DOI：10.7652/xjtuxb201604010

# 子空间域相关特征变换与融合的语音识别方法

陈斌[1]，胡平舸[2]，屈丹[1]
（1.解放军信息工程大学信息系统工程学院，450001，郑州；2.山东大学信息科学与工程学院，250100，济南）

**摘要**：为了提高语音识别准确率，提出了一种子空间域相关特征变换与融合的语音识别方法（MFCC-BN-TC方法）。该方法提取语音短时谱结构特征（BN）和包络特征（MFCC）分别描述语音短时谱结构和包络信息，并采用域相关特征变换的形式分别对BN和MFCC特征进行特征变换；然后对这种变换进行泛化扩展提出子空间域相关特征变换，以采用不同的时间颗粒度（帧和语音分段）进行多层次区分性特征表达；最后，对多种区分性特征变换后的特征进行联合表征训练声学模型，并给出了区分性特征变换与融合的一般框架。实验结果表明：MFCC-BN-TC方法比采用原始BN特征方法和采用MFCC特征基线系统方法，识别性能各自提高了0.98%和1.62%；融合MFCC-BN-TC方法变换以后的语音信号特征，相比于融合原始特征，识别率提升了1.5%。

**关键词**：语音识别；区分性训练；深度神经网络；子空间域相关特征变换
**中图分类号**：TN912 **文献标志码**：A **文章编号**：0253-987X(2016)04-0060-08

# A Continuous Speech Recognition Method Using Dependent Feature Transformation and Combination of Subspace Region

CHEN Bin[1], HU Pingge[2], QU Dan[1]
(1. Institute of Information System Engineering, Information Engineering University, Zhengzhou 450001, China;
2. Institute of Information Science and Engineering, Shandong University, Jinan 250100, China)

**Abstract**: A speech recognition method based on dependent feature transformation and combination of subspace regions (MFCC-BN-TC) is proposed to improve the recognition accuracy. The structure feature (BN) and envelope feature (MFCC) are extracted to separately describe the structure and envelope information of the short speech spectrum, and the region dependent feature transformation is adopted to perform feature transformation for the BN and the MFCC, respectively. The transformation is then generalized to give a subspace region-dependent feature transformation so that different time units (frame and segment) are applied to finish multi-level modeling. Moreover, a feature combination framework is proposed, and the acoustic model is trained using combined multi-features after transformation. Experimental results and comparisons with the method using raw BN and the method based on MFCC feature show that the recognition rate of the MFCC-BN-TC method increases by 0.96% and 1.62%, respectively. The gain in performance of the MFCC-BN-TC method increases by 1.5% through combining the transformed features.

**Keywords**: speech recognition; discriminative training; deep neural network; subspace region-dependent feature transformation

---

收稿日期：2015-08-07。 作者简介：陈斌(1987—)，男，博士生；屈丹(通信作者)，女，博士，副教授。 基金项目：国家自然科学基金资助项目(61175017,61403415)。
网络出版时间：2016-01-13 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20160113.1957.004.html

自动语音识别是将人类自然语音转化为文本或命令的技术，是实现与机器“友好式”交流的重要技术之一。为提高语音识别率，常对特征参数进行某种变换[1-2]，以得到具有鲁棒性和区分性的特征。其中以采用高斯混合模型进行声学空间划分的最小音素错误率（feature minimum phone error，fMPE）方法[3]、域相关特征变换（region dependent linear transform，RDLT）方法[4]和状态绑定（tied-state）的RDLT 方法[5]为典型代表。

近年来，由于具有强大的学习和建模能力，深度神经网络（deep neural network，DNN）[6]受到了广泛关注而成为另一主流语音识别模型，其中，采用深度神经网络-隐马尔科夫模型（deep neural network-hidden Markov model，DNN-HMM）[7]的声学建模方法相比于高斯混合模型隐马尔科夫模型（hidden Markov model-Gaussian mixture model，GMM-HMM）声学模型，在多种语音识别任务下优势非常明显。在特征提取方面，基于 DNN 提取瓶颈（bottleneck，BN）特征[8]的方法提出后，构建的 BN-GMM-HMM 识别系统，可得到与 DNN-HMM 相当的识别性能。鉴于 GMM-HMM 模型下区分性特征变换方法优越的性能，学者们力图在深度神经网络中寻找进行特征变换的方法[9-10]。为进一步提高 BN 特征的区分性，根据传统的 fMPE 方法，文献[11]提出了一种通过调整神经网络中 BN 层的权值进行特征变换的方法。提取 BN 特征时，DNN 的输入为包络特征（MFCC）、感知加权线性预测系数特征（PLP）等短时谱特征，得到的 BN 特征主要侧重于输入特征结构信息的学习和表达，这与侧重于描述谱包络的短时谱特征，具有较好的互补性[8,10,12]。但是，如何在同一框架下，利用特征变换方法寻找到更具有区分性表达的特征，以及有效地将侧重结构信息表达和谱细节等不同方式的特征进行融合表征，仍然是语音识别领域重点关注问题之一[13]。

为进一步增大 BN 和 MFCC 特征的区分性，以及将 2 种侧重不同表征的特征有效融合，本文在同一区分性特征变换目标函数下，分别调整 BN 特征提取网络输出层权值及求得 MFCC 特征变换矩阵集合，进行特征变换。在特征变换过程中，分别对 BN 和 MFCC 特征采用基于帧和分段的方式，在不同的时间层次上进行特征变换，以提高特征间的互补性，然后联合区分性特征变换后的特征进行声学模型训练，得到了区分性特征变换与融合的一般框架。在该框架下，通过设置不同的变换矩阵形式，讨论了不同特征融合方法的识别性能。

