优必选
UBTECH

智能语音信号处理及应用

项目 3
智能小义之学会聆听

CONTENT 目 录

课程导入

AI 技术在日常生活中应用广泛，其中语音技术在很早的时候就尝试走入大家的生活。从亚马逊的 Echo 到微软的 Cortana ，从苹果的语音助手 Siri 到谷歌的 Assistant 等等，语音识别技术的广泛应用让生活便利了很多，它们的工作原理又是什么？你是否会对这些智能化的语音回答感到兴趣满满呢？

语音识别技术

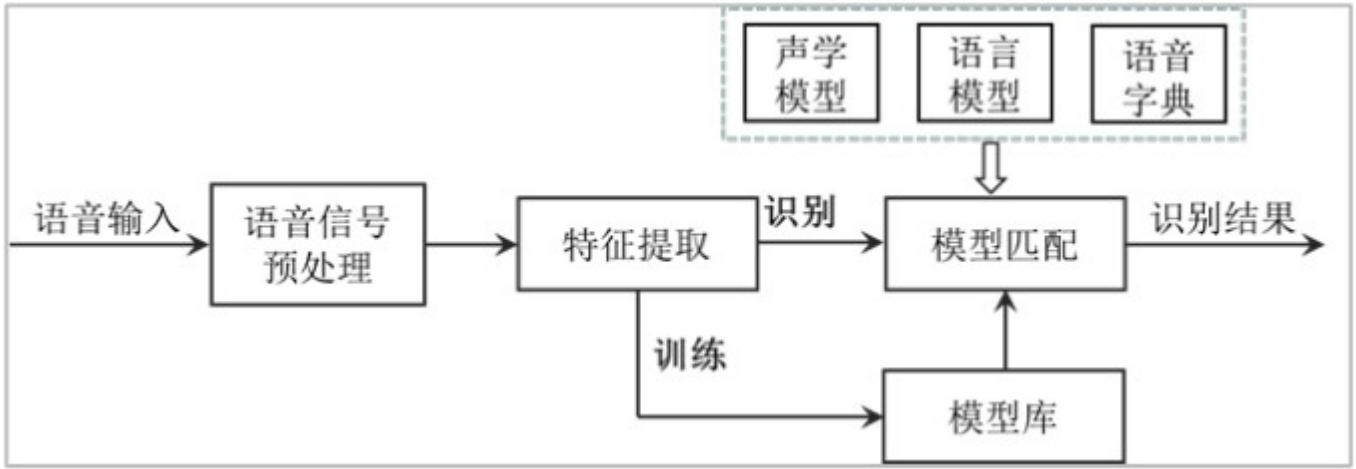
语音识别是试图使机器能“听懂”人类语音的技术。其作用是将语音转换成等价的书面信息，即让计算机听懂人说话。就好比“机器的听觉系统”，其目的就是赋予机器人听觉特性，能将听到的语音信号转变为相应的文本或命令。

商场中和机器人的咨询

语音识别技术

语音识别本质是一种基于语音特征参数的模式识别，主要包含 4 个部分：语音信号预处理、特征提取、模型库和模式匹配。

语音识别原理框图如图所示。



声学模型

声学模型是语音识别模型中用来识别声音的模型，是语音识别系统的重要组成部分，它决定了语音识别中大部分的计算开销与语音识别系统的性能。在识别时可以将待识别的语音的特征参数和声学模型进行匹配，得到识别结果。

目前的主流语音识别系统多采用隐马尔可夫模型 HMM 进行声学模型建模。

声学模型

在深度学习兴起前，传统的统计参数模型是主流的声学模型建模方法。

这是由于传统的统计模型一般具有比较简单的假设，无论从理论还是公式上都能给出非常完整的解释和推导，便于人们的理解。主要包括以下几种：

1. 混合高斯模型
2. 联合概率密度混合高斯模型
3. 隐马尔可夫模型

声学模型

1. 混合高斯模型

假设一个多元随机变量为 γ = (γ1, γ2, ..., γD)T，如果该随机变量符合多元高斯分布，则其概率密度函数为

$$p(\gamma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum|^{\frac{1}{2}}} exp(-\frac{1}{2}(\gamma - \mu)^T \sum^{-1} (\gamma - \mu))$$

