计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于 CTC 准则的普通话识别及改进
作者：张立民，王彦哲，张兵强，朱念斌
DOI：10.19678/j.issn.1000-3428.0051065
网络首发日期：2018-06-27
引用格式：张立民，王彦哲，张兵强，朱念斌．基于 CTC 准则的普通话识别及改进．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0051065

网络首发时间：2018-06-27 11:50:34
网络首发地址：http://kns.cnki.net/kcms/detail/31.1289.TP.20180626.1804.008.html

计 算 机 工 程
Computer Engineering

基于 CTC 准则的普通话识别及改进

张立民[1]，王彦哲[1]，张兵强[1]，朱念斌[2]
（1. 海军航空大学 信息融合研究所，烟台 264000；
2.中国人民解放军 61923 部队，北京 100000）

摘 要：主流神经网络训练的交叉熵准则是对声学数据的每个帧进行分类优化，而连续语音识别是以序列级转录准确性为性能度量。针对这个不同，本文构建了基于序列级转录的端到端语音识别系统。针对低资源语料条件下系统性能不佳的问题，在实验中改进 LSTM 网络的结构，在解码过程结合词典和语言模型，同时前端增加音调特征来丰富声学特征。最后将语音识别的传统技术融入构建的系统，用序列区分度训练技术去提升 CTC 模型的建模效果。实验结果表明系统性能提升约 25%，优于主流语音识别系统。

关键词：序列级；低资源；端到端；解码；声学特征；区分度训练

Mandarin Recognition and Improvement Based on CTC Criterion

ZHANG Limin，WANG Yanzhe，ZHANG Bingqiang，Zhu Nianbin
（1. Institute of Information Fusion，Naval Aeronautical University，Yantai 264000，China；
2. Troops 61923 of PLA，Beijing 100000，China）

【Abstract】The cross-entropy criterion of mainstream neural network training is to classify and optimize each frame of acoustic data，while the continuous speech recognition uses the sequence-level transcription accuracy as a performance measure. In view of this difference, an end-to-end speech recognition system based on sequence level transcription is constructed in this paper. In order to solve the problem of poor system performance under the condition of low resource corpus, the structure of LSTM network is improved in the experiment, the dictionary and language model are combined in the decoding process, and the tone feature is added to the front end to enrich the acoustic feature. Finally，the traditional speech recognition technology is integrated into the constructed system，and the modeling effect of CTC model is improved by using the sequence discrimination training technique. Experimental results show that the system performance is improved by about 25%，better than mainstream speech recognition systems.

【Key words】sequence level；low resource；end-to-end；decode；acoustic feature；discrimination training
DOI:10.19678/j.issn.1000-3428.0051065

0 概述

构建现代自动语音识别（ASR）系统是一项复杂任务，系统基于严格设计的处理流程，包括输入特征、声学模型、语言模型和隐马尔可夫模型（HMM）。深度学习算法的引入使得传统的混合高斯模型（Gaussian Mixture Model，GMM）开始被神经网络（Deep Neural Network，DNN）取代，来对状态输出进行建模，语音识别的准确率开始得到大幅度的提高[错误!未找到引用源。]。在 DNN 的基础上，卷积神经网络（Convolutional Neural Network，CNN）和循环神经网络（Recurrent neural networks，RNN）的应用让语音建模能力得到进一步提高。

以上这些算法通常侧重于改进声学模型，而声学模型只是系统里面的一个部件。语音识别中，神经网络被训练成使用交叉熵准则（Cross-entropy criteria，CE）作为目标函数[错误!未找到引用源。]，来分类声学数据的各个帧，这与以序列级的转录准确性为实际性能指标有很大的不同。帧级别的训练标注必须事先获得，而这需要通过训练好 GMM-HMMs，然后对训练集进行强制对齐来得到，此外还需要相应的专业语音学、语言学知识。由此增加了构建现代语音识别系统的难度，门槛相对较高。

