深度神经网络的语音深度特征提取方法

李 涛，曹 辉，郭乐乐
(陕西师范大学物理学与信息技术学院，陕西西安 710100)

摘要：为了提升连续语音识别系统性能，将深度自编码器神经网络应用于语音信号特征提取。通过堆叠稀疏自编码器组成深度自编码器(Deep Auto-Encoding, DAE)，经过预训练和微调两个步骤提取语音信号的本质特征，使用与上下文相关的三音素模型，以音素错误率大小为系统性能的评判标准。仿真结果表明相对于传统梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)特征以及优化后的 MFCC 特征，基于深度自编码器提取的深度特征更具优越性。

关键词：语音识别；深度自编码器；梅尔频率倒谱系数；
中图分类号：H107 **文献标识码**：A **文章编号**：1000-3630(2018)-04-0367-05
DOI 编码：10.16300/j.cnki.1000-3630.2018.04.013

Speech deep feature extraction method for deep neural network

LI Tao, CAO Hui, GUO Le-le
(*School of Physics and Information Technology*, *Shaanxi Normal University*, *Xian*, 710100, *Shaanxi*, *China*)

Abstrct: In order to improve the performance of continuous speech recognition system, this paper applies the deep auto-encoder neural network to the speech signal feature extraction process. The deep auto-encoder is formed by stacking sparsely the auto-encoder. The neural networks based on deep learning introduce the greedy layer-wise learning algorithm by pre-training and fine-tuning. The context-dependent three-phoneme model is used in the continuous speech recognition system, and the phoneme error rate is taken as the criterion of system performance. The simulation results show that the deep auto-encoder based deep feature is more advantageous than the traditional MFCC features and optimized MFCC features.

Key words: speech recognition; Deep Auto-Encoding (DAE); Mel-Frequency Cepstral Coefficient (MFCC)

0 引 言

语音识别是人类与机器进行语音交流，机器理解、识别人类的语音信号后将其转换成对应的文本或者命令的过程[1]。语音识别过程主要包括 3 个部分：语音特征的提取、建立声学模型与解码[2-3]。语音信号的特征提取在整个语音识别系统中至关重要，对这些特征进行降维、去噪，准确地提取出表示该语音本质的特征参数将使得后面的分类识别更有效，识别率更高。目前表示语音信息主要用的是短时频谱特征，比如梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)、差分倒谱特征(Shifted Delta Cepstra, SDC)、感知线性预测特征(Perceptual Linear Predictive, PLP)等。但这些短时频谱特征在实际的使用中都存在一些不足：以 MFCC 为例，每帧只包含 20～30 ms 语音，不但容易受到噪声干扰，而且还会忽略语音信号的动态特性和语音信号中所含有的类别信息，这些不足都会影响语音识别的准确率[4]。

2006 年 Hinton 等[5]提出基于深度信念网络(Deep Believe Network, DBN)的非监督贪心逐层训练算法，将深度学习算法应用于训练多层神经网络，它特殊的训练方式可以给神经网络提供较优的初始权值与偏置，使得网络能够快速地收敛于合理的极值点，有效避免了传统多层感知器(Multi-Layer Perceptron, MLP)在增加隐含层的同时易陷入局部最优解和需要大量有标记数据的问题。同时 DBN 的深度结构被证明相对于原有的浅层建模方法能够更好地对语音、图像信号进行建模。利用可以有效提升传统语音识别系统性能的深度神经网络 DBN 来进行语音识别[5]，学习到了更能表征原始数据本质的特征。随后 Hinton 等[6-7]提出了自编码器(Auto Encoder, AE)的深层结构：深度自编码器(Deep Auto Encoder, DAE)。自编码神经网络是一种网络误差函数定义与 DBN 不同的典型深度神经网络。

收稿日期：2017-08-04；**修回日期**：2017-10-18
基金项目：国家自然科学基金资助(1202020368、11074159、11374199)。
作者简介：李涛(1992－)，男，新疆伊犁人，硕士研究生，研究方向为信号与信息处理。
通讯作者：曹辉，E-mail: caohui@snnu.edu.cn

当隐含层节点的输入、输出呈线性关系，且训练网络采用最小均方误差(Least Mean Square Error, LMSE)准则时，整个编码过程与主成分分析(Principle Component Analysis, PCA)等效。当隐含层映射呈非线性映射时，即为自动编码器。本文采用这种自编码神经网络结构进行语音信号特征的提取。

