ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 5,Sep. 2017,pp. 921－930
DOI:10.16337/j.1004-9037.2017.05.009

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于非负矩阵分解的语音深层低维特征提取方法

秦楚雄　张连海

(解放军信息工程大学信息系统工程学院,郑州,450001)

摘　要:作为一种基于深层神经网络提取的低维特征,瓶颈特征在连续语音识别中取得了很大的成功。然而训练瓶颈结构的深层神经网络时,瓶颈层的存在会降低网络输出层的帧准确率,进而反过来影响该特征的性能。针对这一问题,本文基于非负矩阵分解算法,提出一种利用不包含瓶颈层的深层神经网络提取低维特征的方法。该方法利用半非负矩阵分解和凸非负矩阵分解算法对隐含层权值矩阵分解得到基矩阵,将其作为新的特征层权值矩阵,然后在该层不设置偏移向量的情况下,通过数据前向传播提取新型特征。实验表明,该特征具有较为稳定的规律,且适用于不同的识别任务和网络结构。当使用训练数据充足的语料进行实验时,该特征表现出同瓶颈特征几乎相同的识别性能;而在低资源环境下,基于该特征识别系统的识别率明显优于深层神经网络混合识别系统和瓶颈特征识别系统。

关键词:连续语音识别;深层神经网络;半非负矩阵分解;凸非负矩阵分解;低维特征

中图分类号:TN912.34　　**文献标志码**:A

Nonnegative Matrix Factorization Based Deep Low-Dimensional Feature Extraction Approach for Speech Recognition

Qin Chuxiong, Zhang Lianhai

(Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou, 450001, China)

Abstract: As a type of deep neural network (DNN) based low-dimensional feature, bottleneck feature (BNF) has achieved great success in continuous speech recognition. However, the existing of bottleneck layer reduces the frame accuracy of output layer when training a bottleneck deep neural network (BN-DNN), which in return has a bad impact on the performance of bottleneck feature. To solve this problem, a nonnegative matrix factorization based low-dimensional feature extraction approach using DNN without bottleneck layer is proposed in this paper. Specifically, semi-nonnegative matrix factorization and convex-nonnegative matrix factorization algorithms are applied to hidden-layer weights matrix to obtain a basis matrix as the new feature-layer weights matrix, and a new type of feature is extracted by forward passing input data without setting a bias vector in the new feature-layer. Experiments show that the feature has a relatively stable pattern around different tasks and network structures. For corpus with enough training data, the proposed features have almost the same recognition performance with conventional bottleneck feature. Under low-resource environment, the recognition accuracy of the new feature-tandem system outperforms both DNN hybrid system and bottleneck-tandem system obviously.

Key words: continuous speech recognition; deep neural network; semi-nonnegative matrix factorization; convex-nonnegative matrix factorization; low-dimensional features

---

基金项目:国家自然科学基金(61175017,61403415)资助项目。
收稿日期:2015-07-30;**修订日期**:2016-02-21

引 言

在传统的连续语音识别(Continuous speech recognition,CSR)中,利用高斯混合模型(Gaussian mixture models,GMM)和隐马尔科夫模型(Hidden Markov models,HMM)进行声学建模是当前语音识别领域最为经典并且最为成熟的建模技术,它具有速度快、复杂度较低以及识别率良好等优点。随着深度学习技术的飞速发展,深层神经网络(Deep neural network,DNN)受到了广泛的关注,并且在语音识别领域取得了巨大成功[1]。

DNN 在语音识别中的应用大致分为两类。一类是使用 DNN 进行语音的声学建模,即使用 DNN 替代 GMM 估计 HMM 的状态发射概率,构建识别率优于传统 GMM-HMM 系统的 DNN-HMM 混合系统[2];另一类是通过训练 DNN 来提取语音中更为抽象的高层特征,如瓶颈特征(Bottleneck features,BNF)[3],这类特征往往具有分布平稳、易于建模等特点,使用这些特征配合传统的 GMM-HMM 建模,可以取得足以媲美 DNN-HMM 的效果,甚至在一些情况下更加优异[4]。

由于低维特征更易于声学建模,因此基于 DNN 提取特征的关键技术在于对隐含层输出特征的降维。传统的方法是通过设置瓶颈(Bottleneck, BN)层实现数据的强制降维。Yu 等[5]提出通过在 DNN 中设置一个节点数很小的隐含层来提取 BNF,并且发现当使用三音子绑定状态作训练目标时,该方法所提取的 BNF 可以有效提升自动语音识别(Automatic speech recognition, ASR)的识别率。但是根据文献[4,5],该方法最大的问题在于 DNN 中 BN 层的存在会增大输出层的帧分类错误率。针对该问题,Gehring 等[6]提出使用 DNN 和自编码结合的方式实现降维并提取 BNF;Yan 等[7]提出使用主成分分析(Principal component analysis, PCA)对 DNN 的最后一个隐含层的输出特征进行降维得到新特征;Zhang 等[8]提出使用低秩矩阵分解(Low-rank matrix factorization)的方法对 DNN 分解权值矩阵提取 BNF 进行建模。实验证明这些方法均得到了同 DNN-HMM 系统相近的识别率。

