基于 MCE 准则的语音识别特征线性判别分析

陈 斌 [1] 张连海 [1] 牛 铜 [1] 屈 丹 [1] 李弼程 [1]

摘 要 提出了一种基于最小分类错误 (Minimum classification error, MCE) 准则的线性判别分析方法 (Linear discriminant analysis, LDA), 并将其应用到连续语音识别中的特征变换. 该方法采用非参数核密度估计方法进行数据概率分布估计; 根据得到的概率分布, 在最小分类错误准则下, 采用基于梯度下降的线性搜索算法求解判别分析变换矩阵. 利用判别分析变换矩阵对相邻帧梅尔滤波器组输出拼接的超矢量变换降维, 得到时频特征. 实验结果表明, 与传统的 MFCC 特征相比, 经过本文判别分析提取的时频特征其识别准确率提高了 1.41 %, 相比于 HLDA (Heteroscedastic LDA) 和近似成对经验正确率准则 (Approximate pairwise empirical accuracy criterion, aPEAC) 判别分析方法, 识别准确率分别提高了 1.14 % 和 0.83 %.

关键词 线性判别分析, 语音识别, 核密度估计, 特征变换

引用格式 陈斌, 张连海, 牛铜, 屈丹, 李弼程. 基于 MCE 准则的语音识别特征线性判别分析. 自动化学报, 2014, **40**(6): 1208−1215

DOI 10.3724/SP.J.1004.2014.01208

A Minimum Classification Error Criterion Based Linear Discriminant Analysis Method for Speech Recognition Feature

CHEN Bin[1] ZHANG Lian-Hai[1] NIU Tong[1] QU Dan[1] LI Bi-Cheng[1]

Abstract A linear discriminant analysis (LDA) method based on the minimum classification error criterion is proposed, and further it is applied to the continuous speech recognition feature transformation. The data probability distribution is estimated using non-parametric kernel density estimation method. According to the obtained probability distribution, a gradient descent based linear search procedure is performed to get the discriminant analysis transformation matrix under the minimum classification error criterion. The dimensionality of super-vector conjoined by the adjacent frames Mel filter bank output is reduced with the transformation matrix, and then after dimensionality reduction the time-frequency feature is acquired. Experimental results show that compared with the traditional MFCC feature, the recognition accuracy rate of the time-frequency feature extracted with the presented discriminant analysis method has a 1.41 % improvement. In contrast with the HLDA and aPEAC discriminant analysis feature transformation method, the recognition accuracy of the presented method increases by 1.14 % and 0.83 % separately.

Key words Linear discriminant analysis (LDA), speech recognition, kernel density estimation, feature transformation

Citation Chen Bin, Zhang Lian-Hai, Niu Tong, Qu Dan, Li Bi-Cheng. A minimum classification error criterion based linear discriminant analysis method for speech recognition feature. *Acta Automatica Sinica*, 2014, **40**(6): 1208−1215

噪声环境下, 语音识别性能急剧下降, 针对此问题, 常采用特征变换的方法[1−2] 以提高特征的鲁棒性和区分性. 根据特征变换是否与分类器参数进行联合调整, 可以将特征变换分为线性判别分析方法和区分性特征提取 (Discriminative feature extraction, DFE) 方法. 其中基于判别分析的特征变换是将特征变换到区分性特征空间的有效方法, 同时可以通过对拼接的多帧特征变换降维得到稳定的长时特征 (如 Spliced feature).

线性判别分析 (Linear discriminant analysis, LDA)[3], 由于求解简单, 且能够通过核函数拓展为非线性判别分析, 进而得到广泛关注. 其目标为在高斯分布的同方差性假设下, 使得类间–类内散度比最大化. HLDA (Heteroscedastic LDA)[4] 在 LDA 的基础上, 假设各类别的方差不相同, 基于最大似然 (Maximum likelihood) 准则进行变换矩阵的求解. HDA (Heteroscedastic discriminant analysis)[5]、PLDA (Power LDA)[6] 同样在异方差性的假设下, 分别对各类别的方差求取几何平均和乘幂平均. 由于后端分类器的非线性和不同类间分离距离的定义, 会出现特征变换后具有较大的分离距离, 而后端分类性能不佳的情况, 此时需要

收稿日期 2013-07-15 录用日期 2013-10-01
Manuscript received July 15, 2013; accepted October 1, 2013
国家自然科学基金 (61175017) 资助
Supported by National Natural Science Foundation of China (61175017)
本文责任编委 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 解放军信息工程大学信息系统工程学院 郑州 450002
1. Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou 450002

根据分离距离和成对 (Pairwise) 经验分类性能对线性判别分析得到的投影向量加权, 即加权 LDA (Weighted LDA, WLDA). Loog 等[7] 提出了近似成对理论正确准则 (Approximate pairwise theoretical accuracy criterion, aPTAC) 替代 Fisher 准则, 利用贝叶斯正确率对投影向量加权. Lee 等[8] 提出了近似成对经验正确率准则 (Approximate pairwise empirical accuracy criterion, aPEAC), 利用经验分类正确率对投影向量加权. 近年来, 不少学者将判别分析的方法应用到说话人识别[9−10]、连续语音识别[11−13] 等语音识别领域中, 并取得了一定的识别性能提升.