为了尽可能地减少系统所需流程，Graves 在论文[错误!未找到引用源。]中提出了基于深度长短时记忆网络（Long

基金项目：国家自然科学基金重大研究计划资助项目（编号 NO.91538201）；泰山学者工程专项经费资助（编号 NO.ts201511020）
作者简介：张立民（1966－），男，教授，博士生导师，主研方向电子系统仿真、人工智能；王彦哲（通信作者），硕士研究生；张兵强，副教授，博士；朱念斌，高级工程师，硕士。
E-mail：iamwyz@foxmail.com

Short-Term Memory，LSTM）和连接时序分类（Connectionist Temporal Classification，CTC）错误!未找到引用源。错误!未找到引用源。目标函数的组合，可以不需要中间的语音表示，直接把音频数据转录成文本，而且无需在输入序列和目标序列之间进行任何预先对齐。虽然端到端语音识别代表了以后发展的趋势，CTC 的鲁棒性需要足够的语料进行充分训练，而且由于解码时语言模型的缺失，相对于传统的深度学习的建模方法，在识别性能和准确率上存在着一定的差距错误!未找到引用源。。

之前的许多工作主要集中在英语语音识别，因此本文结合深度双向长短时记忆网络（Deep Bidirectional LSTM，DBLSTM）和 CTC 技术，建立一个中文普通话的端到端语音识别系统，据中文普通话的语音特点，以声韵母作为建模单元，解码时结合词典和语音模型，最后通过序列区分性训练提升声学模型区分能力。实验结果表明，在资源受限条件下的中文普通话识别，本系统取得与传统深度学习声学模型相当的识别率。

1 长短时记忆网络

标准的前馈网络通常仅考虑帧的固定长度滑动窗口中的信息，因此不能利用语音信号中的长范围相关性。RNN 可以将序列历史编码为它们的内部状态，并且可以基于当前帧之前观察到的所有语音特征来预测音素，具有学习序列的复杂动态时间的优势。

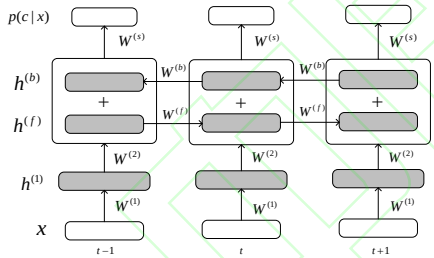

图 1 循环神经网络（RNN）在时间上展开

然而标准的 RNN 在时序上对序列进行处理时通常忽视了未来的上下文信息。在一些情况下，当前时刻的输出同时需要之前和之后的状态，而语音的前后相关性十分明显，于是双向 RNN 网络（如图 1）就被提出来充分利用语音的未来信息。假设给定一个输入序列 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2...,\boldsymbol{x}_t)$，一个循环层通过从 t=1 到 t=T 来计算隐藏状态的前向序列：

$$\overrightarrow{\boldsymbol{h}}_t=\sigma(\overrightarrow{\boldsymbol{W}}_{hx}\boldsymbol{x}_t+\overrightarrow{\boldsymbol{W}}_{hh}\overrightarrow{\boldsymbol{h}}_{t-1}+\overrightarrow{\boldsymbol{b}}_h) \tag{1}$$

其中 $\overrightarrow{\boldsymbol{W}}_{hx}$ 是输入层到隐藏层的权重矩阵，$\overrightarrow{\boldsymbol{W}}_{hh}$ 是隐藏层之间的权重矩阵，σ 是逻辑 sigmoid 函数。除了输入 $\boldsymbol{x}_t$ 之外，前一时刻的隐藏激活 $\overrightarrow{\boldsymbol{h}}_{t-1}$ 被传递来影响当前时刻的隐藏输出。附加的循环层通过从 t=T 到 t=1 来计算隐藏状态的反向序列：

$$\overleftarrow{\boldsymbol{h}}_t=\sigma(\overleftarrow{\boldsymbol{W}}_{hx}\boldsymbol{x}_t+\overleftarrow{\boldsymbol{W}}_{hh}\overleftarrow{\boldsymbol{h}}_{t-1}+\overleftarrow{\boldsymbol{b}}_h) \tag{2}$$

