网络出版时间：2018-05-17 13:10:07
网络出版地址：http://kns.cnki.net/kcms/detail/51.1307.TP.20180517.1309.002.html

Journal of Computer Applications 计算机应用 | ISSN 1001-9081 CODEN JYIIDU | 2018-05-17 http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2018020402

基于 BLSTM-CTC 和 WFST 的端到端中文语音识别系统

姚煜，Ryad Chellali*

(南京工业大学 电气工程与控制科学学院，南京 211816)

(*通信作者电子邮箱 rchellali@njtech.edu.cn)

摘　要: 针对隐马尔科夫模型在语音识别中存在的不合理条件假设，进一步研究循环神经网络的序列建模能力，提出了基于双向长短时记忆神经网络的声学模型构建方法，并将联结时序分类训练准则成功地应用于该声学模型训练中，搭建出不依赖于隐马尔科夫模型的端到端中文语音识别系统；同时设计了基于加权有限状态转换器的语音解码方法，有效解决了发音词典和语言模型难以融入解码过程的问题。与传统 GMM-HMM 系统和混合 DNN-HMM 系统对比，实验结果显示该端到端系统不仅明显降低了识别错误率，而且大幅提升了语音解码速度，表明了文中提出的声学模型可以有效地增强模型区分度和优化系统结构。

关键词: 语音识别；长短时记忆神经网络；联结时序分类；加权有限状态转换器；端到端系统

中图分类号: TN912.34　　**文献标志码**：A

End-to-end Chinese speech recognition system using bidirectional long short-term memory networks and weighted finite-state transducers

YAO Yu , RYAD Chellali*

(College of Electrical Engineering and Control Science, Nanjing Tech University, Nanjing 211816, China)

Abstract: In view of the unreasonable conditional hypotheses of hidden Markov model in speech recognition, an acoustic model based on bidirectional long short-term network was proposed after further studying the ability of recurrent neural network. We have successfully applied the training criterion based on connectionist temporal classification to the training of this acoustic model, and built our end-to-end Chinese speech recognition system without hidden Markov model. Meanwhile, a speech decoding method base on weighted finite-state transducer was designed to effectively solve the problem that lexicon and language model are difficult to integrate into the decoding process. Compared with the traditional GMM-HMM system and the hybrid DNN-HMM system, the experimental results show our end-to-end system significantly reduces the recognition error rate, while at the same time speeding up decoding dramatically. It is shown that the acoustic model proposed in this paper can effectively enhance the model discrimination and optimize the structure of speech recognition system.

Keywords: speech recognition; long short-term memory; connectionist temporal classification; weight finite-state transducer; end-to-end system

0　引言

在过去三十年，自动语音识别系统被各大高校和研究机构广泛地研究，在性能上基本满足了日常使用的要求。在此期间，基于混合高斯模型/隐马尔科夫模型(Gaussian Mixture Model/Hidden Markov Model, GMM/HMM)的声学模型范式一度成为自动语音识别的主流框架。其中，HMM 用来处理语音信号在时序上的变化性，GMM 用来完成声学输入到隐马尔科夫状态之间的映射[1]。然后多层感知器被用来替代 GMM 完成 HMM 发射概率的计算，在一定程度上优化了这套识别框架[2]。随后，Geoffrey Hinton 领导的深度学习开始兴起，使得深度神经网络(Deep Neural Network，DNN)被引入到自动语音识别的声学模型建模当中[1,3-5]。依靠 DNN 深度抽象和强大表示学习的能力，语音识别系统的识别准确性又一次获得了大幅提升。

然而在混合 DNN/HMM 系统的训练过程中，依然需要利用 GMM 来对训练数据进行强制对齐，以获得语音帧层面的标注信息进一步训练 DNN。这样显然不利于针对整句发音进行全局优化，同时也相应地增加了识别系统的复杂度和搭建门槛。另外由于 HMM 属于生成模型，其中存在与实际发音不符的条件独立性假设[6]，导致了这套基于 HMM 的识别框架在理论上就存在重大缺陷，并不十分完美。同时在 DNN 发展基础上，循环神经网络(Recurrent Neural Network，RNN)和长短时记忆神经网络(LSTM)依靠强大的序列输入建模能