上述判别分析方法一般都假设数据服从单高斯分布, 在最大似然准则下求取变换矩阵. 无论是同方差性还是异方差性, 均较难有效地处理分布复杂的数据. 最大似然准则只考虑增大正确类别的似然度, 并未考虑错误分类的信息, 可能会造成在提高正确类别相似度的同时, 错误类别的似然度变得更高. 声学模型的区分性训练[14] 是目前提高语音识别性能的有效方法, 区分性训练准则不仅要求数据在正确类别的相似度越高越好, 同时要求在错误类别的相似度越低越好. 其中最小分类错误 (Minimum classification error, MCE)[15] 准则可以有效地应用于特征变换和声学模型的调整.

为了提高特征变换和后端分类器的关联性, 提出了基于 MCE 的区分性特征提取方法, 该类方法期望对特征变换和分类器参数进行联合优化, 求取全局最优解, 如滤波器组的设计[16]、结合声学模型的特征变换[17−18]、动态特征的提取[19] 等方法. 但该类方法需要同时优化较多的参数, 算法复杂度较高, 主要用于数字、人名等孤立词识别上, 这种情况下类别和声学模型数较少. 随着数据的增大和识别任务复杂度的提高 (如大词汇量连续语音识别), 算法复杂度会迅速的增大, 参数较难同时收敛.

由于现有判别分析方法对数据分布作假设, 使得分布复杂的数据识别性能不高, 区分性特征提取方法将特征变换和声学模型参数联合优化, 算法复杂度较高, 针对上述问题, 提出了一种基于 MCE 准则的线性判别分析 (MCE linear discriminant analysis, MCELDA) 方法, 该方法不对数据的分布作假设, 采用区分性准则进行前端特征变换, 将特征变换与模型分开优化, 有效地降低算法复杂度, 同时提高类别之间的区分性. 通过采用核密度估计 (Kernel density estimation) 方法进行概率分布的估计, 在估计得到数据真实分布的基础上, 采用最小分类错误的区分性准则进行特征变换, 使得所有数据经过特征变换后的分类损失函数最小. 并利用本文方法对相邻帧梅尔滤波器组输出拼接的超矢量特征变换降维, 得到时频特征进行语音识别, 最后对识别性能进行了讨论.

1 基于 MCE 准则特征变换的分类损失函数

本文采用 MCE 准则的主要思想, 即不仅考虑正确分类样本同时考虑错误分类样本, 使得错误分类损失函数最小, 进行特征变换. 而不直接利用传统的 MCE 准则进行声学模型参数的调整, 以降低复杂度. 即本文根据估计得到的概率分布, 先基于 MCE 准则得到所有数据的分类损失函数, 再寻求一个特征变换矩阵 W, 使得变换后所有数据的分类损失函数最小.

设原始空间中的特征向量为 $\boldsymbol{x} \in \mathbf{R}^n$, 经过线性变换 W 后为 $\boldsymbol{y} \in \mathbf{R}^p$, 可以描述为

$$\boldsymbol{y} = W\boldsymbol{x} \tag{1}$$

其中, $W \in \mathbf{R}^{p\times n}$ 为一个由 n_w 列向量组成的投影变换矩阵, 不同的特征变换方法主要体现在 W 矩阵上.

假设训练数据总共有 M 类, $m = \{1, \cdots, M\}$, $g(i)$ 也为类别标识号, 让 $g(i) \to \{1, \cdots, M\}$ 对应, $n^{g(i)}$ 表示类别 $g(i)$ 所对应的训练样本集 $\{y_i\}$ 的总数, N 为训练样本的总数, $\sum_{g(i)=1}^{M} n^{g(i)} = N$, $\sum_{m=1}^{M} n^m = N$, 变换后的特征 y_i 基于 MCE 的分类损失函数可以定义为

$$\begin{aligned} d_{g(i)}(y_i, \Lambda) = & -\log P(y_i|\Lambda_{g(i)}) + \\ & \frac{\varepsilon}{M-1} \sum_{m=1, m\neq g(i)}^{M} \log P(y_i|\Lambda_m) \end{aligned} \tag{2}$$

其中, y_i 属于第 $g(i)$ 类, $\log P(y_i|\Lambda_{g(i)})$ 表示特征 y_i 在所属类别下的概率, $\Lambda_{g(i)}$ 表示 $g(i)$ 类别特征的概率分布参数信息, $\frac{\varepsilon}{M-1}\sum_{m=1,m\neq g(i)}^{M} \log P(y_i|\Lambda_m)$ 表示平均分类错误的概率, ε 用来控制平均分类错误的权重, ε 越大越重视平均分类错误所带来的影响, 分类损失函数 $d_{g(i)}(y_i, \Lambda)$ 越小表示分类越正确.

目标函数为使得所有数据的分类损失函数值最小, 具体的推导过程见附录 A, 可得:

$$\begin{aligned} R_{\mathrm{MCE}} = & \sum_{i=1}^{N} d_{g(i)}(y_i|\Lambda) = \\ & \sum_{g(i)=1}^{M} \sum_{i=1}^{n_{g(i)}} \Bigg[-\log P(y_{i,g(i)}|\Lambda_{g(i)}) + \\ & \frac{\varepsilon}{M-1} \sum_{m=1, m\neq g(i)}^{M} \log P(y_{i,g(i)}|\Lambda_m) \Bigg] = \end{aligned}$$

$$-\left(1+\frac{\varepsilon}{M-1}\right)\sum_{g(i)=1}^{M}\sum_{i=1}^{n_{g(i)}}\log P(y_{i,g(i)}|\Lambda_{g(i)})+$$

$$\frac{\varepsilon}{M-1}\sum_{m=1}^{M}\sum_{i=1}^{N}\log P(y_i|\Lambda_m) \tag{3}$$

