相关向量机及在说话人识别应用中的研究

杨成福[1,2]，章 毅[1]

(1. 电子科技大学计算智能实验室 成都 610054; 2. 四川文理学院理工系 四川 达州 635000)

【摘要】对基于相关向量机和高斯混合模型的说话人识别算法的模型和特征空间进行了一系列的研究。与一些基于语音帧的说话人识别算法相比，该算法将GMM算法作为底层的语音特征提取，从而实现对语音整体上的处理，对常用的两种语音特征美尔频率倒频系数和瞬时频率的表现进行了对比研究；同时，该算法充分利用了相关向量机的所提供的高泛化性、核函数功能和结果的高稀疏性。基于Chains和AHUMADA两个专门用于说话人识别的语音库的仿真表明，该算法在减少相对误差和减少计算量方面有较大的优势。

关 键 词 高斯分布； GMM超向量核； 瞬时频率； 相关向量机； 语音分析

中图分类号 TP391.42 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.02.034

Study to Speaker Recognition Using RVM

YANG Cheng-fu[1,2] and ZHANG Yi[1]

(1. Computational Intelligence Laboratory, University of Electronic Science and Technology of China Chengdu 610054;
2. Officer of Education Administrator, Sichuan University of Arts and Science Dazhou Sichuan 635000)

Abstract A series of studies on speaker recognition algorithm based on relevance vector machine (RVM) and gaussian mixture model (GMM) was proposed in this paper. The sparseness and probability prediction of RVM make the algorithm suitable for speaker recognition in applications. The robust speech features based on GMM are investigated. In contrast to the most current systems based on frame-level discrimination, the approach has two outstanding merits. The first is the system provides direct discrimination between whole sequences by combining GMM as underlying generative models in feature-space. The paper focused on two main feature space: mel-frequency cepstrum coefficient (MFCC) and instantaneous frequencies (IF). The second combines the high generalization, kernel tricks, and sparser performance of RVM to generate more robust classification results and to reduce the computational complexity. The simulations using the Chains database and the AHUMADA database show that the proposed algorithm outperforms the other systems on reducing the relative error rates and reducing the computational complexity in high dimensionality space and big scale data.

Key words gaussian distribution; GMM super-vector kernel; instantaneous frequencies; relevance vector machine; speech analysis

说话人识别是指从说话人的语音中提取能代表个人的特征参数，然后用分类算法对说话人的身份进行辨认(identification)或确认(verification)。根据训练和测试时所使用的语音内容的方式可以分为文本相关(Text-Dependent)说话人识别、文本无关(Text-Independent)说话人识别和文本提示(Text-Prompted)说话人识别[1]。与其他生物特征识别技术(如人脸识别、指纹识别)相比，说话人识别是最自然、最经济的方式。因语音具有很高的不稳定性和混叠性，从而使说话人识别的准确性受外界环境的影响。找到更鲁棒的说话人语音特征和更具泛化性的分类算法是目前说话人识别领域正在探索的方向。本文从这两个方面出发，在GMM超级向量空间[2]中探索基于RVM的说话人识别算法，同时比较研究算法中现在普遍使用的美尔频率倒谱系数和瞬时频率两种主要语音特征，及其组合对提高说话人识别的准确率和效率方面的作用。

现在常用的说话人识别算法归纳起来有两种：基于模板(如高斯混合模型和隐马尔可夫模型)和基于模式分类(如神经网络和支持向量机)[1-4]。基于模

收稿日期：2008－09－28； 修回日期：2009－06－16
基金项目：国家863计划(2007AA01Z321)；四川省教育厅自然科学重点项目(08ZA037)
作者简介：杨成福(1972－)，男，在职博士生，主要从事语音处理及支持向量机等方面的研究.

