ISSN 1000-0054 清华大学学报（自然科学版）2013 年 第 53 卷 第 6 期 15/40
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2013, Vol. 53, No. 6 804-807,812

基于辨别性深度信念网络的说话人分割

马 勇[1,2]， 鲍长春[1]， 夏丙寅[1]

（1. 北京工业大学 电子信息与控制工程学院，北京 100124；2. 江苏师范大学 物理与电子工程学院，徐州 221009）

摘 要：该文基于语音信号的超矢量特征空间，提出了一种基于 Fisher 准则的可辨别性深度信念网络（discriminative deep belief network，DDBN）训练方法，得到了优于传统深度信念网络（deep belief network，DBN）的说话人码本矢量特征，并利用这些码本特征对多说话人的音段进行了聚类与分割。由 TIMIT 数据库生成的多说话人语音分割的实验结果表明，该基于 Fisher 准则函数的 DDBN 说话人分割算法的性能明显好于传统的 Bayes 信息判决（Bayesian information criterion，BIC）法和 DBN 法。

关键词：说话人分割；辨别性深度信念网络；Fisher 准则

中图分类号：TP 391.4 **文献标志码**：A

文章编号：1000-0054(2013)06-0804-04

Speaker segmentation based on discriminative deep belief networks

MA Yong[1,2], BAO Changchun[2], XIA Bingyin[1]

(1. School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China;
2. School of Physics and Electronic Engineering, Jiangsu Normal University, Xuzhou 221009, China)

Abstract: A discriminative deep belief network (DDBN) based on the Fisher criterion is used here to calculate the super-vector feature space of speech signals. The network extracts the feature codebook of the speaker that is superior to the one from the traditional deep belief network (DBN) algorithm for multi-speaker clustering and segmentation. Evaluations on the multi-speaker audio stream corpus generated from the TIMIT database show that the speaker segmentation algorithm based on the DDBN with the Fisher criterion performs better than the traditional Bayesian information criterion (BIC) method and the DBN method.

Key words: speaker segmentation; discriminative deep belief network; Fisher criterion

说话人分割研究的是如何把连续的语音分割成若干单一说话人的语音片断[1]，这项技术常用于多说话人识别和自适应语音识别等任务的前端处理。目前，说话人分割主要有基于模型的方法和基于距离度量的方法[1]，常用的距离度量有 Bayes 信息判决（Bayesian information criterion，BIC）距离和交叉似然比（cross likelihood ratio，CLR）距离等[2]。经典方法中，语音内容信息的存在影响了说话人分割的精度，文[3]提出了基于说话人因子特征的说话人分割算法，克服了以上问题，获得较好的分割效果。

2006 年，Hinton 等提出了深度信念网络（deep belief network，DBN）的学习方法[4]。DBN 是由多层受限的 Boltzmann 机（restricted Boltzmann machine，RBM）构成的复杂神经网络，它包括一个显层和若干隐层。Hinton 等首先利用无监督的预训练初始化网络参数，然后通过 BP（back propagation）算法调整网络参数。目前，DBN 广泛应用于手写体识别[4]、图像识别[5]、音频分类[6]和语音识别[7]等领域。2011 年，Chen 等首次把深度信念网络用于说话人分割的研究[8]，并取得了一定的成功。

本文研究在语音信号的超矢量特征空间中，利用 Fisher 准则构造目标函数，训练具有辨别性的 DBN，提取表征说话人因子的码本矢量特征，并利用这些码本特征分割多说话人语音。

1 辨别性深度信念网络

辨别性深度信念网络（discriminative DBN，DDBN）是一种能够提取不同类别样本区别性信息的深层信念网络结构。DDBN 的训练过程主要包括 3 个部分：预训练，调整训练，基于 Fisher 准则的辨别性训练。

收稿日期：2013-04-27
基金项目：北京市教育委员会科技计划重点项目（KZ201110005005）；
国家自然科学基金项目（61072089）
作者简介：马勇（1977—），男（汉），江苏，博士研究生。
通信作者：鲍长春，教授，E-mail：baochch@bjut.edu.cn

