基于PCA和SVM的普通话语音情感识别

蒋海华[1] 胡 斌[2]
(北京工业大学计算机学院 北京 100124)[1] (北京工业大学电子信息与控制工程学院 北京 100124)[2]

摘 要 在语音情感识别中,情感特征的选取与抽取是重要环节。目前,还没有非常有效的语音情感特征被提出。因此,在包含6种情感的普通话情感语料库中,根据普通话不同于西方语种的特点,选取了一些有效的情感特征,包含Mel频率倒谱系数、基频、短时能量、短时平均过零率和第一共振峰等,进行提取并计算得到不同的统计量;接着采用主成分分析(PCA)进行抽取;最后利用基于支持向量机(SVM)的语音情感识别系统进行分类。实验结果表明,与其他一些重要的研究结果相比,该方法得到了较高的平均情感识别率,且情感特征的选取、抽取及建模是合理、有效的。

关键词 语音情感识别,主成分分析,支持向量机
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.055

Speech Emotion Recognition in Mandarin Based on PCA and SVM

JIANG Hai-hua[1] HU Bin[2]
(College of Computer Science, Beijing University of Technology, Beijing 100124, China)[1]
(College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China)[2]

Abstract Feature selection and extraction play a vital role in speech emotion recognition. At present, no effective speech emotion features are proposed. For these reasons, according to the characteristic of Mandarin which is different from western languages, some effective emotional features and related statistics, including Mel-frequency cepstral coefficients, pitch frequency, short-time energy, short-time average zerocrossing rate, the first formant and so on, were analyzed on a Mandarin emotional corpus which contains 6 kinds of emotions. Then we chose principal component analysis (PCA) for extraction and presented a speech emotion recognition method based on support vector machine (SVM) for classification. The experimental results show that the proposed method achieves high emotion recognition accuracy compared with several significant methods, and the emotion extraction and modeling are reasonable and effective.

Keywords Speech emotion recognition, PCA, SVM

1 引言

语音是人们交流的主要方式,语音信号不仅传递语义信息,而且通过高低强弱、抑扬顿挫来表达说话人丰富的情感信息。当说话人愤怒时,音调升高、语速加快;悲伤时,语调低沉、语速缓慢。听者可以通过语音信号感受说话人的情感变化。如何使计算机从语音信号中自动识别出说话人的情感状态及其变化,是实现自然人机交互界面的关键前提,具有很大的研究价值和应用价值。例如:可以用于对电话服务中心用户紧急程度的分拣,从而提高服务质量;用于对汽车驾驶者的精神状态进行监控,从而在驾驶员疲劳时进行提醒,避免交通事故的发生;用于对抑郁症患者的情感变化进行跟踪,从而作为疾病诊断和治疗的依据等。计算机的语音情感识别能力是计算机情感智能的重要组成部分。目前,国内外学者在这方面进行了大量的研究。

对语音情感相关特征的有效提取是该领域内十分重要的研究课题。由于语音情感本身自有的复杂性,还没有占据重要地位的有效情感特征被提出[1]。寻找合适的情感识别算法,也是本领域研究者一直以来努力的目标。

Origlia等人[2]使用基频和能量相关特征的最大值、最小值、均值、标准差组成了一个31维的韵律特征集,在一个包含有意大利语、法语、英语、德语在内的多语种情感语料库上取得了接近60%的识别率。Seppänen等人[3]使用基频、能量、时长相关的43维全局韵律特征进行芬兰语的情感识别,在说话人不相关的情形下取得了60%的识别率。Li等人[4]提取了频率微扰和振幅微扰并将其作为声音质量参数对SUSAS数据库中的语料数据进行了说话人不相关的情感识别,HMM(Hidden Markov Model)被用作识别器,同仅使用Mel频率倒谱系数(Mel-Frequency Cepstral Coefficients,MFCC)的基线性能65.5%的识别率相比,MFCC和频率微扰的特征组合可以得到68.1%的识别率,MFCC和振幅微扰的特征组合可以得到68.5%的识别率,最佳性能69.1%的识别率由MFCC、频率微扰和振幅微扰的共同组合获得。Nwe等人[5]将HMM用作分类器对包括生气、厌恶、害怕、高兴、悲伤和