本文提出一种基于非负矩阵分解(Non-negative matrix factorization,NMF)的降维方法。NMF 是由 Lee 和 Seung 在 1999 年[9]提出的一种矩阵分解方法,使用该算法对图像处理时可以学习到一些很好的局部特征。Ding 等[10]引入了两种基于 NMF 原理且适用于包含正负元素矩阵的分解算法——半非负矩阵分解(Semi-nonnegative matrix factorization,SNMF)和凸非负矩阵分解(Convex-nonnegative matrix factorization,CNMF)。基于 NMF 的算法对数据具有很好的解释性,在图像检索、信号分离等方面具有很成功的应用。在语音信号处理方面,NMF 在语音增强和语音去噪方面有着较为广泛的应用。2008 年,Wilson 等[11]提出使用基于先验信息的 NMF 对语音进行去噪处理,并通过实验证明 NMF 对多种噪声的去除能力强于维纳滤波器;2013 年,Mohammadiha[12]提出使用基于时域动态的 NMF 处理方法对语音进行去噪、增强,实验证明该方法优于传统的语音去噪方法。

本文提出在连续语音识别中利用 CNMF 和 SNMF 算法提取基于 DNN 的低维特征,以提升该类特征的性能。首先训练 DNN,然后对 DNN 某一层的权值矩阵进行矩阵分解,使用分解得到的基矩阵作为新的权值矩阵,从而形成新的特征提取层,且该层不设置偏移量。实验表明,使用 CNMF 所提取的特征具有比 SNMF 提取特征更为稳定的规律,两者在训练数据充足的条件下,具有和基于传统方法提取的瓶颈特征几乎相同的识别性能;而基于低资源训练数据进行实验时,两者在训练时间代价提升很小的情况下取得了优于传统瓶颈特征的识别性能,且基于 CNMF 的瓶颈特征更为出色。

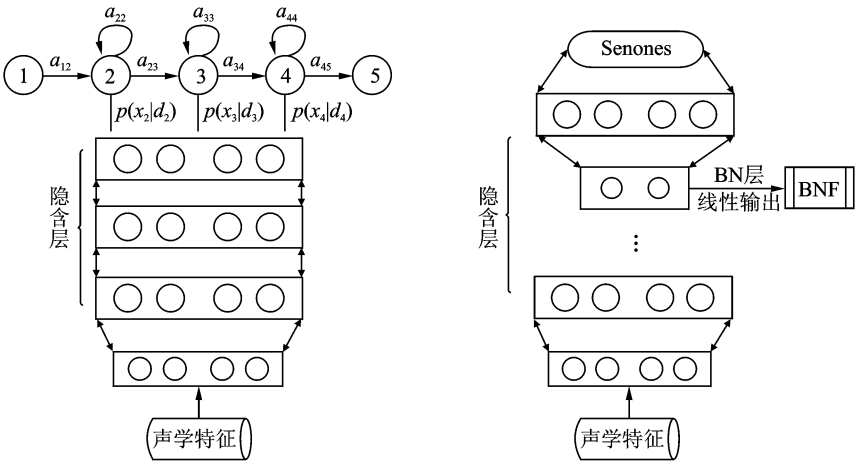
1 基于深层神经网络的声学模型和瓶颈特征提取

DNN 对训练数据要求较低,输入特征既可以是多种特征的融合特征,也可以是联合前后帧的长时语音特征。图 1 给出了 DNN 在语音识别中的两种典型应用的示意图。当前性能最优的 DNN 声学模型使用三音子绑定状态(Senones)作为训练目标,构成上下文相关深层神经网络隐马尔科夫模型(Con-

text-dependent deep neural network hidden Markov model,CD-DNN-HMM)[13],其结构如图 1(a)所示。隐含层激活函数一般使用 Sigmoid 函数,假设共有 L 个隐含层,每一层激活元输出 $\boldsymbol{x}_l$ 的表达式为

$$\boldsymbol{x}_l=\sigma(\boldsymbol{W}_l\boldsymbol{x}_{l-1}+\boldsymbol{b}_l)\quad 1\leqslant l\leqslant L \tag{1}$$

式中 $\boldsymbol{W}_l$ 和 $\boldsymbol{b}_l$ 分别为 L 层和 L-1 层之间的权值矩阵和偏移向量,激活函数定义为 $\sigma(x)=\frac{1}{1+\mathrm{e}^{-x}}$。



(a) CD-DNN-HMM的结构示意图
(a) Structural sketch of CD-DNN-HMM

(b) 基于DNN的BNF提取
(b) BNF extraction based on DNN

图 1 DNN 在语音识别中的两种典型应用

Fig. 1 Two main applications of DNN in speech recognition

在上下文相关(Context-dependent, CD)结构的 DNN 中,首先需要预先训练三音子 GMM 模型,然后采取强制对齐的方式得到 DNN 训练数据的硬性标注,使得输入的每帧声学特征与真实的类别标签信息对应起来。对于 Softmax 输出层,状态 s 的后验概率 $P(s|\boldsymbol{o}_t)$ 为

$$P(s\mid\boldsymbol{o}_t)=\frac{\mathrm{e}_s^{(\boldsymbol{W}_l\boldsymbol{x}_{l-1}+\boldsymbol{b}_l)}}{\sum_{s'}\mathrm{e}_s^{(\boldsymbol{W}_l\boldsymbol{x}_{l-1}+\boldsymbol{b}_l)}} \tag{2}$$

对 DNN 直接使用误差反向传播算法(Back propagation,BP)[14]训练易使 DNN 参数陷入局部最优解,所以往往利用受限玻尔兹曼机(Restricted Boltzmann machine,RBM)对 DNN 实施逐层预训练以得到更好的初始参数值[15,16]。用 $\boldsymbol{v},\boldsymbol{h}$ 分别表示可见层和隐含层,它们之间的能量函数定义为[2]

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=-\sum_{i=1}^{M}\sum_{j=1}^{N}v_iw_{ij}h_j+\frac{1}{2}\sum_{i=1}^{M}(v_i-b_i)^2-\sum_{j=1}^{N}h_ja_j \tag{3}$$