板的主要反映了类内的相似性，而基于模式分类的主要反映了类间的差异性。现在有很多算法将两种思想结合起来使用[2-6]。

相关向量机(relevance vector machine，RVM)[7]具有和支持向量机(support vector machine，SVM)[8]相同的决策函数形式，是基于概率统计的一种学习机。RVM比SVM有更稀疏的表示，并具有概率预测和不用人为凭经验确定参数的优点。RVM在函数拟合及分类应用中在准确度上与SVM基本一致，在测试阶段因其稀疏性比SVM更快，并具有自动考虑噪声影响的功能，从而具有更好的泛化性能。

GMM[3]是一种经典的、高准确性的说话人识别模型。其基本原理是用大量的说话人语音样本进行训练，以提取说话人的语音模板；在测试阶段，将测试语音帧的特征向量与前面形成的语音模板进行比对，获得与模板相似性的得分，然后累加语句段各帧的得分，以形成说话人在某个模板上的得分，说话人的身份最大可能性是得分最多所对应的模板。本文将RVM与GMM结合起来，利用RVM的GMM超向量核函数技术将语音模型投影到高维空间，然后利用RVM的高泛化性和高稀疏性实现分类操作。

说话人识别算法所用的语音特征有很多，现在主要使用美尔频率倒谱系数和瞬时频率[1,5,9]。文献[2]将美尔频率倒谱系数特征向量投影到GMM的模型空间，以提高系统的稳健性。本文在此基础上对两种常用的语音特征及组合在说话人识别中的应用进行对比研究。

1 相关向量机

相关向量机(RVM)是一种与SVM有相似决策函数的稀疏概率模型[7]。

设训练样本为$\{x_i, t_i\}_{i=1\sim N}$，$t_i\in\{0,1\}$为目标输出。在贝叶斯框架下找到输入与目标输出间的关系为：

$$\begin{cases} y(x, w)=\sum_{k=1}^{N} w_k K(x, x_k)+w_0 \\ P(t\mid w)=\prod_{i=1}^{N}\sigma[y(x_i;w)]^{t_i}\{1-\sigma[y(x_i;w)]\}^{1-t_i} \\ P(w\mid\alpha)=\prod_{i=1}^{N}\frac{\alpha_i}{\sqrt{2\pi}}\exp\left(-\frac{\alpha_i w_i^2}{2}\right) \end{cases} \tag{1}$$

式中 $\sigma(x)=\dfrac{1}{1+\mathrm{e}^{-x}}$ 为sigmoid函数。

在贝叶斯框架下通过所给的训练样本找到最优的$P(w|t,\alpha)$，以使整个系统有最优的分界面。最常见的拉普拉斯方法为：

$$\begin{aligned} w_{\mathrm{MP}} &= \arg\max_{w} P(w\mid t,\alpha)= \\ &\arg\max_{w}\ln\{P(t\mid w)P(w\mid\alpha)\} \end{aligned} \tag{2}$$

使用牛顿方法得到w_{MP}更新表达式为：

$$\begin{cases} \boldsymbol{g}=\nabla_w\ln\{P(t\mid w)P(w\mid\alpha)\}= \\ \quad \boldsymbol{\phi}^{\mathrm{T}}\cdot(t-\boldsymbol{y})-\boldsymbol{A}\cdot w \\ \boldsymbol{H}=\nabla_w\nabla_w\ln\{P(t\mid w)P(w\mid\alpha)\}= \\ \quad -(\boldsymbol{\phi}^{\mathrm{T}}\cdot\boldsymbol{B}\cdot\boldsymbol{\phi}) \\ \Delta w=-\boldsymbol{H}^{-1}\cdot g \\ w_{\mathrm{MP}}^{\mathrm{new}}=w_{\mathrm{MP}}+\Delta w \end{cases} \tag{3}$$

式中 $\boldsymbol{y}=[y_1,y_2,\cdots,y_N]^{\mathrm{T}}$；$\boldsymbol{A}=\mathrm{diag}[\alpha_1,\alpha_2,\cdots,\alpha_N]$；$\boldsymbol{B}=\mathrm{diag}[y_1(1-y_1),y_2(1-y_2),\cdots,y_N(1-y_N)]$；$\boldsymbol{\Phi}$为由各特征向量代入核函数所得到的设计矩阵，其具体表达式为：

