一种新的基于瓶颈深度信念网络的特征提取方法及其在语种识别中的应用

李晋徽 杨俊安 王 一

(电子工程学院 合肥 230037) (电子制约技术安徽省重点实验室 合肥 230037)

摘 要 在语种识别中,传统的MFCC特征由于每帧信号上的信息量不足,很容易受到噪声污染,且抗噪能力较弱。同时,目前普遍使用的SDC特征提取方法在参数选择上需要人为设定,这增加了识别结果的不确定性。针对上述不足,将深度学习方法引入特征提取之中,提出了基于瓶颈深度信念网络的特征提取方法。最后在NIST2007数据库上对瓶颈层的大小、隐层数目以及瓶颈层位置进行了相关的对比实验,结果表明,提出的方法相对于传统的特征提取方法能够取得更高的识别率。

关键词 语种识别,瓶颈特征,深度信念网络

中图法分类号 TM344.1 **文献标识码** A

New Feature Extraction Method Based on Bottleneck Deep Belief Networks and its Application in Language Recognition

LI Jin-hui YANG Jun-an WANG Yi

(Electronic Engineering Institute, Hefei 230037, China) (Anhui Key Laboratory of Electronic Restriction, Hefei 230037, China)

Abstract In language recognition, due to the insufficiency of information in each frame, traditional MFCC feature extraction is easily suffered from noise pollution. Meanwhile, the general method of SDC feature extraction depends on artificially setting in parameter selection which increases the uncertainty of recognition performance. In order to overcome these drawbacks, the deep learning method was introduced and a novel feature extraction approach named BN-DBN which is based on deep learning was proposed. Finally, the relevant comparative experiments for the bottleneck layer size, the number of hidden layers and the position of the bottleneck layer were carried out in NIST2007 database. Experimental results show that extraction method of the bottleneck features based on deep belief networks are more effective in language recognition, compared with traditional methods.

Keywords Language recognition, Bottleneck features, Deep belief networks

1 引言

语种识别(Language Identification, LID)是语音识别的一个重要分支,其目的是通过对给定的语音段进行分析处理,识别出所属语言种类。近年来,随着全球一体化进程的不断加快,不同语言间的信息交互日趋频繁,LID在自动转换服务、多语种信息补偿等信息检索领域有着重要的应用。此外,LID还可在军事上用来对目标身份和国籍进行监听或判别。随着信息时代的到来以及因特网的发展,语种识别越来越显示出其应用价值。

语种识别技术的研究最早可追溯到1974年,TI公司采用音素单元序列对不同语种进行分类识别[1]。之后的近40年里,语种识别的发展日新月异,相关技术也日趋成熟,逐渐形成了主流的以并行高斯混合模型进行语种识别的方法[2,3]。目前语种识别系统中常用的梅尔倒谱系数特征(Mel-Frequency Cepstral Coefficient, MFCC),由于每帧信号通常只包含20~30毫秒的语音信号,很容易受到噪声污染,其抗噪能力较弱[4,5]。对于另一种特征提取方法——差分倒谱参数(Shifted Delta Cepstra, SDC)[6]而言,其虽然相对于MFCC参数进行了很大的改进,但是由于SDC的参数均是人为设定,使得其无法通用于所有的语音数据,通用性较差。

针对上述问题,本文将瓶颈结构(Bottle-Neck: BN)和近似人工神经网络(Artificial Neural Networks,以下简称ANN)[7-9]的深度信任神经网络(Deep Belief Network, DBN)相结合,提出了一种新的特征提取方法,称为BN-DBN方法。DBN由于具有对输入数据的内部统计结构和密度函数的要求不严格,可对较长时间段的语音数据进行处理,并且对不同说话人的说话方式、口音、外界噪声等干扰的鲁棒性更强等优点,因此在处理语音数据时,具有更强的建模和表征能力[10]。本文通过NIST07语音数据库数据,使用瓶颈(Bottle-Neck, BN)DBN方法进行了语种识别实验,实验结果表明与传统语种识别方法MFCC、SDC相比,基于BN-DBN方法能够更有

