ISSN 1000-0054 清华大学学报（自然科学版）2012 年 第 52 卷 第 5 期 4/26
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2012, Vol. 52, No. 5 598-601

基于鉴别性 i -vector 局部距离保持映射的说话人识别

栗志意， 何 亮， 张卫强， 刘 加

（清华大学 电子工程系，清华信息科学与技术国家实验室，北京 100084）

摘 要：为了进一步提高 i-vector 说话人识别系统的性能，该文提出了一种鉴别性 i-vector 局部距离保持映射（discriminant i-vector local distance preserving projection，DIVLDPP）的流形学习算法。该算法以 i-vector 间的 Euclid 距离作为度量准则，并以最小化同类点间距离同时最大化异类近邻点间距离的鉴别性准则作为优化目标函数，利用求解广义特征值的方法，得到最终的投影映射矩阵。在美国国家标准技术局 2008 年说话人识别核心数据集上的实验结果表明：该算法可以明显提高目前 i-vector 说话人识别系统的性能。

关键词：流形学习；i-vector；鉴别性；局部距离保持映射；说话人识别

中图分类号：TP 391.4 **文献标志码**：A

文章编号：1000-0054(2012)05-0598-04

Speaker recognition based on discriminant i-vector local distance preserving projection

LI Zhiyi, HE Liang, ZHANG Weiqiang, LIU Jia

(Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: The performance of the popular i-vector based speaker recognition system is improved by a manifold learning algorithm named discriminant i-vector local distance preserving projection (DIVLDPP). This algorithm uses the Euclidean distance to measure the i-vector space. The target function minimizes the distance between the same speaker samples and maximizes the distance between neighbouring samples of different speakers. A linear mapping matrix is obtained by solving a generalized eigenvalue problem. Tests on the speaker recognition evaluation data corpus released by the US National Institute of Standards and Technology in 2008 demonstrate that this i-vector system provides better speaker recognition performance.

Key words: manifold learning; i-vector; discriminative training; local distance preserving projection; speaker recognition

说话人识别是指从说话人的语音信号中辨识或验证说话人身份的一项技术。该技术广泛应用于国家安全、司法鉴定、语音拨号、电话银行等诸多领域[1]。近几年来，研究者们提出了一系列以 Gauss 混合模型-通用背景模型[2]（Gaussian mixture models-universal background models，GMM-UBM）为基础的说话人建模方法，使得说话人识别系统的性能有了突飞猛进的提高。近年来在美国国家标准技术局（American national institute of standards and technology，NIST）组织的评测中占主导地位的 Gauss 混合模型超矢量-支持向量机（Gaussian mixture model super vector-support vector machine，GSV-SVM）、联合因子分析（joint factor analysis，JFA）以及 i-vector（identity vector）等说话人建模技术均以 GMM-UBM 系统为基础，其中 i-vector 系统的性能最好，成为目前国内外研究前沿的主流系统。

i-vector 说话人建模技术的基本思想[3]如式（1）所示，即认为信道和会话的影响均包含在总体变化子空间（total variability space）$\boldsymbol{T}$ 中，通过对包含说话人信息和信道信息的 GMM 均值超矢量 $\boldsymbol{M}$ 在低秩的总体变化子空间 $\boldsymbol{T}$ 上进行投影，得到只包含说话人信息的低维矢量 w（典型维数为 400 维），称之为 i-vector，其中 $\boldsymbol{m}$ 是与说话人和信道均无关的矢量，起到参考坐标的作用，因此通常采用 UBM 均值超矢量来代替。

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}w. \tag{1}$$

在得到初始的 i-vector 矢量后，通过对 i-vector 矢量进行线性区分性分析（linear discriminate analysis，LDA），以求在最小化类内说话人距离和最大化类间说话人距离的鉴别性准则下进一步减低 i-vector 的维数（典型维数为 200 维）。之后进行类内方差归

