语音信号中的情感特征分析和识别的研究

赵 力[1,2],将春辉[1,2],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.东南大学学习科学研究中心,江苏南京 210096)

摘 要: 提出了一种利用全局和时序结构的组合特征以及MMD进行情感特征识别的方法.对于从10名话者中搜集的带有欢快,愤怒,惊奇和悲伤4种情感的1000句语句,利用提出的新的识别方法获得了94%的平均情感识别率.

关键词: 语音信号;情感特征分析;MMD;情感识别
中图分类号: TP391.42 **文献标识码:** A **文章编号:** 0372-2112 (2004) 04-0606-04

A Study on Emotional Feature Analysis and Recognition in Speech

ZHAO Li[1,2], JIANG Chun-hui[1,2], ZOU Cai-rong[1], WU Zhen-yang[1]
(1. *Dept. of Radio Engineering, Southeast University, Nanjing, Jiangsu* 210096, *China*;
2. *Research Center of Learning Science, Southeast University, Nanjing, Jiangsu* 210096, *China*)

Abstract: This paper presents a new emotion recognition method based on the hybrid feature and MMD. The hybrid feature was composed of global and time sequence. Experiments were conducted on a task of 10 speakers' 1000 sentences including happy, angry, surprising and sorrowful emotions to demonstrate has effectiveness of the new method. The average emotion recognition rate was reached as higher as 94%.

Key words: speech signal; emotional feature analysis; MMD; emotion recognition

1 引言

语言是人类交际的最重要的交流工具.人类的话语中不仅包含了文字符号信息,而且还包含了人们的感情和情绪等信息.因此,情感信息特征的人工处理,在信号处理和人工智能等领域具有重要意义[1].在语音信号情感特征分析和识别的研究中,我们曾利用以整个情感语句为单位的全局特征为识别参数,进行了识别实验,并取得了一定的情感识别效果[2~4].然而可以想象,主要反映情感变化动态特性的时序结构特征,也应该对语音情感识别起重要作用.但是现在国内外对于反映情感语音变化的关键词、短语和时序结构的研究还很少[5].

本文研究了语音信号中情感特征分析和识别的问题.针对含有欢快、愤怒、惊奇、悲伤等4种情感的语音信号,分析了它们的时间构造、振幅构造、基频构造和共振峰构造的特征,并通过和不带感情的平静语音信号进行比较,总结了不同情感信号特征的分布规律.同时,根据这些分析,利用整个句子的全局特征和不同区段的时序韵律特征的组合进行了情感识别实验.针对特征维数的增加,还提出了一种新的修改型马氏距离判别式(MMD:Modified Mahalanobis Distance).对于从10名话者中搜集的1000句情感语句,获得了94%的平均情感识别率.

2 情感分析用语音资料的选择

本文对分析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感情进行分析比较.其次,对于语句的长度、辅音以及助词成分的构成,男女的性别差异等都作了考虑.根据上述原则,我们选用了60个语句作为情感分析用语音资料[3].本文把情感类型粗略划分为欢快,愤怒,惊奇,悲伤等4种,并尽可能将通常认为的所有情感纳入此分类中,我们认为这样的分类对于计算机情感分析的研究是合理的.为了获得原始的语音数据,我们让10名善于表演的男性话者对60个语句用欢快,愤怒,惊奇,悲伤等4种情感各发音1遍,同时再让每位话者用不带感情的尽可能平静的方式将每一语句各发音1遍,这样共搜集到3000个实验用语句.在识别实验时其中2000句作为训练用语句,1000句作为识别用语句.为了检验所搜集的实验用情感语音数据的有效性,我们做了一个听取实验.我们要求以上10名发音者以外的另5名实验者坐在电脑终端前,随机播放

收稿日期:2002-06-10;修回日期:2004-01-06
基金项目:教育部《面向21世纪教育振兴行动计划》及教育部科学技术重点项目(No.03082)

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.经过反复听取比较以及数学上的有意性检定(Mcnemar 检定)[4],对其中情感特征不明显的句子进行了删除和重新制作.

3 情感特征的分析和识别特征提取

一般来说,语音信号中的情感特征往往通过语音韵律的变化表现出来[4,5].例如,当一个人发怒的时候,讲话的速率会变快,音量会变大,音调会变高等,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感情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.本文主要对含有 4 种情感的语音信号的时间构造、振幅构造、基频构造和共振峰构造等特征和不含感情的平静语音信号进行比较,寻找不同情感信号特征的构造特点和分布规律.

