计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(3):14−23 [doi: 10.15888/j.cnki.csa.007852]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

语音识别及端到端技术现状及展望①

鱼　昆, 张绍阳, 侯佳正, 张少博

(长安大学 信息工程学院, 西安 710064)
通讯作者: 鱼　昆, E-mail: 624501922@qq.com

摘　要: 通过对语音识别技术的发展梳理, 简单介绍了语音识别的历史和应用现状, 并将传统语音识别的技术和当前的研究进展进行描述. 传统语音识别采用基于统计的方法, 采用声谱特征, 在 GMM-HMM 混合结构上进行训练和匹配. 当前的语音识别模型主要基于深度学习的方法, 采用 CNN、RNN 都可以有效的进行特征提取从而建立声学模型. 进一步的研究采用了端到端的技术, 避免了多个模型间的误差传导. 端到端技术主要有 CTC 技术和 attention 技术, 最新的模型和方法着重研究了 attention 技术, 并在尝试进行与 CTC 的融合以达到更好的效果. 最后结合作者自身的理解, 概括了语音识别当前所面临问题和未来发展方向.

关键词: 语音识别; 隐马尔可夫模型; 深度学习; 端到端; 注意力机制

引用格式: 鱼昆,张绍阳,侯佳正,张少博.语音识别及端到端技术现状及展望.计算机系统应用,2021,30(3):14–23. http://www.c-s-a.org.cn/1003-3254/7852.html

Survey of Speech Recognition and End-to-End Techniques

YU Kun, ZHANG Shao-Yang, HOU Jia-Zheng, ZHANG Shao-Bo

(School of Information Technology Engineering, Chang'an University, Xi'an 710064, China)

Abstract: The paper briefly introduces the history and application of speech recognition, traditional speech recognition techniques, and current research progress. Traditional speech recognition relies on statistics-based methods and sound spectrum features to train Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) hybrid model. Nowadays, speech recognition models are mainly based on deep learning. Generally, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) can effectively extract features to establish acoustic models. Further research depends on end-to-end techniques to avoid error transmission among models, and these techniques mainly include Connectionist Temporal Classification (CTC) and attention. The latest models and methods highlight attention, which are trying to integrate it with CTC to achieve better results. Finally, combined with the authors' understanding, the paper summarizes the existing problems and future development in speech recognition.

Key words: speech recognition; HMM; deep learning; end-to-end; attention

语音是采用一定语言规则通过人体发声器官发出的有规律的声音信号. 语音识别 (Auto Speech Recognition, ASR) 是研究如何将语音信息转化成文本信息. 语音的研究领域可以细分为语音识别、语音合成、声

① 基金项目: 陕西省技术创新引导专项 (S2018-YD-CGRGG-0030); 中央高校基本科研业务费高新技术研究培育项目 (300102240202); 陕西省自然科学基础研究计划面上项目 (2014JM2-5074)
Foundation item: Technology Innovation Guiding Project of Shaanxi Province (S2018-YD-CGRGG-0030); High-tech Research and Development Projects of the Fundamental Research Funds for the Central Universities (300102240202); General Program of Fundamental Research Program of Natural Science of Shaanxi Province (2014JM2-5074)
收稿时间: 2020-07-27; 修改时间: 2020-08-25; 采用时间: 2020-09-08; csa 在线出版时间: 2021-03-03

纹识别. 其涉及到信号处理, 自然语言处理等. 在发展过程中经历了 3 个阶段. 一是孤立词识别, 二是连接词识别, 如连续数字或连续单词, 三是大词汇量下连续语音识别.

自上世纪 50 年代开始, 着手于最简单的数字识别任务, 语音识别领域进入研究者的视野. 到 80 年代, 研究者们采用统计分析的方法使连续语音识别成为可能. 在我国, 50 年代末有研究者采用电子管电路, 对英语中的元音进行尝试识别. 90 年代, 清华大学和中科院自动化所等单位在汉语听写机原理样机的研制方面取得有效成果[1]. 进入 21 世纪, 深度学习的发展极大促进了语音识别技术. 2017 年, 微软宣布了其在 Switchboard 词错率 (Word Error Rate, WER) 降至 5.1%[2], 这意味一定条件下机器已经可以像人类专业速记员一样识别词语了. 2018 年阿里巴巴语音识别模型 DFSMN 采用开源框架 Kaldi 进行构建, 在 Fisher (FSH) 数据集上测试词错率仅为 9.4%[3]. 百度的模型在其自建的中文数据集上训练并测试, WER 低至 7.93%, 取得良好的效果[4]. 但是在复杂多变的应用场景中, 识别准确率会大大下降. 因此, 语音识别领域还有许多问题需要继续研究和解决.

1 语音识别技术研究

1.1 语音信号特征提取技术

早在 1952 年, 首先研究了特定说话人孤立数字, 是由贝尔实验室的 Davis 等进行的[5]. 1956 年, RCA 实验室的奥尔森通过带通滤波器, 实现了一些单音节词的识别[1]. 1959 年, Fry 和 Denes 等通过频谱分析, 对语音的特征进行提取, 然后采用模式匹配的方法, 识别元音和辅音[1].

