Audio-visual Fusion, (Deep) Speech Recognition and Beyond

An Ongoing Survey

Yuan-Hang Zhang

May 6, 2018

University of Chinese Academy of Sciences

目录

基础知识回顾

语音识别的挑战

- **自然** (spontaneous) 语音 vs **朗读** (read) 语音
- 大词表
- 噪声
- 低资源 (low-resource)
- **远场** (far-field)：混响 (reverbation)
- 口音
- 说话人自适应 (speaker-adaptive)

短时傅里叶变换

Short-Term Fourier Transform, STFT

语音信号：局部平稳，全局非平稳

把一段长信号分帧（较短的等长信号）、加窗，再对每一帧做傅里叶变换

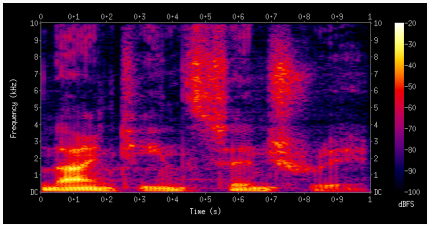
$$\mathbf{STFT}\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}\,dt$$

波形乘**窗函数** (window function) w（在给定区间之外取值均为 0 的实函数）不仅是为了不在窗口边缘两端引起急剧变化，还相当于对信号谱与窗函数的 Fourier 变换进行卷积。

声谱图 (spectrogram)

- **出发点**：经过训练你真的可以读出声谱图里的音素

  见 *How to read a spectrogram*, Robert Hagiwara
- 加窗的离散 STFT：能量 vs 频率 [离散] vs 时间 [离散]



$$\mathbf{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n}$$

$$\text{spectrogram}\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2$$

Mel 频率表示

模型训练的输入特征要求对相位不敏感，传统特征基于原始音频的能量谱得到，如 **MFCC**（Mel 倒谱系数）和 **Mel Filter Bank**（Mel 滤波器组）.

- FFT 维数较高
- $m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$
- 得到对数 Mel 特征

Mel 的名字来源于单词 melody，表示这个刻度是基于音高比较而创造的.

传统特征提取

- **频域**：在傅里叶变换之后使用人工设计的滤波器组来提取特征，造成频域上的信息损失；高频区域尤为明显

传统特征提取

- **频域**：在傅里叶变换之后使用人工设计的滤波器组来提取特征，造成频域上的信息损失；高频区域尤为明显
- **时域**：为了计算量的考虑必须采用非常大的帧移，造成时域上的信息损失；说话人语速较快时问题明显

传统特征提取

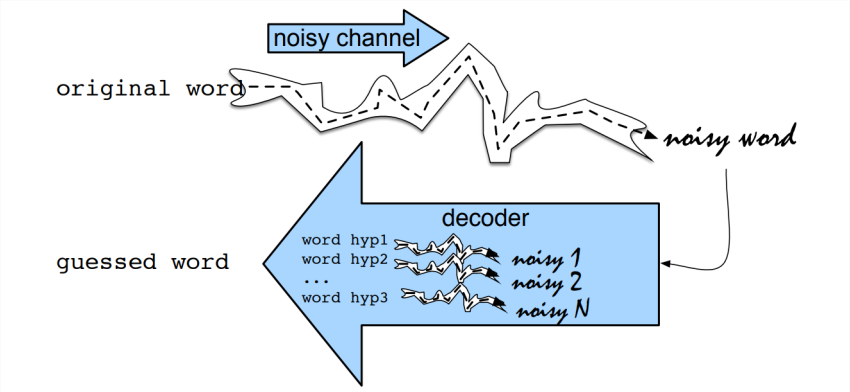
- **频域**：在傅里叶变换之后使用人工设计的滤波器组来提取特征，造成频域上的信息损失；高频区域尤为明显
- **时域**：为了计算量的考虑必须采用非常大的帧移，造成时域上的信息损失；说话人语速较快时问题明显
- 改进方案：将声谱图作为输入

传统特征提取

- **频域**：在傅里叶变换之后使用人工设计的滤波器组来提取特征，造成频域上的信息损失；高频区域尤为明显
- **时域**：为了计算量的考虑必须采用非常大的帧移，造成时域上的信息损失；说话人语速较快时问题明显
- 改进方案：将声谱图作为输入

统计语音识别

噪声信道模型 (Noisy Channel Model)



统计语音识别 (cont'd)

根本问题

给定声学输入 O，语言 L 中最可能对应的句子是什么？

Statistical ASR in One Formula

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{O}) = \arg\max_{\mathbf{W}} \underbrace{p(\mathbf{O}|\mathbf{W})}_{\text{声学模型}} \underbrace{P(\mathbf{W})}_{\text{语言模型}}$$

其中，观察序列 $\mathbf{O}$ 是声学特征矩阵。

- **声学模型**：给出语音属于某个声学符号的概率

  声学符号：音节 (syllable) / 音素 (phoneme)；声韵母等
- **语言模型**：给定历史，预测下一个词的概率

声学模型 (AM)

声学建模单元 (phonetic units)

- **音素** (phoneme, phone): “cat”→ /K/, /AE/, /T/
- **上下文无关的 HMM 状态**: k_1， k_2， $ae_1 \ldots$
  - 分别建模起始态 (onset)、中间态 (middle)、结束态 (end)
- **上下文相关的状态**: $k_{1.17}, \ldots$
- **上下文相关的音素 (CD 音素)**
  - “我吃了”→“w o ch i l e”
  - (双音子, diphone) → “sil-w w-o o-ch ch-i i-l l-e”
  - (**三音子**, triphone) → “sil-w-o w-o-ch o-ch-i ch-i-l i-l-e”
  - 经决策树聚类后得到的绑定三音子 (tied triphone states) 称为 *senone*

声学模型 (AM) (cont'd)

声学建模单元（续）

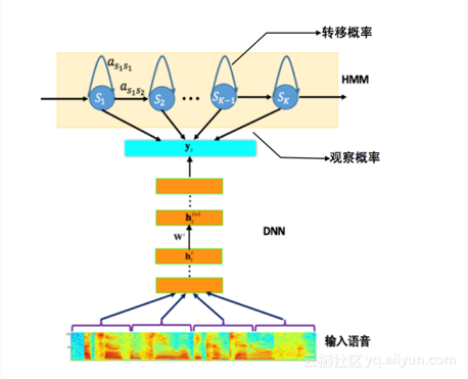
- **音节** (syllables)
- **字素** (grapheme): 一个字符 (character)，如英文字母或汉字（不区分异形、大小写变体）
- **字片** (word piece): 词的组成部分 (sub-word units)
- 整个词: 对生僻词泛化不好

高亮者被用于序列到序列模型，不像音素，它们不需要引入额外的发音模型和语言模型.