1 深度自编码器的工作原理

深度自编码器是一种期望网络得到的输出为其原始输入的特殊深度神经网络。由于令该网络的输出趋近与它的原始输入，所以该网络中间层的编码完整地包含了原始数据的全部信息。但是是以一种不同的形式来对原始输入数据进行分解和重构，逐层学习了原始数据的多种表达。因此整个编码过程可看作是对信号的分解重构。将该网络结构用于特征压缩时，隐含层的神经元个数少于输入层神经元个数；把特征映射到高维空间时，则隐含层神经元个数多于输入层神经元个数。

自编码器是使用了无监督学习与反向传播算法，并令目标值趋近于输入值的前向传播神经网络。可对高维数据进行降维，进而得到低维的特征向量。设 $\boldsymbol{x}$ 向量为输入样本，则隐含层、输出层神经元的激活情况计算公式为

$$y=f(\boldsymbol{E}\boldsymbol{x}+\boldsymbol{b}) \tag{1}$$

$$h=f(\boldsymbol{E}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{b}') \tag{2}$$

其中，$f(x)=1/[1-\exp(-x)]$。$\boldsymbol{E}$ 与 $\boldsymbol{E}^{\mathrm{T}}$ 均为权重矩阵且两者互为转置，$\boldsymbol{b}$ 为隐含层偏置量，$\boldsymbol{b}'$ 为输出层偏置量。与传统神经网络有监督学习算法的区别在于：它是要使输出值尽量接近于输入值，即 $\boldsymbol{h}$ 趋近于 $\boldsymbol{x}$；损失函数 L 与 KL 之间的关系式为

$$L(\boldsymbol{X},\boldsymbol{H})=\sum_{i=1}^{n} KL(x_i\|h_i) \tag{3}$$

其中：$\boldsymbol{X}$ 为 n 个输入样本向量所组成的矩阵；$\boldsymbol{H}$ 为 n 个输出样本向量所组成的矩阵；$KL(x_i\|h_i)$ 表示 h 和 x 之间的 KL 散度，其目的是度量它们之间的差异[8]。预训练使用随机梯度下降法，可推导出其权值的更新公式为[8]

$$\boldsymbol{E}\leftarrow\boldsymbol{E}-\tau\frac{\partial L(\boldsymbol{X},\boldsymbol{H})}{\partial\boldsymbol{E}} \tag{4}$$

其中，τ 表示更新步长，参数 $\boldsymbol{b}$ 和 $\boldsymbol{b}'$ 的更新方法与 $\boldsymbol{E}$ 相同。

在训练自动编码器时，为了确保在处理数据过程中隐层神经元只有少部分被激活，故而限制隐含层的神经元被激活的数量，在损失函数中引入对激活隐层神经元数目的约束项，也就是实现对原始输入数据的稀疏编码，经证明稀疏编码能够有效降低模型的识别错误率[9]。损失函数为

$$L(\boldsymbol{X},\boldsymbol{H})=\sum_{i=1}^{N} KL(x_i\|h_i)+\upsilon\sum_{j=1}^{b} KL(p\|p_j) \tag{5}$$

其中：$KL(p\|p_j)=p\log\frac{p}{p_j}+(1-p)\log\frac{1-p}{1-p_j}$，$p_j$ 为第 j 个隐含层神经元被激活的概率，$p_j=\sum_{j=1}^{n} Y_{ij}/n$，$p$ 为初始值，该值是随机设定的，通常取值较小用以表示小概率事件；式(5)中第一项是对 N 个样本求和处理，第二项是对 b 个隐层神经元求和处理；υ 表示稀疏值惩罚的权重。

深度自编码器网络是由训练好的自编码器自底向上堆叠而成的，通过逐层贪婪训练算法得到自编码神经网络的参数，首先利用深度自编码器的原始输入训练该网络的第一层，得到该层参数 $E_{(1,1)}$、$E_{(1,2)}$、$b_{(1,1)}$、$b_{(1,2)}$，以及该层网络的隐含层神经元激活值组成的向量 M，接着把 M 作为深度自编码器第二层的输入，得到第二层参数 $E_{(2,1)}$、$E_{(2,2)}$、$b_{(2,1)}$、$b_{(2,2)}$；而后继续对余下的各层网络使用同样

的方法：上层的输出参数作为下层的原始输入参数依次训练整个网络；微调阶段利用反向传播算法调整所有层的参数。

常见的自编码器含有一个隐含层，如图 1 所示。文献[10]将深度神经网络定义为隐含层层数超过一层的神经网络。在本文中构建一个含有两层隐含层的深度神经网络来提取语音信号的深度特征。网络结构如图 2 所示。

