ROS与语音识别

华东师范大学
陈宇

什么是语音识别（Automatic Speech Recognition)?

hello world

语音识别的基本原理

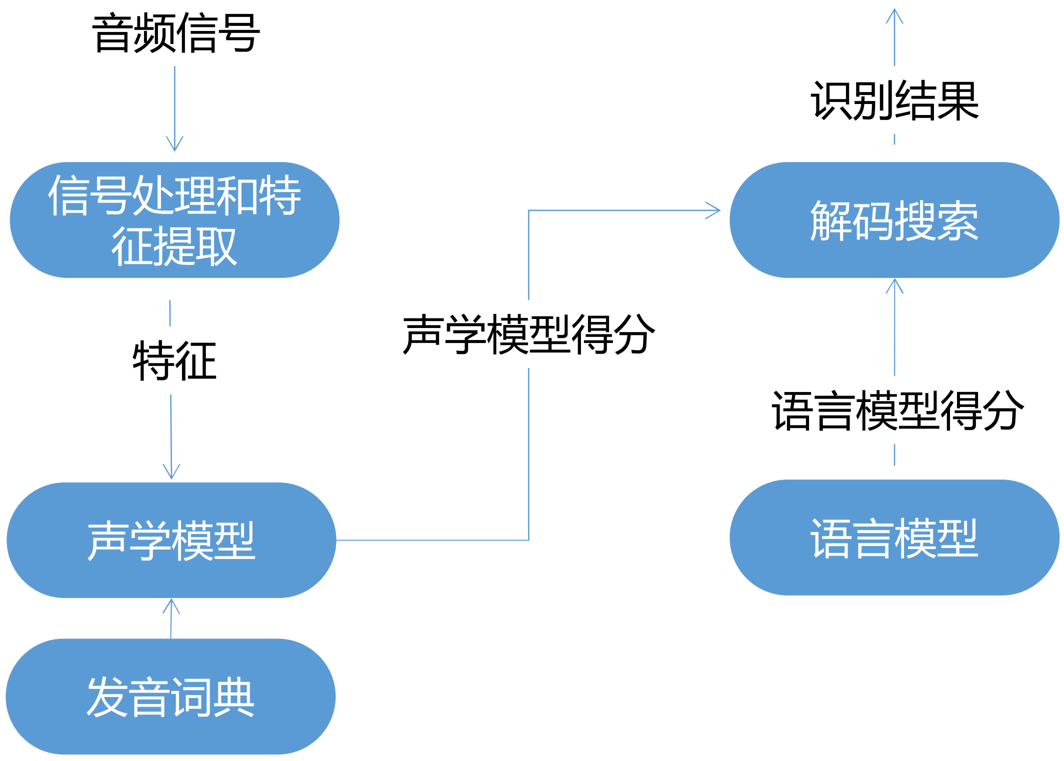
$$W^* = \arg\max_w P(W \mid Y) \tag{1}$$

$$= \arg\max_w \frac{P(Y \mid W)P(W)}{P(Y)} \tag{2}$$

$$\approx \arg\max_w \underbrace{P(Y \mid W)}_{\text{Acoustic Model(AM)}}\underbrace{P(W)}_{\text{Language Model(LM)}} \tag{3}$$

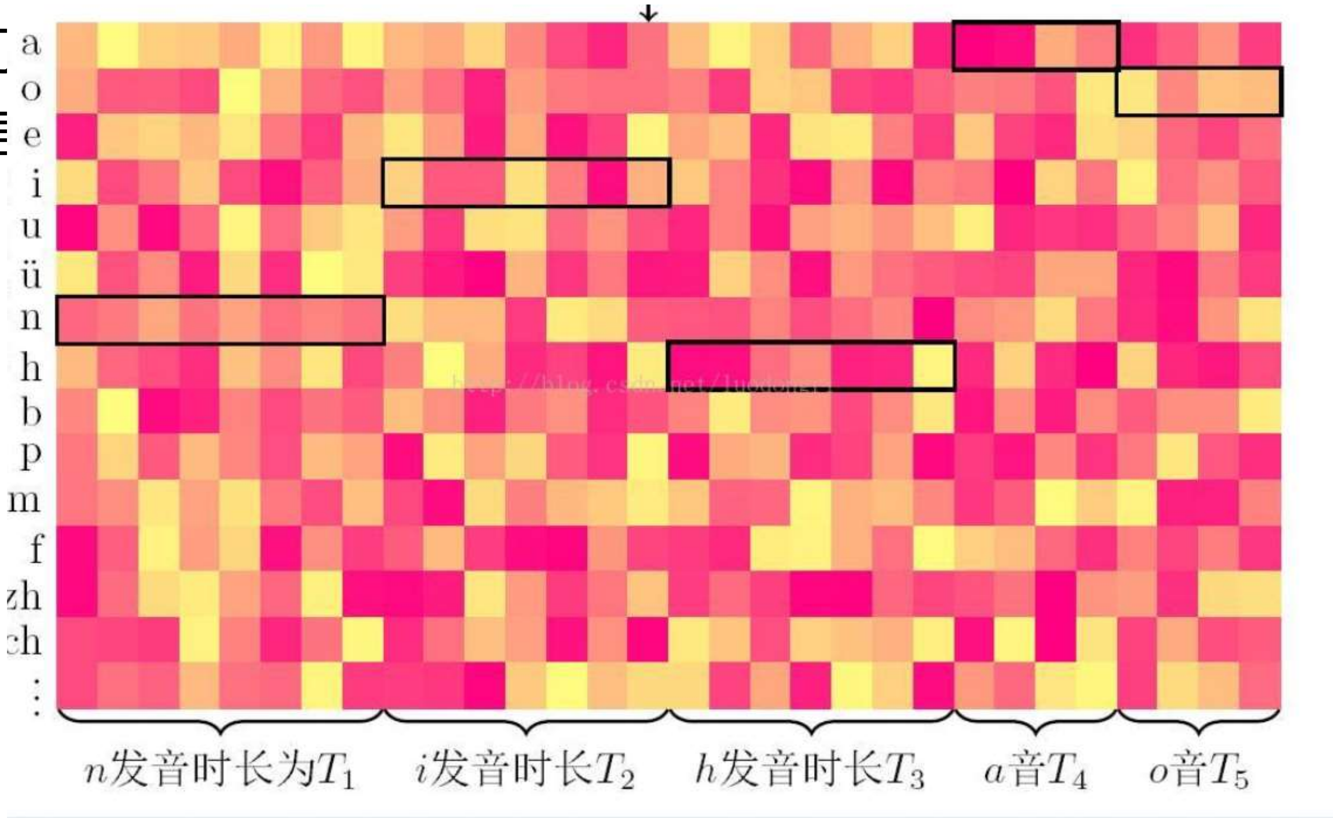
$$P(W) = P(w1, w2, ..., wk) = P(w1)P(w2 \mid w1)...P(wk \mid w1, w2, ..., wk-1)$$

