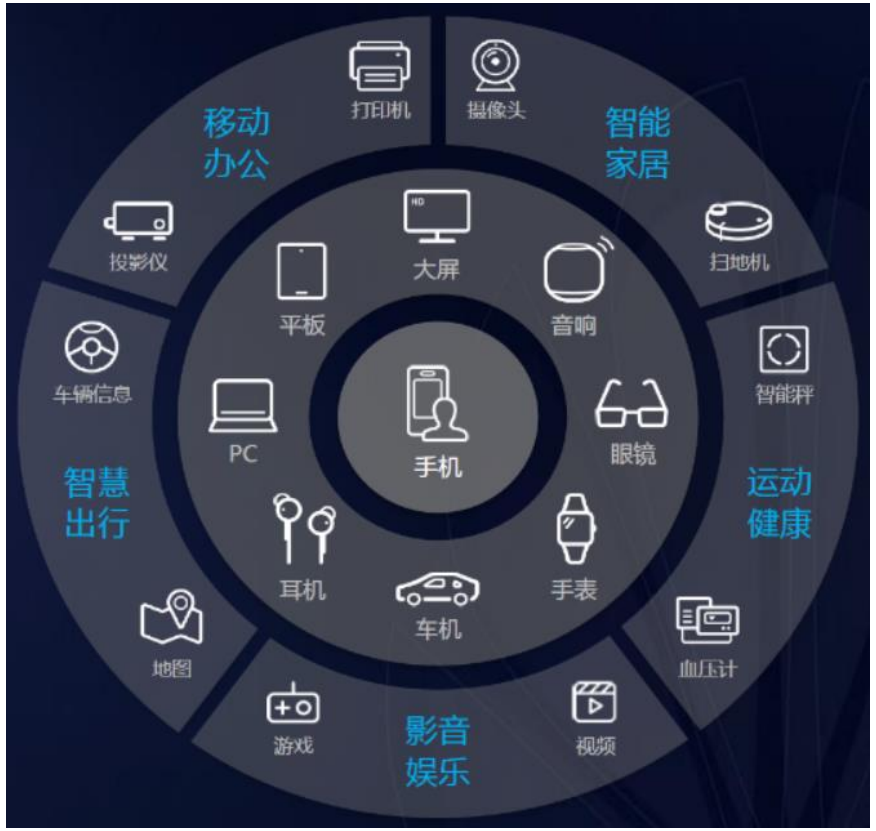
- 基于云端深度学习算法和大数据库支撑的在线语音识别率目前可以做到95%以上，科大讯飞、百度、阿里巴巴都提供了达到商业标准的语音识别服务。

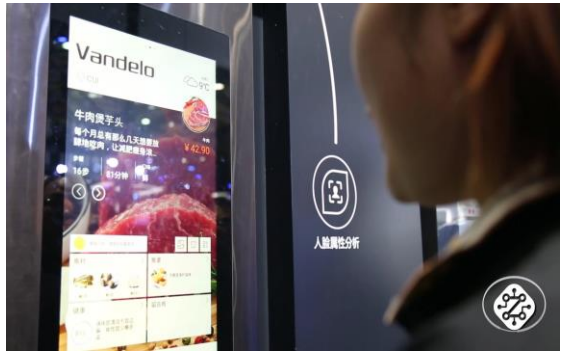
语音云

语音输入法

语音搜索

1.4 国内语音识别现状—产业发展





未来是万物互联的全场景智慧时代，需要适应不同场景多终端的语音交互技术

1.4 国内语音识别现状—产业发展

- 自2017年以来，智能家居逐渐普及，音箱市场开始火热，为抢占语音入口，阿里巴巴、百度、小米、华为等大公司纷纷推出了各自的智能音箱。

- 2019年第1季度，中国市场智能音箱出货量全球占比为51%，首次超过美国，成为全球最大的智能音箱市场。

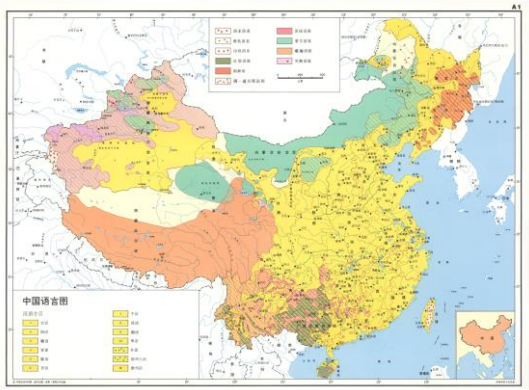
- IDC《中国智能音箱设备市场月度销量跟踪报告》显示，2022年上半年，中国智能音箱市场销量为1483万台。

