第 39 卷 第 2 期
2020 年 3 月

应用声学
Journal of Applied Acoustics

Vol. 39, No. 2
March, 2020

◇ 研究报告 ◇

SE-MCNN-CTC的中文语音识别声学模型*

张 威[1] 翟明浩[1] 黄子龙[1] 李 巍[2] 曹 毅[1†]

(1 江南大学机械工程学院 无锡 214122)

(2 苏州工业职业技术学院 苏州 215104)

摘要: 为了解决传统卷积神经网络在识别中文语音时预测错误率较高、泛化性能弱的问题，首先以深度卷积神经网络(DCNN)-连接时序分类(CTC)为研究对象，深入分析了不同卷积层、池化层以及全连接层的组合对其性能的影响；其次，在上述模型的基础上，提出了多路卷积神经网络(MCNN)-连接时序分类(CTC)，并联合SENet提出了深度SE-MCNN-CTC声学模型，该模型融合了MCNN与SENet的优势，既能加强卷积神经网络的深层信息的传递、避免梯度问题，又可以对提取的特征图进行自适应重标定。最终实验结果表明：SE-MCNN-CTC相较于DCNN-CTC错误率相对降低13.51%，模型最终的错误率达22.21%；算法改进后的声学模型可以有效地提升泛化性能。

关键词: 深度学习；语音识别；声学模型；SE-MCNN-CTC

中图法分类号: TN912.34 **文献标识码:** A **文章编号:** 1000-310X(2020)02-0223-08

DOI: 10.11684/j.issn.1000-310X.2020.02.008

Towards end-to-end speech recognition for Chinese mandarin using SE-MCNN-CTC

ZHANG Wei[1] ZHAI Minghao[1] HUANG Zilong[1] LI Wei[2] CAO Yi[1]

(1 *School of Mechanical Engineering, Jiangnan University, Wuxi* 214122, *China*)

(2 *Suzhou Institute of Industrial Technology, Suzhou* 215104, *China*)

Abstract: In order to solve the problems of high prediction error rate and poor generalization performance with traditional convolutional neural network in Chinese speech recognition, different convolutional layers, pooling layers and fully connected layers on DCNN-CTC are analyzed in this paper. Based on the above model, two kinds of acoustic models referred as MCNN-CTC and SE-MCNN-CTC are proposed, respectively. With the combination of the advantages of MCNN and SENet in the latter model, the deep information transmission is reinforced, and the gradient problems can be effectively avoided simultaneously, the extracted feature maps can be adaptively recalibrated. Compared with DCNN-CTC, the research results show that SE-MCNN-CTC not only yields a 13.51% relative PER reduction, and the final PER is 22.21%, but also the generalization performance of the improved acoustic model can be improved effectively.

Keywords: Deep learning; Automatic speech recognition; Acoustic model; SE-MCNN-CTC

2019-07-02 收稿; 2019-11-28 定稿

*国家自然科学基金项目 (51375209), 江苏省“六大人才高峰”计划项目 (ZBZZ-012), 江苏省研究生创新计划项目 (KYCX18_0630, KYCX18_1846), 高等学校学科创新引智计划项目 (B18027)

作者简介: 张威 (1994–), 男, 江苏宿迁人, 硕士研究生, 研究方向: 语音识别。

†通信作者 E-mail: caoyi@jiangnan.edu.cn

0 引言

自动语音识别(Automatic speech recognition, ASR)技术是人机交互的一项关键技术，近年来，基于深度学习的语音识别技术取得了跨越式的发展[1−2]，在语音搜索、个人数码助理及车载娱乐系统[3]等领域广泛应用。迄今为止，已有不少旨在提高语音识别声学模型准确率的方法，上述方法大致可概括为2类：(1) 深度神经网络-隐马尔科夫模型(Deep neural networks-hidden Markov model, DNN-HMM)声学模型[4]；(2) 端到端(End-to-end)语音识别声学模型[5]。

DNN-HMM是对高斯混合模型-隐马尔科夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM)的改进，由DNN代替GMM来描述语音声学特征的概率分布，弥补了GMM对语音特征建模能力不足的缺点[6]。Li等[7]使用DNN-HMM替代GMM-HMM使得语音识别识别性能得到显著的提升；Peddinti等[8]提出了一种结合时延神经网络(Time delay neural network, TDNN)与长短时记忆网络(Long short-term memory, LSTM)声学模型，其可显著提高声学模型的识别准确率。