本文用的是一个深层结构，其中堆叠着多个双向 RNN，在每个帧 t，前向和后向的输出 $[\overrightarrow{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t]$ 是下一个双向层的输入。假设所有 N 层使用同一种激活函数，隐藏矢量序列 $\boldsymbol{h}^n$ 从 n=1 到 N 和 t=1 到 T 来迭代计算：

$$\boldsymbol{h}_t^n=\sigma(\boldsymbol{W}_{h^{n-1}h^n}\boldsymbol{h}_t^{n-1}+\boldsymbol{W}_{h^nh^n}\boldsymbol{h}_{t-1}^n+\boldsymbol{b}_h^n) \tag{3}$$

其中定义 $\boldsymbol{h}^0=\boldsymbol{x}$，网络的输出 $\boldsymbol{y}_t$ 为：

$$\boldsymbol{y}_t=\boldsymbol{W}_{h^Ny}\boldsymbol{h}_t^N+\boldsymbol{b}_y \tag{4}$$

RNN 很适合序列建模任务，但其本身存在随时间呈指数增加或减小的梯度问题，很难在长时间序列任务上训练，实际上只能模拟短程效应，而长短时记忆网络的提出，很好地克服这个问题，而且已经在多种 ASR 任务上取得优于 DNN 的结果错误!未找到引用源。。

虽然 LSTM 的建模能力很强大，由于要对大量的网络参数进行优化，让 LSTM 网络的训练很缓慢，需要很高的计算能力，尤其是对于 BLSTM。H.Sak 等人错误!未找到引用源。提出了带投影层的长短记忆网络(Long Short Term Memory Projection，LSTMP）结构，即在 LSTM 层的顶部都具有单独的线性投影层，这个投影矢量比 LSTM 的输出具有更低的尺寸，结果表明 LSTMP 可以更好地利用模型参数，有助于降低错误率，并且加快了训练速度错误!未找到引用源。。

给定一个输入序列 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2...,\boldsymbol{x}_t)$，LSTMP 网络从 t=1 到 T 进行迭代计算：

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t+\boldsymbol{W}_{ip}\boldsymbol{p}_{t-1}+\boldsymbol{W}_{ic}\boldsymbol{c}_{t-1}+\boldsymbol{b}_i) \tag{5}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t+\boldsymbol{W}_{fp}\boldsymbol{p}_{t-1}+\boldsymbol{W}_{fc}\boldsymbol{c}_{t-1}+\boldsymbol{b}_f) \tag{6}$$

$$\boldsymbol{a}_t=g(\boldsymbol{W}_{cx}\boldsymbol{x}_t+\boldsymbol{W}_{cp}\boldsymbol{p}_{t-1}+\boldsymbol{b}_c) \tag{7}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\boldsymbol{a}_t \tag{8}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t+\boldsymbol{W}_{op}\boldsymbol{p}_{t-1}+\boldsymbol{W}_{oc}\boldsymbol{c}_t+\boldsymbol{b}_0) \tag{9}$$

$$\boldsymbol{p}_t=\boldsymbol{W}_{pm}(\boldsymbol{o}_t\boldsymbol{h}(\boldsymbol{c}_t)) \tag{10}$$

$$\boldsymbol{y}_t=soft\max(\boldsymbol{W}_{yp}\boldsymbol{p}_t+\boldsymbol{b}_y) \tag{11}$$

