第 39 卷 第 3 期
2020 年 5 月

应用声学
Journal of Applied Acoustics

Vol. 39, No. 3
May, 2020

◇ 研究报告 ◇

基于双向循环神经网络的汉语语音识别*

李 鹏 杨元维† 高贤君 杜李慧 周 意 蒋梦月 张净波

(长江大学地球科学学院 武汉 430100)

摘要: 当前基于深度神经网络模型中,虽然其隐含层可设置多层,对复杂问题适应能力强,但每层之间的节点连接是相互独立的,这种结构特性导致了在语音序列中无法利用上下文相关信息来提高识别效果,而传统的循环神经网络虽然做出了改进,但是只能对上文信息进行利用。针对以上问题,该文采用可以同时利用语音序列中上下文相关信息的双向循环神经网络模型与深度神经网络模型相结合,并应用于语音识别。构建具有5层隐含层的模型,其中第3层为双向循环神经网络结构,其他层采用深度神经网络结构。实验结果表明:加入了双向循环神经网络结构的模型与其他模型相比,较好地提高了识别正确率;噪声对双向循环神经网络汉语识别有重要影响,尤其是训练集和测试集附加噪声类型不同时,单一的含噪声语音的训练模型无法适应不同噪声类型的语音识别;调整神经网络模型中隐含层神经元数量后,识别正确率并不是一直随着隐含层中神经元数量的增加而增加,神经元数量数目增加到一定程度后正确率出现了降低的趋势。

关键词: 语音识别;深度学习;深度神经网络;循环神经网络

中图法分类号: TN912.3 **文献标识码:** A **文章编号:** 1000-310X(2020)03-0464-08

DOI: 10.11684/j.issn.1000-310X.2020.03.020

A study of Chinese speech recognition based on bidirectional recurrent neural network

LI Peng YANG Yuanwei GAO Xianjun DU Lihui ZHOU Yi

JANG Mengyue ZHANG Jingbo

(*College of Geosciences, Yangtze University, Wuhan* 430100, *China*)

Abstract: Within deep neural network (DNN) models, the hidden layer can be set up multi-level, adaptable to complicated problem, but the node connected between each layer is independent of each other, the structure characteristics make it impossible to use contextual information in the speech sequence to improve the effect of recognition, and while a traditional recurrent neural network (RNN) has made the improvement, but only to use the above information. To solve the above problems, the bidirectional RNN (Bi-RNN) model and DNN model were combined in this paper, which can simultaneously utilize the context-related information in speech sequences, and apply them to speech recognition. A model with five hidden layers was constructed, in which the third layer was Bi-RNN structure and the other layers were DNN structure. The experimental results show that: compared with other models, the model with Bi-RNN structure improves the recognition accuracy. Noise plays an important role in Bi-RNN Chinese language recognition. In particular, the training set and test set have different types of additional noise. After adjusting the number of neurons in the hidden layer in the neural network model, the recognition accuracy does not always increase with the increase of the number of neurons in the hidden layer, but decreases after the number of neurons increases to a certain extent.

Keywords: Speech recognition; Deep learning; Deep neural network; Recurrent neural network

2019-03-19 收稿; 2019-11-28 定稿

*湖北省教育厅科学研究计划资助项目 (Q20181317), 长江大学大学生创新创业基金项目 (2018012), 地理国情监测国家测绘地理信息局重点实验室开发基金项目 (2017NGCM07)

作者简介: 李鹏 (1997–), 男, 山西晋城人, 本科在读, 研究方向: 语音识别。

†通信作者 E-mail: yyw_08@whu.edu.cn

0 引言

语音识别是指计算机能够理解人的语言，将音频信息转换成文本信息。随着互联网技术和人工智能技术的飞速发展，语音识别被逐渐应用到各个领域内，因此与之相关的研究也越来越受到重视。特别地，Google、Microsoft、科大讯飞、百度等公司，都争相在语音识别上投入大规模的研发，推出相关的算法、软件及应用。语音识别的产业化也进一步推动着语音识别技术的发展。