# 1　基于区分性训练准则的域相关特征变换

## 1.1　BN 特征区分性变换

基于 DNN 提取 BN 特征的流程如图 1 所示，BN 特征的输出层位于网络的中间层，控制着输入层到输出层的信息流。其中，BN 层的节点数远小于其他隐含层的节点数，因而在训练过程中，会对前一层的输入矢量进行非线性降维，将有利于降低分类错误的信息尽可能地压缩并保留在 BN 层，得到一个低维、紧致且富含信息的特征表示。训练结束后，BN 层以后的隐藏层和输出层将被去除，BN 层的线性输出即为基线系统的 BN 特征。

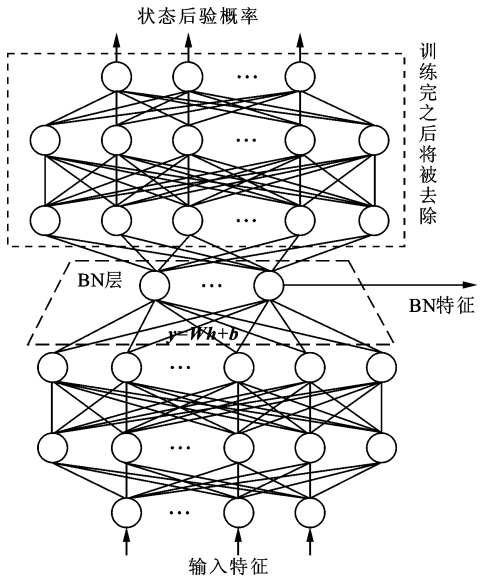


图 1　BN 特征提取流程图

在基线 BN 特征的基础上，进一步根据特征变换目标函数，对 BN 特征进行区分性特征变换。特征变换时，仅调整 BN 特征提取网络中的最后一层（BN 层）权重系数，前 2 层的系数保持不变，如图 1 中梯形虚线框所示。区分性特征变换前后的权重矩阵分别记为 $\boldsymbol{W}_{\mathrm{O}}$ 和 $\boldsymbol{W}^{\mathrm{BN}} \in \mathbf{R}^{n_i \times (n_{i-1}+1)}$，$n_i$ 为第 $i$ 个隐含层的节点数，$\boldsymbol{W}^{\mathrm{BN}}$ 相当于传统特征变换方法中的变换矩阵 $\boldsymbol{M}^{\mathrm{BN}}$，$\boldsymbol{W}_{\mathrm{O}}$ 则对应于 $\boldsymbol{M}_{\mathrm{O}}$。变换后的特征可以表示为

$$\boldsymbol{y}(t) = \boldsymbol{W}^{\mathrm{BN}}\boldsymbol{h} + \boldsymbol{b} \tag{1}$$

式中：$\boldsymbol{h}$ 为 BN 层节点的激励信号，即输入特征 $\boldsymbol{o}(t)$ 经过前 3 层后的输出；$\boldsymbol{b}$ 为偏移矢量。

## 1.2　域相关特征变换

RDLT 方法[4]利用全局的 GMM 模型将声学空

间分成 $R$ 个区域(简称域),通过区分性训练得到一个变换矩阵集合,每个变换矩阵对应于上述声学特征空间划分中的一个区域。用特征向量所属区域对应的变换矩阵对其进行变换,最终变换后的特征有如下形式

$$\boldsymbol{y}(t)=F_{\mathrm{RDLT}}(\boldsymbol{o}(t))=\sum_{i=1}^{R}\kappa_{t,i}(\boldsymbol{A}_i\boldsymbol{o}(t)+\boldsymbol{b}_i) \quad (2)$$

式中:$\boldsymbol{o}(t)$为时刻 $t$ 的输入特征;$\boldsymbol{A}_i$ 为第 $i$ 个域对应的变换矩阵;$\kappa_{t,i}$为 $\boldsymbol{o}(t)$属于第 $i$ 个域的概率,可采用 GMM 高斯分量后验概率来表示。通常,RDLT 方法中变换矩阵 $\boldsymbol{A}_i$ 基于词图(Lattice)信息获得,根据 MPE 准则更新,声学模型参数通过最大似然(maximum likelihood, ML)准则更新。

### 1.3 区分性特征变换目标函数

根据声学模型的区分性训练准则进行变换矩阵的求解,借鉴文献[4-5]中的方法。给定训练语音数据集,最小音素错误准则(minimum phone error, MPE)的目标函数为

$$H_{\mathrm{MPE}}(\boldsymbol{Y},\boldsymbol{\Lambda})=\sum_{z=1}^{Z}\sum_{s_{zi}\in S}P_{\boldsymbol{\Lambda}}(s_{zi}\mid\boldsymbol{y}_z(t))\mathrm{rawacc}(S_{z\mathrm{R}},s_{zi}) \quad (3)$$

式中:$\boldsymbol{y}_z(t)$为语句 $z$ 的特征向量序列;$\boldsymbol{\Lambda}$ 为声学模型参数;$S_{z\mathrm{R}}$为 $\boldsymbol{y}_z(t)$对应的正确识别结果;$s_{zi}$为语句 $z$ 在词图上的候选序列之一;$S$ 为语音识别器产生的所有可能候选词序列所成集合;$\mathrm{rawacc}(S_{z\mathrm{R}},s_{zi})$函数定义为正确识别个数减去插入、删除和替换错误个数;$P_{\boldsymbol{\Lambda}}(s_{zi}\mid\boldsymbol{y}_z(t))$为语句 $z$ 被识别为候选词序列 $s_{zi}$的后验概率。

