知乎 LIVE

语音识别技术的前世今生

王赟 (Maigo)
PhD @ Carnegie Mellon University
2017.5.19

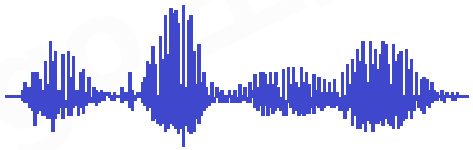
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

什么是语音识别？

- 语音识别：把语音转换成文字



OK, Google

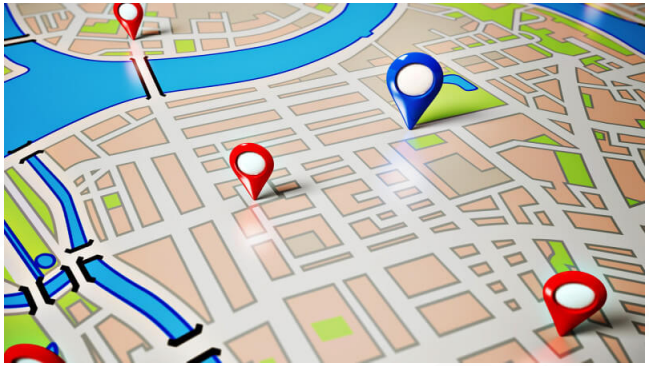
- 相关课题：
  - 元数据识别：语种、说话人、情感等
  - 语音增强与分离
  - 语音合成与转换
  - 自然语言理解、对话系统

前世：GMM + HMM
- 孤立词识别
  - 特征提取
  - 动态弯算法
  - GMM
  - HMM
  - EM训练算法
  - 语音识别基本方程
- 连续语音识别
  - 语言模型
  - 大词汇量
  - 语音识别系统结构
  - 评价指标：WER
- 潘多拉魔盒
  - 上下文有关模型
  - 区分式训练
  - 说话人适应
  - 二次打分

今生：神经网络
- 前馈神经网络
  - Tandem结构
  - Hybrid结构
- 循环神经网络
  - CTC
  - Grapheme系统
  - 注意力机制

现状与未来

语音识别的应用



语音识别之前世

—— GMM + HMM

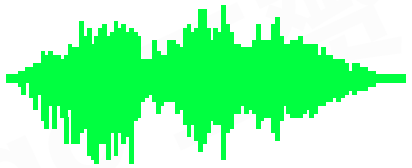
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

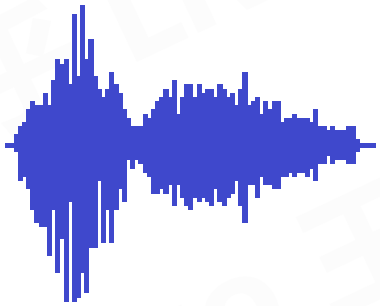
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

孤立词识别

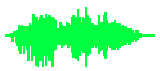


Yes



?

No

- 模板比较法
  - 计算距离：$d($  ,  $), d($  ,  $)$
  - 距离小者为识别结果
- 直接比较波形吗？

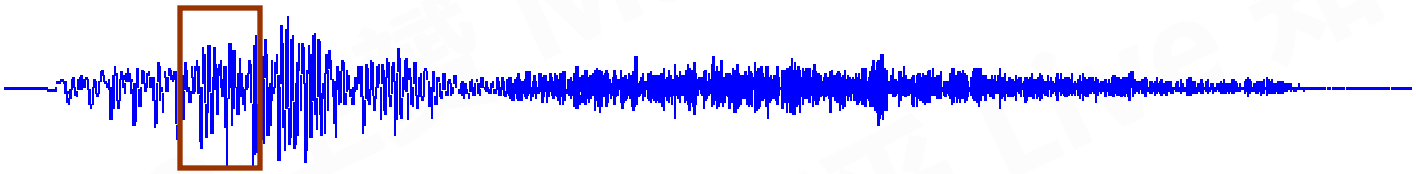
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

特征提取



- 一帧信号，通常为 20 ~ 50 ms
  - 微观上足够长：至少包含 2 ~ 3 个周期
  - 宏观上足够短：在一个音素之内

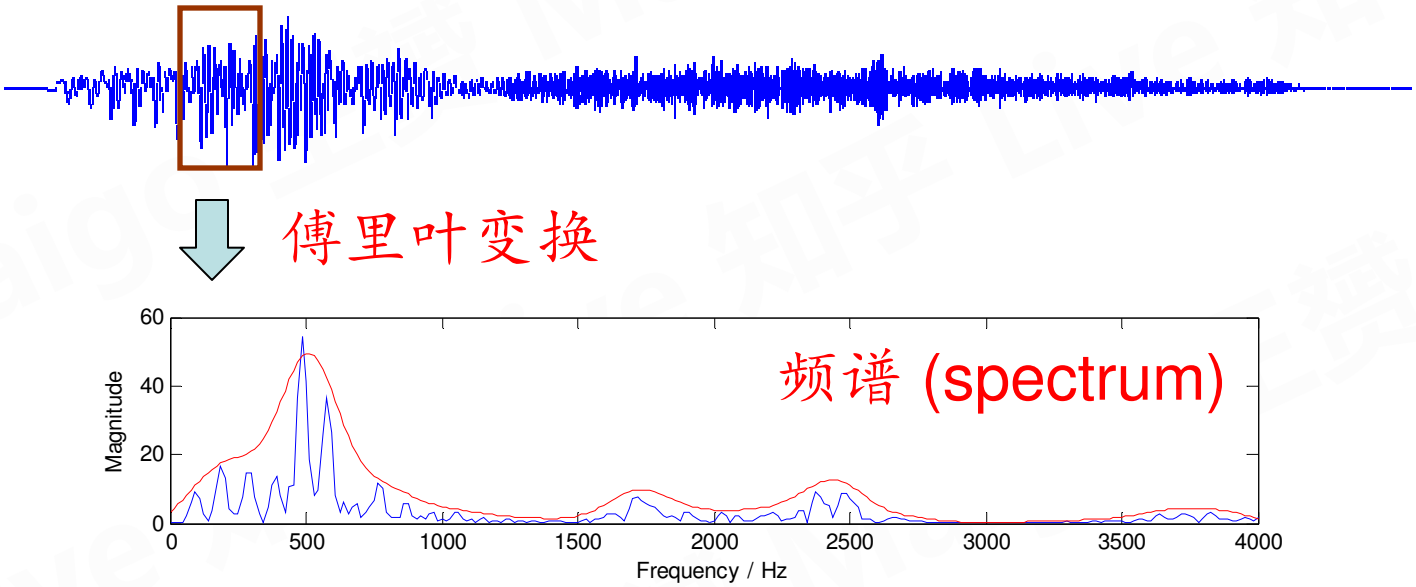
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

特征提取



- 频谱具有精细结构和包络
  - 精细结构反映音高，用处较小
  - 包络反映音色，是主要信息

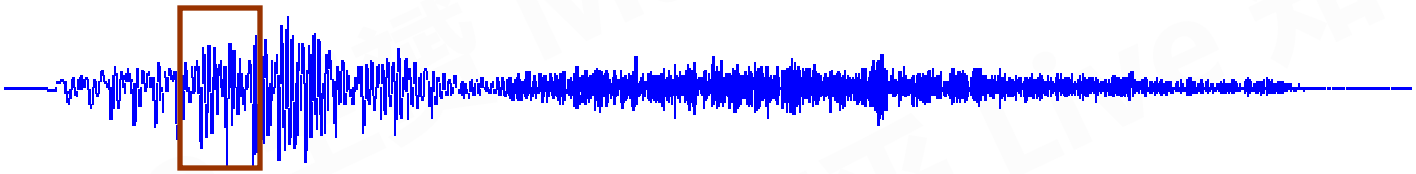
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

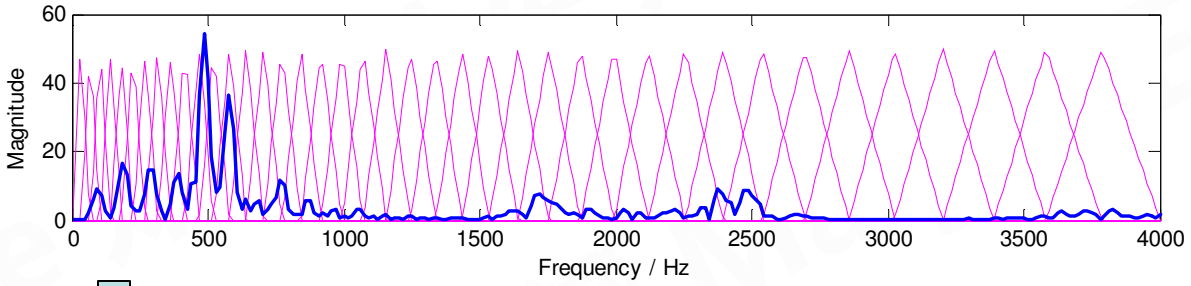
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

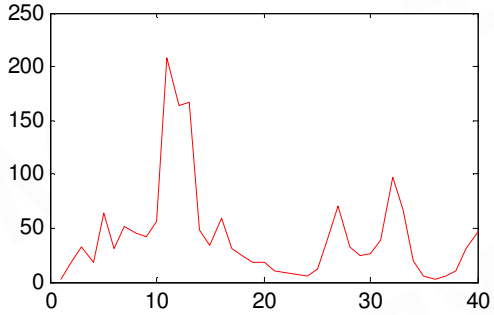
特征提取



傅里叶变换



三角滤波



Filterbank output

滤波器组输出

（近似频谱包络）

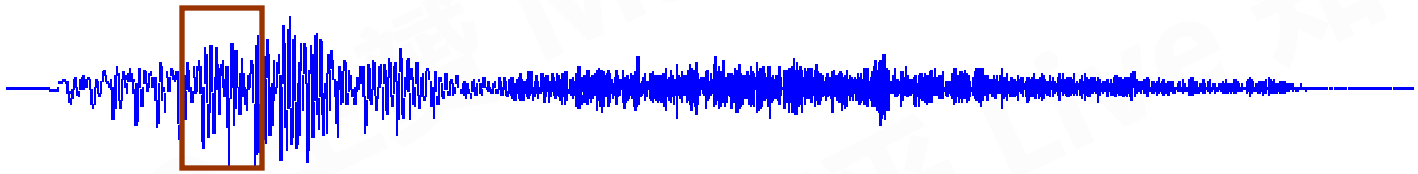
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