收稿日期：2011-09-02

基金项目：国家自然科学基金资助项目（90920302，61005019）；国家“八六三”高技术项目（2008AA040201）

作者简介：栗志意（1981—），男（汉），山西，博士研究生。

通信作者：刘加，教授，E-mail：liuj@tsinghua.edu.cn

一化（within class covariance normalization，WCCN）变换，使得变化后的说话人子空间的基尽可能的正交。最后对变换后所得的最终 i-vector 进行余弦距离打分、判决并给出识别结果。

由目前 i-vector 建模过程可以看出，在总体变化子空间 $\boldsymbol{T}$ 上投影得到的初始 i-vector 矢量中仍然包含说话人以外的信道分量，而文[4]中的实验结果表明，基于流形学习的算法在说话人识别系统中对于信道补偿问题有较好的效果。基于此，本文提出了一种 i-vector 局部距离保持映射的流形算法，该算法首先假设 i-vector 空间中的说话人数据存在于低维流形[5-6]上，然后将最小化同类样本点在低维投影后的距离并最大化异类近邻样本点在低维投影后的距离作为优化目标函数，通过转化为求解广义特征值的问题，从而获得鉴别性 i-vector 局部距离保持映射的投影矩阵。在 NIST 2008 年说话人识别（speaker recognition evaluation 2008，SRE2008）核心测试数据库上，给出基于该算法的 i-vector 说话人系统的实验结果。

1 鉴别性 i-vector 局部距离保持映射

本文算法以 i-vector 空间的 Euclid 距离作为度量准则。在此准则下，该算法的目标是希望使同类 i-vector 样本点间的距离尽可能小，而异类近邻 i-vector 样本点间的距离尽可能大。前者是希望算法可以尽可能地保持同类点的嵌入图结构；后者是希望通过采用鉴别性的建模方法，使得映射后的同类样本点和异类样本点间可以更容易得到区分。基于此，可得到该算法所对应的目标函数为

$$\min \frac{\sum_{i,j} \| \boldsymbol{y}_i - \boldsymbol{y}_j \|_2^2 w_{ij}}{\sum_{i,j} \| \boldsymbol{y}_i - \boldsymbol{y}_j \|_2^2 \delta_{ij}}. \tag{2}$$

其中：$\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$ 分别表示第 i 和 j 个样本点的原始 i-vector 矢量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 投影后的低维矢量，w_{ij} 和 δ_{ij} 分别为加权系数矩阵 $\boldsymbol{W}$ 和 $\boldsymbol{\Delta}$ 中的第 i 行第 j 列的元素。在线性映射的前提下，不妨记线性映射矩阵为 $\boldsymbol{P}$，将映射表达式 $\boldsymbol{y}=\boldsymbol{P}\boldsymbol{x}$ 代入式(2)进行化简和变形可得到式(3)所示的 Rayleigh 商函数：

$$\min \frac{\sum_{i,j} \| \boldsymbol{y}_i - \boldsymbol{y}_j \|_2^2 w_{ij}}{\sum_{i,j} \| \boldsymbol{y}_i - \boldsymbol{y}_j \|_2^2 \delta_{ij}} = \min \frac{\mathrm{tr}(\boldsymbol{P}\boldsymbol{X}\boldsymbol{L}_{\mathrm{w}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}})}{\mathrm{tr}(\boldsymbol{P}\boldsymbol{X}\boldsymbol{L}_{\mathrm{d}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}})}. \tag{3}$$

其中：$\mathrm{tr}(\cdot)$ 表示矩阵的迹。$\boldsymbol{X}$ 为由所有 K 个样本点矢量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_K$ 组成的矩阵，可表示为 $\boldsymbol{X}=[\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \cdots, \boldsymbol{x}_K^{\mathrm{T}}]^{\mathrm{T}}$，$\boldsymbol{L}_{\mathrm{w}}$ 和 $\boldsymbol{L}_{\mathrm{d}}$ 均为对称矩阵，表达式如下：

$$\boldsymbol{L}_{\mathrm{w}} = \mathrm{diag}(\boldsymbol{W} \times 1) - \boldsymbol{W}, \tag{4}$$

$$\boldsymbol{L}_{\mathrm{d}} = \mathrm{diag}(\boldsymbol{\Delta} \times 1) - \boldsymbol{\Delta}. \tag{5}$$