声学模型

2. 联合概率密度混合高斯模型

联合概率密度混合高斯模型（joint-density GMM,JD-GMM）是一种常用来对特征匹配的输入输出之间关系建模的模型 ， JD-GMM 通过对输入和输出向量的联合向量进行统一建模来得到生成式的全概率模型。

在实际应用中，在得到输入输出的联合概率分布以后，直接通过计算给定输入向量，得到的输出数据的条件概率分布，为最终需要的预测模型。

声学模型

3. 隐马尔可夫模型

隐马尔可夫模型 HMM 是指这一马尔可夫模型的内部状态外界不可见，外界只能看到各个时刻的输出值，因此马尔可夫模型的概念可被理解成是一个离散时域有限状态自动机。一般情况下，对于语音识别系统，输出值是指从各个帧计算中得出的声学特征。

语言模型

语言模型是根据语言的客观事实而对语言进行抽象的数学建模。语言模型可以估计一段文本的概率 ，在信息检索，机器翻译，语音识别等方面起着重要的作用。

语言模型分为统计语言模型和神经网络语言模型。

语言模型

统计语言模型

统计语言模型的基本思想是计算条件概率，若要判断某一段文字 w1, w2, w3, ..., wm 是否为一句合理的话，其中 w 为某个词，则可以先计算其联合概率，如公式所示：

$$P(w1, w2, w3, ..., wm) = P(w1)P(w2|w1)P(w3|w1, w2)\cdots P(wm|w1, w2, w3, ..., wm\ 1)$$

语言模型

统计语言模型

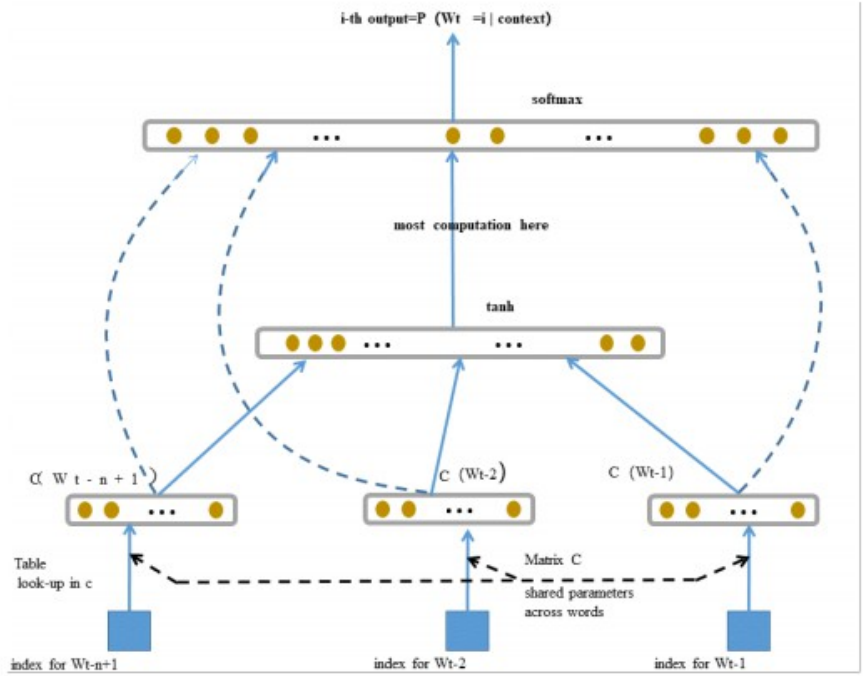
由于在实际操作中会出现文本较长的情况，那么上式中的计算会变得较为复杂，这时应该使用更为简化的模型 N-gram 语言模型，该模型中第 n 个词的出现只与前面 n-1 个词有关，与更前面的词无关，因此公式被改写为：

$$P(wi|w1, w2, ...wi\ 1) = P(wi|wi\ (n\ 1), ..., wi\ 1)$$

语言模型

神经网络语言模型

前馈神经网络模型 FFLM 它共有三层，分别是输入层，隐藏层和输出层，其在计算时利用全连接的神经网络模型来估计给定的 n-1 个上文的情况下，计算第 n 个单词出现的概率。