到稿日期:2013-05-29 返修日期:2013-08-03 本文受国家自然科学基金项目(61272333)资助。

李晋徽(1988—),女,硕士生,主要研究方向为语种识别,E-mail:hhlee88223@yahoo.com.cn;杨俊安(1965—),男,博士,教授,博士生导师,主要研究方向为通信信号处理、智能计算等。

效地提高识别准确率。

2 一种新的基于 BN-DBN 的语音特征提取方法

DBN 由 Hinton[11,12] 等于 2006 年正式提出，从理论上说是一种对具有深层结构(即包含多层非线性运算单元)的模型进行学习的方法，它相对于以往针对“浅层”结构(即只含有单层非线性运算单元)的建模方法，在处理真实世界中的数据(如自然语音、自然图像、视频等)时，具有更强的建模和表征能力。本质上 DBN 仍然是一种多层 ANN，但是它采用了一种监督式和非监督式相结合的训练手段来获取网络参数，解决了 ANN 反向传播算法(Back-propagation Algorithm)很容易陷入局部最优这一问题。

瓶颈的概念由 Grézl 等于 2007 年首次提出并应用于连续语音识别当中[13]，而 BN-DBN 则是将瓶颈的概念与 DBN 相结合的产物。在 BN-DBN 的网络组成结构上，它通常被设定为一个奇数层的多层 ANN，并将其中最中间的一层命名为瓶颈层。顾名思义，瓶颈的意思就是指该层神经元个数相对于其他层要少得多。基于 BN-DBN 的语音特征提取方法可以分为两个步骤实施。

步骤 1　建立神经网络，通过预训练与微调，建立一个 DBN。

从组成上看，DBN 是由一系列受限波尔兹曼机(Restricted Boltzmann Machine，RBM)层叠而成的，一个完整的 DBN 的组成结构如图 1 所示。

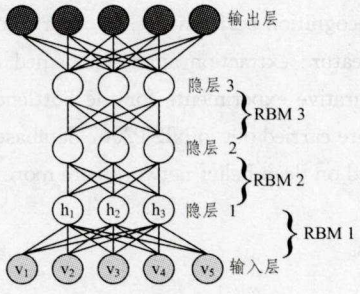

图 1　DBN 组成结构图

如图所示，一个 RBM 由可视层 v_i 和隐层单元 h_j 相互连接而成，在给定模型参数条件下的联合分布可用能量函数来表示：

$$E(v_i,h_j;\theta)=-\sum_{ij}w_{ij}v_ih_j-\sum_i b_iv_i-\sum_j a_jh_j \tag{1}$$

其中，$\theta=\{w,a,b\}$，w_{ij} 为可视单元和隐单元的连接权值。b_i 和 a_j 分别是对应偏置量。概率密度分布可以通过 Boltzmann 分布来确定：

$$P_\theta(v,h)=\frac{1}{Z(\theta)}\exp(-E(v,h;\theta))$$

$$=\frac{1}{Z(\theta)}\prod_{ij}e^{w_{ij}v_ih_j}\prod_i e^{b_iv_i}\prod_j e^{a_jh_j} \tag{2}$$

其中，$Z(\theta)=\sum_{h,v}\exp(-E(v,h;\theta))$，因为隐藏节点之间是条件独立的，即：

$$P(h|v)=\prod_j P(h_j|v) \tag{3}$$

通过上式，可以比较容易地得到在给定可视层 v 的基础上，隐层第 j 个节点为 1 或者为 0 的概率：

$$P(h_j=1|v)=\frac{1}{1+\exp(-\sum_i w_{ij}v_i-a_j)} \tag{4}$$

同理，在给定隐层 h 的基础上，可以得到可视层第 i 个节点为 1 或者为 0 的概率：

$$P(v|h)=\prod_i P(v_i|h) \tag{5}$$

$$P(v_i=1|h)=\frac{1}{1+\exp(-\sum_j w_{ij}h_j-b_i)} \tag{6}$$

最大化以下对数似然函数：

$$L(\theta)=\frac{1}{N}\sum_{n=1}^{N}\log P_\theta(v)-\frac{\lambda}{N}\|w\|_F^2 \tag{7}$$