传统语音识别系统



关于声学模型，主要有两个问题：

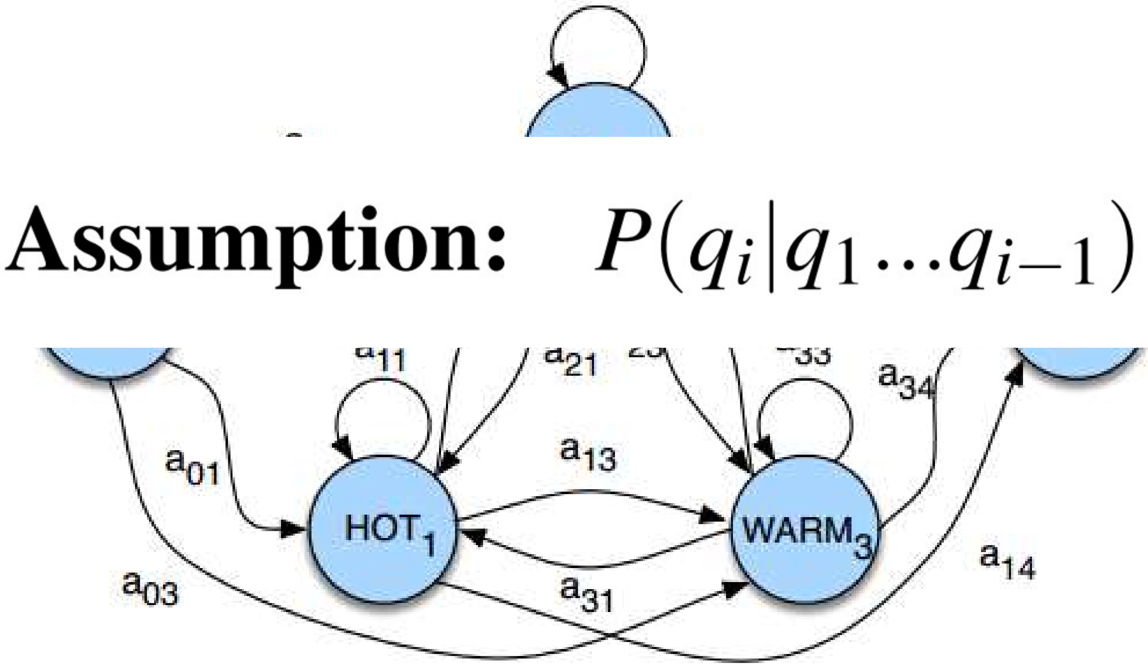
- 1、特征向
- 2、音频信



马尔可夫假设与马尔可夫链

$$P(W) = P(w1, w2, \ldots, wk) = P(w1)P(w2 \mid w1) \ldots P(wk \mid w1, w2, \ldots, wk-1)$$

Markov Assumption: $P(q_i | q_1 \ldots q_{i-1}) = P(q_i | q_{i-1})$



隐马尔科夫模型（Hidden Markov Model，HMM）

1. 隐含状态 S
2. 可观测状态 O
3. 初始状态概率矩阵 π
4. 隐含状态转移概率A
5. 隐含状态到观测状态的发射概率B

Baum-Welch

隐马尔科夫模型（Hidden Markov Model，HMM）

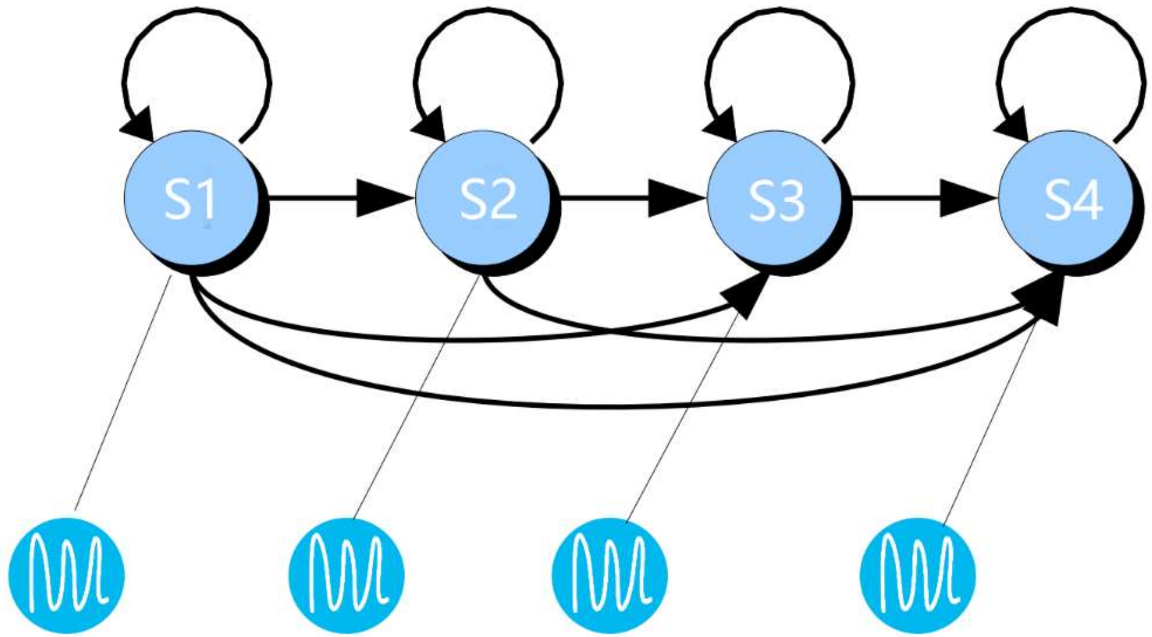
$$P(o_1, o_2, \ldots, o_t, s_1, s_2, \ldots, s_t)$$

$$= P(o_1, o_2, \ldots, o_t \mid s_1, s_2, \ldots, s_t) P(s_1, s_2, \ldots, s_t)$$

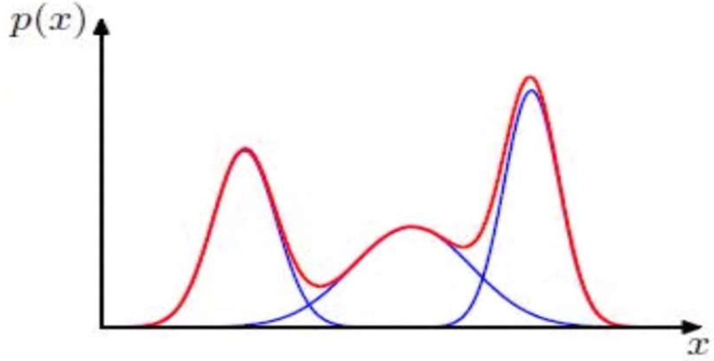
$$= \prod_t P(o_t \mid s_t) P(s_t \mid s_{t-1})$$

