浙江理工大学学报，2019，41(2)：217-223
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(n).2019.02.012

基于深度学习的语音识别模型及其在智能家居中的应用

包晓安[1]，徐　海[1]，张　娜[1]，吴　彪[2]，钱俊彦[3]
(1.浙江理工大学信息学院，杭州 310018；2.山口大学东亚研究科，日本山口 753-8514；
3.桂林电子科技大学计算机科学与工程学院，广西 桂林 541004)

摘　要：为了满足人们对智能家居设备控制便捷化的需求，提出了一种基于降噪自动编码器的深度学习语音识别模型，经过语音识别模型解析出短语控制指令，以实现家居设备控制。该语音识别模型主要包含两个部分：首先进行无监督学习预训练，预训练前随机将一些网络节点置为0，人工模拟噪声数据，然后采用限制玻尔兹曼机权重矩阵依次训练每一个隐含层，通过比较输入数据与输出数据的偏差修改权重，优化参数；然后进行有监督微调，把训练好的参数作为整个网络的初始值，采用误差反向传播算法对整个网络模型调参。实验结果表明：该语音识别模型与深度信念网络对比，在语音识别率和对噪声的鲁棒性都有明显提高。将该语音识别模型和智能家居系统相结合，从普通短语中判断出家居控制指令，实现人机交互非接触式、便捷式控制，从而使系统更加智能化。

关键词：深度学习；语音识别；降噪自动编码器；智能家居
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2019) 03-0217-07

Speech recognition model based on deep learning and its application in smart home

BAO Xiaoan[1], *XU Hai*[1], *ZHANG Na*[1], *WU Biao*[2], *QIAN Junyan*[3]
(1. School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2.Graduate School of East Asian Studies, Yamaguchi University, Yamaguchi-shi 753-8514, Japan; 3.School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: In order to meet the needs of people to control smart home equipment conveniently, a deep learning speech recognition model based on denoising autoencoder was proposed. Through the speech recognition model, the phrase control instruction was parsed to achieve the purpose of home equipment control. The speech recognition model mainly consists of two parts. The first part is unsupervised learning pre－training. Before the unsupervised pre－training, some network nodes were randomly set to 0; the noise data were artificially simulated; then each hidden layer was trained sequentially by using the Boltzmann machine weight matrix. The weight was modified and the parameters were optimized through comparing the deviation between input data and output data. Then, supervised fine adjustment was conducted. The well－trained parameters served as the initial values of the whole network, and error back propagation algorithm was adopted to adjust parameters of the whole network model. The experimental

收稿日期：2018－10－22　　**网络出版日期**：2018－12－28
基金项目：国家自然科学基金项目(61502430，61562015)；广西自然科学重点基金项目(2015GXNSFDA139038)；浙江理工大学521人才培养计划项目
作者简介：包晓安(1973－　)，男，浙江东阳人，教授，硕士，主要从事软件测试、智能信息处理方面的研究。
通信作者：张　娜，E-mail：zhangna@zstu.edu.cn

results showed that speech recognition rate and noise robustness of the speech recognition model improve significantly, compared with deep belief network. The speech recognition model could be combined with smart home system to judge home control command from the common phrase and achieve human－computer interaction non-contact and convenient control so that the system is more intelligent.

Key words: deep learning; speech recognition; denoising autoencoder; smart home

0 引 言

随着信息技术的不断发展,智能家居设备逐渐走进了人们的生活。智能家居通过物联网技术集成智能家电、安防控制等设备,为用户营造一种舒适便捷的家居环境。目前智能家居设备多以手机为主要控制终端,以“Wi-Fi”+4G互相配合的控制模式,完成对智能家居设备的单一化控制。为实现智能家居设备的多样化控制,满足用户对家居设备控制方便快捷的更高需求,将智能家居和语音识别技术结合已经成为必然的趋势。