收稿日期: 2018-03-01;　修回日期:2018-04-23;　录用日期: 2018-05-04。

作者简介: 姚煜（1991—），男，江苏镇江人，硕士研究生，主要研究方向：自动语音识别，深度学习。Ryad Chellali（1964—），男，法国人，教授，博士，主要研究方向：机器学习、计算机听觉、机器人运动学。

力进一步提高了语音识别准确度[7]。并且近几年，有相关研究尝试在语音识别中应用卷积神经网络(CNN)，利用其卷积不变性来克服语言信号本身的多样性来进行语音识别，并获得不错的性能表现[8]。不过这些语音识别系统中仍然保留着 HMM 结构。这也就意味着在序列标记任务中，我们依然需要忍受不合理的 HMM 假设。

对于序列标记任务，Graves 等[9]提出了在循环神经网络训练中引入了联结时序分类(Connectionist Temporal Classification，CTC)目标函数，使得 RNN 可以自动地完成序列输入自动对齐任务。本文在 Graves 工作的基础上，进一步深入研究了深度循环神经网络(Deep RNN)，并针对汉语发音特性，提出了以声韵母为建模单元的基于双向长短时记忆神经网络(Bidirectional Long Short-Term Memory，BLSTM)的声学模型，并成功地将 CTC 函数应用于该声学模型的训练中。另外结合中文语言学知识，创新运用了基于加权有限状态转换器(Weighted Finite-State Transducer，WFST)的中文解码方法[10]，解决了发音词典和语言模型等语言学知识无法顺利融入语音解码过程中的难点问题。该解码方法将 CTC 标签、发音词典和语言模型编码进单独的 WFST 网络中，组成一个完整的搜索图。根据 CTC 网络的输出，利用束搜索技术在搜索图中解码获得最终整体得分最高的识别文字串。实验结果表明了本文设计的基于 BLSTM-CTC 的端到端语音识别系统，在识别性能上不仅大幅超越了传统的 GMM/HMM 系统，而且与同样建模单元的混合 DNN/HMM 系统相比，音素错误率和单词错误率分别降低了 4.7%和 4.43%，同时在语音识别速度上提升了一倍多。

1 基于双向长短时记忆网络的声学模型

相较于前馈神经网络(Feedforward Neural Network, FNN)，RNN 是一种允许隐层神经元存在自反馈通路的神经网络类型。循环链接使得循环神经网络的隐层单元具备了记忆上一层刺激的能力，最终随时序作用于网络输出层。这一特性也使得循环神经网络特别适用于处理序列形式的输入数据[11]。而对于一些序列标记任务，如：手写识别、语音识别，我们不但需要过去时刻的上下文信息，同样也希望获得未来时刻的上下文信息，来进一步预测当前时刻的状态。双向循环神经网络(Bidirectional Recurrent Neural Network，BRNN)就提供了一种优雅的解决方案[12]。它通过在 RNN 基础上增加一套完全独立的后向传播隐层，让前向隐层和后向隐层共同作用于输出层，这样更加准确地预测当前输出状态。图 1 为 RNN 和 BRNN 在时间维度上展开的网络结构对比。

给定一串输入特征序列 $\boldsymbol{X}=(\boldsymbol{x}_1,...,\boldsymbol{x}_T)$，BRNN 前向隐层状态 $\boldsymbol{H}^f=(\boldsymbol{h}_1^f,...,\boldsymbol{h}_T^f)$ 可从 $t=1$ 到 T 迭代计算得出：

$$\boldsymbol{h}_t^f=q(\boldsymbol{W}_{ih}^f\boldsymbol{x}_t+\boldsymbol{W}_{hh}^f\boldsymbol{h}_{t-1}^f+\boldsymbol{b}_h^f) \tag{1}$$

其中 $\boldsymbol{W}_{ih}^f$ 为输入层到前向隐层的权值矩阵，$\boldsymbol{W}_{hh}^f$ 为隐层到前向隐层的权值矩阵，$\boldsymbol{h}_{t-1}^f$ 为 $t-1$ 时刻前向隐层输出向量，$\boldsymbol{b}_h^f$ 为前向隐层偏置向量，$q(.)$ 为神经元激活函数。同时，BRNN 的后向隐层状态可由 $t=T$ 到 1 迭代计算得到：

$$\boldsymbol{h}_t^b=q(\boldsymbol{W}_{ih}^b\boldsymbol{x}_t+\boldsymbol{W}_{hh}^b\boldsymbol{h}_{t+1}^b+\boldsymbol{b}_h^b) \tag{2}$$