定义模型关于可见层节点的边缘概率为

$$P(\boldsymbol{v};\theta)=\frac{\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta))}{Z} \tag{4}$$

式中 Z 为规整因子。通过最大似然准则进行无监督学习,其中训练目标函数为

$$\tilde{\theta}=\arg\max_{\theta}\log P(\boldsymbol{v};\theta) \tag{5}$$

设 $E_{\mathrm{data}}(v_ih_j)$ 为实际数据期望值,$E_{\mathrm{model}}(v_ih_j)$ 为重建数据期望值,设 RBM 预训练时的学习速率为 $\varepsilon_{\mathrm{RBM}}$,目标函数对参数求偏导可得权值的更新式为

$$\Delta w_{ij}=\varepsilon_{\mathrm{RBM}}(E_{\mathrm{data}}(v_ih_j)-E_{\mathrm{model}}(v_ih_j)) \tag{6}$$

预训练之后,采用随机梯度下降(Stochastic gradient descent,SGD)法对训练样本之间的交叉熵(Cross-entropy)代价函数进行优化[13]。设 T 为样本总量,定义代价函数为

$$D=\sum_{t=1}^{T}\log P(s(t)\mid \boldsymbol{o}(t)) \tag{7}$$

设 ε 为学习速率，引入冲量项 α 和衰减因子 η 来控制参数更新值的波动，记 $\boldsymbol{\theta}=\{\boldsymbol{W},\boldsymbol{b}\}$ 统一表示参数，$\Delta\boldsymbol{\theta}^{(i)}$ 为第 i 轮训练参数更新值，则更新过程按式(8)进行修正。

$$\Delta\boldsymbol{\theta}^{(i+1)}=\alpha\cdot\Delta\boldsymbol{\theta}^{(i)}+(1-\alpha)\cdot\left(\varepsilon\cdot\frac{\partial D}{\partial\boldsymbol{\theta}}+\varepsilon\cdot\eta\cdot\boldsymbol{\theta}^{(i)}\right) \tag{8}$$

当使用 DNN 进行声学建模时，DNN 用来估计状态后验概率 $P(s|\boldsymbol{o}_t)$，根据贝叶斯准则，按式(9)计算 HMM 的状态发射概率，然后进行 HMM 的训练、解码。

$$P(\boldsymbol{o}_t\mid s)=\frac{P(s\mid\boldsymbol{o}_t)\cdot P(\boldsymbol{o}_t)}{P(s)} \tag{9}$$

当使用 DNN 提取瓶颈特征时，传统的方法是在 DNN 中设置瓶颈层，如图 1(b)所示。用 $\boldsymbol{F}$ 表述 BNF 向量，训练完之后，$\boldsymbol{F}$ 可表示为

$$\boldsymbol{F}=\boldsymbol{W}_{BN}^{\mathrm{T}}\boldsymbol{u}_{BN-1}+\boldsymbol{b}_{BN} \tag{10}$$

该特征配合传统的 GMM 进行建模可以取得较传统声学特征(如梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCCs)、感知线性预测(Perceptual linear predictive, PLP)系数等)更好的效果。然而，由于 BN 层的存在降低了 DNN 的分类错误率，因此该特征提取方法有待优化。

2 基于非负矩阵分解的低维特征提取

针对第 1 节引出的问题，本节将通过几种 NMF 方法对 DNN 隐含层的线性输出实现降维，以提取性能更优的低维特征，着重介绍基于两种 NMF 算法的特征提取方法。

2.1 非负矩阵分解

2001 年，Lee 和 Seung 在文献[17]中详细介绍了 NMF 分解算法。该算法针对一个全部为非负元素的矩阵，将其近似分解为一个全部为非负元素的矩阵 $\boldsymbol{F}$ 和矩阵 $\boldsymbol{G}$，即

$$\boldsymbol{X}\approx\boldsymbol{F}\boldsymbol{G}^{\mathrm{T}} \tag{11}$$

式中：$\boldsymbol{X}$ 为 $n\times m$ 的矩阵，$\boldsymbol{F}$ 为 $n\times k$ 的基矩阵，$\boldsymbol{G}$ 为 $k\times m$ 的系数矩阵。

为了得到这种分解方式，定义代价函数为 $\min\|\boldsymbol{X}-\boldsymbol{F}\boldsymbol{G}\|^2$，其中 $\|\boldsymbol{A}-\boldsymbol{B}\|^2=\sum_{ij}(\boldsymbol{A}_{ij}-\boldsymbol{B}_{ij})^2$，即两个矩阵欧式距离的平方。NMF 是针对全部元素符号为非负的矩阵分解方法，但实际中矩阵元素的符号未必遵从非负的限制，而 SNMF 和 CNMF 则是两种可以针对含有正负元素的矩阵进行分解的算法。

在 SNMF 和 CNMF 中，目标依然是找到可以通过相乘逼近待分解矩阵 $\boldsymbol{X}$ 的矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{G}$。由于 $\boldsymbol{X}$ 矩阵中的元素有正有负，因此分解时 $\boldsymbol{F}$ 矩阵中的元素符号同样有正有负，仅限定 $\boldsymbol{G}$ 矩阵中元素为非负。

首先介绍 SNMF 算法。对 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 矩阵初始化之后，采用新的准则对 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 进行更新迭代。

固定 $\boldsymbol{G}$ 矩阵不变，更新 $\boldsymbol{F}$ 矩阵为

$$\boldsymbol{F}=\boldsymbol{X}\boldsymbol{G}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1} \tag{12}$$

