语音信号中的情感识别研究*

赵 力[1], 钱向民[2], 邹采荣[1], 吴镇扬[1]

[1](东南大学 无线电工程系,江苏 南京 210096);

[2](南京航空航天大学 电子工程系,江苏 南京 210016)

E-mail: zhaoli@seu.edu.cn

http://www.seu.edu.cn

摘要: 提出了从语音信号中识别情感特征的方法.从5名说话者中搜集了带有欢快、愤怒、惊奇和悲伤的情感语句共300句.从这些语音资料中提取了10个情感特征.提出了3种基于主元素分析的语音信号中的情感识别方法.使用这些方法获得了基本上接近于人的正常表现的识别效果.

关键词: 情感识别;语音信号;韵律特征;主元素分析

中图法分类号: TP391 **文献标识码**: A

随着计算机多媒体技术的不断发展,能处理包含在媒体中的情感信息的柔软的拟人化的多媒体计算机系统的研究越来越引起人们的兴趣.因为语音信号既是多媒体人机交互的主要利用方式,又是传载情感信息的重要媒体,所以,包含在语音信号中的情感信息的计算机处理研究就显得尤为重要.

分析和处理语音信号中的情感特征、判断说话人的喜怒哀乐等方面的研究事例以及相应的研究成果目前还很少[1].我们针对含有欢快、愤怒、惊奇、悲伤这4种情感的语音信号,分析了反映这些情感信息的物理特征,并利用这些特征参数采用多变量分析手法进行了情感识别的初步尝试,得到了令人满意的实验结果.利用该方法,不仅达到了较高的识别率,而且几乎不占用时间和空间资源,具有较强的实用价值.

1 情感识别用语音资料的选择

选择合适的情感分析用语音资料具有重要意义.然而,目前用于情感分析的语音资料的分析条件和标准还没有被提出[2].在我们的情感分析实验中,对实验用语句的选择主要考虑了以下两个方面:第1,所选择的语句必须不包含某一方面的情感倾向;第2,必须具有较高的情感自由度,对同一个语句能施加各种情感进行分析和比较.根据这两个原则,我们选用4个语句作为情感分析用语音资料,见表1.划分情感类型也应该是情感分析研究的一个重要部分.目前从心理学的角度,国外的资料有不同的划分方法[3],然而,从工学处理的角度进行划分在国内外还很少作为课题被提出.本研究作为初步探索,把情感类型粗略划分为欢快、愤怒、惊奇、悲伤这4种.为了获得原始的语音数据,我们让5名善于表演的说话者按表1所示的语句用欢快、愤怒、惊奇、悲伤这4种情感各发音3

* **收稿日期**: 1999-10-15; **修改日期**: 2000-03-23

基金项目: 国家自然科学基金资助项目(69871009)

作者简介: 赵力(1958-),男,江苏南京人,博士,副教授,主要研究领域为语音信号处理;钱向民(1972-),男,陕西西安人,博士,主要研究领域为情感语音信号的特征提取、识别与合成;邹采荣(1950-),男,江苏昆山人,博士,教授,博士生导师,主要研究领域为数字信号处理;吴镇扬(1947-),男,江苏扬州人,教授,博士生导师,主要研究领域为数字信号处理.

遍,同时再让每位说话者用不带感情的尽可能平静的方式将每一语句各发音 3 遍,这样共搜集到 300 个实验用语句,其中 200 句作为训练用语句,100 句作为识别用语句.

Table 1 Emotional texts chosen for experiment

表 1 实验用情感语句

Text① 1	啊,下雨了
Text 2	你真伟大呀
Text 3	快点干
Text 4	这下全完了

①语句.

为了检验所搜集的实验用情感语音数据的有效性,我们做了听取实验.我们要求以上 5 名发音者以外的另 3 名实验者坐在电脑终端前,然后随机播放所搜集到的带有各种情感的语句,让实验者通过主观评价说出所放语音的情感类型.对一些情感类型不明显的语句进行删除和重新制做,最终 4 种情感实验语句的听取实验结果见表 2.我们可以用这个结果和本文后面介绍的方法得出的识别结果做粗略的对比.

Table 2 Listening experimental results to chosen emotional texts

表 2 实验用情感语句的听音实验结果

Emotion①	Happiness②	Anger③	Surprise④	Sorrow⑤	Error rate⑥(%)
Happiness	53	2	4	1	11.7
Anger	4	52	2	2	13.3
Surprise	7	2	51	0	15
Sorrow	2	0	1	57	5

①情感类别,②喜,③怒,④惊,⑤悲,⑥错误率.