到稿日期:2014-11-05 返修日期:2015-03-16 本文受国家重点基础研究发展计划项目(2014CB744600),北京市科学技术研究院创新团队项目(IG201203N)资助。

蒋海华(1984-),男,硕士,讲师,主要研究方向为语音情感计算,E-mail:jianghaihua@bjut.edu.cn;胡 斌(1965-),男,博士,教授,主要研究方向为普适计算、认知科学。

惊讶在内的6类情感进行说话者相关的识别，结果表明LFPC取得了77.1%的识别率。Breazeal等人[6]利用高斯混合模型(Gaussian Mixture Model，GMM)分类器对面向婴儿的KISMET数据库进行情感分类，并使用一种基于峰态模型的选择策略对Gaussian成分的数量进行优化，由基频和能量的相关特征训练得到GMM模型，最优性能可达到78.77%。Nichoson等人[7]基于MLP(multi-layer perceptron)建立了一个OCON网络模型，对8种情感进行识别，所使用的数据库是自行录制的，有100位说话者参与录制。实验表明，该模型的最优识别率为52.87%。

上述研究大多数都是针对西方语种的语音情感。而针对普通话语音情感的研究则较少，基本都是在国内进行，主要研究机构有东南大学、微软亚洲研究院、清华大学、浙江大学及台湾的一些大学和研究所等。由于普通话与西方语种发音的方式不同，普通话是声调语言，而英语等西方语种是重音语言，因此能准确表达情感的特征必然不尽相同。

针对普通话的语音情感，本文在普通话情感语料库上提取了一些与其它研究不同的语音情感特征参数组合，接着对原始特征集合利用主成分分析(Principal Component Analysis，PCA)来降低维数，然后采用多类分离支持向量机(Support Vector Machine，SVM)进行建模，最后进行大量识别测试实验。

2 语音情感识别系统结构

本文的语音情感识别系统结构如图1所示，每个步骤之间通过样本数据进行关联。情感特征的提取、PCA降维和SVM建模是其中的关键环节。

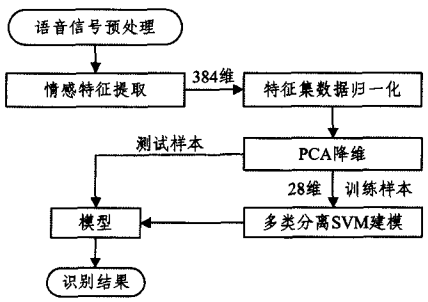

图1 语音情感识别系统结构

3 语音情感特征的提取

情感语音库能为情感建模和情感语音声学特征分析提供统计数据支持，同时也为情感语音识别提供必要的训练和测试语料数据。情感语音库的质量高低，决定了由它训练得到的情感识别系统的性能好坏。目前，领域内存在的情感语音库类型多样，并没有统一的建立标准[1]，多数研究会建立自己的语料库。为了使实验结果有更好的横向比较依据，本文采用国内具有代表性的中科院自动化所模式识别实验室提供的普通话情感语料库CASIA[8]。该语料库由中文普通话情感短句组成，是在实验室受控环境下，对两位女性和两位男性分别录制所得。语料的采样率为16KHz，采样精度为16bit，音质清晰，无明显噪声干扰。语料涵盖了愤怒(angry)、高兴(happy)、悲伤(sad)、害怕(fear)、惊讶(surprise)和中性(neutral)这6种类别。每个说话人每个类别均包括50个已标记情感类别的短句，整个语料库由1200个短句组成。

实验中，首先对所有的1200句语料进行预处理、分帧和加窗，再基于各语音分析帧提取声学特征。用于语音情感识别的声学特征主要有韵律学特征、基于谱的相关特征和声音质量特征这3种类型。这些特征以帧为单位进行提取，以全局特征统计值的形式参与情感的识别。全局统计的单位是听觉上独立的语句或者单词，本文采用的统计指标有极值、极值范围、均值和方差等。

