

支持向量机应用于语音情感识别的研究

张石清^{1,2}, 赵知劲¹, 戴育良², 杨广映²

(1. 杭州电子科技大学, 通信工程学院, 浙江杭州 310018; 2. 台州学院物理与电子工程学院, 浙江临海 317000)

摘要: 为了有效识别包含在语音信号中情感信息的类型, 提出一种将支持向量机应用于语音情感识别的新方法。利用支持向量机把提取的韵律情感特征数据映射到高维空间, 从而构建最优分类超平面实现对汉语普通话中生气、高兴、悲伤、惊奇 4 种主要情感类型的识别。计算机仿真实验结果表明, 与已有的多种语音情感识别方法相比, 支持向量机对情感识别取得的识别效果优于其他方法。

关键词: 支持向量机; 情感识别; 韵律情感特征

中图分类号: TN912.34

文献标识码: A

文章编号: 1000-3630(2008)-01-0087-04

A study of support vector machine for speech emotion recognition

ZHANG Shi-qing^{1,2}, ZHAO Zhi-jin¹, DAI Yu-liang², YANG Guang-ying²

(1. School of Telecommunication Engineering, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China;

2. School of Physics and Electronic Engineering, Taizhou College, Linhai 317000, Zhejiang, China)

Abstract: A new method of speech emotion recognition in speech signal via Support Vector Machine (SVM) is proposed. SVM maps the extracted prosody emotional feature data into a high dimensional space and constructs the optimum classifying hyper-plane to recognize the four main speech emotions in Chinese mandarin such as anger, happiness, sadness and surprise. Computer simulation results show that SVM can obtain better recognition rate for emotion by comparing with other existing methods for speech emotion recognition.

Key words: support vector machine; emotion recognition; prosody emotional features

1 引 言

语音是人类相互交流的最重要工具之一, 也是传递情感信息的一种重要媒介。传统的语音信号处理技术仅仅着眼于语音词汇传达的准确性, 而大多忽视了其中的情感信息。因此, 如何让计算机从语音中自动地识别出说话者的情感状态, 是当前国内外倍受关注的一个新的热点研究课题。该研究在人工智能^[1]、机器人技术^[2]、新型人机交互技术^[3]等领域具有重要的应用价值。

目前, 研究者识别语音情感类型普遍采用的方法有主成分分析^[4](Principal Component Analysis,

PCA)、高斯混合模型^[5](Gaussian Mixtures Model, GMM)、K 最近邻法^[6](K-Nearest Neighborhood, KNN)等。这些方法应用于语音情感识别取得的情感识别率还不太令人满意, 有待进一步提高。

本文提出利用支持向量机 (Support Vector Machine, SVM) 分类技术, 实现对汉语普通话生气、高兴、悲伤和惊奇四种主要情感类型的语音情感识别。计算机仿真实验结果表明, SVM 应用于语音情感识别, 与上述方法相比较能取得更高的识别率。

2 支持向量机分类器

支持向量机是建立在统计学习理论基础上的, 在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势。

对于两类分类问题, 假定二类训练样本集:

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

收稿日期: 2007-01-27; 修回日期: 2007-04-20

基金项目: 浙江省教育厅高校青年教师资助(2005)

作者简介: 张石清(1980-), 男, 湖南人, 硕士, 研究方向: 语音信号处理和识别。

通讯作者: 张石清, E-mail: tzcqsq@163.com

考虑到有些样本不能被超平面正确分类,引入松弛变量 $\xi_i \geq 0$ 。此时超平面约束变成

$$y_i(w \cdot x_i + b) - 1 + \xi_i \leq 0, i=1, \dots, n \quad (1)$$

对于线性可分和线性不可分情况,最优分类超平面是当且仅当是 (w, b) 下述优化问题的解

$$\begin{aligned} \min & \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right) \\ \text{st.} & \Rightarrow y_i(w \cdot x_i + b) - 1 + \xi_i \leq 0 \end{aligned} \quad (2)$$

式中 $C > 0$ 为权值系数,它控制对错误样本惩罚的程度,又称惩罚系数。

相应的决策函数为

$$D_i(x) = \text{sgn} \left[\sum_{i=1}^n a_i y_i(x_i, x) + b \right] \quad (3)$$

为了优化问题(2)可转化为一个二次规划问题,如下式所示:

$$\begin{aligned} \max W(a) &= \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j (x_i \cdot x_j) \\ \text{st.} & \Rightarrow \sum_{i=1}^n y_i a_i = 0, i=1, \dots, n; 0 \leq a_i \leq C \end{aligned} \quad (4)$$