然而，训练一个DNN-HMM系统过程尤为复杂，并且模型的优劣很大程度上依赖人为经验[9]。相较于训练上述系统，端到端语音识别系统的训练过程非常简单。目前，端到端语音识别系统主要有3种：连接时序分类(Connectionist temporal classification, CTC)模型[10]、循环神经网络转换机制(Recurrent neural network transducer, RNN Transducer)[11]以及基于注意力机制(Attention-based)模型[12]。CTC由于其建模过程简单被广泛关注。于重重等[13]基于BLSTM(Bidirectional long short-term memory)-CTC对濒危语音识别进行研究，相较于混合系统取得了较好的实验结果；姚煜等[14]基于BLSTM-CTC-WFST(Weighted finite-state transducer)构建中文语音识别系统，明显降低了识别错误率。但上述声学模型多使用RNN网络结构，该结构参数繁多且容易出现梯度问题，卷积神经网络(Convolutional neural network, CNN)由于权值共享、局部连接以及池化等操作[15]，有效地弥补了RNN的缺点，使其区别于DNN、RNN等网络架构成为神经网络的一个重要分支，并在图像识别[16−18]、视频动作识别[19]等领域取得到显著成功。Abdel等[20]首次结合CNN与HMM，构建CNN-HMM混合系统取得了开创性的进展；Sainath等[21]采用深度卷积神经网络(Deep convolutional neural networks, DCNN)应用于声学模型并取得显著成功；Zhang等[22]结合DCNN-CTC构建了端到端语音识别，取得了相对较好的结果；Hu等[23]提出SENet (Squeeze-and-excitation networks)网络结构，对DCNN结构提取的特征权值进行重标定，进而提高网络性能。

综上所述，本文首先在深入研究DCNN网络的基础上，结合CTC损失函数，构建DCNN-CTC声学模型。然后，在上述模型基础上对DCNN模型在宽度上进行增加，从而提出多路卷积神经网络(Multipath convolutional neural network, MCNN)-CTC声学模型。最后，综合考虑SENet与MCNN网络优势构建深度SE-MCNN-CTC语音识别声学模型，并通过实际数据集对上述声学模型有效性进行验证，模型最终错误率降至22.21%。

1 深度卷积神经网络连接时序分类

1.1 卷积神经网络

CNN主要包括卷积层、池化层以及全连接层，层与层之间通过局部连接、权值共享操作使得CNN参数相较于DNN以及RNN网络架构得到极大的减少，并在一定程度上可以避免梯度问题[16]。

图1给出了卷积神经网络用于语音识别声学模型建模时，卷积层与池化层的结构图，其中卷积层通过卷积核对特征局部进行加权计算，并且不断移动卷积窗口得到不同位置的特征；池化层对前一层提取的特征进行降采样，每一个特征图与相邻前一层的卷积层特征图唯一对应。池化层旨在通过降采样操作得到特征图空间不变性特征，同时降低网络的参数与计算量[15]，相应的计算如式(1)和式(2)所示：

$$h^{(l)} = \sigma\left(\boldsymbol{W}^{(l)} * h^{(l-1)} + b^{(l)}\right), \tag{1}$$

$$h^{(l+1)} = f_{\text{pool}}\left(h^{(l)}\right). \tag{2}$$

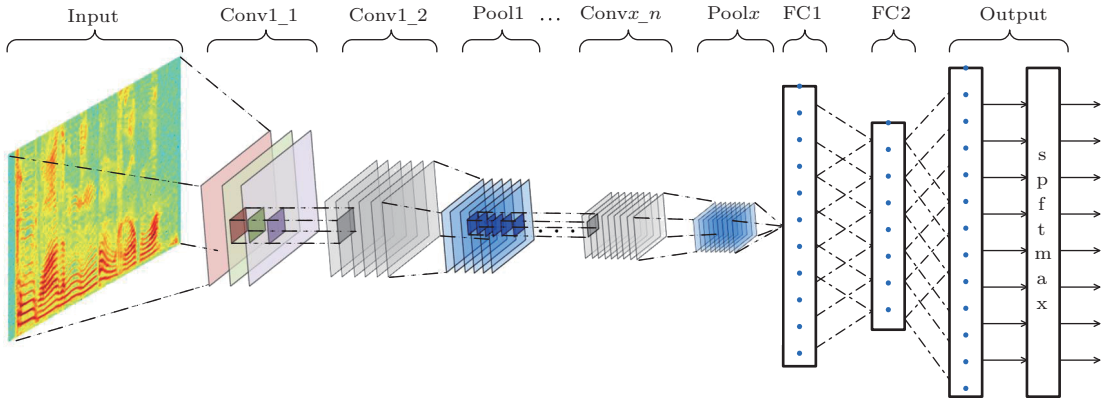


图 1　卷积神经网络结构图

Fig. 1　The structure of convolutional neural networks

1.2　连接时序分类

CTC 是由 Graves 等[10−11]提出的一种时序分类方法。CTC 与传统的基于 DNN-HMM 声学模型不同，其不需要在时间维度上帧级别对齐标签，输入语音特征即可预测结果，通过训练降低 CTC 损失值进而降低预测值与真实标签差异，该过程极大地简化了声学模型的训练流程。必须指出的是，CTC 额外引入“blank”标签对静音、字间重叠等建模，简化建模过程。因此 CTC 尤其适合序列建模，其模型结构如图 2 所示。