由式(1)和(2)的特征变换表达形式可知,两者变换矩阵的求解过程相类似,仅从表达形式上看,单纯考虑变换的作用,忽略输入的差别,可以认为式(1)为式(2)的一种特殊情况。采用导数链式法则求解式(3)所示的 MPE 目标函数关于 RDLT 第 $i$ 个区域参数 $\boldsymbol{A}_i$ 的偏导,即

$$\frac{\partial H_{\mathrm{MPE}}(\boldsymbol{Y},\boldsymbol{\Lambda})}{\partial\boldsymbol{A}_i}=\sum_{z,t}\kappa_{t,i}\frac{\partial H_{\mathrm{MPE}}(\boldsymbol{Y},\boldsymbol{\Lambda})}{\partial\boldsymbol{y}_z(t)}\boldsymbol{o}_z^{\mathrm{T}}(t) \quad (4)$$

式中:$\boldsymbol{o}_z(t)$为语句 $z$ 在时刻 $t$ 的特征向量。获得式(4)偏导数后,通过设置步长,采用梯度下降法来更新特征变换参数。对 MFCC 和 BN 特征分别得到 $\boldsymbol{A}_i^{\mathrm{RDLT}}$ 和 $\boldsymbol{A}_i^{\mathrm{BN}}$,进一步根据 $\boldsymbol{A}_i^{\mathrm{RDLT}}$ 和 $\boldsymbol{A}_i^{\mathrm{BN}}$ 构成矩阵 $\boldsymbol{M}^{\mathrm{RDLT}}$和 $\boldsymbol{M}^{\mathrm{BN}}$,即

$$\begin{aligned}\boldsymbol{M}^{\mathrm{RDLT}}&=[\boldsymbol{A}_1^{\mathrm{RDLT}},\boldsymbol{A}_2^{\mathrm{RDLT}},\cdots,\boldsymbol{A}_R^{\mathrm{RDLT}}]\\\boldsymbol{M}^{\mathrm{BN}}&=[\boldsymbol{A}_1^{\mathrm{BN}},\boldsymbol{A}_2^{\mathrm{BN}},\cdots,\boldsymbol{A}_R^{\mathrm{BN}}]\end{aligned} \quad (5)$$

在对 BN 和 MFCC 特征变换过程中,均未考虑偏置项 $\boldsymbol{b}$,实验结果表明去除该项对识别结果影响不大[4-5]。

## 2 子空间域相关的段级特征变换及其正则化方法

上一节中的特征变换方法以帧为处理单元,帧级特征对应时长较小、颗粒度过细,因此特征变换后的区分性提升空间有限,且会出现测试与训练失配问题,为此学者们提出了段级特征。通常段级特征以帧级特征为基础,分段方式可采用固定长度分段和自适应长度分段 2 种,自适应长度分段按照某种准则进行划分更能体现信号特征空间内在的关联关系,常用的方法是采用强制对齐方式进行分段。以语音段为单位进行特征变换的方法即为段级特征变换方法。传统帧级方法中经验设定变换矩阵个数,再根据每一语音帧后验概率值的大小进行选择和加权,而本文的段级特征变换方法是采用子空间变换空间的思想,在区域相关特征变换方法上进行讨论,因此称之为子空间域相关段级特征变换方法。

子空间域相关段级特征变换方法的基本思想是,先利用训练语料中的所有语音段,构造 RDLT 变换子空间字典,即对每一语音段根据其声学统计量信息,利用最大似然准则,得到变换矩阵集合 $\boldsymbol{D}^{\mathrm{RDLT}}=\{\boldsymbol{A}_1^{\mathrm{RDLT}},\boldsymbol{A}_2^{\mathrm{RDLT}},\cdots,\boldsymbol{A}_R^{\mathrm{RDLT}}\}$完成字典构造;每个语音段的特征变换认为是变换子空间的投影,对于每个语音段,由于只是特征空间非常有限的一部分,因此只需要采用字典中的部分变换矩阵即可完成特征变换。在这个过程中需要自动选取变换矩阵并确定相应的系数,本文采用一种自适应变换矩阵选择方法进行区分性特征变换。

设经过域划分后总共有 $R$ 个域,其每一个域对应的变换矩阵为 $\boldsymbol{A}_i$,语音信号被分成 $S$ 段,其中第 $s$ 个语音段的特征变换可以描述为

$$\boldsymbol{o}'_s(t)=\sum_{i=1}^{R}x_{i,s}\boldsymbol{A}_i\boldsymbol{o}_s(t) \quad (6)$$

式中:$x_{i,s}$为所选择的特征变换矩阵 $\boldsymbol{A}_i$ 对应的权重系数。由于以下论述中,均在语音段 $s$ 内求解相关参数,为了行文简化,将下标 $s$ 略去。为了提高特征变换后的识别性能,依据最大似然准则,要使得变换后特征的似然度最大,其目标函数为

$$\hat{\boldsymbol{x}}=\arg\max_{\boldsymbol{x}}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\lg p(\boldsymbol{o}'(t)\mid\boldsymbol{\Lambda}_m)=\arg\max_{\boldsymbol{x}}\left\{-\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)\left[\sum_i x_i\boldsymbol{A}_i\boldsymbol{o}(t)-\boldsymbol{\mu}_m\right]^{\mathrm{T}}\cdot\right.$$

$$\boldsymbol{\Sigma}_m^{-1}\Big[\sum_i x_i \boldsymbol{A}_i \boldsymbol{o}(t)-\boldsymbol{\mu}_m\Big]\Big\} \tag{7}$$