式中：$\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$ 为一个 $k\times k$ 且元素为正值的半正定矩阵，多数情况下 $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$ 是非奇异的，当 $\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$ 奇异时，取其伪逆。

然后固定 $\boldsymbol{F}$ 矩阵，更新 $\boldsymbol{G}$ 矩阵为

$$\boldsymbol{G}_{ik}\leftarrow\boldsymbol{G}_{ik}\sqrt{\frac{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{F})_{ik}^{+}+[\boldsymbol{G}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-}]_{ik}}{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{F})_{ik}^{-}+[\boldsymbol{G}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{+}]_{ik}}} \tag{13}$$

式中 $\boldsymbol{A}_{ik}^{+}=(|\boldsymbol{A}_{ik}|+\boldsymbol{A}_{ik})/2$，$\boldsymbol{A}_{ik}^{-}=(|\boldsymbol{A}_{ik}|-\boldsymbol{A}_{ik})/2$，分别为 $\boldsymbol{A}$ 的正分量和负分量。

CNMF 不同于 SNMF。在 CNMF 中，将基矩阵 $\boldsymbol{F}$ 定义为待分解矩阵的列的凸组合。即 $\boldsymbol{f}_l=w_{1l}\boldsymbol{x}_1+\cdots+w_{nl}\boldsymbol{x}_n$，或写为 $\boldsymbol{F}=\boldsymbol{X}\boldsymbol{W}$。根据文献[10]，因子矩阵 $\boldsymbol{W}$ 和系数矩阵 $\boldsymbol{G}$ 具有稀疏的性质。

CNMF 的初始化分为两种方法,一是基于 K-means 聚类方法,二是基于已有 NMF 解或 SNMF 解的初始赋值方法,本文选用 K-means 方法。首先对待分解矩阵 $\boldsymbol{X}$ 作一次 K-means 聚类,得到隶属度矩阵 $\boldsymbol{H}=(h_1,\cdots,h_k)$,$\boldsymbol{H}_{ik}=\{0,1\}$,基于 $\boldsymbol{H}$ 对 $\boldsymbol{G}$ 矩阵初始化,即

$$\boldsymbol{G}^{(0)}=\boldsymbol{H}+0.2\boldsymbol{E} \tag{14}$$

式中 $\boldsymbol{E}$ 为全 1 矩阵。使用聚类的类心矩阵作为 $\boldsymbol{F}$ 矩阵,即

$$\boldsymbol{F}=\boldsymbol{X}\boldsymbol{H}\boldsymbol{D}_n^{-1} \tag{15}$$

式中 $\boldsymbol{D}_n=\mathrm{diag}(n_1,\cdots,n_k)$。根据 $\boldsymbol{F}=\boldsymbol{X}\boldsymbol{W}$ 与式(15),得 $\boldsymbol{W}=\boldsymbol{H}\boldsymbol{D}_n^{-1}$,但为了平滑处理,设 $\boldsymbol{W}^{(0)}=(\boldsymbol{H}+0.2\boldsymbol{E})\boldsymbol{D}_n^{-1}$。

初始化之后进行迭代运算,更新 $\boldsymbol{G}$ 的值为

$$\boldsymbol{G}_{ik}\leftarrow\boldsymbol{G}_{ik}\sqrt{\frac{\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{+}\boldsymbol{W}\right]_{ik}+\left[\boldsymbol{G}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-}\boldsymbol{W}\right]_{ik}}{\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-}\boldsymbol{W}\right]_{ik}+\left[\boldsymbol{G}\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{+}\boldsymbol{W}\right]_{ik}}} \tag{16}$$

更新 $\boldsymbol{W}$ 的值为

$$\boldsymbol{W}_{ik}\leftarrow\boldsymbol{W}_{ik}\sqrt{\frac{\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{+}\boldsymbol{G}\right]_{ik}+\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-}\boldsymbol{W}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right]_{ik}}{\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-}\boldsymbol{G}\right]_{ik}+\left[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{+}\boldsymbol{W}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}\right]_{ik}}} \tag{17}$$

2.2 基于凸非负矩阵分解和半非负矩阵分解的低维深层特征提取

对于一个不包含 BN 层的 DNN,它的第 l 隐含层的线性输出具有维数大、相关性大的特点,不适合直接进行 GMM 的声学建模,因此需要对 DNN 特征降维。若直接利用矩阵分解算法对 DNN 特征数据作降维,理论上并不好,其原因是,首先无法针对一帧特征向量做矩阵分解变换;其次,当通过组合多帧特征形成特征矩阵时,矩阵分解会产生结构性的变化,进而破坏语音特征的时序信息,也就无法训练出良好的声学模型。

由于无法直接对隐含层的线性输出作变换,因此本文采用一种间接降维的方法。在计算 DNN 隐含层的线性特征时,层与层之间的权值矩阵作用于每一帧特征,权值矩阵可以看作是一种广义的映射函数,具有一定的整体分布性。又由于权值矩阵元素有正有负,因此适合于作矩阵分解。而由于同一层的偏移向量和权值矩阵并没有整体联系,因此很难对偏移向量实施与权值矩阵相同的操作。本方法中,对权值矩阵分解的目的在于实现矩阵列的降维,进而实现对特征数据的降维。而低资源数据条件下训练的 DNN 隐含层权值矩阵具有收敛极不充分的特点,因此该矩阵对特征数据的线性变换效果很有限。针对低资源训练的权值矩阵中存在大量冗余项,可以通过矩阵分解算法从矩阵中提取更基本的分类变换信息。一些常用的矩阵分解方法对于本方法而言具有较大的局限性。首先,特征值分解只能针对方阵,且分解得到的矩阵与原矩阵尺寸相同,因此无法实现降维;其次,奇异值分解将矩阵分解为左奇异分量、奇异值矩阵和右奇异分量,它们都包含了原矩阵中很多的有效信息,无论使用哪一个分量作为新的特征矩阵,浪费的矩阵分量都过多。而对于 SNMF 和 CNMF,根据式(11),所分解得到的基矩阵包含正负元素,它作为主分量包含了原权值矩阵中的主要分类信息,适合构建新的特征提取层,而系数矩阵作为次分量仅包含非负元素,不具备形成权值矩阵的条件。当然,系数矩阵包含对基矩阵的组合系数,所以一定包含一些有效信息,但是对本文所提出的特征提取过程的贡献不大,所以在此舍弃了系数矩阵的使用。