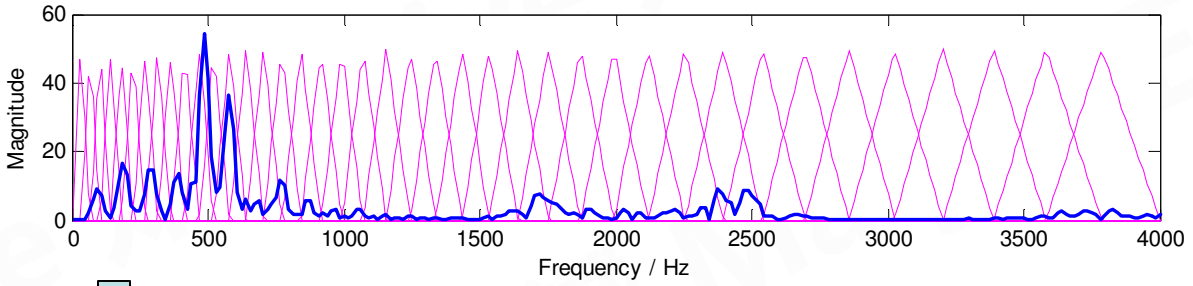
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

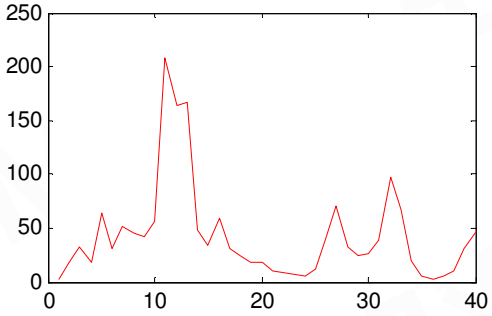
特征提取



傅里叶变换

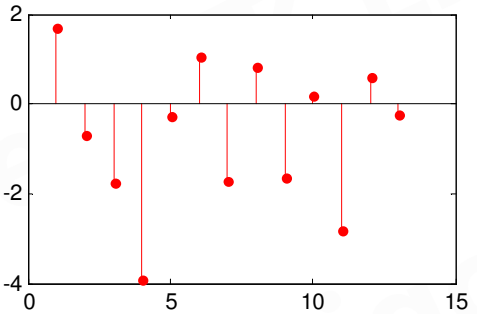


三角滤波



Filterbank output

log, DCT



MFCC (Mel frequency cepstral coefficients)

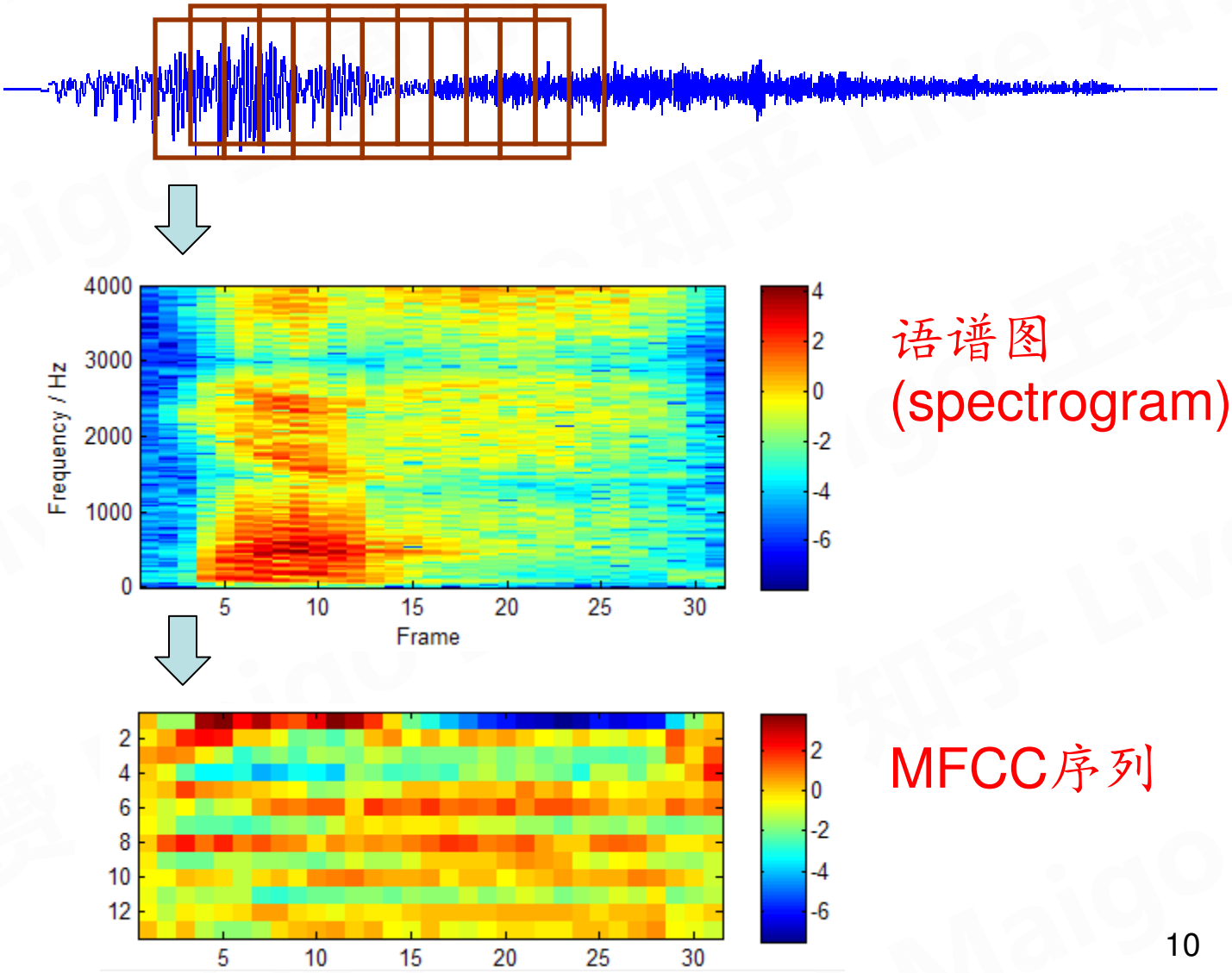
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

特征提取



语谱图
(spectrogram)

MFCC序列

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

特征提取

- MFCC序列是最常用的特征
  - 主要描述频谱包络
  - 优点：排除基频，符合听觉，维度低
  - 缺点：视野小，受噪声、回声、滤波影响严重
- 改进：
  - 加入一阶、二阶差分
  - 各种归一化
- 曾经有过各种其它改进特征
  - 在上一个十年是研究热点
  - 随着神经网络的兴起偃旗息鼓

引子

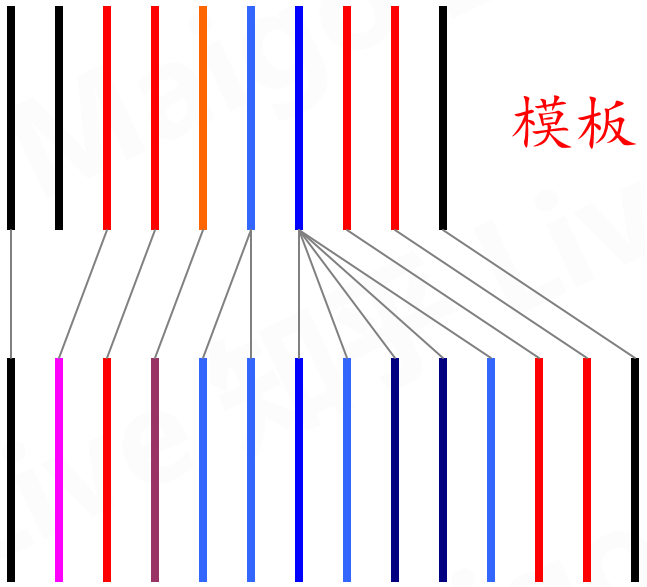
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

“动态弯”算法

- 怎样计算两个特征序列的距离？
- Dynamic Time Warping (DTW)
  - 让待识别语音中的每一帧与模板中最相似的一帧匹配
  - 但要保持顺序
  - 动态规划算法
- 总距离为各帧的欧氏距离之和



引子

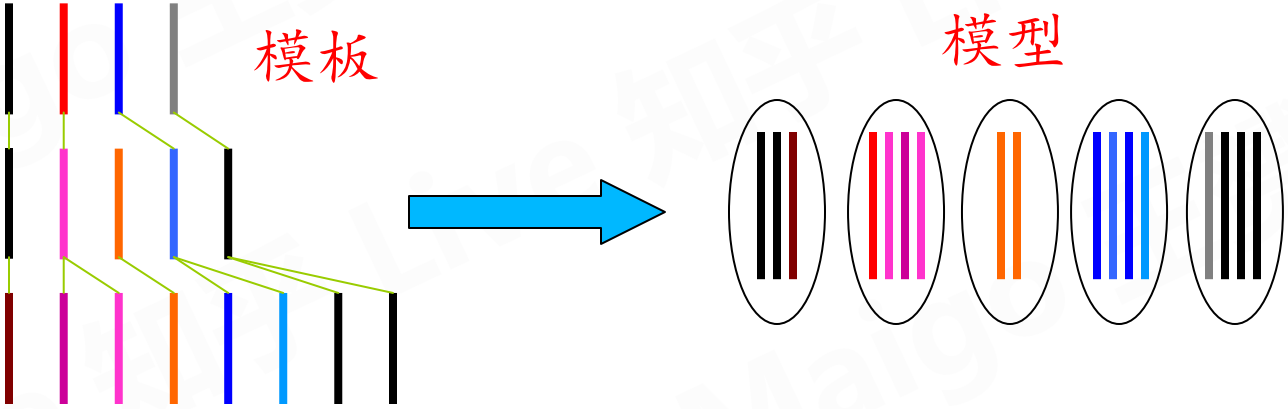
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

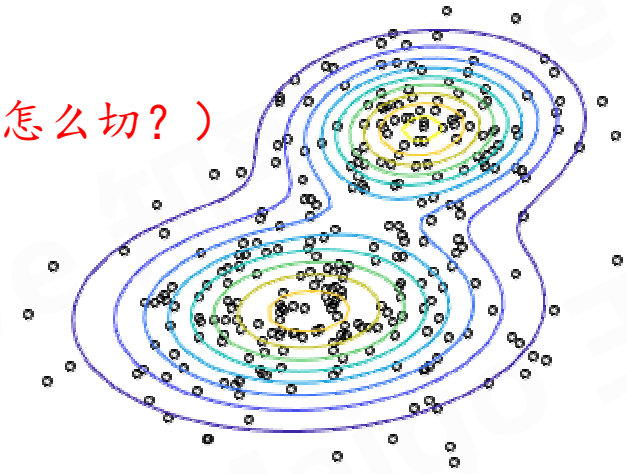
现状与未来

GMM（高斯混合模型）

- 如果每个词有多个模板，怎么办？



- 训练模型！
  - 把模板切分成多个段落（怎么切？）
  - 用高斯分布的叠加拟合每段中特征向量的分布
    - 对任一特征向量，可给出概率密度



引子

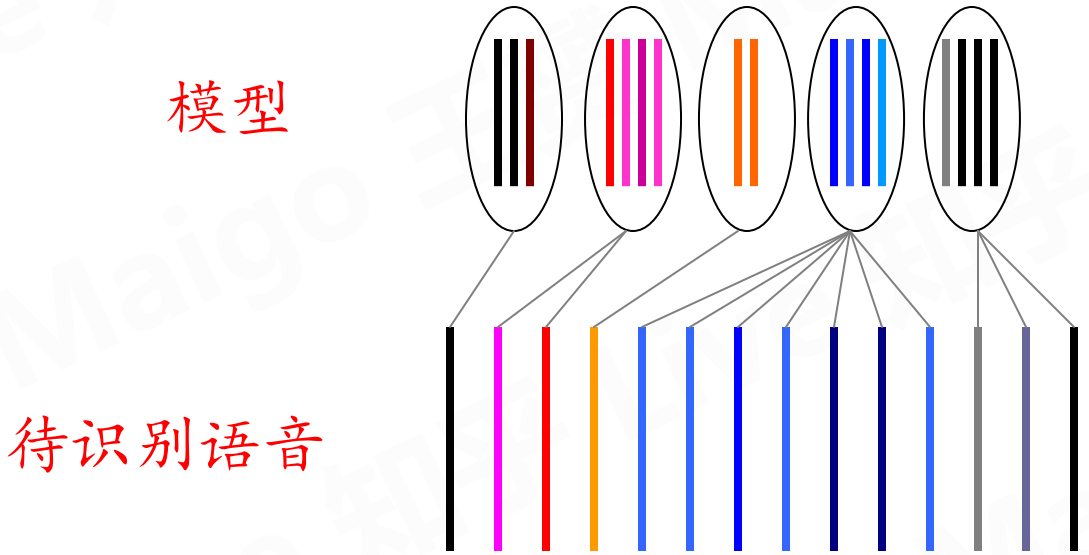
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