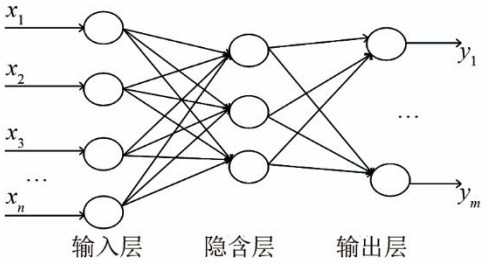

图 1 单隐含层神经网络
Fig.1 Single hidden layer neural network

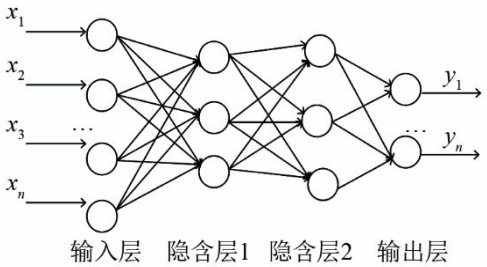

图 2 深度神经网络
Fig.2 Deep neural network

2　基于 DAE 模型的深度特征提取

因说话人、说话方式不同及噪声等影响，可能使在实验环境下表现优异的语音识别系统在实际应用中的识别性能不稳定。因此，使用改善系统的鲁棒性和自适应能力的方法来优化声学特征参数，增强识别系统的抗干扰能力，使其性能更加稳定，能够应对多种环境。目前常用解决方法是：为增强特征参数的适应能力而对其进行特征变换处理；或为提高特征参数的鲁棒性而对语音信号进行增强、滤波、去噪等处理。

提取深度特征之前，先对提取的 MFCC 特征进行特征变换，再作为深度自编码器的原始输入，进而得到识别率更高的语音深度特征，对原始 MFCC 特征依次进行线性判别分析、最大似然线性变换和最大似然线性回归变换处理。

考虑到协同发音的影响，将已提取的 39 维 MFCC 特征向量(静态、一阶、二阶差分)进行前后 5 帧的拼接，得到 39 ×11=429 维的特征向量。对这 429 维特征向量进行线性判别分析(Liner Discriminant Analysis, LDA)抽取分类信息，同时降低维度至 40 维从而得到 LDA 特征。然后对这 40 维 LDA 特征向量进行最大似然线性变换(Maximum Likelihood Linear Transformation, MLLT)来去除相关性得到 LDA+MLLT 特征，最后对经过去除相关性的 40 维 LDA+MLLT 特征在特征空间上进行最大似然线性回归(Feature-space Maximum Likelihood Linear Regression, fMLLR)说话人自适应训练，实现特征参数自适应，减小测试声学特征与声学模型参数之间的不匹配，得到了 40 维的 LDA+MLLT+fMLLR 特征。仿真结果表明，以上特征变换均能有效降低音素识别的错误率。

深度自编码器能够更好地对语音信号中与音素相关的信息进行逐层表征，基于深度自编码器提取的语音深度特征过程，实质上是一种非线性的特征变换和降维过程。利用神经网络的层次化提取信息过程来作为对原始输入特征的非线性特征提取与转换，使得特征维度与神经网络训练目标尺度分离。相对网络首层输入层而言，隐层的神经元个数要少得多，所以隐层在通过学习到原始输入样本的低维表示的同时，还可以最大限度地包含与高维表示相同的信息。并且可以通过更精细的子音素类别来表示音素目标，最终由原始输入向量经过逐层映射得出对应隐含层的输出向量。由此就得到能够最大限度地包含输入向量信息的一个低维编码，这使得输出的深度特征具有比传统底层声学语音特征参数相近或更好的特性区分性，还带有类别信息，加强了特征表示声学单元的能力，得到更有效的特征表达，进而提高后期语音识别系统的性能。使用 DAE 提取深度特征的流程图如图 3 所示。

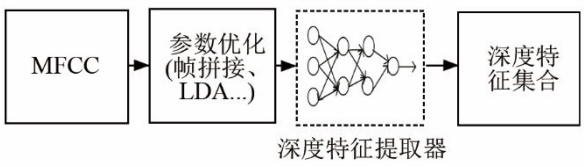

图 3　深度特征提取流程

Fig.3　Deep feature extraction process

本文使用的原始输入特征是进行前后 5 帧拼接的 40 维 LDA+MLLT+fMLLR 特征，形成 40×11=440 维的输入特征向量，这 11 帧拼接的 LDA+MLLT+fMLLR 特征相对于传统的单帧特征更具优势[11]：一个音素持续的时间大约在 9 帧左右，所以大约 9 帧的信息量就能够包含一个完整的音素，同时也含有其他音素的部分信息，它可以提供单帧特征所体现不出的更细致更丰富的音素变化信息。