在每个 t 时刻，我们计算当前循环隐层的状态，并把 $[\boldsymbol{h}^f,\boldsymbol{h}^b]$ 作为下一层隐层的输入。这样不断迭代，直到计算出输出层的最终状态。

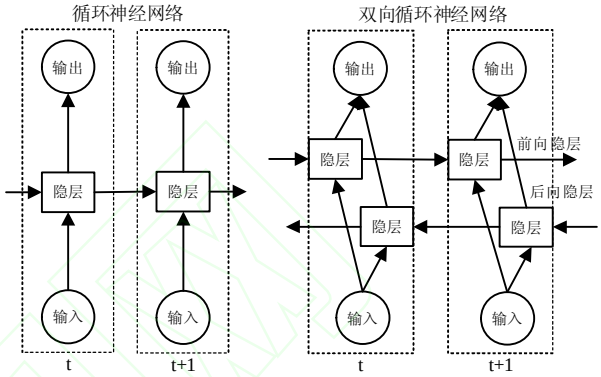


图 1 标准和双向 RNN 对比

Fig.1 Comparison of standard and bidirectional RNN

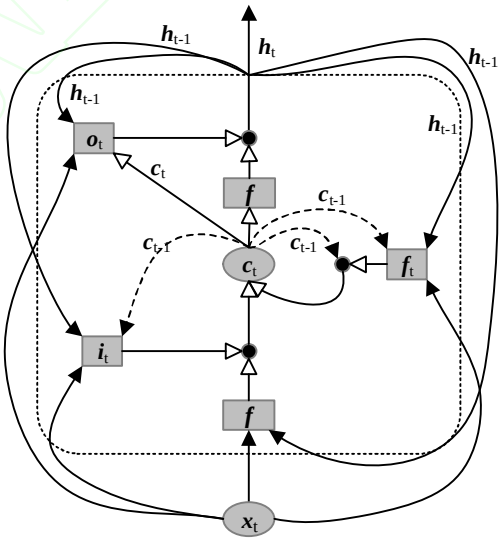


图 2 一个 LSTM 记忆块

Fig.2 A memory block of LSTM

本文采用时序反向传播算法(Back-propagation through time, BPTT)来学习循环神经网络中各层间的连接权值。在实际应用中，因为各层梯度会随反向传播不断减小而出现梯度消失现象，这样很难让 RNN 充分学习到上下文的长时依赖[13]。针对此种情况，Hochreiter 等[14]引入 LSTM 模块作为构建 RNN 隐层的单元。一个 LSTM 记忆模块主要包含：一个记忆单元，用来存储网络时序状态；和三个控制门，分别为：输入门、输出门和遗忘门，用来控制信息流。具体网络结构如图 2 所示。其中虚线箭头是将记忆单元和各个控制门联系

到一起进行精确定时输出的的窥视孔连接，实心黑体圆表示乘法单元。在 t 时刻，LSTM 的输出计算如下：

$$\boldsymbol{i}_t = s\,(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ic}\boldsymbol{c}_{t-1} + \boldsymbol{b}_j) \quad (3)$$

$$\boldsymbol{i}_t = s\,(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{fc}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f) \quad (4)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \quad \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \quad f(\boldsymbol{W}_{cx}\boldsymbol{x}_t + \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \quad (5)$$

$$\boldsymbol{o}_t = s\,(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{oc}\boldsymbol{c}_t + \boldsymbol{b}_o) \quad (6)$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \quad f(\boldsymbol{c}_t) \quad (7)$$

其中 $\boldsymbol{i}$，$\boldsymbol{o}$，$\boldsymbol{f}$，$\boldsymbol{c}$ 分别代表输入门、输出门、遗忘门和记忆单元；$\boldsymbol{W}_x$ 为与输入层连接的权值矩阵，$\boldsymbol{W}_h$ 为与上一层隐层连接的权值矩阵，$\boldsymbol{W}_c$ 为与记忆单元连接的权值矩阵；s(.) 为 sigmoid 激活函数，f(.) 为 tanh 激活函数。以同样的方式，可以计算得到后向隐层的 LSTM 的输出。