1.4 国内语音识别现状—应用挑战

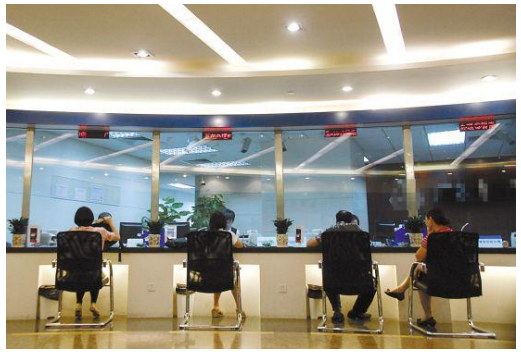
语音识别仍然面临的挑战：

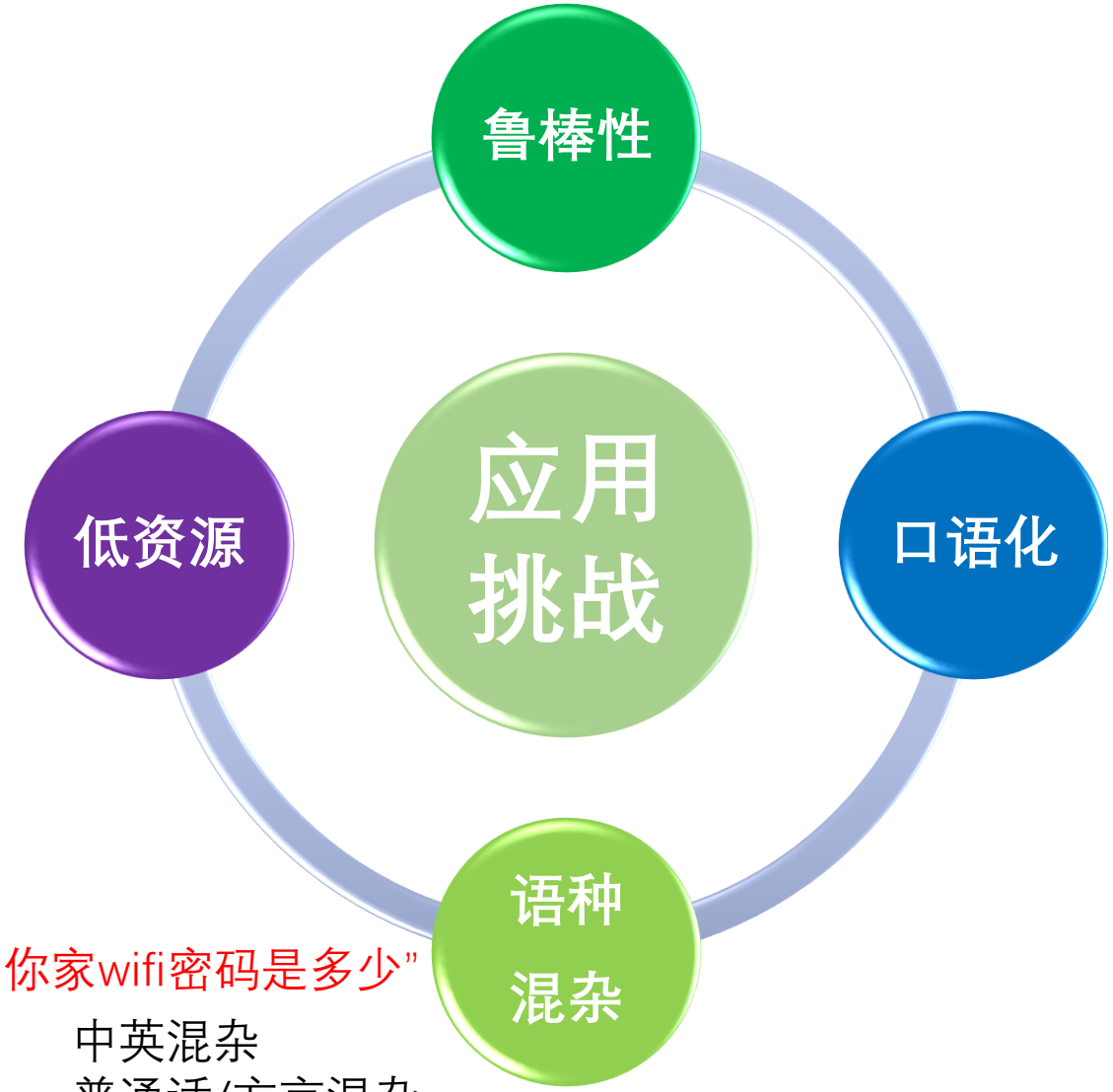
- 多数语音厂商在安静状态下的语音识别准确率大都达到95%以上，支持5米远场交互，如今均在向非标准环境下提高远场识别准确率、降低误唤醒率及加快响应速度的方向发展。
- 如何在复杂场景（包括非平稳噪声、混响、远场）下，提高语音识别鲁棒性，研发“能用=>好用”的语音识别产品，仍然是要重点解决的问题。