2 情感特征量的选择和提取

一般来说,语音中的情感特征往往通过语音韵律的变化表现出来[3,4].例如,当一个人发怒的时候,讲话的速率会变快,音量会变大,音调会变高,这些都是人们直接可以感觉到的.另外,由于语音信号中的情感信息多少受到语句词汇内容的影响,所以,为了使分析结果消除这方面的影响,一般都是通过分析情感语音和不带情感的平静语音的相对关系,找出这种相对特征的构造、特点和分布规律,用来处理和识别不同的情感语音信号.本文首先按 12KHz,16bit 对输入信号进行 A/D 变换,然后对抽样信号加上窗长 23.22ms(256 点),窗移 10ms 的汉明窗.为了尽可能地利用语音信号中所包含的有关情感方面的信息,我们选取了语句发音持续时间、平均基音频率、最大基音频率、基音频率的平均变化率、平均振幅能量、振幅能量的动态范围、共振峰频率的平均值、共振峰频率的平均变化率、共振峰峰值点回归直线的平均斜率以及共振峰峰值的平均值这 10 个情感特征,作为情感识别用参数.

2.1 语句发音持续时间

我们计算出每一情感语句从开始到结束的持续时间.这个时间包括句中的无声部分,因为无声部分本身对情感是有贡献的.在识别时,我们把情感语句的持续时间和相应的平静语句持续时间的比值作为识别用特征参数.

2.2 基音频率

我们利用倒谱法逐帧求出基音频率,并对基频曲线进行中值滤波和线性平滑处理[5],然后提取

情感信号基频轨迹曲线的的最大值、整个曲线的基频平均值以及平均变化率等特征.这里的基频平均变化率是指各帧语音信号基频的差分的绝对值的平均值.在识别时,我们把情感语句的基频平均值、最大值和相应的平静语句的基频平均值、最大值的差值,以及情感语句的基频变化率和相应的平静语句的基频变化率的比值作为识别用特征参数.

2.3 振 幅

在本文中,我们主要针对振幅平均能量以及动态范围等特征进行分析比较.为了避免发音中无声部和噪音的影响,我们只考虑短时能量超过某一阈值时的振幅的绝对值的平均值.在识别时,我们把情感语句的振幅平均能量、动态范围和相应的平静语句的振幅平均能量、动态范围的差值作为识别用特征参数.

2.4 共振峰

共振峰是反映声道特性的一个重要参数.本文首先用线性预测法(LPC)求出14阶预测系数,然后用预测系数估计出声道的频响曲线,再用峰值检出法(peak picking)计算出各共振峰的频率[6].本文分析了第一共振峰频率的平均值、第一共振峰频率的变化率、前4个共振峰峰值点回归直线的平均斜率以及前4个共振峰峰值的平均值.选择情感语句各帧的第一共振峰频率的平均值、前4个共振峰峰值点回归直线的平均斜率以及前4个共振峰峰值的平均值和相应的平静语句的这些参数的差值以及第一共振峰频率的变化率和相应的平静语句的比值作为识别用特征参数.

3 情感识别方法

本文提出了基于多变量解析中主元素分析的3种情感识别方法[7,8].针对N个10维原始特征矢量的训练语句矢量集,首先求出相关矩阵,然后求出相关矩阵的特征值和特征向量,由特征向量可以组成变换阵.对于任一语句的10维原始特征矢量可以利用变换阵转变为主元素特征矢量.变换阵中和一个主元素相对应的向量叫做该主元素的基向量.一般选择前P个主元素作为有效主元素使用.这样,对于一个给定的样本$\vec{X}$,我们可以根据各基向量求出其各个有效主元素.这些有效主元素组成的矢量被用作情感训练和识别用特征矢量.另外,由于训练和识别用原始样本矢量中各维元素的单位不统一,所以,在主元素分析之前应该做归一化处理.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.