GMM（高斯混合模型）

- 如何用模型识别未知语音？
  - 用“动态弯”算法对齐待识别语音与模型
    - 用GMM概率密度代替特征向量间的欧氏距离
    - 相乘得到 *P*(待识别语音|模型)
      - 这里隐含了一个独立性假设
  - 取概率最大的模型为识别结果



引子

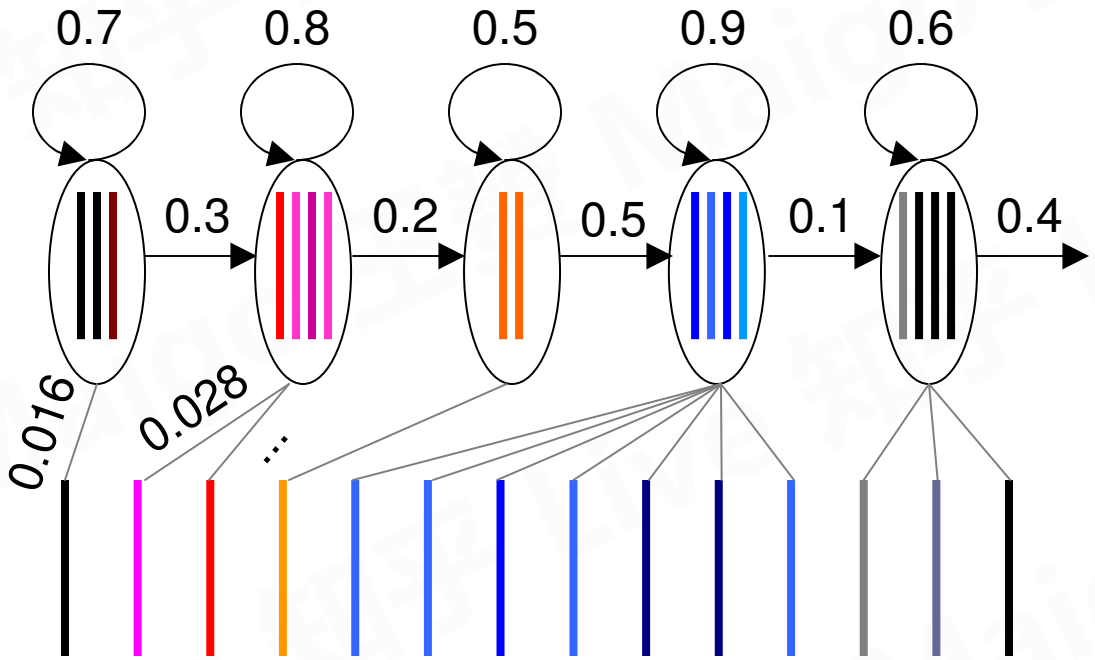
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

HMM（隐马尔可夫模型）

- 把模型完全概率化
  - 添加状态间的转移概率
  - *P*(语音，对齐方式|模型) = GMM观测概率 * HMM转移概率



引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

HMM（隐马尔可夫模型）

- 模型特点：
  - 隐：特征序列由隐状态产生，对齐方式未知
  - 马尔可夫：
    - 转移概率与观测概率都只由当前状态决定
    - 推论：状态持续时间服从指数分布
- 模型参数：
  - 转移概率
  - 观测概率（GMM）
  - 模型是单向的，不必讨论初始概率

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

HMM（隐马尔可夫模型）

- HMM三大问题：
  1. 求值：给定模型参数和语音，求*P*(语音|模型)
     - 把 *P*(语音,对齐|模型) 对所有对齐方式求和
     - 动态规划算法：Forward algorithm
  2. 解码：给定模型参数和语音，求最佳对齐方式
     - 动态规划算法：Viterbi decoding
     - 这是“动态弯”算法的升级版
     - 最佳对齐方式的概率，可以作为总概率的近似
  3. 训练：给定语音和模型结构，求模型参数
     - 如果知道了对齐方式，就好办了……

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

EM训练算法

- 鸡生蛋、蛋生鸡问题：
  - 如果知道了对齐方式，则容易求模型参数
  - 如果知道了模型参数，也容易求对齐方式
- 解决方法：
  - 先瞎猜一种对齐方式（如均匀分割）
  - 由此求出模型参数（M步）
  - 然后更新对齐方式（E步）
    - 可用Viterbi，实际中用Forward-backward
  - 循环直至收敛
- 最大似然估计：最大化*P*(训练语音|模型)

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

语音识别基本方程

$$W^* = \arg\max_W P(W \mid X) = \arg\max_W \frac{P(X \mid W)P(W)}{P(X)}$$

$$= \arg\max_W P(X \mid W)P(W)$$

- W^*：识别结果
- W：任一单词
- X：待识别语音信号
- $P(X|W)$：声学模型（GMM + HMM）
- $P(W)$：单词的先验概率

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

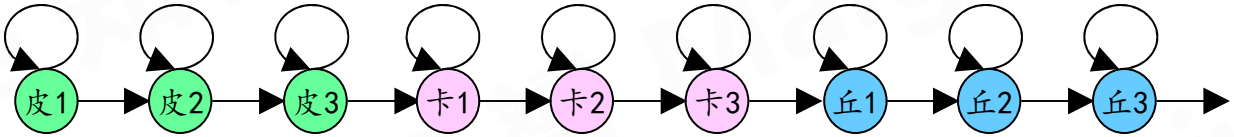
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

连续语音识别

$$W^* = \arg\max_{W} P(X \mid W) P(W)$$

- W 和 W^* 由单词变成了句子
- $P(X|W)$：句子的声学模型
  - 可由单词的声学模型串起来



- $P(W)$：句子的先验概率 —— 语言模型
  - 通俗理解：一句话是否“像话”

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

语言模型

- 链式法则：
  - P(皮卡皮卡丘) = P(皮) * P(卡|皮) * P(皮|皮卡) * P(卡|皮卡皮) * P(丘|皮卡皮卡)
  - 根据半句话猜下一个词
- 最常见形式：n-gram
  - 每个词只与前 n-1 个词有关
    - Bigram: P(皮卡皮卡丘) = P(皮) * P(卡|皮) * P(皮|卡) * P(卡|皮) * P(丘|卡)
    - Trigram: P(皮卡皮卡丘) = P(皮) * P(卡|皮) * P(皮|皮卡) * P(卡|卡皮) * P(丘|皮卡)
  - 容易训练和使用
- 其它形式：最大熵、神经网络……

引子

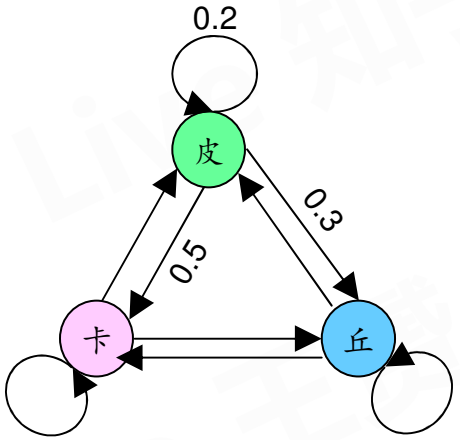
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

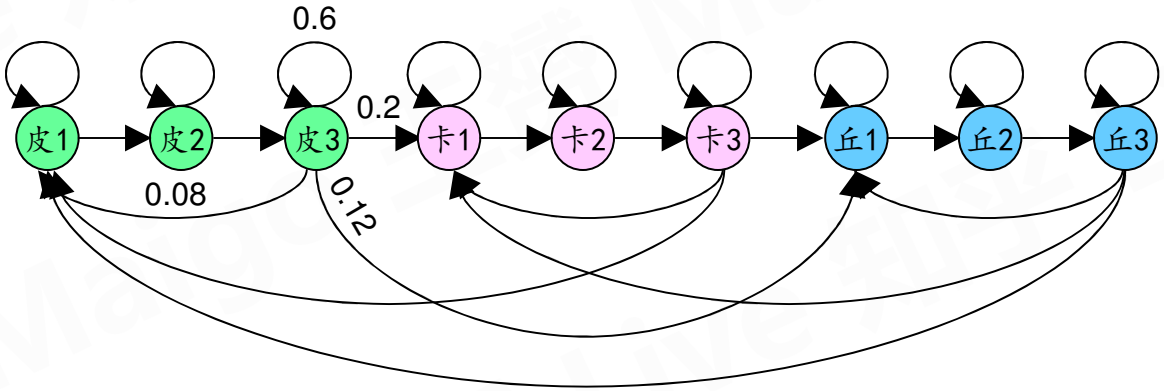
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