3.1 时间构造的分析

分析情感语音的时间构造主要着眼于不同情感语音的发话时间构造的差别,对由情感引起的持续时间等的变化进行分析和比较.本文中我们计算出每一情感语句从开始到结束的持续时间,这一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.然后就情感语句的发话持续时间长度(简称为 T)以及平均发话速率(音节/秒)和情感的关系进行了分析和比较.在发话的持续时间上,愤怒、惊奇的发音长度和平静发音相比压缩了,而欢快、悲伤的发音长度却伸长了.在被压缩的愤怒、惊奇中,愤怒的发音最短,其次是惊奇.欢快和悲伤相比,悲伤伸长很多,而欢快只是稍稍伸长.通过进一步的观察可知,这些现象的产生是由于和平静语音相比,在情感语音中一些音素被模糊地发音、拖长或省略掉了的缘故.

3.2 振幅构造的分析

信号的振幅特征与各种情感信息具有较强的相关性[4,5].在我们的实际活动中也会有这样的感觉,就是当人们愤怒或者惊奇的时候,发音的音量往往变大,而当人们沮丧或悲伤的时候,往往讲话的声音很低.因此,在一些有关情感分析的研究中,振幅构造特性都是作为重要特征加以分析研究的.在本文中,我们主要针对振幅平均能量以及动态范围(分别简称为 A 和 A range)等特征量进行分析比较.我们求取语音信号每一帧的短时能量,分析它们随时间的变化情况.而且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绝对值的平均值.从分析结果可知,欢快、愤怒、惊奇 3 种情感发音信号和平静发音信号相比振幅将变大,相反地,悲伤和平静相比,振幅将减小.而且从听取实验可知,情感信号具有这样的倾向,即,欢快、愤怒、惊奇的平均振幅越大,悲伤的平均振幅越小,其情感效应表现的越明显.利用振幅特征,我们可以很清楚地把欢快、愤怒、惊奇和悲伤区分开来,另外,振幅特性也具有一定的区分欢快、愤怒和惊奇情感信号的能力.

3.3 基频构造的分析

基音频率也是反映情感信息的重要特征之一[4,5].为了分析情感语音信号基频构造的特征,我们首先求出情感语音信号平滑的基频轨迹曲线[6],然后分析不同情感信号基频轨迹曲线的变化情况,找出不同的情感信号各自具有的基频构造特征.本文分析了不同情感信号轨迹曲线的动态范围、整个曲线的基频平均值以及变化率(分别简称为 F_0, $F_{0\,\mathrm{range}}$ 和 $F_{0\,\mathrm{rate}}$)等特征,这里的基频变化率是指各帧语音信号基频的差分的绝对值的平均值.与平静语音信号相比、欢快、愤怒和惊奇的平均基频、动态范围、平均变化率比较大,而相反,悲伤语音信号的则较小.对比较大的欢快、愤怒、惊奇来讲,惊奇语音信号的特征量最大,其次是欢快和愤怒.另外,通过观察语音信号的基频轨迹曲线,我们发现了一个区分惊奇和其它情感信号的重要特征,那就是惊奇情感信号的基频轨迹曲线在句尾的地方往往有上翘的特征.

3.4 共振峰构造的分析

共振峰是反映声道特性的一个重要参数.因为不同情感的发音可能使声道有不同的变化,所以,我们能够预料到不同情感发音的共振峰的位置不同.本文首先用 LPC 法求出声道的功率谱包络,再用峰值检出法(Peak Picking)[7]算出各共振峰的频率.本文只研究了第一共振峰频率平均值,动态范围和变化率(分别简称为 F_1, $F_{1\,\mathrm{range}}$ 和 $F_{1\,\mathrm{rate}}$).通过分析可知相对于平静发音,欢快和愤怒的第一共振峰频率略微地升高了,而悲伤的第一共振峰频率有明显的降低.通过进一步的观察,我们发现,这是因为人们在表达欢快和愤怒时,嘴比平静发音时张得更大的缘故.而在表达悲伤时,除了嘴张得比平时更小以外,还伴有模糊不清的鼻音.4 种情感的第一共振峰频率的动态范围均比平静时要大,其中,惊奇最大.而 4 种情感的第一共振峰频率的变化率均比平静时要小,其中悲伤最小.

3.5 分析结论

综合以上从 4 个方面对含有 4 种情感的语音信号进行的分析比较,我们可以归纳出如表 1 所示的情感信号的特征规律.