逐帧分类

Framewise classification (HMM-GMM/DNN)

- 语音波形经过加窗、分帧，提取出频谱特征
- 将输入的声学特征映射得到不同输出建模单元的后验概率
- 结合 HMM 进行解码，得到最终的识别结果
- **缺点**：声学模型的训练未与最终目标挂钩，逐帧准未必最后准



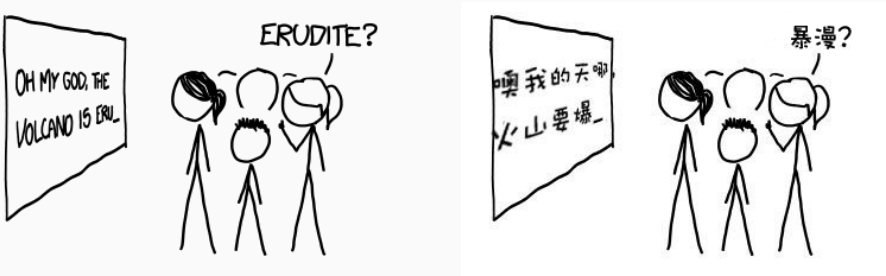
强制对齐 (forced alignment)

- GMM 如何训练？——**EM 算法**
  - 用模型将声学特征与语音状态进行极大似然对齐
  - 训练时，用 Viterbi 算法强制要求搜索经过真实的词序列
  - **E 步**：确定状态序列，从而为每帧给出音素标注；**M 步**：估计参数 (μ, σ)
- 前面提到的 DNN 用什么训练？
  - “输出层的标注一般采用 GMM-HMM 基线系统经强制对齐得到”
  - 天道好轮回，苍天饶过谁
- 题外话：**注意力机制** (attention mechanism)
  - 完全注意力没有*单调对齐* (monotonic alignment)
  - 后面讨论 Deep Speech 3 时还会再谈

语言模型 (LM)

噢我的天哪，火山要爆_

“a volcano threatens to flood a town with information”



扇贝听了都想鼓掌

语言模型 (cont'd)

概率估计

$$P(w_1, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1 \ldots w_{i-1})$$

通常采用 **N 元文法** (N-gram) 估计右侧连乘的概率：

$$P(w_i \mid w_1 \ldots w_{i-1}) = P(w_i \mid w_{i-n+1} \ldots w_{i-1})$$

更多内容请参见自然语言处理教材，如 [Jurafsky 2009] 等

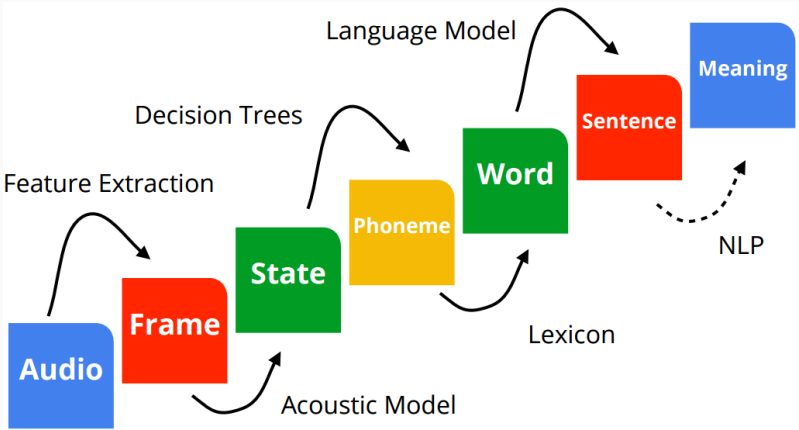
总结：语音识别做什么？

- 把**帧** (frame) 识别成**状态** (state)
  - 以前是 GMM 来做这件事，现在是 DNN 来做
  - **取代原因**：DNN 可以拼接相邻语音帧作为输入，描述长时结构信息；输入特征可以是多种离散或连续特征的融合；不需要对语音数据分布进行假设
- 把**状态**组合成**音素** (phenome)
- 把**音素**组合成**单词** (word)
  - 发声词典可由 Grapheme-to-Phoneme (G2P) 产生
- 对不同**单词**组成的**句子** (sentence) 进行重评分

