基于 DNN 的低资源语音识别特征提取技术

秦楚雄 [1] 张连海 [1]

摘 要 针对低资源训练数据条件下深层神经网络 (Deep neural network, DNN) 特征声学建模性能急剧下降的问题, 提出两种适合于低资源语音识别的深层神经网络特征提取方法. 首先基于隐含层共享训练的网络结构, 借助资源较为丰富的语料实现对深层瓶颈神经网络的辅助训练, 针对 BN 层位于共享层的特点, 引入 Dropout, Maxout, Rectified linear units 等技术改善多流训练样本分布不规律导致的过拟合问题, 同时缩小网络参数规模、降低训练耗时; 其次为了改善深层神经网络特征提取方法, 提出一种基于凸非负矩阵分解 (Convex-non-negative matrix factorization, CNMF) 算法的低维高层特征提取技术, 通过对网络的权值矩阵分解得到基矩阵作为特征层的权值矩阵, 然后从该层提取一种新的低维特征. 基于 Vystadial 2013 的 1 小时低资源捷克语训练语料的实验表明, 在 26.7 小时的英语语料辅助训练下, 当使用 Dropout 和 Rectified linear units 时, 识别率相对基线系统提升 7.0 %; 当使用 Dropout 和 Maxout 时, 识别率相对基线系统提升了 12.6 %, 且网络参数数量相对其他系统降低了 62.7 %, 训练时间降低了 25 %. 而基于矩阵分解的低维特征在单语言训练和辅助训练的两种情况下都取得了优于瓶颈特征 (Bottleneck features, BNF) 的识别率, 且在辅助训练的情况下优于深层神经网络隐马尔科夫识别系统, 提升幅度从 0.8 %～3.4 % 不等.

关键词 低资源语音识别, 深层神经网络, 瓶颈特征, 凸非负矩阵分解

引用格式 秦楚雄, 张连海. 基于 DNN 的低资源语音识别特征提取技术. 自动化学报, 2017, **43**(7): 1208−1219

DOI 10.16383/j.aas.2017.c150654

Deep Neural Network Based Feature Extraction for Low-resource Speech Recognition

QIN Chu-Xiong[1] ZHANG Lian-Hai[1]

Abstract To alleviate the performance degradation that deep neural network (DNN) based features suffer from transcribed training data is insufficient, two deep neural network based feature extraction approaches to low-resource speech recognition are proposed. Firstly, some high-resource corpuses are used to help train a bottleneck deep neural network using a shared-hidden-layer network structure and dropout, maxout, and rectified linear units methods are exploited in order to enhance the training effect and reduce the number of network parameters, so that the overfitting problem by irregular distributions of multi-stream training samples can be solved and multilingual training time can be reduced. Secondly, a convex-non-negative matrix factorization (CNMF) based low-dimensional high-level feature extraction approach is proposed. The weight matrix of hidden layer is factorized to obtain the basis matrix as the weight matrix of the newly formed feature-layer, from which a new type of feature is extracted. Experiments on 1 hour′s Vystadial 2013 Czech low-resource training data show that with the help of 26.7 hours′ English training data, the recognition system obtains a 7.0 % relative word error rate reduction from the baseline system when dropout and rectified linear units are applied, and obtains a 12.6 % relative word error rate reduction while reduces 62.7 % relative network parameters and 25 % training time as compared to other proposed systems when dropout and maxout are applied. Matrix factorization based features perform better than bottleneck features (BNF) in both low-resource monolingual and multilingual training situations. They also gain better word accuracies than the state-of-art deep neural network hidden Markov models hybrid systems, by from 0.8 % to 3.4 %.

Key words Low-resource speech recognition, deep neural network (DNN), bottleneck features (BNF), convex-nonnegative matrix factorization (CNMF)

Citation Qin Chu-Xiong, Zhang Lian-Hai. Deep neural network based feature extraction for low-resource speech recognition. *Acta Automatica Sinica*, 2017, **43**(7): 1208−1219

在训练样本充足的大词汇量连续语音识别 (Large vocabulary continuous speech recognition, LVCSR) 中, 使用传统的声学特征训练高斯混合模型

收稿日期 2015-10-16 录用日期 2016-10-20
Manuscript received October 16, 2015; accepted October 20, 2016
国家自然科学基金 (61673395, 61302107, 61403415) 资助
Supported by National Natural Science Foundation of China (61673395, 61302107, 61403415)
本文责任编委 贾珈
Recommended by Associate Editor JIA Jia

1. 信息工程大学信息系统工程学院 郑州 450001
1. Department of Information and System Engineering, Information Engineering University, Zhengzhou 450001

–隐马尔科夫模型 (Gaussian mixture models hidden Markov models, GMM-HMM) 搭建识别系统可以取得良好的识别率. 然而由于 GMM-HMM 是基于最大似然准则 (Maximum likelihood estimation, MLE) 进行训练的, 因此当特征的分布不平稳或者较为复杂时, 所需建模参数会增多, 在理论上需要使用大量的样本进行训练才能取得良好的效果, 所以在训练数据有限的低资源语音识别任务中, 使用传统声学特征训练 GMM-HMM 的方法并不可行[1].

基于深层神经网络 (Deep neural network, DNN) 模型提取的特征往往具有分布平稳、易于建模等特点, 典型的是瓶颈特征 (Bottleneck features, BNF), 这种使用 DNN 作为特征提取模块并使用 GMM-HMM 进行声学建模所构成的系统称为级联 (Tandem) 系统, Tandem 系统明显优于使用传统特征训练的 GMM-HMM 识别系统[2−3], 在 LVCSR 任务中它可以取得足以媲美深层声学模型 DNN-HMM 的性能, 甚至在一些情况下更加优异[4], 使用 DNN 提取特征具有一定优势, 它可以联合特征的上下文 (Context) 信息形成长时特征矢量, 并且具有深层次的非线性变换能力, 因此 DNN 能够从有限的数据中挖掘出更多的信息[4−5].

然而在低资源条件下, DNN 无法通过有限的训练样本得到有效的训练, 因此所提取特征的性能自然会受到影响. 针对该问题, 研究者们陆续提出了一些强化 DNN 特征提取模块的方法. Lal 等[6] 提出一种通过提取辅助语料 Tandem 特征进行低资源跨语言的声学建模方法, 实验表明新的系统相比于 MFCC 特征的基线系统在识别率方面有了显著提升; Veselý 等[7] 和 Tüske 等[8] 均提出了使用多层感知器 (Multi-layer perceptron, MLP) 对具有相同音素集的多语言提取 BNF 的方法, 并通过实验证明该方法取得了优于单语言训练的效果; Gehring 等[9] 提出使用基于多语言共享隐含层 (Shared-hidden-layer, SHL) 结合自编码 (Autoencoder) 技术提取特征, 该特征在 Tandem 系统和 DNN-HMM 混合系统中均表现出优异的性能; Miao 等[10] 提出使用共享隐含层多语言 DNN (Shared-hidden-layer multilingual deep neural network, SHL-MDNN) 结合卷积神经网络 (Convolutional neural network, CNN) 提取上千维的高维卷积网络神经元输出作为特征, 实验证明该特征优于同维数的 DNN 特征. 改善 DNN 特征提取模块的研究有很多, 但上述研究中仅有少数是针对低资源的情况.

鉴于此, 本文提出两种方法对低资源环境下的 DNN 特征提取过程进行改进.