一般认为, 人们在 10–30 ms 的时段内, 语音是稳定的, 因此它是一个短时的时不变信号. 一般的特征提取方法有: 线性预测编码参数 (LPCC), 感知线性预测系统 (PLP), 梅尔频率倒谱系数 (MFCC) 等.

1980 年, Davis 等在前人研究的基础上, 做了大量生理心理学实验, 得到了一组经验公式[6], 频率转换公式为:

$$f_{\text{Mel}}=2595\times\lg\left(1+\frac{f}{700}\right)$$

对每帧信号进行变换, 采用信号处理中的短时傅里叶变换:

$$X_n(\omega)=\sum_{m=-\infty}^{\infty}x(m)w(n-m)\mathrm{e}^{-j\omega m}$$

其中, $w(n)$ 为窗函数.

1.2 动态时间归正 (DTW)

RCA 实验室的 Martin 等在 1960 年代末提出了时间归正的相关方法. 同时苏联的 Vintsyuk 也提出了采用动态规划方法来解决对齐问题[7]. 最终在 70 年代, 日本学者 Sakoe 给出了动态时间归正 (Dynamic Time Warping, DTW) 算法, 也称动态时间弯折、动态时间规整, 其将时间和距离计算结合起来, 采用动态规划的思想进行归正[8].

假设首先根据统计得来某个语音的模板, 其特征矢量序列为$X=\{x_1,x_2,x_3,\cdots,x_I\}$, 输入语音特征矢量序列为$Y=\{y_1,y_2,y_3,\cdots,y_J\}$, $I\neq J$. 递推公式为:

$$dtw[i][j]=\min(\min(dtw[i-1][j],dtw[i][j-1]),\\ dtw[i-1][j-1])+d[i][j]$$

其中, $d[i][j]=(y[j]-x[i])^2$表示 x_i 和 y_j 之间的欧式距离, $dtw[i][j]$ 表示 DTW 距离, 当算出 $dtw[I][J]$ 时递推结束.

1.3 矢量量化 VQ

70 年代末, Buzo 等[9] 提出了矢量量化 (VQ), 并将其成功应用. 首先采用统计方法, 将某个语音对应的多个信号划为一组, 用中心矢量作为代表值. 这样, 就将 d 维无限空间划分为 K 个区域边界, 每个区域称为一个包腔, 当待识别的输入信号的矢量给定时, 将其与这些包腔的边界进行比较, 当输入信号属于某个包腔时, 就被量化为此包腔的中心矢量值. 包腔的中心称为码字, 码字的组合称为码本. 一般采用 K-means 算法或 LBG 算法获得码字. 采用欧氏距离 (均方差距离) 度量. 这种技术主要用于孤立词的语音识别.

1.4 GMM-HMM

从 1980 年代开始, CMU 使用 VQ/HMM (Hidden Markov Model) 实现了一个语音识别系统 SPHINX. 可以实现 997 词的非特定人连续语音识别[10]. 隐马尔可夫模型是在 70 年代由 Baum 和 Baker 等建立和应用的[11].

HMM 具有无后效性的特征, 参数包含初始概率和概率转移矩阵, HMM 中的观察变量和状态通过一组概率分布相联系. 这个隐变量和观察值的对应的统计规律, 用高斯混合模型 (Gaussian Mixture Model, GMM)

表示. K 阶高斯混合模型是由 k 个高维联合高斯分布加权求和而得:

$$p(x)=\sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\Sigma_k)$$

其中, $\mathcal{N}(x|\mu_k,\Sigma_k)$称为混合模型中的第 k 个分量, π_k称为混合系数, 满足: $\sum_{k=1}^{K}\pi_k=1, 0\le\pi_k\le 1$.

对于语音来讲, 同一个音素可能在不同情态下的发音方式区别很大, 语音特征区别也就很大, 因此需要用多中心的分布来对应一个 HMM 中的状态, 因此两者结合起来, 就形成了 GMM-HMM 方法. 它由一组参数描述: N, 状态数目; π, 初始状态概率; A, 状态转移概率矩阵; B, 观察值概率分布.

GMM-HMM 的训练分为两步, 首先是求 GMM 的参数, 语音字典建立后, 语音和音素状态建立了多对一的对应关系. 将同一个音素状态的所有语音的特征进行分别提取, 用这些数据建立一个 GMM 来对这个音素状态进行拟合. 重复这一过程, 将所有的音素状态分别建模. 第二步是对 HMM 中的参数 π 和 A 进行估计(训练), 即给定一个观察值序列$O=o_1,o_2,\cdots,o_T$, 确定一个$\lambda=(\pi,A,B)$, 使$P(O|\lambda)$最大. 一般使用 Baum-Welch 算法进行, 类似于 EM 算法, 利用递归的思想, 使$P(O|\lambda)$取得最大值, 最后得到模型参数组$\lambda=(\pi,A,B)$. 至此, GMM-HMM 模型的训练完成.

使用 Viterbi 算法进行预测. 即给定观察值序列$O=o_1,o_2,\cdots,o_T$, 和模型$\lambda=(\pi,A,B)$, 确定一个最佳状态序列$Q'=q'_1,q'_2,\cdots,q'_T$. 定义$\delta_t(i)$为时刻 t 状态为 i 的所有单个路径$(i_1,i_2,\cdots,i_t)$中概率的最大值, $\varphi_t(i)$为时刻 t 状态为 i 的所有单个路径$(i_1,i_2,\cdots,i_{t-1},i)$中概率最大的第 t–1 个节点.