语音识别的相关研究最早可以追溯至20世纪50年代AT&T贝尔研究室。该研究室的Audry系统基于简单的孤立词，能够对10个单音节单词进行识别。在60年代提出的动态时间规整(Dynamic time warping, DTW)方法[1]，有效解决了两个不同长度音频片段的对齐问题。随后语音识别研究进一步发展，线性预测分析技术(Linear predictive coding, LPC)被扩展应用[2]，DTW也基本成熟。与此同时，隐马尔科夫模型(Hidden Markov model, HMM)理论被提出。随着HMM技术不断成熟和完善，语音识别从原来的模板匹配的方法转变为概率模型的方法[3]，并且以HMM相关模型为主要研究方法[4]。而后，人工神经网络(Artificial neural net, ANN)逐渐被用于语音识别的研究中[5]，以寻求新的突破。杨华民等[6]采用ANN进行语音识别的原理，给出了求解语音特征参数和典型神经网络的学习过程，通过具体的实例展示了ANN技术的实用化。但传统神经网络本身也存在需要大量标记数据等问题。2006年，Hinton等[7]提出了深度学习的概念。此后，深度学习以其良好的普适性被应用到语音识别领域里，打破了HMM的主导局面，极大地提升了基于传统神经网络的语音识别系统的性能，突破了某些应用情景中的识别瓶颈[8]。

在深度学习的大环境下，最初应用在语音识别里的是深度置信网络(Deep belief network, DBN)[9]，能够对神经网络进行预训练以达到使模型稳定的效果。而后深度神经网络(Deep neural network, DNN)、卷积神经网络(Convolution neural network, CNN)和循环神经网络(Recurrent neural network, RNN)等相继问世，这引发了人们对各类神经网络进行深入研究。张仕良[10]指出基于DNN的训练速度相较于CNN 或RNN的更快，然而利用DNN进行语音识别却未能良好解决其中较为重要的时序问题。DNN 和CNN对输入的音频信号的感受视野相对固定，所以对于与时序相关的问题不具有较好的处理能力。RNN在隐含层存在反馈连接，它能通过递归来挖掘序列中上文的相关信息，在一定程度上克服DNN和CNN的缺点[11]，但是却无法挖掘序列中下文的相关信息。随后，Schuster等[12]提出双向循环神经网络(Bidirectional RNN, Bi-RNN)，并弥补了RNN的缺点，能够同时利用上下文信息，在时序问题上相对于RNN识别正确率取得了进一步的提升。因此本文基于Bi-RNN模型在语音识别方面进行研究，从言语产生与言语感知的角度对Bi-RNN进行更深层次的解读，探讨了Bi-RNN模型在不同噪声环境中的识别效果，并进行大量的实验，选取出一套适合本模型的参数，进一步地降低了语音识别错误率。

在进行语音识别之前，本文首先对音频进行预处理。预处理包括对音频进行预加重、分帧和加窗。对预处理之后的音频做语音特征提取，即将音频转化为梅尔频率倒谱系数(Mel frequency cepstral coefficient, MFCC)。再用训练集迭代训练模型，将训练后的模型对测试集进行实验，最后得到识别结果。

1 循环神经网络结构

1.1 人工神经网络

ANN是一种由大量简单处理单元(神经元)按照不同的连接方式组成的运算模型。一个神经元的模型如图1所示。在结构上可以将人工神经网络划分为3层——输入层、隐含层、输出层(图2)。神经网络的输入/输出关系表示为下列公式：

$$u_i = \sum_{j=1}^{N} w_{ij}\chi_i - \theta_i, \tag{1}$$

$$o_j = f(u_i), \tag{2}$$

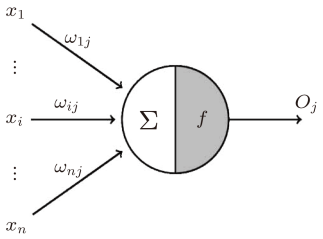

图1　神经元模型

Fig. 1　Neuron model

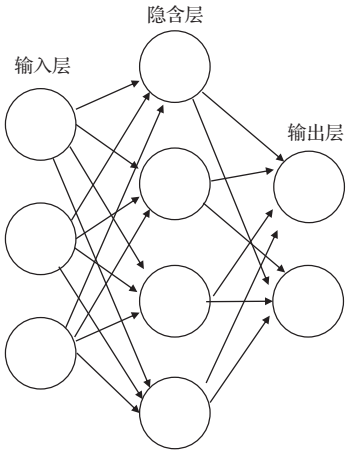

图2 神经元网络

Fig. 2 Neural network

其中：w_{ij} 为连接权重，即神经元 i 与神经元 j 之间的连接强度；χ_j 为神经元 i 的某个状态变量；θ_i 为神经元 i 的阈值；u_i 为神经元 i 的活跃值；o_j 为神经元 i 的一个输出；f 为激活函数。