综上所述,基于 SNMF 和 CNMF 两种算法提取深层低维特征的方法如图 2 所示。首先对某一层的 $n\times m$ 权值矩阵分解,得到 $n\times r$ 的基矩阵和 $r\times m$ 的系数矩阵,然后使用包含正负元素的基矩阵作为新的权值矩阵,形成新的特征提取层,并提取维数为 r 的低维特征。设待分解的权值矩阵为 $\boldsymbol{W}$,经过 $\boldsymbol{W}=\boldsymbol{W}'\boldsymbol{G}^{\mathrm{T}}$ 的分解,计算出新特征 $\boldsymbol{F}'$,即

$$\boldsymbol{F}'=\boldsymbol{W}'^{\mathrm{T}}\boldsymbol{U} \tag{18}$$

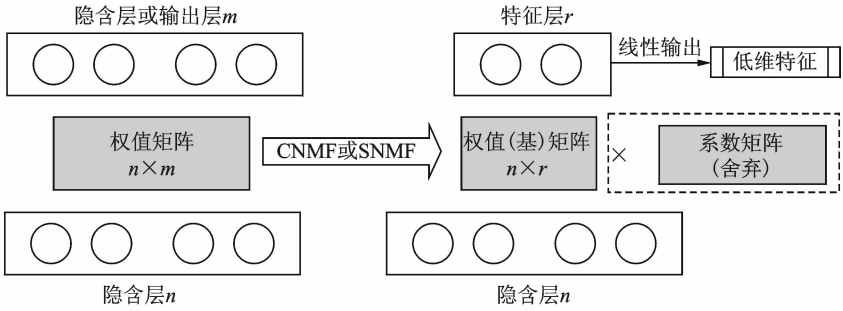


图 2 基于 SNMF/CNMF 的低维特征提取方法

Fig. 2 SNMF/CNMF based low-dimensional feature extraction approach

式中 $\boldsymbol{U}$ 为上一隐含层的激活元输出，不设置特征层偏移量，因此计算不包含该项。

SNMF 和 CNMF 算法对含有正负元素的矩阵进行分解时往往可以得到更好的数据解释性。它们在算法性能上各有千秋，首先，使用 SNMF 比使用 CNMF 的方法得到解的精度更高一些；其次，CNMF 可以得到比 SNMF 更稀疏的解；最后，CNMF 的解比 SNMF 的解具有更好的正交性。进一步进行推导，在 CNMF 中，设 $\boldsymbol{H}$ 为因子矩阵(因子矩阵在 2.1 中用 $\boldsymbol{W}$ 表示，这里为了与权值矩阵区分，用 $\boldsymbol{H}$ 表示)，则有 $\boldsymbol{W}=\boldsymbol{W}'\boldsymbol{H}$，因此式(18)可写为

$$\boldsymbol{F}'=\boldsymbol{H}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{U} \tag{19}$$

可知，基于 CNMF 的方法相当于对权值矩阵作了一次列的线性变换，比基于 SNMF 的方法更符合权值矩阵降维的实际要求，又因为它的初始化是基于 K-means 聚类而不是随机初始化，因此 CNMF 低维特征理应具有更稳定的性能。

3 实验部分

3.1 实验语料

采用训练数据充足的语料和训练数据不足的低资源语料两种语料进行实验，分别测试本文所提出的特征在两种环境下的识别性能表现。一是 RM 语料，RM 语料库是由美国国防部高级研究项目局(Defense advanced research projects agency，DARPA)牵头收集定制的较为早期的英语语料库，语料经过数字采样和文本标注，专门用于设计和评估连续语音识别系统。RM 语料模拟训练语料较为充足的情况。二是 Vystadial 2013 Czech data(Vystadial_cz)，它是开源的捷克语语料库，全部时长 15 h，来源于 3 类数据：Call Friend 电话服务的语音数据、Repeat After Me 的语音数据和 Public Transport Info 的口语对话系统的语音数据。实验随机选取 Vystadial_cz 语料库中约 1 h(1.06 h)的训练语音文件组成训练集，以模拟低资源的条件；再选取语料库中约 30 min 的测试语音数据作为测试集，测试集包含 666 句话，共 3 910 个待识别词；使用 Vystadial_cz 语料库中全部训练语料的标注文本训练语音识别系统的二元语言模型。

3.2 实验工具与评价指标

实验使用 Kaldi 工具包[18]进行数据准备、底层声学特征和高层声学特征的提取、语言模型的声学模型的训练与解码；使用 PDNN 工具包[19]基于 GPU(Quadro 600)进行相关的 DNN 的搭建与训练；使用 PYMF 工具包[20]实现 SNMF 和 CNMF 等分解算法。

实验评价指标采用连续语音识别中的词错误率(Word error rate，WER)，设 N 为语料库人工标注文本中词(全部正确词)的数量，W 为解码连续语音与人工标注作对比统计出的插入词、删除词和替代词