语言模型

- Bigram是马尔可夫模型
  - 下一个词只与当前词有关
  - 模型是遍历的，不是单向的
- 可与单词的声学模型复合
  - 得到一门语言的HMM





引子

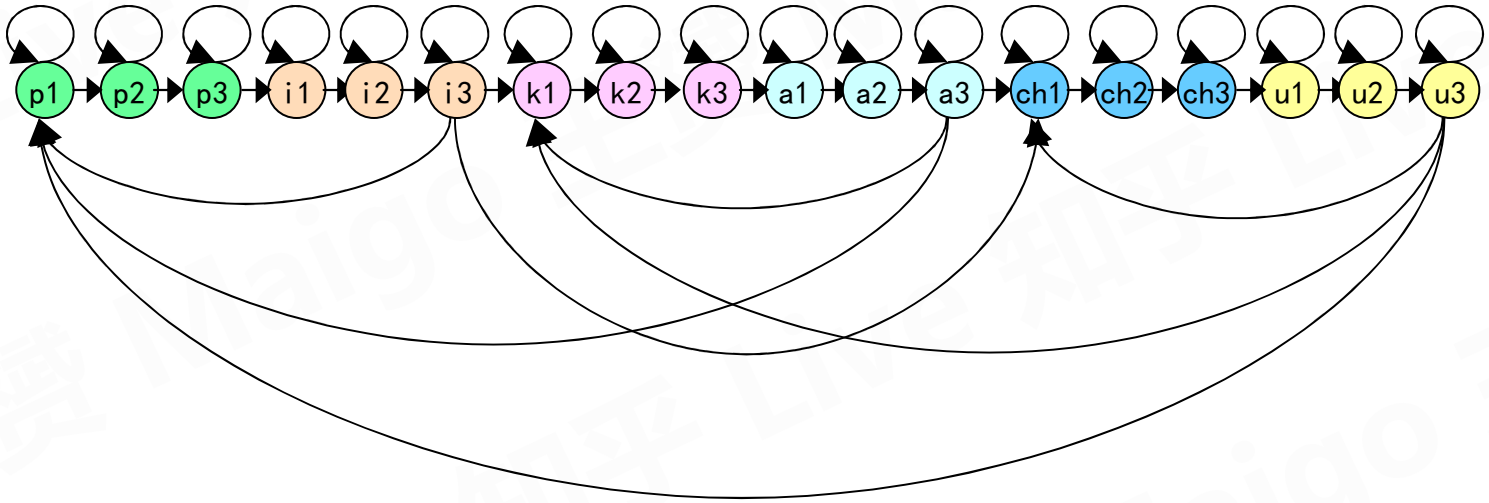
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

大词汇量语音识别

- 不能为每个单词训练单独的HMM
  - 改成为每个音素训练一个HMM
- HMM的复合：
  - 音素HMM按词典拼接成单词HMM
  - 单词HMM与语言模型复合成语言HMM



引子

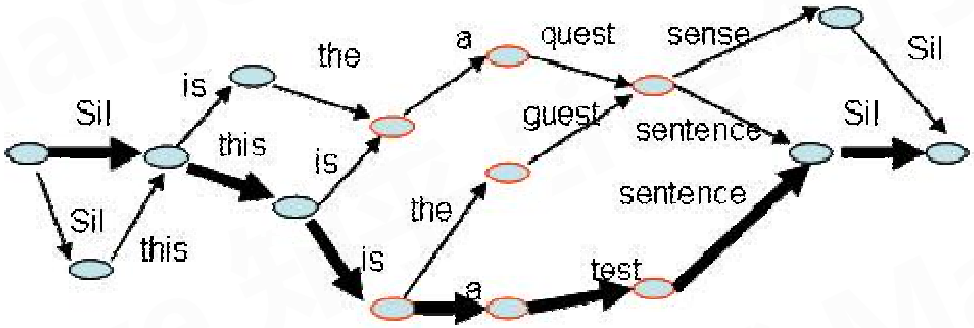
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

大词汇量语音识别

- 训练
  - 给定许多语音和对应的音素串，求模型参数
  - 每个音素串的HMM是单向的，仍用EM算法
- 解码
  - 给定一门语言的HMM和一条语音，求单词串
  - 用Viterbi算法求最佳路径（beam search剪枝）
  - 最佳路径经过的单词为识别结果
    - 也可以得到n-best list或lattice



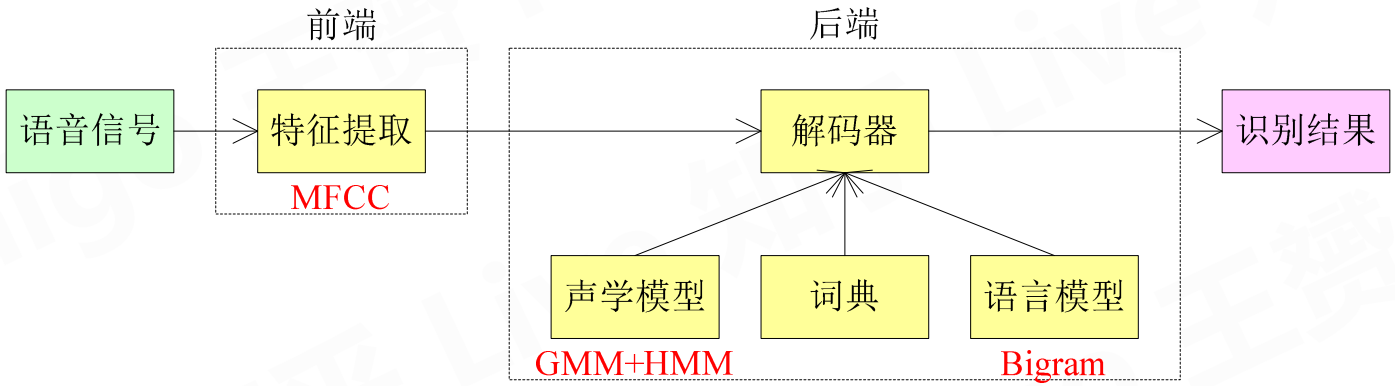
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

语音识别系统结构



引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

评价指标：词错误率

- 计算方法：
  - 将标准答案与识别结果对齐
  - 用插入、删除、替换错误的总数除以标准答案的长度
  - 对齐应使得错误数最少（动态规划）
- 例：
  - 标准答案：too young too simple sometimes naive
  - 识别结果：too young simple some times knife
  - 错误： 删除 替换 插入 替换
  - 词错误率 (word error rate)：4 / 6 = 66.7%

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

评价指标：词错误率

- 最优对齐不一定唯一
  - 标准答案：too young too simple sometimes naive
  - 识别结果：too young simple some times knife
  - 错误：删除 替换 替换 插入
  - 词错误率：4 / 6 = 66.7%
- WER可能高于100%
  - 标准答案：recognize speech
  - 识别结果：wreck a nice beach
  - 错误：替换 插入 插入 替换
  - 词错误率：4 / 2 = 200%

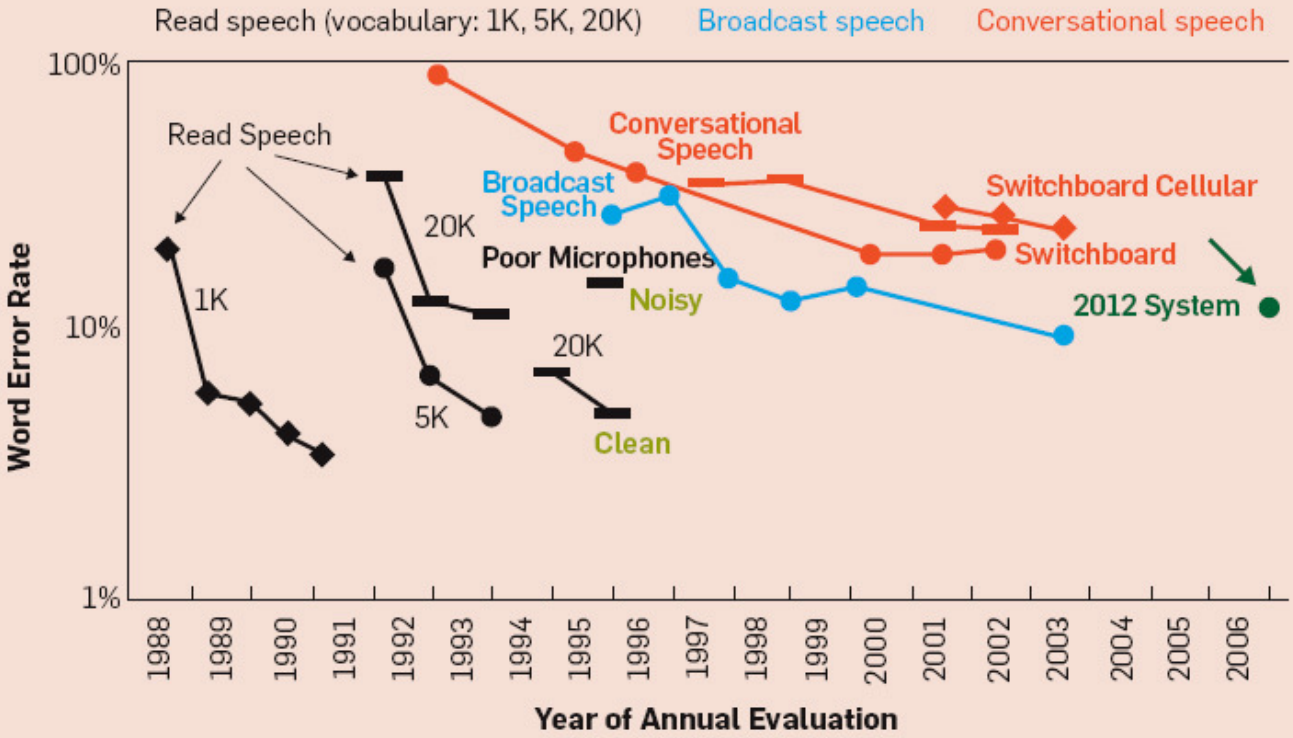
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

评价指标：词错误率



图片来源：http://deliveryimages.acm.org/10.1145/2510000/2500887/figs/f1.jpg

- 人类听写的词错误率范围：2~4%

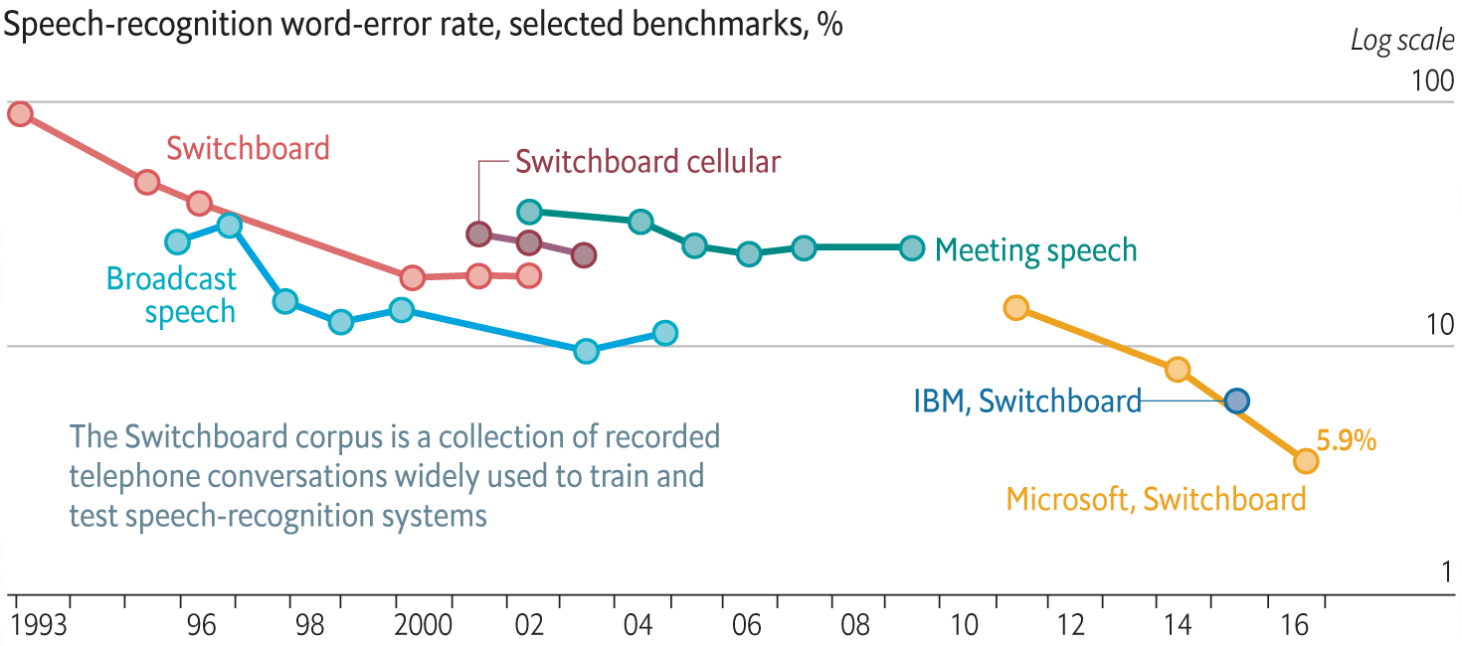
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