1.2 单向循环神经网络

在DNN或者CNN中，它们的基本前提是每层之间的节点连接是相互独立的。这样的结构存在一个潜在的弊端，即无法对具有时间特性的相关信息来建立模型。然而语音识别却是一个典型的具有时间特性的问题[13]，输入顺序是一个非常重要的因素，它不类似于图像识别——对输入的顺序无特殊要求。因此为了解决DNN、CNN的这种弊端，对RNN的研究在20世纪80年代迅速开展起来。

相较于DNN或者CNN，RNN最大的不同之处就是在隐含层中增加了节点之间的连接[14−15]，这使得隐含层的输入不仅来源于输入层，还包含了隐含层前一时刻的输出。RNN是根据人的记忆原理而产生的。比如一句话“我要去饭吃了”，这句话听起来很奇怪，这是因为大脑接收到这段话会受到刺激，进而产生预测功能。如果“我要去”后面跟着“吃”，就感觉很正常。从言语产生和言语感知的角度来理解，这是因为大脑对每个字的先后顺序是有一定的判断的。其模型如图3所示。

在RNN中，上一时间点到当前时间点变换过程中每层的权重 W 是共享的，这样在很大程度上减少了训练参数数目。图3中，W_0 表示输入层与隐含层之间的权重值，W_1 表示上一时刻隐含层到当前时刻隐含层之间的权重值，W_2 表示隐含层与输出层之间的权重值；$S_{(t)}$ 表示隐含层的第 t 个RNN节点的输出状态。

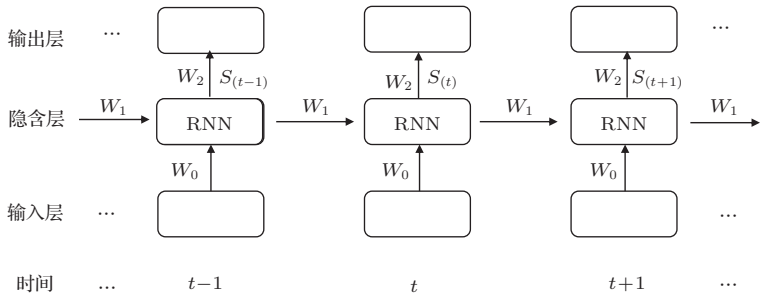

图3 循环神经网络结构

Fig. 3 The structure of RNN

1.3 双向循环神经网络

由1.2节可知，传统的RNN只是利用了上一时刻的信息，而在具有时间特性的语言序列中，有很多需要同时联系过去与未来时刻的信息。同样是这句话“我要去饭吃了”，如果说出“饭”的前面一个字是什么，大脑可能需要时间思考一下，甚至要再默念一遍这句话，而不是反着读这句话“了吃饭去要我”，但最终都会找到这个字。这种现象引发了两个很值得思考的问题：第一，大脑可以通过一定的规则而找到“饭”这个字前面的字，这种现象可以理解为大脑对于信息的存储，并不是简单的单独存储，而是一种链条式的存储方式，这种方法有个极大的好处，大脑只要记住相关的存储规则或者方法就可以，这样大大节省了很多空间。第二，大脑很难进行反方向的搜寻信息。基于这种现象，Bi-RNN应运而生，相对于CNN结构与DNN结构，其最大的特点在于能够将过去与未来的信息作为输入再一次地输入到神经元，这种结构非常适合具有时序性质的数据，但同时也可能需要更长的训练时间。Bi-RNN结构解决了其中较为重要的时序问题，能够对一些有时