1.4 国内语音识别现状—应用挑战



方言识别（七大方言区）
特定场景（银行/证券柜台）





“你家wifi密码是多少”

中英混杂
普通话/方言混杂

电话闲聊
会议场景

1.4 国内语音识别现状—应用挑战

语音识别仍然面临的挑战：

- **口语化（多说话人）：**电话场景和会议场景的语音识别，包含很多口语化表达，如闲聊式的对话，可能还存在多人交替插话，识别效果很不理想。

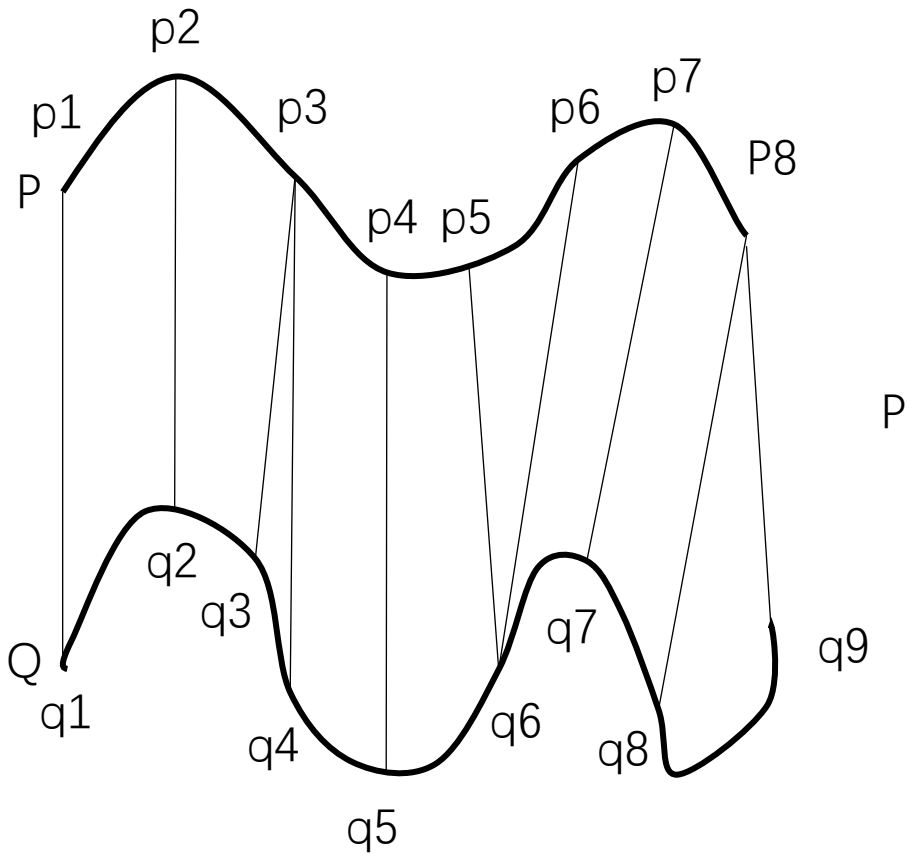
- **低资源：**特定场景、方言识别还存在低资源问题。尤其是方言数据采集相当困难。

