

Chapter 12

Automatic Speech Recognition

语音识别

主要内容

- 语音识别概述
- 三个要素问题
 - 声学模型建模
 - 语言模型建模
 - 解码搜索
- 其他**ASR**相关问题

语音识别概述

语音识别简史

- 语音识别是一门交叉学科
 - 涉及声学、生理学、心理学、信号处理、模式识别、人工智能、语言学、计算机科学等多学科
- 语音识别历史悠久且丰富
 - 1920年代Radio Rex玩具狗：最早的语音识别器(狗的名字)
 - 1952年：Bell Lab实现特定人英文数字识别系统(共振峰)
 - 1960年代
 - 动态规划（DP）算法对齐不同长度语音(Vintsyuk)
 - CMU的Reddy开始进行连续语音识别（CSR）的开创性工作



语音识别简史

- 1970年代

- 模式识别、动态规划、线性预测基础研究发展
- IBM展开大词汇量连续语音识别（Jelinek）
 - 备忘录听写系统Tangorn
- Bell Lab展开非特定人识别系统研究
- DARPA介入语音识别领域:语音理解研究计划
 - 1973年CMU给出演示系统
 - 词典扩展到“千”级别
 - 引入图搜索的概念

语音识别简史

- 1980年代：孤立词识别→连接词识别
 - 建模方案：基于模板方法→统计建模框架
 - HMM理论和应用趋于完善
 - 语音识别相关技术持续进步
 - 一阶二阶差分系数 (Furui)
 - N-gram语言模型 (Jelinek)
 - 引入神经网络 (Hinton, Katagiri)
 - DARPA推出LVCSR系统研究计划
 - CMU的SPHINX系统 (Kai-Fu Lee)
 - BBN的BYBLOS系统 (Chow Y)
 - SRI的DECIPHER系统 (Weintraub)

语音识别简史

- 1990年代

- 模式识别领域涌现大量创新工作

- 贝叶斯框架下的分布估计→最小化经验错误准则

- 语音识别技术继续发展

- 区分性训练: MCE&GPD (B-H Juang)

- MMI (P. Brown, Bahl)

- 模型自适应方法: MLLR (PC Woodland)

- MAP (C-H Lee)

- 噪声鲁棒性问题研究

- DARPA和NIST联合推出更多语音识别评测任务

- RM / NAB / WSJ / Switchboard

- 剑桥推出HTK (Hidden Markov Model Toolkit)并公开化, 大大降低语音识别门槛

语音识别简史

- 20世纪：语音识别研究向广度和深度发展
 - 语音识别技术：局部改进+颠覆式创新
 - 更好的区分性训练方法(Povey等)
 - 更好的语言模型建模方法 (NN LM)
 - 鲁棒性、置信度、自适应等技术
 - 革命性变革：**深度神经网络**（微软等）
 - DARPA推出EARS和GALE计划
 - **语音识别核心技术整合程度空前**
 - 语音识别技术实用化：新的应用和产品
 - Google Voice Search; Siri
 - 语音输入法和语音助理类产品
 - 多模态语音识别

EARS (Effective Affordable Reusable Speech-to-Text): 放宽待转写音频的范围（包括多信道、多语种），对自动转写精度要求提高，希望人们可以不用听音频信号就可以直接理解转写内容。

GALE(Global Autonomous Language Exploitation): 关注多语言的语音和文本分析。

语音识别分类

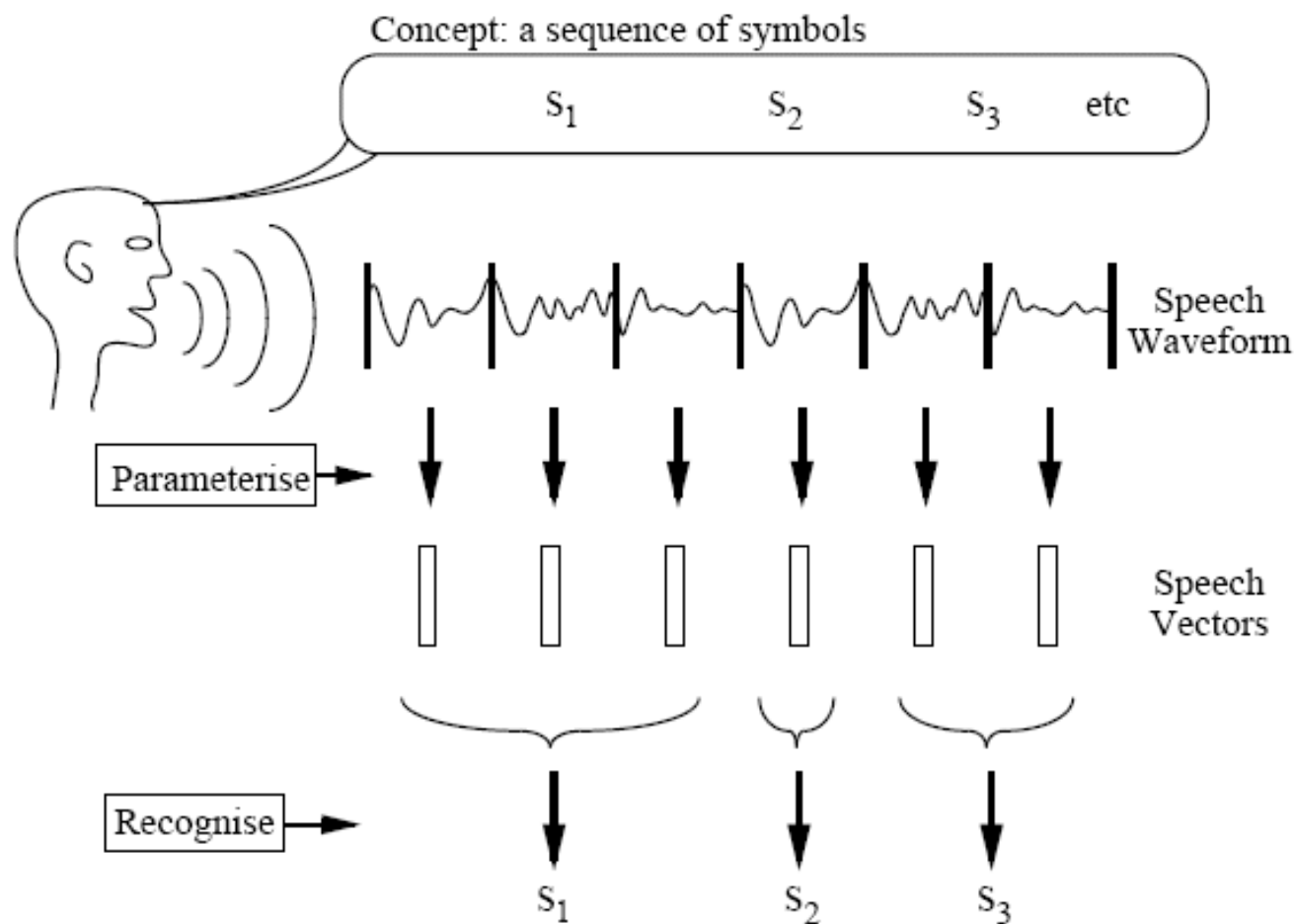
- Isolated vs. continuous ASR
 - Isolated(孤立词): 每个词发音之间要有停顿
 - Continuous(连续识别): 发音不需要停顿
- Small vs. medium vs. large vocabulary
 - 命令控制 / 语音呼叫导航 / 语音输入法
- 发音风格: 朗读/自由发音
- 多语言/口音/方言