式中：$T$ 为语音段 $s$ 中含有的总帧数；声学模型采用 HMM-GMM 模型，共含有 $M$ 个高斯分量；$\boldsymbol{\mu}_m$ 和 $\boldsymbol{\Sigma}_m$ 分别为第 $m$ 个高斯分量的均值矢量及协方差矩阵；$\gamma_m(t)$ 表示第 $t$ 帧特征属于第 $m$ 个高斯分量的后验概率。

令 $\boldsymbol{O}_i(t)=\boldsymbol{A}_i\boldsymbol{o}(t)$ 为经过域 $i$ 的变换矩阵 $\boldsymbol{A}_i$ 变换后的矢量，则 $\boldsymbol{o}(t)$ 经过所有 $R$ 个域变换后的矢量矩阵为 $\boldsymbol{\xi}_t\in\mathbf{R}^{D\times R}$，则有

$$\begin{aligned}\boldsymbol{\xi}_t &= [\boldsymbol{A}_1\boldsymbol{o}(t),\boldsymbol{A}_2\boldsymbol{o}(t),\cdots,\boldsymbol{A}_R\boldsymbol{o}(t)]=\\ &\quad [\boldsymbol{O}_1(t),\boldsymbol{O}_2(t),\cdots,\boldsymbol{O}_R(t)]\end{aligned} \tag{8}$$

则式(7)可转换为

$$\hat{\boldsymbol{x}}=\arg\max_{\boldsymbol{x}}\Big\{-\frac{1}{2}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_m(t)[\boldsymbol{\xi}_t\boldsymbol{x}-\boldsymbol{\mu}_m]^{\mathrm{T}}\cdot \boldsymbol{\Sigma}_m^{-1}[\boldsymbol{\xi}_t\boldsymbol{x}-\boldsymbol{\mu}_m]\Big\}=\arg\max_{\boldsymbol{x}}\Big[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}+\boldsymbol{f}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{C}\Big] \tag{9}$$

由式(9)可知，子空间域相关段级特征变换方法是一个典型的二次优化问题。对式(9)关于 $\boldsymbol{x}$ 求导，并令导数等于 0，可得

$$\hat{\boldsymbol{x}}=\boldsymbol{G}^{-1}\boldsymbol{f} \tag{10}$$

式中：$\boldsymbol{f}=\sum_{m=1}^{M}\Big[\sum_{t=1}^{T}\gamma_m(t)\boldsymbol{\xi}_t^{\mathrm{T}}\Big]\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\mu}_m$ 和 $\boldsymbol{G}=\sum_{m=1}^{M}\Big[\sum_{t=1}^{T}\gamma_m(t)\boldsymbol{\xi}_t^{\mathrm{T}}\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\xi}_t\Big]$ 分别为广义观测数据的加权一阶和二阶统计量。由于在一次变换过程中数据量有限，得到的方程数常会小于自变量的个数，压缩感知技术为解决这类问题提供了理论支持，可采用正则化方法来解决。引入正则化函数后，特征变换的目标函数变为

$$\hat{\boldsymbol{x}}=\arg\max_{\boldsymbol{x}}\Big[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}\boldsymbol{x}+\boldsymbol{f}^{\mathrm{T}}\boldsymbol{x}-J_{\lambda}(\boldsymbol{x})\boldsymbol{I}\Big] \tag{11}$$

式中：$J_\lambda(\boldsymbol{x})$ 是一个正则化函数，其参数为正则化权重矢量 $\boldsymbol{\lambda}(\boldsymbol{\lambda}>0)$；$\boldsymbol{I}$ 为单位矩阵。当 $J_\lambda(\boldsymbol{x})=\lambda_1\|\boldsymbol{x}\|_1$ 时为 $L_1$ 正则化，当 $J_\lambda(\boldsymbol{x})=\lambda_2\|\boldsymbol{x}\|_2^2$ 时为 $L_2$ 正则化，当 $J_\lambda(\boldsymbol{x})=\lambda_1\|\boldsymbol{x}\|_1+\lambda_2\|\boldsymbol{x}\|_2^2$ 时为弹性网正则化(elastic net regularization)[14]，本文分别讨论了这 3 种正则化。

在目标函数的优化过程中，直接进行变换矩阵子集的选取，非零系数项所对应的变换矩阵将通过线性加权的方式构成最后的变换矩阵，而零系数所对应的变换矩阵不会参与构建，可采用快速迭代收敛阈值算法(fast iterative shrinkage thresholding algorithm, FISTA)[15]进行求解。

## 3　区分性变换后的特征融合方法

不同识别系统可采用不同的特征参数、声学模型、在不同的单位层级(如帧、段)[16]等来构造。通过融合不同的语音识别特征和系统输出结果是提高识别性能的有效方法之一。本文对不同的特征，在不同的时间层次上进行区分性特征变换后，也进一步有效地结合，进行声学模型的训练。本文采用级联(concatenate)的方式融合 2 种特征，融合后的特征可以表示为

$$\boldsymbol{y}_{\mathrm{con}}(t)=[\boldsymbol{M}^{\mathrm{BN}}\quad\boldsymbol{M}^{\mathrm{RDLT}}]\begin{bmatrix}\boldsymbol{h}\\ \boldsymbol{x}\boldsymbol{o}(t)\end{bmatrix} \tag{12}$$

式中：$\boldsymbol{h}$ 为 BN 层节点的激励信号，如式(1)所给出；$\boldsymbol{M}^{\mathrm{BN}}$ 和 $\boldsymbol{M}^{\mathrm{RDLT}}$ 分别为 BN 和 MFCC 特征采用 RDLT 进行特征变换后得到的变换矩阵。