间依赖性的数据进行更好的学习，如语音识别、情感分类、文本分类、机器翻译、词向量的生成等，将Bi-RNN展开后，可看出在网络结构中有一部分参数是共享的，这在一定程度上大大减少了所训练的神经网络参数个数，同时也带来了另一个优势——Bi-RNN输入可以是不固定长度的序列。因此基于传统的RNN计算原理，可对结构进行一定程度的改进，推导出Bi-RNN结构。Bi-RNN 可以同时利用过去与未来时刻的信息，将时间序列信息分为前后两个方向，输入到模型里，并构建向前层与向后层用来保存两个方向的信息，同时输出层需要等待向前层与向后层完成更新[16]，才能进行更新。其模型结构如图4所示。

Bi-RNN的整个计算过程与单向循环神经网络类似，即在单向循环神经网络的基础上增加了一层方向相反的隐含层。从输入层到输出层的传播过程中，共有6个共享权值。图4中，W_0表示输入层与向前层之间的权重值，W_1表示上一时刻隐含层到当前时刻隐含层之间的权重值，W_2表示输入层与向后层之间的权重值，W_3表示向前层与输出层之间的权重值，W_4表示下一时刻隐含层到当前时刻隐含层之间的权重值，W_5表示向后层与输出层之间的权重值。Bi-RNN结构向前传播的计算过程如下列公式：

$$S_{(t)} = f(W_0 X_{(t)} + W_1 S_{(t-1)} + b), \tag{3}$$

$$H_{(t)} = f(W_2 X_{(t)} + W_4 H_{(t+1)} + b_1), \tag{4}$$

$$O_{(t)} = g(W_3 S_{(t)} + W_5 H_{(t)}), \tag{5}$$

其中，$X_{(t)}$表示在t时刻的输入，$S_{(t)}$表示向前层的第t个RNN节点的输出，$H_{(t)}$表示向后层的第t个RNN节点的输出，$O_{(t)}$表示在t时刻的输出，b和b_1表示偏置参数，f和g均表示激活函数。相对于传统的RNN而言，Bi-RNN实现了同时利用过去与未来时刻的信息，因此记忆效果比之前更佳。

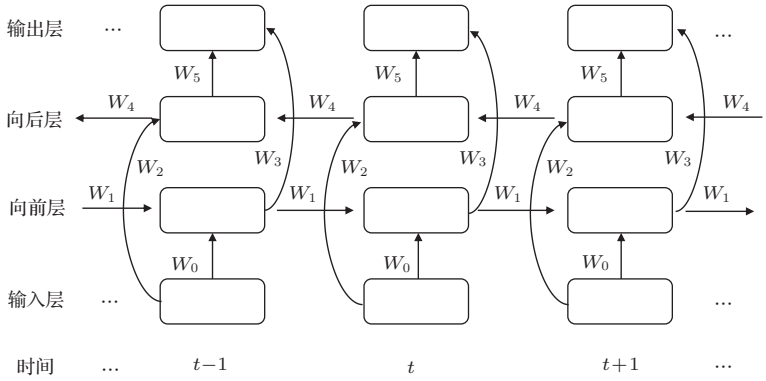

图4　双向循环神经网络结构

Fig. 4　The structure of Bi-RNN

2　汉语识别实验

2.1　实验设计

本文基于tensorflow深度学习平台，使用Anaconda软件中自带的spyder编译器进行编译，并进行仿真实验。共设置了3组实验：

实验1：为了说明Bi-RNN在语音识别上的优越性，分别用DNN模型与Bi-RNN模型对不带噪声的训练集进行实验，并与文献[17]所提出的改进CNN算法进行比较；

实验2：为了测验基于某一个环境训练出的模型在不同背景噪声的音频识别效果，首先根据训练音频类型共设置了3组实验，每组实验下再根据测试音频类型分别设置3个实验；先用Bi-RNN模型对3个训练集分别进行实验，再基于3种训练集所训练出的模型对其他噪声类型的测试集进行实验；

实验3：为了研究隐含层中神经元数量对实验效果的影响，本实验基于Bi-RNN模型，通过调整隐含层神经元个数，设置8组实验，再使用不带噪声的训练集进行实验。实验流程图如图5所示。