目前智能家居设备和语音识别技术相结合已经出现了许多研究成果。王爱芸[1]设计了基于NL6621嵌入式语音识别的智能家居平台,在实验室环境中语音识别率达到了90%,但该平台采用基于隐马尔科夫模型(Hidden Markov model,HMM)的算法进行声学模型训练和语音识别,相对于深度学习神经网络模型,需要设置更多的参数。闵梁[2]提出了基于动态时间规整算法(Dynamic time warping,DWT)的非特定人孤立词语音识别系统,该系统不能进行实时的语音识别操作,没有体现出智能家居系统中语音实时控制家居的便捷性。刘文强[3]运用基于半监督学习的分类识别方法对智能家居系统进行总体设计,但是该系统主要针对的是面向特定人的短语识别,不利于该智能家居语音识别系统的推广。戴礼荣等[4]采用前馈序列记忆神经网络(Feedforward sequential memory networks,FSMN)进行语音识别声学建模和语言模型建模,显著提升了语音识别系统的性能和训练效率,但是在远场以及噪声比较强的情况下,语音识别系统的性能依然不够理想。居治华等[5]提出了基于反向卷积的双向长短时记忆(Bidirectional long short term memory,Bi-LSTM)网络,加速了网络计算过程,节省了模型训练时间,但是该模型仅经过较小数据量的训练和测试,还有许多理论和应用问题需要继续研究。杨洋等[6]改进了卷积神经网络(Convolutional neural network,CNN)算法,采用新型log激活函数,有效提高了CNN的语音识别性能,缓解了语音识别时出现的过拟合问题,但是该模型识别结果和其它深度学习语音识别模型相比词错率偏高。

传统的人工神经网络(Artificial neural networks,ANN)在进行语音信号处理方面存在易陷入局部最优、需要大量标签样本等问题,本文针对这些不足提出了一种基于降噪自动编码器(Denoising autoencoder,DA)的深度学习声学特征提取模型。该模型采用降噪自动编码器,以非监督方式进行每个神经网络层的特征提取和训练,在特征训练时对特定信息更加敏感,比如特定的家居语音控制指令,并且在每一层之间利用反向传播算法,通过标签样本采用有监督的训练方式对整个模型进行微调,提高复杂环境下语音识别对噪声的鲁棒性,提高特定语音控制指令的处理能力以及最终的语音识别结果的识别率。该系统将基于深度学习的语音识别技术与智能家居结合,能够有效解决“鸡尾酒会”问题,满足了人们对于智能家居便捷化的要求[7-9]。

1 语音识别流程

语音识别是计算机能够自动地将人类语言内容转化为相应的文字或控制指令,达到人机交互目的的技术。随着语音识别的不断发展,从20世纪60年代初贝尔实验室的Audrey语音识别系统到现在的语音技术本、语音智能玩具[10],人们的生活变得越来越方便。语音识别的方法基本可以概括为三种:基于语音学和声学、模板匹配和神经网络的方法[11]。第一种方法虽然起步较早,但是由于其模型较复杂,还没有到实用的阶段。第二种方法中的隐马尔科夫模型的出现,能很好地描述语音信号的整体非平稳性和局部平稳性,使语音识别有了质的飞跃。对于神经网络,使用浅层神经网络学习训练易造成梯度不稳定,且样本特征由人工抽取,费时费力,识别效果一直不好。对比之前的语音识别方法,采用深度神经网络的方法识别错误率大大降低,彻底改变了语音识别模型框架。深度神经网络类似人的大脑,模仿人类的

思考方式，当面对大量感知数据时，通过低维特征的组合形成更加抽象的高维特征，与浅层神经网络相比，提取到的特征是网络自动完成的而不是随机初始化，不需要人工参与。

语音识别根据是否针对指定发音人分为特定人和非特定人的识别，根据发音方式分为孤立词和连续词识别等，但不管哪一种语音识别系统，其系统流程的内部原理都是相似的。语音识别流程主要包括语音信号的预处理、特征提取、模式匹配几个部分[12]。本文提出的用于智能家居的语音识别模型是针对非特定人的连续少词汇量识别。该语音识别模型对智能家居语音控制指令的特性更加敏感，类似于人类听觉系统神经单元，对特定的语音比较敏感，能够在声音信息复杂或是在噪声干扰的情况下解析出家居控制指令。语音识别流程如图 1 所示，首先对语音信号预处理以消除人类发声器官和语音采集设备对语音信号的影响，然后根据提取到的语音特征建立语音识别声学模型并进行模式匹配，最后由相应的语音识别决策输出识别结果。