基于谱的相关特征被认为是声道形状变化和发声运动之间相关性的体现[9]，已经在包括语音识别、说话者识别等在内的语音信号处理领域有着成功的运用。Nwe等人[5]通过对情感语音的相关谱特征进行研究发现，语音中的情感内容对频谱能量在各个频谱区间的分布有着明显的影响。例如，表达高兴情感的语音在高频段表现出高能量，而表达悲伤的语音在同样的频段却表现出差别明显的低能量。本文提取其中的一种倒谱特征——MFCC。

韵律是指语音中在语义符号之上的音高、音长、快慢和轻重等方面的变化，是对语音流表达方式的一种结构性安排。它的存在与否并不影响我们对字、词、句的听辨，却决定着一句话听起来是否自然顺耳，它的情感区分能力已得到语音情感识别领域研究者们的广泛认可。普通话的声调特性决定了它的韵律特征在情感识别中的重要作用。本文提取3种韵律特征：基频、短时能量和短时平均过零率。

声音质量是语音的一种主观评价指标，用于衡量语音是否纯净、清晰、容易辨识等[10]。对声音质量产生影响的声学表现有喘息、颤音、哽咽等，并且常常出现在说话者情绪激动、难以抑制的情形下。声音质量的变化被多数研究者认定为同语音情感的表达有着密切的关系。本文提取的共振峰频率就是其中之一。

3.1 MFCC特征

MFCC是语音情感识别中常用的一种特征参数。它根据人耳的听觉特性，将频谱最终转化为倒谱域上的系数[11]。MFCC具有较好的识别性能和抗噪能力，它的值大体上对应于实际频率的对数分布关系，具体关系可用式(1)表示：

$$Mel(f)=2595\lg(1+f/700) \tag{1}$$

3.2 基音频率特征

基音是指发浊音时声带振动所引起的周期性。声带振动频率称为基频。因为普通话是一种有调语言，基音的变化模式称为声调，它携带着非常重要的信息，基音频率反映了语音情感的重要特征。本文采用短时自相关函数来检测基音：

$$R_n(k)=\sum_{m=0}^{N-k-1}S_n(m)S_n(m+k) \tag{2}$$

3.3 短时能量

设语音时域信号为$x(l)$，窗函数为$w(m)$，加窗后的第n帧语音信号为$x_n(m)$，则该帧的短时能量为：

$$E_n=\sum_{m=0}^{N-1}[w(m)x(n+m)]^2 \tag{3}$$

3.4 短时过零率

短时过零率指帧语音信号时域波形穿过时间轴的次数。加窗后第n帧语音信号$x_n(m)$的短时过零率为：

$$Z_n=\frac{1}{2}\sum_{m=0}^{N-1}|sgn[x_n(m)]-sgn[x_n(m-1)]| \tag{4}$$

3.5 共振峰特征

共振峰是反映声道信息的重要参数，代表了发音信息最

直接的来源。本文采用线性预测法求取第一共振峰。首先估计出预测系数，然后估计声道的功率谱，再用峰值检测法检测出各共振峰的频率。

提取出每句语料的各个特征后，再计算每个特征的 12 个统计量。这些共同组成 384 维的统计特征向量，用来表征对应语句的情感信息。提取的情感特征及相对应的统计量如表 1 所列。

表 1　384 维统计特征

声学特征类别	MFCC（12 维）	基频	短时能量均方根	短时平均过零率	第一共振峰
特征数量	288	24	24	24	24
统计量	最大值、最小值、范围、最大值所在帧、最小值所在帧、均值、线性近似斜率、线性近似偏移、线性近似二次偏差、标准偏差、偏斜度、峰度及其一阶差分的相关统计量				

4　基于 PCA 的特征抽取

表 1 中提取的特征向量属于高维空间，为了降低运算开销同时提高分类正确率，需要利用特征抽取方法将原始空间的特征属性转移到新的低维空间。目前，语音情感分析中常用的特征抽取算法有主成分分析 PCA、线性判别分析（Linear Discriminant Analysis，LDA）、核主成分分析（Kernel PCA）、等度规映射（Isometric Feature Mapping，ISOMAP）等。

