

基于核函数的 IVEC-SVM 说话人识别系统研究

栗志意¹ 张卫强¹ 何亮¹ 刘加¹

摘要 在说话人识别研究中, 基于身份认证向量 (Identity vector, IVEC) 的说话人建模方法可以有效地提取说话人信息, 是目前处于国际前沿的建模方法. 本文对身份认证向量后接支持向量机 (Identity vector followed by support vector machine, IVEC-SVM) 的说话人识别系统进行了研究, 对比了该系统在十种不同核函数下的识别性能, 并与文献中身份认证向量后接余弦距离打分 (Identity vector followed by cosine distance scoring, IVEC-CDS) 系统进行了比较. 在美国国家标准技术局 (American National Institute of Standards and Technology, NIST) 组织的 2010 年电话信道—电话信道说话人识别核心评测数据库上的实验结果显示, 基于核函数的 IVEC-SVM 系统性能明显优于 IVEC-CDS 的系统性能. 此外, 实验结果表明基于 Spline 核的 IVEC-SVM 系统可取得最好的识别性能, 与 IVEC-CDS 系统相比, 其等错点 (Equal error rate, EER) 在分数归一化前后分别降低了 10% 和 3%.

关键词 身份认证向量后接余弦距离打分, 身份认证向量后接支持向量机, Spline 核, 说话人识别

引用格式 栗志意, 张卫强, 何亮, 刘加. 基于核函数的 IVEC-SVM 说话人识别系统研究. 自动化学报, 2014, 40(4): 780–784

DOI 10.3724/SP.J.1004.2014.00780

Speaker Recognition with Kernel Based IVEC-SVM

LI Zhi-Yi¹ ZHANG Wei-Qiang¹ HE Liang¹ LIU Jia¹

Abstract In the text-independent speaker recognition research area, identity vector (IVEC) based modeling has been recently proved to be the most efficient method of extracting speaker information. This paper explores and compares the performances of ten different kernel functions in identity vector followed by support vector machines (IVEC-SVM) system and identity vector followed by cosine distance scoring (IVEC-CDS). Experiments corpora the speaker recognition evaluation data, telephone-telephone corpus released by American National Institute of Standard and Technology (NIST) in 2010, demonstrate that the kernel function based IVEC-SVM system performs better than the IVEC-CDS system. Among all the kernel function based IVEC-SVM systems, the spline kernel function performs the best, and it has relative decreases of 10% and 3% in EER compared to the IVEC-CDS system before and after doing score normalization, respectively.

Key words Identity vector followed by cosine distance scoring (IVEC-CDS), identity vector followed by support vector machine (IVEC-SVM), spline kernel, speaker recognition

Citation Li Zhi-Yi, Zhang Wei-Qiang, He Liang, Liu Jia. Speaker recognition with kernel based IVEC-SVM. *Acta Automatica Sinica*, 2014, 40(4): 780–784

收稿日期 2012-09-12 录用日期 2013-01-18
Manuscript received September 12, 2012; accepted January 18, 2013

本文责任编辑 宗成庆
Recommended by Associate Editor ZONG Cheng-Qing
国家自然科学基金 (61005019, 61273268, 90920302, 61370034) 资助
Supported by National Natural Science Foundation of China (61005019, 61273268, 90920302, 61370034)

1. 清华大学电子工程系清华信息与科学技术国家实验室 北京 100084
1. Tsinghua National Laboratory for Information Science and Tech-

说话人识别是指通过从说话人的语音信号中提取声纹特征从而进行辨识或确认说话人身份的一项技术. 作为一种重要的基于生物特征的身份鉴定技术, 目前说话人识别已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. 近几年来, 以高斯混合模型–通用背景模型 (Gaussian mixture model – universal background model, GMM-UBM)^[1] 为基础的说话人建模技术取得了非常大的成功, 使得说话人识别系统的系统性能有了显著提升^[2–3].

目前在由美国国家标准技术局 (American National Institute of Standards and Technology, NIST) 组织的国际说话人评测中, 占主导地位的说话人识别系统如高斯混合模型超向量–支持向量机 (Gaussian mixture model super vector – support vector machine, GSV-SVM)^[4]、联合因子分析 (Joint factor analysis, JFA)^[5–6] 以及身份认证向量–余弦距离打分 (Identity vector – cosine distance scoring, IVEC-CDS)^[7] 等均是高斯混合模型–通用背景模型建模技术为基础. 文献研究表明目前 IVEC-CDS 说话人系统的性能最好, 成为国内外研究的前沿技术^[7].