语音识别学习和研究工具

- HTK
Hidden Markov Model Toolkit
- Kaldi
<https://github.com/kaldi-asr/kaldi>
Hybrid HMM
- Espnet
<https://github.com/espnet/espnet>
End-to-End Speech Processing Toolkit

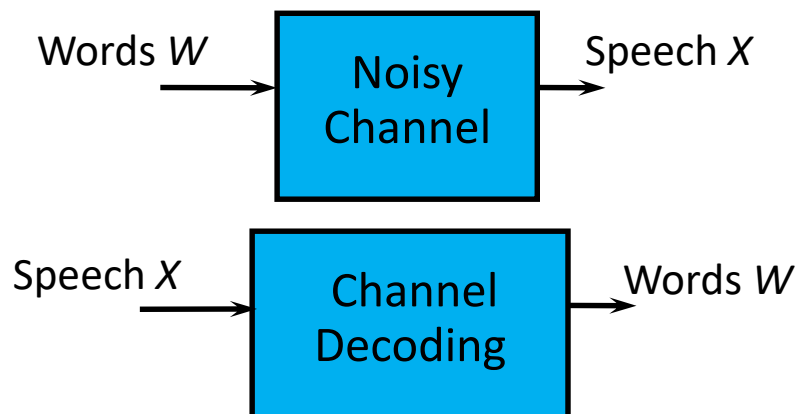
自动语音识别 (ASR)

- 语音识别的过程



语音识别理论描述

- 语音识别可以看成是一个(噪声)信道解码问题，或模式分类问题



- 语音识别问题的理论解决方案

– **plug-in MAP decision rule**

$$\hat{W} = \operatorname{argmax}_{W \in \Omega} p(W | X) = \operatorname{argmax}_{W \in \Omega} P(W) \cdot p(X | W)$$

$$= \operatorname{argmax}_{W \in \Omega} \bar{P}_T(W) \cdot \bar{p}_\Lambda(X | W)$$