$$\hat{w} = \arg\max_{w \in \Sigma^*} P(w) \cdot \sum P(o|s)P(s|p)P(p|w)$$

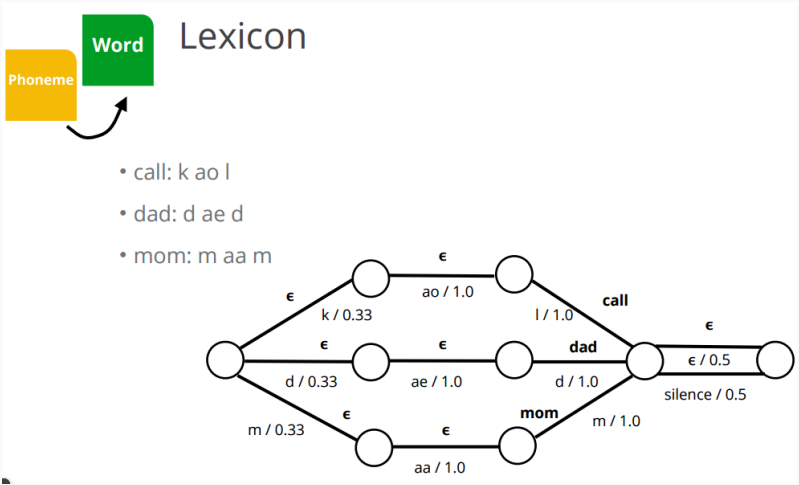
重新审视……

把语音识别视为一个**转换** (transduction) 问题



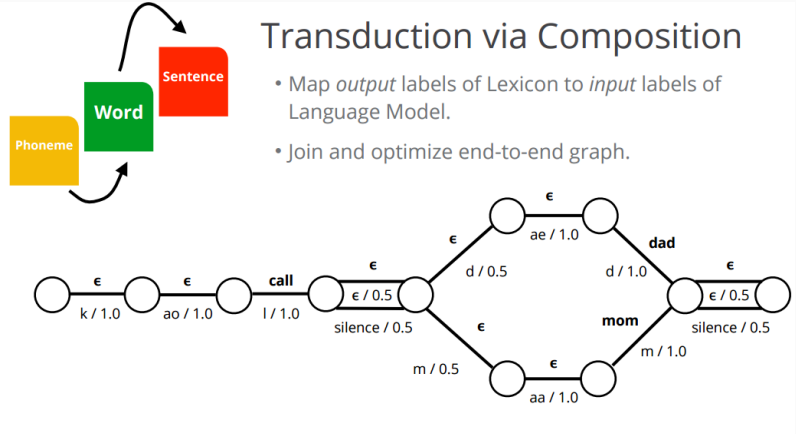
重新审视……

- **音素－词**：词表 (lexicon)，**加权有限状态转换器** （WFST）



重新审视……

- **音素－句子**：将词表 FST 与文法 FST 合成：$L \circ G$



联结主义时序分类

Connectionist Temporal Classification (CTC), [Graves et al. 2006]

观点：声学模型真正应该关心的是输出的词或音素序列，而不是传统交叉熵（CE）训练中优化的逐帧标注

- 为解决标签数量少于输入语音帧数量的问题，引入了空符号，并允许标签重复，从而迫使输出和输入序列的长度相同
- DeepSpeech 和 EESEN 作为端到端的语音识别系统直接预测字符而非音素，不再需要使用词典和决策树
  注：只有以词 (*word*) 为输出单元的 *CTC* 模型才是真正的端到端

序列区分性训练

Sequence-discriminative training

记 X 为训练数据中的语音信号，W 为训练数据的文本，θ 为声学模型参数：

最大似然训练

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} P_{\theta}(X \mid W)$$

序列区分性训练 (MMI 准则为例)

$$\hat{\theta}_{\mathrm{DT}} = \arg\max_{\theta} P_{\theta}(W \mid X)$$

序列区分性训练 (cont'd)

定义

$$\hat{\theta}_{\mathsf{ML}} = \arg\max_{\theta} P_\theta(X \mid W)$$

$$\hat{\theta}_{\mathsf{DT}} = \arg\max_{\theta} P_\theta(W \mid X)$$

用一次 Bayes 公式：

$$\hat{\theta}_{\mathsf{DT}} = \arg\max_{\theta} \frac{P_\theta(X \mid W)P(W)}{P_\theta(X)} = \arg\max_{\theta} \frac{P_\theta(X \mid W)P(W)}{\sum_w P_\theta(X \mid w)P(w)}$$

分子上的 $P_\theta(X|W)$ 是最大似然的目标函数，而分母则是所有文本产生训练语音的概率（按语言模型）的加权和。

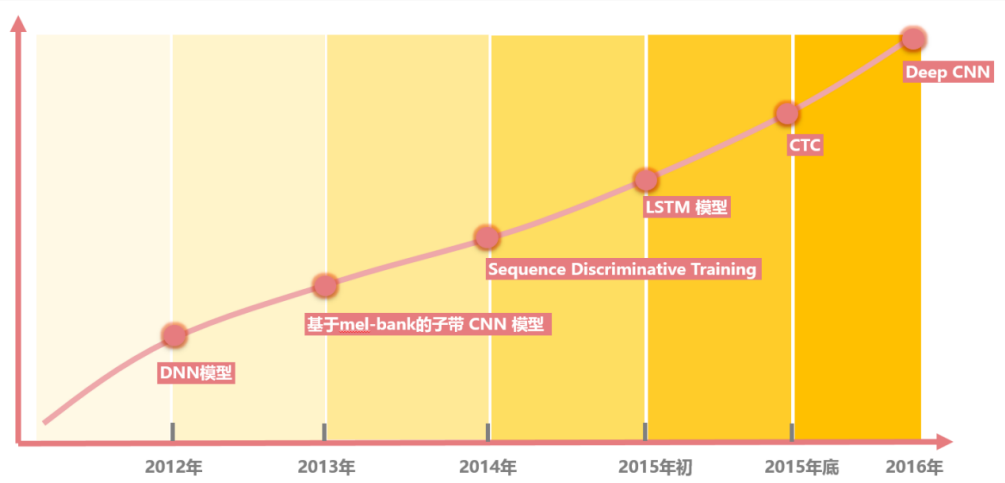
序列到序列模型

Sequence-to-sequence (Seq2Seq)

- 基本的 Seq2Seq（简单的 LSTM 编解码等）并不适合语音识别
  - 语音信号太长，记不下来
  - 不同于机器翻译，语音有单调性
- **注意力机制** + Seq2Seq
  - **Listen, Attend and Spell (LAS)**
- 不断输出**字符**，直到 EOS
- 语言模型只能基于训练集来做，很难加入外部的语言模型；不过可以在做 beam search 的时候借助语言模型做重评分 (rescoring)

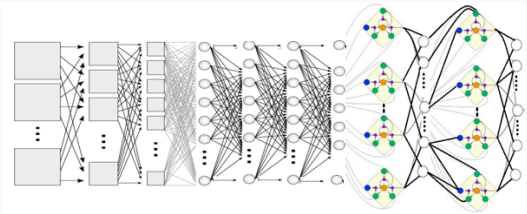
工业界的语音识别

百度的迭代史


DNN模型
基于mel-bank的子带 CNN 模型
Sequence Discriminative Training
LSTM 模型
CTC
Deep CNN
2012年
2013年
2014年
2015年初
2015年庙
2016年

CLDNN (CNN+LSTM+DNN) / CDL

约 2015 年初，语焉不详 [Jia 2015], [Liu 2015]