第一种方法从提高训练效果的角度出发, 提出一种基于 SHL 结构的改进的 BN-DNN 特征提取模型. 不同于一般的 SHL 多语言训练过程, 由于此时辅助语料的数量远多于低资源目标语料, 且 BN 层位于共享层, 因此容易出现训练不平衡的现象. 本文提出借助 Dropout 技术的子模型平均原理, 降低对某类特征的过拟合程度, 改善多语言训练效果, 并使用 Maxout, ReLU (Rectified linear units) 替代传统的 Sigmoid 激活函数, 在最大化 Dropout 训练效果的同时, 降低训练时间. 实验表明, 当加入一定辅助语料时, 该特征的性能明显优于单语言训练得到的特征; 当引入 Dropout, Maxout 和 ReLU 改进技术后, 特征性能得到较明显的提升, 训练效果得到进一步改善, 训练时间得到一定降低.

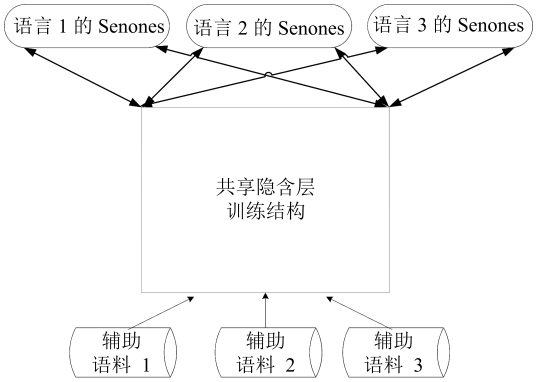
第二种方法从改善 DNN 特征性能的角度出发, 提出一种基于矩阵分解算法的低维高层特征提取方法. 传统的通过设立 BN 层提取 DNN 特征 (BNF) 的方法存在一个缺陷, 即 BN 层的存在降低了 DNN 的分类准确率, 因此 BNF 并不能充分体现 DNN 的性能. 对此, 本文提出一种 “先训练、后降维” 的思想对该问题进行改善. 具体来说, 通过使用一种凸非负矩阵分解 (Convex-nonnegative matrix factorization, CNMF) 的算法对 DNN 某一层的权值矩阵进行分解, 得到基矩阵作为特征层的权值矩阵, 在不设立偏移量的情况下从该层提取线性输出作为一种新的低维特征. 在两种不同语言的低资源实验中, 该特征均取得了优于传统 BNF 的识别率, 且具有稳定的规律. 当结合 SHL-MDNN 提取特征时, 该特征所训练的 Tandem 系统的识别率在某些实验中优于 BNF-tandem 系统和 DNN-HMM 系统.

本文组织结构如下: 第 1 节介绍基于 SHL 的 BN-DNN 训练与提取特征的原理; 第 2 节介绍基于 CNMF 算法的特征提取方法; 第 3 节介绍实验设置以及结果分析; 第 4 节为本文的结论部分.

1 一种改进的基于 SHL 结构的 BN-DNN

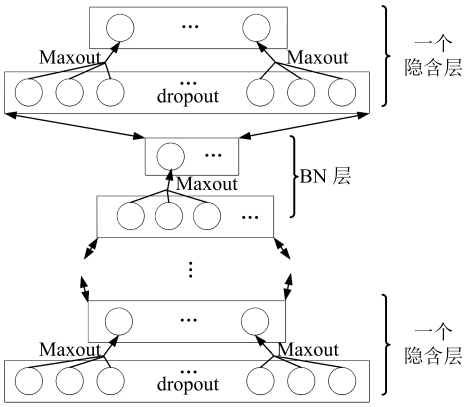
SHL-MDNN 是 Huang 等[11] 提出的一种较为新颖的多语言训练网络结构, 如图 1 (a) 所示. 它可以实现 N 种语料的并行式训练, 在该训练过程中它们相互补充. 本文将共享隐含层改造为 BN 结构的隐含层, 提出基于 SHL 结构的低资源 BN-DNN 训练方式.

在 SHL 结构的网络中, 每一个 Softmax 层对应各自训练语料的三音子绑定状态 (Senones), 仅有隐含层的参数在训练中是共享的, 各输出层的参数的更新计算与其他输出层的参数不相关. 然而当普通的隐含层结构换成 BN 层结构后, 不容易取得良好的训练效果, 原因在于 DNN 用作声学建模和特征提取时, 最终使用的层是不同的. 当构建 DNN-HMM



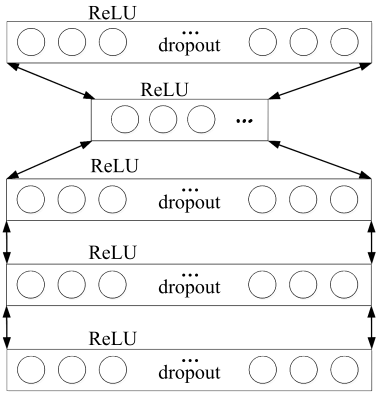
(a) SHL-MDNN

(a) SHL-MDNN



(b) 基于 dropout-maxout 的 BN-DNN 隐含层

(b) Dropout-maxout BN-DNN hidden layers



(c) 基于 dropout-ReLU 的 BN-DNN 隐含层

(c) Dropout-ReLU based BN-DNN hidden layers

图 1 基于 SHL 的网络结构示意图

Fig. 1 SHL based network structures

时, Softmax 层不参与共享训练, 与其他辅助语料没有直接关联, 因而不容易产生偏向性的问题; 而构建 BN-DNN 提取特征时, BN 层位于共享层, 其训练效果与所有参与训练的语料相关联, 虽然各语言特征具有一定声学相似性, 但是特征流的具体分布是不同的, 网络参数会受到来自其他语言数据的直接干扰. 此外, SHL-MDNN 的网络参数规模较大, 多语言训练耗时较多也是一个亟待解决的问题.

针对这些问题, 本文对基于 SHL 的 BN-DNN 的训练做出改进. 在多流特征的训练中, 某一流特征数据量较大造成的语言偏向性问题本质是训练过拟合问题. Dropout 训练技术可以有效防止 DNN 在训练时出现过拟合现象, 进而提升识别的准确性[12]. Dropout 通过概率控制 DNN 隐含层节点在训练中是否被 "激活", 在每次训练中, 每个节点在每轮训练时都以一个隐含层遗弃因子 (Hidden drop factor, HDF) 作为概率, 决定是否参与当次的参数更新. 由于每次随机忽略的节点都有所不同, 故最终模型相当于由不同的子模型叠加而成, 且每种子模型的参数都是相关的. 这种对子模型做平均的方法使得 DNN 不过分依赖于某一种特征分布, 保证 BN 层最大化获取多语言特征流的声学相似性信息. 因此在多流数据交杂的 SHL 训练中, 该技术可以有效提升 BN 层的训练效果. 此外, 为了配合 Dropout 技术, 往往舍弃传统的 Sigmoid 函数, 采用其他激活函数. Maxout 是一种可以最大化 Dropout 训练效果的激活函数[13], 该函数通过降低实际的激活的隐含层节点数量来降低网络参数规模; ReLU (Rectified linear units) 则是一种可以提升 DNN 的泛化性能的激活函数, 并且根据文献 [14], 使用该函数可以加速 DNN 的训练过程.