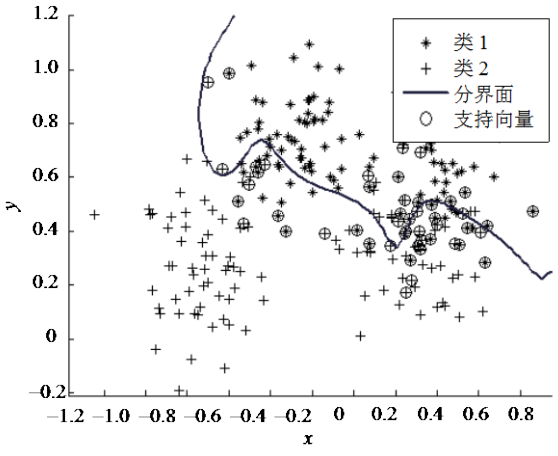
$$\boldsymbol{\Phi}=\begin{pmatrix} 1 & K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_N) \\ 1 & K(x_2,x_1) & K(x_2,x_2) & \cdots & K(x_2,x_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & K(x_N,x_1) & K(x_N,x_2) & \cdots & K(x_N,x_N) \end{pmatrix} \tag{4}$$

拉普拉斯方法将$P(w|t,\alpha)$近似为高斯分布，其均值 $\mu=w_{\mathrm{MP}}$，变异矩阵 $\boldsymbol{\Sigma}=(-\boldsymbol{H})^{-1}$，有：

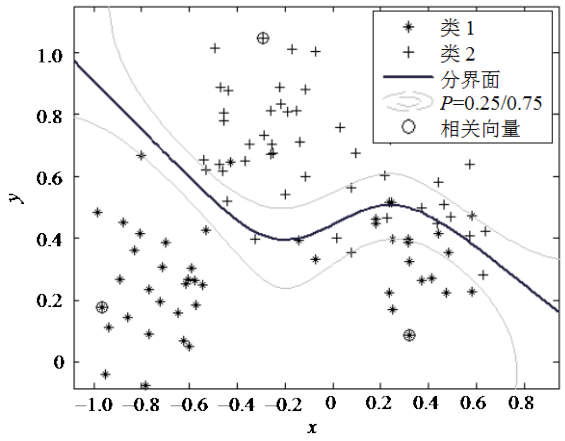
$$\begin{cases} \alpha_i^{\mathrm{new}}=\dfrac{\gamma_i}{\mu_i^2} \\ \gamma_i=1-\alpha_i\boldsymbol{\Sigma}_{i,i} \end{cases} \tag{5}$$

对于多分类问题，采用one-against-one和采用one-against-all算法均可，影响不大。

图1为SVM和RVM在相同数据集上的分类结果。其中SVM用的参数C=12，所得到的支持向量的个数为81；RVM所得到的相关向量的个数为7。从图1可以看出RVM和SVM在准确度上相差无几，但相关向量的数目比支持向量的数目少得较多，使得RVM比SVM在测试阶段所用的时间少得多，并且不用人为确定参数。



a. SVM



b. RVM

图1　SVM和RVM在相同数据集上的表现

2　基于RVM的说话人识别算法

2.1　语音数据库描述

本文使用AHUMADA[10]和Chains Corpus两个免费的语音数据库[11]，预处理采用文献[12]中所述的算法实现。

AHUMADA包括来自不同国家的30位男性和30位女性在相同情景下的西班牙语音，每位发音者的语音放在一个目录中。每位发音者的目录中包括两段语音，一段用于训练系统，一段用于测试系统。

Chains Corpus语音库包括36位发音者的语音，其中每位发音者的语音在两种情景下录制：一种使用专业的设备和录音室进行录制；另一种则用一般设备在安静的办公室内录制，每一种情景下发音者提供6种不同发音方式的语音。本文主要使用其中的NORM方式(正常语速)、FAST(较快语速)和WHSP(特快语速)对系统进行训练和测试以验证系统的鲁棒性。