对最大对数似然函数求导，得到 L 最大时对应的参数 w。

$$\frac{\partial L(\theta)}{\partial w_{ij}}=E_{P_{data}}[v_ih_j]-E_{P_\theta}[v_ih_j]-\frac{2\lambda}{N}w_{ij} \tag{8}$$

如果把隐藏层的层数增加，我们可以得到深度波尔兹曼机(Deep Boltzmann Machine，DBM)；如果在靠近可视层的部分使用贝叶斯信念网络，而在最远离可视层的部分使用 RBM，便可以得到 DBN，然后就可以采用类似传统 BP 神经网络的监督式学习方式，对整个 DBN 进行由后至前的回调(在 DBN 中称为微调)，最终建立 DBN[14,15]。

步骤 2　如图 2 所示，将瓶颈层之后的网络去除，把原来的瓶颈层作为输出层。

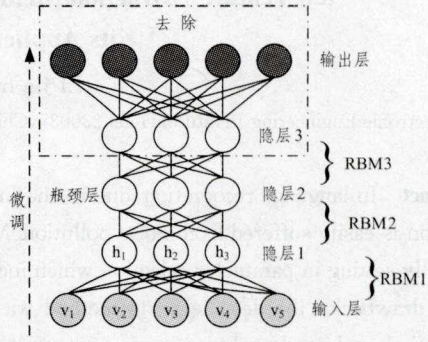

图 2　BN-DBN 组成结构图

以一个 5 层网络为例[16,17]，图 3 是一个基于 BN-DBN 的语音特征提取方法实施示意图。隐层 2 是瓶颈层，输入数据经过 2 个显层和 3 个隐层对网络进行训练，并采用类似传统 BP 神经网络的监督式学习方式，对整个 DBN 进行由后至前的微调，最终建立 DBN。在得到训练好的隐层 2 后，去除隐层 3 和输出层，将瓶颈层作为新的输出层，即输入数据通过输出层、隐层 1、瓶颈层，最终得到输出数据。

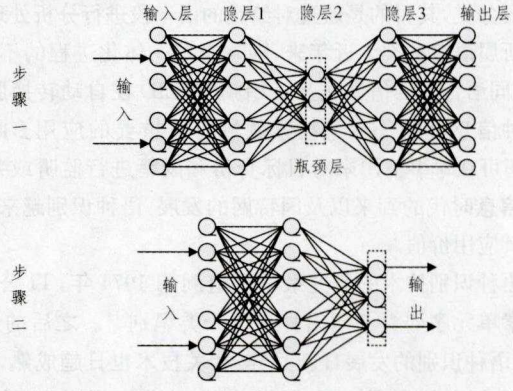

图 3　基于 BN-DBN 的语音特征提取方法示意图

在实际应用中，我们利用神经网络对多帧信号的处理能力，将多帧特征拼接后送入网络(在实验中选取连续 10 帧语音信号送入神经网络)，因此输入层神经元个数等同于帧数×每帧特征的维数；在设定隐层 1 的神经元个数时，我们通常将其神经元个数设得尽量大，使其能够提供强大的建模能力，保

证其能够获取数据的内部结构，隐层 3 的神经元个数等同于隐层 1；而对于瓶颈层，通常将其神经元个数设定为等同于单帧的维数。

3 BN-DBN 在语种识别中的应用实验及分析

3.1 实验数据分析及环境

实验语音库来自 NIST LRE 2007[18]，数据为电话录音，真实对话风格，包含噪音、停顿、呼吸、重复、不完整的发音、口音，等等，采样频率为 8kHz。训练集 R 使用英语、法语、俄语、日语、韩语 5 个语种，包含 3 分钟至 6 分钟的语音片段共 1000 条，通过端点检测技术将这些语料切分为 6268 句语音段，每段有效语音时长不少于 30 秒。测试集 T 采用 LRE 2007 的闭集测试，共 5 个语种（英语、法语、俄语、日语、韩语），包括 30 秒、10 秒语音段各 600 句，且保证 $T \cap R = \varnothing$。