基于 Dropout 训练, 第 l 个隐含层的真实样本输出可以写为

$$\boldsymbol{x}_l(t) = \boldsymbol{u}_l(t) \otimes D(t), \quad 1 \le l \le L \tag{1}$$

其中, $\boldsymbol{u}_l(t)$ 为该层在 Dropout 处理之前的激活元输出, $D(t)$ 和 $\boldsymbol{u}_l(t)$ 维度相同, $D(t)$ 中元素是以 HDF 为概率分布的二值采样, 通过该二值矩阵的变换, 得到每一层的真实的激活元输出. 设第 $l-1$ 层有 1 个节点, 对于不同的激活函数, $\boldsymbol{u}_l(t)$ 计算方式不同.

$$\begin{cases} \boldsymbol{u}_l(t) = \Big[\max\limits_i \left(\boldsymbol{x}_{l-1}^1(t), \cdots, \boldsymbol{x}_{l-1}^i(t)\right), \cdots, \\ \quad \max\limits_i \left(\boldsymbol{x}_{l-1}^{(j-1)\times i+1}(t), \cdots, \boldsymbol{x}_{l-1}^{j\times i}(t)\right)\Big], \\ \qquad\qquad\qquad\qquad \text{Maxout} \\ \boldsymbol{u}_l(t) = \\ \quad \Big[\max\left(0, \boldsymbol{x}_{l-1}^1(t)\right), \cdots, \max\left(0, \boldsymbol{x}_{l-1}^{j\times i}(t)\right)\Big], \\ \qquad\qquad\qquad\qquad \text{ReLU} \end{cases} \tag{2}$$

通过误差反向传播 (Back propagation, BP) 算法对参数进行全局微调. 根据文献 [15], 训练 DNN 时目标函数为

$$D=\sum_{t=1}^{T}\log P\left(s\left(t\right)|\boldsymbol{o}\left(t\right)\right) \tag{3}$$

其中, $\boldsymbol{o}\left(t\right)$ 是一帧的训练特征向量, $s\left(t\right)$ 是 $\boldsymbol{o}\left(t\right)$ 对应的状态标签, T 是训练特征总量, 根据式 (4), 利用 Softmax 计算

$$P\left(s\left(t\right)|\boldsymbol{o}\left(t\right)\right)=\frac{\exp_{s}\left(W_{L}\boldsymbol{u}_{L-1}+\boldsymbol{b}_{L}\right)}{\sum\limits_{s'}\exp_{s'}\left(W_{L}\boldsymbol{u}_{L-1}+\boldsymbol{b}_{L}\right)} \tag{4}$$

基于目标函数 D, 利用随机梯度下降 (Stochastic gradient descent, SGD) 法更新权值 W_l 与偏移量 $\boldsymbol{b}_l$.

$$\left(W_{l},\boldsymbol{b}_{l}\right)+\varepsilon\frac{\partial D}{\partial\left(W_{l},\boldsymbol{b}_{l}\right)}\rightarrow\left(W_{l},\boldsymbol{b}_{l}\right),\ \ 1\leq l\leq L \tag{5}$$

ε 为学习速率 (Learning rate). 训练时引入冲量项 (Momentum) α 和衰减因子 η 来控制参数更新值的波动, 记 $\theta=\{W,\boldsymbol{b}\}$ 统一表示参数, $\Delta\theta^{(i)}$ 为第 i 轮训练参数更新值, 更新过程按式 (6) 进行修正:

$$\begin{aligned}\Delta\theta^{(i+1)}=&\ \alpha\times\Delta\theta^{(i)}+(1-\alpha)\times\\&\left(\varepsilon\times\frac{\partial D}{\partial\theta}+\varepsilon\times\eta\times\theta^{(i)}\right)\end{aligned} \tag{6}$$

对于 BN-DNN, 训练完成之后, 将输入特征在 DNN 中前向传播, 从 BN 层提取线性特征, 如式 (7) 所示:

$$F=W_{BN}^{\mathrm{T}}\boldsymbol{u}_{BN-1}+\boldsymbol{b}_{BN} \tag{7}$$

基于 Dropout-maxout 和 Dropout-ReLU 的 BN-DNN 隐含层结构分布如图 1 (b) 和图 1 (c) 所示. 依据上述原理, 结合低资源训练优先的要求, 在 GPU 的硬件条件下, SHL-BN-MDNN 的训练流程如图 2 所示. 首先按比例将多流训练样本组成数据分组, 然后再进行 SGD 的计算, 这样可以保证数据训练的并行性和平衡性, 该训练方式可以保证多个 DNN 几乎同时收敛、结束训练.

2　基于 CNMF 的低维高层特征提取

使用 DNN 提取特征, 本质在于对隐含层的输出进行降维、去相关. 对于典型的高层特征 BNF 的提取, 是通过在隐含层中设立 BN 层进行数据的强制降维来实现的, 然而该方法在训练过程中降低了 DNN 的分类准确率, 对于训练不够充分的低资源 DNN, 会进一步降低所提取特征的性能. 本小节提出一种新的 DNN 特征提取方法, 具体来说, 首先保留完整的 DNN 训练结构 (舍弃 BN 层的设置), 然后

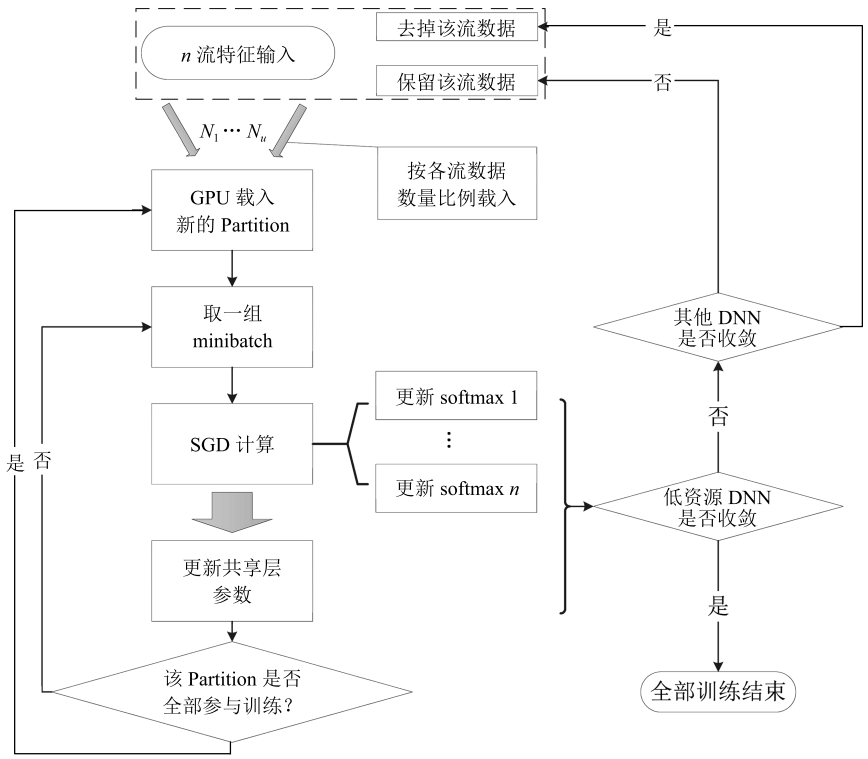


图 2　SHL-BN-MDNN 的训练流程图

Fig. 2　Diagram of SHL-BN-MDNN training scheme

对数据采取有效的降维算法实现特征提取, 这样就能避免前面提到的 BN 层偏向性问题.

