



语音识别原理

Automatic Speech Recognition Principle

清华大学基础工业训练中心

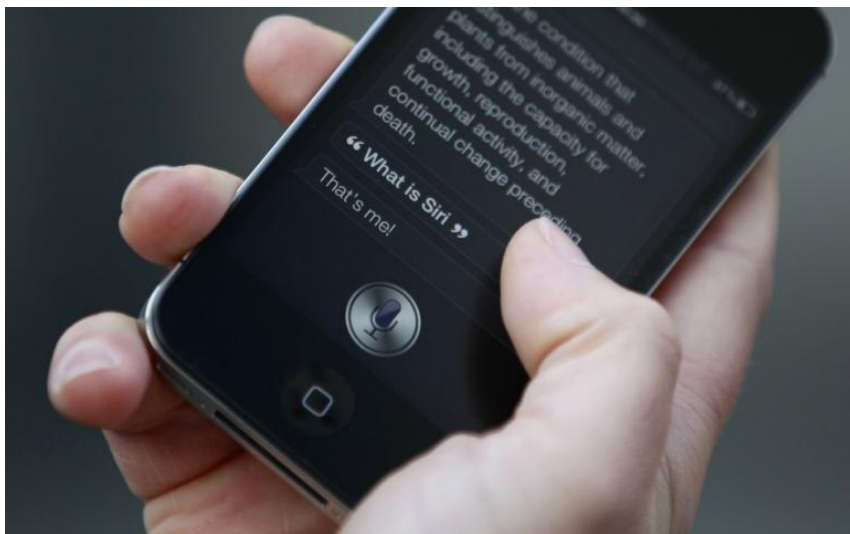
智能音箱 (Smart Speaker)

- 2014年11月，亚马逊推出了全新概念的Echo智能音箱。
- 截止到2017年年底，全美共有超过4000万台智能音箱正在使用，其中亚马逊是3000万（Cirp数据）
- 2018年2月1日，美国智能音箱的线上市场份额，Echo全系列占比约70%，Echo show在Echo阵营中占比7%（intelligence 数据）



语音助手（Voice Assistant）

- 语音识别是智能助手的第一步
- 苹果Siri，微软Cortana，谷歌Home，亚马逊 Alex
- 语音识别ASR与问答系统QA

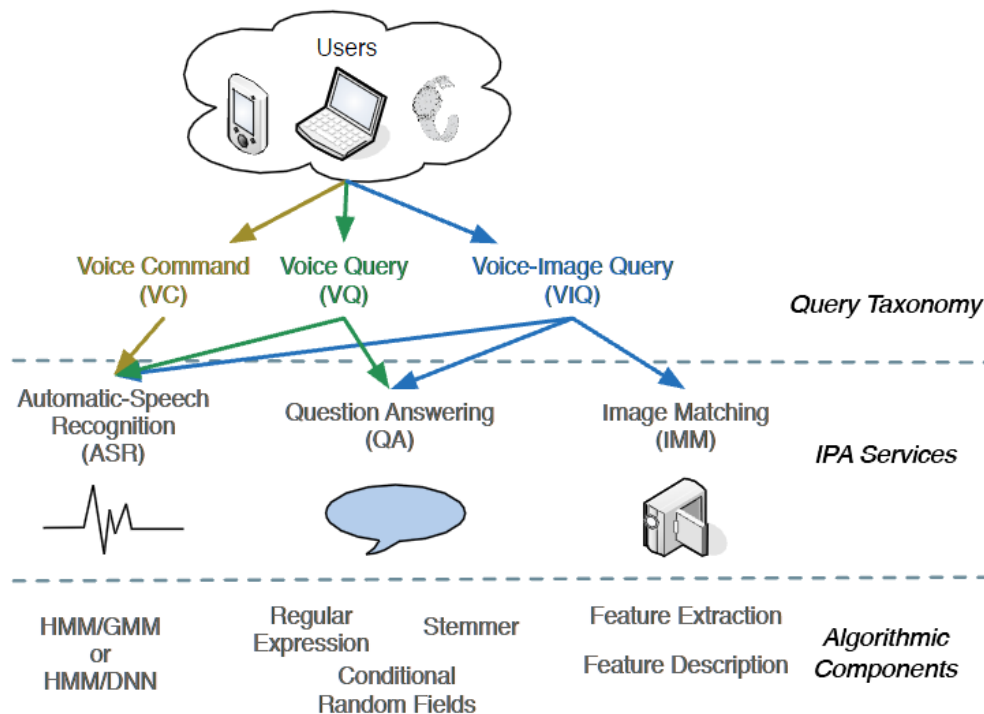


语音交互(voice interaction)

- 语音控制，语音查询，语音图像查询（演示视频）



语音



[x] Hauswald, Johann, et al. "Sirius: An open end-to-end voice and vision personal assistant." ACM PLOS, 2015.