其中 $\boldsymbol{i}$，$\boldsymbol{f}$，$\boldsymbol{o}$，$\boldsymbol{a}$ 和 $\boldsymbol{c}$ 分别代表了输入门，忘记门，输出门，单元输入激活和单元状态向量。$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示权重矩阵和偏置向量，$\boldsymbol{W}_{ic}$，$\boldsymbol{W}_{fc}$ 和 $\boldsymbol{W}_{oc}$ 是各个门和窥孔（Peephole）连接的对角线权重矩阵。g 和 h 是输入和输出激活函数，本文采用 tanh 函数。

2 基于连接时序分类的语音识别系统

在自动语音识别（ASR）中用神经网络进行声学建模的方法有许多种，必须选择声学建模的基本单元（例如 HMM 状态、音素、字符等等），训练的时候通常需要预先和标签进行对齐，然后以目标函数作为训练的标准。与混合深度学习语音识别系统不同，如图 2 所示，本文 DBLSTM 系统未使用交叉熵准则来训练帧级标签，而是专注于端到端的训练，采用 CTC 的目标函数，解码时用基于加权有限状态机(Weighted Finite State Transducers，WFST）的方法[错误!未找到引用源。]。

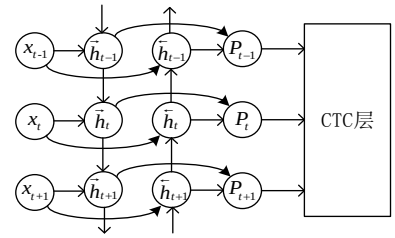

图 2 DBLSTM-CTC 语音识别系统

2.1 连接时序分类（CTC）

CTC 使用 RNN 来学习序列标记，可以直接建模语音特征和标签之间的映射，而不必依赖音频序列和标签序列之间的对齐，其输入来自于 RNN 的 softmax 层的输出，假设训练数据中的标签序列包含 K 个标签，softmax 层中的节点与标签序列相对应，额外添加一个空白标签，用于估计特定时刻不输出标签概率。因此，来自网络的输出标签概率定义了包括空白标签在内所有可能的输入序列标签的概率分布。

给定一个长度为 T 的系统的输入序列 $\boldsymbol{X}$，在 t 时刻 softmax 层输出标签或者空白的索引 k 的概率为：

$$P(k\mid t,\boldsymbol{x})=\frac{\exp(y_t^k)}{\sum_{k'}\exp(y_t^{k'})} \tag{12}$$

概率 $p(\boldsymbol{z}^l\mid\boldsymbol{x})$ 是每个时间步的输出概率的乘积：

$$p(\boldsymbol{z}^l\mid\boldsymbol{x})=\prod_{t=1}^{T}P(k\mid t,\boldsymbol{x}) \tag{13}$$

其中 $\boldsymbol{z}^l$ 是网格编码的 $\boldsymbol{x}$ 和 $\boldsymbol{l}$ 所有可能的对齐，其中允许标签重复和空白标签的存在。因此一个音频序列，存在许多可能的路径与其对应，要把这些路径映射到转录，去掉重复的标签和路径里面的空白。标签序列 $\boldsymbol{l}$ 是由映射到 $\boldsymbol{l}$ 的所有可能 CTC 路径的集合表示，可能性是所有路径概率的总和：

$$p(\boldsymbol{l}\mid\boldsymbol{x})=\sum_{p\in\phi(\boldsymbol{l})}p(\boldsymbol{z}\mid\boldsymbol{x}) \tag{14}$$

$\phi(\boldsymbol{l})$ 是与 $\boldsymbol{l}$ 对应的CTC路径的集合，这是一个多对一的映射，因为多个 CTC 路径可以对应相同的标签序列。而 $p(\boldsymbol{l}\mid\boldsymbol{x})$ 可以通过前向-后向算法来得到[错误!未找到引用源。]。