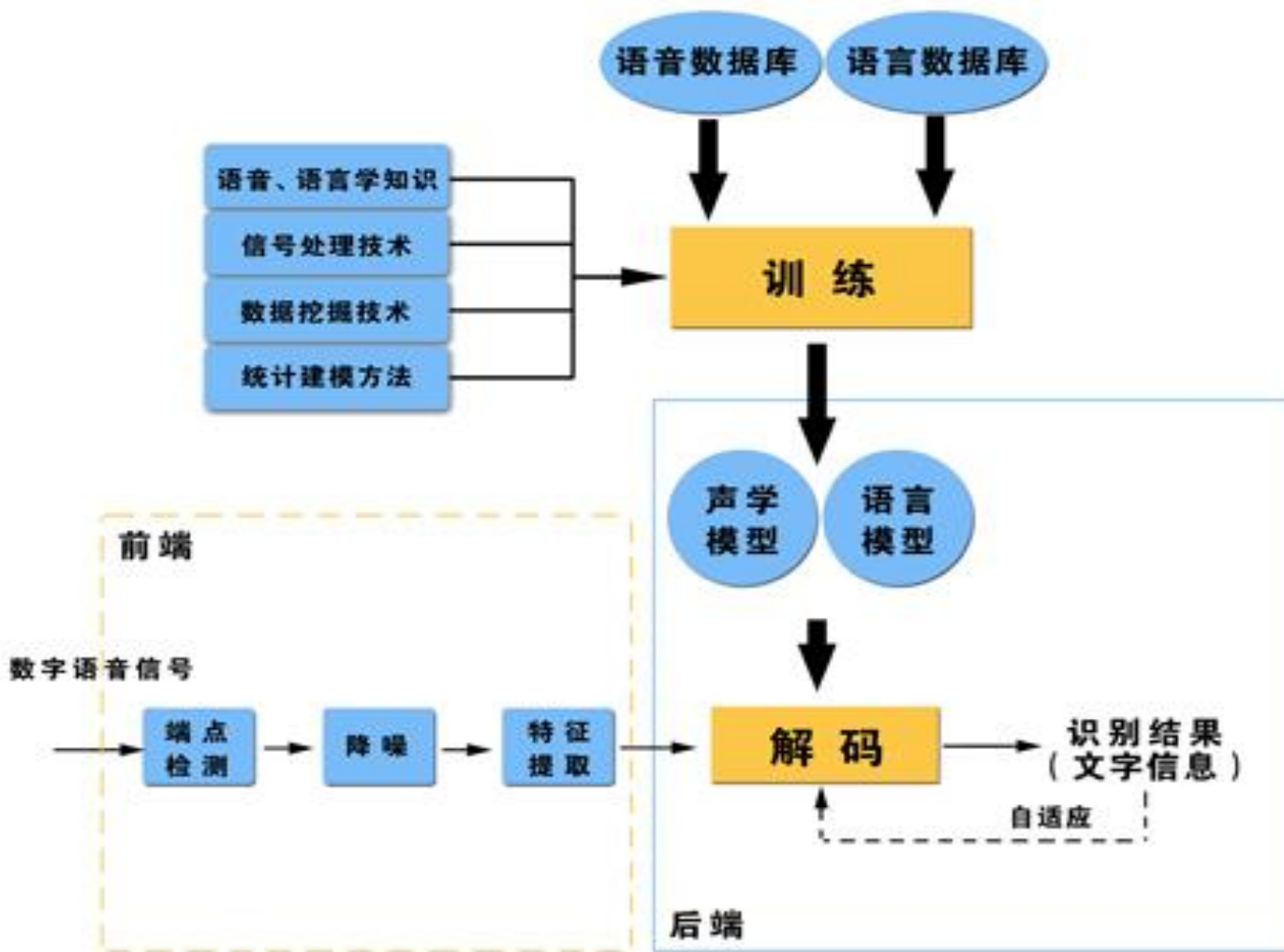
语音识别理论描述

$$\hat{W} = \operatorname{argmax}_{W \in \Omega} p(W | X) = \operatorname{argmax}_{W \in \Omega} P(W) \cdot p(X | W)$$

$$= \operatorname{argmax}_{W \in \Omega} \bar{P}_{\Gamma}(W) \cdot \bar{p}_{\Lambda}(X | W)$$

- $\bar{p}_{\Lambda}(X | W)$ ---- 声学模型 (Acoustic Model, AM)
 - 词W发音时生成特征X的概率
 - 需要考虑建模单元、模型选择等问题
- $\bar{P}_{\Gamma}(W)$ ---- 语言模型 (Language Model, LM)
 - (人们) 会说某词或词序列的概率
 - 需要灵活的模型来对所有可能出现的词组合进行描述
- 解码搜索空间 Ω

语音识别框架图



ASR问题中最重要的三个要素

- **声学模型建模**

- 如何选择合适的语音建模单元，并利用可用的语音数据可靠、高效的估计相应HMM参数

- **语言模型建模**

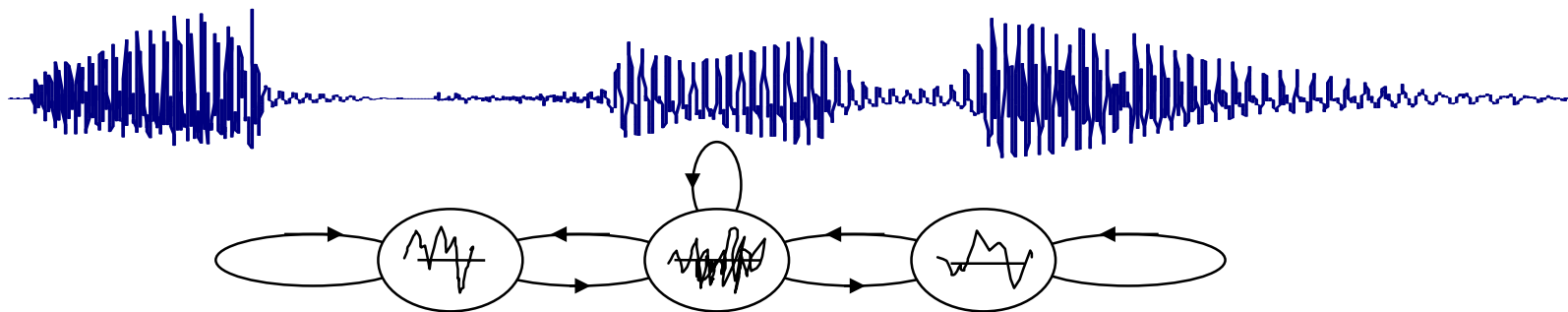
- 如何基于文本训练数据合理估计n-gram语言模型，并较好的处理数据稀疏问题

- **解码搜索**

- 给定声学HMM模型和语言n-gram模型，如何从一个大语法网络中高效的找出最优路径
 - 语音识别的搜索空间非常巨大，迫切需要高效的裁剪策略

声学模型建模

HMM: 一种适用于语音识别的声学模型

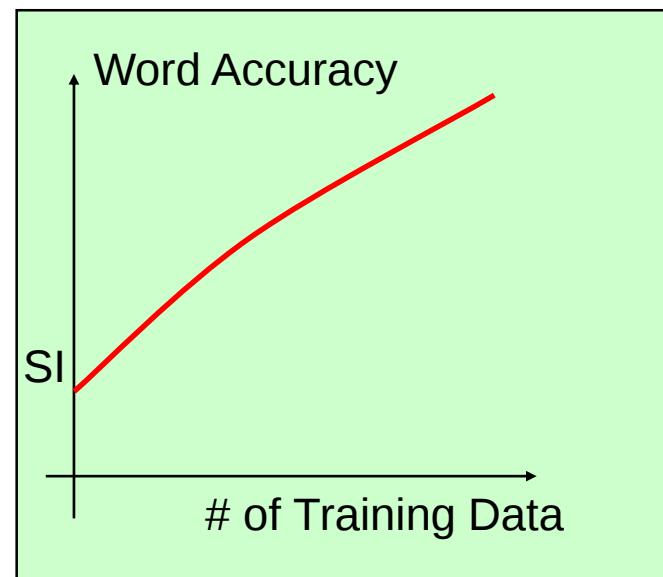
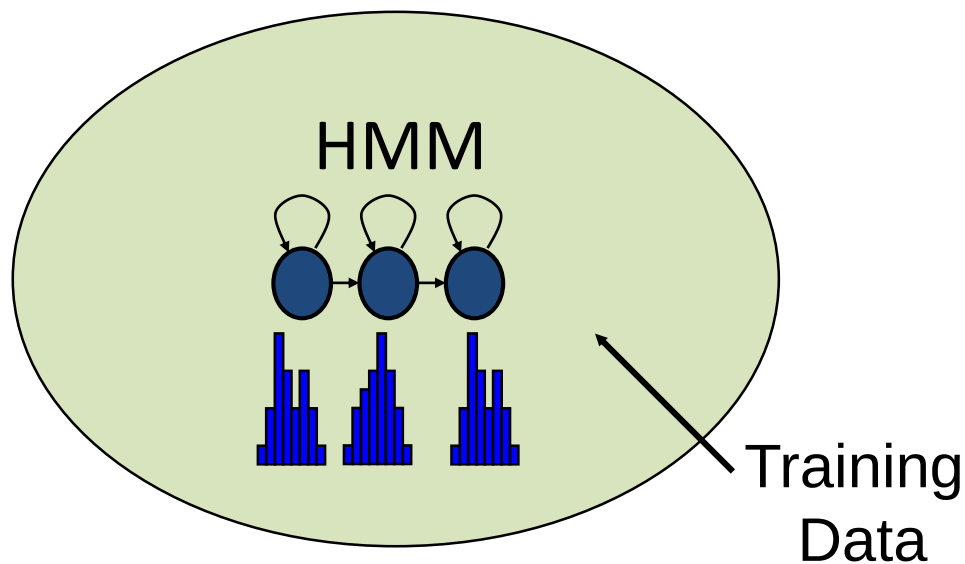