基于身份认证向量 IVEC 的说话人建模方法是目前说话人识别领域处于国际研究前沿的说话人建模方法. 该建模方法认为高斯混合模型高维均值超向量中同时包含有说话人信息以及信道信息, 通过因子分析的方法, 获得总体变化子空间, 并在该子空间上进行投影, 得到低维的总体变化因子向量, 也即身份认证向量 IVEC (典型维数为 400 维). 并将得到的低维 IVEC 进行线性鉴别性分析 (Linear discriminant analysis, LDA) 和类内协方差归一化 (Within class covariance normalization, WCCN). 前者 LDA 的目的在于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优化准则下进一步降低 IVEC 的维数 (降维后的典型维数为 200 维), 后者类内协方差归一化技术 WCCN 的目的是通过白化协方差矩阵, 尽量使得变换后的子空间的基正交, 从而可有效抑制信道信息的影响. 最终将经过 LDA 和 WCCN 变换过后的 IVEC 向量, 作为新的特征送入后续的分类器进行分类判决.

文献 [7] 中实验结果显示, 目前基于余弦距离打分 (Cosine distance scoring, CDS) 的分类器可以取得比基于余弦距离核函数的支持向量机 (Support vector machine, SVM) 分类器更好的分类性能. 但实际上, SVM 分类器的性能受多方面因素的影响, 诸如反模型 (Negative pool model, NEG) 数据选择是否合适、核函数及其参数选择是否合适等. 文献中的实验并没有深入对 IVEC-SVM 的系统性能进行研究. 一般来说, 鉴别性的分类器往往可比非鉴别性的分类器取得较好的分类效果, 因此虽然余弦距离打分是一种比较简单快捷的对称式打分方法, 但由于支持向量机分类器采用了结构风险最小化的鉴别式准则, 理论上可以取得更好的分类性能.

基于此研究现状, 本文针对 IVEC-SVM 说话人识别系统进行了深入研究, 着重研究了不同核函数选择下 IVEC-SVM 的分类性能. 通过对比选择得出在该系统中更合适的核函数.

1 身份认证向量 IVEC

1.1 IVEC 因子分析

高斯混合模型–通用背景模型建模的基本思想是: 首

nology, Department of Electronic Engineering, Tsinghua University, Beijing 100084

先利用一些训练数据通过期望最大化 (Expectation maximum, EM) 算法训练得到一个通用背景模型 (Universal background model, UBM). 该 UBM 提供了一个统一的参考坐标空间, 并在一定程度上解决了由于说话人训练数据较少导致的小样本问题. 然后通过训练数据在该 UBM 上面进行最大后验概率 (Maximum a posterior, MAP) 自适应, 从而得到了代表说话人的高斯混合概率密度函数, 并将所有 C 个高斯模型的均值向量拼接成一个高维的均值超向量作为说话人向量 (假设单高斯模型的维数为 F , 则最终生成的均值超向量的维数为 FC).

基于身份认证向量的建模方法的基本思想就是在上述高斯混合模型均值超向量基础上利用因子分析的算法从其中估计出总体变化矩阵 (Total variability matrix) T , 通过将高维的高斯混合模型均值超向量在矩阵 T 所表示的总体变化子空间中进行投影, 最终得到投影后的总体变化因子^[7].

总体变化矩阵 T 的估计类似于联合因子分析 JFA 中的本征音估计过程^[8], 如式 (1) 所示.

$$\mathbf{M} = \mathbf{m} + T\mathbf{w} \quad (1)$$

其中, $\mathbf{M}$ 表示高斯混合模型的均值超向量, $\mathbf{m}$ 表示一个与特定说话人和信道都无关的超向量, 通常用通用背景模型均值超向量来替代. 对于说话人 s 的声学特征 $\mathbf{x}_{s,t}$, 其总体变化因子 $\mathbf{w}$ 的估计过程如下:

步骤 1. 计算该声学特征 $\mathbf{x}_{s,t}$ 相对于 UBM 均值超向量 $\mathbf{m}$ 的零阶统计量 $N_{c,s}$ 、一阶统计量 $\mathbf{F}_{c,s}$ 以及二阶统计量 $S_{c,s}$, 如式 (2) 所示.