CTC 的损失函数被定义为每个训练样本正确标记的负对数概率之和，而且函数可微，因此可通过反向传播算法训练 CTC 网络：

$$\mathcal{L}_{CTC}=-\sum_{(\boldsymbol{x},\boldsymbol{l})}\ln p(\boldsymbol{z}^l\mid\boldsymbol{x})=-\sum_{(\boldsymbol{x},\boldsymbol{l})}\mathcal{L}(\boldsymbol{x},\boldsymbol{z}^l) \tag{15}$$

CTC 和传统的标签框架主要有两点区别：1）当输出不确定时，额外添加的空白标签可以减少网络在一帧进行标签预测；2）训练标准是对状态序列的对数概率进行优化，而不是输入的对数似然。

2.2 基于 WFST 的解码方法

网络经过 CTC 的训练后，需要有算法来对其进行解码。最开始的贪婪搜索方法无需添加任何语言信息，选取每一个时间节点最大概率输出，搜索最佳的路径 $p\in L'^T$：

$$\arg\max_{p}\prod_{t=1}^{T}P_{AM}^{t}(p_t\mid\boldsymbol{X}) \tag{16}$$

在 CTC 训练过程中，声学模型本身具有语言模型的属性，不添加外部语言信息的情况下解码效果良好，但用大型文本语料库训练的外部语言模型是获得最佳结果所必需的，然而存在的一个问题是两种模型都包含语言模型的特性，可能会出现相互影响性能的情况，这与基于 HMM 的解码框架不一样。以往的工作引入多种方法进行解码，但这些方法要么不能整合单词级语言模型，要么在约束条件下实现整合[错误!未找到引用源。]。此外，由于空白标签的存在，CTC 声学模型的高效解码一直是一个挑战。

所以本文采用基于加权有限状态机 (WFST) 的解码方法，包括 CTC 标签、词典和语言模型都被编码成 WFST，然后组成一个全面的搜索图。

首先用扩展标签集 L'，每个单元的先验概率来对概率序列进行预处理[错误!未找到引用源。]：

$$p(\boldsymbol{x}\mid\boldsymbol{k})\propto P(\boldsymbol{k}\mid\boldsymbol{x})/P(\boldsymbol{k}) \tag{17}$$

WFST 的搜索图由三个单独的部分组成：1）语法（Grammar）WFST 基于 n-gram 形式的语言模型，对符合的单词序列进行编码；2）标注符号（Token）WFST 通过多对一的映射函数 $\phi(\boldsymbol{l})$，将扩展标签集 L' 的单元映射到标签序列 L 中的每个单元；3) 词典 (Lexicon) WFST

将标签序列 L 的单元序列映射到单词。

编译完三个独立的 WFST，再组合成一个全面的搜索图：

$$S = T \circ \min(\det(L \circ G)) \tag{18}$$

其中∘、min、det 分别代表构图、最小化、确定化算法，搜索图 S 将由语音帧得到的 CTC 标签序列映射为字序列。此方法提供一种处理空白标签的便捷方式，可以有效并高效地将词语模型并入 CTC 解码中，提升了解码性能，解码效率优于传统的 HMM 模型。

3 实验结果及分析

实验采用清华大学开源的中文普通话语料库 THCHS-30 [错误!未找到引用源。]，该语料库使用单个麦克风，在室内安静的环境下以 16Khz 的采样频率和 16bits 的采样大小进行录制。声学模型的训练集选取时长 25 小时共 10000 句发音，测试集时长 6.24 小时共 2495 句发音，而验证集时长 1.9 小时共 893 句发音用来交叉验证训练时的效果。语言模型为三元语音模型（3-gram），由一个从中文 Gigaword 语料库中随机选取的文本集合训练而成。

3.1 基线系统

实验总共设置 GMM-HMMs、DNN-HMMs 以及 BLSTM-HMMs 三种语音识别系统

3.1.1 GMM-HMMs

基于混合高斯模型的隐马尔科夫语音识别系统首先用标准的 13 维梅尔倒谱系数（MFCC）加一阶和二阶导数来训练一个单音素的模型，采用倒谱均值归一化（CMN）来减轻信道噪声的影响，再基于单音素模型通过线性判别式分析 (LDA) 和最大似然线性变换 (MLLT) 进行特征转换构建三音素的系统。