- HMM: Hidden Markov Model
 - 隐马尔科夫模型
- 语音信号的多变性
 - temporal(时域) & spectral(频域)
- HMM可以同时描述时域和频域的变化

建模单元选择

- whole vs. sub word
 - 整词建模：对于每个词采用一个对应的HMM
 - 需要收集所有词对应的语音数据
 - 受限制较多：无法在不同词之间共享数据；扩展(加词)困难
 - 子词单元建模：对于每个phone或者syllable构建相应HMM
 - 能较好的解决上述问题，共享性较好
 - 不易解决协同发音(co-articulation)问题→上下文相关sub-word建模(bi-phone, tri-phone, ...)

声学模型建模及系统识别性能



在一个典型系统中，每个音素 (phoneme) 一般用3-5个状态、自左向右的连续概率高斯混合HMM来描述，而背景噪声一般用1-3个状态的HMM来描述

在当今语音识别系统中，成千上万小时的语音数据用于训练HMM模型

HMM训练

- 最大似然（Maximum Likelihood）准则
 - 通过最大化 $P(\mathbf{o}|\boldsymbol{\lambda})$ 优化求解模型参数
 - Baum-Welch算法，迭代更新模型参数
- 区分性训练
 - 以分类准确性作为模型训练目标
 - Minimum Classification Error (MCE)
 - Minimum Phone Error (MPE)
 - Maximum Mutual Information (MMI)

语言模型建模

语言模型

- 什么是语言模型？
 - 通俗的讲：一种判断文本是否合理的方式
 - 中国科大 vs. 科大中国
 - A cat on a desk vs. A desk on a cat
 - 科学的讲： $W \rightarrow \bar{P}_r(W)$
- 语言模型的常用方法
 - Bag of words (multinomial model)
 - Multinomial mixture model (mixture model)
 - Latent Dirichlet Allocation (Bayesian model)
 - **N-gram (Markov Chain model)**
 - **Neural network (RNN)**

N-gram语言模型

- N-gram语言模型本质上属于Markov Chain Model
 - 有限历史假设：当前词的条件概率仅与前N-1个词（历史）有关
- 给定词序列 $W=w_1, w_2, \dots, w_M$, LM概率可以表示为

$$P(W) = P(w_1, w_2, \dots, w_M) = \prod_{m=1}^M p(w_m \mid h_m)$$

- $h_m = w_{m-N+1}, \dots, w_{m-1}$ 表示 w_m 的历史
- 对于unigram, $h_m = \text{null}$ (参数数目 $\sim |V|$, $|V|$ 词典大小)
- 对于bigram, $h_m = w_{m-1}$ (参数数目 $\sim |V| * |V|$)
- 对于trigram, $h_m = w_{m-2}w_{m-1}$ (参数数目 $\sim |V| * |V| * |V|$)
- 对于4-gram, $h_m = w_{m-3}w_{m-2}w_{m-1}$ (参数数目 $\sim |V| * |V| * |V| * |V|$)

语言模型的评估指标

- **Perplexity(混淆度)**是最常用的评估LM性能的方式
- Perplexity定义
 - 给定一个词典规模为 $|V|$ 的LM- $\{P(\cdot)\}$ ，以及一条足够长的测试词序列 $W=w_1, w_2, \dots, w_M$
 - Perplexity定义为**词串概率几何平均的倒数**

$$PP = [P(w^M)]^{-\frac{1}{M}} = \left[\prod_{m=1}^M p(w_m \mid w_{\max[m-N+1, 1]}^{m-1}) \right]^{-\frac{1}{M}}$$

- Perplexity的含义
 - PP值越小，则说明LM的预测能力越好

LM中的词典选择

- 词典并不是越大越好
 - N-gram组合数目以及LM模型参数随词典规模指数级增长
 - 需要更多的训练数据和运算资源
- 如何控制词典规模
 - 根据训练语料中的词频确定
 - 在词典规模一样的前提下，该方法形成的集外词(Out-of-Vocabulary , OOV)比例最低
- 举例：英文Wall Street Journal任务
 - 文本训练集：来自3年报纸的37 million词
 - 词典规模：2万词
 - OOV rate: 4%
 - 2-gram PP 114; 3-gram PP 76

语言模型训练中的数据稀疏问题

- 数据稀疏(Data Sparseness)问题普遍存在
 - 导致ML估计不准确: **ML中, 0频率”意味着“0概率”**
 - 举例: 在1.2million词的训练文本中(词典规模1000)
 - 20%的bi-gram和60%的tri-gram仅仅出现一次
 - 85%的tri-gram出现少于5次
- 如何解决数据稀疏问题?
 - 加数据? 仅通过获取更多训练数据不能解决该问题
 - 在自然语言中, n-gram的分布是极其不均匀的
 - 当数据达到一定数量后, 通过加数据来减少OOV或者新N-gram的速度大大减慢
 - 需要更好的参数估计策略(对ML估计进行平滑处理)
 - back-off策略: discounting + redistribution
 - 线性插值策略

解码搜索

解码问题

$$\begin{aligned}\hat{W} &= \operatorname{argmax}_{W \in \Omega} P(W \mid X) = \frac{\operatorname{argmax}_{W \in \Omega} P(W) \cdot P(X \mid W)}{P(X)} \\ &= \operatorname{argmax}_{W \in \Omega} \bar{P}_{\Gamma}(W) \cdot \bar{P}_{\Lambda}(X \mid W)\end{aligned}$$

- $\bar{P}_{\Lambda}(X \mid W)$ --- 声学模型 (Acoustic Model, AM)
- $\bar{P}_{\Gamma}(W)$ --- 语言模型 (Language Model, LM)
- Ω --- 解码搜索空间