$$\begin{aligned} N_{c,s} &= \sum_t \gamma_{c,s,t} \\ \mathbf{F}_{c,s} &= \sum_t \gamma_{c,s,t} (\mathbf{x}_{s,t} - \mathbf{m}_c) \\ S_{c,s} &= \text{diag} \left\{ \sum_t \gamma_{c,s,t} (\mathbf{x}_{s,t} - \mathbf{m}_c)(\mathbf{x}_{s,t} - \mathbf{m}_c)^T \right\} \end{aligned} \quad (2)$$

式中, $\mathbf{m}_c$ 代表 UBM 均值超向量 $\mathbf{m}$ 中的第 c 个高斯均值分量. t 表示时间帧索引. $\gamma_{c,s,t}$ 表示 UBM 第 c 个高斯分量的后验概率. $\text{diag}\{\cdot\}$ 表示取对角运算.

步骤 2. 估计因子 $\mathbf{w}$ 的一阶和二阶统计量. 估计过程如式 (3) 所示. 其中超向量 $\mathbf{F}_s$ 是由 $\mathbf{F}_{c,s}$ 向量拼接成的 $FC \times 1$ 维的向量. N_s 是由 $N_{s,c}$ 作为主对角元拼接成的 $FC \times FC$ 维的矩阵.

$$\begin{aligned} L_s &= I + T^T \Sigma^{-1} N_s T \\ E[\mathbf{w}_s] &= L_s^{-1} T^T \Sigma^{-1} \mathbf{F}_s \\ E[\mathbf{w}_s \mathbf{w}_s^T] &= E[\mathbf{w}_s] E[\mathbf{w}_s^T] + L_s^{-1} \end{aligned} \quad (3)$$

式中, L_s 是临时变量, Σ 是 UBM 的协方差矩阵.

步骤 3. 更新 T 矩阵和协方差矩阵 Σ . T 矩阵的更新过程可利用式 (4) 来实现, 也可根据文献 [8] 中的快速算法来实现.

$$\sum_s N_s T E[\mathbf{w}_s \mathbf{w}_s^T] = \sum_s \mathbf{F}_s E[\mathbf{w}_s] \quad (4)$$

对 UBM 协方差矩阵 Σ 的更新过程如式 (5) 所示.

$$\Sigma = N^{-1} \sum_s S_s - N^{-1} \text{diag} \left\{ \sum_s \mathbf{F}_s E[\mathbf{w}_s^T] T^T \right\} \quad (5)$$

式中 S_s 是由 $S_{c,s}$ 进行矩阵对角拼接成的 $FC \times FC$ 维的矩阵, $N = \sum N_s$ 为所有说话人的零阶统计量之和.

利用式 (2) ~ (5) 对上述步骤反复迭代, 进行 6 ~ 8 次后, 可近似认为 T 和 Σ 收敛.

1.2 线性鉴别性分析

线性鉴别性分析 (LDA)^[7] 是模式识别领域广泛采用的一种鉴别性降维技术. 在基于身份认证向量 IVEC 的说话人系统中, 由于前述基于因子分析的方法没有采用鉴别性的准则, 因此通常采用 LDA 对初始通过因子分析得到的身份认证向量进行鉴别性降维, 同时抑制信道信息的影响. 训练 LDA 矩阵的过程主要通过优化如式 (11) 所示的目标函数, 即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, 求该目标函数的最优解.

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (6)$$

式 (6) 中类间协方差矩阵 S_B 和类内协方差矩阵 S_W 的计算过程如式 (7) 和式 (8) 所示.

$$S_B = \sum_{s=1}^S (\mathbf{w}_s - \bar{\mathbf{w}})(\mathbf{w}_s - \bar{\mathbf{w}})^T \quad (7)$$

$$S_W = \sum_{s=1}^S \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T \quad (8)$$

其中, $\bar{\mathbf{w}}_s = (1/n_s) \sum_{i=1}^{n_s} \mathbf{w}_i^s$ 代表第 s 个说话人的 IVEC 均值, S 表示说话人的个数, n_s 表示第 s 个说话人的 IVEC 段数. 最终, 求解式 (6) 中优化目标函数的过程可转换为求解如式 (9) 所示广义特征值的过程.