- **卷积层**：描述说话人频谱偏移带来的变化
- **七层全连接层**：提取抽象高层信息（害怕!）
- **LSTM 层**：描述时间序列变化信息

结论：2000h，双层 LSTM 优于 CNN；10000h，相反（1024 个节点）

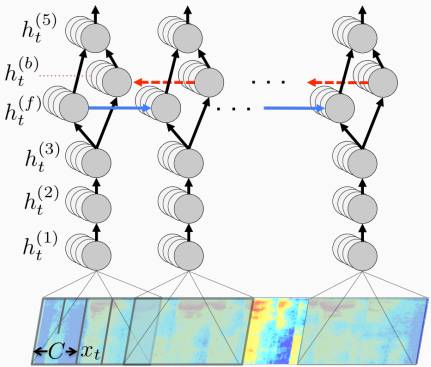
CDL (cont'd)

- 单向 LSTM 在做建模单元的整体建模上有诸多好处，但一直以来，因其存在解码路径右边信息的不完整性，导致识别效果较低，始终超不过传统的三状态建模。
- 在大数据、大模型条件下，在固定边界的 CE 训练之后，采用 CTC 训练，可能对多层双向 LSTM 模型的性能提升很有限，但对多层单向 LSTM 模型的改善是显著的。
- CTC 的空白吸收机制和动态边界尖峰学习能力，可以动态自适应地形成 'target delay'，从而解决单向 LSTM 模型的右边信息不完整的问题，而这个作用对双向 LSTM 模型的价值就小很多。

Deep Speech: Scaling up end-to-end speech recognition

[Hannun et al. 2014]

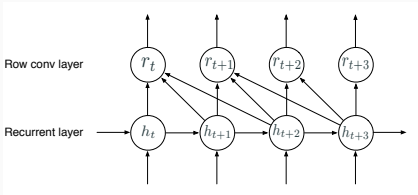
- 前三层为全连接层，第一层的输入为语谱图
- 第四层为双向 RNN
- 第五层将前向 RNN 和反向 RNN 求和作为隐单元的输入
- softmax 给出每个时间段内，将语音识别为各字母的概率



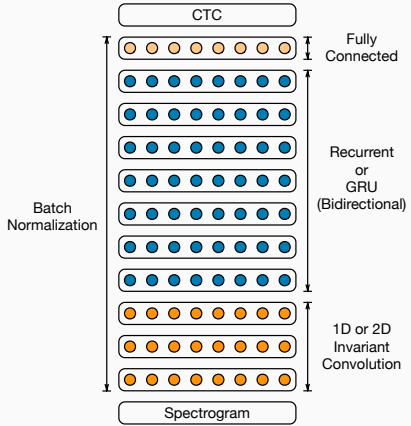
Deep Speech 2: End-to-End SR in English and Mandarin

[Amodei et al. 2015]

- 底层的 2D （时频）不变卷积对含噪声数据有很大帮助
- **行卷积** (row convolution)：向前单向看 τ 步，控制时延

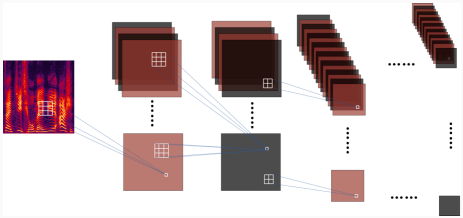


$$r_{t,i} = \sum_{j=1}^{\tau+1} W_{i,j} h_{t+j-1,i}, \quad \text{其中 } 1 \le i \le d.$$



Deep CNN

约 2016 年末，语焉不详

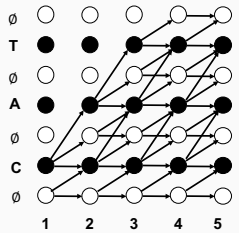


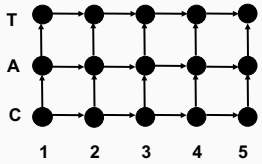
- 基于 Deep Speech 2，采用 VGGNet 中的 3×3 小 kernel 和残差连接，卷积加到 10 层
- CNN 的平移不变性可以帮助克服语音信号本身的多样性（说话人自身、以及说话人间、环境等），但是“单用深层 CNN 端到端建模性能较差”（讯飞：我不服）

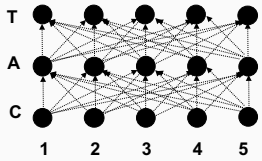
Exploring Neural Transducers for End-to-end Speech Recognition

[Battenberg et al. 2017]

	CTC	RNN-转换器	Seq2Seq (full attention)
给定音频时，不同时间步 所作预测之间的条件独立性	是	否	否
输入和输出单元间的对齐是单调的	是	是	否
硬对齐 vs. 软对齐	硬	硬	软



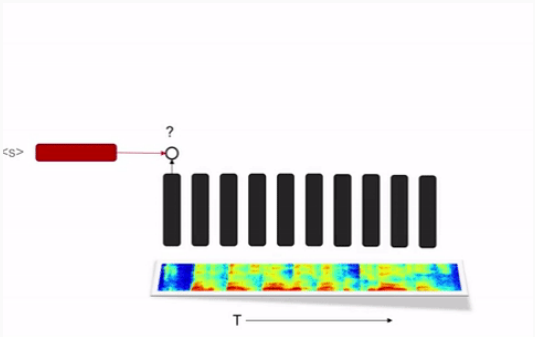




Exploring Neural Transducers for End-to-end Speech Recognition

[Battenberg et al. 2017]

RNN 转换器工作示意如下：



附录：Cold Fusion

[Sriram et al. 2017]

将语言模型带入 Seq2Seq 的训练过程

没看完，不讲了

Deep Peak 2

2018 年初，语焉不详

- “基于 LSTM 和 CTC 的上下文无关音素组合建模”
- 上下文相关 (CD) 的三音子建模单元有一万个音素组合
- **上下文无关** (CI) 的建模上，组合缩小到一千个，不再需要决策树聚类
- 声学和语言文本分开训练
- 口语和朗读语音组合训练，中英文混合、多风格混合输入

Achieving Human Parity in Conversational Speech Recognition

[Xiong et al. 2016]