- **语种混杂：**在日常交流中，还可能存在语种混杂现象，如中英混杂（尤其是城市白领）、普通话与方言混杂，但商业机构在这方面的投入还不多。

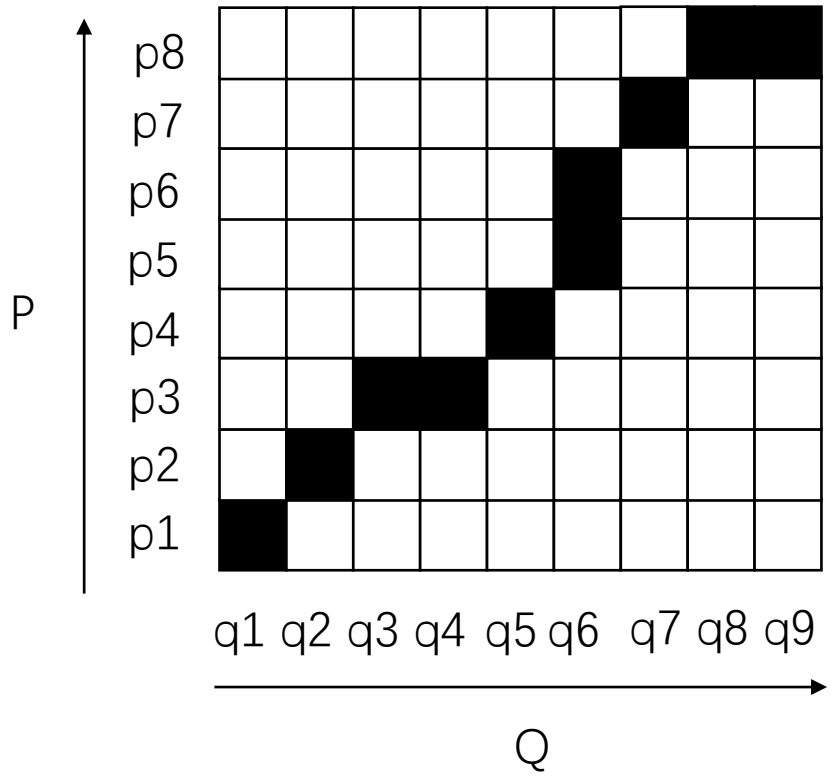
1.5 语音识别建模方法

- 模板匹配：DTW
- 统计模型：GMM-HMM
- 深度学习：DNN-HMM、End-to-End(E2E)

动态时间规整（DTW）



(a)



(b)

- DTW采用动态规划的算法思想，通过时间弯折，实现P和Q两条语音的不等长匹配，将语音匹配相似度问题转换为最优路径问题。
- DTW是模板匹配法中的典型方法，非常适合用于小词汇量孤立词语音识别系统。
- 但DTW的识别效果过分依赖于端点检测，不适合用于连续语音识别，DTW对特定人的识别效果较好。