在诸多降维方法中, 非负矩阵分解 (Nonnegative matrix factorization, NMF) 是一种性能较为出色的方法, 最初用于人脸识别中对局部特征的学习[16]. 该算法可以从矩阵中挖掘更为本质的信息, 简单来说, 它将一个待分解的具有非负元素的矩阵 X 分解为非负的基矩阵 F 和系数矩阵 G, 如式 (8) 所示:

$$X = FG^{\mathrm{T}} \tag{8}$$

NMF 在语音处理领域中主要应用于语音的去噪[17−18]. 然而在语音特征提取过程中, 所面对的实际数据是有正有负的, 因此一般的 NMF 算法是不适用的. CNMF 是 NMF 衍生出的一个基于聚类原理的重要算法[19], 在 CNMF 中, 将基矩阵 F 定义为待分解矩阵的列的凸组合. 即 $f_l = w_{1l}x_1 + \cdots + w_{nl} + x_n$. 或写为 $F = XW$, 其中 W 为因子矩阵, 允许 X 和 F 矩阵为半非负性质. 根据文献 [19], 因子矩阵 W 和系数矩阵 G 具有稀疏的性质, 且使用 CNMF 方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.

2.1 凸非负矩阵分解算法

CNMF 的初始化大致分为两种方法. 第一种是基于 K-means 聚类的方法, 第二种是基于已有的 NMF 的解, 本文选用 K-means 方法. 首先对待分解矩阵 X 做一次 K-means 聚类, 得到隶属度矩阵 $H = (h_1, \cdots, h_k)$, $H_{ik} = 0, 1$, 然后按式 (9), 基于 H 对 G 矩阵初始化:

$$G^{(0)} = H + 0.2E \tag{9}$$

E 为全 1 矩阵. 使用聚类的类心矩阵作为 F 矩阵, 如式 (10) 所示:

$$F = XHD_n^{-1} \tag{10}$$

其中, $D_n = \mathrm{diag}\{n_1, \cdots, n_k\}$. 根据 $F = XW$ 与式 (10), $W = HD_n^{-1}$, 但此处为了平滑处理, 设 $W^{(0)} = HD_n^{-1}$.

$$G_{ik} \leftarrow G_{ik} \times \sqrt{\frac{\left[(X^{\mathrm{T}}X)^{+}W\right]_{ik} + \left[GW^{\mathrm{T}}(X^{\mathrm{T}}X)^{-}W\right]_{ik}}{\left[(X^{\mathrm{T}}X)^{-}W\right]_{ik} + \left[GW^{\mathrm{T}}(X^{\mathrm{T}}X)^{+}W\right]_{ik}}} \tag{11}$$

再根据式 (12) 更新 W 的值:

$$W_{ik} \leftarrow W_{ik} \times \sqrt{\frac{\left[(X^{\mathrm{T}}X)^{+}W\right]_{ik} + \left[GW^{\mathrm{T}}(X^{\mathrm{T}}X)^{-}W\right]_{ik}}{\left[(X^{\mathrm{T}}X)^{-}W\right]_{ik} + \left[GW^{\mathrm{T}}(X^{\mathrm{T}}X)^{+}W\right]_{ik}}} \tag{12}$$

2.2 基于凸非负矩阵分解的特征提取

对于不包含 BN 层的 DNN 而言, 它的第 l 个隐含层的线性输出具有维数大、相关性大的特点, 将其直接作为特征进行高斯混元建模会得到很差的结果, 因此需要进行降维和去相关等处理. 若直接利用矩阵分解算法对 DNN 特征做降维, 理论上行不通, 因为语音不同于图像, 一幅图像具有整体的平稳性且不具有时变性, 因此易于对其提取整体特征, 而语音仅具有短时平稳性. 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, 矩阵变换会破坏语音特征的时序信息, 导致无法训练出良好的声学模型.

本文采用一种间接的方法. 由于在计算 DNN 隐含层的线性输出时, 层与层之间的权值矩阵作用于每一帧原始声学特征, 因此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. 而由于同一层的偏移向量和权值矩阵并没有整体性, 因此很难对偏移向量与权值矩阵实施相同的操作, 本方法在提取特征时舍弃偏移向量的使用. 该特征提取方法如图 3 所示.

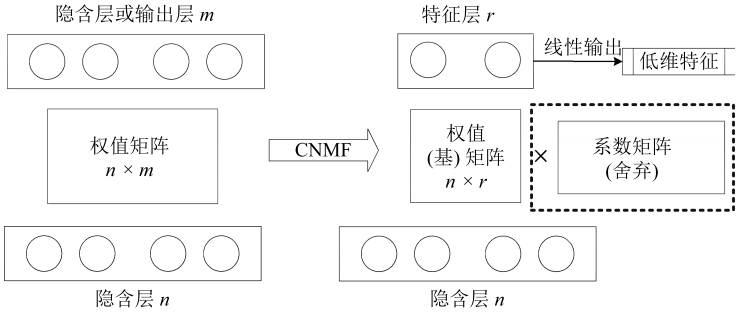


图 3 基于 CNMF 的低维特征提取方法

Fig. 3 CNMF based low-dimensional feature extraction approach

首先将某一层的 $n \times m$ 权值矩阵进行分解, 得到 $n \times r$ 的基矩阵和 $r \times m$ 的系数矩阵, 然后取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.

待分解权值矩阵记为 W, 经过分解, 得到 $W = W'G^{\mathrm{T}}$, 由于不设置特征层偏移量, 因此新的低维特征计算如式 (13):

$$F = W'^{\mathrm{T}}X \tag{13}$$

其中, X 为上一隐含层的激活元输出. 而由于 $W' = WH$, 其中 H 为因子矩阵, 因此式 (13) 可以写为

$$F = H^{\mathrm{T}}W^{\mathrm{T}}X \tag{14}$$

将式 (14) 和式 (7) 作对比可以发现, 利用 CNMF 提取特征时, 实质上是对 DNN 的线性输出做了一次基于聚类原理的降维. 由于权值矩阵以最后的分类输出为目标进行训练, 而训练目标又是音素建模单元, 因此该矩阵通过基于聚类原理的矩阵分解之后, 最优特征维数应当与训练语料的音素数量相关, 而不像 BNF 一样仅仅与输入的特征维数相关 (BN 层起到对输入特征非线性压缩的作用).

3 实验

3.1 实验语料与评价指标

实验采用 RM、TIMIT、Vystadial 2013 English data 和 Vystadial 2013 Czech data 四种语料. RM 语料库[20] 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; TIMIT 语料库[21] 由 630 个说话人的语料组成, 每个说话人包含 10 个语句, 共涵盖了美式英语的 8 种主要方言, 语料库包含了音素级标注; Vystadial 2013 English data (Vystadial_en) 是一类开源的英语语料库[22], 全部时长 41 小时, 来源于人工信息服务系统的对话语音数据; Vystadial 2013 Czech data (Vystadial_cz) 是开源的捷克语语料库[23], 全部时长 15 小时, 来源于三类数据: 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.

实验评价指标为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、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. 如式 (15) 所示:

$$r = \frac{W}{N} \times 100\,\% \tag{15}$$

3.2 实验工具与硬件条件

实验使用 Kaldi 工具包[24] 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[25] 进行相关的 DNN 的搭建与训练; 使用 PYMF 工具包[26] 实现 CNMF 算法.

声学模型训练、解码矩阵分解的过程基于 12 核 3.07 GHz Xeon CPU 实现, DNN 的训练则是基于单核 Quadro 600 GPU 进行.

3.3 基于 SHL 结构的 BNF 特征的低资源捷克语识别实验

该部分实验中, 基于 Vystadial_cz 语料构建低资源语音数据环境. 选取 Vystadial_cz 中的 1 小时训练语料作为训练集, 总共 1 504 句话; 再选取 Vystadial_cz 语料库测试语音部分的 30 分钟左右的数据作为测试集, 包含 666 句话, 共 3 910 个待识别词. 基于 Vystadial_cz 语料库中全部训练语料的标注文本构建发音字典并训练二元语言模型 (Bigram language model). 辅助训练语料为全部 RM 的训练语料 (3.82 小时)、全部 TIMIT 训练语料 (3.15 小时) 和一半的 Vystadial_en 训练语料 (19.7 小时).