Viterbi Search

- 识别问题本质上是一个Viterbi Search问题
 - 整个网络看成一个“组合式”HMM: Λ
 - 对于输入的语音（特征） X ，我们需要遍历整个语法网络寻找最优的状态路径 S^*

- Viterbi path

$$S^* = \operatorname{argmax}_{S \in \Theta} \Pr(S) \cdot p(X \mid S, \Lambda)$$

$$= \operatorname{argmax}_{S \in \Theta} \Pr(W_S) \cdot p(X \mid S, \Lambda)$$

- 在获得最优状态路径 S^* 后，可通过回溯 (backtracking) 的方式获得相应的词序列，以完成识别过程

Viterbi Search

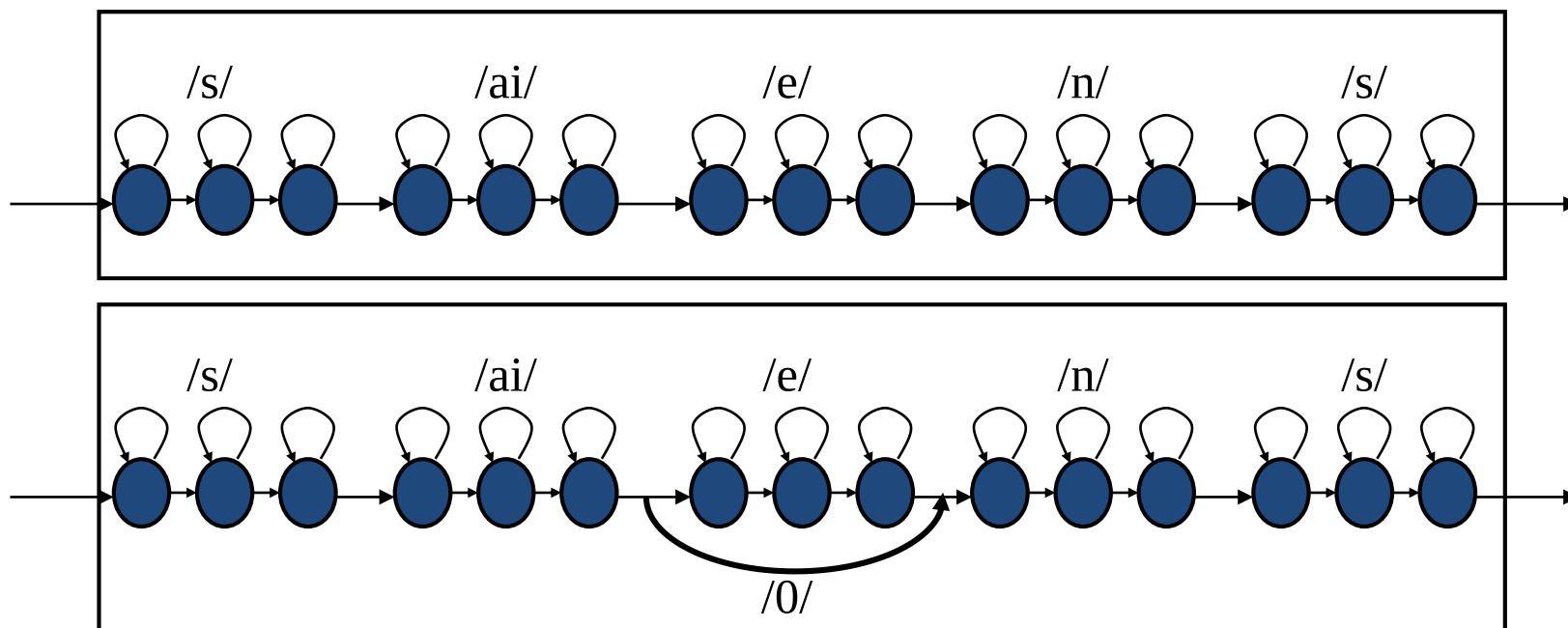
- 解码空间——所有可能的识别结果
 - 最自然的想法：逐一列举所有可能的词序列
 - 解决方案：建立一个包含所有可能词路径的解码网络，识别就是在解码网络中搜索最佳路径
- 建立HMM描述的音素声学模型到解码空间中词序列的映射关系
 - Lexicon Modeling
 - Word Juncture Modeling
 - Grammar Network

Lexicon Modeling

- 如何将“词”和“发音”对应起来的问题
 - 假设每个HMM是一个上下文无关的单音素模型 (context-independent mono-phone model)
 - 对于英语来说：42个mono-phone对应42个HMM模型
 - 每个词由mono-phone序列连接组成
 - Lexicon: /science/= /s/+ /ai/+ /e/+ /n/+ /s/
或 /s/+ /ai/+ /n/+ /s/
 - 实际应用中可能存在多发音现象