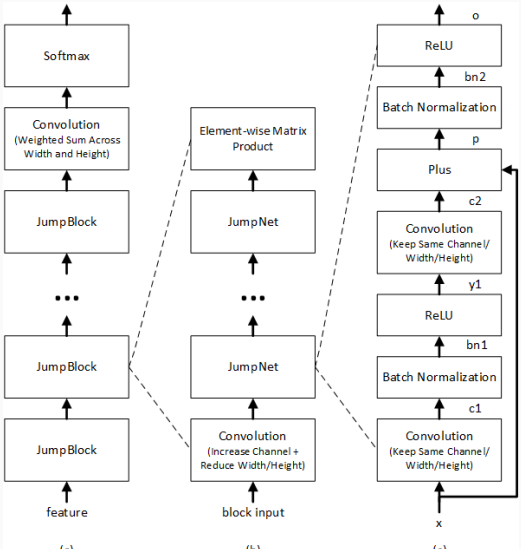
- 换 CNN，疯狂 ensemble；用 i-vector 实现说话人自适应 (speaker adaptation)
  - VGGNet / ResNet / LACE (TDNN-variant)
  - **LACE**: 逐层语境扩展和注意 (layer-wise context expansion and attention)

“语音模型我们基本上用了 6 个不同的神经网络，并行的同时识别（……）在此基础之上再用 4 个神经网络做语言模型，然后重新整合。所以基本上10 个神经网络在同时工作，这就造就了我们历史性的突破。”

——黄学东，微软首席语音科学家，ACM Fellow

- 《微软研究院新成果！对话语音识别水平超人类，错误率仅为 5.9%》

LACE 网络架构


Softmax
Convolution
(Weighted Sum Across Width and Height)
JumpBlock
JumpBlock
JumpBlock
feature
(a)
Element-wise Matrix Product
JumpNet
JumpNet
Convolution
(Increase Channel + Reduce Width/Height)
block input
(b)
o
ReLU
bn2
Batch Normalization
p
Plus
c2
Convolution
(Keep Same Channel/ Width/Height)
y1
ReLU
bn1
Batch Normalization
c1
Convolution
(Keep Same Channel/ Width/Height)
x
(c)

The Microsoft 2017 Conversational Speech Recognition System

[Xiong et al. 2017]

- 换 CNN，疯狂 ensemble
- “不行啊，好像咱们去年就是这么玩的?”
- “那你在 CNN 那加个 BLSTM 试下吧!”
- “不错不错，还能怎么改吗?”
- “你在 ensemble 那再放一个 rescoring!”
- 《微软语音识别技术里程碑：错误率降至 5.1%，超过专业速记员》

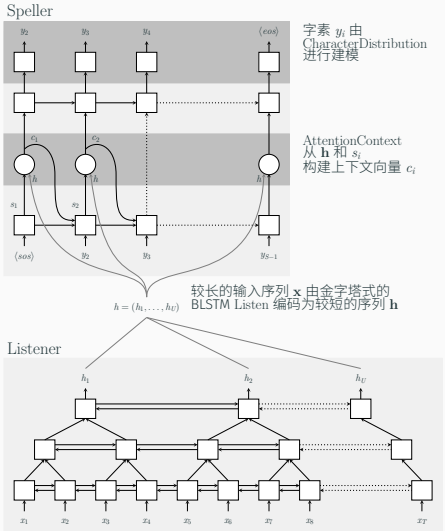
Listen, Attend and Spell

[Chan et al. 2015]

- Listener：声学模型**编码器**，Listen 将原始信号 $\mathbf{x}$ 编码为高层表示 $\mathbf{h} = (h_1, \ldots, h_U)$, $U \leq T$
- Speller：基于注意力的字符**解码器**，AttendAndSpell 接收 $\mathbf{h}$ 并产生一个字符列上的概率分布

$$\mathbf{h} = \text{Listen}(\mathbf{x})$$

$$P(\mathbf{y}|\mathbf{x}) = \text{AttendAndSpell}(\mathbf{h}, \mathbf{y})$$



State-of-the-art Speech Recognition With Seq2Seq Models

[Chiu et al. 2017]

为了让端到端发威，作了很多结构和训练方面的优化：

- 字片取代字素
- **多头注意力** (multi-head attention)
- 定时采样 (scheduled sampling)、标签平滑 (label smoothing)
- **目标函数**：最小化期望字错误率 (MWER)

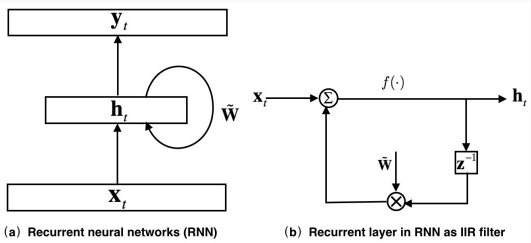
$$\mathcal{L}_{\text{MWER}} = \mathbb{E}_{P(\mathbf{y}|\mathbf{x})}[\text{WordErrors}(\mathbf{y}, \mathbf{y}^*)] + \lambda \mathcal{L}_{\text{CE}}$$

$$\mathcal{L}_{\text{MWER}}^{\text{Nbest}} = \frac{1}{N} \sum_{y_i \in \text{NBest}(\mathbf{x}, N)} [\text{WordErrors}(\mathbf{y}_i, \mathbf{y}^*) - \widehat{\text{WordErrors}}] \widehat{P}(\mathbf{y}_i|\mathbf{x}) + \lambda \mathcal{L}_{\text{CE}}$$

Feed-forward Sequential Memory Network

2015 年 12 月, [Zhang et al. 2015]

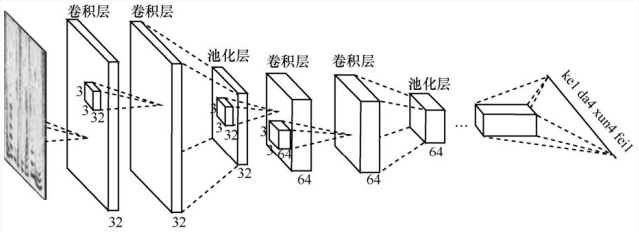
有限冲击响应 (FIR) 与无限冲击响应 (IIR)



RNN 相当于一个 IIR 滤波器，卷积次数足够的 CNN 相当于高阶 FIR 滤波器

Deep Fully Convolutional Neural Network

2016 年 8 月



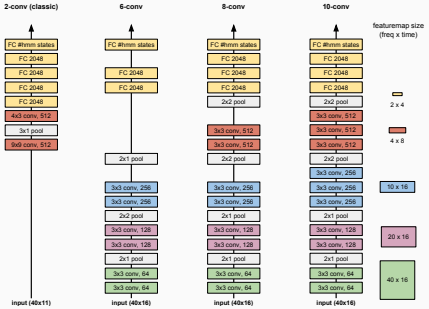
- 深度全序列卷积；直接对整句语音信号进行建模
- 使用 3×3 的小卷积核，**多个卷积层之后再加上池化层**
- 可以更好地表达语音的长时相关性，相比 RNN 网络结构更鲁棒；池化层等特殊结构可以使得 CTC 的端到端训练变得更加稳定

Deep Fully Convolutional Neural Network (cont'd)

ICASSP, Very Deep Multilingual ConvNets for LVCSR [Sercu et al. 2016]

这个想法是从哪来的?