3.1.2 DNN-HMMs

基于 DNN 的语音识别系统通过 GMM-HMMs 获得强制对齐训练语料后的帧级标注，在交叉熵准则下训练。输入特征是 40 维的 filterbank 加一阶和二阶导数，每帧左右各拼接 3 帧，输入为 7 帧来提高 DNN 模型的区分度，拼接特征经过 LDA 处理后减少为 200 维。DNN 由 4 个隐层构成，每层有 1024 个节点，输出层有 3386 个节点。初始学习率设定为 8e-3，最小批处理数为 256 帧。

3.1.3 BLSTM-HMMs

由于语音场景的特殊性，利用未来帧的信息通常能为当前帧带来更高的识别准确率。本次实验要求的实时性不高，双向比单向的神经网络通常可以带来更好的性能[错误!未找到引用源。]，所以采用 BLSTM，其中神经网络单元为 LSTMP。系统输入特征是 40 维的 filterbank 加一阶和二阶导数，网络正向反向隐层各 2 层，每隐层有 1024 个节点，输出层有 3386 个节点。网络初始学习率为 5e-4，动量参数设定为 0.9，采用了多句并行来加快训练速度。

3.2 端到端语音识别系统

端到端的语音识别系统中所用的网络结构和 BLSTM-HMMs 相同，输入特征是 40 维的 filterbank 加一阶和二阶导数，而 CTC 的输出为 217 个节点，其中包括 216 个声韵母标签和 1 个空白标签。因为 CTC 不需要上下文决策树来获得良好的性能，所以采用上下文无关（context-independent，CI）的音素所为目标。

以音素为标签的 CTC 可以允许模型每 30ms 输出一次，而不是传统的 10ms，以较低的帧率可以减少输入和输出序列的长度，降低了解码期间的计算成本并且提供了延迟的改进[错误!未找到引用源。]。所以系统的输入将 3 个 10ms 的帧堆叠在一起，处理完一次堆叠帧可以跳过 3 帧，同时解码期间的声学得分评估每 30ms 发生一次，加快速度。

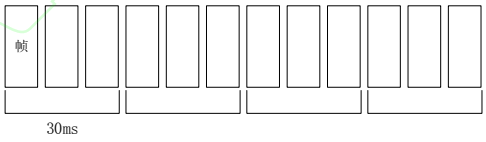

图 3 输入帧的堆叠

神经网络的模型参数初始值从[-0.1,0.1]的均匀分布范围随机抽取，初始学习率设定为 4e-5，网络中的错误由 CTC 反向传播。通常，LSTM 参数在一个区间内被初始化为小的随机值，遗忘门的偏置矢量在现有的大多数工作初始化为 0 或小的随机权重，然而这是次优选择，其中小重量有效地关闭门，防止单元记忆及其梯度及时流入[错误!未找到引用源。]。本实验中把遗忘门的偏置初始化为 1，允许信息更加轻易地流动传递。

3.3 实验结果

实验使用的机器配置 CPU 为 Intel Xeon E5-2640 v4，GPU 为 Nvidia Tesla M40，内存 128GB。每组实验分别进行 5 次，取平均值作为最后的实验结果。

3.3.1 与基线系统的比较

表 1 呈现了端到端系统与基线系统的性能对比，识别率为词错误率（Word Error Rate，WER），可以得出几个结论。首先基线系统中 BLSTM-HMMs 相对于

GMM-HMMs、DNN-HMMs 表现更好，WER 分别提升 25%、11%，说明 LSTM 的建模能力更强。此外端到端系统性能稍差于基线系统，而 LSTM 的遗忘门偏置初始化为 1 比随机初始化 WER 要提高 3.1%。