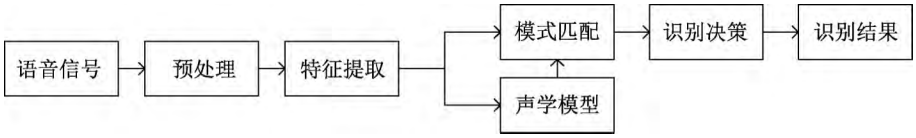

图 1　语音识别流程

2　深度学习语音识别模型

2.1　典型自动编码器神经网络

典型自动编码器神经网络是一种无监督的机器学习算法，只有一层隐含层，具有输入和输出相同节点数的神经网络，使用反向传播算法，目的是求函数 $h_{w,b}(x)\approx x$（x 表示输入向量，w 表示权重，b 表示偏置），使自动编码器的输出能够代表输入数据的核心特征，降低数据的维度，使神经网络的输出与原始输入误差尽量减少，这就要求神经网络需要学习到原来输入样本的特征，找到可以代表原信息的主要成分。

典型自动编码器网络结构如图 2 所示，其中 x_1、x_2 等表示网络层节点，+1 表示偏置节点。自动编码器除去偏置节点，输入和输出节点数相同，每一层的节点激励函数采用 sigmoid，使用反向传播算法进行训练，对矩阵 B^1 编码得到数据的压缩特征，达到降维的目的，先将隐含层 L_1 进行压缩提取有用特征到隐含层 L_2，然后将隐含层 L_2 解压为 L_3，使输入输出节点数相同。

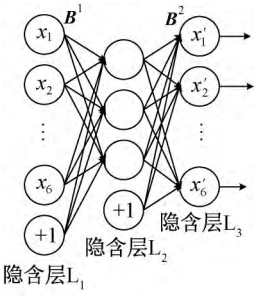

图 2　典型自动编码器网络结构

2.2　限制玻尔兹曼机

由于深度学习模型比较复杂，相邻两层之间的建模本文采用限制玻尔兹曼机（Restricted Boltzmann machine，RBM）模型[13]。RBM 是只有两层的浅层神经网络，第一层称为可见层，也叫输入层，第二层称为隐含层。输入层各数据节点间是没有关联的，输出层的各节点之间也是没有关联的，节点值为 0 或 1，并且所有节点满足波尔曼兹分布，用 $p(v|n)$ 表示，其中：v 表示可见单元，n 表示隐藏单元。在该模型中，可以根据输入节点通过 $p(v|n)$ 获得输出节点，同样，已知输出节点通过 $p(v|n)$ 获得输入节点。因此，由输入层数据 v 推出输出层数据，再根据输出层数据倒推出 v'，将 v 和 v' 比较进行对数似然函数计算，从而调整 $p(v|n)$ 参数，获取最优表达式。限制玻尔兹曼机数据处理如图 3 所示。

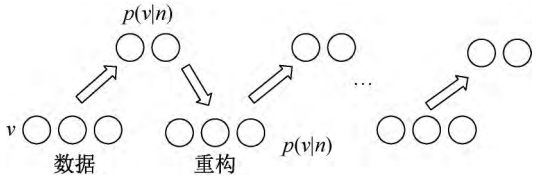

图 3　限制玻尔兹曼机数据处理示意图

2.3　误差反向传播算法

误差反向传播算法（Backpropagation algorithm，BP）是传统网络算法中的一种，它的提出解决了多层神经网络隐含层连接权重学习的问题，由一个输入层，一个输出层和一个或多个隐含层构成，激活函数是 sigmoid 函数，采用有监督的特征训练方式，在本质上是将网络误差平方和作为目标函

数,采用梯度下降算法来计算目标函数的最小值。输入特征在各隐含层中训练,在输出层得到输出数据,将输出数据与期望数据进行比较并计算偏差,然后再对偏差值从输出层经过各个隐含层,在输入层输出进行偏差的各隐含层分摊,在这个过程中利用梯度下降算法对神经元权值进行调整,从而达到优化特征训练的目的。sigmoid 激活函数如式(1)所示:

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{1}$$

2.4 语音识别模型

典型自动编码器的隐含层只有一层,为了有效提取到原始数据在隐含层的表示形式和增强学习能力,本文采用包含有多隐含层的深度自动编码器。在深度自动编码器中,采用 RBM 对每一层的输入数据编码后再进行解码,与原始输入数据对比进行收敛计算,从而得到每一层的最优参数调整。经过这个步骤完成一层的特征训练,并把该层编码后的数据作为下一层的输入数据,采取相同的方式进行数据处理,如此完成所有层的特征训练。

由于自动编码器提取到的语音特征不能保证很好的噪声鲁棒性,自动编码器的变体结构降噪自动编码器,通过语音样本的充分训练,对特定的智能家居语音控制指令敏感性提高,极大地增加了语音特征提取的准确性。降噪自动编码器是在自动编码器的基础上,使用二项分布,随机擦除原始输入矩阵的部分特征,人为制造噪声,形成破损数据。在自动编码器进行训练时,为了使隐含层学习到更具有代表性的特性,需要从噪声数据中还原原始数据,使自动编码器必须进行降噪处理。

在深度学习中,多层次的神经网络具有较强的学习能力,但是层次过多易产生梯度消失或梯度爆发的问题。经过多次模型试验后,本文采用的降噪自动编码器如图 4 所示,由一个输入层、五个隐含层和一个输出层组成,各层节点数依次是 390×680×680×50×680×680×390。其中输入层包含有 390 个节点,对应语音输入特征;与输入层相连的两个隐含层 H_1 和 H_2 的节点数为 680,依次构造出两个高维的隐含层特征空间;网络中间节点数为 50 的隐含层 H_3 为编码输出;两个隐含层 H_4、H_5 和输出层共同完成解码的工作。在网络节点类型的选择上,输入层、中间编码层和输出层采用高斯型(Gaussian)线性激励节点,其他隐含层采用 sigmoid 非线性激励的伯努利型(Bernoulli)节点。

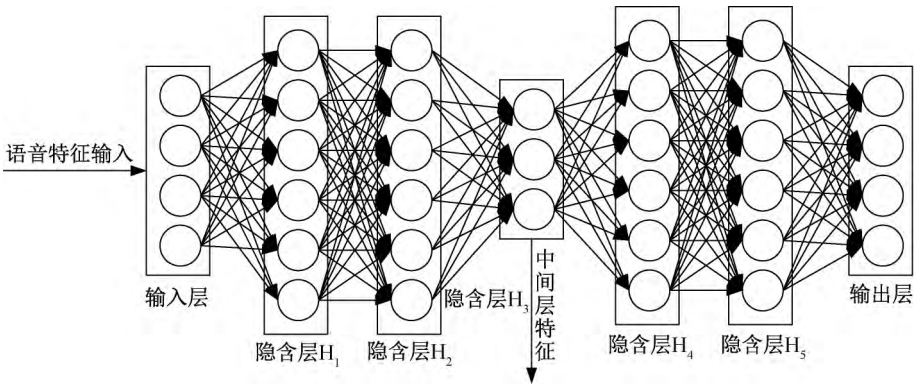

图 4 深度降噪自动编码器模型图

语音特征经过去燥自动编码器每层训练后,得到每一层的最优参数。但神经网络每层之间难以避免会有数据损失,这种损失经过层层训练,最终结果与原始的输入特征相比偏差较大,因此需要带有标签的有监督学习来对神经网络进行微调。在无监督训练顶层加上一层分类器进行模式分类,logistic 回归易出现分类和回归精确度不高、易出现欠拟合的问题,本文采用的激活函数是 Softmax 方式,它可以完成多类别划分且各类别之间是严格互斥的。Softmax 激活函数如式(2)所示:

$$S_{Z_i}=\frac{e^{Z_j}}{\sum_{i=1}^{n}e^{Z_i}} \tag{2}$$

其中Z_j为上一隐含层的输出,n 表示隐含层层数。经过 Softmax 分类器后的数据与样本标签数据进行比较,使用反向传播算法逐层进行参数调整,设定权值区间为$[-1,1]$,达到整个网络的最优化。收敛计算以原始输入和解码后数据的均方误差为目标函数,进行参数调整,降噪自动编码器的损失函数为:

$$J(\boldsymbol{w},\boldsymbol{b})=\Big[\frac{1}{m}\sum_{i=1}^{m}J(\boldsymbol{w},\boldsymbol{b};\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})\Big]+\frac{l}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_i}\sum_{j=1}^{s_{j+1}}(\boldsymbol{w}_{ji}^{(l)})^2+\sum_{\boldsymbol{x}ls}L(\boldsymbol{x},g(f(\boldsymbol{x}))) \tag{3}$$

其中:$g(f(x))$表示在输入数据中模拟的噪声数

据。式(3)第一项表示平均重构误差；第二项表示权重约束项，防止过拟合；第三项表示降噪的约束表达式。

语音识别模型训练主要算法流程如下：

a) 构建降噪自动编码器初始化映射矩阵，进行参数初始化，如式(4)所示，d_{in}和 d_{out}分别表示该层输入和输出的权重参数个数并且服从均匀分布，均匀分布使用 U 表示：

$$U \sim \left[-\frac{\sqrt{6}}{\sqrt{d_{\text{in}}+d_{\text{out}}}}, \frac{\sqrt{6}}{\sqrt{d_{\text{in}}+d_{\text{out}}}} \right] \tag{4}$$

b) 完成映射矩阵及偏置向量，利用字典类型存储数据，调用 a)中公式完成输入层到隐含层矩阵映射。

c) 构建 RBM 网络，完成每一层函数的训练。

d) 输入数据和高斯噪声系数进行 batch 训练。

e) 按式(3)构建损失函数。

f)对训练数据和测试数据进行标准化处理，使语音特征映射到 0－1 空间，如式(5)所示：

$$x_{\text{new}} = \frac{x-u}{\sigma} \tag{5}$$

其中：u 表示语音样本的均值，σ 表示语音样本的方差。

g) 循环载入训练数据。

3　实验与结果分析

3.1　语音特征提取及预处理

实验的第一步是样本数据的采集。从学校选取 40 人，20 男 20 女，发音清楚且能辨别发音内容，完成常规智能家居控制指令语音数据采集，例如关灯、开灯、关门、开门、开窗、关窗等 10 个常用短词语，采样率为 8 kHz，编码方式为 16 位，单声道，每人单控制短语采集 3 遍。共有 1200 个语音样本作为训练和识别的语音库，其中 900 个样本作为训练集，300 个样本作为测试集。

语音信号的特征训练之前必须经过数据的预处理[14]。首先增加采样样本高频频谱分辨率，通过预加重高频频谱的方式，来消除低频的影响，用公式可以表示为：

$$H(z) = 1 - az^{-1} \tag{6}$$

其中：a 为预加重系数。

为了获得实验所需的语音段样本信号，去除采集样本的静音段，将语音段和静音段分开需要端点检测技术[14]。本文采用时域端点检测方法，参数配置为帧长 256，帧移 128，短时平均能量的高门限为语音输入信号采样点最大能量值的 0.25 倍，短时平均能量的低门限为高门限的 0.25 倍，短时过零率高门限为 10，短时过零率低门限为 5，静音确定长度为 6，静音最短时间门限为 15。

语音特征提取使用梅尔频率倒谱系数(Mel-frequency cepstral coefficients，MFCC)[15]。对采集到的语音进行快速傅里叶变换(Fast Fourier transformation，FFT)得到语音频谱、幅度谱，然后对幅度谱进行梅尔滤波器对数运算再求余弦变换，求得 MFCC。但是此时得到的样本参数由于采集到的声音因人而异，提取到的样本参数长短不一致，为了保证之后的神经网络学习过程中参数一致，还需对初步特征参数进行压缩规整，降低单个样本的数据量，最终得到每一帧 12 维的特征参数。