3.3.1 基于单语言训练的低资源语音识别基线系统

首先对 1 小时的 Vystadial_cz 提取 39 维 MFCC 特征 (13+△+△△) 训练一个三音子 GMM 模型, 进行强制对齐后, 训练基于线性判别分析 (Linear discriminant analysis, LDA) 和最大似然线性变换 (Maximum likelihood linear transform, MLLT) 的三音子 GMM 声学模型 (13 维 MFCC 特征进行 9 帧拼接, LDA 降到 40 维), 该模型高斯混元数为 19 200 个. 然后再利用特征空间最大似然线性回归 (Feature-space maximum likelihood linear regression, fMLLR) 技术进行说话人自适应训练 (Speaker adaptive training, SAT), 从而构成 LDA + MLLT + SAT 的 GMM 声学模型.

通过对该模型强制对齐的方式, 得到 BN-DNN 中 softmax 层的训练目标. DNN 的训练特征使用效果较好的 fbanks 特征[5], 首先提取 40 维的 fbanks 特征, 进行 11 帧的拼接 (5-1-5), 将所得到的超矢量作为 DNN 的输入特征. 对于单语言的 BN-DNN, 仅使用 1 小时的低资源训练语料进行训练. 隐含层有 5 层, 每层节点有 1 024 个, BN 层有 40 个节点, softmax 层节点数同 LDA + MLLT + SAT 的 GMM 的 senones 数量一致, 为 915 个. 借鉴文献 [3−4, 9, 27] 的经验, 本实验将 BN 层置于隐含层的中后层位置. 因此, 该 BN-DNN 的结构为 “440-

1 024-1 024-1 024-40-1 024-915". 训练集和交叉验证 (Cross-validation) 集各占训练数据的 95 % 和 5 %.

对每个隐含层 (包含 BN 层) 进行 10 轮的 RBM 预训练, 然后利用 BP 算法进行全局参数的微调, 在训练过程中, 学习速率设置初始值为 0.08, 每当相邻两轮训练的验证误差小于 0.1 % 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 0.1 % 时训练停止 (如果一直大于 0.1 %, 则最多衰减 8 次. 此外冲量值设为 0.5, Minibatch 尺寸设为 256. 训练完成之后, 从 BN 层提取 BNF, 使用 BNF 训练基于 LDA、MLLT 的三音子 GMM 声学模型 (9 帧拼接, LDA 降至 40 维), 该模型的高斯混元数量设定为 22 000, 识别结果如表 1 第 1 行所示.

表 1 不同训练方法下 BNF 的 WER (%)

Table 1 WER of BNF based on different training methods (%)

训练方案	WER	DNN 参数数量 (MB)
单语言 BNF	67.42	3.57
SHL + BNF	63.25	8.34
SHL + Dropout + Maxout + BNF	**58.95**	**3.11**
SHL + Dropout + ReLU + BNF	62.74	8.34

3.3.2 基于多语言训练的低资源语音识别系统

本小节中, 使用 SHL 结构对低资源的 BN-DNN 进行辅助训练, 引入 RM、TIMIT、Vystadial_en 等一共 26.7 小时的英语语料进行辅助训练. 对三种辅助语料分别训练三个基于 LDA + MLLT + SAT 的 GMM 模型, 通过强制对齐得到各自 DNN 的 softmax 层的训练目标, 各自 DNN 输出层节点数分别为 1 487、2 009 和 1 031, DNN 的输入均为 440 维拼接的 fbanks 特征 (40 × 11), 隐含层结构为 "1 024-1 024-1 024-40-1 024". 本实验对 SHL 结构的 DNN 不进行预训练而是随机进行参数的初始化, 然后直接通过 SGD 的计算调整网络参数. 冲量值和学习速率的设置与基线系统保持一致.

训练完之后, 得到 4 个 BN-DNN, 使用低资源 Vystadial_cz 的 DNN 对低资源语料提取 BNF, 然后训练 LDA + MLLT 的 GMM 声学模型, 参数设置与基线系统保持一致. 识别结果如表 1 第 2 行所示, 可以看出, 该系统比基线系统的 WER 相对降低了 6.2 % (67.42 % → 63.25 %).

然后利用 Dropout、Maxout 和 ReLU 技术对 BN-DNN 进行改进. 已知对于一般的 BN-DNN, 激活函数不作用于 BN 层的特征提取过程; 而对于 Maxout-DNN, 激活函数需要作用于 BN 层, 因为 Maxout 函数并未对函数幅值作归一化, 并且 Maxout 保证了 BN 层的维数为 40, 因此具有可比性. 此外, 在参数设置方面, Dropout 的 HDF 与 Maxout 的 Pooling 尺寸都需要进行设置. 根据文献 [28−29] 的经验, Dropout 为 0.2、Pooling 尺寸为 3 时效果最佳, 为此进行实验验证, 其中 HDF 分为 0.1, 0.2, 0.3 三种情况进行讨论, Pooling 尺寸分为 512 × 2, 342 × 3 和 256 × 4 三种情况进行讨论, 这样使得隐含层原始尺寸与基线系统基本一致 (节点数在 1 024 左右), 实验结果如表 2 所示.

实验可得到的第一个结论是最佳的 HDF 为 0.2、最合适的 Pooling 尺寸为 3; 第二个结论是隐含层使用 Dropout 技术可以有效增强训练效果, 而 BN 层不宜使用 Dropout 训练技术. 对于第二个结论, 可以作如下解释, 由于数据经隐含层映射和经 BN 层映射得到的是两种不同分布的数据, 普通隐含层由于节点数较多, 因此映射时对输入数据的分布细节要求更多, 训练中的过拟合现象会在一定程度上影响映射效果; 而 BN 层对输入数据的压缩可以看作是一种广义的聚类, 因此对数据分布的细节要求较少, 过拟合现象影响不大. 理论而言, Dropout 对 BN 层的训练效果不会有增益.

另外, 对比表 1 的第 2 ∼ 4 行可以看出, 当引入 Maxout、ReLU 等激活函数后, BNF 的性能在多语言训练的基础上得到进一步的明显提升, 其识别系统比基线系统的 WER 分别相对降低了 12.6 % (67.42 % → 58.95 %) 和 7.0 % (67.42 % → 62.74 %), 说明多语言的辅助训练效果更好了, BN-DNN 的语言偏向性降低了. 从结果不难看出, Max-

表 2 不同 dropout 和 maxout 参数下的 WER (%)

Table 2 WER under different dropout and maxout parameters (%)

Dropout-maxout 参数	HDF = 0.1 BN-DF = 0	HDF = 0.1 BN-DF = 0.1	HDF = 0.2 BN-DF = 0	HDF = 0.2 BN-DF = 0.2	HDF = 0.3 BN-DF = 0	HDF = 0.3 BN-DF = 0.3
Pooling 尺寸: 512 × 2 (40 × 2)	62.11		60.77		61.89	
Pooling 尺寸: 342 × 3 (40 × 3)	59.72	61.14	**58.95**	60.32	60.13	61.5
Pooling 尺寸: 256 × 4 (40 × 4)	61.23		60.36		61.84	

out 配合 Dropout 的训练效果是最优的, 而且基于 SHL 结构的 BN-DNN 参数规模相对其他 SHL 网络参数的数量降低了 62.7 % (8.34 MB → 3.11 MB), 从训练时间的角度来说, 降低了约 25 % (在第 3.2 节所描述硬件条件下记录 DNN 训练耗时, 大约从 12 小时降至 9 小时).