表 1　情感语音中各特征参数的变化情况

	T	F_0	$F_{0\,\mathrm{range}}$	$F_{0\,\mathrm{rate}}$	A	A_{range}	F_1	$F_{1\,\mathrm{range}}$	$F_{1\,\mathrm{rate}}$
喜	+	+	+	+	+	+ +	+	+	−
怒	−	+	+	+ +	+	+ +	+	+	−
惊	−	+ +	+ +	+ +	+ +	+ +	−	+	−
悲	+ +	·	·	− −	·	+	− −	+	− −

(上表中符号意义 +:增加 + +:较大增加 −:减小 − −:较大减小 ·:无明显变化)

3.6 识别特征提取

根据以上分析,利用整个句子的全局特征和时序结构特征的组合进行了情感识别实验.本文选取的全局特征是:语句发音持续时间长度、F_0 的平均变化率、F_1 平均变化率和相应的平静语句的值的比值;情感语句平均振幅能量、振幅能量的动态范围、F_0 的平均值、F_0 的最大值、F_1 平均值、F_1 的动态

范围和相应的平静语句的值的差值.

通过对情感语音信号的分析可知,语音信号中的情感信息主要反映在有基音部分的变化上.所以为了提取时序结构特征,我们把语音信号中有基音的部分抽取出来,按时序分割成 M 等份的时间序列,分别选取各区间的基音频率、振幅能量和第一共振峰频率 F_1 这 3 个参数的平均值、最大值和动态范围作为识别用时序结构特征参数.

4 情感识别方法

利用上述情感语句的全局特征和各时序结构特征的组合,可采用式(1)所示的马氏距离进行语音信号中情感特征的识别[8].

$$d^2(X)=(X-\mu)^t\sum^{-1}(X-\mu) \tag{1}$$

这里 $X(x_1,x_2,\cdots,x_p)$是维数为 P 的输入特征向量,μ 是参考样本的均值向量,$\sum$是参考样本的协方差矩阵.在式(1)中,计算$\sum$所必要的乘法次数是 p^2+p,所以当 X 的维数增加时,计算量和内存所占容量会变得很大.而且更重要的是,随着 X 的维数的增大,协方差矩阵的推定误差将增大,从而降低判别的性能.为了避免直接计算$\sum$所引起的计算量和计算误差增大的问题,可以对式(1)进行修改.设式(1)中$\sum$的第 k 个特征值为λ_k,其对应的特征向量为 ϕ_k,则式(1)可变换为如下式(2):

$$d^2(X)=\sum_{k=1}^{P}\frac{1}{\lambda_k+b}(X-\mu,\phi_k)^2 \tag{2}$$

这里偏置 b 是为了缓和由于特征值误差而引起的识别率下降,其值可由实验确定.由于采用整个情感句子的全局特征和各时序韵律特征组合成的情感特征作为识别特征,其特征维数很大,而且训练数据相对较少.所以改良后的马氏距离仍然不足以大幅度降低计算量和提高识别性能.对此,本文提出了一种新的矢量分割型马氏距离判别式.假定标准马氏距离的协方差矩阵可分割成如式(3)所示的形式,其中 $\sum_1,\sum_2,\cdots,\sum_M$ 为 $K\times K$ 的方阵($K\times M=P$),矩阵中其他元素为 0.则标准马氏距离可化成式(4)所示的形式.

$$\sum=\begin{bmatrix}\sum_1 & & & 0\\ & \sum_2 & & \\ & & \ddots & \\ 0 & & & \sum_M\end{bmatrix} \tag{3}$$

$$\begin{aligned}d^2(X)&=(X-\mu)^t\sum^{-1}(X-\mu)\\&=\sum_{i=1}^{M}(X_i-\mu_i)^t\sum_i^{-1}(X_i-\mu_i)\\&=\sum_{i=1}^{M}\sum_{k=1}^{K}\frac{1}{\lambda_{ik}+b}(X_i-\mu_i,\phi_{ik})^2\end{aligned} \tag{4}$$

其中 X_i 和μ_i 分别是向量X 和μ 的第$K_{i-1}+1$ 到 K_i 的元素构成的 k 维向量.λ_{ik}是$\sum_i$ 的第 k 个特征值.ϕ_{ik}是λ_{ik}所对应的特征向量.显然式(4)的计算量 $O(K^2M)=O(PK)$是式(1)计算量 $O(n^2)$的 $1/M$ 倍.由于 P 维的矢量被分割成M 个 k 维的矢量,求取 M 个 $K\times K$ 的协方差矩阵,这样使得对于相同数量的学习数据,维数和学习数据的比是原来的 M 倍,所以求取得协方差矩阵的可靠性进一步提高.