3.2 实验设计及评价指标

特征部分对各语种提取 MFCC 参数，将其作为输入进入 BN-DBN 网络，得到新的语音特征。模型训练阶段使用 GMM-UBM 模型[19]，对提取出的特征按照最大似然估计（Maximum Likelihood Estimation，MLE）准则训练出一个 512 阶的通用背景模型（Universal Background Model，UBM），以 UBM 模型为基础，从中自适应出各语种的高斯混合模型（Gaussian Mixture Model，GMM）作为初始模型，进一步采用最大互信息量准则（Maximum Mutual Information，MMI）继续训练，迭代 10 次，得到最终的 MMI 模型。测试阶段使用测试集中语句提取特征后分别与 UBM 模型和 MMI 模型进行对数似然率计算得分，归一化后得分最大的即为识别语种。(BN-DBN)-GMM-UBM 系统框图如图 4 所示。

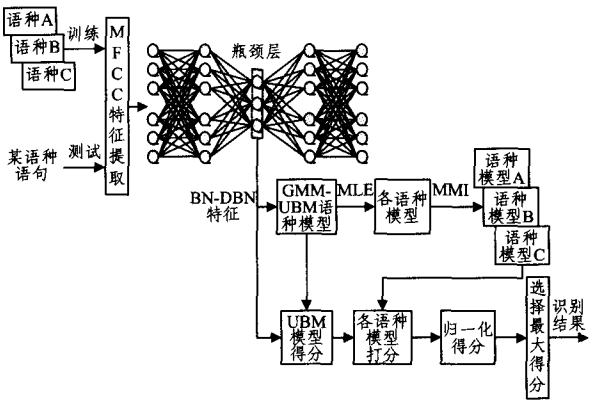

图 4 (BN-DBN)-GMM-UBM 系统框图

为了充分验证 BN-DBN 相对于其他语音特征的有效性，本文设计了 3 个实验，分别为对采用不同 BN 层维数的 BN-DBN 进行比较、对采用不同 DBN 网络架构的 BN-DBN 进行比较、对 4 种不同语音特征进行比较，以此来验证瓶颈层神经元个数对识别结果的影响，以及不同 DBN 结果对识别结果的影响。实验表明 BN-DBN 相对于其他特征提取算法能够有效提高识别率。

由于语种识别是一个典型的分类的问题，评价指标由等概率错误（Equal Error Rate，EER）表示，一般的研究都是集中在两个量来表示，即虚警概率 E_{fa} 和漏警概率 E_{ma}：

$$E_{fa} = \frac{n_{fp}}{n_{tp} + n_{fp}} \tag{9}$$

$$E_{ma} = \frac{n_{fn}}{n_{tp} + n_{fn}} \tag{10}$$

其中，E_{fa} 反映了被判为目标语种的样本中有多少是冒认的，E_{ma} 反映了有多少目标语种不被接受。EER 表示 E_{fa} 和 E_{ma} 曲线交点的错误率，在该点处取 $E_{fa} = E_{miss}$ 时 E 来衡量系统性能。

3.3 实验结果

实验 1 基于不同 BN 层的 BN-DBN 性能比较

改变瓶颈层神经元个数，依次设置为 20、25、30、35、39、50、60，结果如表 1 所列。

表 1 基于不同 BN 层的 BN-DBN 性能比较(EER%)

特征	30s	10s
BN-DBN20	1.91%	6.62%
BN-DBN25	1.86%	6.42%
BN-DBN30	1.86%	6.41%
BN-DBN35	1.85%	6.39%
BN-DBN39	1.82%	6.34%
BN-DBN50	1.87%	6.47%
BN-DBN60	1.94%	6.61%

实验 2 基于不同结构的 BN-DBN 性能对比

实验分别采用 3 层、5 层和 7 层 BN-DBN39 网络，并变换隐层神经元个数。

3 BN-DBN39：3 层神经元组合为 1024-39-1024

5 BN-DBN39：5 层神经元组合为 1024-1024-39-1024-1024

5 BN-DBNGR39：5 层神经元组合为 1024-512-39-512-1024

7 BN-DBN39：7 层神经元组合为 1024-1024-1024-39-1024-1024-1024

7 BN-DBNGR39：7 层神经元组合为 1024-1024-512-39-512-1024-1024

结果如表 2 所列。

表 2 基于不同结构的 BN-DBN 性能对比(EER%)