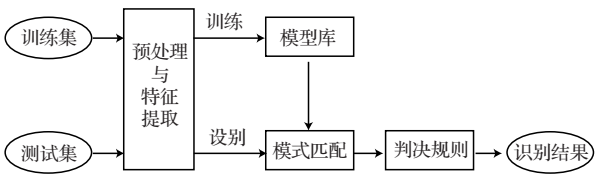

图5　实验流程图

Fig. 5　Flow chart of experiment

2.2 数据集描述

本文采用了两个版本的THCHS-30语料库：第一个是通过单个碳粒扬声器，在安静的办公室环境下录制的无噪声音频；第二个是通过简单的波形混合，在第一个版本的数据加上了白噪声和咖啡馆噪声，噪声和音频的能量相等。THCHS-30的文本是从大容量的新闻选取出1000句，音频总时长超过30 h。参与该语料库录音的人员，大部分是会说流利普通话的大学生。

由于计算机性能的限制，本文没有对整个数据集进行训练。选用句子的发音人数目为22人，包括15名女生和7名男生，每句话在30字左右，其中陈述句居多，约为95%左右。双音素占35%左右，三音素占53%左右，单音素与四音素共占12%左右，双音素与三音素覆盖率较好。本文共建立了3个训练集以及3个相对应的测试集，每个训练集包括2241句话，测试集包括249句话，这3个训练集的差别只是在于带噪声的类型，其他方面设置保持一致，并且训练集与测试集的文字内容是相一致的。

2.3 模型的构建

基于上述Bi-RNN的优点，本文采用Bi-RNN构建模型。在文献[18]中，DNN的性能并不是随着层数增加而增加的，并表明3~5个隐层的DNN结构是合适的。据此本文所构建的模型共包括5层，其中第1层、第2层与第4层都为852个单元的全连接层，激活函数采用ReLU；第3层为852维的双向循环神经网络，为了减小模型产生过拟合现象，在每层后面加一个Dropout层；第5层为全连接层，并采用$(X+1)$个单元的Softmax用于分类，其中X表示字体的个数，1表示空白符号，$X+1$表示字体与空白符号的概率分布。语音识别属于神经网络中的时序类分类，通过联结主义时间分类(Connectionist temporal classification, CTC)来解决输入与输出的序列长度不等的问题。使用ctc_loss方法来计算损失值。模型如图6所示。

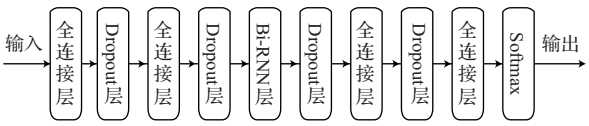

图6 模型结构示意图

Fig. 6 Schematic diagram of model structure

2.4 实验结果与分析

实验1

用上述Bi-RNN模型对无噪声的训练集进行训练，测试集也使用无噪声的音频；同时对DNN与RNN构建模型，并采用相同的方法进行实验，其中DNN的模型结构是将上述Bi-RNN模型的第3层Bi-RNN层换成全连接层。Bi-RNN与DNN实验训练集的损失函数值和正确率分别如图7与图8所示。