1.1　预训练

预训练过程就是通过无监督学习初始化网络参数的过程。Hinton 提出了逐层贪婪优化训练 RBM (restricted Boltzmann machine) 的策略。首先，训练第一层 RBM 模型。RBM 包括一个显层和一个隐层，由于 RBM 是一种能量模型[9]，因此其显层和隐层之间关系可以用能量函数表示为

$$E(v,h;\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j. \quad (1)$$

其中：v_i，h_j 分别为显层节点和隐层节点的状态，一般取 0 或 1；a_i 和 b_j 为对应的偏置，它们之间的连接权重为 w_{ij}。模型产生显层矢量的联合概率为

$$p(v;\theta) = \sum_h \mathrm{e}^{-E(v,h)} / \sum_u \sum_h \mathrm{e}^{-E(u,h)}. \quad (2)$$

显层和隐层之间的条件概率计算如下：

$$p(h_j = 1 \mid v) = \sigma\Big(\sum_{i=1}^{V} w_{ij}v_i + a_j\Big), \quad (3)$$

$$p(v_i = 1 \mid h) = \sigma\Big(\sum_{j=1}^{H} w_{ij}h_j + b_i\Big). \quad (4)$$

其中：$\sigma(x)=(1+\mathrm{e}^{-x})^{-1}$ 为 Sigmoid 函数。通过对概率的对数求偏导，可以得到 RBM 模型权重参数的更新值，

$$\Delta w = \varepsilon \frac{\partial \ln p(v)}{\partial w_{ij}} = \varepsilon(\langle v_i h_j\rangle_{\mathrm{data}} - \langle v_i h_j\rangle_{\mathrm{model}}). \quad (5)$$

ε 表示学习率，$\langle\cdot\rangle$ 表示对数据求期望。由于实际模型的无偏样本很难获得，一般采用对比散度 (contrastive divergence，CD) 方法对重构数据的采样近似来更新网络权重[9]。

第一层 RBM 训练完成后，以其隐层节点的激活概率矢量作为输入数据训练下一层的 RBM，以此类推来训练若干层的 RBM，这样就完成预训练过程。

1.2　调整训练

经过预训练之后，每层 RBM 都得到初始化的参数，把所有 RBM 按训练的顺序串联起来构成一个深层信念网络，网络包括一个显层和若干隐层。调整训练过程就是根据输入数据和重构数据的损失函数，利用 BP 算法重新调整网络的参数，最终使网络达到全局最优的过程。输入数据和重构数据的损失函数为

$$L(x,x') = \| x - x' \|_2^2. \quad (6)$$

其中：x 为输入数据，x' 为重构数据，$\|\cdot\|_2$ 表示重构误差的 2 范数形式。对误差损失函数求权重的偏导，可以获得权重的更新值。

1.3　辨别训练

DBN 能够提取表征输入数据内在结构特点的码本信息。为了提高码本的辨别能力，文[10]提出了改进的深度网络的训练方法。在此基础上，本文采用 Fisher 准则函数训练 DDBN。Fisher 准则函数是一种常用的辨别性判决函数，其主要思路是通过选择合适的投影方向，使异类样本的类间散度最大，而同类样本的类内散度最小[11]。在辨别训练过程中，针对深度网络的码本层构造 Fisher 准则函数，通过优化网络的权重等参数，使得同类样本的码本散度尽量小，而异类样本的码本散度尽量大，以此提高网络的辨别能力，如图 1 所示。