特征	30s	10s
3BN-DBN39	2.12%	6.96%
5BN-DBN39	1.82%	6.34%
5BN-DBNGR39	1.88%	6.56%
7BN-DBN39	1.91%	6.68%
7BN-DBNGR39	2.01%	6.82%

实验 3 基于 4 种不同特征提取的性能比较

这 4 种不同特征为：

MFCC39：各语种样本先经过预加重滤波器 $H(z) = 1 - 0.97z^{-1}$ 然后进行多帧平均，每帧长为 256 点，帧移 128 点，窗函数采用 Hamming 窗，并且滤波器组使用 24 个 Mle 三角滤波器，并将 24 维 MFCC 进行差量倒频谱参数等变化，形成最终的 39 维参数。

MFCC39-11：基于 MFCC39，将每一帧前后各扩展 5 帧(5+1+5)，变为新的 11 帧参数。

SDC：基于 MFCC39，得到前 7 阶系数(C0-C6)，MFCC 按(7,1,3,7)(N,d,P,k)扩展为 49 维特征，将 49 维 SDC 和 7 阶 MFCC 系数拼接起来，得到最终使用的 56 维 SDC 特征参数。通过 VTLN 正规、RASTA 滤波、VAD 检测、高斯化、倒谱域减均值(CMS)等预处理方法去除噪声和说话人影响等。

BN-DBN39：提取的 MFCC39-11 特征进入 5 层 BN-DBN 网络，其中 5 层神经元个数为 1024-512-39-512-1024 组合，通过大量逐层学习自底向上的 RBM 可以构建一个初始的

DBN，最后采用类似传统 BP 神经网络的监督式学习方式，对整个 DBN 进行由后至前的微调。结果如表 3 所列。

表 3　基于不同特征提取的性能比较(EER%)

特征	30s	10s
MFCC39	5.63%	10.45%
MFCC39-11	2.74%	7.28%
SDC	1.93%	6.79%
BN-DBN39	1.82%	6.34%

3.4　结果分析

结果表明：

(1)瓶颈层的神经元个数在一定范围内对于 BN-DBN 识别性能的影响并不太明显

实验 1 为了测试瓶颈层神经元个数的变化对识别效果的影响，依次选取 20、25、30、35、39、50、60 个神经元作为瓶颈层的个数，由实验结果可以发现，瓶颈层的神经元个数在一定范围内对于识别性能的影响并不太明显，因此建议瓶颈的大小为 24 到 50 个神经元。

(2)网络隐层数和瓶颈层位置对 BN-DBN 识别性能具有一定影响

实验 2 在测试 BN-DBN 结构变化对识别结果影响时分别将隐层数设置为 3 层、5 层和 7 层，通过实验 3，认为当网络隐层数增加时，中间的瓶颈层在监督过程中只能得到较少的信息，此外还能发现，当瓶颈层向前放置时，前面的隐层提取特征的能力非常有限。

(3)BN-DBN 相比 MFCC、SDC 方法具有更好的鲁棒性

实验 3 选取了 4 种不同的特征提取方法，可以看出 BN-DBN 特征的识别率相比传统的 MFCC 以及帧扩展的 MFCC 有明显提高。这是因为 BN-DBN 对传统的 MFCC 做了帧扩展，在相对较长的一帧信号中提取出了更多的语种信息，并且监督式的训练方法不但提高了鲁棒性还克服了 SDC 人为规定特征参数的局限性，且效果与 SDC 基本相当。

正如前文所述，传统的 MFCC 特征不能很好地提取有效语种信息，并且容易受到噪声污染。而 SDC 特征的计算受控于其 4 个参数：*N-d-P-k*(*N*：倒谱参数的个数；*d*：计算差分倒谱的帧间间隔；*P*：计算差分倒谱的相邻块的帧移；*k*：差分倒谱块的个数)。通常，*N-d-P-k* 需要通过实验的结果来人工确定最优组合，过程繁琐且耗费系统资源大。由于 BN-DBN 具有 DBN 和瓶颈特征的共同优点，在具有神经网络优点的同时，能够有效克服现有特征提取算法的缺陷。主要特点在于：