令 $g_{\Lambda_{g(i)}}(y_i) = \sum_{i=1}^{n_{g(i)}}\log P(y_{i,g(i)}|\Lambda_{g(i)})$, $G_{\Lambda_m}(y_i)=\sum_{i=1}^{N}\log P(y_i|\Lambda_m)$. 式 (3) 中概率分布 $p_y(\boldsymbol{y}|\Lambda)$ 可以根据相应类别所对应的数据集, 采用核密度估计 (Kernel density estimation, KDE)[20−21] 方法进行估计:

$$p_y(\boldsymbol{y}|\Lambda)=\frac{1}{l}\sum_{i=1}^{l}K_h(\boldsymbol{y}-\boldsymbol{y}_i) \tag{4}$$

其中, l 为数据集中的样本数, K 为核函数, h 为核宽 (Bandwidth). h 为一个常数, 影响着所估计概率分布的平滑度, 本文采用文献 [22] 的核宽估计方法, K 采用高斯形式:

$$K_h(\boldsymbol{y}-\boldsymbol{y}_i)=\alpha\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{y}_i)^{\mathrm{T}}\Psi^{-1}(\boldsymbol{y}-\boldsymbol{y}_i)\right) \tag{5}$$

其中, α 因子用来保证 $p_y(\boldsymbol{y}|\Lambda)$ 的积分为 1, Ψ 为对角矩阵, 每一对角元素为

$$\psi_{k,k}=\xi\frac{1}{l-1}\sum_{i=1}^{l}(y_{ik}-\overline{\boldsymbol{y}}_k)^2 \tag{6}$$

其中, $\overline{\boldsymbol{y}}_k$ 是 $\boldsymbol{y}_k$ 的经验均值, 根据平滑策略[23], 将参数 ξ 设置成 $\xi=(\frac{4}{3l})^{0.1}$.

2 基于梯度下降法的变换矩阵求解

进一步需要求得一个变换矩阵 W 使得式 (3) 的分类损失函数最小, 由于目标函数较为复杂, 较难直接得到闭式的解析解, 本文采用基于梯度下降的线性搜索算法进行变换矩阵的求解. 式 (3) 中第 k 维投影向量为 $\boldsymbol{w}_k$, 具体的推导过程见附录 B:

$$\nabla_{\boldsymbol{w}_k}R_{MCE}=-(1+\frac{\varepsilon}{M-1})\sum_{g(i)=1}^{M}\nabla_{\boldsymbol{w}_k}g_{\Lambda_{g(i)}}(y_i)+$$

$$\frac{\varepsilon}{M-1}\sum_{m=1}^{M}\nabla_{\boldsymbol{w}_k}G_{\Lambda_m}(y_i)=$$

$$\left[-(1+\frac{\varepsilon}{M-1})\sum_{g(i)=1}^{M}\boldsymbol{F}^{g(i)}+\frac{\varepsilon}{M-1}\sum_{m=1}^{M}\boldsymbol{F}^m\right]\boldsymbol{w}_k \tag{7}$$

由于式 (7) 的表达形式较为复杂, 本文采用沿着梯度方向, 利用迭代的方式进行变换矩阵 W 求取:

$$\boldsymbol{w}_k^{(\mathrm{iter}+1)}=\boldsymbol{w}_k^{(\mathrm{iter})}+\lambda\nabla_{\boldsymbol{w}_k}R_{\mathrm{MCE}}^{(\mathrm{iter})} \tag{8}$$

式中 λ 为步长.

根据上述的推导过程, 变换矩阵的求解步骤如算法 1 所示. 先初始化初值 $W(0)$, 计算类别的协方差矩阵 $\phi=(\boldsymbol{x}-\overline{\boldsymbol{x}})(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathrm{T}}$, 其中 $\overline{\boldsymbol{x}}$ 为 $\boldsymbol{x}$ 均值; 接着根据变换矩阵 $W^{(\mathrm{iter})}$, 分别计算正确和错误类别的梯度 $\boldsymbol{F}^{g(i)}$、$\boldsymbol{F}^m$, 在此基础上求解 $W^{(\mathrm{iter})}$ 的梯度, 设置步长和增加值沿着梯度方向搜索局部最小值, 如果没有找到局部最小值则步长和增加值减小一倍, 重新搜索; 最后如果分类损失函数值的变化率小于门限值 10^{-4} 或是迭代次数大于设定值 I, 则停止迭代.

算法 1. 基于最小分类错误准则判别分析的变换矩阵求解过程

Require. M 个类别的训练样本 $\{x_i\}_{i=1}^N$;

Ensure. 线性变换矩阵 W;

1) 初始化. 计算 $\phi^{g(i)}$, ϕ^m, 设置迭代次数 iter = 0, 利用 LDA 得到初始化变换矩阵 $W(0)$, $W(0)$ 为全变换矩阵 (即不降维);

2) **repeat**

3) 计算 $\boldsymbol{F}^{g(i)}$, $\boldsymbol{F}^m$, 根据式 (7) 求取 $W^{(\mathrm{iter})}$ 的梯度;

4) 沿着梯度方向进行线搜索, 设置步长 $\lambda=0.002$, 每次增加值 0.002, 步长变化范围为 $[0, 0.02]$, 如果没有找到局部最小值, 则将步长初值和增加值减小一倍, 直到找到局部最小值, 将此时的变换矩阵设置为 $W^{(\mathrm{iter}+1)}$;

5) 迭代次数 iter = iter + 1;

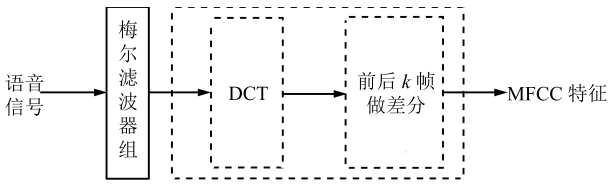
6) **until** 分类损失函数值 R_{MCE} 的变化率小于 10^{-4} 或迭代次数大于 I.