融合后特征所对应的特征变换目标函数为

$$H_{\mathrm{MPE}}(\boldsymbol{Y}_{\mathrm{con}},\boldsymbol{\Lambda};[\boldsymbol{M}^{\mathrm{BN}},\boldsymbol{M}^{\mathrm{RDLT}}])=\sum_{z=1}^{Z}\sum_{s_{zi}\in S}P_{\boldsymbol{\Lambda}}(s_{zi}\mid\boldsymbol{y}_{z,\mathrm{con}}(t))\mathrm{rawacc}(S_{z\mathrm{R}},s_{zi}) \tag{13}$$

式中：变换矩阵 $\boldsymbol{M}^{\mathrm{BN}}$、$\boldsymbol{M}^{\mathrm{RDLT}}$ 分别以式(13)为目标函数进行参数优化和求解，其参数优化过程与 1.3 节类似。需要指出的是，式(13)并不对 $[\boldsymbol{M}^{\mathrm{BN}},\boldsymbol{M}^{\mathrm{RDLT}}]$ 进行联合求解，而是针对每一种变换方法单独进行优化，之所以这样做，是因为联合优化相对单独优化而言更复杂，容易陷入局部极值。采用式(12)的特征融合方式，可以较为方便地讨论单一特征以及特征融合后对识别结果的影响，同时还能加入其他的特征，如噪声谱估计、说话人自适应信息等。

式(12)给出了特征融合的通用性框架，根据 $\boldsymbol{M}^{\mathrm{BN}}$、$\boldsymbol{M}^{\mathrm{RDLT}}$ 选择方式的不同，对应着不同的特征融合方式，区分性变换后的特征融合方法如图 2 所示。在训练阶段，根据特征变换目标函数得到不同时间长度和特征的变换集合或参数；在识别阶段，测试语音基于训练阶段获得的特征变换集合或参数，根据融合方式进行相应特征变换，得到变换后的特征序列进行模型训练和识别。特征有 2 种：一是传统 MFCC 特征及其段级特征；二是帧级 BN 特征。根据这些特征，研究 2 种特征变换方式：一是域相关特征变换，帧级 BN、帧级 MFCC、段级 MFCC 均进行该特征变换；二是网络权值调整特征变换，只采用 BN 特征。

图 2 中，通过设定矩阵 $\boldsymbol{M}^{\mathrm{BN}}$、$\boldsymbol{M}^{\mathrm{RDLT}}$ 的形式，可以研究不同特征融合方法的识别性能。当设定 $\boldsymbol{M}^{\mathrm{BN}}$ 为原 BN 层权值 $\boldsymbol{M}_{\mathrm{O}}$，$\boldsymbol{M}^{\mathrm{RDLT}}$ 为单位阵时，即为帧级的

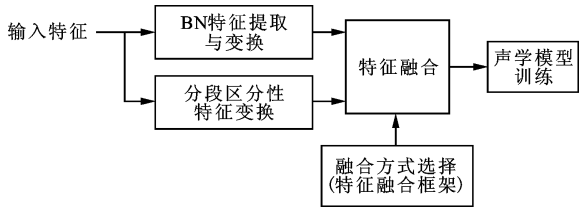


图 2 区分性变换特征融合示意图

MFCC 特征和帧级的 BN 特征级联；当 $\boldsymbol{M}^{\mathrm{BN}}$ 为原 BN 层权值 $\boldsymbol{M}_0$，而 $\boldsymbol{M}^{\mathrm{RDLT}}$ 采用基于子空间域相关段级特征变换方法获得时，则为另外一种融合方法，可表示为 S-RDLT+BN；本文在以下 4.2 节的实验中会继续讨论多种不同的融合方式。在区分性特征变换过程中，当采用基于帧和基于分段相结合的方式时，其具体的实现过程为：在利用强制对齐进行语音分段的基础上，对每一语音段先进行基于分段的特征变换；然后在每一语音分段时间内，逐帧进行特征变换；最后，在每一语音分段内，融合两者变换后的特征进行声学模型参数的调整。

## 4 测试评估

### 4.1 实验设置

本部分具体研究本文区分性特征变换与融合的方法在连续语音识别中的性能。实验语料采用中文微软语料库 Speech Corpora(Version 1.0)，其全部语料在安静办公室环境下录制，采样率为 16 kHz，16 bit 量化。训练集共有 19 688 句，共 454 315 个音节，测试集共 500 句。选择声韵母作为模型基元，零声母(_a、_o、_e、_i、_u、_v)，加上静音(sil)以及常规的声韵母，一共有 69 个模型基元，在此基础上将模型基元扩展为上下文相关的交叉词三音子。特征矢量采用 13 维的美尔倒谱系数(MFCC)及其一阶、二阶差分系数，共 39 维。基于 HTK 3.4 建立基线系统，声学模型采用 3 状态的 HMM 模型，通过决策树对三音子模型进行状态绑定，绑定后的模型有效状态数为 2 843 个。先得到最大似然声学模型，进一步采用增进的互信息准则(boosted maximum mutual information，BMMI)[17]进行区分性训练。

利用得到的 GMM-HMM 系统，对数据进行强制对齐，得到每个状态对应的数据，训练 DNN 模型。DNN 含有 5 个隐含层(共 7 层)，其中 BN 层节点数为 42 个，其他隐含层节点数为 2 048 个，其输入特征由当前帧以及联合前后 5 帧，共 11 帧 MFCC 特征组成，输出节点对应于 GMM-HMM 系统的 2 843个状态。采用 RBM 算法预训练时，每一层的学习率为 0.01，精调整时，前 5 轮的学习率为 0.08，后 5 轮的学习率为 0.02。在使用 BP 算法联合调整参数过程中，计算随机梯度下降时，每一次(mini-batch)的数据样本数为 1 024 个。最后利用得到的 BN 特征训练 BN-GMM-HMM 模型。采用识别准确率作为实验结果的评估标准。