3.2　模型实验与结果分析

本文实验在 TensorFlow 人工智能学习系统上完成的，分别针对深度信念网络(Deep Belief Network，DBN)和深度自动编码器模型进行特征训练。

本文通过深度信念网络将语音信号的提取特征转换为一个 96 维的向量，并进行归一化处理，使输入数据范围为[0，1.0]的闭区间，采用批处理梯度下降法来调整权值，控制权值更新为权值的 0.005 倍，本文的 RBM 学习率为 0.002。因为神经网络增加节点数比增加隐含层的单元要简单，且能有效提高识别度，所以本文采用双隐含层的 RBM 神经网络，对比实现的隐含层节点数 n 分别为 10、20、30、40、50，分别进行 100 次迭代，在每次迭代完成后计算 loss 值，并且每 10 次迭代计算识别准确度。

深度信念网络不同隐含层数下节点数 loss 值变化曲线和识别精度曲线如图 5 和图 6 所示，开始时随着节点数的增加，语音识别准确度不断上升，但是在节点数 30 之后，随着节点数的增加识别度反而下降，训练时间都是随着节点数的增加而增加的。所以在 DBN 神经网络中隐含层节点数目为 30 时网络性能最佳。

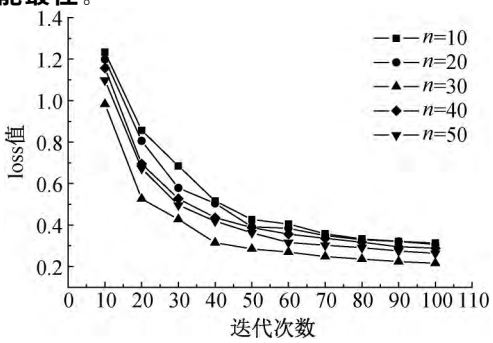

图 5　不同隐含层节点数下 loss 值变化曲线

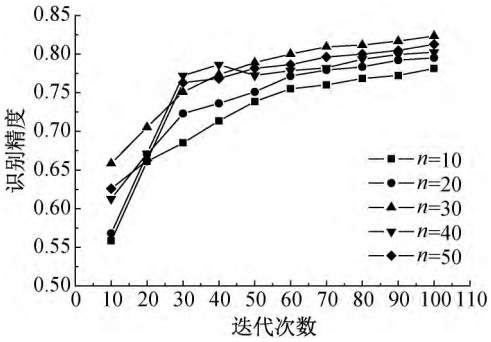

图 6 不同隐含层节点数下识别精度曲线

根据降噪自动编码器模型中的 RMB 初始化网络参数。随机的将一些网络节点值置为 0 来模拟噪声,RBM 模型参数的初始值采用随机函数 0.0001×rand()得到一个较小的值,偏置设为 0,学习率设为 0.002,隐含层之间链接权重的学习率为 0.005,进行 60 次迭代。网络模型每层训练完毕之后,采用 Softmax 激活函数对网络进行微调,设置最大的迭代次数为 100 次,并且当相邻两次迭代结果的均方误差变化率小于 0.001 时停止迭代。

三种不同的自动编码器 loss 值变化曲线和识别精度曲线如图 7 和图 8 所示。从图中可以看出,5 层的深度自动编码器比 3 层的自动编码器有更好的识别率,而深度降噪自动编码器与自动编码器相比,性能又提高了 2.02%。多层编码器网络由于其内部是非线性结构,能更好地进行特征学习,同时,降噪的引入,对语音识别结果有了更强的鲁棒性,进而提升了整体系统性能。