利用深度自编码器神经网络进行深度特征参数提取的步骤如下:

(1) 以 11 帧拼接 LDA+MLLT+fMLLR 特征作为输入，经训练得出第一层隐含层的网络参数，并以此计算第一层隐含层输出；

(2) 把第一层隐含层的输出作为第二层的输入，再用同样的训练方法得出第二层隐含层的网络参数以及该层的输出；

(3) 继续把上一层的输出作为第三层的输入，再用同样的方法训练该层网络的参数，而后利用反向传播算法微调所有层的参数。最后将输出层输出的深度特征参数作为最终音素识别系统的输入。

3　仿真结果与分析

3.1　数据库与仿真环境

仿真使用标准 TIMIT 语音库(The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus)证明本文深度特征的有效性，TIMIT 语音数据库采集 630 人(其中男性 438 人，女性 192 人)的美式语音数据，每人只录 10 句，共 6 300 个语音。语音库中包含测试集和训练集，两集之间无相同说话人，其中训练集由 462 个人所讲的 3 696 个句子组成，测试集包括由 168 个人所讲的 1 344 个句子组成，

两集之间无相同说话人。输入语音采样使用汉明窗，设定：窗长为 25 ms、窗移为 10 ms。利用开源工具包 kaldi 在 LINUX 系统上使用图形处理器进行此次实验仿真。建立含有两个隐含层的深度自编码器，隐含层偏执为 1.0，神经元激活函数为 $f(x)=1/(1-\exp(-x))$。

以 11 帧拼接的 LDA+MLLT+fMLLR 特征作为原始输入，经过归一化之后，所有输入数据大小都在 0～1 之间。为保证实验的准确性和客观性，音素识别的基线系统选择常用的混合隐马尔科夫模型(Hidden Markov Model, HMM)+深度神经网络模型(Deep Neural Network, DNN)音素识别系统。

3.2 分析

本文设计 2 个实验来验证深度特征的优越性，用音素错误率(Phoneme Error Rate, PER)作为评价特征有效性的标准。

3.2.1 最优神经网络配置

隐层单元数与隐层数的选择将影响后期识别的音素错误率。若神经元过少，学习的容量有限，网络所获取的解决问题的信息不足，难以存储训练样本中蕴含的所有规律。若神经元过多就会增加网络训练时间，还可能把样本中非规律性的内容存储进去，反而会降低泛化能力。通过改变隐层层数与每层神经元个数来确定网络最佳配置，设置隐层层数从 1 到 3 层变化，每个隐层所含神经元个数以 50 的偶数倍增加，最多为 400 个。为降低计算量，减少训练时间，将每层隐含层的神经元设置成相同个数。对比不同网络结构配置下音素识别率的变化，进而选定最优参数配置。图 4 显示了改变隐含层的层数与神经元个数对最终音素识别错误率的影响。

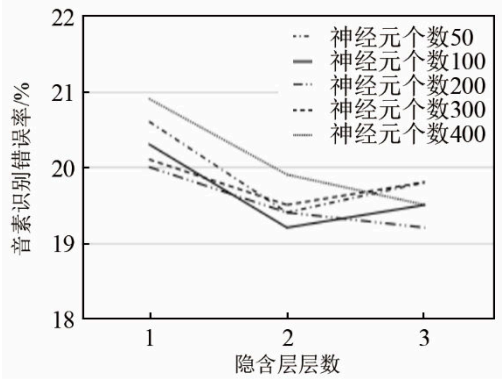

图 4 不同隐层层数与神经元个数对音素识别错误率的影响

Fig.4 Effects of the number of hidden layers and the number of neurons on phoneme recognition error rate

从图 4 可以看出，改变隐层层数和隐层神经元个数对降低音素识别错误率有一定影响，当隐含层为两层且隐含层神经元为 100 时与隐含层为三层隐含神经元为 200 时错误率最小，并不是隐含层数与神经元个数越多越好。当隐层层数与隐层神经元个数增加至一定数量时，音素错误率不会降低反而上升，同时由于计算参数的增加使得训练时间增长，为减少计算参数及训练时间，同时确保音素识别正确率，本文选择建立含有两个隐含层的深度神经网络。深度自编码器的输入神经元个数即为输入特征的维数 440，每一隐含层神经元个数为 100，输出层神经元个数设置为 40，则该深度自编码器结构可表示为 440-[100-100]-40，“[ ]”中数字为隐层神经元的个数。