2 端到端系统的训练及解码

2.1 基于联结时序分类目标函数的端到端训练

在大词汇量连续语音识别中，声学模型训练通常采用嵌入式训练方式，即利用 HMM 让模型自动对齐语音分割与音素标记，进而来训练声学特征映射模型，如：GMM，DNN。这样无疑增加了混合 DNN-HMM 系统的复杂度。而且 HMM 的局限性也不利于语音识别技术的进一步发展。本文应用 CTC 技术对基于 BLSTM 的声学模型进行端到端地训练，彻底摆脱对于隐马尔科夫模型的依赖问题。

CTC 训练是在 RNN 网络输出层应用 CTC 目标函数，自动完成输入序列与输出标签之间的对齐。对于序列标记任务，假设存在一个大小为 K 的标签元素表 L (如：音素集，或字符集)。假设给定输入序列 $\boldsymbol{X} = (\boldsymbol{x}_1,...,\boldsymbol{x}_T)$，和对应输出标签序列 $\boldsymbol{z} = (z_1,...,z_U)$，CTC 训练的目标就是在给定输入序列下，通过调整 RNN 内部参数最大化输出标签序列的对数概率，即 $\max(\ln P(\boldsymbol{z}\,|\,\boldsymbol{X}))$。其中输出标签 $\boldsymbol{z}_u \in L \cup \{<blank>\}$，输入序列长度大于输出标签长度，即 $T \geq U$。$<blank>$ 为一个空标签，用来表示那些不属于标签元素表 L 的映射，如静音、字间停顿等。

RNN 网络的输出层是一个 Softmax 输出层,包含 $K+1$ 输出节点，分别对应扩展标签表 $L' = L \cup \{<blank>\}$ 中的每一个元素。在 t 时刻，Softmax 层得到的输出向量 $\boldsymbol{y}_t$，其中 $\boldsymbol{y}_t^k$ 代表第 k 个标签对应的后验概率。因为输出标签序列 $\boldsymbol{z}$ 和输入序列 $\boldsymbol{X}$ 没有对齐，所以很难根据 RNN 输出直接计算输出标签序列 $\boldsymbol{z}$ 的似然度。为了解决 RNN 输出和标签序列间的对齐关系，我们引入一个与输入序列在帧层面上一一对应的 CTC 路径 $\boldsymbol{p} = (p_1,...,p_T)$。在 CTC 路径中允许 $<blank>$ 标签和非 $<blank>$ 标签连续重复出现。整个 CTC 路径的概率可以由每一帧对应标签的概率组合而成，

$$P(\boldsymbol{p}\,|\,\boldsymbol{X}) = \prod_{t=1}^{T} y_t^{p_t} \quad (8)$$

通过定义映射 $\Phi: L^{\leq T}$ **a** L'^{T}，再将标签序列 $\boldsymbol{z}$ 映射到 CTC 路径 $\boldsymbol{p}$ 上。这是个1到 n 的映射，也就是一个输出标签可以对应于多个 CTC 路径。例如："A A - - B C -" 与 "- A A B - C C" 都可以被映射到标签序列"A B C"。因此可以用所有 CTC 路径的概率来表示输出标签 $\boldsymbol{z}$ 的概率，

$$P(\boldsymbol{z}\,|\,\boldsymbol{X}) = \sum_{\boldsymbol{p}\in\Phi(\boldsymbol{z})} P(\boldsymbol{p}\,|\,\boldsymbol{X}) \quad (9)$$

不过 CTC 路径的所有可能情况会随输入序列规模呈指数式增长，导致计算复杂度太大。幸运的是，借鉴传统语音识别中前向后向算法(Forward-backward Algorithm)就能够在篱笆网络中高效地计算路径似然度。首先，扩展输出标签序列 $\boldsymbol{z}$，在首尾插入索引值为 0 的 $<blank>$ 标签，并且在每个输出标签 $\boldsymbol{z}_u$ 间也插入 $<blank>$ 标签，得到一个增广式标签序列 $\boldsymbol{l} = (l_1,...,l_{2U+1})$。在由 $\boldsymbol{l}$ 构成的如图 3 所示的篱笆网络中，计算前向变量 a_t^u —在 t 时刻以 l_u 为结束标签的所有前向路径的总概率，和后向变量 b_t^u —在 t 时刻以 l_u 开始的所有后向路径的总概率。