评价指标：词错误率



图片来源：http://cdn.static-economist.com/sites/default/files/external/tq2016/charts/03/WIDE_TQC003.png

- IBM最新结果：5.5% (https://arxiv.org/pdf/1703.02136.pdf)

引子

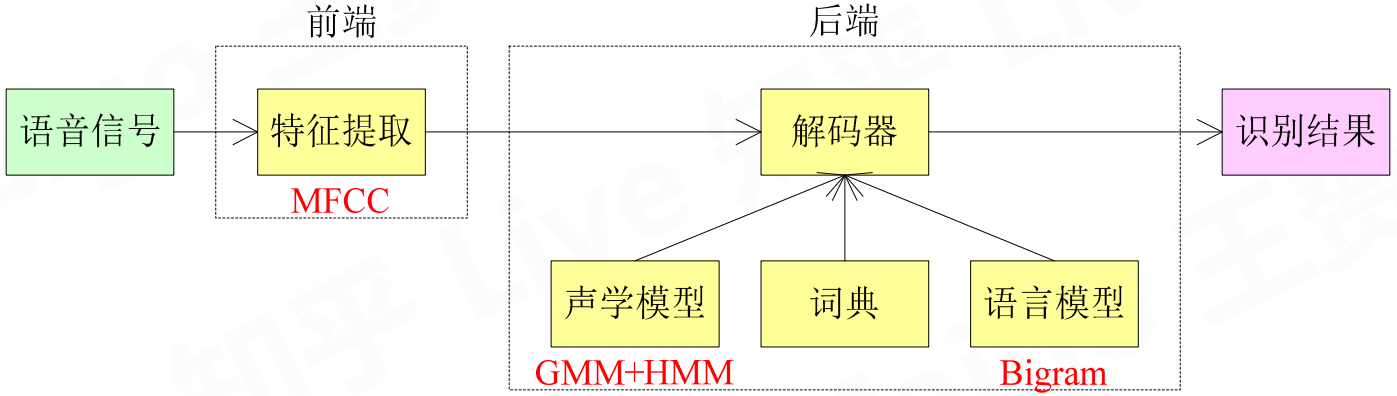
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

潘多拉魔盒

- 1990 ~ 2010 年，下图的框架没有变化



- 人们给这个框架打了好多补丁……

引子

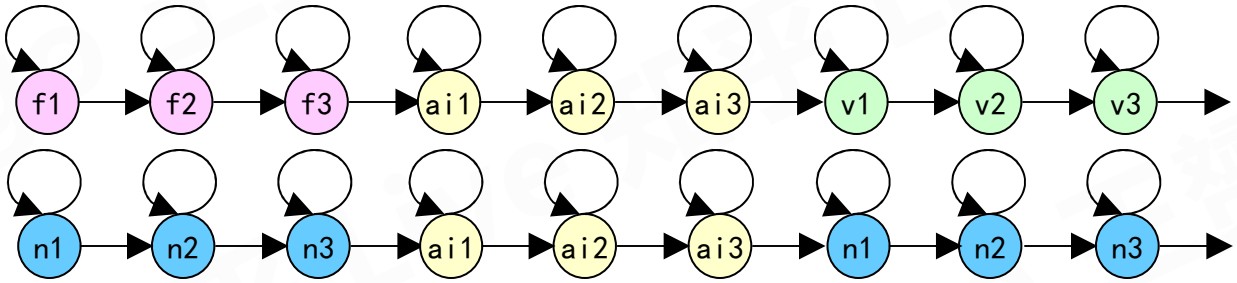
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

上下文有关模型

- Five和nine的单词HMM



- 上下文有关模型：
  - 两个ai不一样，分别是ai(f,v)和ai(n,n)
  - 两个n不一样，取决于前后的单词
- 上下文聚类：
  - 按上述方法，状态数会爆炸！(N → N^3)
  - 聚类：ai(f,v)和ai(n,n)可能仍不同，但ai(m,n)与ai(n,n)可能就一样了

引子

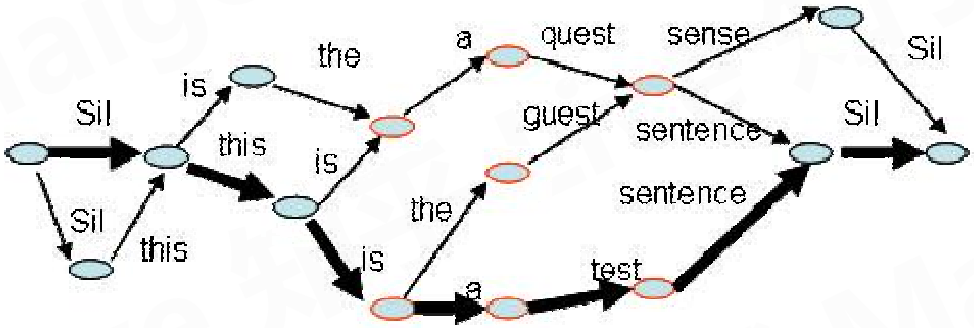
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

区分式训练

- EM算法是最大似然估计
  - 最大化$P(X|W)$，W是训练文本
  - 但$P(X|W')$可能更大了，W'是W的竞争者
  - 导致$P(W|X)$不一定最大化
- 区分式训练 (discriminative training)
  - 让$P(X|W)$大，同时让$P(X|W')$小
  - 竞争者从哪儿来？
    - 来自最大似然系统输出的n-best list或lattice



引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

说话人适应

- 说话人相关训练 (speaker dependent training)
  - 在训练时，专门收集特定说话人的数据
- 说话人适应 (speaker adaptation)
  - 在测试时，把模型参数或待识别语音的特征向量整体平移，使二者互相适应
  - 模型参数可以不断更新，逐步适应说话人
- 说话人适应训练 (speaker adaptive training)
  - 在训练时，提取说话人的特征（如i-vector），与声学特征一同作为模型的输入

引子

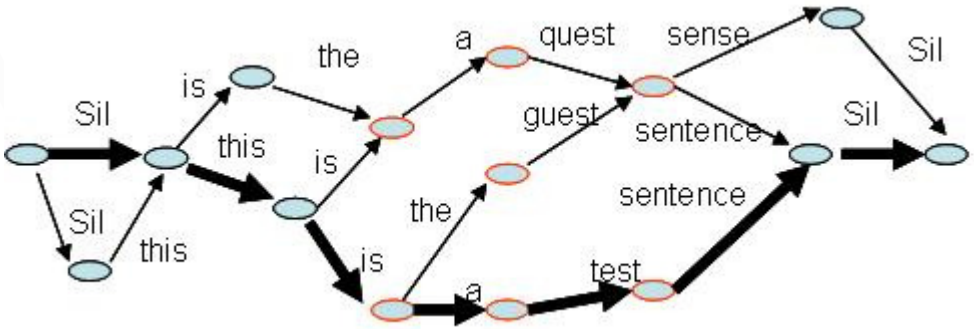
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

二次打分

- 解码器只能利用n-gram语言模型
  - HMM的状态数随n指数增长
  - 还要考虑编程复杂度
  - 实际一般只用bigram
- 如何利用更好的语言模型？
  - 用bigram识别得到n-best list或lattice
  - 用更好的语言模型对这些句子重新打分，选出最优解



引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

潘多拉魔盒

- 这些补丁确实降低了WER
- 但是：
  - 系统变得复杂到难以驾驭
  - 各模块是单独训练的，用力并不统一
- 天下大势，分久必合


语音识别之今生
——神经网络

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

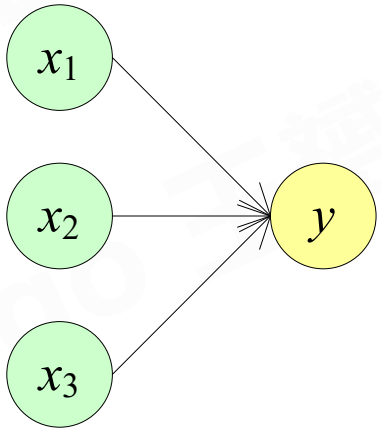
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

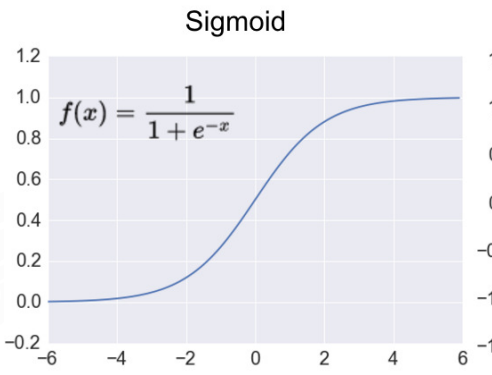
什么是神经网络

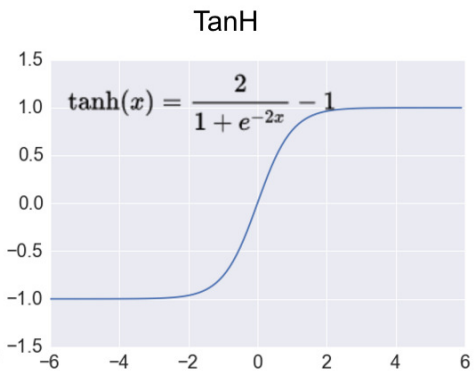
- 神经网络是一个复杂的函数
- 神经元：

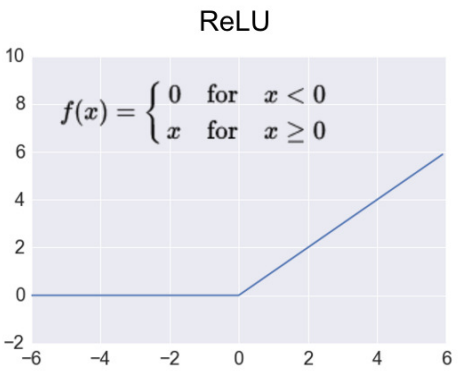
$$y = \sigma(w_1 x_1 + w_2 x_2 + w_3 x_3 + b)$$