对于非线性问题,SVM通过非线性变换将原始的数据映射到一个高维特征空间,并在此空间实现最优分类。此时的目标函数为

$$W(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j K(x_i \cdot x_j) \quad (5)$$

相应的决策函数为

$$D_i(x) = \text{sgn} \left[\sum_{i=1}^n a_i y_i K(x_i, x) + b \right] \quad (6)$$

式中 $K(x_i, x_j)$ 称为核函数。目前较常用的核函数有线性核函数、多项式核函数、径向基核函数。

对于多类分类问题,常用的方法有“一对多”和“一对一”方法。“一对多”方法是对于 k 类问题构造 k 个支持向量机子分类器,其中的每一个子分类器都把其中的一类同余下的各类分开。而“一对一”方法,是分别选取 2 个不同类别构成一个 SVM 子分类器,这样共有 $k(k-1)/2$ 个 SVM 子分类器。当对一个未知样本进行分类时,每个分类器都对其类别进行判断,为相应的类别“投上一票”,最后得票最多的类即作为该未知样本所属的类。“一对一”方法训练速度较“一对多”快。

3 情感语音数据库

本文选用了如表 1 所示的 5 个语句作为情感分析语音料,并请 10 名测试者分别用生气、高兴、悲伤

和惊奇 4 种情感对每个句子发音 5 遍,在安静的环境里录制,采集到采样率为 11 025Hz, 16bit 的 wav 格式的实验用情感语音 1000 句,建立实验数据库。

表 1 实验所用情感语句

Table 1 Emotional sentences in test

语句	1	2	3	4	5
内容	学校就要开学了	外边下雨了	你真伟大呀	这件事是他干的	这下子全完了

4 情感特征参数提取

提取何种有效的语音情感特征是语音情感识别研究最重要的问题之一,情感特征的优劣直接影响到情感最终识别结果的好坏。目前, Petrushin^[7]、Ang^[8] 等人已经证明语音信号中情感信息,主要是通过语音信号中的韵律特征参数,如基音频率、振幅、发音持续时间、语速等来表现的。因此,本文提取韵律情感特征参数用于语音情感识别。

4.1 基音频率

基音频率 (pitch), 简称基频, 是声带的振动频率, 是反映情感信息的重要特征之一。本文采用自相关算法提取情感语句的基频轨迹曲线, 如图 1 所示, 列出了同一说话人在不同情感下说同一语句(学校就要开学了)时的基频轨迹曲线。由图 1 可知: 愤怒、高兴和惊奇时的基频的动态变化范围都比较大, 而在悲伤时的变化范围比较小; 愤怒、高兴和惊奇时基频的平均值都明显高于悲伤时基频的平均值; 惊奇时的基频曲线在末端结束时通常上翘。可见, 基频特征能很好地把悲伤、惊奇和其它情感区分开来。基于基频特征, 本文提取出整个基音频率轨迹曲线的最大值 P_{\max} 、最小值 P_{\min} 、极差 $P_d = P_{\max} - P_{\min}$ 、上四分位数 $P_{0.75}$ 、中位数 $P_{0.5}$ 、下四分位数 $P_{0.25}$ 、内四分极值 $P_i = P_{0.75} - P_{0.25}$ 、基频平均值 m_p 、基频标准差 σ_p 、平均绝对斜度 M_s (mean absolute slope)、以及基频抖动值 P_j (pitch jitter) 等共 11 个参数。

4.2 振幅

语音信号的振幅特征与各种情感信息具有较强的相关性。在实际生活中我们可以发现, 当人们愤怒或惊奇时, 发音的音量往往变大; 而当沮丧或悲伤时, 讲话声音往往很低。因此, 振幅特征是情感分析研究中不可或缺的重要特征。如图 2 所示, 列出了同一说话人在不同情感下说同一语句(外边下雨了)时的振幅曲线。从图 2 显然得知, 悲伤与其它三种情感相比, 一个明显的差别是悲伤的振幅曲线在语