### 4.2 实验结果

基于 MFCC 和 BN 特征，对声学模型分别采用最大似然(ML)准则和区分性训练 BMMI 准则的识别结果如表 1 所示，其中(BN+D)表示在 BN 特征基础上加入 BN 特征的一阶差分，(BN+D+DD)为在 BN 特征基础上加入 BN 特征的一阶、二阶差分构建识别系统。

表 1 采用 2 种准则对 4 种特征的识别性能比较

| 特征 | 识别率/% | |
|---|---|---|
| | ML | BMMI |
| MFCC-GMM-HMM | 73.04 | 75.79 |
| BN-GMM-HMM | 80.42 | 81.37 |
| (BN+D)-GMM-HMM | 80.58 | 81.50 |
| (BN+D+DD)-GMM-HMM | 79.95 | 80.81 |

由表 1 的识别结果可知，采用 BN 特征得到的识别性能明显优于基于 MFCC 特征的识别系统，说明基于深度神经网络能获得有利于识别分类的特征。声学模型区分性训练后识别性能得到进一步的提升。加入 BN 特征的一阶差分对识别结果的影响不大，加入二阶差分之后识别性能反而会降低。这主要是由于在提取 BN 特征过程中，其输入为当前帧以及上下文共 11 帧的长时信息，得到的 BN 特征已经获得了上下文相关信息。另外，BN 特征的提取过程是一个非线性的学习过程，简单的进行线性差分可能不足以描述 BN 特征的这种相关性。因此，在接下来的实验中，未加入 BN 特征的差分系数。

先讨论 MFCC 特征进行区分性特征变换后的识别性能，其全局 GMM 模型由声学模型状态中的高斯聚类得到，最终共有 800 个高斯混元，基于帧 MFCC 特征 RDLT 变换的识别率为 76.92%。将 RDLT 得到的变换矩阵构造一个字典 $\boldsymbol{D}^{\mathrm{RDLT}}=\{\boldsymbol{A}_1^{\mathrm{RDLT}},\boldsymbol{A}_2^{\mathrm{RDLT}},\cdots,\boldsymbol{A}_R^{\mathrm{RDLT}}\}$，共有 800 个字典项，采用 $L_1$、$L_2$ 和弹性网正则化方法选取字典项，进行分段特征变换。在这个过程中，正则化参数 $\lambda_1$、$\lambda_2$ 以及语音信号的分段时长对识别结果具有较大的影响，本文选取了实验性能较好的参数设置，即弹性网正则化方法 $\lambda_1=5$，$\lambda_2=10$。另外，实验发现，固定长度

的分段方法中，段长为 2 s 能得到相对较优的识别结果，说明该长度能得到相对稳定的统计特性。采用强制对齐的分段方法，能得到最高的识别性能。这主要是因为对齐到相同状态的数据具有相类似的声学特性，利用这些数据能估计得到稳健的参数信息。

表 2 给出了 MFCC 和 BN 特征经区分性特征变换后，分别采用 ML 和 BMMI 准则训练声学模型的认别性能。表中 F-RDLT、S-RDLT 分别表示基于帧和基于分段的 RDLT 特征变换方法，G-BN 为基于 BN 特征采用 RDLT 的方法进行特征变换，即将原来 RDLT 方法中的 MFCC 特征替换为 BN 特征进行特征变换，N-BN 表示在 BN 特征提取网络上调整权值进行区分性特征变换。

表 2　区分性特征变换后 2 种准则的识别性能

| 特征 | 识别率/% | |
|---|---|---|
| | ML | BMMI |
| F-RDLT | 76.92 | 78.66 |
| S-RDLT | 78.22 | 80.32 |
| G-BN | 80.91 | 81.83 |
| N-BN | 81.47 | 82.35 |

表 2 中的识别结果表明，基于分段的 RDLT 特征变换方法 S-RDLT 的识别性能优于传统基于帧的特征变换方法 F-RDLT。这主要是由于基于分段的方法利用的是一个相对稳定的语音段信息进行参数的估计，而 F-RDLT 方法仅基于一帧信号，根据后验概率选取变换矩阵，其声学性质易受噪声或是一些发音现象的影响，较难得到稳定的变换矩阵。纵观 MFCC 和 BN 特征，BN 参数的总体性能均优于 MFCC 特征。对 BN 特征进行特征变换时，采用调整网络权值的方法(N-BN)得到最好的识别性能，因为特征提取网络具有更好的特征变换表达和学习能力，因此将区分性目标函数作为目标对特征提取网络的变换矩阵进行学习时，整个 BN 特征提取与变换具有较好的相容性和连贯性，而对 BN 特征直接进行域相关特征变换(即 G-BN)，即将传统的 MFCC 特征替换为 BN 特征，然后采用传统的方法对 BN 特征进行变换，相当于将特征提取和变换分割成较为独立的部分单独进行考虑，较难描述 BN 特征的内在连贯性，进而影响识别性能。