3 特征参数提取

在得到变换矩阵 W 的基础上, 进行特征参数的提取. 本文分别提取了传统的 MFCC 特征和时频特征, 其提取示意图如图 1 所示, 二者的差别主要在于对梅尔滤波器组输出的处理方式不同.

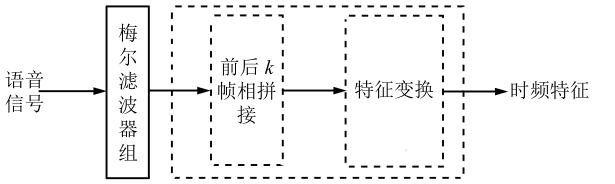
图 1 (a) 中 MFCC 特征提取是将 18 个梅尔频谱滤波器组 (Filter banks) 的输出经离散余弦转换 (Discrete cosine transform, DCT) 后求得 13 维梅尔倒频谱系数, 以及它们的一阶、二阶差分共 39 维作为特征矢量. 图 1 (b) 中时频域特征的提取是拼接前后相邻 k 个梅尔滤波器组的输出, 组成一个超矢量, 再对超矢量进行判别分析特征变换降维, 具体的提取过程如图 2 所示. 通过强制对齐得到状态序列, 将每一帧特征依据其状态占有率分配到相应的状态中, 进行判别分析. 为了能够定性地与现有方法进行比较, 文中选取 $k=2$, 特征变换前维数为

$18 \times (2 \times 2 + 1) = 90$ 维, 经过特征变换降维后特征矢量为 39 维.



(a) MFCC 特征提取示意图

(a) The illustration of MFCC feature extraction



(b) 时频特征提取示意图

(b) The illustration of time-frequency feature extraction

图 1　MFCC 和时频特征提取示意图

Fig. 1　The illustration of MFCC and time-frequency feature extraction

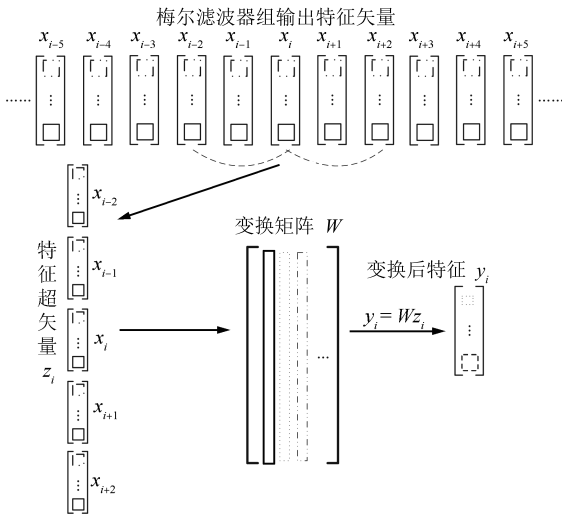


图 2　时频特征的提取过程

Fig. 2　The flow chart of time-frequency feature extraction

隐马尔科夫模型 (Hidden Markov model, HMM) 假设状态观测概率分布的协方差矩阵为对角矩阵, 经过特征变换后的特征并不一定能满足该假设, 这会造成 HMM 模型参数有较大的估计失真. 为此文中在特征变换后进一步进行最大似然线性转换 (Maximum likelihood linear transformation, MLLT)[5], 将变换后特征的协方差矩阵对角化, 其变换方式为

$$W_{\mathrm{MLLT}} = \arg\max_{W} \sum_{i=1}^{C} -\frac{n_{g(i)}}{2} \times \left|\mathrm{diag}\{W^{\mathrm{T}} \tilde{\phi}^{g(i)} W\}\right| + N \log |W| \tag{9}$$

4　测试评估

4.1　实验设置

接下来将本文判别分析方法进行特征变换应用到连续语音识别中. 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 16 000 Hz, 16 bit 量化. 训练集共有 19 688 句, 共 454 315 个音节, 测试集共 500 句.

对语料进行统计, 共有 404 个无调音节, 文中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 (Cross-word tri-phone) 基元, 简称为三音子.

基于 HTK 3.4 建立基线系统, 声学模型采用 3 状态的自左至右无跨越的 HMM 来建模无调音节的声韵母单元, 在三音子扩展时, 根据前后所有可能声韵母进行扩展, 通过决策树对三音子模型进行状态绑定, 解决数据稀疏的问题, 最终绑定后的模型有效状态数为 2 843 个. 其中单音子模型的训练是通过计算全局均值与方差作为全局初始模型, 采用 HMM 模型参数嵌入式重估 (B-W) 算法, 训练所有单音子基元, 然后添加 Sil 模型和 Sp 模型, 重新训练模型; 最后对齐训练语料, 获取语料的最优标注, 再次训练模型. 在训练完单音子模型后, 将单音子扩展为上下文相关的三音子模型; 设定问题集, 利用脚本生成决策树, 根据决策树执行状态绑定后重估四遍, 从而得到跨词边界的上下文相关三音子模型; 最后, 加入 Sp 模型, 将其与 Sil 模型的第三个状态共享. 利用 SRILM 工具根据语料库中自有的标注文件训练得到语言模型, 采用 Katz 语言模型平滑技术.