语音识别与交互

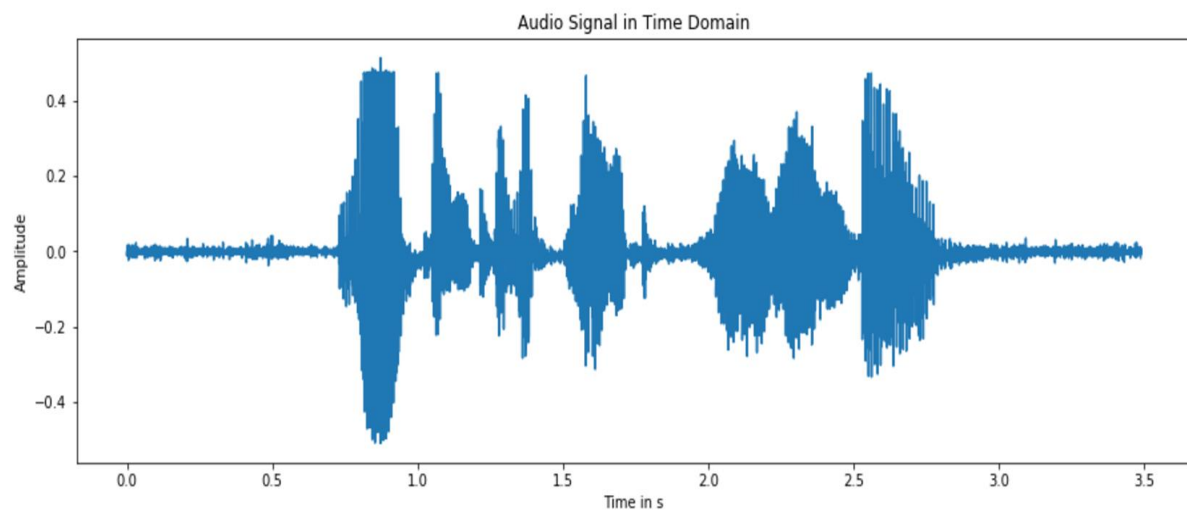
- 语音识别（Automatic Speech Recognition (ASR)）
 - 语音（信号波形）→文字（自然语言）
- 语音合成（Text To Speech (TTS)）
 - 文字（自然语言）→语音（信号波形）
- 科大讯飞/出门问问/思必驰/捷通华声/普强
Pachira