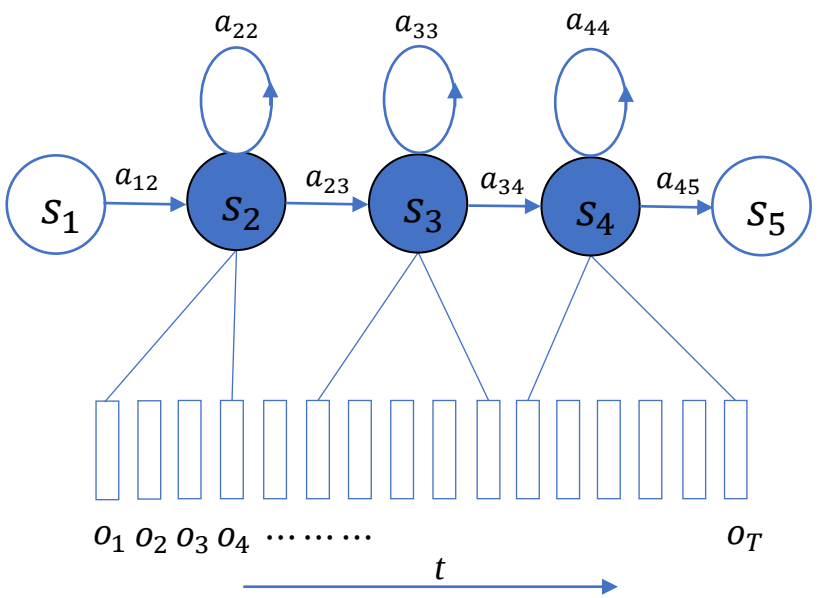
隐马尔可夫模型（HMM）

- 隐含Markov模型

  状态$\{s_1, s_2, s_3, s_4, s_5\}$隐含，
  观察值$O = o_1, o_2, \cdots, o_T$可见。

- HMM的模型要素包括：

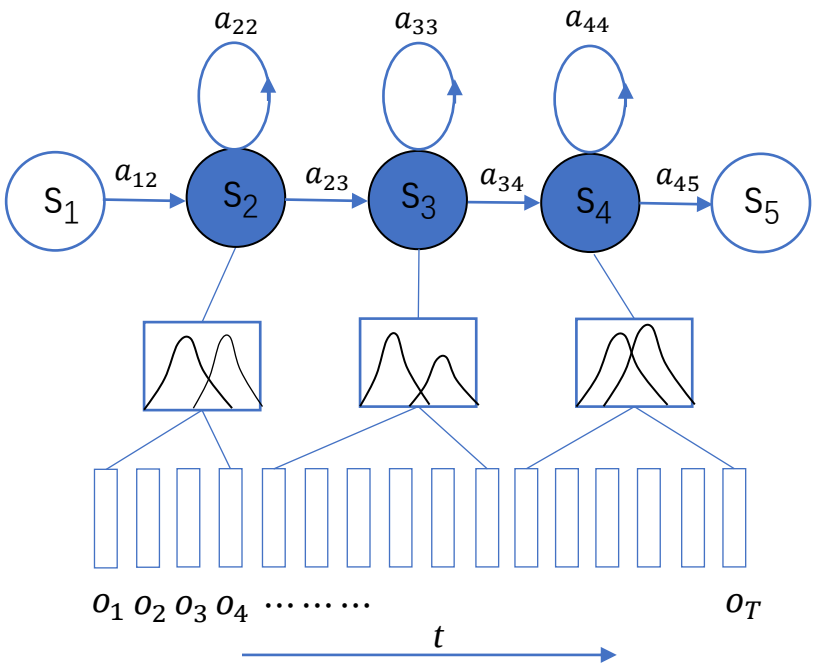
  N：模型中的状态数目
  M：每个状态可能输出的观察符号的数目
  $A = \{a_{ij}\}$：状态转移矩阵
  $B = \{b_j(k)\}$：输出概率矩阵
  $\pi = \{\pi_i\}$：初始状态概率分布