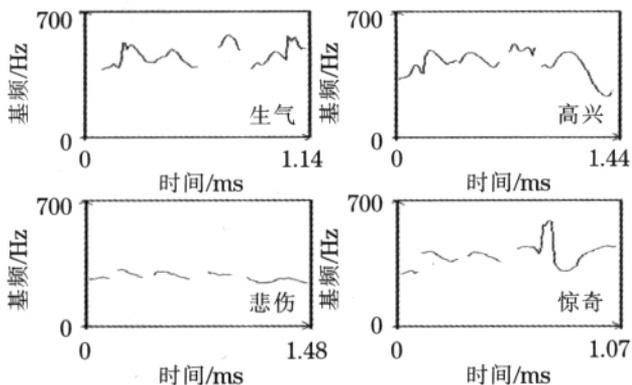


图 1 不同情感下说同一语句的基频轨迹曲线比较
Fig.1 Pitch contour curves of speaking the same sentence under different emotions

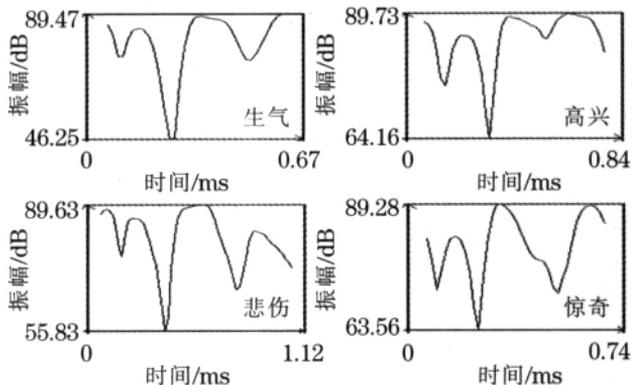


图 2 不同情感下说同一语句的振幅曲线比较
Fig.2 Intensity curves of speaking the same sentence under different emotions

句发音快结束时的一小段时间间隔内的幅值是最小的。因此,使用振幅特征很容易区分悲伤和另外三种情感。

本文对每条语句基于振幅提取其平均值 m_a 、标准差 σ_a 、最大值 A_{max} 、最小值 A_{min} 、极差 $A_d = A_{max} - A_{min}$ 、上四分位数 $A_{0.75}$ 、中位数 $A_{0.5}$ 、下四分位数 $A_{0.25}$ 和内四分极值 A_i , 共 9 个参数作为情感识别用的与振幅有关特征。

4.3 发音持续时间

发音持续时间特征参数着眼于不同情感语音的发音时间构造的差别。设每条情感语句的整句发音总帧数为 F_s 、有声语音总帧数为 F_v 、无声语音总帧数为 F_u , 帧长取 10ms。作为情感识别用的与发音持续时间有关的 5 个特征参数为:

- (1) 发音持续总时间 $T_s = \text{语句发音总帧数 } F_s \times \text{帧长}$;
- (2) 有声发音持续时间 $T_v = \text{有声语音总帧数 } F_v \times \text{帧长}$;
- (3) 无声发音持续时间 $T_u = \text{无声语音总帧数 } F_u \times \text{帧长}$;
- (4) 有声发音持续时间与发音持续总时间的比

值 $T_{VR} = T_v / T_s$;

(5) 无声发音持续时间与发音持续总时间的比值 $T_{UR} = T_u / T_s$

对于同一语句(外边下雨了),同一说话人在不同情感下发音,其时间特征参数比较如表 2 所示。

表 2 不同情感下说同一语句的时间特征比较
Tab2 Time features of speaking the same sentence under different emotions

	TS/ms	TV/ms	TU/ms	TVR	TUR
生气	650	590	60	0.91	0.09
高兴	810	770	40	0.95	0.05
悲伤	1090	1020	70	0.94	0.064
惊奇	710	660	50	0.93	0.07

从表 2 可知,悲伤情感的发音持续总时间最长,有声发音持续时间最长,无声发音持续时间也最长;而生气情感的发音持续总时间最短,有声发音持续时间也最短,无声发音持续时间与发音持续总时间的比值最大。可见,使用发音持续时间参数能较好地地区分悲伤和生气两种情感。

4.4 语速

语速即一个人说话的快慢,它反映了一个人在说话时的心情的急切程度,这显然会随着情感状态的不同而不同。本文所提到的语速(S_r)定义为:

$$S_r = \frac{\text{语句所含文字个数}}{\text{发音持续总时间}}$$

一般在愤怒和惊奇的情况下,人说话的语速比较快,而在高兴和悲伤的情况下,人说话的语速比较慢。本文取平均语速(M_{S_r})作为情感识别用的语速特征参数。

5 情感识别实验及结果分析

本文的情感识别问题,属于多类分类,采用“一对一”方法,实验用的核函数为径向基函数,即:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \quad (7)$$