- conv-conv-pool-conv-conv-pool...
- 显然，VGGNet
- 改进点在哪?
- **长时相关性**：输入整张语谱图，不再切帧留上下文!



LC-LFR-BLSTM

ICASSP, [Kun et al. 2017] / 支付宝 95188 热线

- 时延控制 (latency-controlled)
- 降帧率 (lower frame rate)
- 不细讲了，和我们关系不大

Deep FSMN

ICASSP, [Zhang et al. 2018]

- 层和层之间加了跳转连接 (skip connection)
- 实习生带着算法跑路了?

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多
- 大佬们的计算资源很多

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多
- 大佬们的计算资源很多
- 总而言之大佬们的钱很多

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多
- 大佬们的计算资源很多
- 总而言之大佬们的钱很多
- 如果不是公司的实习生就别做语音识别了，反正你一无所有

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多
- 大佬们的计算资源很多
- 总而言之大佬们的钱很多
- 如果不是公司的实习生就别做语音识别了，反正你一无所有

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多
- 大佬们的计算资源很多
- 总而言之大佬们的钱很多
- 如果不是公司的实习生就别做语音识别了，反正你一无所有
- “……数据、计算力和算法，这三者加起来才是最后系统的性能。三个要素中，如果缺少任何一个，系统的性能就会差很多。如果算法比别人好一点，但是数据比别人少很多，那么算法的有事很可能弥补不了数据的缺失，反之亦然。”

  ——俞栋，腾讯 *AI Lab* 副主任，西雅图实验室负责人

小 (po) 总 (leng) 结 (shui)

- 大佬们的数据很多
- 大佬们的计算资源很多
- 总而言之大佬们的钱很多
- 如果不是公司的实习生就别做语音识别了，反正你一无所有
- “……数据、计算力和算法，这三者加起来才是最后系统的性能。三个要素中，如果缺少任何一个，系统的性能就会差很多。如果算法比别人好一点，但是数据比别人少很多，那么算法的有事很可能弥补不了数据的缺失，反之亦然。”

  ——俞栋，腾讯 *AI Lab* 副主任，西雅图实验室负责人
- 社会，社会，告辞

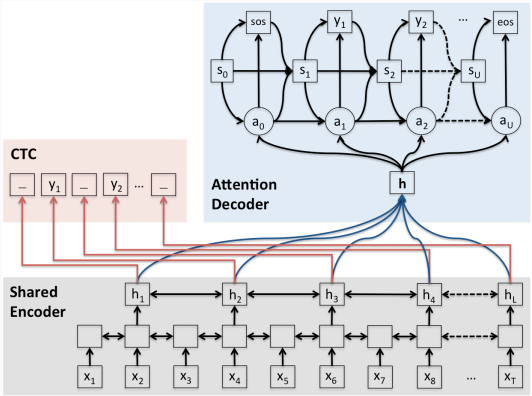
CTC + Attention

[Kim et al. 2017]

多任务学习

Attention 有不能单调地从左到右对齐和收敛缓慢的缺点。

通过将 CTC 目标函数用作辅助成本函数，注意训练和 CTC 训练以一种多任务学习的方式结合到了一起。这种训练策略极大地改善了基于注意的模型的收敛，并且缓解了对齐问题。



$$\mathcal{L}_{\text{MTL}} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{Attention}}, 0 \leq \lambda \leq 1$$

语音活动检测

视听觉融合

人说话（通常）不止出声，嘴也会跟着动（废话，用你说）

基于视觉的语音活动检测

2017 年 9 月～

- 支持向量机 + (傅里叶变换 + 人脸特征点检测)
- **改进**：CNN 提取图像特征，用词袋模型聚类后与前面的特征拼接
- 有什么问题吗?

不鲁棒！

- 特征点检测器对大角度人脸效果不好
- 张嘴未必出声，嘴基本不动也未必没出声

可以用滤波平滑预测结果，但是要考虑时间步长

- 未考虑帧移问题，样本少；不能精准定位端点

V-VAD as Action Recognition

动作识别 (action recognition)

根据视频序列判别画面中人物的行为

- 双流网络 (two-stream)
- 3D 时空卷积 (Spatiotemporal 3D-CNN)
- 光流 (optical flow) 等局部特征

Visual Voice Activity Detection in the Wild

[Patrona et al. 2015]

把 V-VAD 任务作为动作识别问题来考虑

- Dense Trajectory
- STIP
- 词袋 (Bag of Words, BoW)

音视频说话人区分

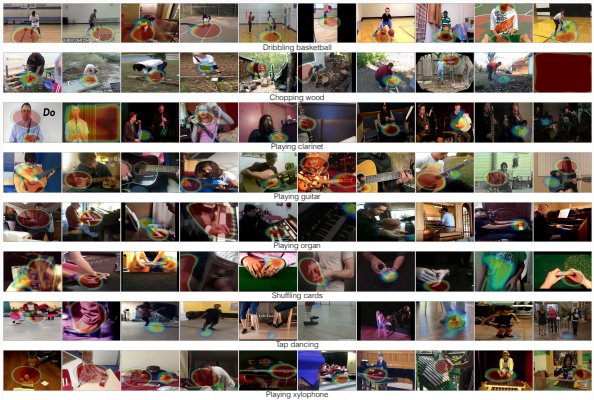
TristouNet: Triplet Loss for Speaker Turn Embedding

[Bredin et al. 2016]