- σ是非线性函数，称为“激活函数”(activation function)







引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

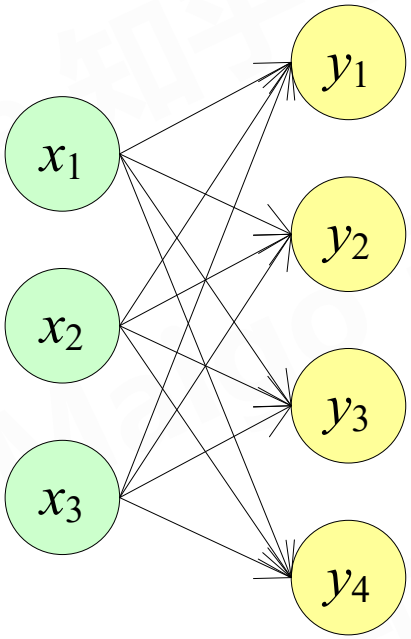
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

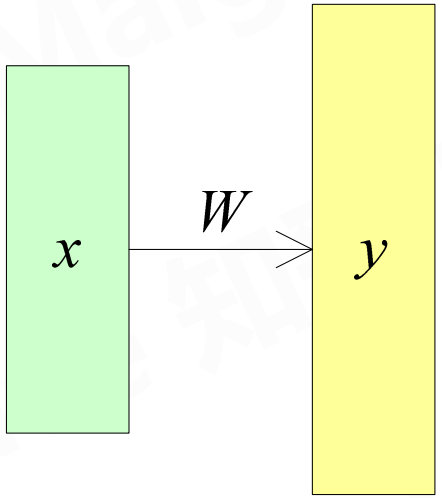
现状与未来

什么是神经网络

- 一层神经网络：
  - 写成向量形式：

$$y = \sigma(Wx + b)$$





引子

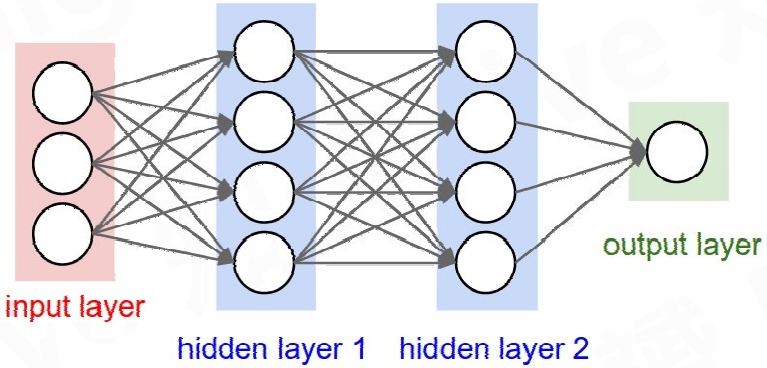
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

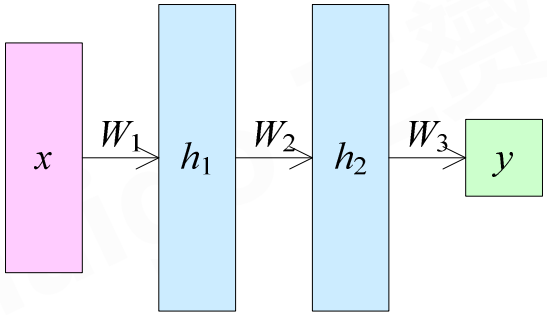
今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

什么是神经网络

- 前馈神经网络 (feed-forward neural network)
  - 许多层神经网络摞起来





$$y = \sigma_3(W_3\sigma_2(W_2\sigma_1(W_1x + b_1) + b_2) + b_3)$$

- 多层非线性让网络具有强大的拟合能力
  - 但需要大量训练数据

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

什么是神经网络

- 神经网络能干什么？
  - 回归：输出任意实数，尽量接近标准答案
  - 两类判别：输出概率(0~1)，尽量接近标答
  - 多类判别：输出多项分布，让标答项最大
- 神经网络发展的关键
  - 大数据
  - GPU

引子

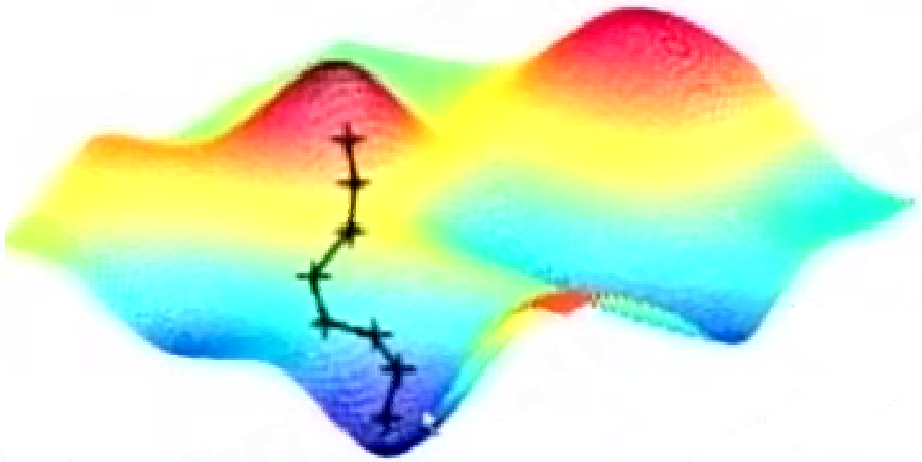
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

什么是神经网络

- 如何训练神经网络？
  - 神经网络是一个带参数(W, b)的函数
  - 设计损失函数 (loss function)
  - 梯度下降法 (gradient descent)
  - 反向传播 (back-propagation)



引子

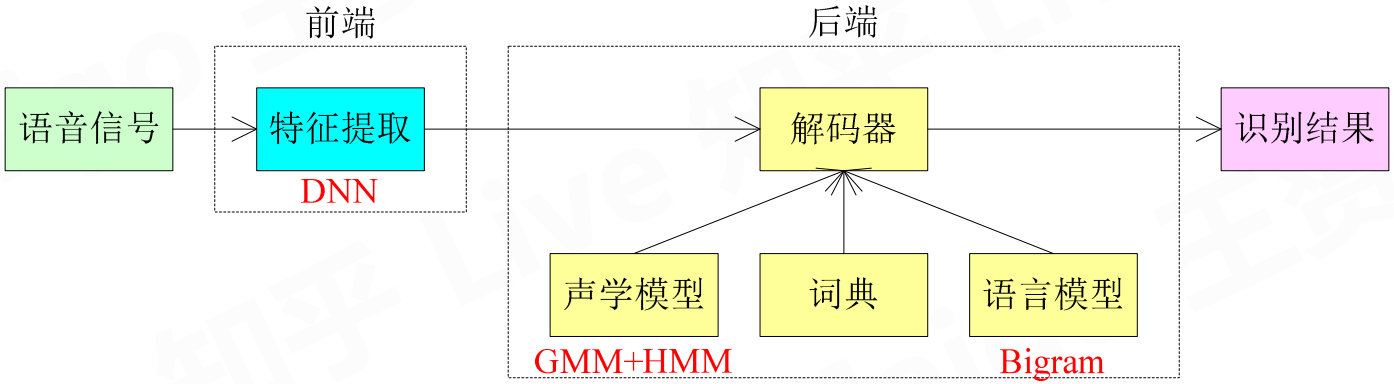
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

Tandem结构

- 神经网络开始蚕食GMM+HMM框架



tandem

['tændəm]

n. 双人自行车

引子

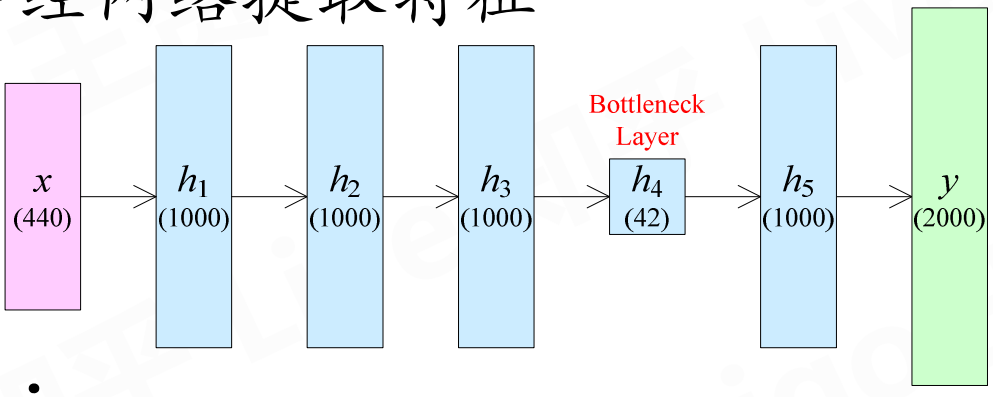
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

Tandem结构

- 用神经网络提取特征



- 输入：
  - 连续若干帧的滤波器组输出
  - 甚至直接输入波形
- 输出：
  - 上下文有关音素的分布（多类判别问题）
  - 标准答案由GMM+HMM系统提供
- 特征来自瓶颈层：醉翁之意不在酒

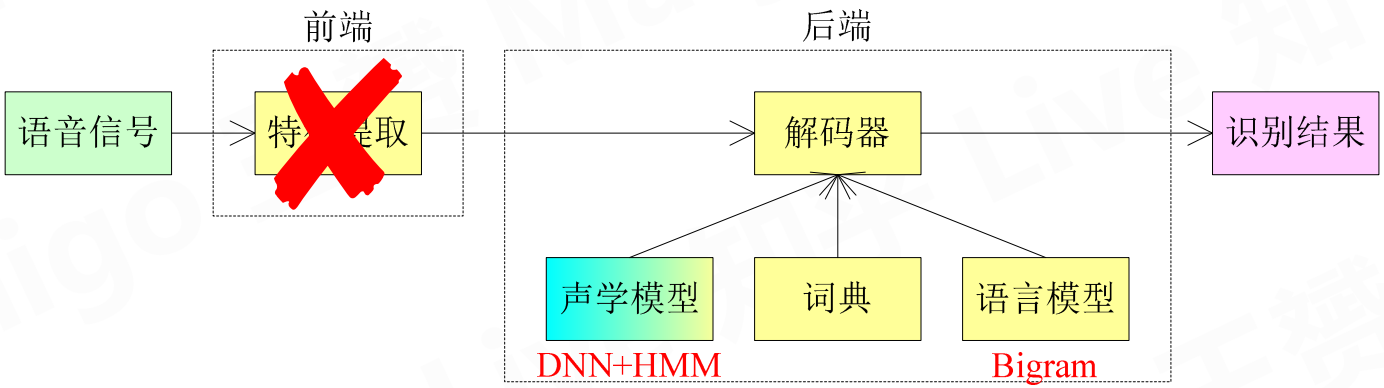
引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