采用准确率进行实验结果的评估, 假设测试语料实际词总数为 N, 正确识别的词总数记为 H, 删除错误的词总数记为 D, 插入错误的词总数记为 I, 替代错误的词总数记为 S, 准确率 $A\%$ 定义如式 (10):

$$A\% = \frac{H - I}{N} \times 100\% \tag{10}$$

4.2　实验结果

由 MCELDA 变换矩阵的求解过程可知, 主要有两个参数可以调整, 一个是控制平均分类错误权重的 ε 因子和迭代次数 I. 首先确定 ε 因子, 由图 3 可知, ε 设定为 400 较为合理. 固定 $\varepsilon = 400$ 后, 再确定迭代次数, 由图 4 可知, 设定为 $I = 7\,000$ 识别

率最高.

接下来将本文方法与现有的 LDA[3]、HLDA[4]、aPEAC[8] 特征变换方法的识别性能进行比较, 所有特征变换方法均提取时频特征, 后端的声学建模和训练采用相同的方法, 如第 4.1 节所述.

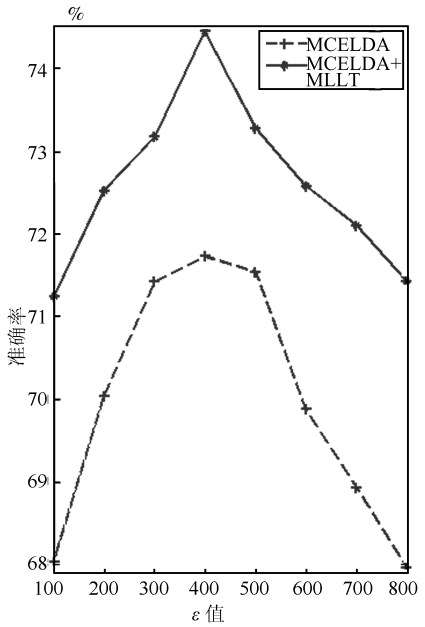


图 3 ε 因子对识别率的影响

Fig. 3 The influences of the ε parameters on the recognition rate

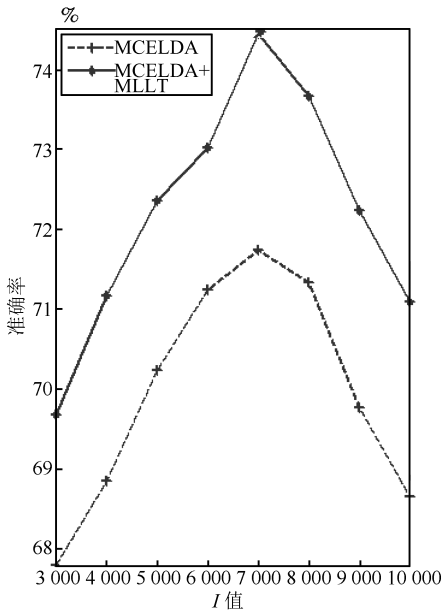


图 4 迭代次数 I 对识别率的影响

Fig. 4 The influences of the iteration times I parameters on the recognition rate

由表 1 可以看出特征变换后, 进行 MLLT 对角变换的识别率都有较为明显的提高, 没有 MLLT 对角变换准确率甚至可能会低于传统 MFCC 特征的方法. 这主要是因为特征变换时主要是考虑建模单元之间的区分性, 并不能保证每一类别的协方差矩阵对角化, 会造成后端 HMM 声学建模参数估计的较大失真, 这也表明经过 MLLT 变换后再进行识别性能的比较较为合理. 由实验结果可知, 采用所有判别分析方法得到的时频特征, 其识别准确率均会优于采用 MFCC 特征的识别方法, 说明时频特征具有区分性的同时也能部分刻画时间的相关性. 本文方法的识别性能会优于其他方法, HLDA 与 aPEAC 方法的性能相当, aPEAC 方法性能略优. 这主要是现有的方法大多数都是采用单高斯分布进行数据的描述, 只考虑了数据的一阶、二阶统计信息, 而没有考虑数据的实际分布情况. 而本文方法在估计数据真实分布的基础上, 采用最小分类错误的准则进行特征变换, 能进一步提高类别之间的区分性和描述数据分布结构. 另外 aPEAC 方法完全用经验正确率替代类别间的马氏距离作为分类度量, 这就要求训练和测试环境相同, 当训练集和测试集类别间的混淆度不一致时, 会对识别性能有较大的影响.

表 1 不同特征变换方法的识别准确率 (%)

Table 1 Accuracy of different feature transform methods (%)

采用方法	无 MLLT 准确率	有 MLLT 准确率
MFCC	73.06	73.04
LDA	69.74	73.11
HLDA	71.19	73.31
aPEAC	71.28	73.62
MCELDA	71.73	74.45

由于本文方法依据估计的数据分布进行特征变换, 可以根据数据样本进行变换矩阵的适应和调整. 而其他方法均假设数据服从高斯分布, 要求数据量足够的大, 当数据减少训练不充分时, 对识别性能会有较大的影响. 表 2 为训练语句减少后的识别性能, 其中训练语句减少为 13 143 句, 共 303 014 个音节, 在减少训练语料的同时保证了所有的声学单元都得到了训练, 后端采用与前面相同的声学建模方法, 测试语句 500 句. 由表 2 的实验结果可知, 本文方法相比于其他方法识别性能下降的更少.