2.2　语音特征空间

现在的说话人识别系统常用的语音特征主要是美尔频率倒谱系数(MFCC)，本文主要对MFCC和瞬时频率(IF)[9]及其组合对系统的影响进行比较性研究。瞬时频率的提取算法如图2所示。

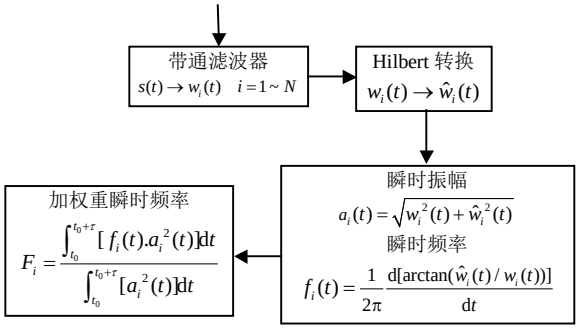


图2　瞬时频率的提取步骤

为了充分利用RVM的核函数技术，本文将GMM的超向量核和RVM结合起来进行比较。GMM超向量的形成如图3所示[2]。

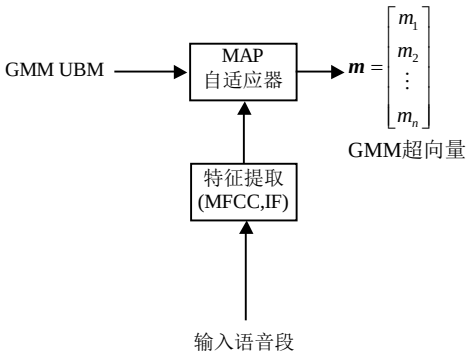


图3　GMM超向量的形成

2.3　算法描述

本文所使用的算法主要是将RVM与MFCC及瞬时频率等语音特征结合起来进行相对于其他算法的对比研究，其算法步骤如图4所示。

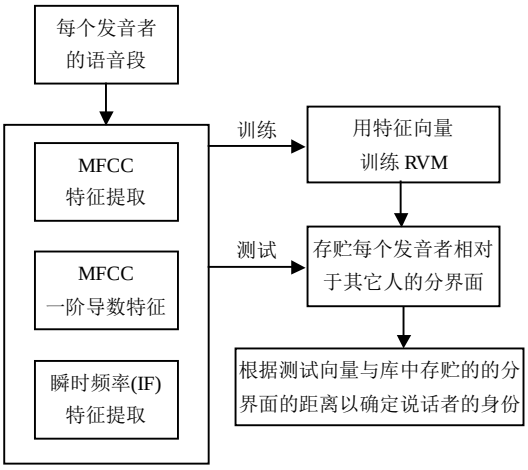


图4　基于RVM的说话人识别框图

3　实验与仿真

为了验证系统的可行性和实用性，本文设计了一系列的仿真实验。

第一个实验的主要目的是比较SVM和RVM在训练和测试时所需要的时间与数据点的数目之间的关系。从AHUMADA数据库中依次取出2位、5位、10位、20位、30位、40位、50位、60位说话人的声音语段，用本文所述方法提取12维的MFCC特征并形成GMM超向量作为训练向量，每位说话人的训练向量数目是10，分别用SVM和RVM进行训练；对每一种情况选择20个有标识的特征向量作为测试数据，以比较RVM和SVM在测试阶段的表现，其结果如图5所示。从仿真结果可以看出，训练阶段随着数据规模的增加，RVM相比SVM所需要的时间多，但