Hybrid结构



- 不再进行特征提取
  - 输入为滤波器组输出或波形
- DNN+HMM声学模型
  - 原先，GMM提供*P*(特征|状态)
  - 现在，DNN提供*P*(状态|输入)
    - 需用贝叶斯公式转换一下
- 成品系统中没有GMM
  - 但训练DNN时需要GMM+HMM系统提供标答

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

循环神经网络

- HMM对上下文的建模能力有限
  - 源于马尔可夫性
- 补救：
  - MFCC特征中的差分
  - DNN声学模型输入连续多帧滤波器组输出
  - 上下文有关的音素模型
- 神经网络能不能取代HMM？

引子

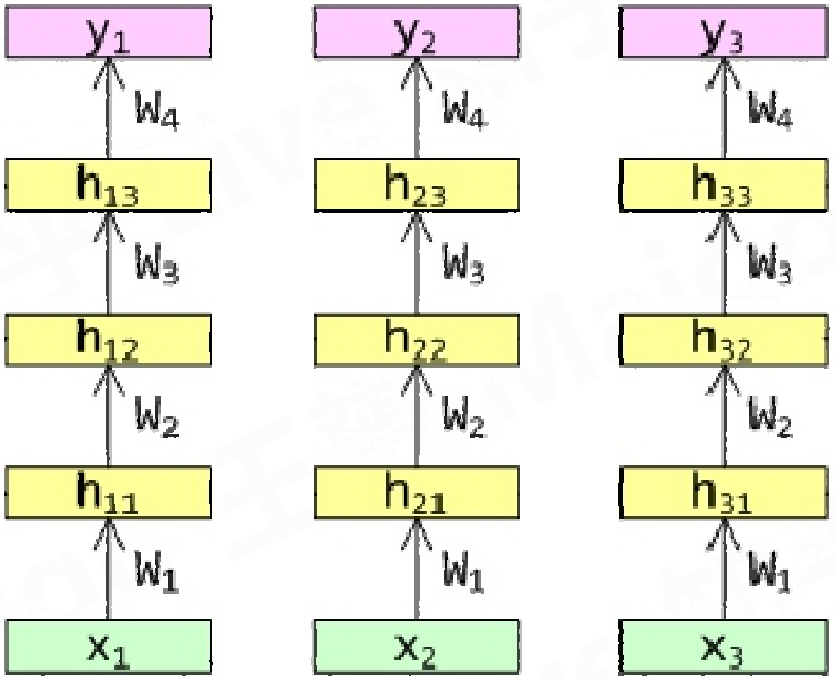
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

循环神经网络

- 前馈神经网络处理时间序列



引子

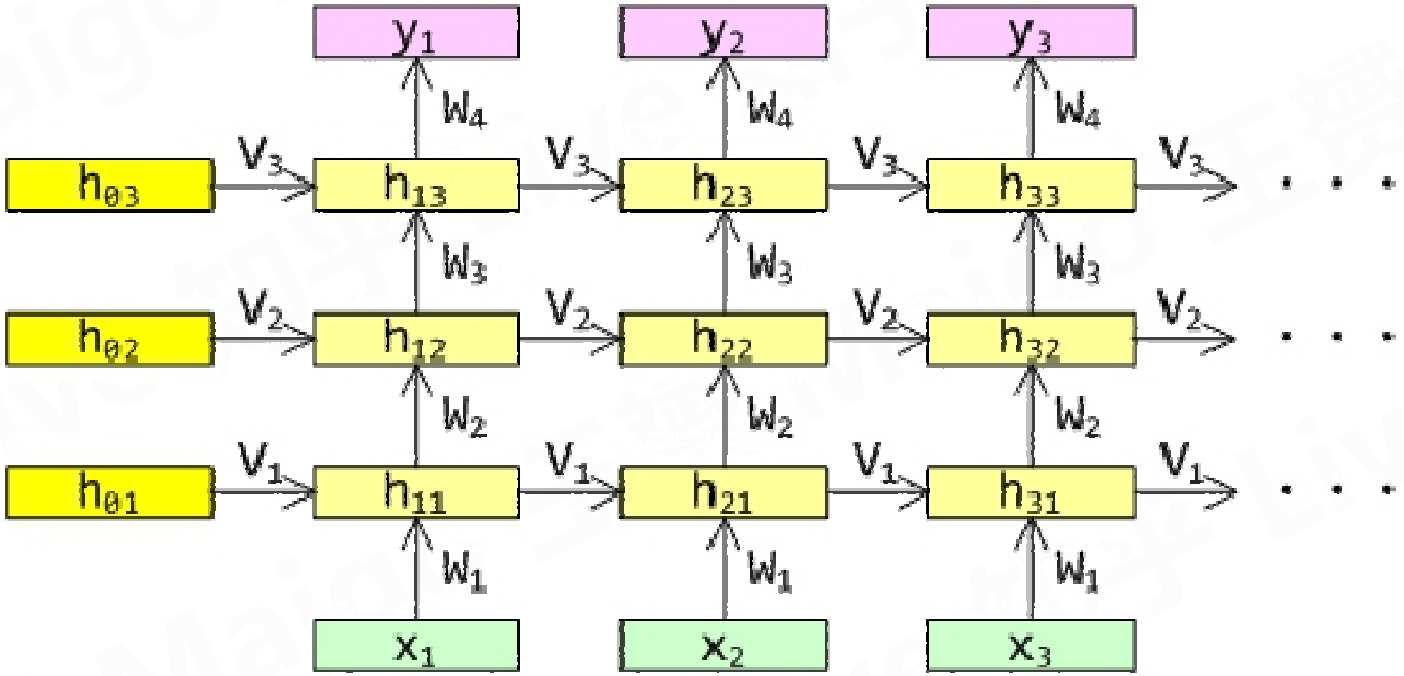
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

循环神经网络

- 循环神经网络 (recurrent neural network)



引子

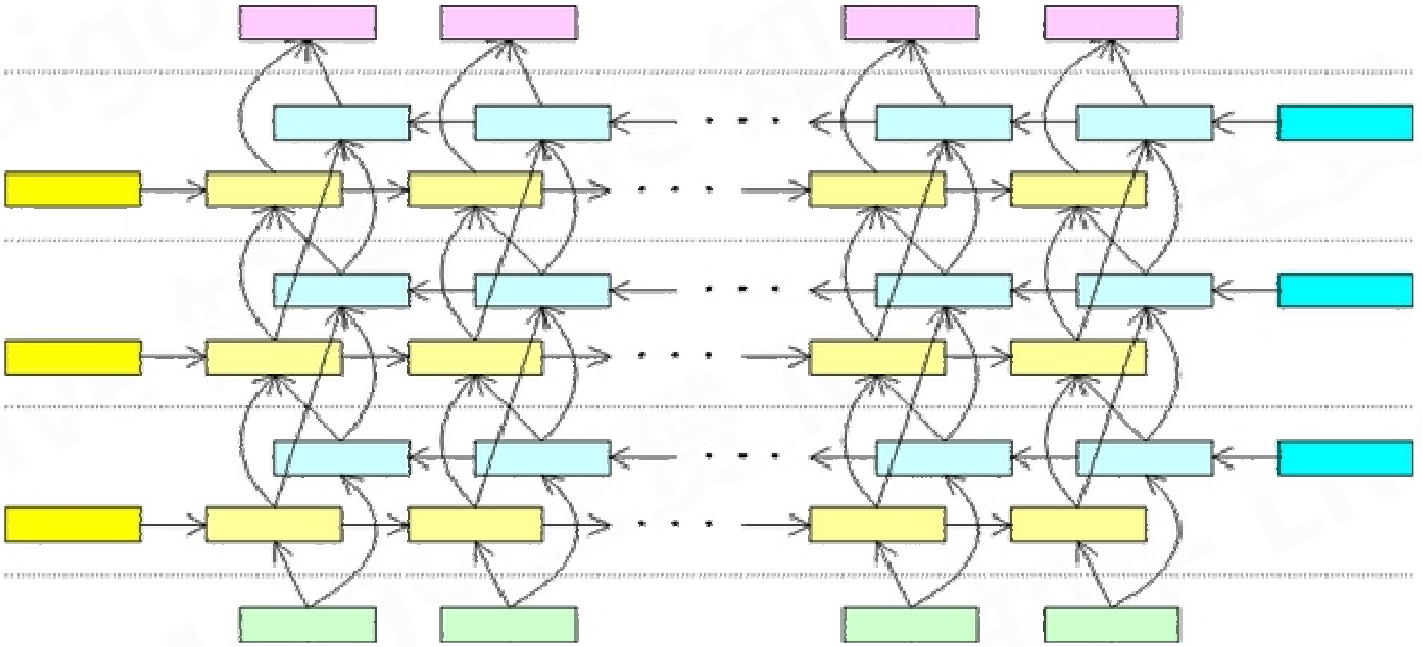
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

循环神经网络

- 双向循环神经网络 (bidirectional RNN)



引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

循环神经网络

- 梯度消失或爆炸问题
  - 导致RNN记忆力有限
  - 解决：LSTM / GRU（复杂的非线性函数）
- RNN在语音识别中的用途：
  - 代替DNN用于特征提取或声学模型
- 为什么一直留着HMM？
  - 神经网络只进行逐帧判别
  - 训练时，需要由HMM系统提供各音素起止时间
  - 解码时，需要考虑状态转移概率

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

CTC

- 不再进行逐帧判别
  - 大部分帧输出为空，小部分帧输出音素
  - 只要求输出音素串与标答相同，不要求位置
    - 实际上，输出音素的位置往往与真实位置接近
- 例：
  - 波形： 
  - 普通声学模型输出：

    p p i i k k a a a a a ch ch u u u u
  - CTC输出：

    - p - i - k - - a - - - ch - u - -

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

CTC

- 别了，HMM！
  - 假设各帧输出相互独立
  - 认为上下文已由RNN处理
- 训练：
  - 目标函数：所有能缩成标答音素串的输出的总概率
  - 动态规划算法
  - 比HMM简单，因为没有转移概率
- 解码：
  - CTC输出音素串，仍需词典和语言模型