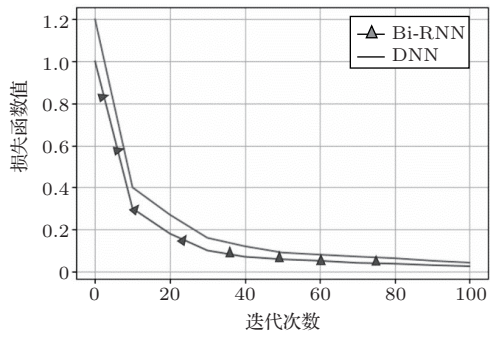

图7 两种不同模型的损失函数

Fig. 7 Loss function of two different models

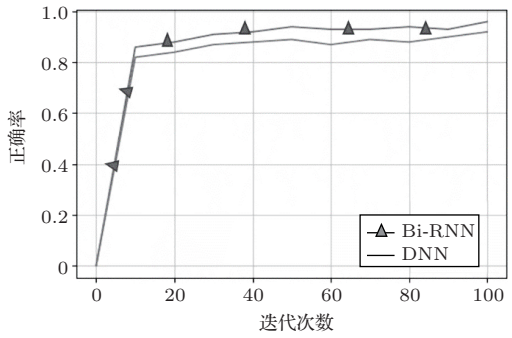

图8 两种不同模型的识别正确率

Fig. 8 Recognition accuracy of two different models

由图7和图8可以看出，Bi-RNN模型的损失函数值下降到稳定的速度最快，且训练集的正确率也高。两种模型的训练集的正确率相差不大，正确率都在93%左右。但测试集的效果显示Bi-RNN模型远强于DNN模型。在用DNN模型进行训练时，其在训练集上的效果很好，但在测试集上错误率大大增加。从数据上表现出DNN模型产生了“过拟合”。

Bi-RNN结构相对于DNN结构更加复杂，Bi-RNN对上下文相关性的拟合较强，理论上Bi-RNN相对于DNN更应该陷入过拟合的问题，而结果显示Bi-RNN的识别错误率更低，因此单纯用“过拟合”来解释是自相矛盾的。通过对DNN的神经元进

行多次调整，当神经元数量到612 时，其错误率最低为53.26%，相比Bi-RNN还是很高，因此并不能简单地通过“过拟合”来解释，说明产生这种现象根本原因在于Bi-RNN与DNN结构的差异性。受到协同发音的影响，语音中的各帧之间有着很强的相关性，每一个字的发音受到前后几个字的影响。在进行输入时，DNN是把相邻的几帧进行拼接，并且其输入窗口是固定的。而Bi-RNN在时序问题上能够更好地体现长时相关性，可以将过去与未来的信息同时输入得到输出结果，以作为预测当前的输入，能够更加深刻地了解其内在联系，因此降低了错误率。本文又与文献[17]所提出的改进CNN算法相比较，错误率也比其提出的方法较低，可见本文的Bi-RNN模型要比文献[17]所提出的改进CNN模型在语音识别方面性能要好。其实验结果如表1所示。

表1　两种模型的实验结果

Table 1　Experimental results of two models

模型类型	识别错误率/%
DNN	54.76
Bi-RNN	19.32
改进CNN	22.19

实验2

在现实生活中，环境因素是动态易变的。为了测试模型在不同环境下的识别效果，首先将Bi-RNN模型在不同类型且带噪音频的、信噪比为0 dB的条件下进行训练再测试，实验结果如表2所示。

表2　基于不同音频训练实验结果

Table 2 Based on the experimental results of different audio training

训练音频类型	测试音频类型	识别错误率/%
加白噪声	加白噪声	27.16
	加咖啡馆噪声	68.75
	净语音	64.33
加咖啡馆噪声	加咖啡馆噪声	24.25
	加白噪声	65.56
	净语音	56.23
净语音	净语音	19.32
	加咖啡馆噪声	53.31
	加白噪声	66.12

由表2可看出，Bi-RNN模型对3种不同环境下的语音库进行训练以及测试。首先通过对表2识别错误率中第1、4、7三个数据的比较，表明训练和测试音频类型相同时带有噪声的音频的错误率要比无噪声的音频错误率要高，其中白噪声的错误率最高，错误率为27.16%，这是因为白噪声和咖啡馆噪声同属于加性噪声，白噪声属于平稳噪声，咖啡馆噪声属于缓变噪声。白噪声是明确定义的，因为其宽带与均匀连续特点，噪声信号与语音信号重合度很大，导致了对语音识别影响很大，其语谱图如图9所示。咖啡馆噪声的频谱分析虽和语音类似，而噪声信号与语音信号重合度相对较小，对语音识别影响相对较小，其语谱图如图10所示。通过与纯净语音语谱图(图11)进行比较，可以看出白噪声共振峰轨迹的干扰要比咖啡馆噪声大，因此白噪声的识别错误率更高。然后通过对每组内的3个实验进行比较时，即当训练音频与测试音频的类型不同时，其识别错误率大大增加，这是因为用于训练音频的背景噪声与测试语音的背景噪声不一致，训练环境与识别环境有着巨大的差异，最终导致了识别语音特征与模板特征之间的失配，系统的性能大大降低。