3.4 基于 CNMF 低维特征的低资源语音识别实验

本节主要对 CNMF 提取的低维特征进行低资源条件下的识别性能测试. 由于 CNMF 的效果与 DNN 隐含层的训练水平相关, 因此实验分为两部分, 一部分是基于低资源单语言训练的 DNN 进行的实验, 另一部分是与 SHL-MDNN 相结合的实验.

3.4.1 基于低资源单语言训练的英语和捷克语实验

该部分实验针对 Vystadial_en 和 Vystadial_cz 两种语料搭建两个低资源识别系统. 分别选取 1 小时训练集和 30 分钟测试集, 并且使用标注文本构建发音字典和训练二元语言模型. 该实验的 BNF 基线系统与第 3.3 节中基线系统的设置基本相同, 唯一不同之处在于本实验中采用 40 维 fMLLR 特征 (13 维 MFCC 特征进行 9 帧拼接, LDA 降至 40 维, 并经过 MLLT 和 SAT 训练) 对 DNN 进行训练.

对于 CNMF 的实现, 首先通过 50 轮的 K-means 训练对矩阵分解初始化, 然后对分解过程进行 500 轮训练, 得到特征层权值矩阵并提取低维特征, 详细方法如第 2.2 节所描述. 使用该特征训练 LDA + MLLT 的 GMM 声学模型搭建识别系统. 该实验中, 待分解权值矩阵的层位置与分解维数是两个很重要的参数指标, 它们与系统识别率的关系如图 4 所示. 其中, 为了便于与 BNF 作对比, 只对 40 左右的维数进行研究, 并且由于高层特征的性能优于底层特征性能, 因此对 DNN 的后三层进行探讨, 图中 “第 5 层” 表示最后一层 (输出层) 的权值矩阵, 以此类推 “第 3 层” 和 “第 4 层” 所代表的分解位置.

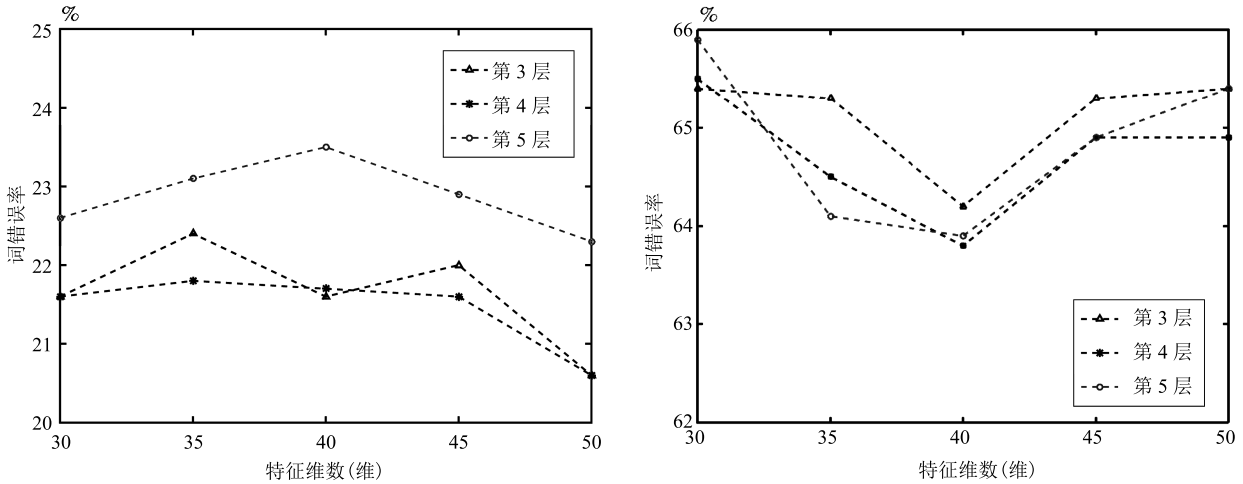
可以看出, 在两个识别任务中, 待分解权值矩阵层位置均为倒数第二层时效果最好, 且英语语料的最优分解维数为 50 维, 捷克语的最优分解维数为 40 维. 由于英语有 48 个音素, 而捷克语有 38 个音素, 因此实验验证了第 2.2 节中的结论.

为了进一步验证 CNMF 提取特征的有效性, 将 CNMF 算法与传统的奇异值分解 (Singular value decomposition, SVD) 算法进行比较. 使用两种算法, 均针对 DNN 的倒数第二层, 且分解维数为 40. 对于 CNMF, 按照本文方法使用基矩阵作为特征层矩阵; 对于 SVD, 利用与本文所提方法相同的思路, 使用左奇异分量作为特征层分解矩阵, 从而对该层线性输出实现降维, 提取特征. 两种算法所提取特征的识别性能对比如表 3 所示.

表 3　基于单语言训练时各特征的识别性能 WER (%)

Table 3　Recognition performance WER each type of feature based on monolingual training (%)

识别任务	BNF	CNMF 低维特征	SVD 低维特征
低资源 Vystadial_en	21.6	20.6	21.51
低资源 Vystadial_cz	64.8	63.76	64.43



(a) Vystadial_en 低资源识别任务
(a) Vystadial_en low-resource task

(b) Vystadial_cz 低资源识别任务
(b) Vystadial_cz low-resource task

图 4　不同分解参数下基于 CNMF 的低维特征词错误率

Fig. 4　WER of CNMF based low-dimensional features under different factorization parameters

由实验结果可知, 基于矩阵分解的特征优于传统的 BN 特征, 且 CNMF 算法优于 SVD 算法. 从原理上解释, CNMF 基于聚类原理, 基矩阵作为权值矩阵的列的线性组合, 包含了原矩阵的主要信息, 冗余信息主要存在于舍弃的系数矩阵中, 因此基于 CNMF 的方法对训练不足的 DNN 的权值矩阵起到聚类、去冗余等作用; 而 SVD 将矩阵分解为左奇异分量和右奇异分量, 无论舍弃哪一个都会浪费有效的矩阵分量, 因此基于 SVD 的特征在识别性能上劣于基于 CNMF 的特征, 实验结果进一步验证了 CNMF 算法在提取高层特征时的有效性.

3.4.2 基于 SHL 结构的多语言训练的捷克语实验

在低资源单语言 DNN 的条件下测试之后, 再对 CNMF 低维特征在辅助训练条件下进行测试, 本节中的 DNN 基于 SHL 结构进行多语言辅助训练, 除了不设置 BN 层之外, 声学参数、网络参数与第 3.3.2 节中基于 SHL 的 BN-DNN 完全一致. 使用 CNMF 技术结合辅助训练的 DNN 提取低维特征, 首先通过 50 轮的 K-means 训练对矩阵分解初始化, 然后进行 500 轮训练得到分解矩阵. 由于参与 DNN 权值矩阵训练的不仅有捷克语还有英语, 因此最优分解维数需要在 40 和 50 之间进行讨论, 结果如表 4 所示. 其中 "第 5 层" 表示对最后一层 (输出层) 做分解, 以此类推.

从结果可知, 由于隐含层是由英语和捷克语共同训练的, 且英语训练语料的数量更多, 训练相对充分, 所以对共享隐含层分解时的最优维数为 50; 而输出层仅有捷克语参与训练, 训练相对不充分, 因此对输出层分解时的最优维数为 40. 这些结果进一步验证了基于 CNMF 的低维特征与语言音素的相关性. 此外, 对倒数第二层的权值矩阵进行分解依然可以得到最优的特征, 这一点与第 3.4.1 节中的结论一致.