1)能够提取多帧数据的特征，而多帧数据中已经被证明包含不同语种特点的重要信息；

2)DBN 具有强大的对数据内部结构和统计特征的表征能力，提取的特征更能反映不同语种的本征特征；

3)BN-DBN 在预训练部分采用了无监督学习方式，因此可以有效利用未标签数据；

4)BN-DBN 的微调部分采用监督式学习的方法，可以根据标签信息对神经网络参数进行调整，使得提取的特征更具有判决性，利于分类；

5)BN-DBN 等同于一种降维方法，能够提高系统运算速度；

6)DBN 是一种神经网络方法，需要人为设定的参数少，可以由机器主动学习进行训练。

结束语　本文针对目前大多数语音特征提取方法不能充分利用多帧信息、对外界干扰敏感、需要人工设定较多参数等问题，提出采出瓶颈深度信念网络的方法来解决这些问题，从而达到提高识别准确率的最终目的。在 NIST2007 数据库中的 3 个实验表明，瓶颈深度信念网络算法的识别正确率相对于文中所比对的另外 3 种算法均取得了较好的结果。

参 考 文 献

[1] Rabiner L R, Sambur M R. An algorithm for determining the endpoints of isolated utterances [J]. The Bell System Technical Journal, 1975, 54(2): 297-315

[2] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[C]// Digital Signal Processing. 2000: 19-41

[3] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification [J]. IEEE Signal Processing Letters, 2006, 13: 308-11

[4] Bilmes J A. Maximum mutual information based reduction strategies for cross-correlation based joint distribution modeling[C]// IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP). Seattle, USA, May 1998

[5] Yang H H, Sharma S, van Vuuren S, et al. Relevance of time-frequency features for phonetic and speaker-channel classification[J]. Speech Communication, 2000, 31(1): 35-50

[6] 付强. 基于高斯混合模型的语种识别的研究 [D]. 合肥：中国科学技术大学，2009

[7] Fousek P, Lamel L, Gauvain J-L. Transcribing Broadcast Data using MLP Features[C]// Proceedings of Interspeech. 2008

[8] Park J, Diehl F, Gales M, et al. Training and Adapting MLP Features for Arabic Speech Recognition [C] // Proc. of IEEE Conf. Acoust. Speech Signal Process(ICASSP). 2009

[9] Picheny M, Nahamoo D, Goel V, et al. Trends and Advances in Speech Recognition[J]. IBM Journal of Research and Development, 2011, 55(5): 2

[10] Deng L. An Overview of Deep-Structured Learning for Information Processing[C]// APSIPA ASC 2011. Xi'an: 2011

[11] Hinton G E, Osindero S, Teh Y. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18: 1527-1554

[12] Hinton G E, Salakhutdinov R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, Recognition, Ph. D. thesis, OGI, Portland, USA, 2006, 313(5786): 504-507

[13] Grézl F, Karaflat M, Kontar S, et al. Probabilistic and bottleneck features for LVCSR of meetings [C] // Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. Honolulu, HI, USA, 2007: 757-760

[14] Hinton G E, Osindero S, Teh Y. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18: 1527-1554

[15] Hinton G E, Salakhutdinov R. Reducing the Dimensionality of Data with Neural Networks[J]. Science, 2006, 313(5786): 504-507

[16] Pinto J, Sivaram G S V S, Doss M M, et al. Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator[C]// IEEE Transcations on Audio, Speech, and Language Processing. 2010

[17] Grezl F, Karafiat M, Kontar S, et al. Probabilistic and Bottleneck Features for LVCSR of Meetings [C] // Proc. of IEEE Conf. Acoust. Speech Signal Process(ICASSP). 2007: 757-760

[18] The 2007 NIST Language Recognition Evaluation Plan[OL]. http://www.itl.nist.gov/iad/mig//tests/lre/2007/LRE07EvalPlan-v8b.pdf

[19] 李思一，戴蓓蒨，王海祥. 基于子带 GMM-UBM 的广播语音多语种识别[J]. 数据采集与处理，2007，22(1)：14-18