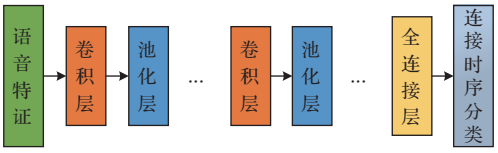


图 2　DCNN-CTC 声学模型结构图

Fig. 2　The structure of DCNN-CTC

设给定序列 $X=(x_1,x_2,\cdots,x_T)$ 表示输入 T 帧语音特征，经过神经网络输出的每帧的预测为 $Y=(y_1,y_2,\cdots,y_{T'})$，由于 CNN 中池化函数的存在，使得序列的长度成倍的变短 $T=nT'$，n 为经过池化计算后特征图减小的倍数，其中 $y_i=(y_i^1,y_i^2,\cdots,y_i^k,\cdots,y_i^m)$，$m$ 为建模单元总数，y_i^k 为第 i 帧的第 k 个建模单元位置。则给定输入序列 X，t 时刻第 k 个建模单元由神经网络 softmax 函数输出的后验概率为

$$P(k|t,X)=\frac{\exp\left(y_t^k\right)}{\sum_{k'}\exp\left(y_t^{k'}\right)}. \tag{3}$$

由式 (3)，依次得到 T' 帧中对应的建模单元的概率分布：

$$P(\pi|X)=\prod_{t=1}^{T'}P(\pi_t|t,X), \tag{4}$$

式 (4) 中，π 为生成预测 T' 序列的路径，通过累积得到对应路径 π 的概率；由于 π 与 y 为多对一关系，ψ 为路径与预测值转换函数，由式 (5) 得到路径对应预测值序列的概率：

$$P(y|X)=\sum_{\pi\in\psi^{-1}(y)}P(\pi|X). \tag{5}$$

式 (6) 通过给定真实标签 y^*，得到最终 CTC 损失函数，通过训练不断降低 CTC 损失值使得预测结果逐步朝着真实标签逼近：

$$\mathrm{CTC}(X)=-\lg\left(P\left(y^*|X\right)\right). \tag{6}$$

目前 CTC 解码主要有三种：最大路径解码、前缀束解码以及束解码[10]。最大路径解码旨在寻找每个概率最大的前 $z(z\leqslant m,m$ 为建模单元个数) 条路径对应的标签，无需字典、语言模型等先验知识，解码过程极其简单，式 (7)、式 (8) 代表其计算过程，y' 为最终的解码结果：

$$\pi^*=\underset{\pi}{\mathrm{Arg}\max}\left(P\left(\pi|X\right)\right),\quad z\leqslant m, \tag{7}$$

$$y'\approx\psi(\pi^*). \tag{8}$$

2　多路卷积神经网络

近些年，卷积神经网络大多在深度方向对网络进行优化，“串联式”连接所有层，通过提取更高维、抽象的特征以达到更佳的分类性能[16−18]。然而，对

于更宽 CNN 研究却相对匮乏。因此，本文在上述研究的基础上对 DCNN 宽度上进行深入研究，进而提出 MCNN 网络结构，即通过“并联”方式将网络进行融合构建既深又宽的网络，最终结合 CTC 目标函数构建 MCNN-CTC 声学模型，其结构如图 3 所示。

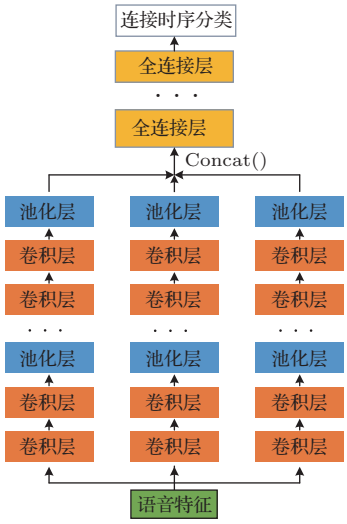


图 3 多路卷积神经网络语音识别声学模型结构图

Fig. 3 Acoustic model structure diagram of speech recognition based on multipaths convolutional neural network