1.5 基于深度学习的语音识别

1980 年代, 人工神经网络 (ANN) 被引入到语音识别[2]. 但是由于计算能力的限制和人工神经网络的理论不完备, 采用人工神经网络方法的语音识别并没有更加蓬勃的发展起来.

1.5.1 引入 DNN 到语音识别

2006 年, Hinton 等提出了深度置信网络 (DBN)[12], 贪婪的逐层无监督学习算法是其核心. 通过先使用 DBN 来对多层感知机进行预训练, 然后通过反向传播算法来进行微调, 提供了一种解决深层网络优化过程中过拟合和梯度消失问题的有效途径. Deng 等促成了这一实践的成功[13]. 他们使用深度神经网络 DNN (Deep Neural Network) 代替传统的 GMM-HMM 系统中的 GMM, 以音素状态为建模单位, 提出了 DNN-HMM 的识别方法 (如图 1), 显著降低了误识率, 使其进入到真实用户可以接受的范围[14]. 和 GMM-HMM 相比, DNN 替换了 GMM, 语音信号的状态与观察值的对应采用深度神经网络来进行建模拟合.

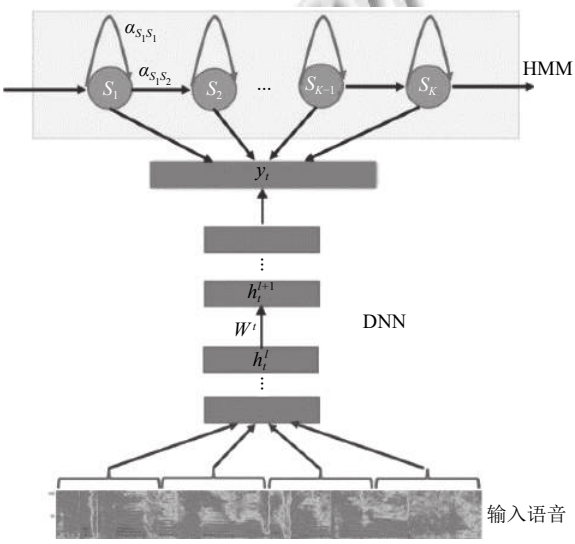

图 1 DNN-HMM 结构图

DNN 的输入可以是 MFCC 特征, 也可以是更底层的滤波器组 (Filter BanK, FBK) 声学特征. 输出矢量的维度对应到 HMM 的状态的个数.

1.5.2 CNN

使用 CNN 来进行语音识别, 主要是将卷积层和池化层堆叠起来以获取更高级别的特征, 这些层的顶部, 有一个标准的全连接层, 代表 HMM 状态, 它对网络中训练出来的特征进行整合. LeCun 等首先提出了沿时间轴进行卷积的语音数据 CNN[15]. 这样可以获得相对较小的时间偏移, 获得具有鲁棒性的特征.

Abedel-Hamid 等[16] 通过沿频率轴而不是时间轴应用卷积和最大池化, 实现了重大的提升. 发现沿频率轴的卷积会产生对小频移具有较高的鲁棒性, 这对于说话者或情绪变化具有较好的适应能力. 越来越多的研究人员在时间轴和频率轴上探索了卷积的方法.

这些探索和试验表明, 在 DNN-HMM 模型中, CNN 的性能优于完全连接的 DBN. 这是由于 DBN 以

任何顺序解释输入, 但实际上语音的特征与频率和时间上紧密相关, 权重共享使得 CNN 可以捕获这些局部相关性. 其次, 权重共享和合并有助于 CNN 捕获等变差异并获得更好的鲁棒性及稳定性. 而对于 DBN, 若要在较小的频率和时间偏移上捕获此类不变性, 需要大量的参数.

Sainath 等[17] 证明, 对于大型词汇任务, CNN 的性能比 DBN 更好. 这些实验进行了细致的优化手段, 包括超参数调整, 有限的权重分配和序列训练. Chan 等[18] 对低资源语言基于 CNN 的声学模型进行的研究得出, 在低资源语言条件下, CNN 能比 DBN 提供更好的鲁棒性和更好的泛化性能.

1.5.3 RNN

语音信号是一种时序信号, 模型如果能够对其时序动态信息进行有效表示, 将大大提升效果. DNN-HMM 的系统中声学模型是 DNN 和 HMM 的混合. 而 RNN (循环神经网络) 通过在隐层上增加反馈连接, 当前时刻的输入分为两部分, 一是当前时刻输入序列产生的输入, 这部分和普通的前馈神经网络是一样的, 传递的神经网络获取的特征表示, 二是由上一时刻保留的记忆信息, 产生的输入. 通过这种机制, RNN 可以利用到之前的信息.

研究人员对 HMM-RNN 混合模型进行了实验[19], 但结果与基于 DBN 的 CNN 声学模型无法相提并论. Graves 等[20] 提出 CTC (Connectionist Temporal Classification) 损失函数, 使神经网络能够学习字符序列和未分段语音之间的对齐关系, 从而避免了使用 HMM 来进行强制对齐, 实验中, 在 TIMIT 数据集上表现优于 HMM-RNN 混合方式. 文献 [21] 在 HMM-RNN 的基础上, 提出使用深度双向 LSTM 作为其声学模型, 并在 TIMIT 数据集中取得了较好结果. 文献 [22] 中对这种声学模型的进行了进一步的研究, 通过使用上下文相关的语音单元, 使用 LSTM 输出空间的上下文相关状态和采用分布式训练方式等方法, 取得了一些进展.