语音信号探究

- 语音信号是声波
- 信号的表示
 - 波形图（Waveform）
 - 频率图（Frequency）
 - 时频谱（Spectrogram）

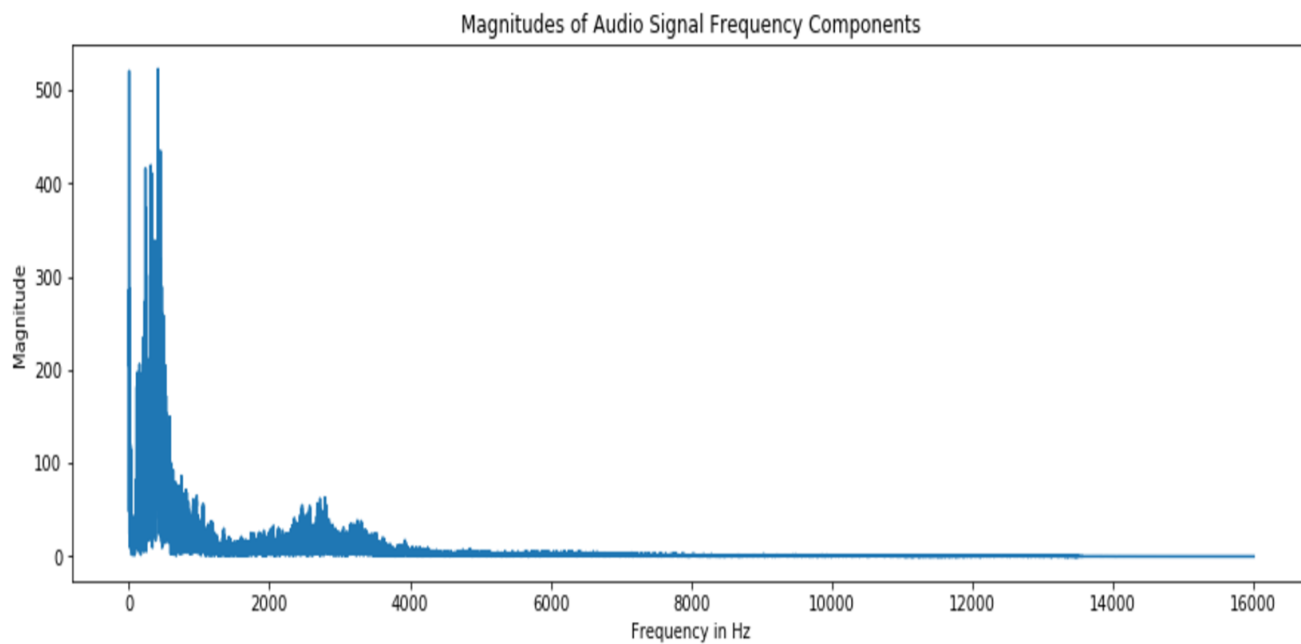
波形图

- 语音信号是时变信号



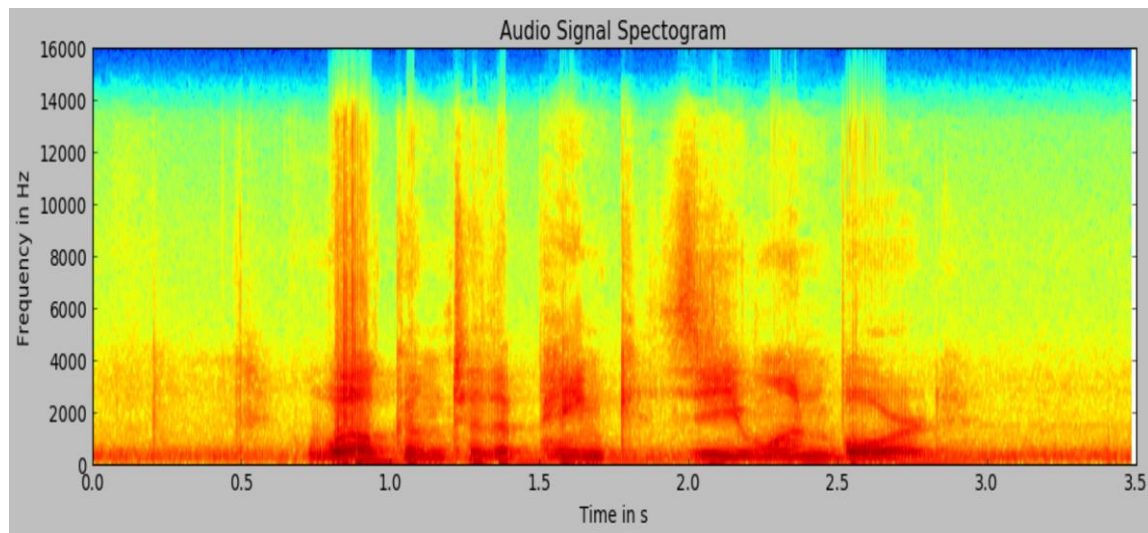
频域图

- FFT（快速傅立叶变换）

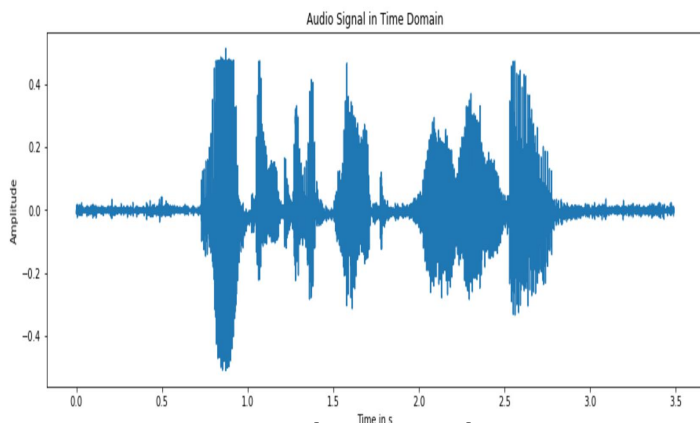


时频谱

- 将波形文件通过STFFT变换
- 优点：保留了时间与频域的信息



语音识别的本质



$$v=f(t, h1)$$



拨打电话

$$w=F(v, h2)=F(f(t, h1), h2)$$

- $w = F(v, h)$ 函数，其中：
 - v 代表语音信号，是 t 的函数 $f(t, h1)$ ；其中 $h1$ 是隐藏变量
 - w 代表文字输出； h 同样代表权重参数；

语音识别的本质

- 映射（ Mapping ）：两个集合之间的关系，如满射、单射、一一对应即双射（满射和单射）
- 函数（Function）：定义 x, y 变量，对于每个 x 数值，按照一定法则总有确定的数值 y 和它对应，则称 y 为 x 的函数， x 叫自变量， y 叫因变量

语音识别的本质

- 函数逼近问题
- 理论问题：就是如何找到F函数和H参数？
- 实际中，如果找不到精确的F函数和H参数，是否可以找到这些函数的近似
- 解决方案：用深度神经网络逼近一个 $F(h, x)$ 函数

语音识别

深度学习方法