以及：“Improving Speaker Turn Embedding by Crossmodal Transfer Learning from Face Embedding”

A-V Scene Analysis with Self-Supervised Multisensory Features

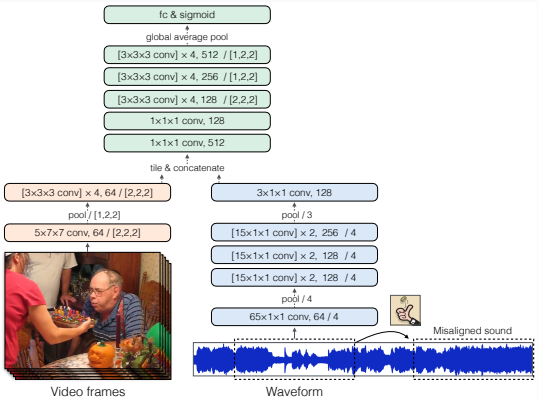
[Owens and Efros 2018]



可用于：音源定位、多模态动作识别、声音分离……

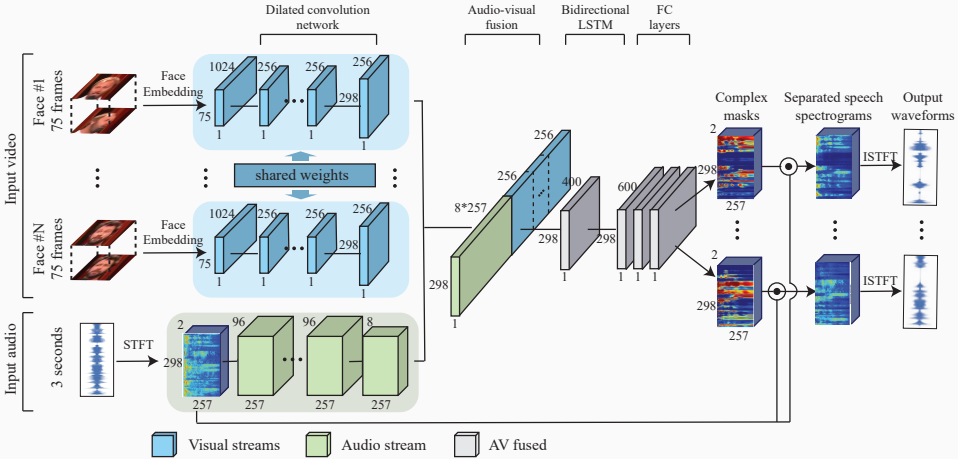
A-V Scene Analysis with Self-Supervised Multisensory Features

自监督特征的学习过程：送入对齐和不对齐的片段



Looking to Listen at the Cocktail Party

[Ephrat et al. 2018] (Google)

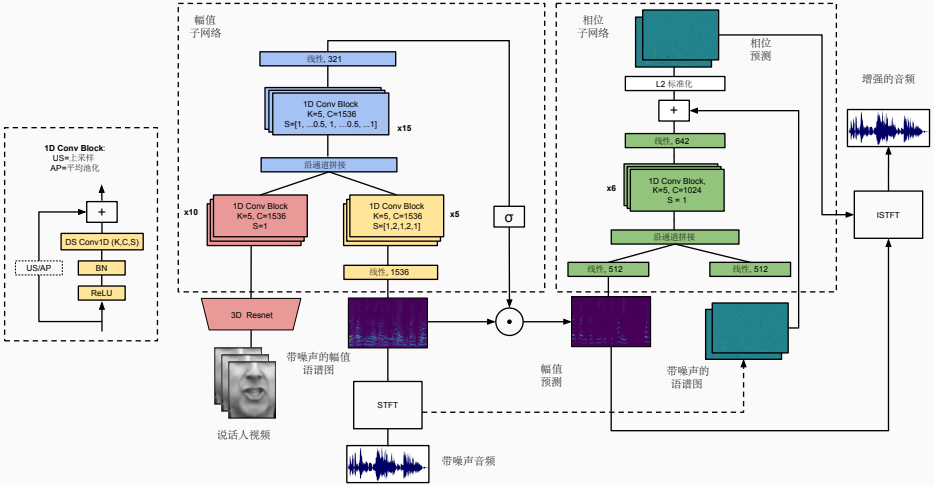


Looking to Listen at the Cocktail Party

- 视频侧的 CNN 卷积是在时间维度上做的
- 利用了**空洞**（扩张）卷积
- 用来做逆 Fourier 变换的幅值谱来自 STFT 得到的带噪幅值谱

The Conversation: Deep Audio-Visual Speech Enhancement

[Afouras et al. 2018] (VGG)



The Conversation: Deep Audio-Visual Speech Enhancement

改进点：不再只预测幅值谱，而是也预测相位谱

- 低噪声，高 SNR（信噪比）的情况下，OK；低 SNR
- 全卷积网络，计算用滑动窗口进行
- 语谱图不作为一张图像，而是视为有频率段数量个通道的时序信号列

二作 Samuel Albanie 的 CV 分享

Poetry (诗作)

The following piece was inspired by real life experiences and was submitted to a poetry competition in 2016:

Computer you piece of #$%&! 神 TM 垃圾电脑,
Now is not the time. 这个节骨眼犯病!
The deadline approaches. ddl 越来越近……

Source: `http://www.robots.ox.ac.uk/~albanie/curriculum-vitals.html`

说话人识别

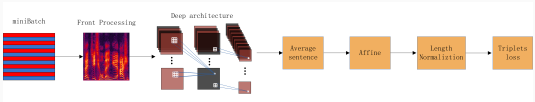
说话人识别中的术语

- **全局背景模型** (UBM, Universal Background Model)：用来训练与说话人无关的特征分布
- **i-vector**：在由 GMM-UBM 模型得到的统计量基础上，将每条不定时长的语音用一个固定维度的低维向量来表示，同时尽可能多的保留了其中的说话人相关信息

Deep Speaker

[Li et al. 2017]

- 基于语谱图构建的说话人嵌入 (embedding)
- 残差连接，小 kernel



VoxCeleb: Deep Speaker Recognition

[Chung et al. 2017]