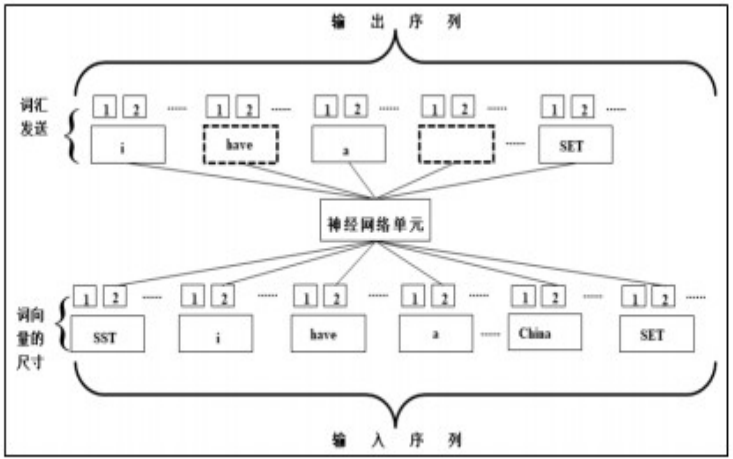


语言模型

循环神经网络语言模型 RNNLM

最常用的循环神经网络有双向循环神经网络和长短期记忆网络，其预测的当前词都与之前的词有关。

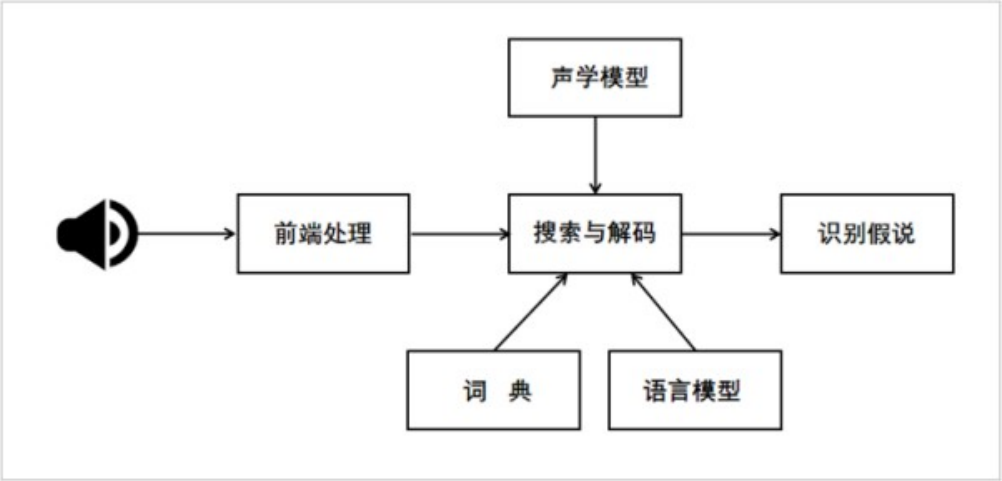
在绘制该模型的结构图时，往往只需要将隐藏层进行更改即可，例如将神经网络语言模型 NNLM 的隐藏层转换成循环神经网络 RNN 神经元。



语音词典

语音词典包含了从单词（words）到音素（phones）之间的映射，作用是用来连接声学模型和语言模型的。

语音词典在语音识别过程中的位置如图所示。



关键函数功能说明

字体“楷体”，字号 24 号（可依据需要微调），段落行距 1.5 倍。尽可能图文并茂，每页 PPT 最好都用有图片配合文字。

如果需要使用动画，尽可能不要有堆叠覆盖的情况发生，如果堆叠覆盖，将来生成 pdf 将影响整体效果。

如果插入了动图、视频、音频等，请保证换电脑能够正常运行。

总结

1. 理解语言识别的概念；
2. 理解语言识别的原理和方法；
3. 理解语音识别模块中声学模型、语音字典和语言模型的概念；
4. 理解开源语音识别工具包 pocketsphinx 的原理

谢谢大家！