语音识别原理

- 经典的语音识别过程

- 语音信号经过 **傅里叶变换STFFT**，把 **连续语音** 时间上分解成一组 **短期向量**，然后应用各种变换把这个 **向量序列** 变换为一个 **音素序列**，然后变换到 **字母序列**，然后到 **词汇序列**。

- 基于 **深度学习** 语音识别

- 深度学习模型需要训练数据和评价驱动的方法来进行参数优化。

语音识别（经典方法）

- 语音信号处理（传统）：
 - 经过短时傅里叶变换（STFFT），把连续语音分解成一组短期向量序列
 - 变换向量序列为一个或多个音素序列
 - 变换音素序列为一个或多个字母序列
 - 变换字母序列为一个或多个词汇序列
- 无深度学习技术，主要采用
 - 隐藏马尔科夫模型（HMMs），
 - 上下文相关模型（context-dependent phone models），
 - n-gram 语言模型（n-gram language models）
 - 维特比搜索算法（Viterbi search algorithms）

混合模型（合金模型）

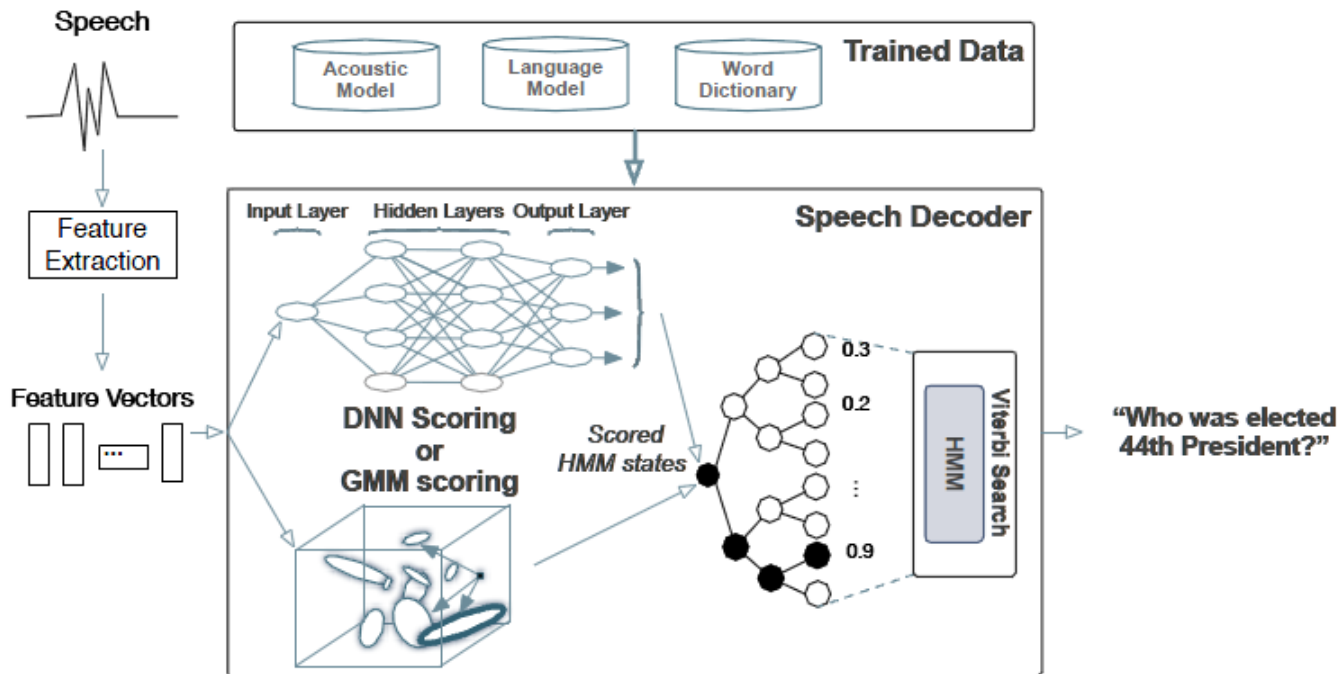
- 引入深度学习技术，深度神经网络（DNN）与隐藏马尔科夫模型（HMMs），上下文相关模型（context-dependent phone models），n-gram 语言模型（n-gram language models）和维特比搜索算法（Viterbi search algorithms）进行混合使用。
- 这个混合模型比较复杂，需要复杂的训练方法，很多的专业知识来搭建语音识别模型。