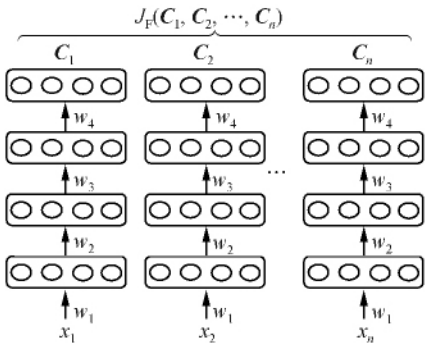

图 1　DBN 码本层的辨别训练

C_n 为 DBN 最顶层的隐层节点，它可以看作是输入数据 x_n 通过 DBN 编码得到的码本特征向量。$J_F(C_1,C_2,\cdots,C_n)$ 为利用 Fisher 准则构造的码本向量类内和类间散度矩阵的迹比准则函数，

$$J_F = \mathrm{tr}(\boldsymbol{S}^w)/\mathrm{tr}(\boldsymbol{S}^b). \quad (7)$$

其中：$\boldsymbol{S}^w$ 和 $\boldsymbol{S}^b$ 分别为码本向量的类内和类间散度矩阵，计算方法如下：

$$\boldsymbol{S}^w = 1/2\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^w (\boldsymbol{C}_i - \boldsymbol{C}_j)(\boldsymbol{C}_i - \boldsymbol{C}_j)^{\mathrm{T}}, \quad (8)$$

$$\boldsymbol{S}^b = 1/2\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^b (\boldsymbol{C}_i - \boldsymbol{C}_j)(\boldsymbol{C}_i - \boldsymbol{C}_j)^{\mathrm{T}}. \quad (9)$$

$\boldsymbol{C}_i$ 和 $\boldsymbol{C}_j$ 分别是第 i 和第 j 个输入数据的码本向量，$\boldsymbol{A}_{i,j}^w$ 和 $\boldsymbol{A}_{i,j}^b$ 分别是类内和类间码本向量之间的仿射矩阵。

J_F 的计算中需要用到网络权重参数 W，最小化式(7)得到最优权重参数 W^*，

$$W^* = \arg\min_W J_F = \arg\min_W (\mathrm{tr}(\boldsymbol{S}^w)/\mathrm{tr}(\boldsymbol{S}^b)). \quad (10)$$

对式(7)求 W 的偏导可得 $\frac{\partial J_F}{\partial \boldsymbol{W}} = \frac{\partial J_F}{\partial \boldsymbol{C}}\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{W}}$。其中，$\frac{\partial J_F}{\partial \boldsymbol{C}} =$

$\left(\mathrm{tr}(\boldsymbol{S}^{\mathrm{b}})\cdot\frac{\partial \mathrm{tr}(\boldsymbol{S}^{\mathrm{w}})}{\partial \boldsymbol{C}}-\mathrm{tr}(\boldsymbol{S}^{\mathrm{w}})\cdot\frac{\partial \mathrm{tr}(\boldsymbol{S}^{\mathrm{b}})}{\partial \boldsymbol{C}}\right)/(\mathrm{tr}(\boldsymbol{S}^{\mathrm{b}}))^2$，而$\frac{\partial \boldsymbol{C}}{\partial W}$可根据链式法则利用BP算法反向传递来计算。