GMM-HMM

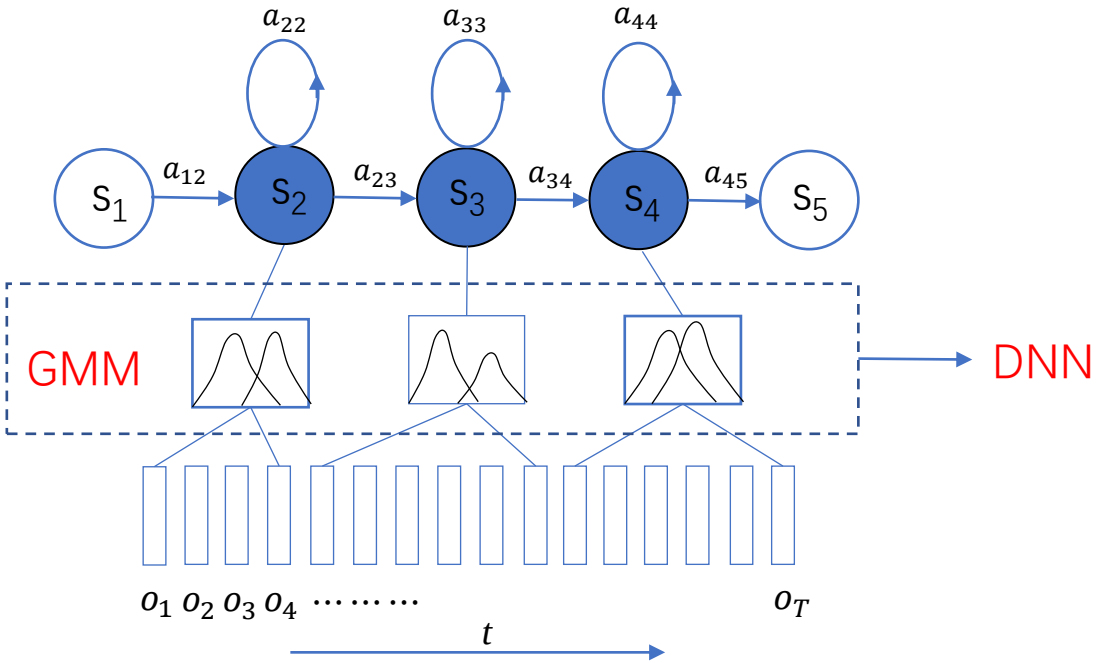
- 高斯混合概率密度函数(pdf):
  - 每个状态j表示为若干函数$N(o_t)$的线性组合
  - $N(o_t)$是连续高斯概率密度函数

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} N\left(o_t | \mu_{jk}, \Sigma_{jk}\right)$$