的个数,r 表示 WER,将 WER 定义为两者的比值,并化为百分率,即

$$r=\frac{W}{N}\times 100\% \tag{20}$$

3.3 基线系统

首先对于 RM 语料,基于 13 维的 MFCC 特征,训练一个基于线性判别分析(Linear discriminant analysis,LDA)(9 帧拼接,LDA 降到 40 维)、最大似然线性变换(Maximum likelihood linear transform,MLLT)和说话人自适应训练(Speaker adaption training,SAT)且高斯混元数为 9 000 的三音子 GMM 声学模型。根据该模型提取 40 维的特征空间最大似然线性回归(feature-space Maximum likelihood linear regression,fMLLR)特征,再通过前后 5 帧的拼接得到 DNN 的输入特征,这样一来 DNN 的输入层节点数为 440 个。之后使用该模型中的 senones 强制对齐得到 DNN 的硬性标注,实验中 DNN 的 Softmax 层一共有 1 487 个节点。基于 RM 语料设置两个提供对比的基线系统。第一个是基于 CD-DNN-HMM 的识别系统,其 DNN 结构设置为"440-1 024-1 024-1 024-1 024-1 024-1 024-1 487";第二个是基于 BNF 训练的子空间高斯混合模型(Subspace Gaussian mixture models,SGMM)的识别系统。经多次实验表明,BN-DNN 设置为"440-1 024-1 024-1 024-1 024-40-1 024-1 487"时 BNF 的性能最佳,该系统声学模型为基于 LDA(9 帧拼接后降至 40 维)、MLLT 且高斯混元数为 9 000 的三音子 GMM,经过强制对齐后,训练高斯混元数为 400 的通用背景模型(Universal background model,UBM),然后训练子状态数为 9 500的 SGMM。

对于 Vystadial_cz 的低资源语料,其训练过程类似于 RM 识别系统。首先基于 13 维的 MFCC 特征训练一个基于 LDA(9 帧拼接,降至 40 维),MLLT 和 SAT 且高斯混元数为 22 000 的三音子 GMM,同样基于此模型提取 40 维的 fMLLR 特征,将其进行 11 帧拼接,作为 DNN 的输入。Vystadial_cz 的基线系统同样为 CD-DNN-HMM 和 BNF-SGMM。对于 CD-DNN-HMM,DNN 的结构设置为"440-1 024-1 024-1 024-1 024-1 024-915";对于 BNF-SGMM 系统,同样经实验验证 DNN 设置为"440-1 024-1 024-1 024-40-1 024-915"时最佳,然后提取 BNF 训练基于 LDA(9 帧拼接后降至 40 维),MLLT 且高斯混元数为 22 000 的三音子 GMM,经过强制对齐后,训练高斯混元数为 400 的 UBM,并训练子状态数为 5 000的 SGMM。

所有 DNN 的训练参数设置均相同。设置学习速率初始值为 0.08,每当相邻两轮训练的验证误差小于 0.2%时,就将学习速率衰减一半,当衰减之后相邻两轮的验证误差再次小于 0.2%时训练停止(如果一直大于 0.2%,则最多衰减 8 次);冲量值设为 0.5;minibatch 尺寸设为 256。

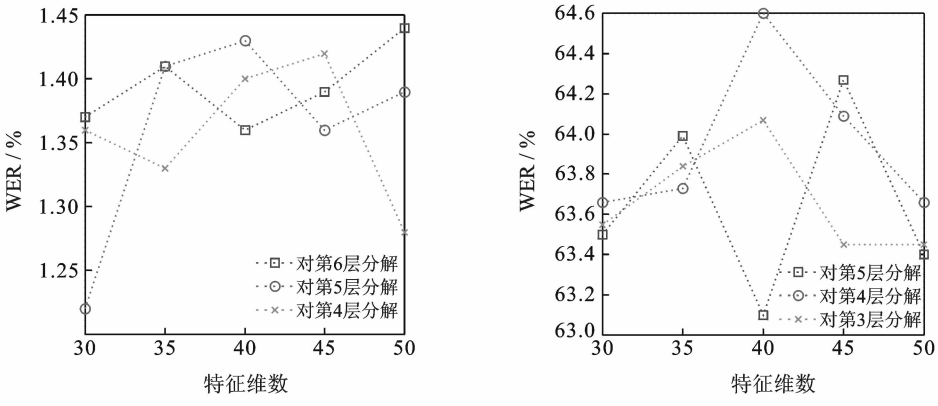
3.4 对比实验

基于基线系统的 DNN 模型,对 DNN 的权值矩阵做 CNMF 和 SNMF。根据第 2 节中的算法,通过 50 轮 K-means 聚类方法对 CNMF 进行初始化,通过随机赋值的方法对 SNMF 进行初始化。两种方法都作 500 次迭代训练从而实现分解。将该低维特征简记为其英文缩写 LDF(Low-dimensional feature),则实验中的两种特征分别记为 CNMF-LDF 和 SNMF-LDF。

使用新的特征训练 SGMM 声学模型并解码,其训练过程、参数设置与基线系统相同。对于分解过程来说,针对 DNN 的哪一层进行分解、分解维数为多少,这两个变量因素都会影响所提取特征的性能。设第 1 个隐含层与第 2 个隐含层之间的权值矩阵所在位置称为第 1 层,对于 RM 语料的 DNN 而言,第 6 个隐含层与输出层间的权值矩阵所在位置为第 6 层,共 6 个待分解位置;对于 Vystadial_cz 语料的 DNN 而言,共 5 个待分解的位置。由于较高层的输出特征明显优于较低层的特征,因此本实验针对后 3 个待分解位置进行对比;此外,根据经验,实验主要针对分解维数为 30,35,40,45,50 的情况进行研究。