现在语音识别模型训练的标准流程中, 常会对特征进行变化后, 再做模型的区分性训练. 接下来讨论经过本文特征变换后, 再进行声学模型区分性训练的识别性能. 主要采用常用的最大互信息 (Maximum mutual information, MMI), 最小音素错误 (Minimum phone error, MPE) 区分性训练准则和

最大似然准则 (Maximum likelihood, ML), 各种特征变换方法经过模型区分性训练后的识别结果如表 3 所示, 均经过十次迭代训练.

表 2　训练数据减少后不同特征变换方法的识别准确率 (%)

Table 2　Accuracy of different feature transform methods after training data reduction (%)

采用方法	无 MLLT 准确率	有 MLLT 准确率
MFCC	60.91	60.98
LDA	60.37	61.64
HLDA	59.98	62.15
aPEAC	59.76	61.93
MCELDA	62.72	65.17

表 3　不同特征变换方法经过模型区分性训练后的识别准确率 (%)

Table 3　Accuracy of different feature transform methods after model discriminant training (%)

采用方法	ML	MMI	MPE
MFCC	73.04	75.79	76.16
LDA	73.11	76.52	76.91
HLDA	73.31	76.32	76.61
aPEAC	73.62	76.20	76.67
MCELDA	74.45	77.62	78.12

由表 3 可知, 在本文特征变换的基础上, 进行区分性训练能进一步提高语音识别的性能, 识别准确率会高于其他特征变换方法. 从识别准确率提高的相对幅度来看, LDA 变换方法的提高幅度最大, 其次是本文方法, 也说明了 LDA 和本文方法特征变换与区分性训练方法均具有较好的互补性.

现有的识别系统不可避免地会受到噪声的影响, 进一步验证本文方法克服噪声的能力. 对微软语料库加入了 Aurora 2.0 不同信噪比下的白噪声、粉红噪声、火车噪声, 不同方法的鲁棒性测试结果如表 4 所示.

由表 4 可知本文方法在所有方法中下降的相对幅度是最低的, 其中 aPEAC 受噪声的影响最大, 这主要是因为其对训练集和测试集类别间混淆度的一致性要求较高. 由识别结果可知, 本文方法在干净语音和带噪语音中均能得到较高的识别率, 说明本文方法能够得到具有较好区分性的特征, 同时具有较好的鲁棒性.

5　结论

本文提出了一种基于最小分类错误准则的线性判别分析方法, 该方法没有对数据样本的分布作高斯或其他分布的假设, 采用非参数核密度的方法估计数据样本的概率分布, 在此基础上采用最小分类错误的准则进行特征变换, 提高了类别之间的区分性, 并将其应用到连续语音识别中, 实验结果表明了本文方法的有效性. 文中变换矩阵的求取采用的是基于梯度下降法, 后续的研究可以采用更好的优化算法以提高收敛速度和性能. 另外本文提出的判别分析方法也可以应用到其他的模式分类问题中.

附录 A　分类损失函数

$$
\begin{aligned}
R_{\mathrm{MCE}} = & \sum_{i=1}^{N} d_{g(i)}(y_i|\Lambda) = \\
& \sum_{g(i)=1}^{M} \sum_{i=1}^{n_{g(i)}} \Big[-\log P(y_{i,g(i)}|\Lambda_{g(i)}) + \\
& \frac{\varepsilon}{M-1} \sum_{m=1, m\neq g(i)}^{M} \log P(y_{i,g(i)}|\Lambda_m) \Big] = \\
& \sum_{g(i)=1}^{M} \sum_{i=1}^{n_{g(i)}} \Bigg[-\log P(y_{i,g(i)}|\Lambda_{g(i)}) + \\
& \frac{\varepsilon}{M-1} \Bigg(\sum_{m=1, m\neq g(i)}^{M} \log P(y_{i,g(i)}|\Lambda_m) +
\end{aligned}
$$

表 4　不同方法的鲁棒性测试结果 (%)

Table 4　Robust test results of different feature transform methods (%)

采用方法	干净语音	白噪声			粉红噪声			火车噪声		
		20 dB	15 dB	10 dB	20 dB	15 dB	10 dB	20 dB	15 dB	10 dB
MFCC	73.04	72.70	70.64	65.43	71.95	68.44	63.15	72.88	70.92	67.95
LDA	73.11	72.89	71.18	66.30	72.72	70.65	66.09	72.91	71.14	67.36
HLDA	73.31	73.02	71.59	66.84	72.89	70.76	66.27	73.10	71.32	67.02
aPEAC	73.62	73.41	70.13	64.87	73.25	70.06	64.11	73.51	70.25	65.02
MCELDA	74.45	74.23	72.68	68.21	74.17	72.19	67.87	74.29	72.16	68.88

$$\log P(y_{i,g(i)}|\Lambda_{g(i)}) - \log P(y_{i,g(i)}|\Lambda_{g(i)})\Big)\Big] =$$
$$-\left(1+\frac{\varepsilon}{M-1}\right)\sum_{g(i)=1}^{M}\sum_{i=1}^{n_{g(i)}}\log P(y_{i,g(i)}|\Lambda_{g(i)}) +$$
$$\frac{\varepsilon}{M-1}\sum_{g(i)=1}^{M}\sum_{i=1}^{n_{g(i)}}\left(\sum_{m=1}^{M}\log P(y_{i,g(i)}|\Lambda_m)\right) =$$
$$-\left(1+\frac{\varepsilon}{M-1}\right)\sum_{g(i)=1}^{M}\sum_{i=1}^{n_{g(i)}}\log P(y_{i,g(i)}|\Lambda_{g(i)}) +$$
$$\frac{\varepsilon}{M-1}\sum_{m=1}^{M}\sum_{i=1}^{N}\log P(y_i|\Lambda_m) \tag{A1}$$