3.1 识别方法1

首先由主元素分析,把每一个训练用D维矢量$\vec{X}_i=\{x_{i1},x_{i2},\ldots,x_{iD}\}$变换成由有效主元素组成的矢量$\vec{Y}=\{y_{i1},y_{i2},\ldots,y_{ip}\}$,$p\leqslant D$.然后,分别对各情感类别求出有效主元素特征矢量集的重心$\vec{\mu}_k$和相应的方差.这样,对于某一个语音情感主元素特征矢量$\vec{Y}$,由式(1)求出它与其他各类别的马氏距离,距离最近的情感类别即为识别结果.

$$D_k=(\vec{Y}-\vec{\mu}_k)'\Sigma^{-1}(\vec{Y}-\vec{\mu}_k) \tag{1}$$

3.2 识别方法2

利用混合高斯分布模型(GMM)进行识别.GMM是只有一个状态的模型,在这个状态里具有多个高斯分布函数.作为一个例子,对于一个特征矢量,其累积概率可由如式(2)所示的三混合的高斯分布模型求得.

$$P_k=w_1f_1(\vec{Y})+w_2f_2(\vec{Y})+w_3f_3(\vec{Y}) \tag{2}$$

在式(2)中,$f_i(\cdot)(i=1,2,3)$是高斯分布函数,$w_1+w_2+w_3=1$是权系数.在训练阶段,首先由矢量量化法(vector quantity,简称VQ)求出各情感类别有效主元素矢量集$\vec{Y}$的码本,并对每一码字$\vec{C}_f=(c_{f_1},c_{f_2},\ldots,c_{f_p})'(f=1,2,\ldots)$求出相应的方差$\vec{\sigma}_f=(\sigma_{f_1},\sigma_{f_2},\ldots,\sigma_{f_p})'(f=1,2,\ldots)$.这样,每一码字和相应的方差即可组成一个高斯分布函数.在本文中,高斯分布函数中的协方差矩阵采用对角阵.在识别时,对于某一个语音情感主元素特征量$\vec{Y}$,由式(2)求出其针对各类别的概率值,概率最大的情感类别即为识别结果.

3.3 识别方法3

为了更合理地选择有效主元素,我们利用主元素基向量,求出任意抽样$\vec{X}_i$的各主元素的得分值.例如,第j主元素的得分值Z_{ij},如式(3)所示.

$$Z_{ij}=\sum_{k=1}^{D}\frac{a_{jk}(x_{ik}-\mu_k)}{\sqrt{S_{kk}}} \tag{3}$$

这里,a_{jd}是第j主元素基向量$\vec{A}_j=(a_{j1},\ldots,a_{jd},\ldots,a_{jD})'$的元素,$\mu_k$是$N$个$\vec{X}_i$中第$k$维元素的均值,$S_{kk}$是第$k$维元素的方差.然后,按式(4)~(6)分别求出$L_j$,$M_j$和$H_j$值.这里,$\mu_{|kj|}$是情感类别$k$的第$j$主元素得分的平均值,$\sigma_{|kj|}$是情感类别$k$的第$j$主元素得分值的标准偏差.所以,$L_j$反映了第$j$主元素在情感类别间的分离性,$M_j$反映了第$j$主元素在各自情感类别中的集中性,$H_j$则反映了第$j$主元素在情感类别中的辨别性.$H$越大,所选该主元素参数的辨别性能就越好.我们对应各主元素分别求出L,M和H,然后按H的大小顺序选择识别用有效主元素.

$$L_j=\frac{1}{8}\sqrt{(\mu_{|aj|}-\mu_{|dj|})^2+(\mu_{|aj|}-\mu_{|suj|})^2+\ldots+(\mu_{|soj|}-\mu_{|hj|})^2} \tag{4}$$

$$M_j=\frac{1}{4}(\sigma_{|aj|}+\sigma_{|suj|}+\sigma_{|soj|}+\sigma_{|hj|}) \tag{5}$$

$$H_j=\frac{L_j}{M_j} \tag{6}$$

判别时,我们求出输入特征矢量$\vec{X}_i$各有效主元素的得分值,对于情感类别k的第j主元素的平均值$\mu_{j|k|}$和标准偏差$\sigma_{j|k|}$,可按式(7)求出得分值Z_{ij}的输出概率.

$$P_{j,|k|}(Z_{ij})=\frac{1}{\sqrt{2\pi}\sigma_{j,|k|}}\exp\left(-\frac{(Z_{ij}-\mu_{j,|k|})^2}{2\cdot\sigma_{j,|k|}^2}\right) \tag{7}$$

各情感类别的各有效主元素的综合概率可按式(8)计算.

$$P_{i,|k|}=\prod_{j=1}^{p}P_{j,|k|}(Z_{ij}) \tag{8}$$

满足式(9)的综合概率最大的类别k即为判别结果.

$$\max_{\text{all }k}(P_{i,|k|}) \tag{9}$$

4 识别实验和结果

我们利用上述3种情感识别方法,针对100句情感测试语句进行了情感识别实验.识别结果见表3.