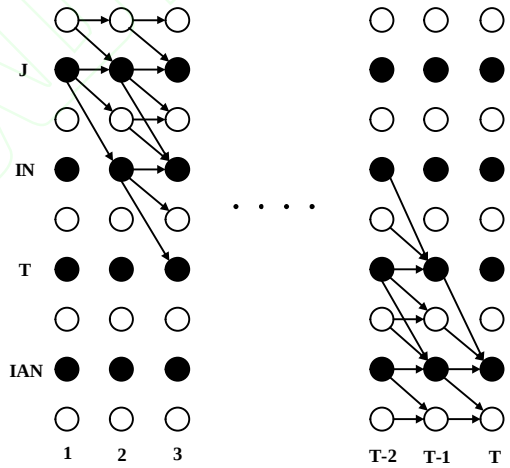


图 3 中文“今天”构成的篱笆图

Fig.3 Lattice graph constructed with Chinese word “J ING T IAN”

然后计算标签序列 $\boldsymbol{z}$ 的似然度：

$$P(\boldsymbol{z}\,|\,\boldsymbol{X}) = \sum_{u=1}^{2U+1} a_t^u b_t^u \quad (10)$$

其中 t 可以为1到 T 中的任意时刻。根据 CTC 目标函数 $\ln P(\boldsymbol{z}\,|\,\boldsymbol{X})$，对网络输出 y^t 求微分，得：

$$\frac{\partial \ln P(\boldsymbol{z}\,|\,\boldsymbol{X})}{\partial y_t^k} = \frac{1}{P(\boldsymbol{z}\,|\,\boldsymbol{X})}\frac{1}{y_t^k}\sum_{u\in g(\boldsymbol{l},\mathrm{k})} a_t^u b_t^u \quad (11)$$

其中 $g(\boldsymbol{l},k) = \{u\,|\,l_u = k\}$ 表示返回在扩展标签序列 $\boldsymbol{l}$ 中标签为 k 的下标。由于目标函数可微，可通过 BPTT 训练算法进一步计算 RNN 网络内部权值的梯度。

2.2 基于加权有限状态转换器的中文解码

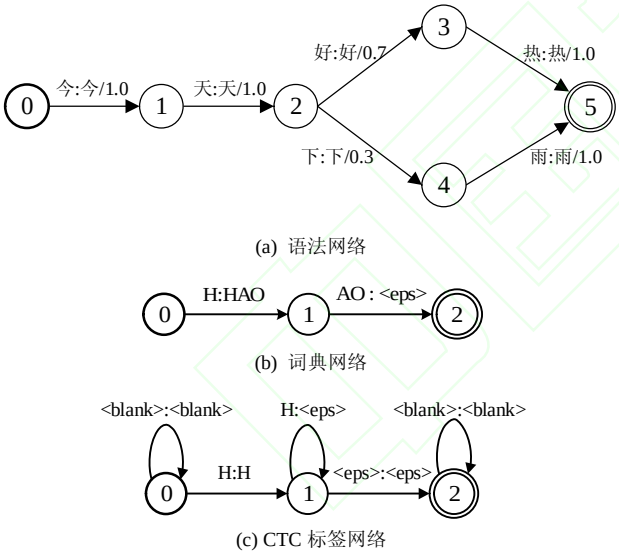
语音解码过程非常依赖于语言学知识，文献[15,16]都没能很好地融合语言文法规则。本文根据汉语的发音特点和语言学知识，将声学模型输出、发音词典和语言模型用 WFST 形式表示，构建一个基于 WFST 的综合搜索图来进行语音解码，有效地保证了语音语言学知识的完整性，同时大幅提高解码效率。

WFST 解码网络由标记(Token)转换器 、词典(Lexicon)和语法(Grammar) 三部分构成，表示形式如图 4 所示。其中，

(1) 语法转换器 G ，主要编码了符合文法规则的单词序列，可以通过手工或数据学习的方式获得。图 4(a)表示了由“今天好热”和“今天下雨”构成的语言模型，其中转换器输入、输出用“:”分隔，弧上权值对应语言模型概率。