附录 B 分类损失函数梯度的求解

将式 (4)～(6) 代入目标函数式 (3), 式 (3) 中的第 1 项可以表示为

$$g_{\Lambda_{g(i)}}(y_i) = \sum_{i=1}^{n_{g(i)}}\log P(y_i|\Lambda_{g(i)}) =$$
$$\sum_{i=1}^{n_{g(i)}}\log\left(\frac{1}{n_{g(i)}}\sum_{i=1}^{n_{g(i)}}K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})\right) \tag{B1}$$

式中第 k 维投影向量 $\boldsymbol{w}_k$ 为

$$\nabla_{\boldsymbol{w}_k} g_{\Lambda_{g(i)}}(y_i) =$$
$$\sum_{i=1}^{n_{g(i)}}\gamma_i^{g(i)}\frac{1}{n_{g(i)}}\sum_{i=1}^{n_{g(i)}}\frac{\partial K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})}{\partial \boldsymbol{w}_k} \tag{B2}$$

其中, $\gamma_i^{g(i)} = \left(\frac{1}{n_{g(i)}}\sum_{i=1}^{n_{g(i)}}K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})\right)^{-1}$,

$$\frac{\partial K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})}{\partial \boldsymbol{w}_k} =$$
$$-\frac{1}{2}K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})\frac{\partial(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})^{\mathrm{T}}(\Psi^{g(i)})^{-1}(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})}{\partial \boldsymbol{w}_k} \tag{B3}$$

令 $\theta_{ij}^{g(i)} = (\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})^{\mathrm{T}}(\Psi^{g(i)})^{-1}(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})$, 则 $\theta_{ij}^{g(i)}$ 的梯度可以表示为

$$\frac{\partial\theta_{ij}^{g(i)}}{\partial \boldsymbol{w}_k} = \frac{\partial(\psi_{kk}^{g(i)})^{-1}}{\partial \boldsymbol{w}_k}(y_{ik}-y_{jk}^{g(i)})^2 +$$
$$(\psi_{kk}^{g(i)})^{-1}\frac{\partial(y_{ik}-y_{jk}^{g(i)})^2}{\partial \boldsymbol{w}_k} =$$
$$\tau^{g(i)}\phi^{g(i)}\boldsymbol{w}_k + 2(\boldsymbol{x}_i-\boldsymbol{x}_j^{g(i)})(\boldsymbol{x}_i-\boldsymbol{x}_j^{g(i)})^{\mathrm{T}}\boldsymbol{w}_k \tag{B4}$$

其中, $\bar{\boldsymbol{x}}_i^{g(i)}$ 为 $\boldsymbol{x}_i^{g(i)}$ 的经验平均值, 令 $\phi^{g(i)} = (\boldsymbol{x}_i^{g(i)} - \bar{\boldsymbol{x}}_i^{g(i)})(\boldsymbol{x}_i^{g(i)} - \bar{\boldsymbol{x}}_i^{g(i)})^{\mathrm{T}}$ 为 $\boldsymbol{x}_i^{g(i)}$ 的协方差矩阵, $\tau^{g(i)} = -2\xi^{-1}(\boldsymbol{w}_k^{\mathrm{T}}\phi^{g(i)}\boldsymbol{w}_k)^{-2}$. 则

$$\nabla_{\boldsymbol{w}_k} g_{\Lambda_{g(i)}}(y_i) = \boldsymbol{F}^{g(i)}\boldsymbol{w}_k =$$
$$-\frac{1}{2n_{g(i)}}\sum_{i=1}^{n_{g(i)}}\gamma_i^{g(i)}\sum_{i=1}^{n_{g(i)}}K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^{g(i)})\times$$
$$\Big[\tau^{g(i)}\phi^{g(i)}(y_{ik}-y_{jk}^{g(i)})^2+$$
$$2(\psi_{kk}^{g(i)})^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j^{g(i)})(\boldsymbol{x}_i-\boldsymbol{x}_j^{g(i)})^{\mathrm{T}}\Big]\boldsymbol{w}_k \tag{B5}$$

类似地, 可以得到式 (3) 中后一项的梯度,

$$\nabla_{\boldsymbol{w}_k} G_{\Lambda_m}(y_i) = \boldsymbol{F}^m\boldsymbol{w}_k =$$
$$-\frac{1}{2n_m}\sum_{i=1}^{N}\gamma_i^m\sum_{i=1}^{n_m}K_h(\boldsymbol{y}_i-\boldsymbol{y}_j^m)\times$$
$$\Big[\tau^m\phi^m(y_{ik}-y_{jk}^m)^2+$$
$$2(\psi_{kk}^{m)})^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j^m)(\boldsymbol{x}_i-\boldsymbol{x}_j^m)^{\mathrm{T}}\Big]\boldsymbol{w}_k \tag{B6}$$

References

1 Abbasian H, Nasersharif B, Akbari A, Rahmani M. Optimized linear discriminant analysis for extracting robust speech features. In: Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing. St Julians: IEEE, 2008. 819−824