1.5.4 端到端技术

传统的语音识别模型通常包含声学模型 (Acoustic Model, AM)、发音词典 (Lexicon) 和语言模型 (Language Model, LM) 三部分组成. 每一部分都需要单独的学习训练, 端到端 (end-to-end) 的机制可以使得模型的训练摒弃发音词典和语言模型, 真正实现直接从语音转录成文本. 端到端主要有两种实现, 其中一种是上文提到的 CTC. 另一种是基于注意力机制 (attention) 的编码器-解码器 (encoder-decoder) 模型, 由 Chorowski 等于 2014 年首先应用到语音中的音素识别上面[23].

如图 2 所示的 CTC 方法最为常用, 是对 RNN 的一种改进. 一般来说, 输入特征序列与音素的对齐关系并不确定, 而且, 按照划分, 音素序列长度远远小于语音按照 10–30 ms 分帧后的序列长度, 然而, RNN 模型中的标注序列和输入序列必须是对应的. 这样的结果就是不管是基于 DNN-HMM 模型还是 RNN-HMM 模型都得首先采用 GMM-HMM 训练进行强制对齐. CTC 在标注符号集中加入了一个空白符号 (blank), 它意味着此帧没有预测值输出. 因而在模型的预测输出中就包含了很多空白符号, 一个音素对应的一整段语音中只有一个尖峰被识别器确认, 其他都被识别为空白, 结果相当于自动的切分了音素边界, 实现了将空白符号和连续出现的状态进行了消除, 就能得到最终预测的字符序列. Hannun 等[24] 采用了带有双向递归层的 5 层 RNN, 经过 CTC 损失训练以及语言模型来纠正, 在 Switchboard 数据集上获得了当时最好的结果. 同时他们还提出了一些优化方案. Amodei 等[4] 在这基础上, 使用有 13 个隐层 (包含卷积层) 的模型取得了更好的结果.

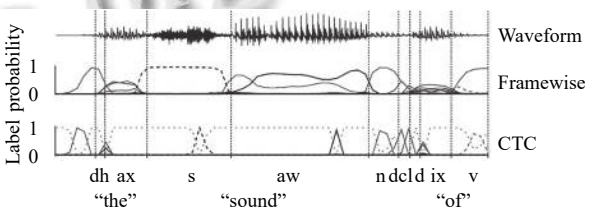

图 2 CTC 方法

Attention 机制最先应用于机器翻译中, 并在机器翻译中取得了较好的效果. 其主要思想就是通过编码器 (encoder) 将原序列转换成一个固定长度的隐层表示, 然后解码器 (decoder) 再根据这个隐层表示生成解码序列, 生成解码序列过程中考虑当前解码输出与隐层表示中哪一部分最相关, 这部分就是注意力机制, 其结构如图 3.

在这个模型结构中, 每一个条件概率的输出定义为:

$$p(y_i|y_1,\cdots,y_{i-1},X)=g(y_{i-1},s_i,c_i)$$

其中, y_i 表示第 i 时刻解码输出标记, X 表示编码器输入, y_{i-1} 表示上一时刻解码输出, s_i 表示 i 时刻的隐层状态, c_i 表示上下文向量. 其中 s_i 计算公式为:

$$s_i=f(s_{i-1},y_{i-1},c_i)$$

上下文向量 c_i 是编码器输出隐变量 h_i 的加权和.

$$c_i=\sum_{j=1}^{T_x}\alpha_{ij}h_{ij}$$

其中, α_{ij}即注意力权重. 其计算过程如下:

$$\alpha_{ij}=\frac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ij})}$$

其中, $e_{ij}=score(s_{i-1},h_j)$, 表示注意力机制的打分结果, 实际上相当于一个相关性计算, 具体的分数计算有多种方式, 其反映了上一时刻隐层状态 s_{i-1} 与向量表示 h_j 之间的相关性.

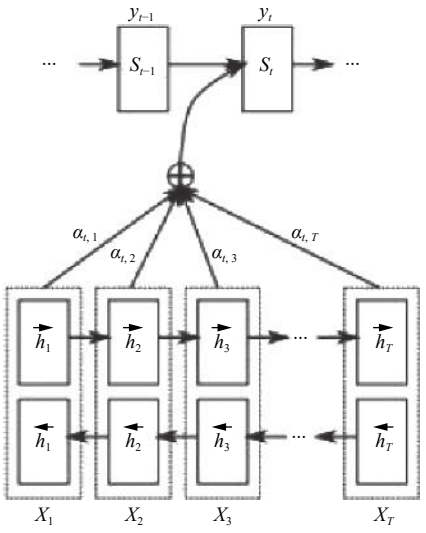

图 3 Attention 基础结构

Encoder-decoder 结构是针对序列到序列的问题提出的, 一般采用 RNN 及其变体 (LSTM 等), 目前常用的就是采用 Bi-LSTM 作为 encoder. 由于 decoder 是对上一时刻输出的依赖, 对于 decoder 的改进较难, 但对 encoder 的研究取得了一定的进展.