在测试阶段却少。

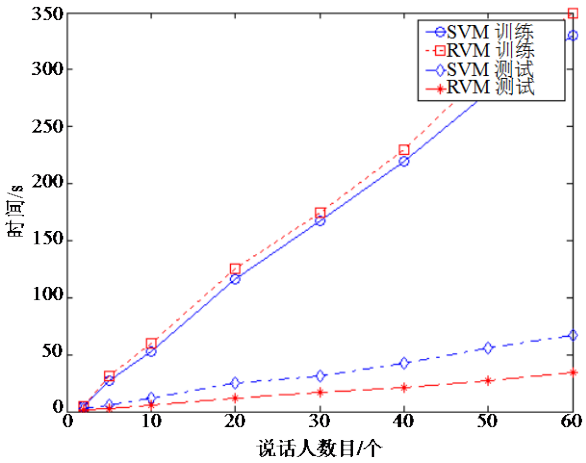


图5 RVM和SVM训练及测试时间比较

第二个实验的主要目的是比较RVM、SVM和GMM在准确度方面的表现。从Chains Corpus中依次取出2位、5位、10位、20位、30位、40位、50位、60位说话人的声音语段，其中训练阶段用NORM、FAST和WHSP三种方式的语段。用本文所述方法对每位说话者提取10个瞬时频率特征向量训练SVM、RVM和GMM，测试阶段对每一位说话者用FAST和WHSP两种方式的语段各提取10个有标识的瞬时频率特征向量进行测试，其结果如图6所示。从实验结果可以看出，三种算法在准确度方面相差无几，但RVM在测试的时间上优于其他两者，在线式测试有很大的优势。

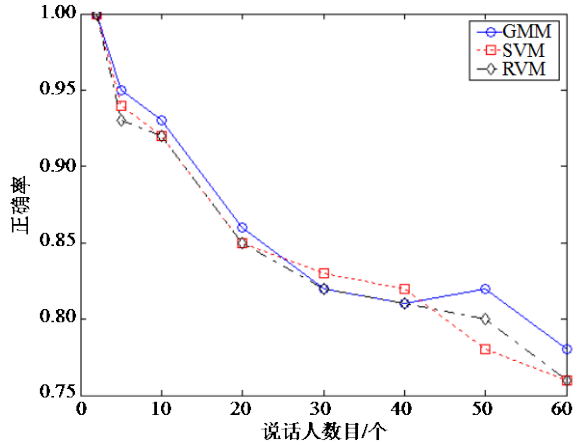


图6 RVM、SVM和GMM准确性比较

第三个实验的主要目的是测试RVM在不同的语音特征空间中的表现。从AHUMADA数据库中取出5位说话者的语段，从Chains Corpus中的NORM、FAST和WHSP三种发音方式分别取出5位说话者的语段。提取MFCC、瞬时频率、瞬时频率投影到GMM超向量空间作为特征向量，分别送到RVM中进行训练后用相应的5位说话人的语段进行测试，其结果如表1所示。从表中数据可以看出，瞬时频率相比MFCC有较好的结果，瞬时频率投影到GMM超向量空间的WHSP发音方式下有较好的结果。

表1 RVM在不同特征空间的表现

	AHUMADA	NORM	FAST	WHSP
MFCC	0.92	0.93	0.89	0.82
IF	0.93	0.94	0.92	0.90
IF+GMM SuperVector	0.94	0.95	0.93	0.90

4 结论与展望

本文比较研究了基于RVM和GMM超向量的说话人识别算法。RVM基于贝叶斯框架，从而使用概率归属的方式给出预测的结果，更符合人的思维习惯；此外，RVM不用人为设定模型的参数，对核函数要求的放宽以及支持向量的更稀疏性是相对于SVM的最大特点。从实验结果可以看出，在准确度方面，RVM和其他算法不相上下，但在测试阶段其占用计算机的时间上有较大的优势。

现阶段有很多的学者在探索说话人识别系统中所使用的语音特征的问题，本文使用了常用的两种语音特征MFCC和瞬时频率及其组合，并将这两种特征向量投影成GMM超向量后，比较研究其在本系统中的表现。从实验结果可以看出，这种超向量在系统的提高系统鲁棒性方面有较大的作用。

以后的研究工作主要集中在如何将多分类框架直接嵌入RVM框架；如何选择RVM核以适应不同的应用；如何改进RVM的训练过程以缩短其占用的计算机时间。

参 考 文 献

[1] CAMPBELL J P. Speaker recognition: a tutorial[J]. Proc IEEE, 1997, 85(9): 1437-1462.