由于各个特征之间具有互补性，因此在区分性特征变换的基础上，进一步在特征和识别候选结果层进行融合。为了便于比较，给出了 7 种融合方式的结果：①MFCC＋BN，即帧级 MFCC 和帧级 BN 进行特征融合；②F-RDLT＋BN，即对 MFCC 进行帧级特征变换后与 BN 特征进行特征融合；③F-RDLT＋N-BN，即对 MFCC 进行帧级特征变换，再对 BN 特征进行网络权值调整变换后，将二者进行特征融合；④S-RDLT＋BN，对 MFCC 进行段级特征变换后与 BN 特征进行特征融合；⑤MFCC＋N-BN，对 BN 特征进行网络权值调整变换后与 MFCC 进行特征融合；⑥S-RDLT＋N-BN，对 MFCC 进行段级特征变换，对 BN 特征进行网络权值调整变换后，将二者进行特征融合；⑦HY-S-RDLT＋HY-N-BN，采用 S-RDLT 和 N-BN 特征变换后得到 2 种候选识别结果，分别为 HY-S-RDLT 和 HY-N-BN，然后对候选识别结果进行融合。表 3 给出了 2 种准则对上述 7 种融合方式下特征的识别性能。

表 3　2 种准则对不同融合方式下特征的识别性能

| 特征 | 识别率/% | |
|---|---|---|
| | ML | BMMI |
| MFCC＋BN | 80.84 | 81.89 |
| F-RDLT＋BN | 80.96 | 82.04 |
| F-RDLT＋N-BN | 81.53 | 82.71 |
| S-RDLT＋BN | 81.15 | 82.20 |
| MFCC＋N-BN | 81.43 | 82.57 |
| S-RDLT＋N-BN | 82.21 | 83.39 |
| HY-S-RDLT＋HY-N-BN | 81.87 | 82.91 |

由表 3 的识别结果可知，采用本文提出的特征融合方式，能通过设置变换矩阵 $\boldsymbol{M}^{\mathrm{BN}}$、$\boldsymbol{M}^{\mathrm{RDLT}}$ 的形式，较为简便地讨论区分性特征变换前后、单一特征以及特征融合对识别结果的影响。当设定 $\boldsymbol{M}^{\mathrm{BN}}$ 为原 BN 层权值 $\boldsymbol{M}_0$，$\boldsymbol{M}^{\mathrm{RDLT}}$ 为单位阵时，即未对特征进行区分性特征变换，直接融合原始 MFCC 和 BN 特征训练声学模型，则识别率会优于采用单一特征的识别系统，说明 MFCC 和 BN 特征具有较好的互补性。虽然 BN 特征的提取过程其输入也为 MFCC，但得到的 BN 特征与 MFCC 特征会有差异性，2 种特征侧重于描述语音的不同声学特性。接着，通过不断调整 $\boldsymbol{M}^{\mathrm{BN}}$、$\boldsymbol{M}^{\mathrm{RDLT}}$ 为区分性特征变换矩阵和单位阵，可以得到不同特征融合方式的识别性能。从实验结果可以看出，融合后的识别率均有不同程度的提升，其中融合 S-RDLT 和 N-BN 这 2 种经过区分性特征变换的特征能获得最佳的识别性能，相对于融合前的 S-RDLT 系统，融合后不同声学模型的识别性能在训练准则下分别提升了 5.1% 和 3.82%；相对于融合前的 N-BN 系统，不同准则下的识别性

能分别提升了 0.9%和 1.26%。这说明通过在帧级和段级不同时间层次上进行特征变换,能进一步提高 2 种特征的互补性。同时发现,特征融合后的识别率会优于融合候选结果得到的识别率。

## 5 结 论

本文提出了一种子空间域相关特征变换与融合的语音识别方法,在区分性特征变换的目标函数下,通过调整神经网络中 BN 层的权值,对 BN 特征利用基于帧的方式进行特征变换,而 MFCC 特征则是采用基于分段的区分性特征变换方法。进一步提出了区分性特征变换与融合的一般框架,通过设置其中特征变换矩阵的形式,得到了不同特征融合方式下的识别性能。实验结果表明,本文基于子空间域相关特征变换与融合的方法能够对多层次的不同特征(如 BN 特征和 MFCC 短时谱)进行区分性表达,融合区分性特征变换前后的特征均能有效地提高识别性能,联合区分性特征变换后的特征进行声学模型训练能得到最高的识别率。由于本文方法是一种通用框架下的联合表示,因此具备多层次不同特征的联合表征能力,后续的研究可以在此框架下加入其他的特征信息(如环境噪声信息、说话人信息等)来提升多信息的联合表征能力。

**参考文献:**


[1] NASERSHARIF B, AKBARI A. SNR-dependent compression of enhanced Mel subband energies for compensation of noise effects on MFCC features [J]. Pattern Recognition Letters, 2011, 28(11): 1320-1326.

[2] 刘晓明,班超帆,冯晓荣. 失真控制下的短时谱估计语音增强算法 [J]. 西安交通大学学报, 2011, 45(8): 78-84.
LIU Xiaoming, BAN Chaofan, FENG Xiaorong. A short time spectrum estimation algorithm of speech enhancement under the distortion control [J]. Journal of Xi'an Jiaotong University, 2011, 45(8): 78-84.

[3] POVEY D, KINGSBURY B, MANGU L, et al. fMPE: Discriminatively trained features for speech recognition [C]// Proceedings of the International Conference on Audio, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2005: 961-964.

[4] ZHANG B, MATSOUKAS S, SCHWARTZ R. Recent progress on the discriminative region-dependent transform for speech feature extraction [C]// Proceedings of the Annual Conference of International Speech Communication Association. Baixs, France: ISCA, 2006: 1495-1498.

[5] YAN Z, HUO Q, XU J, et al. Tied-state based discriminative training of context-expanded region-dependent feature transforms for LVCSR [C]// Proceedings of the International Conference on Audio, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2013: 6940-6944.