对于上述 Rayleigh 商函数的求解可通过求解如式(6)所示的最小的 k 个广义特征值对应的特征向量得到[7]：

$$\boldsymbol{L}_{\mathrm{w}}\boldsymbol{P}\boldsymbol{X} = \boldsymbol{L}_{\mathrm{d}}\boldsymbol{P}\boldsymbol{X}. \tag{6}$$

对于 $\boldsymbol{W}$ 和 $\boldsymbol{\Delta}$ 的构建，本文采用简单连接的方法，即如果样本 i 和 j 相连，则对应加权系数矩阵的元素为 1，否则为 0。对于 $\boldsymbol{W}$，本文将属于同一说话人的数据集构成一个子集，同子集内的样本点之间是相连的，故 $w_{ij}=1$；而不同子集内的样本点间是不相连的，故 $w_{ij}=0$。而对于 $\boldsymbol{\Delta}$，本文将每个样本点与所有异类说话人集合中距离最近邻的 k_{d} 个样本点构成一子集。为保证矩阵 $\boldsymbol{\Delta}$ 的实对称性，若样本 i 是样本 j 的最近邻异类样本，但样本 j 不是样本 i 的最近邻异类样本时，取两者的并集作为各自的最近邻异类样本，以此类推，得到每个样本点的异类近邻样本点子集。在此子集中的样本点之间是相连的，故 $\delta_{ij}=1$；不同子集中的样本点之间是不相连的，故 $\delta_{ij}=0$。

2 基于鉴别性 i-vector 局部距离保持映射的说话人识别系统

将本文提出的算法应用到 i-vector 说话人识别系统，其实现框图如图 1 所示。图 1 中训练语音包括训练说话人模型和各矩阵的所有语音数据。

系统中所用 UBM（universal background model）可根据 EM（expectation maximization）算法[8]进行训练得到。根据训练所得的 UBM 提取 Baum-Welch 统计量：

$$N_m = \sum_t \gamma_{m,t}, \tag{7}$$

$$\boldsymbol{F}_m = \sum_t \gamma_{m,t}(\xi_t - \boldsymbol{\mu}_m). \tag{8}$$

其中：N_m 和 $\boldsymbol{F}_m$ 分别代表零阶和一阶统计量，t 表示离散的时间帧序列的第 t 帧，m 代表 UBM 的第 m 个混合分量，$\gamma_{m,t}$ 是占有率，计算方法如下：

$$\gamma_{m,t} = \frac{N(\xi_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^{M} N(\xi_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}. \tag{9}$$

$N(;\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ 代表均值为 $\boldsymbol{\mu}_m$、方差为 $\boldsymbol{\Sigma}_m$ 的 Gauss 分布，ξ_t 表示第 t 帧的随机矢量，M 表示 UBM 的混合分量个数。