表 1 基线系统与端到端系统的性能

模型	建模单元	遗忘门偏置	识别率%（WER）
GMM-HMMs	状态	-	28.07 ± 0.12
DNN-HMMs	状态	-	23.65 ± 0.08
BLSTM-HMMs	状态	-	21.12 ± 0.06
BLSTM-CTC	音素	小的随机值	26.07 ± 0.13
BLSTM-CTC	音素	1.0	25.35 ± 0.11

图 4 呈现了相同解码搜索空间裁剪（beam）下，不同语音识别系统对测试集解码耗时。在相同硬件配置条件下，端到端系统的解码速度快于基线系统。端到端系统使用的基于 WFST 解码方法不再需要 HMM 模型，构建的解码搜索空间要比传统解码方法小，使得解码效率大幅度提升。而且 CTC 的解码在空白段的时候，beam 值可以大幅度缩小，加快速度。

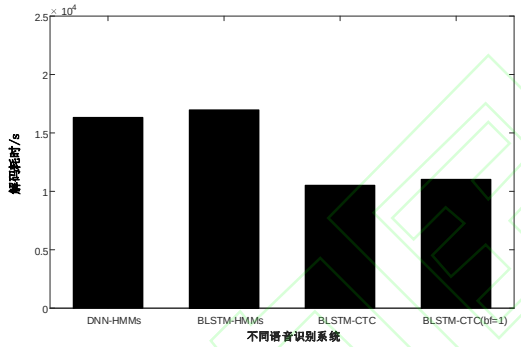

图 4 不同声学模型的解码时间

3.3.2 不同 LSTM 隐藏层的比较

表 1 的结果显示 2 层 BLSTM 的端到端系统比基线系统的性能要差，提高神经网络的建模能力的关键是让网络层数加深。表 2 结果中表明随着网络层数从 2 层增加到 3 层，系统的 WER 约有 10.6%的提升，但当网络层数增加到 4 层时，系统性能减弱，这很大程度是因为训练语料的不足，参数训练不充分导致网络欠拟合，鲁棒性下降。

表 2 不同网络层数端到端系统的性能

模型	LSTM 层数	遗忘门偏置	
		小的随机值	1
BLSTM-CTC	2	26.07 ± 0.13	25.35 ± 0.11
	3	23.29 ± 0.17	22.68 ± 0.08
	4	28.35 ± 0.06	27.53 ± 0.11

3.3.3 普通话的音调特征

在基线系统中，输入通常可以用 GMM 模型学习特征或者额外附加特征来增强模型性能。声音中的音调特征（Pitch Features）对带有声调的语音（例如普通话和粤语）识别是有利[错误!未找到引用源。]，因此实验中，将音高特征结合到具有 3 个隐藏层的 CTC 模型的训练中，在每一帧上，将三维音高添加到 40 维滤波器组特征中，由此得到 43 维特征向量。表 3 实验结果表明音调特征的添加对系统性能有提升。

表 3 添加音调特征前后系统的性能

模型	特征	识别率%（WER）
BLSTM-CTC (遗忘门偏置为 1)	filterbank	22.68 ± 0.08
	filterbank+pitch	21.87 ± 0.12

3.3.4 序列区分度训练

交叉熵和 CTC 准则对于语音识别中词错误率的最小化而言是次优的，已经证明区分度训练准则在解码过程中可以引入词汇和语音模型约束，提高由 CE 准则训练的声学模型的性能[错误!未找到引用源。]。实验继续使用状态级最小风险贝叶斯 (state level minimum Bayes risk，sMBR) 准则[错误!未找到引用源。]对用 CE 和 CTC 准则初始化的声学模型进行序列区分度训练，sMBR 准则的梯度可以通过最短路径算法来有效地计算。