5 识别结果

针对 1000 句情感测试语句利用上述方法进行了情感识别实验,实验中 b 取 1.3.本文首先分析比较了 M 取不同等份时对识别结果的影响.从表 2 可以看出,M 等份取得过细,识别效果并不一定好.这是因为对于一定长度的语句,取得过细将使每一等份的数据量变少.

表 2 情感识别结果[%]

情感类型	喜	怒	惊	悲	平均
M 取 4 等份	89	91	87	98	91
M 取 6 等份	90	92	88	100	93
M 取 8 等份	92	95	90	100	94
M 取 10 等份	89	92	89	98	92

为了进行比较,表 3 给出了几种识别方法的识别结果.方法 1 是只利用全局特征和直接利用式(1)所示的马氏距离进行分析判别的识别结果.方法 2 是全局特征和时序结构特征并用后,直接利用马氏距离进行分析判别的识别结果(M 取 8 等份).方法 3 是全局特征和时序结构特征并用后,利用提出的 MMD 进行分析判别的识别结果(M 取 8 等份).从表 3 可以看出方法 3 的识别效果明显高于其他 2 种方法.

表 3 情感识别结果[%]

情感类型	喜	怒	惊	悲	平均
识别方法 1	84	85	83	95	87
识别方法 2	82	83	79	92	84
识别方法 3	92	95	90	100	94

6 结论

本文对含有欢快、愤怒、惊奇和悲伤 4 种情感的语音信号进行了分析比较,找出了不同情感信号特征的分布规律,并提出了基于 MMD 进行情感特征识别的方法.经过对情感测试语句识别实验结果表明,使用该识别方法获得了基本上接近于人的正常表现的识别效果.今后的工作主要集中在寻找更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分析和识别实验.

参考文献:

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

[2] 赵力,钱向民,邹采荣,吴镇扬.语音信号中的情感识别的研究[J].软件学报 2001,12(7):1050 - 1054.

[3] 赵力,钱向民,邹采荣,吴镇扬.语音信号中的情感特征分析和识别的研究[J].通信学报,2000,12(10):18 - 25.

[4] 赵力,钱向民,邹采荣,吴镇扬.从语音信号中提取情感特征的研究[J].数据采集与处理,2000,15(1):120 - 123.

[5] Cowie R. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 2001, 18(1): 32 - 80.

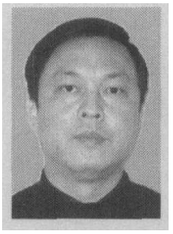
[6] Zhao L, Y Kobayashi, Y Niimi. Tone recognition of Chinese continuous

speech using continuous HMMs. 日本音响学会论文志, 1997, 53 (12):933 - 940.

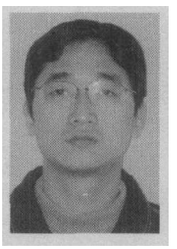
[7] 周迪伟,等,译. 计算机语音处理[M]. 北京: 国防工业出版社, 1987. 139 - 149.

[8] M Shigenaga. Features of Emotionally Uttered Speech Revealed by Discriminant Analysis(Ⅵ)[M]. The preprint of the acoustical society of Japan, 1999. 2 - 18.

作者简介:



赵　力　男, 1958 年 1 月生于江苏南京, 1982 年毕业于南京航空航天大学自动化系, 1988 年取得东南大学工学硕士学位, 1992 年受日本丰桥科学技术大学中川圣一教授的邀请, 赴日参加中日文语音翻译系统项目合作的研究, 1995 年从师于日本京都理工大学(Kyoto Institute of Technology)新美康永教授攻读博士学位, 研究汉语连续语音识别和汉语连续语音声调识别技术, 获工学博士, 1998 年底回国, 现为东南大学无线电工程系教授, 研究语音信号处理、自然语言处理、情感信息处理等, 中国电子学会、日本音响学会、日本电子情报通信学会会员.



蒋睿晖　男, 1977 年 10 月生于江苏泰州, 1999 年毕业于东南大学无线电工程系, 现为东南大学无线电工程系硕士研究生, 研究方向是语音信号处理和情感信息处理.