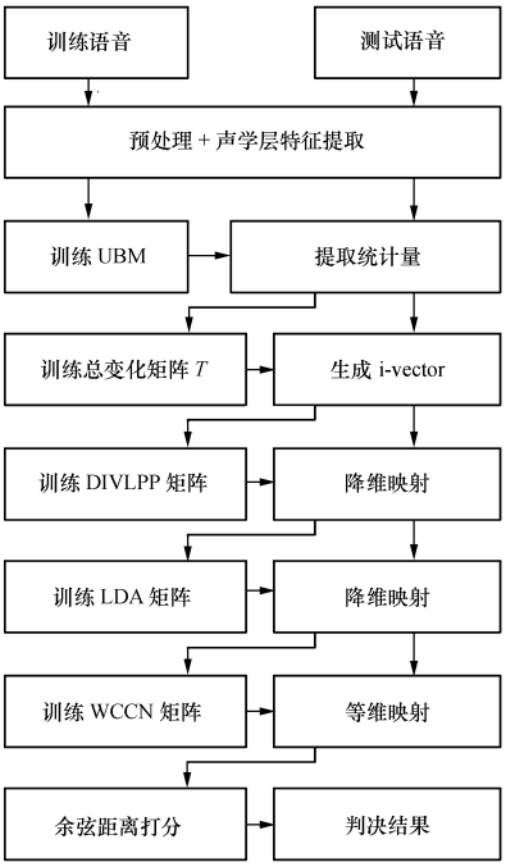

图 1 基于鉴别性 i-vector 局部距离保持映射的说话人识别系统框图

总变化子空间矩阵 T 的训练可采用如下所示的迭代公式训练[9-10]得到:

$$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{T}, \tag{10}$$

$$\mathrm{E}(\boldsymbol{x}) = \boldsymbol{L}^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}. \tag{11}$$

其中 $\boldsymbol{L}$ 为隐变量,$\boldsymbol{F}$ 为 $\boldsymbol{F}_m$ 的矢量排列,$\boldsymbol{N}$ 和 $\boldsymbol{\Sigma}$ 分别为 N_m 和 $\boldsymbol{\Sigma}_m$ 的矩阵对角排列[4]。DIVLDPP 矩阵的训练过程如节 1 中所述。LDA 矩阵的训练可以通过求解式(12)所示目标函数得到[3]:

$$J(\boldsymbol{v}) = \frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{v}}. \tag{12}$$

其中 v 表示待选择的矢量,$\boldsymbol{S}_{\mathrm{B}}$ 和 $\boldsymbol{S}_{\mathrm{W}}$ 分别为类间协方差矩阵和类内协方差矩阵,各自的计算公式如下[3]:

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{r=1}^{R}(\boldsymbol{v}_r - \bar{\boldsymbol{v}})(\boldsymbol{v}_r - \bar{\boldsymbol{v}})^{\mathrm{T}}, \tag{13}$$

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{r=1}^{R}\frac{1}{n_r}\sum_{i=1}^{n_r}(\boldsymbol{v}_i^r - \overline{\boldsymbol{v}_r})(\boldsymbol{v}_i^r - \overline{\boldsymbol{v}_r})^{\mathrm{T}}. \tag{14}$$

训练 WCCN 矩阵的计算公式为

$$\boldsymbol{W} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_r}\sum_{i=1}^{n_r}(\boldsymbol{v}_i^r - \overline{\boldsymbol{v}_r})(\boldsymbol{v}_i^r - \overline{\boldsymbol{v}_r})^{\mathrm{T}}. \tag{15}$$

其中:R 表示训练数据中说话人的个数,$\boldsymbol{v}_r$ 代表第 r 个训练样本,$\boldsymbol{v}_i^r$ 代表第 r 个说话人中的第 i 个样本,n_r 表示第 r 个说话人的样本个数。本文系统中采用与文[2]一致的余弦距离打分作为系统的判决分数。

3 实 验

3.1 实验设置

本文针对 NIST SRE2008 中的核心训练和测试电话对电话测试集上进行实验。训练 UBM 和各矩阵参数的数据均来自于 NIST SRE2004、2005、2006 以及 2008followup 数据集。

前端预处理技术和特征提取技术包括使用 G.723.1 进行有效语音端点检测(voice activity detection, VAD)、倒谱均值减(cepstral mean subtraction, CMS),设置了 3 s 窗长用于特征弯折(feature warping, FW),进行了前 25% 低能量删减以及预加重。在此基础上提取 13 维 MFCC((Mel frequency cepstrum coefficient)基本特征,并与 1、2 阶差分特征构成最终的 39 维 MFCC 特征。

UBM 的混合度为 1 024,单 Gauss 概率密度的方差采用对角阵。i-vector 总变化子空间矩阵的列数是 400。DIVLDPP 矩阵训练过程中的近邻点数设为 128,DIVLDPP 矩阵降维后的新维数为 350。LDA 矩阵降维后的新维数设置为 200。