Table 3 Recognition rates achievec by using above three methods (%)

表 3 3种识别方法的识别结果 (%)

Emotion①	Happiness②	Anger③	Surprise④	Sorrow⑤	Average⑥
Method⑦ 1	80	85	70	100	83.75
Method 2	80	90	80	100	87.5
Method 3	85	95	80	100	90

①情感类型,②喜,③怒,④惊,⑤悲,⑥合计,⑦识别方法.

从表3可以看出,3种判别方法的识别率都在80~90%左右.其中,识别方法3的性能最好,方法2次之.3种方法对"欢快"和"惊奇"的识别率比其余两种情感要低,而对"悲伤"做到了完全识别.

5 结 论

本文提出了3种基于主元素分析的语音情感类别识别方法.经过对100句情感测试语句的识别实验结果表明,使用这些方法,获得了基本上接近于人的正常表现的识别效果,证明了本文提出的方法的有效性.另一方面,实验中采用的说话者仅限于男性,而且人数、实验语句数也偏少,所选择的韵律等情感特征参数对"欢快"和"惊奇"的识别效果还不是很理想.今后的工作主要集中在寻找更为有效的情感特征参数和识别方法,在更广的范围进行进一步的识别实验.

References:

[1] Niimi, Y. Emotional Robot World. Tokyo: Talk and Speak Press, 1995. 67~96.

[2] Muraka, S. Emotional constituents in text and emotional components in speech [Ph. D. Thesis]. Kyoto: Kyoto Institute of Technology, 1998.

[3] Kawanami, H. Considerations on the prosodic features of utterances with attitudes and emotions. Technical Report, sp97', Kokyo: Institute of Electronics, Information and Communication Engineers, 1997.

[4] Cowie, R., Douglas, E. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: IEEE ed. Proceedings of the 14th International Congress of Phonetic Sciences, Vol 3. San Francisco: Academic Press, 1999. 2327~2330.

[5] Zhao, Li, Kobayashi, Y., Niimi, Y. Tone recognition of Chinese continuous speech using continuous HMMs. Journal of the Accustical Society of Japan, 1997,53(12):933~940.

[6] Zhou, Di-wei. Computer Speech Signal Processing. Beijing: National Defense Industry Press, 1987. 130~146 (in Chinese).

[7] Wang, Xue-ren, Wang, Song-gui. Practical Multivariate Statistical Analysis. Shanghai: Shanghai Science and Technology Press, 1995. 150~187 (in Chinese).

[8] Tang, Shou-zheng. Multivariate Statistical Analysis Methods. Beijing: China Forestry Press, 1987. 20~37 (in Chinese).

附中文参考文献:

[6] 周迪伟.计算机语音处理.北京:国防工业出版社,1987.130~146.

[7] 王学仁,王松桂.实用多元统计分析.上海:上海科学技术出版社,1995.150~187.

[8] 唐守正.多元统计方法.北京:中国林业出版社,1987.20~37.

A Study on Emotional Recognition in Speech Signal *

ZHAO Li[1], QIAN Xiang-min[2], ZHOU Cai-rong[1], WU Zhen-yang[1]

[1](*Department of Radio Engineering, Southeast University, Nanjing* 210096, *China*);

[2](*Department of Electronic Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing* 210016, *China*)

E-mail: zhaoli@seu.edu.cn

http://www.seu.edu.cn

Abstract: In this paper, some methods are proposed to discriminate utterances from the speech signal. A corpus containing emotional speech of happiness, anger, surprise and sorrow with over 300 utterances from five speakers is recorded. Ten emotional features are extracted from the speech signal. Three emotion recognition methods are introduced based on principal component analysis. Using these methods, recognition performance is obtained, which is close to human performance on the task.

Key words: emotional recognition; speech signal; prosodic feature; principal component analysis

* Received October 15, 1999; accepted March 23, 2000

Supported by the National Natural Science Foundation of China under Grant No. 69871009