2 基于辨别性深度信念网络的说话人分割

基于辨别性深度信念网络的说话人分割聚类主要包括两个过程：训练过程和分割过程。如图2所示，虚线所示路径是DDBN训练的过程，而实线所示路径是说话人分割过程。

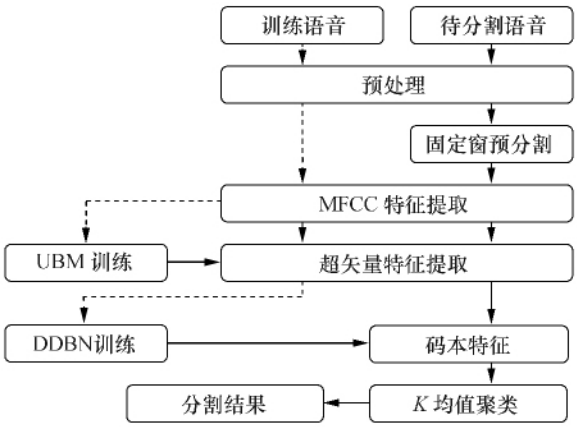

图2 基于DDBN的说话人分割

2.1 训练过程

训练过程包括两个阶段：1）通用背景模型(universal background model，UBM)训练和超矢量特征提取。训练语音经过话音激活检测(voice activity detection，VAD)等预处理之后，提取13维Mel频率倒谱系数(Mel frequency cepstral coefficient，MFCC)特征和其一阶、二阶差分，利用期望最大化(expectation maximization，EM)算法训练一个非特定人的UBM，再利用最大后验(maximum a posteriori，MAP)估计方法得到各个音段的Gauss混合模型(Gaussian mixture model，GMM)[12]。GMM是从UBM适应得到的，在实验中只对UBM的均值参数进行适应，

$$\hat{\mu}_i=\alpha_i^m E_i(x)+(1-\alpha_i^m)\mu_i. \tag{11}$$

其中：μ_i为UBM的均值参数；$\hat{\mu}_i$为适应所得GMM模型的均值参数；α_i^m为自适应系数，控制GMM和UBM之间的平衡；$E_i(x)$是训练数据均值参数的充分统计量。最后，把适应得到的每个音段GMM模型的均值拼接在一起就构成这个音段的超矢量特征。

2）利用超矢量特征训练DDBN。本文以DBN对超矢量特征编码后的码本矢量作为输入数据，以基于Fisher准则的迹比函数作为损失函数，训练能够提取说话人区别性信息的DDBN。

图3显示的是基于Fisher准则DDBN训练的收敛曲线。采用混合度为8的UBM提取TIMIT数据库语音的超矢量特征作为训练和测试数据。图3中纵轴表示经过DDBN编码之后同类样本和异类样本散度矩阵的迹比系数，系数越小，说明同类样本的码本之间越相似，而异类样本的码本之间差别越大。从图3中可以看出，随着迭代次数的增加，迹比系数不断下降，当迭代次数达到20次时，收敛速度变慢。

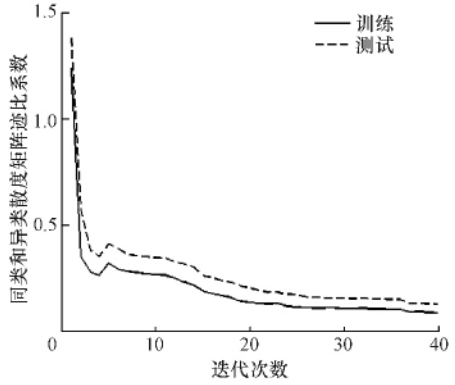

图3 Fisher辨别训练曲线

2.2 分割过程

分割过程就是根据前面训练的DDBN模型提取待分割语音的超矢量特征中的说话人因子特征，然后根据说话人的特点对音段进行聚类，从而分割出连续语音中不同说话人的语音。首先，把经过预处理的多说话人语音利用固定长度窗分割成若干语音片断，分割窗的长度为100帧，相邻两个窗之间滑动1帧。利用训练好的UBM模型提取每个预分割音段的超矢量特征，而后利用DDBN模型提取这些音段在超矢量特征空间的辨别性码本特征。然后，利用无监督的K均值聚类方法对所有预分割音段的码本特征进行聚类，以每个音段的聚类标号作为预分割音段的中间帧的类别标号。最后，根据连续语音中类别标号的变化寻找说话人的变换点，分割不同说话人的语音。

3 实 验

3.1 实验设计

本文的实验是在TIMIT语音数据库上进行的。TIMIT库包括630个说话人的语音数据。本文随机选取其中400个说话人的语音数据训练DDBN，再利用余下的数据生成多说话人语音作为测试数据。测试数据中说话人个数分别为2至9