3.4.3 CNMF 低维特征与 BNF 在 GMM 识别系统中的对比

分别对单语言和辅助训练两种情况选取最好的识别结果, 将 BNF 与 CNMF 低维特征的识别系统进行对比, 如表 5 所示. 对于英语和捷克语的低资源 DNN 的实验, CNMF 低维特征分别相对 BNF 特征的识别性能相对提高了 4.6 % (21.6 % → 20.6 %) 和 1.6 % (64.80 % → 63.76 %). 此实验说明, 当 DNN 训练相对不充分时, 在使用 GMM 搭建识别系统的情况下, CNMF 低维特征优于传统的高层特征 BNF. 而在辅助训练语料充足的条件下, 对于 ReLU-DNN, CNMF 低维特征优于 BNF; 对于 Sigmoid-DNN、Maxout-DNN, BNF 则显示出了优

表 4 基于 SHL 多语言训练的 CNMF 低维特征的 WER (%)

Table 4 WER of SHL multilingual training CNMF based low-dimensional features (%)

CNMF 特征提取方案	第 3 层	第 4 层	第 5 层
Sigmoid + 40 维分解	64.27	64.94	64.71
Sigmoid + 50 维分解	63.86	**63.81**	64.99
Dropout + Maxout + 40 维分解	60.33	60.13	59.59
Dropout + Maxout + 50 维分解	59.59	**59.12**	59.95
Dropout + ReLU + 40 维分解	63.71	61.59	61.28
Dropout + ReLU + 50 维分解	62.15	**60.26**	61.84

表 5 BNF 与 CNMF 低维特征的 GMM tandem 系统 WER (%)

Table 5 WER of BNF and CNMF based low-dimensional features on GMM tandem system (%)

实验配置	BNF	CNMF 低维特征
Vystadial_en (单语言 fMLLR) + Sigmoid-DNN	21.6	**20.6**
Vystadial_cz (单语言 fMLLR) + Sigmoid-DNN	64.8	**63.76**
Vystadial_cz (单语言 fbanks) + Sigmoid-DNN	**63.25**	63.81
Vystadial_cz (单语言 fbanks) + Dropout-maxout-DNN	**58.95**	59.12
Vystadial_cz (单语言 fbanks) + Dropout-ReLU-DNN	62.74	**60.26**

于 CNMF 低维特征的识别性能, 且训练耗时更少. 该实验说明, 使用 GMM 建模时, 在训练相对充分的 DNN 结构中, BNF 优于 CNMF 低维特征.

总体来说, BN-DNN 中, 由于 BN 层权值矩阵的训练过程与训练样本完全相关, 因此 BNF 的最优维数与输入 DNN 的声学特征维数是密切相关的, 所以 BN 层的功能在于实现了对输入特征的非线性压缩. 而对于 CNMF 的特征提取方法, 声学特征训练只是与特征层权值矩阵的原始矩阵直接相关, 最终的特征层权值矩阵还与 CNMF 的迭代训练有关, 由于 DNN 的权值矩阵以音素状态为训练目标, 因此基于聚类原理的 CNMF 算法可以从原始权值矩阵中分解得到更为本质的包含分类信息的矩阵, 使得特征包含了更多 DNN 对该语言的分类信息, 所以该特征的最优维数与 DNN 训练语料的音素个数息息相关.

3.5 两种方法与 DNN-HMM 在低资源捷克语识别实验中的对比

由于 DNN-HMM 识别系统往往能在训练语料相对充足的情况下取得所有识别系统中最优的性能[5, 15], 因此将本文的两种方法与 DNN-HMM 识别系统作对比. 根据文献 [30−31], 采用子空间高斯混合模型 (Subspace Gaussian mixture models, SGMM) 搭建识别系统可以得到优于 GMM-HMM 系统的识别率, 尤其适用于低资源环境, 因此分别使用 BNF 和 CNMF 低维特征训练各自的 SGMM 进行识别系统的搭建. 首先对 LDA + MLLT 的声学模型做强制对齐, 然后训练高斯混元数为 400 的通用背景模型 (Universal background model, UBM), 基于此模型, 训练子状态数量为 5 000 的 SGMM.

此外, 由于基线系统的参数规模达到了 3.57 MB, 而低资源训练数据的数据量约为 20 000 帧, 所以存在一定过拟合的风险. 为了进一步验证本文方法的有效性, 在两种较小参数规模的 DNN 结构下进行实验. 由于隐含层节点数至少达到与输出层节点数同一个量级 (本实验中约为 1 000) 时才能保证 DNN 有较好的分类性能, 否则无法估计出性能良好的后验概率, 且在低资源条件下, 以极大的牺牲分类性能来避免过拟合现象是得不偿失的. 因此, 通过降低网络节点数降低参数规模是不现实的, 实验主要通过降低隐含层数量来降低网络参数规模. 在此增加了两种网络结构的对比实验: 隐含层层数降为 3 层 (BN 层设置在三层中的第 2 个隐含层), 节点数设为 1 024 和 512 (基线系统的参数规模降为 1.47 MB 和 0.74 MB), 使用 Maxout 时, 隐含层分别设置为 171×3 和 342×3. 实验结果如表 6 所示.

从表 6 中可知, 当引入 SGMM 时, CNMF 低维特征在各 Tandem 系统中几乎均优于 BNF, 且基于该特征的 Tandem 识别系统在各实验中取得了最优的结果.

将基线系统的网络尺寸的缩小理论上可以相对降低低资源单语言训练时的过拟合风险. 但是在多语言训练时, 比较不同网络尺寸的实验结果可以发现, 较小的网络尺寸并不能取得更好的识别结果, 这是因为训练数据量与 DNN 的参数量之间的规模差距得到减小, 层数较多、节点数较多的网络可以估计出更准确的后验概率分布, 即具有更强的非线性变换能力. 同时, 实验结果表明, 本文提出的两种方法适用于不同的网络结构, 都取得了相对基线系统识别率的提升, 且 Dropout 和 Maxout 在不同网络结构下均改善了多语言训练的效果, 使得所提取特征性能得到提高.

根据表 6 结果, 网络结构设置为 5 个隐含层时,

表 6　基于 SHL 多语言训练时 SGMM tandem 系统和 DNN-HMM 系统的 WER (%)

Table 6　WER of SGMM tandem systems and DNN-HMM hybrid systems based on SHL multilingual training (%)

DNN 隐含层结构		BNF	CNMF 低维特征	DNN-HMM
Sigmoid	5 层 1 024 (BN: 40)	63.15	**61.79**	63.94
	3 层 1 024 (BN: 40)	63.09	**61.85**	63.99
	3 层 512 (BN: 40)	63.5	**61.84**	63.96
Dropout + Maxout	5 层 342 (*3, BN: 40)	58.03	**57.8**	58.24
	3 层 342 (*3, BN: 40)	60.61	**60.4**	63.99
	3 层 171 (*3, BN: 40)	**62.61**	64.72	68.77
Dropout + ReLU	5 层 1 024 (BN: 40)	60.72	**58.82**	59.57
	3 层 1 024 (BN: 40)	64.35	**59.16**	59.92
	3 层 512 (BN: 40)	63.43	**61.68**	62.2

系统可以取得最优识别率, CNMF-SGMM 取得了最优识别率, 且分别相对 DNN-HMM 提高了 3.4 % (63.94 % → 61.79 %), 0.8 % (58.24 % → 57.80 %), 1.3 % (59.57 % → 58.82 %).