$$\arg\max_w \underline{P(Y \mid W)}\,\underline{P(W)}$$

隐马尔科夫模型（Hidden Markov Model，HMM）



混合高斯模型GMM

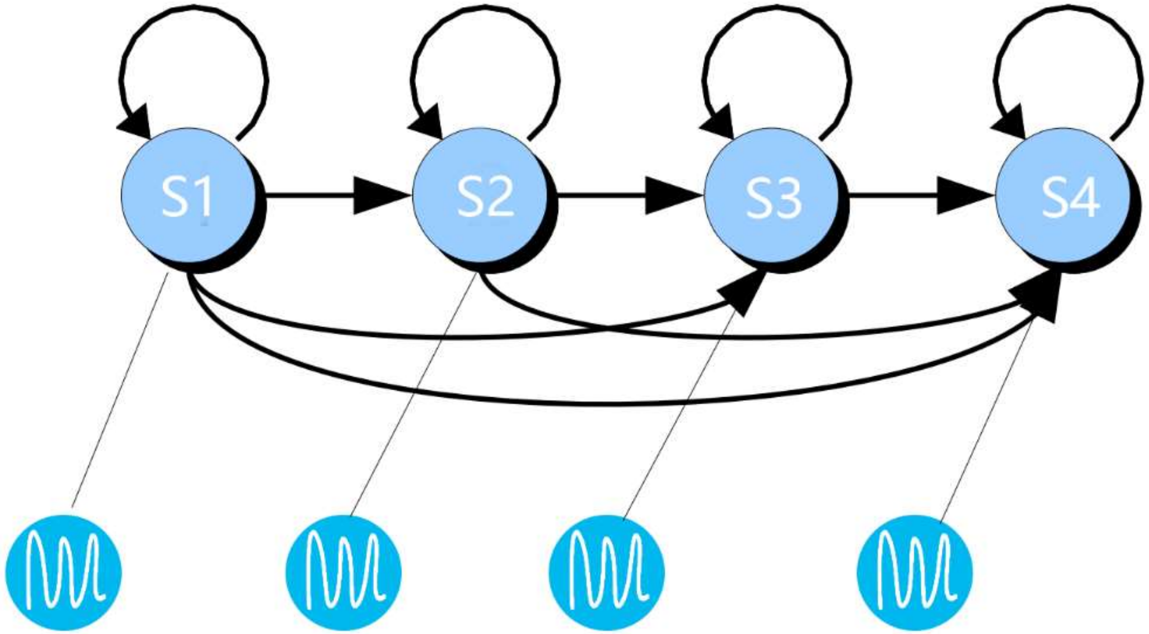


$$b_i(o_t) = \sum_{m=1}^{M} \frac{c_{i,m}}{2\pi^{D/2} \left|\Sigma_{i,m}\right|^{1/2}} \exp[-\frac{1}{2}(o_t - \mu_{i,m})^T \Sigma_{i,m}^{-1} (o_t - \mu_{i,m})]$$