$$S_B \mathbf{w} = \lambda S_W \mathbf{w} \quad (9)$$

1.3 类内协方差归一化

类内协方差归一化 (WCCN)^[9] 通过白化说话人因子使得变化后的说话人子空间的基尽可能正交. WCCN 矩阵可由式 (10) 估计.

$$W = \frac{1}{S} \sum_{s=1}^S \frac{1}{n_s} \sum_{i=1}^{n_s} (\mathbf{w}_i^s - \bar{\mathbf{w}}_s)(\mathbf{w}_i^s - \bar{\mathbf{w}}_s)^T \quad (10)$$

其中, $\bar{\mathbf{w}}_s = (1/n_s) \sum_{i=1}^{n_s} \mathbf{w}_i^s$ 代表第 s 个说话人的 IVEC 均值, S 表示说话人的个数, n_s 表示第 s 个说话人的 IVEC 段数.

2 余弦距离打分

余弦距离打分 (CDS)^[7] 是一种对称式的分类器, 即说话人模型向量与测试段向量交换后打分结果不变. 在对 IVEC 进行分类时, 该分类器将说话人向量 $\mathbf{w}_{\text{tar}}$ 和测试段向量 $\mathbf{w}_{\text{tst}}$ 的余弦距离分数直接作为判决分数, 并与阈值 θ 进行比较, 给出判决结果, 如式 (11) 所示.

$$\text{score}(\mathbf{w}_{\text{tar}}, \mathbf{w}_{\text{tst}}) = \frac{\langle \mathbf{w}_{\text{tar}}, \mathbf{w}_{\text{tst}} \rangle}{\|\mathbf{w}_{\text{tar}}\| \cdot \|\mathbf{w}_{\text{tst}}\|} \geq \theta \quad (11)$$

本质上讲, 该分类器通过归一化向量的模去除了向量幅度的影响, 将两个向量的余弦角度作为分类的依据, 因此计算简单快捷.

3 支持向量机

支持向量机 (SVM)^[10-11] 是建立在统计学习理论和结构风险最小化理论基础上的两类监督分类器, 它在解决小样本、非线性分类及高维模式识别中表现出许多特有的优势. Campbell 等在文献 [4] 中首次将支持向量机成功应用于说话人识别领域, 提出了目前实际应用中广泛采用的高斯混合模型超向量-支持向量机说话人识别系统. 支持向量机的训练过程是在给定的训练样本集 $X = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_M, \mathbf{y}_M)\}$ 中找出区分正样本点和负样本点的最优线性区分性分类面 H , 其分类函数如式 (12) 所示.

$$f: \mathbf{R}^N \mapsto \mathbf{R}$$

$$\mathbf{x} \mapsto f(\mathbf{x}) = \omega^T \mathbf{x} + b \quad (12)$$

式中, $\mathbf{x}$ 代表输入向量, (ω, b) 为 SVM 训练得到的权重和偏置. 当给定一个核函数 $K(\mathbf{x}, \mathbf{y})$ 时, 对于新测试样本 $\mathbf{x}$, 则分类函数如式 (13) 所示.

$$f(\mathbf{x}) = \sum_i \omega_i^T K(\mathbf{x}, \mathbf{x}_i) + b \quad (13)$$

SVM 的关键在于核函数的选择. 低维空间向量集通常难于线性划分, 而解决方法之一是将其映射到可线性划分的高维空间. 但这个方法带来的直接问题之一就是计算复杂度的增加, 而核函数的引入正好巧妙地解决了这个问题. 只要选用适当的核函数, 就可以得到相应的高维再生希尔伯特空间中的分类函数, 而无需进行显式的非线性映射. 在 SVM 中, 采用不同的核函数往往可以得到不同的 SVM 算法.

4 基于核函数的 IVEC-SVM 说话人识别系统

核函数^[10] 的基本思想是, 在低维空间线性不可分的非紧致模式往往可通过非线性映射到高维特征空间实现紧致, 从而可实现线性可分, 但是如果显式地研究非线性映射, 则难度很大而且存在高维特征空间分类时的“维数灾难”问题. 核函数提供了解决上述问题的有效方法. 假设 $\mathbf{x}, \mathbf{z}$ 是输入空间 $X \in \mathbf{R}^n$ 的向量, 非线性函数 ϕ 为实现输入空间到再生希尔伯特空间 $F \in \mathbf{R}^m$ 的映射, 其中 $n \ll m$. 则相应的核函数定义为

$$K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle \quad (14)$$

式中, $\langle \cdot \rangle$ 表示内积运算, $K(\mathbf{x}, \mathbf{z})$ 为核函数.