主成分分析 PCA 是经典的线性降维算法，其基本思想是在样本空间中寻找若干个变化最为显著的方向，通过投影达到有效降低特征维数的目的。Lee[12] 在分离对话语料的两种情感状态时采用了 PCA 技术；Chuang[13] 采用 PCA 从 33 个声学特征中抽取了 14 个主元特征。

在特征抽取过程中，对于给定的 n 个 d 维的训练样本特征向量 $x_1, x_2, \cdots, x_n$，x_i 均为列向量形式，可将其构成一个 d 行 n 列的数据矩阵 $X_{d\times n}=[x_1, x_2, \cdots, x_n]$，则 PCA 的计算流程如下：

1. 计算特征向量 $x_1, x_2, \cdots, x_n$ 的均值 μ 和协方差矩阵 $COV_{d\times d}$。

2. 计算矩阵 $COV_{d\times d}$ 的本征值和本征向量，每个本征向量都对应一个本征值，组成多个本征值-本征向量对 (λ_i, e_i)，按照本征值从大到小排序为：$(\lambda_1, e_1), (\lambda_2, e_2), \cdots, (\lambda_d, e_d)$，其中 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d \geqslant 0$。

3. 选取前 k 个本征值（$k \ll d$）所对应的本征向量 $e_1, e_2, \cdots, e_k$ 作为主成分方向，即低维子空间的基向量，构造出大小为 $d\times k$ 的映射矩阵 A，其中 A 的第 i 列（$1 \leqslant i \leqslant k$）就是所选取的第 i 个本征向量。

4. 将高维原始数据按照下式投影到低维子空间：$PCA(x)=A^{\mathrm{T}}(x-\mu)$。

根据所选取的主成分本征值之和占协方差矩阵所有本征值总和的百分比大于 95%，确定参数 k 的值为 28。因此，通过 PCA 方法将原来的 384 维的特征向量降为 28 维。由于特征向量中各维元素的单位不统一，因此在主成分分析前对所有的特征向量做了归一化处理。本文的归一化处理方法是把各维元素都化为均值为 0、方差为 1 的正态分布参数。

5　SVM 分类模型的构建

常用的分类模型的构建方法有支持向量机 SVM、K 近邻法（K-Nearest Neighbor，KNN）、隐马尔可夫模型 HMM、高斯混合模型 GMM、人工神经网络（Artificial Neural Network，ANN）。本文采取的分类算法是 SVM，它在解决小样本数据集、非线性以及高维模式识别问题中表现出特有的优势。

SVM 是建立在结构风险最小化准则的基础上的，它根据有限的样本信息，通过对推广误差上界的最小化达到最大的泛化能力。对于线性可分的样本空间，该算法寻找最优分类超平面，能够同时最小化经验误差与最大化几何边缘区，最优分类超平面能够尽可能多地将两类样本正确地分离，同时使分离的两类样本距离超平面最远，这是一个受限的二次规划问题求解。

对线性可分的样本集 (x_i, y_i)，$i=1,2,\cdots,n$，$x \in R^d$，$y \in [-1,+1]$，利用 Lagrange 优化方法可以把最优分类面问题转化为对偶问题。即在约束条件

$$\sum_{i=1}^{n} y_i \partial_i = 0 \tag{5}$$

和 $\partial_i \geqslant 0$，$i=1,2,\cdots,n$ 下，对 ∂_i 求解下列函数的最大值：

$$Q(\partial)=\sum_{i=1}^{n}\partial_i - \frac{1}{2}\sum_{i,j=1}^{n}\partial_i\partial_j y_i y_j (x_i \cdot x_j) \tag{6}$$

其中，∂_i 为与每个样本对应的 Lagrange 算子。若 ∂_i^* 为目标函数最优解，则解得最优分类函数为：

$$f(x)=\mathrm{sgn}(wx+b)=\mathrm{sgn}\left(\sum_{i=1}^{n}\partial_i^* y_i (x_i \cdot x)+b^*\right) \tag{7}$$