来自 YouTube 上的视频，SyncNet 检出说话人

比较了用 CNN 和用 i-vector 识别的效果（有意思吗……主要就是发数据集吧）

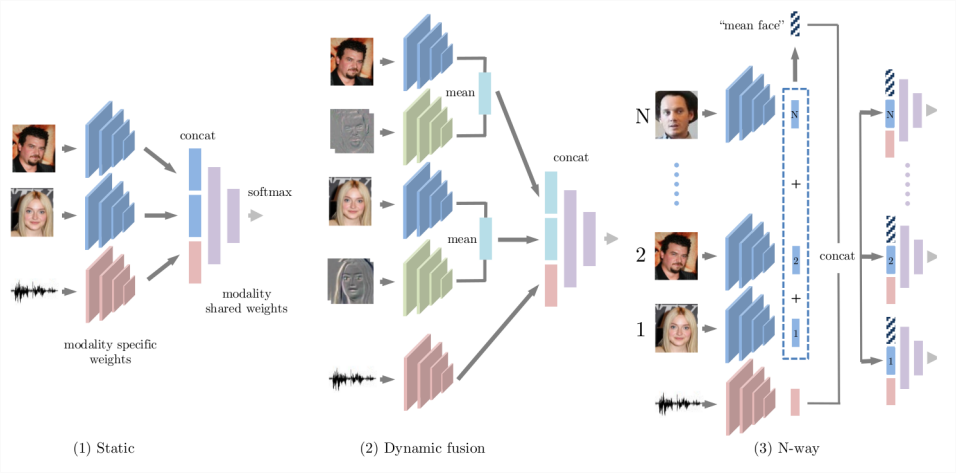
Seeing Voices and Hearing Faces: Cross-modal biometric matching

[Nagrani et al. 2018]

- **找不同网络** (Odd-one-out Network)：多个分支间共享权重
- 区别：不同模态分支间不共享权重
- 特征融合之后模态间共享全连接层
- **动态图像网络** (Dynamic Image Network)：单帧图像综合整个序列信息

- 查询池化 (query pooling)：解决多个输入分支间不能互相感知的问题

Seeing Voices and Hearing Faces: Cross-modal biometric matching



谜之致谢

论 Joon 是如何用两篇论文撑起了所有的后续研究

Acknowledgements: The authors gratefully acknowledge the support of EPSRC CDT AIMS EP/L015897/1 and the Programme Grant Seebibyte EP/M013774/1. The authors would also like to thank Erika Lu for help with the AMT study, Hakan Bilen and Joe Levy for useful discussions, and Joon Son Chung for being a living legend.

“感谢传奇人物 Joon”

音视频同步检测

问题提出

人有明显不同步感的音视频偏移范围：-125ms $\sim$ $+45$ms

应用：判断视频中的人是否是音频中的说话人（过滤配音和群像镜头等）；判断画面里哪一个是当前说话人（active speaker detection, ASD）

注：显然后者可以用于部分解决前面的说话人区分问题

Detecting Audio-Visual Synchrony Using Deep Neural Networks

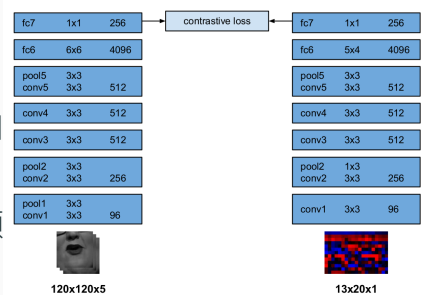
[Marcheret et al. 2015]

- **网络输入**：拼接后的音视频**瓶颈特征** (bottleneck features)，510ms 的上下文窗口
- **输出**：一个类别标签，1 个“同步”类（0ms），12 个“不同步”类（$\pm$180ms）

SyncNet

[Chung and Zisserman 2016]

- 双流 (two-stream) 网络，损失函数：**对比损失** (contrastive loss)
- **网络输入**：5 帧唇部图像（120×120）和对应的音频帧 MFCC 热图（13×20）
- 计算时，为（平移后的）音频特征和视频特征寻找最小 Euclid 距离
- **学到的 CNN 特征提升了唇读的准确率**；回顾 [Owens et al. 2018]



Multi-view SyncNet

[Chung and Zisserman 2017]

- 调整了一下 RoI 大小，其余变化不大
- 训练策略：**递进学习** (curriculum learning)

对比损失；孪生网络 (siamese network)

$$L = \frac{1}{2N}\sum_{n=1}^{N} yd^2 + (1-y)\max(\text{margin} - d, 0)^2$$

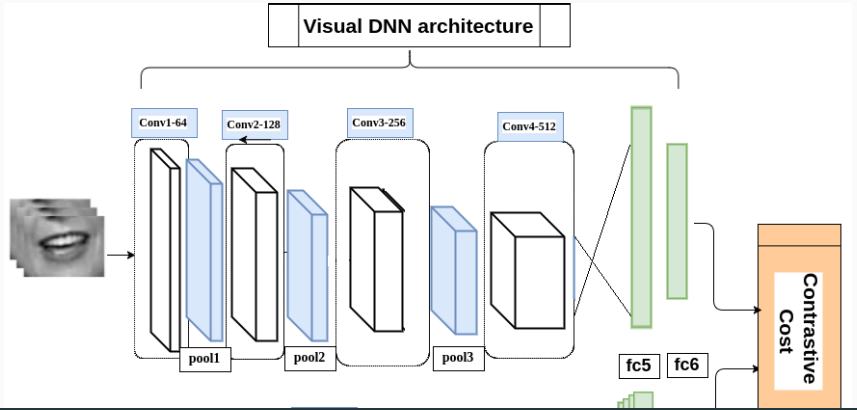
$$d = ||a_n - b_n||_2$$

其中 $y \in \{0, 1\}$ 表示两个样本是否匹配，margin 为预设的阈值。

3D CNN for Cross Audio-Visual Matching Recognition

[Torfi et al. 2017]

网络类似，特征直接拼接，只不过两侧的 CNN 都换成了 3D-CNN



结语

苟利国家生死以，岂因祸福避趋之

“那么人呐就都不知道，自己就不可以预料。一个人的命运啊，当然要靠自我奋斗，但是也要考虑到历史的行程。我绝对不知道，我一个视觉组的实习生怎么调研起语音识别来了?”

Thank you!

Questions?

张远航 | me@caszhang.cn