核函数应用广泛, 满足 Mercer 定理的函数都可以作为核函数. 在支持向量机中引入核函数有很多优点: 1) 利用核函数可避免 SVM 训练时的“维数灾难”, 大大减小计算复杂度; 2) 无需知道非线性映射函数的具体形式和参数, 采用不同的核函数即代表了不同的映射函数和映射空间.

本文选取了模式识别领域较为成熟的十种核函数进行验证, 通过实验研究了不同核函数对 IVEC 向量的非线性映射分类性能. 基于核函数的 IVEC-SVM 说话人识别系统的系统框图如图 1 所示.

1) Linear 核

$$K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle + c \quad (15)$$

2) Cosine 核

$$K(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \quad (16)$$

3) Polynomial 核

$$K(\mathbf{x}, \mathbf{y}) = (\alpha \mathbf{x}^T \mathbf{y} + c)^d \quad (17)$$

4) Gaussian 核

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \quad (18)$$

5) Multiquadric 核

$$K(\mathbf{x}, \mathbf{y}) = \sqrt{\|\mathbf{x} - \mathbf{y}\|^2 + c^2} \quad (19)$$

6) Wave 核

$$K(\mathbf{x}, \mathbf{y}) = \frac{\theta}{\|\mathbf{x} - \mathbf{y}\|} \sin \frac{\|\mathbf{x} - \mathbf{y}\|}{\theta} \quad (20)$$

7) Log 核

$$K(\mathbf{x}, \mathbf{y}) = -\log \|\mathbf{x} - \mathbf{y}\|^d + 1 \quad (21)$$

8) Spline 核

$$K(\mathbf{x}, \mathbf{y}) = 1 + \mathbf{x}\mathbf{y} + \mathbf{x}\mathbf{y}\min(\mathbf{x}, \mathbf{y}) - \frac{\mathbf{x} + \mathbf{y}}{2} \min(\mathbf{x}, \mathbf{y})^2 + \frac{\min(\mathbf{x}, \mathbf{y})^3}{3} \quad (22)$$

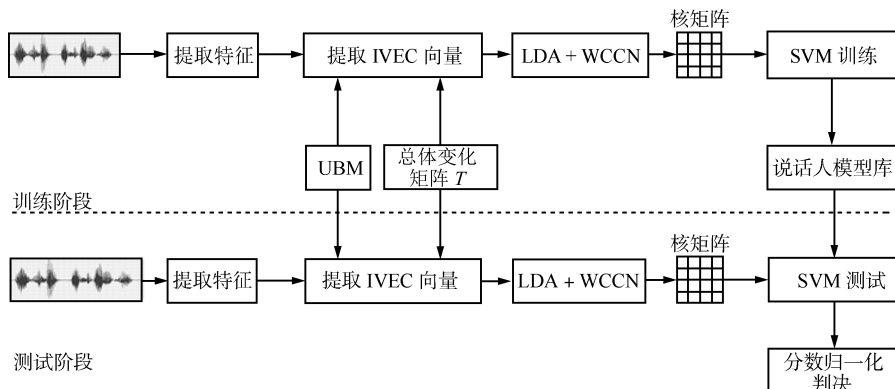


图 1 基于核函数的 IVEC-SVM 说话人系统框图

Fig. 1 Diagram of speaker recognition system with kernel based IVEC-SVM

9) Cauchy 核

$$K(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma}} \quad (23)$$

10) Tstudent 核

$$K(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \|\mathbf{x} - \mathbf{y}\|^d} \quad (24)$$

5 实验配置与实验结果

本节针对上一节中所列十种核函数在 IVEC-SVM 说话人系统中的性能进行实验, 得出性能最优的 IVEC-SVM 的核函数形式, 并与基于 SVM-CDS 的说话人识别系统的性能进行了对比, 最后分析并得出结论. 本实验中采用 SHOGUN 机器学习工具箱^[12] 以及 LIBSVM 工具箱^[13] 来实现核函数以及 SVM 系统搭建.