纯深度学习模型

- 深度学习基于通用的神经网络（DNN），调整网络参数（训练）来优化可微分的成本、代价或损失函数（loss/cost function）。
- 纯 DNN 方法在语音识别上取得了巨大的成功。
- 拥有了相当多的训练数据和足够的计算资源，就可以构建一个高水准的大词汇量连续语音识别（Large Vocabulary Continuous Speech Recognition (LVCSR)）系统。
- <https://www.intel.ai/end-end-speech-recognition-neon/>

基于深度学习方法的语音识别

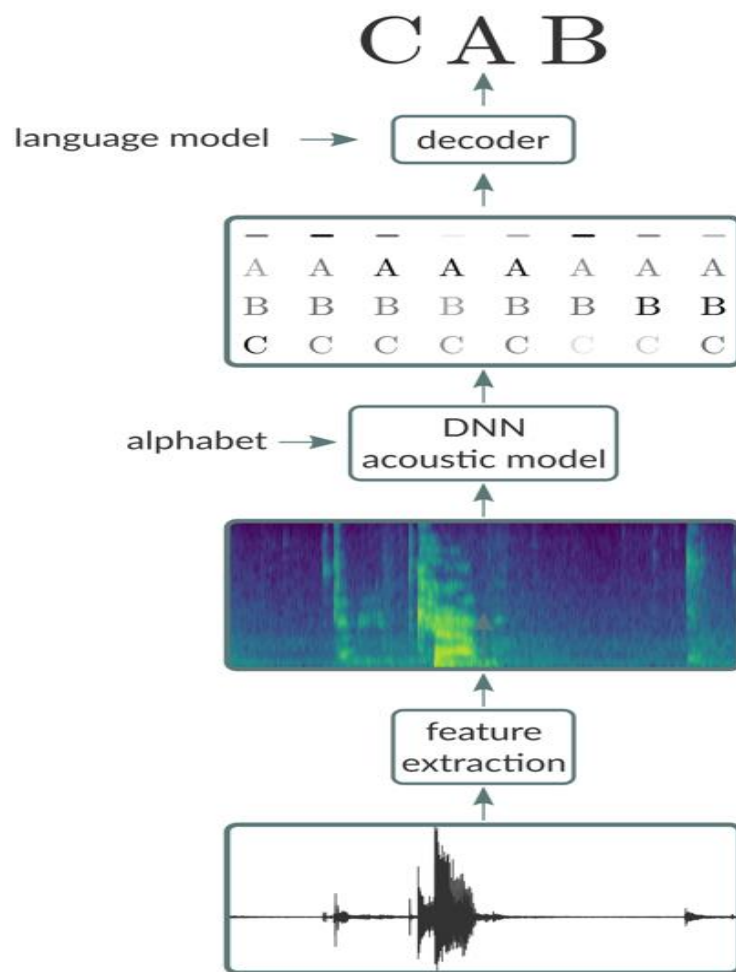
- 基于深度学习的语音识别：语音信号特征提取，神经网络输入，神经网络运算，神经网络输出
- 深度学习使得语音识别错误率在以往最好系统的基础上**相对下降了30%或更多，突破了语音识别真正实用的临界点**