2 Nasersharif B, Akbari A. SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features. *Pattern Recognition Letters*, 2011, **28**(11): 1320−1326

3 Li Bi-Cheng, Shao Mei-Zhen, Huang Jie. *Pattern Recognition Theory and Application*. Xi′an: Xi′an University Press, 2008. 45−52
(李弼程, 邵美珍, 黄洁. 模式识别原理与应用. 西安: 西安电子科技大学出版社, 2008. 45−52)

4 Kumar N, Andreou A G. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 1998, **26**(4): 283−297

5 Saon G, Padmanabhan M, Gopinath R, Chen S. Maximum likelihood discriminant feature spaces. In: Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul: IEEE, 2000, **2**: 1129−1132

6 Sakai M, Kitaoka N, Nakagawa S. Linear discriminant analysis using a generalized mean of class covariances and its application to speech recognition. *IEICE Transactions on Information and Systems*, 2008, **E91-D**(3): 478−487

7 Loog M, Duin R P W, Haeb-Umbach R. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(7): 762−766

8 Lee H S, Chen B. Empirical error rate minimization based linear discriminant analysis. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, China: IEEE, 2009. 1801−1804

9 Kenny P, Stafylakis T, Ouellet P. PLDA for speaker verification with utterances of arbitrary duration. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013. 7649−7653

10 Kanagasundaram A, Dean D, Vogt R. Weighted LDA techniques for I-vector based speaker verification. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012. 4781−4784

11 Ye J X, Kobayashi T, Murakawa M. Kernel discriminant analysis for environmental sound recognition based on acoustic subspace. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013. 808−812

12 Senior A, Cho Y M, Weston J. Learning improved linear transforms for speech recognition. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012. 1957−1960

13 Tomar V S, Rose R C. Efficient manifold learning for speech recognition using locality sensitive hashing. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013. 6995−6999

14 Heigold G, Ney H, Schluter R, Wiesler S. Discriminative training for automatic speech recognition. *IEEE Signal Processing Magazine*, 2012, **29**(5): 58−69

15 Juang B H, Chou W, Lee C H. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1997, **5**(3): 257−265

16 Biem A, Katagiri S, McDermott E, Juang B H. An application of discriminative feature extraction to filter-bank-based speech recognition. *IEEE Transaction on Speech and Audio Processing*, 2001, **9**(2): 96−110

17 Ruske G, Faltlhauser R, Pfau T. Extended linear discriminative analysis (ELDA) for speech recognition. In: Proceedings of the 1998 ICSLP Sydney. Australia: ISCA, 1998. 1473−1476

18 Li X B, Li J Y, Wang R H. Dimensionality reduction using MCE-optimized LDA transformation. In: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing. Quebec, Canada: IEEE, 2004, **1**: 137−140

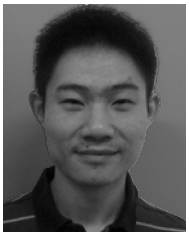
19 Chengalvarayan R, Deng L. Use of generalized dynamic feature parameters for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1997, **5**(3): 232−242

20 Wang Jun, Wang Shi-Tong, Deng Zhao-Hong, Ying Wen-Hao. Fast kernel density estimator based image thresholding algorithm for small target images. *Acta Automatica Sinica*, 2012, **38**(10): 1679−1689
(王骏, 王士同, 邓赵红, 应文豪. 面向小目标图像的快速核密度估计图像阈值分割算法. 自动化学报, 2012, **38**(10): 1679−1689)

21 Scott D W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley and Sons, 1992. 125−190

22 Botev Z I, Grotowski J F, Kroese D P. Kernel density estimation via diffusion. *The Annals of Statistics*, 2010, **38**(5): 2916−2957

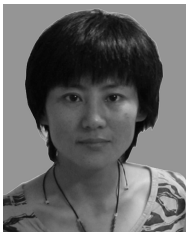
23 Simonoff J S. *Smoothing Methods in Statistics*. New York: Springer-Verlag, 1996. 53−64



陈 斌 解放军信息工程大学信息系统工程学院博士研究生. 主要研究方向为连续语音识别, 区分性训练. 本文通信作者. E-mail: chenbin873335@163.com
(**CHEN Bin** Ph. D. candidate at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers continuous speech recognition and discriminative training. Corresponding author of this paper.)

张连海 解放军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, 语音编码与语音识别.
E-mail: lianhaiz@sina.com
(**ZHANG Lian-Hai** Associate professor at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers speech signal processing, speech coding and speech recognition.)

牛 铜 解放军信息工程大学信息系统工程学院博士研究生. 主要研究方向为语音增强, 语音识别.
E-mail: niutong0072@gmail.com
(**NIU Tong** Ph. D. candidate at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers speech enhancement and speech recognition.)



屈 丹 解放军信息工程大学信息系统工程学院副教授. 2005 年获解放军信息工程大学博士学位. 主要研究方向为语音信号处理与模式识别.
E-mail: qudanqudan@sina.com
(**QU Dan** Associate professor at the Institute of Information System Engineering, PLA Information Engineering University. Her research interest covers speech signal processing and pattern recognition.)

李弼程 解放军信息工程大学信息系统工程学院教授. 主要研究方向为文本分析与理解, 语音处理与识别, 图像/视频处理与识别, 信息融合.
E-mail: lbclm@163.com
(**LI Bi-Cheng** Professor at the Institute of Information System Engineering, PLA Information Engineering University. His research interest covers text analysis and understanding, speech/image/video processing and recognition, and information fusing.)