5.1 实验配置

本实验中的测试数据库采用 NIST SRE 2010 电话信道-电话信道 (Tel-tel) 的核心测试集^[14]. 训练语音和测试语音均来自 5 分钟左右的电话对话信道中抽取的单声道说话人声音, 实际的语音长度约 3 分钟左右. 训练 UBM 和各矩阵的数据均来自于 NIST SRE 2004、2005、2006 以及 2008 数据集中含有多段语音说话人的数据. SRE 2004 的 1side 和 8sides 的数据用于 SVM 的反模型训练. 各训练数据集的配置如表 1 所示.

表 1 用于估计矩阵 UBM, T, LDA, WCCN, NEG, ZT-Norm 的数据集

Table 1 Corpora used to estimate the UBM, total variability matrix T, LDA, WCCN, NEG, and ZT-Norm

矩阵类型	Switchboard	NIST 2004	NIST 2005	NIST 2008
UBM	×	×	×	×
T	×	×	×	×
LDA	×	×	×	×
WCCN	×	×	×	×
NEG		×		
ZT-Norm	×		×	

实验中采用 Mel 频率倒谱特征 (Mel-frequency cepstral coefficient, MFCC) 作为声学层特征. 在预处理阶段采用 G.723.1 进行有效语音端点检测 (Voice activity detection, VAD) 以及倒谱均值减 (Cepstral mean subtraction, CMS) 技术来有效去除或抑制信道的卷积噪声, 并设置了 3s 窗长用于特征弯折 (Feature warping), 进行了 25% 低能量删减以及预加重 (因子为 0.95). 在上述预处理基础上, 首先提取 13 维的基本特征, 并与一阶、二阶差分特征一起构成最终的 39 维 MFCC 声学层特征. 实验中使用的 UBM 混合数设置为 1024, 高斯概率密度的方差采用对角阵. IVEC 中总体变化子空间矩阵 T 的子空间维数即列数设置为 400, 训练时迭代次数取 6 次. LDA 矩阵降维后的维数设置为 200. 本实验采用 NIST 等错误率 (Equal error rate, EER) 和 NISTSRE 2008 规定的最小检测代价函数 (Minimum detection cost function, MinDCF) 作为评测指标. 实验中还对测试分数进

行零测试规整 (Zero test normalization, ZT-Norm) 后的性能进行了评估.

5.2 实验结果

从表 2 中的实验结果可以看出: 首先, 基于 Cosine 核的 IVEC-SVM 系统在分数归一化后的 EER 和 MinDCF 性能要差于基于余弦距离打分的 IVEC-CDS 系统的性能, 这一结论与文献中结果一致. 而从不做分数归一化的性能可以看出, 前者的性能要优于后者, 因此对于 IVEC-CDS 的性能很大程度上依赖于分数归一化操作, 而 IVEC-SVM 系统性能依赖性较小. 其次, 通过对不同核函数下的 IVEC-SVM 性能进行对比, 可以看出除基于 Polynominal 核外, 其余核函数下的 IVEC-SVM 系统的原始分性能均优于 IVEC-CDS 系统, 而其中基于 Spline 核函数的性能最好, 在分数归一化前后不仅优于基于 Cosine 核的 IVEC-SVM, 而且优于 IVEC-CDS 系统性能, 可见 Spline 核函数与其他核函数相比对 IVEC 向量有较好的非线性映射能力.

表 2 基于核函数的 IVEC-SVM 实验结果
Table 2 Performance of kernel function based IVEC-SVM

序号	系统	核函数	分数归一化	EER	MinDCF
1	IVEC-CDS	—	—	5.24	0.2279
2	IVEC-CDS	—	ZT-Norm	4.18	0.2005
3	IVEC-SVM	Linear 核	—	5.06	0.2642
4	IVEC-SVM	Linear 核	ZT-Norm	4.15	0.2142
5	IVEC-SVM	Cosine 核	—	4.86	0.2167
6	IVEC-SVM	Cosine 核	ZT-Norm	4.77	0.2051
7	IVEC-SVM	Polynominal 核	—	5.36	0.2965
8	IVEC-SVM	Polynominal 核	ZT-Norm	4.50	0.2435
9	IVEC-SVM	Gaussian 核	—	4.95	0.2522
10	IVEC-SVM	Gaussian 核	ZT-Norm	4.17	0.2091
11	IVEC-SVM	Multiquadric 核	—	4.77	0.2314
12	IVEC-SVM	Multiquadric 核	ZT-Norm	4.32	0.2117
13	IVEC-SVM	Wave 核	—	5.05	0.2549
14	IVEC-SVM	Wave 核	ZT-Norm	4.17	0.2104
15	IVEC-SVM	Log 核	—	4.89	0.2395
16	IVEC-SVM	Log 核	ZT-Norm	4.30	0.2077
17	IVEC-SVM	Spline 核	—	4.67	0.2486
18	IVEC-SVM	Spline 核	ZT-Norm	4.02	0.1901
19	IVEC-SVM	Cauchy 核	—	5.03	0.2148
20	IVEC-SVM	Cauchy 核	ZT-Norm	4.18	0.2105
21	IVEC-SVM	Tstudent 核	—	4.85	0.2318
22	IVEC-SVM	Tstudent 核	ZT-Norm	4.49	0.2151