传统的深度卷积神经网络仅在单条分支上提取语音序列中代表性特征 [24]，由于语音序列的多样性，造成 DCNN 在提取特征时遗漏重要特征，从而降低整体的识别准确率。为解决上述问题，本文提出了多路卷积神经网络，即采用 3 条不同的分支分别提取语音序列特征，弥补了单条分支提取特征的不足，降低了由于特征缺乏对模型识别率的影响。最终，采用反向传播 (Back propagation, BP) 算法对模型中可训练参数进行调整 [15]。

MCNN 先提取语音特征，分别将其无差别的输入到 3 条不同分支的 DCNN 中，既能在深度方向提取网络的重要特征，又可在宽度方向通过不同的卷积核提取具有代表性的语音特征，加强模型的非线性化程度，从而使得网络具有更优越的拟合性能 [15,24]，最后对提取的高维特征进行拼接，得到全部的特征序列：

$$H^l = \mathrm{Concat}\left(h_i^l, h_j^l, h_k^l\right), \tag{9}$$

式 (9) 中，h_i^l、h_j^l、h_k^l 分别代表 3 条不同支路的第 i、j、k 张特征图，Concat(·) 函数表示拼接特征图得到第 l 层的总特征图 H^l。

2.1 SE 模块

图 4 中，$\boldsymbol{X}_i$ 表示对应层的输入特征矩阵，$\boldsymbol{X}_o$ 表示经过 SENet 模型输出的特征矩阵。H、W、C 分别表示特征矩阵的三维信息；$F_{\mathrm{sq}}(\cdot)$、$F_{\mathrm{ex}}(\cdot,\boldsymbol{W}_i)$、$F_{\mathrm{scale}}(\cdot,)$ 分别代表 SENet 内部变换，计算公式如下：

$$z_c = F_{\mathrm{sq}}\left(u_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c\left(i,j\right), \tag{10}$$

$$s = F_{\mathrm{ex}}\left(z, \boldsymbol{W}\right) = \sigma\left(\boldsymbol{W}_2 f\left(\boldsymbol{W}_1 z\right)\right), \tag{11}$$

$$\boldsymbol{X}_o = F_{\mathrm{scale}}\left(u_c, s_c\right) = s_c \cdot u_c, \tag{12}$$

其中，u_c 表示经过卷积变换后第 c 个特征；z_c 表示经过全局平均池化变换后的第 c 个特征映射；$\sigma(\cdot)$ 表示 sigmoid 激活函数；s_c 表示经过全连接之后相应特征图对应的权值；$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 分别代表两层全连接层的权值矩阵，其中 $\boldsymbol{W}_1 \in \mathbf{R}^{c/\gamma \times c}$、$\boldsymbol{W}_2 \in \mathbf{R}^{c \times c/\gamma}$，其中 γ 为第一层全连接层的维度变换率；通过上述计算，最终自适应得到特征图对应的权重 [25]。

2.2 SE-MCNN 模型

综合利用 SENet 与 MCNN 各自的优势，构建了 SE-MCNN-CTC 模型，使用 SENet 模型对 MCNN 提取的特征进行概率重标定，在合适的参数范围内减少 MCNN 模型特征冗余现象。SE-MCNN 模型如图 5 所示。