Facebook 在 2017 年提出 ConvS2S 结构, 把卷积结构首先引入机器翻译问题中, 并且一度获得最好的效果[25]. 由于没有时序结构, 因此需要在 embedding 的基础上面增加位置信息, 模型中将 position embedding (p_i) 与原来的 embedding(w_i) 直接进行相加, 因此模型的输入序列为 $\boldsymbol{e}=\{e_1,\cdots,e_m\}$, 其中 $e_i=w_i+p_i$. Decoder 在每一时刻的输入 $g=\{g_1,\cdots,g_m\}$, 同样由两部分组成, 分别是上一时刻输出的 word embedding 以及对应的 position embedding. Decoder 中第 l 个 block 的输出定义为$h_l=\left(h_1^l,\cdots,h_n^l\right)$, encoder 中第 l 个 block 的输出定义为$z_l=\left(z_1^l,\cdots,z_n^l\right)$. 通过堆叠卷积结构, 能够扩大感受野的面积, 越高层结构获得的上下文信息越多. 模型中使用了残差连接和 GLU (Gated Linear Units). 不同于一般的 encoder-decoder 结构, 这里的注意力机制使用了 multi-step attention, 在 decoder 的每层中都计算注意力机制分数.

ConvS2S 将 CNN 引入到 Seq2Seq 中, 这样既可以处理序列变长的问题, 又可以实现在序列不同位置的并行计算. RNN 的另一个缺陷在于, 对于一个长度为 n 的序列, 要建立长时相关, 需要经过 O(n) 次运算, 而对于卷积核宽度为 k 的多层 CNN 来说, 则需要 O(n=k) 次运算.

受限于 LSTM 的计算速度问题, 常见的 Seq2Seq 结构都采用的是浅层结构, Zhang 等[26]受 Very Deep CNN 在 ASR 任务中的优秀表现启发, 提出使用更深层的网络来进行序列编码, 代替浅层 encoder. 使用了 Network-in-Network(NiN), Batch Normalization (BN), Residual Networks(Res-Nets) 和 Convolutional LSTM(ConvLSTM) 等方法构建模型. 借鉴 NiN 中的 1×1卷积, 来增加网络的深度和模型的表征能力. BN 和 ResNets 方法有助于训练更深层的结构. ConvLSTM 中使用卷积操作代替 LSTM 内部的内积操作. 实验在 WSJ 数据集上进行, 结果显示模型获得了 WER 为 10.53%.

Chan 等[27] 提出新的 ASR 结构, 即 LAS (Listen, Attend and Spell). 主要包含两部分, Listener 是金字塔型的循环网络编码器, 接受滤波后的频谱作为输入. Speller 是基于注意力机制的循环网络解码器, 以之前的字符和声音序列为条件预测字符. 提高了编码的速度, 每层都会将时间步减少一半.

$$\begin{cases}h_i^j=pBLSTM\left(h_{i-1}^j,\left[h_{2i}^{j-1},h_{2i+1}^{j-1}\right]\right)\\ c_i=AttentionContext(s_i,h)\\ s_i=RNN(s_{i-1},y_{i-1},c_{i-1})\\ P(y_i|x,y<i)=CharacterDistribution(s_i,c_i)\end{cases}$$

这里的金字塔结构采用每层合并上一层相邻的

2 或 3 个时间步, 其中 character distribution 是前馈网络结构.

对于 attention 机制的探索也是一个重要的研究方向. Attention 的核心思想就是计算当前要解码序列需要的输入信息与上下文信息之间的相关性. 相关性的计算就是: $e_{ij}=score(s_{i-1},h_j)$.

文献 [28] 中直接进行计算, 这种方式没有考虑两个向量位于不同的特征空间, 直接计算打分结果. 常见的有:

Dot:

$$score\left(h_t,\bar{h}_s\right)=h_t^{\mathrm{T}}\bar{h}_s$$

General:

$$score\left(h_t,\bar{h}_s\right)=h_t^{\mathrm{T}}W_a\bar{h}_s$$

Concat:

$$score\left(h_t,\bar{h}_s\right)=v_a^{\mathrm{T}}\tanh\left(W_a\left[h_t^{\mathrm{T}};\bar{h}_s\right]\right)$$

文献 [29] 指出打分过程应参考上一时刻的注意力权重, 那么打分过程成为: $score(s_{u-1},\alpha_{u-1},h_t)$.

Hard-attention 是文献 [30] 在 image caption generation 任务中提出的. 常见的注意力机制是经过 Softmax 层输出之后有不同的权重, 是一个向量, 里面元素都是范围在 [0,1] 之间的小数, 和为 1. 而采用 hard-attention 之后, 注意力向量中的元素只有一个是 1, 其余的都是 0, 也就是在每一个时间步, 模型只关注一个位置. 向量是 One-hot 形式.

而 soft-attention 更常见一些, 即注意力向量中的不同位置的权重值不同, 这样的 soft-attention 是光滑的且是可微的. 文献 [30] 中还对注意力机制进行了微调.

$$\widehat{z}_t=\beta\sum_i\alpha_{t,i}h_i$$

其中, $\beta=\sigma\left(f_\beta(s_{u-1})\right)$, 用来调节上下文向量在 LSTM 中的比重.