由于传统的 SVM 是针对二分类问题的，在本文中对二类分离 SVM 模型进行扩展，构造 15 个 SVM 子分类器。每个子分类器负责分离其中的两个类别。比如分离类别 i 和 j 的子分类器将属于类别 i 的数据标记为 1 类，属于类别 j 的数据标记为 2 类。测试时，每个子分类器都对测试数据进行分类标记，最后的结果由所有的子分类器参加投票，得票最高的类别就是测试数据的类。

本文通过非线性变换将原问题转化为线性问题，转化过程采用 RBF 核函数：

$$R_i(P)=\exp\left(-\frac{\|P-C_i\|^2}{\sigma_i^2}\right) \tag{8}$$

对于参数值的选取，本文根据多次实验测试，选取实验结果最好的情况：惩罚因子 $C=10$，参数 $r=0.01$。

6　实验结果与分析

为了验证生成模型的可靠性和实用性，本文采用 10 折交叉验证。重复实验 10 次，每次将 90% 的数据用于训练，而 10% 留作测试数据。因此，每次实验中，1080 个样本用于训练，120 个样本用于测试。

经过测试，实验得到的结果如表 2 所列。从表 2 中的对角线上观察到，6 种情感的识别率分别为 76.0%、73.5%、64.5%、61.5%、66.5% 和 70.5%，都处于合理的范围内。另外可以看到，生气与高兴、惊讶间的误判率，还有害怕与难过间的误判率，是相对较高的。原因是这两组情感之间相似度较高，以至于容易发生误判。

表 2　测试样本识别结果

测试样本	中性	生气	害怕	高兴	难过	惊讶
中性	76.0%	0%	6.0%	18.0%	0%	0%
生气	4.0%	73.5%	3.0%	9.0%	2.5%	8.0%
害怕	2.5%	8.0%	64.5%	5.0%	17.0%	3.0%
高兴	9.0%	12.0%	4.0%	61.5%	3.5%	10.0%
难过	4.0%	0%	21.0%	8.5%	66.5%	0%
惊讶	0%	15.5%	3.0%	8.0%	3.0%	70.5%

本文选取了其它一些有代表性的研究结果与上述实验方法及结果进行了对比，如表3所列。从表3可以看出，各个研究使用的语料库、语音情感特征、特征维数和分类模型都不尽相同。在有使用普通话情感语料的研究中，本文的平均识别率68.8%处于中等水平。在这些研究当中，本文最终选取的特征维数属于中等，但平均识别率还是较高，处于中上水平。通过对比表明本文的方法及实验结果是有效的。

表3 测试样本识别结果对比

作者	语料库	特征	特征维数	分类模型	识别率
Origlia[2]	包含意大利语、法语、英语、德语等语料库	基频、能量等	31	MLP-ANN	60%
Seppänen[3]	芬兰语	基频、能量、时长等	43	KNN	60%
Li[4]	SUSAS语料库	MFCC、频率微扰、振幅微扰等	26	HMM	69.1%
Nwe[5]	缅甸语和普通话	LFPC、MFCC、LPCC等	28	HMM	77.1%
Luo[14]	Berlin语料库	基频、强度、共振峰、语音停顿、信噪比、基频微扰、振幅微扰等	9	SVM	67.5%
You[15]	普通话情感语料库CASIA	能量、基频、共振峰等	29	SVM	68.3%
本文	普通话情感语料库CASIA	MFCC、基频、短时能量、短时过零率、共振峰	28	SVM	68.8%

从表3看出，本文与You[15]选取的语料库相同，但特征不完全相同，特征维数和分类模型基本一致。为了有更强的对比和更深入的研究，本文与You[15]的方法进行详细对比，结果如表4所列。

表4 识别结果详细对比

作者	中性	生气	害怕	高兴	难过	惊讶
本文	76.0%	73.5%	64.5%	61.5%	66.5%	70.5%
You[15]	77.6%	71.7%	66.0%	57.1%	68.2%	69.3%