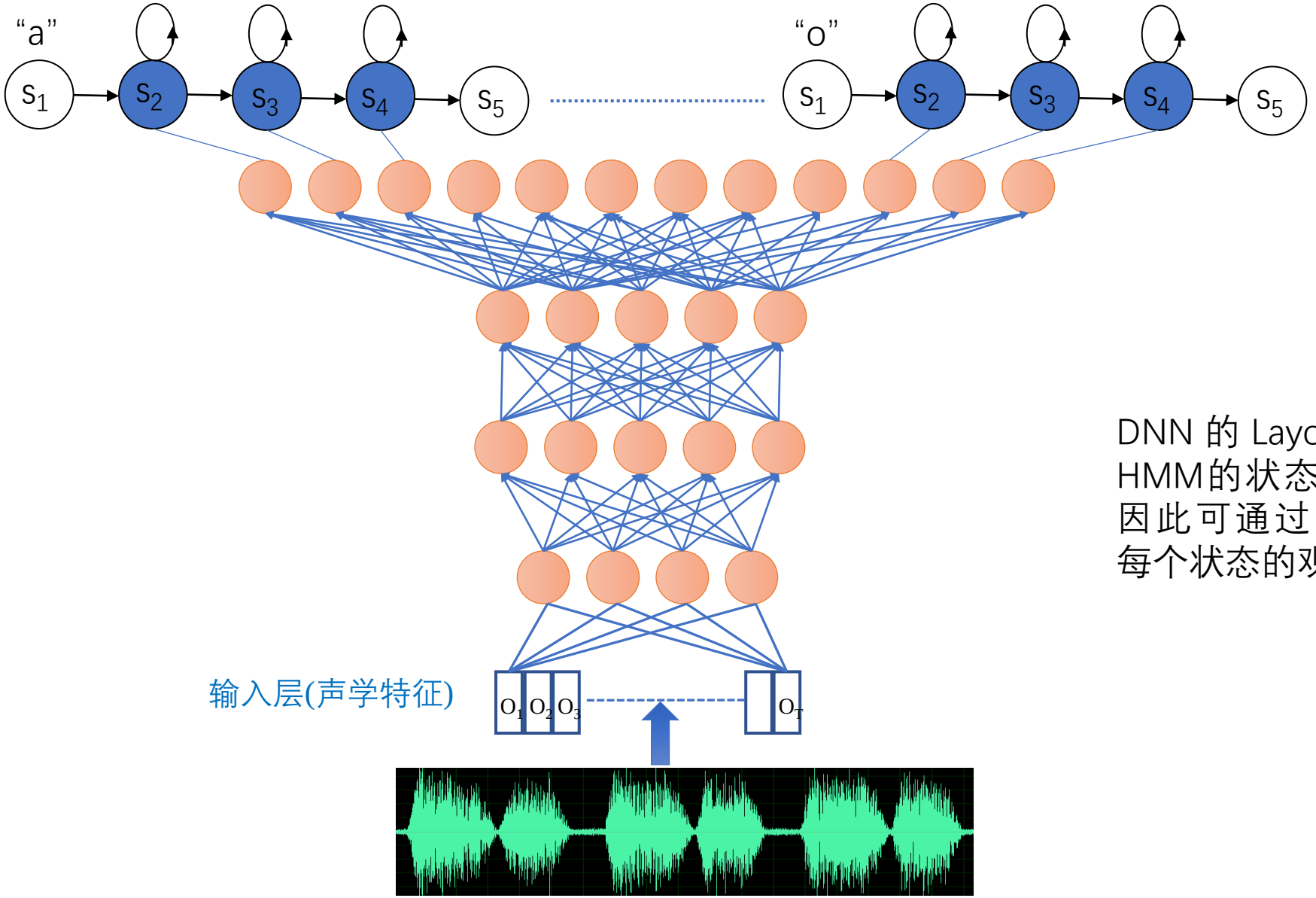


DNN-HMM

- GMM用深度神经网络（DNN）替代，HMM的转移概率和初始状态概率保持不变。
- DNN能够对包含多帧数据的特征进行处理，可以表征更丰富的语音变化特征；
- 相对于 GMM，DNN 对 HMM 中后验概率的估计不需要很苛刻的数据分布假设，条件更宽泛。

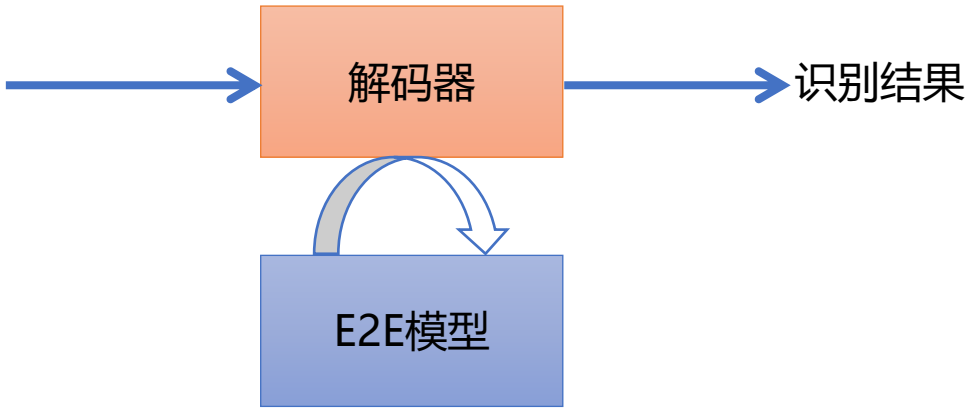


DNN-HMM



DNN 的 Layout 输出节点与HMM的状态节点一一对应，因此可通过DNN的输出得到每个状态的观察概率。

端到端(E2E)语音识别



2015年以来，端到端（E2E）模型开始流行，并被应用于语音识别领域。传统语音识别系统的发音词典、声学模型和语言模型三大组件被融合为一个E2E模型，直接实现输入语音到输出文本的转化，得到最终的识别结果。