文献 [29] 中考虑为了使得 h_u 中的元素更加具有区分性, 可以考虑把权重换成矢量$\alpha_{u,t}$. 文献 [28] 中提到 global-attention 和 local-attention, global 就是在 decoder 计算注意力权重的每一时刻都考虑全部的上下文信息, 赋予不同位置的上下文信息不同的权重, 并加权求和. 一方面这样有很大的计算量, 另一方面在语音识别中, 两种序列时序一致, 注意力只需要集中在时序对应的位置, 因此采用 local-attention 有助于实现 sequence-to-sequence. Local 方式就是上下文向量的计算每次都只关注到几个源隐藏状态. 它是可微的, 因此更加容易训练, 分为 local-m (local monotonic alignment) 和 local-p (local predictive alignment) 两种计算方式. *pt* 是关注的焦点, 距离中心 *pt* 越远, 其位置上的源隐藏状态对应的权重则被压缩的越厉害.

文献 [31] 最先提出了 Multi-Head Attention (MHA). MHA 在传统注意力机制的基础上扩展了多个 head, 每个 head 能够生成不同的注意力分布. 这个允许每个 head 在对应编码器输出的时候, 可以扮演不同的角色. 这种方式能够帮助解码器更容易的从编码输出中检索出所需要的信息. 传统的 single-head attention 更加依赖于编码器提供清晰的隐层表示以使得 Decoder 能够挑选出重要的信息. MHA 趋向于分配一个 head 去关注语句的开头, 语句的开头往往包含大部分的背景噪声. 为了确保 MHA 在训练过程中确定能够关注到不同的位置, 一些研究者在损失函数中增加正则项, 以确保多个 MHA 之间存在差异.

纯 attention 方法虽然取得了不错的效果, 但是在训练过程中存在着明显的收敛速度慢, 震荡幅度大等问题. 这很大程度上在于一开始 attention 注意范围太广, 难以收敛. 文献 [29] 提出使用 CTC 辅助 attention 模型的训练, 实验表明这种方法能够极大的提高模型的收敛速度. 模型成功的关键在于在损失函数中引入 CTC Loss:

$$Loss=\lambda L_{\mathrm{CTC}}+(1-\lambda)L_{\mathrm{Attention}}$$

在 CTC 辅助训练的情况下, 原本需要 9 个 epoch 才能收敛的模型在 5 个 epoch 的时候已经收敛了. 在解码阶段, 如果对应于 attention 的 decoder 中非 OOV (Out Of Vocabulary) 的词汇, 则使用对应的输出. 如果最大概率的输出是 OOV 标记, 则使用 CTC 中的结果进行代替. 为了实现混合解码, CTC 部分除了增加 blank, 还应该增加一个词边界标记 wb.

Transformer 是最初在机器翻译领域中获得了成功. 其解决的问题主要是提高 encoder 的并行度. 其中关键的点就是 self-attention 和 MHA 两种机制. Self-attention 是每个词都要和所有的词计算 attention, 可以捕获长距离的依赖关系. MHA 中不同 head 学习不同

的子空间语义, 关注编码器输出的不同部分.

Self-attention:

$$Attention(Q,K,V)=softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$

Multi-head attention:

$$\begin{cases}MultiHead(Q,K,V)=Concat(head_1,\cdots,head_h)W^o\\ head_i=Attention\left(QW_i^Q,K_i^Q,V_i^Q\right)\end{cases}$$

其中, Q 表示 query, K 表示 key, V 表示 value, 在 self-attention 时候, Q=K=V=*inputs*.

同时, 经过对 attention 的一系列探索, 一些优化手段被总结出来: 采用大的建模单元, 如子词或词等, 这样的建模单元更加稳定并且有助于语言建模. 文献 [32] 采用 label smoothing 方法来避免模型对于预测结果过于自信. 文献 [33] 使用最小化词错误率的方式进行区分性训练. 模型除了训练和推理过程训练时通常使用 CE, 而在评价阶段使用 WER 等.

近两年, 虽然在学术领域语音识别已经取得了接近极限的实验结果, 但研究人员对端到端语音识别的研究仍然在不断拓展和尝试. 文献 [34] 采用深层 Transformer, 认为其具有的高学习能力能够超越以前的端到端方法, 甚至可以比肩传统的混合系统. 对编码器和解码器使用 48 个 Transformer 层训练, 使用随机残差连接, 极大地提高了模型泛化性能和训练效率. 文献 [35] 提出 Jasper 模型, 其使用了一维卷积, 批量归一化, ReLU 激活, dropout 和残差连接, 同时引入了一个称为 NovoGrad 的分层优化器. 通过实验, 最多使用了 54 个卷积层的模型系取得了良好的结果. 文献 [36] 针对于在线应用问题, 认为 BLSTM 虽然代表了端到端 ASR 的先进技术, 但不适合流应用程序, 因此提出时延 LSTM (TDLSTM) 和并行时延 LSTM (PTDLSTM) 流, 它们都限制了时延大小, 保证了在线应用的效果.