本文选用第 3 节介绍的情感语音数据库中 1000 句,250 句作为训练样本,750 句用于情感识别样本。首先计算提取出上述第 4 节介绍的 26 个韵律情感特征参数,并将特征参数数据归一化;然后采用 10 次交互验证方法训练样本数据,搜索到最优的惩罚系数 C 和径向基函数参数 γ ;接着用最优的 C 和 γ 训练建立 SVM 模型;最后,进行 SVM 情感识别,识别结果如表 3 所示。

由表 3 可知,生气和悲伤的识别率较为令人

表3 基于SVM的情感识别结果

Tab3 Emotion recognition results based on SVM

	生气	高兴	悲伤	惊奇
生气	0.858	0.031	0.022	0.089
高兴	0.044	0.741	0.035	0.18
悲伤	0.047	0.032	0.848	0.073
惊奇	0.058	0.176	0.037	0.729

满意,分别达到了 85.8%和 84.8%,总体平均识别率为 79.4%。但是高兴和惊奇的识别率略低,主要原因是这两种情感在发音时,许多生理特征相似,较易混淆。

为检验本文提出方法的有效性,分别用 PCA、GMM、KNN 三种不同的方法对同样的情感特征数据进行识别实验。其中,PCA 利用识别语句的特征参数向量到主成分的投影来计算该识别语句相对于某种情感的综合概率,选取概率最大的情感作为识别结果。GMM 利用对于某个识别语句的情感主元素特征矢量来求取它相对于每个情感类别的概率值,概率最大的即为识别情感,试验中采用一个高斯分量即 $M=1$ 时效果最好。KNN 作为传统的无参数分类器,试验中采用一个最近邻训练模式即 $K=1$ 时效果最好。三种方法的实验结果与 SVM 相比,如表 4 所示。

从表 4 可以看出,与其它三种方法相比,SVM 对生气、高兴、悲伤、惊奇四种情感类型的识别率都是最高的。其中平均识别率比 PCA 提高了 11% 比

表4 不同识别方法的识别结果比较(%)

Tab4 Emotion recognition results based on different classification ways

	生气	高兴	悲伤	惊奇	平均识别率
SVM	85.8	74.1	84.8	72.9	79.4
PCA	74.5	68.3	70.2	60.6	68.4
GMM	79.4	71.3	76.5	64.7	72.9
KNN	80.2	72.8	78.9	69.3	75.3

GMM 提高了 6.5%,比 KNN 提高了 4.1%,这说明 SVM 的识别性能是最好的。

6 结 论

本文提出了利用 SVM 的语音情感识别方法,实验结果表明,采用径向基函数的 SVM 对汉语普通话生气、高兴、悲伤、惊奇四种情感类型的识别率都要高于 PCA、GMM 和 KNN,可见 SVM 应用于语音情感识别取得了很好的识别率。

参 考 文 献

- [1] Juan Martínez-Miranda, Arantza Aldea. Emotions in human and artificial intelligence[J]. Computers in Human Behavior, 2005, 2(21): 323-341.
- [2] Cynthia Breazeal. Emotion and sociable humanoid robots[J]. International Journal of Human-Computer Studies, 2003, 1-2(59): 119-155.
- [3] Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal Processing magazine, 2001, 18(01): 32-80.
- [4] 王治平, 赵力, 邹采荣. 利用模糊熵进行参数有效性分析的语音情感识别[J]. 电路与系统学报, 2003, 3(08): 109-112.
WANG Zhiping, ZHAO Li, ZOU Cairong. Emotion recognition of speech using fuzzy entropy effectiveness analysis[J]. Journal of Circuits and Systems, 2003, 3(08): 109-112.
- [5] 赵力, 钱向民, 邹采荣, 等. 语音信号中的情感识别研究[J]. 软件学报, 2001, 12(07): 1050-1055.
ZHAO Li, QIAN Xiangming, ZOU Cairong, et al. A Study on emotional recognition in speech signal[J]. Journal of Software, 2001, 12(07): 1050-1055.
- [6] Chul Min Lee, et al. Classifying emotions in human-machine spoken dialogs[A]. Multimedia and Expro Proceeding. 2002 IEEE International Conference[C]. 2002, 737-740.
- [7] V Petrushin. Emotion recognition in speech signal: experimental study, development, and application[A]. Proceedings of the ICSLP[C]. Beijing, 2000, 222-225.
- [8] J Ang, R Dhillon, A Krupski, et al. Prosody-based automatic detection of annoyance and frustration in human-computer dialog[A]. Proceedings of the ICSLP[C]. Denver, 2002, 2037-2039.