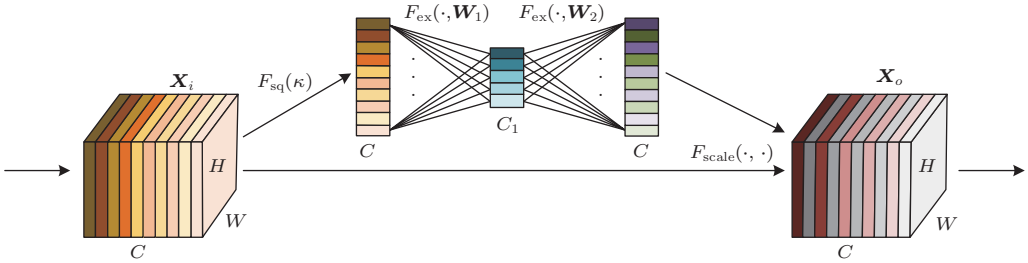


图 4 SENet 模型结构图

Fig. 4 The structure of SENet

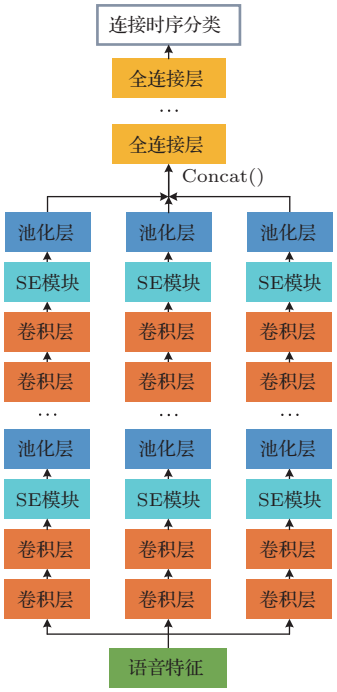


图 5　SE-MCNN-CTC 声学模型结构图

Fig. 5　Structure diagram of acoustic model for SE-MCNN-CTC

值得指出的是，SENet 模型与 MCNN 模型结合主要有三种优点：(1) 使得网络具有更强的非线性，可以更好地拟合数据；(2) 通过巧妙地设置全连接层数中的维度变换率，在提升模型的拟合能力的同时，极大地减小了 SENet 模型中全连接层神经元数目；(3) 通过对特征图的概率重标定，最大程度地利用特征图的信息，减小对冗余特征的依赖 [26]。

3　实验结果及分析

3.1　实验数据

本文使用的数据集为清华大学开源的约 30 h 数据集 (Thchs30) 和北京冲浪科技公司开源的约 150 h 中文语音数据集 (ST-CMDS)。其中 Thchs30 数据集中训练、验证集以及测试集分别为 10000 句、893 句以及 2495 句；ST-CMDS 语音数据集共 102600 句，在训练阶段采用文献 [27] 对数据进行划分；两种数据集训练集、验证集以及测试集之间均无交叠。

3.2　实验平台

实验所用硬件配置为 I7-8700K 处理器，32 GB 运行内存，GPU 显卡为 NVIDIA GTX-1080Ti；软件运行环境为 64 位 Ubuntu16.04 操作系统下搭建的 Keras+Tensorflow 深度学习框架。

3.3　数据预处理

该文以帧长 25 ms、帧移为 10 ms 提取语音原始信息。其中，Thchs30 数据集提取语谱图 (spectrogram) 为输入特征，共 200 维；ST-CMDS 数据集以 FBank(Filter Bank) 作为语音的输入特征，加上其一阶、二阶差分统计量，前后拼接一帧 [28]，共 360 维。在训练阶段选取适应性动量估计算法 (Adaptive moment estimation, Adam) 作为模型的优化器，该算法不仅能够对不同参数计算适应性学习率，而且能够加速网络收敛速度 [29]；在每层卷积层之后添加批量归一化 (Batch normalization, BN) 对网络中的权重进行自适应调整，以此提高网络的训练速度和泛化能力 [30]；在池化层之后使用丢弃法 (Dropout)[31] 以此有效地降低网络的过拟合风险，初始学习率设置为 1×10^{-3}；在微调阶段，以随机梯度下降算法 (Stochastic gradient descent, SGD) 作为模型的优化器，通过设置更小的学习率使得网络在后期优化更为稳定，微调学习率设置为 1×10^{-5}。

表 1 是对图 2、图 4 所示的声学模型参数进行配置，其中 $[3\times3 \times 32k] \times m$ 表示使用 3×3 卷积核初始数目为 32 个，每经过一个池化层，卷积核数目成倍增加；对于偶数层卷积层，则 $m = 2$，奇数层则 $m = 3$；512-1422 表示最后全连接层神经元数目依次为 512、1422。

MCNN 网络由于宽度增加而造成参数繁多，为此，将 MCNN 每层的卷积核数目相较于 DCNN 减小一半。最终实验表明：上述参数设置策略不但没有造成参数繁多难以训练现象，而且使得该配置的网络在参数减小的情况下，MCNN 模型的泛化性均无影响，所设计的声学模型如表 1 所示。

表 1　卷积神经网络配置参数信息

Table 1　Convolutional neural network configuration parameter information

模型结构	DCNN-CTC	MCNN-CTC	SE-MCNN-CTC
卷积层	$[3\times3\times32k] \times m$	$[3\times3 \times 16k]\times m$	$[3\times3 \times 16k]\times m$
SENet 模型	—	—	SE($16k/\gamma, 16k$)
池化层	2×2 最大池化		
全连接层	512-1422	512-1024-1422	

3.4 实验分析

为验证提出的MCNN-CTC以及SE-MCNN-CTC声学模型的性能，在Thchs30以及ST-CMDS数据集上对上述模型进行实验。

3.4.1 Thchs30 实验结果分析

首先以音节为建模单元在Thchs30数据集进行声学模型实验，参考文献[32]构建神经网络模型，然后采用7层卷积层，并联合CTC损失函数，构建了DCNN(7)-CTC声学模型。在上述声学模型基础上，增加卷积层数目到8层和9层，研究不同卷积层数目对声学模型的影响。最终，DCNN-CTC声学模型结构如图2所示，其实验结果如表2所示。