$$C_{i,m}, \mu_{i,m}, \Sigma_{i,m}$$

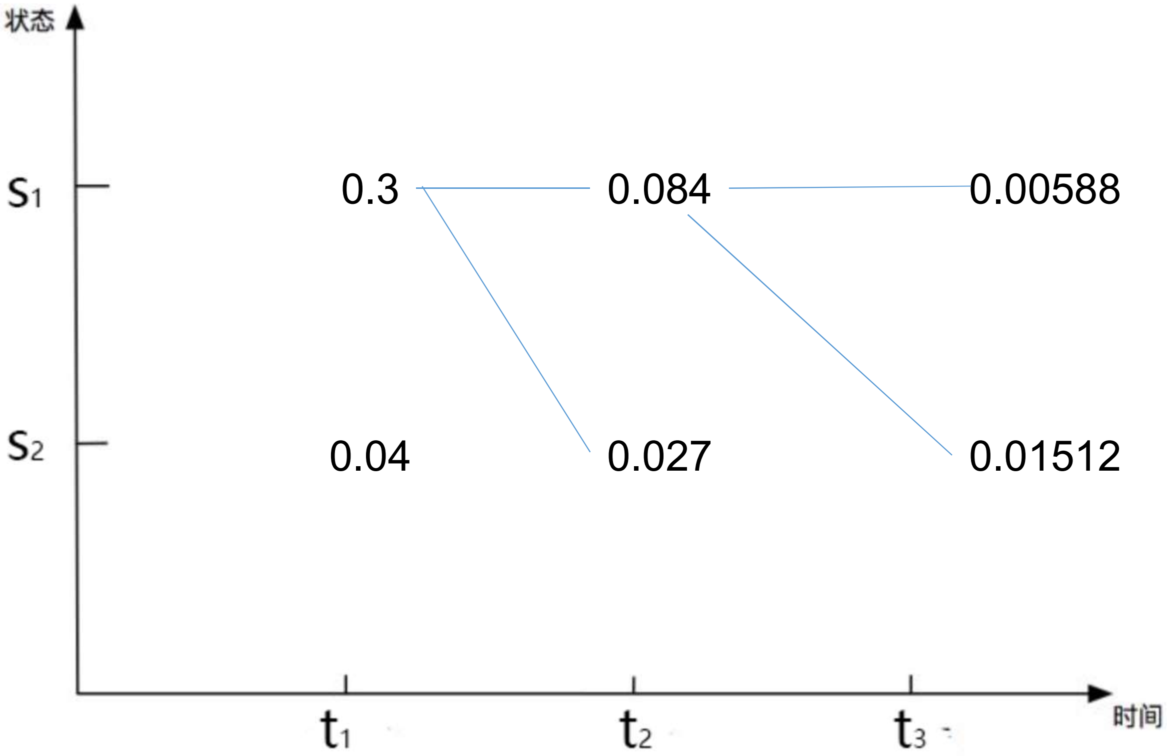
训练方法：Baum-Welch算法

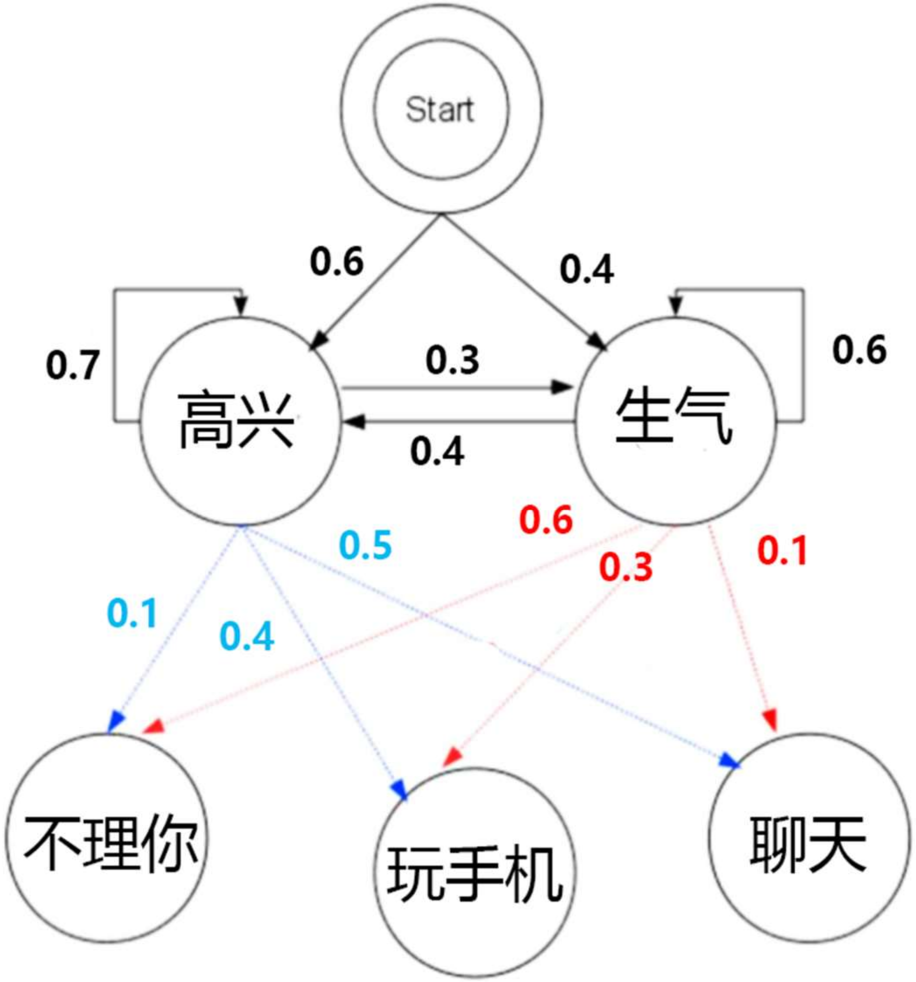
GMM-HMM 模型



维特比算法

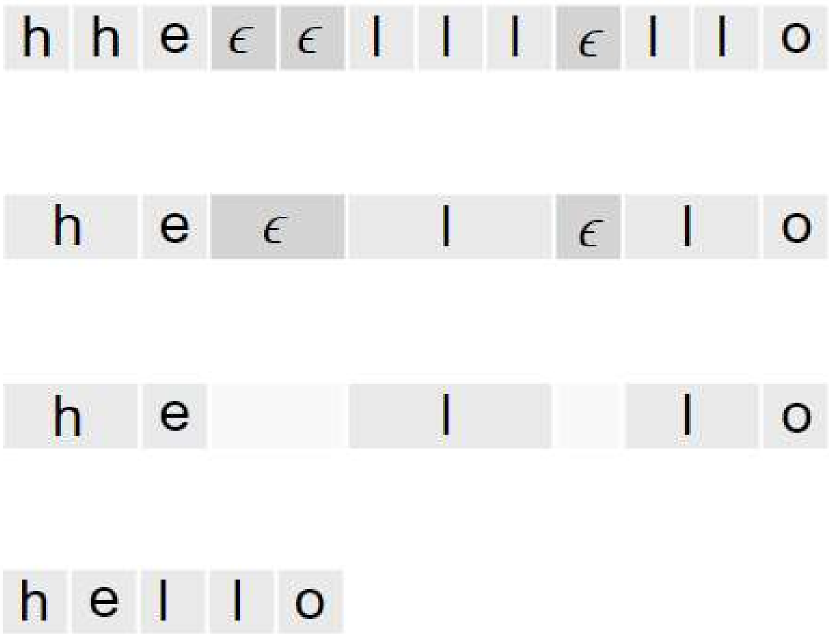
输入观测序列O





最终得到最优路径：S1,S1,S2

一种简单的分割



First, merge repeat characters.

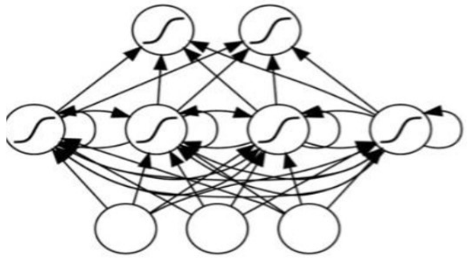
Then, remove any ϵ tokens.

The remaining characters are the output.