(2) 词典转换器 L ，主要编码了发音词典构建单元与单词之间的映射关系，图 4(b)为“好”字发音的 WFST 形式，其中标记 $<eps>$ 表示一个空的输入或输出。

(3) CTC 标记转换器 T ，主要编码了语音帧级的 CTC 标签到词典单元的映射关系；图 4(c)对应了发音单元“ H ”的 CTC 转换器，其中 $<blank>$ 为 CTC 空标签；同时自循环连接可以有效地处理 CTC 网络输出重复标签的情况。



(a) 语法网络

(b) 词典网络

(c) CTC 标签网络

图 4 三种 WFST 网络示例

Fig.4 Three types of WFST network examples

实际应用中，先分别生成各自层面的标记 T 网络、字典 L 网络和语法 G 网络，然后利用组合、最小化和确定化算法将它们组合在一起。首先，通过组合操作将词典网络 L 和语法网络 G 合并得到 LG 网络；其次对 LG 网络做确定化和最小化操作，进行权重推移以优化 WFST 网络；最后与 CTC 标签网络合并，生成一个完整搜索图 S ：

$$S = T \circ min(det(L \circ G)) \tag{12}$$

其中 $\circ$, det 和 min 分别表示组合，确定化和最小化操作[17,18]。搜索图 S 可以将本文声学模型输出的 CTC 标签映射为对应的文字序列，并且得到每种可能文字序列的概率，最终选择整体得分最高的文字序列作为识别结果。

基于 BLSTM-CTC 的端到端中文识别系统的具体识别流程如图 5 所示。图 5 的下半部分为双向长短时记忆神经网络的输入部分，是原始语音信号经过加窗分帧操作和特征提取后的声学参数序列；中间部分描述的是一个深度长短时记忆神经网络，分别由输入层、输出层和多个的循环隐层构成，能够根据上下文信息输出当前语音帧对应的 CTC 标签概率；上半部分描述了一个 CTC 篱笆网络，最终可以通过网络输出概率和完整的 WFST 转换器在 CTC 网络中搜索得到输入语音序列对应的字符串识别结果。

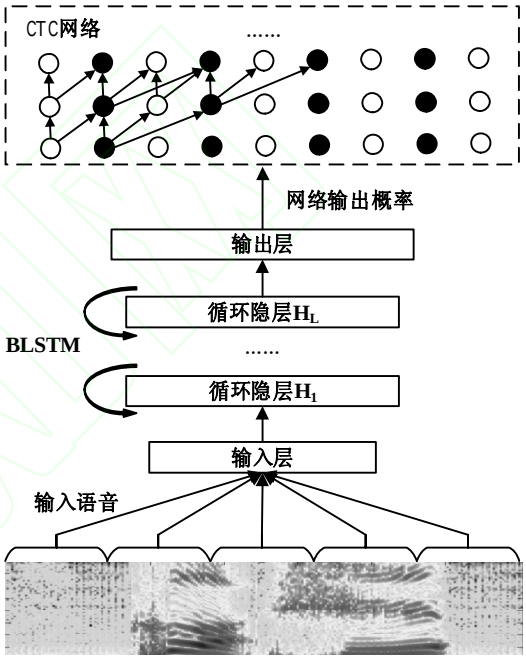


图 5 BLSTM-CTC 声学模型流程

Fig.5 Flowchart of BLSTM-CTC-based acoustic model

3 实验与分析

3.1 实验数据

本文实验在 THCHS-30 中文数据集上进行。该数据集包含来自 50 人的 35 小时录音数据，采样频率为 16kHz，量化位数为 16bit。其中 25 小时(10000 句)录音数据作为训练集，大约 2 小时(893 句)作为开发集，剩下的大约 6 小时(2495 句)作为测试集。同时该数据集还包含一个基于单词的 3-gram 文法模型和一个基于音素的 3-gram 文法模型，及对应的词典和音素词典。由于汉语以音节为发声单元，包含 23 个声母和 24 个韵母，并且每个发音单元包含阴平、阳平、上声、去声和轻音五种声调。于是本文采用包含声调的共 218 个声韵母为基本建模单元，测试各个模型的性能。