[6] 高莹莹,朱维彬. 深层神经网络中间层可见化建模 [J]. 自动化学报, 2015, 41(9): 1627-1637.
GAO Yingying, ZHU Weibin. Deep neural networks with visible intermediate layers [J]. Acta Automatica Sinica, 2015, 41(9): 1627-1637.

[7] 袁胜龙,郭武,戴礼荣. 基于深层神经网络的藏语识别 [J]. 模式识别与人工智能, 2015, 28(3): 209-213.
YUAN Shenglong, GUO Wu, DAI Lirong. Speech recognition based on deep neural networks on Tibetan corpus [J]. Pattern Recognition and Artificial Intelligence, 2015, 28(3): 209-213.

[8] SAINATH T N, KINGSBURY B, RAMABHADRAN B. Auto-encoder bottleneck features using deep belief networks [C]// Proceedings of the International Conference on Audio, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2012: 4153-4156.

[9] SAON G, KINGSBURY B. Discriminative feature-space transforms using deep neural networks [C]// Proceedings of the Annual Conference of International Speech Communication Association. Baixs, France: ISCA, 2012: 14-17.

[10] PAULIK M. Lattice-based training of bottleneck feature extraction neural networks [C]// Proceedings of the Annual Conference of International Speech Communication Association. Baixs, France: ISCA, 2013: 89-93.

[11] LIU D Y, WEI S, GUO W, et al. Lattice based optimization of bottleneck feature extractor with linear transformation [C]// Proceedings of the International Conference on Audio, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2014: 5617-5621.

[12] YU D, SELTZER M L. Improved bottleneck features using pretrained deep neural networks [C]// Proceedings of the Annual Conference of International Speech Communication Association. Baixs, France: ISCA, 2011: 237-240.

[13] HOFFMEISTER B, KLEIN T, SCHLUTER R, et al. Frame based system combination and a comparison with weighted ROVER and CNC [C]// Proceedings of the International Conference on Spoken Language Pro-

cessing. Piscataway, NJ, USA: IEEE, 2006: 537-540.

[14] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. Journal of the Royal Statistical Society: Series B Statistical Methodology, 2005, 67(2): 301-320.

[15] BECK A, TEBOULLE M. A fast iterative shrinkage thresholding algorithm for linear inverse problems [J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202.

[16] CHEN X, ZHAO Y. Building acoustic model ensembles by data sampling with enhanced trainings and features [J]. IEEE Transactions on Audio Speech and Language Processing, 2013, 21(3): 498-507.

[17] POVEY D, KANEVSKY D, KINGSBURY B, et al. Boosted MMI for model and feature space discriminative training [C] // Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ, USA: IEEE, 2008: 4057-4060.

（编辑　刘杨）

（上接第 40 页）

LUO Guangming, HUANG Xiaoyu, ZHU Jianlin. Computer simulation of fuzzy self-tuning PID control based on MATLAB [J]. Machinery and Electronics, 2001(2): 23-26.

[11] KARASAKAL O, GUZELKAYA M, EKSIN I, et al. Online tuning of fuzzy PID controllers via rule weighing based on normalized acceleration [J]. Engineering Applications of Artificial Intelligence, 2013, 26(1): 184-197.

[12] WAEWSAK Y C, NOPHARATANA A, CHAIPRASERT P, et al. Neural-fuzzy control system application for monitoring process response and control of anaerobic hybrid reactor in wastewater treatment and biogas production [J]. Journal of Environmental Sciences, 2010, 22(12): 1883-1890.

[13] 封琦. 基于模糊控制的 Boost 有源功率因数校正器设计与研究 [D]. 南京：河海大学，2006.

[14] 倪骏康，刘崇新，庞霞. 电力系统混沌振荡的等效快速终端模糊滑模控制 [J]. 物理学报，2013，62(19)：190507.
NI Jun-Kang, LIU Chong-Xin, PANG Xia. Fuzzy fast terminal sliding mode controller using an equivalent control for chaotic oscillation in power system [J]. Acta Physica Sinica, 2013, 62(19): 190507.

（编辑　刘杨）

（上接第 53 页）

[9] PILLAY N, XU H. Comments on antenna selection in spatial modulation systems [J]. IEEE Communications Letters, 2013, 17(9): 1681-1683.

[10] NTONTIN K, DI RENZO M, PEREZ-NEIRA A I, et al. A low-complexity method for antenna selection in spatial modulation systems [J]. IEEE Communications Letters, 2013, 17(12): 2312-2315.

[11] 龚丽莎，尹露，龚赛丹，等. 基于空间调制的天线选择和能效优化算法 [J]. 电子科技大学学报，2014，43(4)：497-501.
GONG Li-sha, YIN Lu, GONG Sai-dan, et al. Transmit antenna selection and energy efficiency improvement for spatial modulation [J]. Journal of University of Electronic Science and Technology of China, 2014, 43(4): 497-501.

[12] ZHENG J, CHEN J. Further complexity reduction for antenna selection in spatial modulation systems [J]. IEEE Communications Letters, 2015, 19(6): 937-940.

[13] WANG Wen-Hsin, CHANG R Y. Signal-spatial constellation optimization for generalized spatial modulation [C] // IEEE 79th Vehicular Technology Conference. Piscataway, NJ, USA: IEEE, 2014: 1-5.

[14] NING M, ANGUO W, CHANGCAI H, et al. Adaptive Joint Mapping Generalised Spatial Modulation [C] // IEEE International Conference on Communications in China. Piscataway, NJ, USA: IEEE, 2012: 520-523.

（编辑　刘杨）