1.6 语音识别开源工具

- HTK（HMM Toolkit）是一个专门用于建立和处理隐马尔科夫模型的实验工具包，由剑桥大学的Steve Young开创，非常适合GMM-HMM系统的搭建；2015年也推出了DNN-HMM。

- Kaldi是一个开源的语音识别工具箱，是基于C++编写的，可以在Windows和Unix平台上编译，主要由Daniel Povey博士在维护。Kaldi适合DNN-HMM系统（包括Chain模型）的搭建，支持TDNN/TDNN-F等模型，基于有限状态转换器（FST）进行解码，它也可用于x-vector等声纹识别系统的搭建。
- Kaldi的下一代版本（Next-gen Kaldi）结合Kaldi和PyTorch的优点，包含了多个子项目，其中k2是核心算法库，除了支持端到端模型，还特别实现了可微分的FST，其目标是在端到端模型进行LF-MMI区分性训练。

- ESPnet是一个端到端语音处理工具集，它将PyTorch作为主要的深度学习引擎，兼容Kaldi风格的数据处理方式，为语音识别、语音合成和其他语音处理实验提供完整的设置，包括CTC/Attention、RNN-T等模型。

- WeNet是出门问问联合西北工业大学音频语音与语言处理研究组推出面向产品和工业界的端到端语音识别开源工具，致力于消除从端到端模型研究到产品落地中的鸿沟，探索更适合工业级产品的端到端解决方案。

1.7 语音识别常用数据库

- **TIMIT**—经典的英文语音识别库，包含630人，来自美国8个主要口音地区，每人10句，包含了词和音素级的标注。
- **SwitchBoard**—对话式电话语音库，采样率8kHz，包含来自美国各个地区543人的2400条通话录音。
- **LibriSpeech**—免费的英文语音识别数据库，总共1000小时，采样率16kHz，包含朗读式语音和对应的文本。
- **Thchs-30**—清华大学提供的一个中文示例，并配套完整的发音词典，其数据集有30小时，采样率16kHz。
- **AISHELL-1**—希尔贝壳开源的178小时中文普通话数据，采样率16kHz。包括400位来自中国不同口音区域的发音人，语料内容涵盖财经、科技、体育、娱乐、时事新闻。
- **WenetSpeech**—由西北工业大学音频语音与语言处理研究组、出门问问、希尔贝壳联合发布，包含超过10000小时的高质量标注数据、2400多小时的弱标注数据，以及约10000小时的无标注数据，覆盖各种互联网音视频、噪声背景条件、讲话方式，来源领域包括有声书、解说、纪录片、电视剧、访谈、新闻、朗读、演讲、综艺和其他10大场景。

1.8 语音识别评价指标

假设“我们明天去动物园”的语音识别结果如下：

原句：	**我们**	**明天**	**去**	**动物园**
识别：	**我×**	**明后天**	**去**	**公园×**

删除错误 Deletion　**插入错误 Insertion**　**替换错误Substitution**

度量语音识别性能的指标有许多，通常是使用在测试集上的**词错误率（Word Error Rate，WER）**来判断整个系统的性能，其公式定义如下：

$$WER = \frac{N_{Del} + N_{Sub} + N_{Ins}}{N_{Ref}}$$

其中，N_{Ref}表示测试集所有的词数量，N_{Del}表示识别结果相对于实际标注发生删除错误的词数量，N_{Sub}代表发生替换错误的词数量，而N_{Ins}则表示发生插入错误的词的数量。

针对中文普通话，评价指标也经常采用**字符错误率（CER）**，即用单字而不是词，来计算错误率。

技术体验

厦门大学
XIAMEN UNIVERSITY

普通话识别Android版

AI语音合成小程序

声纹识别Android版本

闽南话识别Android版

闽南话识别小程序

说咱闽南话APP

网页版演示：https://tc.talentedsoft.com:58120/speech_stream_recognition

Thank you!

厦门大学智能语音实验室

speech.xmu.edu.cn