对这两个变量研究的实验结果分别如图 3,4 所示。

从图 3(a,b)中可知,使用 SNMF 时,对最后一层分解 40 维、对倒数第二层分解 30 维、对倒数第 3 层分解 50 维时效果相对更好,但是最好的分解位置和维数并不随网络结构的变化而具有固定的规律;由图 4(a,b)可知,使用 CNMF 时,对每一层普遍进行 40 维分解时效果更好,且对倒数第 2 层权值矩阵分解 40 维时可以取得最好的效果。总体来说,CNMF-LDF 比 SNMF-LDF 具有更稳定的建模规律,其主要原因在于 CNMF 通过 50 轮的 *K*-means 聚类进行初始化而 SNMF 通过随机赋值进行初始化。而在 40 维左右时取得最好的性能这一结果,说明 CNMF 起到了与传统方法中 BN 层相类似的作用,即对输入特征实现了非线性压缩。至于在倒数第 2 层取得最优性能这一结果,说明该特征和 BNF 有类似的结论,都是将特征层置于 DNN 中间靠后的隐含层时效果最佳。

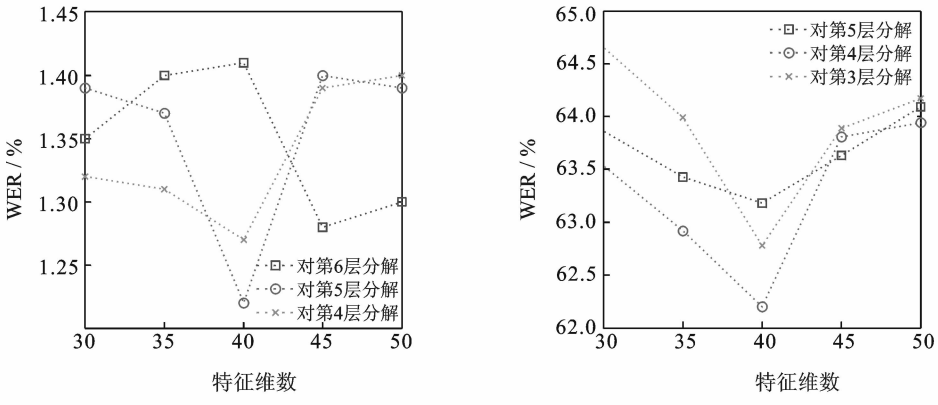


(a) 基于RM语料的实验 (b) 基于1 h Vystadial_cz语料的实验

(a) Experiments on RM (b) Experiments on 1 h Vystadial_cz

图 3 不同分解位置和维数对 SNMF 特征的影响

Fig. 3 Effects of different factorization locations and dimensions on SNMF



(a) 基于RM语料的实验 (b) 基于1 h Vystadial_cz语料的实验

(a) Experiments on RM (b) Experiments on 1 h Vystadial_cz

图 4 不同分解位置和维数对 CNMF 特征的影响

Fig. 4 Effects of different factorization locations and dimensions on CNMF

基于前文对参数的研究,实验选取 SNMF 与 CNMF 中的最优结果与基线系统进行对比,结果如表 1 所示。

由表 1 结果可知,在训练语料较为充足的 RM 实验中,CNMF-LDF 与 SNMF-LDF 可以在建模时取得相同的识别率,但两者的性能与传统方法所提取的特征几乎相同。

而文献[21]表明,引入 dropout 训练技术可有效提高低资源下的 DNN 训练效果。在此基于前一实验设置,对 DNN 的隐含层引入 dropout 进行训练,以进一步验证该方法对于 DNN 的有效性。实验中,保持 DNN 的其他训练参数设置和不引入 dropout 时一致,仅仅在每一轮的训练中,对每个隐含层引入遗弃因子(Hidden drop factor,HDF),根据文献[21]的经验,将该值设置为 0.2,即每一轮训练中,每一个隐含层都有约 20%的节点参数不参与本次更新。各系统的实验结果如表 2 所示。由表 2 可知,在训练语料不足的条件下,两种特征均优于传统的 BNF,且 CNMF-LDF 的识别性能更突出一些,它相对 CD-DNN-HMM 提高了 7.4%(67.16%→62.20%),相对基线 BNF-SGMM 提高了 2.0%(63.45%→62.20%)。当引入 dropout 技术提升 DNN 性能时,仍然得到了类似的对比结果,SNMF-LDF 和 CNMF-LDF 依然优于 BNF 的识别性能,且基于 CNMF-LDF 的识别系统依然取得了最好的识别率,分别相对基线 CD-DNN-HMM 和 BNF-SGMM 提高了 5.6%(65.32%→61.66%)和 1.7%(62.73→61.66%)。基于以上实验结果,本文认为使用 CNMF 提取特征为更优的方法。

表 1 在 RM 语料下不同系统的对比结果

Tab. 1 Recognition results of different systems for RM

语音识别系统	WER/%
CD-DNN-HMM	1.70
BNF-SGMM	1.22
CNMF-LDF-SGMM	1.22
SNMF-LDF-SGMM	1.22

表 2 基于 dropout 训练的不同低资源识别系统的对比结果

Tab. 2 Recognition results of different low-resource recognition systems based on dropout training

语音识别系统	WER/%	
	DNN	Dropout-DNN
CD-DNN-HMM	67.16	65.32
BNF-SGMM	63.45	62.73
CNMF-LDF-SGMM	62.20	61.66
SNMF-LDF-SGMM	63.10	62.14