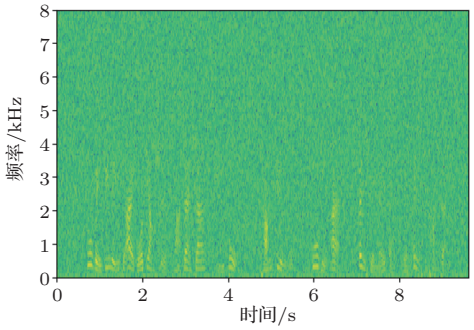

图9　加白噪声的音频语谱图

Fig. 9　Audio spectrum with white noise

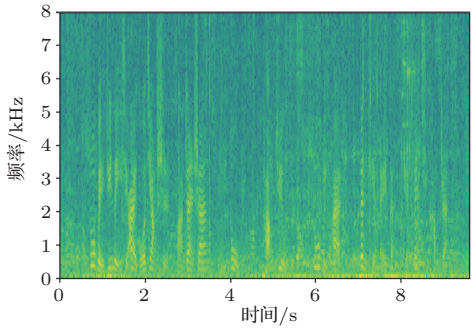

图10　加咖啡馆噪声的音频语谱图

Fig. 10　Audio spectrum with cafe noise

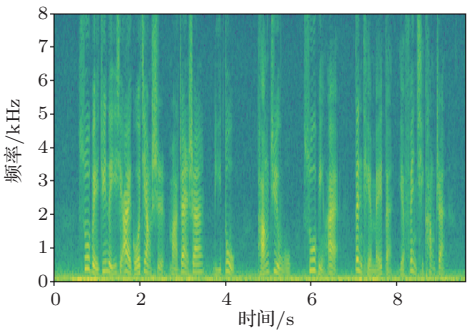

图 11　纯净音频语谱图

Fig. 11　Pure audio spectrum

实验 3

为了研究隐含层中神经元数量对实验效果的影响，采用 Bi-RNN 模型，通过对隐含层神经元个数调整，进行识别。

实验结果如表 3 所示，当神经元数量增加到 512 时，识别错误率大幅减少，这是因为隐含层节点数量过少，导致网络的学习与处理能力较差；而当神经元数量大于 512 时，识别错误率的减少程度较缓，说明了神经元的数量将趋于饱和状态；当神经元数量大于等于 1024 时，错误率出现增加趋势，说明再增加神经元数量，就会出现在训练集上有很好的识别效果，但是在测试集上的识别效果变差的现象，即出现过拟合现象。

表 3　不同神经元数量的实验结果

Table 3　Experimental results for different numbers of neurons

隐含层每层神经元数量	识别错误率/%
64	54.53
128	47.91
256	39.85
512	19.32
682	18.24
852	18.92
1024	21.73
2048	29.67

从这 3 个实验可看出，Bi-RNN 相对于 DNN 在语音识别方面效果更加良好，两个模型在无噪声的训练集上效果相差不大。但是在测试集上，DNN 模型错误率在 54.76%，文献 [17] 所提出的改进 CNN 错误率在 22.19%，而 Bi-RNN 模型错误率为 19.32%，相对于 DNN 模型与改进的 CNN 模型都有了降低。由此可以看出，Bi-RNN 可同时利用上下文信息，发挥出其独特的优势。当使用 Bi-RNN 模型对 3 种不同类型的音频进行实验时，在无噪声的测试集上错误率为 19.32%，在带咖啡馆噪声的测试集的错误率为 24.25%，在带白噪声的测试集的错误率为 27.16%，在无噪声的音频条件下实验效果最好；当采用基于某一语音库所训练的模型对其他两个环境下的音频进行测验时，效果很差，说明采用单个训练集训练的模型无法适应不同噪声类型的音频，在以后的研究中将考虑联合训练。在探索隐含层的神经元数量对识别效果的实验中，当隐含层每层神经元数量在 682~852 时，效果最好。同时，识别错误率并不是随着隐含层每层神经元的增加而降低，甚至当神经元个数增加到一定程度时，识别错误率不下降反而上升。