表2中，对比GMM-HMM以及DNN-HMM的实验结果，其中建模单元为音素，最终得到测试集字错误率；本文建模单元均为音节，最终得到音节错误率；DCNN(7)-CTC表示声学模型中卷积层数目为7层，表2不仅验证了DCNN-CTC用于声学建模的有效性，而且可得出随着CNN深度增加，错误率在不断降低，最终DCNN-CTC错误率降至25.42%。

表2 DCNN-CTC声学模型实验结果

Table 2 The experimental results of DCNN-CTC acoustic model

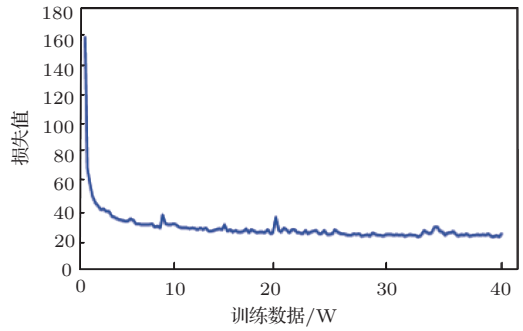
声学模型	建模单元	参数数量	测试集错误率/%
GMM-HMM[27]	音素	—	30.53
DNN-HMM[27]	音素	—	25.16
CNN-HMM[33]	音素	—	24.10
BLSTM-CTC[34]	音素	—	25.35
DCNN(7)-CTC	音节	1.95 M	26.65
DCNN(8)-CTC	音节	2.10 M	25.66
DCNN(9)-CTC	音节	2.25 M	25.42

表3 MCNN-CTC与SE-MCNN-CTC声学模型实验结果

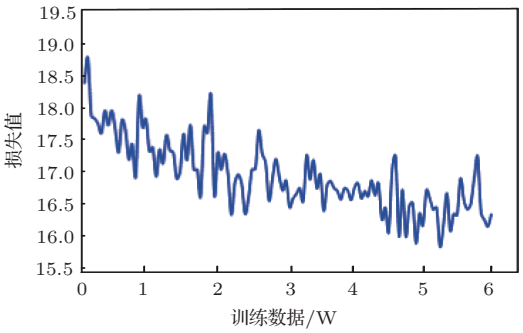
Table 3 The experimental results of MCNN-CTC and SE-MCNN-CTC acoustic model

声学模型	建模单元	参数数量	测试集错误率/%
MCNN(7)-CTC	音节	4.77 M	23.43
MCNN(8)-CTC	音节	3.61 M	24.85
MCNN(9)-CTC	音节	3.72 M	25.18
SE-MCNN(7)-CTC	音节	4.82 M	23.05

表3列出了本文提出的两种声学模型实验结果，其中MCNN(7)-CTC表示MCNN层数为7层；SENet模型在7层MCNN模型上进行改进，其中SE-MCNN(7)-CTC训练、微调训练曲线如图6(a)与图6(b)所示。



(a) Thchs30 训练损失变化曲线



(b) Thchs30 微调损失变化曲线

图6 Thchs30实验损失值曲线图

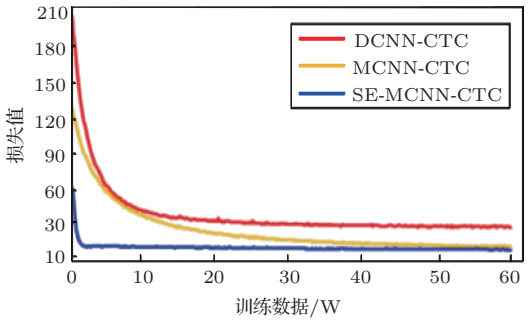
Fig. 6 The loss curve of Thchs30 dataset

综上所述：(1) MCNN(7)-CTC、MCNN(8)-CTC以及MCNN(9)-CTC相较于DCNN(7)-CTC、DCNN(8)-CTC、DCNN(9)-CTC音节错误率分别相对降低12.08%、3.16%以及1.89%，由此可以得出MCNN-CTC声学模型相较于DCNN-CTC声学模型效果更佳；(2) SE-MCNN(7)-CTC声学模型参数仅相对增加1.04%，最终识别结果相较于DCNN(7)-CTC错误率相对降低13.51%，融入SENet模型的声学模型识别效果更强。