个，每个多说话人语音材料长约 1 min，每段语音大约有 20～40 个说话人的变换点。实验中训练的 DDBN 为 4 层网络结构，网络的隐层节点数从低层到高层依次为 1 000、500、200 和 100。经过 DDBN 编码之后，输入的超矢量特征映射为 100 维的码本矢量特征。在预训练中，采用 1 步 CD 方法训练 RBM，学习率为 0.001，最大迭代次数为 100 次。调整训练最多迭代 50 次，辨别训练最大迭代次数为 20 次。本文采用虚警率(false alarm rate，FAR)和漏检率(miss detection rate，MDR)衡量说话人分割性能。其中：FAR 是指检测到的非真实变换点占检测到的所有变换点的比率；MDR 是指未检测到的变换点占所有真实变换点的比率，反映了说话人分割的准确程度。

3.2 实验结果和分析

实验 1 在两个说话人的情况下，对比了本文所提方法和基于 BIC 方法的说话人分割性能。表 1 数据显示，基于 DDBN 方法的 MDR 较小，而 FAR 偏大，这说明 DDBN 方法能够更准确地检测说话人变换点，但也产生了一些过切分的情况。

表 1 两人情况下 DDBN 和 BIC 方法的分割性能对比

说话人数目	BIC		DDBN	
	MDR/%	FAR/%	MDR/%	FAR/%
2	14.3	25.3	11.2	29.5

实验 2 在多说话人情况下，对比了 DDBN 和 DBN 对说话人分割性能，其中 UBM 模型的混合数为 8。从表 2 数据可以看出，DDBN 总体上要优于 DBN 方法，特别是在 MDR 指标上，本文方法具有明显的优势，而且随着语流中混合说话人的数目增多，基于 DDBN 方法的稳定性更好。

表 2 多人情况下 DDBN 和 DBN 方法的分割性能对比

说话人数目	DBN		DDBN	
	MDR/%	FAR/%	MDR/%	FAR/%
3	17.5	37.2	15.3	33.2
5	21.8	35.4	18.1	36.3
7	24.1	40.4	21.5	39.8
9	28.4	43.8	22.8	42.1

实验 3 本实验分析了 UBM 模型的 Gauss 混合数对 DDBN 方法分割性能的影响，实验中每个音段的混合说话人数目为 9 个，结果如表 3 所示。随着 UBM 中 Gauss 混合数的增加，说话人分割的 MDR 降低，而 FAR 也有所下降，这说明 DDBN 能够从高维的超矢量特征中提取更多的辨别性信息，因此分割的正确率更高。

表 3 UBM 模型的 Gauss 混合数对分割性能的影响

UBM 中 Gauss 混合数	DDBN 分割的错误率	
	MDR/%	FAR/%
8	22.8	42.1
16	23.1	35.9
32	21.3	37.4
64	20.7	38.5

总之，通过以上实验可看出，基于 DDBN 的说话人分割技术能够更有效地分割多说话人的语音，而且随着所用超矢量特征维数的增加，分割的效果有一定提高。

4 结 论

本文提出了基于 Fisher 准则函数的 DDBN 说话人分割算法，这种方法能有效地提取不同说话人在超矢量特征空间的区别性码本信息，从而提高说话人分割的精度。由 TIMIT 数据库生成的多说话人语音分割的实验结果表明：与经典的基于 BIC 距离的方法相比，DDBN 方法对说话人变换点的漏检率要低，而在多说话人分割任务中，DDBN 方法比 DBN 方法在漏判率和虚警率方面都具有一定的优势，而且随着说话人数的增多，本文方法性能的变化不大。另外，随着超矢量特征维数的增加，基于 DDBN 的说话人分割方法的漏检率有一定的降低。

参考文献 (References)

[1] Tranter S, Reynolds D. An overview of automatic speaker diarization systems [J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2006, **14**: 1557-1565.

[2] Kotti M, Moschou V, Kotropoulos C. Speaker segmentation and clustering [J]. *Signal Processing*, 2008, **88**: 1091-1124.

[3] Castaldo F, Colibro D, Dalmasso E, et al. Stream-based speaker segmentation using speaker factors and eigenvoices [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, NV, USA: IEEE Press, 2008: 4133-4136.

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. *Science*, 2006, **313**: 504-507.