3　结论

自深度学习的概念提出后，深度学习在语音识别方面相较于传统的方法，如混合高斯模型，在性能有了很大的提升。其中基于 Bi-RNN 模型在语音识别方面更是具其独特的优势。本文使用 Bi-RNN 进行语音方面了探索，并与 DNN 和改进的 CNN 进行比较，初步验证了 Bi-RNN 在语音识别方面的独特优势。同时对含有噪声的音频的识别效果进行测试，以及隐含层神经元数量对识别效果的影响方面，做了初步的探索。结果如下：(1) 在汉语语音识别中采用 Bi-RNN 模型得到了在同样条件下高于 DNN 和改进的 CNN 的识别率，成功地构建了一个汉语识别模型；(2) 初步考察了噪声对 Bi-RNN 汉语识别模型的影响，分析了白噪声的影响大于咖啡馆噪声的原因；(3) 研究了 Bi-RNN 汉语识别模型中隐含层中神经元数量对识别率的影响，提出了该模型中核心层神经元数量为 682~852 的最优设计。

本文由于一些软件与硬件资源上的限制，有许多问题还需要进一步的探索。主要有：

(1) 在进行探讨隐含层神经元的数量对识别效果的实验中，只是提出了神经元数量并不是越多越好，但是对不同结构的神经网络结构神经元数量的合理设定的范围，并未给出结果，需要进一步的探索。

(2) 在本文中使用DNN与Bi-RNN相结合用以构建模型。在使用DNN时，由于参数太多，易出现过拟合现象，为了更好地解决这一问题，在接下来的学习与探索中，将CNN与Bi-RNN相结合来构建模型，并进行实验。

参 考 文 献

[1] Vensko G, Lieu K B, Meloche S A, et al. Dynamic time warping (DTW) apparatus for use in speech recognition systems: US, 5073939[P]. 1991–12–17.

[2] Itakura F. Mnimum prediction residual principle applied to speech recognition[J]. IEEE Transactions on Acoustics, Speech & Signal Processing, 1975, 23(1): 67–72.

[3] 赵力. 语音信号处理 [M]. 北京: 机械工业出版社, 2006.

[4] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257–286.

[5] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1989, 3(3): 328–339.

[6] 杨华民, 姜会林, 李平. 基于神经网络的语音识别技术应用研究 [J]. 电子技术应用, 1997(9): 7–9.

[7] Hinton G E, Osindero S, Teh Y. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(3): 1527–1554.

[8] 戴礼荣, 张仕良, 黄智颖. 基于深度学习的语音识别技术现状与展望 [J]. 数据采集与处理, 2017, 32(7): 221–231.
Dai Lirong, Zhang Shiliang, Huang Zhiying. Deep learning for speech recognition: review of state-of-the-arts technologies and prospects[J]. Journal of Data Acquisition and Processing, 2017, 32(7): 221–231.

[9] Deng L, Yu D, Dahl G E. Deep belief network for large vocabulary continuous speech recognition: US, 8972253[P]. 2015–03–03.

[10] 张仕良. 基于深度神经网络的语音识别模型研究 [D]. 合肥: 中国科学技术大学, 2017.

[11] 邢吉亮. 结合注意力机制的Bi-LSTM循环神经网络对关系分类的研究 [D]. 长春: 吉林大学, 2018.

[12] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673–2681.

[13] 石颖. 基于循环神经网络的语音识别方案的优化与设计 [D]. 北京: 北京交通大学, 2017.

[14] 汪优升. 基于深度学习的语音识别及其交互应用研究 [D]. 长沙: 湖南大学, 2017.

[15] 梁静. 基于深度学习的语音识别研究 [D]. 北京: 北京邮电大学, 2014.

[16] 黄积杨. 基于双向LSTMN神经网络的中文分词研究分析 [D]. 南京: 南京大学, 2016.

[17] 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 algorithm[J]. Journal of Applied Acoustics, 2018, 37(6): 940–946.

[18] Maas A L, Qi P, Xie Z, et al. Building DNN acoustic models for large vocabulary speech recognition[J]. Computer Speech & Language, 2015, 41: 195–213.