本文还对余弦距离打分后的原始分数进行了零规整(zero normalization, ZNORM)和零测试规整(zero test normalization, ZTNORM)。

本实验评测指标采用等错误率(equal error rate, EER)和 NIST SRE2008 规定的最小检测代价函数(minimum detection cost function, MinDCF)[11]。

3.2 实验结果与分析

表 1 是原始 i-vector 系统与加入 DIVLDPP 算法改进后的本文 i-vector 系统在 NIST SRE2008 核心数据库上的对比实验结果。其中,ZNORM 表示利用 ZNORM 伙伴集的测试分数来规整原始分数,而 ZTNORM 是指利用 ZNORM 和测试规整(test normalization, TNORM)的伙伴集的测试分数对原始分数进行联合规整。本文实验中 ZNORM 和 TNORM 伙伴集数据均来源于训练矩阵的数据集。

表 1 原始 i-vector 系统和本文 i-vector 系统在 NIST SRE2008 数据库上的对比实验结果

分数种类	EER(%)		MinDCF	
	原始 i-vector 系统	本文 i-vector 系统	原始 i-vector 系统	本文 i-vector 系统
原始分数	4.76	4.49	0.20	0.19
ZNORM	4.52	4.13	0.18	0.17
ZTNORM	3.99	3.60	0.16	0.14

实验结果表明：通过在原始 i-vector 系统中加入鉴别性流形学习算法 DIVLDPP，能够更有效地区分同类样本点和距离较近的异类样本点，从而进一步提高了 i-vector 系统的分类性能。

4 结 论

本文提出了一种鉴别性 i-vector 距离局部保持映射算法 DIVLDPP，用于进一步提高目前国际前沿的 i-vector 说话人识别系统的性能。通过加入局部鉴别性分析，DIVLDPP 能够更有效地区分同类样本点和距离较近的异类样本点，因此能进一步提高系统的分类性能。基于该算法的 i-vector 说话人系统在 NIST2008 核心测试标准库上的实验结果证明了该算法的有效性。

参考文献 (References)

[1] Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors [J]. *Speech Communication*, 2010, **52**(1): 12-40.

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

[3] Dehak N, Kenny P, Ouellet P, et al. Front-end factor analysis for speaker verification [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **4**: 788-798.

[4] 何亮，栗志意，蔡猛，等. 集合分类中的鉴别式局部信息距离保持映射 [J]. 清华大学学报(自然科学版)，2011，**51**(7)：1010-1016.
He L, Li Z Y, Cai M, et al. Discriminant local information distance preserving projection for set classification [J]. *J Tsinghua Univ (Sci and Tech)*, 1994, **34**(2): 1-7. (in Chinese)

[5] He X, Niyogi P. Locality preserving projections [J]. *Advances in Neural Information Processing System*, 2004, **16**: 6-10.

[6] He X, Cai D, Yan S, et al. Neighborhood preserving embedding [C]// Proceedings of the 10th IEEE International Conference on Computer Vision, Beijing: IEEE Press, 2005: 1208-1213.

[7] 张贤达. 矩阵分析与应用 [M]. 北京：清华大学出版社，2004.
ZHANG Xianda. Matrix Analysis and Application [M]. Beijing: Tsinghua University Press, 2004. (in Chinese)

[8] Ghahramani Z, Hinton G. The EM algorithm for mixtures of factor analyzers, Technical Report CRG-TR-96-1 [R]. Toronto: University of Toronto, Department of Computer Science, 1996.

[9] Kenny P, Ouellet P, Dehak N, et al. A study of interspeaker variability in speaker verification [J]. *IEEE Transactions on Audio, Speech and, Language Processing*, 2008, **16**: 980-988.

[10] Kenny P, Boulianne G, Ouellet P, et al. Speaker and session variability in GMM-based speaker verification [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**: 1448-1460.

[11] NIST. NIST speaker recognition evaluation [EB/OL]. [2010-11-08]. http://www.itl.nist.gov/iad/mig/tests/spk/2008/index.html.