端到端的深度学习语音识别模型

- 端到端语音识别流水线由三个主要部分组成：特征提取、声学模型、语言模型
1. 将原始音频信号（如 **wav** 文件）作为输入，并产生特征向量序列，给定音频输入帧的特征向量。特征提取的输出示例包括原始波形、频谱图和时频谱图。
 2. 将特征向量序列作为输入，并产生以特征向量输入为条件的字符或音素序列的概率的声学模型。
 3. 采用两个输入（声学模型的输出以及语言模型）的解码器，在受到语言模型中编码的语言规则约束的声学模型生成的不同序列，搜索其中最可能的转录。

端到端的深度学习模型

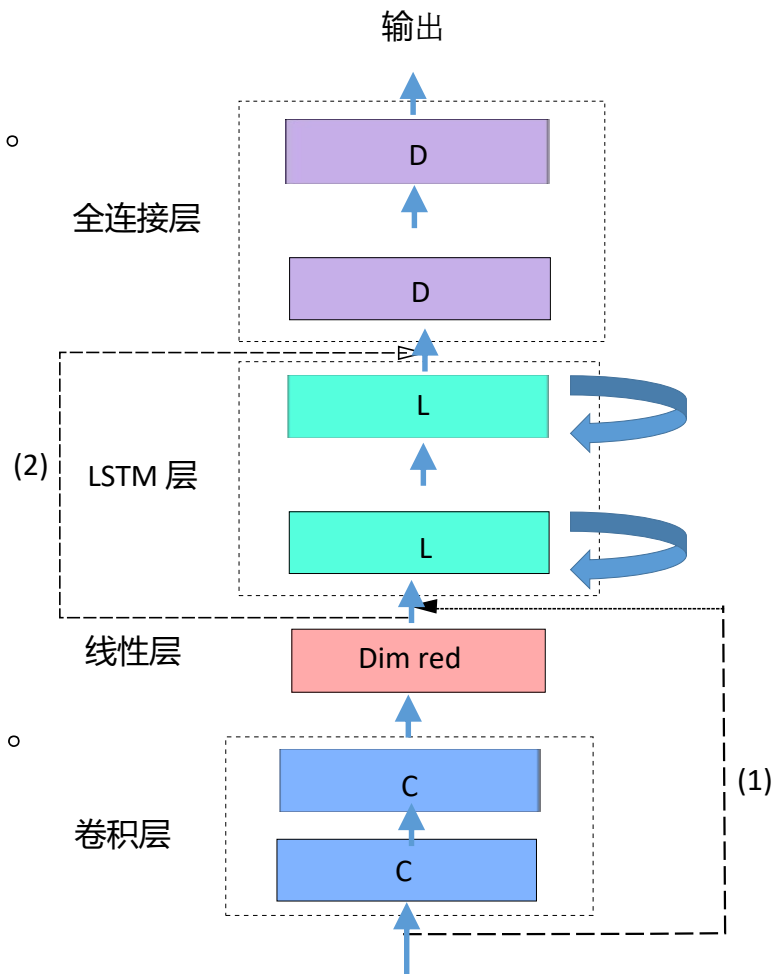


<https://www.intel.ai/end-end-speech-recognition-neon/>

谷歌CLDNN

- CLDNN结合了卷积网络，LSTM和DNN。
 - 当输入信号进行时间域的卷积操作之后，输出数据再进行一次频率域的卷积操作以减少频谱的变化，之后再通过三层LSTM，最后再通过一层DNN。
 - 训练过程中，时间卷积层和其他层会一起进行训练。
- 输入数据为以时间为下标的连续向量。

此次DNN专指为全连接网络



参考文献

- Recent progresses in deep learning based acoustic models, 2017
- Highway-LSTM and Recurrent Highway Networks for Speech Recognition, 2017
- Anchored Speech Detection, 2016
- Connectionist Temporal Classification Labelling Unsegmented Sequence Data with Recurrent Neural Networks, 2006

谢谢指正！

zhenchen@Tsinghua.edu.cn