3.2 声学模型对比

1) GMM-HMM. 输入特征参数采用包含一阶、二阶差分共 39 维的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient，MFCC)，分别以单音子、三音子为建模单元进行实验，其中 HMM 状态经过基于决策树的状态聚类处理。单音子模型的状态数为 656 个，搜索图大小为 718MB；三音子模型的上下文相关状态数为 1658 个，搜索图大小为 747MB。

2) DNN-HMM. 特征参数为 40 维的梅尔标度滤波器组特征参数(Mel-scale Filter Bank，FBank)，同样分别以单音子、三音子为建模单元。其中 DNN 模型包括 4 个隐层，每层 1024 个神经元节点；输入层经拼帧操作后总共有 440(40×11)个节点；在对单音子建模时，DNN 输出层包含 656 个输出节点；对三音子建模时，DNN 输出层包含 165 8 个输出节点。训练数据通过 GMM-HMM 系统进行强制对齐，以获得了语音帧级别的标注信息。另外，搜索图与上述 GMM-HMM 系统相同。

3) BLSTM-CTC. 输入特征参数为包含一阶、二阶差分共 120 维梅尔标度滤波器组特征(Mel-scale Filter Bank，FBank)，建模单元为单音节，其中 BLSTM 声学模型包括 4 个隐层，每层包含正向和后向传播两部分共 640 个 LSTM 单元；输入层包括 120 个输入节点；输出层包括 220 个输出节点。将 CTC 目标函数作为 BLSTM 网络的目标函数，和利用 BPTT 算法来训练循环神经网络。搜索图大小为 439MB。

3.3 结果对比

本文所有实验都是基于相同的训练集、开发集、测试集和语言模型，在 GTX 1070 显卡和 i7 CPU 构建的硬件计算平台上，利用 Kaldi 语音识别工具包完成了对以上各种声学模型的测试。图 6 展示了各系统性能对比结果，其中本文提出的 BLSTM-CTC 端到端系统获得了 11.16%的 PER 和 24.92%的 WER。与基于单音节和基于上下文相关状态的传统 GMM-HMM 识别系统相比，在 PER 上分别降低了 21.04%和 9.13%，在 WER 上分别降低了 25.91%和 11.02%。另外，与基于单音节的混合 DNN-HMM 系统相比，端到端系统在 PER 和 WER 上分别降低了 4.7%和 4.43%；同时识别率上非常接近了基于上下文相关状态的混合 DNN-HMM 识别系统的 10.27%的 PER 和 23.69%的 WER。实验结果表明了基于 BLSTM 的声学模型在应用 CTC 训练准则后，充分挖掘循环神经网络对序列数据的建模能力，在模型表示能力上要明显优于 GMM 和 DNN，同时也使得语音识别系统摆脱了不合理 HMM 条件假设。

另外，我们对语音识别系统的另一个重要性能指标——解码速度做了对比分析。由 3.2 节中给出的各模型搜索图和图 6 中实时率的对比，可以发现端到端系统在解码时间上相比于 GMM-HMM 系统减少了 2 倍，并且相比于混合 DNN-HMM 系统减少了 1.3 倍；同时搜索图(TLG)大小只有 GMM-HMM 系统的 GMM-HMM 和 DNN-HMM 系统的 0.61 倍。证实了端到端技术在有效降低识别错误率的同时优化了系统结构，并获得了巨大的存储空间和解码时间上的节省。

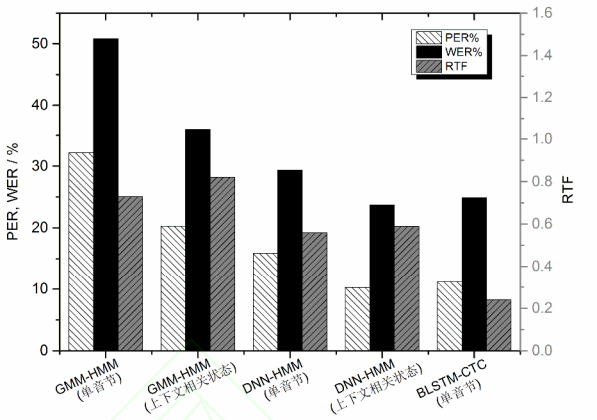


图 6 不同系统音素错误率(PER)、单词错误率(WER)和实时率(RTF)对比

Fig.6 Comparisons of phoneme error rate (PER), word error rate (WER) and real time factor (RTF) between different system.