从表4中观察到，两个研究中6种情感识别率之间互有高低，但本文的6个识别率之间相差较小，并且总识别率略高。原因是本文选取了更为合理有效的情感特征。另外，虽然两者都是采用SVM构建分类模型，但参数设定不一样。

结束语 本文在中科院的普通话情感语料库上进行语音情感识别实验，并与相关成果进行了比较。结果表明，本文选取的特征值是合理的，进行的抽取方法是有效的、可行的。在今后的研究工作中可以提取更多不同、有效的特征值并选取不同的抽取方法进行实验，对于相似度较高的情感可以有针对性地研究，减少相互的误判率，以达到更高的识别率。

参考文献

[1] 韩文静，李海峰，阮华斌，等. 语音情感识别研究进展综述[J]. 软件学报，2014，25(1)：37-50
Han Wen-jing, Li Hai-feng, Ruan Hua-bin, et al. Review on speech emotion recognition[J]. Journal of Software, 2014, 25(1):37-50

[2] Origlia A, Galata V, Ludusan B. Automatic classification of emotions via global and local prosodic features on a multilingual emotional database[C]// Proc. of the 2010 Speech Prosody. Chicago, 2010

[3] Seppänen T, Vayrynen E, Toivanen J. Prosody-Based classifycation of emotions in spoken finnish[C]//Proc. of the 2003 European Conf. on Speech Communication and Technology (EUROSPEECH). Geneva: ISCA, 2003: 717-720

[4] Li Xi, Tao Ji-dong, Johnson M T, et al. Stress and emotion classification using jitter and shimmer features[C]// Proc. of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Honolulu: IEEE Computer, 2007: 1081-1084

[5] Nwe T L, Foo S W, Silva L C D. Speech emotion recognition using hidden Markov models[J]. Speech Communication, 2003, 41(4): 603-623

[6] Breazeal C, Aryananda L. Recognition of affective communicative intent in robot-directed speech[J]. Autonomous Robots, 2002, 12(1): 83-104

[7] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks[J]. Neural Computing & Applications, 2000, 9(4): 290-296

[8] Institute of Automation, Chinese Academy of Sciences. CASIA Mandarin emotional corpus [DB/OL]. http://www.chinese-ldc.org/resource_info.php? rid=76Casis

[9] Benesty J, Sondhi M M, Huang Y. Springer Handbook on Speech Processing[M]. Berlin: Springer-Verlag, 2008

[10] Gobl C, Chasaide A N. The role of voice quality in communicating emotion, mood and attitude[J]. Speech Communication, 2003, 40(1/2): 189-212

[11] 赵力. 语音信号处理(第2版)[M]. 北京：机械工业出版社，2011：51-52
Zhao Li. Speech signal processing(Second Edition)[M]. Beijing: China Machine Press, 2011: 51-52

[12] Lee C M, Narayanan S S, Pieraccini R. Classifying emotion-ns in human-machine spoken dialogs[C]//Proceedings of IEEE International Conference on Multimedia and Expo. Lusanne, Switzerland, 2002: 737-740

[13] Chuang Ze-jing, Wu Chung-hsien. Emotion recognition using acoustic features and textual content[C]// Proceedings of IEEE International Conference on Multimedia and Expo. Taipei, Taiwan, 2004: 53-56

[14] 罗宪华，杨大利，徐明星，等. 面向非特定人语音情感识别的PCA特征选择方法[J]. 计算机科学，2011，38(8)：212-213
Luo Xian-hua, Yang Da-li, Xu Ming-xing, et al. PCA Based Feature Selection Algorithm on Speaker-independent Speech Emotion Recognition[J]. Computer Science, 2011, 38(8): 212-213

[15] 尤鸣宇. 语音情感识别的关键技术研究[D]. 杭州：浙江大学，2007
You Ming-yu. Research on Key Techniques of Speech Emotion Recognition[D]. Hangzhou: Zhejiang University, 2007