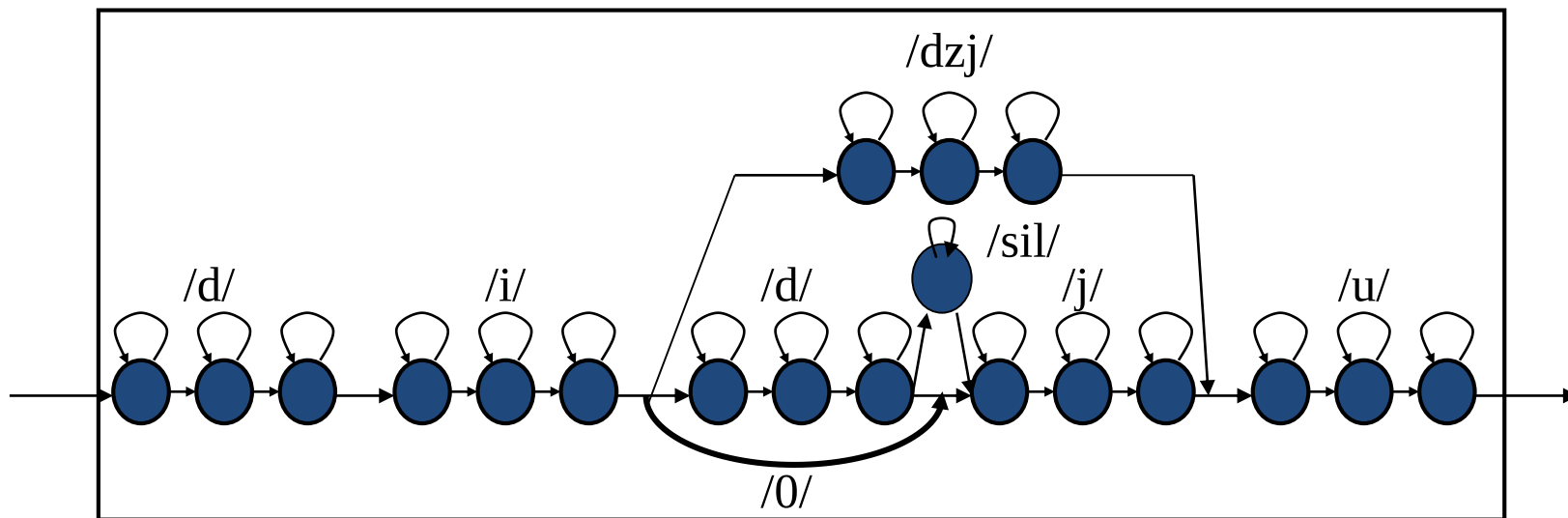
Lexicon Modeling

- 如何将“词”和“发音”对应起来的问题
 - 假设每个HMM是一个上下文无关的单音素模型 (context-independent mono-phone model)



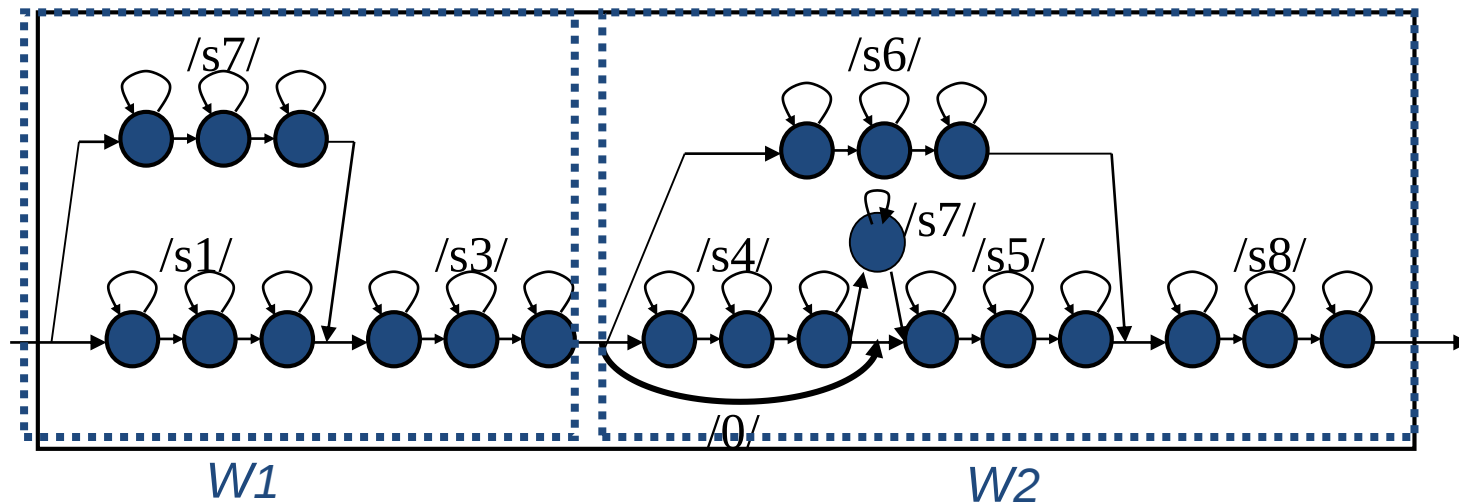
Word-Juncture Modeling

- 协同发音现象 (Co-articulation effect)
 - 语音识别错误的重要来源之一
 - 硬处理: "did you" = /d/+ /i/+ /dzj/+ /u/
 - 软处理: 在原有phone序列基础上加入部分可变部分
 - 对于有明确音节(Syllable)边界的语种较容易处理 (例如中文、日语、意大利语等)