$$y \in \{a, b, \ldots, z, ?, ., !, \ldots, blank\}$$

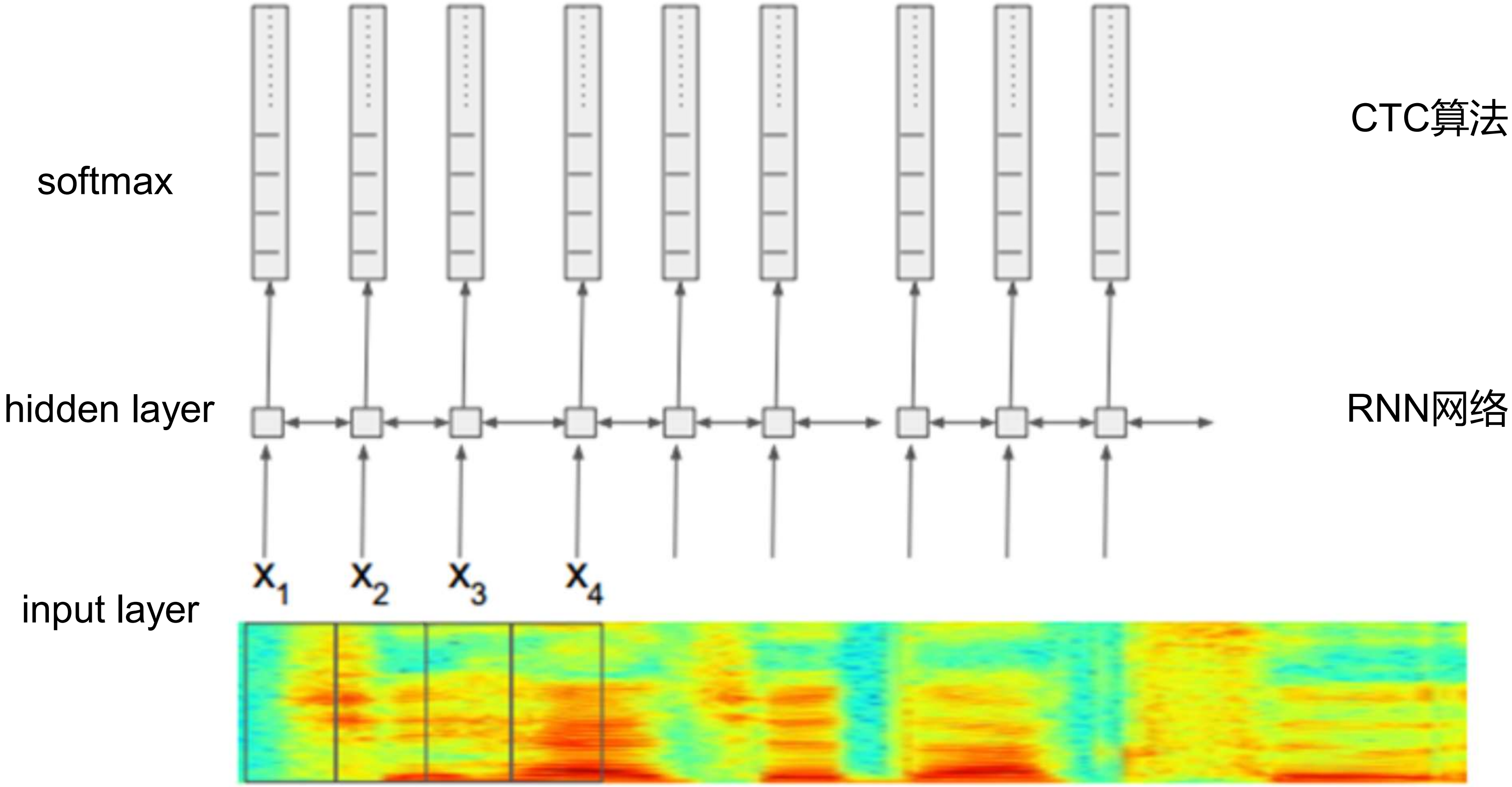
基于神经网络的END-TO-END模型

声学特征向量序列

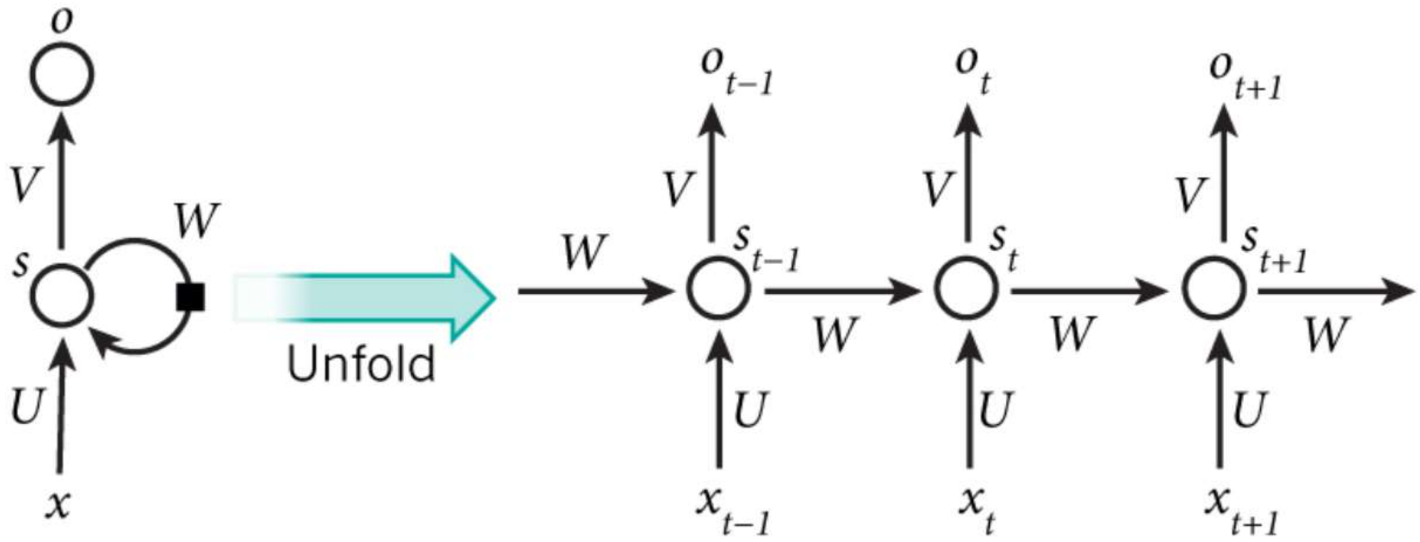


预测序列

神经网络结构



循环递归神经网络RNN



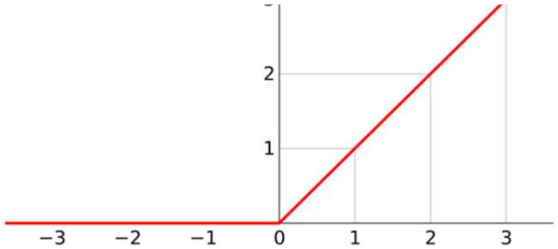
$$\mathrm{s}_{\mathrm{t}} = g(W * \mathrm{s}_{\mathrm{t-1}} + U * x_t + Bias)$$

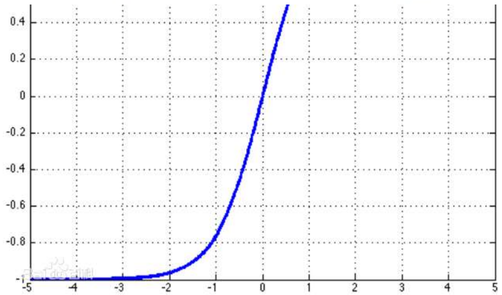
$$o_t = g(V * s_t + Bias)$$

梯度消失和梯度爆炸

$$\frac{\partial g_3}{\partial s_3} W \quad * \quad \frac{\partial g_2}{\partial s_2} W \quad * \quad \frac{\partial g_1}{\partial s_1} \frac{\partial s_1}{\partial W}$$

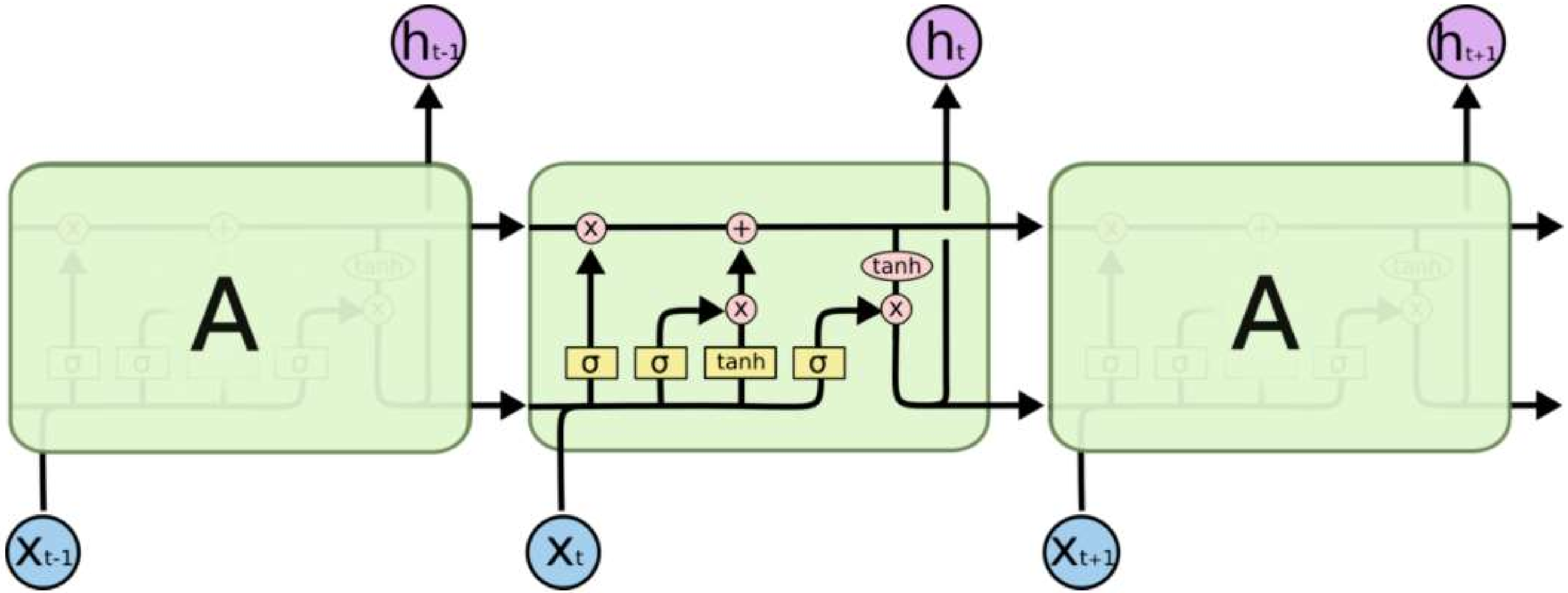
$$\frac{\partial g_3}{\partial W} = \frac{\partial g_3}{\partial s_3} g_2 + \frac{\partial g_3}{\partial s_3} W * \frac{\partial g_2}{\partial s_2} g_1 + \frac{\partial g_3}{\partial s_3} W * \frac{\partial g_2}{\partial s_2} W * \frac{\partial g_1}{\partial s_1} \frac{\partial s_1}{\partial W}$$