分析认为,权值矩阵训练充分与否直接影响了对输入特征的变换能力的好坏。当训练数据充足时,传统方法中 BN 层的权值矩阵收敛充分,因此 BN 层对输入数据可以进行充分的非线性变换,从而提取性能出色的 BNF,此时使用基于 NMF 的方法体现不出优越性;当训练数据不足时,DNN 的权值矩阵训练不足,BN 层对输入特征的变换不充分,而此时使用 NMF 算法可以从高维权值矩阵提取出更为本质的低维矩阵,相当于对权值矩阵实施了进一步的训练,因此该方法体现出了比传统方法更优越的性能,尤其当使用聚类性能更好的凸非负矩阵分解时,其基矩阵具有更好的稀疏特性和更好的变换效果,因此所提取的特征性能更出色。而在低资源下,SNMF-LDF、CNMF-LDF 基于 DNN 和 Dropout-DNN 都取得了最好的识别结果,也说明了本文方法对于低资源数据条件下多种结构 DNN 的适用性。

通过实验验证,本文所提出低维特征与 BNF 在模型训练和解码时所用时间基本相同,主要的时间代价差别存在于特征提取过程。虽然从网络参数规模的角度来说,设置 BN 层的 DNN 参数比不设置 BN 层的 DNN 参数更少一些,但在训练数据量很小的低资源情况下,两者在训练时间上几乎没有差别(尤其是使用高性能 GPU 时),加之非负矩阵分解的迭代训练速度很快,因此该方法和设置 BN 层的方法在训练时间代价上相差无几。本文所提出的方案以微小的时间代价换取了可观的识别性能提升。

4 结束语

在连续语音识别中利用瓶颈特征进行高斯混元建模的过程中,针对 DNN 提取瓶颈特征时,设置 BN 层会影响 DNN 的训练效果这一问题,本文提出一种新的利用 DNN 提取特征的方法。使用已训练好的 DNN 模型,基于 CNMF 和 SNMF 算法对隐含层权值矩阵做分解,将得到的基矩阵作为新的特征提取层的权值矩阵,在不设置偏移量的前提下,提取新的低维特征。实验表明,CNMF 的方法比 SNMF

的方法具有更为稳定的规律，两种方法在训练数据充足的情况下，取得了同基线系统相同的性能；而在权值矩阵得不到充分训练的低资源语料环境下，这两种方法均以极小的时间代价换取了相对传统方法的识别率的提升，且 CNMF 的效果更为明显。

参考文献：

[1] 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 Process Mag, 2012, 29(6): 82-97.

[2] 戴礼荣，张仕良. 深度语音信号与信息处理：研究进展与展望[J]. 数据采集与处理，2014，29(2)：171-179.
Dai Lirong, Zhang Shiliang. Deep speech signal and information processing: Research progress and project[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 171-179.

[3] Grézl F, Karafiat M, Kontar S, et al. Probabilistic and bottle-neck features for LVCSR of meetings [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu, HI: IEEE, 2007: 757-760.

[4] Bao Yebo, Jiang Hui, Dai Lirong, et al. Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC: IEEE, 2013: 6980-6984.

[5] Yu D, Seltzer M L. Improved bottleneck features using pretrained deep neural networks[C]//Proc Interspeech. Florence, Italy: International Speech Communication Association, 2011: 237-240.

[6] Gehring J, Miao Y J, Metze F, et al. Extracting deep bottleneck features using stacked auto-encoders [C]//IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC: IEEE, 2013: 3377-3381.

[7] Yan Z J, Huo Q, Xu J. A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR [C]//In Proc INTERSPEECH. Lyon, France: International Speech Communication Association, 2013: 104-108.

[8] Zhang Y, Chuangsuwanich E, Glass J R. Extracting deep neural network bottleneck features using low-rank matrix factorization [C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 185-189.

[9] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401(6755): 788-791.

[10] Ding C, Li T, Jordan M I. Convex and semi-nonnegative matrix factorizations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(1): 45-55.

[11] Wilson K W, Raj B, Smaragdis P, et al. Speech denoising using nonnegative matrix factorization with priors[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008: 4029-4032.

[12] Mohammadiha N. Speech enhancement using nonnegative matrix factorization and hidden markov models[D]. Stockholm: KTH Royal Institute of Technology, 2013.

[13] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition [J]. IEEE Transactions on Audio, Speech and Language Processing, 2012, 20(1): 30-42.

[14] Rumelhart D E, Hinton G E, William R J. Learning representations by back-propagating errors[J]. Cognitive Modeling, 2002, 1: 213.

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

[16] Mohamed A, Dahl G E, Hinton G. Acoustic modeling using deep belief networks [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 14-22.

[17] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2000: 556-562.

[18] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit [C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Waikoloa, Hawaii, USA: IEEE, 2011: 1-4.

[19] Miao Y J. Kaldi+PDNN: Building DNN-based ASR systems with Kaldi and PDNN [J]. Eprint Arxiv, 2014: 1401.6984.

[20] Thurau C. Python matrix factorization module[EB/OL]. https://pypi.python.org/pypi/PyMF/0.1.9, 2011-04-28.

[21] Miao Y J, Metze F. Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training [C]//Proc Interspeech. Lyon, France: International Speech Communication Association, 2013: 2237-2241.

作者简介：

秦楚雄（1991-），男，硕士研究生，研究方向：智能信息处理与语音信号处理，E-mail：qinchuxiong911@163.com。

张连海（1971-），男，副教授，研究方向：语音信号处理、语音识别。