4 结语

本文深入研究深度循环神经网络，搭建了基于 BLSTM-CTC 的端到端中文语音识别系统。该端到端系统将 CTC 训练准则成功地应用在 BLSTM 声学模型训练中，摆脱了对 HMM 的依赖；结合汉语语言学知识，设计了基于 WFST 的解码方法，解决了声学模型、发音词典和语言模型难以融合的问题。在 THCHS-30 数据集上，实验表明该端到端技术不仅明显地降低了识别错误率，而且有效地简化系统复杂度，大幅提升了识别速度。当然语音识别是一个受外部环境和说话人因素影响非常大的多重识别过程。如何减少这些因素的影响，进一步提升模型的区分度和鲁棒性，将是下一阶段的研究工作重点。另外通过迁移学习，将端到端系统应用于不同领域或不同环境，也是接下来研究的方向。

参考文献

[1] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6):82-97.

[2] VALENTE F, MAGIMAI-DOSS M, WANG W. Analysis and comparison of recent MLP features for LVCSR systems[C]// INTERSPEECH 2011, Conference of the International Speech Communication Association, Florence, Italy, August. 2011:28-31.

[3] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on audio, speech, and language processing, 2012, 20(1): 30-42.

[4] MOHAMED A R, HINTON G, PENN G. Understanding how deep belief networks perform acoustic modelling[C]// ICASSP 2012, IEEE

International Conference on Acoustics, Speech and Signal Processing. IEEE, 2012:4273-4276.

[5] VESELY K, GHOSHAL A, BURGET L, et al. Sequence-discriminative training of deep neural networks[C]// Interspeech. 2013: 2345-2349.

[6] BLASIAK S, RANGWALA H. A hidden markov model variant for sequence classification[C]// IJCAI Proceedings-International Joint Conference on Artificial Intelligence. 2011, 22(1): 1192.

[7] HAYASHI T, WATANABE S, TODA T, et al. Duration-controlled LSTM for polyphonic sound event detection[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2017, 25(11): 2059-2070.

[8] SAON G, KUO H K J, RENNIE S, et al. The IBM 2015 english conversational telephone speech recognition system[C]// Interspeech. 2015:6-10.

[9] GRAVES A, FERNANDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// ICML, international conference on Machine learning. ACM, 2006: 369-376.

[10] MOHRI M, PEREIRA F, RILEY M. Speech recognition with weighted finite-state transducers[M]// Springer Handbook of Speech Processing. Springer Berlin Heidelberg, 2008: 559-584.

[11] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]// ICASSP, Acoustics, speech and signal processing, 2013 IEEE international conference on. IEEE, 2013: 6645-6649.

[12] MORILLOT O, LIKEFORMANSULEM L. New baseline correction algorithm for text-line recognition with bidirectional recurrent neural networks[J]. Journal of Electronic Imaging, 2013, 22(2):023028.

[13] WOLLMER M, SCHULLER B, EYBEN F, et al. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening[J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(5): 867-881.

[14] SAINATH T N, VINYALS O, SENIOR A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]// ICASSP, Acoustics, Speech and Signal Processing, 2015 IEEE International Conference on. IEEE, 2015: 4580-4584.

[15] MAAS A, XIE Z, JURAFSKY D, et al. Lexicon-free conversational speech recognition with neural networks[C]// ACL, Conference of the Association for Computational Linguistics: Human Language Technologies. 2015: 345-354.

[16] HANNUN A, CASE C, CASPER J, et al. Deep Speech: scaling up end-to-end speech recognition[J]. Computer Science, 2014.

[17] POVEY D, GHOSHAL A, BOULIANNE G, et al. The kaldi speech recognition toolkit[C]// ASRU, Conference of the Automatic Speech Recognition and Understanding. 2011, EPFL-CONF-192584.

[18] DROSTE M, KUICH W, VOGLER H. Handbook of weighted automata[J]. Monographs in Theoretical Computer Science An Eatcs, 2009, 380(1-2):69-86.

Backgroud：

YAO Yu, born in 1991, M. S. candidate. His research interests include automatic speech recognition, deep learning.

Ryad Chellali, born in 1964, Ph. D., professor. His research interests include machine learning, computer vision, robot kinematics.