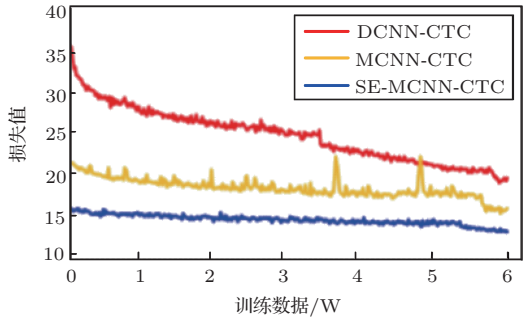
3.4.2 ST-CMDS 实验结果分析

为验证MCNN-CTC以及SE-MCNN-CTC声学模型的泛化能力，选取Thchs30数据集中DCNN(7)-CTC、MCNN(7)-CTC以及SE-MCNN(7)-CTC三个不同的声学模型在ST-CMDS数据集上进一步实验。图7(a)和图7(b)分别表示声学模型训练与微调的损失值变化曲线。可以看

出，随着训练数据量的增加，声学模型逐渐趋于收敛，最终损失值减小到固定的范围内。DCNN-CTC 损失值减小至 19 左右，MCNN-CTC 可降至 16 附近，最终改进后的 SE-MCNN-CTC 可以减小至 14 以下，证明 SE-MCNN-CTC 较传统的 DCNN-CTC 及 MCNN-CTC 能更好地训练深层模型以及上述模型能够训练声学模型的有效性。



(a) ST-CMDS训练损失变化曲线



(b) ST-CMDS微调损失变化曲线

图 7　ST-CMDS 实验损失值曲线图

Fig. 7　The loss curve of ST-CMDS dataset

由表 4 可得出，MCNN-CTC 相较于 DCNN-CTC 参数量得到了极大的降低，可训练参数相对降低 13.60%，在验证集和测试集错误率分别相对降低 3.94%、3.49%；SE-MCNN-CTC 相较于 DCNN-CTC 在验证集和测试集错误率分别有 4.11%、6.68% 的相对降低，错误率最低。

表 4　ST-CMDS 数据集的实验结果

Table 4　The experimental results of ST-CMDS dataset

声学模型	参数数量	验证集错误率/%	测试集错误率/%
DCNN(7)-CTC	7800110	23.86	23.80
MCNN(7)-CTC	6738014	22.92	22.97
SE-MCNN(7)-CTC	6767342	22.88	22.21

4　结论

本文提出了 MCNN-CTC 和 SE-MCNN-CTC 两种端到端声学模型，并且通过大量的实验验证了声学模型的错误率以及泛化性能，得出结论如下：

(1) 以音节为建模单元构建了 DCNN-CTC 声学模型，验证了其对于声学建模的优越性；

(2) 提出了 MCNN-CTC 声学模型，不但在识别错误率相较于 DCNN-CTC 声学模型取得了较大的降低，而且具有较强的泛化性能；

(3) 融合了 SENet 与 MCNN 模型，提出了 SE-MCNN-CTC 声学模型，通过特征通道自适应重标定既减小特征冗余的影响，又实现了声学模型性能的进一步提升。

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] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]// InterSpeech. Canada, 2013: 6645–6649.

[3] Seltzer M L, Ju Y C, Tashev I, et al. In-car media search[J]. IEEE Signal Processing Magazine, 2011, 28(4): 50–60.

[4] 李云红, 梁思程, 贾凯莉, 等. 一种改进的 DNN-HMM 的语音识别方法 [J]. 应用声学, 2019, 38(3): 371–377.
Li Yunhong, Liang Sicheng, Jia Kaili, et al. An improved DNN-HMM speech recognition method[J]. Journal of Applied Acoustics, 2019, 38(3): 371–377.

[5] Parinia B, Albert Z, Ralf S, et al. On using 2D sequence-to-sequence models for speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, 2019: 5671–5675.

[6] 余栋, 邓力. 解析深度学习: 语音识别实践 [M]. 余凯, 钱彦旻, 译. 第 5 版. 北京: 电子工业出版社, 2017: 78–89.

[7] Li J, Yu D, Huang J, et al. Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM[C]// Spoken Language Technology Workshop. Miami, 2013: 131–136.

[8] Peddinti V, Wang Y, Povey D, et al. Low latency acoustic modeling using temporal convolution and LSTMS[J]. IEEE Signal Processing Letters, 2018, 25(3): 373–377.

[9] Wang P, Li J, Xu B. Applying connectionist temporal classification objective function to Chinese mandarin speech recognition[C]// International Symposium on Chinese Spoken Language Processing. Tianjin, 2016: 1–5.

[10] Graves A, Fernández S, Gomez F. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// International Conference on Machine Learning. Pittsburgh, 2006: 369–376.

[11] Graves A. Sequence transduction with recurrent neural networks[J]. Computer Science, 2012, 58(3): 235–242.

[12] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[J]. arXiv: 1609.06773, 2017.