[2] CAMPBELL W M, STURIM D E, REYNOLDS D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE Signal Processing Letters, 2006, 13(5): 308-311.

[3] REYNOLDS D A, QUATIERI T F, DUNN R. Speaker verification using adapted gaussian mixture models[J]. Dig Signal Process, 2000, 10(1-3): 19-41.

[4] WAN V. Speaker verification using support vector machines [D]. Sheffield, U.K: Univ Sheffield, 2003.

[5] KINNUEN T. Spectral features for automatic text-independent speaker recognition[D]. Joensuu, Finland: Univ Joensuu, 2003.

[6] BIMBOT F, MAGRIN C I, MATHAN L. Second-order statistical measures for text-independent speaker identification[J]. Speech Commun, 1995, 17(1-2): 177-192.
[7] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research, 2001, 1(3): 211-244.
[8] VAPNIK V. Statistical learning theory[M]. New York: John Wiley, 1998.
[9] GRIMALDI M, CUMMINS F. Speaker identification using instantaneous frequencies[J]. IEEE Transaction on Audio, Speech, and Language Processing, 2008, 16(6): 1097-1111.
[10] ORTEGA G J, GONZALEZ R J, MARRERO A V, et al. A large speech corpus in Spanish for speaker characterization and identification[J]. Speech Communication, 2000, 31: 255-264.
[11] CUMMINS F, GRIMALDI M, LEONARD T, et al. The Chains corpus: characterizing individual speaker[C]//Proc SPECOM'06. Petersburg, Russian: [s.n.], 2006.
[12] ZHANG De-xiang, GAO Qing-wei, CHEN Jun-ning. Single channel speech enhancement by de-noising using stationary wavelet transform[J]. Journal of Electronic Science and Technology of China, 2006, 4(1): 39-42.

编 辑 税 红

(上接第301页)

[9] ZENG L, DUMAS M. QoS-aware middleware for Web[J]. IEEE Transactions on Software Engineering, 2004, 30(5): 311-327.
[10] CANFORA G, PENTA M D, ESPOSITO R, et al. A lightweight approach for QoS-aware service composition[C]//Proceedings of the 2nd International Conference on Service Oriented Computing. Newyork, USA: [s.n.], 2004.
[11] CANFORA G, PENTA M D, ESPOSITO R, et al. An approach for QoS-aware service composition based on genetic algorithms[C]//Genetic and Evolutionary Computation Conference. Washington, D. C. USA: [s.n.], 2005.
[12] 蔡美玲, 高春鸣. 基于树型编码的遗传算法在Web服务选择中的应用[J]. 计算机工程与应用, 2007, 43(31): 214-218.
CAI Mei-lin, GAO Chun-ming. Application of tree-coding genetic algorithm on Web services selection[J]. Computer Engineering and Applications, 2007, 43(31): 214-218.
[13] 张成文, 苏 森, 陈俊亮. 基于遗传算法的QoS感知的Web服务选择[J]. 计算机学报, 2006, 29(7): 1029-1037.
ZHANG Cheng-wen, SU sen, CHEN Jun-liang. Genetic algorithm on Web services selection supporting QoS[J]. Chinese Journal of Computers, 2006, 29(7): 1029-1037.

编 辑 黄 莘

·我校科研成果介绍·

宽带集成光波导电场传感器

传统的电磁波接收装置是各种各样的天线，新型集成宽带集成光波导电场传感系统有激光器、保偏光纤、电场传感器、普通光纤、光探测器组成。电场传感器是整个系统的核心器件，它是一种利用光波导技术实现的非功能型传感器，应用晶体的电光效应，通过晶体基片中光波导来检测空间中电磁波信号的电场分量，使电场信号调制到光载波上，光强度随着被检测电场相应地变化，经过光探测器后的输出信号电流即反映了被检测的是电场信号。

应用范围：电磁兼容方面的电场测量；天线近场电场测量、远场方向图测量；机载、弹载和星载的电磁信号接收。