3.2.2 特征有效性对比

将本文特征解码的结果与以下四种特征解码得出的音素错误率进行对比，结果如表 1 所示。作为对比的四种特征分别为：(1) 原始 MFCC 特征参数；(2) LDA+MLLT 特征：MFCC 在三音素模型的基础上进行 LDA+MLLT 变换；(3) LDA+MLLT +fMLLR 特征：在(2)的基础上进行基于特征空间的最大似然线性回归(fMLLR)的说话人自适应训练；(4) bottleneck 特征：以 11 帧拼接的 MFCC 特征作为原始输入，建立含有五个隐含层的 DBN 网络，输入输出层神经元个数为 440，第四隐含层为瓶颈层且其神经元个数为 40，其余隐含层神经元个数为 1 024，提取出 bottleneck 特征。

由表 1 可知，与传统特征以及特征变换后的优化特征作为 HMM+DNN 系统的输入相比，将深度特征作为系统原始输入时，系统的音素错误率明显下降，同时相对于使用 DBN 网络提取 bottleneck 特征，其网络参数的计算量和训练时长较少。表 1 中的结果也证明了本文提取的深度特征的有效性。

表 1 传统特征与深度特征的音素错误率对比

Table 1 Comparison of phoneme error rate between traditional and deep features

各类语音特征	音素错误率/%
MFCC	25.6
LDA+MLLT	23.8
LDA+MLLT+fMLLR	21.6
MFCC+bottleneck	20.1
深度特征	19.2

4 结 语

针对传统语音特征的不足，本文对原始 MFCC 特征参数优化之后，建立含有两个隐层的深度自编码器，将优化后的 MFCC 参数作为其输入，实现原

始输入的特征变换与降维，提取了可以更好地反应语音本质特征的深度特征参数，作为 HMM+DNN 系统的输入。实验证明了本文特征的有效性。下一步研究将在本研究基础上与 DBN 结合，提取更优异的声学特征，进一步提高语音识别系统的性能。

参 考 文 献

[1] 韩纪庆, 张磊, 郑铁然. 语音信号处理[M]. 北京: 清华大学出版社, 2005.
HAN Jiqing, ZHANG Lei, ZHENG Tieran. Speech Signal Processing[M]. Beijing: Tsinghua University Press, 2005.

[2] 陈雷, 杨俊安, 王一, 等. LVCSR 系统中一种基于区分性和自适应瓶颈深度置信网络的特征提取方法[J]. 信号处理, 2015, **31**(3): 290-298.
CHEN Lei, YANG Junan, WANG Yi, et al. A feature extraction method based on discriminative and adaptive bottleneck deep confidence network in LVCSR system[J]. Signal Processing, 2015, **31**(3): 290-298.

[3] SCHWARZ P. Phoneme Recognition Based on Long Temporal Context[EB/ OL]. [2013-07-10]. http://speech. Fit. Vutbr. cz/software/Phoneme-recognizer-based-long-temporal-context.

[4] GREZL F, FOUSEK P. Optimizing bottleneck feature for LVCSR[C]//IEEE International Confe rence on Acoustics, Speech and Signal Processing, 2008: 4792-4732.

[5] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, **18**(7): 1527-1554.

[6] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, **313**(5786): 504-507.

[7] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. 计算机应用研究, 2012, **29**(8): 2806-2810.
SUN Zhijun, XUE Lei, XU Yangming, et al. Review of deep learning research[J]. Journal of Computer Applications, 2012, **29**(8): 2806-2810.

[8] 张开旭, 周昌乐. 基于自动编码器的中文词汇特征无监督学习[J]. 中文信息学报, 2013, **27**(5): 1-7.
ZHANG Kaixu, ZHOU Changle. Unsupervised learning of Chinese vocabulary features based on automatic encoder[J]. Journal of Chinese Information Processing, 2013, **27**(5): 1-7.

[9] COATES A, NG A Y, LEE H. An analysis of single- layer networks inunsupervised feature learnin[C]//Proc of International Conferenceon Artificial Intelligence and Statistics. 2011: 215-223.

[10] HINTON G E, 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.

[11] SIVARAM G, HERMANSKY H. Sparse multilayer per-ceptron for phoneme recognition[J]. IEEE Transac-tions on Audio, Speech, and Language Processing, 2012, **20**(1): 23-29.