6 结论

本文探索性地研究了基于 IVEC-SVM 的说话人识别系统, 并对比了十种不同核函数下该系统的识别性能. 在 NIST 组织的 2010 年电话信道-电话信道说话人识别核心评测数据库上的实验结果表明, 基于核函数的 IVEC-SVM 系统性能明显优于目前文献中基于 IVEC-CDS 说话人系统的性能. 此外, 实验结果表明基于 Spline 核的 IVEC-SVM 系统可取得最好的识别性能, 与 IVEC-CDS 系统相比, 其等错点在分

数归一化前后分别降低了 10% 和 3%, 可有效提高目前说话人识别的性能。

References

- 1 Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, **10**(1–3): 19–41
- 2 Kinnunen T, Li H Z. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 2010, **52**(1): 12–40
- 3 Li Zhi-Yi, He Liang, Zhang Wei-Qiang, Liu Jia. Speaker recognition based on discriminant i-vector local distance preserving projection. *Journal of Tsinghua University (Science and Technology)*, 2012, **52**(5): 598–601
(栗志意, 何亮, 张卫强, 刘加. 基于鉴别性 i-vector 局部距离保持映射的说话人识别. 清华大学学报 (自然科学版), 2012, **52**(5): 598–601)
- 4 Campbell W M, Campbell J P, Reynolds D A, Singer E, Torres-Carrasquillo P A. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 2006, **20**(2–3): 210–229
- 5 Kenny P, Boulianne G, Ouellet P, Dumouchel P. Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(4): 1448–1460
- 6 Kenny P, Boulianne G, Ouellet P, Dumouchel P. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2007, **15**(4): 1435–1447
- 7 Dehak N, Kenny P J, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, **19**(4): 788–798
- 8 Kenny P, Boulianne G, Dumouchel P. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 2005, **13**(3): 345–354
- 9 Hatch A O, Kajarekar S S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition. In: *Proceedings of the International Conference on Spoken Language*. Pittsburgh, PA, 2006. 1471–1474
- 10 Bishop C M. *Pattern Recognition and Machine Learning*. Berlin: Springer, 2008
- 11 Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, **20**(3): 273–297
- 12 Sonnenburg S, Rätsch G, Henschel S, Widmer C, Behr J, Zien A, de Bona F, Binder A, Gehl C, Franc V. The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, 2010, **11**: 1799–1802
- 13 Chang C C, Lin C J. LIBSVM: a library for support vector machines, 2001 [Online], available: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, September 12, 2012
- 14 NIST. The NIST Year 2010 Speaker Recognition Evaluation Plan [Online], available: <http://www.nist.gov/speech/tests/sre/2010/index.html>, September 12, 2012

栗志意 清华大学电子工程系博士研究生. 主要研究方向为说话人识别与语种识别. 本文通信作者.

E-mail: lizhiyi06@mails.tsinghua.edu.cn

(LI Zhi-Yi Ph.D. candidate in the Department of Electronic Engineering, Tsinghua University. His research interest covers speaker recognition and language recognition. Corresponding author of this paper.)

张卫强 清华大学电子工程系助理研究员. 主要研究方向为说话人识别与语种识别. E-mail: wqzhang@tsinghua.edu.cn

(ZHANG Wei-Qiang Assistant professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speaker recognition and language recognition.)

何亮 清华大学电子工程系助理研究员. 主要研究方向为说话人识别与语种识别. E-mail: heliang@tsinghua.edu.cn

(HE Liang Assistant professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speaker recognition and language recognition.)

刘加 清华大学电子工程系教授. 主要研究方向为语音识别和信号处理. E-mail: liuj@tsinghua.edu.cn

(LIU Jia Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition and signal processing.)