Grammar Network

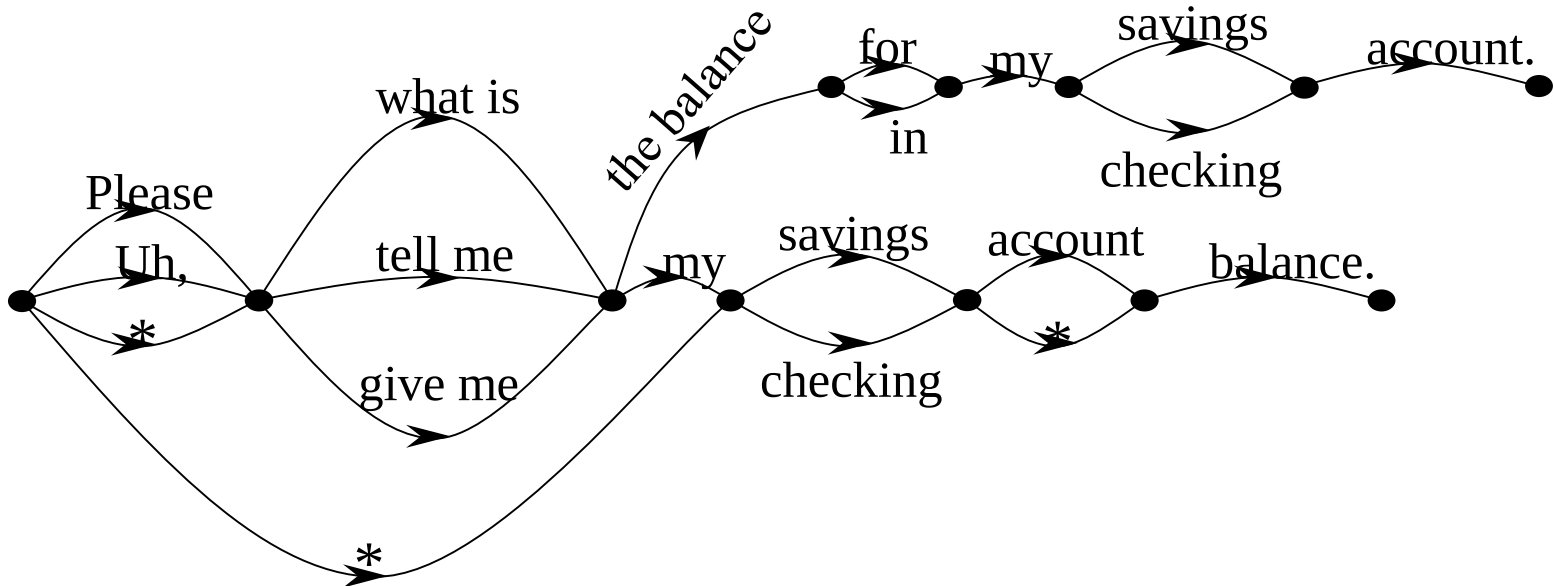
- 词 → 词序列 → 更进一步



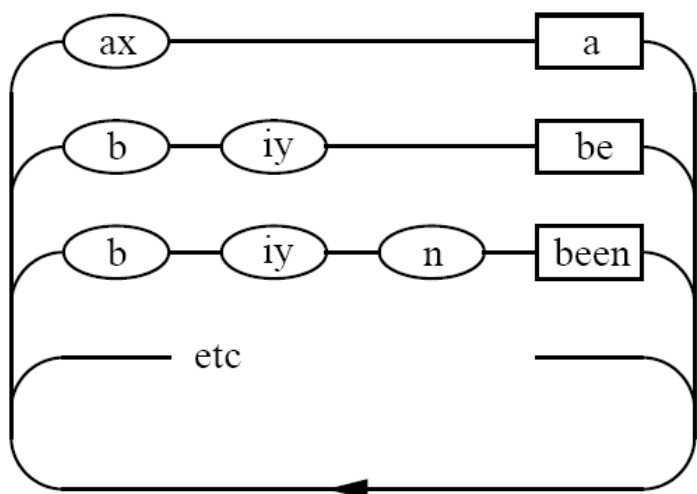
- 语法模型（语法网络）：用一个大的“**组合式**” HMM网络来表示所有可能、合法的词序列
 - 用有限状态语法(Finite State Grammar)网络来表示词之间的关系（及限制）
 - 对于大词汇量来说语法网络会很大

有限状态语法样例

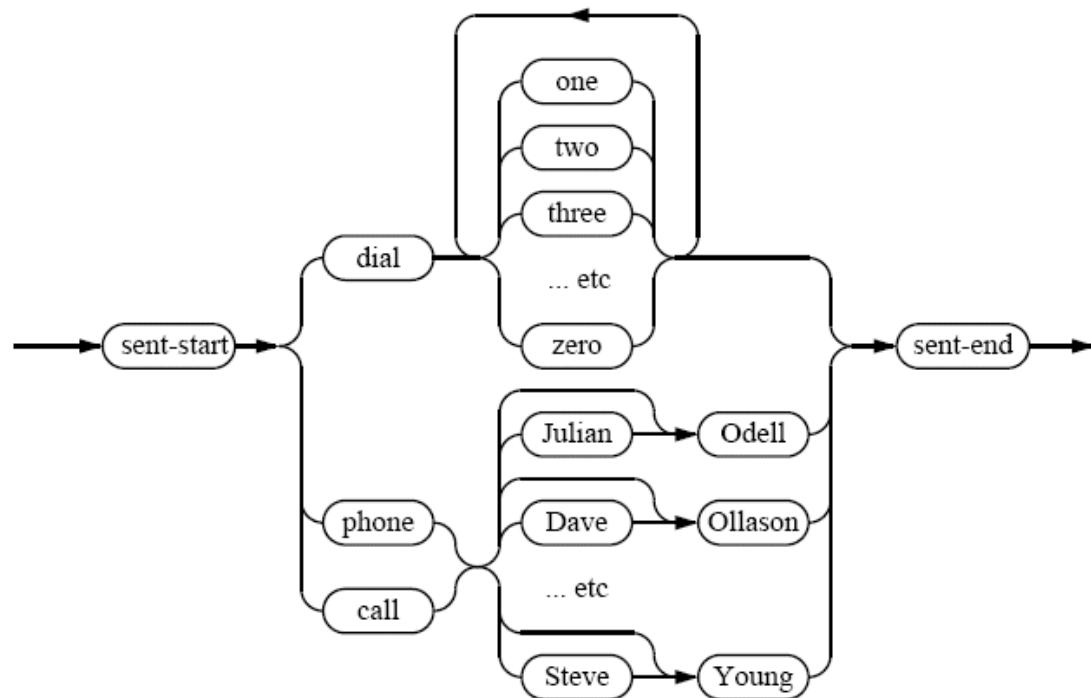
- 用有限状态语法表示一个账户咨询任务
 - 每条弧(arc)表示一个词或者短语
 - “*” 代表可跨越的空弧
 - “Please tell me my checking account balance” 是其中的一条合法路径