1.5.5 复杂噪声环境下的语音识别

语音识别系统往往受到噪声干扰, 使其性能大大下降. 在前端添加可以把目标说话人的声音和其它干扰分开的处理过程, 就可以提高语音识别系统的鲁棒性, 从而提高准确率, 因而这已成为 ASR 系统中无法缺少的一环. 这种对语音进行去噪、分离、解混响的任务统称为语音分离. 根据干扰的不同可对领域进行细分, 当干扰为非语音噪声时, 称为语音增强; 当其为其他说话人的语音时, 称为多说话人分离, 当其为目标说话人自身的反射波时, 称为解混响.

传统的语音分离方法有谱减法、维纳滤波法、卡尔曼滤波法等. 深度学习方法流行后, 研究人员采用了 DNN、LSTM 和 CNN 等进行模型构建, 取得了良好的效果. 文献 [37] 采用 GAN 作为模型, 将生成器全部设置为卷积层, 减少了参数和训练时间, 判别器负责向生成器提供生成数据的真伪信息, 从而使模型参数逐渐向生成纯净语音方向变化. 训练目标一般包括两类, 一类是基于 Mask 的方法, 另一类是基于频谱映射方法. 基于 Mask 方法主要有理想二值掩蔽 (IBM) 和理想比率掩蔽 (IRM), 都是根据听觉感知特性, 将音频信号分成不同的子带, 根据不同的信噪比, 将能量设置为 0、1 或者相应比率. 频谱映射采用谱特征, 让模型通过有监督学习, 使其自己能够学习到有干扰谱和无干扰谱之间的映射关系[38]. 文献 [39] 在 CHiME-5 挑战数据集上取得了良好的成绩. 其首先进行多通道解混响与增强, 再进行单通道去噪, 采用调整的波束成型方法和说话人相关训练, 测试达到了 60%WER 的效果. 文献 [40] 结合了频谱特征和空间特征训练网络, 从估计的方向和特定的频谱中提取目标语音, 可以不需要已知麦克风数量和位置. 采用目标语音的时频单元估计方向, 结合深度聚类及采用置换不变训练目标函数的 Chimera++网络, 集成时频掩蔽的波束成型技术, 使系统有了强大的分离随机排列的麦克风场景语音和解混响能力.

2 当前面临的挑战及发展趋势

对于语音识别和端到端系统来说, 学界已经研究的相当深入, 当前主要的研究热点在于, 一方面是将已经成熟的机器视觉和自然语言处理方向的方法理论迁移到语音领域, 一方面是继续深挖已有端到端技术下的各种微调和优化手段, 不断提升识别性能和鲁棒能力. 当前主要的挑战有两方面, 一是技术方面的, 另一个则是数据和工具方面的.

从技术方面来说, 首先, attention 应用到 ASR 中, 和原来应用在机器翻译领域不同, ASR 问题语音信号和文本序列之间存在着明显的时序对应关系, 需要考虑如何在模型中应用这种时序对应关系帮助我们进行

模型训练. 同时, 翻译问题中文本中存在着明显的词边界, 其 encoder 能够提供更加清晰的隐层表示, 对于 ASR, 需要考虑怎么获取更加清晰和更加有区分性的隐层表示. 其次, 相比于传统 AM, LM, 发音词典独立的模型结构, attention 方法在建模语言之间的关联关系方面存在着缺陷, 怎样能够在不增加整体语音语料和语言模型的情况下, 提高模型对于表征单词之间联系的能力. 第三, 随着建模单元的逐渐增加, 怎么更加高效的解决诸多未登录词问题. 因此, 目前 CTC 和 attention 方法可能都不是最优的端到端建模的方法, 探索新的建模方法也是未来的重点之一.

从数据和工具方面来说, 深度语音识别的实践存在阻碍. 一是高质量数据集较难获取, 语音数据的收集和标注费时费力, 只有大企业才有获得这些数据的入口, 而更多的研究者在进行研究时因为数据问题而无法得到较好结果, 只能转向传统的 GMM-HMM 方法; 二是深度语音识别的框架和工具还有待更新和简化, 这样才会将技术壁垒进一步消除, 使得更多的研究者能将语音识别应用到更广阔的场景; 三是当前识别的评价指标的指向不够泛化和实用, 往往在一个数据集上表现好的模型不一定在其他数据集上同样优秀.

因此, 当前深度语音识别研究的主要趋势就是不断深入研究端到端模型及其各种优化方法, 同时, 探索用于迁移学习的工具和数据策略, 使得语音识别也能像机器视觉领域那样遍地开花, 大大提高整个社会的人工智能化水平.

3 结论与展望

本文简要介绍了语音识别技术发展历史并详细阐述了语音识别中端到端技术的进展, 同时分析了当前语音识别所面临的挑战与趋势. 在现有深度语音识别的研究基础上, 应当继续探索端到端技术的潜力, 并着力解决数据和框架工具等影响实践的障碍, 从而使其能更广泛更方便的应用到实际任务中.

参考文献

1 韩纪庆, 张磊, 郑铁然. 语音信号处理. 北京: 清华大学出版社, 2004.

2 Yu D, Deng L. 解析深度学习-语音识别实践. 俞凯, 钱彦旻, 译. 北京: 电子工业出版社, 2016.

3 Zhang SL, Lei M, Yan ZJ, *et al*. Deep-FSMN for large vocabulary continuous speech recognition. Porceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. 2018. 5869–5873. [doi: 10.1109/ICASSP.2018.8461404]

4 Amodei D, Ananthanarayanan S, Anubhai R, *et al*. Deep speech 2: End-to-end speech recognition in English and Mandarin. Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA. 2016. 173–182.