综合第 3.4.3 节和第 3.5 节中的结论, 在 GMM 建模的情况下, 基于 SHL 结构的 BNF 提取方法更省时, 且识别率更高; 而在 SGMM 建模的情况下, 基于 CNMF 的低维特征提取方法更优, 且取得了优于 DNN-HMM 系统的识别性能. 总的来说, 本文提出的第一种方法的主要优势在于训练时间较短, 第二种方法的优势在于提取的特征识别性能更为出色.

4 结论

本文针对低资源训练数据下 DNN 特征建模识别性能不佳的问题, 首先从训练的角度, 提出利用 SHL 结构对 BN-DNN 进行辅助训练, 为缓解 BN 层的语言偏向问题和多语言训练耗时问题, 引入 Dropout、Maxout、ReLU 等技术对 DNN 的训练进行改进; 然后从特征提取方法的角度, 提出利用 CNMF 算法对权值矩阵进行聚类降维, 进而提取一种新的基于 DNN 的特征. 实验证明, 在 DNN 训练不充分的低资源条件下, CNMF 特征优于 BNF 的识别性能; 而在 SHL 辅助训练的情况下, 基于改进训练技术的 BNF 相比低资源训练的 BNF 有了明显提升, 且网络参数得到了大幅降低, 使用 GMM 建模时, BNF 更优, 使用 SGMM 建模时, CNMF 特征更优, 且取得了优于 DNN-HMM 系统的识别性能.

References

1 Thomas S. Data-driven Neural Network Based Feature Front-ends for Automatic Speech Recognition [Ph. D. dissertation], Johns Hopkins University, Baltimore, USA, 2012.

2 Grézl F, Karaát M, Kontár S, Černocký J. Probabilistic and bottle-neck features for LVCSR of meetings. In: Proceedings of the 2007 International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hawaii, USA: IEEE, 2007. 757−760

3 Yu D, Seltzer M L. Improved bottleneck features using pretrained deep neural networks. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH). Florence, Italy: Curran Associates, Inc., 2011. 237−240

4 Bao Y B, Jiang H, Dai L R, Liu R. Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition. In: Proceedings of the 2013 International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, 2013. 6980−6984

5 Hinton G E, Deng L, Yu D, Dahl D E, Mohamed A R, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, 2012, **29**(6): 82−97

6 Lal P, King S. Cross-lingual automatic speech recognition using tandem features. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(12): 2506−2515

7 Veselý K, Karafiát M, Grézl F, Janda M, Egorova E. The language-independent bottleneck features. In: Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT). Miami, Florida, USA: IEEE, 2012. 336−341

8 Tüske Z, Pinto J, Willett D, Schlüter R. Investigation on cross-and multilingual MLP features under matched and mismatched acoustical conditions. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, 2013. 7349−7353

9 Gehring J, Miao Y J, Metze F, Waibel A. Extracting deep bottleneck features using stacked auto-encoders. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, 2013. 3377−3381

10 Miao Y J, Metze F. Improving language-universal feature extraction with deep maxout and convolutional neural networks. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore: International Speech Communication Association, 2014. 800−804

11 Huang J T, Li J Y, Dong Y, Deng L, Gong Y F. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, 2013. 7304−7308

12 Hinton G E, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R R. Improving neural networks by preventing co-adaptation of feature detectors. *Computer Science*, 2012, **3**(4): 212−223

13 Goodfellow I J, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout networks. In: Proceedings of the 30th International Conference on Machine Learning (ICML). Atlanta, GA, USA: ICML, 2013: 1319−1327

14 Zeiler M D, Ranzato M, Monga R, Mao M, Yang K, Le Q V, Nguyen P, Senior A, Vanhoucke V, Dean J, Hinton G H. On rectified linear units for speech processing. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, 2013. 3517−3521

15 Dahl G E, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(1): 30−42

16 Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, **401**(6755): 788−791

17 Wilson K W, Raj B, Smaragdis P, Divakaran A. Speech denoising using nonnegative matrix factorization with priors. In: Proceedings of the 2008 International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Las Vegas, NV, USA: IEEE, 2008. 4029−4032

18 Mohammadiha N. Speech Enhancement Using Nonnegative Matrix Factorization and Hidden Markov Models [Ph. D. dissertation], KTH Royal Institute of Technology, Stockholm, Sweden, 2013.

19 Ding C H Q, Li T, Jordan M I. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(1): 45−55

20 Price P, Fisher W, Bernstein J, Pallett D. Resource management RM1 2.0 [Online], available: https://catalog.ldc.upenn.edu/LDC93S3B, May 16, 2015

21 Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N, Zue V. TIMIT acoustic-phonetic continuous speech corpus [Online], available: https://catalog.ldc.upenn.edu/LDC93S1, May 16, 2015

22 Korvas M, Plátek O, Dušek O, Žilka L, Jurčíček F. Vystadial 2013 English data [Online], available: https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-4671-4, May 17, 2015

23 Korvas M, Plátek O, Dušek O, Žilka L, Jurčíček F. Vystadial 2013 Czech data [Online], available: https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-4670-6?show=full, May 17, 2015

24 Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y M, Schwarz P, Silovsky J, Stemmer G, Vesely K. The Kaldi speech recognition toolkit. In: Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA: IEEE Signal Processing Society, 2011. 1−4

25 Miao Y J. Kaldi + PDNN: Building DNN-based ASR Systems with Kaldi and PDNN. arXiv preprint arXiv: 1401.6984, 2014.

26 Thurau C. Python matrix factorization module [Online], available: https://pypi.python.org/pypi/PyMF/0.1.9, September 25, 2015

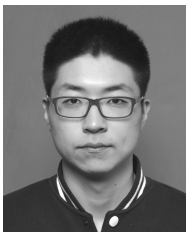
27 Sainath T N, Kingsbury B, Ramabhadran B. Auto-encoder bottleneck features using deep belief networks. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Kyoto, Japan: IEEE, 2012. 4153−4156

28 Miao Y J, Metze F. Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France: Interspeech, 2013. 2237−2241

29 Miao Y J, Metze F, Rawat S. Deep maxout networks for low-resource speech recognition. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Olomouc, Czech: IEEE, 2013. 398−403

30 Povey D, Burget L, Agarwal M, Akyazi P, Feng K, Ghoshal A, Glembek O, Goel N K, Karafiát M, Rastrow A, Rastrow R C, Schwarz P, Thomas S. Subspace Gaussian mixture models for speech recognition. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Texas, USA: IEEE, 2010. 4330−4333

31 Wu Wei-Lan, Cai Meng, Tian Yao, Yang Xiao-Hao, Chen Zhen-Feng, Liu Jia, Xia Shan-Hong. Bottleneck features and subspace Gaussian mixture models for low-resource speech recognition. *Journal of University of Chinese Academy of Sciences*, 2015, **32**(1): 97−102
(吴蔚澜, 蔡猛, 田垚, 杨晓昊, 陈振锋, 刘加, 夏善红. 低数据资源条件下基于 Bottleneck 特征与 SGMM 模型的语音识别系统. 中国科学院大学学报, 2015, **32**(1): 97−102)



秦楚雄 信息工程大学信息系统工程学院博士研究生. 主要研究方向为智能信息处理. 本文通信作者.
E-mail: chuxiongq313@gmail.com
(**QIN Chu-Xiong** Ph. D. candidate in the Department of Information and System Engineering, Information Engineering University. His main research interest is intelligent information processing. Corresponding author of this paper.)

张连海 信息工程大学信息系统工程学院副教授. 主要研究方向为语音信号处理与智能信息处理.
E-mail: lianhaiz@sina.com
(**ZHANG Lian-Hai** Associate professor in the Department of Information and System Engineering, Information Engineering University. His research interest covers speech signal processing and intelligent information processing.)