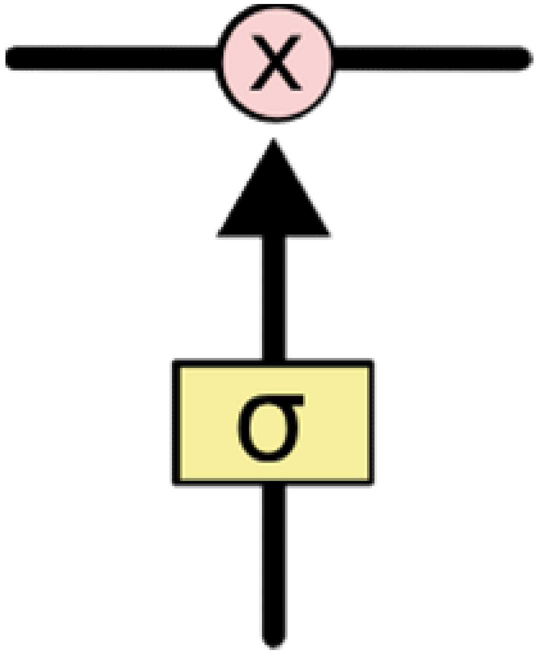


Today I want to make a steak, first of all I want to go to the farms to buy beef, then salt, vinegar, and finally go home to cook it.