5 Davis KH, Biddulph R, Balashek S. Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 1952, 24(6): 637–642. [doi: 10.1121/1.1906946]

6 Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 357–366. [doi: 10.1109/TASSP.1980.1163420]

7 Vintsyuk TK. Speech discrimination by dynamic programming. Cybernetics, 1968, 4(1): 52–57.

8 Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics Speech and Signal Processing, 1978, 26(1): 43–49. [doi: 10.1109/TASSP.1978.1163055]

9 Buzo A, Gray A, Gray R, *et al*. Speech coding based upon vector quantization. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(5): 562–574. [doi: 10.1109/TASSP.1980.1163445]

10 Lee KF, Hon HW, Hwang MY, *et al*. The SPHINX speech recognition system. International Conference on Acoustics. IEEE, 1989.

11 Juang BH, Rabiner LR. Hidden markov models for speech recognition. Technometrics, 1991, 33(3): 251 –272. [doi: 10.1080/00401706.1991.10484833]

12 Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18(7): 1527–1554. [doi: 10.1162/neco.2006.18.7.1527]

13 Dahl GE, Yu D, Deng L, *et al*. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic. 2011. 4688–4691.

14 Dahl GE, Yu D, Deng L, *et al*. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and

Language Processing, 2012, 20(1): 30–42. [doi: 10.1109/TASL.2011.2134090]

15 LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. In: Arbib MA, ed. The Handbook of Brain Theory and Neural Networks. Cambridge: MIT Press, 1998. 255–258.

16 Abdel-Hamid O, Mohamed Ar, Jiang H, *et al*. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan. 2012. 4277–4280.

17 Sainath TN, Kingsbury B, Mohamed Ar, *et al*. Improvements to deep convolutional neural networks for LVCSR. Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic. 2013. 315–320.

18 Chan W, Jaitly N, Le QV, *et al*. Listen, attend and spell. arXiv: 1508.01211, 2015.

19 Vinyals O, Ravuri SV, Povey D. Revisiting recurrent neural networks for robust ASR. Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan. 2012. 4085–4088.

20 Graves A, Fernández S, Gomez F, *et al*. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA, USA. 2006. 369–376.

21 Graves A, Jaitly N, Mohamed Ar. Hybrid speech recognition with deep bidirectional LSTM. Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic. 2013. 273–278.

22 Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore. 2014. 338–342.

23 Chorowski J, Bahdanau D, Cho K, *et al*. End-to-end continuous speech recognition using attention-based recurrent NN: First results. arXiv: 1412.1602, 2014.

24 Hannun A, Case C, Casper J, *et al*. Deep speech: Scaling up end-to-end speech recognition. arXiv: 1412.5567, 2014.

25 Gehring J, Auli M, Grangier D, *et al*. Convolutional sequence to sequence learning. Proceedings of the 34th International Conference on Machine Learning (ICML'17). Sydney, NSW, Australia. 2017. 1243–1252.

26 Zhang Y, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. 2017. 4845–4849.

27 Chan W, Jaitly N, Le Q, *et al*. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. 2016. 4960–4964. [doi: 10.1109/ICASSP.2016.7472621]

28 Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. arXiv: 1508.04025, 2015.

29 Das A, Li JY, Zhao R, *et al*. Advancing connectionist temporal classification with attention modeling. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. 2018. 4769–4773.

30 Xu K, Ba J, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning. Lille, France. 2015. 2048–2057.

31 Chiu CC, Sainath TN, Wu YH, *et al*. State-of-the-art speech recognition with sequence-to-sequence models. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. 2018. 4774–4778.

32 Kannan A, Wu YH, Nguyen P, *et al*. An analysis of incorporating an external language model into a sequence-to-sequence model. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. 2018. 1–5828.

33 Prabhavalkar R, Sainath TN, Wu YH, *et al*. Minimum word error rate training for attention-based sequence-to-sequence models. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, AB, Canada. 2018. 4839–4843.

34 Pham NQ, Nguyen TS, Niehues J, *et al*. Very deep self-attention networks for end-to-end speech recognition. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria. 2019. 66–70.

35 Li J, Lavrukhin V, Ginsburg B, *et al*. Jasper: An end-to-end convolutional neural acoustic model. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria. 2019. 71–75.

36 Moritz N, Hori T, Le Roux J. Unidirectional neural network architectures for end-to-end automatic speech recognition. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria. 2019. 76–80.

37 Pascual S, Bonafonte A, Serrà J. SEGAN: Speech enhancement generative adversarial network. Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden. 2017. 3642–3646.

38 Williamson DS, Wang DL. Speech dereverberation and denoising using complex ratio masks. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA. 2017. 5590–5594.

39 Wang ZQ, Wang DL. Combining spectral and spatial features for deep learning based blind speaker separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(2): 457–468. [doi: 10.1109/TASLP.2018.2881912]

40 Wu J, Xu Y, Zhang SX, *et al*. Improved speaker-dependent separation for CHiME-5 challenge. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria. 2019. 466–470.