[13] 于重重, 陈运兵, 孙沁瑶, 等. 基于动态 BLSTM 和 CTC 的濒危语言语音识别研究 [J]. 计算机应用研究, 2019, 36(11): 3334–3337.
Yu Chongchong, Chen Yunbing, Sun Qinyao, et al. Research on endangered language speech recognition based on dynamic BLSTM and CTC[J]. Application Research of Computers, 2019, 36(11): 3334–3337.

[14] 姚煜, Ryad Chellali. 基于双向长短时记忆 -联结时序分类和加权有限状态转换器的端到端中文语音识别系统 [J]. 计算机应用, 2018, 38(9): 2495–2499.
Yao Yu, Ryad C. End-to-end Chinese speech recognition system based on bidirectional long-term memory-join timing classification and weighted finaite-state transducer[J]. Journal of Computer Applications, 2018, 38(9): 2495–2499.

[15] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述 [J]. 计算机学报, 2017, 40(6): 1229–1251.
Zhou Feiyan, Jin Linpeng, Dong Jun. A review of convolutional neural networks[J]. Chinese Journal of Computers, 2017, 40(6): 1229–1251.

[16] Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[17] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Computer Vision and Pattern Recognition, Las Vegas, 2016: 770–778.

[18] Huang G, Liu Z, Laurens V D M, et al. Densely connected convolutional networks[C]// Computer Vision and Pattern Recognition. Hawaii, 2017: 2261–2269.

[19] 王珂, 武军, 周天相, 等. 一种融合全局时空特征的 CNNs 动作识别方法 [J]. 华中科技大学学报 (自然科学版), 2018, 46(12): 36–41.
Wang Ke, Wu Jun, Zhou Tianxiang, et al. A CNNs motion recognition method based on global spatiotemporal features[J]. Journal of Huazhong University of Science and Technology, 2018, 46(12): 36–41.

[20] Abdel H O, Mohamed A R, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, 2012: 4277–4280.

[21] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, 2013: 8614–8618.

[22] Zhang Y, Pezeshki M, Brakel P, et al. Towards end-to-end speech recognition with deep convolutional neural networks[J]. arXiv: 1701.02720, 2017.

[23] Hu J, Li S, Samuel A, et al. Squeeze-and-excitation networks[J]. arXiv: 1709.01507, 2018.

[24] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用 [J]. 计算机学报, 2019, 42(3): 453–482.
Zhang Shun, Gong Yihong, Wang Jinjun. Development of deep convolutional neural networks and its application in computer vision[J]. Chinese Journal of Computers, 2019, 42(3): 453–482.

[25] 吴仁彪, 赵婷, 屈景怡. 基于深度 SE-DenseNet 的航班延误预测模型 [J]. 电子与信息学报, 2019, 41(6): 1510–1517.
Wu Renbiao, Zhao Ting, Qu Jingyi. Flight delay prediction model based on deep SE-DenseNet[J]. Journal of Electronics and Information Technology, 2019, 41(6): 1510–1517.

[26] 仇利克, 郭忠文, 刘青, 等. 基于冗余分析的特征选择算法 [J]. 北京邮电大学学报, 2017, 40(1): 36–41.
Qiu Like, Guo Zhongwen, Liu Qing, et al. Feature selection algorithm based on redundancy analysis[J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40(1): 36–41.

[27] Wang D, Zhang X. THCHS-30: a free chinese speech corpus[J]. arXiv: 1512.01882, 2015.

[28] Li Jie, Zhang Heng, Cai Xinyuan, et al. Towards end-to-end speech recognition for Chinese mandarin using long short-term memory recurrent neural networks[C]// InterSpeech. Dresden, 2015: 615–3619.

[29] Kingma D, Ba J. Adam: a method for stochastic optimization[J]. arXiv: 1412.6980, 2015.

[30] Sergey I, Christian S. Batch normalization: accelerating deep network training by reducing internal covariate shift[J]. arXiv: 1502.03167, 2015.

[31] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

[32] Tan T, Qian Y, Zhou Y, et al. Adaptive very deep convolutional residual network for noise robust speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(8): 1393–1405.

[33] 杨洋, 汪毓铎. 基于改进卷积神经网络算法的语音识别 [J]. 应用声学, 2018, 37(6): 940–946.
Yang Yang, Wang Yuduo. Speech recognition based on improved convolutional neural network[J]. Journal of Applied Acoustics, 2018, 37(6): 940–946.

[34] 张立民, 王彦哲, 张兵强, 等. 基于 CTC 准则的普通话识别及改进 [J]. 计算机工程, 2019,45(6): 249–253, 266.
Zhang Limin, Wang Yanzhe, Zhang Bingqiang, et al. Mandarin recognition and improvement based on CTC criterion[J]. Computer Engineering, 2019, 45(6): 249–253, 266.