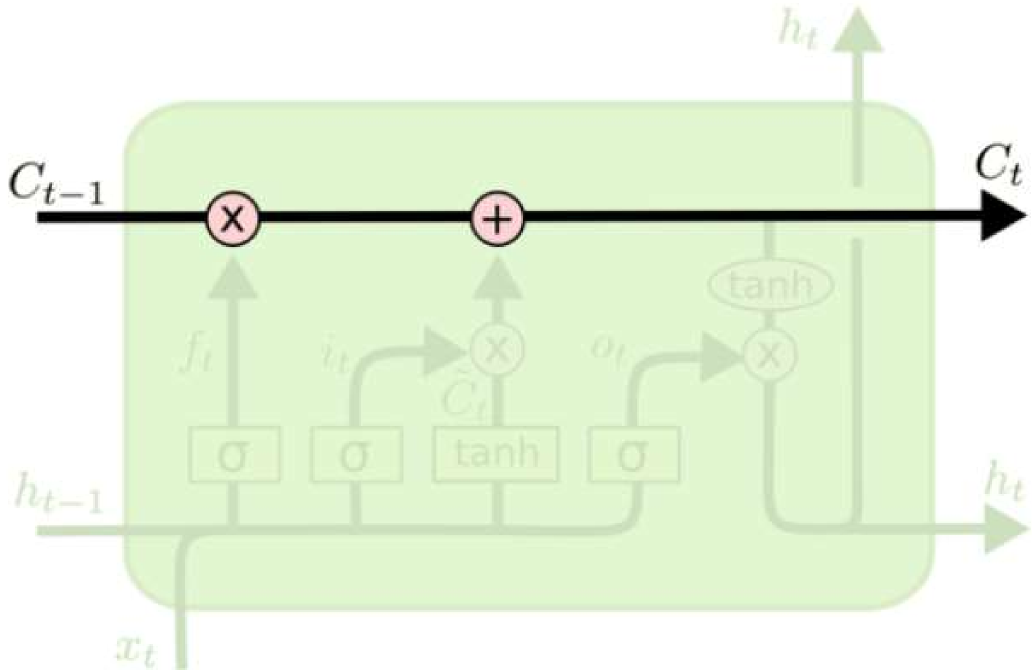
LSTM（Long Short-Term Memory）



LSTM 中的重复模块包含四个交互的层

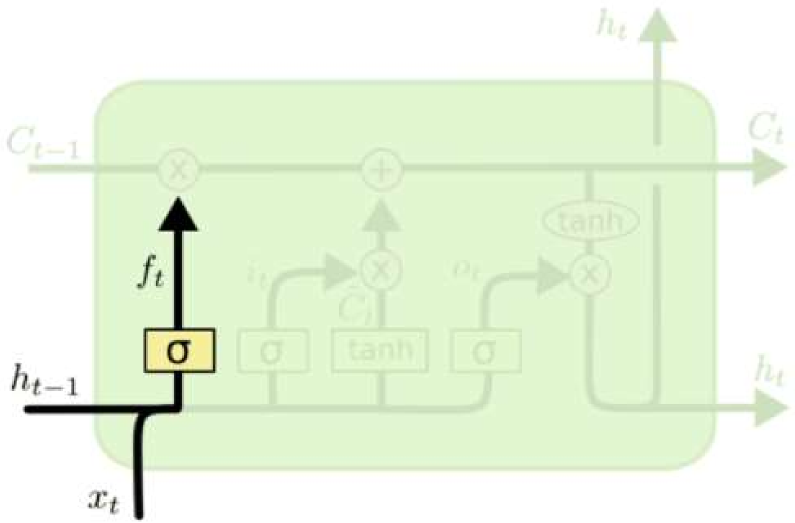

x
σ

细胞状态



忘记门

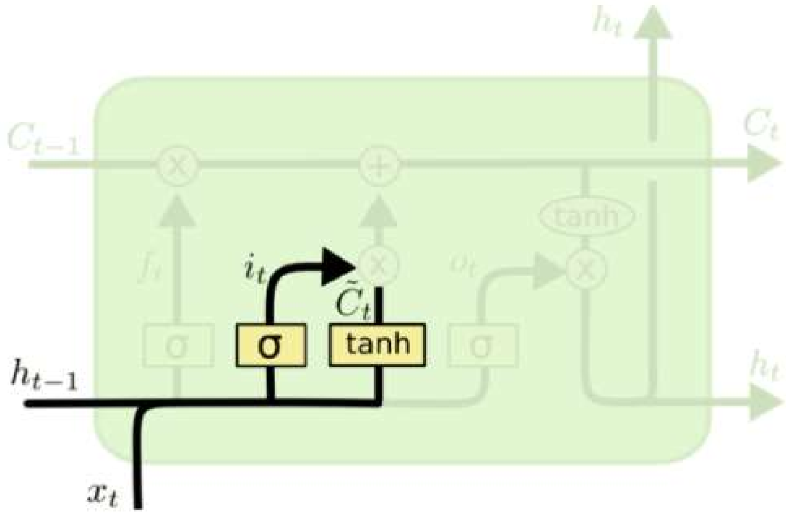
我说要吃饭，他说要吃面。



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

决定丢弃信息

输入门

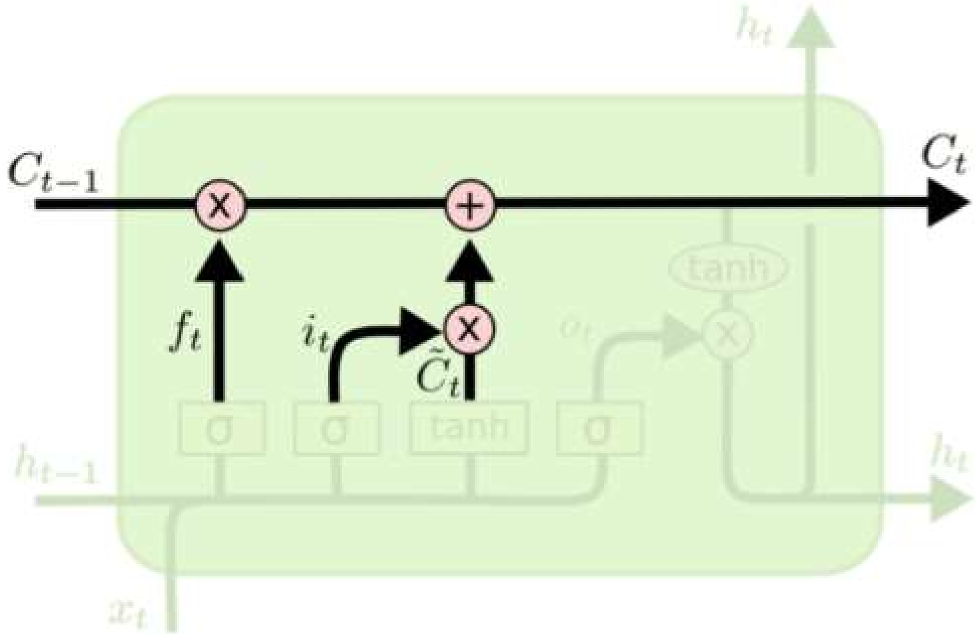


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

确定更新的信息

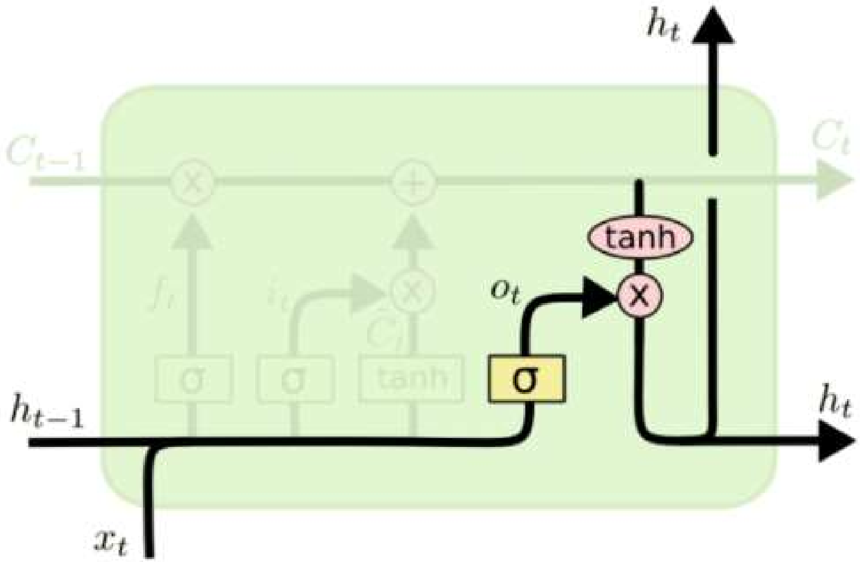
更新细胞状态



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

更新细胞状态

输出门

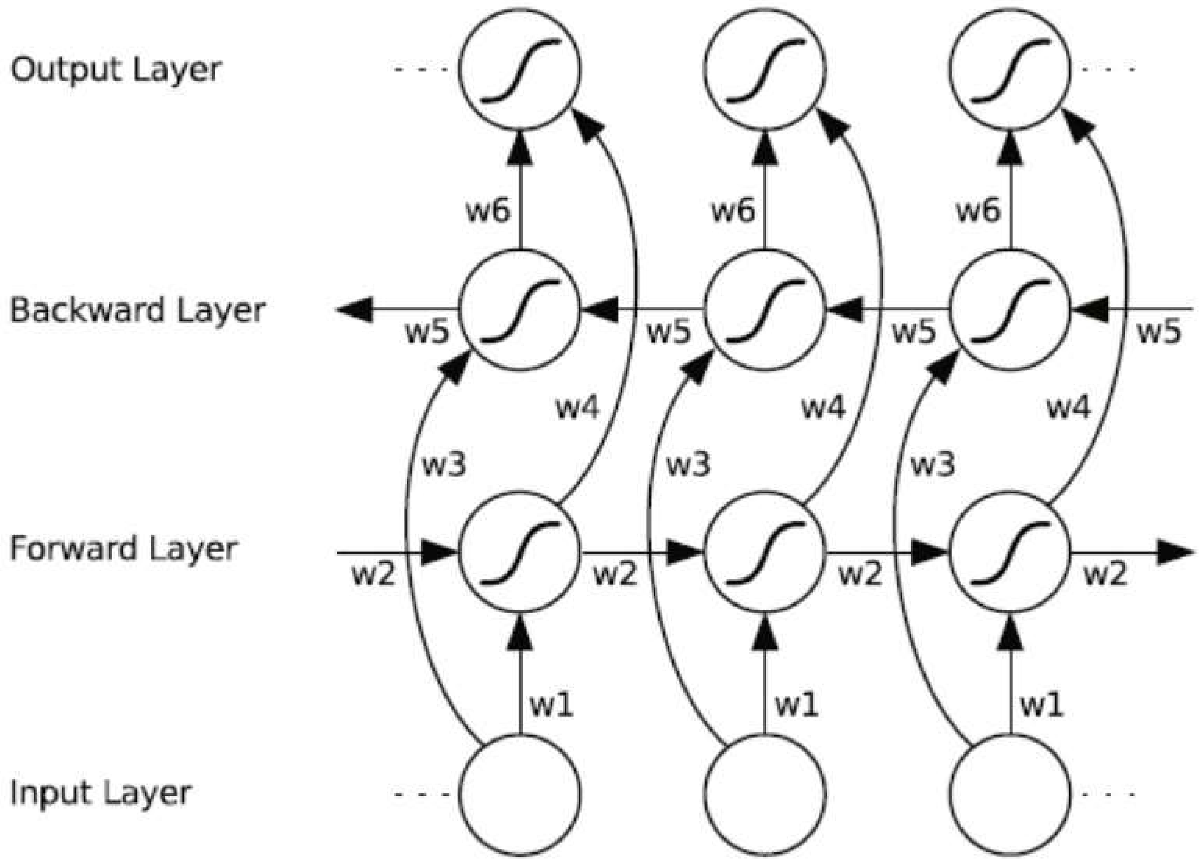


$$o_t = \sigma\left(W_o \left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

输出信息

BRNN



Output Layer
Backward Layer
Forward Layer
Input Layer
w6
w6
w6
w5
w5
w5
w5
w4
w4
w4
w3
w3
w3
w2
w2
w2
w2
w1
w1
w1

End to End的输入输出

- 输入是音频或者处理后的特征向量 $X=x_1x_2\dots x_T$
- Y是输出的序列 $Y=y_1y_2\dots y_L$ $y \in \{a, b, \dots, z, ?, ., !, \dots, blank\}$

$T>=L$

产生了问题：

1、X和Y都是可变长的

2、我们无法对X和Y进行精确的对齐

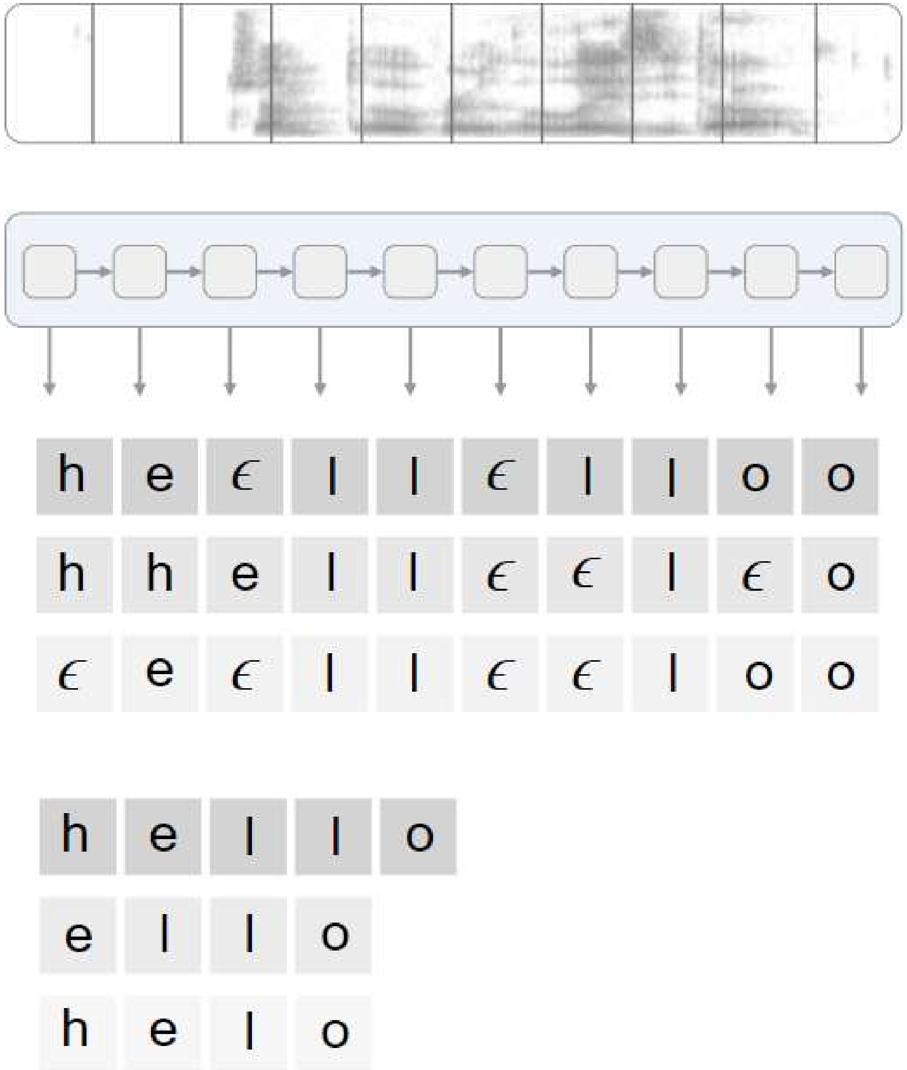
CTC (connectionist temporal classification)

- 不需要预先对数据进行对齐
- 直接输出序列预测的概率，不需要额外的路径搜索

CTC LOSS

Valid Alignments

ϵ	c	c	ϵ	a	t