表 4 序列区分度训练前后系统的性能

模型	建模单元	识别率%（WER）	+sMBR(%)
DNN-HMMs	状态	23.65 ± 0.08	21.50 ± 0.13
BLSTM-HMMs	状态	21.12 ± 0.06	19.53 ± 0.06
BLSTM-CTC (遗忘门偏置为 1)	音素	21.87 ± 0.12	19.09 ± 0.16

表 4 显示了各个声学模型进行 sMBR 训练后系统性能的变化，相对于初始模型，系统性能提升了大约 11%。说明区分度训练，能够真正根据语音识别最后过程，结合声学模型和语言模型来优化，同时相对于其他系统，CTC 系统的效果提升明显。

4 结束语

本文搭建了基于 BLSTM 的 CTC 端到端语音识别系统，用于中文普通话的识别。创新采用 WFST 的解码方法，将语言模型结合到解码过程中，提高解码效率和识别性能。前端声学特征处理额外加入了音调特征，更加符合普通话的语音特点。此外还对 LSTM 结构本身进行了探究和改进，最后经过序列区分度训练后的系统和

主流成熟的基于 HMM 的混合系统的性能相同。未来的研究主要有两点：1）改善因为训练语料的不足带来模型欠拟合的问题，同时采用融合多流特征的输入[错误!未找到引用源。]和新颖的 dropout 策略[错误!未找到引用源。]，提升系统的鲁棒性；2）加强声学模型在带噪条件下的性能表现，结合多种不同背景带噪语料进行训练[错误!未找到引用源。]。

参考文献

[1] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition [J]. IEEE Signal Processing Magazine, 2012, 29[6]: 82-97.

[2] 李伟林，文剑，马文凯．基于深度神经网络的语音识别系统研究[J]．计算机科学, 2016 (S2): 45-49.

[3] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]//International Conference on Machine Learning. 2014: 1764-1772.

[4] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[5] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 369-376.

[6] Li J, Zhang H, Cai X, et al. Towards end-to-end speech recognition for chinese mandarin using long short-term memory recurrent neural networks[C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[7] Sak H, Senior A, Beaufays F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition[J]. Computer Science, 2014: 338-342.

[8] Yu D, Li J. Recent progresses in deep learning based acoustic models[J]. IEEE/CAA Journal of Automatica Sinica, 2017, 4(3): 396-409.

[9] Miao Y, Gowayyed M, Metze F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding[J]. Automatic Speech Recognition & Understanding, 2016: 167-174.

[10] 黎长江，胡燕．基于循环神经网络的音素识别研究[J]．微电子学与计算机, 2017, 34(8): 47-51.

[11] Zenkel T, Sanabria R, Metze F, et al. Comparison of decoding strategies for ctc acoustic models[J]. arXiv Preprint arXiv:1708.04469, 2017.

[12] Wang D, Zhang X. Thchs-30: A free chinese speech corpus[J]. arXiv Preprint arXiv:1512.01882, 2015.

[13] Pundak G, Sainath T N. Lower Frame Rate Neural Network Acoustic Models[C]//Interspeech. 2016: 22-26.

[14] Miao Y, Gowayyed M, Na X, et al. An empirical exploration of CTC acoustic models[C]//Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 2623-2627.

[15] Ghahremani P, BabaAli B, Povey D, et al. A pitch extraction algorithm tuned for automatic speech recognition[C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 2494-2498.

[16] Kingsbury B. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling[C]//Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. IEEE, 2009: 3761-3764.

[17] 秦楚雄，张连海．低资源语音识别中融合多流特征的卷积神经网络声学建模方法[J]．计算机应用，2016, 36(9): 2609-2615.

[18] Billa J. Improving LSTM-CTC based ASR performance in domains with limited training data[J]. arXiv Preprint arXiv:1707.00722, 2017.

[19] 胡文君，傅美君，潘文林．基于 Kaldi 的普米语语音识别[J]．计算机工程, 2018, 1: 034.