[5] Honglak Lee, Grosse R, Ranganath R, et al. Convolution deep belief networks for scalable unsupervised learning of hierarchical representations [C]// The 26th International Conference on Machine Learning. Montreal, Canada, 2009: 609-616.

(下转第 812 页)

iVector与SVM-GSV语种识别系统上达到了较好的识别性能,而且与一对多分类方法线性融合后识别性能有明显提升。实验结果表明,该方法不失为一种有效的语种识别分类方法。

参考文献 (References)

[1] WANG Haipeng, Leung C C, Lee T, et al. Shifted-delta MLP features for spoken language recognition [J]. *Signal Processing Letters*, 2013, **20**(1): 15-18.

[2] Torres-Carrasquillo P A, Singer E, Kohler M A, et al. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features [C]// Proc ICSLP. Denver, CO, USA, 2002: 33-36.

[3] Campbell W M, Campbell J P, Reynolds D A, et al. Support vector machines for speaker and language recognition [J]. *Computer Speech and Language*, 2006, **20**(2): 210-229.

[4] LEI Yun, Hansen J H L. Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese [J]. *Audio, Speech, and Language Processing*, 2011, **19**(1): 85-96.

[5] Torres-Carrasquillo P A, Reynolds D A, Deller J R. Language identification using Gaussian mixture model tokenization [C]// 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Orlando, FL, USA: 2002, **1**: 757-760.

[6] Ronan C, Samy B. SVMTorch: Support vector machines for large-scale regression problems [J]. *Journal of Machine Learning Research*, 2001, **1**: 143-160.

[7] Cumani S, Laface P. Analysis of large-scale SVM training algorithms for language and speaker recognition [J]. *Audio, Speech, and Language Processing*, 2012, **20**(5): 1585-1596.

[8] 2011 Language Recognition Evaluation [Z/OL]. (2011-08-21). [2013-03-06]. http://www.nist.gov/itl/iad/mig/lre11.cfm.

[9] Martin A, Le A. NIST 2007 language recognition evaluation [C]//Proc of Odyssey. Stellenbosch, South Africa, 2008.

[10] Castaldo F, Colibro D, Dalmasso E, et al. Compensation of nuisance factors for speaker and language recognition [J]. *Audio, Speech, and Language Processing*, 2007, **5**(7): 1969-1978.

[11] Martinez D, Plchot O, Burget L, et al. Language recognition in iVectors space [C]// Proceedings of Interspeech. Firenze, Italy, 2011.

[12] Dehak N, Kenny P J, Dehak R, et al. Front-end factor analysis for speaker verification [J]. *Audio, Speech, and Language Processing*, 2011, **19**(4): 788-798.

[13] Cambell W M. A covariance kernel for SVM language recognition [C]// 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Las Vegas, NE, USA: 2008: 4141-4144.

(上接第807页)

[6] Lee H, Largman Y, Pham P, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks [C]// Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2009: 1-9.

[7] DONG Yu, LI Deng. Deep learning and its applications to signal and information processing [J]. *IEEE Signal Process Mag*, 2011, **28**: 145-154.

[8] Chen K, Salman A. Learning speaker-specific characteristics with a deep neural architecture [J]. *IEEE Trans Neural Networks*, 2011, **22**: 1744-1756.

[9] Hinton G E. A Practical Guide to Training Restricted Boltzmann Machines [R]. UTML Tech Report 2010-003. Toronto, Canada: Univ of Toronto, 2010.

[10] Wong W K, SUN Mingming. Deep learning regularized Fisher mappings [J]. *IEEE Trans Neural Networks*, 2011, **22**: 1668-1675.

[11] Sugiyama M. Local Fisher discriminant analysis for supervised dimensionality reduction [C]// The 23rd International Conference on Machine Learning. Pittsburgh, PE, USA, 2006: 905-912.

[12] Reynolds D, Quatieri T, Dunn R. Speaker verification using adapted Gaussian mixture models [J]. *Digital Signal Processing*, 2000, **10**: 19-41.