c	c	a	a	t	t
c	a	ϵ	ϵ	ϵ	t



We start with an input sequence, like a spectrogram of audio.

The input is fed into an RNN, for example.

With the per time-step output distribution, we compute the probability of different sequences

By marginalizing over alignments, we get a distribution over outputs.

ROS中的集成语音识别包

- ROS中集成了CMUSphinx开源项目的代码，有适用于嵌入式的独立语音识别包pocketsphinx

indigo可以直接安装，其他版本需要先下载pocketsphinx包

对recognizer.py做修改添加lm，dict，hmm

```
self.asr.set_property('lm', '~/pocketsphinx/model/lm/en/tidigits.DMP')
self.asr.set_property('dict', '~/pocketsphinx/model/lm/en/tidigits.dic')
self.asr.set_property('hmm', '~/pocketsphinx/model/hmm/en/tidigits')
```

调用其他其他SDK

- 修改SDK的运行文件，添加需要的ROS接口

```
ros::Subscriber voiceSub = n.subscribe(..)
ros::Publisher wordSub =  n.advertise<std_msgs::String>(..)
while(ros::ok(){
    检测到语音{
            调用SDK
        }
    语音识别完成
        {
            发布结果
        }

    }
```

不要英音不要美音！就要我大中华纯正中式口音！

只要15分钟，蹲个坑就录完了

各位爸爸们来录个语料库吧！

录完所有句子通过审核就有10块软妹币

可以买杯奶茶啊！

邀请好友一起来录音，你就能拿5元红包啊，

每天安利2个好友，**天天**喝奶茶啊！

微信扫一扫小程序码就能录了，随时随地

只要环境安静就行，在家里在寝室在天台都可以录啊！

英语语料库

初中英语水平就够了