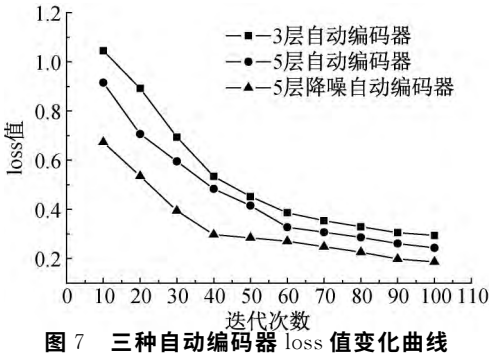

图 7 三种自动编码器 loss 值变化曲线

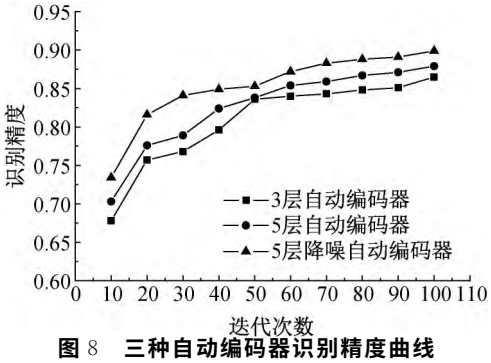

图 8 三种自动编码器识别精度曲线

综上分析,深度降噪自动编码器模型的语音识别准确率达到了 90.51%,高于 DBN 神经网络的 83.51%。在相同训练样本下,由于 DBN 神经网络在隐含层进行编码训练,在输出层与期望输出进行误差分析来调整网络参数。而降噪自动编码器通过人工增加噪声,每一层训练先是使用 RBM 扩展了语音输入特征,将语音特征从低维提升到高维后再压缩表示,得到语音核心特征,减少了样本差异带来的影响,再通过每层之间使用有监督的训练进行微调从而提高了语音识别率。

4 结 语

本文采用基于降噪自动编码器的声学模型进行语音识别训练,通过实验对比了 DBN 神经网络模型和普通自动编码器模型,结果表明:本文的声学模型不管是在语音识别率还是在噪声的鲁棒性方面都要优于其他的声学模型。但是在语音特征提取和识别训练中,训练样本和测试样本的数量有限,要想提高实验准确性,还需要大量的训练样本和测试样本;并且在编码器节点类型的设计和每层的节点数目上,还可以做更多的测试实验来进行优化,通过不断的深入研究,提高语音识别的准确度和效率,从而设计出用户交互性更好的系统。

参考文献:

[1] 王爱芸.语音识别技术在智能家居中的应用[J].软件,2015(7):104-107.
[2] 闵梁.面向智能家居的语音识别技术研究与实现[D].哈尔滨:哈尔滨工业大学,2013:25-49.
[3] 刘文强.语音识别技术在智能家居中的研究与应用[D].大连:大连海事大学,2013:15-62.
[4] 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(2):221-231.
[5] 居治华,刘罡,陈琦岚,等.基于反向卷积的 Bi-LSTM 语音识别[J]. 软件导刊,2017(7):27-30.
[6] 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(6):940-946.
[7] 包晓安,常浩浩,徐海,等.基于 LSTM 的智能家居机器学习系统预测模型研究[J].浙江理工大学学报,2018,39(2):224-231.
[8] 包晓安,林辉,周建平,等.基于智能家居的一致性模型融合技术研究[J].浙江理工大学学报,2015,33(1):109-114.
[9] 童江松,李仁旺,钱小燕.基于 ARM 的智能家居红外控制系统设计[J].浙江理工大学学报,2015,33(1):124-129.
[10] 侯一民,周慧琼,王政一.深度学习在语音识别中的研

究进展综述[J].计算机应用研究，2017，34(8)：2241-2246.

[11] Li Q，Huang Y. An auditory－based feature extraction algorithm for robust speaker identification under mismatched conditions[J]. IEEE Transactions on Audio，Speech，and Language Processing，2011，19(6)：1791-1801.

[12] 王山海，景新幸，杨海燕.基于深度学习神经网络的孤立词语音识别的研究[J].计算机应用研究，2015，32(8)：2289-2291.

[13] 蒋泰，张林军.语音识别自适应算法在智能家居中的应用[J].计算机系统应用，2017，26(3)：150-155.

[14] Zhang S X，Liu C，Yao K，et al. Deep neural support vector machines for speech recognition[C]// IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2015：4275-4279.

[15] Maas A L，Qi P，Xie Z，et al. Building DNN acoustic models for large vocabulary speech recognition[J]. Computer Speech & Language，2017，41：195-213.

（责任编辑：康　锋）