其他格式的语法网络



“词循环”语法网络



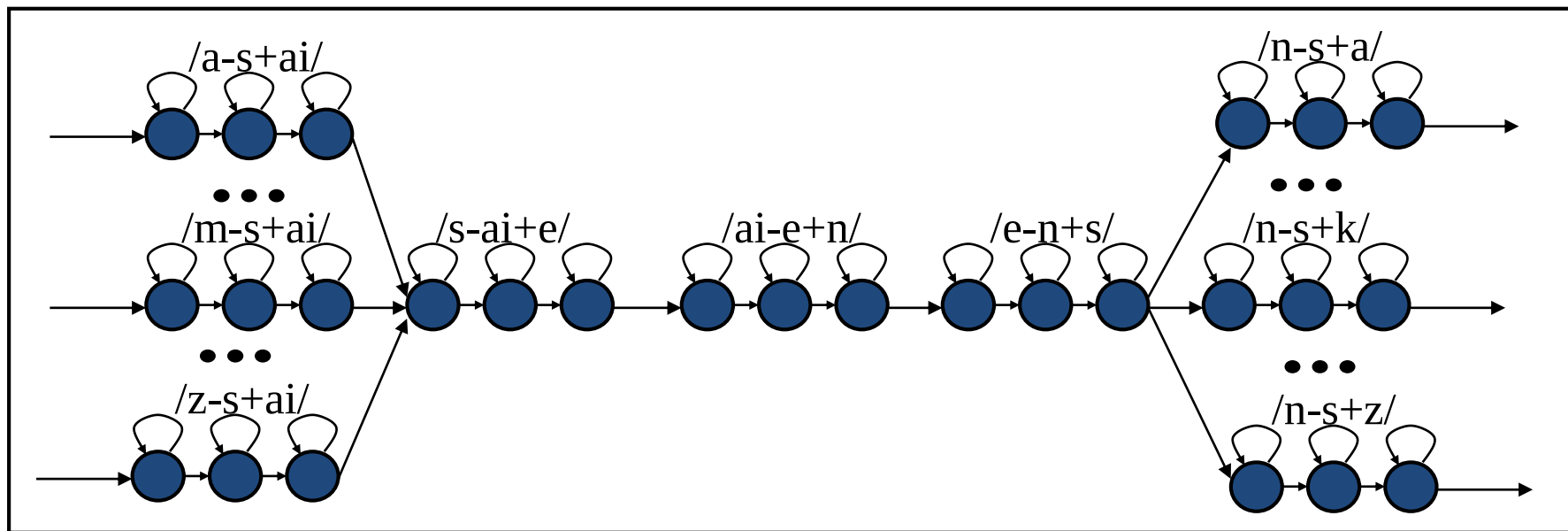
“打电话拨号”语法网络

Tri-phone (Bi-phone) 建模

- Mono-phone建模无法精细模拟协同发音现象
- 需要引入上下文相关的建模方式
 - Bi-phone, tri-phone, etc...
 - 对于英文来说, 有 $42*42*42=74088$ 个tri-phones
- 使用上下文相关模型会让语法网络及其扩展变得复杂

Tri-phone (Bi-phone) 建模

- Mono-phone建模无法精细模拟协同发音现象
- 需要引入上下文相关的建模方式



其他ASR相关问题

其他ASR相关问题

- **端点检测** (Voice Activity Detection)
 - 定义：检测出输入语音的有效端点
 - 目的：提升识别准确性和解码有效性
 - 主流方法
 - 基于能量的：能量双门限或四门限方法
 - 基于模型的：GMM模型 / DNN模型
 - 更普适性的问题：Diarization
 - 语音段包含多个说话人、各种干扰语音
 - 噪声检测及去除、端点检测、说话人分离

其他ASR相关问题

- 声学特征提取

- 好的声学特征应具备的属性
 - 对声学建模单元具有良好的区分性
 - 特征维数要适中
 - 具备一定“抗干扰性”：环境、信道、说话人
- 两大经典特征：MFCC、PLP
- 提升特征性能的其他方法
 - 特征规整：CMN、MVN
 - 差分等特征扩展方案：一阶二阶差分
 - 区分性特征：fMPE、TANDEM
 - 特征降维技术：LDA、PCA
 - 噪声鲁棒性特征：RASTA、VTS

其他ASR相关问题

- 噪声鲁棒性问题

- 如何提升噪声环境下的识别效果？

- 环境噪声和信道畸变

- 主要方法

- 鲁棒性特征：PLP+RASTA, PNCC

- 信号和特征增强：维纳滤波, CMN, VTS

- 模型补偿：PMC, MAP, MLLR

- Multi-Training

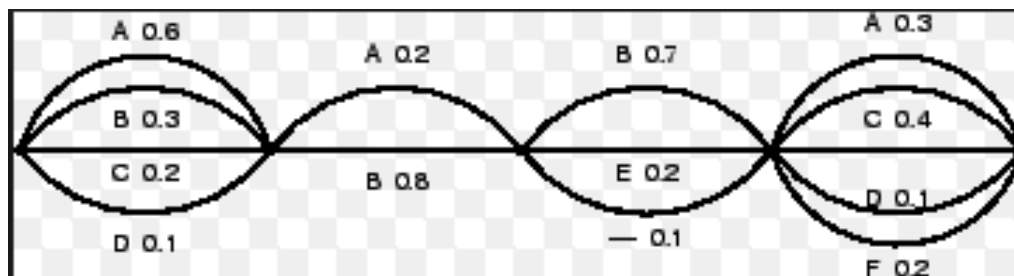
- 其他方法：基于麦克风阵列的降噪

其他ASR相关问题

- **置信度判决** (Confidence Measure)
 - 定义：对识别结果可靠性的判断
 - 用途：
 - 检测集外词 / 数据挑选 / 多候选
 - 主流方法
 - 基于特征及特征组合
 - 基于后验概率
 - 基于似然比检验
 - 参考文献
 - Confidence Measure for Speech Recognition: A Survey, Hui Jiang

其他ASR相关问题

- 系统融合 (System Combination)
 - 基于多个系统产生的多个识别结果，进一步提升识别效果
 - 各系统的识别结果具有互补性：特征 / 框架
 - 用于系统融合的各系统性能一般差别不大
 - 常用方法
 - ROVER：基于1-best的融合
 - CNC：基于n-best的融合



影响ASR识别效果的因素

- 说话人发音方式

- 发音不连贯、发错音等现象

- 犹豫、重读、语气词

- 连音现象

- this supper = this upper

- 说话人特性

- 语速、音量大小、口音方言、是否配合等

- 文字内容

- 同音字现象: blue <-> blew

- 声学层面混淆: 黄<->王

影响ASR识别效果的因素

- **环境和信道影响**
 - 信道失真
 - 麦克风录音效果
 - 各种环境噪声
- **普适性问题**
 - 特定受限领域内容能获得较好效果
 - ASR要做到普适性很难（LM覆盖问题）
- **硬件限制**
 - 原则上模型越大越好、数据越多越好
 - 硬件运算资源和内存等限制模型大小

ASR系统效果评估准则

- 词（字）正确率（Word Accuracy）

$$\text{Word Accuracy} = 1 - \frac{\#sub + \#ins + \#del}{\#words \text{ in correct trascriptions}}$$

- 在连续语音识别中，如何统计各类错误(替换/插入/删除)?
- Minimum Edit Distance 准则：最小化总体的“替换+插入+删除”错误

- 句子正确率：一句话所有词都识别正确才算对

- 语义正确率：

- 正确的理解了一句话的意思；实现了正确的动作；对所有重要的语义信息内容都识别正确

DARPA ASR Benchmark