引子

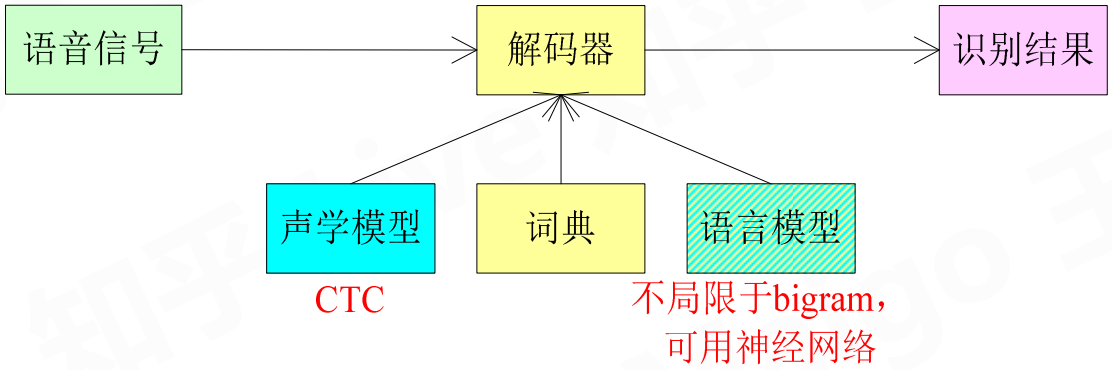
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

Grapheme系统

- CTC系统结构：



- Phoneme vs grapheme
  - 训练时，已知字符串，需要转换成音素串
  - 解码时，CTC输出音素串，需要结合词典和语言模型转换成字符串
  - 费那劲干嘛！

引子

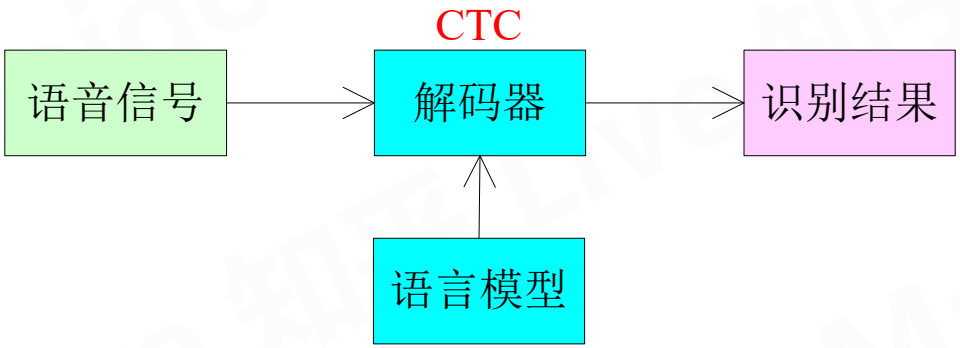
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

Grapheme系统

- 优点：
  - 简洁！
  - 不需要语言知识
  - 不怕生词
  - 可进行端到端训练
- 缺点：
  - 需要大量训练数据
  - 难以利用纯文本数据训练语言模型
- 往往外接语言模型



引子

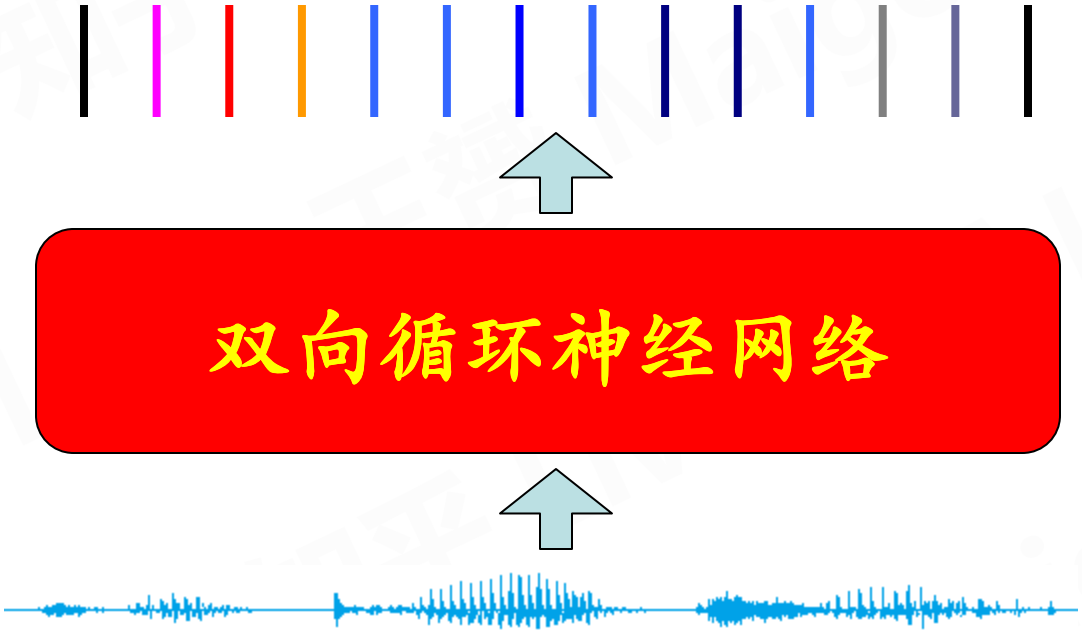
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

注意力机制

- CTC：每步输入一帧，可能只思考不输出
- 注意力：
  - 每步主动选择输入，并输出一个音素或字符
  - 由编码器和解码器两部分组成
- 编码器（可理解成特征提取）：



引子

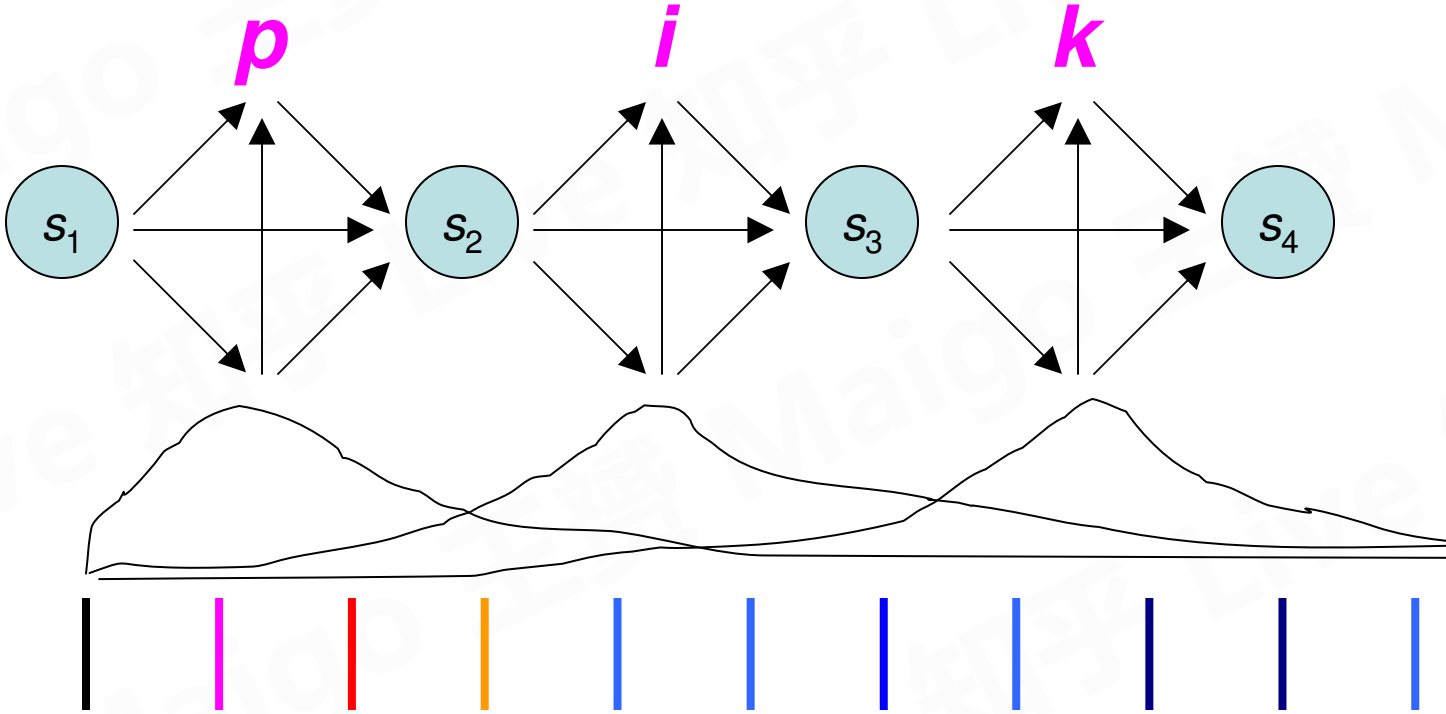
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

注意力机制

- 解码器：



- 能解决语序不单调的问题！
  - e.g. $2000

引子

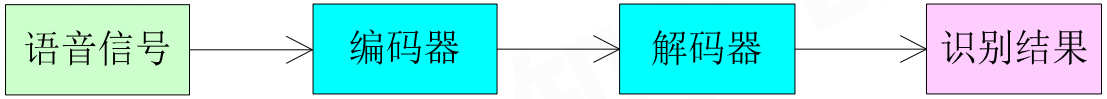
前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

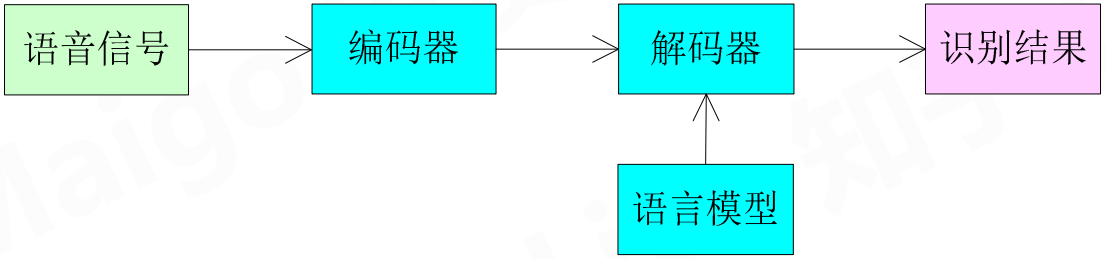
注意力机制

- 注意力系统结构：



  - 看似变复杂了，其实无伤大雅

- 也可以外接语言模型




语音识别之未来

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

语音识别的现状

- 模型简洁，容易训练和使用
- 在理想情况下的性能可与人媲美
- 在恶劣条件下不堪一击
  - 噪声
  - 信道特性（如手机）
  - 远场
  - 口音

引子

前世：GMM + HMM
孤立词识别
特征提取
动态弯算法
GMM
HMM
EM训练算法
语音识别基本方程
连续语音识别
语言模型
大词汇量
语音识别系统结构
评价指标：WER
潘多拉魔盒
上下文有关模型
区分式训练
说话人适应
二次打分

今生：神经网络
前馈神经网络
Tandem结构
Hybrid结构
循环神经网络
CTC
Grapheme系统
注意力机制

现状与未来

语音识别的未来

- 有针对性地应对恶劣条件
  - 除噪、语音增强
  - 麦克风阵列
- 收集大数据，让神经网络“长见识”
  - 有助于理解口音
- 相关领域的配合
  - 对话系统中对于打断的处理
  - 真实环境中信息的利用
    - e.g. 660 First Street vs 6 61st Street

Thank you!

欢迎关注我的知乎及知乎专栏：
https://www.zhihu.com/people/maigo
https://zhuanlan.zhihu.com/maigo